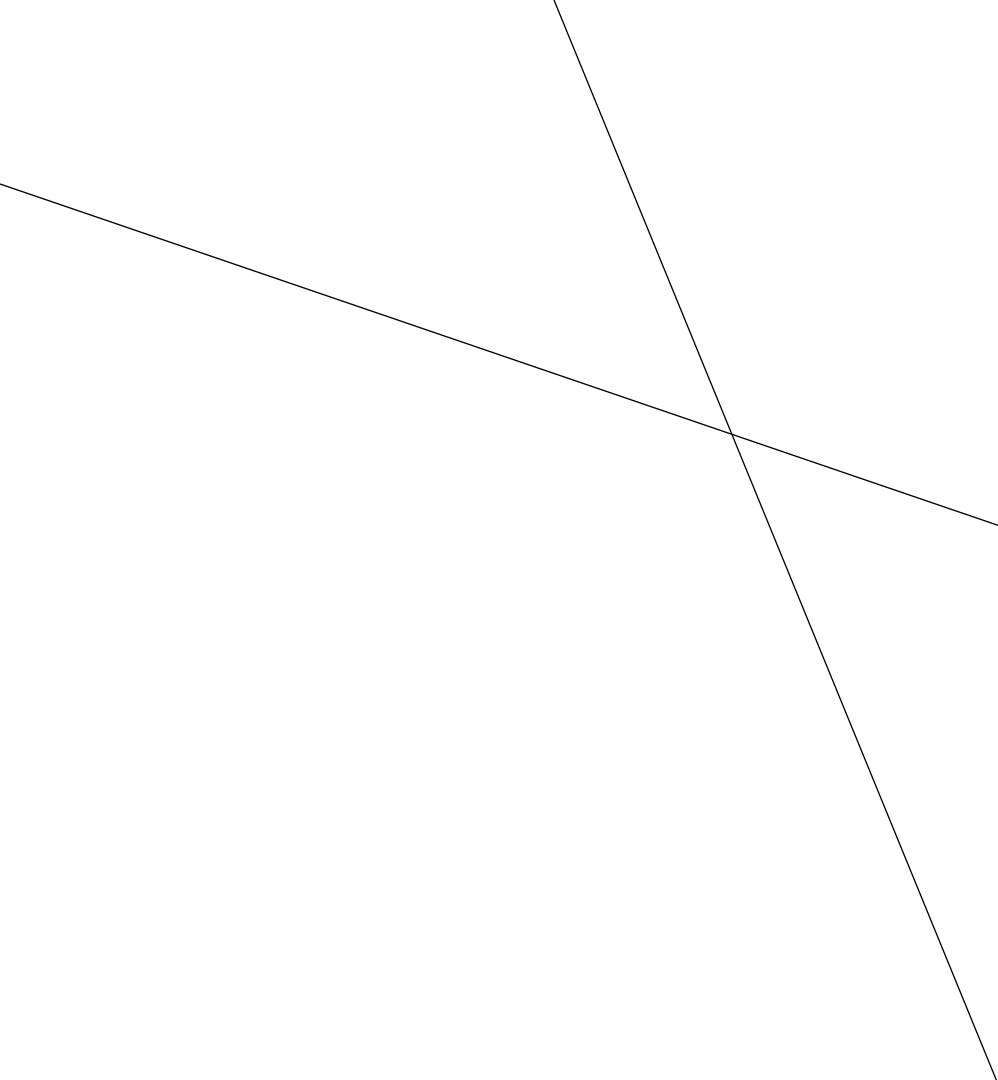


# QUIMIOINFORMÁTICA

Fernán Agüero  
Instituto de Investigaciones Biotecnológicas, UNSAM

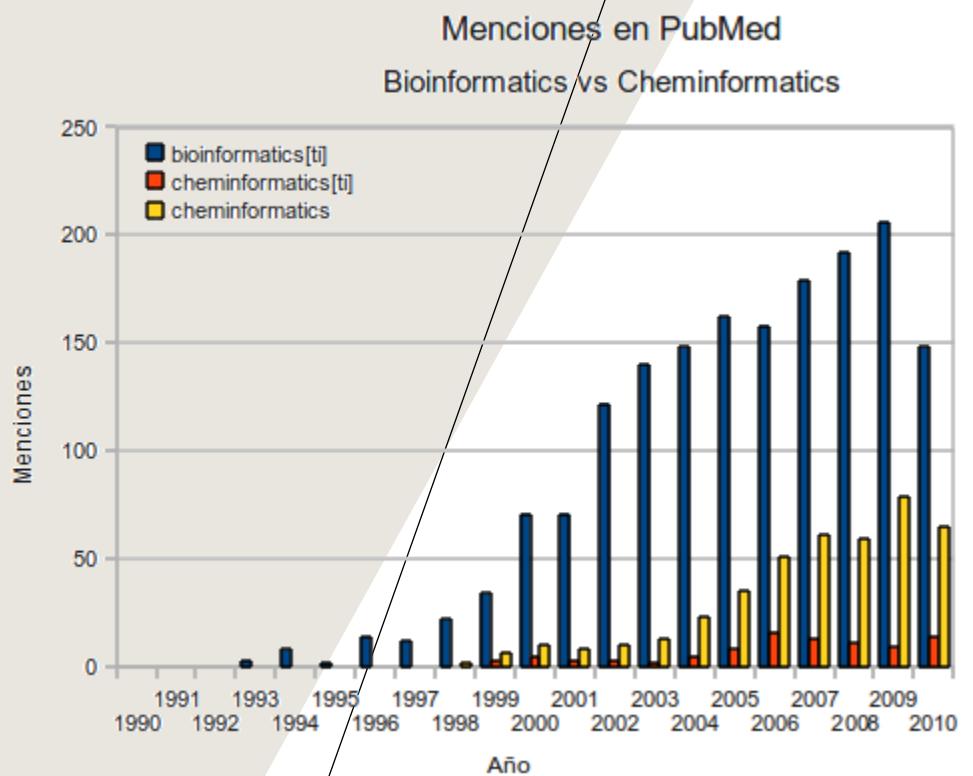


# INTRODUCTION TO CHEMINFORMATICS

Cheminformatics is a relatively new field of information technology that focuses on the collection, storage, analysis, and manipulation of chemical data. The chemical data of interest typically includes information on small molecule formulas, structures, properties, spectra, and activities (biological or industrial). Cheminformatics originally emerged as a vehicle to help the drug discovery and development process, however cheminformatics now plays an increasingly important role in many areas of biology, chemistry, and biochemistry. The intent of this unit is to give readers some introduction into the field of cheminformatics and to show how cheminformatics not only shares many similarities with the field of bioinformatics, but that it can also enhance much of what is currently done in bioinformatics.

-- David Wishart

# CHEMINFORMATICS – QUÉ ES?



*“The application of computational techniques to the discovery, management, interpretation and manipulation of chemical information and data extracted therefrom”.*

Chemistry plans a structural overhaul.  
Nature 419:4-7 (2002)

Se la conoce como:

- Computational chemistry
- Theoretical chemistry
- Molecular modeling

Nace con el desarrollo de la mecánica cuántica a principios del siglo XX

Parece haber pasado desapercibida en la revolución “omica”

En activo desarrollo y expansión a partir de la introducción de las computadoras

# CHEMINFORMATICS EN LA LITERATURA

Term	Google	Google Scholar	Web of Knowledge	Scopus
Chemical documentation	695,000	66	1	34
Chemical informatics	50,400	129	20	39
Chemical information management	978	42	4	28
Chemical information science	779	17	2	5
Chemiinformatics	2,230	2	2	2
Cheminformatics	320,000	447	83	250
Chemoinformatics	191,000	5636	99	473

Table 1. Occurrences of search terms in *Google*, *Google Scholar*, the *Web of Knowledge* and *Scopus*

## Google:

- “Bioinformatics” (2023): ~ 240 millones de páginas
- “Cheminformatics” (2023): ~ 1.6 millones de páginas

Willett P (2007). *A bibliometric analysis of the literature of chemoinformatics*. **Aslib Proceedings**, 60: 4-17

# CUESTIONES QUÍMICAS

La química se ocupa de esto

estructura  propiedades

## Compuestos

- Propiedades Físicas (→Energía)
- Propiedades Químicas (Estructura, Reactividad)
- Propiedades Biológicas (→Actividad)
- Separaciones de mezclas de compuestos
- **Aspectos estáticos**

## Transformaciones

- Reacciones químicas
- **Aspectos dinámicos**

Y tiene estos desafíos

propiedades  estructura

## Inferencia

- Qué compuestos (estructuras) van a mostrar una determinada propiedad?
  - Inhibición de una actividad enzimática X (ej. drogas)
  - Propiedades mecánicas y elásticas definidas (ej. polímeros)
- Definir caminos óptimos para la síntesis de compuestos
  - Reacciones
  - Materiales iniciales
- Predecir estructuras
  - A partir de datos experimentales (ej NMR)
  - Compuestos desconocidos

Síntesis, Abstracciones,  
Predicciones

Utilización de información para  
aplicaciones (memorización de  
datos)

Datos ordenados, refinados  
y puestos en contexto

Experimentos, Mediciones

# Predecir EL DESAFÍO DE LA QUIMIOINFORMÁTICA

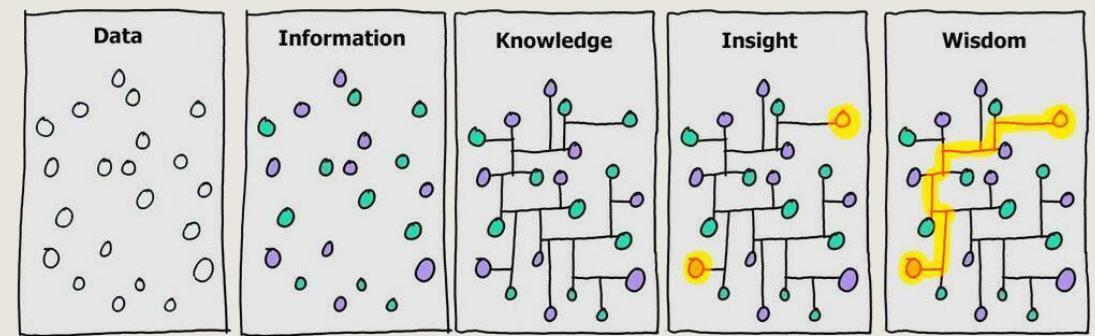
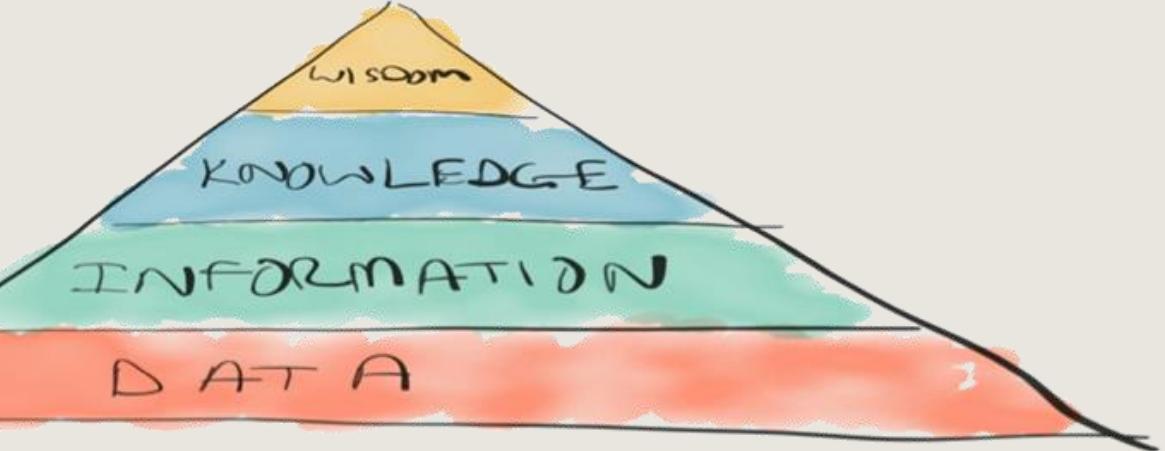
Transformar datos  
en conocimiento

Insight, Wisdom

Knowledge

Information

Data



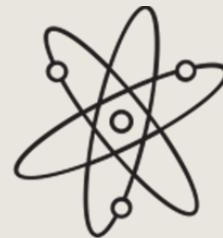
- El curso de una reacción química en un solvente determinado, a una temperatura dada y usando un catalizador definido
- La actividad biológica de un compuesto X contra una proteína target Y

# TEORÍA VS MODELOS



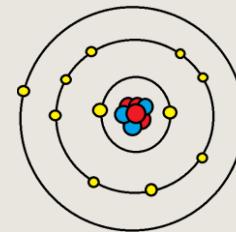
**En esencia son lo mismo, pero**

Una **teoría** suele ser general  
Mientras que los **modelos** introducen particularidades para facilitar la interpretación y el entendimiento  
En algunos casos los **modelos** son aproximaciones, con error medible.



**Mecánica cuántica**

**Teoría** fundamental de la química  
Permite describir un sistema (por ej una molécula) en forma completa, usando funciones de onda, formación y ruptura de enlaces, reacciones químicas, etc.



**Modelo de valencia, capas de electrones y repulsión**

Todos estudiamos este modelo en cursos básicos de química  
Es un modelo o aproximación  
Permite entender los mismos sistemas fácilmente  
Pero tiene problemas para describir comportamientos de algunos sistemas químicos

# REPRESENTACIÓN DE COMPUESTOS QUÍMICOS

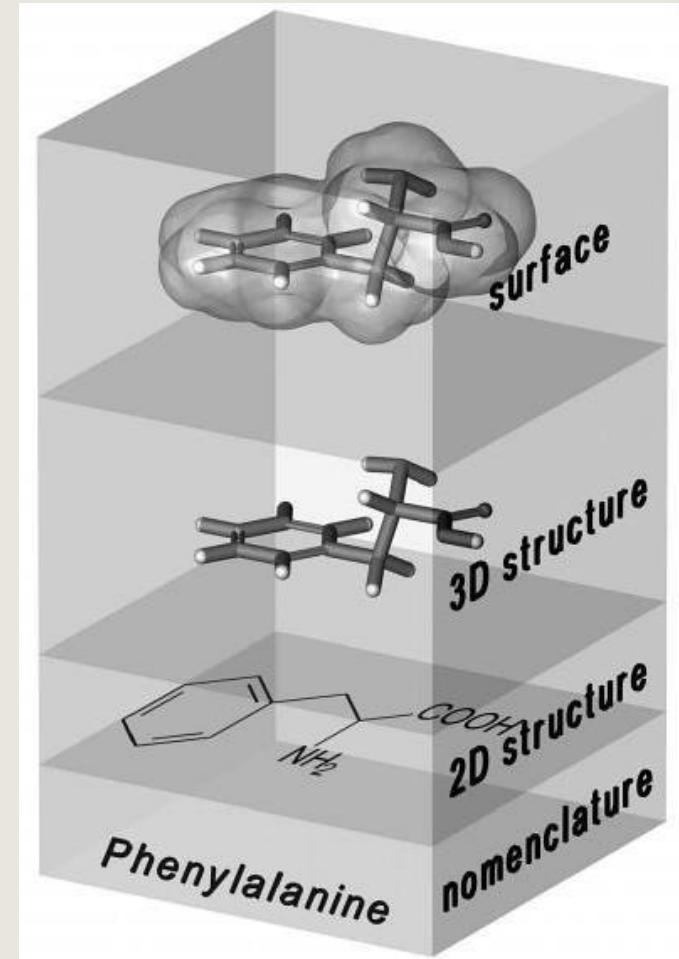
## 2D Structure vs 3D Structure

2D: Lenguaje natural “universal” entre químicos

- Explica la topología de una molécula
- Qué átomos están conectados mediante qué enlaces
- No explica el arreglo tridimensional de los átomos

3D: Requiere datos adicionales

- Posición de los átomos en el espacio
- Ángulos y distancias de los enlaces

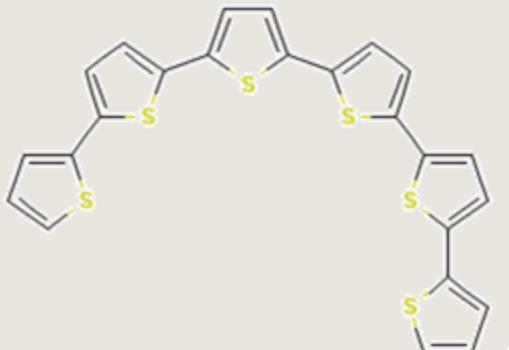
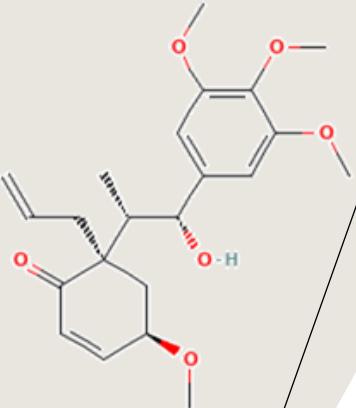


Hierarchical scheme for representations of a molecule with different content of structural information.

Tomado de J Gasteiger & T Engel (2003).

Moléculas con  
nombres  
populares raros:

Traumatic acid  
Erotic acid  
Commic acid  
Diabolic acid  
Megaphone  
Sextiophene



# NOMENCLATURA QUÍMICA

## Histórica

aqua fortis (nitric acid)  
oil of vitriol (sulfuric acid)  
sweet oil of vitriol (diethyl ether)

## Trivial

Fenilalanina  
Ibuprofeno

Popular, pero difícil de sistematizar

## IUPAC

2-amino-3-phenylpropanoic acid  
2-[4-(2-methylpropyl)phenyl]propanoic acid

Sistématico, pero los nombres pueden ser largos!

## Fórmula empírica

C<sub>9</sub>H<sub>11</sub>NO<sub>2</sub>  
C<sub>13</sub>H<sub>18</sub>O<sub>2</sub>

Ambiguo: varios compuestos pueden tener la misma fórmula

# REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: SMILES

## SMILES (Simplified Molecular Input Line Entry System)

Introducido en 1986 por David Weininger

Representa moléculas en forma lexicográfica

**Usa conceptos de grafos | Nodos conectados a través de aristas o arcos**

### Reglas:

Los átomos se representan con sus respectivos símbolos:

C, N, Br, Na, Cl, O, F

MAYUSCULAS → alifáticos; minúsculas → aromáticos

Los hidrógenos son implícitos

Los átomos vecinos aparecen juntos

Se usan paréntesis cuando hay más de un vecino: ramificaciones

Enlaces dobles se representan usando '='

Enlaces triples se representan usando '#'

Quiralidad: '@' (contrario a las agujas del reloj)

'@@' (en el sentido de las agujas del reloj)

Anillos: números a continuación de los átomos que abren/cierran el ciclo

Más información y reglas en:

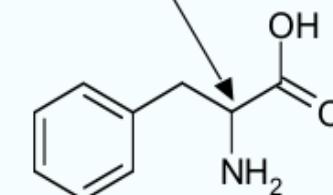
<https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>

[https://en.wikipedia.org/wiki/Simplified\\_molecular-input\\_line-entry\\_system](https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system)



NC(Cc1ccccc1)C(O)=O

phenyl ring



### Otros ejemplos:

Ciclohexano: C1CCCCC1

Benceno: C1=CC=CC=C1 (Kekulé)

Benceno: c1ccccc1

Etanol: CCO

Piridina: C1=CC=NC=C1 (Kekulé)

Piridina: c1ccncc1

Ácido acético: CC(=O)O

Ácido cianhídrico: C#N

L-alanina: N[C@@H](C)C(=O)O

L-alanina (sin especificar quiralidad): N[CH](C)C(=O)O

Cloruro de Sodio: [Na+].[Cl-]

# ANILLOS EN SMILES

## Linealizar y Etiquetar

**Linealizar el anillo en cualquier parte**

Benceno: cccccc (C=CC=CC=C)

Dioxano: occocc, ccocco, coccoc

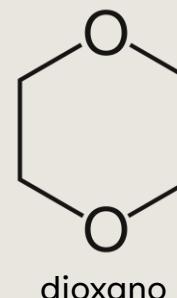
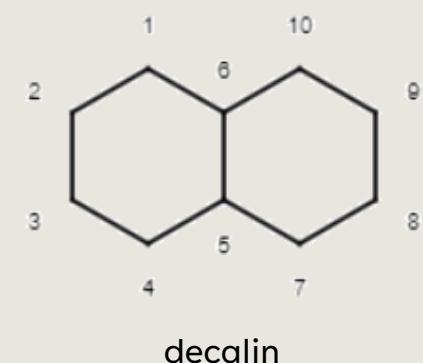
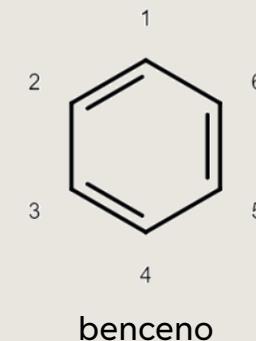
**Agregar etiquetas numéricas para indicar el inicio y cierre del anillo**

Benceno: c1ccccc1

Dioxano: O1CCOCC1, C1COCCO1, C1OCCOC1

*Las etiquetas numéricas pueden empezar en cero (0) pero rara vez se usa*

Decalin: C1CCCC2C1CCCC2, C1CCCC2CCCCC12



# REPRESENTACIÓN DE PATRONES EN MOLECULAS

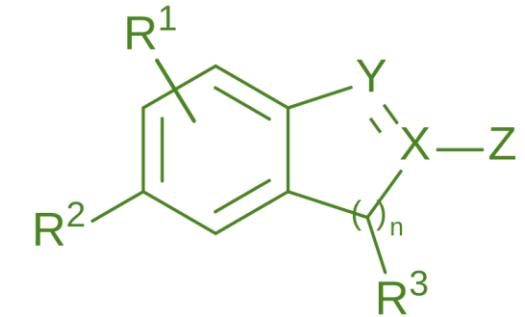
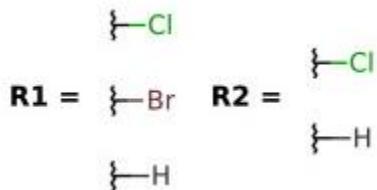
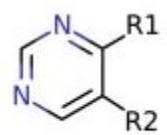
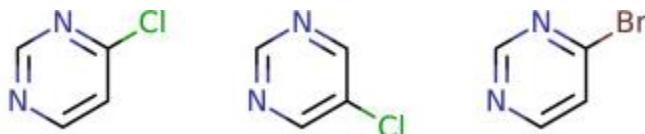
Al principio hubo **Markush structures**:

Representan varias estructuras posibles

Grupos R variables

Descripción general de una molécula con **ambigüedad** en algunas posiciones

Son comunes en patentes, y en libros de texto.



[https://es.wikipedia.org/wiki/Estructura\\_de\\_Markush](https://es.wikipedia.org/wiki/Estructura_de_Markush)



Eugene A. Markush

# REPRESENTACIÓN DE PATRONES: SMARTS

## SMARTS - A Language for Describing Molecular Patterns

Representación lexicográfica de partes de una molécula

Es una extensión de **SMILES**

Concepto similar al de **expresiones regulares** (regex) en texto.

[https://en.wikipedia.org/wiki/Regular\\_expression](https://en.wikipedia.org/wiki/Regular_expression)

**Reglas (las mismas que SMILES), y además:**

Representación de patrones para átomos:

- \* cualquier átomo

- a aromático

- A alifático ... hay más reglas para átomos

Representación de patrones para enlaces:

- ~ cualquier enlace

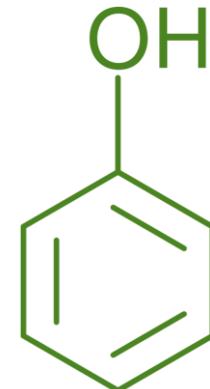
- @ cualquier enlace en un anillo

- / enlace dirigido “arriba”

- \ enlace dirigido “abajo”

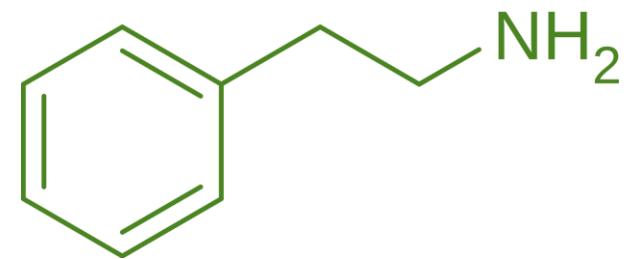
**Más información y reglas en:**

<https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>



**SMARTS: [OH]c1ccccc1**

hydroxyl-group attached to 6 aromatic carbons in a ring

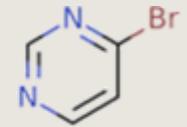
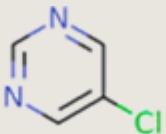
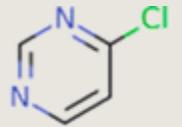
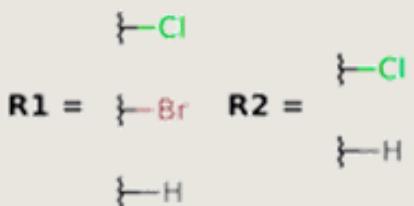
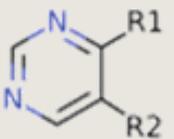


**SMARTS: NCCc1ccccc1**

Aliphatic nitrogen attached to 2 aliphatic carbons attached to 6 aromatic carbons in a ring

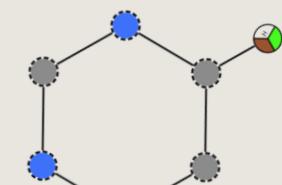
# FROM MARKUSH TO SMARTS

Original molecules

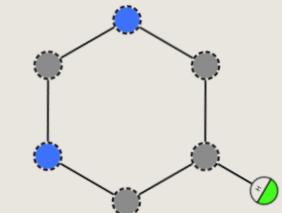


SMARTS PATTERNS

`n1cnc([Cl,Br,H])cc1`



`n1cncc([Cl,H])c1`



SMARTS.PLUS SmartView: <https://smarts.plus/>

# SMARTS – A LANGUAGE FOR DESCRIBING MOLECULAR PATTERNS

[Cl,Br,F,I]C([Cl,Br,F,I])([Cl,Br,F,I])CCC1=CC=CC=C1

Una representación SMILES es un patrón SMARTS válido

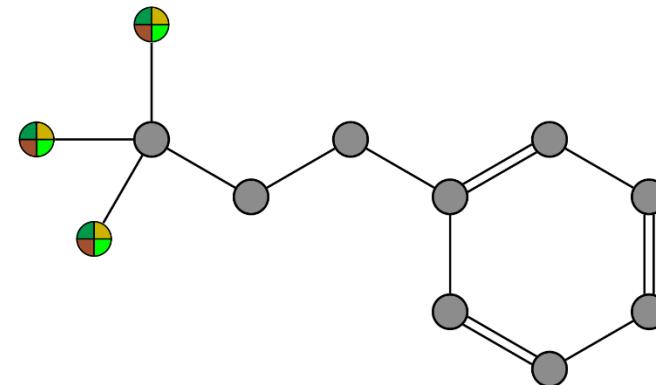
[OH]c1ccccc1 (phenol)

Patrones SMARTS simples

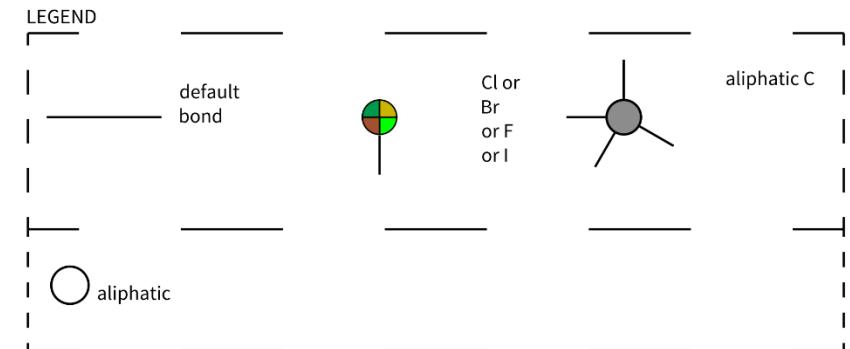
[C,N]1CCCCC1

[Cl,Br,F,I]C([Cl,Br,F,I])([Cl,Br,F,I])CCC1=CC=CC=C1

C-C=C-C=C~\*~[++]



Picture created by the SMARTSviewer [<https://smarts.plus/>].  
Copyright: ZBH - Center for Bioinformatics Hamburg.

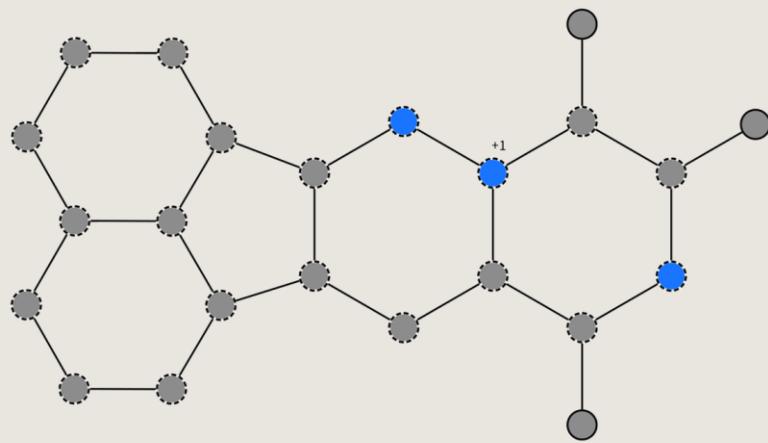


Referencias

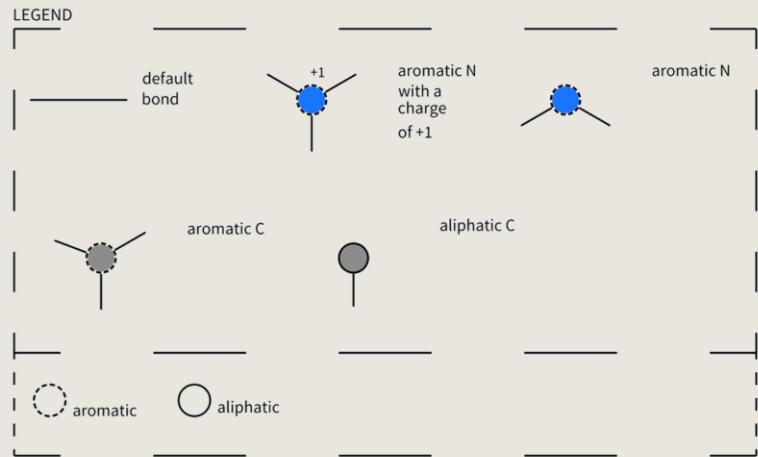
<https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>

<https://smarts.plus/>

# PATRONES SMARTS PARA BÚSQUEDAS



Picture created by the SMARTSviewer [<https://smarts.plus/>].  
Copyright: ZBH - Center for Bioinformatics Hamburg.



<https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>  
<https://smarts.plus/>

## Referencias

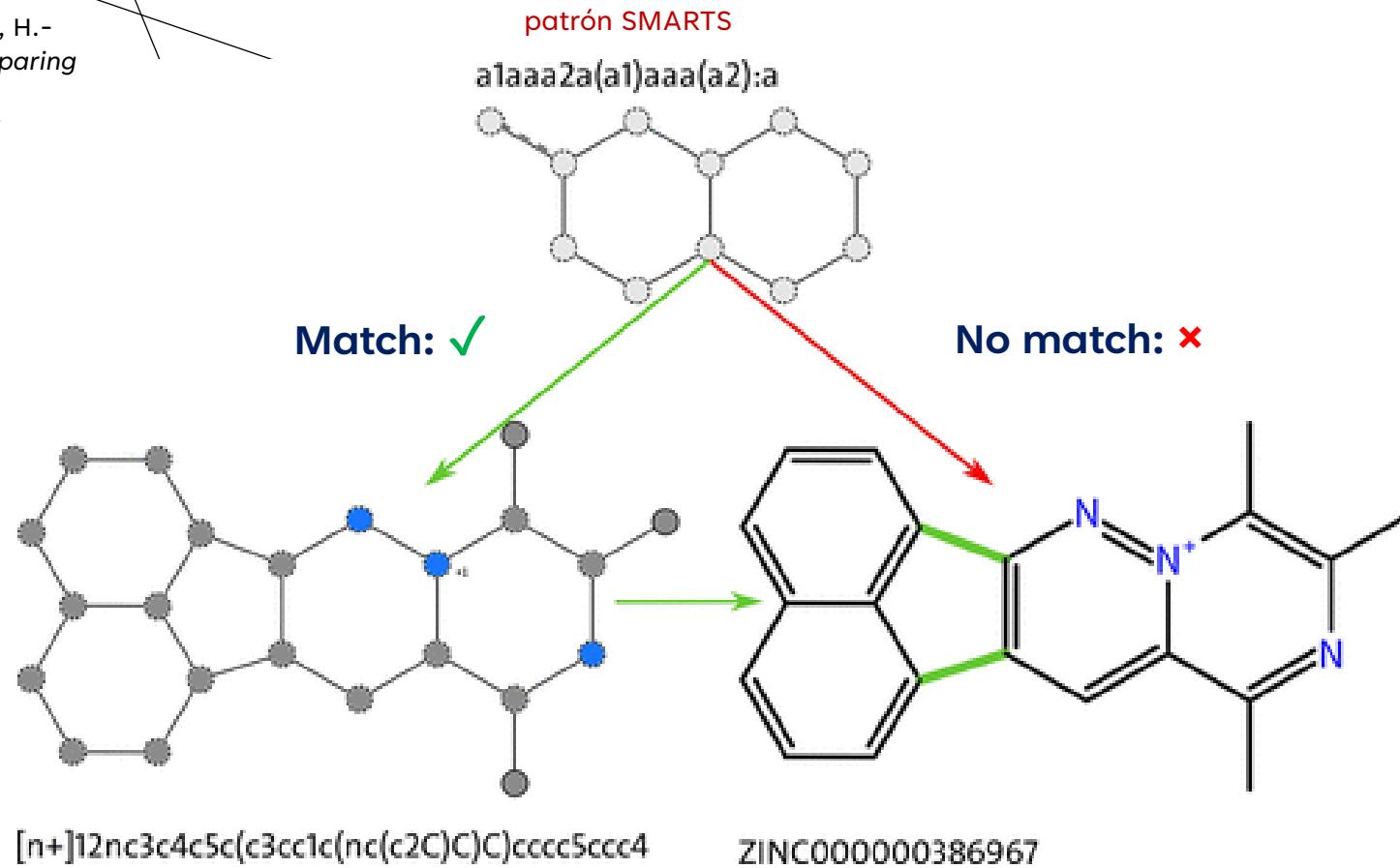
[n+]12nc3c4c5c(c3cc1c(nc(c2C)C)CCCC5CC4

C = carbono alifático

c = carbono aromático

# MATCHING SMARTS PATTERNS

Schmidt, R., Ehmki, E. S. R., Ohm, F., Ehrlich, H.-C., Mashychev, A., & Rarey, M. (2019). Comparing Molecular Patterns Using the Example of SMARTS: Theory and Algorithms. *Journal of Chemical Information and Modeling*. doi:10.1021/acs.jcim.9b00250



# TESTING AT SMARTS.PLUS

<https://smarts.plus/>



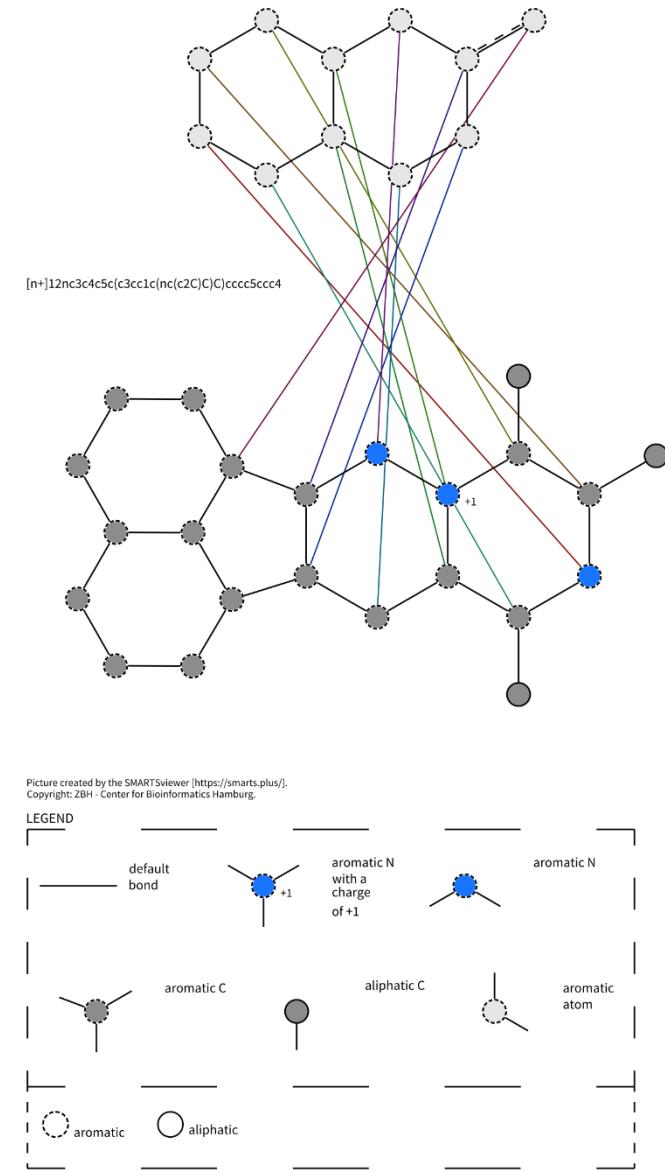
View   Compare   Search   Create

Compare two SMARTS expression with respect to subset relation (Does expression A match whenever B matches?) or similarity and receive a visualization of the node mapping.

SMARTS pattern:

SMARTS to compare:

[More Options](#)   [Go!](#)



[NX3,NX4+][CX4H]([\*])[OX3](=[OX1])[O,N]  
[NX3,NX4+][CX4H]([\*])[CX3](=[OX1])[O,N]

## SMARTS EXAMPLES

### Amino Acids

Generic amino acid: low specificity:

[NX3,NX4+][CX4H]([\*])[CX3](=[OX1])[O,N]

### Other interesting examples

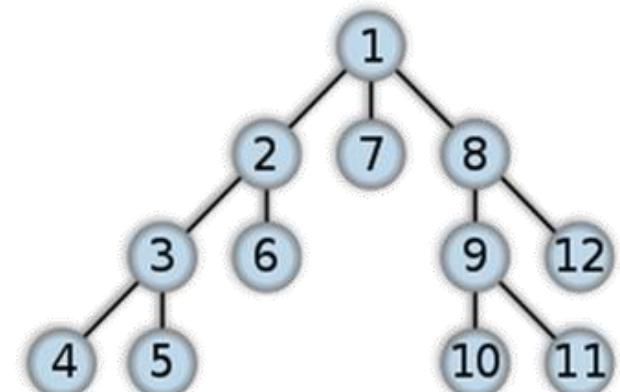
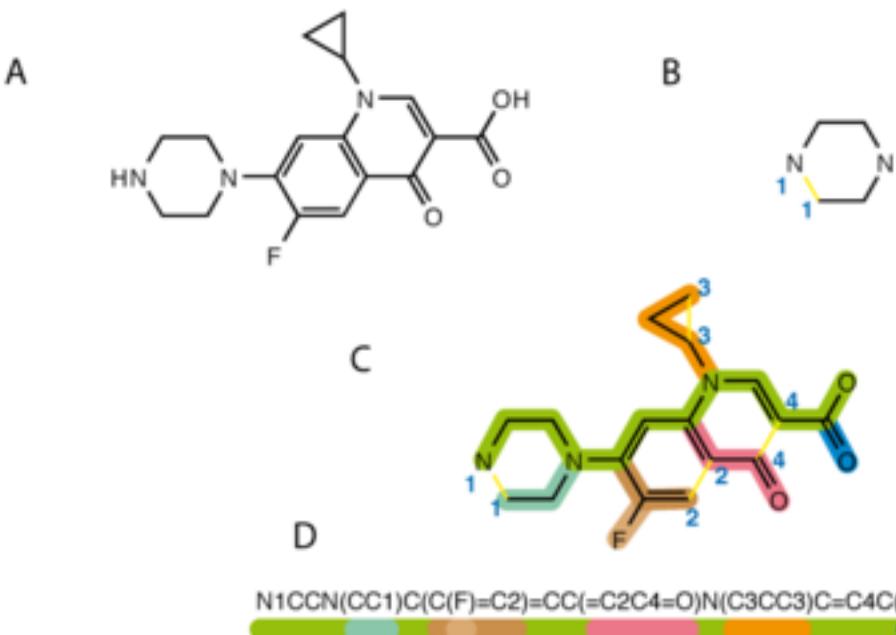
[https://www.daylight.com/dayhtml\\_tutorials/languages/smarts/smarts\\_examples.html](https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html)

# REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: SMILES

SMILES, relación con Teoría de Grafos

SMILES es una cadena de texto (ASCII)

Es el producto de escribir los símbolos  
(átomos) a medida que se recorre el grafo  
químico (la molécula) de modo *depth-first*



Order in which the nodes are expanded

Class Search algorithm

Data structure Graph

Worst case performance  $O(|V| + |E|)$  for explicit graphs traversed without repetition,  $O(b^d)$  for implicit graphs with branching factor  $b$  searched to depth  $d$

Worst case space complexity  $O(|V|)$  if entire graph is traversed without repetition,  $O(\text{longest path length searched})$  for implicit graphs without elimination of duplicate nodes

Depth-first Tree/Graph Traversal:

[http://en.wikipedia.org/wiki/Depth-first\\_search](http://en.wikipedia.org/wiki/Depth-first_search)

# CANONIZACIÓN DE MOLECULAS: ALGORITMO DE MORGAN

Canonización: Representar la conectividad de una molécula de manera uniforme

Una estructura con  $n$  átomos puede ser descripta de  $n!$  maneras diferentes

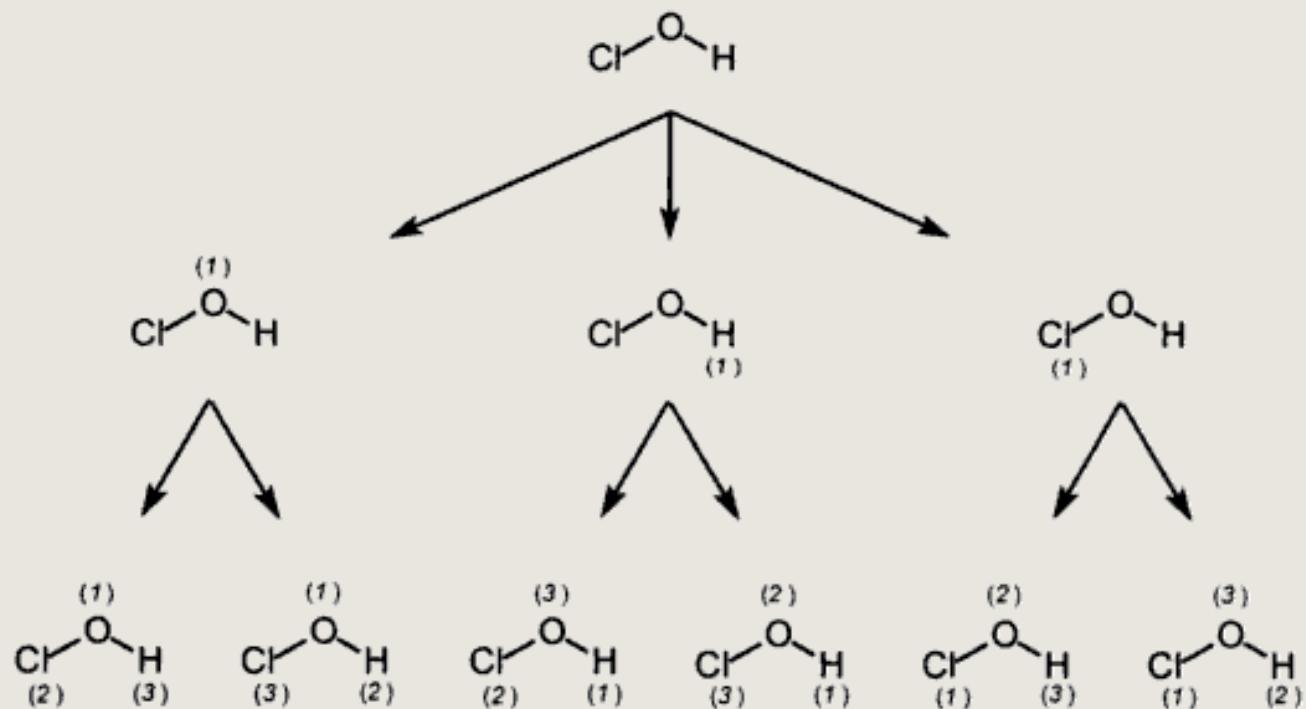


Figure 2-41. Six different possibilities for numbering the atoms in a hypochlorous acid molecule.

El algoritmo de Morgan es viejo pero lo vamos a usar para aprender el concepto de **canonización**!

Hay variantes nuevas!

Schneider N, Sayle RA, Landrum GA. Get Your Atoms in Order--An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J Chem Inf Model*. 2015 Oct;26(55)(10):2111-20. doi: 10.1021/acs.jcim.5b00543. Epub 2015 Oct 15. PMID: 26441310.

# CANONIZACIÓN: ALGORITMO DE MORGAN

**Paso 1:** clasificar átomos de acuerdo a conectividad (vecindad)

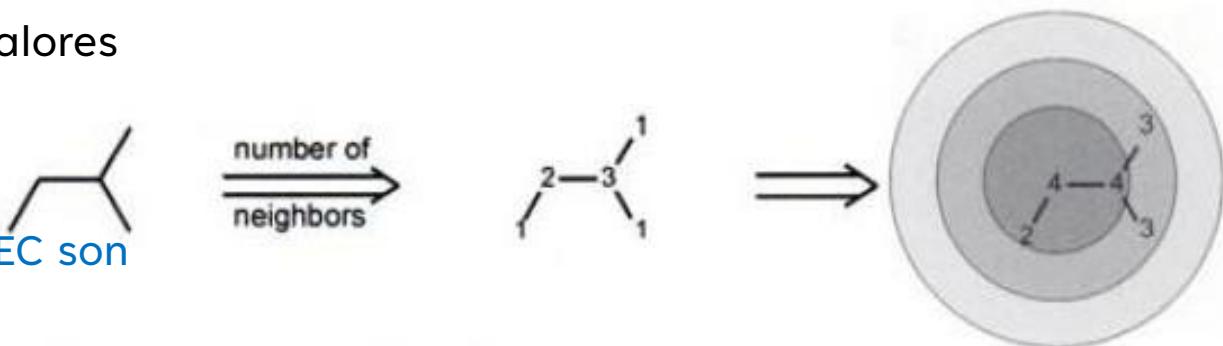
Estructuras conteniendo C, N, O, H y halógenos se clasifican en cuatro categorías dependiendo del número de enlaces (no H)

**Paso 2:** Iteraciones

En una segunda iteración los valores de conectividad de cada átomo se incrementan de acuerdo al de los vecinos siguiendo una serie de reglas:

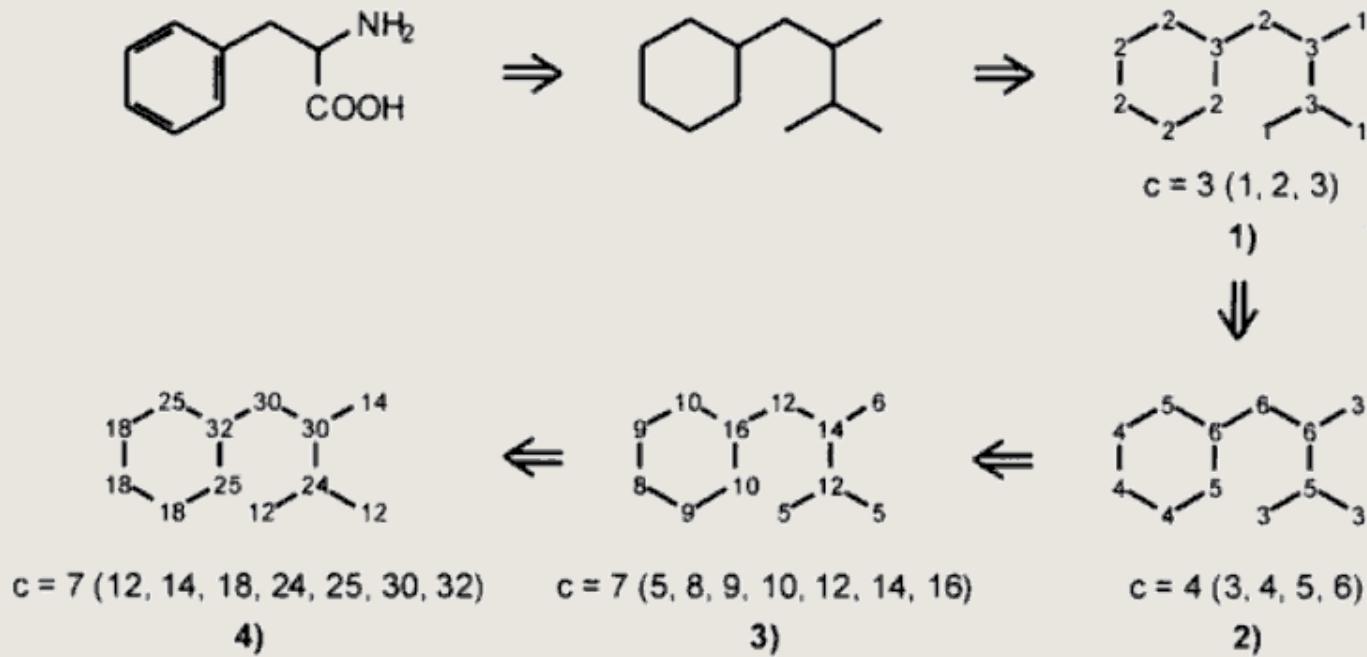
- Sumas (átomos internos) o transferencia de valores (átomos terminales)
- *Extended connectivity*

Las iteraciones siguen hasta que los valores de EC son iguales o menores a los de la iteración anterior



**Figure 2-43.** The EC value or the atom classification of each atom, respectively, is calculated by summing the EC values of the directly connected neighboring atoms of the former sphere (relaxation process).

# CANONIZACIÓN: ALGORITMO DE MORGAN

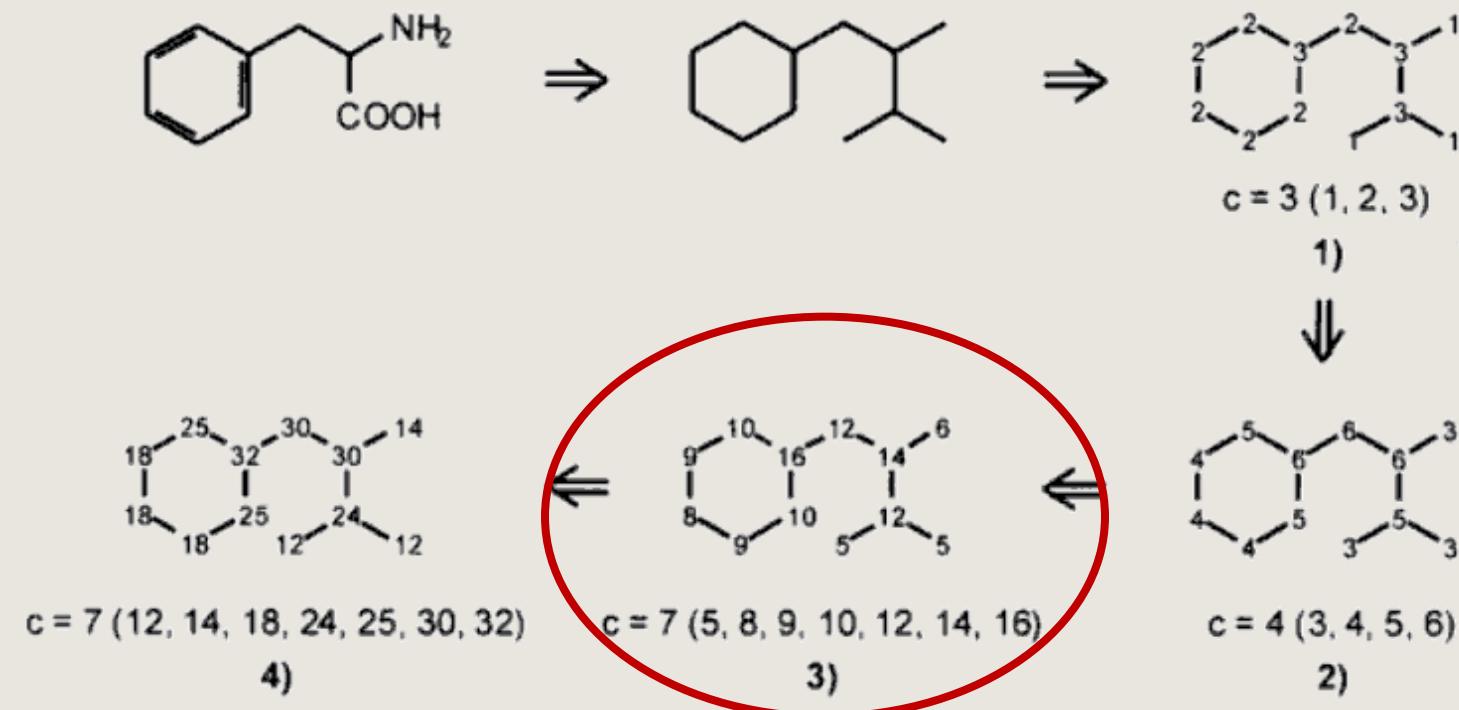


**Figure 2-44.** The EC values of the atoms of phenylalanine (without hydrogens) are calculated by considering the class values of the neighboring atoms. After each relaxation process, *c*, the number of equivalent classes (different EC values), is determined.

The process is repeated until the number of different EC values is lower than or equal to the number of different EC values in the previous iteration.

# CANONIZACIÓN: ALGORITMO DE MORGAN

Paso 3: Asignación de números de átomos únicos



**Figure 2-44.** The EC values of the atoms of phenylalanine (without hydrogens) are calculated by considering the class values of the neighboring atoms. After each relaxation process,  $c$ , the number of equivalent classes (different EC values), is determined.

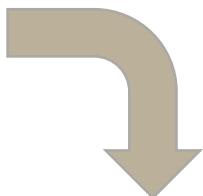
Se comienza por el paso en el que se obtiene el mayor EC por primera vez.

El atomo número 1 es el que tiene el mayor valor de EC en este paso.

El atomo 2 es el que sigue en la secuencia de valores EC.

# CANONIZACIÓN DE MOLECULAS

El que implementa  
RDKit (Python)



Schneider N, Sayle RA, Landrum GA. Get Your Atoms in Order--An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J Chem Inf Model.* 2015 Oct;26;55(10):2111-20. doi: 10.1021/acs.jcim.5b00543. Epub 2015 Oct 15. PMID: 26441310.

Krotko DG. Atomic ring invariant and Modified CANON extended connectivity algorithm for symmetry perception in molecular graphs and rigorous canonicalization of SMILES. *J Cheminform.* 2020 Aug 20;12(1):48. doi: 10.1186/s13321-020-00453-4. PMID: 33431026; PMCID: PMC7439248.

El algoritmo es de 1965! Es viejo!

Hay moléculas problemáticas que no son fáciles de canonizar.

El problema general que intenta resolver es el de  
**Canonizacion de Grafos**

- Es un problema computacional complejo
- Relacionado con problemas de isomorfismo de grafos
- Hay muchas otras maneras (algoritmos) de resolverlos:  
[http://en.wikipedia.org/wiki/Graph\\_canonization](http://en.wikipedia.org/wiki/Graph_canonization)

En resumen:

Después de aplicar un método de canonización

# REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: INCHI

InChI – *International Chemical Identifier*

Introducido recientemente (2005) por IUPAC  
(International Union of Pure and Applied Chemistry)

Heller SR, McNaught A, Pletnev I, Stein S,  
Tchekhovskoi D. InChI, the IUPAC International  
Chemical Identifier. J Cheminform. 2015 May  
30;7:23. doi: 10.1186/s13321-015-0068-4.  
PMID: 26136848; PMCID: PMC4486400.

## Objetivos

Establecer un identificador (nomenclatura, etiqueta) **único** y  
**no propietario** para cada molécula

Que pueda ser utilizado tanto en medios impresos como  
electrónicos y que facilite la búsqueda de compuestos

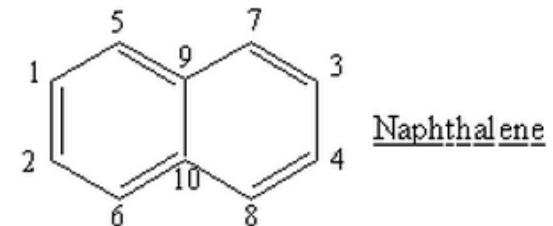
# REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: INCHI

Formato de un identificador InChI

Es una cadena de texto (ASCII) compuesta por *segmentos* (layers) separada por *delimitadores* (/)

Cada capa contiene distintos tipos de información estructural

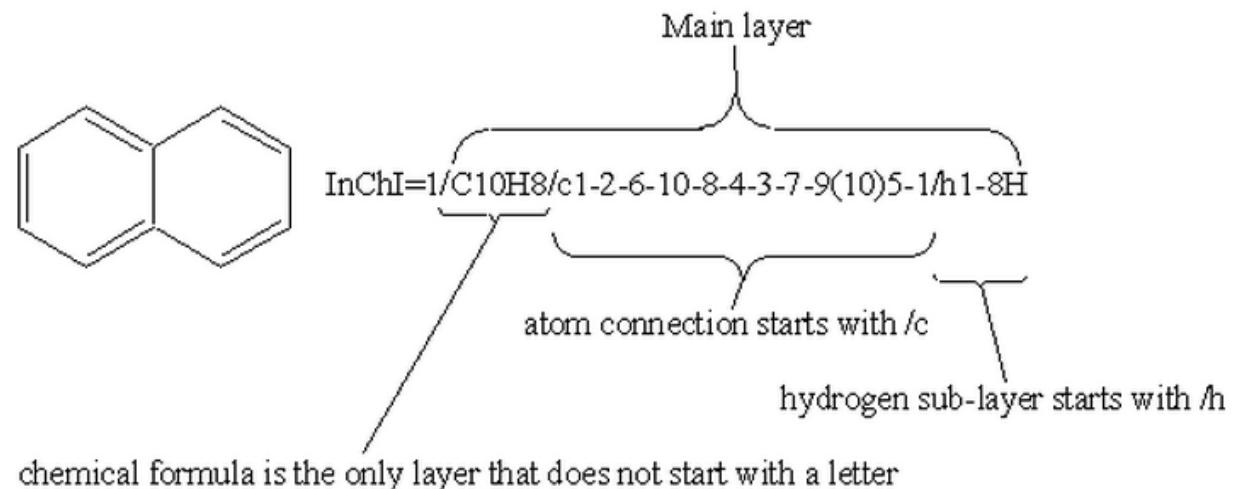
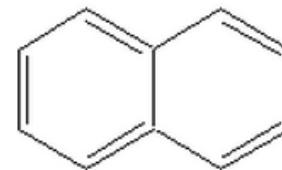
Los números dentro de una capa representan la numeración canónica de los átomos de la primera capa (fórmula) excepto los hidrógenos.



## Ejemplos:

Agua: InChI = 1/H2O/h1H2

Benceno: InChI = 1/C6H6/c1-2-4-6-5-3-1/h1-6H



# INCHI IDENTIFIER: MAIN LAYER

InChI = 1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3.(H,11,12)

Conectividad

Fórmula empírica

Hidrógenos



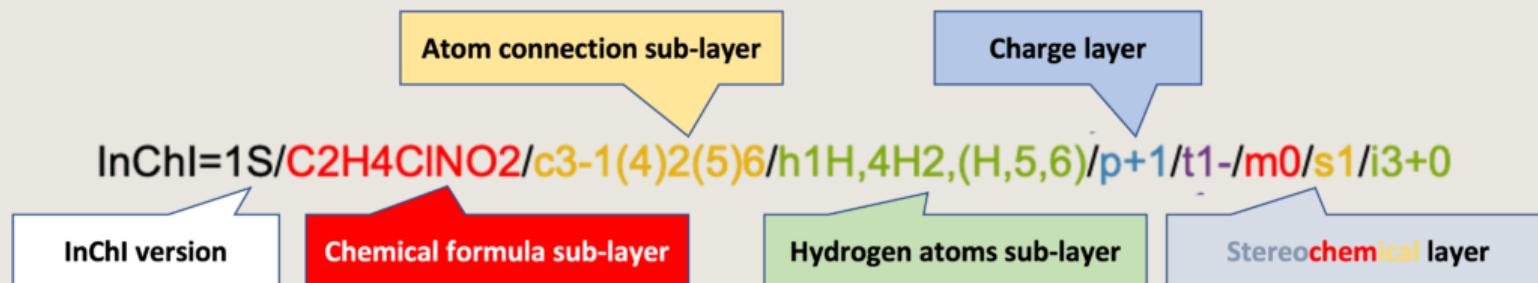
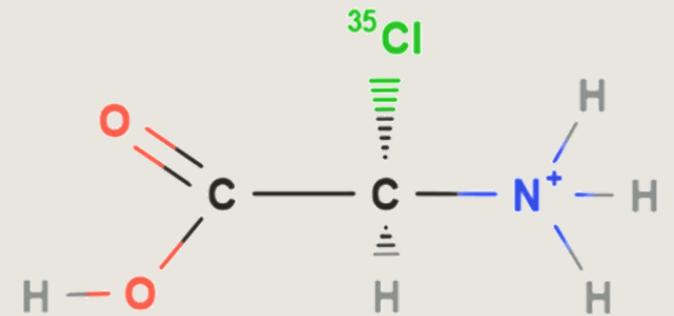
InChI = 1S/H2O/h1H2

# REPRESENTACIÓN DE COMPUESTOS: INCHI

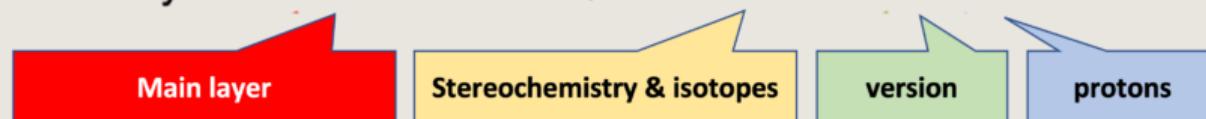
Representa la información en capas (layers)

Permiten elegir el nivel de detalle que uno quiere incluir

Sólo la capa principal es mandatoria



InChIKey=UWPWWENWLZPQGU-WRFRXMDISA-O

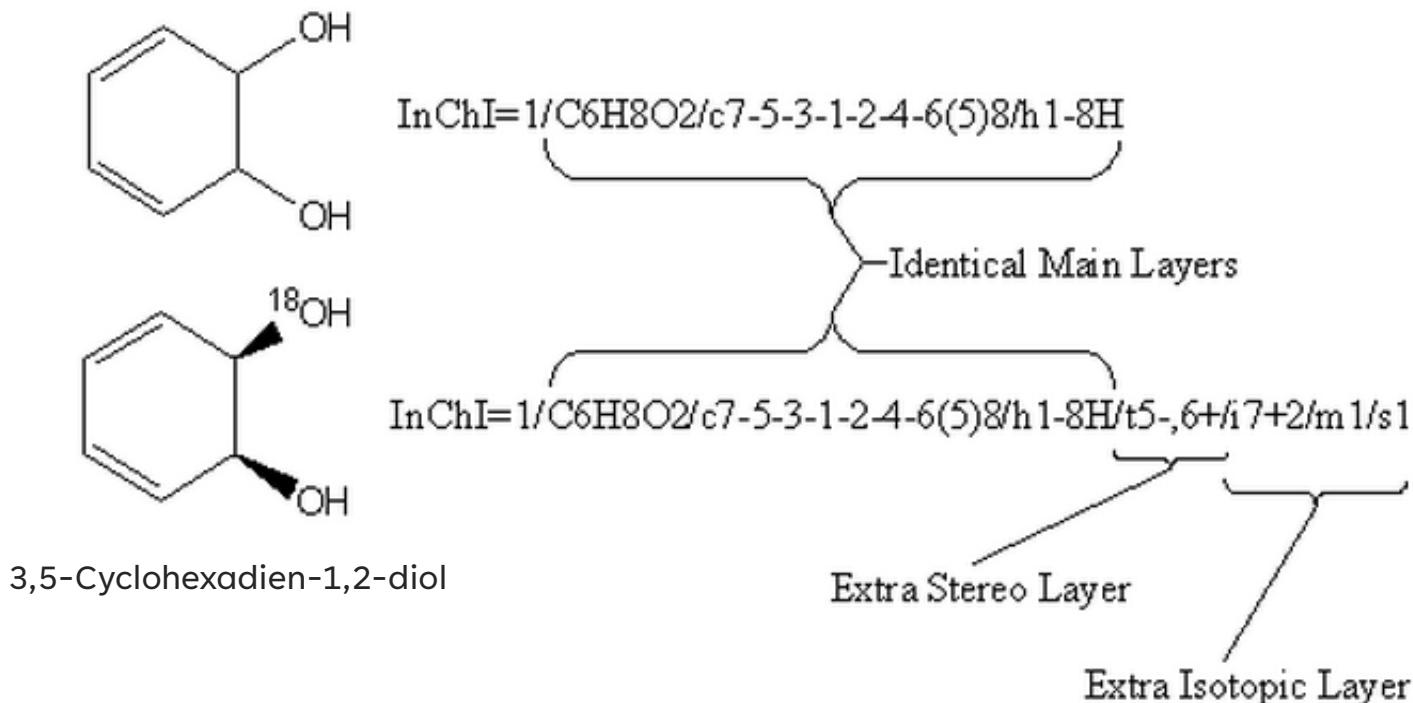


Tomado de: InChI Trust,  
<https://www.inchi-trust.org/>

# REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: INCHI

Si dos InChIs son iguales, los compuestos también lo son.

Pero los compuestos pueden estar representados con diferente nivel de detalle



# INCHI VS SMILES



Caffeine

Tomado de: InChI Technical FAQ  
<https://www.inchi-trust.org/technical-faq-2>

**Valid SMILES** for Caffeine (not complete)

```
[c]1([n+]([CH3])[c]([c]2([c]([n+]1[CH3])[n][cH][n+]2[CH3]))[O-])[O-]  
CN1C(=O)N(C)C(=O)C(N(C)C=N2)=C12  
Cn1cnc2n(C)c(=O)n(C)c(=O)c12  
Cn1cnc2c1c(=O)n(C)c(=O)n2C  
O=C1C2=C(N=CN2C)N(C(=O)N1C)C  
CN1C=NC2=C1C(=O)N(C)C(=O)N2C
```

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

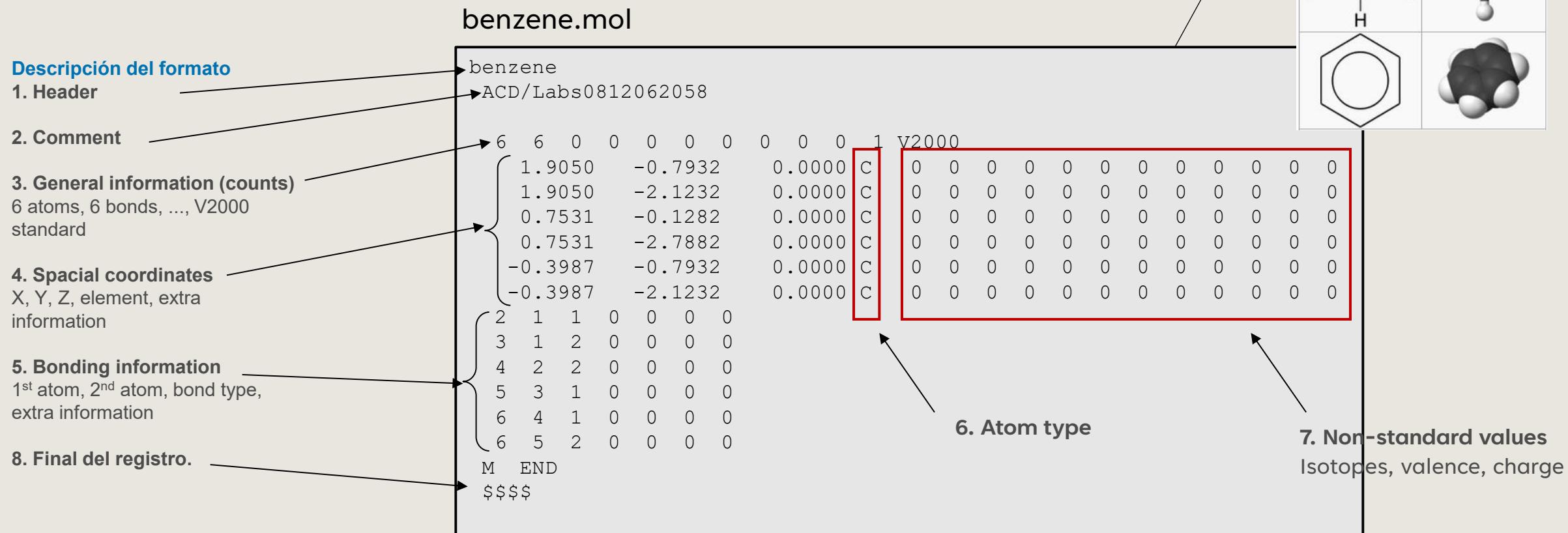
**InChI Key:** RYYVLZVUVIJVGH-UHFFFAOYSA-N

# REPRESENTACIÓN DE COMPUESTOS: MOLFILES

MDL, Molfile | Formato creado por MDL (ahora Symyx)

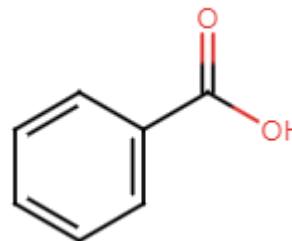
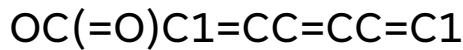
Contiene información sobre: Átomos, enlaces, conectividad y *coordenadas espaciales*

Permite representar moléculas tanto en **2D** como en **3D**



# MOLFILES: BOND BLOCK

Anatomy of a MOL file  
ChemInformatics 2017 (LibreTexts Chemistry)



		chemdraw-Dec-2016.cdx				ChemDraw12011615112D			
First atom row number	Second atom row number	0	0	0	0	0	0	0	0
		-0.0000	0.00	-0.8250	0.00	-1.2375	0.00	0.0000	0.00
1	2	2	3	2	1	0	0	0	0
2	3	3	4	2	0	0	0	0	0
3	4	4	5	1	0	0	0	0	0
4	5	5	6	2	0	0	0	0	0
5	6	6	1	2	0	0	0	0	0
6	1	7	5	1	0	0	0	0	0
5	7	7	8	1	0	0	0	0	0
7	8	8	9	2	0	0	0	0	0
7	9	9							
M END		0.7145	1.2375	0.00	0.00	0.00	0.00	0.00	0.00

Bond  
type

Bond  
stereochemistry

# REPRESENTACIÓN DE COMPUESTOS: SDF FILES

NGC00015959-03.sdf

MOLFILE

```
NCGC00015959-03
Marvin 07111412562D

25 30 0 0 0 0          999 V2000
 3.4098 -1.3130 0.0000 N 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.8329 -1.3130 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 3.4098 -2.1380 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.1248 -2.5436 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 2.6948 -2.5436 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.8329 -2.1380 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 4.1248 -0.8937 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 5.5547 -0.8937 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

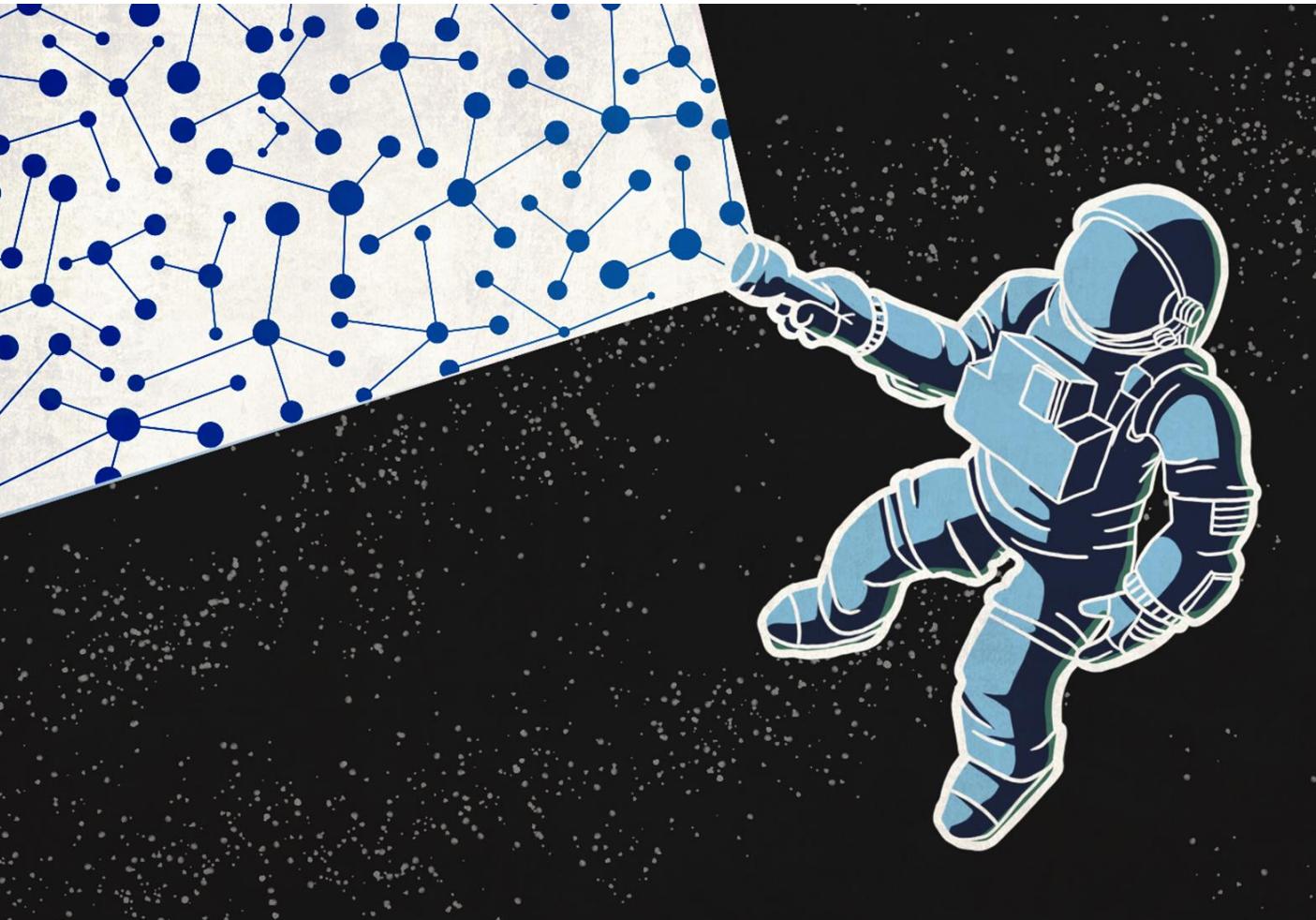
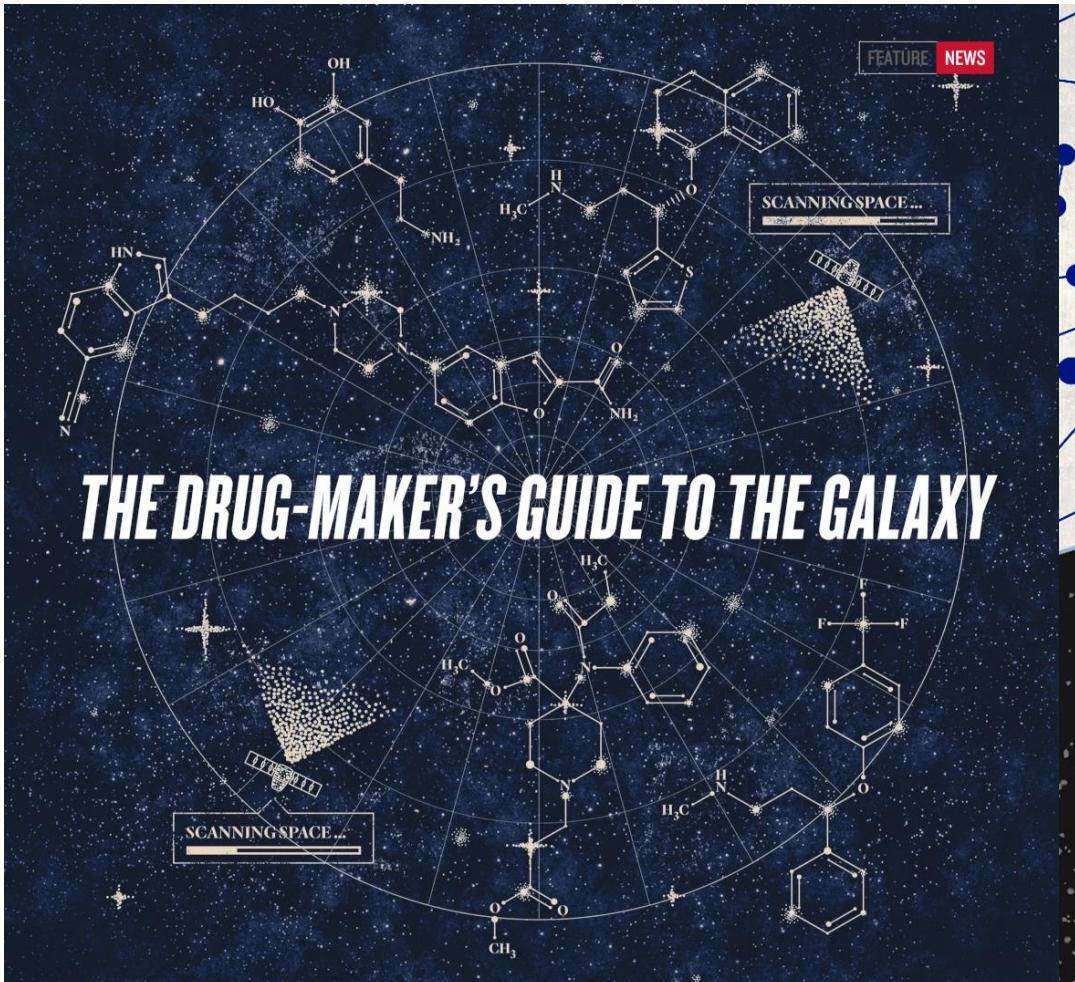
```
1 3 1 0 0 0 0
1 7 2 0 0 0 0
1 25 1 0 0 0 0
2 7 1 0 0 0 0
2 6 2 0 0 0 0
2 8 1 0 0 0 0
3 4 2 0 0 0 0
3 5 1 0 0 0 0
4 13 1 0 0 0 0
4 6 1 0 0 0 0
5 9 1 0 0 0 0
:
M CHG 1 1 1
M END
```

```
> <Formula>
C20H14NO4
> <FW>
332.3289
> <DSSTox_CID>
25204
> <Active>
1
```

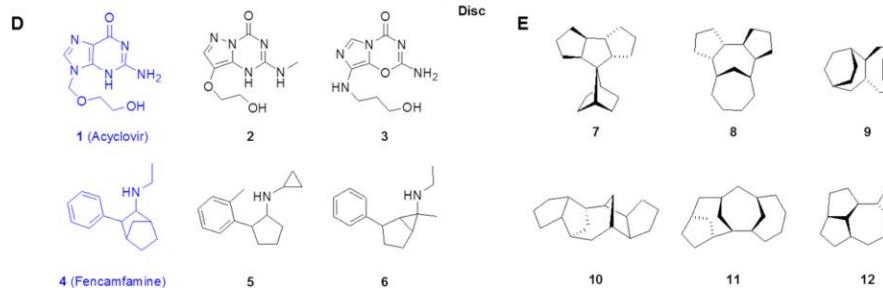
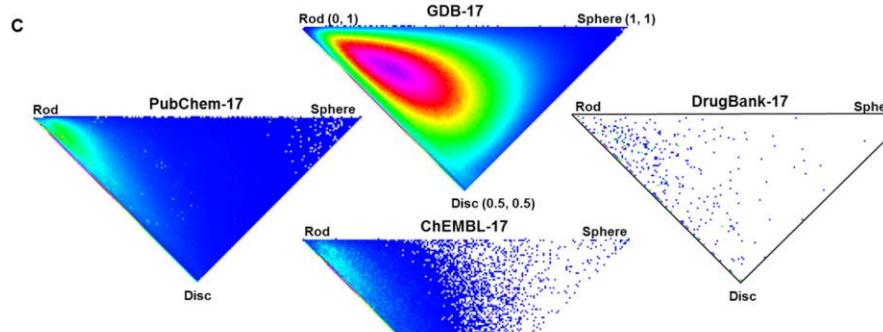
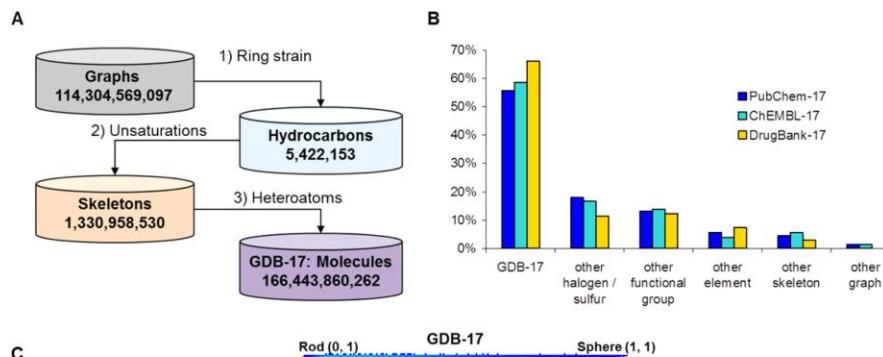
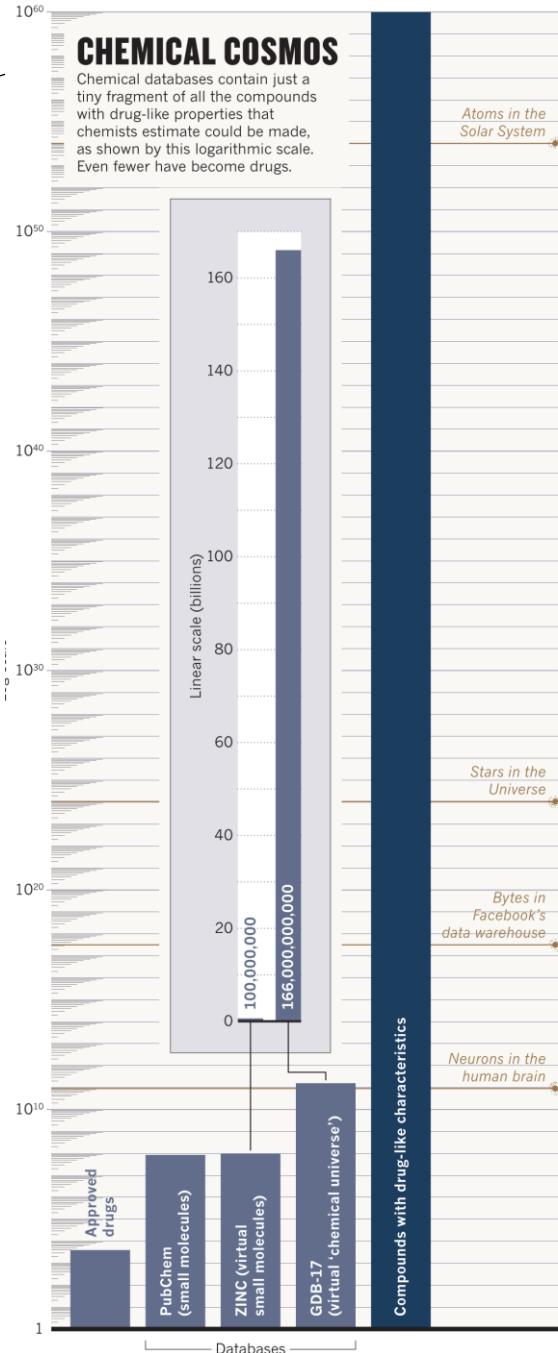
Anotaciones

# QUE ES EL ESPACIO QUIMICO?

Mullard A. The drug-maker's guide to the galaxy. Nature. 2017  
549(7673):445-447. doi: 10.1038/549445a. PMID: 28959982.



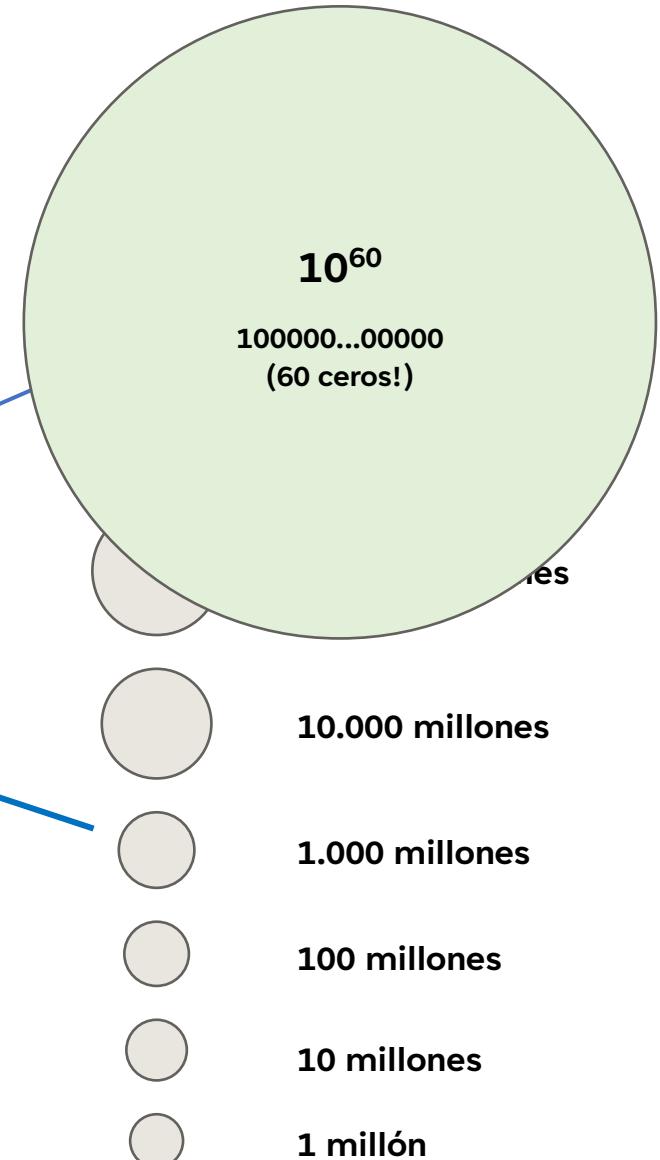
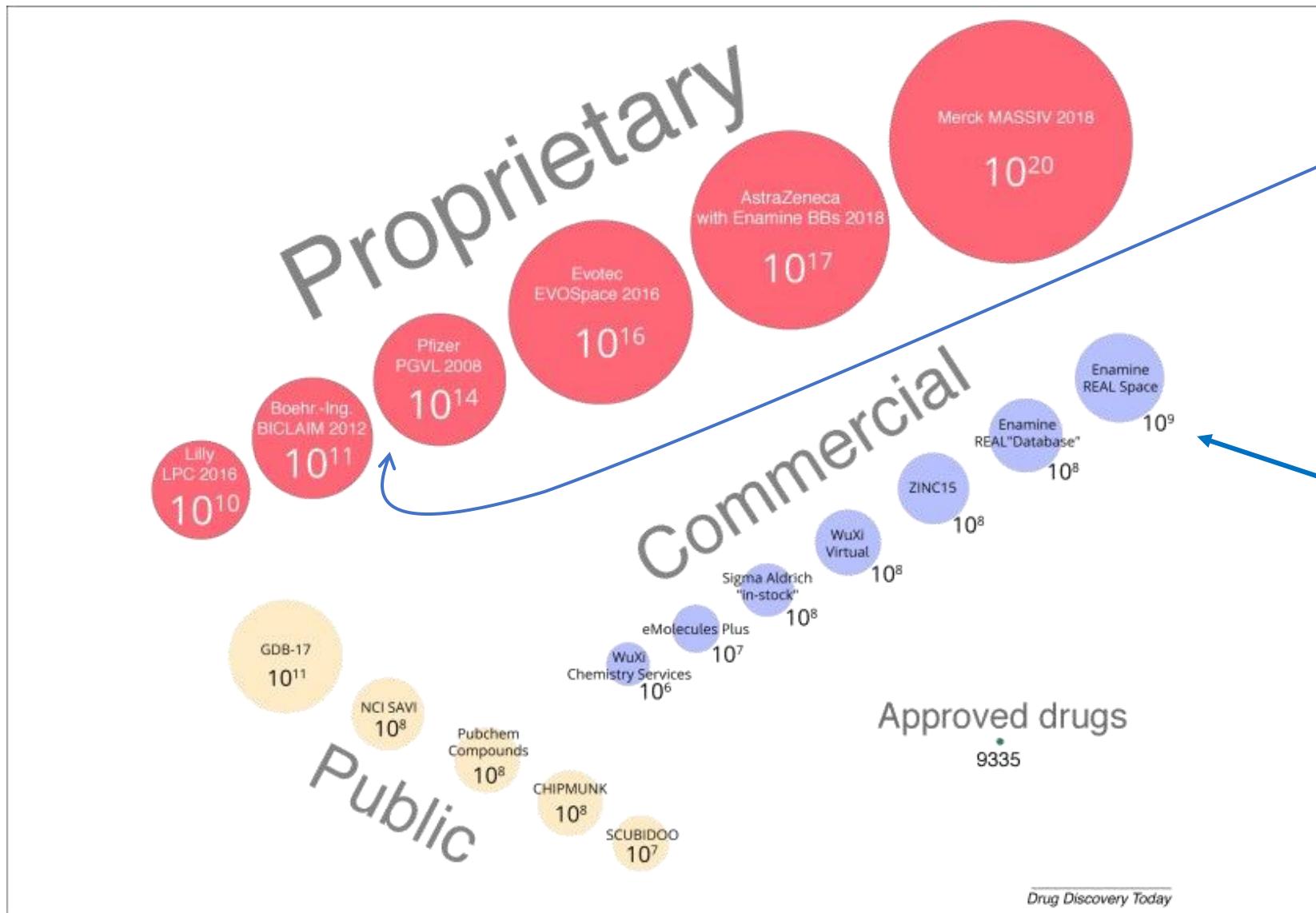
# THE CHEMICAL SPACE PROJECT



Reymond JL. Chemical space as a unifying theme for chemistry. J Cheminform. 2025 Jan 16;17(1):6. doi: 10.1186/s13321-025-00954-0. PMID: 39825400; PMCID: PMC11740331.

Reymond JL. The chemical space project. Acc Chem Res. 2015 Mar 48(3):722-30. doi: 10.1021/ar500432k. PMID: 25687211.

# CUÁN GRANDE ES EL ESPACIO QUÍMICO?



Gorse, A.-D. (2006). Diversity in Medicinal Chemistry Space. *Current Topics in Medicinal Chemistry*, 6(1), 3–18. doi:10.2174/156802606775193310

Hoffmann T, Gastreich M. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov Today*. 2019

# CHEMICAL DATABASES

- **PubChem, NCBI** | <https://pubchem.ncbi.nlm.nih.gov/>
  - repositorio abierto de información sobre moléculas y sus actividades biológicas
- **ChEMBL, EBI** | <https://www.ebi.ac.uk/chembl/>
  - Repositorio abierto de bioactividades de moléculas, extraídas de la literatura
- **ChemSpider, Royal Society of Chemistry** | <http://www.chemspider.com/>
  - Free chemical structure database providing access to >130 million structures from hundreds of data sources.
- **NIST Chemistry Web Book** | <https://webbook.nist.gov/>
- **DrugBank** | <http://www.drugbank.ca/>
- **GDB-17, GDBMedChem**
- **Zinc Databases** | <https://zinc.docking.org/>
  - commercially-available compounds for virtual screening



# ZINC20

# PUBCHEM

Open chemistry database at the National Institutes of Health (NIH).

Data is submitted by Academic Labs, Governmental Agencies, Chemical Supply and Pharmaceutical Companies, Journal Publishers, Individual Researchers

Mostly contains small molecules, but also larger molecules such as nucleotides, carbohydrates, lipids, peptides, and chemically-modified macromolecules.

Has information on chemical structures, identifiers, chemical and physical properties, biological activities, patents, health, safety, toxicity data, etc.

PubChem

About Docs Submit Contact

Search PubChem

COMPOUND SUMMARY

(1) Search with what you have

Aspirin

(2) Then, navigate to the information you need

PubChem CID 2244

Structure

2D 3D Crystal

Chemical Safety

! Irritant

Laboratory Chemical Safety Summary (LCSS) Datasheet

Molecular Formula  $C_9H_8O_4$   
 $CH_3COOC_6H_4COOH$

Synonyms aspirin  
ACETYLSALICYLIC ACID  
50-78-2  
2-Acetoxybenzoic acid  
2-(Acetoxy)benzoic acid

Cite Download

CONTENTS

- Title and Summary
- 1 Structures
- 2 Names and Identifiers
- 3 Chemical and Physical Properties
- 4 Spectral Information
- 5 Related Records
- 6 Chemical Vendors
- 7 Drug and Medication Information
- 8 Pharmacology and Biochemistry
- 9 Use and Manufacturing
- 10 Identification
- 11 Safety and Hazards
- 12 Toxicity
- 13 Associated Disorders and Diseases
- 14 Literature
- 15 Patents
- 16 Interactions and Pathways
- 17 Biological Test Results

# PUBCHEM SEARCH

**Explore**  
Quickly find chemicals

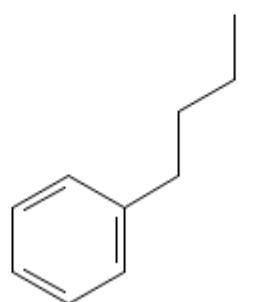
Try aspirin EGFR C9H8O4  Use Entrez

 Draw Structure

DRAW STRUCTURE

Broadband ▾ SMILES ▾ C1=CC=CC=C1CCCC

New Udo Cln Sfy Del Qry 



Export MDL Molfile ▾ Done

Hydrogen Keep AsIs ▾ Help

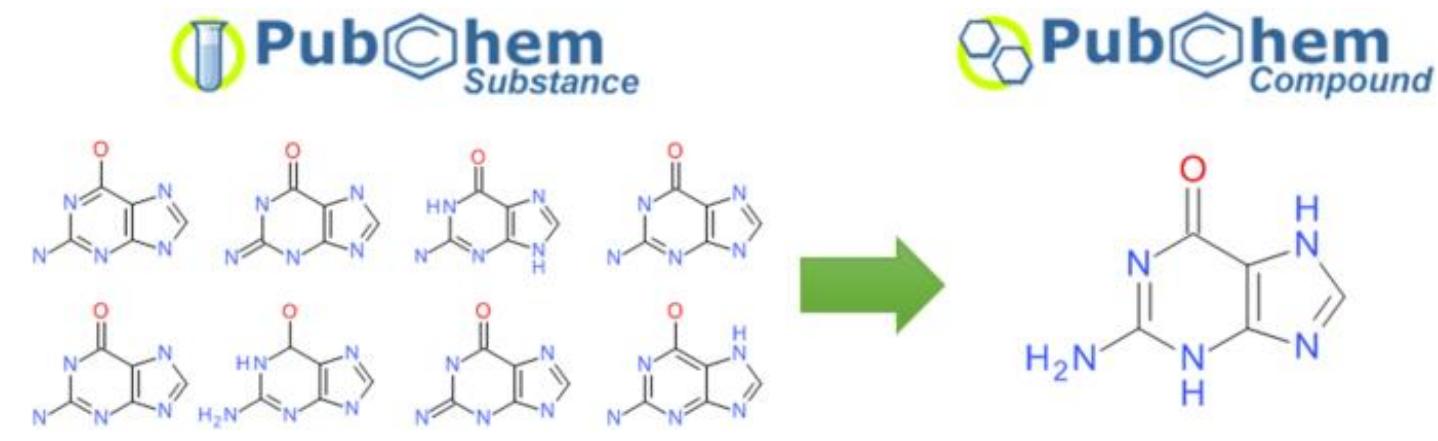
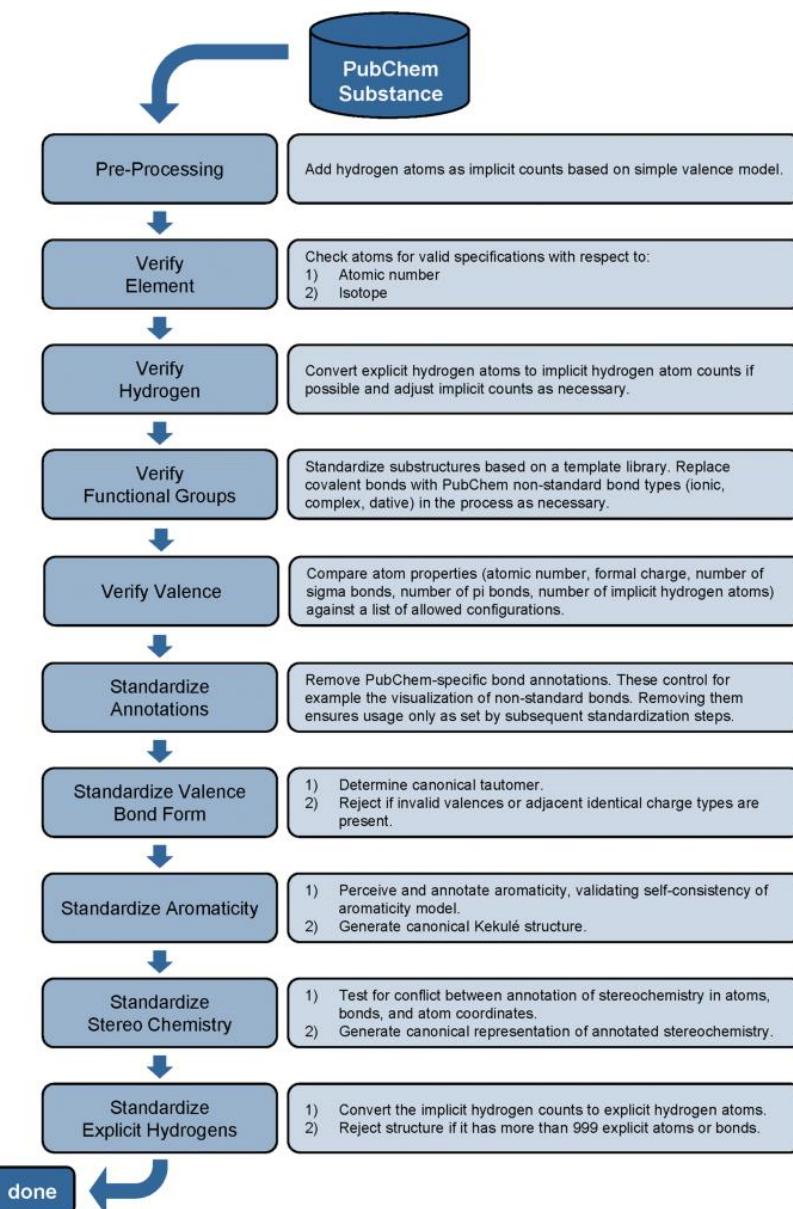
Import Choose File No file chosen

Search for This Structure

# PUBCHEM DATA

Collection	Live Count	Description	Last Updated
Periodic Table of Elements	118	Interactive periodic table with up-to-date element property data collected from authoritative sources	May 9, 2025
Compounds	119,097,936	Unique chemical structures extracted from contributed PubChem Substance records	May 10, 2025
Substances	329,647,545	Information about chemical entities provided by PubChem contributors	May 10, 2025
BioAssays	1,768,327	Biological experiments provided by PubChem contributors	May 10, 2025
Pathways	250,769	Interactions between chemicals, genes, and proteins	May 9, 2025
Proteins	248,623	Proteins in PubChem including those found in BioAssays, Pathways, and Patents	May 9, 2025
Genes	166,889	Genes in PubChem including those found in BioAssays, Pathways, and Patents	May 9, 2025
Taxonomies	108,432	Organisms in PubChem including those found in BioAssays, Pathways, and Patents	May 9, 2025
Patents	53,726,453	Patents with links in PubChem	May 9, 2025
Literature	43,177,340	Scientific publications with links in PubChem	May 5, 2025
Cell Lines	2,009	Information about cell lines	May 9, 2025
BioActivities	296,622,527	Biological activity data points reported in PubChem BioAssays	May 9, 2025
Data Classifications	76	Browse the distribution of PubChem data among nodes in the hierarchy of interest	May 7, 2025
Data Sources	1,047	Organizations contributing data to PubChem	May 11, 2025

# PUBCHEM CHEMICAL STRUCTURE STANDARDIZATION



Hähnke VD, Kim S, Bolton EE. PubChem chemical structure standardization. *J Cheminform.* 2018 10(1):36.  
doi: 10.1186/s13321-018-0293-8. PMID: 30097821; PMCID: PMC6086778.

# REPRESENTACIÓN DE COMPUESTOS QUÍMICOS: GRAFOS

Un grafo es una estructura *abstracta* que contiene *nodos* conectados con *aristas* (o *arcos*)

“Los grafos son redes (*networks*) de puntos y líneas”

En inglés: *nodes*, *edges*

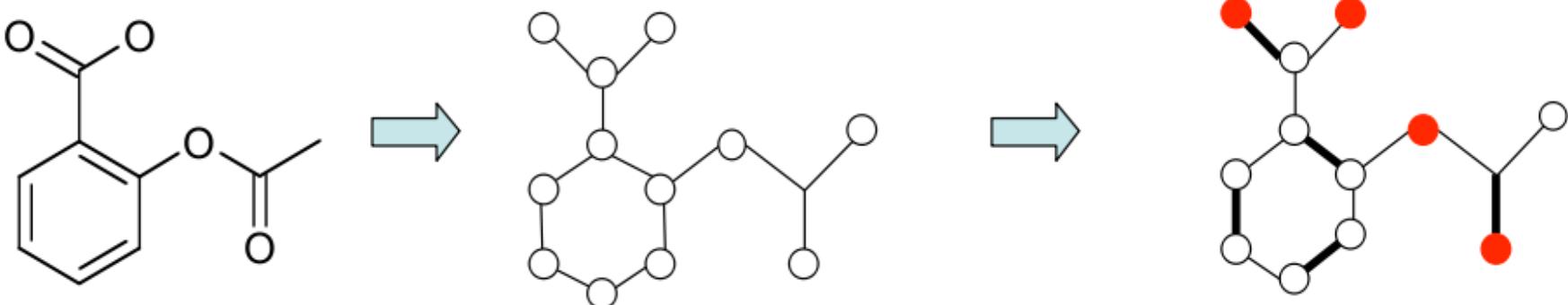
Moléculas químicas pueden representarse como *grafos*:

Los átomos como nodos

Los enlaces como aristas

Se pueden asociar propiedades a cada nodo (ej número atómico), y a cada arista (ej número y/o tipo de enlace)

En el grafo final pueden entonces distinguirse distintos tipos de nodos y aristas



# UN DESVÍO: HISTORIA DE LOS GRAFOS

El problema de **los 7 puentes de Königsberg**.

La ciudad de Königsberg se encuentra dividida por el río Pregel

Incluye 2 islas que se conectan con tierra mediante 7 puentes

**El problema:** Encontrar un camino a través de la ciudad que cruce cada puente una sola vez. Hay que cruzar todos los puentes. Sólo se puede acceder a las islas cruzando un puente.

En 1735 Leonard Euler demostró que el problema no tiene solución.

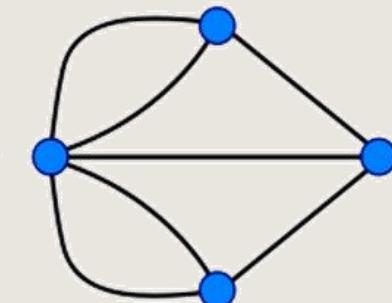
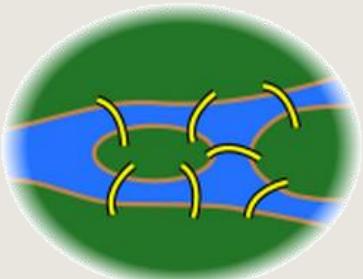
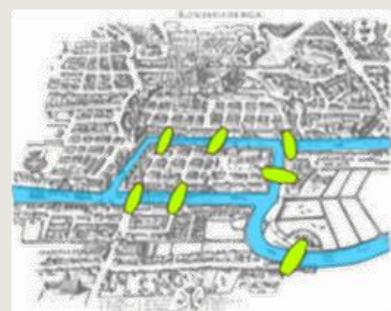
**El razonamiento:**

La elección del camino dentro de cada porción de tierra era **irrelevante**

La única característica de la ruta elegida importante era la secuencia de puentes cruzados



Leonard Euler (1707-1783)



**Abstracción del problema:**

En una lista de porciones de tierra (**nodos**)

Y una lista de puentes (**aristas**)

Sólo la información de **conectividad** era relevante!

Tomado de Wikipedia

[https://en.wikipedia.org/wiki/Seven\\_Bridges\\_of\\_K%C3%B6nigsberg](https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg)

# GRAFOS: PROPIEDADES Y OPERACIONES

## Propiedades de los grafos:

Grado de conectividad de los nodos (degree)

Direccionalidad de las aristas

Intensidad (sentido vectorial) de cada arista

Las aristas pueden tener asociado un valor numérico (peso, largo, costo)

## Posibilidad de identificar los nodos

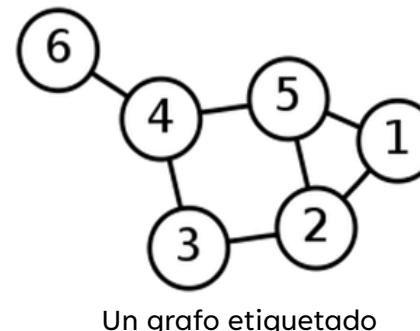
Son elementos de un conjunto

Grafos etiquetados (labeled)

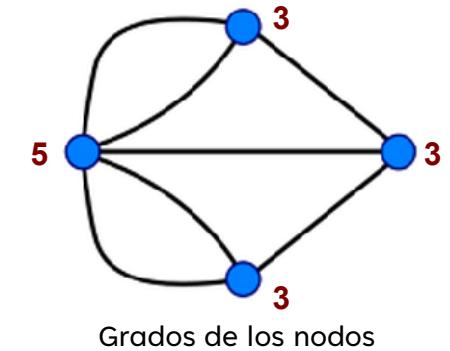
vs no-etiquetados (unlabeled)

## Operaciones con grafos (algunos ejemplos):

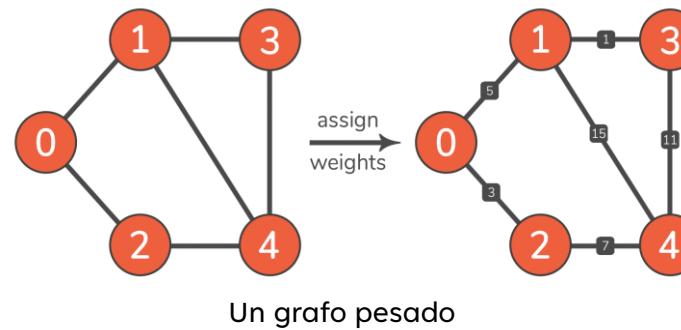
Complementación, Unión, Suma, Intersección, Diferencia, ...



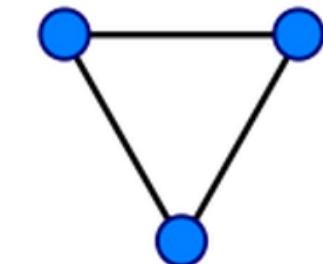
Un grafo etiquetado



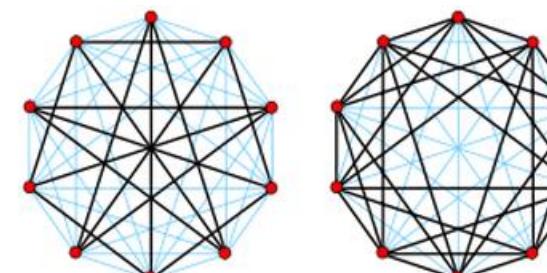
Grados de los nodos



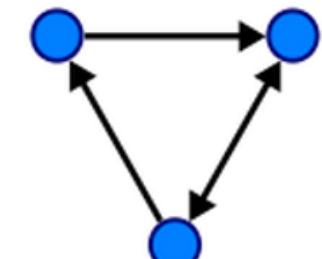
Un grafo pesado



Grafo simple o regular



Un grafo y su complemento



Grafo dirigido (red)

# PROBLEMA: ENCONTRAR MOLÉCULAS IGUALES

Problema  
frecuente en  
química

Si representamos moléculas como **grafos**:

- dos moléculas son la misma si es posible redibujar una de ellas de manera que se vea idéntica a la otra: **Isomorphic graphs**

Problema visualmente interesante, pero la solución es obvia: **solo la conectividad es relevante!**

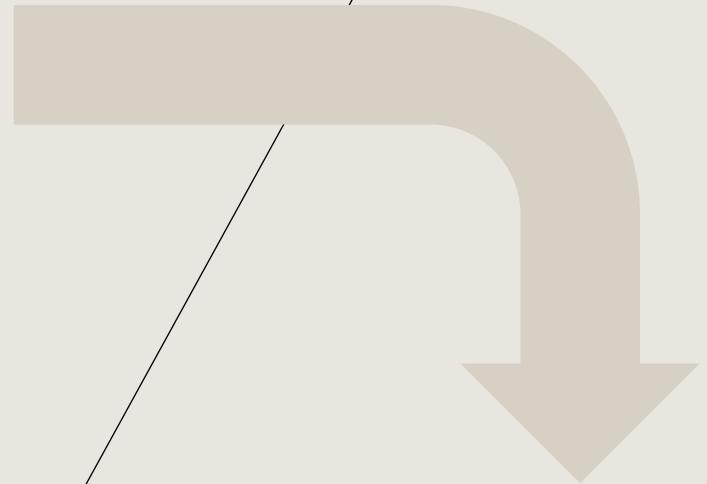
$G_1$ : nodos = { $u_1, v_1, w_1, x_1$ }

aristas = {  $\{u_1, v_1\}, \{u_1, w_1\}, \{u_1, x_1\}, \{v_1, x_1\}, \{v_1, w_1\}, \{x_1, w_1\}$  }

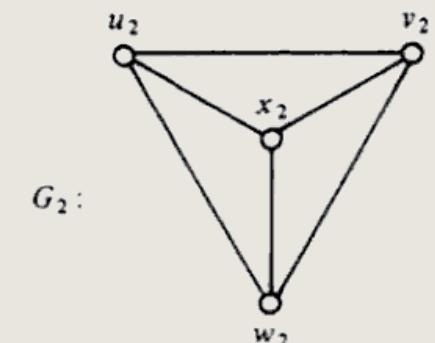
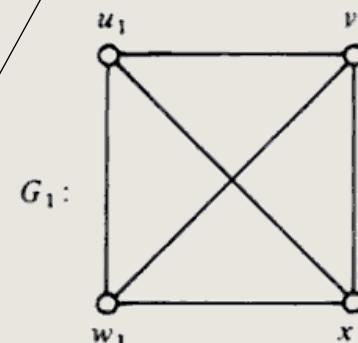
$G_2$ : nodos = { $u_2, v_2, w_2, x_2$ }

aristas = {  $\{u_2, v_2\}, \{u_2, w_2\}, \{u_2, x_2\}, \{v_2, x_2\}, \{v_2, w_2\}, \{x_2, w_2\}$  }

Problema computacionalmente sencillo (usualmente)



Two isomorphic graphs. Tomado de “Introductory Graph Theory”. G. Chartrand (1977). Dover Publications.



# PROBLEMA MÁS DÍFICIL: ENCONTRAR MOLÉCULAS CON GRUPOS SIMILARES

Foye's Principles of Medicinal Chemistry (2008).  
T Lemke, DA Williams. Wolters Kluwer

## Otro problema común

Identificar compuestos que comparten grupos químicos similares

**Farmacóforos** – grupos químicos responsables de actividad farmacológica

**Grupos reactivos** – carbonilos, aldehidos, cetonas,

## Aplicaciones

Agrupar compuestos químicos en familias

Desarrollo de nuevas Drogas

Inferencia

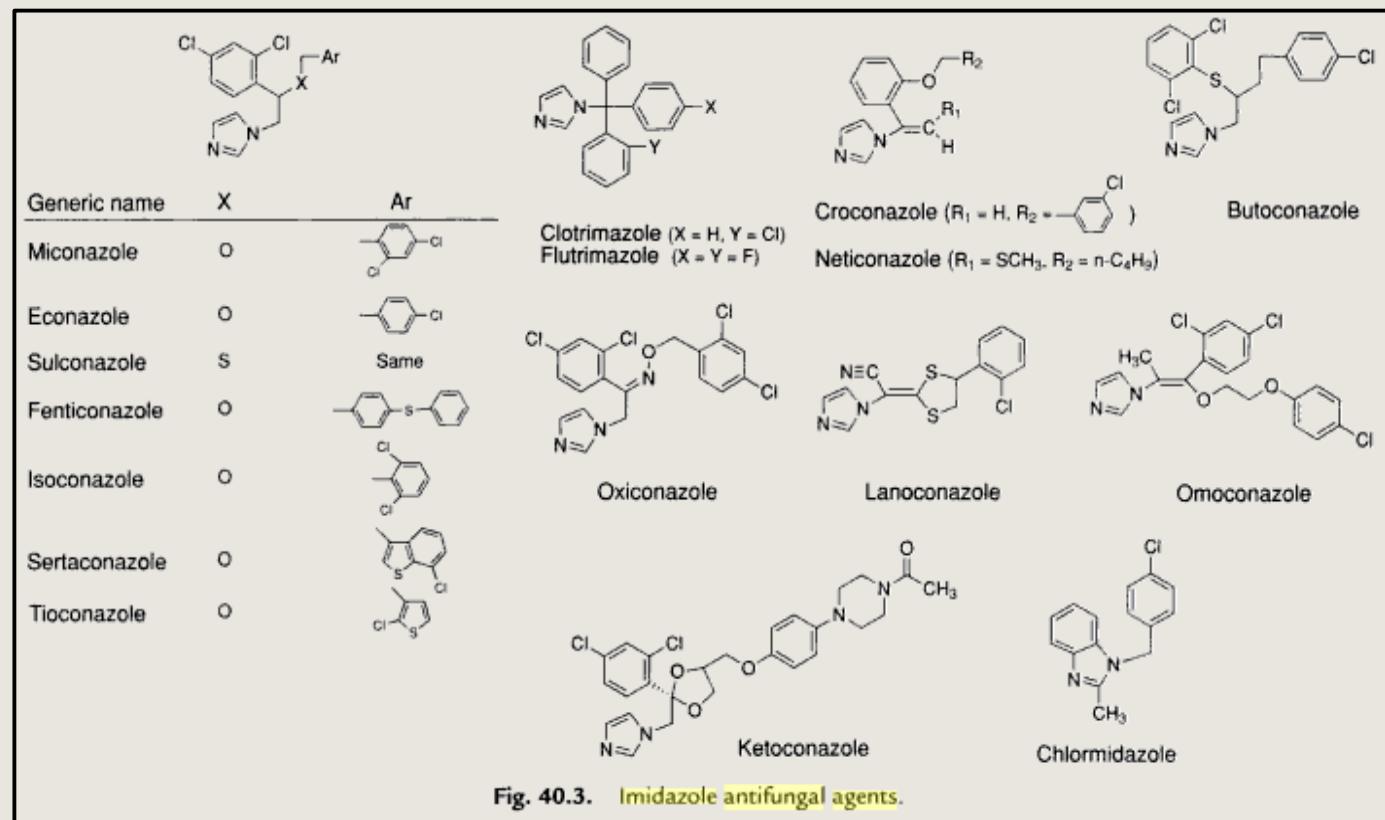


Fig. 40.3. Imidazole antifungal agents.

# PROBLEMA MÁS DÍFICIL: ENCONTRAR MOLÉCULAS CON GRUPOS SIMILARES

Computacionalmente: *subgraph isomorphism problem*

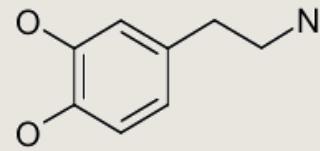
Encontrar un grafo determinado (fijo) dentro de otro grafo

Encontrar el **máximo subgrafo compartido** entre dos grafos

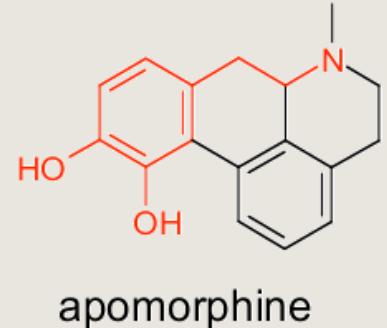
Es un problema computacionalmente difícil!

El tiempo se incrementa exponencialmente con el tamaño del problema (en este caso el número de nodos del grafo)

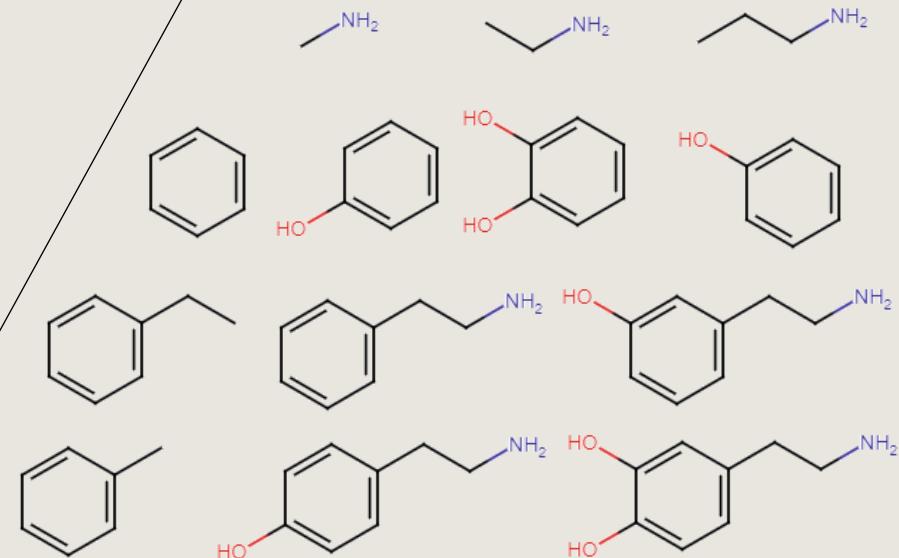
Query:



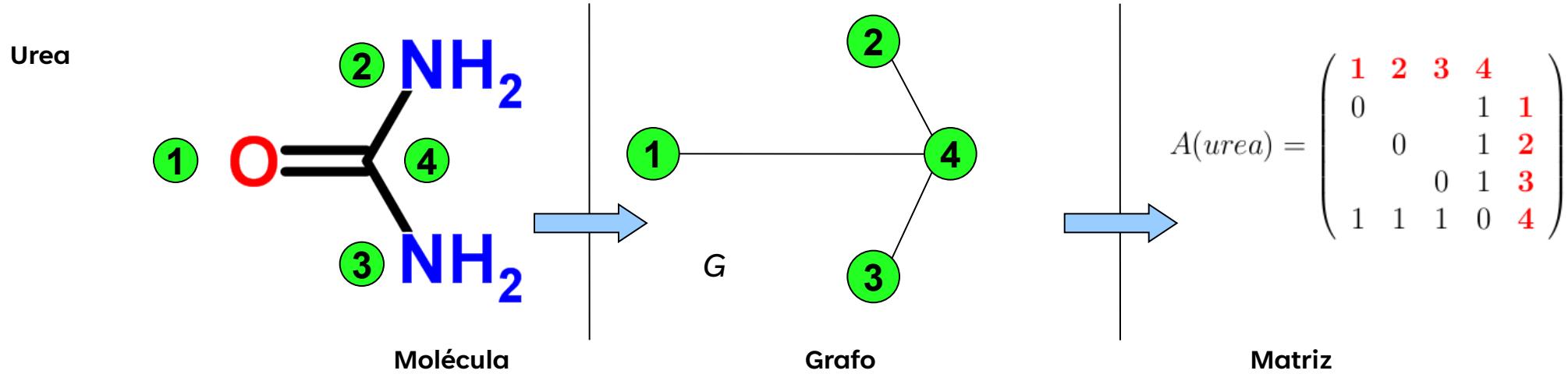
Hits:



**Subgrafos compartidos**



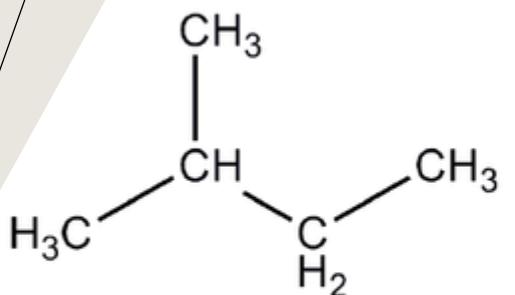
# BÚSQUEDA DE SUBESTRUCTURAS: MATRICES DE ADYACENCIA



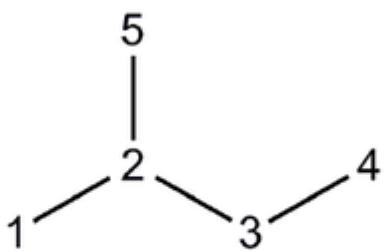
Dado un grafo, es posible construir una **matriz de adyacencia**

Es una aproximación (heurística) a la búsqueda de subestructuras: localizar coincidencias en una matriz de adyacencias

# ADJACENCY MATRICES



Molecule



Graph

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

adjacency matrix

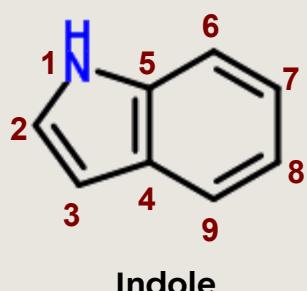
The chemical graph and adjacency matrix of the isopentane.

# BÚSQUEDA DE SUBESTRUCTURAS: MATRICES DE ADYACENCIA

**Indol:** compuesto heterocíclico aromático, precursor de muchas drogas

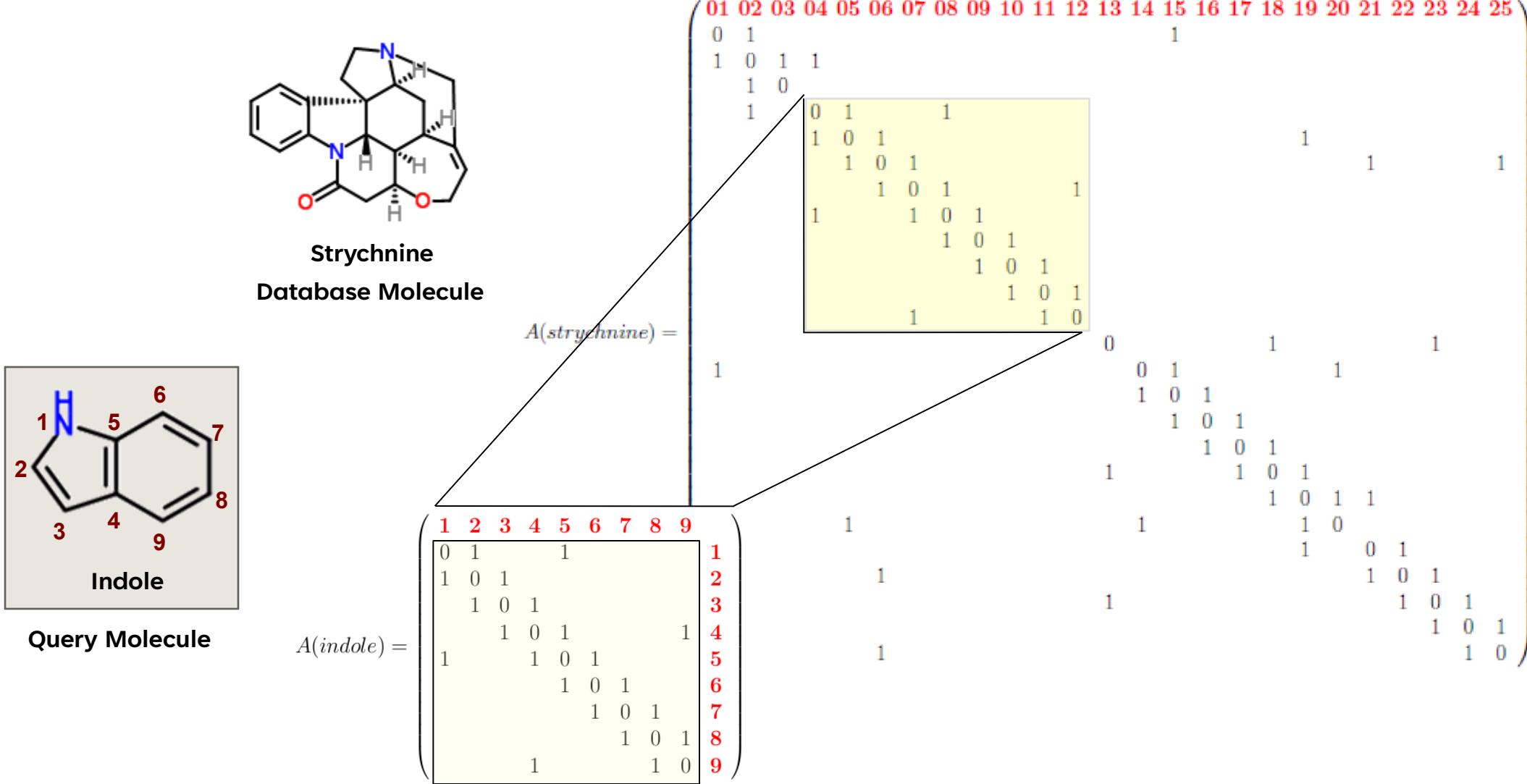
Búsqueda de compuestos que contengan el grupo **indol**

1. Calcular la matriz de adyacencia para la molécula 'query'
2. Calcular las matrices de adyacencia para todas las moléculas a testear (la base de datos)
3. Buscar coincidencias en las matrices de adyacencia



$$A(indole) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 0 & 1 & & & & 1 & & & 1 \\ 1 & 0 & 1 & & & & & & 2 \\ & 1 & 0 & 1 & & & & & 3 \\ & & 1 & 0 & 1 & & & & 4 \\ & & & 1 & 0 & 1 & & & 5 \\ & & & & 1 & 0 & 1 & & 6 \\ & & & & & 1 & 0 & 1 & 7 \\ & & & & & & 1 & 0 & 1 \\ & & & & & & & 1 & 0 \\ 1 & & & & & & & & 8 \\ & & & & & & & & 9 \end{pmatrix}$$

# BÚSQUEDA DE SUBESTRUCTURAS: MATRICES DE ADYACENCIA



# BÚSQUEDA DE SUBESTRUCTURAS: MATRICES DE ADYACENCIA

---

**Problema de esta estrategia (hasta acá):**

Puede dar falsos positivos

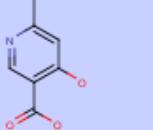
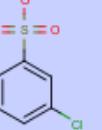
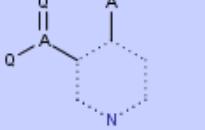
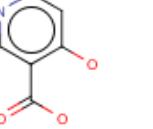
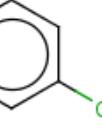
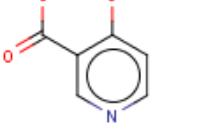
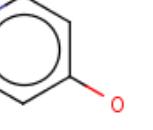
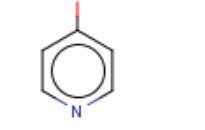
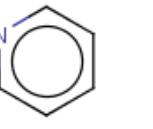
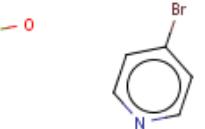
Grafos que tienen el mismo número de nodos, con la misma adyacencia, pero cuyos nodos están compuestos por distintos átomos (en el caso de moléculas)

**Possible solución:**

**Screening** – realizar la búsqueda sólo sobre un subconjunto de moléculas (grafos) compatibles

Ej: (query = indol) filtrar la base de datos: seleccionar solamente moléculas que tengan al menos 1 átomo de nitrógeno

# MAXIMUM COMMON SUBSTRUCTURE SEARCH

	Query 1	Query 2	Query 3
Target 1			
Target 2			
Target 3			
Target 4			

[https://docs.chemaxon.com/display/docs/jklustor\\_maximum-common-substructure-mcs-search.md](https://docs.chemaxon.com/display/docs/jklustor_maximum-common-substructure-mcs-search.md)

# INTERVALO

## 15 minutos

May your morning coffee  
give you the strength  
to make it to your  
mid-morning  
coffee.



someecards

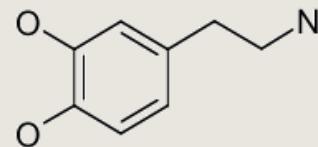
# BÚSQUEDA DE SUBESTRUCTURAS

Screenings

Simple:

- Usa la fórmula molecular (ej C<sub>8</sub>O<sub>2</sub>N)
  - La fórmula de todos los compuestos está almacenada en la base de datos
  - La fórmula de la molécula *query* se calcula al inicio de la búsqueda
  - Se descartan moléculas a las que les faltan átomos requeridos

Query:

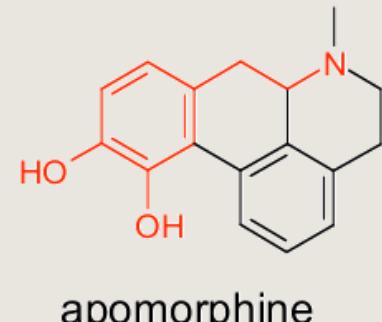


MF: C<sub>8</sub>O<sub>2</sub>N (H implícito)

Hits:



adrenaline



apomorphine

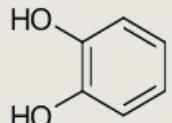


morphine

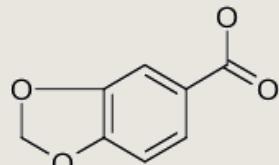
# BÚSQUEDA DE SUBESTRUCTURAS: FINGERPRINTS

**Fingerprint:** representación abstracta de características o propiedades de una molécula (features)

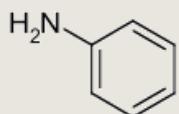
- Presencia/ausencia de cada elemento
- Configuraciones electrónicas inusuales (carbono sp<sub>3</sub>, nitrógeno unido con un triple enlace)
- Anillos y sistemas de anillos (naftaleno, piridina, cyclohexano)
- Grupos funcionales (alcoholes, aminas, carboxilos, etc.)
- Se suelen utilizar tanto para búsquedas de subestructuras como para detectar similitud



1	0	0	0	1	1	0
---	---	---	---	---	---	---



1	0	1	1	1	1	0
---	---	---	---	---	---	---

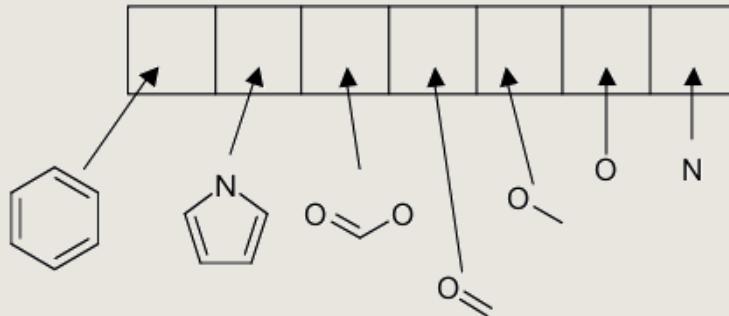


1	0	0	0	0	0	1
---	---	---	---	---	---	---

Query

✓ passes

✗ does not pass



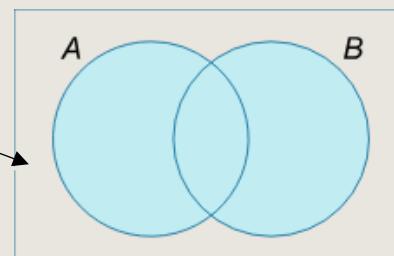
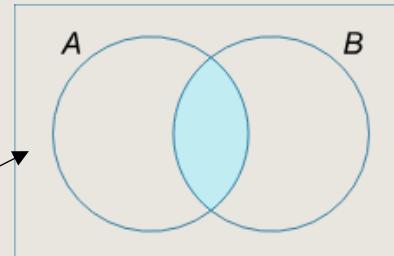
# BÚSQUEDA DE SUBESTRUCTURAS Y SIMILITUD: FINGERPRINTS

**Ventajas:** screening extremadamente rápido

Se evalúa equivalencia entre conjuntos de bits usando el operador AND binario

Se pueden calcular distancias de similitud a partir de los bits significativos

<b>X</b>	10001101
<b>Y</b>	01010111
-----	
<b>X AND Y</b>	00000101
-----	
<b>X OR Y</b>	11011111



# DISTANCE METRICS: SIMILARITY, DISIMILARITY

Cociente entre el tamaño de la intersección y el tamaño de la unión de los conjuntos de datos

**Jaccard index (J) = Jaccard similarity coefficient = Tanimoto Index = Tanimoto similarity coefficient**

(tambien llamado “Intersection Over Union”)

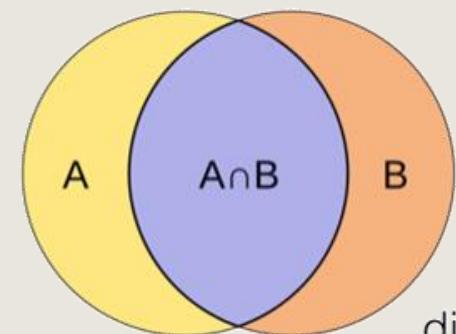
Compara similitudes entre conjuntos de datos finitos

**Jaccard distance ( $d_J$ )**

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

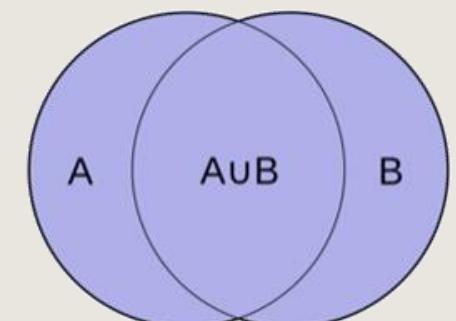
$$d_J(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

The intersect of A & B



division

The union of A & B



Safizadeh H, et al. Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical-Genetic Interactions. *J Chem Inf Model* **61**, 4156–4172 (2021).  
<https://doi.org/10.1021/acs.jcim.0c00993>.

Raymond, J.W., Willett, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J Comput Aided Mol Des* **16**, 59–71 (2002).  
<https://doi.org/10.1023/A:1016387816342>.

$\text{sqrt}$  = square root



X = number of bits **set** in **both fingerprints**

Y = number of bits **set** in the **first fingerprint**

Z = number of bits **set** in the **second fingerprint**

W = total number of bits in the bitstring

# DISTANCE METRICS

*There are many ways to measure distances...*

Name	Measurement	Range
Braun-Blanquet	$x / \max(x,y)$	0 – 1
Cosine	$x / \sqrt{yz}$	0 – 1
Dice	$2x / (y + z)$	0 – 1
Dot product	$x$	$0 - \infty$
Euclidean	$1 / 1 + \sqrt{y + z - 2x}$	0 – 1
Kulczunksi		0 – 1
McConaughey	$x(y + z) - yz / yz$	-1 – 1
Russel / Rao	$x / w$	0 – 1
Simpson	$x / \min(y,z)$	0 – 1
Sokal / Sneath	$x / (2y + 2z - 3x)$	0 – 1
Tanimoto / Jaccard	$x(y + z - x)$	0 – 1
Tullos	$xyz$	0 – 1
Tversky	$x / \alpha(y-x) + (1-\alpha)(z-x) + x$	0 – 1

# MOLECULAR FINGERPRINTS

*There are many ways to fingerprint molecules ...*

**Table 1. Molecular Fingerprints<sup>a</sup>**

ID	name	description	features	reference(s)
FP1	AP2D	topological atom pairs	1211	(44).
FP2	ASP	all-shortest paths	26,194	(45).
FP3	AT2D	topological atom triplets	56,963	(44).
FP4	DFS	all-paths (depth-first search)	48,448	(46).
FP5	ECFP	extended connectivity fingerprints	42,672	(47).
FP6	LSTAR	local path environments	85,232	(48).
FP7	MACCS	MDL public keys (166 keys)	155	(49).
FP8	PHAP2POINT2D	topological pharmacophore pairs	17	(50).
FP9	PHAP3POINT2D	topological pharmacophore triplets	302	(50).
FP10	RAD2D	topological molprint-like fingerprints	92,191	(48).
FP11	RDKit	topological daylight-like fingerprints	65,183	(43,51).

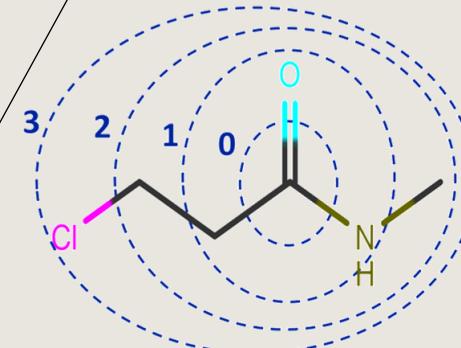
Safizadeh H, et al. Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical-Genetic Interactions. J Chem Inf Model **61**, 4156-4172 (2021).  
<https://doi.org/10.1021/acs.jcim.0c00993>.

# EXTENDED CONNECTIVITY FINGERPRINTS

## *Circular fingerprints*

Concepto similar al de “**extended connectivity**” de Morgan

1. Assign each atom with an identifier
2. Update each atom's identifiers based on its neighbors
3. Remove duplicates
4. Fold list of identifiers into a 2048-bit vector (a Morgan fingerprint)



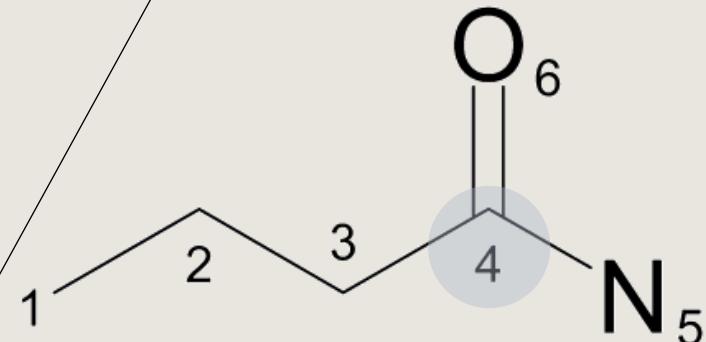
Extended Connectivity  
Circular Fingerprints  
**ECFP6 (radius = 3)**  
1024 or 2048 bits

# EXTENDED CONNECTIVITY FINGERPRINTS

## 1. Assign each atom with an identifier

We choose an atom in the molecule (e.g. #4) and take note of:

- number of nearest-neighbour non-hydrogen atoms: **3**
- number of bonds attached to the atom (not including bonds to hydrogens): **4**
- atomic number: **6**
- atomic mass: **12**
- number of hydrogens connected to the atom: **0**
- is the atom in a ring (1) or not (0)?: **0**
- **Resulting list of numbers is (3,4,6,12,0,0)**
- **Hash this list of numbers into an integer (identifier)**
  - In Python: `hash((3, 4, 6, 12, 0, 0, 0))` → -5700861834356229464



A beginner's guide for understanding Extended-Connectivity Fingerprints(ECFPs). Manish Kumar (2021).  
<https://chemicbook.com/2021/03/25/a-beginners-guide-for-understanding-extended-connectivity-fingerprints.html>

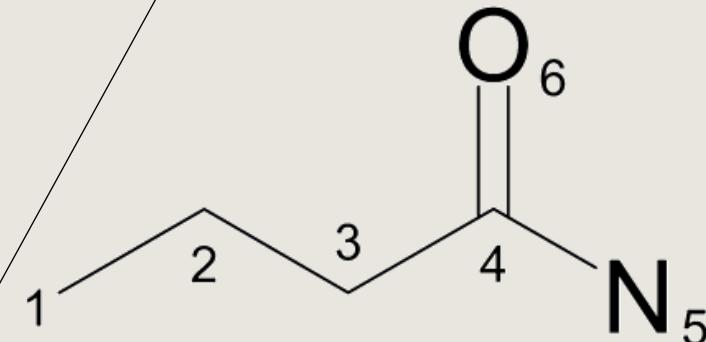
# EXTENDED CONNECTIVITY FINGERPRINTS

```
# identificadores para cada atomo
atomo1 = hash((1, 1, 6, 12, 0, 3, 0)) # -CH3
atomo2 = hash((2, 2, 6, 12, 0, 2, 0)) # -CH2
atomo3 = hash((2, 2, 6, 12, 0, 2, 0)) # -CH2
atomo4 = hash((3, 4, 6, 12, 0, 0, 0)) # -C
atomo5 = hash((1, 2, 7, 14, 0, 0, 0)) # -NH2
atomo6 = hash((1, 2, 8, 16, 0, 0, 0)) # =O
```

```
atomo 1 4940186308562569707
atomo 2 -7815985147897826576
atomo 3 -7815985147897826576
atomo 4 -5700861834356229464
atomo 5 -6296387744277800866
atomo 6 8618411755682373892
```



**List of  
features  
(6)**



<https://andrewbrookins.com/technology/pythons-default-hash-algorithm/>

A beginner's guide for understanding Extended-Connectivity Fingerprints(ECFPs). Manish Kumar (2021).  
<https://chemicbook.com/2021/03/25/a-beginners-guide-for-understanding-extended-connectivity-fingerprints.html>

# EXTENDED CONNECTIVITY FINGERPRINTS

Update each atom's identifiers based on its neighbors

Each atom collects its identifier and the identifiers of its immediately neighboring atoms, into an array (list)

And we hash this list again into a new identifier.

## Paso anterior

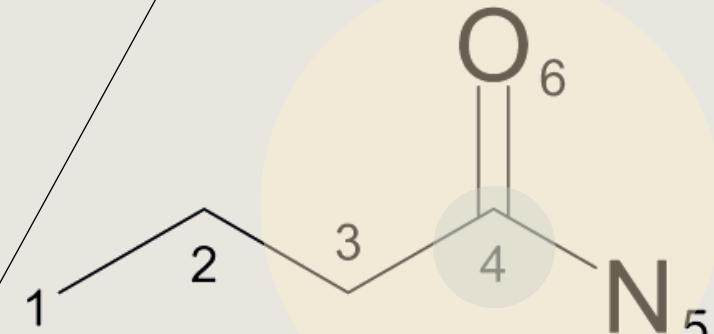
```
atomo 1 4940186308562569707  
atomo 2 -7815985147897826576  
atomo 3 -7815985147897826576  
atomo 4 -5700861834356229464  
atomo 5 -6296387744277800866  
atomo 6 8618411755682373892
```

```
atomo4_updated = hash((  
    1, -5700861834356229464,  
    1, -7815985147897826576,  
    1, -6296387744277800866,  
    2, 8618411755682373892  
))
```

-6784272694619664722

repetimos para los 6 átomos

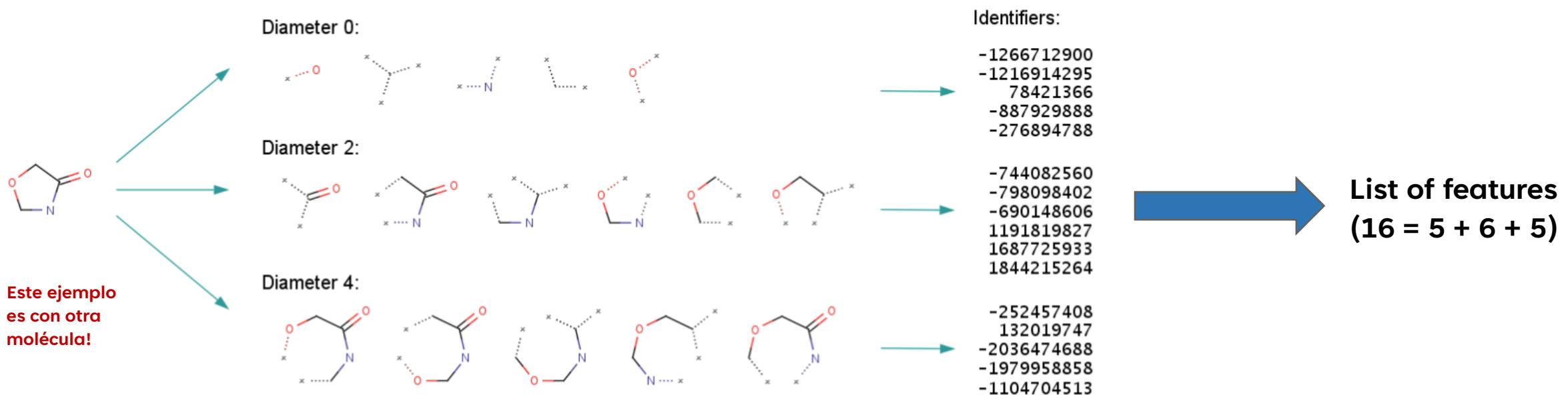
List of features (12)



A beginner's guide for understanding Extended-Connectivity Fingerprints(ECFPs). Manish Kumar (2021).  
<https://chemicbook.com/2021/03/25/a-beginners-guide-for-understanding-extended-connectivity-fingerprints.html>

# EXTENDED CONNECTIVITY FINGERPRINTS

- After that, several iterations are performed to combine the initial atom identifiers with identifiers of neighboring atoms **until a specified diameter is reached**. Each iteration captures larger and larger circular neighborhoods around each atom
- ECFP4 = Extended Circular Fingerprint with **diameter = 4 (radius = 2)**
- ECFP6 = Extended Circular Fingerprint with **diameter = 6 (radius = 3)**



# FINGERPRINTS: FOLDING AND BIT COLLISIONS

Para acomodar estos **features** en un fingerprint de 1024 bits

- Inicializar el fingerprint con **todos los bits en 0 (OFF)**
- Dividir cada identificador por 1024, y anotar el **resto de la división**
  - En Python: operador módulo (%)
- **Ese es el número de bit → que se pone en 1 (ON)**

**Resto**

$$\begin{array}{r} 24 \underline{\quad} 11 \\ 2 \quad \quad \quad 2 \end{array}$$

**Ejemplos:**

$$132019747 \% 1024 = 547$$

$$1687725933 \% 1024 = 877$$

$$-798098402 \% 1024 = \textcolor{red}{30}$$

**Folding**

**Fixed-length binary representation**

000100000001000001000001100 $\textcolor{red}{1}$ 000010001000000000000000100000[...]0000 $\textcolor{red}{1}$ 00000000000010

**Bit Collision:**

$$-14439656419269748 \% 1024 = \textcolor{red}{908}$$

$$-4080868480043360372 \% 1024 = \textcolor{red}{908}$$

**Solution: increase fingerprint size**

$$-14439656419269748 \% 2048 = \textcolor{green}{908}$$

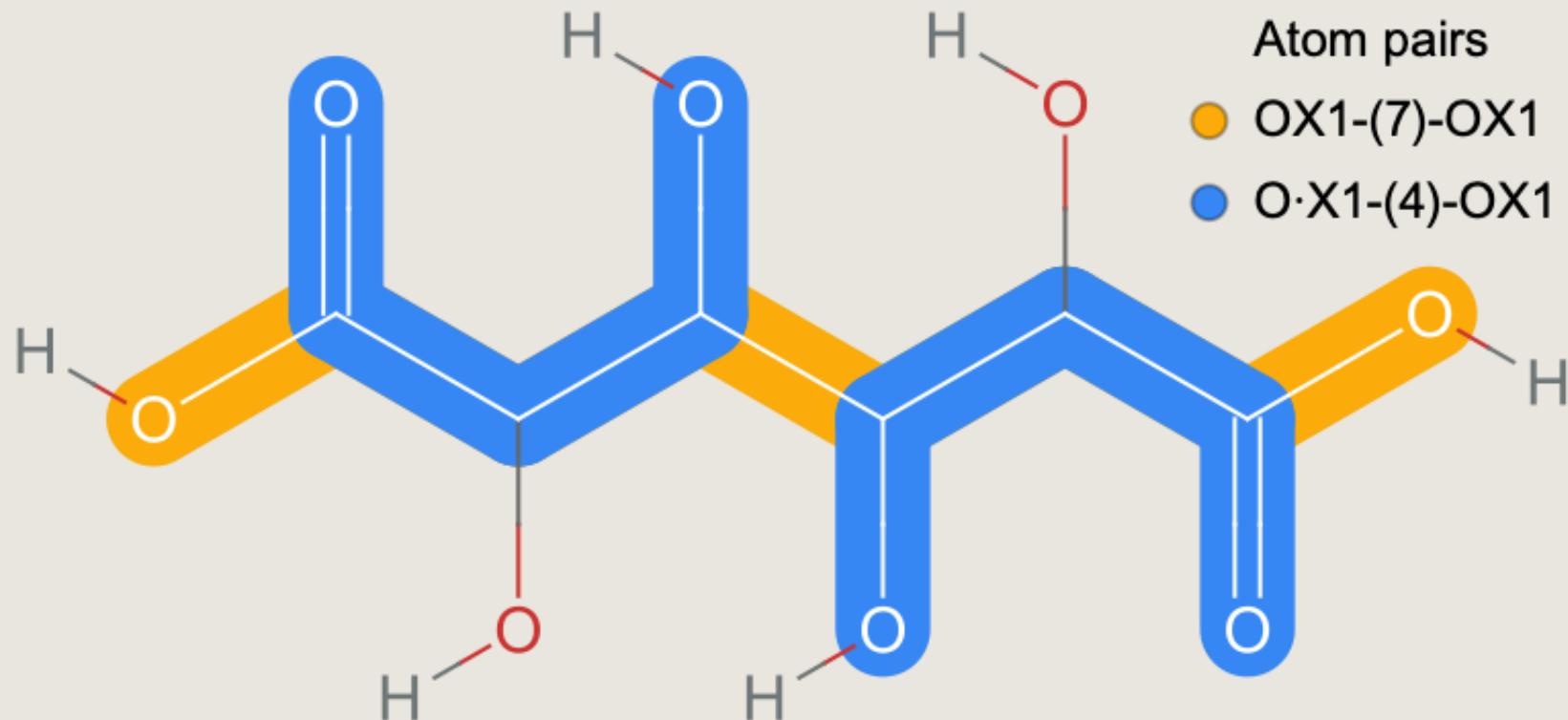
$$-4080868480043360372 \% 2048 = \textcolor{green}{1932}$$

# TYPES OF FINGERPRINTS

**Table 1. Molecular Fingerprints<sup>a</sup>**

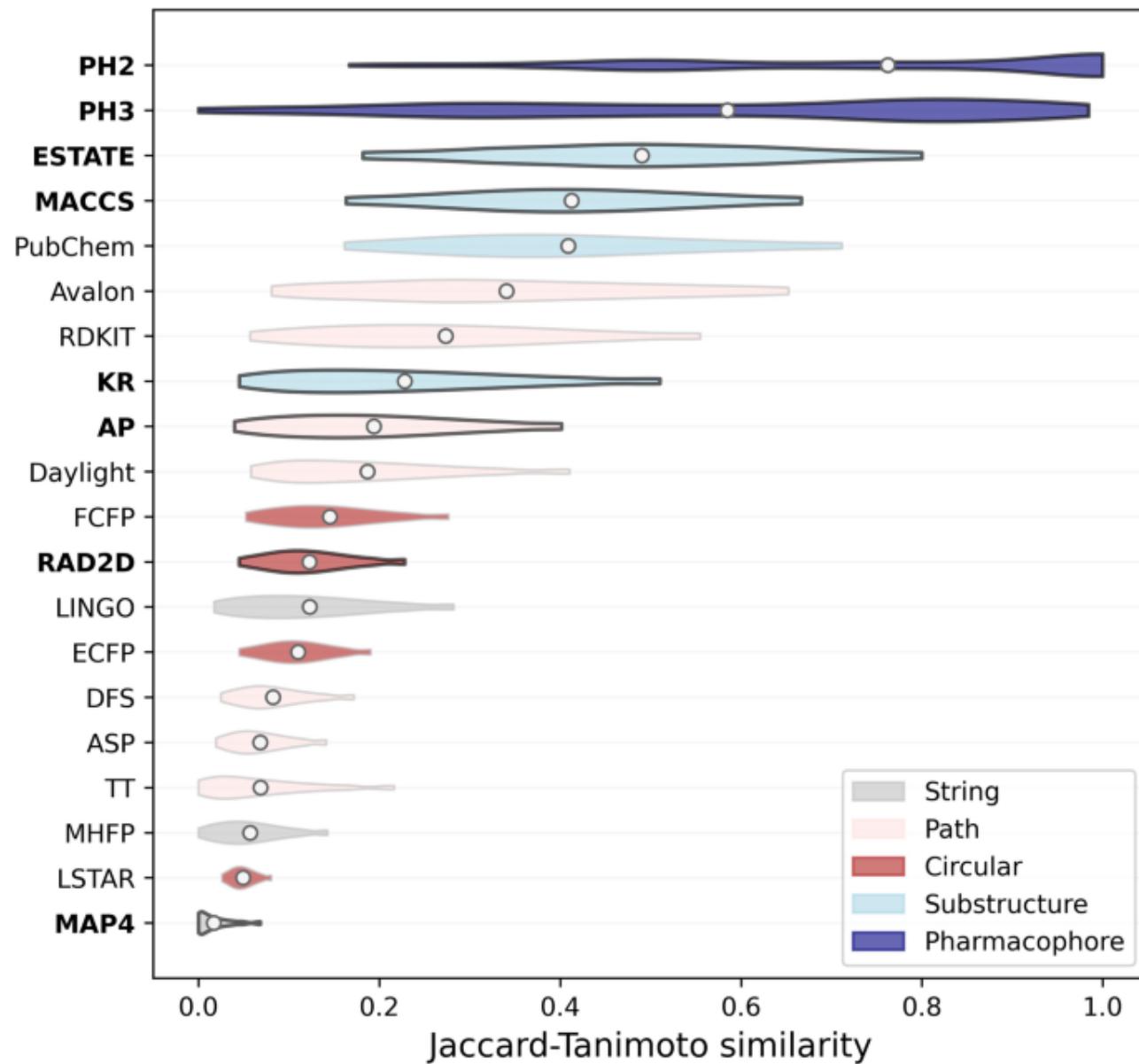
ID	name	description	features	reference(s)
FP1	AP2D	topological atom pairs	1211	(44).
FP2	ASP	all-shortest paths	26,194	(45).
FP3	AT2D	topological atom triplets	56,963	(44).
FP4	DFS	all-paths (depth-first search)	48,448	(46).
FP5	ECPFP	extended connectivity fingerprints	42,672	(47).
FP6	LSTAR	local path environments	85,232	(48).
FP7	MACCS	MDL public keys (166 keys)	155	(49).
FP8	PHAP2POINT2D	topological pharmacophore pairs	17	(50).
FP9	PHAP3POINT2D	topological pharmacophore triplets	302	(50).
FP10	RAD2D	topological molprint-like fingerprints	92,191	(48).
FP11	RDKit	topological daylight-like fingerprints	65,183	(43,51).

# TOPOLOGICAL ATOM PAIRS



<https://www.wolfram.com/language/12/molecular-structure-and-computation/topological-similarity-searching.html.en>

# ARE ALL FINGERPRINTS EQUAL?



Special case: *Natural compounds*

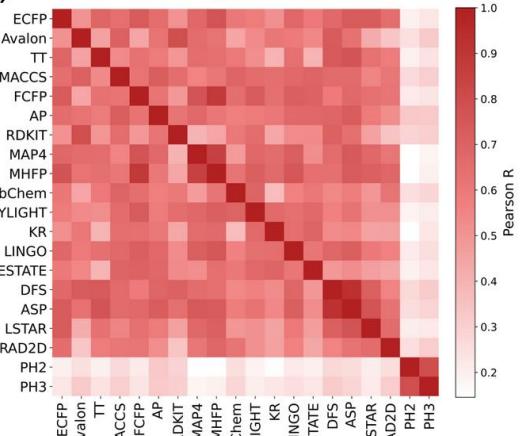
Different structural motifs in comparison with typical drug-like compounds, e.g.

- a wider range of molecular weight,
- multiple stereocenters
- higher fraction of  $sp^3$ -hybridized carbons

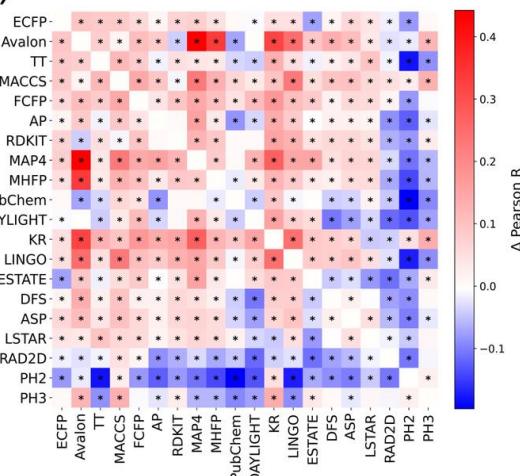
Boldini D, Ballabio D, Consonni V, Todeschini R, Grisoni F, Sieber SA. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *J Cheminform*. 2024 Mar 25;16(1):35. doi: 10.1186/s13321-024-00830-3. PMID: 38528548; PMCID: PMC10964529.

# CORRELATION SIMILARITY ACROSS FINGERPRINT TYPES

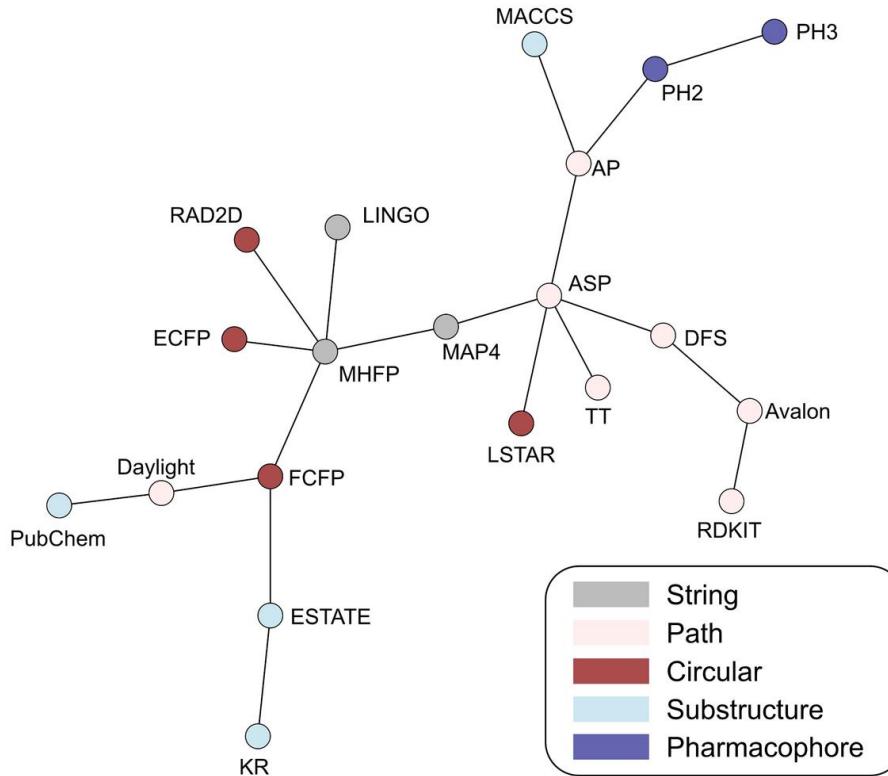
a)



b)



c)

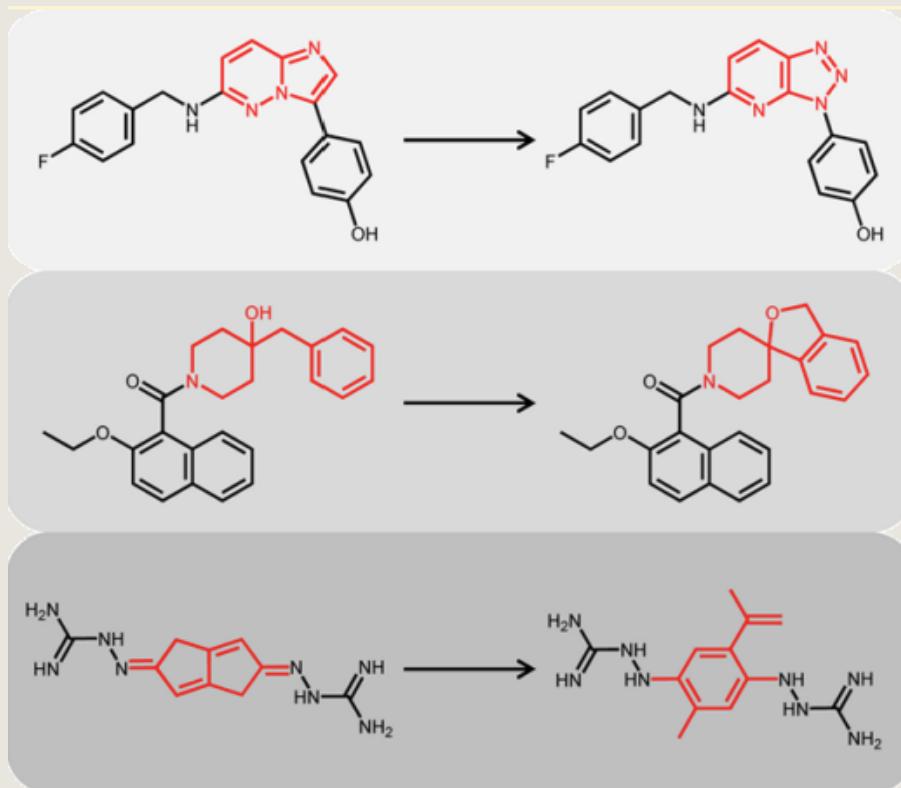


Jaccard-Tanimoto similarity correlation analysis for all fingerprints. **a** Correlation matrix for all fingerprints evaluated in this study on the COCONUT dataset. **b** Difference between the correlation matrix obtained for the COCONUT dataset and for the Drug Repurposing Hub. Positive values indicate higher fingerprint correlation in the NP space, while negative values denote higher correlation in the drug-like space. Asterisks denote statistical significance according to one-sample Mann Whitney tests with Benjamini–Hochberg correction ( $\alpha = 0.05$ ). **c** MST constructed from the fingerprint correlation matrix obtained for the NP chemical space. Each encoding is colored on the basis of its category

Boldini D, Ballabio D, Consonni V, Todeschini R, Grisoni F, Sieber SA. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *J Cheminform*. 2024 16(1):35. doi: 10.1186/s13321-024-00830-3. PMID: 38528548; PMCID: PMC10964529.

# SCAFFOLD HOPING

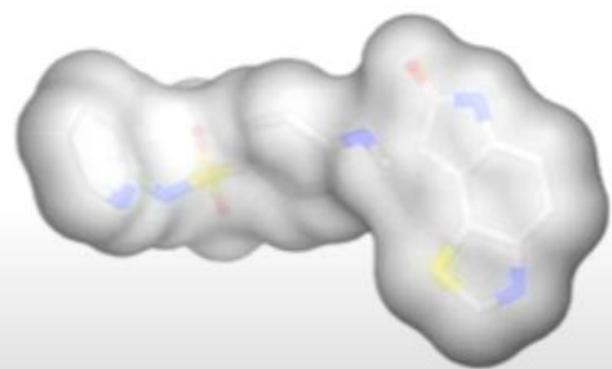
*“the identification of **isofunctional** molecular structures  
with chemically completely different core structures”*



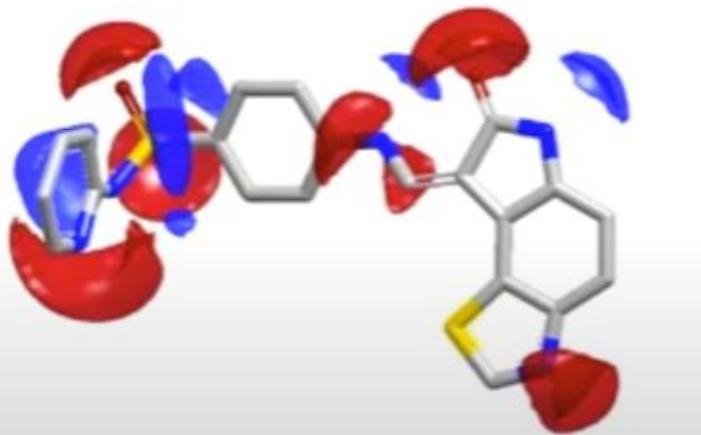
*computer-aided search for  
active compounds containing  
different core structures*

Hu Y, Stumpfe D, Bajorath J. Recent Advances in Scaffold Hopping. *J Med Chem.* 2017 Feb; 50(4):1238-1246. doi: 10.1021/acs.jmedchem.6b01437. Epub 2016 Dec 21. PMID: 28001064.

# OTRAS REPRESENTACIONES DE MOLÉCULAS



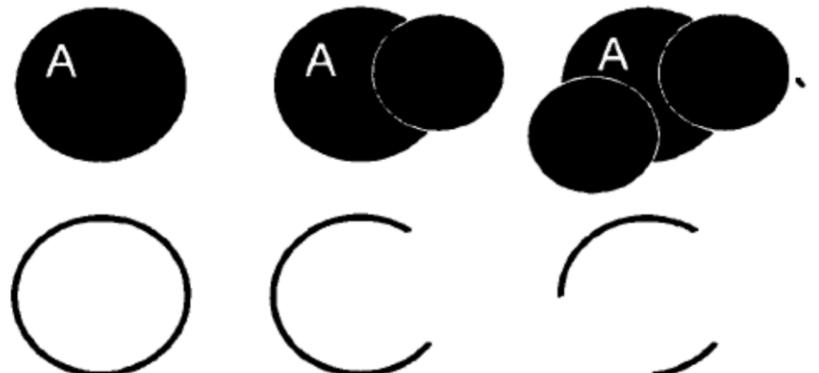
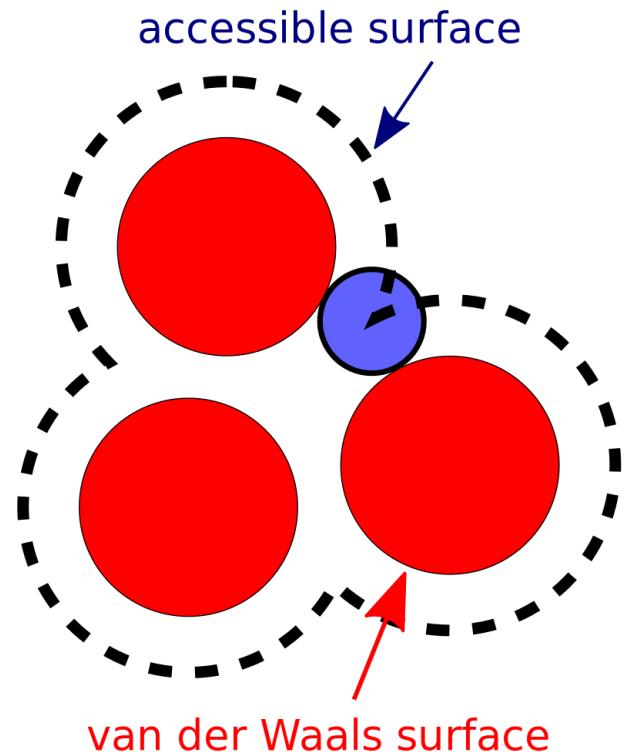
Shape



Electrostatics

# SOLVENT ACCESSIBLE SURFACE AREA CALCULATION

- VSA = van der Waals Surface Area
- AS = Accessible Surface Area

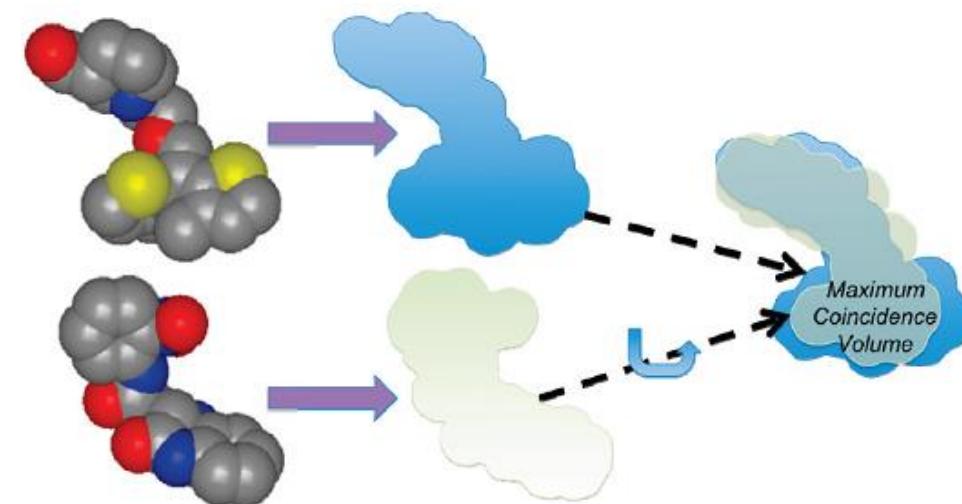
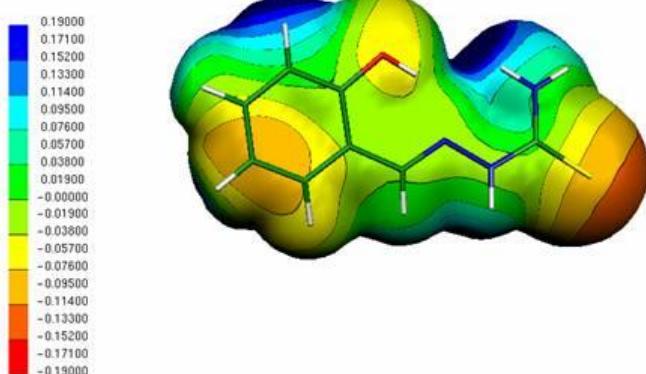


*Figure 1. Assuming spherical atoms, the surface area of atom A is the amount of surface area not contained in other atoms.*

Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Res. 2016 Feb 18;5:189. doi: 10.12688/f1000research.7931.1. PMID: 26973785; PMCID: PMC4776673.

# REPRESENTACIÓN DE MOLECULAS: 3D

- Una representación tridimensional de la molécula requiere no sólo especificar coordenadas espaciales de átomos
  - También hay que especificar
    - **Volumen**
      - Fused spheres
      - Atom-centered Gaussians
    - **Superficie**
    - **Forma**
      - Coincidencia de volúmenes

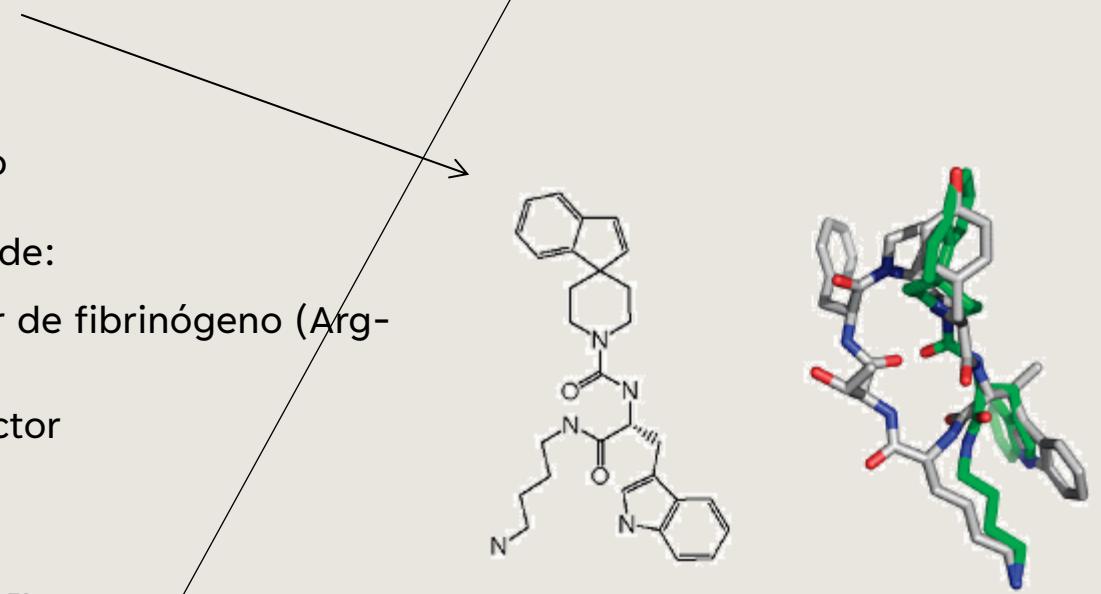


Molecular shape and medicinal chemistry: a perspective.  
2010. A Nicholls et al. J Med Chem 53: 3862

# REPRESENTACIÓN DE FORMA (SHAPE)

Varias aplicaciones posibles:

- Búsqueda de moléculas similares
  - En este caso la similitud es a nivel de forma
  - Se pueden agregar adicionalmente limitaciones
- Varias implementaciones en la industria farmacéutica
  - Virtual screening
    - Varios casos de éxito conocidos
    - Merck, primer aplicación publicada del método
      - Identificación de análogos no-peptídicos de:
        - antagonista endógeno del receptor de fibrinógeno (Arg-Gly-Pro)
        - Somatotrophin release inhibitor factor



# REPRESENTACIÓN DE FORMA (SHAPE)

Varias aplicaciones posibles:

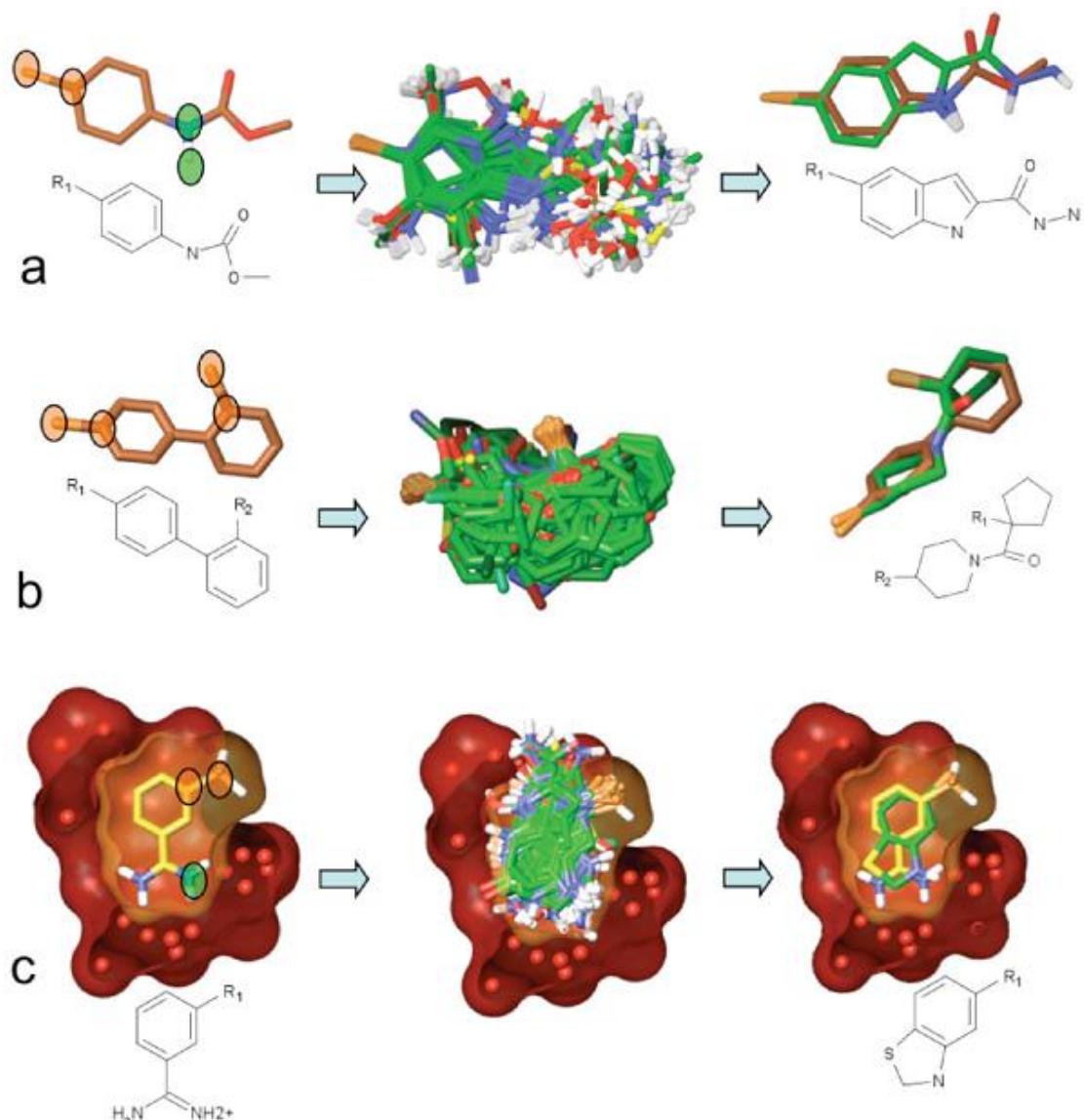
## Lead optimization

Uno cuenta con una molécula activa que quiere optimizar

## Scaffold Hoping

Facilmente explorable  
utilizando métodos  
computacionales

KIN: Bristol-Myers Squibb



Molecular shape and medicinal chemistry: a perspective. 2010. A Nicholls et al. J Med Chem 53: 3862

# CALCULO DE PROPIEDADES

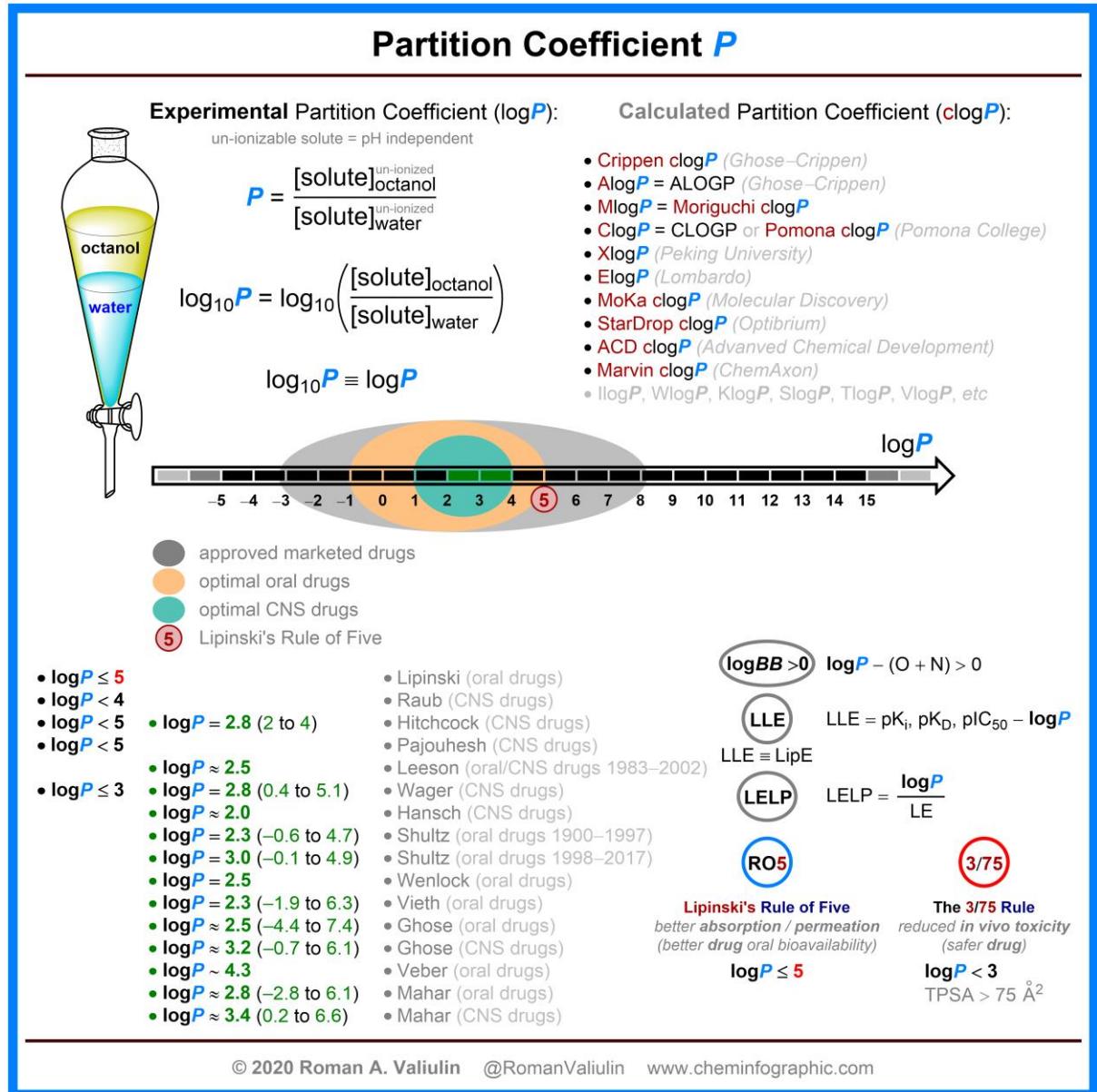
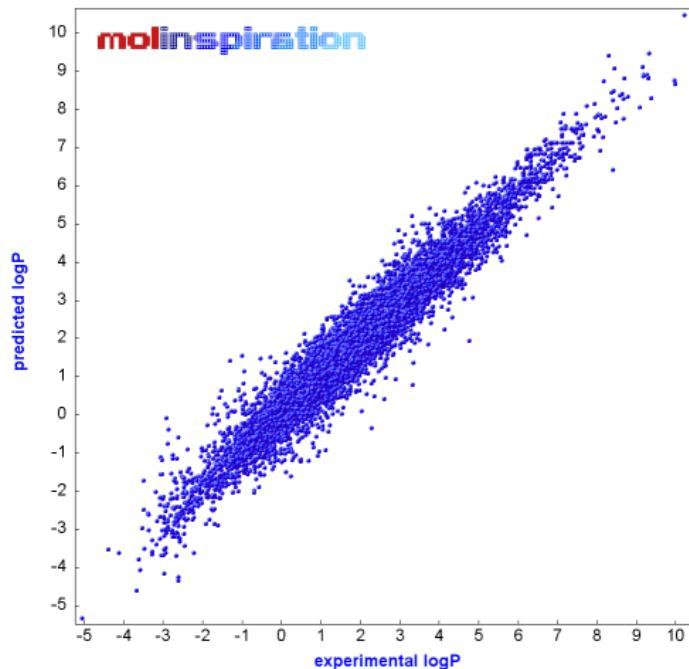
Enlaces rotables

Dadores / Aceptores de puentes de hidrógeno

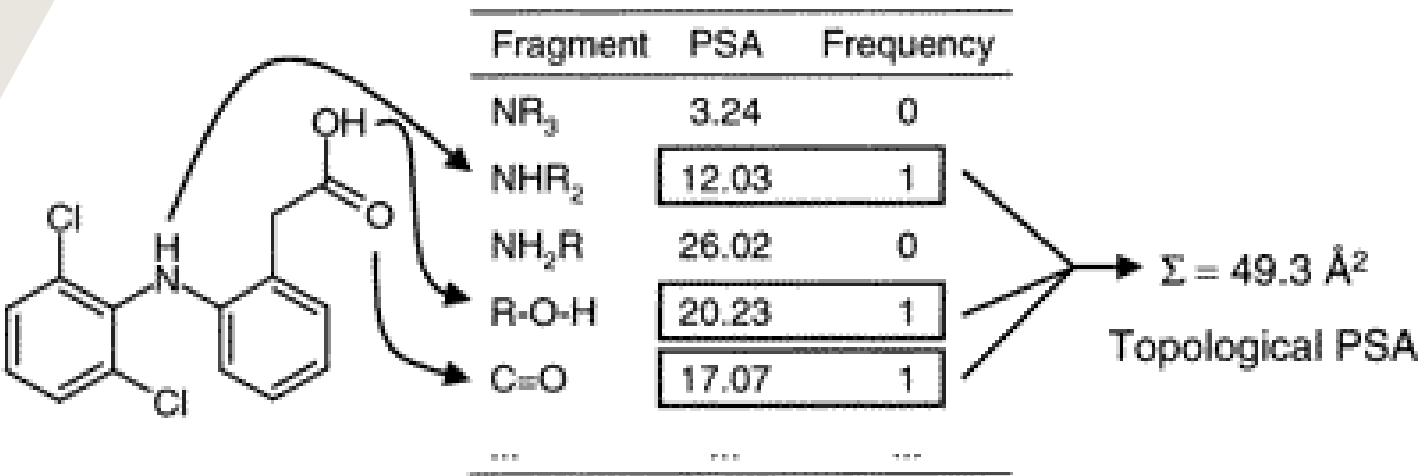
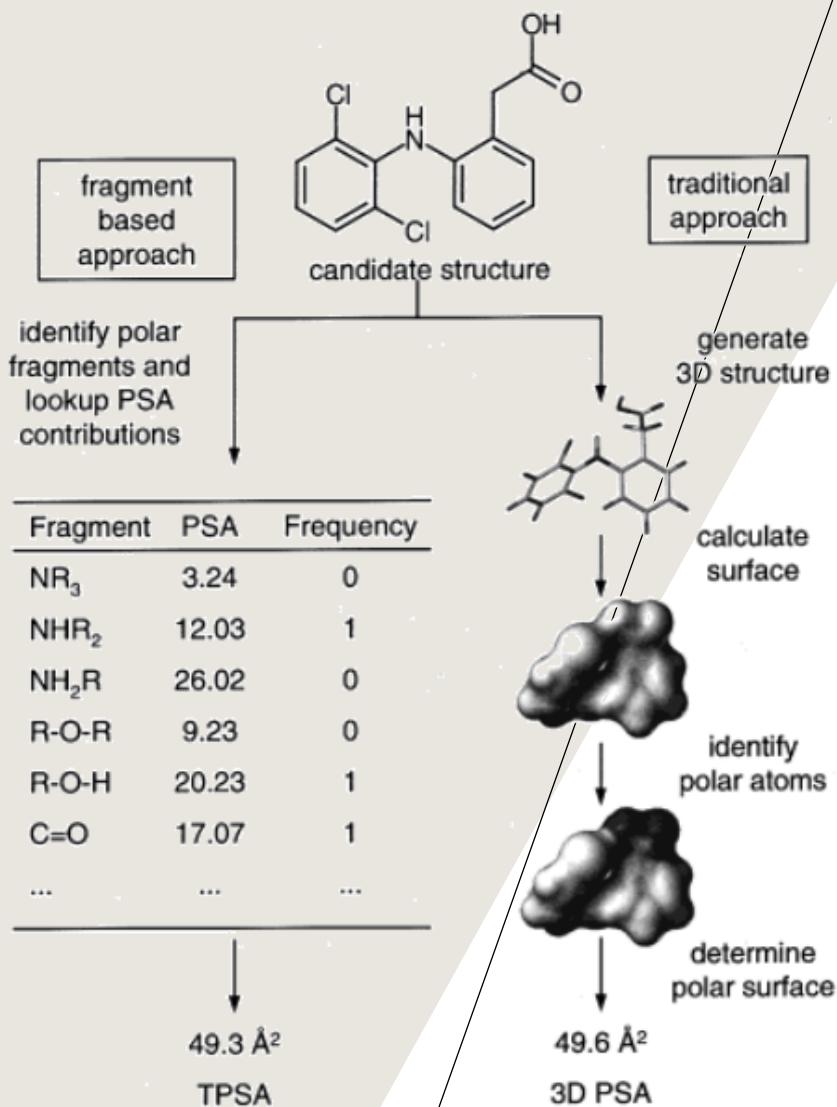
cLogP (coeficiente de partición octanol / agua)

PSA (polar surface area) / TPSA (topological surface area)

# LOGP PARTITION COEFFICIENT



# PSA / TPSA



## Polar Surface Area (PSA, costoso)

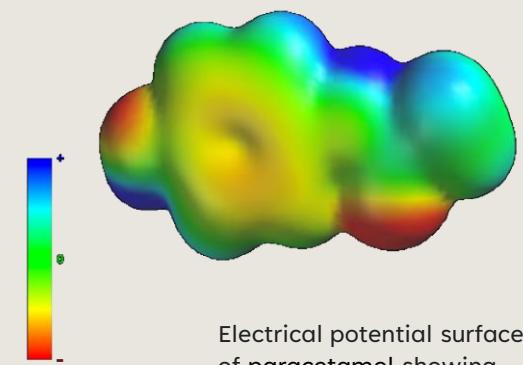
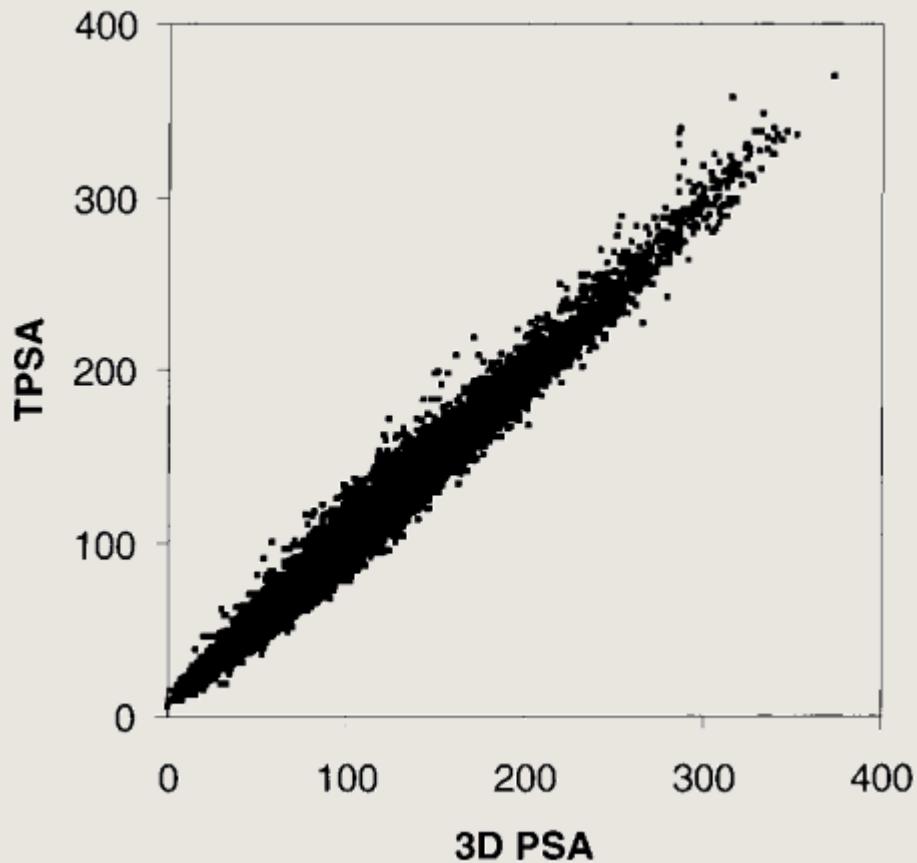
- Requiere generar conformeros 3D para calcular SA (Surface Area)

## Topological Polar Surface Area (TPSA)

- Sumatoria de contribuciones tabuladas de fragmentos polares

Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem. 2000 Oct 5;43(20):3714-7. doi: 10.1021/jm000942e. PMID: 11020286.

## TPSA VS PSA (3D)

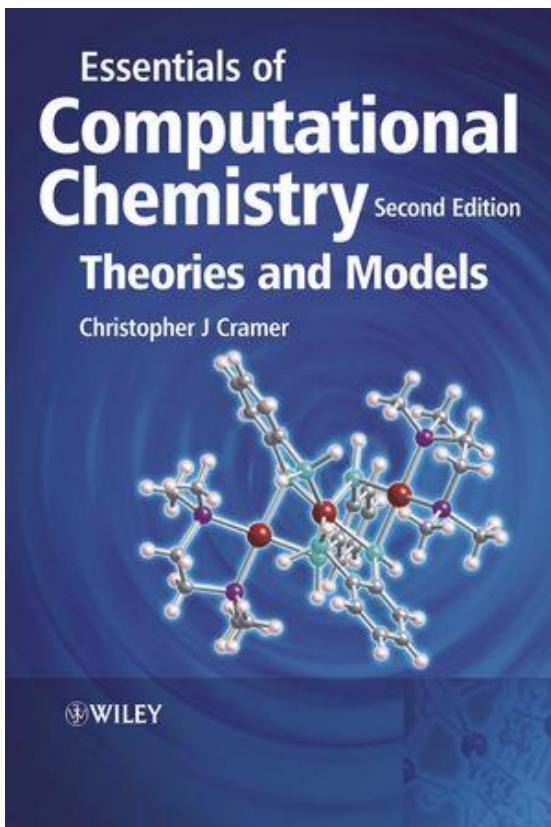


Electrical potential surface  
of paracetamol showing  
polar areas in red and blue

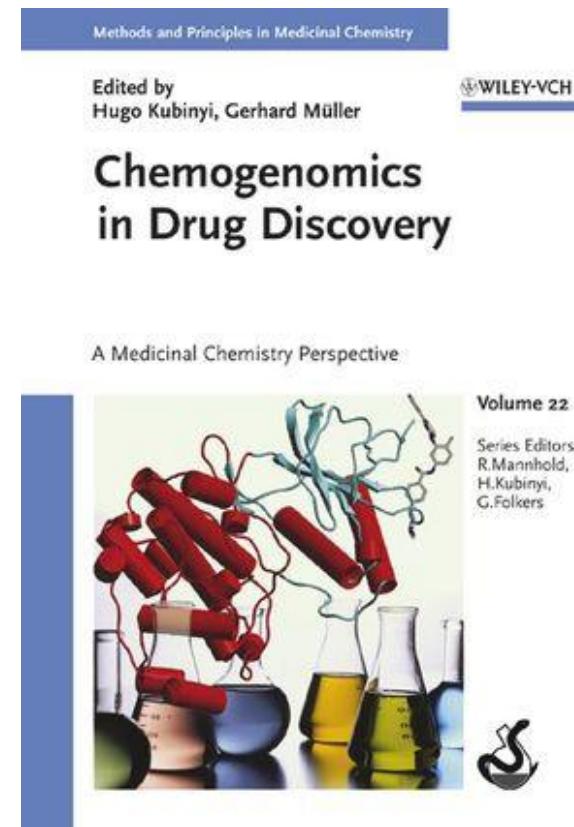
Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J Med Chem.* 2000 Oct 5;43(20):3714-7. doi: 10.1021/jm000942e. PMID: 11020286.



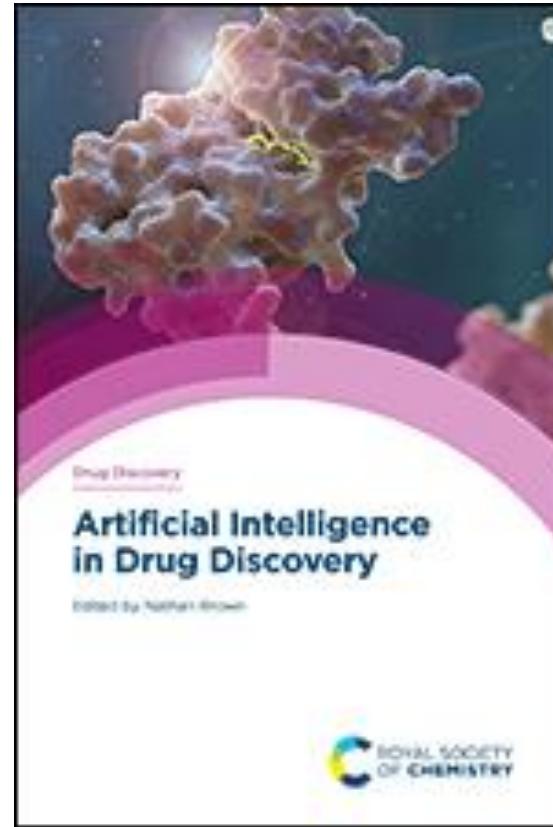
# BIBLIOGRAFÍA | MATERIAL DE LECTURA



Essentials of Computational Chemistry (2004), 2<sup>nd</sup> Ed, CJ Cramer. Wiley.  
<https://www.wiley.com/en-sg/Essentials+of+Computational+Chemistry:+Theories+and+Models,+2nd+Edition-p-9780470091821>



Chemogenomics in Drug Discovery: A Medicinal Chemistry Perspective (2006). Edited by Hugo Kubinyi & Gerhard Müller, Wiley-VCH.  
<https://www.wiley.com/en-us/Chemogenomics+in+Drug+Discovery%3A+Medicinal+Chemistry+Perspective-p-9783527604029>



Artificial Intelligence in Drug Discovery (2020). Edited by Nathan Brown. Royal Society of Chemistry.  
<https://doi.org/10.1039/9781788016841>



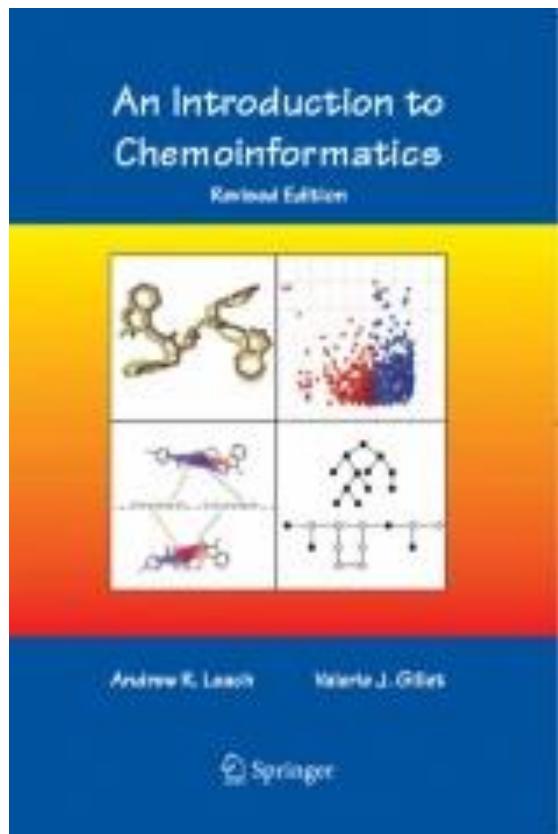
Open-Source Cheminformatics and Machine Learning

The RDKit Book (2023).  
[https://www.rdkit.org/docs/RDKit\\_Book.html](https://www.rdkit.org/docs/RDKit_Book.html)

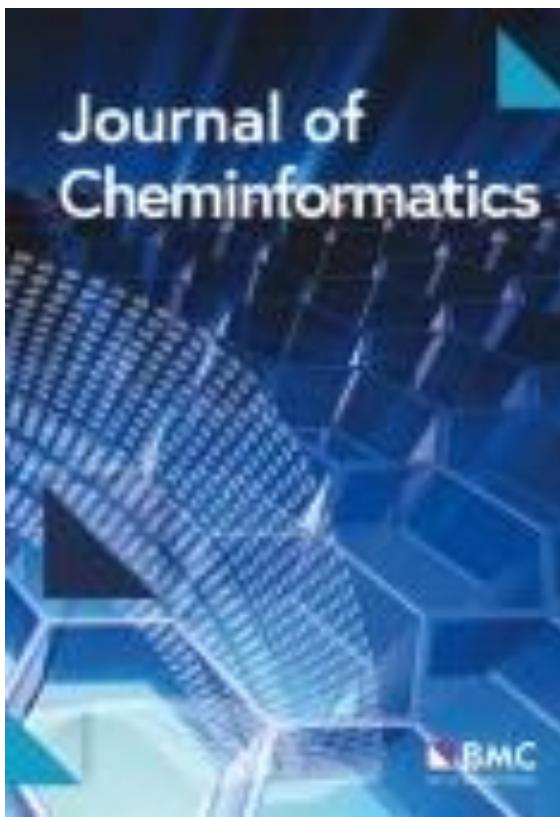
# BIBLIOGRAFÍA | MATERIAL DE LECTURA



<https://www.ebi.ac.uk/chebi/aboutChebiForward.do>



An Introduction to Chemoinformatics (2007). Andrew Leach and Valerie Gillet. Springer.  
<https://link.springer.com/book/10.1007/978-1-4020-6291-9>



Químicoinformática UNSAM

Dalke J Cheminform (2019) 11:76  
<https://doi.org/10.1186/s13321-019-0398-8>

Journal of Cheminformatics

**METHODOLOGY**  
The chemfp project  
Andrew Dalke\*

**Abstract**  
The chemfp project has had four main goals: (1) promote the FPS format as a text-based exchange format for dense binary chemoinformatics fingerprints, (2) develop a high-performance implementation of the BitBound algorithm that could be used as an effective baseline to benchmark new similarity search implementations, (3) experiment with funding a pure open source software project through commercial sales, and (4) publish the results and lessons learned as a guide for future implementors. The FPS format has had only minor success, though it did influence development of the FFP binary format, which is faster to load but more complex. Both are summarized. The chemfp benchmark and the no-cost/open source version of chemfp was proposed as a reference baseline to evaluate the effectiveness of other similarity search tools. They are used to evaluate the faster commercial version of chemfp, which can test 130 million 1024-bit fingerprint Tanimoto per second on a single core of a standard x86-64 server machine. When combined with a GPU, chemfp achieves a 100x speedup. The GPU version can search 1.8 billion 1024-bit fingerprints of CHEMBL 24 averages 27 ms/q. The same search of 970 million PubChem fingerprints averages 220 ms/q, making chemfp one of the fastest CPU-based similarity search implementations. Modern CPUs are fast enough that memory bandwidth and latency are now important factors. Single-threaded search uses most of the available memory bandwidth. Sorting the fingerprints by popcount improves memory coherency, which when combined with 4 OpenMP threads makes it possible to construct an N x N similarity matrix for 1 million fingerprints in about 30 min. These observations may affect the interpretation of previous publications which assumed that search was strongly CPU bound. The chemfp project funding came from selling a purely open-source software product. Several product business models were tried, but none proved sustainable. Some of the experiences are discussed. In order to contribute to the ongoing conversation on the role of open source software in cheminformatics.

**Keywords:** Molecular fingerprints, Similarity searching, Tanimoto, High-performance, Format, Open source, FOSS, Performance benchmark

**Introduction**  
Molecular similarity search is a fundamental concept in cheminformatics. The most common form is almost certainly a Tanimoto similarity search of bitstring fingerprints. Complete search systems are available from many vendors, or a good programmer can implement a basic system in a few hours. High-performance search systems, which combine fast popcount evaluation and pruning algorithms, require significantly more development effort. This paper starts with a review of those approaches, many of which are either described in the cheminformatics literature in an incremental fashion which make them difficult to discover, or only published in the specialist literature of other fields.

The chemfp project started in order to develop a de

facto file format for chemical fingerprints. This requires some consideration of why such a format did not already exist, in order to understand which factors to focus on. The chemfp project was funded by selling a software package, the text-based FPS exchange format, which is simple to read and write, easily compressed, and appropriate for streaming workflows, and the binary FFP application format which is more complex and requires random-access reads, but has significantly shorter load times.

The chemfp package for Python includes optimized threshold and k-nearest implementations FPS file scan search implementations, highly-optimized implementations of the BitBound pruning method to search data

\*Correspondence: [dalke@dalkescientific.com](mailto:dalke@dalkescientific.com)  
Andrew Dalke Scientific AB, Trollhättan, Sweden

**BMC**

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

The ChemFP Project (2019).  
<https://link.springer.com/article/10.1186/S13321-019-0398-8>

# PREGUNTAS?

Fernán Agüero

[fernan@iib.unsam.edu.ar](mailto:fernan@iib.unsam.edu.ar)

