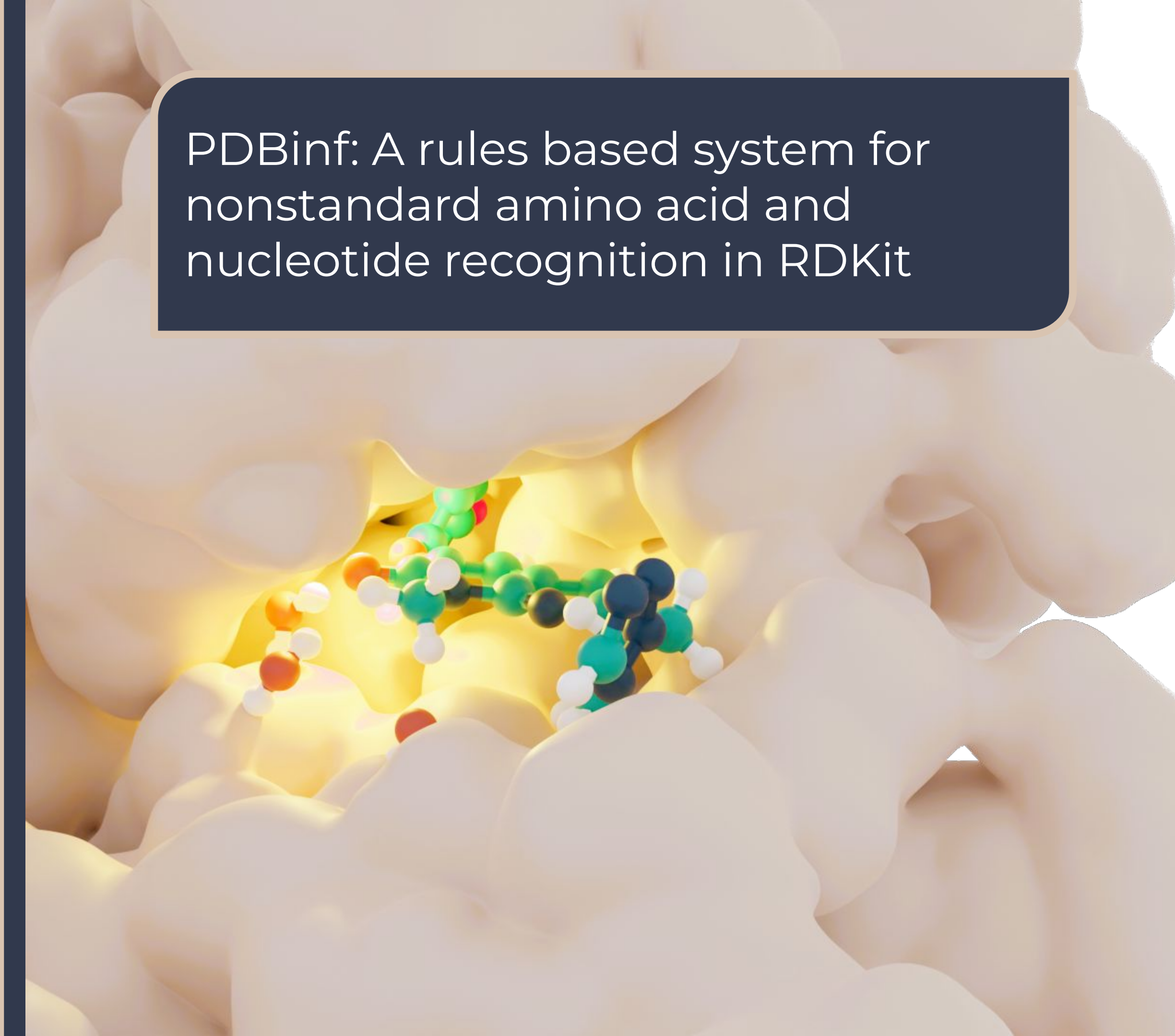
Open
Free
Energy

free software free energies

https://openfree.energy/



PDBinf: A rules based system for nonstandard amino acid and nucleotide recognition in RDKit

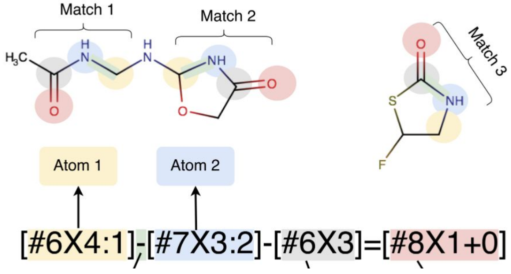
Problem statement

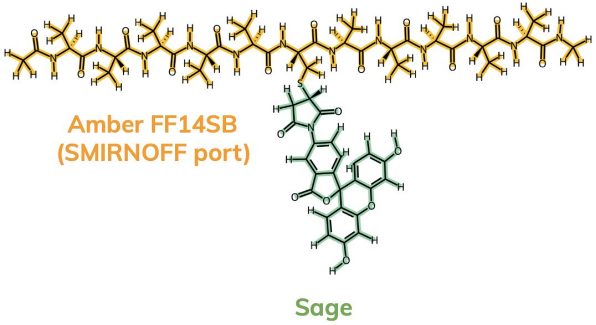
Loading proteins/biopolymers into rdkit sort of works, except for nonstandard amino acids etc.

Having “cheminformatically correct” representations is important for analysis, but also the next generation of force fields, which are using SMARTS patterns to assign parameters.

E.g. if I want to assign force field parameters to a PTM.

open forcefield



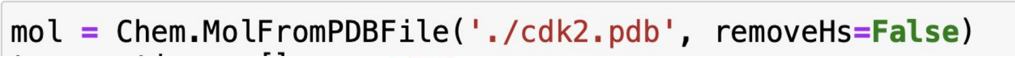


PDBinf for bond order assignment

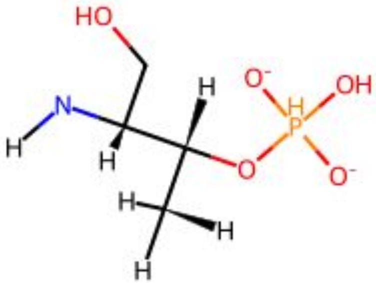
PDBinf is a little package for reading PDB(x) files, and trying to apply extra information from monomer templates to build cheminformatically-accurate models.

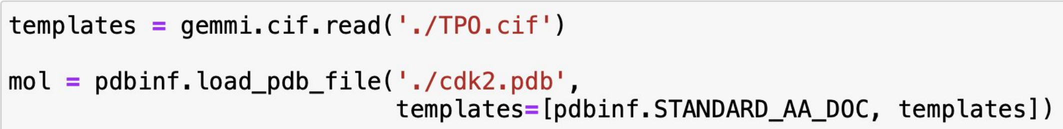
This can handle standard amino acids/RNA/DNA... And also all of the nonstandard residues in the PDB chemical component dictionary (CCD).

Templates come from this PDB CCD. The dynamic nature of the CCD make it unsuitable for the current “compile everything” approach.



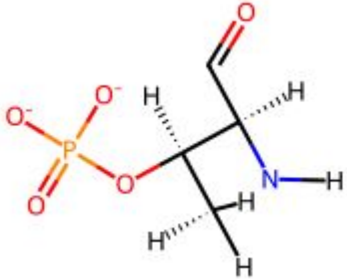
```
mol = Chem.MolFromPDBFile('./cdk2.pdb', removeHs=False)
```





```
templates = gemmi.cif.read('./TPO.cif')

mol = pdbinf.load_pdb_file('./cdk2.pdb',
                           templates=[pdbinf.STANDARD_AA_DOC, templates])
```



MolHash for amino acid lookup

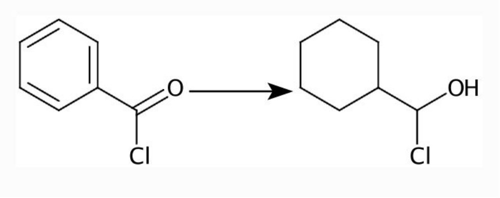
Each monomer (residue) in the biopolymer is assigned a residue name. These names are used to lookup the correct template.

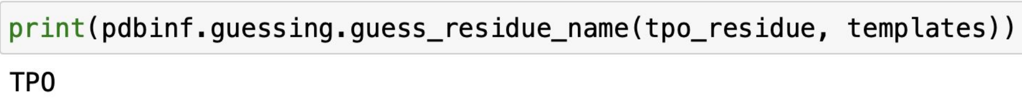
It is however traditional to mislabel things in PDB files.

`rdMolHash` with element graphs turns a nasty graph matching problem into an easy string lookup problem.

See:

https://www.nextmovesoftware.com/talks/OBoyle_MolHash_ACS_201908.pdf





```
print(pdbinf.guessing.guess_residue_name(tpo_residue, templates))
```

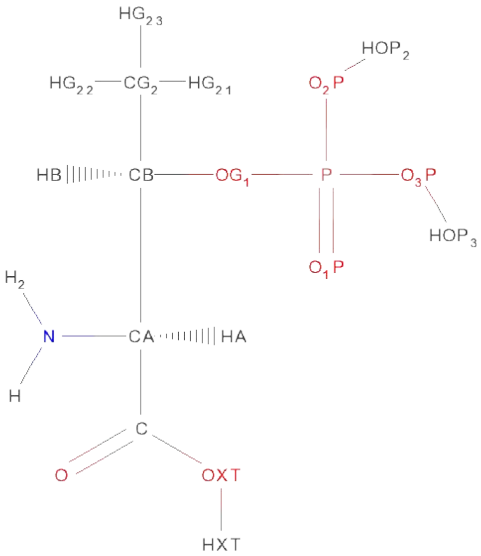
```
TPO
```

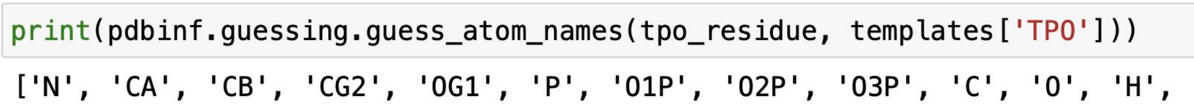
SMARTS lookup to assign atom names

Within a residue, each atom has a unique name, occasionally these are very useful.

Again these are often mislabelled, making some things more difficult than they should be.

SMARTS w/ atom mapping can allow you to rename atoms to their canonical names.





```
print(pdbinf.guessing.guess_atom_names(tpo_residue, templates['TPO']))
['N', 'CA', 'CB', 'CG2', 'OG1', 'P', 'O1P', 'O2P', 'O3P', 'C', 'O', 'H',
```

Conclusion

PDBinf is a hacky package for loading PDB(x) files into RDKit that I've put together.

It can:

- Load data from PDB(x) files into rdkit (with bonds assigned from templates, not guessed from geometry)
- Figure out what your mislabelled residues are.
- Figure out what your atom names should be.

https://github.com/OpenFreeEnergy/pdbinf

Xyz2mol module serves a similar function, but in contrast:

- Is a stack of heuristics so will work without templates
- Rules based system offers different tradeoff in precision vs recall