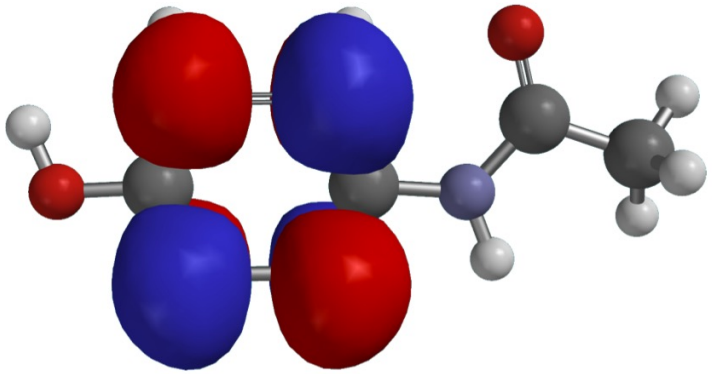
Representing molecules in the computer

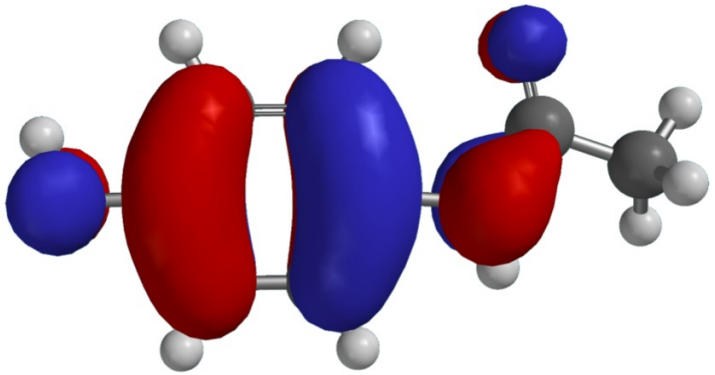
1. Molecular formats
   - SMILES
   - InChi
   - InChiKey
   - Mol-blocks
   - SDF
   - PDB
2. Visualisation
   - Graphical software
   - Representations
   - Molecular motion

Molecular representation

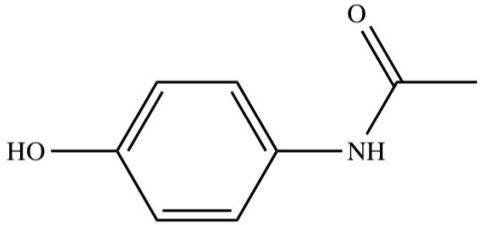
LUMO



HOMO



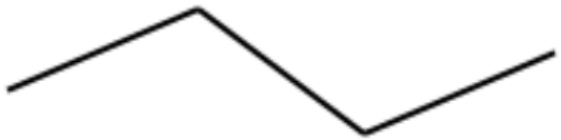
(a)



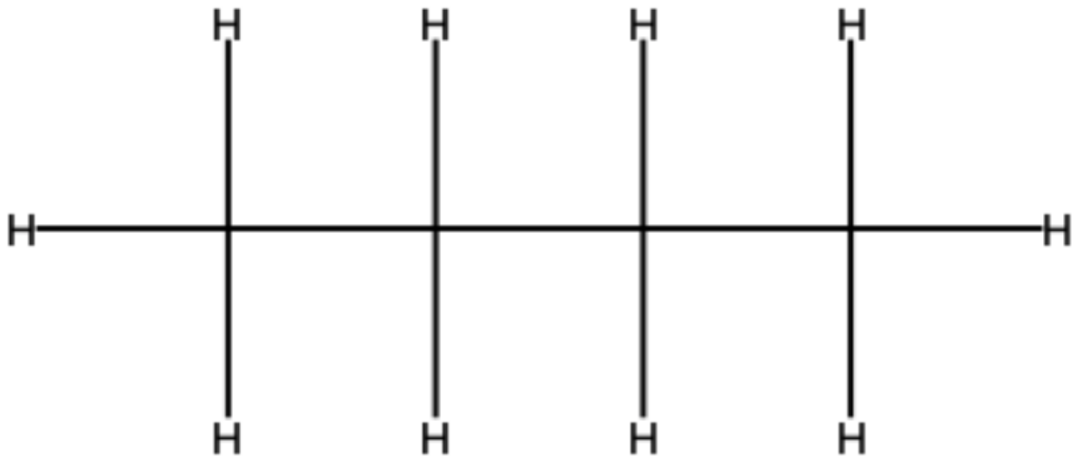
(b)

Molecular representation: SMILES

- cccc



cccc



cccc

Atoms in SMILES

- `[Ag]`
- `C`, `N`, `O`, `Br`, `I`, `F`, `Cl`, `B`, `P`, `S`
  - Implicit hydrogens `[CH3-]`
  - Formal charges `[O+]`
  - Isotopes `[14C]`
  - Chirality
  - Aromaticity

Bonds in SMILES

- Implicit (single or aromatic)
- -
  =
  #
  :

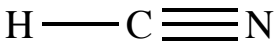
Rings in SMILES

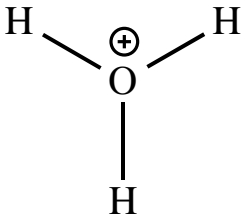
- Digits to denote start and end of ring closure
- Digits may be reused

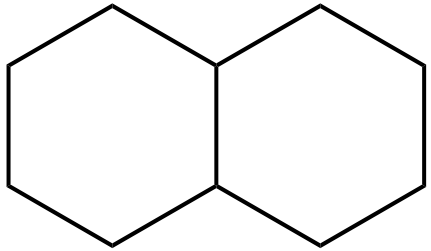
Disconnected structures

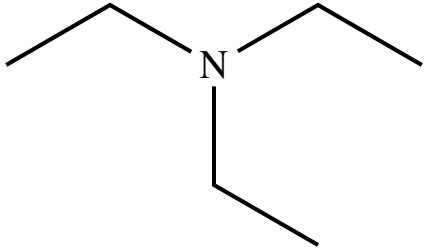
- Separated by a dot (`.`)

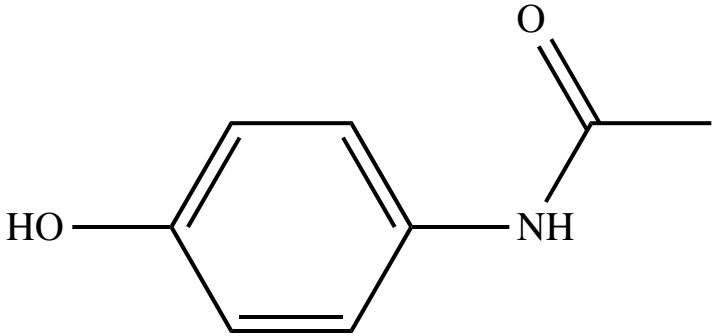
Examples

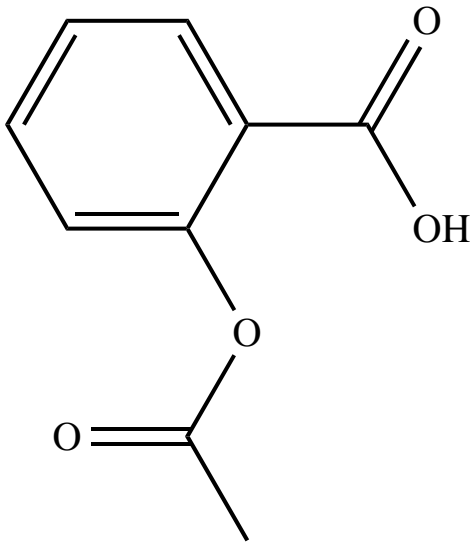


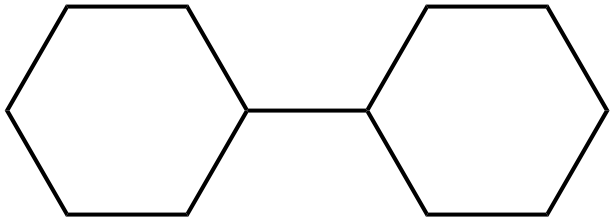


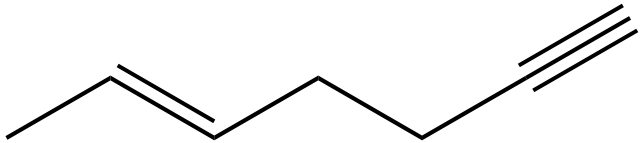






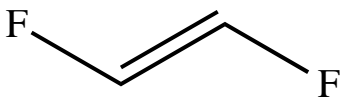




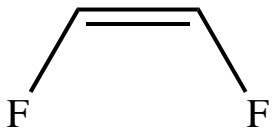


Stereochemistry around bonds

- Double bond configuration is denoted with / and \



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

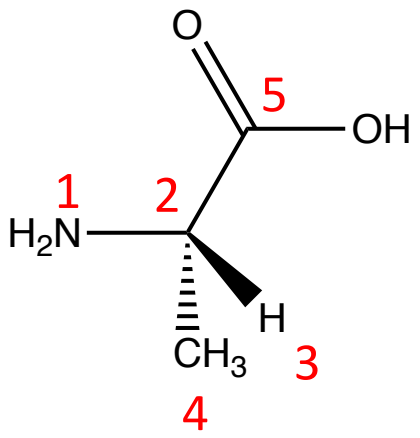


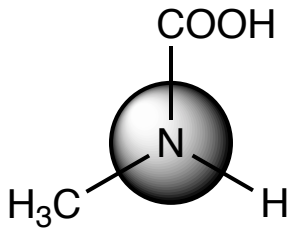
F/C=C\F
F/C=C\F

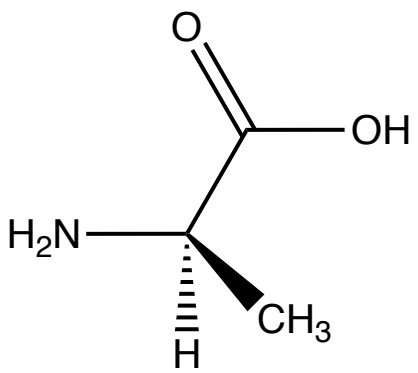
C/C=C/C=C/C

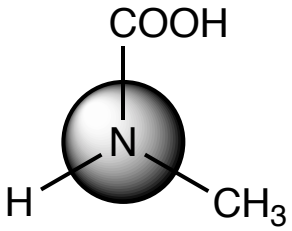
Stereochemistry around sp^3 carbons

- Use @ (counter-clockwise) and @@ (clockwise)









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

1 2 3 4 5

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

Canonical SMILES

- Methane
  - C
  - [CH4]
  - [H][C]([H])([H])
  - [#1][#6]([H])([#1])[#1]
  - ....

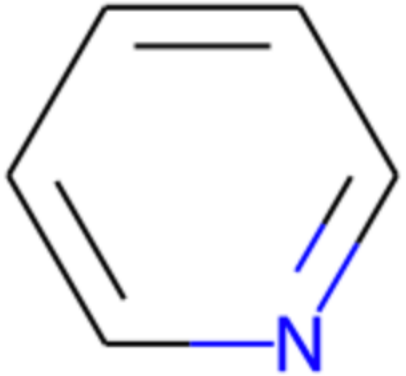
```python
all_methanes = ['C',
                '[CH4]',
                '[H][C]([H])([H])[H]',
                '[#1][#6]([H])([#1])[#1]']
for methane in all_methanes:
  mol = Chem.MolFromSmiles(methane)
  print(Chem.MolToSmiles(mol))
```

```
C
C
C
C
```

InChi: main layer

```
mol = Chem.MolFromSmiles("c1ncccc1")
inchi = Chem.MolToInchi(mol)
print("InChi: ", inchi)
mol
```

```
InChi:  InChI=1S/C5H5N/c1-2-4-6-5-3-1/h1-5H
```

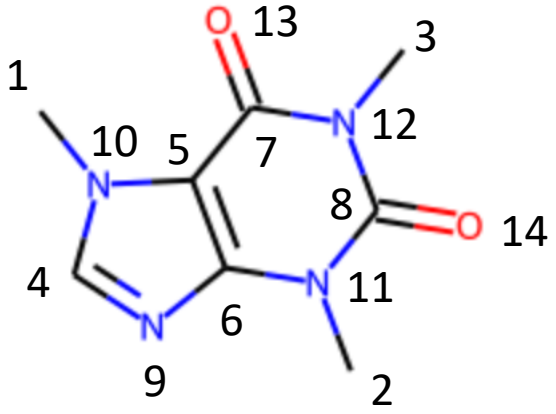


1) InChi version
2) “Chemical formula” sublayer
3) “Atom connection” sublayer (starts with /c)
4) “Hydrogen atom: sublayer (starts with /h)

InChi: branching

```
mol = Chem.MolFromSmiles("O=C1C2=C(N=CN2C)N(C)C(N1C)=O")
inchi = Chem.MolToInchi(mol)
print("InChi: ", inchi)
mol
```

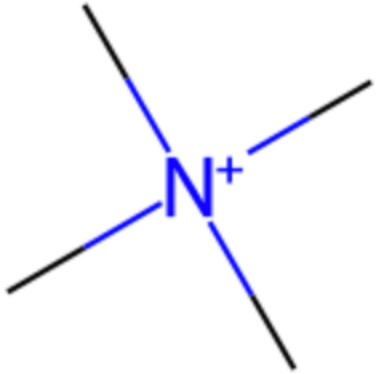
```
InChi:  InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
```



InChi: charge layer
Total charge sublayer for permanent formal charges

```
mol = Chem.MolFromSmiles("C[N+](C)(C)C")
inchi = Chem.MolToInchi(mol)
print("InChi: ", inchi)
mol
```

```
InChi:  InChI=1S/C4H12N/c1-5(2,3)4/h1-4H3/q+1
```



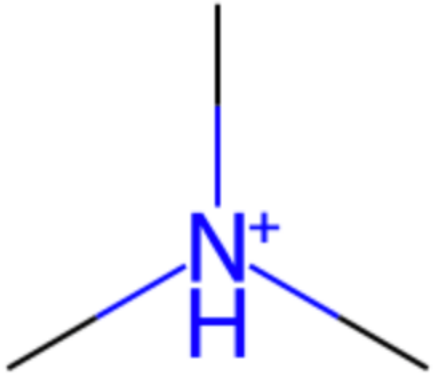
"Total charge" sublayer (starts with /q)

InChi: charge layer
Protonation/deprotonation sublayer

+ Code + Text

```
mol = Chem.MolFromSmiles("C[NH+](C)C")
inchi = Chem.MolToInchi(mol)
print("InChi: ", inchi)
mol
```

```
InChi:  InChI=1S/C3H9N/c1-4(2)3/h1-3H3/p+1
```

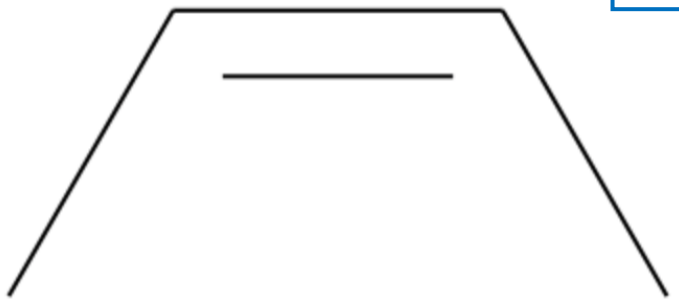


"Protonation/deprotonation" sublayer (starts with **/p**)

InChi: stereochemistry layer
Double bonds

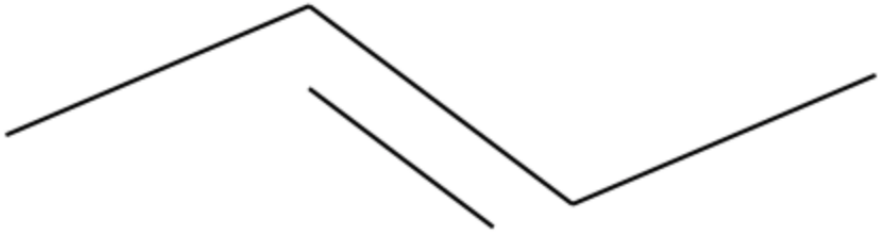
```
mol = Chem.MolFromSmiles("C/C=C\C")
inchi = Chem.MolToInchi(mol)
print("InChi: ", inchi)
mol
```

```
InChi:  InChI=1S/C4H8/c1-3-4-2/h3-4H,1-2H3/b4-3-
```



```
mol = Chem.MolFromSmiles("C/C=C/C")
inchi = Chem.MolToInchi(mol)
print("InChi: ", inchi)
mol
```

```
InChi:  InChI=1S/C4H8/c1-3-4-2/h3-4H,1-2H3/b4-3+
```

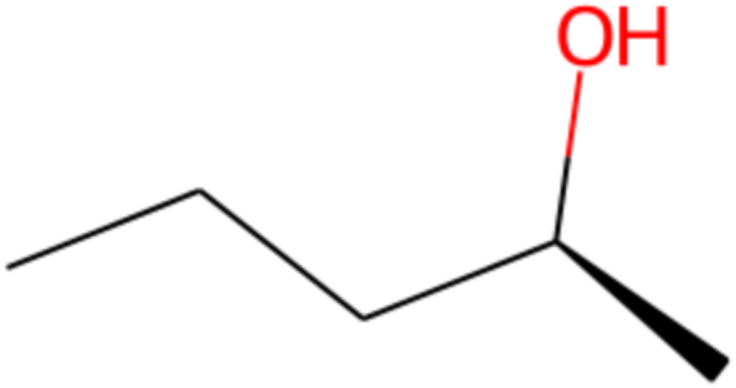


"Double bonds" sublayer (starts with **/b**)

InChi: stereochemistry layer
sp^3 systems

```
[9]  mol = Chem.MolFromSmiles("CCC[C@@]([H])(O)C")
     inchi = Chem.MolToInchi(mol)
     print("InChi: ", inchi)
     mol
```

```
InChi:  InChI=1S/C5H12O/c1-3-4-5(2)6/h5-6H,3-4H2,1-2H3/t5-/m0/s1
```



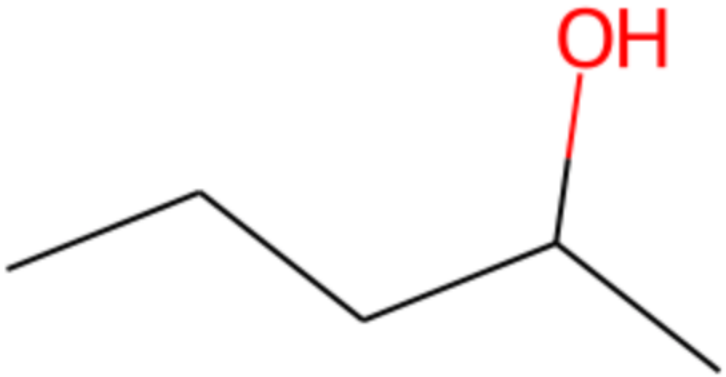
"sp3 stereo" sublayer: indicated with `/t`, `/m` and `/s`

InChiKey

```
mol = Chem.MolFromSmiles("CCCC(O)C")
inchikey = Chem.MolToInchiKey(mol)
print("InChiKey: ", inchikey)
mol
```

```
InChiKey:  JYVLIDXNZAXMDK-UHFFFAOYSA-N
```

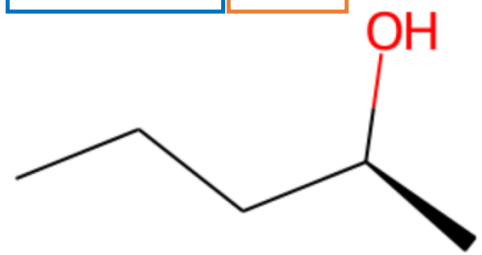
27 characters



InChiKey
How does it look like?

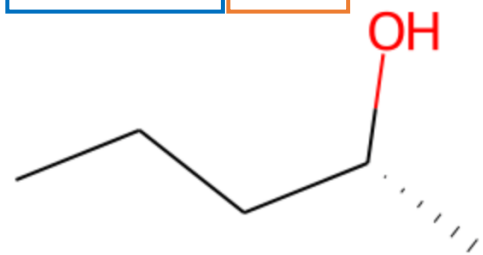
```
mol = Chem.MolFromSmiles("CCC[C@@]([H])(O)C")
inchikey = Chem.MolToInchiKey(mol)
print("InChiKey: ", inchikey)
mol
```

```
InChiKey:  JYVLIDXNZAXMDK-YFKPBYRVSA-N
```




```
mol = Chem.MolFromSmiles("CCC[C@]([H])(O)C")
inchikey = Chem.MolToInchiKey(mol)
print("InChiKey: ", inchikey)
mol
```

```
InChiKey:  JYVLIDXNZAXMDK-RXMQYKEDSA-N
```



Skeleton description: 14 characters

Stereochemistry, charges and isotopes: 8 characters

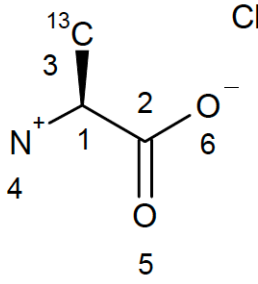
S: derived from standard InChi

A: InChiKey version 1

N: neutral

Mol-blocks

L-Alanine



Chiral

```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0
  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Header Block

Counts Line

Atom Block

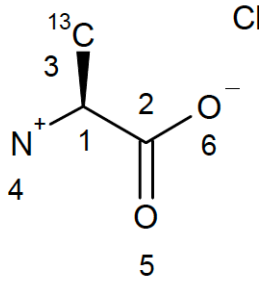
Bond Block

Properties Block

Connection Table (Ctab)

Mol-blocks: atoms

L-Alanine



Chiral

x,y,z-coordinates
atom symbol
mass difference flag
charge flag (5 = **-1**, 3 = **+1**)
chirality flag

```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0

  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Header Block

Counts Line

Atom Block

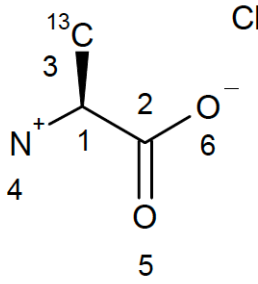
Bond Block

Properties Block

Connection Table (Ctab)

Mol-blocks: bonds

L-Alanine



Chiral

begin atom
end atom
bond type
bond stereo

```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0

  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Header Block

Counts Line

Atom Block

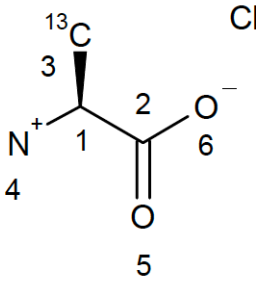
Bond Block

Properties Block

Connection Table (Ctab)

Mol-blocks: properties

L-Alanine



Chiral

charge property
isoptope property
required END card

```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0

  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Header Block

Counts Line

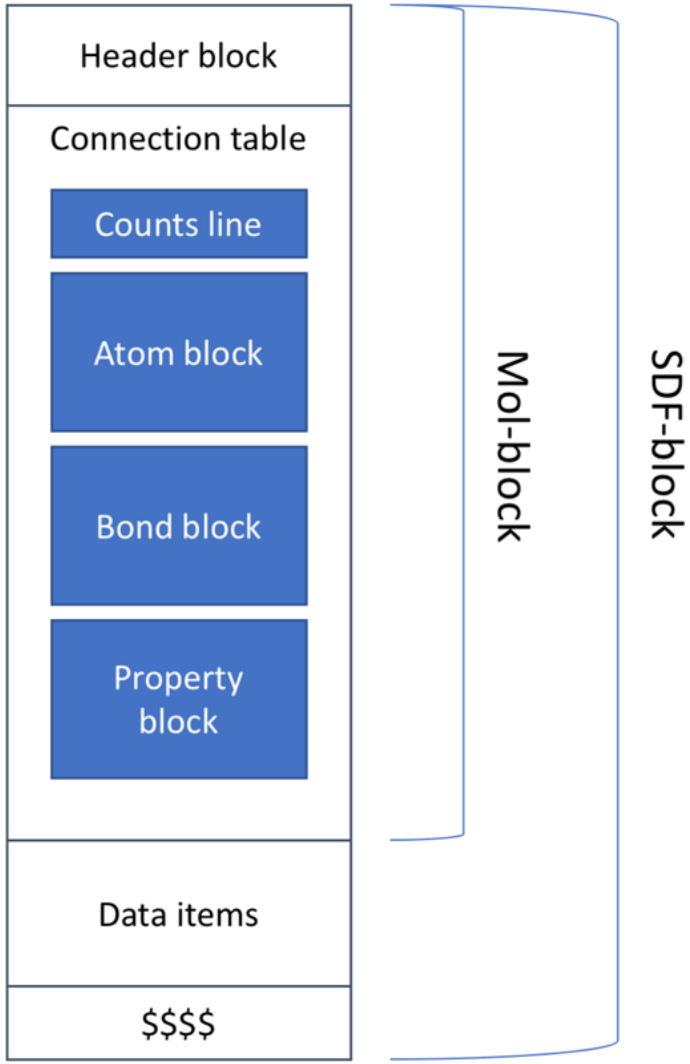
Atom Block

Bond Block

Properties Block

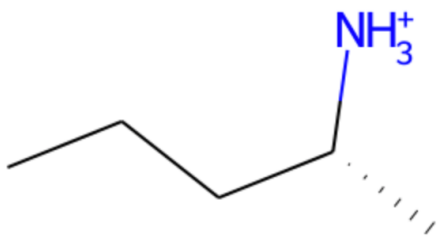
Connection Table (Ctab)

Structure-Data-Format (SDF)



SDF: Data items

```python
mol = Chem.MolFromSmiles("CCC[C@]([H])([NH3+])C")
mol.SetProp("_Name", "Compound_name")
mol.SetProp("Data_1", "0.90")
mol.SetProp("Data_2", "Charged compound")
mol.SetProp("Warranty", "Do not eat")
sdfWriter = Chem.SDWriter("out.sdf")
sdfWriter.write(mol)
sdfWriter.close()
mol
```



+ Code + Text

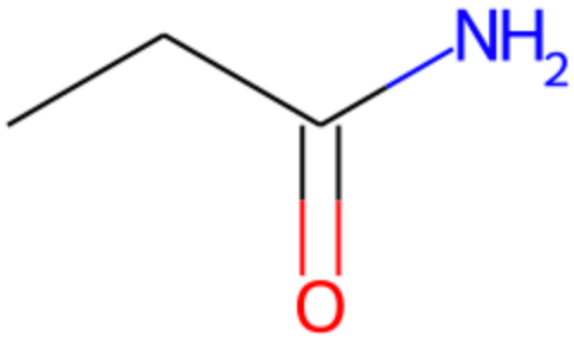
```python
!cat out.sdf
```

```
Compound_name
     RDKit          2D

  6  5  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2990    0.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5981   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8971    0.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8971    2.2500    0.0000 N   0  0  0  0  0  4  0  0  0  0  0  0
    5.1962   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0
  2  3  1  0
  3  4  1  0
  4  5  1  0
  4  6  1  6
M  CHG  1   5   1
M  END
>  <Data_1>  (1)
0.90

>  <Data_2>  (1)
Charged compound

>  <Warranty>  (1)
Do not eat

$$$$
```

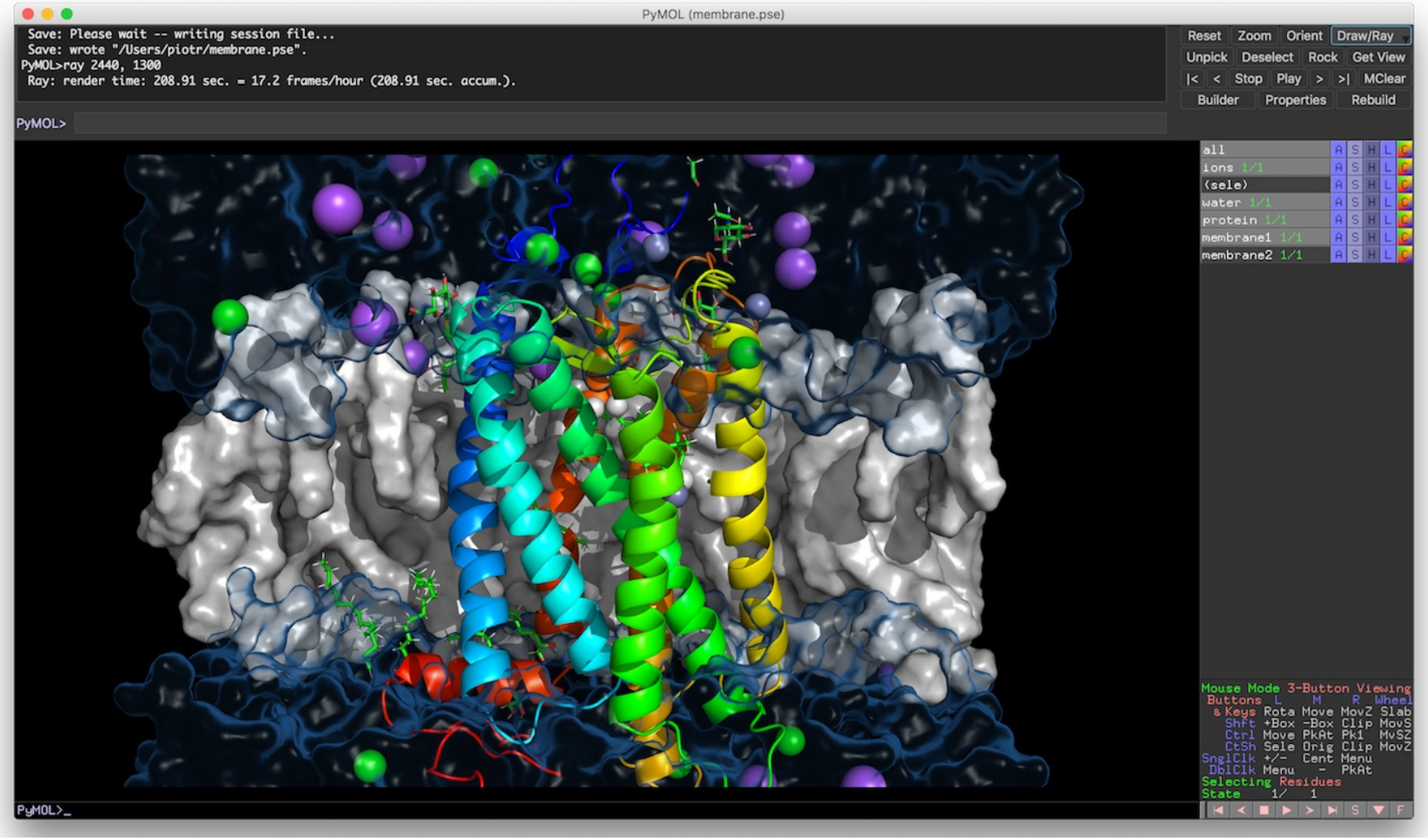
PDB

```python
mol = Chem.MolFromSmiles("CCC(=O)N")
mol.SetProp("_Name", "Propanamide")
print(Chem.MolToPDBBlock(mol))
mol
```

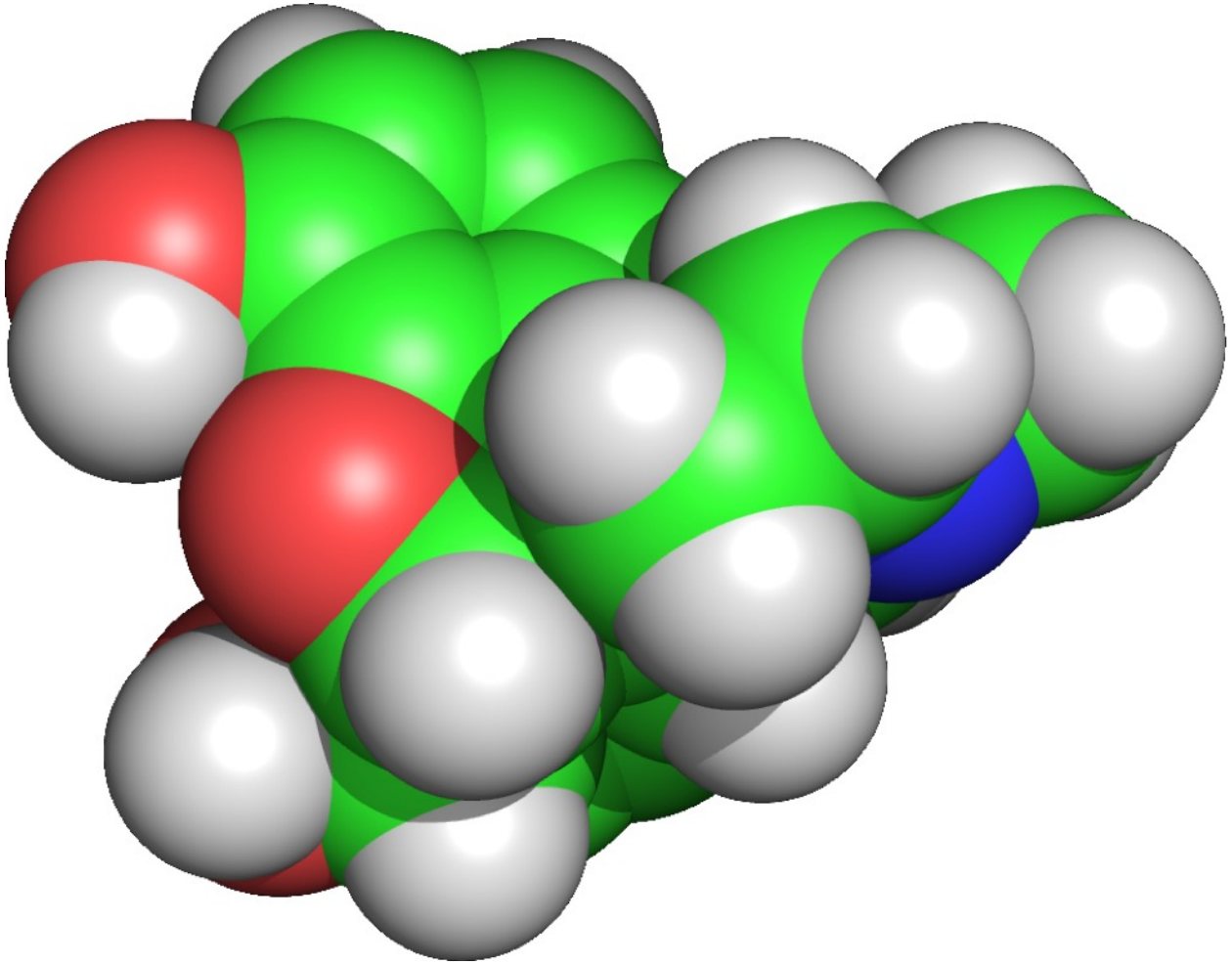
```
COMPND     Propanamide
HETATM     1  C1   UNL      1       0.000   0.000   0.000  1.00  0.00           C
HETATM     2  C2   UNL      1       0.000   0.000   0.000  1.00  0.00           C
HETATM     3  C3   UNL      1       0.000   0.000   0.000  1.00  0.00           C
HETATM     4  O1   UNL      1       0.000   0.000   0.000  1.00  0.00           O
HETATM     5  N1   UNL      1       0.000   0.000   0.000  1.00  0.00           N
CONECT     1     2
CONECT     2     3
CONECT     3     4     4     5
END
```

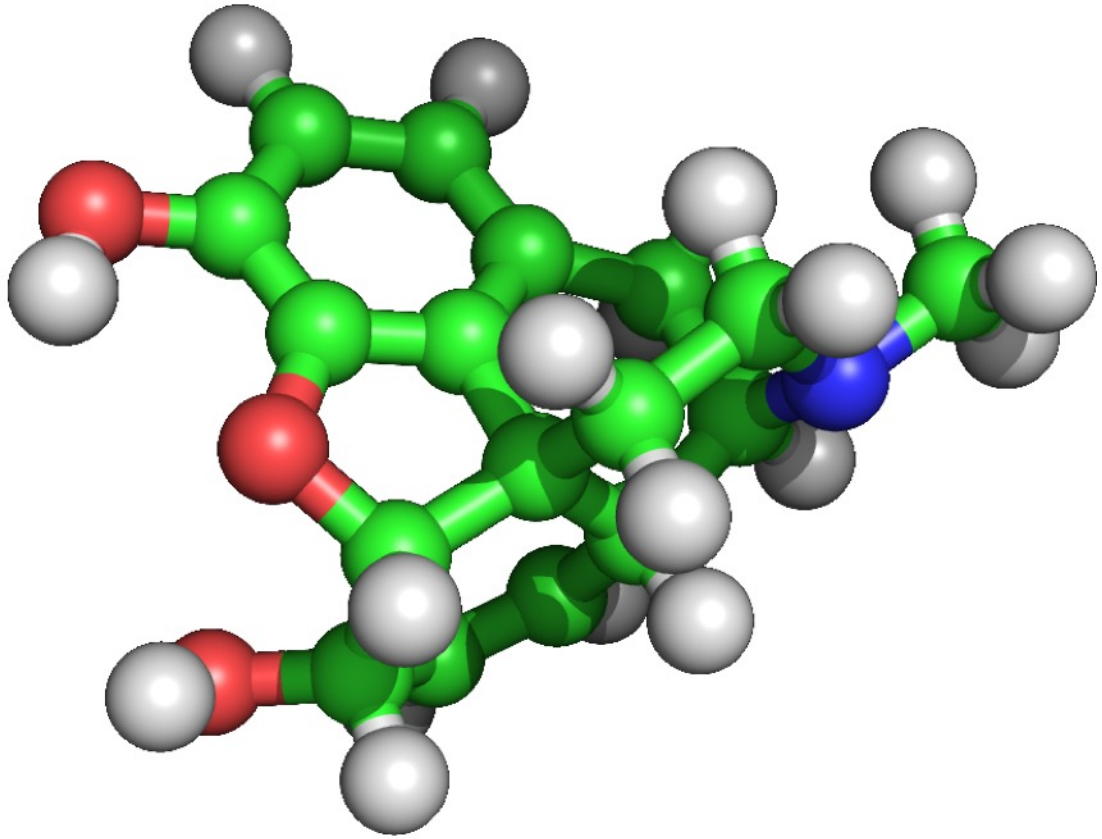


Graphical software: VMD and PyMol

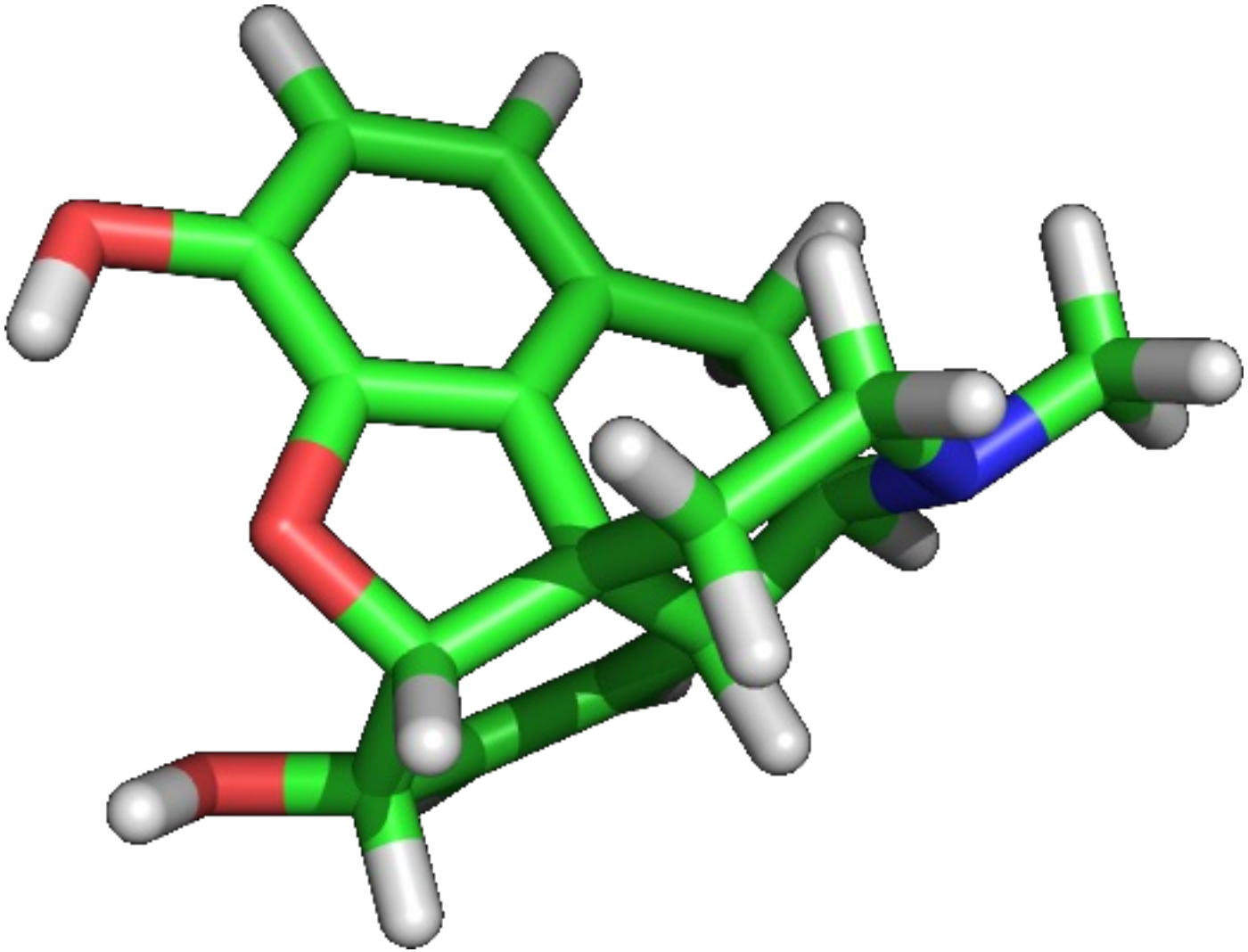


CPK: Corey-Pauling-Koltun

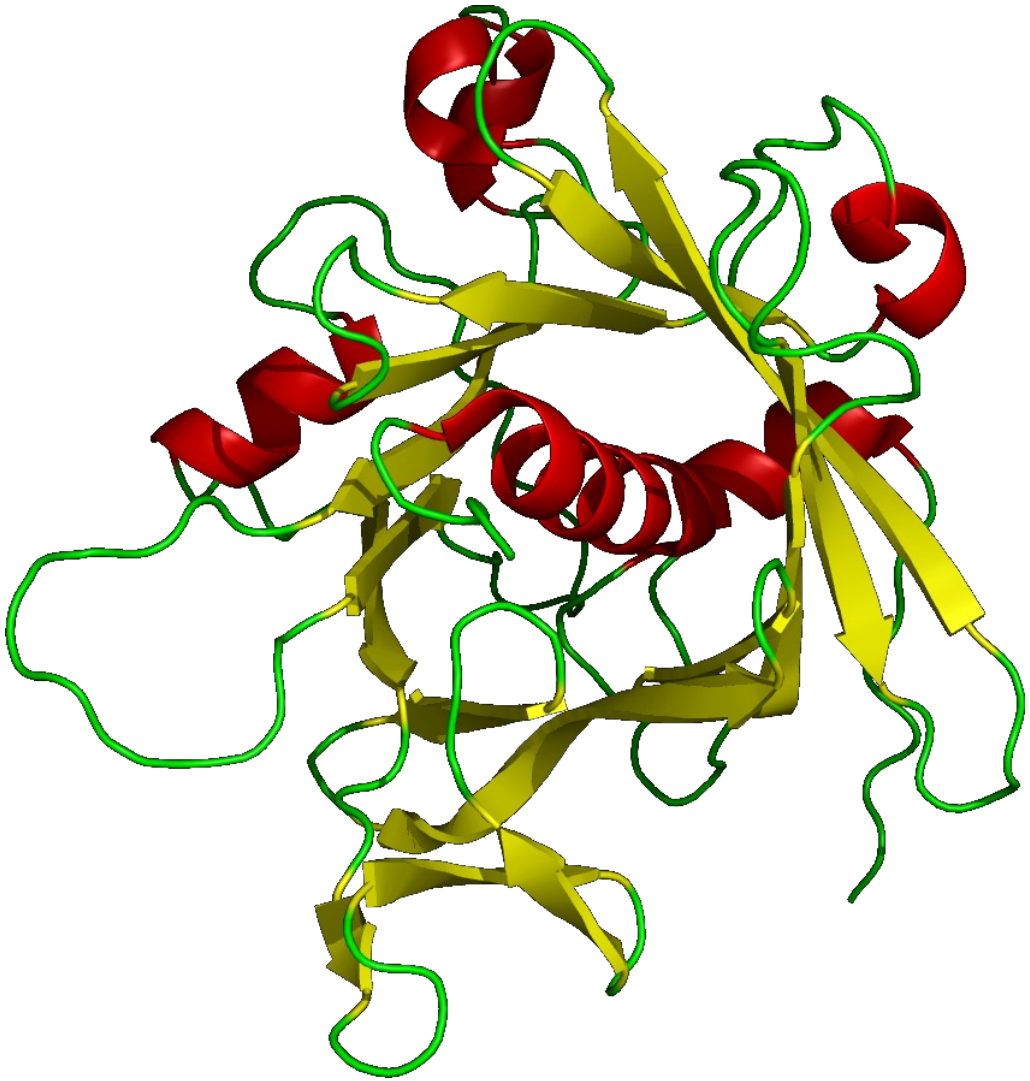




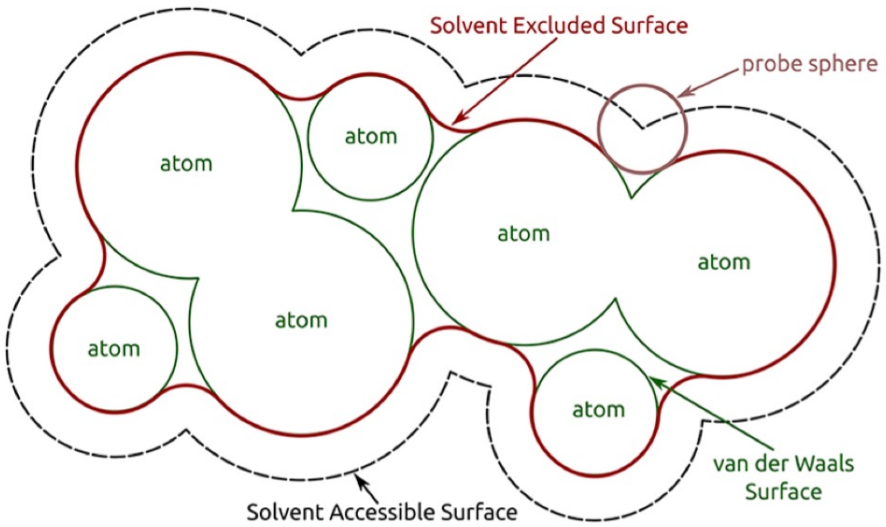
Sticks

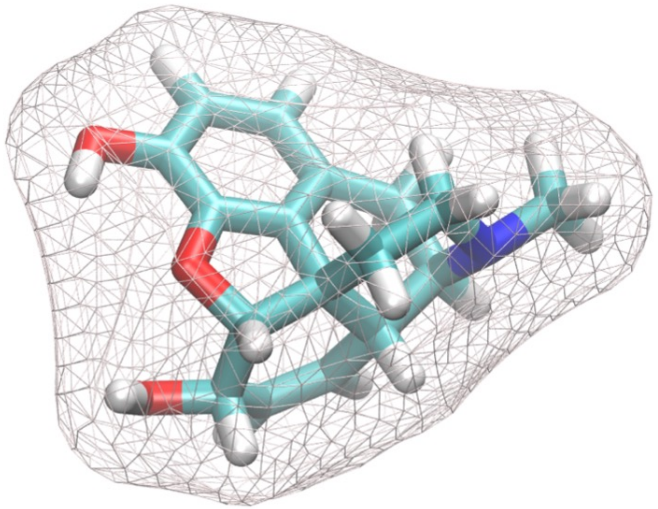


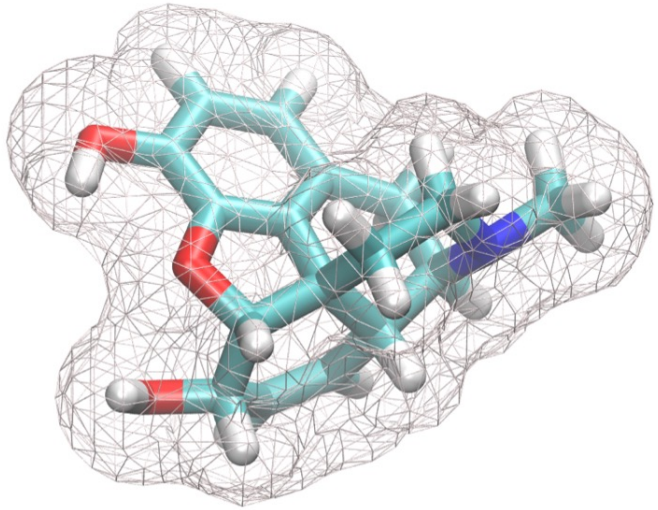
Ribbon

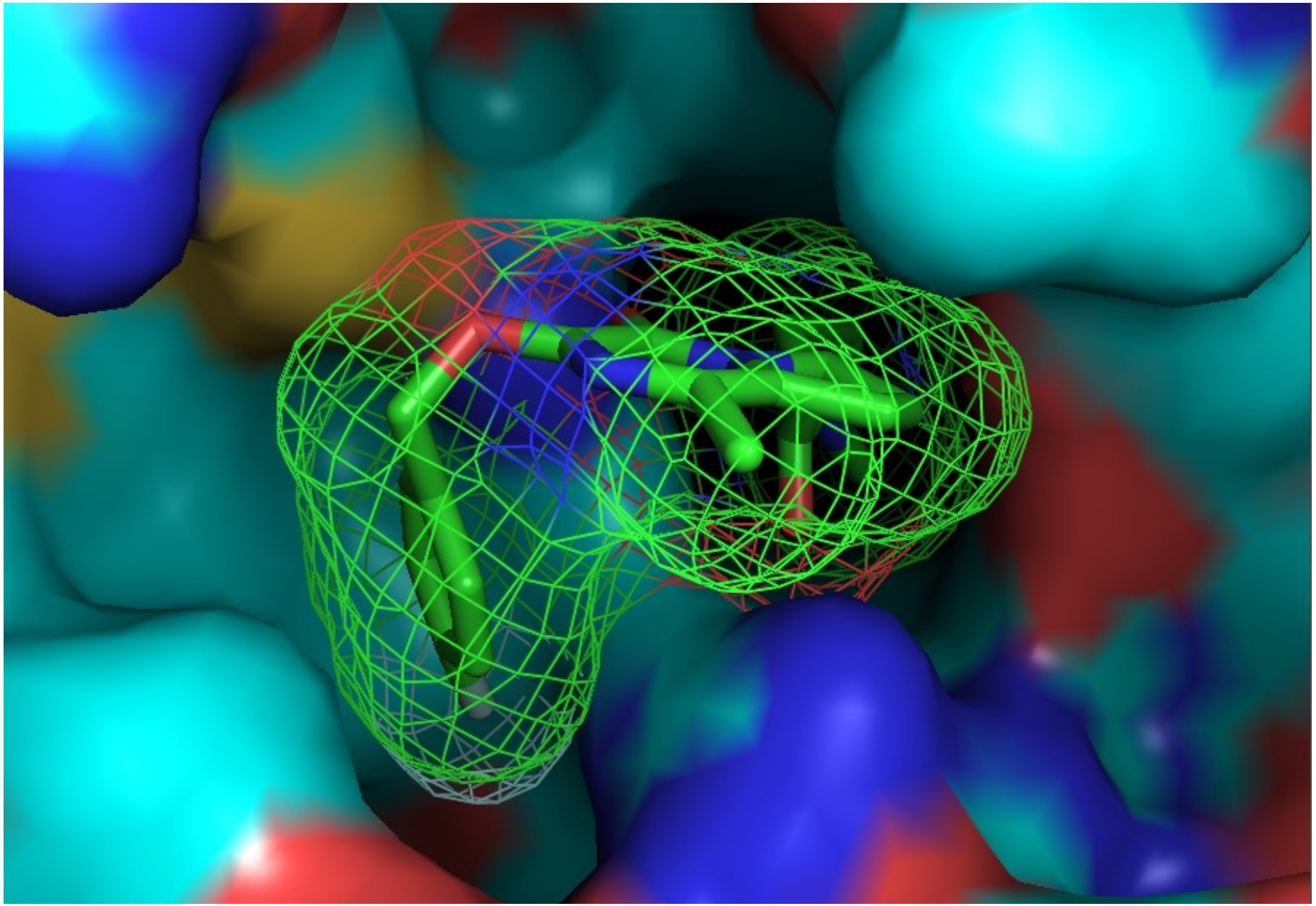


Solvent-accessible surface

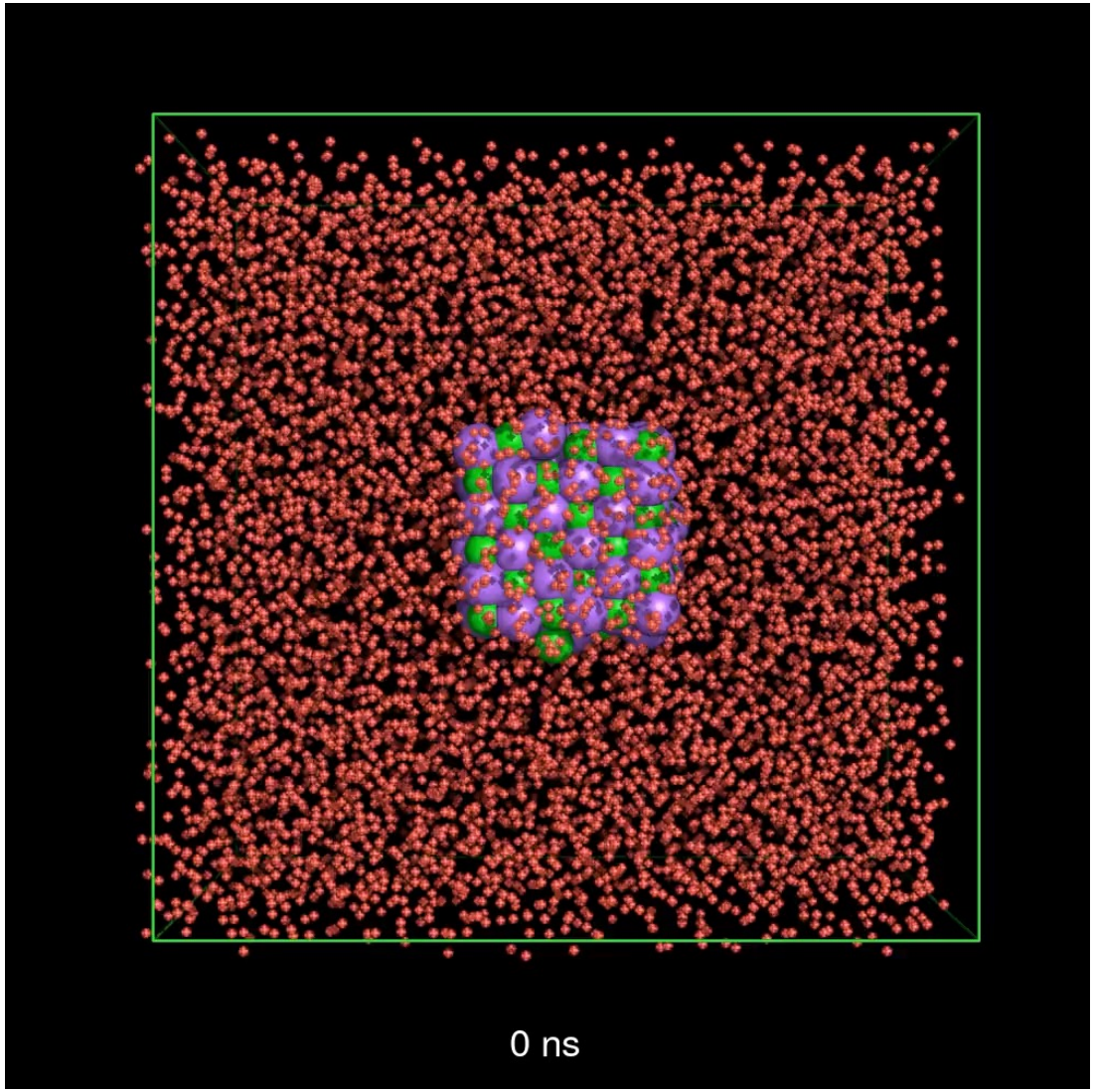








Visualising molecular motion
NaCl dissolution



Visualising molecular motion
Ligand-protein dissociation