
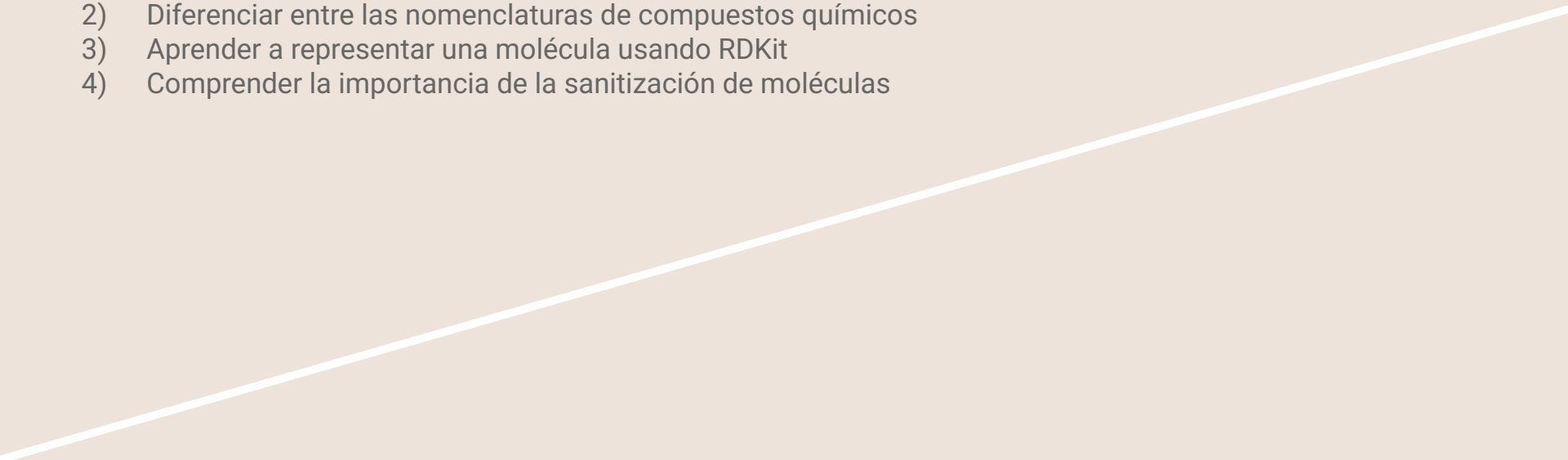


TP1

Introducción a Bases de Datos y Softwares Quimioinformáticos



Objetivos

- 1) Conocer las bases de datos Quimioinformáticas
 - 2) Diferenciar entre las nomenclaturas de compuestos químicos
 - 3) Aprender a representar una molécula usando RDKit
 - 4) Comprender la importancia de la sanitización de moléculas
- 

Organización de la clase

9:00 a 9:30	Introducción al TP
9:30 a 10:30	Trabajo en la guía de ejercicios (Parte 1: Bases de Datos Quimioinformáticas)
10:30 a 11:00	Recreo
11:00 a 12:00	Trabajo en la guía de ejercicios (Parte 2: Softwares Quimioinformáticos)
12:00 a 13:00	Lectura de paper y puesta en común

Base de Datos

PubChem

ChEMBL



Softwares Quimioinformáticos



Open-Source Cheminformatics
and Machine Learning

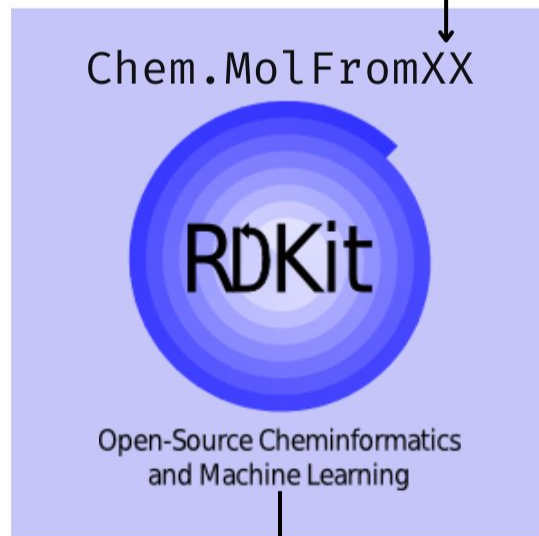
Softwares Quimioinformáticos

Tipo de formato	Es admitido para generar Mol?
SMILES	
InChi	
InChiKey	
SMARTS	

Softwares Quimioinformáticos

Tipo de formato	Es admitido para generar Mol?
SMILES	Si
InChi	Si
InChiKey	No
SMARTS	Si

Smiles = Inchi = Smarts = MolFile =
'Cc1ccccc1' 'InChI=1S/C...' 'ccO' 'input.mol'

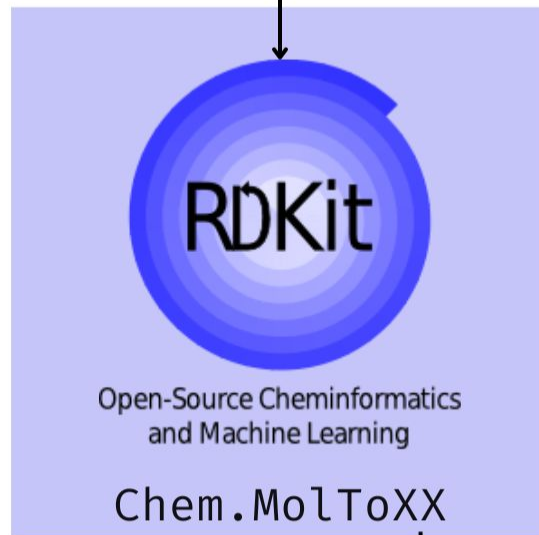


Mol

<rdkit.Chem.rdchem.Mol object at 0x...>

Mol

<rdkit.Chem.rdchem.Mol object at 0x...>



Smiles =

InchiKey =

Inchi =

Smarts =

MolFile =

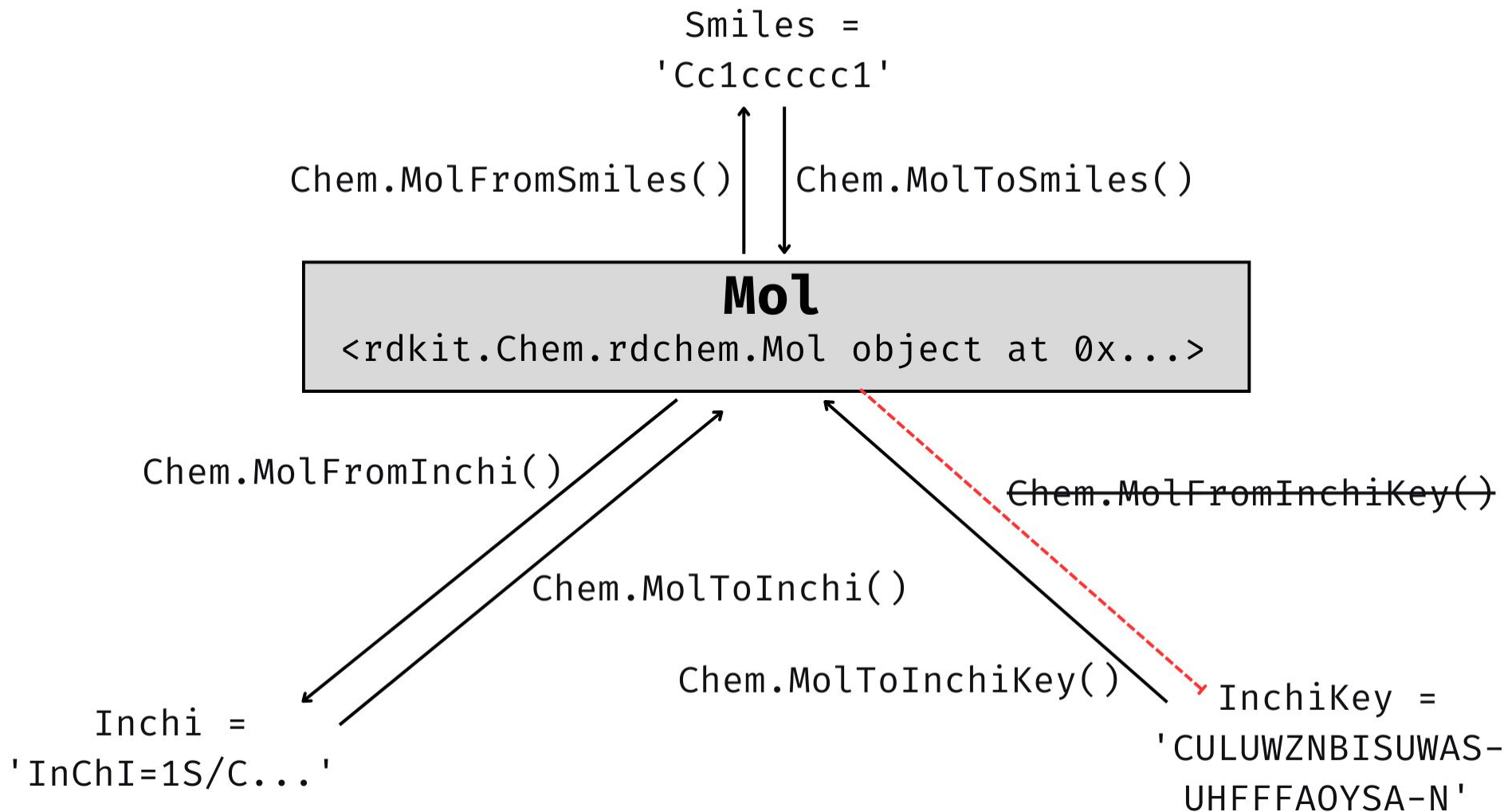
'Cc1ccccc1'

'CULUWZNBISUWAS-
UHFFFAOYSA-N'

'InChI=1S/C...'

'cc0'

'output.mol'



Inchi vs InchiKey

El nomenclador Inchi o InchiKey son los más recomendados para usar a la hora de hacer análisis computacionales.

InChI permite que la estructura química esté escrita de manera tal que sea fácil de almacenar y encontrar en la web y las plataformas informáticas. InChI promueve los "Principios rectores FAIR para la gestión y administración de datos científicos". FAIR se publicó en 2016 para proporcionar pautas para mejorar la Encontrabilidad, Accesibilidad, Interoperabilidad y Reutilización (Findability, Accessibility, Interoperability, and Reuse) de activos digitales.

InChI proporciona 'Encontrabilidad' para estructuras químicas y extiende la interoperabilidad entre plataformas, las cuales fomentan la accesibilidad y la reutilización.

Softwares Quimioinformáticos

Búsqueda de subestructuras

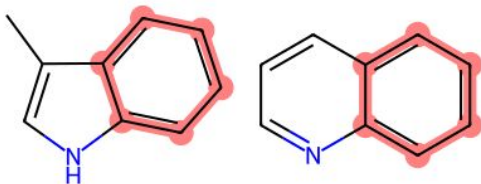
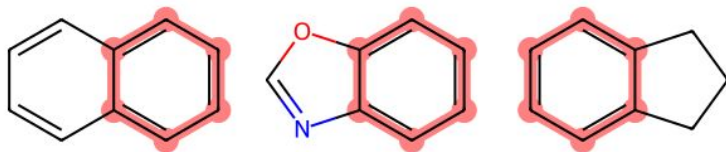
SMILES

Vs

SMARTS

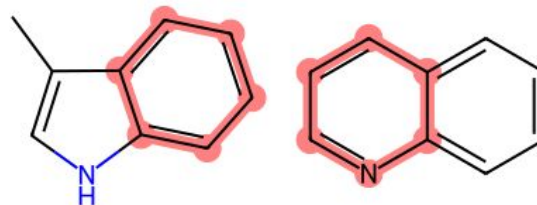
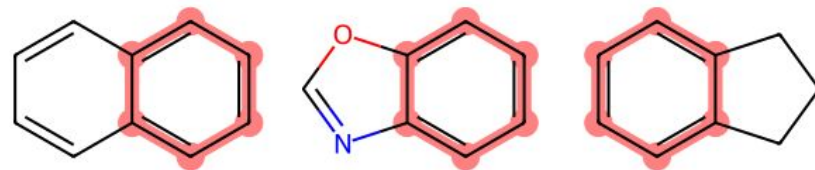
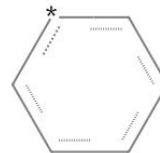
SMILES

benceno = 'C1=CC=CC=C1'



SMARTS

benceno = 'c1cccc[*]c1'



Cierre

Bento et al. *J Cheminform* (2020) 12:51
<https://doi.org/10.1186/s13321-020-00456-1>

Journal of Cheminformatics

METHODOLOGY

Open Access

An open source chemical structure curation pipeline using RDKit



A. Patrícia Bento¹ , Anne Hersey¹, Eloy Félix¹, Greg Landrum², Anna Gaulton¹, Francis Atkinson^{1,3}, Louisa J. Bellis^{1,4}, Marleen De Veij¹ and Andrew R. Leach^{1*}

<https://doi.org/10.1186/s13321-020-00456-1>

Background

The ChEMBL database is one of a number of public databases that contain bioactivity data on small molecule compounds curated from diverse sources.

Incoming compounds are typically not standardised according to consistent rules. In order to maintain the quality of the final database and to easily compare and integrate data on the same compound from different sources it is necessary for the chemical structures in the database to be appropriately standardised.

Keypoints

The compound structures and associated bioactivity data are extracted on a regular basis primarily from the medicinal chemistry literature.

Chemical structures from the primary scientific literature are mostly manually drawn from the structural information in the papers prior to loading into ChEMBL. These structures are often represented in the publication as Markush structures with different R-groups shown in SAR (structure–activity relationship) tables.

The challenge was the registration of compounds in a database and determining chemical uniqueness

The decision was made to build a curation pipeline around the widely used open-source RDKit toolkit and its implementation of the MolVS molecule validation and standardisation tool

MolVS: Molecule Validation and Standardization

MolVS is a molecule validation and standardization tool, written in Python using the RDKit chemistry framework.

Methods

The new ChEMBL curation pipeline comprises three processes:

1. Checker: identifies and validates structures and identifies problems before they are added to the database
2. Standardizer: processes (standardised) chemical structures according to a set of predefined rules
3. GetParent: generates parent structures based on a set of rules and defined lists of salts and solvents

This was performed on ChEMBL v26

Checker component

The Checker component validates structures prior to the compounds being loaded into ChEMBL.

If an error or problem is detected in the structure a score is reported for the molecule; the score increases with the severity of the perceived problem.

In the majority of cases compounds are loaded into the database even if a warning flag is identified.

The scores are recorded but at this point errors are not corrected. Instead, they are prioritised and subjected to subsequent manual curation, as time and degree of seriousness permits.

Checker component

Table 1 Penalty scores and annotation that are output from the *Checker* module

Penalty score	Penalty explanation
7	Error-9986 (Cannot process aromatic bonds) Illegal input InChI: Unknown element(s)
6	All atoms have zero coordinates InChI: Accepted unusual valence(s) InChI: Empty structure Molecule has 3D coordinates Molecule has a radical that is not found in the known list Molecule has six (or more) atoms with exactly the same coordinates Number of atoms less than 1 Polymer information in mol file V3000 mol file
5	InChI_RDKit/Mol stereo mismatch Mol/Inchi/RDKit stereo mismatch RDKit_Mol/InChI stereo mismatch Molecule has a bond with an illegal stereo flag Molecule has a bond with an illegal type Molecule has a crossed bond in a ring Molecule has two (or more) atoms with exactly the same coordinates
2	InChI_Mol/RDKit stereo mismatch Molecule has a stereo bond in a ring Molecule has an atom with multiple stereo bonds Molecule has a stereo bond to a stereocenter Molecule has the 3D flag set for a 2D conformer Other InChI Warnings

7 is the most serious penalty score and 2 the least important

Standardizer component

The standardisation rules implemented in the ChEMBL database are based largely on the FDA/IUPAC guidelines.

Whilst the aim is to adhere to these rules as closely as possible, the practical reality is that submitted compounds are sometimes drawn imperfectly or the structures are ambiguously defined in the original publication or by the depositor.

An automated standardiser can only safely correct some of the potential issues and the standardisation rules.

Standardizer component

1. Standardise unknown stereochemistry

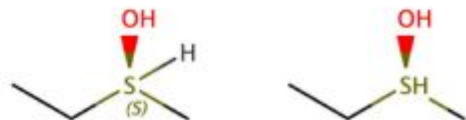
- Change "wiggly" bonds on sp^3 carbons denoting unknown stereo to show no stereo
- Set either or unknown cis/trans bonds to crossed bonds instead of showing them as "wiggly" bonds

Before Standardisation **After Standardisation**



- Clear S Group data from the molfile
- Generate a kekulé form of the structure
- Remove explicit H atoms except:
 - Hs where an isotope of hydrogen has been specifically set
 - Hs that have a wedged or dashed bond to them

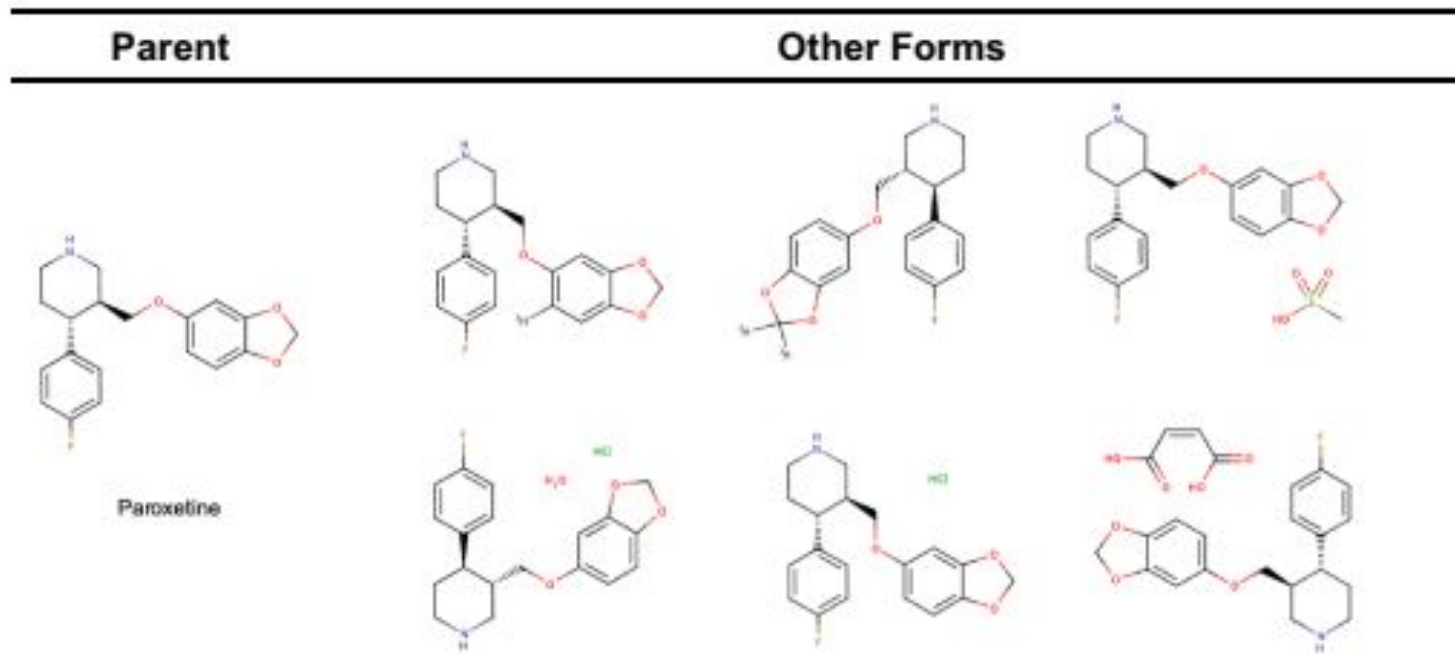
- Hs bonded to atoms with tetrahedral stereochemistry set ("Chiral Hs"). This is an example:



chemistry set ("Chiral Hs"). This is an example:

GetParent component

Many compound registration systems, including the ChEMBL database, identify compounds that are related by virtue of being a salt form of a common parent structure.



Availability of structure curation pipeline

The code for the pipeline has all been developed using the RDKit toolkit (version 2019.09.2.0). It is open source and **publicly available in GitHub**, currently as version 1.0.0. A conda package is also available to facilitate installation.

The Standardizer, Checker and GetParent functions are also integrated in the ChEMBL Beaker web services and can be used in this way via the 'check', 'getParent' and 'standardize' endpoints.

Any new features developed by the ChEMBL group will be added to the repository and **comments and suggestions from others are welcomed**.

Results

A chemical curation pipeline has been developed using the open source toolkit RDKit.

It comprises three components:

- a Checker to test the validity of chemical structures and flag any serious errors;
- a Standardizer which formats compounds according to defined rules and conventions
- and a GetParent component that removes any salts and solvents from the compound to create its parent.

This pipeline has been applied to the latest version of the ChEMBL database as well as uncured datasets from other sources to test the robustness of the process and to identify common issues in database molecular structures.

Conclusion

All the components of the structure pipeline have been made freely available for other researchers to use and adapt for their own use.

The code is available in a GitHub repository (https://github.com/chembl/ChEMBL_Structure_Pipeline) and it can also be accessed via the ChEMBL Beaker webservice.

It has been used successfully to standardise the nearly 2 million compounds in the ChEMBL database and the compound validity checker has been used to identify compounds with the most serious issues so that they can be prioritised for manual curation.

Objetivos

- 1) Conocer las bases de datos Quimioinformáticas
 - 2) Diferenciar entre las nomenclaturas de compuestos químicos
 - 3) Aprender a representar una molécula usando RDKit
 - 4) Comprender la importancia de la sanitización de moléculas
- 