A part of this presentation was created

with tools using

Artificial Intelligence

Pr Alain Chavanieu

Where do the therapeutic drugs come from ?

- **From a natural product** with known biological/clinical activity. Sometimes it is possible to start from a natural product or the natural ligand for the target of interest such as a hormone or a substrate. Historically this has also been a very successful strategy in drug discovery, although it is somewhat out of fashion today.

- **From an existing lead or drug** (sometimes referred to as patent busting or best-in-class). Here, the main challenges are to identify a compound with some significant biological/clinical advantage and a good intellectual property position relative to the original compound. This approach has been highly successful in generating incremental improvements to medicines for patients.

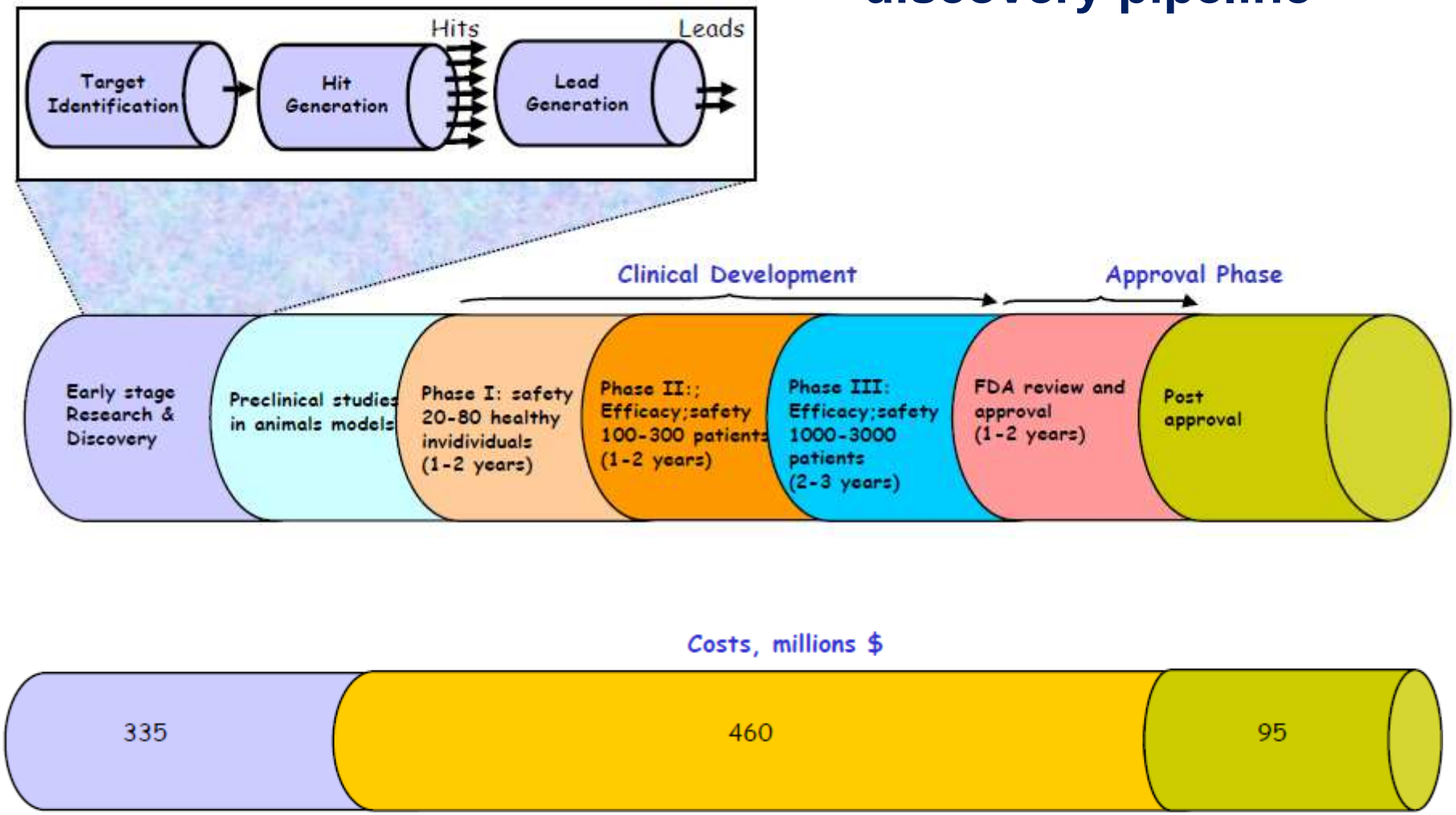
- **High-throughput screening** (HTS) is currently one of the dominant paradigms for lead identification in the pharmaceutical industry. In this method up to a few million compounds are screened against the target of interest; any identified hits are then followed up and optimized into chemical leads. Recent approaches for HTS have moved away from the idea that drug discovery is a numbers game, and have focused on screening 'drug-like' or 'lead-like' compounds that are representative of pharmacophores known to exhibit desirable biological activity.

Chemical leads and technologies

- 1960/70s: natural-product-derived leads
- 1970/80s: computational methods such as quantitative structure–activity relationships (QSAR)
- 1980/1990s: Structure Based Drug-Discovery (with a strong impact around 2000s)
- 1990s: combinatorial chemistry and HTS
- 2000s : FBDD
- 2000s : computational methods such as virtual screening

- New target areas

The classical…drug discovery pipeline



2010 FDA drug approvals

The US Food and Drug Administration approved slightly fewer new drugs than in recent years, and the industry's focus on specialty-care products continued to shine through.

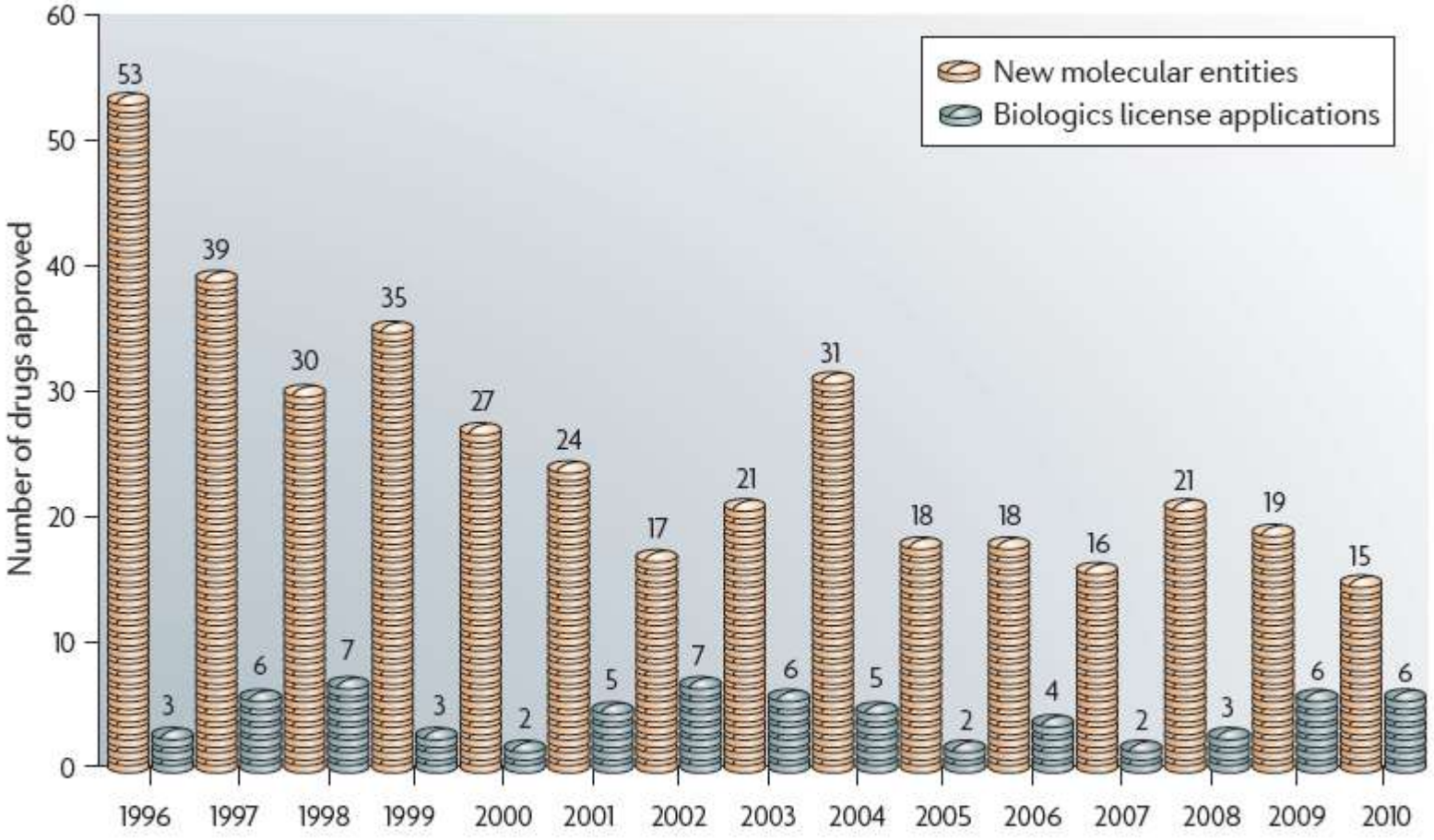


Figure 1 | **FDA drug approvals since 1996.** New molecular entities and biologics license applications approved by the US Food and Drug Administration's (FDA's) Center for Drug Evaluation and Research, by year.

Nature *Asher Mullard* | february 2011 | Volume 10, 82-85

2019 FDA drug approvals

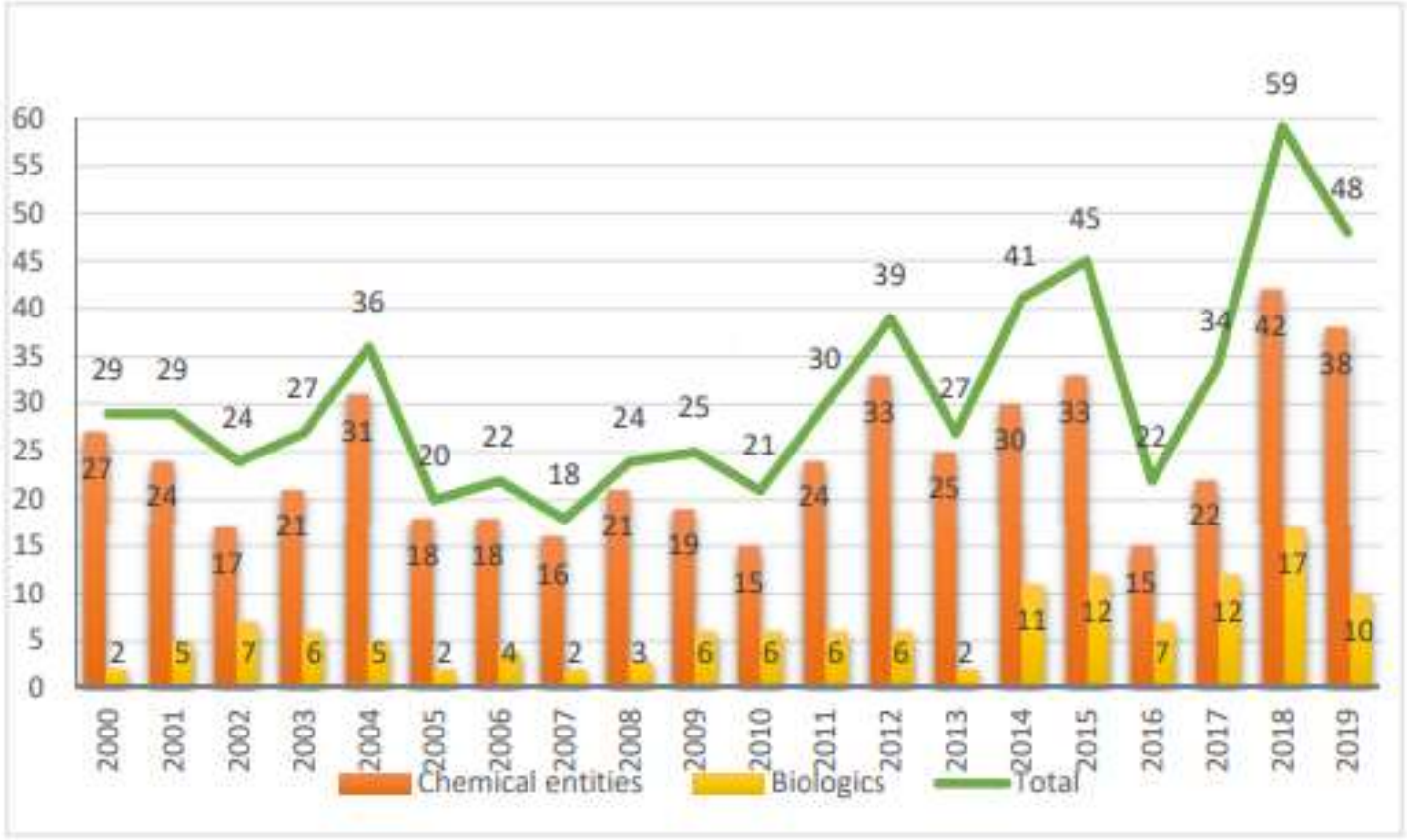
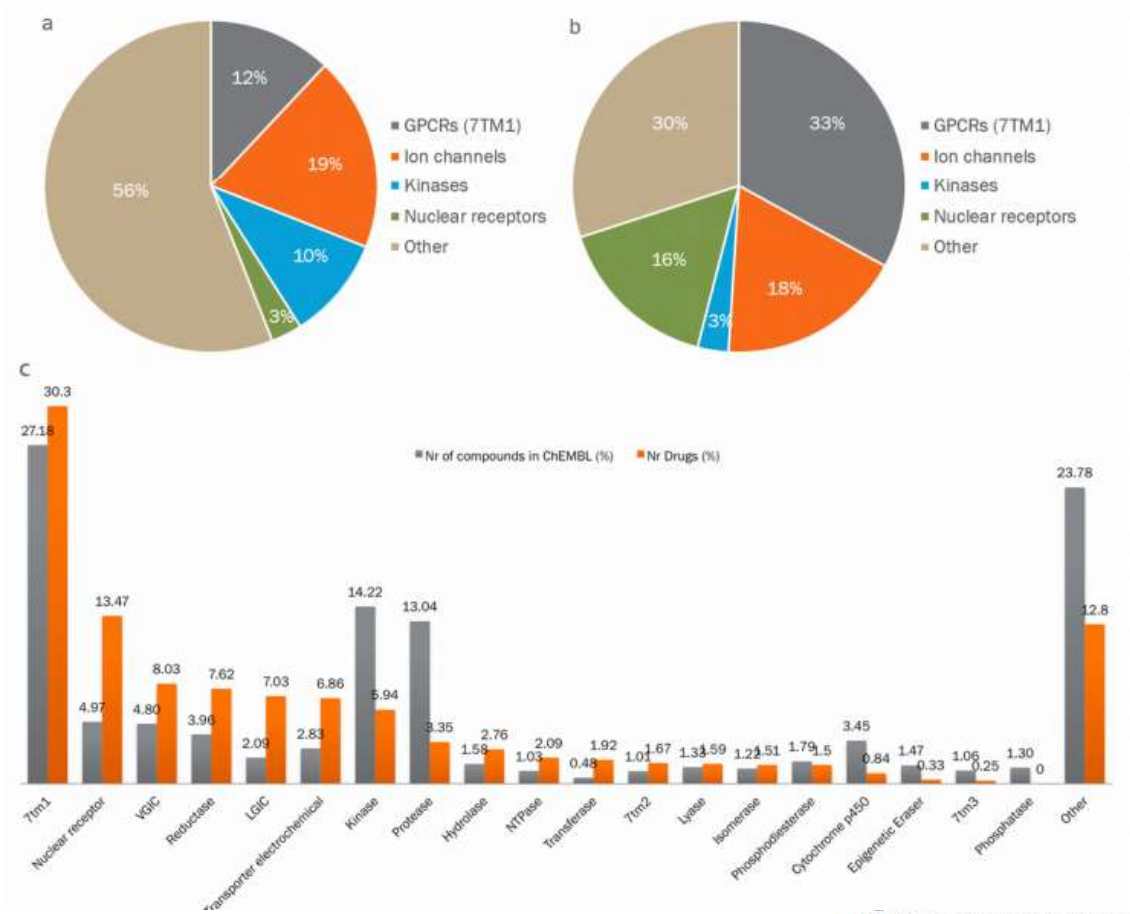


Figure 1. New chemical entities and biologics approved by the FDA in the last two decades [1,6].



Major protein families as drug targets.
(a) Distribution of human drug targets by gene family. (b) distribution by the fraction of drugs targeting those families; the historical dominance of four families is clear. (c) Clinical success of privileged protein family classes. Distribution of non-approved compounds in ChEMBL 20 (extracted from the medicinal chemistry literature, with bioactivity tested against human protein targets) per family class, and distribution of approved drugs (small molecules and biologics) per human protein family class. 7TM, seven transmembrane family; GPCR, G protein-coupled receptor; LGIC, ligand-gated ion channel; NTPase, nucleoside triphosphatase; VGIC, voltage-gated ion channel.

A comprehensive map of molecular drug targets

Nat Rev Drug Discov. 2017 January ; 16(1): 19–34.
doi:10.1038/nrd.2016.230

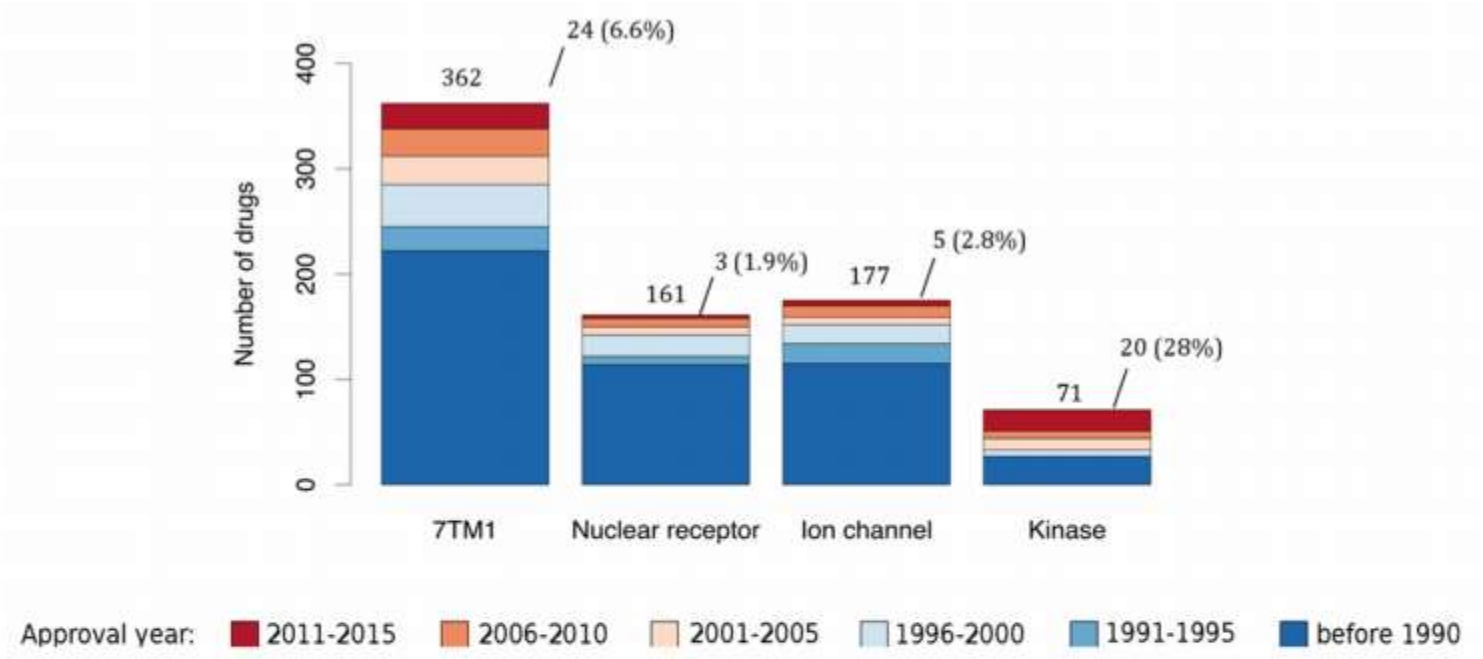


Figure 3.

Innovation patterns in privileged protein classes. Histogram depicting the number of drugs (small molecules and biologics) that modulate the four privileged families, distributed per year of first approval. On top of each bar, the total number of approved drugs is shown, together with the number and percentage of drugs approved since 2011 in respect to the total drugs modulating these four families. A spreadsheet view of this data is provided in supplementary information S6 (table). 7TM1: G-protein coupled receptor 1 family; Ion channel: Voltage-gated ion channel and Ligand-gated ion channel. Drugs without an ATC code (U – Unclassified) were excluded from this analysis.

HTS: High Throughput Screening

Hit: Key molecule (active compound)

Lead: molecule of interest in the process of discovering.

Pharmacophore Constraints : Active part of a compound (3D)

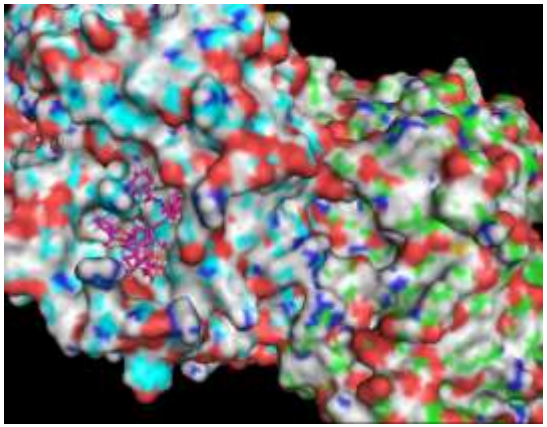
A compound, an interaction with the target and something else

A

10^{60}

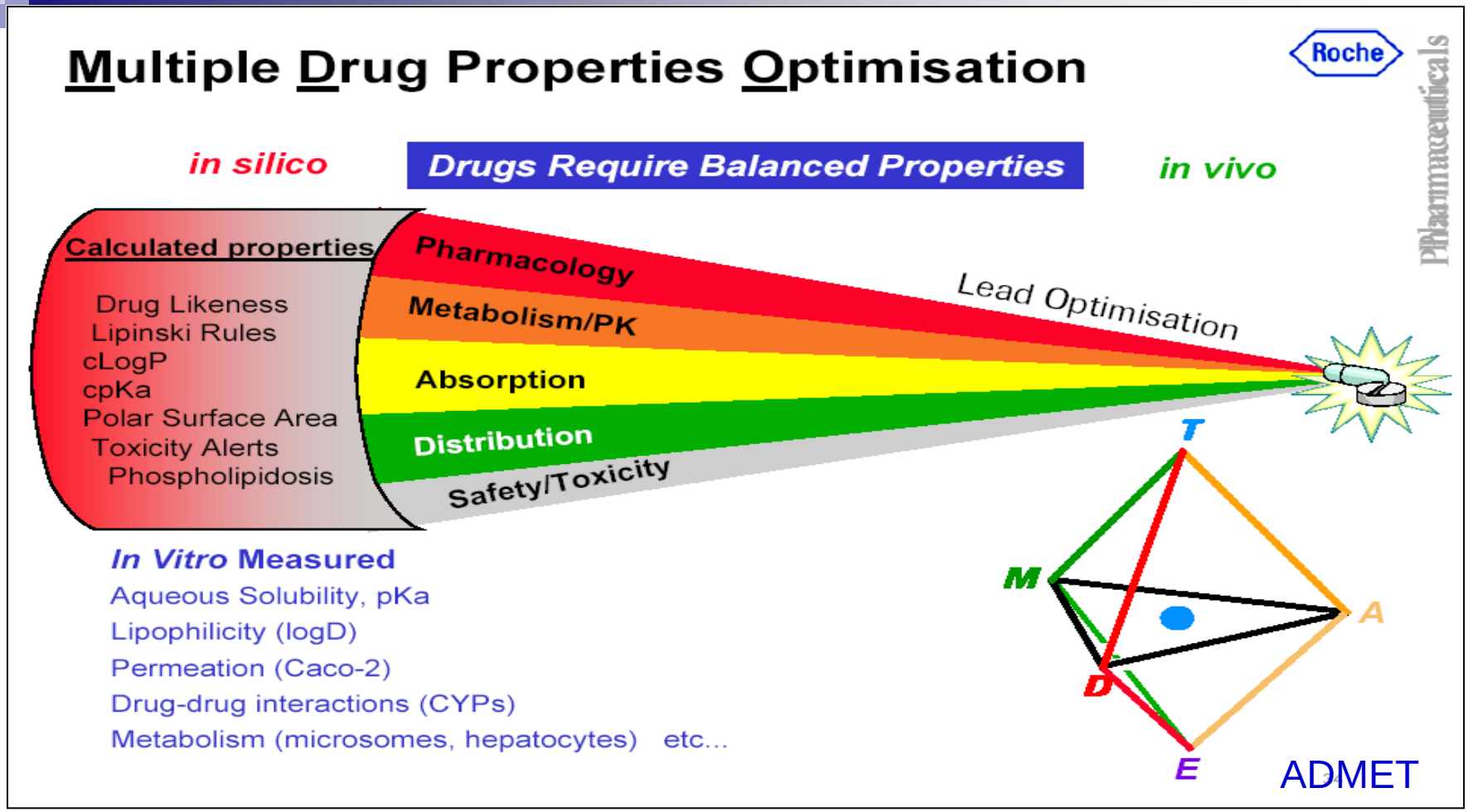
< PM
500 da

Molécules Chimiques



Structure of alpha-beta tubulin from zinc-induced sheets stabilized with taxol

Première étape, l'interaction avec la cible



Example of properties :

Affinity, specificity, solubility, toxicity

Bioavailability, cell permeability, BBB, Binding to serum proteins, Half-life

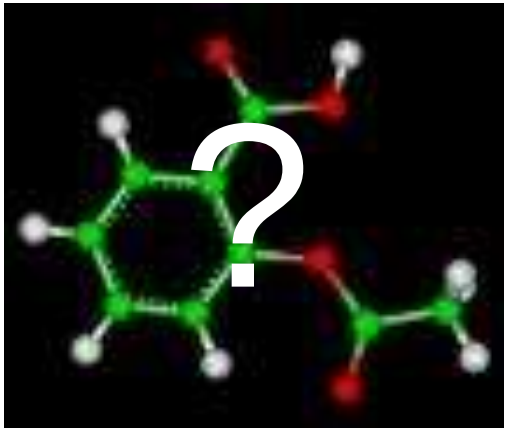
Synthesis

Patentability ...

Chemoinformatic

Definition

Chemoinformatic is to combine computers and chemical informations to a range of problems in the fields of Chemistry and Biology.



The gathering
(development, creation, organization, storage),

The systematic use of chemical information,
(analysis, visualization ...)

-> to predict *In Silico* the behavior of unknown compounds.

Integrated with research and development strategies
in the pharmaceutical industry

Domains of Chemoinformatics

Chemoinformatics is an interdisciplinary field that combines chemistry, computer science, and information technology to analyze and model chemical data.

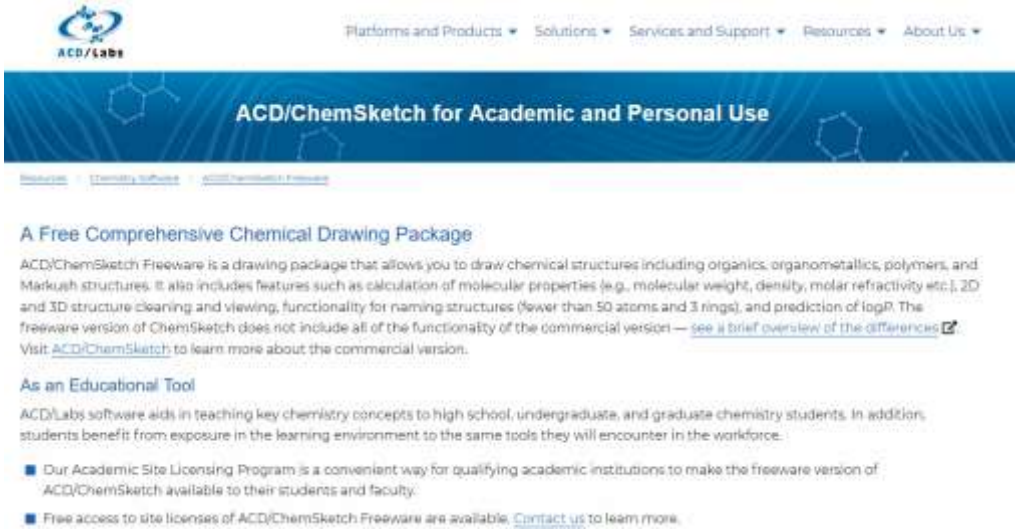
1. Chemical Structure Databases, visualization, properties, searching and retrieving :

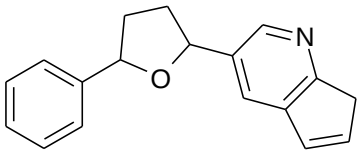
- Creating 2D and 3D visual representations of molecules : ChemSketch…
- Managing vast databases of chemical structures : PubChem, ChemSpider. PDB, CCSD
- Representation and research structures and substructures : Molecular Graphes.
- Similarity search (2D / 3D), clustering and diversity analysis : Tanimoto, clustering.
- Search chemical molecules, patent databases or chemical reactions
- QSAR (Quantitative Structure-Activity Relationship) modeling.

2. Molecular Modeling ant Interactions :

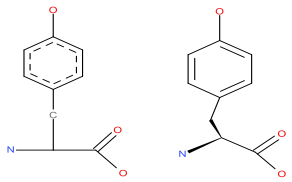
- Predicting molecular structures and properties : homology modeling.
- Molecular dynamics simulations.
- Molecular docking simulations.
- Pharmacophore modeling.
- AI.

Draw and visualise





AND Enantiomer



AutoNom Name:
2-Amino-3-(4-hydroxy-phenyl)-propionic acid

(S)-2-Amino-3-(4-hydroxy-phenyl)-propionic acid

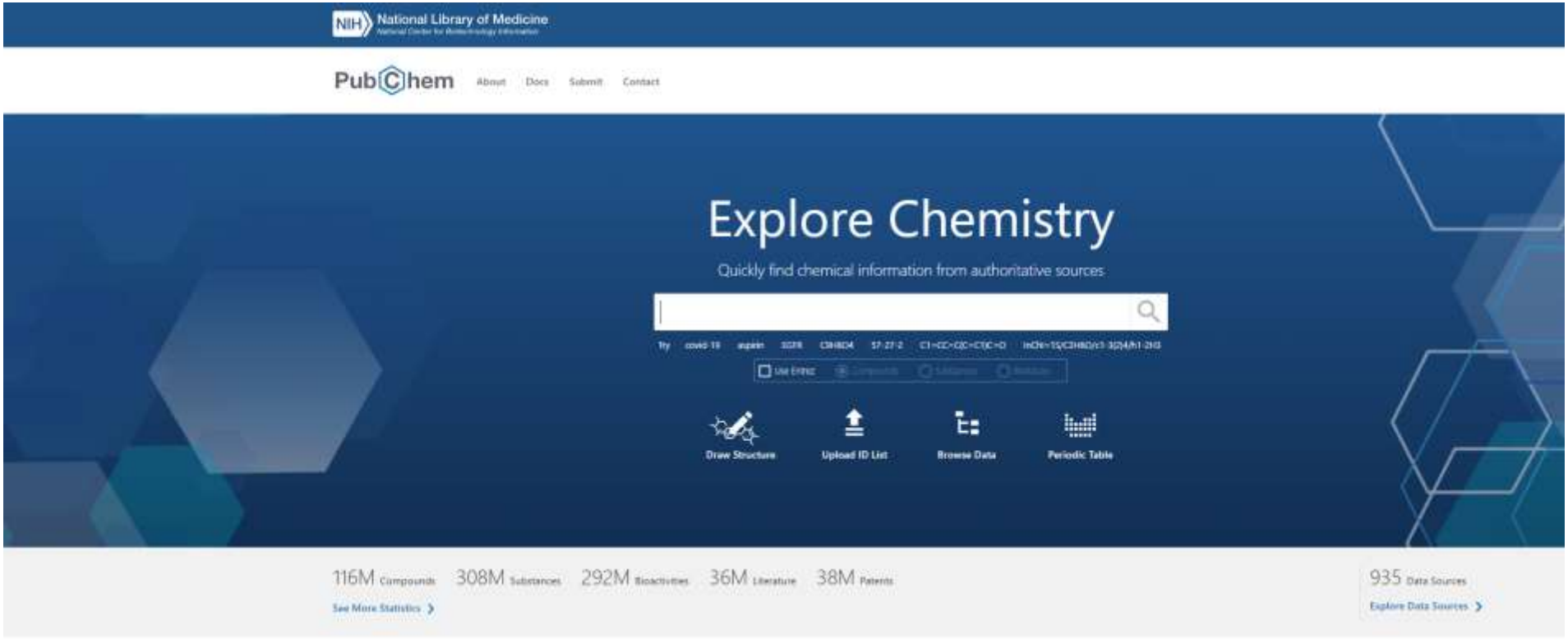
For small chemical coumpounds
Passage 2D→3D; CONCORD, CORINA

1) The input structure is analyzed and separated into rings systems and acyclic atoms.

2) Bond lengths and bond angles are taken from a table. They depend on atom type and bond order. The atom types are rather detailed. For carbon, e.g., 21 atoms types are considered.

3) Ring systems are processed by the assignment of a general conformation (e.g., chair, boat, etc) to each ring. The rings are ordered according to a certain priority and are optimized in steps in this order by the minimization of a special strain function in internal coordinate space. The coordinates of rings already previously processed (on a higher level of priority) remain unchanged.

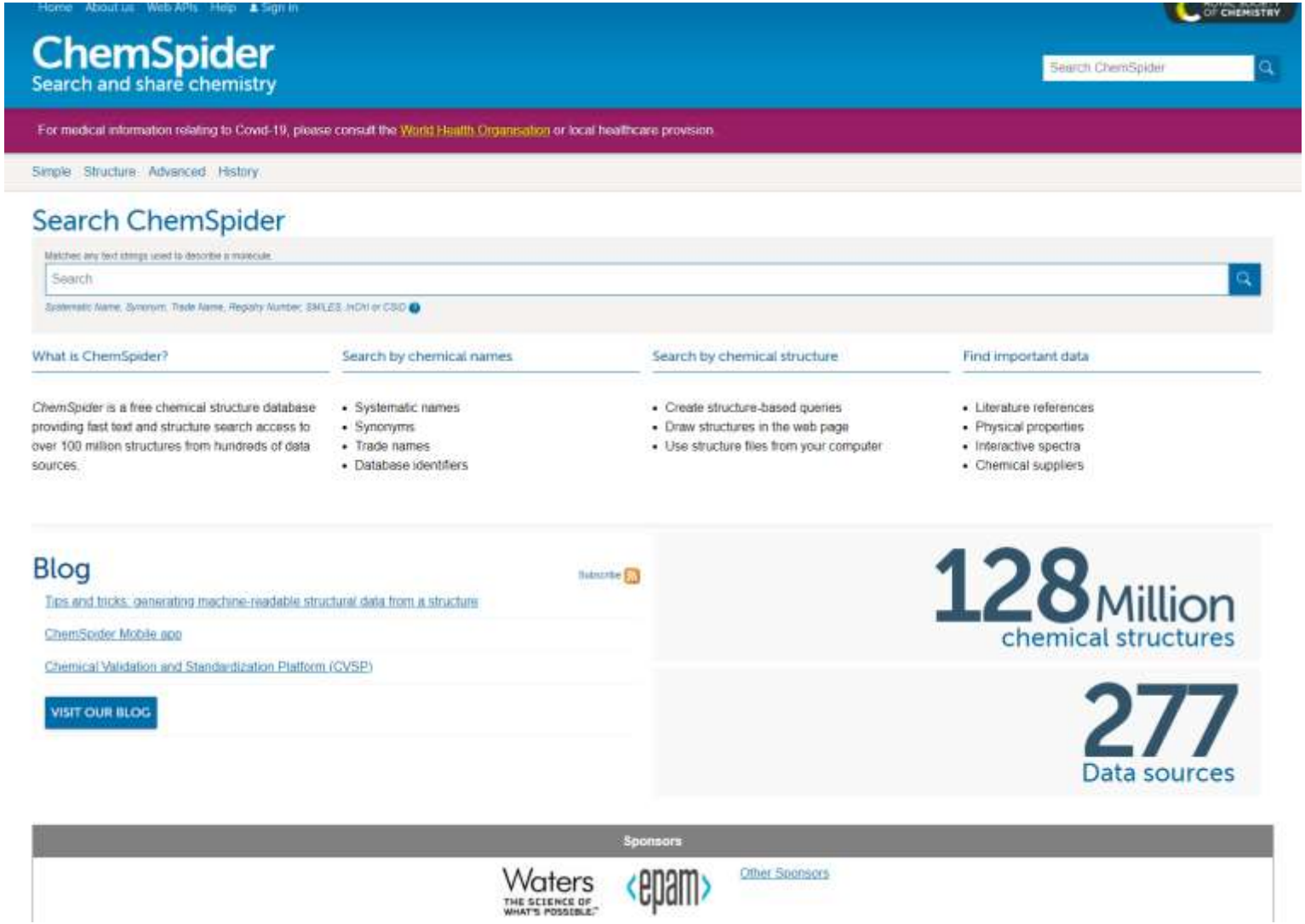
4) Finally, the torsional angles of the acyclic parts are set to values, which minimize the steric interactions of all 1-4, 1-5, and 1-6 interactions. Close contacts are relaxed by a limited energy minimization.



What is PubChem?

PubChem is the world's largest collection of freely accessible chemical information. Search chemicals by name, molecular formula, structure, and other identifiers. Find chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations and more.

We are constantly adding new data and working on improving interfaces to chemical information. Please check back often!


Home About us Web APIs Help Sign in
ChemSpider
Search and share chemistry
Search ChemSpider
For medical information relating to Covid-19, please consult the World Health Organisation or local healthcare provision
Simple Structure Advanced History
Search ChemSpider
Matches any text strings used to describe a molecule.
Search
Systematic Name, Synonym, Trade Name, Registry Number, SMILES, InChI or CSID
What is ChemSpider?
ChemSpider is a free chemical structure database providing fast text and structure search access to over 100 million structures from hundreds of data sources.
Search by chemical names
Systematic names
Synonyms
Trade names
Database identifiers
Search by chemical structure
Create structure-based queries
Draw structures in the web page
Use structure files from your computer
Find important data
Literature references
Physical properties
Interactive spectra
Chemical suppliers
Blog
Subscribe
Tips and tricks: generating machine-readable structural data from a structure
ChemSpider Mobile app
Chemical Validation and Standardization Platform (CVSP)
VISIT OUR BLOG
128 Million
chemical structures
277
Data sources
Sponsors
Waters
THE SCIENCE OF WHAT'S POSSIBLE.
epam
Other Sponsors

From organic chemical space to identification and therapeutic compounds

Chemical space, 500daltons : 10^{60}

Organic chemistry : Less than 200 million compounds

Commercial compounds : More than 50 millions ?

Pharmaceutical Industry : 1 million

HTS : 10000 (oriented) to 500.000

→ ***25 NME / year.***

The CAS database (102 million compounds with inorganic compounds, 65 million of polymers…).
The Beilstein database (10 million substances).
PubChem contains (111 million entries).

There is a great number of more specialized databases for diverse branches of organic chemistry.

some websites and tools for calculating and predicting chemical properties:

1.ChemAxon Property Calculator (https://www.chemaxon.com/products/calculators/**)**: ChemAxon offers a set of tools for calculating various chemical properties such as logP (partition coefficient), solubility, polarity, and more.

2.ChemSpider (https://www.chemspider.com/**)**: In addition to being a chemical structure database, ChemSpider also provides calculations of chemical properties for listed compounds.

3.ChemDraw (https://www.perkinelmer.com/lab-solutions/resources/docs/CHU/CHU_MDL_CalculationCard.pdf**)**: ChemDraw, developed by PerkinElmer, includes modules for calculating chemical properties and predictions, including molecular weight, polarity, solubility, and more.

4.RDKit (https://www.rdkit.org/**)**: RDKit is an open-source collection of cheminformatics software that offers a wide range of chemical property calculations.

5.ACD/ChemSketch (https://www.acdlabs.com/products/draw_nom/draw/**)**: ACD/ChemSketch is a molecular drawing software that includes tools for predicting various chemical properties.

6.Marvin (https://chemaxon.com/products/marvin**)**: Marvin, also developed by ChemAxon, is another cheminformatics software that provides chemical property calculation capabilities.

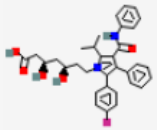
7.PubChem (https://pubchem.ncbi.nlm.nih.gov/**)**: PubChem, an NCBI resource, offers information on chemical properties for thousands of compounds, although online calculations are limited.

8.ChemDoodle (https://www.chemdoodle.com/**)**: ChemDoodle offers a suite of tools for calculating and predicting chemical properties.

Atorvastatin

Cite this Record

Vendors | Drug Information | Pharmacology | Literature | Patents | Bioactivities



PubChem CID:	60823
Chemical Names:	Atorvastatin; 134523-00-5; Cardyl; Sotis; Lipitor; Tozalip; More...
Molecular Formula:	$C_{33}H_{35}FN_2O_5$
Molecular Weight:	558.639803 g/mol
InChI Key:	XUKUURHRXDUEBC-KAYWLYCHSA-N
UNII:	A0JWA85V8F
Modify Date:	2016-09-17
Create Date:	2005-06-24

Atorvastatin is a pyrrole and heptanoic acid derivative,HYDROXYMETHYLGLUTARYL-COA REDUCTASE INHIBITOR (statin), and ANTICHOLESTEREMIC AGENT that is used to reduce serum levels of LDL-CHOLESTEROL; APOLIPOPROTEIN B; AND TRIGLYCERIDES and to increase serum levels of HDL-CHOLESTEROL in the treatment of HYPERLIPIDEMIAS and prevention of CARDIOVASCULAR DISEASES in patients with multiple risk factors.

from MeSH

Atorvastatin is a HMG-CoA Reductase Inhibitor. The mechanism of action of atorvastatin is as a Hydroxymethylglutaryl-CoA Reductase Inhibitor.

FDA Pharmacology Summary from FDA Pharm Classes

Atorvastatin Base is a synthetic lipid-lowering agent. Atorvastatin competitively inhibits hepatic hydroxymethyl-glutaryl coenzyme A (HMG-CoA) reductase, the enzyme which catalyzes the conversion of HMG-CoA to mevalonate, a key step in cholesterol synthesis. Atorvastatin also increases the number of LDL receptors on hepatic cell surfaces to enhance uptake and catabolism of LDL and reduces LDL production and the number of LDL particles. This agent lowers plasma cholesterol and lipoprotein levels and modulates immune responses by suppressing MHC II (major histocompatibility complex II) on interferon gamma-stimulated, antigen-presenting cells such as human vascular endothelial cells. (NCI04)

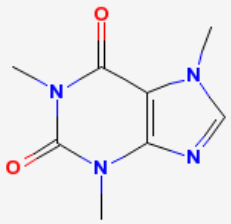
Pharmacology from NCIt

http://fortune.com/2016/03/25/new-blockbuster-drugs-to-watch/

Atorvastatin, Like all statins, atorvastatin works by inhibiting HMG-CoA reductase, an enzyme found in liver tissue that plays a key role in production of cholesterol in the body.

Caffeine on Pubchem

MW: 194.1906 | MF: C8H10N4O2
IUPAC Name: 1,3,7-trimethylpurine-2,6-dione
Canonical SMILES: CN1C=NC2=C1C(=O)N(C(=O)N2C)C
InChI: 1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3



A methylxanthine naturally occurring in some beverages and also used as a pharmacological agent. Caffeine's most notable pharmacological effect is as a central nervous system stimulant, increasing alertness and producing agitation. It also relaxes smooth muscle, stimulates cardiac muscle, stimulates diuresis, and appears to be useful in the treatment of some types of headache. Several cellular actions of caffeine have been observed, but it is not entirely clear how each contributes to its pharmacological profile. Among the most important are inhibition of cyclic nucleotide phosphodiesterases, antagonism of adenosine receptors, and modulation of intracellular calcium handling.

SMILES™

Simplified Molecular Input Line Entry System

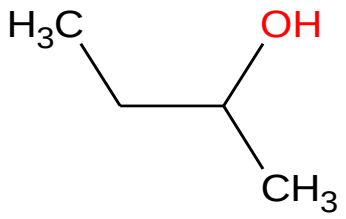
Some simple SMILES™ examples:

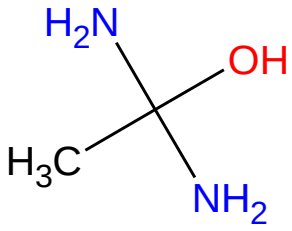
Ethanol	CCO
Acetic acid	CC(=O)O
Cyclohexane	C1CCCCC1
Pyridine	c1cnccc1
Trans-2-butene	C/C=C/C
L-alanine	N[C@@H](C)C(=O)O
Sodium chloride	[Na+].[Cl-]
Displacement reaction	C=CCBr>>C=CCI

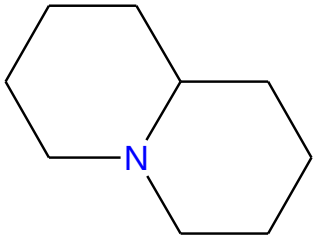
Canonical SMILES: CN1C=NC2=C1C(=O)N(C(=O)N2C)C

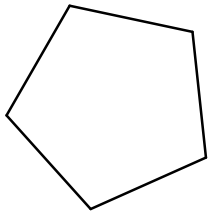
http://www.daylight.com/smiles/

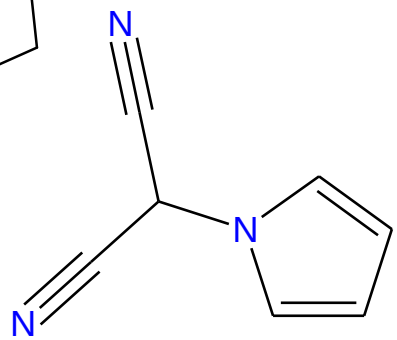
SMILES Basics











- Branching: Parentheses
- Cycles: Numerical annotations
  - CCC(O)C
  - CC(N)(N)O
  - C1CCCC1
  - N12CCCCC1CCCC2
  - N#CC(C#N)N1C=CC=C1
- Extensions for
  - Inorganic atoms, unusual valence, formal charges, stereochemistry, aromaticity, reactions, etc.

InChI: InChI=1/C8H10N4O2/c1-10-4-9-6 5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

INCHI

Code IUPAC International Chemical Identifier (InChI) pour coder de manière unique les molécules, basé sur les tables de connexion, de façon canonique et compactée. Ce code prend en compte la stéréochimie sp2 et sp3, les isotopes, et la tautomérie.

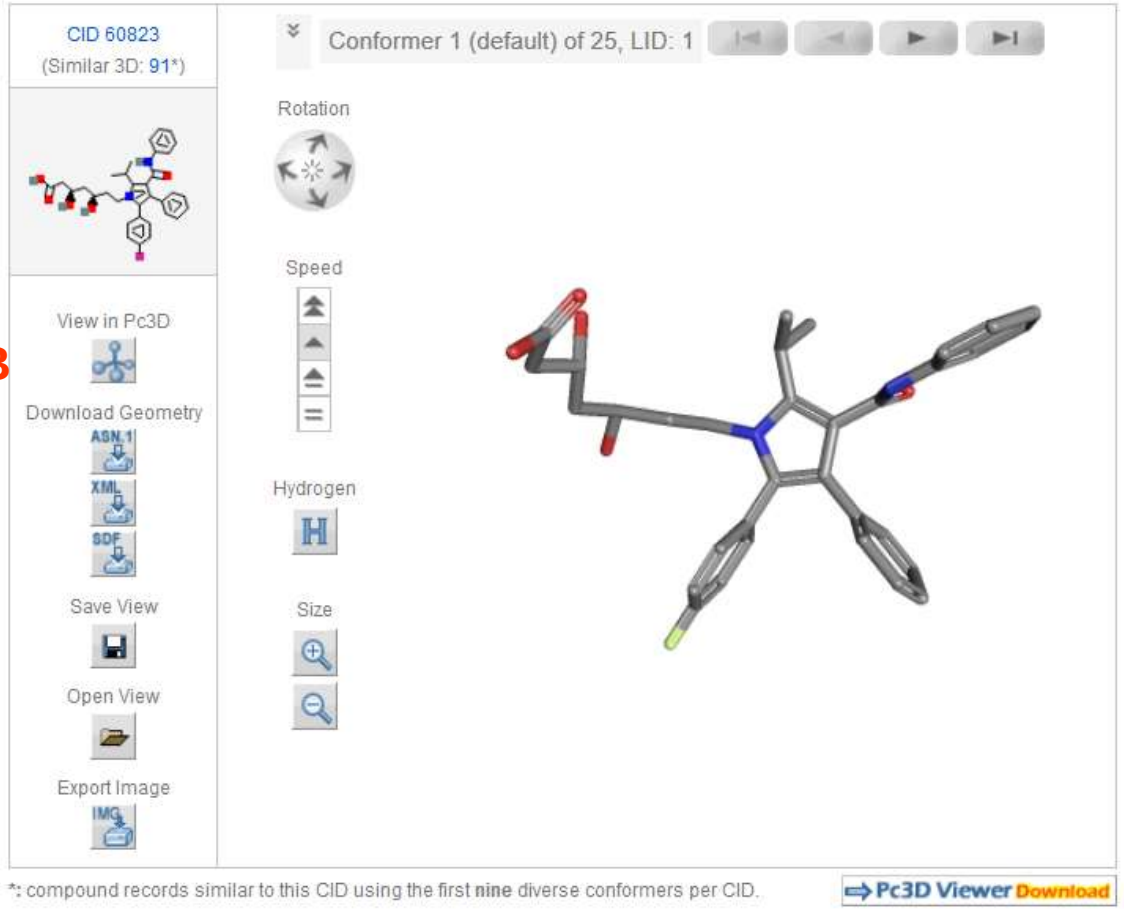
Un codage unique par SMILES était dépendant du logiciel utilisé, InChI est généré par un algorithme gratuit.

Ce code possède un avantage par rapport au numéro CAS : il ne nécessite pas la centralisation des données.

Files of molecules
in molecular modelling programs
and in exchanges

Files of 2D/3D molecules in molecular modelling programs and in exchanges

.SDF .MOL .MOL2 .PDB



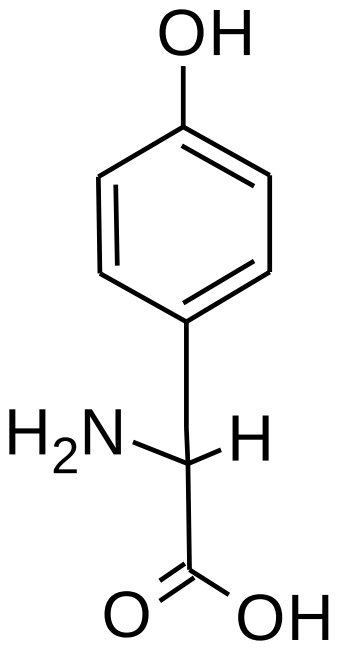
https://pubchem.ncbi.nlm.nih.gov/vw3d/vw3d.cgi?cmd=crtvw&reqid=3273753637728925955

tyrosine.mol

OMFMANAGER09240711423D 0.00000

```
 14 14  0  0  0  0  0  0  0  0  1 V2000
  -0.0167    1.3770    0.0096 C   0  0  0  0  0  0
   0.0021   -0.0041    0.0020 C   0  0  0  0  0  0
   1.2082   -0.6803   -0.0131 C   0  0  0  0  0  0
   2.3980    0.0237   -0.0207 C   0  0  0  0  0  0
   2.3846    1.4052   -0.0127 C   0  0  0  0  0  0
   1.1761    2.0859    0.0020 C   0  0  0  0  0  0
   1.1603    3.4448    0.0095 O   0  0  0  0  0  0
   1.2257   -2.1872   -0.0209 C   0  0  0  0  0  0
   1.2452   -2.7044    1.4189 C   0  0  1  0  0  0
   0.0466   -2.2323    2.1250 N   0  0  0  0  0  0
   1.2627   -4.2113    1.4111 C   0  0  0  0  0  0
   2.2716   -4.8677    0.8171 O   0  0  0  0  0  0
   0.3685   -4.8286    1.9396 O   0  0  0  0  0  0
   2.1356   -2.3333    1.9263 H   0  0  0  0  0  0
  1  2  2  0  0  0  0
  6  7  1  0  0  0  0
  3  4  2  0  0  0  0
  3  8  1  0  0  0  0
  4  5  1  0  0  0  0
  9 10  1  0  0  0  0
  2  3  1  0  0  0  0
  9 11  1  0  0  0  0
  5  6  2  0  0  0  0
 11 12  1  0  0  0  0
  6  1  1  0  0  0  0
 11 13  2  0  0  0  0
  8  9  1  0  0  0  0
  9 14  1  0  0  0  0
M  END
$$$$
```

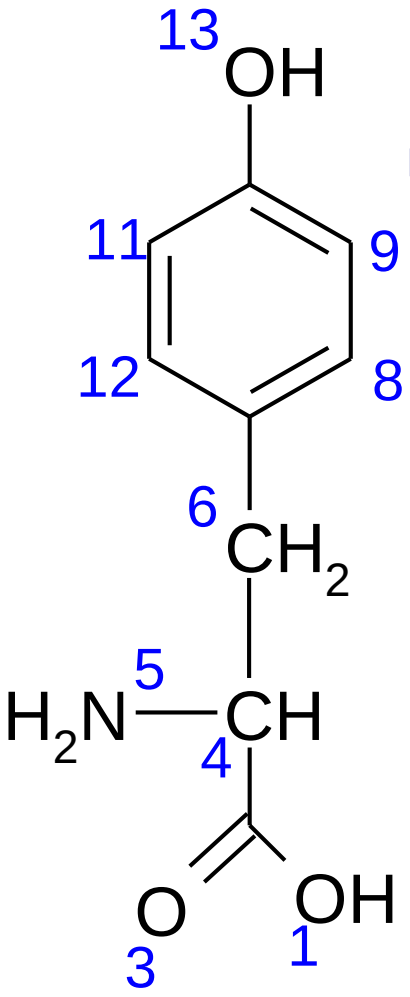
3D



Notion de Multi fichiers

MDL Connection Table

.mol, .mol2…



- Header Block
  - data on molecule name and file origin
  - counts of atoms and bonds etc.

```
Tyrosine
  -ISIS-  08220120432D

 13 13  0  0  0  0  0  0  0  0999 V2000
```

MDL Connection Table

- Atoms block
  - one line per atom
  - specifies X,Y,Z-coords, atom symbol, isotope, charge, stereo code etc.

```
    0.2459   -1.4736    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.5815   -1.4724    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.9944   -2.1872    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.5810   -2.9037    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.2495   -2.9008    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.6586   -2.1854    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4836   -2.1830    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.9042   -2.1792    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.1027   -2.1870    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
   -3.1359   -1.1516    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -3.9070   -2.1847    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -4.4070   -2.6845    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -4.4989   -1.5618    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
```

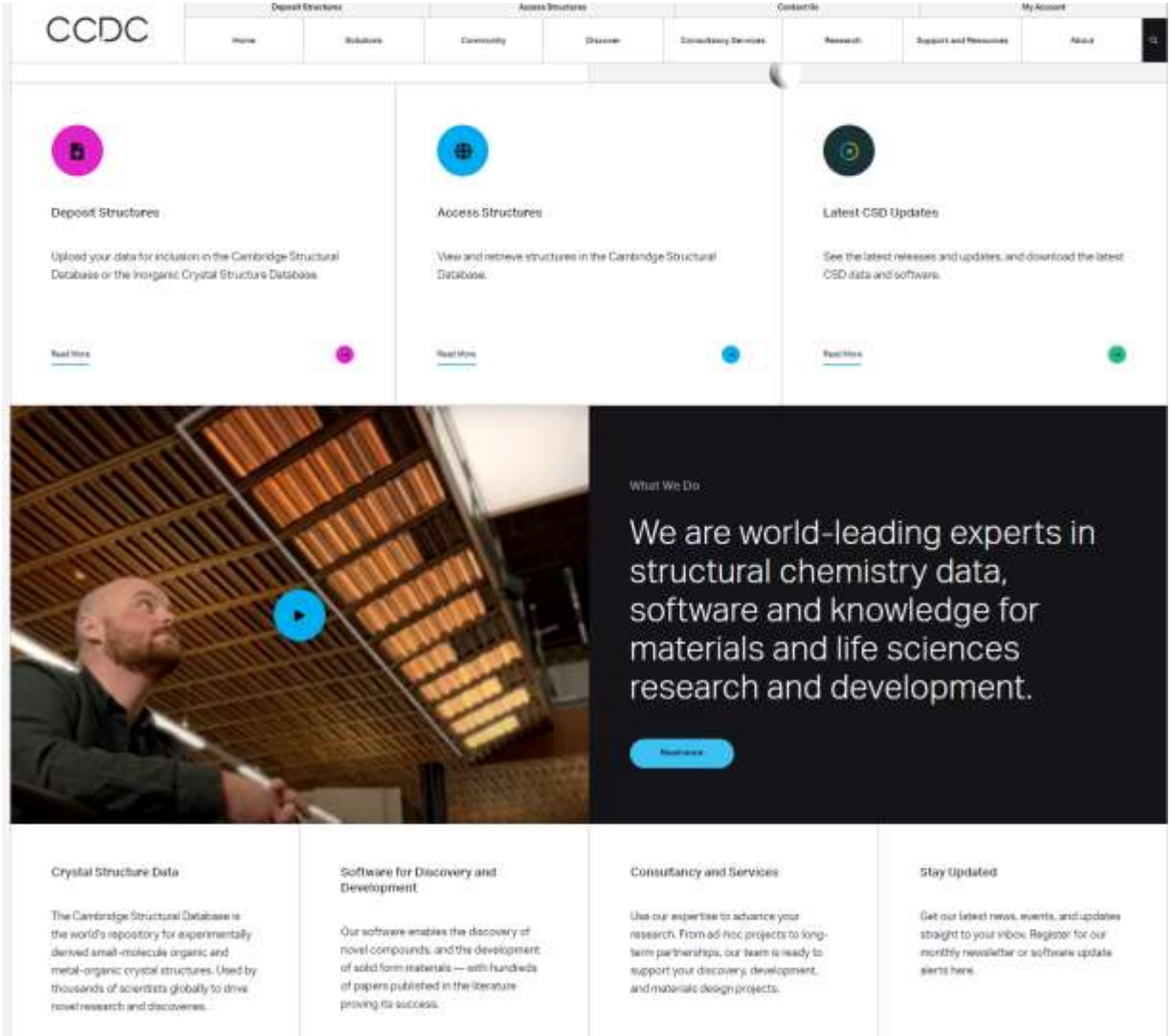
Stéréochimie inconnue

MDL Connection Table

- Bonds Block
  - one line per bond (each bond shown once)
  - specifies row numbers for atoms, and codes for bond type, bond stereochemistry etc.

```
  1  2  2  0  0  0  0
  6  7  1  0  0  0  0
  3  4  2  0  0  0  0
  3  8  1  0  0  0  0
  4  5  1  0  0  0  0
  9 10  1  0  0  0  0
  2  3  1  0  0  0  0
  9 11  1  0  0  0  0
  5  6  2  0  0  0  0
 11 12  1  0  0  0  0
  6  1  1  0  0  0  0
 11 13  2  0  0  0  0
  8  9  1  0  0  0  0
M  END
```

3D DATABASE



http://www.ccdc.cam.a

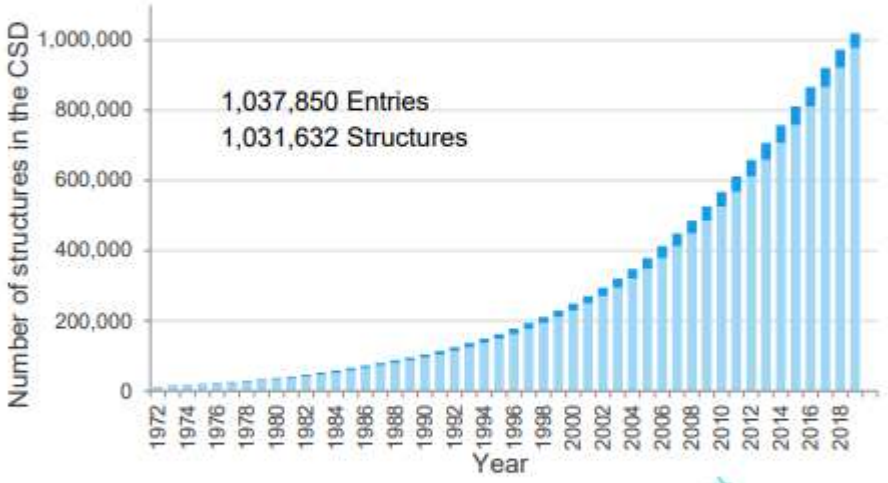
The Cambridge Structural Database (CSD) in Numbers

Yu Gan, Suzanna Ward, Matthew Lightfoot, Clare Tovee, Philip Andrews
The Database Team, The Cambridge Crystallographic Data Centre (CCDC)

Last updated: December 2019 for AsCA2019

DATABASE

A database of organic and metal-organic experimental structures



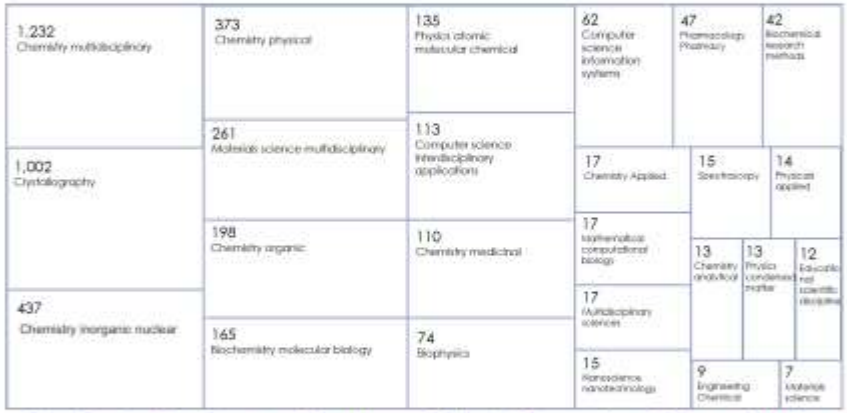
STRUCTURES

- 940,410 Refcode families
- 11,054 Polymorph families
- 169,968 Melting points
- 865,982 Crystal colours
- 734,175 Crystal habits
- 23,871 Bioactivity data

56: Highest Z' value OGUROZ

SCIENCE

The diagram below shows a breakdown of the fields of research citing the latest CSD paper.



The Cambridge Structural Database. Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, Suzanna C. Ward, *Acta Crystallogr. ,Sect. B:Struct. Sci. ,Cryst. Eng. and Mat.*, 2016, **72**, 171-179, DOI: 10.1107/S2052520616003954

PEOPLE AND PLACES

- > 10,000 Different depositors a year
- 402,180 Unique authors in the CSD
- 498,004 Publications in the CSD

2019 Global leaderboard

China
United States
Germany

Molecule of the Month: Histone Deacetylases

Histone deacetylases regulate access to genetic information by modifying histones

This article was written and illustrated by Jessica Damanski, Paloma Munguia Salazar, Mihika Shah and Rajiv Shape as part of a seven-long boot camp for undergraduate and graduate students hosted by the Rutgers Institute for Quantitative Biomedicine. The article is presented as part of the 2023-2024 PDB-101 health focus on "Cancer Biology and Therapeutics."

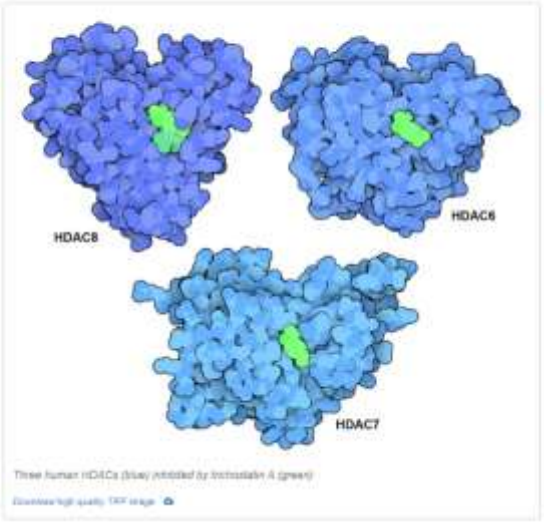
The human genome contains much of the information needed for every cell in the body to function. However, different types of cells often need different types of information. For example, neurons need to build voltage-gated ion channels and muscle cells need to build lots of actin, myosin and fibrin. Cells get their unique characteristics by selectively expressing the genes that they need. Access to DNA is controlled in part by how tightly it is wrapped around histones in nucleosomes. The affinity of histones towards DNA is tuned by the addition and removal of charged groups such as acetyl groups on histone tails, determining where transcription factors can bind and begin expression of particular genes.

The Deacetylation Supervisor

Histone deacetylases (HDACs), typically found in the nucleus, catalyze removal of acetyl groups from histone tails. There are two structurally distinctive families of these enzymes: the histone deacetylase family shown here depends on a zinc cofactor, whereas sirtuins use a NAD cofactor to catalyze the chemical reaction. When the acetyl group is removed by an HDAC, histones become more positively charged and wrap around the negatively-charged DNA more tightly. This gives transcription factors less access to the DNA and represses nearby gene expression. Acetylation and HDACs are also used to modulate the function of many other proteins, such as p53 tumor suppressor and microtubules, and even charged molecules such as polyamides.

Targeting HDAC

HDACs play a role in the development and progression of several different types of cancer by affecting transcription of oncogenes and tumor suppressor genes. Typically, high levels of HDACs correlate with poor outcomes in cancer patients, so HDAC inhibitors are attractive candidates for use as antitumor drugs. A bacterial molecule, trichostatin A, provided a place to start for development of HDAC-blocking drugs. It binds in the active sites of many HDACs, as shown in the structures of HDAC8, HDAC7, and HDAC6 (PDB ID 1t64, 3c10, and 5edu). Several anticancer drugs were developed that improve the action of trichostatin A, and are currently in use to treat patients.



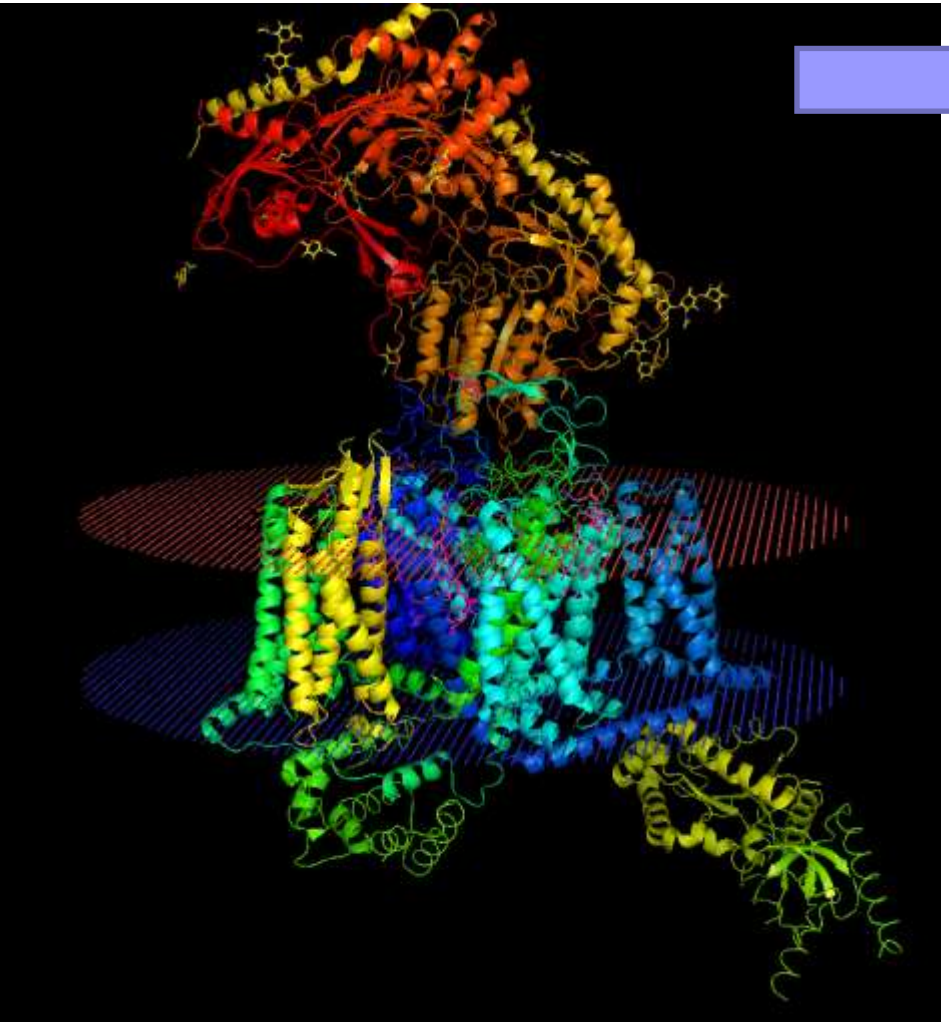
Three human HDACs (blue) inhibited by trichostatin A (green)

Download high quality TIFF image

http://www.rcsb.org/

Dr Stefano Trapani

Software to visualize 3D



Registration For Educational-Use-Only PyMOL Builds

Schrödinger offers **Educational-use-only** PyMOL builds available at no cost to **teachers and high school and college students** (including online courses, homeschooling, etc.) for classroom instruction, homework assignments, and to provide a means for creating high quality figures. Please note that it is not provided for the purposes of academic research or publication.

-> FAQ (Frequently Asked Questions)

The Educational-use-only PyMOL builds are provided "AS IS" with no obligation to grant download access, fix bugs, furnish updates, provide documentation, or meet any other need related to the educational-use PyMOL builds.

If you intend to use PyMOL products for academic research or publication, please purchase an Academic PyMOL subscription, which includes access to technical support, screencasts, and additional resources. See http://pymol.org/academic.

I am a:

Your First Name:

Your Last Name:

Your Email Address:

Your Telephone Number:

Institution:

Comments (optional):

Continue

PyMOL 1.7.4 (August 2015)

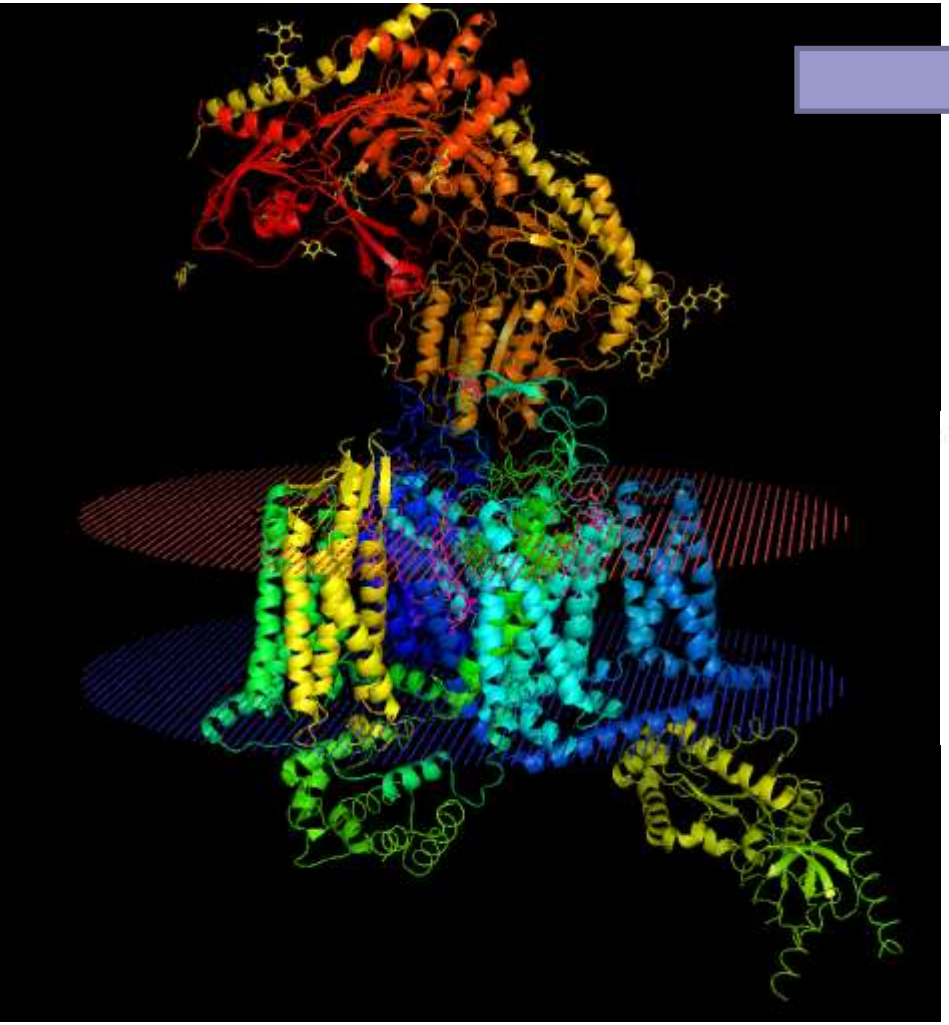
Windows 64bit: EduPyMOL-v1.7.4.5-Win64.msi

Windows 32bit: EduPyMOL-v1.7.4.5-Win32.msi

OS X 10.7+: EduPyMOL-v1.7.4.5r1.dmg

Linux 64bit: EduPyMOL-v1.7.4.5-Linux-x86_64.tar.bz2

Need a license if you want to use in publication



NIH CENTER FOR MACROMOLECULAR MODELING & BIOINFORMATICS | UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

THEORETICAL *and* COMPUTATIONAL BIOPHYSICS GROUP

Home | Research | Publications | Software | Instruction | News | Galleries

Home

Overview

VMD Visual Molecular Dynamics

Version 1.9.3 (2016-11-30) Platforms:

We recommend that all users upgrade to VMD 1.9.3

- Source Code
- LINUX_64 OpenGL, CUDA, OptiX, OSPRay (Linux (RHEL 6.7 and later) 64-bit Intel/AMD x86_64 SSE, with CUDA 8.x, OptiX, OSPRay)
- LINUX_64 Text-mode w/ EGL (Linux (RHEL 6.7 and later) 64-bit Intel/AMD x86_64 w/ SSE, Text-mode w/ EGL)
- LINUX_64 Text-mode (Linux (RHEL 6.7 and later) 64-bit Intel/AMD x86_64 w/ SSE, Text-mode)
- LINUX MIC-AVX512 Text-mode (Linux (RHEL 6.7 and later) 64-bit Intel Xeon Phi MIC w/ AVX-512, Text-mode, OSPRay)
- LINUX MIC-AVX512, OpenGL, CUDA, OptiX, OSPRay (Linux (RHEL 6.7 and later) 64-bit Intel Xeon Phi MIC w/ AVX-512, OpenGL, CUDA7.5, OptiX, OSPRay)
- LINUX OpenPOWER Text-mode (Linux 64-bit IBM OpenPOWER w/ VSX, Text-mode)
- MacOS X OpenGL (32-bit Intel x86) (Apple MacOS-X (10.4.7 or later) with hardware OpenGL (native bundle))
- Windows OpenGL, CUDA (Windows XP/Vista/7/8/10 (32-bit) with OpenGL and CUDA)
- Windows OpenGL (Microsoft Windows XP/Vista/7/8/10 (32-bit) using OpenGL)
- NCSA Blue Waters (Cray XK7 w/ OpenGL) (NCSA Blue Waters (Cray XK7) MPI, CUDA, OpenGL Pbuffers, TachyonL-OptiX)
- ORNL Titan (Cray XK7) (ORNL Titan (Cray XK7) MPI, CUDA, TachyonL-OptiX)
- CSCS Piz Daint (Cray XC50 w/ EGL) (CSCS Piz Daint (Cray XC50) MPI, CUDA, EGL Pbuffers, TachyonL-OptiX)

The overall structure is approximately 170 Å in height
and 100 Å in the longest dimension of the width

UCSF ChimeraX

UCSF ChimeraX (or simply ChimeraX) is the next-generation molecular visualization program from the Resource for Biocomputing, Visualization, and Informatics (RBVI), following UCSF Chimera. ChimeraX can be downloaded free of charge for academic, government, nonprofit, and personal use. Commercial users, please see ChimeraX commercial licensing.

ChimeraX is developed with support from National Institutes of Health R01-GM129325, Chan Zuckerberg Initiative grant EOSS4-0000000439, and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Feature Highlight

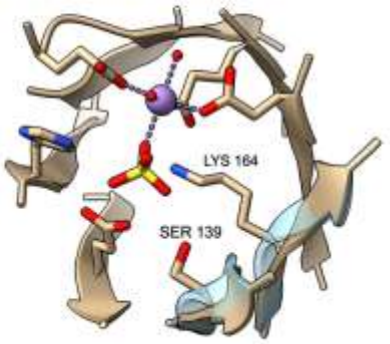
Adjustable β-Strand Smoothing

Cartoon ribbons are drawn to pass through the exact positions of peptide α-carbons in helix and coil, but by default, strands are smoothed to appear less ripply. This means that α-carbons in strands may not fall on the ribbon. A tether is drawn wherever a sidechain is displayed but would otherwise be detached from the ribbon. (Tether appearance can be adjusted with **cartoon tether**.)

However, β-strand ribbon smoothing can be tuned continuously between zero (off) and 1.0 (default, fully on), and this can be done for all residues or for specific residues only. The image compares ribbon positions for residues 139 and 164 in the active site of mandelate racemase (PDB 2mnr) with and without smoothing. The unsmoothed position (transparent blue) can be obtained with:

cartoon :139,164 smooth 0

This leaves ribbon smoothing at other residues unchanged. Values between 0 and 1 would give intermediate positions.



More features...

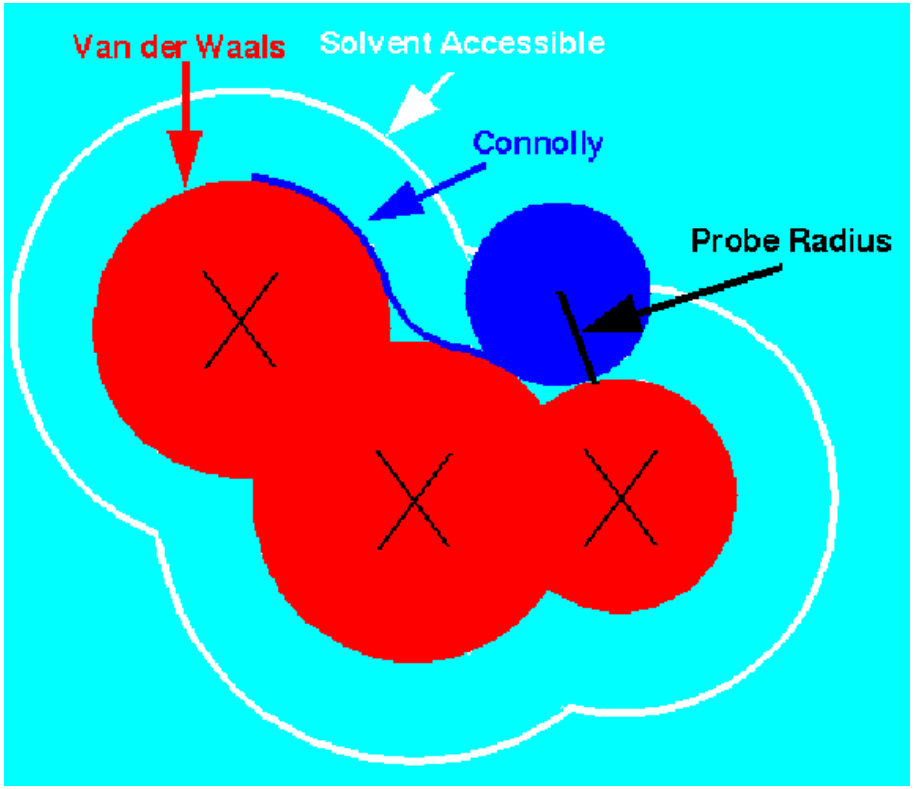
Example Image

Calmodulin and Target Peptide

Calmodulin (CaM) acts as a calcium sensor. When its four Ca++ sites are fully occupied, it binds and modulates the activity of various downstream proteins, including CaM-dependent protein kinase I (CaMKI). Here, a complex between CaM and its target peptide from CaMKI (PDB 1mxe) is shown with cartoons, a transparent molecular surface, silhouette outlines, and light soft ambient occlusion. (If you prefer a less smudgy/rustic appearance, try using light gentle instead.) For image setup other than positioning, see the command file cam.cxc.

More images...

Importance de la surface



La surface d'une molécule représentée par des boules de van der Waals est constituée par la frontière de l'ensemble de ces boules (*a*) : c'est la surface de van der Waals. La surface accessible par le solvant est l'ensemble des positions possibles du centre d'une boule représentant le solvant, qui roulerait sur la surface de van der Waals (*b*). La surface de Connolly prend en compte le recouvrement des creux par la boule représentant le solvant (*c*).
Infographie : Pour la Science

Image sur http://www.chem.leeds.ac.uk

Domains of Chemoinformatics

Chemoinformatics is an interdisciplinary field that combines chemistry, computer science, and information technology to analyze and model chemical data.

1. Chemical Structure Databases, visualization, properties, searching and retrieving :

- Creating 2D and 3D visual representations of molecules : ChemSketch…
- Managing vast databases of chemical structures : PubChem, ChemSpider. PDB, CCSD.
- Representation and research structures and substructures : Molecular Graphes.
- Similarity search (2D / 3D), clustering and diversity analysis : Tanimoto, clustering.
- Search chemical molecules, patent databases or chemical reactions
- QSAR (Quantitative Structure-Activity Relationship) modeling.

2. Molecular Modeling ant Interactions :

- Predicting molecular structures and properties : homology modeling, A.I.
- Molecular dynamics simulations.
- Molecular docking simulations.
- Pharmacophore modeling.

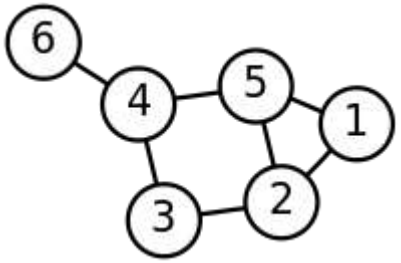
Théorie des graphes moléculaires

Théorie des graphes (wikipedia)

Le terme de **graphe** désigne en mathématiques une opération d'application.

•le graphe d'une fonction (à distinguer de sa *représentation graphique*)

•un objet représentant une relation binaire, orientée ou non, entre des éléments d'un ensemble (dans le cas de plusieurs relations entre éléments, on parle d'hypergraphe).



Un exemple de graphe non orienté avec 6 sommets et 7 arêtes

Chemical Structure Representation and Search Systems

Lecture 4. Nov 11, 2003

John Barnard

Barnard Chemical Information Ltd

Chemical Informatics Software & Consultancy Services

Sheffield, UK

pdf

J Cheminform. 2012; 4: 13.

Published online 2012 Jul 31. doi: 10.1186/1758-2946-4-13

PMCID: PMC3586954

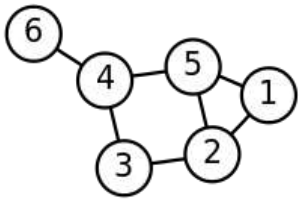
Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2

Hans-Christian Ehrlich[1] and Matthias Rarey[1]

Author information ► Article notes ► Copyright and License information ►

Chemical Structure Representation and Search Systems

- full structure search
  - query is is complete molecule
  - is this molecule in the database?
    or tautomers, stereoisomers etc. of it,
- substructure search
  - query is a pattern of atoms and bonds
  - does this pattern occur as a substructure of any of the molecules in my database?



Graph isomorphism is an important concept in graph theory and chemistry, particularly when studying molecular graphs.

Molecular Graph: A molecular graph is a representation of a chemical molecule in the form of a graph.
Atoms are represented by vertices (nodes), and chemical bonds between atoms are represented by edges (edges) connecting the vertices.

Graph Isomorphism: Graph isomorphism is a relation between two graphs, indicating that they are "essentially the same" in terms of their structure. In other words, two graphs are isomorphic if they have the same topology, meaning they can be transformed into each other while preserving the connectivity between vertices and edges, possibly by renaming the vertices. In the context of molecular graphs, graph isomorphism is essential because it allows us to compare different molecules to determine if they share the same underlying structure, even if they may have different types of atoms. Here's how it works:

Suppose you have two molecules, A and B, each represented as a molecular graph. To determine if these two molecules are isomorphic, you can follow these steps:

1. Assign Labels to Vertices: Assign a label (or tag) to each vertex of the graph, indicating the type of atom it represents (e.g., C for carbon, H for hydrogen, O for oxygen, etc.).
2. Compare Graph Structure: Compare the structure of the graphs by examining the bonds between atoms. Ensure that the chemical bonds (edges) between vertices are the same, meaning they connect the same types of atoms and have the same spatial configuration.
3. Verify Isomorphism: If the graphs have the same basic structure, even if the atoms have different labels, then the two molecules are considered isomorphic on a structural level.
It's important to note that graph isomorphism does not consider the specific chemical properties of atoms (such as their charge or electronegativity) but only how they are connected within the molecule.

In summary, graph isomorphism for molecular graphs allows us to determine if two molecules share the same underlying structure while ignoring specific atomic details. This can be useful in chemical research for comparing molecules and understanding their structural similarity.

Adjacency Matrix: A mathematical way to represent a graph is through its adjacency matrix. This matrix is a square matrix where each element A[i][j] indicates whether there is an edge between vertices i and j. If A[i][j] is nonzero, it means there is a connection between vertices i and j. In the context of molecules, this matrix is often binary (0 for absence of a bond, 1 for presence of a bond).

Permutation Invariance: To compare two molecular graphs, you can begin by checking if they share the same fundamental structure, meaning they have the same adjacency matrix. If both graphs have the same adjacency matrix, it implies they have the same connectivity between atoms.

Graph Isomorphism: Formally, two graphs are considered isomorphic if they can be transformed into each other by permuting their vertices. In other words, if you can rearrange the vertices of one graph in such a way that it has the same adjacency matrix as the other graph, then the two graphs are isomorphic.

Mathematically, if you have two graphs G1 and G2 with adjacency matrices A1 and A2, they are isomorphic if there exists a permutation P of the vertices of G1 such that:

$A1[i][j] = A2[P(i)][P(j)]$

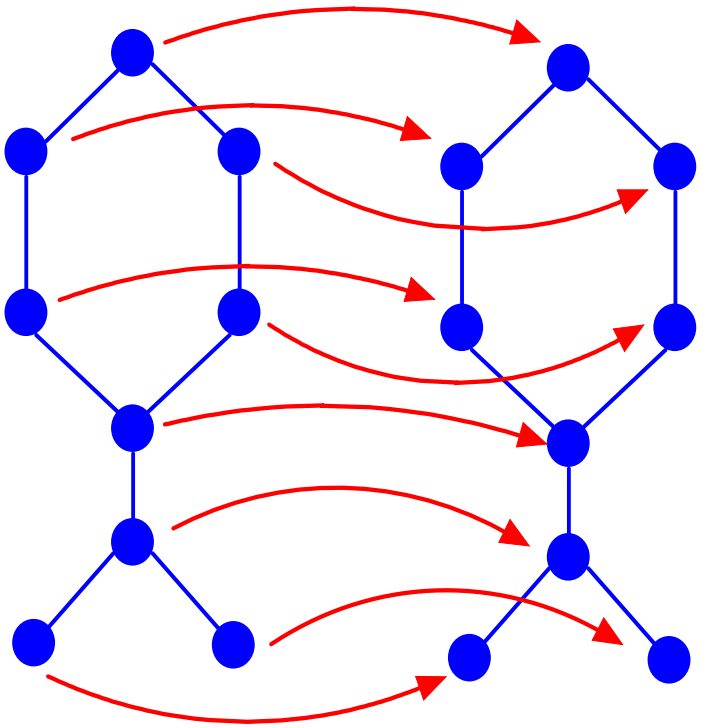
This means that the bond relationships between atoms are preserved, even if the vertices (atoms) are rearranged.

In summary, to compare the structure of two molecular graphs, you can use their adjacency matrices and check for isomorphism by finding a permutation of the vertices that preserves the chemical bonds. If such a permutation exists, then the two graphs are considered isomorphic in terms of structural properties.

Graph Isomorphism

In graph theory terms, when two full structures match, their graphs are said to be *isomorphic*

- each node N_1 in G_1 must be mapped to a node N_2 in G_2
- neighbours of N_1 must map to neighbours of N_2

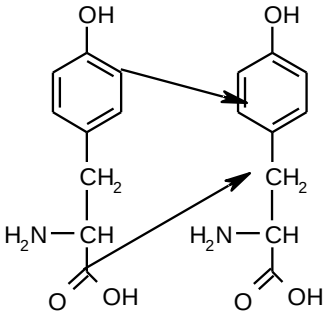


John Barnard

Graph isomorphism by brute force

- for each node in G_1
  - map it against an unmapped node in G_2
- check that neighbours of each node map appropriately in the two graphs

- if each graph has n nodes there are *n*! ways of doing this
  - $n \times (n\text{-}1) \times (n\text{-}2) \times (n\text{-}3) \ldots \times 3 \times 2 \times 1$
  - this is a big number if *n* is anything non-trivial
  - 9! = 362 880
  - 10! = 3 628 800



Some algorithms may have complexity $O(n^3)$, $O(n^4)$, $O(\log n)$, $O(n \log n)$ etc.
these are all "polynomial" time algorithms

Some algorithms have exponential complexity, e.g. $O(2^n)$
this is much slower than polynomial

Brute-force graph isomorphism is $O(n!)$
–this is even slower than simple exponential

non-polynomial - The set or property of problems for which no polynomial-time algorithm is known.
This includes problems for which the only known algorithms require a number of steps which increases exponentially with the size of the problem, and those for which no algorithm at all is known. Within these two there are problems which are "provably difficult" and "provably unsolvable".

polynomial-time algorithm - A known algorithm (or Turing Machine) that is guaranteed to terminate within a number of steps which is a polynomial function of the size of the problem.

polynomial - An arithmetic expression composed by summing multiples of powers of some variable.

John Barnard

NP-complete problems

- **graph isomorphism** is probably NP-complete
- **subgraph isomorphism** is a generalisation of graph isomorphism
  - nodes in G_1 (query structure) must be mapped to subset of nodes in G_2 (database structure)
  - i.e. G_1 is a subgraph G_2
- **subgraph isomorphism** has been proven to be NP-complete
- Much effort has been expended on this problem over the past 40+ years
  - closely-related problems remain an active area of research

Speeding up subgraph isomorphism

How to ?

1. use a faster computer
2. use tricks to avoid exploring potential solutions that are bound to fail
3. do most of the work in a pre-processing of the database structures, independently of the query

Screening using fingerprints

Aim : Remove compounds that are not a solution

Fingerprints are a very abstract representation of certain structural features of a molecule. Unlike a structural key with its pre-defined patterns, the patterns for a molecule's fingerprint are generated from the molecule itself. The fingerprinting algorithm examines the molecule and generates the following:

a pattern for each atom
a pattern representing each atom and its nearest neighbors (plus the bonds that join them)
a pattern representing each group of atoms and bonds connected by paths up to 2 bonds long
... atoms and bonds connected by paths up to 3 bonds long
... continuing, with paths up to 4, 5, 6, and 7 bonds long.

2em solution. **Screening using Structural keys and fingerprints**

Aim : Remove compounds that are not a solution

Fingerprints are a very abstract representation of certain structural features of a molecule.

For example, the molecule **OC=CN** would generate the following patterns:

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

The list of patterns produced is exhaustive: *Every* pattern in the molecule, up to the pathlength limit, is generated. For all practical purposes, the number of patterns one might encounter by this exhaustive search is infinite, but the number produced for any *particular* molecule can be easily handled by a computer.

Screening using fingerprints

Aim : Remove compounds that are not a solution

OC=CN: Fingerprints are a very abstract representation of certain structural features of a molecule. Fingerprints address the lack of generality of Structural keys by eliminating the idea of pre-defined patterns.

Nb C >=6	0	Présence de S	0
Présence de N	1	Présence de C-C	0
Présence de c=O	1	Présence de N-O	0
Présence de O	1	Présence de C=N	1

Screening using fingerprints

Nb C >=6	0	Présence de S	0
Présence de N	1	Présence de C-C	0
Présence de c=O	1	Présence de N-O	0
Présence de O	1	Présence de C=N	1

b) « bit strings » with Dictionary fingerprints

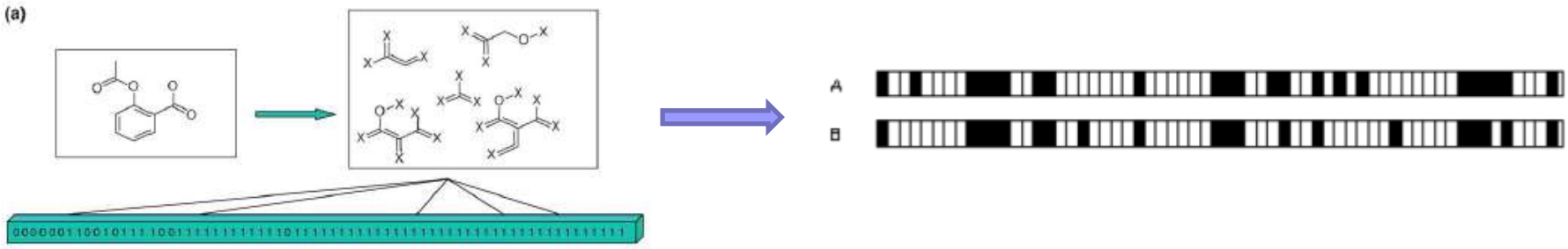
Binairy

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

Continuous

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

http://www.daylight.com/dayhtml/doc/theory/theory.finger.html



Query:	00000100010101000001010011010100	
DB struct 1:	00010100010101000101010011110100	MATCH
DB struct 2:	00000001001001010010010000111000 00	NO MATCH

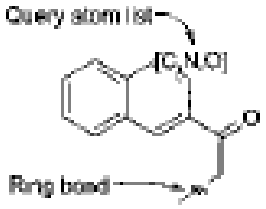
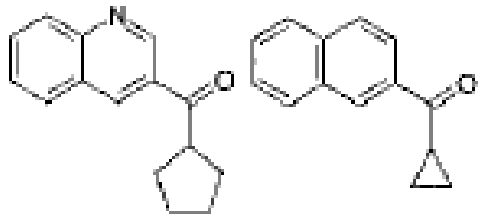
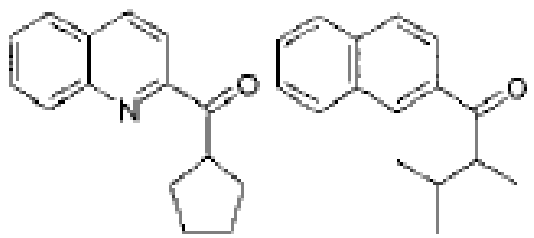
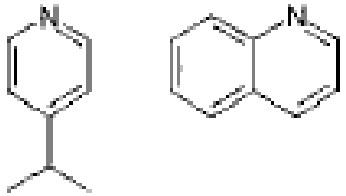
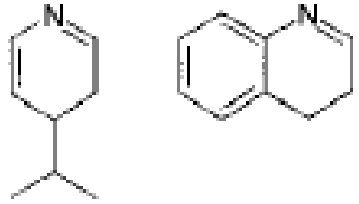
"Ultimately we still have to use a real substructure search to get a 100%-confident" (Chemical information system, Inc.)

That means = at the end, with selected compounds one will perform a Graph Isomorphism.

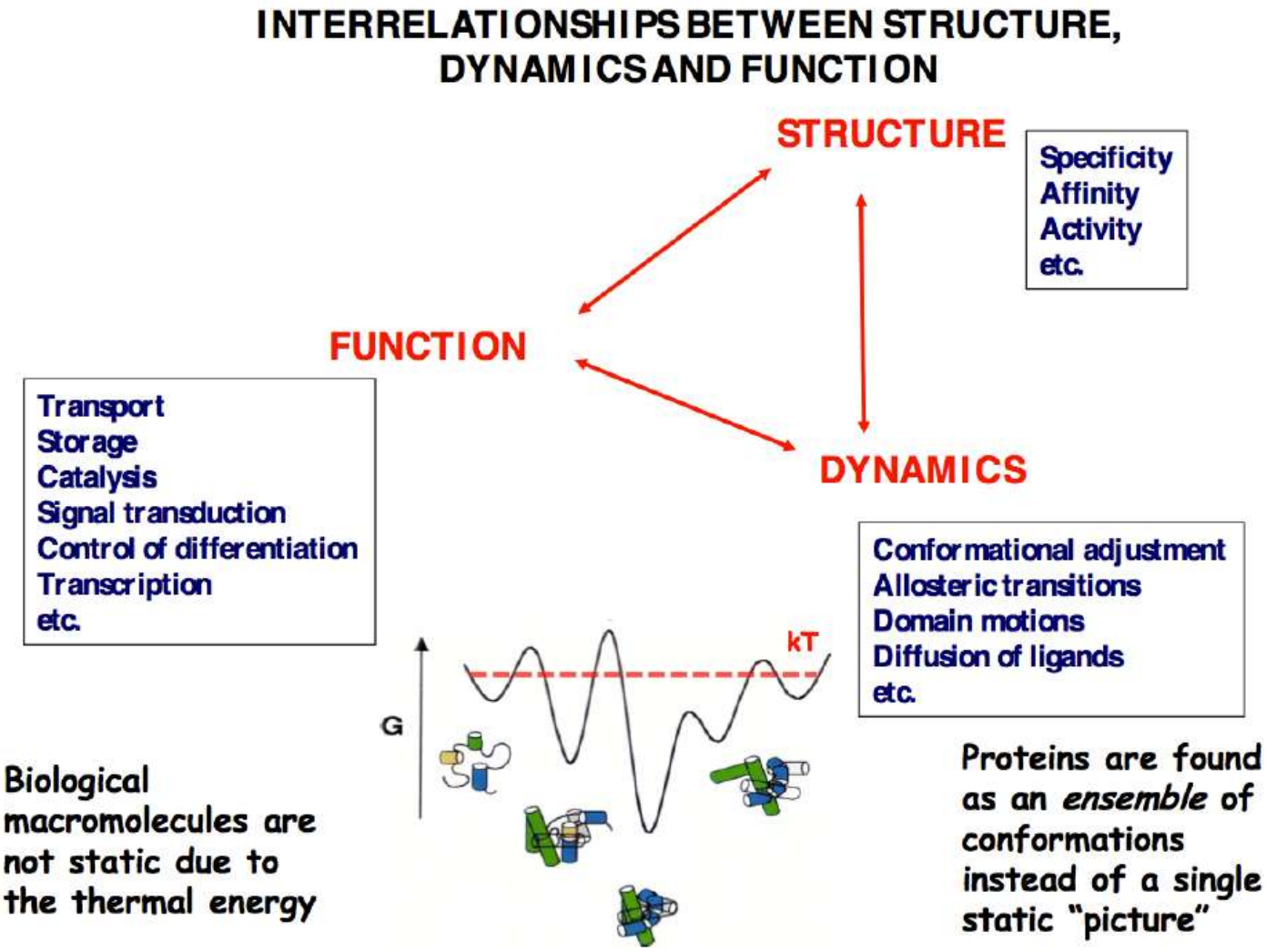
wikipedia

Bit *(acr. angl.) (contraction de Binary Digit)*
C'est l'unité binaire de quantité d'information qui peut représenter deux valeurs distinctes : 0 ou 1. Un champ de 8 bits constituant ce qu'on appelle 1 byte ou 1 octet. Un **mot de longueur l construit sur cet alphabet est une suite finie de l bits.** En anglais on appelle une telle suite finie un *bit string.*

Tab. 1-3 Examples of query atoms, bonds, their hits and non-hits

Query	*Hits*	*Non-hits*
		
		
		

Conformational space of molecules

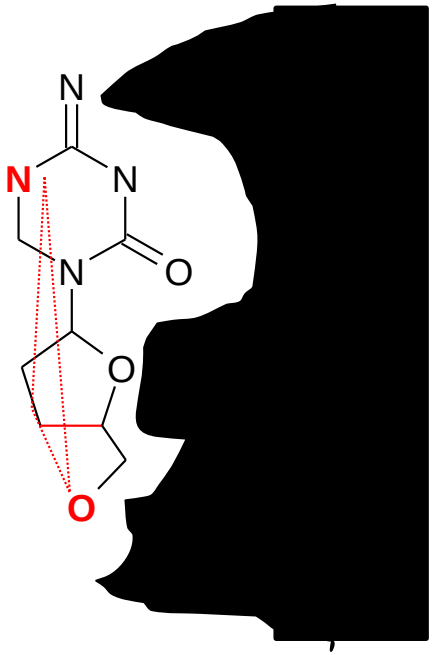


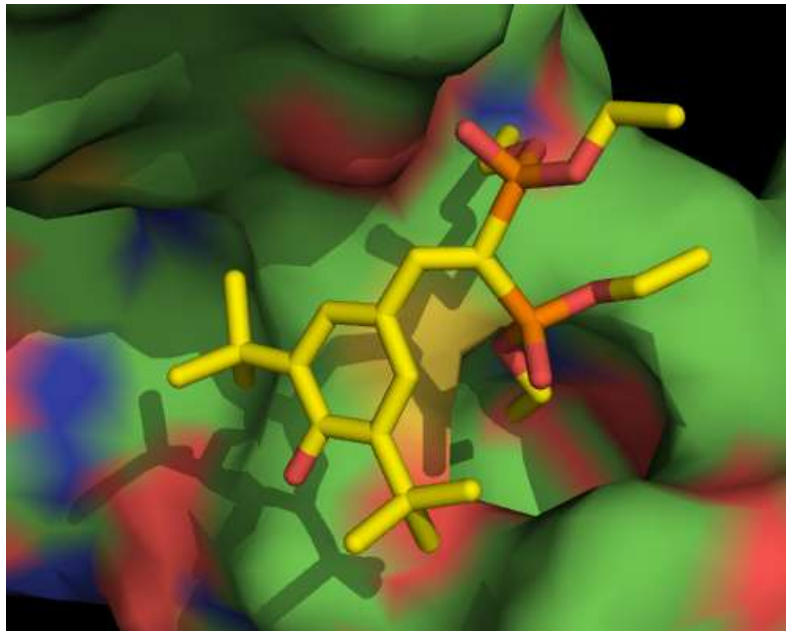
Recherche à partir d'un Pharmacophore

Un **pharmacophore** est l'ensemble des groupements fonctionnels disposés selon un arrangement spatial adéquat, assurant la fixation sur le récepteur et donc capable d'induire la réponse physiologique.

Xtalo, RMN, docking.
Conformation bioactive ligand-protéine

Mesure des distances des groupements nécessaires à l'activité biologique.



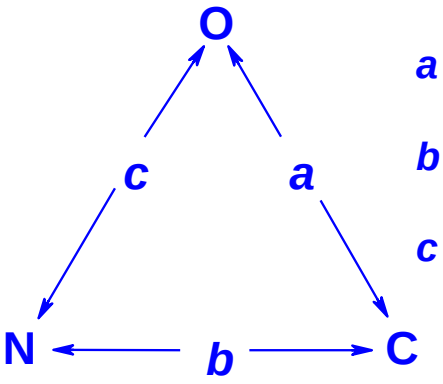


La requête de sous structure en tenant compte de distances.

- Les molécules sont des objets en 3D. Extension des méthodes 2D de recherche de sous structures (isomorphisme de graphe, screening) pour permettre de tenir compte de cette 3D (Pfizer and Lederle, 1987).

- Connaître/stocker les molécules de la database sous forme 3D.
  - Donnée expérimentale (Cambridge Structure Database)
  - Modélisation moléculaire (mécanique quantique, dynamique moléculaire)
  - Passage 2D -> 3D via CONCORD (Texas, 1987) ou CORINA (Munich/Erlangen, 1990)

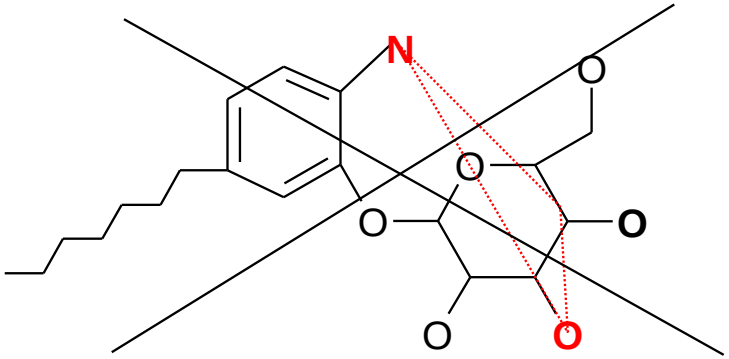
- Par la suite, recherche en tenant compte de la flexibilité (Tripos, 1994).

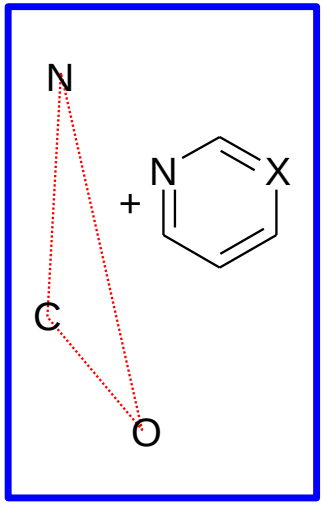

O
c
a
N
b
C

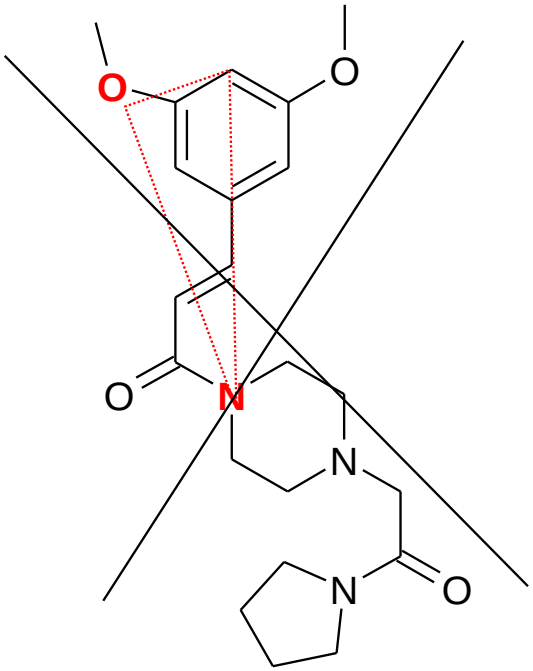
a = 7 A

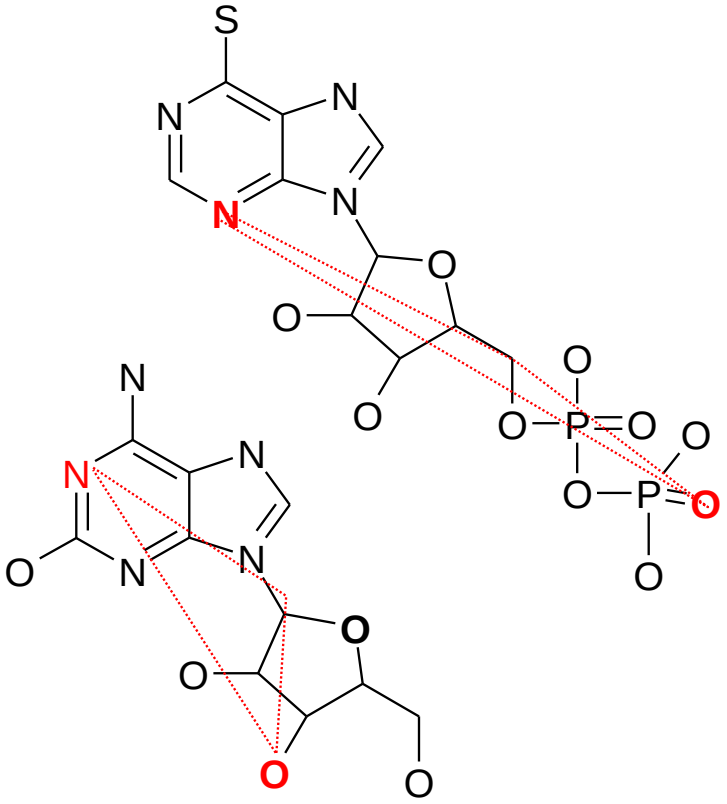
b = 5 A

c = 5 A


N
O
O
O
O
O
O


N
N
X
+
C
O


O
O
O
N
N
N
O

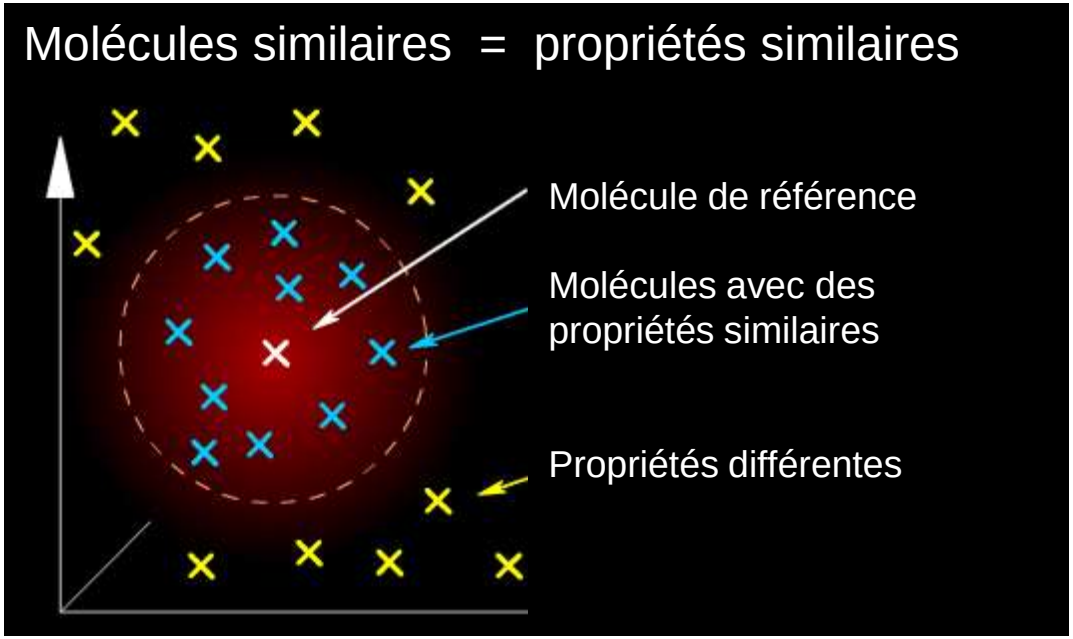

S
N
N
N
N
O
O
O
O
P
O
O
O
P
O
O
N
N
N
N
N
O
O
O
O

Exemple de recherche sur une banque commerciale

Les grands domaines de la chemoinformatique.

- Représentation et recherche de structures et de sous-structures.
- Recherche de similarité (2D/3D), clustering et analyse de diversité.

Similarité : Caractère similaire.
Synonyme ressemblance



Recherche de similarité, pourquoi

- Recherche de composés similaires à une molécule d'intérêt ou entre eux (clustering).

- Évaluation de la diversité d'une Database.

Ces composés seront classés par ordre (rank) de similarité.

- Techniques de criblage virtuel pour trouver des molécules susceptibles de présenter une activité biologique.

Exemple : - Une seule molécule active
- SAR peu précise
- Ne permet pas une recherche par fragment.

Recherche par similarité

Recherche de similarité. Structure ? Activité ? 2D ? 3D ?

Similarité = structure chimique ?

Similarité = Activité ?

Similarité = structure + activité ?

Similarité = même propriétés physico-chimiques à la surface de la molécule (3D) ?

Pour la similarité biologique il faudra tenir compte de la protéine cible

Target structure = molécule d'intérêt avec des spécifications chimiques, biologiques etc…

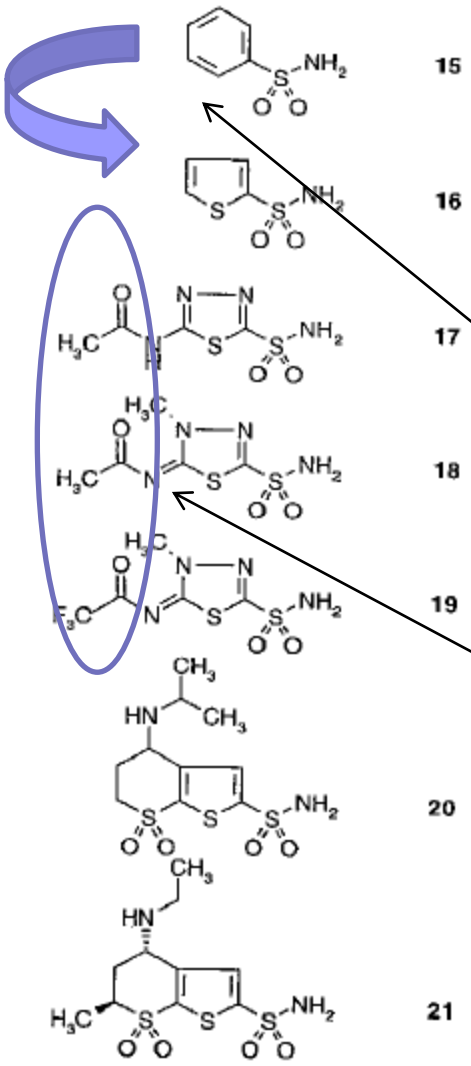


Figure 15. Chemical structures of the carbonic anhydrase inhibitors **15**–**21**. The small aromatic sulfonamides **15** and **16** bind with nanomolar affinity to carbonic anhydrase. Methazolamide **18** was used for a long time to treat glaucoma. **19** was the first topically active inhibitor. Structure-based drug design at Merck first led to **20** and then to the marketed drug, dorzolamide **21**.

Qu'est ce qui définit une ressemblance ?

des fragments communs ?
des volumes, masses ?

Qu'est ce qui détermine une différence ?

Recherche de similarité (2D/3D), comment.

Pour évaluer la similarité entre 2 molécules, il faut :

1) Définir un schéma descriptif commun à toutes les molécules. (molécule d'intérêt et Data base).

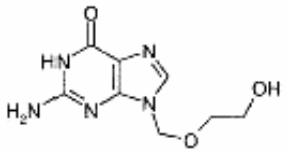
Ce schéma descriptif utilisé pour caractériser les molécules est composé d'éléments comme des sous structures (fragments), des indices de descripteurs topologiques, des éléments en relation avec la 3D…

Ces différents éléments peuvent être pondérés pour moduler certains paramètres, certaines propriétés physico-chimiques…

2) A partir de ces éléments calculer un coefficient de similarité, une distance moléculaire.

Schéma qui doit être adapté en fonction du problème (2D/3D), des outils, des banques.

Les descripteurs moléculaires

	Représentation Type	Descripteurs Type
1D	C8H10N5O3	Masse moléculaire Nombre d'atomes
2D		Fragments Indices Topologiques Connectivité
3D		Surface Moléculaire Volume Moléculaire Energie d'interaction

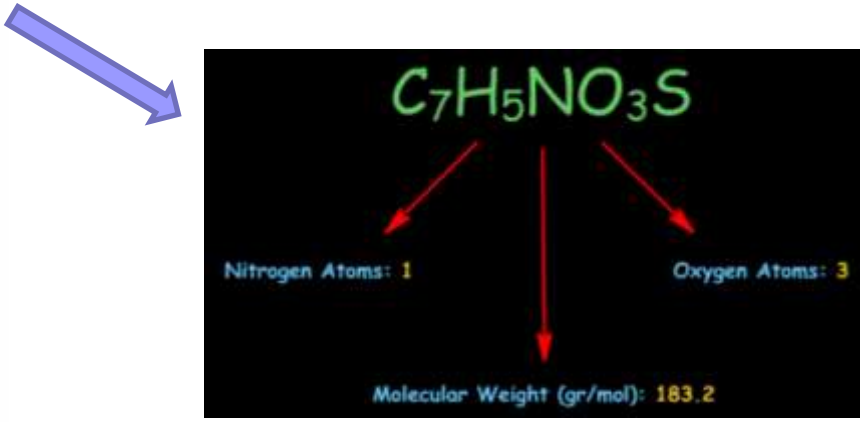


Figure 1: Quelques exemples de descripteurs et leur classification en 1D, 2D et 3D.

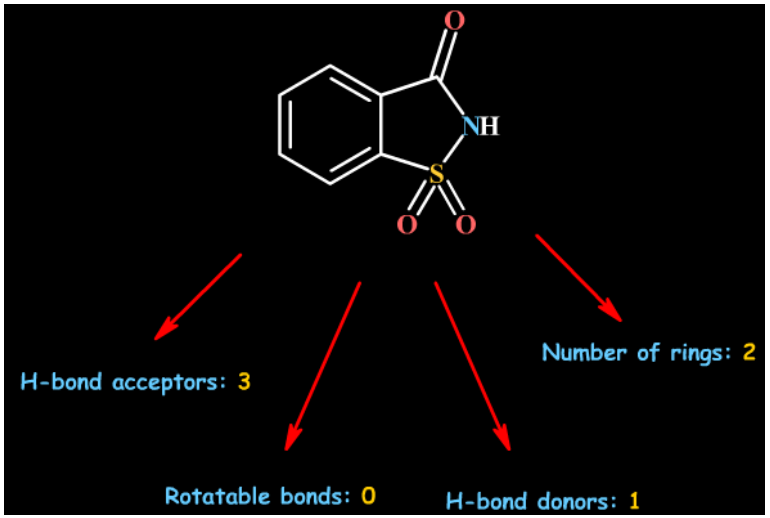
-

Molecular similarity and diversity in chemoinformatics: From theory to applications
Authors: Maldonado, Ana; Doucet, J.; Petitjean, Michel; Fan, Bo-Tao[1]
Source: Molecular Diversity, Volume 10, Number 1, February 2006 , pp. 39-79(41)
Publisher: Springer

Les descripteurs moléculaires

Descripteurs 2D

Les indices

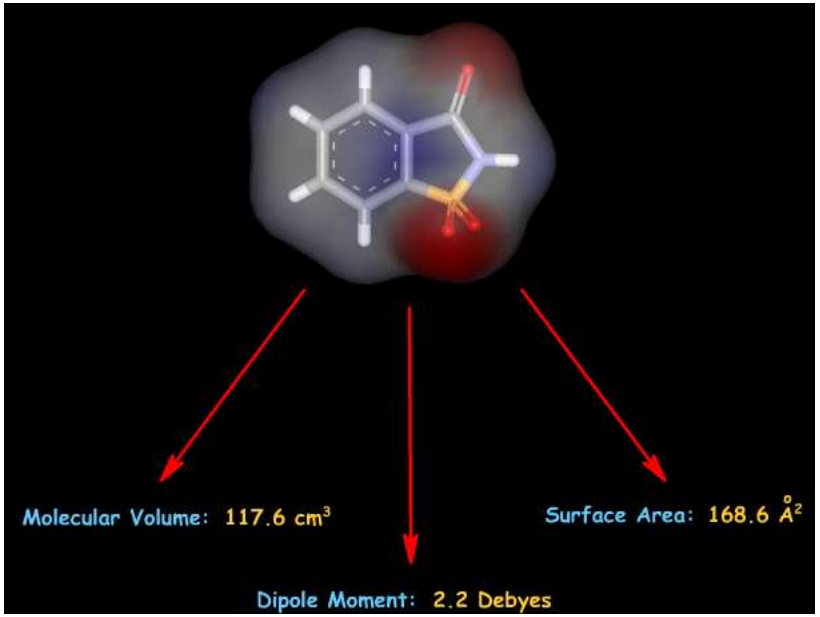


- **constitutionnels:** fc. des différents composants de la molécule (nombre de liaisons, liaisons libres de rotation, doubles, nombres de cycles, fragments pré-définis (voir sous graphes)).
- **topologiques:** Indices de Balaban, de Wiener, de Kier, de Petitjean, de Randic…
- **Propriétés physico-chimiques:** Surfaces de Van der Waals des atomes, clogP, donneur accepteur de liaisons hydrogènes.

Les descripteurs moléculaires

Descripteurs 3D

Les indices

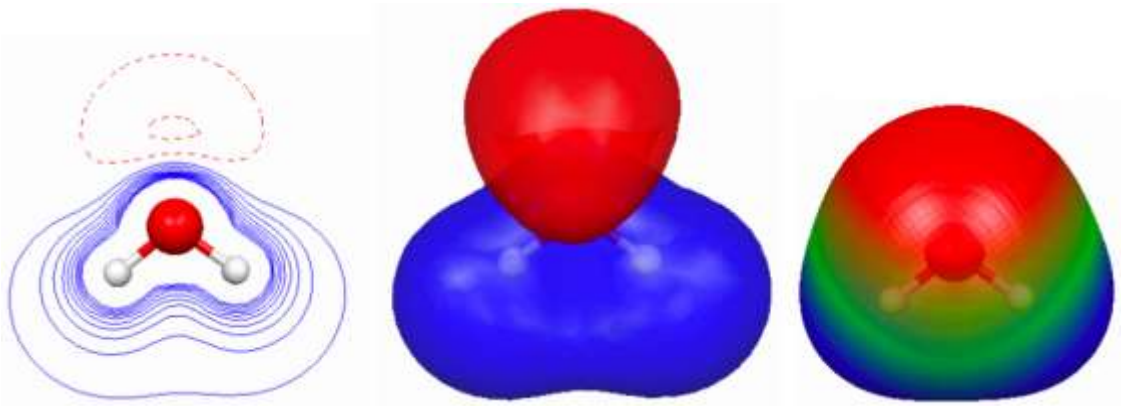


- **géométriques:** Appliqués aux distances géométriques (Wiener ...)
- **Propriétés physico-chimiques:** Fonction de l'agencement spatial des atomes, surface accessible au solvant, le potentiel électrostatique.

Carte ou Surface du Potentiel électrostatique

http://chemwiki.ucdavis.edu/Theoretical_Chemistry/Chemical_Bonding/General_Principles_of_Chemical_Bonding/Electrostatic_Potential_maps

Electrostatic potential maps are very useful three dimensional diagrams of molecules. They enable us to visualize the charge distributions of molecules and charge related properties of molecules

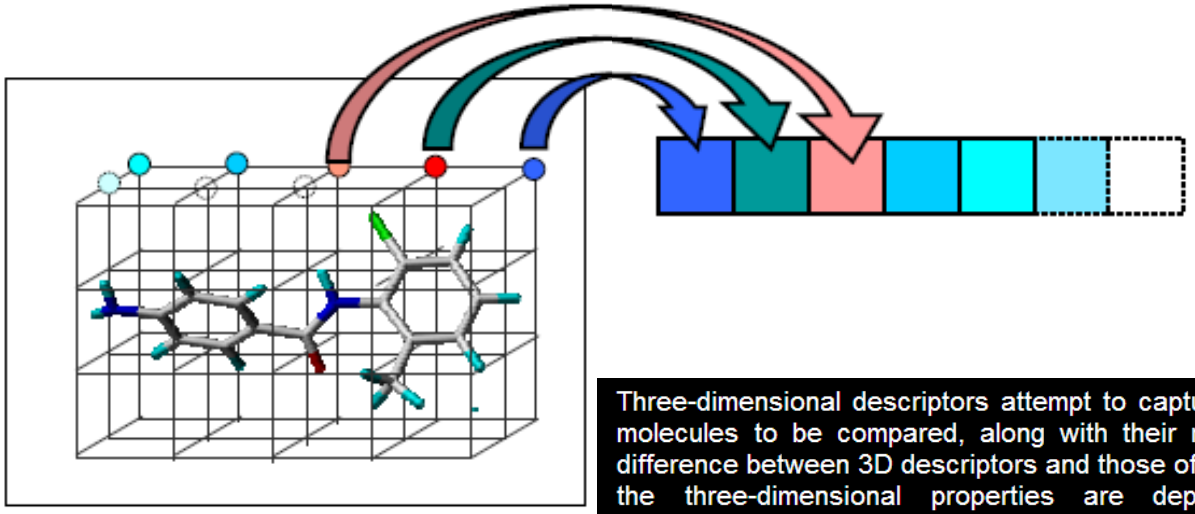


An electric potential (also called the electric field potential or the electrostatic potential) is the amount of electric potential energy that a unitary point electric charge would have if located at any point in space, and is equal to the work done by an electric field in carrying a unit positive charge from infinity to that point.

http://molecularmodelingbasics.blogspot.fr/2009/12/electrostatic-potential.html

Molecular fields

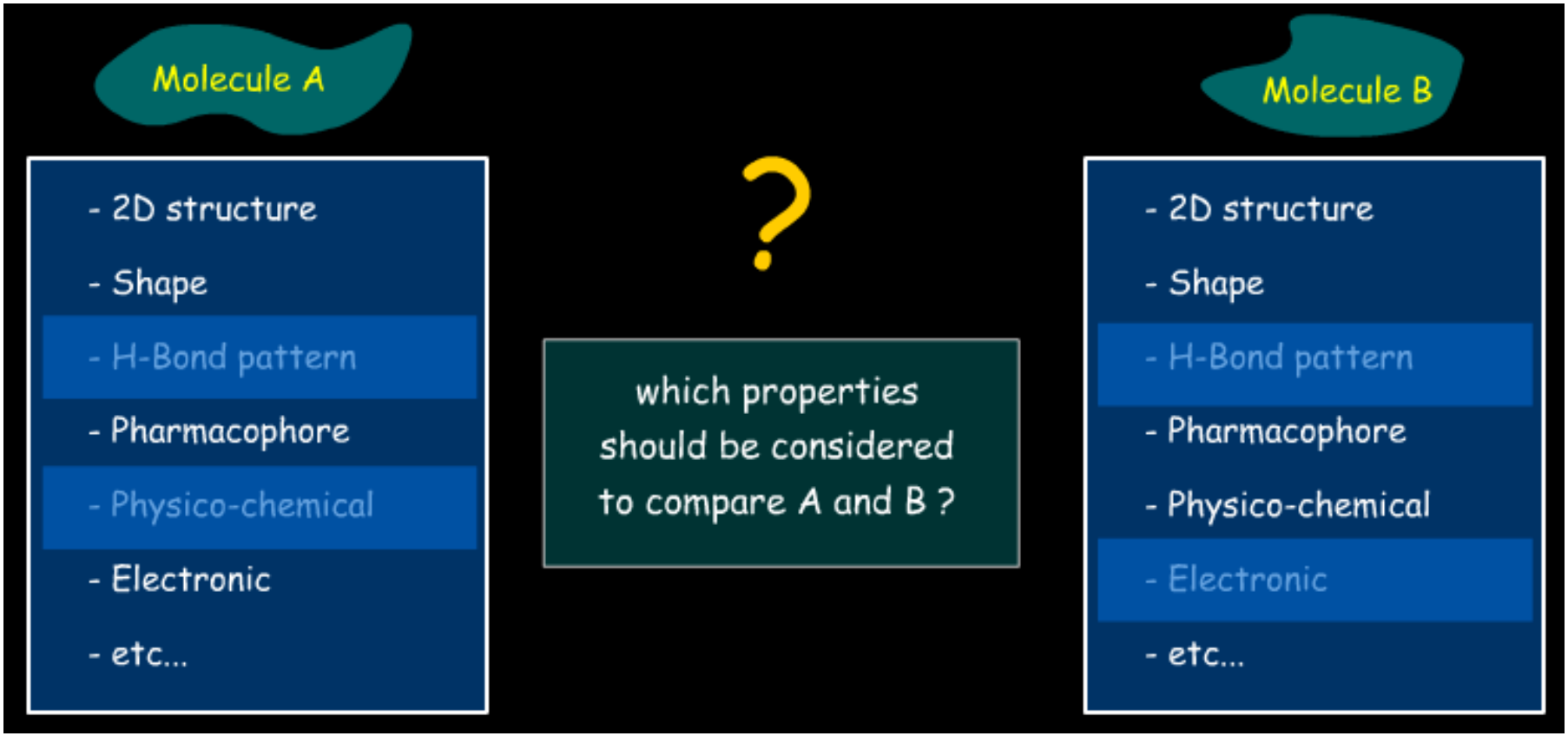
- Descriptors are field strengths around molecules
  - Steric, electrostatic, H-bond, CoMSIA, indicator …
- Fields directly encode shape & property data



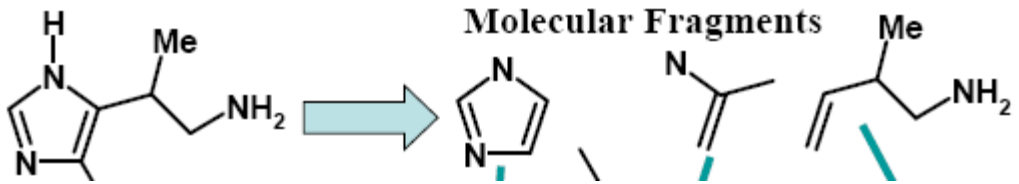
Three-dimensional descriptors attempt to capture the 3D properties of the molecules to be compared, along with their relevant properties. A major difference between 3D descriptors and those of the 1D and 2D types, is that the three-dimensional properties are dependent on the particular conformation of the molecule. Thus, one can either use a low-energy structure of the molecule to calculate its three-dimensional descriptors, or run conformational sampling to explore the conformationally accessible space of the structure. While the second route introduces more information into the descriptor, it should be noted that at the same time it also introduces more noise, which does not necessarily improve the signal-to-noise ratio of the descriptor.

Image taken from ‘Virtual screening and ADME’
György M. Keseru

Recherche de similarité (2D/3D), comment.



Fingerprint, calcul



0	1	1	0	1	0	1	0	1	0

a) Fingerprint binaire

0,3	42	17,3	11	4,8	14,7	0,5	-5,1	57,4	1,02

b) Valeur numérique

Logp Surface

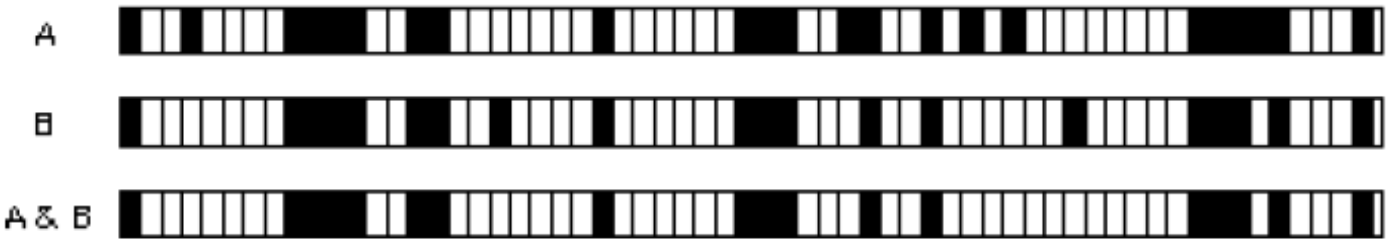
Lien sur le net

Coefficient de Tanimoto

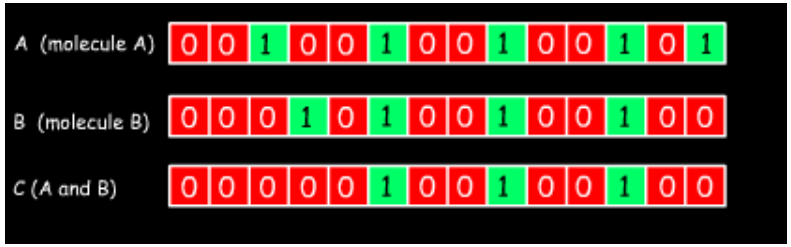
$$T = \frac{A \cap B}{(A+B) - A \cap B}$$

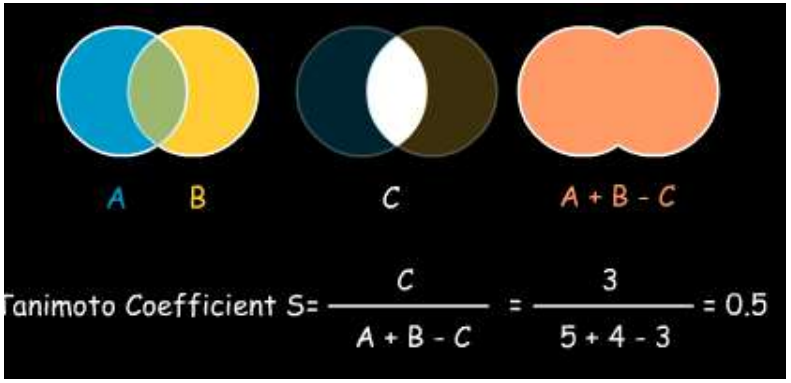
where A, B, A&B, are the number of bits set in fingerprint A, B, and A-AND-B.

Example: A, B,and A&B are 24, 21, and 19, respectively, T = 0.73 (1.00 means identical)

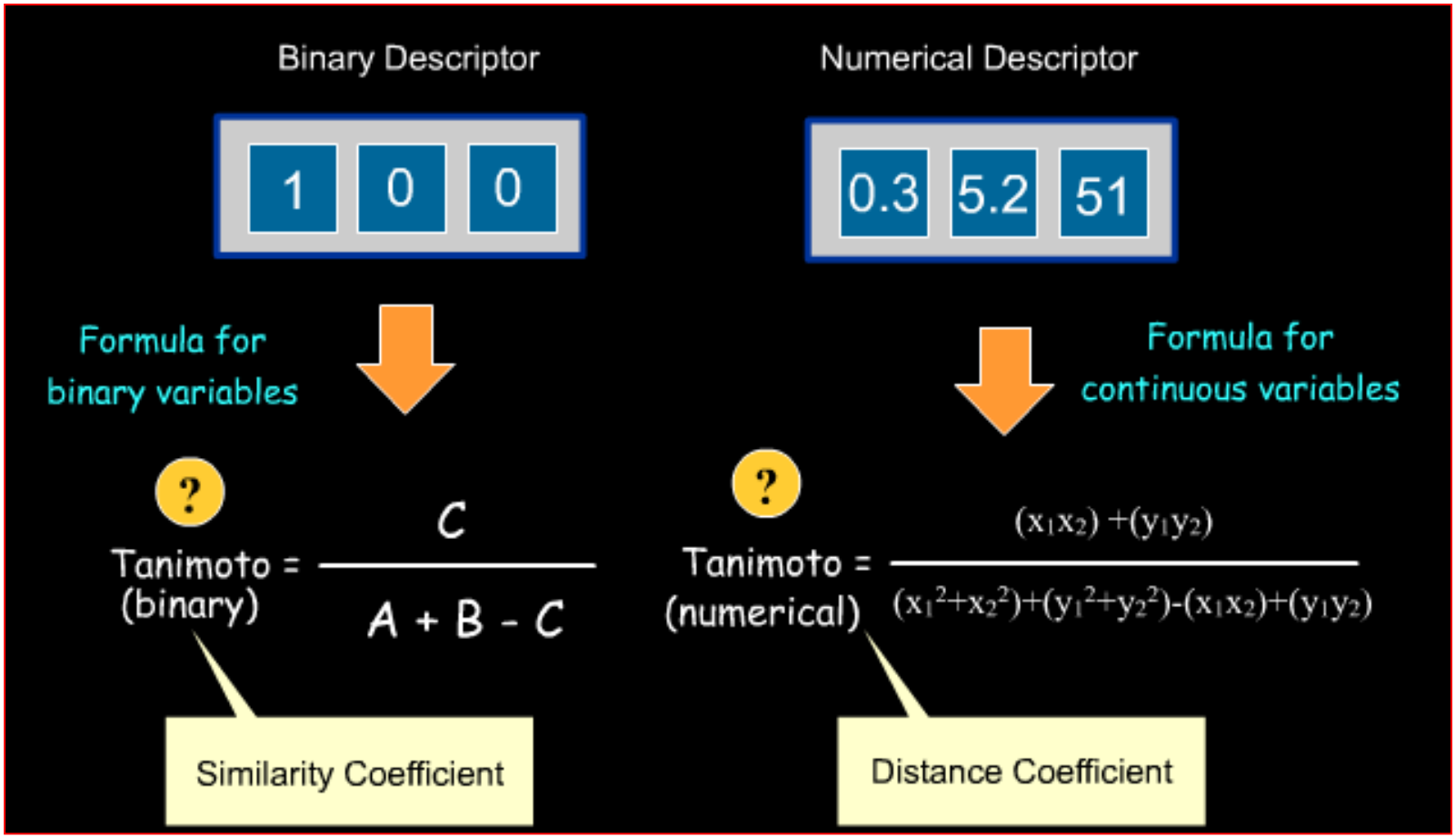


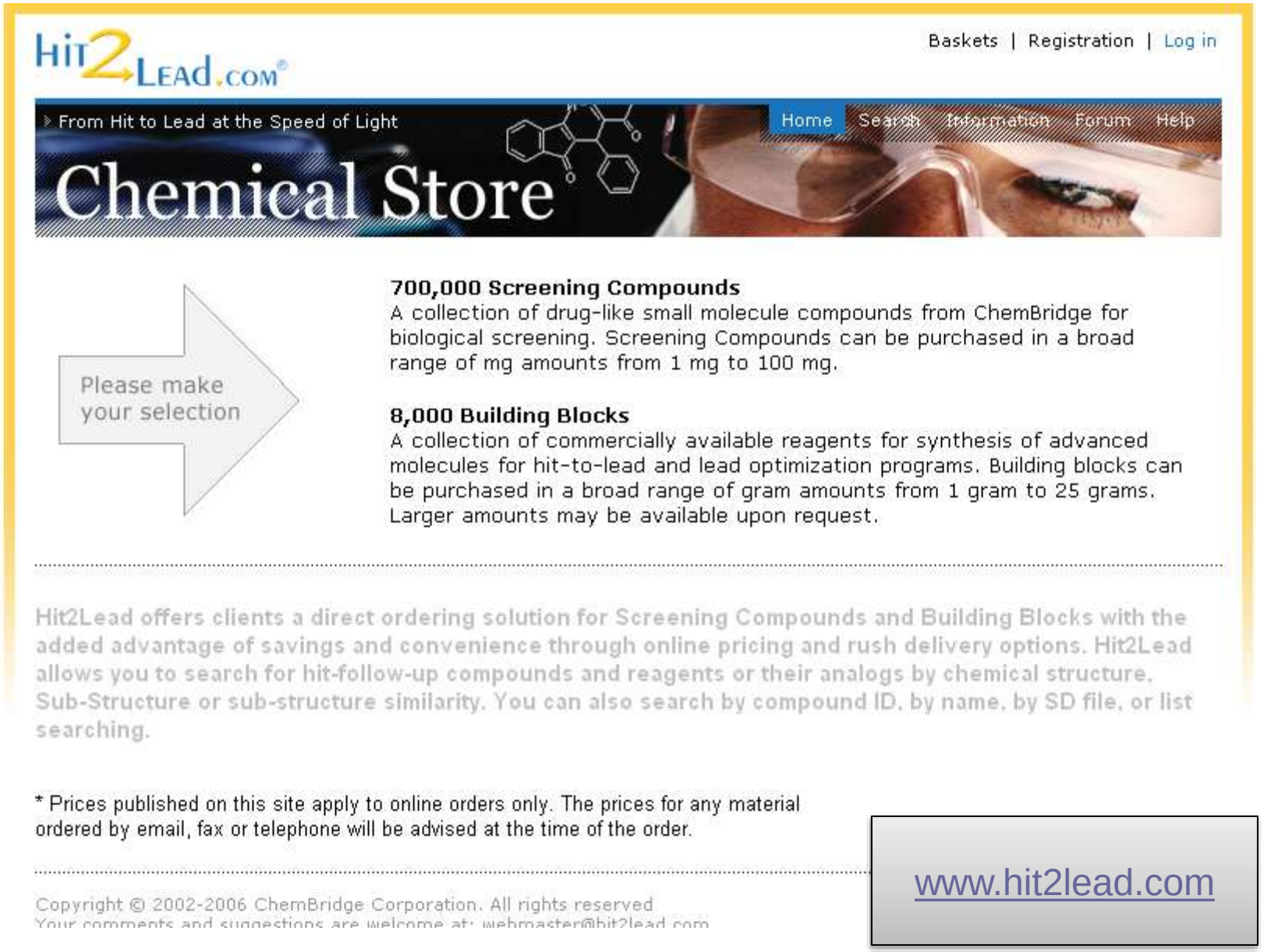
Molecules with T>0.85 are considered as similar.





Coefficient de similarité et distance





www.hit2lead.com

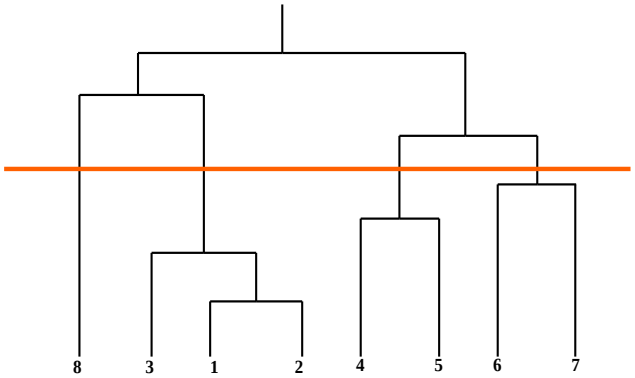
Cluster Analysis

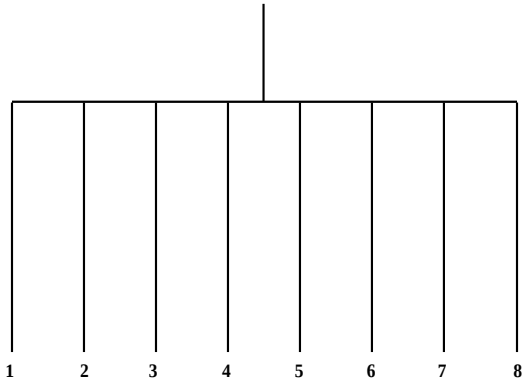
- Classer des molécules en fonction de leur similarité.
- Les molécules dans le même cluster sont similaires et inversement (notion de seuil).
- Nombreux algorithmes.

Downs, G. M., Barnard, J. M., *Rev. Comput. Chem.*, *18* (**2002)**

Clusters Hiérarchiques et non-hiérarchiques

Les molécules sont dans les différents clusters





Notion de seuil pour définir le nombre de clusters

Cluster Analysis

a) Achat de composés différents.
   Enrichissements de banques.
   Limiter les redondances.
   Classification des nouveaux composés acquis.

b) HTS : Un composé par cluster en première approche.
   Composés du même cluster qu'une molécule active.

c) Diversité moléculaire pour optimiser la recherche de 'Hits'.
   (large spectre, représentation des différents clusters)

d) Favoriser une étude QSAR.

Diversité de Chimiothèques

Preparation of a molecular database from a set of 2 million compounds for virtual screening applications :

gathering, structural analysis and filtering.

Université d'Orléans

Introduction

HTS needs that content of the screening libraries is of high quality. Redundant and non appropriate (reactive and toxic) structures have to be removed and a high diversity is required to increase the likelihood that novel active compounds are discovered during the screening process.

In this study, 15 libraries were gathered from diverse origins :

- The uniqueness of the libraries was compared.

- Their diversity was investigated and visualized graphically, using MACCS fingerprints and recently designed 2D surface descriptors.

- The drug-likeness of these databases was also assessed using some common chemical features.

Diversité de Chimiothèques

Company	Library name	www address
AcbBlocks	ACB	http://www.acbblocks.com
Asinex	Asinex	http://www.asinex.com
Key Organics	Bionet	http://www.keyorganics.ltd.uk
ChemBridge	ChemBridge	http://www.chembridge.com
ChemDiv	ChemDiv	http://www.chemdiv.com/main.phtml
ChemStar	Chemstar	http://www.chemstar.ru
InterBioScreen	IBS	http://www.ibscreen.com
MayBridge	MaybBridge	http://www.maybridge.com
Molecular Diversity Preservation International	MDPI	http://www.mdpi.org
NCI/NIH Developmental Therapeutics Program	NCI	http://dtp.nci.nih.gov/index.html
Tripos	Tripos	http://www.tripos.com

Redundancy

The internal duplication rate in each database and the redundancy between databases were evaluated. Non-stereospecific SMILES codes were used because of the lack of any stereochemistry information in some libraries.
All except one stereoisomer of each compound were removed.

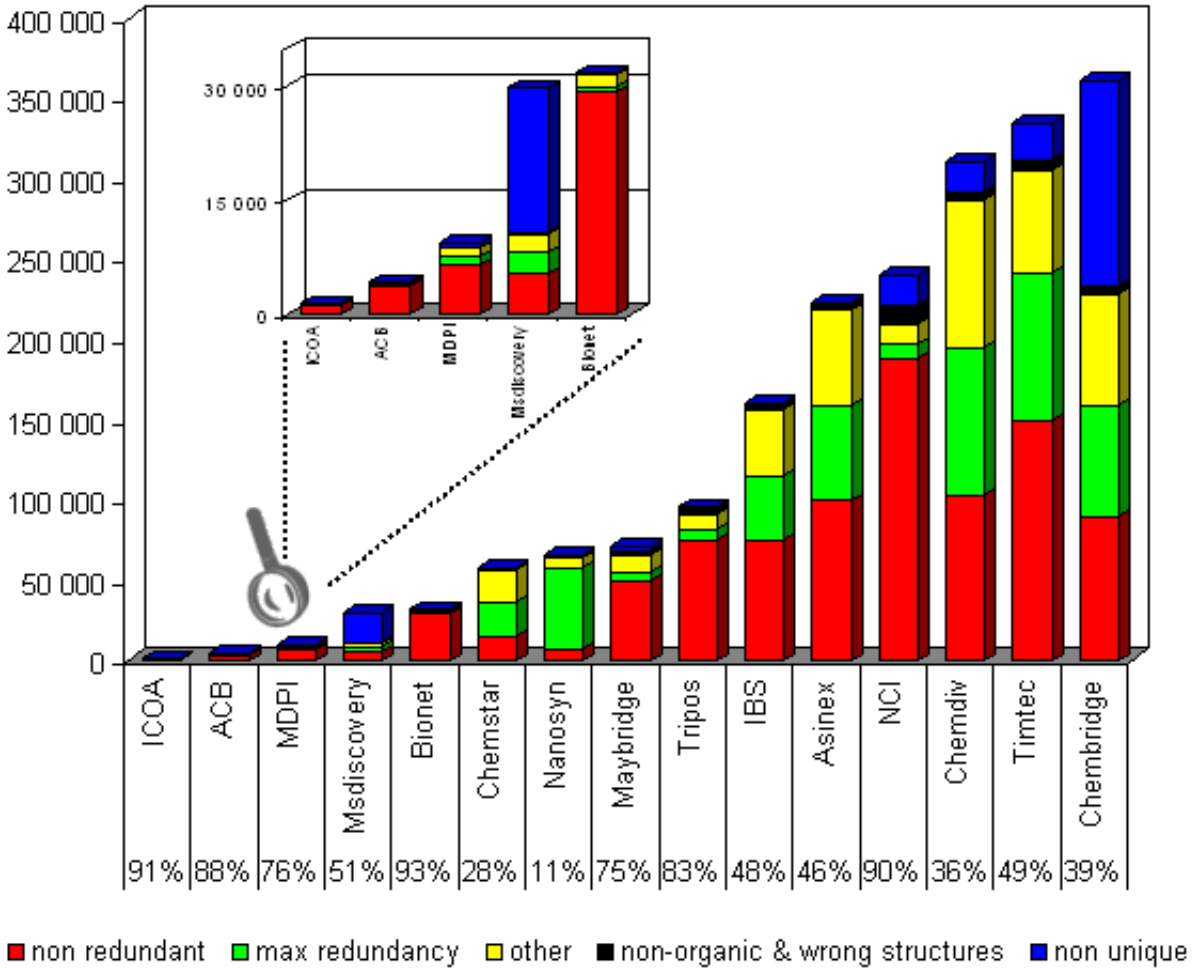
The doubloons in each library (blue)

Non-organic compounds (black)

Compounds found only in this chemical library (red)

Compounds which are common to the one with the largest redundancy among the 14 other libraries (green)

The other redundant compounds with other libraries (yellow)

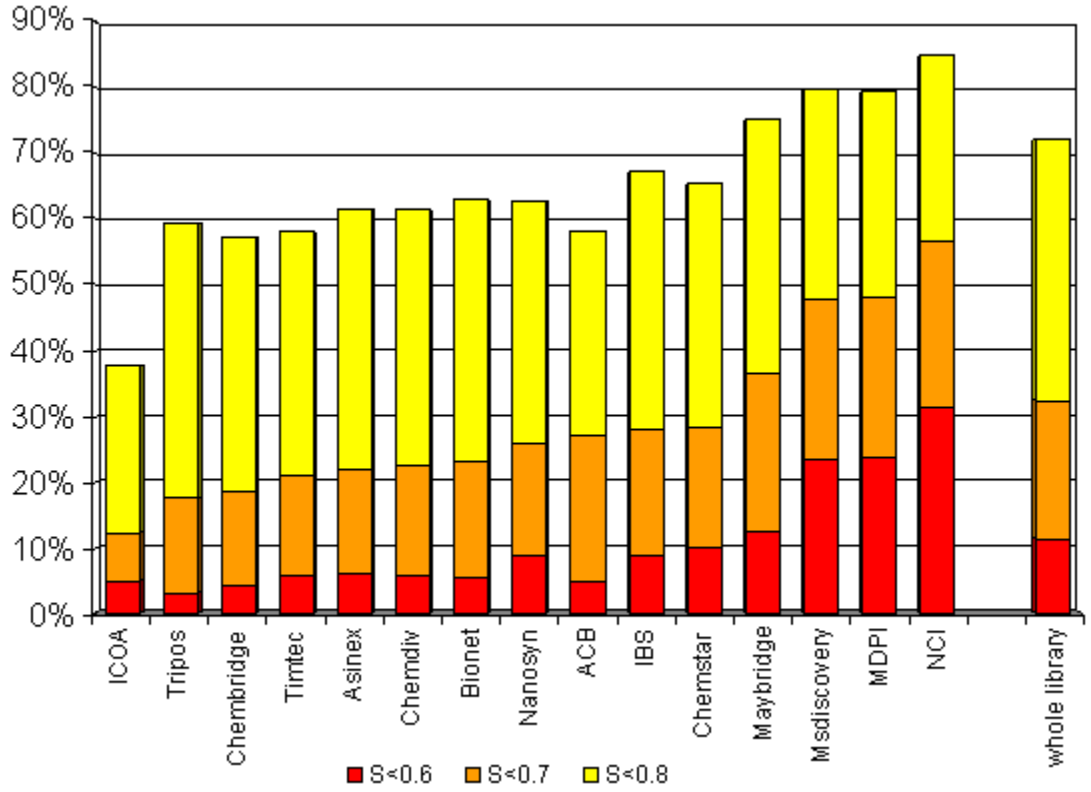


Diversity

The molecular diversity of each database was quantified using 2D molecular descriptors based on atomic contributions to van der Waals surface area, log P, molecular refractivity, and partial charge. They have the advantage of being conformation independent and so very quickly calculated. They are supposed to capture hydrophobic and hydrophilic effects, polarizability and electrostatic interactions.

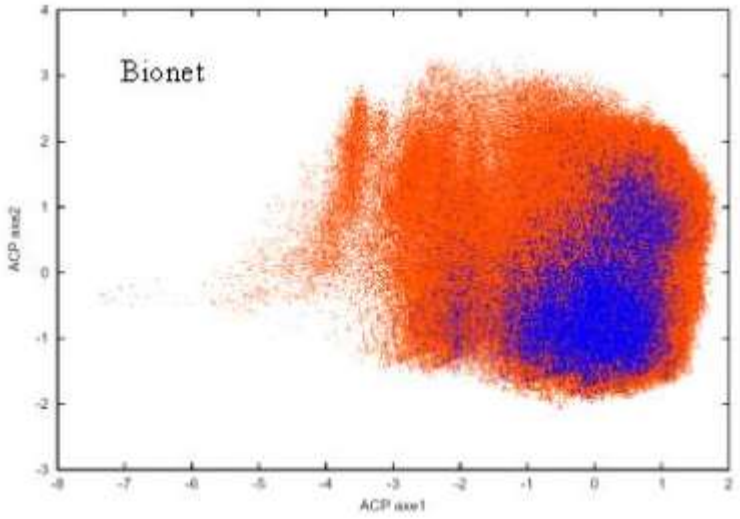
a) Similarity indices between compounds of each database were calculated using the Tanimoto coefficient. The diversity of a database was expressed as the fraction of compounds which has a similarity index < 0.6, 0.7 or 0.8, on average, with the other compounds of the database.

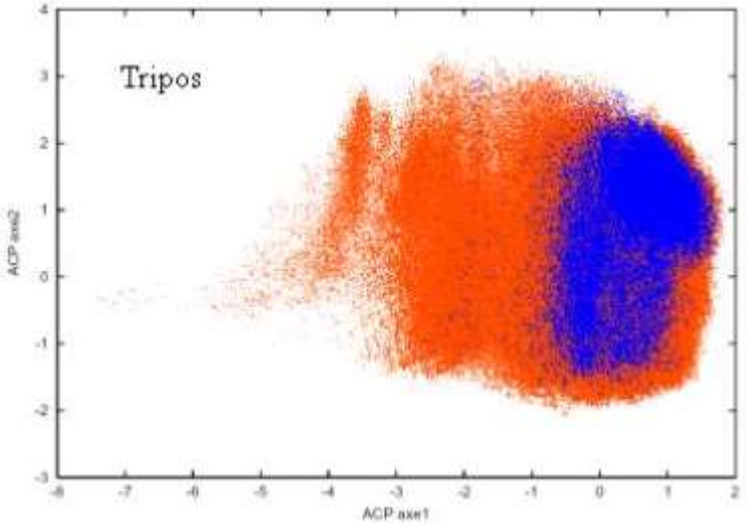
b) The set of 166 MACCS structural keys were also calculated for each compound. Then, it was reduced with principal component analysis (PCA). The compounds were displayed on a graph using the 2 first latent variables. In this way, the diversity and the location of each database in the chemical space were graphically compared.

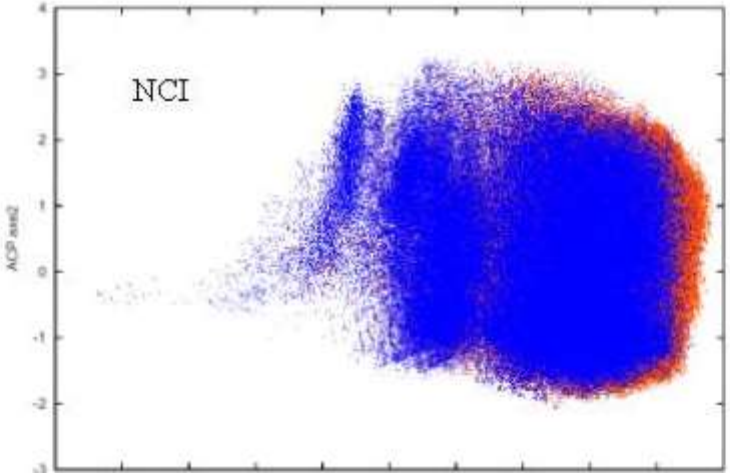


The NCI, MDPI, MsDiscovery and Maybridge databases to a minor extent are the most diverse libraries. For instance, 31 % of the molecules in the NCI library are < 0.6 similar to the other molecules of this database and 95% are < 0.8 similar to the other molecules.

ICOA database is the least diverse. Less than 40 % of the molecules are < 0.8 similar to the other compounds. These products come from one single lab which certainly explains this result.


Bionet
ACP axe2
ACP axe1


Tripos
ACP axe2
ACP axe1

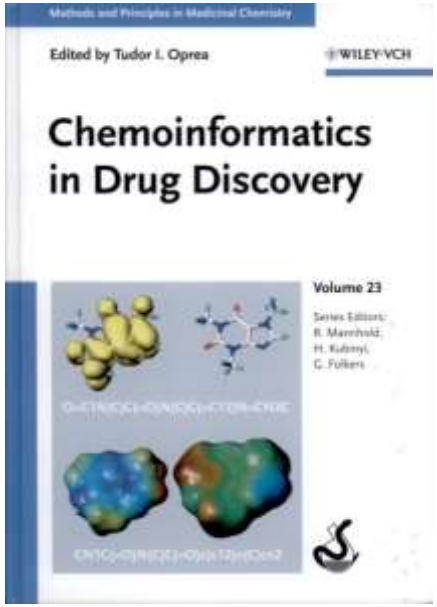

NCI
ACP axe2

Le dogme pour faire un médicament

Lipinski, rules of 5

1.Poids moléculaire < 500 (opt ~= 350)
2.Nbre de liaisons H accepteurs < 10 (opt ~=5)
3.Nbre de liaisons H donneurs < 5 (opt ~=2)
4. -2 < clogP < 5 (opt ~= 3)
5.Nbre d'angles de rotation =< 5/ Pas plus de 5 noyaux attachés

Lipinski *et al*, Adv. Drug. Del. Rev., 23, 3-25 (1997)



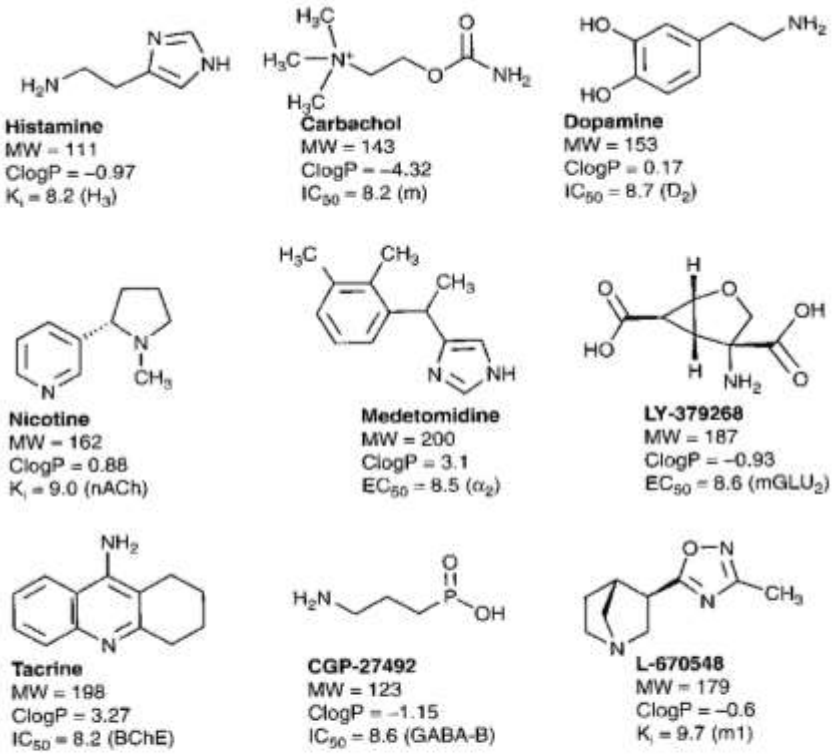


Fig. 2.3 Examples of small molecules (MW ≤ 200) that have biological activity better than 10 nM. Under each molecule, the following information is included: molecule name, MW, ClogP, the biological activity type, value and target. Target names are as follows: H_3 – histamine receptor subtype 3; m – muscarine receptor; D_2 – dopaminergic receptor type 2; nACh – nicotine receptor; α_2 – alpha adrenergic receptor subtype 2; $mGLU_2$ – metabotropic glutamate receptor subtype 2; BChE – butyryl choline esterase; GABA-B – gamma-amino butyric acid receptor subtype B; m1 – muscarinic receptor subtype 1.

Oui mais… 25% des ‘highly active molecules’ ne respectent pas le Rule of 5

Domains of Chemoinformatics

Chemoinformatics is an interdisciplinary field that combines chemistry, computer science, and information technology to analyze and model chemical data.

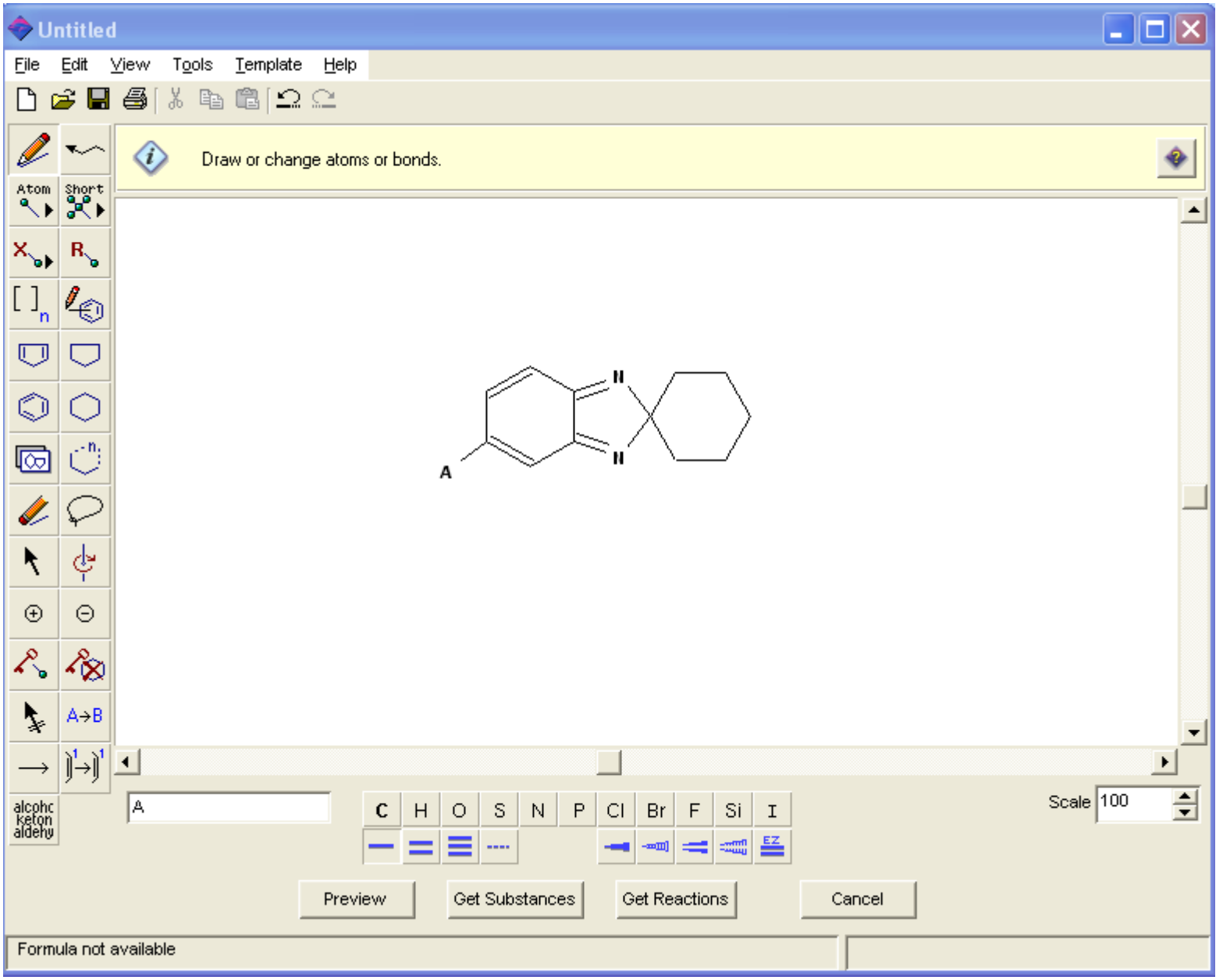
1. Chemical Structure Databases, visualization, properties, searching and retrieving :

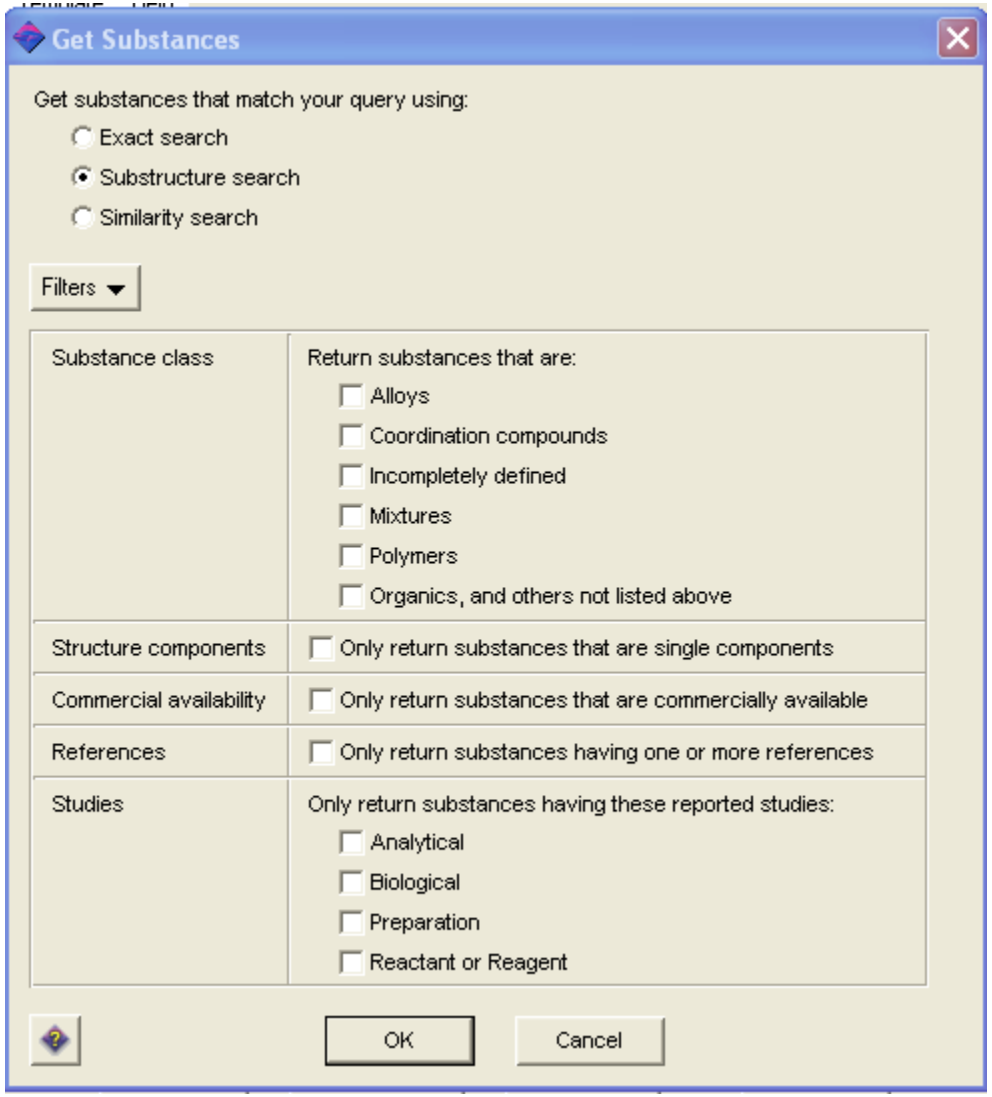
- Creating 2D and 3D visual representations of molecules : ChemSketch…
- Managing vast databases of chemical structures : PubChem, ChemSpider.
- Representation and research structures and substructures : Molecular Graph-structure.
- Similarity search (2D / 3D), clustering and diversity analysis : Tanimoto, clustering.
- Search chemical molecules, patent databases or chemical reactions
- QSAR (Quantitative Structure-Activity Relationship) modeling.

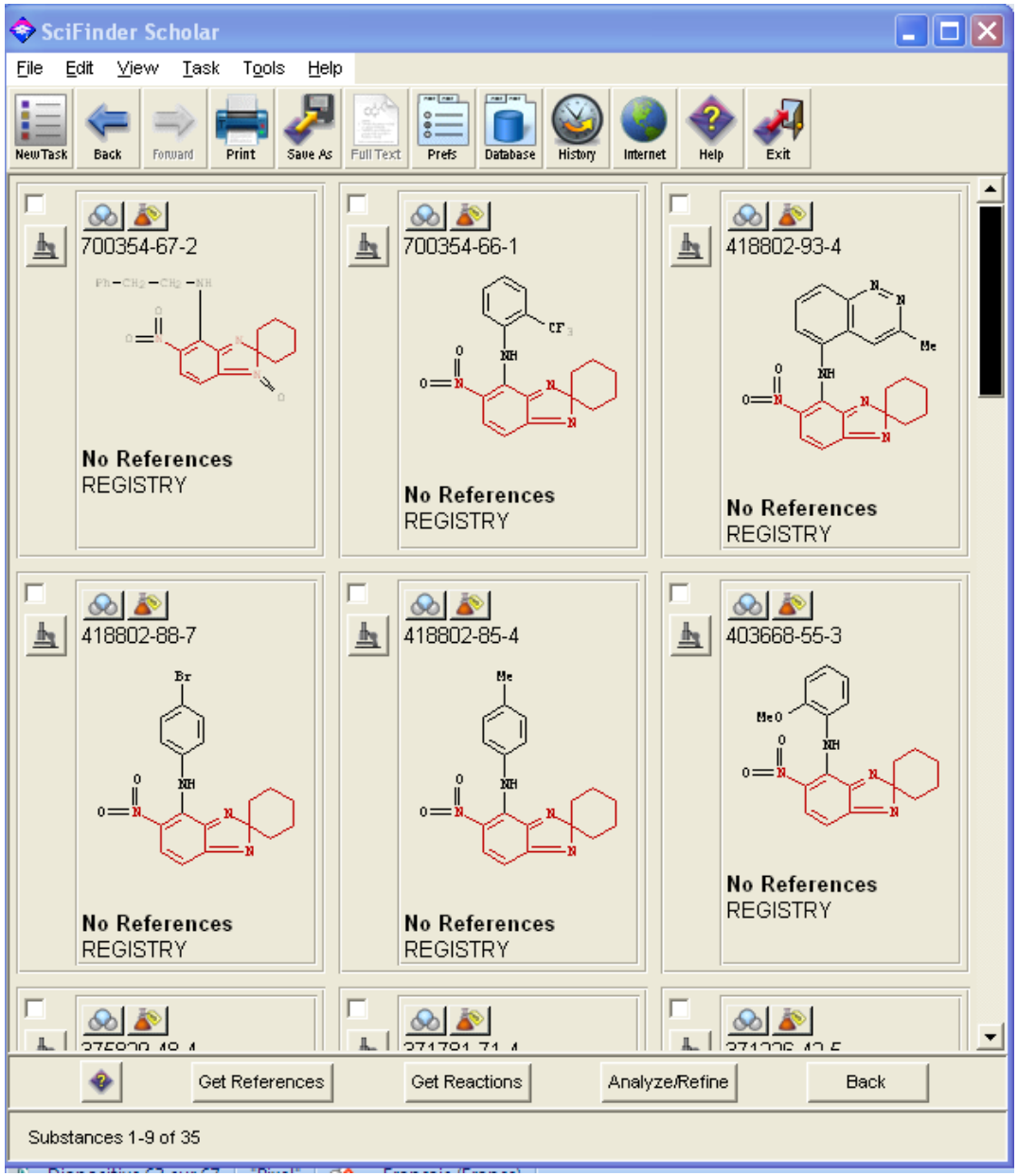
2. Molecular Modeling ant Interactions :

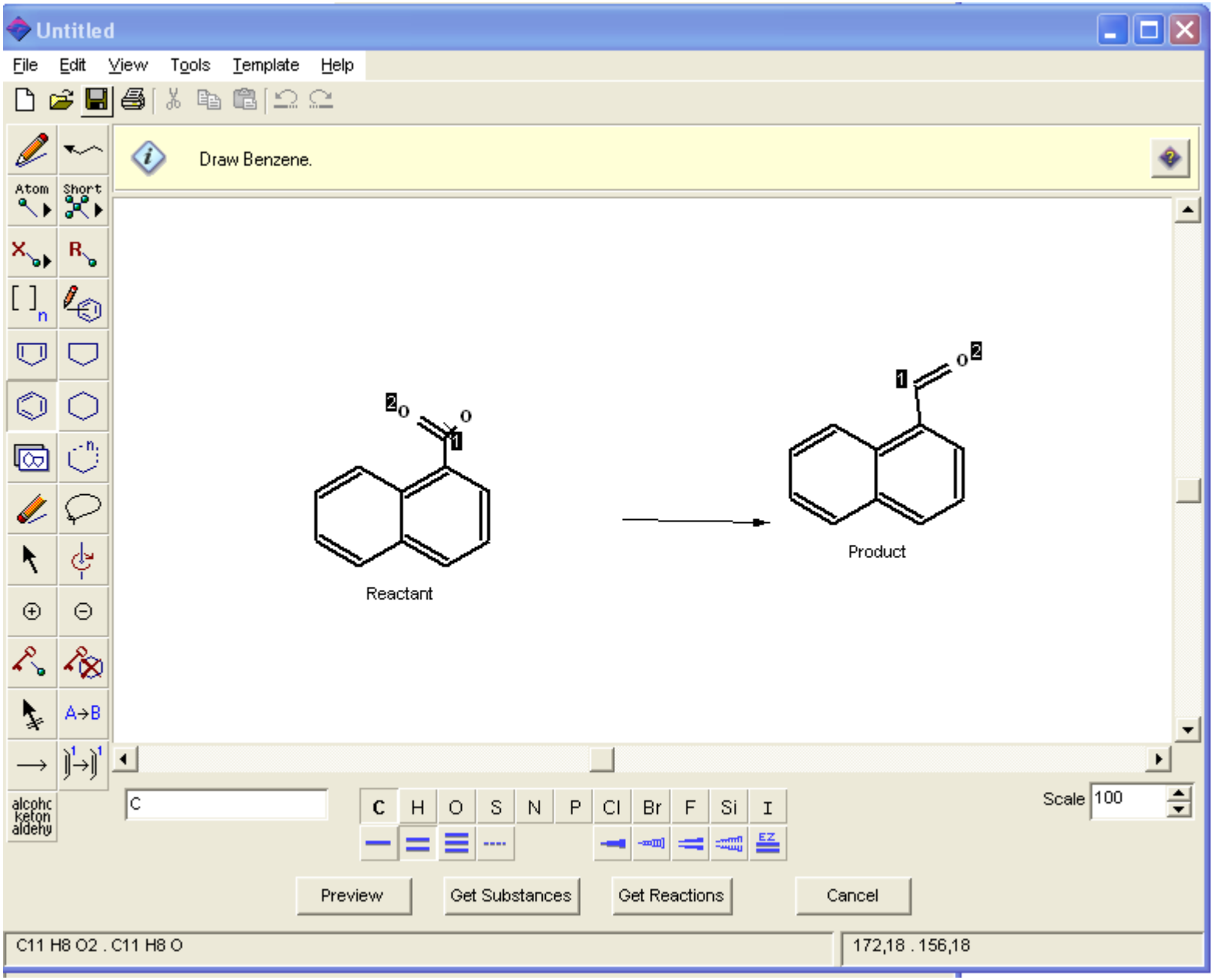
- Predicting molecular structures and properties : homology modeling.
- Molecular dynamics simulations.
- Molecular docking simulations.
- Pharmacophore modeling.
- A.I.

http://www.cas.org/


Untitled
File Edit View Tools Template Help
Draw or change atoms or bonds.
Atom
Short
A→B
alcohc
keton
aldehy
A
C H O S N P Cl Br F Si I
Scale 100
Preview
Get Substances
Get Reactions
Cancel
Formula not available


Get Substances
Get substances that match your query using:
Exact search
Substructure search
Similarity search
Filters
Substance class
Return substances that are:
Alloys
Coordination compounds
Incompletely defined
Mixtures
Polymers
Organics, and others not listed above
Structure components
Only return substances that are single components
Commercial availability
Only return substances that are commercially available
References
Only return substances having one or more references
Studies
Only return substances having these reported studies:
Analytical
Biological
Preparation
Reactant or Reagent
OK
Cancel

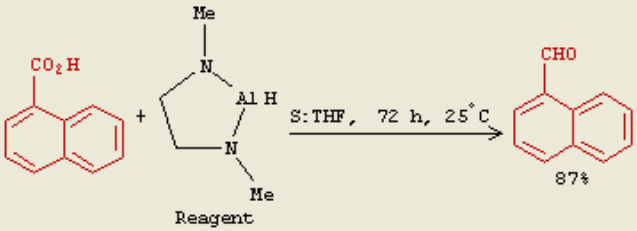

SciFinder Scholar
File Edit View Task Tools Help
New Task
Back
Forward
Print
Save As
Full Text
Prefs
Database
History
Internet
Help
Exit
700354-67-2
No References
REGISTRY
700354-66-1
No References
REGISTRY
418802-93-4
No References
REGISTRY
418802-88-7
No References
REGISTRY
418802-85-4
No References
REGISTRY
403668-55-3
No References
REGISTRY
Get References
Get Reactions
Analyze/Refine
Back
Substances 1-9 of 35


Untitled
File
Edit
View
Tools
Template
Help
Draw Benzene.
Atom
Short
R
A→B
alcohc
keton
aldehy
Reactant
Product
C
H
O
S
N
P
Cl
Br
F
Si
I
Scale
100
Preview
Get Substances
Get Reactions
Cancel
C11 H8 O2 . C11 H8 O
172,18 . 156,18

SciFinder Scholar

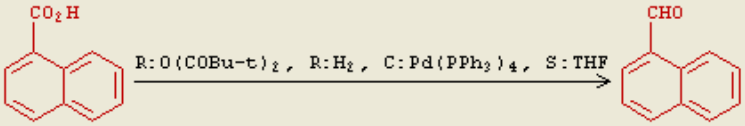
File Edit View Task Tools Help

New Task | Back | Forward | Print | Save As | Full Text | Prefs | Database | History | Internet | Help | Exit



NOTE: Reactants: 1, Reagents: 1, Solvents: 1,
Steps: 1, Stages: 1

Bulletin of the Korean Chemical Society, 23(9), 1340-1342; 2002
CASREACT



NOTE: high pressure,
Reactants: 1, Reagents: 2, Catalysts: 1, Solvents: 1,
Steps: 1, Stages: 1

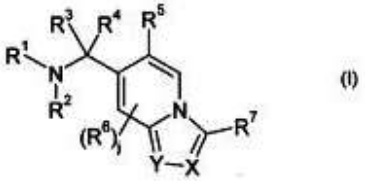
Bulletin of the Chemical Society of Japan, 74(10), 1803-1815; 2001
CASREACT

CO2H CHO

Get References | Analyze/Refine | Back

Reactions 1-3 of 10

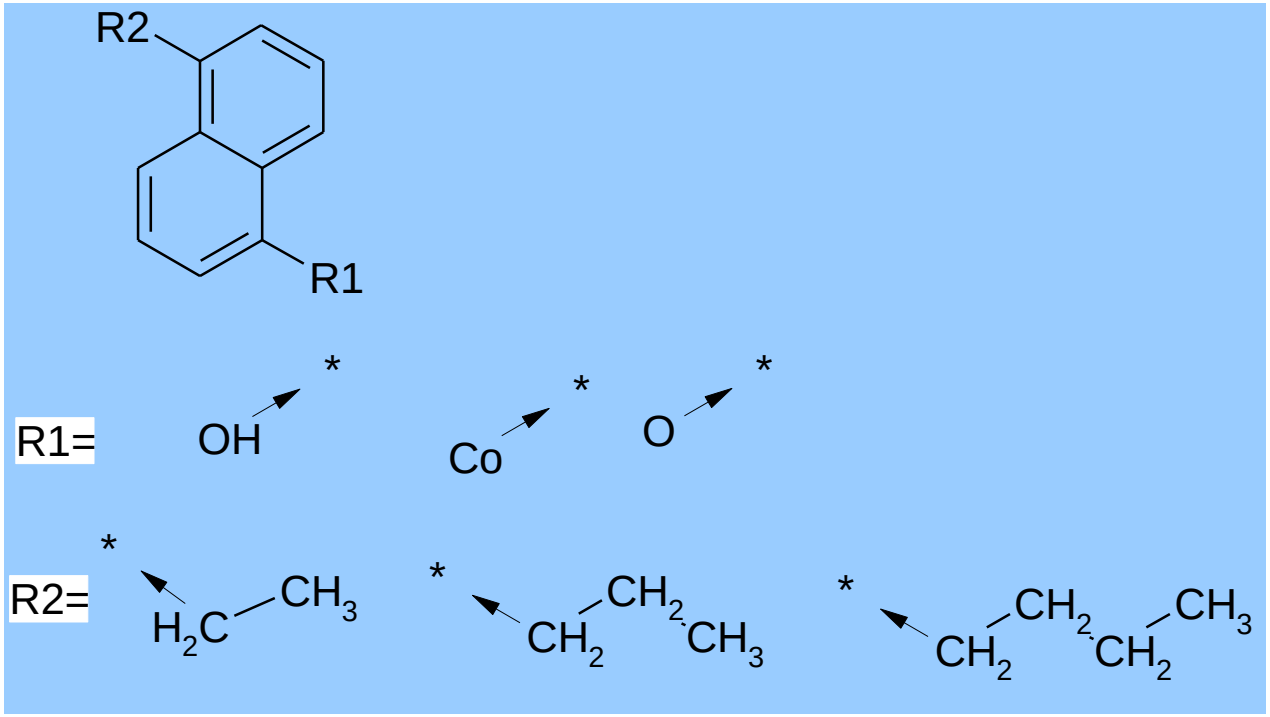
Accessing Patents on Molecules : Scientific research and intellectual property protection



Markush structure (time consuming)

Structures de Markush

- Structures moléculaires qui incluent un degré de variabilité (Structures génériques, R-groupes)



Domains of Chemoinformatics

Chemoinformatics is an interdisciplinary field that combines chemistry, computer science, and information technology to analyze and model chemical data.

1. Chemical Structure Databases, visualization, properties, searching and retrieving :

- Creating 2D and 3D visual representations of molecules : ChemSketch…
- Managing vast databases of chemical structures : PubChem, ChemSpider.
- Representation and research structures and substructures : Molecular Graph-structure.
- Similarity search (2D / 3D), clustering and diversity analysis : Tanimoto, clustering.
- Search chemical molecules, patent databases or chemical reactions
- QSAR (Quantitative Structure-Activity Relationship) modeling.

2. Molecular Modeling ant Interactions :

- Predicting molecular structures and properties : homology modeling.
- Molecular dynamics simulations.
- Molecular docking simulations.
- Pharmacophore modeling.
- A.I.