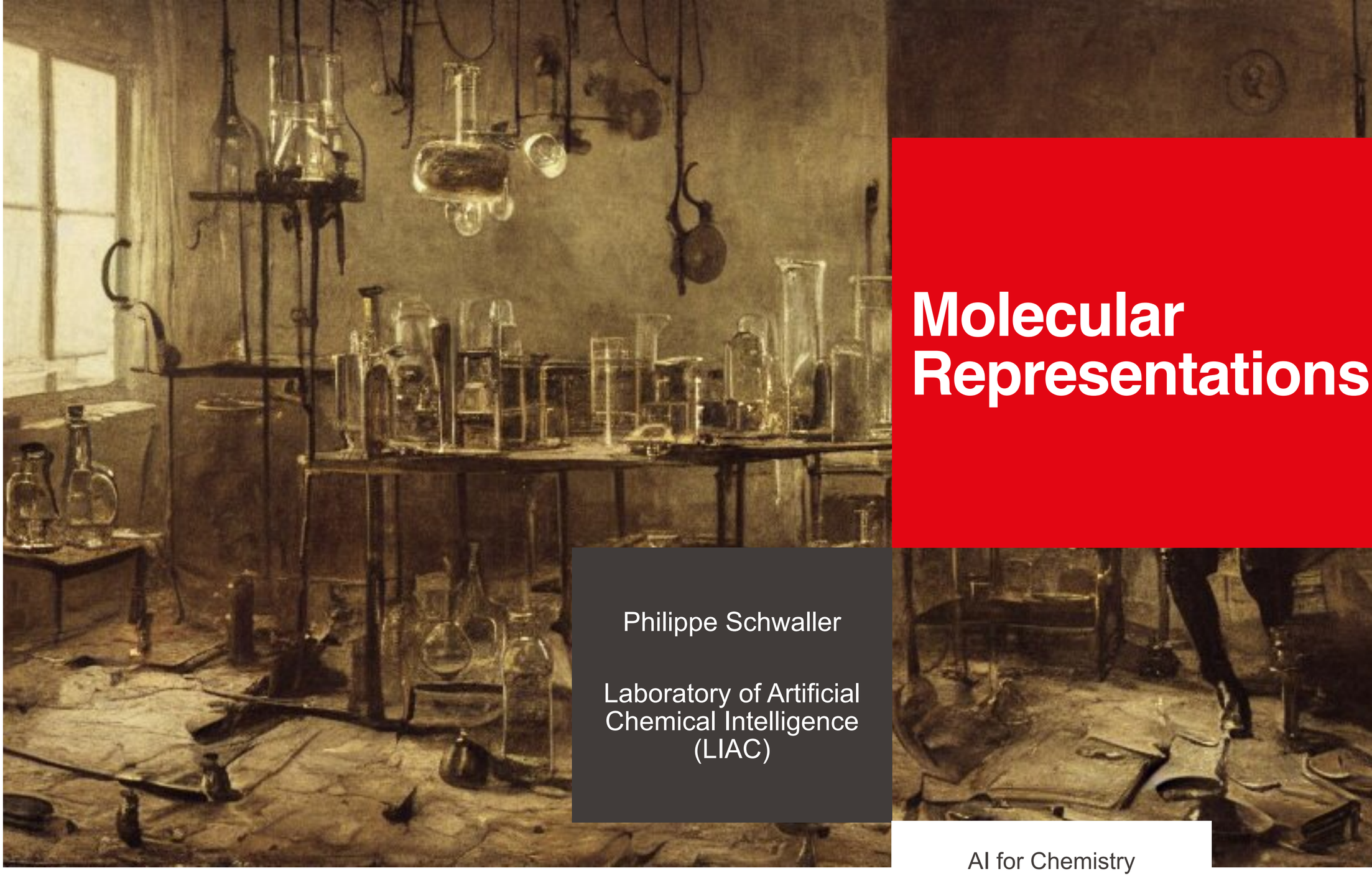


Molecular Representations

Philippe Schwaller

Laboratory of Artificial Chemical Intelligence (LIAC)

AI for Chemistry

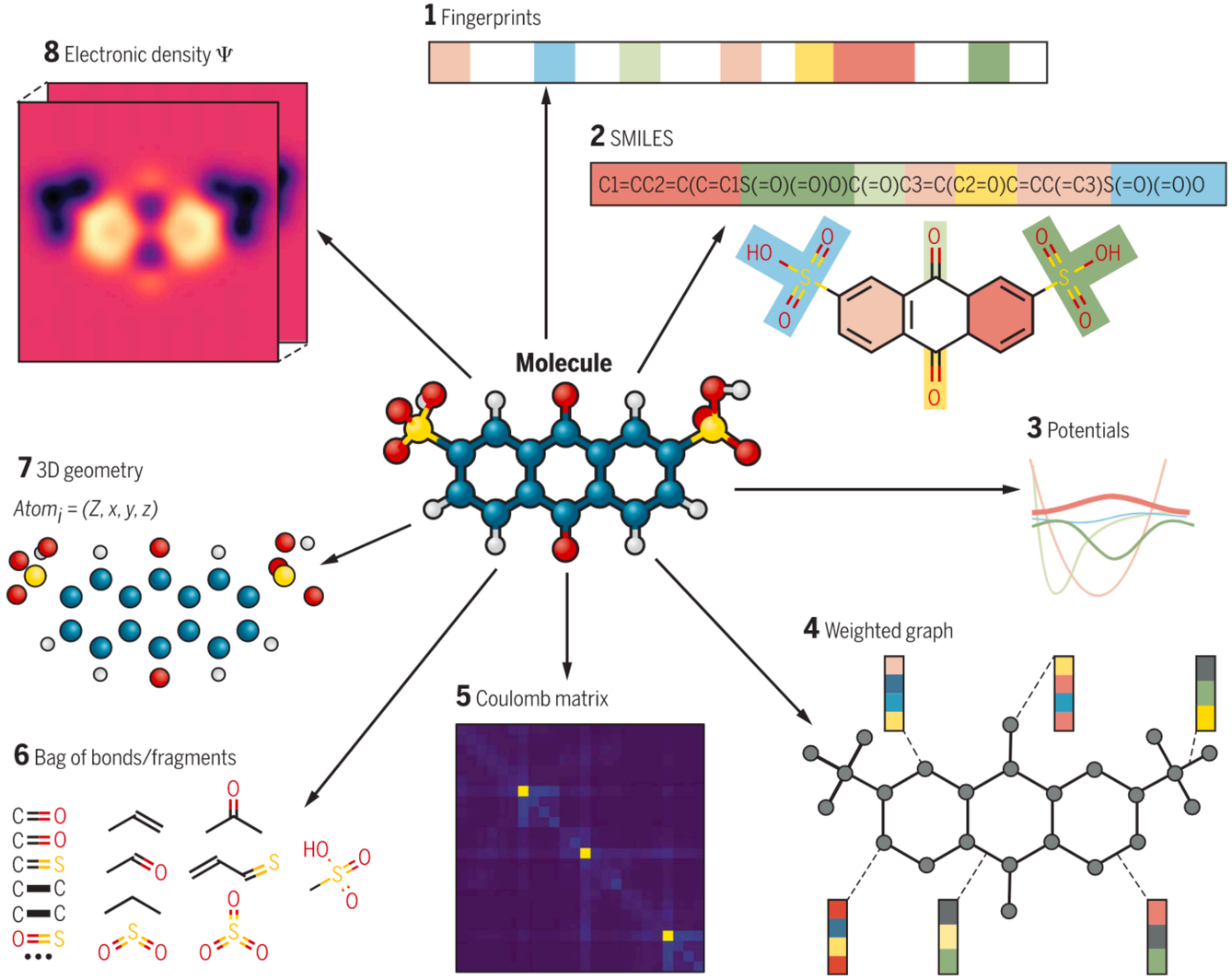
https://lexica.art/prompt/120c4d79-25ad-4155-b658-5f508e230efd

"… we cannot improve the language of any science, without, at the same time improving the science itself; neither can we, on the other hand, improve a science, without improving the language or nomenclature which belongs to it …"

— Lavoisier, 1787

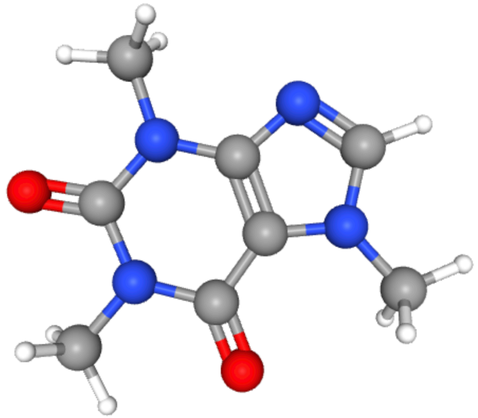
It's all about representations!

How to represent a molecule?



- https://www.science.org/doi/full/10.1126/science.aat2663

Overview



- **Machine-accessible formats** for molecules (how to store molecules on a computer)
- Molecular **representations**
  - Molecular **descriptors** (like molecular weight)
  - Molecular **fingerprints** (based on present/absent structural motifs)

Aim

Given a molecule, what is its activity?

$$f_\theta(\text{molecule}) = \text{activity}$$

ML model

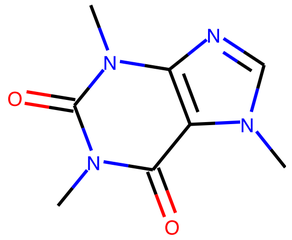
Challenge: Human-written names (e.g. Aspirin) typically do not provide **structural/atomic/molecular** information!

Key ingredients for Machine Learning

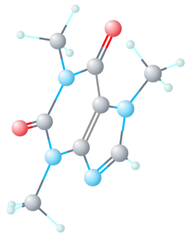
Molecular fingerprints
000010000....0100

Text-based representations
CN1C=NC2=C1C(=O)N(C(=O)N2C)C

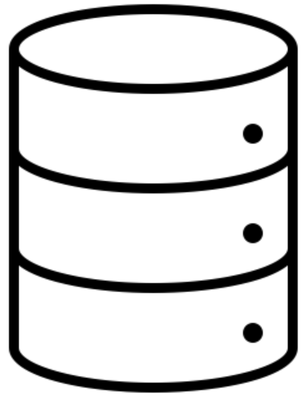
Graph-based representations



3D coordinates & surface



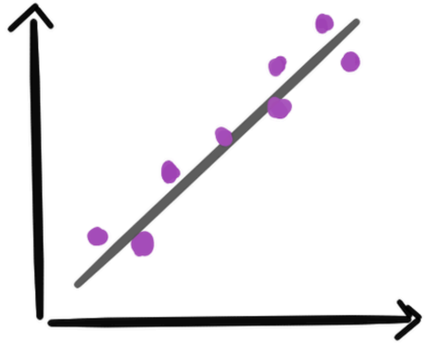
Representations (machine-readable)



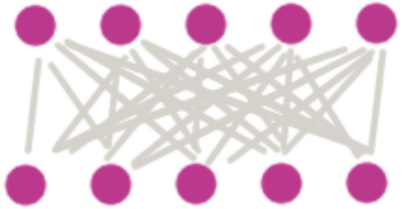
- Molecules & properties
- Chemical reactions
- Synthesis procedures

Data (garbage in = garbage out)

Linear regression model

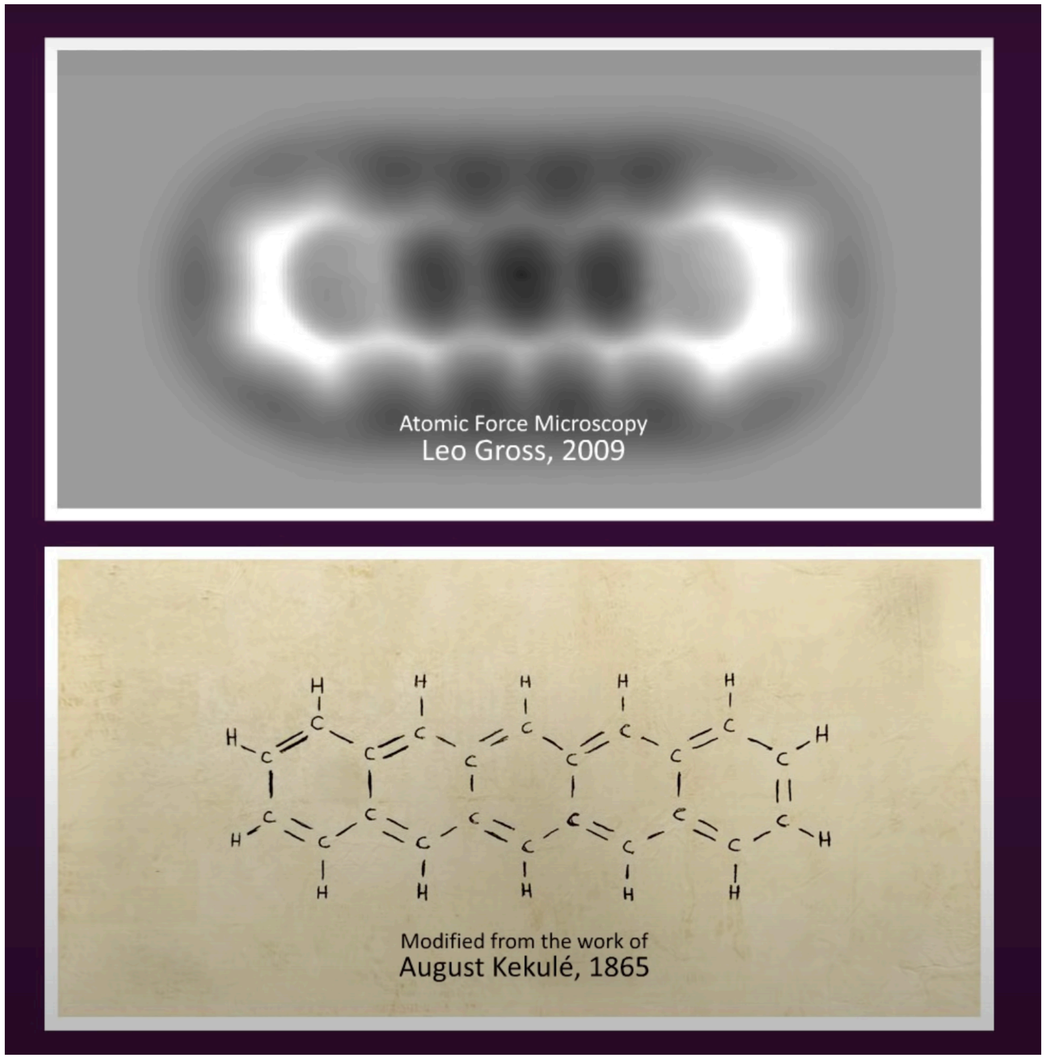


Neural networks



And many more..

Models/algorithms

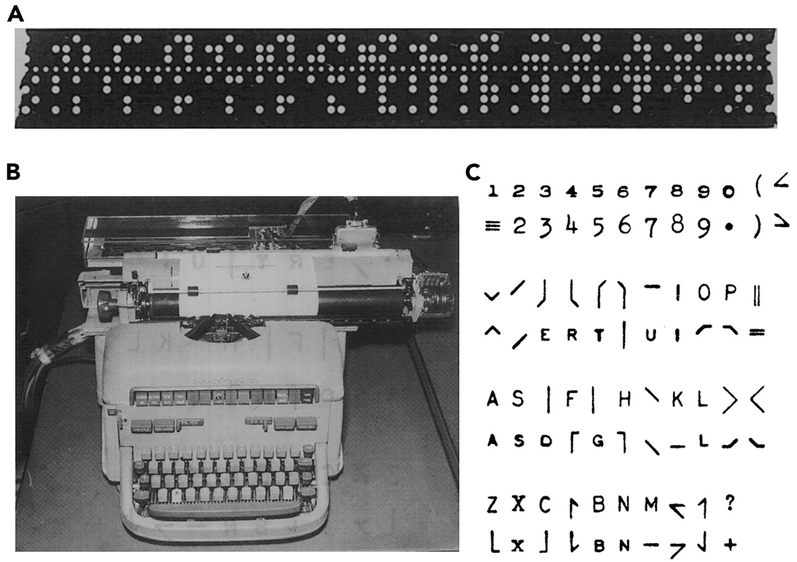


https://www.youtube.com/watch?v=C0Qaf-UJ2XQ

Storing molecular information

Need for machine-readable formats

- Databases of chemicals require machine-readable formats (e.g. for Chemical Abstracts Service, CAS)
- Development from the `60s
  - **Line notations**
    - Encode molecular graph *implicitly*
  - **Connection tables**
    - Encode molecular graph *explicitly*
- Enabled structure, substructure, and similarity searching (early use cases)



Army Chemical Typewriter (ACT)

https://doi.org/10.1016/j.patter.2022.100588

Storing molecular graphs

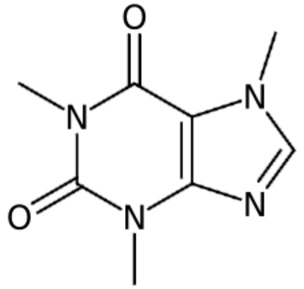
Simplified molecular-input line-entry system (SMILES)

SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules

DAVID WEININGER
Medicinal Chemistry Project, Pomona College, Claremont, California 91711

Received June 17, 1987

SMILES (Simplified Molecular Input Line Entry System) is a chemical notation system designed for modern chemical information processing. Based on principles of molecular graph theory, SMILES allows rigorous structure specification by use of a very small and natural grammar. The SMILES notation system is also well suited for high-speed machine processing. The resulting ease of usage by the chemist and machine compatability allow many highly efficient chemical computer applications to be designed including generation of a unique notation, constant-speed (zeroeth order) database retrieval, flexible substructure searching, and property prediction models.



caffeine

SMILES. 2. Algorithm for Generation of Unique SMILES Notation

DAVID WEININGER, ARTHUR WEININGER, and JOSEPH L. WEININGER*
Daylight Chemical Information Systems, Irvine, California 92714

Received May 4, 1988

The chemical notation language SMILES is designed for the conversion of an arbitrarily chosen description of a chemical structure to one unique notation. This is accomplished in a two-stage algorithm, CANGEN. The first stage involves CANonicalization of structure, whereby the molecule is treated as a graph with nodes (atoms) and edges (bonds). Each atom is canonically ordered and labeled. In the second stage, starting with the lowest labeled atom, a molecular graph is GENerated, which is the unique SMILES structure.

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

SMILES basics

Atoms:
aliphatic_organic 'B' | 'C' | 'N' | 'O' | 'S' | 'P' | 'F' | 'Cl' | 'Br' | 'I'
aromatic_organic 'b' | 'c' | 'n' | 'o' | 's' | 'p'
Other atoms in brackets, e.g. [Li]

-	Single bond	
=	Double bond	
#	Triple bond	
*	Aromatic bond	
.	Disconnected structures	

CC	CH_3CH_3	Ethane
C=C	CH_2CH_2	Ethene
CBr	CH_3Br	Bromomethane
C#N	C≡N	Hydrocyanic acid
Na.Cl	NaCl	Sodium chloride

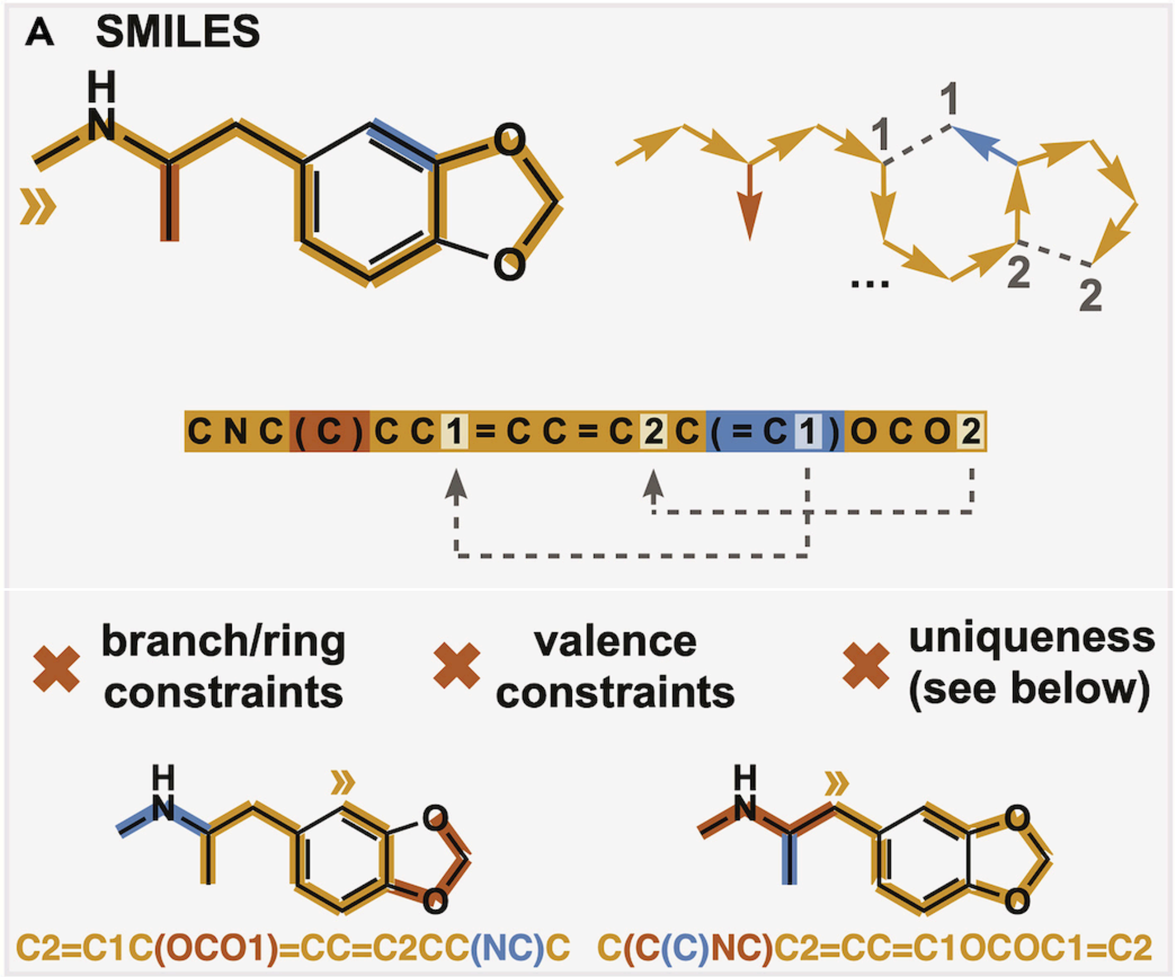
CC(O)C	2-Propanol
CC(=O)C	2-Propanone
CC(CC)C	2-Methylbutane
CC(C)CC(=O)	2-Methylbutanal
c1c(N(=O)=O)cccc1	Nitrobenzene
CC(C)(C)CC	2,2-Dimethylbutane

c1ccccc1	Benzene
C1OC1CC	Ethyloxirane
c1cc2ccccc2cc1	Naphthalene

Hydrogen atoms are often implicit!
=> can be calculated from the atom's valence

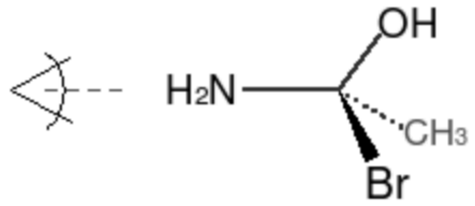
- http://opensmiles.org/opensmiles.html, https://archive.epa.gov/med/med_archive_03/web/html/smiles.html

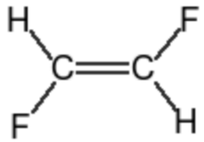
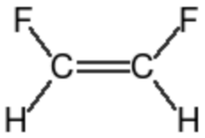
Non-uniqueness of SMILES — canonicalisation



- https://www.cell.com/patterns/fulltext/S2666-3899(22)00206-9

Stereochemistry

Tetrahedral Chirality	
look from N towards C (chiral center)	list the neighbors anticlockwise
	`N[C@](Br)(O)C`
	...or clockwise
	`N[C@@](Br)(C)O`

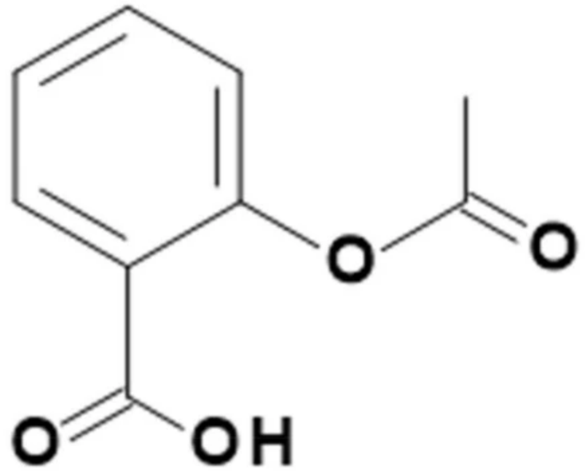
Depiction	SMILES	Name
	`F/C=C/F` `F\C=C\F`	trans-difluoroethane **(both SMILES are equivalent)**
	`F\C=C/F` `F/C=C\F`	cis-difluoroethane **(both SMILES are equivalent)**

SMILES could technically also encode Square Planar Centers, (`@SP1`, `@SP2` and `@SP3`)

Trigonal Bipyramidal Centers, Octahedral Centers, etc.

- http://opensmiles.org/opensmiles.html

InChI, the IUPAC International Chemical Identifier



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

Chemical formula | Atom connections | Hydrogen atoms

- https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00460-5

Connection tables

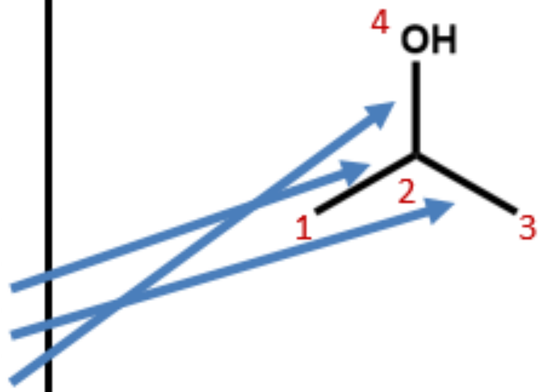
SCT I

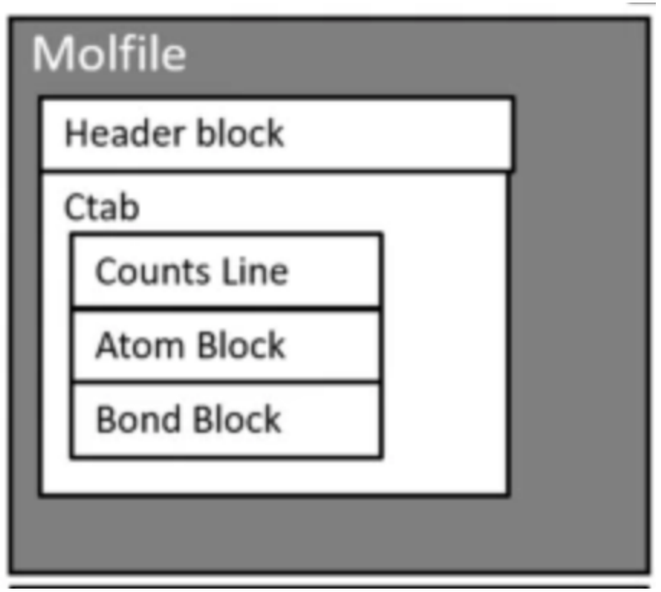
Atom table

Atom #	Atom type
1	C
2	C
3	C
4	O

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	1
2	4	1





- Representing D-alanine
  - using a Molfile takes 612 bytes
  - using InChI 59 bytes
  - using SMILES 15 bytes

- https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/02:_Representing_Small_Molecules_on_Computers/2.02:_Connection_Tables, https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00460-5

Connection tables — non-uniqueness

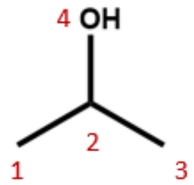
SCT I

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	O

Bond table

Atom 1	Atom 2	Bond order
1	2	1
2	3	1
2	4	1



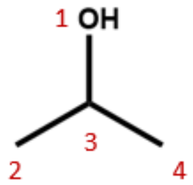
SCT II

Atom table

Atom #	Atom type
1	O
2	C
3	C
4	C

Bond table

Atom 1	Atom 2	Bond order
1	3	1
2	3	1
3	4	1



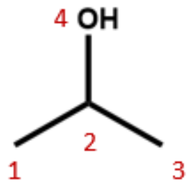
SCT III

Atom table

Atom #	Atom type
1	C
2	C
3	C
4	O

Bond table

Atom 1	Atom 2	Bond order
2	3	1
1	2	1
4	2	1



- https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics/02:_Representing_Small_Molecules_on_Computers/2.02:_Connection_Tables

Storing atomic coordinates

Cartesian coordinates (xyz)


```
12
benzene example
  C          0.00000          1.40272          0.00000
  H          0.00000          2.49029          0.00000
  C         -1.21479          0.70136          0.00000
  H         -2.15666          1.24515          0.00000
  C         -1.21479         -0.70136          0.00000
  H         -2.15666         -1.24515          0.00000
  C          0.00000         -1.40272          0.00000
  H          0.00000         -2.49029          0.00000
  C          1.21479         -0.70136          0.00000
  H          2.15666         -1.24515          0.00000
  C          1.21479          0.70136          0.00000
  H          2.15666          1.24515          0.00000
```

Atomic coordinates (without defined bonds)

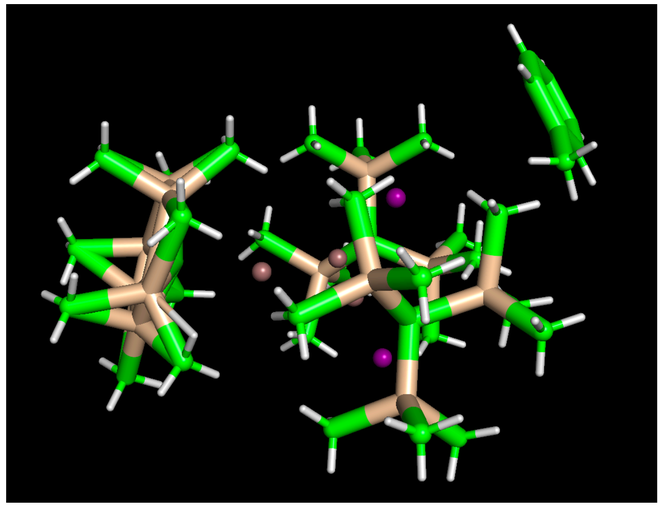
minimal content: element and coordinates => not for crystals

- https://openbabel.org/docs/2.3.1/FileFormats/XYZ_cartesian_coordinates_format.html

Crystal structures (.cif)


```
;
bis(\m~2~-Iodo)-tris(tris(trimethylsilyl)methyl)-tri-indium toluene solvate
;
_space_group_IT_number           4
_symmetry_cell_setting           monoclinic
_symmetry_space_group_name_Hall  'P 2yb'
_symmetry_space_group_name_H-M   'P 1 21 1'
_cell_angle_alpha                90
_cell_angle_beta                 102.491(5)
_cell_angle_gamma                90
_cell_formula_units_Z            2
_cell_length_a                   9.1269(5)
_cell_length_b                   24.8574(10)
_cell_length_c                   13.5416(8)
_cell_volume                     2999.477
_diffrn_ambient_temperature      ?
_exptl_crystal_colour            yellow
_exptl_crystal_density_diffrn    1.534
_exptl_crystal_description       plates
_refine_ls_R_factor_gt           2.38
_refine_ls_wR_factor_gt          2.38
_cod_original_sg_symbol_H-M      'P 21'
_cod_database_code               1100231
loop_
_symmetry_equiv_pos_site_id
_symmetry_equiv_pos_as_xyz
1 x,y,z
2 -x,1/2+y,-z
```

```
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
In1 In 0.30475(3) 0.21073(1) 0.21076(2)
C1 C 0.2164(5) 0.28418(16) 0.1254(3)
Si1 Si 0.31124(14) 0.34775(5) 0.18418(10)
C2 C 0.5024(6) 0.3589(2) 0.1568(5)
H1 H 0.49440 0.35990 0.08490
H2 H 0.56790 0.33010 0.18550
H3 H 0.54240 0.39240 0.18590
C3 C 0.3382(6) 0.3502(2) 0.3248(4)
H4 H 0.40170 0.32100 0.35420
H5 H 0.24270 0.34710 0.34310
H6 H 0.38420 0.38370 0.34940
C4 C 0.2005(7) 0.40959(19) 0.1342(5)
H7 H 0.25200 0.44100 0.16520
H8 H 0.10290 0.40770 0.14970
H9 H 0.19020 0.41150 0.06220
Si2 Si 0.00766(12) 0.28337(5) 0.12276(10)
C5 C -0.0699(5) 0.2130(2) 0.1191(4)
H10 H -0.02880 0.19510 0.18190
H11 H -0.04320 0.19340 0.06450
H12 H -0.17720 0.21450 0.10920
C6 C -0.1046(5) 0.3194(2) 0.0092(4)
H13 H -0.20930 0.31720 0.01000
H14 H -0.08730 0.30280 -0.05130
H15 H -0.07440 0.35640 0.01100
C7 C -0.0411(6) 0.3150(2) 0.2359(5)
H16 H -0.00200 0.35100 0.24360
```



Converting coordinates to molecular graphs

Convert Cartesian coordinates to one or more molecular graphs

Given Cartesian coordinates in the form of a `.xyz` file, the code constructs a list of one or more molecular graphs. In cases where there are several possible resonance forms xyz2mol returns a list of all, otherwise just a list of one.

This code is based on the work of DOI: 10.1002/bkcs.10334

```
Yeonjoon Kim and Woo Youn Kim
"Universal Structure Conversion Method for Organic Molecules:
From Atomic Connectivity to Three-Dimensional Geometry"
Bull. Korean Chem. Soc.
2015, Vol. 36, 1769-1777
```

https://github.com/jensengroup/xyz2mol

cell2mol: Unit Cell to Molecule Interpretation

Sergi Vela, Ruben Laplaza, Yuri Cho & Clémence Corminboeuf

npj Computational Materials **8**, Article number: 188 (2022) | Cite this article

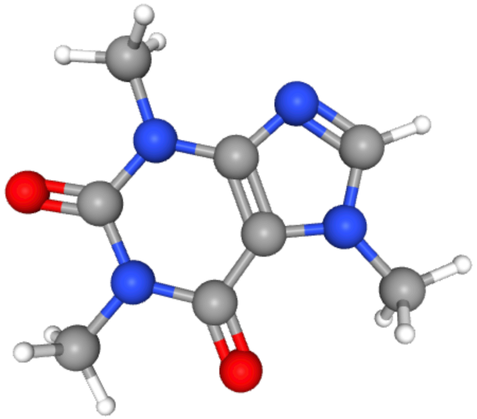
https://github.com/lcmd-epfl/cell2mol

Now that we have **machine-accessible** molecular **formats** to **store molecules**. What's next?

$$f_\theta(\text{molecule}) = \text{activity}$$

Can we represent the molecules as **vectors** of numbers? How do we make them **comparable**?

Molecular descriptors



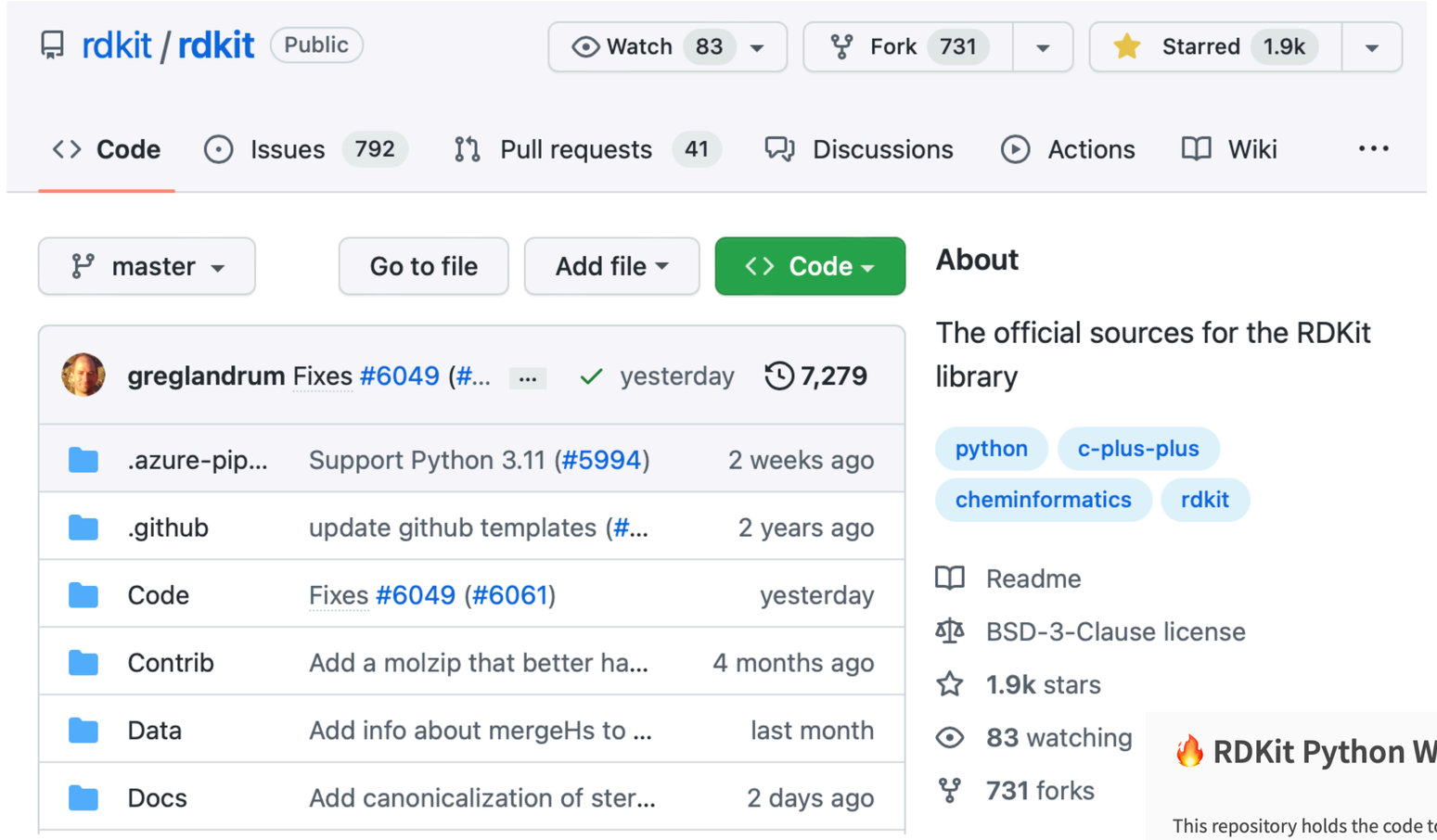
CN1C=NC2=C1C(=O)N(C(=O)N2C)C

Open-Source Cheminformatics and Machine Learning

MolWt
ExactMolWt
HeavyAtomCount
HeavyAtomMolWt
NHOHCount
NOCount
NumHAcceptors
NumHDonors
NumHeteroatoms
NumRotatableBonds
NumValenceElectrons
NumAmideBonds
Num{Aromatic,Saturated,Aliphatic}Rings
Num{Aromatic,Saturated,Aliphatic} {Hetero,Carbo}cycles
RingCount
FractionCSP3

There are many more.

RDKit - the most used Python cheminformatics toolkit



https://github.com/rdkit/rdkit

Mordred: a molecular descriptor calculator

number of descriptors

```
>>> from mordred import Calculator, descriptors
>>> n_all = len(Calculator(descriptors, ignore_3D=False).descriptors)
>>> n_2D = len(Calculator(descriptors, ignore_3D=True).descriptors)
>>> print("2D:    {:5}\n3D:    {:5}\n------------\ntotal: {:5}".format(n_2D, n_all - n_2D, n_all)
2D:     1613
3D:      213
------------
total:  1826
```

- mordred.ABCIndex module
- mordred.AcidBase module
- mordred.AdjacencyMatrix module
- mordred.Aromatic module
- mordred.AtomCount module
- mordred.Autocorrelation module
- mordred.BCUT module
- mordred.BalabanJ module
- mordred.BaryszMatrix module
- mordred.BertzCT module
- mordred.BondCount module
- mordred.CPSA module
- mordred.CarbonTypes module
- mordred.Chi module
- mordred.Constitutional module
- mordred.DetourMatrix module
- mordred.DistanceMatrix module
- mordred.EState module
- mordred.EccentricConnectivityIndex module
- mordred.ExtendedTopochemicalAtom module
- mordred.FragmentComplexity module
- mordred.Framework module
- mordred.GeometricalIndex module
- mordred.GravitationalIndex module
- mordred.HydrogenBond module
- mordred.InformationContent module
- mordred.KappaShapeIndex module
- mordred.Lipinski module

atomic properties

c
: gasteiger charge

dv
: valence electrons

d
: sigma electrons

s
: intrinsic state

Z
: atomic number

m
: mass

v
: vdw volume

se
: sanderson EN

pe
: pauling EN

are
: allred-rocow EN

p
: polarizability

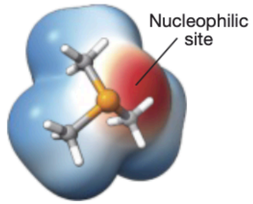
i
: ionization potential

https://github.com/mordred-descriptor/mordred

- https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0258-y,

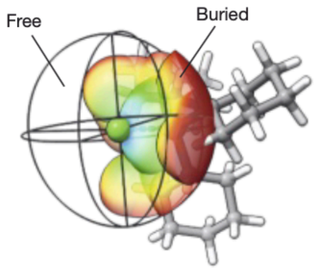
QM descriptors

Electronic



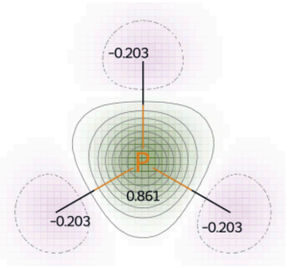
Average local ionization energy

Steric



Buried volume

Topological



LogP atom contributions

MORFEUS

molecular features for machine learning

MORFEUS calculates molecular features from 3D structures with a focus on steric descriptors. It can be used as a Python library or through command line scripts.

Examples

MORFEUS can be imported as a Python module that is easily integrated into workflows. It can also be used from the command line. Here is an example for calculating the exact ligand cone angle:

Module | Command line

```
>>> from morfeus import ConeAngle, read_xyz
>>> elements, coordinates = read_xyz("PdPMe3.xyz")
>>> cone_angle = ConeAngle(elements, coordinates, 1)
>>> print(cone_angle.cone_angle)
117.11012922937584
```

METHODS

- Bite angle
- Buried volume
- Cone angle
- Conformer tools
- Dispersion descriptor
- Local force constants
- Pyramidalization
- Solvent accessible surface area
- Solid angle
- Sterimol
- XTB

Prof. Dr. Kjell Jorner

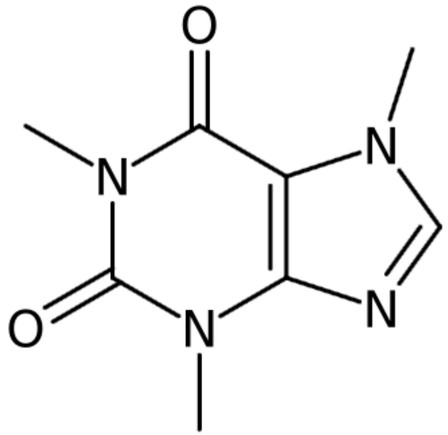
Address

ETH Zürich
Dep. of Chemistry and Applied Biosc.
Prof. Dr. Kjell Jorner
Professur für Digitale Chemie
HCI E 137
Vladimir-Prelog-Weg 1-5/10
8093 Zürich
Switzerland

+41 44 633 35 30
kjell.jorner@chem.ethz.ch

- https://github.com/kjelljorner/morfeus, https://chimia.ch/chimia/article/view/2023_22/2023_22

Molecular fingerprints

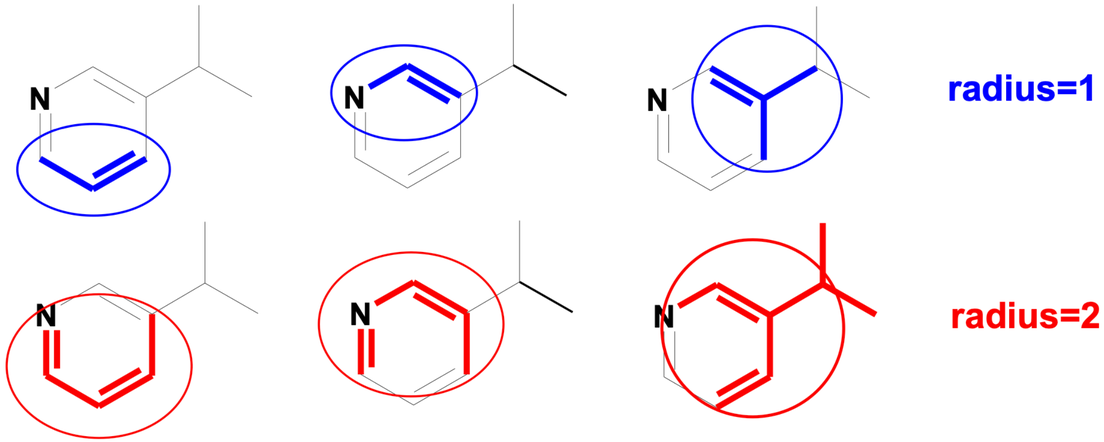


caffeine

[0, 0, 1, 0, …, 0, 1, 1]

Depending on the substructures, present (1) or absent (1) in the molecular structure.

ECFP / Morgan fingerprint (circular)

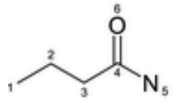


radius=1

radius=2

Atomic properties to create hash

- **Connectivity**: Element, #heavy neighbors, #Hs, charge, isotope, inRing
- **Chemical features**: Donor, Acceptor, Aromatic, Halogen, Basic, Acidic



> <ECFP_0>	> <ECFP_2>	> <ECFP_4>	> <ECFP_6>
734603939	734603939	734603939	734603939
1559650422	1559650422	1559650422	1559650422
-1100000244	-1100000244	-1100000244	-1100000244
1572579716	1572579716	1572579716	1572579716
-1074141656	-1074141656	-1074141656	-1074141656
	863188371	863188371	863188371
	-1793471910	-1793471910	-1793471910
	-1789102870	-1789102870	-1789102870
	-1708545601	-1708545601	-1708545601
	-932108170	-932108170	-932108170
	2099970318	2099970318	2099970318
		-87618679	-87618679
		1112638790	1112638790
		-627599602	-627599602

Modulo fingerprint size →

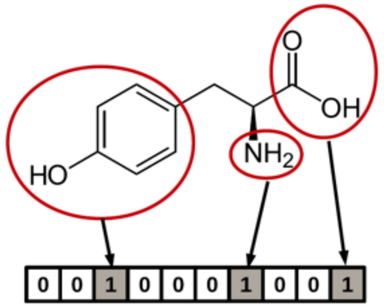
[0, 0, 1, …, 1, 0, 1]

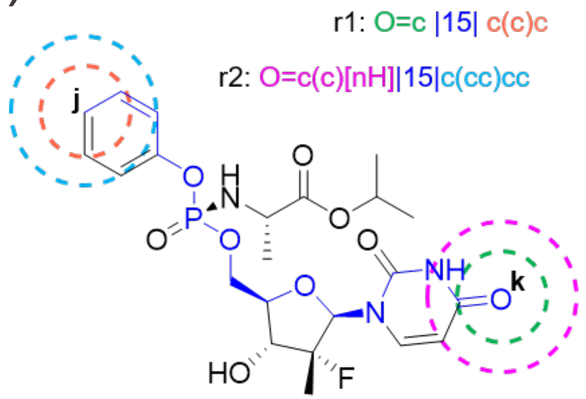
Bit vector

- ECFP https://pubs.acs.org/doi/full/10.1021/ci100050t (2010), Morgan Algorithm https://pubs.acs.org/doi/pdf/10.1021/c160017a018 (1965)
  https://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf

Other molecular fingerprints

- MACCS fingerprint (functional groups)
- Atom pair fingerprints (good for large mols)
- MinHash fingerprint (shingles - good for small mols)
- One molecular fingerprint to rule them all:
  drugs, biomolecules, and the metabolome
  => MAP4 fingerprint (https://tm.gdb.tools/map4/)

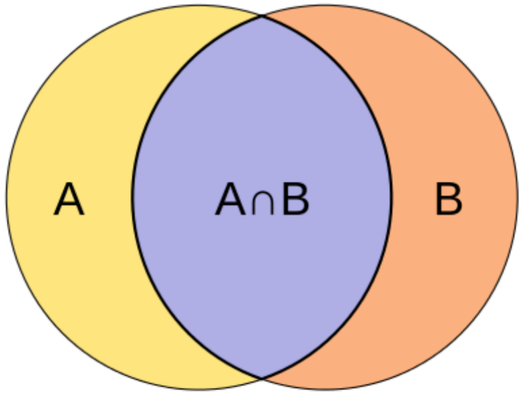




- https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00445-4

Similarity in molecules — Tanimoto coefficient / Jaccard Index

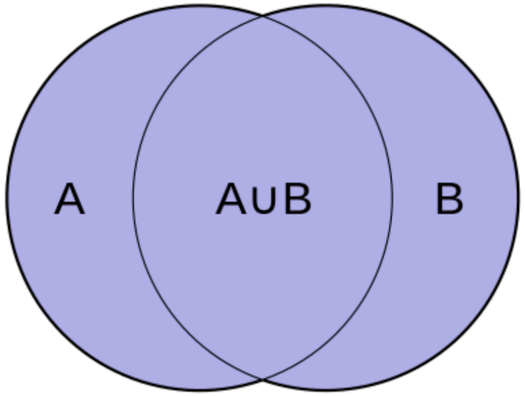
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Intersection of the two molecular fingerprints A and B, **divided by the union**.

A: [1, 1, 0]
B: [0, 1, 1]

$$\frac{1}{3} = 0.333$$



- https://pubs.rsc.org/en/content/articlelanding/2004/OB/b409813g

Research article | Open Access | Published: 12 February 2020

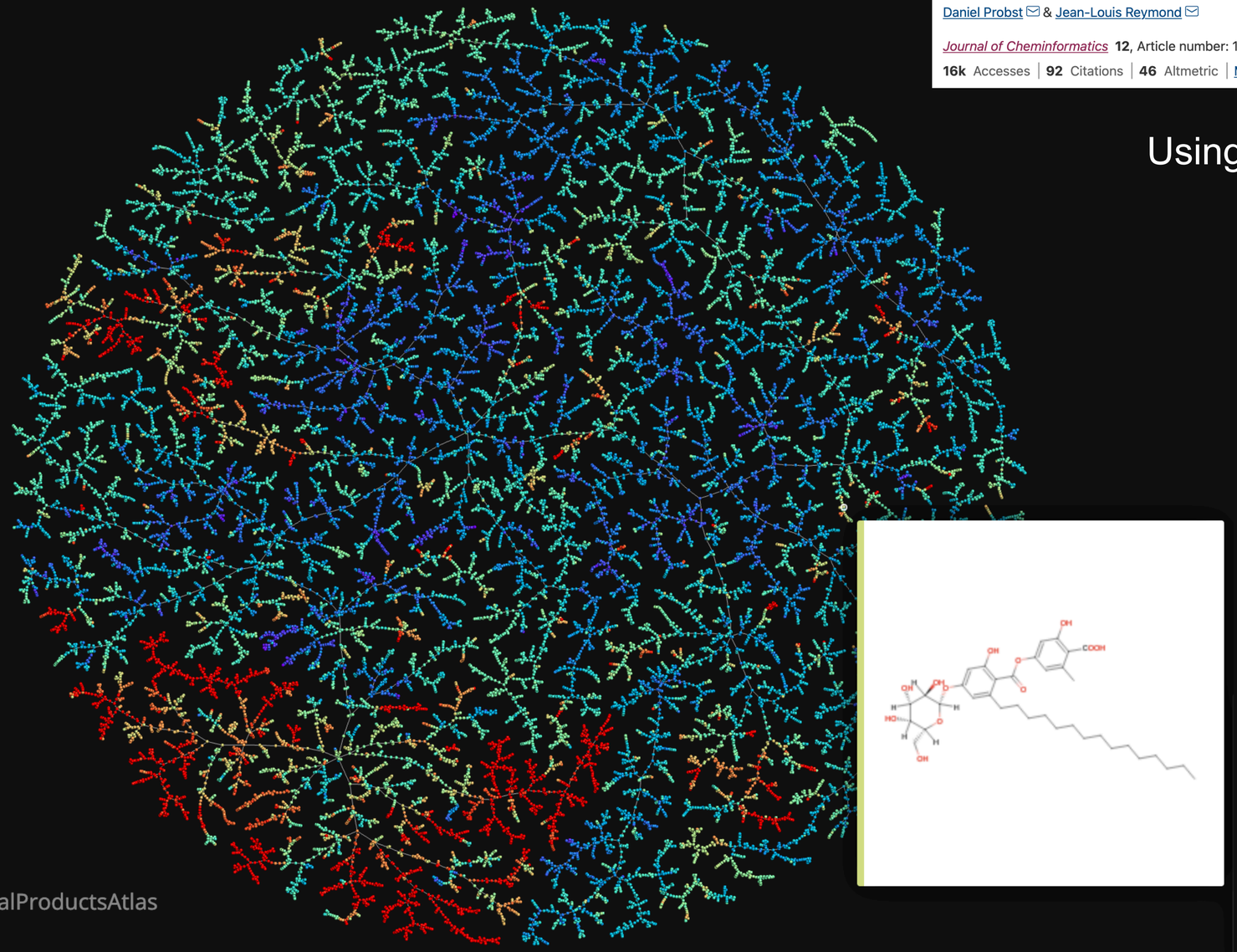
Visualization of very large high-dimensional data sets as minimum spanning trees

Daniel Probst & Jean-Louis Reymond

Journal of Cheminformatics **12**, Article number: 12 (2020) | Cite this article

16k Accesses | **92** Citations | **46** Altmetric | Metrics

Using MAP4 fingerprint



- https://tm.gdb.tools/map4/npatlas_map_tmap/,