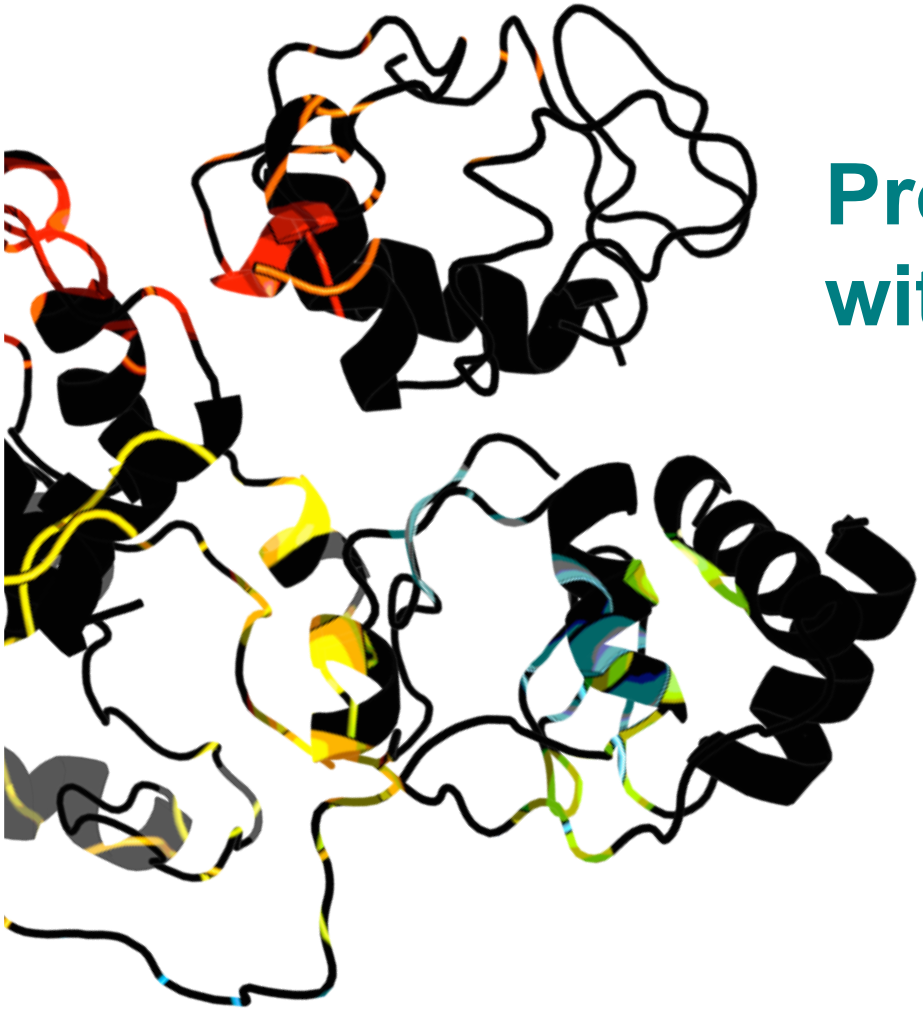
Grzegorz Chojnowski
Lamzin Group

Automated Model Building with ARP/wARP

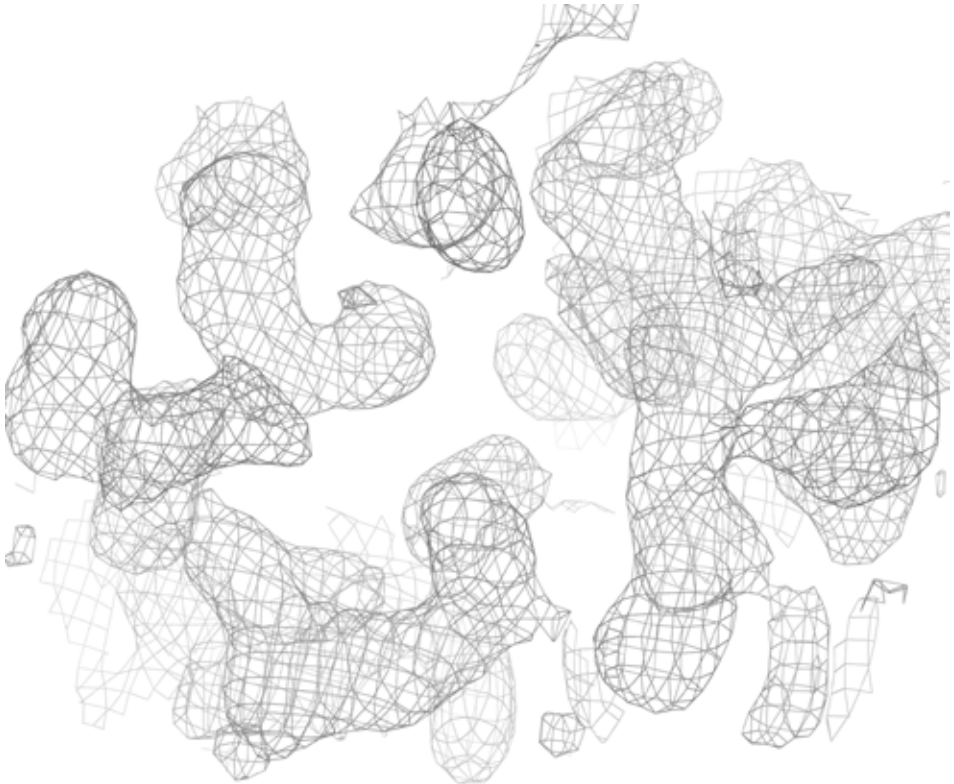
EMBL



Protein Model Building with ARP/wARP

EMBL

Electron density map interpretation



Electron density map interpretation

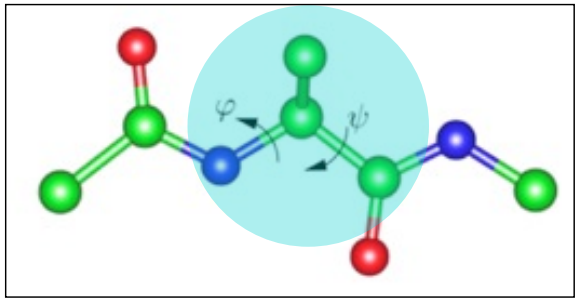
Photo by artwolfe.com

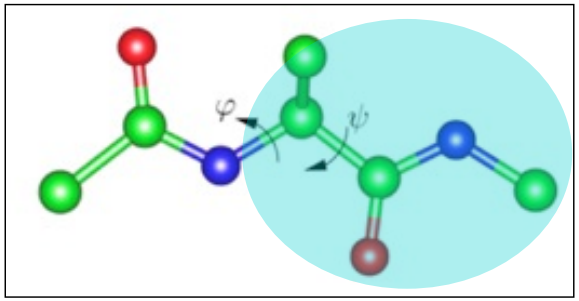
Electron density map interpretation

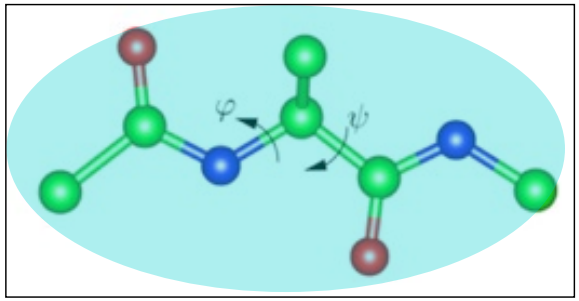
Zebra model by Maria Chojnowska

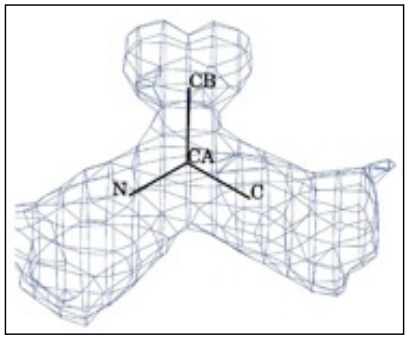
Methods for automated model building

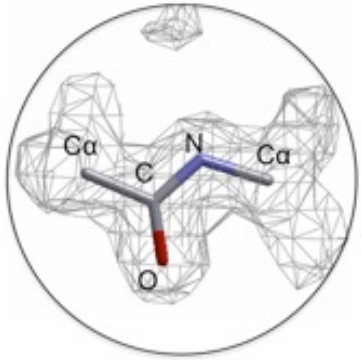
Most current methods search for different patterns that are then used to build and extend a main chain:

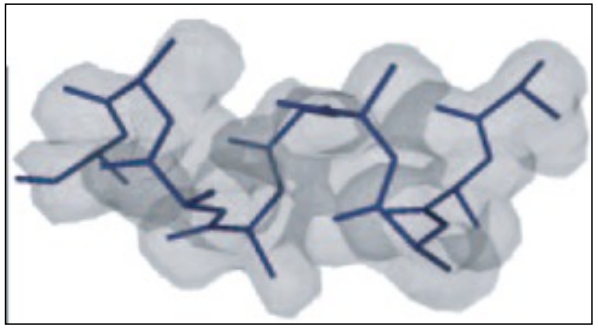






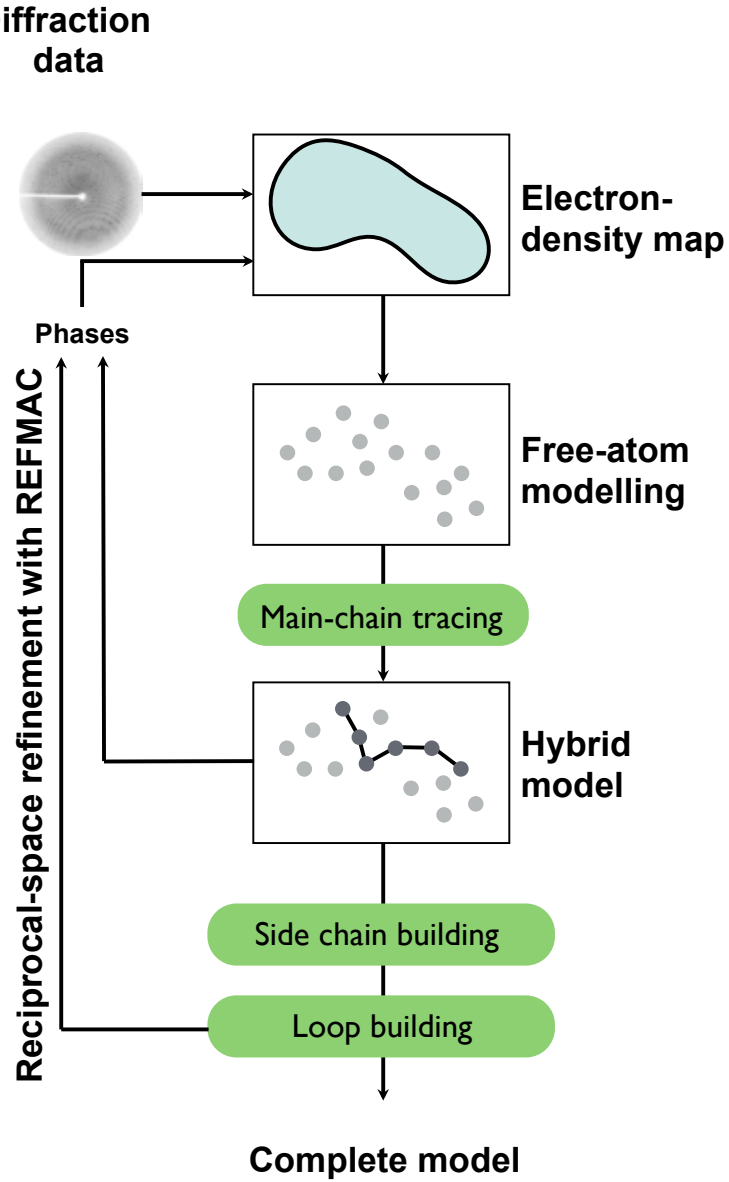




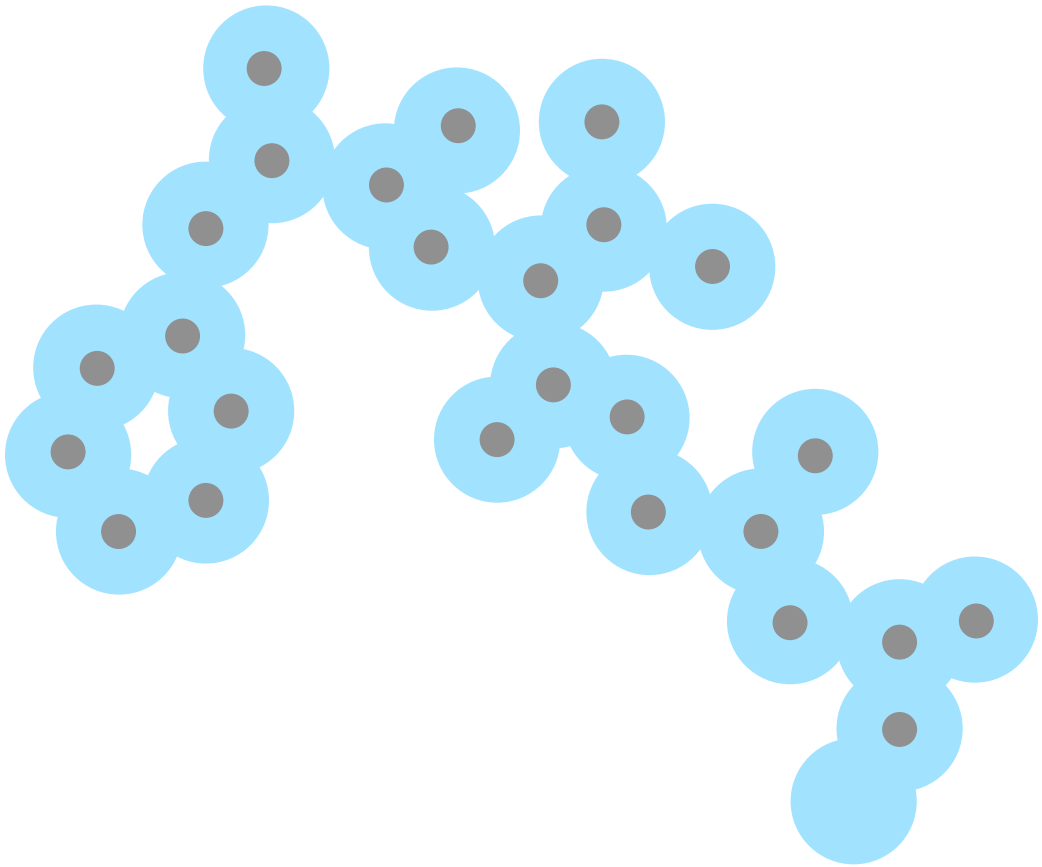


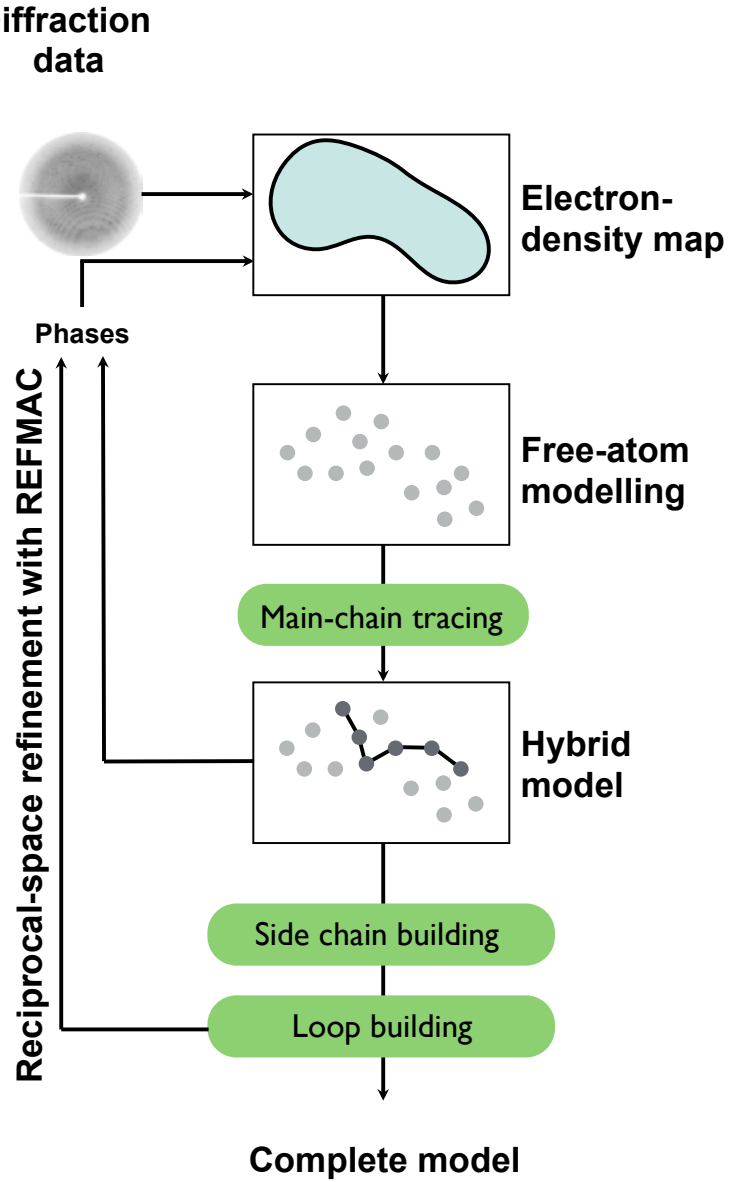
Buccaneer	ARP/wARP	phenix.autobuild
Cα	(Di-) Peptides	Fragments

ARP/wARP workflow

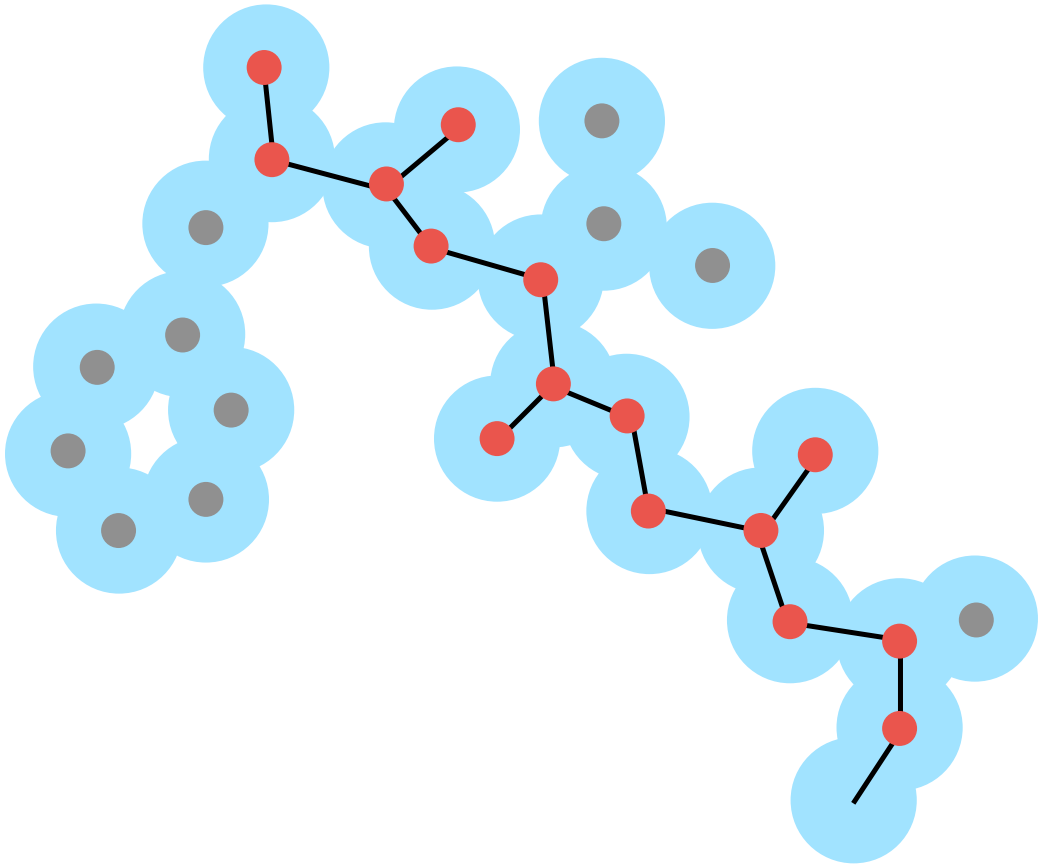


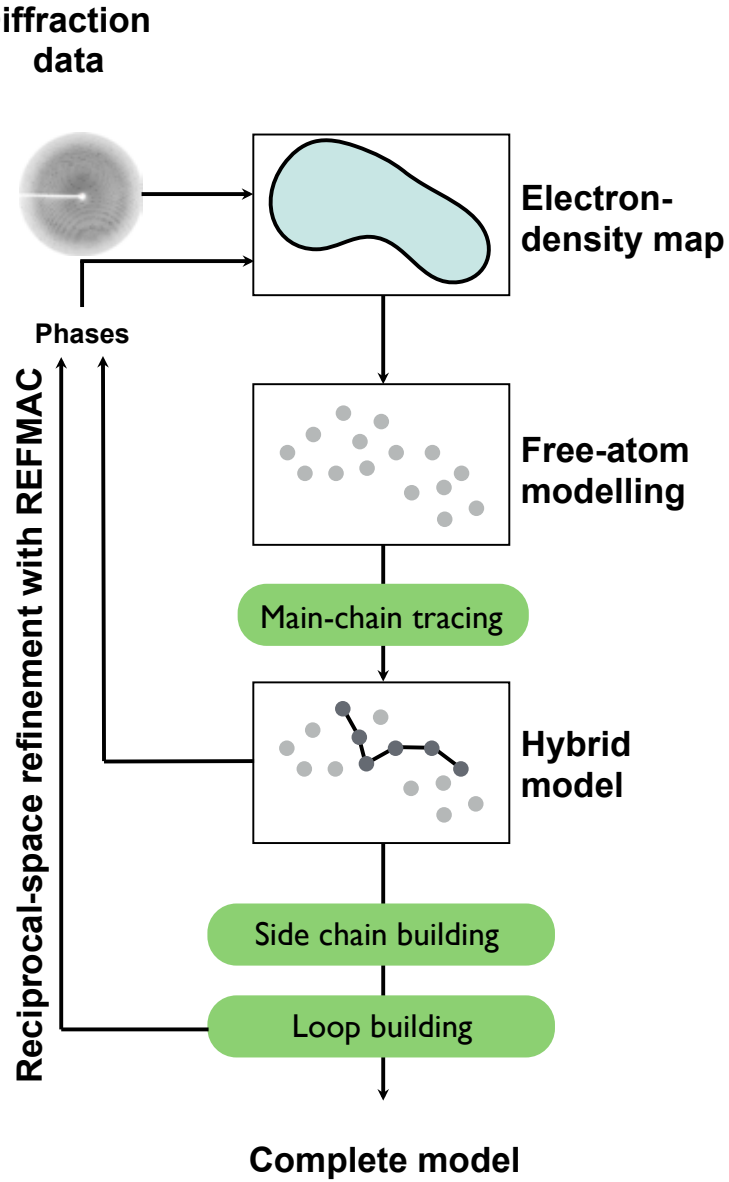
ARP/wARP workflow



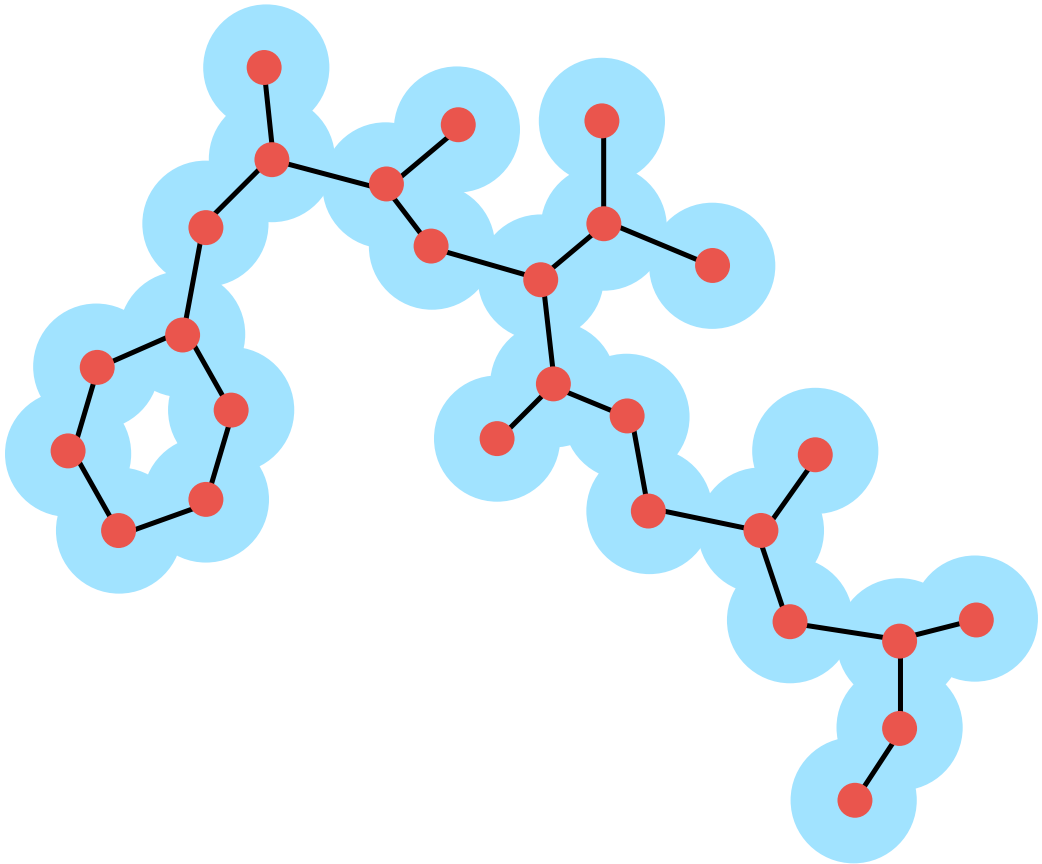


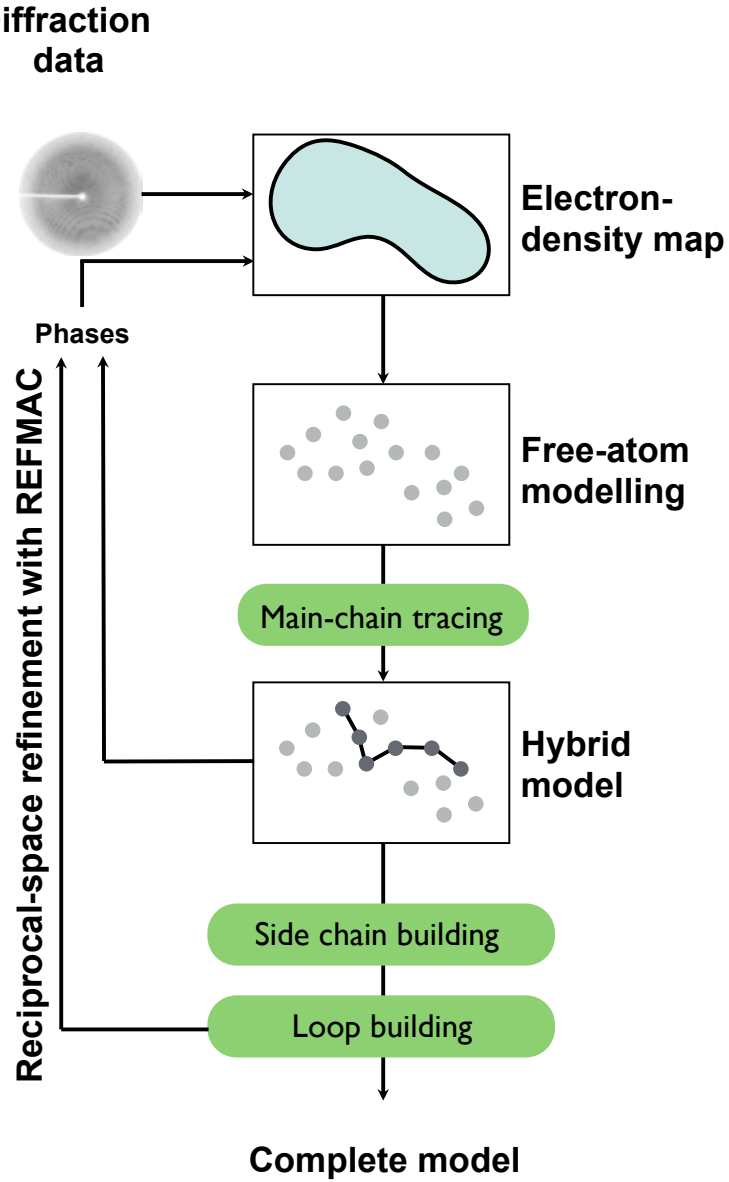
ARP/wARP workflow



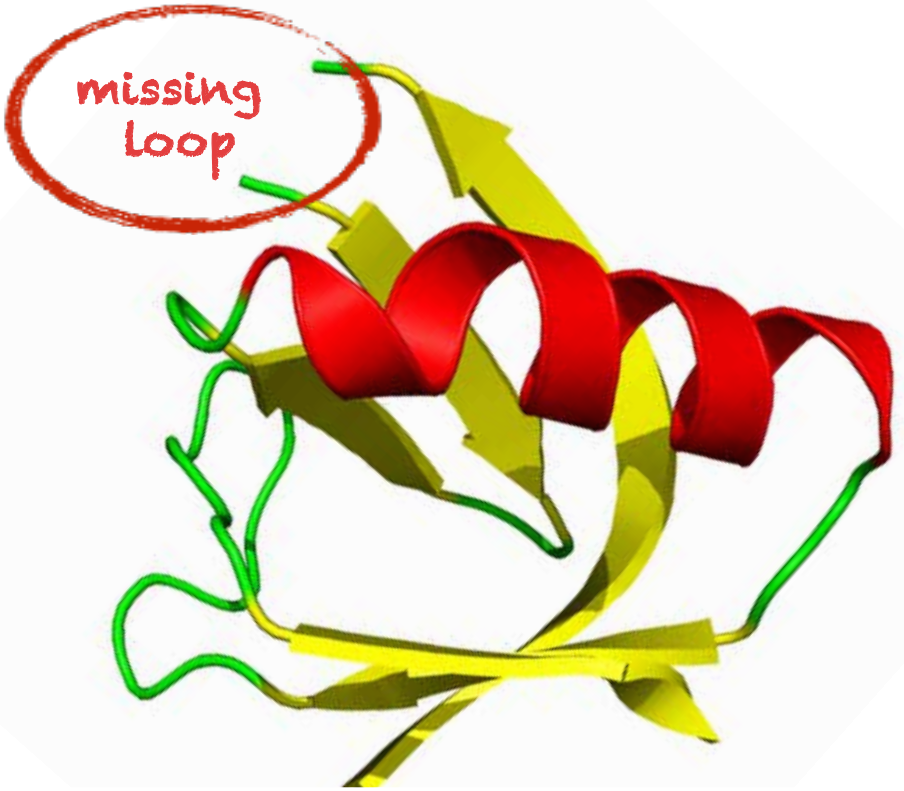


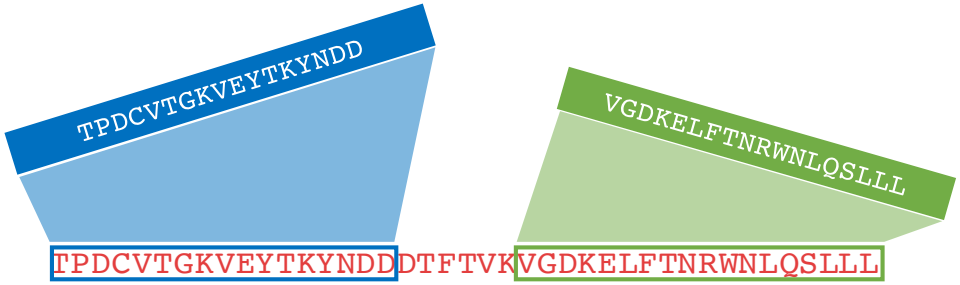
ARP/wARP workflow

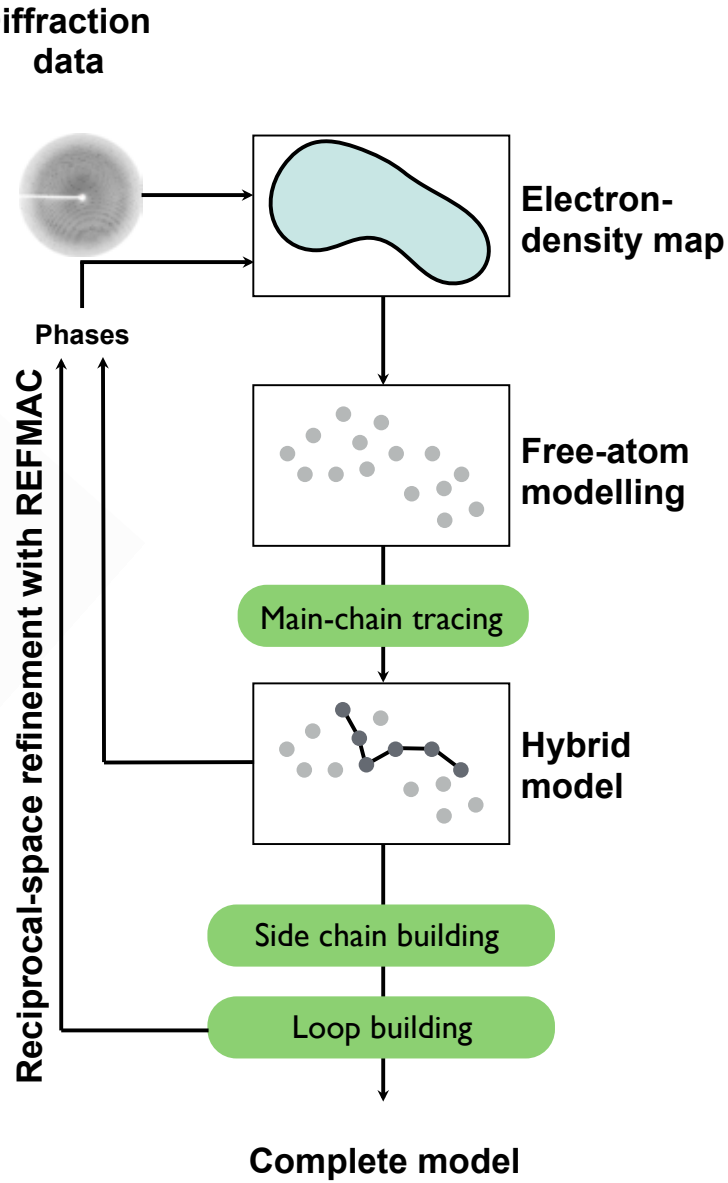




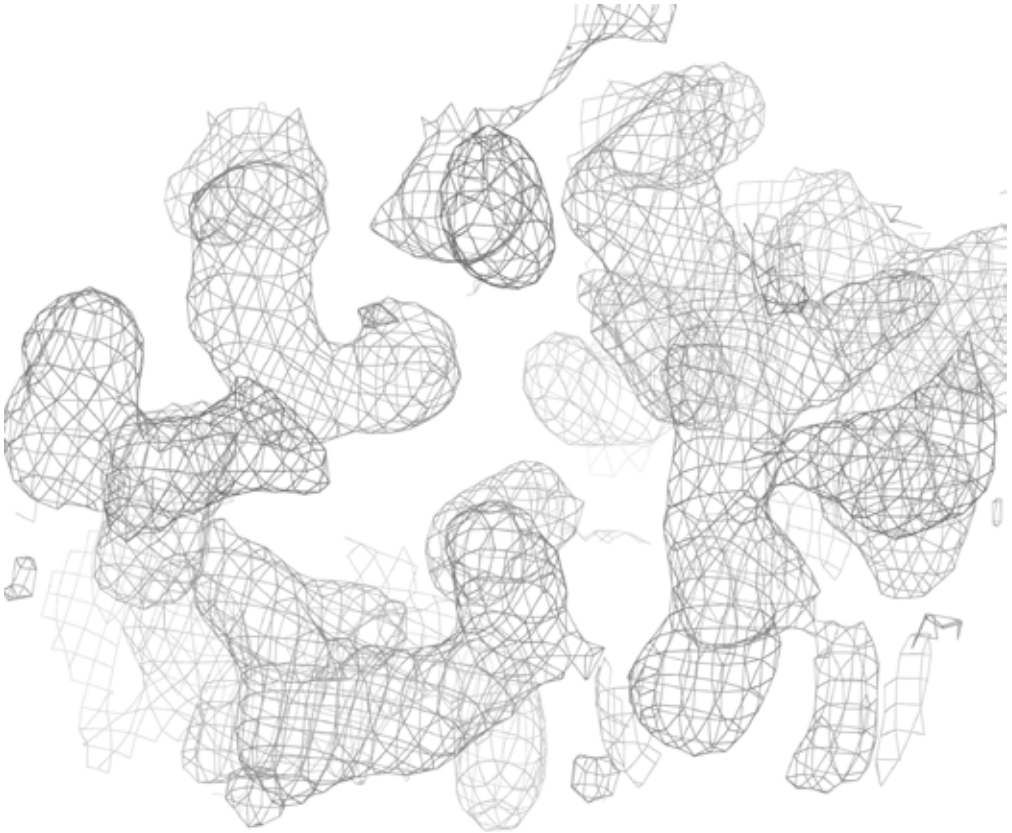
ARP/wARP workflow

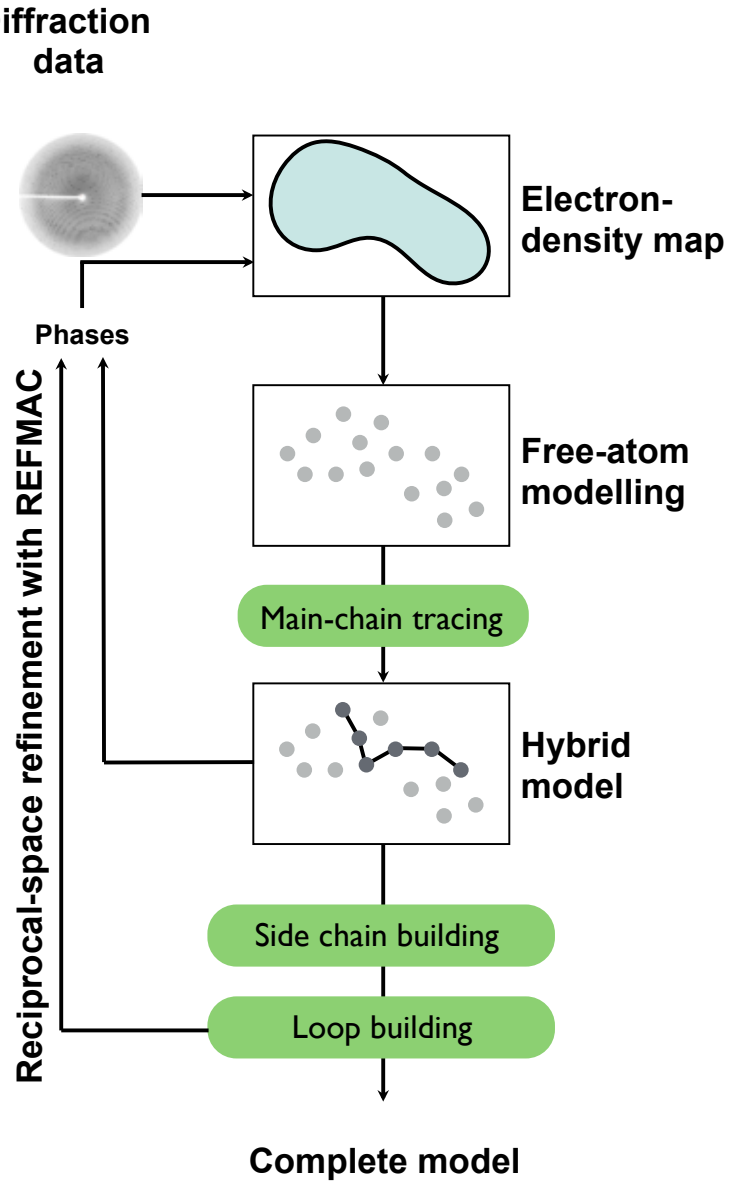




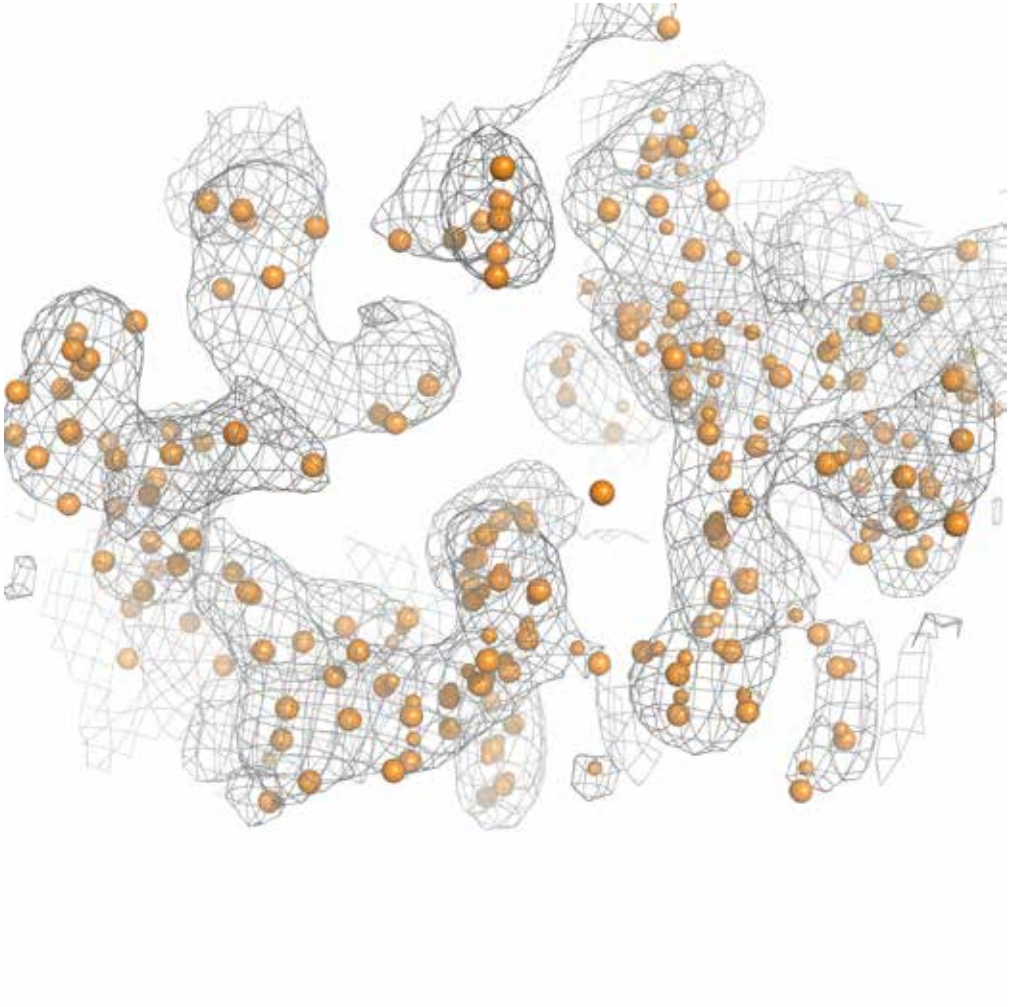


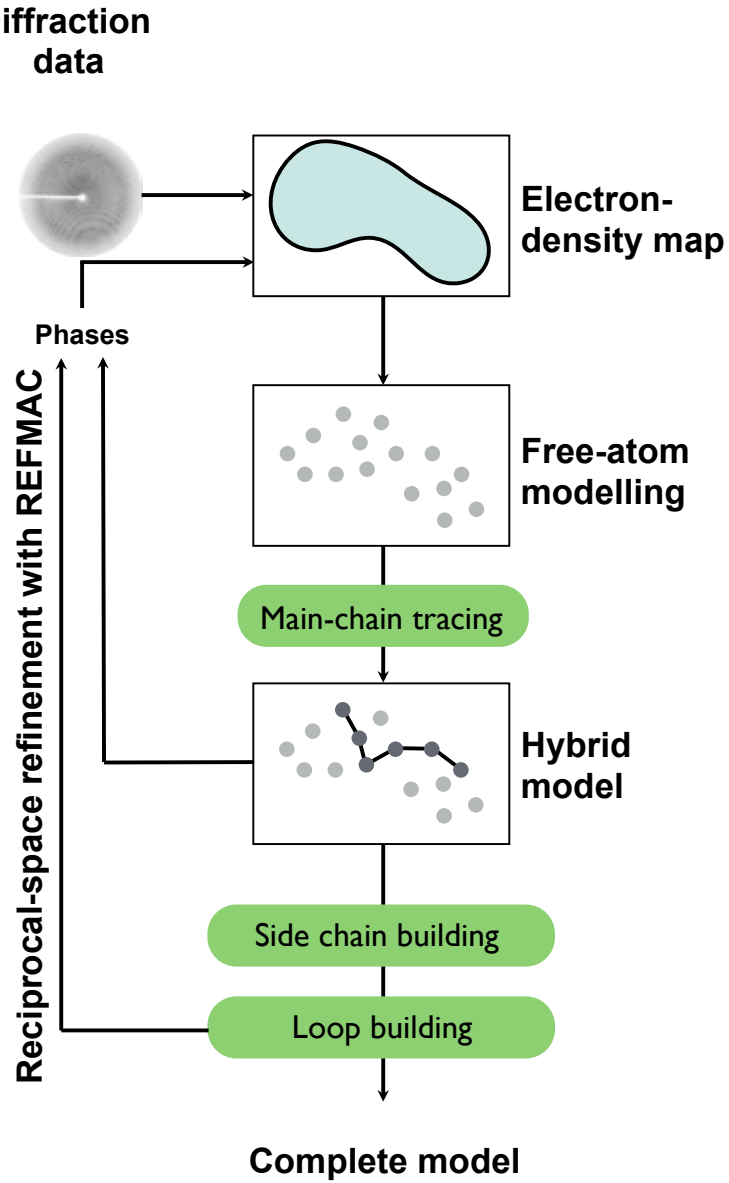
ARP/wARP workflow



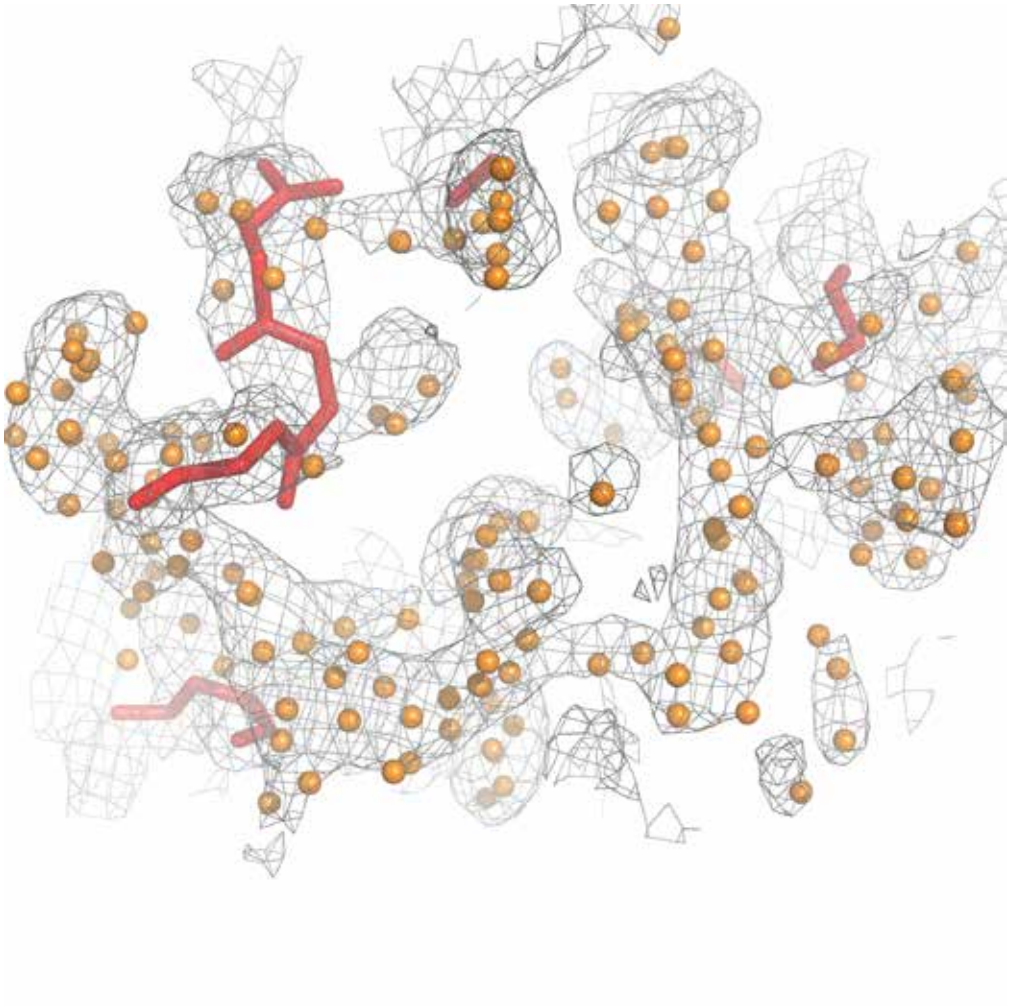


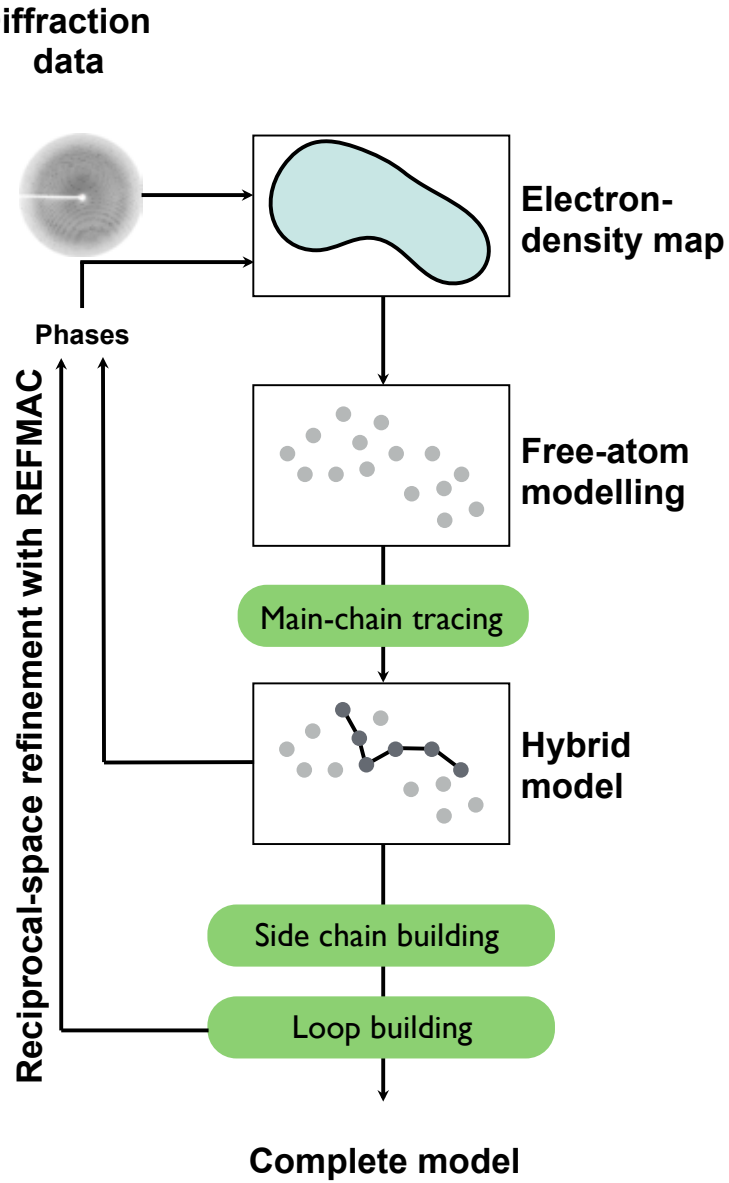
ARP/wARP workflow



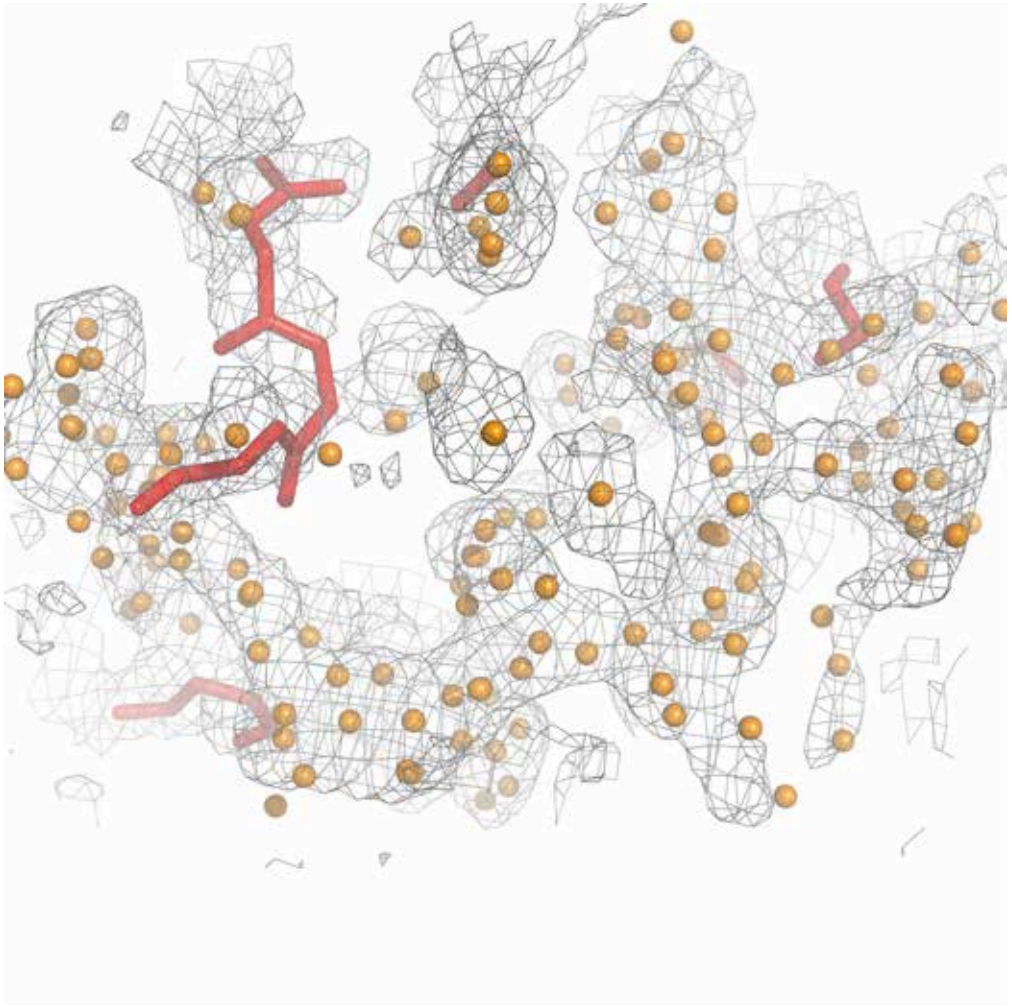


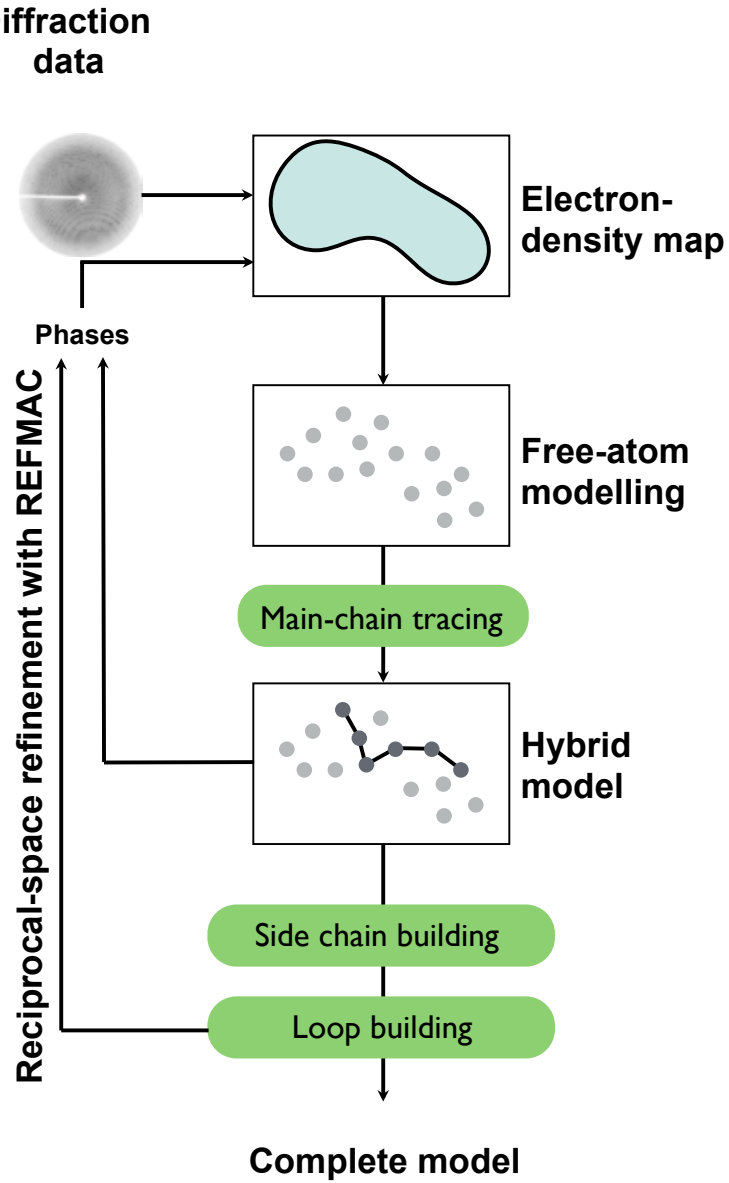
ARP/wARP workflow



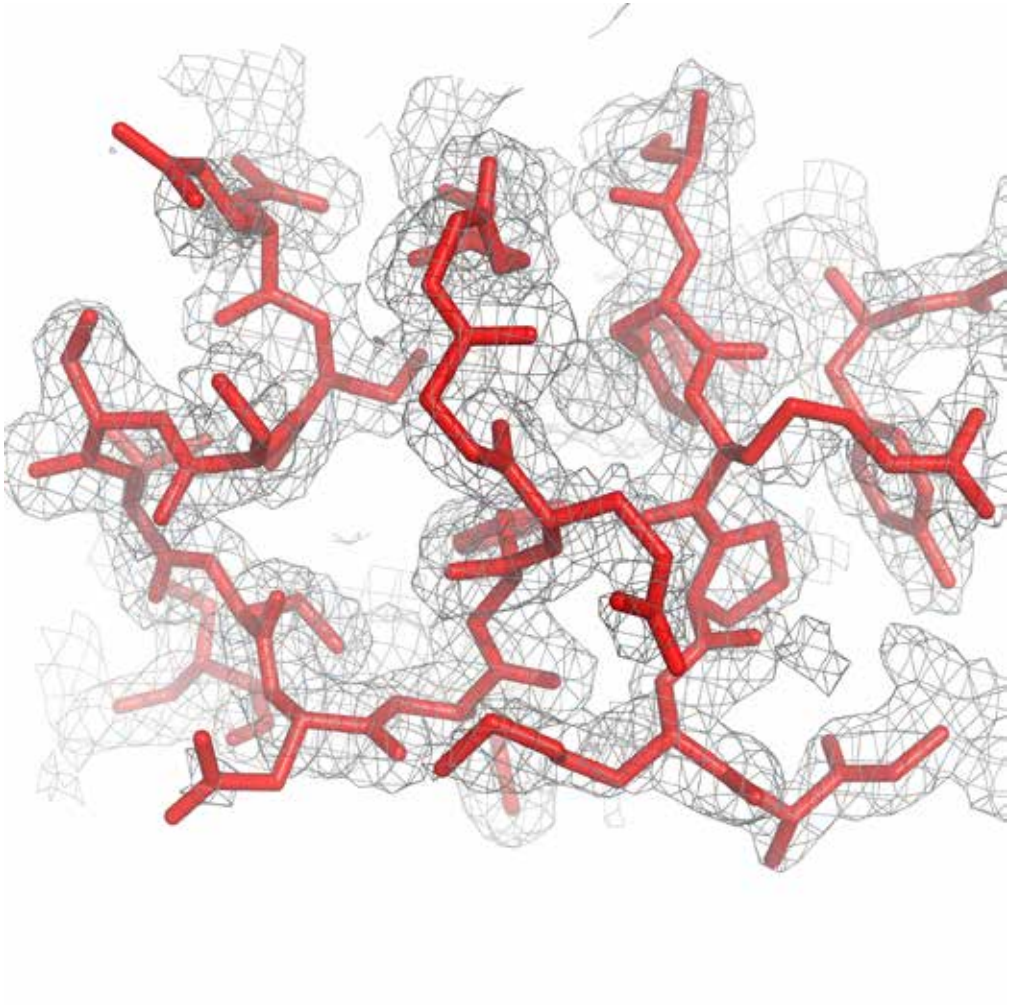


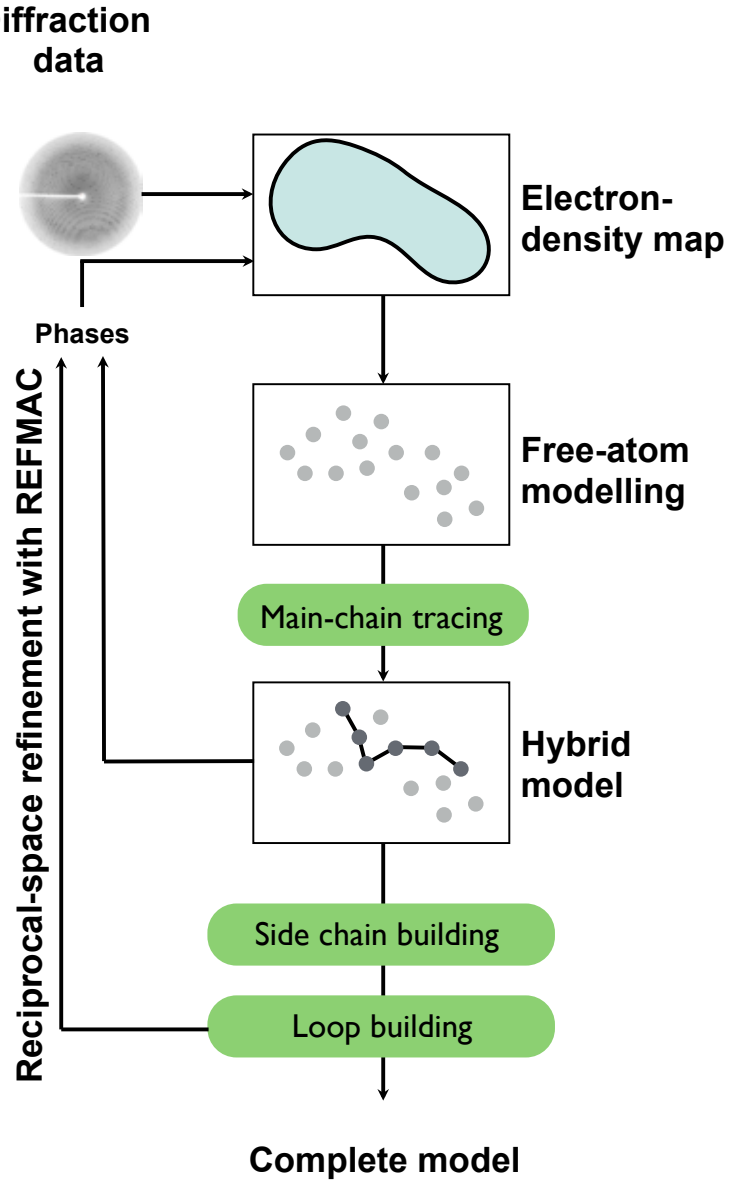
ARP/wARP workflow



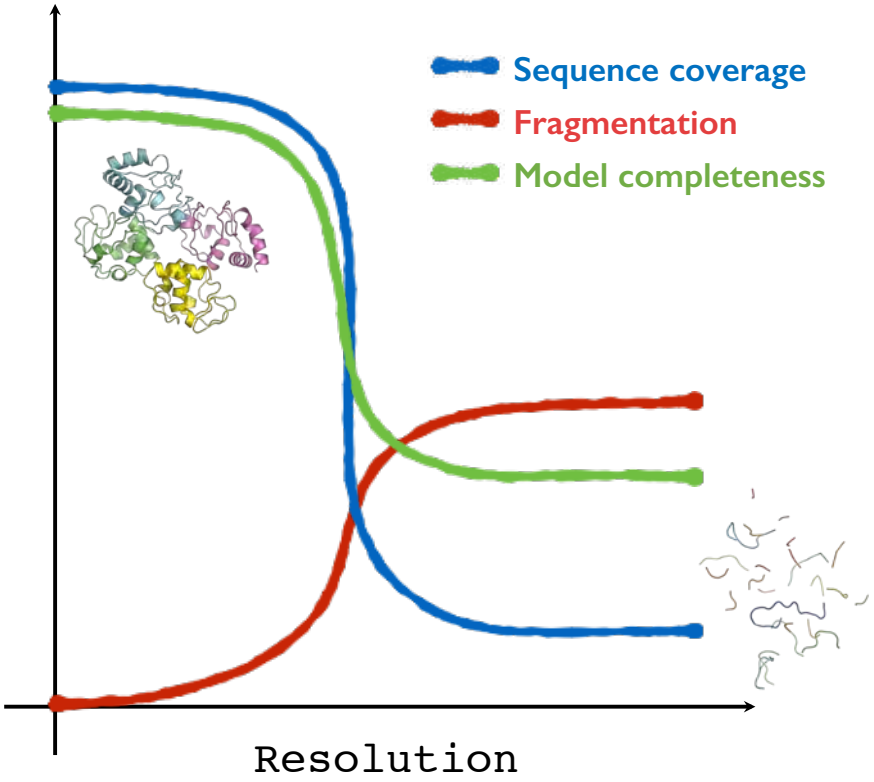


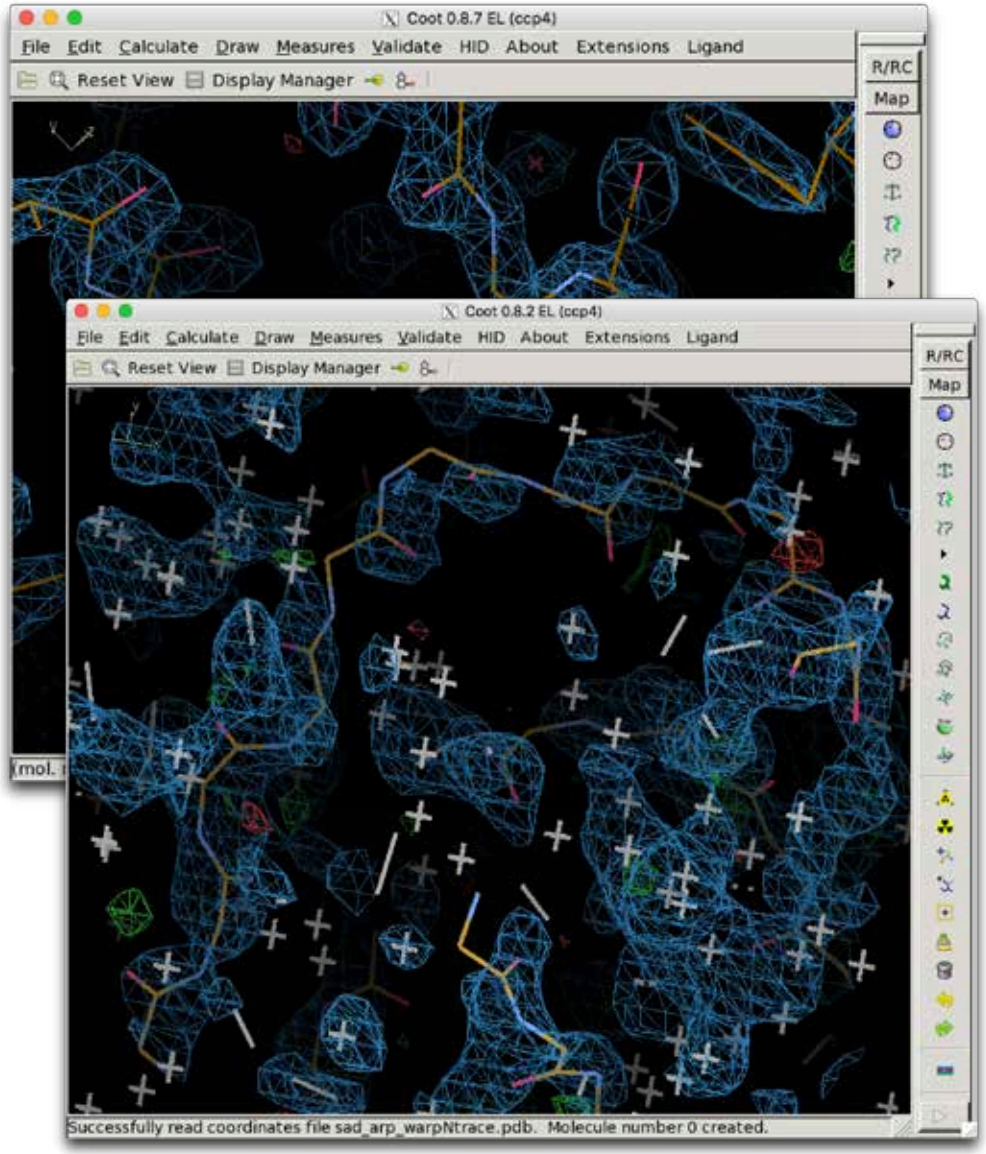
ARP/wARP workflow



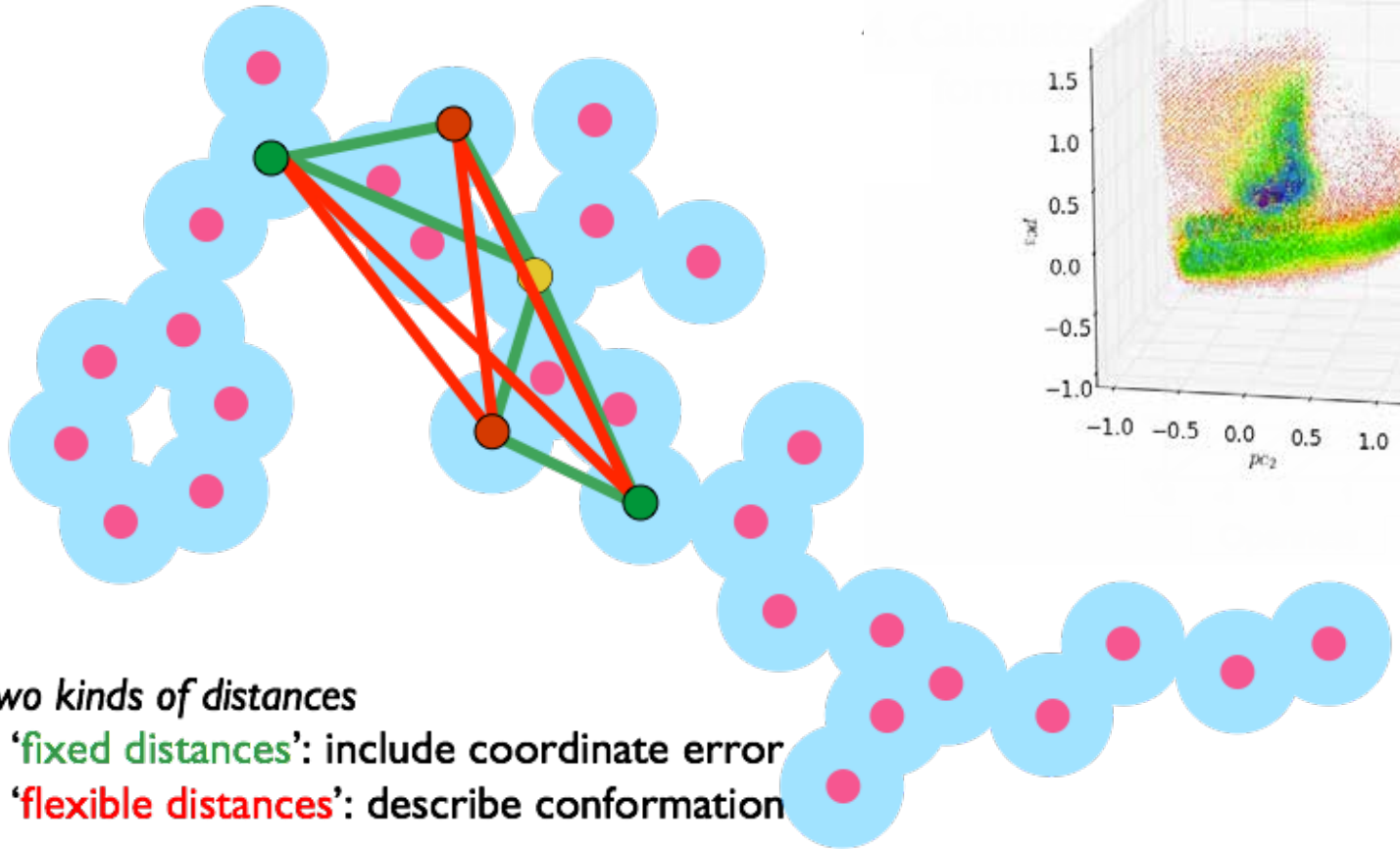


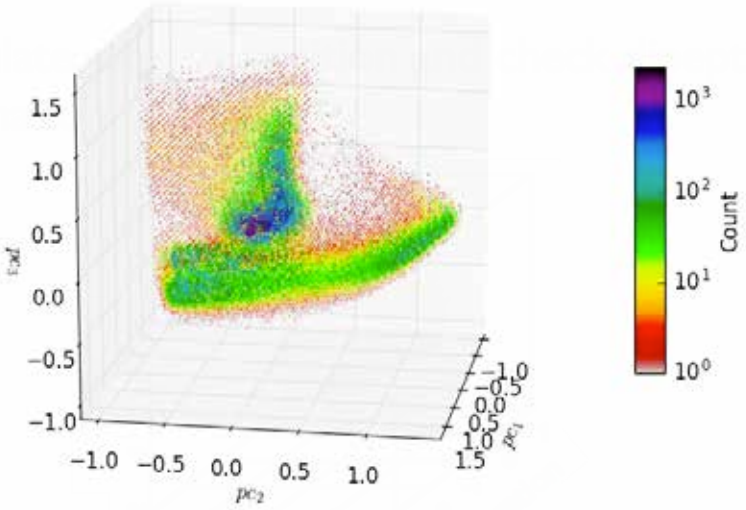
Model building at low resolution





DipCheck: main-chain geometry validation





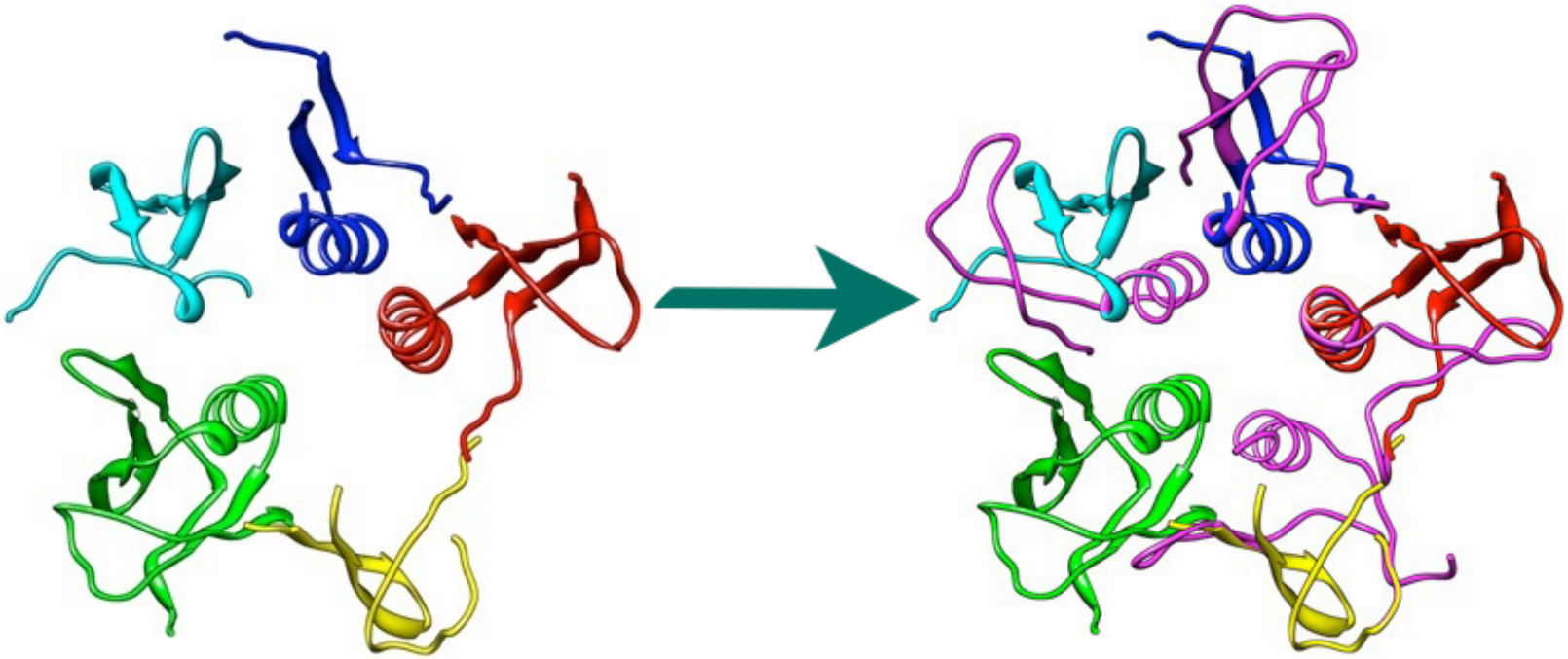
Two kinds of distances

- 'fixed distances': include coordinate error
- 'flexible distances': describe conformation

Pereira, J., & Lamzin, 2017

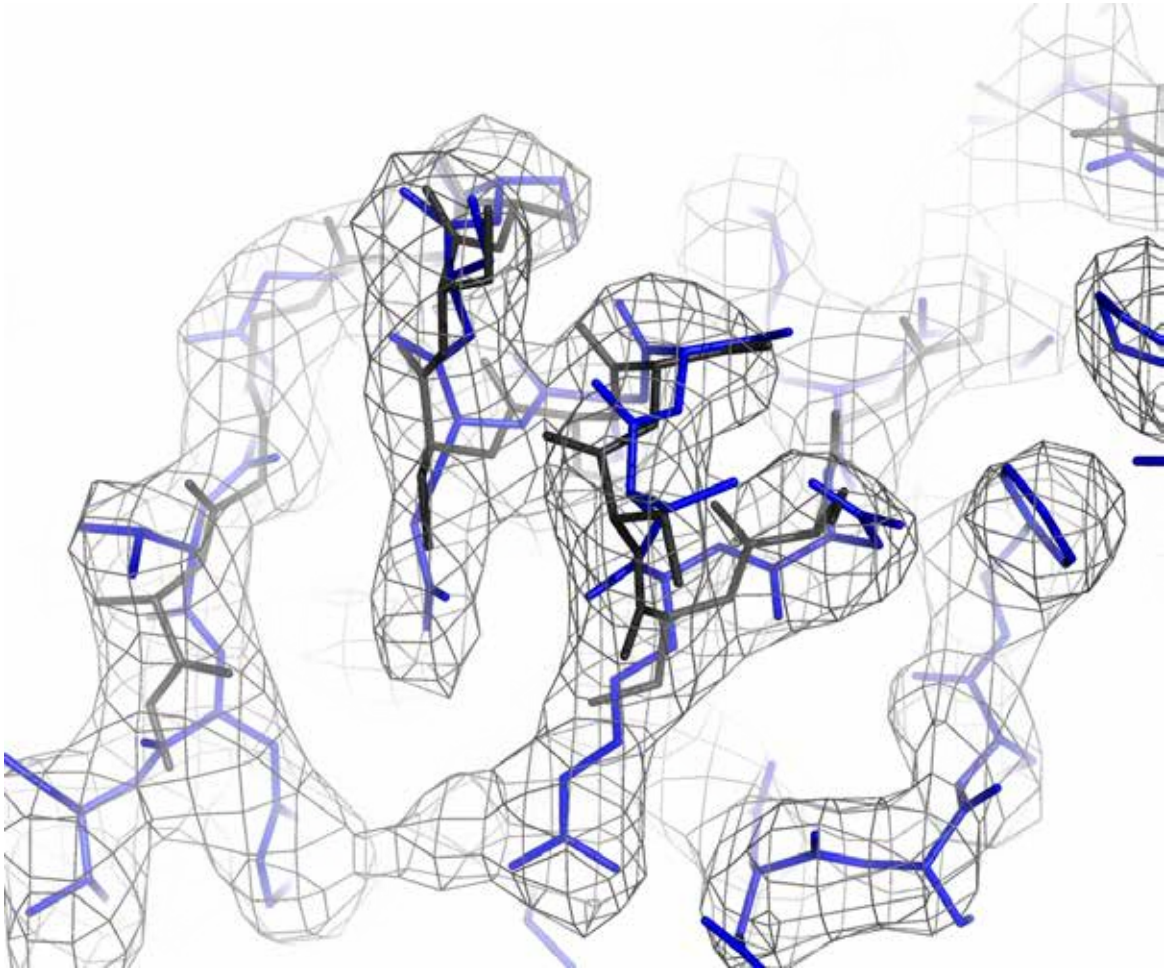
NCS Detection and Extension

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.
We can use this information to extended fragmented models!



Wieggels, T. et al. 2015

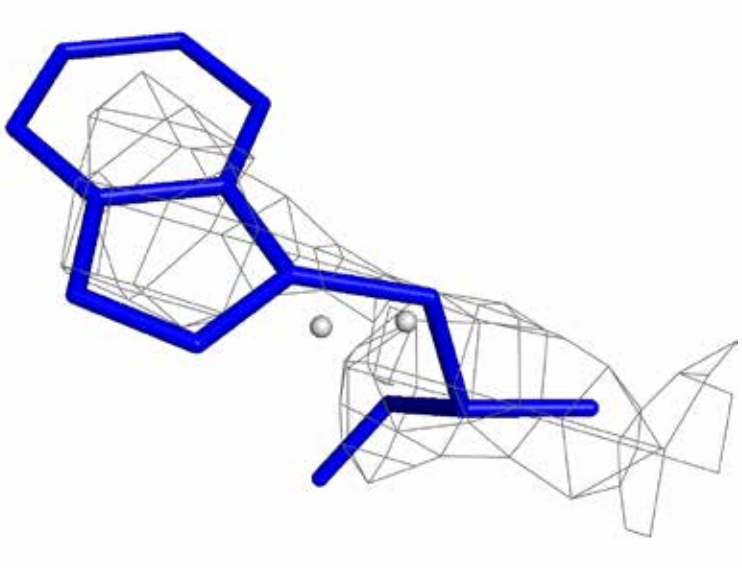
New loop building algorithm

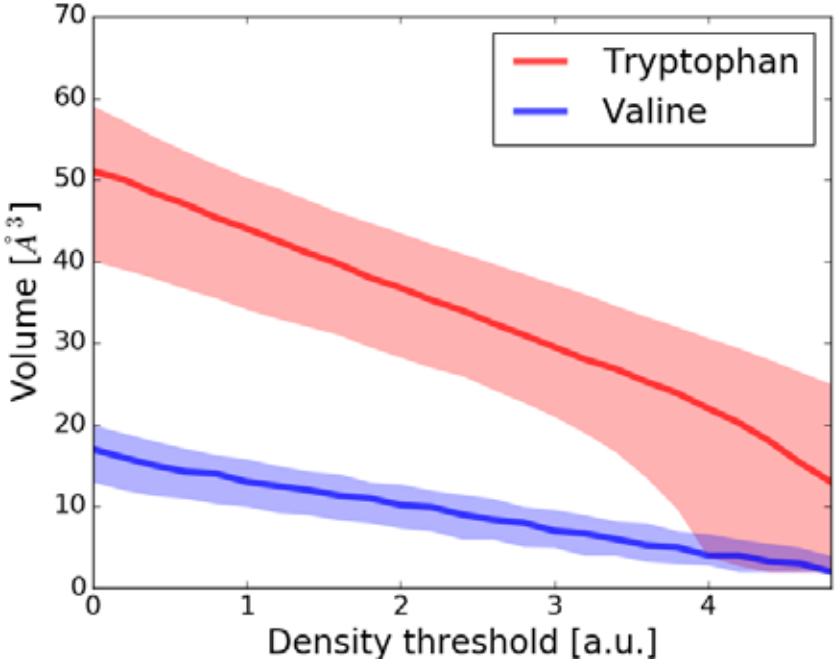


ARP/wARP model
Deposited coordinates

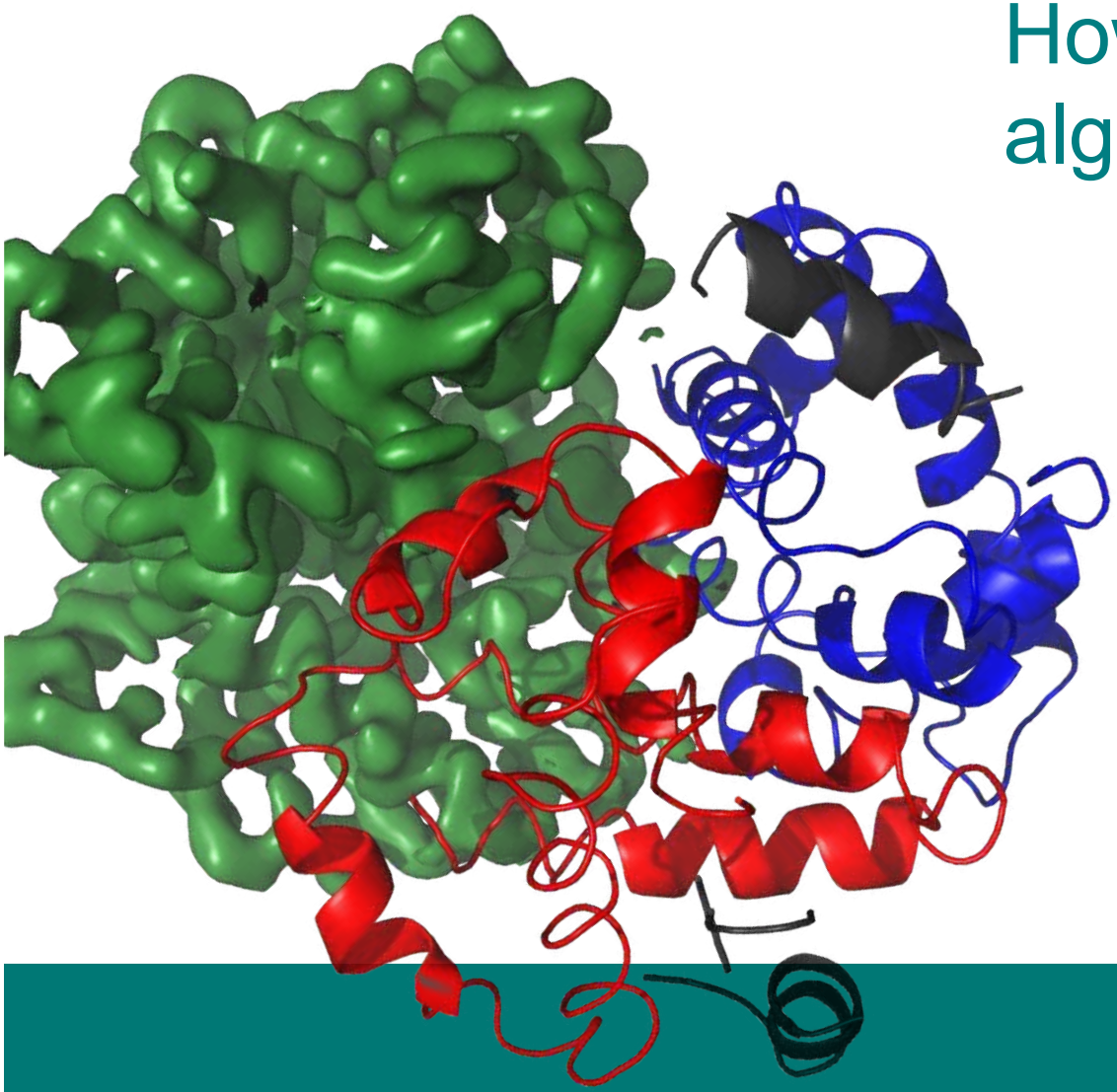
The algorithm merges close fragments
and can correct reversed ones

New side-chain building algorithm





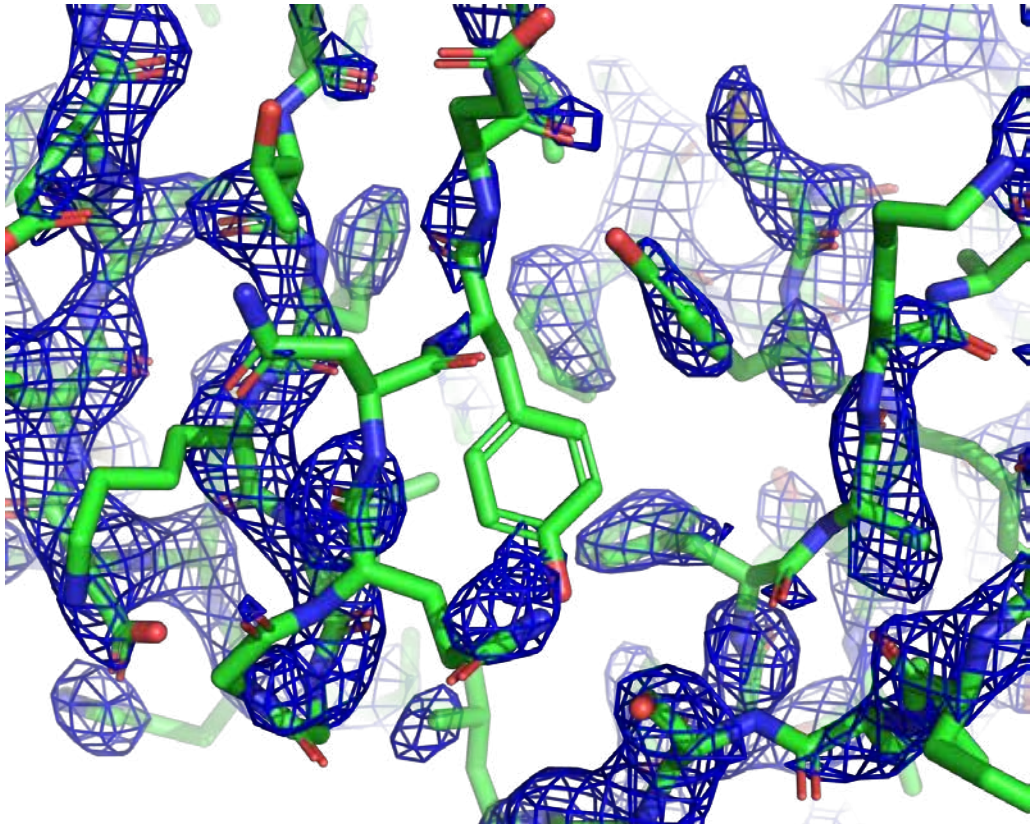
Side-chain electron-density volume changes correlate with residue type.



How do the new algorithms perform?

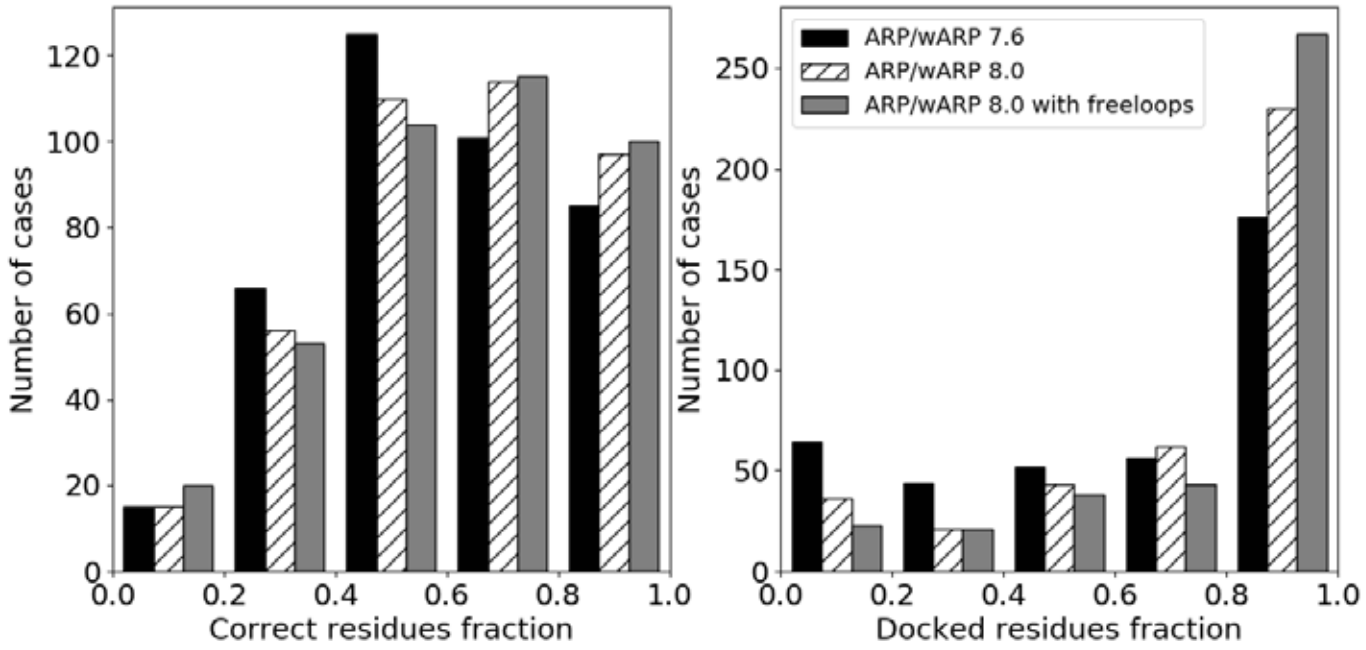
EMBL

Updated ARP/wARP performance



400 ARP/wARP model building simulations starting from maps at 2-3 Å resolution with 40 deg phase error

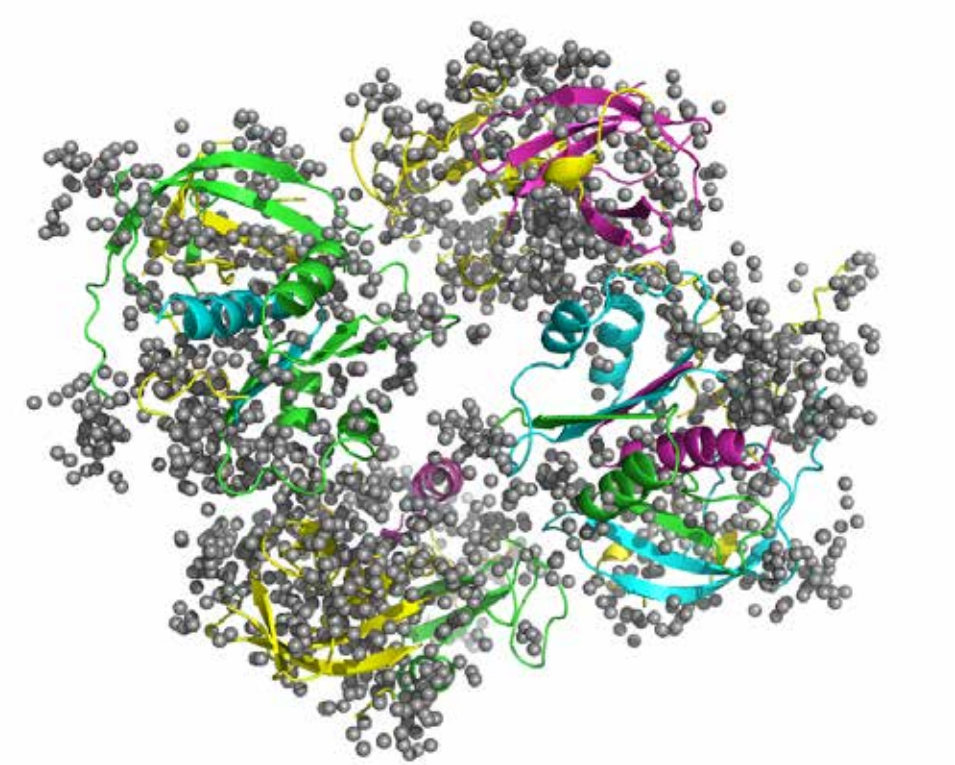
Updated ARP/wARP performance



400 ARP/wARP model building simulations starting from maps at 2-3 Å resolution with 40 deg phase error

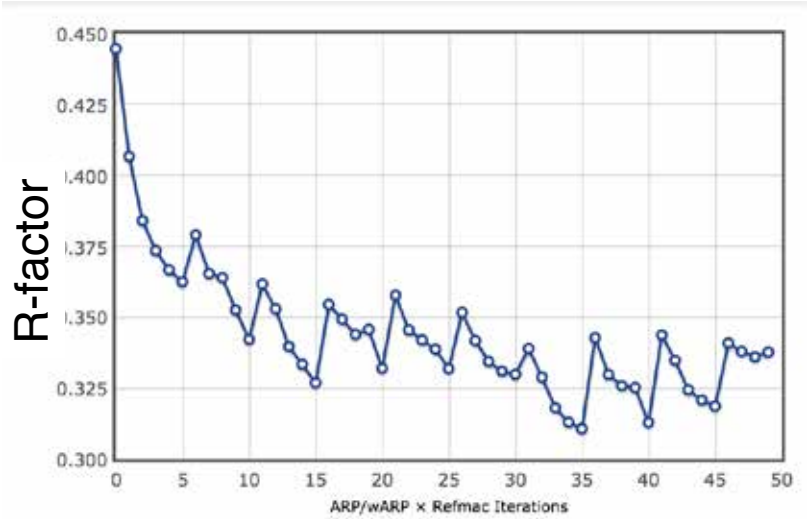
Updated ARP/wARP performance

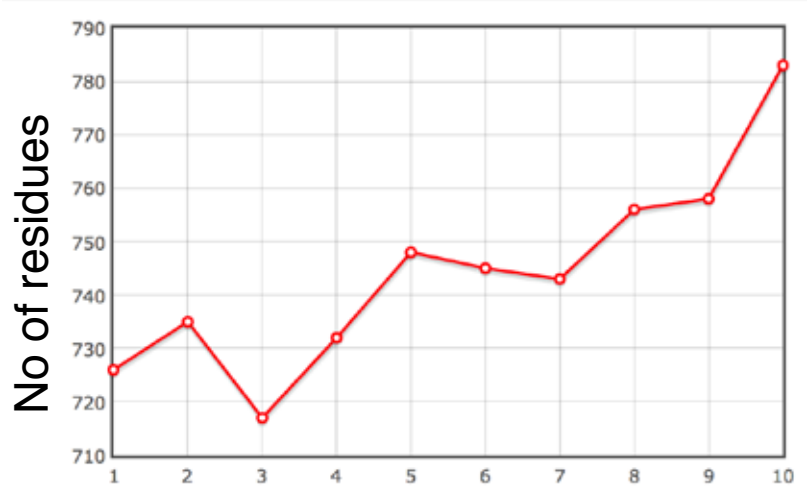
User case: ribonuclease
at 2.5Å resolution



ARP/wARP 7.6

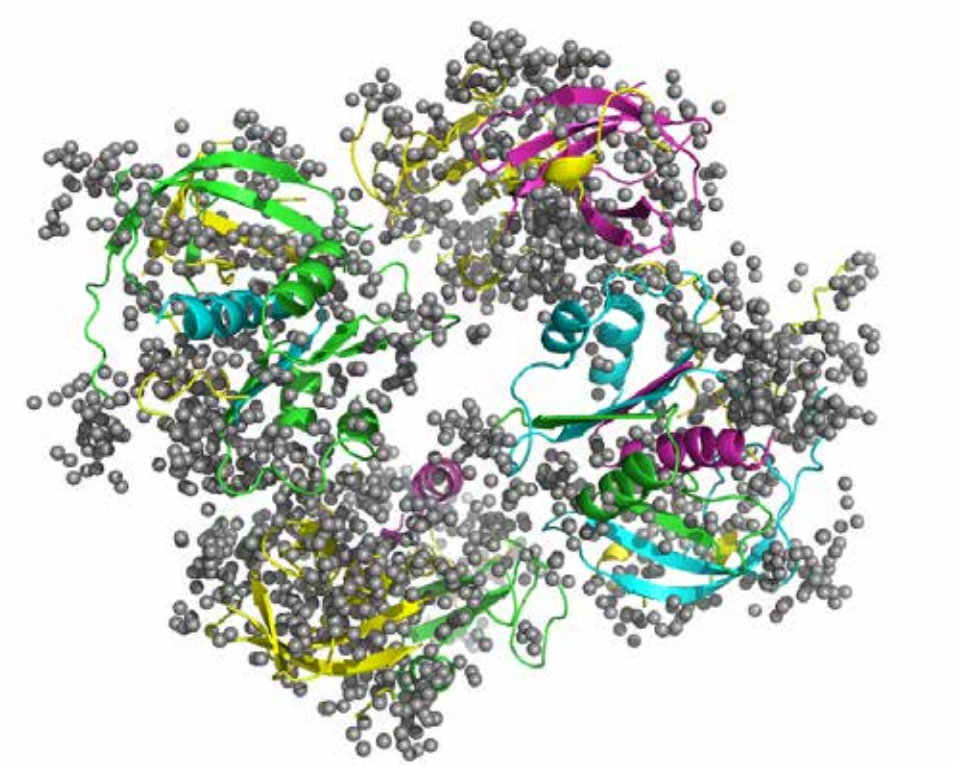
784/1300 residues built
444 docked to the sequence
R 0.34





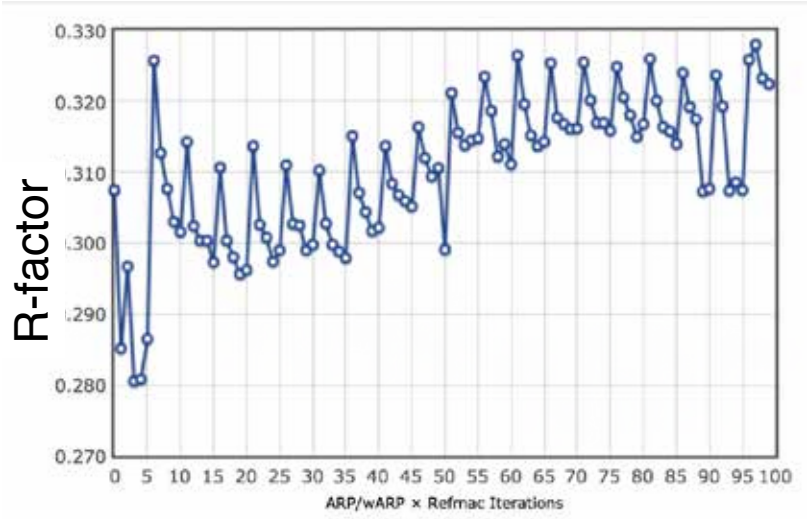
Updated ARP/wARP performance

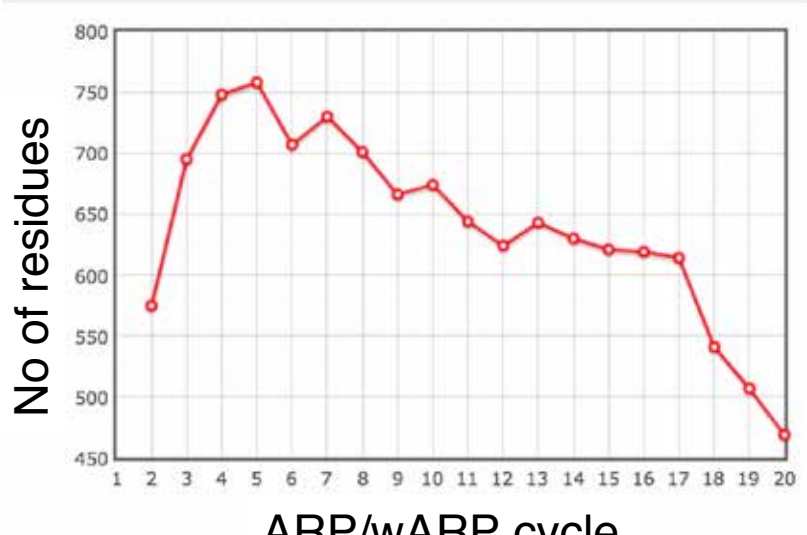
User case: ribonuclease
at 2.5Å resolution



ARP/wARP 7.6

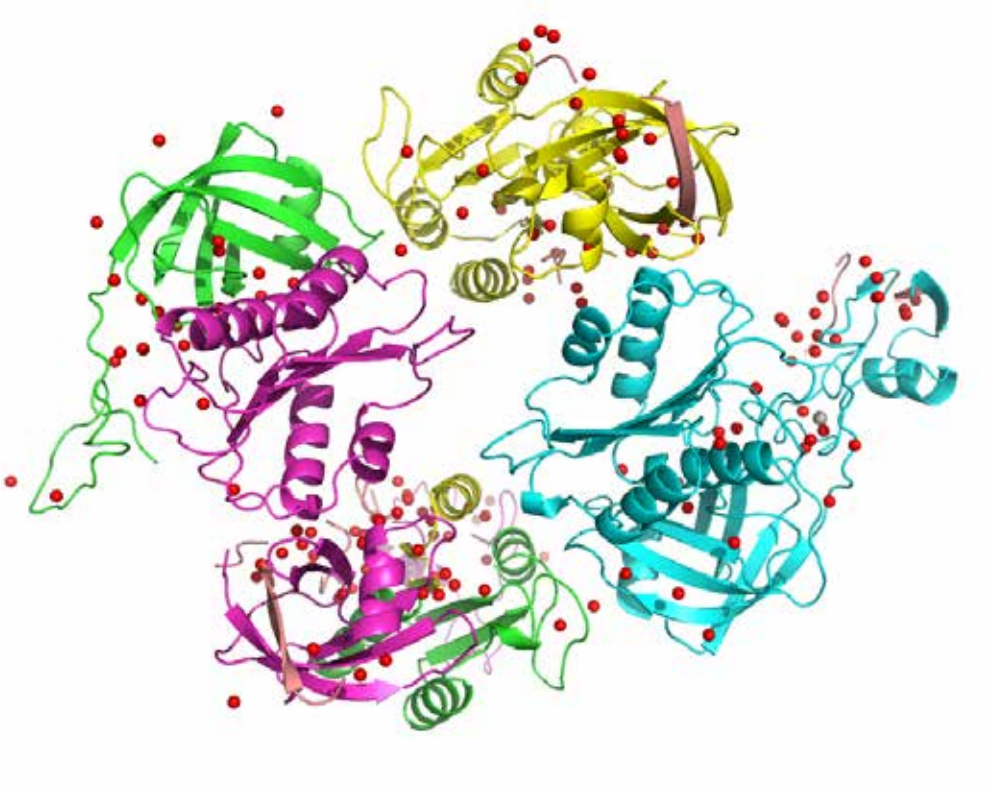
784/1300 residues built
444 docked to the sequence
R 0.34





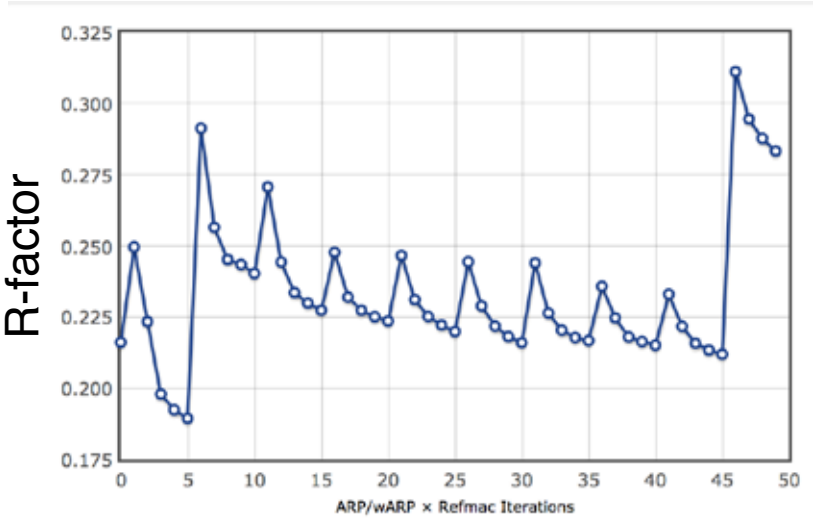
Updated ARP/wARP performance

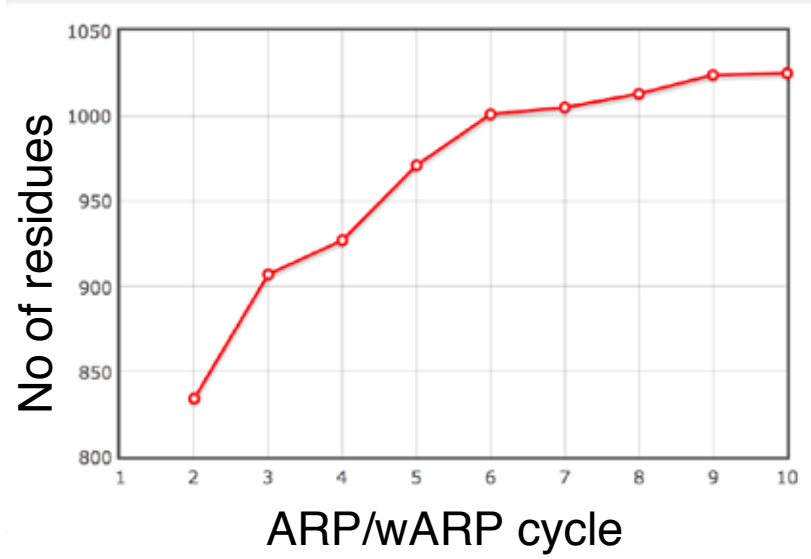
User case: ribonuclease
at 2.5Å resolution

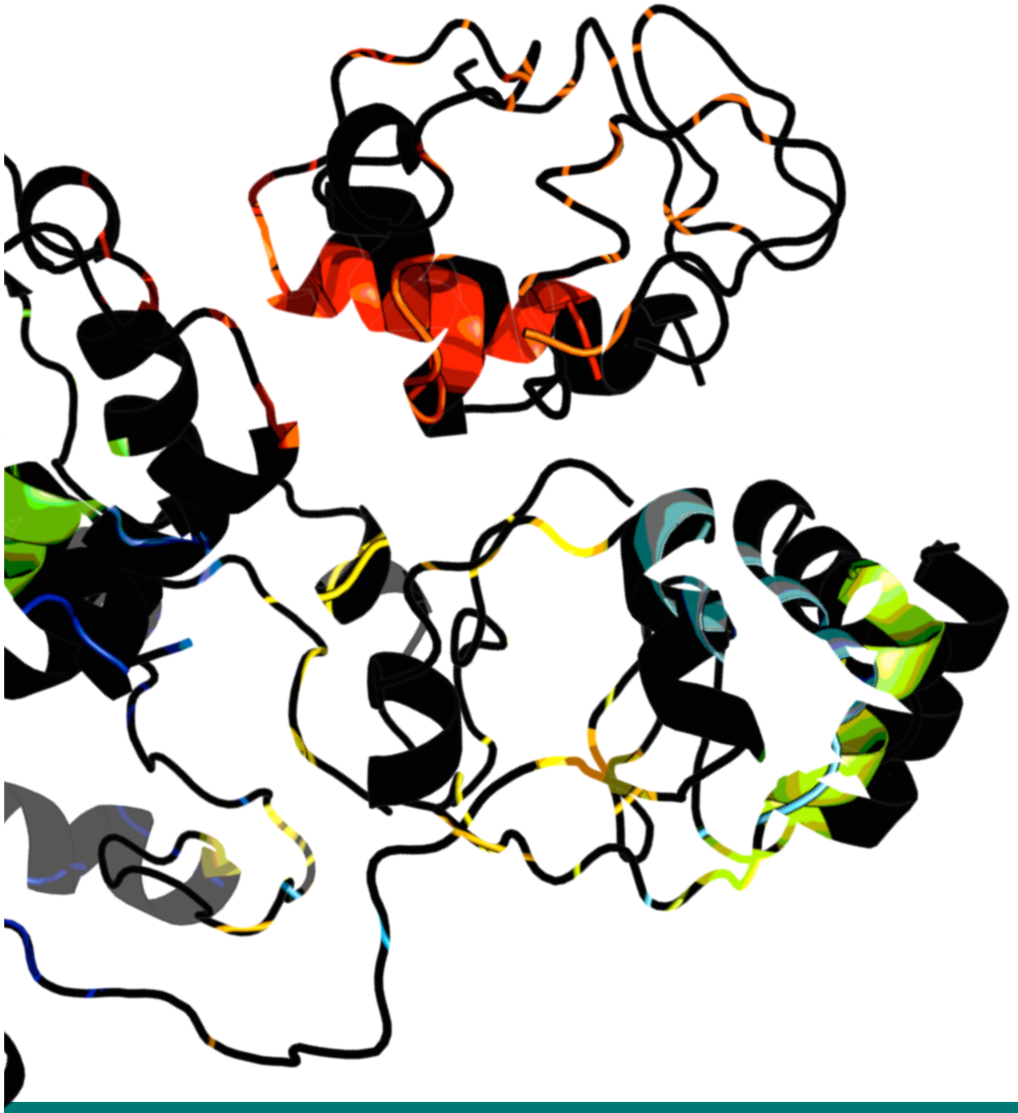


ARP/wARP 8.0

1044/1300 built
1024 docked
R/R_{free} 0.28/0.33







ARP/wARP
tips and tricks

EMBL

ARP/wARP tips and tricks

Resolution
best results above 2.6Å,
but phases can be poor

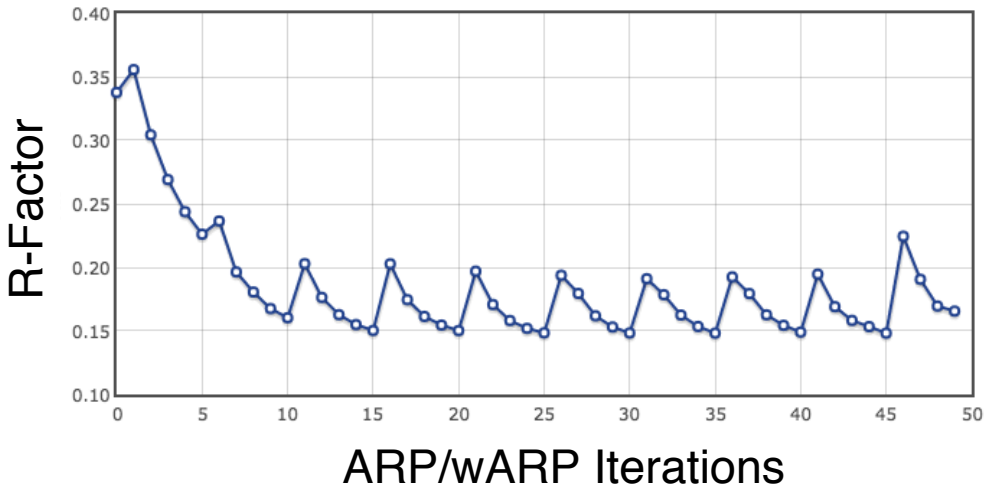
Space group
double check if model
building fails

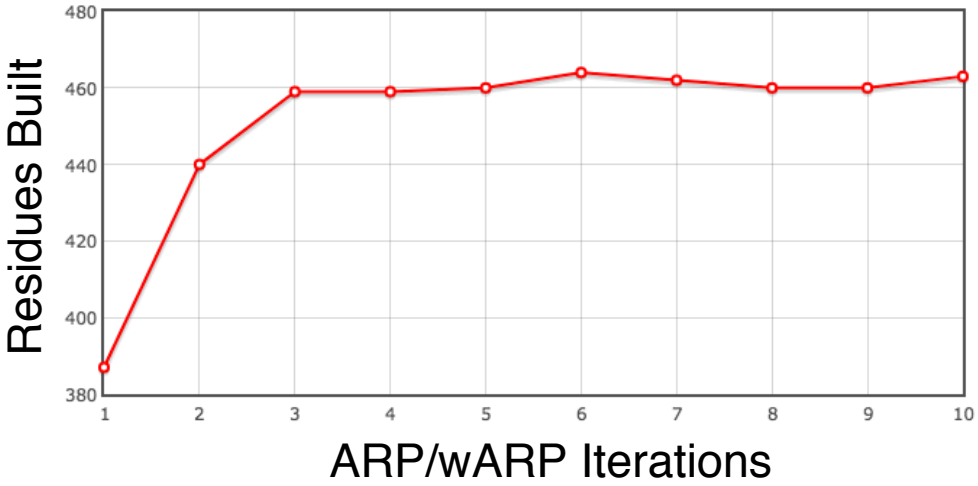
Data completeness
as high as possible.

Data quality (outliers, ice rings)
always check Wilson plot

Use secondary structure modeling
auto_albe.sh or quickfold in CCP4i

Try different strategies





ARP/wARP tips and tricks

Resolution
best results above 2.6Å,
but phases can be poor

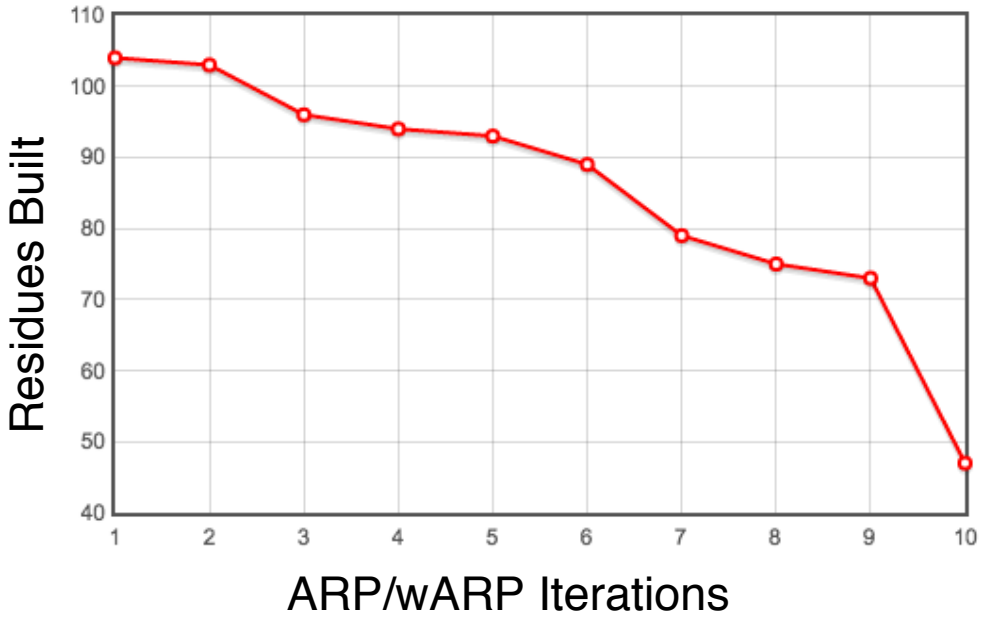
Space group
double check if model
building fails

Data completeness
as high as possible.

Data quality (outliers, ice rings)
always check Wilson plot

Use secondary structure modeling
auto_albe.sh or quickfold in CCP4i

Try different strategies



ARP/wARP tips and tricks

Resolution

best results above 2.6Å,
but phases can be poor

Space group

double check if model
building fails

Data completeness

as high as possible.

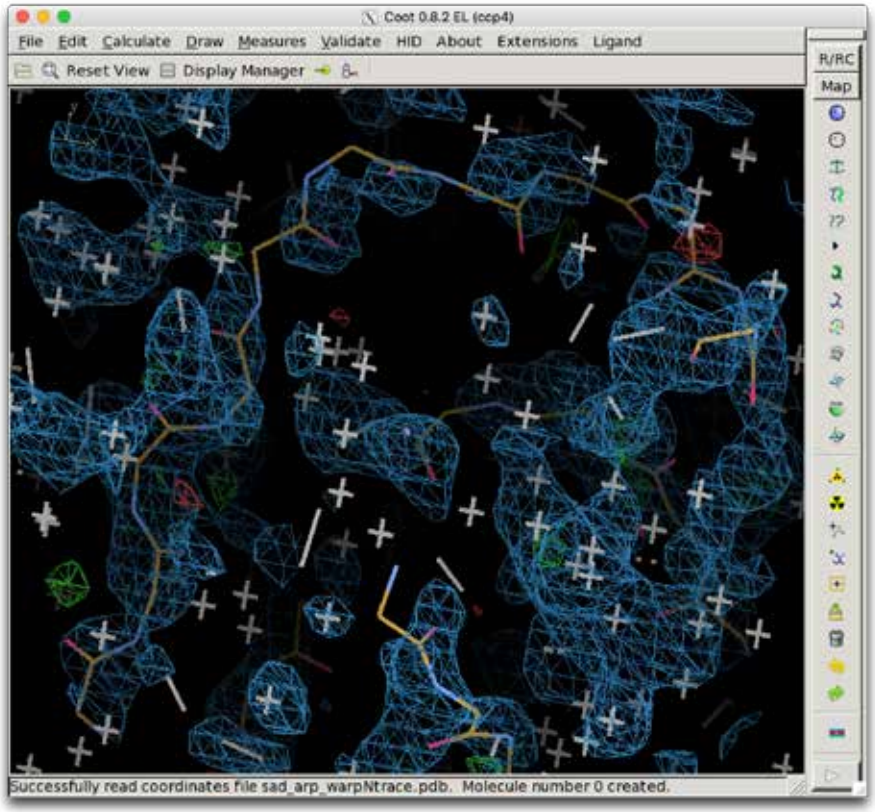
Data quality (outliers, ice rings)

always check Wilson plot

Use secondary structure modeling

auto_albe.sh or quickfold in CCP4i

Try different strategies



ARP/wARP tips and tricks

Resolution

best results above 2.6Å,
but phases can be poor

Data completeness

as high as possible.

Space group

double check if model
building fails

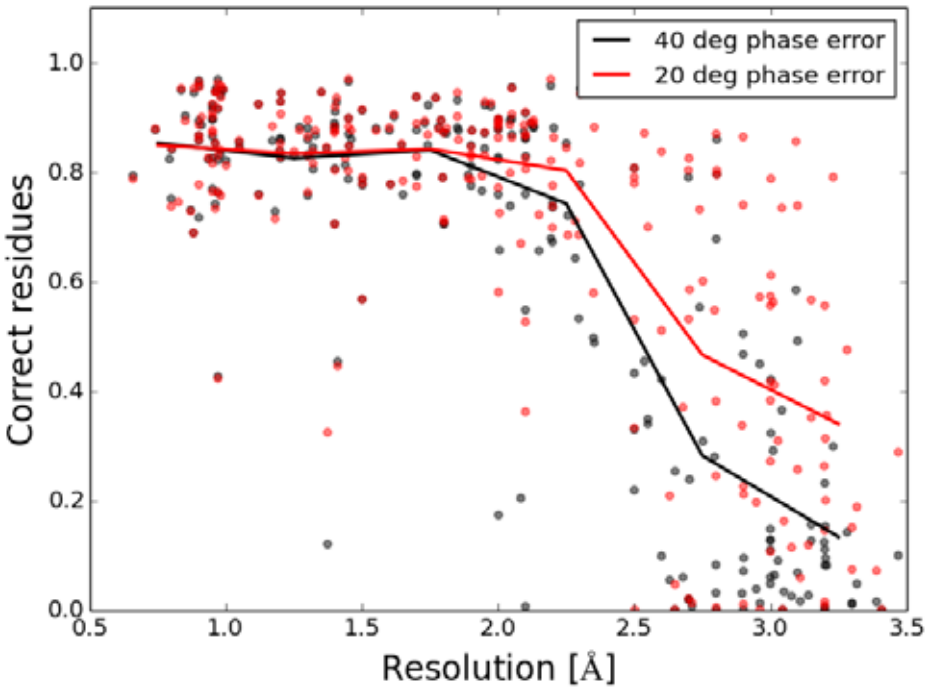
Data quality (outliers, ice rings)

always check Wilson plot

Use secondary structure modeling

auto_albe.sh or quickfold in CCP4i

Try different strategies



ARP/wARP tips and tricks

Resolution

best results above 2.6Å,
but phases can be poor

Data completeness

as high as possible.

Space group

double check if model
building fails

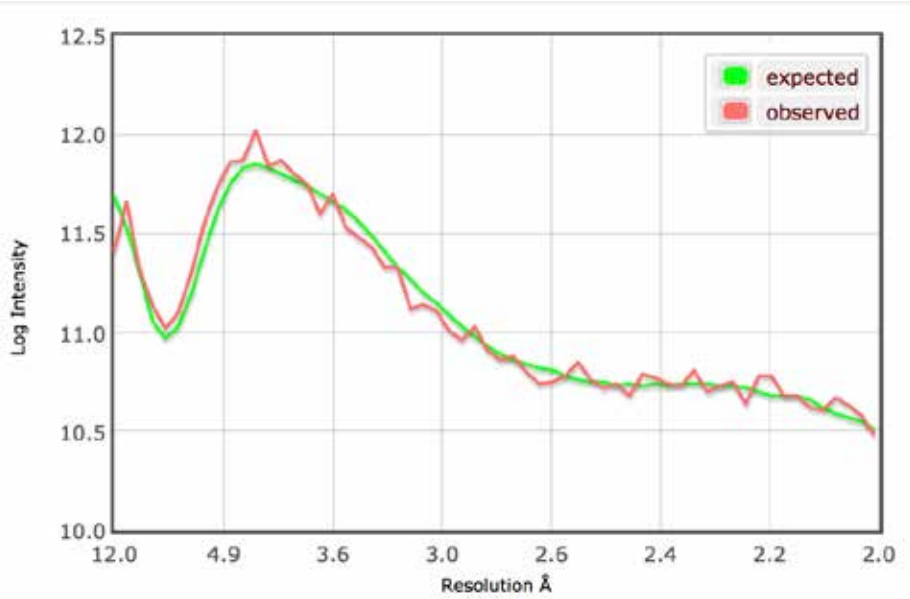
Data quality (outliers, ice rings)

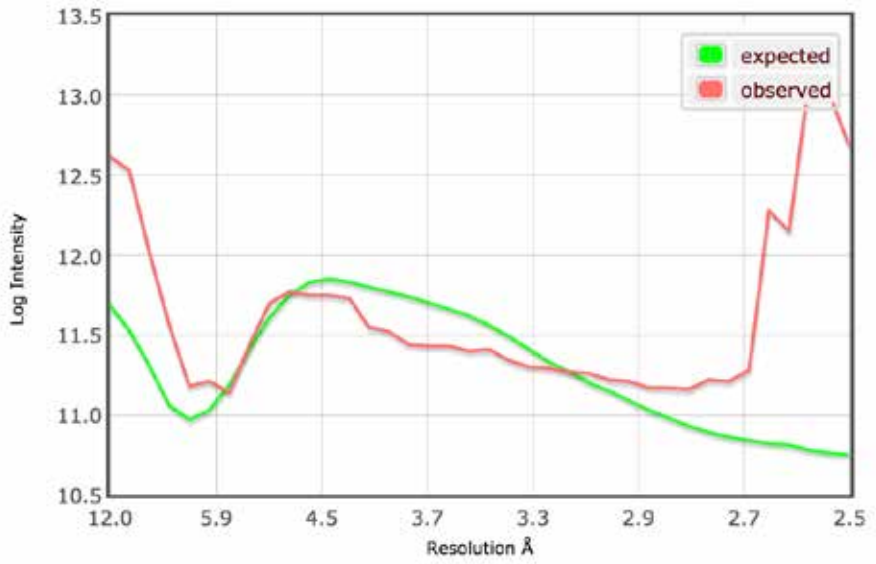
always check Wilson plot

Use secondary structure modeling

auto_albe.sh or quickfold in CCP4i

Try different strategies





ARP/wARP tips and tricks

Resolution
best results above 2.6Å,
but phases can be poor

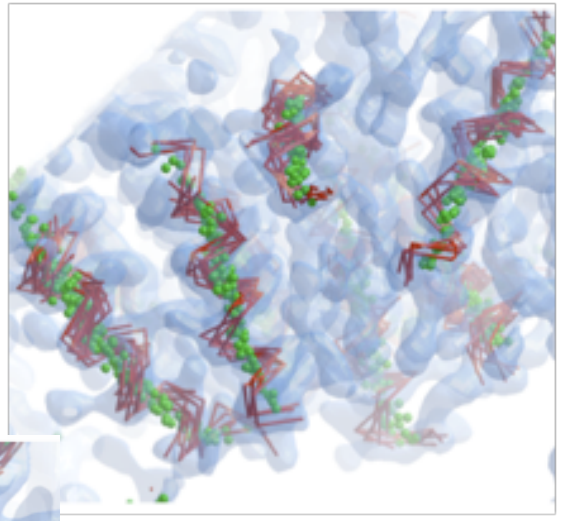
Data completeness
as high as possible.

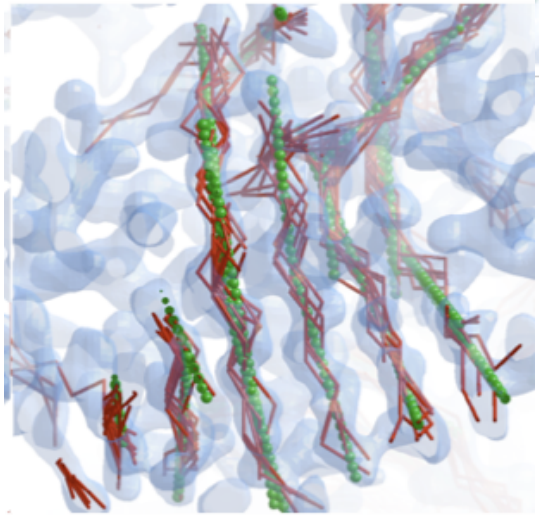
Space group
double check if model
building fails

Data quality (outliers, ice rings)
always check Wilson plot

Use secondary structure modeling
auto_albe.sh or quickfold in CCP4i

Try different strategies





ARP/wARP tips and tricks

Resolution

best results above 2.6Å,
but phases can be poor

Data completeness

as high as possible.

Space group

double check if model
building fails

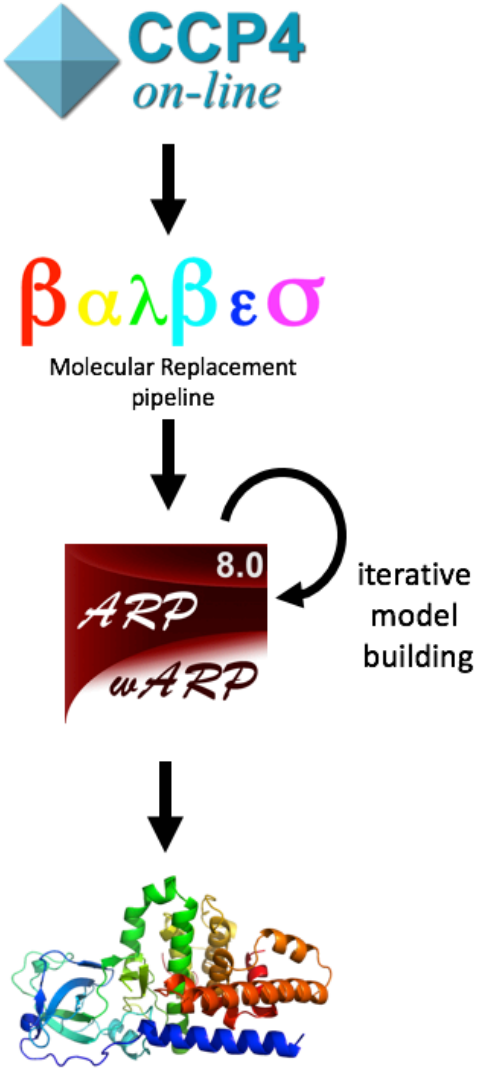
Data quality (outliers, ice rings)

always check Wilson plot

Use secondary structure modeling

auto_albe.sh or quickfold in CCP4i

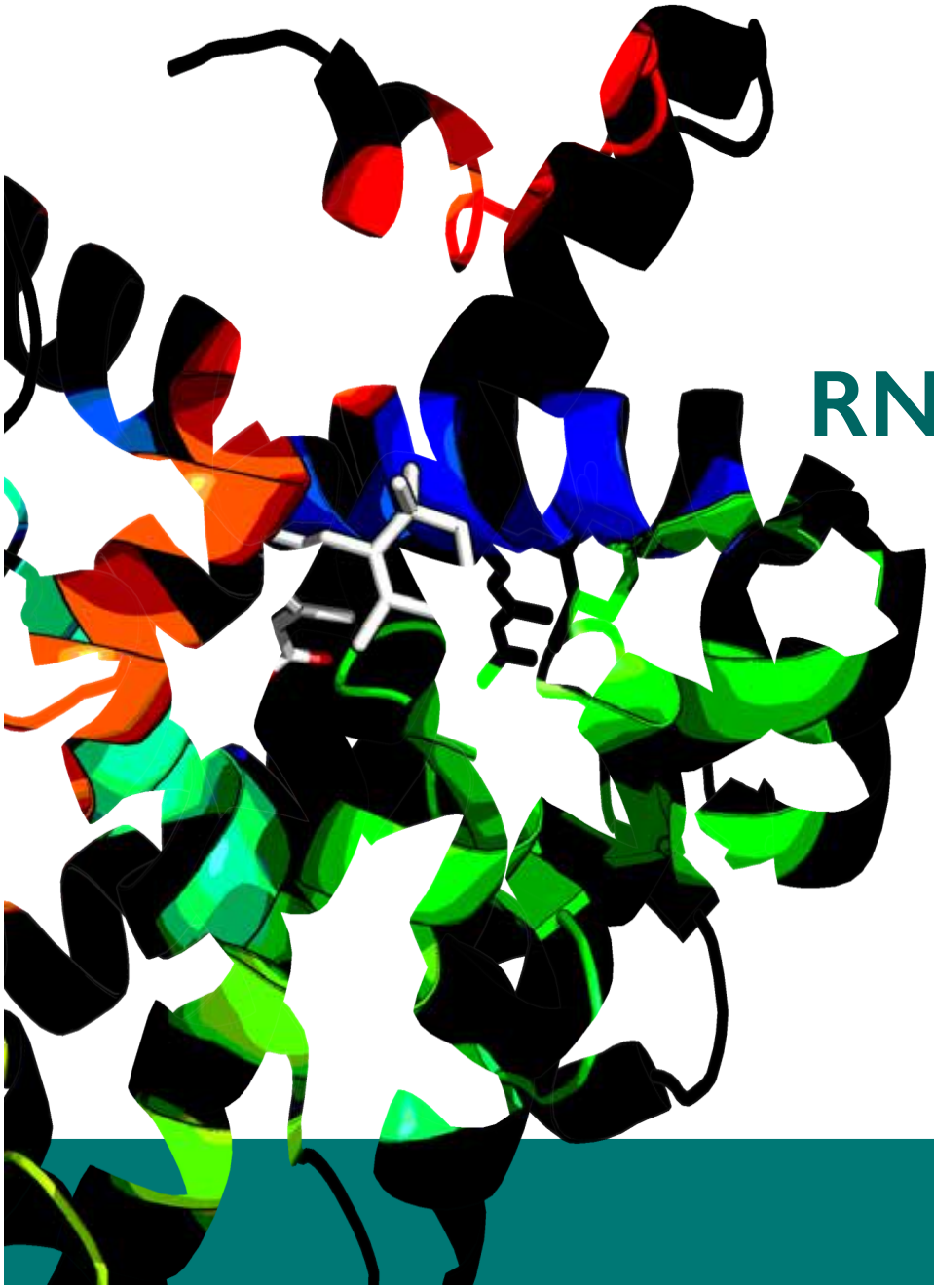
Try different strategies



Byung Hak Ha and Titus J. Boggon JBC 2017

Electron density map interpretation

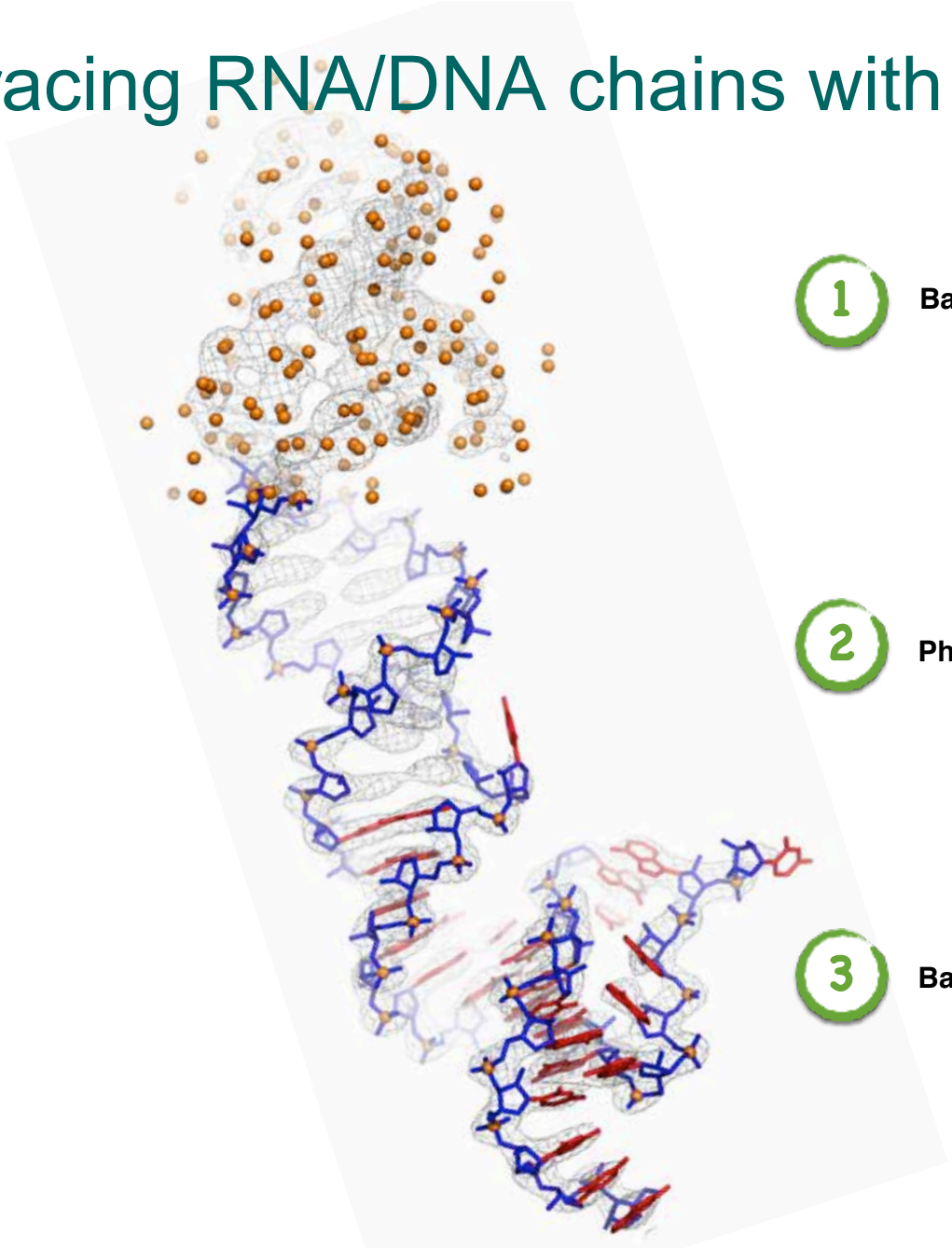
Art Wolfe/der Maus/Maria Chojnowska



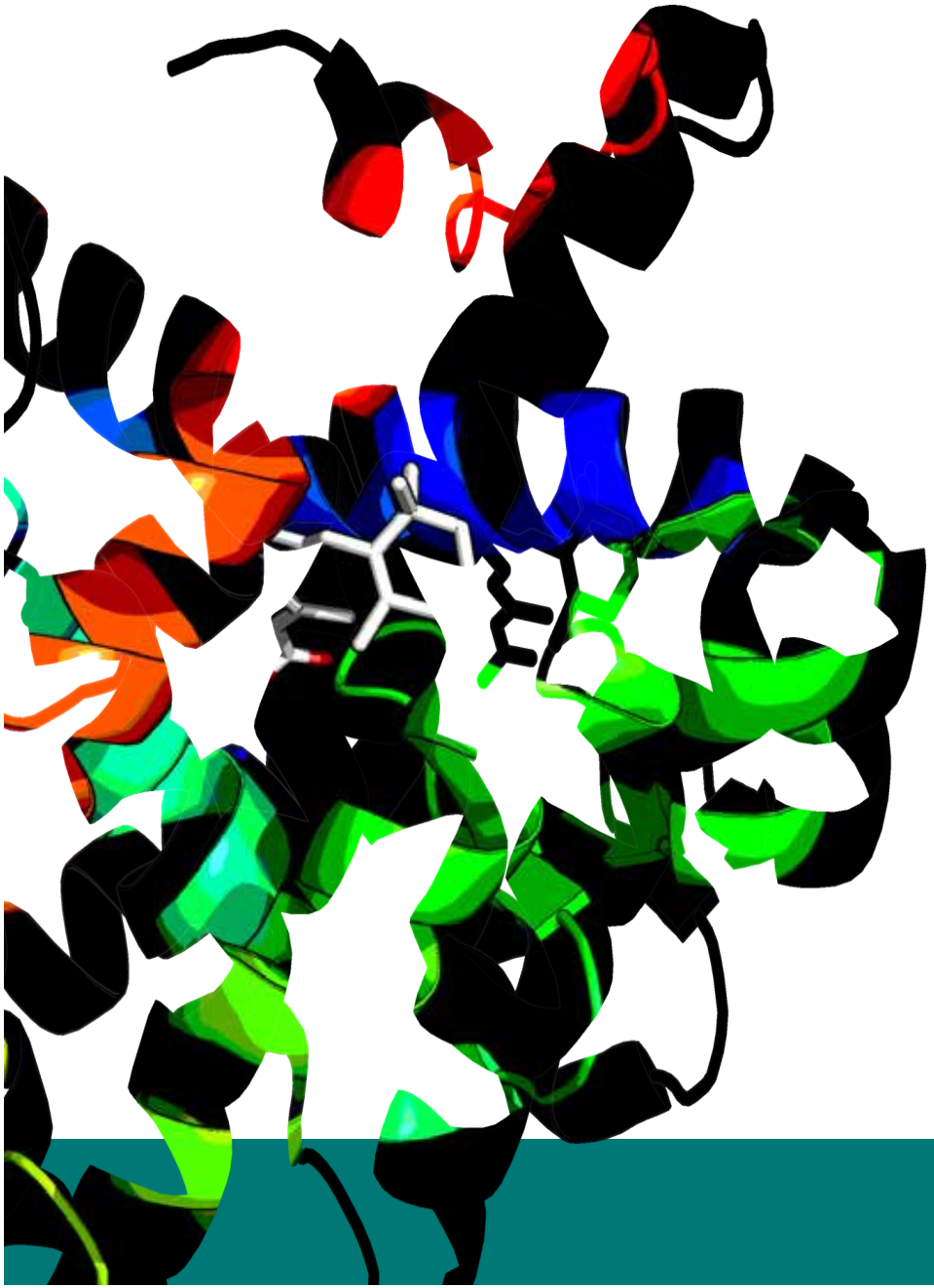
RNA/DNA Model Building with ARP/wARP

EMBL

Tracing RNA/DNA chains with ARP/wARP



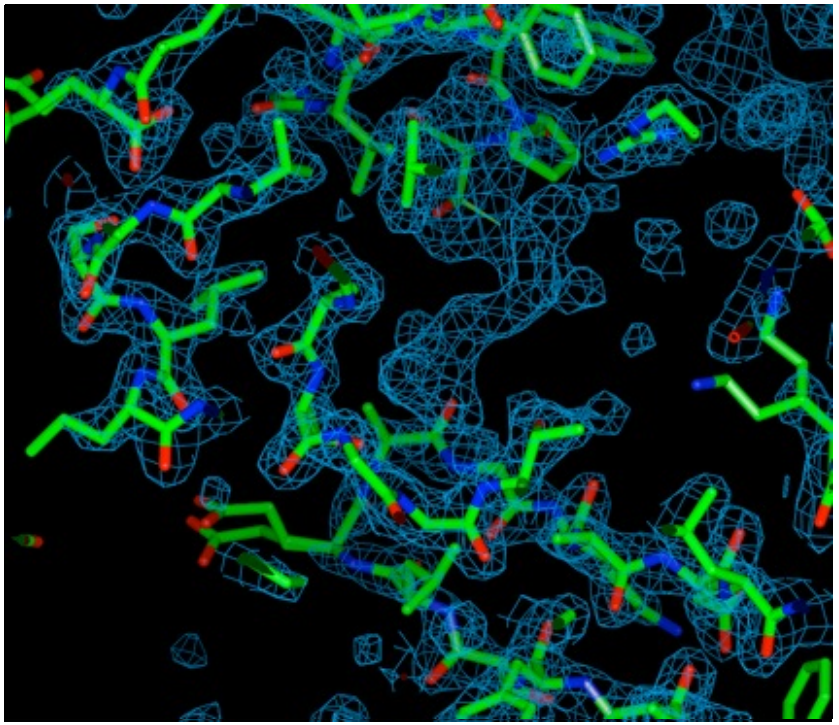
1. **Backbone phosphates detection**
   - NEW **Machine learning algorithm**
2. **Phosphate-ribose-phosphate units placement**
   - NEW **Updated geometry restrains**
   - NEW **Iterative chain tracing (like for proteins)**
3. **Bases building**



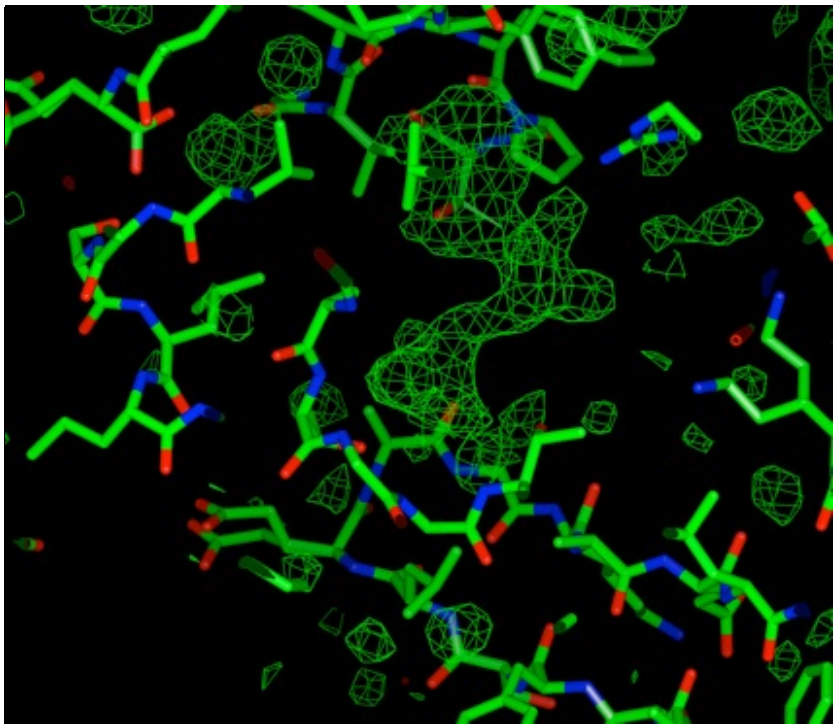
Building ligands with ARP/wARP

EMBL

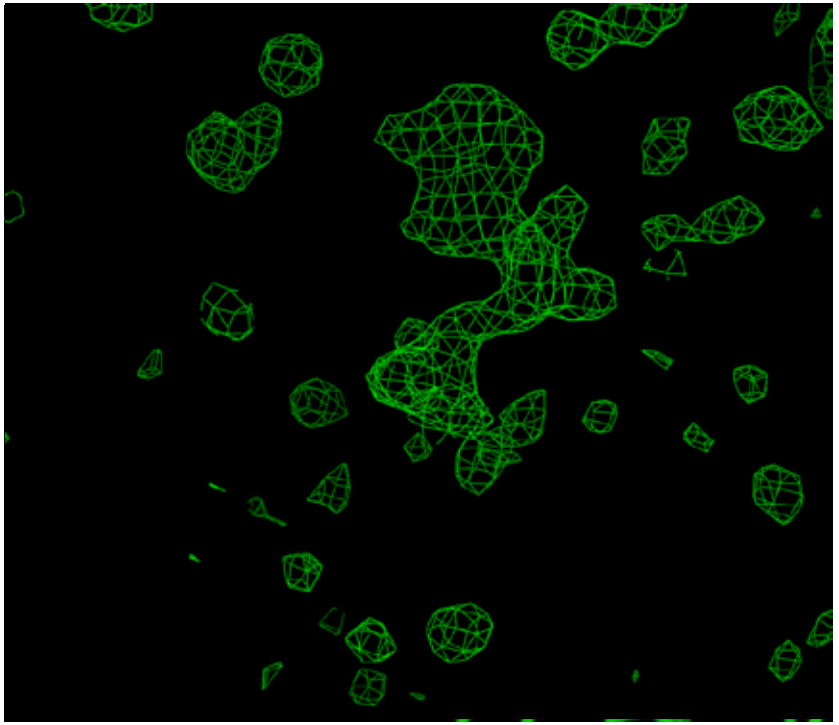
- electron density map (good resolution)
- protein model (refined)
- **difference density map**
  - shows **excess density** not explained by the protein model
  - to fit the appropriate ligand into



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

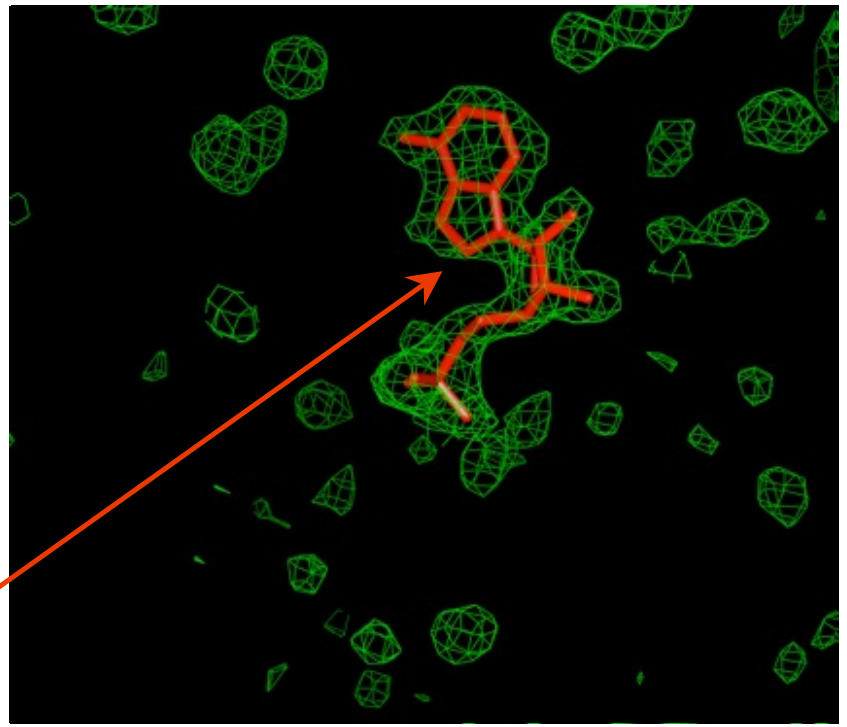


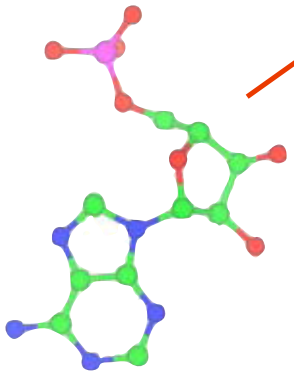
- electron density map (good resolution)
- protein model (refined)
- **difference density map**
  - shows **excess density** not explained by the protein model
  - to fit the appropriate ligand into



Challenges:

- different **resolutions** and data quality
- different ligand **complexity / topology**
- **partial disorder** of a ligand



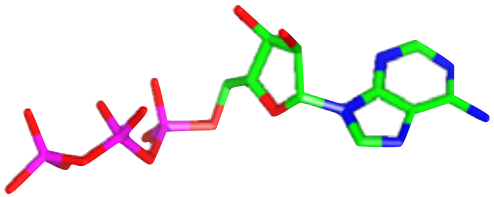


What we know at the beginning will determine the protocol:

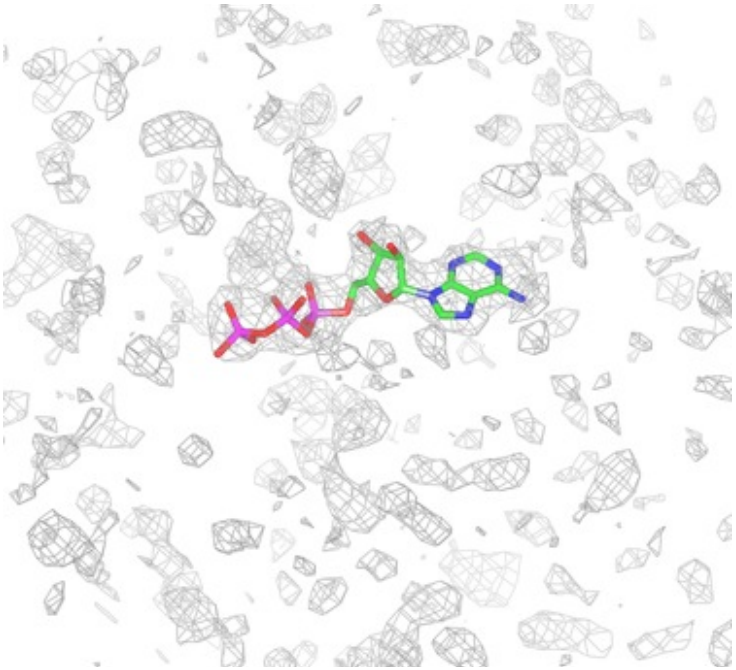
Ligand identity \ Binding site	Known	Unknown
Known	X	
Unknown		

1 ligand density

1 ligand



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



Prepare

- Prepare Sparse density map
- **Generate ligand topology**

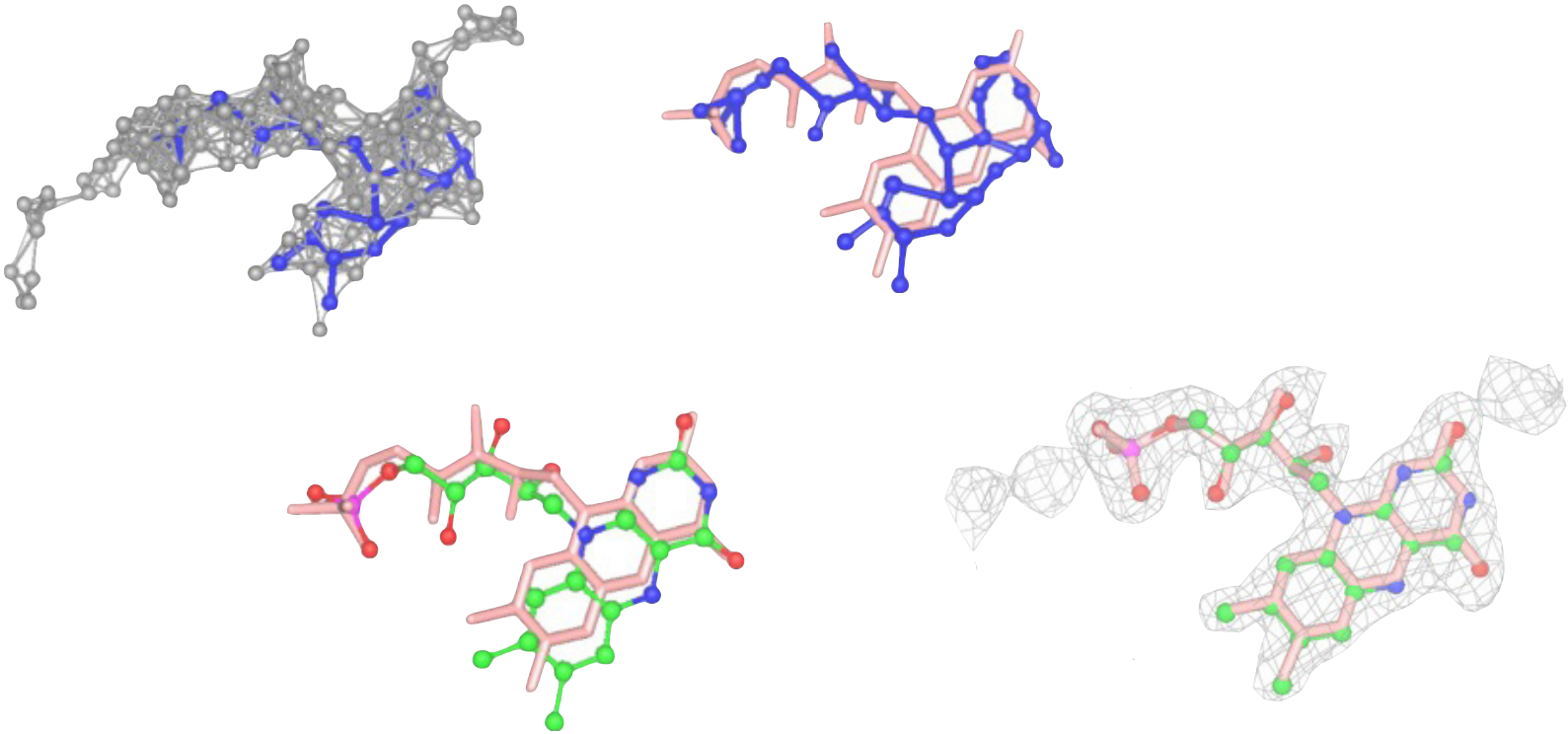
Fit

- Construct **ensemble** of ligand models
- **fit sparse density**

Refine

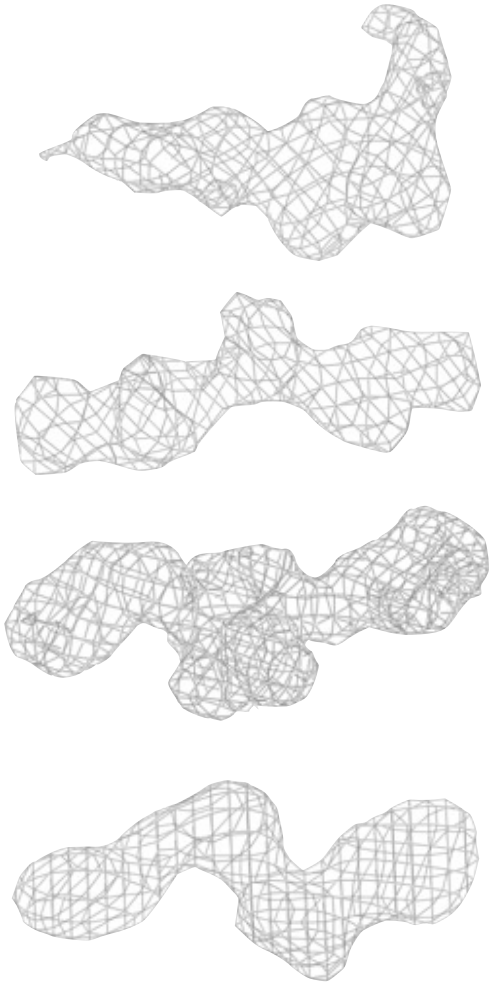
- Refine ligand coordinates
- Choose **best model**

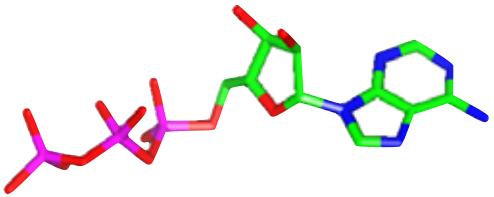
- Uses the **full density** and **topology information**
- Employs a gradient method to **optimise fit to density** and **ligand geometry** simultaneously



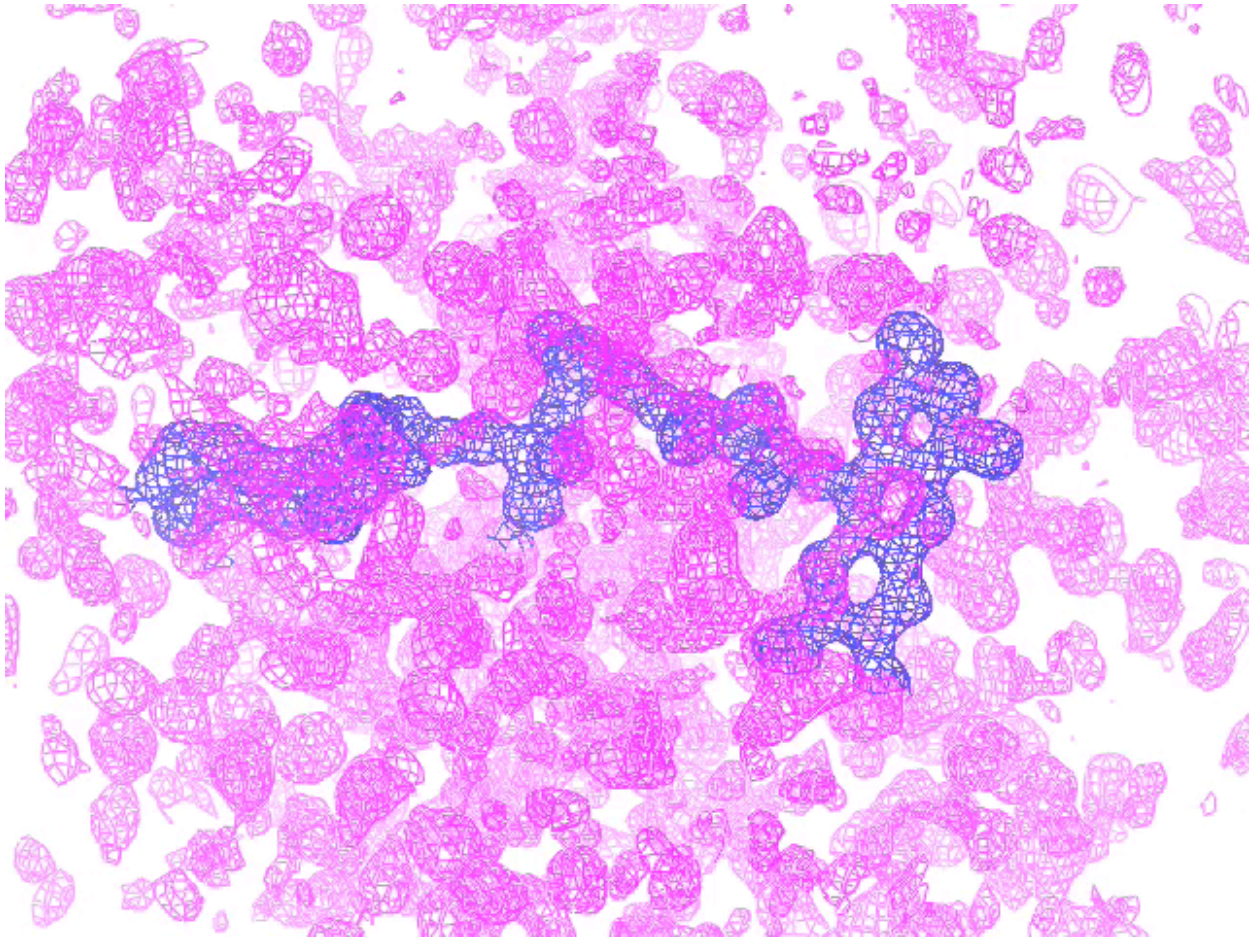
What we know at the beginning will determine the protocol:

Ligand identity \ Binding site	Known	Unknown
Known		X
Unknown		

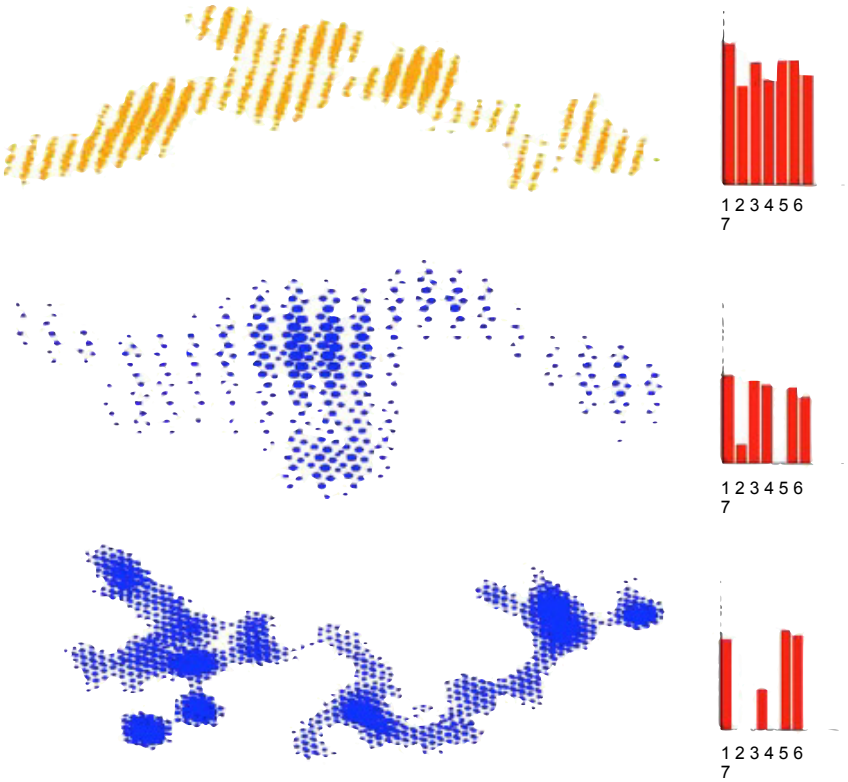




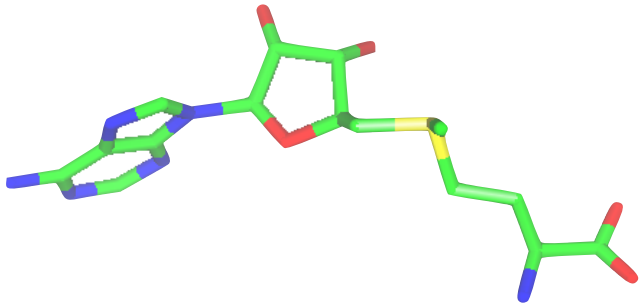
- The difference density map from **low to high** contour threshold



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

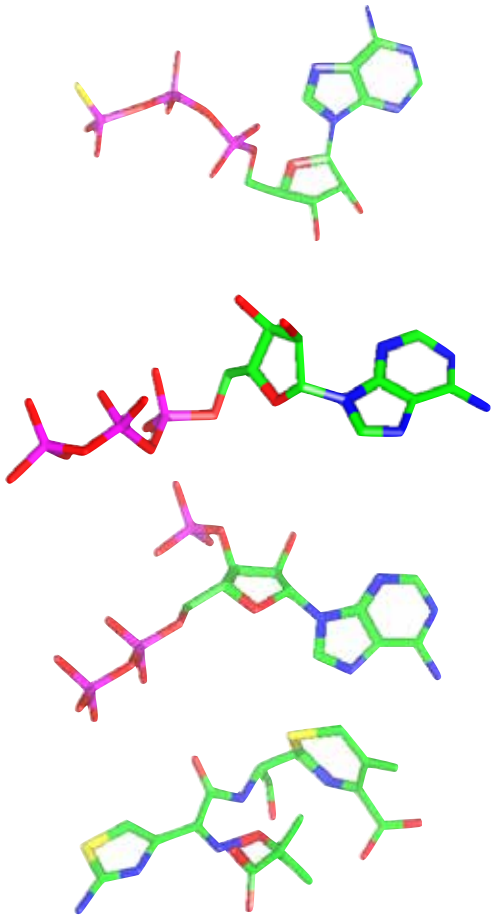
Decision is based on **feature vectors**

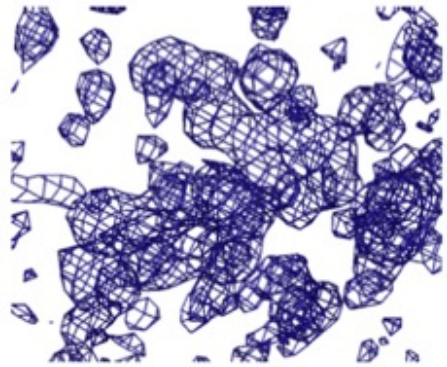
What we know at the beginning will determine the protocol:

Ligand identity \ Binding site	Known	Unknown
Known		
Unknown	X	

1 ligand density

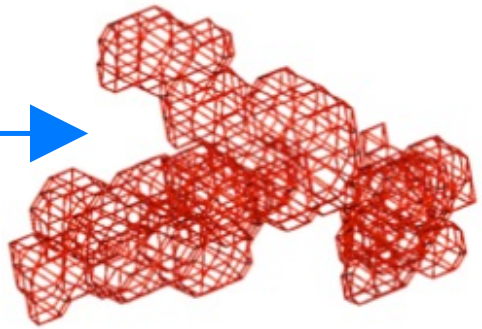
N ligand candidates





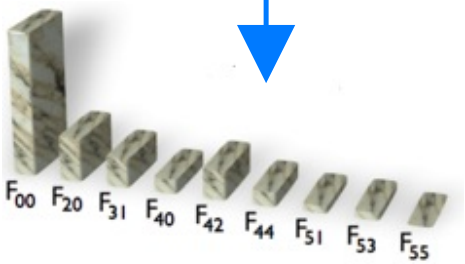
Density map (or otherwise surface of protein pocket)

Segmentation



Projection/ normalisation

Calculation of shape descriptors



Comparison

Database

84 of the most common ligands in crystallography

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

Ranking

Or your own cocktail…

Ligand

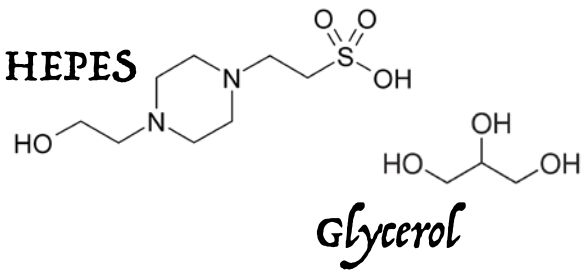
- 84 of the **most common ligands in crystallography**

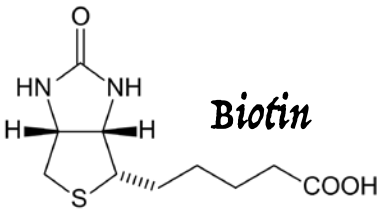
C. Carolan & V. Lamzin, Acta D (July 2014)

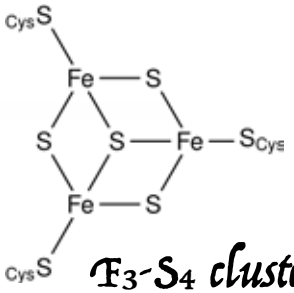
Ligand three-letter code	Ligand common name
017	Darunavir
1PE	Pentaethylene glycol
2GP	Guanosine 2-monophosphate
2PE	Nonaethylene glycol
5GP*	Guanosine 5-monophosphate
A3P*	Adenosine 3′,5′-diphosphate
ACO*	Acetyl coenzyme A
ADE	Adenine
ADN	Adenosine
ADP	Adenosine 5′-diphosphate
AKG	2-Oxoglutaric acid
AMP	Adenosine monophosphate
ATP*	Adenosine 5′-triphosphate
B3P	2-[3-(2-Hydroxy-1,1-dihydroxymethyl-ethylamino)-propylamino]-2-hydroxymethyl-propane-1,3-diol
BCL	Bacteriochlorophyll A
BTB	Bis-tris buffer
BTN	Biotin
C2E*	Cyclic diguanosine monophosphate
CAM	Camphor
CDL	Cardiolipin
CHD	Cholic acid
CIT	Citric acid
CLA	Chlorophyll A
CMP	Adenosine 3′,5′-cyclic monophosphate
COA	Coenzyme A
CXS	3-Cyclohexyl-1-propylsulfonic acid
CYC	Phycocyanobilin
DIO	1,4-Diethylene dioxide
DTT	1,4-Dithiothreitol
EPE	HEPES

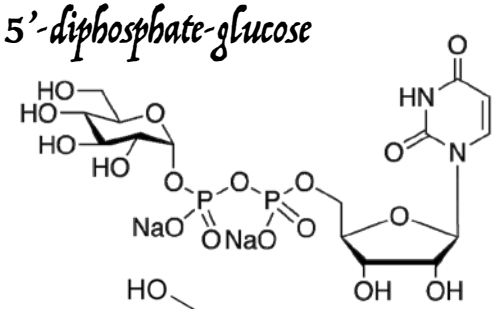
Ligand three-letter code	Ligand common name
F3S	Fe_3–S_4 cluster
FAD*	Flavin-adenine dinucleotide
FMN*	Flavin mononucleotide
FPP	Farnesyl diphosphate
GOL	Glycerol
GSH	Glutathione
H4B	5,6,7,8-Tetrahydrobiopterin
HC4	*para*-Coumaric acid
HEA*	Haem A
HED	2-Hydroxyethyl disulfide
HEM	Haem
IMD	Imidazole
IPH	Phenol
LDA	Lauryl dimethylamine-*N*-oxide
MES	2-(*N*-Morpholino)ethanesulfonic acid
MLI	Malonate ion
MLT	D-Malate
MPD	(4*S*)-2-Methyl-2,4-pentanediol
MTE	Phosphonic acid mono-(2-amino-5,6-dimercapto-4-oxo-3,7,8A,9,10,10A-hexahydro-4H-8-oxa-1,3,9,10-tetraaza-anthracen-7-ylmethyl)ester
MYR	Myristic acid
NAD*	Nicotinamide adenine dinucleotide
NAP*	Nicotinamide adenine dinucleotide phosphate
NCO	Cobalt hexammine(III)
NHE	2-(*N*-Cyclohexylamino)ethanesulfonic acid
OLA	Oleic acid
ORO	Orotic acid
P6G	Hexaethylene glycol
PEG	Di(hydroxyethyl)ether
PEP	Phosphoenolpyruvate
PG4	Tetraethylene glycol
PGA	2-Phosphoglycolic acid
PGO	*S*-1,2-Propanediol
PHQ	Benzyl chlorocarbonate

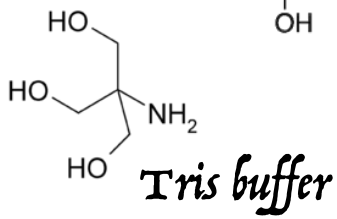
Ligand three-letter code	Ligand common name
PLM	Palmitic acid
PLP	Pyridoxal-5′-phosphate
POP	Pyrophosphate^{2-}
PYR	Pyruvic acid
RET	Retinal
SAM*	*S*-Adenosylmethionine
SF4	Iron–sulfur cluster
SIA	*O*-Sialic acid
SO4	Sulfate ion
SPO	Spheroidene
STU*	Staurosporine
TAM	Tris(hydroxyethyl)aminomethane
THP	Thymidine 3′,5′-diphosphate
TLA	L-(+)-Tartaric acid
TPP	Thiamine diphosphate
TRS	Tris buffer
TYD	Thymidine 5′-diphosphate
U10	Coenzyme Q10
UPG	Uridine 5′-diphosphate-glucose



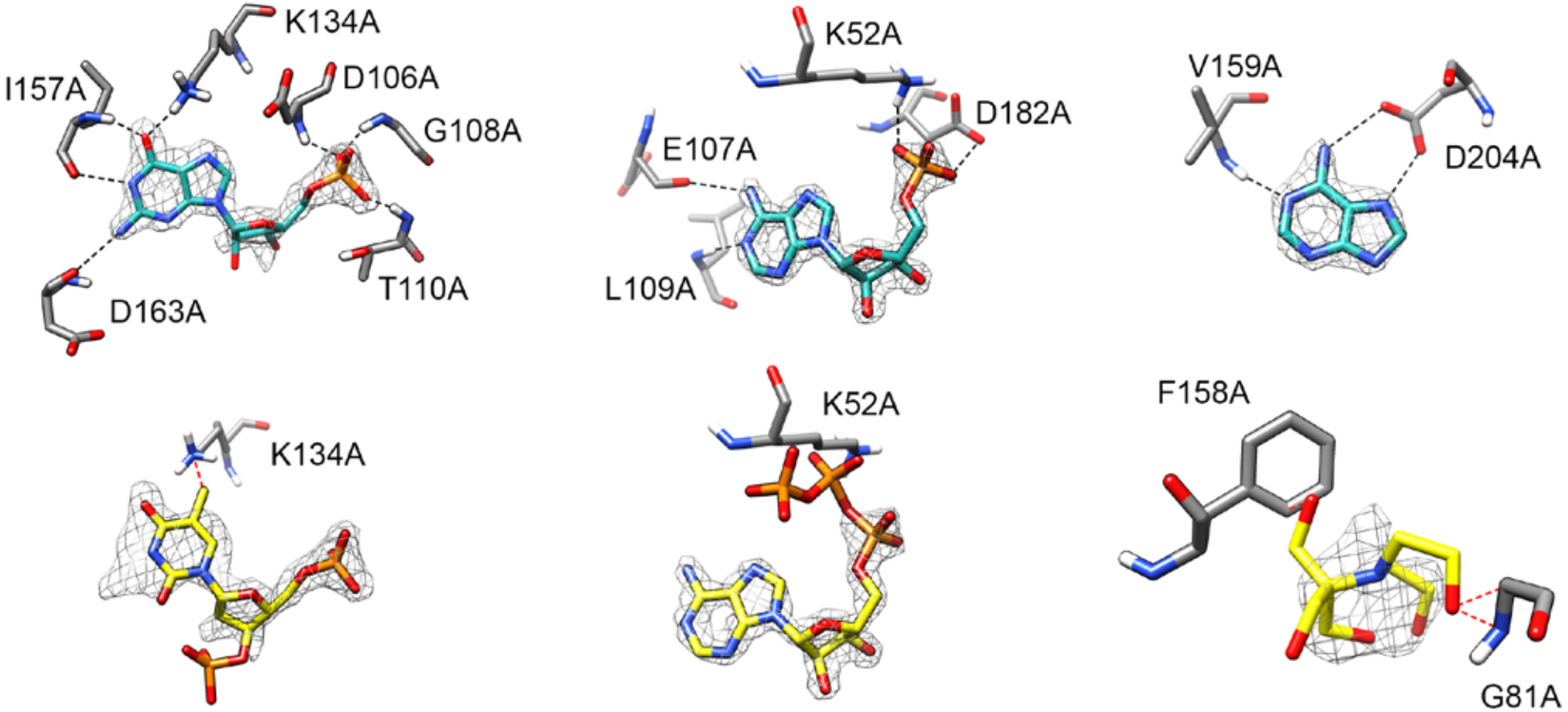






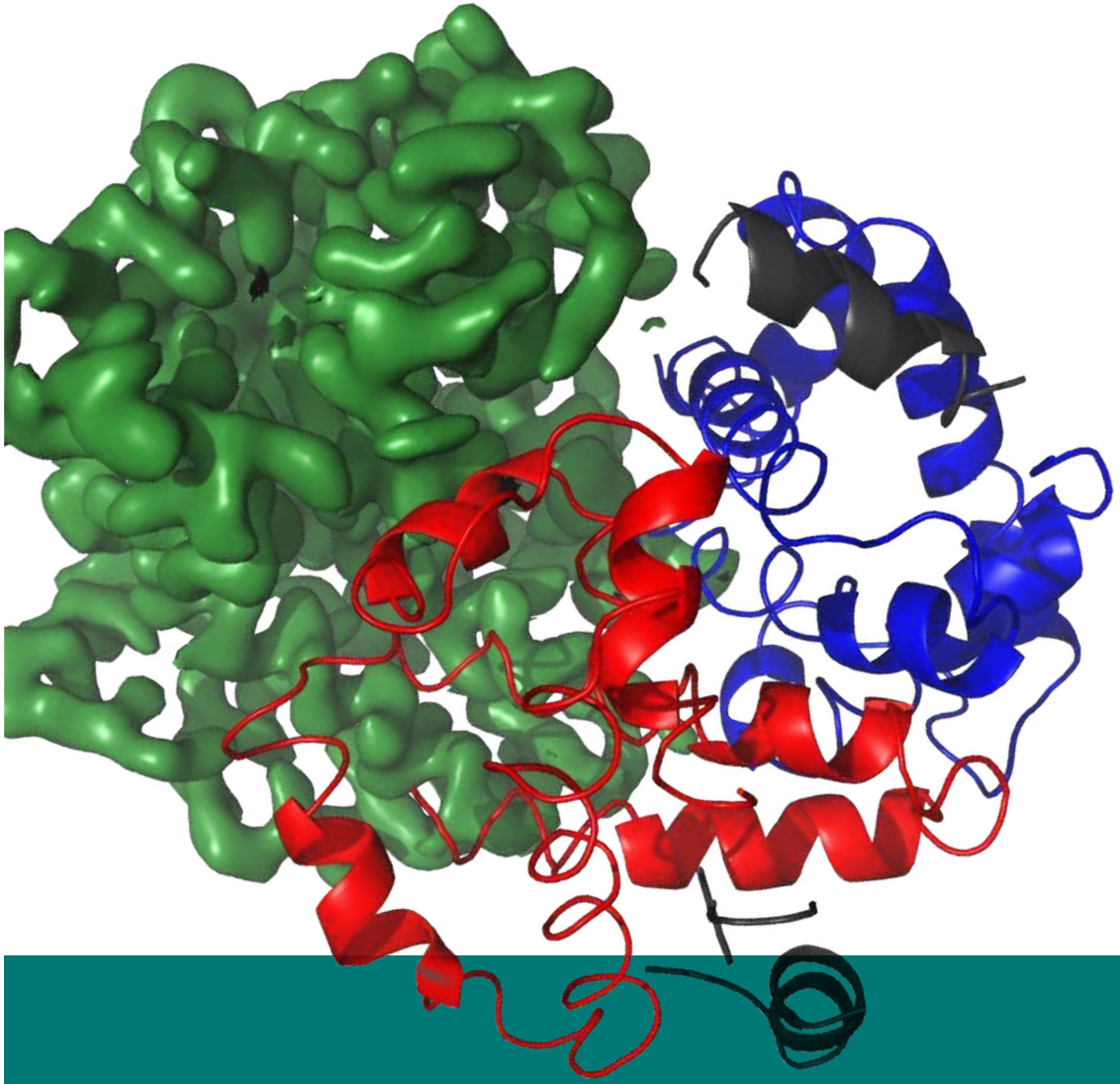


$$V_{bound}^{P-L} = W_{VDW}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + W_{elec}\sum_{i,j}\left(\frac{q_i q_j}{\varepsilon(r_{ij})r_{ij}}\right) + W_{sol}\sum_{i,j}\left(S_i V_j + S_j V_i\right)\exp\left(\frac{-r_{ij}^2}{2\sigma^2}\right)$$



Ligand guessing without (yellow skeleton) and with (blue skeleton) the use of the estimated energy as an additional parameter.

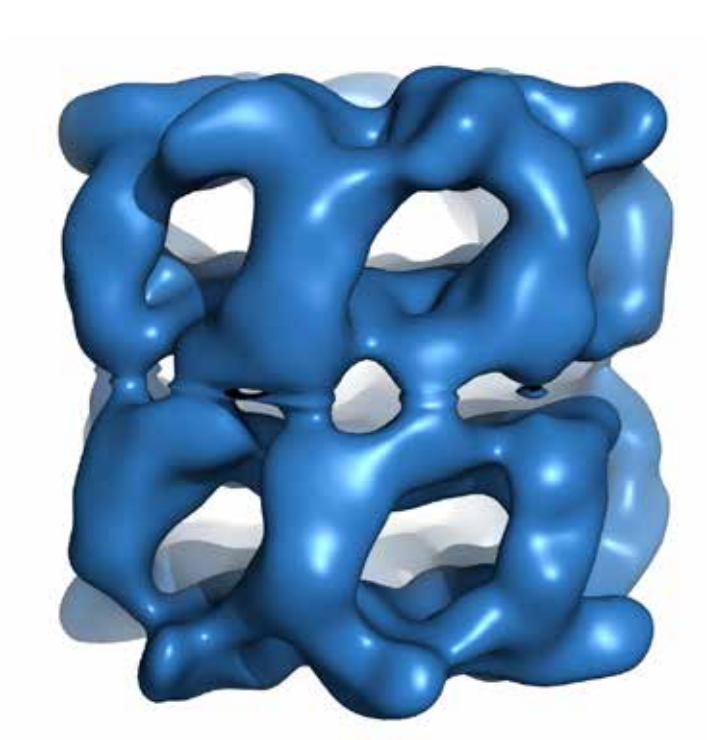
D. Beshnova, J. Pereira, V. Lamzin, Acta D (2017)



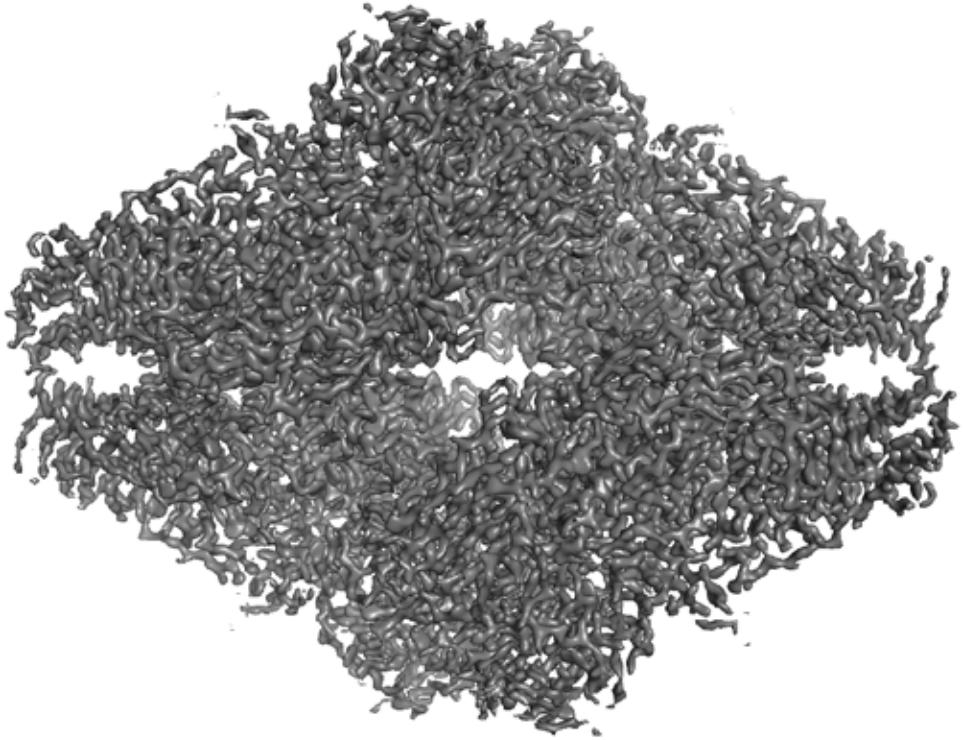
Building models into electron microscopy maps

EMBL

The Resolution Revolution

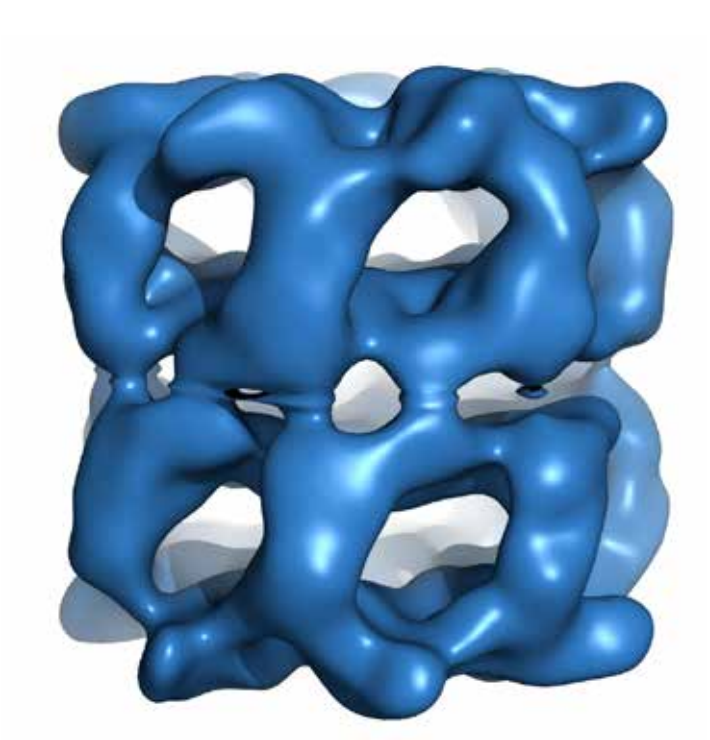


GroEL @ 18 Å resolution (EMD-5143)

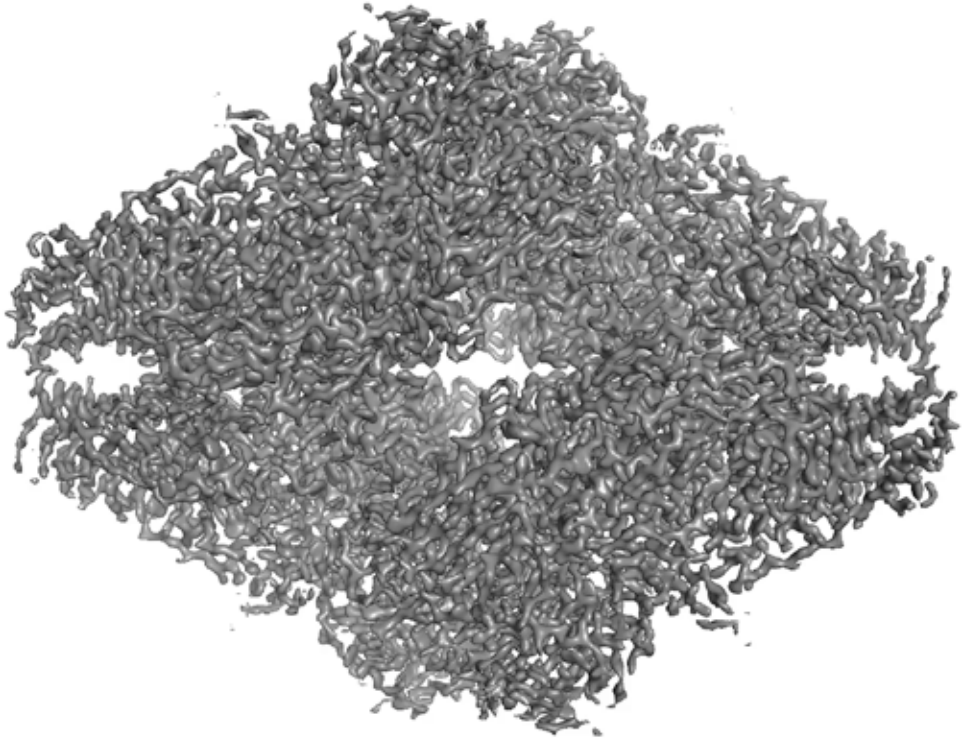


β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

The Resolution Revolution

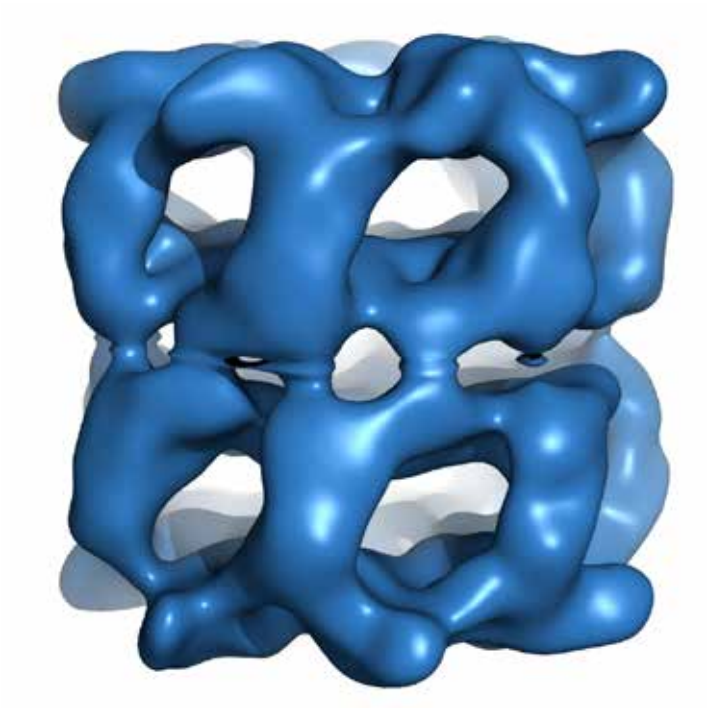


GroEL @ 18A resolution (EMD-5143)

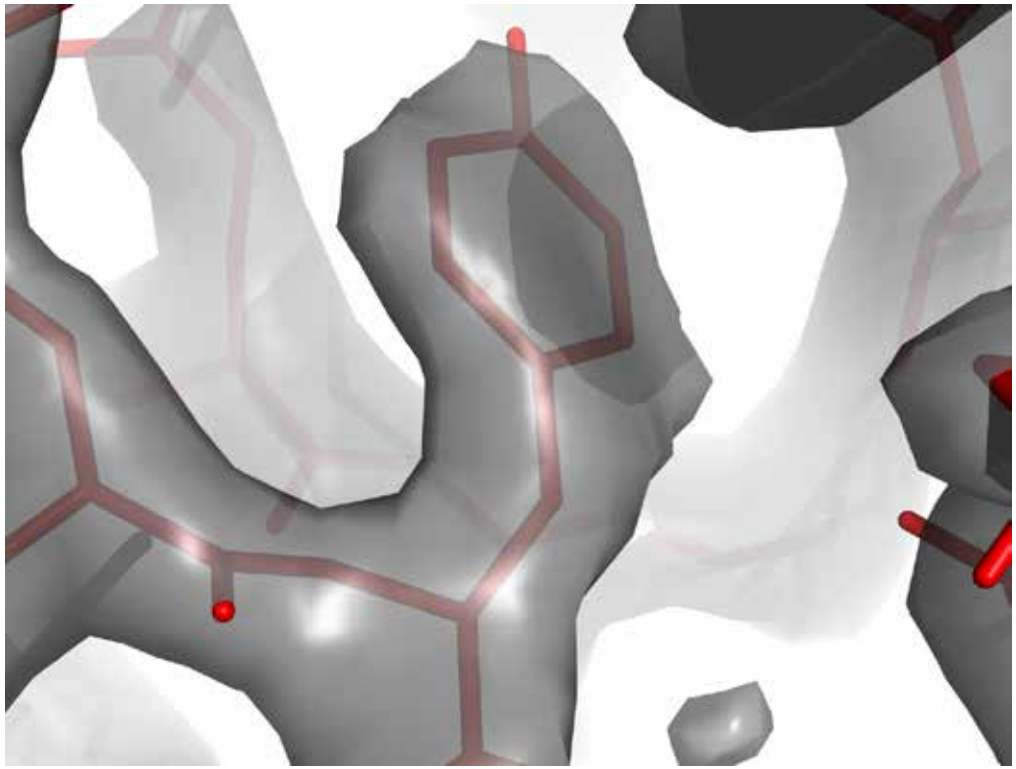


β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

The Resolution Revolution

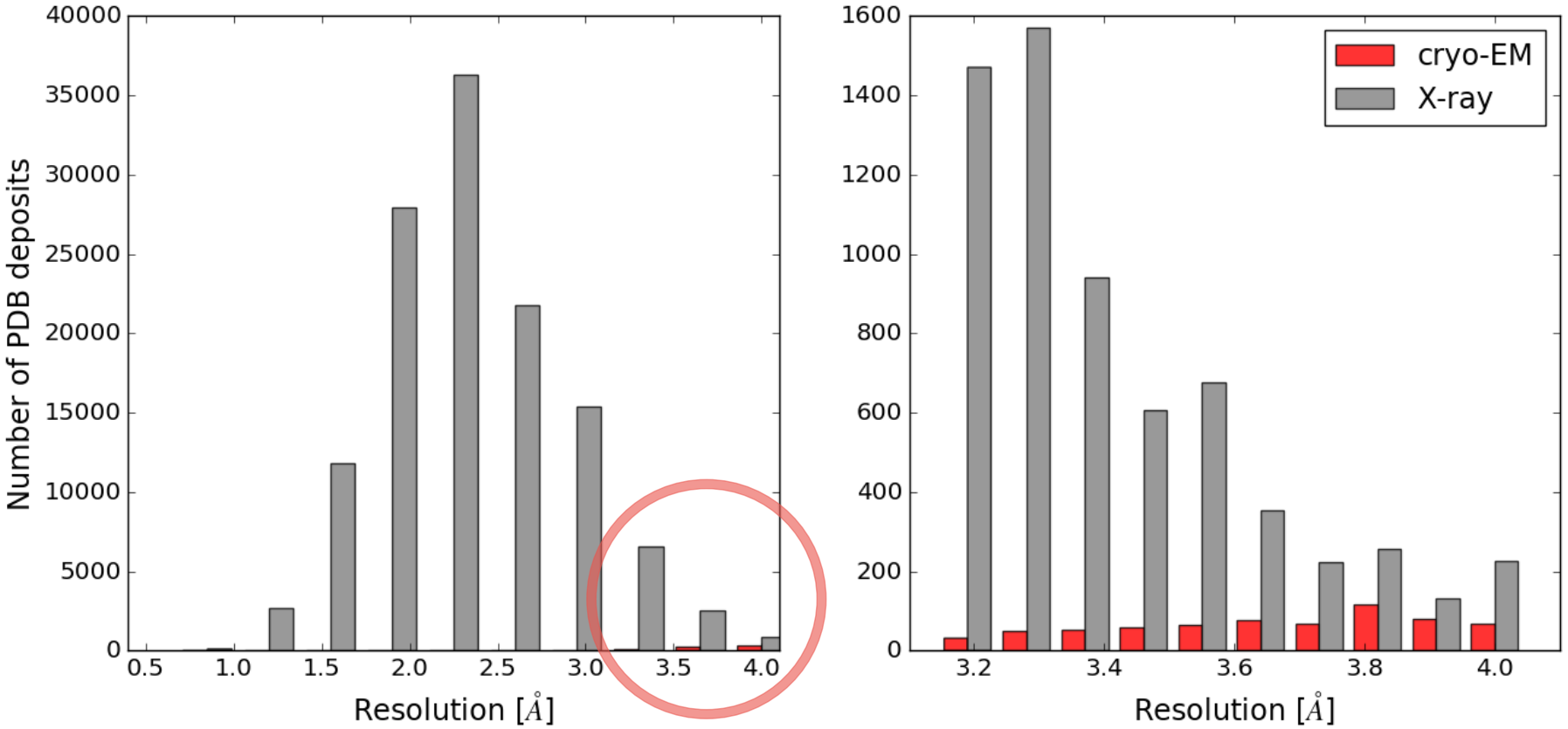


GroEL @ 18A resolution (EMD-5143)



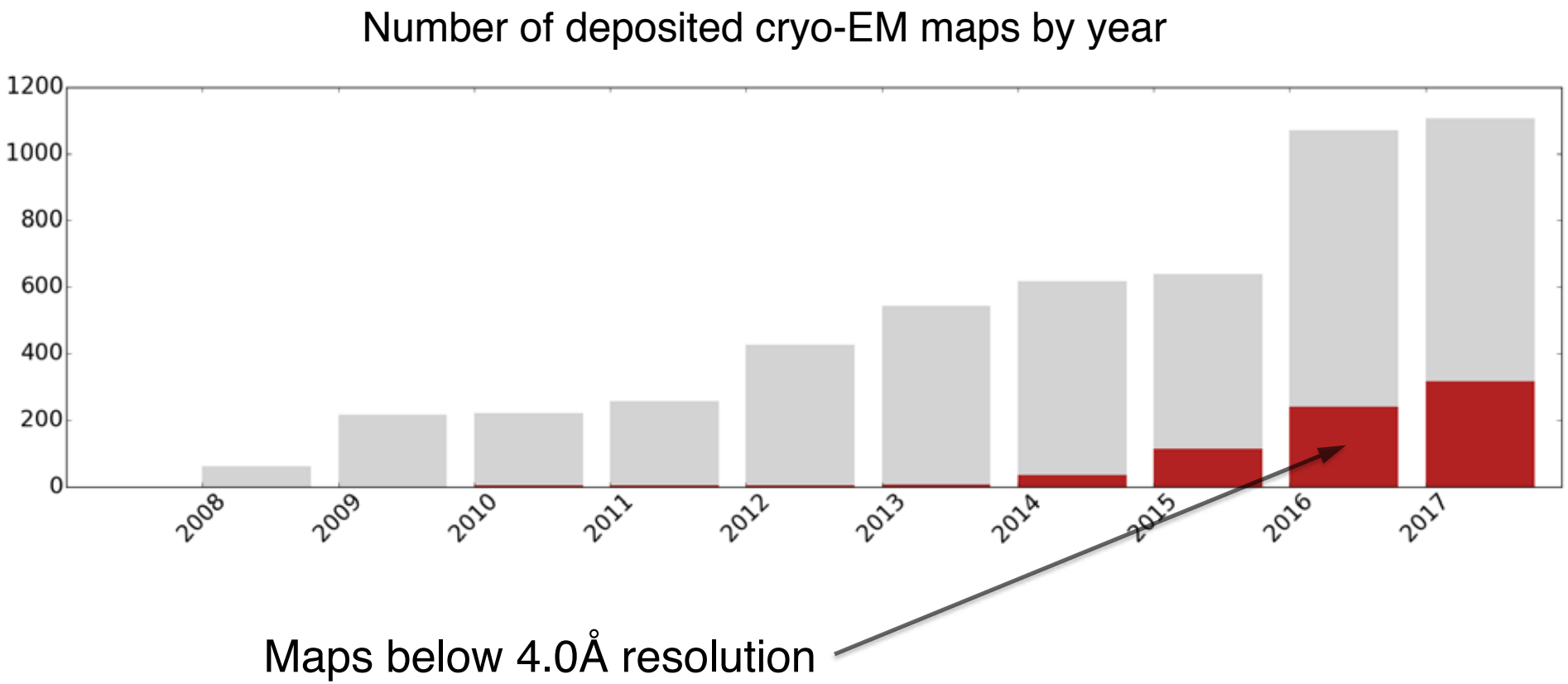
β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

The Resolution Revolution



Data taken from https://www.ebi.ac.uk/pdbe/emdb/

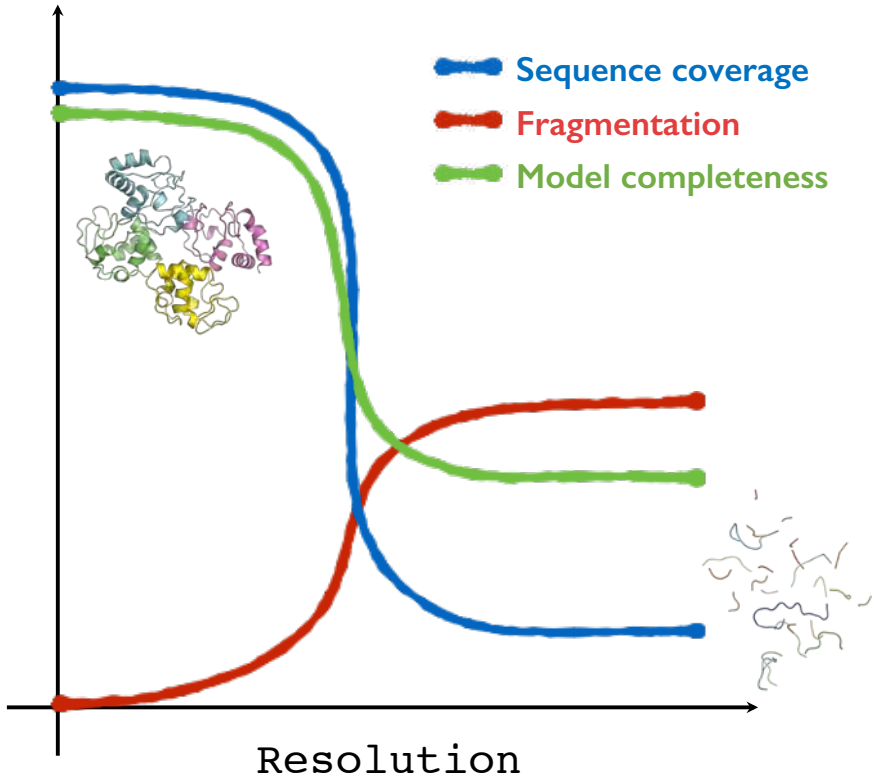
The Resolution Revolution



Data taken from https://www.ebi.ac.uk/pdbe/emdb/

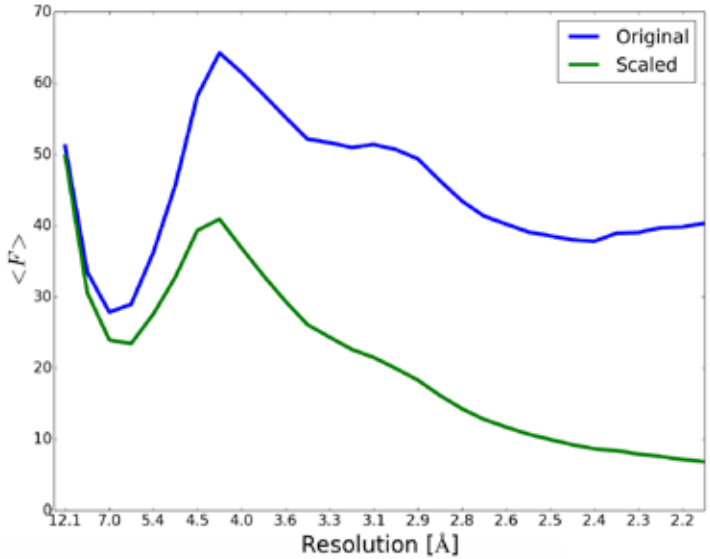
ARP/wARP for cryo-EM

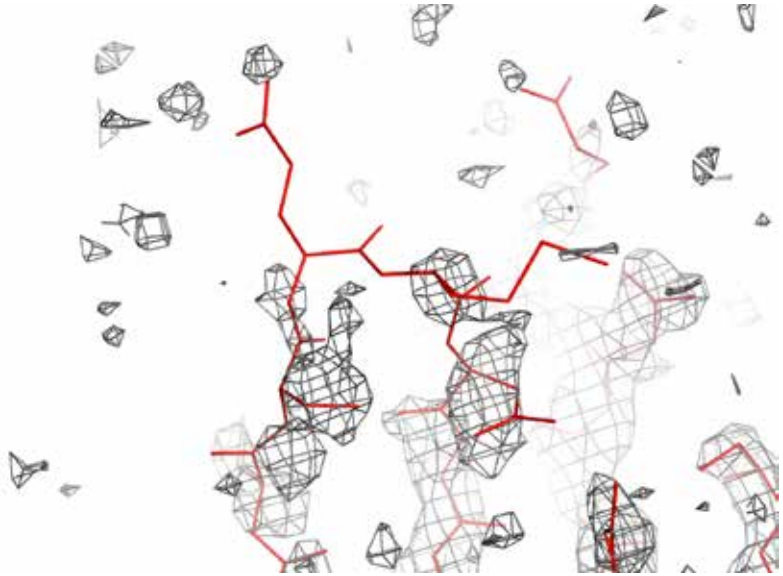
- EM and X-ray map features differ
- EM maps local resolution vary
- EM models may be very large
- EM maps can be over-sharpened
  (unrealistic B-factors)



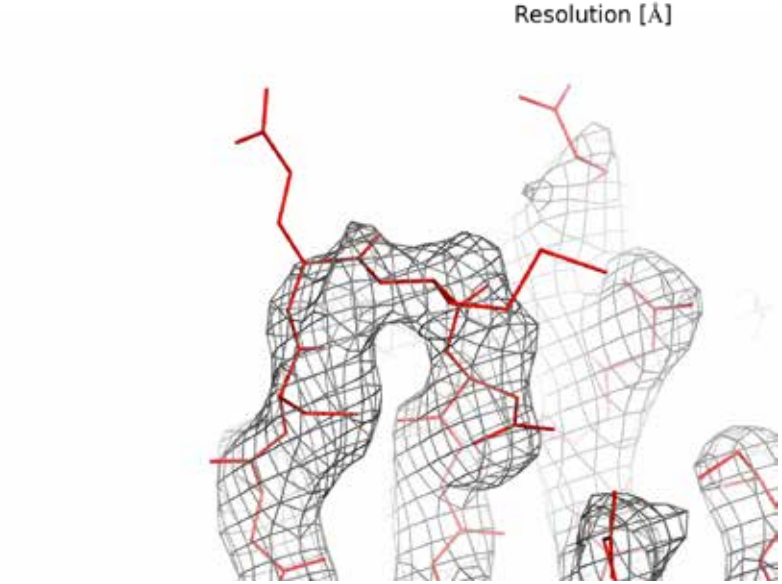
Map preprocessing

β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)



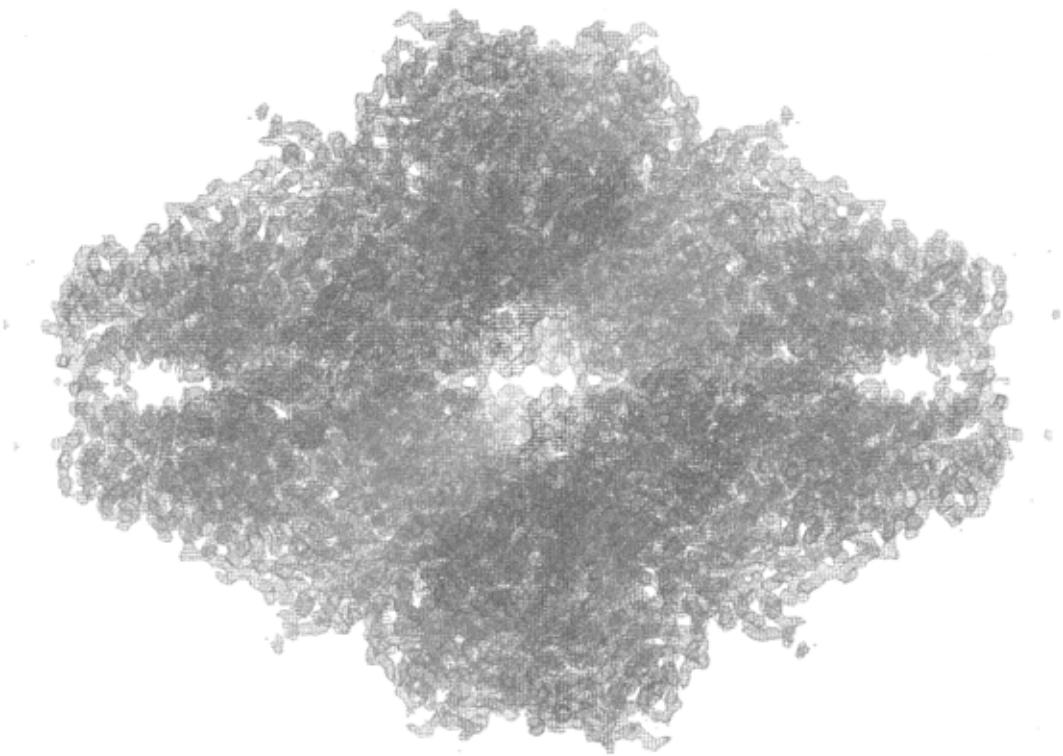


original map



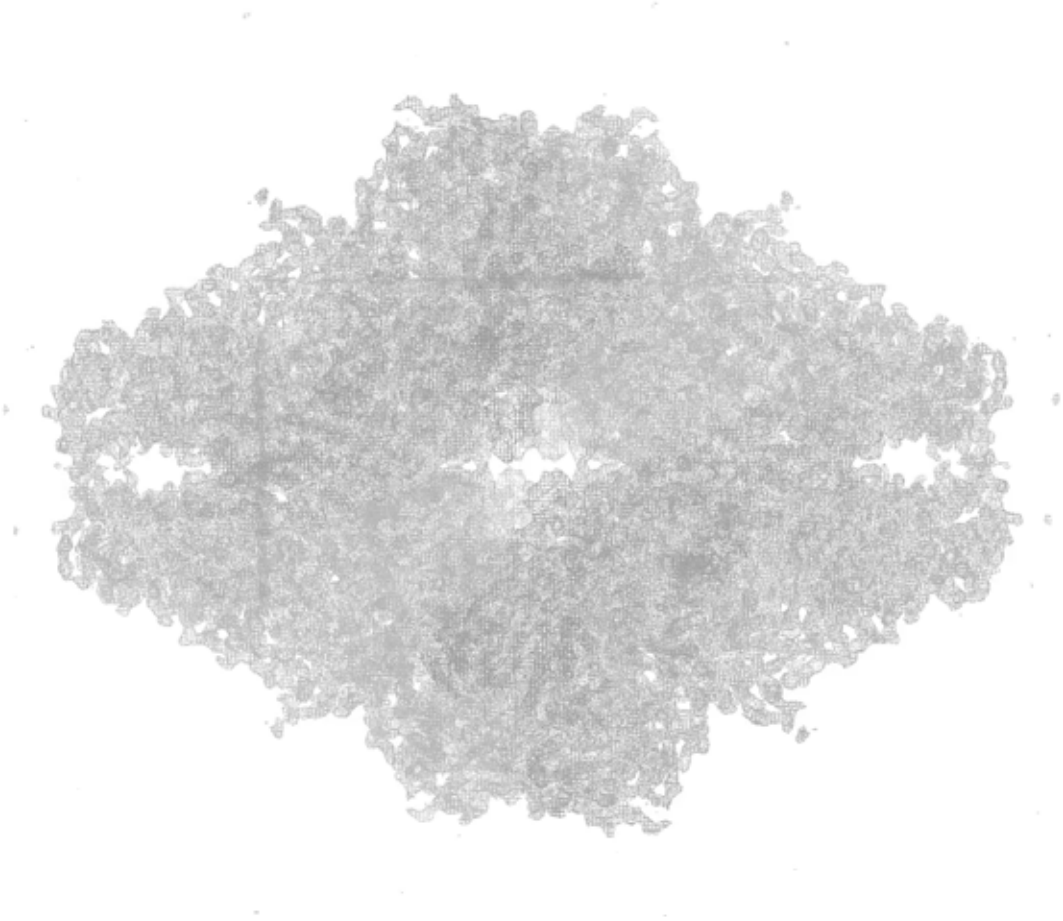
processed map

ARP/wARP for cryo-EM



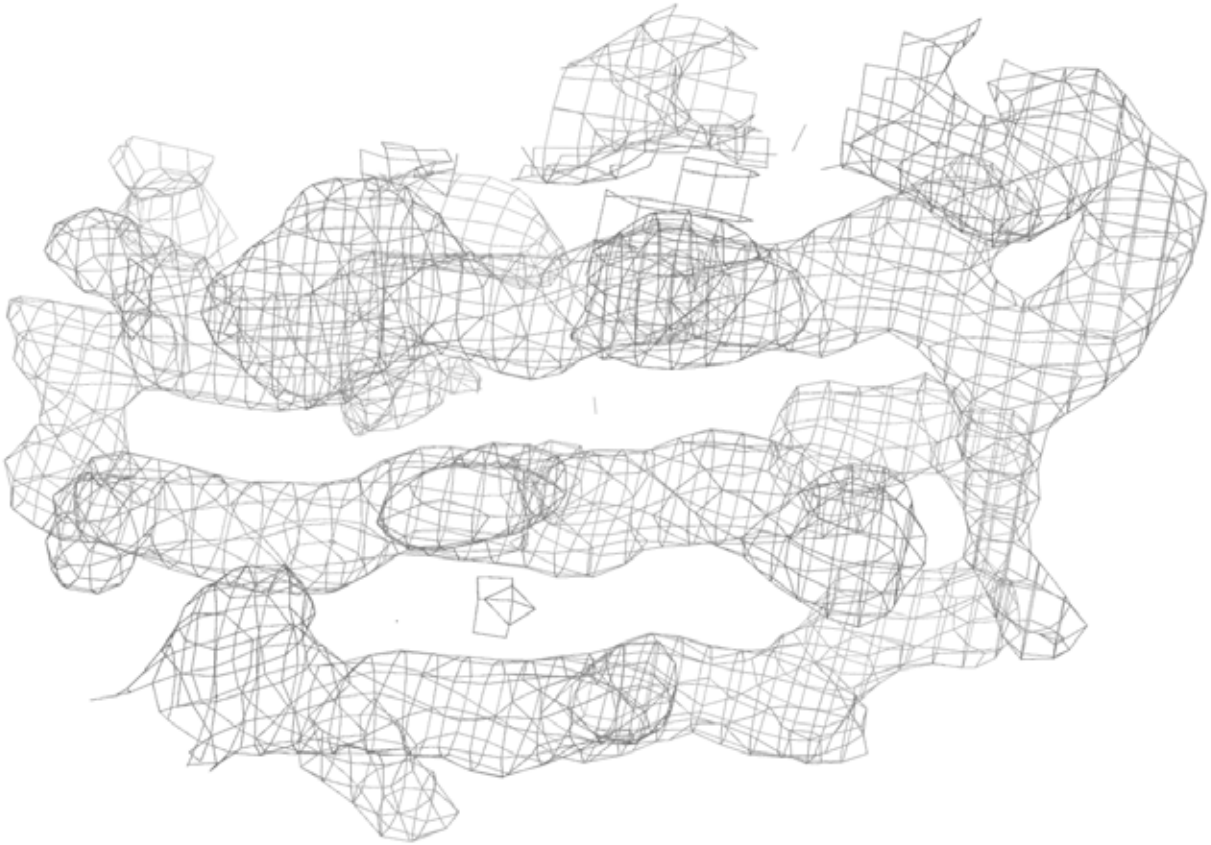
β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

ARP/wARP for cryo-EM



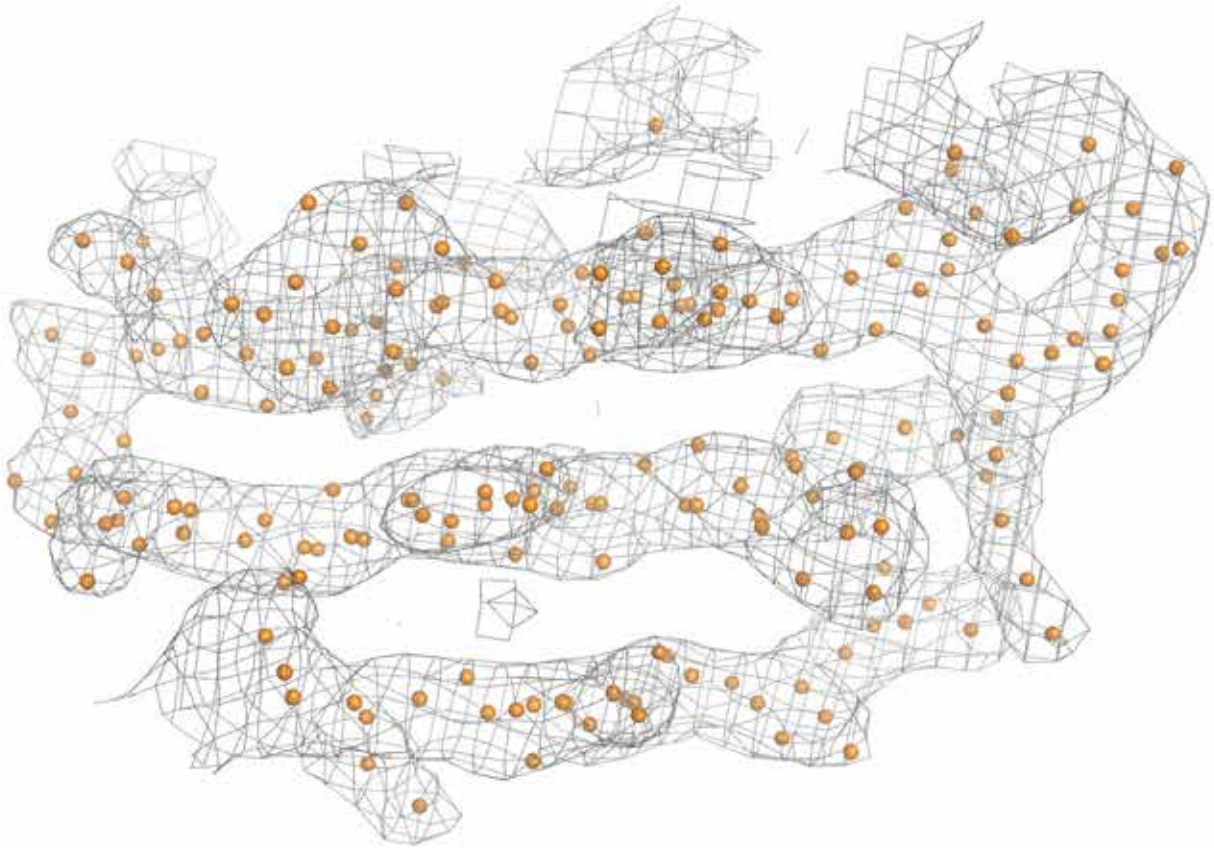
β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

ARP/wARP for cryo-EM



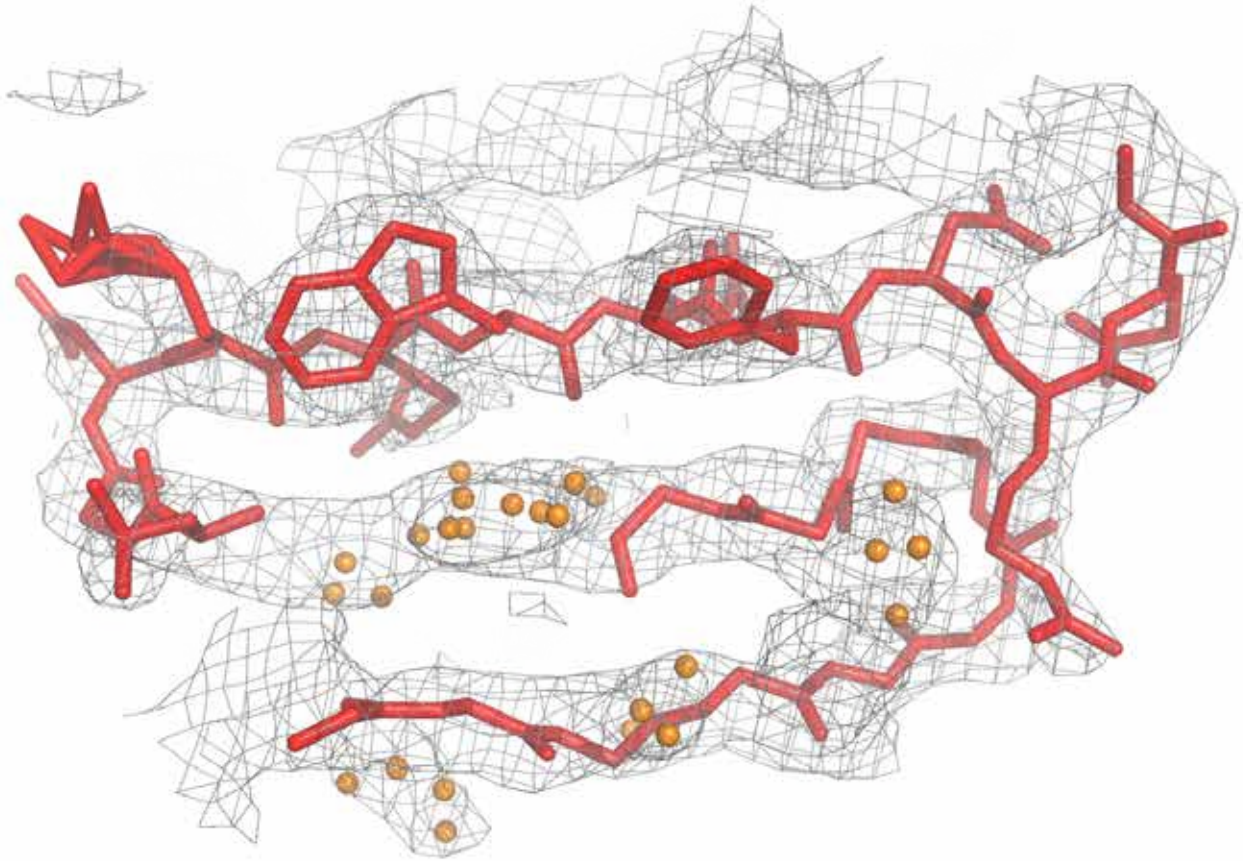
β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

ARP/wARP for cryo-EM



β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

ARP/wARP for cryo-EM

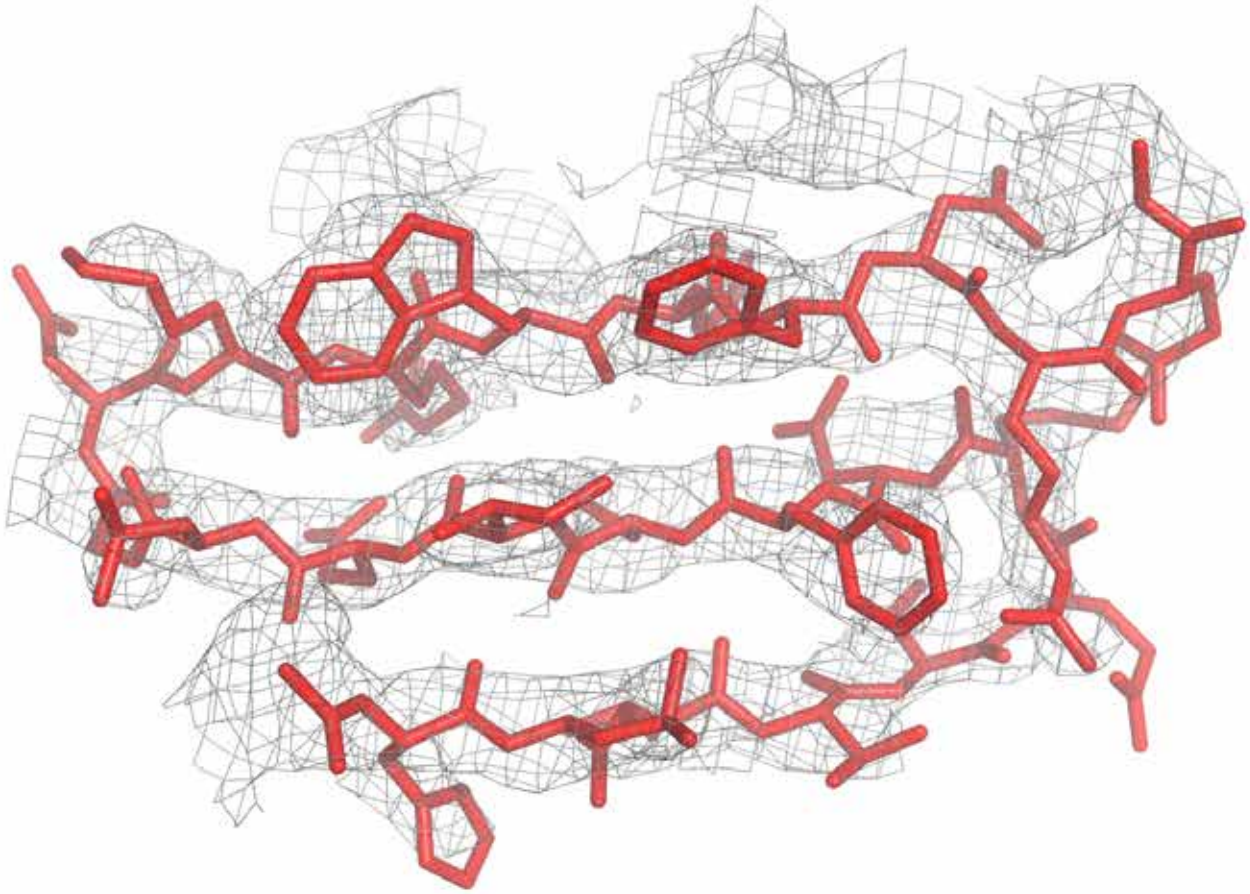


β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

ARP/wARP for cryo-EM

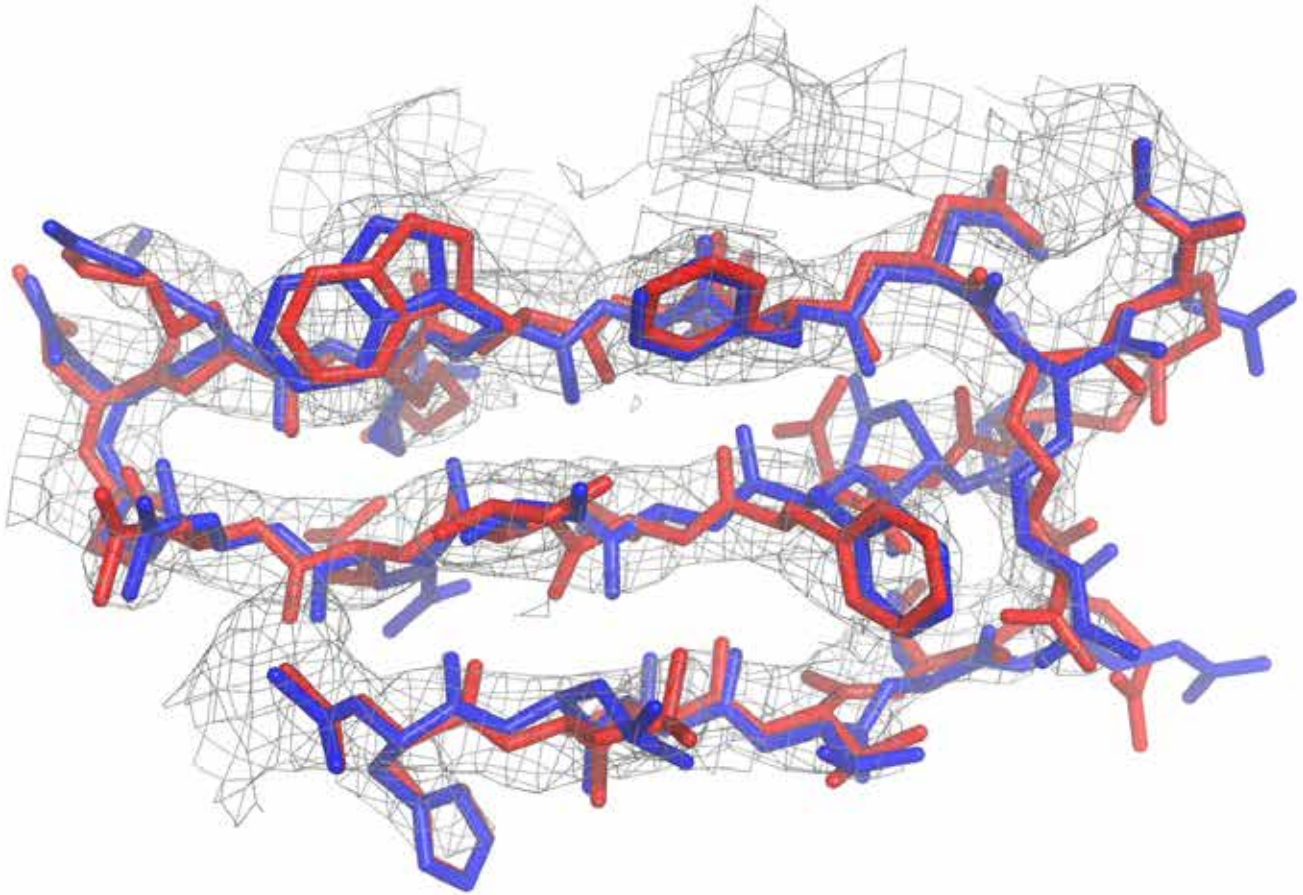
β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

ARP/wARP for cryo-EM



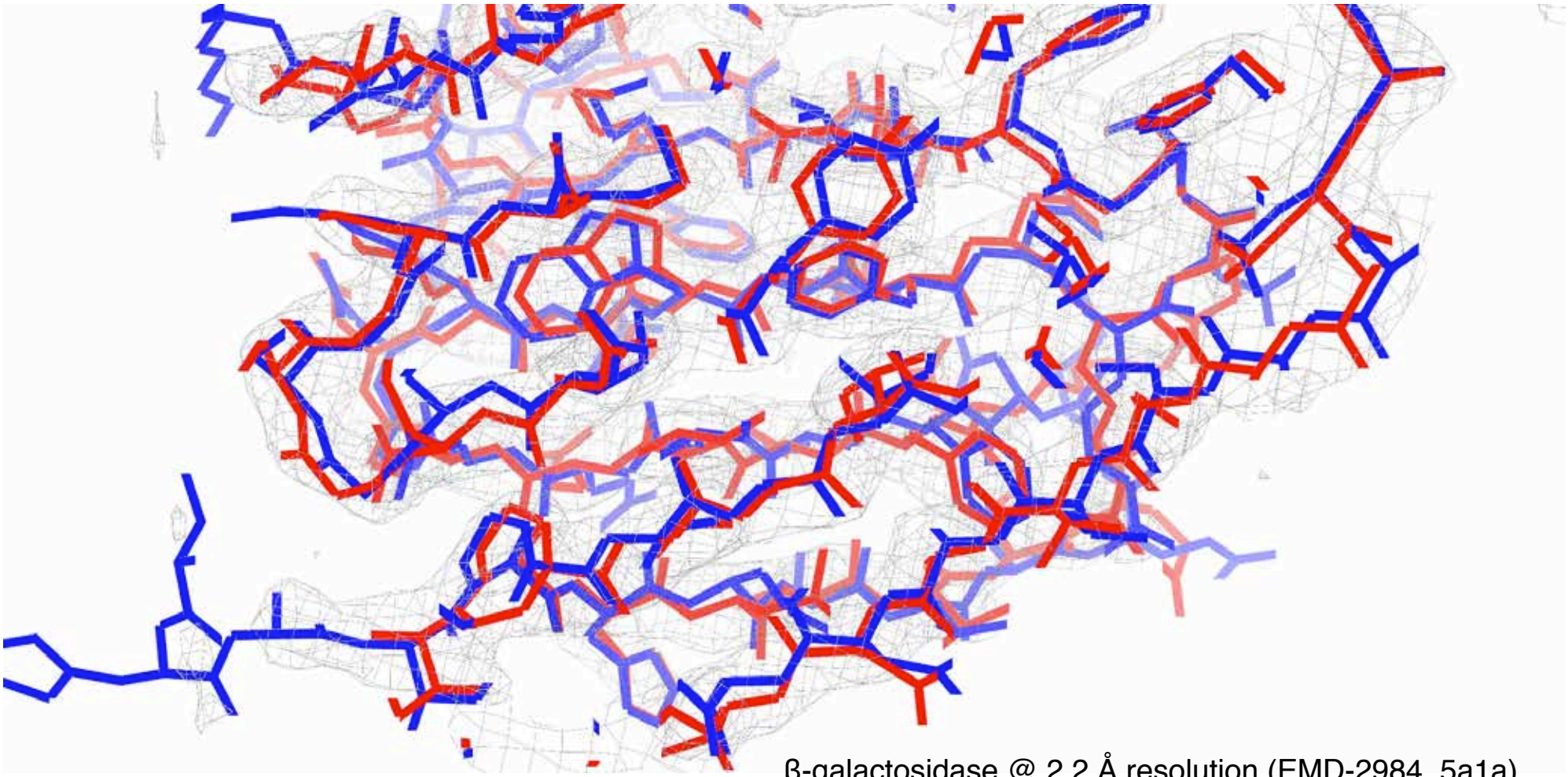
β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)

ARP/wARP for cryo-EM



β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)
ARP/wARP model
deposited coordinates

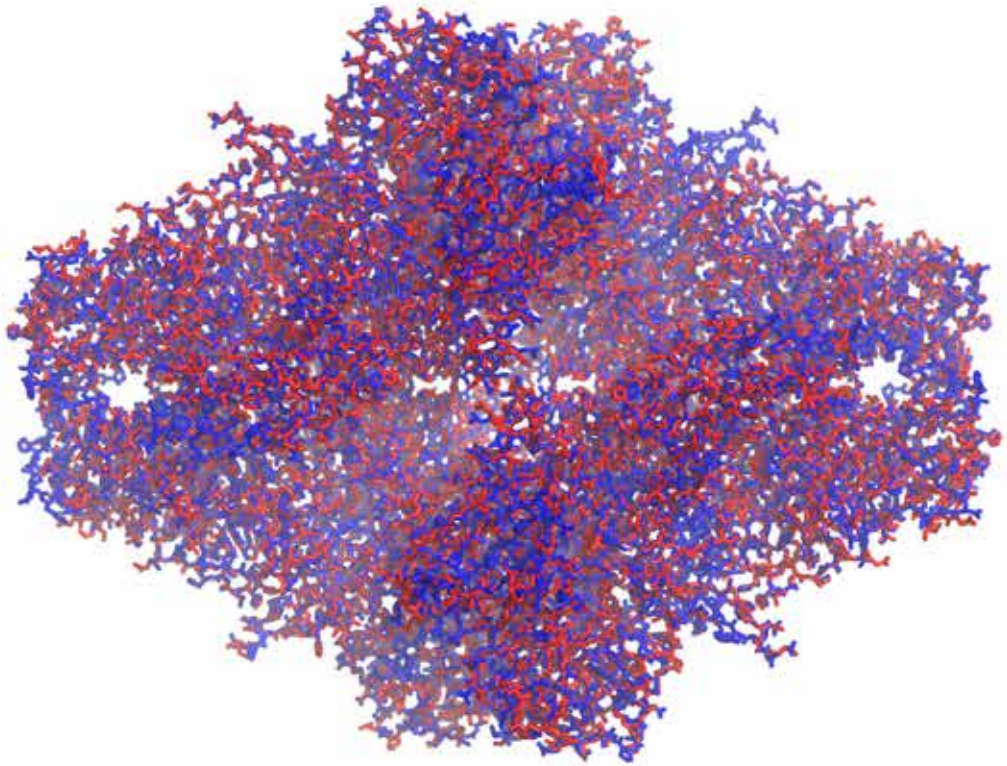
ARP/wARP for cryo-EM



β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)
ARP/wARP model
deposited coordinates

ARP/wARP for cryo-EM

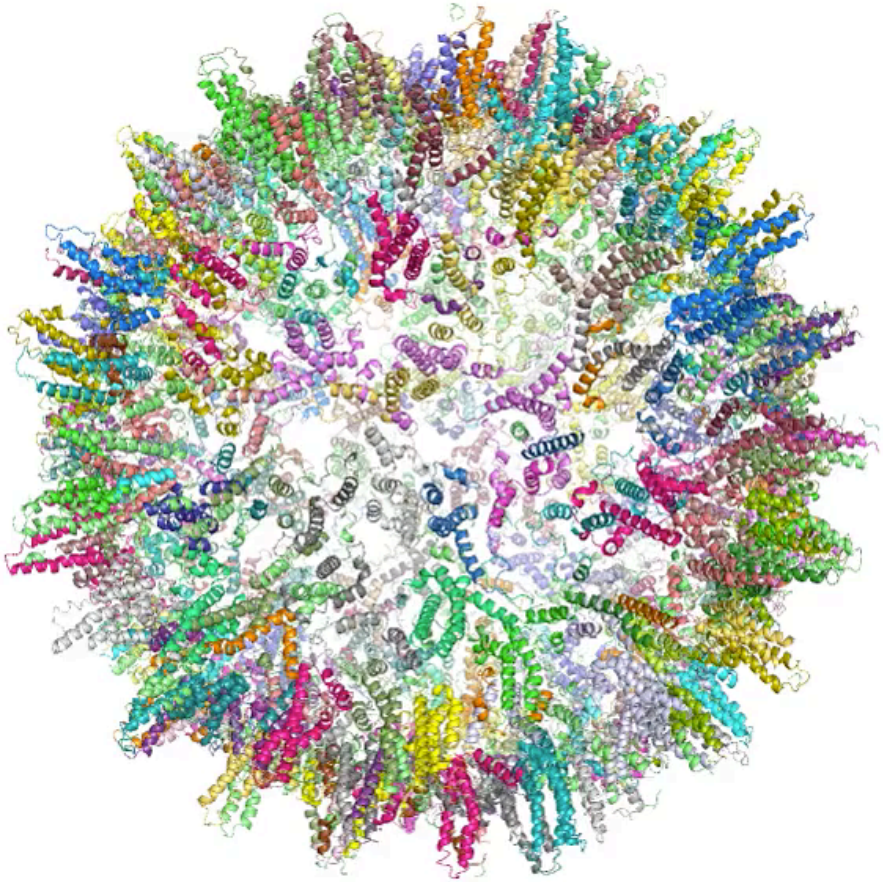
ARP/wARP model (30h of CPU time, first model after 6h)
~ **98%** (4006 out of 4088) of the residues traced and docked



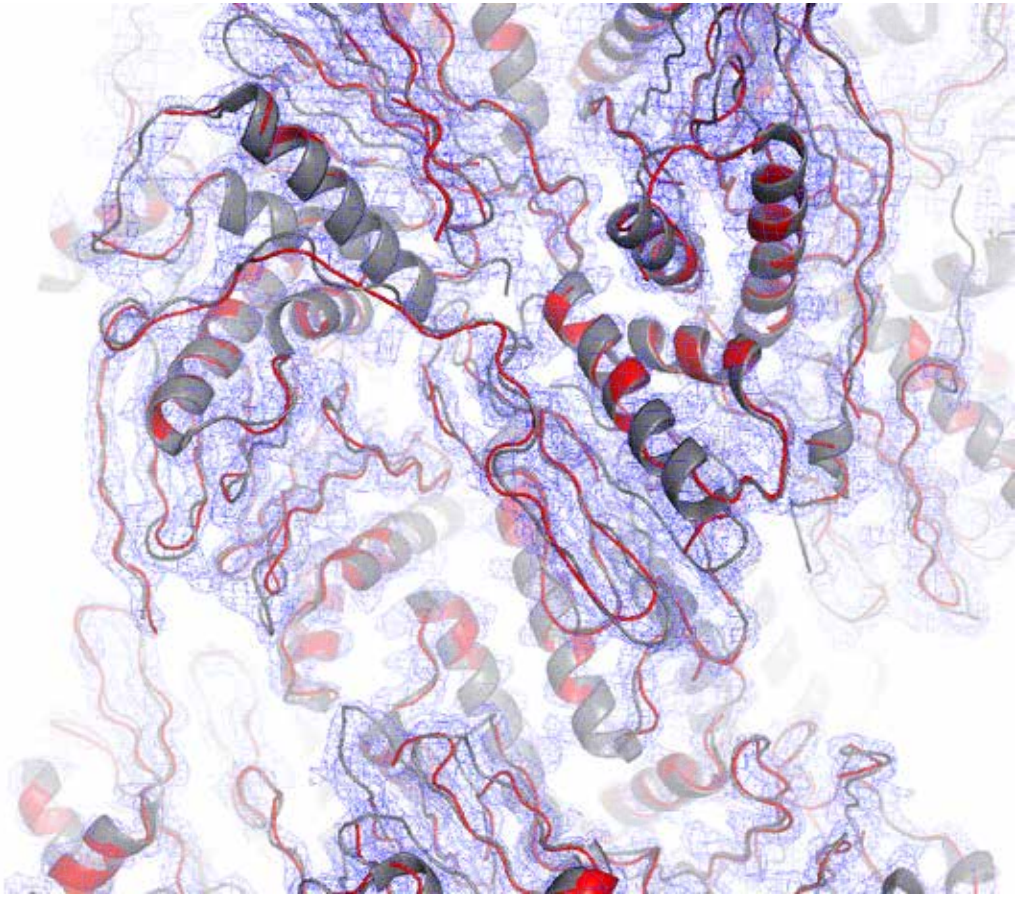
β-galactosidase @ 2.2 Å resolution (EMD-2984, 5a1a)
ARP/wARP model
deposited coordinates

ARP/wARP for cryo-EM

Full model of a viral capsid ready in 48h CPU time
75% of 41,000 residues traced



26S proteasome at 4.0 Å resolution

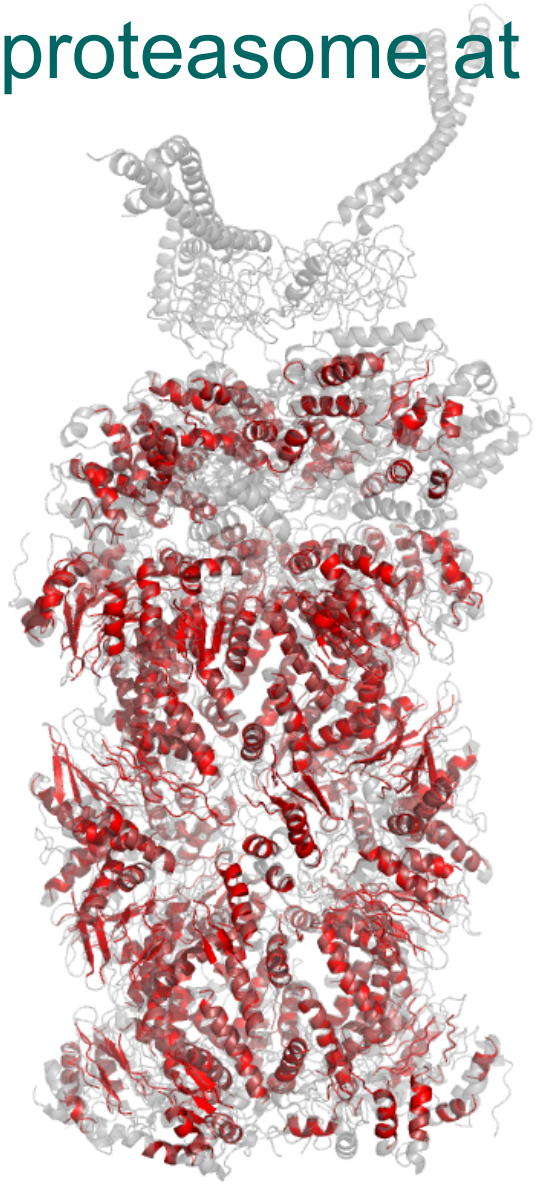


After ARP/wARP (48 hours of CPU time, first model after 6h)

61% residues correct
(3619/5963 CA_{ARP}-CA_{PDB} < 2Å)
27% reversed
(1671/5963 CA_{ARP}-CA_{PDB} < 2Å)

ARP/wARP model / deposited coordinates
26S proteasome @ 4.0 Å resolution (5l4g/EMD-4002)

26S proteasome at 4.0 Å resolution



ARP/wARP model / deposited coordinates
26S proteasome @ 4.0 Å resolution (5l4g/EMD-4002)

ARP/wARP secondary structure modelling

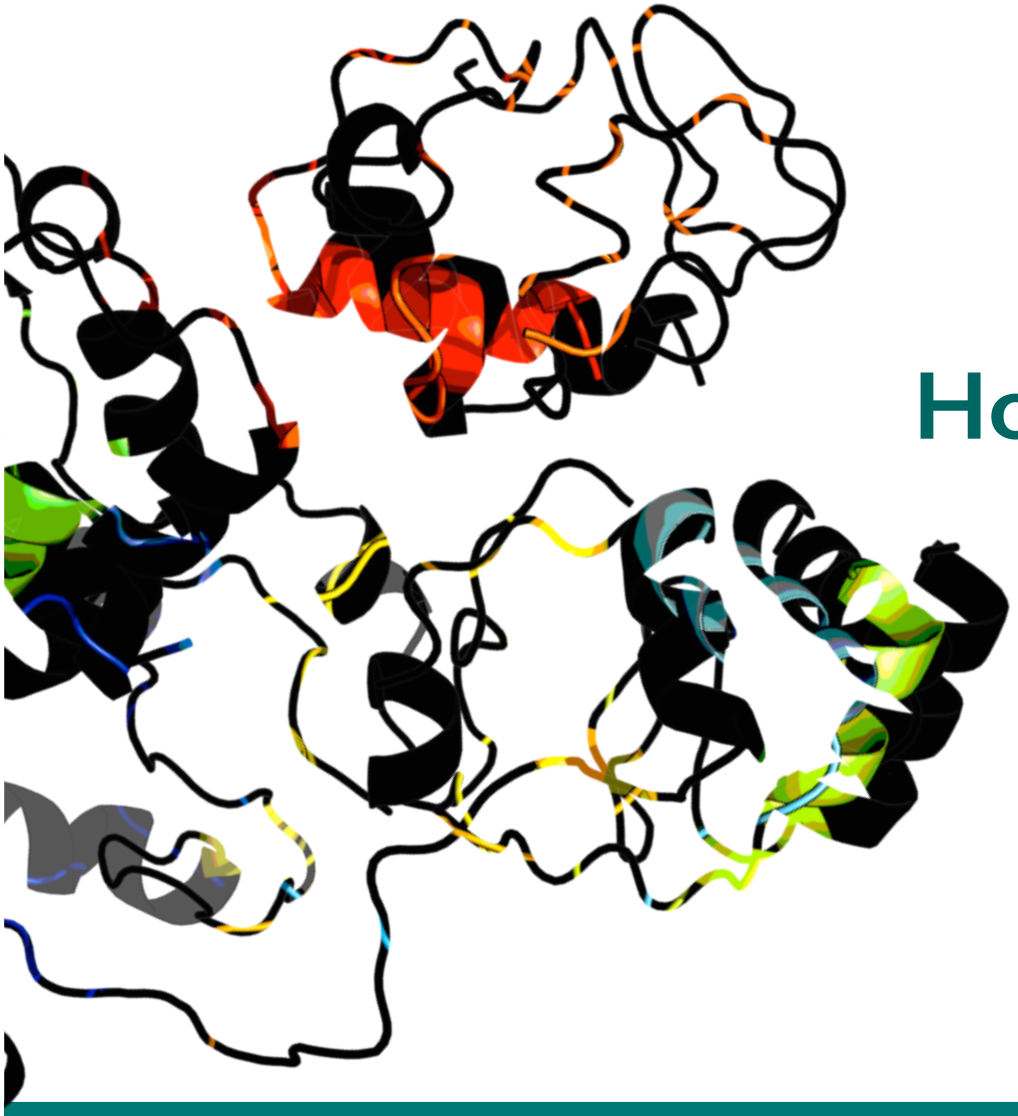
5 minutes on a single CPU

42% residues correct
(2506/5963 CA_{ARP}-CA_{PDB} < 2Å)
15% reversed
(991/5963 CA_{ARP}-CA_{PDB} < 2Å)

The initial model can be further processed with ARP/wARP tools:

- **automated loop building**
- **sequence docking**
- **hybrid real/reciprocal space refinement**

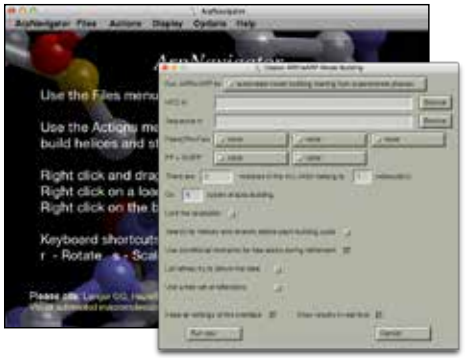
(uses Refmac with ProSMART)



How to use ARP/wARP?

EMBL

How do I use ARP/wARP?



ArpNavigator

Intuitive interaction with automated model building

Real time visualisation of model building

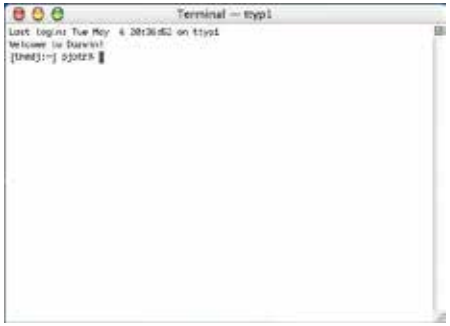
Displays the intermediate steps

CCP4 Suite

Model building job saved in your project directory

Extra options for automated model building

Job submission to Hamburg cluster



Terminal

Good for expert users

Complete control of the parameters to use

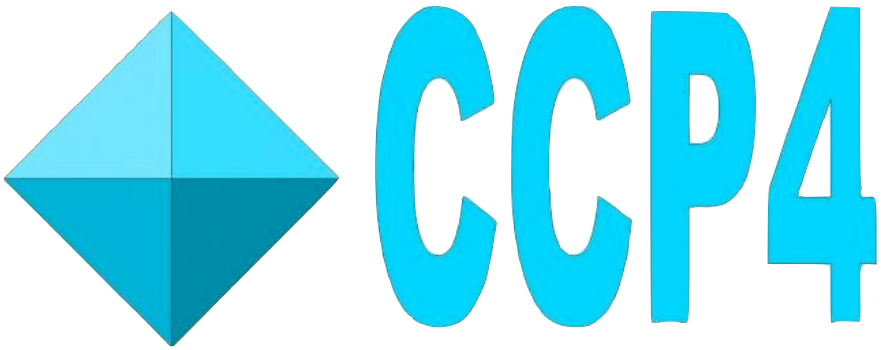
Check user guide for the commands

Web Service

Complete ARP/ wARP interface

Convenient for parallel computing

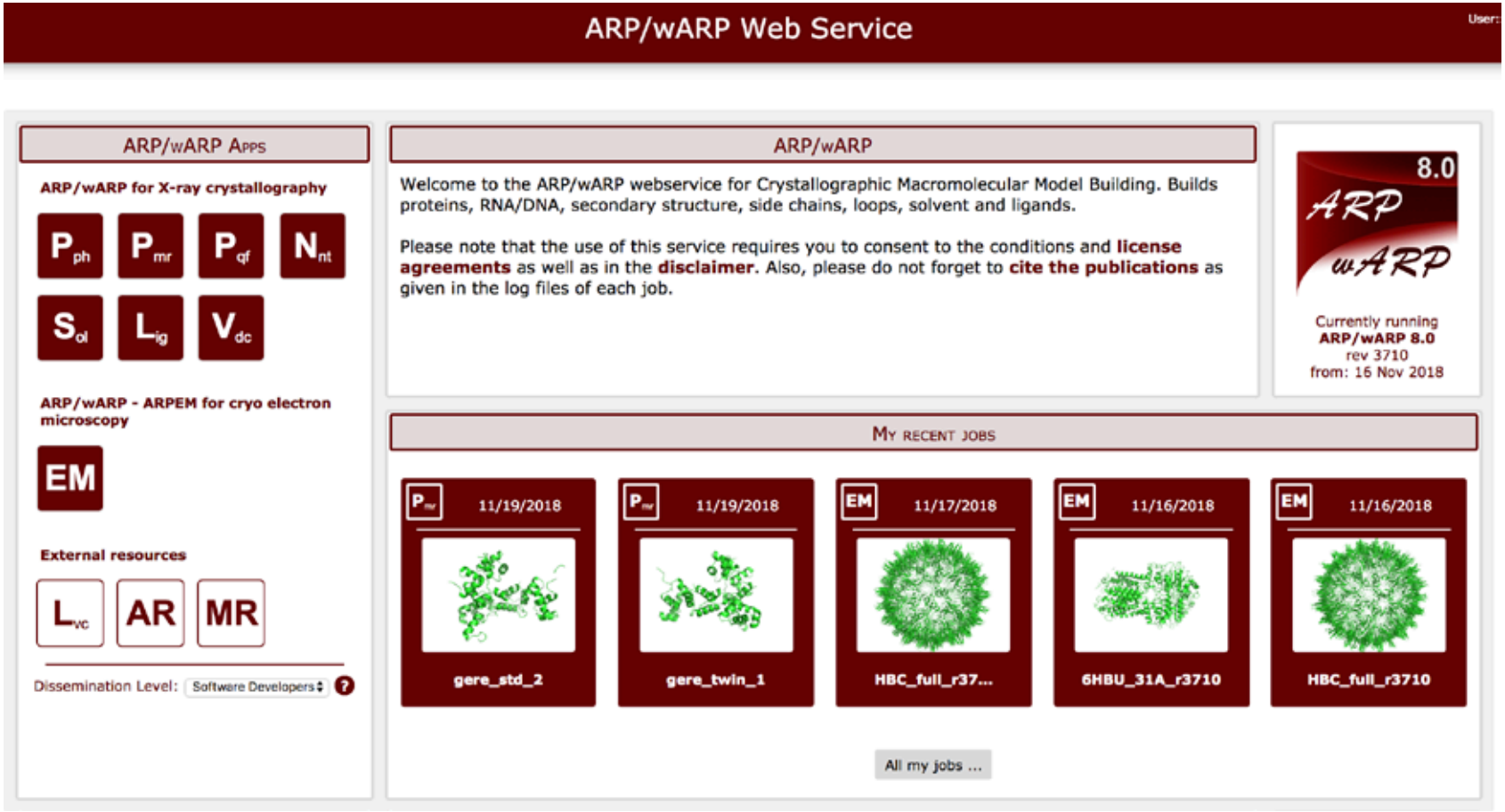
Where can I find ARP/wARP?



Joint release - one installer, no worries!

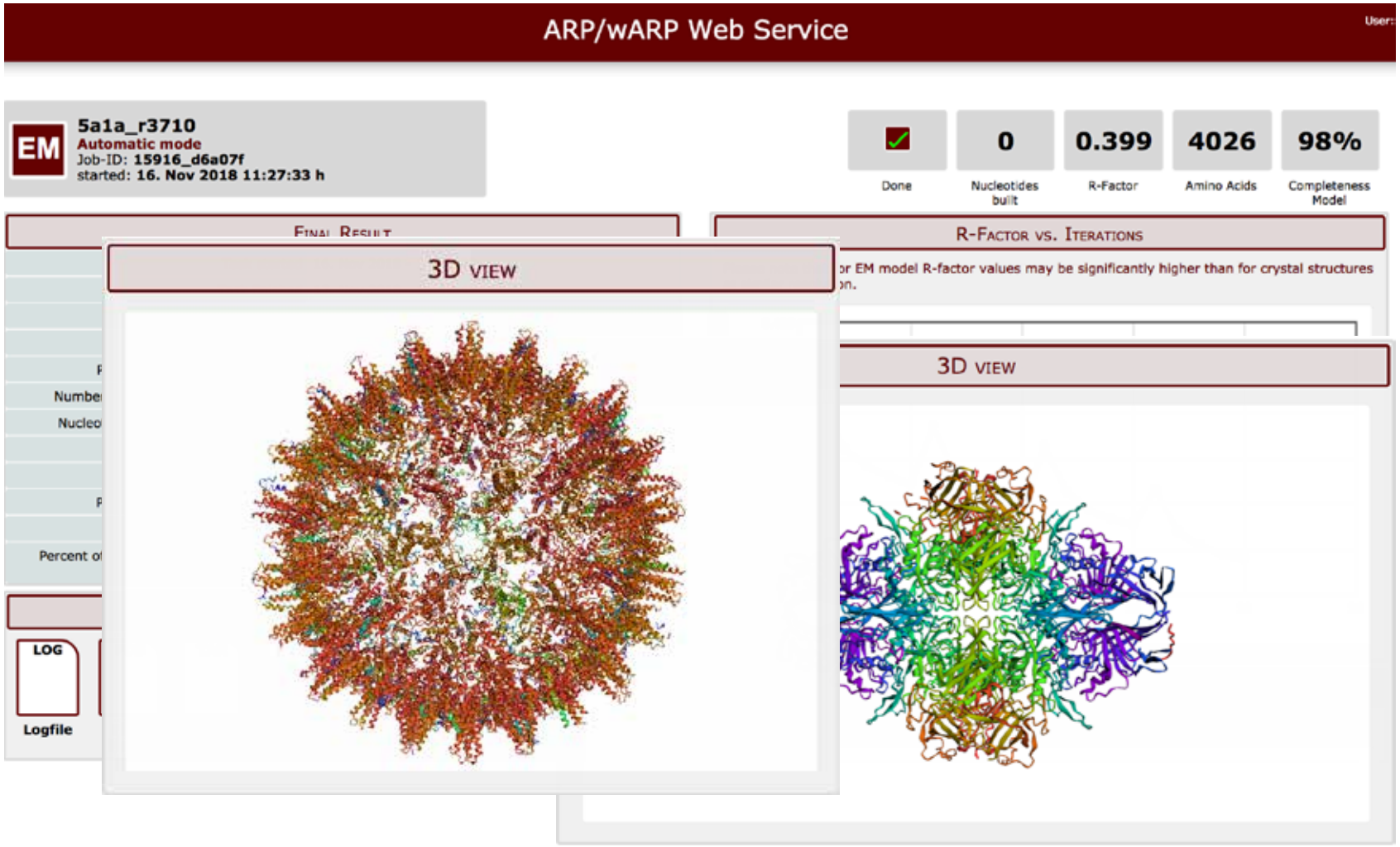
Running ARP/wARP 8.0 on-line

https://arpwarp.embl-hamburg.de/



Running ARP/wARP 8.0 on-line

https://arpwarp.embl-hamburg.de/



Running ARP/wARP 8.0 on-line

Auto-Rickshaw
X-ray structure determination server

MoRDa
Molecular Replacement pipeline

mrBUMP
Molecular Replacement pipeline

BALBES
Molecular Replacement pipeline

PDB_REDO
X-ray model optimisation server

DipCheck
Model geometry validation server

Vici
In-silico ligand-based drug design

Developed within West-Life

ARP/wARP team @ EMBL-Hamburg

Collaborators

MRC Laboratory, Cambridge :

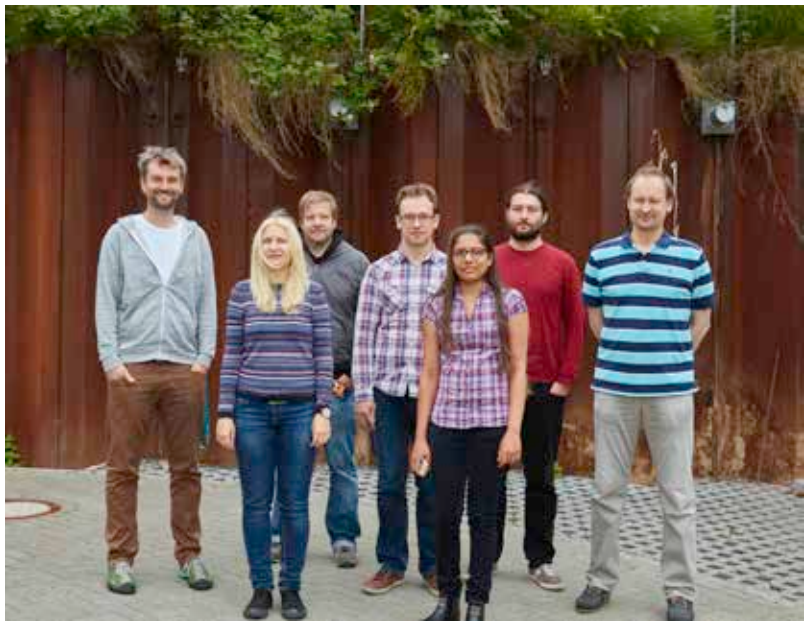
Garib Murshudov's group

Rutherford Appleton Laboratory:

The core CCP4 group

EMBL HH, HD, EMBL EBI:

Thomas Schneider, Carsten Sachse, John Briggs, Gerard Kleywegt, Sameer Velankar, Janet Thornton



Group members

Victor Lamzin
Umut Oezugurel
Sravya Kantamneni
Egor Sobolev
Koushik Choudhury

Former ARP/wARP developers

Philipp Heuser, Daria Beshnova, Ciarán Carolan, Serge Cohen, Zbyszek Dauter, Helene Doerksen, Guillaume Evrard, Francisco Fernandez, Johan Hattne, Saul Hazledine, Marouane Jelloul, Krista Joosten, Matheos Kakaris, Olga Kirillova, Gerrit Langer, Wijnand Mooij, Richard Morris, Venkat Parthasarathy, Joana Pereira, Anastassis Perrakis, Tilo Strutz, Ioan Vancea, Tim Wiegel, Keith Wilson, Peter Zwart

Funding:

Bundesministerium für Bildung und Forschung

West-Life