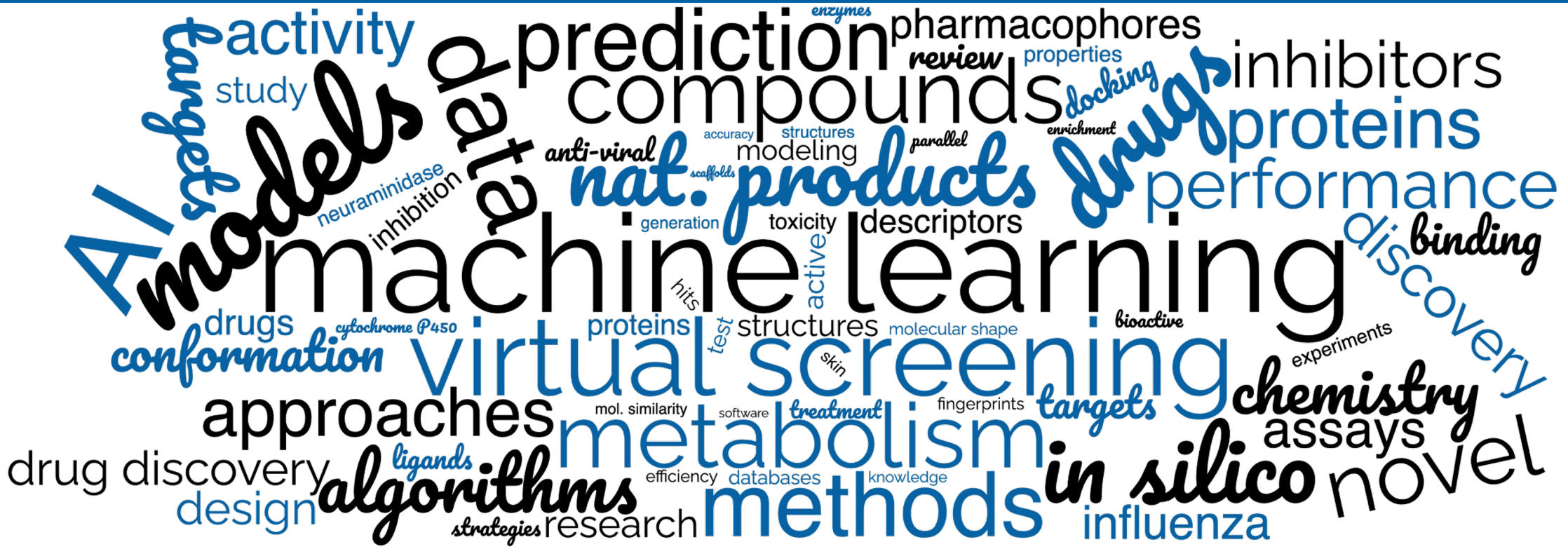
universität wien

Christian Doppler Laboratory for Molecular Informatics in the Biosciences

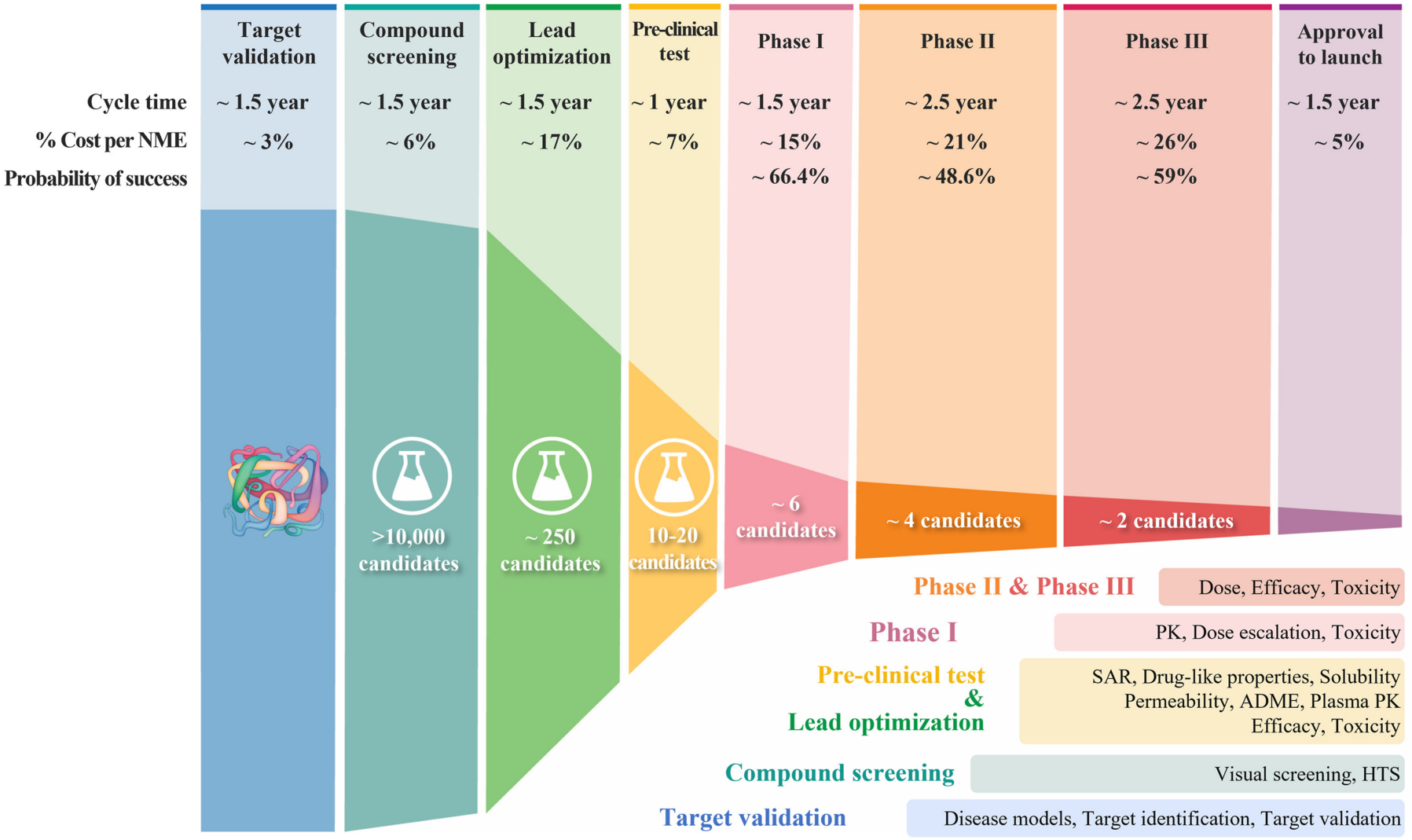
An introduction to Cheminformatics

Johannes Kirchmair

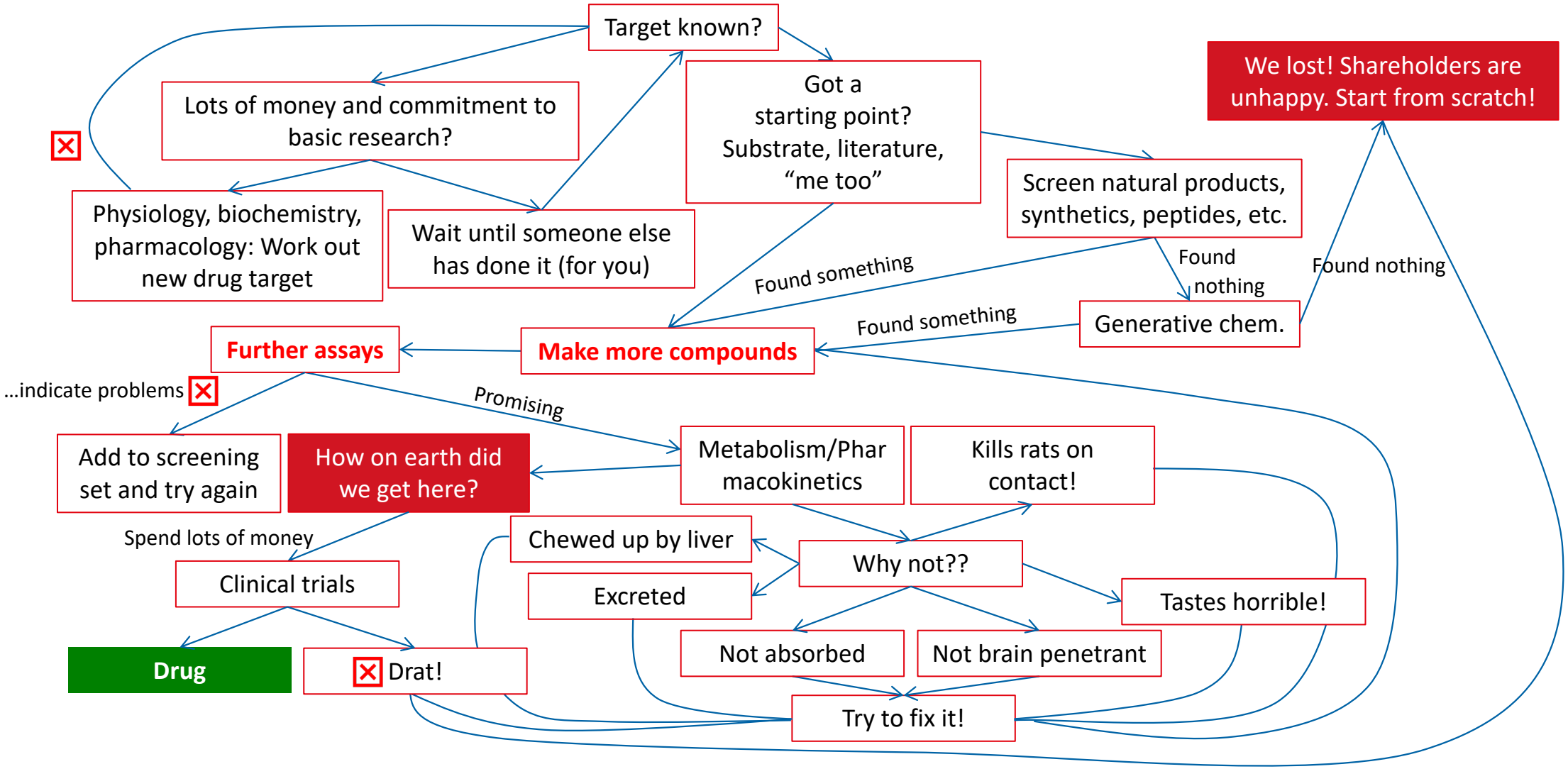


Word-Cloud based on the Abstract texts of our group's publications

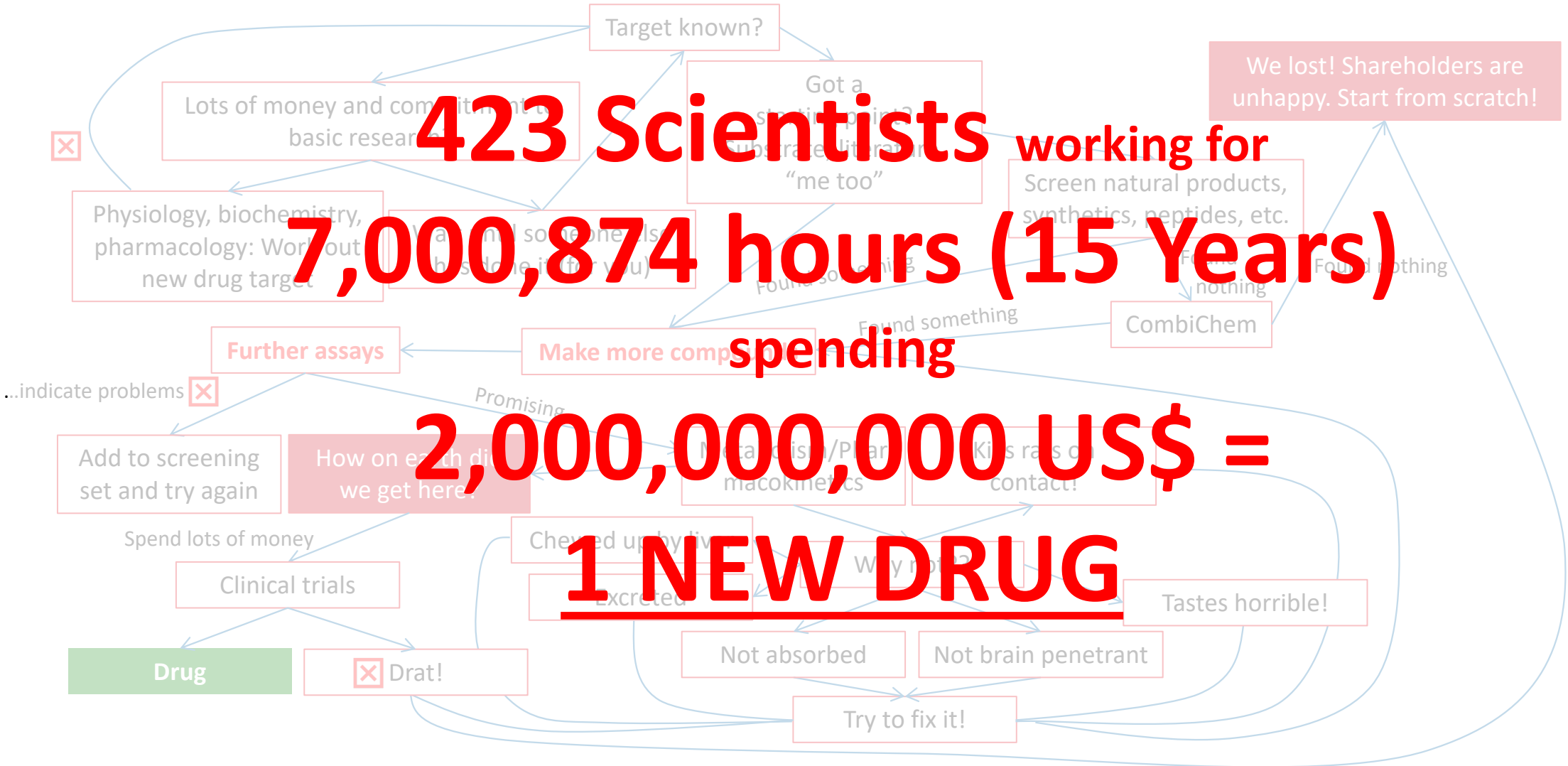
Small-molecule drug discovery and development



The long way to a new drug…



The long way to a new drug…



423 Scientists working for

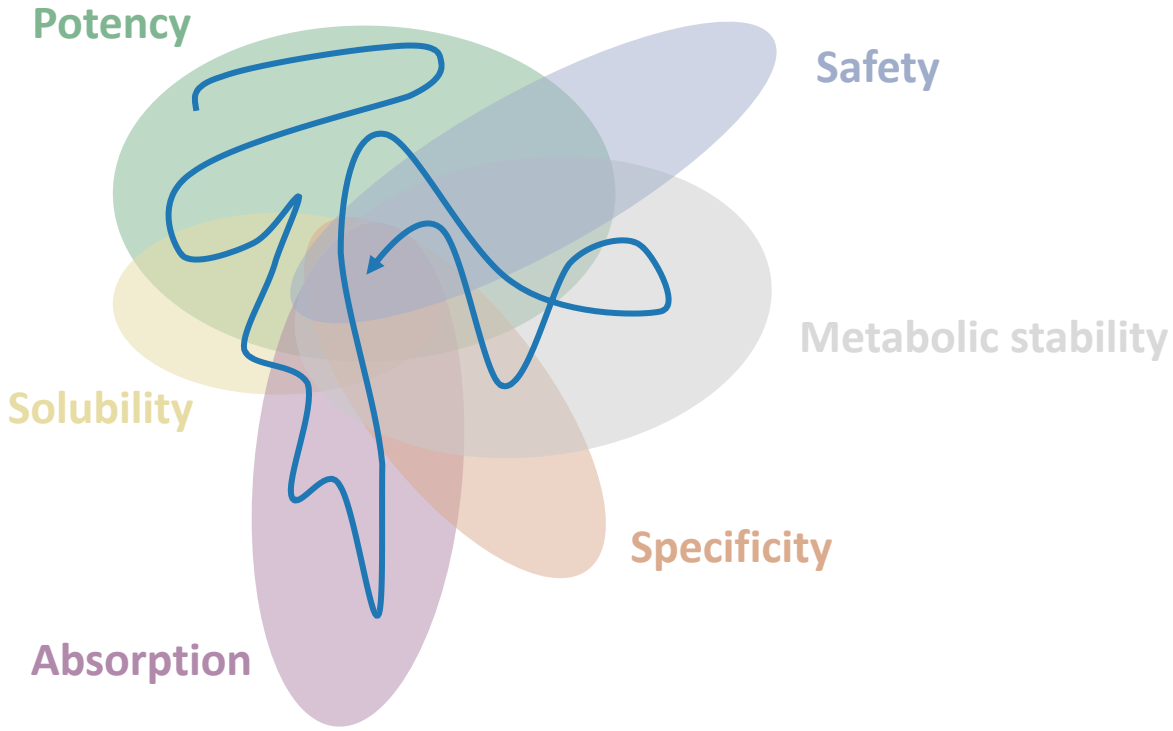
7,000,874 hours (15 Years)

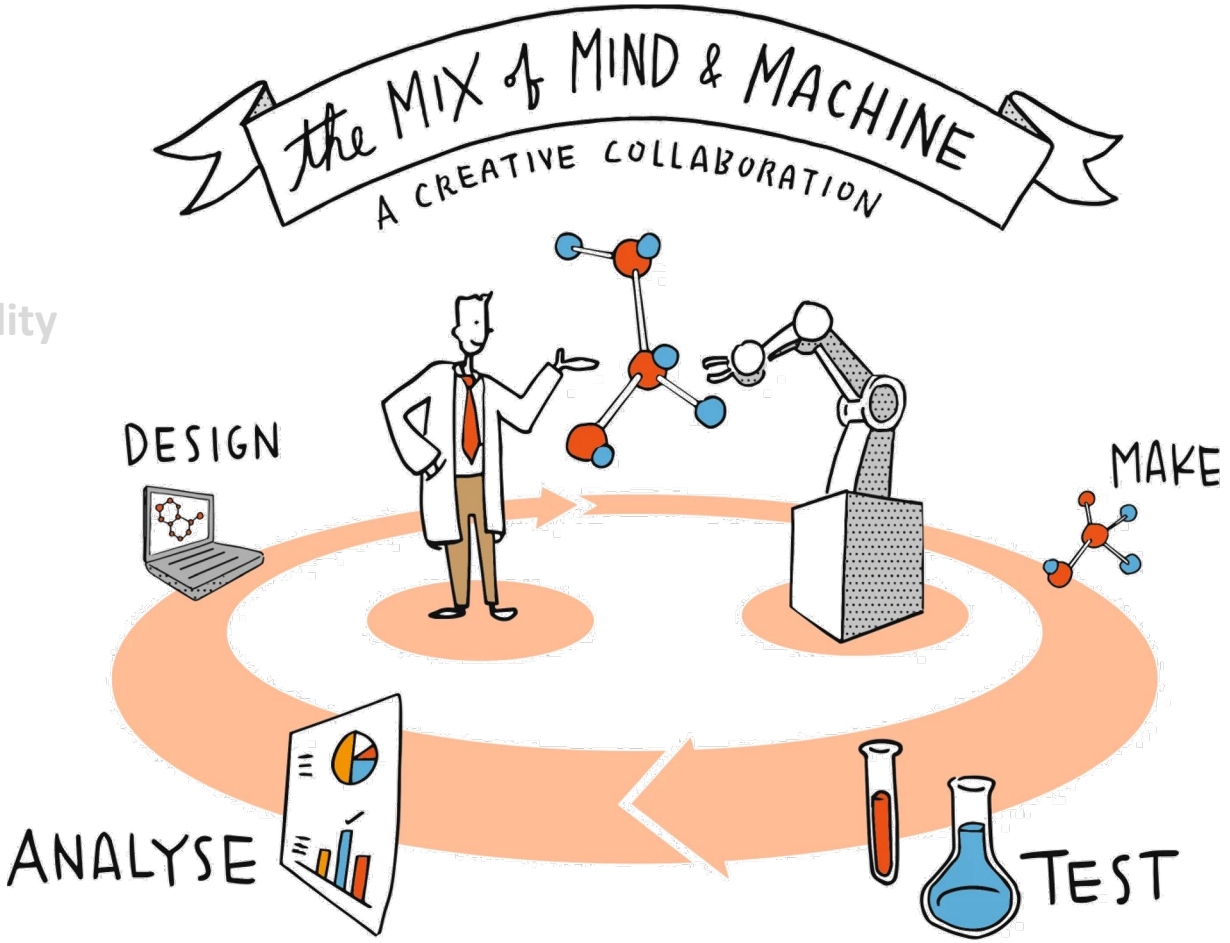
spending

2,000,000,000 US$ =

1 NEW DRUG

Developing a new drug is a multi-objective optimisation (MOO) problem





- Drug discovery and development can be a **chaotic journey** on a multidimensional landscape that is not getting easier
- Pathway is typically **sensitive to initial conditions** and the fact that **many different end points can result from the same starting point**
- **Different teams in different companies will end up with different drugs**, by taking a slightly different perspective on data or ideas generated from the same starting points, which are then influenced by emerging observations as each program evolves

Left: Hann MM, MedChemComm 2011, 2, 349-355. doi: 10.1039/C1MD00017A
Right: Schneider et al., Nature Rev Drug Discov 2020, 19, 353–364.

Computers in drug discovery

- **Bioinformatics**
  - Analysis of genes and genomes, protein structure and function, and interactions of biomolecules
- **Cheminformatics**
  - Design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information
  - Main driver today: machine learning
- **Molecular modeling**
  - Modeling the structure and properties of small molecules, biomacromolecules, and their interactions
  - Term usually used in the context of forcefield-based methods
- **Computational chemistry**
  - Modeling and prediction of physicochemical properties
  - Term usually used in the context of quantum chemistry

Cheminformatics

- Definition by Greg Paris (1998):
  "Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information"
- Definition by Johann Gasteiger (2004):
  - *"Chemoinformatics is the application of informatics methods to solve chemical problems"*
- Definition from Wikipedia:
  "Cheminformatics … is the use of computer and informational techniques applied to a range of problems in the field of chemistry. These in silico techniques are used, for example, in pharmaceutical companies in the process of drug discovery. These methods can also be used in chemical and allied industries in various other forms."

Cheminformatics – Why?

- Urgent need for **more efficiency in drug discovery**

- The relevant **chemical space is of enormous dimensions**. It cannot be systematically explored by experimental or theoretical methods

- Available data are too large for manual processing
- Need for means to store, organize, search, visualize and analyze data
- The **physicochemical and biological properties of small molecules are often unknown**:
  - For only about 1 in 1000 known chemicals the 3D structure has been experimentally determined
  - Even for many approved drugs the mode of action is unknown
  - We are lucky enough if we know a single target of a compound, but we generally do not know the full bioactivity spectrum of compounds yet
- **Structure-activity and structure-property relationships can be highly complex**
- Computers can provide answers quickly, and at low cost

Dimensions of the chemical space

Type	# molecules
Particles in the (observable) universe	10^{82}
Molecules < 1000 Da consisting of C, N, O, P, S, Hal, H	Up to 10^{180}
Drug-like molecules based on extrapolation on GDB-17[2] based on stitching together up to 30 carbon, nitrogen, oxygen, and sulfur atoms in different arrangements	 10^{33} 10^{63}
Merck Accessible Inventory (MASSIV)	10^{20}
Chemical universe database GDB-17: Listing all molecules up to 17 atoms[1]	166,400,000,000
Make-on-demand compounds in the public domain	> 5,000,000,000
On-stock compounds	230,000,000
Known natural products	700,000
Purchasable natural products	25,000[3]

[1] Ruddigkeit L et al. J Chem Inf Model 2012, 52, 2864–2875. doi: 10.1021/ci300415d
[2] Polishchuk PG, Madzhidov TI, Varnek A. J Comput-Aided Mol Des 2013, 27, 675–679.
[3] Chen Y. et al., J Chem Inf Model 2017, 57, 2099–2111.
Figure: https://cen.acs.org/pharmaceuticals/drug-discovery/Hunting-drugs-chemical-space/100/i23

Computational approaches in drug discovery & development

Computers in drug discovery and development

Discovery phase (~3 years)

Development phase (~2 years)

Target selection

- Data acquisition and management
- Bioinformatics
- Protein structure and function prediction
- Binding pocket identification
- Druggability prediction
- …

Hit identification

- Data acquisition and management
- Physicochemical property profiling and filtering
- Virtual screening
- De novo design
- Generative chemistry
- …

Hit to lead

- Multi-objective optimization
- Structure-based modeling
- Similarity-based approaches
- (Q)SAR and QS(P)R modeling
- Activity cliff exploration
- ADME prediction
- Toxicity and safety profiling
- …

Lead optimization

Candidate selection

- Physiologically-based pharmacokinetic (PBPK) modeling
- Dose-response modeling
- Adverse event prediction
- Clinical risk assessment
- …

Pre-clinical studies

Clinical trials (~7 years)

Registration (~1-2 years)

Clinical phase I

Clinical phase II

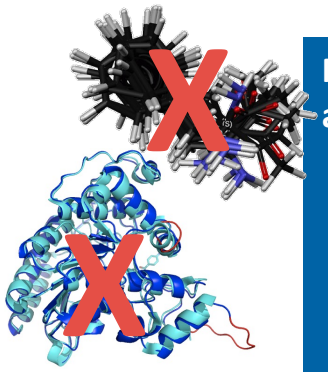
Clinical phase III

Registration

Clinical phase IV

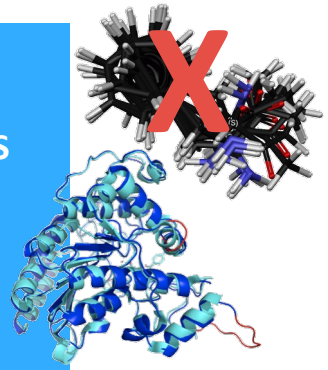
- Study design and data management
- Clinical risk assessment
- Statistical software and AI tools for study design and assessment
- …

Application scenarios for computational methods in early-stage drug discovery



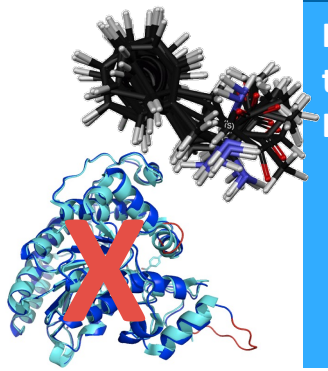
No information on target structures and ligands available

- Selection of a diverse compound library
- Filtering for physicochemical properties, chemical reactivity, etc.



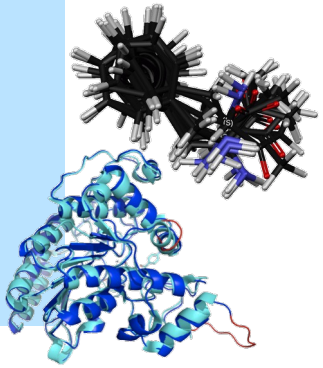
Information on target structures but not on ligands available: STRUCTURE-BASED APPROACHES

- Design of focused libraries
- Structure-based virtual screening (docking)
- *De novo* design at (putative) ligand binding site



Information on ligands but not on target structures available: LIGAND-BASED APPROACHES

- Design of focused libraries
- Similarity- & shape-based screening
- Pharmacophore-based screening
- QSAR
- Virtual compound synthesis



Information on target structures and ligands available

- Any ligand- or structure-based approach and combinations thereof

- There is no in silico approach or method in existence that consistently outperforms all others
- Individual computational models commonly represent only a small fraction of the bioactive conformational space → Combination of methods and models is key!
- Machine learning has become a key technology in all four scenarios

universität wien

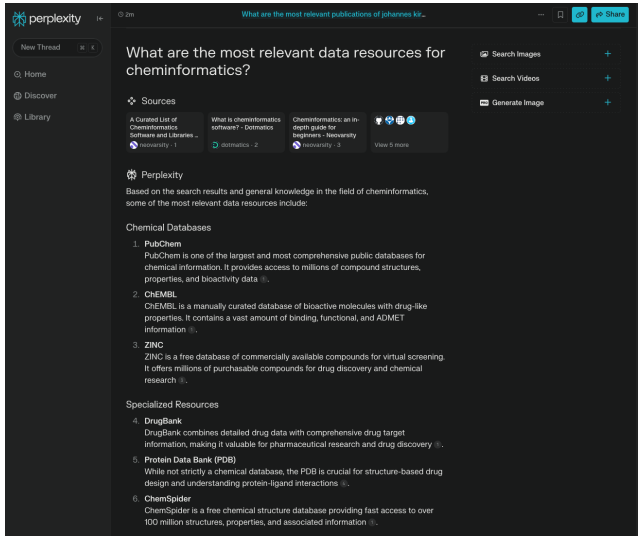
Important data and information sources

Johannes Kirchmair

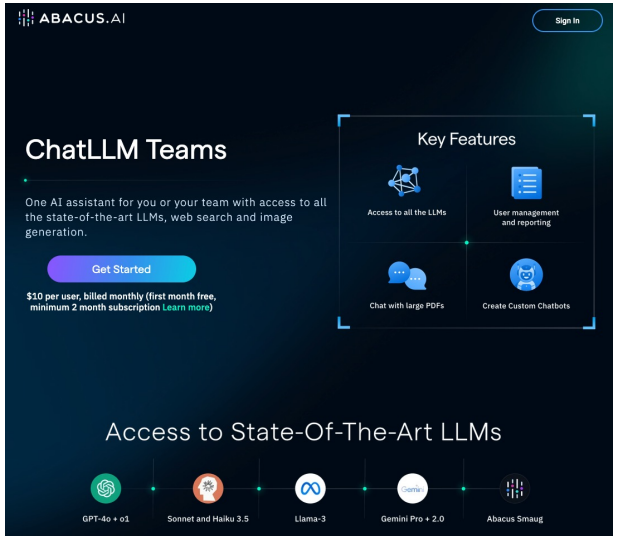
Data for cheminformatics and structural bioinformatics

Data type	Primary resource	Key numbers
Primary literature	The web	>6 k per day
Protein X-ray structures	Protein Data Bank (PDB)	>230 k structures
Small molecule X-ray data	Cambridge Crystallographic Database (CSD)	1.25 million compounds
Bioactivities	PubChem and PubChem BioAssays ChEMBL database	~300 million on 119 million compounds >21 million on 2.5 million compounds, covering >16 k targets
Structure and physicochemical properties of small molecules	CAS SciFinder PubChem	>100 million (+5 million per year) ~110 million
Make-on-demand compounds	Enamine REAL, Wuxi Chemistry, ZINC (meta database)	Several billion
In-stock compounds	ZINC	>12 million
Natural products with known chemical structure	COCONUT	up to 700k
Chemical reactions	CAS SciFinder	>130 million (+5 million per year)

Perplexity.ai, Abacus.ai and related tools...



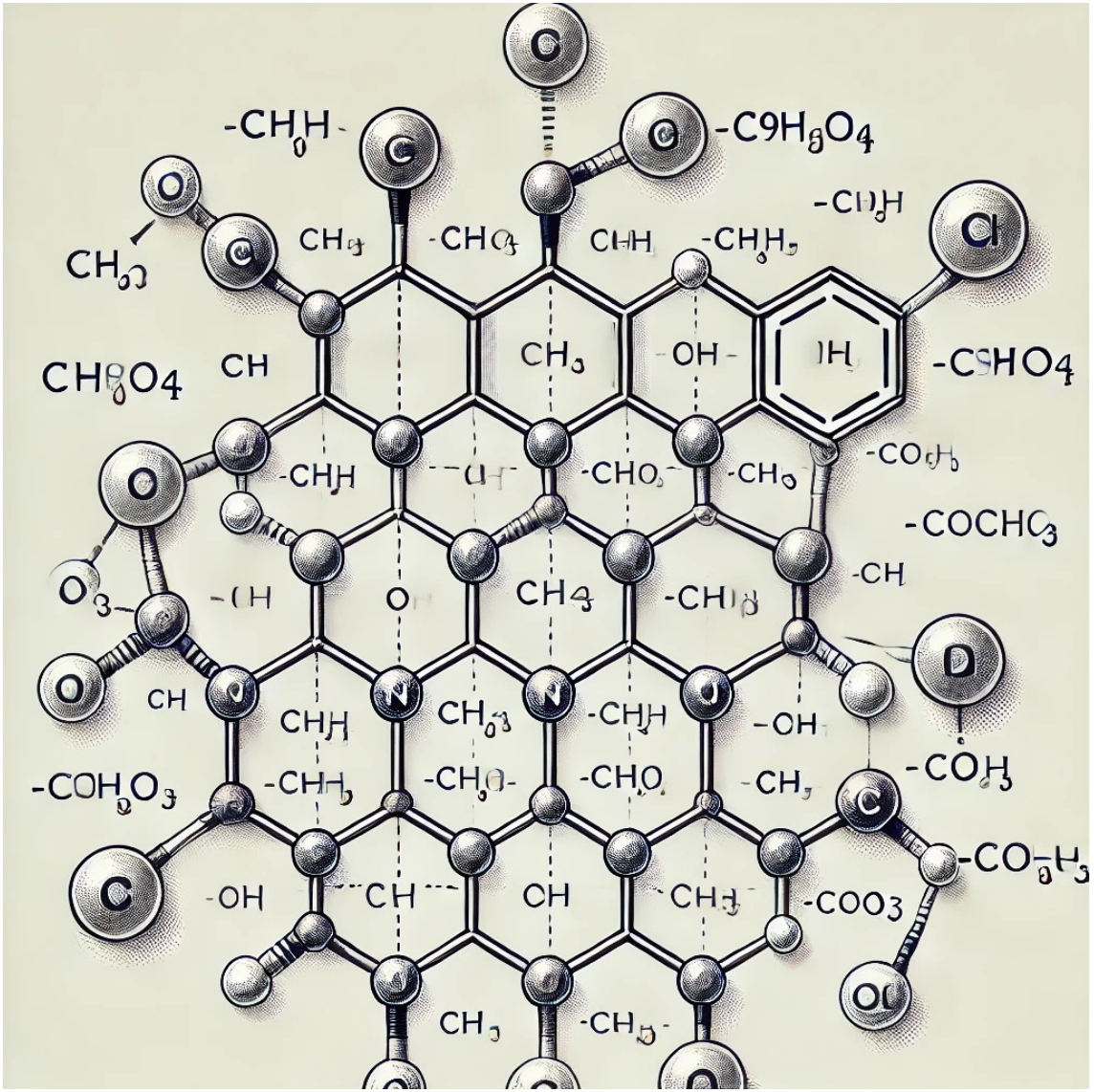
- Rapid literature review
- Up-to-date information
- Cross-disciplinary exploration
- Question answering
- Data analysis assistance
- Source citation
- Hypothesis generation
- Research planning



- Access to most SOTA LLMs
- AI-driven code editor and generator

ChatGPT:

Here is an image of the molecular structure of aspirin with the atoms labeled according to their atomic numbers.



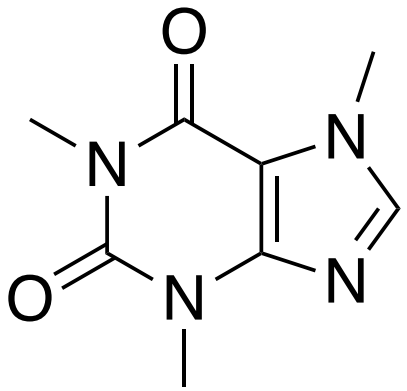
Let me know if you need further details or modifications!

universität wien

Molecular representations

Johannes Kirchmair

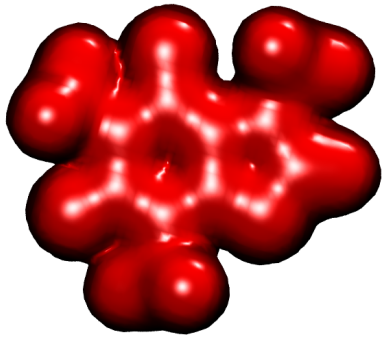
What is a molecule?



«caffeine»

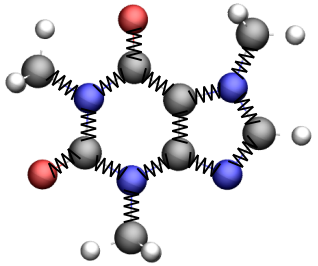
$C_8H_{10}N_4O_2$

Quantum Chemistry:



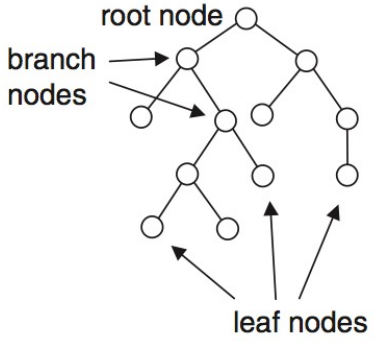
Electron density

Force Fields:

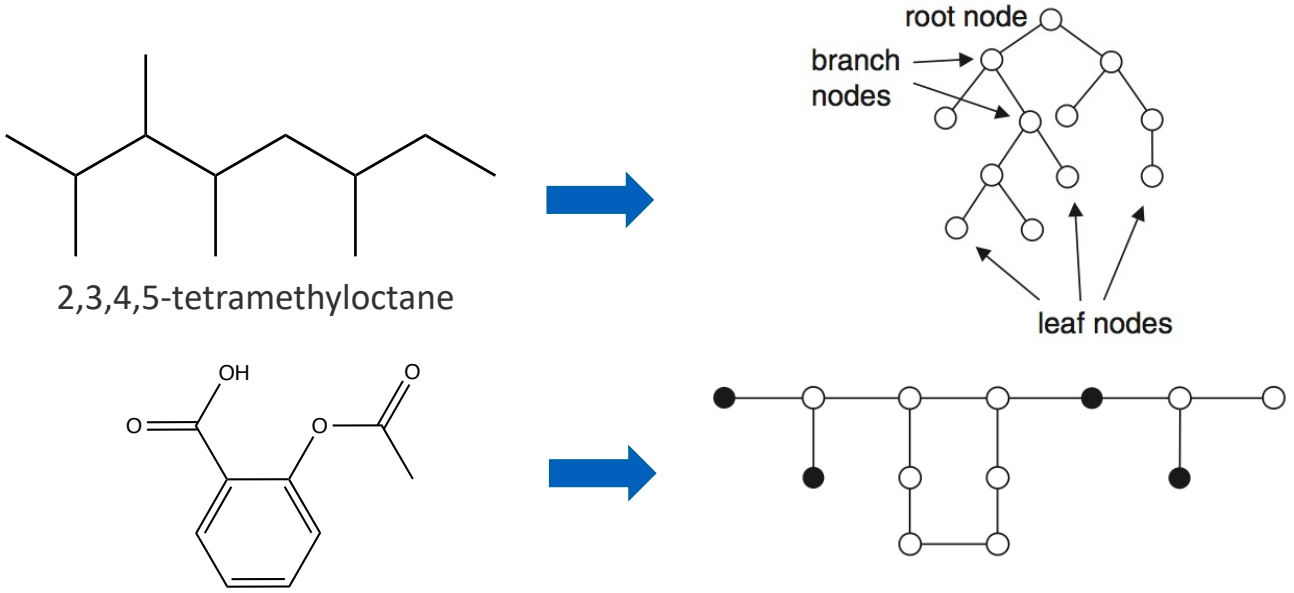


Set of harmonic springs

Graph representations:



Molecular graphs



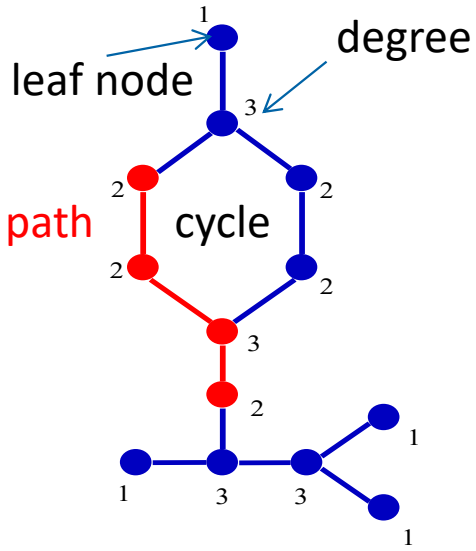
2,3,4,5-tetramethyloctane

automated optical chemical structure identification

- Chemical structures can be stored as images of course, but this is of little value to computers. In fact, it is one of the biggest problems
- One solution: Representation of molecules as molecular graphs: Graph theory
  - A graph represents the topology of a molecule: The way the nodes (or atoms) are connected
  - A molecular graph consists of nodes (atoms) and edges (bonds), often with properties associated with them (e.g. atom type, bond type)
  - Hydrogens often omitted

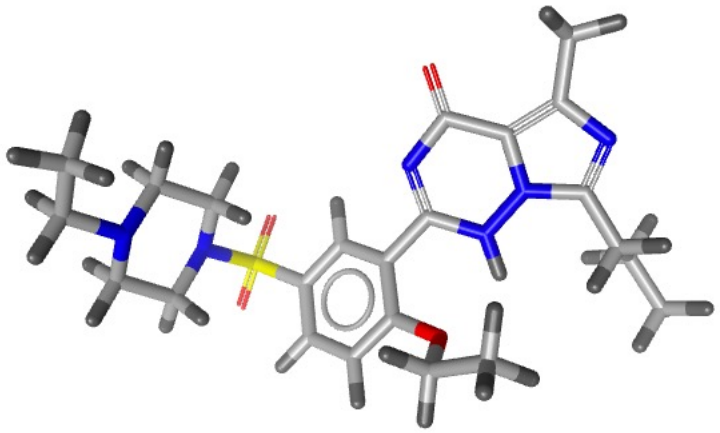
Graph terminology

Term	Explanation
Degree of a node	Number of edges meeting at the node
Leaf node	A node with degree 1
Path	Connected sequence of edges between two nodes
Cycle	Path which returns to its starting node
Tree	Graph without cycles
Subgraph	Subset of nodes and edges of another graph

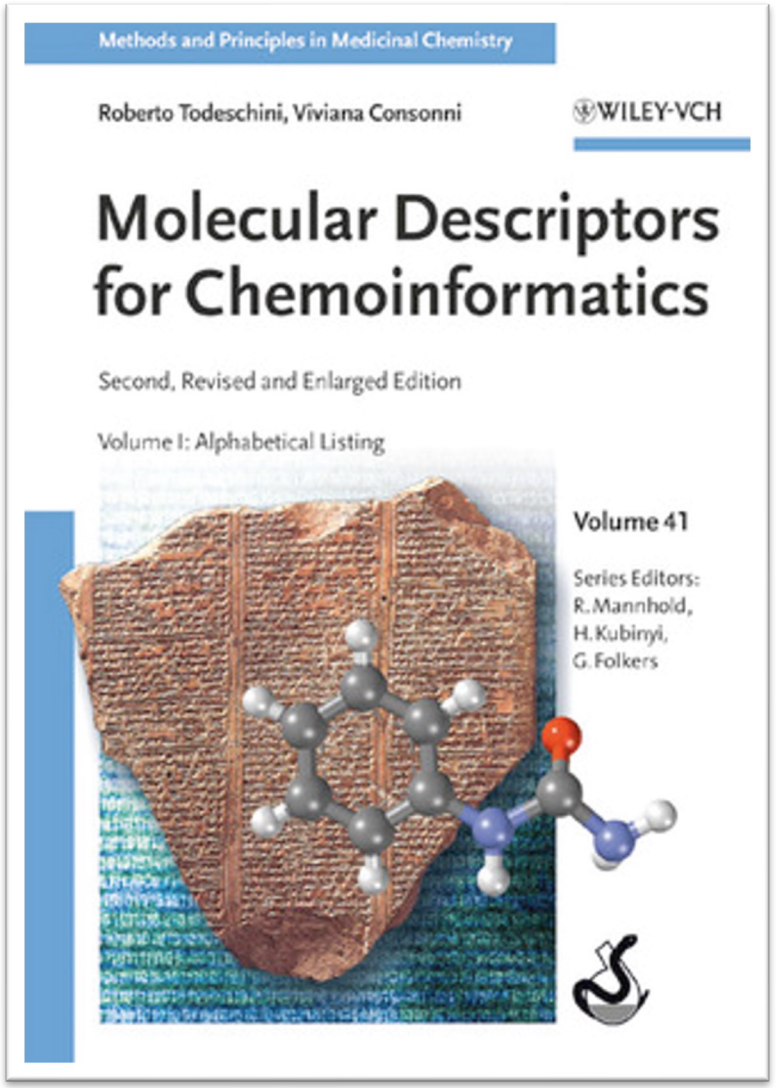


How can we describe the physicochemical properties of molecules?

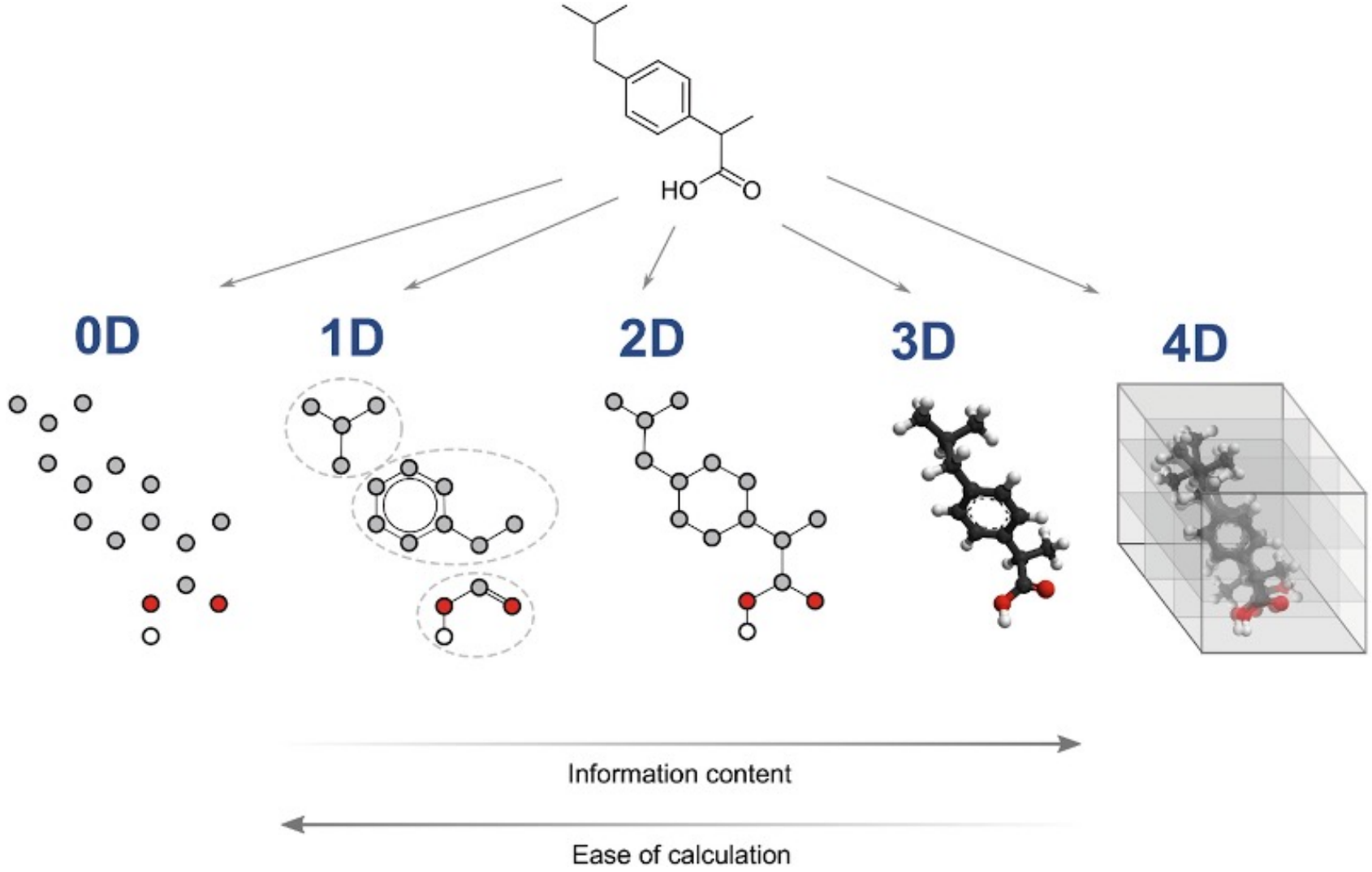
- Nearly 5000 chemical descriptors available today
- Tricky question: Which ones are applicable to my problem?



sildenafil



Types of descriptors



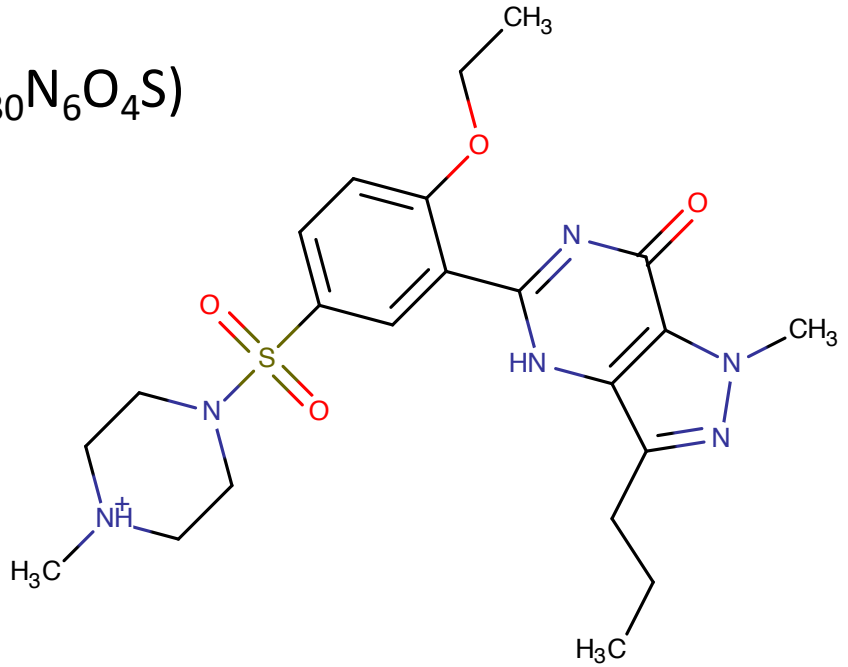
Types ("dimensionality") of descriptors

Descriptor class	Derived from	Definition	Examples
0D-descriptors	Atom list	Represent intrinsic properties of a molecule that do not depend on its structure or connectivity → Can be derived directly from the chemical formula	Molecular weight; number of atoms, bonds, or specific elements
1D-descriptors*	Sequential data	Descriptors calculated from substructural information or sequential representations. Topology information does not need to be complete	H-bond donors; number of specific substructures
2D-descriptors*	Molecular graph	Single-valued descriptors calculated from molecular graph representations (topology). Sensitive to structural features of the molecule (size, shape and symmetry) → most commonly used descriptors	Topological indices
3D-descriptors	Molecular geometry	Descriptors calculated from 3D molecular structures (geometries)	Molecular surface area and volume; dipole moment; 3D pharmacophores; steric and electrostatic fields
4D-descriptors	Molecular geometry + spatio-temporal components	Addn. dimension that captures the dynamic behavior of molecules over time, considering multiple conformations and their transitions (often derived from molecular dynamics simulations)	Ensemble pharmacophore models; molecular dynamics simulations of binding modes; conformational ensemble-based properties

* Note that there are some ambiguities in the definition of the classes of 1D and 2D descriptors in the scientific literature

0D-descriptors: Constitutional descriptors, counts

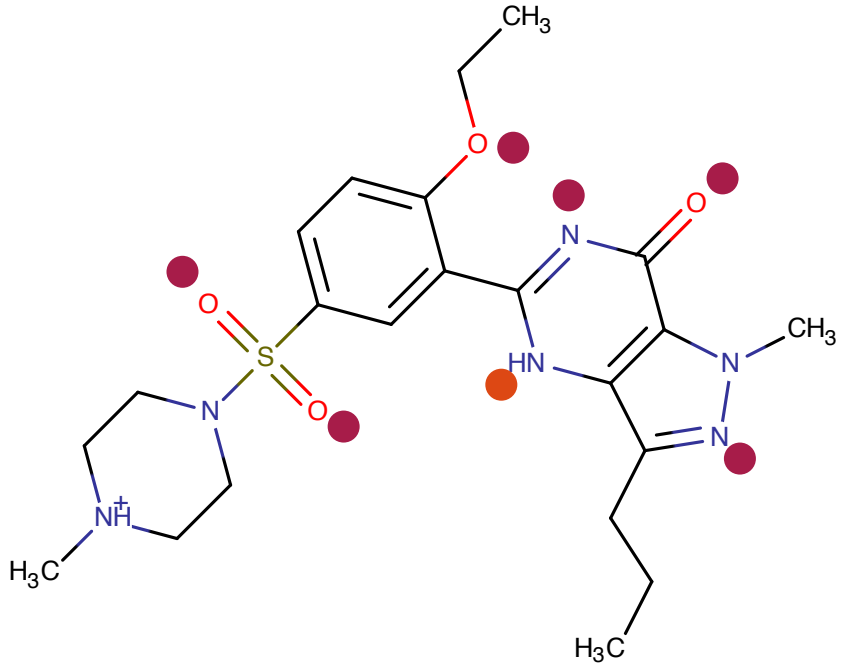
- Any descriptors for which no information about the molecular structure and atom connectivity is required
- Can be derived directly from the chemical formula (e.g. $C_{22}H_{30}N_6O_4S$)
- Examples:
  - Simple atom counts
    - Number of nitrogen atoms
    - Number of oxygen atoms
  - Sum or average of atom properties
  - Molecular weight
  - Simple bond counts



No. of N:	6
No. of O:	4
MW:	474

1D-descriptors: Substructural information

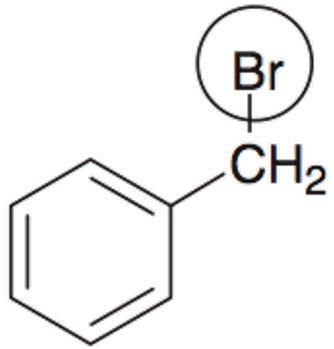
- Any descriptors calculated from substructural information
- Examples:
  - Counts of functional groups and substructure fragments
  - Number of hydrogen bond acceptors (HBA)
  - Number of hydrogen bond donors (HBD)
  - Number of sulfonamide groups
  - Fragment-based descriptors
    - logP descriptor

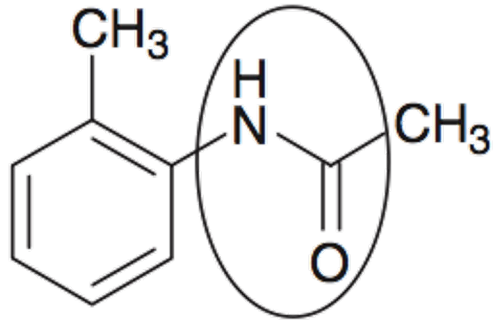


No. of HBA:	6
No. of HBD:	1
No. of piperazines:	1
No. of sulfonamides:	1
No. of ethyl groups:	2

ClogP: Use of structural fragments

- Fragment-based approach (hence 1D descriptor)
- Significant electronic interactions can be taken into account
- Not accounting for intramolecular hydrogen bonds
- Challenge: Estimating contributions of fragments not in the training set





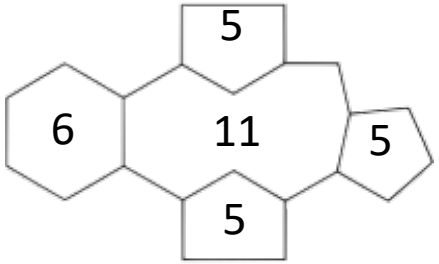
Bromide fragment	0.480
1 aliphatic isolating carbon	0.195
6 aromatic isolating carbons	0.780
7 hydrogens on isolating carbons	1.589
1 chain bond	-0.120
Total	2.924

NH-amide fragment	-1.510
2 aliphatic isolating carbons	0.390
6 aromatic isolating carbons	0.780
10 hydrogens on isolating carbons	2.270
1 chain bond	-0.120
1 benzyl bond	-0.150
ortho substituent	-0.760
Total	0.900

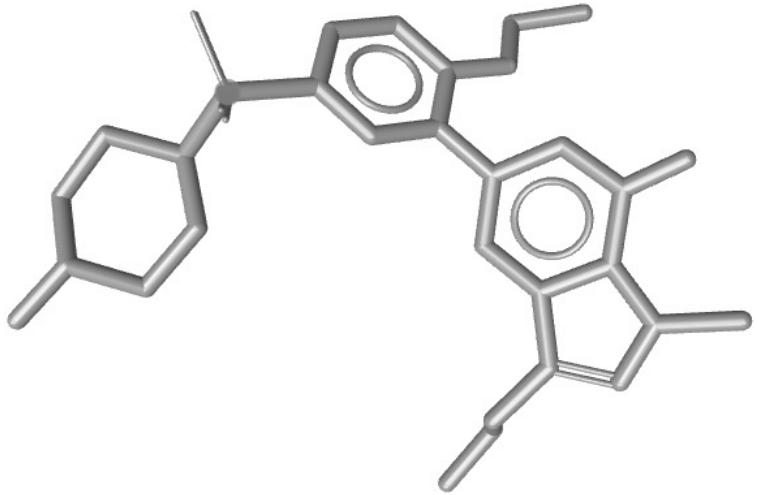
isolating carbon: carbon not doubly or triply bonded to a heteroatom

SSSR (smallest set of smallest rings)

- Set of rings from which all others in the molecular graph can be constructed
- Comprises those rings containing the fewest atoms
- Used for quick structure search (e.g. downsize number of molecules in substructure searches)

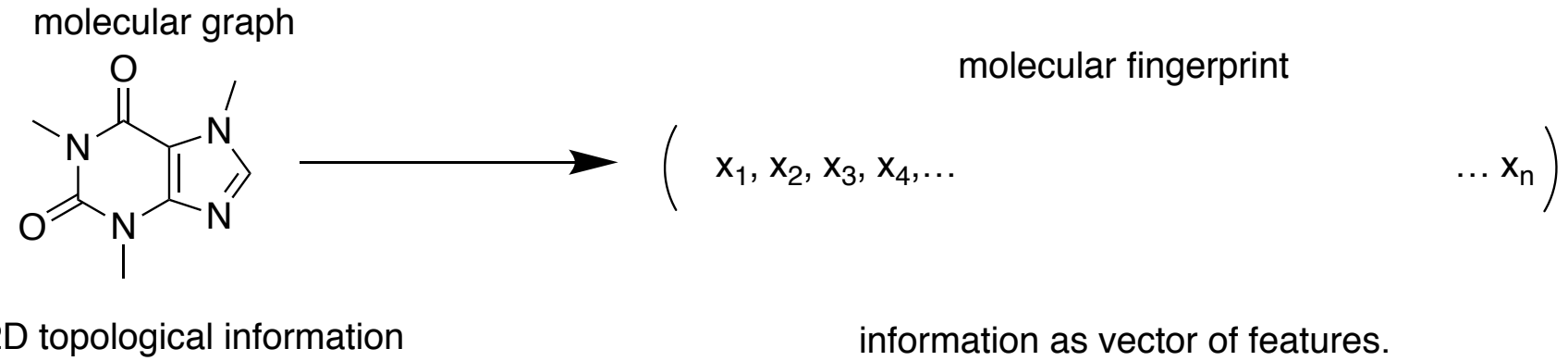


SSSR=(5,5,5,6,11)



SSSR=(5,6,6,6)

Molecular fingerprints



- Helpful encoding of chemical information
- Vectors are usually filled with bits (0 or 1)
- Fingerprints are one of the preferred types of input in machine learning models

Types of molecular fingerprints

Fingerprint types	Description	Examples
Dictionary-based (also “structural”) keys	Encode predefined structural features of molecules	MACCS (Molecular ACCess System) structural keys
Topological and path-based fingerprints	Describe combinations of atom types and paths between atom types	AP (atom pair) fingerprints
Circular fingerprints	Encode circular atom environments up to a certain bond radius from the central atom	ECFP (extended connectivity fingerprint), FCFP (functionality connectivity fingerprints), Morgan fingerprints
3D Pharmacophore fingerprints	Encode the presence of pharmacophore features in molecules in 3D space	

- **Keyed representations (“Keys”):** Each bit position corresponds to the presence or absence of a specific feature (or, albeit much less frequently used, a feature count)
- **Hashed representations (“Fingerprints”):** Features are mapped to overlapping bit segments (hence producing specific bit patterns without 1:1 bit-to-feature correspondence

Dictionary-based structural keys

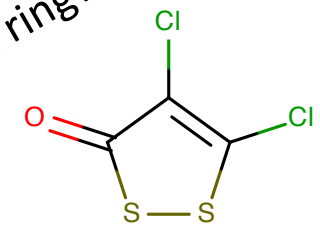
MACCS (Molecular ACCess System) structural keys

- One of the earliest developed and most established dictionary-based structural keys
- Consists of 166 predefined substructures defined as SMARTS
- Developed for substructure screening rather than similarity search. It does work for similarity search, however, not well.

Bitstring:

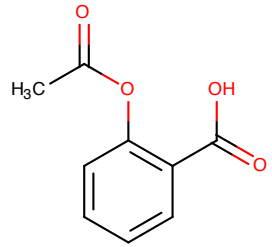
1	1	0	0

Is there a chlorine?
Is there an S-S bond?
Is there a ring size of 6?
Is there an oxygen in a ring?

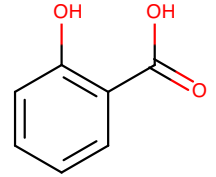


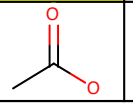
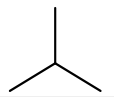
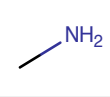
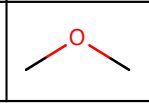
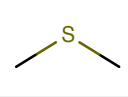
Further example:

A



B



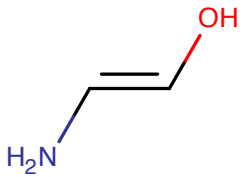
A	1	1	0	1	0	1	0
B	1	1	0	1	0	0	0
							

Dictionary-based structural keys

- Boolean array encoding the presence (TRUE)/absence (FALSE) of predefined structural fragments of a molecule as a bit string
- Boolean arrays can be compared quickly and a similarity score determined
  - Enables fast substructure search, which normally is computationally expensive (NP-complete problem)
  - Enables fast ranking of molecules according to their similarity to query compound(s)
- Originally designed for high-speed search/filtering of large databases, not to quantify chemical similarity or describe chemistry
- Important limitation: Pre-definition of keys in a dictionary leads to a lack lack of generality → applicability depends on the specific dataset

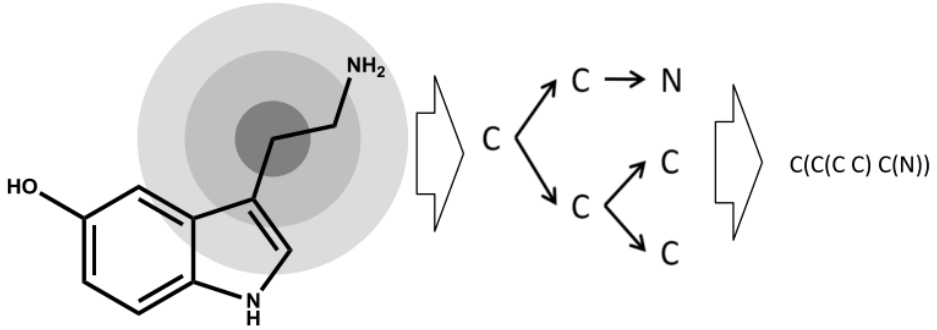
Topological and path-based fingerprints

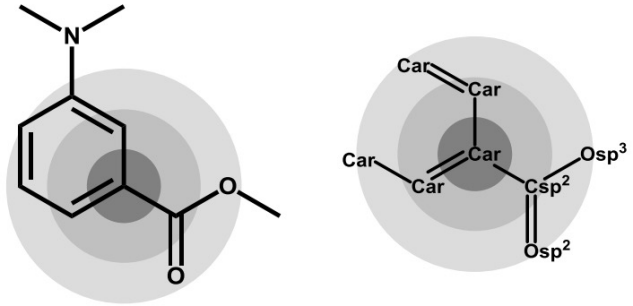
- Example: Atom pair fingerprints

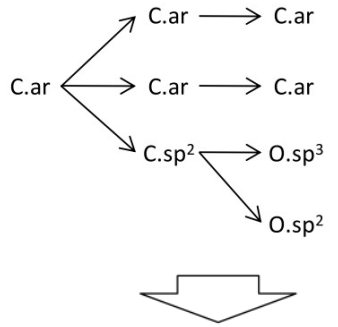


0-bond paths:	**C**	**O**	**N**
1-bond paths:	**OC**	**C=C**	**CN**
2-bond paths:	**OC=C**	**C=CN**	
3-bond paths:	**OC=CN**		

- Example: Circular fingerprints







Fingerprint: C.ar ; 2x C.ar , $C.sp^2$; 2x C.ar , $O.sp^2$, $O.sp^3$

In both examples:

- An exhaustive list of patterns is produced, up to the path length limit
- Number of possible patterns is huge → assignment of a particular bit for each pattern not feasible → use of a “hash function” → “hashed fingerprint”

http://www.daylight.com/dayhtml/doc/theory/theory.finger.html
Samuel E. Adams, Ph.D. Thesis, Univ. of Cambridge, 2010

Hashed fingerprints I

- Hashed fingerprints are a result of the evolution of structural keys, from patterns defined by a dictionary to tailored patterns generated from the molecule itself → hashed fingerprints are independent of a dictionary and –in principle– applicable to all molecular structures
- Hashing algorithm:
  - Each pattern (key) serves as a seed to a pseudo-random number generator (hashing algorithm), the output of which is a set of bits (typically 4 or 5 bits per pattern)
  - The set of bits thus produced is added (with a logical OR) to the fingerprint

Hashed fingerprints II

- Advantages:
  - Generally applicable: One fingerprint serves all databases and all types of queries
  - More effective use is made of the bitmap: Structural keys are usually very "sparse" (mostly zeros) since a typical molecule has very few of the patterns that the structural key's bits represent. Hashed fingerprints can be relatively "dense" (20-40% of bits set to “1”) without losing specificity.
- Disadvantages:
  - (Low) risk of overlapping bits (which means loss of information):
    - Each pattern generates its defined set of bits → as long as at least one of those bits is unique (not shared with any other pattern present in the molecule), we can tell if the pattern is present or not
  - Limited interpretability:
    - Fingerprint cannot be converted back into structural features (since one bit doesn’t encode a specific feature)
    - If a bit of a fingerprint indicates a pattern is missing then it certainly is missing, but if a bit of a fingerprint indicates that a pattern is present it can do this only with some probability (since this bit could also encode other features)

General limitations of molecular fingerprints

- Encode presence and absence of features but not chemistry itself
- Generally do not encode stereochemistry
- Generally do not encode the number of instances of a specific feature present in a molecule
  → “holograms” encode the number of instances as an integer rather than a bit vector
- Connectivity of the individual features is lost:
  Bit strings of molecules A-B-C and C-A-B are identical

2D-descriptors

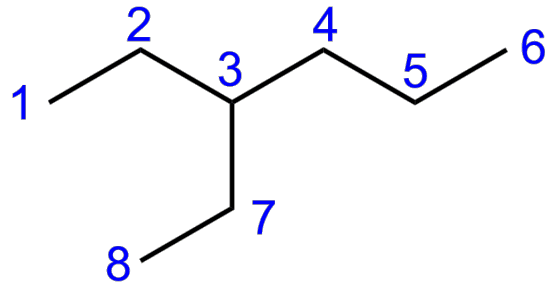
- Single-valued descriptors calculated from the H-depleted molecular graph representation
- Encode adjacency and connectivity
- Sensitive to structural features such as size, shape, symmetry and degree of branching
- Two types:
  - Topostructural descriptors
    - Encode 2D graph information only: Size, branching, overall shape
  - Topochemical descriptors
    - Encode also specific chemical properties of atoms, e.g mass or hybridisation states

Topostructural indices

Wiener Index (1947)

- Oldest topological index related to the **branching of a molecule**
- Closely **correlated with the boiling points** of alkane molecules
- The path number **w is defined as the sum of the distances D (counted as number of bonds) between any two carbon atoms in the molecule, i and j, in terms of heavy atom bonds**
- Wiener Index decreases with a higher degree of branching

$$w = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} D_{ij}$$



3-ethylethane: *w*=72, resulting from the sum of distances between atom pairs:

1-2(1), 1-3(2), 1-4(3), 1-5(4), 1-6(5), 1-7(3), 1-8(4),
2-3(1), 2-4(2), 2-5(3), 2-6(4), 2-7(2), 2-8(3),
3-4(1), 3-5(2), 3-6(3), 3-7(1), 3-8(2),
4-5(1), 4-6(2), 4-7(2), 4-8(3),
5-6(1), 5-7(3), 5-8(4),
6-7(4), 6-8(5),
7-8(1)

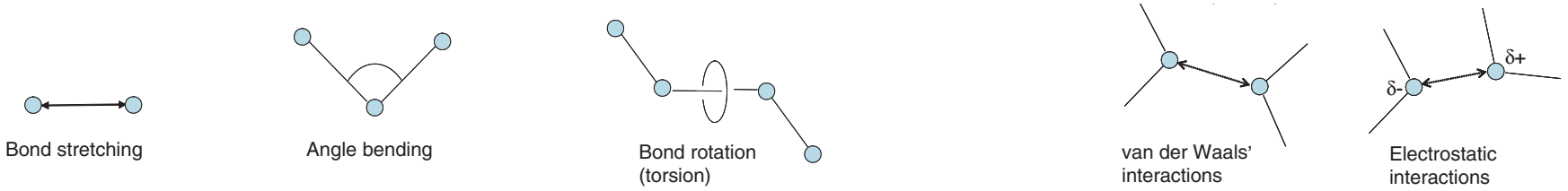
1+2+3+4+5+3+4 +1+2+3+4+2+3 +1+2+3+1+2 +1+2+2+3 +1+3+4 +4+5 +1 = 72

Substance	Wiener-Index
n-hexane	35
2-methylpentane	32
3-methylpentane	31
2,3-dimethylbutane	29

3D Descriptors: Geometrical representation of molecules

- Use Cartesian coordinates of atoms to derive descriptions of molecules
- Advantages:
  - Rich information content
- Disadvantages:
  - Conformation needs to be known or at least determinable
  - Depend (to very different extents) on ligand conformation, which itself depends on the ligand environment (vacuum, solution, protein, etc.)
    - This is why they often do not perform substantially better than less complex descriptors
  - Because the relevant ligand conformation is often unknown, conformer ensembles are used to represent the relevant conformational space. However, this increases complexity but not necessarily accuracy
  - 3D-based approaches for virtual screening tend to obtain enrichment rates that are comparable to those of 2D approaches, but scaffold diversity among the correctly identified active molecules is generally higher

Empirical force fields



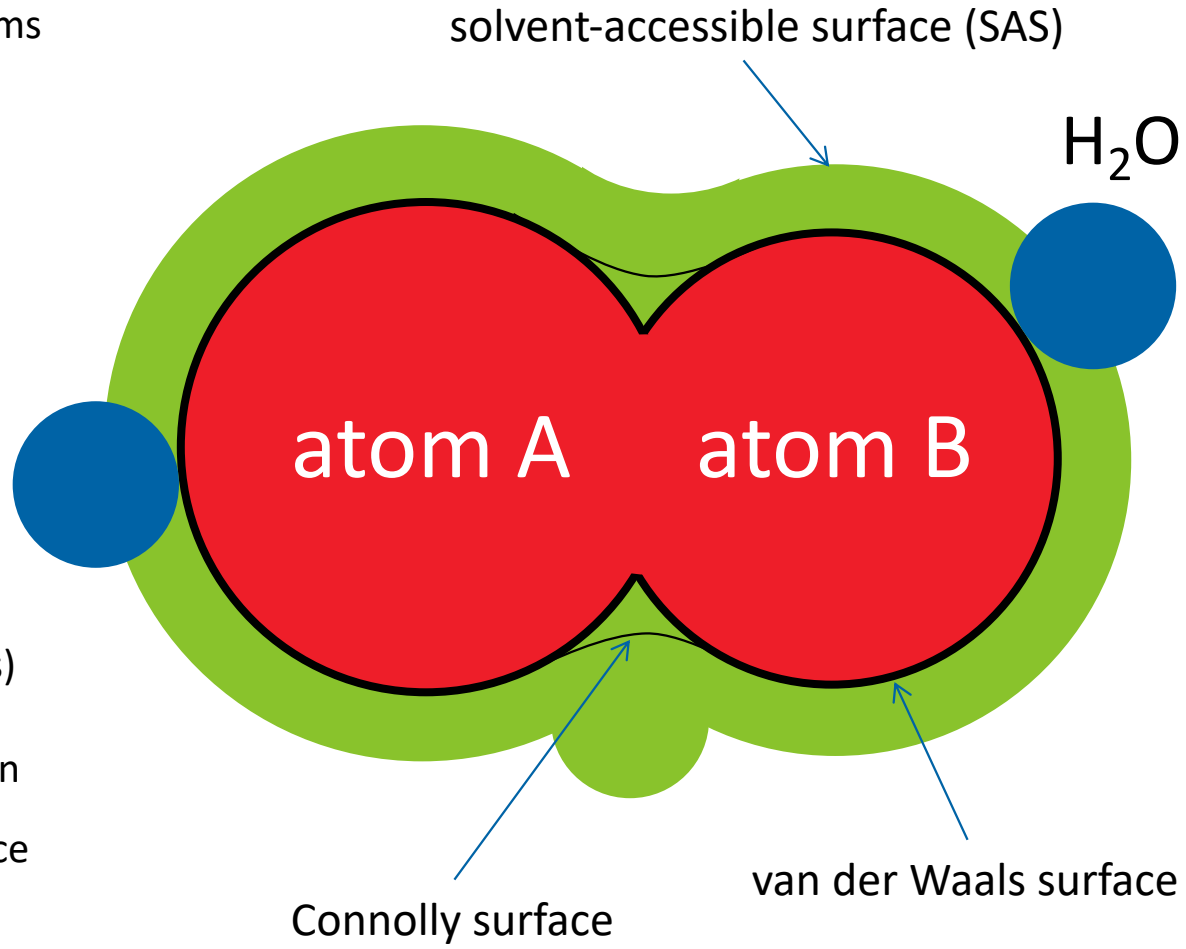
$$E = \underbrace{\sum_{bonds} k_r (r - r_0)^2 + \sum_{angles} k_\Theta (\Theta - \Theta_0)^2 + \sum_{torsions} k\Phi\left[1 + \cos(n\Phi - \Phi_0)\right]}_{\text{bonded}} + \underbrace{\sum_i \sum_{j \neq i} 4\varepsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i Q_j}{4\pi\varepsilon_0 r_{i,j}}}_{\text{non-bonded}}$$

- Force field: Functional form and parameter sets used to calculate the potential energy of a system of atoms
- Molecular mechanics is an **empirical approach**:
  - There is no reason why a specific functional form is necessarily better than any other
  - **Parameters** of the energy functions can be derived from experimental work and quantum mechanical calculations
- **Transferability**: It is assumed that the parameters derived from small sample systems can be applied to much larger molecules and molecular systems
- **Pair-wise additive approximation**: Interaction energy between one atom and the rest of the system is calculated as a sum of pair-wise (on atom to one atom) interactions, or as if the pair of atoms do not see the other atoms in the system
- **Fixed set of atom types**: The number of atom types is kept at a minimum by grouping, which can lead to errors
- Force fields ignore the electronic motions in the system and calculate the energy solely as a function of the atom positions
- Large number of force fields exists. Many specialized force fields: Organic molecules, biomacromolecules, sugars, etc.
- "All-atom" force fields: provide parameters for every type of atom in a system

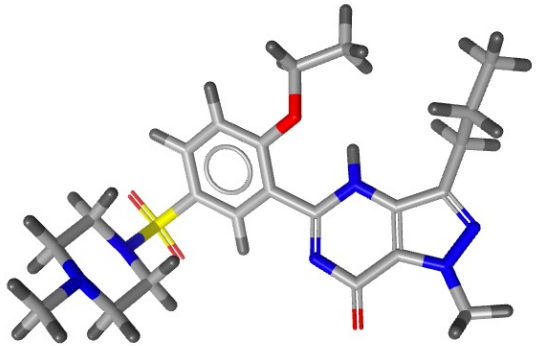
www.ch.embnet.org/MD_tutorial/pages/MD.Part2.html
Leach, A. In "Comprehensive Medicinal Chemistry II", Elsevier, 2006.

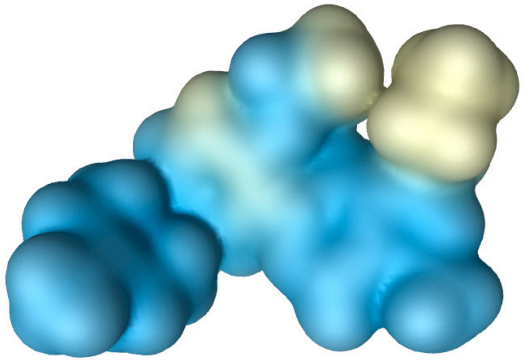
Surfaces

- **Van der Waals surface**
  - Simplest surface that represents the van der Waals radii of all atoms
  - Each atom represented by a sphere
  - The spheres of all atoms are fused
  - The total volume is the van der Waals volume, and the envelope defines the van der Waals surface
  - Quickly calculated
  - Limitation: Small cavities are artifacts and not of interest →
- **Connolly surface** (also "**molecular surface**" or "solvent-excluded surface")
  - Generated by simulating a sphere rolling over the van der Waals Surface
  - The sphere represents the solvent
  - Radius is typically 1.4 Å, which is the effective radius of water
  - Connolly surface has two regions:
    - Convex contact surface (segment of the vdW surface)
    - Concave surface (where the sphere touches two or more atoms)
- **Solvent-accessible surface (SAS)**
  - Defines the surface accessible to the solvent → find space that can accommodate water molecules
  - Path of the center of the probe that generates the Connolly surface
  - SASA is larger than a molecular surface
- Beware! Hydrogens need to be taken into account!

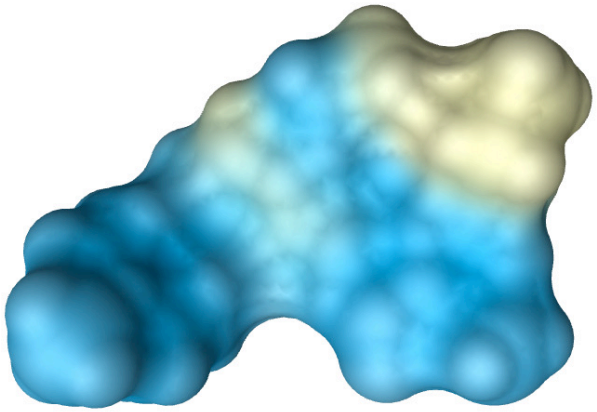


Physicochemical properties can be projected onto surfaces

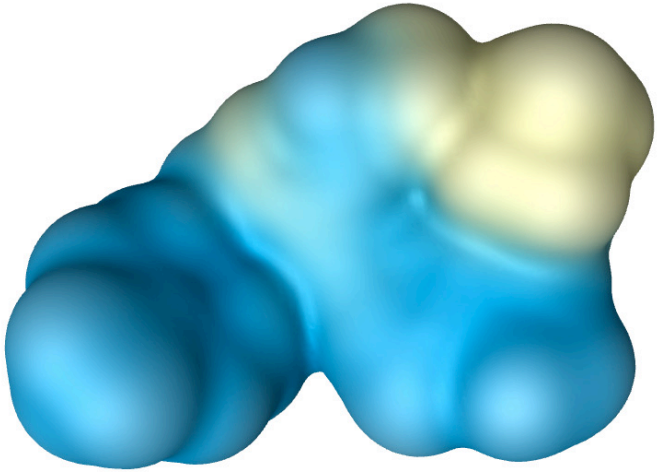




van der Waals surface

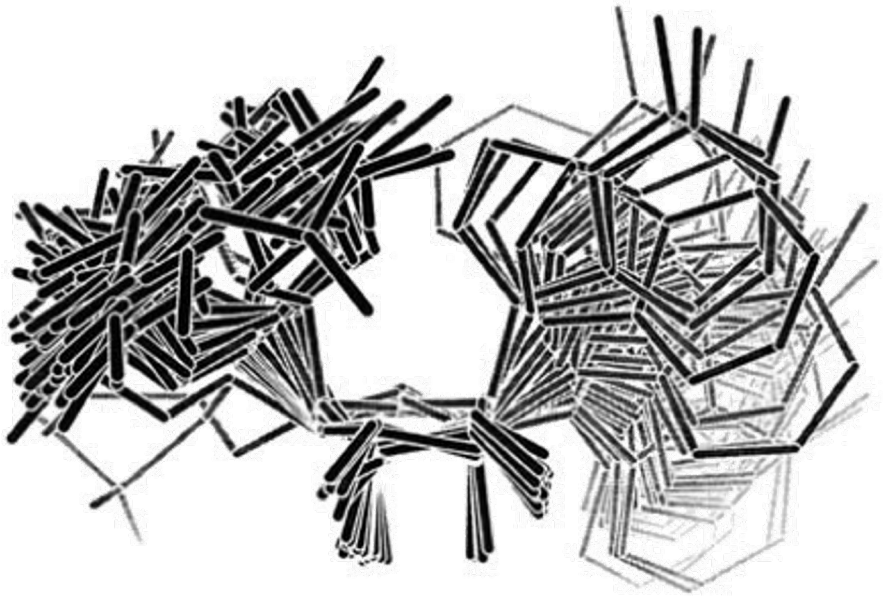


Connolly or molecular surface



Solvent-accessible surface (SAS)

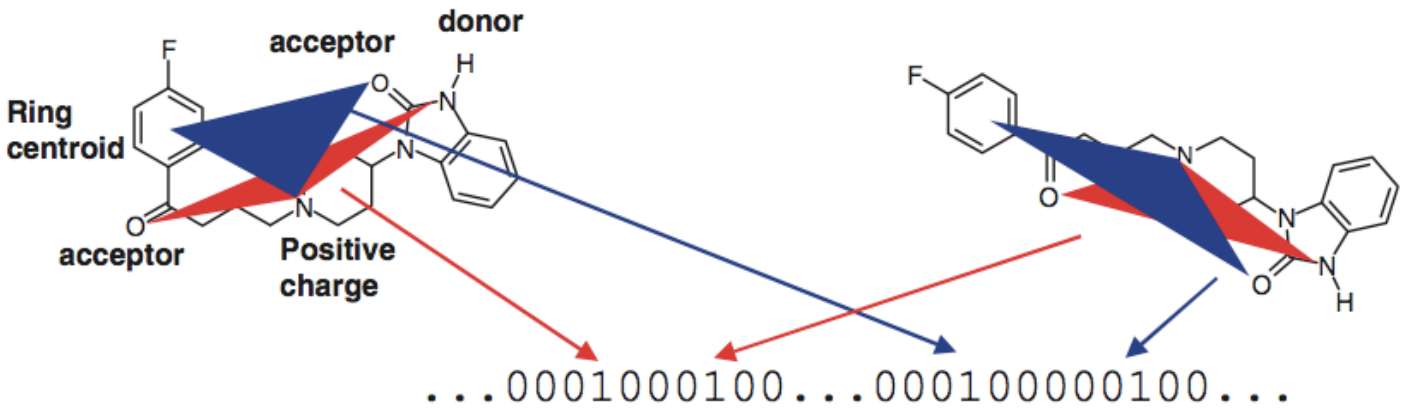
Conformational flexibility and its representation by conformer ensemble generators



Experimentally observed conformations for AMP

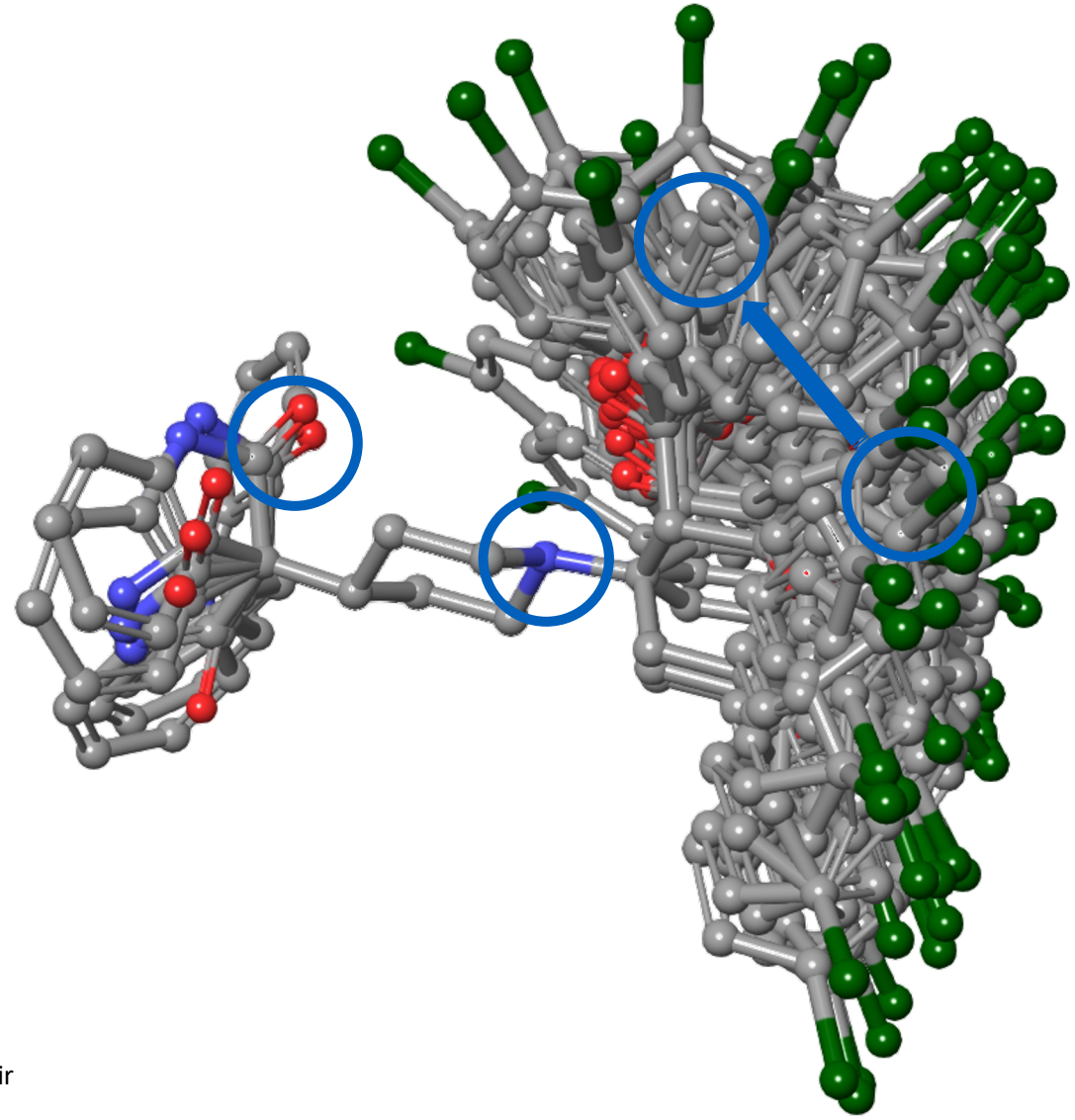
The relevant conformational space is attempted to be represented by a small ensemble of (14) conformers

3D Pharmacophore keys I



- Pharmacophore keys: characterised by the types of pharmacophore features involved and the distances (binned) between them
- Most common: Three or four-point pharmacophore keys
- Each bit in the pharmacophore key bit string thus represents one possible three/four-point pharmacophore.
- Pharmacophore keys are generated for each individual conformation. When comparing two molecules, the fingerprints of the individual conformations are compared and the best-matching fingerprints are used to quantify similarity

3D Pharmacophore keys II



3D pharmacophore keys are conformation-dependent: Example of an HBD-HBA-Aro pharmacophore:

conformer A: …0001010001…
conformer B: …0000011001…

Learned molecular representations

- Definition:
  - Molecular representations learned directly from data using algorithms like neural networks (e.g., GNNs, transformers)
- Aim:
  - Automated extraction of the features most relevant to a specific task without relying on pre-defined descriptor
- Advantages:
  - **Task-specific:** automatically capture features most relevant to the task (e.g., solubility prediction, docking)
  - **Higher expressiveness:** learn patterns beyond traditional descriptors
  - **Scalability:** handle large datasets and integrate multimodal information (e.g., combining 2D and 3D features)
- Examples:
  - **Sequence-based representations:** derived from models like transformers trained on SMILES strings
  - **Graph-based representations:** derived from GNNs, where molecules are treated as graphs (atoms = nodes, bonds = edges)
  - **3D geometric representations:** derived from 3D molecular structures using equivariant neural networks or force-field simulations

universität wien

Data formats

Johannes Kirchmair

The connection tables: MOL and SD file format

number of atoms

number of bonds

```
molecule name
-ISIS-  09270222202D

13 13  0  0  0  0  0  0  0  0  0999 V2000
   -3.4639   -1.5375    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4651   -2.3648    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7503   -2.7777    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0338   -2.3644    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0367   -1.5338    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7521   -1.1247    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7545   -0.2997    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0413    0.1149    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4702    0.1107    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3238   -1.1186    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6125   -1.5292    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6167   -2.3542    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.1000   -1.1125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  6  7  1  0  0  0  0
  3  4  2  0  0  0  0
  7  8  1  0  0  0  0
  7  9  2  0  0  0  0
  4  5  1  0  0  0  0
  5 10  1  0  0  0  0
  2  3  1  0  0  0  0
 10 11  1  0  0  0  0
  5  6  2  0  0  0  0
 11 12  2  0  0  0  0
  6  1  1  0  0  0  0
 11 13  1  0  0  0  0
M  END
>  <property 1>
property value 1

>  <property 2>
property value 2
```

atom block

bond block

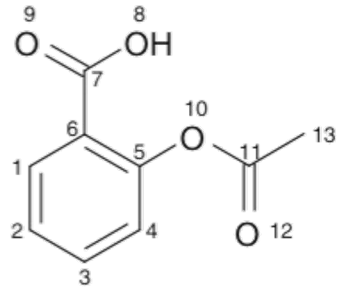
properties

the first atom is a carbon

stereo flags:
- 0 not stereo
- 1 clockwise
- 2 counter-clockwise
- 3 either or unmarked stereo center

the first three numbers are the *x*, *y* and *z* coordinates of the atom

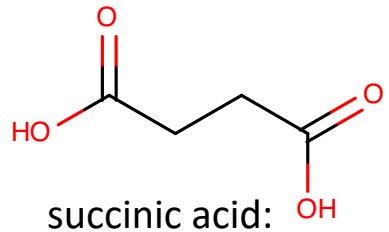
the first bond is between atoms 1 and 2 and has order 2



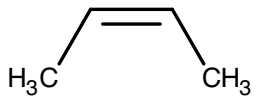
Hydrogens are often suppressed as they can be inferred from the element type, hybridisation and ionization state

SMILES - Simplified Molecular Input Line Entry Specification

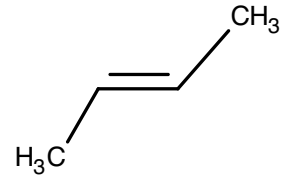
- Single line notation - fairly easy to interpret
- “Walk” through the chemical structure. Each atom is visited only once
- Branch points: Indicated using brackets
- Branches can be nested to any level necessary
- Hydrogens usually omitted
- Aromatic atoms: small letters
- Double bonds: =
- Triple bonds: #
- R/S: In this case, hydrogens are included
- cis/trans: pair of “/X=X/” or “/X=X\”
- “.”: Separator of components,
  e.g. salt components



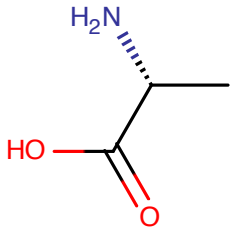
succinic acid:
OC(=O)CCC(=O)O



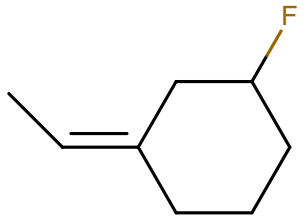
cis-butene
C/C=C\C



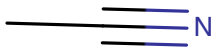
trans-butene
C/C=C/C



N[C@H](C)C(=O)O

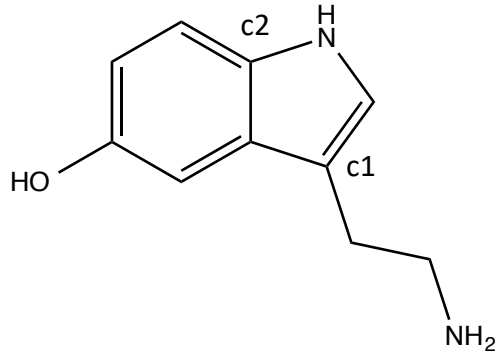


C\C=C1\CCCC(F)C1

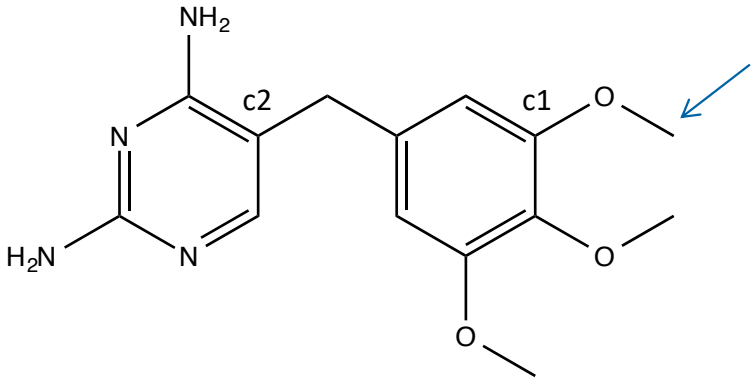


Leach & Gillet, An Introduction to Chemoinformatics, 2007, Springer.
http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html

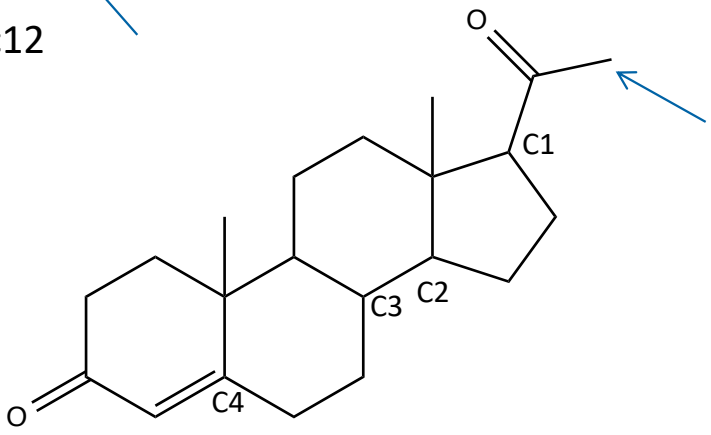
SMILES - Simplified Molecular Input Line Entry Specification



serotonin:
NCCc1cnc2ccc(O)cc12

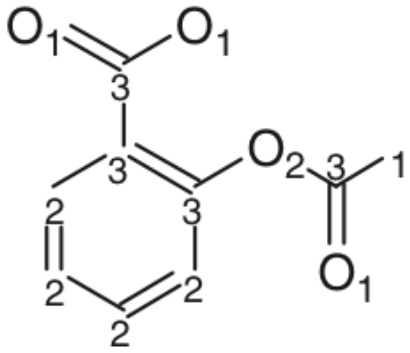


trimethoprim:
COc1cc(Cc2cnc(N)nc2N)cc(OC)c1OC

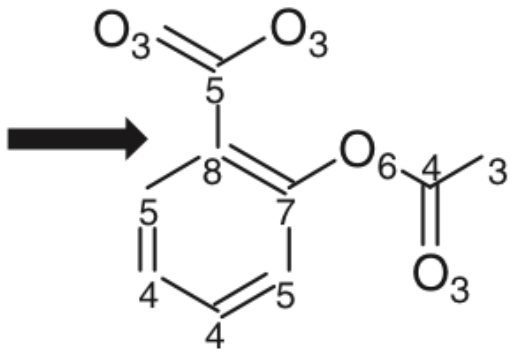


progesterone:
CC(=O)C1CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C

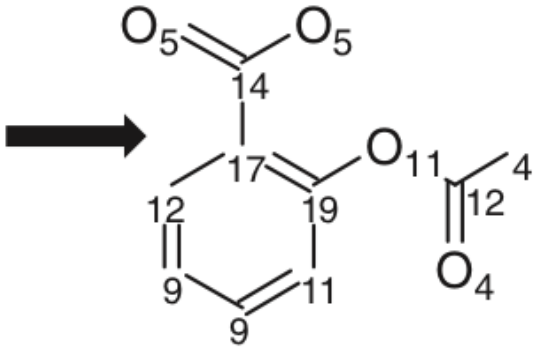
The Morgan algorithm (1/2)



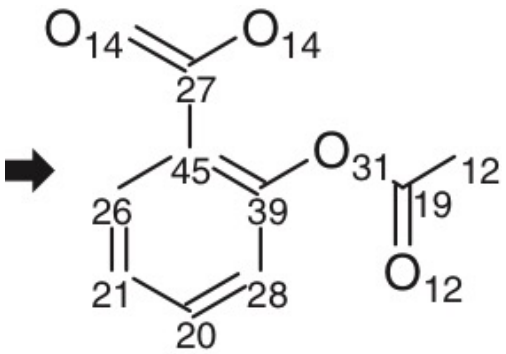
n=3



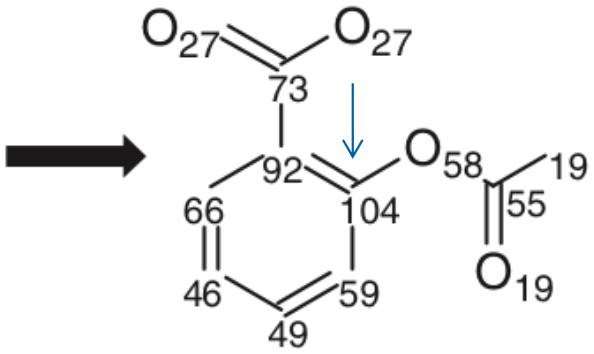
n=6



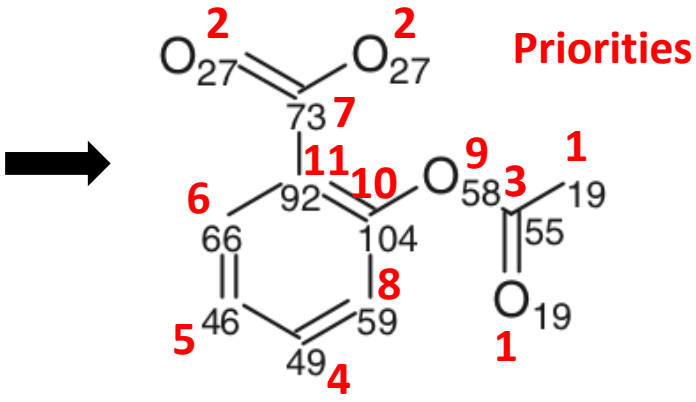
n=8



n=11



n=11



1 1 9 8 5 11 2 2

CC(=O)Oc1ccccc1C(O)=O

3 10 4 6 7

n... number of different connectivity values

The Morgan algorithm (2/2)

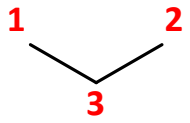
Morgan algorithm:

- Basic idea: Iterative calculation of "**connectivity values**" to enable differentiation of the atoms
- Initially, each atom is assigned a connectivity value **equal to the number of connected atoms**
- In the second and subsequent **iterations** a new connectivity value is calculated as the sum of the connectivity values of the neighbors
- The procedure continues until the number of different connectivity values reaches a maximum
- Most nodes have now **unique connectivity values**

- The atom with the lowest connectivity value (19 in this example) is then chosen as the first atom to assign priorities
- If a "tie" occurs (e.g. the two oxygens of the carboxyl group in the above example), then additional properties are considered such as atomic number and bond order
- Starting from the lowest priority value, we then walk through the molecule, generating the canonical SMILES notation

InChI (IUPAC International Chemical Identifier) I

- Unique chemical identifier developed by IUPAC that includes the full information on molecular structures
- InChI is designed for machine-readability; still human-readable but requires much more practice than SMILES representation
- Key advantages:
  - Layered structure allows the representation of structures at the desired level of detail
  - Allows the description of mobile hydrogens → tautomer-invariant description, meaning that most tautomers can be covered with a single InChI

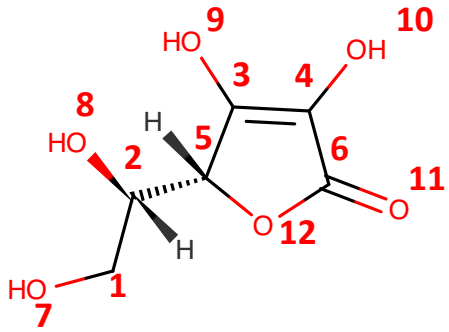


InChI=1S/C3H8/c1-3-2/h3H2,1-2H3

"1" indicating InChI version number; "S" for standard InChI

https://en.wikipedia.org/wiki/International_Chemical_Identifier
Heller et al., J Cheminf 2015, 7:23.

InChI: Example ascorbic acid



InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1

chemical formula | atom connections | hydrogen atoms | stereochemistry layer

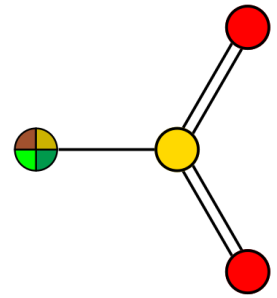
- Main layer:
  - Chemical formula
  - Atom connections: atom 7 connected to atom 1, atom 1 to atom 2, atom 2 to atoms 8 and 5…
- Hydrogen layer:
  - Atoms 2, 5 and 7 to 10 have a single H attached
  - Atom 1 has two H atoms attached
- Stereochemistry layer:
  - t indicates tetrahedral stereochemistry of atoms 2 and 5
  - m indicates that the selected molecule has exactly this configuration
  - s indicates type of stereochemistry information

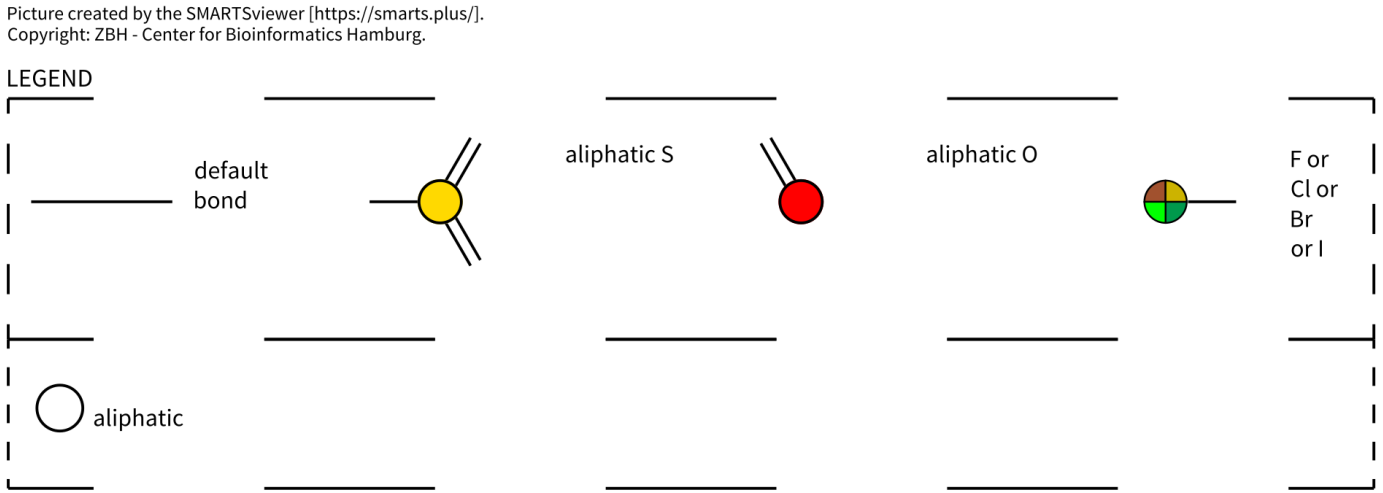
SMARTS

- “Regex for molecules”, built on the idea of SMILES representations
- Allows the use of logical expressions

- Example:
  Pattern describing the reactive group of
  sulfonyl and sulfuryl halogenides

S(=O)(=O)[F,Cl,Br,I]





SMARTS

- “Regex for molecules”, built on the idea of SMILES representations
- Allows the use of logical expressions

[C;!D4;!D1;!R;$(C(=O));$(C([O;D2])[#6,#7;!D1]!=[!D1])];!@[O;D2;$(O(C(=O))[#6;!D1][#6;!D1]);!$(OCO);!$(O[P,S])]

“An acyclic, non-aromatic carbon atom of an ester, adjacent to a non-terminal carbon or nitrogen without a double bond to a non-terminal atom, which is connected via an acyclic single bond to a 2-coordinate oxygen atom of the ester, adjacent to two other non-terminal carbon atoms, but not part of a carboxylate, or adjacent to sulfur or phosphorus.”

Symbol	Description
*	Any atom
$	Used to represent recursive SMARTS expressions
a	Aromatic atom
A	Aliphatic atom
D<n>	Atom with <n> explicit bonds (commonly: bonds to non-H atoms)
X<n>	Atom with a total of <n> bonds
v<n>	Atom with bond order <n>
H<n>	Atom with <n> neighboring H atoms
h<n>	Atom with <n> neighboring implicit H atoms
R<n>	In <n> rings of a SSSR (Smallest Set of Smallest Rings)
r<n>	In the smallest ring of a SSSR with atoms <n>
-<n>	Negative formal charge <n>
+<n>	Positive formal charge <n>
#<n>	Atom with atomic number <n>
@	Atom with local chirality counterclockwise
@@	Atom with local chirality clockwise
<n>	Atom with atomic mass <n>

universität wien

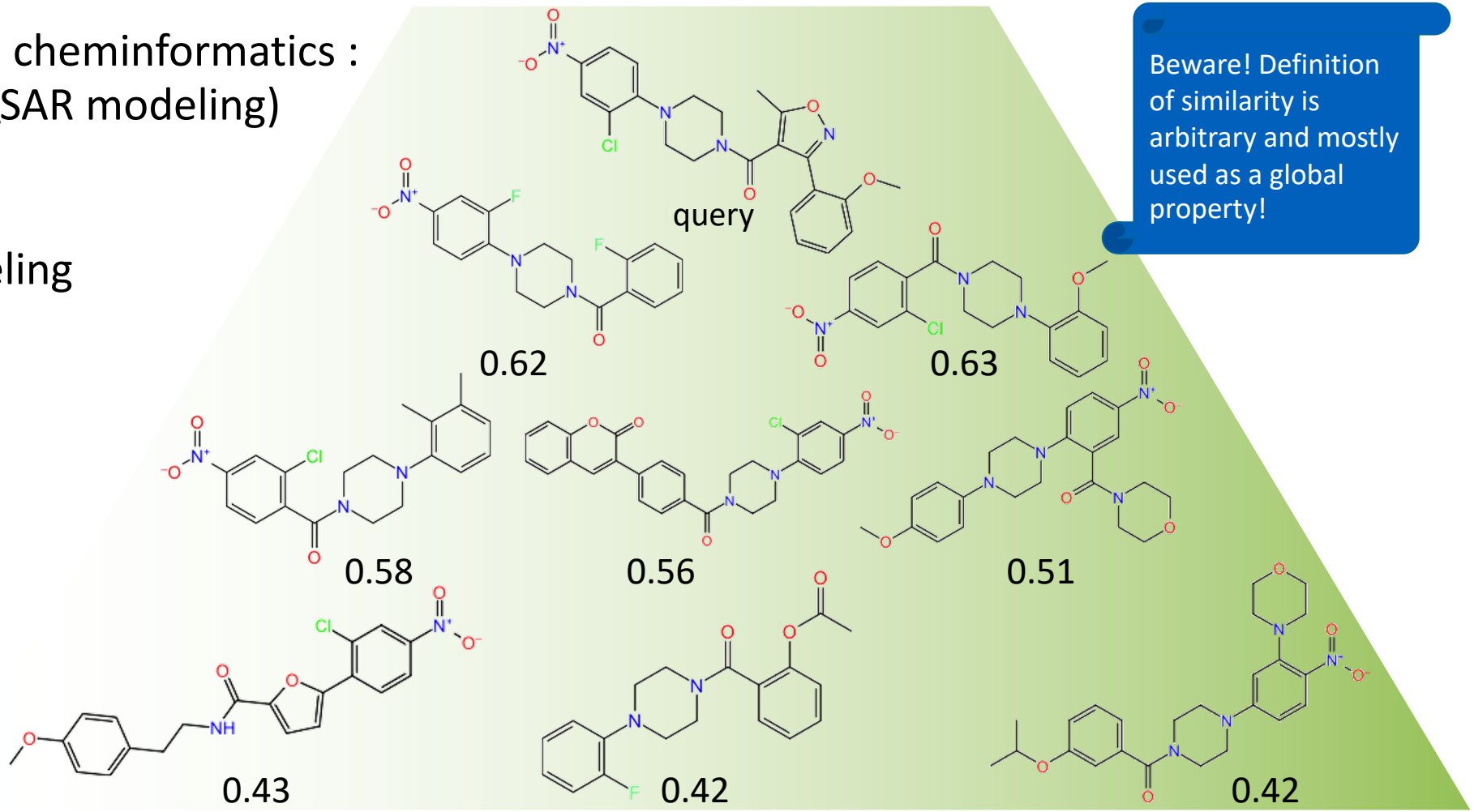
Molecular similarity – molecular diversity

Johannes Kirchmair

The molecular similarity principle:

Compounds which are structurally similar are likely to have similar physicochemical and biological properties

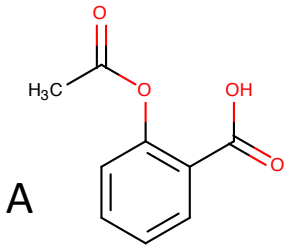
- Widely applied concept in cheminformatics :
  - Bioactivity prediction (QSAR modeling)
  - ADME prediction
  - Toxicity prediction:
    read-across, QSTR modeling
  - Virtual screening
  - Hit expansion
  - Target prediction
  - …

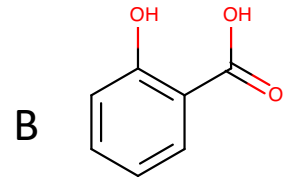


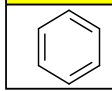
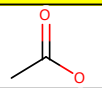
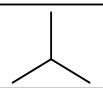
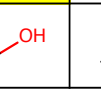
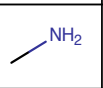
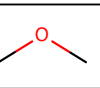
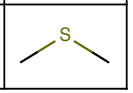
Beware! Definition of similarity is arbitrary and mostly used as a global property!

Similarity coefficients: Tanimoto coefficient

- [0 – 1], where Tanimoto = 1 indicates identical molecules
- **Tanimoto coefficient is the ratio between the number of common bits and the number of bits set (i.e. nonzero) in either sample**
- Tanimoto coefficient is size-dependent:
  - Molecule with few features all of which are shared with a molecule with many features is not evaluated similar
  - Larger molecules tend to have higher Tanimoto coefficient
- Tanimoto coefficient is descriptor-dependent: A Tanimoto of 0.4 may mean two molecules are completely unrelated by MACCS keys or similar by circular fingerprints





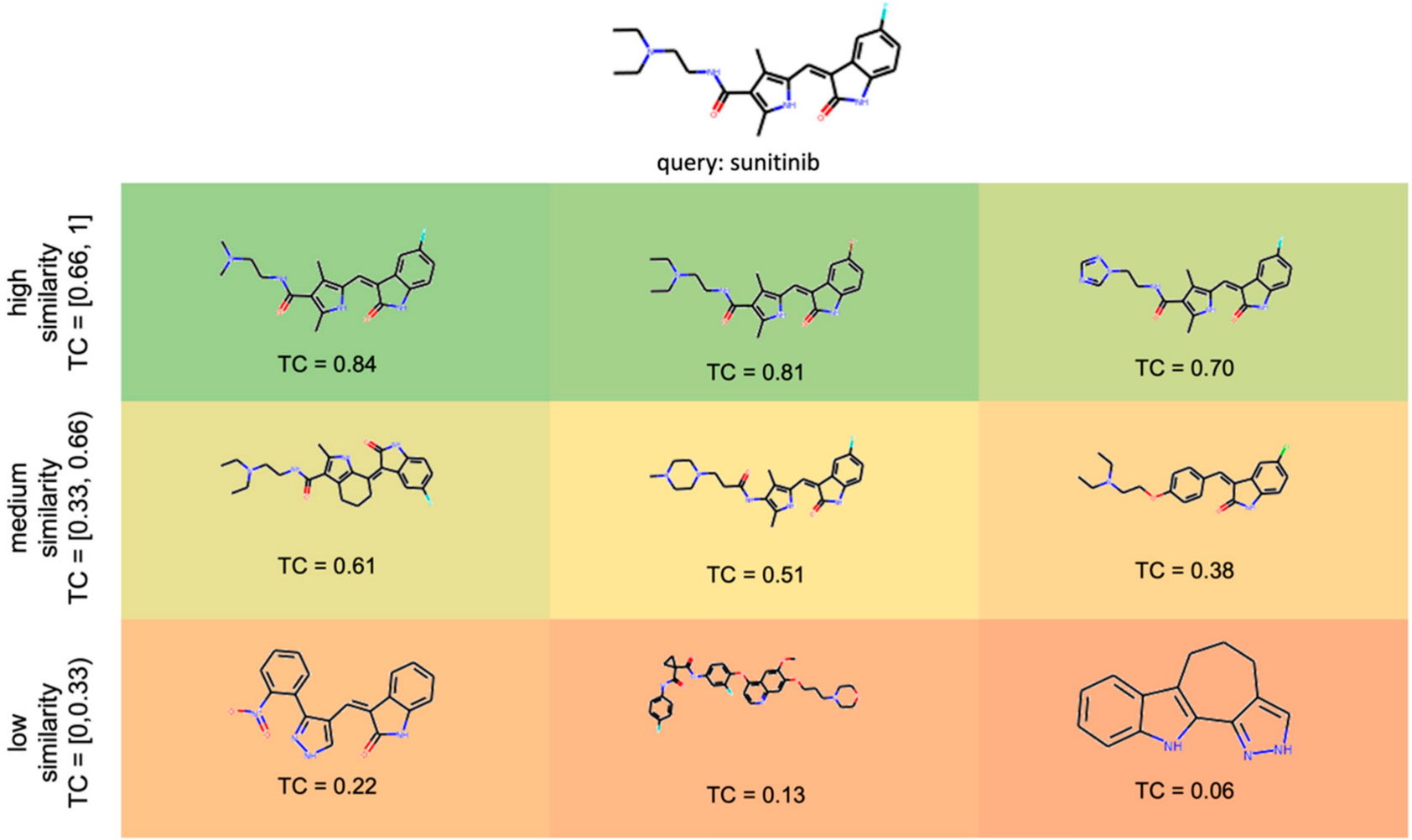
A	1	1	0	1	0	1	0
B	1	1	0	1	0	0	0
							

$$Tanimoto = \frac{c}{a+b-c} = \frac{3}{4+3-3} = 0.75$$

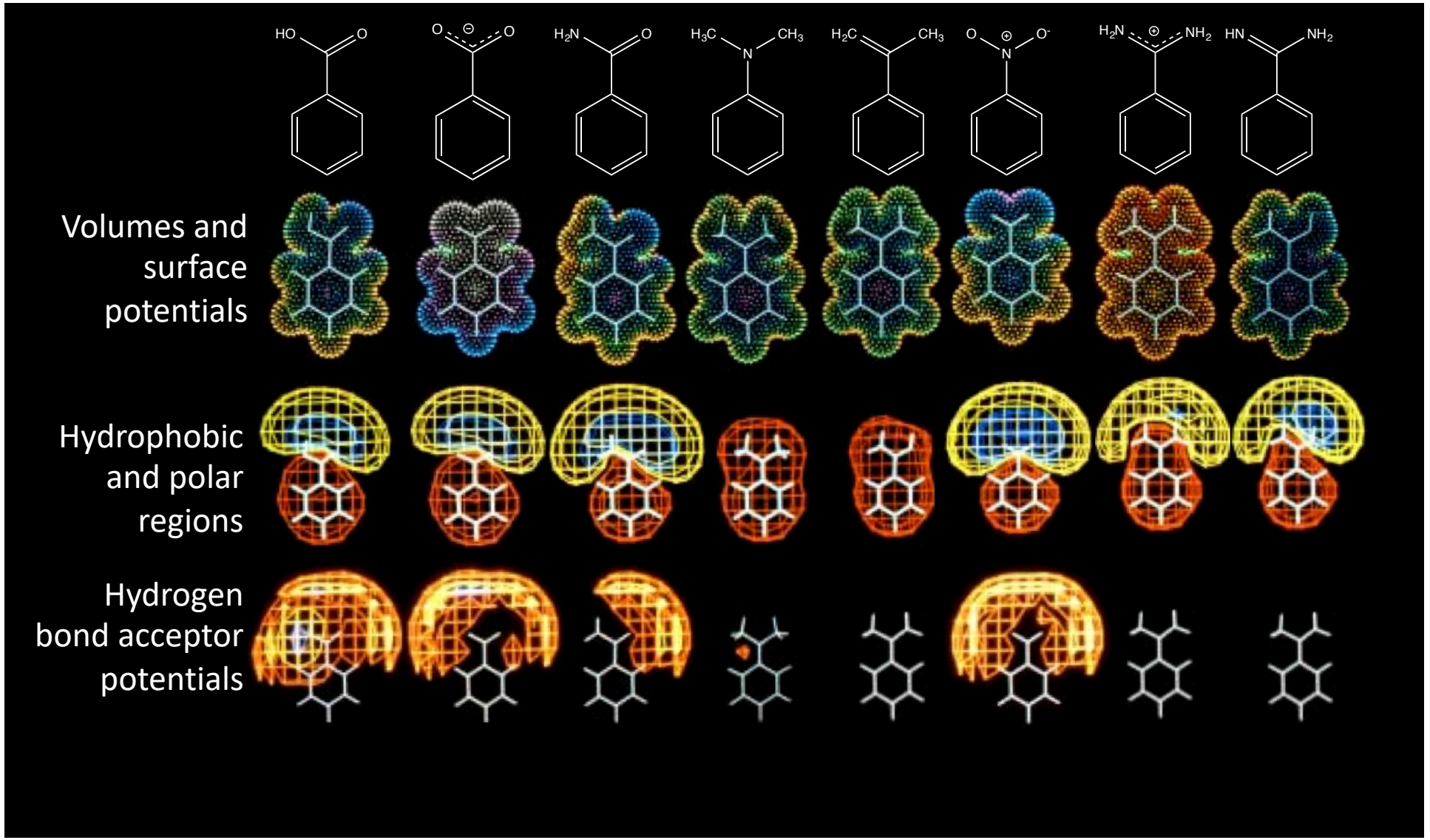
a... Number of bits on in (A)
b... Number of bits on in (B)
c... Number of bits on in (A) AND (B)

Leach & Gillet, An Introduction to Chemoinformatics, 2007, Springer.
Jain, A. In "Chemoinformatics for Drug Discovery", Wiley 2014.

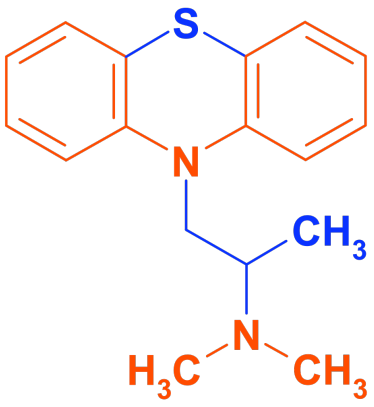
Levels of molecular similarity (visualization)

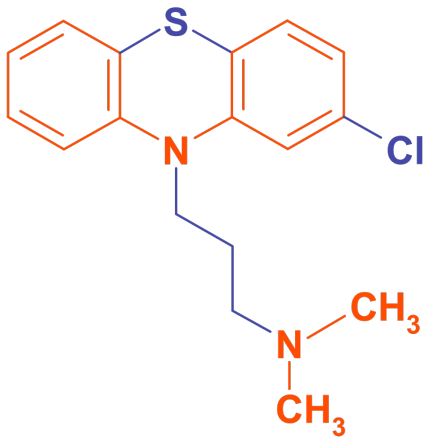


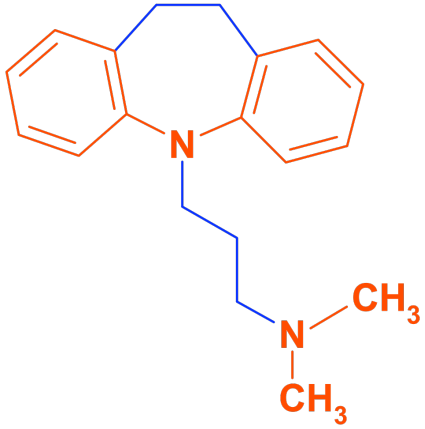
Definition of similarity is arbitrary and mostly used as a global property



Structurally related drugs - distinct pharmacological effects







Promethazine	Chlorpromazine	Imipramine
H_1 antagonist	D_2-antagonist	5-HT & many others
1st-generation antihistamine	neuroleptic	antidepressant

Activity landscapes can be rugged!

continuous SAR

discontinuous SAR

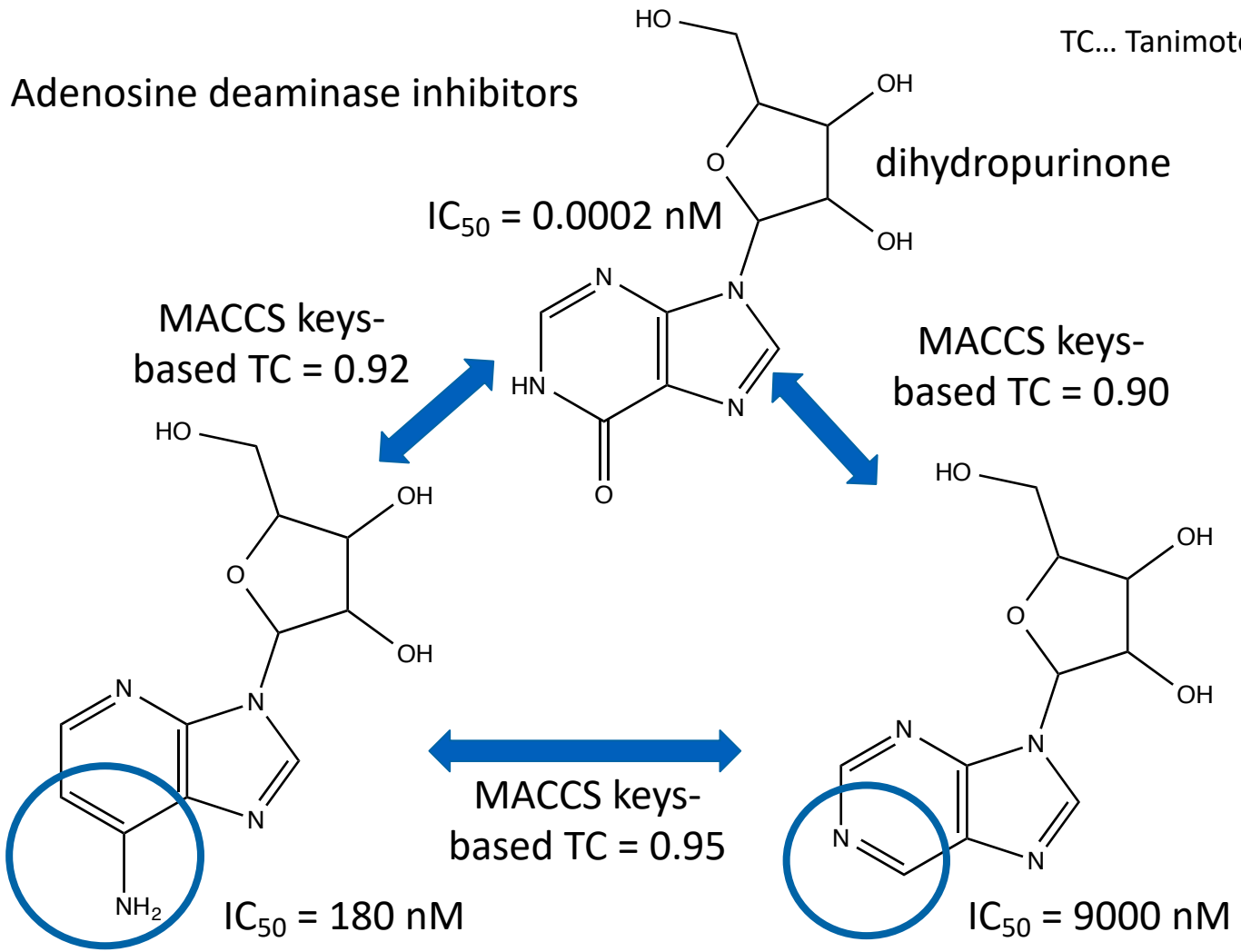
- Activity cliffs: Small structural changes can lead to substantial changes in bioactivity → **non-linearity of structure-activity relationships**
- Molecules may be binding at different locations within the binding site

http://www.wallpaperhere.com/California_Poppies_and_Rolling_Hills_24369/download_1600x1200
http://www.visitsouthernutah.com/Bryce-Canyon-National-Park

Discontinuous SARs: Activity cliffs

TC… Tanimoto coefficient

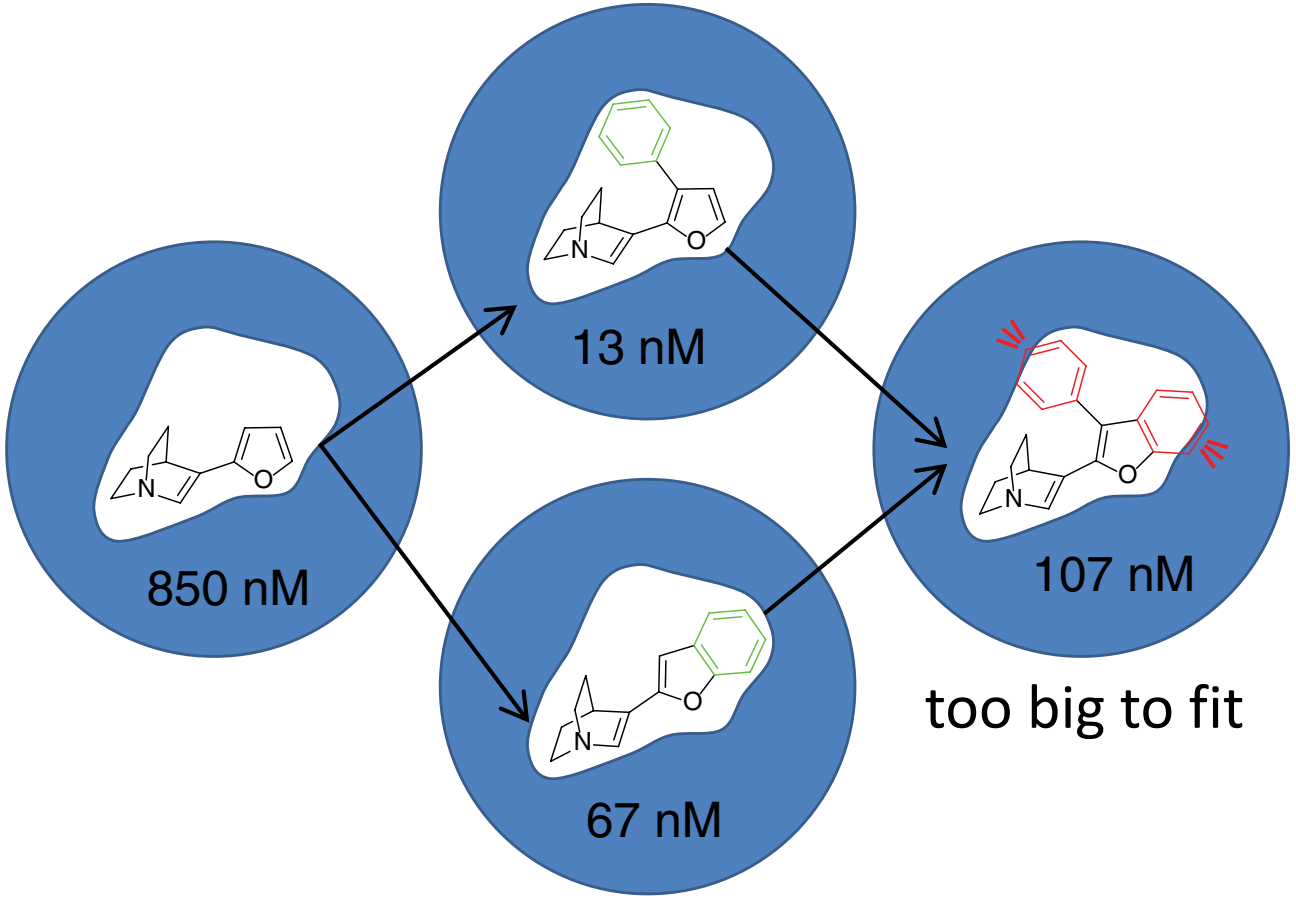
Example: Adenosine deaminase inhibitors



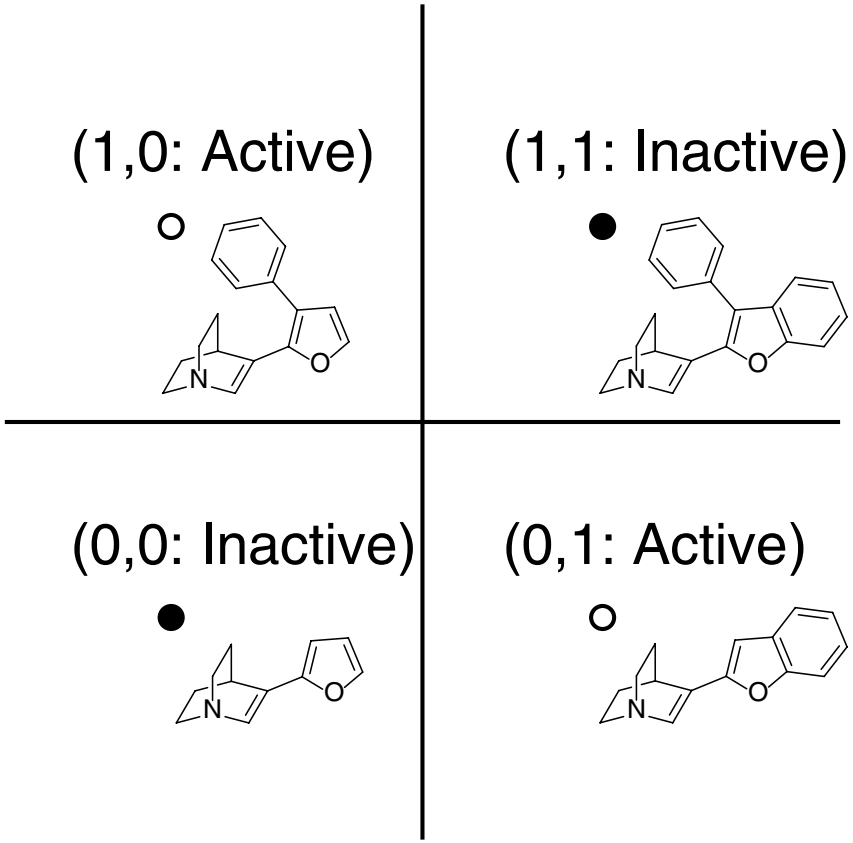
Non-additivity in SAR

Example of an exclusive-OR problem: Four muscarinic antagonists, all based on the quinuclidinene-furan scaffold, exhibit a pattern of potency variation that is highly non-additive

Nonadditivity in SAR

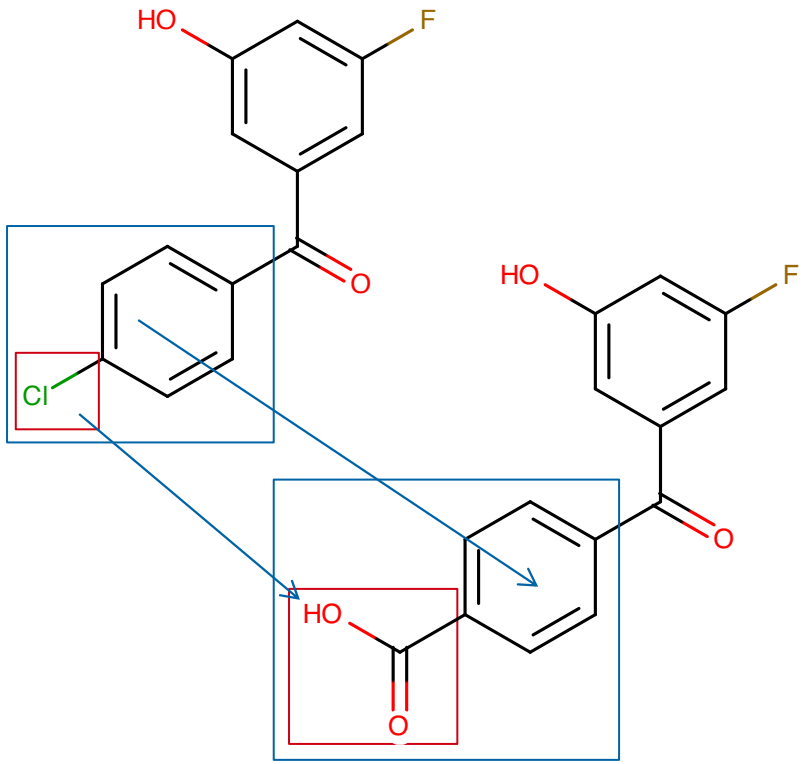


The XOR problem



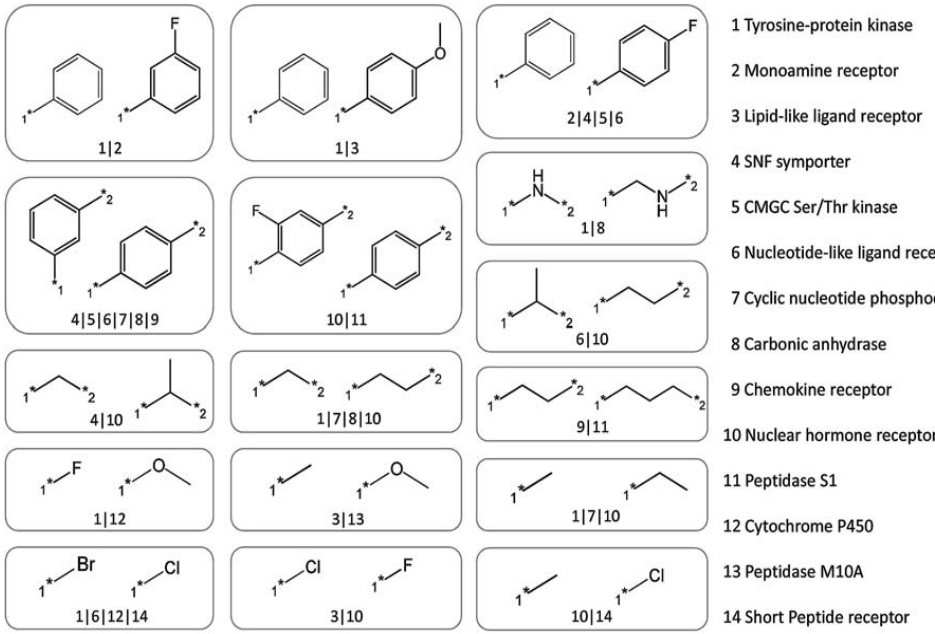
Matched molecular pairs (MMPs)

- Substantial changes in activity are most valuable information, in particular if they can be attributed to a specific fragment
- **MMPs are pairs of compounds that differ only at a single localized site and are distinguished by a defined substituent or molecular fragment**
- Often trends can be derived from analysing a series of MMPs based on the same transformation
- MMPs are straightforward to understand from a medicinal chemistry perspective
- Results are context-specific and the context (fragment size) can be chosen arbitrarily

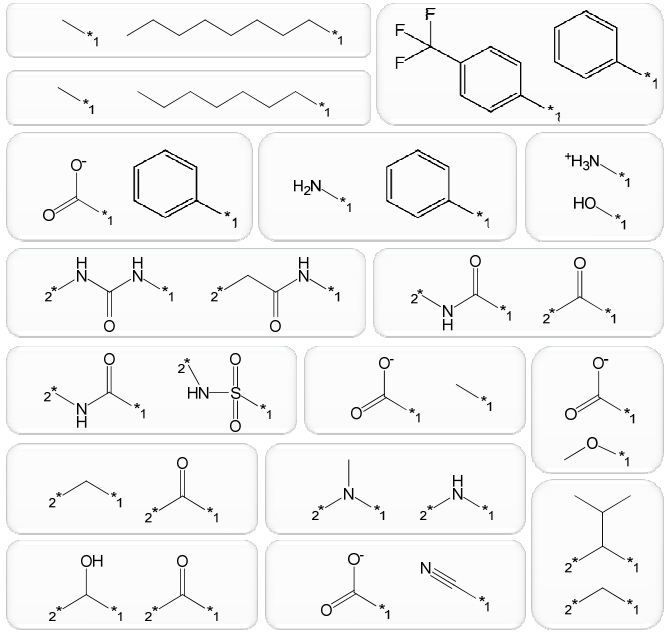


Applications of MMPs: Example bioisosteric replacement

- Bioisosteres are groups or molecules which have chemical and physical similarities producing broadly similar biological effects
- Why search for bioisosteres?
  - Scaffold-hopping
  - Side-chain enumeration
  - Patent protection by hit expansion
  - Patent breaking
  - Property manipulation



Bioisosteres identified through the analysis of a large number of target protein families

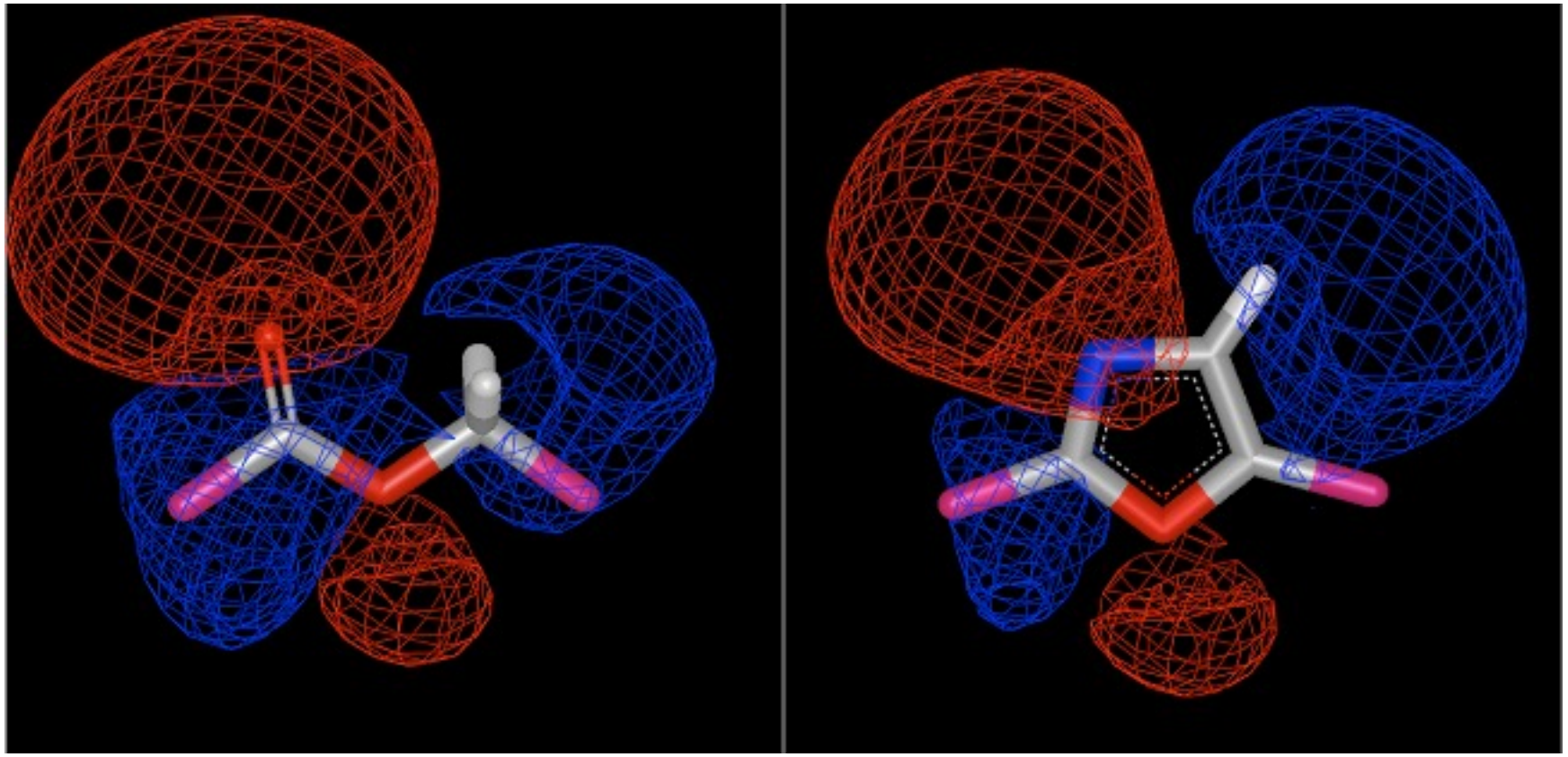


Representative subset of structural transformations that frequently introduce activity cliffs

Wassermann A.M. and Bajorath J., Med Chem Commun 2011, 2, 601-606.
Wassermann A.M., PhD thesis, University of Bonn, 2012.

BROOD: fragment replacement

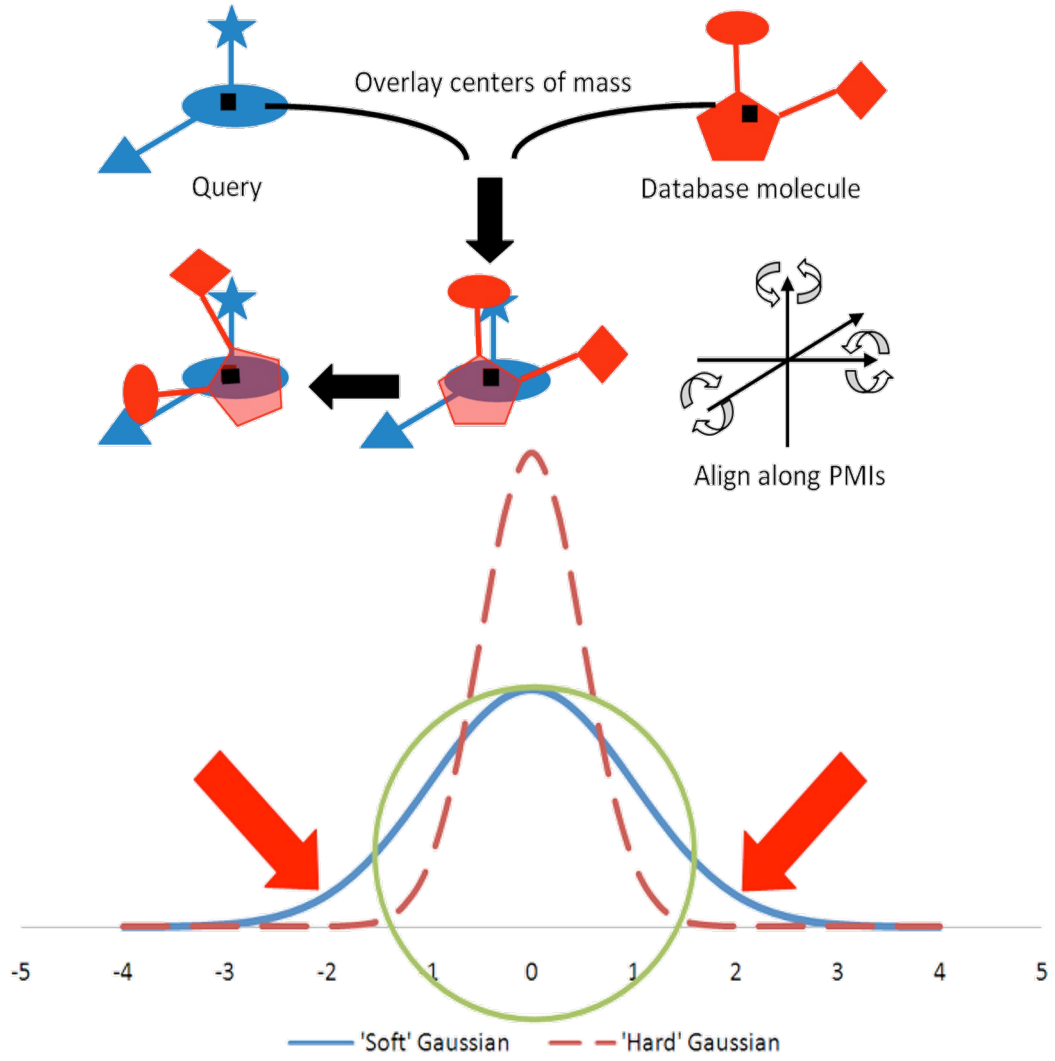
- BROOD generates analogs of the lead by replacing selected fragments in the molecule with fragments that have similar shape and electrostatics, yet with selectively modified molecular properties



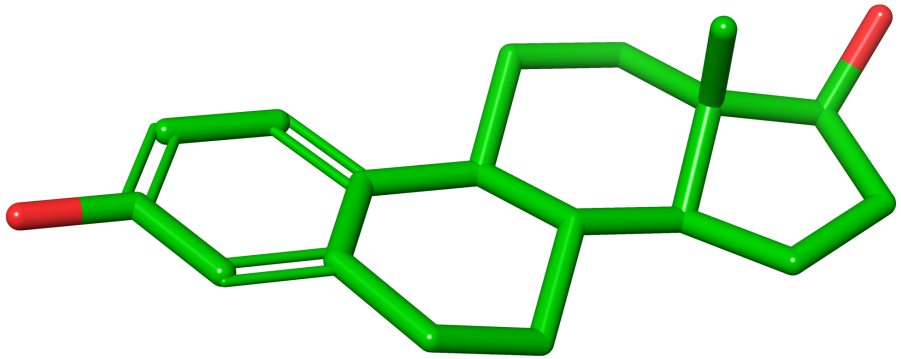
Comparison of an ester fragment and an oxazole fragment showing the electrostatic isopotential contour surfaces. The electrostatic Tanimoto coefficient between the two fragments is 0.54.

Similarity-based approaches: Shape-focused methods

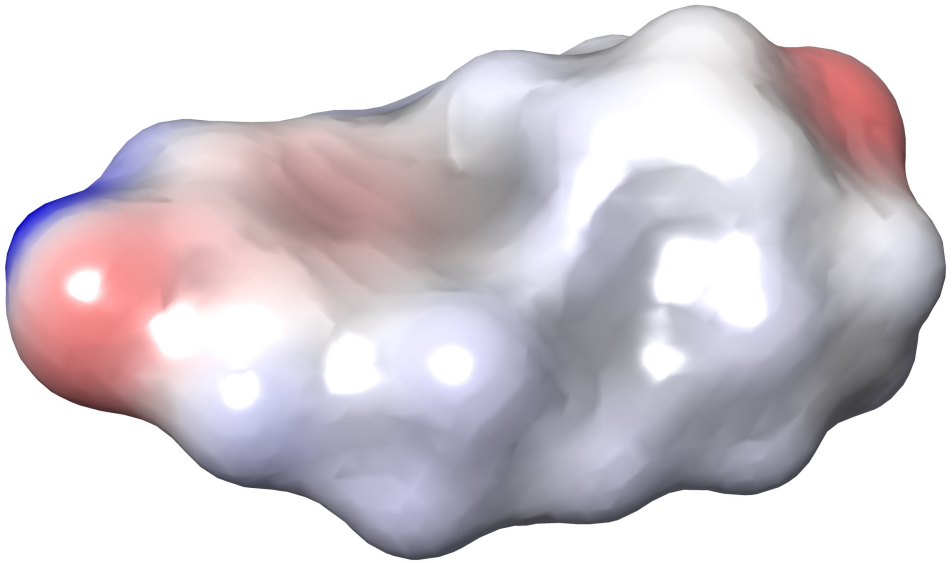
- Representation of a molecules' individual atoms by spherical Gaussians:
  - Shape represented by soft (fuzzy) Gaussians
  - Chemical features (e.g. H-bond donor) represented by hard Gaussians
- Molecules aligned by rigid body optimization, maximizing the overlap of volumes between them:
- Determination of the center of mass, then rotation along the principal moments of inertia
- Chemical features used to "snap in" alignment (→ accurate superposition of hydrogen bonds)
- Hydrogens are ignored
- Can typically analyze 1000 conformers per second
- Companies today screen up to 10^{12} molecules

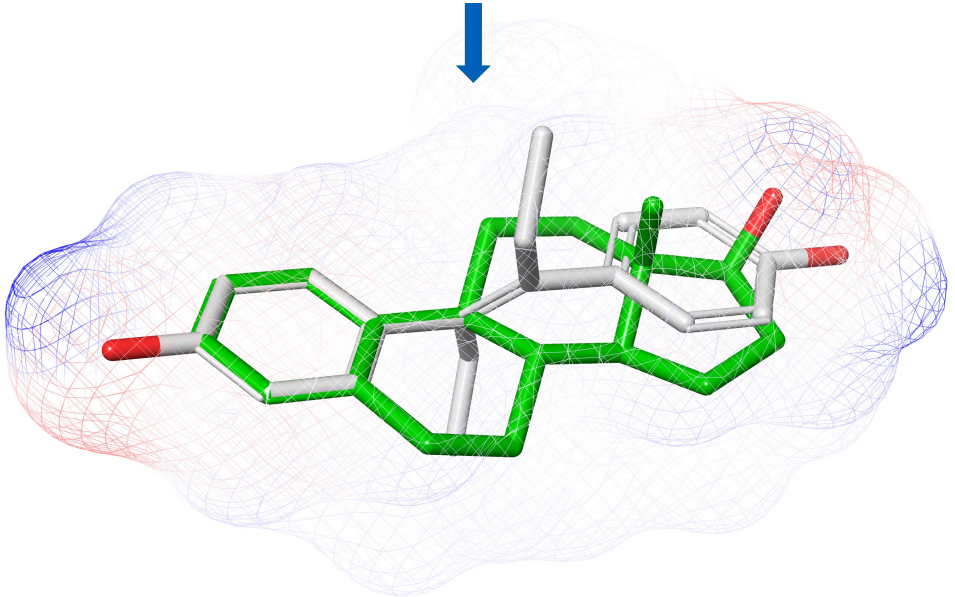


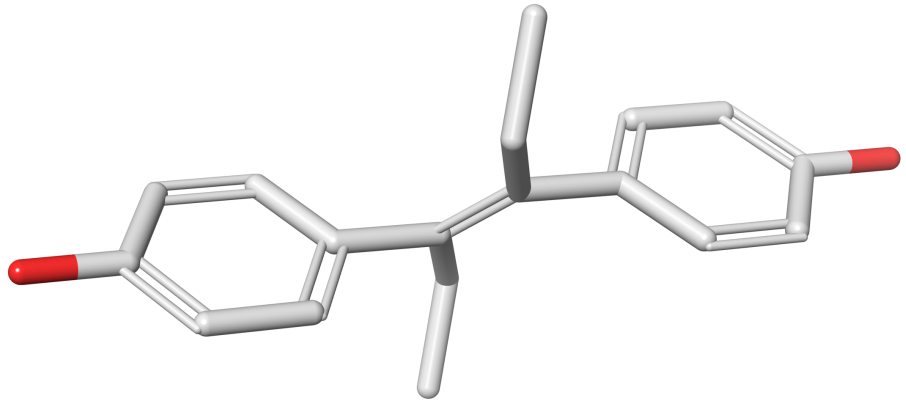
From bioisosteric replacement to “scaffold hopping”



Estrogen







Diethylsilbestrol

universität wien

…an exciting journey lies ahead of you…. …enjoy the trip!

Thanks for your attention!