



Automated protein structure solution for weak SAD data

Pavol Skubak and Navraj Pannu

Biophysical Structural Chemistry,
Leiden University, The Netherlands

<http://www.bfsc.leidenuniv.nl/software/crank/>

Low resolution and/or weak anomalous signal SAD data sets

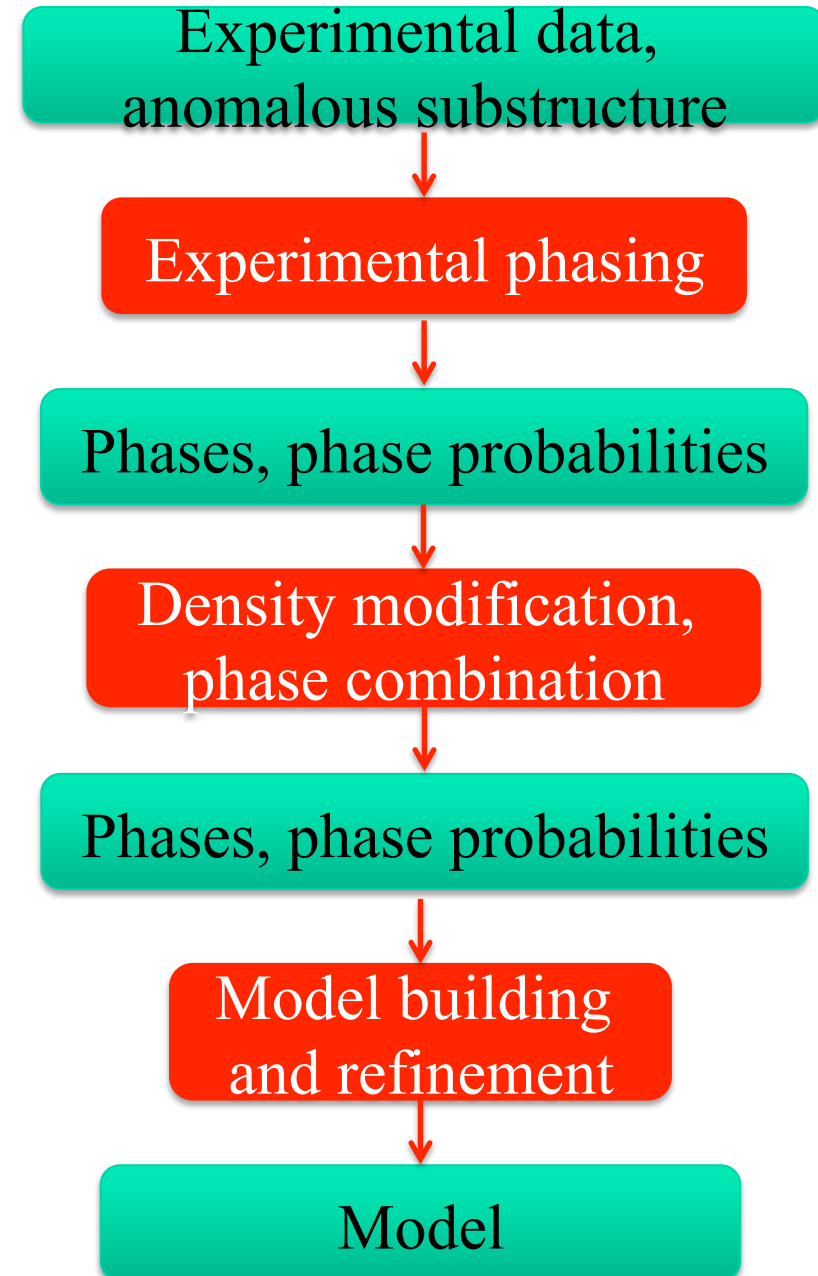
- With a sufficient anomalous signal and resolution better than 3 Angstroms, your structure is likely to be automatically built.
- What can be done if your SAD data is weak?

Simultaneously combining experimental phasing steps to improve structure solution

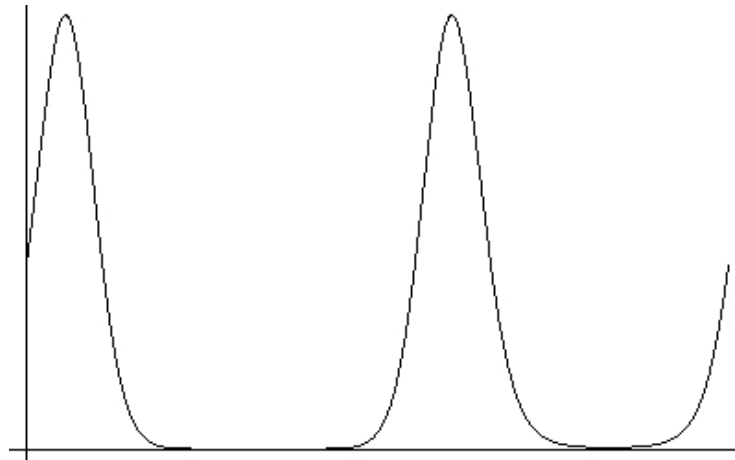
- Traditionally structure solution is divided into distinct steps:
 - Substructure detection
 - Obtain initial phases
 - (Density) modify the initial experimental map
 - Build and refine the model
- By combining these steps, we can improve the process.

Traditional structure solution

- Step-wise
- Information is propagated via 'phase probabilities'



Phase probabilities



$$P(\alpha) = \exp(A \cos(\alpha) + B \sin(\alpha) + C \cos(2\alpha) + D \sin(2\alpha))$$

- The phase distribution can be approximated via 4 “Hendrickson-Lattmann” coefficients, A, B, C, D.
- We rely on programs to estimate these coefficients.
- ‘Even better than the real thing?’

Phase probabilities

How do we choose the phase to use for our electron density map?

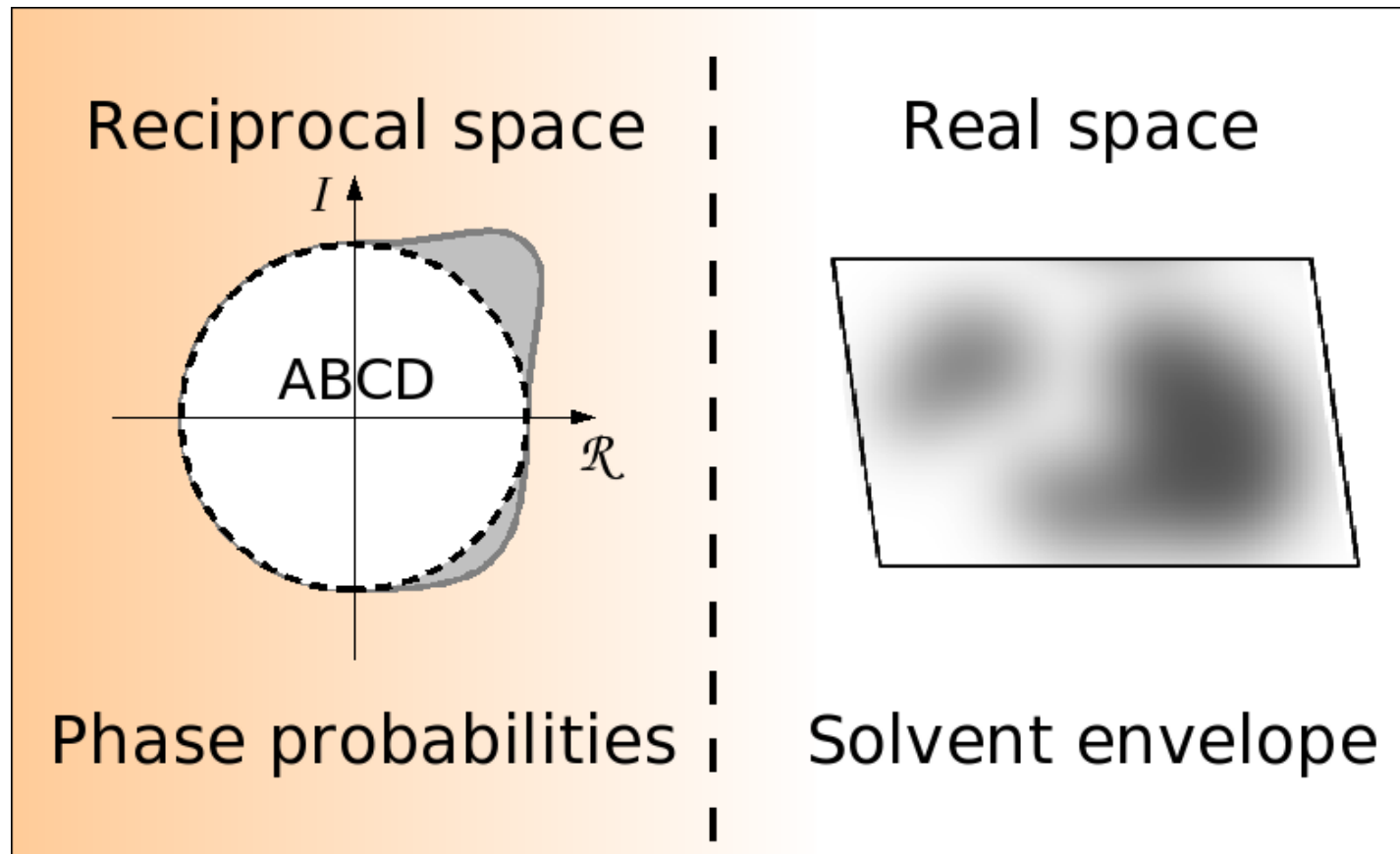
The “best phase” corresponds to mean/average/expected value (Blow and Crick).

What are ways to determine how good and accurate your phases are?

FOM: figure of merit: mean cosine of the phase error (a value of 1 means phases are perfect, a value of 0 means phases is the same as random).

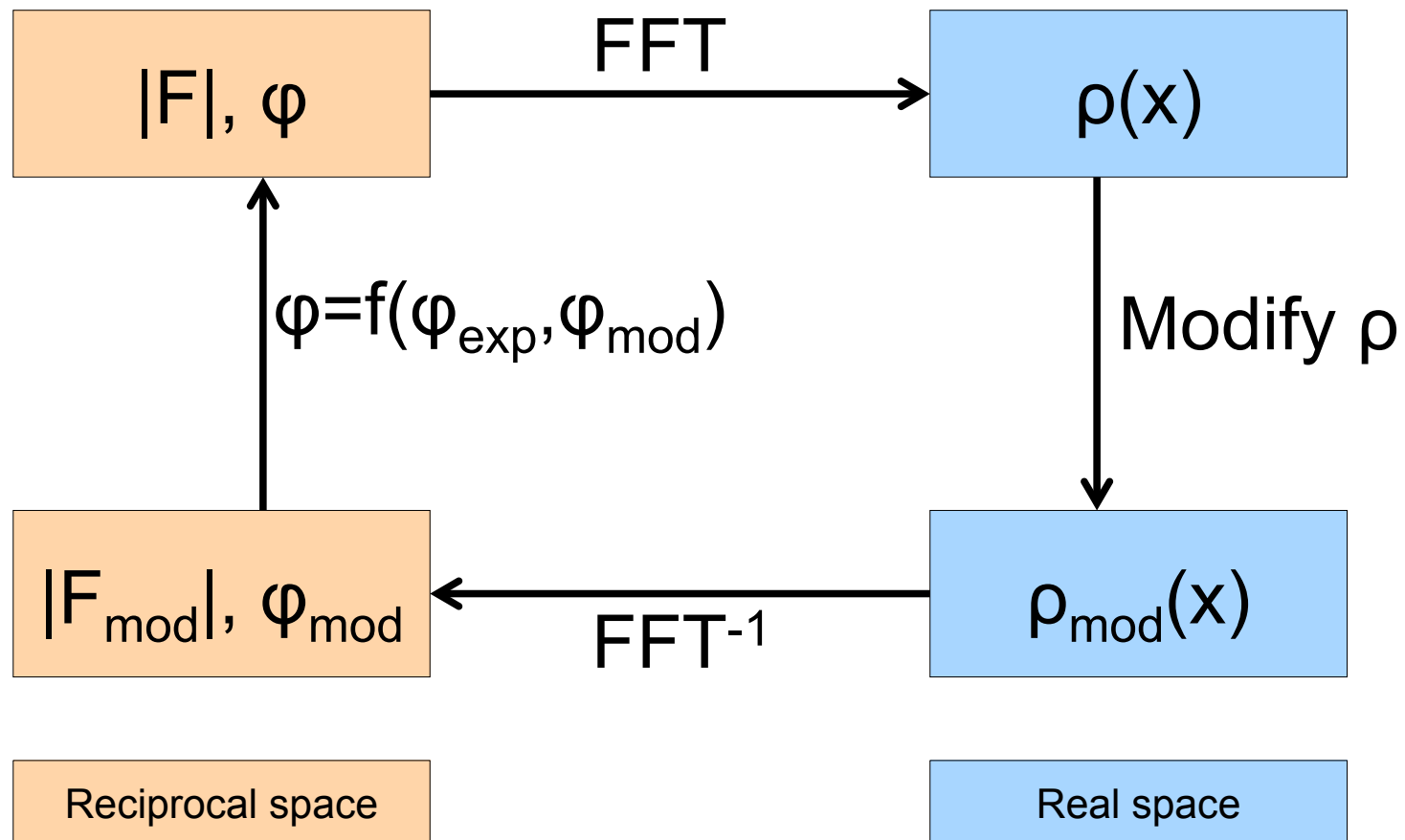
Density modification

- Density modification is a problem of combining information:



Density modification

2. Phase weighting:

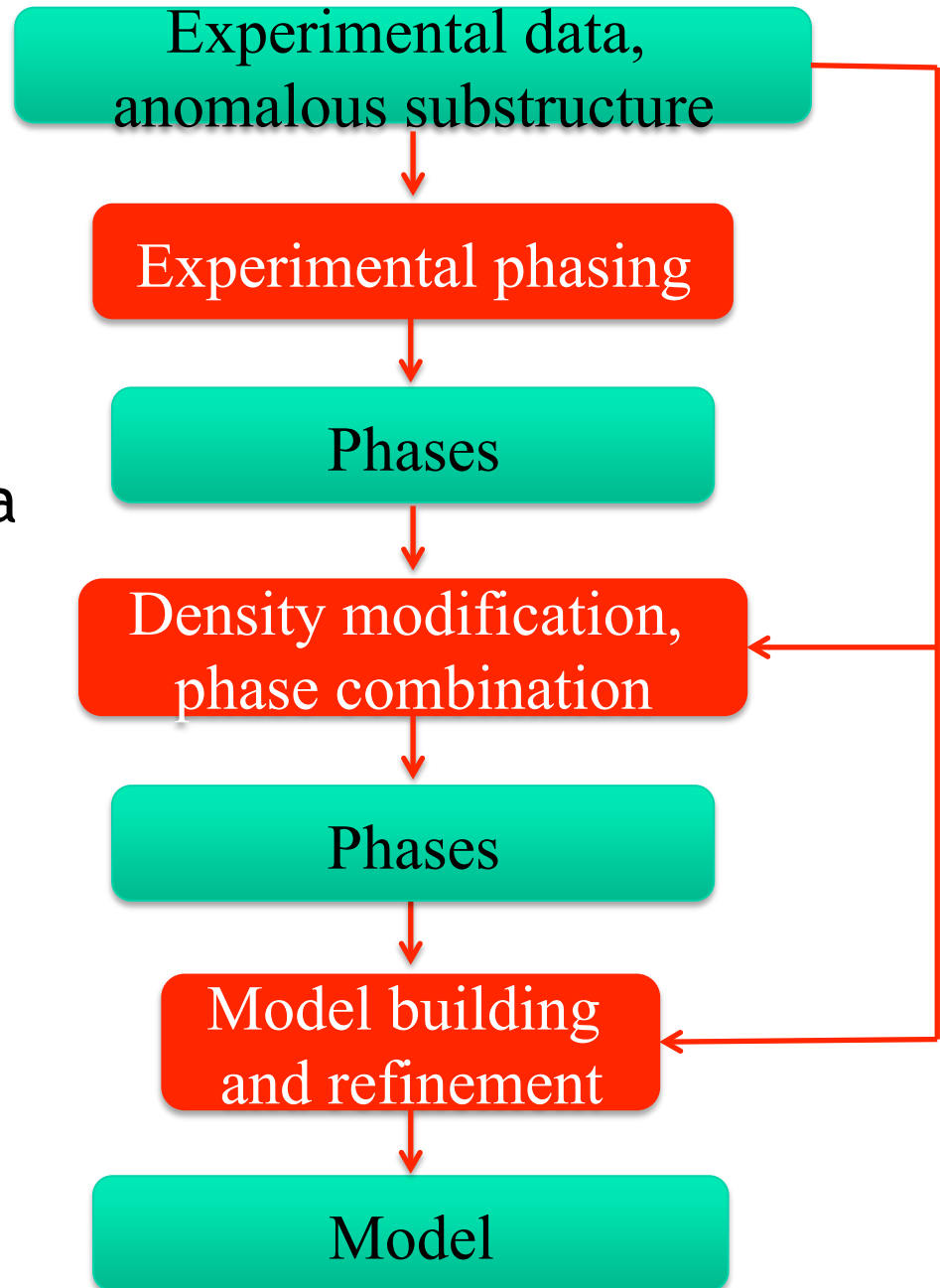


HL propagation and the independence assumption

- Use the experimental data and anomalous substructure directly!
 - Do not need to assume independence or rely on HL coefficients.
 - Need multivariate distributions at each step that take into account correlations between the model and data.

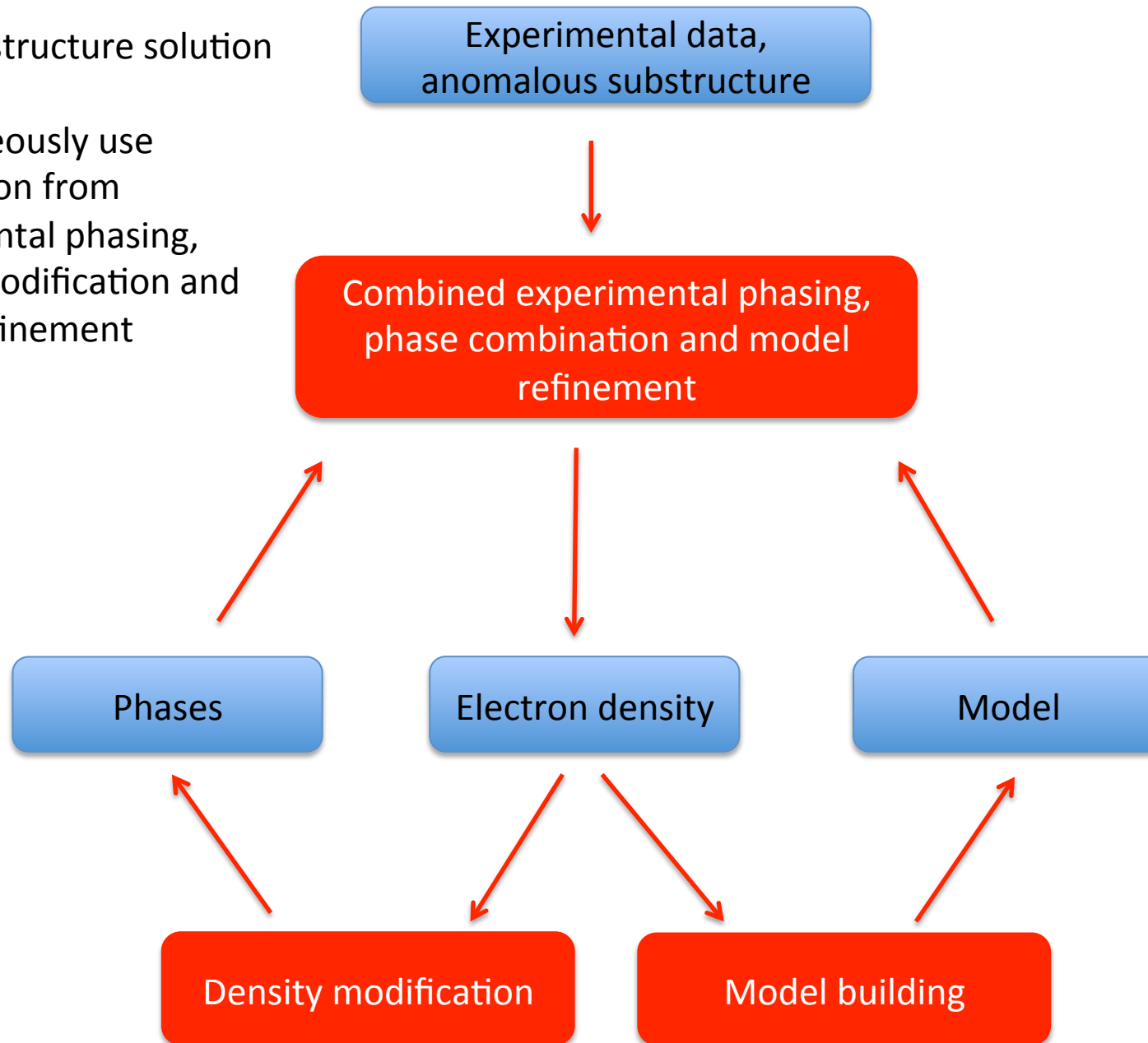
Step-wise multivariate structure solution

- Still step-wise
- Information is propagated via the data and model(s).



Combined structure solution

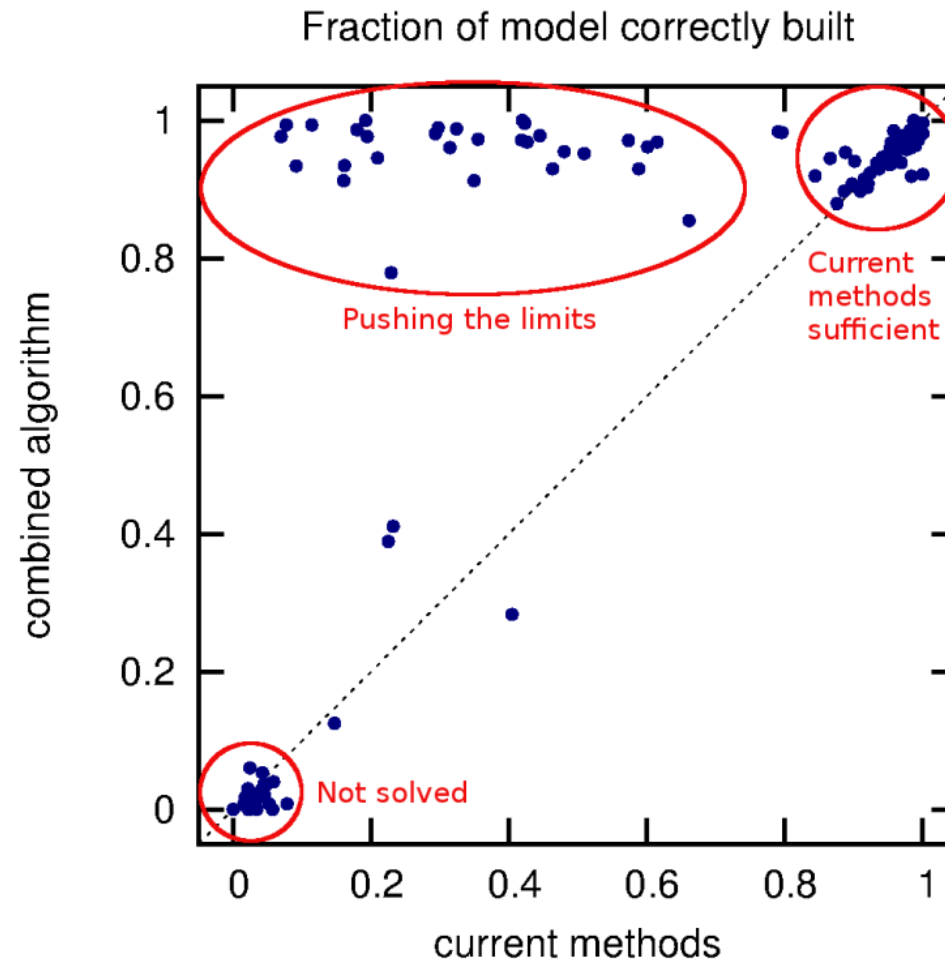
- Simultaneously use information from experimental phasing, density modification and model refinement



Tests of > 140 real SAD data sets

- Resolution range of data sets is 0.94 to 3.88 Angstroms
- Types of anomalous scatterers: selenium, sulfur, chloride, iodide, bromide, calcium, zinc (and others).
- We compare with the step wise multivariate approach (current CRANK) versus the combined approach.

Model building results on over 140 real SAD data sets (using parrot and buccaneer)

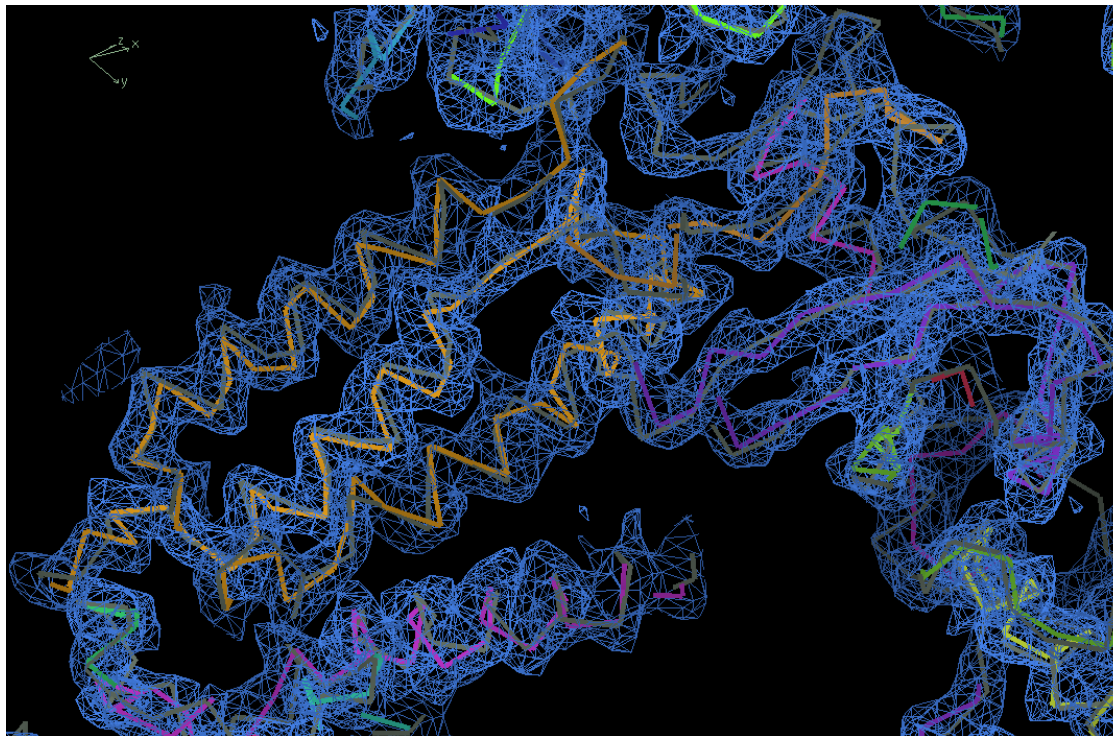


Summary of large scale test

- The average fraction of the model built increased from 60% to 74% with the new approach.
- If we exclude data sets built to 85% by the current approach or where the substructure was not found, 45 data sets remain and the average fraction of the model built increased from 28 to 77%.

3.88 Angstrom RNA polymerase II

- 3.88 Angstrom SAD data with signal from zinc.
- Authors could not solve the structure with SAD data alone, but with a partial model, multi-crystal MAD and manual building.
- > 80% can be built with SAD data alone with the new algorithm automatically to an R-free of 37.6%

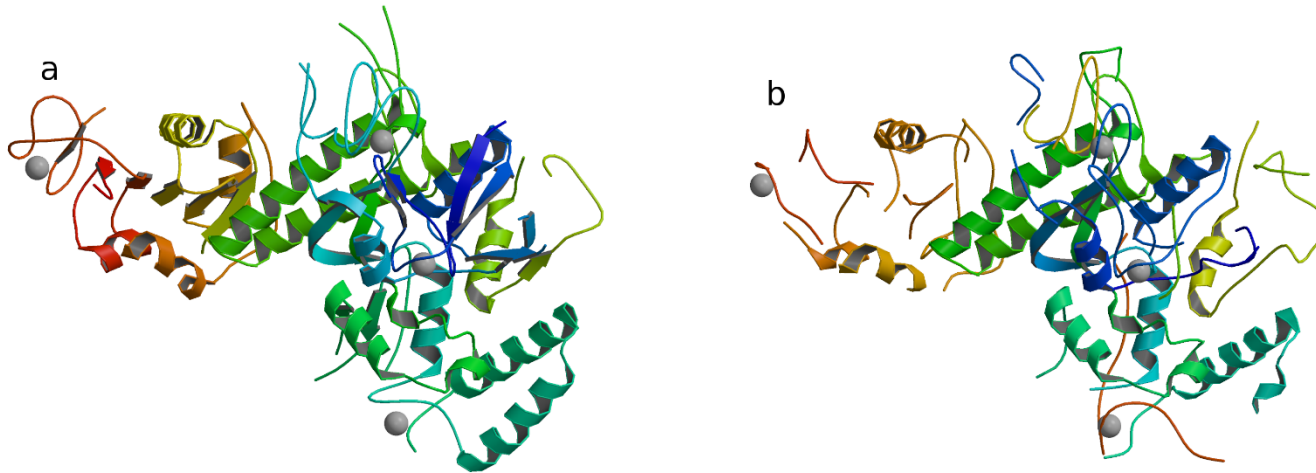


4.5 Angstrom ATPase SecA-SecY complex

- 4.5 Angstrom data set with weak anomalous signal from Se-Met SecY.
- Authors could not solve the structure with SAD data alone, but used a partial MR structure, (2-fold NCS averaging), cross-crystal averaging, and manual model building
- > 60% can be built automatically starting just from selenium positions (obtained from partial MR solution)
- R-free obtained was under 40%.
- If we start with the partial MR structure, results are worse!

Related RNA polymerases complex from Cramer et al.

- 3.3 Angstrom data with signal from zinc.
- Could not solve the structure with anomalous data alone.
- With the new method, a majority can be built automatically in minutes.



4.6 Angstrom SKI2-3-8 complex

- 4.6 Angstrom data set Se-Met data set.
- > 50% can be built automatically starting just from selenium positions.
- R-free obtained was under 40%.
- If we start with the partial MR structure, results are again worse! (MR models are higher resolution and fit 'fairly' well to the final model).

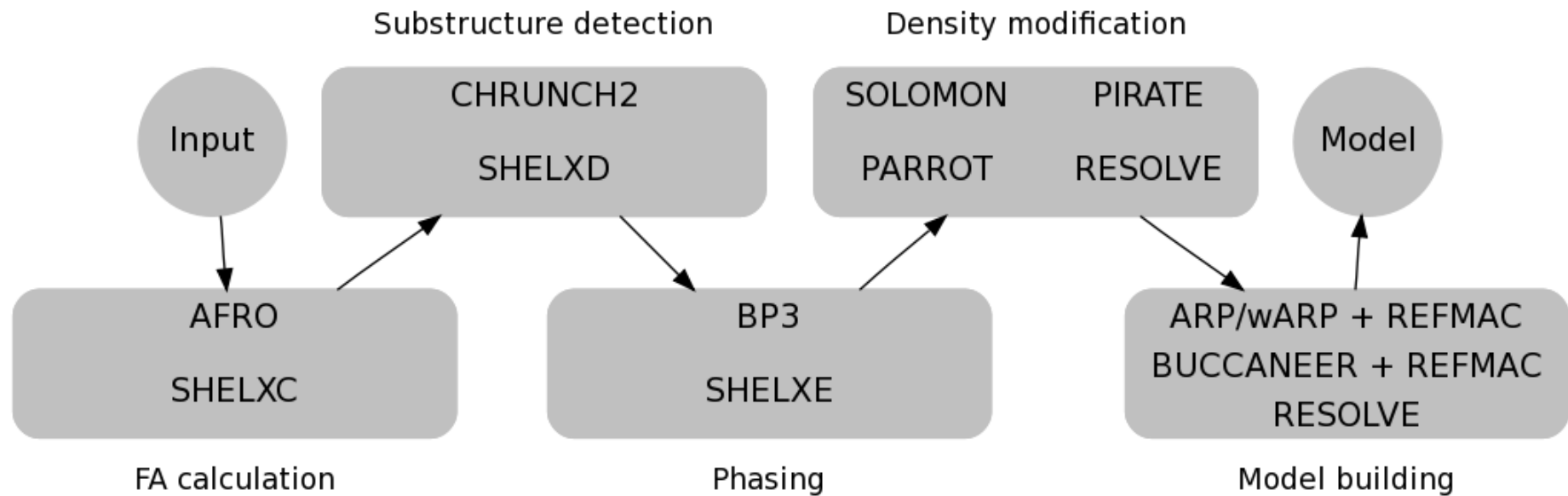
Future work on MR-SAD

- Are we throwing away useful information or is it biased?
- At the moment, if a (partial) MR solution is available, it is best to run *two* crank2 jobs:
 - Input the whole MR solution
 - Input just the heavy atoms

Crank1 vs Crank2

- Crank is suitable for S/MAD and S/MIRAS experiments and implements a stepwise multivariate function.
- Crank2 is its replacement that implements a combined multivariate function for SAD only.
- Both are available in CCP4 6.4.0 at the moment.

Programs in Crank



Important parameters in substructure detection

- The number of cycles run.
- The number of atoms to search for.
 - Should be within 10-20% of actual number
 - A first guess uses a probabilistic Matthew's coefficient
- The resolution cut-off:
 - For MAD, look at signed anomalous difference correlation.
 - For SAD, a first guess is $0.5 + \text{high resolution limit}$.

Is my map good enough?

- Statistics from substructure phasing:
 - Look at FOM from BP3.
 - For SAD, look at Luzzati parameters.
 - Refined occupancies.
- Statistics from density modification:
 - Compare the “contrast” from hand and enantiomorph (output of solomon or shelxe).
- Does it look like a protein? (model visualization)
- For Crank2, look to see if $R\text{-comb} < 40\%$.

References

- Crank
 - Ness et al (2004) Structure 12, 1753-1761.
 - Pannu et al (2011) Acta Cryst D67, 331-337.
- Combined approach and Crank2
 - Skubak and Pannu (2013) Nature Communications 4: 2777.
- Using data directly in refinement
 - Skubak et al (2004) Acta Cryst D60, 2196-2201.
 - Skubak et al (2009) Acta Cryst D65, 1051-1061.
- Multivariate phase combination
 - Waterreus et al (2010) Acta Cryst D66, 783-788.

Acknowledgements

- All dataset contributors (JCSG, Z. Dauter, M. Weiss, C. Mueller-Dieckmann)
- Garib Murshudov, Kevin Cowtan, George Sheldrick, Victor Lamzin
- <http://www.bfsc.leidenuniv.nl/software/crank/>



Cyttron

