Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

43384 – Digital Alchemy

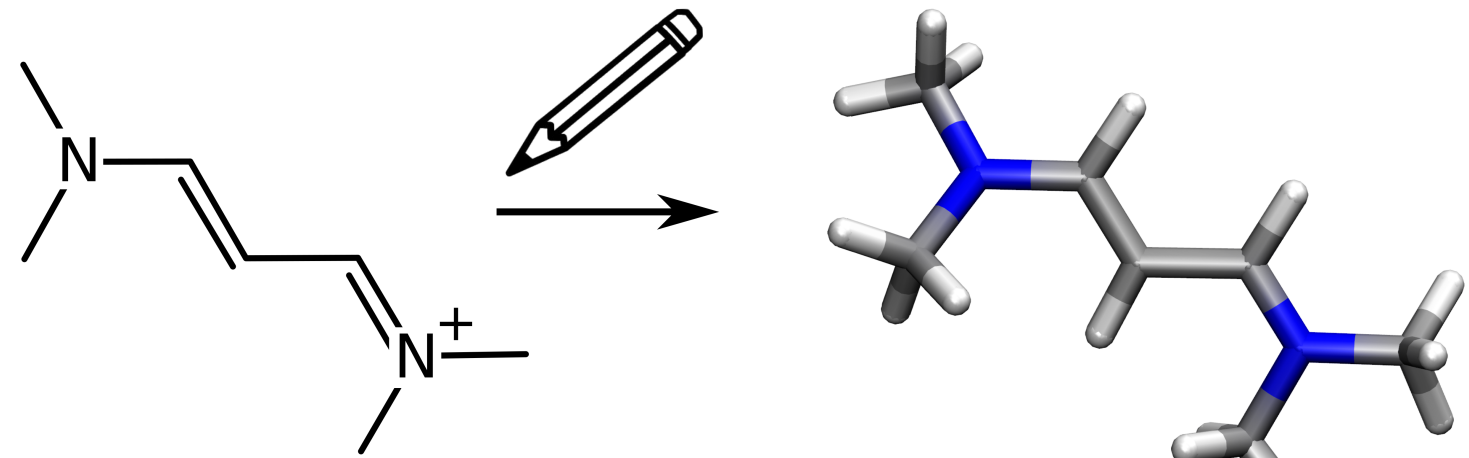
Unit 03 – Processing Configurational and Constitutional Information

Prof. Dr. Carolin Müller

October 27, 2025

3D Structures and Information

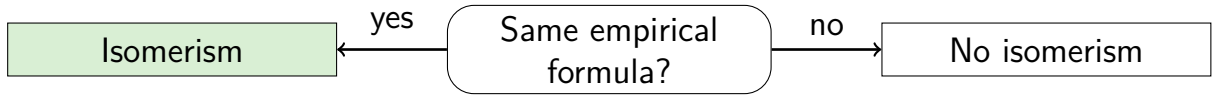
Encoding, Processing and Generation



CN(C)C=CC=[N+](C)C

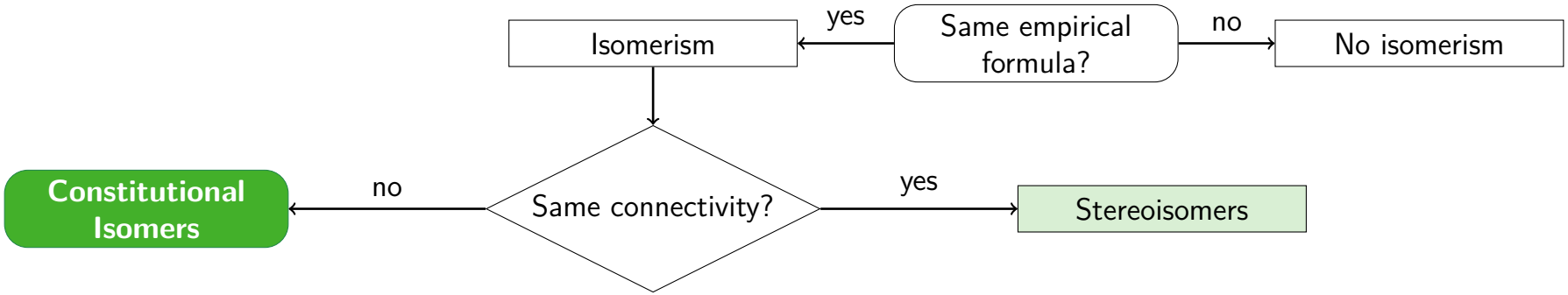
3D Structure – Isomerism

Specific Types of Chemical Structures



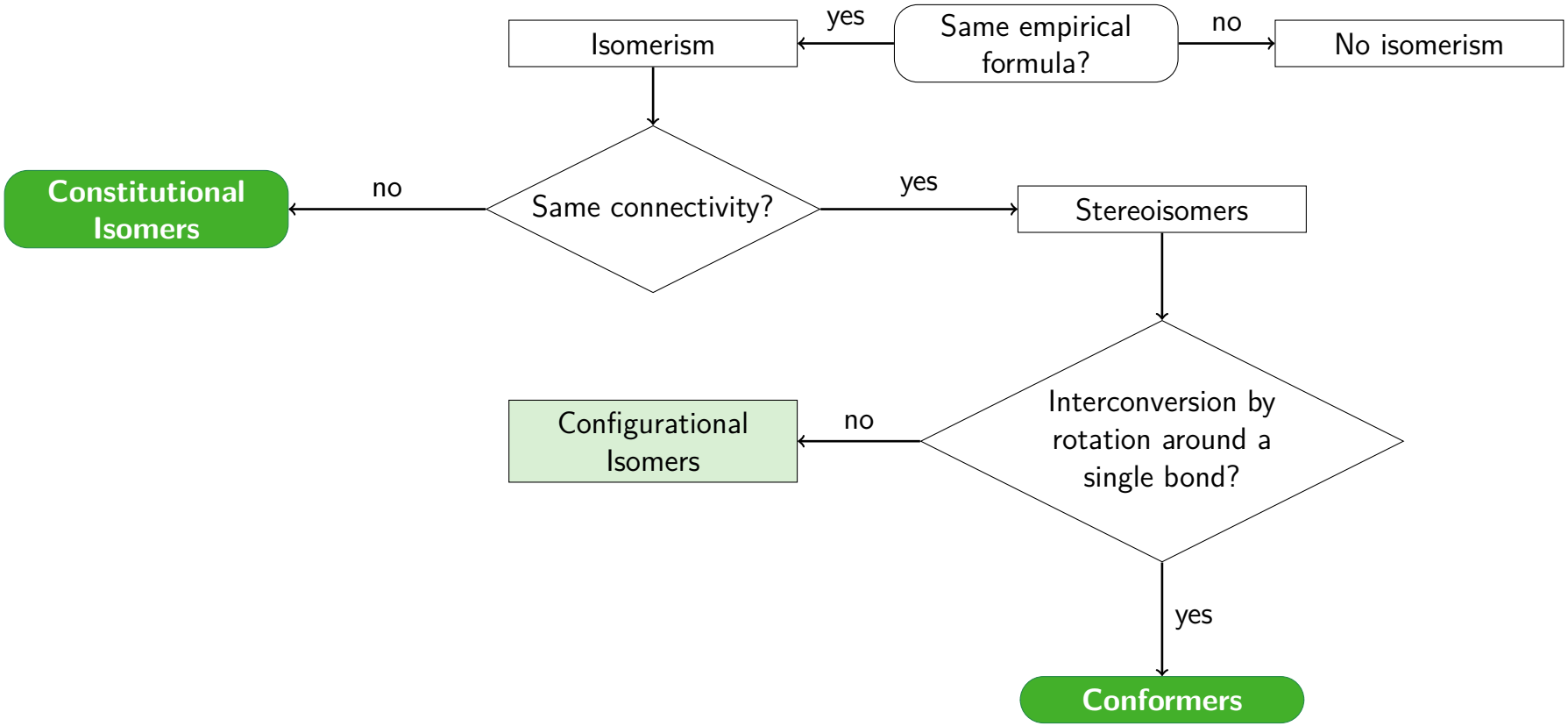
3D Structure – Isomerism

Specific Types of Chemical Structures



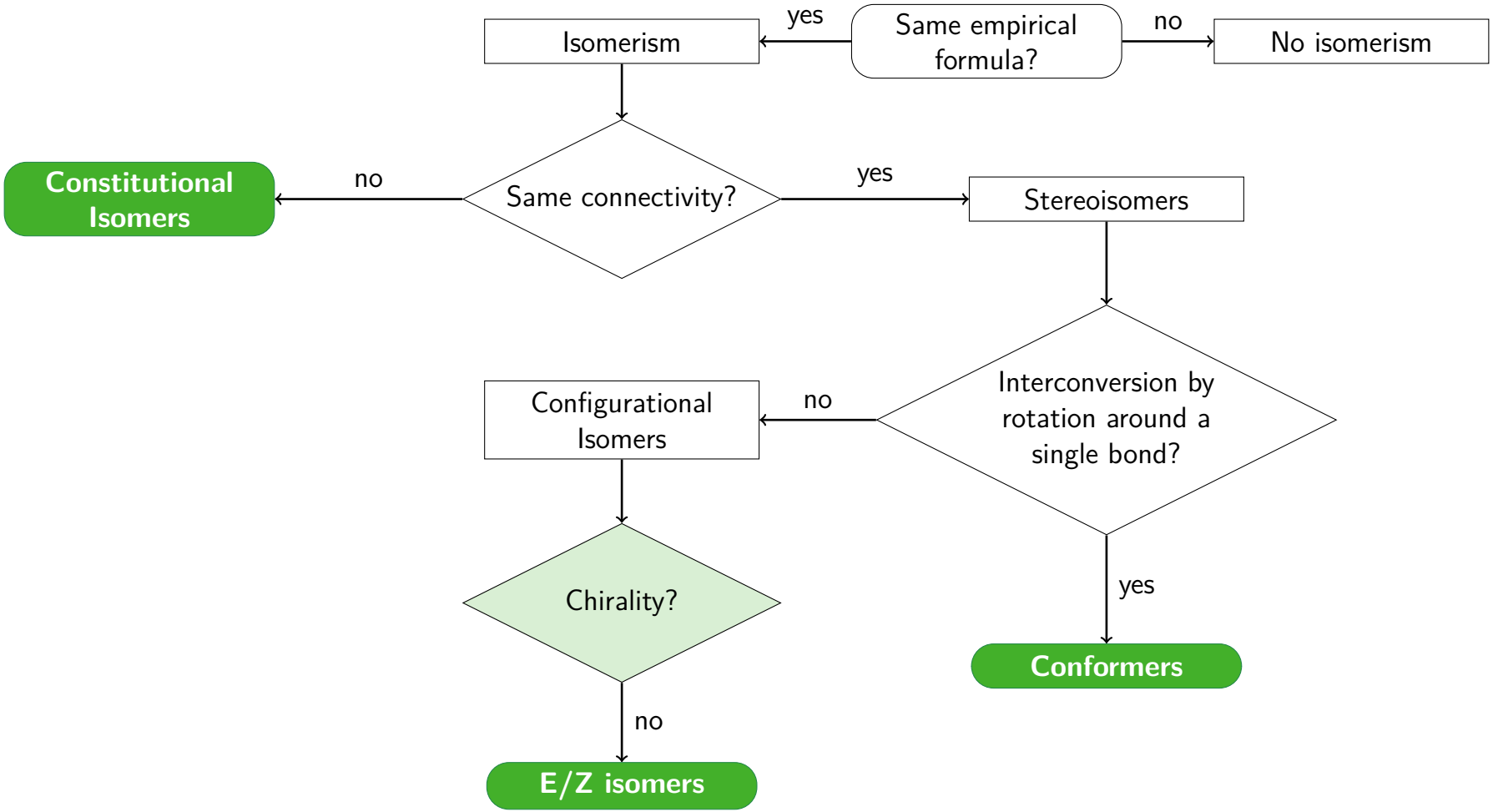
3D Structure – Isomerism

Specific Types of Chemical Structures



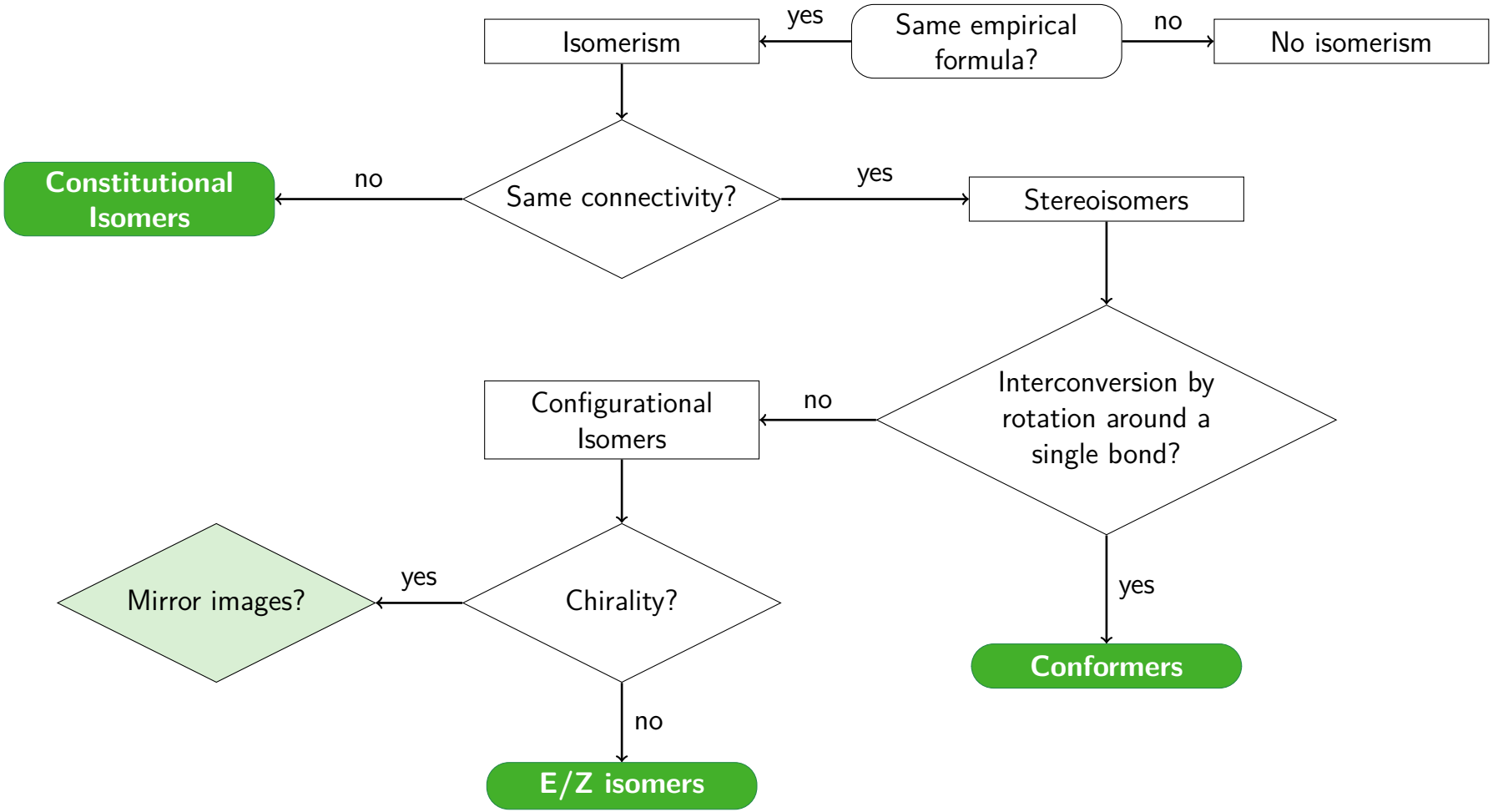
3D Structure – Isomerism

Specific Types of Chemical Structures



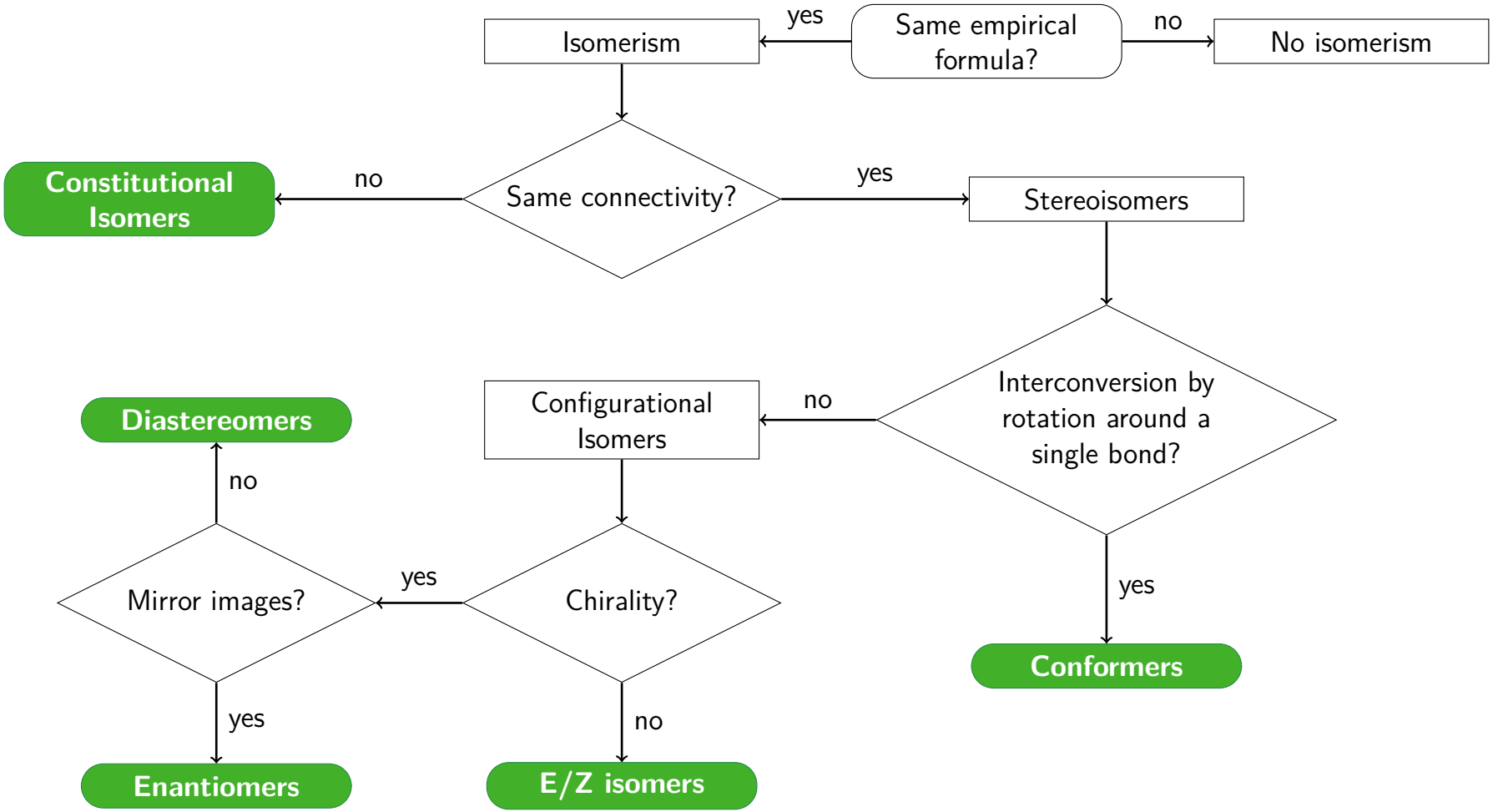
3D Structure – Isomerism

Specific Types of Chemical Structures



3D Structure – Isomerism

Specific Types of Chemical Structures



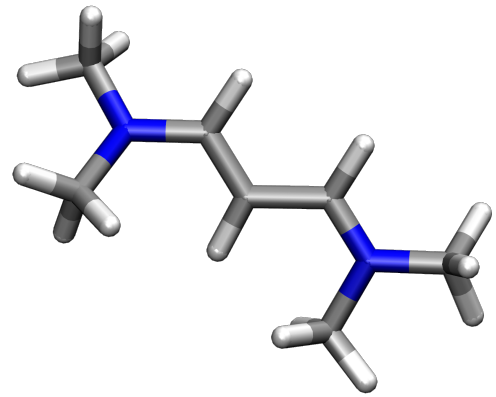
3D Structure – Isomerism

Representation of 3D structures

***Easy* solution:**

XYZ File

```
1  24 # total number of atoms
2  charge=+1 # empty line or description
3    C     3.673812   0.501242  -0.009871
4    H     3.811112  -0.578258  -0.015371
5    C     1.342812  -0.120158   0.004929
6    H     1.726112  -1.136758   0.002529
7    C    -0.030587   0.072442   0.006629
8    H    -0.450588   1.068142   0.004529
9  ...
10   N     2.257812   0.818942   0.004929
11   H     4.159212   0.918142   0.874829
12   H     4.141513   0.924542  -0.901171
```



Disadvantages:

- **constitution**al isomers: implicit

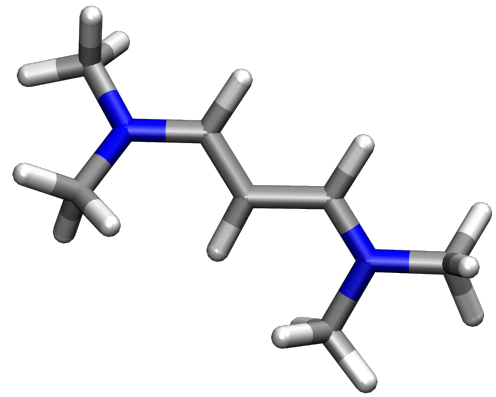
3D Structure – Isomerism

Representation of 3D structures

***Easy* solution:**

XYZ File

```
24 # total number of atoms
charge=+1 # empty line or description
  C     3.673812   0.501242  -0.009871
  H     3.811112  -0.578258  -0.015371
  C     1.342812  -0.120158   0.004929
  H     1.726112  -1.136758   0.002529
  C    -0.030587   0.072442   0.006629
  H    -0.450588   1.068142   0.004529
...
  N     2.257812   0.818942   0.004929
  H     4.159212   0.918142   0.874829
  H     4.141513   0.924542  -0.901171
```



Disadvantages:

- **constitution**al isomers: implicit
- **configuration**al isomers: implicit

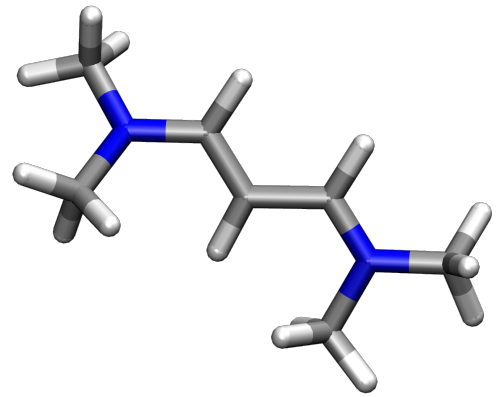
3D Structure – Isomerism

Representation of 3D structures

***Easy* solution:**

XYZ File

```
1  24 # total number of atoms
2  charge=+1 # empty line or description
3    C     3.673812   0.501242  -0.009871
4    H     3.811112  -0.578258  -0.015371
5    C     1.342812  -0.120158   0.004929
6    H     1.726112  -1.136758   0.002529
7    C    -0.030587   0.072442   0.006629
8    H    -0.450588   1.068142   0.004529
9  ...
10   N     2.257812   0.818942   0.004929
11   H     4.159212   0.918142   0.874829
12   H     4.141513   0.924542  -0.901171
```



Disadvantages:

- **constitution**al isomers: implicit
- **configuration**al isomers: implicit
- **not invariant** w.r.t. permutations, translations, and rotations

Isomerism

Representation of 3D structures

Type	File Format	Constitution	Configuration	3D structure
Chemical Notation	**Empirical Formula**	No	No	No
	SMILES			
	InChI			
Math. Notation	**XYZ File**	Implicitly	Implicitly	Yes
	Molecular Graphs			
	MDL Molfile			

Recap SMILES

Simplified Molecular Input Line Entry System

SMILES

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) is a linear representation of the atoms and bonds of a molecule.

- **Acyclic** molecules:
- **Cyclic** molecules:

Recap SMILES

Simplified Molecular Input Line Entry System

SMILES

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) is a linear representation of the atoms and bonds of a molecule.

- **Acyclic** molecules:
  - atoms are represented by their chemical symbols
    - *e.g.*, `B, C, N, O, S, P, Cl`
  - Hydrogens are implicit
- **Cyclic** molecules:

Recap SMILES

Simplified Molecular Input Line Entry System

SMILES

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) is a linear representation of the atoms and bonds of a molecule.

- **Acyclic** molecules:
  - atoms are represented by their chemical symbols
    - *e.g.*, `B, C, N, O, S, P, Cl`
  - Hydrogens are implicit
  - adjacent atoms are connected by a bond
  - special symbols can be used to indicate the order of the bond
    - `-, =, #, :, .`
- **Cyclic** molecules:

Recap SMILES

Simplified Molecular Input Line Entry System

SMILES

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) is a linear representation of the atoms and bonds of a molecule.

- **Acyclic** molecules:
  - atoms are represented by their chemical symbols
    - *e.g.*, `B, C, N, O, S, P, Cl`
  - Hydrogens are implicit
  - adjacent atoms are connected by a bond
  - special symbols can be used to indicate the order of the bond
    - `-, =, #, :, .`
  - branches can be specified by enclosing them in parentheses
    - *e.g.*, iso-butane: `CC(C)C`
- **Cyclic** molecules:

Recap SMILES

Simplified Molecular Input Line Entry System

SMILES

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) is a linear representation of the atoms and bonds of a molecule.

- **Acyclic** molecules:
  - atoms are represented by their chemical symbols
    - *e.g.*, `B, C, N, O, S, P, Cl`
  - Hydrogens are implicit
  - adjacent atoms are connected by a bond
  - special symbols can be used to indicate the order of the bond
    - `-, =, #, :, .`
  - branches can be specified by enclosing them in parentheses
    - *e.g.*, iso-butane: `CC(C)C`
- **Cyclic** molecules:
  - cycles are *broken* by removing a bond
  - broken bonds are marked by numerical indices
  - the resulting tree is linearized
  - atoms of broken bonds are annotaed with the corresponding indices

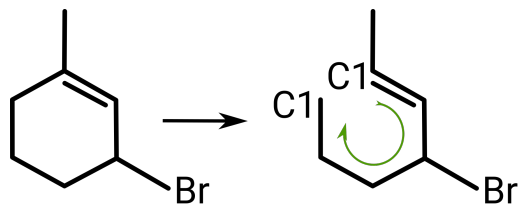
Recap SMILES

Simplified Molecular Input Line Entry System

SMILES

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) is a linear representation of the atoms and bonds of a molecule.

- **Acyclic** molecules:
  - atoms are represented by their chemical symbols
    - *e.g.*, `B, C, N, O, S, P, Cl`
  - Hydrogens are implicit
  - adjacent atoms are connected by a bond
  - special symbols can be used to indicate the order of the bond
    - `-, =, #, :, .`
  - branches can be specified by enclosing them in parentheses
    - *e.g.*, iso-butane: `CC(C)C`
- **Cyclic** molecules:
  - cycles are *broken* by removing a bond
  - broken bonds are marked by numerical indices
  - the resulting tree is linearized
  - atoms of broken bonds are annotaed with the corresponding indices



`CC1=CC(Br)CCC1`

Isomeric SMILES

Encoding isotopism, E/Z isomers and chirality

Isomeric SMILES

SMILES written with isotopic, configurational (double bonds), and chiral specifications are collectively known as **isomeric SMILES**, allowing a complete and rigorous partial specification of chirality.

Isomeric SMILES

Encoding isotopism, E/Z isomers and chirality

Isomeric SMILES

SMILES written with isotopic, configurational (double bonds), and chiral specifications are collectively known as **isomeric SMILES**, allowing a complete and rigorous partial specification of chirality.

Isotopic Specification

- preceding atomic symbol with integral atomic mass

`[12C]`	carbon-12
`[13CH4]`	C-13-methane

Isomeric SMILES

Encoding isotopism, E/Z isomers and chirality

Isomeric SMILES

SMILES written with isotopic, configurational (double bonds), and chiral specifications are collectively known as **isomeric SMILES**, allowing a complete and rigorous partial specification of chirality.

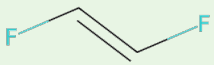
Isotopic Specification

- preceding atomic symbol with integral atomic mass

`[12C]`	carbon-12
`[13CH4]`	C-13-methane

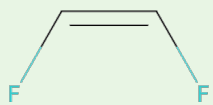
E/Z isomers

- symbols `/` and `\` indicate relative directionality of a bond



`F/C=C/F`

`F\C=C\F`



`F/C=C\F`

`F\C=C/F`

Isomeric SMILES

Encoding isotopism, E/Z isomers and chirality

Isomeric SMILES

SMILES written with isotopic, configurational (double bonds), and chiral specifications are collectively known as **isomeric SMILES**, allowing a complete and rigorous partial specification of chirality.

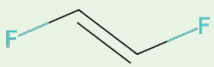
Isotopic Specification

- preceding atomic symbol with integral atomic mass

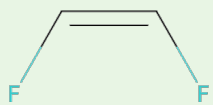
`[12C]`	carbon-12
`[13CH4]`	C-13-methane

E/Z isomers

- symbols `/` and `\` indicate relative directionality of a bond



`F/C=C/F`
`F\C=C\F`



`F/C=C\F`
`F\C=C/F`

Configuration Around Tetrahedral Centers

- orientations are based on the order in which neighbors occur in the SMILES string
- tetrahedral centers may be indicated by a simplified chiral specification (`@` or `@@`)

Isomeric SMILES

Encoding isotopism, E/Z isomers and chirality

Isomeric SMILES

SMILES written with isotopic, configurational (double bonds), and chiral specifications are collectively known as **isomeric SMILES**, allowing a complete and rigorous partial specification of chirality.

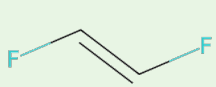
Isotopic Specification

- preceding atomic symbol with integral atomic mass

`[12C]`	carbon-12
`[13CH4]`	C-13-methane

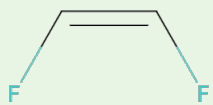
E/Z isomers

- symbols `/` and `\` indicate relative directionality of a bond



`F/C=C/F`

`F\C=C\F`

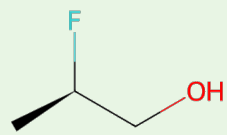


`F/C=C\F`

`F\C=C/F`

Configuration Around Tetrahedral Centers

- orientations are based on the order in which neighbors occur in the SMILES string
- tetrahedral centers may be indicated by a simplified chiral specification (`@` or `@@`)
  - `@`: neighbors are listed anti-clockwise (*S*)



`[C@H](C)(F)CO`

Isomeric SMILES

Encoding isotopism, E/Z isomers and chirality

Isomeric SMILES

SMILES written with isotopic, configurational (double bonds), and chiral specifications are collectively known as **isomeric SMILES**, allowing a complete and rigorous partial specification of chirality.

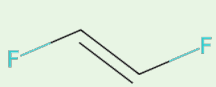
Isotopic Specification

- preceding atomic symbol with integral atomic mass

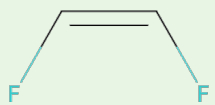
`[12C]`	carbon-12
`[13CH4]`	C-13-methane

E/Z isomers

- symbols `/` and `\` indicate relative directionality of a bond



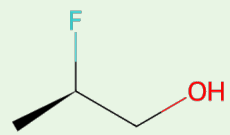
`F/C=C/F`
`F\C=C\F`



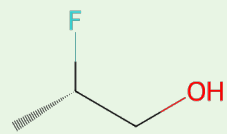
`F/C=C\F`
`F\C=C/F`

Configuration Around Tetrahedral Centers

- orientations are based on the order in which neighbors occur in the SMILES string
- tetrahedral centers may be indicated by a simplified chiral specification (`@` or `@@`)
  - `@`: neighbors are listed anti-clockwise (*S*)
  - `@@`: neighbors are listed clockwise (anti-anti-clockwise, *R*)



`[C@H](C)(F)CO`



`[C@@H](C)(F)CO`

Isomeric SMILES

Encoding isotopism, E/Z isomers and chirality

Isomeric SMILES

SMILES written with isotopic, configurational (double bonds), and chiral specifications are collectively known as **isomeric SMILES**, allowing a complete and rigorous partial specification of chirality.

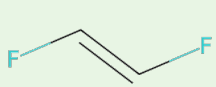
Isotopic Specification

- preceding atomic symbol with integral atomic mass

`[12C]`	carbon-12
`[13CH4]`	C-13-methane

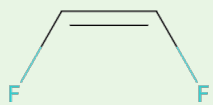
E/Z isomers

- symbols `/` and `\` indicate relative directionality of a bond



`F/C=C/F`

`F\C=C\F`

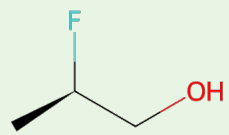


`F/C=C\F`

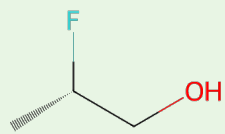
`F\C=C/F`

Configuration Around Tetrahedral Centers

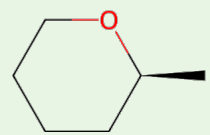
- orientations are based on the order in which neighbors occur in the SMILES string
- tetrahedral centers may be indicated by a simplified chiral specification (`@` or `@@`)
  - `@`: neighbors are listed anti-clockwise (*S*)
  - `@@`: neighbors are listed clockwise (anti-anti-clockwise, *R*)
- ring closure bond's chiral order is implied by the position of the ring closure digit on the chiral atom in lexical order



`[C@H](C)(F)CO`



`[C@@H](C)(F)CO`

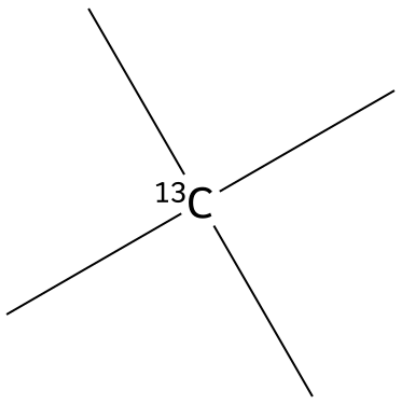


`C[C@H]1CCCCO1`

Isomeric SMILES (Hands On)

Encoding isotopism, E/Z isomers and chirality

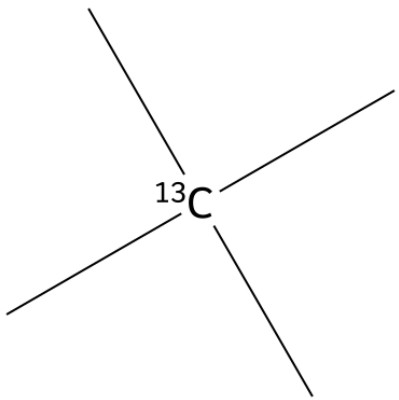
What are the isomeric SMILES of the following compounds?



Isomeric SMILES (Hands On)

Encoding isotopism, E/Z isomers and chirality

What are the isomeric SMILES of the following compounds?

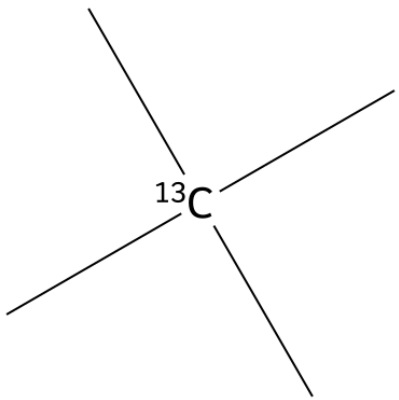


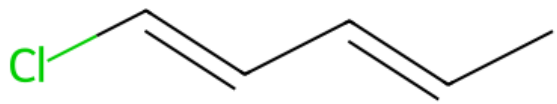

```
C[13C](C)(C)C
```

Isomeric SMILES (Hands On)

Encoding isotopism, E/Z isomers and chirality

What are the isomeric SMILES of the following compounds?



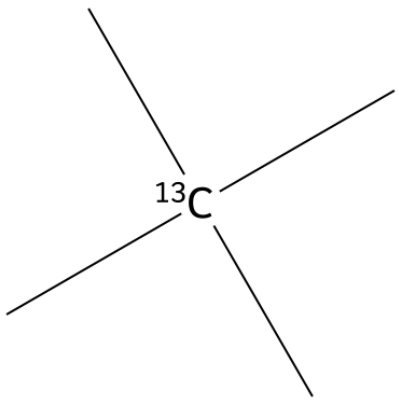


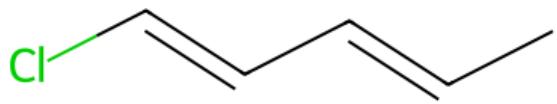
`C[13C](C)(C)C`

Isomeric SMILES (Hands On)

Encoding isotopism, E/Z isomers and chirality

What are the isomeric SMILES of the following compounds?





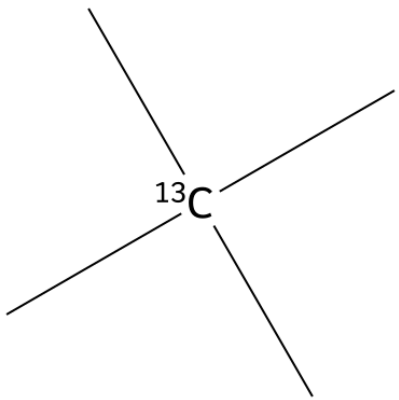
`C[13C](C)(C)C`

`Cl/C=C/C=C/C`

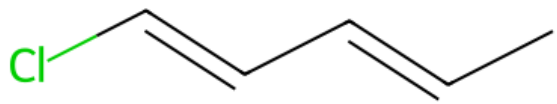
Isomeric SMILES (Hands On)

Encoding isotopism, E/Z isomers and chirality

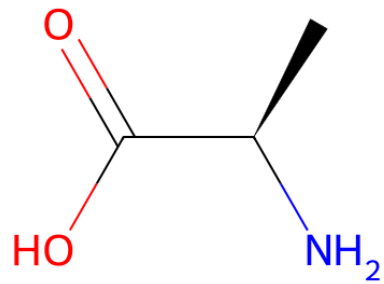
What are the isomeric SMILES of the following compounds?



`C[13C](C)(C)C`



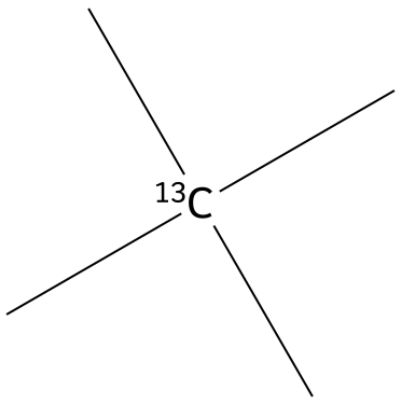
`Cl/C=C/C=C/C`



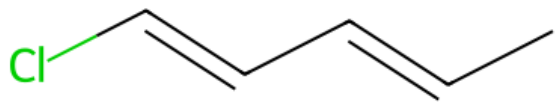
Isomeric SMILES (Hands On)

Encoding isotopism, E/Z isomers and chirality

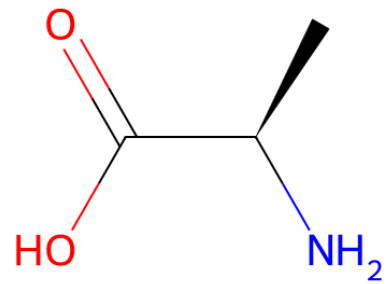
What are the isomeric SMILES of the following compounds?



`C[13C](C)(C)C`



`Cl/C=C/C=C/C`



`N[C@H](C)C(=O)O`

InChI

The International Chemical Identifier

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

InChI

The International Chemical Identifier

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Composition of InChI (**`InChI=1S/`**) in six hierarchical sublayers:

1. Main layer
   - chemical formula
   - connection sublayer **`/c`**
   - H-atom sublayer **`/h`**
2. Charge layer
   - charge sublayer **`/q`**
   - proton sublayer **`/p`**
3. Stereochemical layer
   - double bond sublayer **`/b`**
   - tetrahedral stereochemistry sublayers **`/t, /m, /s`**
4. Isotopic layer
   - **`/i, /h, /b, /t, /m, /s`** sublayer
5. Fixed-H layer (only nonstandard InChI)
6. Reconnected layers for metals (only nonstandard InChI)

InChI

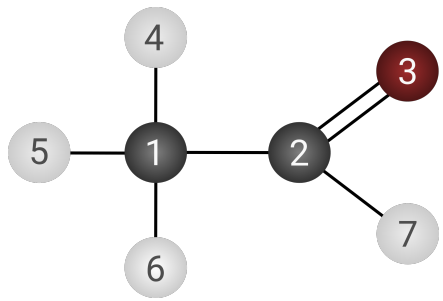
The International Chemical Identifier

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

1. Main layer
   - chemical formula
   - connection sublayer **`/c`**
   - H-atom sublayer **`/h`**



`InChI = 1S`

Main layer

InChI

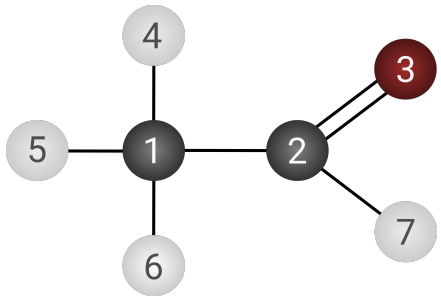
The International Chemical Identifier

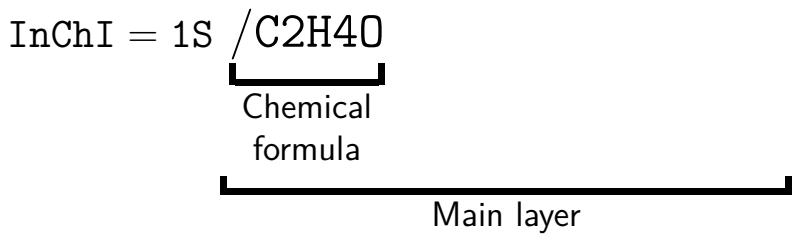
InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

1. Main layer
   - chemical formula
   - connection sublayer **`/c`**
   - H-atom sublayer **`/h`**





InChI

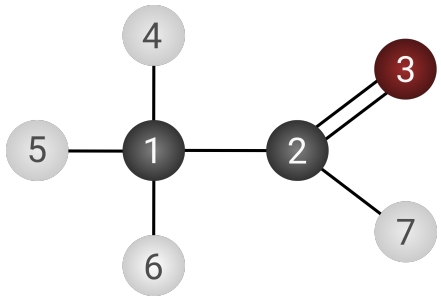
The International Chemical Identifier

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**InChI=1S/**):

1. Main layer
   - chemical formula
   - connection sublayer **/c**
   - H-atom sublayer **/h**



$$\texttt{InChI} = \texttt{1S}\ \underbrace{\underbrace{/\texttt{C2H4O}}_{\text{Chemical formula}}\ \underbrace{/\texttt{c1-2-3}}_{\text{Connection sublayer}}}_{\text{Main layer}}$$

InChI

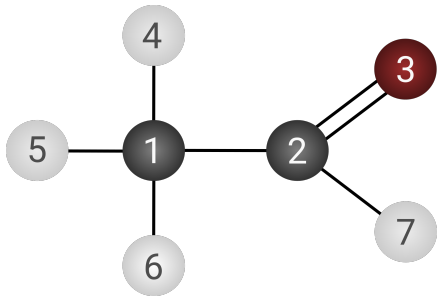
The International Chemical Identifier

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

1. Main layer
   - chemical formula
   - connection sublayer **`/c`**
   - H-atom sublayer **`/h`**



$$\texttt{InChI} = \texttt{1S}\ \underbrace{\underbrace{\texttt{/C2H4O}}_{\text{Chemical formula}}\ \underbrace{\texttt{/c1-2-3}}_{\text{Connection sublayer}}\ \underbrace{\texttt{/h2H,1H3}}_{\text{H atom sublayer}}}_{\text{Main layer}}$$

InChI

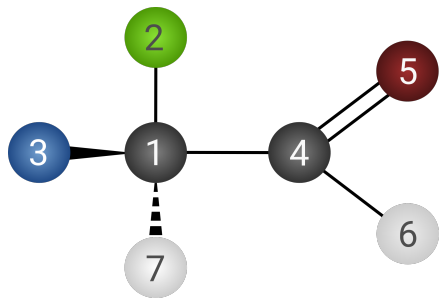
Encoding of 3D information

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

3. Stereochemical layer
   - double bond sublayer **`/b`**
   - tetrahedral stereochemistry sublayers **`/t, /m, /s`**



`InChI = 1S`

Main layer | Stereochemical layer

InChI

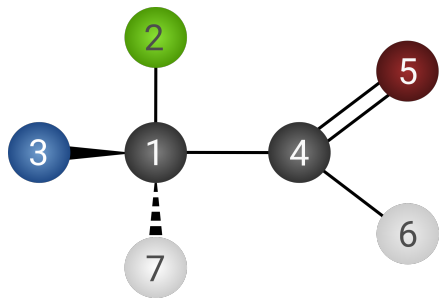
Encoding of 3D information

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

3. Stereochemical layer
   - double bond sublayer **`/b`**
   - tetrahedral stereochemistry sublayers **`/t, /m, /s`**



`InChI = 1S /C2H2ClFO`

Chemical formula

Main layer

Stereochemical layer

InChI

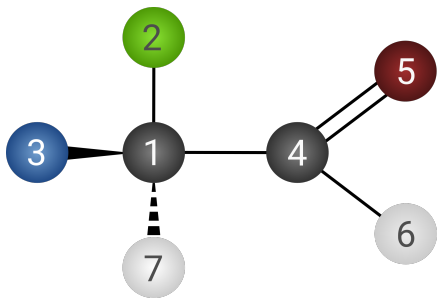
Encoding of 3D information

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

3. Stereochemical layer
   - double bond sublayer **`/b`**
   - tetrahedral stereochemistry sublayers **`/t, /m, /s`**



`InChI = 1S` `/C2H2ClFO` `/c3-2(4)1-5`

`/C2H2ClFO`: Chemical formula

`/c3-2(4)1-5`: Connection sublayer

`InChI = 1S /C2H2ClFO /c3-2(4)1-5`: Main layer

Stereochemical layer

InChI

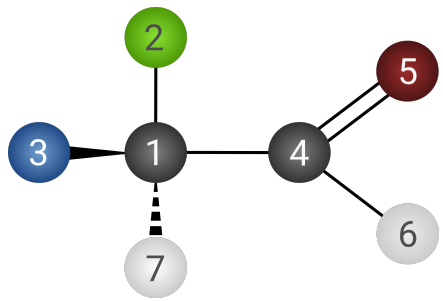
Encoding of 3D information

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

3. Stereochemical layer
   - double bond sublayer **`/b`**
   - tetrahedral stereochemistry sublayers **`/t, /m, /s`**



$$\texttt{InChI} = \texttt{1S}\ \underbrace{\underbrace{\texttt{/C2H2ClFO}}_{\text{Chemical formula}}\ \underbrace{\texttt{/c3-2(4)1-5}}_{\text{Connection sublayer}}\ \underbrace{\texttt{/h1-2H}}_{\text{H atom sublayer}}}_{\text{Main layer}}\ \underbrace{\phantom{\texttt{/h1-2H}}}_{\text{Stereochemical layer}}$$

InChI

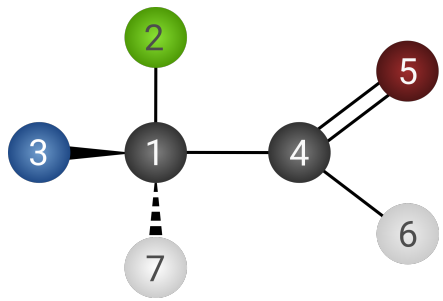
Encoding of 3D information

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

3. Stereochemical layer
   - double bond sublayer **`/b`**
   - tetrahedral stereochemistry sublayers **`/t, /m, /s`**



`InChI = 1S` `/C2H2ClFO` `/c3-2(4)1-5` `/h1-2H` `/t1-`

- `/C2H2ClFO`: Chemical formula
- `/c3-2(4)1-5`: Connection sublayer
- `/h1-2H`: H atom sublayer
- Main layer: `/C2H2ClFO /c3-2(4)1-5 /h1-2H`
- Stereochemical layer: `/t1-`

InChI

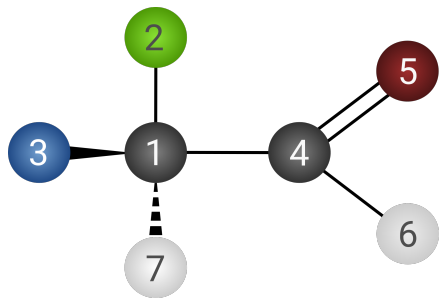
Encoding of 3D information

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

3. Stereochemical layer
   - double bond sublayer **`/b`**
   - tetrahedral stereochemistry sublayers **`/t, /m, /s`**



`InChI = 1S /C2H2ClFO /c3-2(4)1-5 /h1-2H /t1- /m1`

- `/C2H2ClFO`: Chemical formula
- `/c3-2(4)1-5`: Connection sublayer
- `/h1-2H`: H atom sublayer
- `/t1-`: *R*
- `/m1`: *R*

Main layer: `/C2H2ClFO /c3-2(4)1-5 /h1-2H`

Stereochemical layer: `/t1- /m1`

InChI

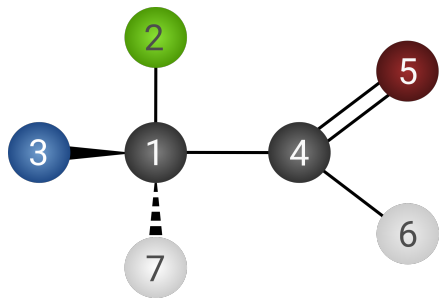
Encoding of 3D information

InChI

The IUPAC **In**ternational **Ch**emical **I**dentifier is a nonproprietary biunique identifier, namely a alphanumeric string, for chemical compounds released in 2000.

Sublayers in InChI (**`InChI=1S/`**):

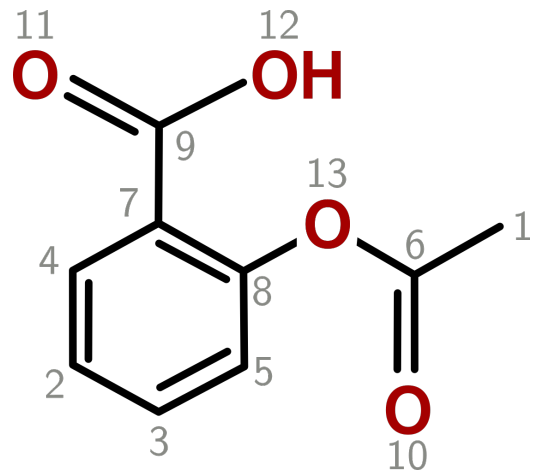
3. Stereochemical layer
   - double bond sublayer **`/b`**
   - tetrahedral stereochemistry sublayers **`/t, /m, /s`**



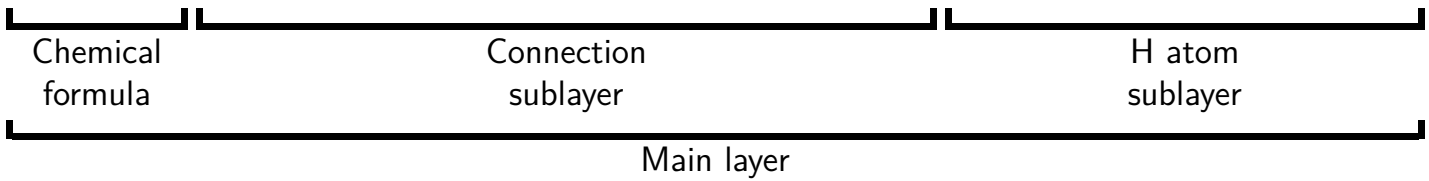
$$\texttt{InChI} = \texttt{1S}\ \underbrace{\underbrace{\texttt{/C2H2ClFO}}_{\text{Chemical formula}}\ \underbrace{\texttt{/c3-2(4)1-5}}_{\text{Connection sublayer}}\ \underbrace{\texttt{/h1-2H}}_{\text{H atom sublayer}}}_{\text{Main layer}}\ \texttt{/t1-}\ \underbrace{\underbrace{\texttt{/m1}}_{R}\ \texttt{/s1}}_{\text{Stereochemical layer}}$$

InChI (Hands On)

Encoding of 3D information

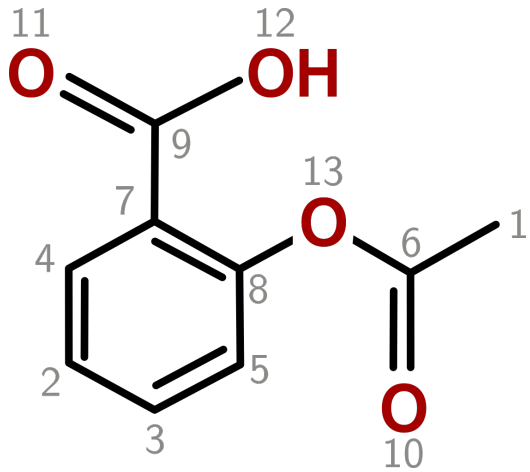


InChI = 1S



InChI (Hands On)

Encoding of 3D information



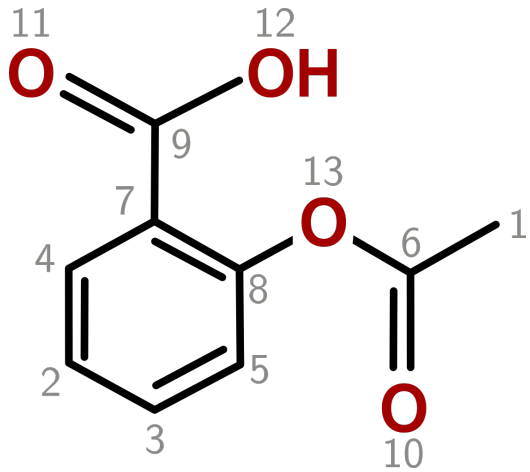
InChI = 1S /C9H8O4

Chemical formula | Connection sublayer | H atom sublayer

Main layer

InChI (Hands On)

Encoding of 3D information



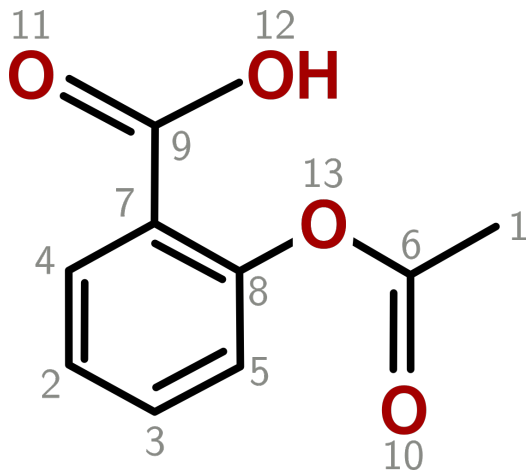
$$\texttt{InChI} = \texttt{1S}\,/\texttt{C9H8O4}\,/\texttt{c1-6(10)13-8-5-3-2-4-7(8)9(11)12}$$

Chemical formula | Connection sublayer | H atom sublayer

Main layer

InChI (Hands On)

Encoding of 3D information



$$\texttt{InChI} = \texttt{1S}\,\underbrace{/\texttt{C9H8O4}}_{\text{Chemical formula}}\,\underbrace{/\texttt{c1-6(10)13-8-5-3-2-4-7(8)9(11)12}}_{\text{Connection sublayer}}\,\underbrace{/\texttt{h2-5H,1H3,(H,11,12)}}_{\text{H atom sublayer}}$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{Main layer}}$$

Isomerism

Representation of 3D structures

Type	File Format	Constitution	Configuration	3D structure
Chemical Notation	**Empirical Formula**	No	No	No
	SMILES	Yes	Yes	No
	InChI	Yes	Yes	No
Math. Notation	**XYZ File**	Implicitly	Implicitly	Yes
	Molecular Graphs			
	MDL Molfile			

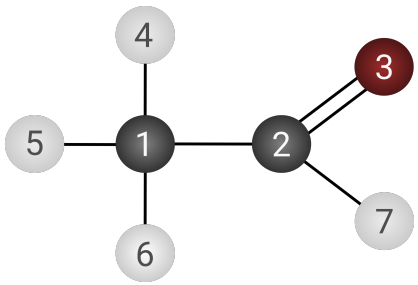
Molecular Graphs

Encoding of 3D information

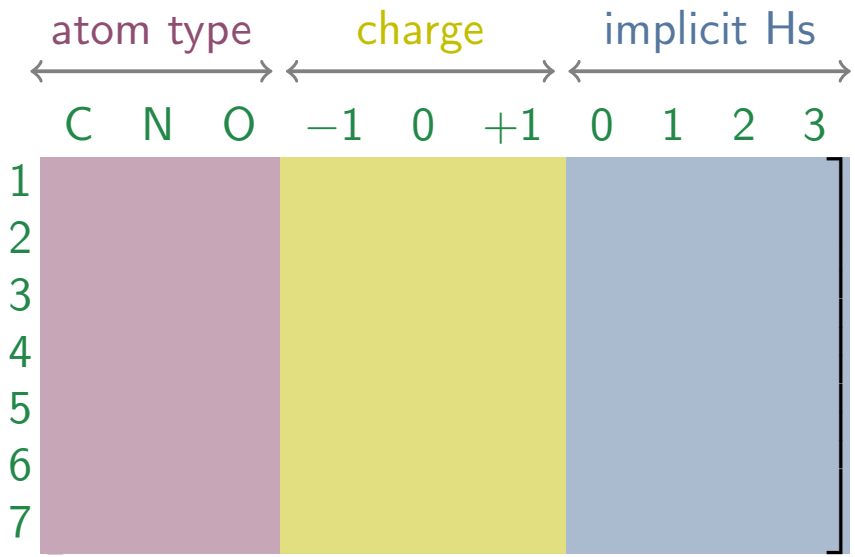
Graphs

Graphs are formally 2D data structures with no spatial relationships between elements. Nonetheless, 3D information (and information that is the result of a 3D structure e.g. stereochemistry) can be encoded into a graph representation.

- node features matrix (X) for node information (such as if a chiral node is *R* or *S*)
- edge features matrix (E) for edge information (such as the length of a bond)



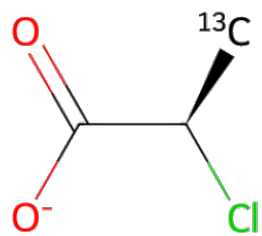
$$X =$$



MDL Molfiles

Format of the 3D mathematical notation

	1	
Header block	2	RDKit
	3	
Counts line	4	6 5 0 0 1 0 0 0 0 0999 V2000
	5	2.5981 1.5000 0.0000 C 1 0 0 0 0 0 0 0 0 0 0 0
	6	1.2990 0.7500 0.0000 C 0 0 1 0 0 0 0 0 0 0 0 0
	7	0.0000 0.0000 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
Atom block	8	-1.2990 0.7500 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
	9	-0.0000 -1.5000 0.0000 O 0 5 0 0 0 0 0 0 0 0 0 0
	10	2.0490 -0.5490 0.0000 Cl 0 0 0 0 0 0 0 0 0 0 0 0
	11	0.5490 2.0490 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0
	12	1 2 1 1
	13	2 3 1 0
Bond block	14	3 4 2 0
	15	3 5 1 0
	16	2 6 1 0
	17	2 7 1 6
Properties	18	M CHG 1 5 -1
block	19	M ISO 1 1 13
	20	M END



MDL Molfiles

Counts Line

```
                1
Header block    2  RDKit
                3
Counts line     4  6  5  0  0  1  0  0  0  0  0999 V2000
                5   2.5981  1.5000  0.0000 C  1  0  0  0  0  0  0  0  0  0  0  0
                6   1.2990  0.7500  0.0000 C  0  0  1  0  0  0  0  0  0  0  0  0
Ato
Bon
Pro
blo
               20  M  END
```

# atoms	# bonds	# atom list	obsolete	chiral flag (0=none, 1=odd, 2=even, 3=either)			other properties		# additional properties	Ctab version
6	5	0	0	1	0	0	0	0	0999	V2000

MDL Molfiles

Atom block

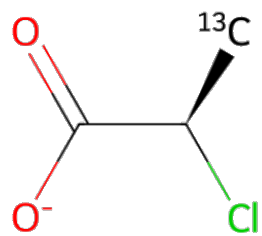
```
                1
Header block    2  RDKit
                3
Counts line     4  6  5  0  0  1  0  0  0  0  0999 V2000
                5   2.5981  1.5000  0.0000 C  1  0  0  0  0  0  0  0  0  0  0  0
                6   1.2990  0.7500  0.0000 C  0  0  1  0  0  0  0  0  0  0  0  0
                7   0.0000  0.0000  0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
Atom block      8  -1.2990  0.7500  0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
                9  -0.0000 -1.5000  0.0000 O  0  0  5  0  0  0  0  0  0  0  0  0
               10   2.0490 -0.5490  0.0000 Cl 0  0  0  0  0  0  0  0  0  0  0  0
               11   0.5490  2.0490  0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
Bon
Pro
blo
```

- atom coordinates
- atom symbol
- mass difference (-3, -2, -1, 0, 1, 2, 3, 4 or 0 if beyond limits)
- charge (0=0, 1=+3, 2=+2, 3=+1, 4=radical, 5=-1, 6=-2, 7=-3)
- atom stereo parity (0=not stereo, 1=odd, 2=even, 3=either)

MDL Molfiles

Bond block

```
                1
                2  RDKit
                3
Counts line     4  6  5  0  0  1  0  0  0  0  0999 V2000
                5   2.5981  1.5000  0.0000 C  1  0  0  0  0  0  0  0  0  0  0  0
                6   1.2990  0.7500  0.0000 C  0  0  1  0  0  0  0  0  0  0  0  0
                7   0.0000  0.0000  0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
Atom block      8  -1.2990  0.7500  0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
                9  -0.0000 -1.5000  0.0000 O  0  5  0  0  0  0  0  0  0  0  0  0
               10   2.0490 -0.5490  0.0000 Cl 0  0  0  0  0  0  0  0  0  0  0  0
               11   0.5490  2.0490  0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
               12  1  2  1  1
               13  2  3  1  0
Bond block     14  3  4  2  0
               15  3  5  1  0
               16  2  6  1  0
               17  2  7  1  6
Properties     18  M  CHG  1   5  -1
block          19  M  ISO  1   1  13
               20  M  END
```



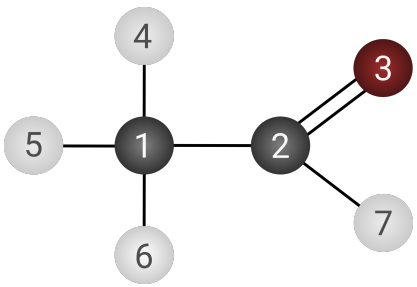
Connection Table

Bond block in Molfiles

Connection table (CT)

Concise molecular representation in form of a table derived from a graph:

1. Atom list: atom names and indices
2. Bond list: connections between all atoms using indices, bond order as 3^{rd} column
3. Merge atom and bond list in a single table, namely the CT



Atom list	
1	C
2	C
3	O
4	H
5	H
6	H
7	H

Bond list		
1	2	1
1	4	1
1	5	1
1	6	1
2	3	2
2	7	1

CT			
1	C	2	1
2	C	1	1
3	O		
4	H		
5	H		
6	H		
7	H		

MDL Molfiles

Bond block

```
                1
Header block    2  RDKit
                3
Counts line     4  6  5  0  0  1  0  0  0
                5   2.5981  1.5000  0.
                6   1.2990  0.7500  0.
                7   0.0000  0.0000  0.
Atom block      8  -1.2990  0.7500  0.
                9  -0.0000 -1.5000  0.
               10   2.0490 -0.5490  0.
               11   0.5490  2.0490  0.
               12  1  2  1  1
               13  2  3  1  0
Bond block     14  3  4  2  0
               15  3  5  1  0
               16  2  6  1  0
               17  2  7  1  6
Properties     18  M  CHG  1   5  -1
block          19  M  ISO  1   1  13
               20  M  END
```

- 1^{st} atom row number
- 2^{nd} atom row number
- bond type

  (1 = Single, 2 = Double, 3 = Triple,
  4 = Aromatic, 5 = Single or Double,
  6 = Single or Aromatic,
  7 = Double or Aromatic, 8 = Any)
- bond stereo

  (single bonds: 0 = not stereo, 1 = Up,
  4 = Either, 6 = Down;
  Double bonds: 0 = use xyz coords,
  3 = *cis* or *trans*)

MDL Molfiles

Properties block

Header block

Cou

Ato

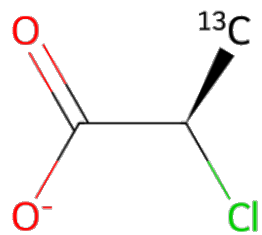
	property	count	atom	value
M	CHG	1	5	−1
M	ISO	1	1	13
M	END			

0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0

Bond block

Properties block

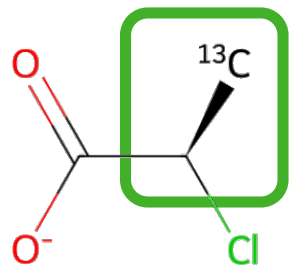
```
 1
 2  RDKit
13  2  3  1  0
14  3  4  2  0
15  3  5  1  0
16  2  6  1  0
17  2  7  1  6
18  M  CHG  1   5  -1
19  M  ISO  1   1  13
20  M  END
```



MDL Molfiles

Encoding of 3D structure and information

```
                1
                2  RDKit
                3
Counts line     4  6  5  0  0  1  0  0  0  0  0999 V2000
                5   2.5981  1.5000  0.0000 C  1  0  0  0  0  0  0  0  0  0  0  0
                6   1.2990  0.7500  0.0000 C  0  0  1  0  0  0  0  0  0  0  0  0
                7   0.0000  0.0000  0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
Atom block      8  -1.2990  0.7500  0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
                9  -0.0000 -1.5000  0.0000 O  0  5  0  0  0  0  0  0  0  0  0  0
               10   2.0490 -0.5490  0.0000 Cl 0  0  0  0  0  0  0  0  0  0  0  0
               11   0.5490  2.0490  0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
               12  1  2  1  1
               13  2  3  1  0
Bond block     14  3  4  2  0
               15  3  5  1  0
               16  2  6  1  0
               17  2  7  1  6
Properties     18  M  CHG  1  5  -1
block          19  M  ISO  1  1  13
               20  M  END
```



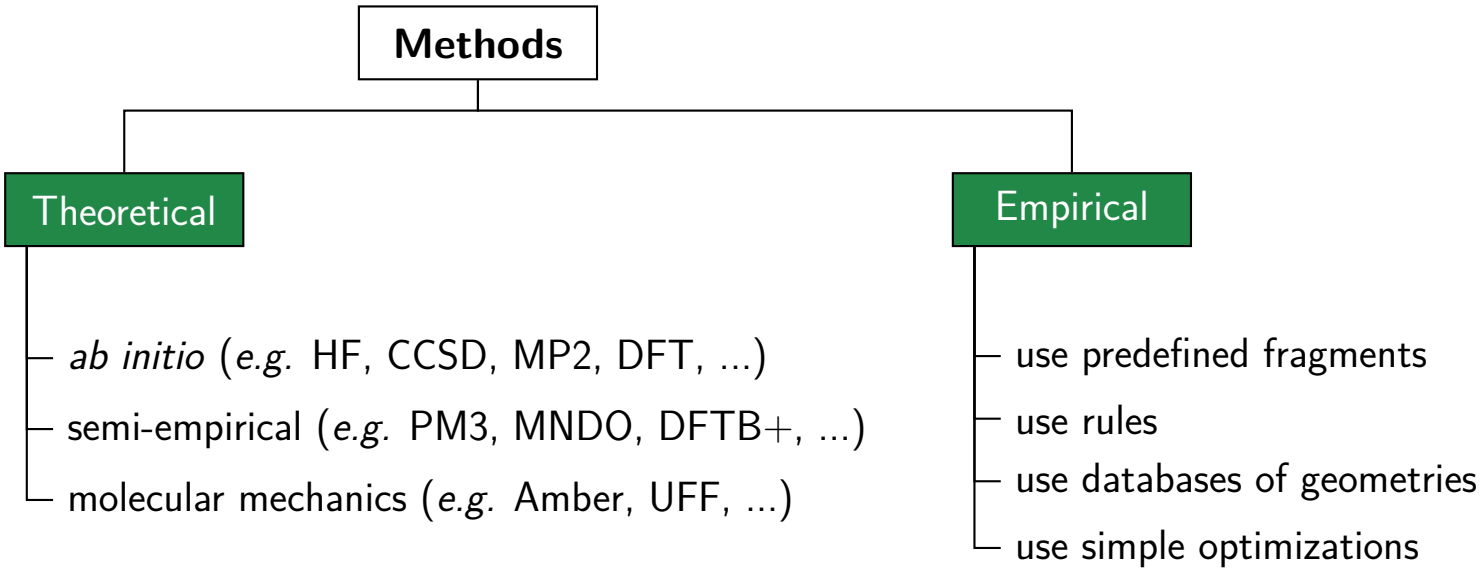
Isomerism

Representation of 3D structures

Type	File Format	Constitution	Configuration	3D structure
Chemical Notation	**Empirical Formula**	No	No	No
	SMILES	Yes	Yes	No
	InChI	Yes	Yes	No
Math. Notation	**XYZ File**	Implicitly	Implicitly	Yes
	Molecular Graphs	Yes	Implicitly	No
	MDL Molfile	Yes	Yes	Yes

Generation of 3D structures

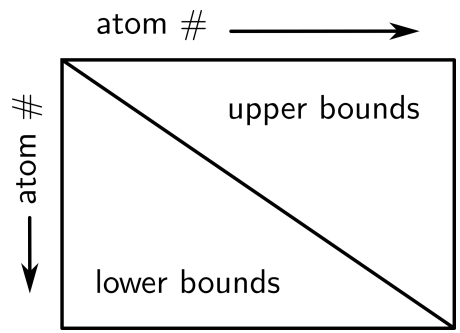
Overview



DG Algorithm

1. Initialization of bounds matrix

- Distance Geometry (DG) algorithm

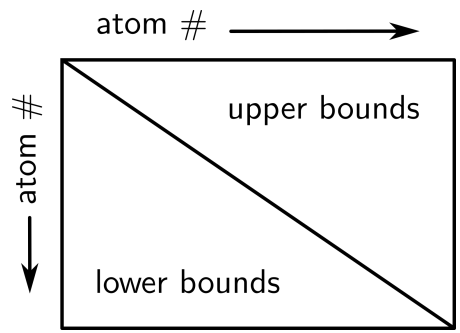


DG Algorithm

1. Initialization of bounds matrix

3. Setting of topological bounds (1,2-, 1,3-, 1,4- and 1,5-distances)

- Distance Geometry (DG) algorithm

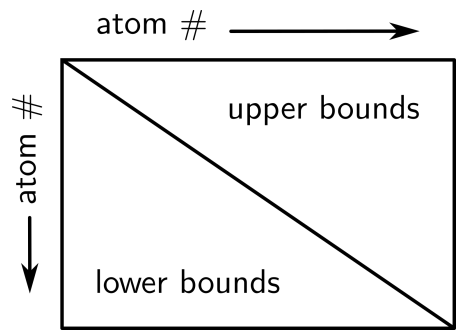


DG Algorithm

1. Initialization of bounds matrix

3. Setting of topological bounds (1,2-, 1,3-, 1,4- and 1,5-distances)
4. Triangle-inequality smoothing of bounds

- Distance Geometry (DG) algorithm

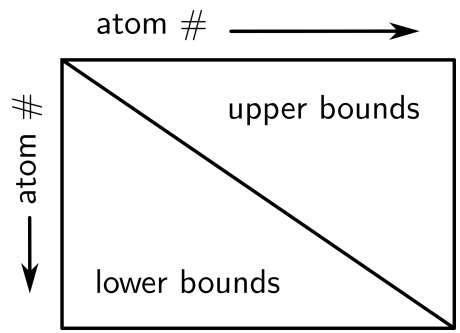


DG Algorithm

1. Initialization of bounds matrix

3. Setting of topological bounds (1,2-, 1,3-, 1,4- and 1,5-distances)
4. Triangle-inequality smoothing of bounds
5. Generating a random distance matrix that satisfies bounds matrix

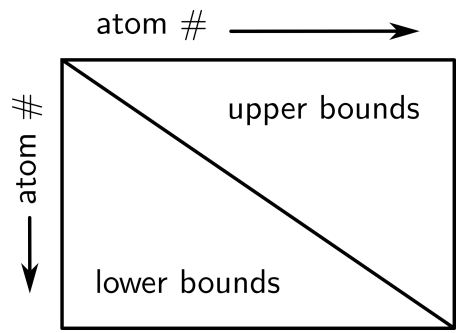
- Distance Geometry (DG) algorithm



DG Algorithm

1. Initialization of bounds matrix
3. Setting of topological bounds (1,2-, 1,3-, 1,4- and 1,5-distances)
4. Triangle-inequality smoothing of bounds
5. Generating a random distance matrix that satisfies bounds matrix
6. Embedding (produce coordinates for each atom)

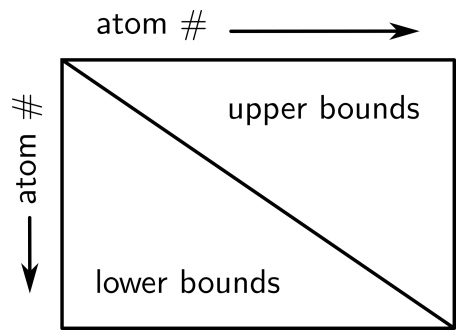
- Distance Geometry (DG) algorithm



DG Algorithm

1. Initialization of bounds matrix
3. Setting of topological bounds (1,2-, 1,3-, 1,4- and 1,5-distances)
4. Triangle-inequality smoothing of bounds
5. Generating a random distance matrix that satisfies bounds matrix
6. Embedding (produce coordinates for each atom)
7. Coordinate refinement by minimizing distance error-function (*e.g. via* a force field)

- Distance Geometry (DG) algorithm

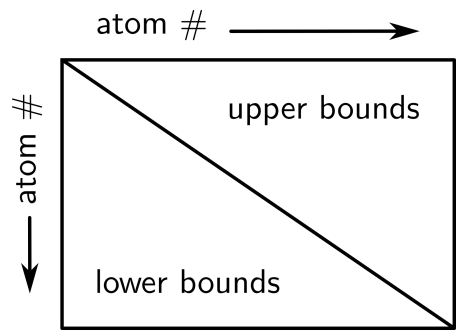


Generation of 3D structures

ETDG Algorithm

1. Initialization of bounds matrix
2. Substructure search for bonds matching torsional patterns, Storage of found torsions in a list
3. Setting of topological bounds (1,2-, 1,3-, 1,4- and 1,5-distances)
4. Triangle-inequality smoothing of bounds
5. Generating a random distance matrix that satisfies bounds matrix
6. Embedding (produce coordinates for each atom)
7. Coordinate refinement by minimizing distance error-function (*e.g. via* a force field)

- Distance Geometry (DG) algorithm
- ETDG algorithm

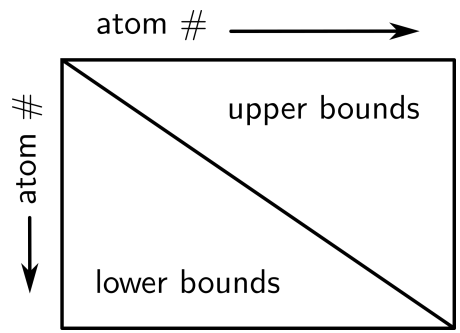


Generation of 3D structures

ETDG Algorithm

1. Initialization of bounds matrix
2. Substructure search for bonds matching torsional patterns, Storage of found torsions in a list
3. Setting of topological bounds (1,2-, 1,3-, 1,4- and 1,5-distances)
4. Triangle-inequality smoothing of bounds
5. Generating a random distance matrix that satisfies bounds matrix
6. Embedding (produce coordinates for each atom)
7. Coordinate refinement by minimizing distance error-function (*e.g. via* a force field)
8. Minimization with torsion potentials, fixed 1,2- and 1,3-distances and bounds for all other atom pairs

- Distance Geometry (DG) algorithm
- ETDG algorithm

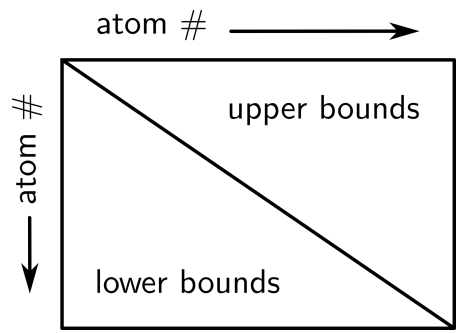


Generation of 3D structures

ETDG Algorithm

1. Initialization of bounds matrix
2. Substructure search for bonds matching torsional patterns, Storage of found torsions in a list
3. Setting of topological bounds (1,2-, 1,3-, 1,4- and 1,5-distances)
4. Triangle-inequality smoothing of bounds
5. Generating a random distance matrix that satisfies bounds matrix
6. Embedding (produce coordinates for each atom)
7. Coordinate refinement by minimizing distance error-function (*e.g. via* a force field)
8. Minimization with torsion potentials, fixed 1,2- and 1,3-distances and bounds for all other atom pairs

- Distance Geometry (DG) algorithm
- ETDG algorithm
- ETKDG algorithm
  - additional knowledge (K) w.r.t. ETDG algorithm in step 8
    - torsional-angle potentials for bonds in aromatic rings
    - constrained angles of 180° for triple bonds
    - UFF inversion terms (N, C, O)



Molecular Force Fields (FFs)

Fast Structure Optimization with FFs

Molecular Force Fields

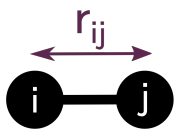
Molecular FFs are mathematical models that describe the potential energy of a molecular system, as:
$V = V_{bond} + V_{angle} + V_{torsion} + V_{coulomb} + V_{vdW}$

Molecular Force Fields (FFs)

Fast Structure Optimization with FFs

Molecular Force Fields

Molecular FFs are mathematical models that describe the potential energy of a molecular system, as:
$V = V_{bond} + V_{angle} + V_{torsion} + V_{coulomb} + V_{vdW}$



Bonded forces

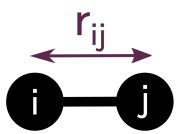
- Bond Stretching
  - $V_{bond} = \sum \frac{1}{2} k_{ij} \left(r_{ij} - r_0\right)^2$

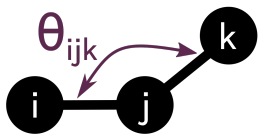
Molecular Force Fields (FFs)

Fast Structure Optimization with FFs

Molecular Force Fields

Molecular FFs are mathematical models that describe the potential energy of a molecular system, as:
$V = V_{bond} + V_{angle} + V_{torsion} + V_{coulomb} + V_{vdW}$





Bonded forces

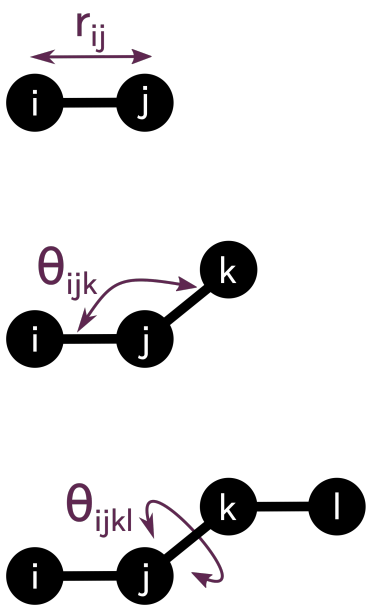
- Bond Stretching
  - $V_{bond} = \sum \frac{1}{2} k_{ij} \left(r_{ij} - r_0\right)^2$
- Angle Bending
  - $V_{angle} = \sum \frac{1}{2} k_{\theta} \left(\theta_{ijk} - \theta_0\right)^2$

Molecular Force Fields (FFs)

Fast Structure Optimization with FFs

Molecular Force Fields

Molecular FFs are mathematical models that describe the potential energy of a molecular system, as:
$V = V_{bond} + V_{angle} + V_{torsion} + V_{coulomb} + V_{vdW}$



Bonded forces

- Bond Stretching
  - $V_{bond} = \sum \frac{1}{2} k_{ij} \left(r_{ij} - r_0\right)^2$
- Angle Bending
  - $V_{angle} = \sum \frac{1}{2} k_{\theta} \left(\theta_{ijk} - \theta_0\right)^2$
- Torsion (Dihedral Angles)
  - $V_{torsion} = \sum \frac{1}{2} V_n \left[1 + \cos\left(n\theta_{ijkl} - \delta\right)\right]$

Molecular Force Fields (FFs)

Fast Structure Optimization with FFs

Molecular Force Fields

Molecular FFs are mathematical models that describe the potential energy of a molecular system, as:
$V = V_{bond} + V_{angle} + V_{torsion} + V_{coulomb} + V_{vdW}$

Non-Bonded forces

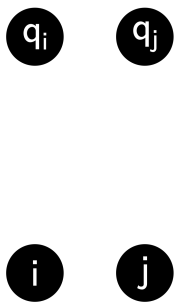
- Coulomb Potential
  - $V_{coulomb} = \sum \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$

Molecular Force Fields (FFs)

Fast Structure Optimization with FFs

Molecular Force Fields

Molecular FFs are mathematical models that describe the potential energy of a molecular system, as:
$V = V_{bond} + V_{angle} + V_{torsion} + V_{coulomb} + V_{vdW}$



Non-Bonded forces

- Coulomb Potential
  - $V_{coulomb} = \sum \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$
- Lennard-Jones Potential
  - $V_{vdW} = \sum 4\pi\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$

Molecular Force Fields (FFs)

Fast Structure Optimization with FFs

Molecular Force Fields

Molecular FFs are mathematical models that describe the potential energy of a molecular system, as:
$V = V_{bond} + V_{angle} + V_{torsion} + V_{coulomb} + V_{vdW}$

Minimization Process:

- *Energy Minimization:*
  - Use force fields to minimize potential energy (V).
  - Algorithms like steepest descent or conjugate gradient are commonly employed.
- *Convergence Criteria:*
  - Ensure energy and gradient tolerances are met.
  - Monitor changes in atomic positions.
- *Validation of Structures:*
  - Assess geometric and energetic stability.
  - Validate against known structural data or experimental results.

RDKit

Generation of ensemble conformations

Usage of ETKDG in RDKit

```
1  from rdkit import Chem
2  from rdkit.Chem import AllChem
3
4  smiles = 'CC(=O)Oc1ccccc1C(=O)O'
5  mol = Chem.MolFromSmiles(smiles)
6
7  # Conformers
8  AllChem.EmbedMultipleConfs(mol, numConfs=12, params=AllChem.ETKDG())
9  results_UFF = AllChem.UFFOptimizeMoleculeConfs(mol, maxIters=10000)
```

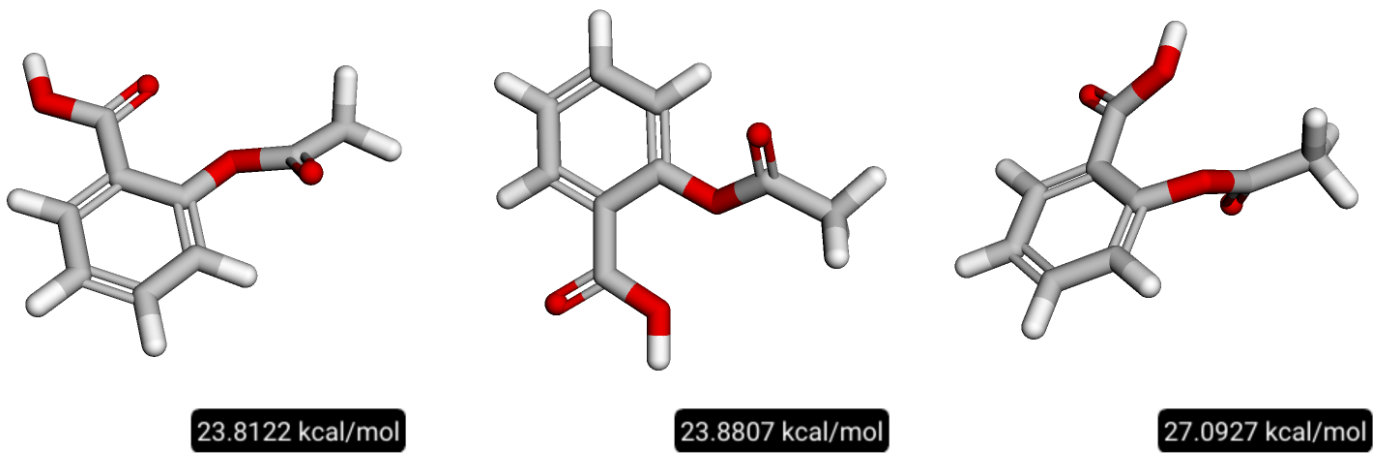
RDKit

Generation of ensemble conformations

Usage of ETKDG in RDKit

```
1  from rdkit import Chem
2  from rdkit.Chem import AllChem
3
4  smiles = 'CC(=O)Oc1ccccc1C(=O)O'
5  mol = Chem.MolFromSmiles(smiles)
6
7  # Conformers
8  AllChem.EmbedMultipleConfs(mol, numConfs=12, params=AllChem.ETKDG())
9  results_UFF = AllChem.UFFOptimizeMoleculeConfs(mol, maxIters=10000)
```

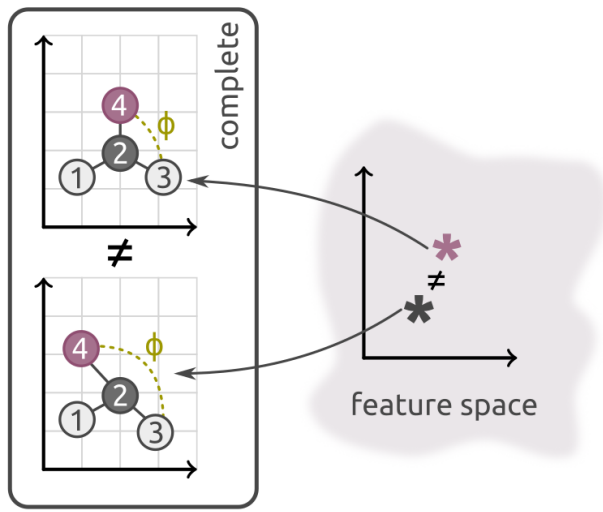
- Structure of three energetically lowest conformers:



Requirements for Effective Structural Representations

Structure representations have to be ...

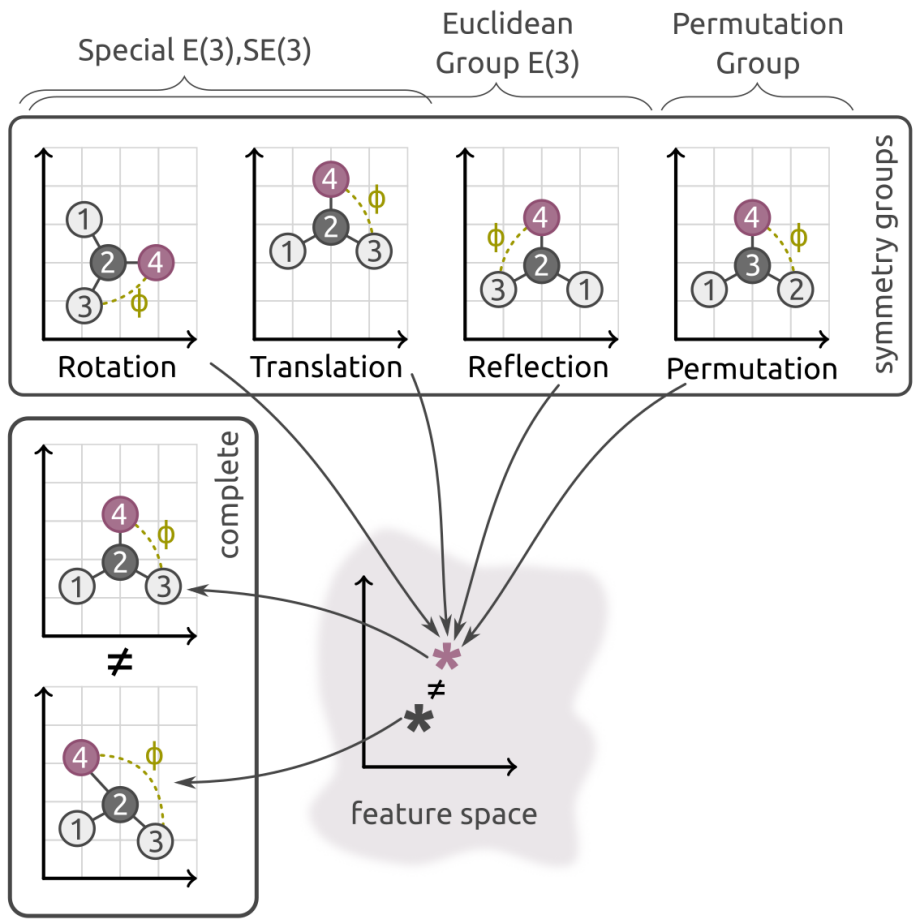
- **complete:** Inequivalent structures should be mapped to distinct features.



Requirements for Effective Structural Representations

Structure representations have to be ...

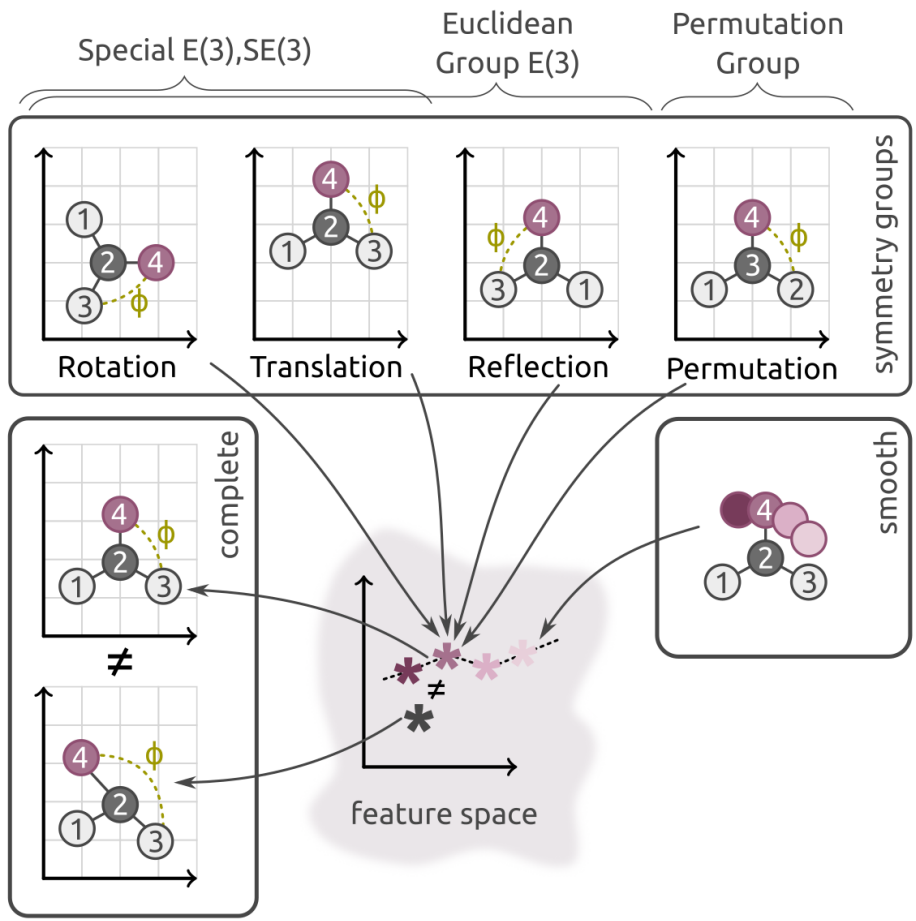
- **complete:** Inequivalent structures should be mapped to distinct features.
- **symmetry-respecting:** Equivalent structures should be mapped to the same features, obeying fundamental physical symmetries.



Requirements for Effective Structural Representations

Structure representations have to be ...

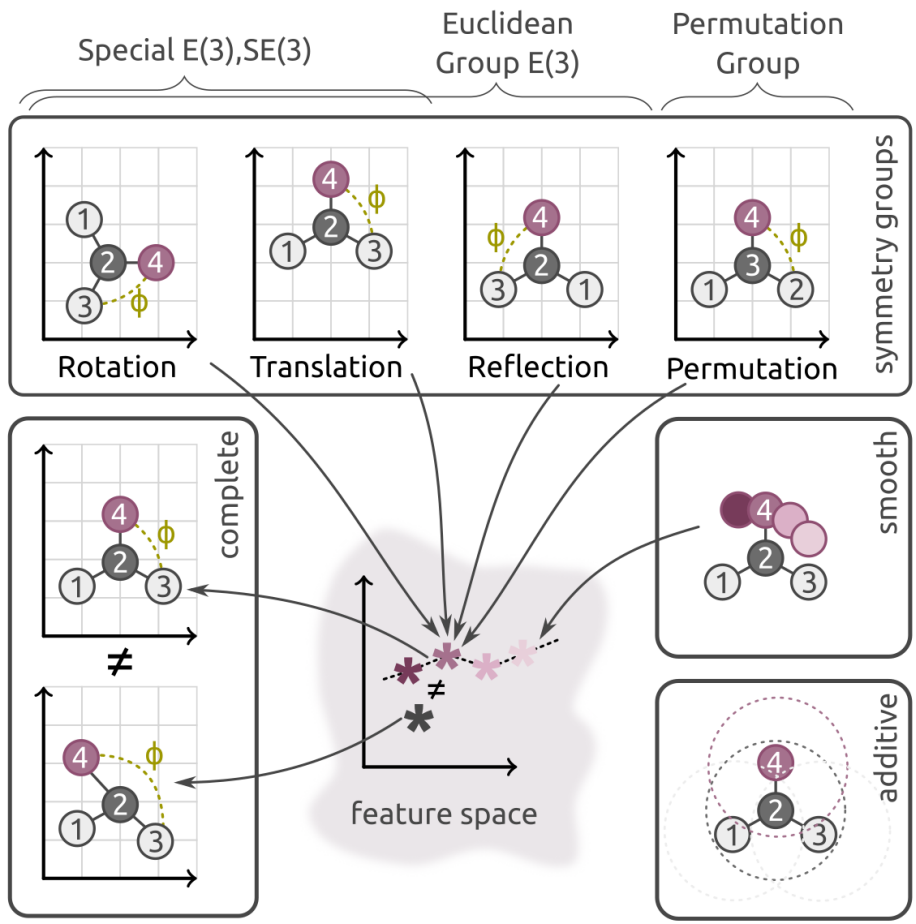
- **complete:** Inequivalent structures should be mapped to distinct features.
- **symmetry-respecting:** Equivalent structures should be mapped to the same features, obeying fundamental physical symmetries.
- **smooth:** Continuous deformations of a structure should result in smooth deformations of the associated features.



Requirements for Effective Structural Representations

Structure representations have to be ...

- **complete:** Inequivalent structures should be mapped to distinct features.
- **symmetry-respecting:** Equivalent structures should be mapped to the same features, obeying fundamental physical symmetries.
- **smooth:** Continuous deformations of a structure should result in smooth deformations of the associated features.
- **additiv:** For heterogeneous datasets (different molecular sizes), the representation should be decomposable into a sum of local environments (usually atom-centered), ensuring transferability and extensivity of predictions.



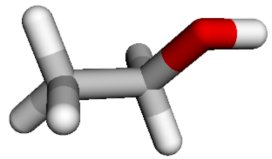
Coulomb Matrix

The **Coulomb matrix** encodes a molecule's structure through the electrostatic interactions between all pairs of atoms. Each entry reflects the nuclear charges and distances between atoms, capturing both the molecule's composition and geometry.

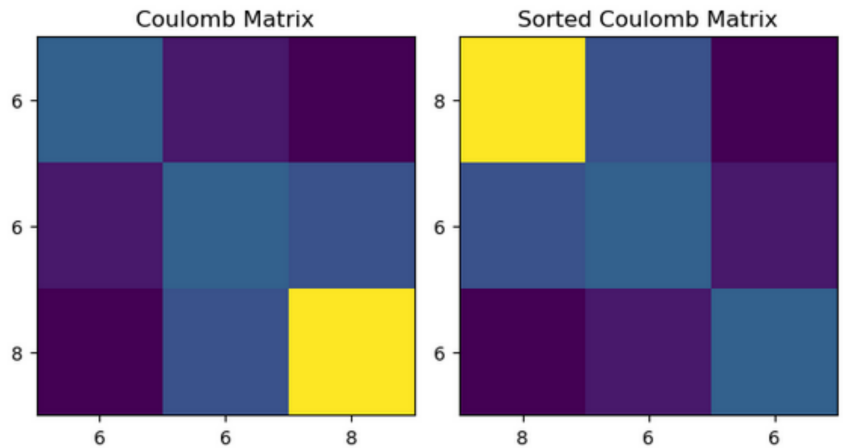
Symmetry considerations:

- rotation:
- translation:
- inversion:
- permutation:

Example: Coulomb Matrix of Ethanol



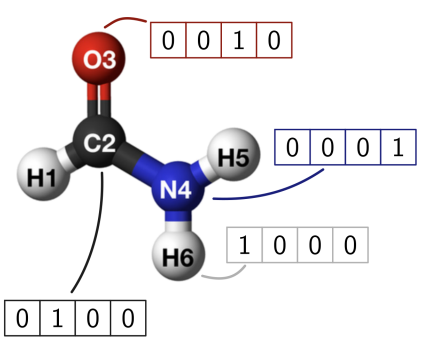
$$M_{ij} = \begin{cases} 0.5 Z_i^{2.4}, & \text{for } j = j \\ \dfrac{Z_i Z_j}{|R_i - R_j|}, & \text{for } i \leq j \end{cases}$$



input: Cartesian coordinates of heavy-atoms N_{at}, namely C and O
output: $N_{at} \times N_{at}$

Bonds-Angles-Dihedrals (BAD)

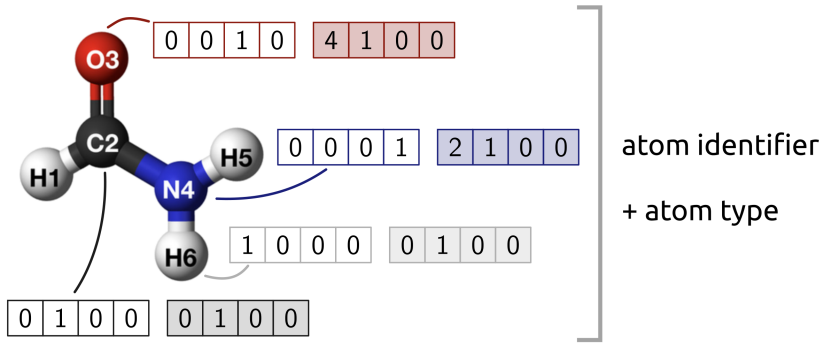
In the **B**onds-**A**ngles-**D**ihedrals molecular representation, each molecule is viewed as consisting of bonded pairs (adjacent atoms); a pair of consecutive bonds forming an angle, and three consecutive bonds forming a dihedral angle.



(image adapted from: https://doi.org/10.1002/jcc.26128)

Bonds-Angles-Dihedrals (BAD)

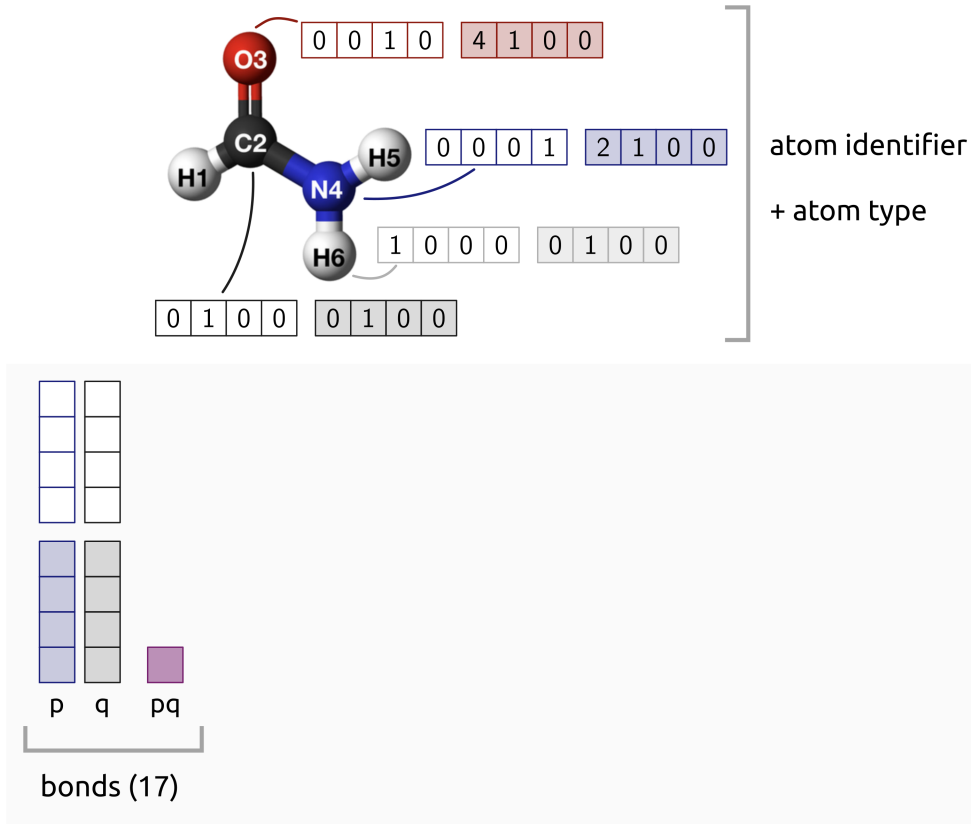
In the **B**onds-**A**ngles-**D**ihedrals molecular representation, each molecule is viewed as consisting of bonded pairs (adjacent atoms); a pair of consecutive bonds forming an angle, and three consecutive bonds forming a dihedral angle.



(image adapted from: https://doi.org/10.1002/jcc.26128)

Bonds-Angles-Dihedrals (BAD)

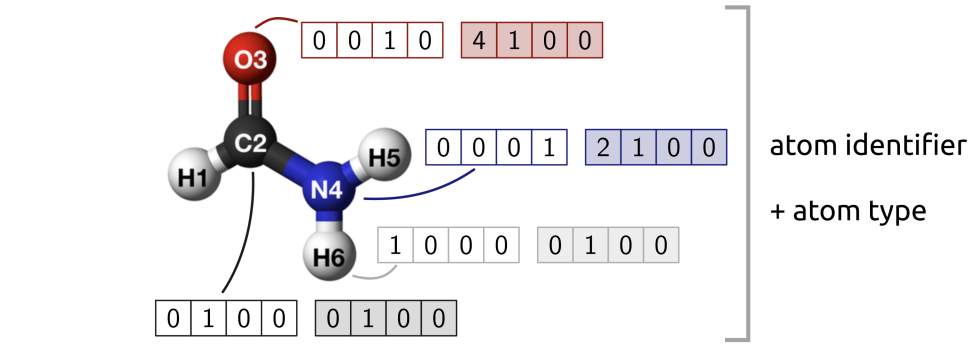
In the **B**onds-**A**ngles-**D**ihedrals molecular representation, each molecule is viewed as consisting of bonded pairs (adjacent atoms); a pair of consecutive bonds forming an angle, and three consecutive bonds forming a dihedral angle.

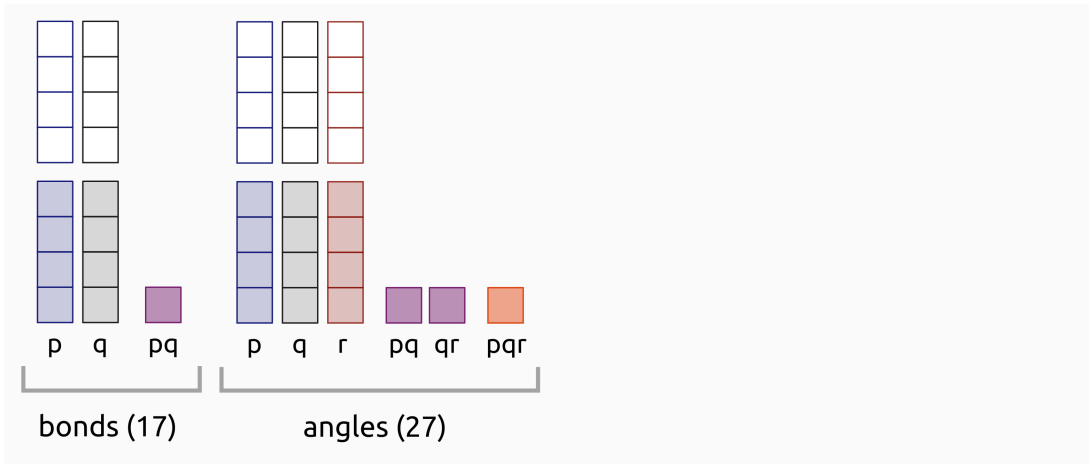


(image adapted from: https://doi.org/10.1002/jcc.26128)

Bonds-Angles-Dihedrals (BAD)

In the **B**onds-**A**ngles-**D**ihedrals molecular representation, each molecule is viewed as consisting of bonded pairs (adjacent atoms); a pair of consecutive bonds forming an angle, and three consecutive bonds forming a dihedral angle.

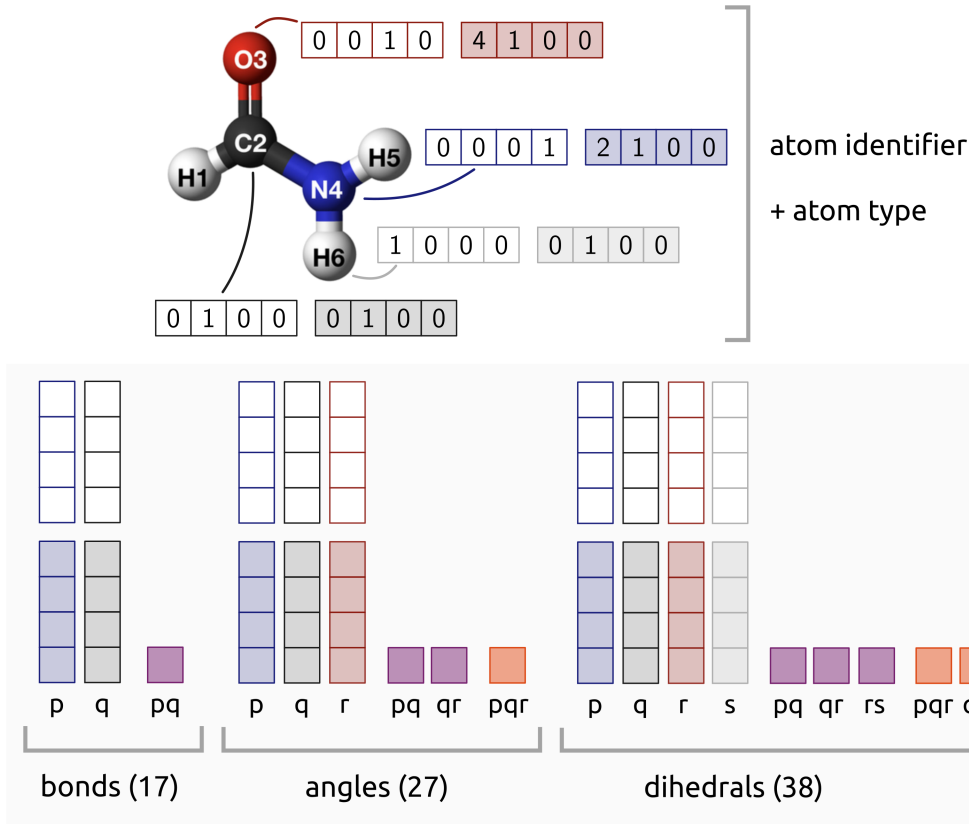




(image adapted from: https://doi.org/10.1002/jcc.26128)

Bonds-Angles-Dihedrals (BAD)

In the **B**onds-**A**ngles-**D**ihedrals molecular representation, each molecule is viewed as consisting of bonded pairs (adjacent atoms); a pair of consecutive bonds forming an angle, and three consecutive bonds forming a dihedral angle.

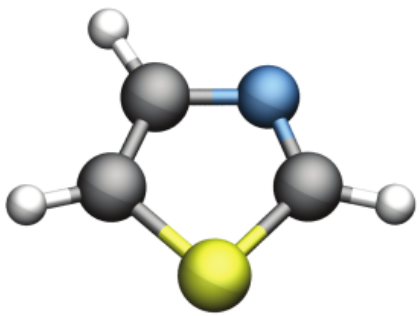


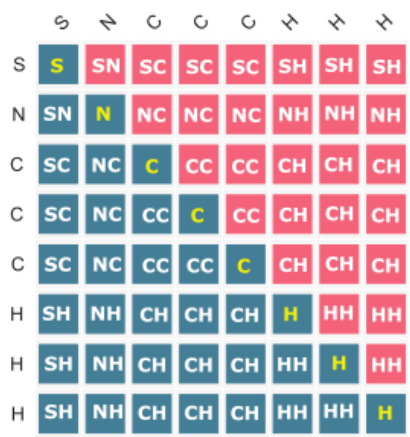
(image adapted from: https://doi.org/10.1002/jcc.26128)

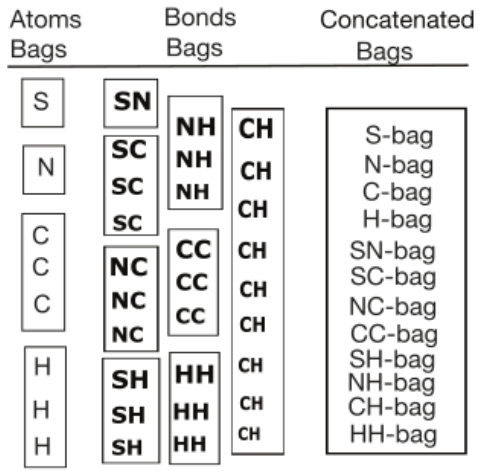
Bag of Bonds (BoB)

The **B**ag **o**f **B**onds (BoB) is structural representation inspired by the *bag-of-words* concept from natural language processing.

- Molecules are decomposed into *bags* corresponding to bond types (*e.g.*, C–O, C–N) and bond orders (single, double, triple).
- each element of a bag is computed as $\frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}$, where Z_I, Z_J are nuclear charges and $\mathbf{R}_I, \mathbf{R}_J$ are atomic positions.
- Bags are concatenated and zero-padded to yield fixed-length feature vectors.





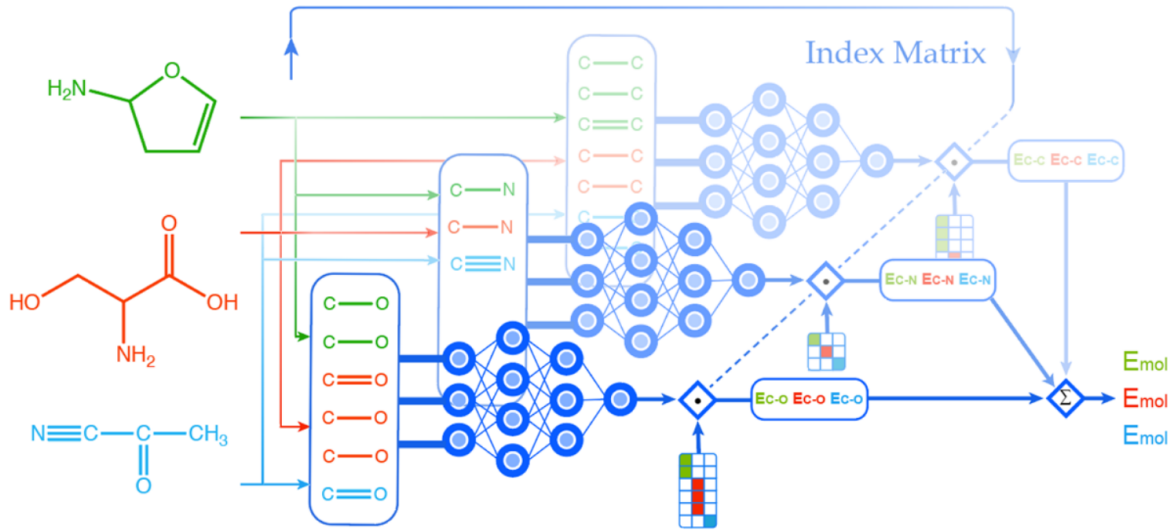


(image taken from: https://doi.org/10.1002/qua.26870)

Bonds In Molecules (BIM)

In the **B**onds **I**n **M**olecules (BIM) is a local, bond-centered structural representation. In this approach, a molecule is decomposed into a set of overlapping bond environments, such as C–H, C–C, or C–O bonds.

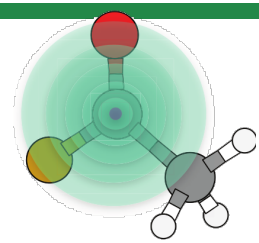
- Covalent bonds are identified using element-specific distance cutoffs.
- Each descriptor encodes:
  - the bond length of the target bond,
  - the nearest-neighbor bonds connected to it, and
  - the angles between those neighboring bonds.

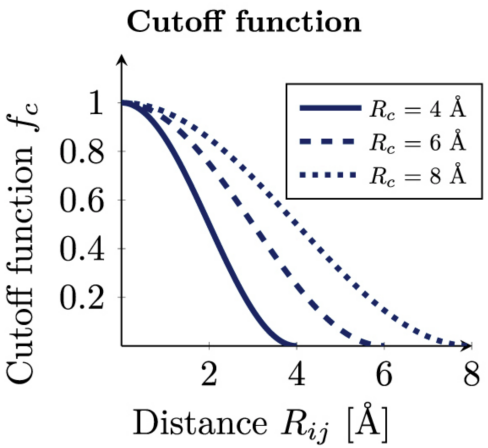


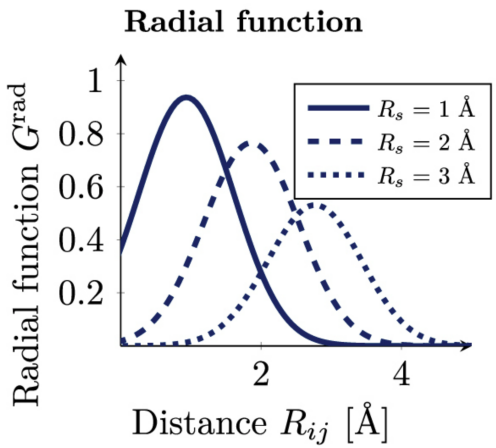
(image taken from: https://doi.org/10.1021/acs.jpclett.7b01072)

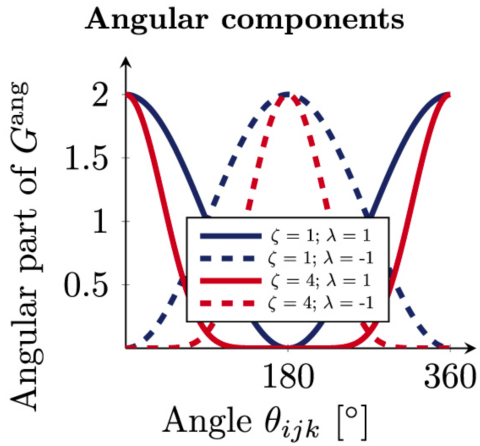
Atom Centered Symmetry Functions (ACSFs)

Atom-Centered Symmetry Functions (ACSFs) provide a numerical description of the local environment of each atom. They encode atomic neighborhoods in a way that is limited to a local region within a cutoff radius R_c.









$$f_c(R_{IJ}) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi R_{IJ}}{R_c}\right) + 1\right], & R_{IJ} \leq R_c, \\ 0, & R_{IJ} > R_c \end{cases}$$

$$G_I^{\text{ang}} = 2^{1-\zeta} \sum_{J \neq I} \sum_{K > J} (1 + \lambda \cos\theta_{IJK})^{\zeta} \exp\left[-\eta(R_{IJ}^2 + R_{IK}^2 + R_{JK}^2)\right] f_c(R_{IJ}) f_c(R_{IK}) f_c(R_{JK})$$

$$G_I^{\text{rad}} = \sum_{J \neq I} \exp\left[-\eta(R_{IJ} - R_s)^2\right] f_c(R_{IJ})$$

Comparing Molecules

Outlook