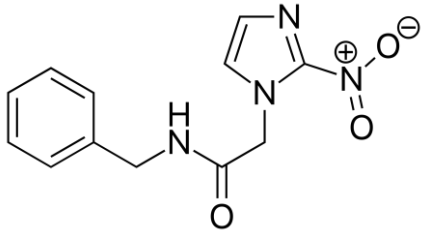
python



# INTRODUCCIÓN A LA QUIMIOINFORMÁTICA

Fernán Agüero
Instituto de Investigaciones Biotecnológicas, UNSAM

# BÚSQUEDA DE SUBESTRUCTURAS/SIMILITUD: FINGERPRINTS

**Cuestiones a tener en cuenta**

El fingerprint debe ser definido de antemano: bits, folding, count vectors...

Distintas aplicaciones pueden generar distintos tipos de fingerprints

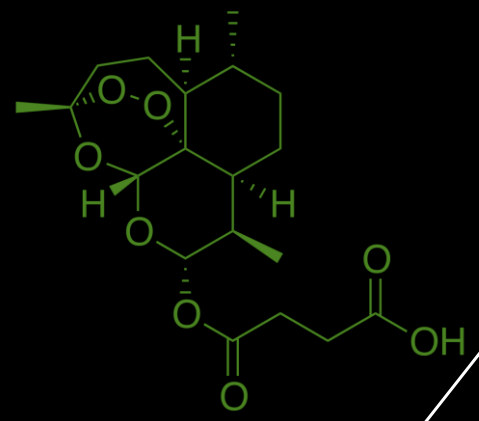
OpenBabel: https://openbabel.org/docs/dev/Features/Fingerprints.html

```
$ babel -L fingerprints
FP2    Indexes linear fragments up to 7 atoms.
FP3    SMARTS patterns specified in the file patterns.txt
FP4    SMARTS patterns specified in the file SMARTS_InteLigand.txt
MACCS    SMARTS patterns specified in the file MACCS.txt
```

Daylight: https://www.daylight.com/dayhtml/doc/theory/theory.finger.html

RDKit: https://www.rdkit.org/docs/GettingStartedInPython.html#fingerprinting-and-molecular-similarity

O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminform. 2011 Oct 7;3:33. doi: 10.1186/1758-2946-3-33. PMID: 21982300; PMCID: PMC3198950.



# DETOUR

Mirar patrones SMARTS en Github

**FP3**
https://github.com/openbabel/openbabel/blob/master/data/patterns.txt
FP4
https://github.com/openbabel/openbabel/blob/master/data/SMARTS_InteLigand.txt
MACCS
https://github.com/openbabel/openbabel/blob/master/data/MACCS.txt

# BÚSQUEDA DE SUBESTRUCTURAS

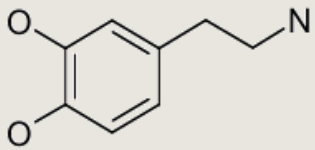
Screenings

Simple:

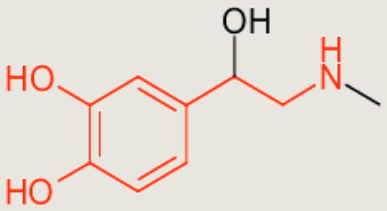
- Usa la fórmula molecular
    - La fórmula de todos los compuestos está almacenada en la base de datos
    - La fórmula de la molécula *query* se calcula al inicio de la búsqueda
    - Se descartan moléculas a las que les faltan átomos requeridos

**MF: C8 O2 N (H implícito)**

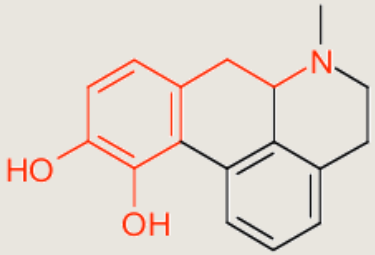
Query:



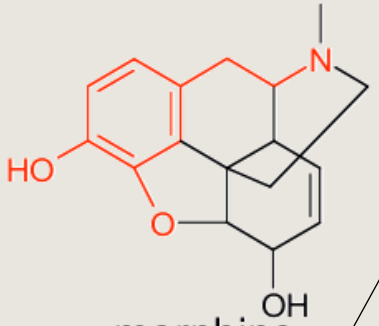

Hits:




adrenaline




apomorphine

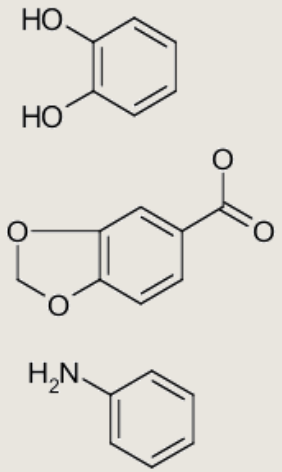



morphine

# BÚSQUEDA DE SUBESTRUCTURAS: FINGERPRINTS

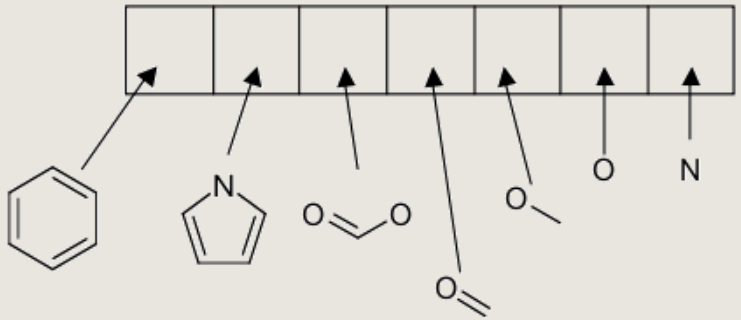
**Fingerprint:** representación abstracta de características o propiedades de una molécula (features)

- Presencia/ausencia de cada elemento
- Configuraciones electrónicas inusuales (carbono sp3, 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 | Query |
| 1 | 0 | 1 | 1 | 1 | 1 | 0 | ✓ passes |
| 1 | 0 | 0 | 0 | 0 | 0 | 1 | X does not pass |

**Un fingerprint**



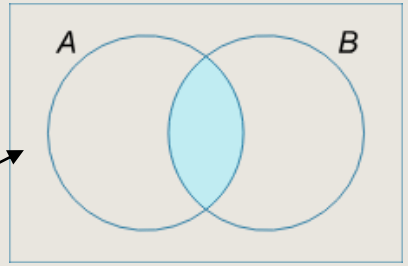
# BÚSQUEDA DE SUBESTRUCTURAS Y SIMILITUD: FINGERPRINTS

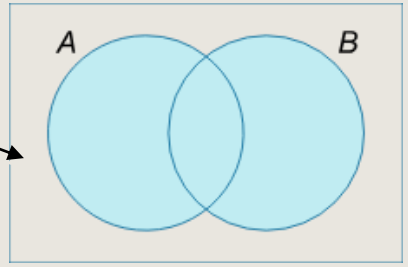
Ventajas: screening extremadamente rápido

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

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

```
X          10001101
Y          01010111
-----------------
X AND Y    00000101
-----------------
X OR  Y    11011111
```



# DISTANCE METRICS: SIMILARITY, DISIMILARITY

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

*(tambien llamado "Intersection Over Union")*

Compara similitudes entre conjuntos de datos finitos

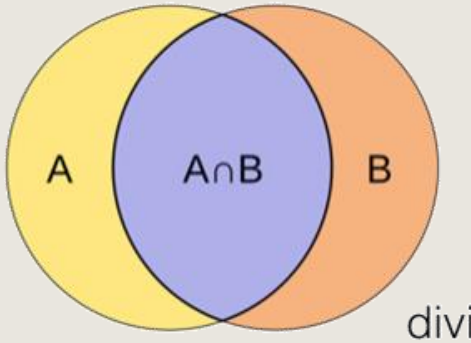
**Jaccard distance ($d_J$)**

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

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

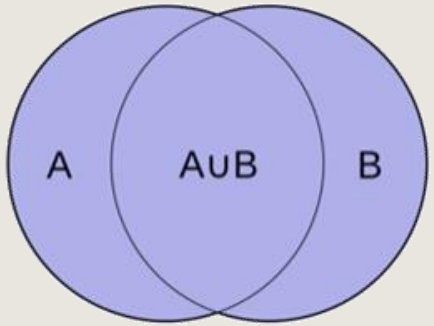
Cociente entre el tamaño de la intersección y el tamaño de la unión de los conjuntos de datos

The intersect of A & B




division

The union of A & B

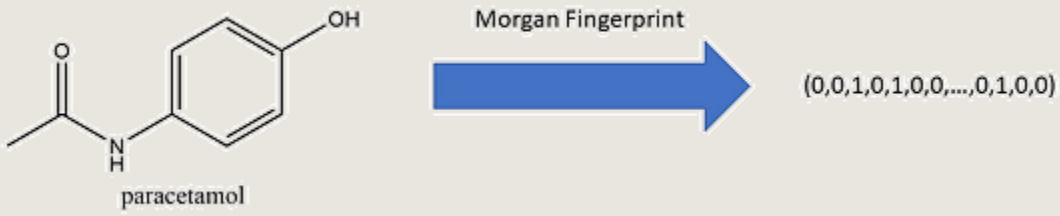


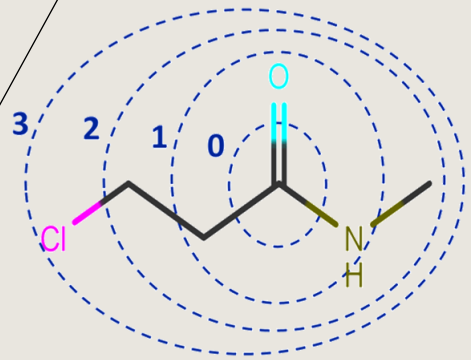

https://en.wikipedia.org/wiki/Jaccard_index

# EXTENDED CONNECTIVITY 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)**

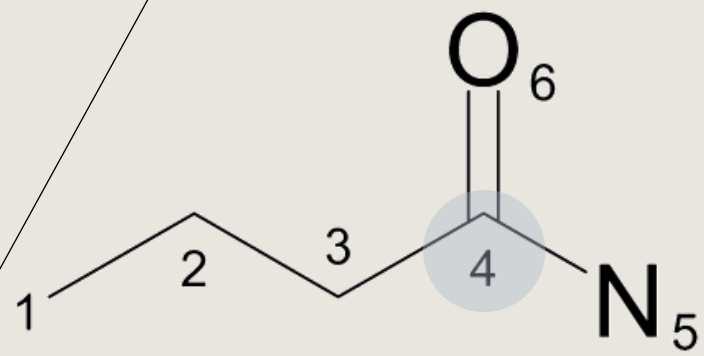
1024 or 2048 bits

# 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`



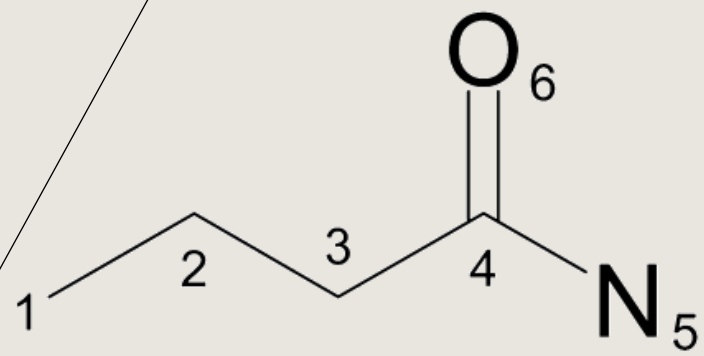

A beginner's guide for understanding Extended-Connectivity Fingerprints(ECFPs). Manish Kumar (2021). https://chemicbook.com/2021/03/25/a-beginners-guide-for-understanding-extended-connectivity-fingerprints.html

# 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)**




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**

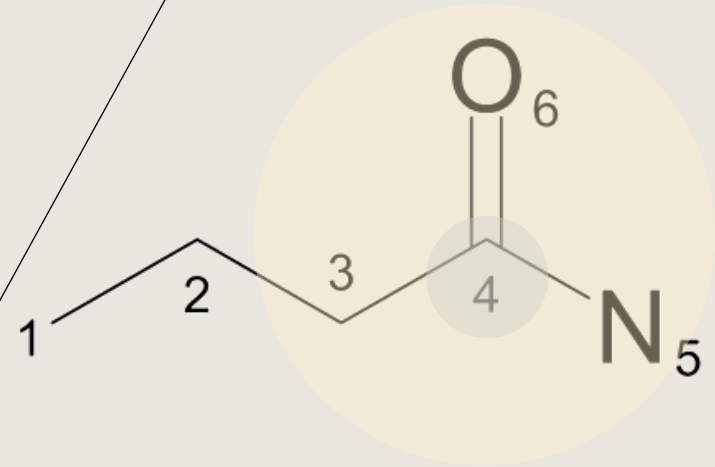
```
atomo 1 4940186308562569707
atomo 2 -7815985147897826576
atomo 3 -7815985147897826576
atomo 4 -5700861834356229464
atomo 5 -6296387744277800866
atomo 6 8618411755682373892
```

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

-6784272694619664722
```

repetimos para los 6 átomos

**List of features (12)**

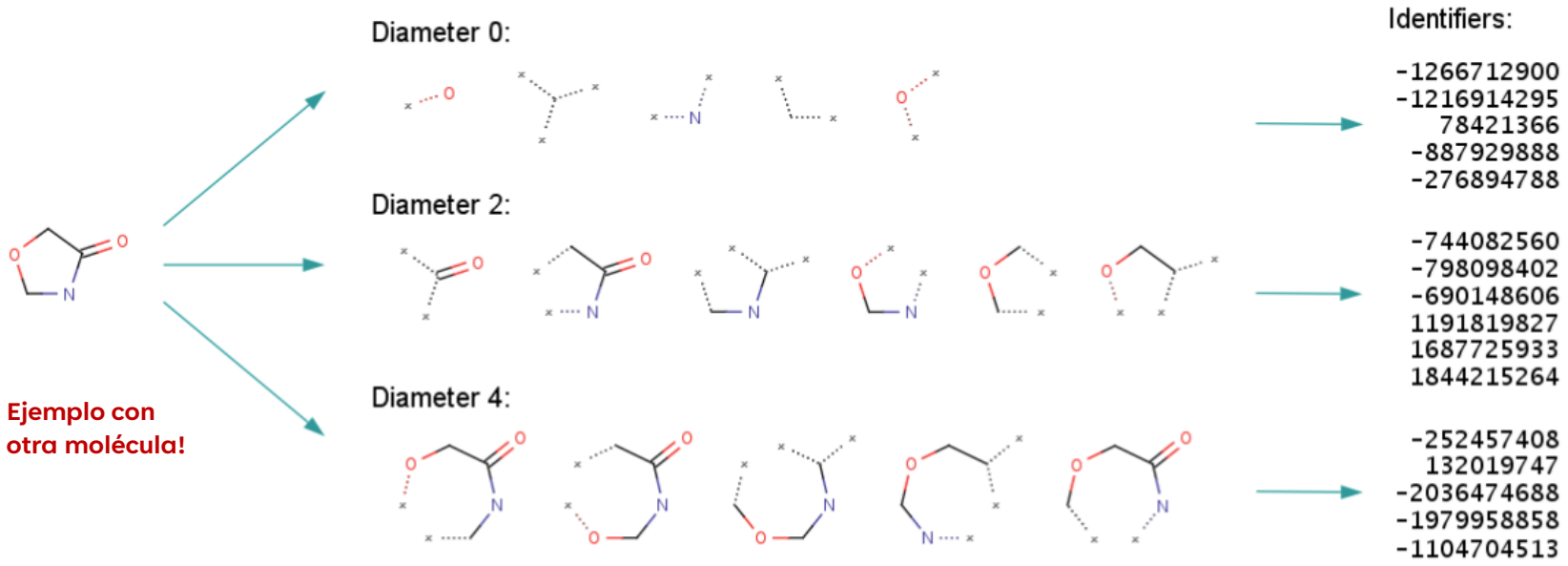



A beginner's guide for understanding Extended-Connectivity Fingerprints(ECFPs). Manish Kumar (2021). https://chemicbook.com/2021/03/25/a-beginners-guide-for-understanding-extended-connectivity-fingerprints.html

# 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 **radius = 4**
- ECFP6 = Extended Circular Fingerprint with **radius = 6**




Ejemplo con otra molécula!

Ejemplo con otra molécula!

**List of features (16 = 5 + 6 + 5)**

# 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**

24 | 11
2 2

**Ejemplos:**

```
132019747  % 1024 = 547
1687725933 % 1024 = 877
-798098402 % 1024 = 30
```

**Bit Collision:**

```
-14439656419269748   % 1024 = 908
-4080868480043360372 % 1024 = 908
```

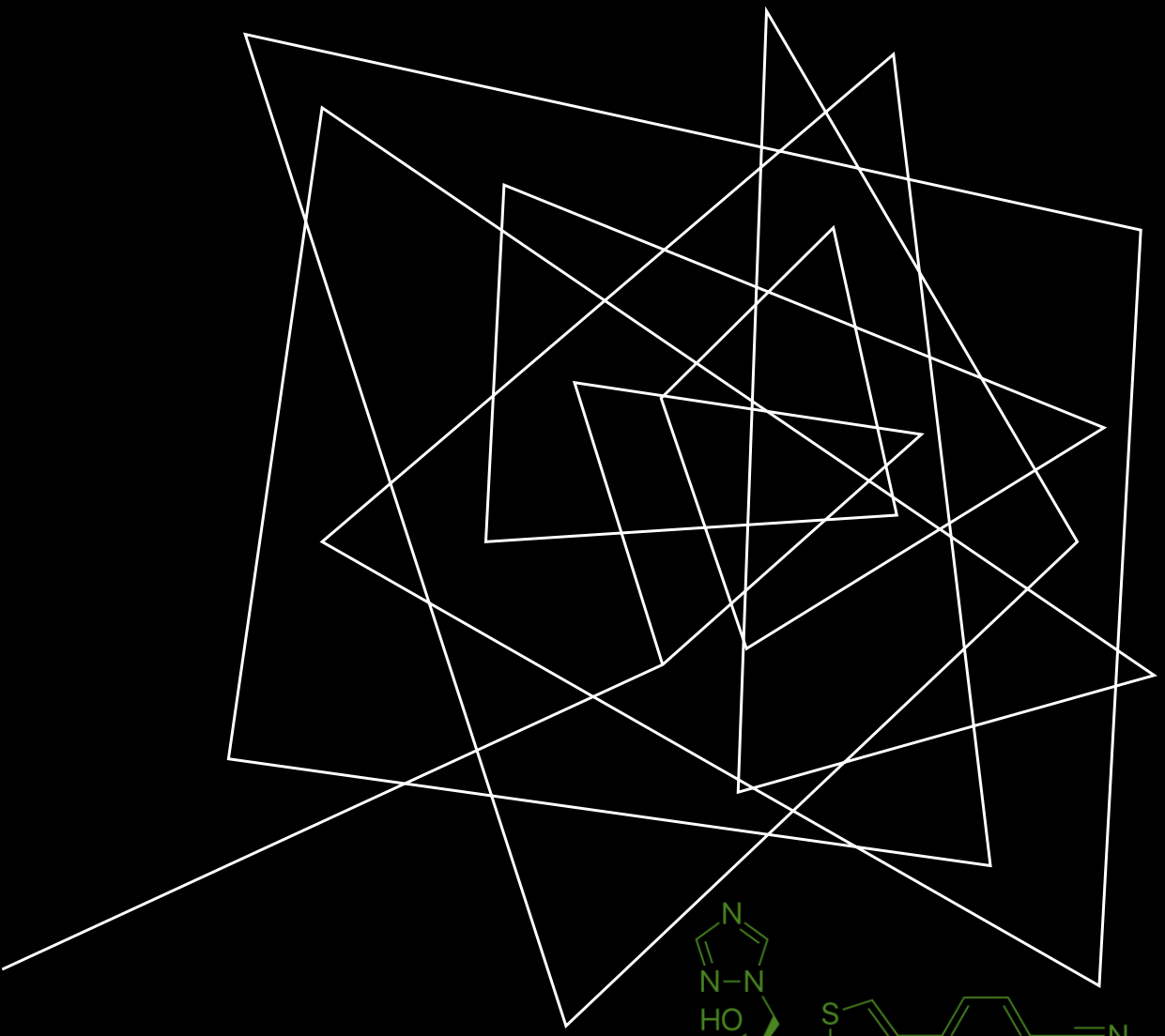
**Solution: increase fingerprint size**

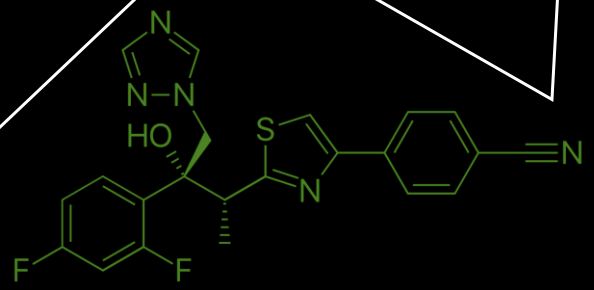
```
-14439656419269748   % 2048 = 908
-4080868480043360372 % 2048 = 1932
```

**Folding**

**Fixed-length binary representation**

0001000000001000001000000011001000010001000000000000000000100000[...]00001000000000000010

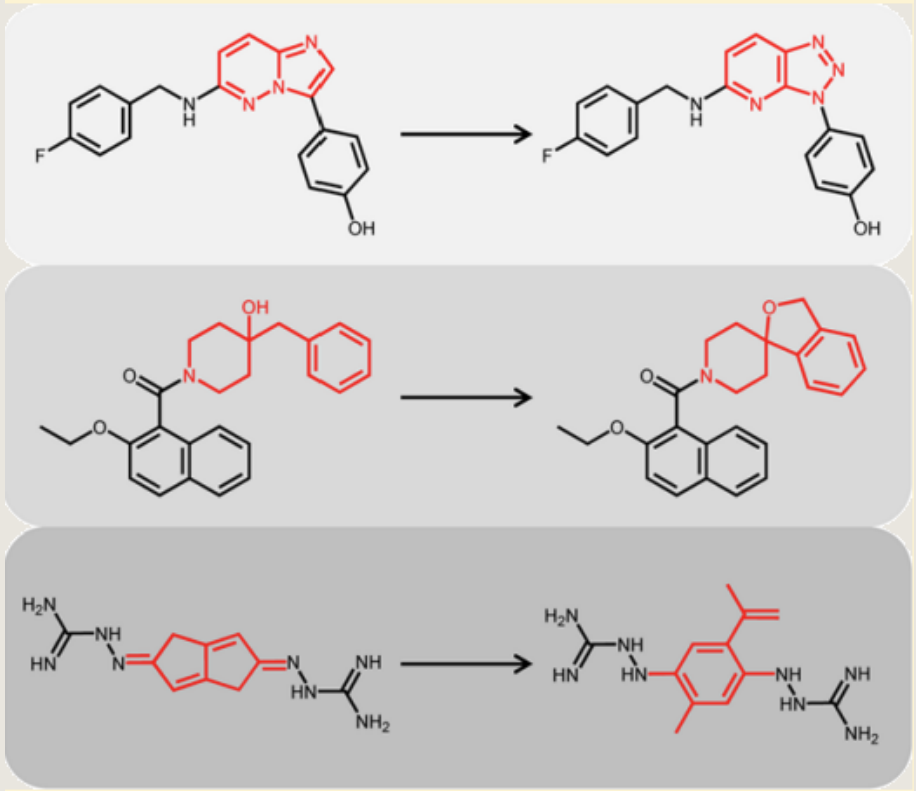





# EJEMPLOS

fingerprinting-with-rdkit.ipynb

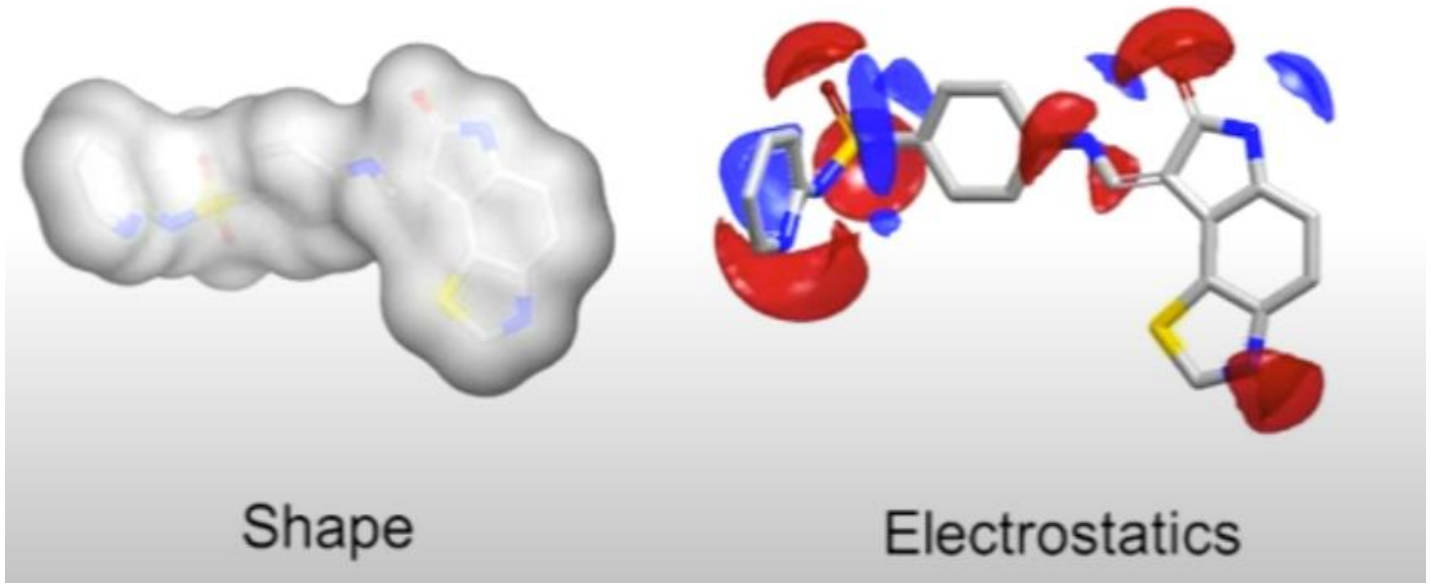
# SCAFFOLD HOPING



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

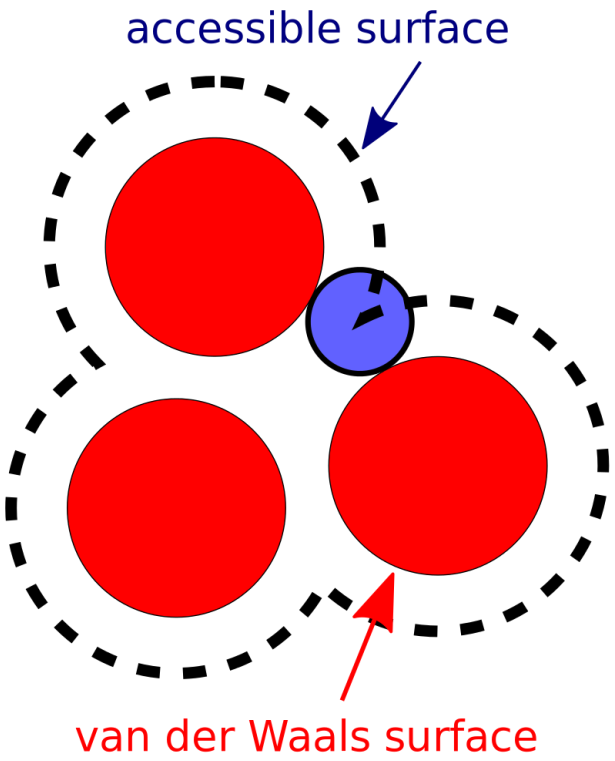
Hu Y, Stumpfe D, Bajorath J. Recent Advances in Scaffold Hopping. J Med Chem. 2017 Feb 23;60(4):1238-1246. doi: 10.1021/acs.jmedchem.6b01437. Epub 2016 Dec 21. PMID: 28001064.

# OTRAS REPRESENTACIONES DE MOLÉCULAS



# SOLVENT ACCESSIBLE SURFACE AREA CALCULATION

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



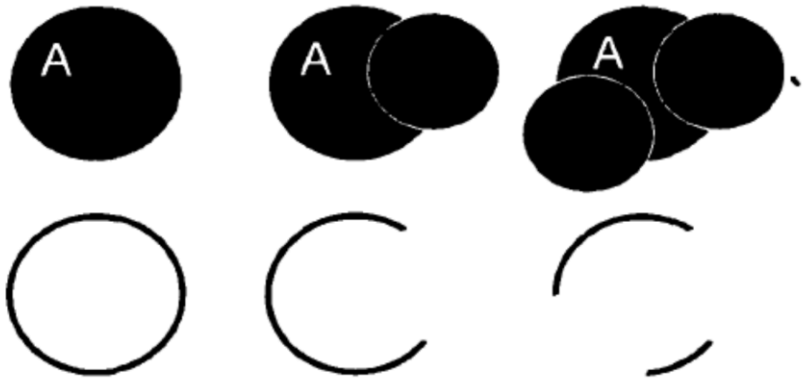



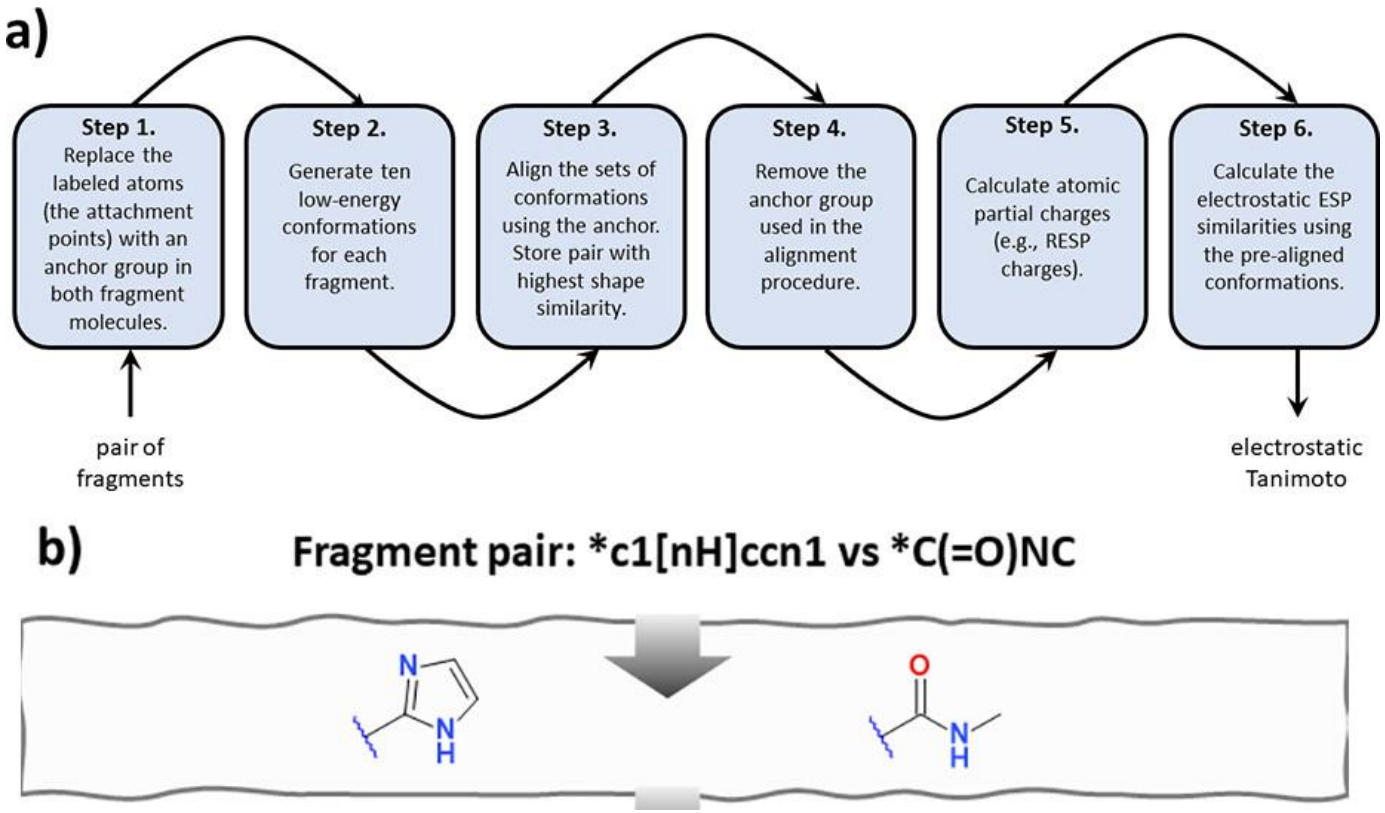

*Figure 1. Assuming spherical atoms, the surface area of atom A is the amount of surface area not contained in other atoms.*

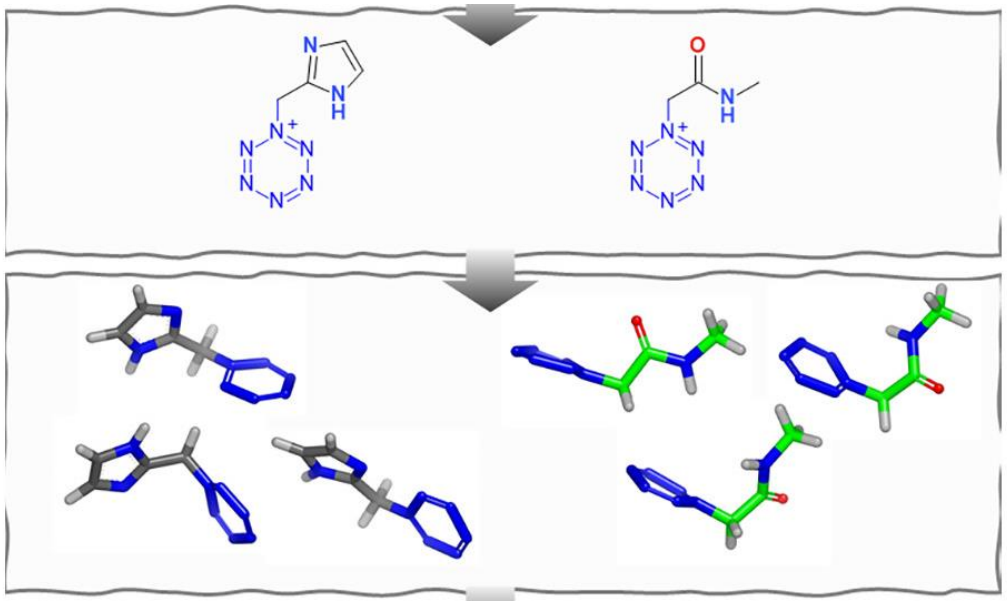
Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Res. 2016 Feb 18;5:189. doi: 10.12688/f1000research.7931.1. PMID: 26973785; PMCID: PMC4776673.

# ESP-SIM: COMPARISON OF ELECTROSTATIC POTENTIAL AND SHAPE

https://github.com/hesther/espsim

Bolcato G, Heid E, Boström J. On the Value of Using 3D Shape and Electrostatic Similarities in Deep Generative Methods. J Chem Inf Model. 2022 Mar 28;62(6):1388-1398. doi: 10.1021/acs.jcim.1c01535. Epub 2022 Mar 10. PMID: 35271260; PMCID: PMC8965872.

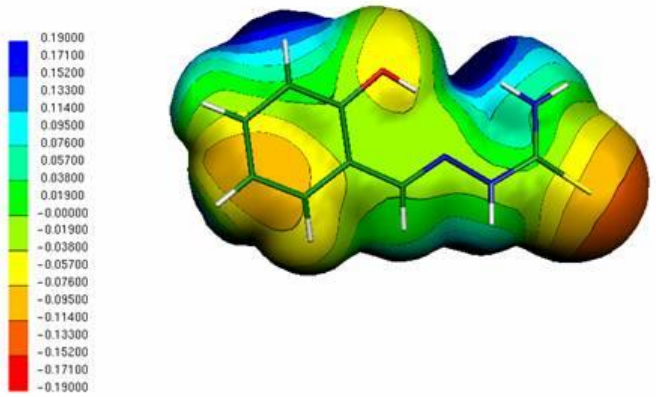


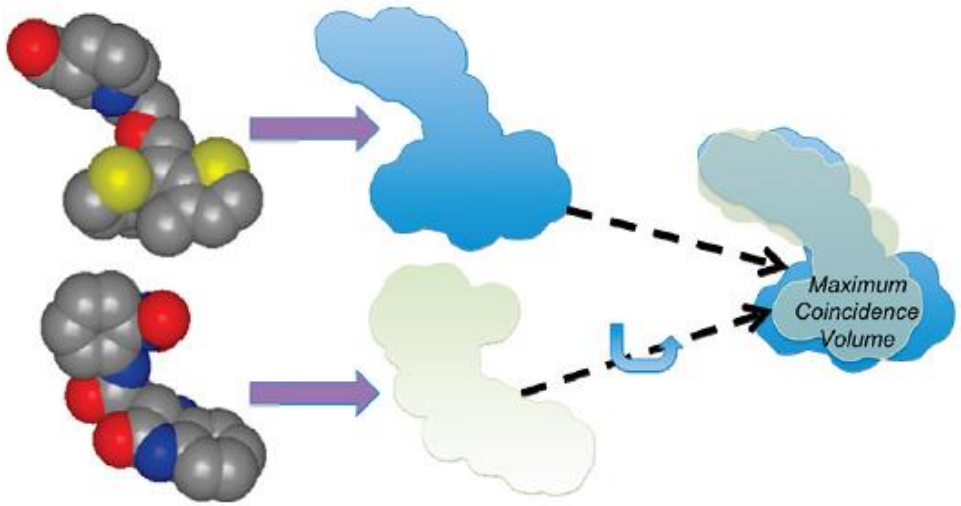



https://github.com/hesther/espsim/blob/master/scripts/short_demonstration.ipynb

# 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 volumenes**



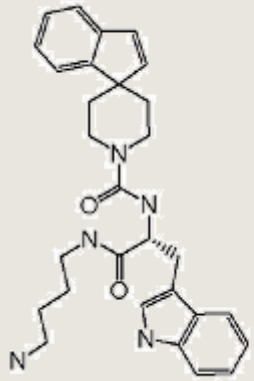



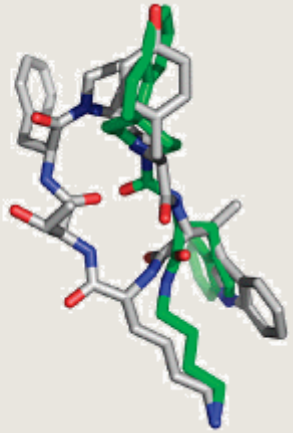

Molecular shape and medicinal chemistry: a perspective. 2010. A Nicholls *et al.* J Med Chem 53: 3862

# REPRESENTACIÓN DE FORMA (SHAPE)

Varias aplicaciones posibles:

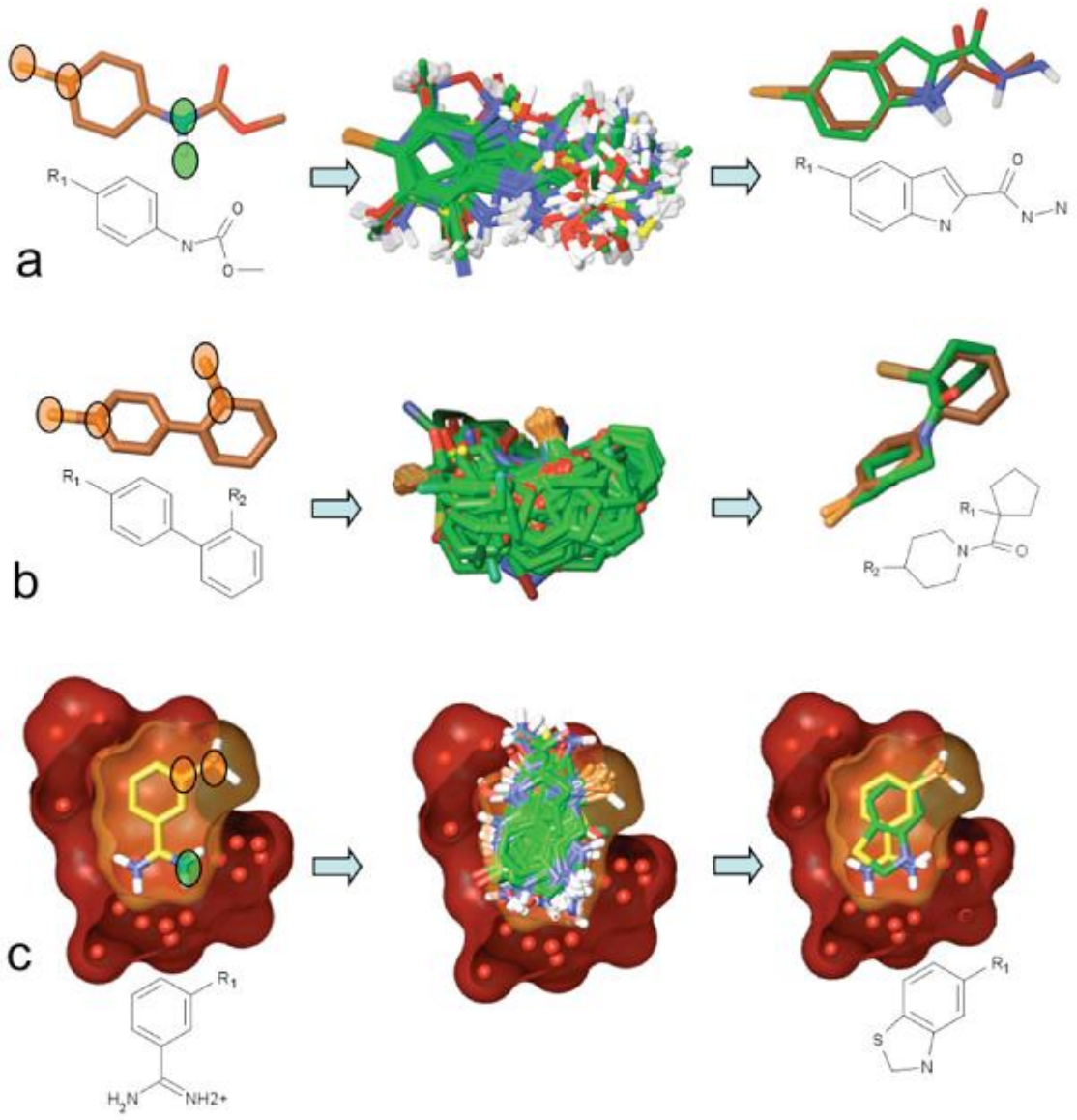
- Búsqueda de moléculas similares
  - En este caso la similitud es a nivel de forma
    - Se pueden agregar adicionalmente limitaciones
  - Varias implementaciones en la industria farmacéutica
    - Virtual screening
      - Varios casos de éxito conocidos
      - Merck, primer aplicación publicada del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



# REPRESENTACIÓN DE FORMA (SHAPE)

KIN: Bristol-Myers Squibb




Varias aplicaciones posibles:

**Lead optimization**

Uno cuenta con una molécula activa que quiere optimizar

**Scaffold Hoping**

Facilmente explorable utilizando metodos computacionales

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

# INTERVALO

15 minutos

# CALCULO DE PROPIEDADES

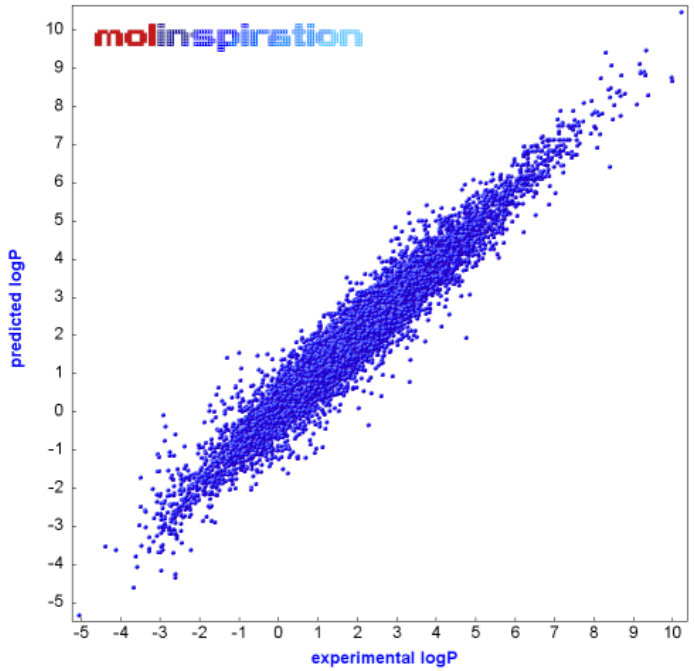
Enlaces rotables

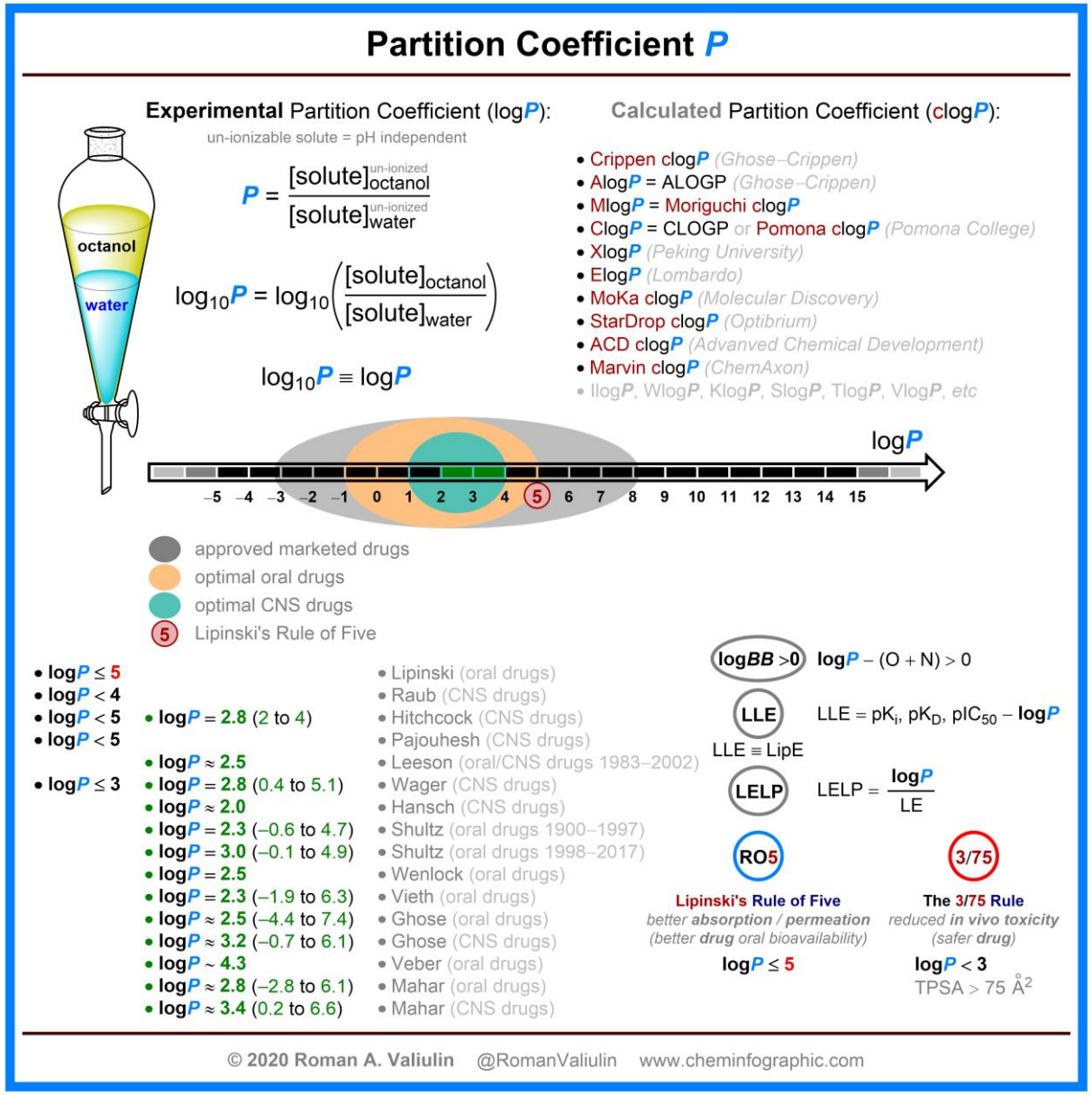
Dadores / Aceptores de puentes de hidrógeno

cLogP (coeficiente de partición octanol / agua)

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

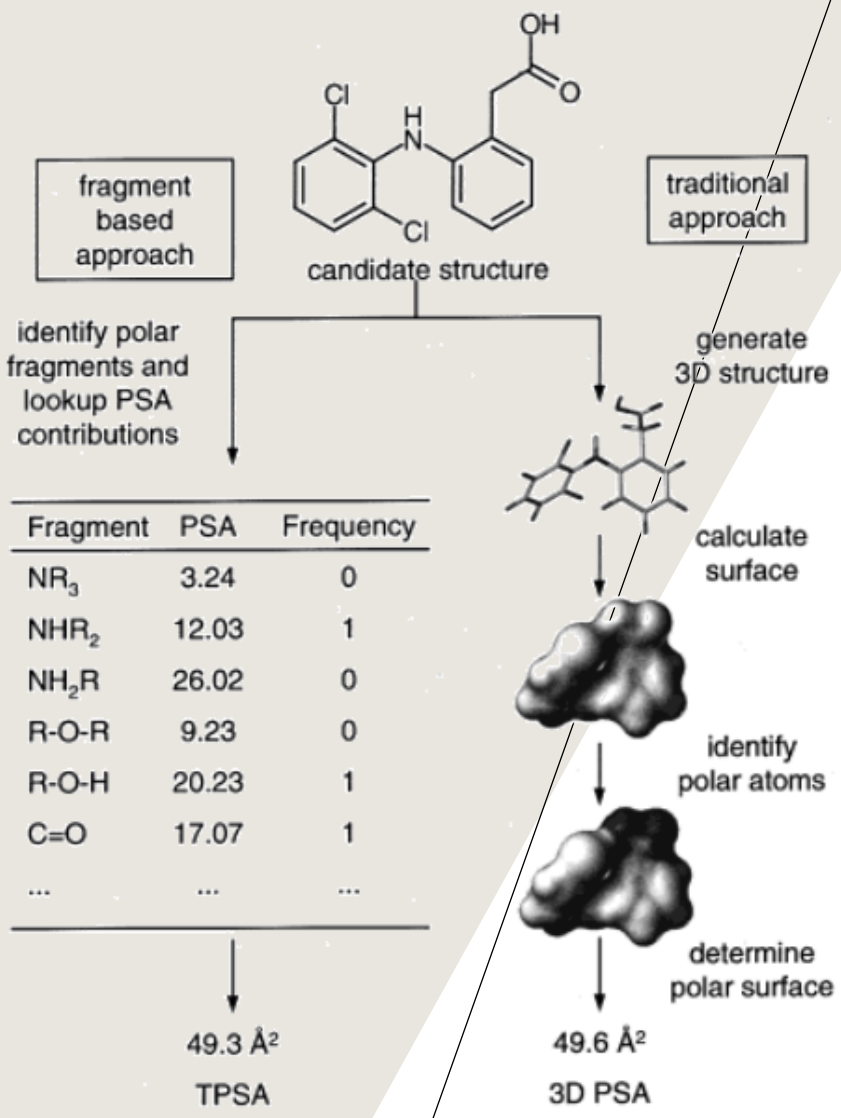
# LOGP PARTITION COEFFICIENT



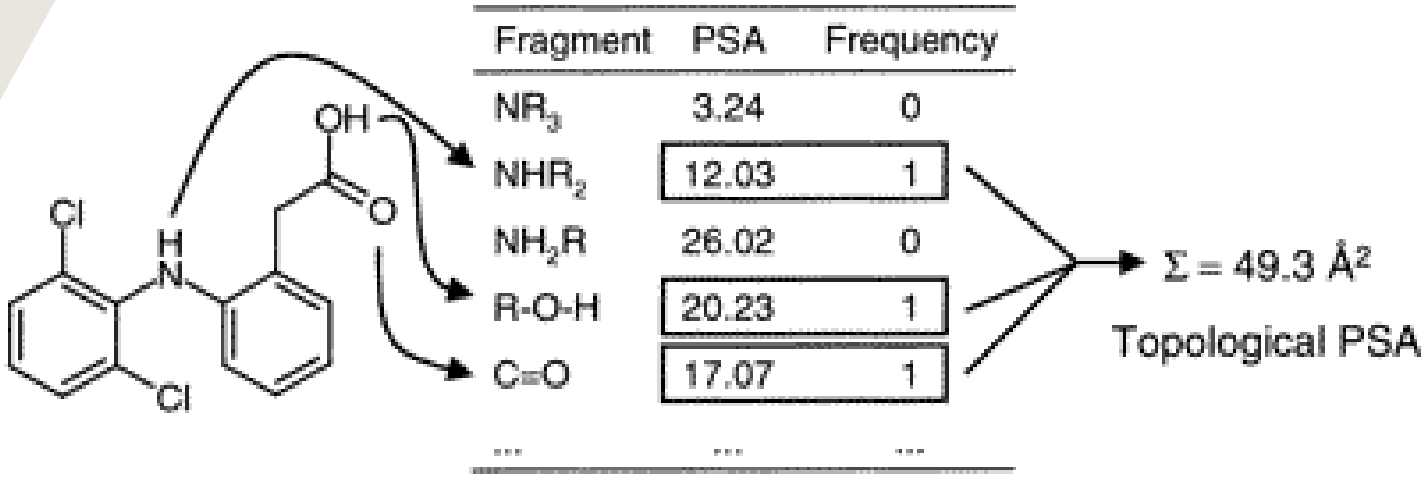




https://cheminfographic.wordpress.com/2020/05/01/partition-coefficient-p-logp/

# PSA / TPSA




| Fragment | PSA | Frequency |
|---|---|---|
| $NR_3$ | 3.24 | 0 |
| $NHR_2$ | 12.03 | 1 |
| $NH_2R$ | 26.02 | 0 |
| R-O-R | 9.23 | 0 |
| R-O-H | 20.23 | 1 |
| C=O | 17.07 | 1 |
| ... | ... | ... |




**Polar Surface Area (PSA, costoso)**

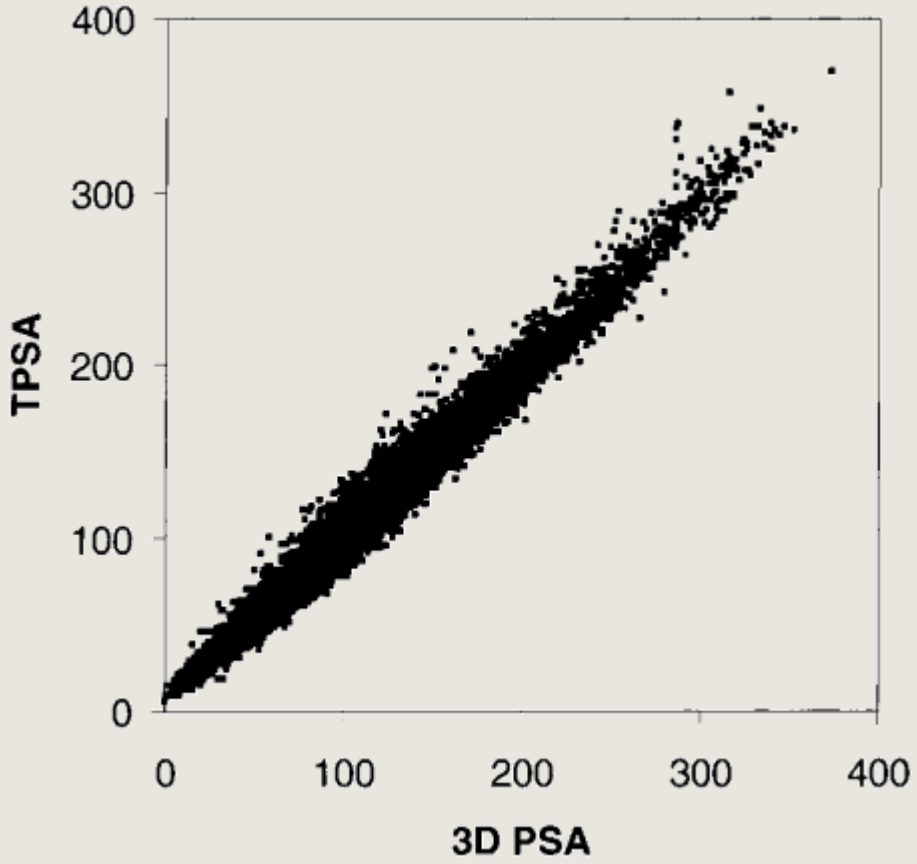
- Requiere generar conformeros 3D para calcular SA (Surface Area)

**Topological Polar Surface Area (TPSA)**

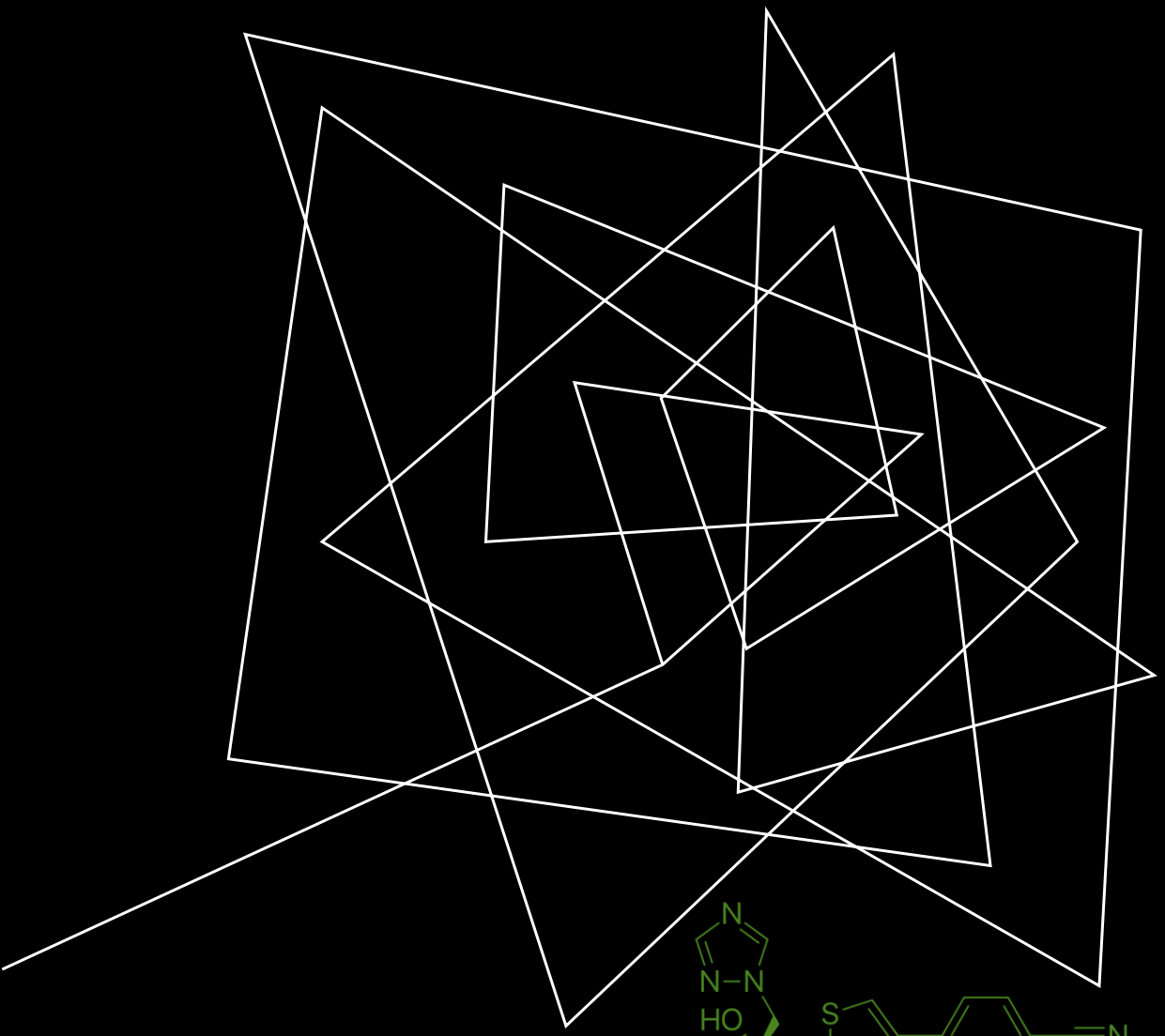
- Sumatoria de contribuciones tabuladas de fragmentos polares

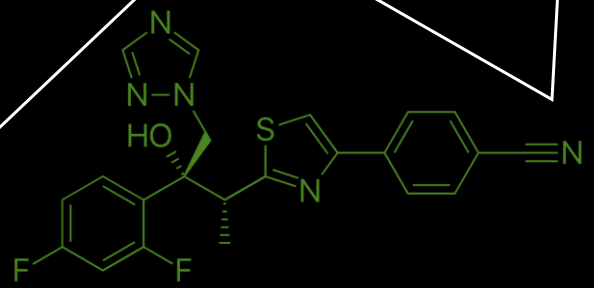
Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem. 2000 Oct 5;43(20):3714-7. doi: 10.1021/jm000942e. PMID: 11020286.

# TPSA VS PSA (3D)




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.






# ONTOLOGIAS

Organizando datos químicos

# QUE ES UNA ONTOLOGIA?

Es una formalización de un área del conocimiento mediante reglas

# QUÉ SIGNIFICA ONTOLOGÍA?

**Webster's Revised Unabridged Dictionary**

**Ontology**

- the things which exist
- *The department of the science of metaphysics which investigates and explains the nature and essential properties and relations of all beings, ...*

**The Free On-Line Dictionary of Computing**

**Ontology**

- Phylosophy: a systematic account of experience
- Artificial Inteligence: an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them. [...]
- Information Science: the hierarchical structuring of knowledge about things by subcategorizing them according to their essential (or at least relevant and/or cognitive) qualities.

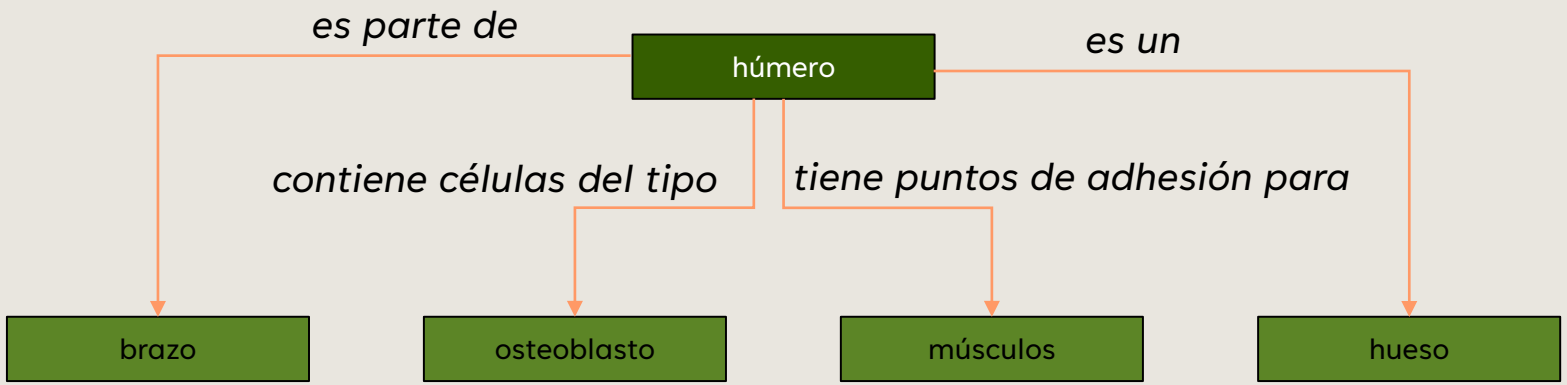
# EJEMPLO: UNA ONTOLOGIA ANATOMICA

**Una ontología es un área del conocimiento que ha sido formalizada:**

- **Términos** (conceptos) individuales
- **Afirmaciones** (reglas) que conectan términos entre sí

**Ejemplo: una ontología anatómica:**

- Términos: húmero, brazo, osteoblasto, músculo, hueso
- Conexiones: es parte de, contiene células del tipo, tiene puntos de adhesión para, es un



# ONTOLOGIAS QUIMICAS

ChEBI – Chemical Entities of Biological Interest

https://www.ebi.ac.uk/chebi/

Son varias ontologías agrupadas:

- **Molecular Structure**
  - **Términos:** hydrocarbons, carboxylic acids, tertiary amines
  - **Afirmaciones:** (reglas): is a (es un), has part (contiene)

caffeine (CHEBI:27732) **is a** purine alkaloid (CHEBI:26385)
caffeine (CHEBI:27732) **is a** trimethylxanthine (CHEBI:27134)

caffeine monohydrate (CHEBI:31332) **has part** caffeine (CHEBI:27732)
sodium caffeine benzoate (CHEBI:32140) **has part** caffeine (CHEBI:27732)

# ONTOLOGIAS QUIMICAS

ChEBI – Chemical Entities of Biological Interest

https://www.ebi.ac.uk/chebi/

Son varias ontologías agrupadas:

- **Role**
  - **Chemical Role**
    - ej ligand, inhibitor, surfactant
  - **Biological Role**
    - antibiotic, antiviral agent, coenzyme, hormone
  - **Application**
    - pesticide, antirheumatic drug, fuel
  - **Términos:** por ej alguna entidad molecular (o una parte)
  - **Afirmaciones:** (reglas): has role (tiene rol de)

caffeine (CHEBI:27732) **has role** environmental contaminant (CHEBI:78298)
caffeine (CHEBI:27732) **has role** adenosine A2A receptor antagonist (CHEBI:53121)
caffeine (CHEBI:27732) **has role** food additive (CHEBI:64047)

# CHEBI: AFIRMACIONES (REGLAS)

**Relationship Types**

- is a
- has part
- is conjugate base of
- is conjugate acid of
- is tautomer of
- is enantiomer of
- has functional parent
- has parent hydride
- is substituent group from
- has role

**CHEBI:32816 pyruvic acid** ← CHEBI:15361 pyruvate

**CHEBI:32816 pyruvic acid** → CHEBI:15361 pyruvate

**CHEBI:17696 isocyanuric acid** ↔ CHEBI:38028 cyanuric acid

**CHEBI:15396 (*R*)-camphor** ↔ CHEBI:15397 (*S*)-camphor

**CHEBI:17026 progesterone**

↑ CHEBI:16973 11-deoxycorticosterone

# CHEBI: AFIRMACIONES (REGLAS)

**Relationship Types**

- is a
- has part
- is conjugate base of
- is conjugate acid of
- is tautomer of
- is enantiomer of
- has functional parent
- has parent hydride
- is substituent group from
- has role

**CHEBI:16482 naphthalene**

CHEBI:50715 methylnaphthalene

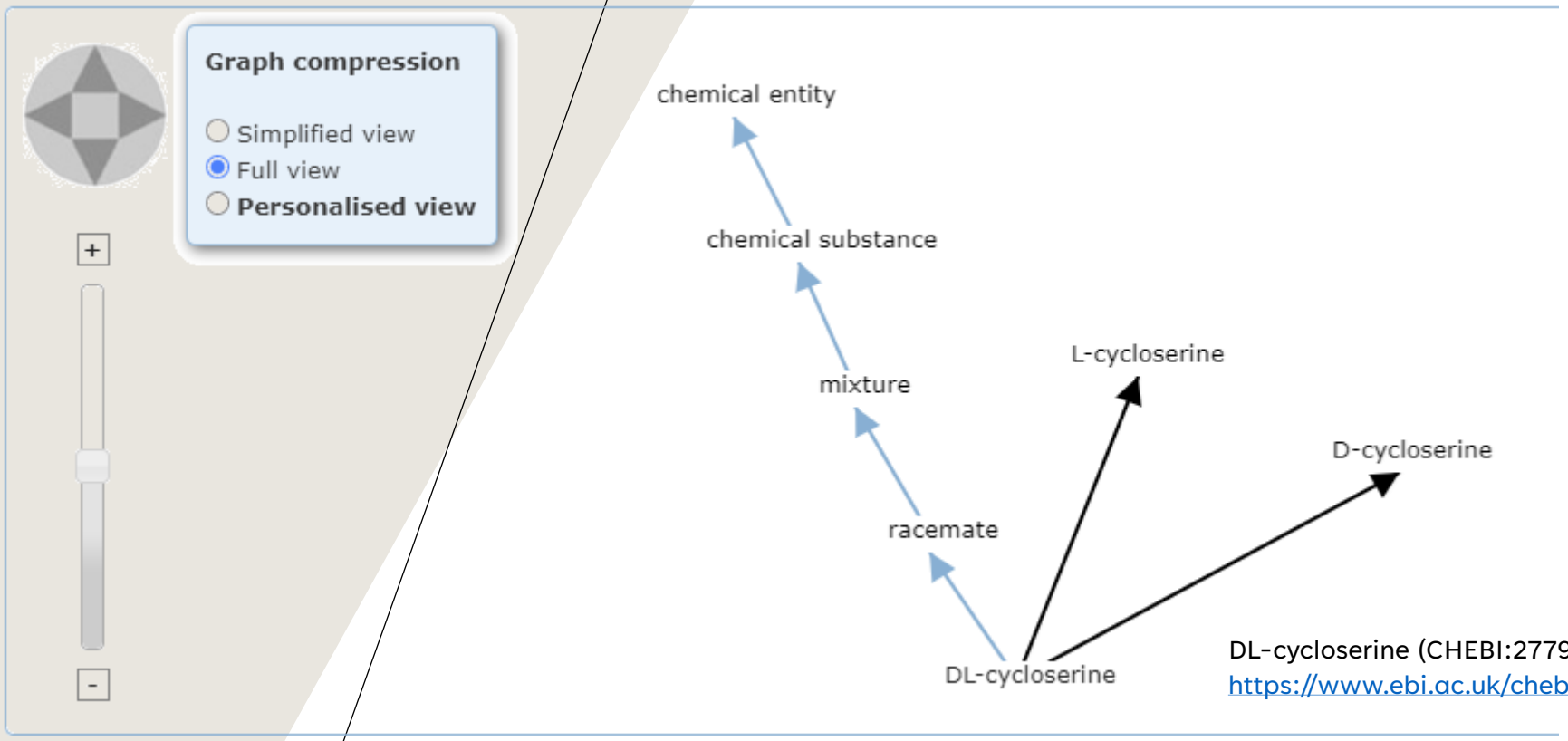
**CHEBI:16414 L-valine**

CHEBI:32853 L-valyl group

CHEBI:35482 opioid analgesic

**CHEBI:17303 morphine**

# NAVEGACIÓN GRAFICA DE LA ONTOLOGIA



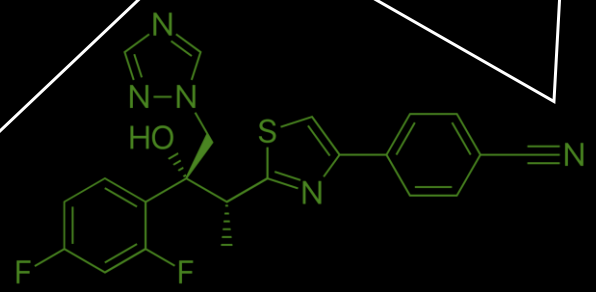

DL-cycloserine (CHEBI:27792)
https://www.ebi.ac.uk/chebi/chebiOntology.do?chebiId=CHEBI:27792

# EJEMPLOS

https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:27732

https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:78298

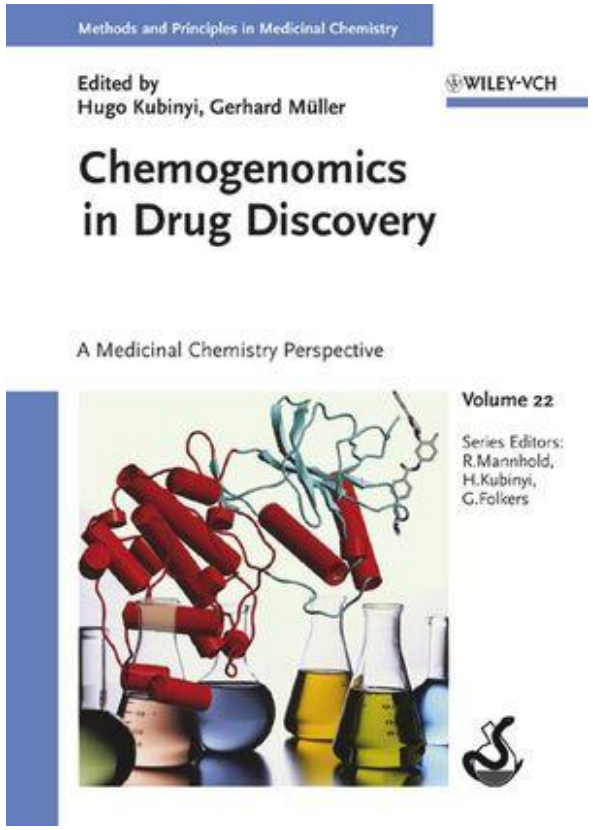
...



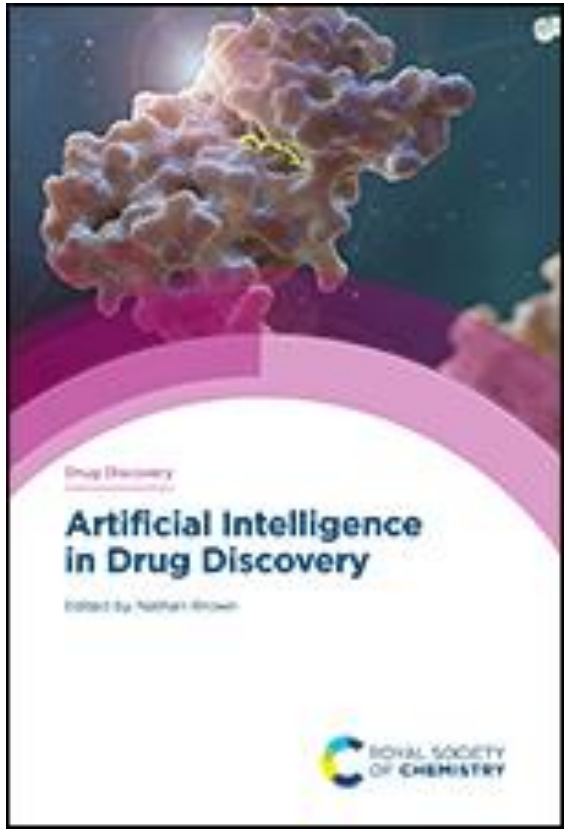
# BIBLIOGRAFÍA | MATERIAL DE LECTURA

ChEBI
USER MANUAL

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




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+A+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

RDKit
Open-Source Cheminformatics and Machine Learning

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