

ARP-Ligand – Locating, identifying and building ligands in electron density

Ciarán Carolan
EMBL-Hamburg



Alternative title...

ARP Ligand

Has my
soaked
ligand
bound??

What is the
binding
mode??

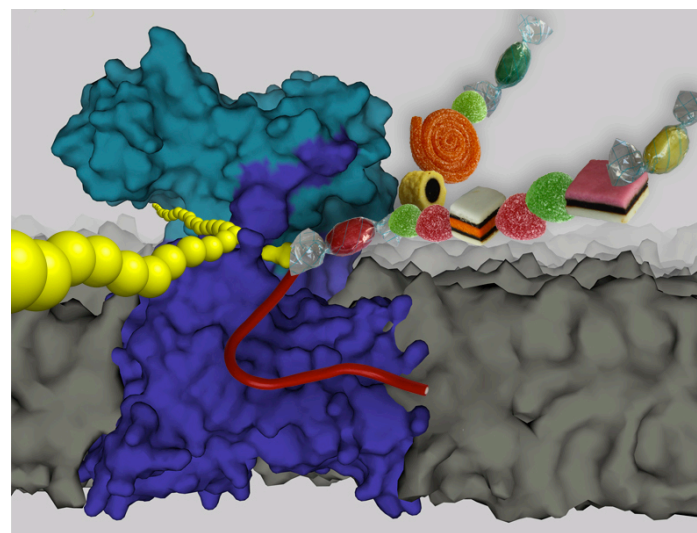
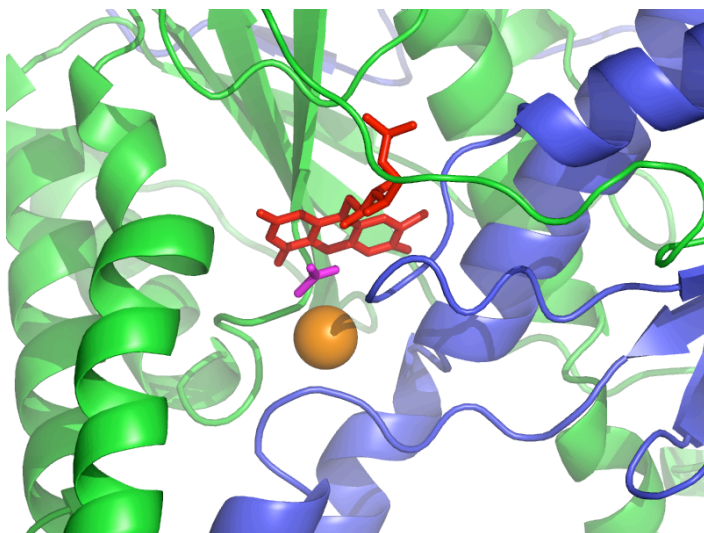
If so, where is
the binding
mode??

Which of my
soaked
ligands has
bound??

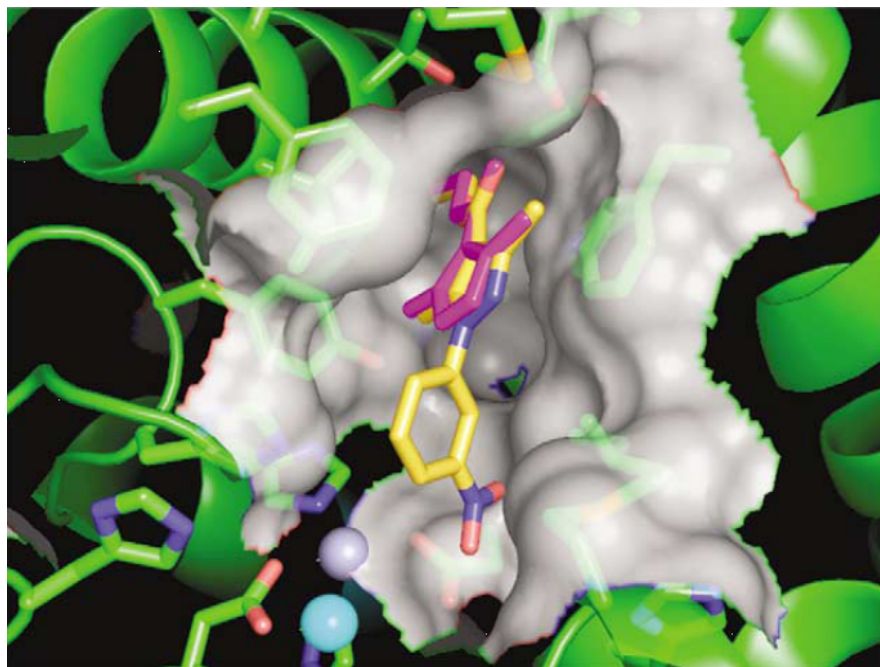
*Fulfilling all your (ligand)
desires since 2004*

Why Ligands?

- Endogenous substrates and analogues thereof
- Known drugs and biological modulators
- Probe molecules
- Post-translational modifications
- Buffer constituents



Structure based drug design



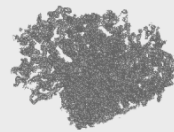
Purple: A lead for the inhibition of PDE4D;

Yellow: the end result from SBDD.

Card et al. (2005) *Nat. Biotechnol.*
23: 201-207

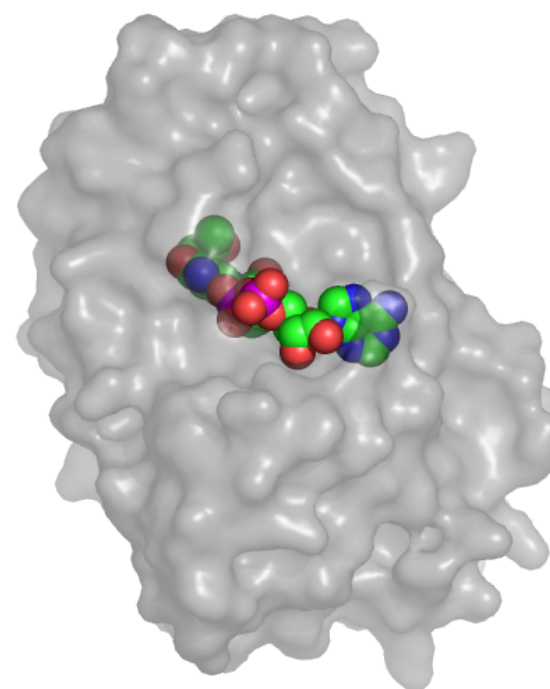
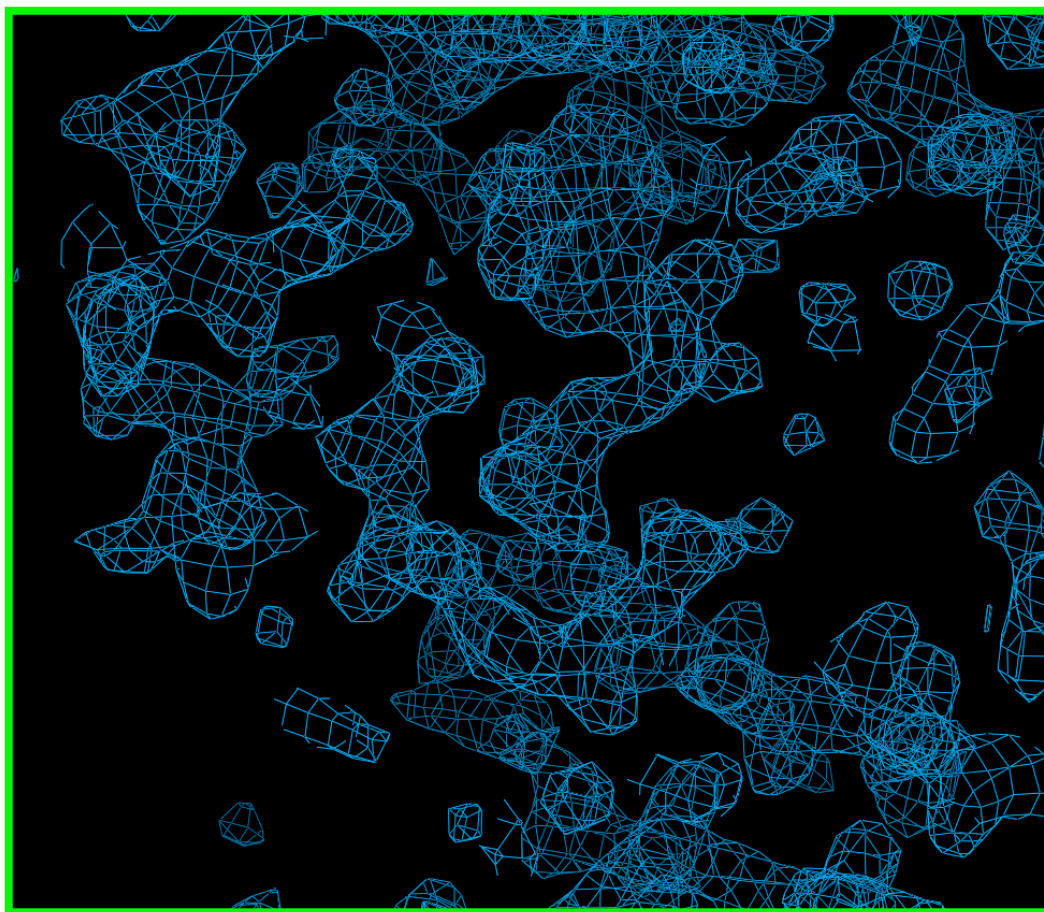
- Structural work, principally involving MX and NMR, allows the elaboration of a lead to a candidate drug.
- High-throughput crystallography is an up-and-coming technique in lead generation (e.g. fragment based drug design). It requires rapid and automated ligand identification and construction, however.
- Elaboration of a lead to produce a drug depends on accurate knowledge of ligand binding modes, and can also be aided by automated ligand building procedures.

The task at hand



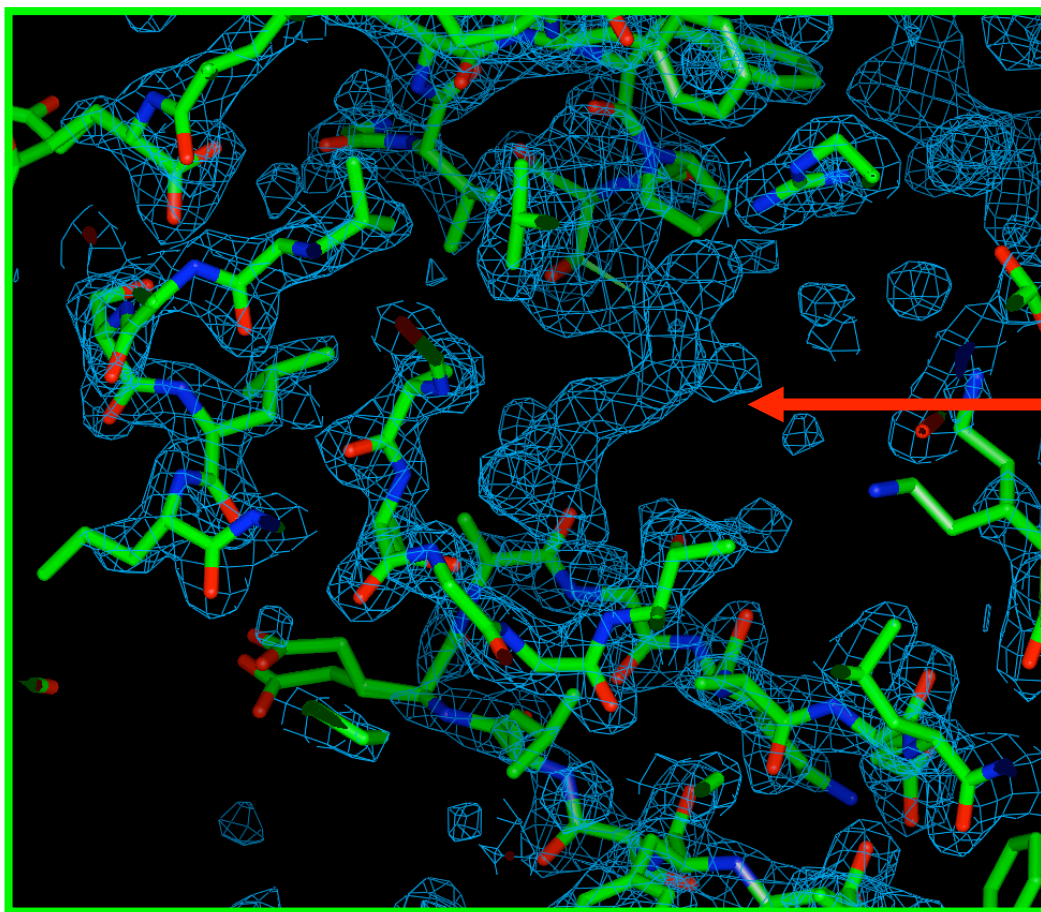
ARP/wARP and Ligand Building

electron density map (good resolution)



ARP/wARP and Ligand Building

electron density map (good resolution) + protein
model

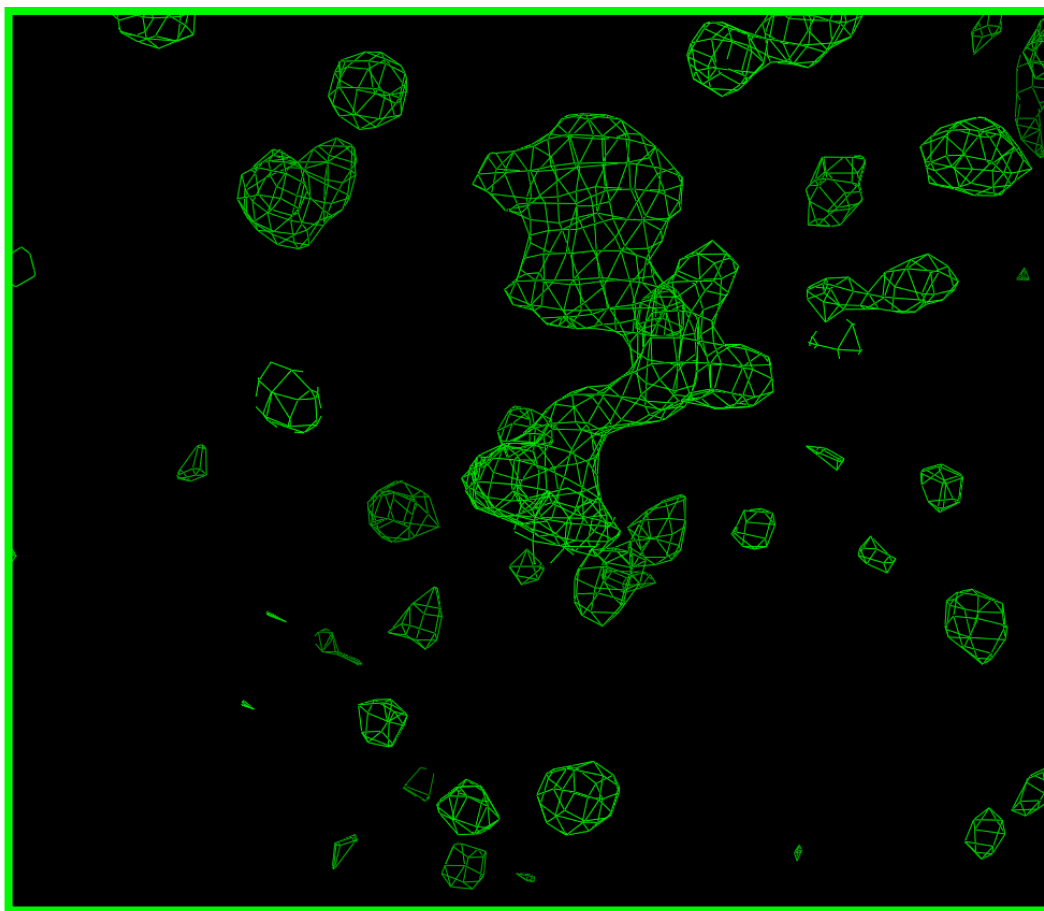


excess density

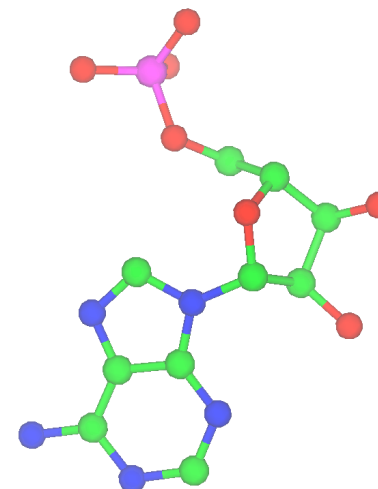
not
accounted for
by the protein

ARP/wARP and Ligand Building

difference density map...

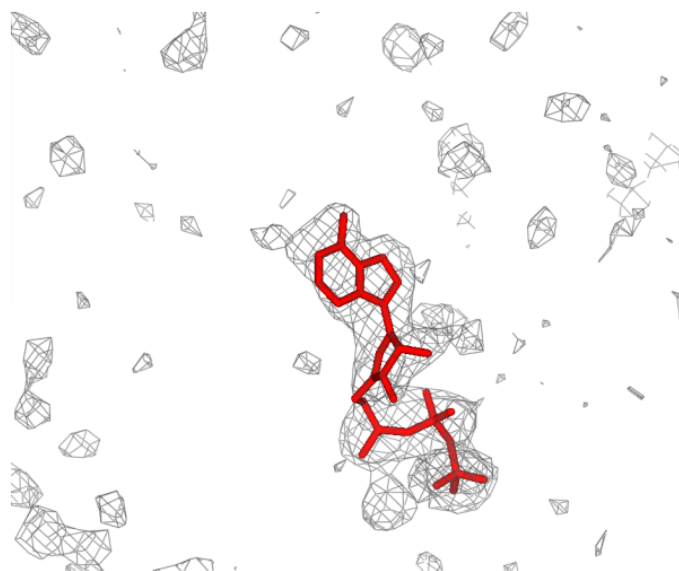


... into which we have
to
FIT
the
APPROPRIATE
ligand



The task of ligand fitting

Challenges to be addressed...



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

The protocol of a modeling task

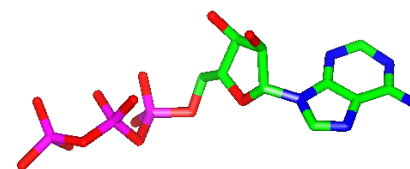
Fundamental initial knowledge determines the protocol.

		Binding site	
		known	unknown
Ligand identity	known	X	
	unknown		

1 density cluster

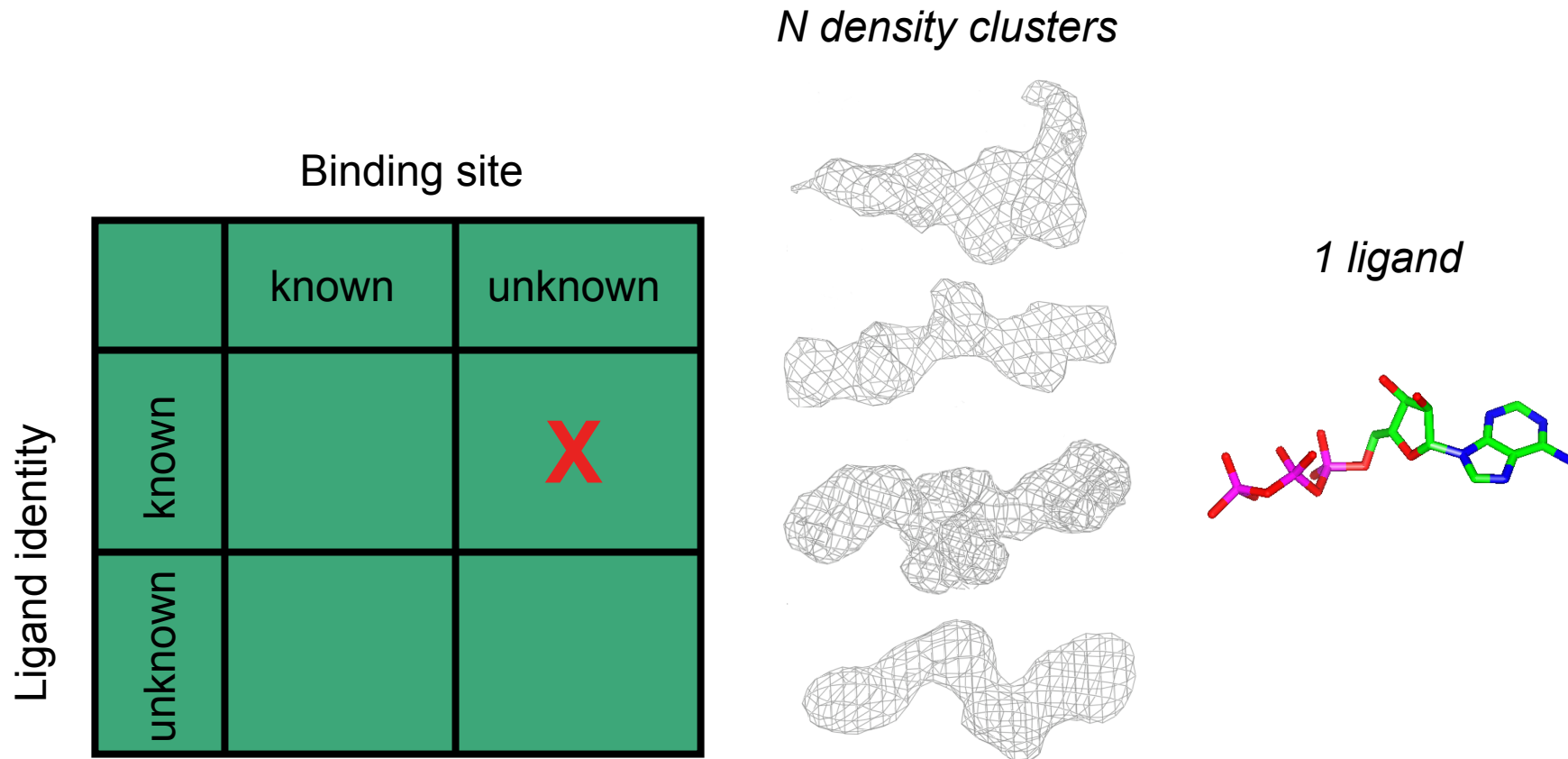


1 ligand



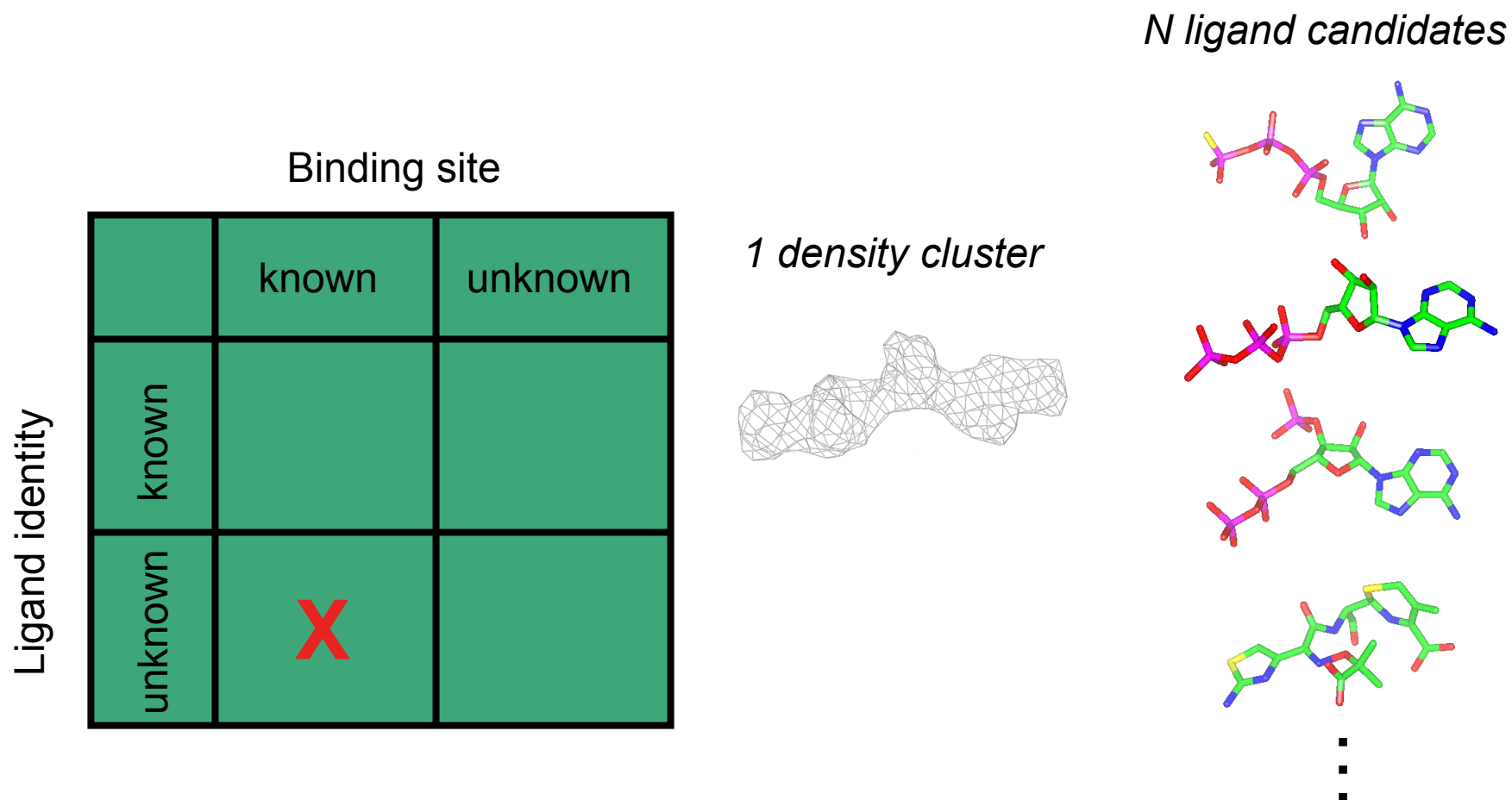
The protocol of a modeling task

Fundamental initial knowledge determines the protocol.



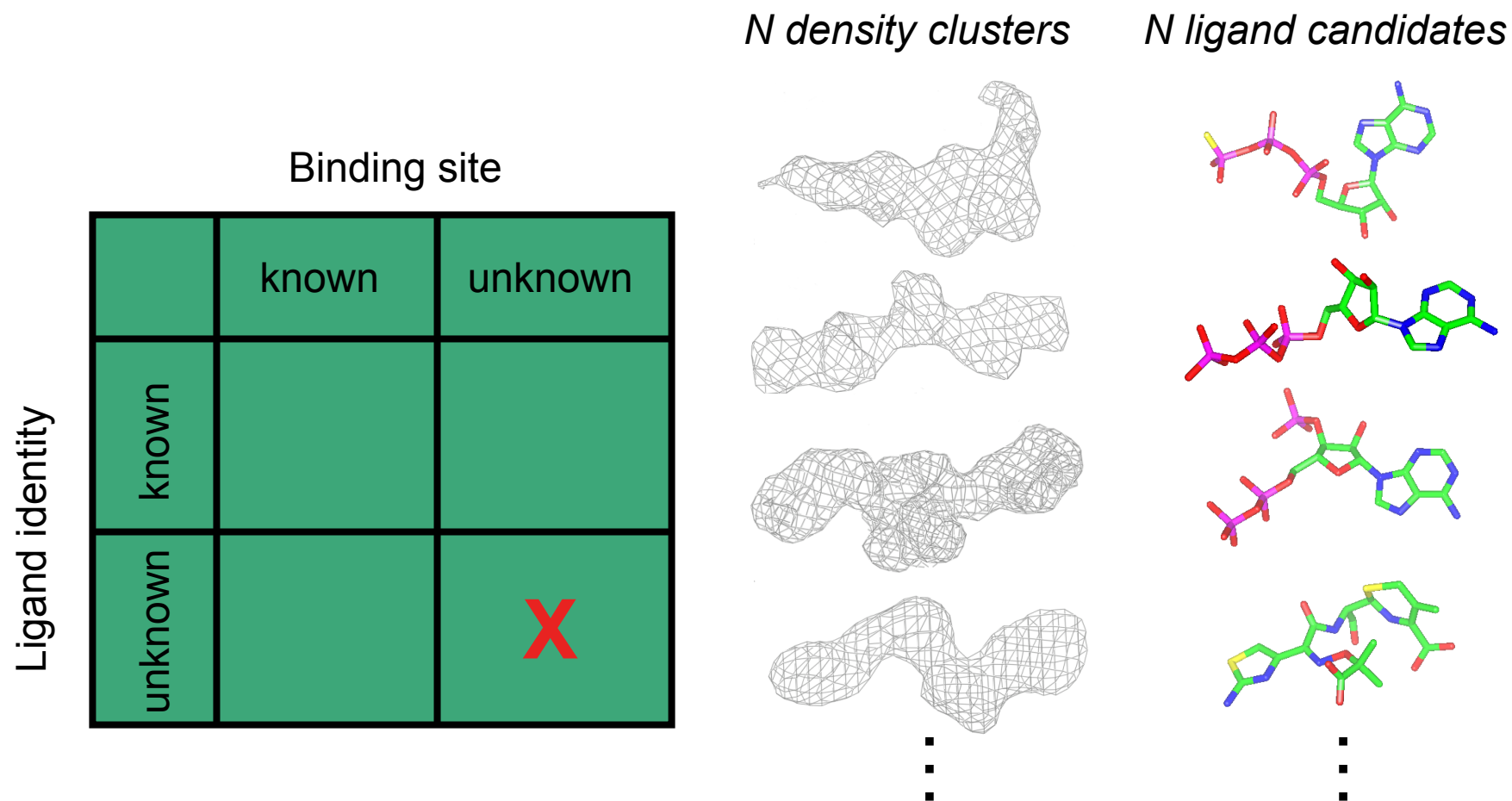
The protocol of a modeling task

Fundamental initial knowledge determines the protocol.



The protocol of a modeling task

Fundamental initial knowledge determines the protocol.



Researcher's needs - #1

CCP4BB: Sep 30th 2008

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.

*Thanks,
Anshul Awasthi*

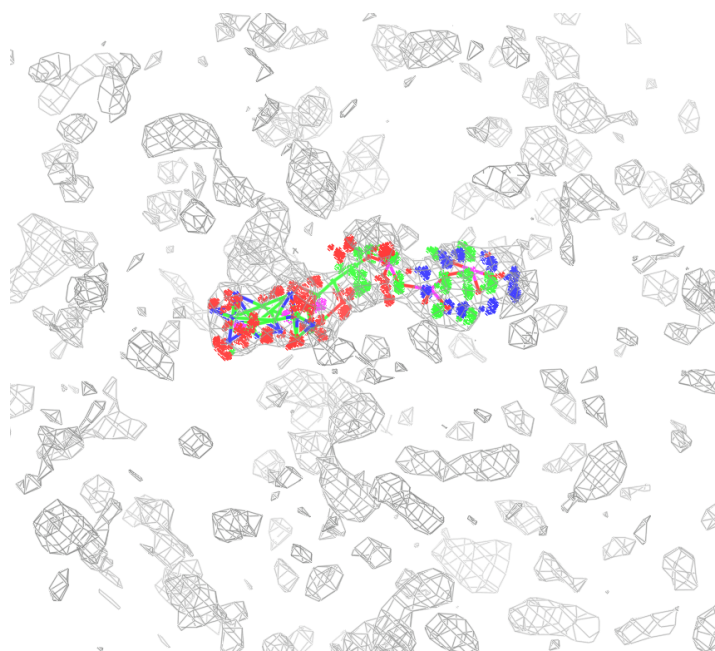
Constructing the ligand

- To address more broadly the different ligand sizes and complexities encountered, to increase the robustness of the software...

2 separate construction methods get used:

- ❑ *Label swapping on a sparsed grid*
- ❑ *Metropolis search in conformation space*

The workflow of ligand fitting in ARP/wARP – The simplest case



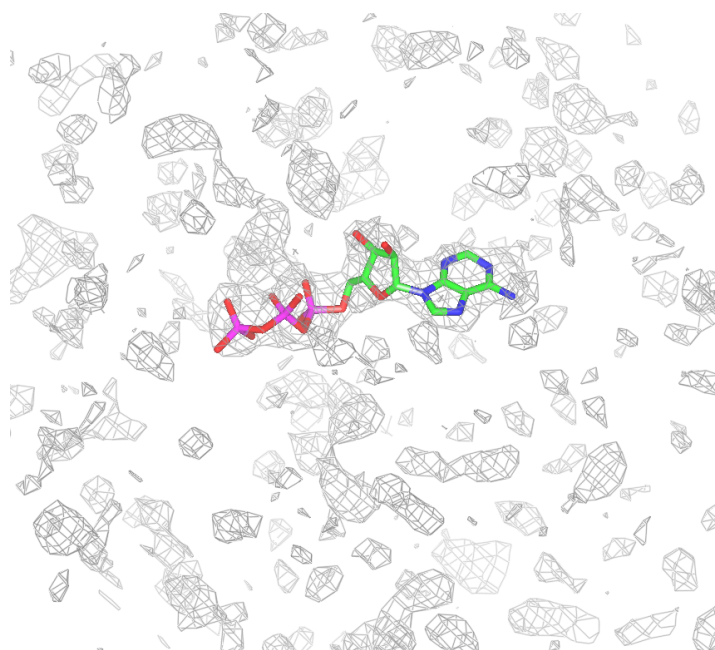
Prepare

- Binding site identified by user.
- Sparse density map and generate ligand topology.

Construct

- Construct ensemble of ligand models in plausible conformations to fit sparsed density.

The workflow of ligand fitting in ARP/wARP – The simplest case



Prepare

- Binding site identified by user.
- Sparse density map and generate ligand topology.

Construct

- Construct ensemble of ligand models in plausible conformations to fit sparsed density.

Refine

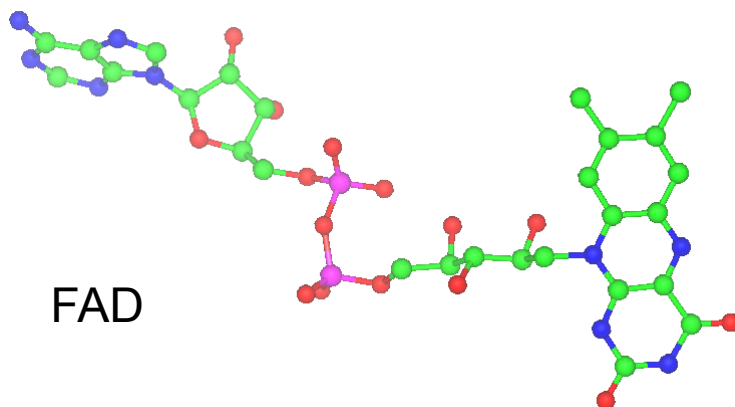
- Refine ligand coordinates to satisfy geometric constraints and to maximise fit to density.
- Choose best model.

Constructing the ligand: Graph search

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

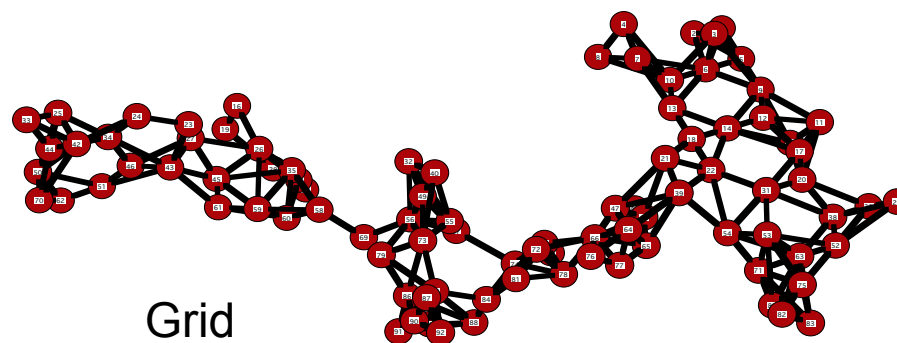
Knowledge about ligand

- A) Connectivity
- B) Distances
- C) Angles
- D) Chirality



Knowledge about grid

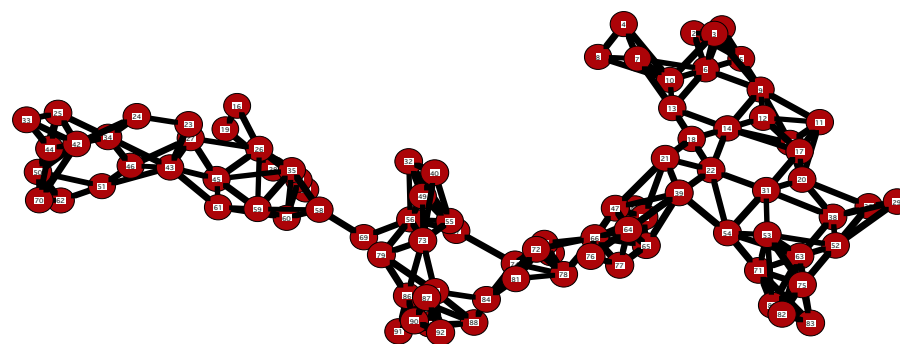
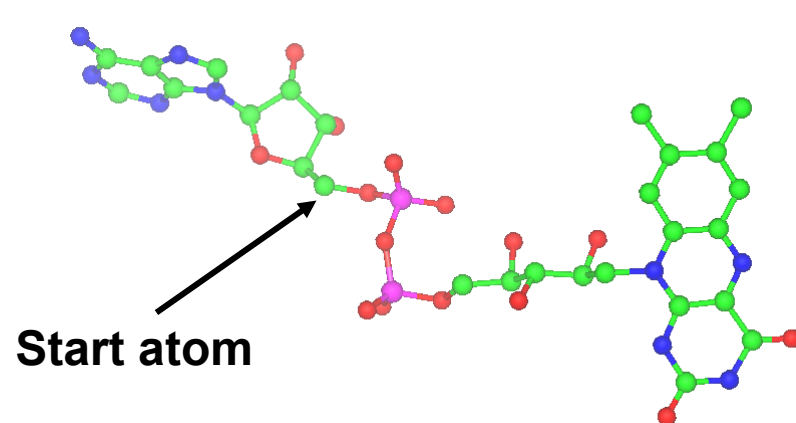
- A) Connectivity
- B) Distances
- C) Density
- D) *NO IDENTITIES!*



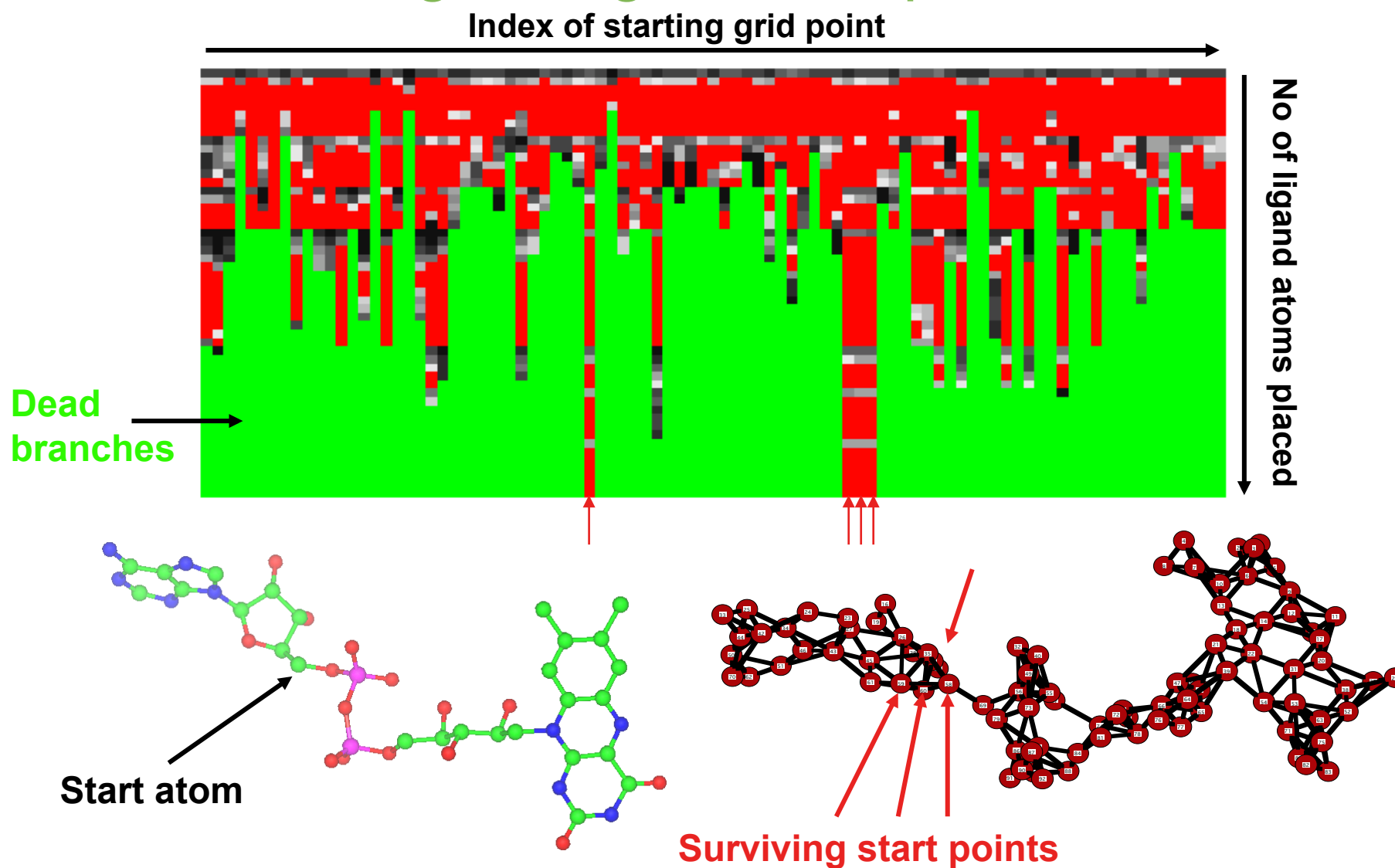
Constructing the ligand: Graph search

‘Label swapping’

- The 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



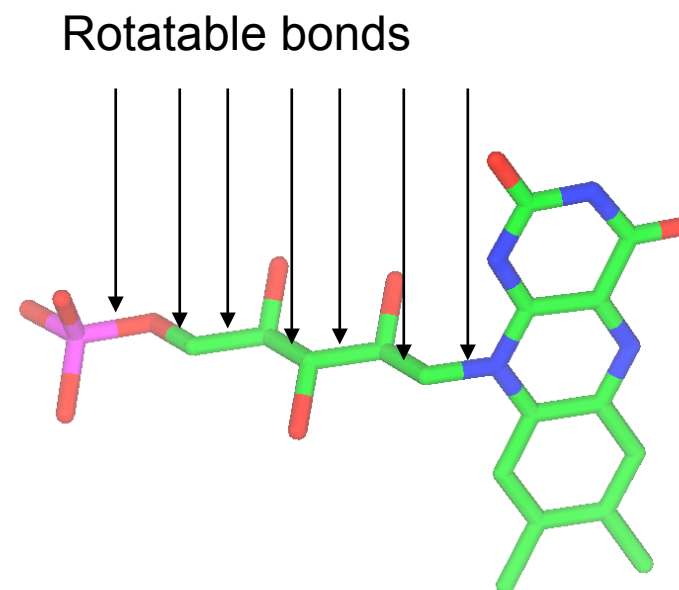
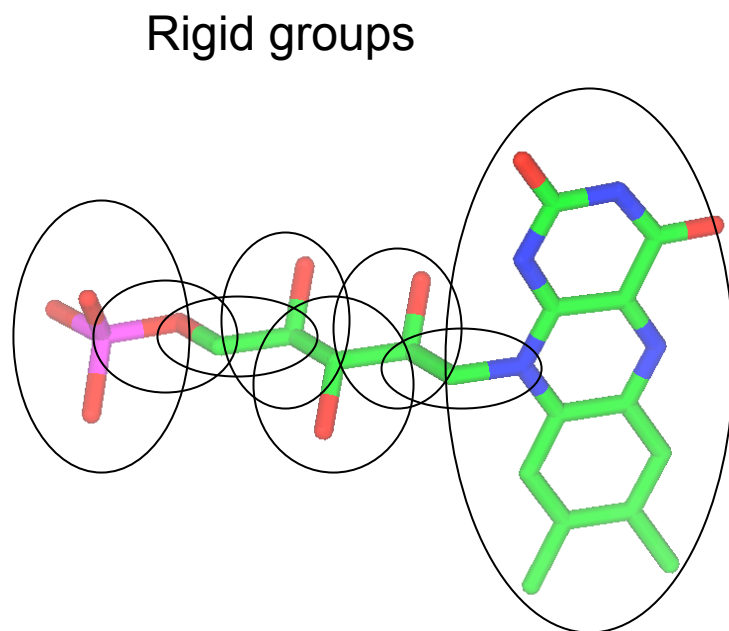
Constructing the ligand: Metropolis search

Perform a random walk in parameter space biased towards the optimum of a score function.

Parameter space: Position, Orientation and Conformation.

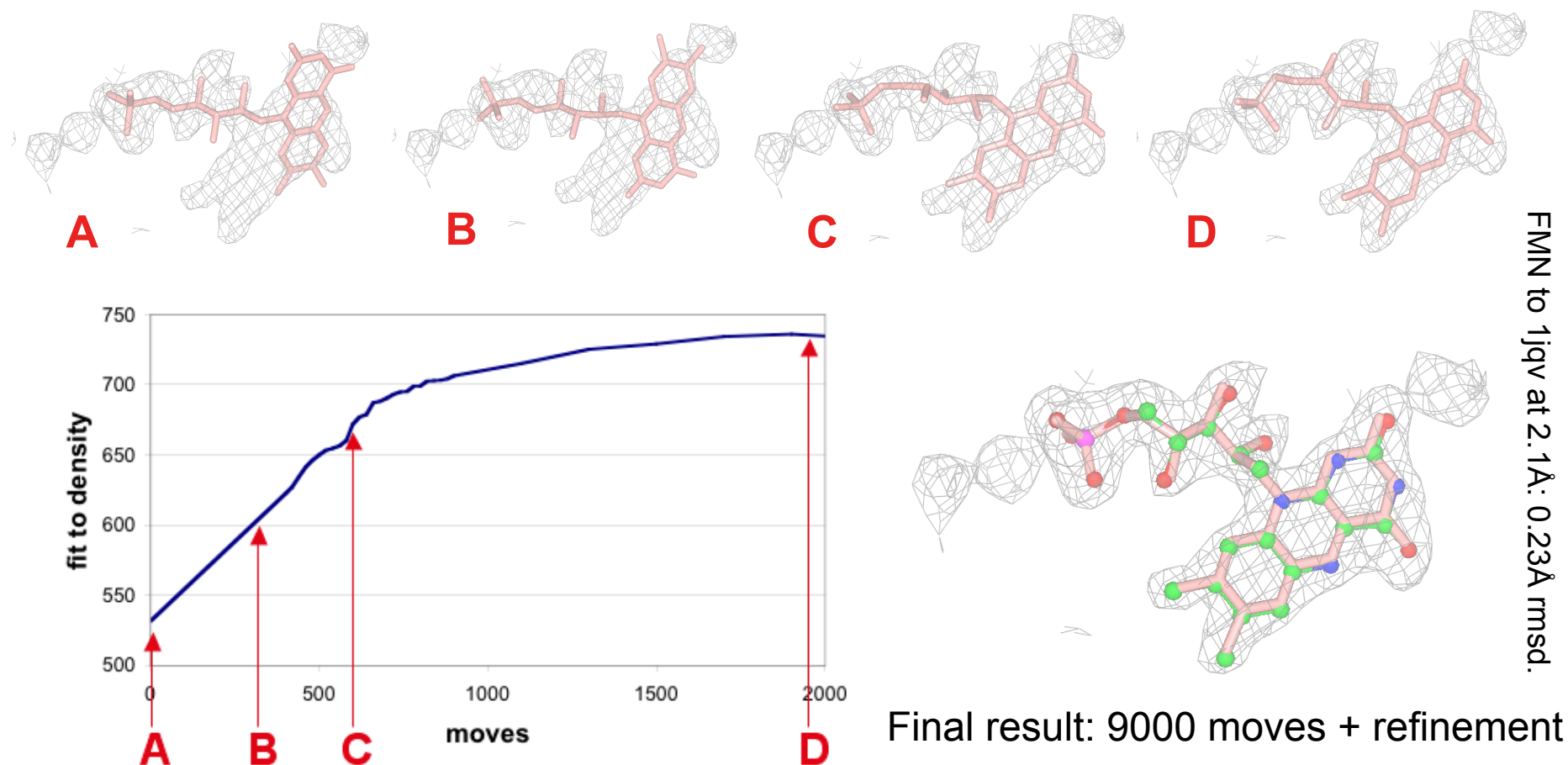
Score function: 'Pseudo' map correlation.

Advantage: Less degrees of freedom.



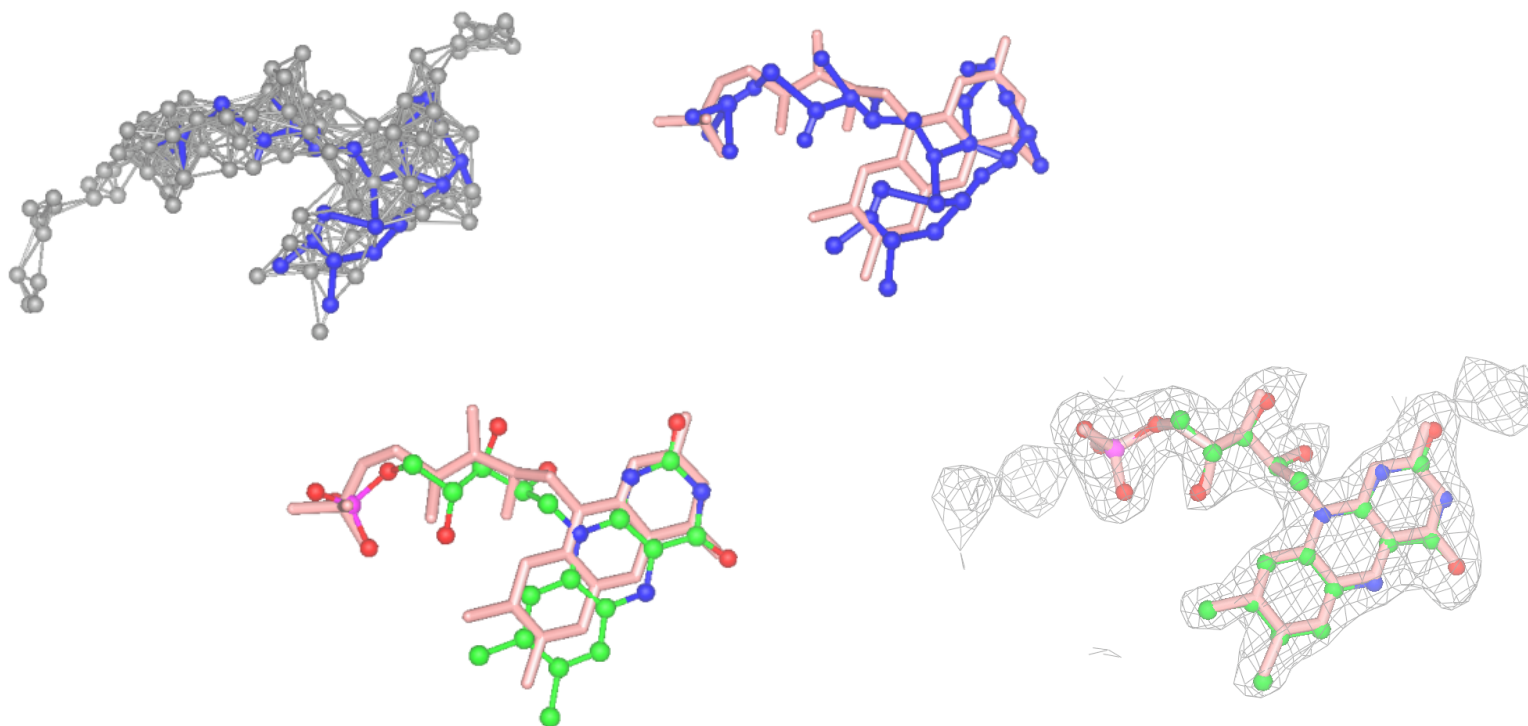
Construction with a Metropolis method

- The evolution of a model during a metropolis optimisation.



Real space refinement

- Uses the full density and topology information
- Employs a gradient method to optimise fit to density and ligand geometry simultaneously



Running a task through the CCP4 interface

The ligand building as part of ARP/wARP has its own GUI.

ARP/wARP Version 7.2.0: Ligand Build

Help

Job title demonstration case

Build the ligand around an approximate XYZ position

MTZ in PROJECT Browse View

Fobs Sigma

Protein model without ligand PROJECT Browse View

Ligand molecule coordinates PROJECT Browse View

Ligand output file PROJECT Browse View

Parameters

Use with REFMAC5 the Fast protocol and do not use the Free R flag

Search for ligand around X 0. Y 0. Z 0. , in a radius of 5.0

2 Ligand building cycles

Assume partial occupancy of ligand

Refmac parameters

Crystal parameters

Test and comparison parameters

Run Save or Restore Close

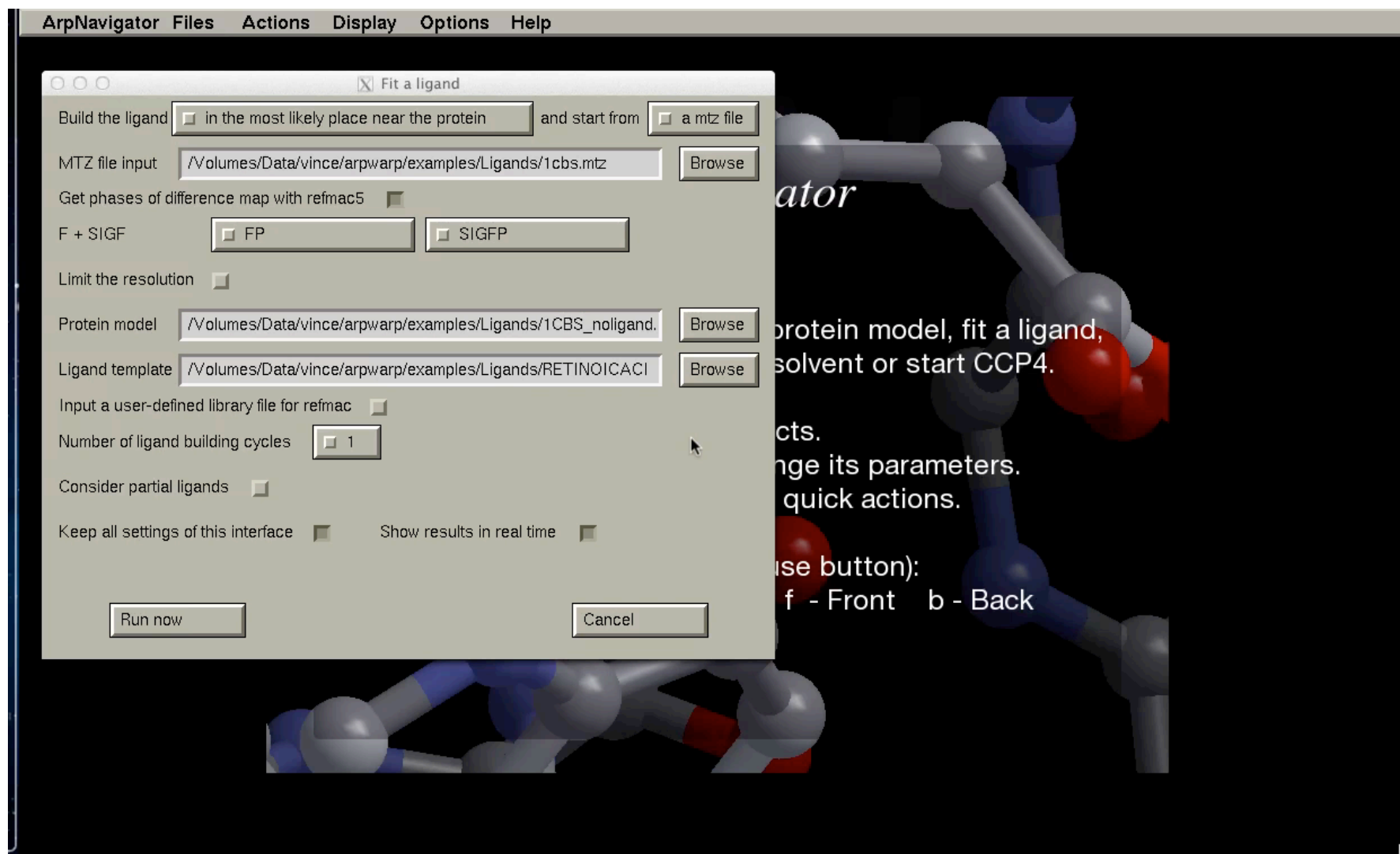
Protocol

Diffraction data

Protein without ligand

Template ligand

Ligand Building via ARPNavigator



Researcher's needs - #2

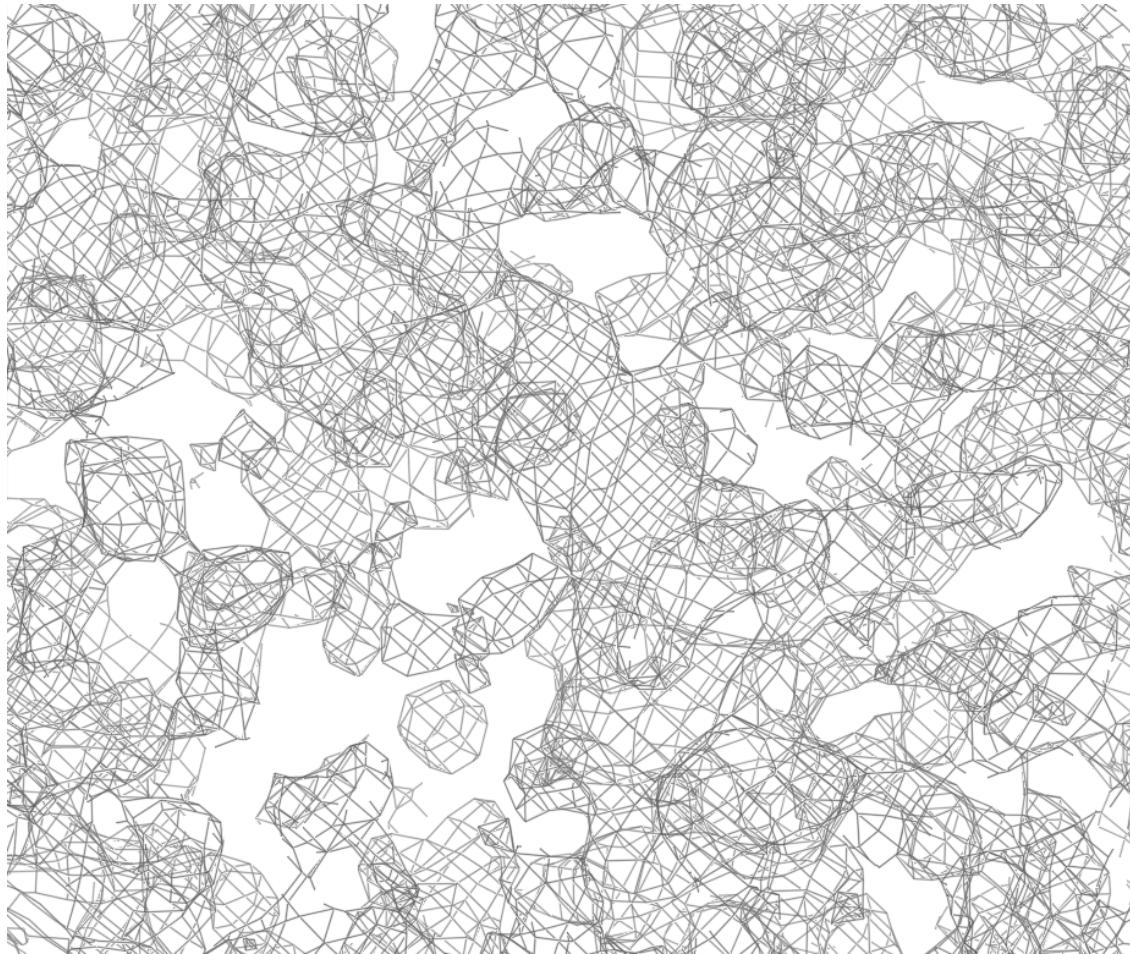
CCP4BB: May 29th 2007

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!

*Best,
Joe*

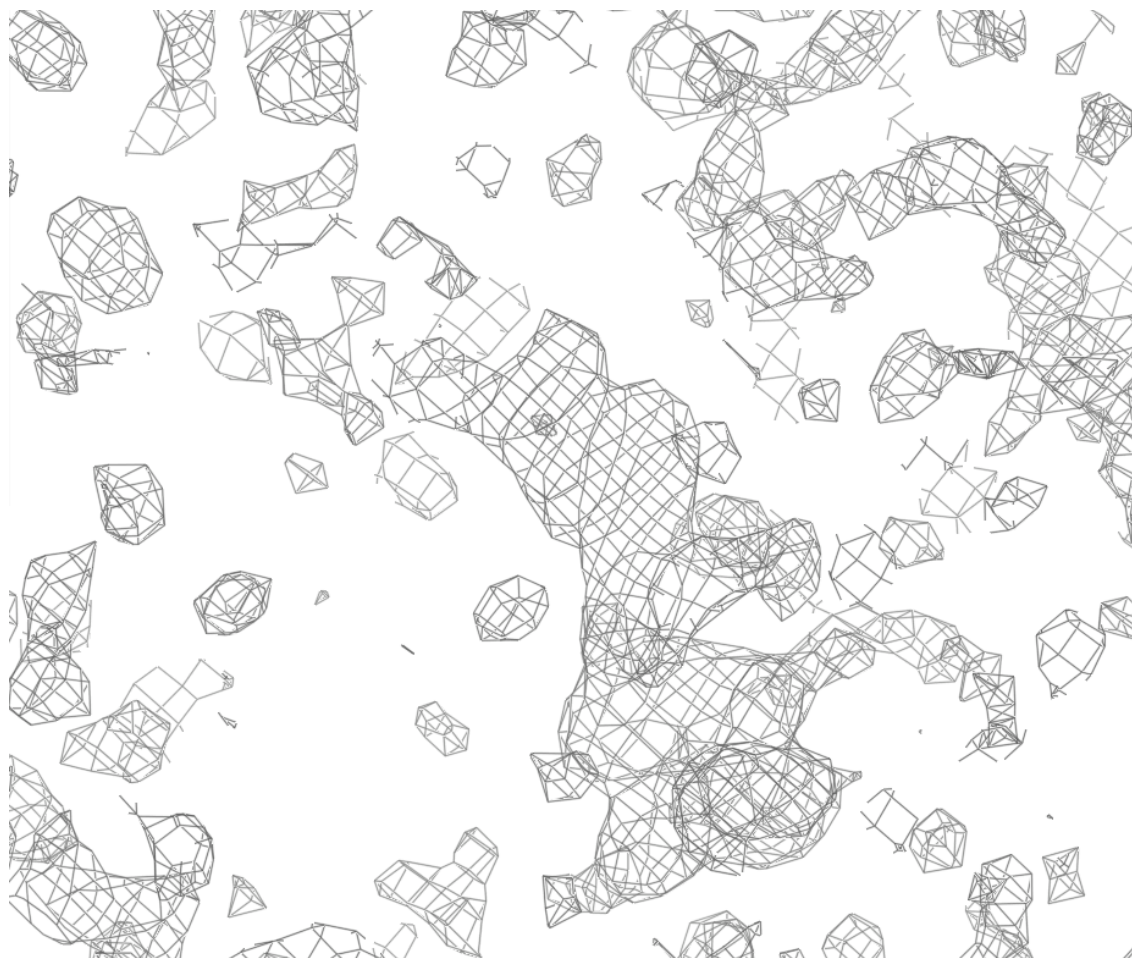
Automatic binding site identification

The difference density map at a low contour threshold.



Automatic binding site identification

The difference density map at a medium contour threshold.



Automatic binding site identification

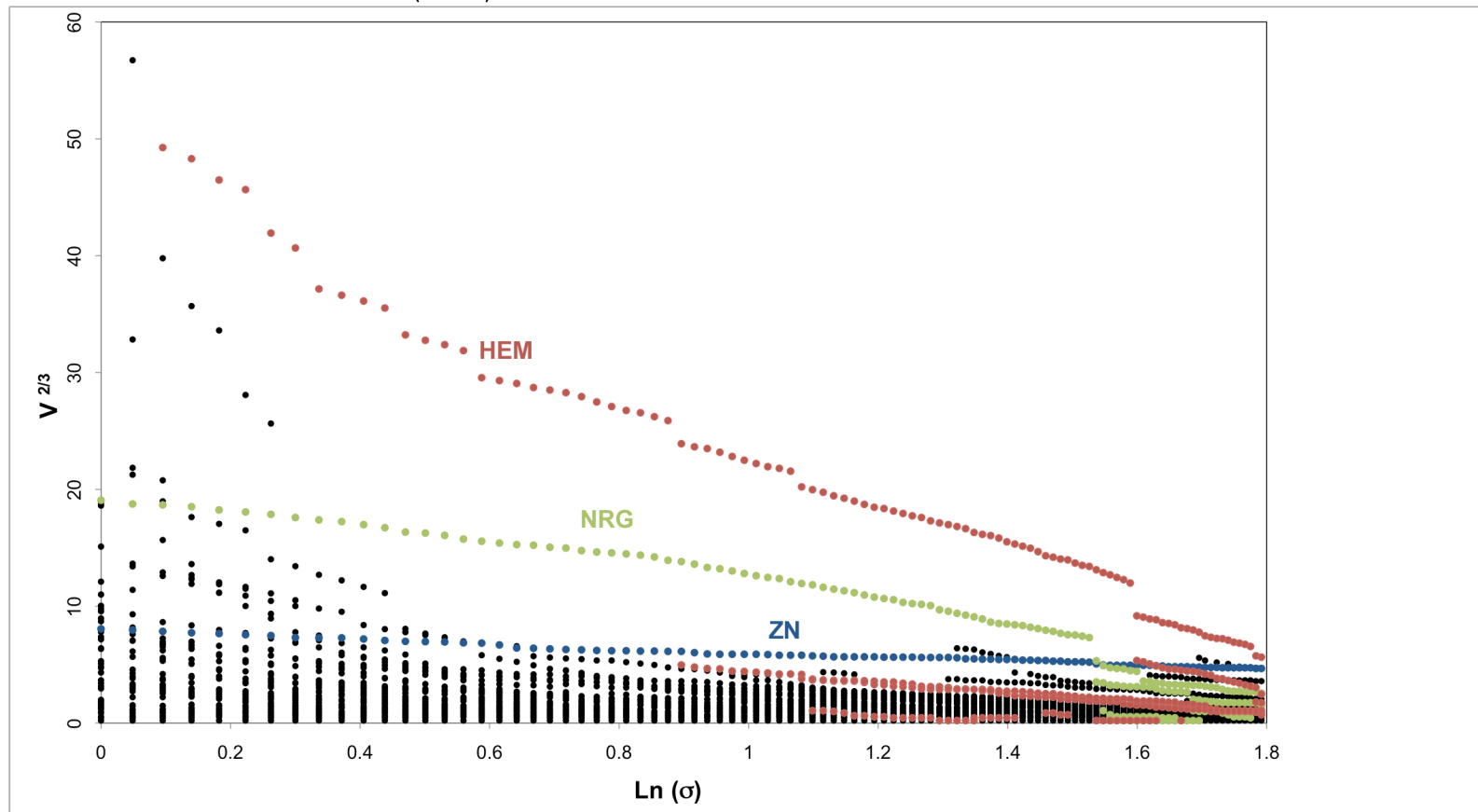
The difference density map at a high contour threshold.



Automatic binding site identification

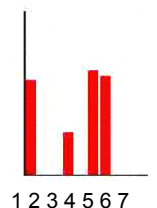
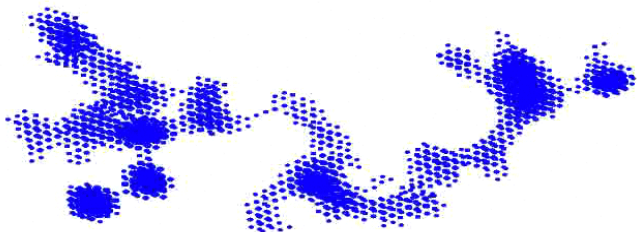
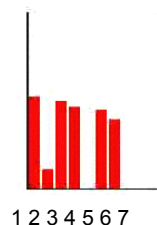
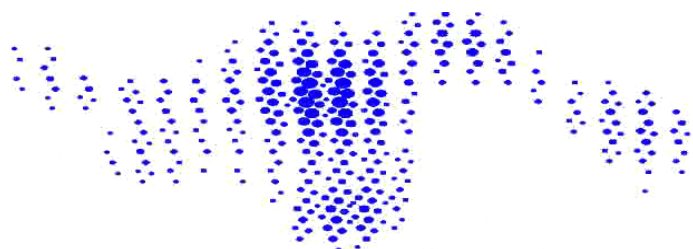
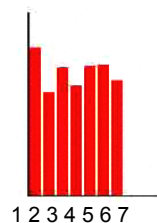
Capturing the dependence of the difference density map on changing the contour thresholds: *Fragmentation tree*.

$$V^{2/3} = \left(\frac{4\pi}{3}\right)^{2/3} \sigma^2 \left[\ln\left((2\pi)^{-3} \sigma^{-6} Z^2\right) - 2\ln(t) \right]$$



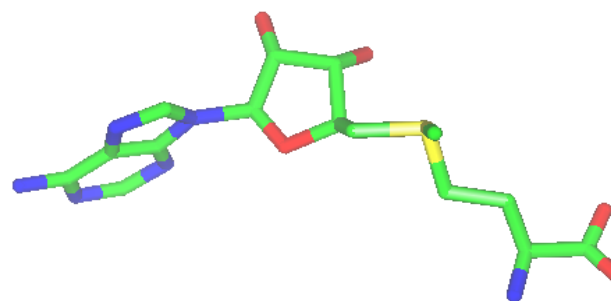
Shape features to measure similarity

To assign a likelihood to a site knowing the ligand.



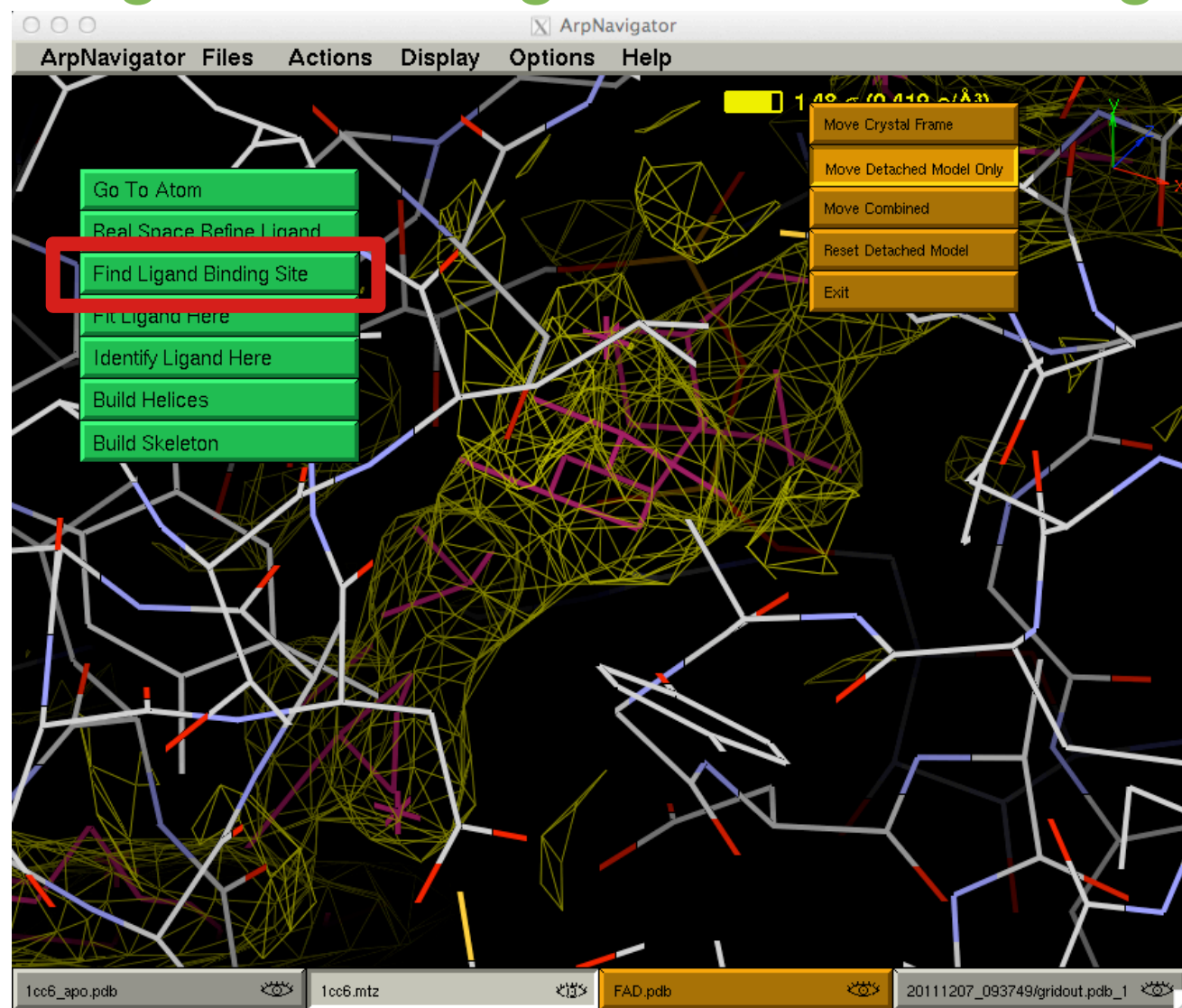
- 1) Surface to volume ratio
- 2) Bounding box limits
- 3) Moments of inertia
- 4) Rotation match score
- 5) Eigenvalues
- 6) Distance histogram
- 7) Geodesic distance histogram

Find the matching cluster for this ligand!



Decision is based on feature vectors.

Locating the binding site in ARPNavigator



Researcher's needs - #3

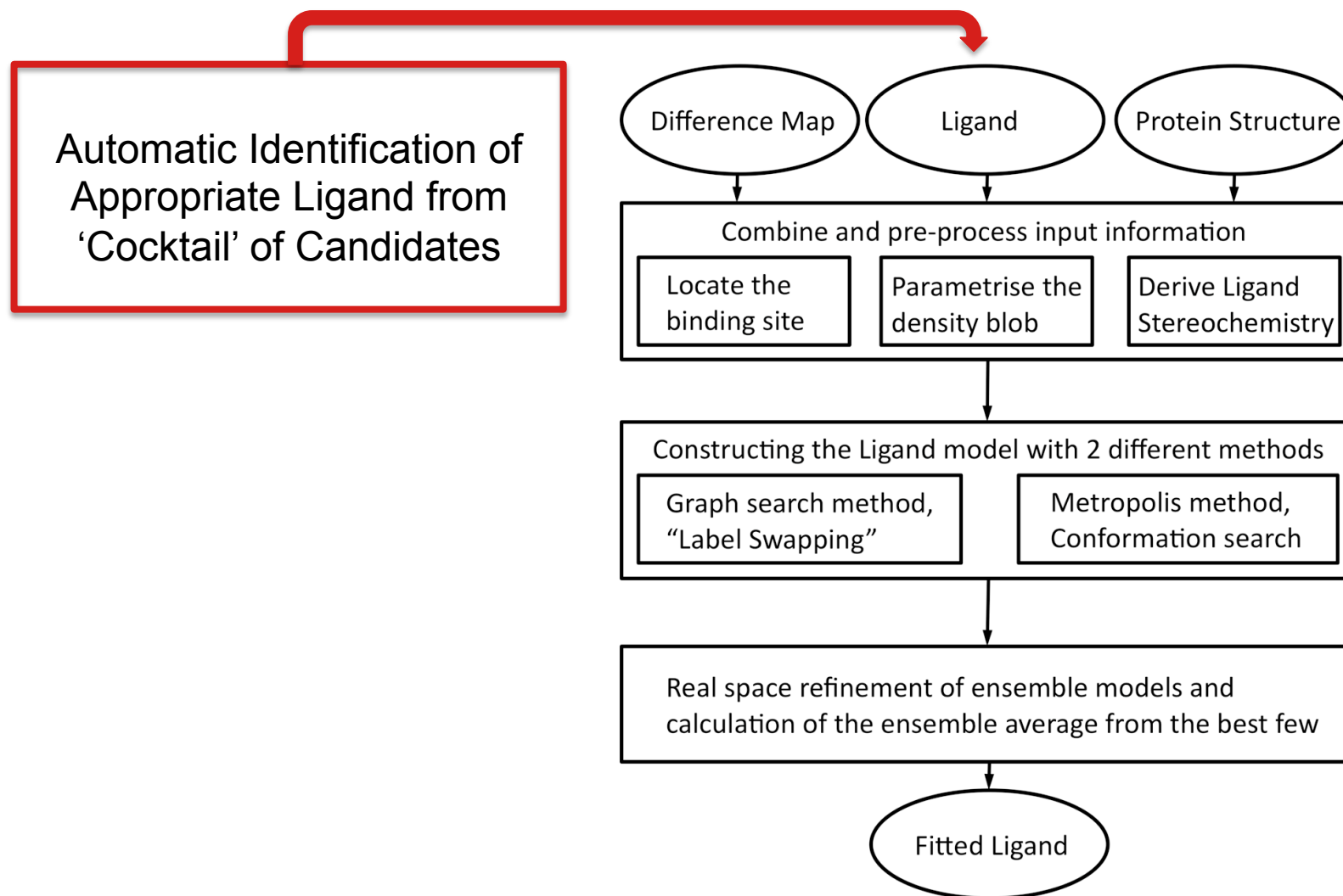
CCP4BB: March 14th 2011

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?

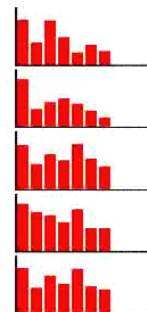
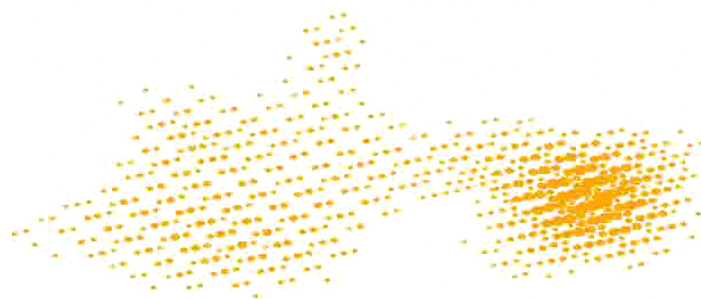
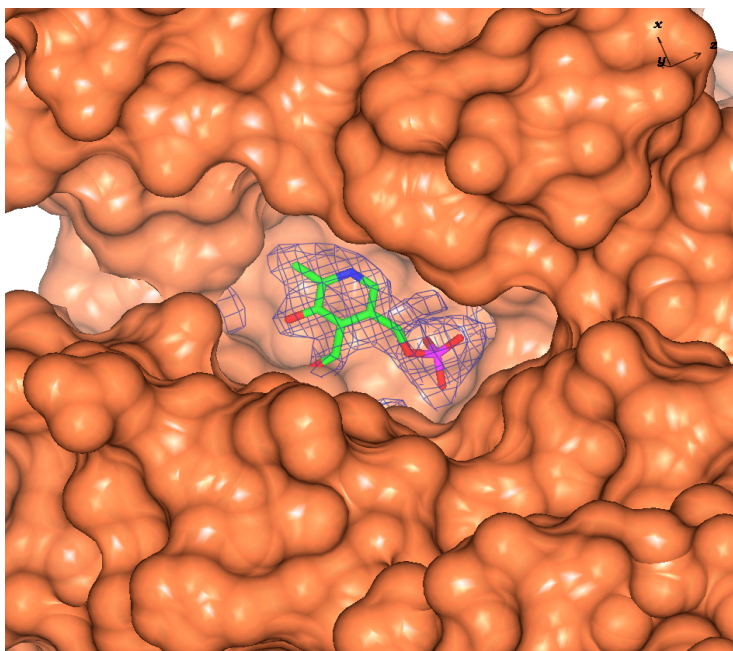
*Many many thanks!!!
Xiaopeng*

Automatic Ligand Identification?

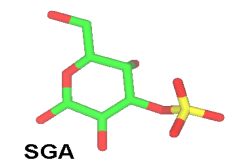
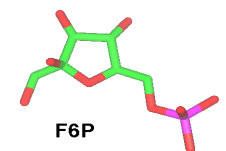
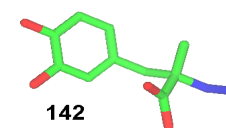
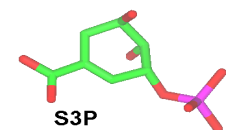
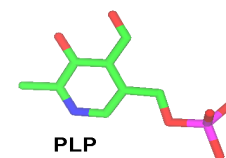


Shape features to measure similarity

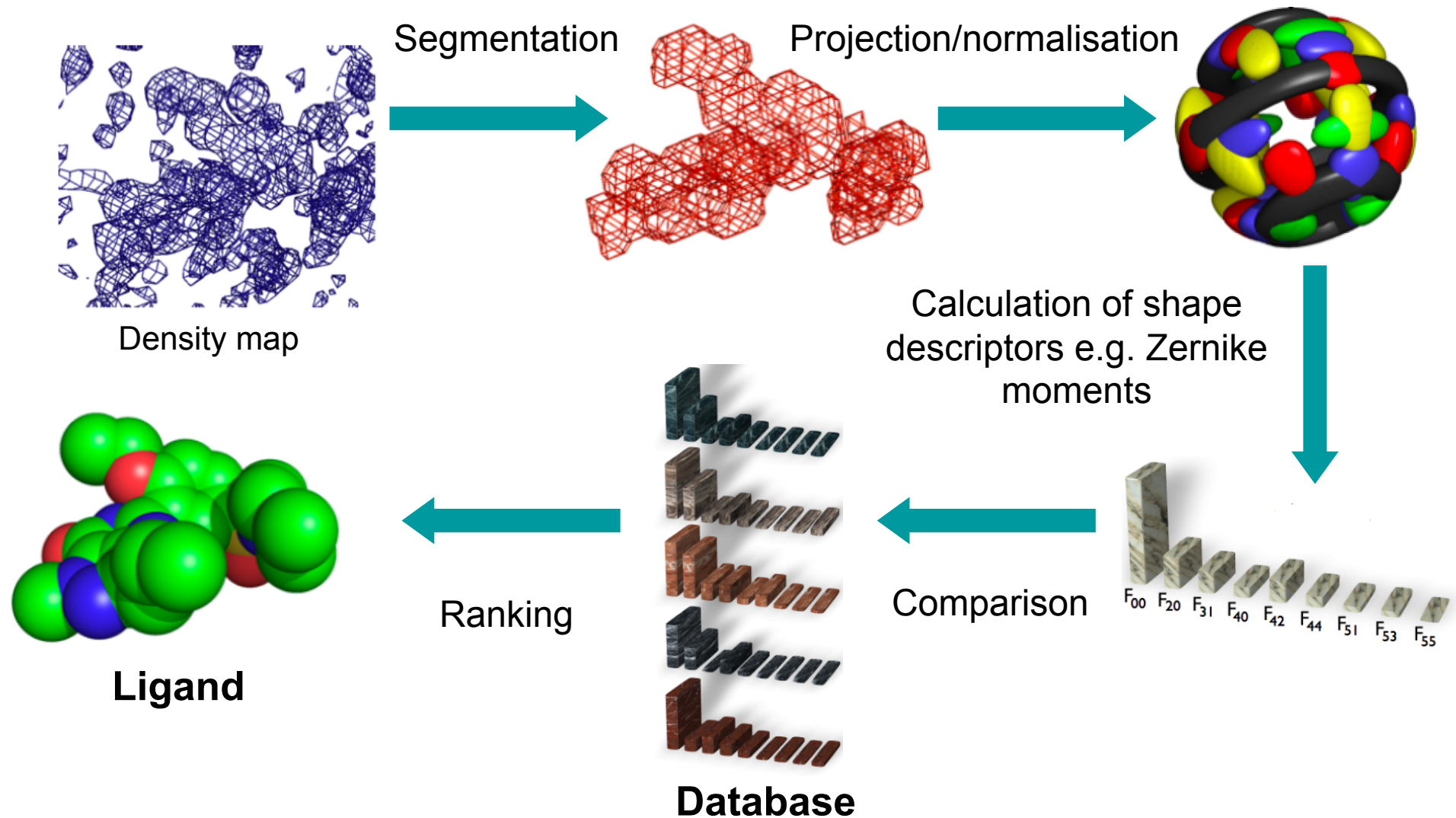
To assign a likelihood to a ligand knowing the binding site.



Final ligand ranking

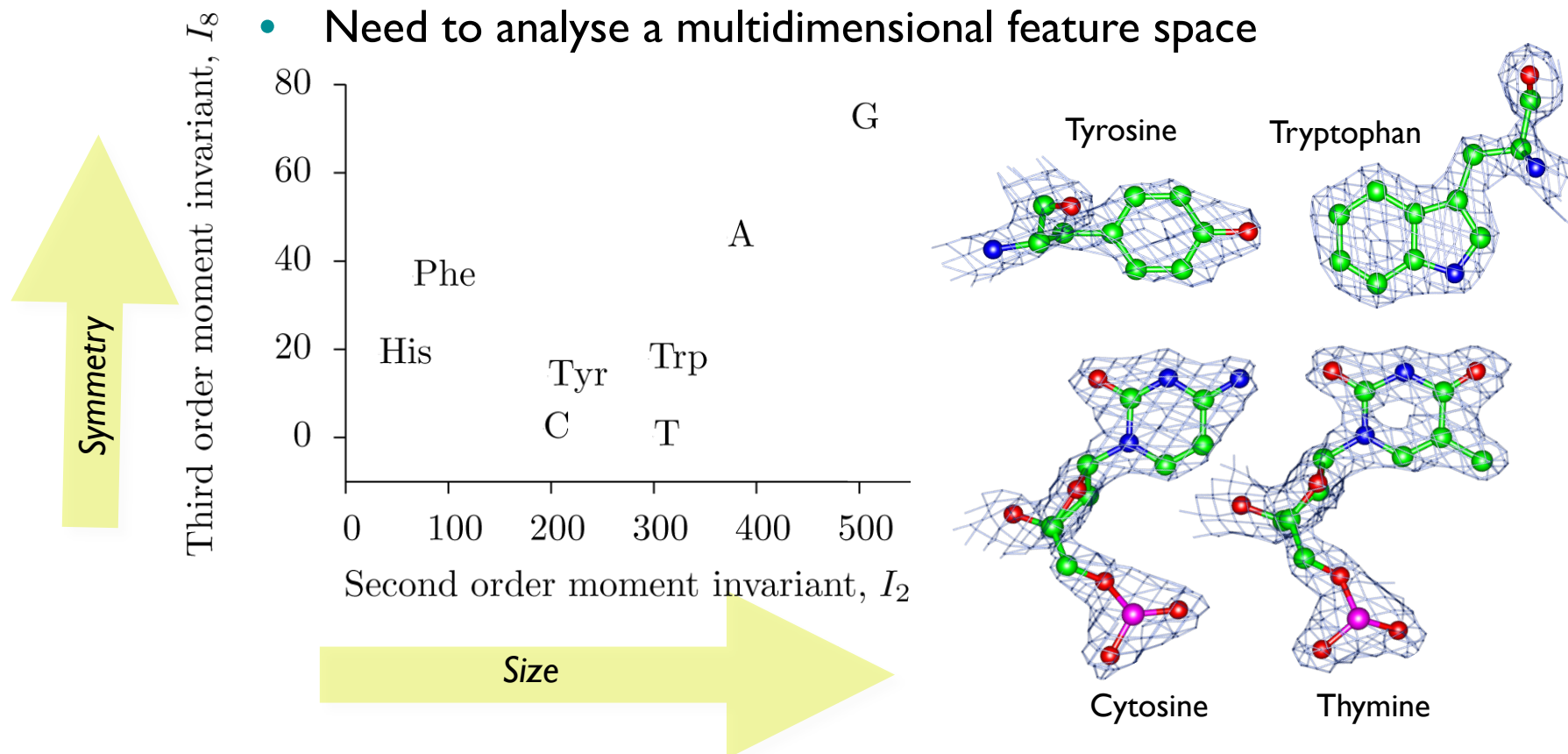


A new feature - High throughput ligand identification

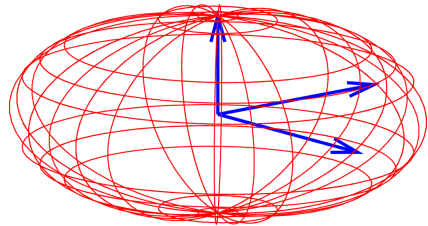


Use a Combination of Various Features

- Any one feature is not discriminative enough
- Need to analyse a multidimensional feature space

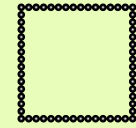
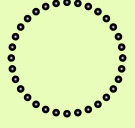
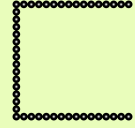
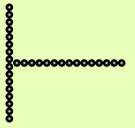
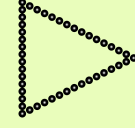


Third Order Moment Invariants

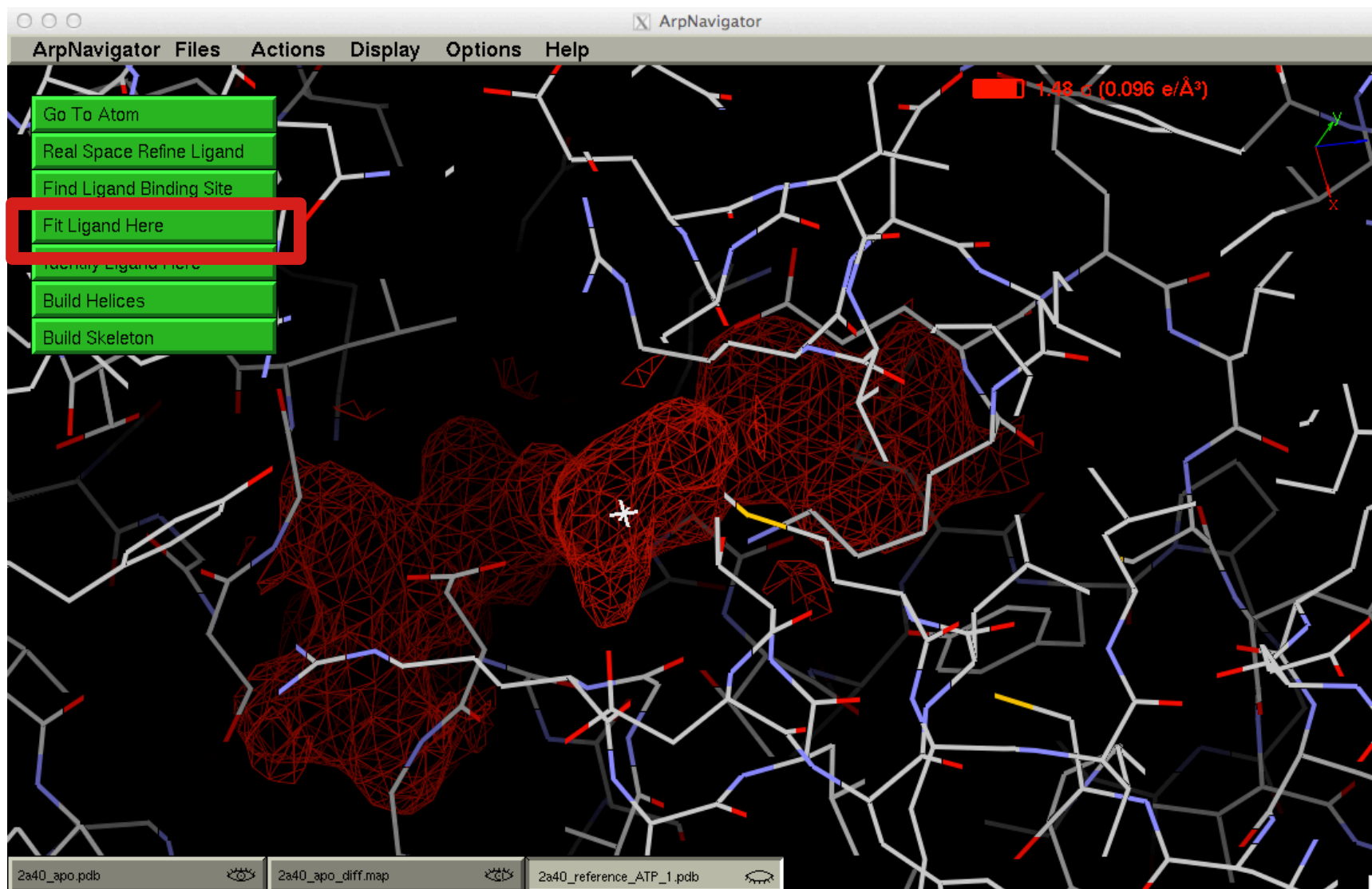


PCA – 2nd order, symmetric shapes

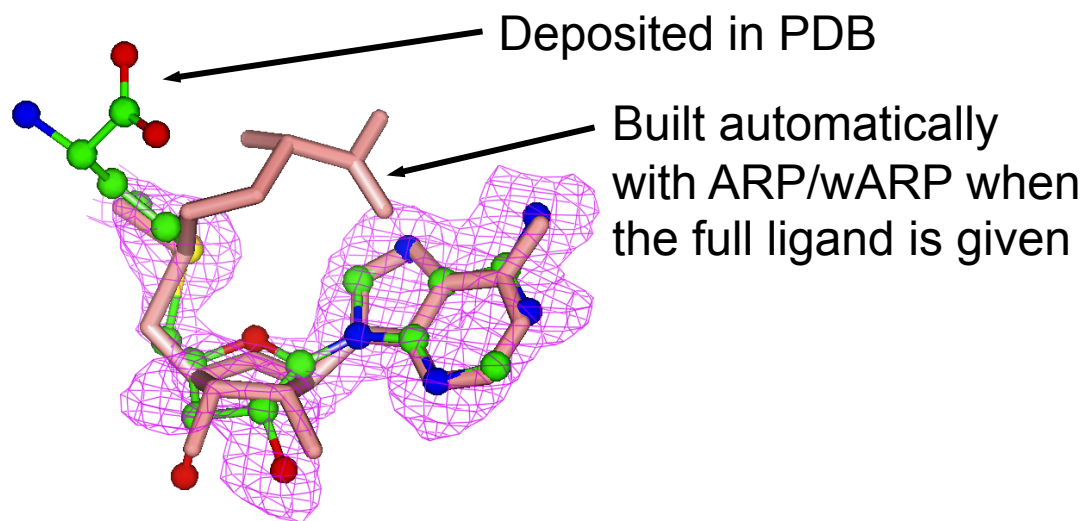
3rd order: Lo & Don (1989) *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 1053–1064

								
μ_{00}	16	32	32	60	32	46	31	45
μ_{20}	170	200	340	2255	900	1228	775	1130
μ_{30}	0	0	0	0	0	3049	3596	2850

Ligand Identification in ARPNavigator

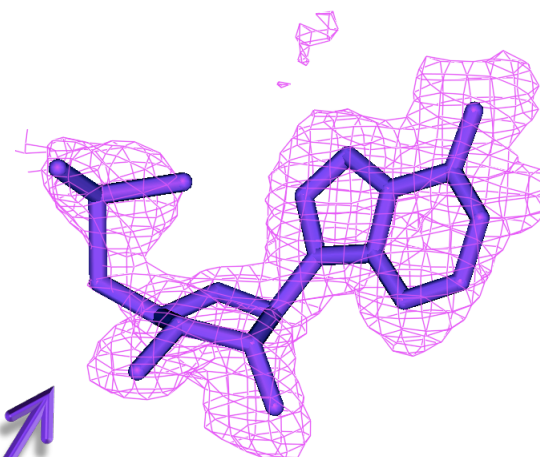
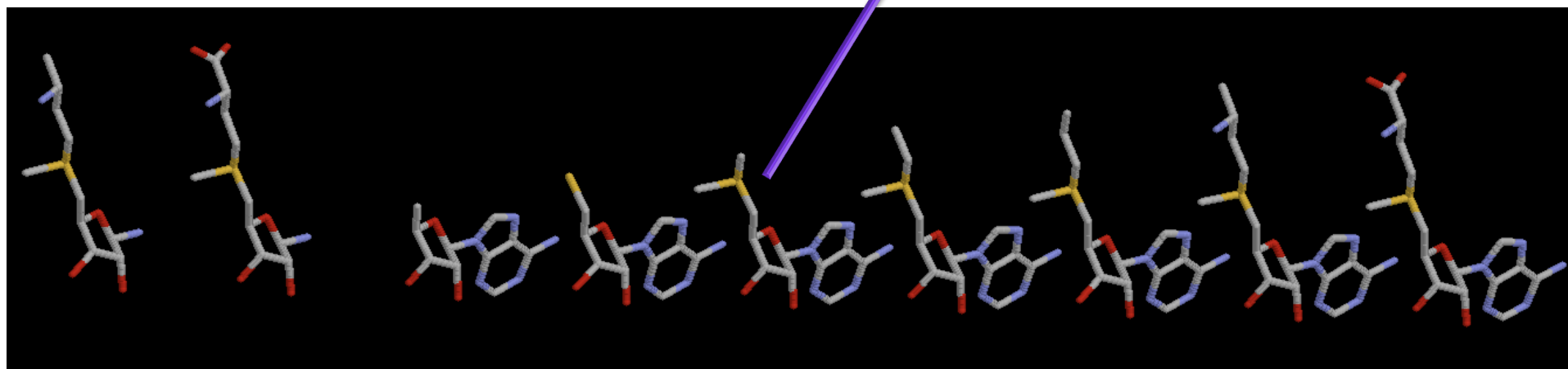


Partial disorder, partially occupied ligands



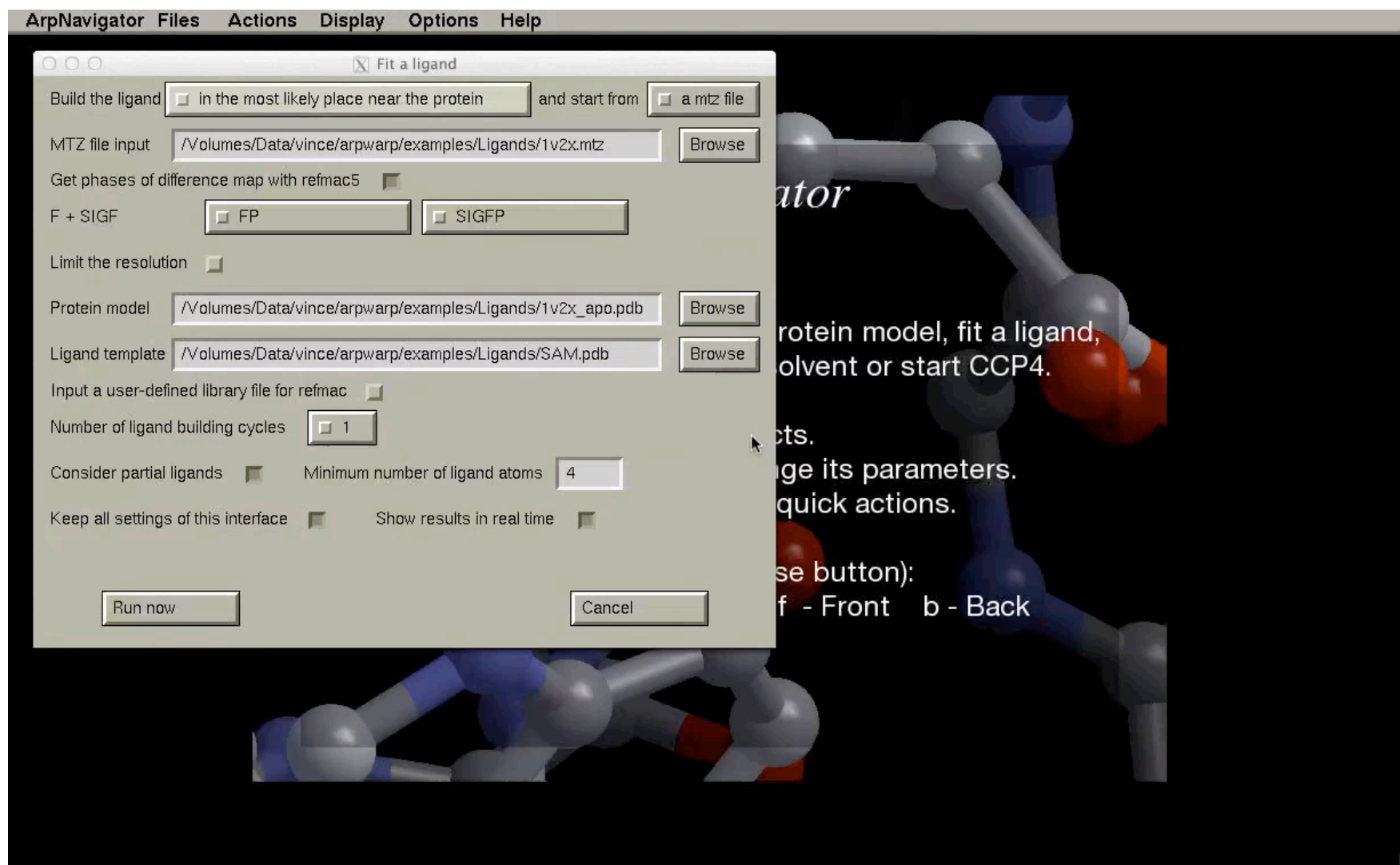
SAM in 1v2x at 1.5Å

Artificial cocktail:



Compound chosen in cocktail case

Partial ligand building in ARPNavigator

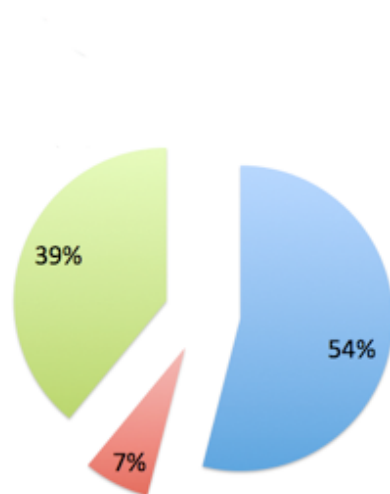
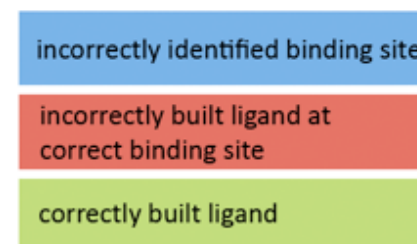


Success rates

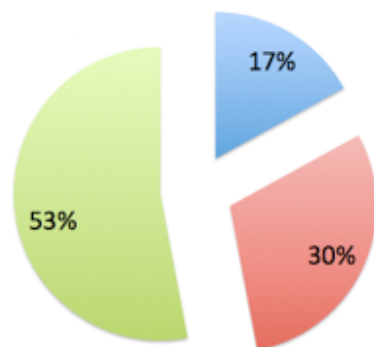
Tested on > 20k PDB entries from EDS

X-ray resolution limits: 1.0 to 3.0Å,
Building the largest fully occupied ligand.
Correctness criterion: rmsd < 1.0Å

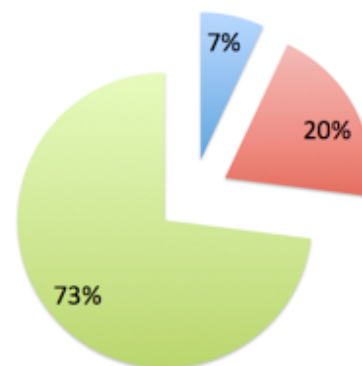
(Current Version 7.2)



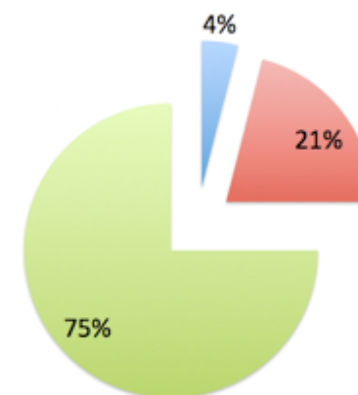
5...6 atoms,
any map corr.



7...100 atoms,
any map corr.

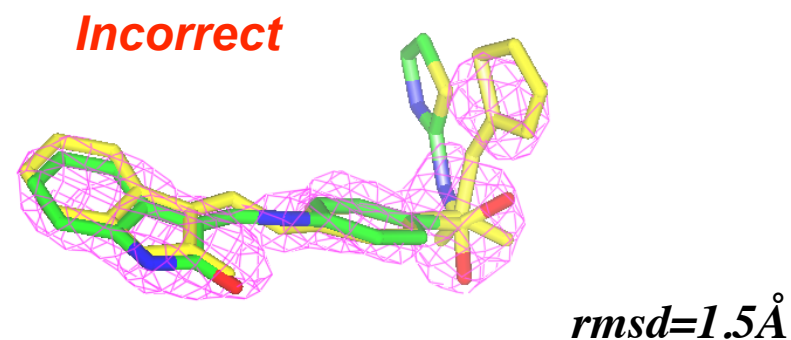
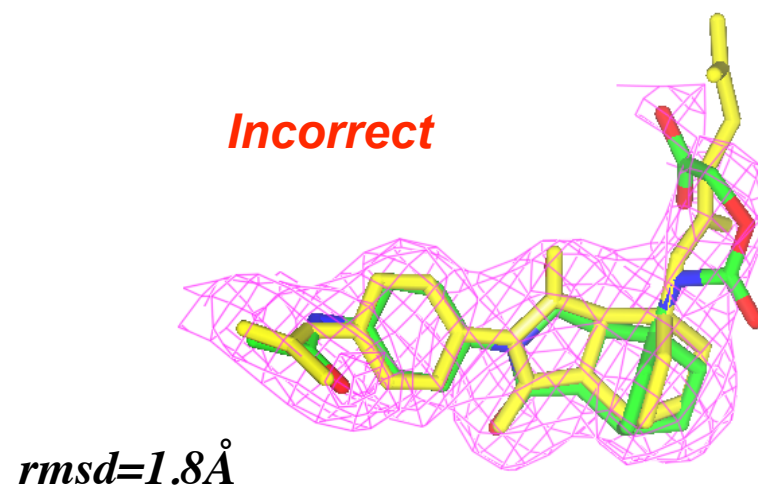
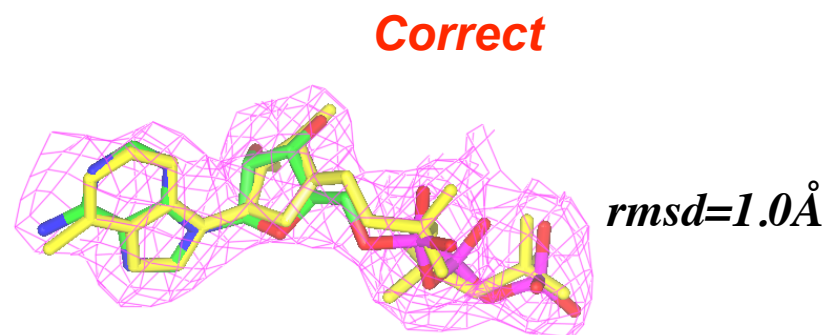
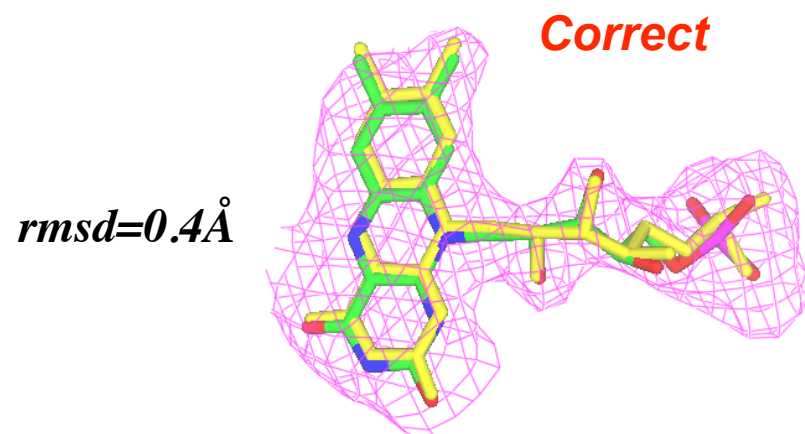


10...40 atoms,
map corr. > 80%



20...40 atoms,
map corr. > 80%

Success criterion: r.m.s.d. < 1.0Å



In yellow: the deposited ligand

Running a task from the command line

Series of jobs can be run this way easily.

```
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}]

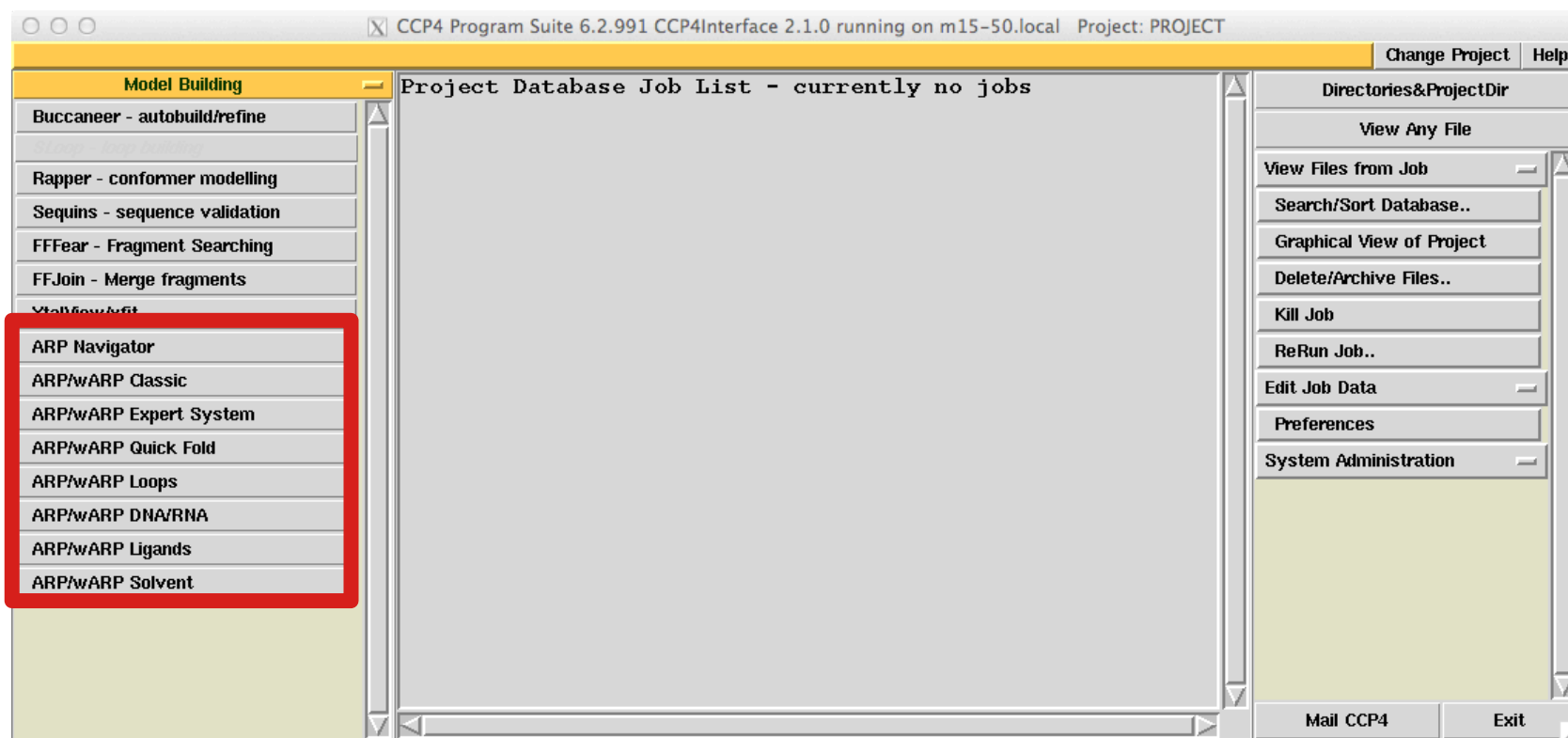
- Optional command line arguments are given in square parentheses
- All input files are assumed to be located in working directory
  unless they are given with full path
- If workdir is not given, the current directory will be assumed
- All output files will be written into workdir/subdirectory
- If no ligand is specified then auto identification of
  the ligand will be attempted provided that a search position
  is given (experimental)

M15-50:~ kabushikigaishatamaya$
```

to launch the
job with the
default scenario
only 3 files are
needed.

Model Completion routines in ARP/wARP

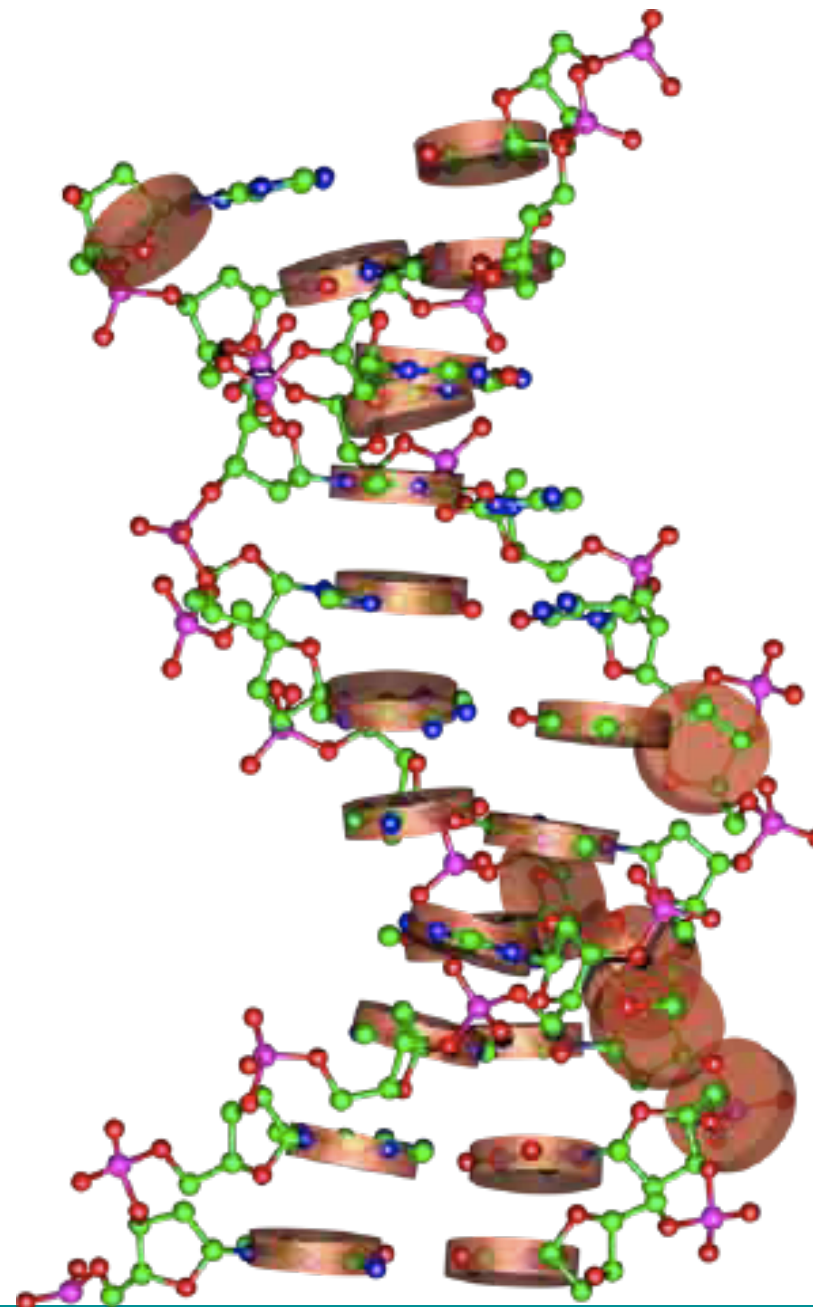
- As well as ligands, there are modules for the modelling of DNA, RNA, solvent and unmodelled loops to aid model completion efforts.



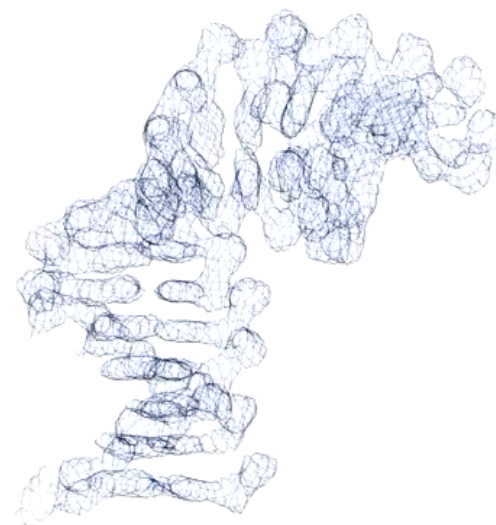
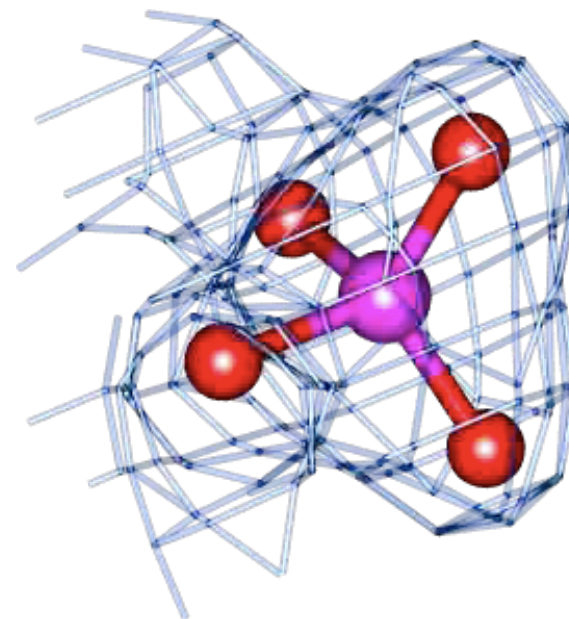
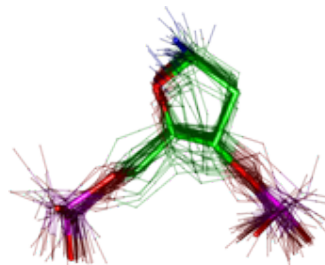
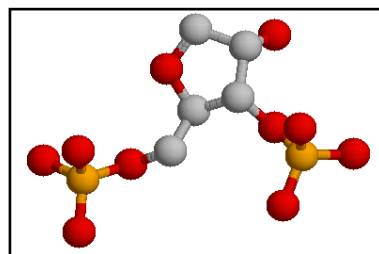
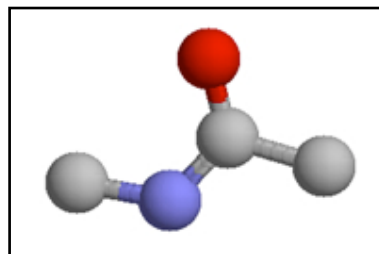
DNA/RNA building

- Detection of phosphate backbone and nucleobase planes using pattern recognition algorithms

1p7h (2.5 Å)	
Completeness:	0.88
Noise level:	0.69
Positional r.m.s.d.:	0.5 Å
Angular difference:	5°

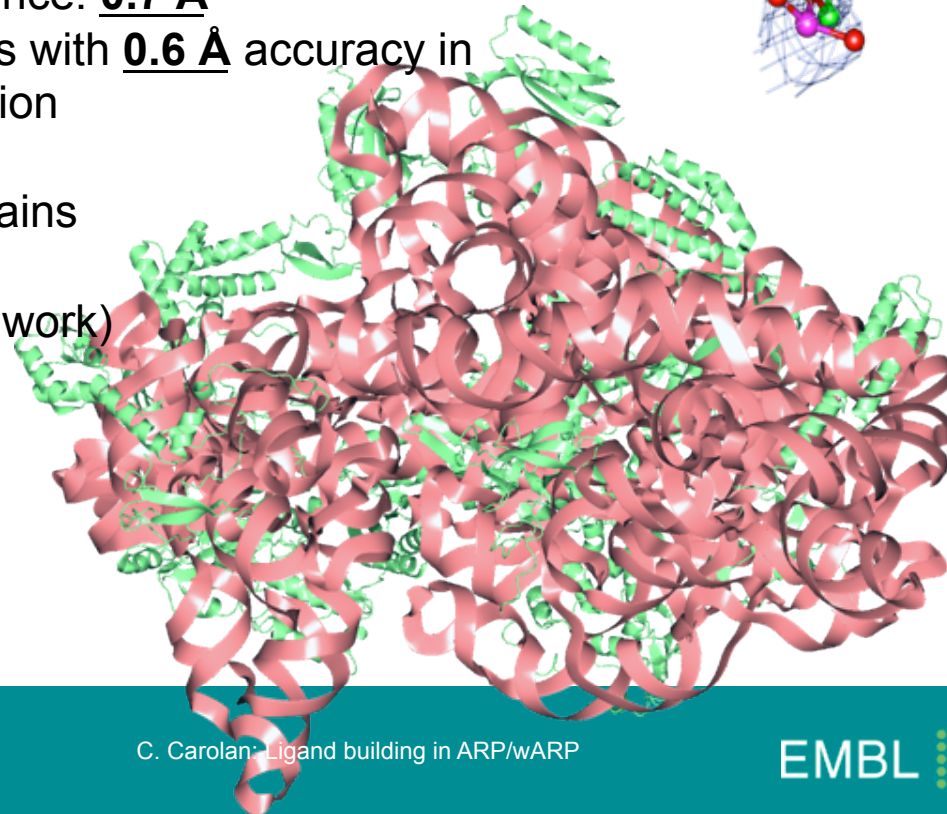
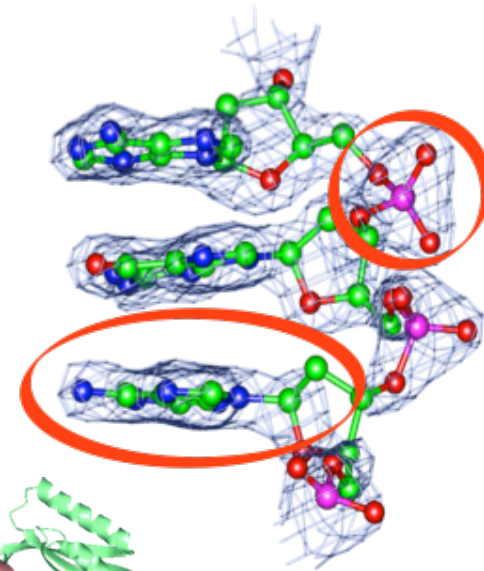


DNA/RNA building



DNA/RNA building

- The **30S ribosomal subunit**
 - Resolution: 3.05 Å
 - Experimental phases (MCC 0.73)
- Automatic model building with ARP/wARP
 - modelled 1,302 out of 1,513 nucleotides (**86%**)
 - backbone r.m.s.d. to reference: **0.7 Å**
 - Located 1,121 nucleobases with **0.6 Å** accuracy in location and 12° in orientation
- Backbone fragmented in 75 chains
- Built in **around 6h**
- (*cf.* several **months** of manual work)



Acknowledgements

Developers

EMBL Hamburg: Ciaran Carolan, Philipp Heuser, Victor Lamzin, Tim Wiegels, Saul Hazledine

NKI Amsterdam: Krista Joosten, Robbie Joosten, Tassos Perrakis

Collaborators

EMBL Hamburg: Gleb Bourenkov, Santosh Panjikar

York/Cambridge: Garib Murshudov's group

Daresbury Laboratory: CCP4 team

Former members

Gerrit Langer, Richard Morris, Peter Zwart, Francisco

Fernandez, Matheos Kakaris, Olga Kirillova, Wijnand Mooij,

Diederick De Vries, Marouane Ben Jelloul, Johan Hattne, Tilo

Strutz, Guillaume Evrard, Serge Cohen, Venkat Parthasarathy,

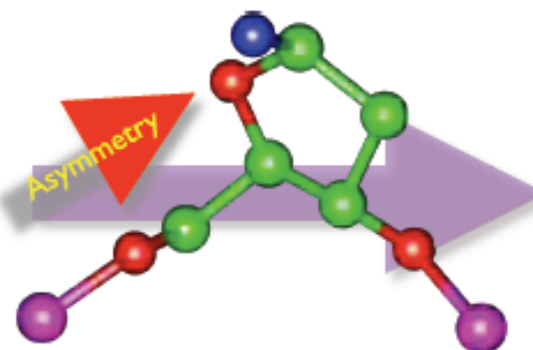
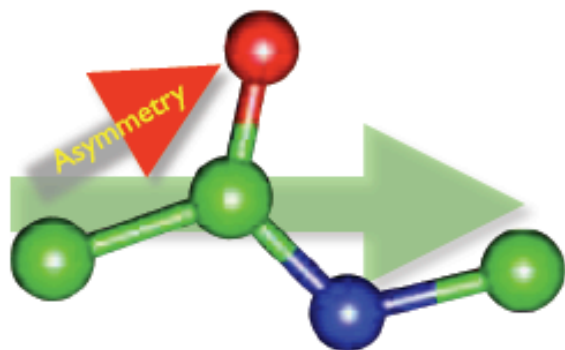
Babu Pothineni, Helene Doerksen

Funding



HMAIN and PEPT

Protein chain tracing in ARP/wARP 7

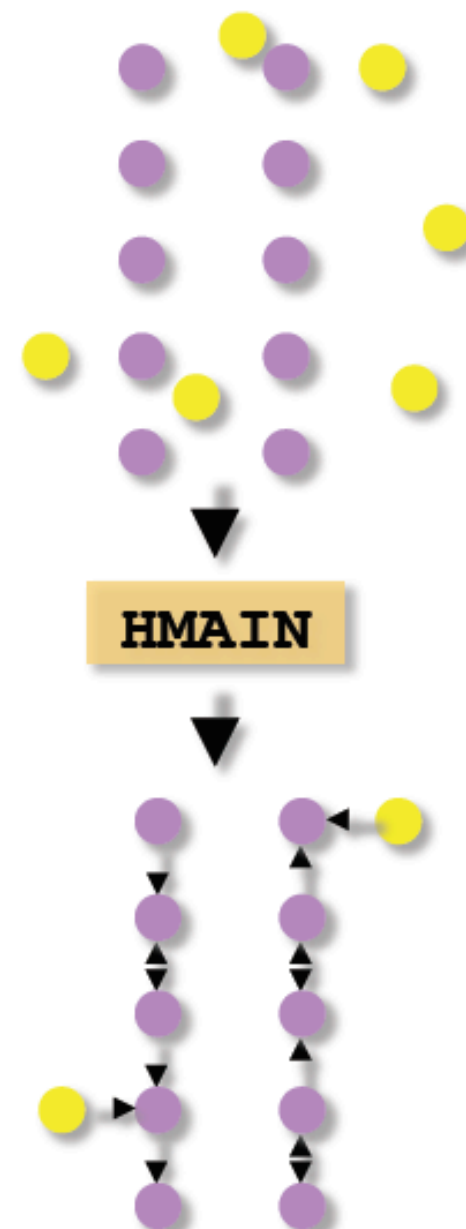


- **HMAIN**

- Locates putative peptide units from C α candidates
- Assigns a score based on peptide density shape
- Resolves topology graph of polypeptide chains

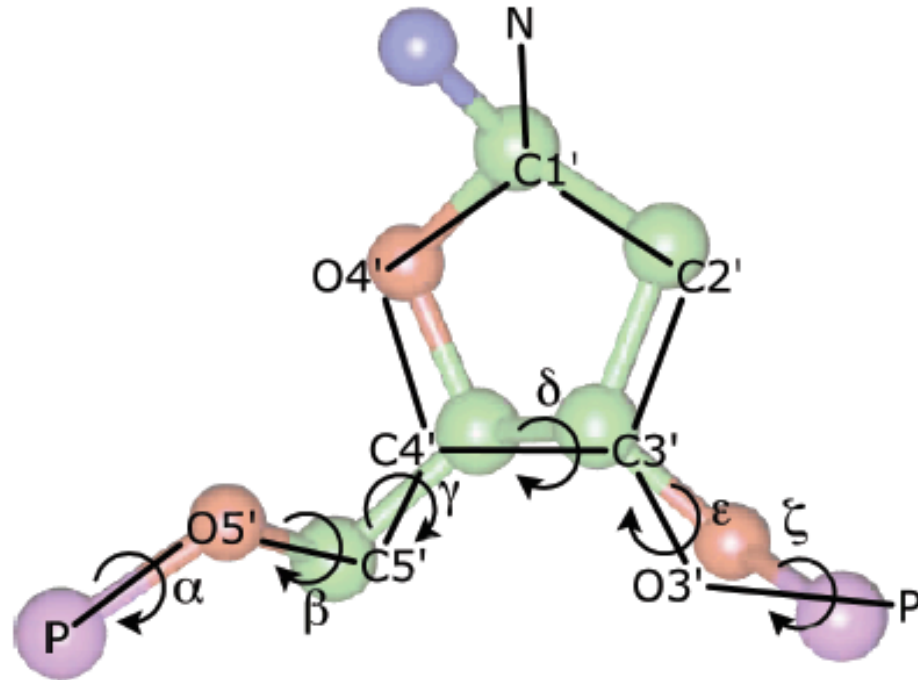
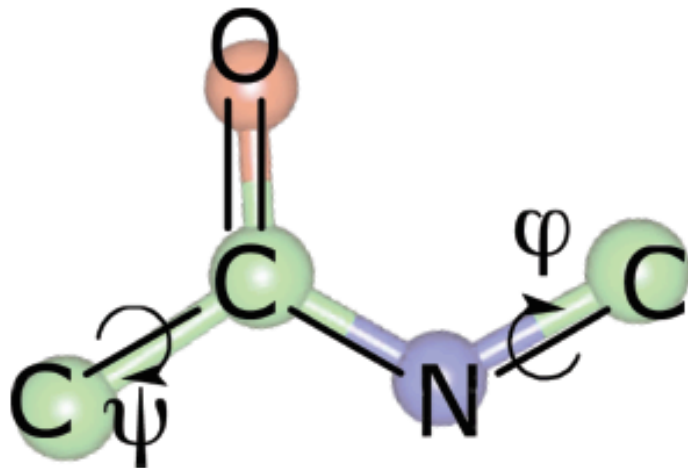
- **PEPT**

- Selects best subset of chains
- Builds main chain and partial side chains
- Fits polypeptides into density map



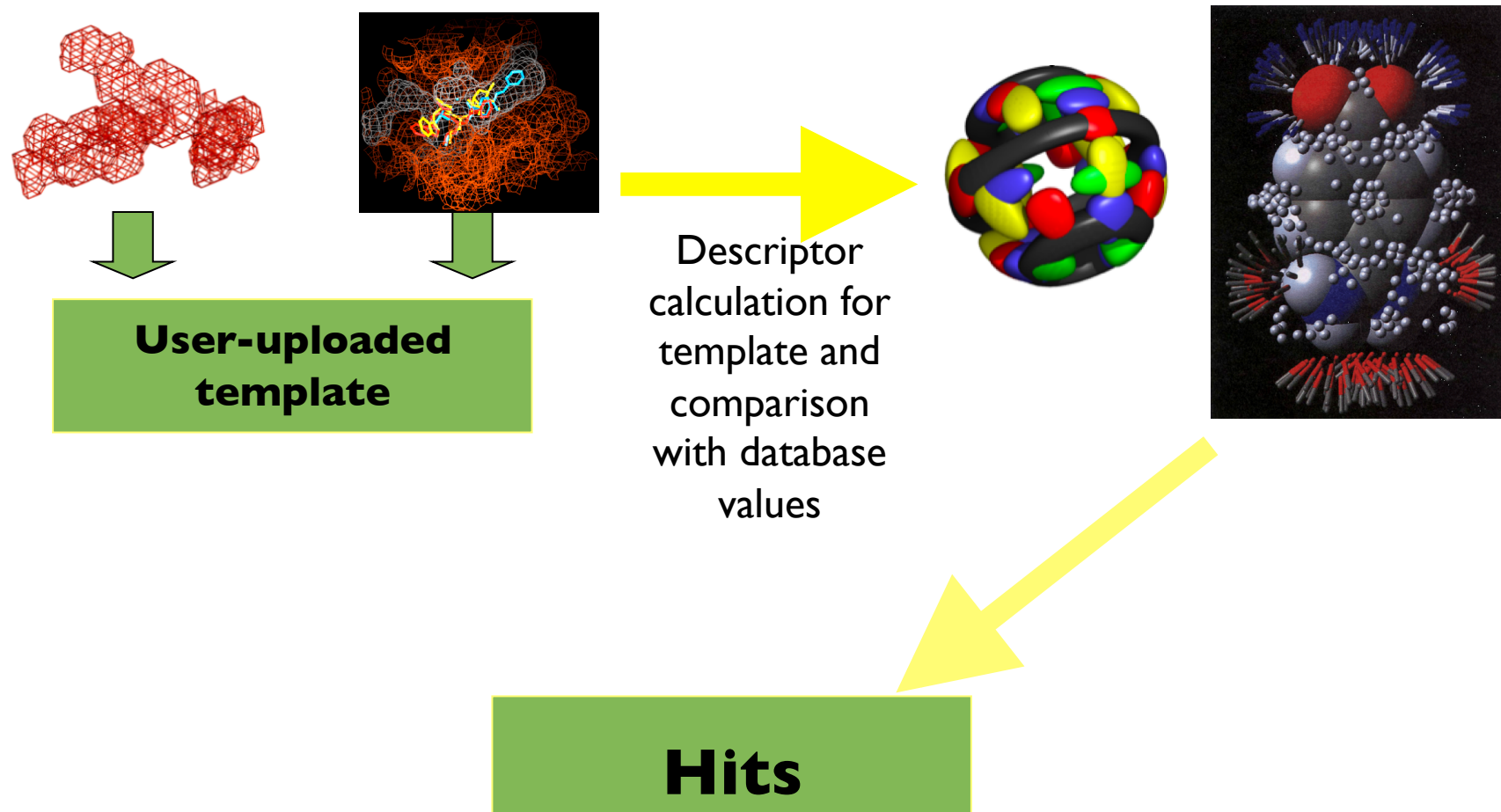
Backbone complexity

Dihedral angles in protein and nucleic acid backbones



- Protein: two dihedral angles
- Nucleic acids: six dihedral angles

The ATOLL Database

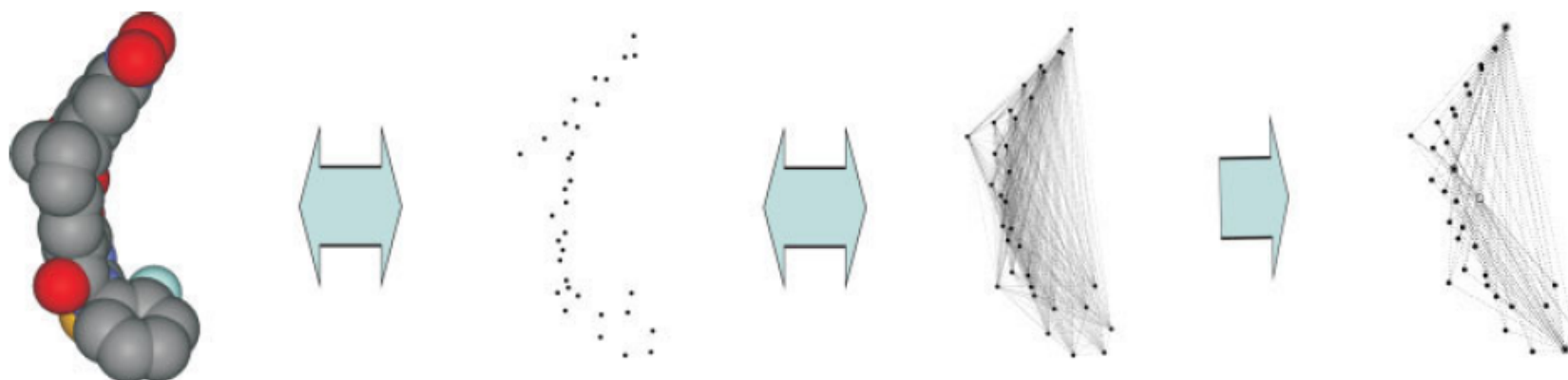


Who???



Tutorial this afternoon.....

Matrix-Based Features



Labeled graph	Laplacian matrix
	$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$

Ligand building with arpnavigator

