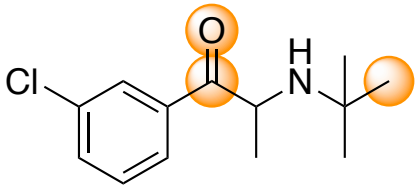
Generation of Bimolecular 3D Complex Structures with RDKit

Christoph Bauer, RDKit UGM

Hamburg, Thursday September 26, 2019

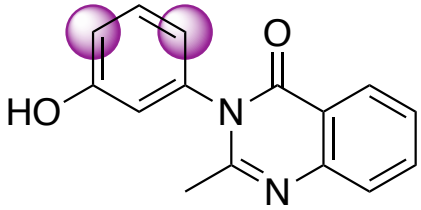
Atomic Descriptors in Practice

Site of metabolism

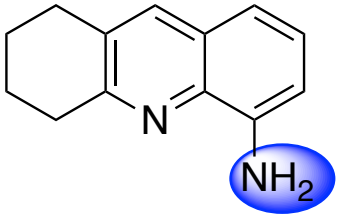


Bupropion

Site of reaction

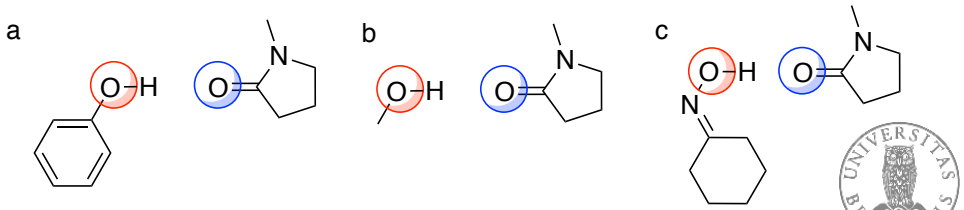


Site of reaction?
(P450 mediated)



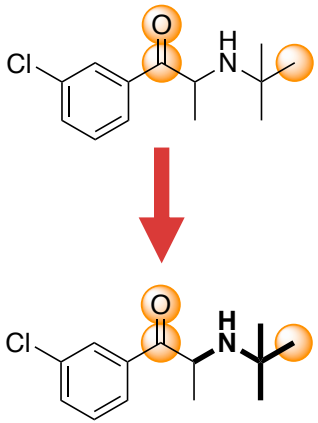
HB donor

HB acceptor



Quantum Chemistry Applications

SmartCYP



Pattern matching – precomputed $\Delta E^{\ddagger}$

Ames mutagenicity

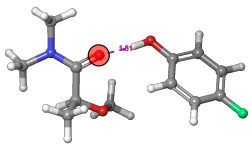
$$\Delta\Delta E = E_{ArNH^+} + E_{PhNH_2} - E_{ArNH_2} - E_{PhNH^+}$$

nitrenium hypothesis

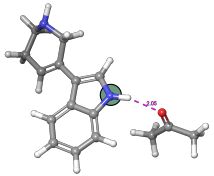
Bentzien J. et al *J. Chem. Inf. Model.* **2010**, *50*, 274-297.

Linking Quantum Chemistry and ML

QC data → ML training → predictions



-12.0 kJ mol^{-1}



-2.3 kJ mol^{-1}

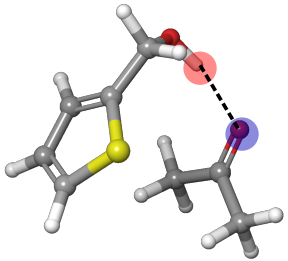
$$\begin{matrix} f \\ f_* \end{matrix} \sim N\left(0 \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*X_*) \end{bmatrix}\right)$$

(Gaussian Process Regression)

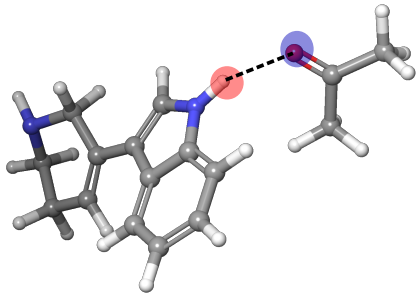
For example:
H-bonding strength

Bauer, C. A.; Schneider, G.; Göller, A. H. *J. Cheminform.* **2019**, *11*, 59.

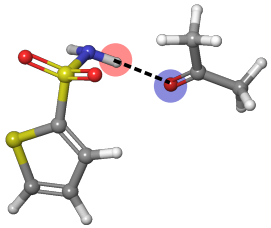
But How Do We Get the QC Data?



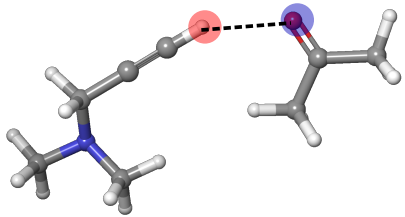
$\Delta G = -2.85$ kJ mol^{-1}



$\Delta G = -2.30$ kJ mol^{-1}



$\Delta G = -4.85$ kJ mol^{-1}



$\Delta G = +2.93$ kJ mol^{-1}

1st example: Hydrogen bonds

Challenges:

- Intermolecular interaction modeling
- Conformations

→ We need a robust and consistent generation of 3D coordinates.

Prerequisites (Hydrogen Bond Example)

How are my 3D generation scripts constructed?

1. 3D coordinates: ETKDG method
2. Explicit hydrogens (usually kékulized structures!)
3. Charge = 0 in this case, but this is extendable at will

Riniker, S.; Landrum G. A. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

Atom Selection by Substructure Matching 1

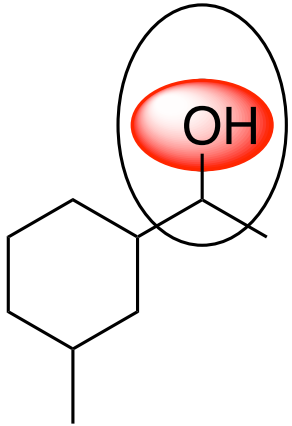
```
Acceptorfunctions = {
'carbonyl': Chem.MolFromSmiles('C=O'),
'pyridine': Chem.MolFromSmiles('c1ccncc1'),
(...)
}

Donorfunctions = {
'alcohol': Chem.MolFromSmiles('C[OH]'),
'pyrrole': Chem.MolFromSmiles('C1=CNC=C1'),
(...)
}
```

→ Sets up a functional group search in HB donor/acceptor space

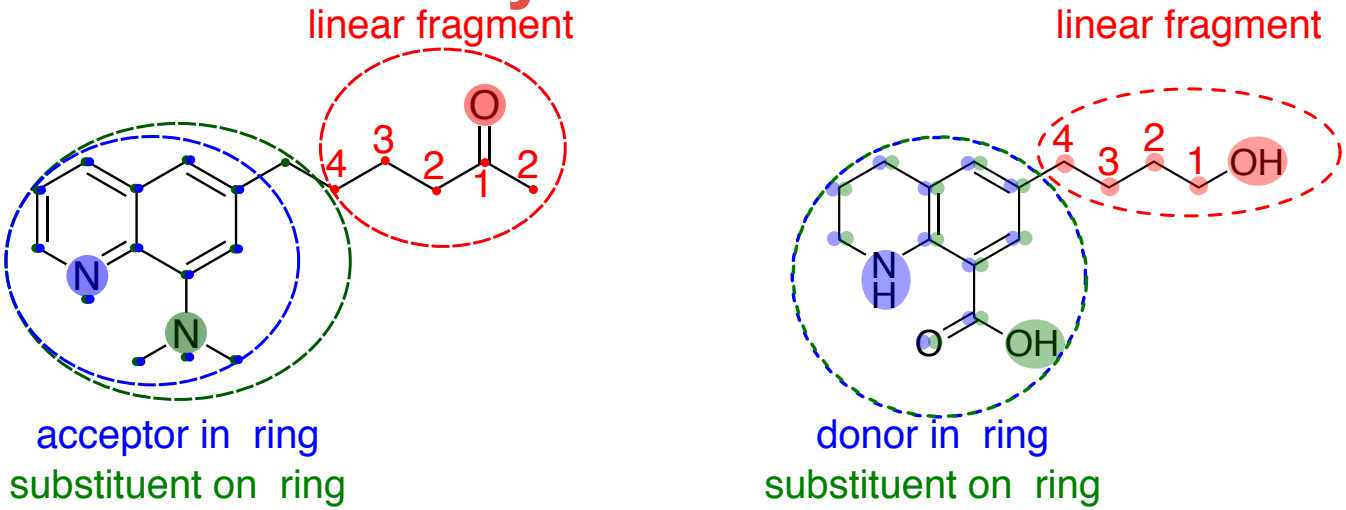
Atom Selection by Substructure Matching 2

```
# iterate over functional groups:
For function in donorfunctions:
  # get substructure matches
  (...)
  for match in matches:
      # get atoms and assign the atom index
        (...)
      for atom in atoms:
          # assign the right atom of the group
          (...)
          donoratomindex = atom.GetIdx() + 1
```



→ Helps keep order but catches only within the functional group space you set up.

Atom Selection by «Brute Force»



- All nitrogens, oxygens
- All OH, NH, SH, C#H

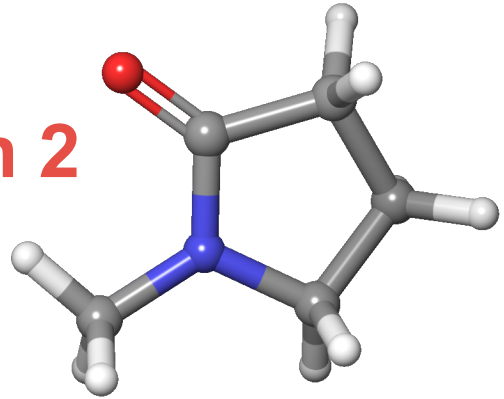
→ In the end, a mixture was applied.

Bimolecular Complex Generation 1

```
# query molecule is loaded
Hbd_indices = find_donors(mol) # as sketched previously
nat = len(mol.GetAtoms())


For i in range(nhbd_of_mol):
   # set up N complexes:
   m2 = Chem.RWMol(mol)
   # find donated H as the neighbor of the donor atom
   h_actindex = (...)
```

Bimolecular Complex Generation 2



```
# add reference molecule
# add the atoms
   for j in range(0,n_atom_ref):
        refatom = refmol.GetAtomWithIdx(j)
        posref = refmol.GetConformer().GetAtomPosition(j)
        m2.AddAtom(refatom)
        m2.GetConformer().SetAtomPosition(j + nat, posref)
   # same for the bonds
```

MMFF94s Force Field Initialization

```
# Initialize MMFF94s

fprs = Chem.MMFFGetMoleculeProperties(m2,
mmffVariant='MMFF94s')

ff= Chem.MMFFGetMoleculeForceField(m2,fprs,
ignoreInterfragInteractions=False)
```

Tosco, P.; Stiefl, G.; Landrum, G.A. *J. Cheminform.* **2014**, *6*, 4.

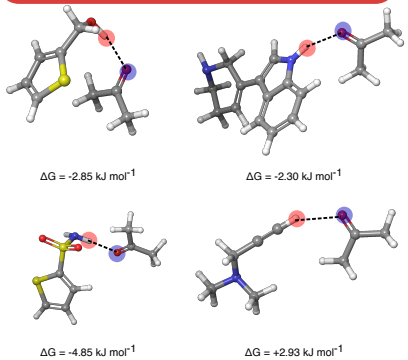
MMFF94s Force Field Optimization

```
# Constraints
ff.MMFFAddDistanceConstraint(h_actindex,
hbaindex, False, 1.98, 2.00, 9999999.)
ff.MMFFAddAngleConstraint(hbaindex,h_actindex,
hbdindex,False, 178., 179.9, 999.)
# Minimize
ff.Minimize(maxIts=100)
```

→ Here are the **parameters** to play with!

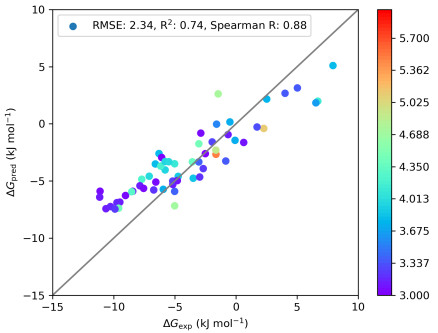
Hydrogen Bond Donor Results

QC data → ML training → predictions



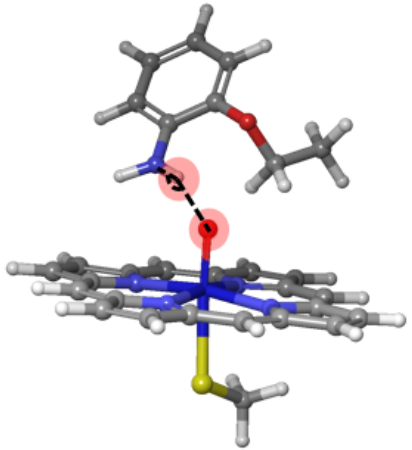
$$\begin{matrix} f \\ f_* \end{matrix} \sim N\left(0 \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*X_*) \end{bmatrix}\right)$$

(Gaussian Process Regression)



Bauer, C. A.; Schneider, G.; Göller, A. H. *J. Cheminform.* **2019**, *11*, 59.

But How Do We Get the QC Data?



Patrik Rydberg, Lars Olsen (SMARTCYP papers)

Foscato M.; Venkatraman V.; Jensen, V. R.
J. Chem. Inf. Model (just accepted),
DOI:10.1021/acs.jcim.9b00516

2nd example: Ames mutagenicity P450 model

Challenges:

- Activation barrier
- Metal involved
- Conformations

→ There is hardly any way to generate robust 3D coordinates.

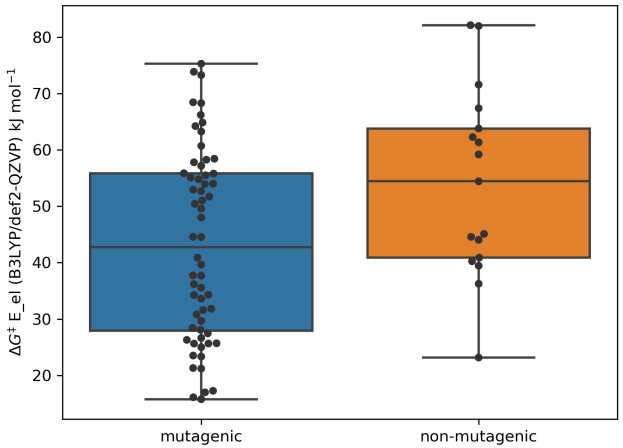
P450 Cpd1 Complex Generation

```
# exactly the same method as for the H-bonds except
fix the cpd1 atoms due to force field collapse for
[Fe^V=O] moiety:


for fixat_idx in range (nat,nat+cpd1nat):
    ff.UFFAddPositionConstraint(fixat_idx, 0.01,
99.)
```

→ Do not move the cpd1 part during optimization (it would go wrong), just optimize the primary aromatic amine.

P450 Cpd1 Primary Aromatic Amines DFT Results



- 77 data points: 60 mutagens, 17 non-mutagens
- High-level DFT (hybrid functional/QZVP basis set)
- Optimized using dispersion corrected DFT
- Mutagenic compounds with lower barriers on average.

→ Barrier height with potential as a QC descriptor.

Conclusions

- Link of QC and Machine Learning world requires consistent 3D coordinate generation.
- RDKit is the software of choice.
- Successful contribution to H-bonding.
- RDKit can even «handle» metals (coordinate generation is meant as preliminary work for DFT input anyway).

Thanks

- Prof. Johannes Kirchmair
- Kirchmair Group
- Andreas Göller (Bayer) and Gisbert Schneider (ETH Zurich)