

# Solved and Unsolved Problems in Chemoinformatics

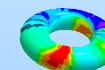
Johann Gasteiger

Computer-Chemie-Centrum

University of Erlangen-Nürnberg

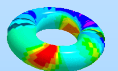
D-91052 Erlangen, Germany

Johann.Gasteiger@fau.de



# Overview

- objectives of lecture
- achievements
- challenges
- unsolved problems
- summary



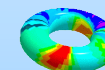
# Objectives

- many problems have been solved

→ let us be proud!

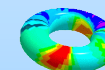
- there is still a lot to be done

→ chemoinformatics is a field of its own,  
is attractive for students



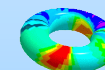
# Chemoinformatics – An Old Discipline

- Structure activity relationships  
1963 Hantsch & Fujita
- structure representation  
1965, Morgan
- structure elucidation  
1965, Sasaki, Munk, DENDRAL
- synthesis design  
1970, Corey & Wipke, Ugi+Gasteiger, Hendrickson
- molecular modeling  
1970, Langridge, Marshall
- data analysis / chemometrics  
1970, Kowalski, Wold



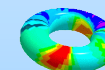
# Achievements

- access to chemical information
- learning from chemical information
- applications

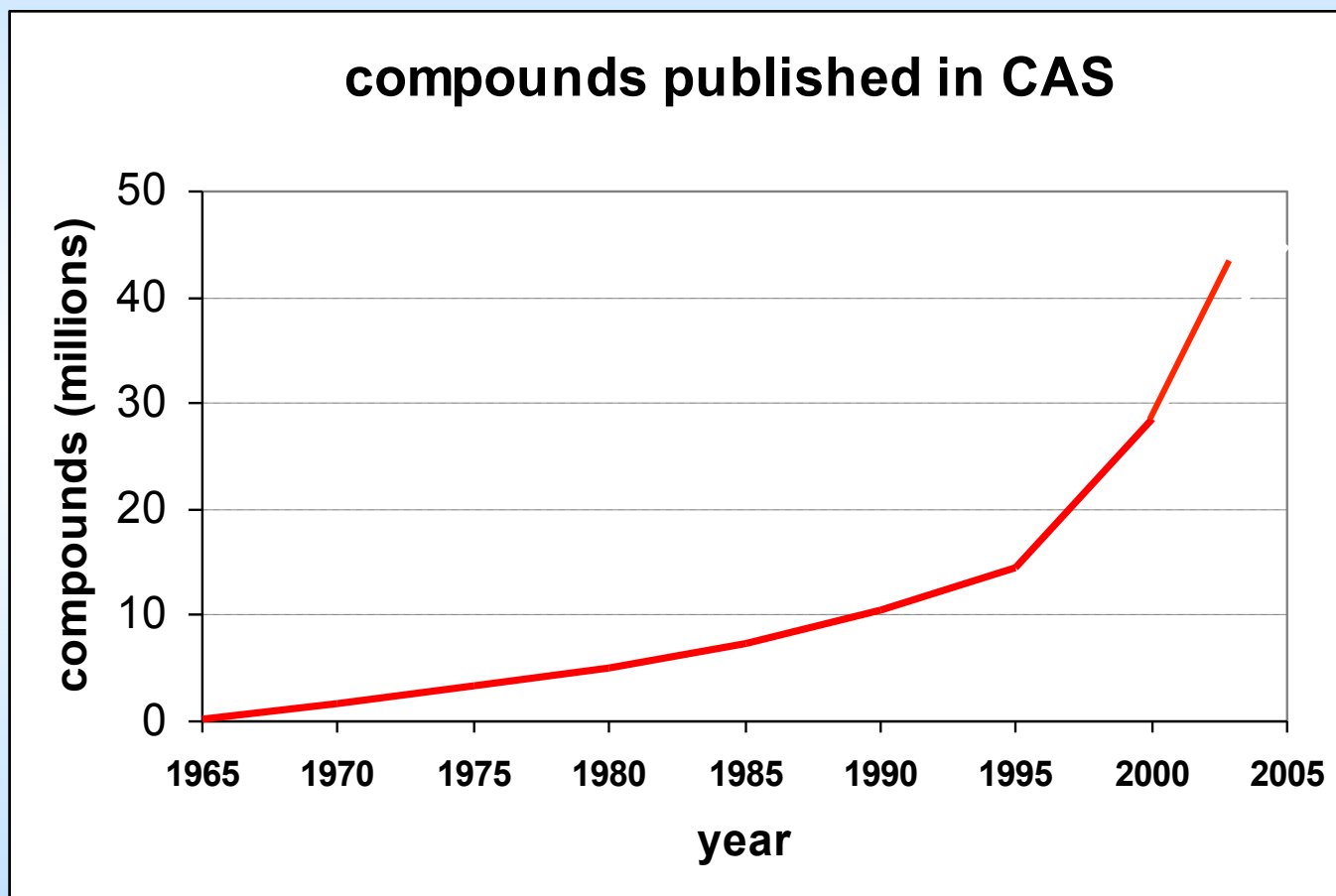


# Databases

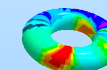
- Chemical Abstracts Service (1975)
- Cambridge Structure Database (1984)
- Beilstein (1990)
- Gmelin (1990)
- ChemInformRX (1991)
- SpecInfo (1991)
- PubChem (2004)
- etc.



# Number of Compounds in Chemistry

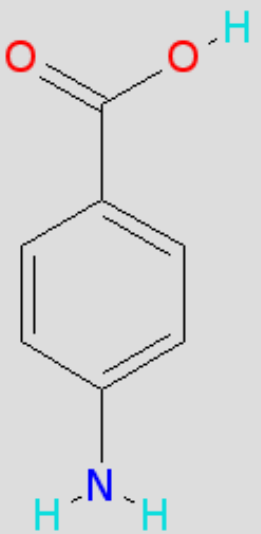


73 million  
compounds,  
64 million  
sequences  
(Sept 2013)



## Database Search ?

**Structure Search** | Upload Molecule File | CAS Registry Number Search | Molecule Name Search





The image shows the chemical structure of 4-aminobenzoic acid. It consists of a central benzene ring. At the top position (12 o'clock), there is a carboxylic acid group (-COOH) with a red oxygen double-bonded to the carbon, and another red oxygen single-bonded to the carbon and a light blue hydrogen atom. At the bottom position (6 o'clock), there is an amino group (-NH2) with a blue nitrogen atom bonded to two light blue hydrogen atoms. The benzene ring is drawn with a hexagon and a circle inside, indicating aromaticity.

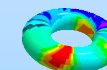
**Search Options**

Full Structure Search

Substructure Search

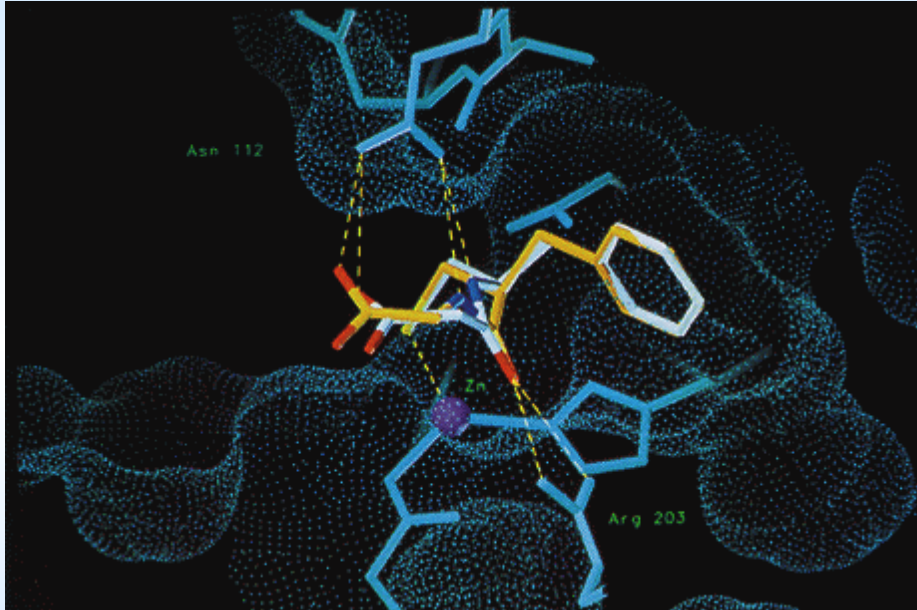
Similarity Search

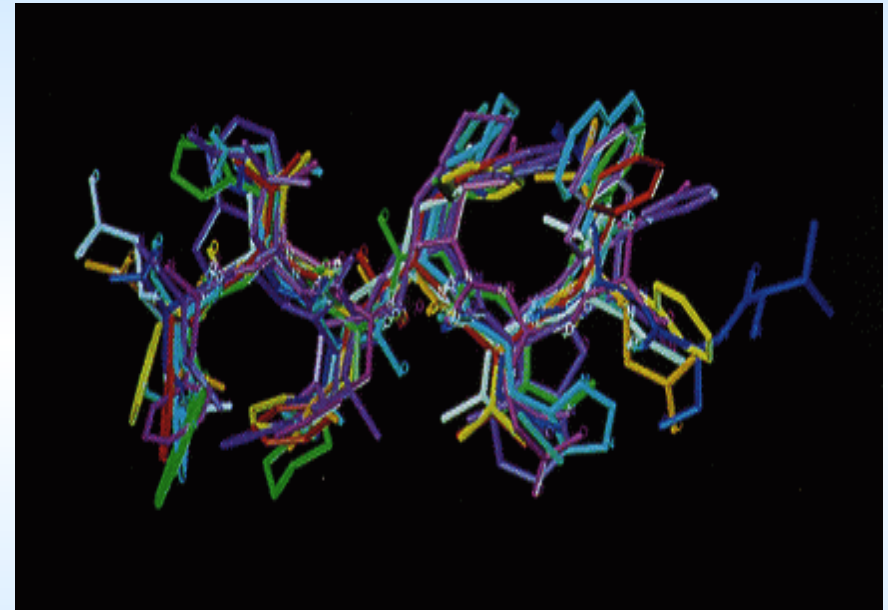




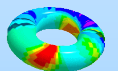
# Search for Cancerostatic Drugs



protein/substrate complex

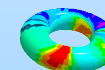


similar substrates



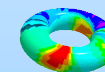
# Chemical Structures

- computers have learnt the language of a chemist
  - communicating by structure diagrams
- molecules are stored with atomic resolution providing access to each atom and bond
  - enabling substructure search
- the complex structures of molecules can be visualized
  - new insights can be gained



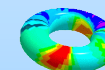
# Summary

- databases have strongly contributed to the **progress** in chemistry and related fields
- without databases **modern research** in chemistry would be inconceivable



# Learning from Chemical Information

- learning from data
- QSAR/QSPR
- representation of chemical structures



# Problem: Not Enough Information

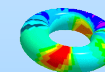
73,000,000

chemical compounds

600,000

3D structures in  
Cambridge Crystallographic Data File

we only have data on the 3D structure for less than  
1% of the known compounds



# Problem: Not Enough Information

73,000,000

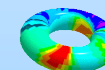
chemical compounds

600,000

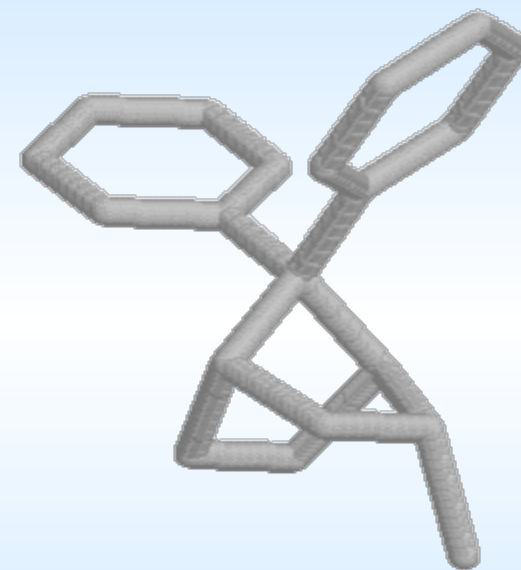
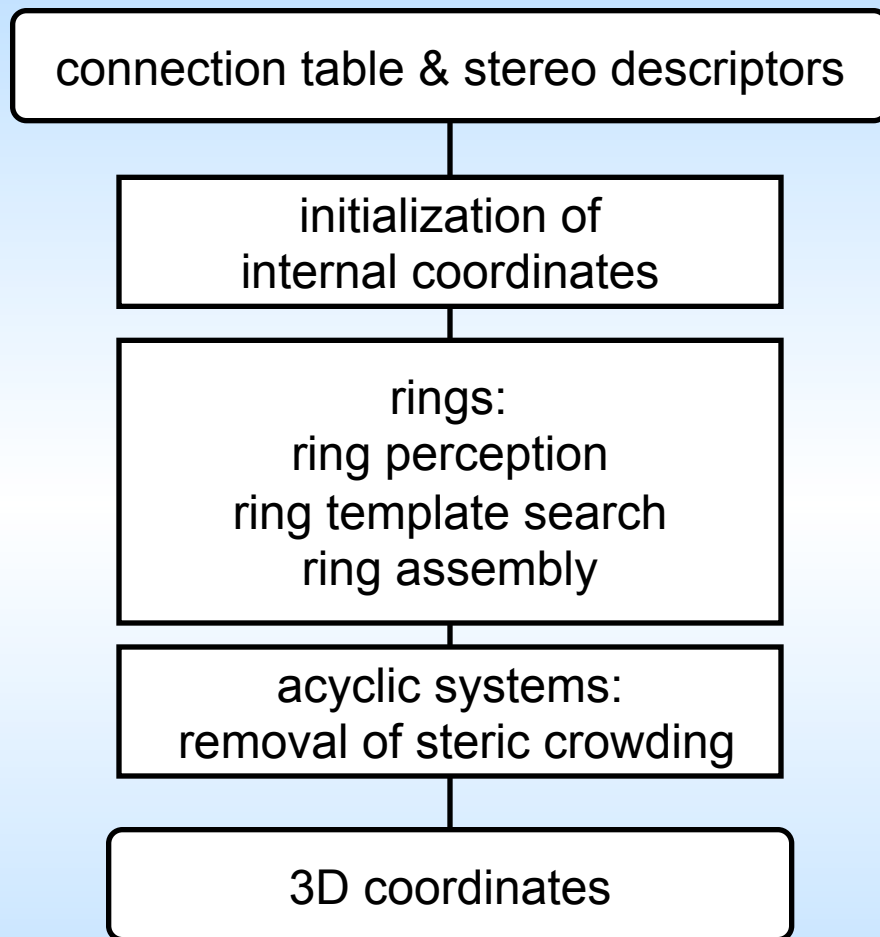
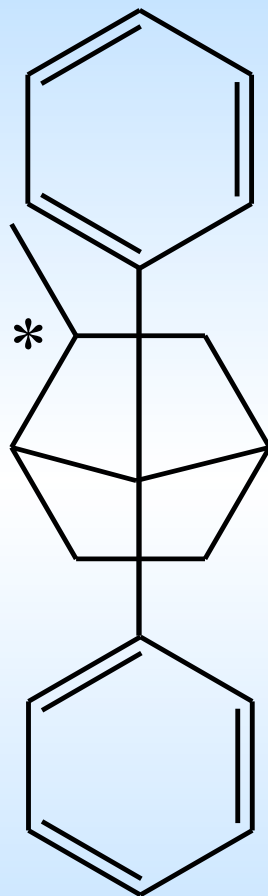
3D structures in  
Cambridge Crystallographic Data File


we have data on the 3D structure for less than  
1% of the known compounds

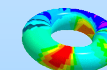
→ can we learn the rules from the known 3D structures



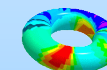
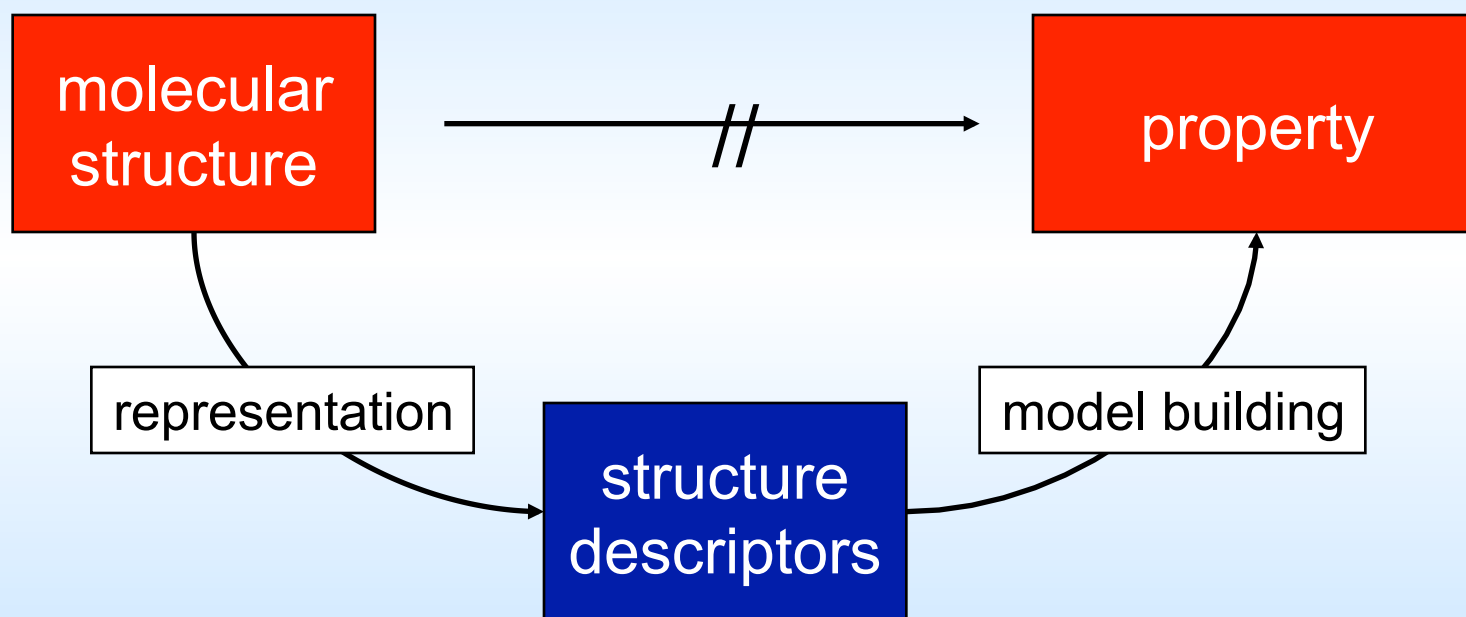
# CORINA: Rule-Based Learning



 generates 3D coordinates for >99.5 % of all organic compounds



# Structure-Property Relationships (QSPR, QSAR) Data-Based Learning



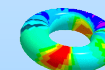


# Representation of Chemical Structures

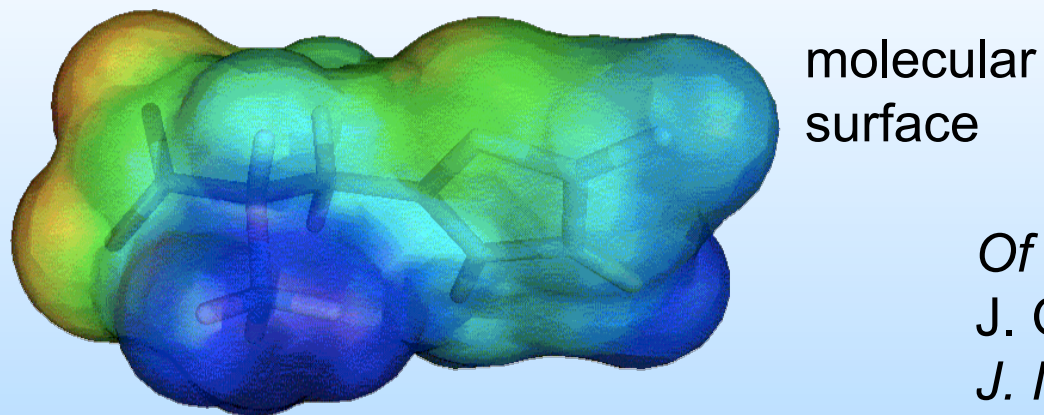
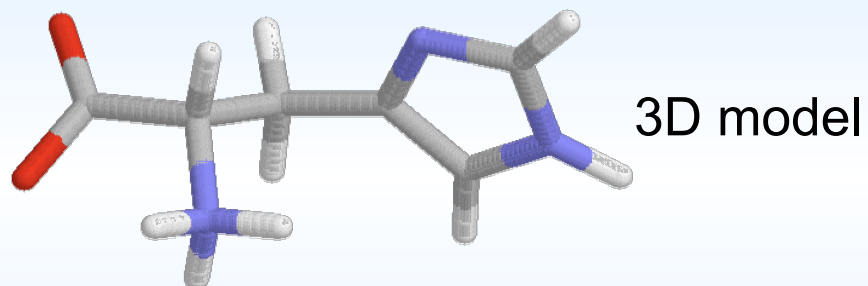
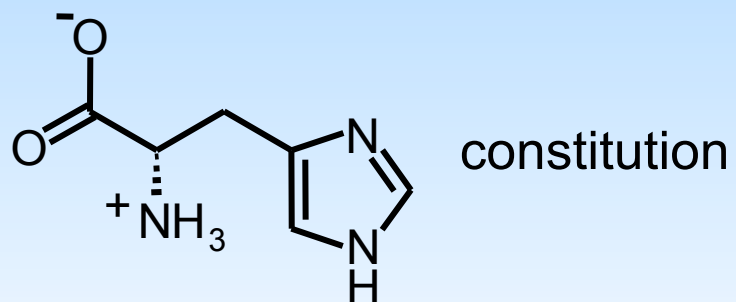
- topological indices
- fragment codes
- fingerprints
- ....

Several thousands of different types of chemical descriptors have been developed

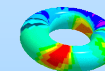
R.Todeschini, V.Consonni, *Molecular Descriptors in Chemoinformatics*, 2 volumes, Wiley-VCH, 2009



# Representation of Chemical Structures

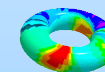


*Of Molecules and Humans*  
J. Gasteiger,  
*J. Med. Chem.*, **2006**, 55, 6429 - 6434



# Prediction of Properties

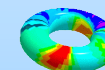
- physical, e.g.
  - aqueous solubility
  - $^{13}\text{C}$  NMR shifts
- chemical, e.g.
  - acidity
- biological, e.g.
  - toxicity

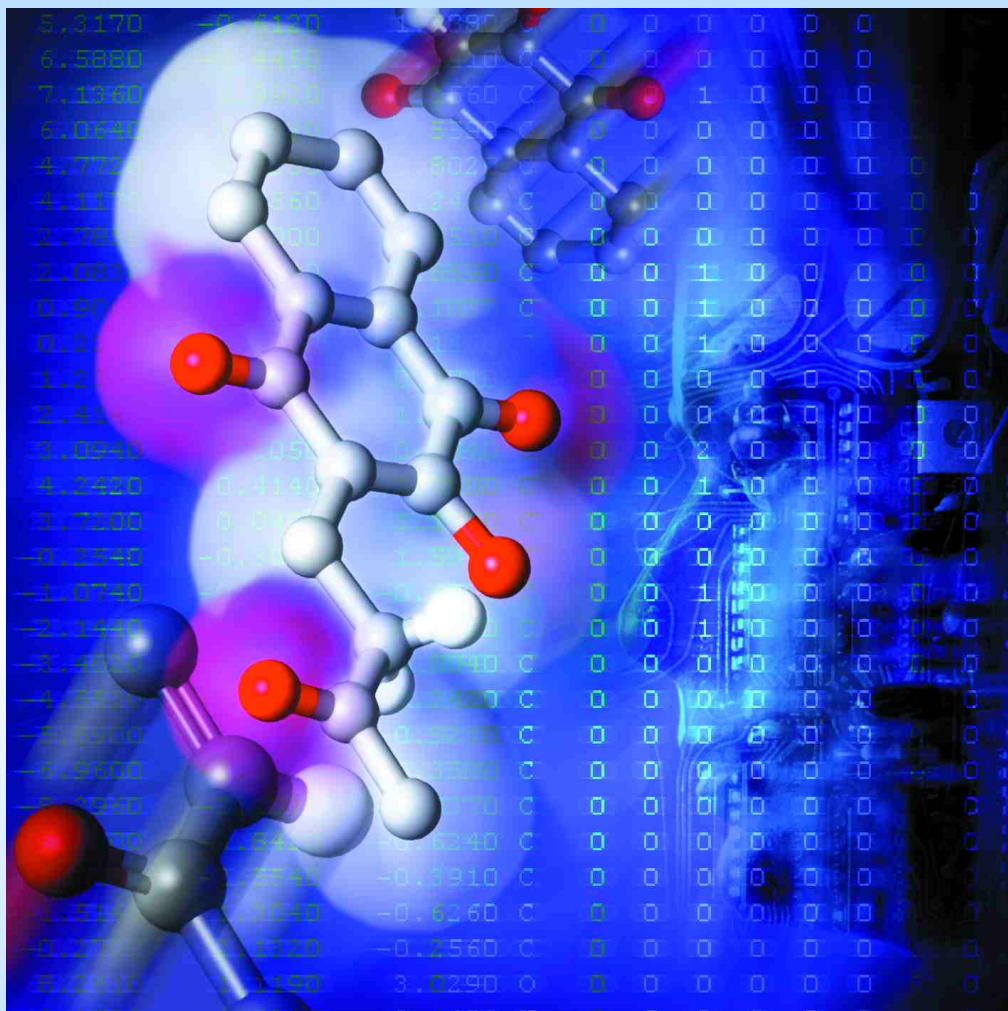


# Applications of Chemoinformatics: Drug Design

Chemoinformatics has become an integral part of the drug design process

- lead discovery
  - virtual screening
  - pharmacophore searching
- lead optimization
  - QSAR
  - molecular docking
- prediction of ADME properties
  - solubility, adsorption, distribution, metabolism, excretion....





# Handbook of Cheminformatics

## From Data to Knowledge

J. Gasteiger (Editor)

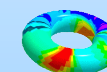
65 authors

73 contributions

4 volumes

1900 pages

Wiley-VCH, Weinheim  
(August 2003)



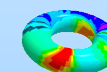


# Chemoinformatics - A Textbook -

J. Gasteiger, T. Engel  
(Editors)

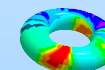
650 pages

Wiley-VCH, Weinheim  
(September 2003)



# Challenges

- applications in all fields of chemistry
- synthesis of properties
- human health
- environmental impact of chemicals
- understanding chemistry
- understanding biological systems



# Synthesis of Properties

The most fundamental and lasting objective of synthesis is not

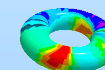
production of new compounds

but

production of properties

George S. Hammond

Norris Award Lecture, 1968





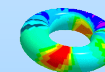
# Fundamental Questions in Chemistry

What structure do I need for a certain property?  
structure-activity relationships

How do I make this structure?  
synthesis design

What is the product of my reaction?  
reaction prediction  
structure elucidation

➔ In all those areas the use of chemoinformatics could help!

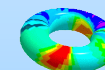


# Risk Assessment of Chemicals

## REACH – Registration, Evaluation, Authorization and restriction of Chemicals

- for those chemicals used with more than 10 tons/year manufactured or imported into the European Union a Chemical Safety Report is needed
- law since June 1, 2007; registration until Dec 1, 2013
- applies to about 35,000 chemicals
- testing on harmful effects on human health or environment, determination of persistence, bioaccumulation and toxicity
- testing is time-consuming, expensive and might need many animals

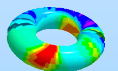
➔ Use chemoinformatics methods for ranking of chemicals



# Cosmetics Directive

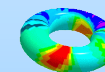
- for chemicals used in cosmetics products
  - no compounds tested on animals are allowed in cosmetics in Europe since 2009.
  - all animal tested cosmetics will eventually be banned on the European market.

→ Use chemoinformatics methods for developing alternatives to animal testing



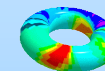
# Problems Still to be Solved

- access to chemical information
- acquisition of chemical information
- learning from chemical information
- applications



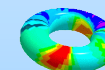
# Access to Chemical Information

- beyond structure editors
  - input by hand-drawing
  - input by voice
- beyond valence bond structures
  - boranes
  - organometallic structures (ferrocene etc.)
  - Markush structures
  - polymers



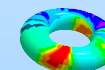
# Acquisition of Chemical Information

- input of chemical structures (hand writing, voice)
- optical character recognition
- text mining
- publishing chemical information (3D structures, spectra)
- publishing and searching on the internet



# Better Databases

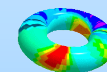
- on compounds
  - store all available information (all properties)
  - store all spectra
- on reactions
  - give the entire stoichiometry of a reaction
  - store all reaction conditions (solvent, temperature, reaction time)
  - kinetic data



# Learning from Chemical Information

From models to interpretation

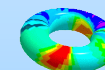
- structure representation by descriptors that can be interpreted
- combine substructures with physicochemical effects
- use data analysis methods that are not black boxes  
(a **forest** of decision trees gives more exact predictions but cannot be interpreted; a **single** decision tree can be interpreted!)





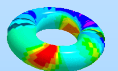
# Applications

- all fields of chemistry
- drug design
- organic synthesis design
- reaction databases
- chemical reactivity
- biochemical reactions
- structure elucidation



# Drug Design

- conformational flexibility of drugs and proteins
- docking into proteins
  - no consensus scoring (science cannot be predicted by voting)
  - try to model the physicochemistry of the process
- protein-protein and protein-DNA interactions
- prediction of ADME-Tox properties
- model the various organs of a human



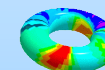
# Computer-Assisted Organic Synthesis Design (CASD)

- CASD was one of the roots of chemoinformatics
- many products can be traced back to CASD work
  - MACCS, REACCS, Beilstein DB, ChemInform RX DB
- however CASD systems are not yet widely accepted by organic chemists
- but it is still true:

“The amount of information to be processed and the decisions between many alternatives suggests the use of computers in synthesis design.”

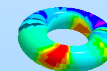
(H.Gelernter, 1973)

➔ The design of organic syntheses should be done more efficiently, using all available information, by using software



# Chemical Reactivity

- needs better data in reaction databases
- reach out to theoretical chemists
- put the results of quantum mechanical calculations into databases

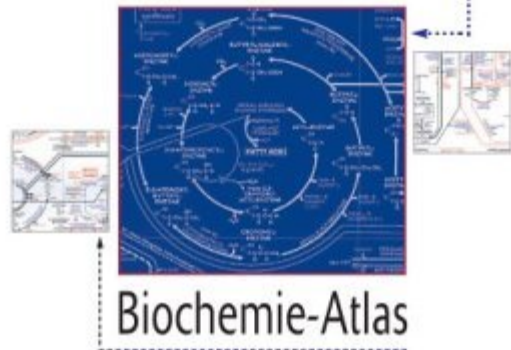


A B C D E F G H I J K L

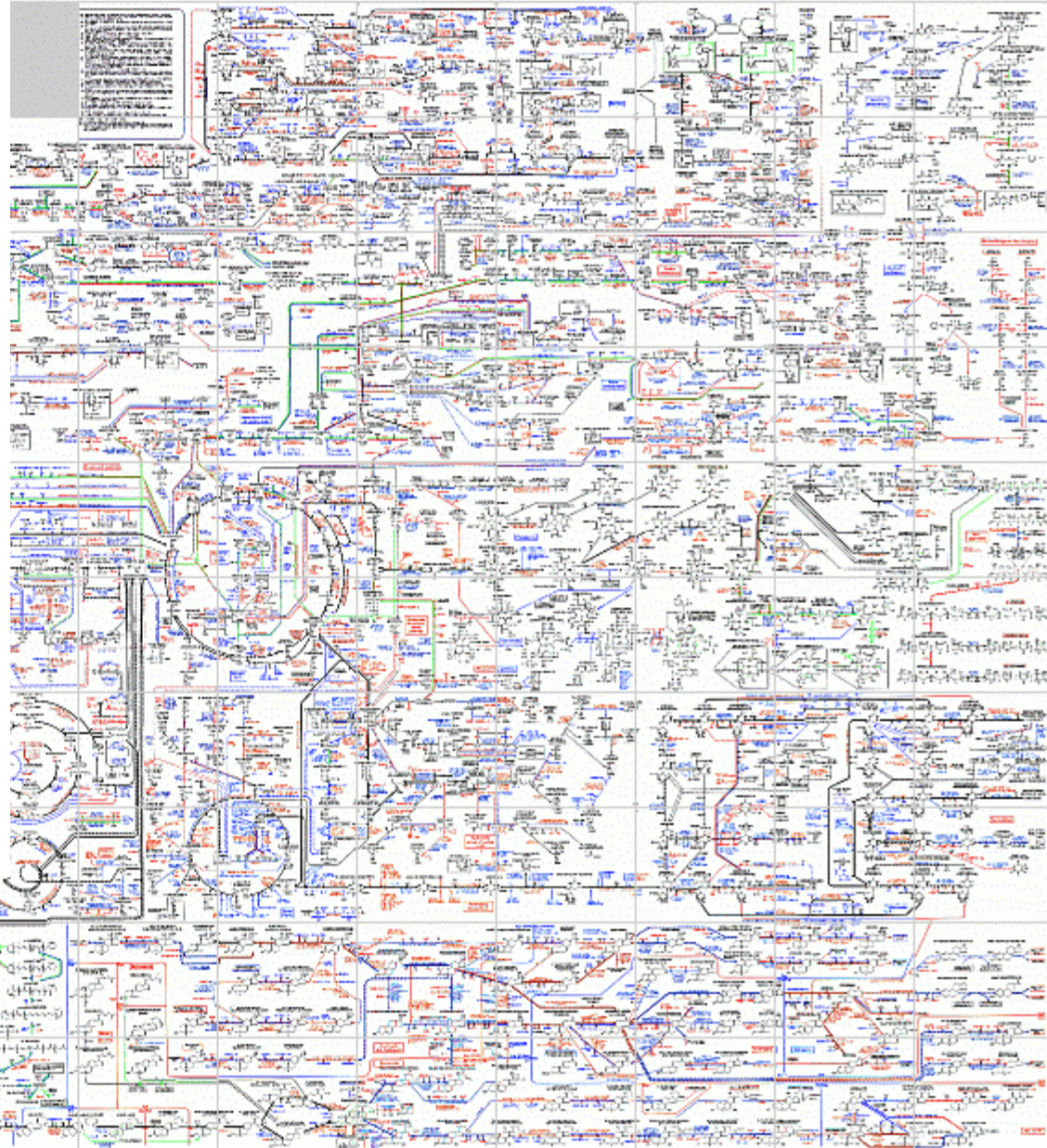
1  
2  
3  
4  
5  
6  
7  
8  
9  
10

Gerhard Michal (Hrsg.)

# Biochemical Pathways



Spektrum

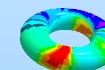


# Application of BioPath.Database

- search for enzyme inhibitors  
M.Reitz, A.von Homeyer, J.Gasteiger,  
J.Chem.Inf.Model., **2006**, 46, 2324-2332
- search for similar enzymes  
O.Sacher, M.Reitz, J.Gasteiger,  
J.Chem.Inf.Model., **2009**, 49, 1525-1534  
X.Hu, A.Yan, T.Tan, O.Sacher, J.Gasteiger  
J.Chem.Inf.Model., **2010**, 50, 1089-1100
- discover essential pathways of diseases  
G.Kastenmüller, J.Gasteiger, H.W.Mewes,  
Bioinformatics, **2008**, 24, i56-i62  
G.Kastenmüller, M.E.Schenk, J.Gasteiger, H.W.Mewes,  
Genome Biology, **2009**, 10, R28

---

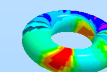
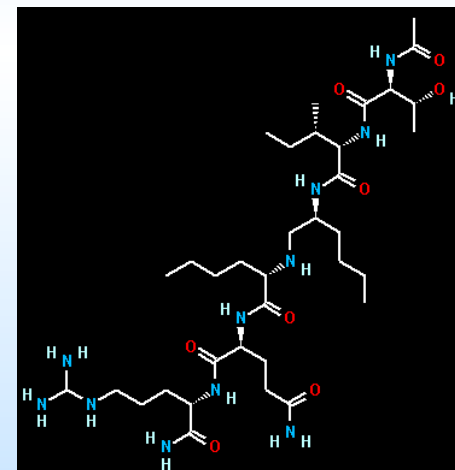
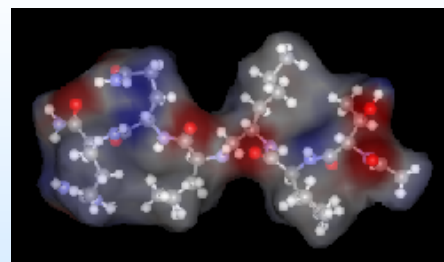
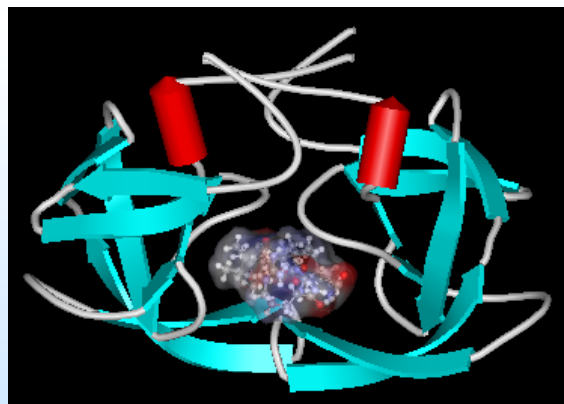
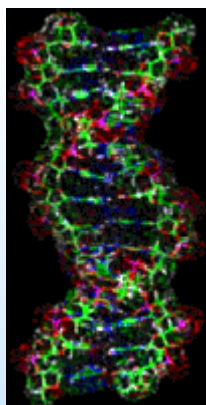
<http://www.molecular-networks.com/databases/biopath>



# Bioinformatics

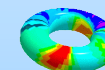
# Chemoinformatics

gene  $\longleftrightarrow$  protein  $\longleftrightarrow$  drug  $\longleftrightarrow$  lead



# Computer-Assisted Structure Elucidation (CASE)

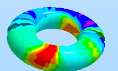
- CASE was one of the roots of chemoinformatics
- various groups worked on it (Munk, Sasaki, Funatsu, Steinrück)
- not much done recently
- no useful general purpose system available
- much time spent by chemists on structure elucidation
- structure elucidation should be done more efficiently, using all available information and using software





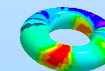
# Toxicity and Risk Assessment

- meeting the challenges posed by REACH, drug design and Cosmetics Directive
- projects funded by the European Union, Innovative Medicine Initiative and Cosmetics Europe
  - eTOX (11 academic groups, 6 SMEs, 13 pharma companies)
  - COSMOS (5 academic groups, 3 public institutions, 5 SMEs, 2 cosmetics companies)



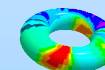
# Conclusions

- Chemoinformatics is a field of its own
  - *many achievements*
  - *still a lot to be done*
- teach chemoinformatics
- increase visibility of chemoinformatics
- publish in widely read journals
- get organized



# Teaching

- several textbooks have been published
  - Gillet+Leach, Gasteiger+Engel, Bajorath, Wild
- define curriculum in chemoinformatics
  - various universities already teach chemoinformatics
  - the number is growing
- integrate chemoinformatics into regular chemistry curricula

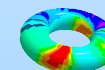


# Outreach

- society expects a lot from chemoinformatics
- cooperation industry (data) – academia (methods)
- funding (go into committees)
- revitalize:

The Cheminformatics and QSAR Society

<http://www.qsar.org/>



# Nobel Prize in Chemistry 2013

- Martin Karplus
- Michael Levitt
- Arieh Warshel

*“Today the computer is just as important a tool for chemists as the test tube.”*

The Royal Swedish Academy of Sciences; *press release*

