Intro to Chemical Data Science

Greg Landrum, ETH Zurich

SCS Spring School on Digital Chemistry

April 2023

What is “Chemical Data Science”?

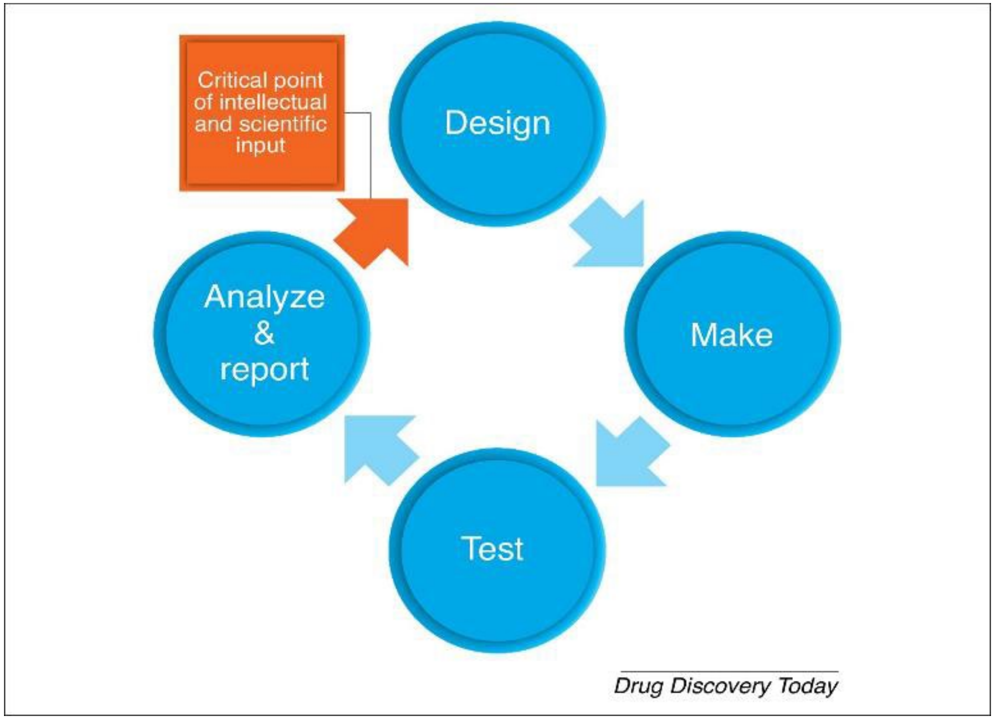
First: what is Data Science?

> Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured, and unstructured data.
>
> https://en.wikipedia.org/wiki/Data_science

Chemical Data science is that, but with the added complication of needing to work with molecules as part of the data

How does this fit into drug discovery?

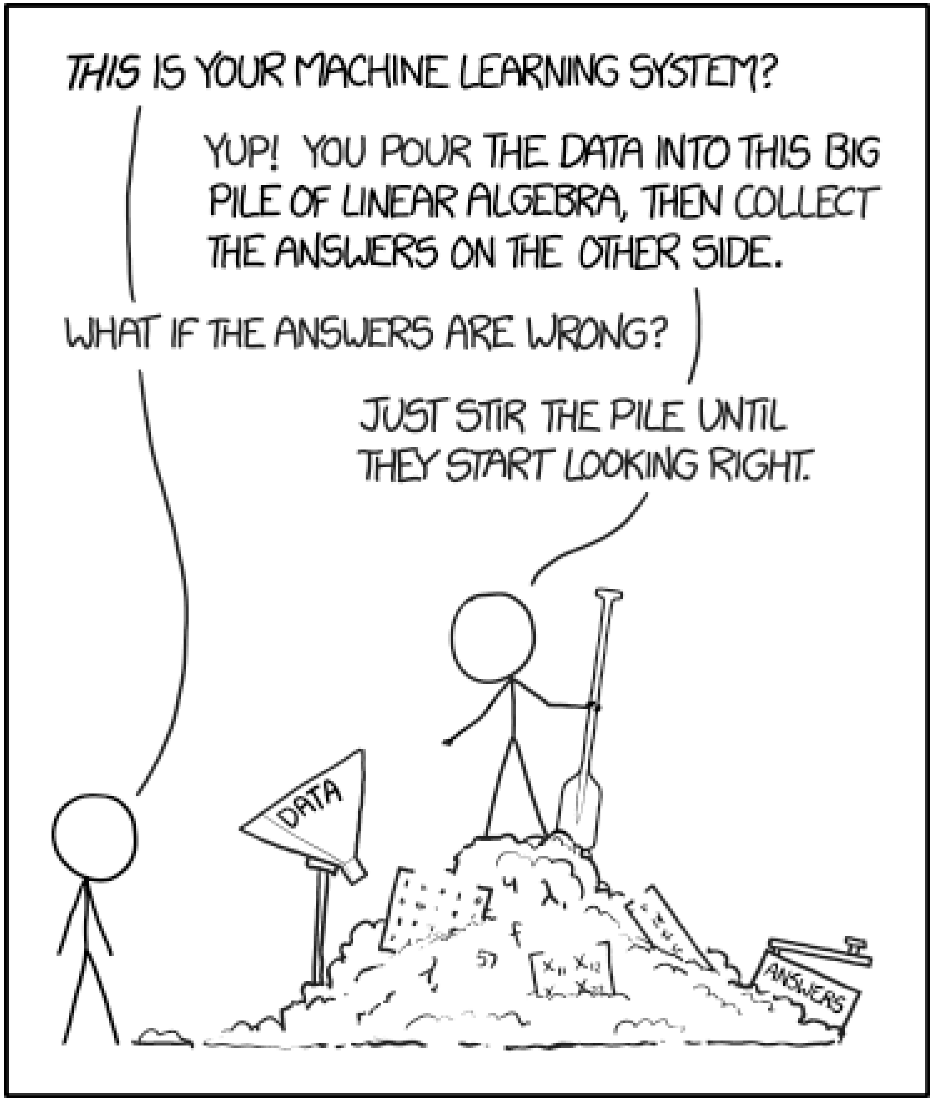
- Drug discovery teams typically follow an iterative approach as they try to discover and optimize drug candidates
- A useful way of thinking about these iterations is the “Design-Make-Test-Analyze” cycle.
- Computer-aided drug design (CADD) is primarily focused on the Design and Analyze stages
- Chemical Data Science runs throughout



https://doi.org/10.1016/j.drudis.2018.09.016

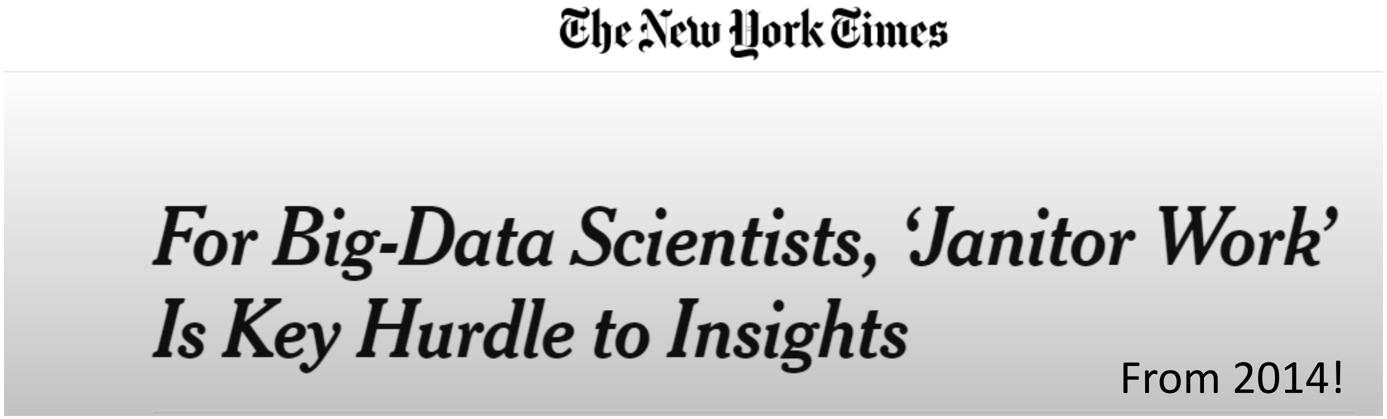
More on data science

More on data science


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://xkcd.com/1838/

More on data science



The New York Times

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

From 2014!

This hasn't changed… there's still a lot of plumbing that needs to be done.

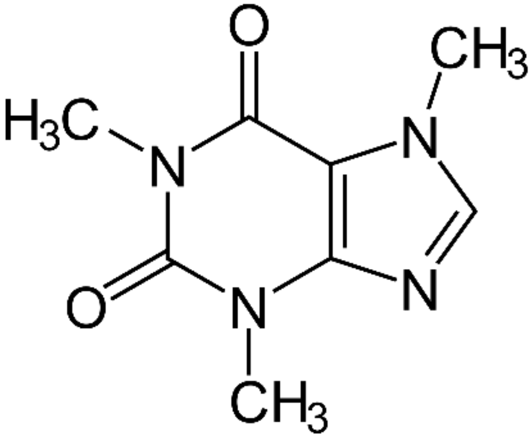
So we're going to talk about the plumbing today

Overview

- Representing molecules in the computer
- Substructure search
- Molecular fingerprints
- Molecular descriptors
- Standardization
- Public resources

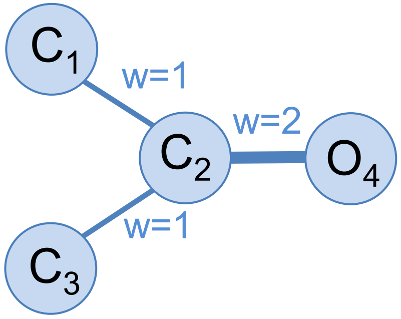
Representation

- How to represent molecules in computers?



Matrix Representation

- Molecule can be drawn as undirected, weighted graph
  - Node = atom:
    - Contains the atom number and information about the element, number of hydrogens (if not explicitly in the graph), isotope, charge, stereochemistry, aromaticity
  - Edge = bond:
    - Contains information about bond type (= weight)
- Example: Acetone



Graph

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 0 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 \end{array}$$

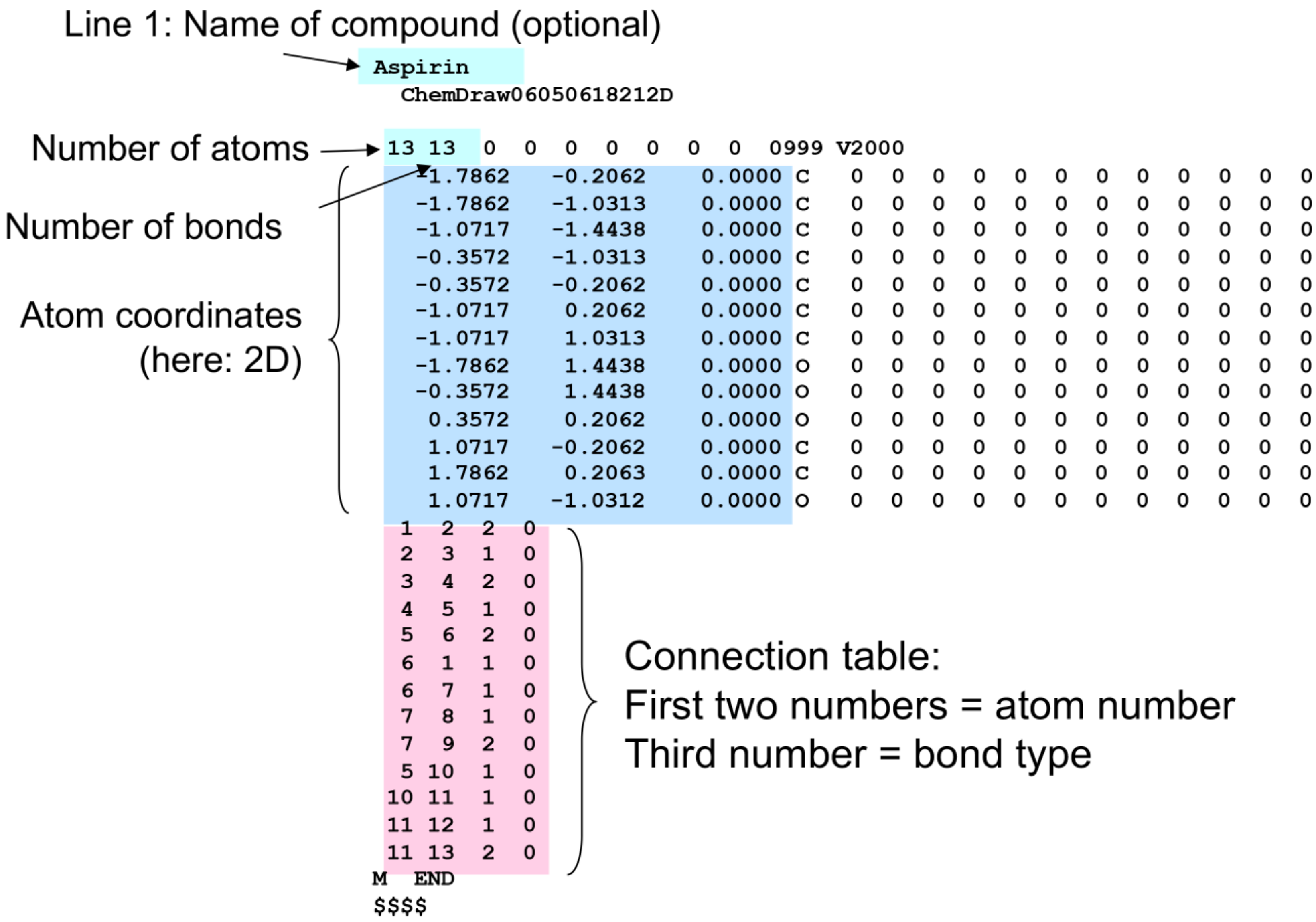
Adjacency matrix

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 0 & 1 & 0 & 0 \\ 4 & 0 & 2 & 0 & 0 \end{array}$$

Connection table

- Problem:
  - Not efficient for storing, retrieving and comparing molecules

Structure Data Format (SDF)



Line 1: Name of compound (optional)

Number of atoms

Number of bonds

Atom coordinates (here: 2D)

Connection table:
First two numbers = atom number
Third number = bond type

```
Aspirin
  ChemDraw06050618212D

 13 13  0  0  0  0  0  0  0  0999 V2000
   -1.7862   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.7862   -1.0313    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0717   -1.4438    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3572   -1.0313    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3572   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0717    0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0717    1.0313    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.7862    1.4438    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3572    1.4438    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.3572    0.2062    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.0717   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.7862    0.2063    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0717   -1.0312    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0
  2  3  1  0
  3  4  2  0
  4  5  1  0
  5  6  2  0
  6  1  1  0
  6  7  1  0
  7  8  1  0
  7  9  2  0
  5 10  1  0
 10 11  1  0
 11 12  1  0
 11 13  2  0
M  END
$$$$
```

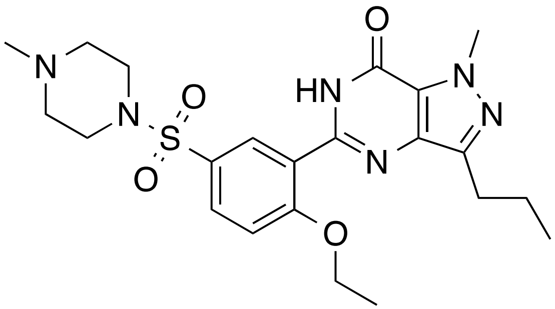
Graphic from G. Schneider

Representation

- 1D (linear) representation of a molecule
  - Compact (less storage)
  - Easy interpretation and parseability (searching) by computers
- Requirements:
  - Includes stereoinformation and aromaticity
  - Reversible (2D → 1D as well as 1D → 2D)
  - Optional: uniquely defined → for use as identifiers in databases
- Examples:
  - IUPAC name
  - Wiswesser Line Notation (WLN)
  - Simplified Molecular Input Line Entry System (SMILES)
  - SYBYL Line Notation (SLN)
  - [International Chemical Identifier (InChi)]

SMILES

- Background:
  - Idea:
    - Encode chemical structure using characters based on a set of rules
  - Introduced in 1988 (D. Weininger, *J. Chem. Inf. Comput. Sci*. **28**, 31–36, 1988)
  - Depending on which atom is taken as root, different valid SMILES are generated
    - Canonicalization needed to generate unique SMILES
  - Example:



Sildenafil

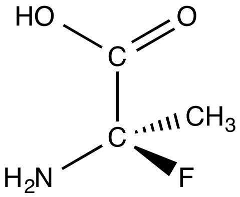
➡ CCCc1nn(C)c2c1nc(-c1cc(S(=O)(=O)N3CCN(C)CC3)ccc1OCC)[nH]c2=O

SMILES

- Rules:
  - Hydrogens as well as single and aromatic bonds are usually omitted, but can be specified explicitly if desired
  - Atoms:
    - General: Atomic symbol in square brackets
    - "Organic" subset (= B, C, N, O, P, S, F, Cl, Br, I) can be written without brackets if the number of attached Hs is "normal"

      e.g. $H_3C\text{-}CH_3$ → CC
    - Attached hydrogens and formal charges always specified inside brackets

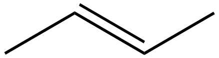
      e.g. [OH-], [NH4+]
    - Atoms in aromatic rings are specified by lower case letters (i.e. c, n)
    - Stereocentres are specified with @ (anti-clockwise writing of neighbors) and @@ (clockwise writing of neighbors) inside brackets



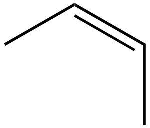
→ N[C@](C)(F)C(=O)O or N[C@@](F)(C)C(=O)O

SMILES

- **Bonds:**
  - Single bond: “-”, double bond: “=“, triple bond: “#”, aromatic bond: “:”
  - Trans/cis double bonds: use of “/” and “\”, e.g.

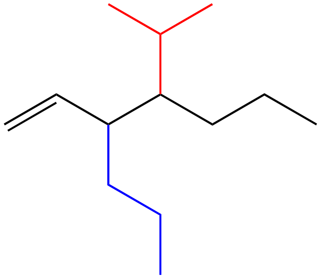


→ C/C=C/C



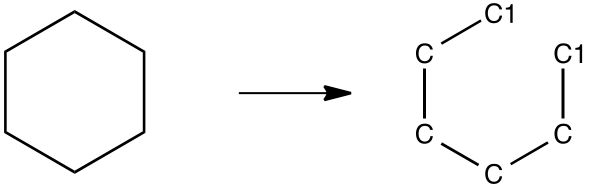
→ C/C=C\C

- **Branches:**
  - Specified by parentheses
  - Can be nested or stacked



→ C=CC(CCC)C(C(C)C)CCC

- **Cyclic structures:**
  - Represented by breaking one bond (to get spanning tree) and numbering the ring-closure atoms
  - Ring-closure digits can be reused

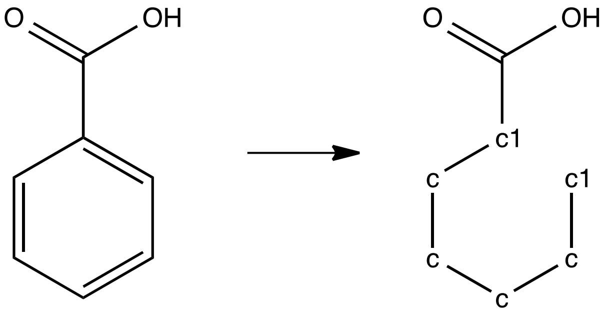


→ C1CCCCC1

SMILES

Aromaticity:

- Aromaticity is a concept → different definitions/algorithms (discussed later)
- Aromatic bonds are usually omitted
- Atoms in an aromatic ring are specified by lower case letters



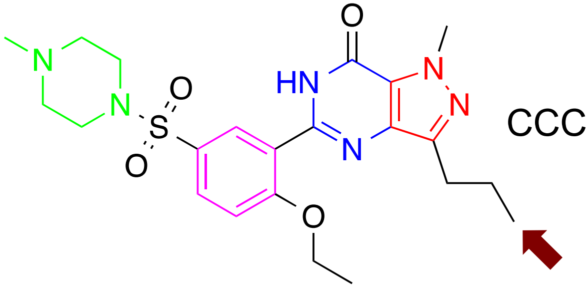
→ c1ccccc1C(=O)O

Four basic rules for organic compounds:

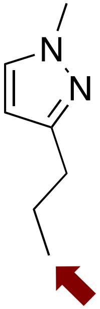
- Atoms are represented by atomic symbols
- Double and triple bonds are represented by = and #
- Branching is indicated by parentheses
- Ring closures are indicated by matching digits appended to the symbol

SMILES

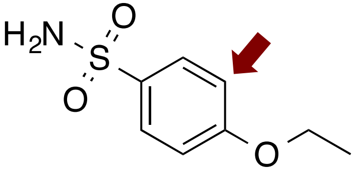
- Back to sildenafil:



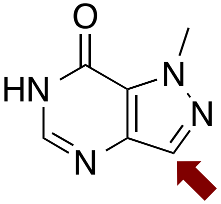
CCCc1nn(C)c2c1nc(-c1cc(S(=O)(=O)N3CCN(C)CC3)ccc1OCC)[nH]c2=O



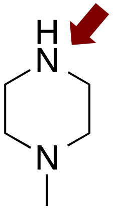
CCCc1nn(C)cc1



c1cc(S(=O)(=O)N)ccc1OCC



c1nn(C)c2c1nc[nH]c2=O



N3CCN(C)CC3

Canonicalization

- Idea:
  - Generate a unique (and reproducible) numbering of the atoms in a molecule
  - Needed to generate a unique SMILES which can be used as identifier

- Identical molecules should have the same canonical SMILES.

- Details are dependent on the details of the algorithm: you can’t compare canonical SMILES from two different toolkits (or possibly from different versions of the same toolkit)

InChI

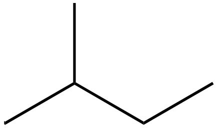
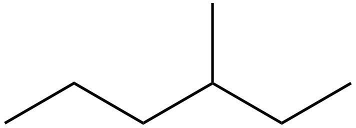
- InChI = international chemical identifier
  - Developed by NIST and IUPAC in 2000-2005
- Structure: Series of layers (specified by a prefix) separated by a “/”
  - Start: InChI=1S
    - 1 → version; S → standard
  - Layers:
    - Main layers:
      - Chemical formula (no prefix)
      - Atom connections (prefix “c”) → hydrogens excluded
      - Hydrogen atoms (prefix “h”)
    - Charge layers:
      - Protons (prefix “p”)
      - Charges (“q”)
    - Stereochemical layers:
      - Bonds (prefix “b”) → double and triple bonds
      - Tetrahedral stereoinfo (prefix “t”, “m”)
      - Type of stereoinfo (prefix “s”)
    - Isotopic layer (prefix “i”, “h”)

InChI

- Examples:
  - CH_3CH_2OH (ethanol) → InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3
  - CH_3CHO (acetaldehyde) → InChI=1S/C2H4O/c1-2-3/h2H,1H3
- Algorithm:
  - Normalization → remove redundant information, pick canonical tautomer, etc.
  - Canonicalization → unique numbering
  - Serialization → generating the string of characters
- Standard InChIKey:
  - Condensed digital representation → not human-readable
  - Hashed to 27 characters:
    - 14 characters from a hash of the connectivity information + Hyphen
    - + 9 characters from a hash of the remaining layers
    - + Version + Hyphen
    - + Checksum (hash sum) character
  - No reversal possible → InChI cannot be restored from its InChIKey
  - Example:
    - Ethanol → InChIKey = LFQSCWFLJHTTHZ-UHFFFAOYSA-N

Substructure Search

- Substructure search:
  - Question: Does this substructure exist in any molecule of my database?

    Does this  occur in  ?

  - Input: query pattern of atoms and bonds (e.g. SMARTS)

  - Mathematical problem is very complex (subgraph isomorphism), but molecules have special properties that allow us to use heuristics to solve the problem
    - Elements of the “organic set” form maximum 4 bonds
    - Hydrogens can be omitted
    - Some elements / bond types are more rare than others (can be searched for first to quit search fast)

SMARTS

- SMILES arbitrary target specification (SMARTS)
  - Extension of SMILES to describe molecular patterns (substructures)
    - Wildcards for atoms and bonds, logical operators
  - Every valid SMILES is also a valid SMARTS, but they may mean something different
- Additional labels:
  - Atoms:
    - Specified by either element symbol or number: e.g. [#6] → any carbon
    - “*” : wild card
    - “A” : any aliphatic atom
    - “a” : any aromatic atom
    - “D” followed by a number : degree (number of explicit connections)
    - “R” followed by a number *n* : in *n* smallest rings
    - “r” followed by a number *n* : in a smallest ring of size *n*
    - “H” followed by a number : number of adjacent hydrogens
    - H has now two meanings:
      e.g. [H] → hydrogen atom, e.g. [*2H] → any atom with two hydrogens
    - Multiple possible matches are separated by a comma: e.g. [C,N] → either aliphatic C or N

SMARTS

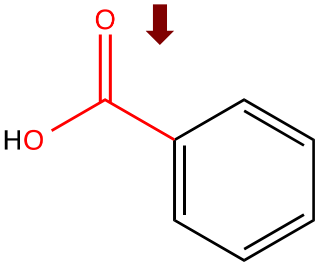
SMARTS viewer:

http://smartsview.zbh.uni-hamburg.de

- Bonds:
  - “~” : any bond
  - “@” : any ring bond
- Logical operators: combinations of atom and bond specifications
  - “!” : NOT, e.g. [!C] → not aliphatic carbon
  - “&” : AND (high priority)
  - “,” : OR
  - “;” : AND (low priority)
  - Operator priority: “!” > “&” > “,” > “;”
- Aromaticity:
  - Note: a double bond is not matched to an aromatic bond!
- Examples:

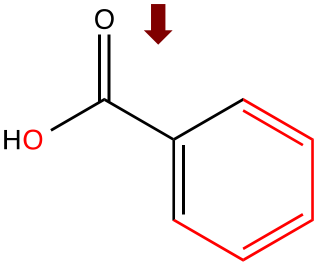
***C(=O)O**

carboxylic acid connected to anything



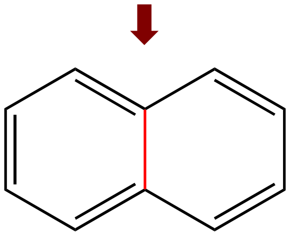
[c,O;H1]

aromatic C or aliphatic O AND connected to a hydrogen



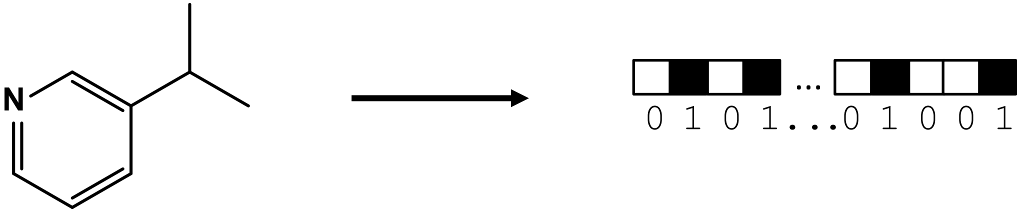
[R2]

any atom in two smallest rings



Molecular Fingerprints

- Abstract representation of 2D topological structure:
  - Idea: Convert chemical structure into a 1D bit string
    - Each bit represents a specific fragment
      - 1 : Structure contains the fragment
      - 0 : Structure does not contain the fragment
    - Typically hundreds or thousands of fragments considered
  - Advantages:
    - Very compact (less storage) and rapid comparison possible
  - Disadvantages:
    - Different structures can have the same fingerprint!



Fingerprints

- 4 types of fingerprints based on 2D structure:
  - Dictionary-based
  - Path-based
  - Circular fingerprints
  - 2D pharmacophores
- Provide a description of molecules in terms of bit- or count-vectors
- Commonly used for molecular similarity and machine learning

Fingerprints

- Dictionary-based fingerprints:
  - ➢ Predefined set of substructures (keys)
    - o Bit position directly connected to a certain pattern
  - ➢ Example: Molecular ACCess System (MACCS) from MDL → 166 structural keys (as SMARTS)

    Bit 13: “[#8]~[#7](~[#6])~[#6]”
  - ➢ Example: PubChem fingerprint [1] → 881 structural keys

Bit 29	>= 2 Si
Bit 123	>= 2 saturated or aromatic carbon-only ring size 3

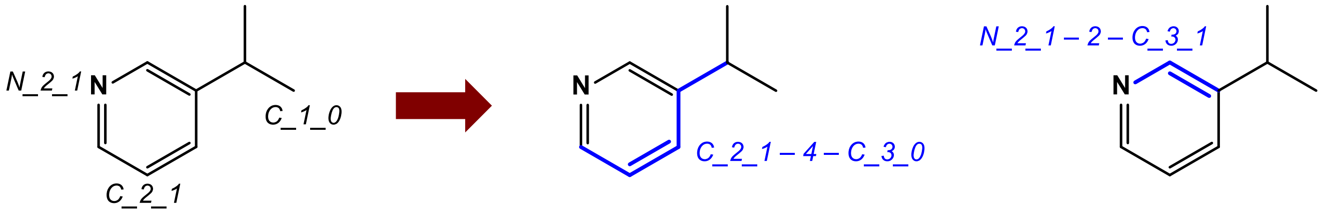
[1] https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf

Fingerprints

- Hashing:
  - In path-based, circular and pharmacophore fingerprints patterns are normally hashed to bit positions
  - No direct connection bit position → pattern (except if stored during hashing)
  - Collisions can occur (different patterns hashed to the same bit position)

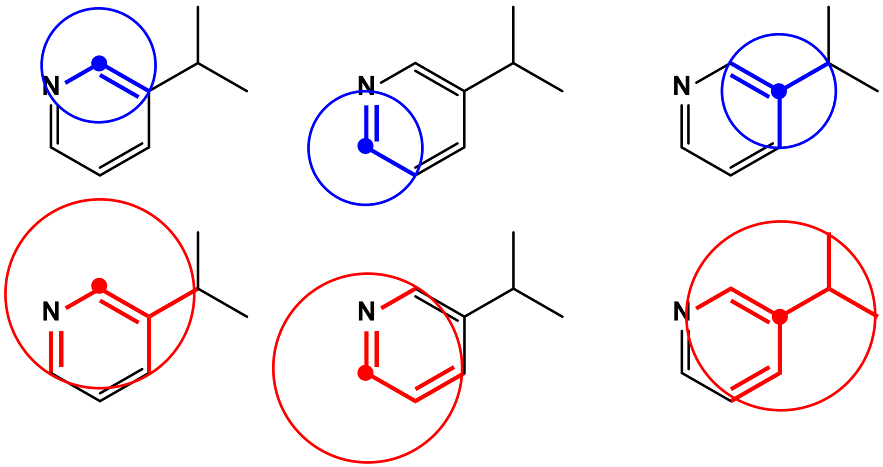
Fingerprints

- Path-based fingerprints:
  - Search for all occurrences of a set of generic substructure patterns in the molecule
  - Example: RDKit fingerprint
    - Patterns: subgraphs with 1-7 bonds
    - Hashing (to get bit positions):
      - The specific atoms and bonds in a given occurrence
  - Example: Atom-pairs fingerprint
    - Atom type: (element)_(# heavy neighbors)_(# π-electrons)
    - Atom pairs: (atom type)_(topological distance)_(atom type)



Fingerprints

- Circular fingerprints (Morgan fingerprints):
  - Fragments = circular environments around the atoms with different radii
  - Circular environments hashed to bit positions
  - Example: Extended connectivity fingerprints (ECFP)
    - Atom type: (element, #heavy neighbors, # Hs, isotope, in-ring flag)
    - Size of max. radius is user-defined (0, 1, 2, ...)



radius = 1 (path length = 2)

radius = 2 (path length = 4)

- Example: Feature connectivity fingerprints (FCFP)
  - Atom type: pharmacophoric features

D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **50**, 742-754, 2010.

Fingerprints

- 2D pharmacophores:
  - Definition [1]: “Ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response”
  - Pharmacophoric features:
    - H-bond donors/acceptors, hydrophobic, pos./neg. charged
    - Aromatic and/or hydrophobic groups: usually a group of atoms mapped to a virtual point
  - 2D pharmacophore fingerprints: Combination of pharmacophoric features and the topological distance between them
    - In the RDKit [2]: Distances are binned and each combination is mapped to a certain bit

Example: Signature from:
2 Patterns
2 - 3 point pharmacophores
2 distance bins (1,3),(3,8)

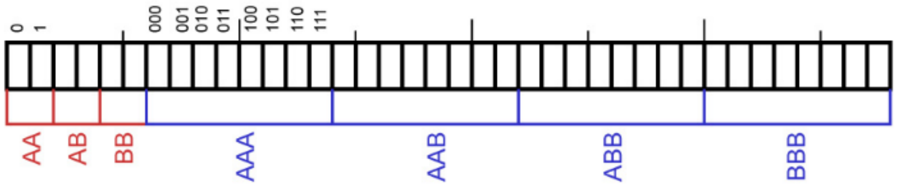
Total Signature Size: 38 bits

2 point pharmacophores:
Combos: AA, AB, BB
2 bits/pharmacophore (1 distance with 2 bins)

Total: 6 bits

3 point pharmacophores:
Combos: AAA, AAB, ABB, BBB
8 bits/pharmacophore (3 distances with 2 bins)

Total: 32 bits



[1] IUPAC recommendations: Pure & Appl. Chem., 70, 1129 (1998).
[2] https://rdkit.org/docs/RDKit_Book.html#representation-of-pharmacophore-fingerprints

Fingerprints

- Technical problem:
  - A huge number of fragments is possible, but a single molecule contains usually only a few
  - Fingerprints are sparsely populated → most of the space is wasted on "zero bits"
- Two solutions:
  - Sparse integer vectors: dictionary where keys = bit position
    - Only on-bits are stored
  - Folding (second hashing):
    - Bit positions are hashed to a user-defined size (e.g. 2048 bits)
    - On- and off-bits are stored
    - Disadvantage: Additional source for collisions
    - Chosen size has to be large enough to avoid too many collisions

Fingerprints

- Similarity metrics: Comparison of fingerprints
  - Resulting values: 1 = all bits in common, 0 = no bits in common
  - Example of a popular similarity metric:
    - Tanimoto coefficient (or Jaccard coefficient):

$$sim_{Tanimoto} = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}}$$

N_X : number of *on*-bits in fingerprint *X* (*X* = *A*,*B*)
$N_{A\&B}$: number of common *on*-bits in *A* and *B*

  - Efficient implementation: Binary logic (boolean algebra)

AND

Bit1	Bit2	Bit1&Bit2
0	0	0
0	1	0
1	0	0
1	1	1

OR

Bit1	Bit2	Bit1|Bit2
0	0	0
0	1	1
1	0	1
1	1	1

XOR

Bit1	Bit2	Bit1^Bit2
0	0	0
0	1	1
1	0	1
1	1	0

Bit operator (C++)	Meaning
&	bitwise AND
|	bitwise OR
^	bitwise excl. OR (XOR)
<<	left shift
>>	right shift
~	NOT

NOT: e.g. ~(1001101) → 0110010

Left shift: e.g. (0001100)<<1 → 0011000

Right shift: e.g. (0001100)>>2 → 0000011

Molecular Property Descriptors

- Interpretable properties:
  - Can be used to try and find simple rules for biological endpoints

- Examples:
  - Molecular weight
  - Calculated octanol/water partition coefficient (ClogP)
  - Topological polar surface area (TPSA)
  - Number of hydrogen-bond acceptors and donors
  - Lipinski's rule-of-five (Ro5)

Molecular Property Descriptors

- Calculated octanol/water partition coefficient (ClogP):
  - Definition:

$$\log P_{octanol/water} = \log\left(\frac{[\text{solute}]_{octanol}}{[\text{solute}]_{water}}\right)$$

  - Used as a metric for lipophilicity
  - One of the most (over-)used descriptors
  - Calculation: Atom/group contribution models
    - Simple, based on 2D topological structure alone
    - Wildman-Crippen ClogP model
      (*J. Chem. Inf. Comput. Sci*., 39, 868, 1999)
      → 68 atom contributions derived from
      training set of 9920 organic molecules

$$\log P_{octanol/water} = \sum_i n_i a_i$$

    - AlogP model
      (*J. Phys. Chem. A*, 102, 3762, 1998)

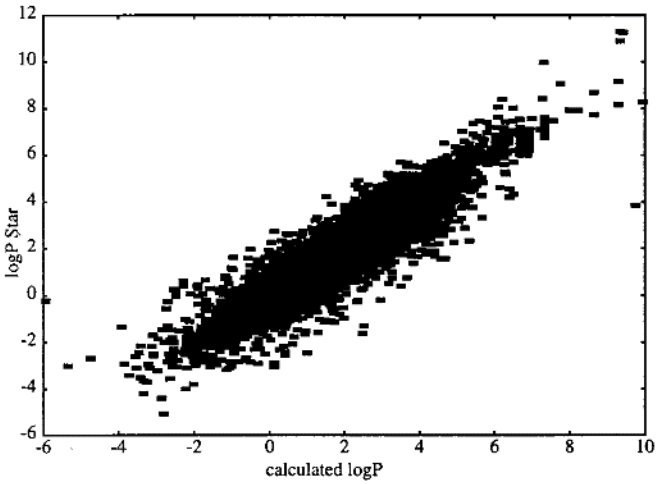


Figure 1. Correlation plot for the fit of 9920 log *P* Star values: $r^2 = 0.918$; $\sigma = 0.677$.

Molecular Property Descriptors

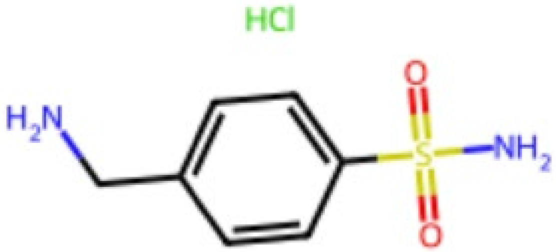
- Lipinski's rules-of-five (Ro5):
  - Original paper: (*Adv. Drug Deliv. Rev*., 23, 3, 1997)
  - Heuristic rules for druglikeness based on analysis of historical data on oral bioavailability of drugs (at Pfizer)
  - "Poor absorption or permeation is more likely when"
    - there are more than 5 hydrogen-bond donors
    - there are more than 10 hydrogen-bond acceptors
    - the molecular weight is greater than 500 g/mol
    - the ClogP is greater than 5
    - (the number of rotatable bonds is greater than 10)
  - Still used as a crude filter
  - Important to keep the limitations of such simple rules in mind!

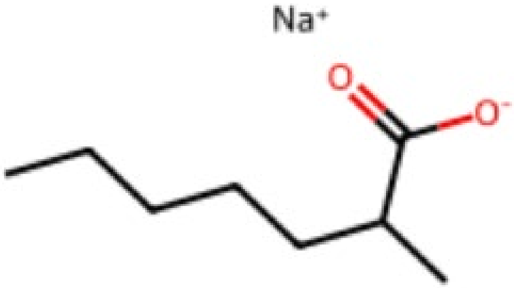
Molecule standardization

- An important (and often forgotten) step when working with chemical data
- Includes things like:
  - Salt stripping
  - Neutralization
  - Charging to assay/physiological pH
  - Tautomer enumeration

- Which steps need to be done depend on the problem at hand.

Salt stripping





- Often compounds are synthesized/tested as salts.
- The salt/counterion is not normally involved in interaction with the protein, so we often remove them

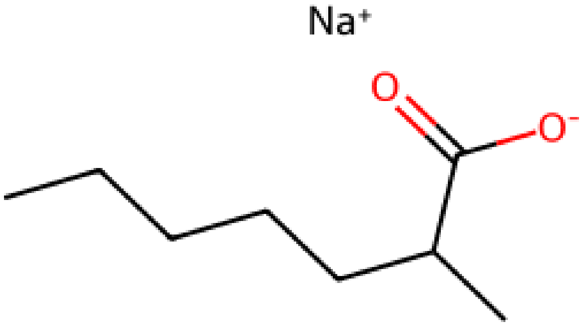
Salts

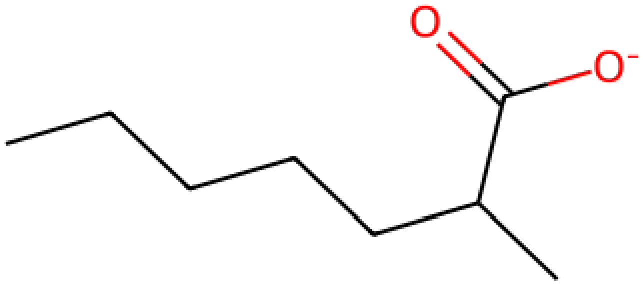
	Mean. Density (g/cm3)	Mp (°C)	Aqueous Solubility (mol/L)
Ibuprofen	1.11±0.001	76-79	3.45x10-4
+ Butylamine	1.07±0.001	104-106	0.583
+ Hexylamine	0.98±0.004	91-94	0.02
+ Octylamine	0.75±0.001	80-82	4.93x10-3
+ Benzylamine	1.08±0.001	107-109	7.99x10-4
+ Cyclohexylamine	1.12±0.001	197-201	3.37x10-3
+ Tert-butylamine	1.02±0.001	185-190	0.018
+ AMP1	1.11±0.001	130-134	0.458
+ AMP2	1.15±0.001	112-116	>0.643
+ Tris	1.20±0.001	160-164	0.0274

Data from: https://doi.org/10.3109/03639045.2011.592530

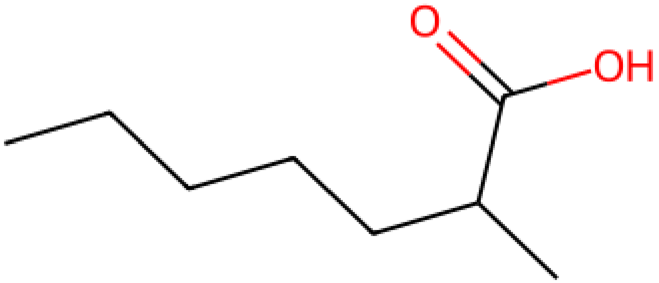
Neutralization

- Salt stripping often results in charged species





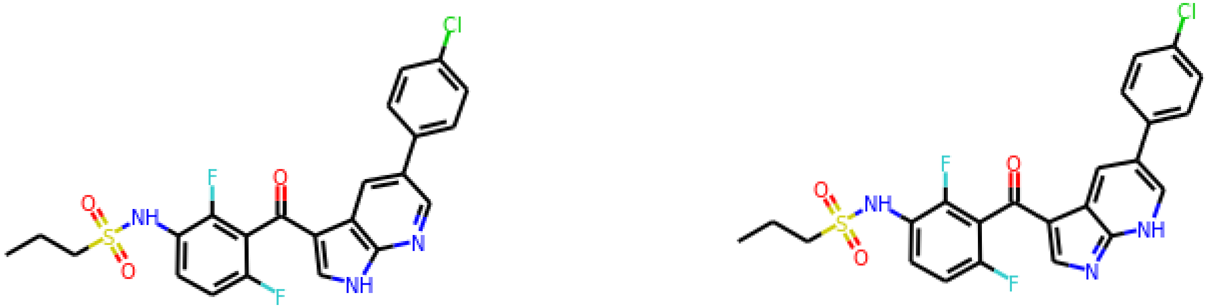
- We often store compounds in their neutral form



- Later, if necessary, we can adjust the protonation whatever is appropriate for our simulation

Why not just pick a tautomer?

If the tautomers can interconvert in solution why not just pick one and use that?

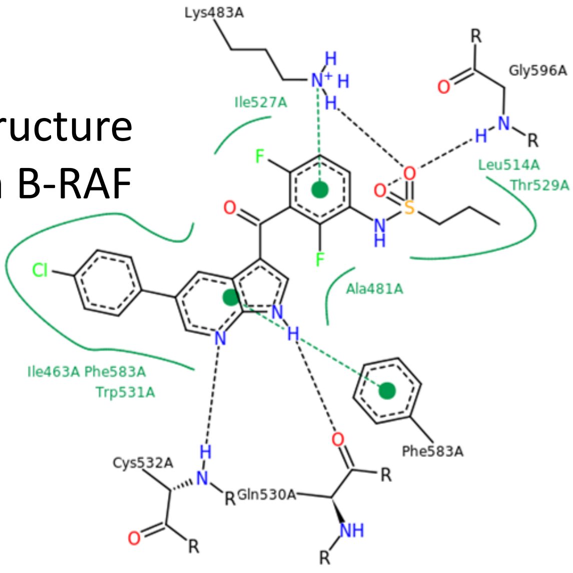


vemurafenib tautomers

Why not just pick a tautomer?

If the tautomers can interconvert in solution why not just pick one and use that?

Here's the crystal structure of vemurafenib with B-RAF



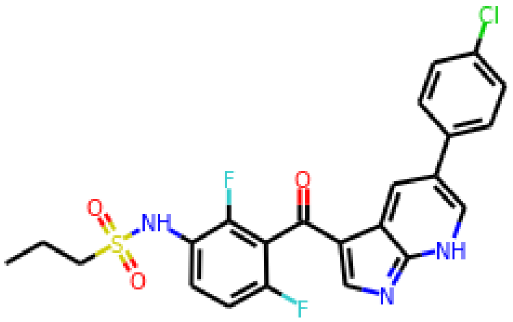
https://www.rcsb.org/structure/4RZV

If we pick the structure which doesn’t corresponds to what binds to the protein, we will have unreliable/incorrect input for whatever else we do with the molecule

Aside: does this actually matter for machine learning?

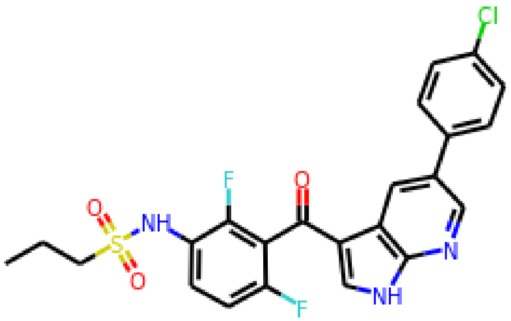
Fingerprint similarities between the two tautomers of vemurafenib:

- Morgan2: 0.79
- Feature Morgan2: 0.79
- RDKit fingerprint: 0.74
- Avalon fingerprint: 0.79
- Atom pairs: 1.0
- Topological torsions: 1.0



Descriptor values:

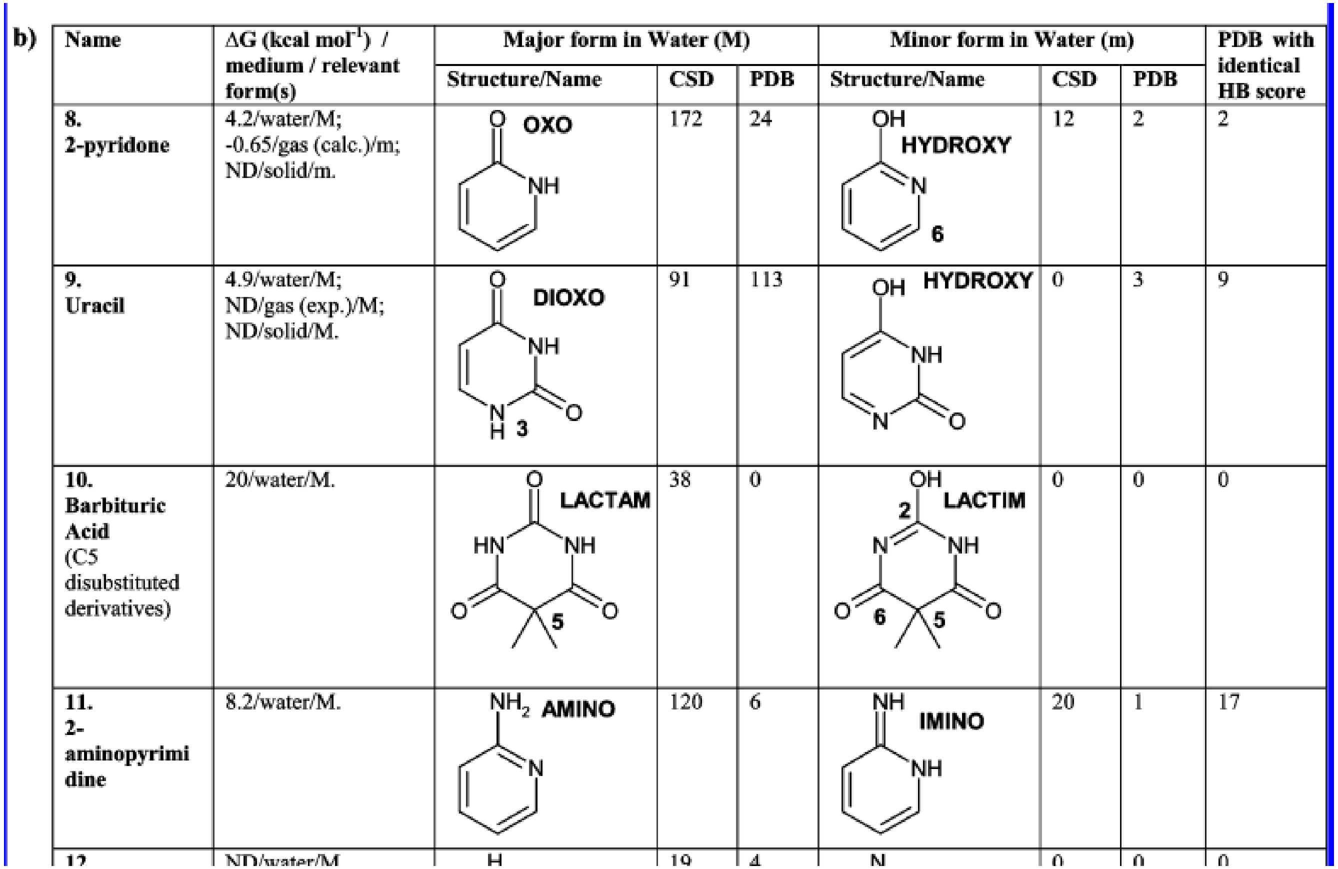
- MolLogP: 5.50, 5.54
- BCUT2D_CHGHI: 2.26, 2.27
- LabuteASA: 193.2, 193.6
- TPSA: 91.92, 91.92
- kappa3: 4.60, 4.60



It depends

More tautomer fun

https://pubs.acs.org/doi/10.1021/ci900501c



b)

Name	ΔG (kcal mol^{-1}) / medium / relevant form(s)	Major form in Water (M)			Minor form in Water (m)			PDB with identical HB score
		Structure/Name	CSD	PDB	Structure/Name	CSD	PDB	
8. 2-pyridone	4.2/water/M; -0.65/gas (calc.)/m; ND/solid/m.	O OXO NH	172	24	OH HYDROXY N 6	12	2	2

9.
Uracil
	4.9/water/M;
	ND/gas (exp.)/M;
	ND/solid/M.
		O
		DIOXO
		NH
		N
		H 3
		O
			91	113	OH
					HYDROXY
					NH
					N
					O
						0	3	9

10. Barbituric Acid (C5 disubstituted derivatives)	20/water/M.	O LACTAM HN NH O 5 O	38	0	OH 2 LACTIM N NH O 6 5 O	0	0	0
11. 2-aminopyrimidine	8.2/water/M.	NH_2 AMINO N	120	6	NH IMINO NH	20	1	17
12.	ND/water/M	H	19	4	N	0	0	0

Examples of rings/groups which have been observed in multiple tautomers in crystal structures

Public resources for chemical data science

- Collections of available compounds
  - ZINC
  - PubChem
- Compounds + bioactivity data
  - ChEMBL
  - PubChem BioAssay
  - ToxCast/Tox21
- This is a partial list… there are many others

ZINC

- “Zinc Is Not Commercial”: https://zinc20.docking.org/
- Free database of commercially-available compounds for virtual screening
- >2 billion molecules
- Many different subsets available: lead-like, drug-like, non-reactive, marketed drugs, etc.

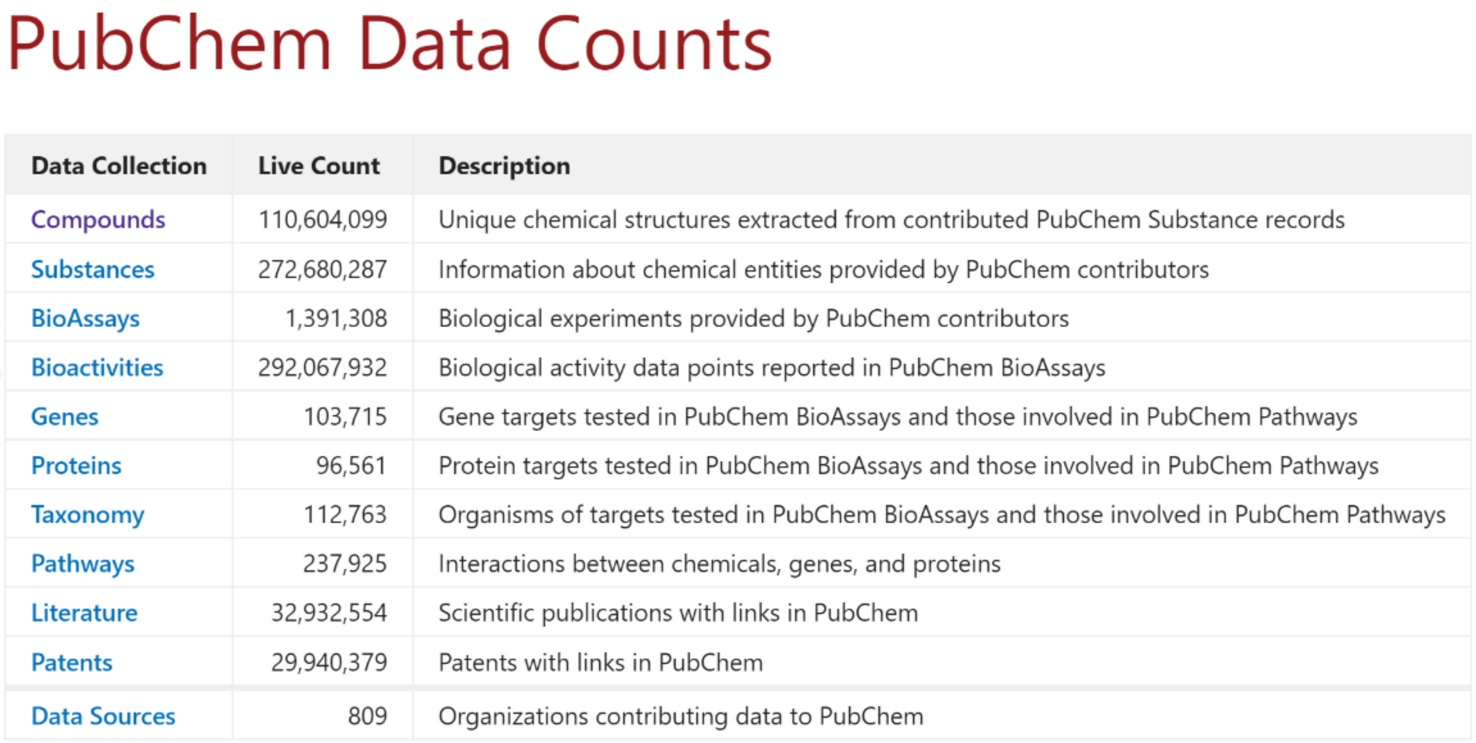
Paper: https://pubs.acs.org/doi/10.1021/acs.jcim.0c00675

PubChem

- https://pubchem.ncbi.nlm.nih.gov/
- “PubChem is the world's largest collection of freely accessible chemical information. Search chemicals by name, molecular formula, structure, and other identifiers. Find chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations and more.”
- PubChem Substance: chemical structures submitted by contributors
- PubChem Compound: standardized (= cleaned up) versions of the structures in PubChem Substance
  - Each PubChem Compound (CID) corresponds to one or more PubChem Substance (SID)
- 2D and 3D structures available
- 110 million compounds, 272 million substances

PubChem BioAssay

- Part of the larger PubChem project
- Includes a large collection of assay data, including full high-throughput screens



PubChem Data Counts

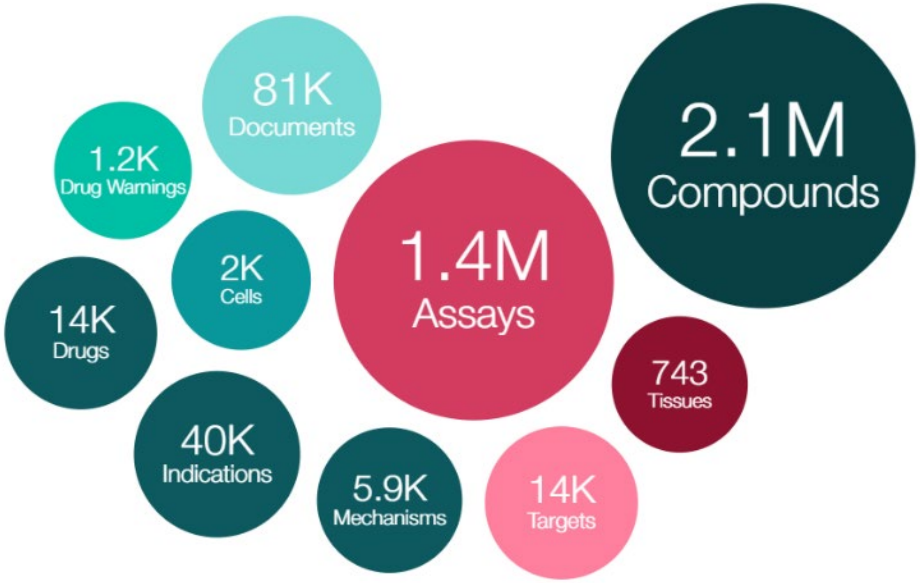
Data Collection	Live Count	Description
Compounds	110,604,099	Unique chemical structures extracted from contributed PubChem Substance records
Substances	272,680,287	Information about chemical entities provided by PubChem contributors
BioAssays	1,391,308	Biological experiments provided by PubChem contributors
Bioactivities	292,067,932	Biological activity data points reported in PubChem BioAssays
Genes	103,715	Gene targets tested in PubChem BioAssays and those involved in PubChem Pathways
Proteins	96,561	Protein targets tested in PubChem BioAssays and those involved in PubChem Pathways
Taxonomy	112,763	Organisms of targets tested in PubChem BioAssays and those involved in PubChem Pathways
Pathways	237,925	Interactions between chemicals, genes, and proteins
Literature	32,932,554	Scientific publications with links in PubChem
Patents	29,940,379	Patents with links in PubChem
Data Sources	809	Organizations contributing data to PubChem

https://pubchem.ncbi.nlm.nih.gov/

ChEMBL

"ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs."

- Open data, freely accessible
- Web interface
- Database dumps
- Web services



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

ToxCast/Tox21

- Tox21: https://tox21.gov/

  Tox21 researchers aim to develop better toxicity assessment methods to quickly and efficiently test whether certain chemical compounds have the potential to disrupt processes in the human body that may lead to negative health effects

- ToxCast: https://www.epa.gov/chemical-research/toxicity-forecasting
  - ToxCast has data for approximately 1,800 chemicals from a broad range of sources including industrial and consumer products, food additives, and potentially green chemicals that could be safer alternatives to existing chemicals.
  - ToxCast screens chemicals in more than 700 high-throughput assay endpoints that cover a range of high-level cell responses.

Acknowledgements

These slides are assembled from a set of lectures that Sereina Riniker originally created and that we've worked together on and expanded for the past couple of years.