

Substance Databases and Bioisosteric Compounds

Problems:

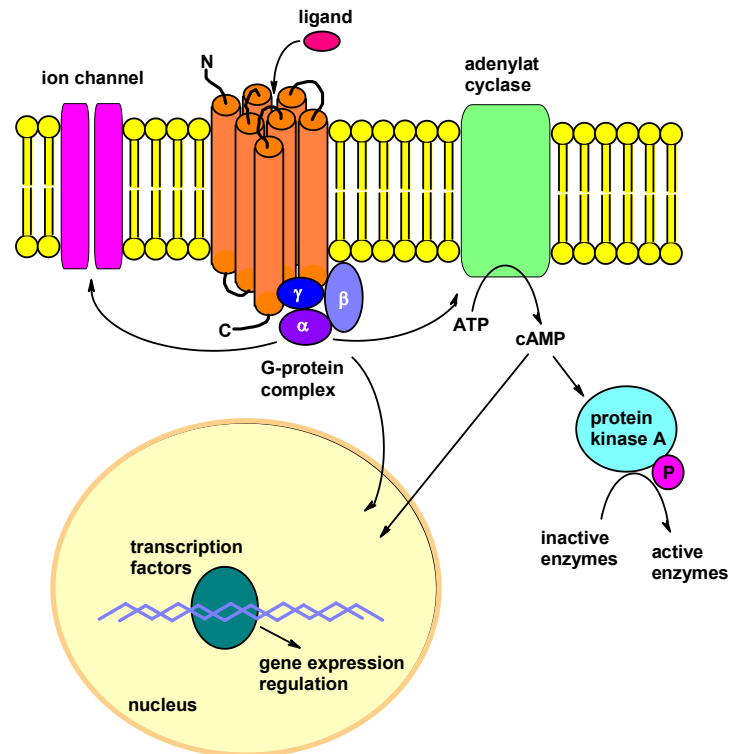
- a) How to choose promising compounds for experimental/virtual screening?
- b) How to automate screening (more compounds tested = more hits?)
 1. step: choice of *target*
 2. step: How much information about the *target* is available?
Are there any *lead compounds* present already?
 3. step: if yes, generate a virtual substance library based on the lead compound(s) → find/generate similar compounds
 4. step: planning of synthesis (combinatorial chemistry)

Setup of substance libraries for high throughput screening (I)

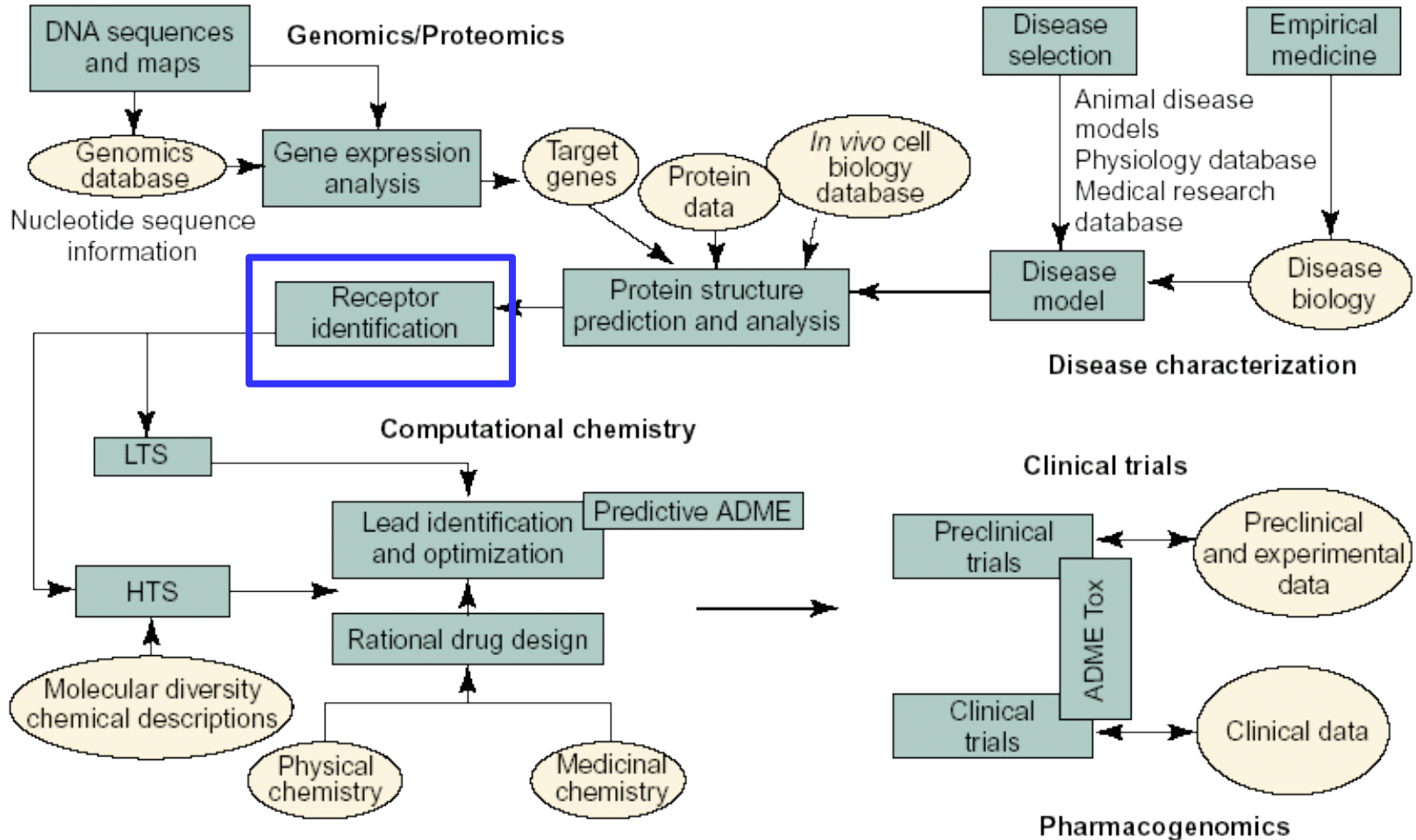
automated test of >1000 compounds on the *target*

Requires the synthesis of the according number of substances and processing of the results

1. step: choice of *target*

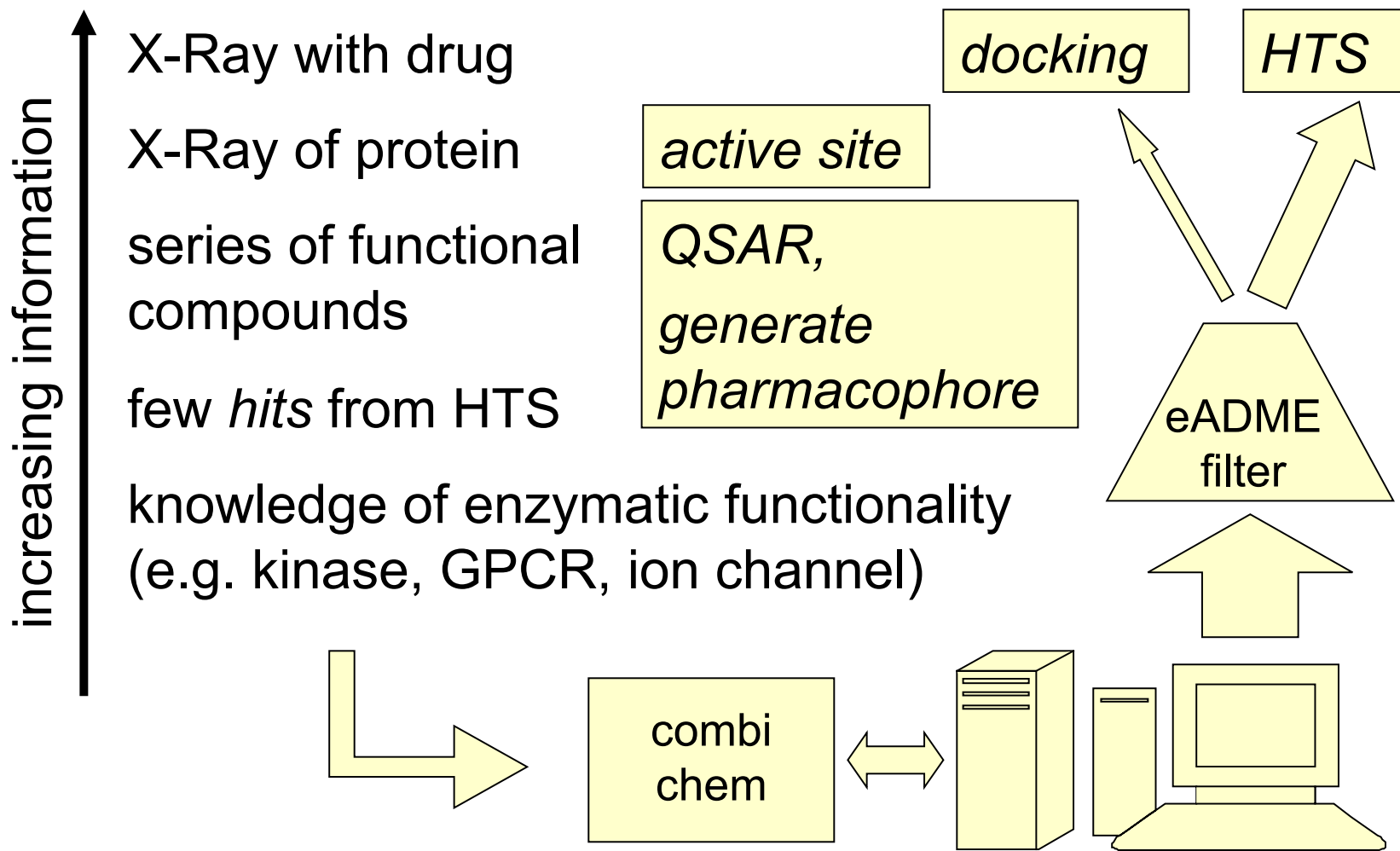


Flow of information in a *drug discovery pipeline*



Compound selection

How much information about the *target* is available?



Setting up a virtual library

Properties of combinatorial libraries

Combinatorial libraries are also tailored to their desired application:

random libraries

drug-like / diverse scaffolds

focused libraries

lead-like / most comprehensive for a certain class of enzymes

targeted libraries

one single enzyme /
substituents as diverse as possible

Chemogenomics

aim: maximum diversity of substance libraries

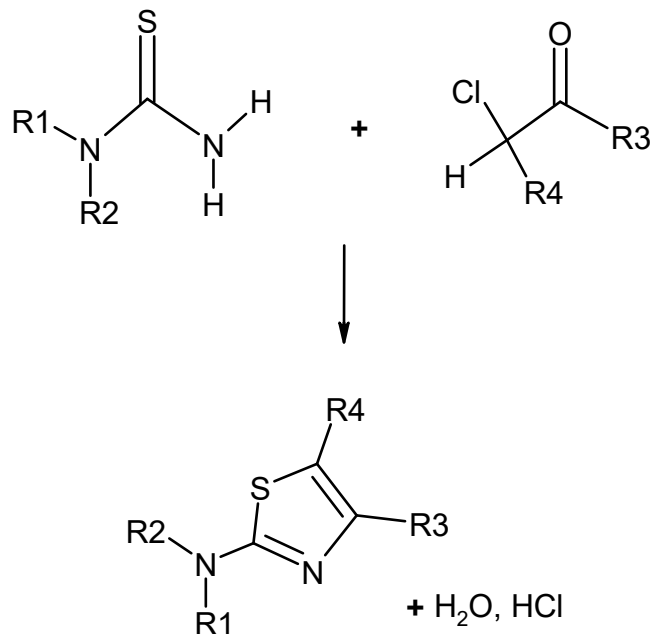
avoiding redundant compounds

improved probability of hits in the HTS

Combinatorial approaches in rational drug design

automated tests of >1000 compounds on a single *target* require particularly effective synthesis and screening strategies:

- synthesis robots
- High Throughput Screening



Original idea: The more compounds being tested, the higher the likelihood of finding a lead compound *should* be.

Setup of substace libraries for the High Throughput Screening (IV)

Synthesis of a multitude of compounds based on a lead compounds required a change in paradigms.

Until the late 80' substances selected for screening were synthesized one by one individually.

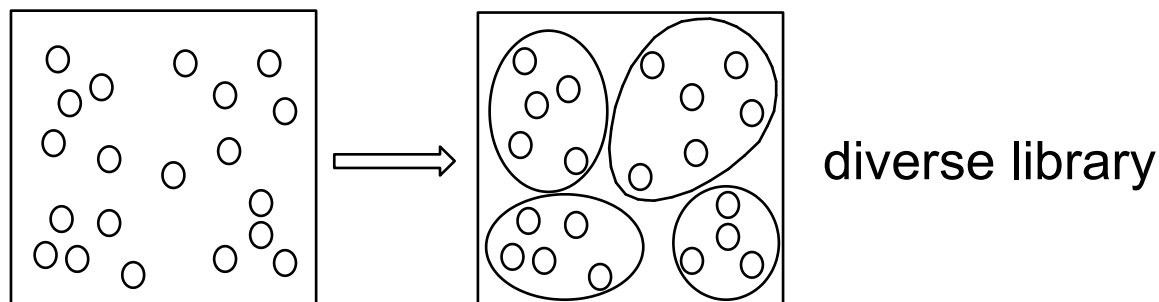
The principles of High Troughput Screening required, however, a different approach.

„If you are looking for the needle in the haystack, it is best not to increase the size of the haystack.“

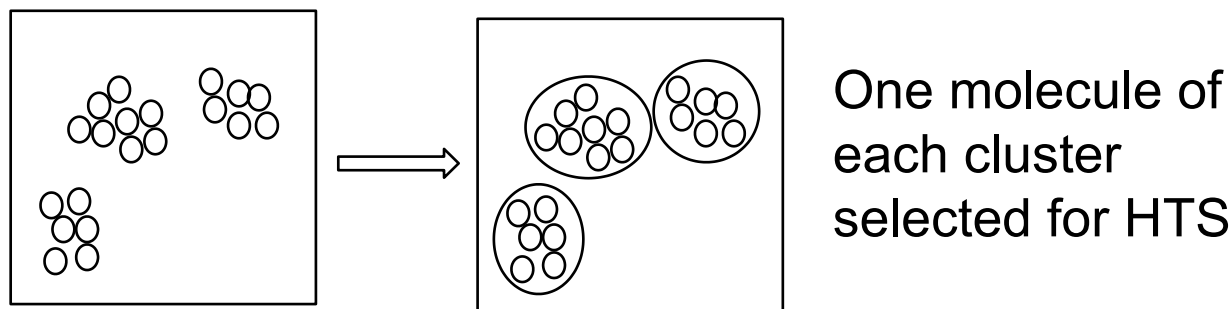


Clustering in sets of data (I)

To evaluate the diversity of a data set, respectively a generated substance library, the obtained compounds have to be grouped to clusters.

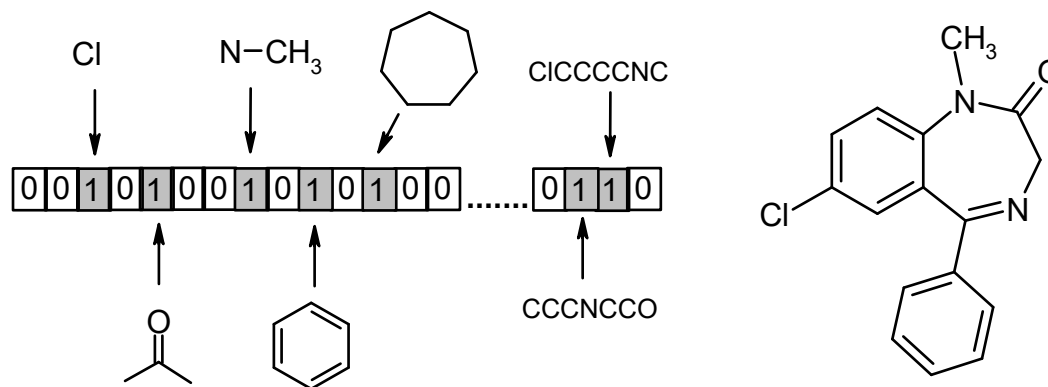


Test further molecules of the same cluster that produced a hit in the HTS



The assignment of the molecules is based on their pair-wise similarity. → Encode molecules in terms of features.

Encoding of Molecules for Data Base Storage



Each present feature set the corresponding bit on
→ binary *fingerprint* of the molecule

Pro : Resulting bit string allows efficient storage, retrieval and comparison (bit-wise AND, OR, EOR operations)

Con: Choice of predefined features is arbitrary and may lead to bias of predefined features

Classification of compounds (I)

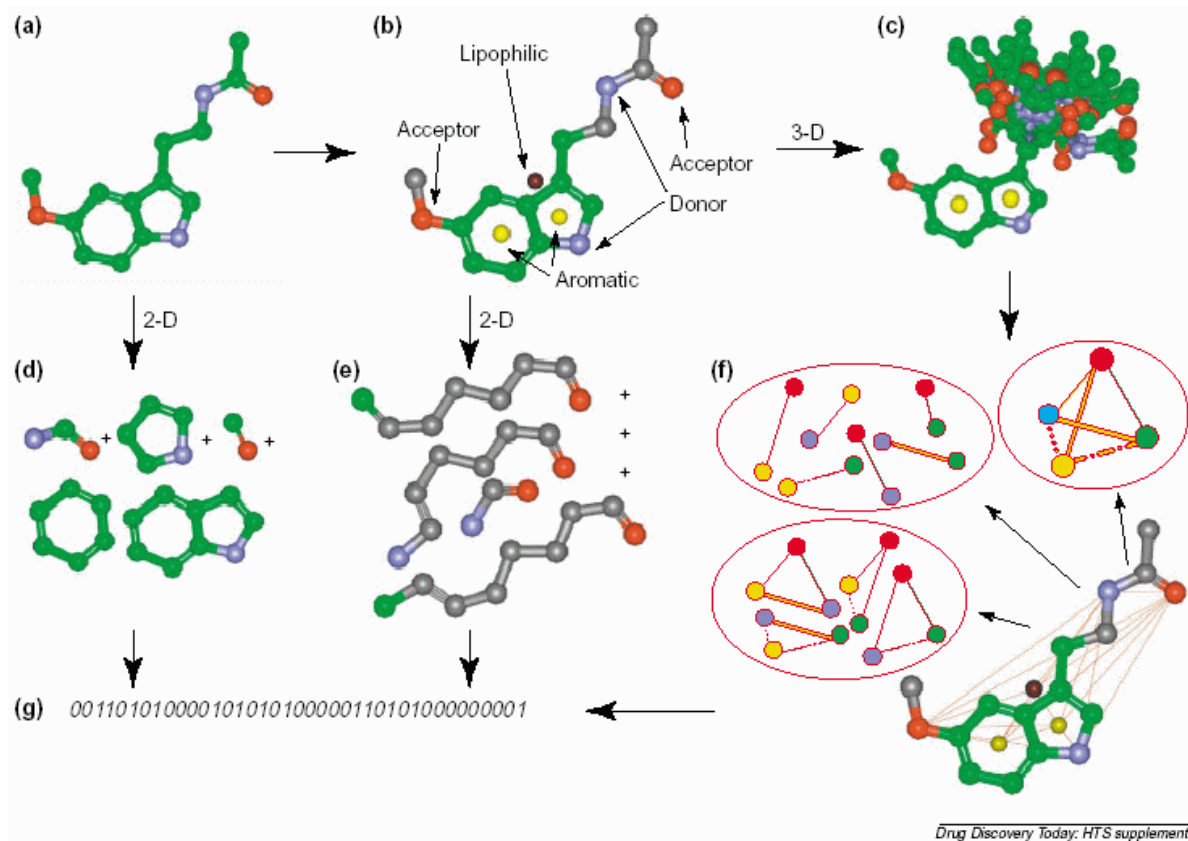


Figure 2. Schematic illustration of primary methods used in molecular fingerprint creation. (a) Create 2-D and 3-D model of molecule; (b) deconstruct the molecule into pharmacophoric elements; (c) generate conformational models; (d) deconstruct the molecule into topological/substructural elements; (e) determine distance between pharmacophoric groups using bond counts; (f) determine 2-, 3- or 4-center distance combinations of pharmacophoric groups for each conformer; and (g) determine the presence or absence of each descriptor element and combine to create a binary fingerprint.

Using pharmacophoric features to obtain a binary *fingerprint* of a molecule

Fingerprints (I)

Similarity of two molecules A and B represented as fingerprints is computed (most frequently) via the Tanimoto coefficient/index (Jaccard index)

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

common bits of A and B (intersection)

all bits of A and B (union, length of fingerprint)

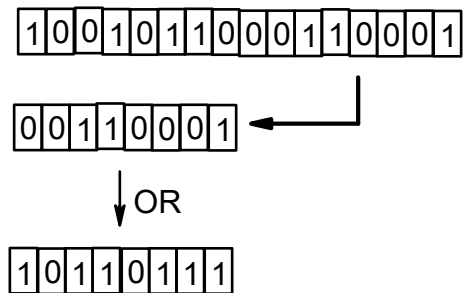
Inherent problem: with increasing length of a fingerprint the chance of common bits is decreasing; so does the information density. As a consequence similarity values become lower than expected. Thus the discriminatory power in virtual screening worsens.

Lit: M.Vogt & J.Bajorath *F1000 Research* **9** (2020) 100.

Fingerprints (II)

How to „increase“ similarity scores

Folding of fingerprints: perform a logical OR operation on both halves, which yields a shorter fingerprint.



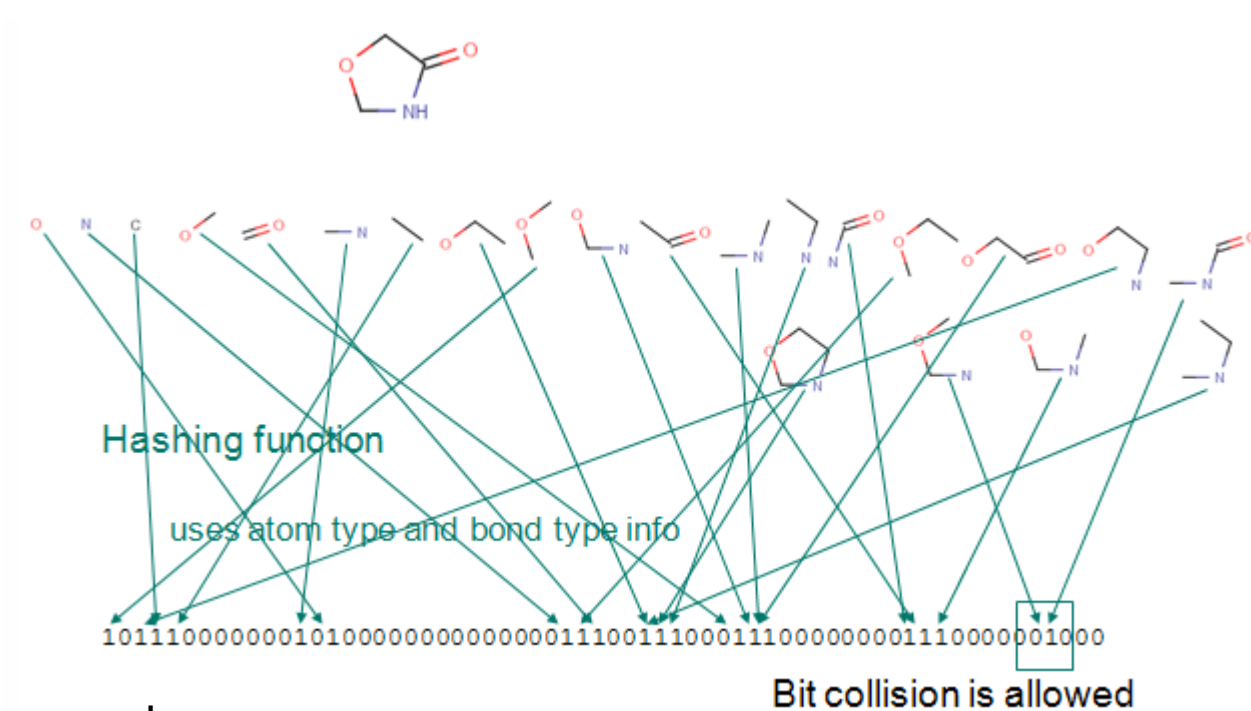
Pro: results in higher bit density and likewise higher similarity Tanimoto coefficients

Con: increases the number of false positive hits (features that are not actually present in the molecule)

Fingerprints (III)

How to „increase“ similarity scores

Hashing of fingerprints: Like usual hashing functions single features are rearranged, resp. grouped together (bit collision)



Source: chemaxon.com

Con: A single feature can no longer be identified unambiguously, but *similar substructures* can be grouped together.

Classification of compounds (II)

Frequently applied fingerprint concepts are:

- Daylight fingerprint (1024 bits) → see also openbabel
- ISIS MOLSKEYS (atom types, fragments of molecules)
- Circular/Morgan/Extended Connectivity Fingerprints
takes the neighborhood of an atom into account
Lit: Rogers & Hahn *J.Chem.Inf.Model.* **50** (2010) 742.

- Topological Torsion

take 1-4 atom type sequences into account

Lit: Nilakatan et al. *J.Chem.Inf.Comput.Sci.* **27** (1987) 82.

- 2D-Pharmacophore Fingerprints

use predefined features

Lit: Gobbi & Poppinger *Biotech.Bioeng.* **61** (1998) 47.

see also RDKit for python implementation www.rdkit.org

Comparison of fingerprints and their performance:

Lit. H.Briem & U.Lessel *Persp.Drug Discov.Des.* **20** (2000) 231.

S.Riniker & G.A.Landrum *J.Cheminf.* **5** (2013) 26.

Classification of compounds (III)

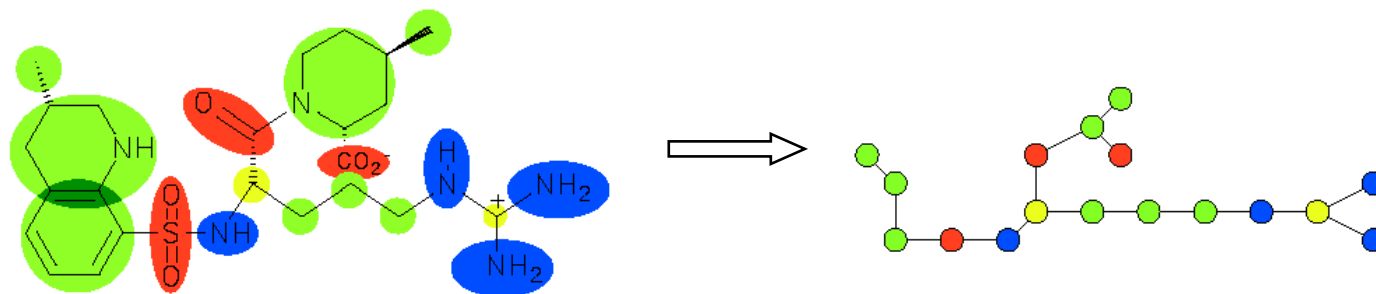
FTREES feature trees concept:

each node (in a molecule) represents a chemical feature

Lit. M.Rarey & J.S.Dixon *J.Comput.-Aided Mol.Des.* **12** (1998) 471.

Allows to search for chemically similar compounds in large virtual substance libraries

Lit. M.Rarey & M.Stahl *J.Comput.-Aided Mol.Des.* **15** (2001) 497.

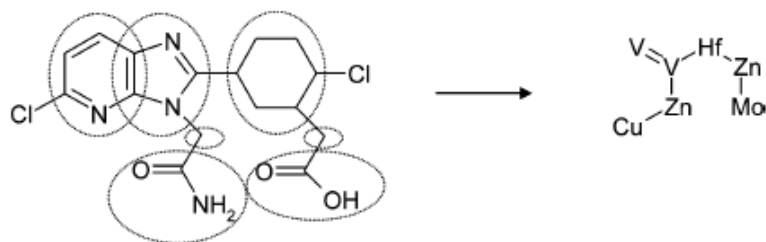


→ the molecule is represented as reduced graph.

The FTREES concept furthermore allows (fast) matching of subtrees to find similar compounds.

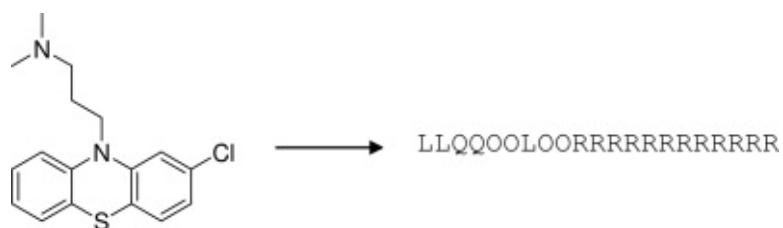
Classification of compounds (IV)

Comparison of molecules using (reduced) graphs:



Lit: V.J.Gillet and co-workers *J.Chem.Inf.Model.* **46** (2006) 577.

Comparison of molecules using alignments: PhAST, LINGO



LINGO freq.	LINGO freq.	LINGO freq.
"N(C)" ... 1	"ccc(" ... 1	"c0Sc" ... 1
"0Sc0" ... 1	"cc00" ... 1	"0c0c" ... 1
"cccc" ... 2	")CCC" ... 1	"CN(C)" ... 1
"N0c0" ... 1	"L)cc" ... 1	"(L)c" ... 1
"C)CC" ... 1	"cc(L)" ... 1	"CN0c" ... 1
"c(L)" ... 1	"Sc0c" ... 1	"(C)C" ... 1
"ccc0" ... 1	")cc0" ... 1	"c0cc" ... 2
"0ccc" ... 2	"cc0S" ... 1	"CCN0" ... 1
	"CCCN" ... 1	

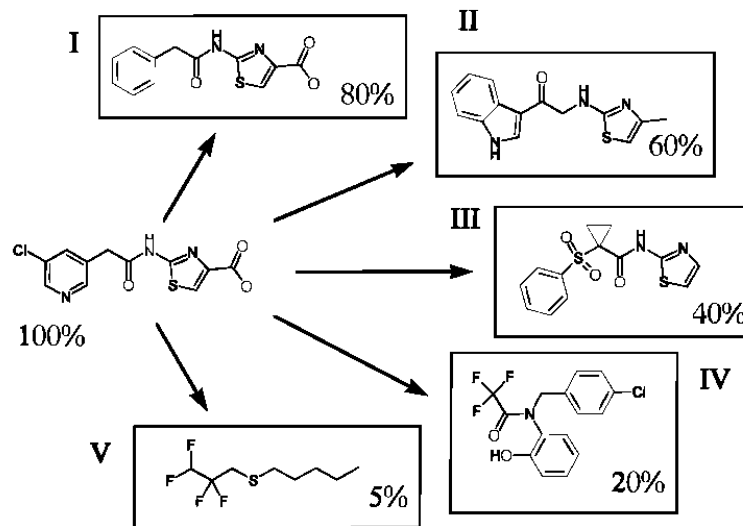
Lit: G. Schneider and co-workers *J.Comput.Chem.* **30** (2009) 761.

Lit: D. Vidal et al. *J.Chem.Inf.Model.* **45** (2005) 386.

Similarity of chemical compounds

The pair-wise similarity of two molecules can be expressed by **similarity indices** computed from their binary fingerprints.

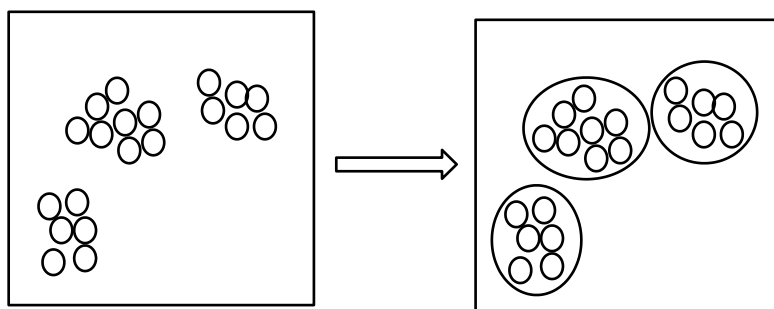
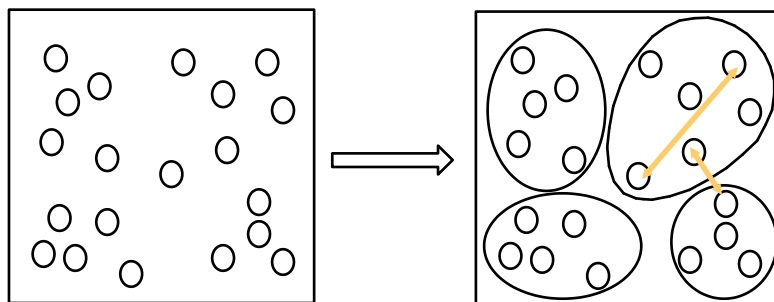
The comparison of binary data is computationally simple, but there are a number of different similarity indices. For the comparison of molecules the **Tanimoto index** is most frequently being used.



More about similarity indices in
lecture 6

Lit. D.R.Flower *J.Chem.Inf.Comput.Sci.* **38** (1998) 379.

Clustering in sets of data (II)



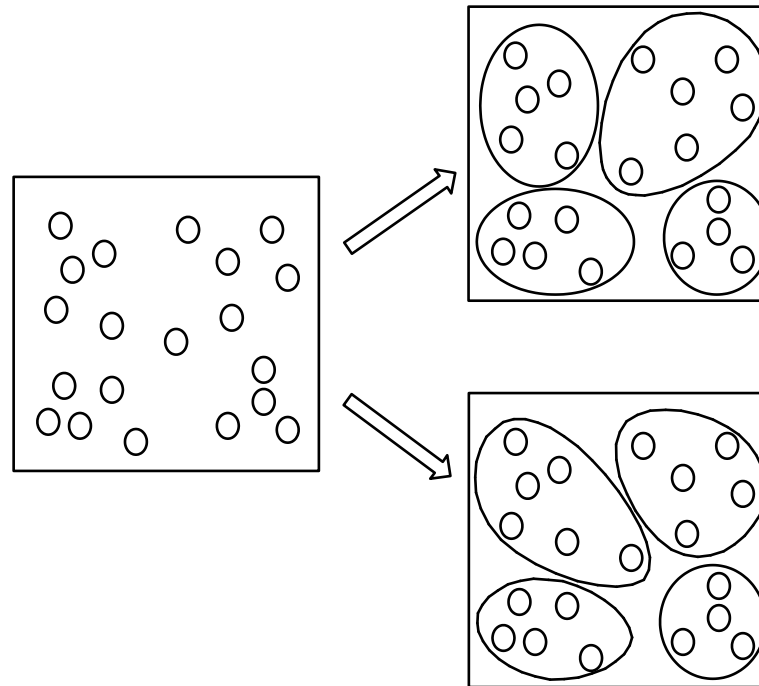
problem: The similarity of two molecules can be higher in between two different clusters than within the same cluster.

→ distance criteria (Euclidean, Manhattan, ...)

→ single linkage vs. complete linkage

Clustering in sets of data (III)

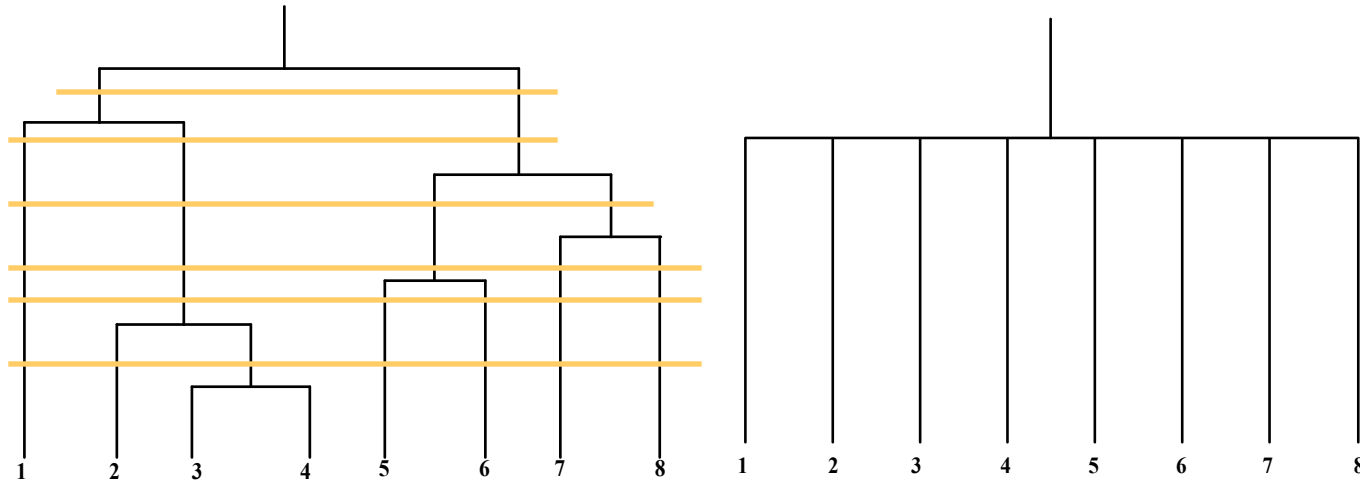
In general: Different algorithms for generating clusters will produce different clusters.



There is a „natural“ clustering in the data set, if different methods produce very similar looking clusters.

Methods of clustering (I)

There are two large groups of clustering algorithms:
hierarchical and non-hierarchical

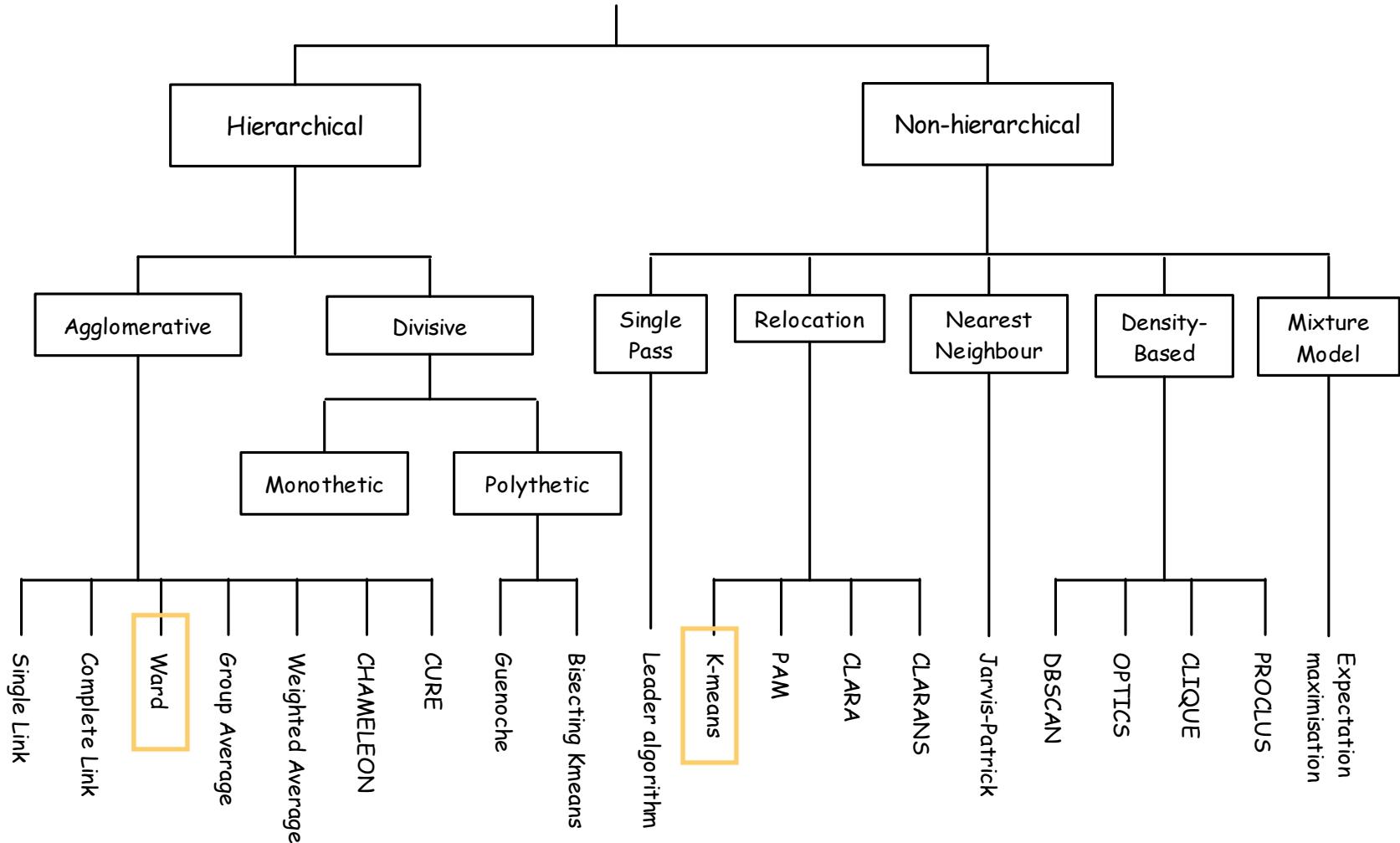


hierarchical clustering methods have the advantage of allowing access a each level.

all methods for clustering are computationally expensive !
runtime: $O(nN)$ to $O(n^2N)$ for n out of N molecules

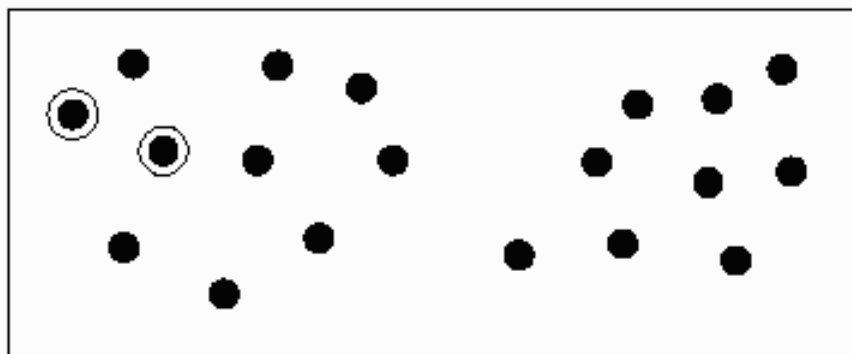
Methods of clustering (II)

„Clustering of clustering methods“- a dendrogram

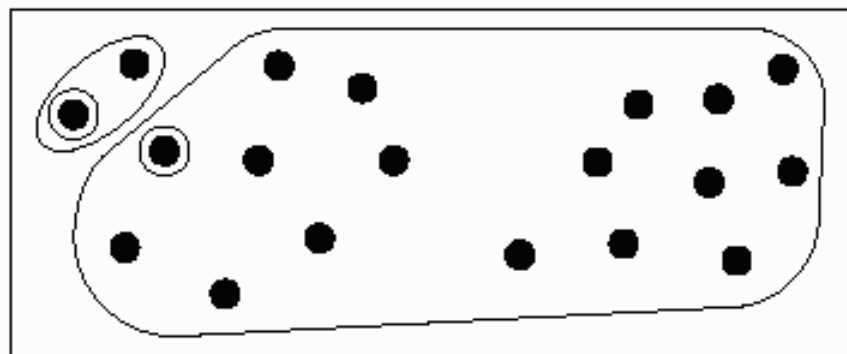


source: John Barnard, Barnard Chemical Information Ltd.

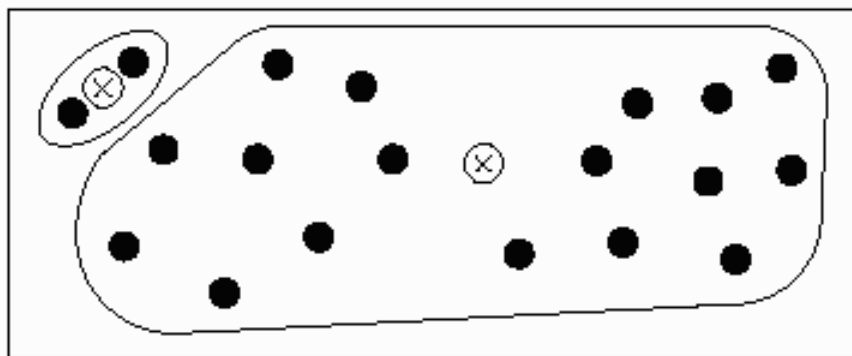
K-means with mobile centroid (I)



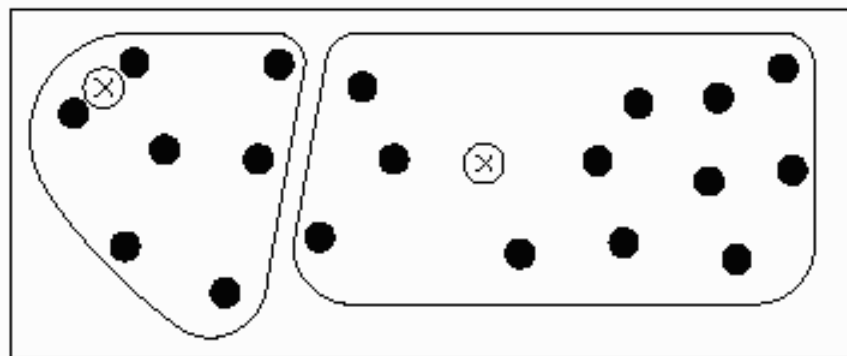
Step 1: K initial centroids are selected



Step 2: Clusters are constructed by affecting each molecule to the closest centroid



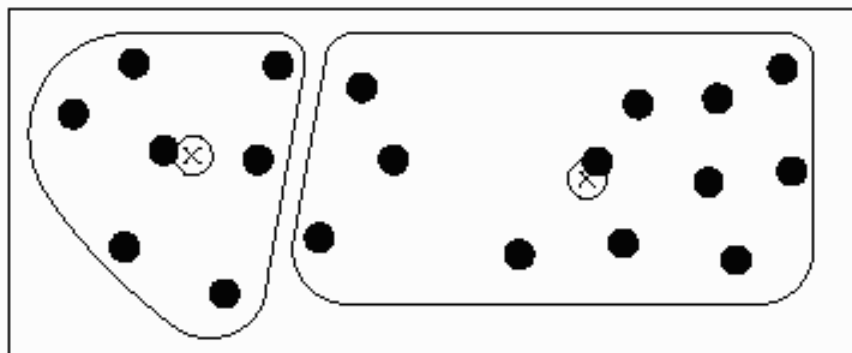
Step 3: Centres of gravity are calculated



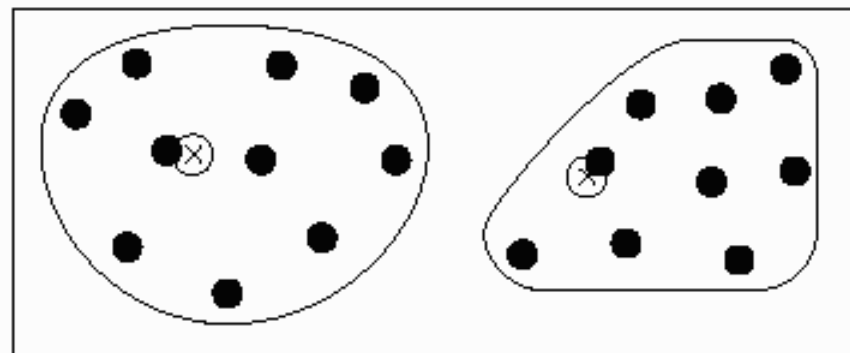
Step 4: Clusters are reconstructed

Lit: D.Gorse et al. *Drug Discovery Today* 4 (1999) 257.

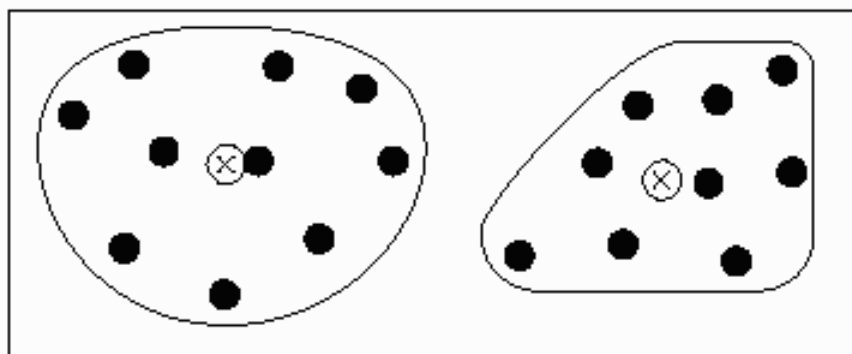
K-means with mobile centroid (II)



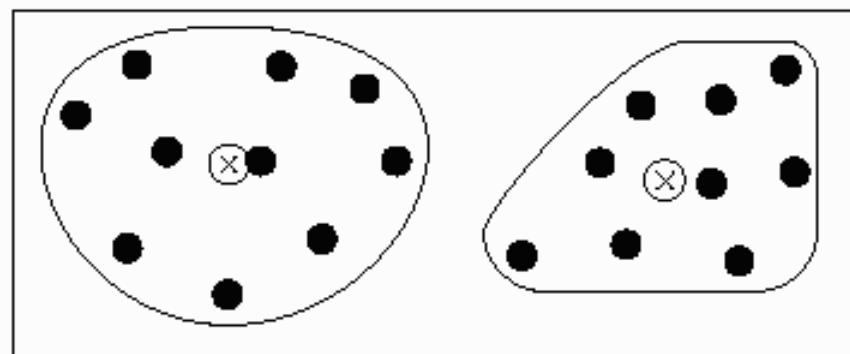
Step 3': Centres of gravity are calculated



Step 4': Clusters are reconstructed



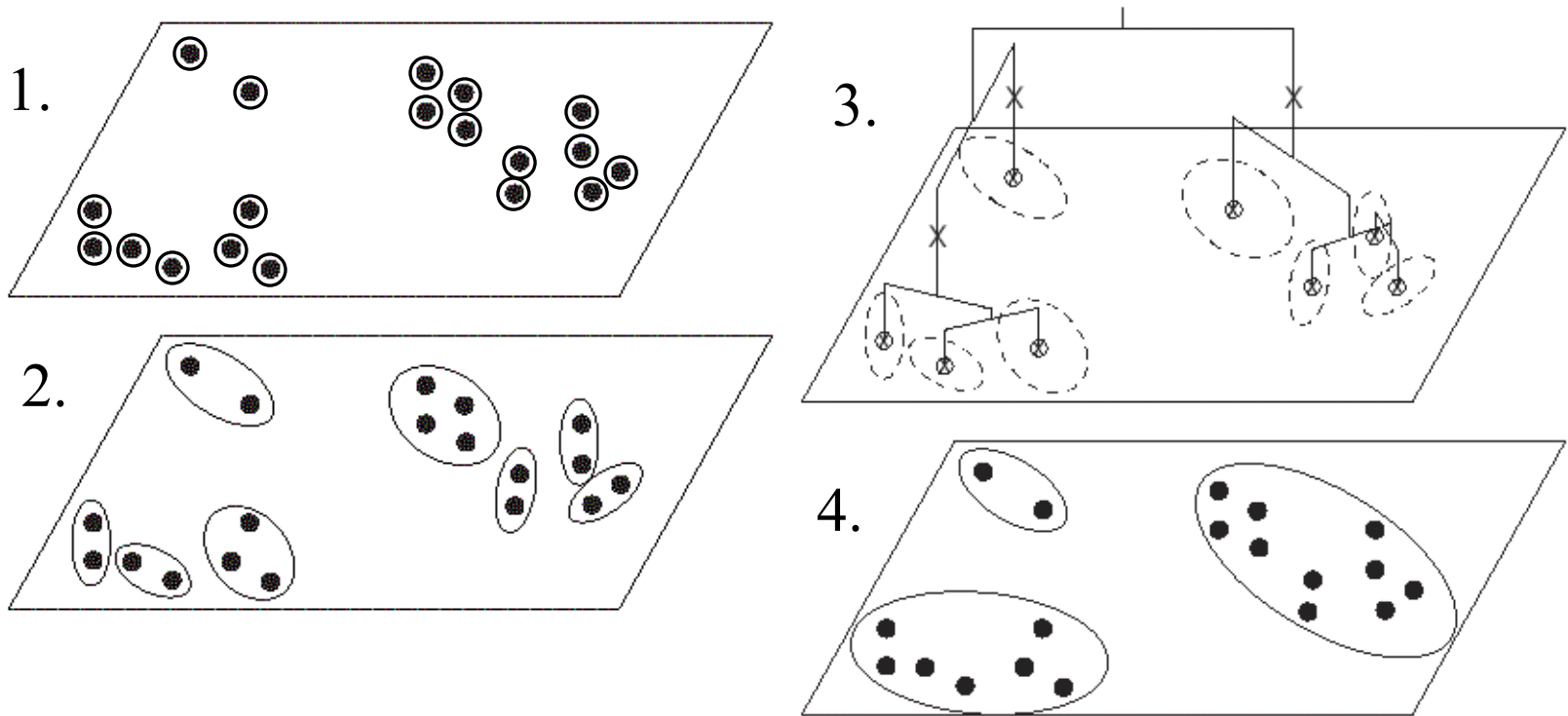
Step 3'': Centres of gravity are calculated



Step 4'': Clusters are reconstructed
Convergence is reached

Disadvantage: spherical clusters are often not adapted optimally regarding the distribution of the molecules in the chemical space

Mobile centres with Ward classification

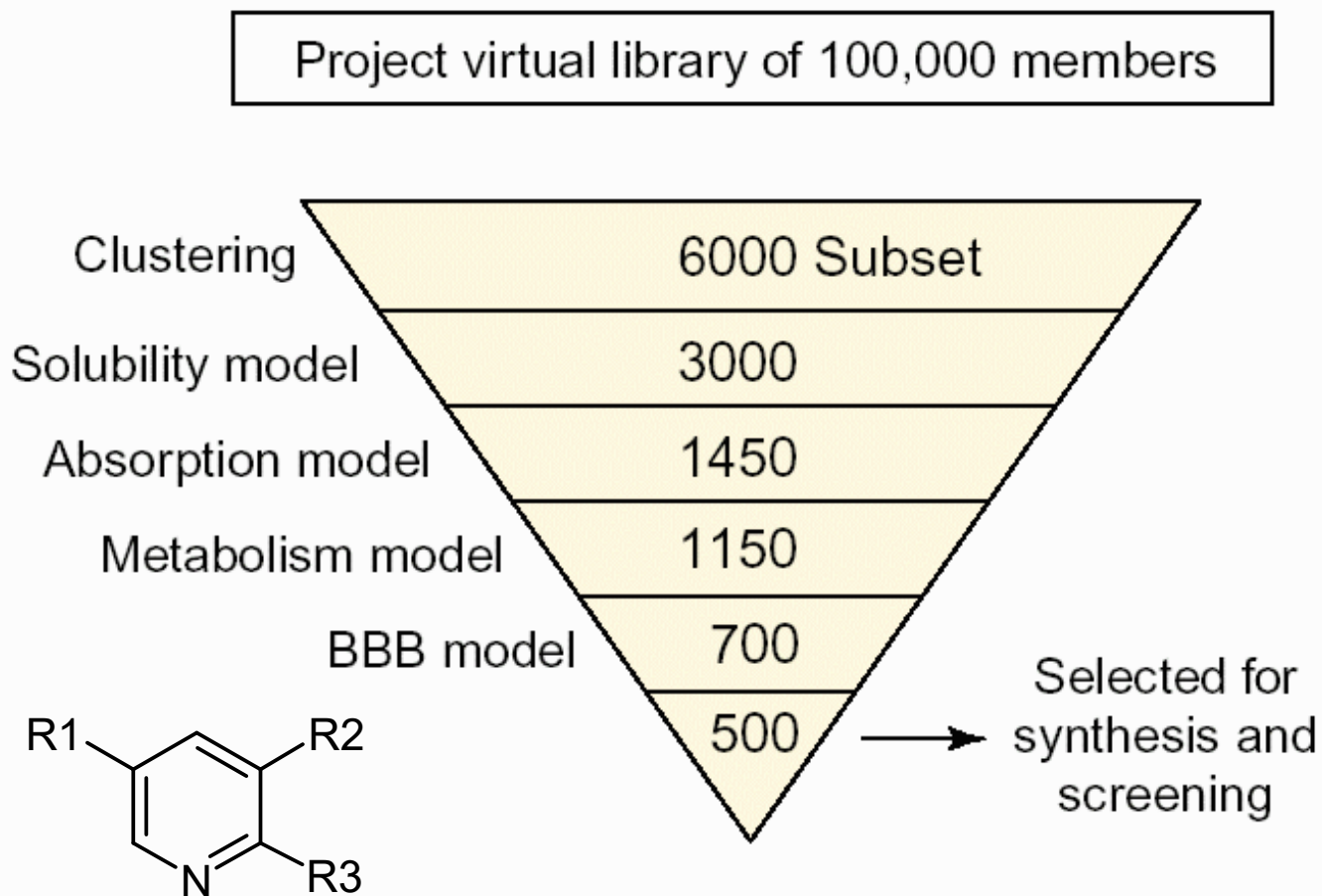


Most similar points of data are grouped to clusters step by step

Advantage: hierarchical, adapted shape of the clusters

Lit: D.Gorse et al. *Drug Discovery Today* 4 (1999) 257.

eADME filter proceeding High Throughput Screening (HTS)



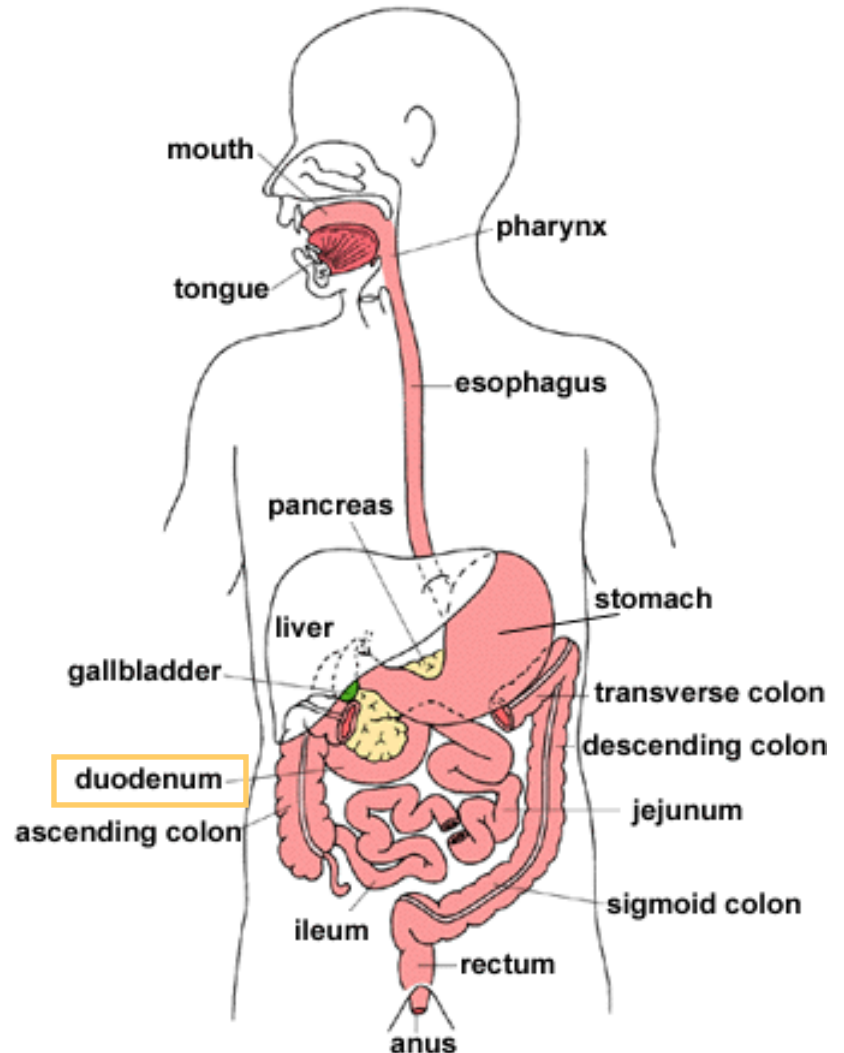
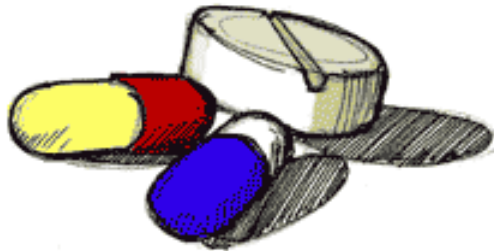
A typical eADME filter

Drug Discovery Today

Absorption

How does the drug reaches its destination ?

During the HTS the bioavailability is neglected first. To ensure the availability of the full dose in the assay, the substances are dissolved in a mixture of water and DMSO instead of pure water.



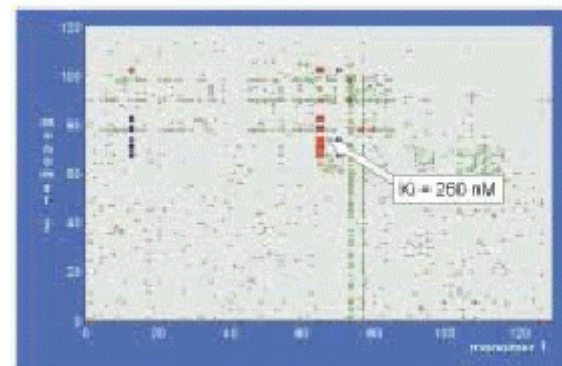
Evaluation of HTS results

Original idea: Automated test of >1000 compound on the target

Requires the synthesis of the according number of compounds, as well as processing of the results.

Sources of uncertainties are:

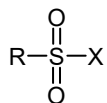
- purity and reliability of the compounds (false negatives)
- colored compounds (false positives)
- colloidal aggregation
- undesired covalent binding
- unspecifically binding compounds (false positives)
e.g. ibuprofen is a promiscuous binder



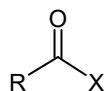
Pan Assay Interference Compounds (PAINS) → in silico filtering

Lit: Aldrich et al. *J.Chem.Inf.Model.* **57** (2017) 387 and references therein

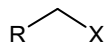
Substructures to be avoided



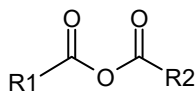
sulfonyl halides



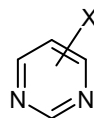
acyl halides



alkyl halides

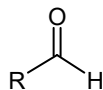


anhydrides

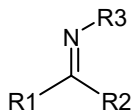


halopyrimidines

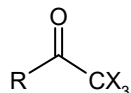
X at any of the carbon atom



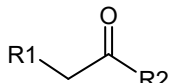
aldehydes



imines



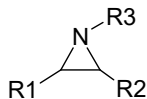
perhalo ketones



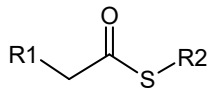
aliphatic ketones



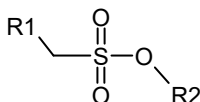
epoxides



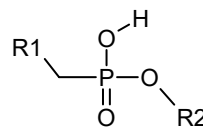
aziridines



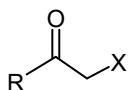
thioesters



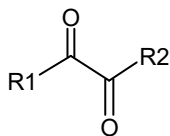
sulfonate esters



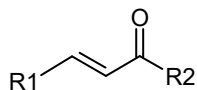
phosphonate esters



α -halocarbonyls



1,2-dicarbonyls

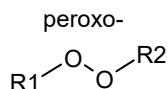


Michael acceptors

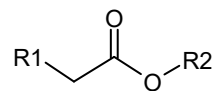
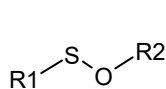
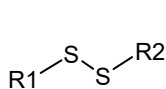
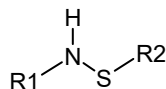
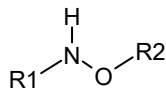
R = carbon

X = F or Cl or Br

labile single bonds between hetero atoms (N, O, S)

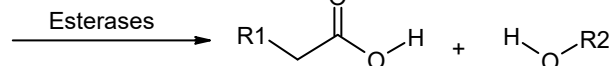


peroxo-



aliphatic esters

present in many prodrugs !



Esterases

source: Hugo Kubinyi,
www.kubinyi.de

Setup of substance libraries for high throughput screening (V)

3. step: if yes, generate a virtual substance library based on the lead compound(s)

systematic variation of the lead compound:

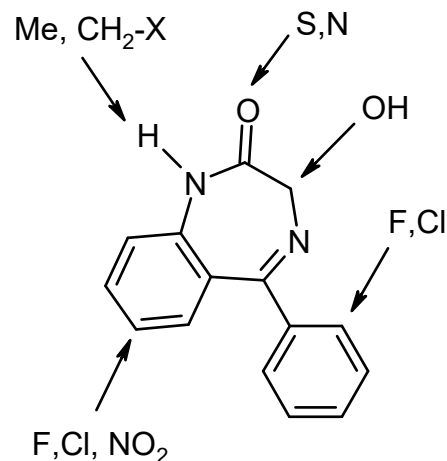
framework

side chains / substituents

bioisosters



Similar biological properties



Publically Available Compound Databases

PubChem > 112,000,000 compounds NCBI
ChEMBL > 2,200,000 compounds EMBL
DrugBank > 500,000 drugs University of Alberta
ZINC15 > 750,000,000 compounds UCSF
(this list is not comprehensive!)

database	actual drugs	drug-like	lead-like	chemicals
PubChem	++	++	+	++
ChEMBL	++	+	+	-
DrugBank	++	+	-	-
ZINC	+	++	++	++

Often compounds are hyper-linked to further information, such as targets and assays.



Setup of substance libraries for high throughput screening (VI)

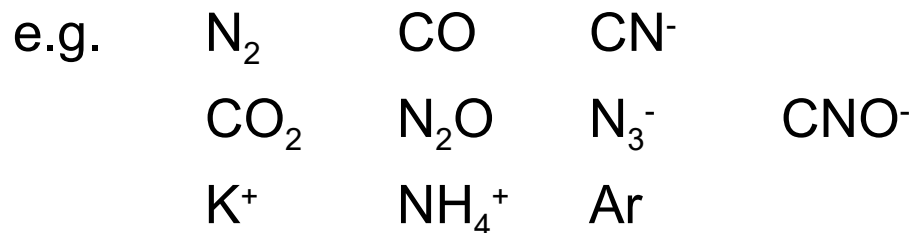
During the optimization from the lead compound to the clinical drug, substances are usually getting larger and more lipophilic (extensive filling of the binding pocket with mostly hydrophobic parts).

Therefore these properties of lead compounds are desirable:

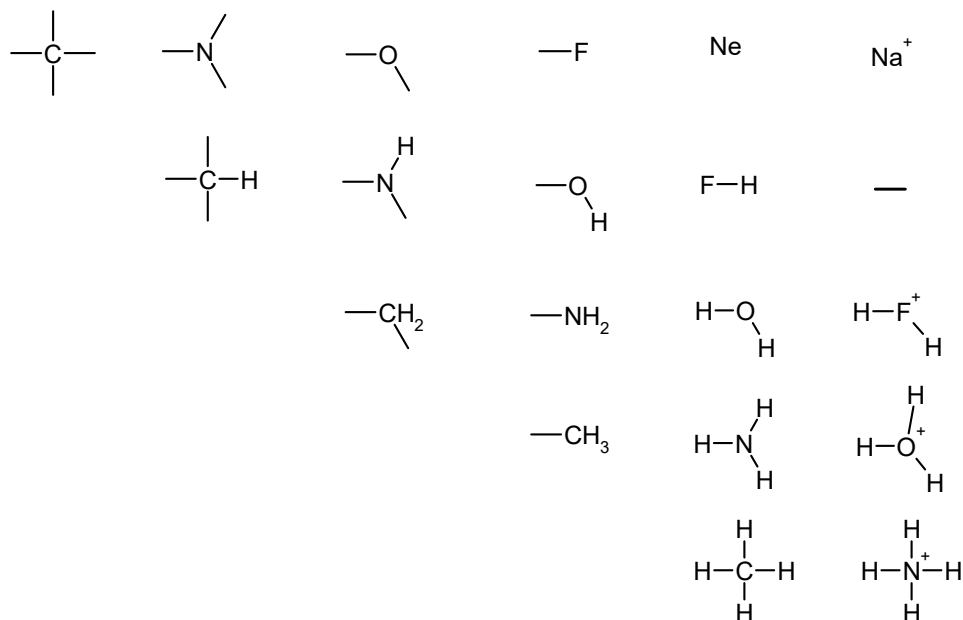
- molecular weight < 250
- low lipophilicity ($\log P < 3$) if orally administered
- enough possibilities for side chains
- sufficient affinity and selectivity

Bioisosters (I)

definition: Same number and arrangement of electrons
(Langmuir 1919)



Grimm's hydride exchange law (1925)

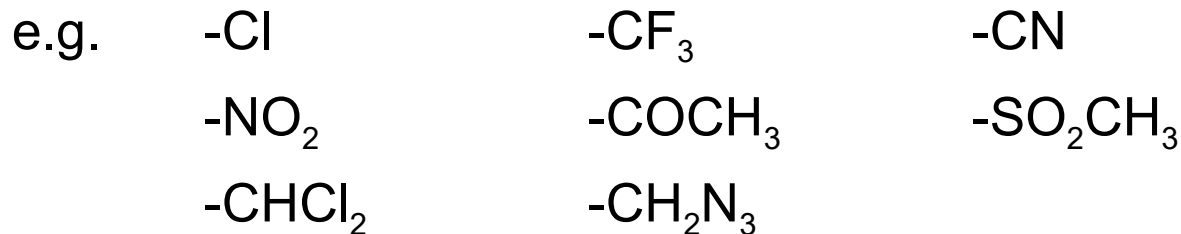


Bioisosters (II)

Definition:

Compounds or groups that possess near-equal, molecular shapes and volumes, approximately the same distribution of electrons, and which exhibit similar physical properties.

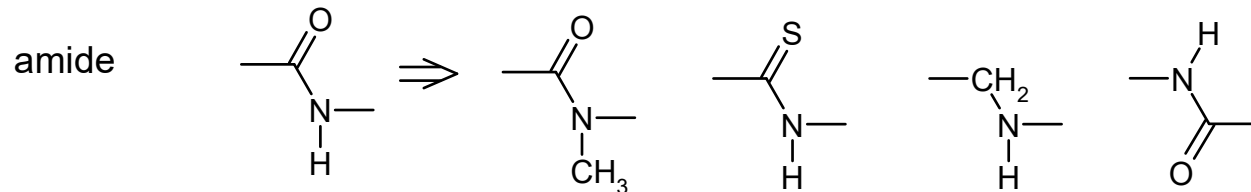
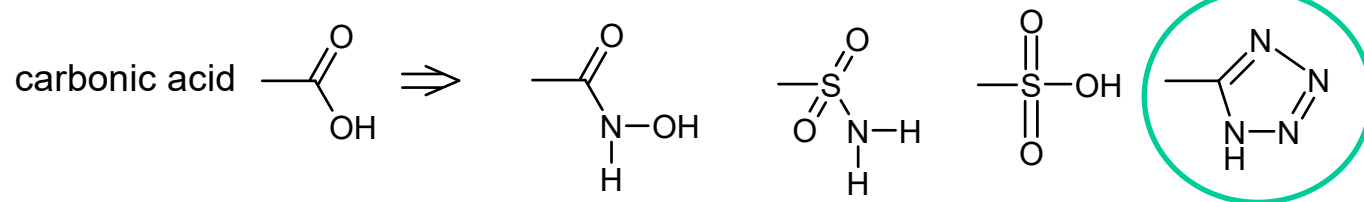
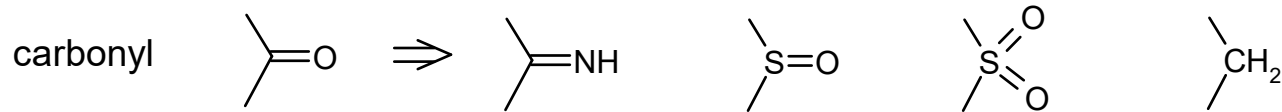
(A. Burger 1970)



Review article: G.A. Patani, E.J. LaVoie, *Chem.Rev.* **96** (1996) 3147.

Bioisosters (III)

classical (bio-)isosters are sterically and electronically similar



Non-classical isosters:

e.g. exchange of cyclic against linear structures
exchangeable groups (no apparent similarity)

Bioisosters (IV)

In the rarest cases bioisosters (similar *chemical space*) will show the same activity profile (similar *biological space*) than the compound they have been derived from.

Aimed are following properties:

better mode of action

improved selectivity

increased bioavailability

less toxic

fewer adverse side effects

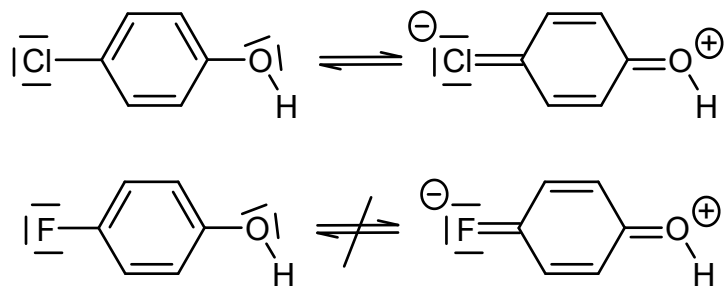
} allows lower dosage

Monovalent Bioisosters (I)

Exchange of (non-polar) H for F

Fluorine has a similar van der Waals radius compared to hydrogen and is thus about the same size. The lipophilic character is retained (fluorocarbons are even less soluble than hydrocarbons).

Fluorine is the most electronegative element, thus it produces an inductive effect (electron pulling) onto the neighboring C atom. In contrast to the other halogens, however, no mesomeric structures are possible. (attributed to the lack of *d*-orbitals)

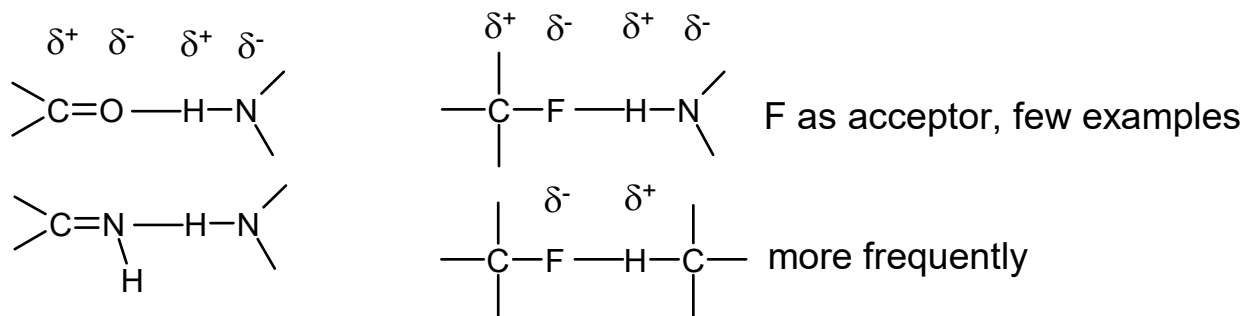


Monovalent Bioisosters (II)

Exchange of –H for –F

The C–F bond is stronger than the corresponding C–H, C–Cl, C–Br, and C–I bonds and therefore also more inert against metabolic reactions.

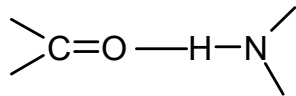
In principle, fluorine should also be a suitable H-bond acceptor like nitrogen or oxygen. However, in X-ray structures this is rarely seen.



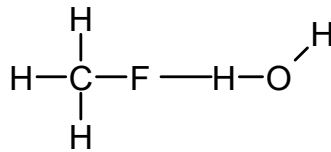
Lit: H.J. Böhm et al., *ChemBioChem* **5** (2004) 637.

Fluorine in Hydrogen Bonds

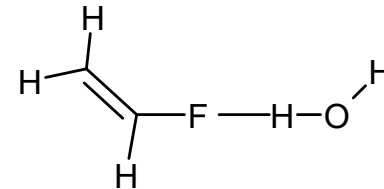
Electronegativity goes along with the tendency to accept electrons, not protons. Covalently bound fluorine is, however, a weak base and an extremely weak proton acceptor. Corresponding H-bonds are very weak.



ca. 5 kcal mol⁻¹



2.4 kcal mol⁻¹



1.5 kcal mol⁻¹

Thus, fluorine is mainly used to block metabolically labile sites in drugs, or to increase lipophilicity without increasing the size at that spot.

Lit: J.A.K. Howard et al. *Tetrahedron* **52** (1996) 12613.

J.D. Dunitz, R. Taylor, *Chem.Eur.J.* **3** (1997) 89.

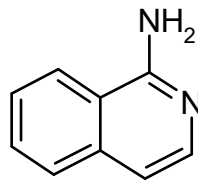
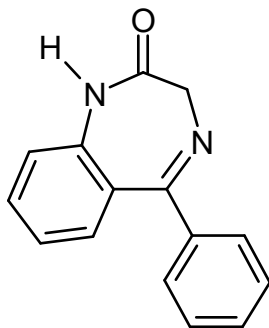
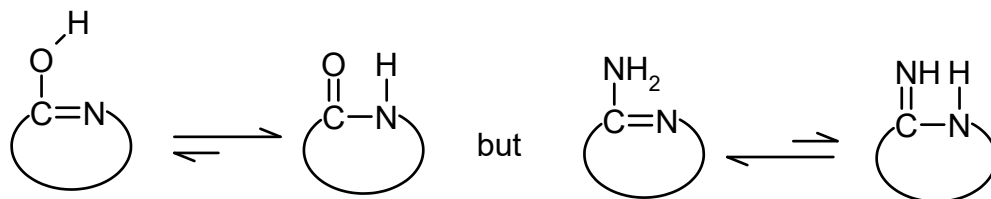
Monovalent Bioisosters (III)

Exchange of $-OH$ for $-NH_2$

Both groups possess similar size and shape

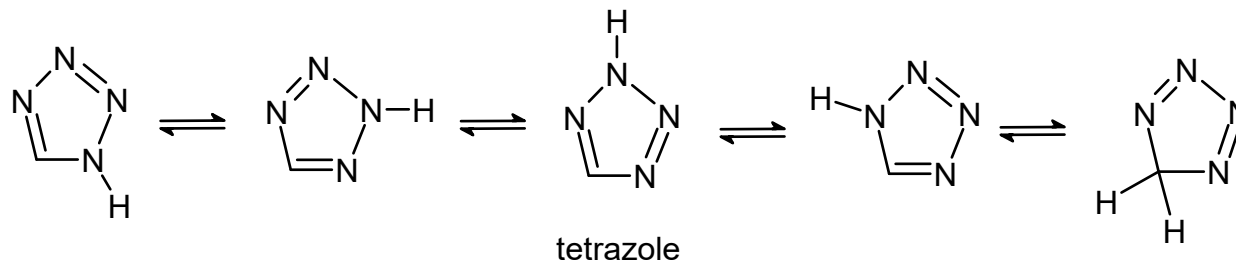
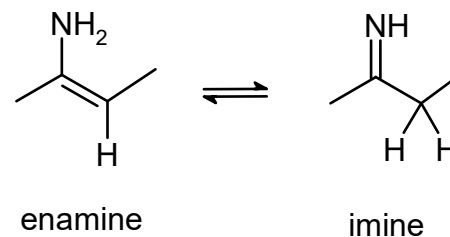
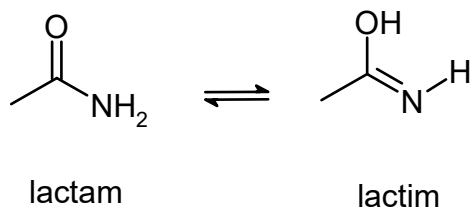
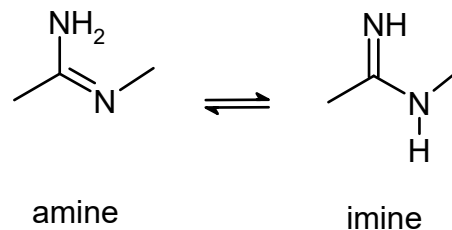
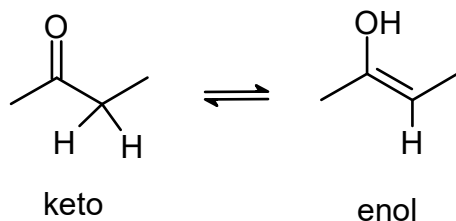
Both are H-bond donors as well as H-bonds acceptors

In heterocyclic rings the equilibrium tautomer is shifted:



Tautomers

Isomers that are interconvertible by the (formal) shift of a hydrogen (atom or proton) along the switch of a single bond and an adjacent double bond. In solution the equilibrium distribution of the possible tautomeric forms is dependent on pH, solvent, ions, ...



Monovalent Bioisosters (IV)

Exchange of –SH for –OH

Sulfur is much larger than oxygen

$$R_{\text{vdw}}(\text{O}) = 1.4 \text{ \AA} \qquad R_{\text{vdw}}(\text{S}) = 1.85 \text{ \AA}$$

and of lower electronegativity

$$\text{O: } 3.5 \qquad \text{S: } 2.4 - 2.6$$

Thus hydrogen bonds to SH are weaker.

Anyhow, thiols are more acidic and stronger dissociated than the corresponding alcohols.

$$\text{Cys-SH} \qquad pK_a \quad 8.3$$

$$\text{Ser-OH} \qquad pK_a \approx 13$$

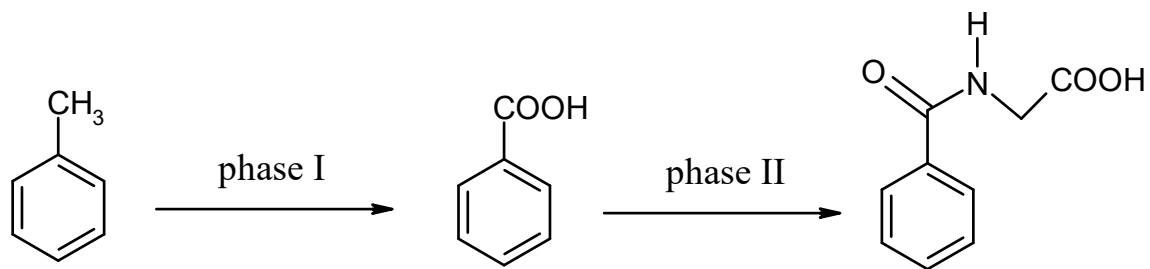
In heterocyclic rings the corresponding thiol can be formed by tautomerization similar to –NH₂

Monovalent Bioisosters (V)

Exchange of $-Cl$ for $-CH_3$

Chlorine and the methyl group possess the same size and lipophilicity.

In contrast to the $C-Cl$ bond the corresponding $C-CH_3$ bond is metabolized and excreted more rapidly.



Monovalent Bioisosters (VI)

Exchange of $-\text{CF}_3$ or $-\text{CN}$ for $-\text{Br}$

The trifluoromethyl and the cyano (=nitrile) group have the same electronic properties, but the $-\text{CN}$ group is much more hydrophilic. Bromine is similar in size and somewhat more lipophilic than the nitrile group.

Rule of thumb concerning **bioavailability**:

Lipophilic compounds are absorbed worse and are increasingly metabolized in the liver.

Usually hydrophilic compounds are easily absorbed but likewise being excreted by the renal pathway more rapidly.

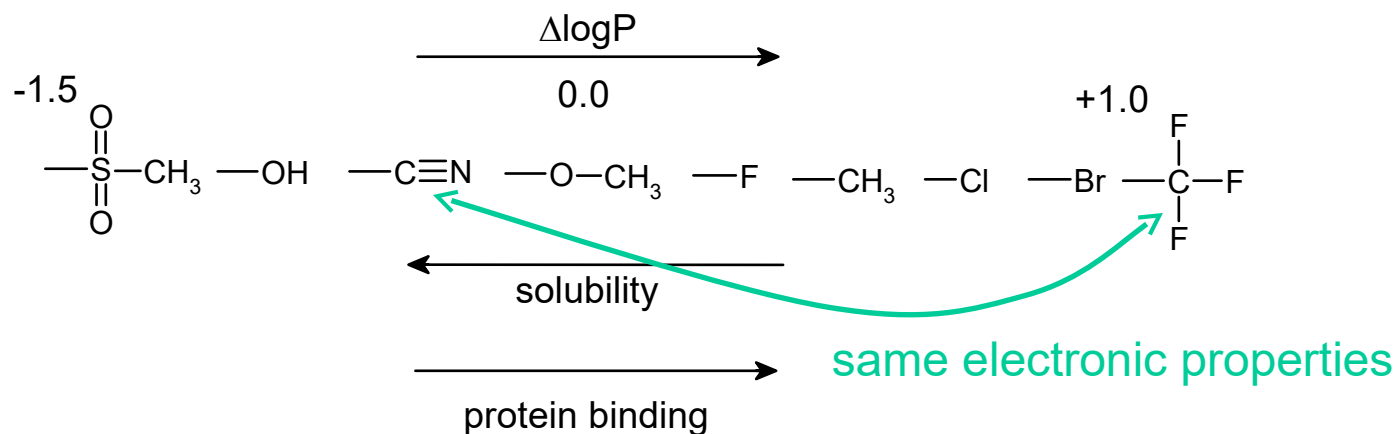
measure: $\log P = n\text{-octanol} / \text{water partition coefficient}$

LogP and Solubility

Rule of thumb concerning **solubility**:

Lipophilic compounds are less soluble than hydrophilic ones

measure: $\log P = n\text{-octanol} / \text{water}$ partition coefficient

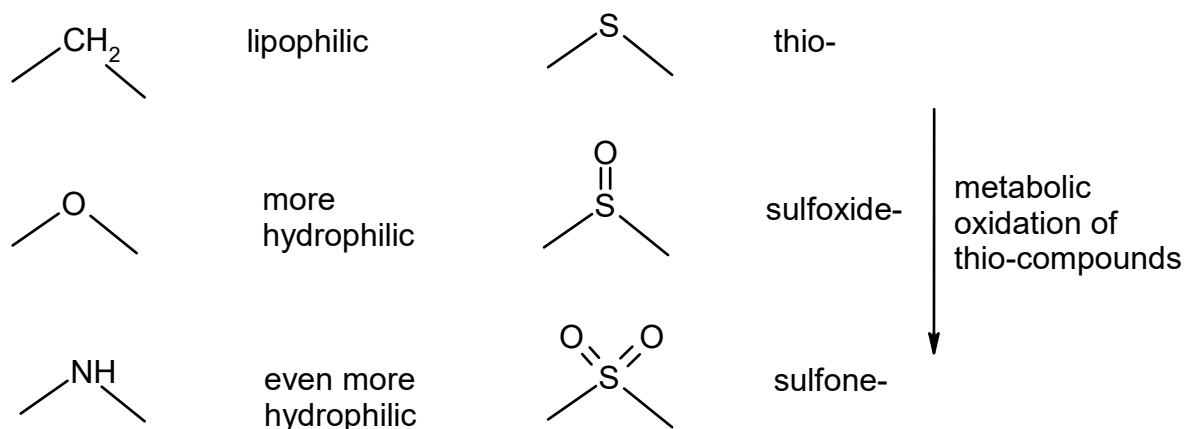


Fragmental contribution of substituents

Lit: A.G. Leach et al. *J.Med.Chem.* **49** (2006) 6672.

Divalent Bioisosters

Exchange of the $-\text{CH}_2-$ (methylene) group



Compounds containing B-H or Si-H bonds are usually to sensitive against hydrolysis.

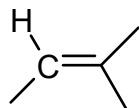
However, here are some examples of actual drugs

Boron: bortezomib, bosentan, dutogliptin, flovagatran

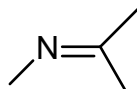
Silicon: flusilazol

Trivalent Bioisosters

Exchange of the $-\text{CH}=\text{}$ group for $-\text{N}=\text{}$ or $-\text{NH}-$

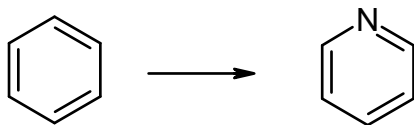


lipophilic

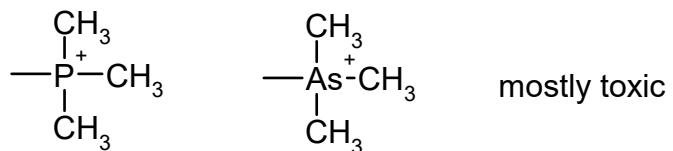
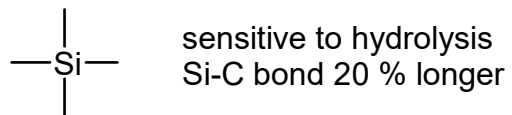
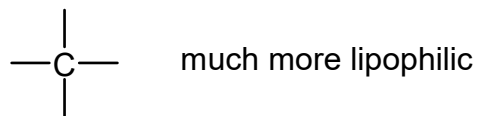
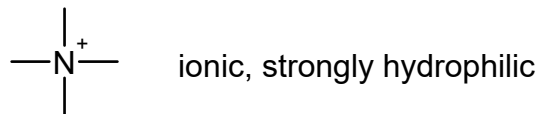


more hydrophilic,
H-bond acceptor

Important and successful especially in heterocyclic ring systems

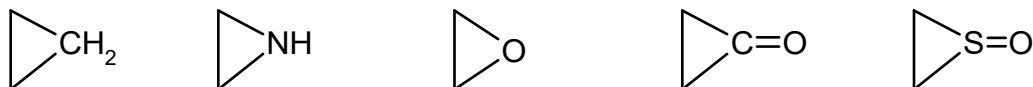


Tetravalent Bioisosters

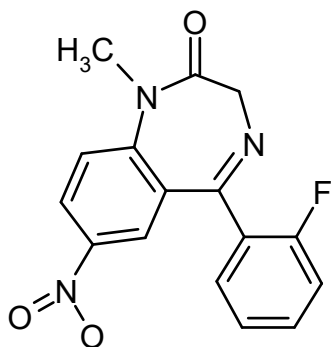


Divalent ring equivalents

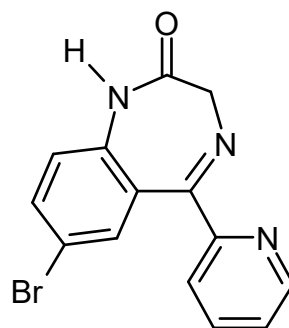
Exchange of the $-\text{CH}_2-$ group



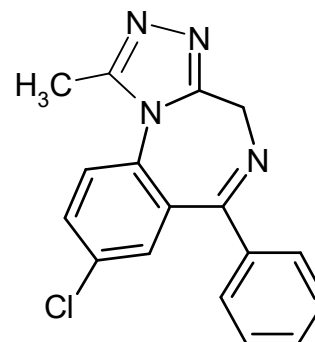
Also possible in larger ring systems (7-membered rings etc, see benzodiazepines):



flunitrazepam



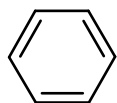
bromazepam



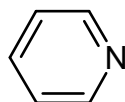
alprazolam

Trivalent ring equivalents

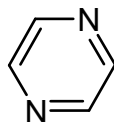
Exchange of the $-\text{CH}=\text{}$ group



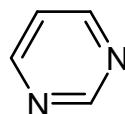
benzene



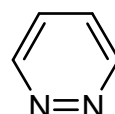
pyridine



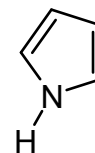
pyrazine



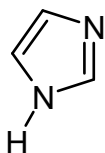
pyrimidine



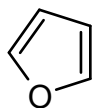
pyridazine



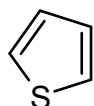
pyrrole



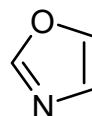
imidazole



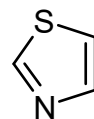
furan



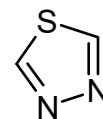
thiophen



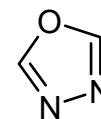
oxazole



thiazole



thiadiazole



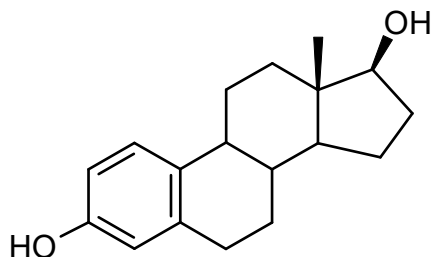
oxadiazole

Enables frequently the fine tuning of the functional and ADME profile

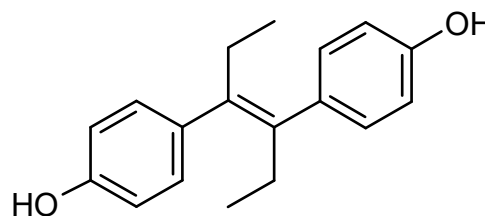
c.f. sildenafil versus vardenafil

Non-classical Isosters (II)

ring opening

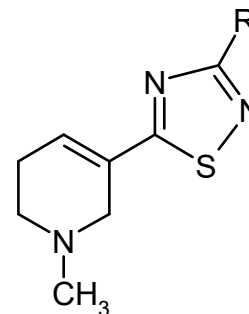
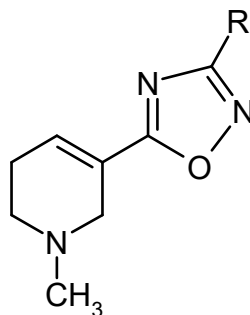
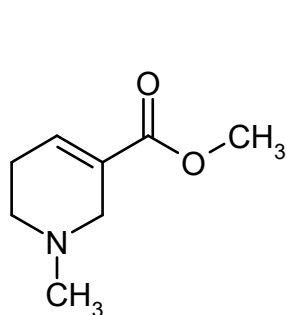


Estradiol



Diethylstilbestrol

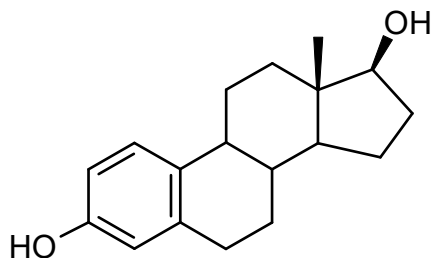
ring closure



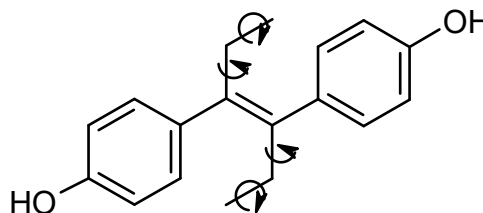
Frequently used to „freeze“ an active conformation

Thermodynamic effects

Ring opening: Generates more degrees of freedom, thus loss of entropy upon binding to the enzyme

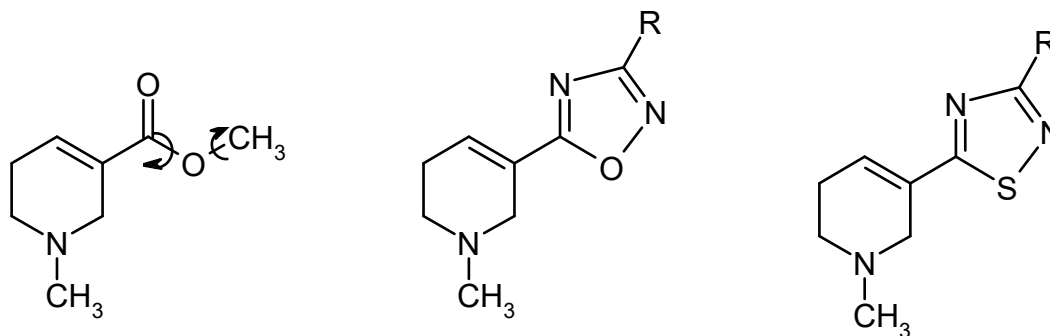


Estradiol



Diethylstilbestrol

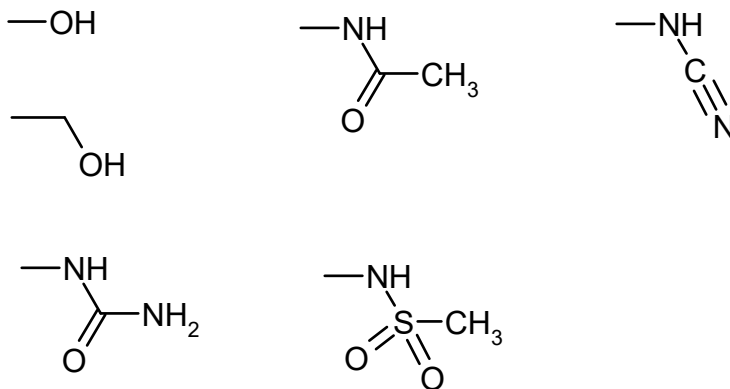
ring closure: Reduced loss of entropy upon binding



Bioisosteric exchange of functional groups

hydroxyl group –OH

Here: Conservation of H-bond properties has priority

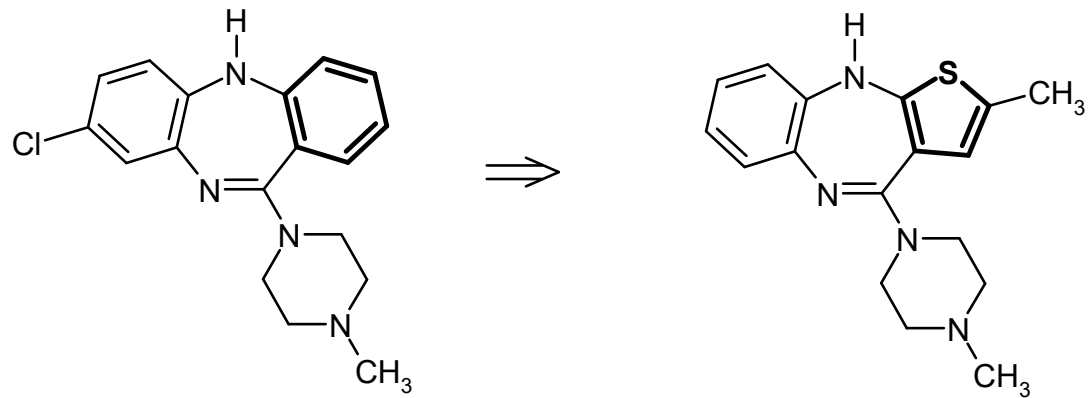


H-bond donors are more conserved than purely H-bond acceptors.

Lit: T.Fehlmann & M.C.Hutter *J.Chem.Inf.Model.* **59** (2019) 1314.

Examples of Bioisosters (I)

Exchange benzene-thiophene



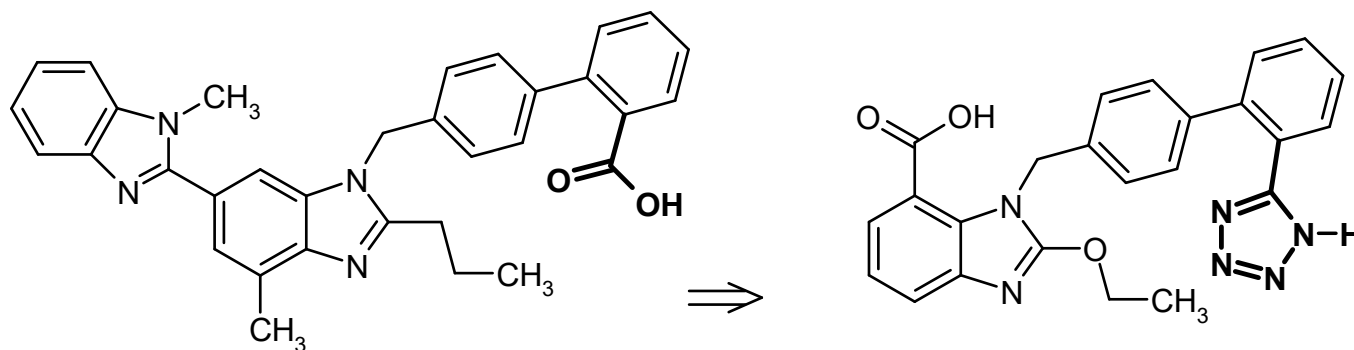
Clozapin

Olanzapin

Avoids exoxidation of the benzene ring, thus reduced hepatotoxicity

Examples of Bioisosters (II)

Exchange carboxylate-tetrazole



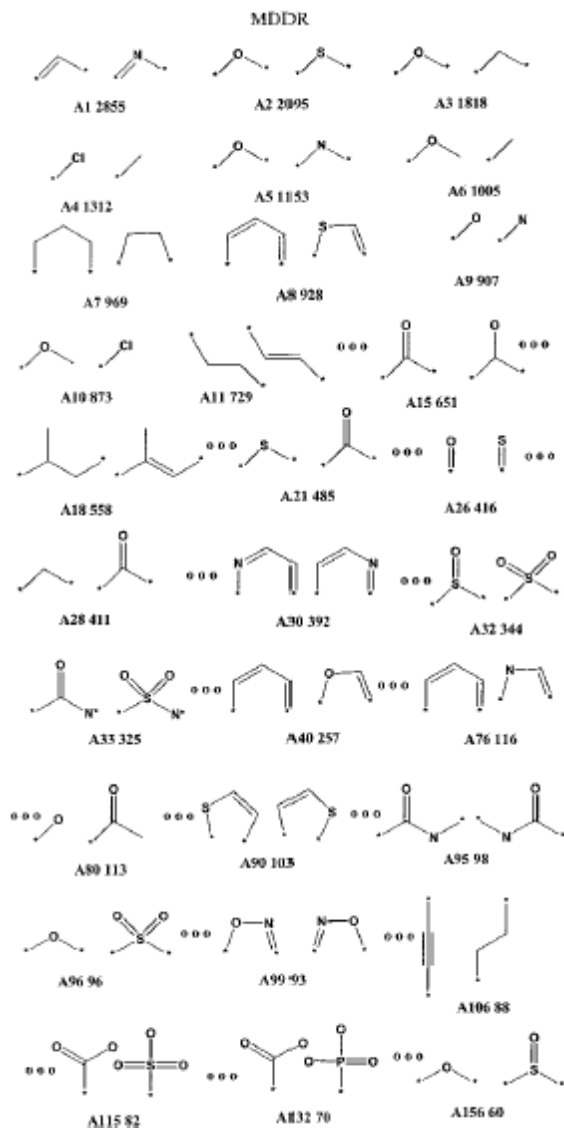
Telmisartan

Candesartan

Comparable acidity along improved solubility

Lit. C.D. Siebert *Chemie in unserer Zeit* **38** (2004) 320.

Distribution of Chemical Replacements (I)



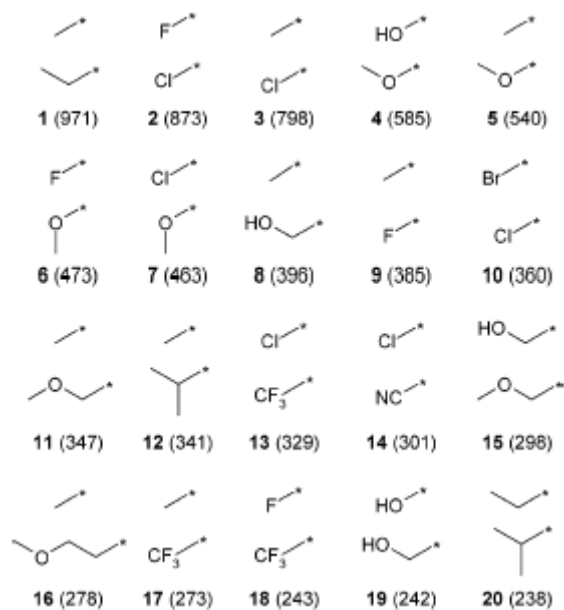
Analysis of the MDL Drug Data Report (>100,000 drugs)

The most common replacements of fragments (starting from top, left)

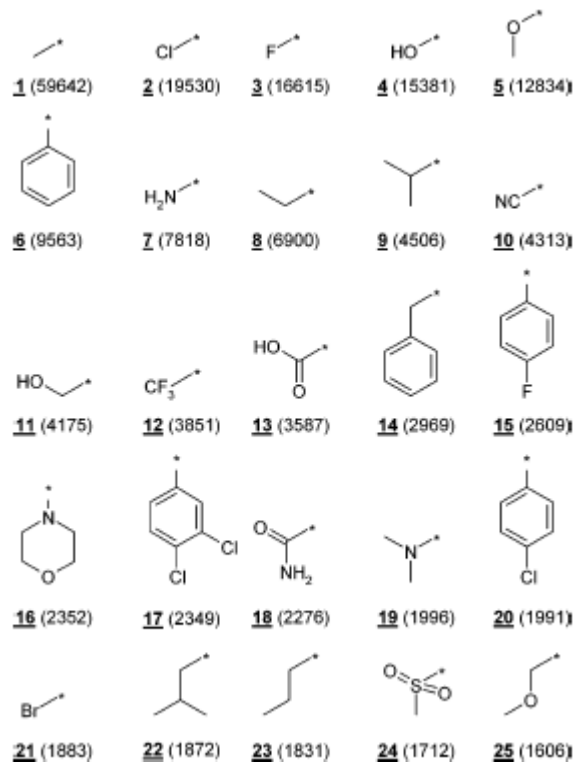
Lit. R.P. Sheridan
J.Chem.Inf.Comput.Sci. **42** (2002) 103.

Distribution of Chemical Replacements (II)

In house database (50,000 drug-like compounds)



Most common replacements
rank (count)

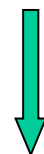
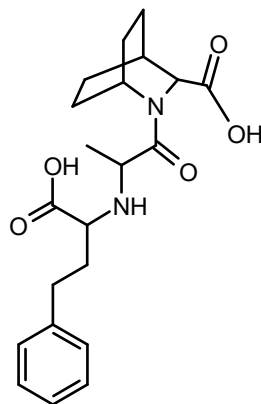
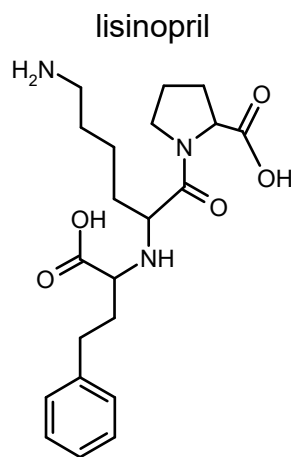


Most common sidechains
rank (count)

Lit. D.Y. Haubertin, P. Bruneau *J.Chem.Inf.Model.* **47** (2007) 1294.

Statistical Evaluation of Bioisosteric Exchanges in Drugs

Align similar drugs of the same target (e.g. ACE-Inhibitors)



Statistical frequencies of chemical exchanges

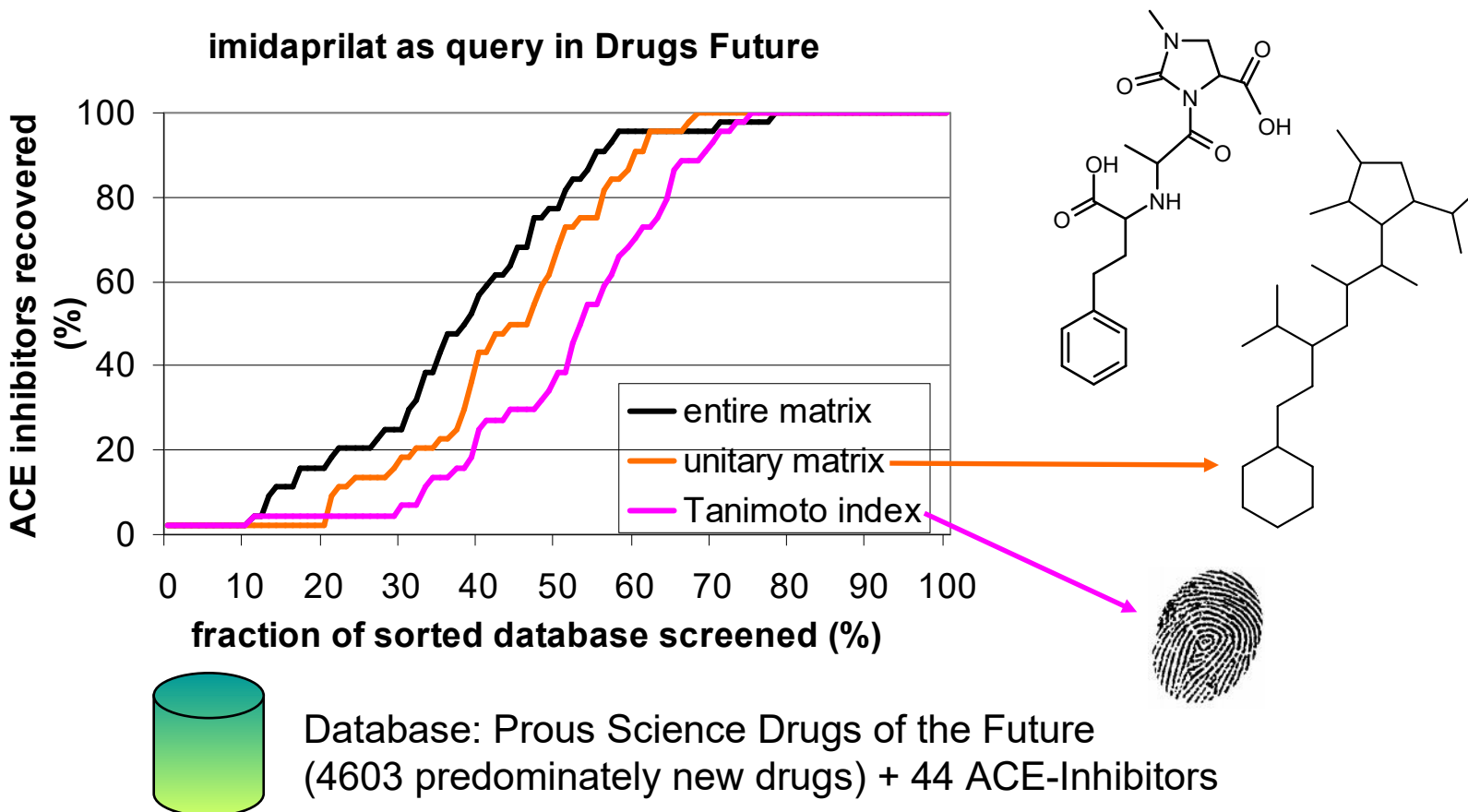
Bioisosteric exchange matrix (similar to amino acid exchange matrices such as PAM250 or BLOSUM62)



Predict similarity of new compounds (in virtual screening)

Lit. M.Krier, M.C.Hutter *J.Chem.Inf.Model.* **49** (2009) 1280.

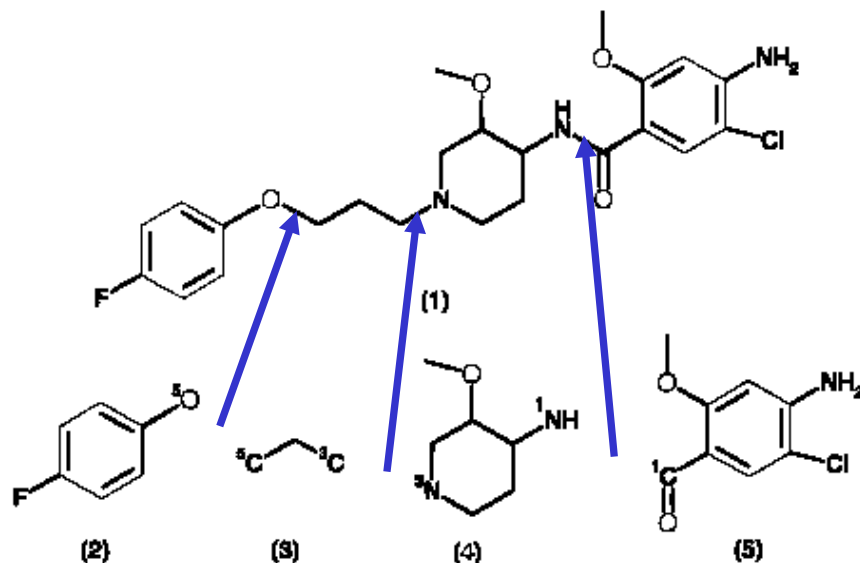
Bioisosteric Similarity vs Substructure matching and fingerprints



Lit. M.Krier, M.C.Hutter *J.Chem.Inf.Model.* **49** (2009) 1280.

Systematic Variation – *in silico* approaches (I)

Analog to the approach used in the feature trees, each molecule is splitted into *nodes* and *linkers*. Each node corresponds to a chemical group and each linker to a bond between such groups.



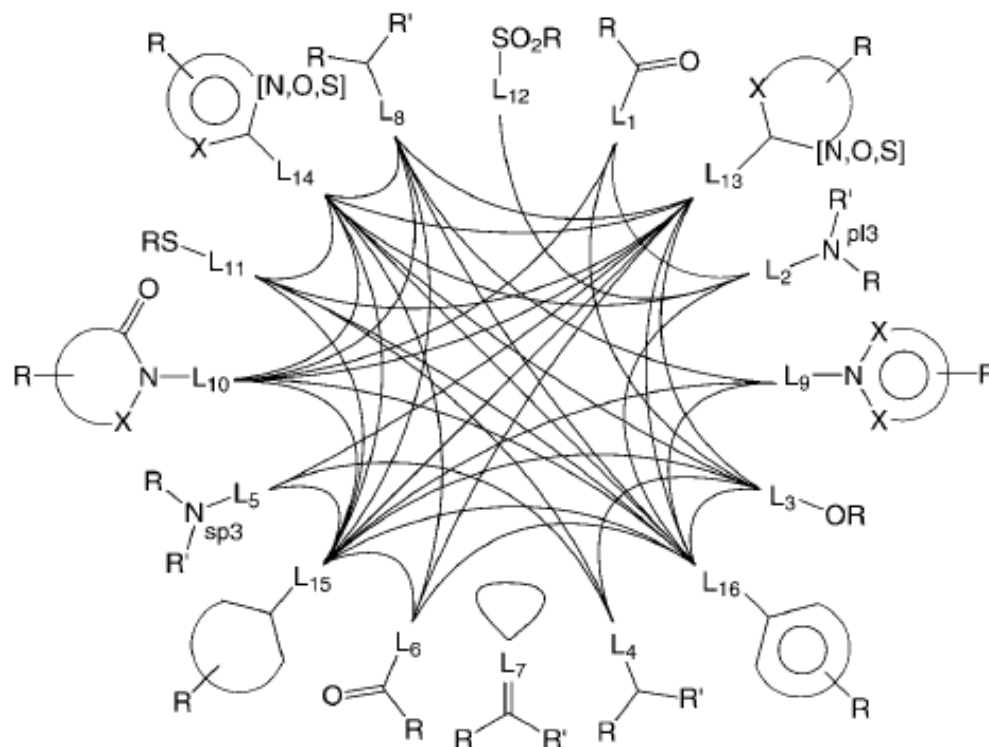
By using defined types of bond cleavages (retro synthesis), matching fragments can be searched in data bases and combined differently.

RECAP concept:

Lit. X.Q.Lewell et al. *J.Chem.Inf.Comput.Sci.* **38** (1998) 511.

Systematic Variation – *in silico* approaches (II)

A more specific set of rules for bond cleavages and reformation of bonds is realized by the **BRICS** concept. Here, information for the synthesis of actual combinatorial libraries was compiled.



Lit. J.Degen et al. *ChemMedChem* **3** (2008) 1503.