In Silico Drug Design

Dragos Horvath

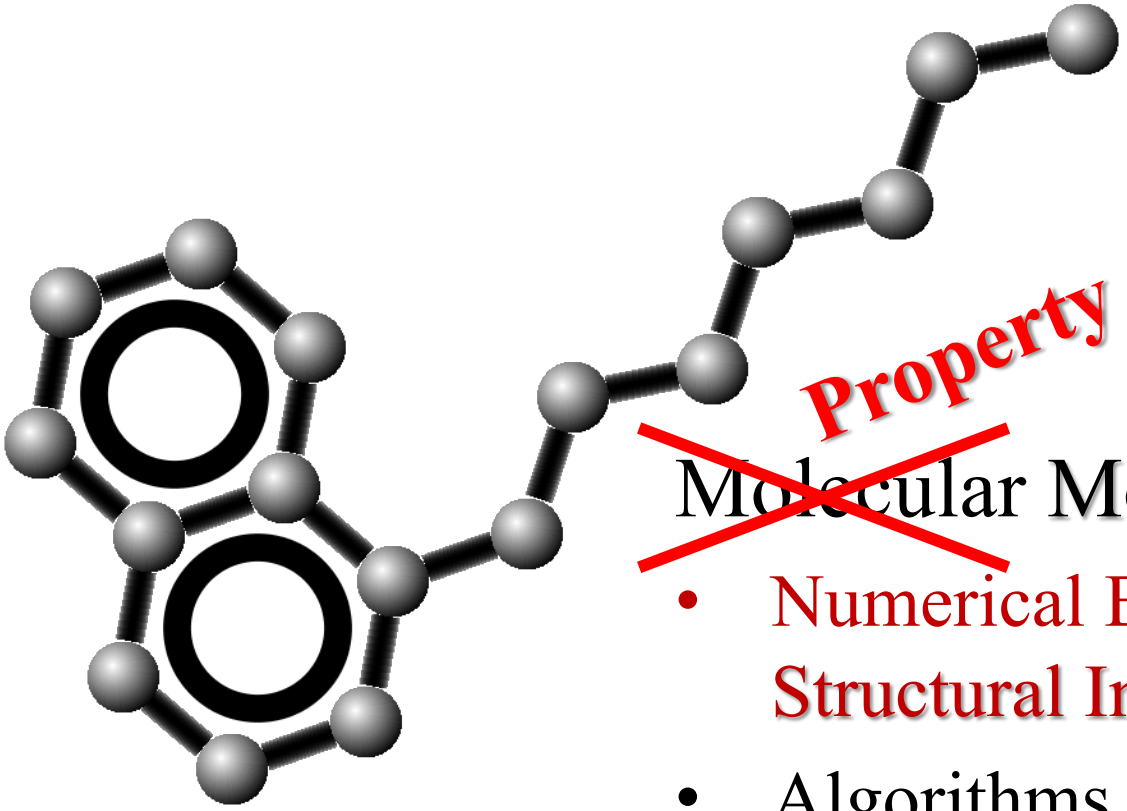
Laboratoire d'InfoChimie, UMR 7177

CNRS – Université de Strasbourg

67000 Strasbourg, France

dhorvath@unistra.fr

1. « Ceci n'est pas une molécule »



Property

~~Molecular~~ Models:

- Numerical Encoding of Structural Information &
- Algorithms relating this to observable Properties

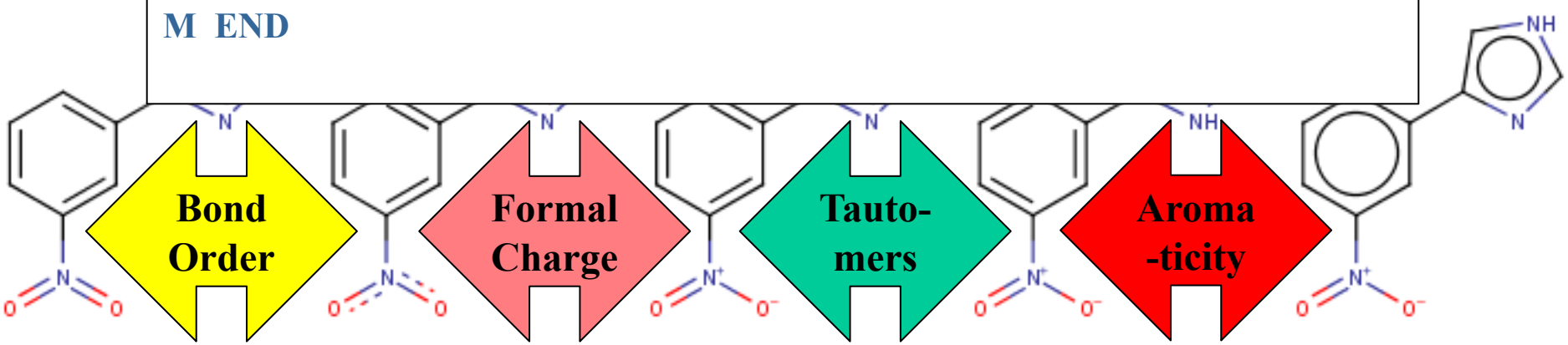
Computer Management of Chemical Structures

- Sto … es

```
Mrv1804 06061810042D   **MDL MOL file header

14 15  0  0  0  0          999 V2000                          ** nr of atoms, bonds
 -19.8631    3.9987    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0 ** atom list
 -20.5776    3.5862    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
 -20.5776    2.7612    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
 -19.8631    2.3487    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  ....
 1  2  1  0  0  0  0  ** bond list (first atom, second atom, bond order, flags)
 2  3  2  0  0  0  0
 3  4  1  0  0  0  0
 4  5  2  0  0  0  0
 ...
M  END
```

- Sta … you see



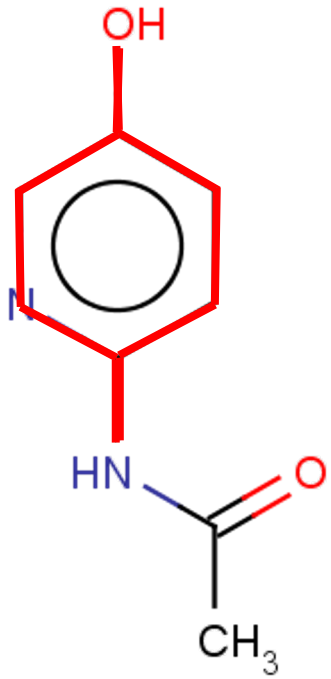
Models… for Human or Artificial Intelligence

- At various levels of possible molecular representations:
  - 2D: based on information available in the molecular graph: "paper chemistry" (*Human*) *vs.* "Chemoinformatics" (*Machine*)
  - 3D: considering molecular geometry: "Stereochemistry" (*Human*) *vs.* "Conformational sampling" (*Machine*)
  - Quantum Chemical (*presenter's IQ insufficient for this topic*)
- Learning from various experimental data sources:
  - **Ligand-based**: From **examples of known ligands/inhibitors** – no knowledge of the target structure: "Structure-Activity Relationships **SAR**" (*Human*) – "**Q**uantitative **SAR**" (*Machine*)
  - **Structure-based**: From **target structures**, hypothesize how the ligand would bind the active site: "Pharmacophore Match, Docking" (*Machine*).

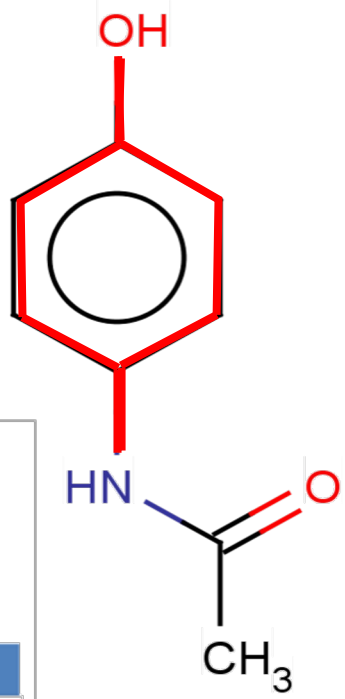
Molecular Descriptors or Fingerprints

- 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)

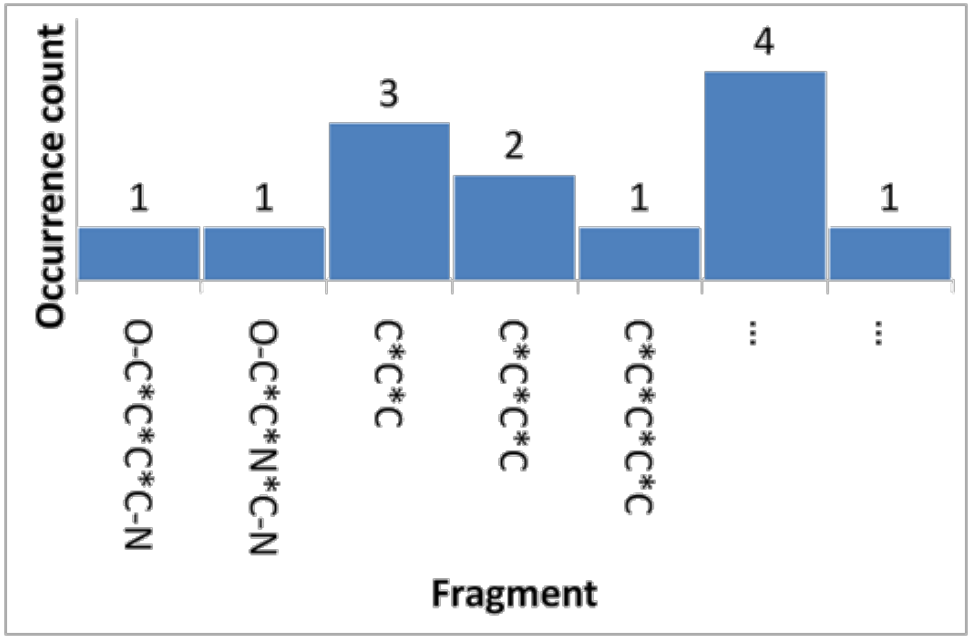
Example 1: ISIDA Sequence Counts



O-C*C*C*C-N	1	2
O-C*N*C*C-N	1	0
…		…

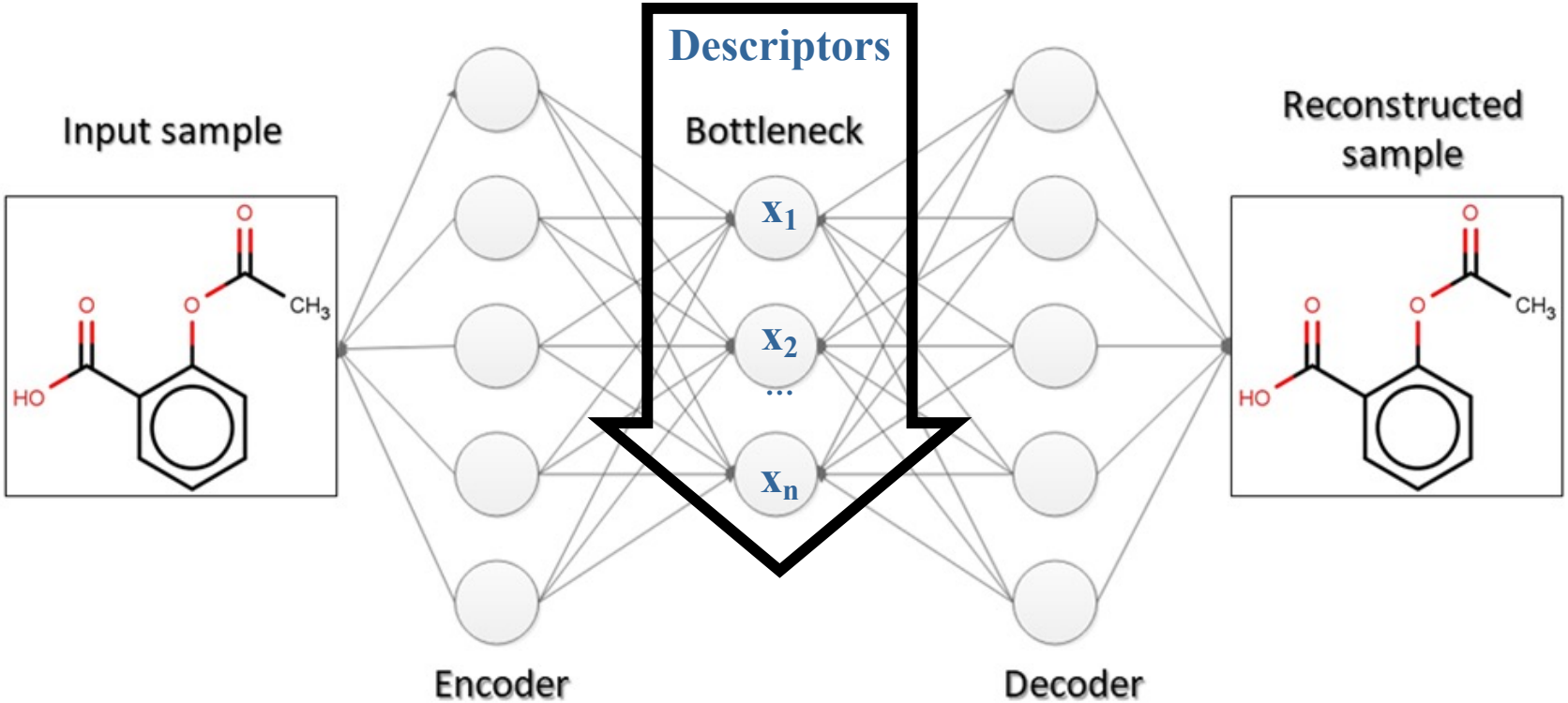


(1,1,…)



(2,0,…)

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…

Training of the Autoencoder

Reconstruction Accuracy:

ChEMBL

~ 1.5 M structures (SMILES strings) of biologically active molecules

Training set:
1.2 M structures

99.98 %

Test set:
300 K structures

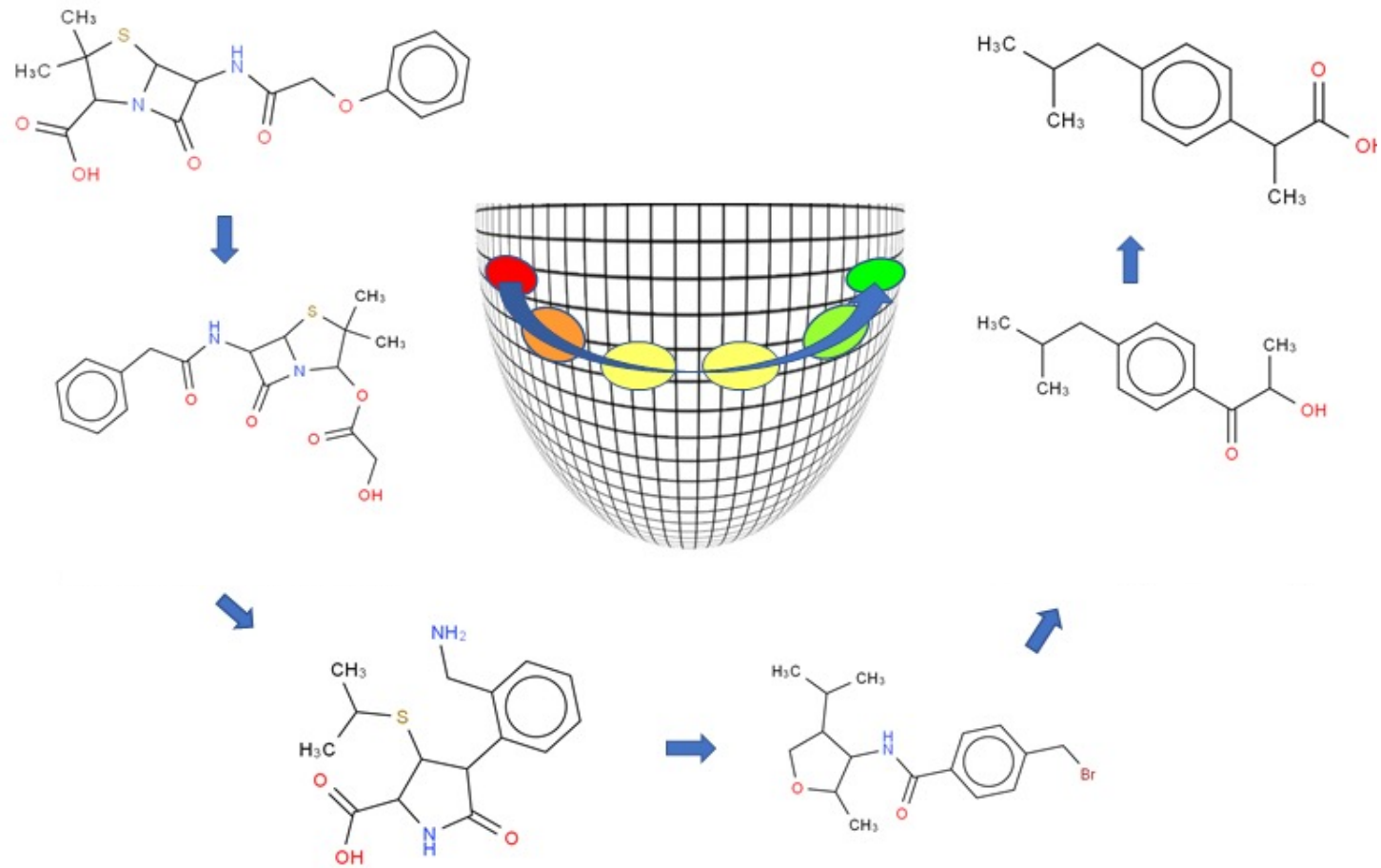
99.71 %

The trained autoencoder model is generalized (it did *not* learn by heart)

Molecular Morphing: walking across chemical space!

Penicillin V

Ibuprofen



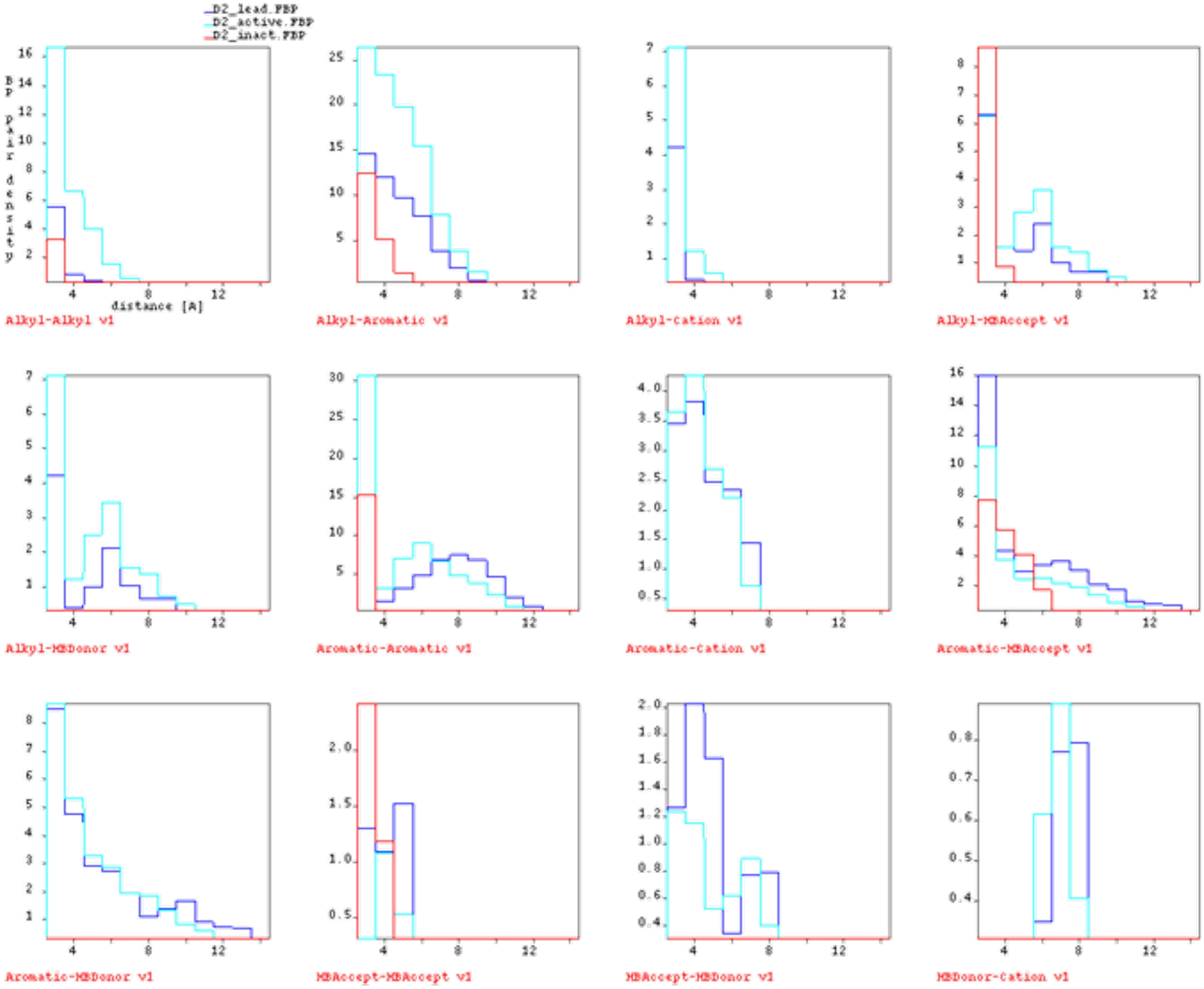
2. Computer-Aided Ligand-Based Design: the « Medicinal Chemistry » of Ligand Fingerprints

« Similar molecules have similar properties » →
« Molecules with similar fingerprints have similar properties »

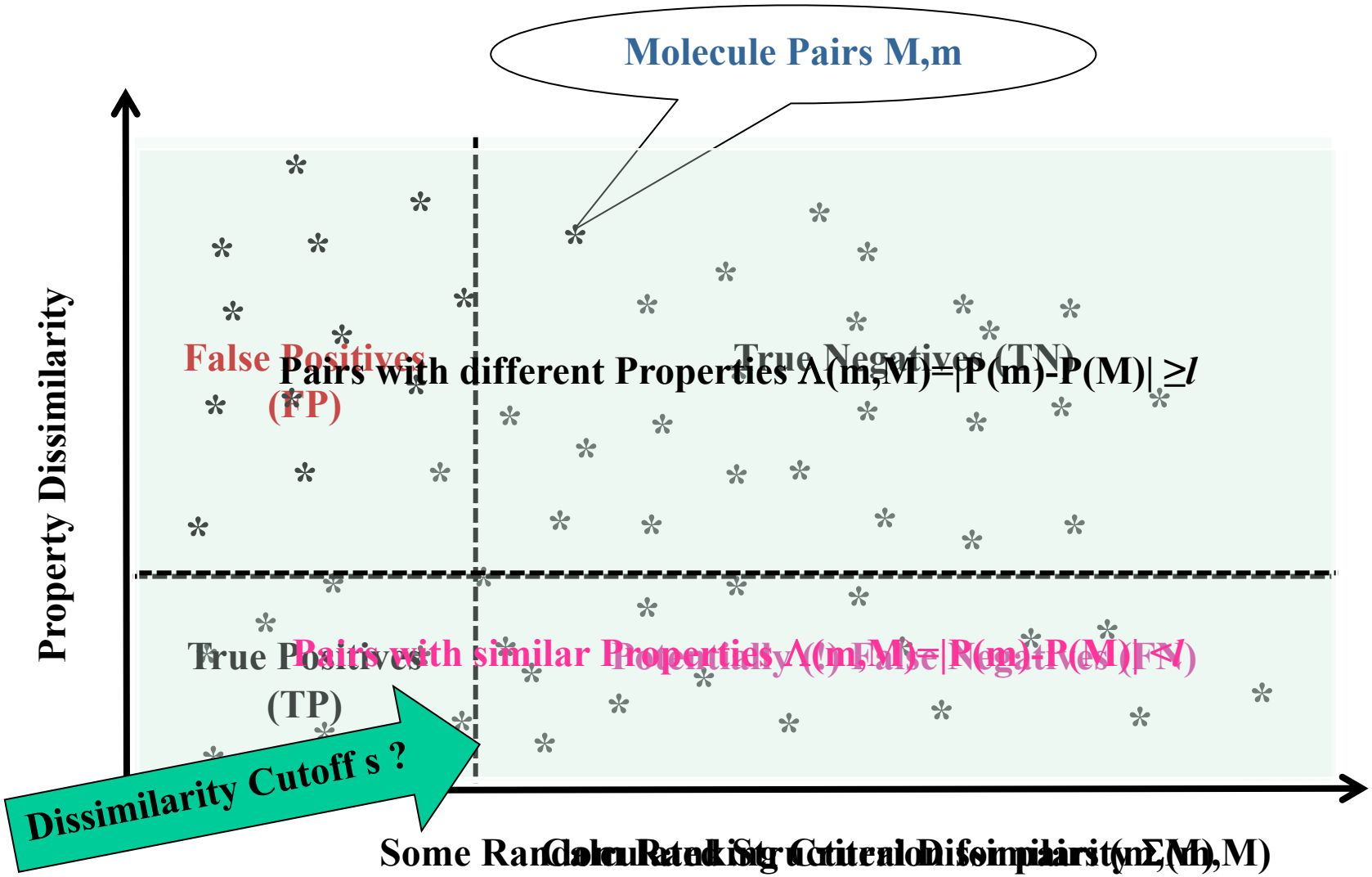
« Structure-Property Relationships » →
« Fingerprint-Property Relationships » (or Quantitative Structure-Property Relationships, QSPR)

2.1 Molecular Similarity in Chemoinformatics

Molecular Similarity Expressed by Fingerprint Similarity

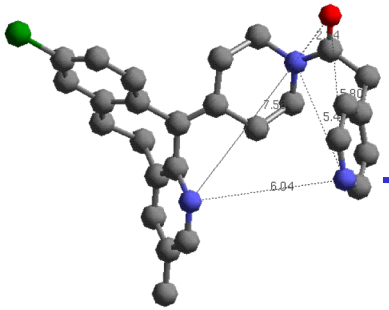


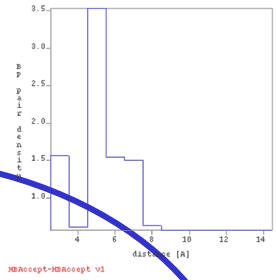
The Similarity Principle – Neighborhood Behavior

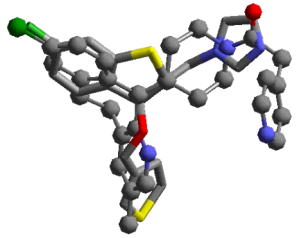


Similarity-Based Virtual Screening…

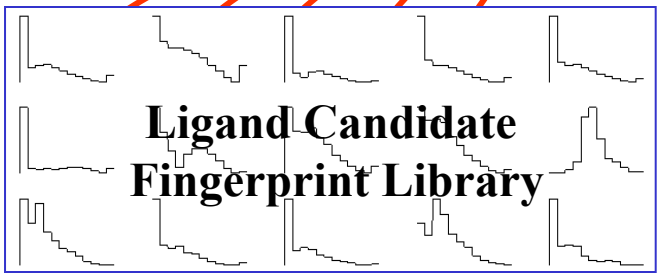
Active Reference → **Reference Fingerprint** → **Nearest Neighbors**

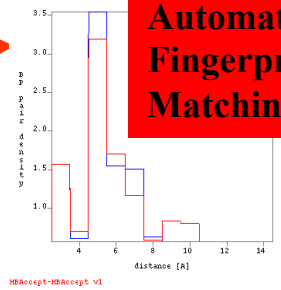




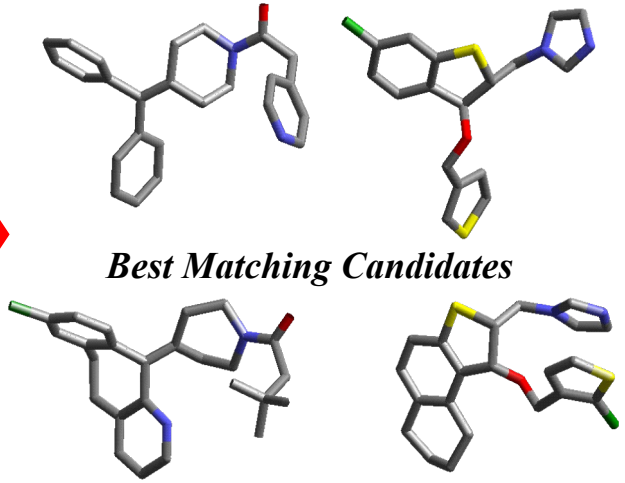


Superposition-based Similarity Scoring





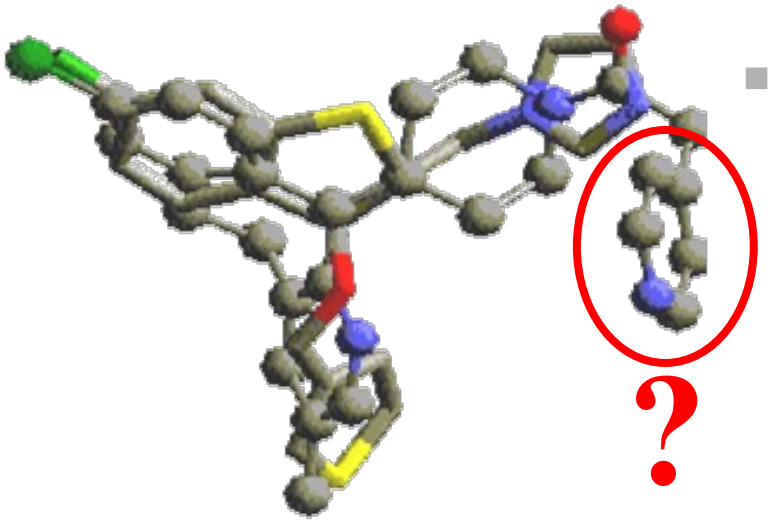
Automated Fingerprint Matching...



Best Matching Candidates

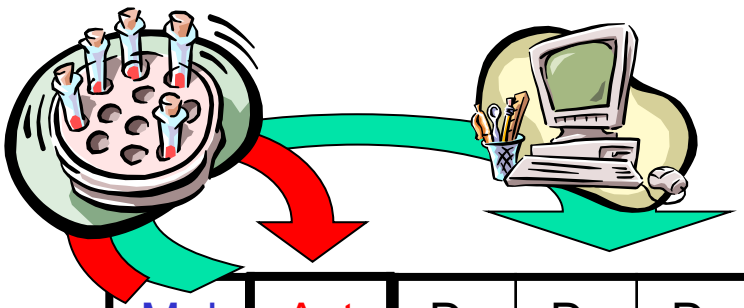
Strenght & Limitations of Similarity-based VS

- (+) Only need ONE active ligand to seek for more like it…
- (+) With appropriate descriptors, calculated similarity may be complementary to the scaffold-based similarity perceived by medicinal chemists
  - → « Scaffold Hopping »: bypassing synthetic bottlenecks and/or pharmacokinetic property problems, patent space evasion, *etc.*



- (--) Within the reference ligand, « all groups are equal, but some are more equal than others » when it comes to controlling activity… *so what if we mismatch the latter*??

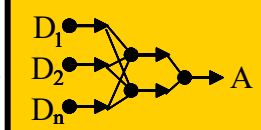
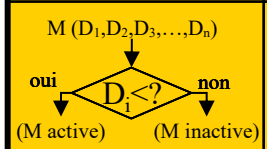
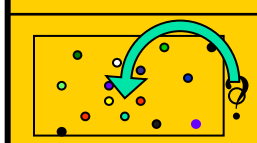
2.2: So, we need to LEARN the features that really matter – building QSPRs



Mol	Act	D_1	D_2	D_3	…	D_n
M_1	A_1	d_{11}	d_{21}	d_{31}	…	d_{n1}
M_2	A_2	d_{12}	d_{22}	d_{32}	…	d_{n2}
M_3	A_3	d_{13}	d_{23}	d_{33}	…	d_{n3}
M_4	A^c_4	d_{14}	d_{24}	d_{34}	…	d_{n4}
$M_{..}$	$A^c_{..}$	$d_{1...}$	$d_{2...}$	$d_{3...}$	…	$d_{n...}$
M_m	[illegible]	[illegible]	d_{2m}	d_{3m}	…	d_{nm}

Model Fitting

A QSPR model expresses observed correlations between *certain* descriptors and activity

$A = \sum \alpha_i' D_i$	linear
	neural net
	decision tree
	neighbor-hood model

Correlations: The Cornerstone of QSPR Philo[illegible] Religion?)

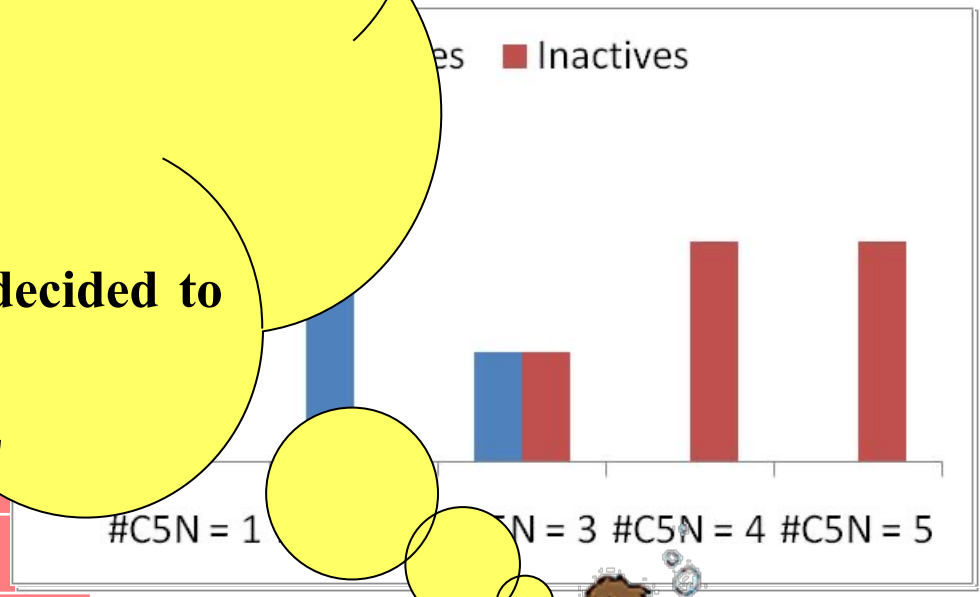
I always end up in this deplorable state, no matter whether I drink:

- Vodka-Soda
- Martini-Soda
- Gin-Soda
- Whisky-Soda…

… therefore, as of tomorrow, I decided to stop drinkin' SODA !!

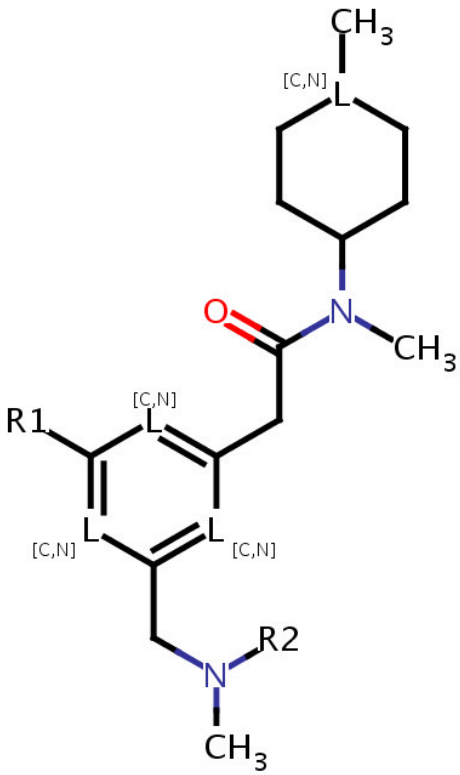
MolID				
3				
5				
6				
7	0			4
8	0	2		
9	0	2	1	5
10	0	1	0	4

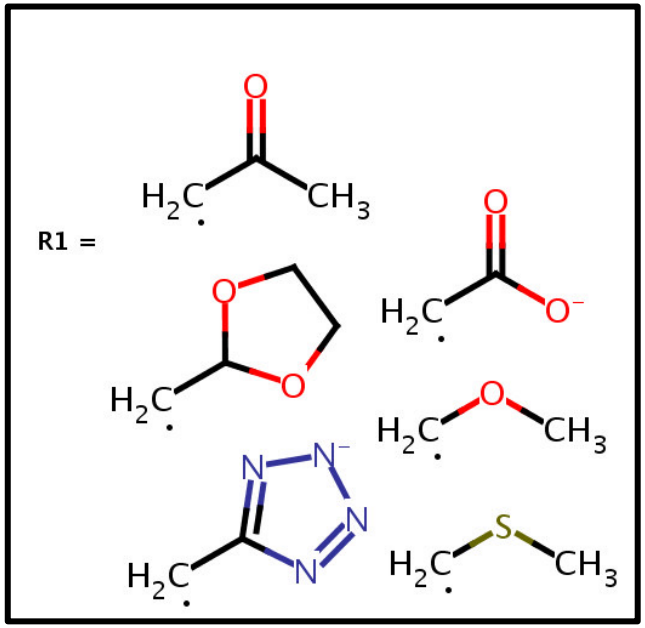
Less is better

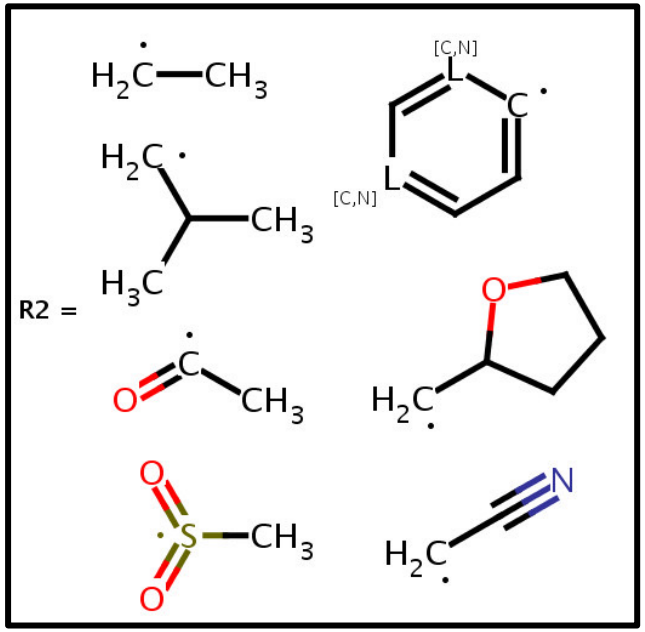


Correlation is not Causality - an Obvious, but Inconvenient Truth…

- SAR sets are **always** limited in diversity and therefore may (and **always will**) accomodate **coincidental** relationships between different features:







Diverse library of 16x6x10=960 compounds… with $N_{PC}=N_{HD}$

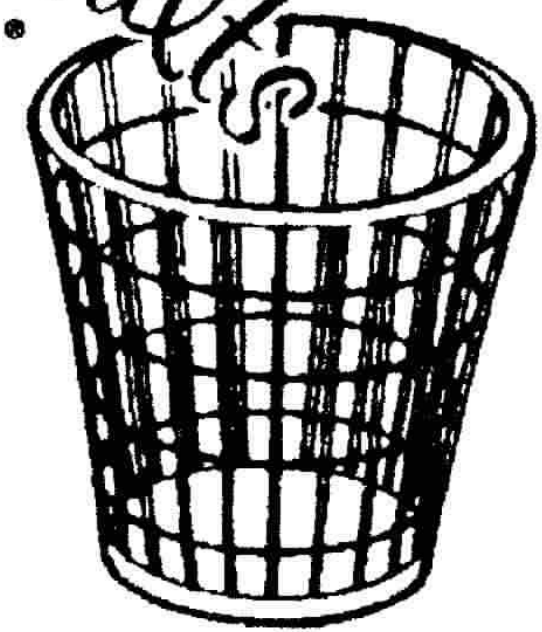
The Descriptor Conspiracy: A Psychedelic μ Opiate Affinity Model…

JOURNAL OF
Irreproducible Results

Official organ of the Society for Basic Irreproducible Research (ISSN 0022-2038)

Welcome to our 48th

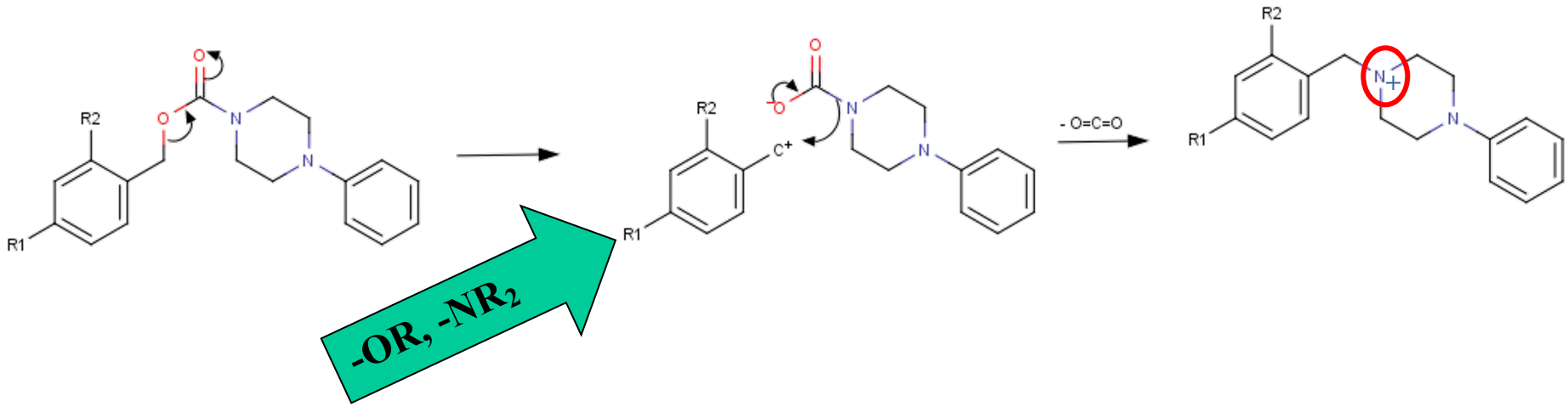
Year of Publication



- HB-acceptor in *para* of benzyl alcohol enhances μ receptor affinity

It's just property covariance – luckily, of the "useful" kind!

- The most "active" carbamates of the training set turned out to be contaminated with ‰ traces of decarboxylation product, featuring the opioid ligand specific tertiary amine and having nanomolar potencies…



- Our QSAR actually explained… the decarboxylation mechanism: *p*-OR or –NR_2 stabilizes the intermediate carbocation… thus rendering contamination possible

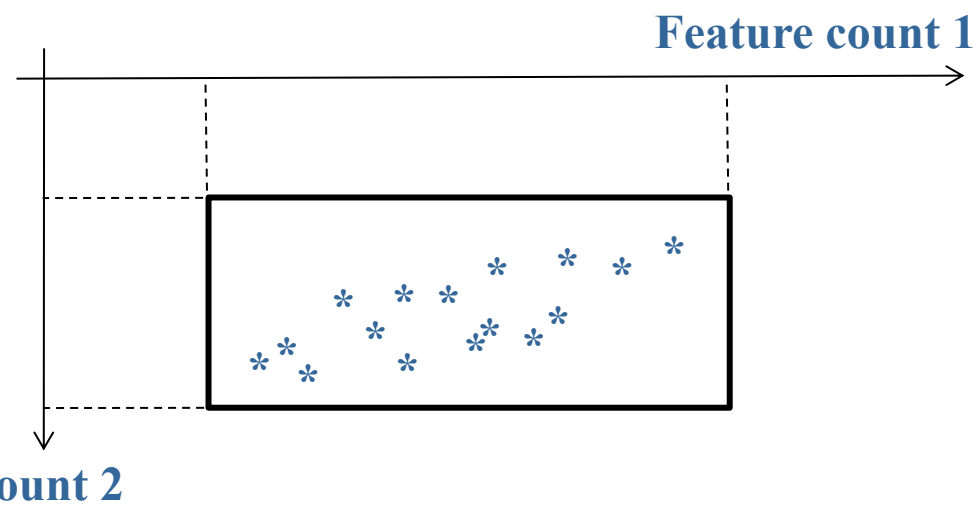
"Let s be a *representative sample* of the set S…"

- It takes a sample of ~10^4 individuals to extrapolate the voting intentions of a population of ~10^7. *What's the representative subset size of 10^{25} drug-like compounds?*
  - If we ever dared to publish QSARs trained on fewer compounds, shame on us!
- If given N=3 coordinate pairs (Y,X), not even Carl Friedrich Gauss could come up with a model more sophisticated than $Y=aX^2+bX+c$
- May your model apply to one million and one molecules – it may still fail for the one million and second!
  - **One cannot <u>validate</u> QSAR – but just <u>fail to invalidate</u> it!**

The Applicability Domain – A Compromise…

- Restrict the applicability of a QSAR model to a well-defined subset of the chemical space – the one populated by the training molecules.
  - Insufficient sampling of chemotypes outside this AD is then irrelevant.
  - How do we define this subset of chemical space to be as large as possible, while nevertheless densely enough populated by training molecules?

Example: the Feature Control Approach



Drawing Confidence from Consensus

infochim.u-strasbg.fr/webserv/VSEngine.html

Prediction of property logP - Page nr. 3 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://infochim.u-strasbg.fr/userdata/dragos/logP/logPP3.html

Most Visited | Personnaliser les liens | Windows Media | Windows

Courrier :: Boîte de réception | Virtual Screening Engine - Laboratoire d... | Prediction of property logP - Pag...

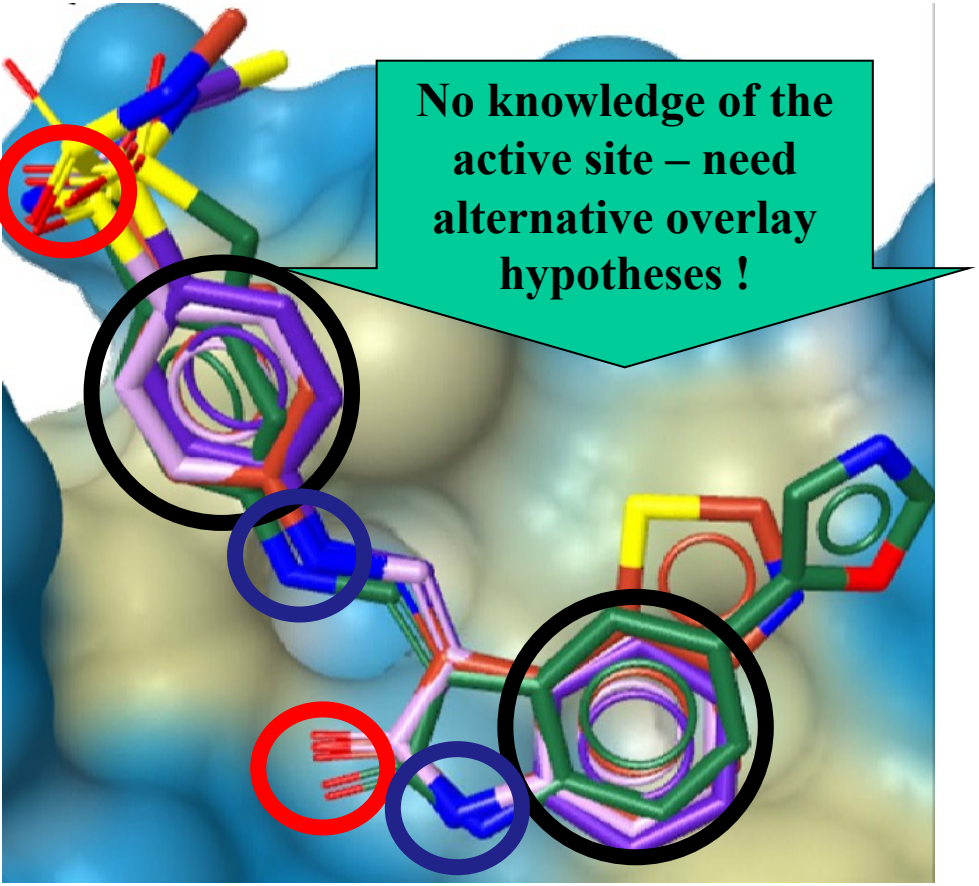
Predicted property **logP** for 9677 compounds *AS A CONSENSUS OF APPLICABILE LOCAL MODELS*

Predictor	Required Trust Level	Number of returned predictions	RMS_E	R^2_E	MaxErr
logP0	Any	9540[c]	0.67	0.866	5.00
logPApp	Any	9325	0.65	0.872	4.50
logP	Any	9540	0.66	0.871	5.00
logP	MEDIUM or better	8562	0.62	0.879	4.25
logP	GOOD or better	6861	0.59	0.883	4.25
logP	OPTIMAL	2318	0.53	0.880	3.75

logP	VAR	TRUST	REASON
3.13	0.127	POOR	containing compound inside their applicability domain - Individual models failed to reach unanimity - prediction variance exceeds 1.0% of the property range width
2.60	0.105	OPTIMAL	-

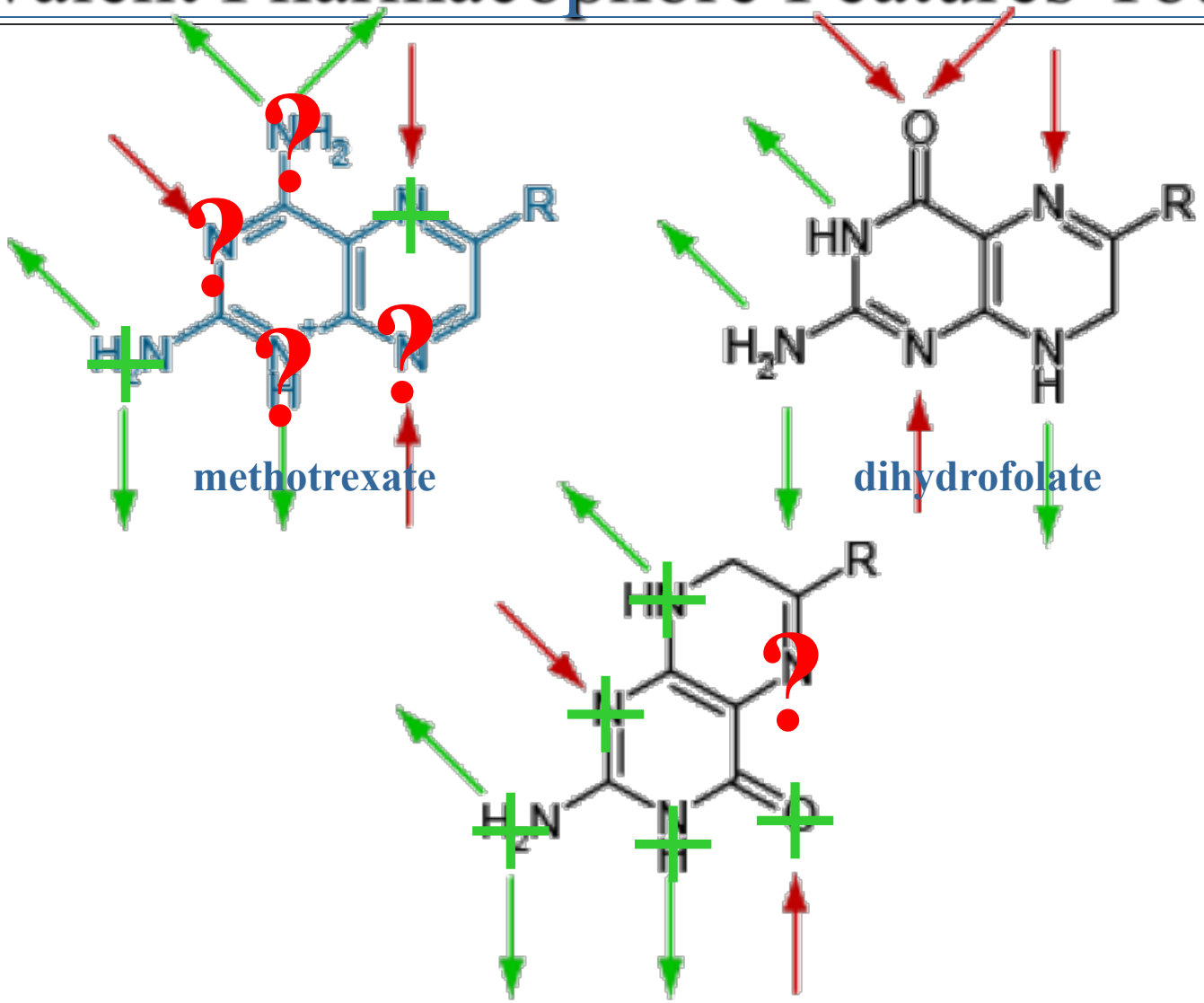
Done

We are Medicinal Chemists – tell us about Pharmacophore Models, forget QSAR!!



- Bad news: Pharmacophore models are just a peculiar type of 3D-QSAR:
  - 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

Kama Sutra with Ligands: Match As Many Equivalent Pharmacophore Features You May!

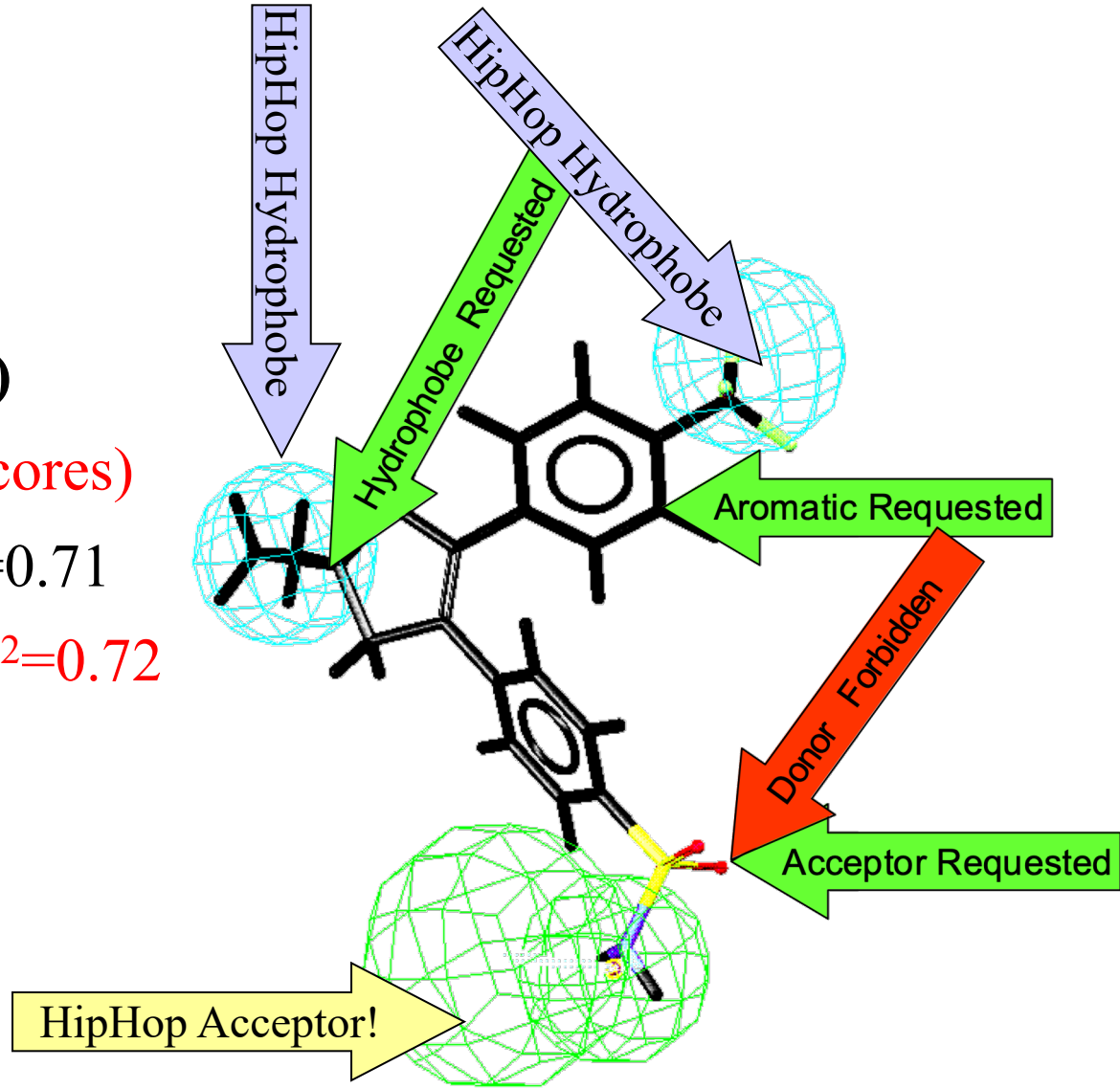


Böhm, Klebe, Kubinyi, "Wirkstoffdesign" (1999)

Cox_2 Minimalistic Overlay-based Model..

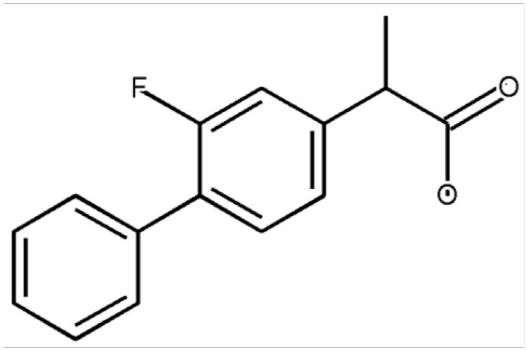
... can't get much better than that!

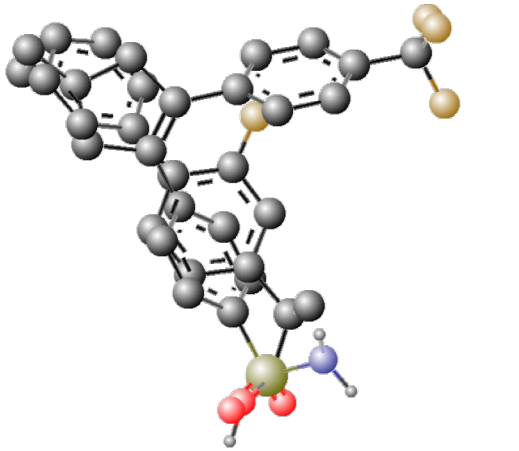
- **~2200 Molecules (pIC_{50})**
- 6 variables (occupancy scores)
- Training RMS=0.71, R^2=0.71
- Validation RMS=0.70, Q^2=0.72



Furthermore, it supports Scaffolfd Hopping !

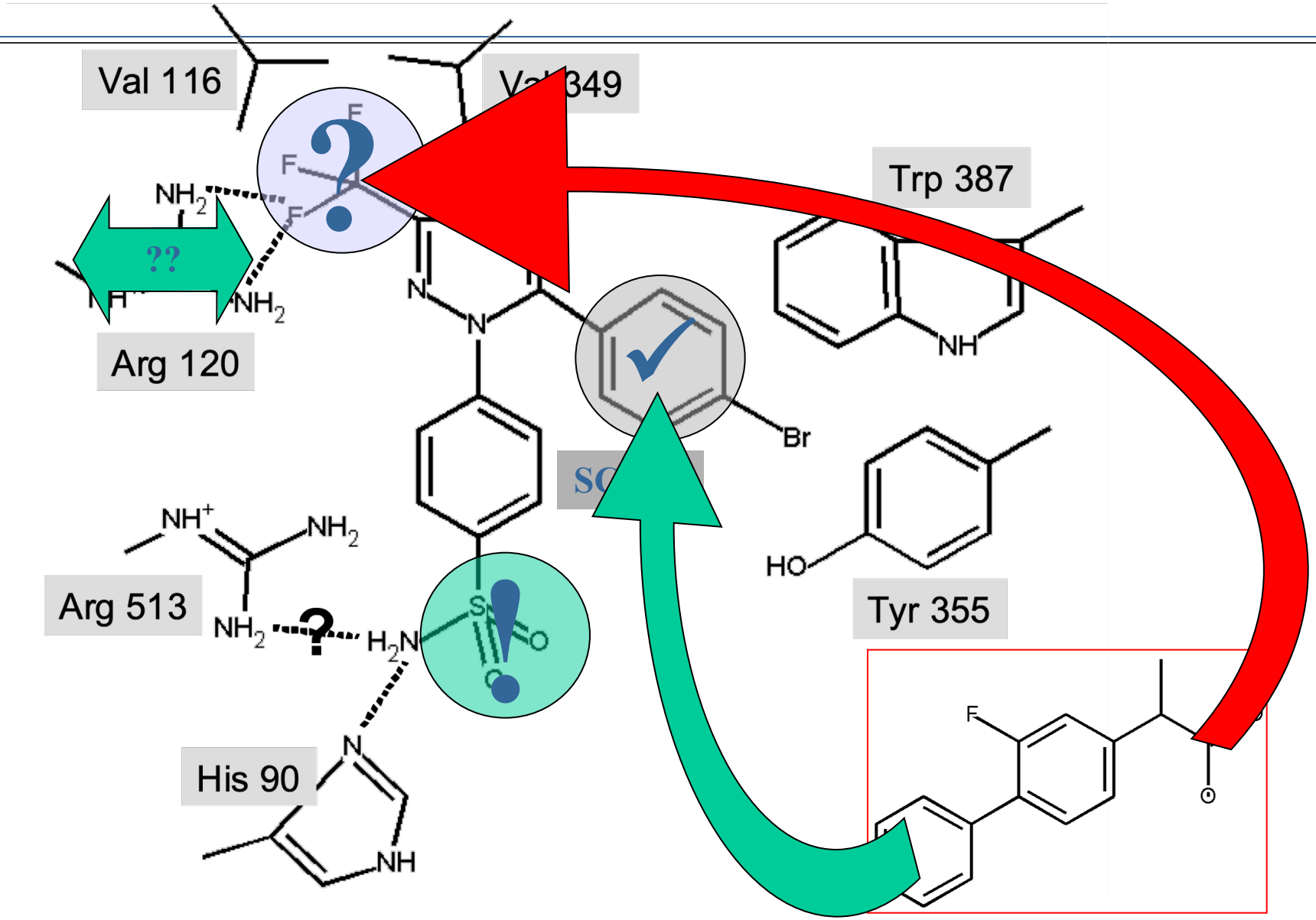
- it manages to explain the Cox_2 activities of the apparently unrelated nonspecific Cox_1/Cox_2 inhibitors:





- This is an ideal scenario – scaffold-independent model trained on thousands of compounds: **so *maybe* the overlay models are *mechanistically relevant* !**

… or maybe not!


Val 116
Val 349
Trp 387
Arg 120
??
Arg 513
His 90
Tyr 355
Br
HO
NH
NH2
NH+
H2N
F

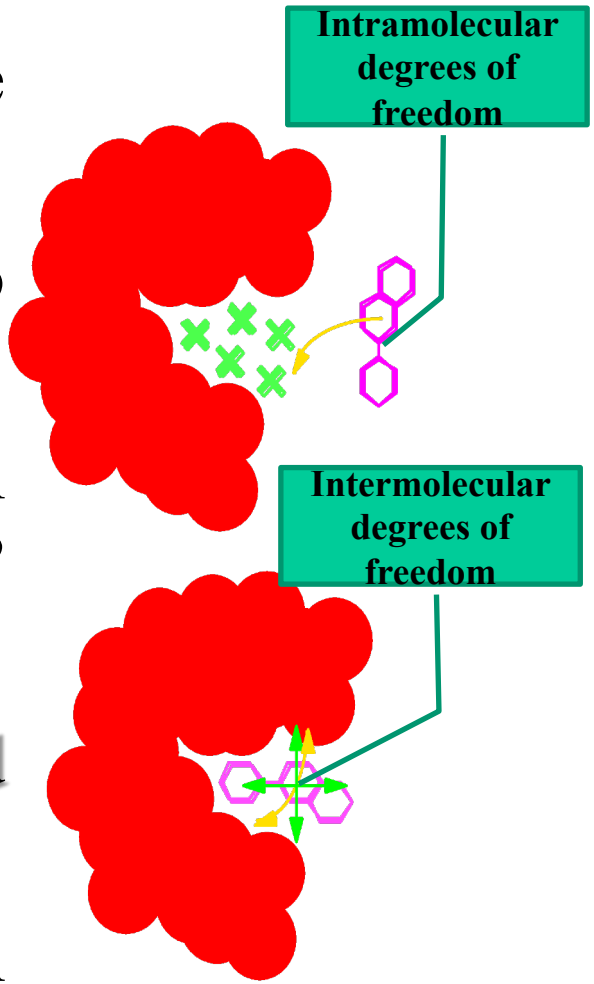
3. Structure-Based Drug Design: Exploiting Knowledge of the Target Structure

Target-Derived Pharmacophore Models

Docking: Simulating the Behavior of the Putative Ligand in Presence of the Target

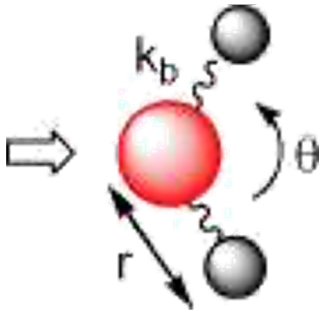
Docking: Conformational Sampling of a Ligand in presence of the target binding site

- place a ligand conformer at the some point of the site.
- rotate & translate ligand with respect to site, and…
- …simultaneously turn rotatable bonds in the ligand (and protein side chains? backbone?)…
- …in order to optimize the site-ligand interaction energy.
- repeat the optimization procedure from other starting points



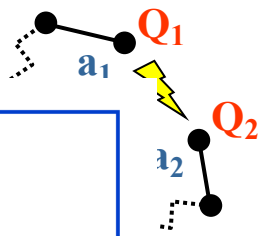
The Energy Function: Molecular Mechanics, Force Fields and Newton's Comeback…

- Quantum chemical calculations are too time-consuming. Atoms are approximated as "classical" interacting spheres.
- Covalent bonds & Valence Angles are modeled as harmonic springs. The energy required to stretch or compress a bond by Δb with respect to its natural length b_0 is expressed as $K_b \Delta b^2$

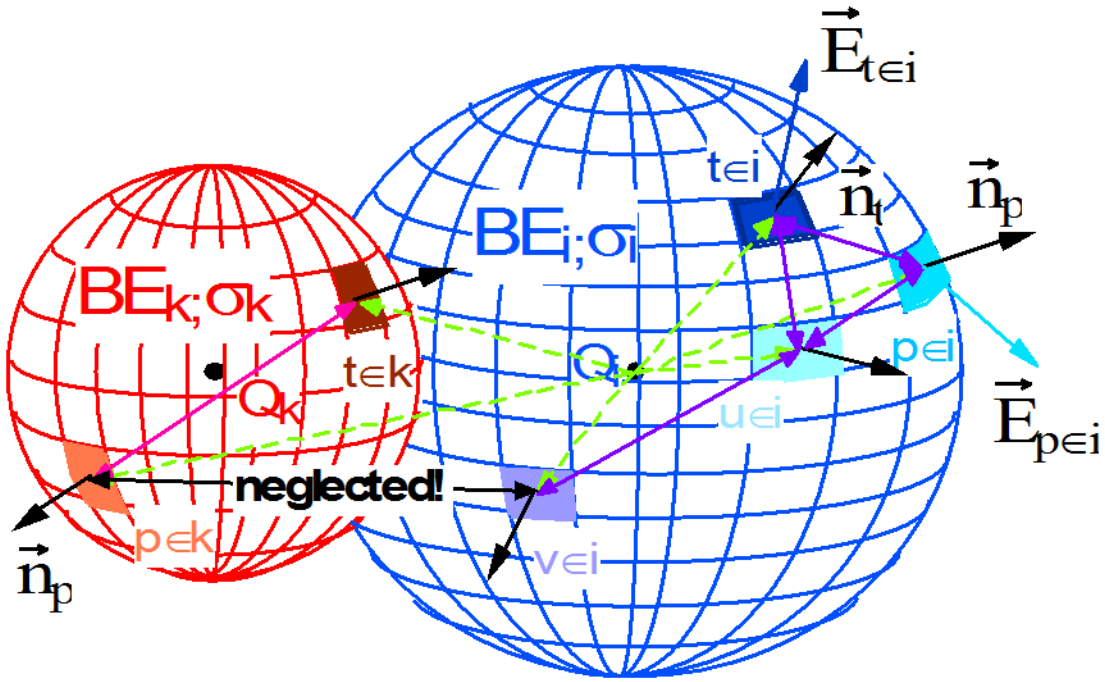


Force Field: Molecular Energy is a (simple?) function of Geometry…

- Non-bonded atoms interact "through space"



E



$$\frac{\sigma_k}{\varepsilon_0} = \langle \vec{E}_p . \vec{n}_p \rangle_{p\in k} \left(1 - \frac{\varepsilon_{ext}}{\varepsilon_{int}}\right) \qquad \frac{\sigma_i}{\varepsilon_0} = \langle \vec{E}_p . \vec{n}_p \rangle_{p\in i} \left(1 - \frac{\varepsilon_{ext}}{\varepsilon_{int}}\right)$$

- Ato … eir che

  – … ge … ay … ers

form the molecular Force Field (FF)

Where do all those parameters come from ?

- Few are directly issued from experimental observations
  - bond & angle deformation constants relate to IR vibration frequencies
  - van der Waals parameters can be measured… for ideal gas atoms.
- Atomic partial charges from electronegativity equilibration, molecular orbital "collapsing".
- Most are *fitted* (did you miss QSPR?), making sure that force field simulations reproduce:
  - experimentally determined geometries & interconformational barriers
  - *Quantum-chemically determined potential energy landscape.*

Docking-driven Virtual Screening…

olecu

¡Viva la Energia *Libre*, camaradas!

$$F_{dock} = -k_B T \ln \sum_{docked} exp\left(-\frac{E_i}{k_B T}\right)$$

$$F_{unbound} = -k_B T \ln \sum_{unbound} exp\left(-\frac{E_i}{k_B T}\right)$$

$$\Delta F = F_{docked} - F_{unbound}$$

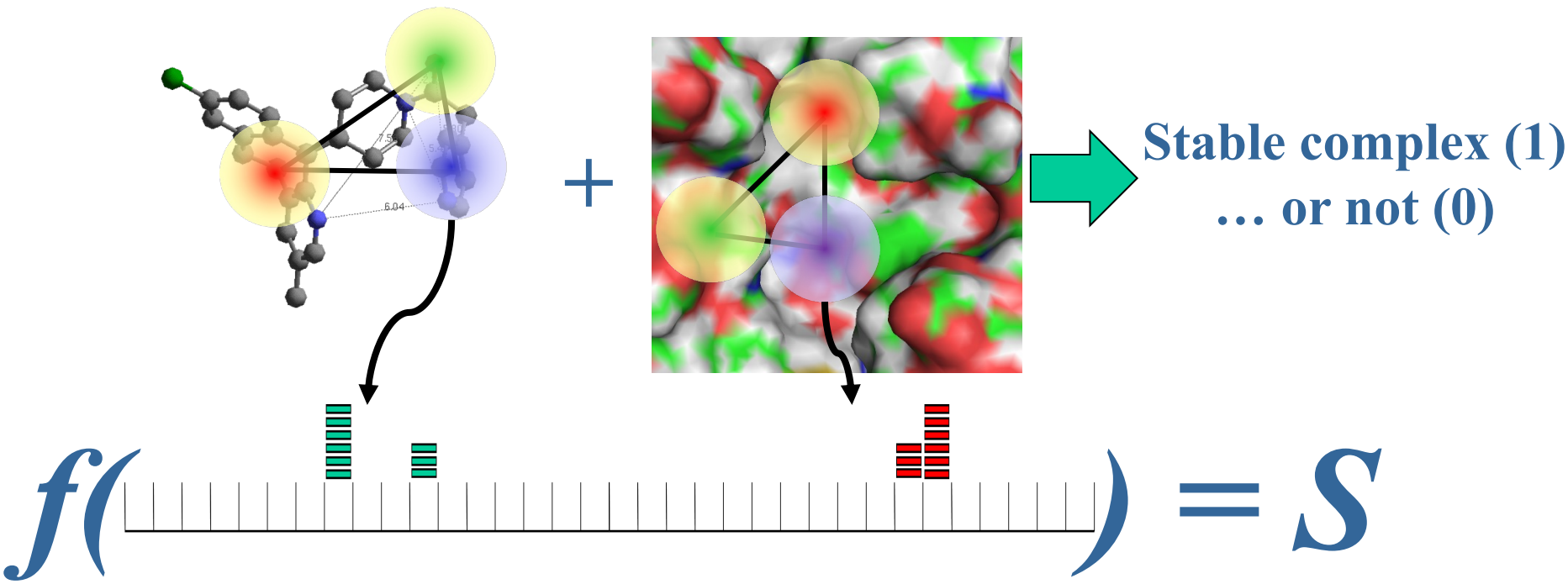
Scoring Functions: The Revenge of QSPR over Boltzmann's Ensemble Physics

- We can **not** apply rigorous statistical physics:
  - We cannot – *at least not without using large-scale computing facilities* – enumerate all the relevant states of the Ensemble.
  - The inaccuracy of force field energies goes way beyond $k_B T \approx 0.6$ kcal/mol – so much for $exp(-E/k_B T)$
- We may try to *fit* a QSPR equation, aiming to predict the binding ΔF from contacts seen in the most stable pose returned by docking, or in crystal structures...
  - Such a construct is called a ***scoring function***.

$$\boldsymbol{\Delta F} = \alpha \times ContactCount_{HBond}^{site-lig} + \beta \times ContactCount_{Hphobic}^{site-lig} + \gamma \times E_{vdW}^{site-lig} + \delta \times BlockedTorsionCount + \ldots$$

Yet, Docking is not the only way to account for the target structure: ProteoChemometrics…

- Since Docking is a sophisticated QSAR, with descriptors based on predicted site-ligand interactions, can't we do this *without* predicting these interactions ?



Conclusions…

- **Molecular modeling is far from first-principle science: its key element is empirical learning (QSPR).**
  - … but then, so is (medicinal) chemistry altogether.
- Correlation is not causality… it's *correlation!*
  - So, if correlations observed within the training set do apply to other molecules, forget metaphysical afterthoughts and exploit them, in successful virtual screening
  - However, an in-depth analysis of the model – *if feasible* – may reveal intrinsic limitations and pitfalls, and help to better delimit the AD.
- Training set information-richness & diversity is the key!
  - what hasn't been taught cannot be known! **Do not blame the machine…***(unless it's Windows-based)*

More Conclusions…

- If a big pharma manager asks you "So, is QSAR useful?", please reply "**Compared to what?**"
- A wrong QSAR model may nevertheless ring a bell in a medicinal chemist's brain, and help to make right decisions
- Rely on the accumulated knowledge, and use QSAR *to discover new combinations of known features & succeed in scaffold hopping*
- There are moments when one should put known things aside, and venture out for random search of new paradigm-breaking ligands – *new scaffold, new binding mode, new action mechanism.*