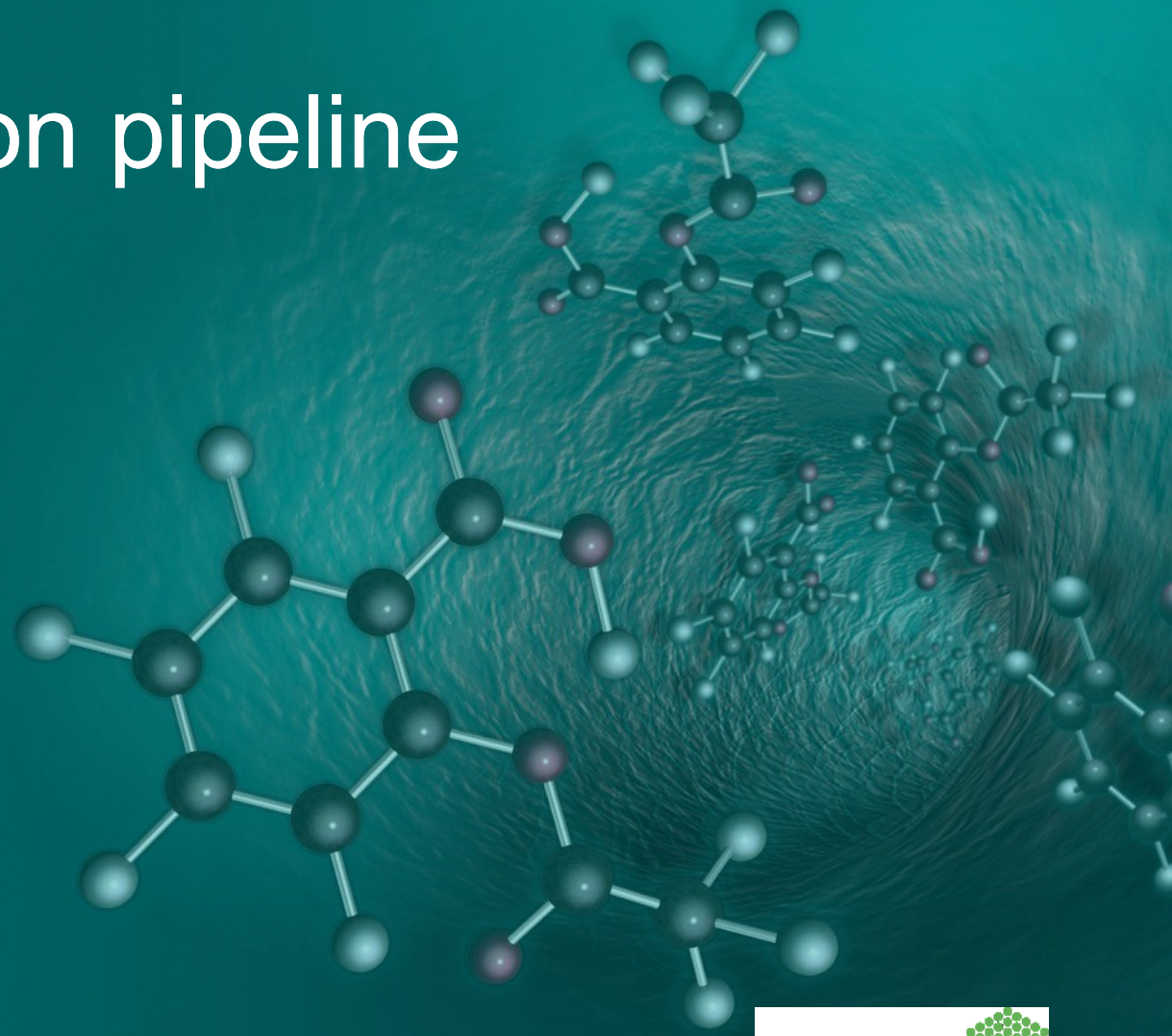
ChEMBL’s open source chemical structure curation pipeline



Anne Hersey

Patricia Bento (patricia@ebi.ac.uk)

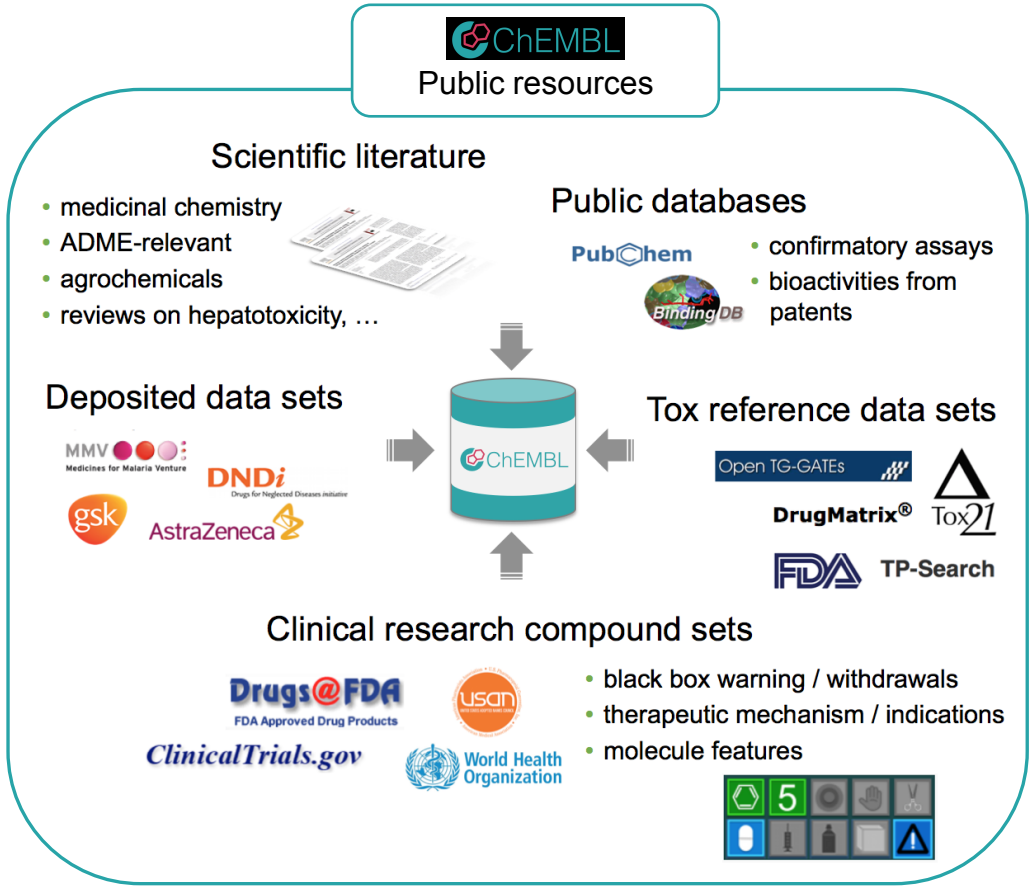
2020 RDKit User Group Meeting

Outline

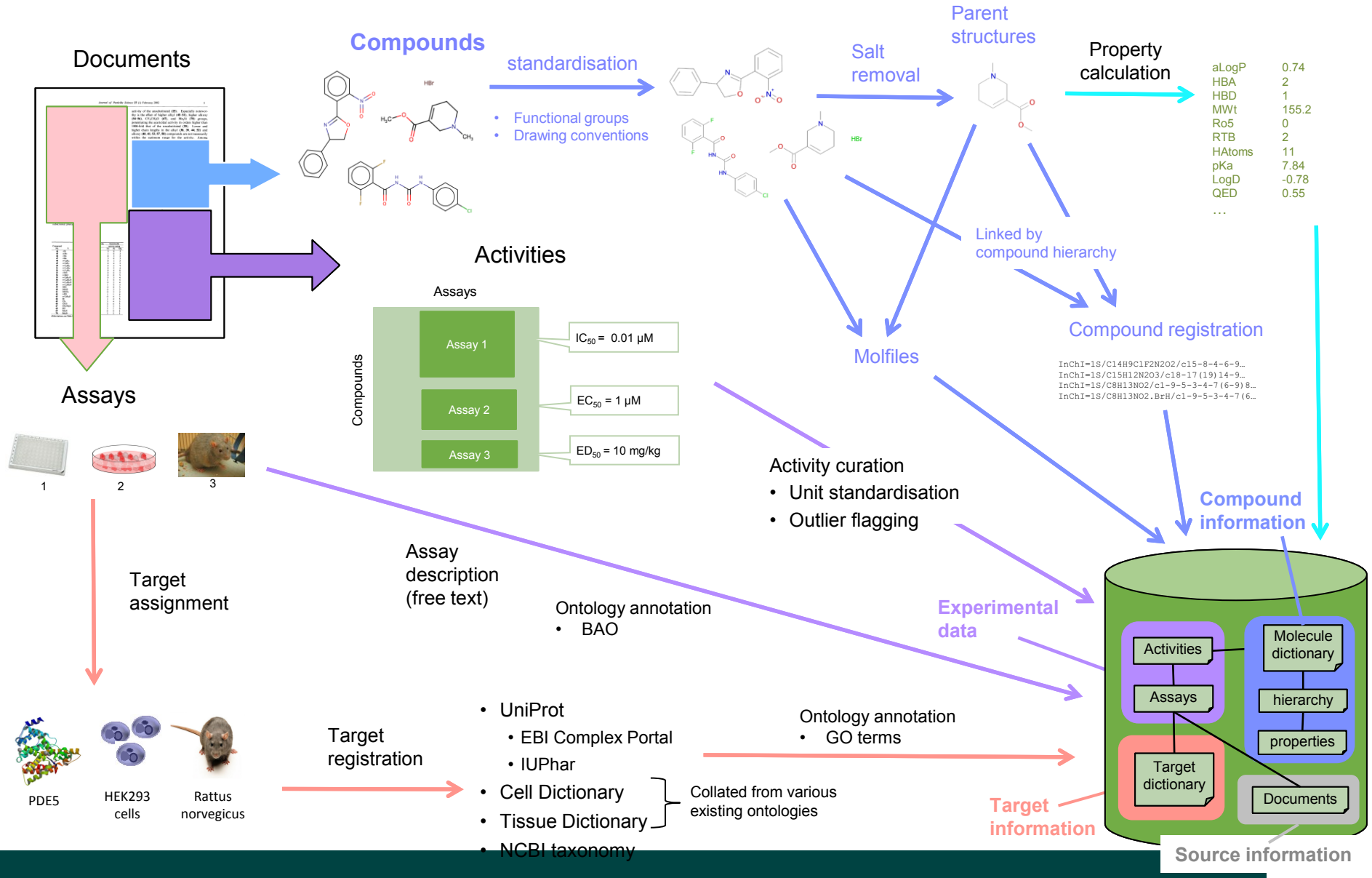
- ChEMBL Database overview and content
- Motivation
- ChEMBL Structure Pipeline
  - *Checker*
  - *Standardizer*
  - *GetParent*
- Code availability
- More information
- Acknowledgements

ChEMBL Database

- Bioactivity data from:
  - Key MedChem journals (e.g. J. Med. Chem., Bioorg. Med. Chem.)
  - Deposited datasets (e.g. MMV, DNDi, CO-ADD antimicrobial screening)
  - Public databases subsets (e.g. PubChem, BindingDB)
  - Contributed datasets (e.g. GSK kinase inhibitors, AstraZeneca DMPK & Physicochemical data)
  - Review articles, book chapters, etc
- Integrated with data on:
  - Clinical development and marketed drugs (from ClinicalTrials.gov, FDA, USANs, INNs)

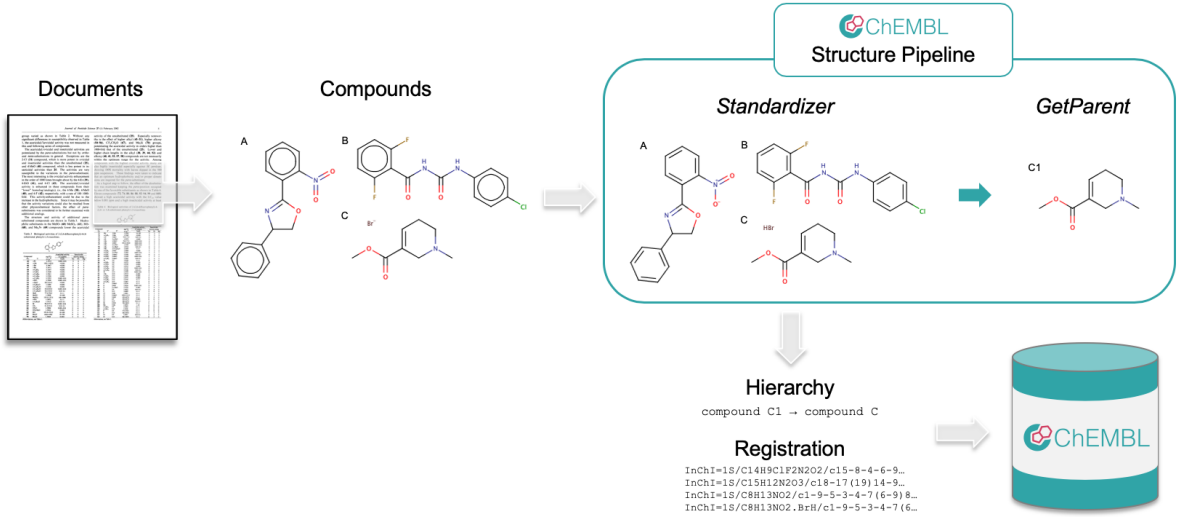


ChEMBL Curation Pipeline



Motivation

- ChEMBL contains 2.5 million compound records on ~2 million unique structures
- Compounds from different sources are not typically standardised according to consistent rules
- In order to maintain the quality of the database and to easily compare data on the same compound from different sources it is necessary for the chemical structures in the database to be appropriately standardized
- A chemical curation pipeline has been developed using RDKit to validate and standardise compounds for ChEMBL, as well as to establish how they are related



ChEMBL Structure Pipeline

- ChEMBL Structure Pipeline is comprised of three processes:
  - *Checker*: identifies and validates structures and identifies problems before structures are loaded into the database
  - *Standardizer*: processes (standardises) chemical structures according to a set of predefined rules
  - *GetParent*: generates parent structures based a set of rules and defined list of salts and solvents
- *Standardizer and GetParent* have been rewritten and adapted from rules originally implemented using a commercial software toolkit
- *Checker* was developed more recently in an attempted to identify problem structures

ChEMBL Structure Pipeline – *Checker*

- *Checker* component validates structures prior to loading compounds into ChEMBL
- If an error or problem is detected in the structure, a score is assigned
- More serious problems have higher scores

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

Score 7

- A fatal error and the data is not loaded into ChEMBL

Score 6

- Compounds are loaded into the database but without a molfile, as it is considered to have a significant issue and it is preferred to fix the problem before loading the molfile

Score <=5

- The structure is loaded but these are prioritized for manual curation

Note: ~75% of ChEMBL no penalty scores

Checker: penalty scores on ChEMBL26

Penalty score	Penalty explanation	No of compounds
6	InChI: Accepted unusual valence(s)	10
	Molecule has a radical that is not found in the known list	9
	Molecule has six (or more) atoms with exactly the same coordinates	50
5	InChI_RDKit/Mol stereo mismatch	810
	Mol/Inchi/RDKit stereo mismatch	6
	RDKit_Mol/InChI stereo mismatch	771
	Molecule has a crossed bond in a ring	632
	Molecule has two (or more) atoms with exactly the same coordinates	259

Compounds where the exclude flag is set are excluded from this analysis

ChEMBL Structure Pipeline – *Standardizer*

- *Standardizer* component processes and standardises chemical structures according to a set of rules
- Rules are based largely on the FDA/IUPAC guidelines
- An ‘exclude’ flag is set and the chemical structure is not standardised. It will appear as a ChEMBL_ID with bioactivity data but no structure in the release version of the database for:
  - Organometallics where V2000 molfile format used by ChEMBL is unable to accurately represent coordination bonds (e.g.[Sc], [Ti], [V], [Cr], [Mn], [Fe], [Co], [Ni], [Cu], [Ga], [Y], [Zr], [Nb], [Mo], [Tc], [Ru], [Rh], [Pd], [Cd], [In], [Sn], [La], [Hf], [Ta], [W], [Re], [Os], [Ir], [Pt], [Au], [Hg], [Tl], [Pb], [Bi], [Po], [Ac], [Ce], [Pr], [Nd], [Pm], [Sm], [Eu], [Gd], [Tb], [Dy], [Ho], [Er], [Tm], [Yb], [Lu], [Th], [Pa], [U], [Np], [Pu], [Am], [Cm], [Bk], [Cf], [Es], [Fm], [Md], [No], [Lr], [Ge], [Sb])
  - Structures that contain > 7 boron atoms

Examples of *Standardizer* rules:

Standardise unknown stereochemistry

- change “wiggly” bonds to sp3 carbons denoting unknown stereo to show no stereo
- set either or unknown cis/transbonds to crossed bonds instead of “wiggly” bonds

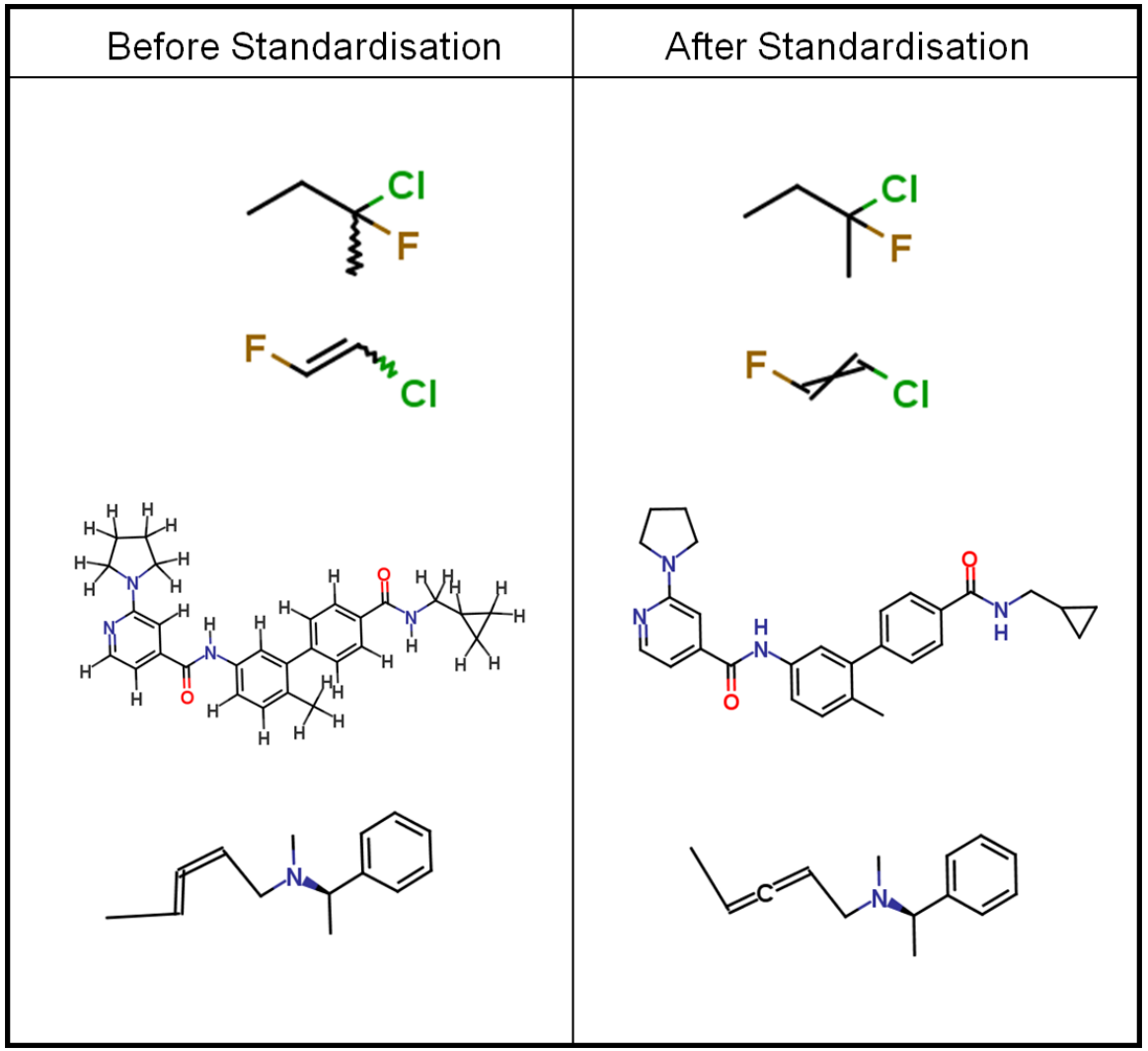
Clear S group data from molfile

Generate 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

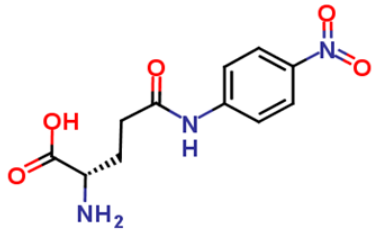
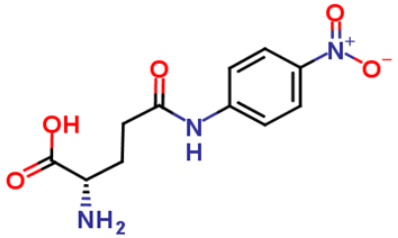
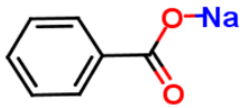
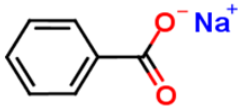
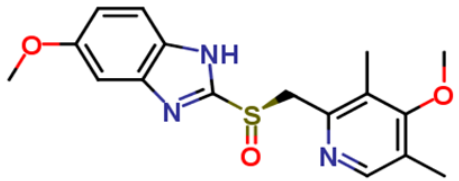
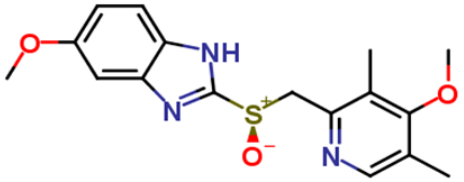
Normalise (straighten) triple bonds and allenes



Examples of *Standardizer* rules:

Normalise structure

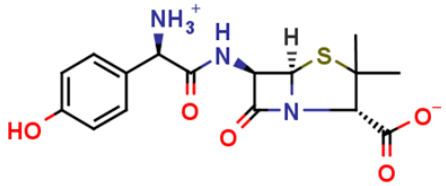
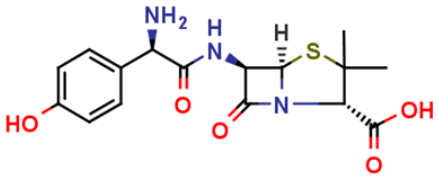
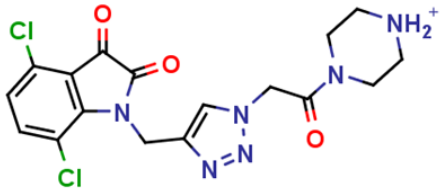
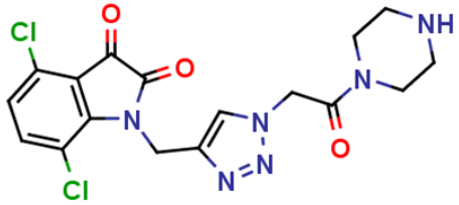
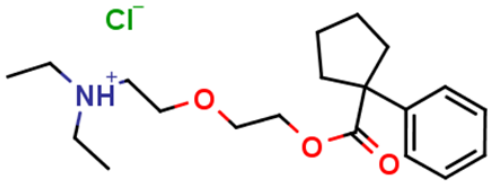
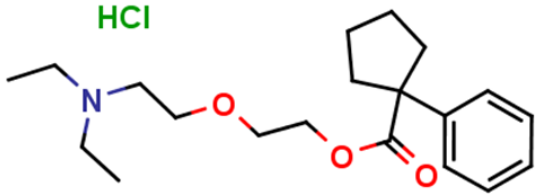
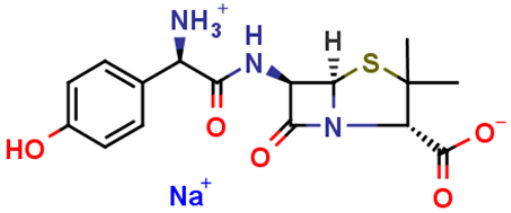
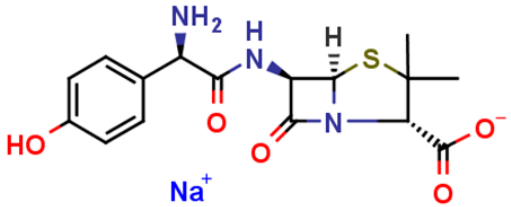
- fix hypervalent nitro groups
- convert covalently drawn alkaline metals connected to O or N to ionic forms (e.g. NaO to Na+ O-)
- fix incorrect amide tautomers, e.g. N=COH to HNC(=O)
- standardise sulphoxides to charge-separated form
- standardise diazonium N to N+
- ensure quaternary N is charged
- ensure trivalent O is charged
- ensure trivalent S is charged
- ensure halogen not bonded to a neighbouring atom is charged

Before Standardisation	After Standardisation
	
	
	

Examples of *Standardizer* rules:

Ensure molecule is neutralized, if possible, by

- adding or removing Hs
- moving Hs from one atom to another (including between components)

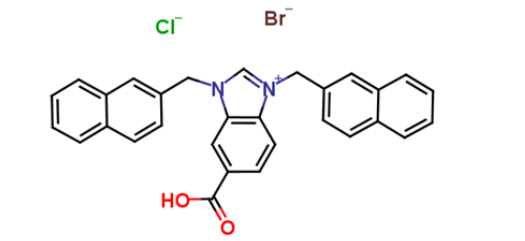
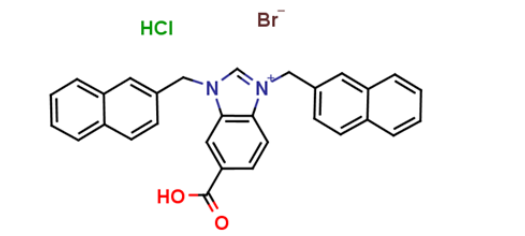
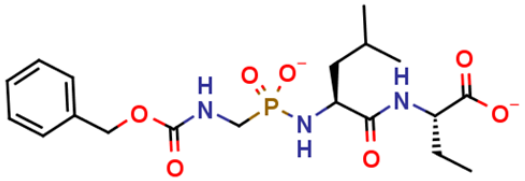
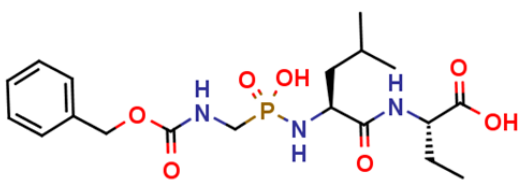
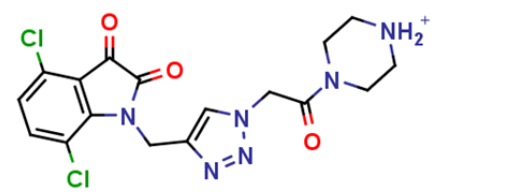
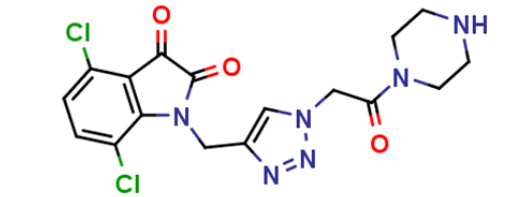
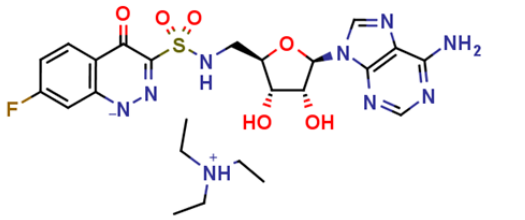
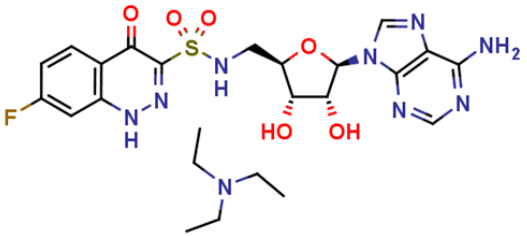
Before Standardisation	After Standardisation
	
	
	
	

Effect of Standardisation on ChEMBL Registration

- ChEMBL uses the Standard InChI/InChIKey to determine a unique chemical structure
- Most standardisations don't result in Standard InChI/InChIKey changes

InChIKey layer change on representative ChEMBL set	No of Compounds
Connectivity	13
Connectivity and Protonation	1
Protonation	297
Stereochemistry	0
Stereochemistry and Protonation	0
Total no of changed InChIKeys after standardisation	311
Total no of compounds	147,008
% changes InChIKeys	.021

Examples of standardisation resulting in InChIKey changes

Before Standardisation	After Standardisation
	
	
	
	

In ChEMBL these are therefore registered with new ChEMBL_IDs

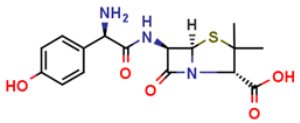
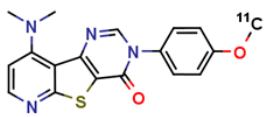
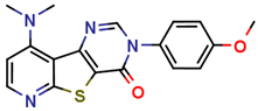
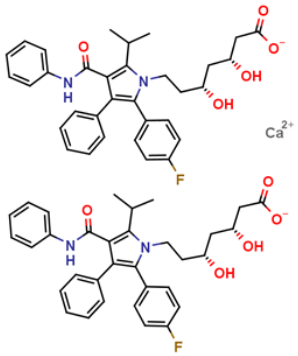
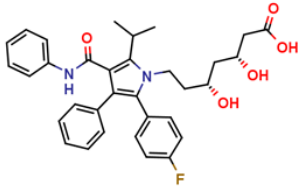
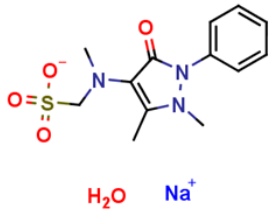
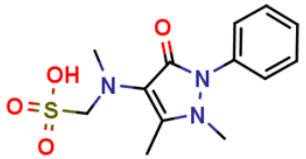
ChEMBL Structure Pipeline – *GetParent*

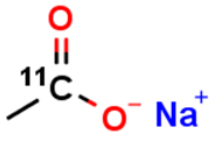
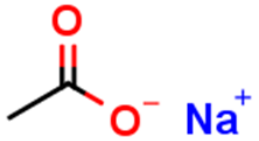
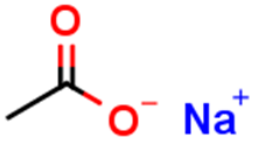
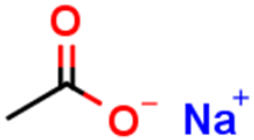
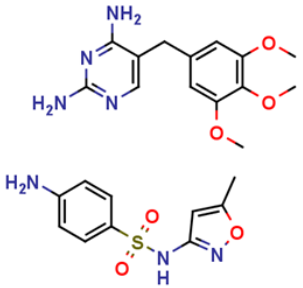
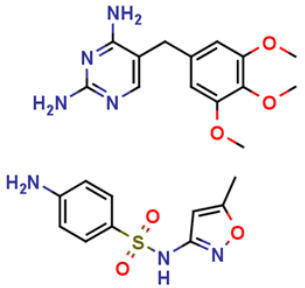
- *GetParent* component generates parent structures based a set of rules and defined list of salts and solvents
- List of salts and solvents is based mainly on the USAN Council's list of pharmacological salts. Additional entities have been added where a significant number of examples are present in ChEMBL datasets
  - currently list contains 162 salts and 9 solvents
- *GetParent* component removes salts regardless of
  - the charge status (*e.g.* acetic acid or acetate)
  - whether or not the stereochemistry is depicted (*e.g.* tartaric acid)
  - cis/trans isomers (*e.g.* maleic and fumaric acid)

ChEMBL Structure Pipeline – *GetParent*

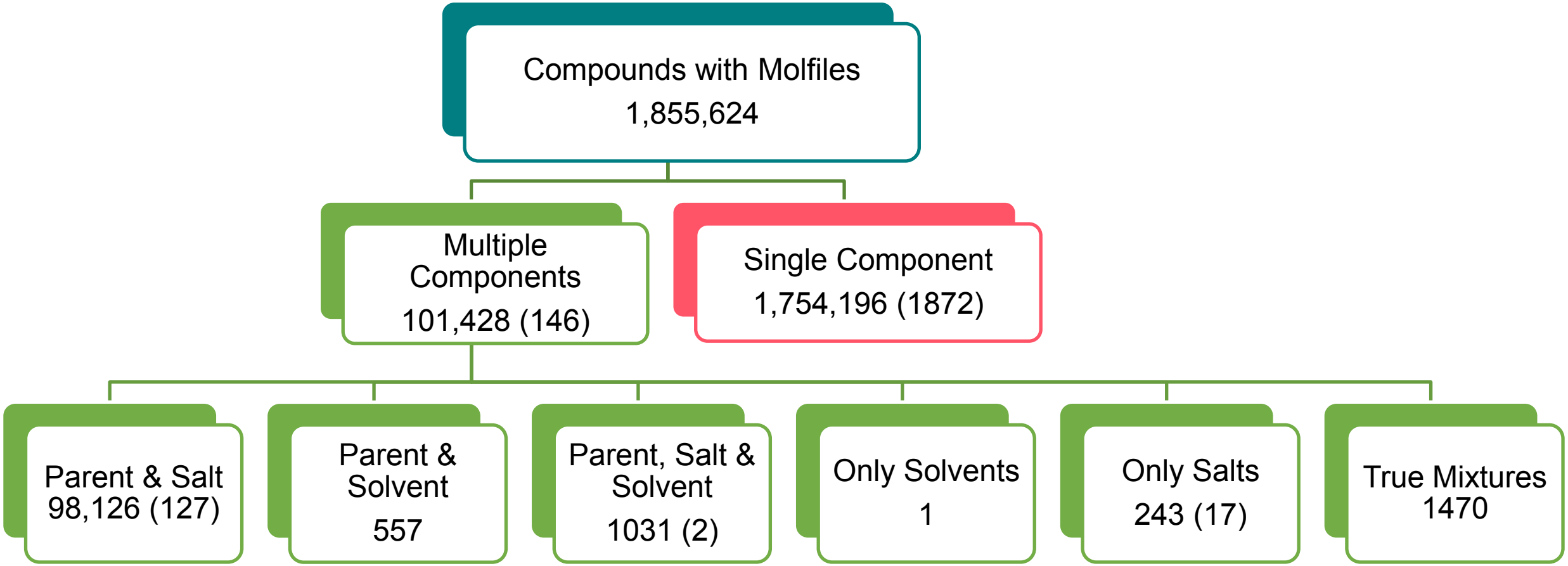
- *GetParent* component is applied to molecules where the molfile contains > 1 component and/or molecules containing atoms with specific isotopes
- All information about specific isotopes is removed
- Solvents and salts are removed that match any of the components in the salts and solvents list
- Having removed all salts, the molecule is neutralized and a new molfile created as the ‘parent’ molecule
- For compounds containing > 1 components (i.e. genuine mixtures), a ‘parent’ molecule is registered as the identical mixture
- For cases where both components are in the salt list, *GetParent* does not remove any component and the parent remains the same as the salt
- For ‘excluded’ compounds, only isotopes and solvents are removed

Applying the *GetParent* module to representative compounds

Example	Child	Parent
Parent & Salt		
Isotope		
2:1 Salt		
Parent, salt & solvent		

Example	Child	Parent
Salt components with isotope		
Salt components		
True Mixture		

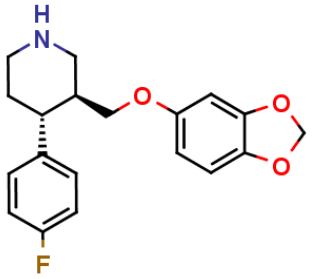
Multiple component compounds in ChEMBL26

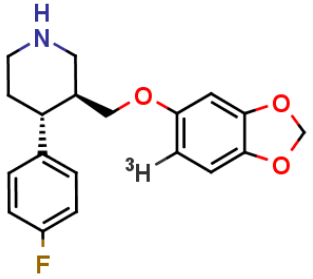


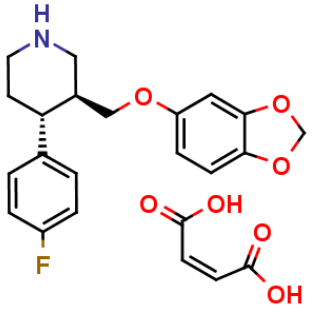
* () no. of isotopes

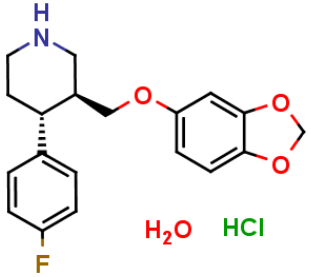
Examples of alternate forms of Paroxetine in ChEMBL26

- Bioactivity data is registered against the form it was measured on.
- Aggregation by parent is undertaken to make it easier to identify all the bioactivity data for salts and isotopes of a common parent

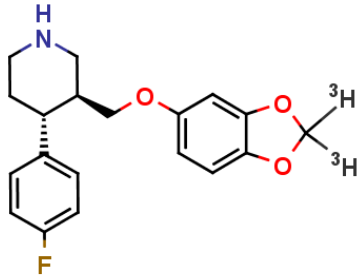


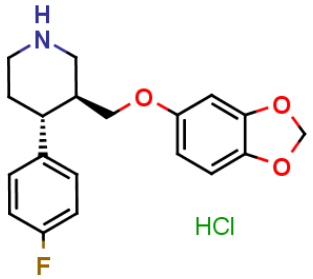


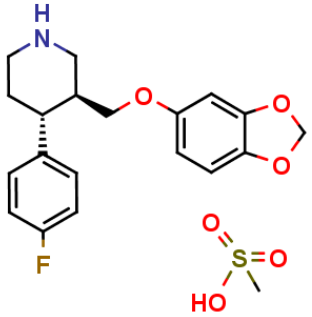




ChEMBL_ID	CHEMBL4901063	CHEMBL16286501	CHEMBL144949076	CHEMBL12569127
Bioactivities				







ChEMBL_ID	CHEMBL31333001	CHEMBL1708100	CHEMBL1200609
Bioactivities			‘drug' only

Code Availability

- ChEMBL Structure Pipeline has been developed using the RDKit toolkit (version 2019.09.2.0)
- It is open source and publicly available (currently as version 1.0.0)
  - GitHub: https://github.com/chembl/ChEMBL_Structure_Pipeline
  - Python Conda Package: https://anaconda.org/chembl/chembl_structure_pipeline
  - ChEMBL Beaker Web Services: https://www.ebi.ac.uk/chembl/api/utils/docs (check, standardize and getParent endpoints)
- Speed: on 100,000 compounds (*Checker* 2 mins, *Standardizer* 4 mins, *GetParent* 11 mins)
- Any new features will be added to the GitHub repository and comments and suggestions are welcome

More information

- Article recently published in Journal of Cheminformatics

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 Bento[1], Anne Hersey[1], Eloy Félix[1], Greg Landrum[2], Anna Gaulton[1], Francis Atkinson[1,3], Louisa J. Bellis[1,4], Marleen De Veij[1] and Andrew R. Leach[1*]

Acknowledgements

ChEMBL

- Anne Hersey
- Patrícia Bento
- Eloy Félix
- Anna Gaulton
- Francis Atkinson
- Louisa Bellis
- Marleen de Veij
- Andrew Leach

RDKit Support and Chemical Data Science - T5 Informatics

- Greg Landrum

Funding:

wellcome trust
Strategic Award

EMBL-EBI