VO Molecular Modelling
WS 2024/25
(Part 2)

Institut für Allgemeine, Anorganische und Theoretische Chemie

Christofer Tautermann
Valentin Egger-Hörschinger
Alesia Yakimchyk

Lecture contents

Overview

Chemoinformatics

1) Introduction to chemoinformatics
2) Representation of molecules
3) Molecular descriptors
4) Molecular similarity and diversity
5) Machine Learning and computational models
6) Chemical spaces
7) Virtual screening and library design

Molecular descriptors

Any molecular feature beyond the chemical structure

Molecular descriptors

- 0D/1D/2D/3D/4D descriptors
  - Definition and examples
    - Topological indices
    - Fingerprints
    - Common descriptors
- Data analysis
  - Pincipal component analysis

Descriptors – or how to compare things

Motivation

Property	hedgehog	hare
class	mammal	mammal
legs	4	4
eyes	2	2
diet	omnivorous	herbivores
spiny	yes	no
length of ears	1 cm	30 cm
max speed	19 km/h	64 km/h

similar?

rather not similar

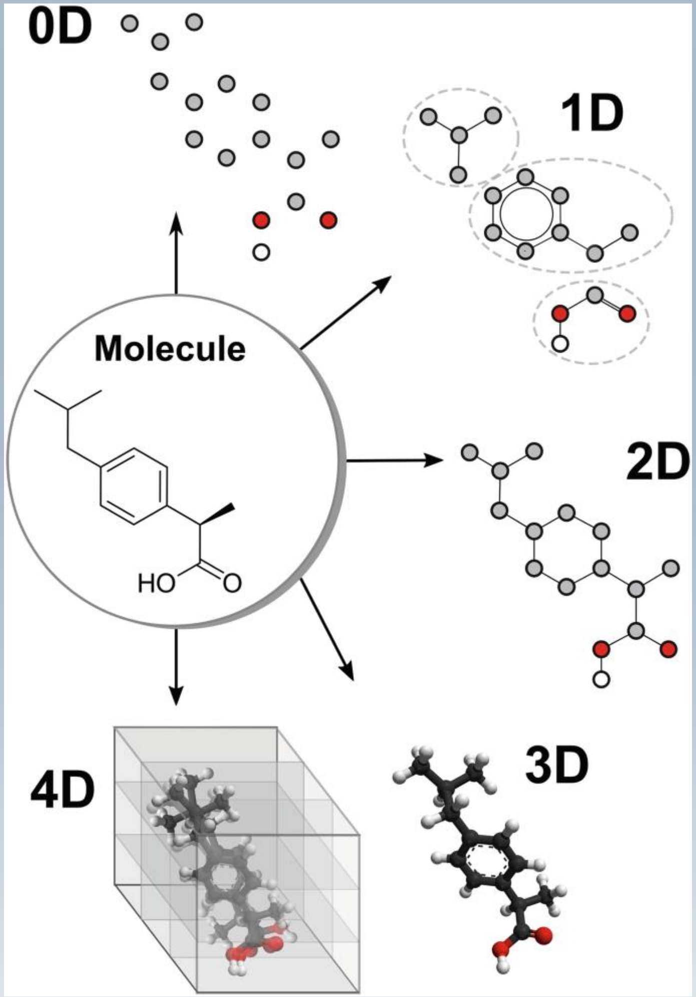
Molecular descriptors

- *"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a **useful** number or the result of some standardized experiment"*
  (Handbook of Molecular Descriptors, Wiley-VCH, 2000)
- Basically the structure of a molecule is transformed into a number which is either
  - An experimental measurement (or a calculated approximation thereof), e.g., logP, molecular weight, dipole moment, ... or
  - A theoretical descriptor which is derived from either the sum formula, the 2D-representation or the 3D representation of the molecule (or more complex), e.g., fingerprints, surface area, volume, QM-descriptors, ...
- Descriptors are numerical values used to characterize molecules
  (Similar as comparing animals – this is usually not done on a genome level, but based on their diet, habitat, physical description, ...)

→ Molecules are assumed to be quite similar if their descriptors are similar

Molecular descriptors
Different dimensionalities

- 0D descriptors: plain counts, such as number of carbons, molecular weight, ...
- 1D descriptors: substructure counts, e.g., how many rings are in the structure? heteroatoms? some kinds of fingerprints (MACCS keys), sp^3 carbons, ...
- 2D descriptors: graph invariants or graph properties. Dependence on the atom connectivity
- 3D descriptors: dependent on the molecule conformation, e.g. QM descriptors
- 4D descriptors: based on conformational molecular ensemble



Impact of Molecular Descriptors on Computational Models | SpringerLink

Molecular descriptors
Different dimensionalities – more examples

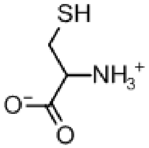
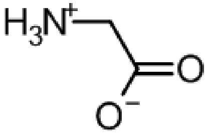
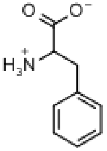
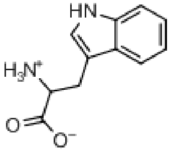
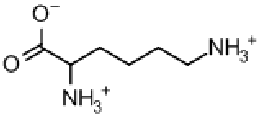
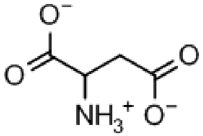
Dim		Examples
0D	Atom counts, bond counts, molecular weight, sum of atomic properties	Molecular weight; number of: atoms, hydrogen atoms, carbon atoms, heteroatoms, non hydrogen atoms, bonds, multiple bonds, double bonds, ...
1D	Counts of atom types Fragment counts	Number of: primary C (sp3), secondary C (sp3), tertiary C (sp3), quaternary C (sp3), secondary C (sp3) in a ring, tertiary C (sp3) in a ring, number of H bond donor atoms, H bond acceptor atoms, Number of rings, 3 membered rings, 4 membered rings, 5 membered rings, 6 membered rings, 7 membered rings, presence of amides (aliphatic/aromatic; primary, secondary, tertiary), amines (aliphatic/aromatic; primary, secondary, tertiary), ammonium, groups, carbamates, hydrazines, ...
2D	Topological descriptors	Zagreb index, Wiener index, connectivity indices chi, kappa shape indices, molecular walk counts, lipophilicity (log *P*), topological polar surface area extended connectivity fingerprints (circular fingerprints), ISIDA fragments, state topological parameter, BCUT descriptors, 2D autocorrelation vector, ...
3D	Geometrical descriptors	Molecular eccentricity, radius of gyration, dipole moment, polar surface area, radial distribution function, 3D autocorrelation vector, HOMO, LUMO
	3D surface properties	Molecular electrostatic potential, hydrophobicity potential, hydrogen bonding potential
	3D grid properties	Comparative molecular field analysis (CoMFA), comparative molecular similarity Indices analysis (CoMSIA)
4D		3D coordinates sampling of conformations

Molecular descriptors
Simple 0D/1D descriptors

- Counts
  - hydrogen bond donors and acceptors
  - rotatable bonds
  - ring systems
  - substructure counts
  - formal charge
  - ...
- Other basic properties
  - Molecular weight
  - fraction sp^3 atoms
  - ...

quickly computed

Simple 1D/2D descriptors
Examples: amino acids

mol	name	a_acc	a_don	a_heavy	a_nCsp3	FCsp3	b_1rotN	FCharge	logP (o/w)
	Cys	0	0	7	2	0,67	2	0	-0,43
	Gly	0	0	5	1	0,50	1	0	-1,00
	Phe	0	0	12	2	0,22	3	0	1,00
	Trp	0	1	15	2	0,18	3	0	1,36
	Lys	0	0	10	5	0,83	5	1	-0,47
	Asp	0	0	9	2	0,50	3	-1	-1,17

Molecular descriptors
Important 2D descriptors – (c)logP

Hydrophobicity/(c)logP

- logP is the partition coefficient of a molecule between water and a non-polar solvent (usually octanol)

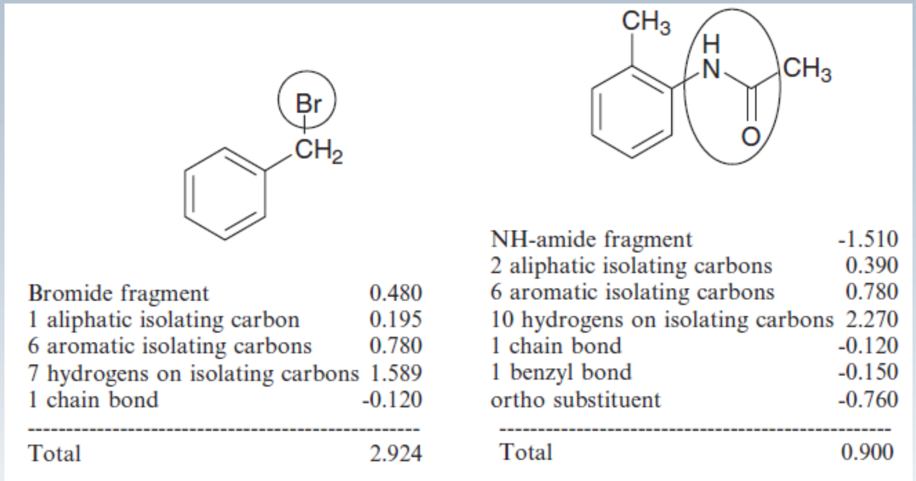
$$logP_{oct/water} = log\left(\frac{[solute]_{octanol}}{[solute]_{water}^{non-ionized}}\right)$$

- Experimentally accessible (companies with > 100k values)
- High importance in drug discovery – correlation to
  - solubility of molecules
  - permeability
  - metabolic stability
  - plasma protein binding
  - ... and often affinity to a target and antitarget ...
- Prediction for new compounds: calculated logP = clogP

Molecular descriptors
Hydrophobicity – calculating clogP

Calculation of clogP:

- additive schemes – fragment or atom based
  - Starting with a measured logP and correct the value for a new substituent
  - substituents assumed to have the same correction value over different series → not true, but applied quite often anyhow
  - primarily used in congeneric series
- other fragmentation schemes
  - by isolating carbons (i.e., carbons not doubly/triply bound to a heteroatom)
  - logP of fragments measured or estimated



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

Molecular descriptors
Topological indices: Wiener Index

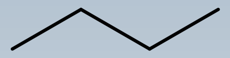
Topological indices are calculated from the 2D graph representation of molecules

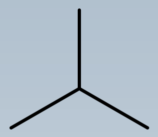
Depend on the size, shape, branching

- Wiener index (oldest topological descriptor, 1947) – also known as „distance of a graph“:

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

where D_{ij} is number of bonds between atom i and atom j





W = 3*1 + 2*2 + 1*3 =10   W = 3*1 + 3*2 + 0*3 = 9

- applied to boiling point properties of alkanes

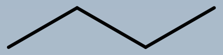
compound	Wiener Index
n-hexane	
2-methylpentane	
3-methylpentane	
2,3-dimethylbutane	

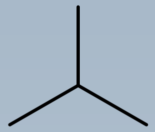
Molecular descriptors
Topological indices: Branching index

- Branching index

  based on the degree δ_i of an atom i = number of the adjacent non-H atoms.

  Branching index $= \sum_{bonds} \frac{1}{\sqrt{\delta_i \delta_j}}$ for all bonds directly connecting atoms i and j

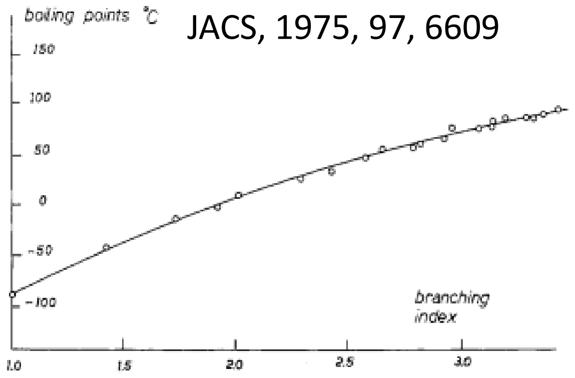




$$B = \frac{1}{\sqrt{1*2}} + \frac{1}{\sqrt{2*2}} + \frac{1}{\sqrt{1*2}} \approx 1.9$$

$$B = \frac{1}{\sqrt{1*3}} + \frac{1}{\sqrt{1*3}} + \frac{1}{\sqrt{1*3}} \approx 1.7$$

- Applied to alkanes



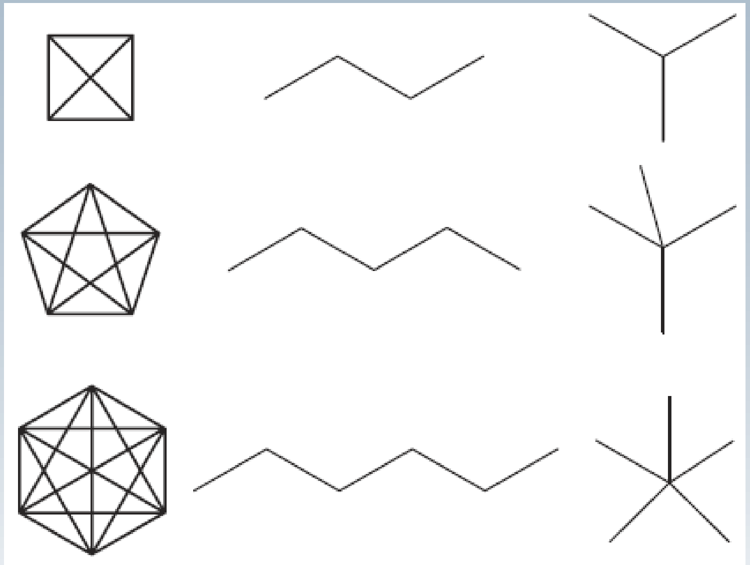
JACS, 1975, 97, 6609

Figure 1. Boiling points of alkane isomers having from two to seven carbon atoms plotted against the topological branching index. (Experimental data are taken from ref 24.)

compound	Branching Index
n-hexane	
2-methylpentane	
3-methylpentane	
2,3-dimethylbutane	

Molecular descriptors
Topological indices: further developments

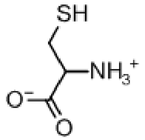
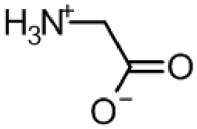
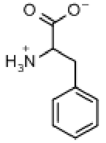
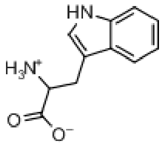
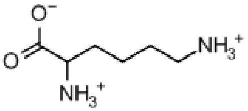
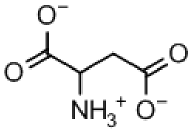
- Further developments of the branching indices
  - inclusion of valence electrons, lone pairs, bound H-atoms in δ_i
  - summation of paths over different lengths → chi indices (${}^{0}\chi$ for summation over atoms, ${}^{1}\chi$ over bonds (=branching index), ${}^{2}\chi$ over paths of length 2, ...)
- Kappa shape indices
  - compare molecules to „extreme“ shapes: count paths of lengths i for ${}^{i}\kappa$ (i=1-3)



$$ {}^{i}\kappa = 2\frac{{}^{i}P_{max}\ \ {}^{i}P_{min}}{(\ {}^{i}P_{molecule})^2} $$

$$ {}^{1}\kappa = \frac{\#atoms(\#atoms-1)^2}{(\#bonds)^2} $$

Topological indices
Example: amino acids

	name	1kappa	2kappa	Wiener Path	chi0	chi1
	Cys	7,00	3,06	46	5,86	3,18
	Gly	5,00	2,25	18	4,28	2,27
	Phe	10,08	4,89	212	8,97	5,70
	Trp	11,48	4,89	369	10,84	7,18
	Lys	10,00	5,76	143	7,98	4,68
	Asp	9,00	3,92	96	7,44	4,04

$$\chi_0 = \frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \frac{1}{\sqrt{1}} + \frac{1}{\sqrt{1}}$$

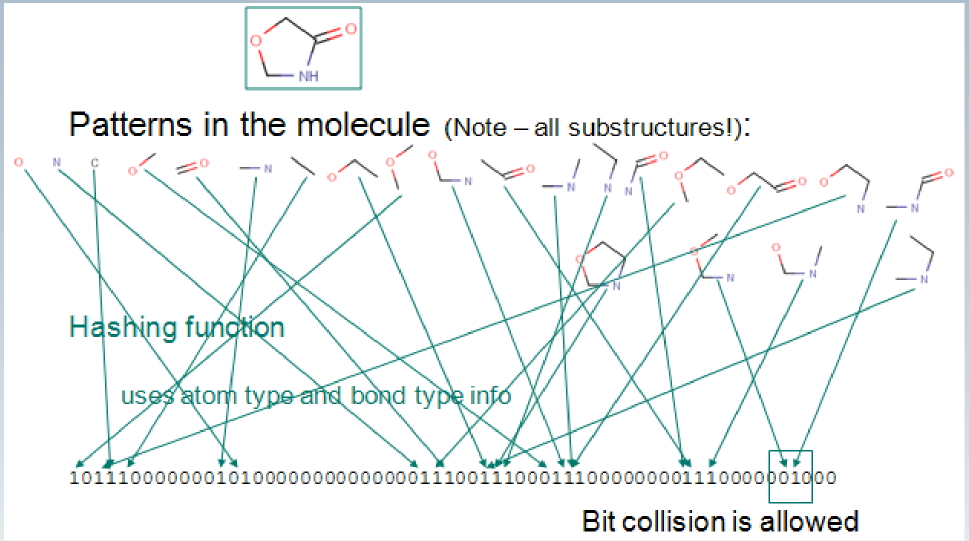
$$\chi_1 = \frac{1}{\sqrt{1 * 2}} + \frac{1}{\sqrt{2 * 3}} + \frac{1}{\sqrt{1 * 3}} + \frac{1}{\sqrt{1 * 3}}$$

$$W = (1 + 2 + 3 + 3) + (1 + 2 + 2) + (1 + 1) + 2$$

$$^{1}\kappa = \frac{5 * 4^2}{4^2}$$

2D fingerprints
Bitstrings

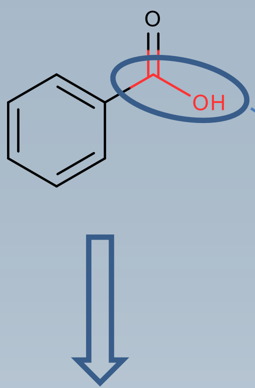
- See Chapter 2 – used and originally developed for *database pre-screening*
  - dictionary based approaches: presence of substructures
  - fingerprints: generate *all* substructures (up to a certain path-length) and apply hashing procedure
    → *not a priori clear that hashed keys/fingerprints should work as descriptors*
- Example: Substructure fingerprints (as implemented by Chemaxon)
  all paths of length up to n atoms annotated (incl. branching and cycles)
  hashing procedure applied



http://gdbtools.unibe.ch:8080/PPB/

2D fingerprints
Substructure keys – MACCS keys

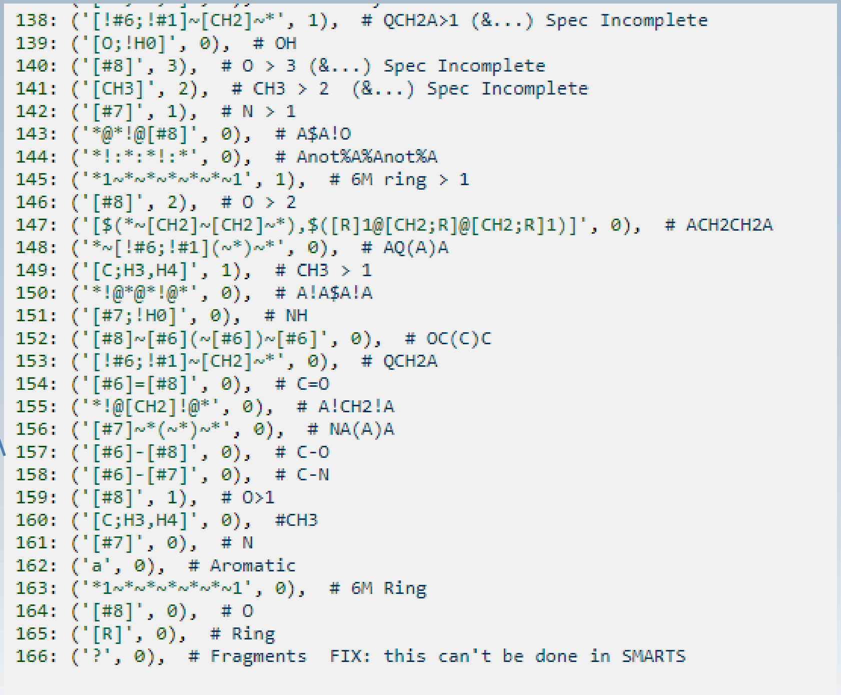
- MACCS: Prominent example for structural keys (166 bit-long) in which each bit is associated with a specific structural pattern.



"011110010100100000000
0000001000000000000000
1000000000000000000000
0000000000000000000000
0000000000000000000000
0000000000000000000000
0000000000000000000000
00000000000000“

Position	Bit	
166	0	
165	1	Ring
164	1	O
163	1	6-memb Ring
162	1	aromatic
161	0	N
160	0	CH3
159	1	O>1
158	0	C-N
157	1	C-O
156	0	
155	0	
154	1	C=O
153	0	
152	0	
151	0	
150	0	
149	0	
148	0	
147	0	
146	0	
145	0	
144	0	
143	0	
142	0	
141	0	
140	0	
139	1	OH
138	0	

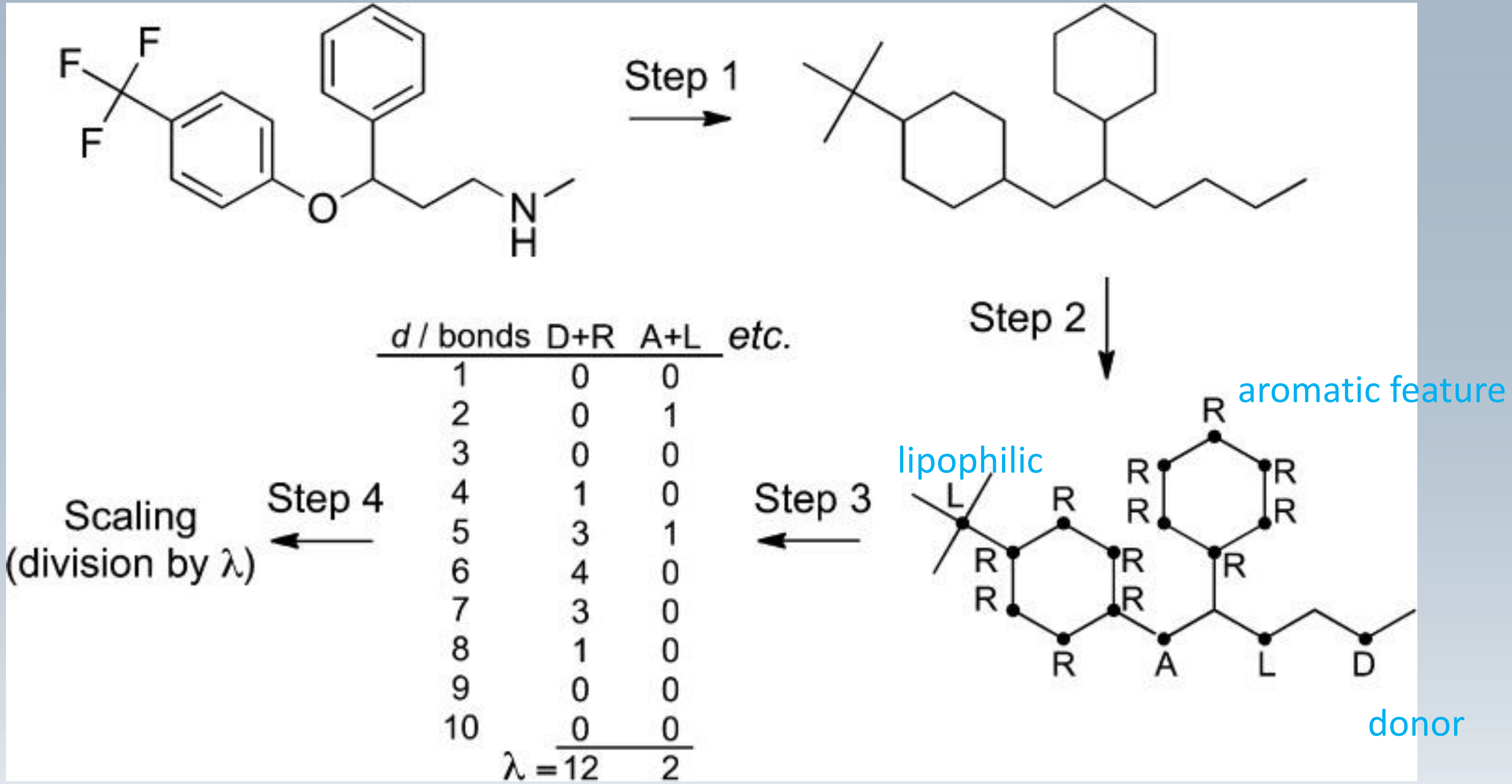
Dictionary (part):



```
138: ('[!#6;!#1]~[CH2]~*', 1),  # QCH2A>1 (&...) Spec Incomplete
139: ('[O;!H0]', 0),  # OH
140: ('[#8]', 3),  # O > 3 (&...) Spec Incomplete
141: ('[CH3]', 2),  # CH3 > 2  (&...) Spec Incomplete
142: ('[#7]', 1),  # N > 1
143: ('*@*!@[#8]', 0),  # A$A!O
144: ('*!:*:*!:*', 0),  # Anot%A%Anot%A
145: ('*1~*~*~*~*~*~1', 1),  # 6M ring > 1
146: ('[#8]', 2),  # O > 2
147: ('[$(*~[CH2]~[CH2]~*),$([R]1@[CH2;R]@[CH2;R]1)]', 0),  # ACH2CH2A
148: ('*~[!#6;!#1](~*)~*', 0),  # AQ(A)A
149: ('[C;H3,H4]', 1),  # CH3 > 1
150: ('*!@*@*!@*', 0),  # A!A$A!A
151: ('[#7;!H0]', 0),  # NH
152: ('[#8]~[#6](~[#6])~[#6]', 0),  # OC(C)C
153: ('[!#6;!#1]~[CH2]~*', 0),  # QCH2A
154: ('[#6]=[#8]', 0),  # C=O
155: ('*!@[CH2]!@*', 0),  # A!CH2!A
156: ('[#7]~*(~*)~*', 0),  # NA(A)A
157: ('[#6]-[#8]', 0),  # C-O
158: ('[#6]-[#7]', 0),  # C-N
159: ('[#8]', 1),  # O>1
160: ('[C;H3,H4]', 0),  #CH3
161: ('[#7]', 0),  # N
162: ('a', 0),  # Aromatic
163: ('*1~*~*~*~*~*~1', 0),  # 6M Ring
164: ('[#8]', 0),  # O
165: ('[R]', 0),  # Ring
166: ('?', 0),  # Fragments  FIX: this can't be done in SMARTS
```

2D fingerprints
Atom pairs fingerprints

- Atom pairs: encode the distance between all pairs of atoms in a molecule
  Similar: pharmacophore distances along bonds (CATS descriptors)

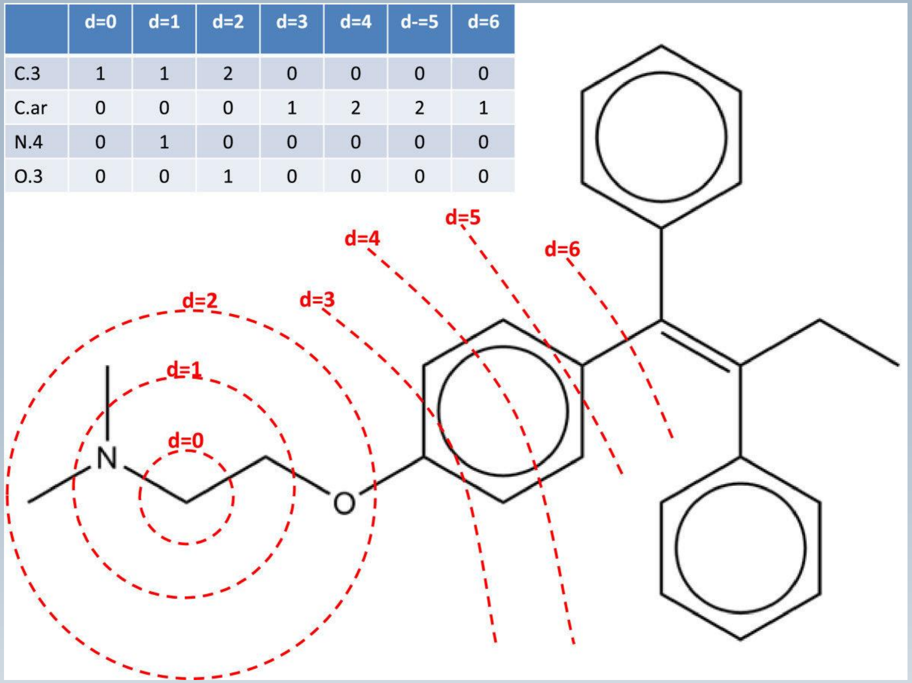


d / bonds	D+R	A+L	*etc.*
1	0	0	
2	0	1	
3	0	0	
4	1	0	
5	3	1	
6	4	0	
7	3	0	
8	1	0	
9	0	0	
10	0	0	
	λ =12	2	

doi: 10.1002/minf.201200141

2D fingerprints
Topological fingerprints

- Topological fingerprints:



	d=0	d=1	d=2	d=3	d=4	d-=5	d=6
C.3	1	1	2	0	0	0	0
C.ar	0	0	0	1	2	2	1
N.4	0	1	0	0	0	0	0
O.3	0	0	1	0	0	0	0

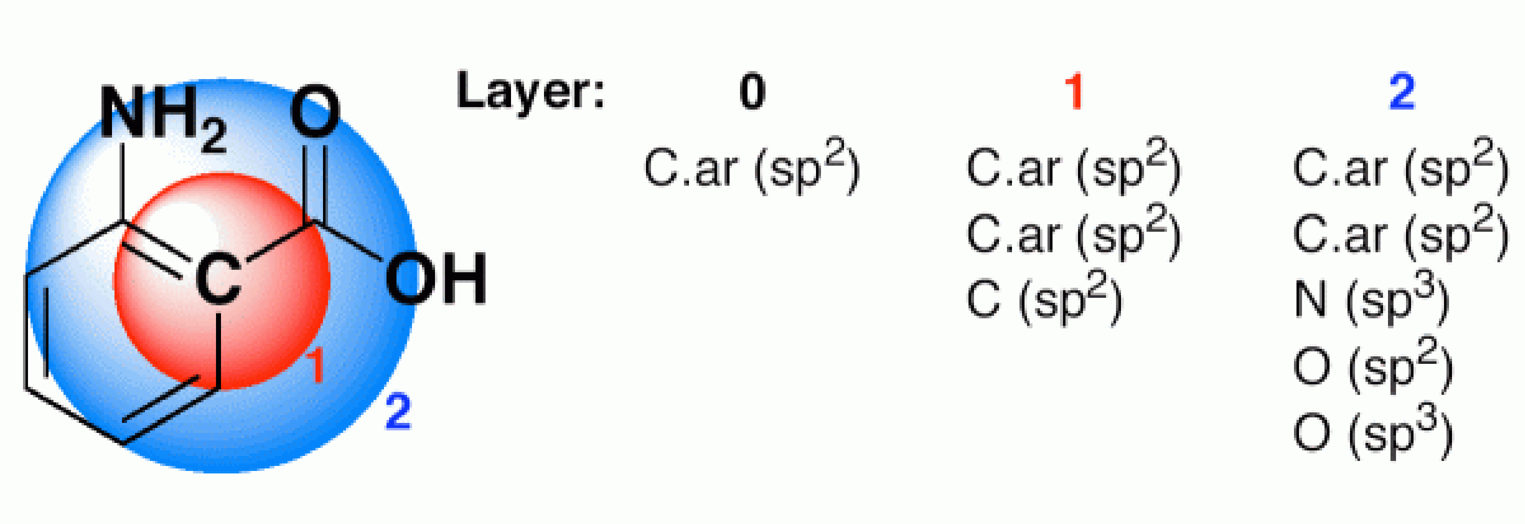
It identifies and hashes topological paths (e.g. along bonds) in the molecule and then uses them to set bits in a fingerprint of user-specified lengths.

https://doi.org/10.1186/1758-2946-6-29

2D fingerprints
Extended connectivity fingerprints

Extended connectivity fingerprints (ECFPs) are an example for circular topological fingerprints and are widely used in pharmaceutical companies (from Scitegic)

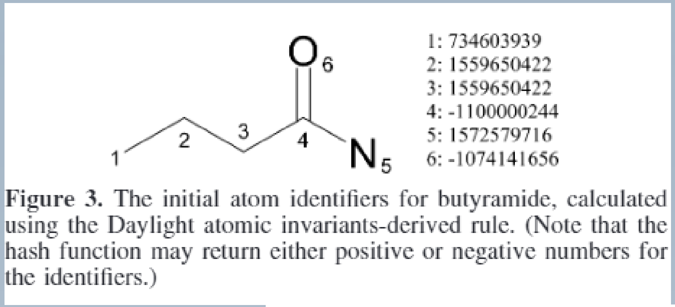
→ circular topological fingerprints
→ radius 4-6 atoms (ECFP4-ECFP6)
→ encode substructure patterns to bit string of length 1024 by hashing



Layer:	0	1	2
	C.ar (sp^2)	C.ar (sp^2)	C.ar (sp^2)
		C.ar (sp^2)	C.ar (sp^2)
		C (sp^2)	N (sp^3)
			O (sp^2)
			O (sp^3)

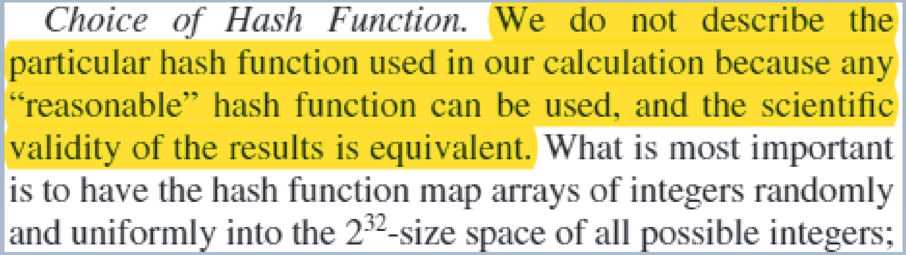
Specific example

- [Extended-Connectivity Fingerprints | Journal of Chemical Information and Modeling (acs.org)]



1: 734603939
2: 1559650422
3: 1559650422
4: -1100000244
5: 1572579716
6: -1074141656

Figure 3. The initial atom identifiers for butyramide, calculated using the Daylight atomic invariants-derived rule. (Note that the hash function may return either positive or negative numbers for the identifiers.)



Choice of Hash Function. We do not describe the particular hash function used in our calculation because any "reasonable" hash function can be used, and the scientific validity of the results is equivalent. What is most important is to have the hash function map arrays of integers randomly and uniformly into the 2^{32}-size space of all possible integers;

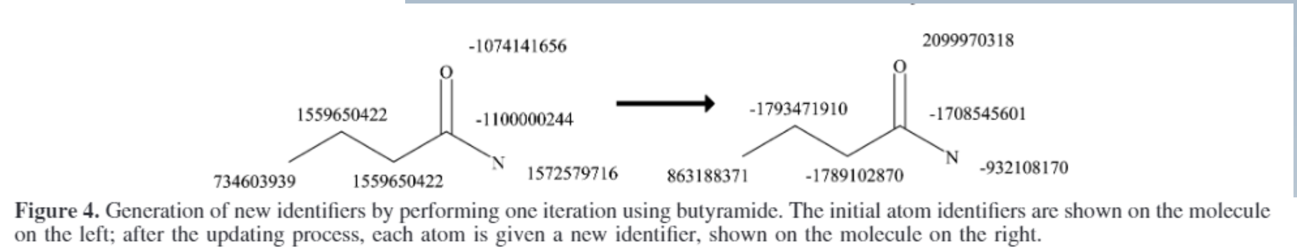
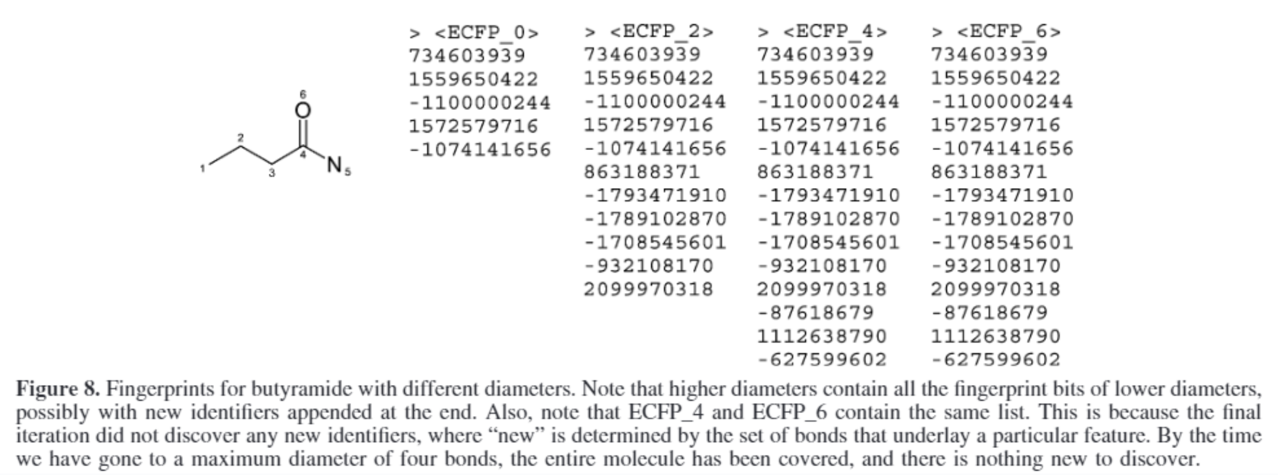


Figure 4. Generation of new identifiers by performing one iteration using butyramide. The initial atom identifiers are shown on the molecule on the left; after the updating process, each atom is given a new identifier, shown on the molecule on the right.

ROGERS AND HAHN



```
> <ECFP_0>      > <ECFP_2>      > <ECFP_4>      > <ECFP_6>
734603939       734603939       734603939       734603939
1559650422      1559650422      1559650422      1559650422
-1100000244     -1100000244     -1100000244     -1100000244
1572579716      1572579716      1572579716      1572579716
-1074141656     -1074141656     -1074141656     -1074141656
                863188371       863188371       863188371
                -1793471910     -1793471910     -1793471910
                -1789102870     -1789102870     -1789102870
                -1708545601     -1708545601     -1708545601
                -932108170      -932108170      -932108170
                2099970318      2099970318      2099970318
                                -87618679       -87618679
                                1112638790      1112638790
                                -627599602      -627599602
```

Figure 8. Fingerprints for butyramide with different diameters. Note that higher diameters contain all the fingerprint bits of lower diameters, possibly with new identifiers appended at the end. Also, note that ECFP_4 and ECFP_6 contain the same list. This is because the final iteration did not discover any new identifiers, where "new" is determined by the set of bonds that underlay a particular feature. By the time we have gone to a maximum diameter of four bonds, the entire molecule has been covered, and there is nothing new to discover.

2D fingerprints
Structural keys vs. fingerprints without dictionary

- Structure keys (= dictionary based) suffer from a lack of generality, because they highly depend on the predefined fragment dictionary.
- Fingerprints address this lack of generality by eliminating the idea of pre-defined patterns.
- Fingerprints are constructed from the molecule itself by an algorithm that examines the molecule and generates a series of patterns in the form of single atoms and bond sequences up to seven bonds long.
- A fingerprint is a Boolean array, or bitmap, but unlike a structural key there is no assigned meaning to each bit.
- Therefore, fingerprints apply to a wider range of molecular structures and have become the preferred type of molecular descriptors.

→ *often in literature there is no clean distinction between “keys” and “fingerprints”*

For further reading: Daylight Theory: Fingerprints

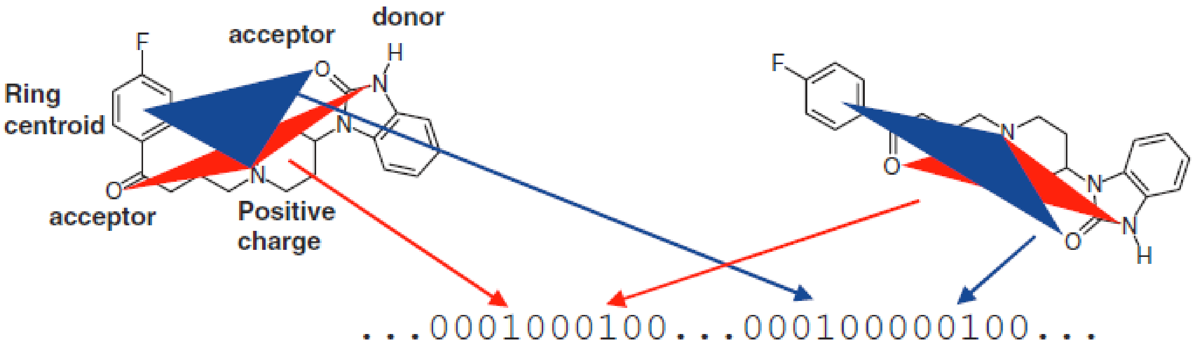
Molecular descriptors
3D fingerprints

Based on the generation of 3D conformations (time consuming for large datasets)

- 3D fragment screens: Originally designed for 3D substructure searching
  → based on distance/angle/dihedral ranges between atom types

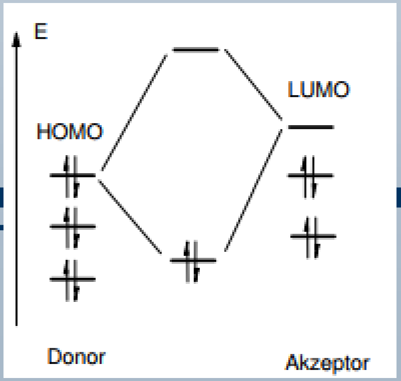
Pharmacophore keys:

- 3 (and 4)-point pharmacophores – most commonly used
  - enumerate all possible combinations of 3 pharmacophore features with all binned distances.

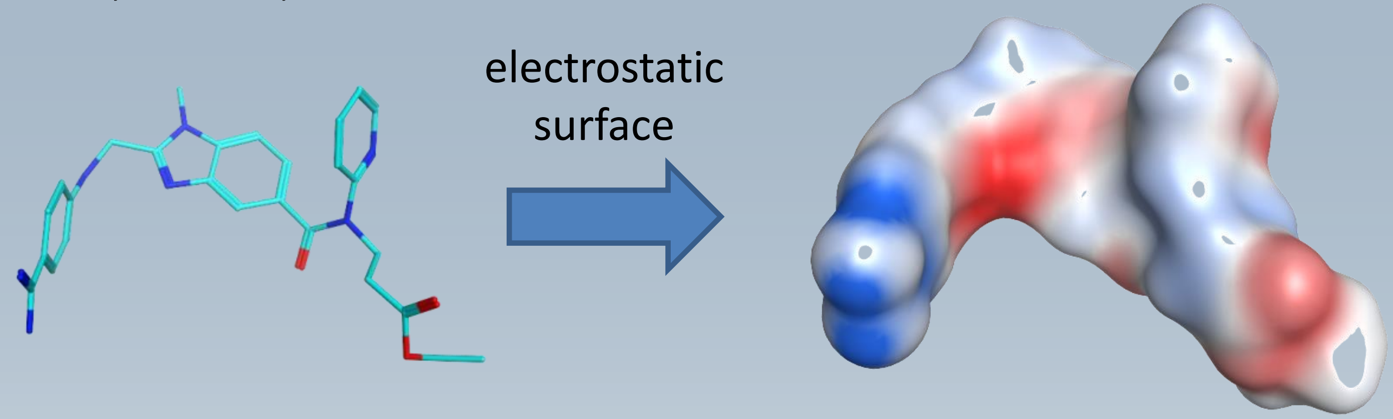


  - Example: Davies 1996
    - 7 feature types and 32 distance ranges
    - ~890.000 feasible different 3-point pharmacophores
      (for 4-point pharmacophores: 350 mio different geometries)

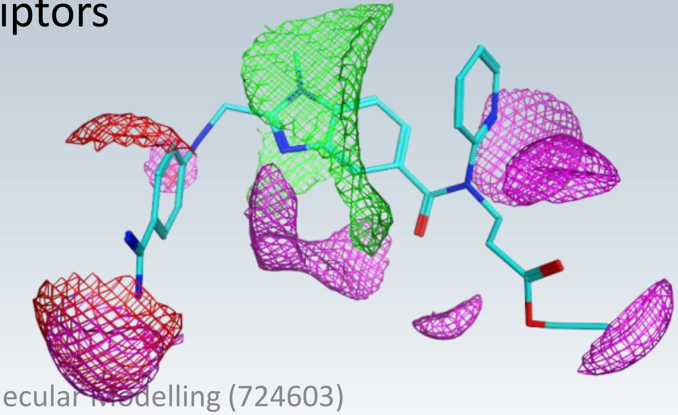
Molecular descriptors
3D descriptors



- Quantum mechanics descriptors
  - expensive to calculate
    Examples: HOMO, LUMO, dipole moment, partial charges, molecular surface properties, volume, ...


electrostatic surface

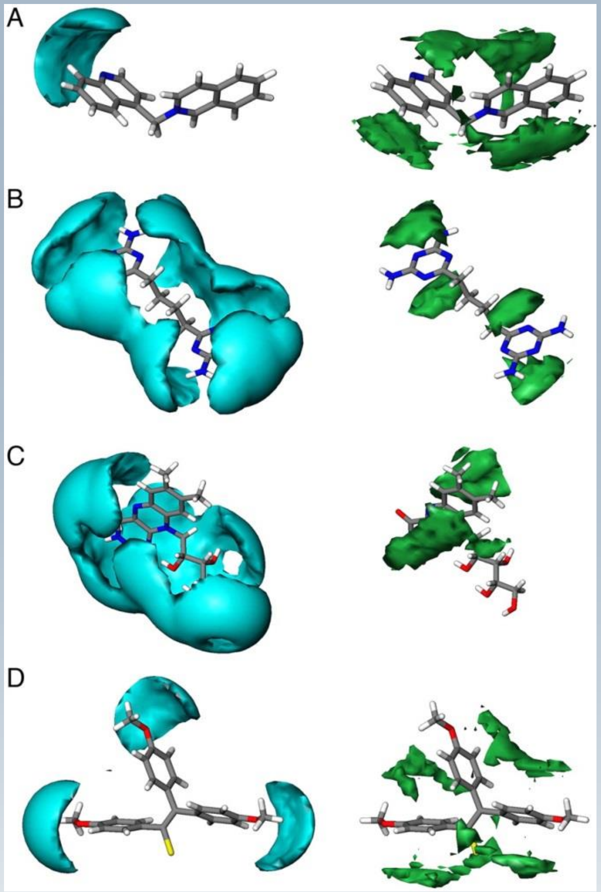
- Field descriptors



Best interaction potentials for: O- (red), H2O (purple), and hydrophobic probes (green)
→ GRID algorithm

Molecular descriptors
3D descriptors - Molecular interaction fields

- Molecular interaction fields



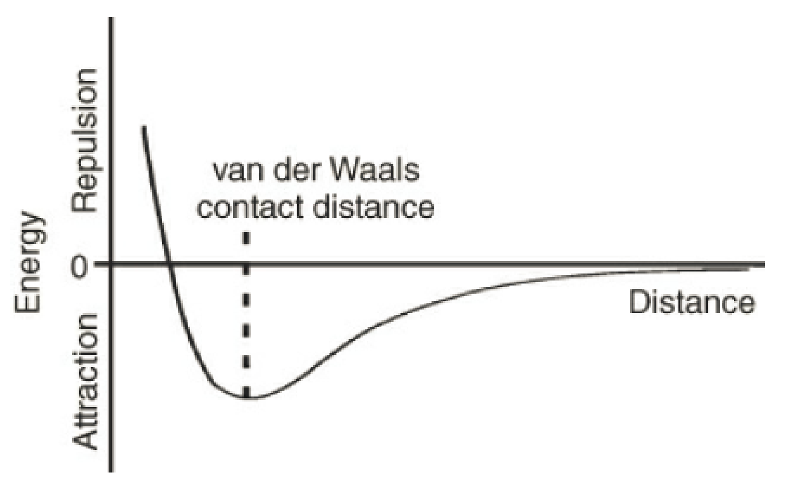
https://doi.org/10.1124/dmd.108.023507

amide probe

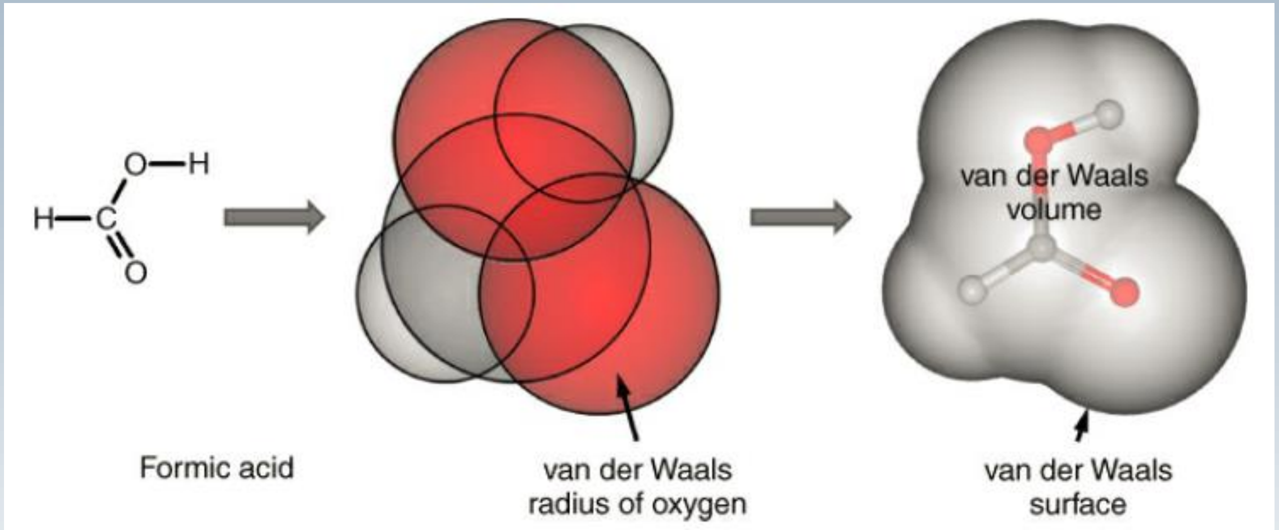
DRY probe

- Place the molecules into a rectangular grid
- put „probes“ on each grid point and calculate the interaction energy with the molecule
- „probes“: water, amide, DRY, charge, carbonyl oxygen, ...
- Display iso-surfaces at certain interaction cutoffs
- Use additional programs (e.g., Volsurf) to turn interaction fields into descriptors

Molecular descriptors
Molecular surfaces: (a) van der Waals surface



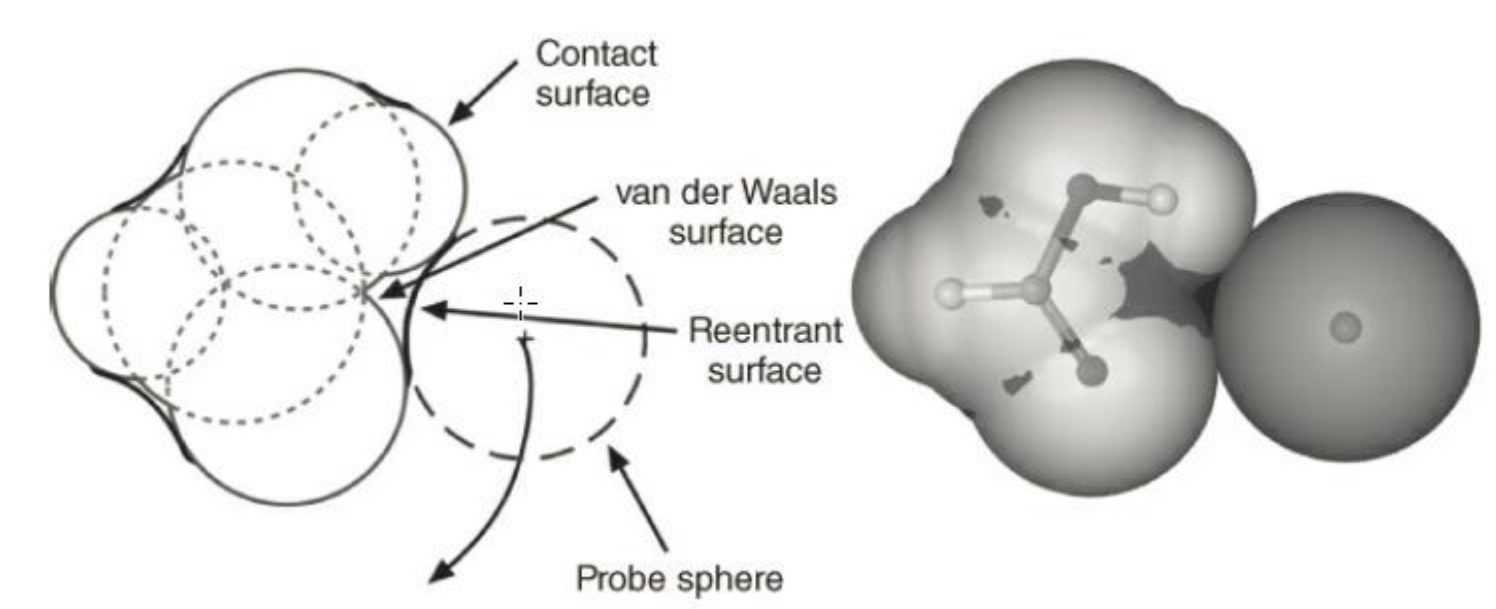
The van der Waals surface is determined by the atomic van der Waals contact distances.
Spheres with these distances are centered on each atom.
→ 3D conformation dependent



Molecular descriptors
Molecular surfaces: (b) Connolly Surface

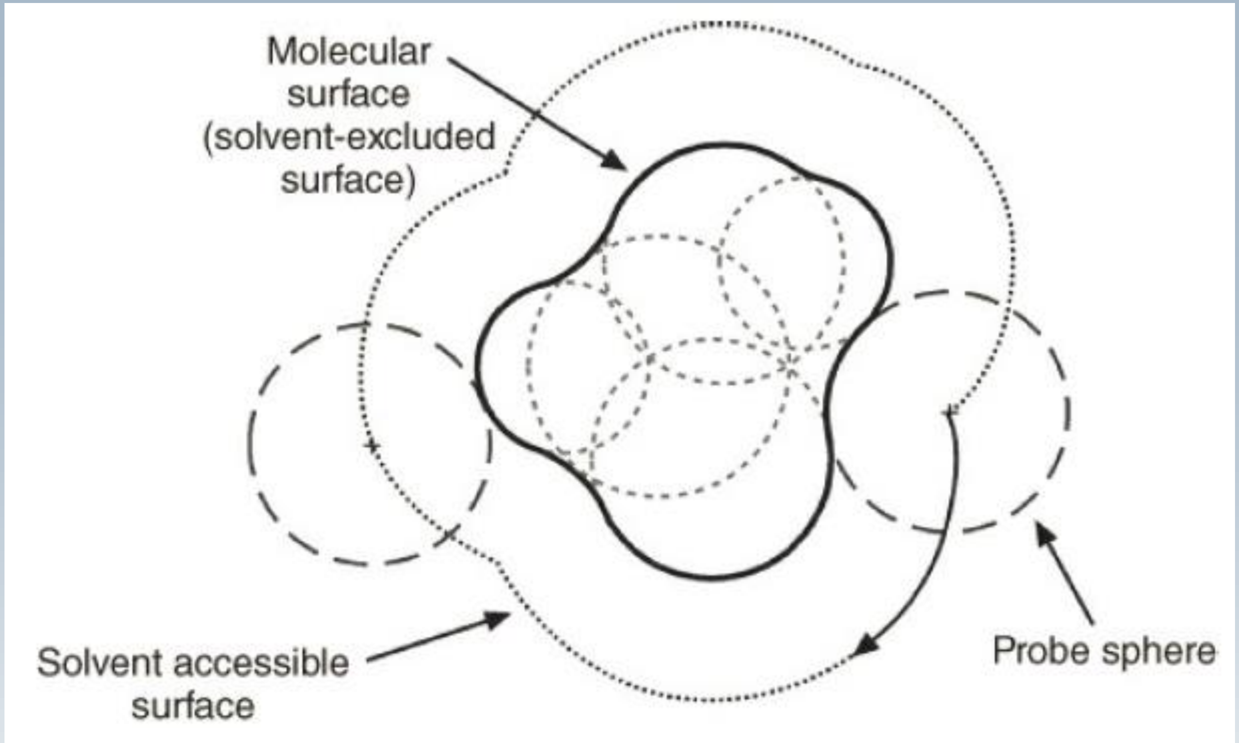
The Connolly or molecular surface is obtained by rolling a spherical probe over the van der Waals surface – usually the radius of 1.4Å (water) is chosen.

→ The Connolly surface is much smoother than the vdW-surface.



Molecular descriptors
Molecular surfaces: (c) Solvent accessible surface

The solvent accessible surface (SAS) : The **center** of the solvent sphere defines the SAS, which is a subtle difference to the Connolly surface


Molecular surface (solvent-excluded surface)
Solvent accessible surface
Probe sphere

Molecular descriptors
Availability

- Most basic descriptors included in all molecular modeling suites
  e.g. MOE or Schrödingers Canvas

- Some software companies have focused on descriptor implementation
  e.g., Kode (Dragon descriptors) provides >5200 different molecular descriptors

Block no.	Block name	Descriptors
1	Constitutional	47
2	Ring descriptors	32
3	Topological indices	75
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	50
7	2D matrix-based descriptors	607
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P-VSA-like descriptors	55
11	ETA indices	23
12	Edge adjacency indices	324
13	Geometrical descriptors	38
14	3D matrix-based descriptors	99
15	3D autocorrelations	80
16	RDF descriptors	210
17	3D-MoRSE descriptors	224
18	WHIM descriptors	114
19	GETAWAY descriptors	273
20	Randic molecular profiles	41
21	Functional groups count	154
22	Atom-centered fragments	115
23	Atom-type E-state indices	172
24	CATS 2D	150
25	2D Atom Pairs	1596
26	3D Atom Pairs	36
27	Charge descriptors	15
28	Molecular properties	20
29	Drug-like indices	28
30	CATS 3D	300

Dragon 7 descriptors

Molecular descriptors
Analysis based on descriptors

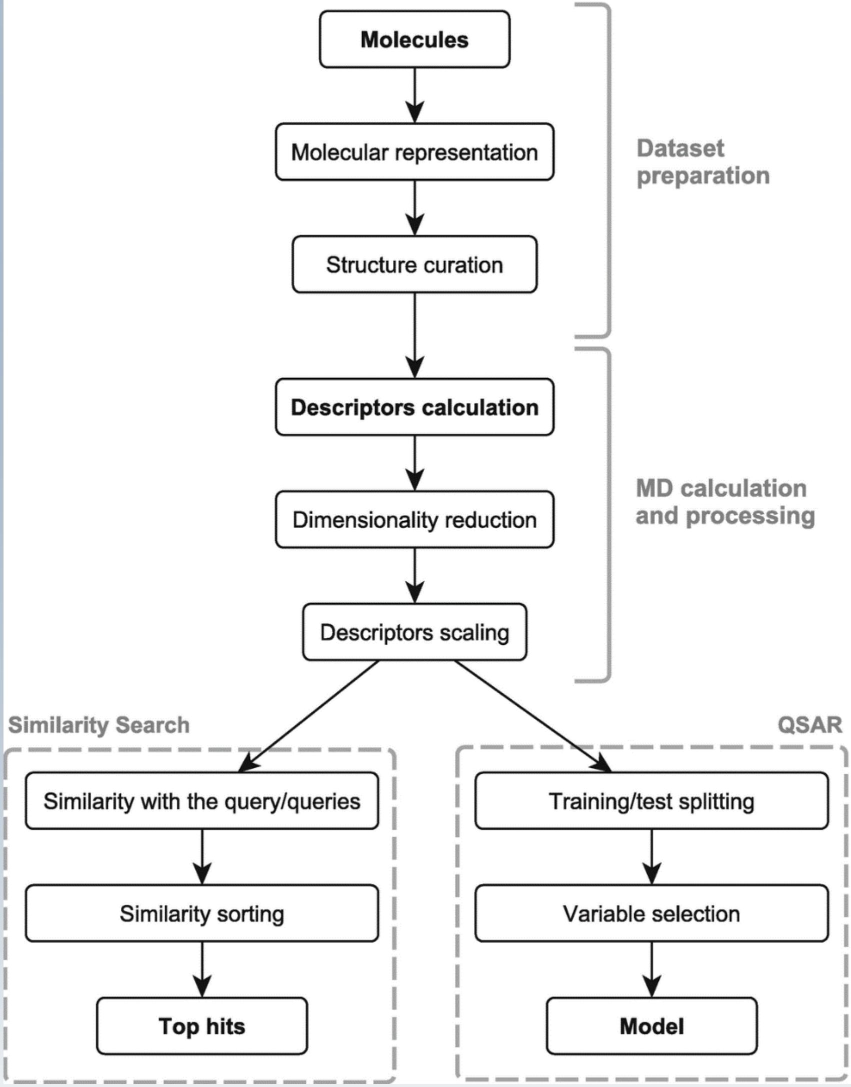
Dataset properties

- Often overcomplete dataset (more descriptors than molecules)
- Descriptors should be able to discriminate molecules
  e.g., descriptors with zero-variance are useless
- Highly correlated descriptors don‘t give extra information

→ Remove descriptors which don’t explain any variance, or which are heavily correlated

Questions to be asked

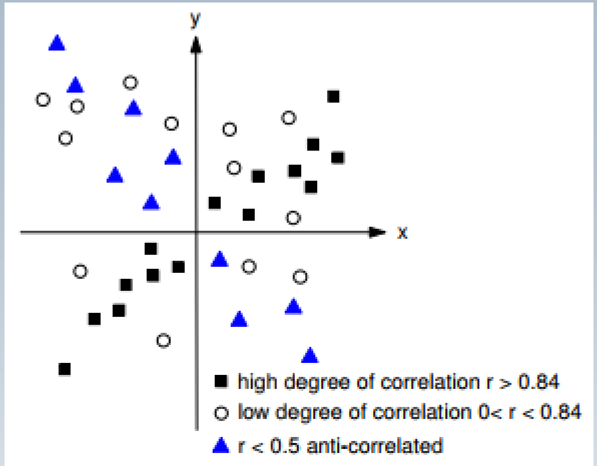
- Diversity of a dataset?
- Can the dataset be partitioned in descriptor space?
- Correlation to properties of interest (affinity, potency, ...)?



Impact of Molecular Descriptors on Computational Models | SpringerLink

Data analysis
Preparation of datasets

- Descriptors vary orders of magnitude
- Scaling and standardization required
  → Centering to the mean value and scaling to unit variance
- Check for correlations in descriptors – e.g., by calculation of the correlation matrix (entry i,j is the correlation coefficient between descriptors x_i and x_j)



$$r = \frac{\sum_{k=1}^{N}\left[\left(x_{i,k} - \langle x_i \rangle\right)\left(x_{j,k} - \langle x_j \rangle\right)\right]}{\sqrt{\sum_{k=1}^{N}\left(x_{i,k} - \langle x_i \rangle\right)^2 \sum_{k=1}^{N}\left(x_{j,k} - \langle x_j \rangle\right)^2}}$$

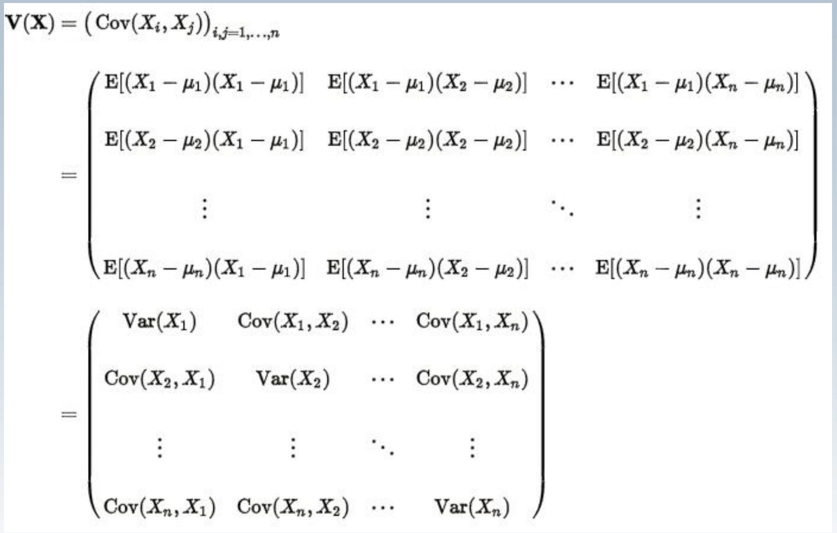
Data analysis
Principal component analysis

- Multivariate method for the reduction of the dimension of a dataset
- Goal: rotation and scaling of the axes in a way that the maximal variance of the dataset is found on the first axis followed by the second highest variance on the second axis ...
- Mathematically:
  Singular value decomposition of the data matrix M

$$M = U\Sigma V^*$$

where U is unitary, Σ is diagonal and V* is adjunct to a unitary matrix V

Equivalent alternative:
U and Σ can be computed as
the eigenvalues and eigenvectors of the
covariance matrix of the dataset.



$$\mathbf{V}(\mathbf{X}) = \left(\operatorname{Cov}(X_i, X_j)\right)_{i,j=1,\dots,n}$$

$$= \begin{pmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{pmatrix}$$

$$= \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1,X_2) & \cdots & \operatorname{Cov}(X_1,X_n) \\ \operatorname{Cov}(X_2,X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_n,X_1) & \operatorname{Cov}(X_n,X_2) & \cdots & \operatorname{Var}(X_n) \end{pmatrix}$$

Data analysis
Principal component analysis

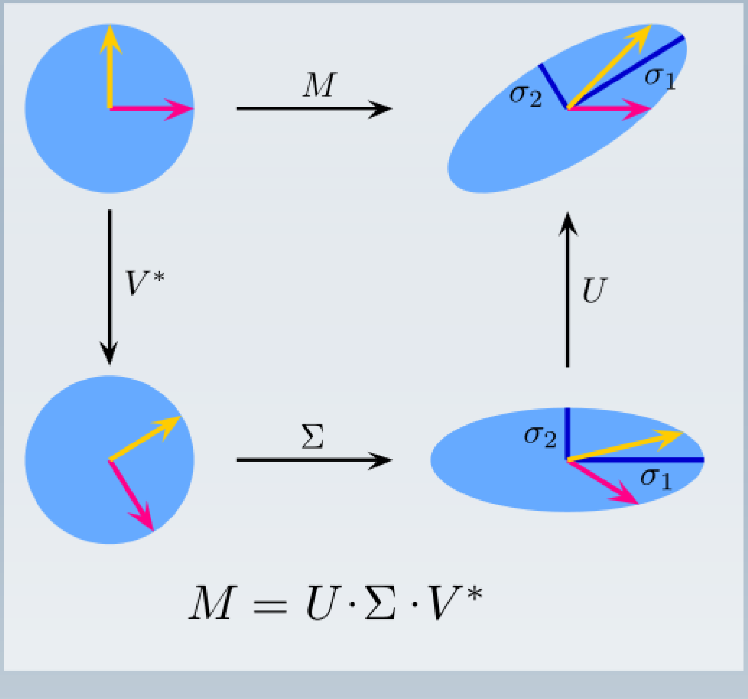
- Geometrical explanation of $M = U\Sigma V^*$
  - M transforms the unit data to a rotated ellipsoid
  - U and V rotate the data and Σ scales along the axis

- PCA rotates and scales the original coordinate system

- New axes are the **linear** combinations of the old variables

$$PC_i = \sum_{j=1}^{n} c_{ij} x_j$$

resulting in a new coordinate system

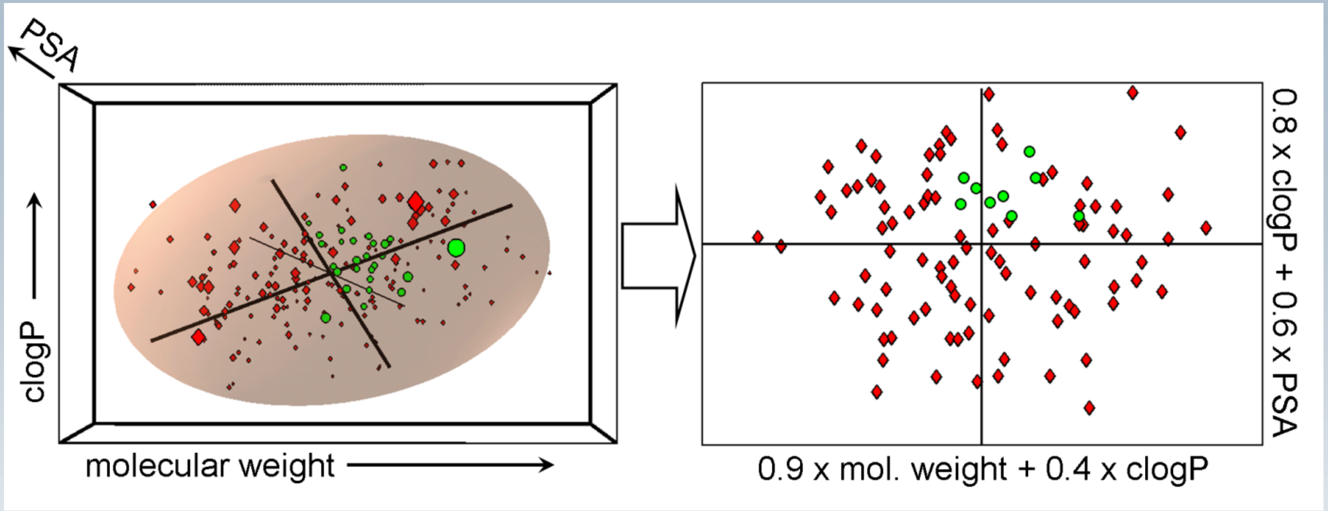


By Georg-Johann - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=11342212

Data analysis
Principal component analysis

Practical results

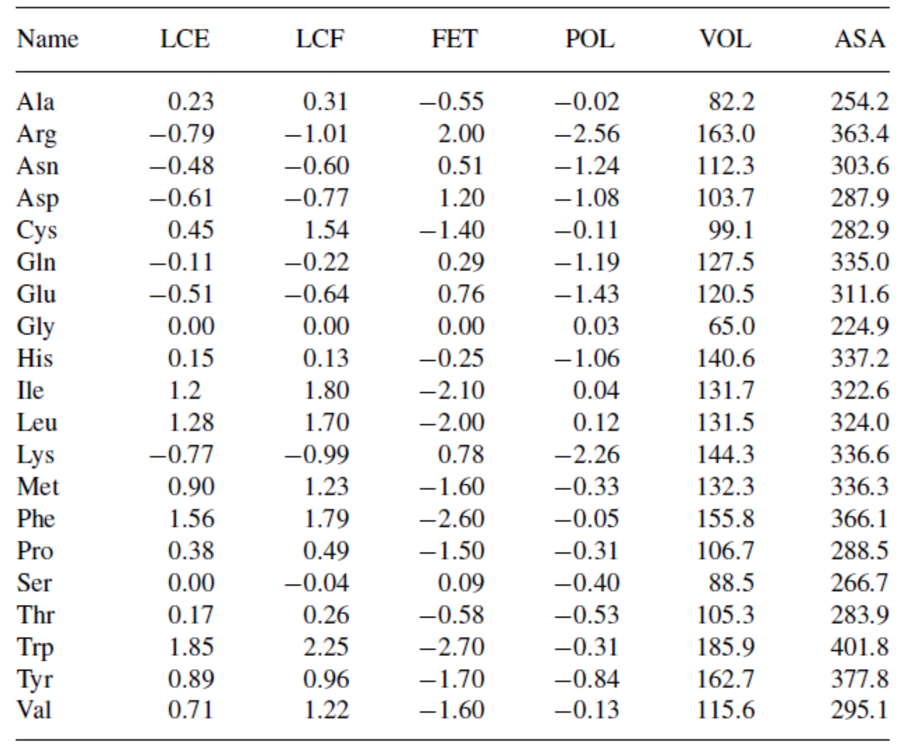
- Reduction of the high dimensional dataset to few (usually 2-6) variables, which explain major parts of the variance in the dataset
- Generation of a set *of linear independent axes*, which can be used as new descriptors Caveat: the interpretation of the new descriptors is difficult (linear combinations of the original descriptors)
- The higher the singular value σ_i, the more variance explained by axis PC_i



Molecular descriptors
PCA example

Both, the coefficients of the new axes (loadings) and the new coordinates (scorings) of the data are important results

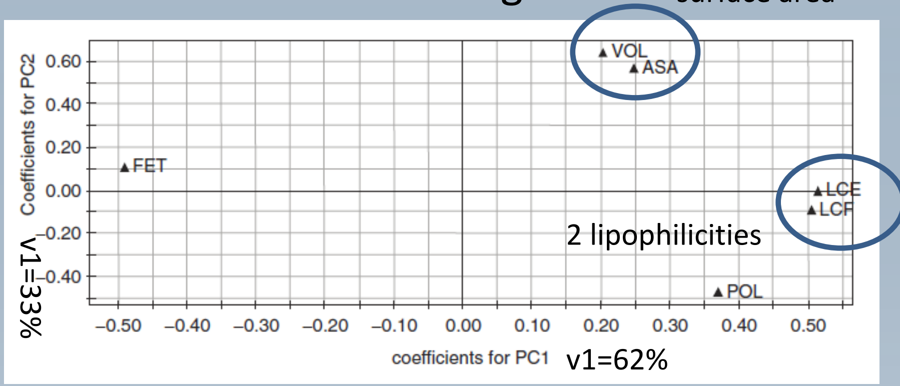
lipophilicity solvation polarity volume surface



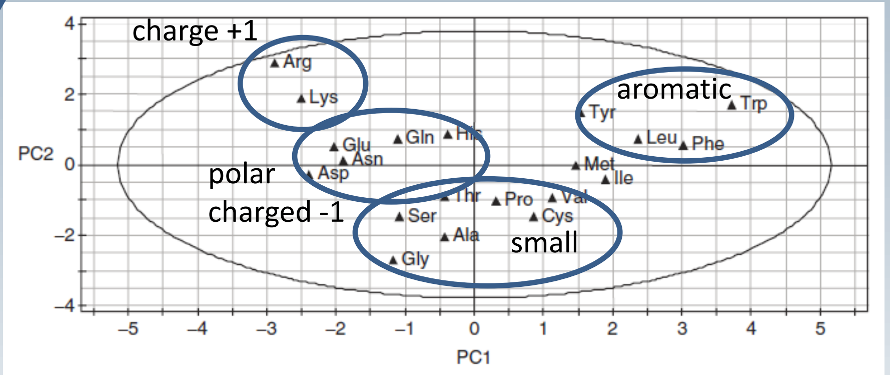
Name	LCE	LCF	FET	POL	VOL	ASA
Ala	0.23	0.31	−0.55	−0.02	82.2	254.2
Arg	−0.79	−1.01	2.00	−2.56	163.0	363.4
Asn	−0.48	−0.60	0.51	−1.24	112.3	303.6
Asp	−0.61	−0.77	1.20	−1.08	103.7	287.9
Cys	0.45	1.54	−1.40	−0.11	99.1	282.9
Gln	−0.11	−0.22	0.29	−1.19	127.5	335.0
Glu	−0.51	−0.64	0.76	−1.43	120.5	311.6
Gly	0.00	0.00	0.00	0.03	65.0	224.9
His	0.15	0.13	−0.25	−1.06	140.6	337.2
Ile	1.2	1.80	−2.10	0.04	131.7	322.6
Leu	1.28	1.70	−2.00	0.12	131.5	324.0
Lys	−0.77	−0.99	0.78	−2.26	144.3	336.6
Met	0.90	1.23	−1.60	−0.33	132.3	336.3
Phe	1.56	1.79	−2.60	−0.05	155.8	366.1
Pro	0.38	0.49	−1.50	−0.31	106.7	288.5
Ser	0.00	−0.04	0.09	−0.40	88.5	266.7
Thr	0.17	0.26	−0.58	−0.53	105.3	283.9
Trp	1.85	2.25	−2.70	−0.31	185.9	401.8
Tyr	0.89	0.96	−1.70	−0.84	162.7	377.8
Val	0.71	1.22	−1.60	−0.13	115.6	295.1

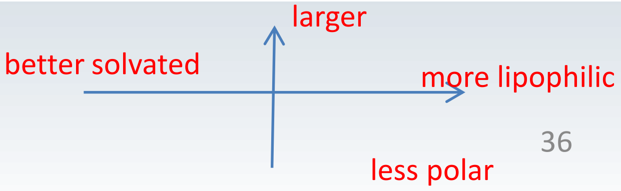
loadings

volume and accessible surface area



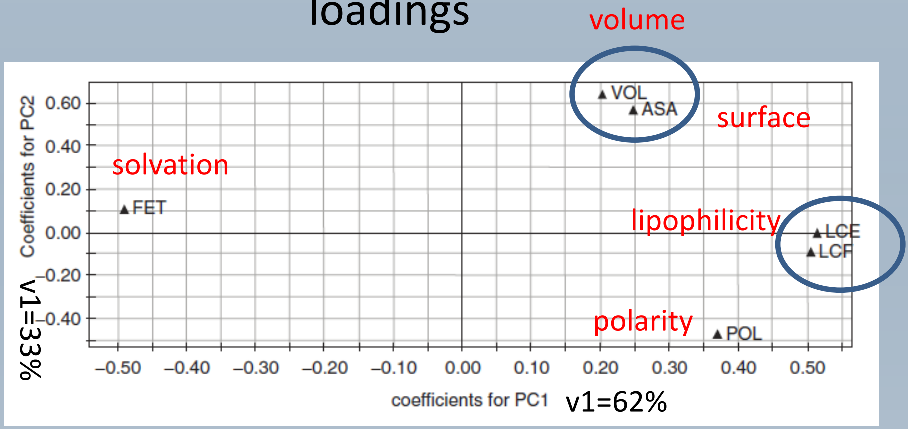
scorings



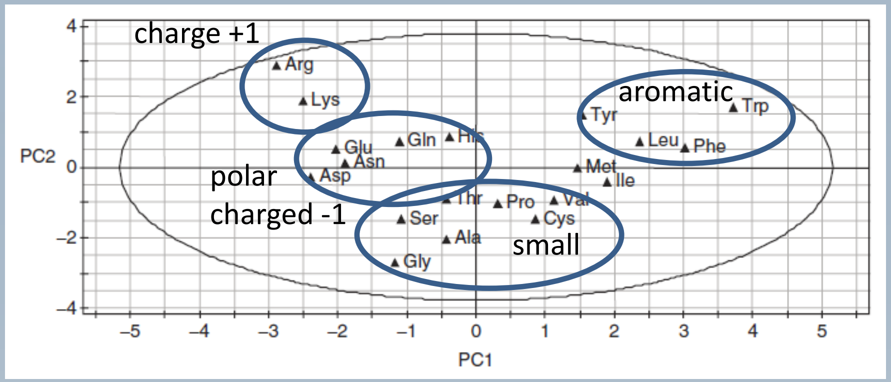


Molecular descriptors
PCA example - interpretation

loadings



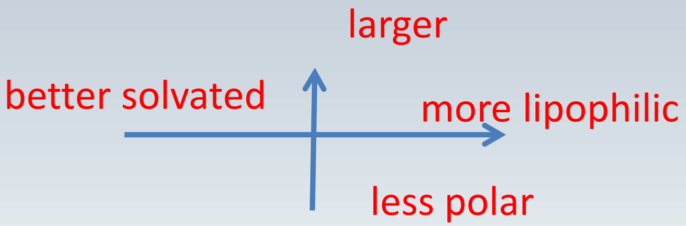
scorings



- Correlated descriptors are found closely together in loading plot
  → contribution of the original descriptors to the new axes (i.e., the principal components)

- Similar compounds are found closely together in scoring plot
  → scorings= coefficients in the new coordinate system

- from the loading plot a qualitative picture of compound properties in the scoring plot can be deduced


larger

better solvated

more lipophilic

less polar

Summary

- Molecular descriptors
  - calculated properties of molecules of different complexity
  - based on 1D, 2D, 3D structure of molecule
  - several thousand different descriptors available
  - fingerprints are often used due to increased generality and practical reasons

- Analysis
  - scaling and standardization
  - correlated descriptors
  - principal component analysis to
    - analyze dataset – see trends and clusters
    - reduce dimensionality of descriptor space

Molecular similarity and diversity

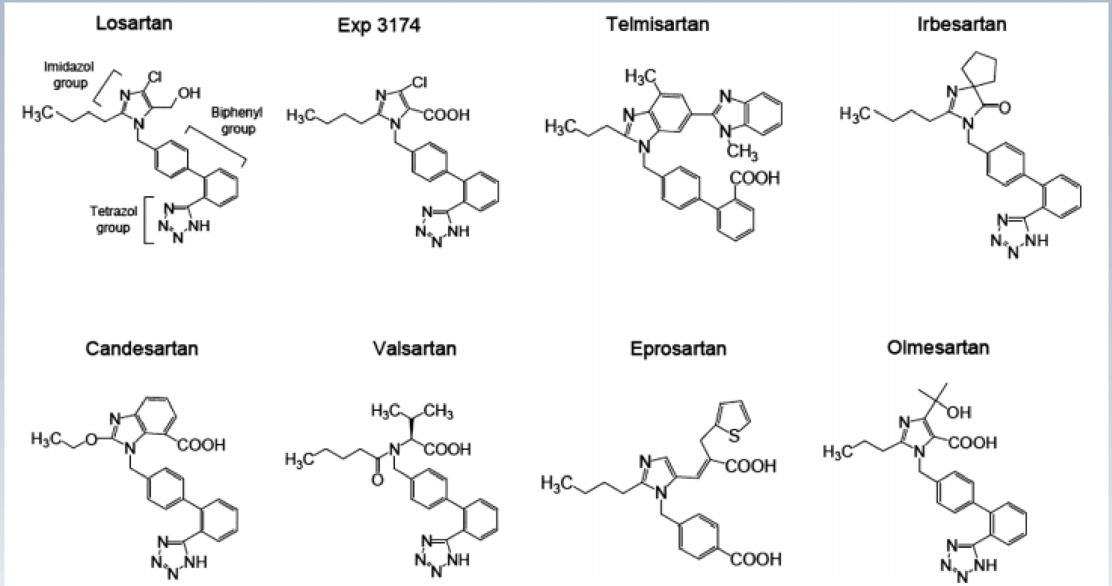
How do you compare molecular structures?

Molecular similarity and diversity

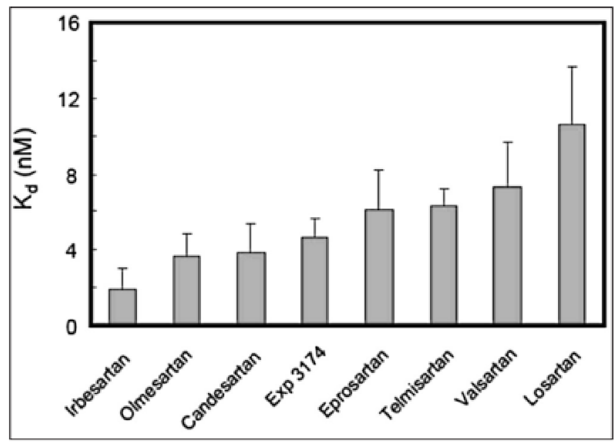
- Fingerprint based similarity
- Similarity indices
- 3D-similarities (non-fingerprint based)

Why similarity?

- Pharmacophore searches and substructure searches rely on exact matches
- Biological activity can be achieved by exchanges of small groups → similarity of molecules is more relevant than exact matches
- Example AT_1 receptor inhibitors for blood pressure lowering → compounds are similar and show potency in similar range



doi:10.1177/1470320310370852



Why similarity?

- „Similar property principle“ or „Neighborhood behavior”
  - structurally similar molecules tend to have similar properties (Johnson and Maggiora 1990 or Patterson et al. 1996)
  - Can be expanded to „chemogenomics“:
    Binding sites, which are phylogenetically related should accommodate similar ligands, and known ligands for a certain target are valid starting points for identifying ligands that bind to closely related targets.

- Advantages of similarity considerations
  - No substructure or pharmacophore to be defined
    → only an active compound is required as starting point
  - User can determine number of hits by adjusting the similarity score
  - Similarity depends on the descriptors
    → different descriptors yield other similarities
  - Similarity relations needed to identify diverse subsets of molecules

Molecular similarity
Quantification of chemical similarity

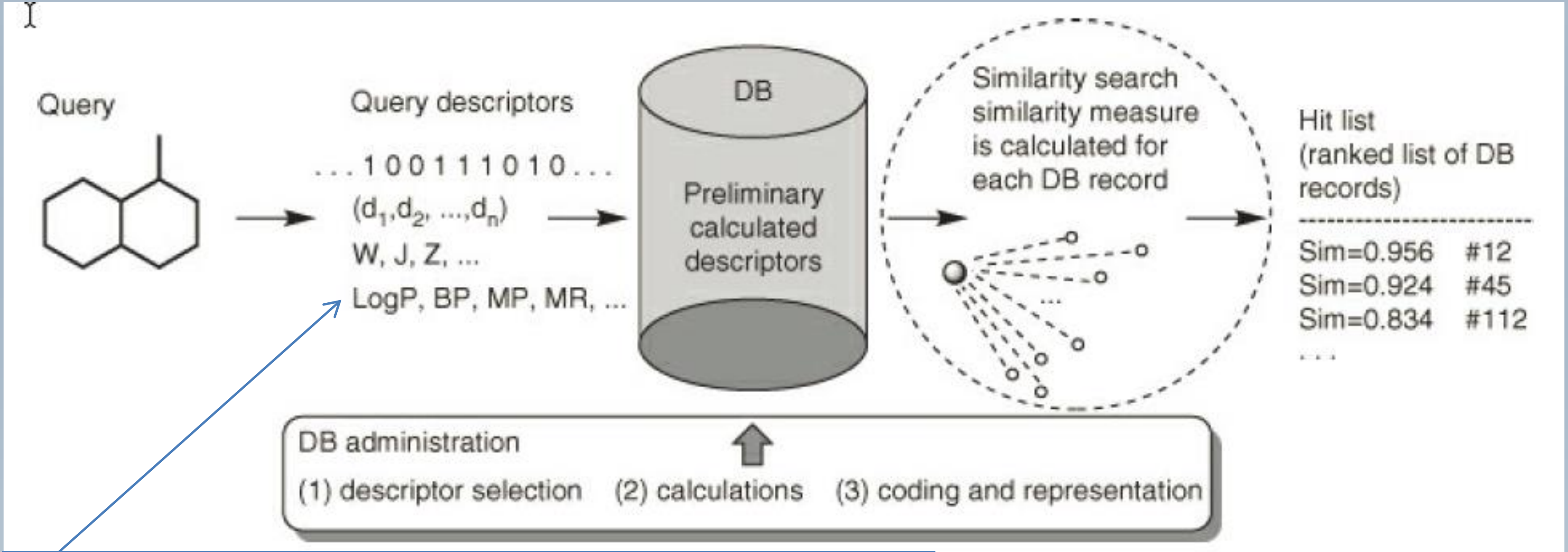
Methods to search for similar molecules

- Substructure Searching
  - Result: Match/Mismatch → size of hitlist can’t be influenced
- Pharmacophore Searching
  - Result: Match/Mismatch → size of hitlist can’t be influenced
- Similarity Searching in Chemical Databases
  - Result: Rank by Similarity → size of hitlist user determined

Usual approach to calculate similarity

- Molecules are represented by a set of the same numerical descriptors or fingerprints
- The distance D in the descriptor space is calculated
- Similarity S = 1 – D (if D is normalized)

Molecular similarity
Descriptors useful for similarity comparisons



Query

Query descriptors

. . . 1 0 0 1 1 1 0 1 0 . . .

$(d_1, d_2, \ldots, d_n)$

W, J, Z, ...

LogP, BP, MP, MR, ...

DB

Preliminary calculated descriptors

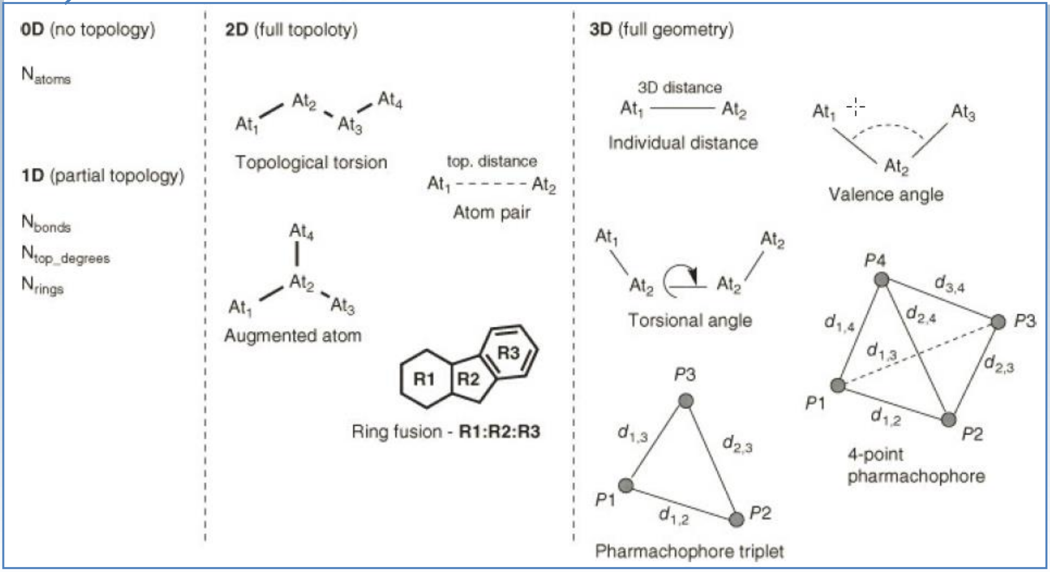
Similarity search similarity measure is calculated for each DB record

Hit list (ranked list of DB records)

Sim=0.956	#12
Sim=0.924	#45
Sim=0.834	#112
. . .	

DB administration

(1) descriptor selection (2) calculations (3) coding and representation



0D (no topology)

N_{atoms}

1D (partial topology)

N_{bonds}

$N_{top\_degrees}$

N_{rings}

2D (full topoloty)

Topological torsion

top. distance

Atom pair

Augmented atom

Ring fusion - **R1:R2:R3**

3D (full geometry)

3D distance

Individual distance

Valence angle

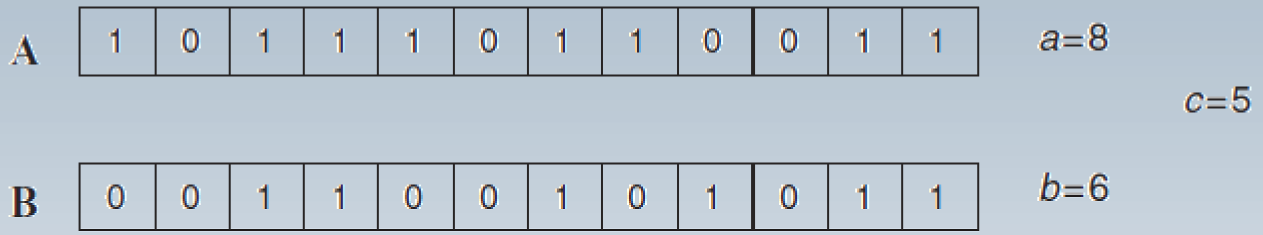
Torsional angle

Pharmachophore triplet

4-point pharmachophore

Molecular similarity
Fingerprint similarities

- Oldest (mid 1980s) and most common approach
- Based on fingerprint vectors of same length
- Similarity of 2 fingerprint vectors assessed on the presence/absence of identical bits
- Most common: Tanimoto similarity $S_{AB} = \frac{c}{a+b-c}$ where
  - a ... number of bits set to “1” in A,
  - b ... number of bits set to “1” in B, and
  - c ... number of common “1”s in A and B:



A	1	0	1	1	1	0	1	1	0	0	1	1	a=8
													c=5
B	0	0	1	1	0	0	1	0	1	0	1	1	b=6

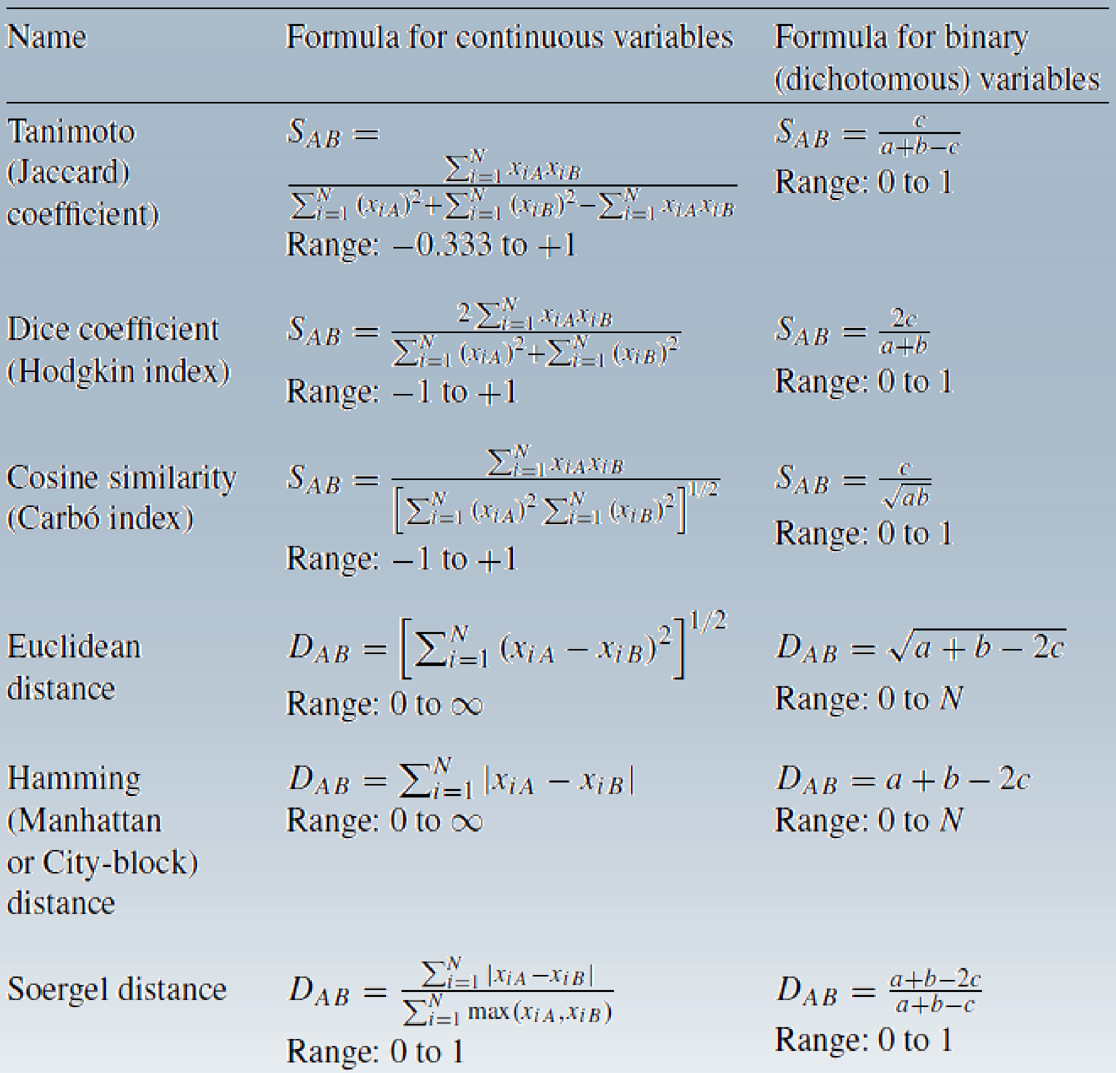
$$S_{AB} = \frac{5}{8+6-5} = 0.56$$

Molecular similarity
Fingerprint similarities – various indices

- Frequently used measures for distance or similarity:

- Asymmetric index: (Tversky index)

$$S_{\mathrm{Tversky}} = \frac{c}{\alpha\,(a-c)+\beta\,(b-c)+c}$$



Name	Formula for continuous variables	Formula for binary (dichotomous) variables
Tanimoto (Jaccard coefficient)	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA}x_{iB}}{\sum_{i=1}^{N}(x_{iA})^2+\sum_{i=1}^{N}(x_{iB})^2-\sum_{i=1}^{N} x_{iA}x_{iB}}$ Range: -0.333 to $+1$	$S_{AB} = \frac{c}{a+b-c}$ Range: 0 to 1
Dice coefficient (Hodgkin index)	$S_{AB} = \frac{2\sum_{i=1}^{N} x_{iA}x_{iB}}{\sum_{i=1}^{N}(x_{iA})^2+\sum_{i=1}^{N}(x_{iB})^2}$ Range: -1 to $+1$	$S_{AB} = \frac{2c}{a+b}$ Range: 0 to 1
Cosine similarity (Carbó index)	$S_{AB} = \frac{\sum_{i=1}^{N} x_{iA}x_{iB}}{\left[\sum_{i=1}^{N}(x_{iA})^2\sum_{i=1}^{N}(x_{iB})^2\right]^{1/2}}$ Range: -1 to $+1$	$S_{AB} = \frac{c}{\sqrt{ab}}$ Range: 0 to 1
Euclidean distance	$D_{AB} = \left[\sum_{i=1}^{N}(x_{iA}-x_{iB})^2\right]^{1/2}$ Range: 0 to ∞	$D_{AB} = \sqrt{a+b-2c}$ Range: 0 to N
Hamming (Manhattan or City-block) distance	$D_{AB} = \sum_{i=1}^{N} \lvert x_{iA}-x_{iB}\rvert$ Range: 0 to ∞	$D_{AB} = a+b-2c$ Range: 0 to N
Soergel distance	$D_{AB} = \frac{\sum_{i=1}^{N} \lvert x_{iA}-x_{iB}\rvert}{\sum_{i=1}^{N}\max(x_{iA},x_{iB})}$ Range: 0 to 1	$D_{AB} = \frac{a+b-2c}{a+b-c}$ Range: 0 to 1

Molecular similarity
Fingerprint similarities – discussion of indices

- Some of the coefficients (Hamming, Euclidian, Soergel) obey conditions of a metric *d*. 1-$S_{Tanimoto}$ on binary fingerprints as well.

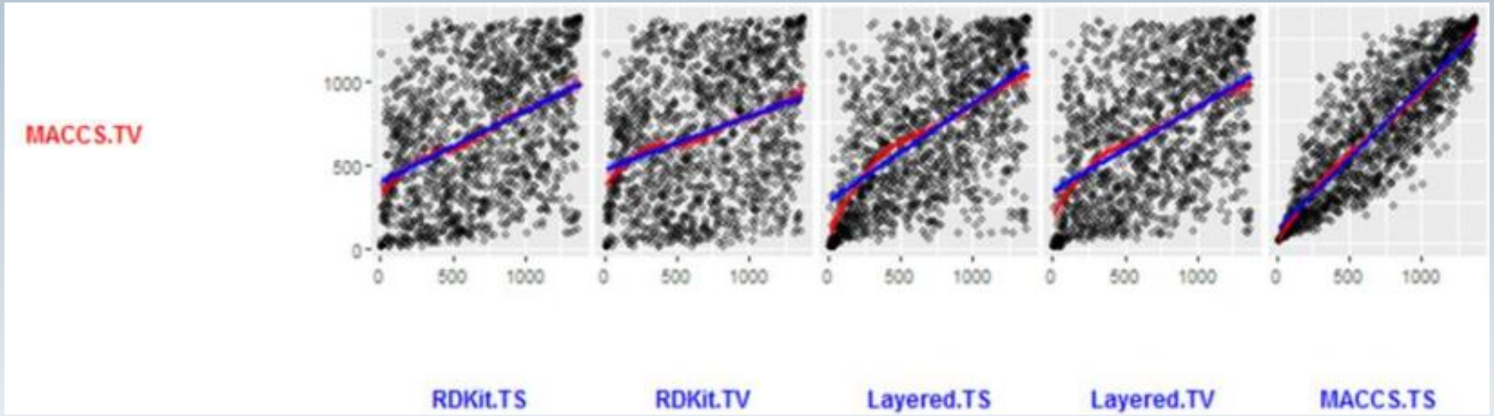
  1. $d(x,y) \geq 0$
  2. $d(x,y) = 0 \Leftrightarrow x = y$
  3. $d(x,y) = d(y,x)$
  4. $d(x,z) \leq d(x,y) + d(y,z)$

  Advantage of a metric: a metric space enables relations of distances

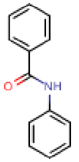
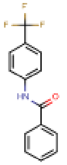
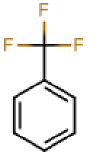
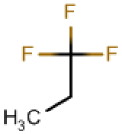
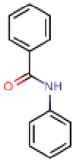
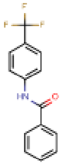
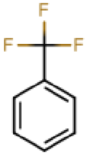
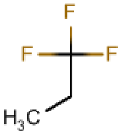
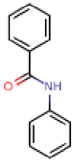
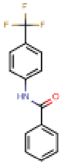
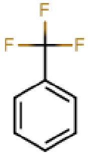
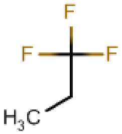
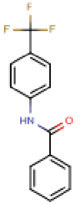
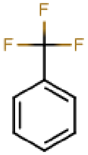
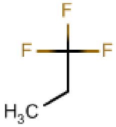
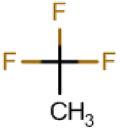
- The coefficients are monotonic with each other (not the asymmetric ones)
  → they produce the same similarity ranking

Molecular similarity
Fingerprint similarities – discussion of indices

- Other features:
  - Tanimoto, Dice, Cosine directly dependent on number of bits in common → smaller molecules often get smaller similarities
  - Hamming and Euclidian distances regard the common absence of features as similar
- Comparison of molecules of different sizes:
  - asymmetric indices appropriate for comparison of molecules of different size
- Dependence on fingerprints larger than the dependence on similarity index (ADMET & DMPK 5(2) (2017) 85-125)



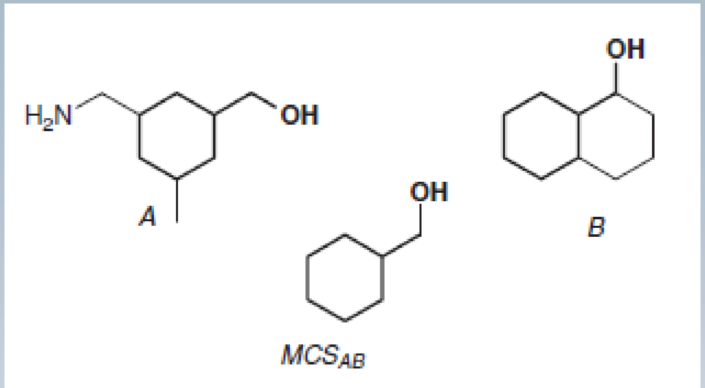
Molecular similarity
Fingerprint similarities – comparison of indices

												
	0,72	Tversky (0.1/2.0)			0,84	Tanimoto			0,91	Dice		
	0,40	0,80			0,20	0,29			0,33	0,45		
	0,11	0,38	0,53		0,03	0,08	0,24		0,06	0,15	0,39	
	0,11	0,43	0,76	0,96	0,03	0,07	0,24	0,73	0,05	0,13	0,39	0,84

→ Comparison small/large molecules requires asymmetric index

Molecular similarity
Maximum common substructures

- Fingerprint based indices are global measures
  - Bit-strings describe the whole molecule and similarity is based on the comparison of two whole molecules
- Alternative: Look for a mapping between molecules
  - Maximum common substructure (MCS)
  - Similarity can be calculated based on matching bonds/overall bonds.
  - MCS determination NP-complete and thus, time consuming
    → for a larger set of molecules prescreening techniques have to be applied
  - Extreme case: Substructure search



Molecular similarity
3D similarity

Why another similarity method?

- 2D searches tend to find common substructures
- Pharmacophore searches find „exact matches“ – and don‘t return similarity values

→ shape and spatial feature location neglected in 2D

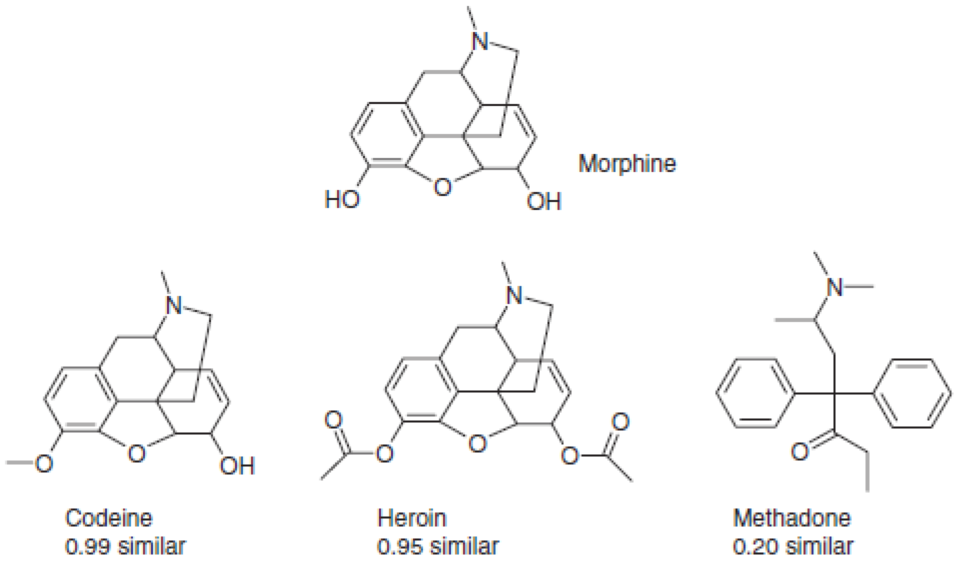
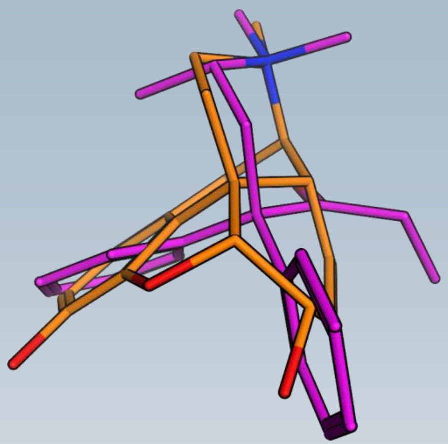


Figure 5-5. Similarities to morphine calculated using Daylight fingerprints and the Tanimoto coefficient.



3D overlay methadone/morphine (shape Tanimoto 0.72)

Molecular similarity
3D similarity

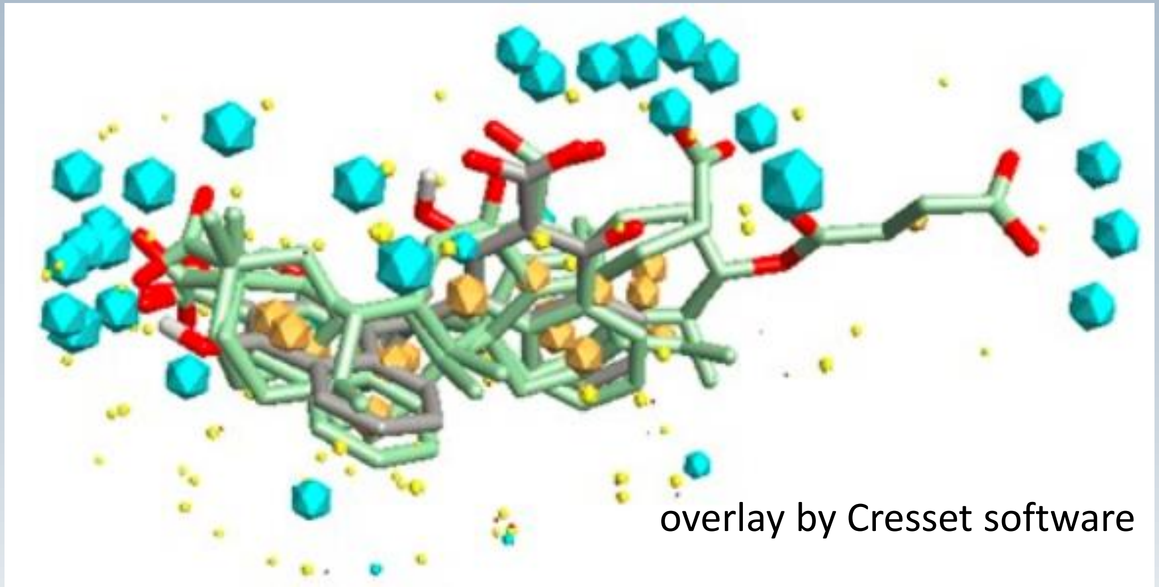
- Conformational properties of molecules required for 3D similarity calculation
- Hypothesis that molecules bind the target in the same way and should occupy the same volume and interact similarly

- Different approaches for 3D similarity
  - Alignment independent:
    → compare similarity between ensembles (3D fingerprints, pharmacophore keys)
  - Alignment dependent
    → similarity dependent on an alignment step
    - Common procedure
      → define one molecule as rigid (e.g., bioactive conformation – critical step) and compare to ensemble of conformations of query
      → pre-calculation of conformational ensemble of structures
      → shape and pharmacophore-feature overlay

Molecular similarity
3D similarity – alignment methods

Overlay of features rather than atoms:

- Carbó proposed alignment by the electron density (1980)
- More common to overlay the interaction maps of molecules (e.g., from 3D grids, electrostatics)
  - grid overlays very time consuming
  - overlay of extreme values (maxima, minima, defined field points)

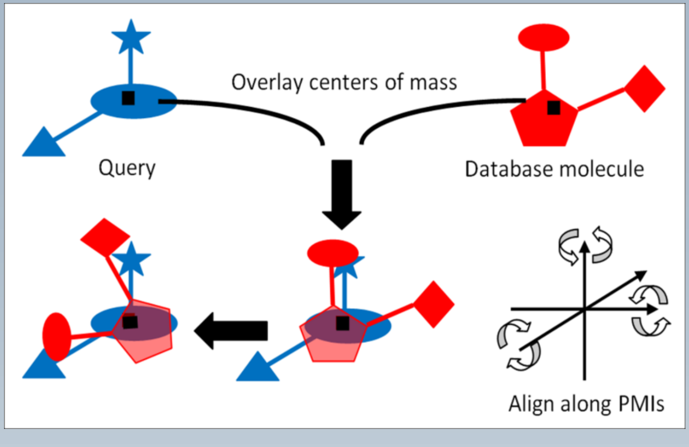


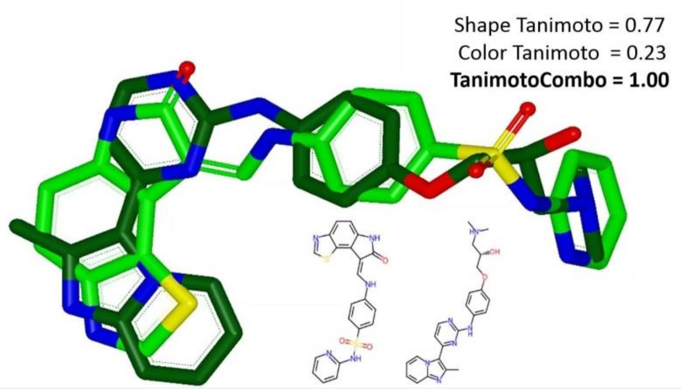
overlay by Cresset software

Molecular similarity
3D similarity - ROCS

Example: 3D overlay procedure by OpenEye

- ROCS (Rapid Overlay of Chemical Structures) is a fast shape comparison application, based on the idea that molecules have similar shape if their volumes overlay well, and any volume mismatch is a measure of dissimilarity
- Volume is Gaussian based rather than hard spheres → overlap quickly computed
- Inputs
  - rigid query molecule
  - database of conformations
- Output: Shape+pharmacophore similarity
- Computation intensive part: generation of the conformations





2D sim: 0.11 (Tanimoto on ECFP-like FP)

Molecular similarity
Evaluation of similarity measures

Comparison of similarity measures

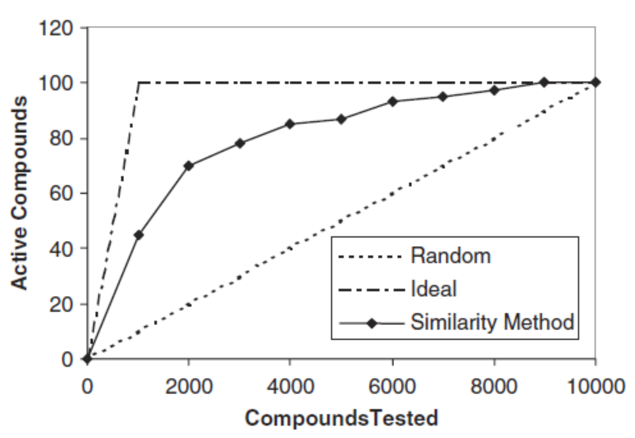
- There is obviously no *a priori* better or worse similarity method
  → the quality depends on the task the index is used for
- Comparison is often done based on the usefulness for identification of compounds of similar properties:
  - Identification of new hits on a drug target (=„virtual screening“, VS)
    - Identification of few bioactive compounds in a huge database (of mostly inactive compounds)
    - Only a small fraction of the database can be experimentally screened, thus the „enrichment“ of hits in the tested set is crucial
    - Datasets with few actives and many random molecules are available as “standard” test cases to ensure fair comparison – e.g., the DUD datasets for docking (http://dud.docking.org/)

    → VS will be treated in more detail later

Molecular similarity
Virtual screening (VS)

Virtual screening (continued)

- Given *n* hits in a database of *N* molecules (n ~ 1000, N~1000000)
- Rank database by similarity to known potent hit
- Probability to draw a bioactive hit randomly = n/N
  Expectation value of hits when drawing m molecules = m*n/N
- VS aims to increase the number of hits in the m molecules above random
- Success of VS is usually measured by its enrichment = hitrate/random hitrate



- Several annotated datasets published which are used for VS evaluation

Molecular diversity

From similarity to diversity

- Similarity measures are introduced
  - Application have been so far: similarity searches
- Why would diversity analysis be important?
  - Missing starting point for similarity searches
  - Coverage of chemical space with few compounds
    → selection for biological testing
    → useful if assays have a low throughput (or are expensive)
  - A diverse subset is assumed to have diverse properties
    → reduce redundancy in a set
  - Important for synthesis planning
    → diverse library members to cover a large feature space

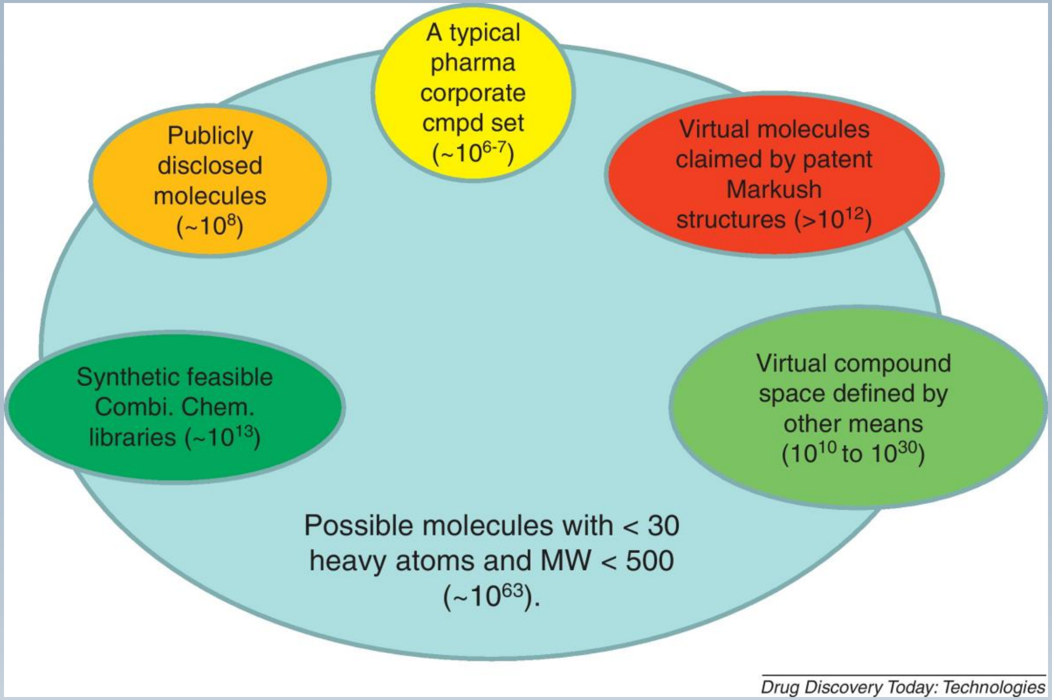
Molecular diversity
Chemical spaces

How many molecules are out there?

- Generally possible molecules
- Molecules described by virtual procedures
- Molecules described virtually with synthesis procedure
- Molecules covered by patent claims
- Molecules synthetized

Which compounds are relevant?

- Purchase (vendors like emolecules, Sigma Aldrich, ...)
- Synthesis feasible (lit, pat, ...)
- (virtual compounds which can serve as templates for similarity searches)


A typical pharma corporate cmpd set ($\sim 10^{6-7}$)
Publicly disclosed molecules ($\sim 10^{8}$)
Virtual molecules claimed by patent Markush structures ($>10^{12}$)
Synthetic feasible Combi. Chem. libraries ($\sim 10^{13}$)
Virtual compound space defined by other means (10^{10} to 10^{30})
Possible molecules with < 30 heavy atoms and MW < 500 ($\sim 10^{63}$).

Drug Discovery Today: Technologies

https://doi.org/10.1016/j.ddtec.2013.01.004

Molecular diversity
Approaches to select diverse compounds

- Brute force enumeration:

  Select a subset of n compounds from a library of N molecules: $\binom{N}{n} = \frac{N!}{n!(N-n)!}$

  → Too high number, procedure not possible for relevant cases (n>10, N>100)
- Approximate selection methods
  - Optimization methods
  - Cluster analysis
    - Hierarchical clustering methods:
      - Agglomerative clustering
    - Non-hierarchical clustering methods:
      - Centroid based (k-means)
      - Density based (DBSCAN)
  - Dissimilarity-based approaches
  - Cell based approaches

Molecular diversity
Optimization methods

- Optimization procedure requires a „diversity function“ with the following properties:
  - Addition of redundant molecule → no increase in diversity
  - Addition of non-redundant molecule → increase in diversity
  - If a molecule is moved away from others → increase
  - Bounded (function values can‘t get infinite)
  - Favoring space filling behavior rather than selecting only outliers

- Typical functions for adding a new member i to a set of m molecules
  - MaxSum: $\sum_{j=1}^{m} D_{i,j}$ ... sum of distance to all members of set
  - MaxMin: $\min\limits_{j=1 \ldots m}(D_{i,j})$ ... closest distance to any member of the set

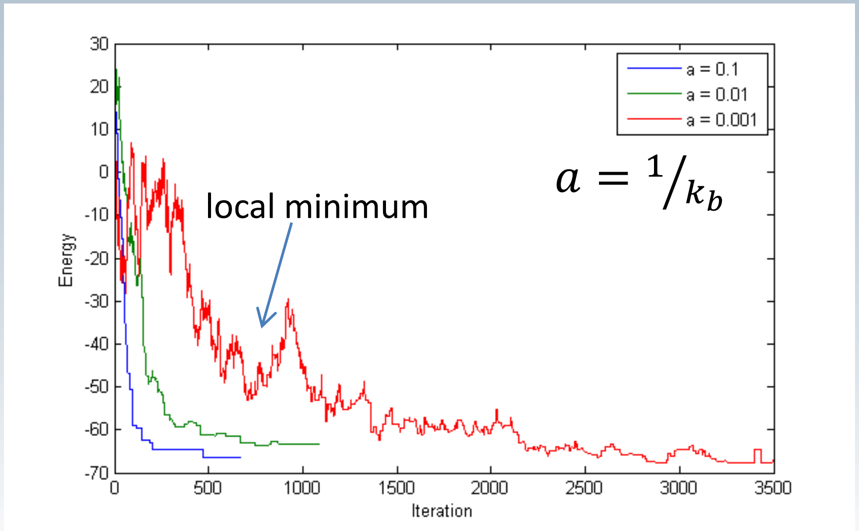
where $D_{i,j}$ is the distance between molecule i and j.

Molecular diversity
Optimization methods

- General procedure has to be a global optimization procedure – like Monte Carlo procedures or genetic algorithms
- Example: Simulated Annealing (a Monte Carlo procedure)
  - Select initial subset of n (out of N) molecules – often random selection
  - Modify the subset and evaluate diversity function
    - if diversity increased: accept modified set
    - if diversity did not increase: accept modified set with following probability: $\exp\left({}^{-\Delta E}/_{k_B T}\right)$ - where ΔE corresponds to the change in diversity, k_B is a scaling factor and T is called temperature (eventually resembling a Boltzmann factor)
  - Lower temperature T and iterate

→ Simulated annealing results highly dependent on the parameter k_B and the temperature schedule

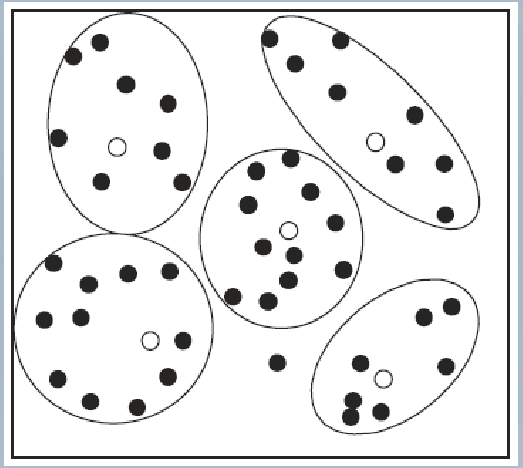
→ can overcome local minima



local minimum

$a = {}^{1}/_{k_b}$

Molecular diversity
Cluster analysis

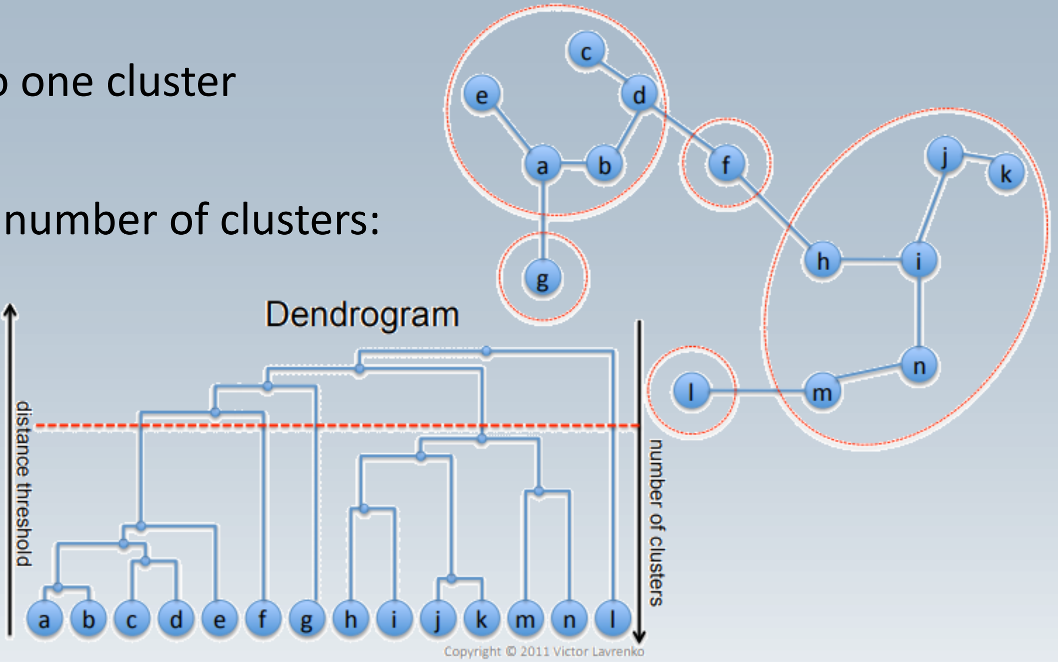


- Cluster?
  - Objects in a cluster are similar
  - Objects from other clusters are dissimilar
- Choosing representative set
  - small number of representatives (often n=1) from each cluster
- Several approaches
  - Connectivity-based clustering (hierarchical clustering)
  - Centroid-based clustering
  - Distribution-based clustering
  - Density-based clustering
- General procedure
  - Calculate descriptors
  - Determine similarity/distance between molecules
  - Group compounds to clusters
  - Select cluster reps

Molecular diversity
Hierarchical clustering

- Objects are connected into clusters based on their distance
- Compounds relations are visualized by a dendrogram
- Example: Agglomerative Clustering
  - start from the bottom (single compounds)
  - identify pair of closest clusters
    and merge to new cluster
  - iterate until all cpds belong to one cluster

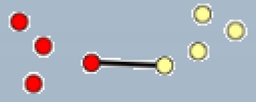
- Level of hierarchy corresponds to number of clusters:
  How to select?
  - Visually
  - Variance Ratio Criterion
    variances within/between clusters
  - Kelley criterion
    balances spread at a particular
    level with the number of clusters



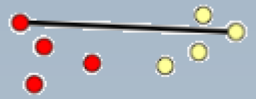
Copyright © 2011 Victor Lavrenko

Molecular diversity
Hierarchical clustering – distance measures

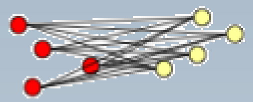
Cluster distance measures



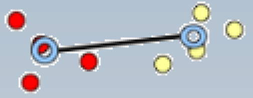
single linkage:
distance of **closest** elements



complete linkage:
distance of **most distant** elements

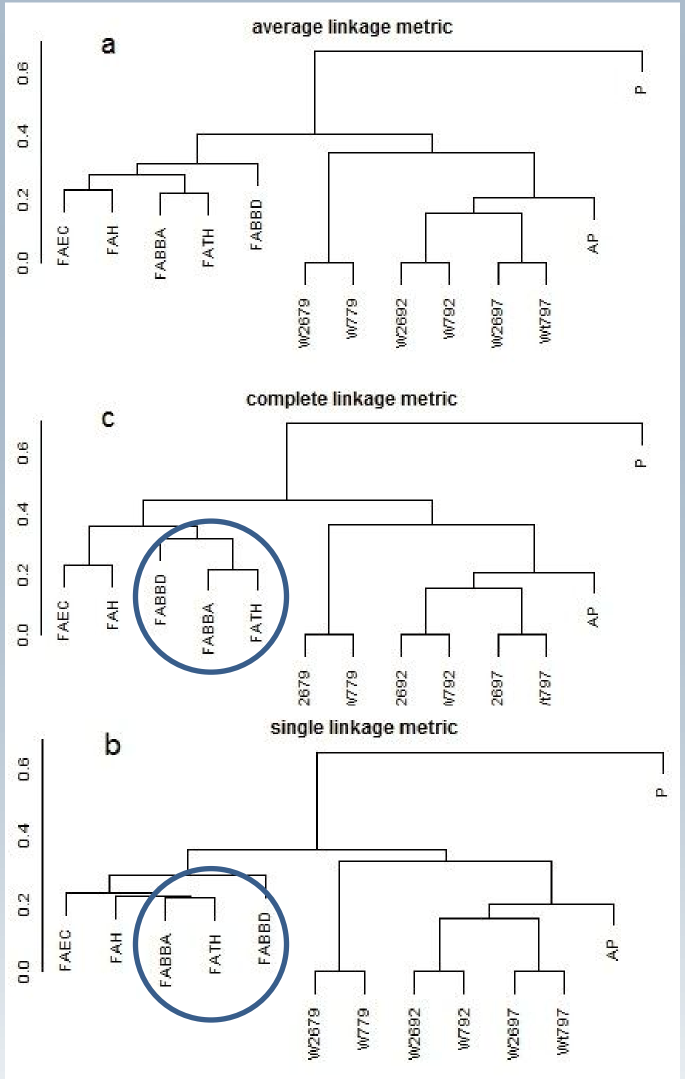


average linkage:
average of pairwise distances



centroids:
distance between the **means**

→ Distance measures influence the clustering outcome:



BMC Bioinformatics 2013, 14:351

Molecular diversity
Non-hierarchical clustering

No hierarchical relationship between clusters

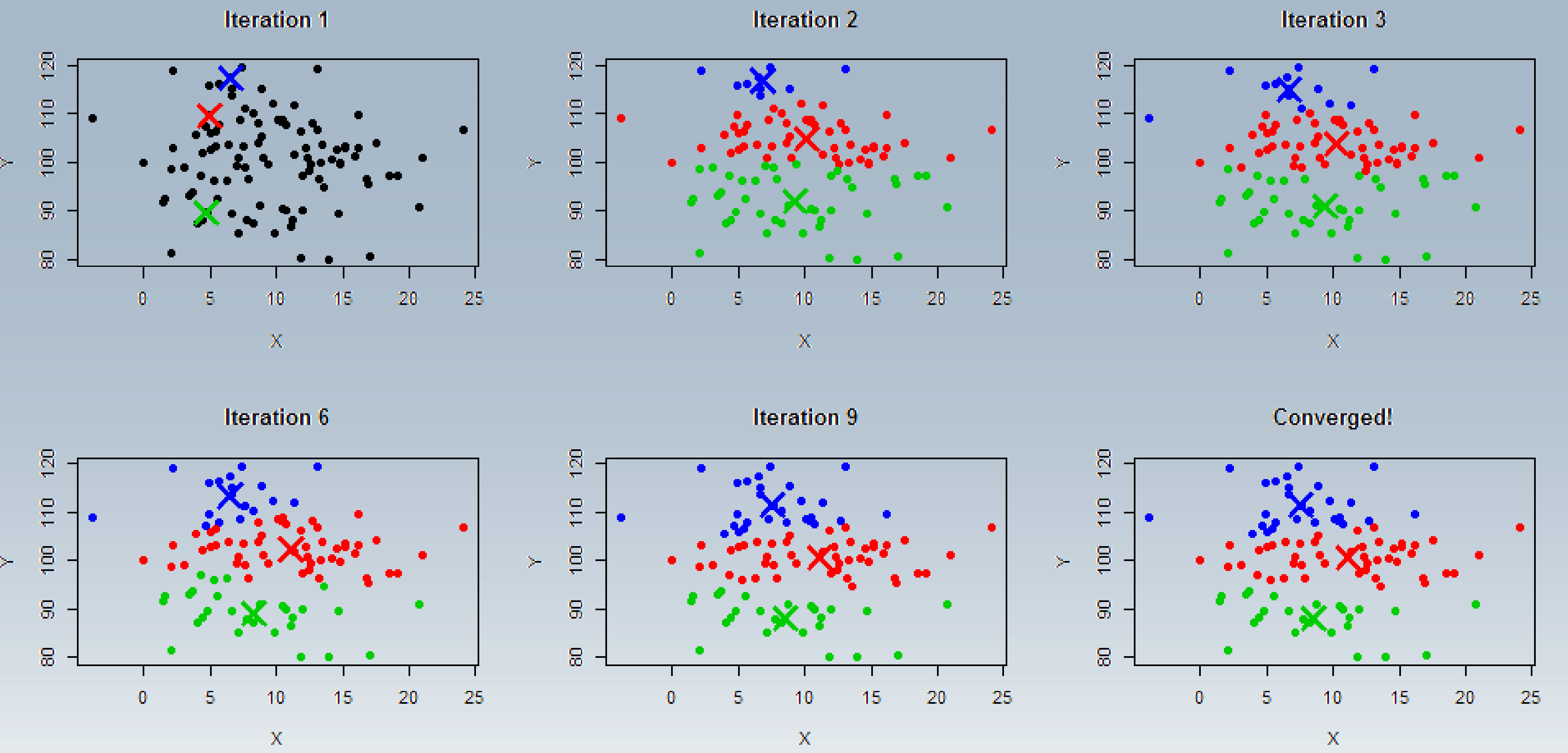
- Single pass methods
  only a single pass through dataset: very fast, but dependent on compound order
  → a pre-defined similarity decides if a new element is put in an existing cluster or if it is assigned to a new cluster.
- Nearest neighbor method (1973)
  - Determine matrix of pairwise distances of all compounds
  - Cluster based on amount of common neighbors

  → often leads to large clusters and many singletons (modified versions available)
- k-means clustering: The number of clusters k has to be pre-defined
  - Select k „seeds" = starting molecules (e.g. random)
  - Assign all remaining compounds to the closest seed
  - Calculate the centroid of the clusters and reassign all compounds to the nearest centroid
  - Iterate

  → Non deterministic (check for stability by clustering with different seeds)

Molecular diversity
k-means clustering - example

- Example for the convergence of k-means clustering



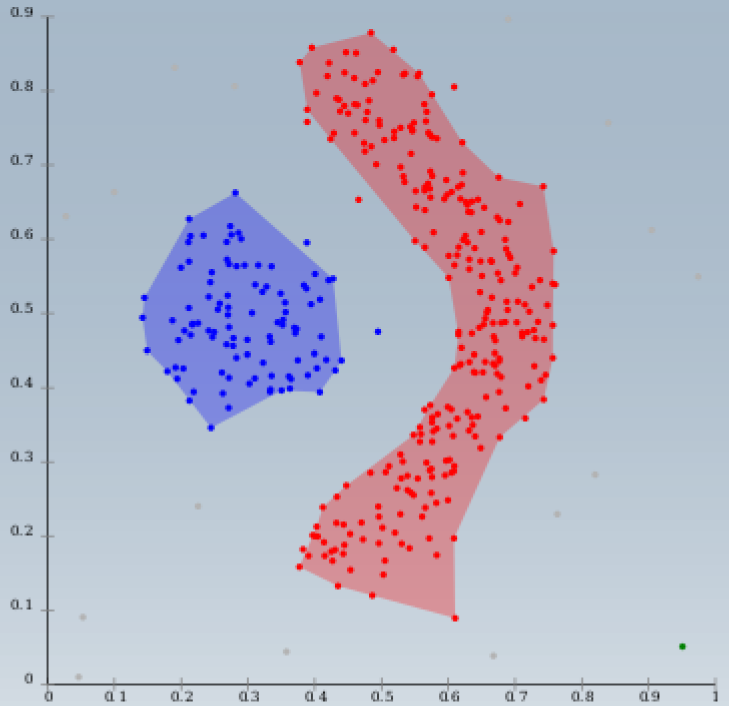
Molecular diversity
Density based clustering

Clusters are defined as areas of higher density than the remainder of the data set

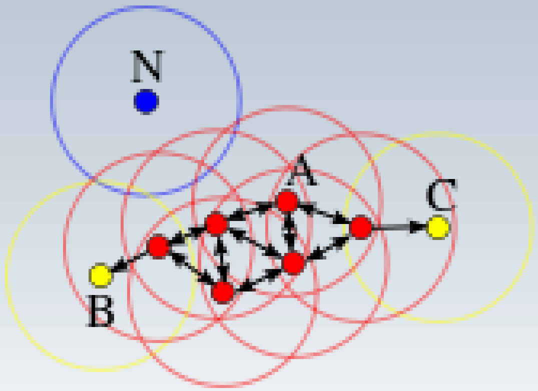
DBSCAN popular algorithm (density-based spatial clustering of applications with noise)

- an object is defined as "core" if at least ***minPts*** points are within distance ***ε***
- objects are „directly reachable" from core points if the distance < *ε*
- 2 objects are „density connected" if there is a chain of core objects connecting them

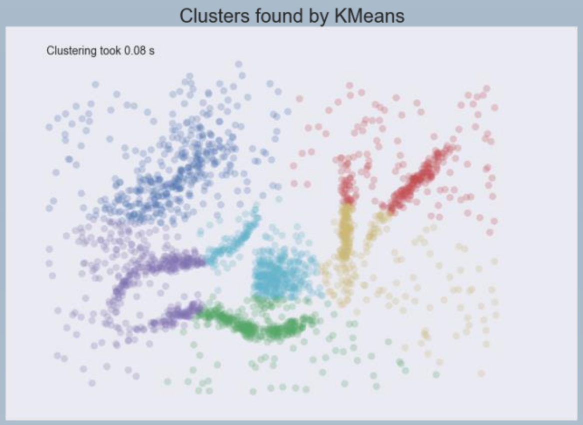
→ a core point forms a cluster with all objects that are density connected to it.



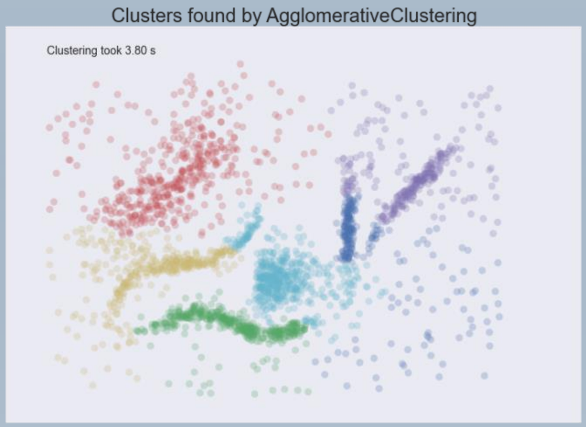
By Chire - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17085332



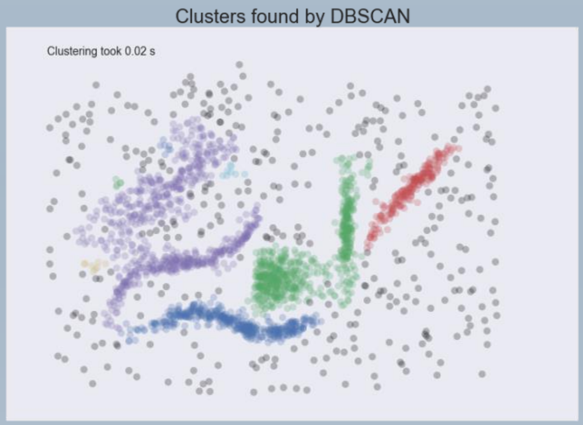
Comparison of clustering algorithms


Clusters found by KMeans

k-means (6 clusters)


Clusters found by AgglomerativeClustering

hierarchical


Clusters found by DBSCAN

density based

- k-means also includes outliers and separates obvious clusters
- DBSCAN clustering much more flexible with cluster-shapes, but many unclustered elements

→ all algorithms need parameters which have to be adjusted accordingly

http://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html

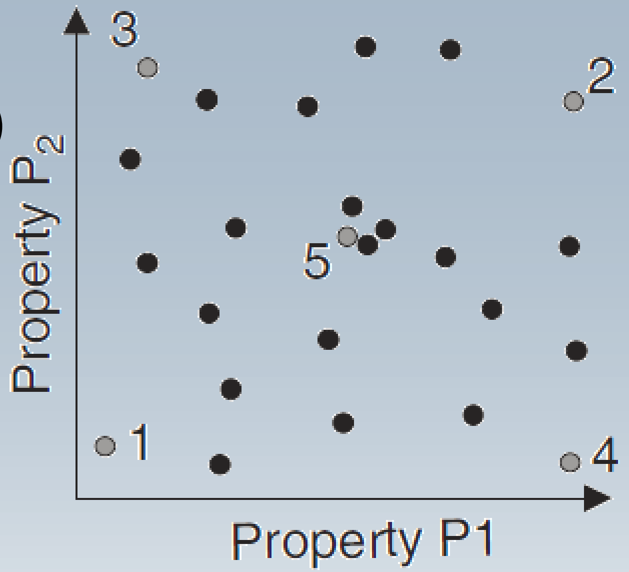
Molecular diversity
Dissimilarity based selection methods

What‘s the difference to clustering?

- Clustering methods first group elements into a cluster and then a subset is chosen
- Dissimilarity-based compound selection (DBCS) attempt to identify a diverse set of compounds directly

General steps of DBCS

- Select the first compound (e.g. random or centroid)
- Calculate the dissimilarity to the rest
- Choose most dissimilar compound (e.g. MaxMin or MaxSum scores)
- Iterate until sufficient compounds collected



→ Results strongly depend on initial compound and how „dissimilarity“ is calculated

Molecular diversity
Sphere exclusion algorithm

Steps of Sphere exclusion:

- Define a threshold dissimilarity parameter t
- Select first compound (e.g. random, mean) and move into subset
- Remove all objects with a dissimilarity < t to selected molecule
- Select new compound and iterate (e.g. choose as closest or as most distant object)

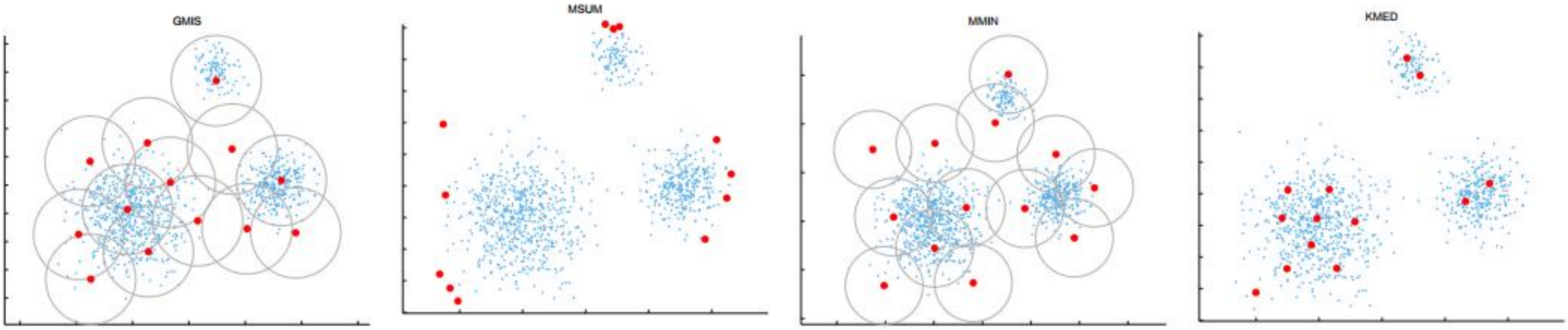
→ usually DBCS results more diverse than sphere exclusion results (but sphere exclusion gives subsets which represents the data)

Sphere exclusion

MaxSum

MaxMin

k-medioids
(similar to k-means)



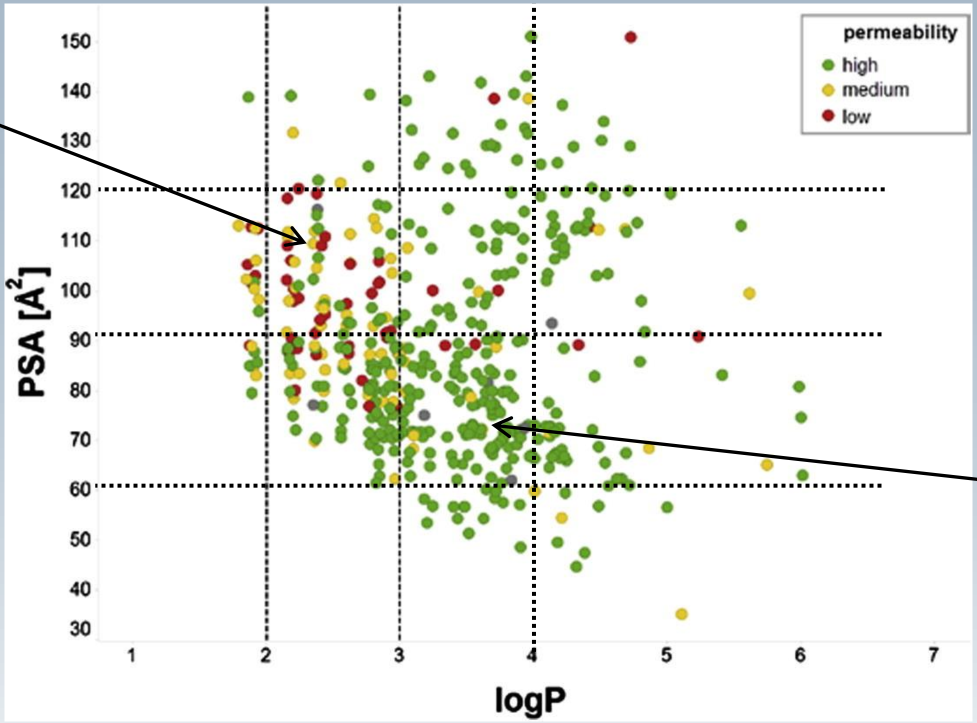
M. Drosou, Proceedings of the VLDB Endowment 6(1)

Molecular diversity
Cell based methods

Difference to clustering and DBCS methods:

- Cell based methods require a pre-defined binned descriptor space e.g.: clogP binned in 4 bins: $(-\infty, 2], (2,3], (3,4], (4, \infty)$
- Low number of dimensions possible (3-5, PCA)

on average low permeability



on average high permeability

https://doi.org/10.1016/j.bmcl.2016.10.069

Molecular diversity
Cell based methods

Key features of cell based approaches

- No pairwise distance calculation required
- Chemical space defined independently of molecules
- Density of molecules per cell easily computed
  - under-represented chemical spaces identified by empty bins (some combinations often not feasible like clogP↑ with H-donor count↑)
  - comparison of different subsets trivial through density comparison per bin
- Drawback is the low dimensionality (number of cells = $\prod_{i=1}^{dim} bins_i$)
  - PCA helps to reduce dimensions → use first 3-4 PCs (makes the method again dependent on the dataset)
- Diverse compound sets straightforwardly generated
  e.g. select one representative from each cell

Molecular diversity
Cell based methods – BCUT descriptors

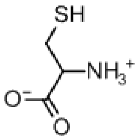
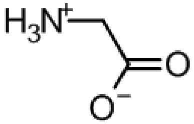
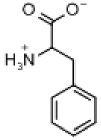
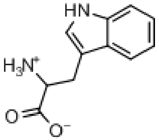
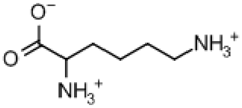
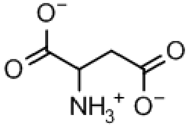
BCUTs are descriptors to generate a low dimensional space

- BCUT descriptors are based on matrix representations of molecules connection tables
  - off-diagonals related to b_{ij} represent the bond order between atoms i and j (BCUT) or the graph distance between atoms i and j (GCUT)
  - diagonal elements p_i correspond to atomic properties like the partial atomic charge of atom i or the atoms contribution to logP (BCUT_SLOGP) or other atomic properties
  - the descriptors themselves are the *eigenvalues* of the matrices above → highest/lowest or other distinctly defined eigenvalues

$$\begin{pmatrix} p_1 & \cdots & \frac{1}{\sqrt{b_{m1}}} \\ \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{b_{1m}}} & \cdots & p_m \end{pmatrix}$$

BCUTs depend on atom properties and connectivity in molecules ...

Molecular diversity
BCUT descriptors

	name	BCUT_PEOE_0	BCUT_PEOE_1	BCUT_PEOE_2	BCUT_PEOE_3
	Cys	-2,503	-0,548	0,501	2,530
	Gly	-2,379	-0,544	0,345	2,370
	Phe	-2,525	-0,584	0,592	2,548
	Trp	-2,516	-0,562	0,599	2,552
	Lys	-2,654	-0,544	0,345	2,662
	Asp	-2,513	-0,560	0,420	2,541

BCUT_PEOE's are calculated from the eigenvalues of a modified adjacency matrix. Each *ij* entry of the adjacency matrix takes the value 1/sqrt(b_{ij}) where b_{ij} is the formal bond order between bonded atoms *i* and *j*. The diagonal takes the value of the PEOE partial charges. The resulting eigenvalues are sorted and the smallest, 1/3-ile, 2/3-ile and largest eigenvalues are reported.

Molecular similarity and diversity
Summary

- Similarity of molecules is calculated based on features (descriptors)
  Can be 2D or 3D based – conformation/overlays have to be considered
- The definition of similarity is not unique – several similarity measures exist
  Depending on the task the appropriate sim. measure is selected
- Similarity is often used for virtual screening
  The assumption is that similar molecules bind the target in a similar way

- Dissimilarity is used to identify diverse subsets of compound collection
  This can be achieved by
  - clustering
  - dissimilarity based compound selections
  - global optimization of dissimilarity
  - cell based methods
- The subset is often dependent on the starting points as well as the definition of distance and/or bins
  → similarity and diversity often a very subjective property