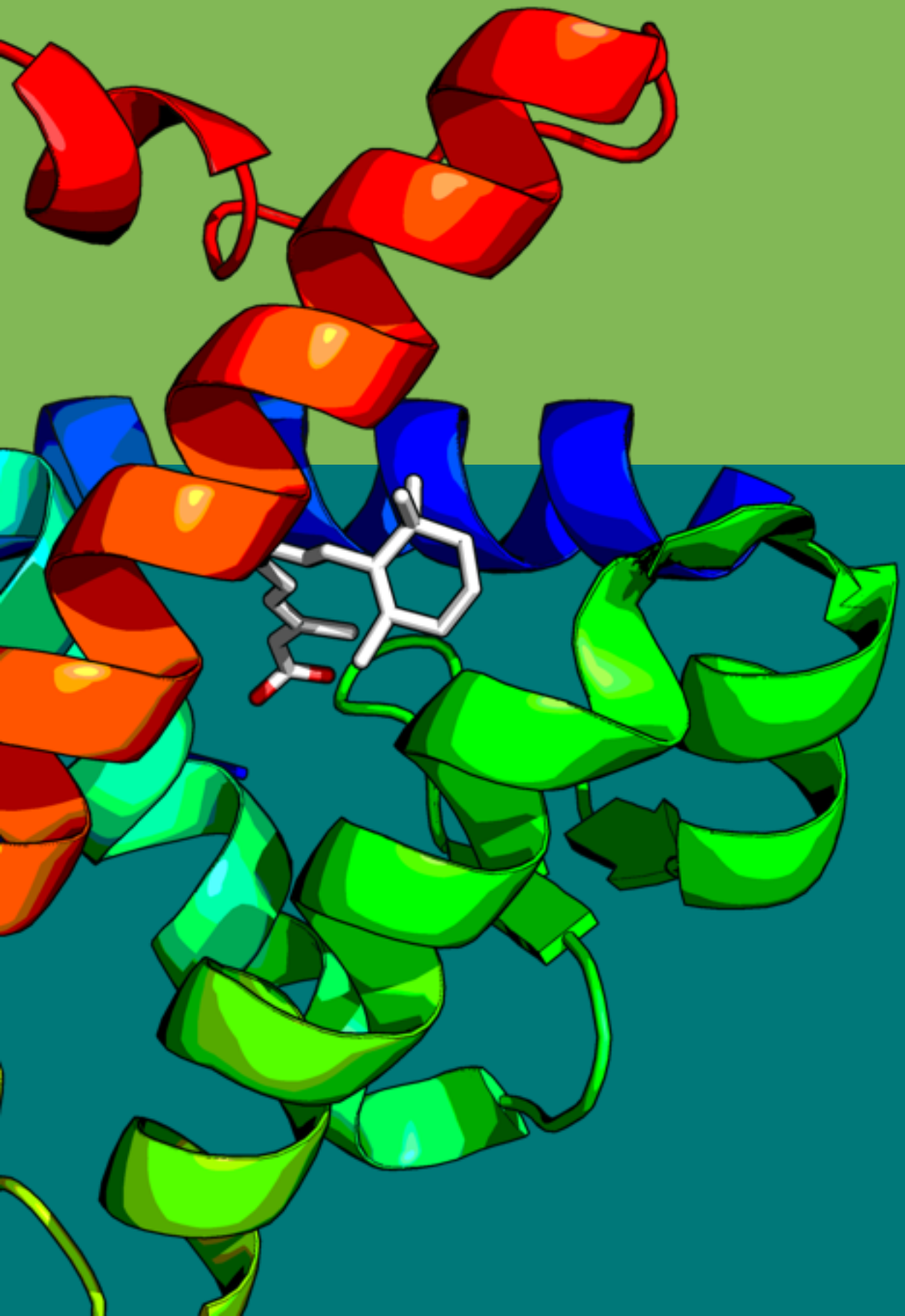
Joana Pereira
Lamzin Group
EMBL Hamburg, Germany



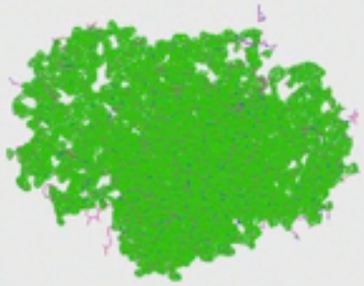
# Small molecules
# How to identify and build them
## (with ARP/wARP)

EMBL 40 YEARS | 1974–2014

7.5 ARP wARP

# The task at hand

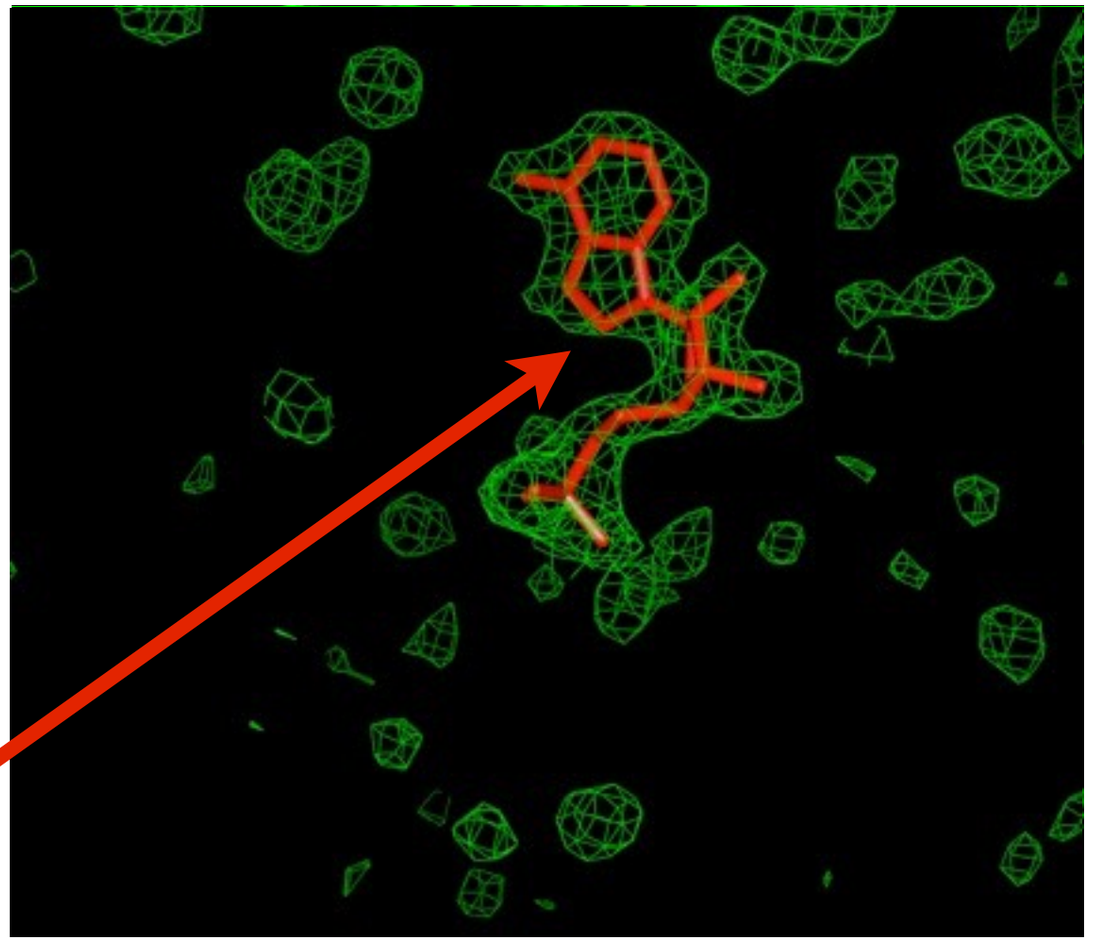
*To find ligand density and build it!*

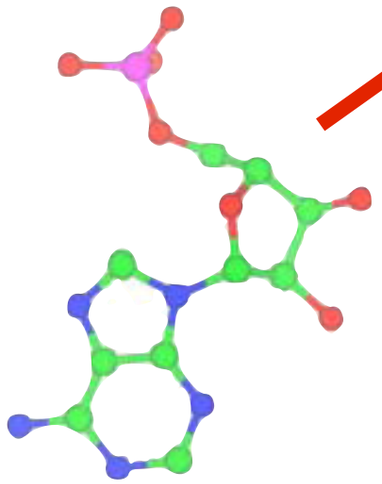


# Fitting a ligand

We have:

- electron density map (good resolution)
- + protein model
  - shows **excess density** not accounted for by protein
- **difference density map**
  - to fit the appropriate ligand into

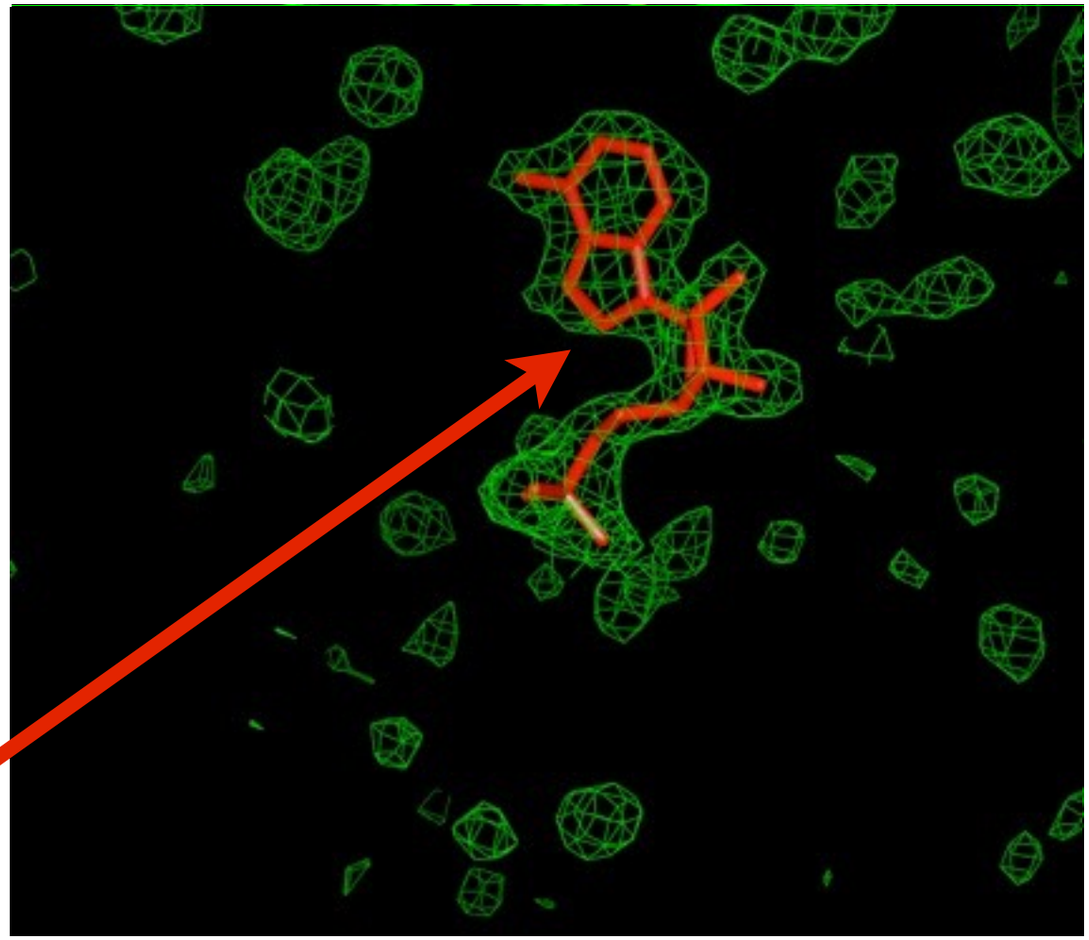


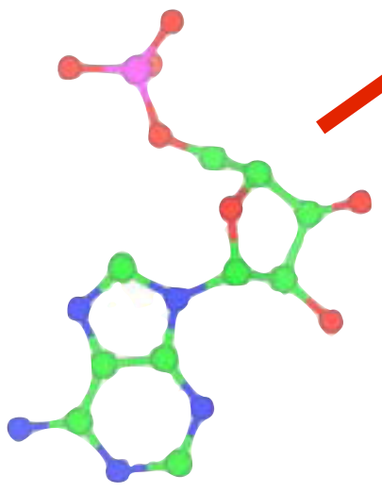


# Fitting a ligand

Challenges:

- different **resolutions** and data quality
- different ligand **complexity / topology**
- **partial disorder** of a ligand
- different ligands at the same site?





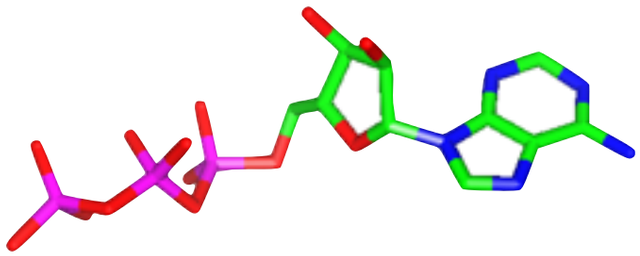
# The protocol of a modelling task

What we know at the beginning will determine the protocol:

| Ligand identity \ Binding site | | Known | Unknown |
|---|---|---|---|
| | Known | X | |
| | Unknown | | |

1 density cluster

1 ligand

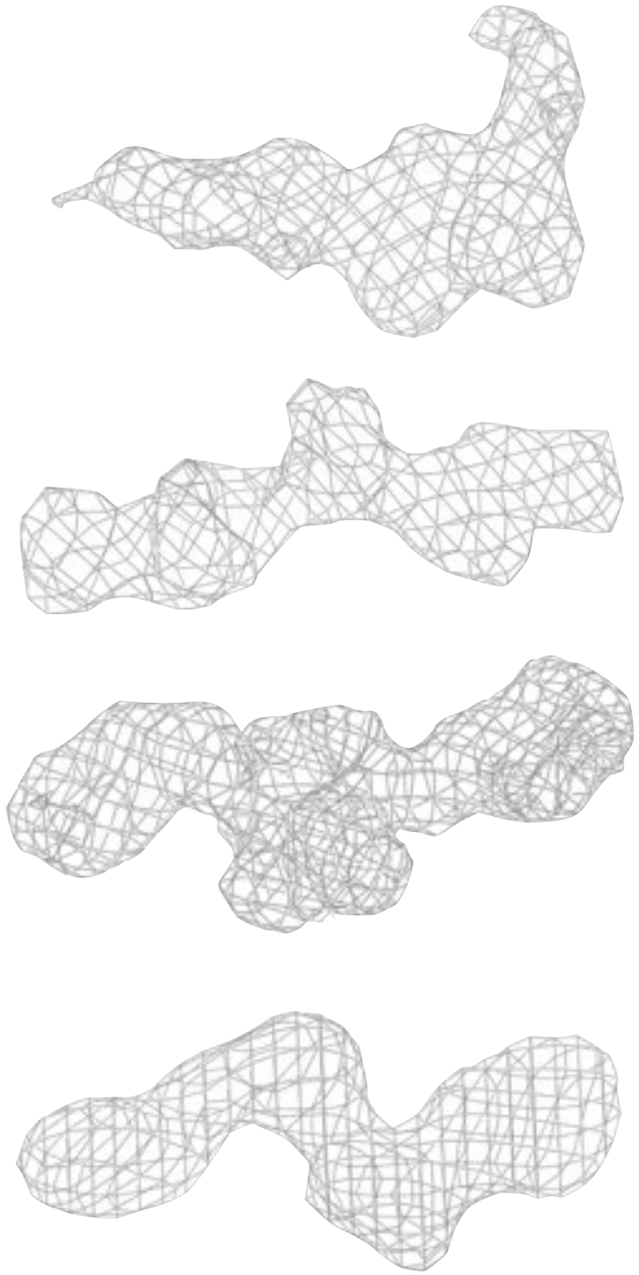


# The protocol of a modelling task

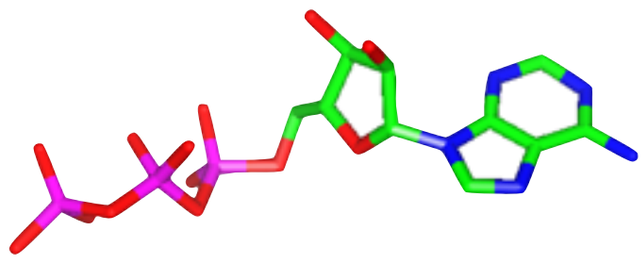
What we know at the beginning will determine the protocol:

| Ligand identity \ Binding site | Known | Unknown |
|---|---|---|
| Known | | X |
| Unknown | | |

N density clusters



1 ligand

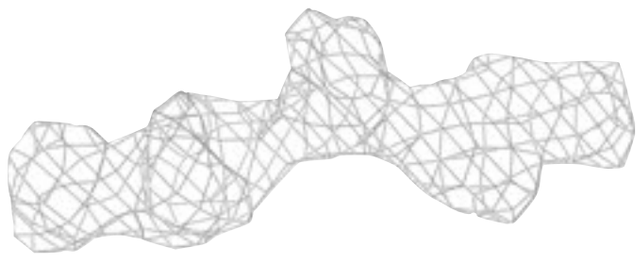


# The protocol of a modelling task

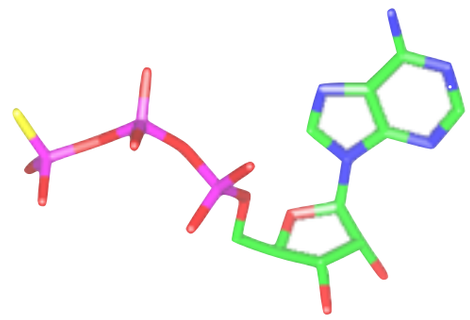
What we know at the beginning will determine the protocol:

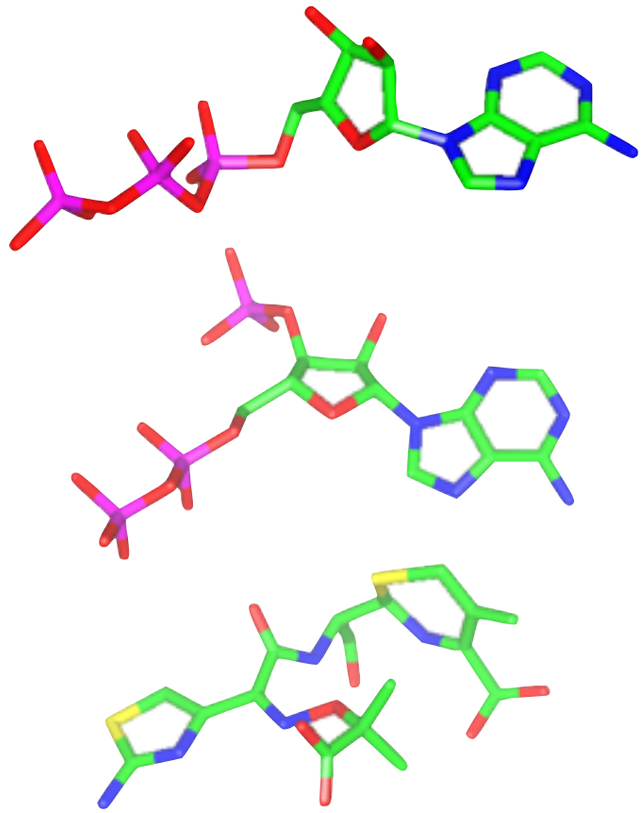
| Ligand identity \ Binding site | Known | Unknown |
|---|---|---|
| Known | | |
| Unknown | X | |

1 density cluster



N ligand candidates





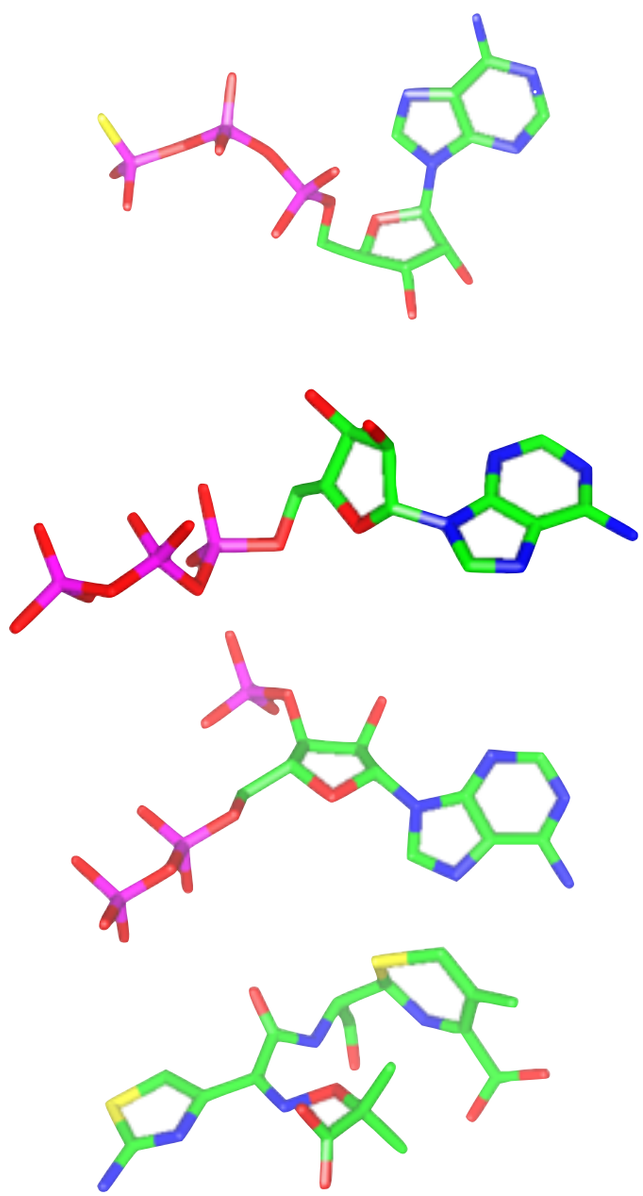
# The protocol of a modelling task

What we know at the beginning will determine the protocol:

| Ligand identity \ Binding site | Known | Unknown |
|---|---|---|
| Known | | |
| Unknown | | X |

N density clusters

N ligand candidates



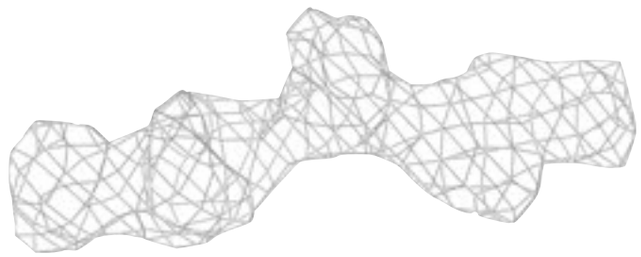
# Researcher's needs #1

I am trying to solve a structure of a protein with some inhibitor. **I want to know how I can put in my inhibitor in the density map of the data I got.** I can see some density in the active site where the inhibitor should be…

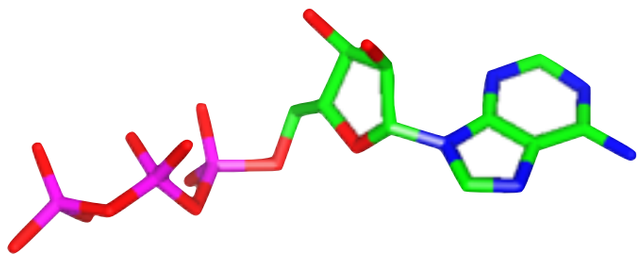
I am not sure of how to do it.

**Which case is this?**

1 density cluster

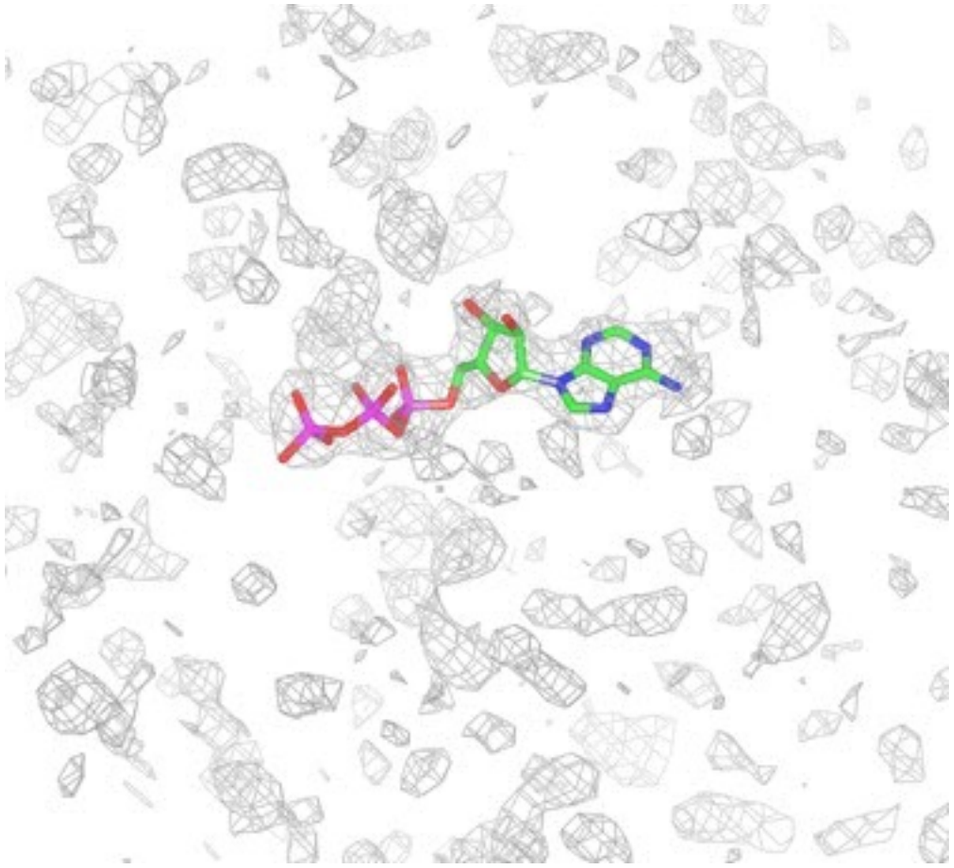


1 ligand



# Ligand fitting with ARP/wARP

Organised as a **pipeline** of core modules for **specific sub-tasks** following an intuitive approach



Prepare

- Identify ligand **binding site** and /or **ligand**
- Sparse density map and **generate ligand topology**

Construct

- Construct **ensemble** of ligand models in plausible conformation to **fit sparsed density**

Refine

- Refine ligand coordinates to satisfy **geometric** constraints and **maximise** fit to density
- Choose **best model**

# Constructing the ligand

- To **address the different ligand sizes** and complexities encountered
  - to increase the **robustness** of the software

- We use 2 separate construction methods:

Label swapping on the sparsed grid

Metropolis search in conformation space

# Constructing the ligand: graph search

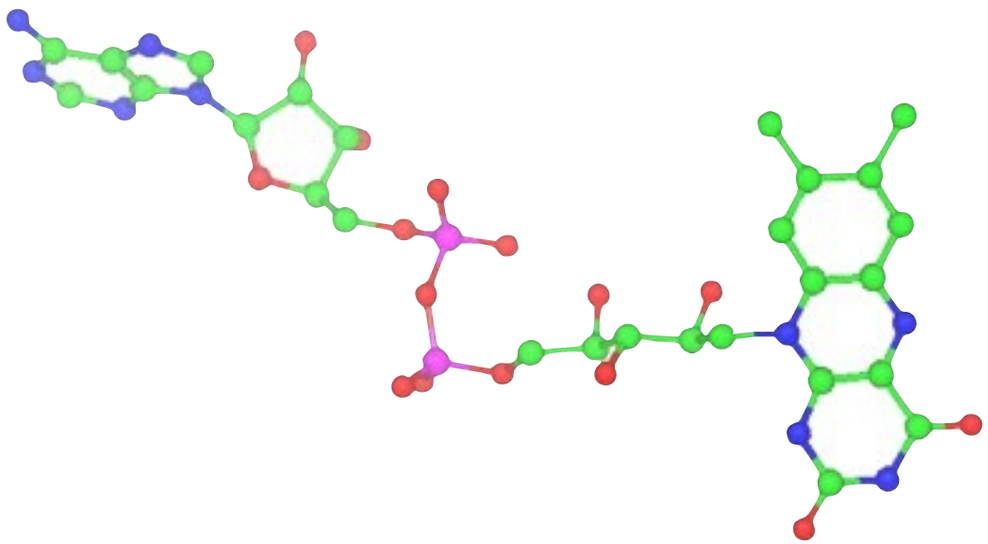
- Uses the **sparse grid representation** of the electron density at the chosen **binding site** & **topology of known ligand**

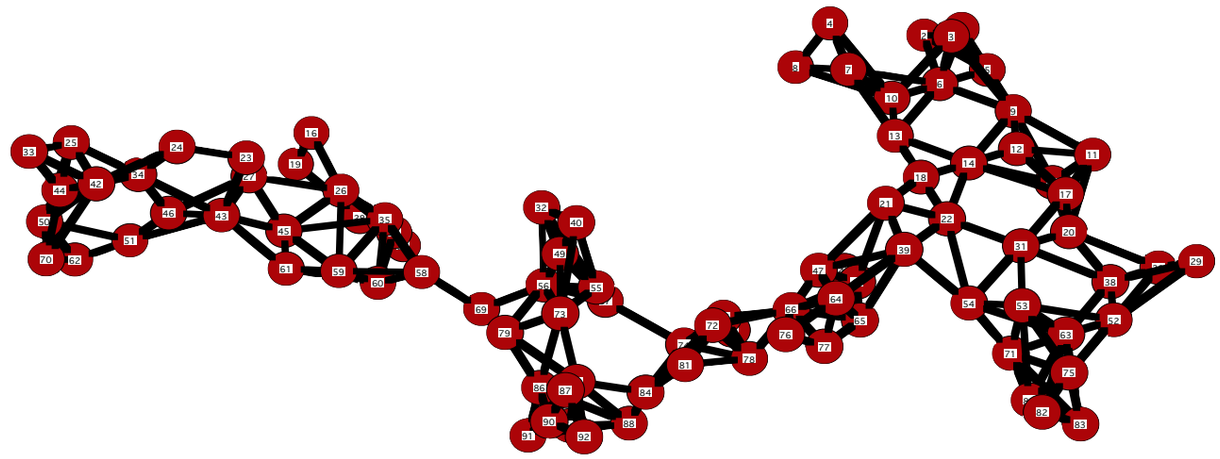
Knowledge about ligand

- Connectivity
- Distances
- Angles
- Chirality

Knowledge about grid

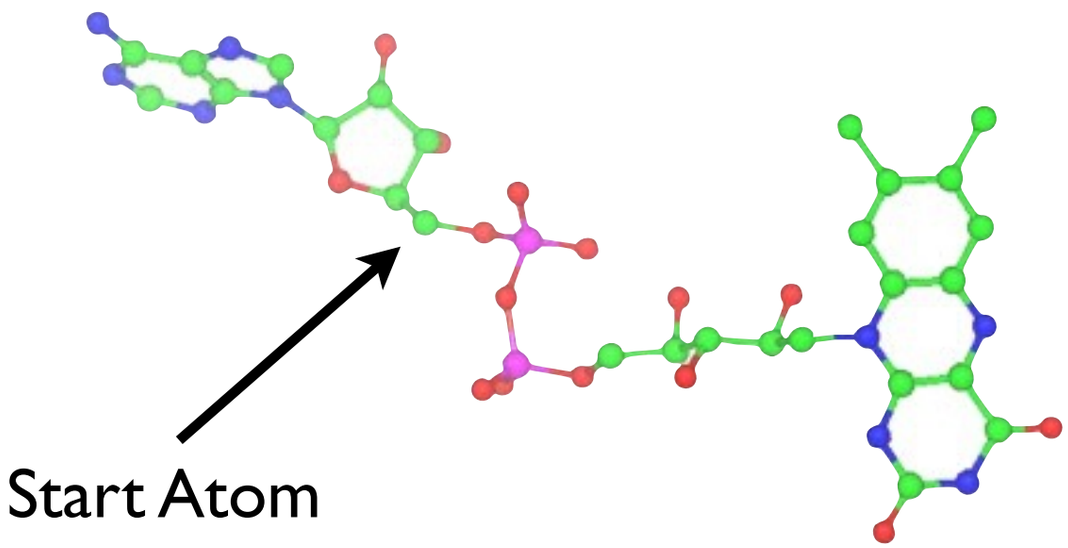
- Connectivity
- Distances
- Density
- *NO IDENTITIES*

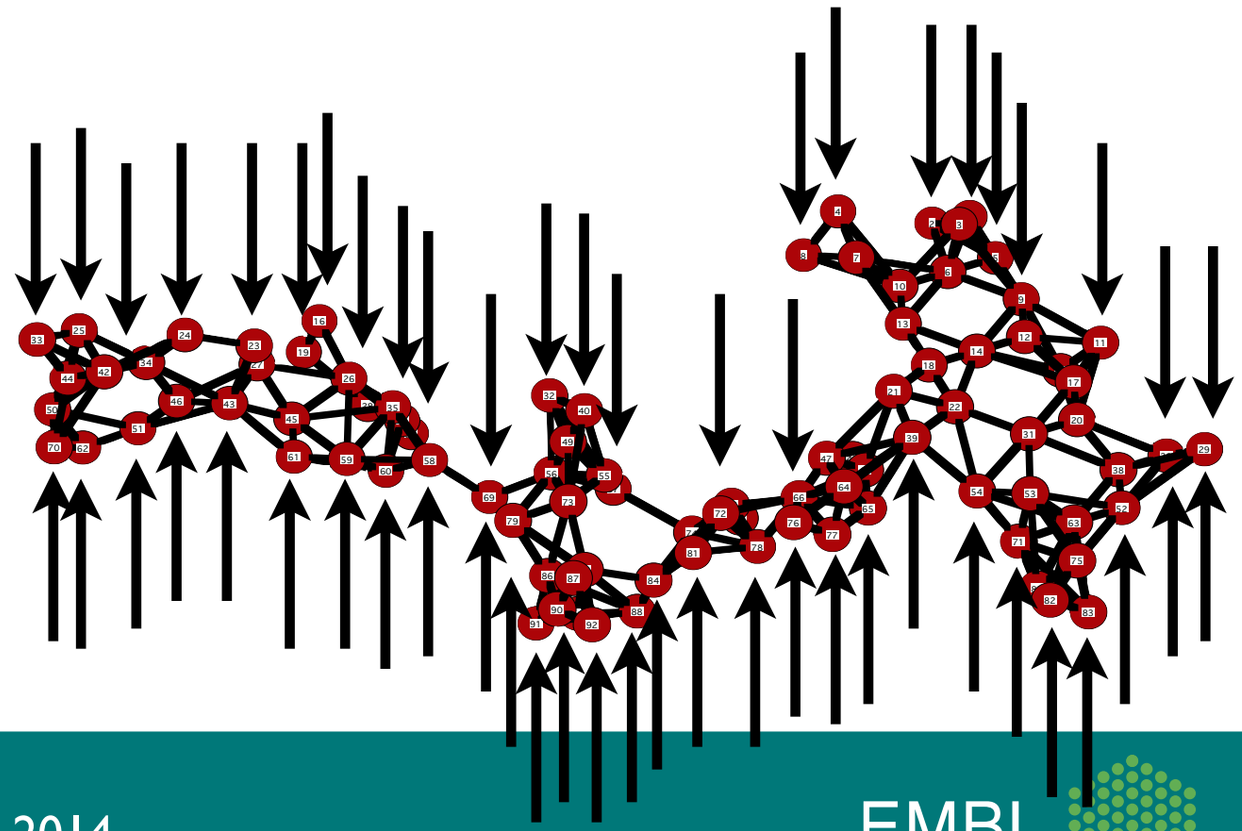




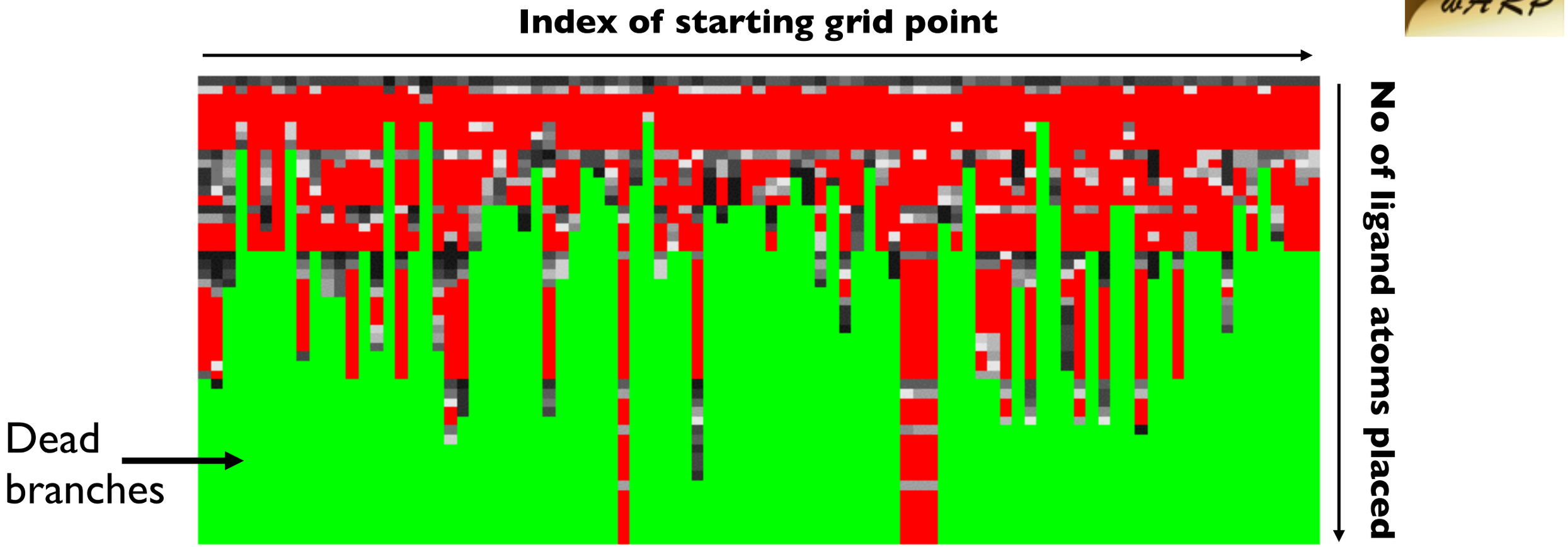
# Constructing the ligand: graph search

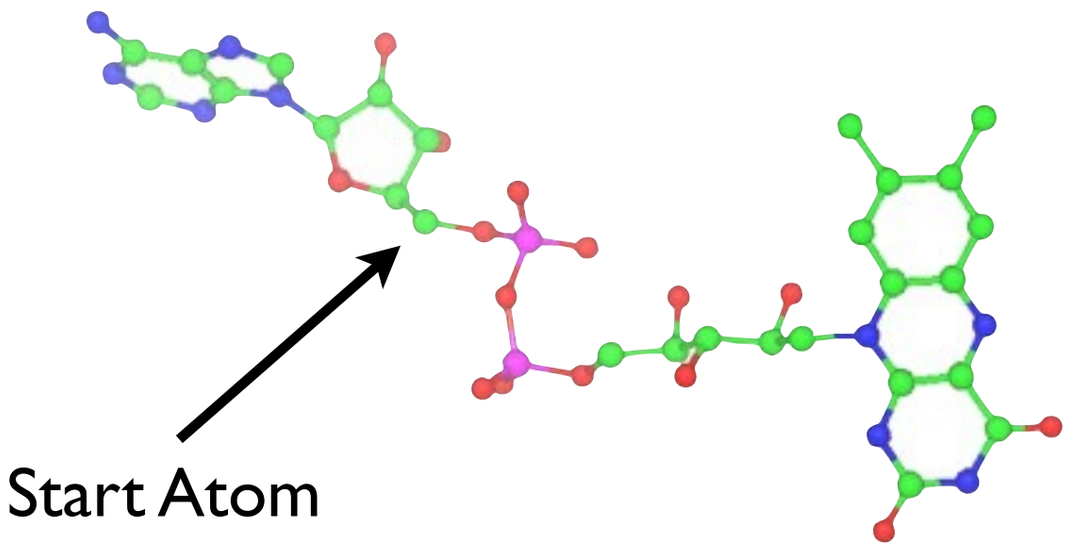
- Label swapping
  - Ligand is **expanded** on the sparse density, **preserving connectivities**
  - Every 'dummy atom' of the sparse grid is tried as a **start point**
  - An **exhaustive graph search** is performed
  - Models are scored by their **fit to density** and expected **stereochemical features**

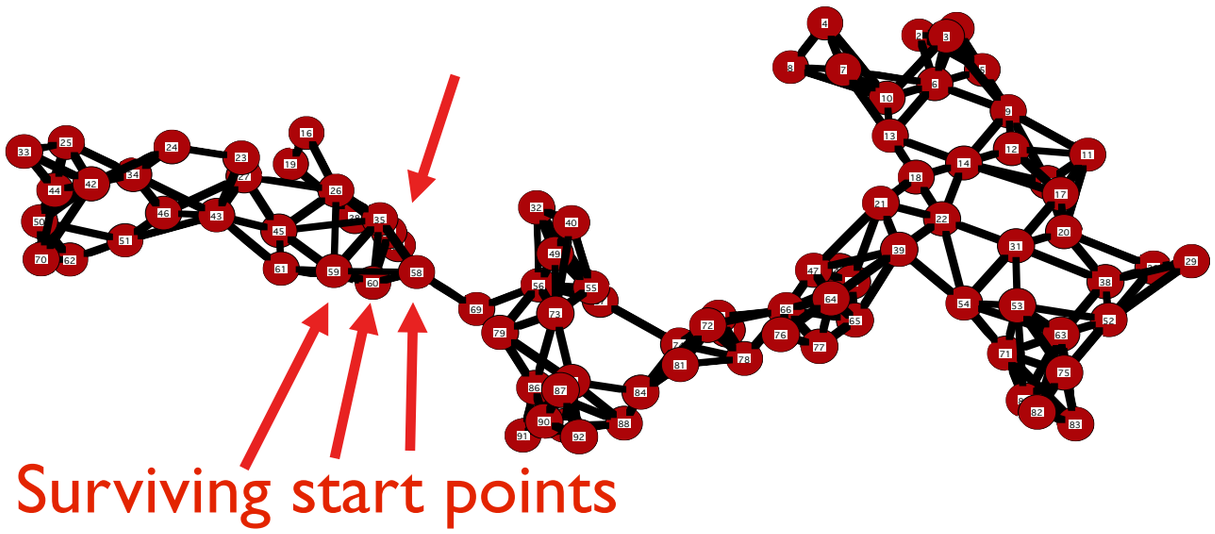


# Constructing the ligand: graph search



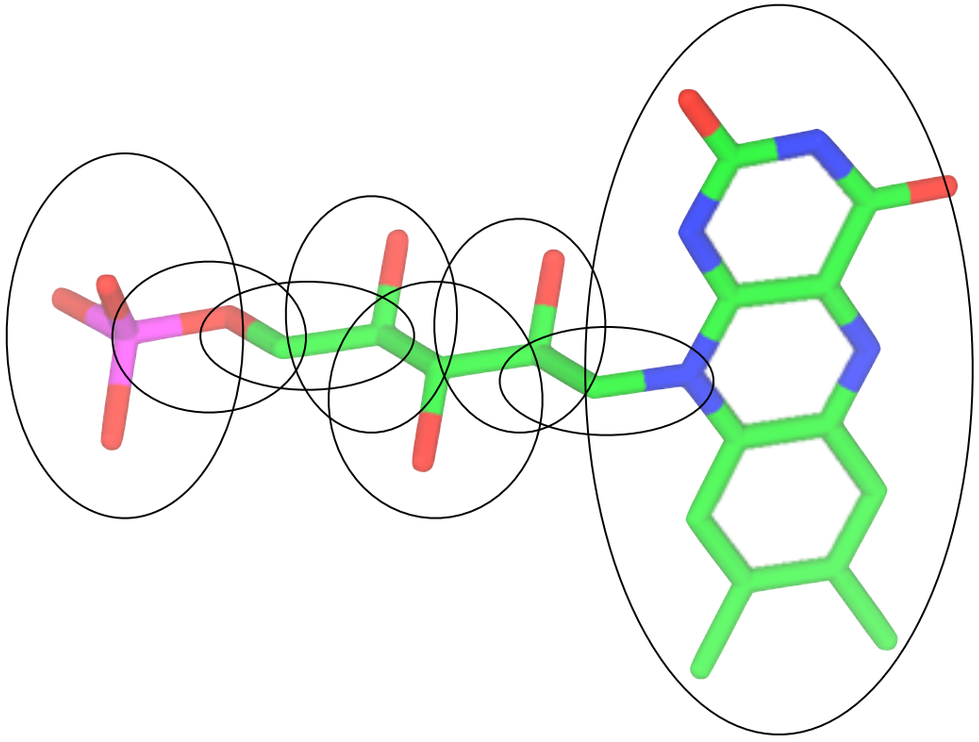



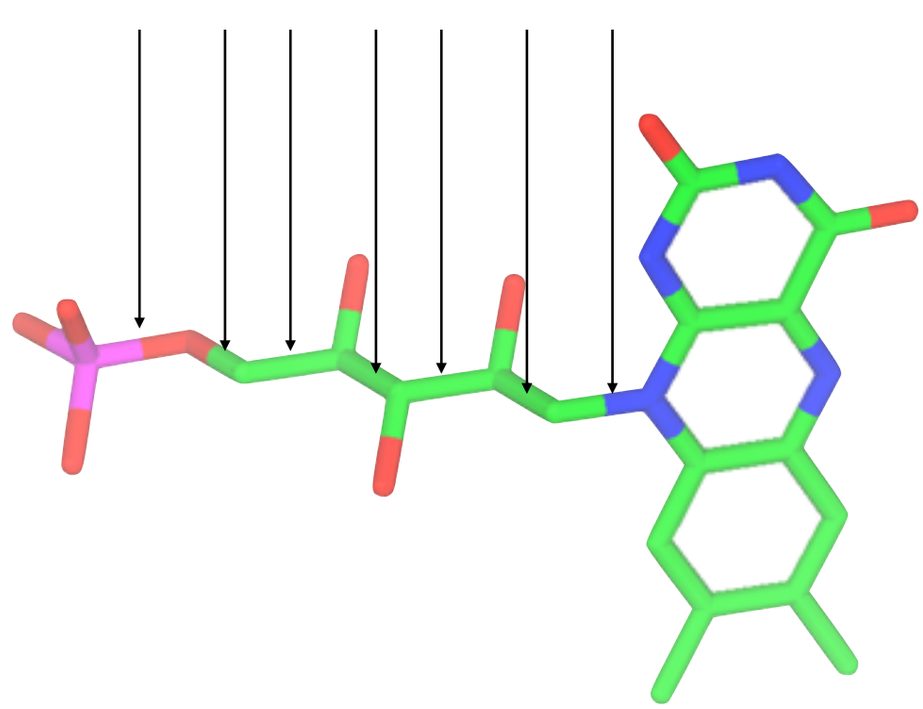
# Constructing the ligand: metropolis search

- Perform a **random walk** in parameter space biased towards the **optimum of a score function**
  - Parameter space: **position, orientation** and **conformation**
  - Score function: **'pseudo' map correlation**
  - Advantage: **less degrees of freedom**

Rigid groups

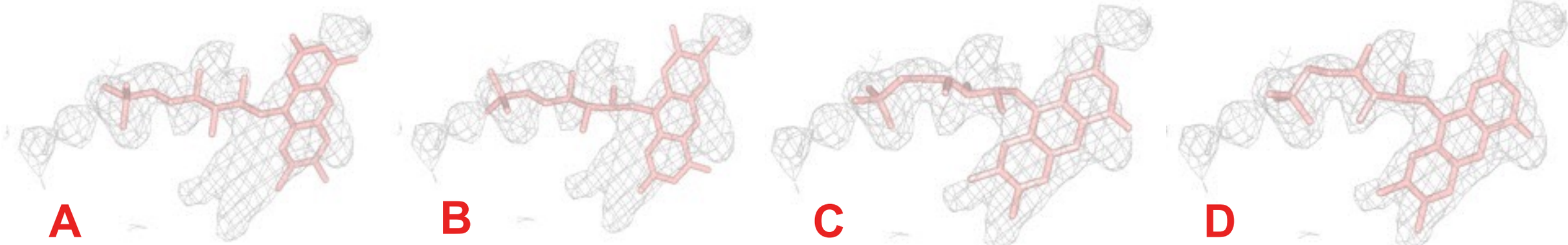


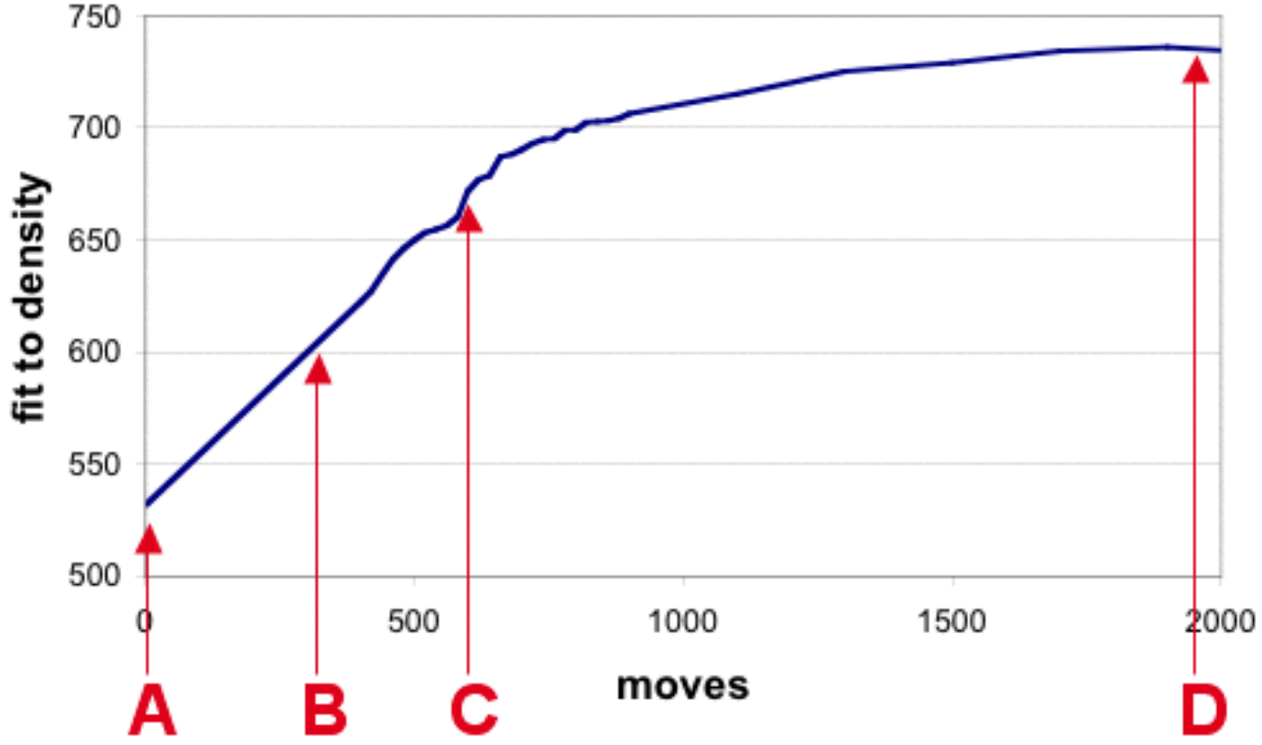
Rotatable bonds

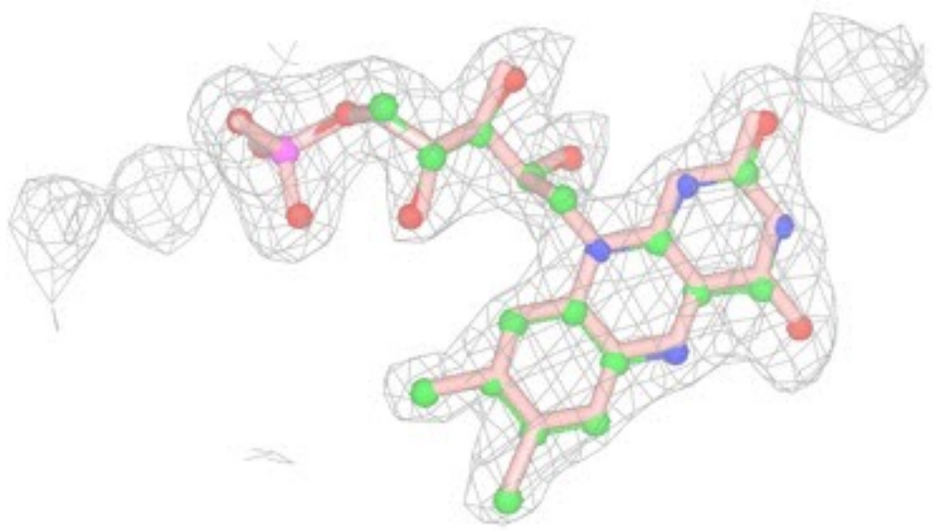


# Constructing the ligand: metropolis search

- Evolution of a model during a metropolis optimisation



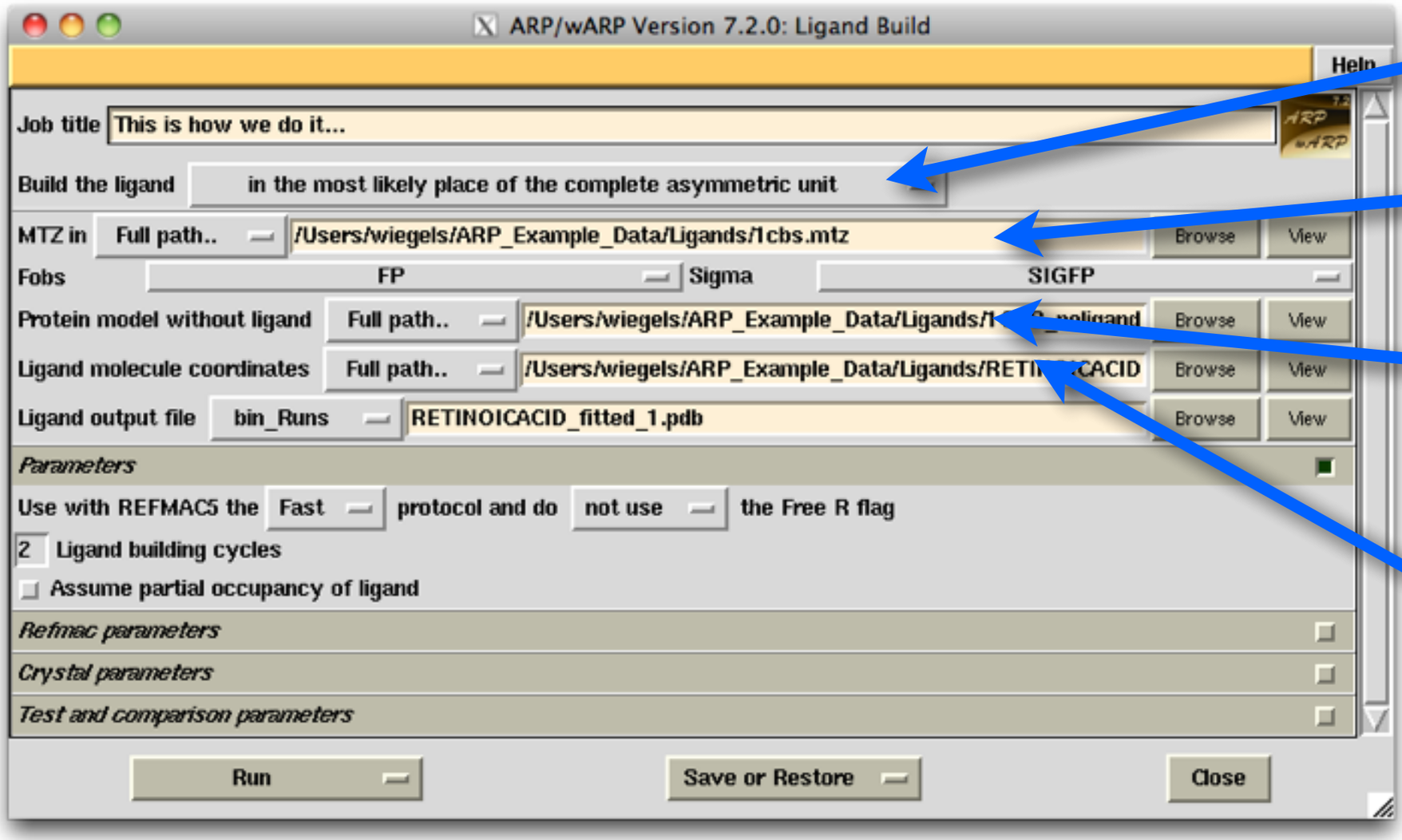



FMN bound to 1jqv at 2.1Å: 0.23Å rmsd.

Final result: 9000 moves + refinement

# Ligand building in the CCP4 interface

- Ligand building GUI as a part of ARP/wARP in CCP4i




- For batch jobs - use the command line...

# Ligand building from the command line

- For batch jobs - use the command line...

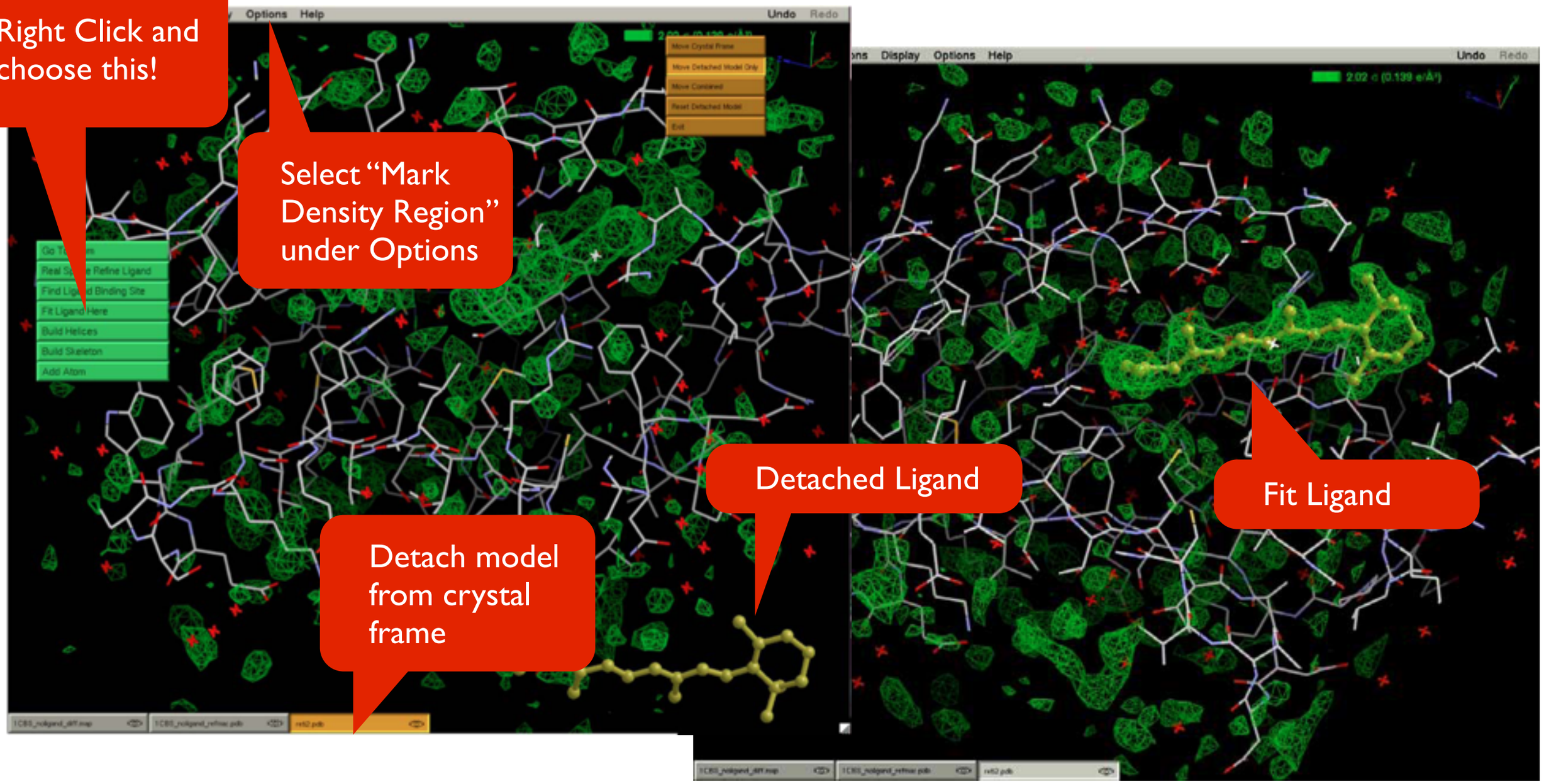
```
Terminal — tcsh — 102×32
[MENORCA:~] wiegels% auto_ligand.sh

Wed Nov  9 16:08:53 CET 2011



Usage:
auto_ligand.sh                                                          \
               datafile {either mtzfile or mapfile}                     \
               protein {starting_PDB_file_without_ligand}               \
               ligand {PDB_file_with_ligand_to_fit}                     \
               [workdir {FULLPATH_WORKING_DIRECTORY}]                   \
               [ligandfileout {output_PDB_file}]                        \
               [fp {fp_label}] [sigfp {sigfp_label}] [freer {freer_label}]  \
               [nligandcycles {number_of_ligandbuild_cycles (default is 2)}]  \
               [search_model {PDB_file_with_model_at_expected_ligand_site}]   \
               [search_position {X Y Z}]                                \
               [search_radius {radius_in_angstroms}]                    \
               [reflist {textfile_with_FULLPATHnames_of_fitted_ligands_for_comparison}] \
               [extralibrary {user_defined_library_for_Refmac5}]        \
               [partial {0 for modelling the whole ligand and 4 or higher number to     \
                  model partially occupied ligand (giving 4 would mean to consider      \
                  4-atoms as the smallest ligand fragment)]             \
               [parfile {parfilename_if_only_parfile_is_to_be_created}]

- Optional command line arguments are given in square parentheses
- All input files are assumed to be located in working directory
  unless they are given with full path
- If workdir is not given, the current directory will be assumed
- All output files will be written into workdir/subdirectory

[MENORCA:~] wiegels%
```

to launch job with default scenario:

- only 3 files needed

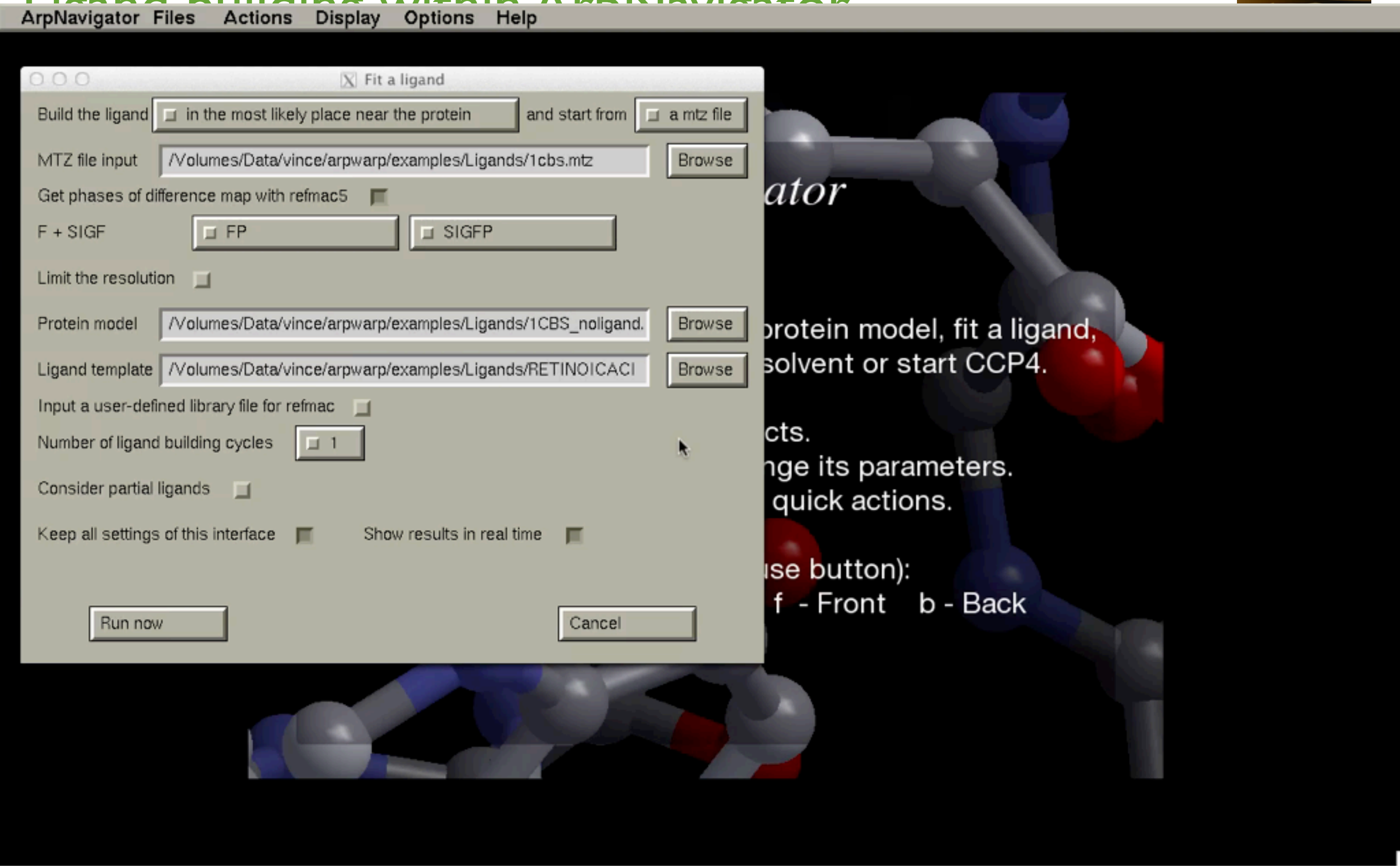
# Ligand building within ArpNavigator

Load the density map and the protein. Load the ligand you want to build.
Mark a density blob, detach and fit the ligand right there. Easy!



# Ligand building within ArpNavigator





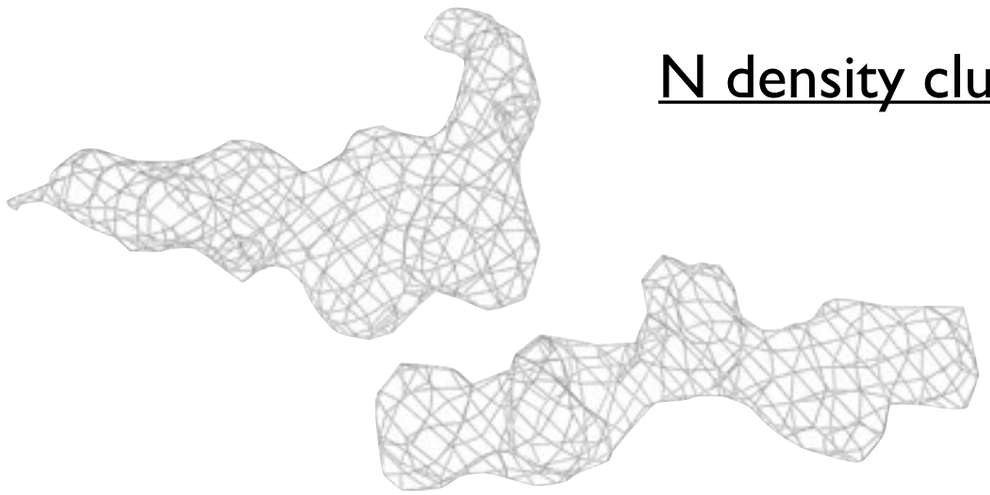
# Researcher's needs #2

Is there a simple way to determine whether ligand is bound or not by comparing the diffraction patterns between ligand-free (structure known) and ligand-soaked protein crystals? I would like to solve the ligand bound protein structure, but before I do so, **I have to find out if the ligand is actually bound and if so, where**.

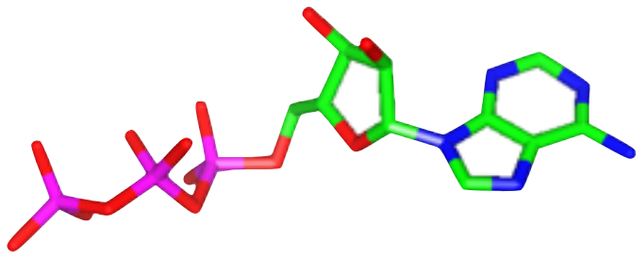
Thank you very much!

## Which case is this?

N density clusters

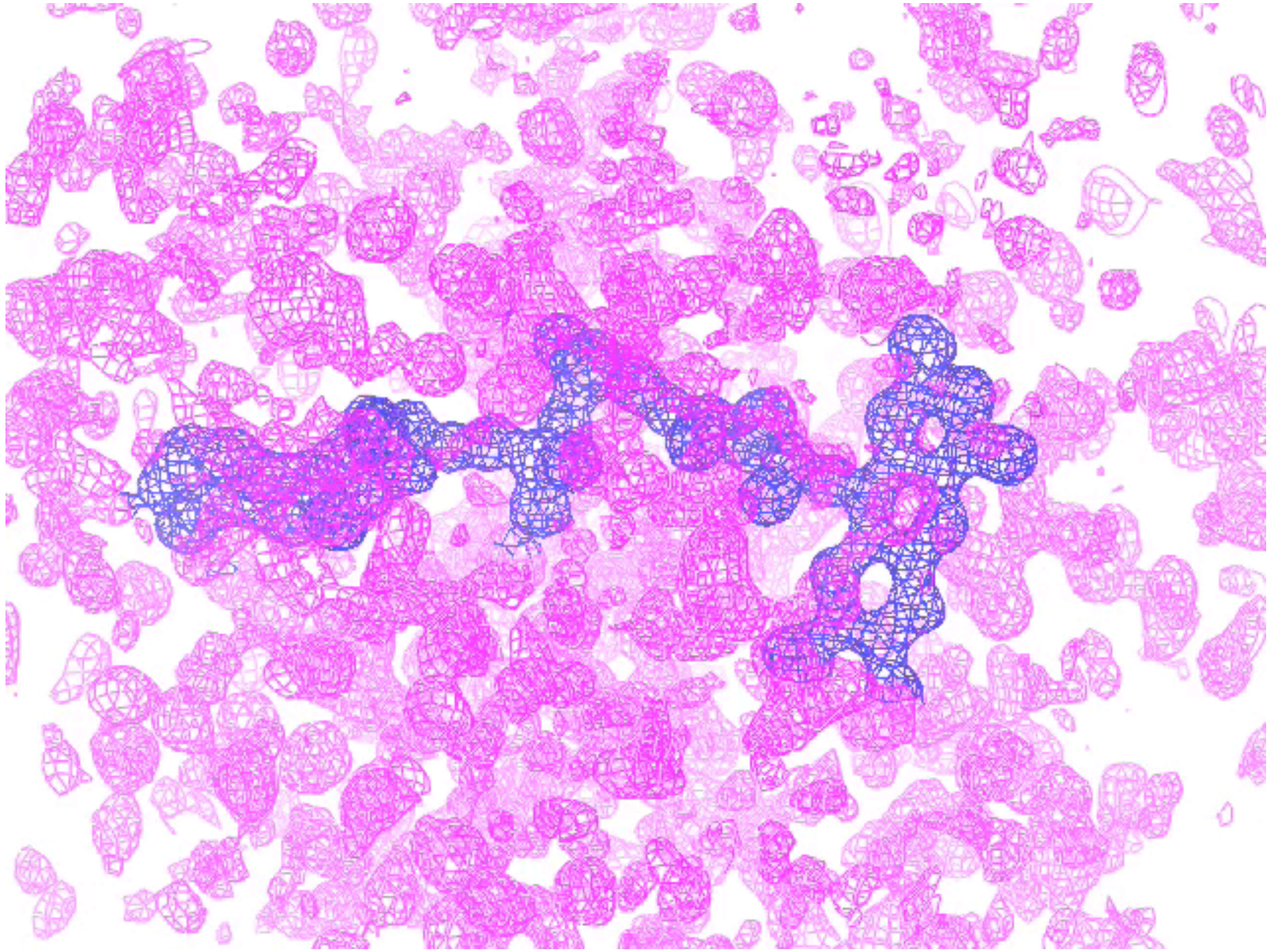


I ligand



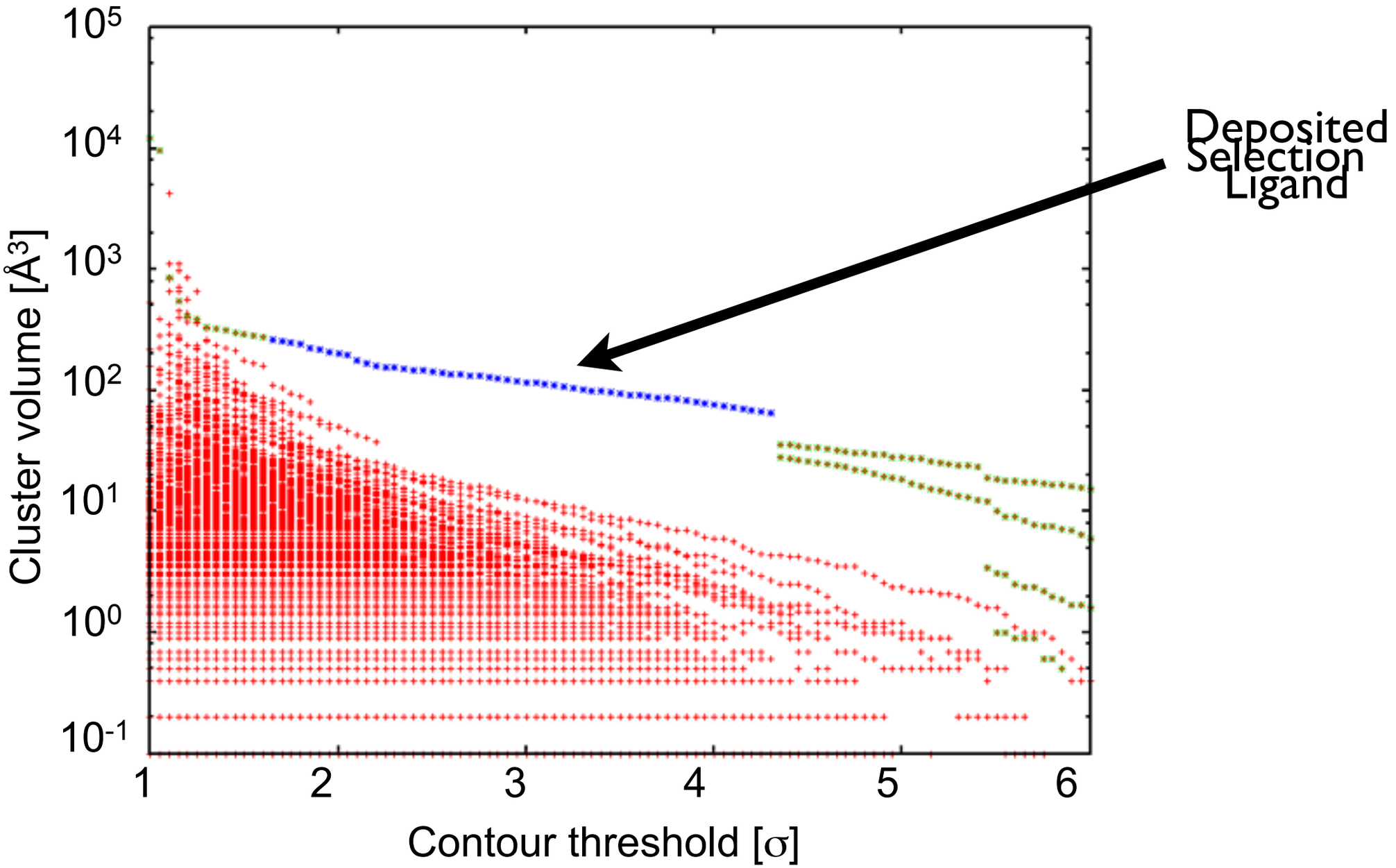
# Automatic binding site identification

- The difference density map from **low to high** contour threshold



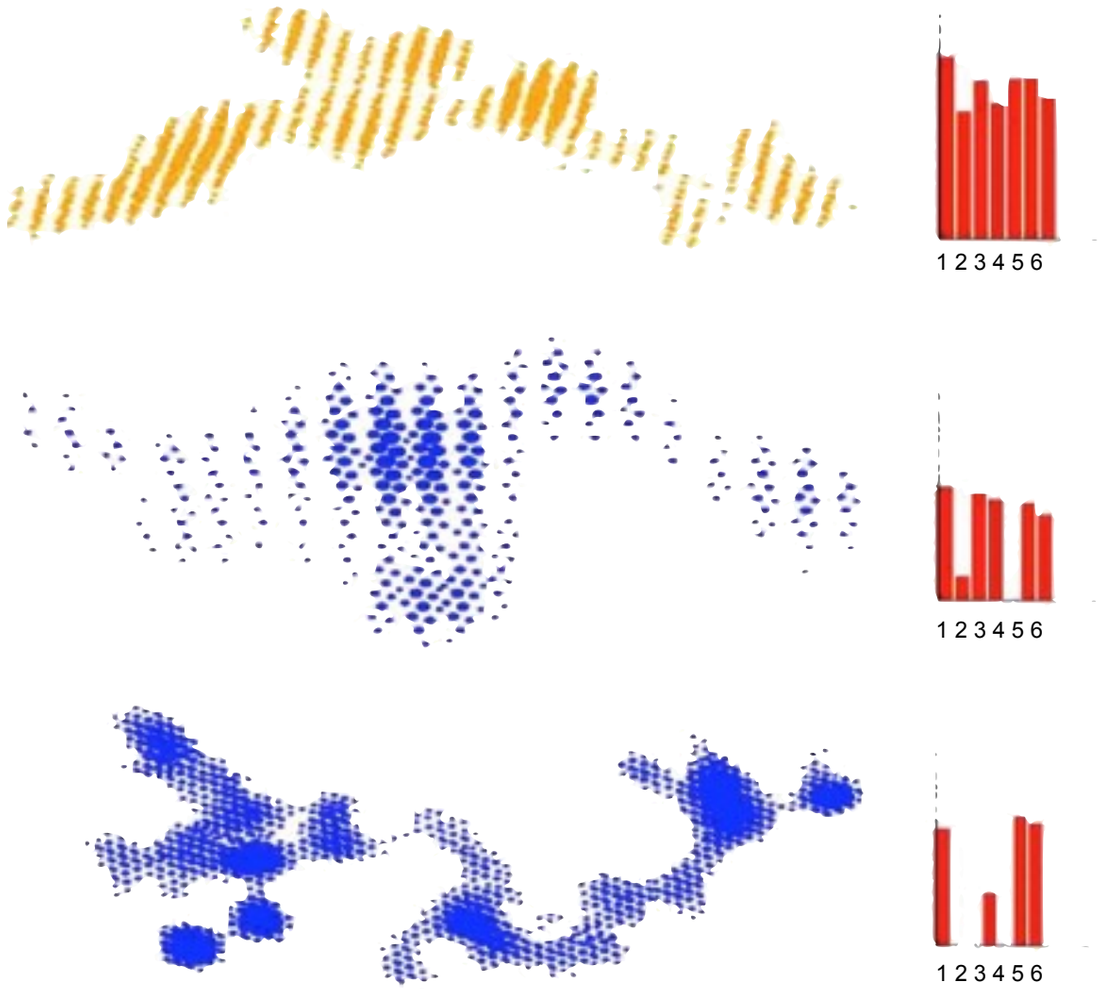
# Automatic binding site identification

- Capturing the dependence of the difference density map on changing the contour thresholds: ***the fragmentation tree***

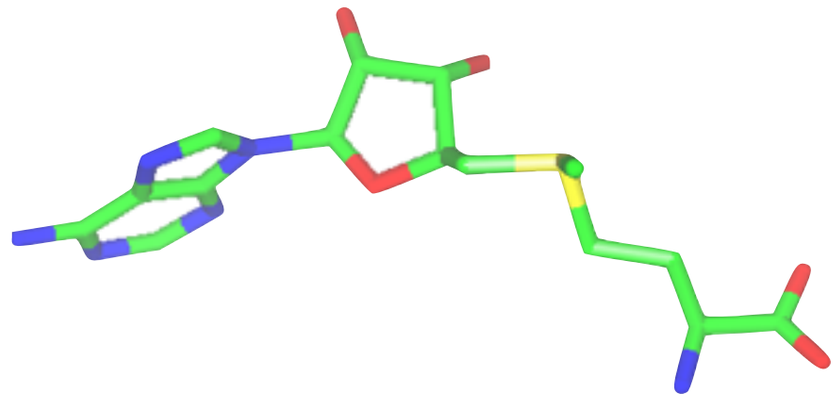


# Shape features to measure similarity

An electron density map calculated from the ligand to be fit is compared to each potential density cluster:



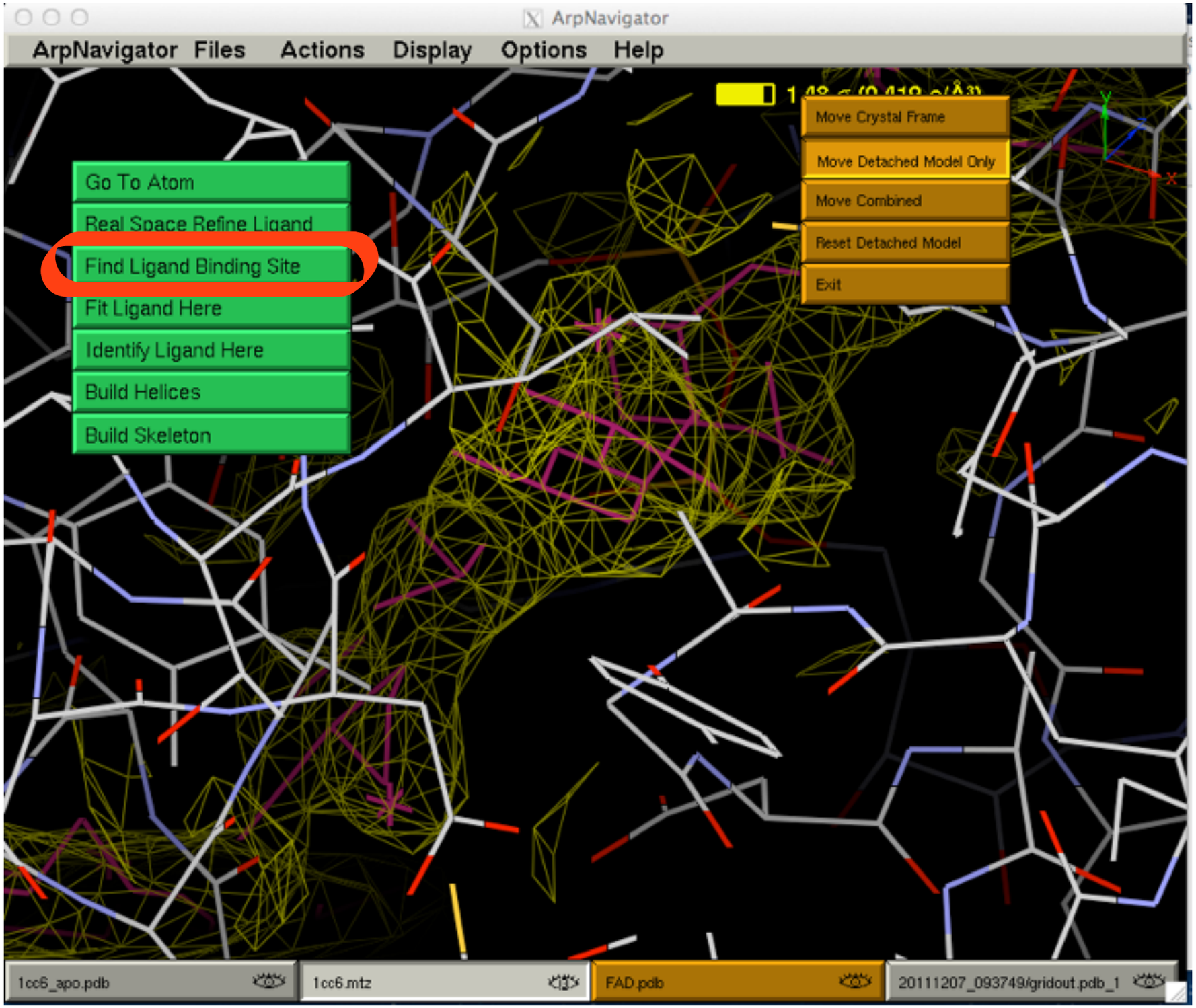

Find the matching cluster for this ligand!



1) Surface to volume ratio
2) Bounding box limits
3) Moments of inertia
4) Rotation match score
5) Eigenvalues
6) Distance histogram
7) Geodesic distance histogram

Decision is based on **feature vectors**

# Locating the binding site in ArpNavigator



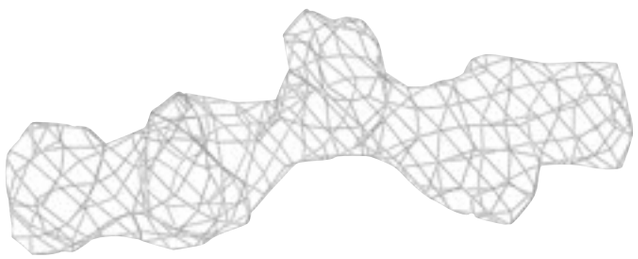
# Researcher's needs #3

We are working on protein/inhibitor complex structure. However, we did find a strange density at the active site, it looks really like GSH, the natural co-enzyme of this protein. We tried to use very simple solution to get crystal then exclude the possibility of buffer molecules, but that density is always there.

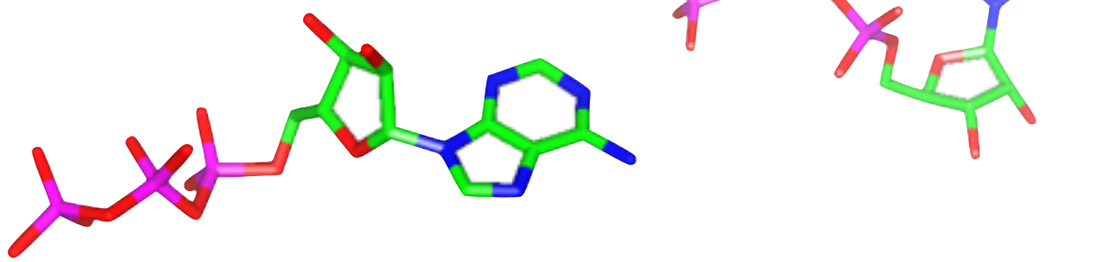
**I want to identify this ligand. Are there any methods to do this work?**

## Which case is this?

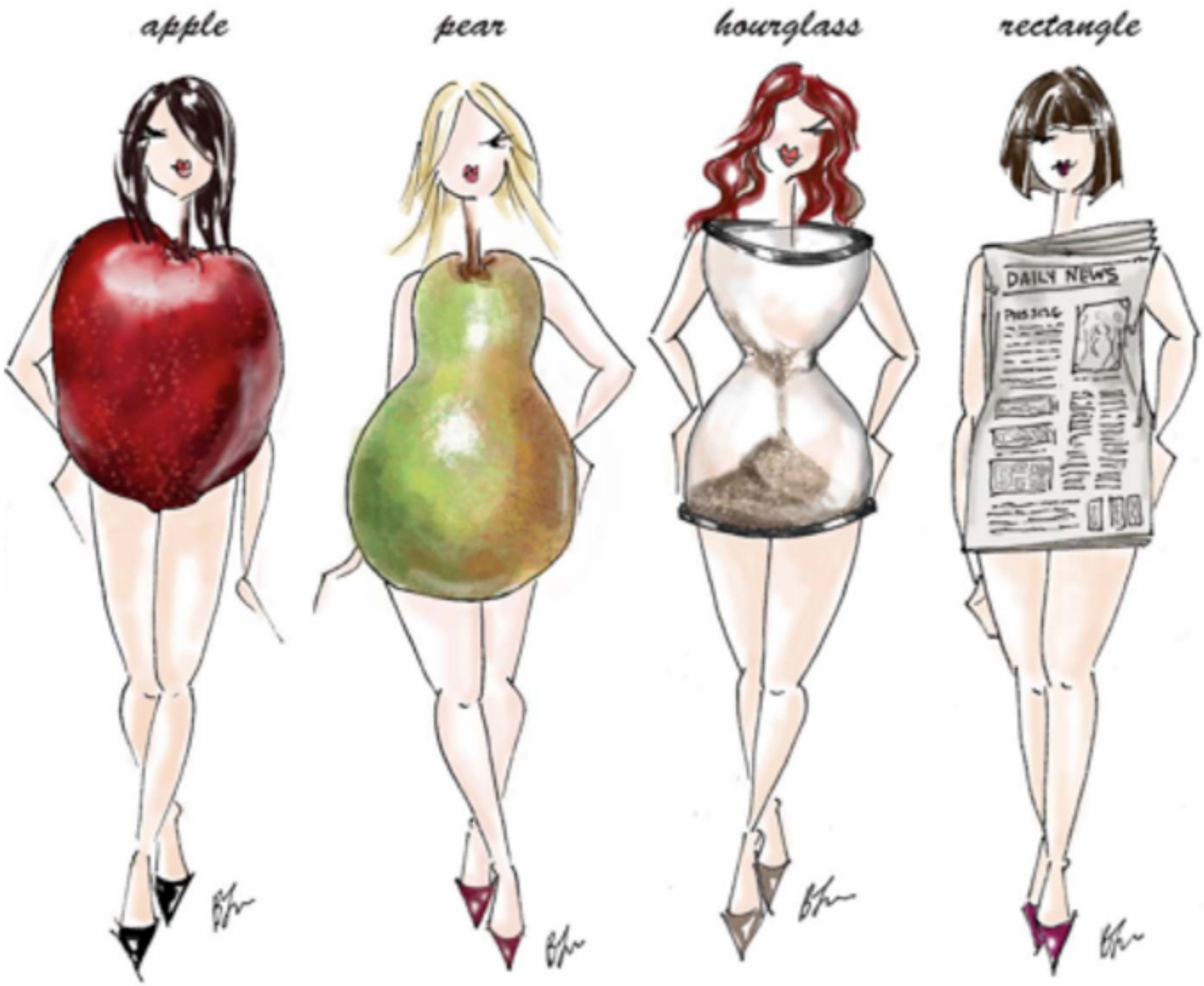
1 density cluster



N ligand candidates

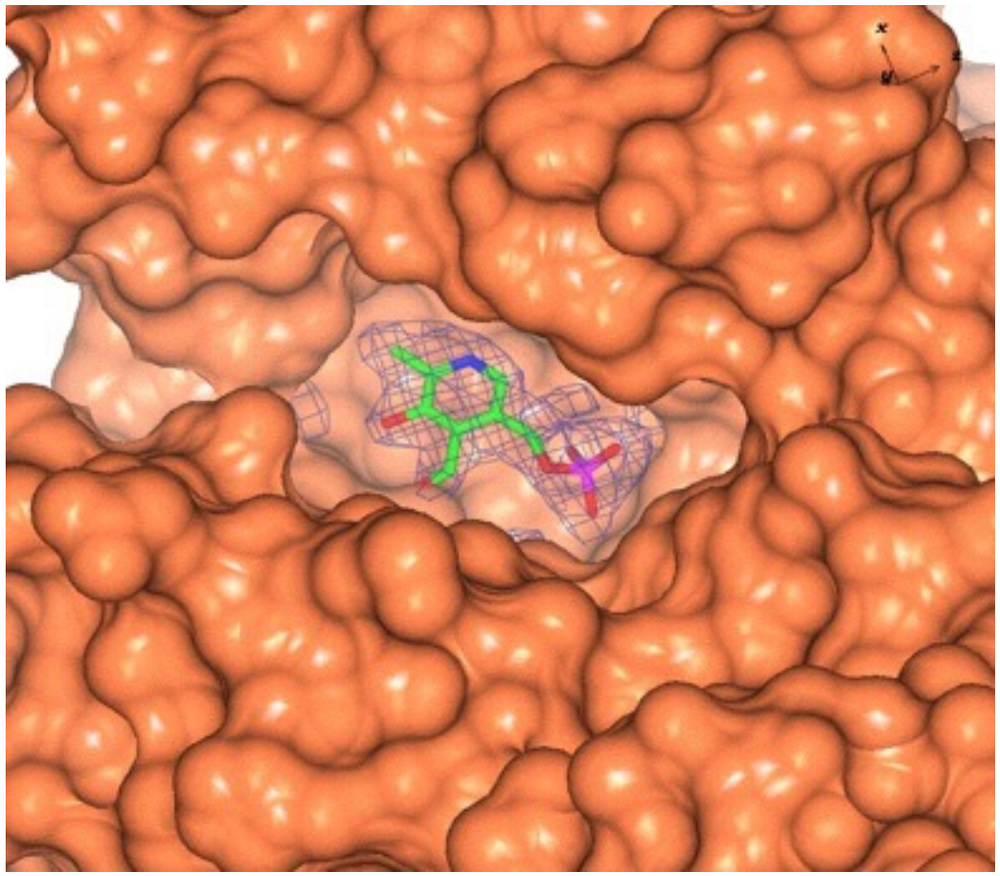


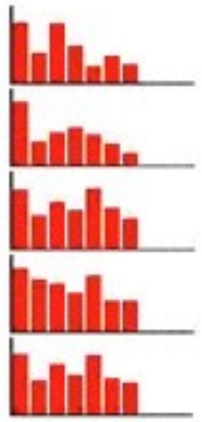
# It's all in the shape, baby!



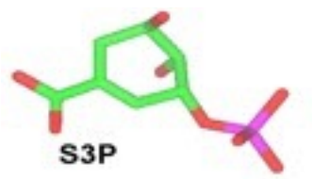
# Shape features to measure similarity

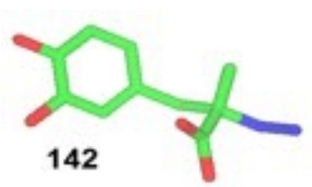
- To **assign a likelihood to a ligand** knowing the binding site



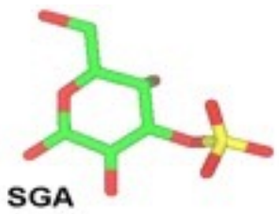


Final ligand ranking

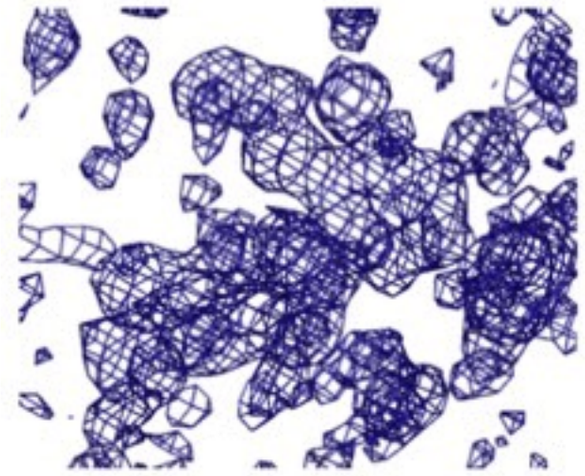





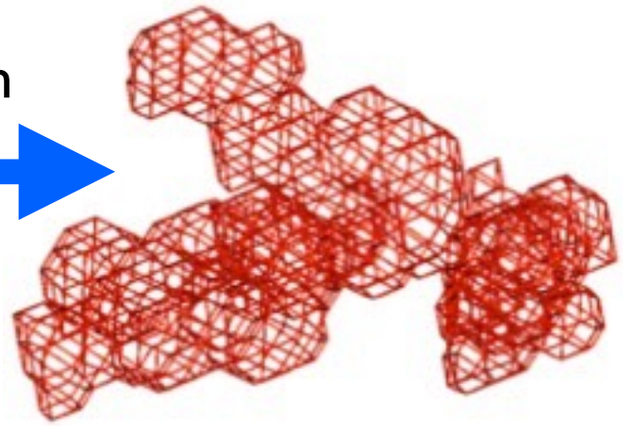


# High throughput ligand identification



Density map (or otherwise surface of protein pocket)

Segmentation



Projection/ normalisation

Calculation of shape descriptors

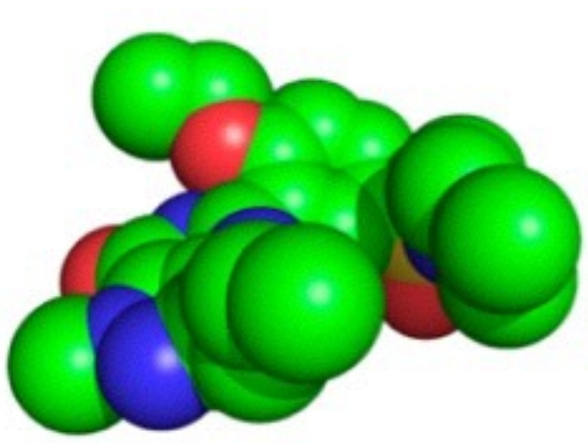



Comparison

84 of the most common ligands in crystallography

Database

Ranking



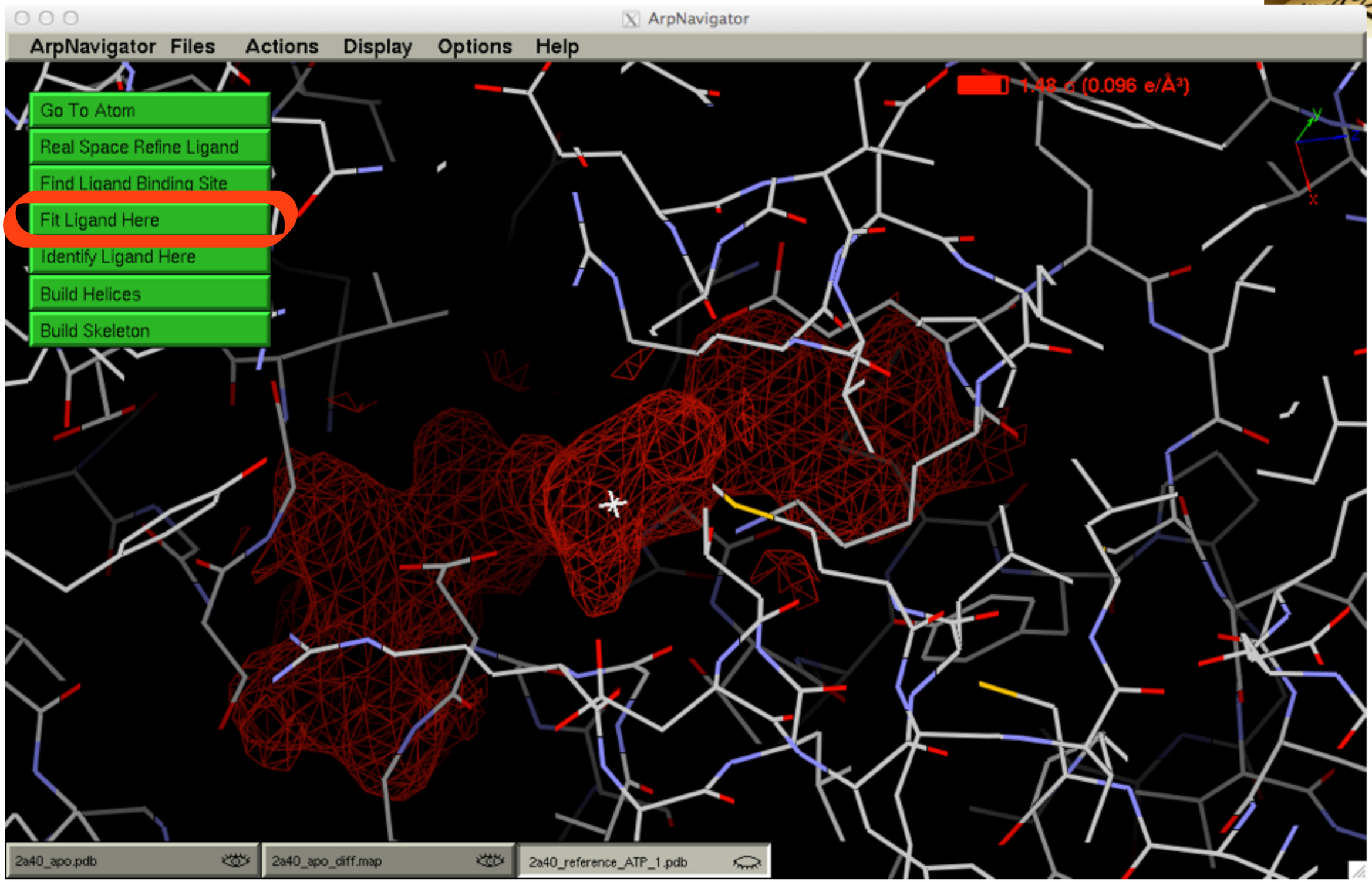
Ligand

*C. Carolan & V. Lamzin, Acta D (July 2014)*

# Ligand identification with ARP/wARP

- **<u>Flexibility - Ligand Conformations!</u>**
  - full ligand flexibility is considered

# Ligand identification in ArpNavigator



# Researcher's needs #4

I have a ligand in the binding pocket and mediocre data. **The core of the ligand is well defined, however there are chains/tails of the ligand which are not**….The question here is how the chains/tails should be modelled (if at all)

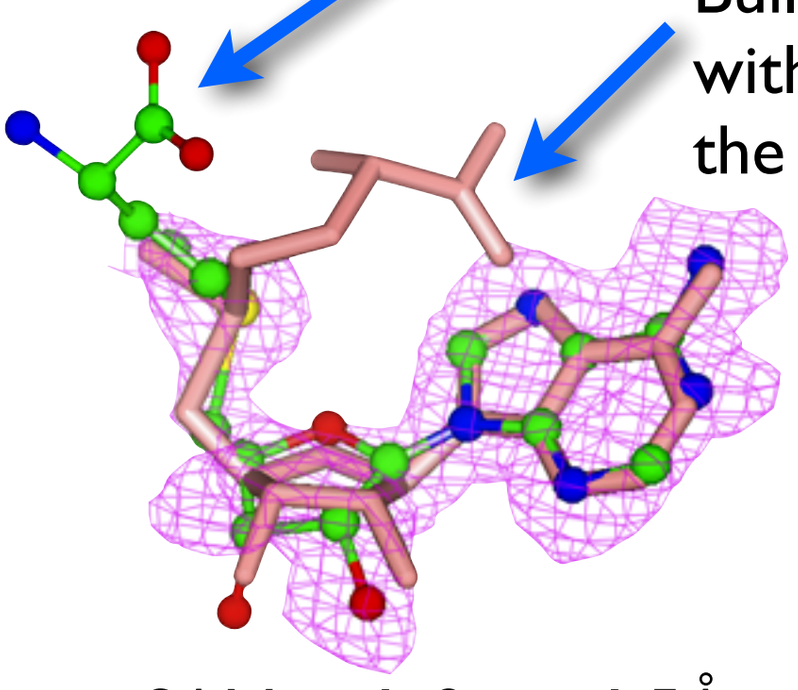
**Which case is this?**

None of the Above

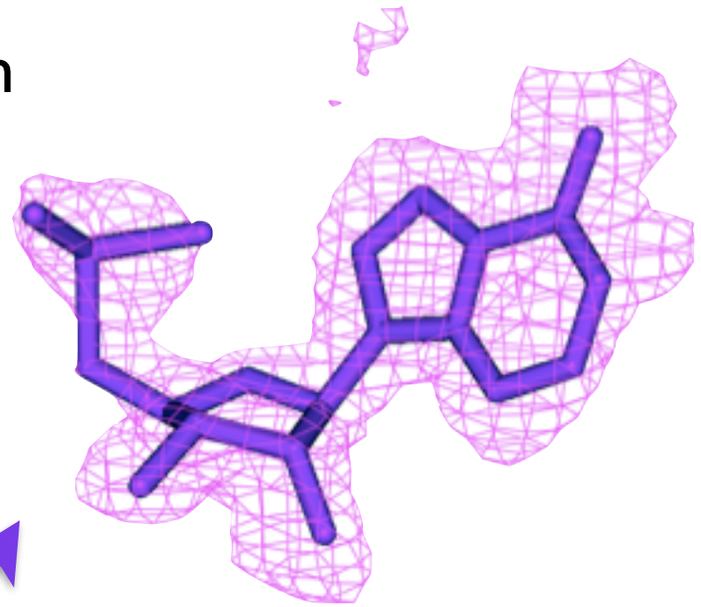
# Partial Disorder, Partially Occupied Ligands

Deposited in PDB

Built automatically with ARP/wARP when the full ligand is given

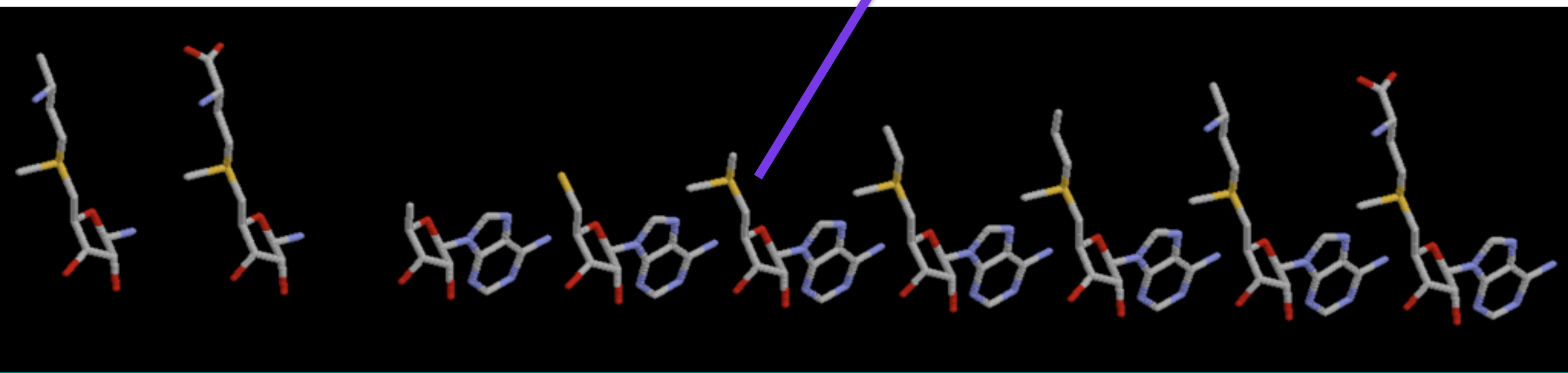


SAM in 1v2x at 1.5Å

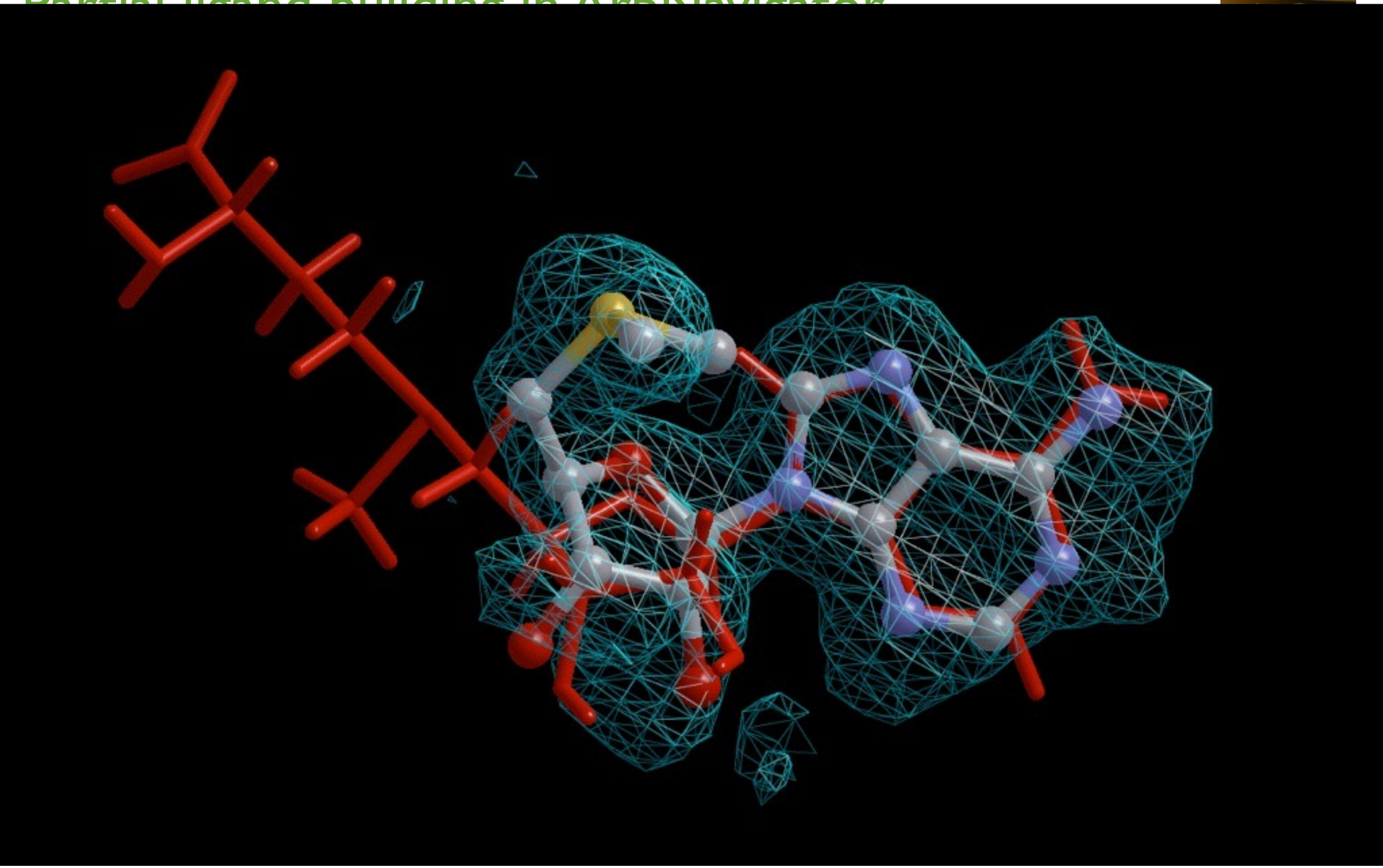


Compound chosen in cocktail case

Artificial Cocktail



# Partial ligand building in ArpNavigator



# Success rate

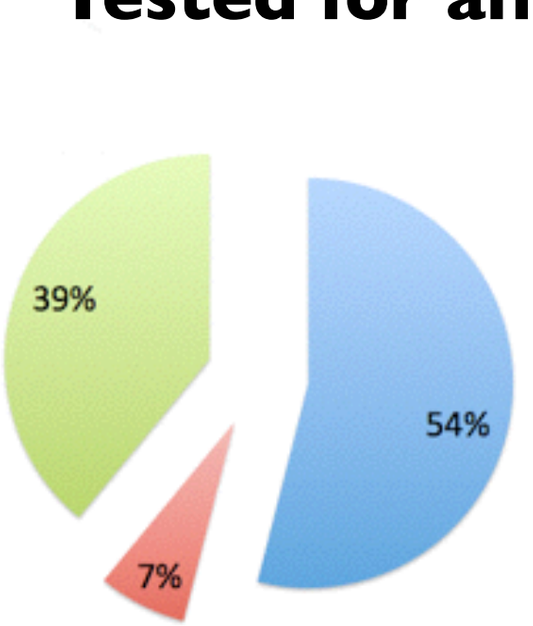
- Tested on > 20k PDB entries from EDS
  - X-ray resolution limits: 1.0 to 3.0 Å
  - Building the **largest fully occupied ligand**
  - Correctness criterion: r.m.s.d. < 1.0 Å

incorrectly identified binding site

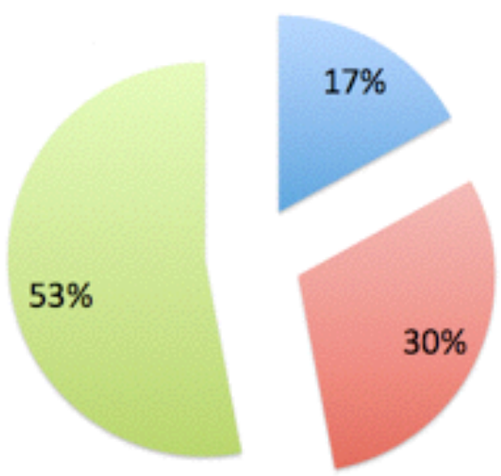
incorrectly built ligand at correct binding site

correctly built ligand

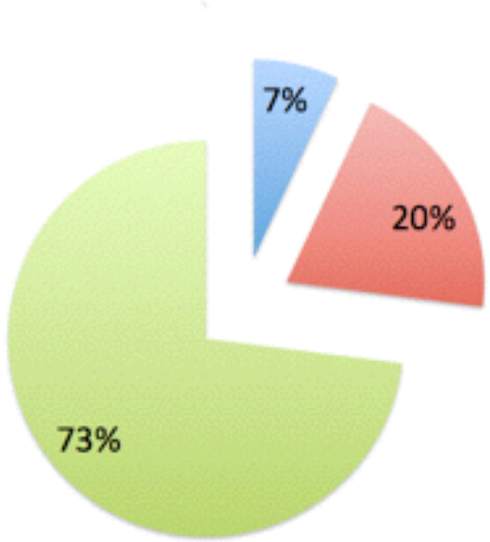
**Tested for an older version…**



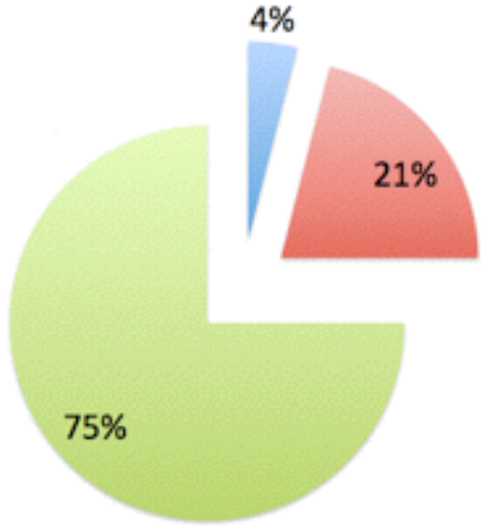

5…6 atoms,
any map corr.




7…100 atoms,
any map corr.



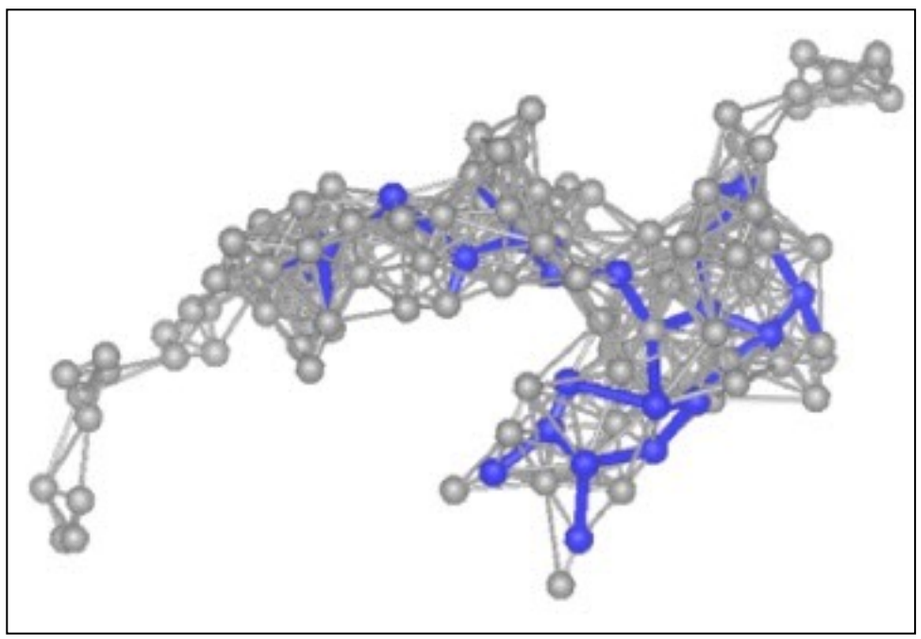

10…40 atoms,
map corr. > 80%



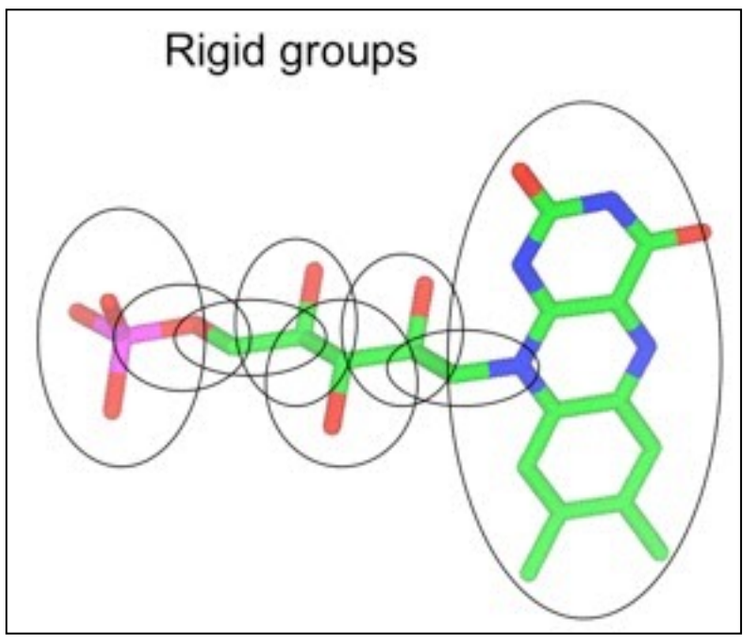

20…40 atoms,
map corr. > 80%

# Ligand building methods

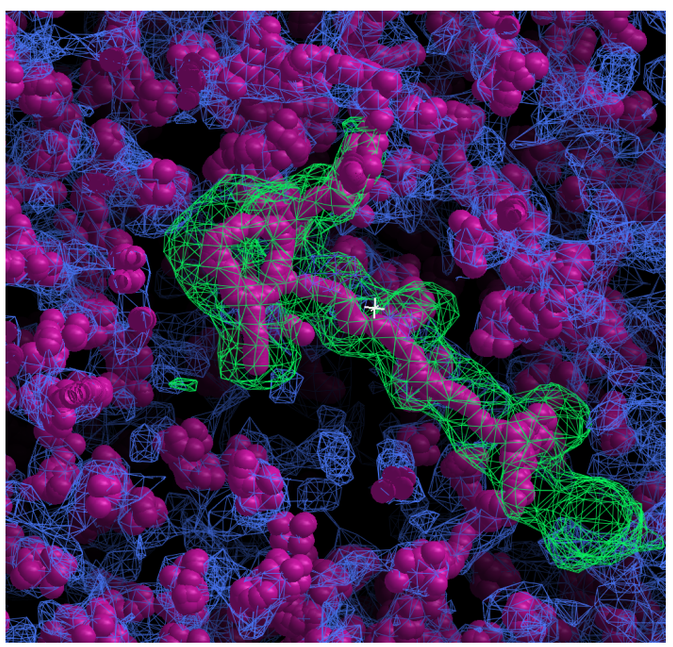
## Sparse grids

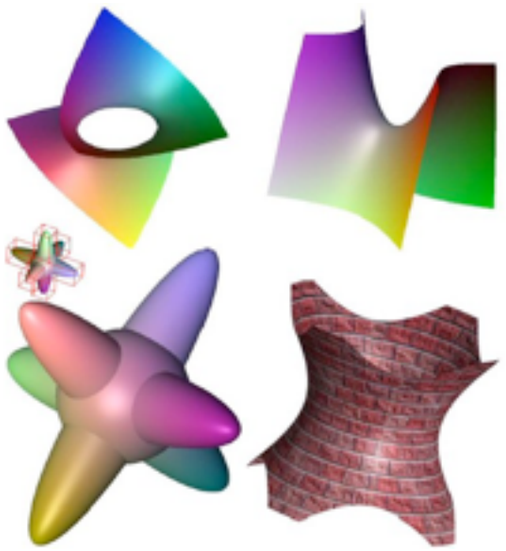


## Conformational fit




## Fine skeletons





- High-throughput ligand identification using shape descriptors

**Building models which do NOT have repetitive motifs like peptides or nucleotides**

# The people - The power!


## Collaborators

**EMBL Hamburg:** Gleb Bourenkov, Santosh Panjikar

**MRC Laboratory, Cambridge :** Garib Murshudov's group

**Rutherford Appleton Laboratory:** CCP4 team

## Previous Members

Ciaran Carolan, Tim Wiegels

## Funding