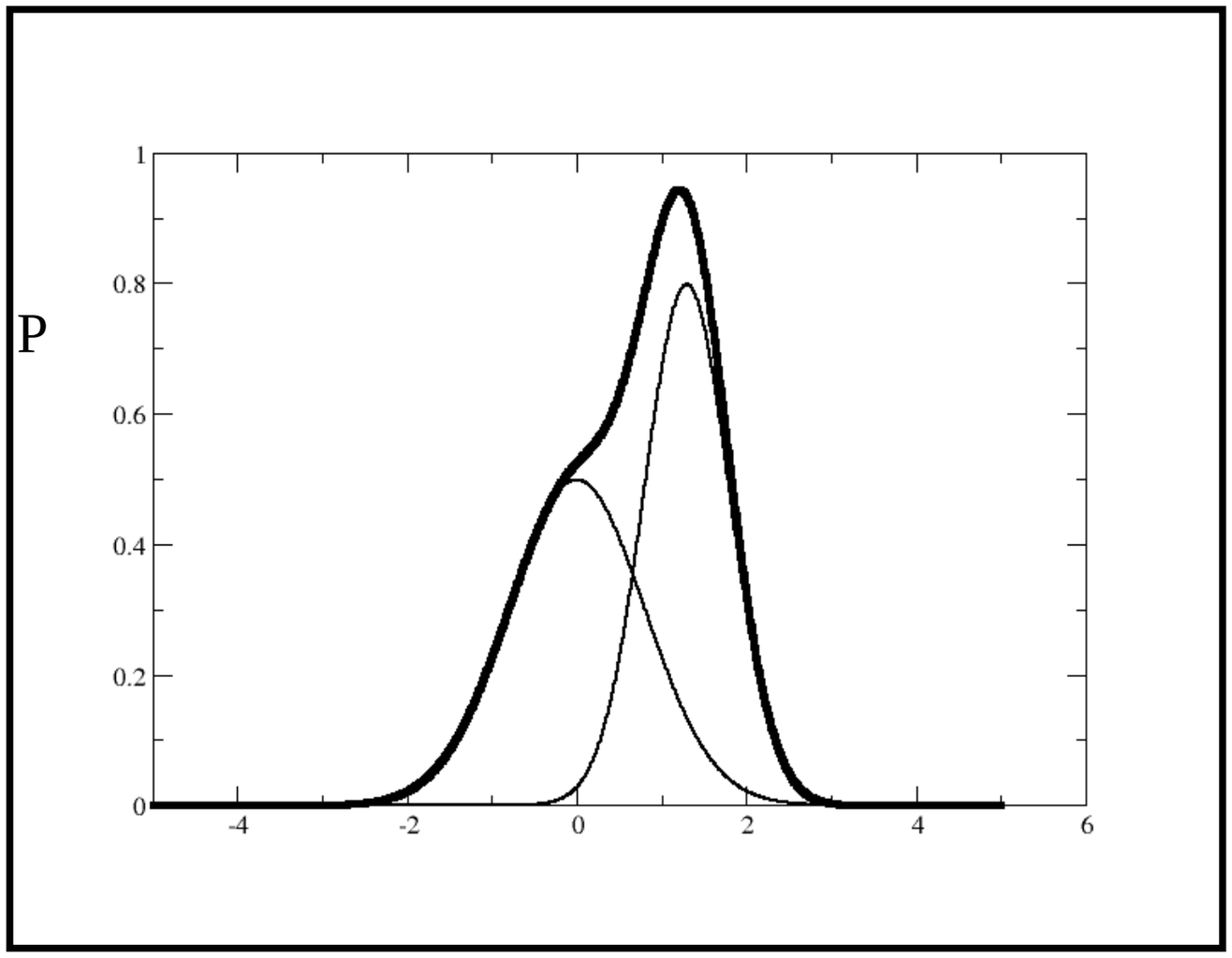
Dictionary of ligands

- Garib Murshudov
- MRC-LMB, Cambridge

Contents

Why restraints:
Two atoms ideal case

- Distance between atoms 1.3Å. B values 20 and 50



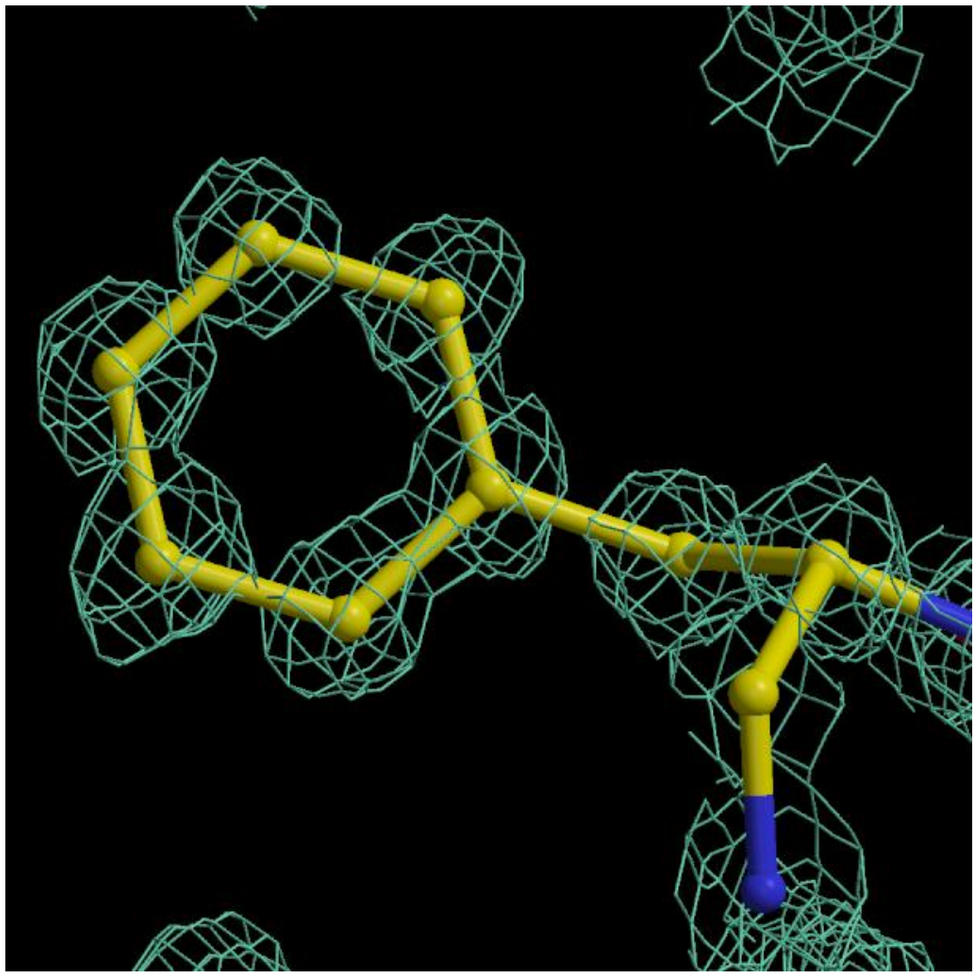
- Thin lines – single atoms
- Bold line - sum of the two atoms

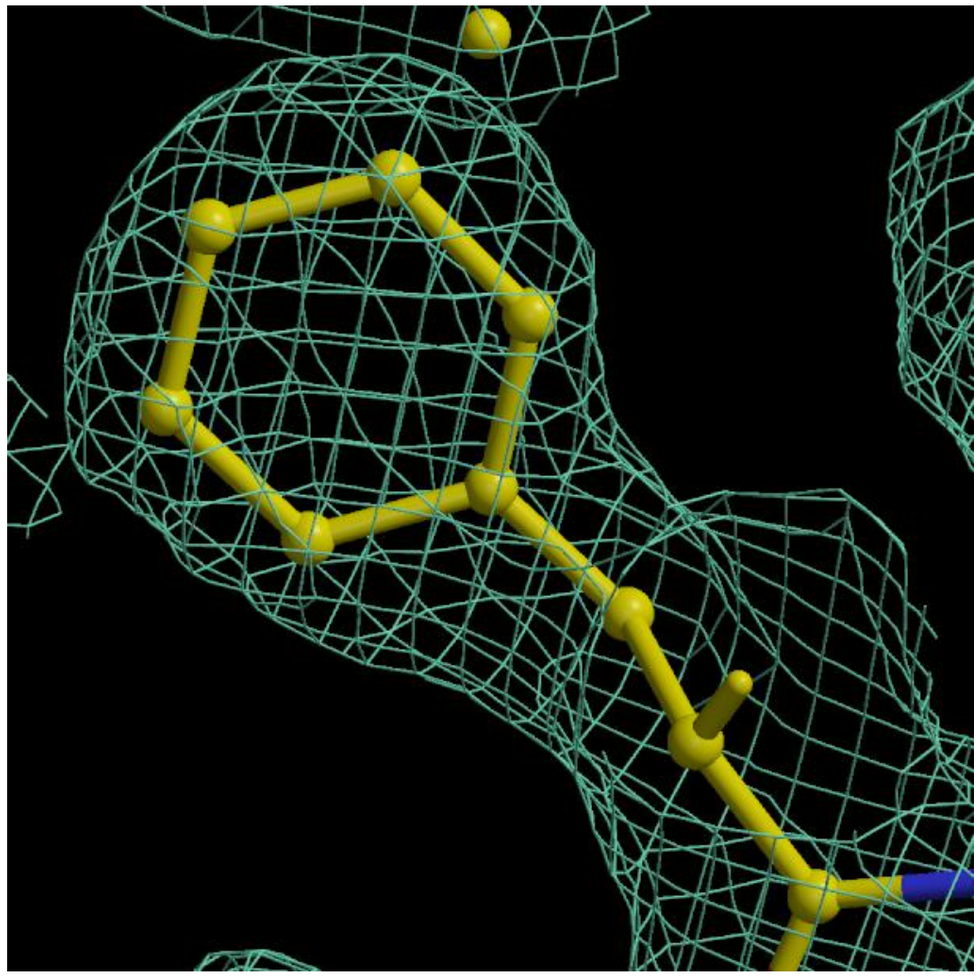
x

Chemical information: Phe at two different resolutions

- 0.88 Å

2 Å and High mobility





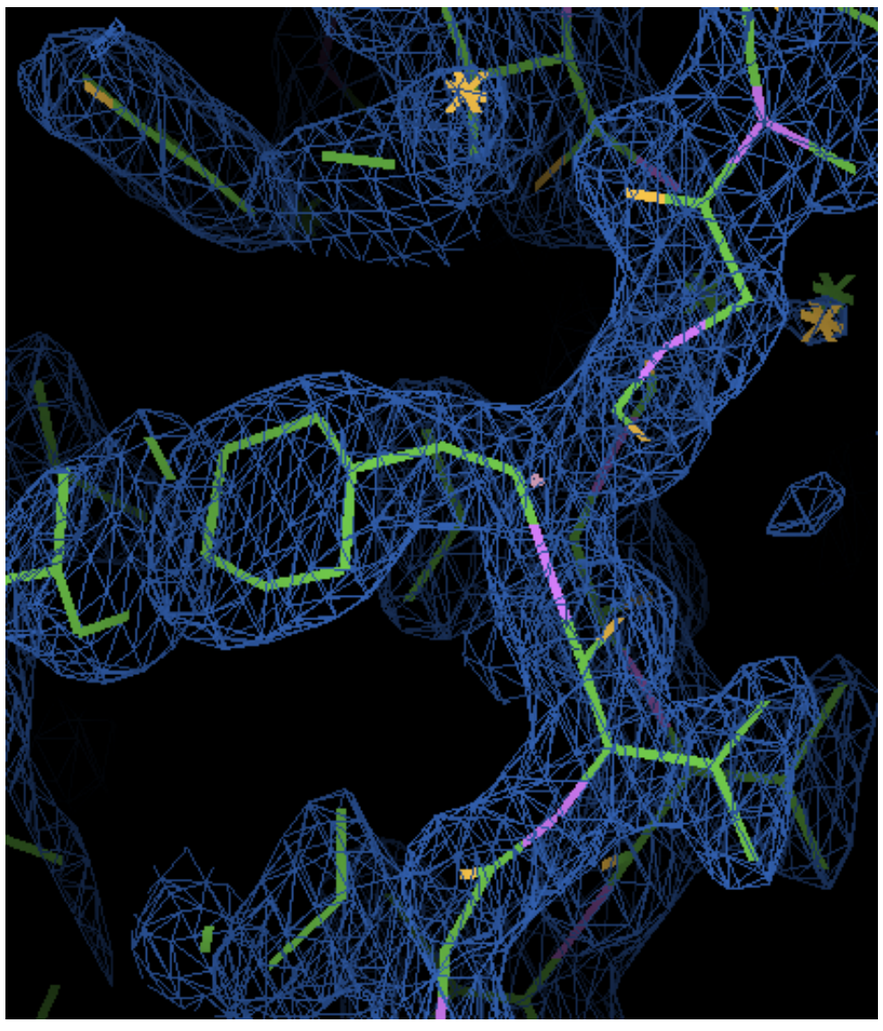
Role of restraints

- When atoms have high B values and/or data are at low resolution then electron density may not show separate peaks
- If restraints would not be used then chemistry of molecule would be unreasonable.
- Role of restraints is that to retain chemistry of atoms and at the same time describe electron density optimally.
- If atoms are close to each other it is unlikely that they will have hugely different B values

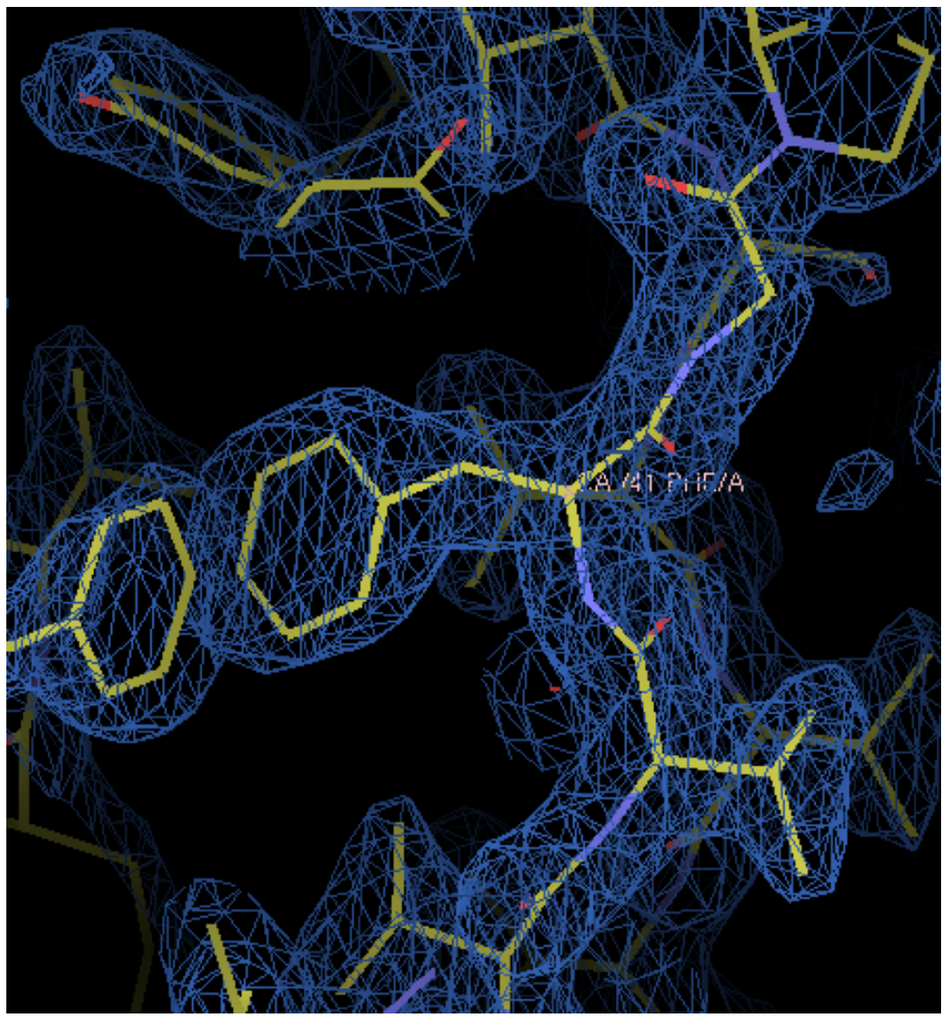
Example

- Data - 1.9A

Unrestrained



Restrained



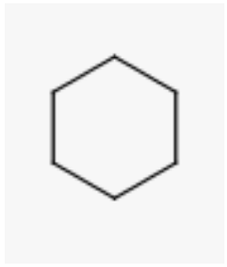
Using restraints

Standard dictionary has description of more than10000 small molecules. If one of them is in your crystal then the will be used automatically.

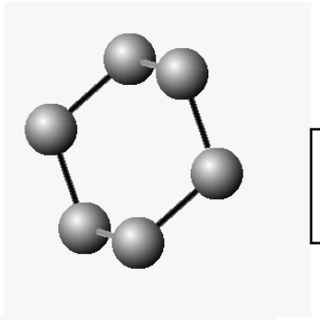
What happens if you have a ligand that is not in the dictionary. Then it is your responsibility to create chemically sensible description.

Before starting to create a description you need to study bonding structure of your ligand.

2D



3D



These two molecules will refine very differently (oviously)

SMILES

•SMILES notation is the most popular notation and almost all computational chemical websites, programs use this notation. They can read and write SMILES.

•It is based on several simple rules. Full description of SMILES can be find from daylight websites.

• http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html

•SMILES stands for Simplified Molecular Input Line Entry System.

•It is concise and widely spread. It is very easy to learn. It was originally designed for manual input using text only editors. SMILES has become as a standard and it is a useful thing to know about.

SMILES

•SMILES uses several very simple rules (these rules are sufficient to generate SMILES from structure and structure from SMILES).

•Rules:

• Atomic symbols used for atoms

• Hydrogen atoms as a rule are implicit. They are deduced using valence information about atoms

• Neighbouring atoms stand one after another

• Single, double, triple and aromatic bonds are denoted using “-”, “=”, “#” and “:” respectively. Single and aromatic bonds are usually not shown.

•Branches represented by parentheses

•Cycles are added by using matching digits on connecting atoms

• Aromatic atoms are denoted using lower cases.

•These rules are sufficient to describe most of the cases. Let us consider some examples

PDB Chemical Component Dictionary

wwPDB maintains dictionary descriptions for all unique chemical components from the structures deposited with the PDB

- Name, synonyms, formula, SMILES, …
- Atoms and bonds
- Ideal and representative coordinates

Each new component assigned a unique 3-letter identifier

- Release coincides with the release of the parent PDB entry

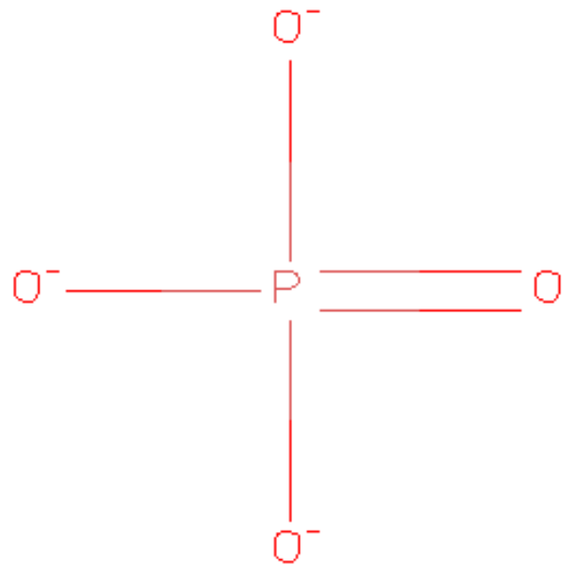
Chemical Component Dictionary

The Chemical Component Dictionary[a] is as an external reference file describing small molecule components found in PDB entries. This dictionary contains detai descriptions for standard and modified amino acids/nucleotides, small molecule solvent molecules. Each chemical definition includes descriptions of chemical p as stereochemical assignments, chemical descriptors (SMILES & InChI), syste names, and idealized coordinates (generated using Molecular Networks' Corina, issues, OpenEye's OMEGA).

The dictionary is organized by the 3-character alphanumeric code that PDB assi chemical component. New chemical component definitions appear in the dictiona entries in which they are observed are released in the PDB archive; consequentl dictionary is updated with each weekly PDB release. The dictionary is regularly

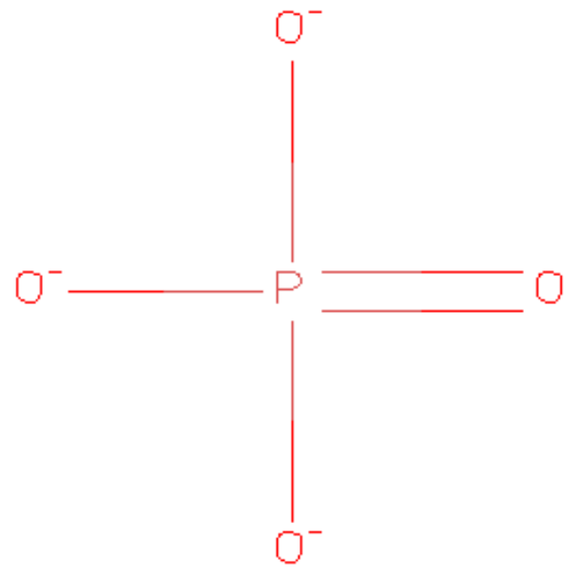
PDB Chemical Component Dictionary

data_PO4
#
_chem_comp.id PO4
_chem_comp.name "PHOSPHATE ION"
_chem_comp.type NON-POLYMER
_chem_comp.pdbx_type HETAI
_chem_comp.formula "O4 P"
_chem_comp.mon_nstd_parent_comp_id ?
_chem_comp.pdbx_synonyms ?
_chem_comp.pdbx_formal_charge -3
_chem_comp.pdbx_initial_date 1999-07-08
_chem_comp.pdbx_modified_date 2011-06-04
_chem_comp.pdbx_ambiguous_flag N
_chem_comp.pdbx_release_status REL
_chem_comp.pdbx_replaced_by ?
_chem_comp.pdbx_replaces IPS
_chem_comp.formula_weight 94.971
_chem_comp.one_letter_code ?
_chem_comp.three_letter_code PO4
_chem_comp.pdbx_model_coordinates_details ?
_chem_comp.pdbx_model_coordinates_missing_flag N
_chem_comp.pdbx_ideal_coordinates_details ?
_chem_comp.pdbx_ideal_coordinates_missing_flag N
_chem_comp.pdbx_model_coordinates_db_code 1IXG
_chem_comp.pdbx_subcomponent_list ?
_chem_comp.pdbx_processing_site EBI



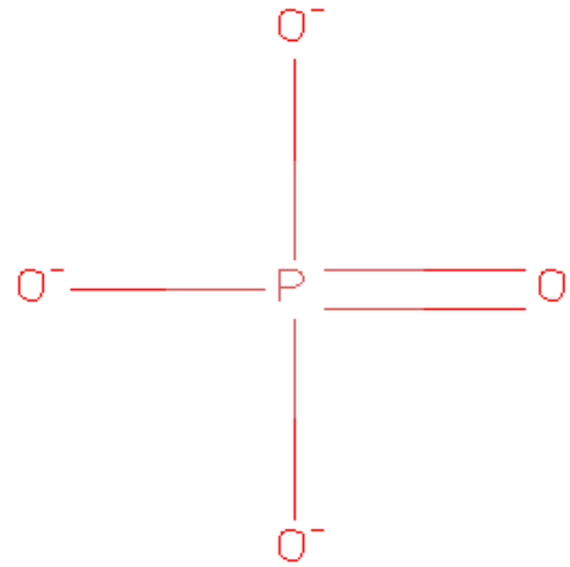
PDB Chemical Component Dictionary

loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.alt_atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.charge
_chem_comp_atom.pdbx_align
_chem_comp_atom.pdbx_aromatic_flag
_chem_comp_atom.pdbx_leaving_atom_flag
_chem_comp_atom.pdbx_stereo_config
_chem_comp_atom.model_Cartn_x
_chem_comp_atom.model_Cartn_y
_chem_comp_atom.model_Cartn_z
_chem_comp_atom.pdbx_model_Cartn_x_ideal
_chem_comp_atom.pdbx_model_Cartn_y_ideal
_chem_comp_atom.pdbx_model_Cartn_z_ideal
_chem_comp_atom.pdbx_component_atom_id
_chem_comp_atom.pdbx_component_comp_id
_chem_comp_atom.pdbx_ordinal
PO4 P P P 0 1 N N N 29.995 23.516 13.249 0.000 0.000 0.000 P PO4 1
PO4 O1 O1 O 0 1 N N N 31.092 22.988 14.164 0.000 -1.288 -0.911 O1 PO4 2
PO4 O2 O2 O -1 1 N N N 30.404 24.896 12.647 0.000 1.288 -0.911 O2 PO4 3
PO4 O3 O3 O -1 1 N N N 29.646 22.518 12.126 -1.288 0.000 0.911 O3 PO4 4
PO4 O4 O4 O -1 1 N N N 28.727 23.744 14.161 1.288 0.000 0.911 O4 PO4 5



PDB Chemical Component Dictionary

```
loop_
_chem_comp_bond.comp_id
_chem_comp_bond.atom_id_1
_chem_comp_bond.atom_id_2
_chem_comp_bond.value_order
_chem_comp_bond.pdbx_aromatic_flag
_chem_comp_bond.pdbx_stereo_config
_chem_comp_bond.pdbx_ordinal
PO4 P O1 DOUB N N 1
PO4 P O2 SING N N 2
PO4 P O3 SING N N 3
PO4 P O4 SING N N 4
#
loop_
_pdbx_chem_comp_descriptor.comp_id
_pdbx_chem_comp_descriptor.type
_pdbx_chem_comp_descriptor.program
_pdbx_chem_comp_descriptor.program_version
_pdbx_chem_comp_descriptor.descriptor
PO4 SMILES           ACDLabs              10.04 "[O-]P([O-])([O-])=O"
PO4 SMILES_CANONICAL CACTVS               3.341 "[O-][P]([O-])([O-])=O"
PO4 SMILES           CACTVS               3.341 "[O-][P]([O-])([O-])=O"
PO4 SMILES_CANONICAL "OpenEye OEToolkits" 1.5.0 "[O-]P(=O)([O-])[O-]"
PO4 SMILES           "OpenEye OEToolkits" 1.5.0 "[O-]P(=O)([O-])[O-]"
PO4 InChI            InChI                1.03  "InChI=1S/H3O4P/c1-5(2,3)4/h(H3,1,2,3,4)/p-3"
PO4 InChIKey         InChI                1.03  NBIIXXVUZAFLBC-UHFFFAOYSA-K
```



CCP4 library of restraints (Refmac monomer dictionary)

Contains monomers from Chemical Component Dictionary

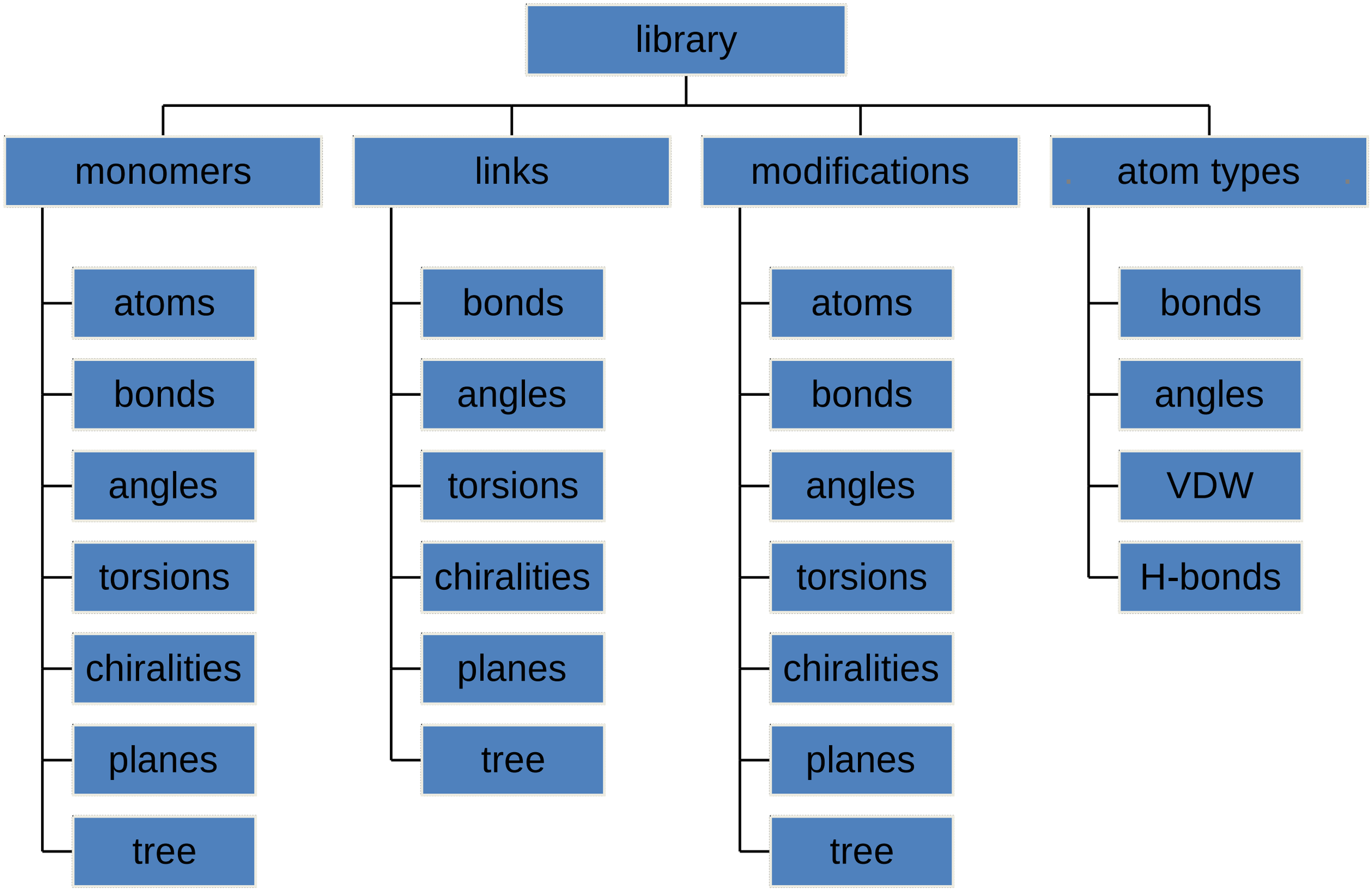
- the same monomer IDs
- the same atom IDs
- synchronised with CCD at the CCP4 releases
- recently added to CCD monomers are absent
- missing monomers can be added to user's additional library

Less annotations than in Chemical Component Dictionary

Accent is made on complete description of restraints

- Refmac and Coot are able to refine structures downloaded from PDB

CCP4 library of restraints



Monomers

A *monomer* entry describes an individual compound

CCP4 library contains:
- All standard amino acids
- All standard nucleic acids
- Common sugars
- Other organic and inorganic compounds 10,500

Monomer and atom names are as in
PDB Chemical Component Dictionary

Jligand and Links

CCP4 monomer library: modfications and links

New link description

Modifications and links

The idea of this mechanism is that
- while *monomer* records describe individual compounds
- *modifications* and *links* describe changes resulted from chemical reactions

Modification formalism allows to change a monomer

Link formalism allows to join modified monomers together

Generic links for peptides

Generic peptide modification "DEL-OXT":

```
NH3-(C-CH3)-COO   -> NH3-(C-CH3)-CO
```

Generic peptide modification "NH1":

```
NH3-(C-CH3)-COO   -> NH-(C-CH3)-COO
```

Generic peptide link "TRANS":

```
NH3-(C-CH3)-CO   +   NH-(C-CH3)-COO -> NH3-(C-CH3)-CO–NH-(C-CH3)-CO
```

Thiese define: bond length, angles and a plane associated with the bond C-N

Specialised monomers vs. generic links

-CO-NH-()-CO-NH-()-CO-NH-()-CO-NH-()-CO-NH-()-CO-NH-()-CO-NH-

Specialised monomers:

NH3-()-CO2
-(trans)-NH-()-CO2
-(cis)-NH-()-CO2
NH3-()-CO-(trans)-
NH3-()-CO-(cis)-
-(trans)-NH-()-CO-(trans)-
-(trans)-NH-()-CO-(cis)-
-(cis)-NH-()-CO-(trans)-
-(cis)-NH-()-CO-(cis)-

---

9 versions
×
20 aminoacids
=
180 library entries

Generic links:

NH3-()-CO2

)-CO-(trans)-NH-(
)-CO-(cis)-NH-(
)-CO2
NH2-(

---

20 aminoacids
+
2 links
+
7 modifications
=
29 library entries

Links for peptides

generic

– peptide-peptide: `TRANS, CIS`

generic from one side

– peptide-PRO: `PTRANS, NMTRANS, PCIS, NMCIS`
– C-terminal modification: `FOR_C-C, DFO_N-C, STA_N-C, ...`
– N-terminal modification: `FOR_C-N, ACE_C-N, DFO_C-N, ...`
– pyranose-(ASP, THR, SER): `NAG-SER, NAG-THR, NAG-ASN`

specialised

– S-S bridge: `CYS-CYS`
– pyranose-peptide: `XYS-SER, XYS-THR, XYS-ASN, ...`
– metal-peptide: `ZN-CYS, FE-CYS`

Standard modifications and links (generic and specialised)

CCP4 library contains modifications for:

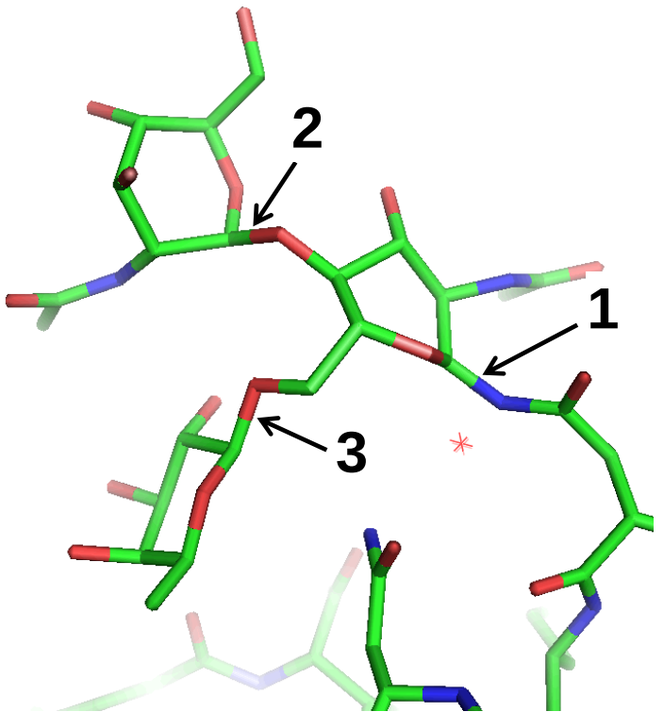
- terminal peptides and nucleotides
- methylated nucleotides
- deprotonated states

CCP4 library contains links and corresponding modifications for:

- polypeptide chains (CIS,TRANS), S-S bridges
- polynucleotide chains
- glycosylated proteins

Generic links for sugars

2xmb



For typical glycosylation cases

- necessary modifications and links are there in the standard ccp4 library

- by default REFMAC uses these library descriptions and does not need any additional instructions

NAG – NAG – ASN
|
FUL

FUL = Beta-L-Fucose
NAG = N-Acetyl-D-Glucosamine

Standard links used here:

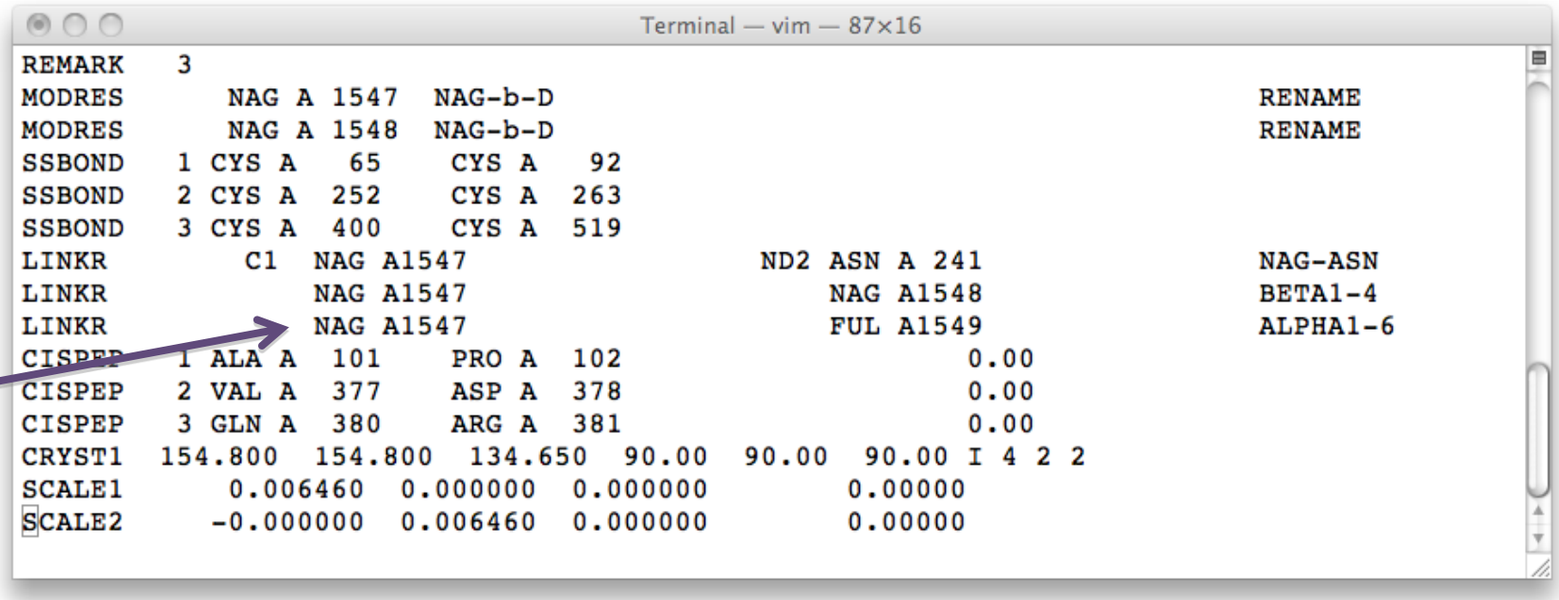
(1) "NAG-ASN"
(2) "BETA1-4"
(3) "ALPHA1-6"

Sugar links: refmac checkpoints

✓ refmac terminated normally

✓ output pdb-fail contains expected LINKR records, e.g.

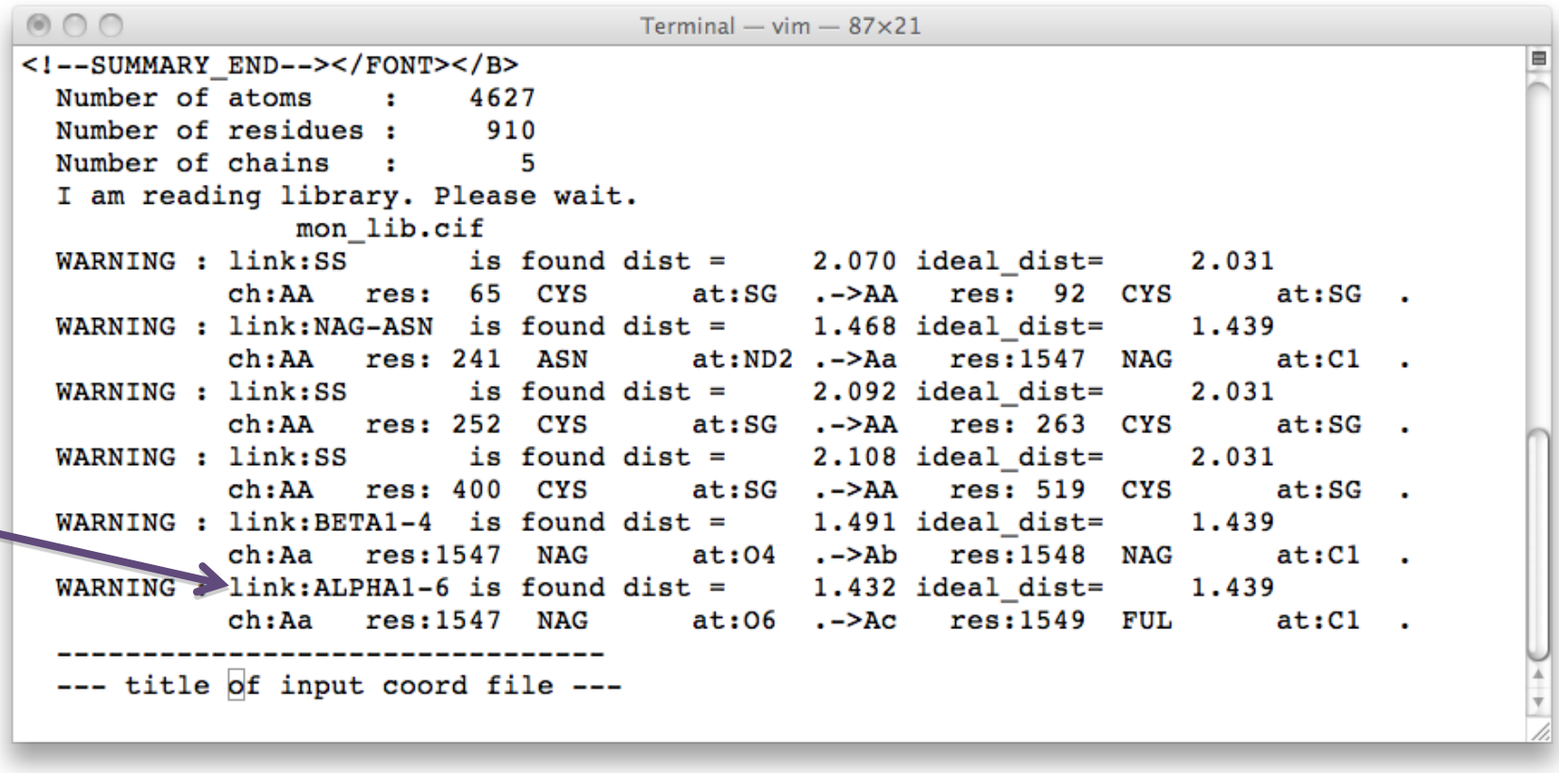
```
LINKR...
...NAG A1547  FUL
A1549...
...ALPHA1-6
```



✓ log-file contains warnings saying e.g. that

```
... link:ALPHA1-6 is
        found
... res:1547  NAG  ...
... res:1549  FUL  ...
```

(WARNING = OK)



User-defined links

When new link descriptions are needed:

side chain – side chain (e.g. TYR – TYR on the figure)

side chain – main chain (e.g. LYS – Ubiquitin)

side chain – ligand (e.g. LYS – PLP)

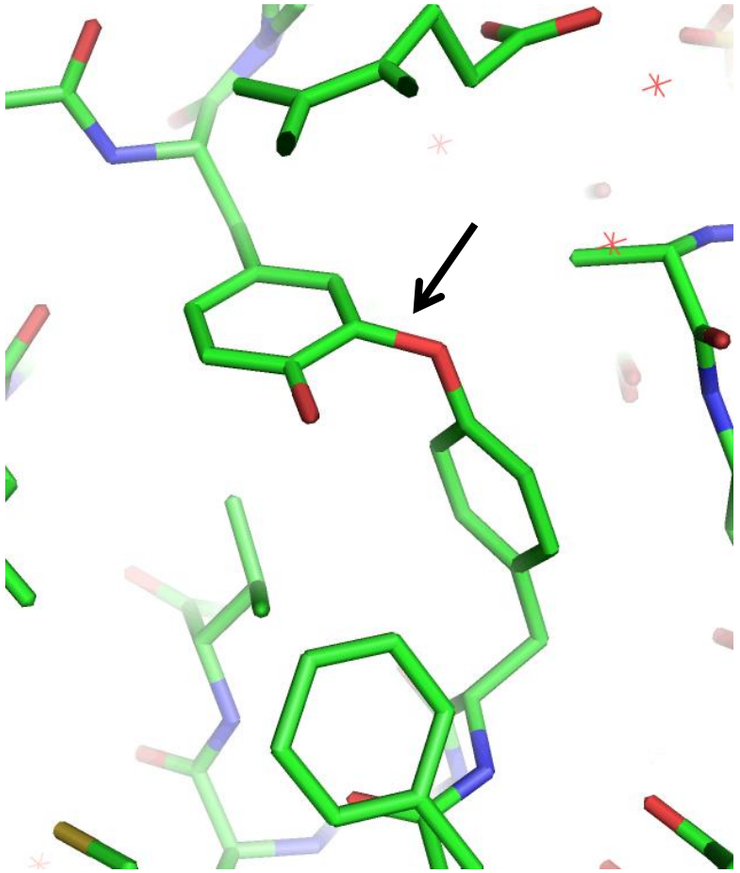
JLigand:

– new GUI for LIBCHECK

– descriptions of monomers (functionality of SKETCHER)

– descriptions of links and corresponding modifications

TYR–TYR covalent link in
M. tuberculosis Hemoglobin O
PDB id 1ngk



CCP4 monomer library: modfications and links

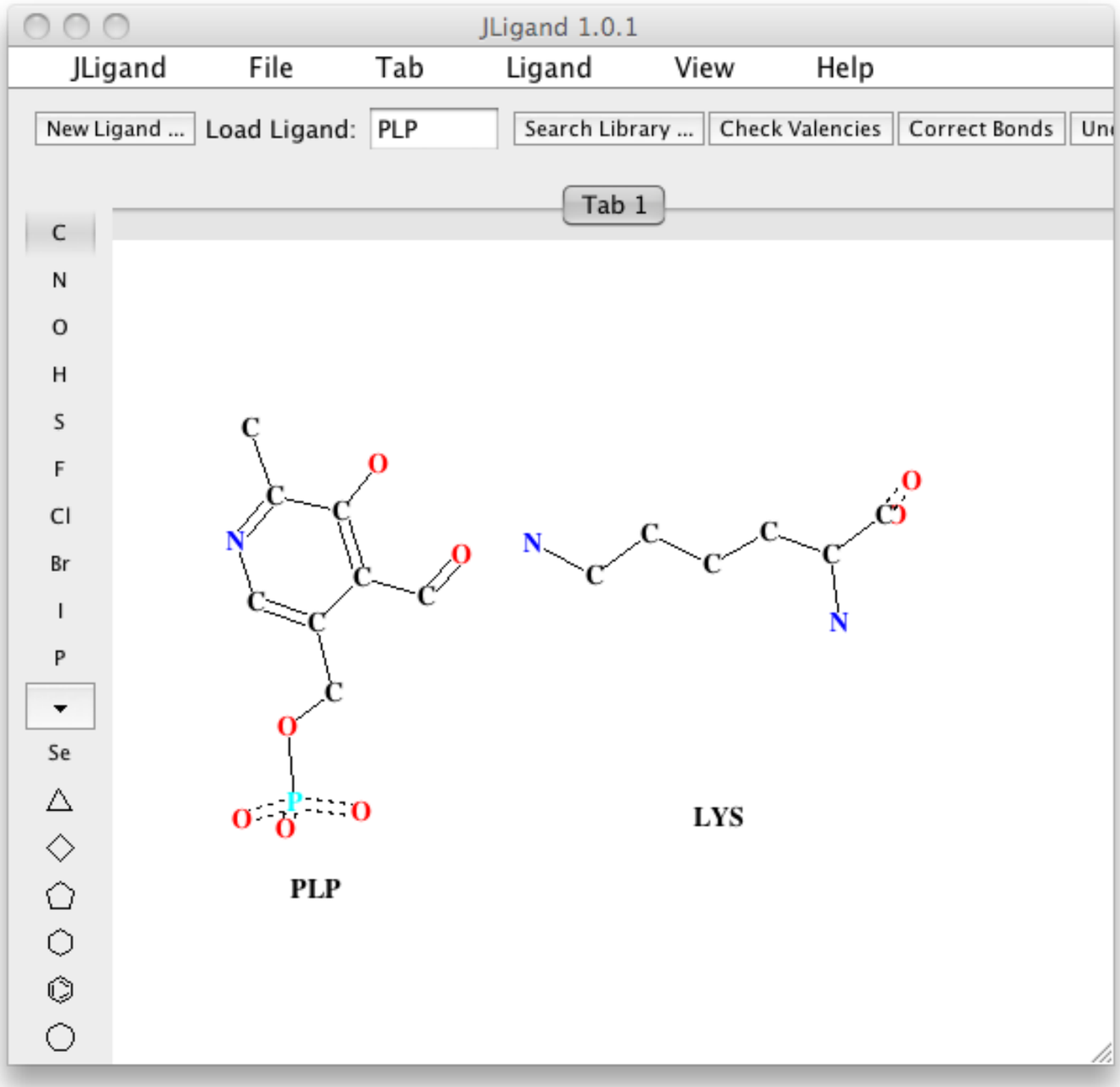
New link description

New link

Example:
 – covalent linkage between LYS and Pyridoxal phosphate (PLP).
 – describes PLP forming internal aldimine in aminotransferases.

Given:
- descriptions of LYS and PLP
  from the standard library

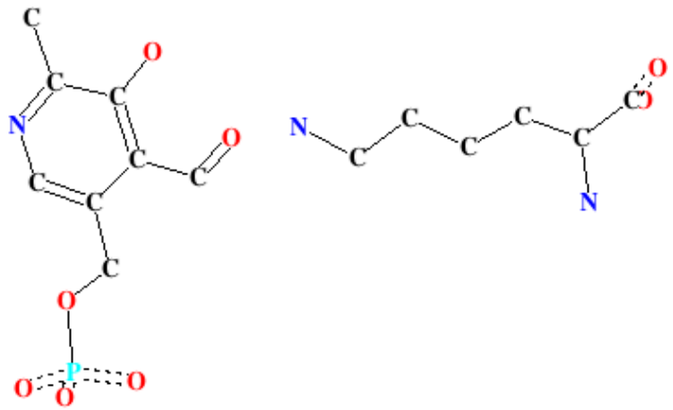
Needed:
- additional library file with
  the description of link LYS–PLP



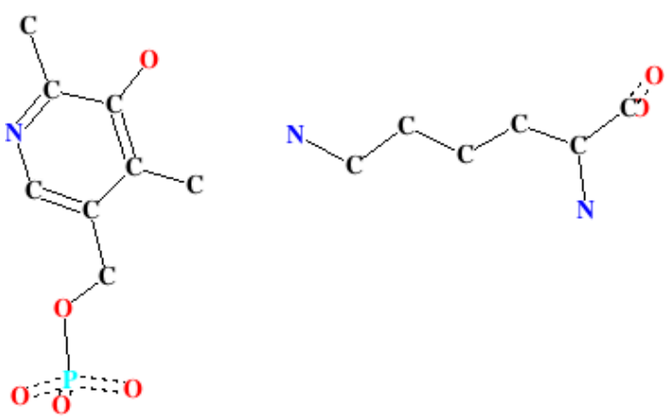
Creating a new link, as seen in JLigand GUI

The two monomers are in effect reacted in silico
Hydrogen atoms are dealt with automatically*)

*) it is also possible to visualise H-atoms and deal with them explicitly

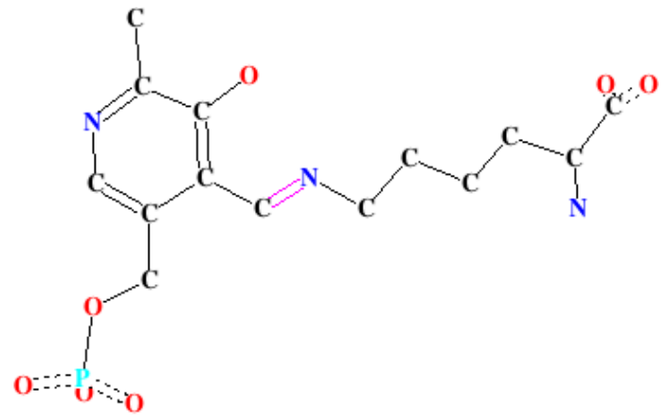


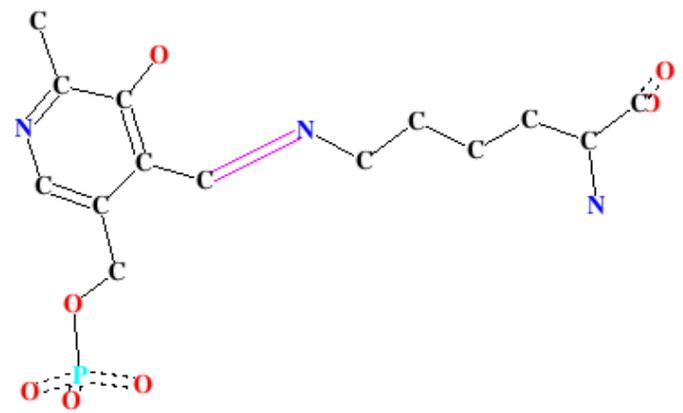
leaving O is removed



double bond C=N is added

the new compound
is regularised



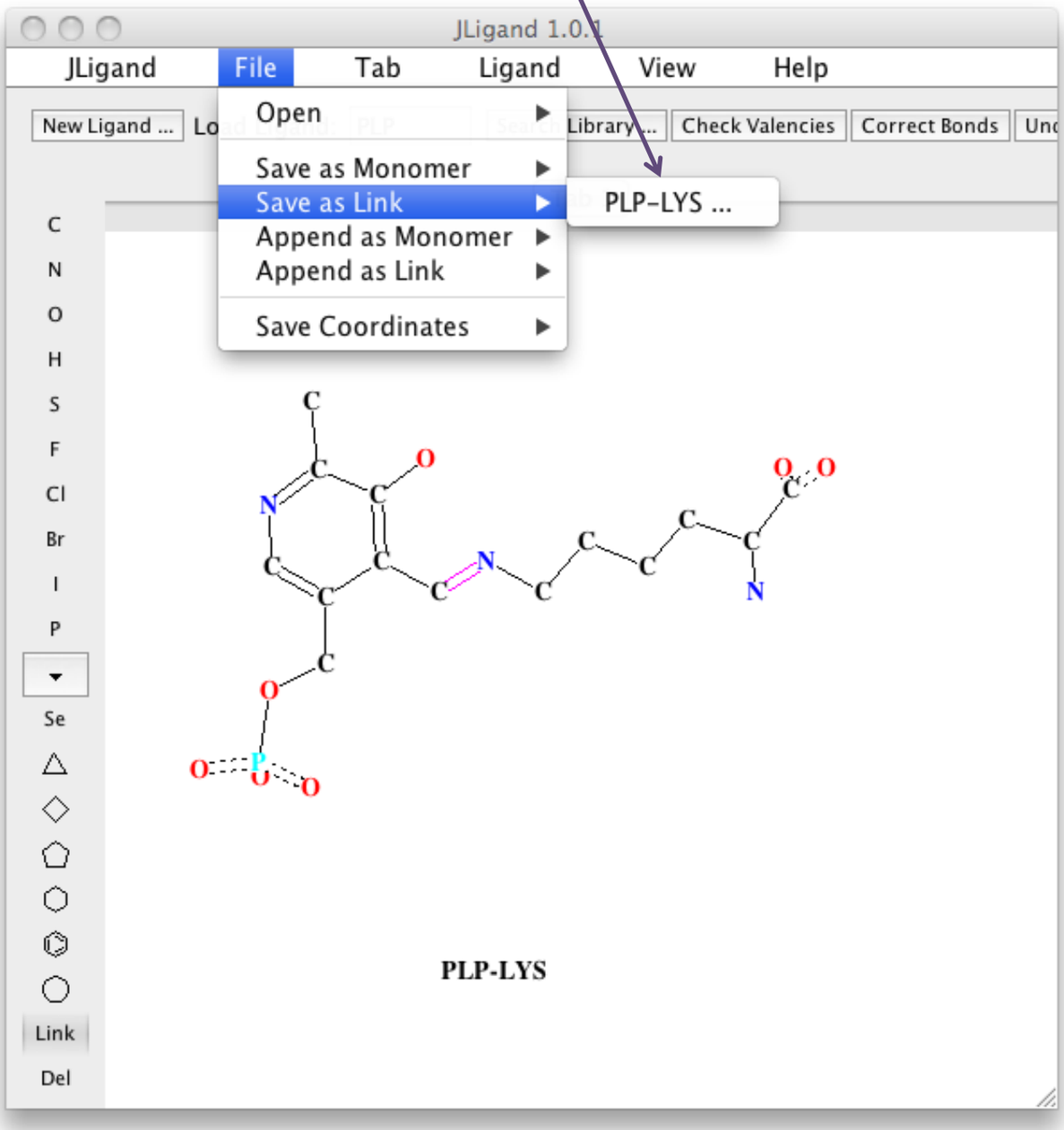


The new link, "file view"

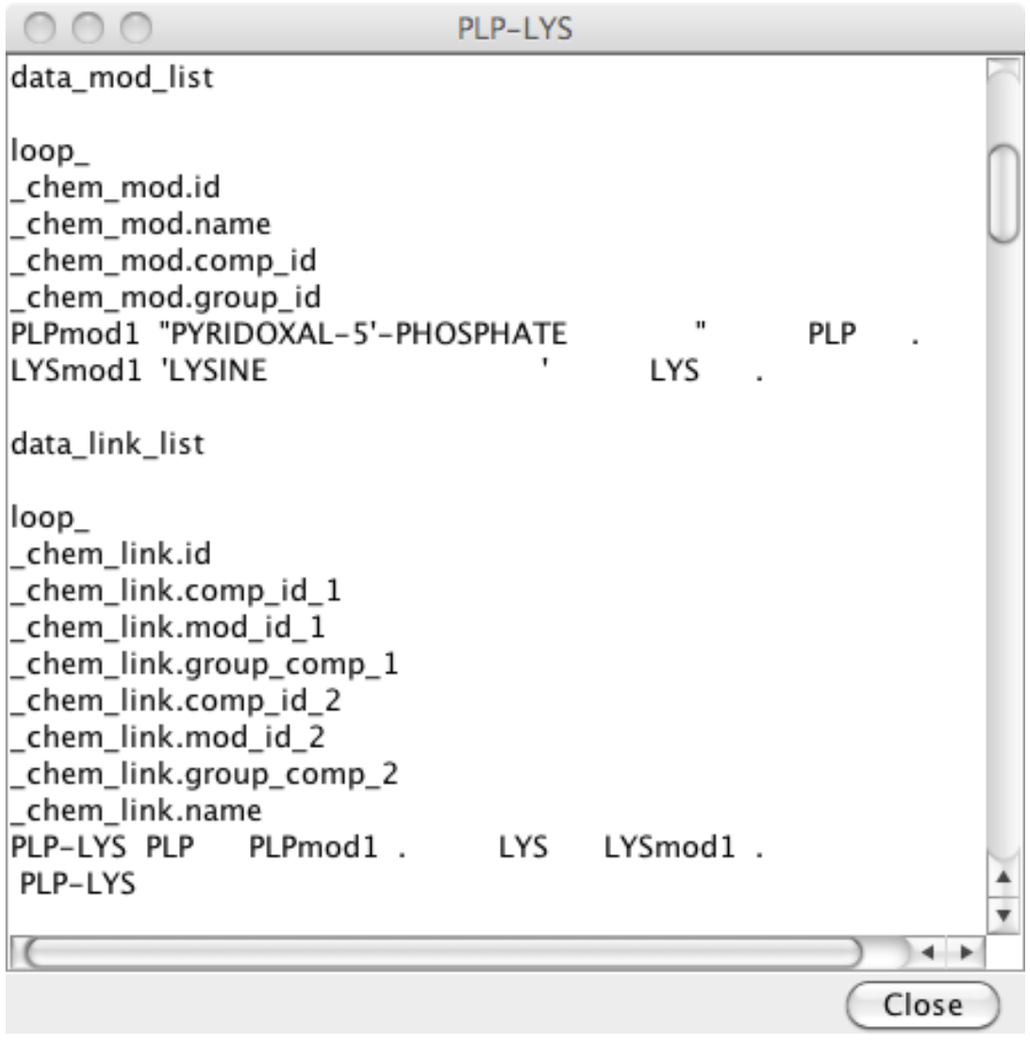
To save into CIF-file (additional library)

Contents:
(1) modification "PLPmod1"
(2) modification "LYSmod1"
(3) link "PLP-LYS"

No monomers

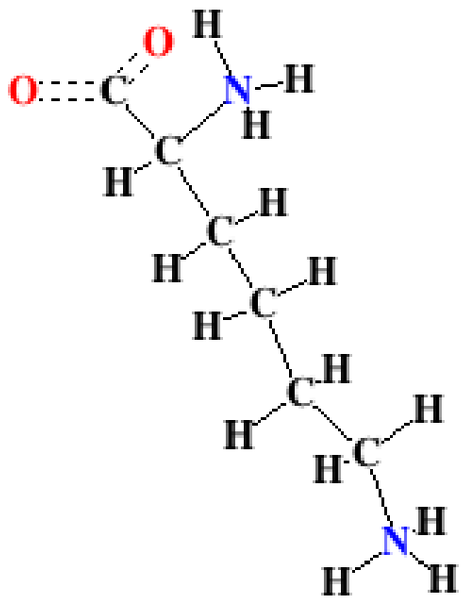


1
2
3

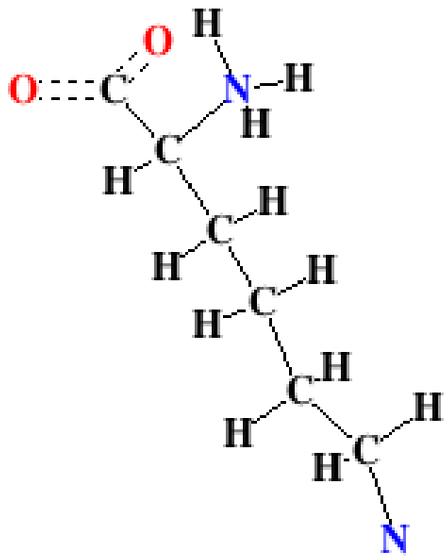


The new link, "file view"

LYS



LYSmod1



Modification "LYSmod1": changes to LYS

```
data_mod_LYSmod1

loop_
_chem_mod_atom.mod_id
_chem_mod_atom.function
_chem_mod_atom.atom_id
_chem_mod_atom.new_atom_id
_chem_mod_atom.new_type_symbol
_chem_mod_atom.new_type_energy
_chem_mod_atom.new_partial_charge
 LYSmod1  change NZ    .    .  N       0.000
 LYSmod1  delete HZ1   .    .  .         .
 LYSmod1  delete HZ2   .    .  .         .
 LYSmod1  delete HZ3   .    .  .         .
```

Atoms

```
loop_
_chem_mod_bond.mod_id
_chem_mod_bond.function
_chem_mod_bond.atom_id_1
_chem_mod_bond.atom_id_2
_chem_mod_bond.new_type
_chem_mod_bond.new_value_dist
_chem_mod_bond.new_value_dist_esd
 LYSmod1  change CE   NZ     .       1.455   0.020
 LYSmod1  delete NZ   HZ3    .         .       .
 LYSmod1  delete NZ   HZ2    .         .       .
 LYSmod1  delete NZ   HZ1    .         .       .
```

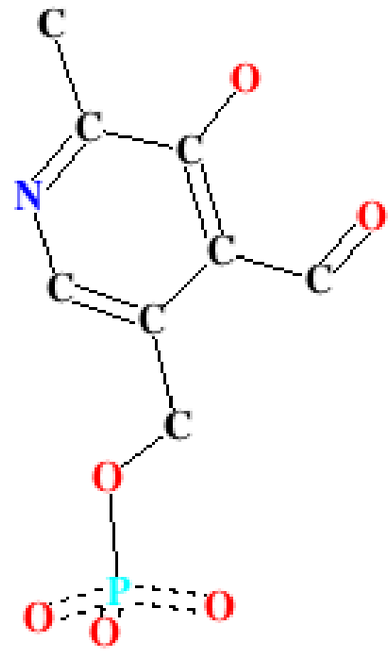
Bonds

Angles

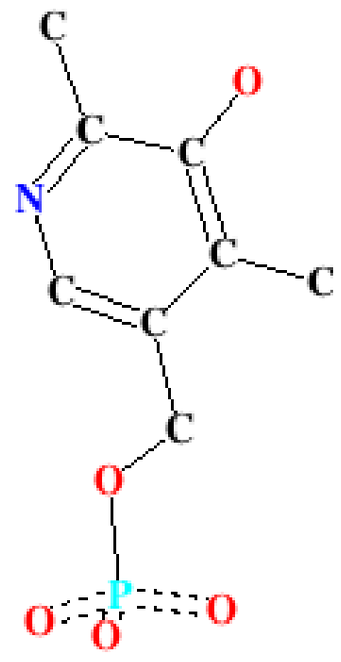
.....

The new link, "file view"

PLP



PLPmod1



Modification "PLPmod1": changes to PLP

```
data_mod_PLPmod1

loop_
_chem_mod_atom.mod_id
_chem_mod_atom.function
_chem_mod_atom.atom_id
_chem_mod_atom.new_atom_id
_chem_mod_atom.new_type_symbol
_chem_mod_atom.new_type_energy
_chem_mod_atom.new_partial_charge
 PLPmod1  delete O4A    .     .    .          .
```

Atom

```
loop_
_chem_mod_bond.mod_id
_chem_mod_bond.function
_chem_mod_bond.atom_id_1
_chem_mod_bond.atom_id_2
_chem_mod_bond.new_type
_chem_mod_bond.new_value_dist
_chem_mod_bond.new_value_dist_esd
 PLPmod1  delete C4A    O4A      .          .      .
```

Bond

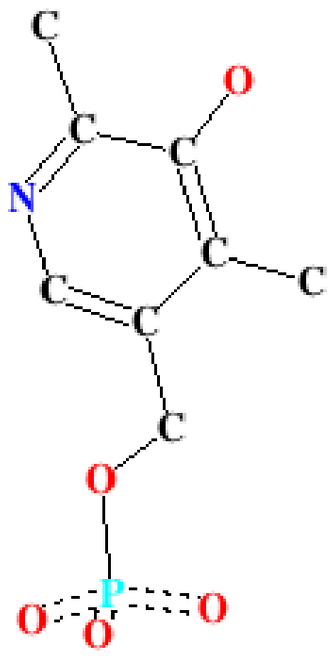
```
loop_
_chem_mod_angle.mod_id
_chem_mod_angle.function
_chem_mod_angle.atom_id_1
_chem_mod_angle.atom_id_2
_chem_mod_angle.atom_id_3
_chem_mod_angle.new_value_angle
_chem_mod_angle.new_value_angle_esd
 PLPmod1  delete H4A    C4A    O4A        .      .
 PLPmod1  delete C4     C4A    O4A        .      .
```

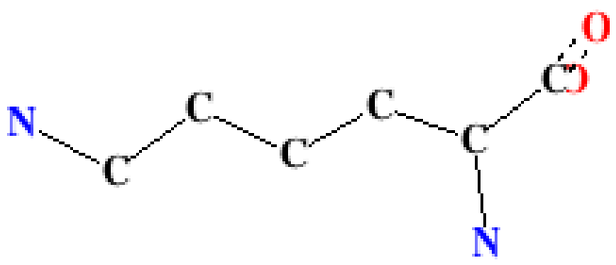
Angles

```
loop_
_chem_mod_plane_atom.mod_id
_chem_mod_plane_atom.function
_chem_mod_plane_atom.plane_id
_chem_mod_plane_atom.atom_id
_chem_mod_plane_atom.new_dist_esd
 PLPmod1  delete plan-2    C4         .
 PLPmod1  delete plan-2    C4A        .
 PLPmod1  delete plan-2    H4A        .
 PLPmod1  delete plan-2    O4A        .
```

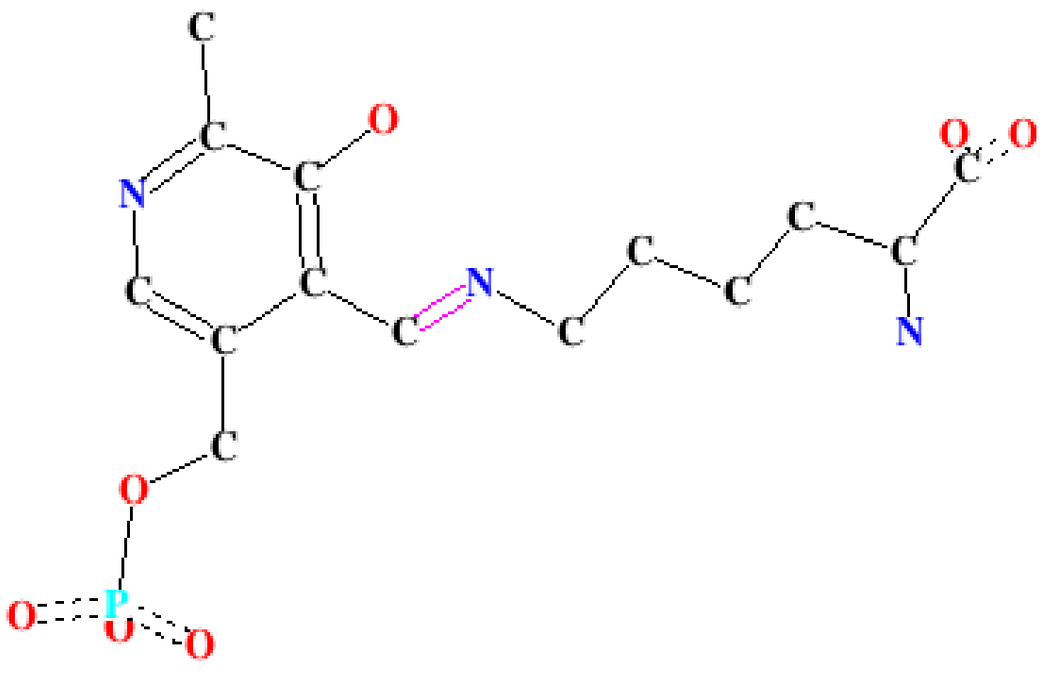
Plane

The new link, "file view"





PLP-LYS



Link "PLP-LYS": changes associated with covalent linkage between modified PLP and LYS

```
data_link_PLP-LYS

loop_
_chem_link_bond.link_id
_chem_link_bond.atom_1_comp_id
_chem_link_bond.atom_id_1
_chem_link_bond.atom_2_comp_id
_chem_link_bond.atom_id_2
_chem_link_bond.type
_chem_link_bond.value_dist
_chem_link_bond.value_dist_esd
 PLP-LYS  1 C4A    2 NZ       double      1.260    0.020
```

Bond

```
loop_
_chem_link_angle.link_id
_chem_link_angle.atom_1_comp_id
_chem_link_angle.atom_id_1
_chem_link_angle.atom_2_comp_id
_chem_link_angle.atom_id_2
_chem_link_angle.atom_3_comp_id
_chem_link_angle.atom_id_3
_chem_link_angle.value_angle
_chem_link_angle.value_angle_esd
 PLP-LYS  1 C4A    2 NZ     2 CE      120.000    3.000
 PLP-LYS  1 H4A    1 C4A    2 NZ      120.000    3.000
 PLP-LYS  1 C4     1 C4A    2 NZ      120.000    3.000
```

Angles

```
loop_
_chem_link_plane.link_id
_chem_link_plane.plane_id
_chem_link_plane.atom_comp_id
_chem_link_plane.atom_id
_chem_link_plane.dist_esd
 PLP-LYS  plan-2    1 C4       0.020
 PLP-LYS  plan-2    1 C4A      0.020
 PLP-LYS  plan-2    1 H4A      0.020
 PLP-LYS  plan-2    2 CE       0.020
 PLP-LYS  plan-2    2 NZ       0.020
```

Plane

Utilising new link description

Three remaining steps:

– docking monomer(s) into electron density

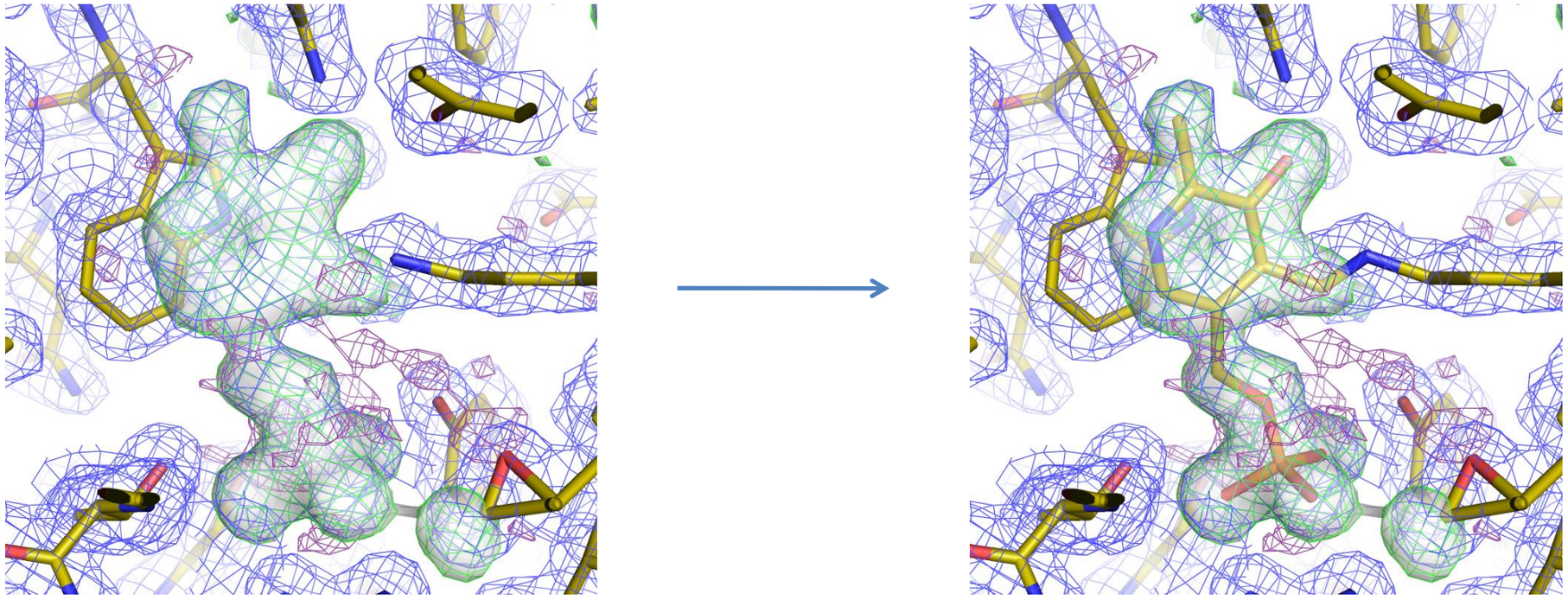
– defining link in the pdb-file

– refinement of the structure with linked ligand using additional library

(1) Docking into the electron density

In our example, this is completely independent step: the additional library is not used.

– non-modified monomer is taken from the standard library
– docking is performed, e.g. using coot:

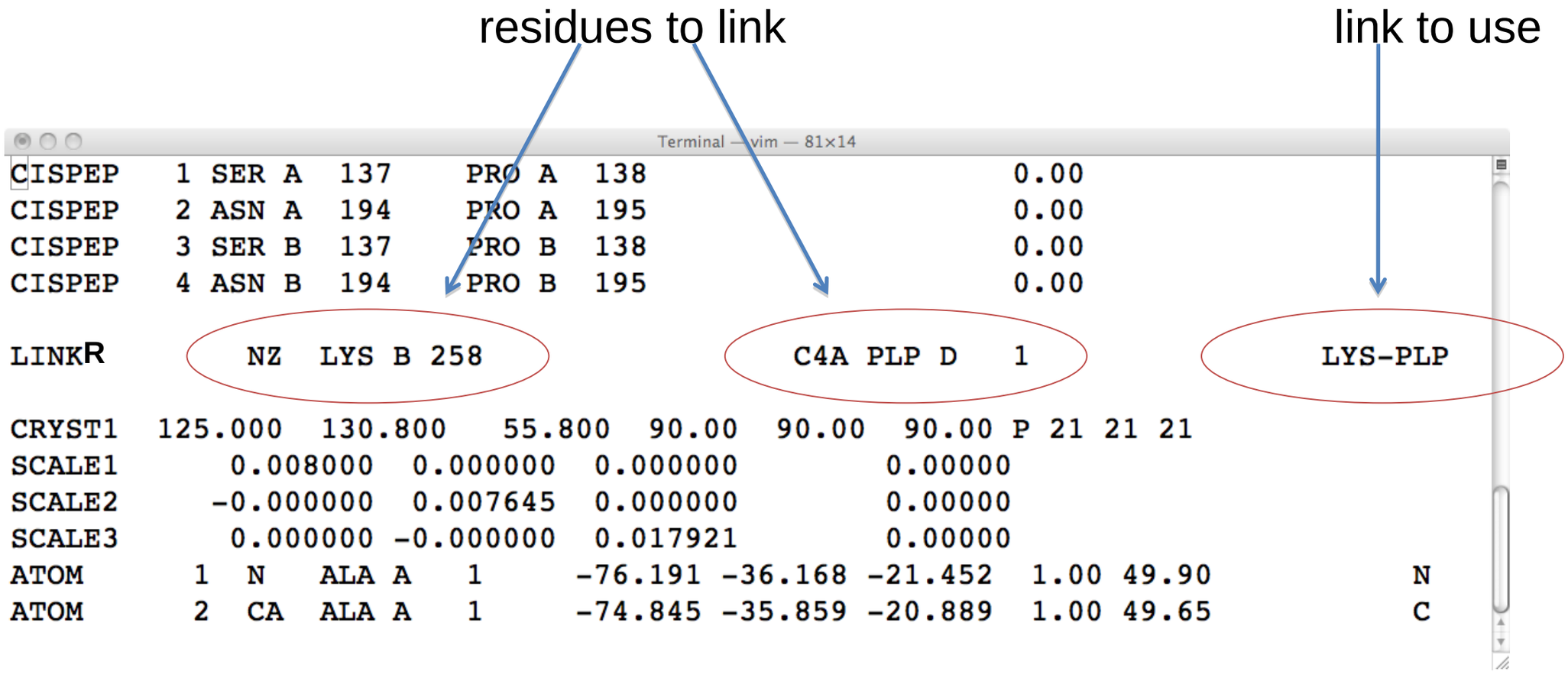


– leaving atoms (O4A of PLP in this example) are removed
– in our example, one of the monomers (LYS) is already in the model

(2) Defining link in the pdb-file

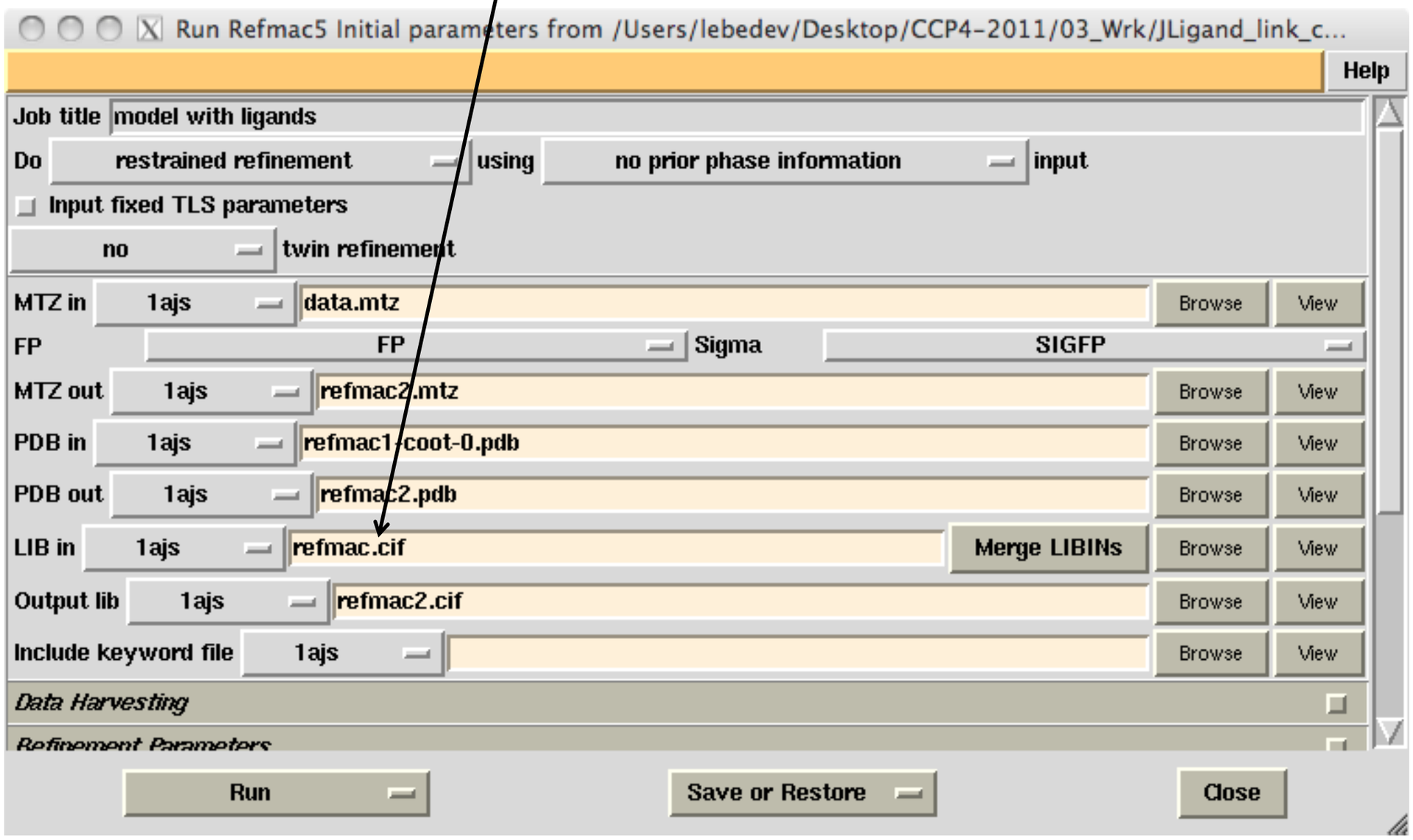
In general case, link cannot be applied automatically.
For example:
– e.g. the same two atoms of the same two compounds can form single or double bond
– H-atom are not defined in the PDB-file

Therefore REFMAC needs additional instructions:



(3) Refinement using additional library

Additional library is defined here



New Ligand description

<u>One solution:</u>

Updatable empirically-derived dictionary

Source: small molecule database

COD – Crystallography Open Da

•Crystal structures of organic, inorganic, metal-organic compounds and minerals

•Freely accessible

•>300,000 structures

Atom types can be applied to other databases

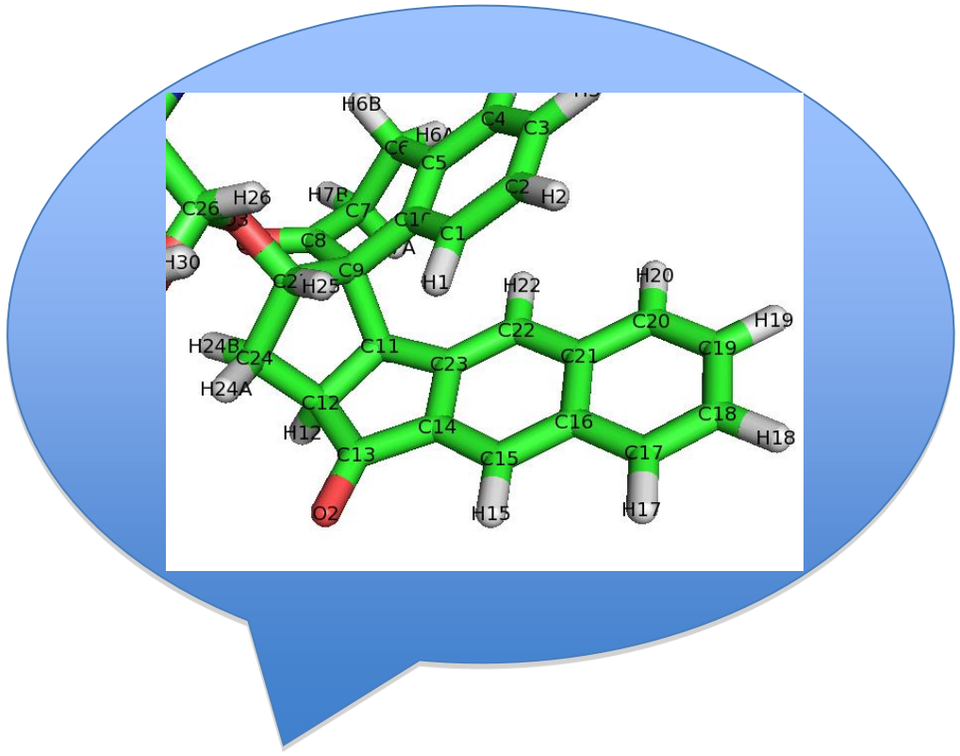
Also utilise data from other sources (e.g. industry)

ACEDRG: A new dictionary generator

Organization of the ACEDRG

The database in ACEDRG:

- Unique atom types as local graphs to describe the specific environment of bonded atoms.



Atom C23: C[5,6a](C[5,5]C[5,5]C[5,6]H)(C[5,6a]C[6a]C[5])(C[6a]C[6a,6a]H)
Atom C18: C[6a](C[6a]C[6a,6a]H)(C[6a]C[6a]H)(H)

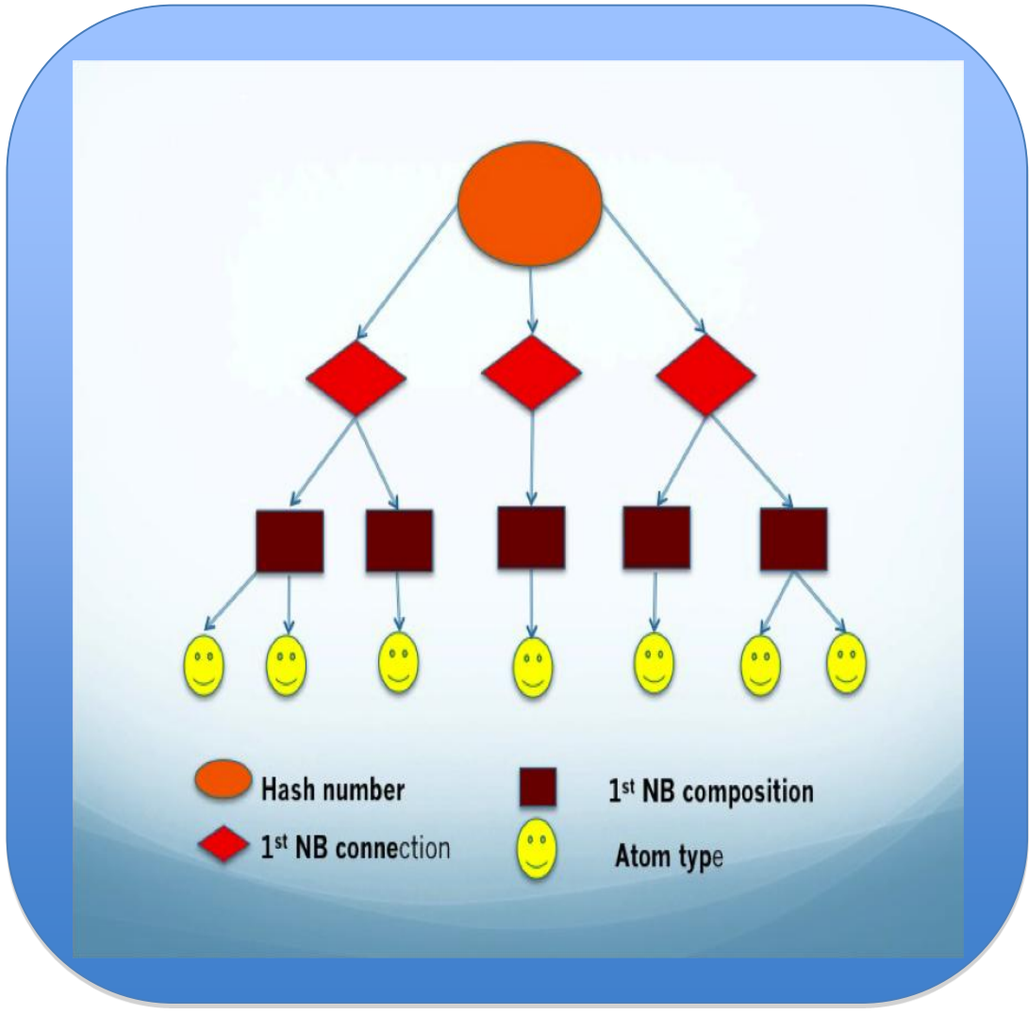
- We have more than 300,000 atom types for “organic set” of elements
- The atom types could be applied to other databases

Organization of the ACEDRG

The database in ACEDRG:

❖Hierarchical tree for different levels of approximations of atom types

❖Isomorphism-mapping algorithms if user's atom types do not exist in the database.



- There are about 10 million bond length values
- There are about 25 million bond angle values

Organization of ACEDRG

Metal-organic compounds:

- Metal-organic compounds are clustering according to their coordination numbers and geometries
- New dictionary includes 26 coordination geometries and the angles within these geometries are stores as tables
- For an organic atom that is connected to metal atoms, its non-metal neighbor atoms are treated as described before

Functionality of ACEDRG

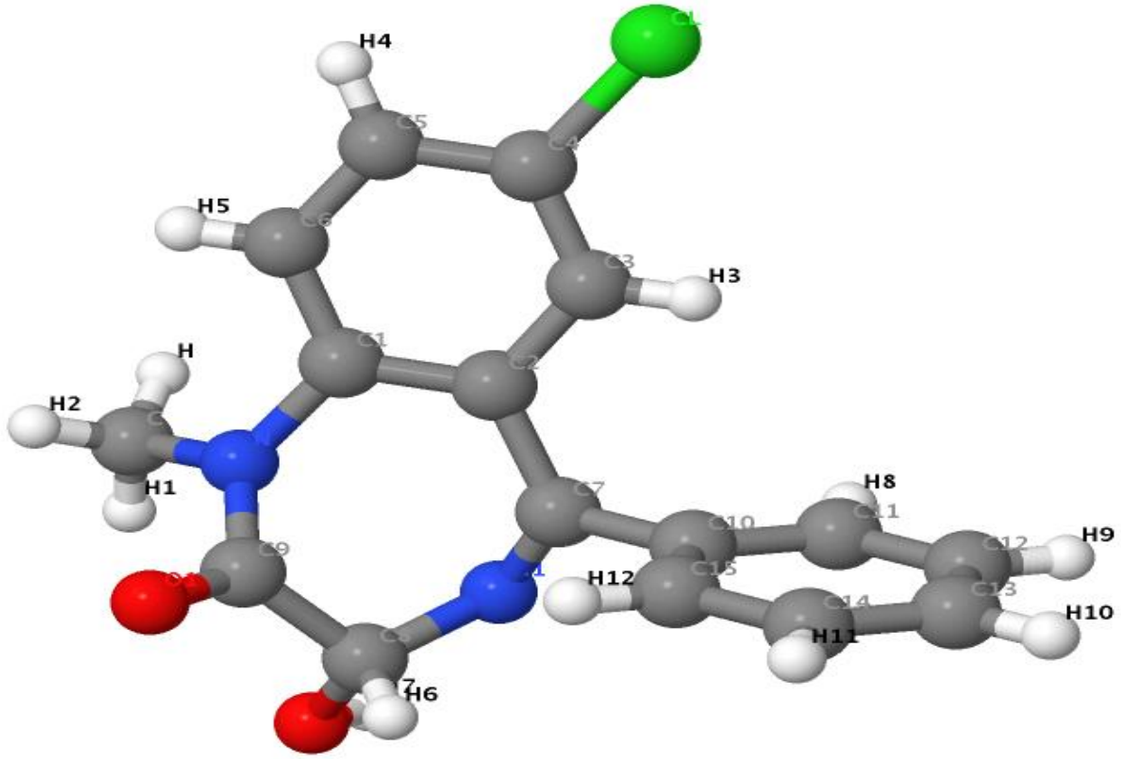
Ligand/Dictionary generator

- Take input files of SMILES, SYBIL/MOL2, SDF/MDL, mmCIF formats
- Output Dictionary files of mmCIF format, which contain atom types, bond lengths and angles, torsion angles, planes and chirality centers, and are used as restraint files for refinements.
- Output coordinate files of PDB format, which represent one of lower energy conformations for the ligands under consideration.

Functionality of ACEDRG

Ligand/Dictionary generator

SMILES STRING:
CN1C2=C(C=C(Cl)C=C2)C(=NC(O)C1=O)C1=CC=CC=C1



Acknowledgment

Fei Long

Rob Nicholls

Paul Emsley

Andrea Thorn

David Tourigny

Andrey Lebedev

Alexei Vagin

Paul Young

CCP4, MRC-LMB people

Jligand and ACEDRG are available from CCP4