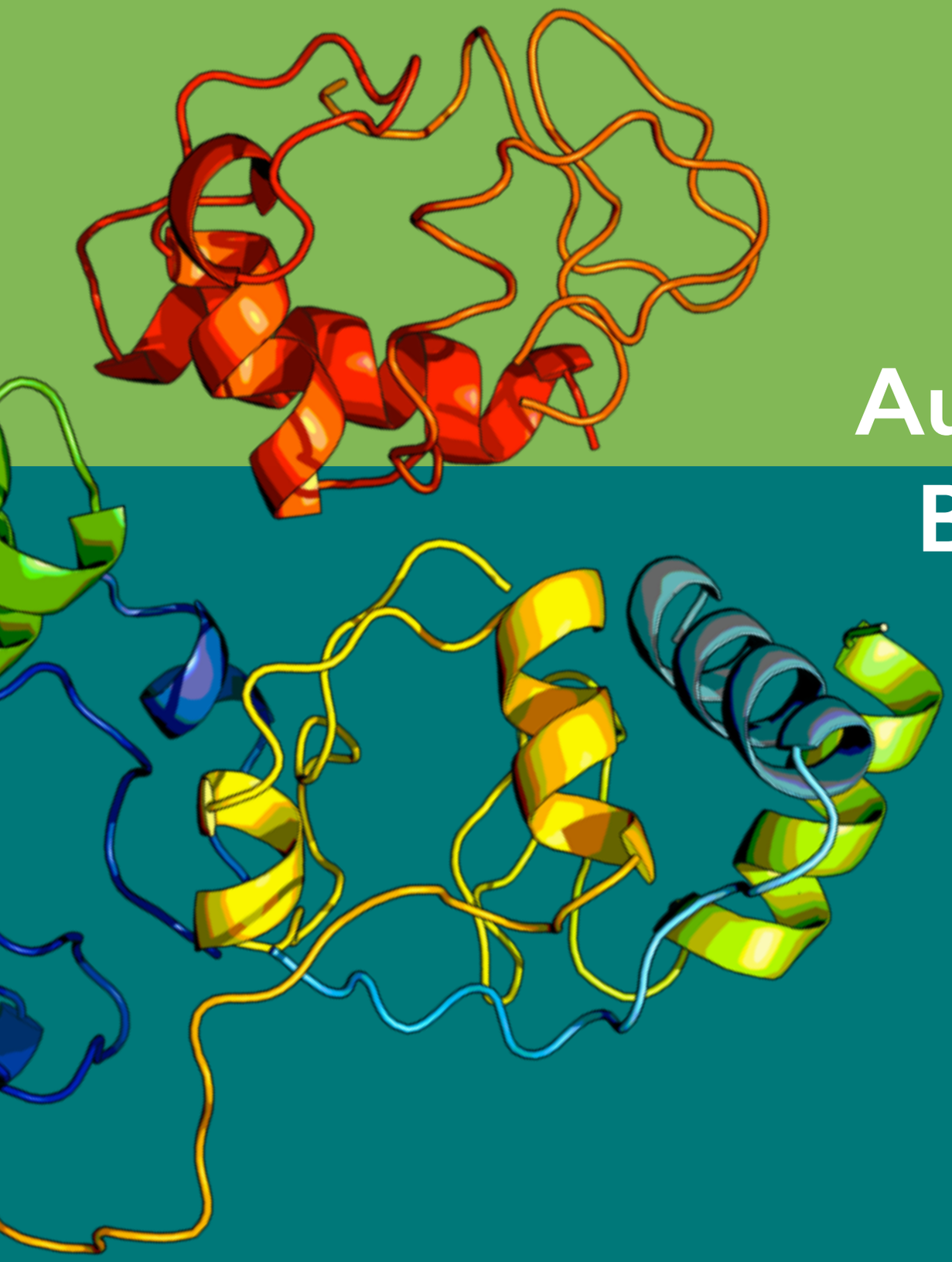
Joana Pereira
Lamzin Group
EMBL Hamburg, Germany



Automated Protein Model Building with ARP/wARP

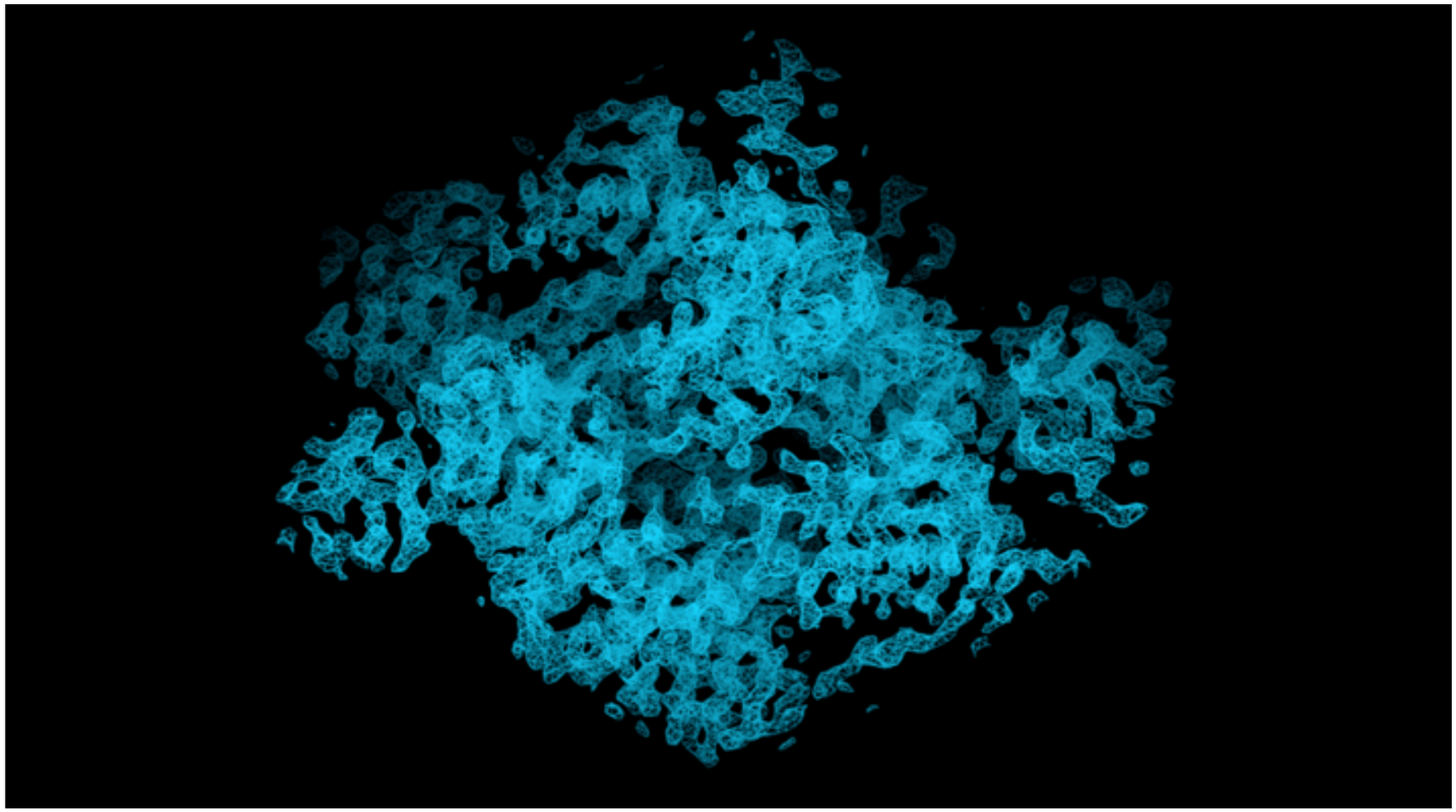
(and nucleic acids too)

EMBL 40 YEARS | 1974–2014

ARP wARP 7.4

Electron density map:
the ultimate result of a crystallography experiment

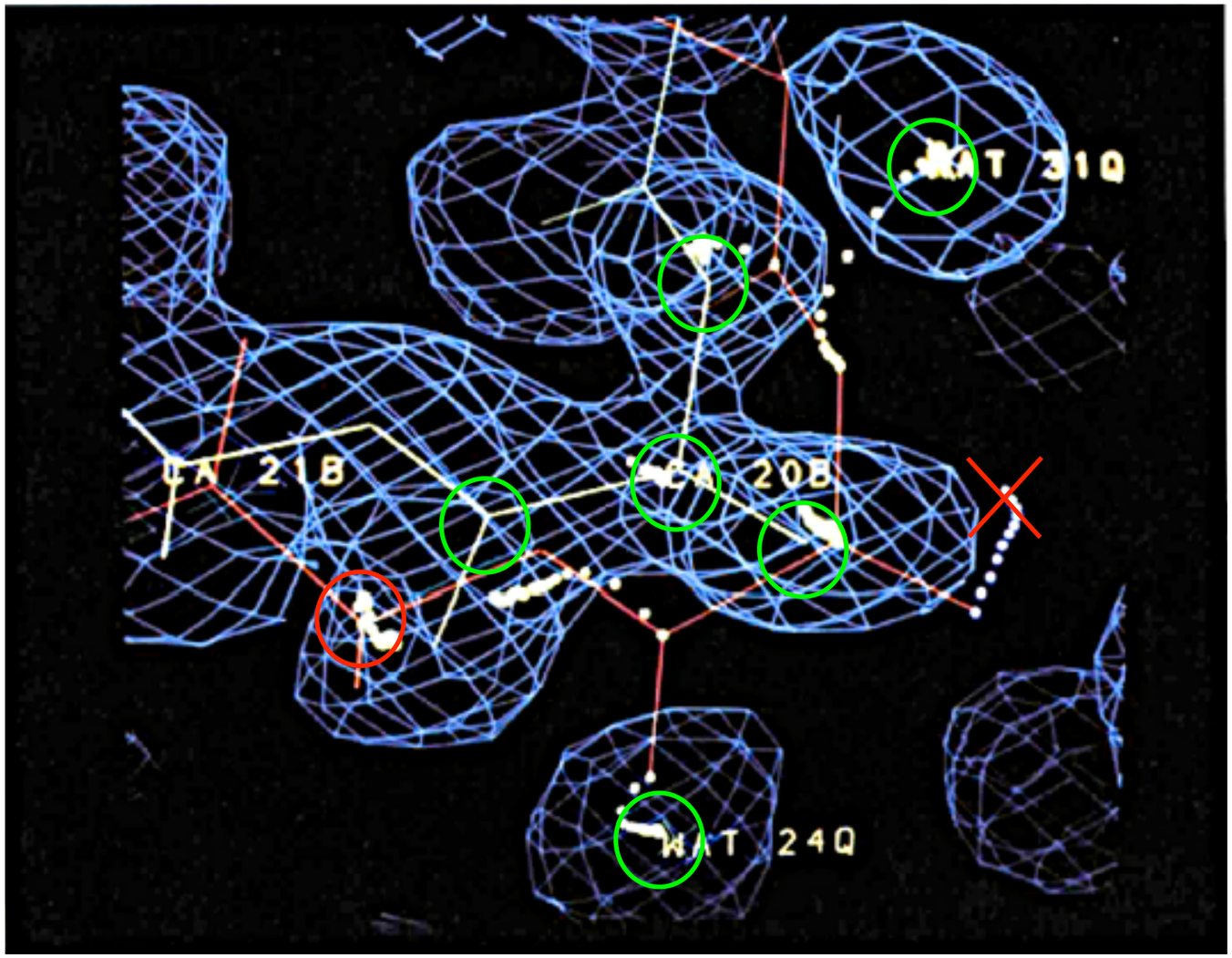
How would you do to interpret this map?



Building a molecular model from the electron density map calls for automation, it can take a long time when done by hand!

ARP/wARP

Developed in Hamburg by the Lamzin group, it is:

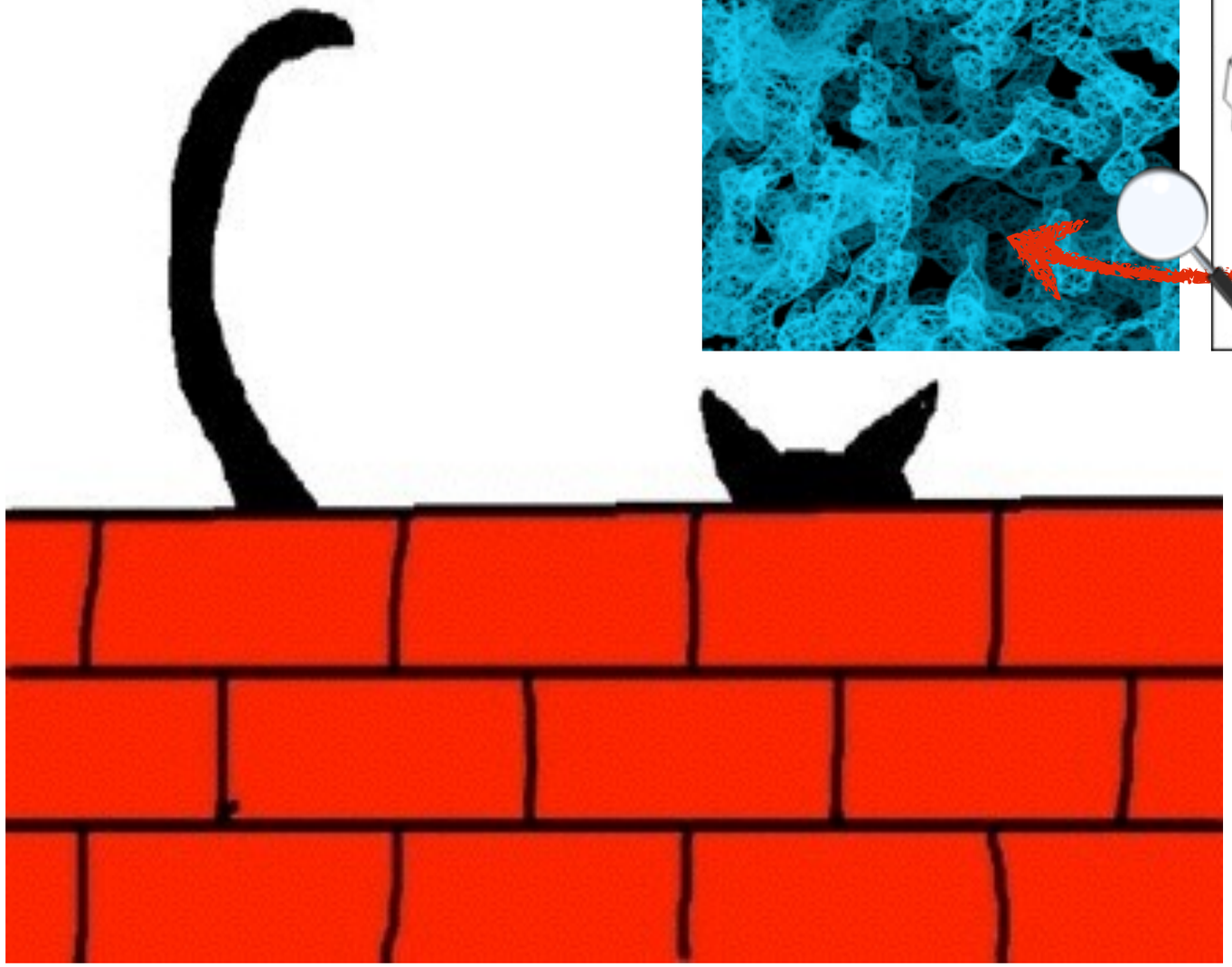


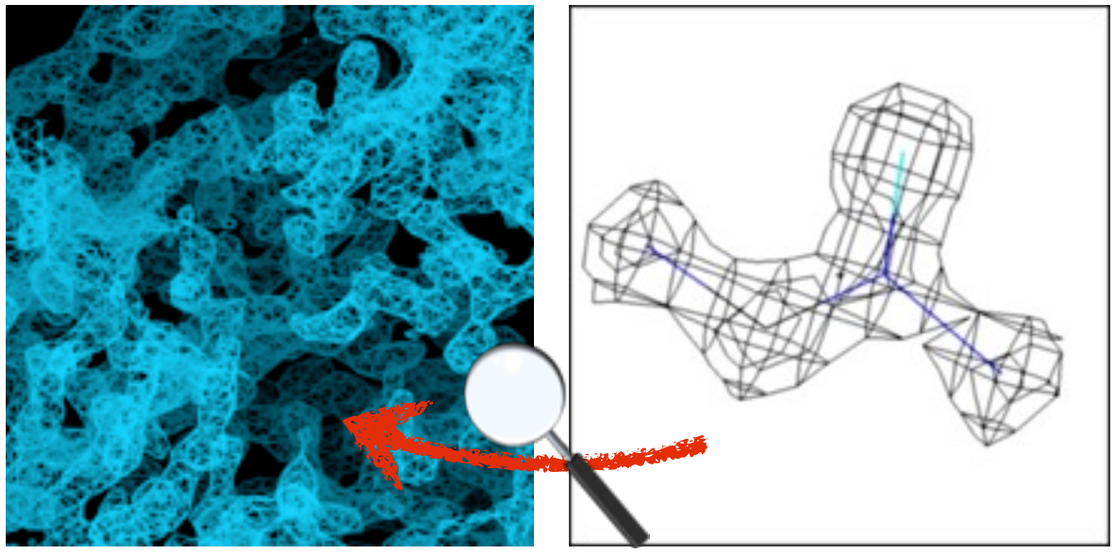
- one of the leading **MX** software projects
- used to build structures of **proteins**, **nucleotides**, **ligands** and their **complexes**
- making use of pattern recognition to build **models** from MX **electron density maps**

Lamzin, V.S. & Wilson, K.S. (1993) Automated refinement of protein models. *Acta Cryst.* **D49**, 129-147.

Pattern Recognition

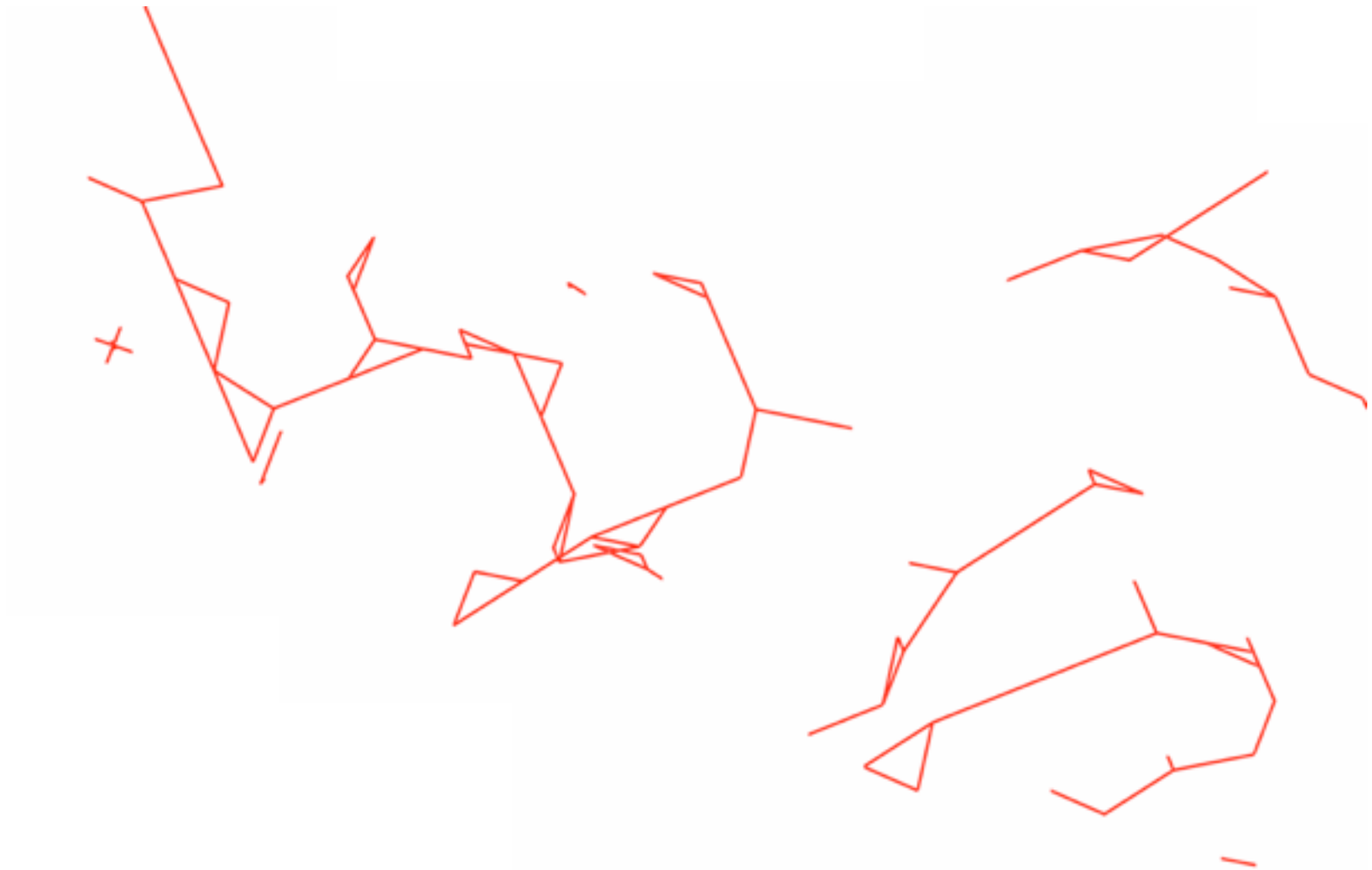
It allows the transformation (reduction) of information from **signals** to **meaning** given **a *priori* available knowledge:**





Methods for electron density map interpretation

Skeletonisation: tracing the path of the macromolecular chain using the connectivity of regions of high electron density



It gives a picture of how one can connect the electron density and where the main chain may run.

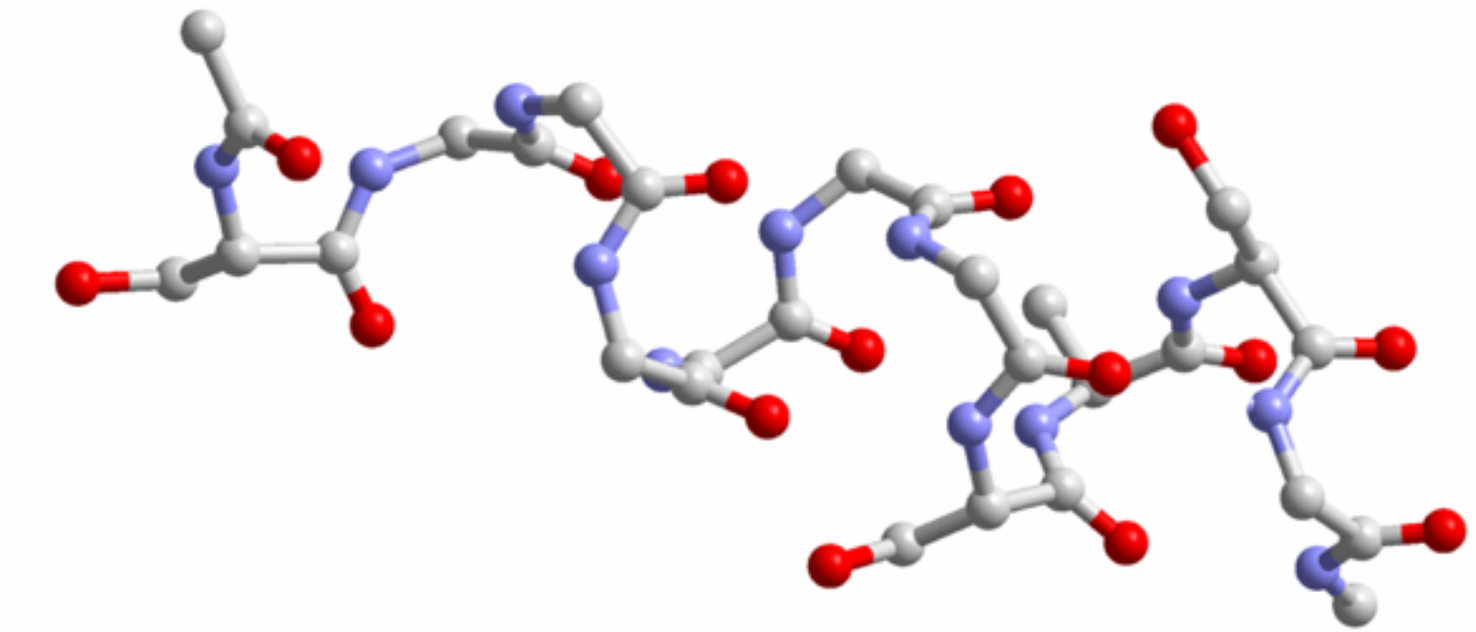
Methods for electron density map interpretation

Secondary structure templates: using penta-alanine templates in ideal α-helix and β-sheet conformations to find secondary structure elements in the density

It shows where we can find the secondary structure elements and then one has to try to connect them.

Methods for electron density map interpretation

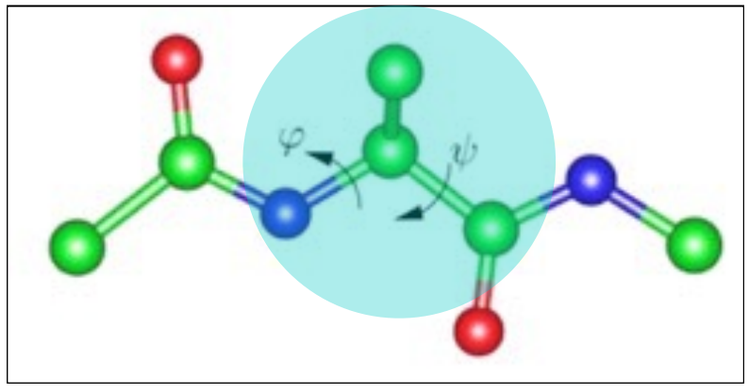
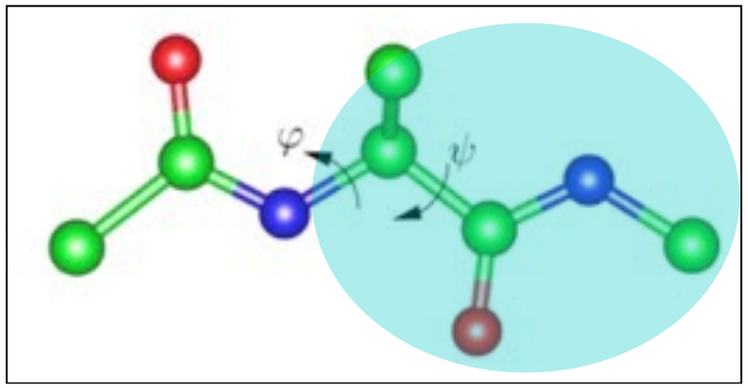
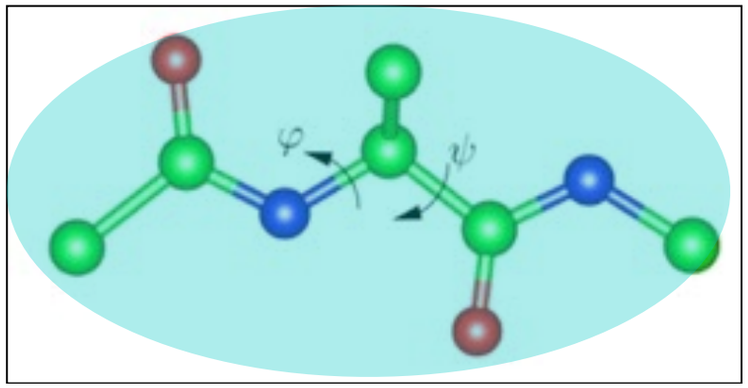
Free atoms: describing the electron density by a set of free atoms without chemical identity that can then be connected to form a main chain

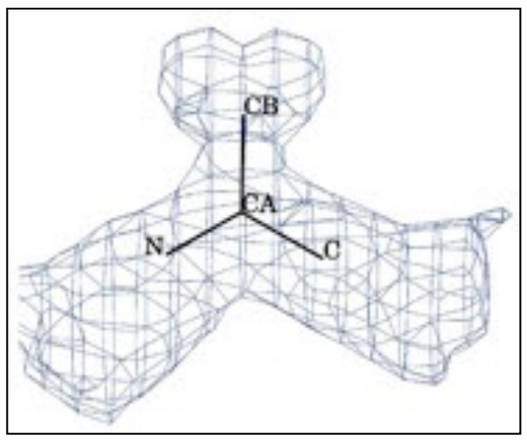


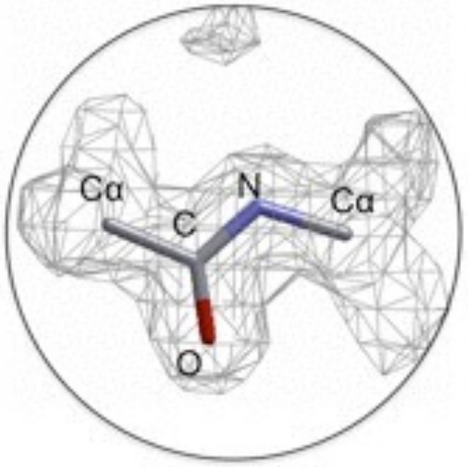
Searching for free atoms in possible Cα positions, the main chain can be found.

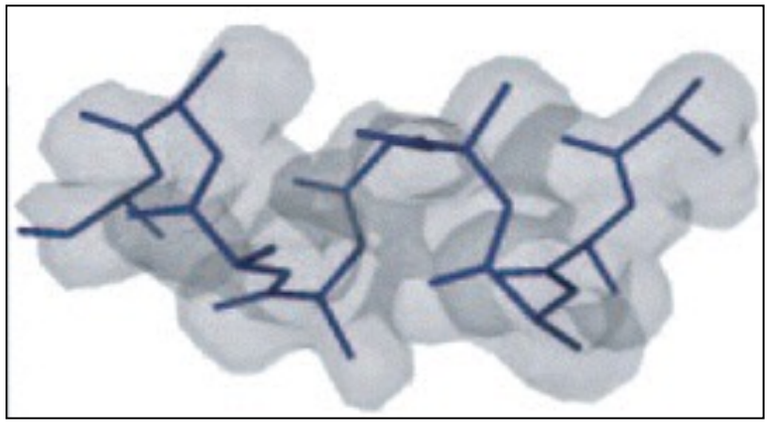
Methods for automated model building

Most current methods search for different patterns that are then used to build and extend a main chain:



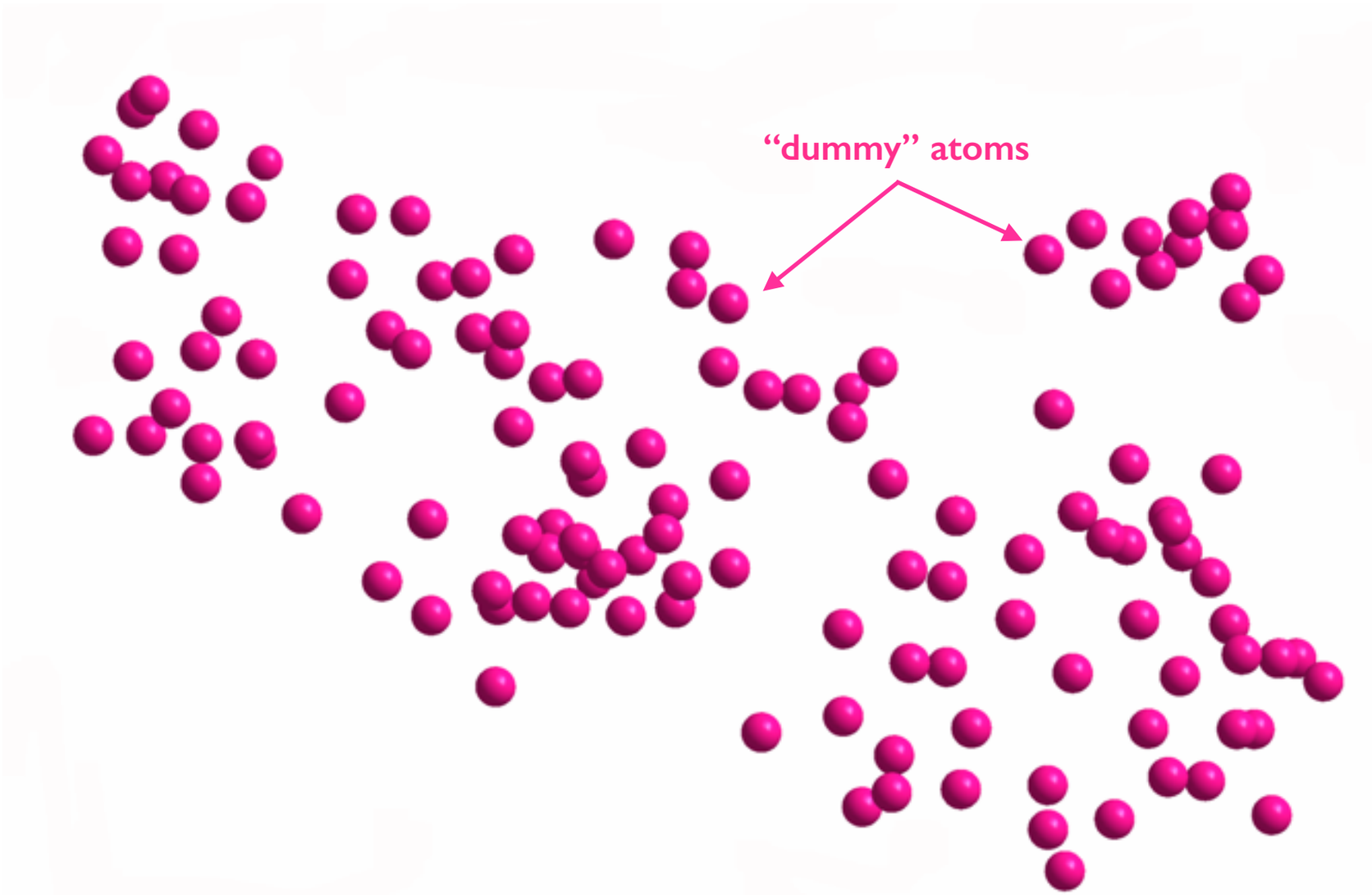




TEXTAL / Buccaneer	ARP/wARP	RESOLVE / ACMI
Cα	(Di-) Peptides	Fragments

The initial model

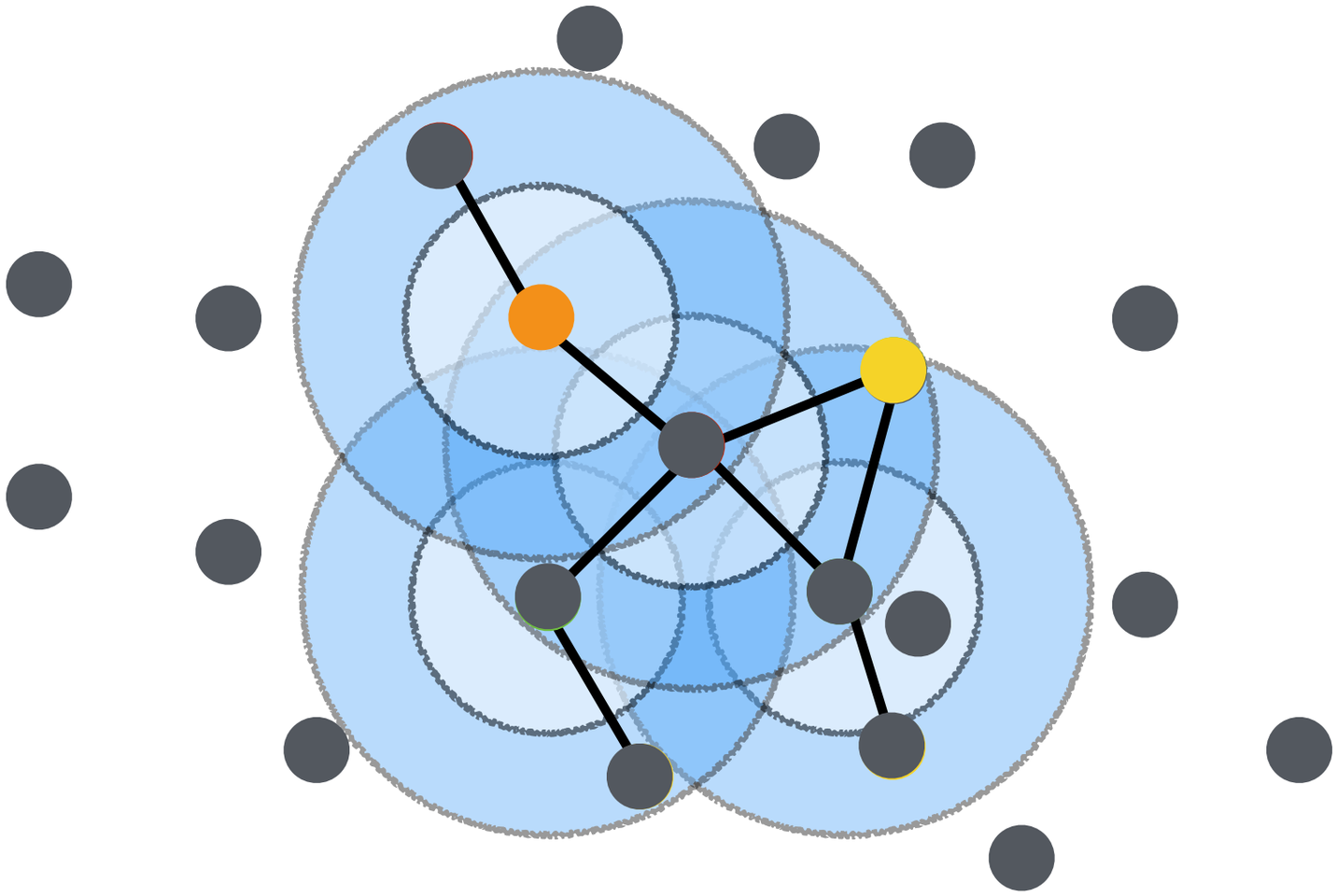
Given an electron density map and the protein sequence, ARP/wARP describes the electron density map as a set of free atoms without chemical identity:



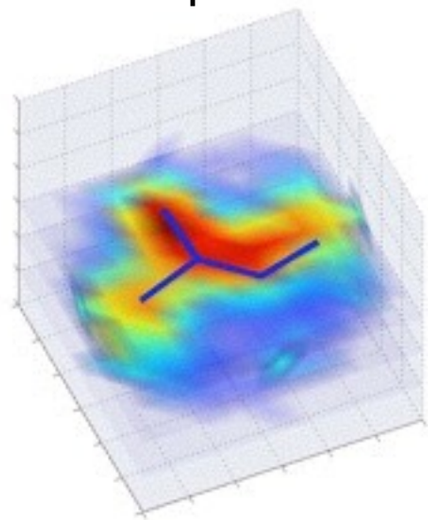
How to connect the free atoms?

Finding peptides

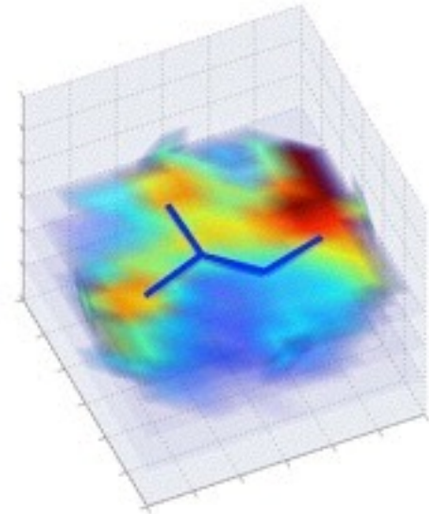
Free atoms at distances close to the ones observed in peptides are searched:

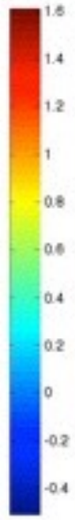


Peptide



No peptide



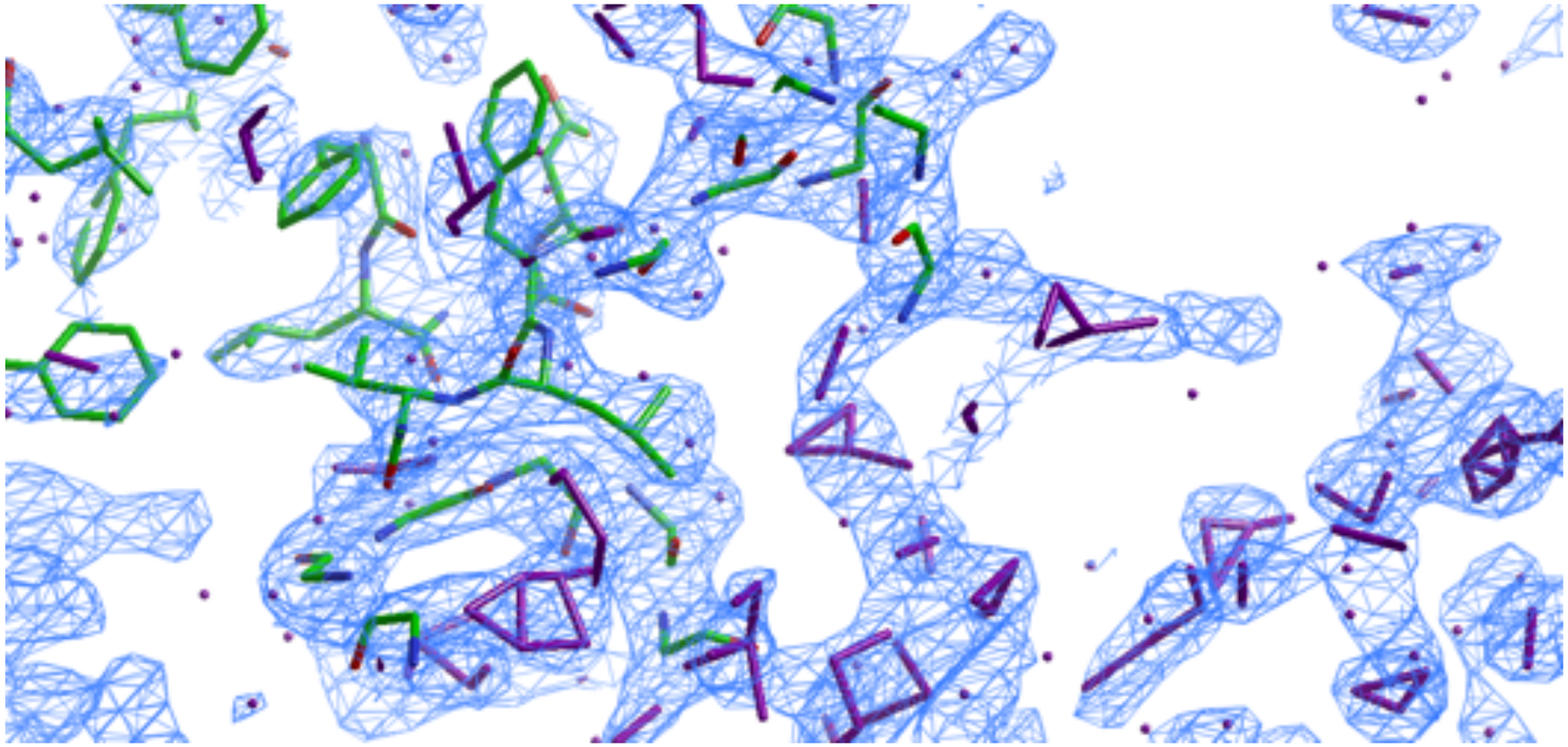


Peptides that share one Cα atom can be connected and dipeptides with correct conformation connected to build the longest fragment

The hybrid model

As only a small part of the initial free atoms model is now recognised as a putative protein, we have an hybrid model that combines 2 sources of information:

- **Chemical knowledge** from the partially built model
- **Free atoms** continue to describe the electron density

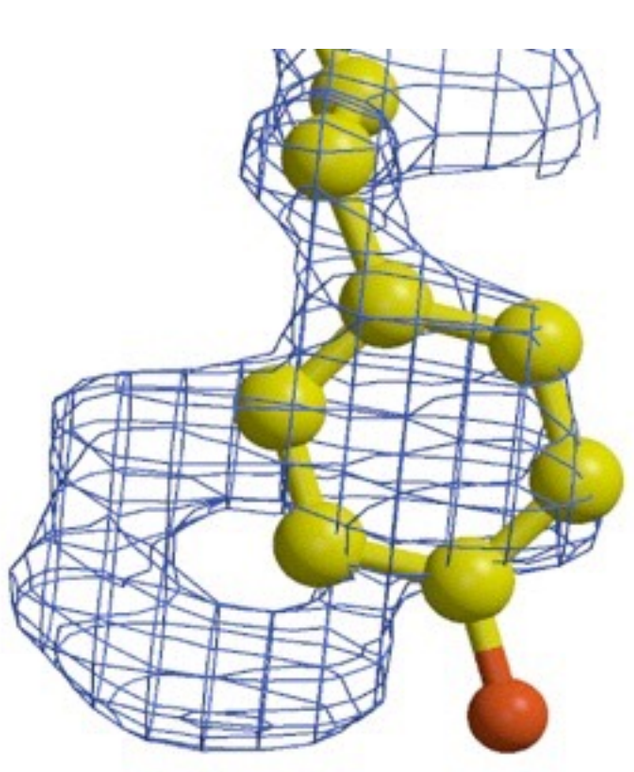


It can provide restraints for further **refinement**

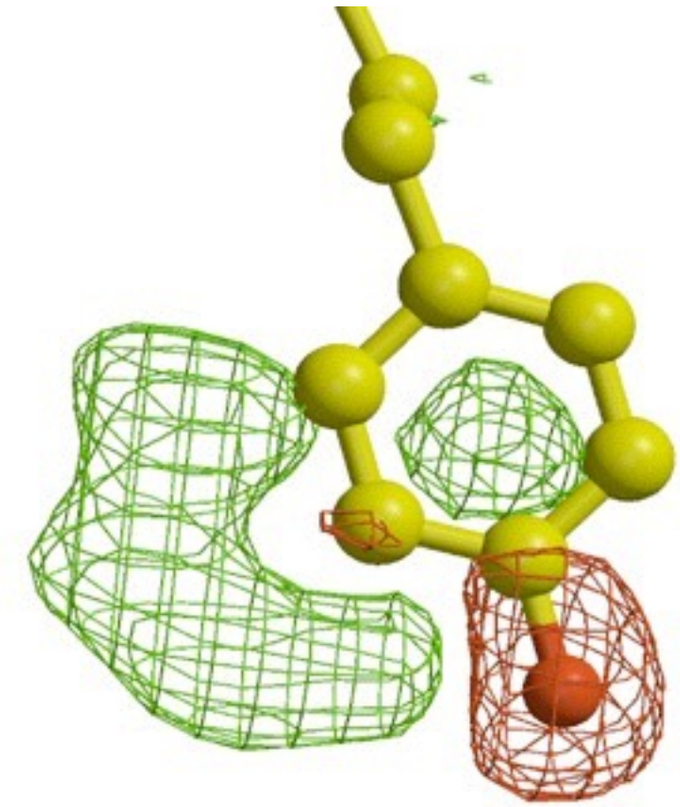
Allows the improvement of the map and consequent better model building

Model update

As the density is improved by refinement, the model can be updated by removing and addition of atoms:



2Fo-Fc map

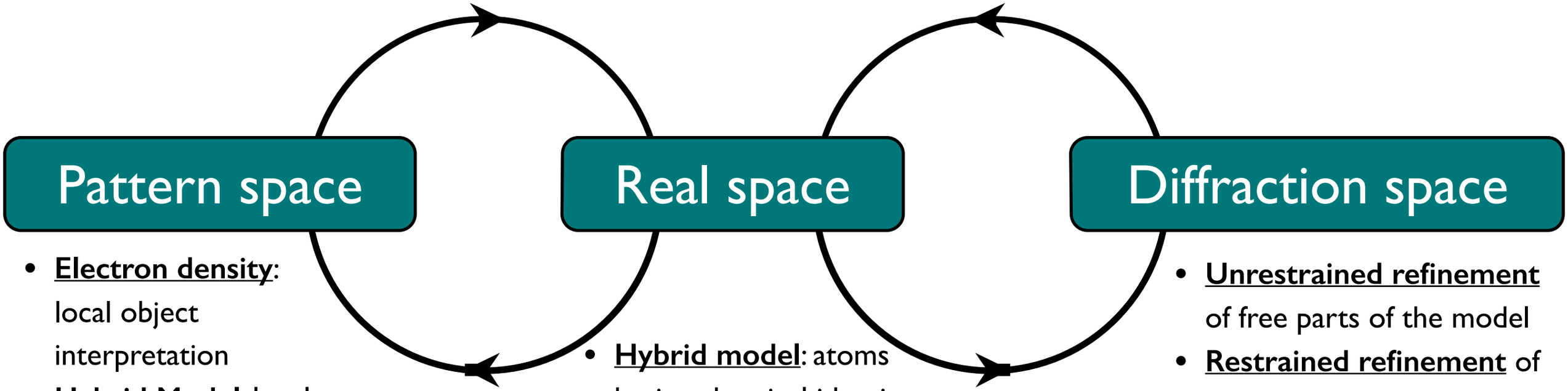


Fo-Fc map

```
ARP/wARP will be iterated with REFMAC5
50 refinement / model update cycles will be run in total.
Atoms will be removed below 1.0 sigma in 2mFoDFc map and added above 3.2 sigma in mFoDFc
map.
```

The fundamental ARP/wARP concept

ARP/wARP circulates between three spaces, forming a unified process of model building:



- **Electron density**: local object interpretation
- **Hybrid Model**: local motif interpretation

- **Hybrid model**: atoms having chemical identity and free atoms
- **Model update**: removing and adding parts of the model

- **Unrestrained refinement** of free parts of the model
- **Restrained refinement** of chemically assigned atoms

Scheme of restraints and free atoms are iteratively updated

Hybrid model is converging to the final model

ARP/wARP workflow

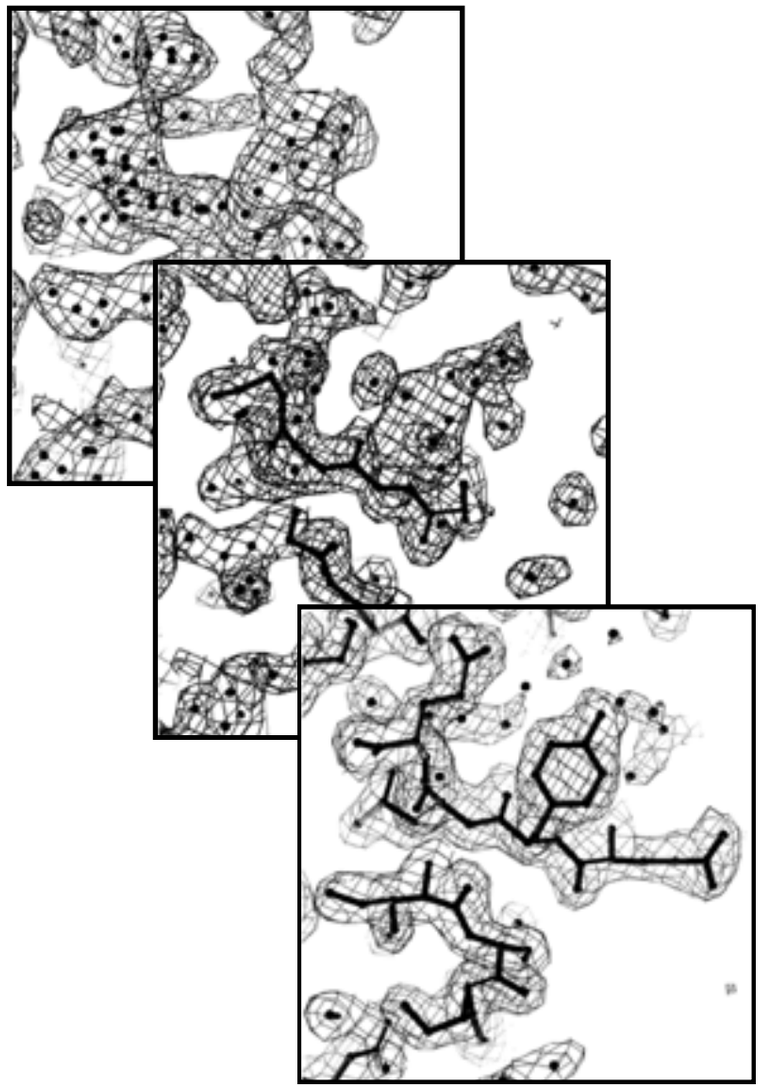
ARP/wARP iterates between main chain tracing, refinement and model update to build models that converge to the final model:

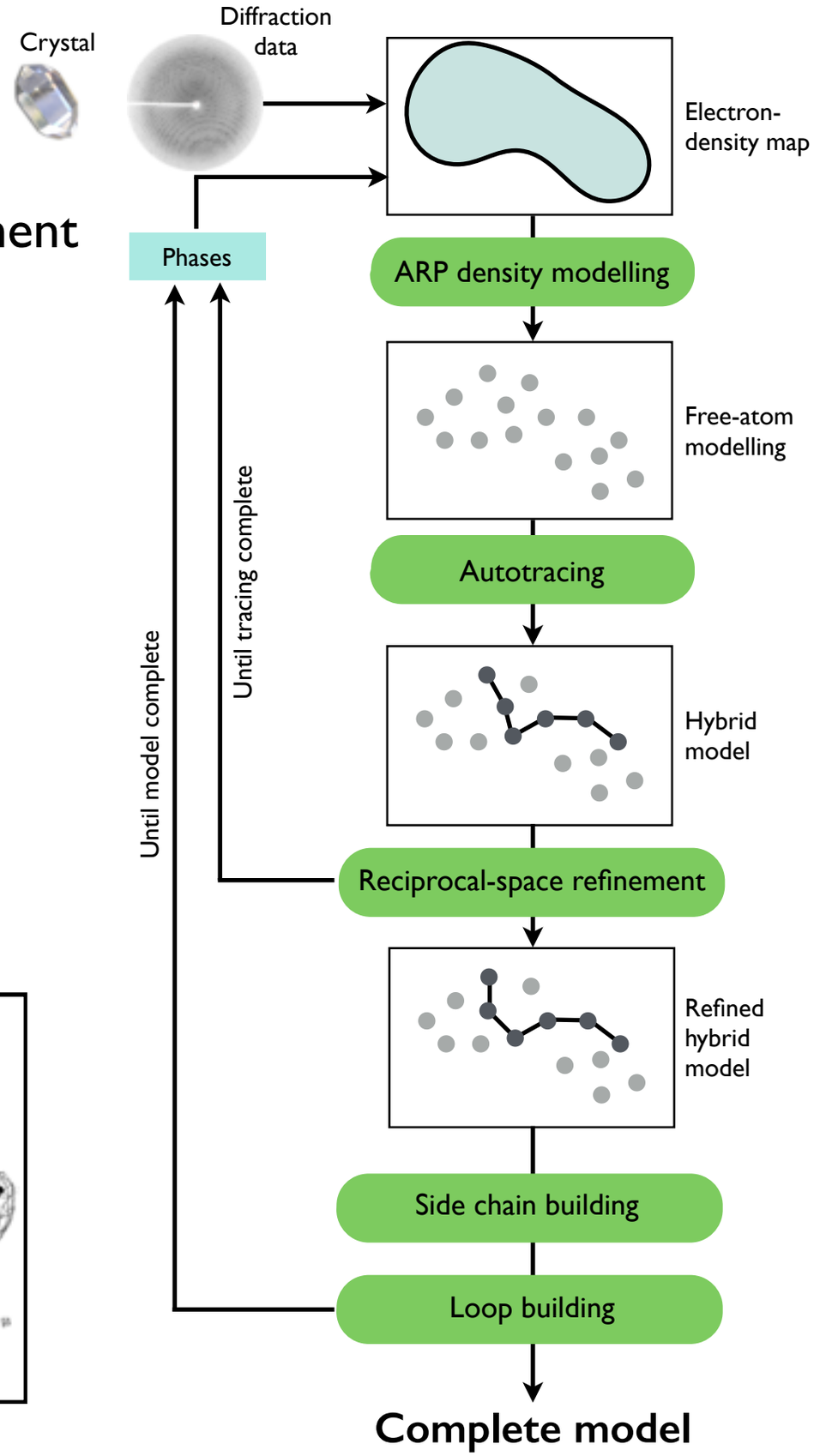
Free Atoms

Pattern Recognition

Hybrid Models

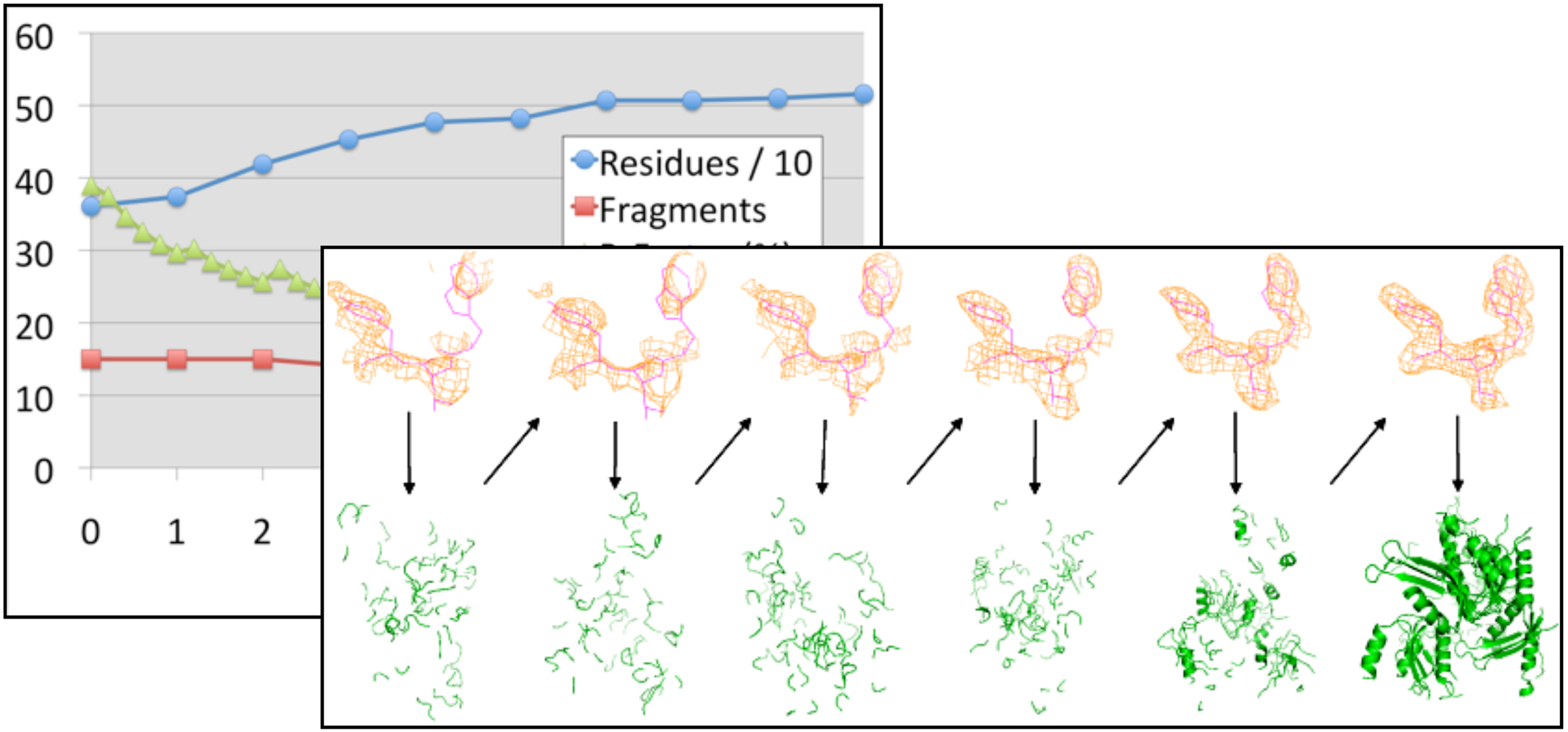
Iterations





ARP/wARP workflow

ARP/wARP iterates between main chain tracing, refinement and model update to build models that converge to the final model:

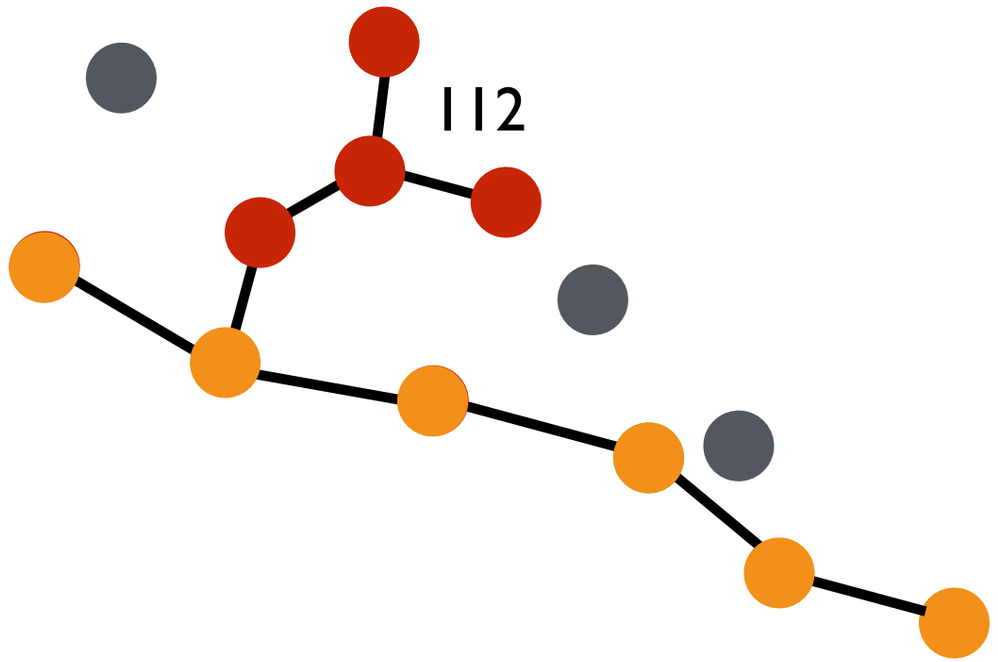


Sequence docking and side chain building

After main-chain tracing, one gets a set of main-chain fragments and the side chain-building module has to:

1. **Sequence docking** using connectivity vectors
   (e.g., serine "11", aspartate "112")

2. **Side chain building** using the Richardson's rotamer database

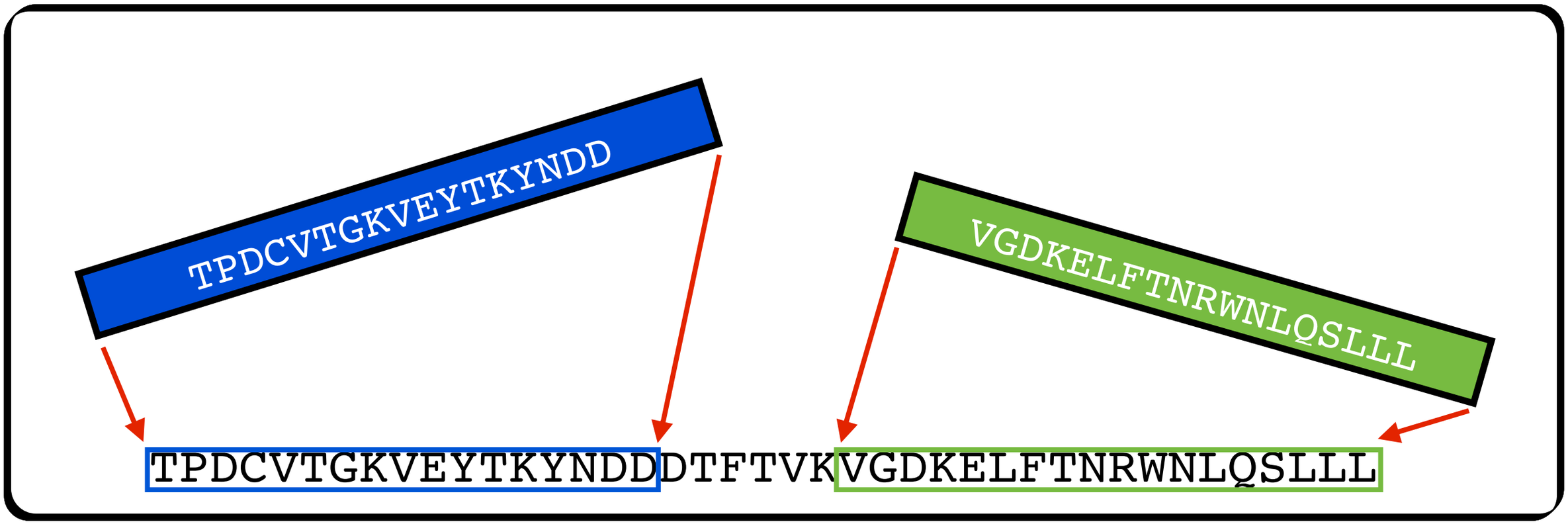


This allows the construction of loops!

Loop building

Even after a successful automated model-building run, some regions of known start, end, length and amino-acid sequence remain to be built, and ARP/wARP builds these **loops** by:

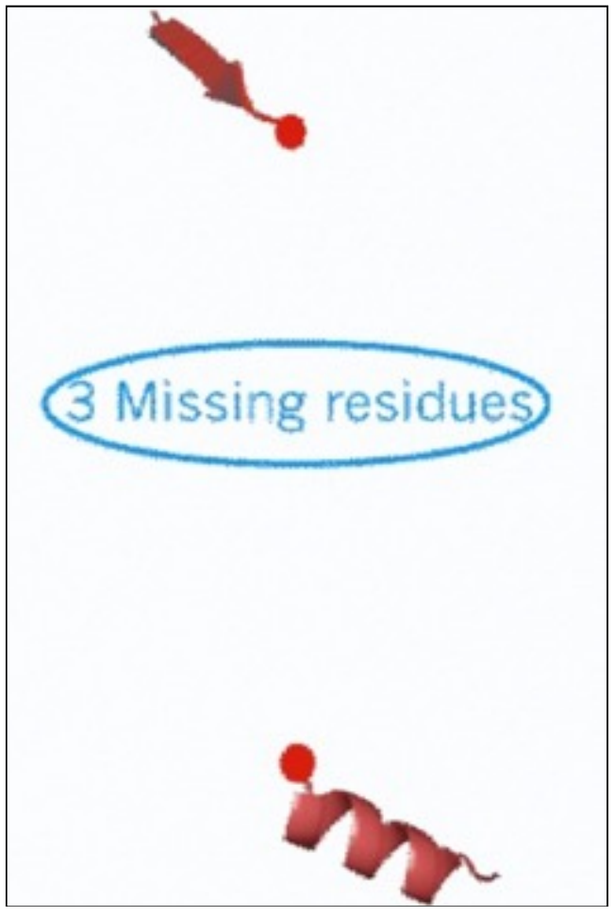
- Using **knowledge of angles** between anchors
- **Check with density** to see if modelled loops are valid

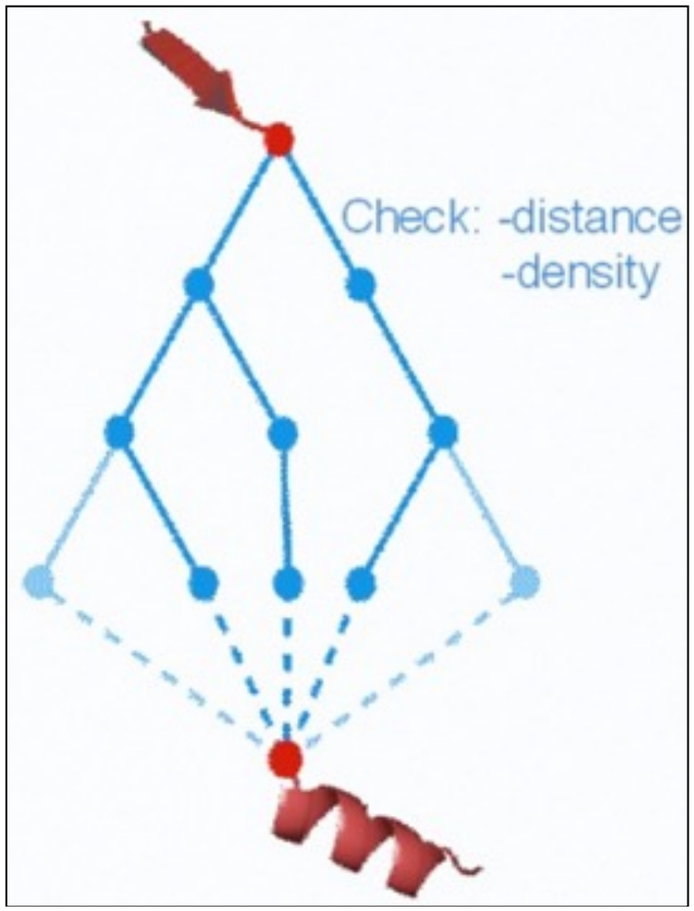


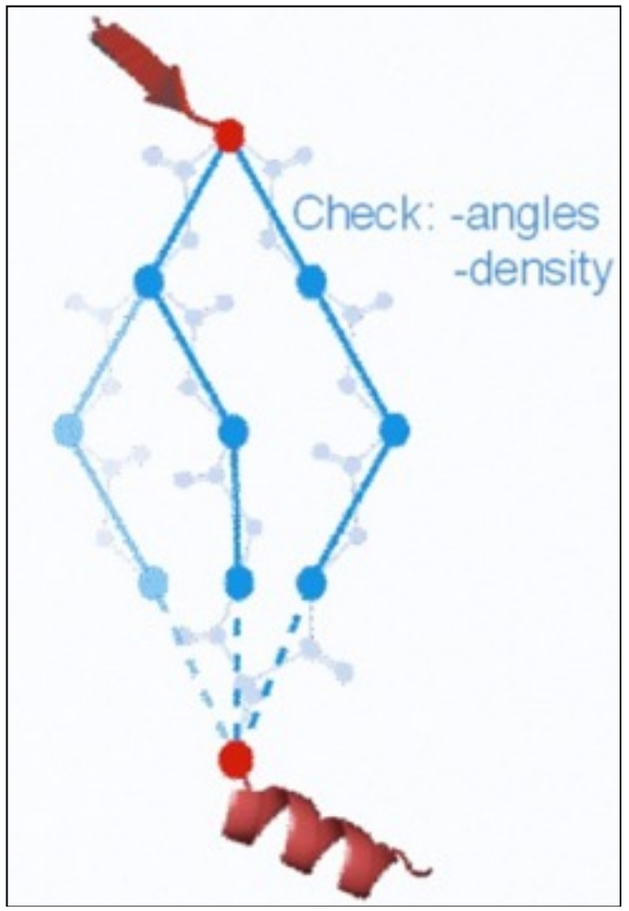
Loop building

Even after a successful automated model-building run, some regions of known start, end, length and amino-acid sequence remain to be built, and ARP/wARP builds these **loops** by:

- Using **knowledge of angles** between anchors
- **Check with density** to see if modelled loops are valid

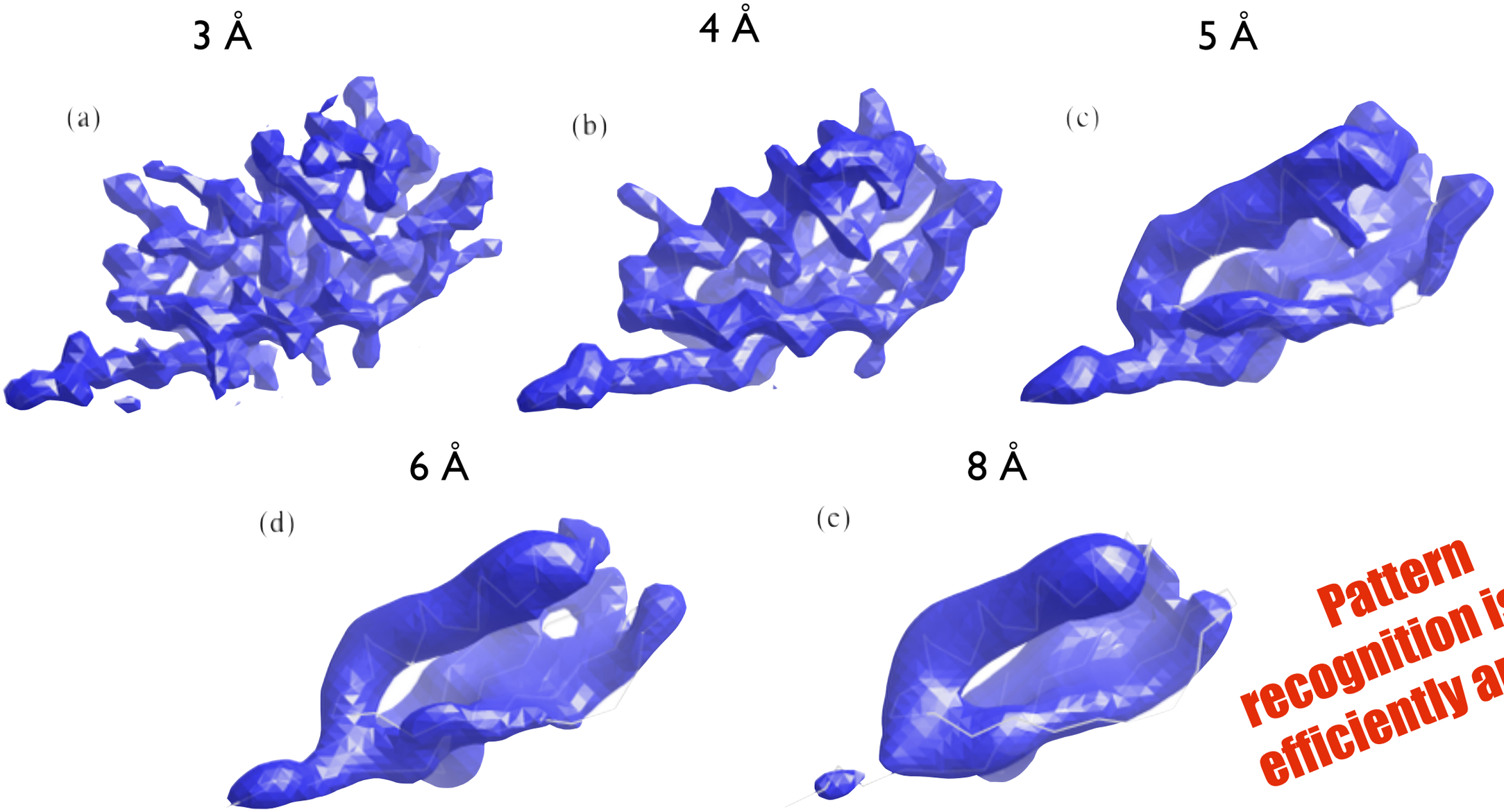






Low resolution as a limit

The primary factor determining the accuracy of an experimental protein structure is the **resolution of the X-ray data**:



Pattern recognition is not efficiently applied

ARP/wARP and resolution

Model building methods were historically developed for and work considerably better with high-resolution diffraction data:

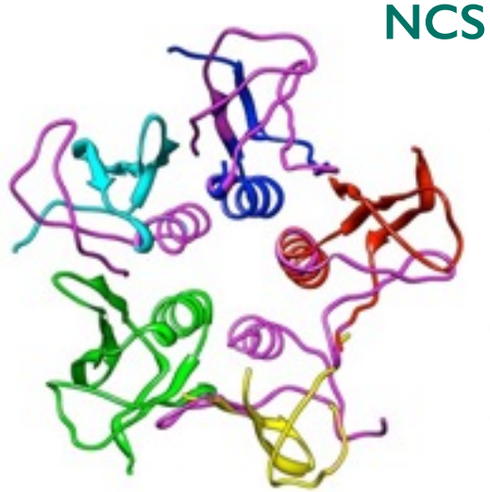
Results from the ARP/wARP webservice (tracing performance)

Resolution	Estimated fraction of automatically built protein structure (October 2011 – March 2012)
1.5 Å and higher	98% and higher
1.5 - 2.0 Å	91 – 98%
2.0 - 2.5 Å	81 - 91%
2.5 - 3.0 Å	67 - 81%
3.0 – 3.5 Å	51 - 67%

Protein model building at medium-to-low resolution in the best case generates incomplete and fragmented macromolecular models!

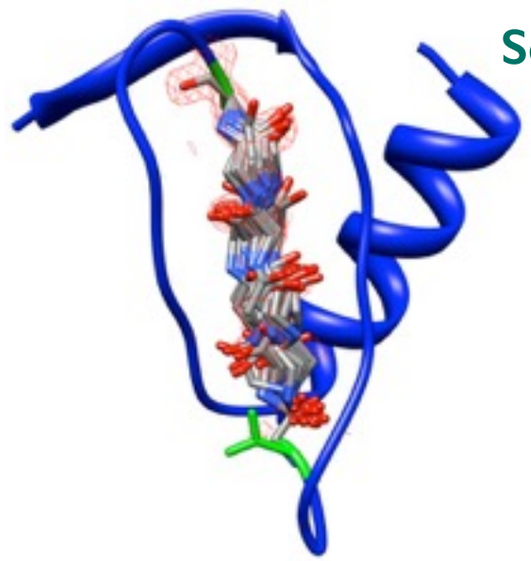
Methods to improve modelling at lower resolution

There's a big effort now to improve automated model building at medium-to-low resolution:



NCS detection and extension

- Detection of NCS related fragments and extension by "copy-paste"
- Improvements at 2.0-3.5 Å
- PNSextender



Sequence-independent loop fitting

- Secondary structure prediction docking to find fragments that can be connected
- Search for links based on the gap size probability and a database of fragments
- Improvements at 3.0-3.8 Å

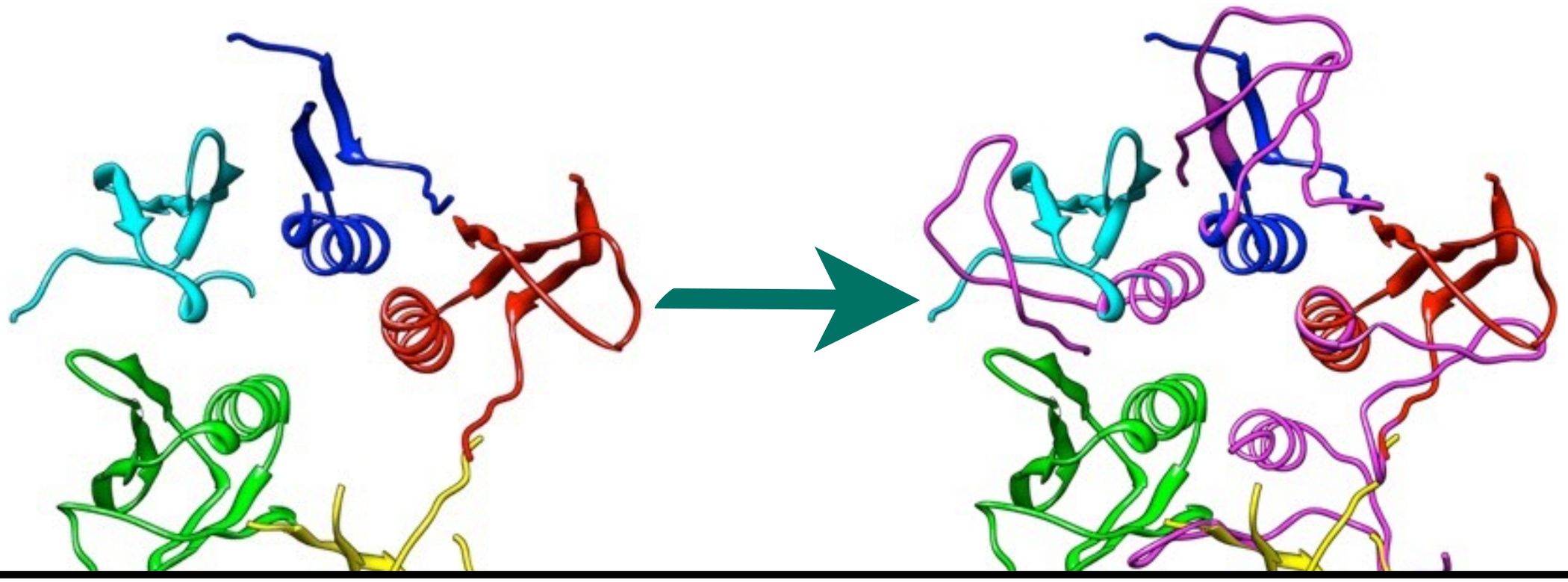
NCS Detection and Extension

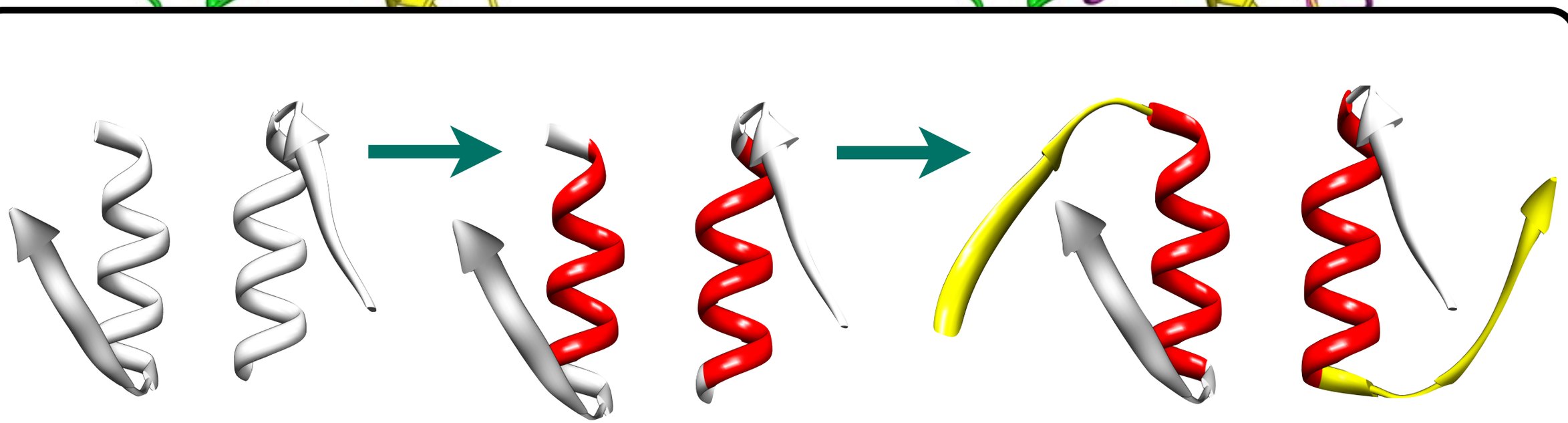
Non-Crystallographic Symmetry occurs if there are multiple subunits in the asymmetric unit of a crystal and some of them adopt almost the same tertiary structure:

Basically like dogs!

NCS Detection and Extension

We can use this information to extended fragmented models!

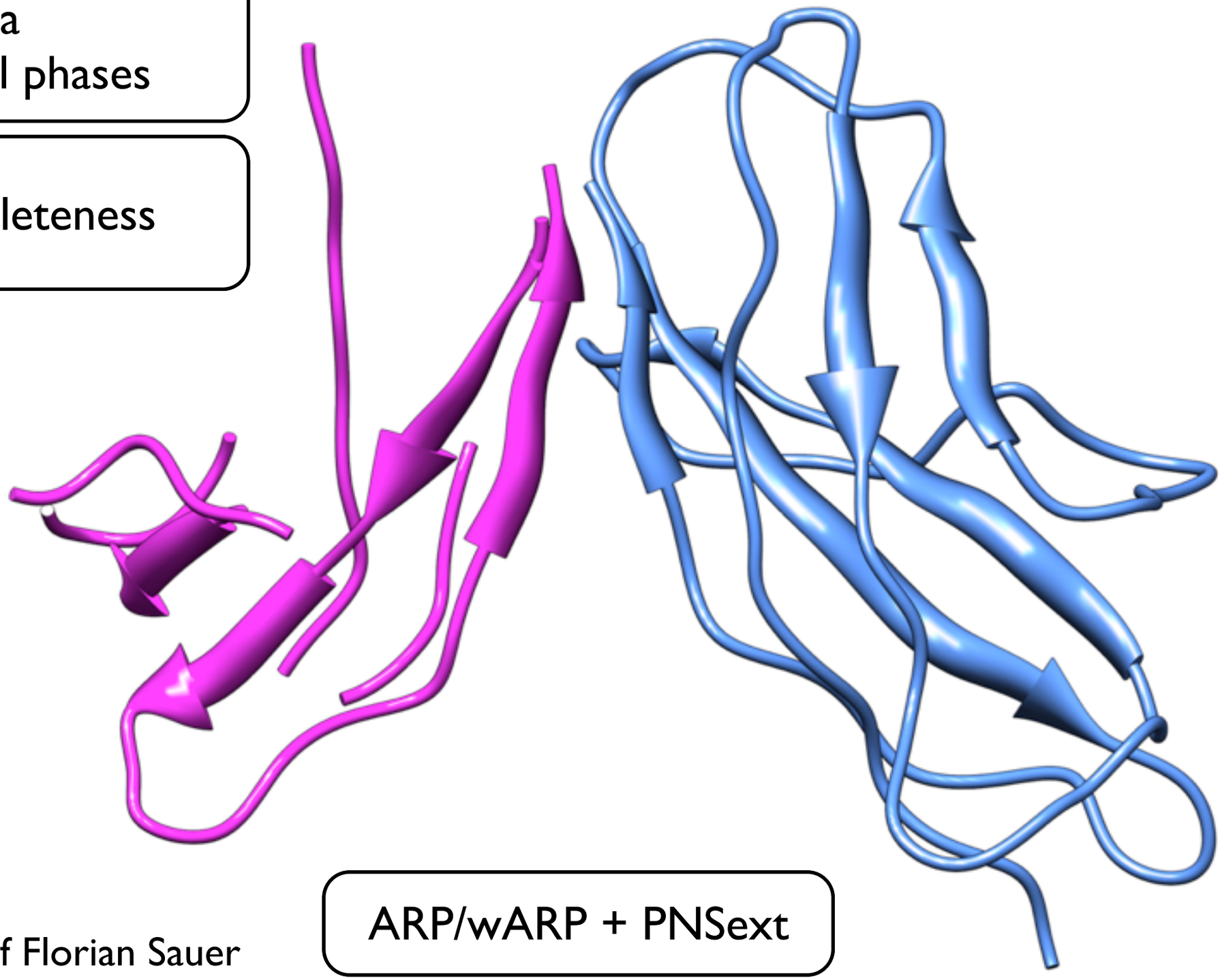




NCS Detection and Extension

- 1.95 Å data
- poor initial phases

- 75% completeness



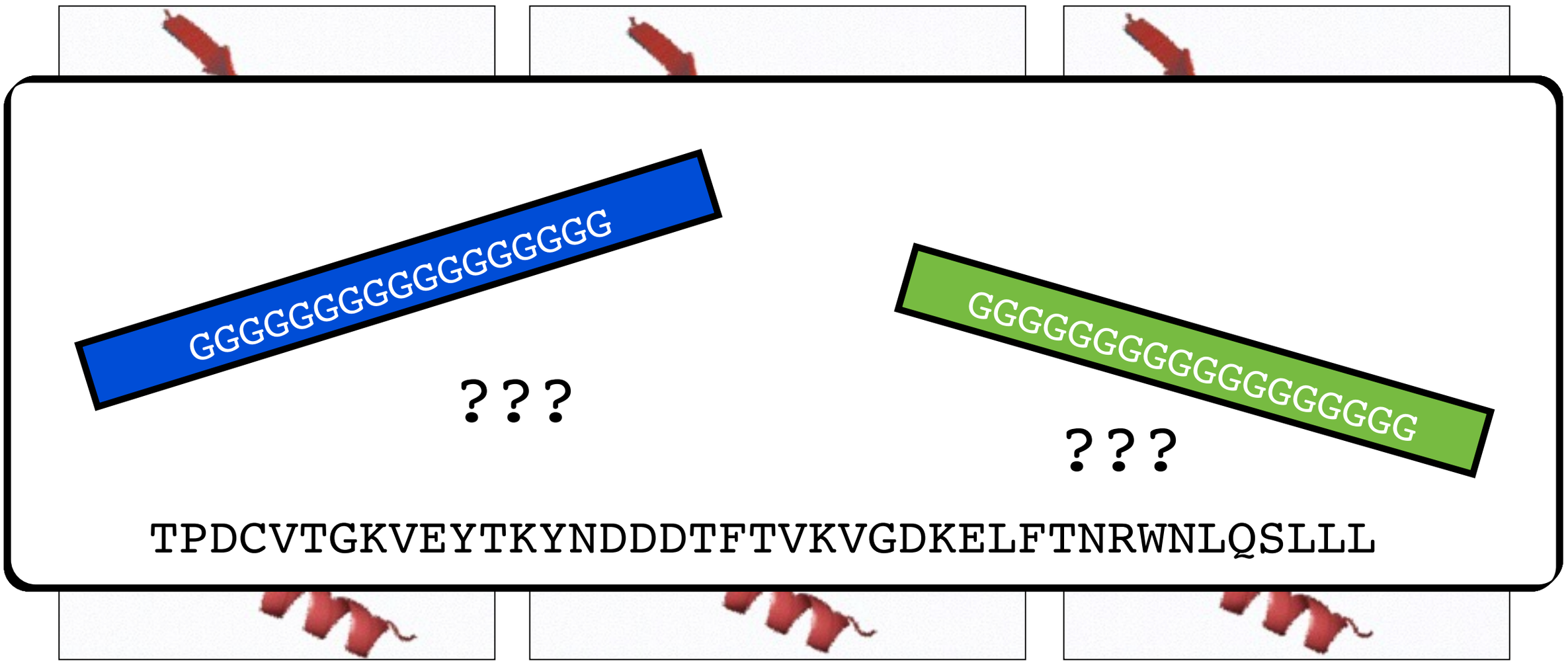
ARP/wARP + PNSext

Data courtesy of Florian Sauer

Modelling Loops With *a priori* Knowledge

Chain fragments are built and **docked to sequence**

- using **knowledge of angles** between anchors
- **check with density** to see if modeled loops are valid

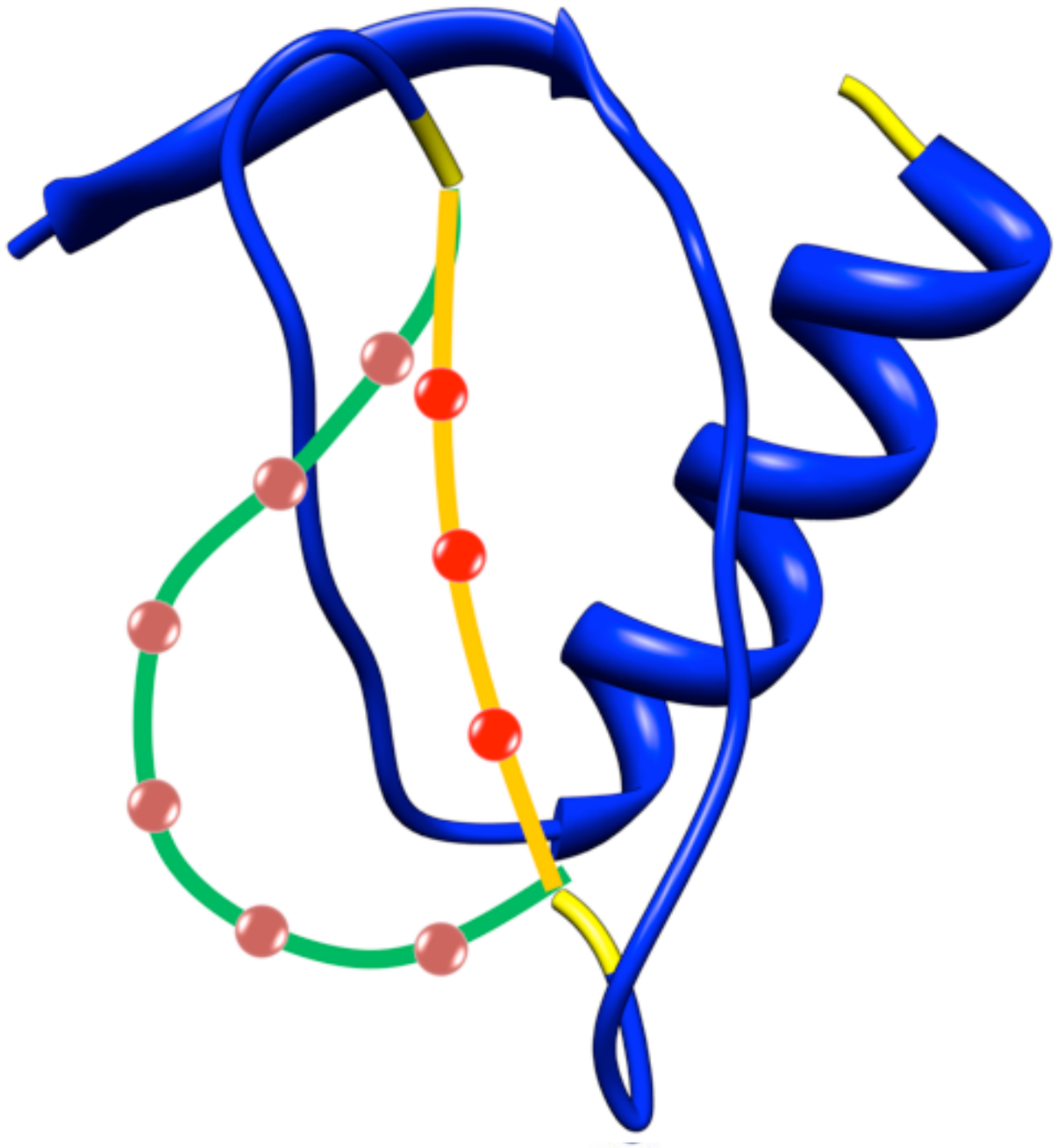


Modelling Loops With *a priori* Knowledge

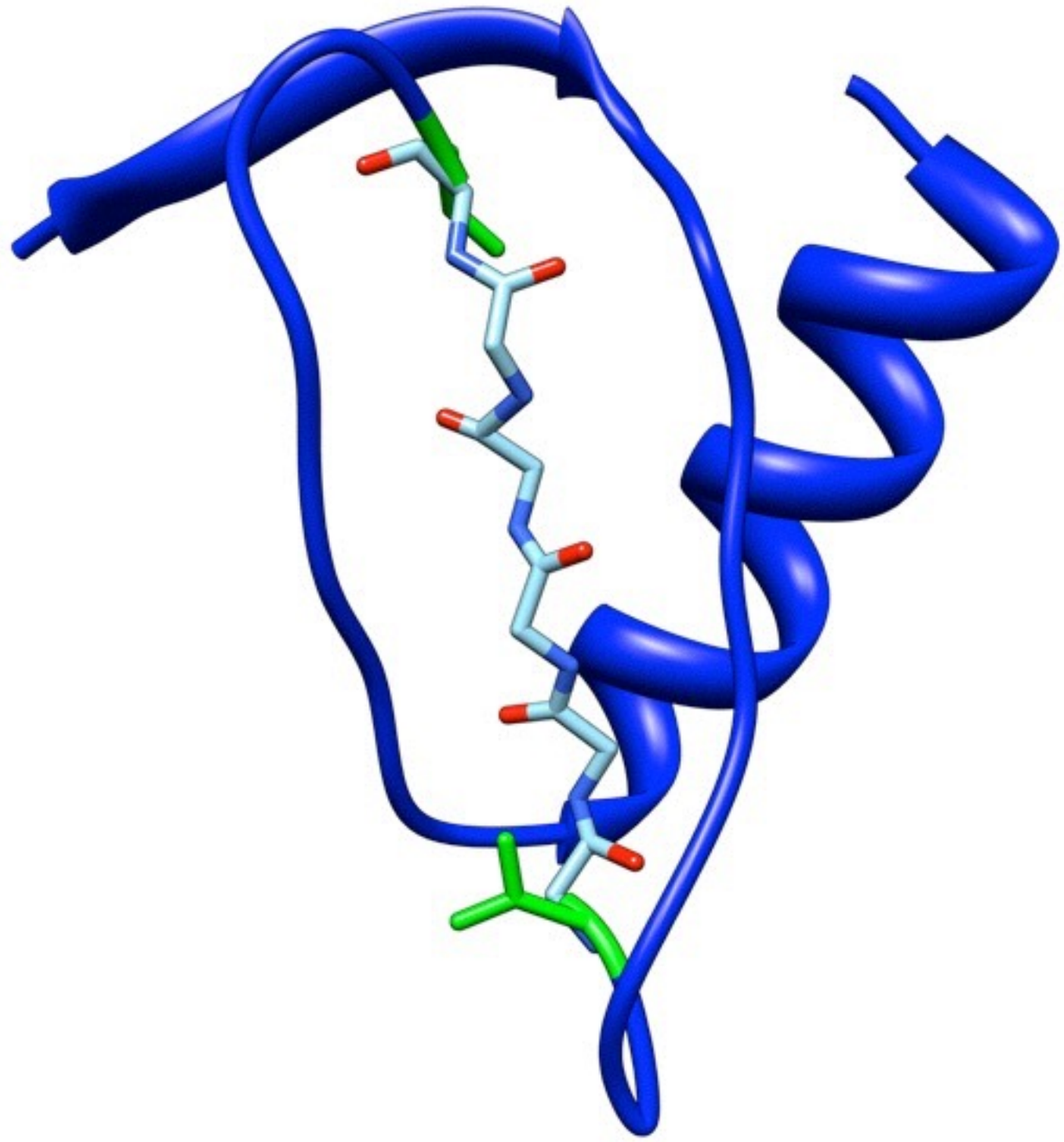
When **no sequence** has been **docked** - **no loop building!**

- Search for **missing fragments** in a **structural database**

- *In development* together with Marco Biasini (Schwede group, SWISSmodel)
  - Using the given **density information**

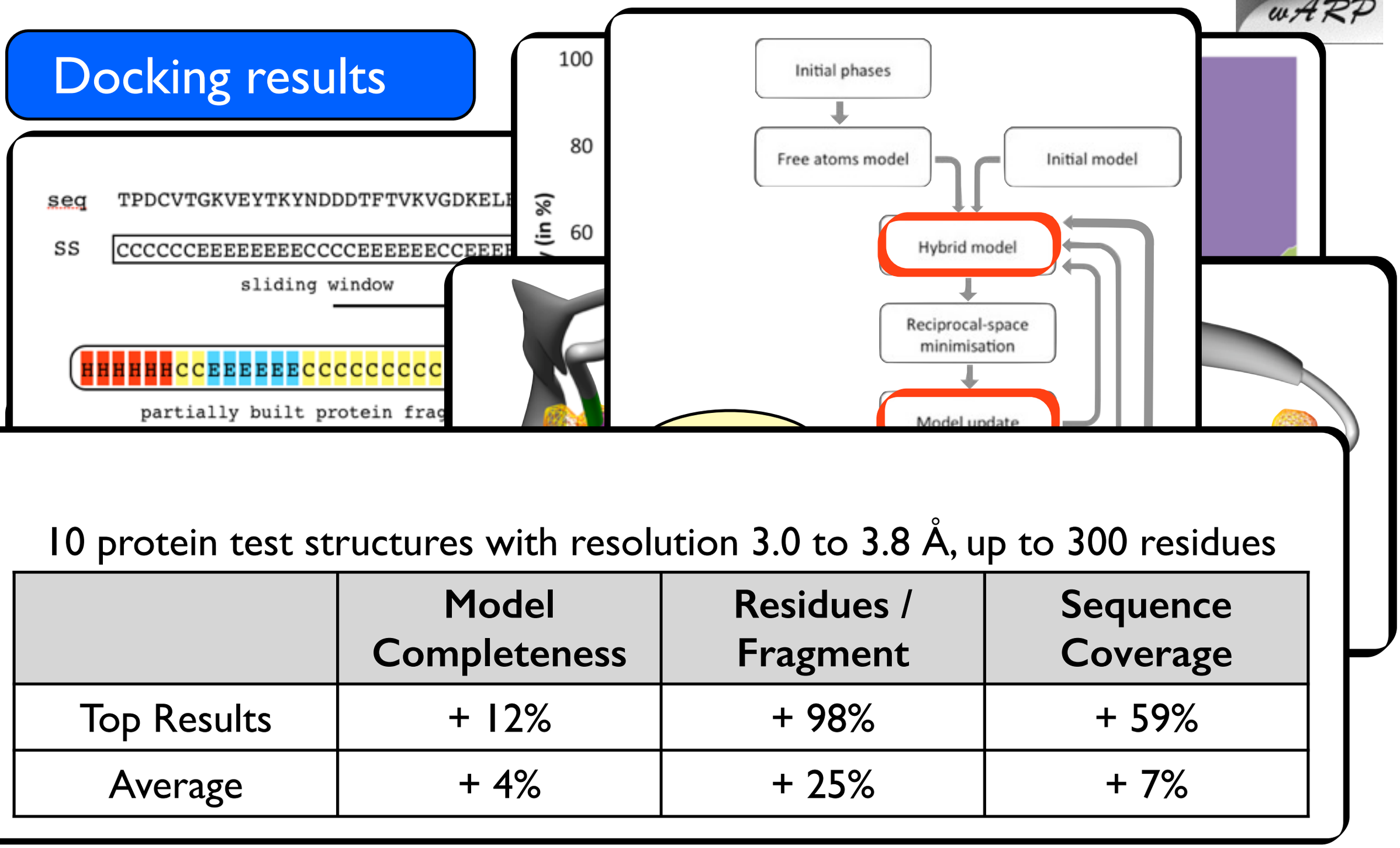


Modelling Loops With *a priori* Knowledge



Modelling Loops With *a priori* Knowledge

Docking results



10 protein test structures with resolution 3.0 to 3.8 Å, up to 300 residues

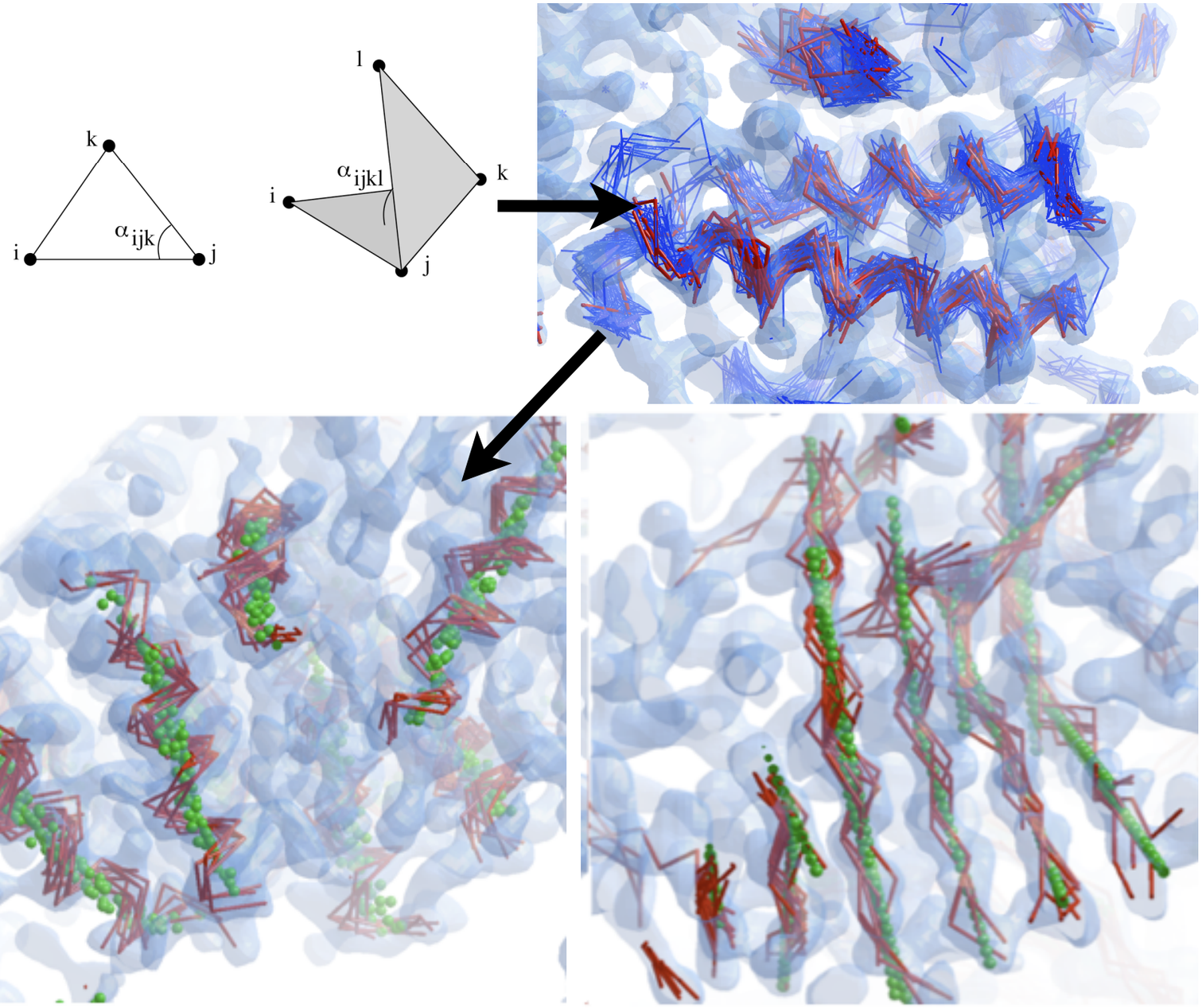
	Model Completeness	Residues / Fragment	Sequence Coverage
Top Results	+ 12%	+ 98%	+ 59%
Average	+ 4%	+ 25%	+ 7%

Secondary structure modelling

At low resolution, one can try to find **secondary structure elements** on the map:

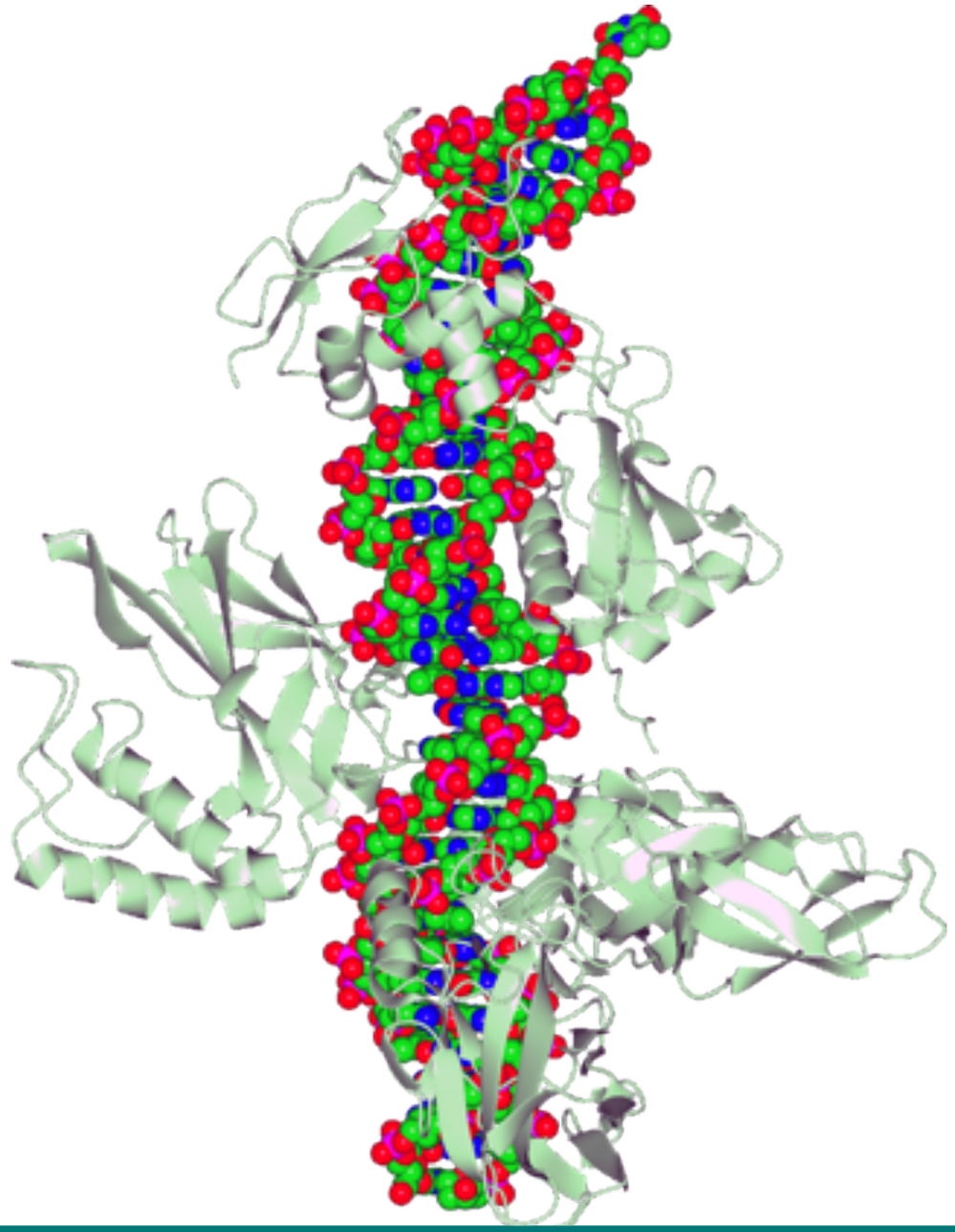
- **Short helix/strand fragments** (3 to 5 Cα candidates) are built.
- Longer traces are formed and the **best are kept** (in red)
- Traces are **clustered**
- Assemblies are **averaged**

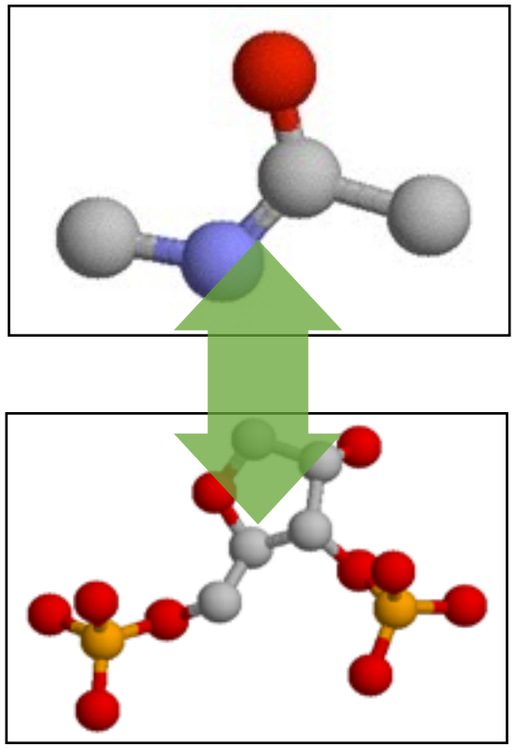
Good to evaluate map quality!

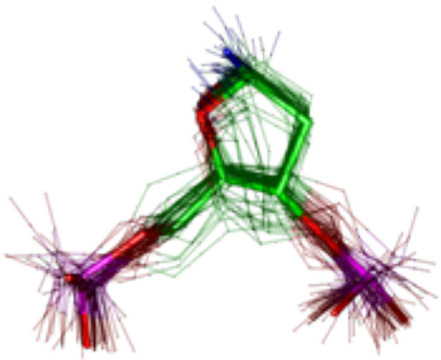


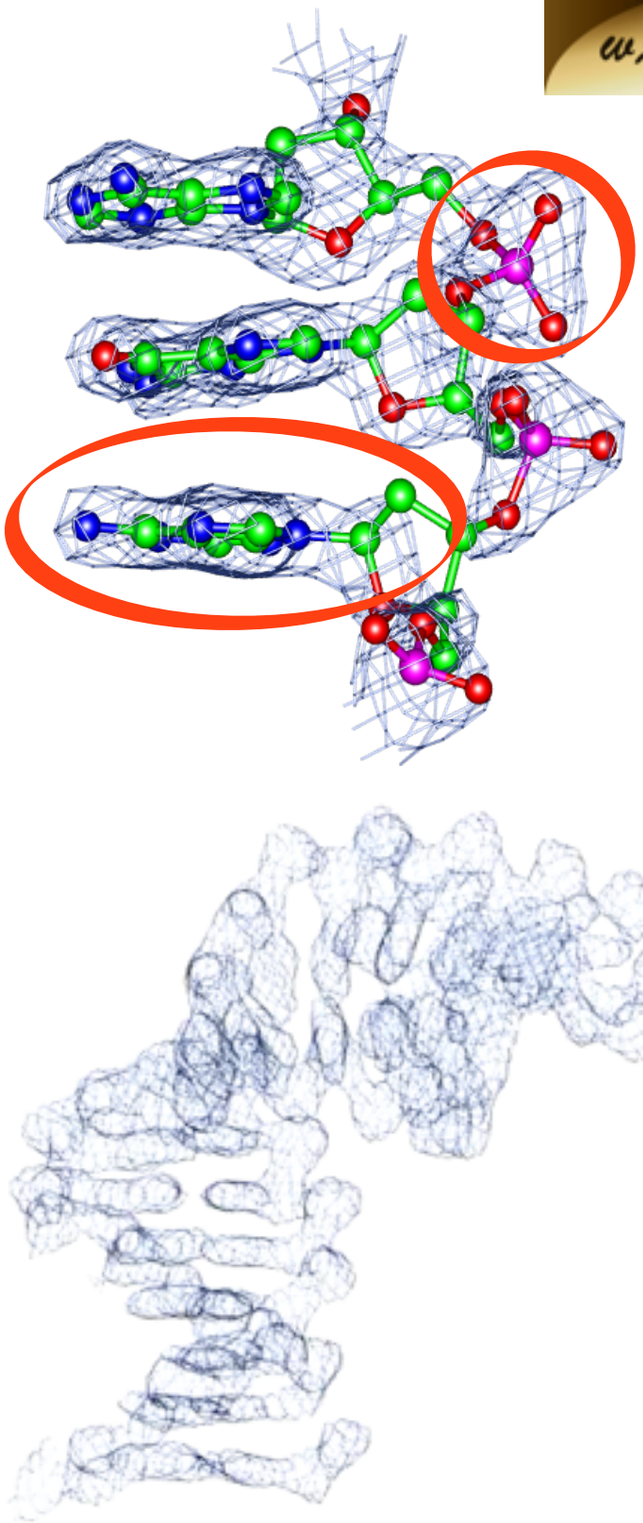
Tracing RNA/DNA Chains

ARP/wARP is just not about building protein models, it can also build **nucleic acids**:





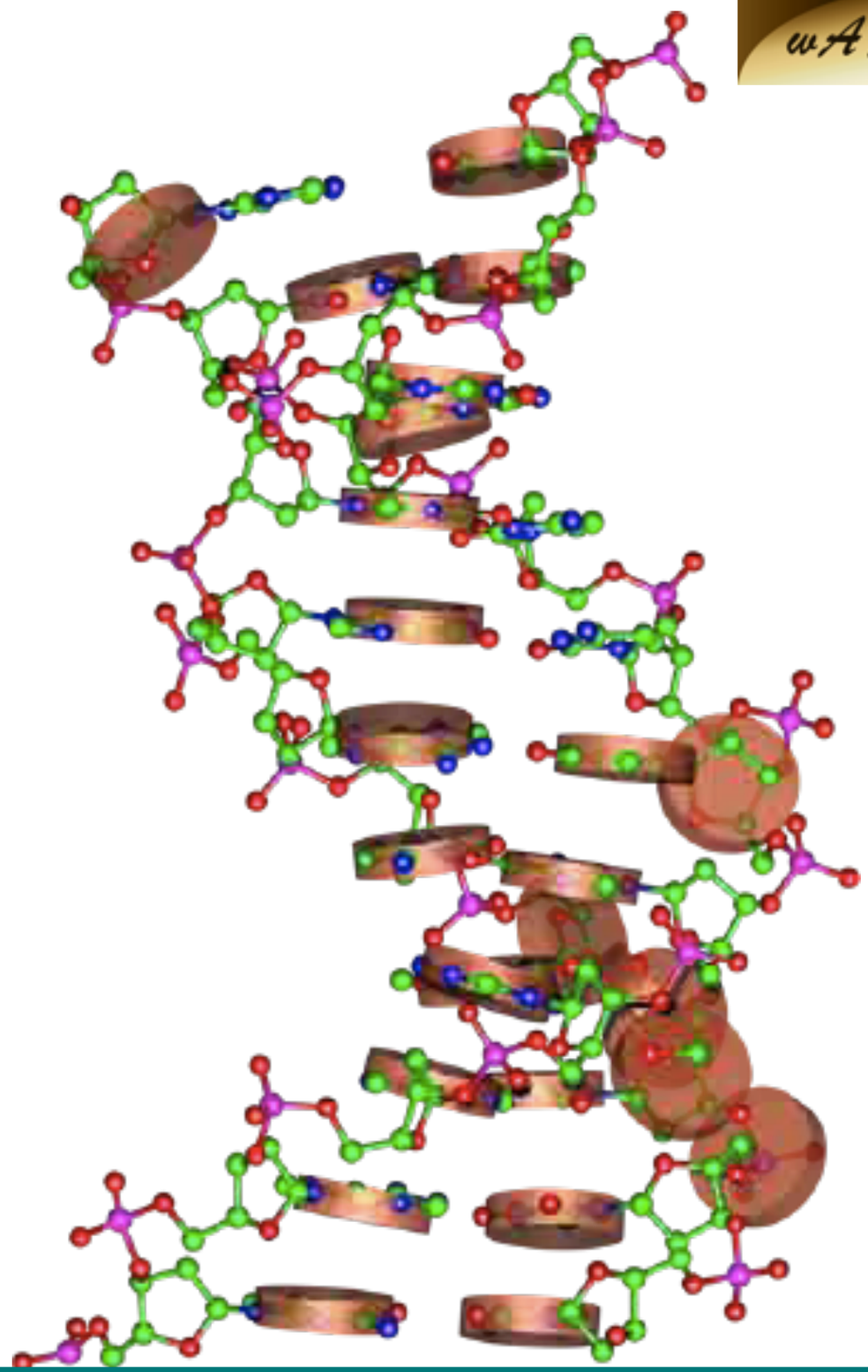




Tracing RNA/DNA Chains

- 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°

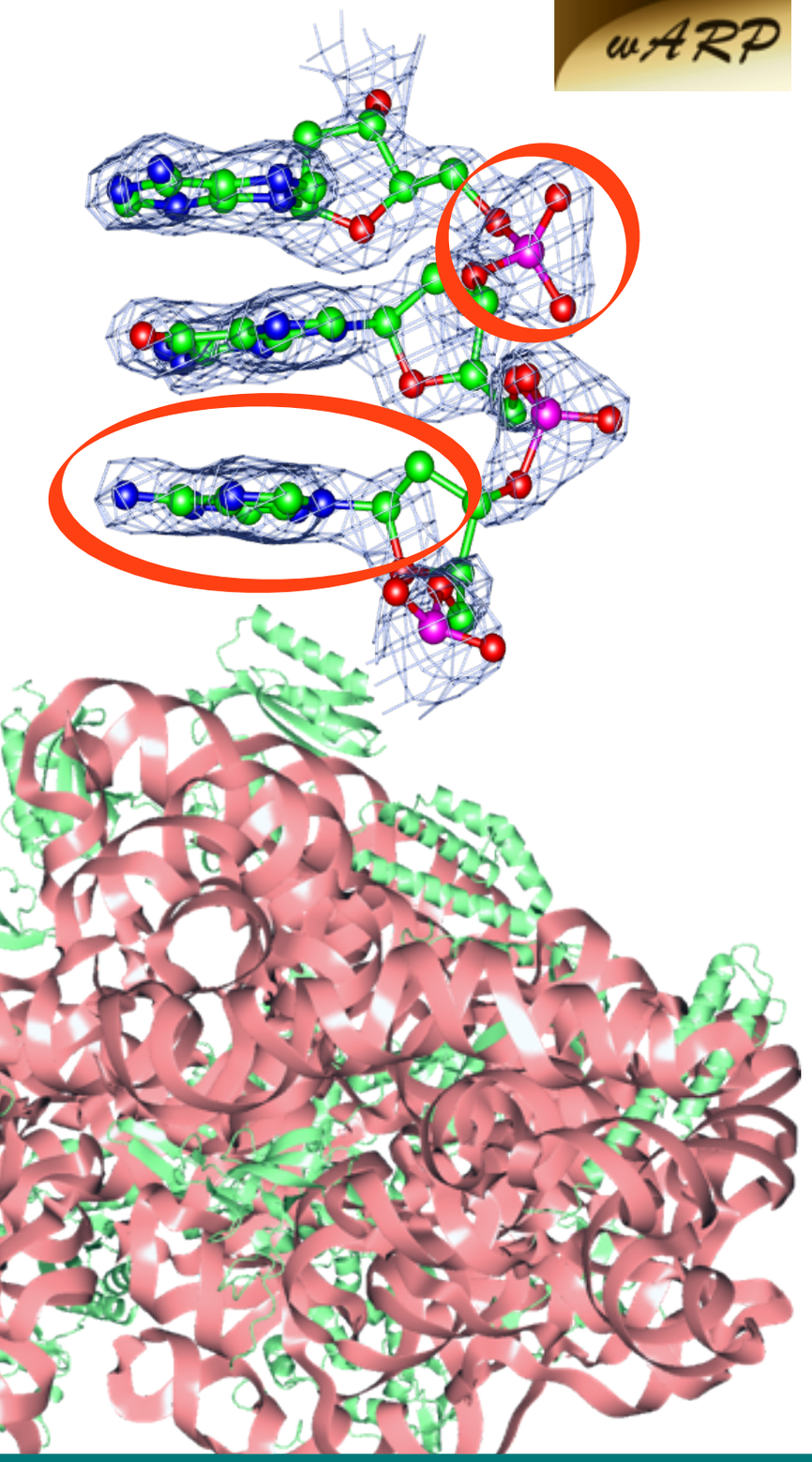


Tracing RNA/DNA Chains

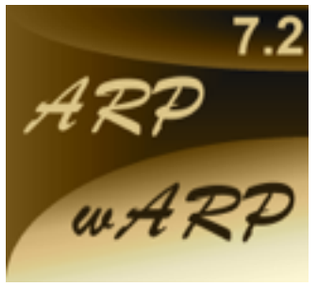
- The **30S ribosomal subunit**
  - Resolution: 3.05 Å
  - Experimental phases (MCC 0.73)

- 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)

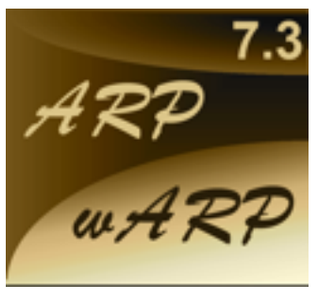


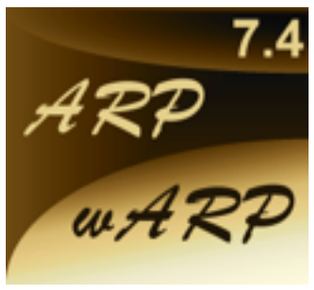
What can I find in ARP/wARP right now?



Changes in Version 7.4

- Accuracy of estimated model correctness has been improved
- Updated method for fitting of partial ligands
- "Fit Ligand" incorporates 84 most common ligands now
  - and uses cif files defining bond, torsion and plane restraints

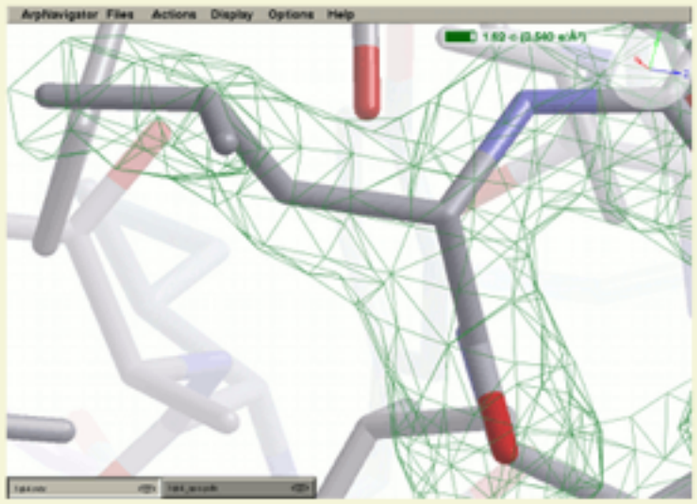




Where can I find it?

http://www.embl-hamburg.de/ARP/

ARP/wARP Copyright 2013
by European Molecular Biology Laboratory



Click on the images above to zoom.

ARP/wARP: Crystallographic Macromolecular Model Building Version 7.4

Builds proteins, RNA/DNA, secondary structure, side chains, loops, solvent and ligands

Download available for Mac and Linux.

Model building also available over the web.

Download/License

Web Service

Only for Mac or Linux!!

Good for Windows users!!

DOCUMENTATION

Download User Guide (pdf)
View User Guide (html)
Download Tutorial (pdf)

MAILING LISTS

Subscribe to ARP/wARP list
Unsubscribe from ARP/wARP list
Send message to ARP/wARP list

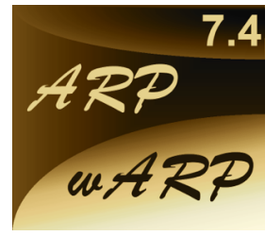
HELP

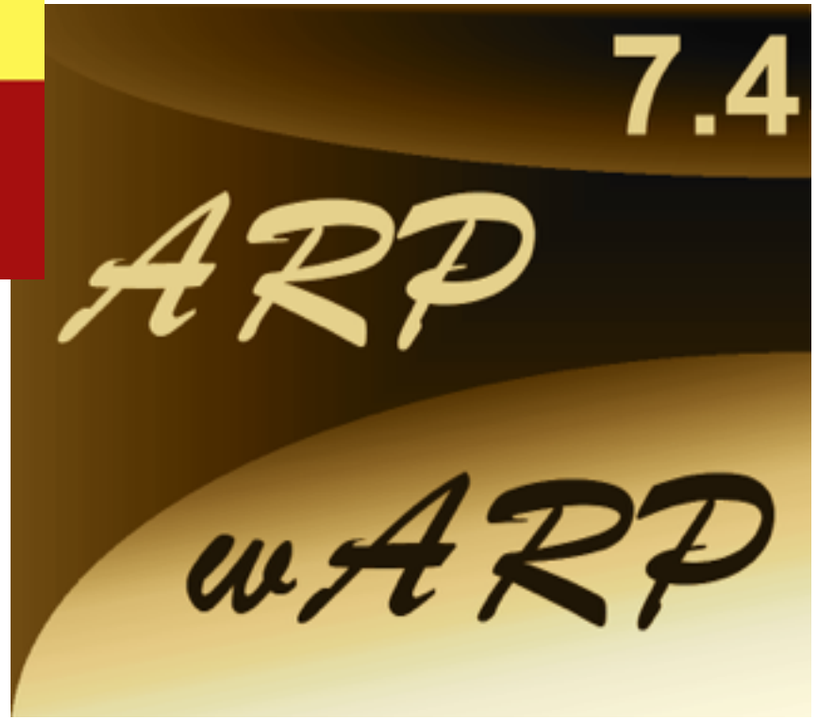
Send email to ARP/wARP developers
Frequently Asked Questions

Useful Links

Auto-Rickshaw	Automated Crystal Structure Determination Platform
ViCi	In-Silico Ligand-Based Drug Design
CCP4	Software for Macromolecular X-Ray Crystallography
Refmac	Macromolecular Refinement Program

Collaboration with CCP4





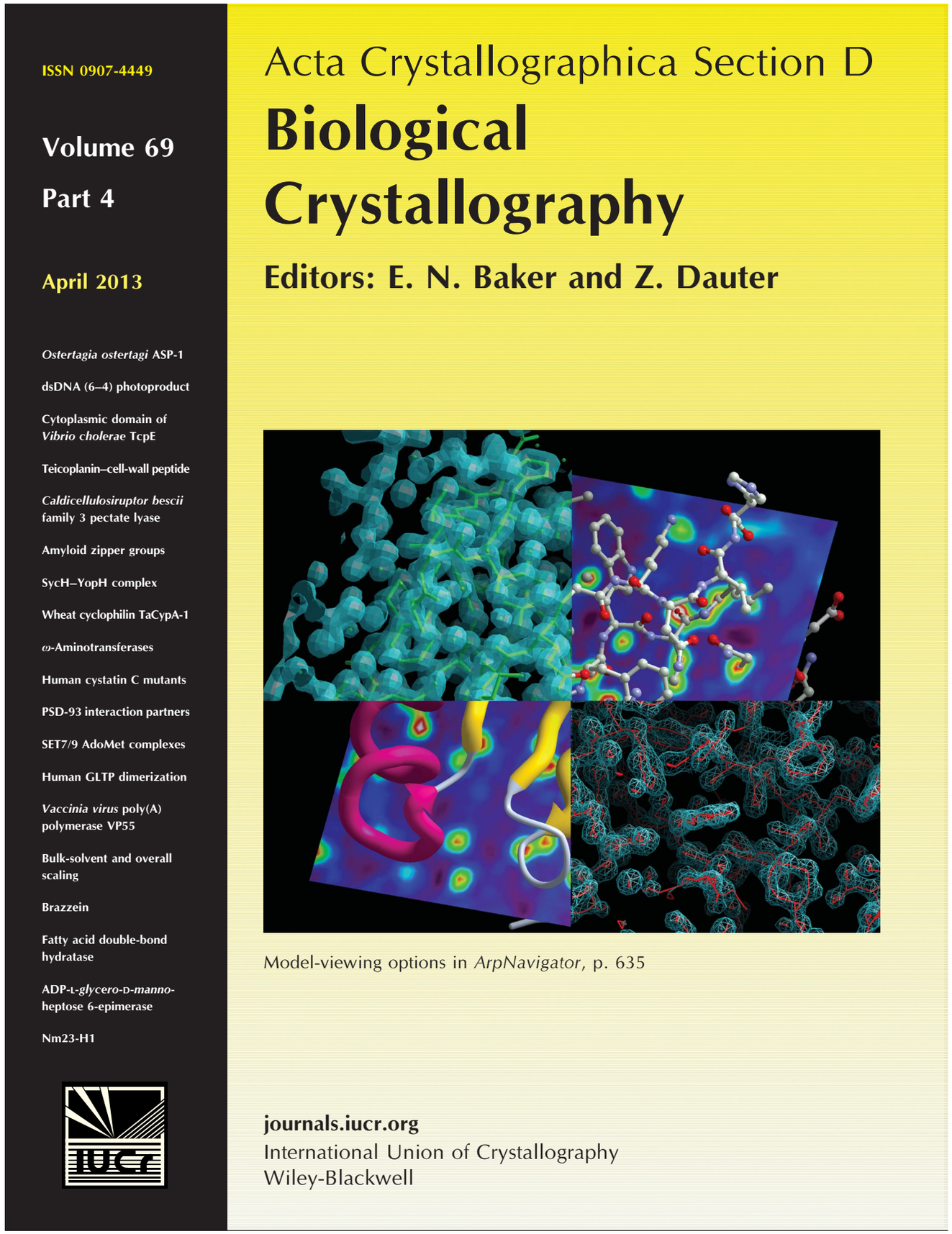
- Joint release
  - CCP4 6.4 and ARP/wARP 7.4
  - one installer, no worries ;)

ArpNavigator

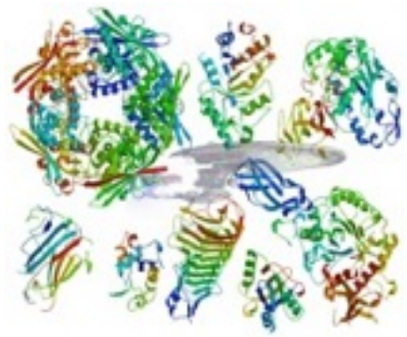
Last year, we made the cover of Acta Crystallographica D because we launched **ArpNavigator**:

- The primary intuitive interface for interacting with the automated model building processes
- Real time visualisation of model building
- Can be used by both experienced and novice users
- It controls the setup of computational tasks and displays the intermediate steps
- Allows model visualisation, map quality check and figure production

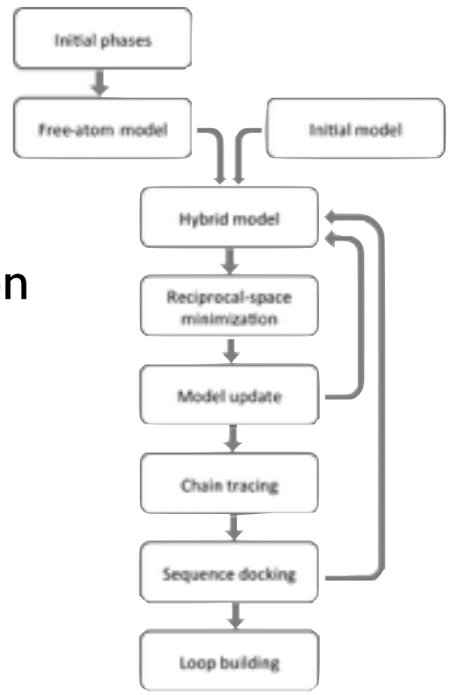
Wait for the tutorial!



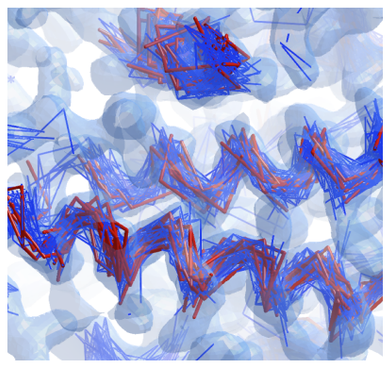
Tutorial content

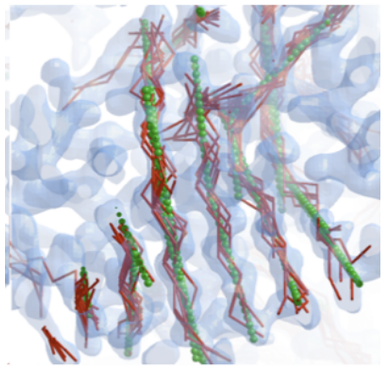


Iterative protein-model building
ARP/wARP – Perrakis *et al.*, (1999)
'Free atoms' – Agarwal and Isaacs (1985)
REFMAC5 – Murshudov et al., (1997)
Model extension by automatic NCS detection



Recognition of secondary structure:
fast approach for helices and sheets at low resolution (< 4.0 Å)





... and all this in the **ArpNavigator**

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

If you are ever in town

The EMBL

EBML Heidelberg ("Main Laboratory")
- Gene Expression
- Cell Biology
- Structural and Computational Biology
- Developmental Biology

EMBL Monterotondo ("Rome")
- Mouse Genetics

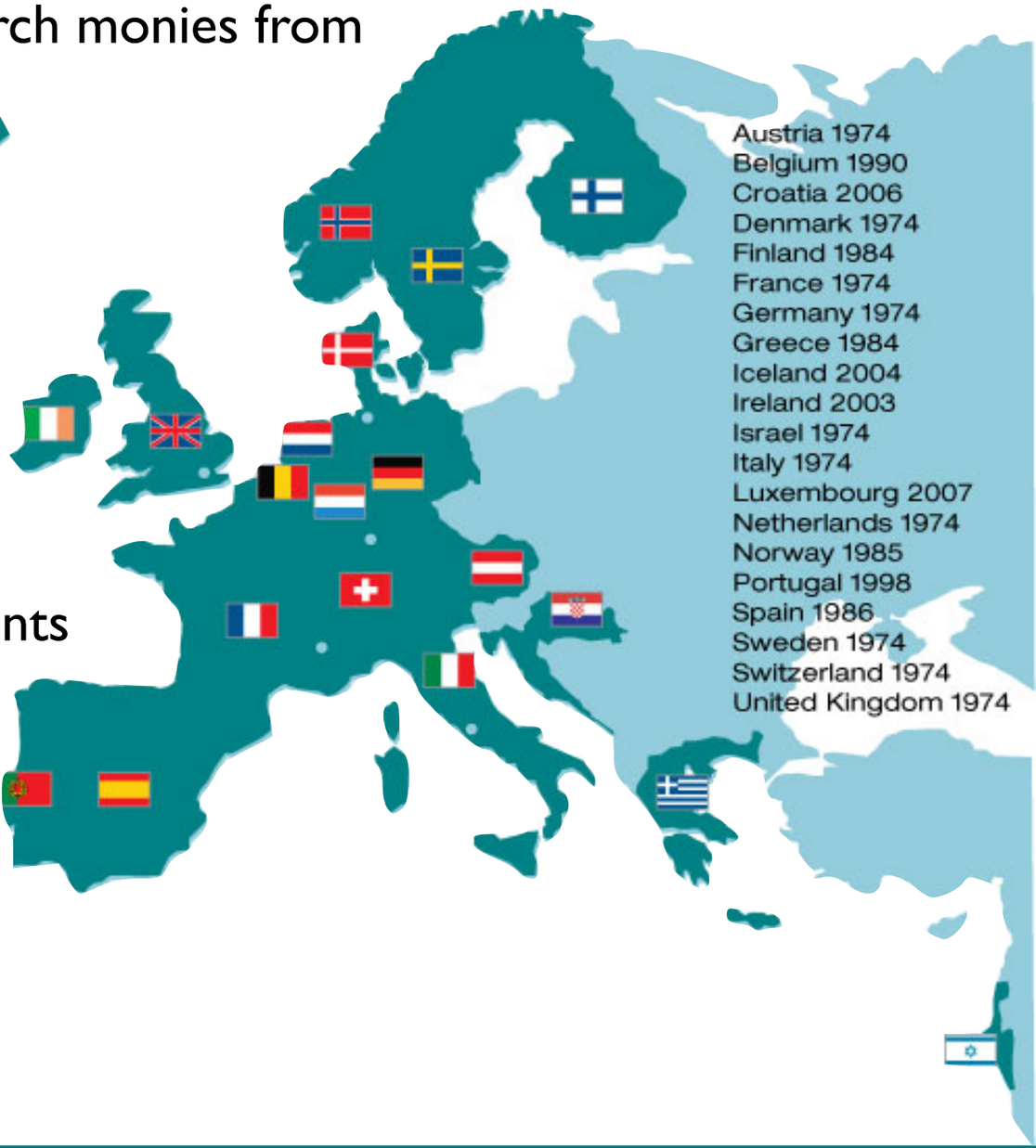
EMBL Grenoble
- Structural Biology

EMBL / EBI Hinxton ("Cambridge")
- Bioinformatics

EMBL Hamburg
- Structural Biology

The EMBL

- Established in **1974** to reflect the need for a European centre for research and training in **molecular biology**
- A **basic research** institute funded by public research monies from **20 member states**
- Over **1300 employees** from 60 nations
- Approximately **80** independent research **groups**



EMBL's mission

- **Basic Research** in Molecular Biology
- **Advanced Training** at all Levels
- **Development** of New Technologies and Instruments
- **Services** and **Facilities** for the Community
- **Technology Transfer Discoveries**

Check www.embl.de for job opportunities