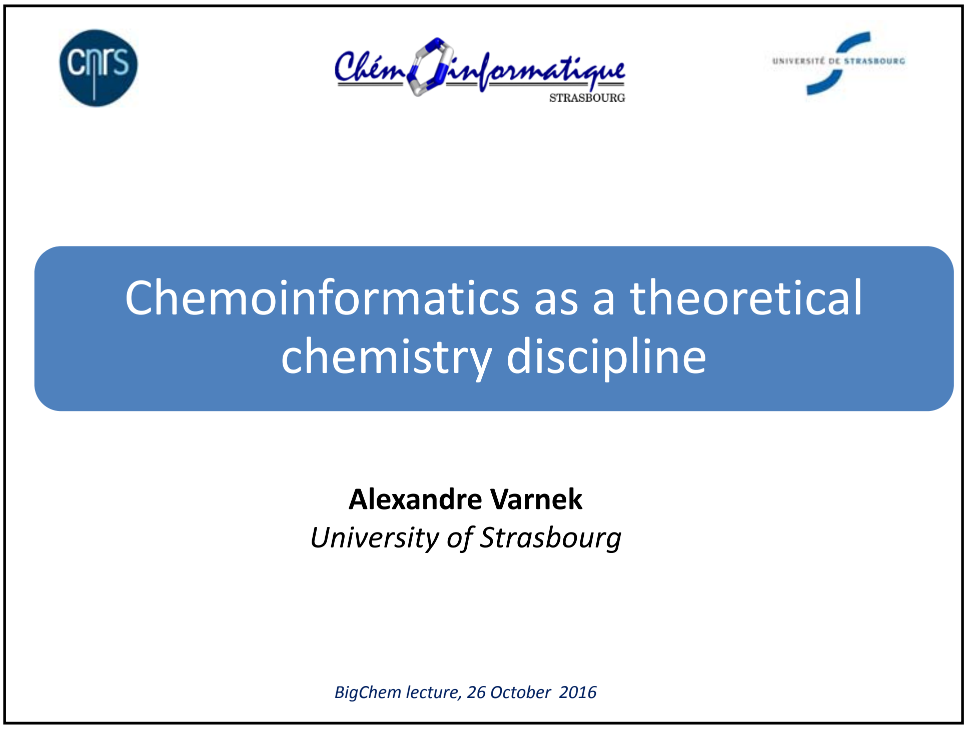
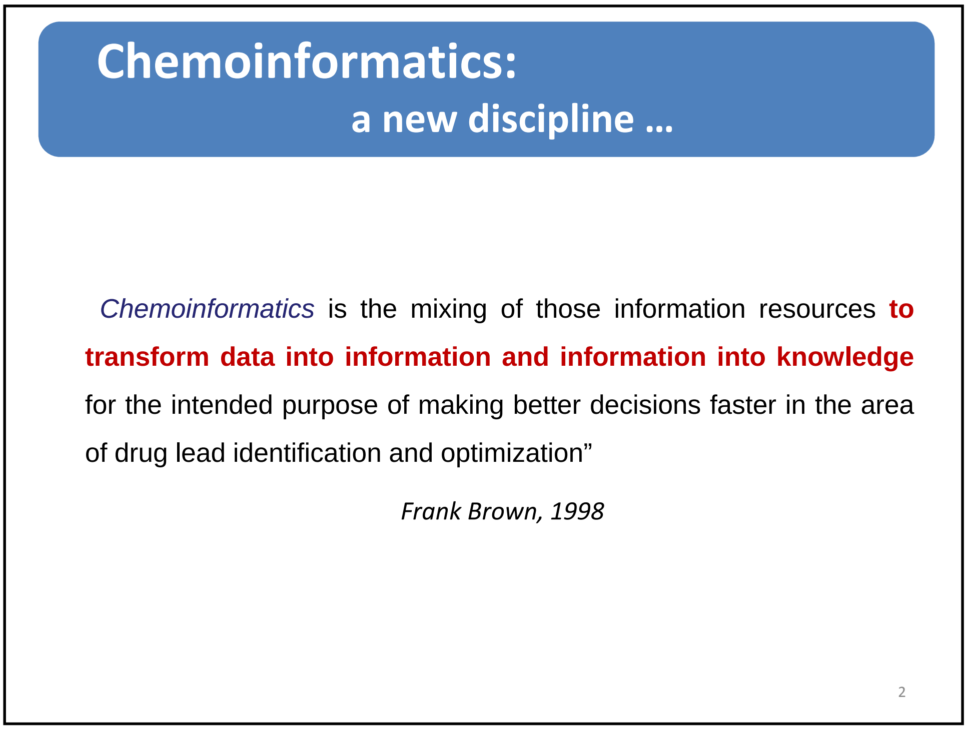


Chémoinformatique STRASBOURG

Chemoinformatics as a theoretical chemistry discipline

Alexandre Varnek
University of Strasbourg

BigChem lecture, 26 October 2016



Chemoinformatics:
a new discipline …

Chemoinformatics is the mixing of those information resources **to transform data into information and information into knowledge** for the intended purpose of making better decisions faster in the area of drug lead identification and optimization"

Frank Brown, 1998

Chemoinformatics: definition

Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information

G. Paris, 1998

Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization"

F.K. Brown, 1998

- *Chemoinformatics* is the application of informatics methods to solve chemical problems

J. Gasteiger, 2004

Chemoinformatics is a field based on the representation of molecules as objects (graphs or vectors) in a chemical space

A. Varnek & I. Baskin, 2011

Chemoinformatics: new disciline combining several „old“ fields

- **Chemical databases**
- **Structure-Activity modeling (QSAR)**
- **Structure-based drug design**
- **Computer-aided synthesis design**

Michael Lynch

Peter Willett

Corwin Hansch

Johann Gasteiger

Irwin D. Kuntz

Hans-Joachim Böhm

Elias Corey

Ivar Ugi

4

Selected books in chemoinformatics

An Introduction to Chemoinformatics
Handbook of Chemoinformatics Algorithms
Chemoinformatics
Handbook of Chemoinformatics
Chemoinformatics
Concepts, Methods, and Tools for Drug Discovery
Jürgen Bajorath
Chemoinformatics
Chemoinformatics Approaches to Virtual Screening
Molecular Descriptors for Chemoinformatics
Chemoinformatics for Drug Discovery
Advanced Methods and Applications in Chemoinformatics
Research Progress and New Applications
HANDBOOK OF RESEARCH ON CHEMOINFORMATICS AND CHEMICAL ENGINEERING
A. K. Haghi
Springer Protocols
Chemoinformatics and Computational Chemical Biology
Edited by Jürgen Bajorath
WILEY
RSC Publishing

***Chemoinformatics*:**

intersection of chemistry, computer science, mathematics, biology, material science, …

Is Chemoinformatics an individual scientific discipline or just a mixture of methods and concepts imported from different fields ?

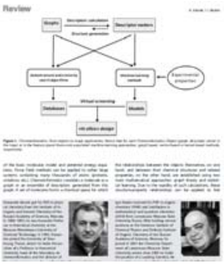
Review

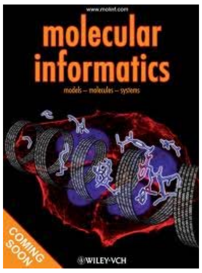
DOI: 10.1002/minf.201000100

Chemoinformatics as a Theoretical Chemistry Discipline

Alexandre Varnek*[a] and Igor I. Baskin[b]

Mol. Inf. 2011, 30, 20 – 32





Chemoinformatics is defined as individual discipline characterized by its own molecular model, basic concepts, major applications and learning approach

7

OUTLOOK

- **Needs in chemoinformatics**
- **3 complementary modeling disciplines**
  - o Quantum Chemistry, FF modeling and Chemoinformatics –
- **Fundamentals of Chemoinformatics**
  - o Chemical Space paradigm: graphs-based and descriptors based CS
  - o Modeling background: *Machine learning methods.*
- **Chemoinformatics and "Sister" Disciplines**
  - o Machine Learning, Chemometrics and Bioinformatics

8

Needs in Chemoinformatics

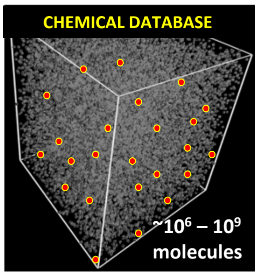
Big Data Challenge

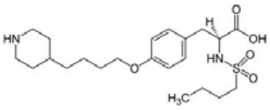
- > 10^8 compounds are currently available
- 10^{33} drug-like molecules could be synthesized

(see P. Polischuk, T. Madzidov , A. Varnek., JCAMD, 2013)

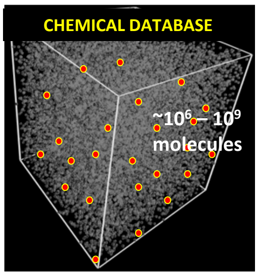
Goal: to select few useful compounds from huge chemical database

Screening: finding the needle in the haystack

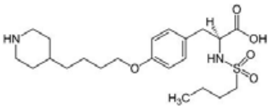




Chemoinformatics: pattern recognition in chemistry



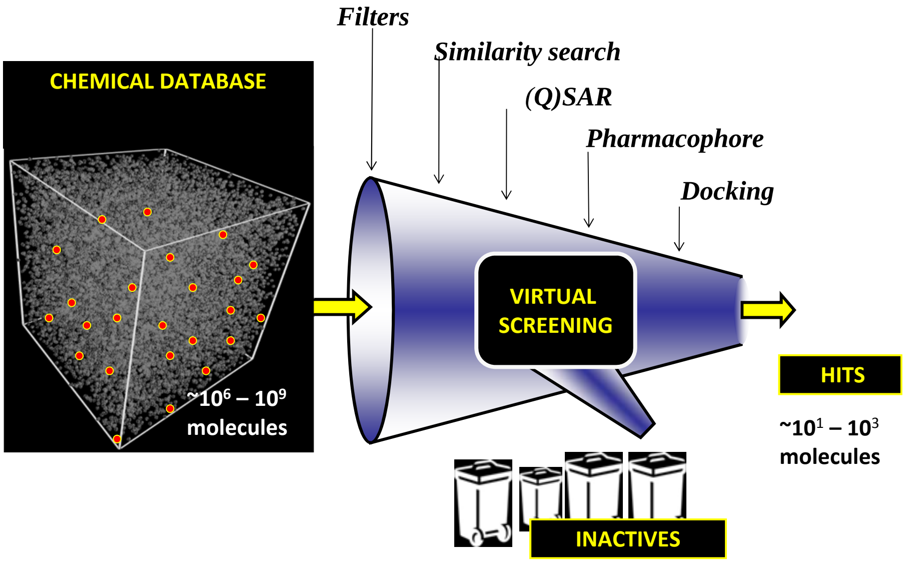
model



- Specific structural motifs,
- Selected molecular properties (shape, fields, …),
- Interaction patterns,
- Mathematical equations

Property= F (structure)

Virtual screening "funnel"



Filters

Similarity search

(Q)SAR

Pharmacophore

Docking

CHEMICAL DATABASE

~10^6 – 10^9 molecules

VIRTUAL SCREENING

HITS

~10^1 – 10^3 molecules

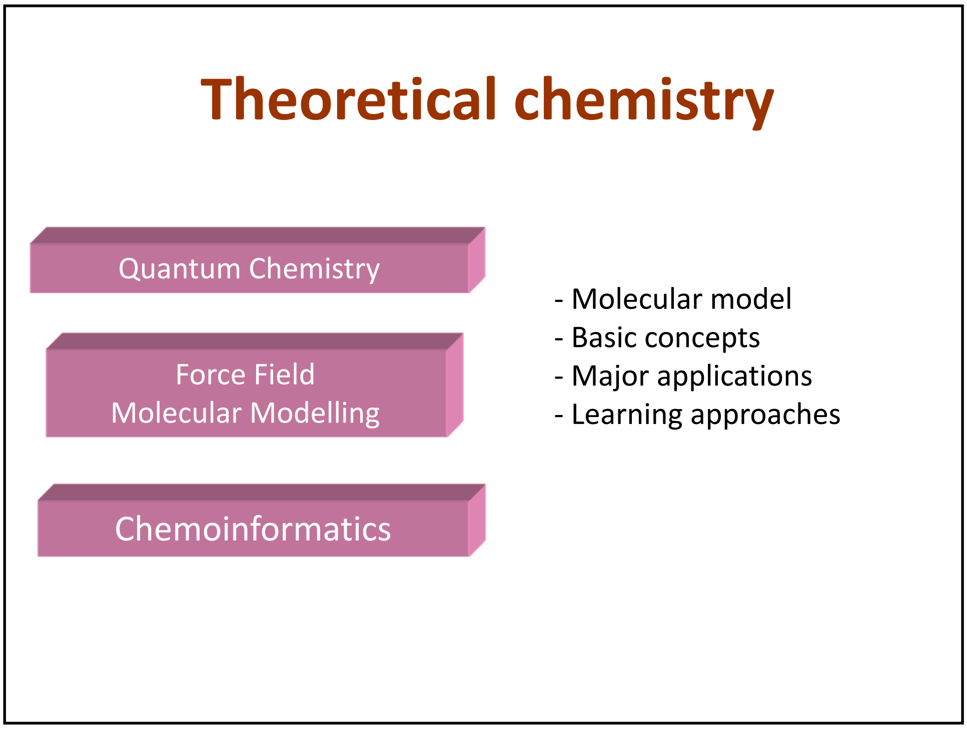
INACTIVES

Theoretical chemistry

Quantum Chemistry

Force Field
Molecular Modelling

Chemoinformatics



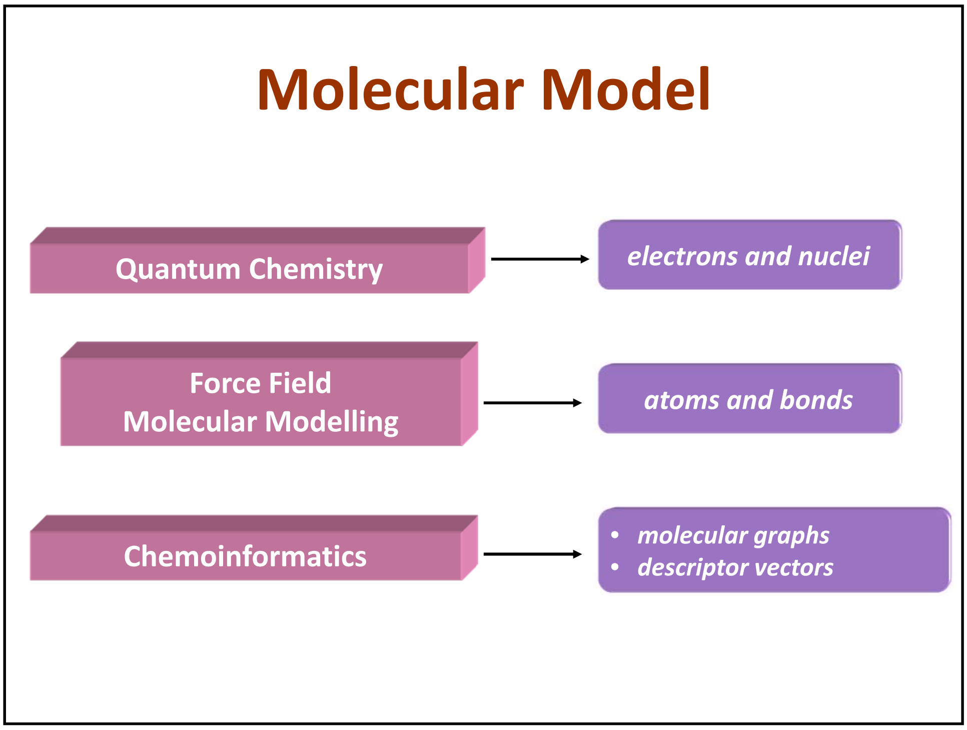
Theoretical chemistry

Quantum Chemistry

Force Field
Molecular Modelling

Chemoinformatics

- Molecular model
- Basic concepts
- Major applications
- Learning approaches



Molecular Model

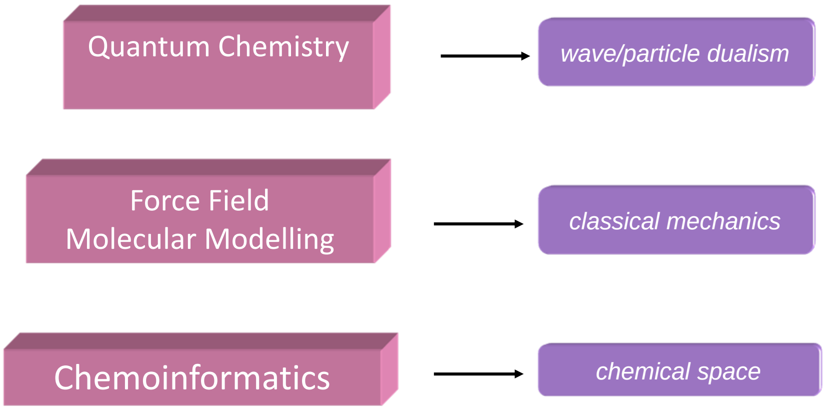
Quantum Chemistry → ***electrons and nuclei***

Force Field
Molecular Modelling → ***atoms and bonds***

Chemoinformatics →

- ***molecular graphs***
- ***descriptor vectors***

Basic concepts

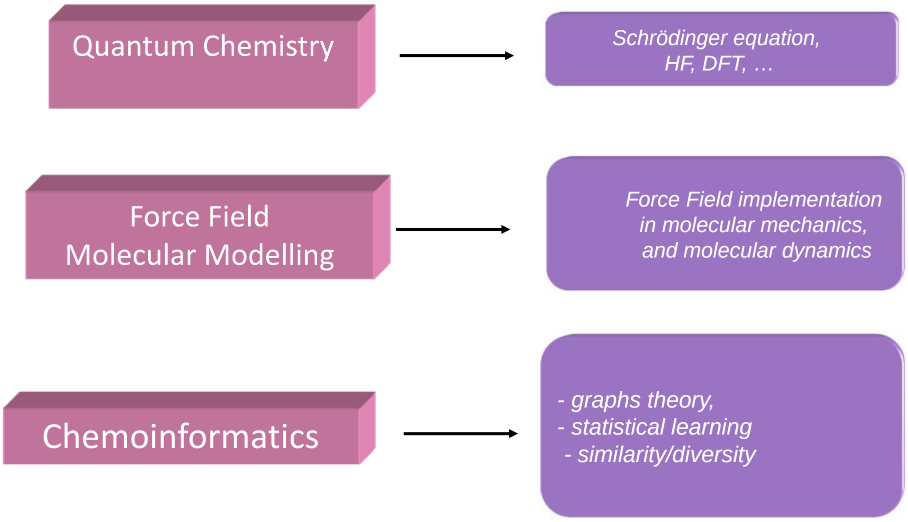


Quantum Chemistry → *wave/particle dualism*

Force Field Molecular Modelling → *classical mechanics*

Chemoinformatics → *chemical space*

Basic approaches

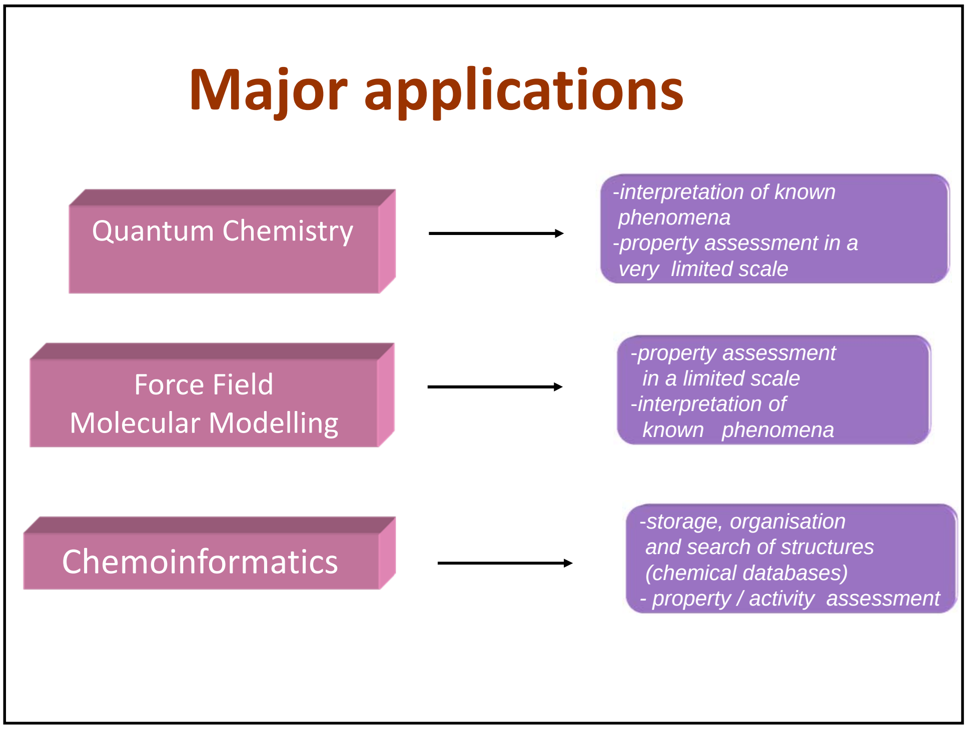


Quantum Chemistry → *Schrödinger equation, HF, DFT, …*

Force Field Molecular Modelling → *Force Field implementation in molecular mechanics, and molecular dynamics*

Chemoinformatics →

- *graphs theory,*
- *statistical learning*
- *similarity/diversity*

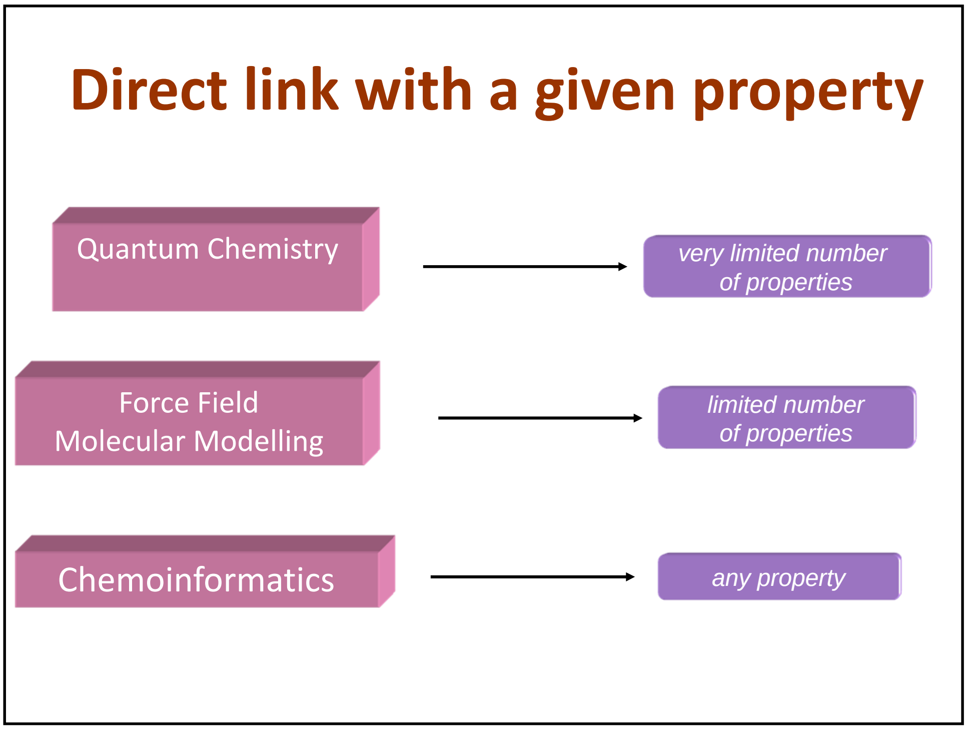


Major applications

Quantum Chemistry → -interpretation of known phenomena
-property assessment in a very limited scale

Force Field Molecular Modelling → -property assessment in a limited scale
-interpretation of known phenomena

Chemoinformatics → -storage, organisation and search of structures (chemical databases)
- property / activity assessment



Direct link with a given property

Quantum Chemistry → very limited number of properties

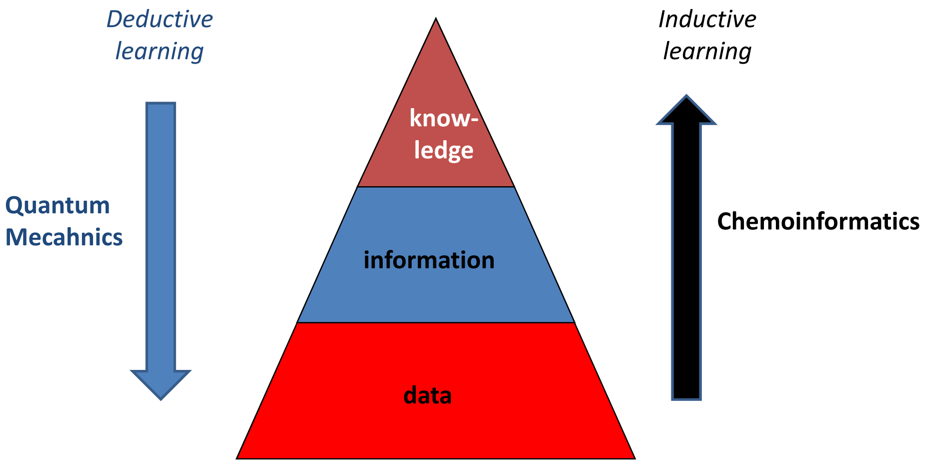
Force Field Molecular Modelling → limited number of properties

Chemoinformatics → any property

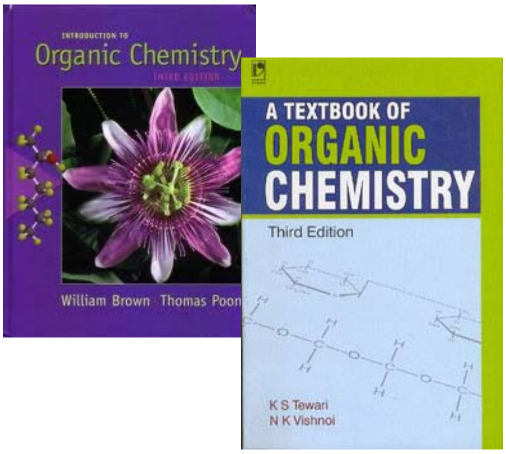
Learning approach

- In chemoinformatics the logic of learning is not based on existing physical theories. **Chemoinformatics considers the world too complex to be *a priori* described by any set of rules.** Thus, the rules (models) in chemoinformatics are not explicitly taken from rigorous physical models, but learned inductively from the data.

Chemoinformatics: From Data to Knowledge



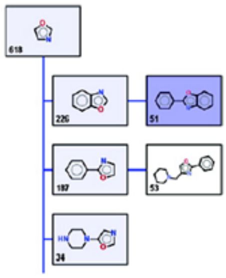
Organic chemistry:
exercise of « intuitive » chemoinformatics



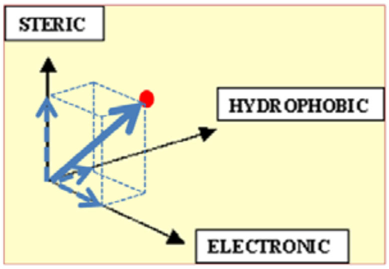
Chemical Space paradigm

Chemoinformatics **is a field dealing with molecular objects (graphs, vectors) in chemical space**

Chemical Space paradigm



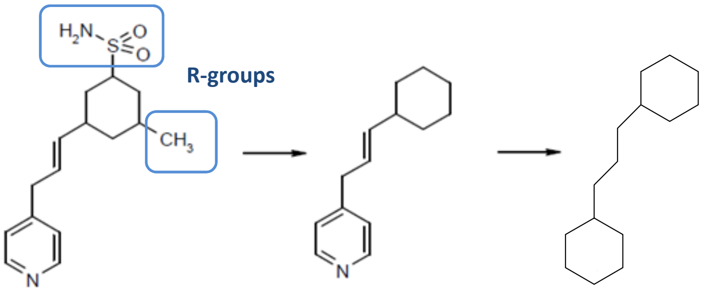
graphs-based



descriptors -based

SPACE = objects + relations between them

Scaffolds and Frameworks



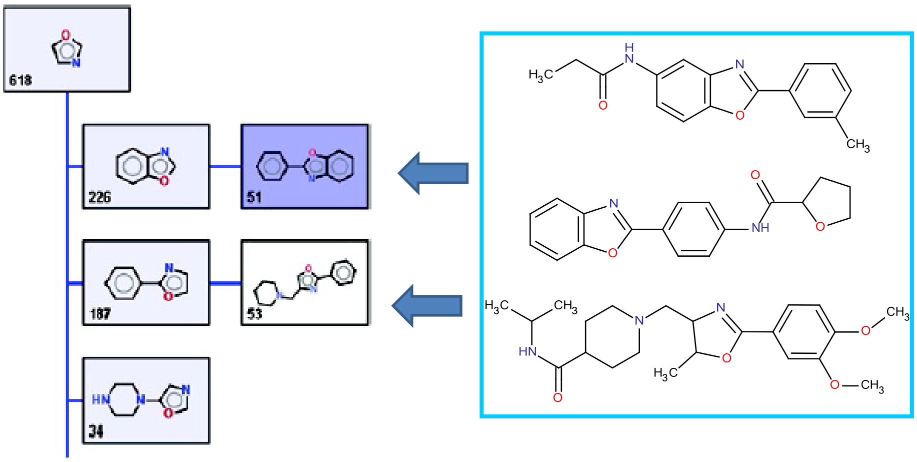
Scaffold

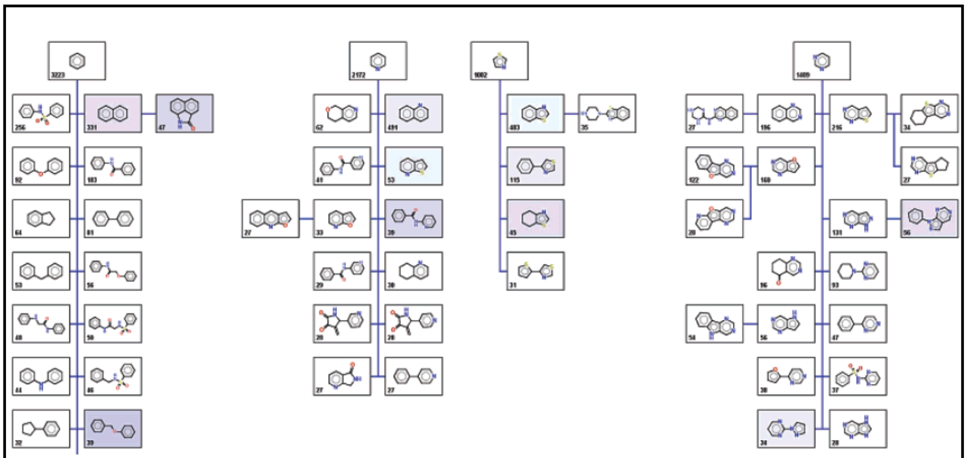
Framework

Bemis, G.W.; Murcko, M.A. *J.Med.Chem* ***1996,*** *39, 2887-2893*

The Scaffold Tree – Visualization of the Scaffold Universe by Hierarchical Scaffold Classification

A. Schuffenhauer, P. Ertl, et al. *J. Chem. Inf. Model.*, ***2007, 47 (1), 47-58***



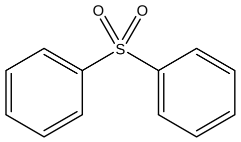


Scaffold tree for the results of pyruvate kinase assay. Color intensity represents the ratio of active and inactive molecules with these scaffolds.

A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, and H.Waldmann *J. Chem. Inf. Model.*, ***2007, 47 (1), 47-58***

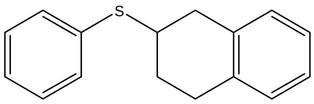
Maximal Common Substructure (MCS) similarity index

$$\textit{Graph Similarity} = \frac{N_{MCS}}{\min(N_1, N_2)}$$



Mol 1

$N_1 = 16$

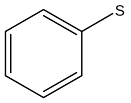


Mol 2

$N_2 = 18$

MCS

$N_{MCS} = 7$



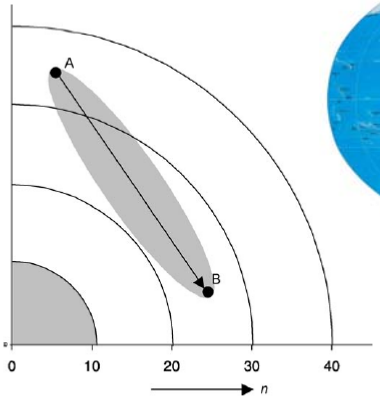
$$\textit{Graph Similarity} = \frac{7}{16}$$

T. R. Hagadone *J. Chem. Inf. Comput. Sci.* 1992,**32**, 515-521

Chemical Space Travel

Ruud van Deursen and Jean-Louis Reymond[

ChemMedChem **2007**, *2*, 636 – 640



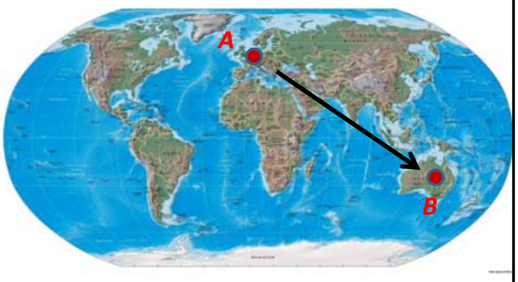
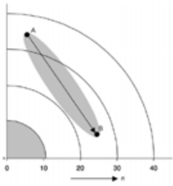


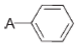
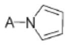
Figure 1. Travelling between A and B for targeted exploration of unknown chemical space (shaded area). The shaded area under n ≤ 11 has been explored by extensive enumeration.[3b] n is the number of non-hydrogen atoms in a molecule. The area is proportional to log *N* for *N* = the total number of molecules in chemical space up to n atoms per molecule.[3b]

Chemical Space Travel

Ruud van Deursen and Jean-Louis Reymond[

ChemMedChem **2007**, *2, 636 – 640*



Nearest neighbour mutations[a]		
Atom type exchange[b,c]	Replaces any atom by another atom type	
Atom inversion[c]	Inverts two neighbouring atoms	
Atom removal[c]	Primary: A–X→A Secondary: A–X-A→A–A Tertiary: XA_3→A–A–A (max. 6 combinations if 3 different A's) $A_2CH–CHA_2$ or $A_2C=CA_2$→CA_4 Quaternary: XA_4→A–A–A–A or $A(A)_3$ (max. 16 combinations if 4 different A's)	
Atom addition[b,c]	On terminal atoms: A→A–X In any bond: A–A→A–X–A In chains: A–A–A→XA_3; A–A–A–A→XA_4 Quaternary centres: CA_4→$A_2CH–CHA_2$ and $A_2C=CA_2$ (max. 6 combinations if 4 different A's)	
Bond saturation[c]	Breaks a cyclic σ- or any π-bond	
Bond unsaturation	Makes a cyclic σ- or π-bond	
Bond rearrangement[c]	Breaks a σ- or π-bond and inserts it anywhere else in the molecule	
Non-nearest neighbour mutations		
	$A-CH_3$→	
Aromatic ring addition[c,d]	$A-NH_2$→	
	H_2O→	

Chemical Space Travel

Ruud van Deursen and Jean-Louis Reymond[

ChemMedChem **2007**, *2, 636 – 640*

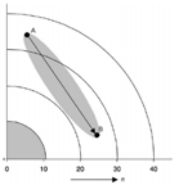
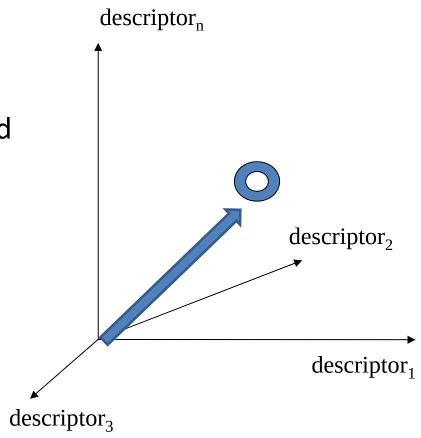


Table 3. Examples of chemical space travel between different molecules.[a]

From:	To:	Cubane	Aspirine	VX	Adenosine	Sucrose
Cubane		–	10	18	23 (1)	19
Aspirine		10*	–	14	21	15
VX		13	17 (1)	–	31 (1)	18
Adenosine		17*	27	18*	–	14
Sucrose		18*	22 (1)	22*	29 (1)	–
Penicillin G		19*	13*	14*	23	19*
Strychnine		21*	17*	20	26	22
Colchicine		27	22*	21	26	18
Tetracycline		28*	20	25*	49	19
Vitamin K		30*	24*	30*	34*	28*

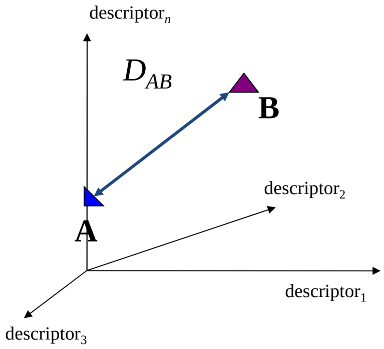
Descriptors-based chemical space

Each object (molecule, reaction, interaction pattern) is represented by a vector whereas the metrics is defined by distance or similarity measures



Descriptors-based chemical space

Distance in chemical space is used as a measure of molecular “similarity“ and “dissimilarity“



Popular Similarity / Distance measures

- **Similarity :**
  - Tanimoto coefficient
  - Dice coefficient
  - Cosine coefficient
- **Distance :**
  - Euclidean
  - Manhattan
  - Soergel

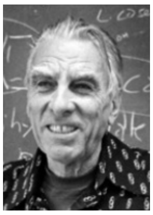
Descriptors-based chemical space

Biological Activity = *f* (Physicochemical properties)

$$\log 1/C = a\cdot(\log P)^2 + b\cdot\log P + c\cdot\sigma + d\cdot E_s + \text{const}$$

Physicochemical properties can be broadly classied into three general types:

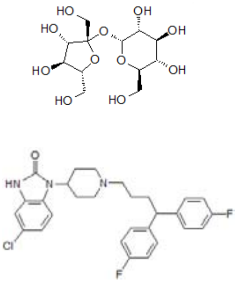
- Electronic (σ)
- Steric (E_s)
- Hydrophobic (**logP**)

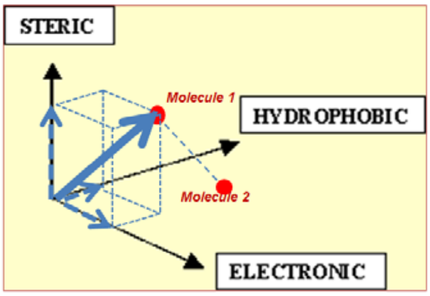


Corwin Hansch

Descriptors-based chemical space

$$\log 1/C = a\,(\log P)^2 + b \log P + \rho\sigma + \delta E_s + \text{const}$$



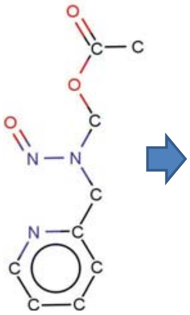


Molecular Descriptors :

ensemble of topological, electronic, geometry parameters calculated directly from molecular structure

Molecular graph

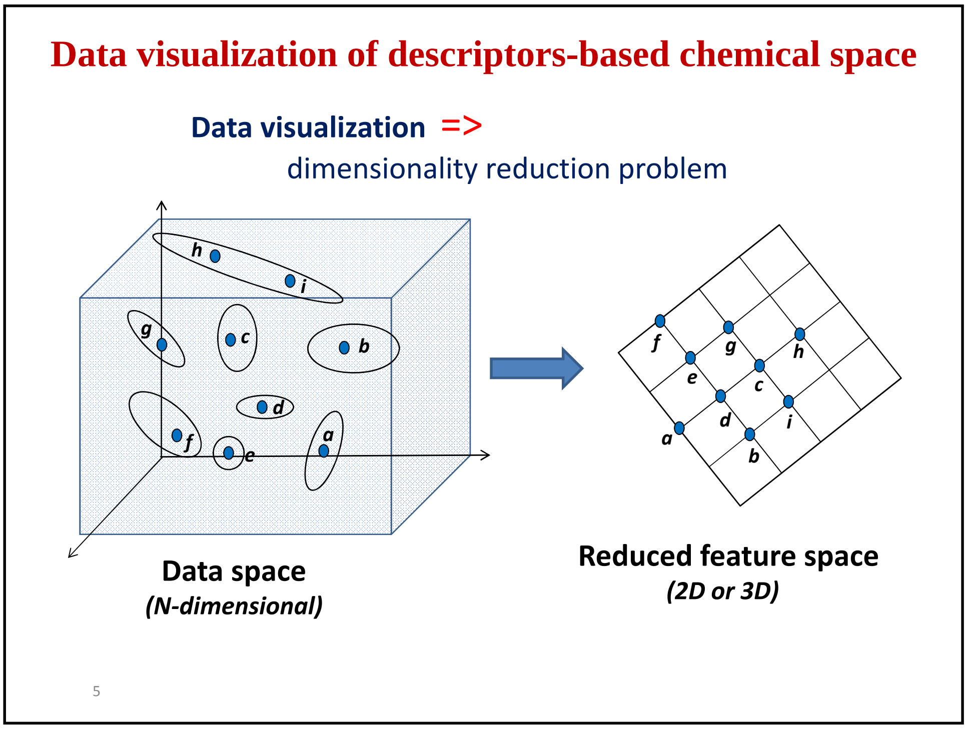
Descriptor vector



- Topological indices,
- Atomic charges,
- Inductive descriptors,
- Substructural fragments,
- Molecular volume and surface, …

Descriptors
D_1
D_2
…
D_i
…

> 5000 types of descriptors are reported



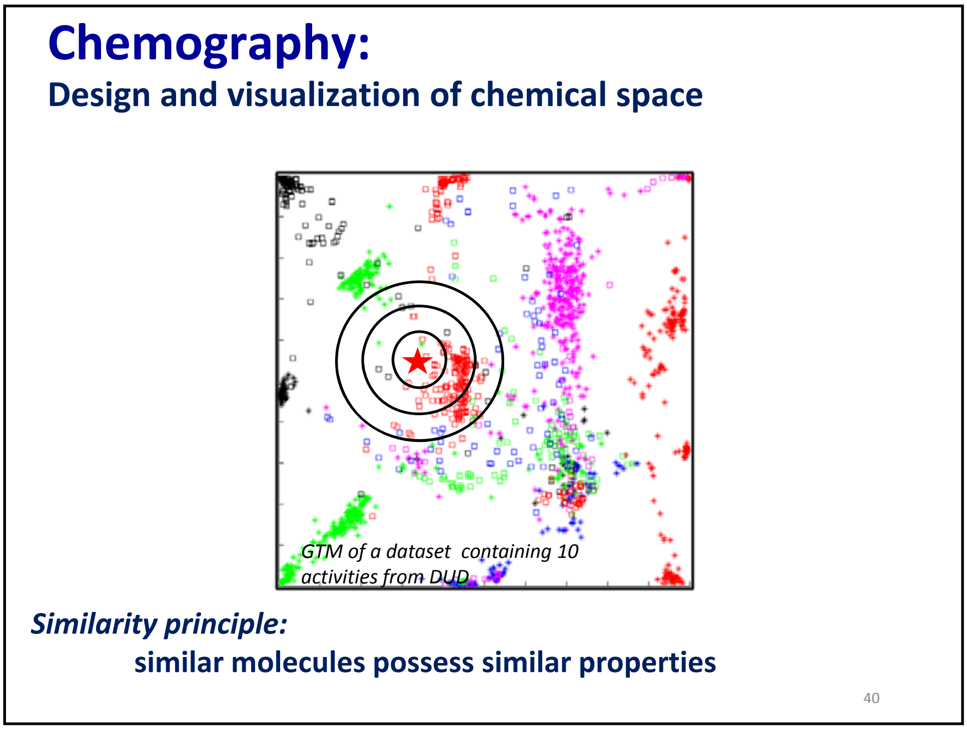
Data visualization of descriptors-based chemical space

Data visualization =>
dimensionality reduction problem

Data space
(N-dimensional)

Reduced feature space
(2D or 3D)

5



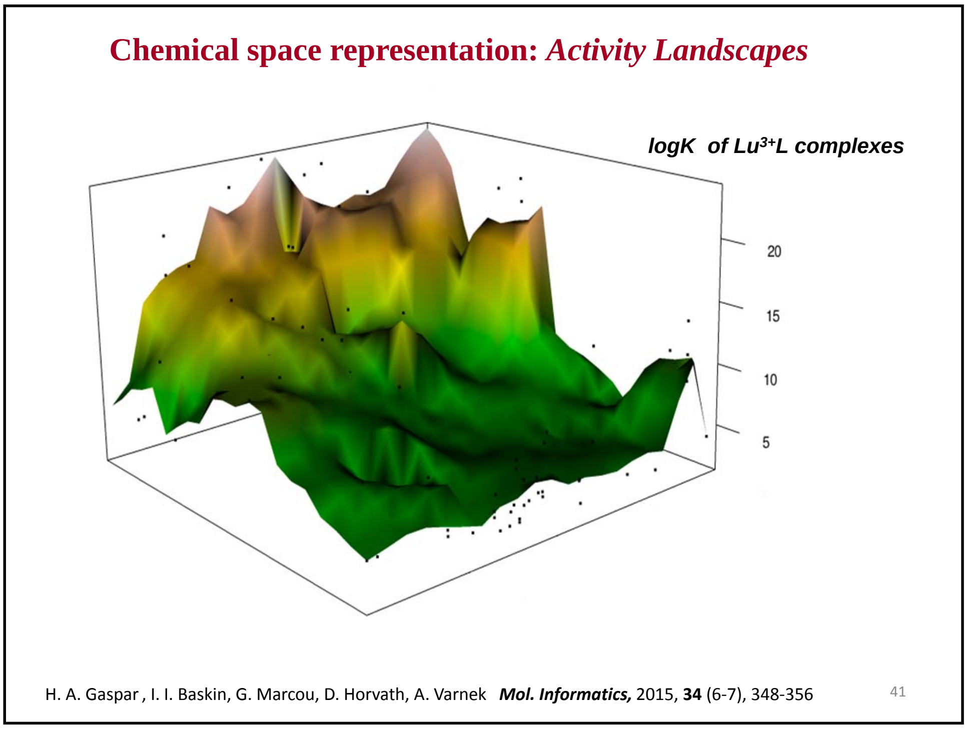
Chemography:

Design and visualization of chemical space

GTM of a dataset containing 10 activities from DUD

Similarity principle:
similar molecules possess similar properties

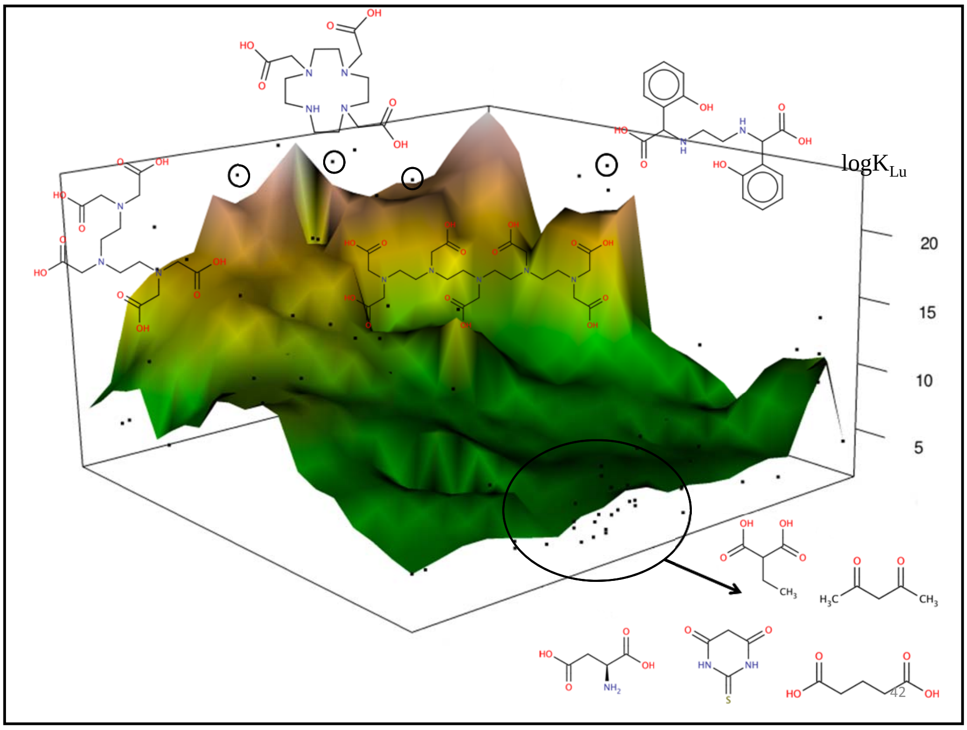
40



Chemical space representation: *Activity Landscapes*

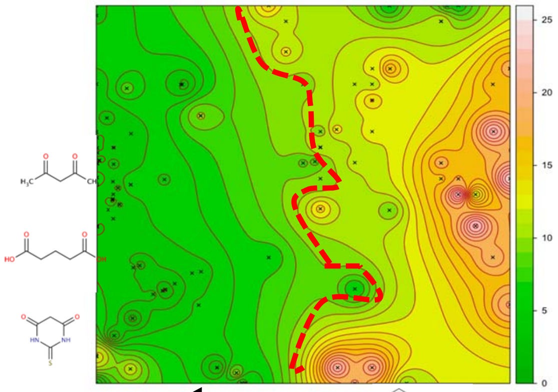
H. A. Gaspar , I. I. Baskin, G. Marcou, D. Horvath, A. Varnek ***Mol. Informatics,*** 2015, **34** (6-7), 348-356

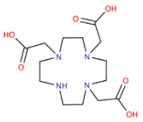
41



42

Chemical space visualization

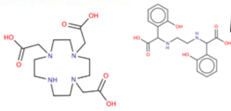




+ *Lu^{3+}*

Generative Topographic Mappping of the set of Lu^{3+} binders

Contours correspond to different logK values



Weak binders

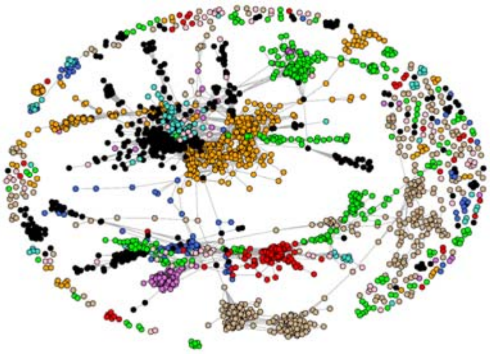
Strong binders

H. A. Gaspar , I. I. Baskin, G. Marcou, D. Horvath, A. Varnek ***Mol. Informatics,*** 2015, **34** (6-7), 348-356

Network-like Similarity Graphs

Representation of the database as a graph

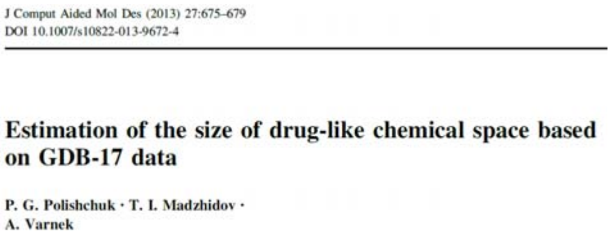
- each molecule is presented as as a node,
- two nodes are connected if they are similar enough (T > T_0)



Database containing > 2700 ligands against 10 different targets extracted from DUD

Wasserman et al. *J. Med. Chem.*, 2010, Vol. 53, No. 23

Chemical Space: how large is it ?



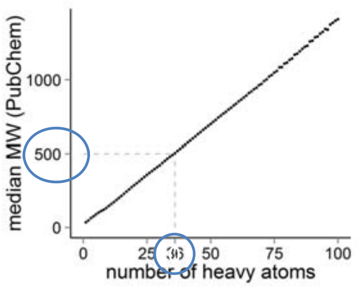
J Comput Aided Mol Des (2013) 27:675–679
DOI 10.1007/s10822-013-9672-4

Estimation of the size of drug-like chemical space based on GDB-17 data

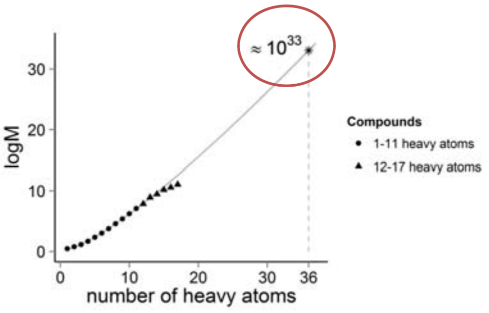
P. G. Polishchuk · T. I. Madzhidov · A. Varnek

- **GDB-17** – computer-generated set of 1.66 *10^{11} structures containing up to *N* = 17 heavy atoms (L. Ruddigkeit et al. J Chem Inf Model 2012, **52,** 2864–2875)
- The number of structures corresponding to *N*= 1, 2, 3, …, 17 is available. This alllows one to establish relationships between the number of structures (*M*) and *N*
- What is a limited value of *N ?*

Chemical Space: how large is it ?

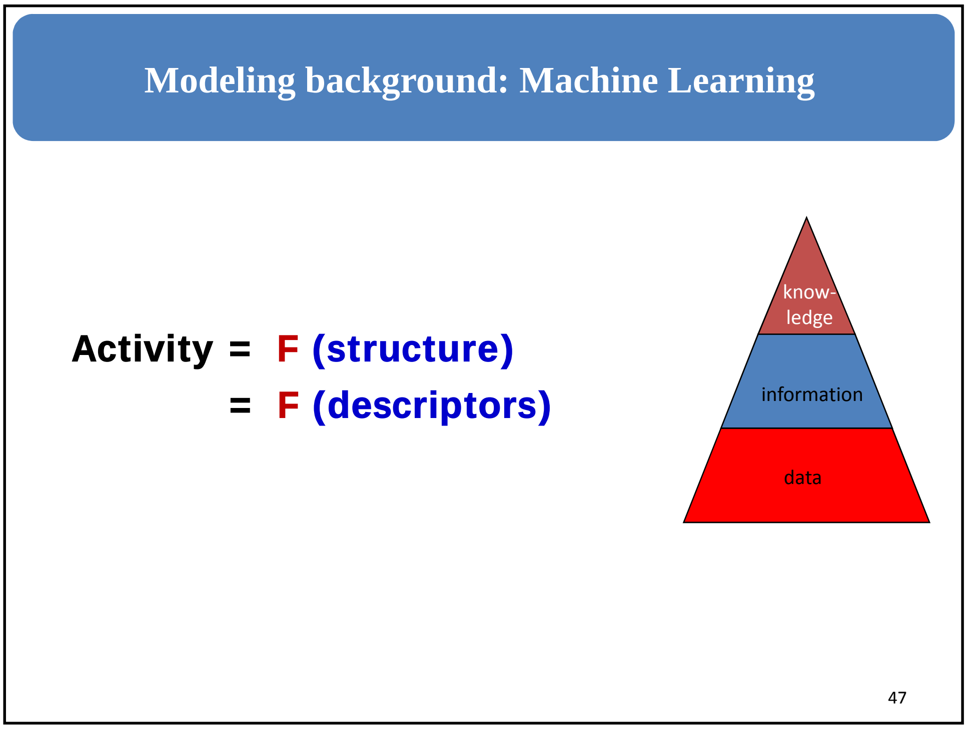


Median MW vs number of heavy atoms for the PubChem database



Extrapolation of the compounds number (***M***) as a function of the number of heavy atoms (***N***) based on data taken from GDB-17

$$\log M = 0.584 \times N \times \log N + 0.356$$

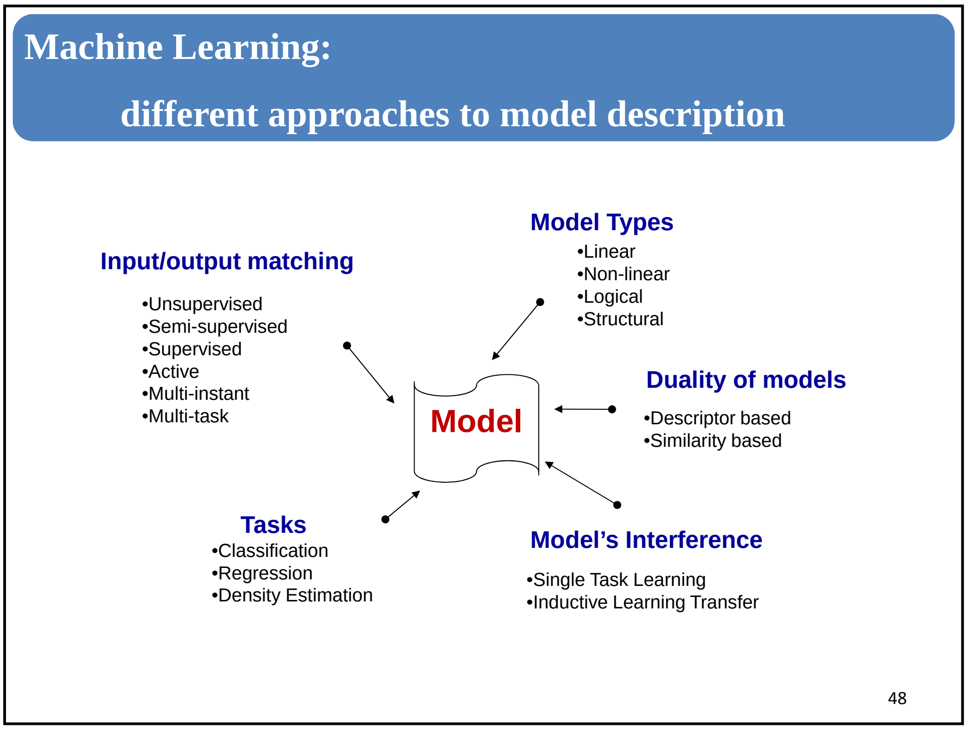


Modeling background: Machine Learning

Activity = F (structure)

= F (descriptors)

know-
ledge
information
data



Machine Learning: different approaches to model description

Input/output matching

- Unsupervised
- Semi-supervised
- Supervised
- Active
- Multi-instant
- Multi-task

Model Types

- Linear
- Non-linear
- Logical
- Structural

Model

Duality of models

- Descriptor based
- Similarity based

Tasks

- Classification
- Regression
- Density Estimation

Model's Interference

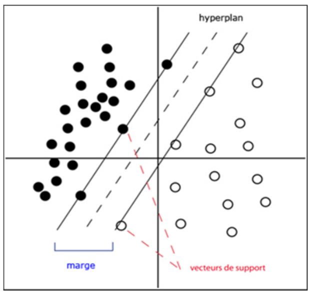
- Single Task Learning
- Inductive Learning Transfer

Machine learning methods

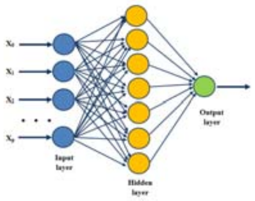
Multiple Linear Regression (MLR)

$$\text{Property} = a_0 + \sum_{i=1}^{k} a_i \cdot X_i$$

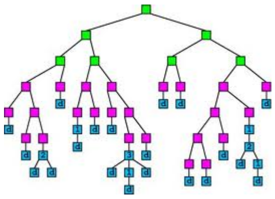
Support Vector Machine



Neural Networks



Decision Trees

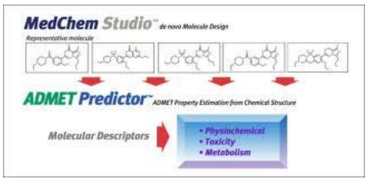


49

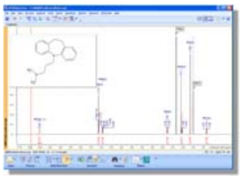
Predictors:

Commericial and Public Software

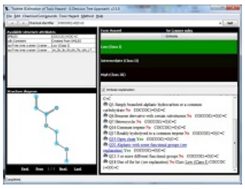
simulations plus, inc.
integrating science and software


MedChem Studio™ de novo Molecule Design
Representative molecule
ADMET Predictor™ ADMET Property Estimation from Chemical Structure
Molecular Descriptors
• Physiochemical
• Toxicity
• Metabolism

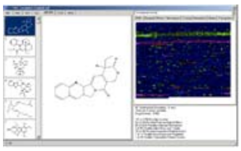
ACD/Labs



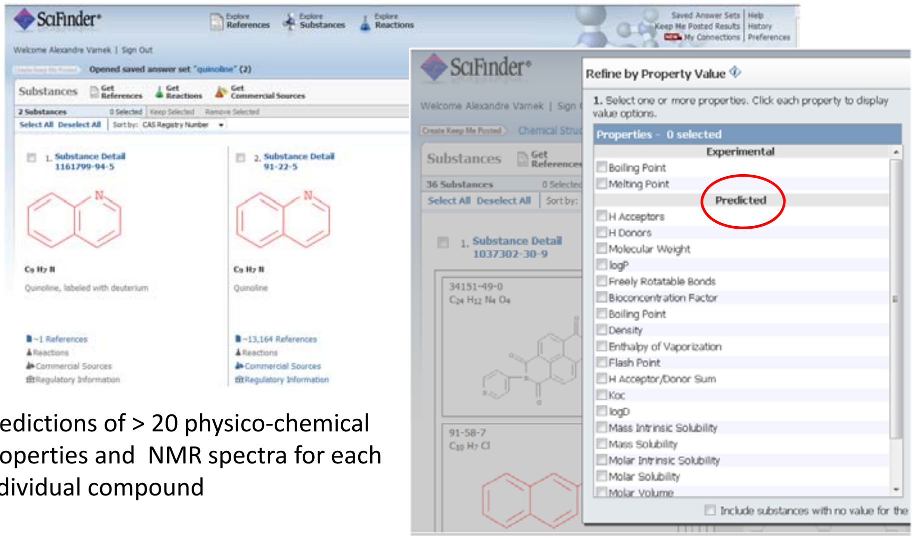
Toxtree



PASS



Predictive tools in SciFinder



Refine by Property Value

1. Select one or more properties. Click each property to display value options.

Properties - 0 selected

Experimental

- □ Boiling Point
- □ Melting Point

Predicted

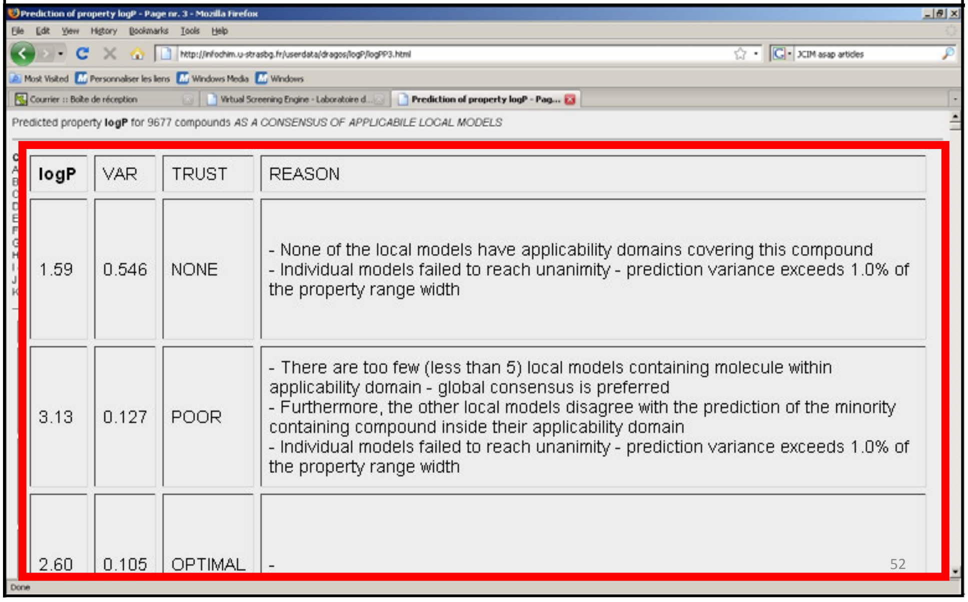
- □ H Acceptors
- □ H Donors
- □ Molecular Weight
- □ logP
- □ Freely Rotatable Bonds
- □ Bioconcentration Factor
- □ Boiling Point
- □ Density
- □ Enthalpy of Vaporization
- □ Flash Point
- □ H Acceptor/Donor Sum
- □ Koc
- □ logD
- □ Mass Intrinsic Solubility
- □ Mass Solubility
- □ Molar Intrinsic Solubility
- □ Molar Solubility
- □ Molar Volume

□ Include substances with no value for the

predictions of > 20 physico-chemical properties and NMR spectra for each individual compound

ISIDA virtual screening server

infochim.u-strasbg.fr/webserv/VSEngine.html

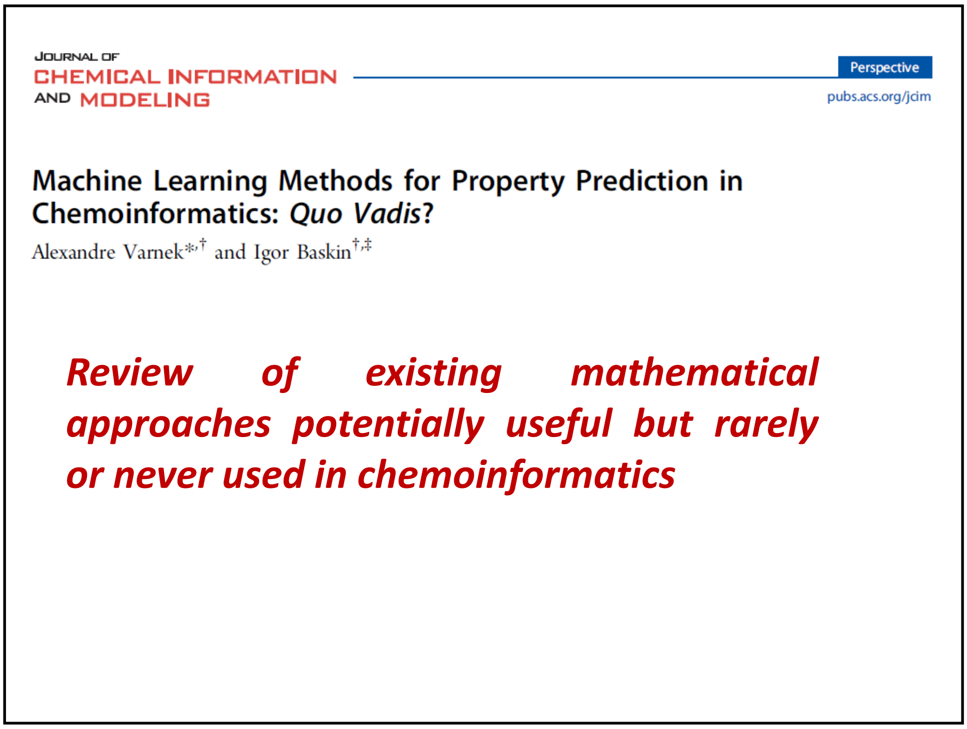


Prediction of property logP - Page nr. 3 - Mozilla Firefox

http://infochim.u-strasbg.fr/userdata/dragos/logP/logPP3.html

Predicted property **logP** for 9677 compounds *AS A CONSENSUS OF APPLICABILE LOCAL MODELS*

logP	VAR	TRUST	REASON
1.59	0.546	NONE	- None of the local models have applicability domains covering this compound - Individual models failed to reach unanimity - prediction variance exceeds 1.0% of the property range width
3.13	0.127	POOR	- There are too few (less than 5) local models containing molecule within applicability domain - global consensus is preferred - Furthermore, the other local models disagree with the prediction of the minority containing compound inside their applicability domain - Individual models failed to reach unanimity - prediction variance exceeds 1.0% of the property range width
2.60	0.105	OPTIMAL	-



JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

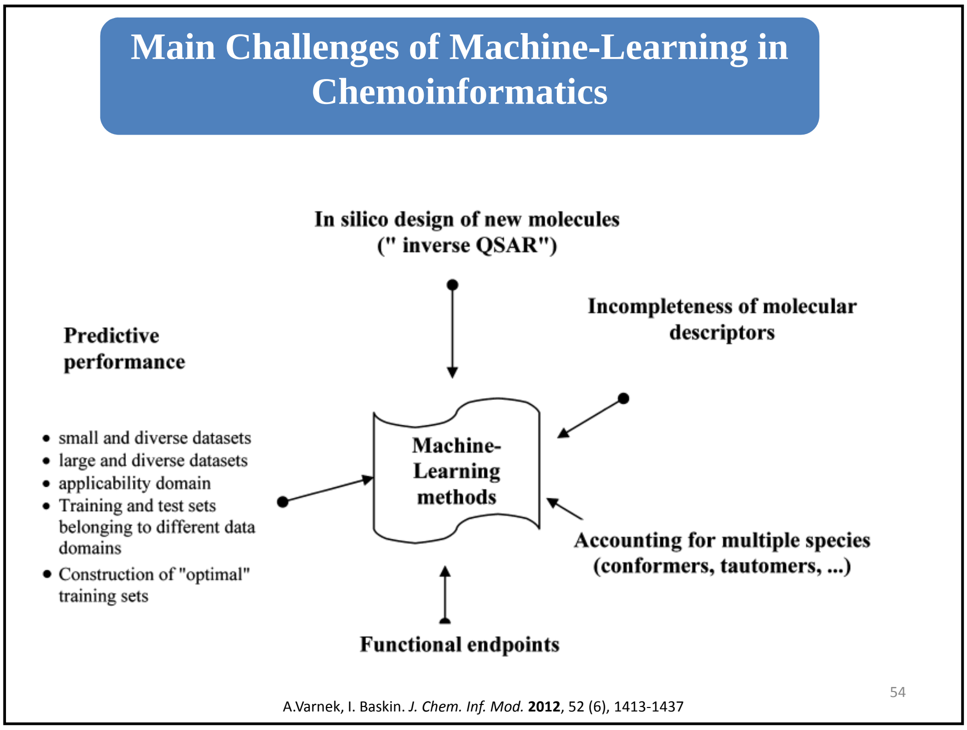
Perspective

pubs.acs.org/jcim

Machine Learning Methods for Property Prediction in Chemoinformatics: *Quo Vadis*?

Alexandre Varnek*,† and Igor Baskin†,‡

Review of existing mathematical approaches potentially useful but rarely or never used in chemoinformatics



Main Challenges of Machine-Learning in Chemoinformatics

In silico design of new molecules
(" inverse QSAR")

Incompleteness of molecular descriptors

Predictive performance

- small and diverse datasets
- large and diverse datasets
- applicability domain
- Training and test sets belonging to different data domains
- Construction of "optimal" training sets

Machine-Learning methods

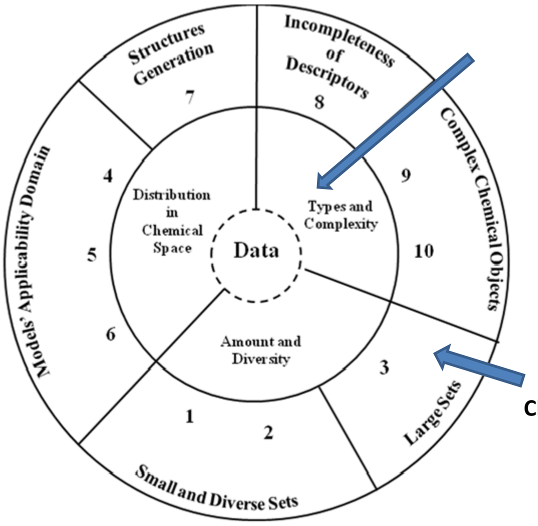
Accounting for multiple species (conformers, tautomers, ...)

Functional endpoints

54

A.Varnek, I. Baskin. *J. Chem. Inf. Mod.* **2012**, 52 (6), 1413-1437

Guide to choose machine learning method to solve chemical problems



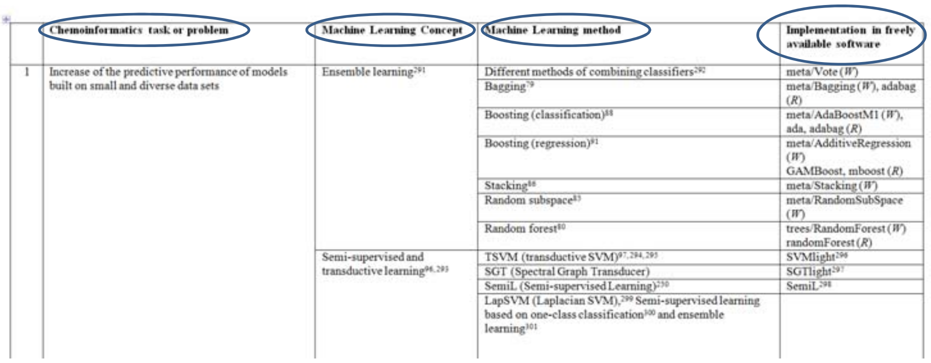
Different features of data

(*inner circle*)

Challenges of chemoinformatics

(*outer circle*)

Chemoinformatics Tools and the Appropriate Machine Learning Concepts and Methods



	Chemoinformatics task or problem	Machine Learning Concept	Machine Learning method	Implementation in freely available software
1	Increase of the predictive performance of models built on small and diverse data sets	Ensemble learning[291]	Different methods of combining classifiers[292]	meta/Vote (*W*)
			Bagging[79]	meta/Bagging (*W*), adabag (*R*)
			Boosting (classification)[88]	meta/AdaBoostM1 (*W*), ada, adabag (*R*)
			Boosting (regression)[91]	meta/AdditiveRegression (*W*) GAMBoost, mboost (*R*)
			Stacking[88]	meta/Stacking (*W*)
			Random subspace[83]	meta/RandomSubSpace (*W*)
			Random forest[80]	trees/RandomForest (*W*) randomForest (*R*)
		Semi-supervised and transductive learning[96,293]	TSVM (transductive SVM)[97,294,295]	SVMlight[296]
			SGT (Spectral Graph Transducer)	SGTlight[297]
			SemiL (Semi-supervised Learning)[230]	SemiL[298]
			LapSVM (Laplacian SVM),[299] Semi-supervised learning based on one-class classification[300] and ensemble learning[301]	

A. Varnek, I. Baskin. *J. Chem. Inf. Mod.* **2012**, *52* (*6*), 1413-1437

Chemoinformatics and "Sister" Disciplines:

Machine Learning, Chemometrics and Bioinformatics

Chemoinformatics *vs* Machine Learning

Chemoinformatics is a very specific area of ML application. The specificity of chemoinformatics results from:

- the nature of chemical objects (molecular graphs),
- the complexity of the chemical universe,

chemical data result from an explorative process rather than from specially organized sampling. Hence, they cannot be considered as representative, independent and identically distributed sampling from a well defined distribution. Special approaches are needed: applicability domain, active learning, …

- a possibility to account for an extra-knowledge, i.e., relationships between different properties issued from physicochemical theory.

Chemoinformatics *vs* Chemometrics

Chemometrics is the chemical discipline that uses mathematical, statistical and other methods

- to design or select optimal measurement procedures and experiments, and
- to provide maximum relevant chemical information by analysing chemical data.

L. Massart, Chemometrics: a textbook, Elseler, NY, 1988

Chemoinformatics *vs* Chemometrics

Generally, *chemometrics* **requires no information** about chemical structure and, therefore it overlaps with *chemoinformatics* only in the area of application of machine learning methods.

It is widely used in experiment design, chemical engineering, analytical chemistry and treatment of spectra – fields where an exhaustive treatment of multivariate data is needed.

Chemoinformatics *vs* Bioinfoormatics

Chemoinformatics - small molecules (2D molecular graphs)

Bioinformatics - large biological molecules (1D and 3D representation)

Combination of bio- and chemo-informatics approaches

- **Docking**: protein structures could be generated by bioinformatics tools, whereas some scoring functions involve vector representation of ligands

- **Protein-Ligand descriptors or fingerprints** based on available 3D information about protein-ligand complexes,

Chemoinformatics: intersection of chemistry, computer science, mathematics, biology, material science, …

Is Chemoinformatics an individual scientific discipline or just a mixture of methods and concepts imported from different fields ?

***Chemoinformatics* is an individual scientific discipline characterizing by its own molecular representations and basic concept – chemical space paradigm. It interfaces with graphs theory, machine-learning, QM and FF approaches in its various applications.**

Chémoinformatique
STRASBOURG

63