Cheminformatics and its Role in the Modern Drug Discovery Process

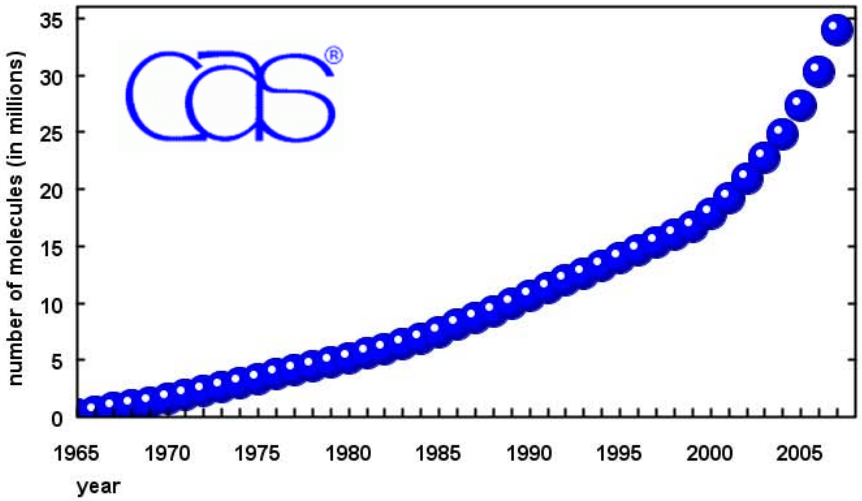
Peter Ertl

Novartis Institutes for BioMedical Research
Basel, Switzerland

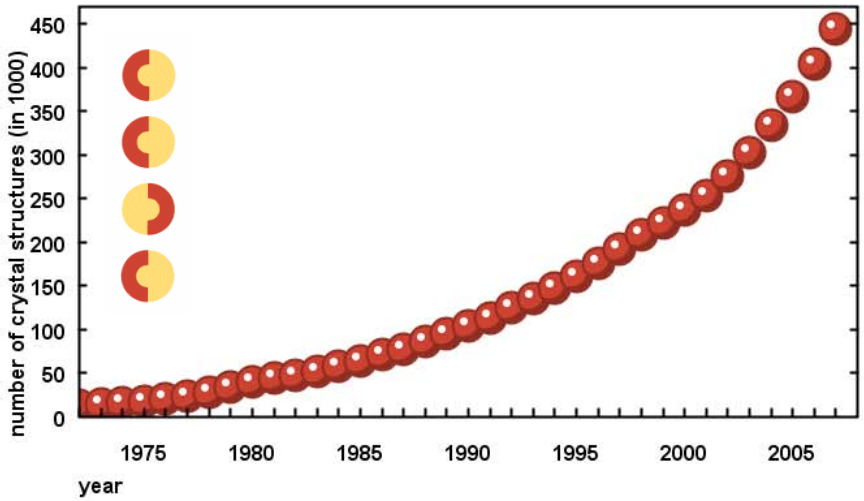
With thanks to my colleagues:
J. Mühlbacher, B. Rohde, A. Schuffenhauer and P. Selzer

NOVARTIS
NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH

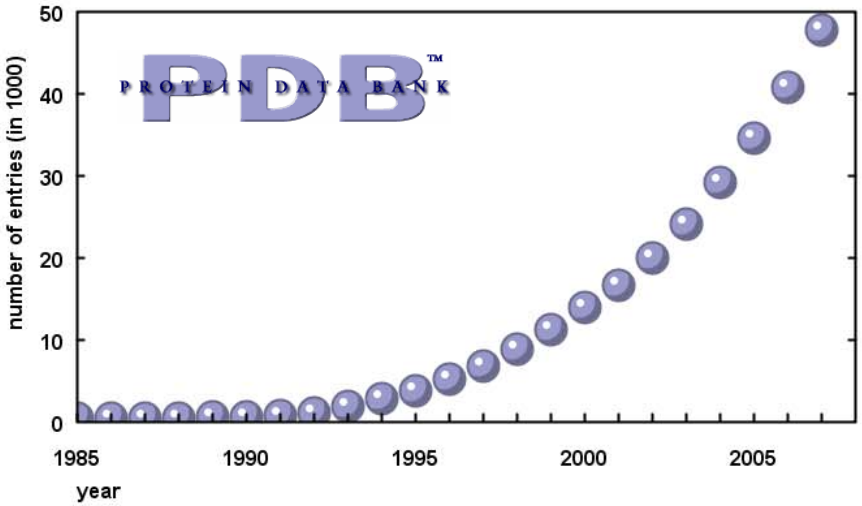
Data Explosion in Chemistry



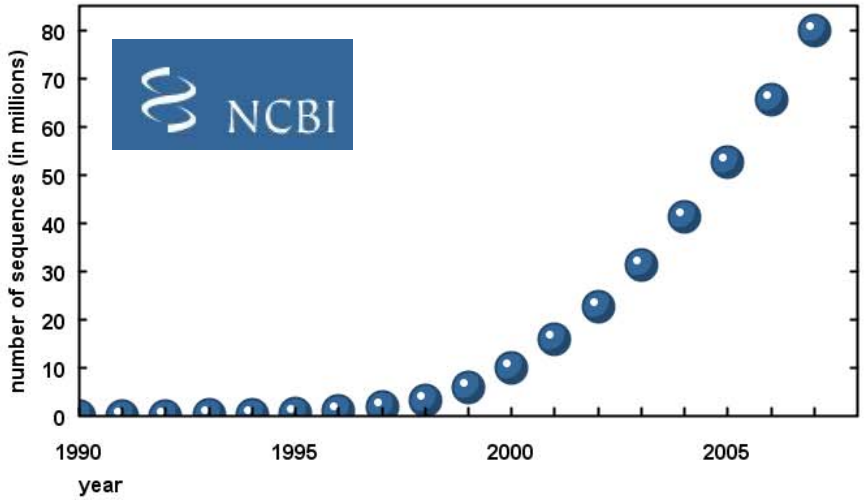
CAS – 34 million molecules



CCDC – 445'000 structures



PDB – 48'000 proteins

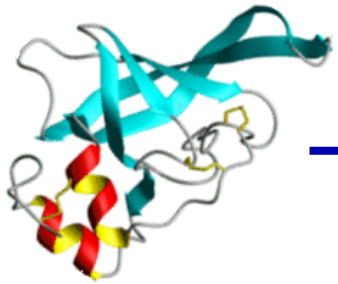


GenBank – 80 million sequences

Drug Discovery & Development Process

Discovery →

Disease



Identify a protein involved in disease (target)

Find a molecule effective against the target

Preclinical testing

Human clinical trials

Approval by regulatory authorities

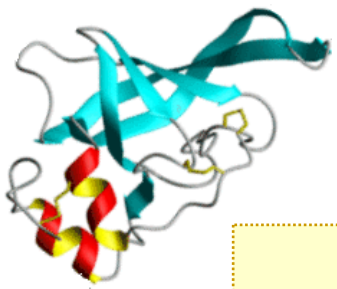
Drug on the market

Development →

The whole process takes 10 – 15 years and costs ~1 billion USD !

Cheminformatics in the Drug Discovery

Target identification



Lead finding

Lead optimization

Chemical Genomics

Bioinformatics

Genomics
Proteomics
Systems biology
Pathway analysis

Cheminformatics

Molecular databases
Combinatorial chemistry
HTS screening support
Data mining
Virtual screening
Property Calculation

QSAR
***In silico* ADME**
Toxicity alerting
Bioisosteric design

Cheminformatics in the Pharma Industry

- **“applied” cheminformatics – ultimate goal = design of new drugs**
- **processing of very large data sets - millions of structures + related information (screening results, experimental and calculated properties, spectra, availability, synthesis information …)**
- **high requirements on methodology validation**
- **direct feedback by experiment (chemistry, biology, experimental properties)**
- **large number of users, operation in a complex global environment**
- **security / confidentiality issues**

Typical Cheminformatics Activities at Pharmaceutical Industry

1. **Molecular databases**
2. **Large-scale data analysis, knowledge discovery**
3. **Calculation of molecular properties / descriptors**
4. **Estimation of ADME characteristics, toxicity alerting**
5. **Navigation in chemistry space**
6. **Virtual screening**
7. **Support for HTS – hitlist triaging**
8. **Support for combinatorial chemistry and molecule optimization**

Novartis Web-based Cheminformatics System

Easy to use "do it yourself" cheminformatics and molecular processing tools for synthetic chemists, available on the company intranet.

- **first tools introduced in 1995**
- **currently more than 20 tools available**
- **open, modular, platform and vendor independent architecture**
- **integration with other scientific applications**
- **more than 1'800 registered users**
- **used from all Novartis research sites (Tsukuba, Wien, Basel, Horsham, Cambridge, San Diego, Singapore)**
- **over 5'000 jobs submitted each month**
- **20 million molecules processed per year**

Web-based cheminformatics tools deployed via corporate Intranets,
P. Ertl, P. Selzer, J. Mühlbacher, BIOSILICO 2, 201, 2004

1. Molecular Databases

Databases in pharmaceutical companies :

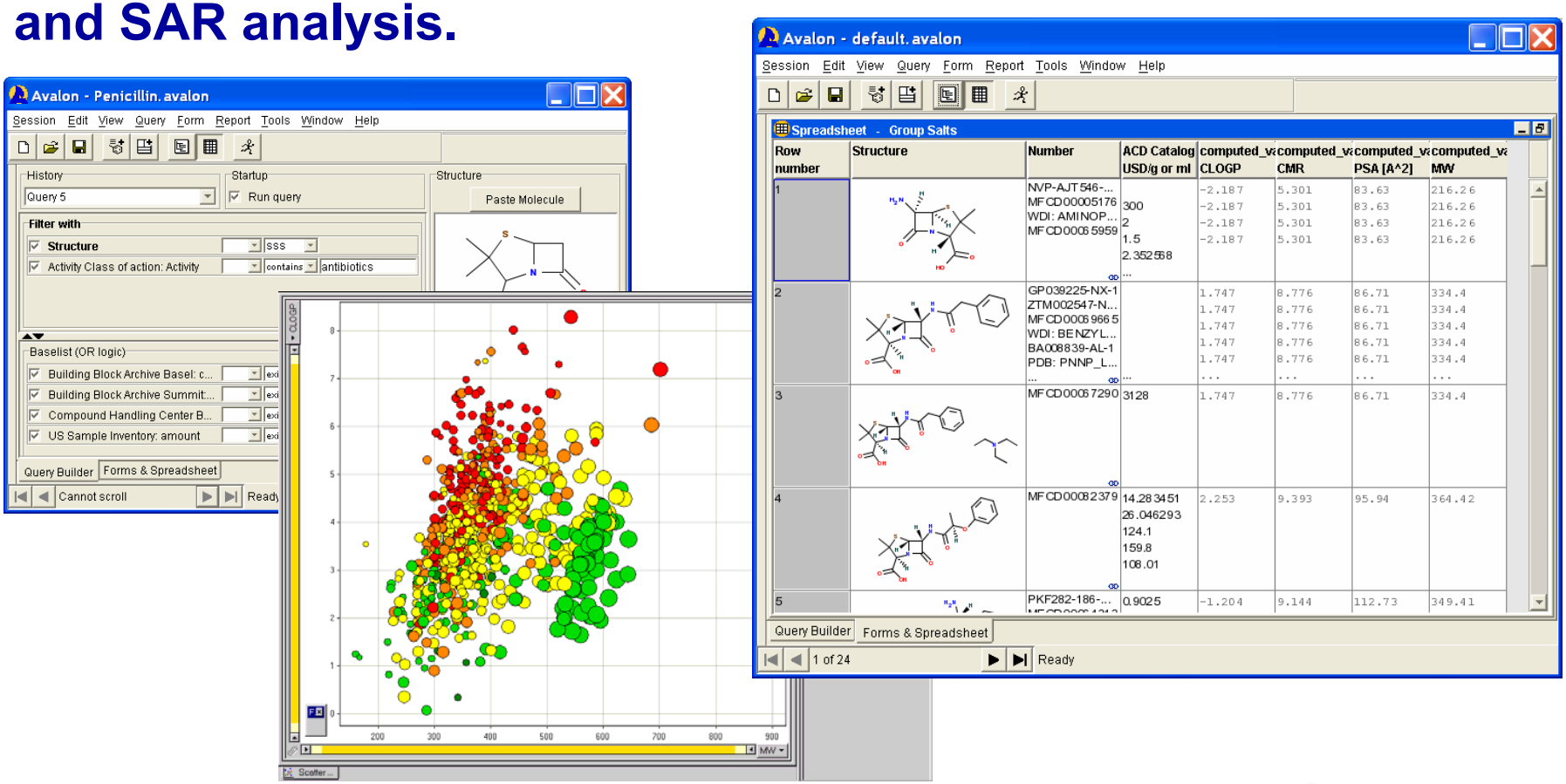
- **millions of structures + related data**
- **normalization of chemical structures (nitro, tautomers …)**
- **all data need to be validated and checked for correctness**
- **interface must support user-friendly data mining and visualisation of large datasets**
- **responsiveness - substructure and similarity searches within seconds**
- **chemically interpretable results - pharmacophore searches, pharmacophore fingerprints**

Current trends :

- **data warehouses**
- **chemistry cartridges**

Novartis Data Warehouse - Avalon

In-house database written in Java, containing all in-house and many reference structures, results of biological screens and many additional data. Allows efficient data-mining, reporting and SAR analysis.



2. Large-Scale Data Mining

Data Mining = Knowledge Discovery in Large Databases

analyzing large amount of data to obtain useful information (in a form of pattern, rule, cluster …) leading to understanding of relationships within data and correct decisions

Data mining techniques used in cheminformatics:

- **classical QSAR, regression analysis**
- **Bayesian statistics**
- **clustering**
- **neural networks**
- **decision trees**
- **…**

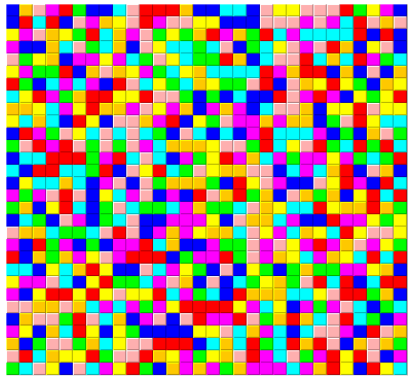
Self-Organizing Neural Networks

Self-organizing (Kohonen) NN is a mathematical tool used to simplify complex multidimensional data by reducing their dimensionality, allowing thus visual processing.

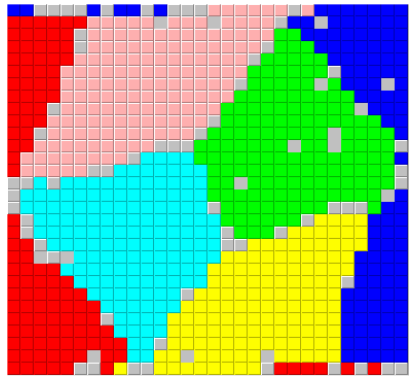
Processed data are expressed as a 2-dimensional map.

■	3.45	4.56	2.38	6.78	9.45	…
■	5.78	9.45	6.45	4.23	3.45	….
■	2.38	3.45	5.44	6.45	5.78	…
■	6.45	5.78	5.44	1.23	4.78	…
■	5.44	4.78	6.23	5.28	3.45	…
■	6.23	5.44	4.67	6.34	5.78	…
■	6.23	5.44	4.67	6.34	5.78	…
■	6.45	5.78	5.44	1.23	4.78	…
■	3.45	4.56	2.38	6.78	9.45	…
■	6.45	5.78	5.44	1.23	4.78	…
■	5.44	4.78	6.23	5.28	3.45	…
■	6.23	5.44	4.67	6.34	5.78	…
■	6.23	5.44	4.67	6.34	5.78	…
■	6.45	5.78	5.44	1.23	4.78	…
■	3.45	4.56	2.38	6.78	9.45	…
…	…	…	…	…	…	…

→



→

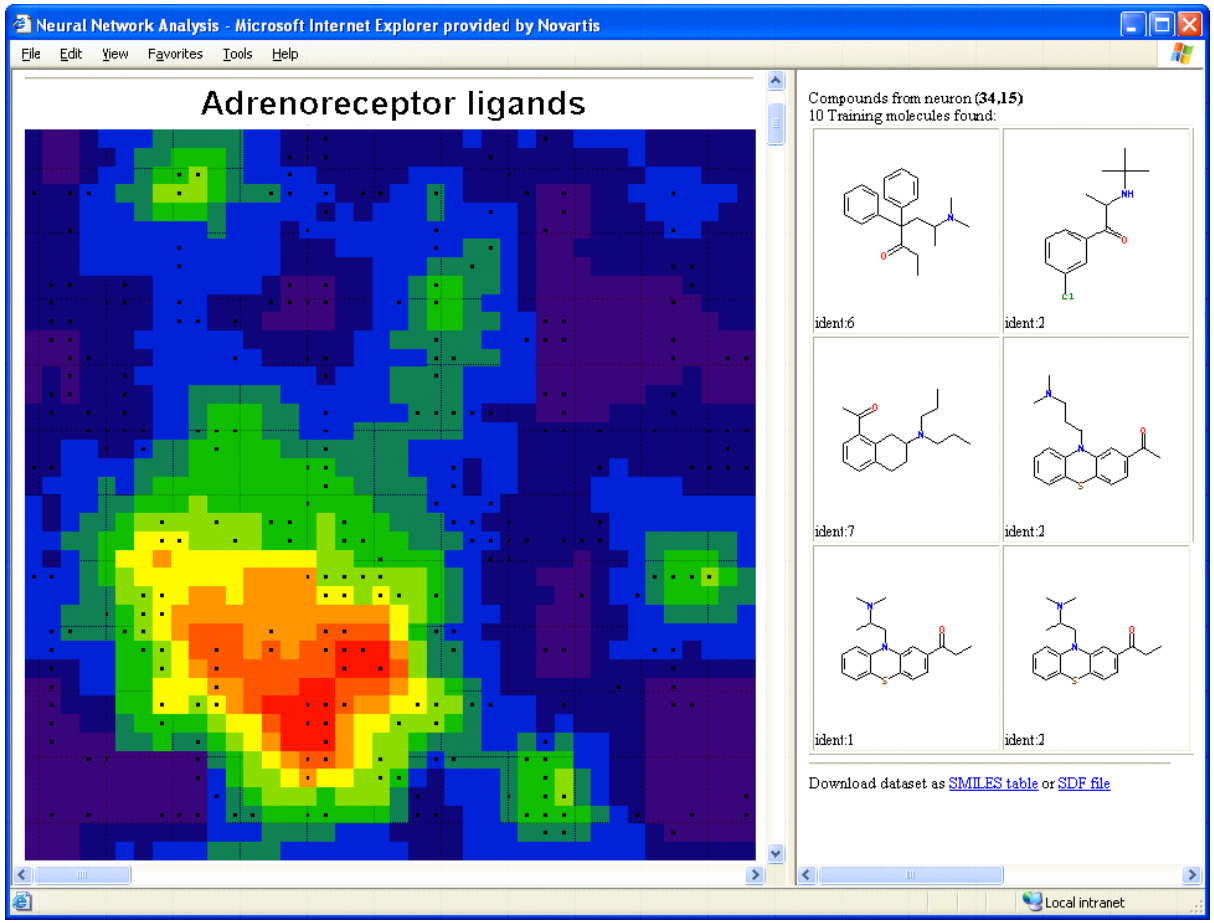


Data table **Initial network** **Trained network**

Display data on a 2D map **Unsupervised training**

Classification of GPCR Ligands

Identification of properties and structural features typical for GPCR ligands by self-organizing neural networks.



Identification and Classification of GPCR Ligands using Self-Organizing Neural Networks

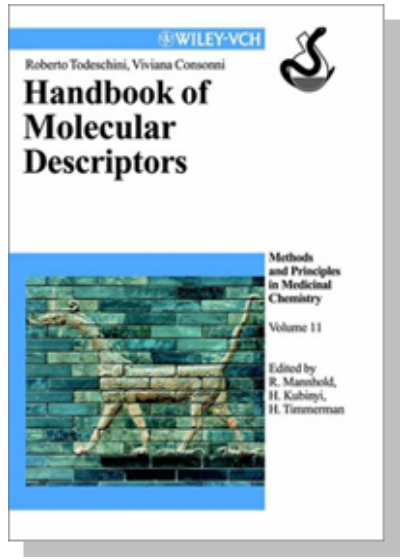
P. Selzer, P. Ertl

QSAR & Comb. Sci. *24*, 270, 2005

Calculation of Molecular Properties

- properties need to be calculated for datasets containing ~10^6 molecules (in-house data, virtual libraries, catalogues)
- calculations need to be fast
- descriptors should be interpretable, physically meaningful
- properties should cover all important types of protein-ligand interactions

Currently the most useful global properties are logP, MW, PSA (polar surface area), HBD and HBA counts, number of rotatable bonds. Many others are used, but they are less interpretable + highly intercorrelated.



R. Todeschini, V. Consonni,
Handbook of Molecular Descriptors, Wiley, 2000
Lists >8000 various molecular descriptors

Novartis *In Silico* Profiling

InSilico Profiling - Microsoft Internet Explorer provided by Novartis

File Edit View Favorites Tools Help

InSilico Profile [v3.2]

Show all results — Peter Ertl (18 Apr 2006 16:13:24)

Global Menu

Contact
Feedback

Navigation
Tools
Home

Tools
Profiling
ToxCheck
Filtering
Portal

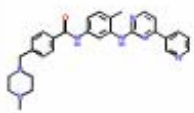
Tool Menu

Actions
Recalculate
New Query

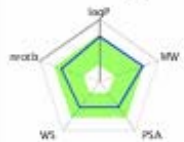
Documentation
Radar plot poster for Vienna
Radarplot Presentation
T2ABS plot publication
User Guide to the T2 Absorption Model

Download
SDF
smiles
TSV (for Excel)

CGP057148-NX: [top] [first] [previous] [next] [last] [bottom]



Bioavailability plot



(Touch with mouse pointer to enlarge)

$C_{29}H_{31}N_7O$

CGP057148-NX

Calculated Properties:

CLogP:	4.529	H-bond acceptors:	8
Molecular Weight:	493.62	H-bond donors:	2
CMR:	14.632	Amide groups:	1
Rule of 5 violations:	0	Violation Classes:	No
PSA:	86.28 Å^2	Rotatable Bonds:	7
Volume:	550.31 Å^3	Flexibility Index:	14.18

Structure Check:

inSilico ToxCheck: NEW

Metabolism alerts:
no alerts

Predictions: (changed 2002-2-22)

Caco2 permeation properties:	global model
Pm [10^{-5} cm/min]:	40.74
logPm:	-3.39
Absorption:	97.71
T2ABS: [Show T2ABS Plot]	good (3.85)
logBB:	-0.35
Water solubility:	2

Visualization

Show 3D Hydrophobicity | 3D Structure Visualizer | JM WebLab Viewer / Get mol2 file | Open ISISDraw

Pageowner: Joerg Muehlbacher — brought to you by the Cheminformatics group

Local intranet

4. ADME-related Properties

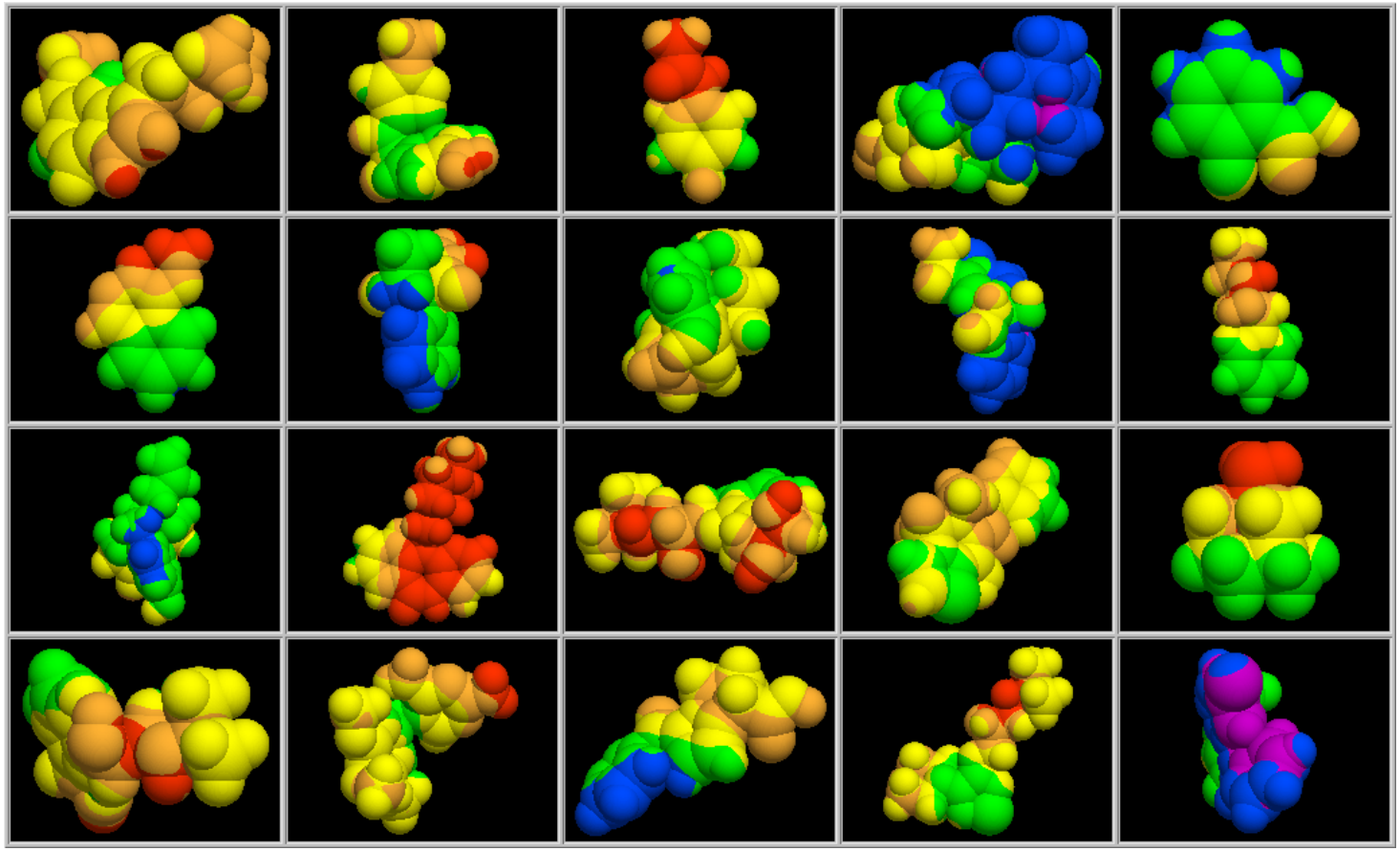
Properties related directly to the biological effect of drugs and their fate in organism, and therefore frequently needed in medicinal chemistry.

- **water solubility**
- **pKa - acidity / basicity estimation**
- **drug transport characteristics**
  - **intestinal absorption**
  - **blood-brain barrier penetration**
  - **Caco-2 permeability**
  - **plasma-protein binding**
  - **efflux**
- **toxic and metabolic characteristics**

Challenges:

- **these properties describe complex physical and biological processes**
- **not enough experimental data to build reliable models**

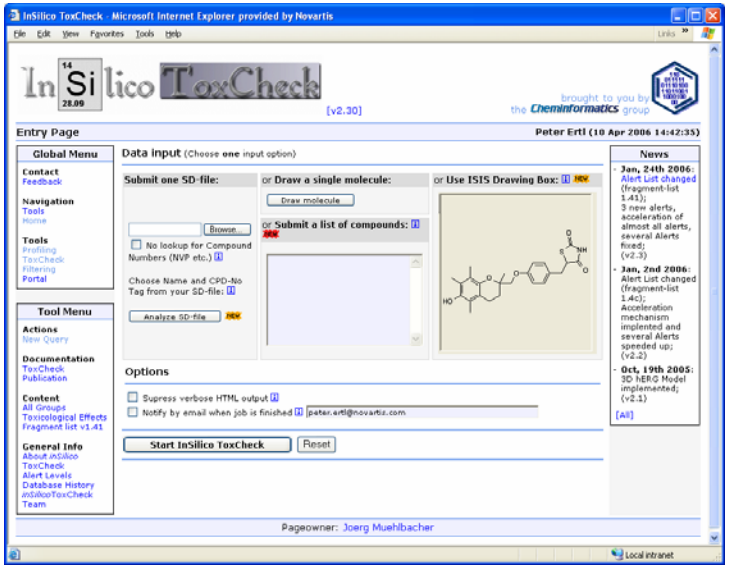
3D Hydrophobicity

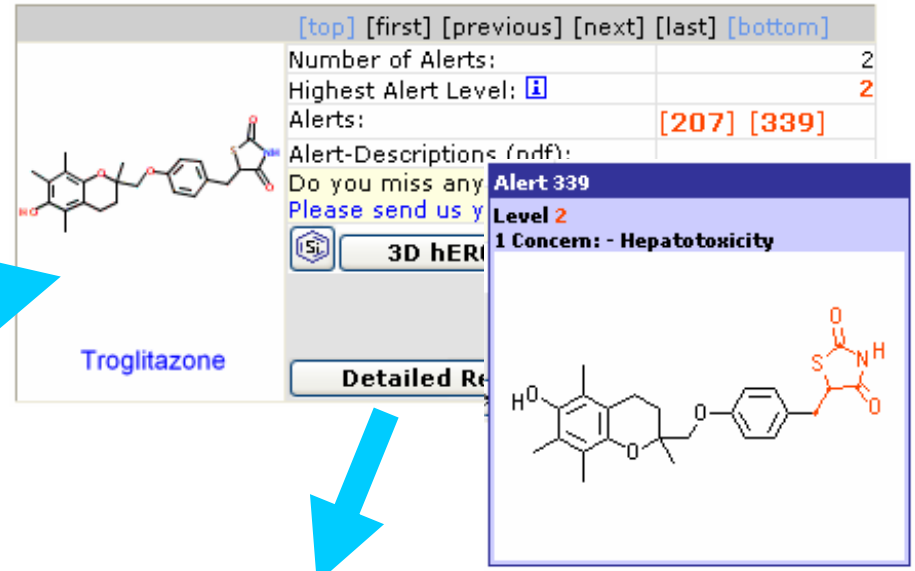


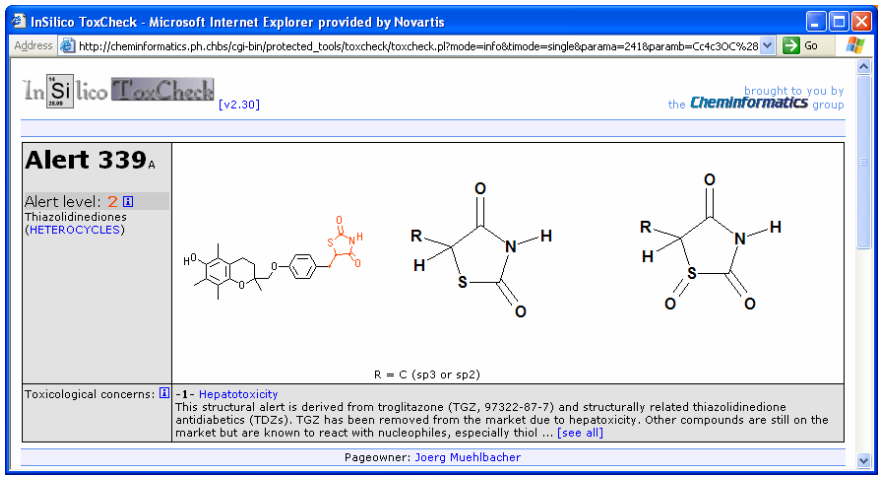
hydrophobic hydrophilic

All molecules have the same logP ~1.5, but different 3D MLP pattern.

Novartis *In Silico* ToxCheck







5. Navigation in Chemistry Space

Size of the known chemistry space:

- **35 million molecules registered in CAS**
- **19 million compounds in PubChem**
- **36 million entries in the Chemical Structure Lookup service**
- **~500,000 molecules with (known) biological activity**

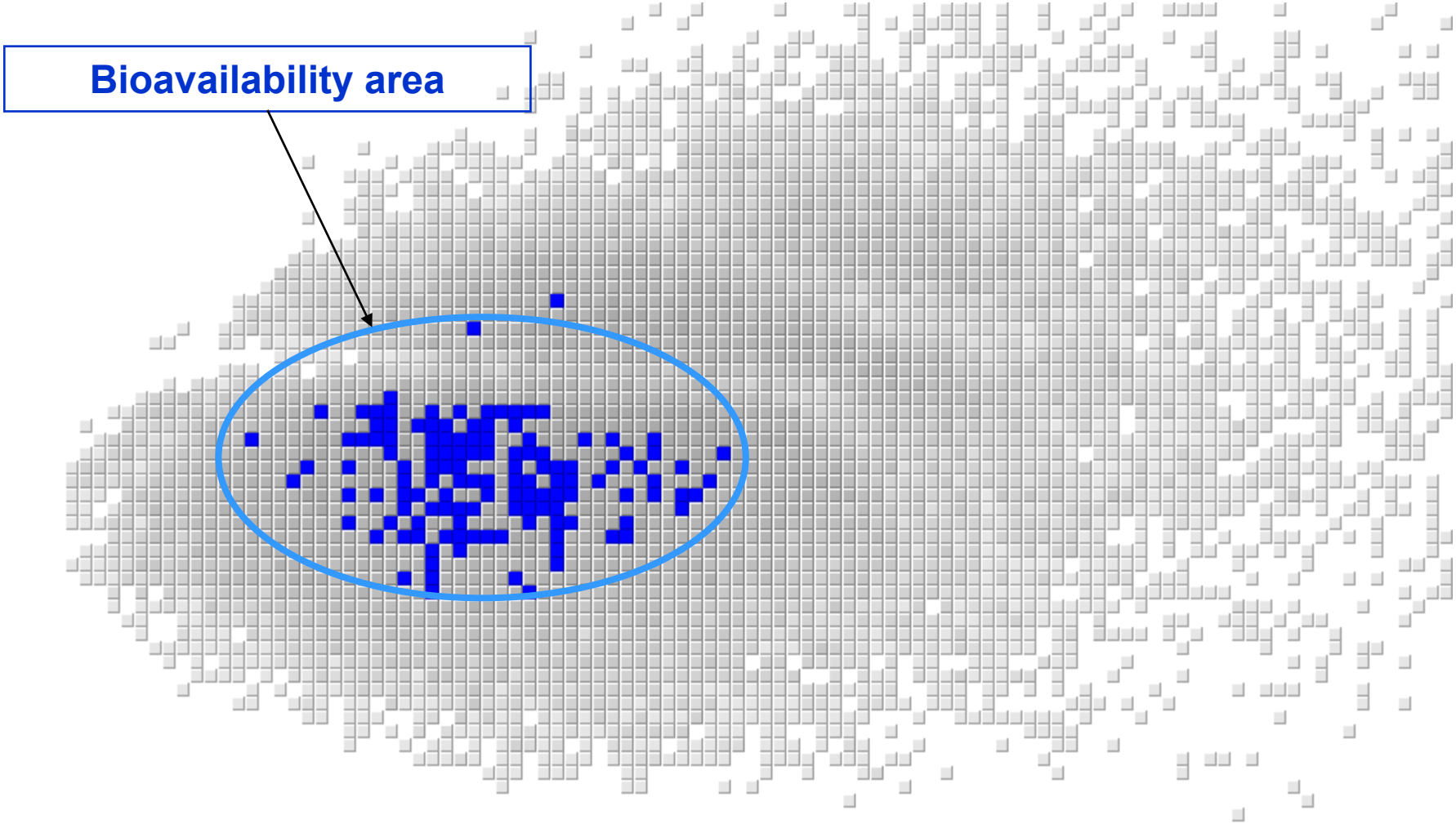
And VERY large number of possible (virtual) molecules

Chemistry space is multidimensional; to process / understand it, we need to characterize it and to reduce its dimensionality.

Chemistry space may be characterised by:

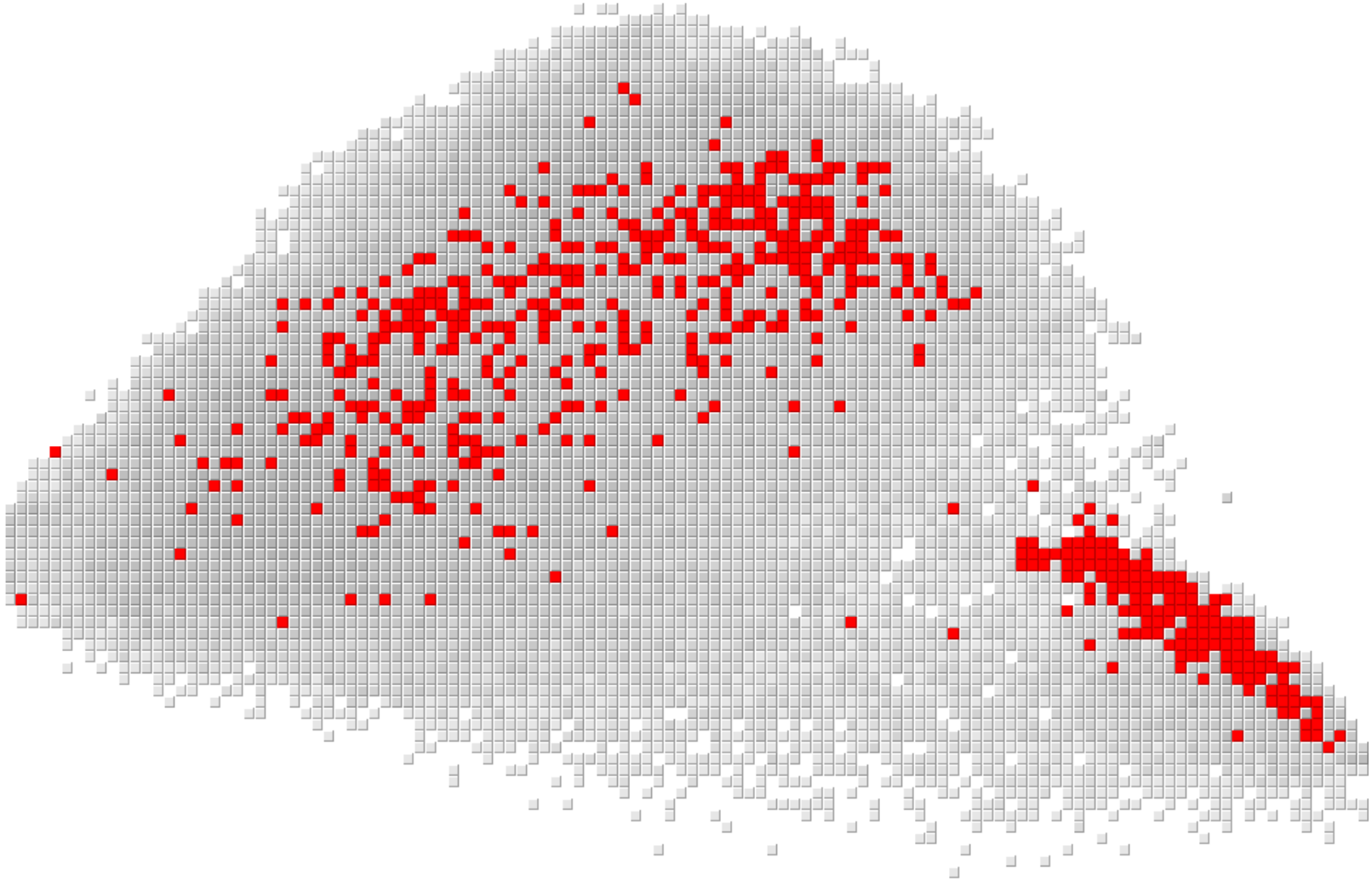
- **physicochemical global molecular properties (logP, PSA …)**
- **substructural features (fragments, fingerprints, pharmacophores ...)**

Molecular Property Space



■ organic molecules, ■ drugs

Structural Diversity Space



organic molecules, drugs

The Scaffold Tree

Separation of molecule universe into smaller parts – clusters:

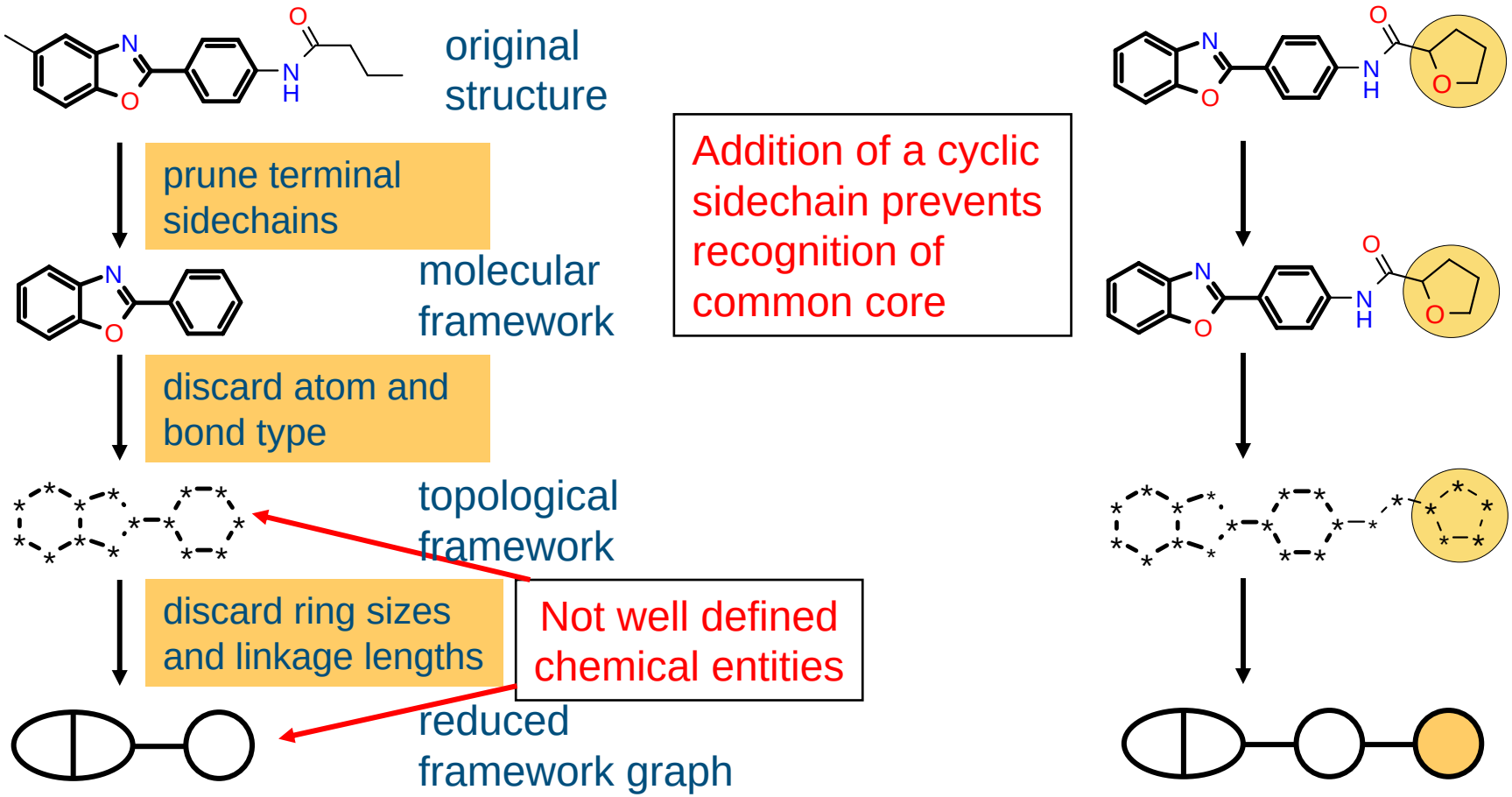
Clustering

- classification derived from unsupervised machine-learning
- information of complete dataset is required for classification
- no incremental updates possible
- n^2 or n.log(n) time scaling
- not easy interpretable

Rule-based

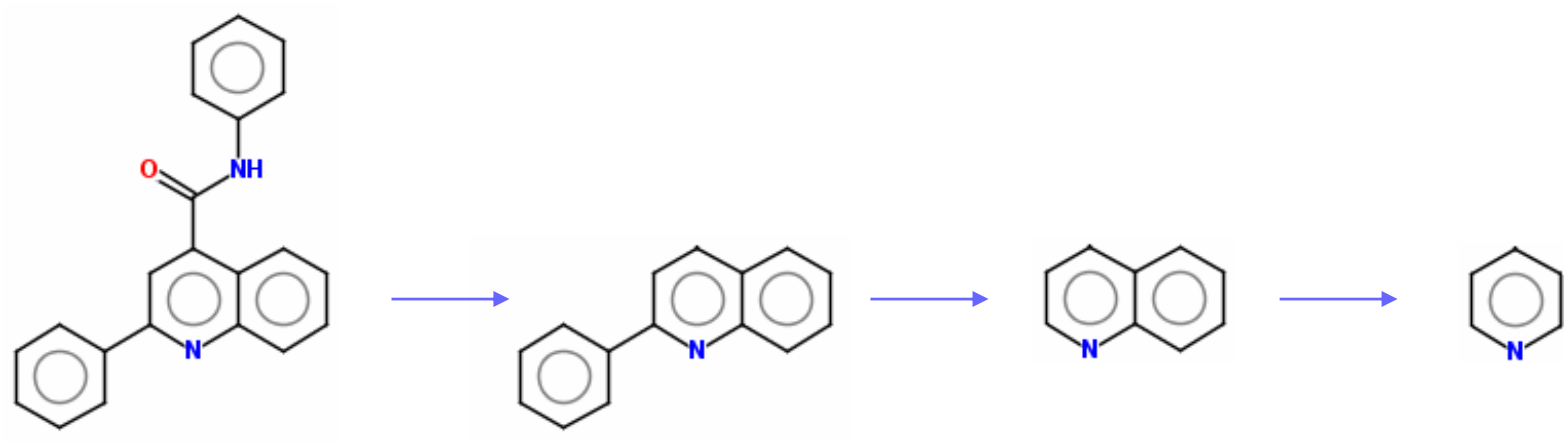
- explicitly formulated rules encode “expert knowledge”
- class assignment is derived for each structure independently - scales linearly with number of molecules in dataset
- incremental updates possible
- better perceived by chemists

The Molecular Framework and its Generalizations



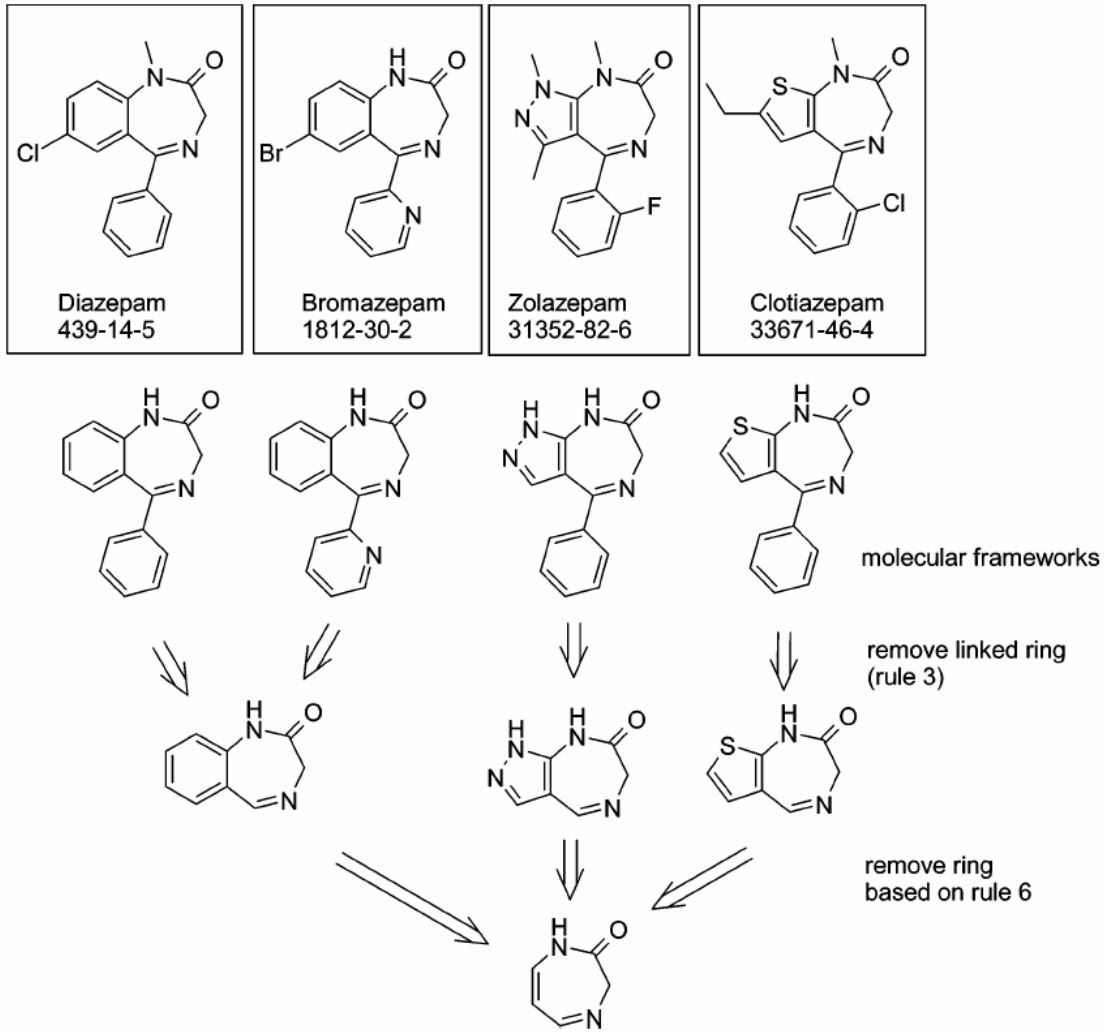
The Scaffold Tree – Basic Algorithm

- retain the molecular framework as classification element
- exocyclic and “exolinker” double bonds are part of the molecular framework
- instead removing atom & bond type and ring size information prune less important rings one by one
- use prioritization rules to decide which ring to remove first
- use small, generic set of rules, no lookup “dictionary”



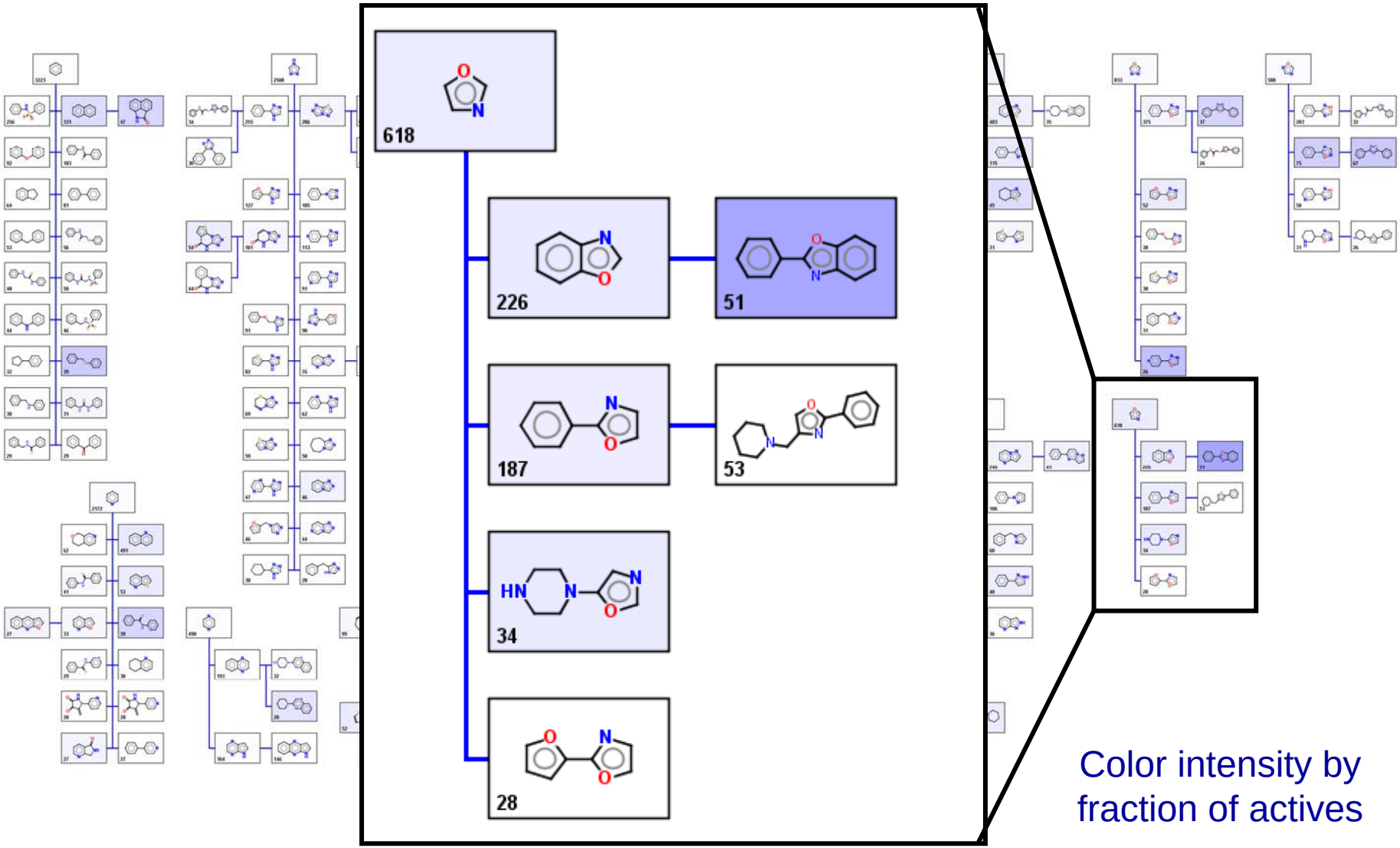
The Scaffold Tree – Visualization of the Scaffold Universe by Hierarchical Scaffold Classification
A. Schuffenhauer, P. Ertl *et al.* J. Chem. Inf. Model., *47*, 47, 2007

Classification of Diazepinenones



Scaffold Tree Example for HTS results

PubChem Pyruvate Kinase Data Set



Color intensity by fraction of actives

6. Virtual Screening

Selection of molecules having the highest probability to be active and to be developed to successful drugs from a large collection of screening samples or virtual molecules.

In-house company archives contain 2-5 million molecules (in house synthesis, acquisitions, mergers, combichem libraries).

20-30 million screening samples available commercially →

Selection criteria:

- **reliable properties (solubility, stability, absence of too reactive fragments) - drug-likeness**
- **no toxicity / adverse effects**
- **diversity / novelty**
- **target focus**

Virtual Screening Workflow

No information about target
~ 10^6 structures / h

Cleaning the data - neutralizing, removing counterions, smaller parts, duplicates

Removal of "junk" – uninteresting, organometallic, toxic, reactive, unstable ...

Bioavailability screen - PSA, Rule of 5 ...

Screen based on the active substructures / Bayesian methods

Cheminformatics

3D pharmacophore known
~ 10^4 – 10^5 / h

3D pharmacophore search

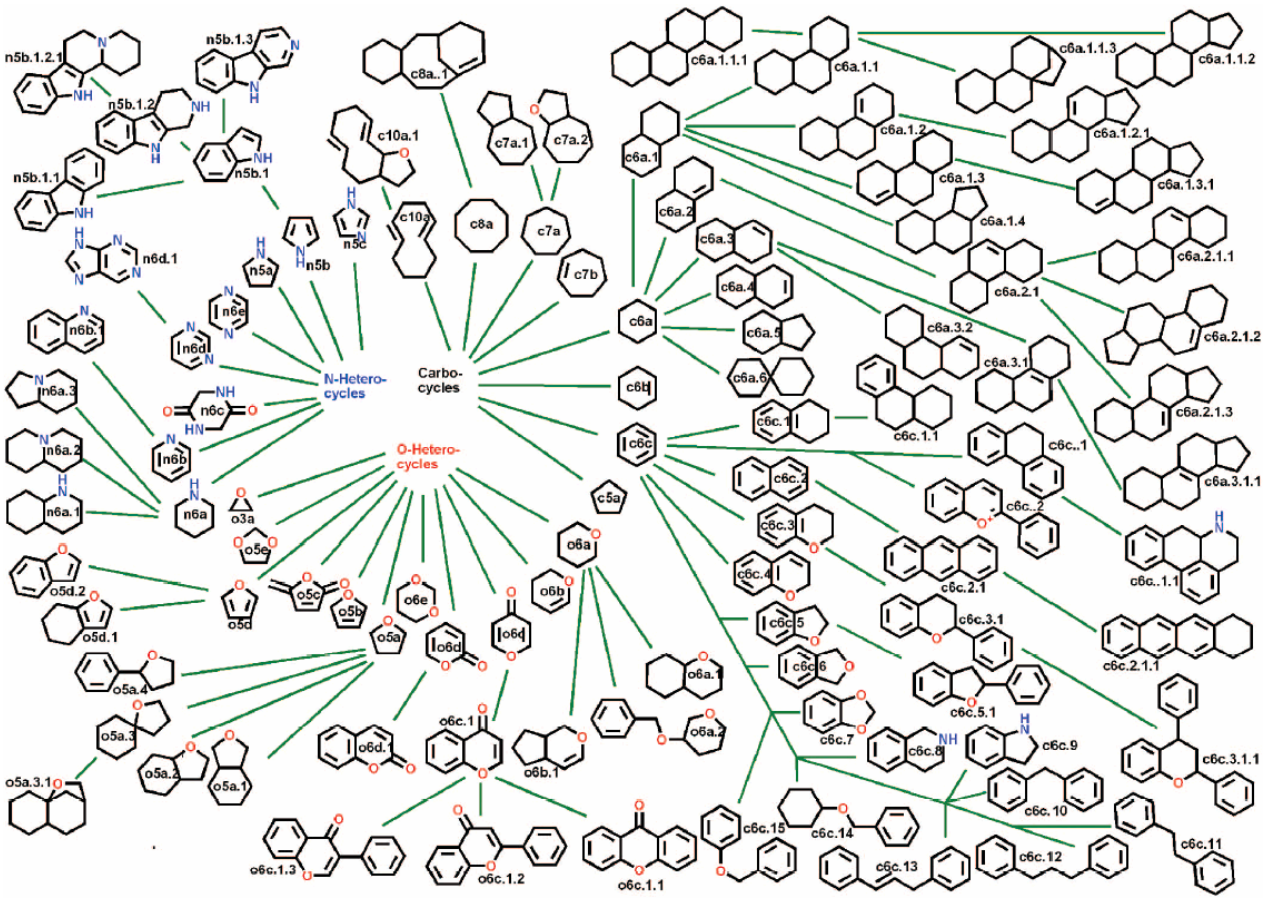
Receptor known
~ 10^3 – 10^4 / h

Virtual receptor screening / docking

Molecular modelling

Learning from the Nature

Natural products (NPs) have been optimized in a several billion years long natural selection process for optimal interaction with biomolecules.



NP molecules are therefore an excellent source of substructures for the design of new drugs.

Charting biologically relevant chemical space: A structural classification of natural products (SCONP)

M.A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl and H. Waldmann

PNAS *102*, 17272-17277, 2005.

7. High-Throughput Screening - HTS

Screening of >1 million molecules on many targets routinely in an automatic way.

Challenges for cheminformatics are

- **to process screening results and identify hits, worth of further follow-up - lead identification, hitlist triaging**
- **support of new types of screening (high content screening, pathways)**

HTS Workflow

- run HTS, collect the data
- identify "active" compounds (based on % inhibition cut-off)
- organize actives into groups (clustering, maximal substructure analysis, common scaffold)
- visualize clusters of actives
- analyze inactives to identify those related to active series
- selected actives (primary hits) are further confirmed in dose/response assays to get EC_{50} values, secondary assays and chemical validation to get validated hits
- use machine learning techniques to develop SAR models for validated hits

8. Combinatorial Chemistry Molecule Design

- synthesis of compounds as ensembles (libraries)
- technology was introduced in the early 90s
- advantages : speed & economics - combination of scaffolds and Rgroups allows creation of very large number of molecules quickly in automatic manner

Cheminformatics issues – library design:

- how large should be combichem libraries?
- which Rgroups and scaffolds to combine?
- diverse (DOS) libraries or targeted libraries?
- how to fill the “holes” in the chemistry space?

Early CombiChem

Results of early combichem were quite a disappointment.

Early combichem libraries were:

- **very large (100'000s molecules)**
- **molecules were large, hydrophobic, not diverse**
- **low hit rates**

This led to:

- **introduction of "drug likeness" – design of compounds with good physicochemical properties**
- **targeted libraries - design of smaller, more focused libraries when information about target is available (i.e. kinase libraries)**
- **use diverse libraries covering broadly chemistry space when little information about target is available – "primary screening" libraries**

Library Design Strategies

Two basic design strategies:

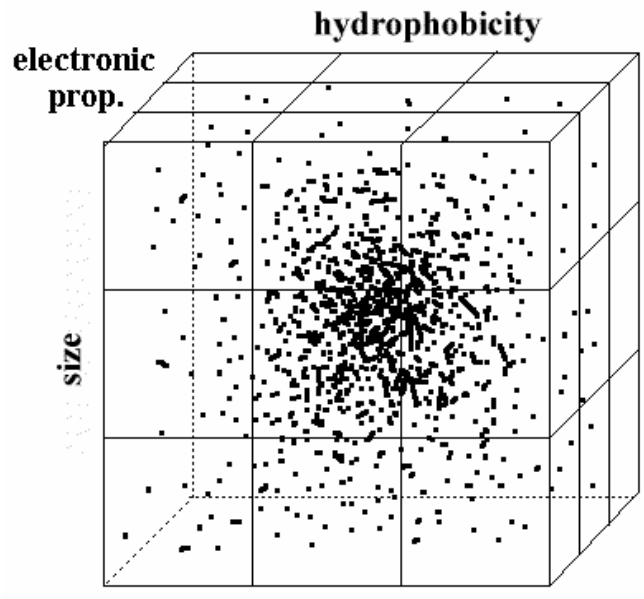
- **reactant-based – building blocks are selected based only on their properties not considering properties of products**
- **product-based – selection of monomers based on the properties of final products. This approach is much more computationally demanding but is more effective**

Trends in modern CombiChem:

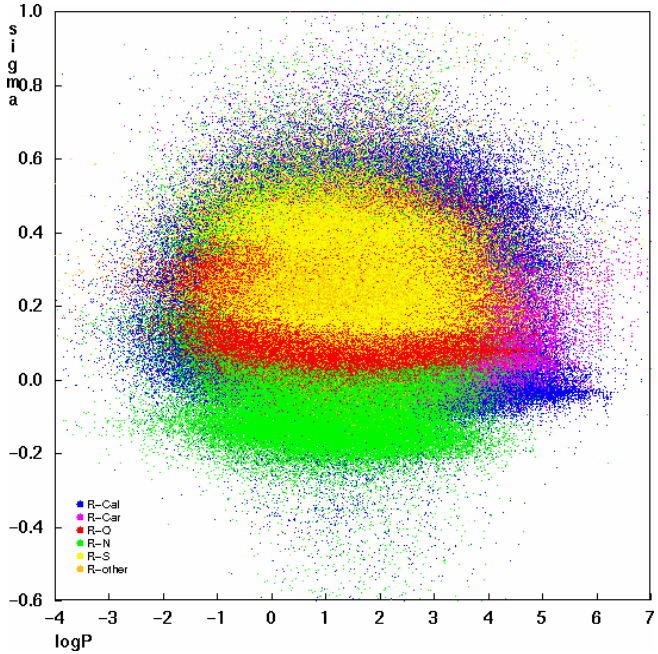
- **smaller (1000s molecules), targeted libraries**
- **multiobjective optimization (Pareto optimization) – optimize at the same time properties, coverage of chemical space, price …**
- **information from pharmacophore search or docking used in design**
- **natural product-like libraries**

Database of Organic Substituents

850'000 substituents extracted from organic molecules and characterised by their calculated hydrophobicity (Hansch *π* constant), donating/accepting power (Hammett σ) and size.



substituent "property cube"



logP / σ plot for 850'000 Rgroups

Cheminformatics analysis of organic substituents, P. Ertl, J. Chem. Inf. Comp. Sci., *43*, 374, 2003

Bioisosteric Design

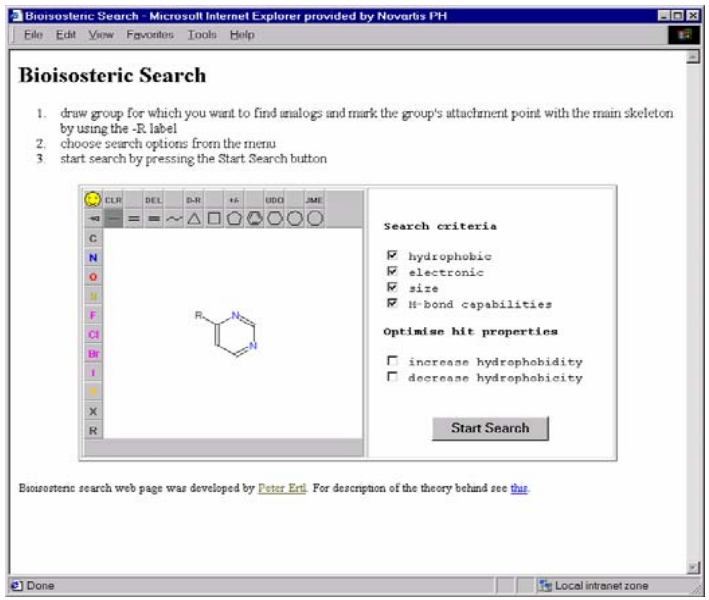
Bioisosteric replacement - replacement of a functional group or spacer in a bioactive molecule by another functionality having similar size and physicochemical properties.

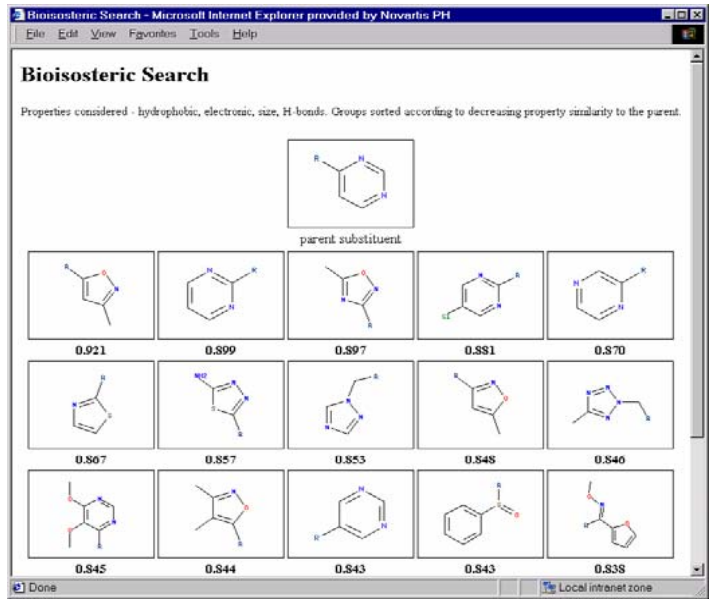
Bioisosteric transformation are used to :

- **optimise properties of drug candidates (activity, selectivity, transport characteristics)**
- **remove side effects (toxicity)**
- **design molecules easier to synthesise**
- **avoid patented structural features**

Substituent Bioisosteric Design

- **identification of substituents and spacers bioisosteric (i.e. physicochemically compatible) with the target**
- **based on > 10’000 drug-like fragments with calculated properties**
- **results may be used as “idea generator” for the design of new non-classical bioisosteric analogs.**





Cheminformatics – Future Trends

- **global databases, integration of multiple data sources, public (Wikipedia-like) curation**
- **use of large chemogenomics databases (WOMBAT, GVK …)**
- **text and image mining, automatic extraction of useful information from publications and patents**
- **integration with bioinformatics, with focus on ligand protein interactions and pharmacophores**
- **disappearing border between cheminformatics and computational chemistry**
- **in technology area – modularization, web services**
- **open source collaborative software development**