Molecular Descriptors

Dragos Horvath, Gilles Marcou, Alexandre Varnek

Laboratoire de ChemoInformatique, UMR 7140

CNRS – Université de Strasbourg

67000 Strasbourg, France

dhorvath@unistra.fr

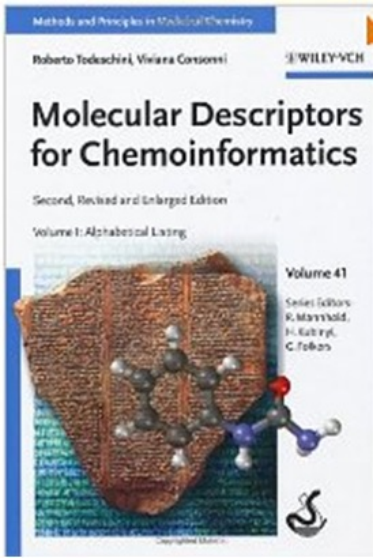
Molecular Descriptors or Fingerprints

- Need to represent a structure by a **characteristic** bunch (**vector**) of numbers (**descriptors**).
  - Example: (Molecular Mass, Number of N Atoms, Total Charge, Number of Aromatic Rings, Radius of Gyration)
- Should include **property-relevant** aspects:
  - the "**nature**" of atoms, including information on their **neighbor-hood-induced properties**, and their **relative arrangement**.
  - Number of N Atoms ⇔ (Primary Amino Groups, Secondary Amino Groups, … , … , Amide, … , Pyridine N, …)
  - **…** unless being a **H bond acceptor** is the key (O or N alike)!
  - Arrangement in **space** (**3D**, conformation-dependent distances in Å) or in the **molecular graph** (**2D**, topological distance = separating bond count)

Definition of 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.

Roberto Todeschini and Viviana Consonni



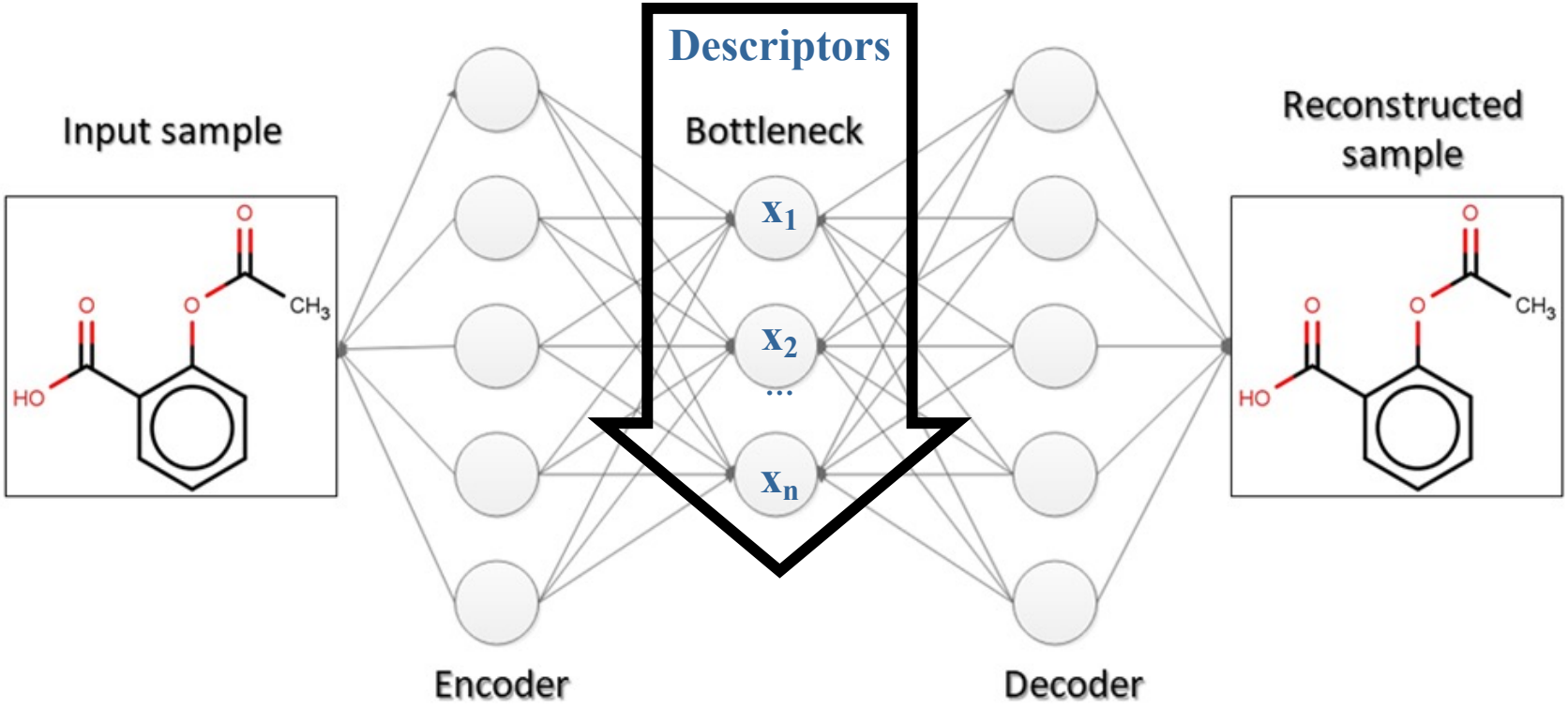
Molecular Descriptors

Classification based on the origin of descriptors

- "experimental"
  - logP, aqueous solubility, Abraham's H-bond parameters, solvent parameters NMR shift, …. Often *predicted* by computer models

- **calculated**
  - **Assessed *in Silico* from 1D, 2D or 3D molecular structure**
  - **Expert-designed… or AI-designed!**

Advertising: Position of Molecular Descriptor Designer. *Humans need not apply!*



- An AutoEncoder/Decoder is a Deep Neural Network producing an efficient dense representation of the input, by performing specific compression of learned data.
- The states of Bottleneck Neurons fully characterize the object!
- It's *reversible*: provide *any* vector (x_1, x_2, …, x_n) and the Decoder will return a chemical structure associated to those coordinates…

Molecular Descriptors

Classification based on described object

- Global

describing the whole molecule (molecular volume, molecular surface, dipole moment, topological indices, …)

- **Local**

describing particular atoms or molecular fragments (atomic charges, bonds polarizabilities, CATS descriptors, ISIDA descriptors, …)

- **Field**

describing molecular fields in the area surrounding the molecule (electrostatic potential, COMFA descriptors, …)

Molecular Descriptors

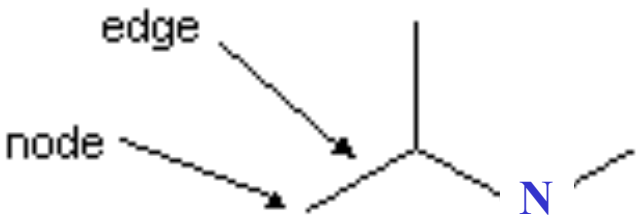
Classification based on the dimensionality of structure representation

- 1D: constitutional descriptors: atom & bond counts, MW
- **2D: based on molecular topology: topological indices, fragment counts**
- **3D: geometrical parameters: molecular surfaces & fields, parameters calculated in quantum chemistry programs**

2D

2D - Topological Descriptors

Molecular colored graph



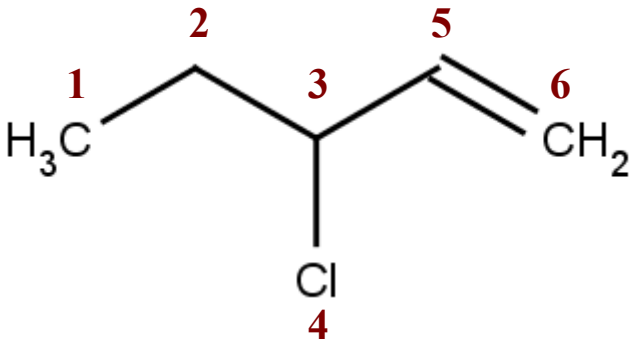
Descriptors based on the molecular graph representation are widely used because they incorporate precious chemical information:

- size,
- degree of branching,
- neighborhood of atoms → electronic & steric effects,
- flexibility
- overall shape,

Matrix representations

A molecular structure with *n* atoms may be represented by an n × n matrix (H atoms are often omitted).

Adjacency matrix : indicates which atoms are bonded.
Bond order matrix : adjacency + bond orders.

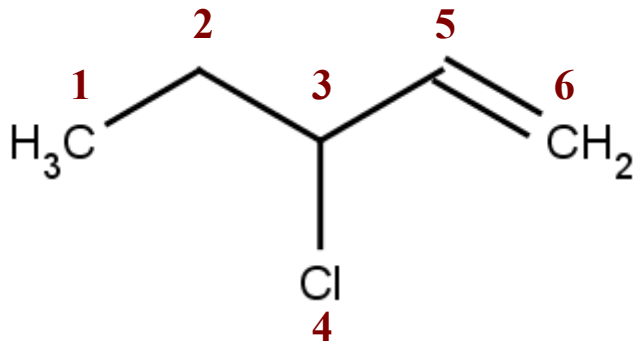


	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	0	**2**
6	0	0	0	0	**2**	0

Matrix representations

Distance matrix : encodes the distances between atoms.

Topological distance is defined as the number of bonds between atoms on the shortest possible path.

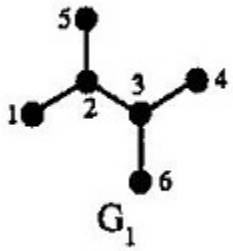


	1	2	3	4	5	6
1	0	1	2	3	3	4
2	1	0	1	2	2	3
3	2	1	0	1	1	2
4	3	2	1	0	2	3
5	3	2	1	2	0	1
6	4	3	2	3	1	0

It is a cheap and robust alternative to actual geometric distances, in Å

TI based on the adjacency matrix :

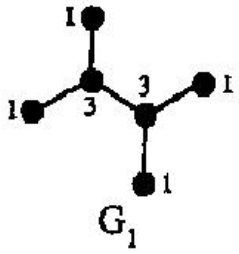
Zagreb group indices



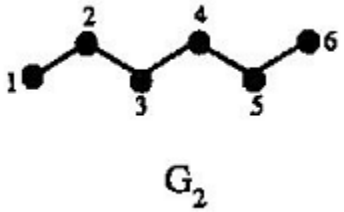
G_1

$A(G_1) =$

	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	1	0
3	0	1	0	1	0	1
4	0	0	1	0	0	0
5	0	1	0	0	0	0
6	0	0	1	0	0	0



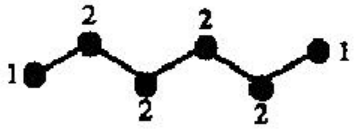
G_1



G_2

$A(G_2) =$

	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	0	0
4	0	0	1	0	1	0
5	0	0	0	1	0	1
6	0	0	0	0	1	0



G_2

- $\mathbf{M_1} = \sum_{i=1}^{n} \delta_i^2 \quad \mathbf{M_2} = \sum \delta_i \, \delta_j$

where the *vertex degree* δ_i is a number of σ bonds involving atom *i* excluding bonds to H atoms.

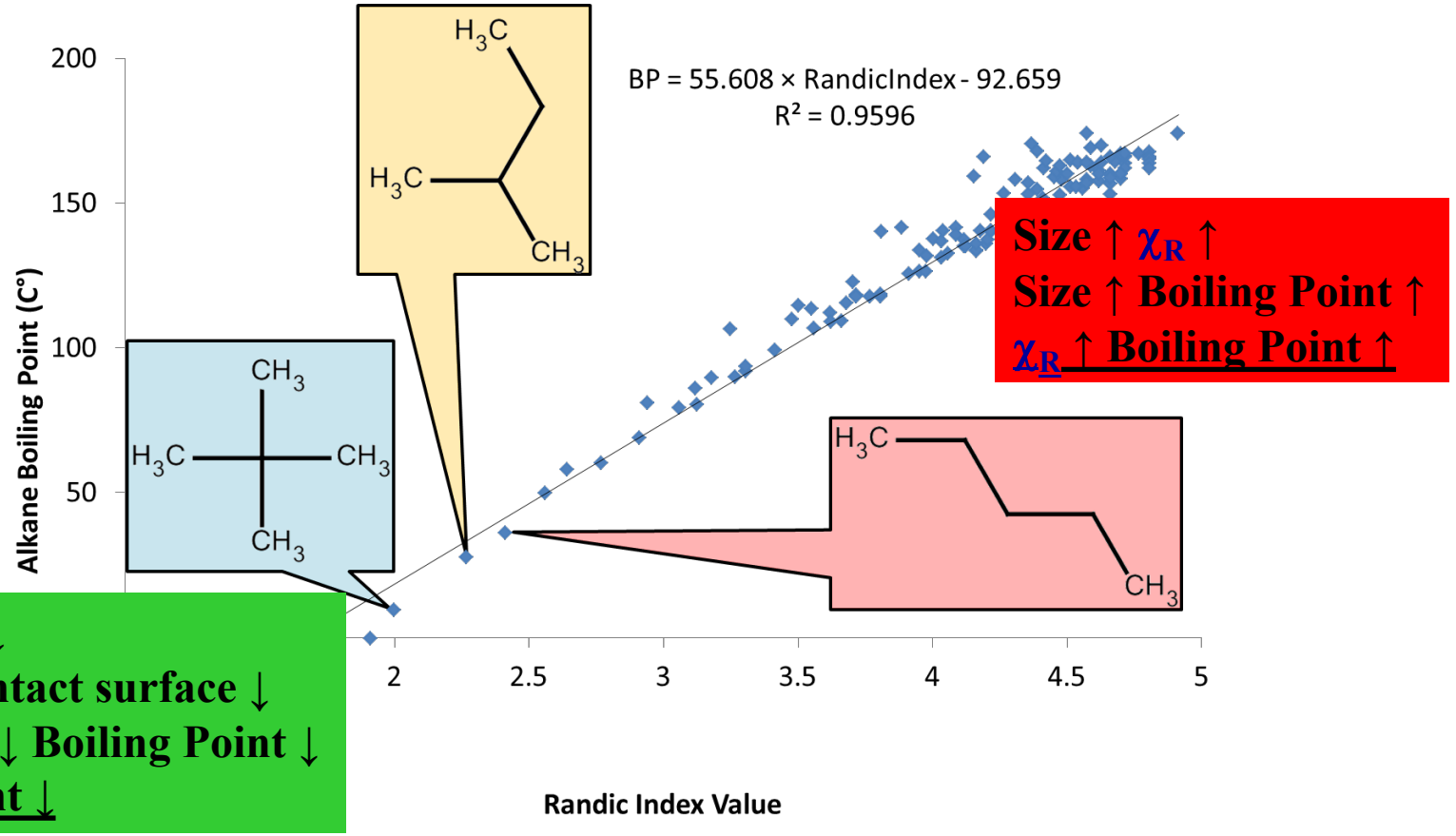
Zagreb group indices were introduced to characterize branching

So why should an obscure topological formula explain chemical properties?

Randic introduced a *connectivity index* similar to M_2

M. Randić, *J. Am. Chem. Soc.*, 97, 6609 (1975)

$$\chi_R = \sum (\delta_i \, \delta_j)^{-1/2}$$

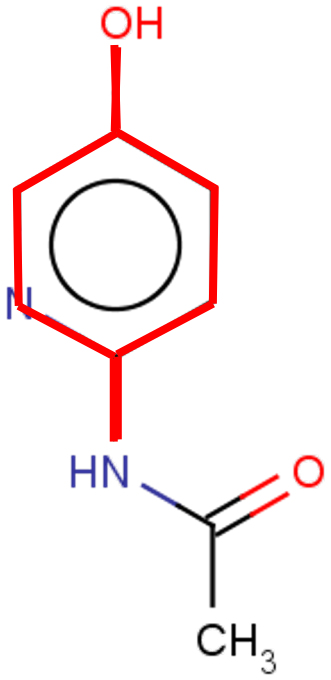


Size ↑ χ_R ↑
Size ↑ Boiling Point ↑
χ_R ↑ Boiling Point ↑

Branching ↑ χ_R ↓
Branching ↑ Contact surface ↓
Contact surface ↓ Boiling Point ↓
χ_R ↓ Boiling Point ↓

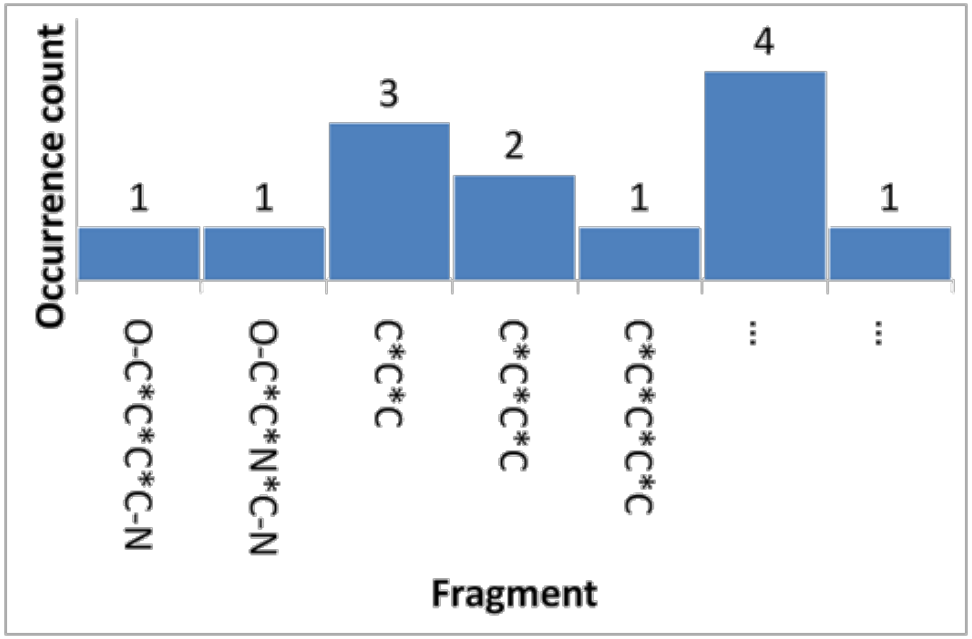
Capturing Topology by Fragment Counts

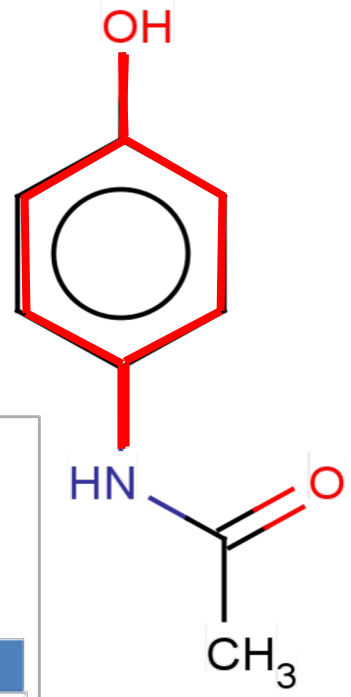
ISIDA Fragmentor, Laboratoire de Chemoinformatique Strasbourg,
http://infochim.u-strasbg.fr/spip.php?rubrique49



(1,1,…)

O-C*C*C*C-N	**1 1**
O-C*N*C*C-N	**1 0**
…	**…**

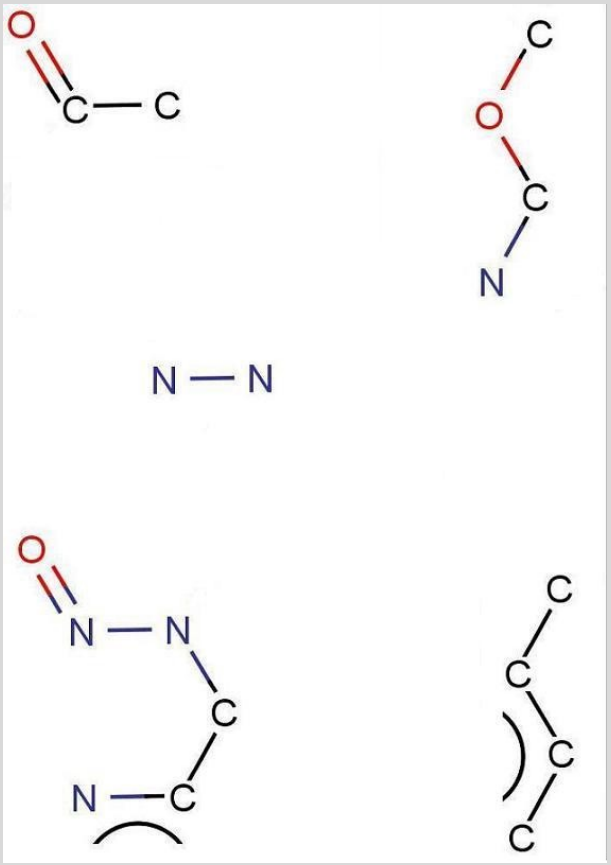




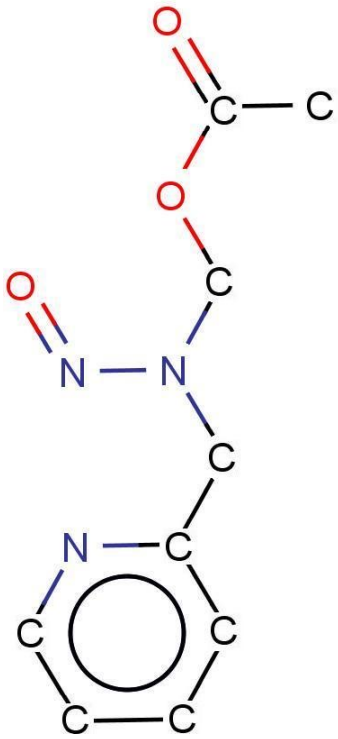
(2,0,…)

ISIDA fragments

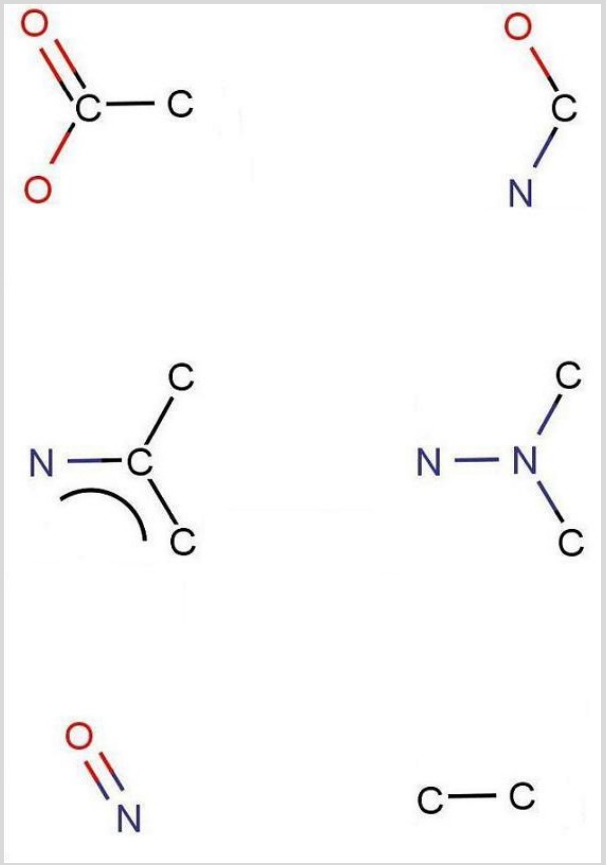
Sequences
containing $2 < N < 15$ atoms

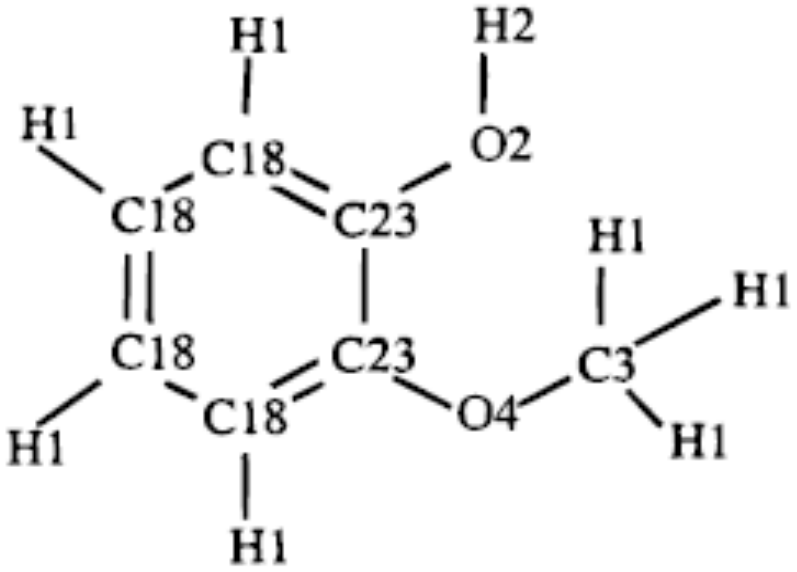


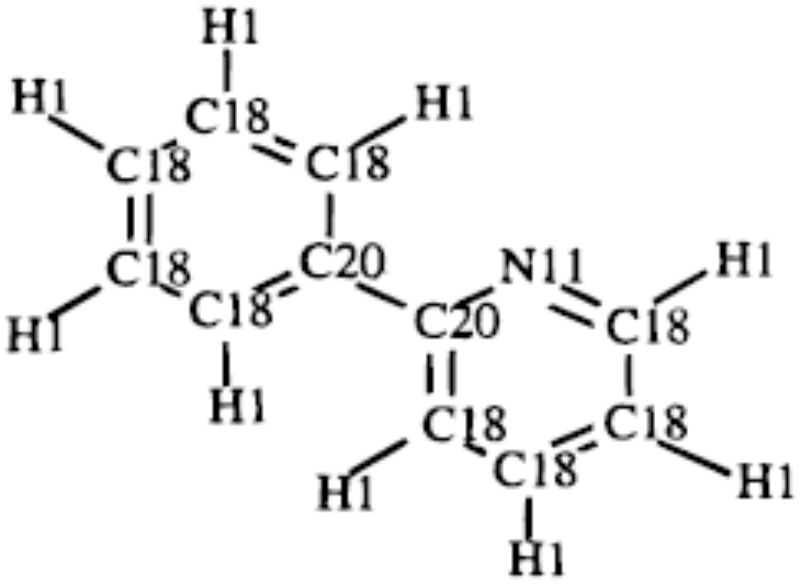
atoms and bonds



Augmented Atoms:
selected atoms with their closest neighbours



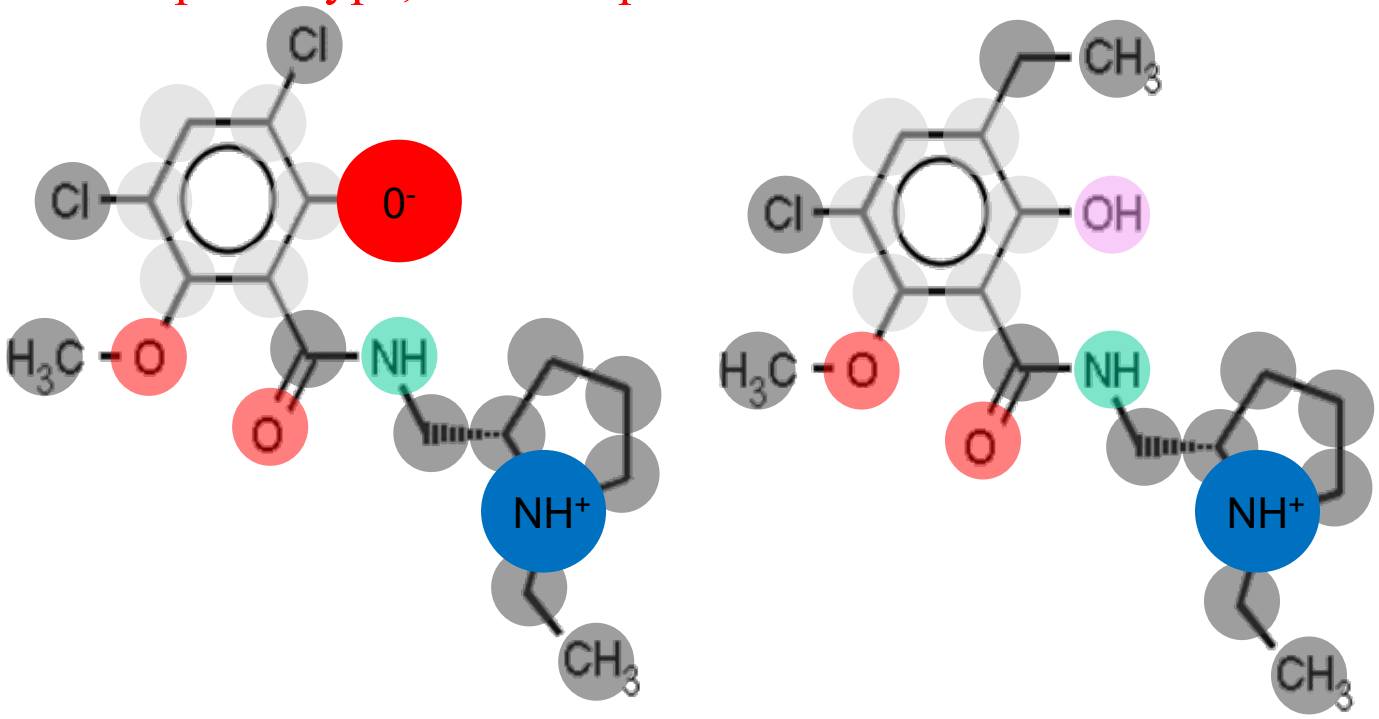




type	log *P*	MR	type	log *P*	MR
C3	−0.2035	2.753	9 × C18	0.1581	3.350
4 × C18	0.1581	3.350	2 × C20	0.2713	3.904
2 × C23	0.5437	3.853	9 × H1	0.1230	1.057
7 × H1	0.1230	1.057	N11	−0.3239	2.202
H2	−0.2677	1.395	calcd	2.75	50.39
O2	−0.2893	0.8238	expt	2.63	49.67
O4	−0.4195	1.182			
calcd	1.40	34.66			
expt	1.32	34.66			

Chemical Relevance: 1. - Go beyond the obvious information in the graph

- Are these compounds nearly identical?
  - Yes, if you mechanically check the "brute" graph
  - No, if you "color" their graphs by relevant chemical properties – pharmacophore type, for example

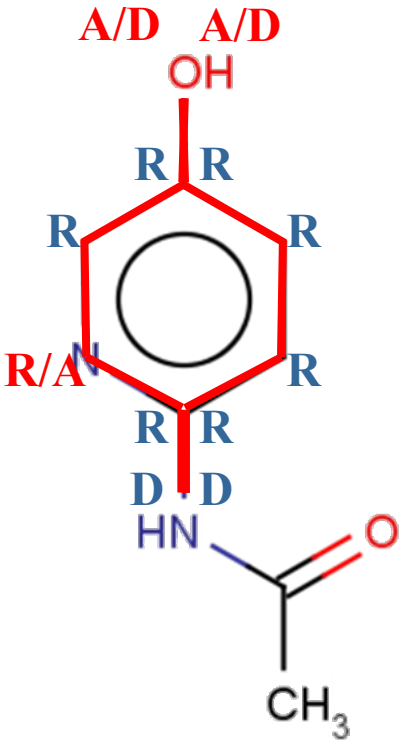


Note – the information you need to do the coloring is contained in the graph too: it' s 2D!

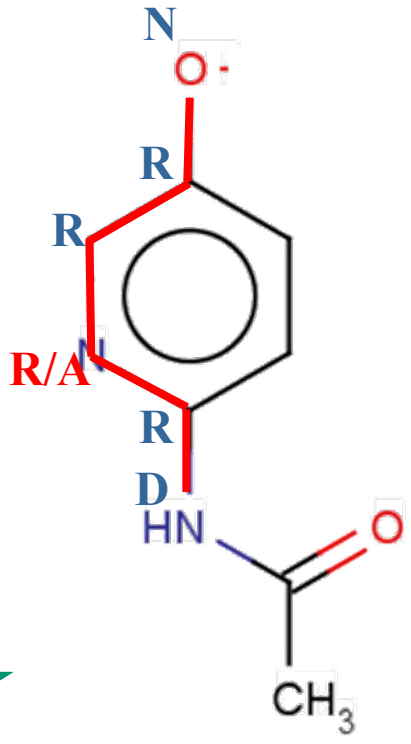
ChemAxon pKa plugin: https://docs.chemaxon.com/display/docs/pKa+Plugin

pH-dependent Labeling of ISIDA Pharmacophore Fragments…

MicroSpecies increment counters of contained fragments by their population levels

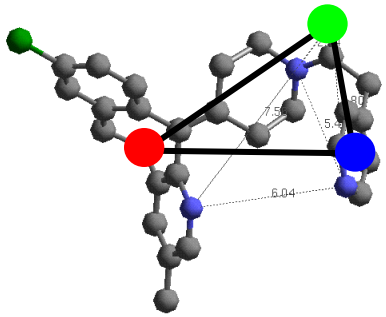


A-R*R*R*R-D	+95
D-R*R*R*R-D	+95
A-R*R*R*R-D	+95
D-R*R*R*R-D	+95
A-R*R*A*R-D	+95
D-R*R*A*R-D	+95
…	
N-R*R*R*R-D	+5
N-R*R*A*R-D	+5
…	



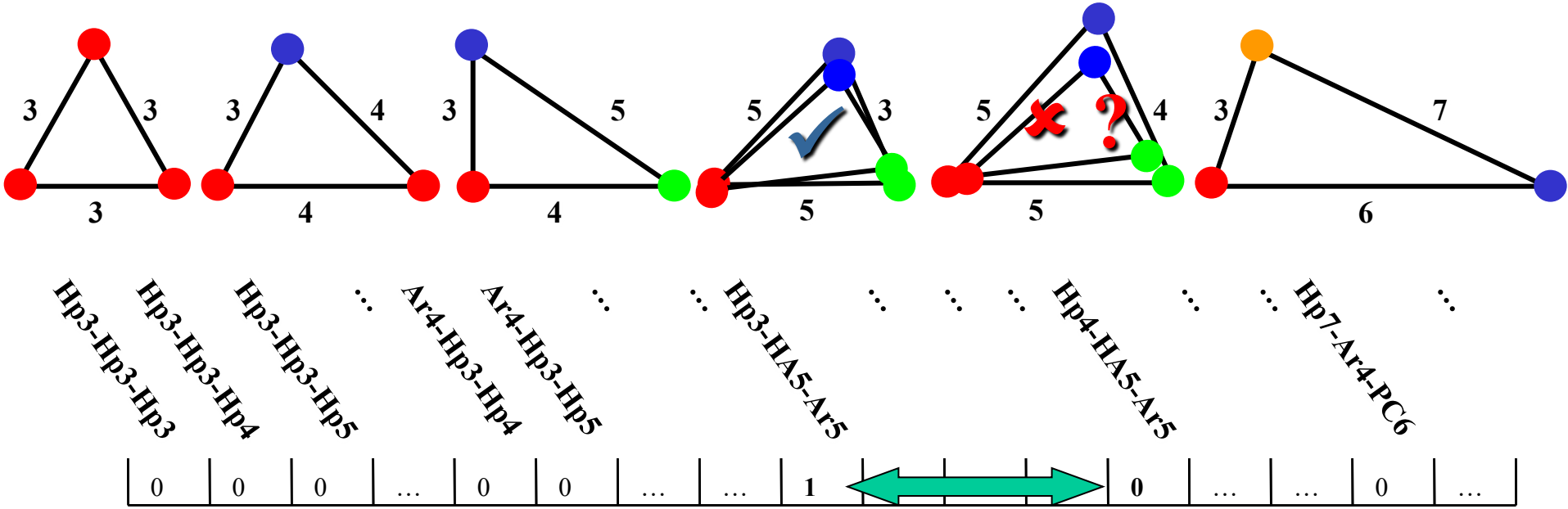
Population: 95% **Molecular Fingerprint** **5%**

Chemical Relevance: 2 - Mother Nature is fuzzy – what about our descriptors? The Triplet Case



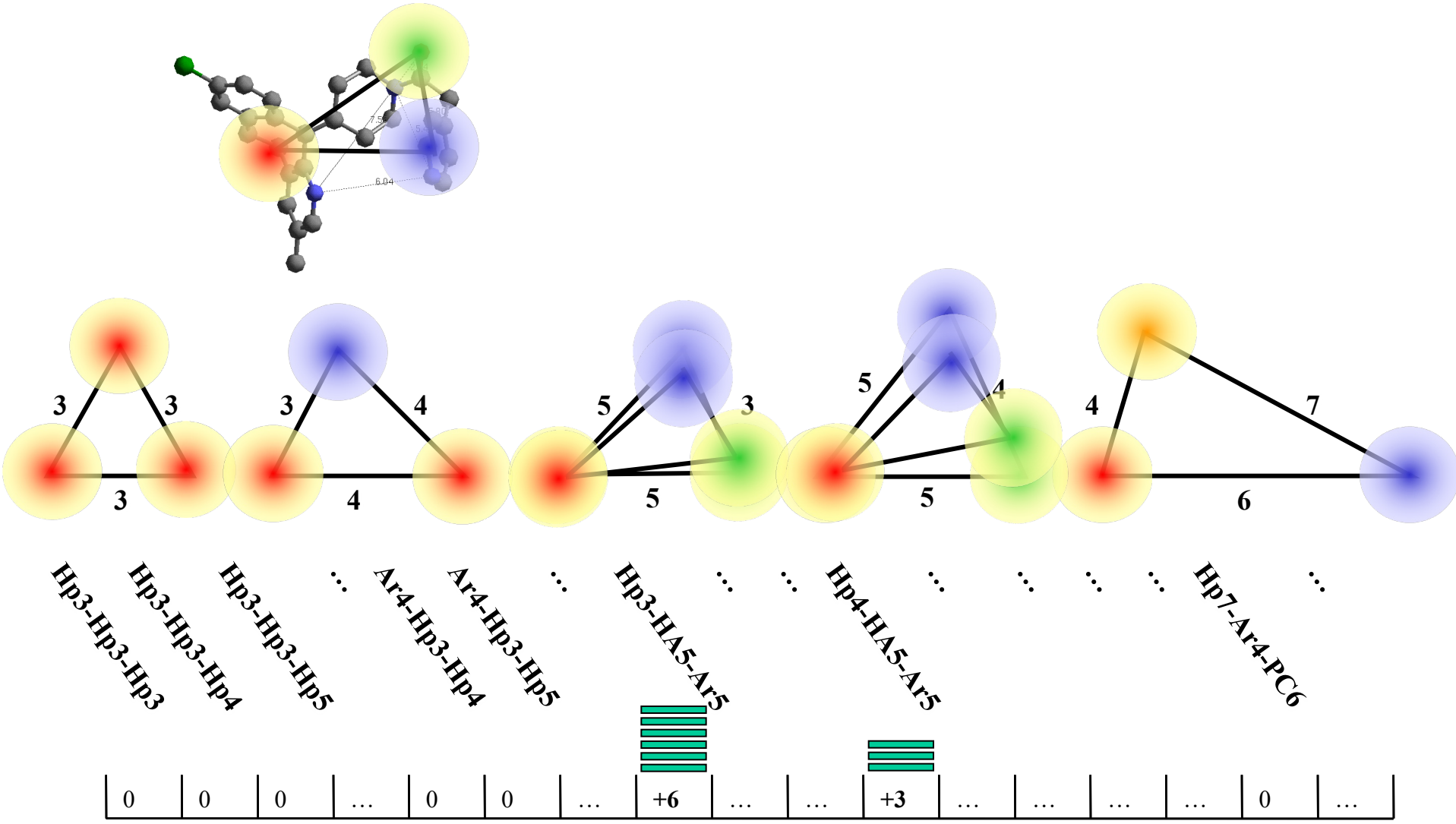
Basis Triplets:

- all possible feature combinations
- at a given series of distances…



Pickett, Mason & McLay, *J. Chem. Inf. Comp. Sci.* 36:1214-1223 (1996)

Fuzziness – blurring the bin borders…



$D_i(m)$ = total occupancy of basis triplet i in molecule m.


3D

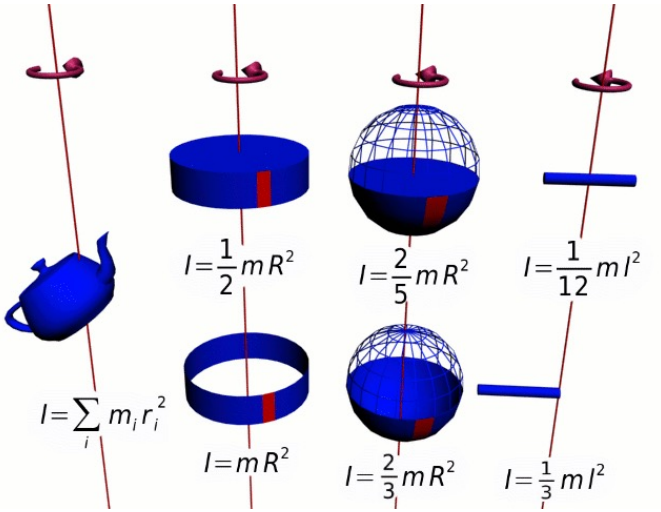
Quantum Chemical Descriptors

Quantitative values calculated in QUANTUM MECHANICS (semi-empirical, HF *Ab Initio* or DFT) calculations

- **LUMO** - Lowest occupied molecular orbital energy
- **HOMO** - Highest occupied molecular orbital energy
- **DIPOLE** moment
- Components of dipole moment along inertial axes (D_x, D_y, D_z)
- **Hf** - Heat of formation
- **Mean Polarizability** - $\alpha = 1/3(\alpha_{xx}+\alpha_{yy}+\alpha_{zz})$
- **EA** – Electron Affinity
- **IP** – Ionization Potential
- **ΔE** – Energy of Protonation
- **Electrostatic Potential** -

$$V(r) = \sum_A \frac{Z_A}{|R_A - r|} - \int \frac{\rho(r')dr'}{|r'-r|}$$

Geometric Indices



Moments of inertia

(value of the moment, principal components)

- The moments of inertia characterize the mass distribution in the molecule

$$I = \sum_i m_i d_i^2$$

Inertia matrix ⇒ $\begin{vmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & I_3 \end{vmatrix}$ **principal moments of inertia**

Radius of gyration

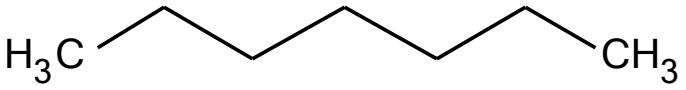
$$Rog = \sqrt{\left(\sum \frac{\left(x_i^2 + y_i^2 + z_i^2\right)}{N} \right)}$$

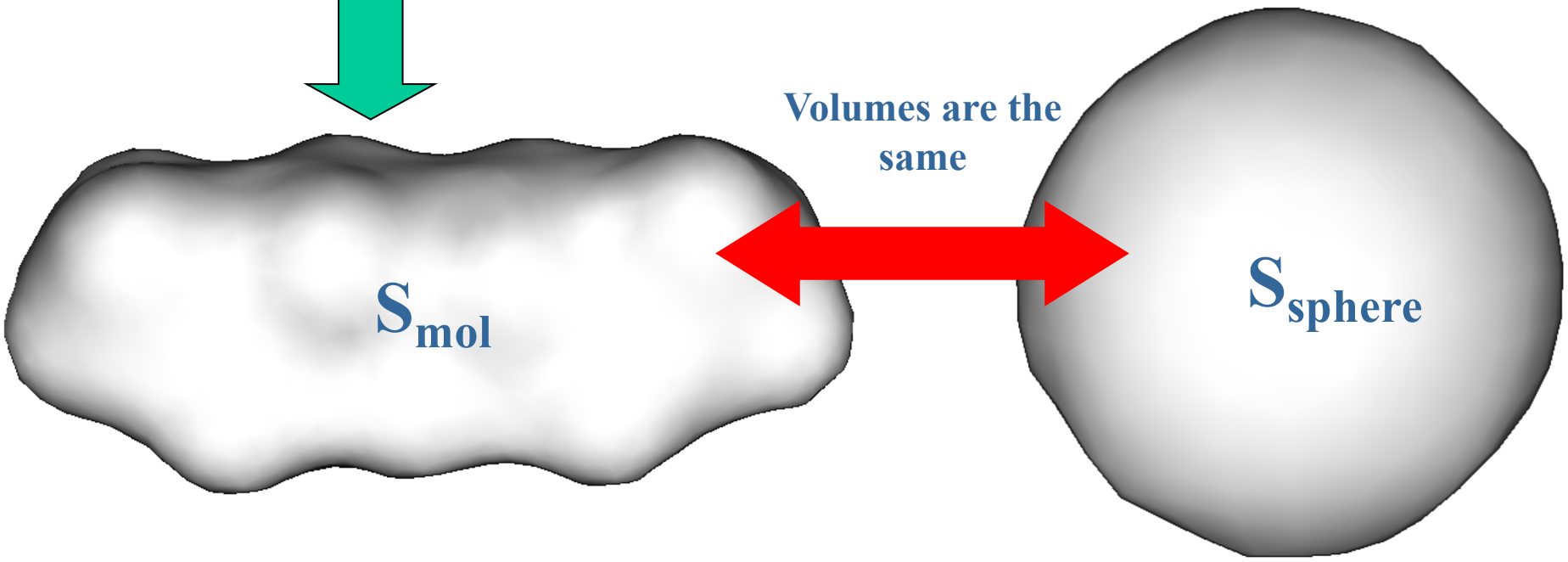
N : number of atoms

x, y, z : the atomic coordinates relative to the center of mass

Ovality

$$Ovality = \frac{S_{mol}}{S_{sp}} = \frac{S_{mol}}{4\pi\left(\frac{3V_{mol}}{4\pi}\right)^{2/3}}$$

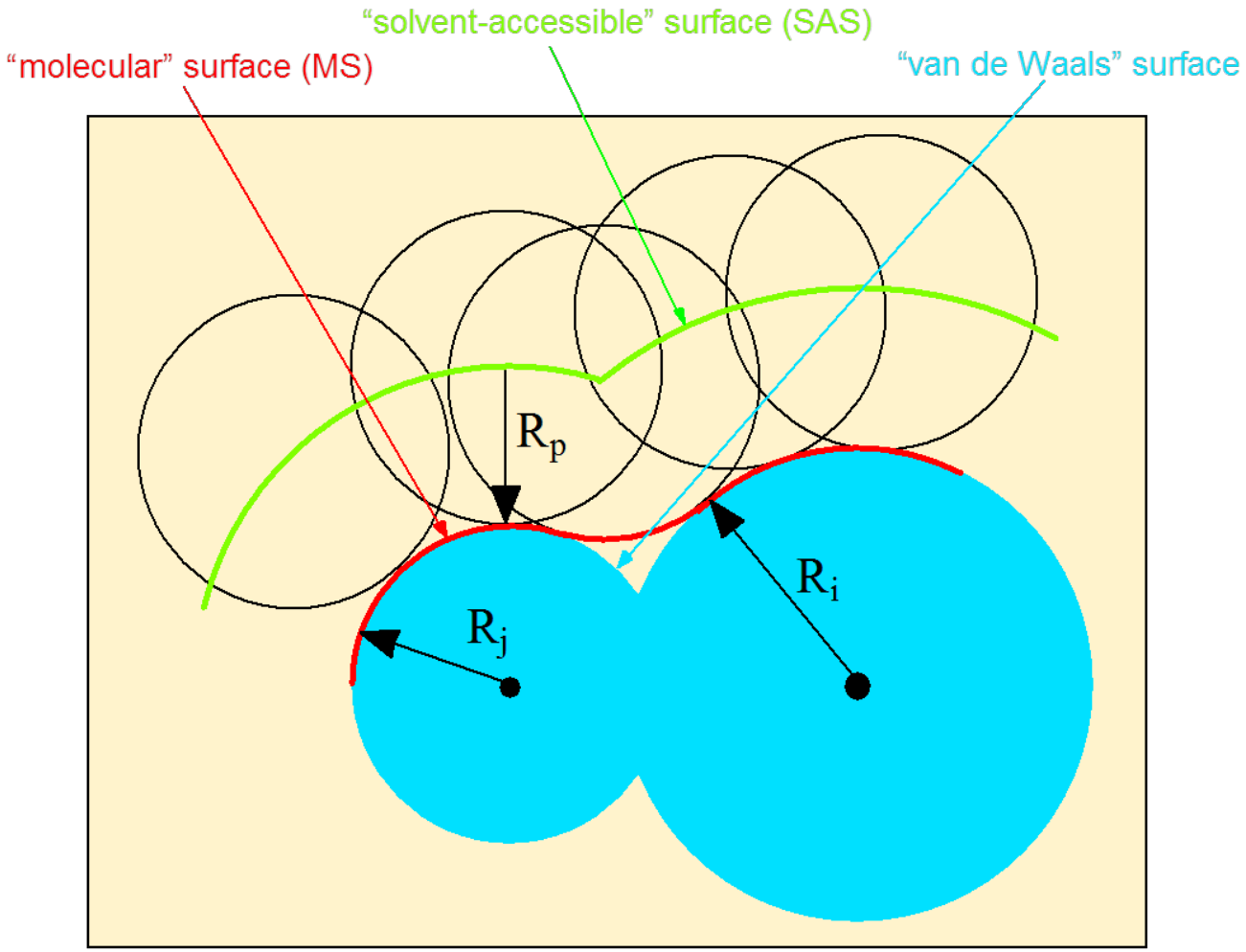




$$S_{sp} = 4\pi\left(\frac{3V_{sp}}{4\pi}\right)^{2/3} = 4\pi\left(\frac{3V_{mol}}{4\pi}\right)^{2/3}$$

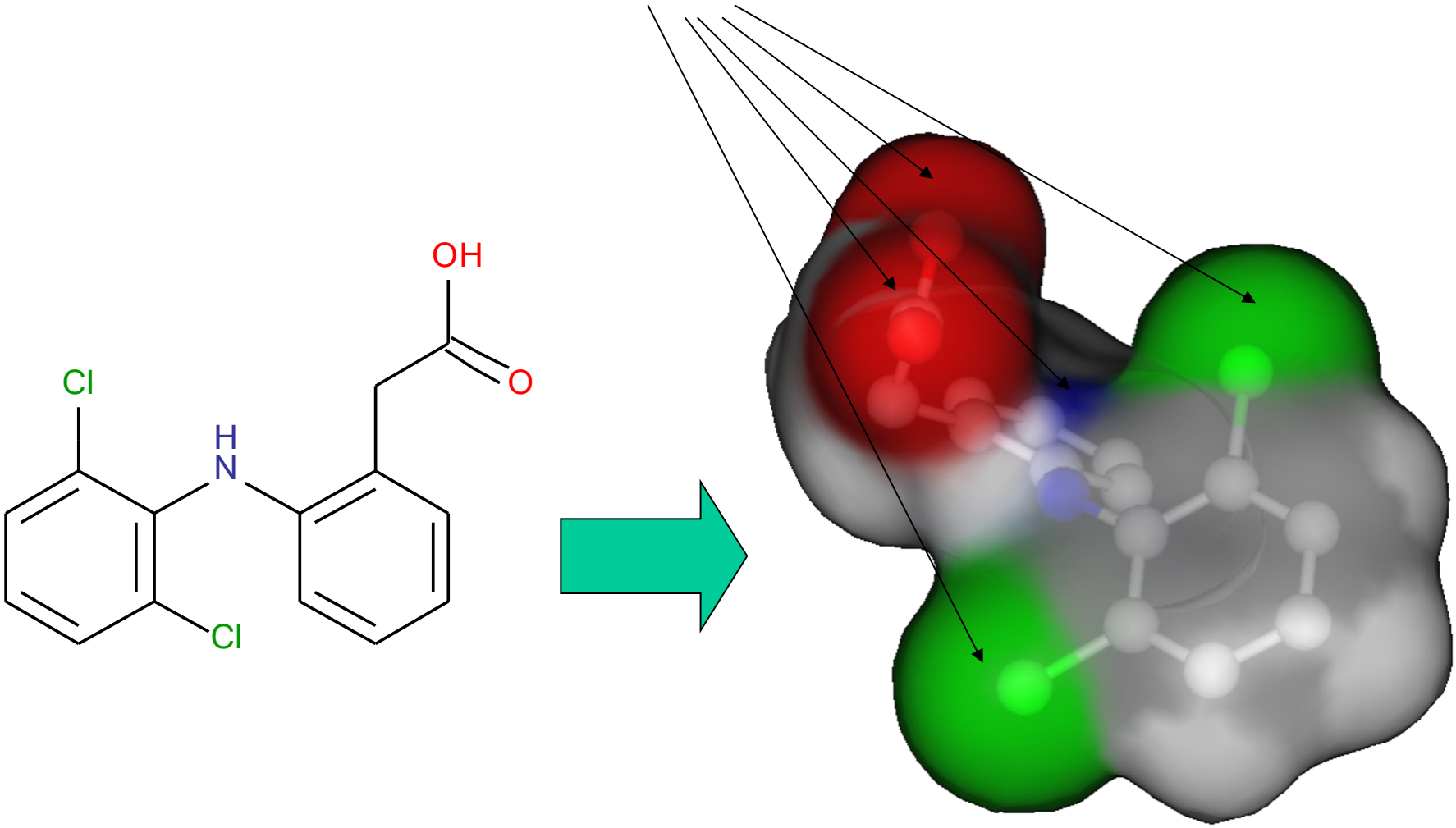
Surface-based descriptors

- Surface area
  - Van der Waals, Solvent-Accessible, Molecular (Connolly) surface area



Surface Polarity descriptors

Polar Surface Area: Total area of the part of the molecular surface that corresponds to polar atoms: O, N, halogens



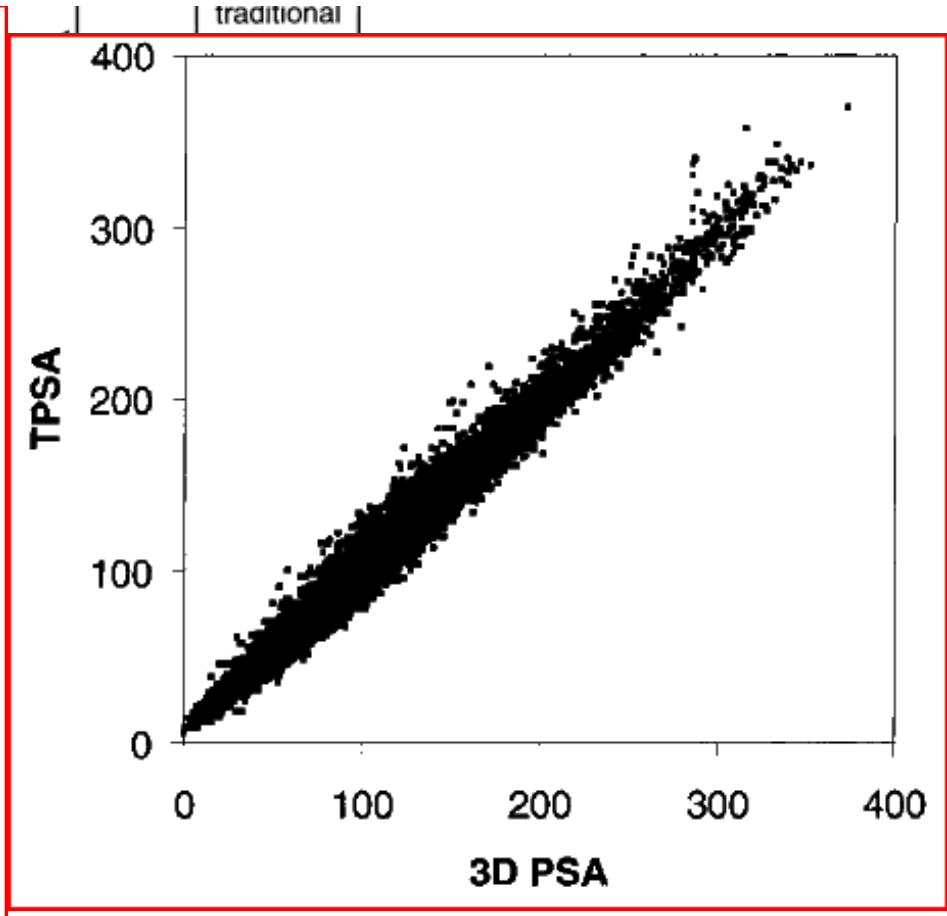
Topological Polar Surface Area: back to 2D!

Peter Ertl, Bernhard Rohde, and Paul Selzer, *J. Med. Chem. 2000, 43, 3714-3717*

$$3DPSA \approx \sum_{groups} (Number\ of\ groups) \times (Fitted\ group\ contribution)$$

Table 1. Atomic Contributions (Å^2) to PSA

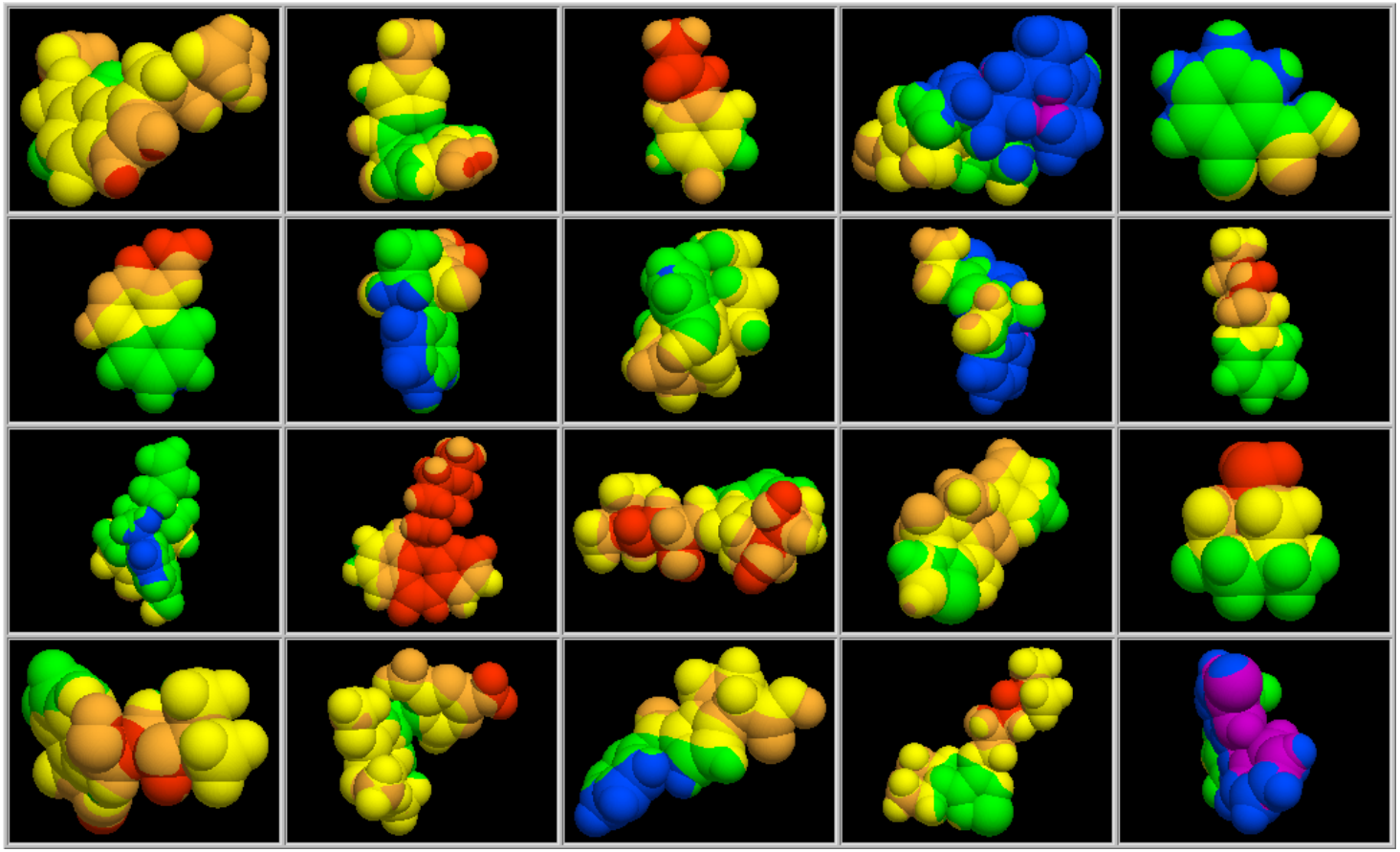
atom type[a]	PSA contrib	atom type[a]	PSA contrib
[N](−*)(−*)−*	3.24	[nH](:*):*	15.79
[N](−*)=*	12.36	[n+](:*)(:*):*	4.10
[N]#*	23.79	[n+](−*)(:*):*	3.88
[N](−*)(=*)=* [b]	11.68	[nH+](:*):*	14.14
[N](=*)#* [c]	13.60	[O](−*)−*	9.23
[N]1(−*)−*−*−1[d]	3.01	[O]1−*−*−1[d]	12.53
[NH](−*)−*	12.03	[O]=*	17.07
[NH]1−*−*−1[d]	21.94	[OH]−*	20.23
[NH]=*	23.85	[O−]−*	23.06
[NH2]−*	26.02	[o](:*):*	13.14
[N+](−*)(−*)(−*)−*	0.00	[S](−*)−*	25.30
[N+](−*)(−*)=*	3.01	[S]=*	32.09
[N+](−*)#* [e]	4.36	[S](−*)(−*)=*	19.21
[NH+](−*)(−*)−*	4.44	[S](−*)(−*)(=*)=*	8.38
[NH+](−*)=*	13.97	[SH]−*	38.80
[NH2+](−*)−*	16.61	[s](:*):*	28.24
[NH2+]=*	25.59	[s](=*)(:*):*	21.70
[NH3+]−*	27.64	[P](−*)(−*)−*	13.59
[n](:*):*	12.89	[P](−*)=*	34.14
[n](:*)(:*):*	4.41	[P](−*)(−*)(−*)=*	9.81
[n](−*)(:*):*	4.93	[PH](−*)(−*)=*	23.47
[n](=*)(:*):* [f]	8.39		



3D PSA

3D Lipophilicity Potential (Rozas)

$$MLP(j) = \sum_{i=1}^{n} \frac{f_i}{1 + d_{ij}}$$

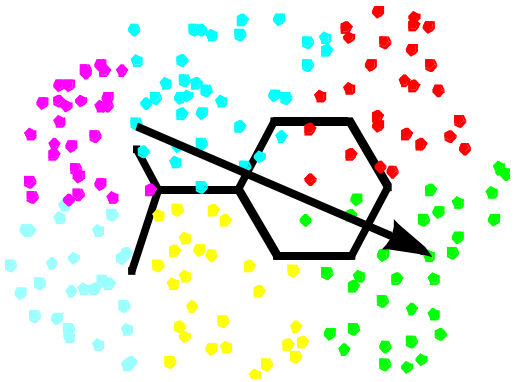


hydrophobic hydrophilic

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

Autocorrelation of Molecular Surface Properties

$$A(d) = \frac{1}{L} \sum_{x,y} p(x) \cdot p(y) \Big|_{d < \|x-y\| \leq d+\varepsilon}$$

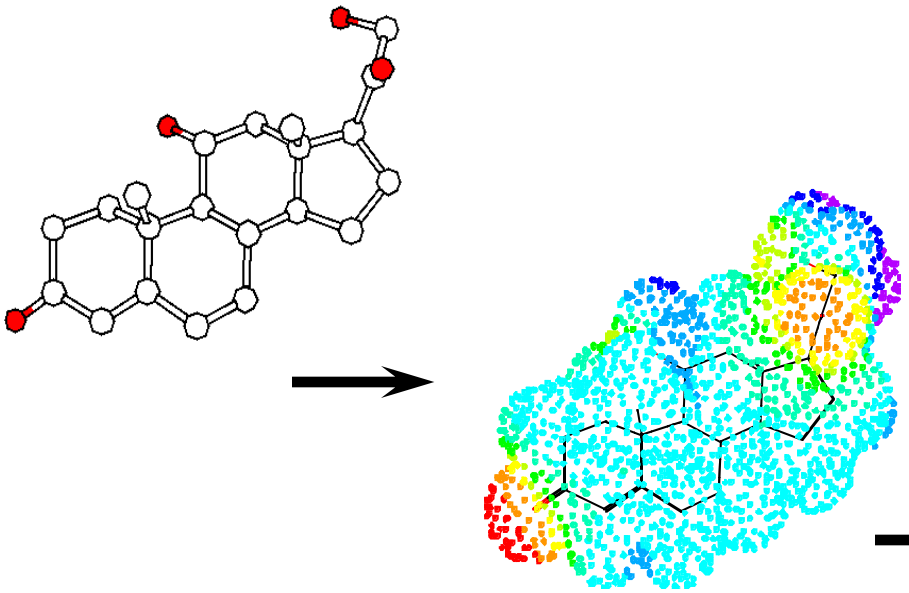


p(x),p(y)	**property at points *x,y***
d	**distance**
L	**number of point pairs**

***d* = [4.0, 5.0 [Å**

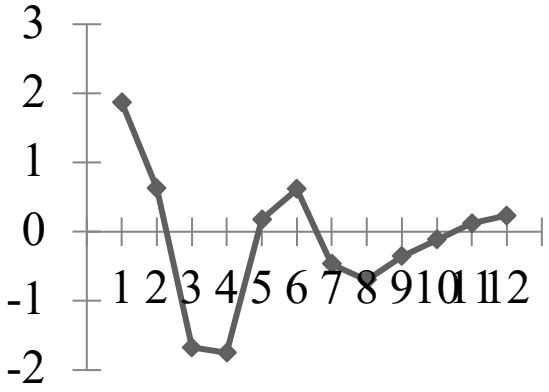
- **Orientation-independent** description: distances do not change upon roto-translation of molecules
- Example: *p=Interaction energy with a molecular probe* (such as water); GRIND descriptors (Pastor *et. al., J. Med. Chem.,* **2000**, *43,* 3233–3243)

Autocorrelation of Molecular Surface Properties



ca. 3000 points
(10 points / Å^2)

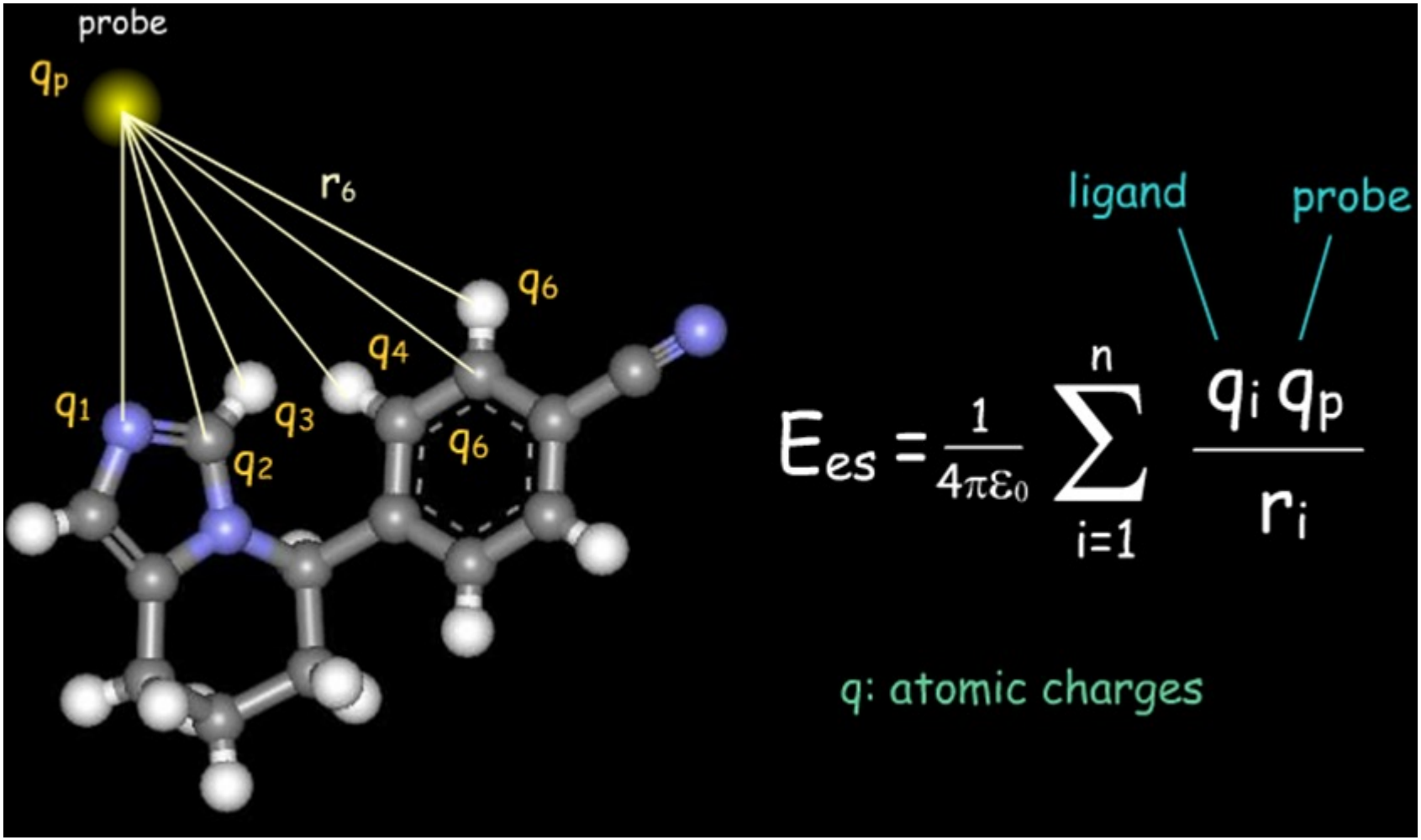
12 autocorrelation coefficients



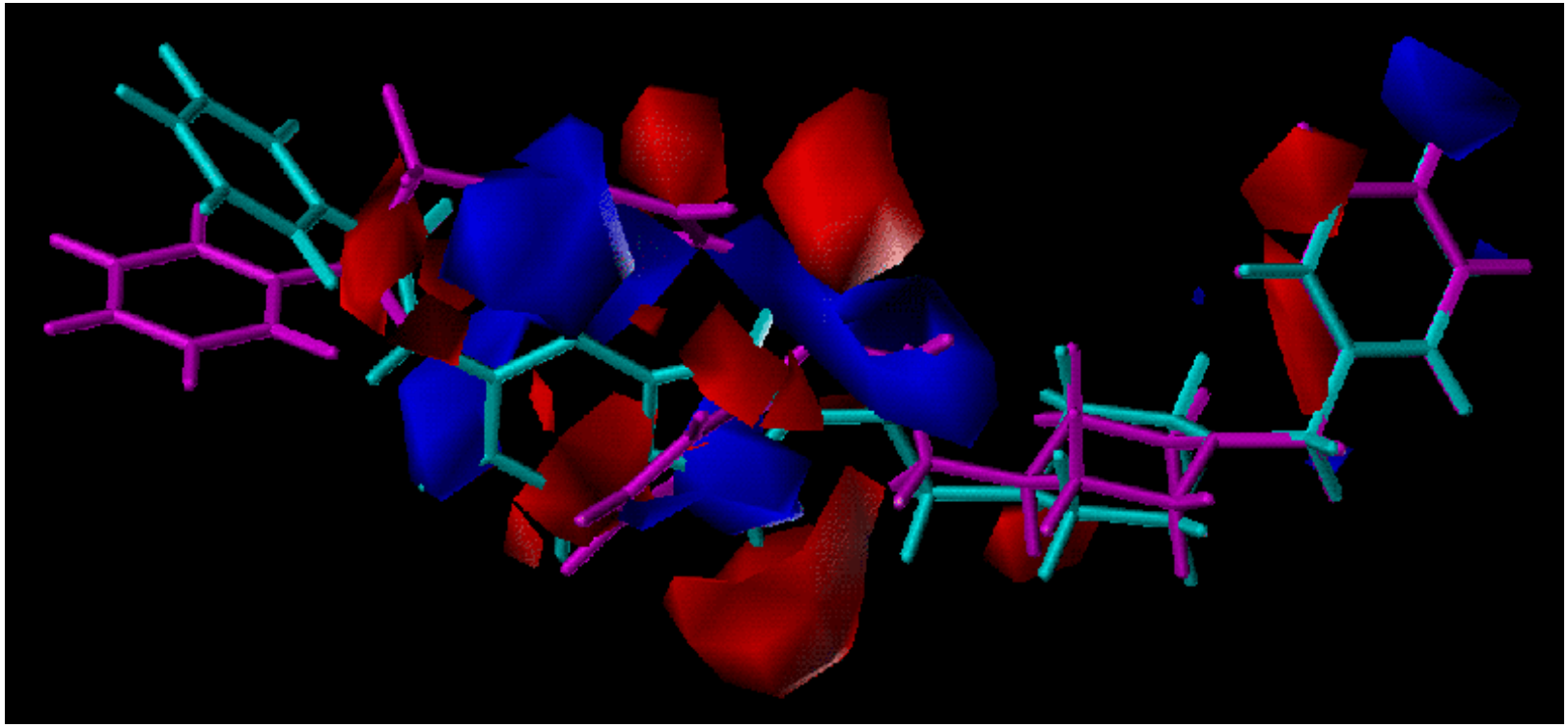
d [Å]

M. Wagener, J. Sadowski, J. Gasteiger, *J. Am. Chem. Soc.* 1995, *117*, 7769.

Field Intensity Descriptors in Surrounding Space are Reference System-Dependent



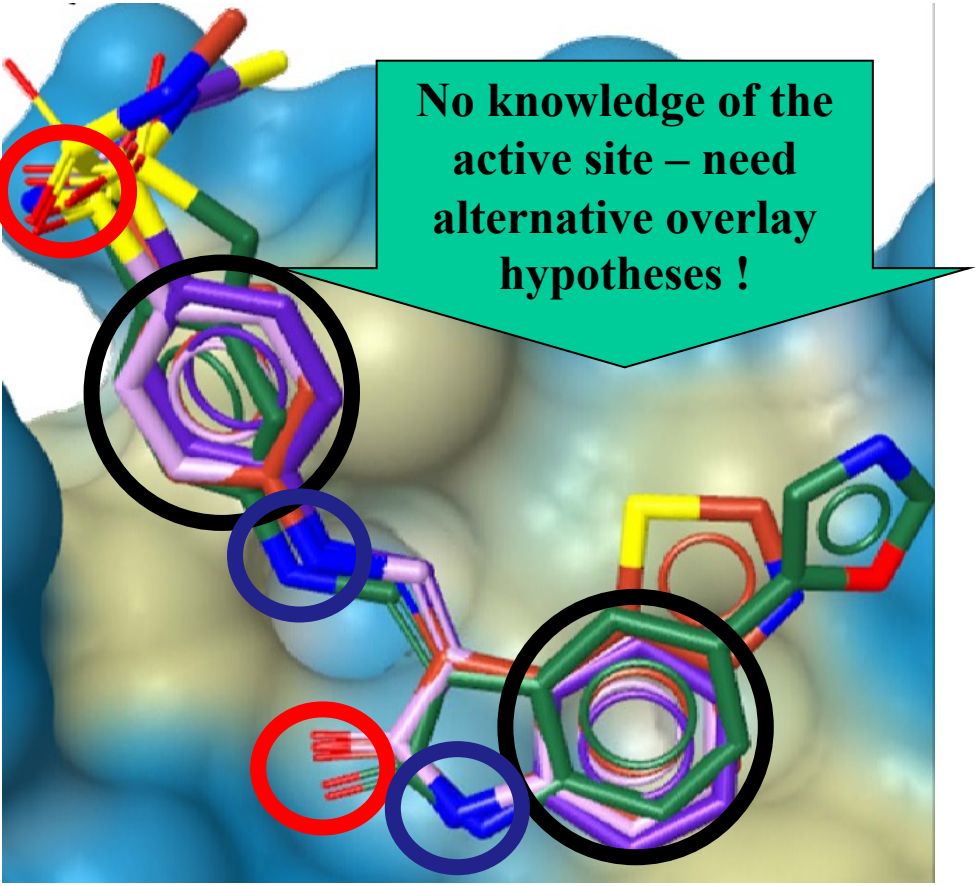
Fields are Orientation-Dependent: to compare them, molecules must first be ALIGNED in 3D



CoMFA: Comparative Molecular Field Analysis

- Red zones are favorable for interactions with the positively charged fragments
- Blue zones are favorable for interactions with the negatively charged fragments

Overlay-Dependent Descriptors: Pharmacophore Occupancy



- Pharmacophore models represent binding mode hypotheses:
  - use overlay models to "bind" descriptors to specific spots in space
  - Pharmacophore hot spots are defined by the consensual presence of groups of similar type, throughout the series of known actives
  - Descriptors are occupancy levels of these spots

CONCLUDING REMARKS

For Each Case Study, Suited Descriptors...

There's no difference between theory and practice, but in practice there is

- **In theory**, molecular topology is all you need to know…
- … but often, the implicit information present in the topology should be made "explicit" by the description strategy:
  - Geometry is rather reliably "written" in the topology
  - The preferred protonation status is "written" in the topology as well – **but not always easy to read**…
- **In practice**, no descriptor provides a complete characterization of a molecular object
  - If you describe the pharmacophore, you should not expect predicting reactivity… unless a lucky correlation makes you believe in it.
  - For modeling *in vivo* properties, need to understand binding (pharmacophore), metabolism (reactivity), bioavailability (lipophilicity, *etc*). It's Mission Impossible…

A Descriptor MUST Have ...

- an unambiguous algorithmically computable definition
- invariance with respect to labeling and numbering of atoms
  - **Make Autoencoder Latent Spaces numbering-independent!**
- invariance with respect to roto-translation, unless based on an unambiguous molecular overlay procedure
- values in a suitable numerical range for the set of molecules where it is applicable to

A Descriptor Should Have ...

- a structural interpretation
- a good correlation with at least one property
- no trivial correlation with other molecular descriptors
- gradual change in its values with gradual changes in the molecular structure
- no dependence on experimental properties
- no restriction to small classes of molecular structures
- if possible, some discrimination power among isomers
- preferably, no dependence on other molecular descriptors
- decodability ? (back from the descriptor value to the structure)