



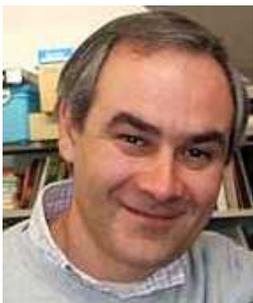
xx.000 to 1?

Is this ratio really acceptable
for a pharma company?

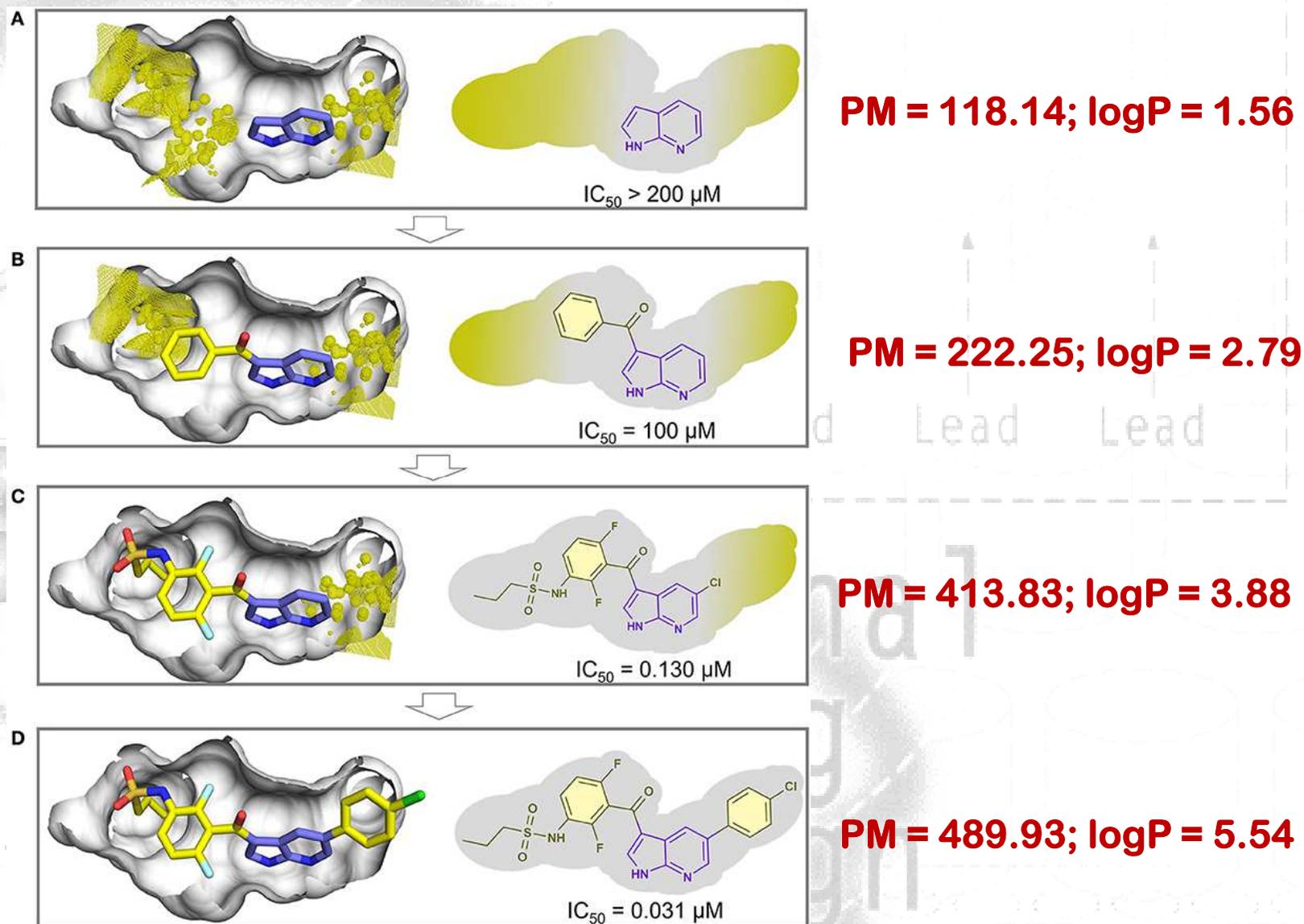


Hit2Lead

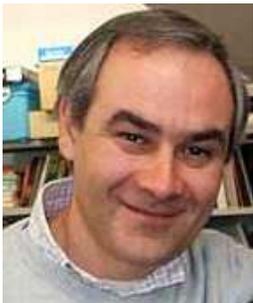




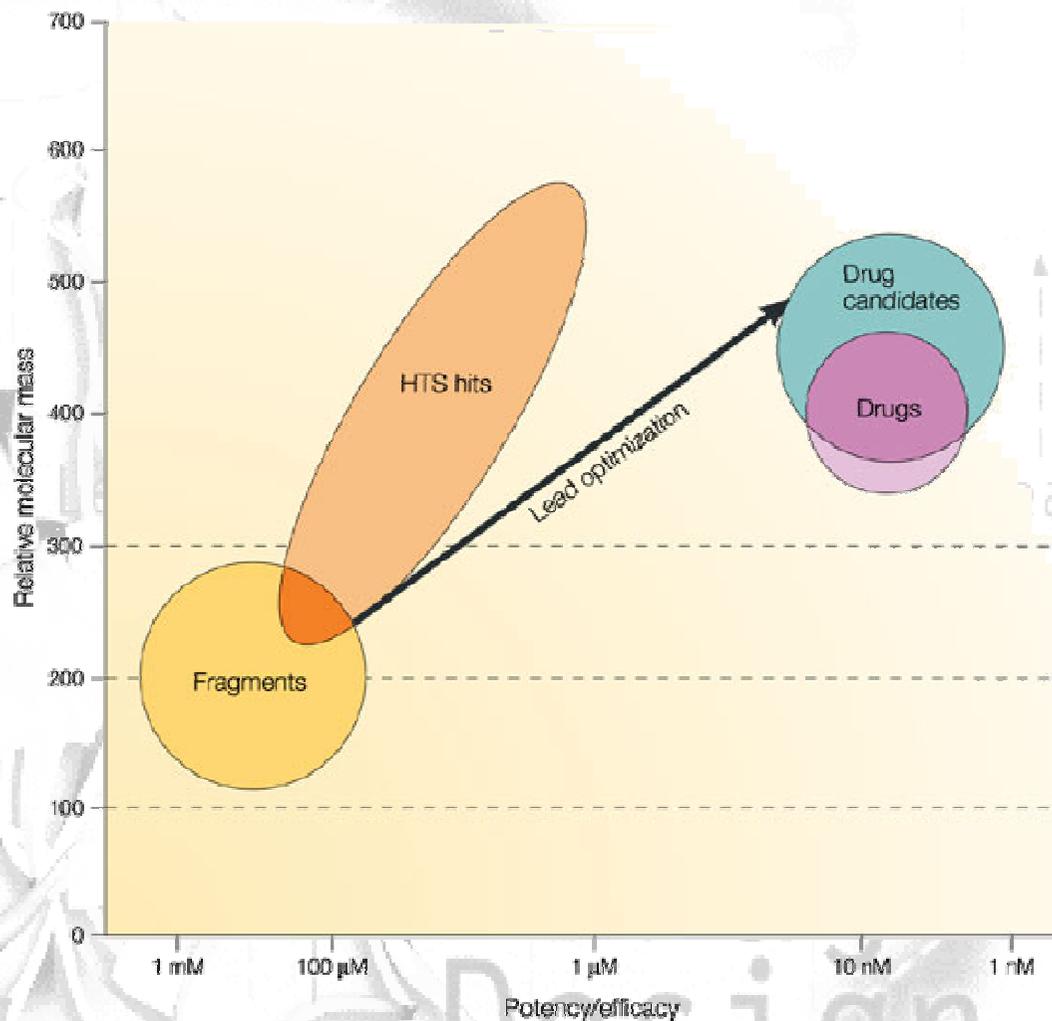
Hit2Lead... a pragmatic view:



Front. Chem., 18 February 2020 | <https://doi.org/10.3389/fchem.2020.00093>



when potency is a good potency?



David C. Rees, Miles Congreve,, Christopher W. Murray & Robin Carr Nature Reviews Drug Discovery 3, 660-672, 2004

Nature Reviews | Drug Discovery

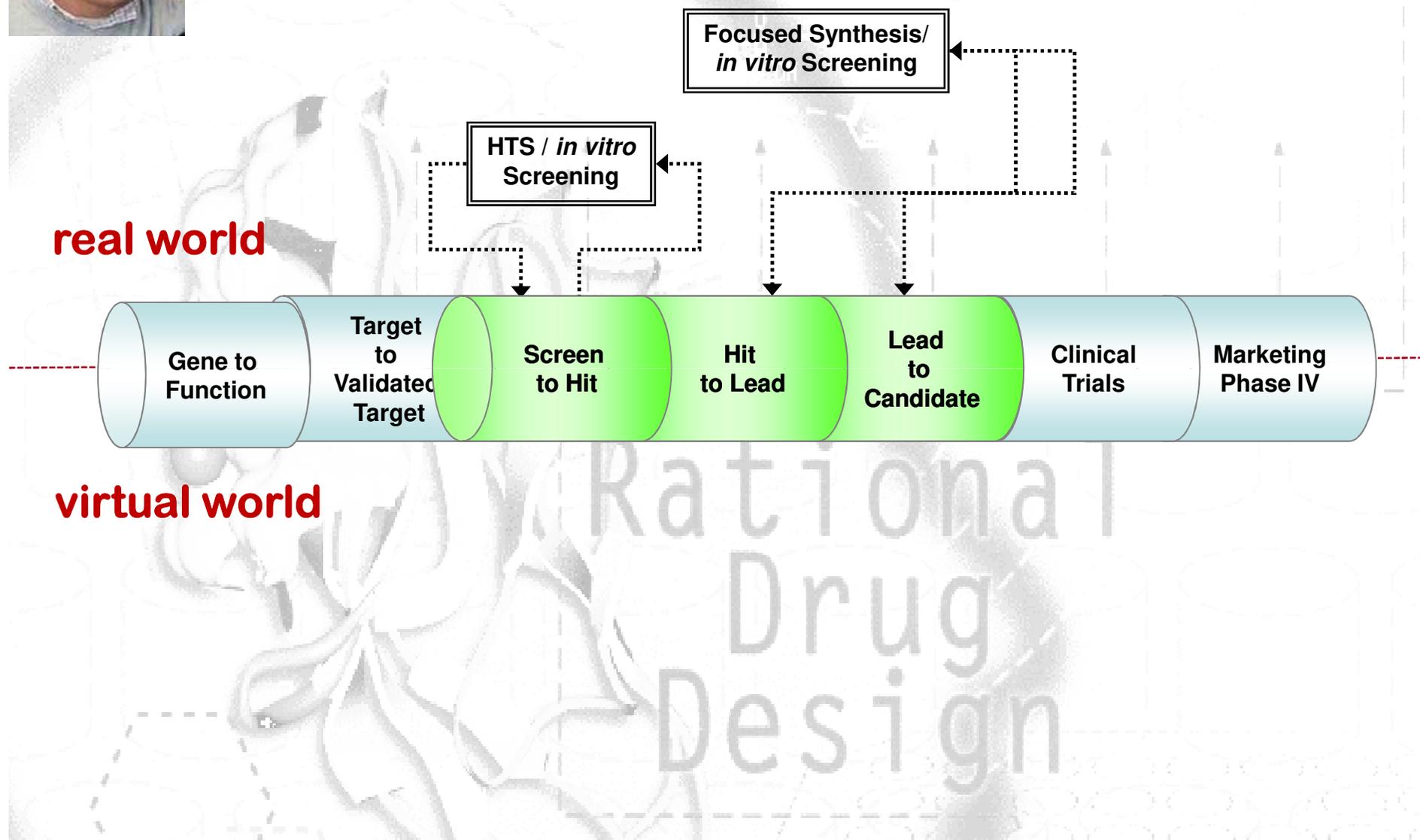


Flowcharts in Drug Discovery:



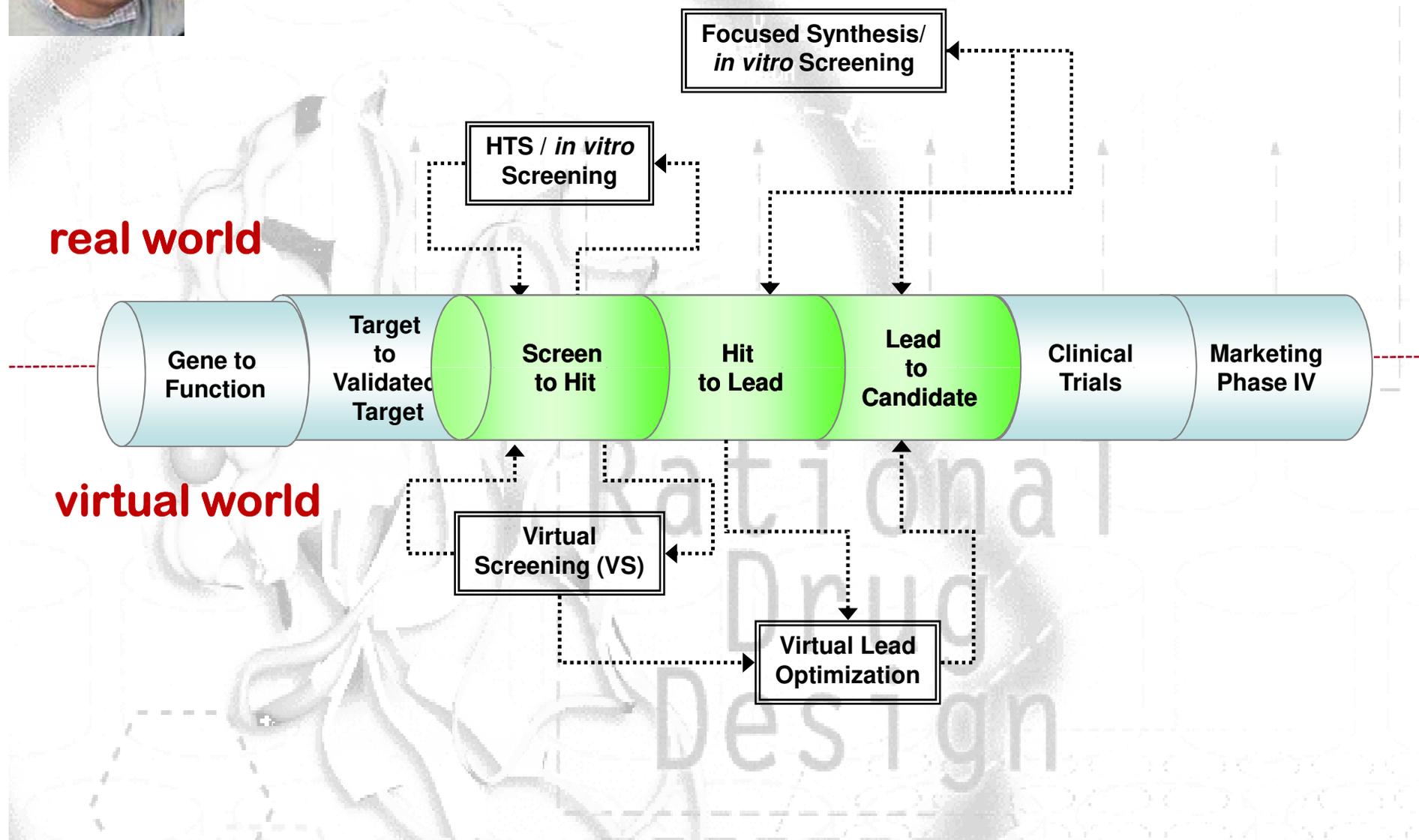


Flowcharts in Drug Discovery:





Flowcharts in Drug Discovery:





Virtual Screening (VS)

Target Structure Determination

Ligand-based Virtual Screening (LBVS)
*Similarity search
Pharmacophore matching*

Structure-based Virtual Screening (SBVS)
HT Molecular Docking





Virtual Lead Optimization

Target Structure Determination

Ligand-based Optimization
Pharmacophore
QSAR



Structure-based Optimization
Molecular Docking
Molecular Dynamics

Menu (touristic) of our PSF course:

- ✓ Ligand-based “Drug” Design (LBDD)
- Structure-based “Drug” Design (SBDD)





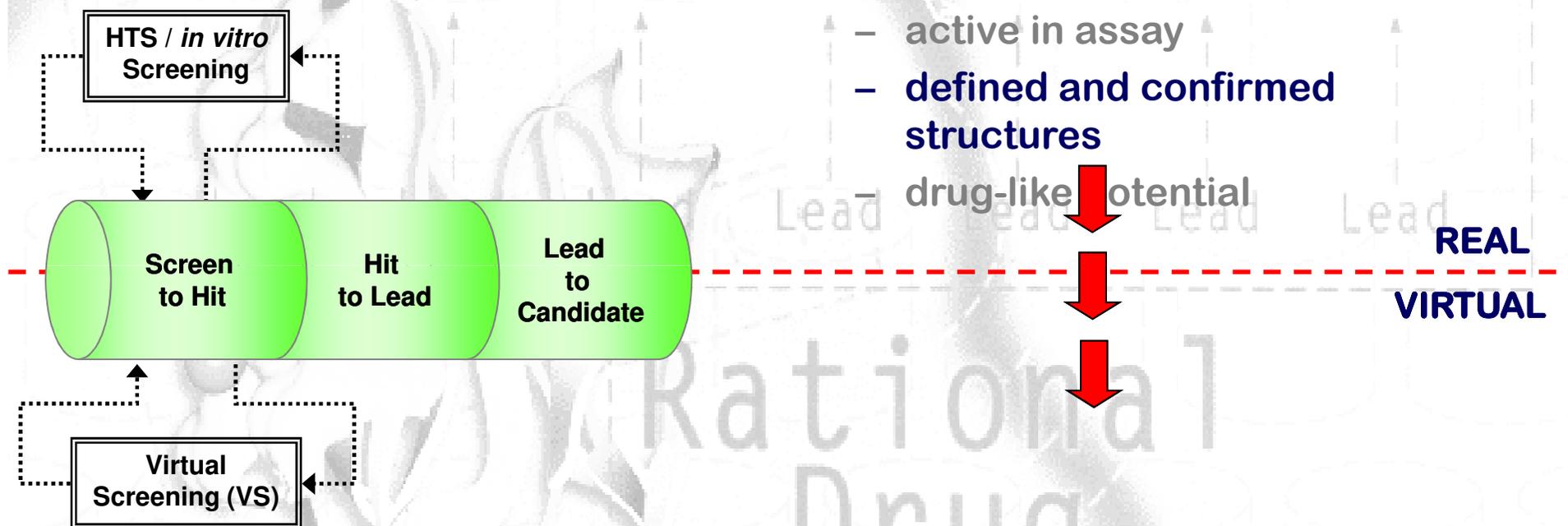
... It is time to start

Hit2Lead

Rational
Drug
Design



We do this first fundamental consideration together:



- **hits**

- active in assay
- defined and confirmed structures
- drug-like potential



The first real help that informatics gives to chemistry and in particular medicinal chemistry is to '*virtualize*' molecular structures.

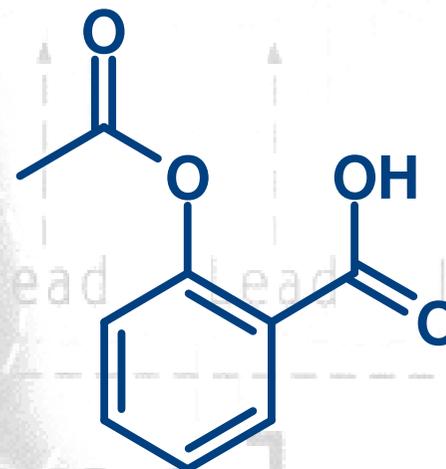
but...



We do NEVER forget:



**This is a
chemical**



**This is one of its
possible chemical
representations**

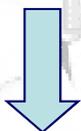


... follow me in this logic comparison:

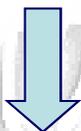
Real world

Virtual world

Chemical Compound (CC)



Chemical Structure (CS)



Chemical Properties (CP)

**Numerical
representations of CS**

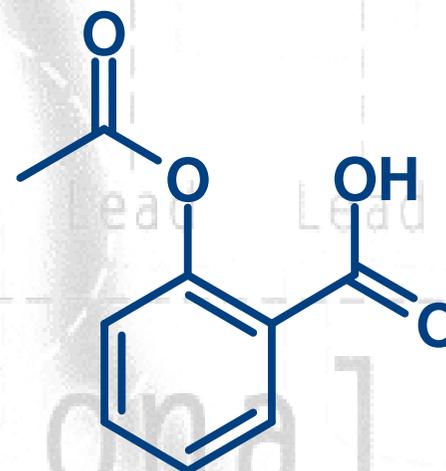


Molecular Descriptors (MD)



With how many chemical representations we can deal?

C₉H₈O₄



...again the salicylic acid?



The crucial informatics differences:

salicylic acid

C9H8O4

... these are simple *strings* (sequences) of alphanumeric characters and they are very easy to manage... informatically speaking!!!

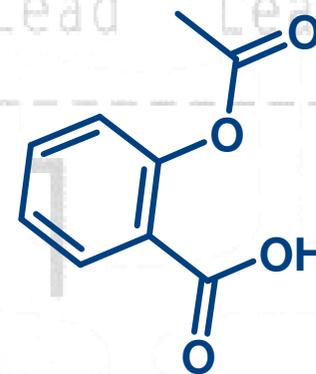
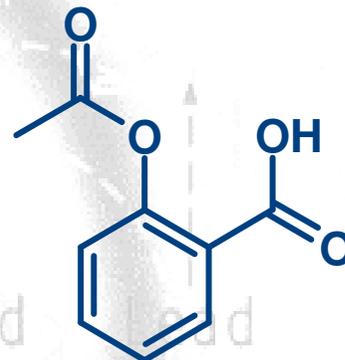


Just a simple example: are these two representations identical?

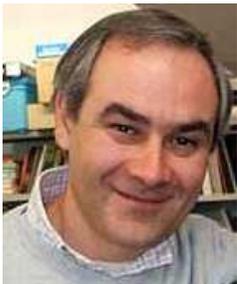
C₉H₈O₄

C₉H₈O₄

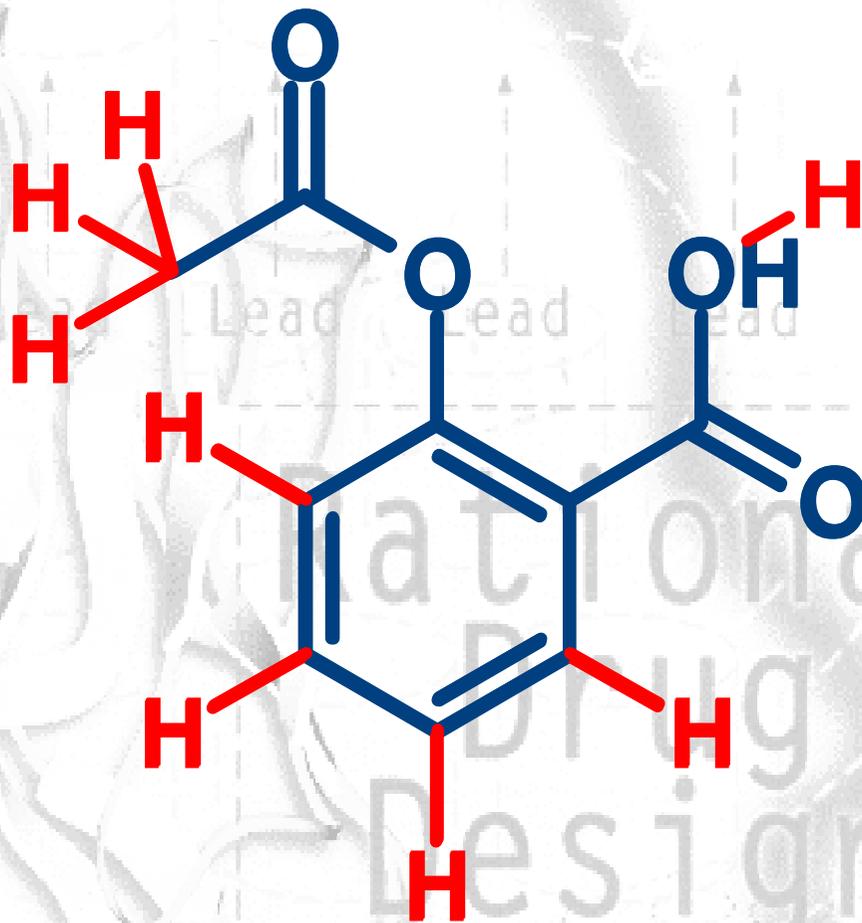
Time of answer (sec):



Time of answer (sec):

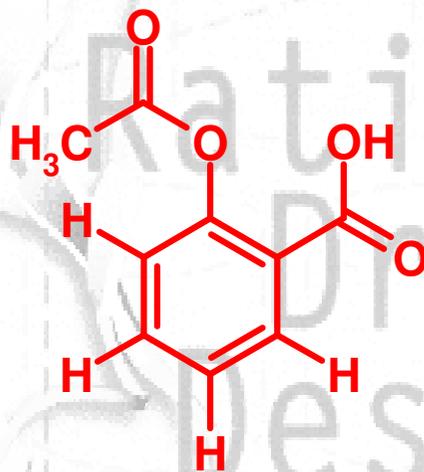
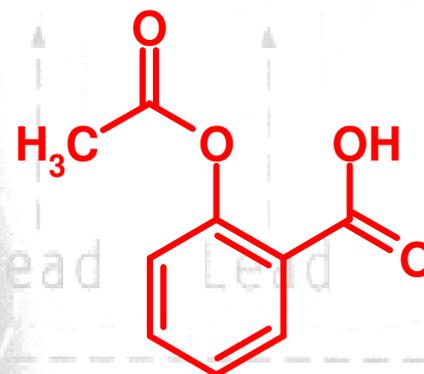
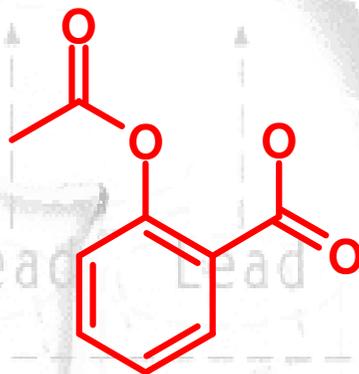
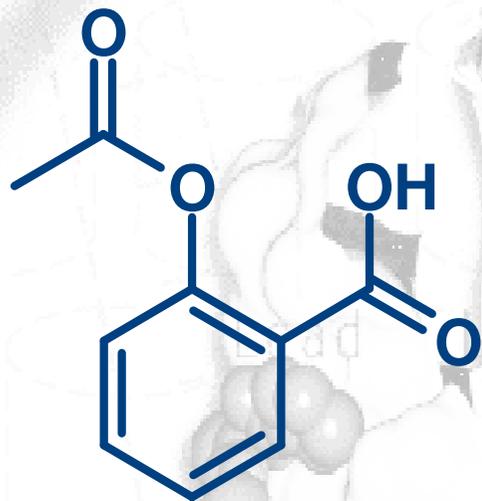


Be careful to the chemical *slang*...





Remember, all of these are not identical... informatically specking!

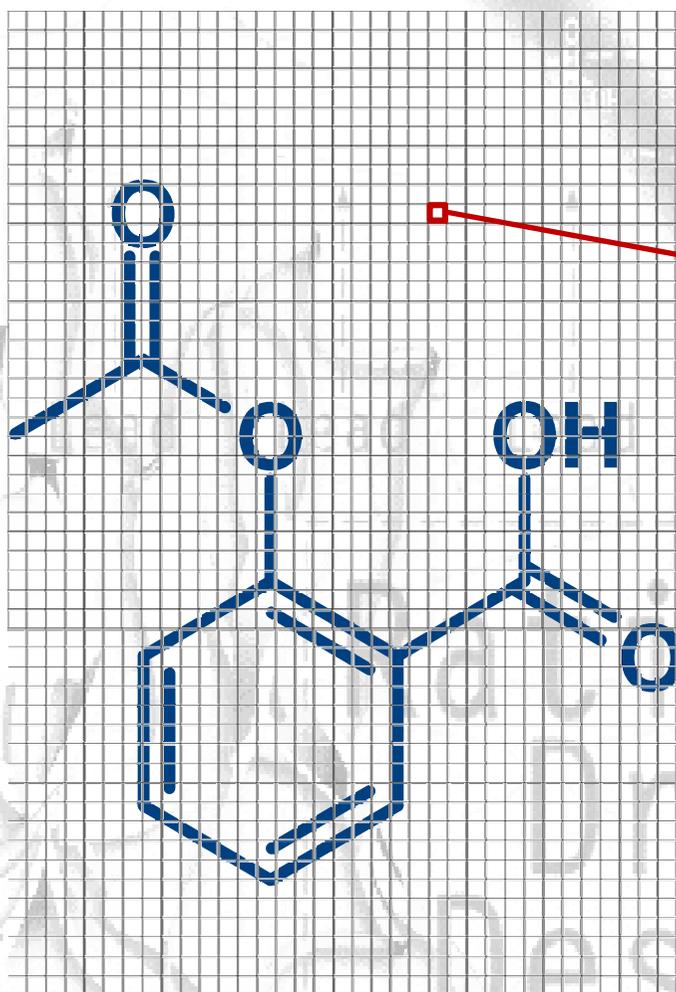


...

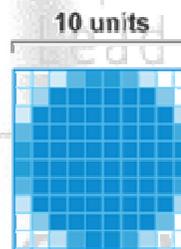


What is the informatics anatomy of this representation:

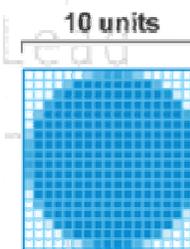
Image
(gif, jpg...)



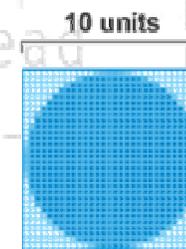
PIXEL



Low Density



Medium Density

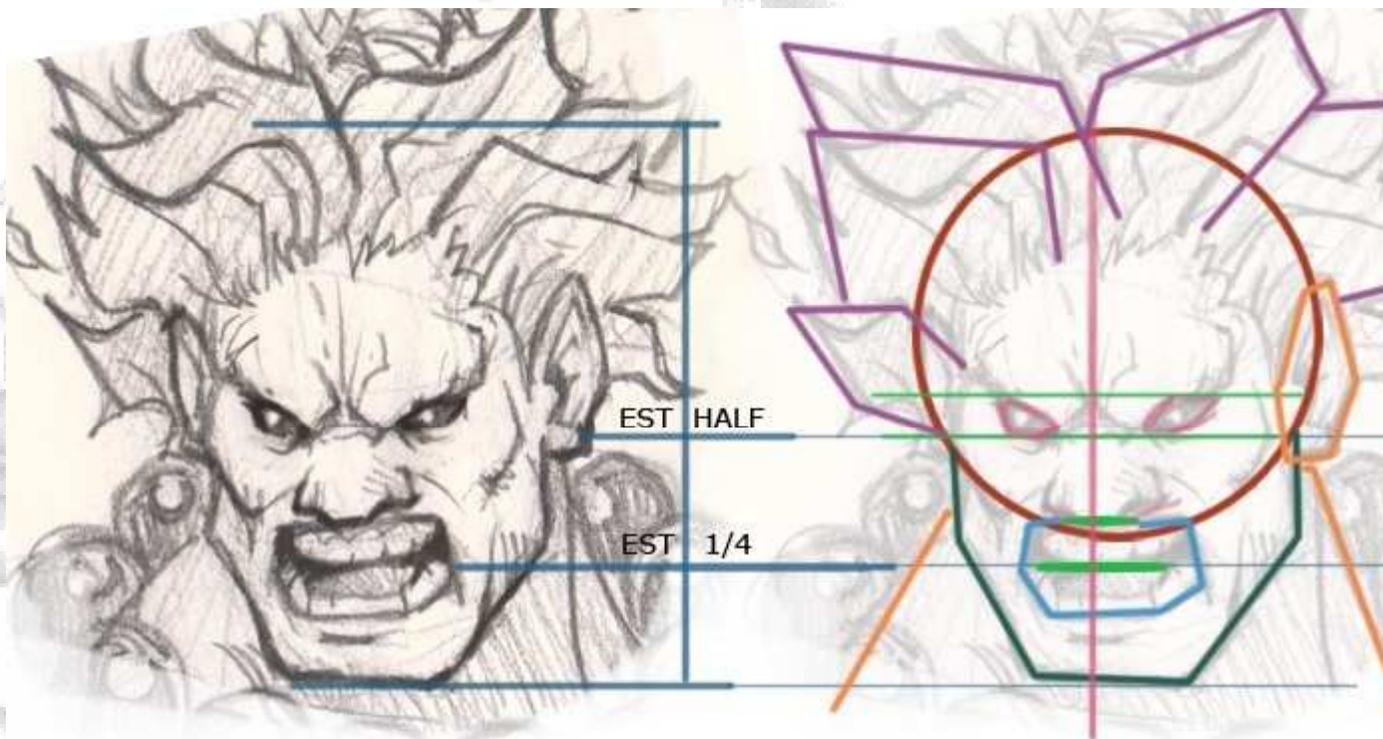


High Density

A PIXEL is generally thought of as the smallest single component of a digital image. Pixels per inch (ppi) and pixels per centimetre (ppcm or pixels/cm) are measurements of the pixel density of an electronic image device.



or...

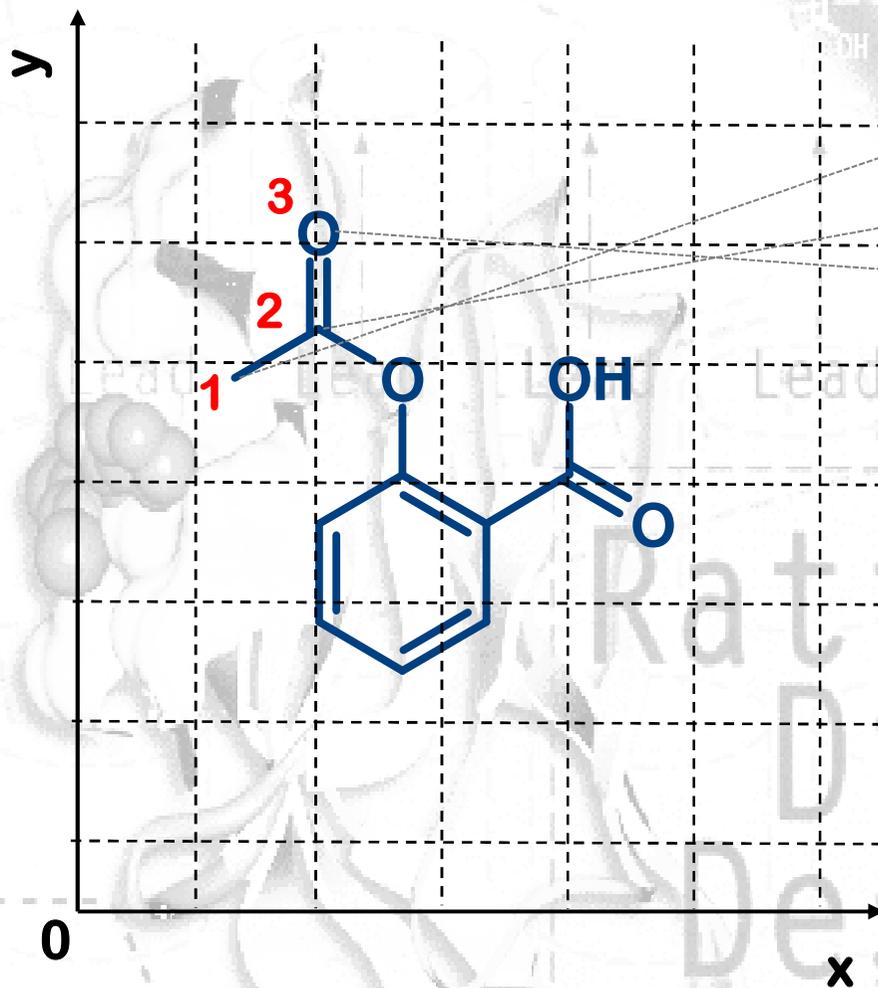


... draw is geometry...

... geometry is numbers in space!!!



geometry is numbers in the space:



Cartesian coordinate table:

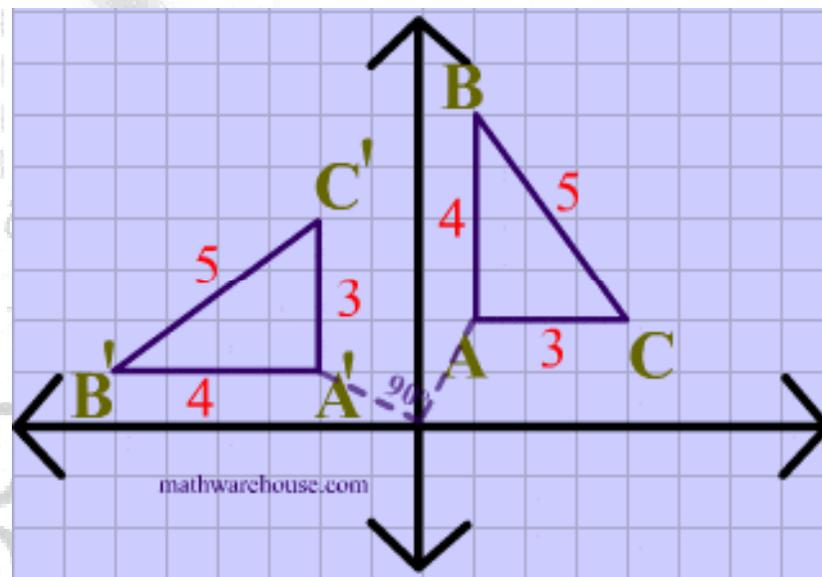
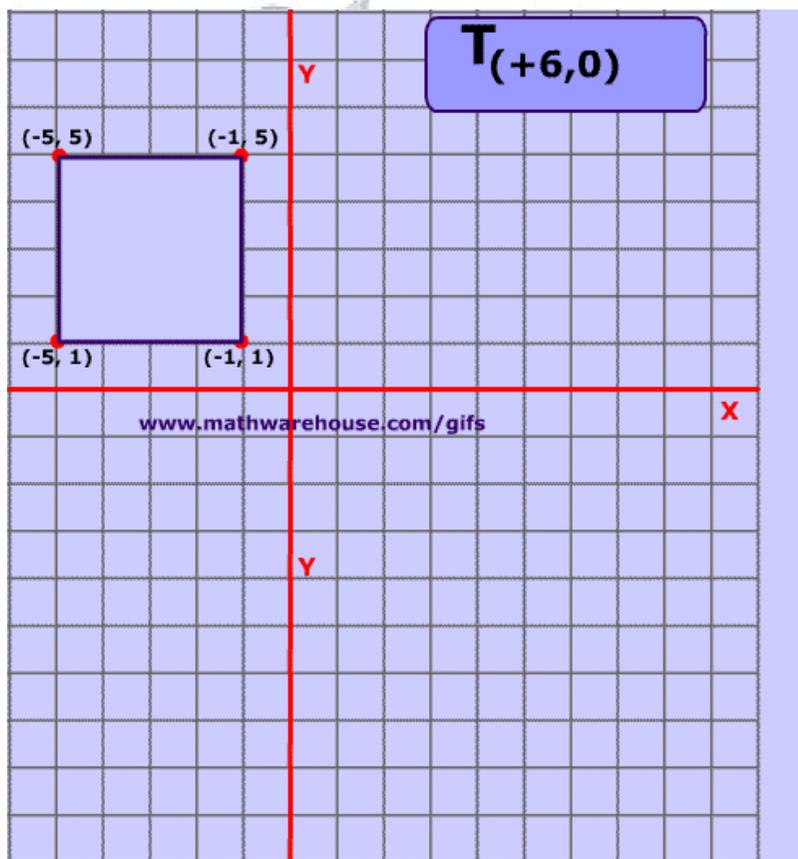
#	Atom Type	x	y
1	C	x_1	y_1
2	C	x_2	y_2
3	O	x_3	y_3
4

Chemical connection table:

#1	#2	Type of bond
1	2	1 (single)
2	3	2 (double)
...



... and any object in a Cartesian plane can be easily translate, rotate, ...:



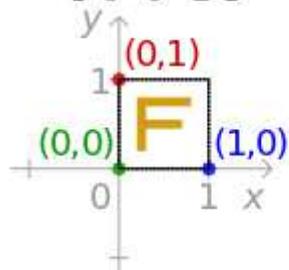


**just for
informatics
addicted!**

Transformations:

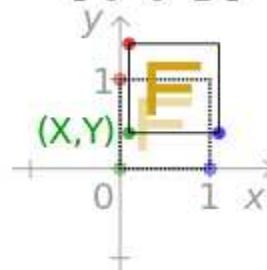
No change

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



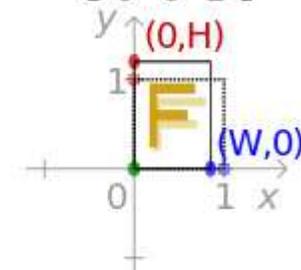
Translate

$$\begin{bmatrix} 1 & 0 & X \\ 0 & 1 & Y \\ 0 & 0 & 1 \end{bmatrix}$$



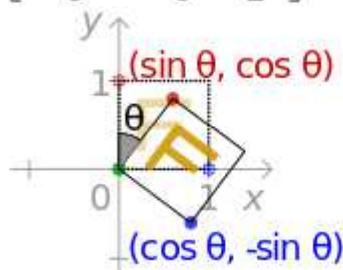
Scale about origin

$$\begin{bmatrix} W & 0 & 0 \\ 0 & H & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



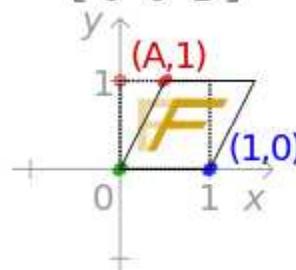
Rotate about origin

$$\begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



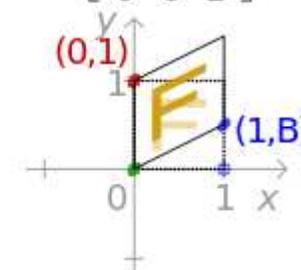
Shear in x direction

$$\begin{bmatrix} 1 & A & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



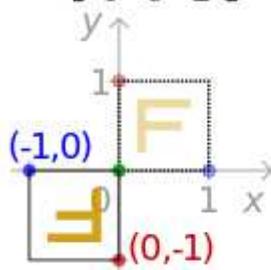
Shear in y direction

$$\begin{bmatrix} 1 & 0 & 0 \\ B & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



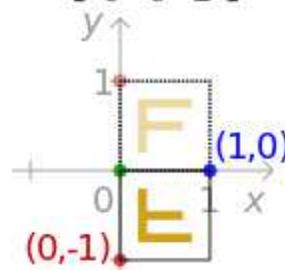
Reflect about origin

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



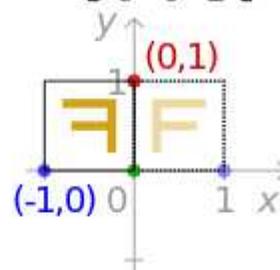
Reflect about x-axis

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



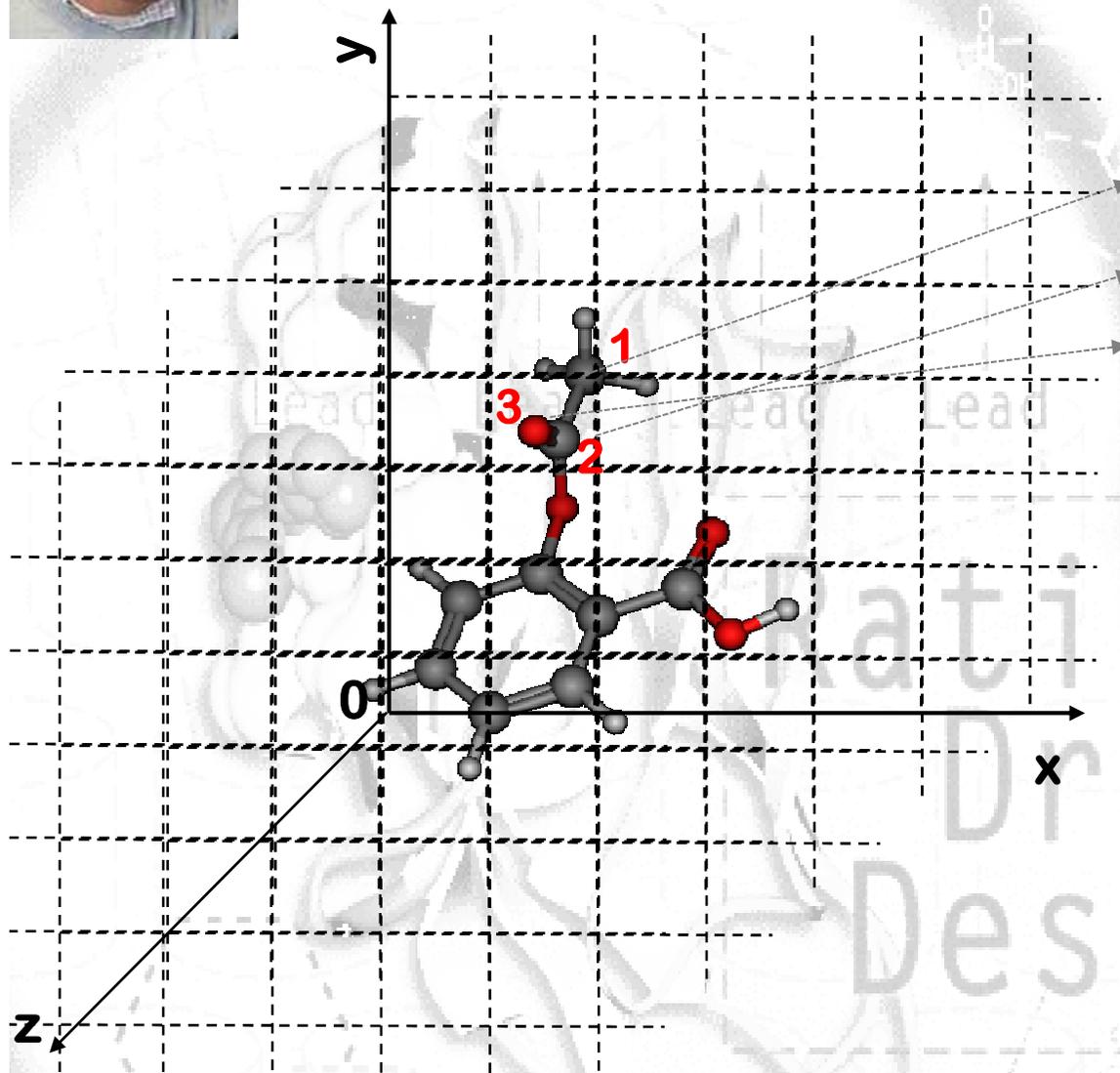
Reflect about y-axis

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$





Now, this is perfect understandable!



Cartesian coordinate table:

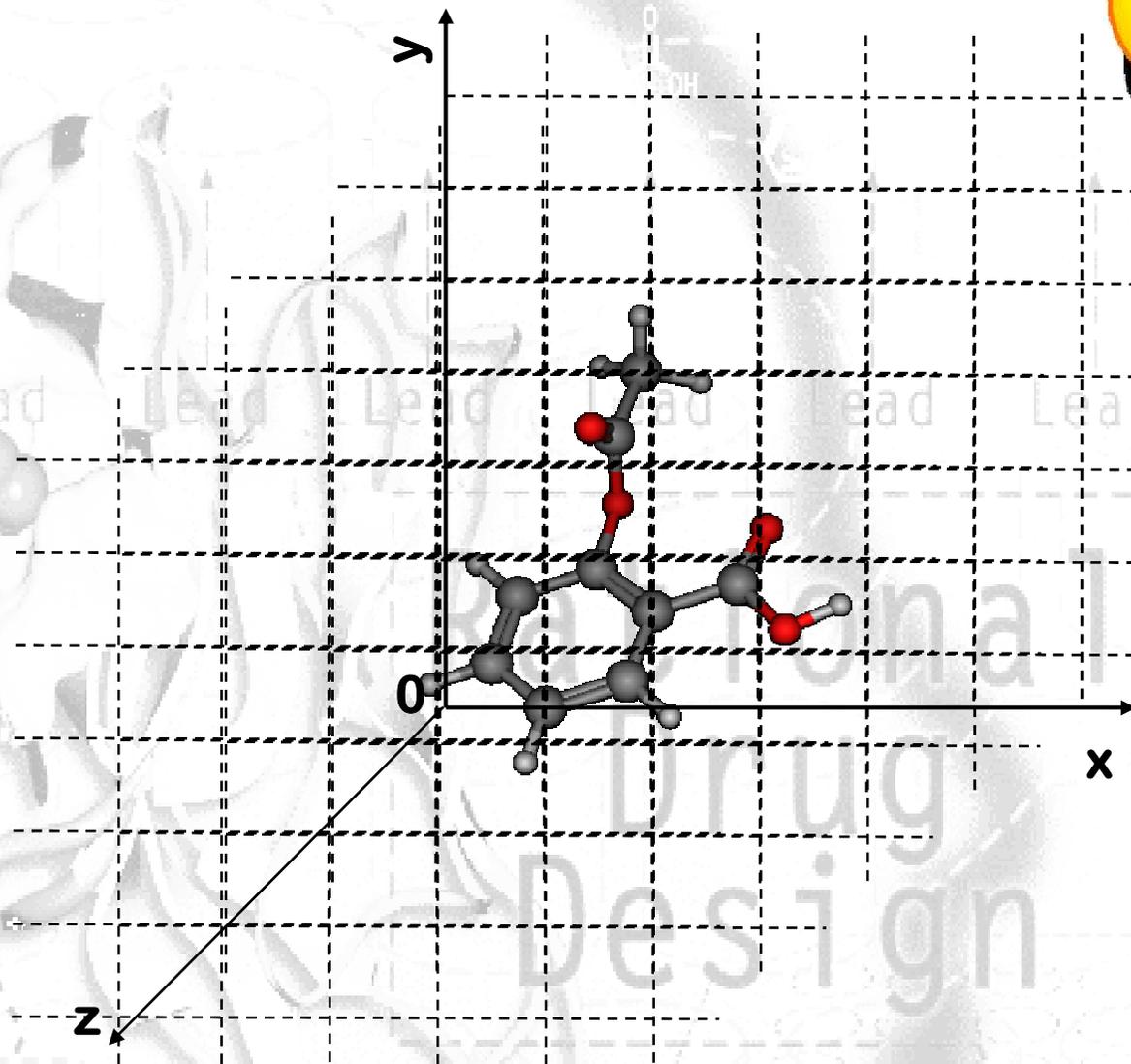
#	Atom Type	x	y	z
1	C	x_1	y_1	z_1
2	C	x_2	y_2	z_2
3	O	x_3	y_3	z_3

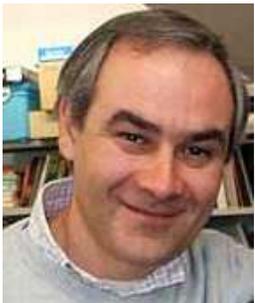
Chemical connection table:

#1	#2	Type of bond
1	2	1 (single)
2	3	2 (double)
...



The crucial question: who gives us the “good” coordinates?

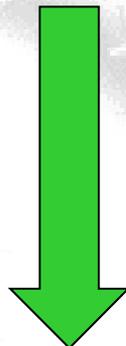




My heroes...



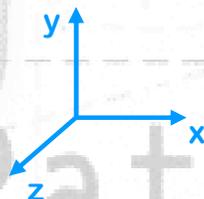
NMR Spectroscopy



X-Ray Crystallography



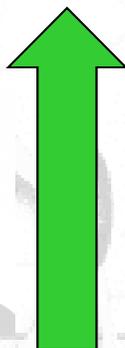
3D

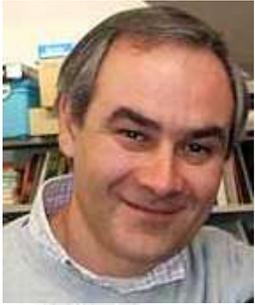


Comparative/Homology Modeling

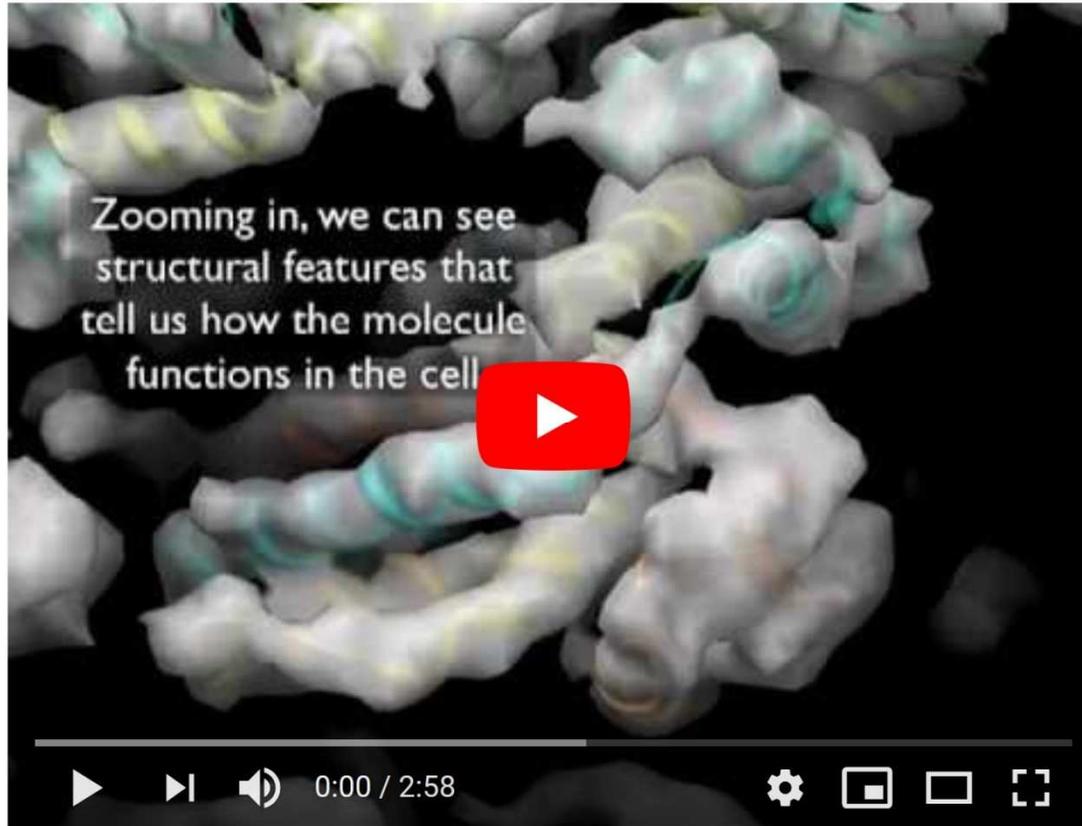


Cryo-Electron Microscopy (Cryo-EM)



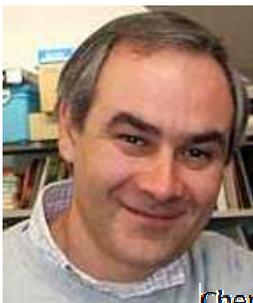


Cryo-EM: the future of structural biochemistry is today!



A 3 minute introduction to CryoEM

credits: <https://www.youtube.com/watch?v=BJKkC0W-6Qk>

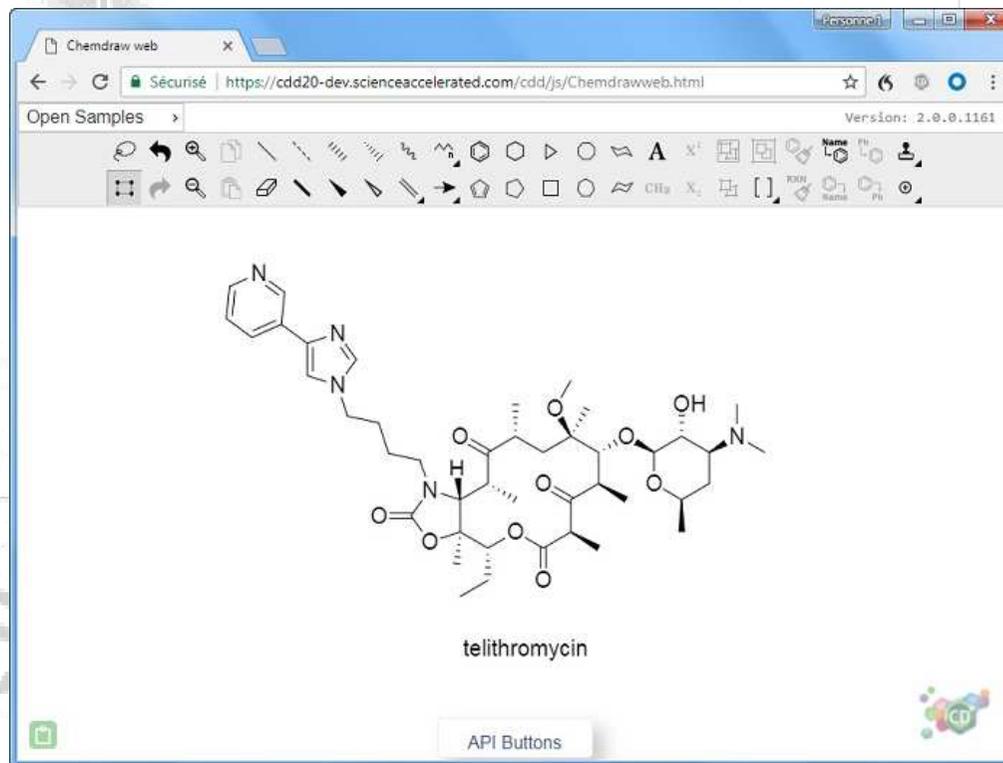
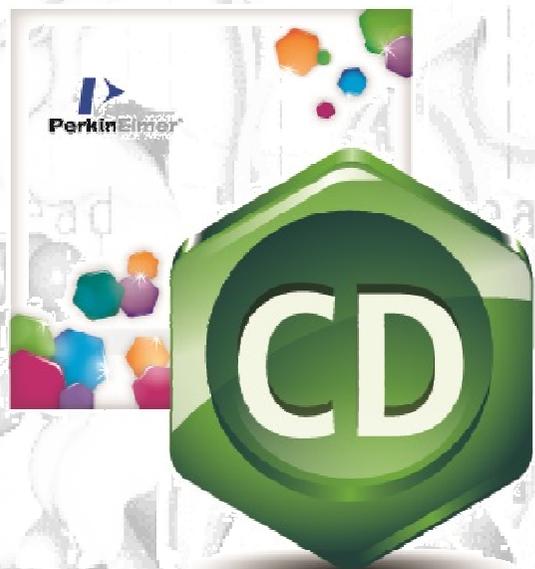


Here is our working platform...

<http://mms.dsfarm.unipd.it>

ChemDraw Prime

ultima modifica 31/10/2019 13:43



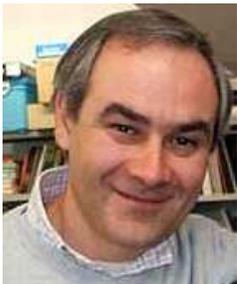
ChemDraw® Prime è un software per il disegno delle strutture molecolari.

Licenza sv rinnovata con contratto triennale: 09.03.2019 - 09.03.2022

Può essere utilizzato dagli [utenti istituzionali](#) collegandosi al seguente link: <http://sitesubscription.cambridgesoft.com/>

[Requisiti sw e hw e consigli per l'installazione](#)

<http://bibliotecachimica.cab.unipd.it/documenti-download/chemdraw-prime>



List of Commonly Used Abbreviations of coordinates archives (files):

xxx.mol: *MDL (Molecular Design Limited, Inc) Molfile*

xxx.mol2: *SYBYL Molfile*

xxx.sdf: *structure-data file*

xxx.pdb: *protein-data bank file*

xxx.chm: *ChemDraw file*

...



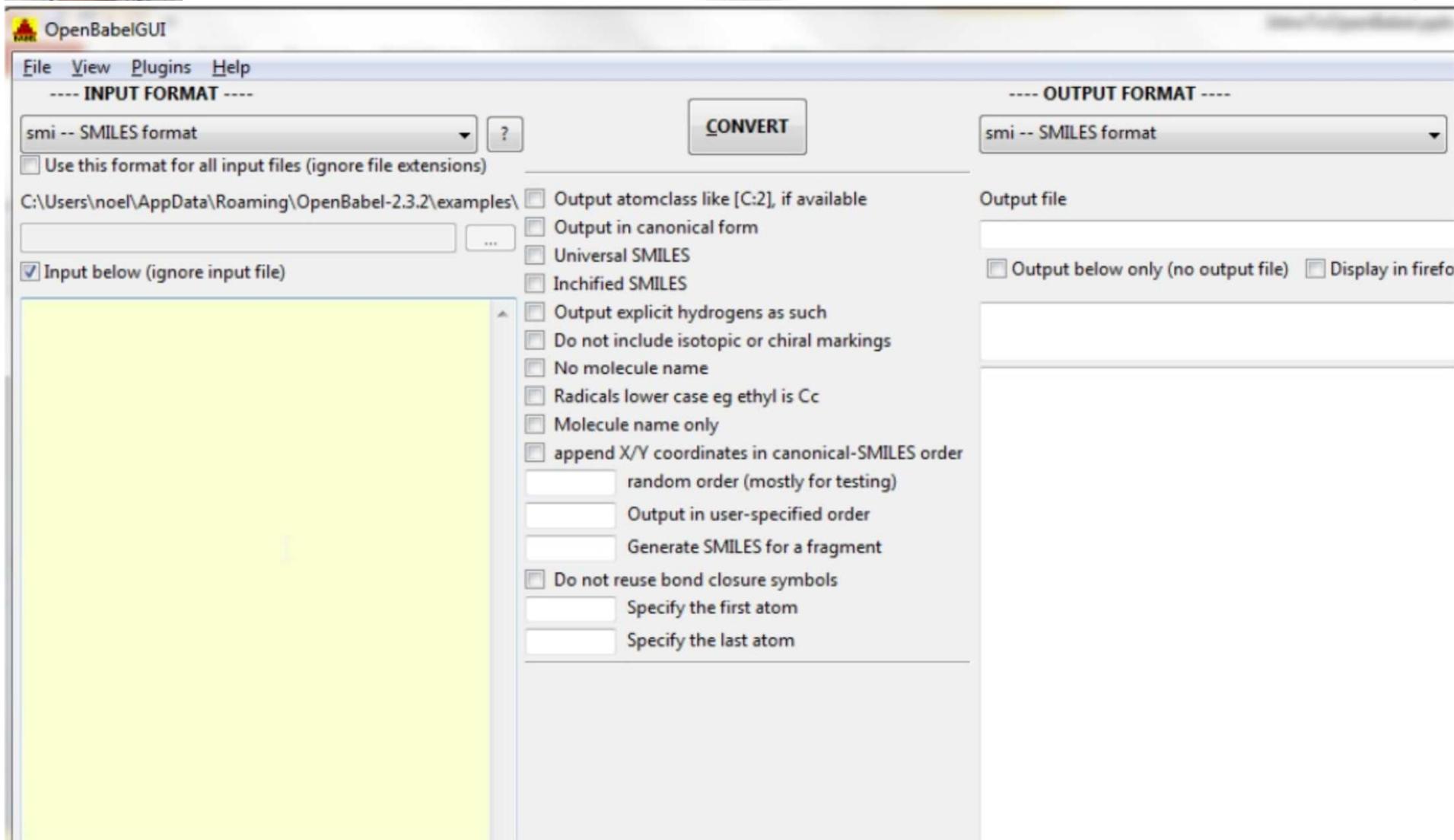
OpenBabel: a precious tool to convert one format to another... and much more!!

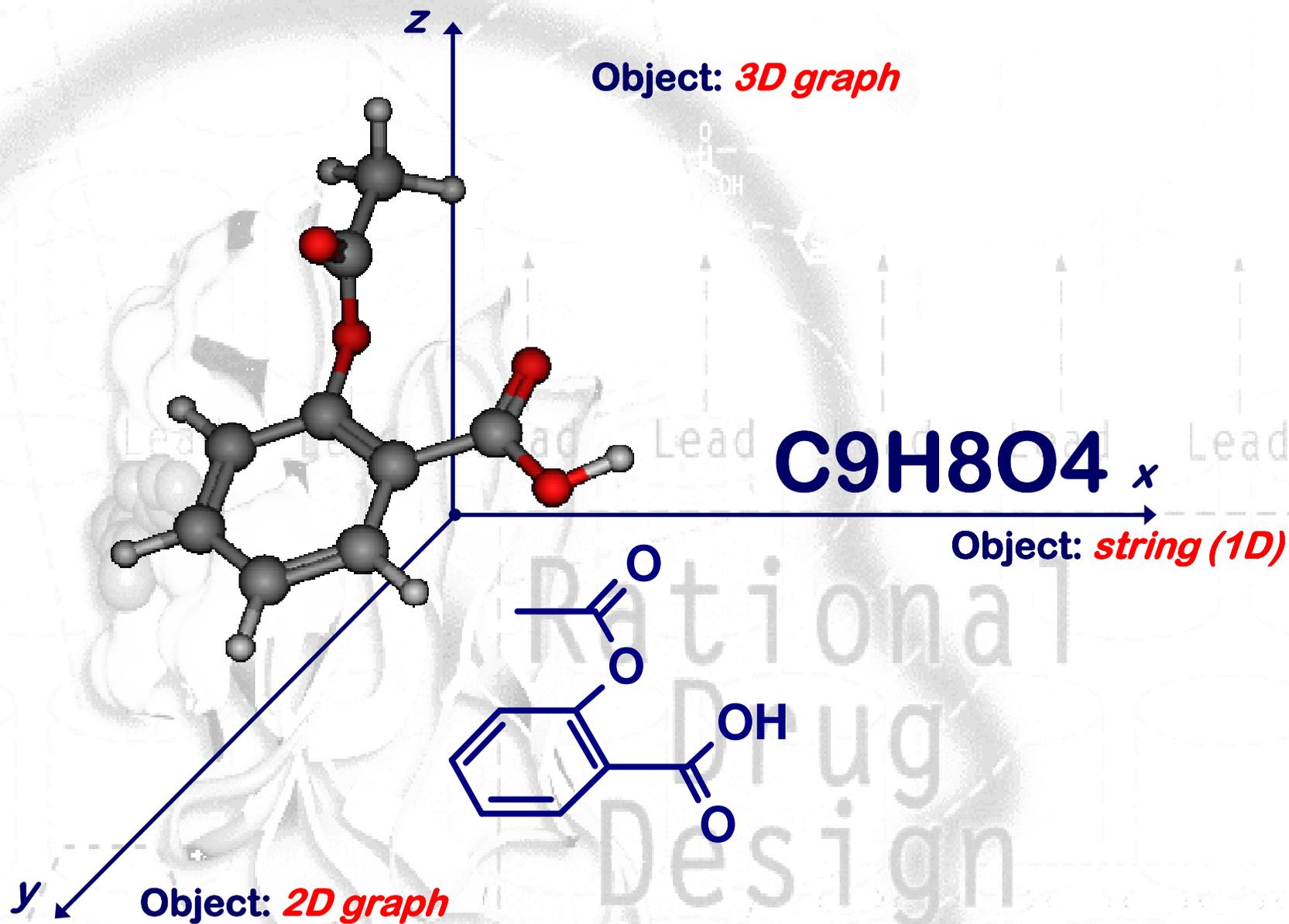


http://openbabel.org/wiki/Main_Page



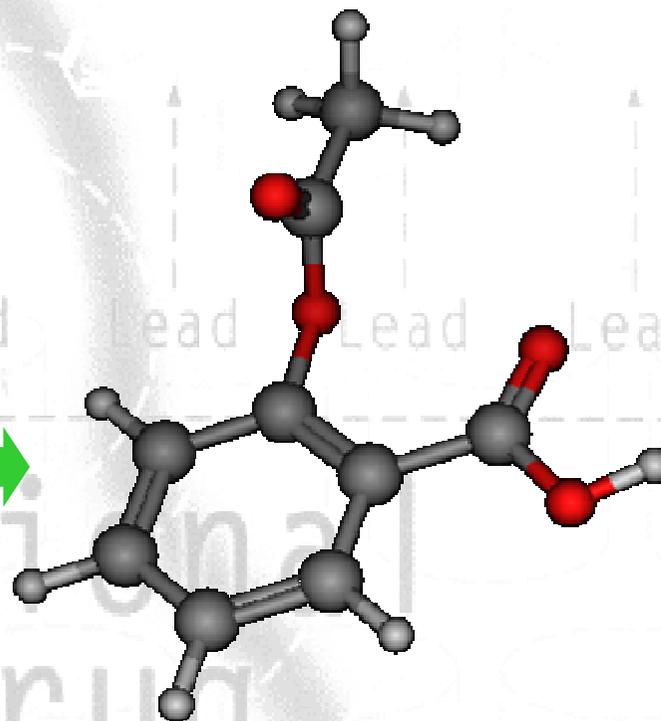
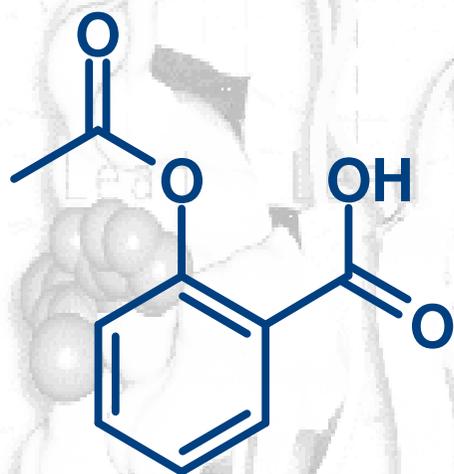
OpenBabel: a precious tool to convert one format to another... and much more!!

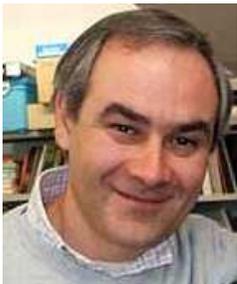




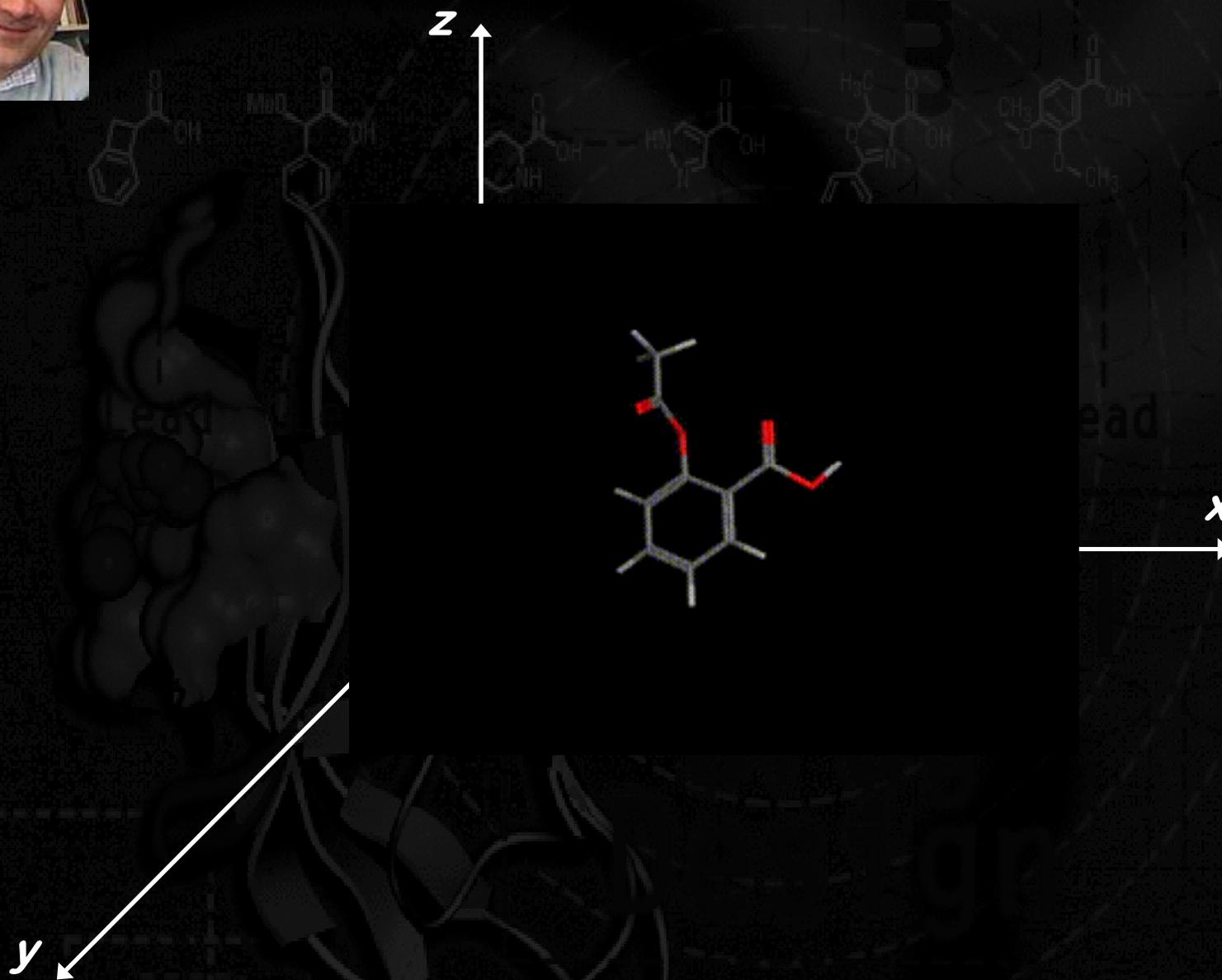


Are we fully satisfied?





What we are really missing...



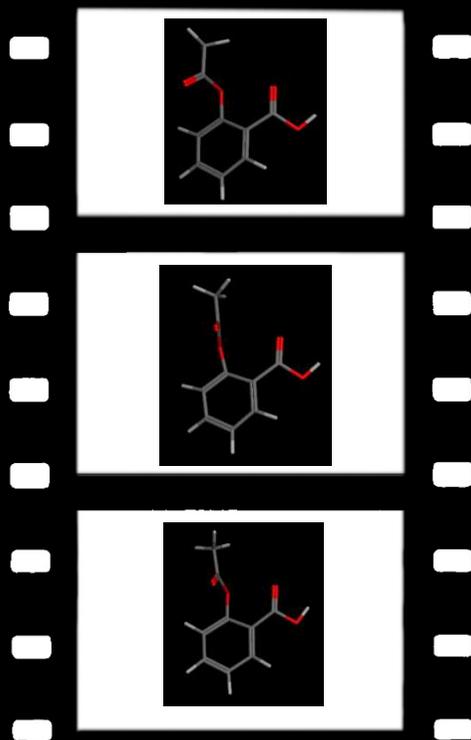
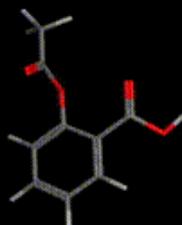


... the illusion of time!





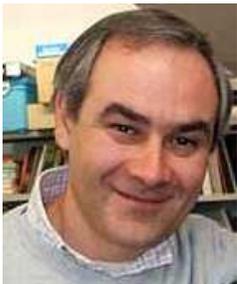
Can you understand what is this?



```
21 21
1.1522 0.5113 -0.0537 C
0.2183 1.2057 -0.8346 C
0.5465 2.3445 -1.5679 O
-1.0985 0.7524 -0.9667 C
-1.8050 1.2920 -1.5911 H
-1.5038 -0.2836 -0.2868 C
-2.5276 -0.7473 -0.3775 H
-0.5938 -1.0870 0.5099 C
-0.9101 -0.9016 1.0411 H
0.7249 -0.6400 0.6244 C
1.4145 -1.2086 1.2449 H
0.5056 3.5056 -0.8113 C
0.9559 4.6613 -1.6596 C
0.9409 5.5777 -1.0629 H
0.2793 4.7842 -2.5073 H
1.9783 4.4900 -2.0055 H
0.1372 3.6103 0.3837 O
2.5419 0.8700 0.0441 C
3.0443 1.8108 -0.5504 O
3.2755 0.2254 0.8972 O
4.1677 0.6310 -0.8800 H
1 10 2
1 2 1
1 18 1
...
20 21 1
M END
```

```
21 21
1.1522 0.5113 -0.0537 C
0.2183 1.2057 -0.8346 C
0.5465 2.3445 -1.5679 O
-1.0985 0.7524 -0.9667 C
-1.8050 1.2920 -1.5911 H
-1.5038 -0.2836 -0.2868 C
-2.5276 -0.7473 -0.3775 H
-0.5938 -1.0870 0.5099 C
-0.9101 -0.9016 1.0411 H
0.7249 -0.6400 0.6244 C
1.4145 -1.2086 1.2449 H
0.5056 3.5056 -0.8113 C
0.9559 4.6613 -1.6596 C
0.9409 5.5777 -1.0629 H
0.2793 4.7842 -2.5073 H
1.9783 4.4900 -2.0055 H
0.1372 3.6103 0.3837 O
2.5419 0.8700 0.0441 C
3.0443 1.8108 -0.5504 O
3.2755 0.2254 0.8972 O
4.1677 0.6310 -0.8800 H
1 10 2
1 2 1
1 18 1
...
20 21 1
M END
```

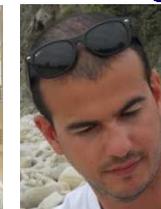
```
21 21
1.1522 0.5113 -0.0537 C
0.2183 1.2057 -0.8346 C
0.5465 2.3445 -1.5679 O
-1.0985 0.7524 -0.9667 C
-1.8050 1.2920 -1.5911 H
-1.5038 -0.2836 -0.2868 C
-2.5276 -0.7473 -0.3775 H
-0.5938 -1.0870 0.5099 C
-0.9101 -0.9016 1.0411 H
0.7249 -0.6400 0.6244 C
1.4145 -1.2086 1.2449 H
0.5056 3.5056 -0.8113 C
0.9559 4.6613 -1.6596 C
0.9409 5.5777 -1.0629 H
0.2793 4.7842 -2.5073 H
1.9783 4.4900 -2.0055 H
0.1372 3.6103 0.3837 O
2.5419 0.8700 0.0441 C
3.0443 1.8108 -0.5504 O
3.2755 0.2254 0.8972 O
4.1677 0.6310 -0.8800 H
1 10 2
1 2 1
1 18 1
...
20 21 1
M END
```



... but this illusion is incredibly exciting!!!

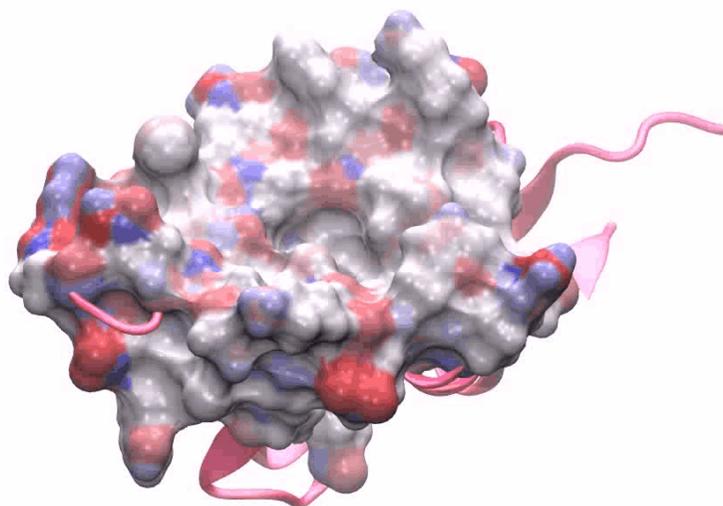
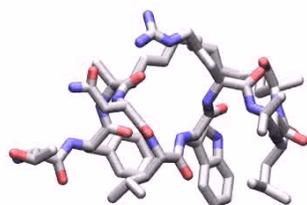


V. Salmaso



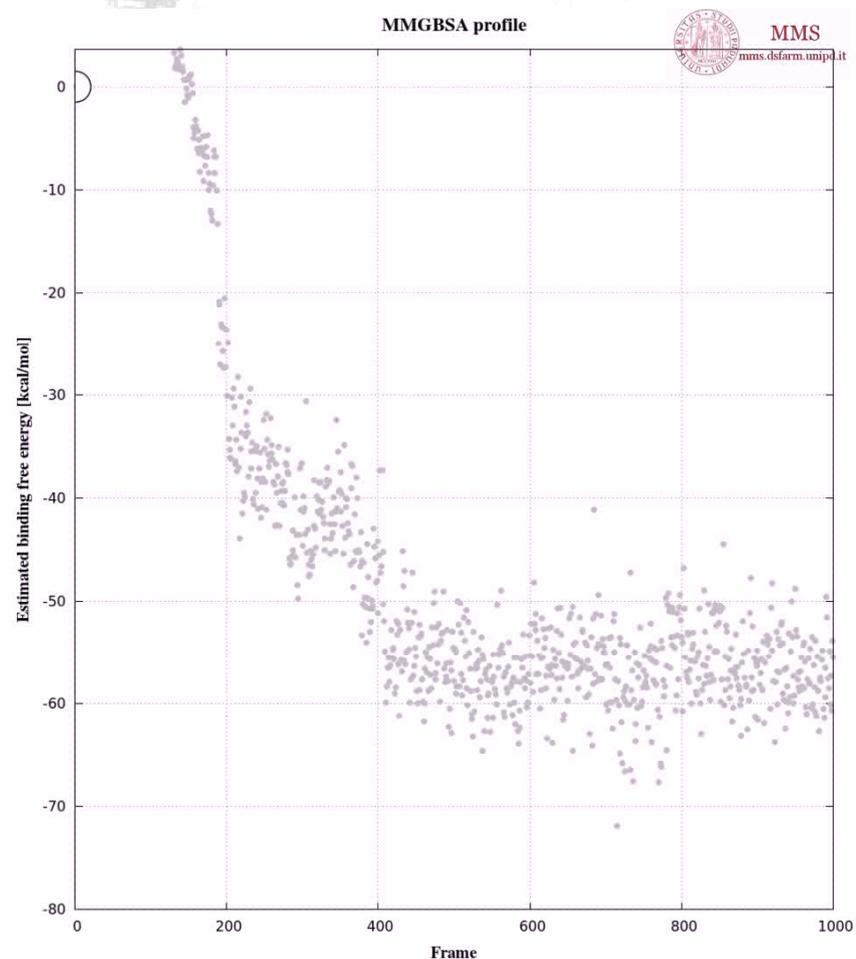
M. Sturlese

SuMD simulation time: 0.01 ns



Stapled p53 Peptide Bound to Mdm2; PDB-ID: 3V3B

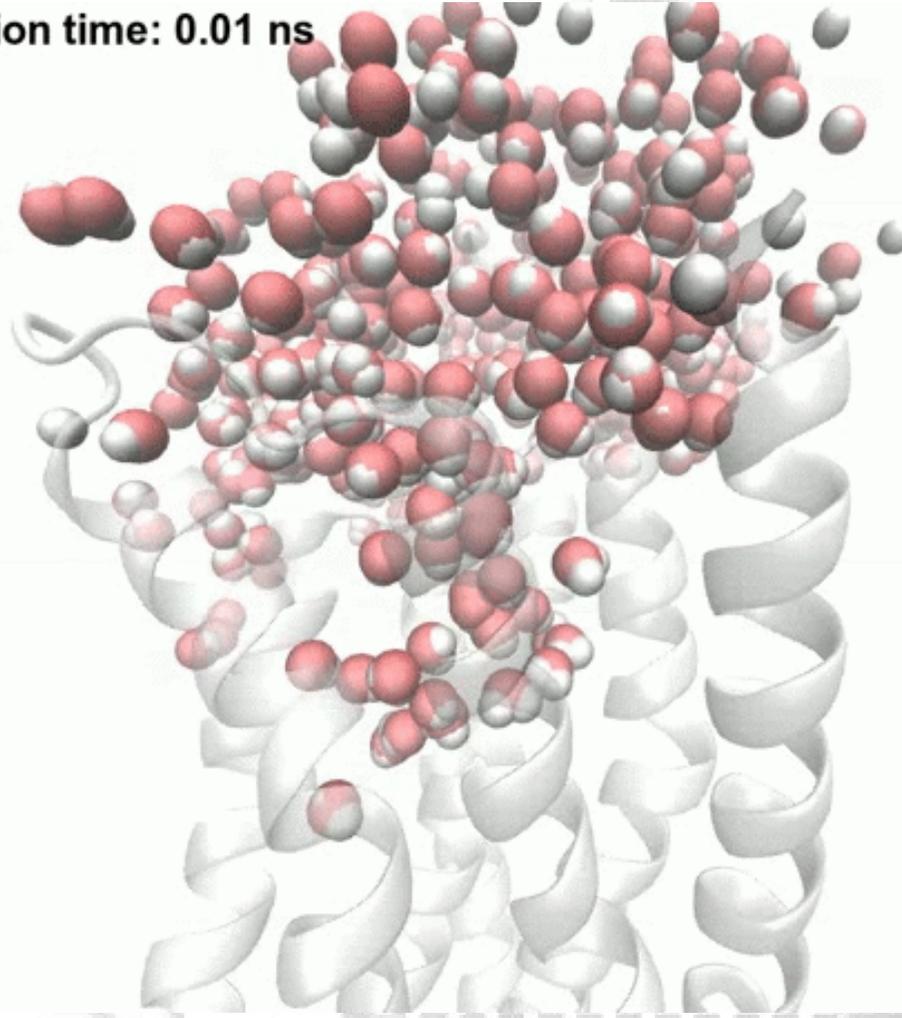
MMGBSA profile





... but this illusion is incredibly exciting!!!

SuMD simulation time: 0.01 ns

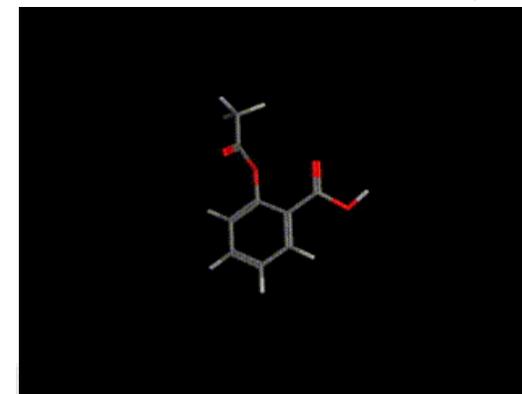
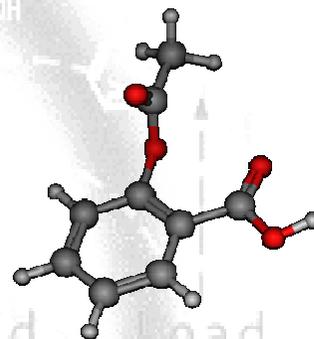
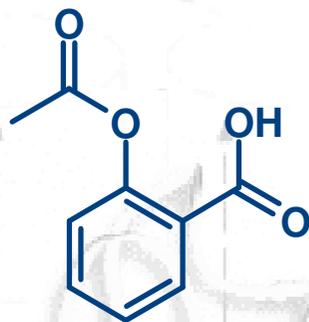


G. Deganutti A. Cuzzolin



Another important informatics difference!

C₉H₈O₄



6 byte



1.299 byte



2.051 byte



73.728 byte

1 byte = 8 bit (1 bit = 0 o 1, true o false)



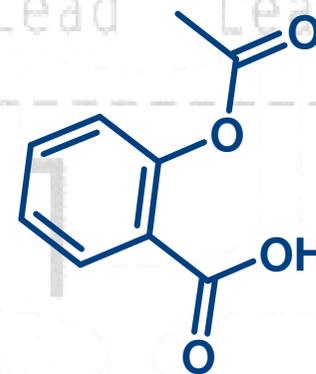
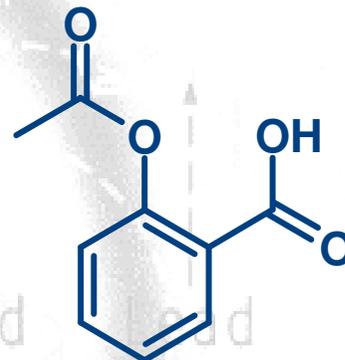


Do you remember?

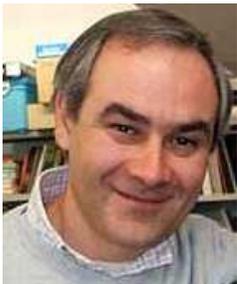
C₉H₈O₄

C₉H₈O₄

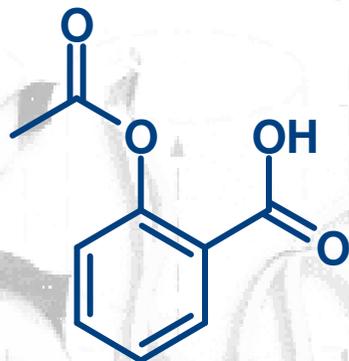
Time of answer (sec):



Time of answer (sec):



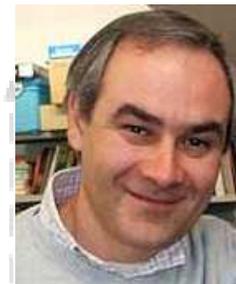
Think about...



C9H8O4



????



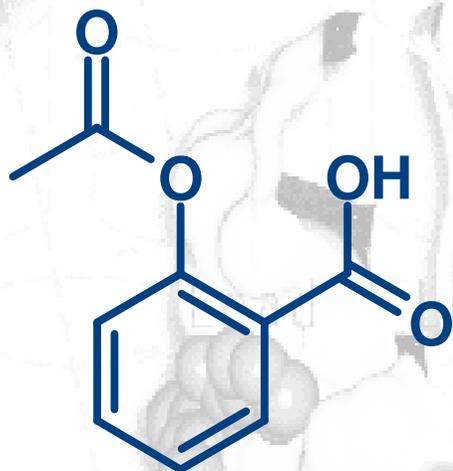
Stefano Moro



MROSFN65B05X407Y



Combining business with pleasure ?



String_of_characters

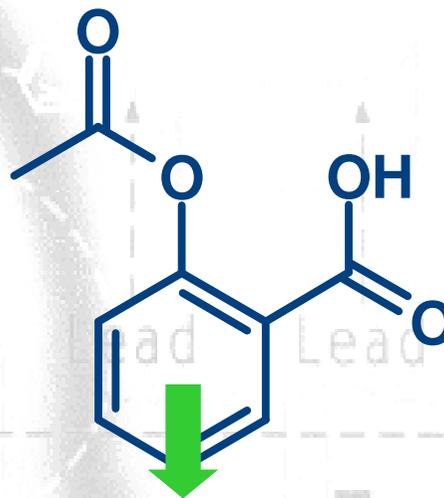


... possibly, using only a keyboard?



Combining business with pleasure ?

String_of_characters



SMILES (Simplified Molecular Input Line Entry Specification)

SMILES (Simplified Molecular Input Line Entry Specification)



The original SMILES specification was initiated by *David Weininger* at the USEPA Mid-Continent Ecology Division Laboratory in Duluth in the 1980s.

Anderson E, Veith GD, Weininger D (1987). SMILES: A line notation and computerized interpreter for chemical structures. Duluth, MN: U.S. EPA, Environmental Research Laboratory-Duluth. Report No. EPA/600/M-87/021.

Weininger D "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". Journal of Chemical Information and Modeling. 28 (1): 31-6.

Using simple rules it is possible to represent the “**connections**” between “**molecular fragments**” (as in the *structural formula*) in a simple “**string**” of “**alphanumeric characters**” (as in the *bruta formula*).

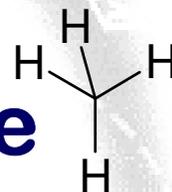
Here is some examples:

SMILES (Simplified Molecular Input Line Entry Specification)

Some SMILES rules:

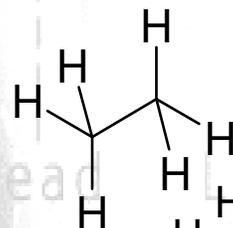
C

methane



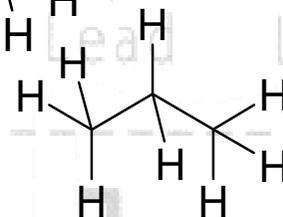
CC

ethane



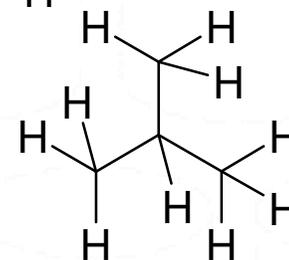
CCC

propane



CC(C)C

2-methyl-propane



C1CCCCC1

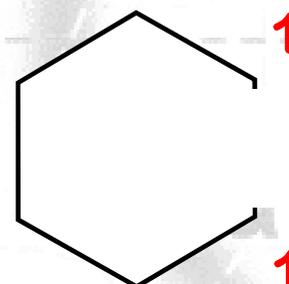
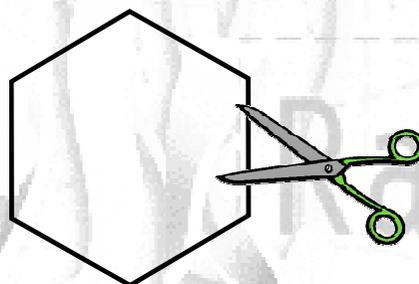
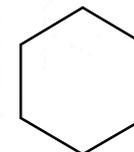
cycloesane



SMILES (Simplified Molecular Input Line Entry Specification)

C1CCCCC1

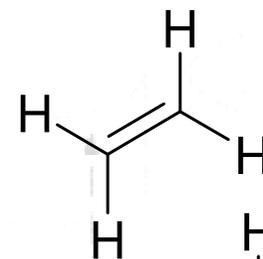
cycloesane



SMILES (Simplified Molecular Input Line Entry Specification)

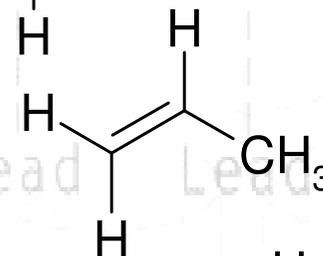
C=C

ethene (ethylene)



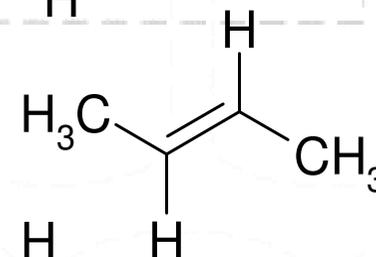
C=CC

propene



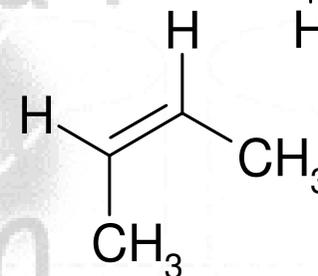
C/C=C/C

trans (E)-2-butane



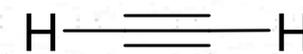
C/C=C\C

cis (Z)-2-butane



C#C

ethyne (acetylene)

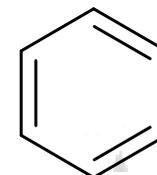


SMILES (Simplified Molecular Input Line Entry Specification)

Caps Lock

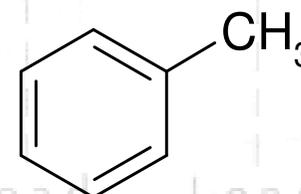
c1ccccc1

benzene



Cc1ccccc1

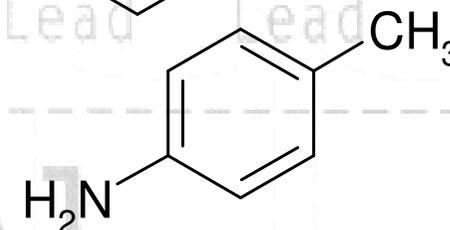
toluene



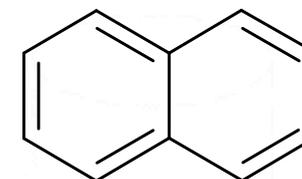
Cc1ccc(N)cc1

4-methyl-aniline

Nc1ccc(C)cc1

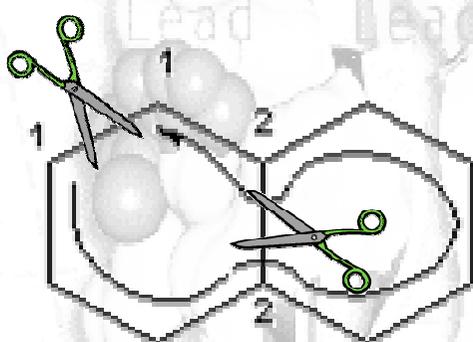
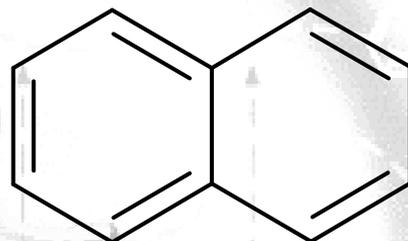


c12c(ccc1)cccc2 naphthalene

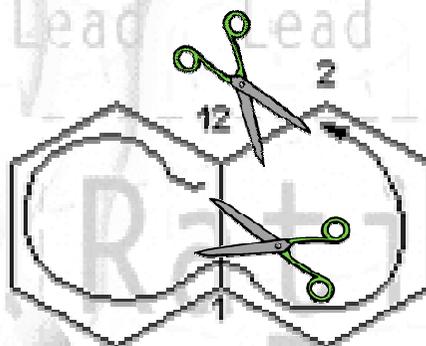


SMILES (Simplified Molecular Input Line Entry Specification)

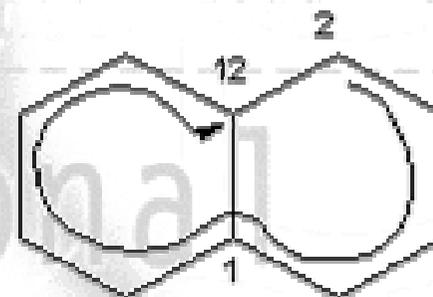
Examples for naphthalene:



c1ccc2ccccc2c1



c12ccccc1cccc2

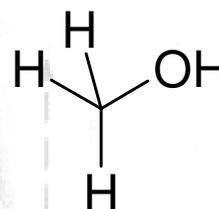


c2ccccc1ccccc12

SMILES (Simplified Molecular Input Line Entry Specification)

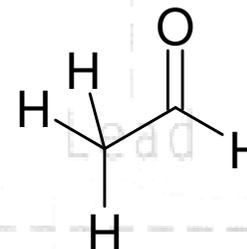
CO

methanol



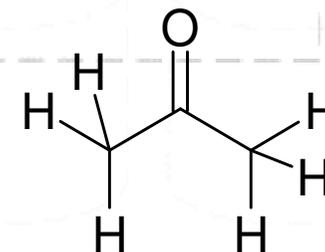
CC=O

ethanal



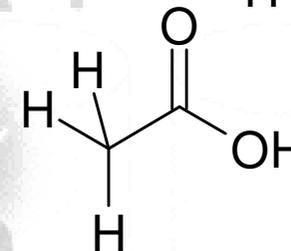
CC(=O)C

acetone



CC(=O)O

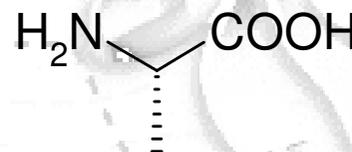
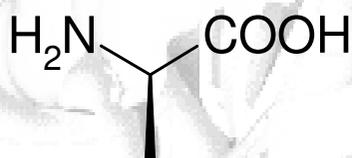
acetic acid



SMILES (Simplified Molecular Input Line Entry Specification)

In SMILES, tetrahedral centers may be indicated by a simplified chiral specification (@ or @@) written as an atomic property following the atomic symbol of the chiral atom.

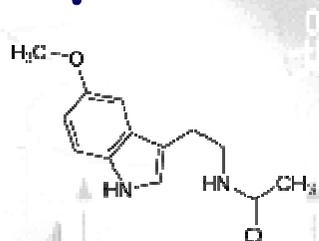
Looking at the chiral center from the direction of the "from" atom (as per atom order in SMILES), @ means "the other three atoms are listed *anti-clockwise*"; @@ means *clockwise*.



SMILES (Simplified Molecular Input Line Entry Specification)

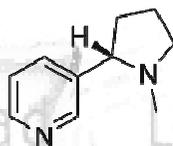
Some medchem examples:

Melatonin ($C_{13}H_{16}N_2O_2$)



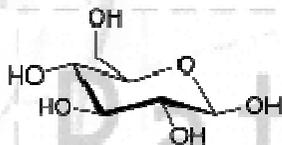
CC(=O)NCCc1c[nH]c2ccc(OC)cc12

Nicotine ($C_{10}H_{14}N_2$)



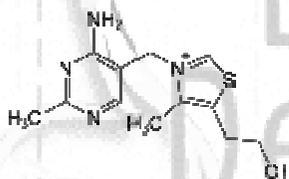
CN1CCC[C@H]1c2cccnc2

Glucose (glucopyranose) ($C_6H_{12}O_6$)



OC[C@H](O1)[C@H](O)[C@H](O)[C@@H](O)[C@@H](O)1

Thiamine ($C_{12}H_{17}N_4OS^+$) (vitamin B1)

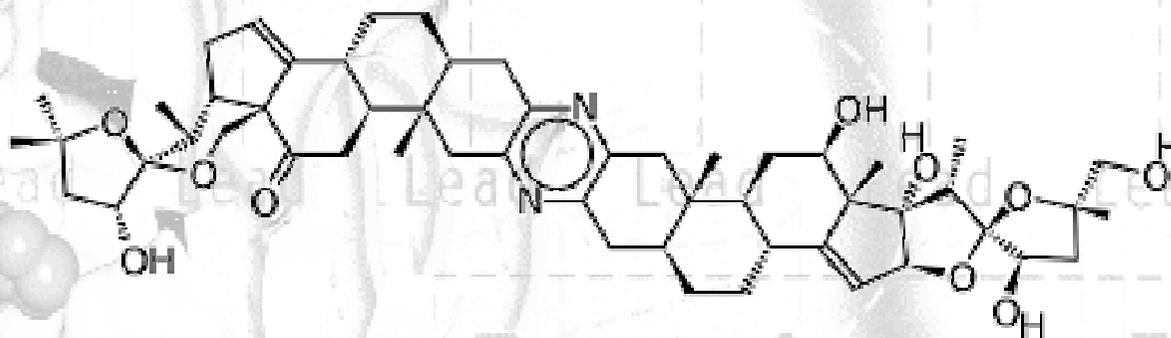


OCCc1c(C)[n+](cs1)Cc2cnc(C)nc2N

SMILES (Simplified Molecular Input Line Entry Specification)

Some examples:

Cephalostatin-1, a steroidal trisdecacyclic pyrazine with the empirical formula $C_{54}H_{74}N_2O_{10}$



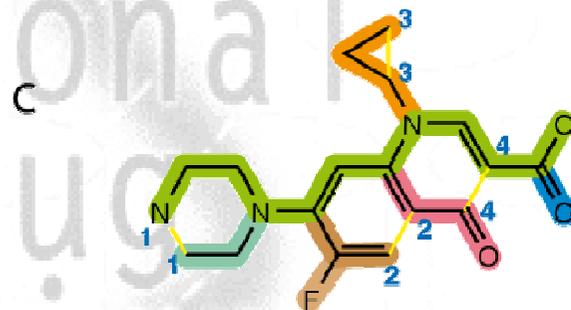
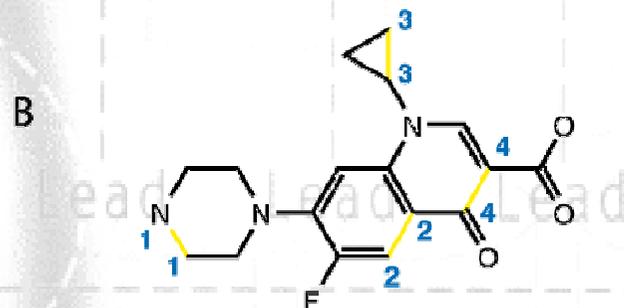
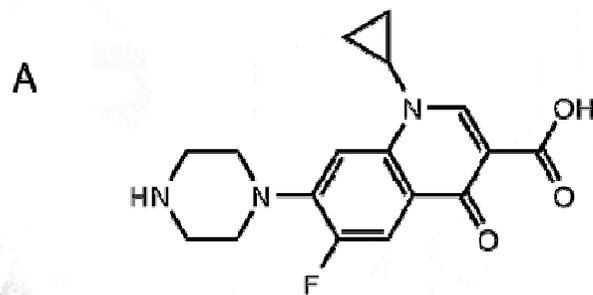
Starting with the left-most methyl group in the figure:

```
C[C@H]1[C@H]2CC=C3[C@]2(CO[C@]14[C@@H](CC(O4)(C)C)O)C(=O)C[C@H]5[C@H]3CC[C@@H]6[C@@]5(CC7=NC8=C(C[C@]9([C@H](C8)CC[C@@H]1[C@@H]9C[C@H]([C@]2(C1=C[C@H]1[C@@]2([C@@H]([C@@]2(O1)[C@@H](C[C@](O2)(C)CO)O)C)O)C)O)C)N=C7C6)C
```

SMILES (Simplified Molecular Input Line Entry Specification)

Generation of SMILES:

Break cycles, then write as branches off a main backbone. (Ciprofloxacin)

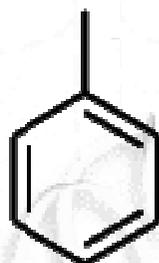


D

```
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O
```

SMILES (Simplified Molecular Input Line Entry Specification)

Toluene SMILES Enumeration



Cc1ccccc1
c1ccccc1C
c1(C)ccccc1
c1c(C)cccc1
c1cc(C)ccc1
c1ccc(C)cc1
c1cccc(C)c1

Canonical SMILES is a unique way of writing a SMILES for a molecule, where some rules about numbering defines the ordering of the atoms. This ensures that there is only one unique SMILES corresponding to one unique molecule. It is often useful to have this 1:1 correspondence:

- One chemical one SMILES string;
- Same SMILES string coming from different programs;
- Improving search process in chemical databases.

SMILES (Simplified Molecular Input Line Entry Specification)

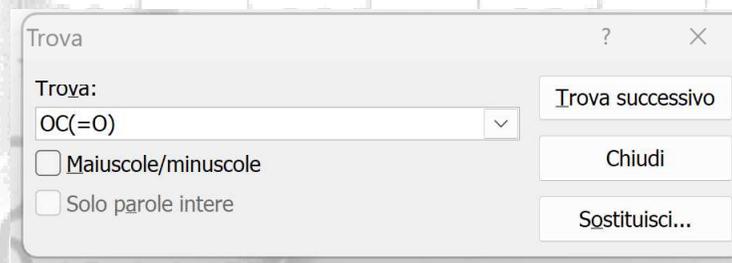
NERDS ONLY

CANONALISING SMILES: please check at the end for this file... and enjoy the Morgan's algorithm!

SMILES (Simplified Molecular Input Line Entry Specification)

A powerful “searching” strategy:

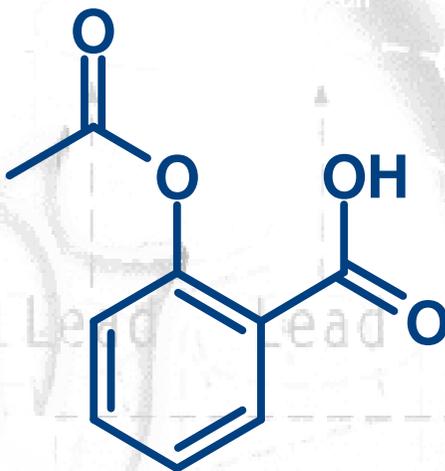
O(O=)Cc1ccccc1OC(=O)C



O(O=)Cc1ccccc1OC(=O)C



reassuring:



O(O=)Cc1ccccc1OC(=O)C

Two faces of the same medal!!!



Do you need a SMILES's translator?

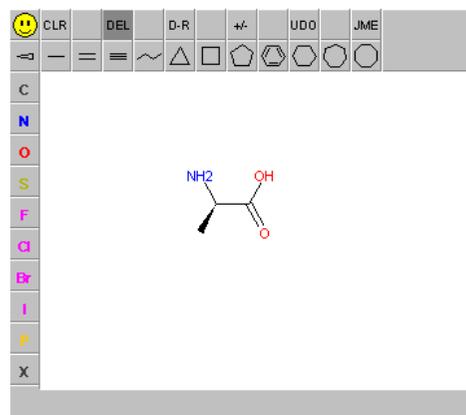
<http://mms.dsfarm.unipd.it/MMsINC/search/>



UNIVERSITA' DEGLI STUDI DI PADOVA
Molecular Modeling Section

MMsINC Search: Structure Search Similarity to PDB ligands

Structure Search



Create SMILES

Clear Editor

Search

Reset

Query Type:

Search using: Identical Structure Search

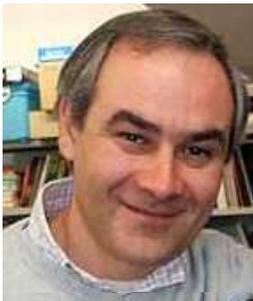
Query Data:

Input: SMILES MMscore InChI Molecular Formula

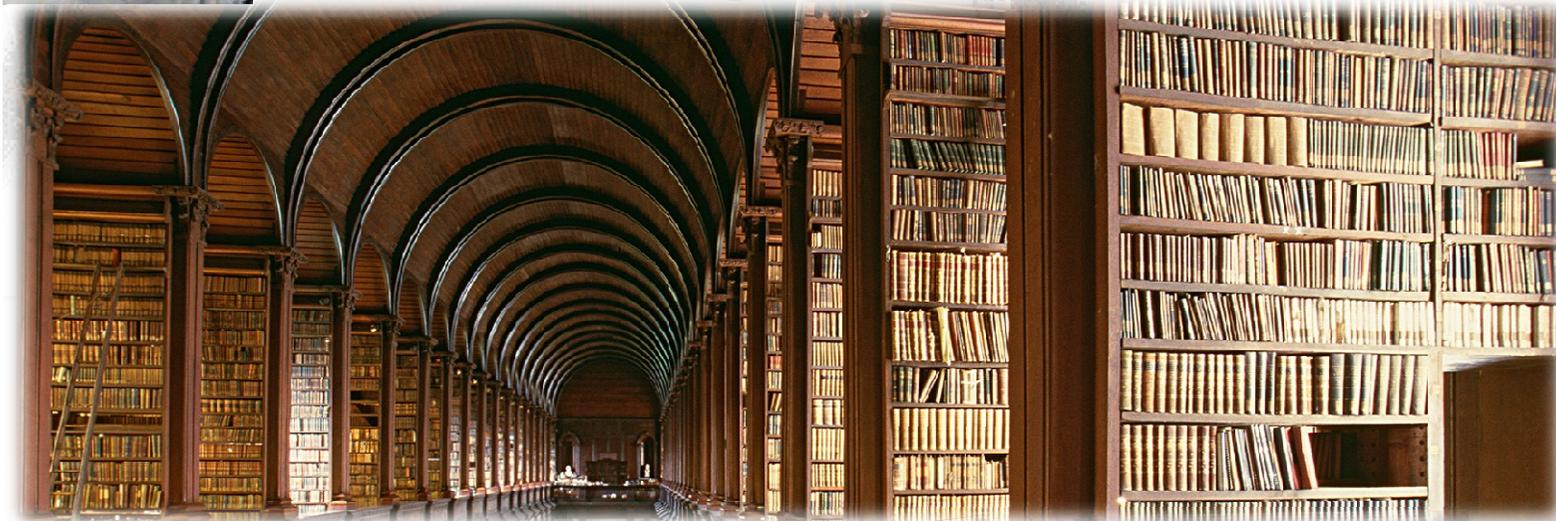
CC([NH2])C(=O)O

Search

Reset



Chemical archives...



Lead



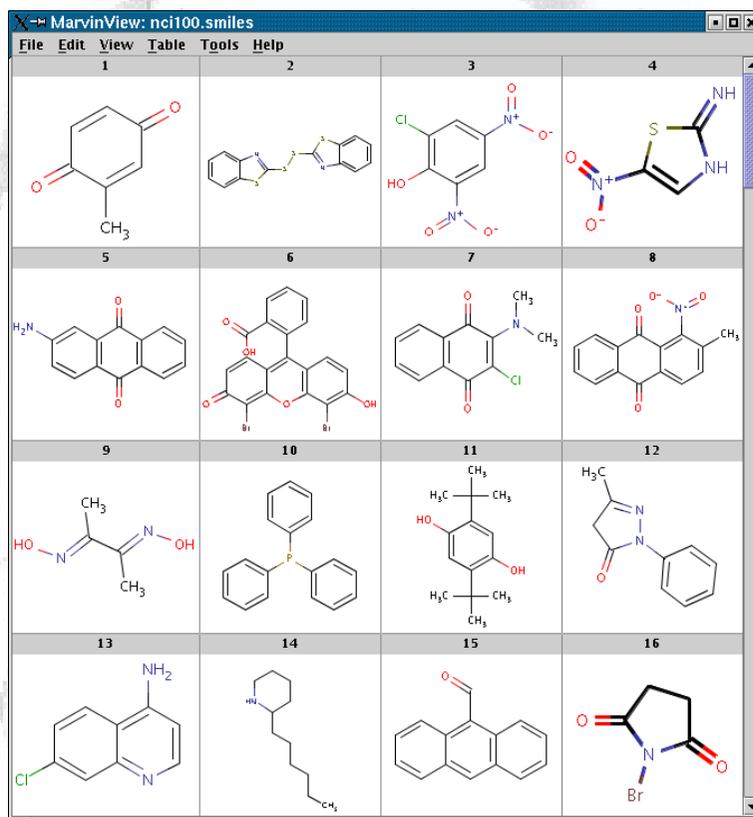
MS

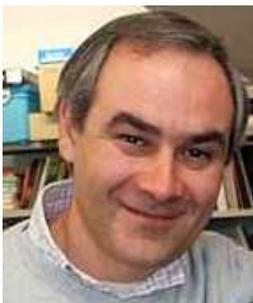
Confidential and Property of ©2005 Molecular Modeling Section
Dept. Pharmaceutical and Pharmacological Sciences – University of Padova - Italy

S.MORO – PSF: LBDD_1



The second major aid that informatics gives to the medicinal chemistry is the *'intelligent storage'* of molecular structures.





archiviatio... the first crucial virtualization process.

MODULARIO a.r.t. n. 531 Mod. AT

REPUBBLICA ITALIANA e della
RICERCA SCIENTIFICA e TECNOLOGICA
UNIVERSITÀ DEGLI STUDI DI PADOVA
timbro lineare dell'Amministrazione rilasciante

TESSERA N. 7527363

MORO Stefano
cognome e nome

Ricercatore
qualifica

Foto
Firma del Titolare

Nato a Treviso
il 05.02.1965
Residenza Villafranca (PD)
Via Molini 4A
Stato civile coniugato

La presente tessera vale cinque anni dalla data di rilascio o di convalida.

CONNOTATI E CONTRASSEGNI SALIENTI

Statura cm. 180
Capelli neri
Occhi neri
Segni particolari

Padova li 25.01.1999

Timbro a umido IL FUNZIONARIO RESPONSABILE Baulis

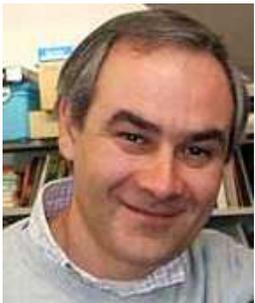
CONVALIDA

La presente tessera è convalidata fino al

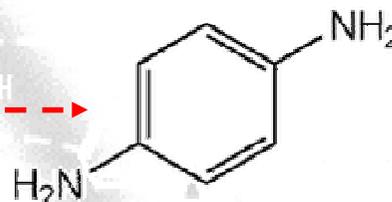
li

Timbro a umido IL FUNZIONARIO RESPONSABILE

Why: to ensure the uniqueness of representation in the virtual world of an object in the real world!



and it is clear, too!



MODULARIO a.r.f. n. 531 Mod. AT

REPUBBLICA ITALIANA e dell'UNIVERSITÀ DEGLI STUDI DI PADOVA
RICERCA SCIENTIFICA e TECNOLOGICA
timbro lineare dell'Amministrazione Rilasciante

TESSERA N. 7527363

MORO Stefano
cognome e nome

Ricercatore
qualifica

Foto: [Portrait of Stefano Moro]

Firma del Titolare: [Signature]

Nato a Treviso
il 05.02.1965

Residenza Villafranca (PD)
Via Molini 4A

Stato civile coniugato

La presente tessera vale cinque anni dalla data di rilascio o di convalida.

CONNOTAZIONE E CONTRASSEGNI SALIENTI

Statura cm. 180

Capelli neri

Occhi neri

Segni particolari

Padova li 25.01.1999

IL FUNZIONARIO RESPONSABILE
[Signature]

CONVALIDA

La presente tessera è convalidata fino al

IL FUNZIONARIO RESPONSABILE

p-Phenylenediamine

CAS 106-50-3

PM 108.14



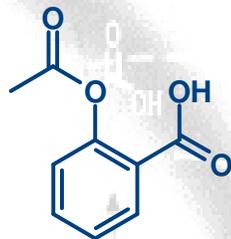
Let's start building our molecular database?

	F.BRUTA	mol	SMILES
1	C7H6O2	1.0986 -13.6500 0.0000 C 1.0974 -14.4773 0.0000 C 1.8122 -14.8902 0.0000 C 2.5287 -14.4769 0.0000 C 2.5258 -13.6463 0.0000 C 1.8104 -13.2372 0.0000 C 1.8080 -12.4122 0.0000 O 1.0923 -12.0018 0.0000 C 1.0898 -11.1768 0.0000 O 0.3790 -12.4165 0.0000 C 3.2387 -13.2311 0.0000 C 3.9547 -13.6409 0.0000 O 3.2356 -12.4061 0.0000 O 1 2 2 6 7 1 3 4 2	<chem>OC(=O)c1ccccc1</chem>
2	C7H6O3	1.0986 -13.6500 0.0000 C 1.0974 -14.4773 0.0000 C 1.8122 -14.8902 0.0000 C 2.5287 -14.4769 0.0000 C 2.5258 -13.6463 0.0000 C 1.8104 -13.2372 0.0000 C 1.8080 -12.4122 0.0000 O 1.0923 -12.0018 0.0000 C 1.0898 -11.1768 0.0000 O 0.3790 -12.4165 0.0000 C 3.2387 -13.2311 0.0000 C 3.9547 -13.6409 0.0000 O 3.2356 -12.4061 0.0000 O 1 2 2 6 7 1 3 4 2	<chem>OC(=O)c1ccccc1(O)</chem>
3	C9H8O4	1.0986 -13.6500 0.0000 C 1.0974 -14.4773 0.0000 C 1.8122 -14.8902 0.0000 C 2.5287 -14.4769 0.0000 C 2.5258 -13.6463 0.0000 C 1.8104 -13.2372 0.0000 C 1.8080 -12.4122 0.0000 O 1.0923 -12.0018 0.0000 C 1.0898 -11.1768 0.0000 O 0.3790 -12.4165 0.0000 C 3.2387 -13.2311 0.0000 C 3.9547 -13.6409 0.0000 O 3.2356 -12.4061 0.0000 O 1 2 2 6 7 1 3 4 2	<chem>OC(=O)c1ccccc1(OC(=O)C)</chem>

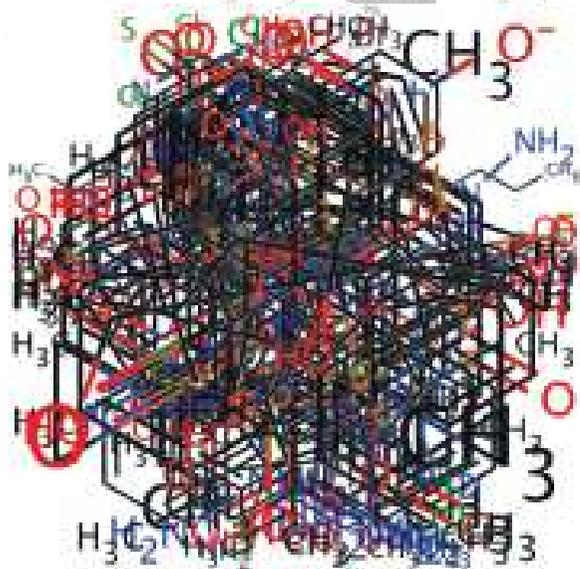


Searching in a database is now very easy!

Query



O(=O)Cc1ccccc1OC(=O)C



	F.BRUTA	mol	SMILES	
1	C7H6O2		<chem>OC(=O)c1ccccc1</chem>	
2	C7H6O3		<chem>OC(=O)c1ccccc1(O)</chem>	
3	C9H8O4		<chem>OC(=O)c1ccccc1OC(=O)C</chem>	





A bit of... sequence alignment.

1. Identity search:

Sequence A: C_3H_6O



Sequence B: C_3H_6O

a. If the number of characters the same?

NO

YES

not

identical

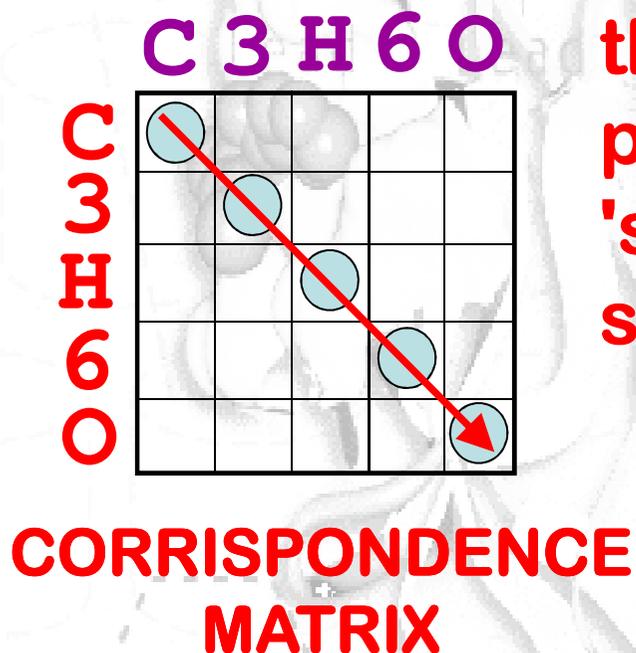
Go on!



A bit of... sequence alignment.

b. If the number of characters is the same:

The *main diagonal* represents in the correspondence matrix the place where we find the highest 'similarity' between the two sequences.



C3H6O

|||
|||

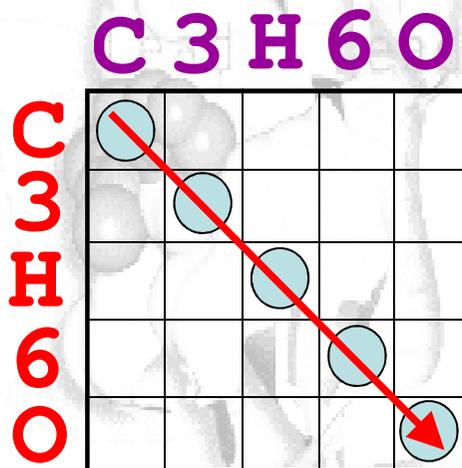
C3H6O

Identity or
100% similarity

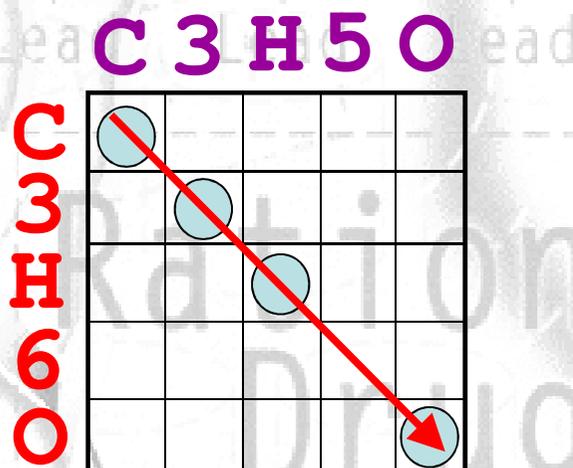


Can we start thinking about the relation between **IDENTITY** and **SIMILARITY**...
even if this is a very bad example!

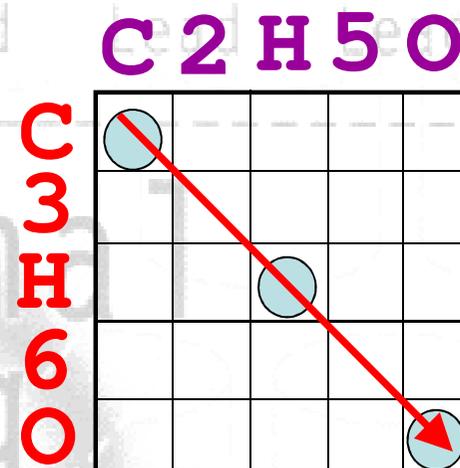
IDENTITY is equal to 100% of **SIMILARITY**



100%



80%

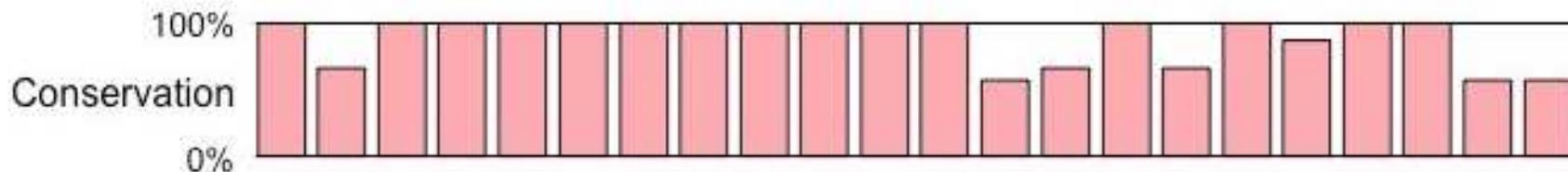


60%



Can you also find another important application of this “string similarity approach?”

mouse	F	S	T	A	A	F	R	F	G	H	A	T	V	H	P	L	V	R	R	L	N	T
rat	F	S	T	A	A	F	R	F	G	H	A	T	V	H	P	L	V	R	R	L	N	T
human	F	S	T	A	A	F	R	F	G	H	A	T	I	H	P	L	V	R	R	L	D	A
pig	F	S	T	A	A	F	R	F	G	H	A	T	I	H	P	L	V	R	R	L	D	A
dog	F	S	T	A	A	F	R	F	G	H	A	T	V	H	P	L	V	R	R	L	D	A
chicken	F	A	T	A	A	F	R	F	G	H	A	T	I	Q	P	I	V	R	R	L	N	A
frog	F	T	T	A	A	F	R	F	G	H	A	T	I	P	P	M	V	H	R	L	D	S
Consensus	F	S	T	A	A	F	R	F	G	H	A	T	I	H	P	L	V	R	R	L	D	A





... important molecular Haystack!



Chemical Abstract Service

(www.cas.org)

Date 21/11/2022 10:37:40 EST

Count 183,000,000 available organic and inorganic chemicals

71 million commercially available substances

136 million reactions



... other important molecular Haystacks!

PubChem

<http://pubchem.ncbi.nlm.nih.gov/>

eMolecules

<http://www.emolecules.com/>

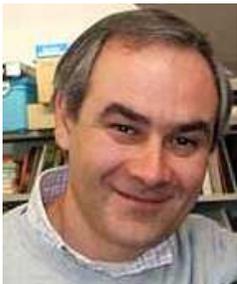


ChemSpider
Building community for chemists

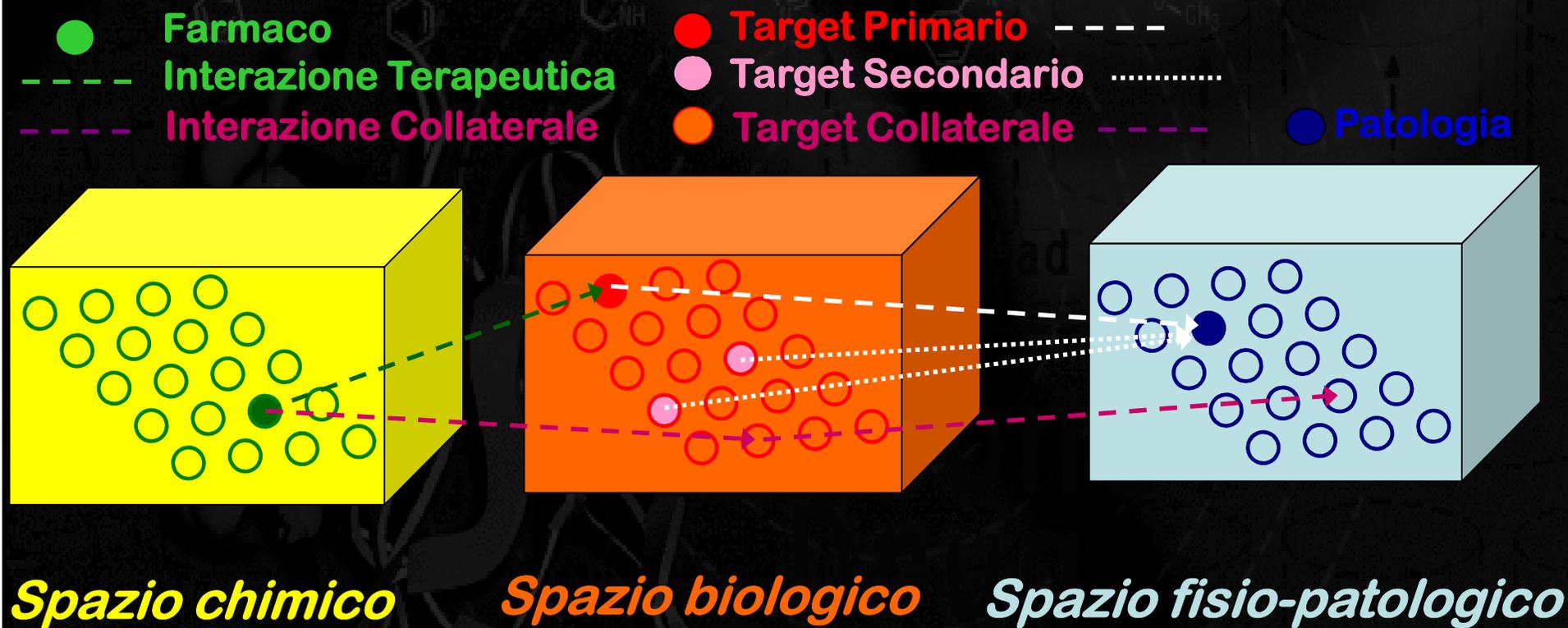
<http://www.chemspider.com/>

MMsINC

<http://mms.dsfarm.unipd.it/MMsINC.html>



How large is our chemical space?

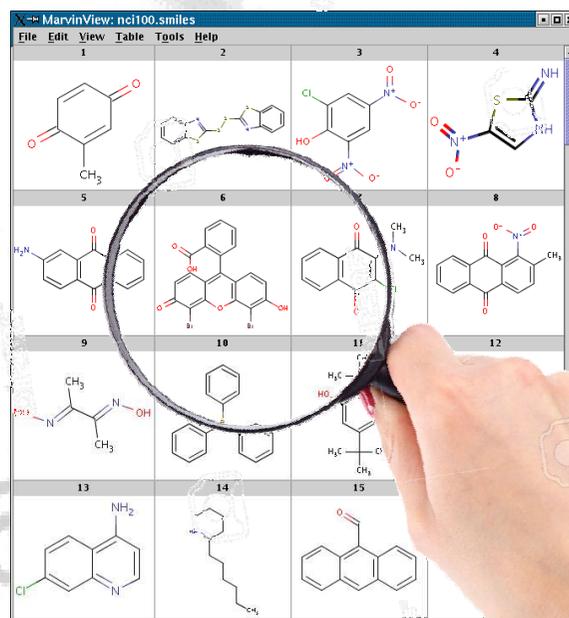
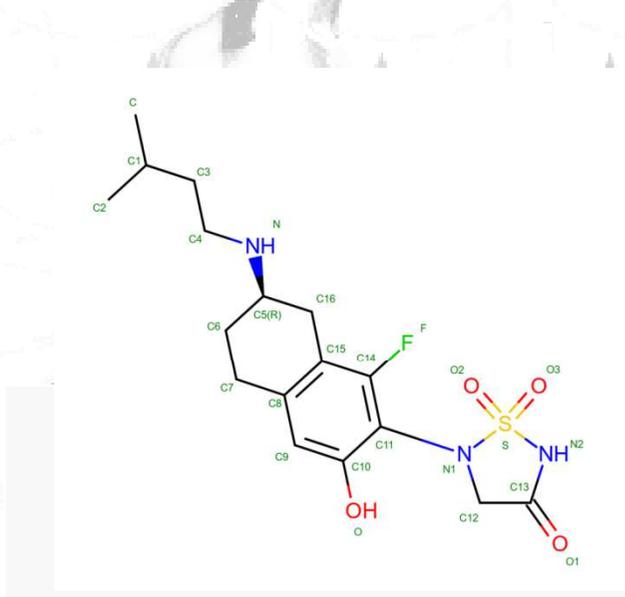


The number of chemical compounds consisting of C, H, N, O, P, S, F, Cl, Br, I and with PM <500 is estimated in the order of:

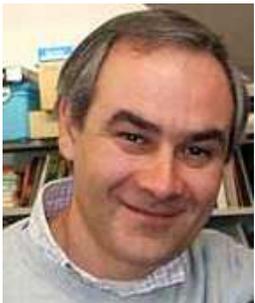
$10^{40} - 10^{120}$



A very common situation: we have only one compound that we know to be '*active*' against our therapeutic target:

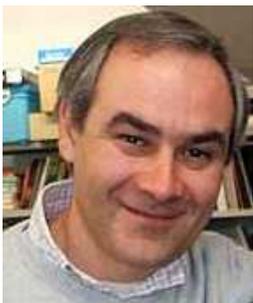


Could you suggest a strategy to find a novel possible candidate?

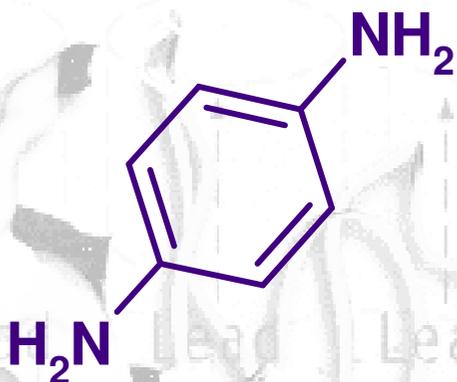


Similarity: structure *versus* properties...

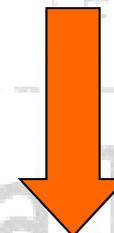




Similarity: structure *versus* properties...



Structure



Properties

PM = 108,14

pKa = 6.2

Volume = 93,9

MP = 142

PSA = 52

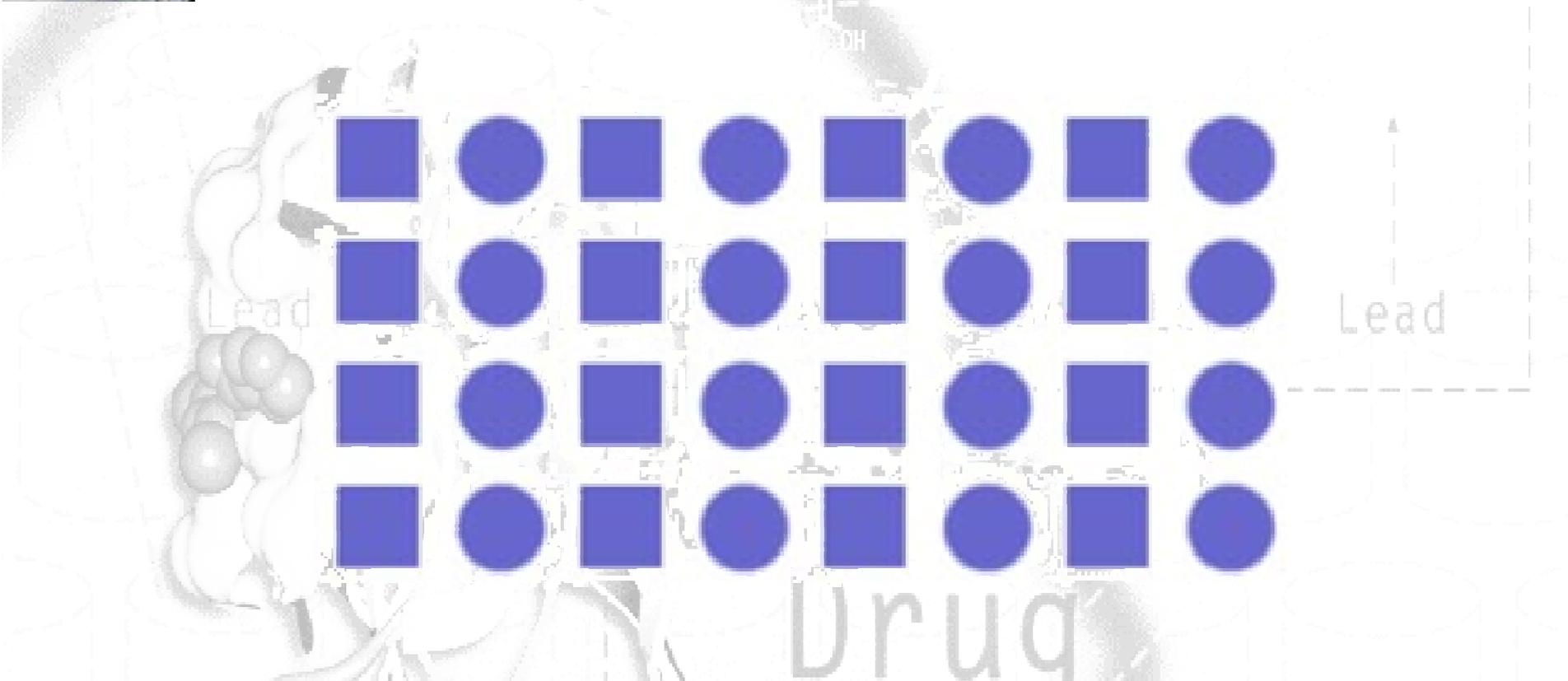
nC = 6

logP = -0.3

...



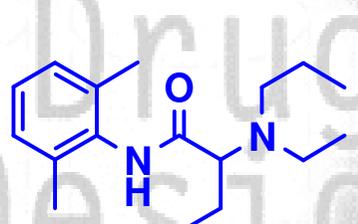
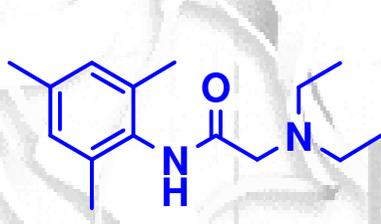
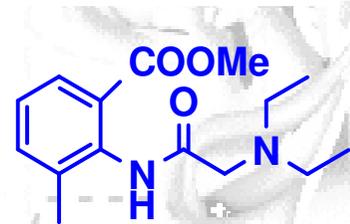
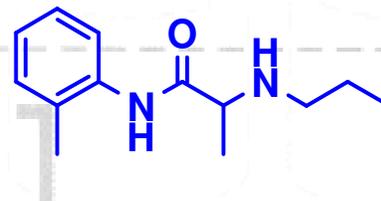
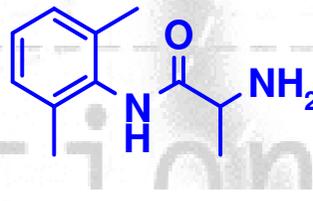
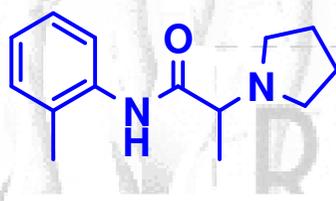
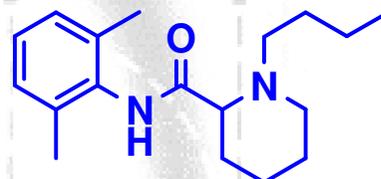
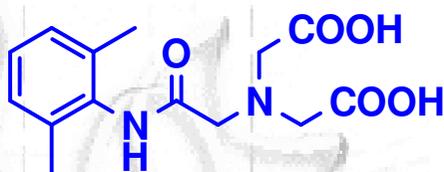
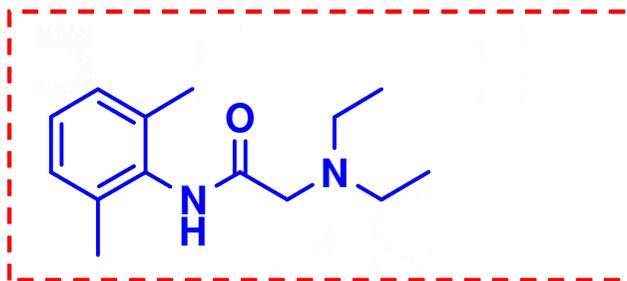
Do together this simple exercise: can you describe what you are looking ?

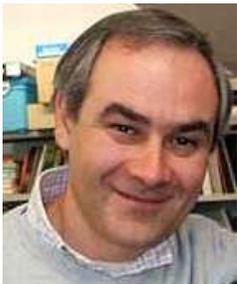


Similarity law: *similar items tend to be grouped together.*



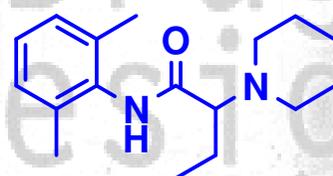
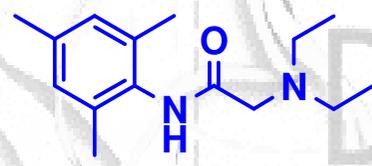
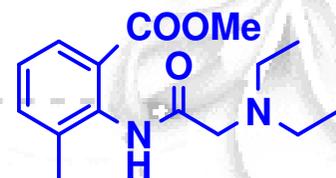
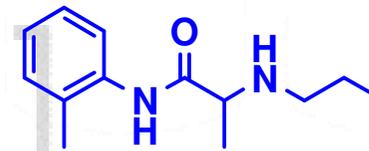
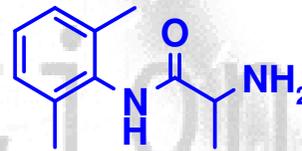
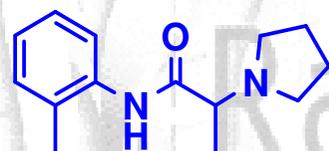
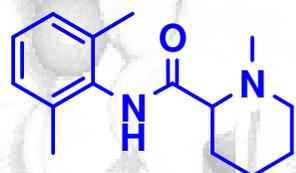
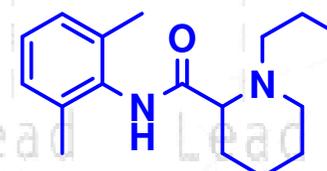
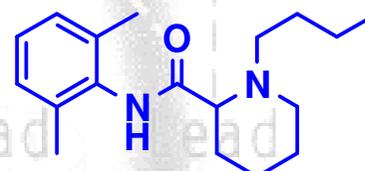
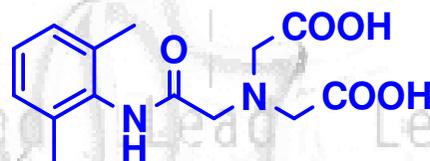
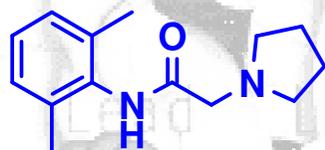
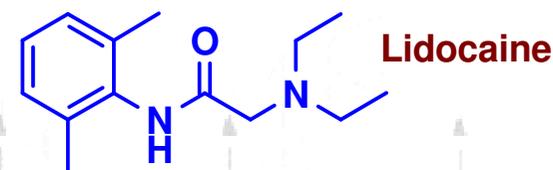
and, sometimes, this is true also in medicinal chemistry:

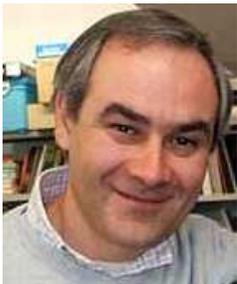




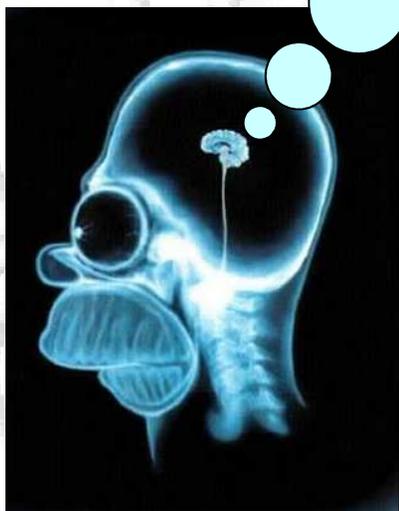
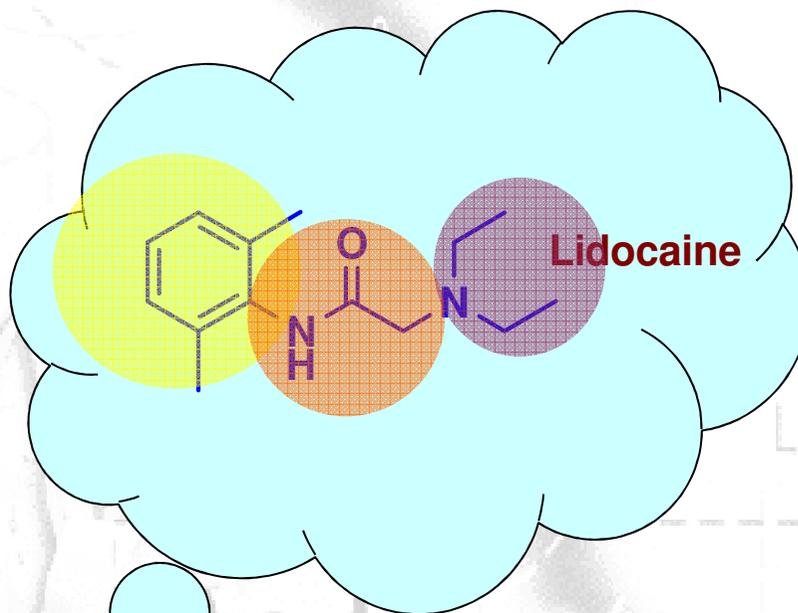
If we want to virtualize *chemical similarity* searching, we have to describe exactly what similarity is!

Are you able to scale *chemical similarity* of these analogs respect lidocaine?



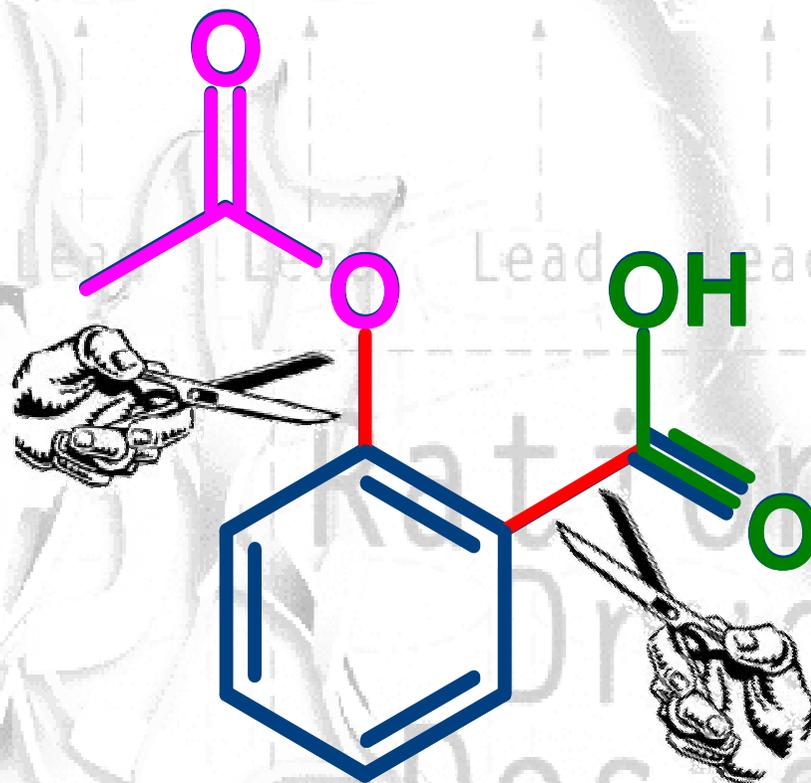


Tell me what you are thinking:



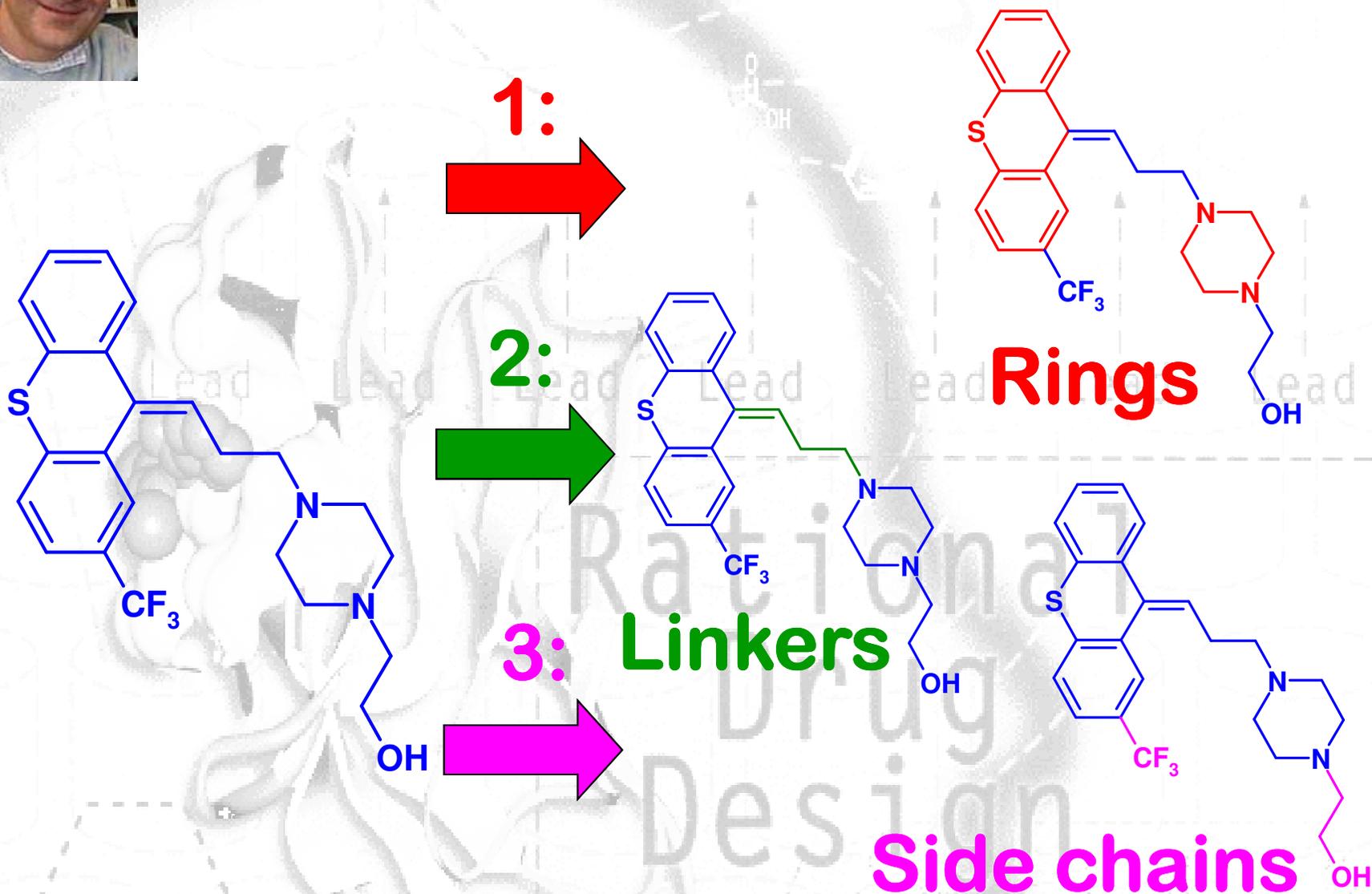


Exactly “*molecular fragmentation*”
could be a possible solution!

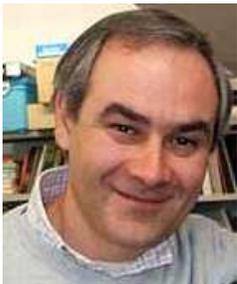




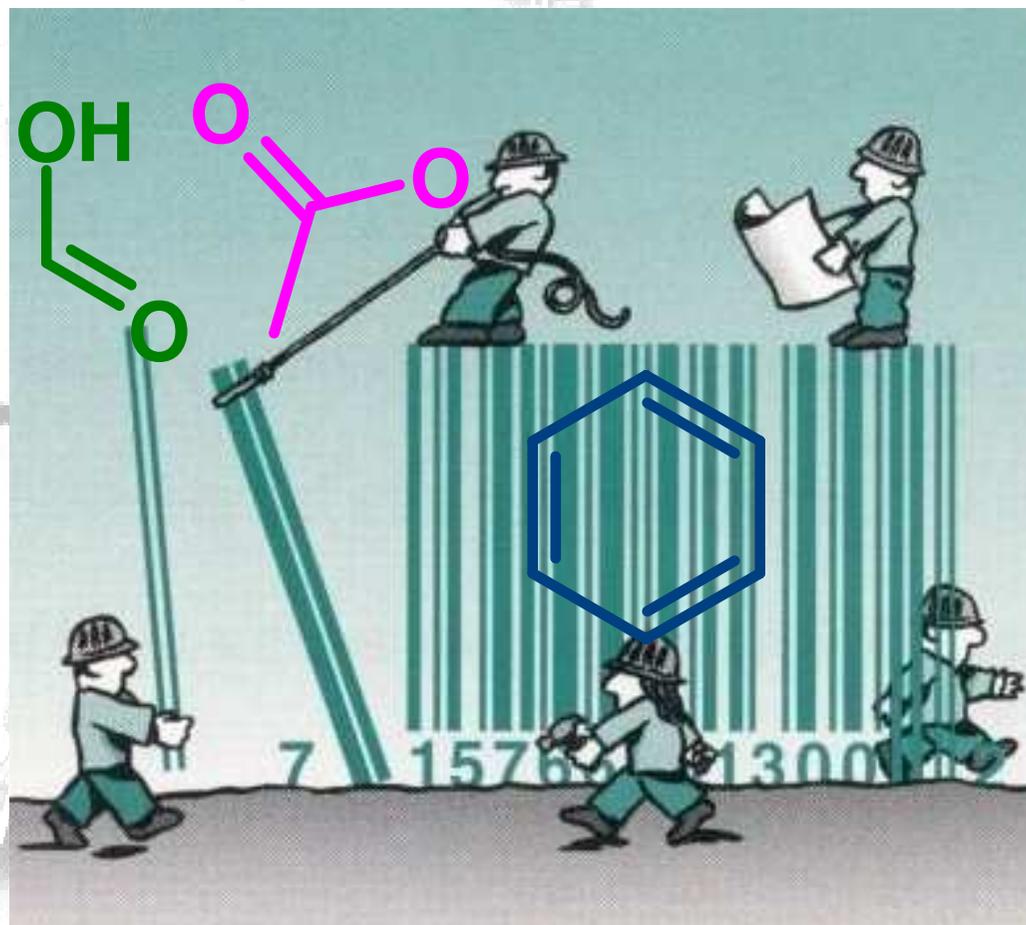
We need some rules:



G.W. Bemis, M. A. Murcko, *J. Med. Chem.* 1996, 39, 2887–2893.



The sequence of fragments is like a molecular bar code...



How we can build it up...



Any standard?

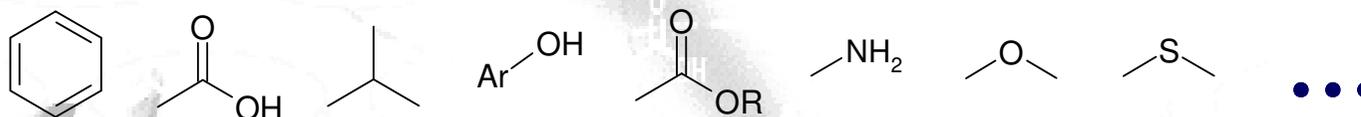
MACCS (Molecular ACcess System) a collection of 166 functional groups representing the most accessible medchem chemical space.

MACCS keys (1979), MDL Information Systems Inc. (originally named Molecular Design Limited, Inc.), San Leandro, CA.

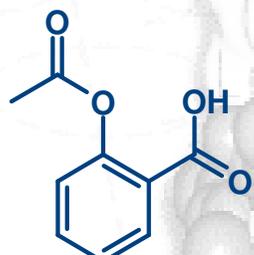
There is no publication describing the MACCS 166 public keys. All of the citations for it either say a variation of "*MDL did it*"

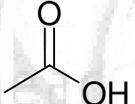
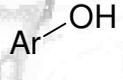
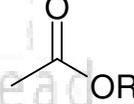
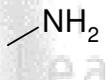
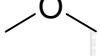
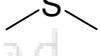
A useful concept: *structural keys*

1. Define all possible chemical fragments (*structural keys*):

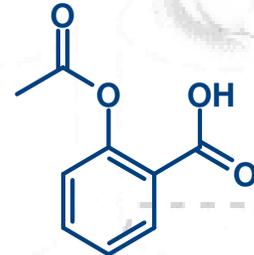


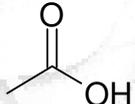
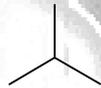
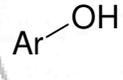
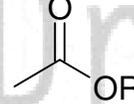
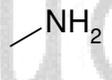
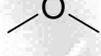
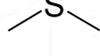
2. Assigne un *bit* for the each corrispondence in our molecule:

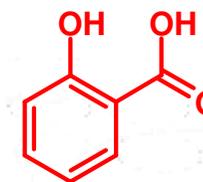


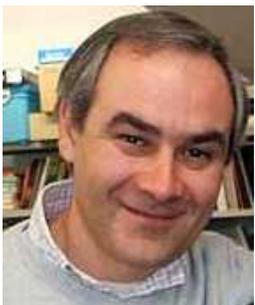
							
1	1	0	0	1	0	0	0

3. Reiterate this procedure for each molecule of the database :

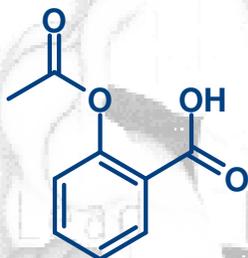


							
1	1	0	0	1	0	0	0
1	1	0	1	0	0	0	0





what about this parallelism?!



0000000100000001000000100010000000000010000010000100001000001000

Lead

Lead

Lead

Lead

Lead

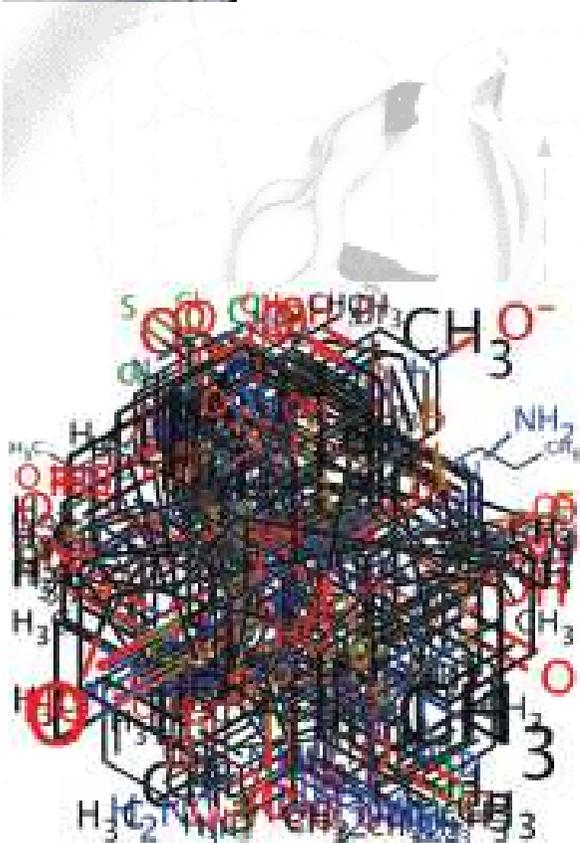
Lead



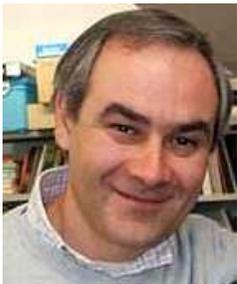
Rational
Drug
Design



... so, now it is very easy to transform my molecular database in a collection of “*structural keys*”!



```
0000001000011010000001010100000000011000001000010000100001000
0100010110010010010110011010011100111101000000110000000110001000
0100010100011101010000110000101000010011000010100000000100100000
000110111001110111110100000100010000110110110000000100110100000
0100010100110100010000000010000000010010000000100100001000101000
0100011100011101000100001011101100110110010010001101001100001000
0101110100110101010111111000010000011111100010000100001000101000
0100010100111101010000100010000000010010000010100100001000101000
000100010001010001010010000000000000001010000010000100000100000000
010001010001001100000000000000000000000101000000100000000000000000
01000101000101000000000000000000000101000010010000000001000000000000
01010101011111001111101000000000000011010100011100100001100101000
01000101000111000010000011000000000010001000000110000000001100000
000000010000000001000010000000000000001010100000000100000100100000
010001010001010000000010000000000000000000000000000000000000000000
0001000100001100010010100000010100101011100010000100001000101000
0100011100010100010000100001001110010010000010001100000000101000
010101010001010001010010000000000000000000000000000000000000000000
```



... well, but why this transformation could help us in measuring chemical similarity?

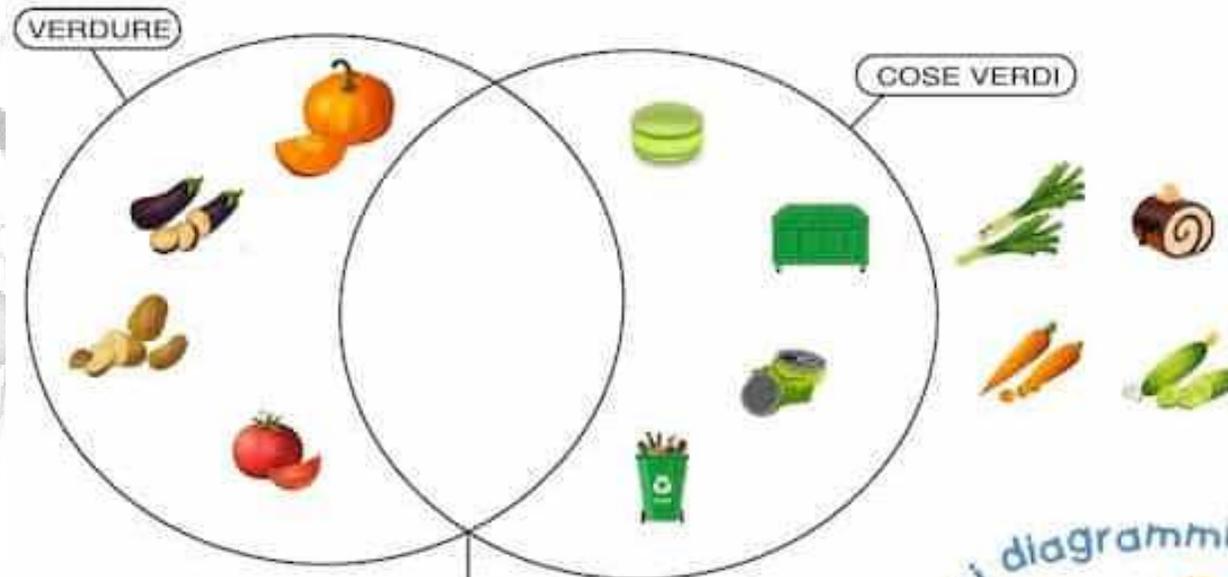
000000100001101000000101010000000000110000010000100001000001000
0100010110010010010110011010011100111101000000110000000110001000
0100010100011101010000110000101000010011000010100000000100100000
000110111001110111110100000100010000110110110000000100110100000
0100010100110100010000000010000000010010000000100100001000101000
0100011100011101000000001011100110110010010001101001100001000
0101110100110101010111111000010000011111100010000100001000101000
01000101001111010100001000100000000100100000101000100010000101000
00010001000101000101000000000000000000010100000100001000001000000000
010001010001001100000000000000000000101000000100001000000000000000
01000101000101000000000000001010000100100000000010000100001000000000
0101010101111100111110100000000000011010100011100100001100101000
0100010100011000010000011000000000010001000000110000000001100000
00000001000000000100001000000000000010100000000100000100100000
0100010100010100000000100000000000010000000000000100001000011000
0001000100001100010010100000010100101011100010000100001000101000
0100011100010100010000100001001110010010000010001100000000101000
0101010100010100010100100000000000010010000010010100100100010000





Let's go back to elementary school for a moment...

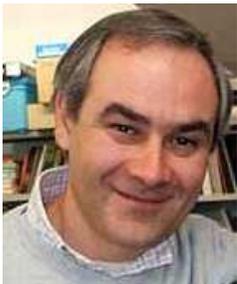
INTERSEZIONE DEGLI INSIEMI



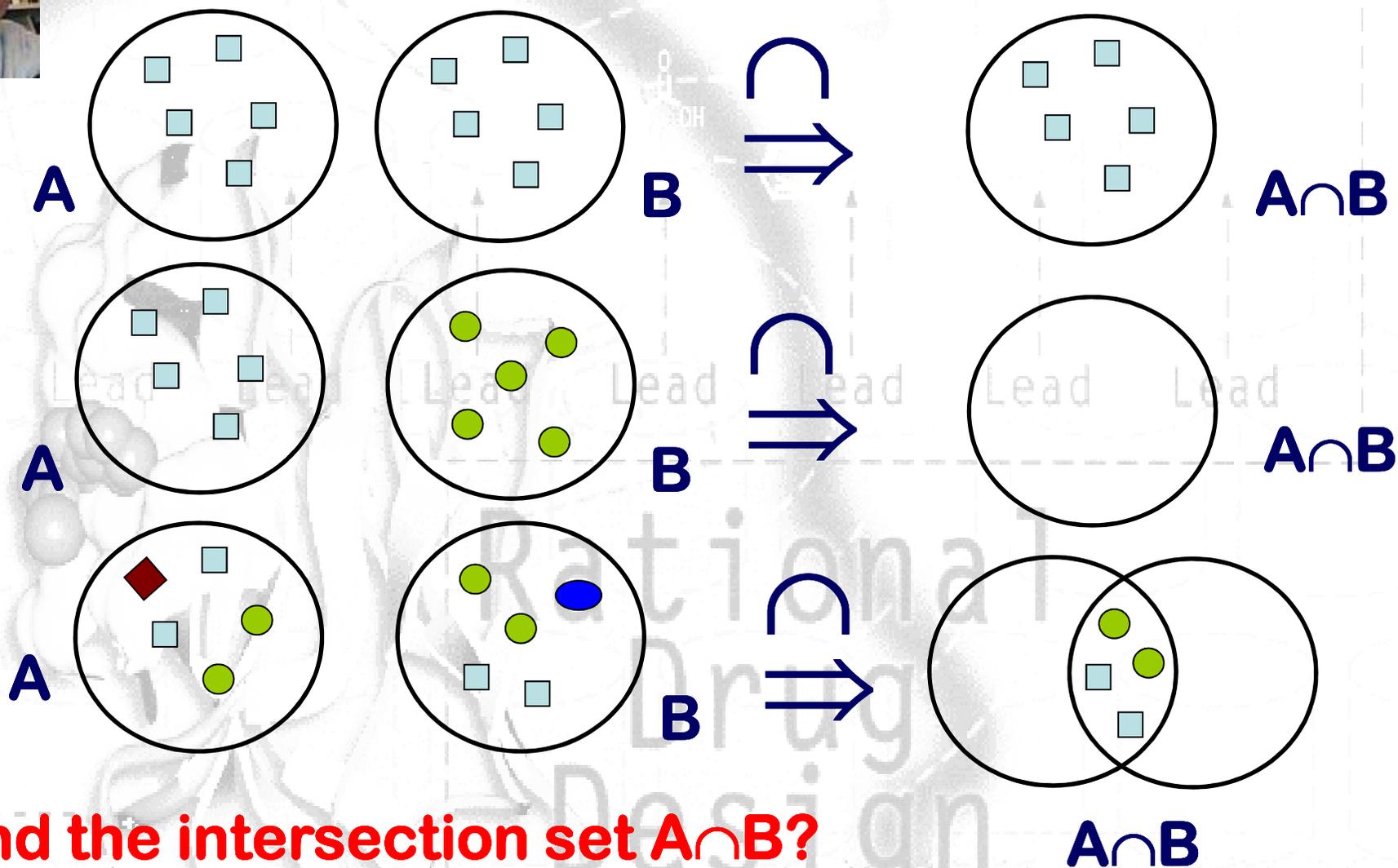
Osserva i diagrammi e...

1. Scegli tra quelli a lato, gli elementi da inserire nell'intersezione e completa il cartellino.

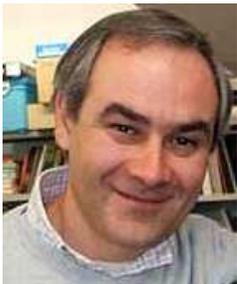




Do you remember the Eulero-Venn diagrams?

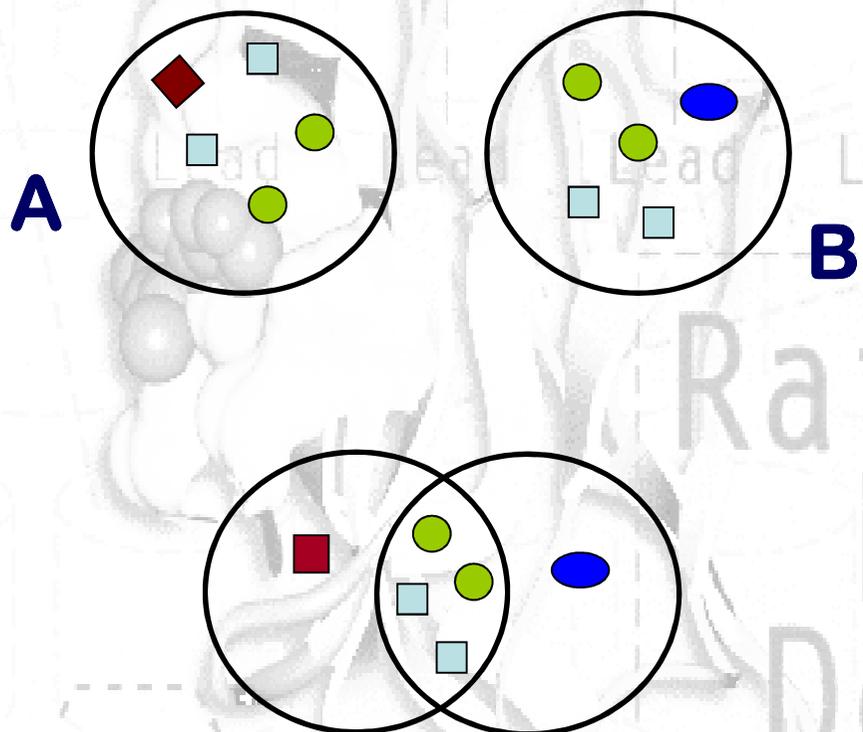


... and the intersection set $A \cap B$?

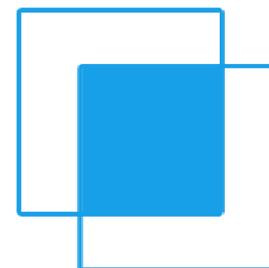


Now, I can introduce you the Jaccard index!

$$\text{Jaccard Index} = \frac{A \cap B}{A_{solo} + B_{solo} + A \cap B}$$



$A \cap B$



Area of Overlap

Area of Union

$A \cup B$



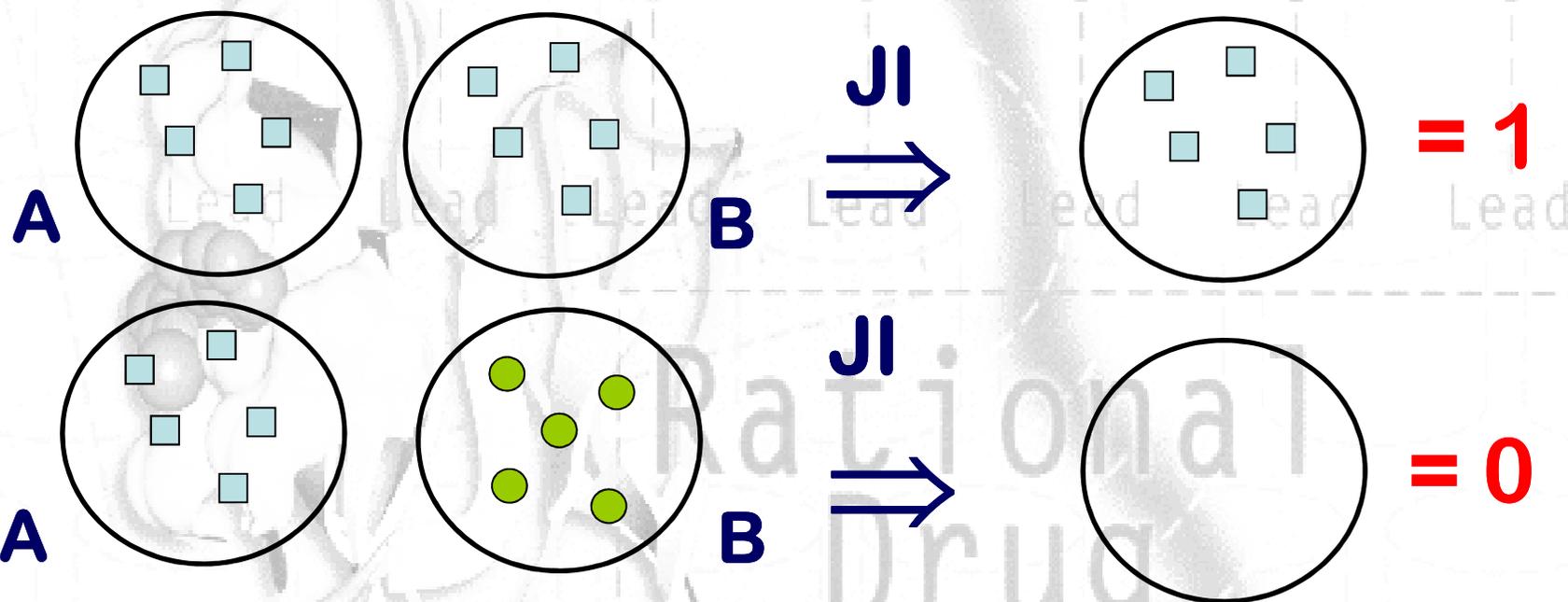
$$= \frac{4}{1 + 1 + 4} = 0.66$$

Originally coined *coefficient de communauté* by Paul Jaccard,
Jaccard, P. (1901) *Bulletin del la Société Vaudoisedes Sciences Naturelles* 37, 241-272.



Now, I can introduce you the Jaccard index!

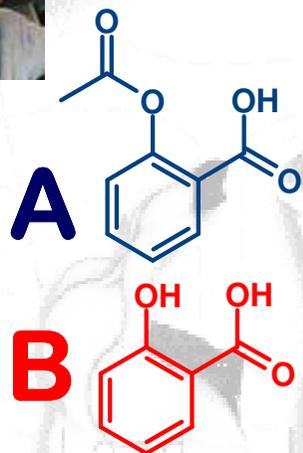
$$\text{Jaccard Index} = \frac{A \cap B}{A_{\text{solo}} + B_{\text{solo}} + A \cap B}$$

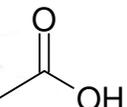
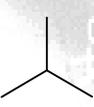
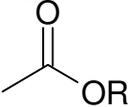


Originally coined *coefficient de communauté* by Paul Jaccard,
Jaccard, P. (1901) *Bulletin del la Société Vaudoisedes Sciences Naturelles* 37, 241-272.



From Jacard... to Tanimoto index!



			Ar-OH				
1	1	0	0	1	0	0	0
1	1	0	1	0	0	0	0

Fragments only in A = 1

Fragments only in B = 1

Fragments in $A \cap B$ = 2

Tanimoto similarity index

$$= \frac{2}{1 + 1 + 2} = 0.5$$

In that paper, a "similarity ratio" is given over *bitmaps*, where each bit of a fixed-size array represents the presence or absence of a characteristic in the plant being modelled.

Tanimoto, T.T. (1957) IBM Internal Report 17th Nov



Similarity... is a symmetrical property?

NERDS ONLY

Usually we think that if A is similar to B, then B is similar to A. An example:



but it not always the case in particular when the two ensembles are very different populated. In this case:

$$\text{Tversky Index} = \frac{A \cap B}{\alpha A_{\text{solo}} + \beta B_{\text{solo}} + A \cap B}$$

where α and β are user-defined parameters.

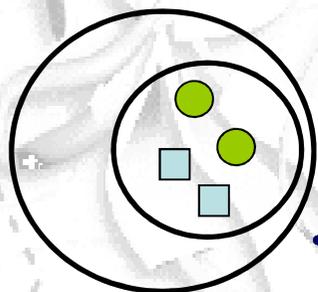


Asymmetric (Tversky) Similarity

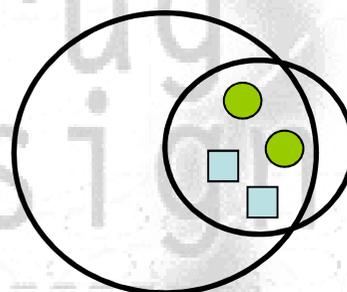
NERDS ONLY

$$\text{Tversky Index} = \frac{A \cap B}{\alpha A_{\text{solo}} + \beta B_{\text{solo}} + A \cap B}$$

- if $\alpha = \beta = 1$, equation reduces to Tanimoto coefficient;
- if $\alpha \neq \beta$, T becomes asymmetric
 - where $\alpha = 1$ and $\beta = 0$, $T = A \cap B / A$
i.e. the fraction of A which it has in common with B
 - when $T = 1.0$, it indicates that A is a substructure of B (at the level of fingerprint matching)
 - when $T \rightarrow 1.0$ it indicates that A is almost a substructure of B



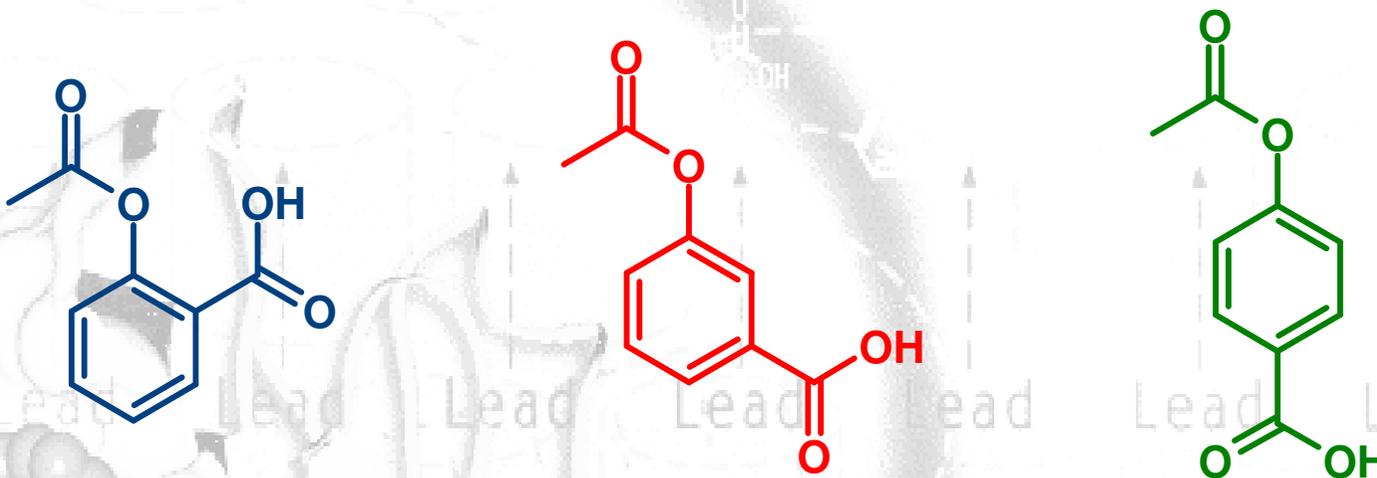
$T = 1$

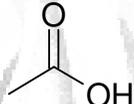
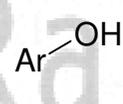
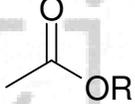
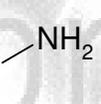
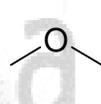
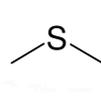


$T \rightarrow 1$



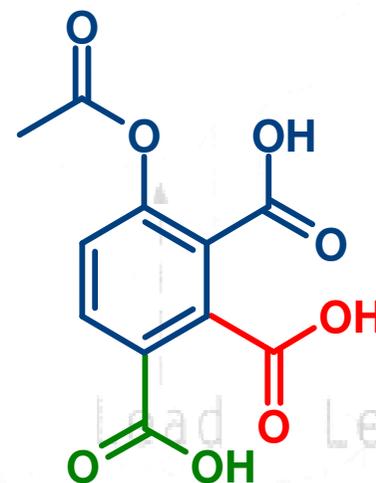
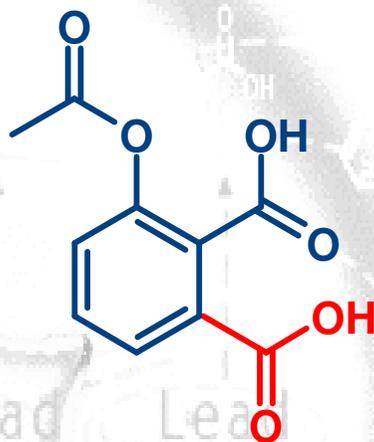
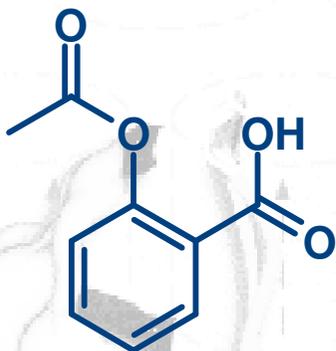
... as for any tool, there are limitations!

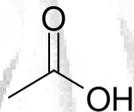
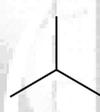
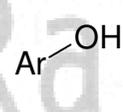
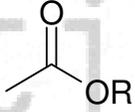
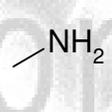
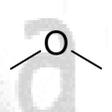
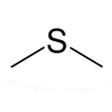


							
1	1	0	0	1	0	0	0
1	1	0	0	1	0	0	0
1	1	0	0	1	0	0	0



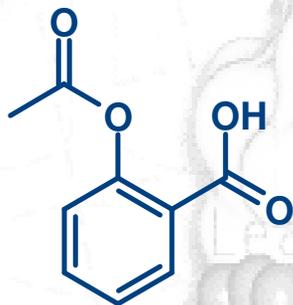
... and if we have more than 1 of the same substituent?



							
1	1	0	0	1	0	0	0
1	2	0	0	1	0	0	0
1	3	0	0	1	0	0	0



Before asking the chemical meaning of Tanimoto similarity index, let's see a clever use:



	Tanimoto
0000000100001101000000101010000000000110000010000100001000001000	0.88
0100010110010010010110011010011100111101000000110000000110001000	0.88
0100010100011101010000110000101000010011000010100000000100100000	0.89
0001101110011101111110100000100010000110110110000000100110100000	0.64
0100010100110100010000000010000000010010000000100100001000101000	0.86
0100011100011101000100001011101100110110010010001101001100001000	0.34
0101110100110101010111111000010000011111100010000100001000101000	0.73
0100010100111101010000100010000000010010000010100100001000101000	0.88
00010001000101000101001000000000000001010000010000100000100000000	0.86
010001010001001100000000000000000001010000001000000000000000000	0.84
01000101000101000000000000000001010000100100000000001000000000000	0.63
0101010101111100111110100000000000011010100011100100001100101000	0.56
0100010100011000010000011000000000010001000000110000000001100000	0.58
00000001000000000100001000000000000001010100000000100000100100000	0.84
01000101000101000000001000000000000100000000000000100001000011000	0.65
0001000100001100010010100000010100101011100010000100001000101000	0.84
0100011100010100010000100001001110010010000010001100000000101000	0.63
0101010100010100010100100000000000010010000010010100100100010000	0.88



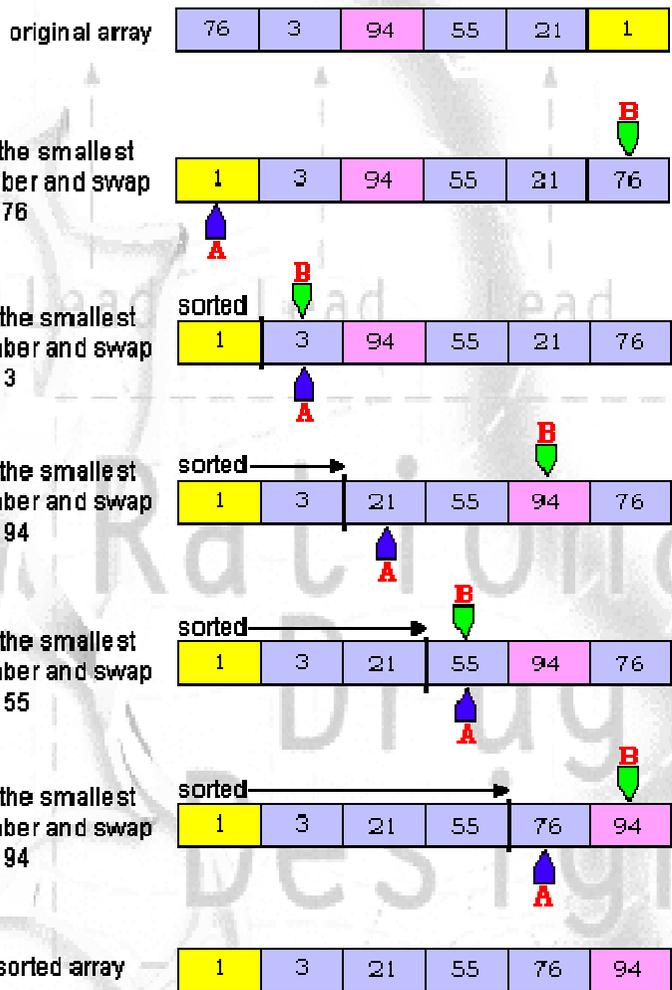
search all compounds that have a Tanimoto similarity index > 0.85!



A bit of... sorting algorithm.

A sorting algorithm is an algorithm that puts elements of a list in a certain order.

... and example:



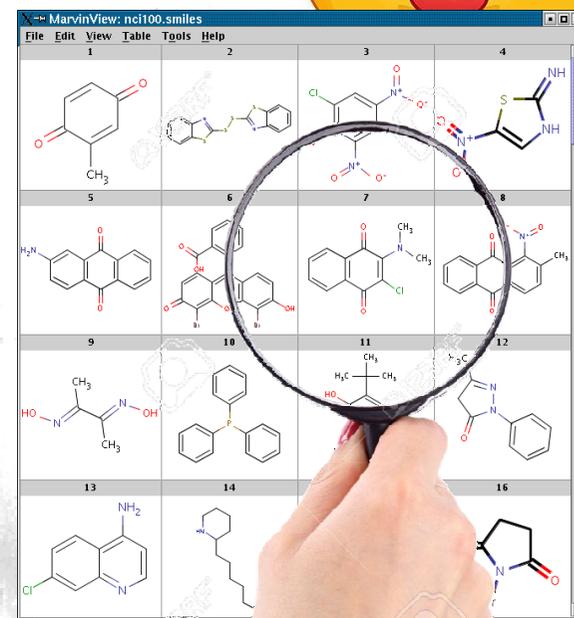
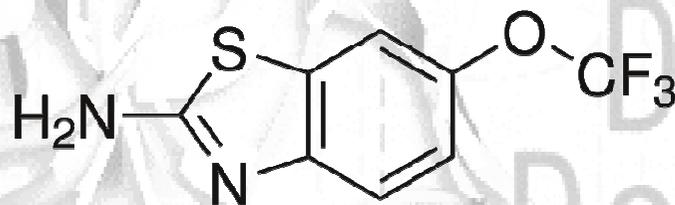
NERDS ONLY

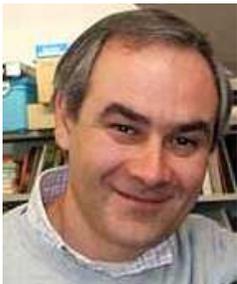


Here is the first important informatics application in medicinal chemistry:

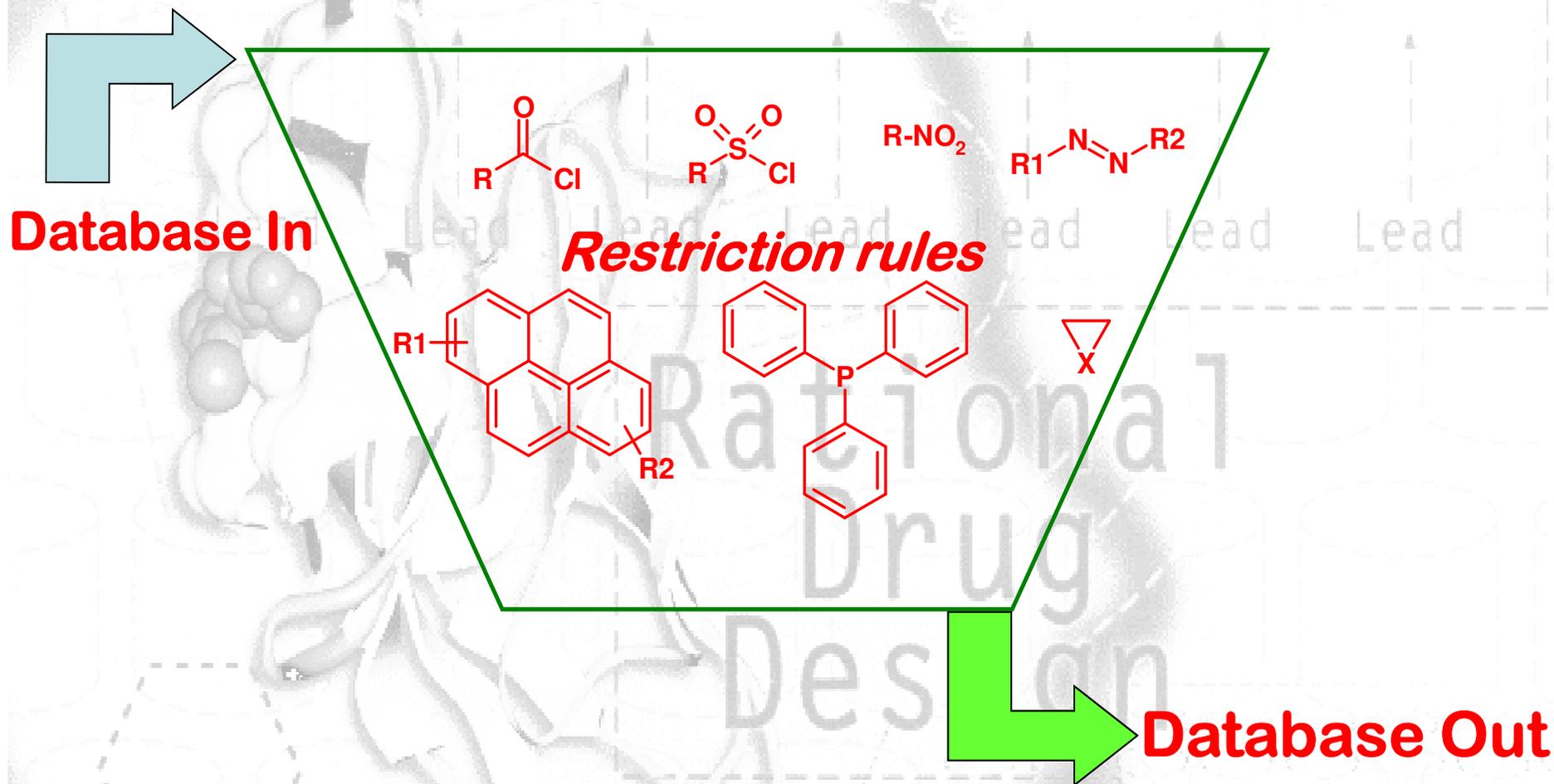


Search all chemical representation close to the target one with a Tanimoto similarity index > 0.85 based on structural keys:





... or we can easily REMOVE compound with undesirable fragments (chemical or biological reactivity, toxicity etc.)!





A very nice example!

INTRODUCING... THE PAINS



Pan-Assay Interference compouNdS

Baell J. and Walters M.A. *Nature* 513, 481-483 (2014)

MS

Confidential and Property of ©2005 Molecular Modeling Section
Dept. Pharmaceutical and Pharmacological Sciences – University of Padova - Italy

S.MORO – PSF: LBDD_1

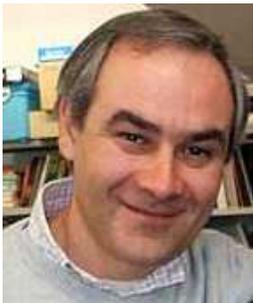
WORST OFFENDERS

Pan-assay interference compounds (PAINS) fall into hundreds of chemical classes, but some groups occur much more frequently than others. Among the most insidious are the eight shown here (reactive portions shown in red and purple). These and related compounds should set off alarm bells if they show up as 'hits' in drug screens.



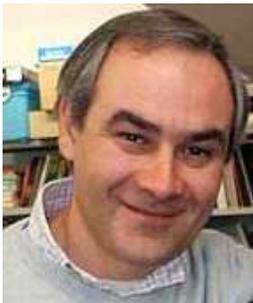
Pan-assay interference compounds (PAINS) encompass some 400 structural classes, but more than half of PAINS in a typical library fall into just 16 easily recognizable categories.

Baell J. and Walters M.A. *Nature* 513, 481-483 (2014)

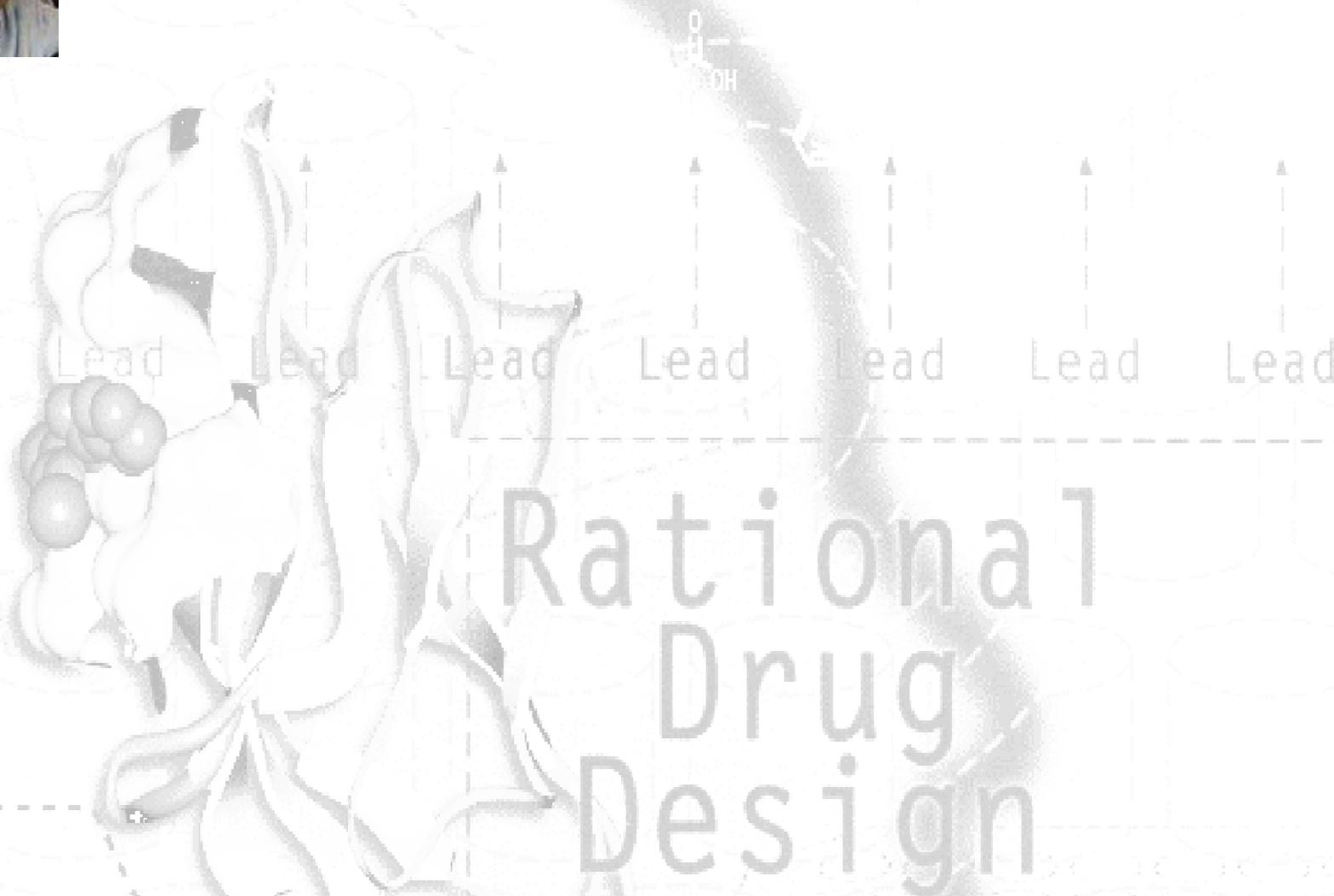


Pharmacophore definition:

A "***pharmacophore***" is a three-dimensional substructure of a molecule that carries ("**phoros**") the essential features responsible for a drug's ("**pharmacon**") biological activity. Alternatively described as an ensemble of interactive functional groups with a defined geometry. Basically, one tries to talk the protein language by finding the "structural and chemical complementaries" (pharmacophore hypothesis) to target receptors.

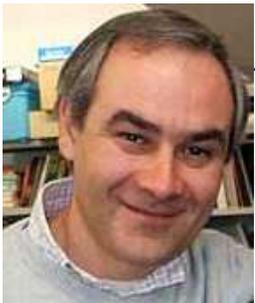


From structure to interaction:

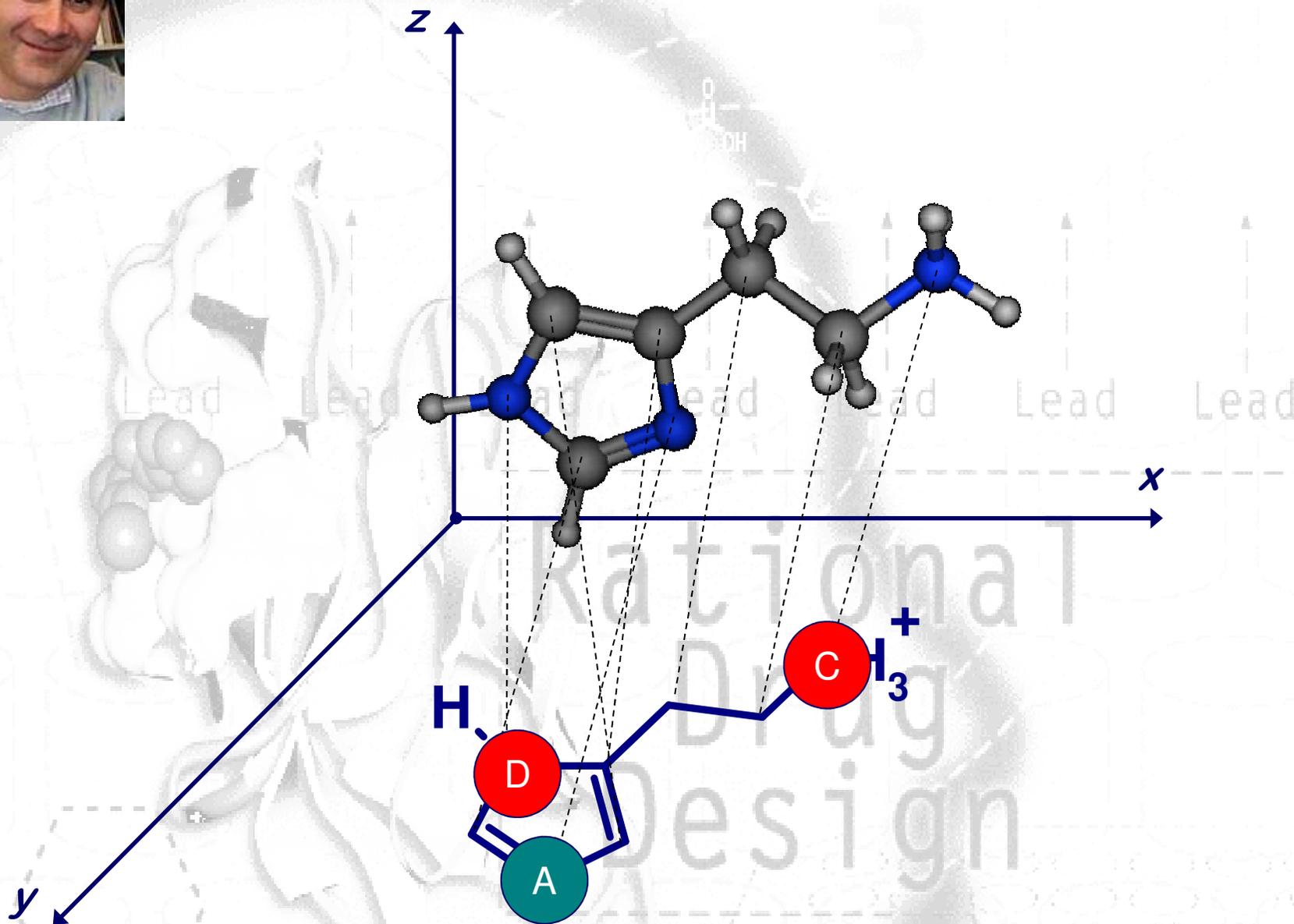


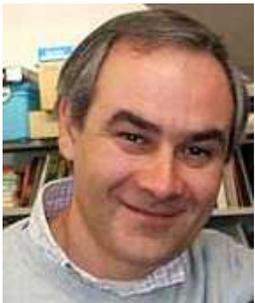
**... a quick refresh: what is
the goal of every SAR study?**

**The generation of
pharmacophoric hypothesis
(models)!!!!**

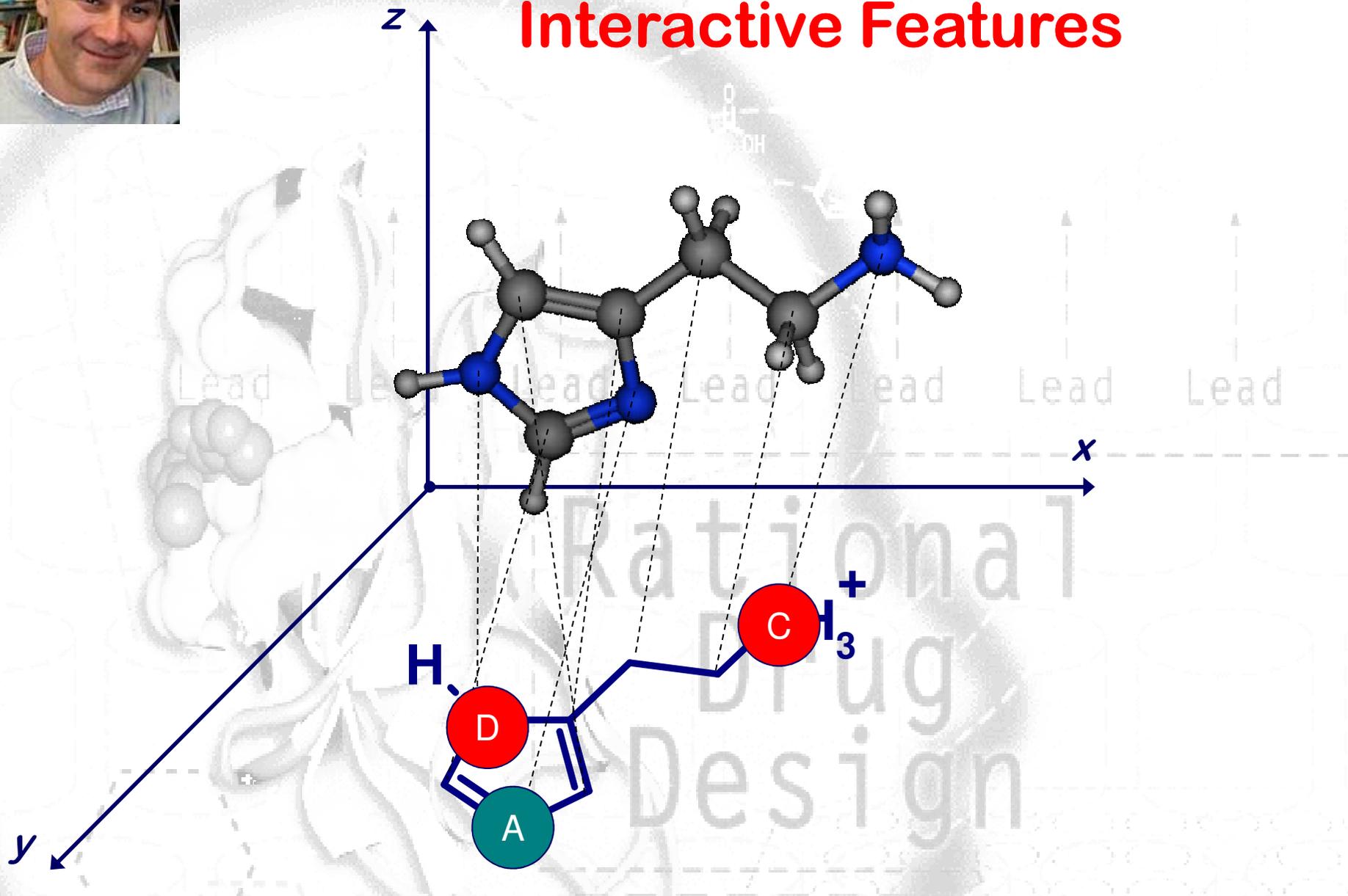


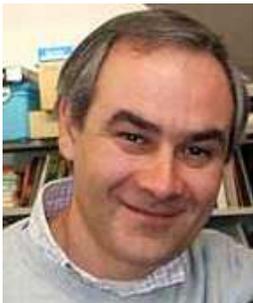
the shadow of the reality:



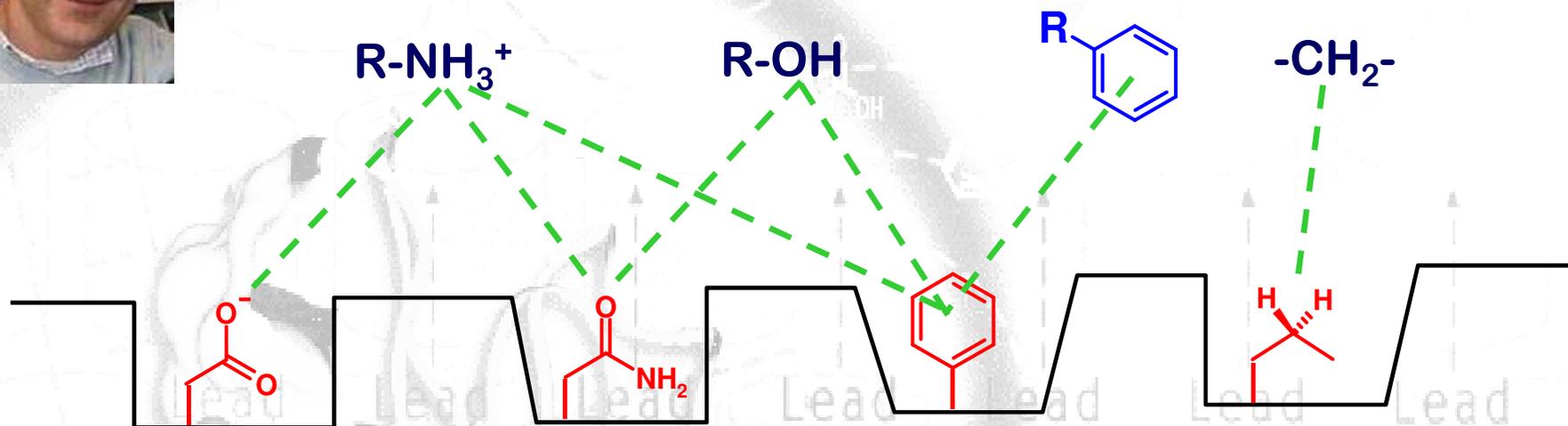


Functional groups can be replaced by Interactive Features





Do you remember...



charge-charge interaction (*ionic bond*):

$$-\Delta G^0 \cong 5 \div 10 \text{ (kcal/mol)}$$

charge-dipole interaction:

$$-\Delta G^0 \cong 1 \div 7$$

charge- π interaction:

$$-\Delta G^0 \cong 8 \div 10$$

hydrogen bond:

$$-\Delta G^0 \cong 1 \div 7$$

charge transfer interaction:

$$-\Delta G^0 \cong 1 \div 6$$

π - π interaction:

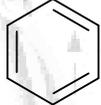
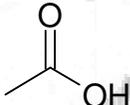
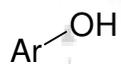
$$-\Delta G^0 \cong 1 \div 2$$

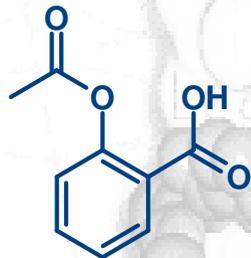
dipole-dipole interaction (van der Waals):

$$-\Delta G^0 \cong 0.5 \div 1$$

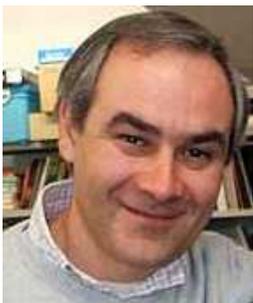


Pharmacophore definition: *From structural key to pharmacophoric key*

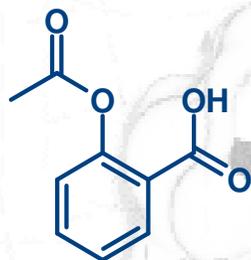
							
Ar	Ac	H	D	A	AD	A	A



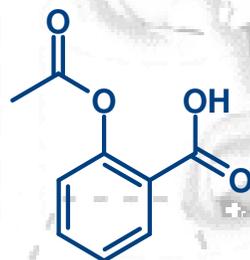
Ar = aromatic
Ac = acid
H = hydrophobic
D = H-bonding donor
A = H-bonding acceptor
C = cation
An = anion
...



From *structural* key to *pharmacophoric* key: expanding the searching potentiality



Ar	Ac	H	D	A	AD	A	A
1	1	0	0	1	0	0	0
1	1	0	0	0	0	1	0
1	1	0	0	0	0	0	1

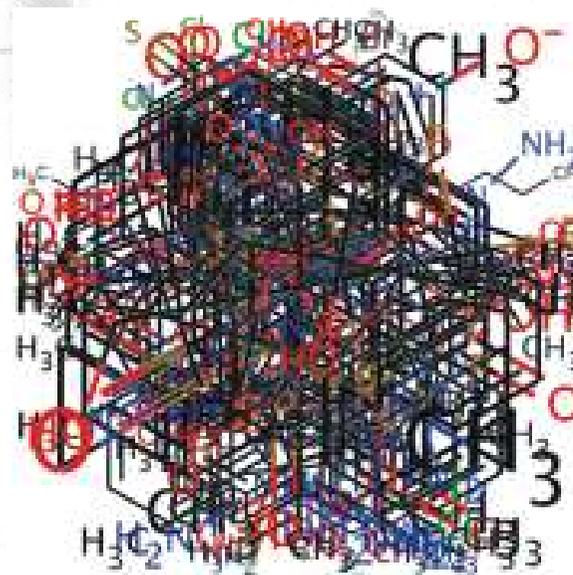


Ar	Ac	A
-----------	-----------	----------



From *structural* key to *pharmacophoric* key: expanding the searching potentiality

Ar Ac A

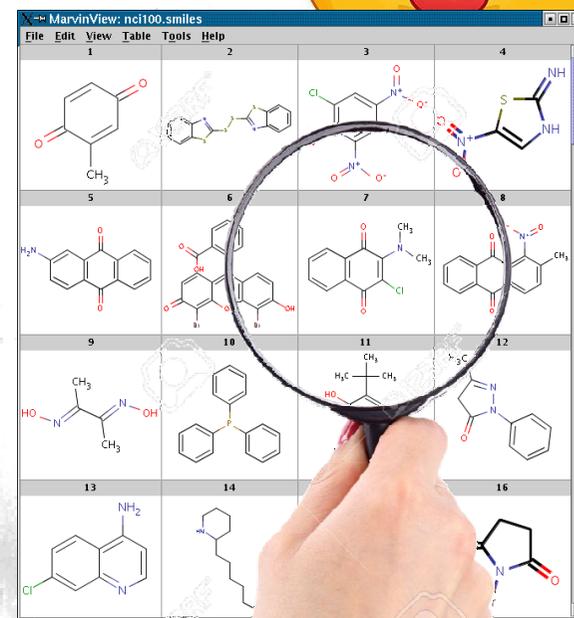
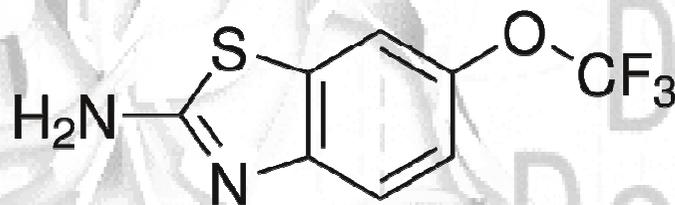


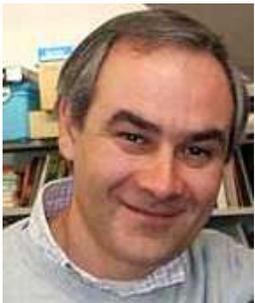


Here is the second important informatics application in medicinal chemistry:

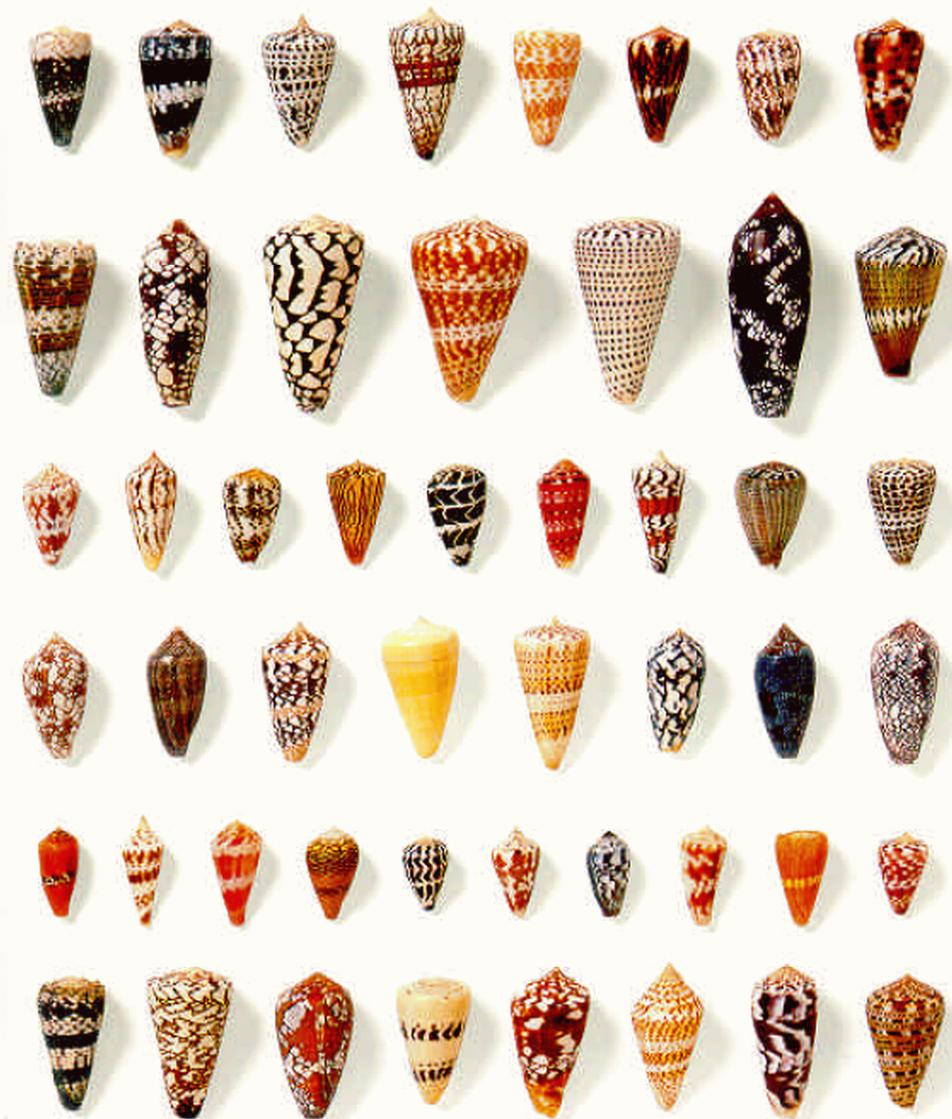


Search all chemical representation close to the target one with a Tanimoto similarity index > 0.85 based on pharmacophore keys:



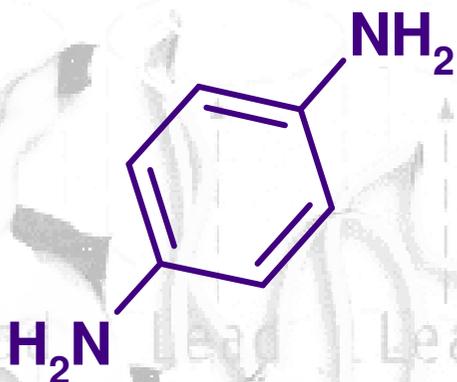


Similarity: structure *versus* properties...

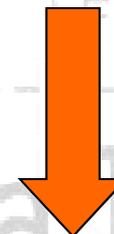




Similarity: structure *versus* properties...



Structure



Properties

PM = 108,14

pKa = 6.2

Volume = 93,9

MP = 142

PSA = 52

nC = 6

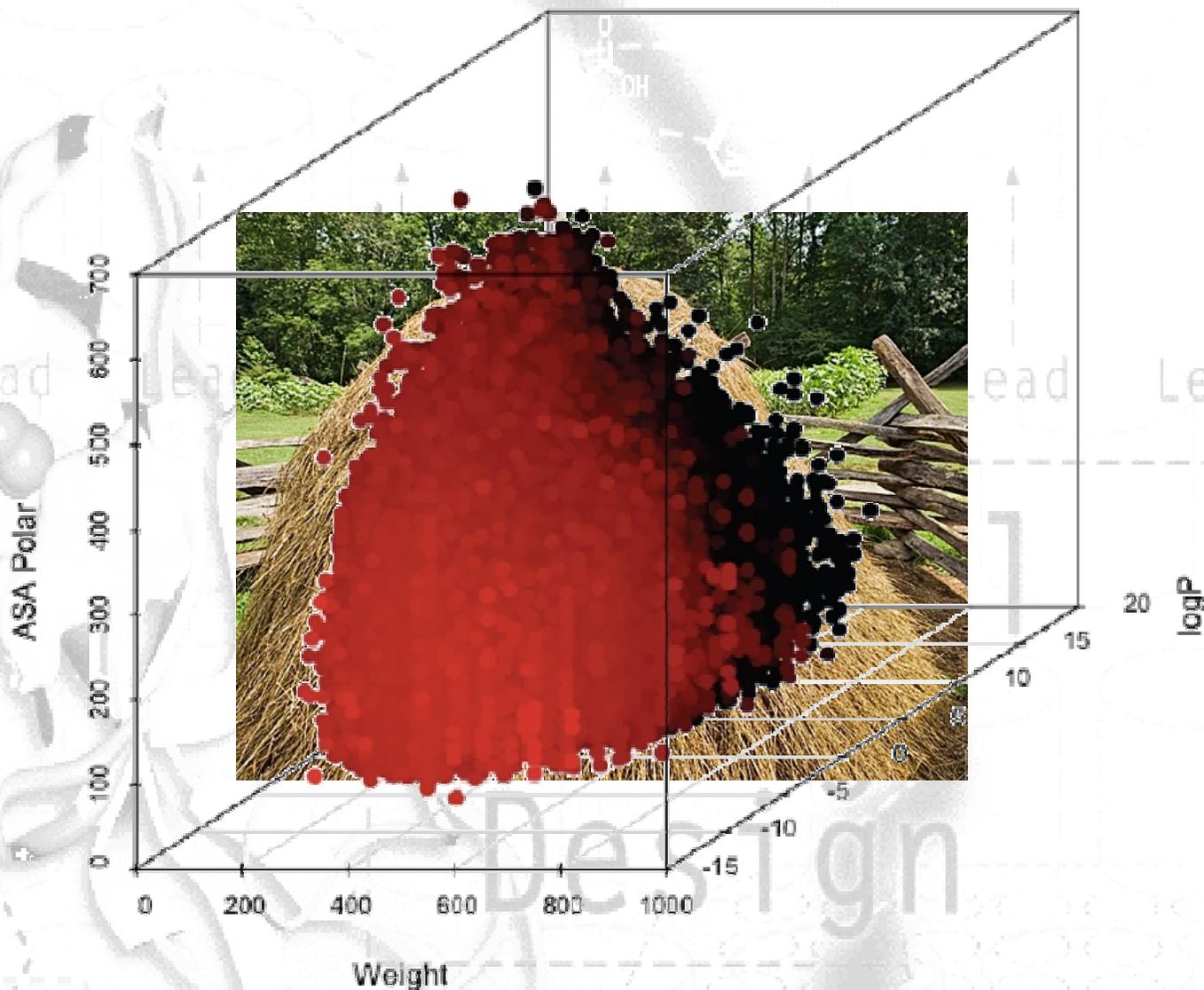
logP = -0.3

...



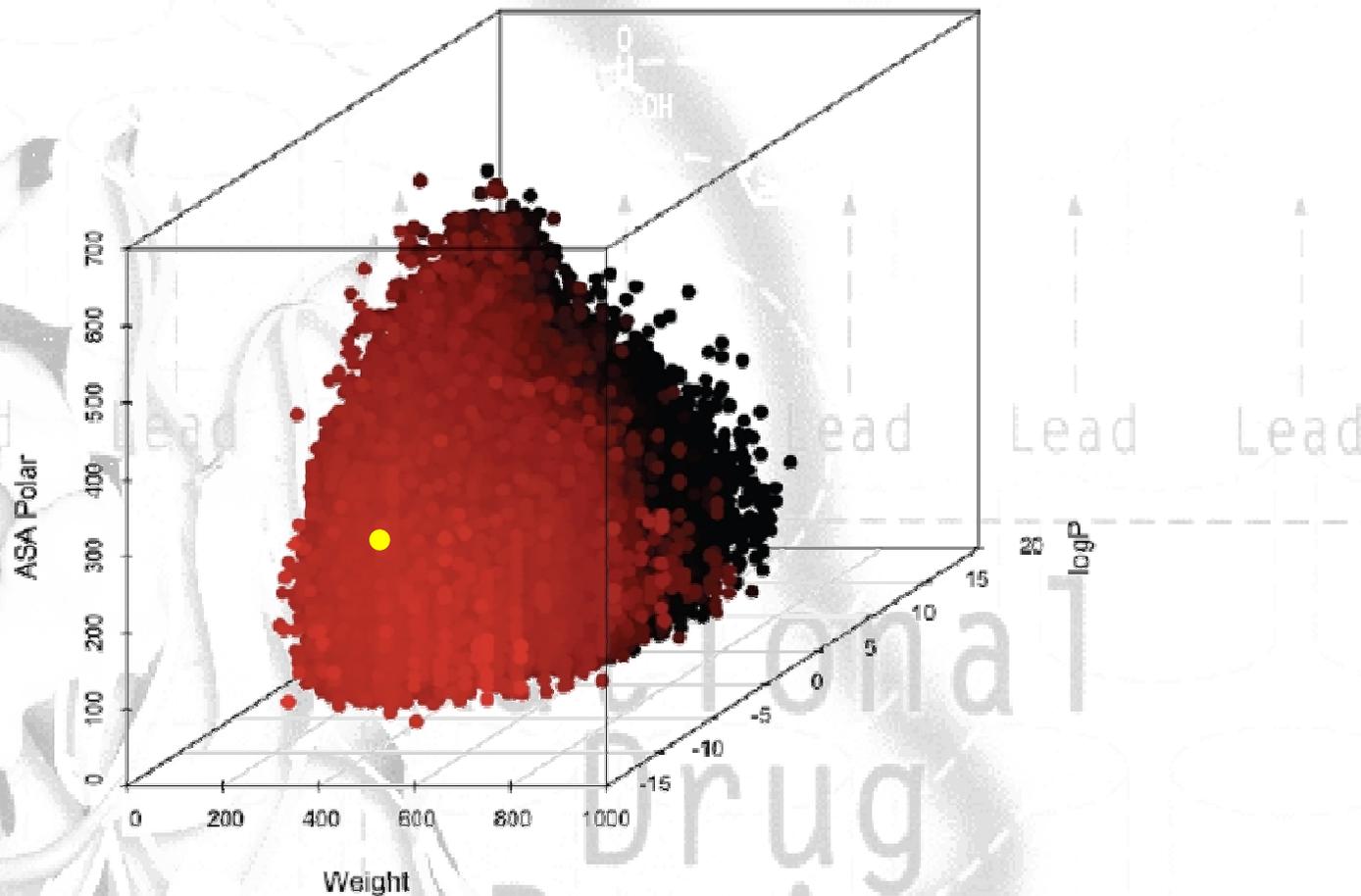
Back to the horrible ratio xx.000:1

Discover a new drug is equal to find a...

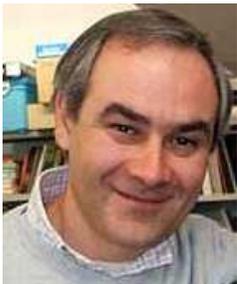




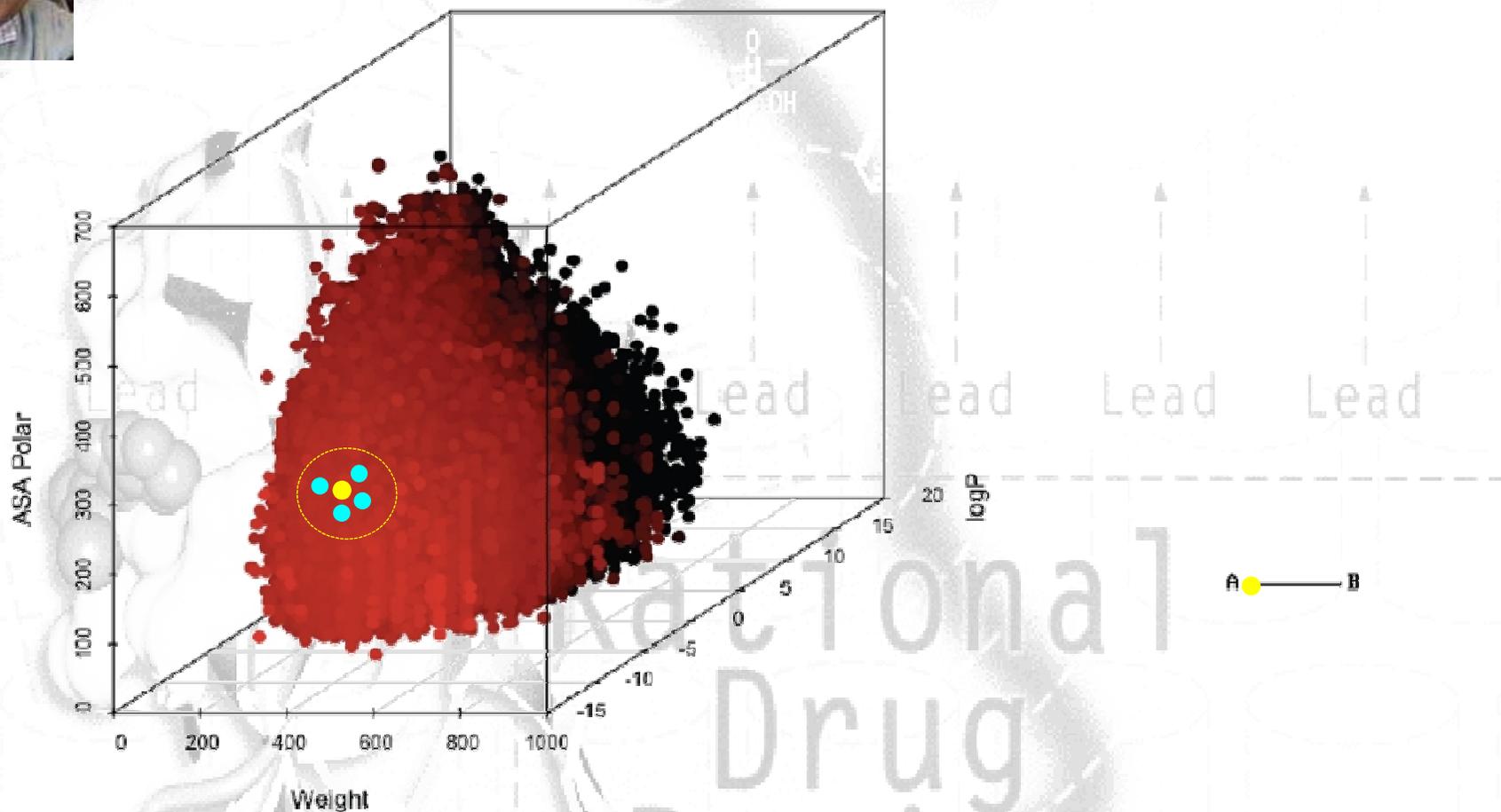
Analyze the nearest chemical space?



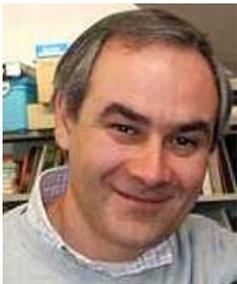
Similar + chemical properties similar chemical structures and similar chemical properties?



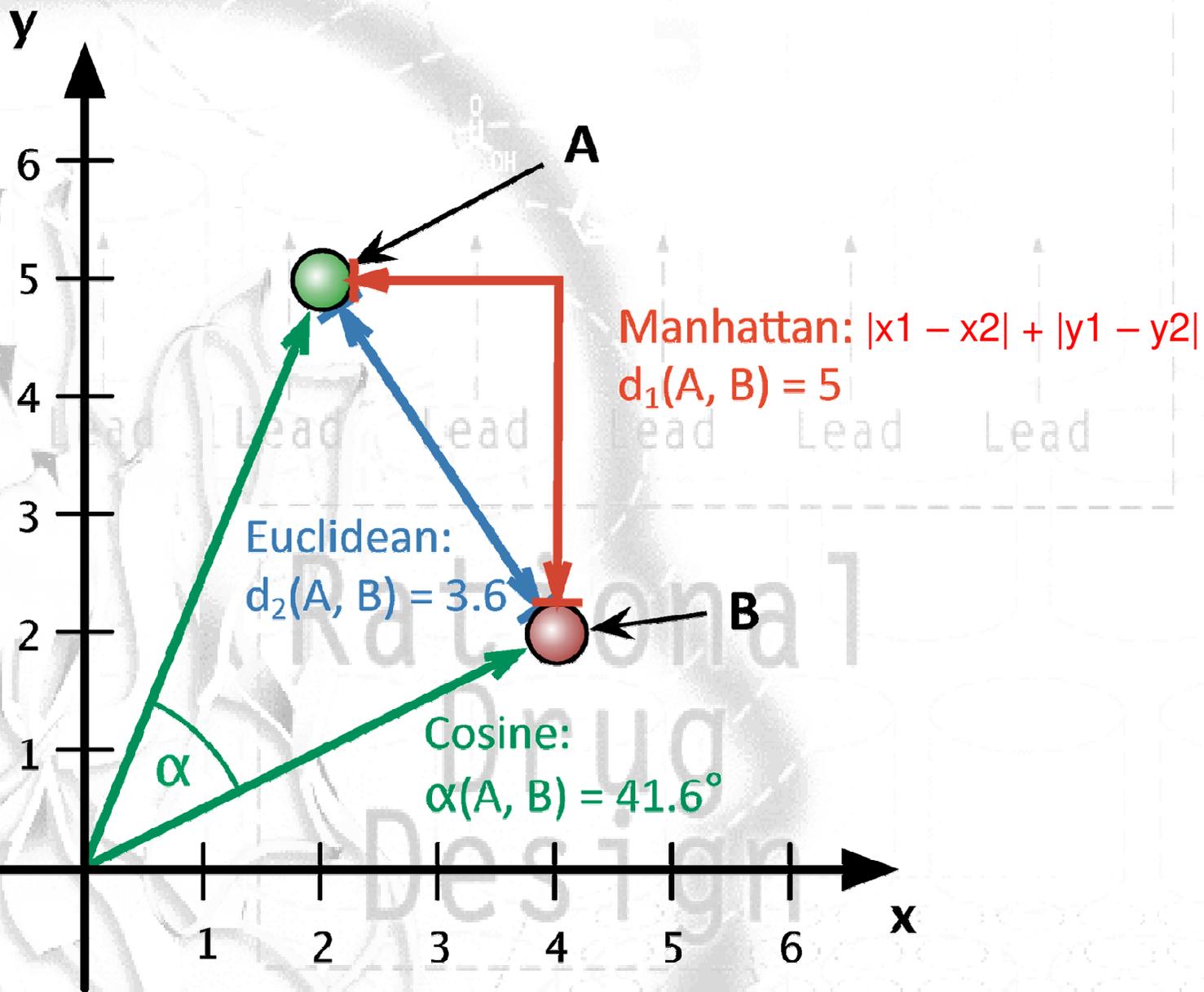
Analyze the nearest chemical space?

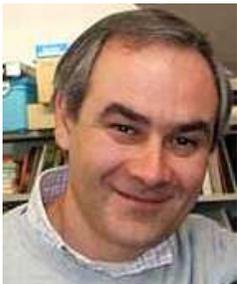


Similar chemical properties similar chemical structures and similar chemical properties?



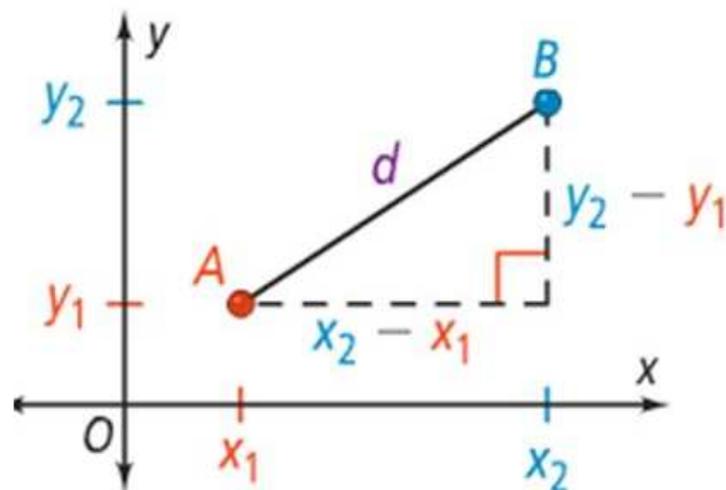
Analyze the nearest chemical space?

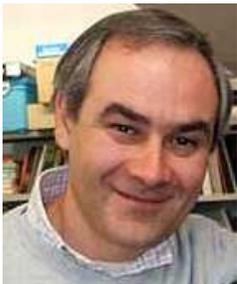




Do you remember how to calculate the distance between two points $A(x_1, y_1)$ and $B(x_2, y_2)$? Facile!!!!

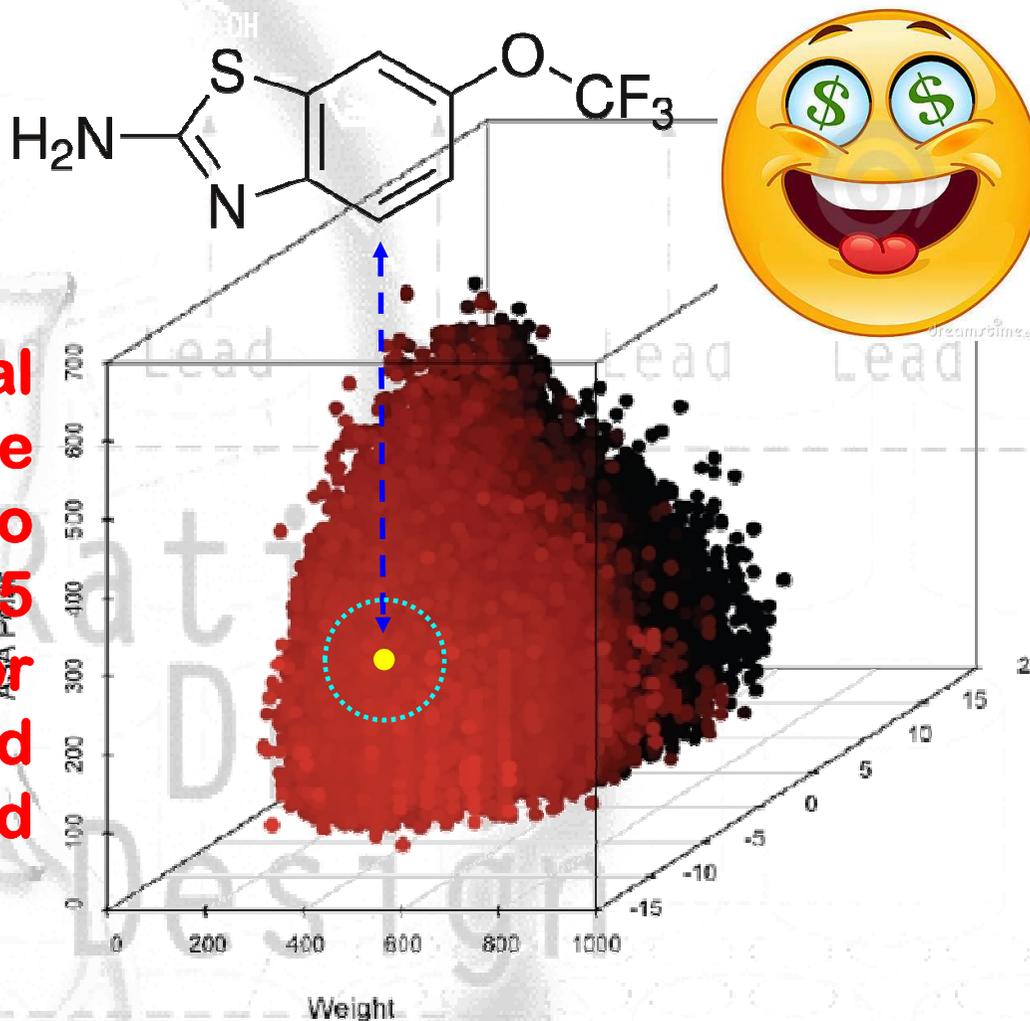
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$





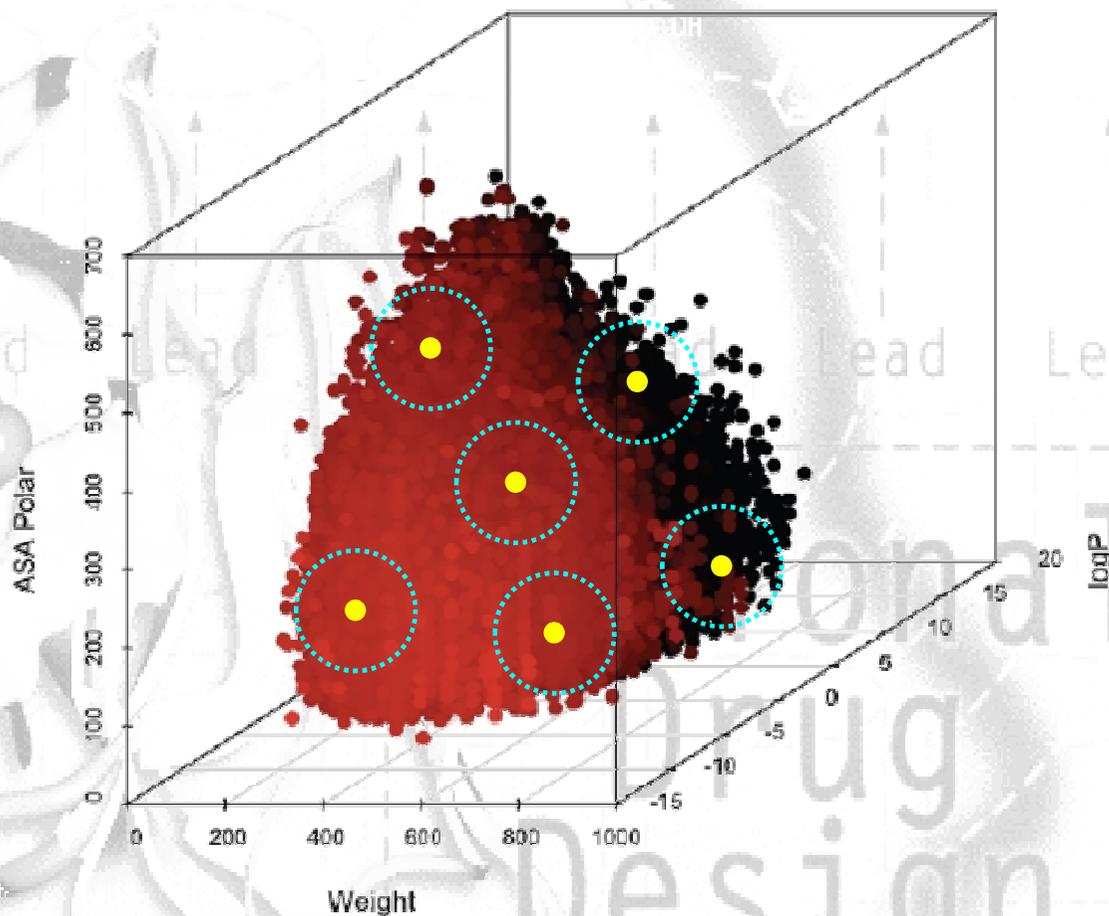
Here is the third important informatics application in medicinal chemistry:

Search all chemical representation close to the target one with a Tanimoto similarity index > 0.85 based on structural or pharmacophore keys, and with a PM = 360 ± 100 and a logP = 3.8 ± 1





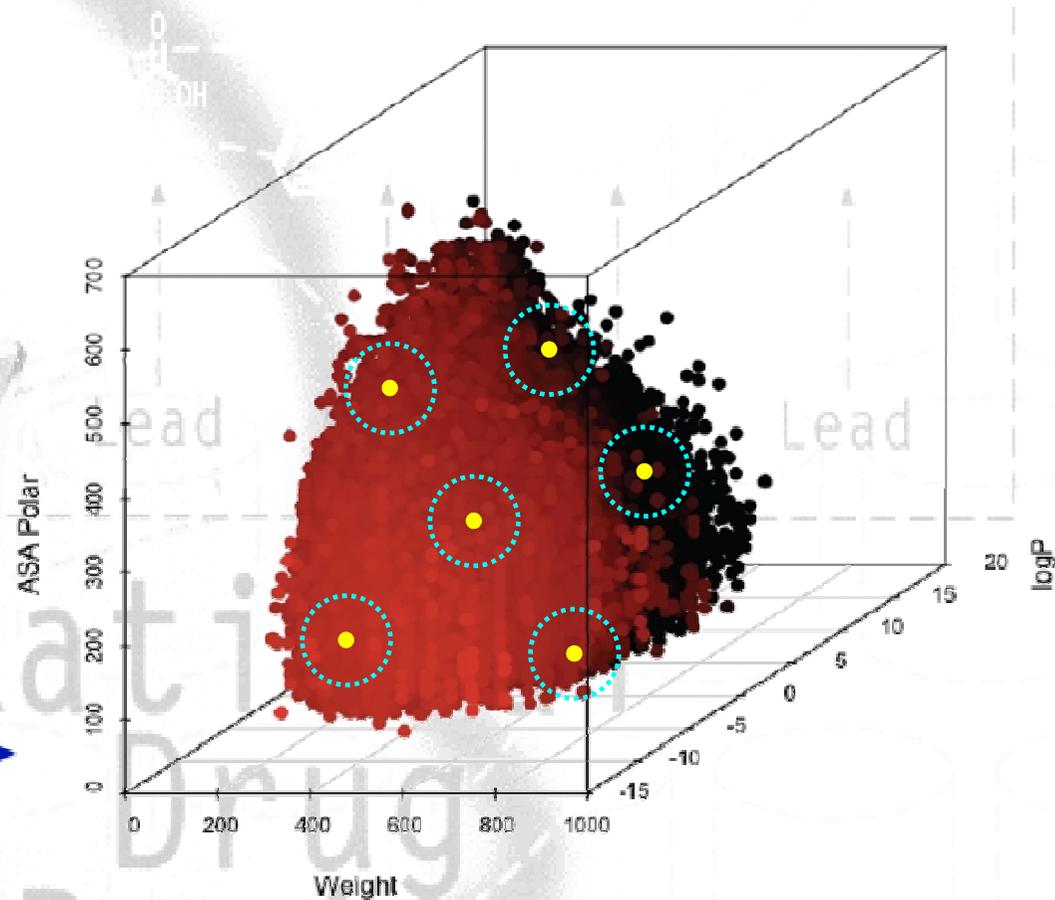
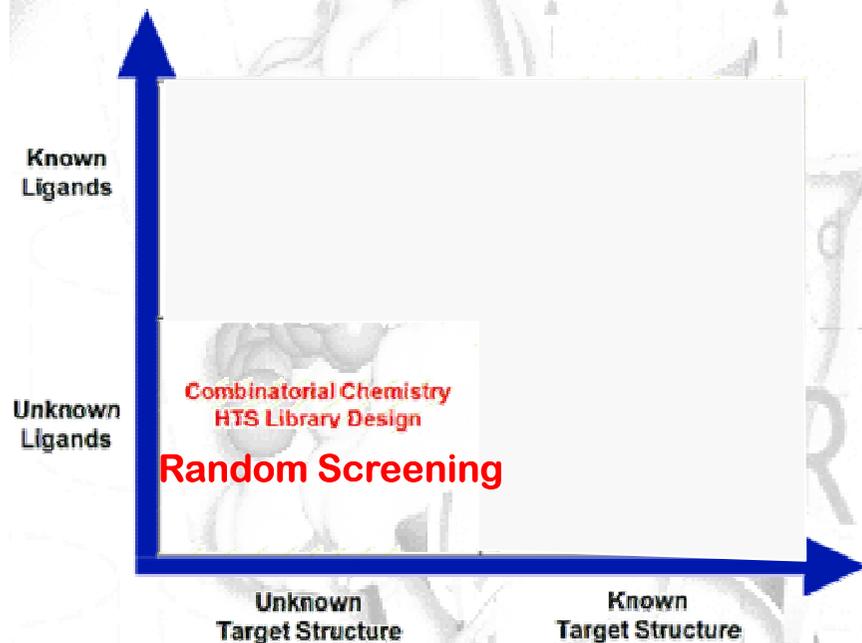
...and don't forget:



1- similarity = diversity



An also the **RANDOM SCREENING** is less **RANDOM**, now!





Lead