Structure solution with Crank2

LUMC, the Netherlands

Pavol Skubak

Crank2

- software for automated structure solution: anomalous data → model.
- It requires minimal input, but is highly configurable.
- User friendly gui/pipelines incorporating our latest developments in substructure detection, phasing, density modification and model building & refinement as well as plugins to externally developed programs.

Crank2

- Several pipelines:
  - SAD pipeline using a multivariate SAD function in all steps
  - MR-SAD pipeline (for SAD after MR and for model rebuilding using anom. data)
  - SIRAS/MAD pipelines
  - SHELX pipeline via Crank2

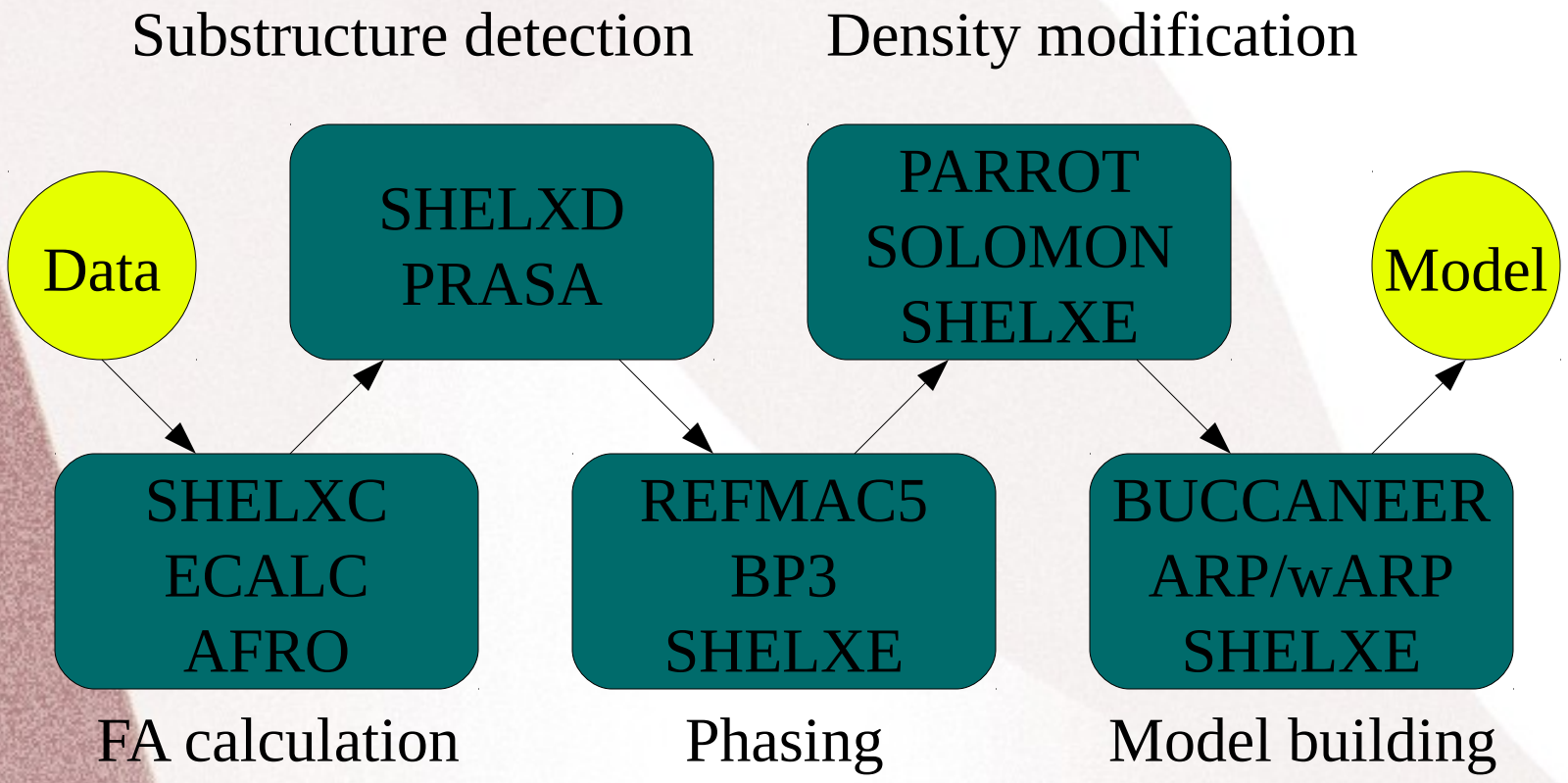
Availability

- ccp4i2 (from CCP 7.0):
  - CRANK2 for SAD,MR-SAD,SIRAS,MAD
- jsCoFe (CCP4 cloud):
  - CRANK2 for SAD,MR-SAD,SIRAS,MAD
- CCP4 Online:
  - CRANK2 for SAD,SIRAS,MAD
- ccp4i (CCP4 7.0):
  - CRANK2 for SAD and MR-SAD
  - CRANK for SIRAS, MAD

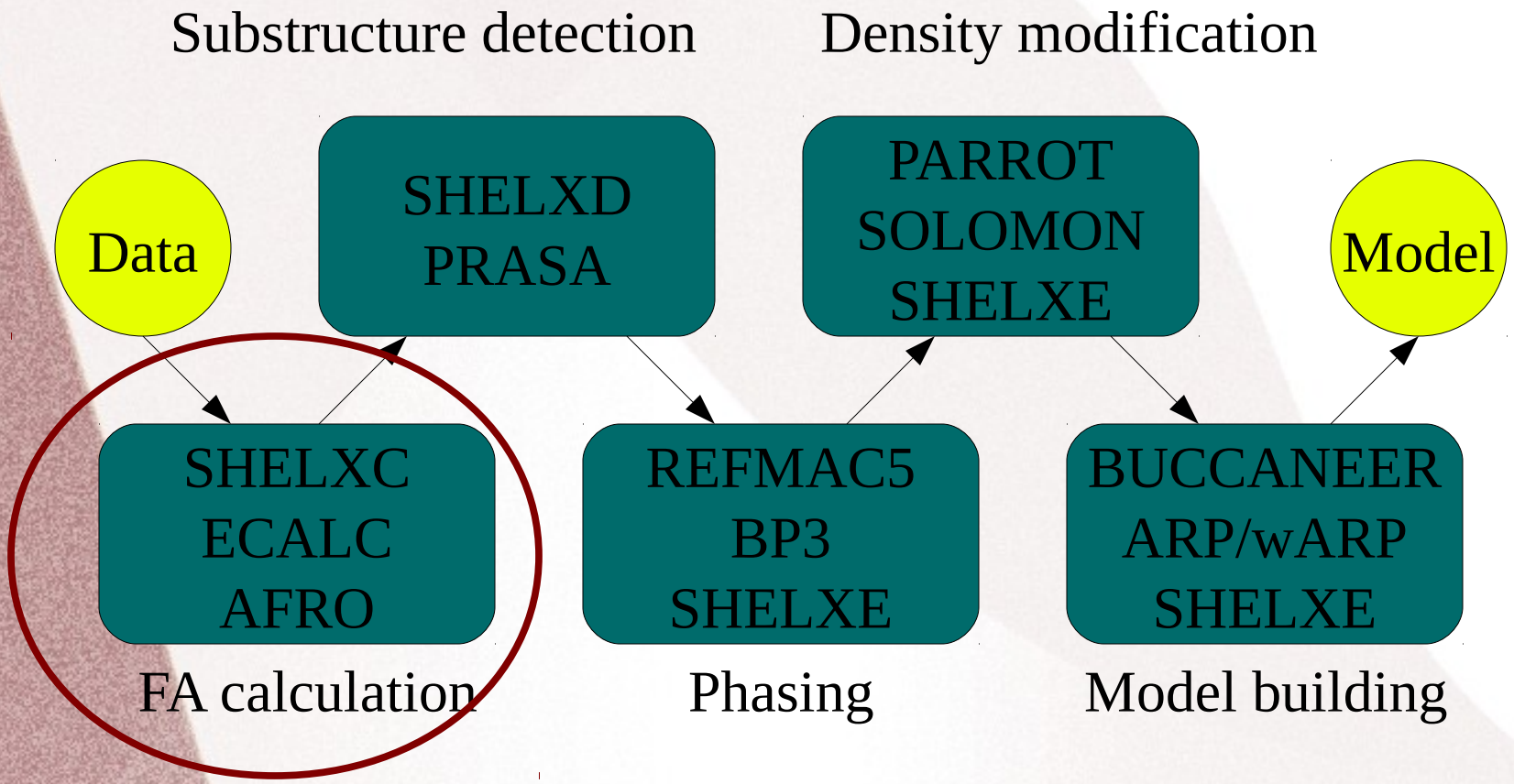
Availability

- ccp4i2 (from CCP 7.0):
  - CRANK2 for SAD,MR-SAD,SIRAS,MAD
- jsCoFe (CCP4 cloud):
  - CRANK2 for SAD,MR-SAD,SIRAS,MAD
- CCP4 Online:
  - CRANK2 for SAD,SIRAS,MAD
- ccp4i (CCP4 7.0):
  - CRANK2 for SAD and MR-SAD
  - CRANK for SIRAS, MAD

Structure solution from experimental phases with Crank2.



Structure solution from experimental phases with Crank2.

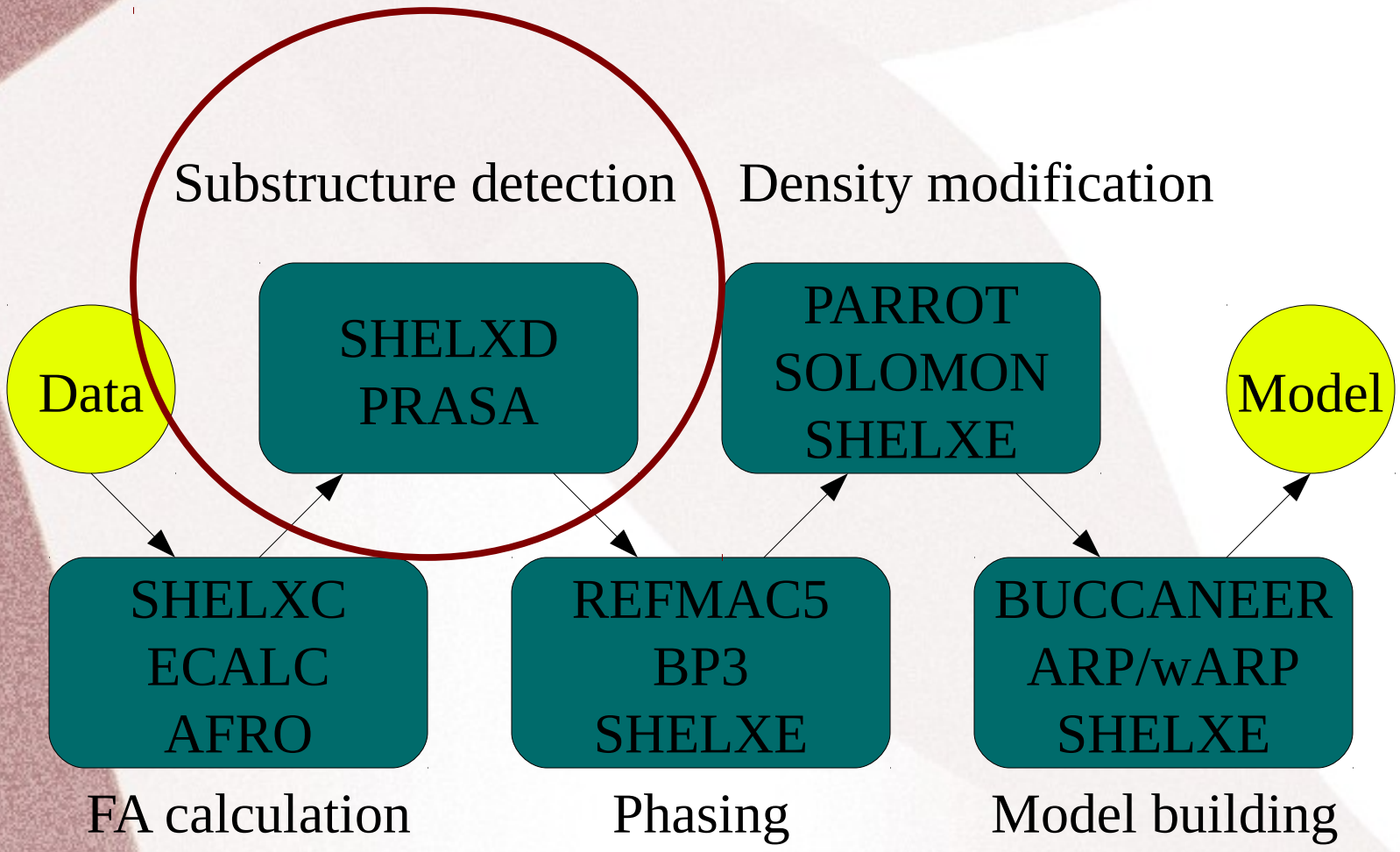


F_A estimation

- Substructure detection methods need to first estimate "F_A" or "substructure factor amplitude"
- Improving the estimates can improve hit rates of substructure detection and solve things that could not be otherwise solved.
- The simplest estimation of F_A for SAD data is

  $\Delta F = \big|\, |F^+| - |F^-| \,\big|$
- E values: normalized F_A values
- Large observation/param. ratio: data exclusion
- Available programs in Crank2: SHELXC, AFRO, ECALC

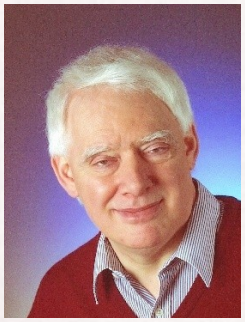
Structure solution from experimental phases with Crank2.



Determination of anomalous scatterers

- A crucial step in structure solution from anom. data
- Current programs (eg SHELXD, HySS):
  - Direct methods: obtain/optimize phases using triplet relationship, tangent formula
  - Patterson methods: obtain positions of anomalous scatterers from Patterson function (or use it for seeding)
  - Space recycling: apply direct methods in reciprocal space and density modification in real space

SHELXD



- Author: George Sheldrick
- SHELX slides: Andrea Thorn
- Originally for *ab initio* solution of big small molecules
- **E-Values** – from SHELXC
- **Patterson seeding;** start atoms consistent with the anomalous/isomorphous Patterson maps.
- **Dual space direct methods** recycle and modify trial substructures by peak search in the density map and refining phases in reciprocal space.
- Substructure is determined - inital phases:

$$\varphi_{A} + \alpha = \varphi T$$

Phase retrieval methods

- Group of algorithms attempting to solve general phase retrieval problem
- Routinely used to solve phase problem eg in astronomy
- Space recycling: iterative application of operations in the reciprocal space and the crystal space
- Do not make any use of direct / Patterson methods
- The operations in the reciprocal space alone or in the real space alone are principally NOT able to solve the phase problem

PRASA: A new program for substructure detection

- Phase Retrieval for Anomalously Scattering Atoms
- C++ program built using the CCP4 Clipper libraries, integrated in CRANK2
- charge flipping and RAAR phase retrieval algorithms
- automatic multiple high resolution cutoffs
- input of number of substructure atoms not needed (RAAR only)
- parallelized (openMP)
- beta stage, released and usable, not default

Charge flipping

- The simplest usable phase retrieval technique
- ***CHF***$(\rho) = R_D P_M(\rho)$
- ***CHF***$(\rho) = \begin{cases} \rho & \rho >= \delta \\ -\rho & \rho < \delta \end{cases}$
- Oszlanyi & Suto (2004, 2007, 2011)
- Palatinus (2007,2012) - program Superflip
- Dumas & van der Lee (2008)

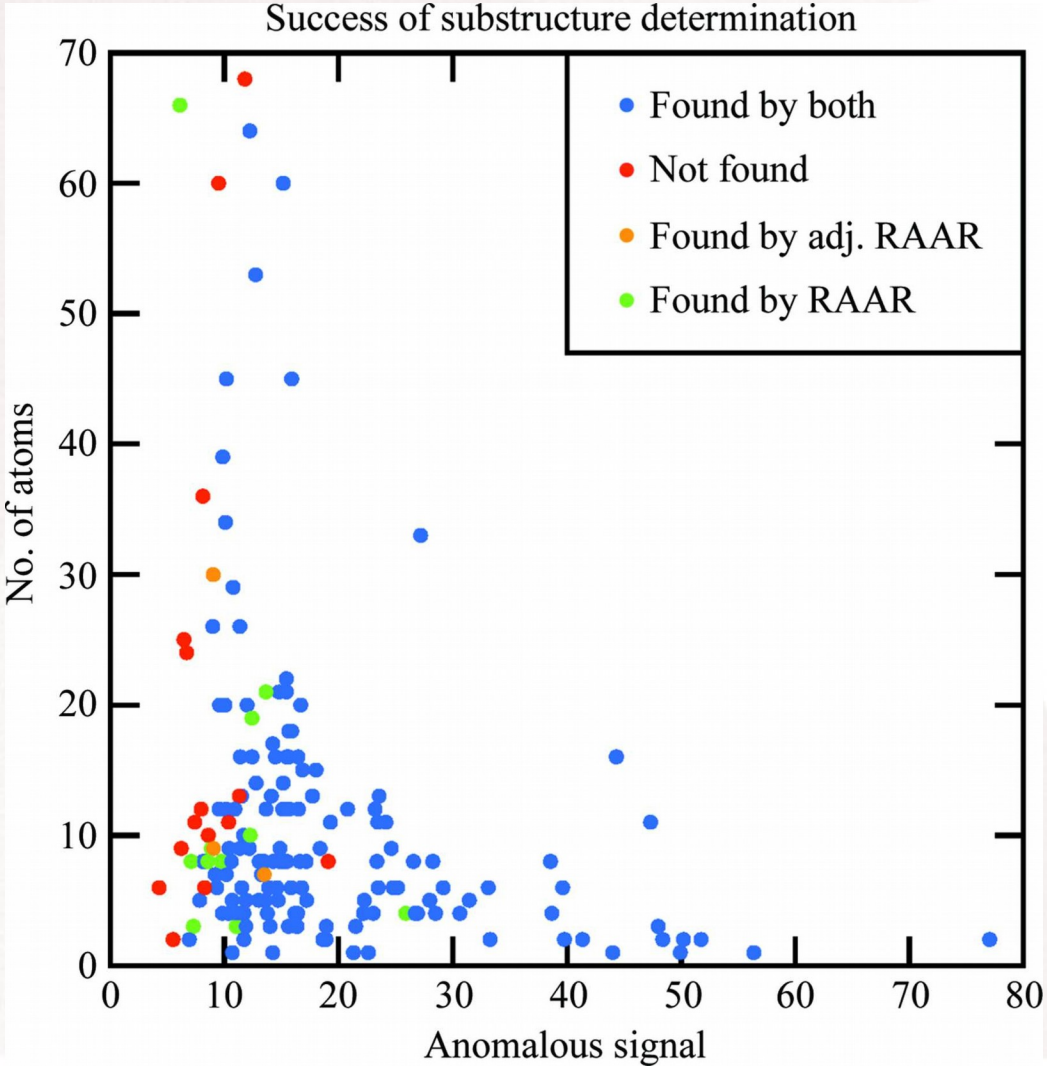
RAAR (Relaxed Averaged Alternating reflections)

- $\boldsymbol{RAAR}(\rho) = \tfrac{1}{2}\beta(R_D R_M + I)(\rho) + (1-\beta)P_M(\rho)$
- $\boldsymbol{RAAR}(\rho) = \rho \qquad R_M(\rho) >= \delta$

  $(1-2\beta)\rho+\beta\rho_{old} \qquad R_M(\rho) < \delta$
- $\beta = 0.8 \Rightarrow \quad -0.6\rho + 0.8\rho_{old}$

Massive testing results

- Automatic structure solution with Crank2+PRASA on ~150 SAD datasets using charge flipping and RAAR



Example: Novel SAD data solved by PRASA

- difficult SAD data from the McGill lab / Quebec, Canada by Juliana Munoz
- 3.2A dataset collected at the APS/CCP4 School workshop
- All structure solution attempts at the APS School failing
- PRASA was able to obtain a clear and complete substructure solution and a highly complete protein model has been eventually built by several iterations of Crank2 combined model building (Rfree=33)

Important parameters in substructure detection

- The number of trial cycles to run.
- The number of atoms to search for (SHELXD)
  - Should be within ~20% of the actual number
  - First guess: Estimate per monomer and multiply
- The high resolution cut-off:
  - For MAD, a good guess comes from anomalous difference correlation.
  - For SAD, a first guess is high resolution limit + 0.5A; if unmerged data are available, a better guess might be derived from CCanom1/2
  - Several jobs with differing values may be needed

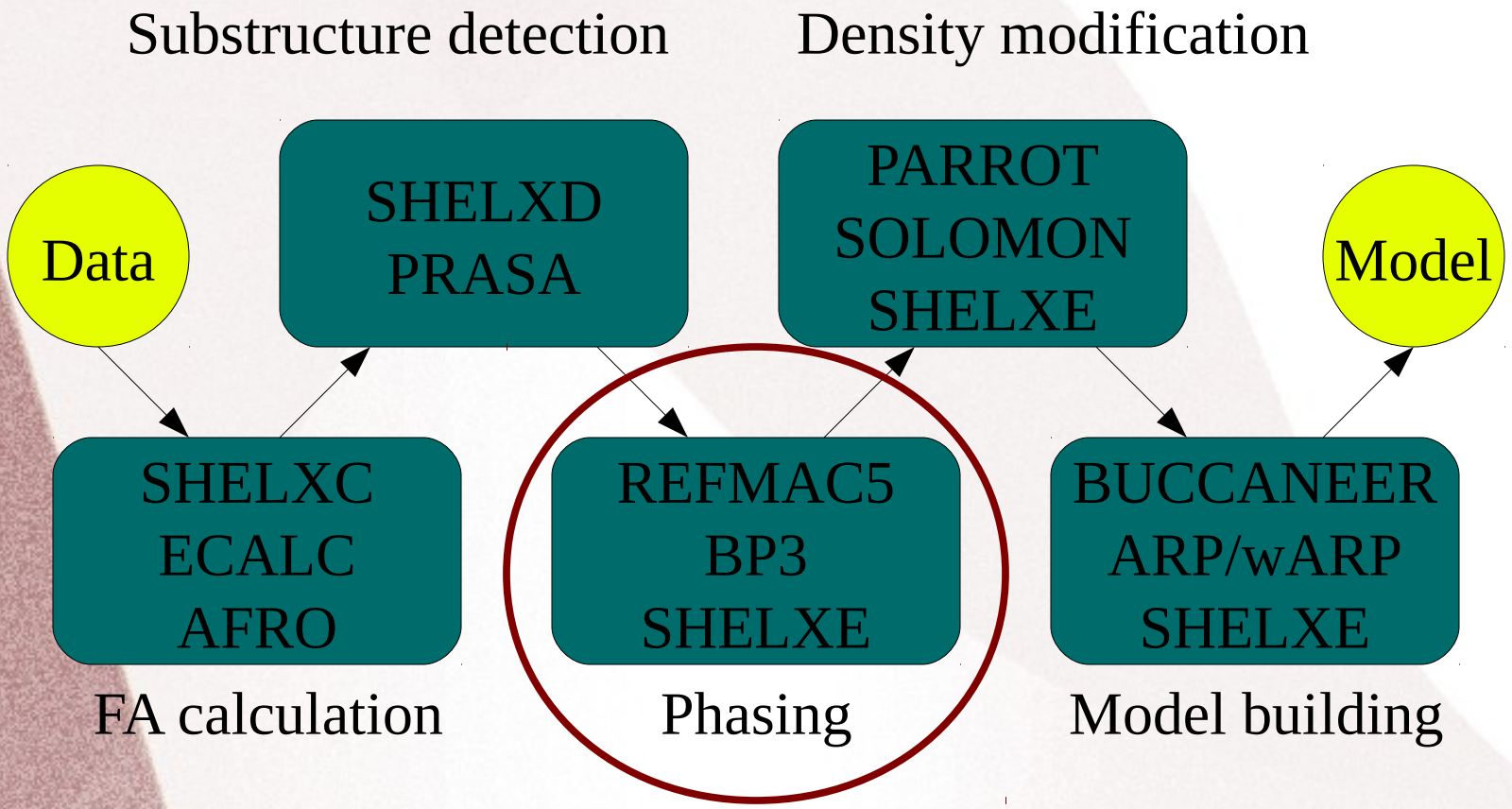
Substructure determination validation

- Usually either a highly complete solution is found or the solution is incorrect
- Indicators of a correct solution:
  - CFOM > 75 for SHELXD (120 for MAD)
  - CLD*FOM*CC > 0.3 for PRASA

(conservative criteria for a solution -

half or even less may be still a solution)

Structure solution from experimental phases with Crank/Crank2.



Substructure refinement and phasing

- Refinement of the substructure parameters and error parameters
- The refined parameters are used for initial estimation of phases
- Available programs in Crank2: REFMAC5, BP3

Substructure refinement and phasing by BP3 and Refmac

- Can be used for SAD, SIRAS, MAD (BP3 only)
- Using multivariate SAD/SIRAS phasing function
- Outputs the estimates of “best” initial phases PHIB and their reliability FOM and four Hendrickson-Lattman (HL) coefficients “encoding” the phase distribution (all in the output MTZ file)
- REFMAC is usually faster except if there are just a few substructure atoms (FFT vs direct summation)

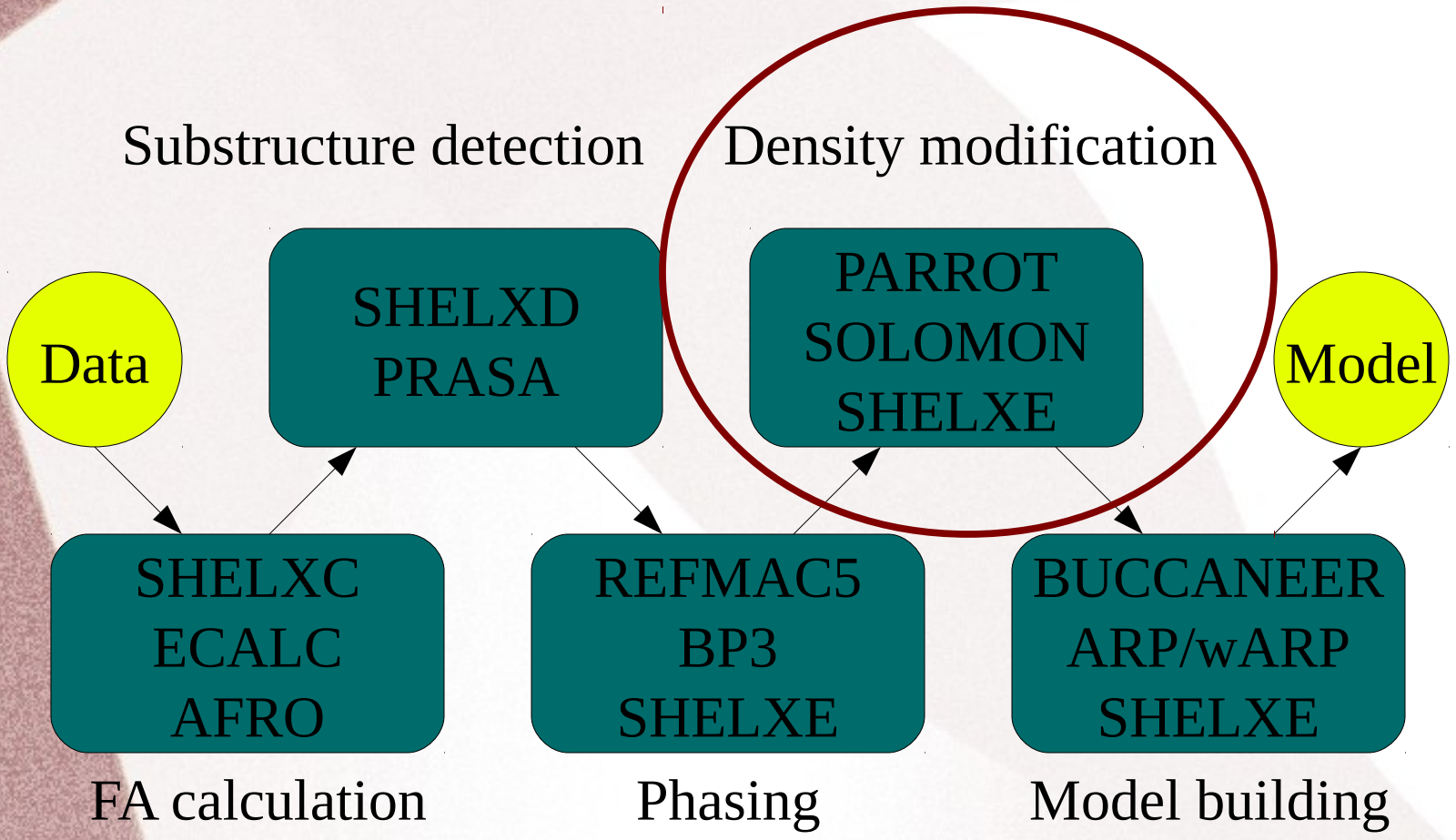
Multivariate distribution for a SAD experiment

- Include effect of model and measurement errors and correlation between observed and calculated Friedel pairs.
- Required multivariate joint probability distribution

$$P_{ph}\left(F_o^+, F_o^- \middle| F_H^+, \varphi_H^+, F_H^-, \varphi_H^-\right)$$

- The function can be further extended for phase combination and refinement in density modification and model building stages

Structure solution from experimental phases with Crank/Crank2.



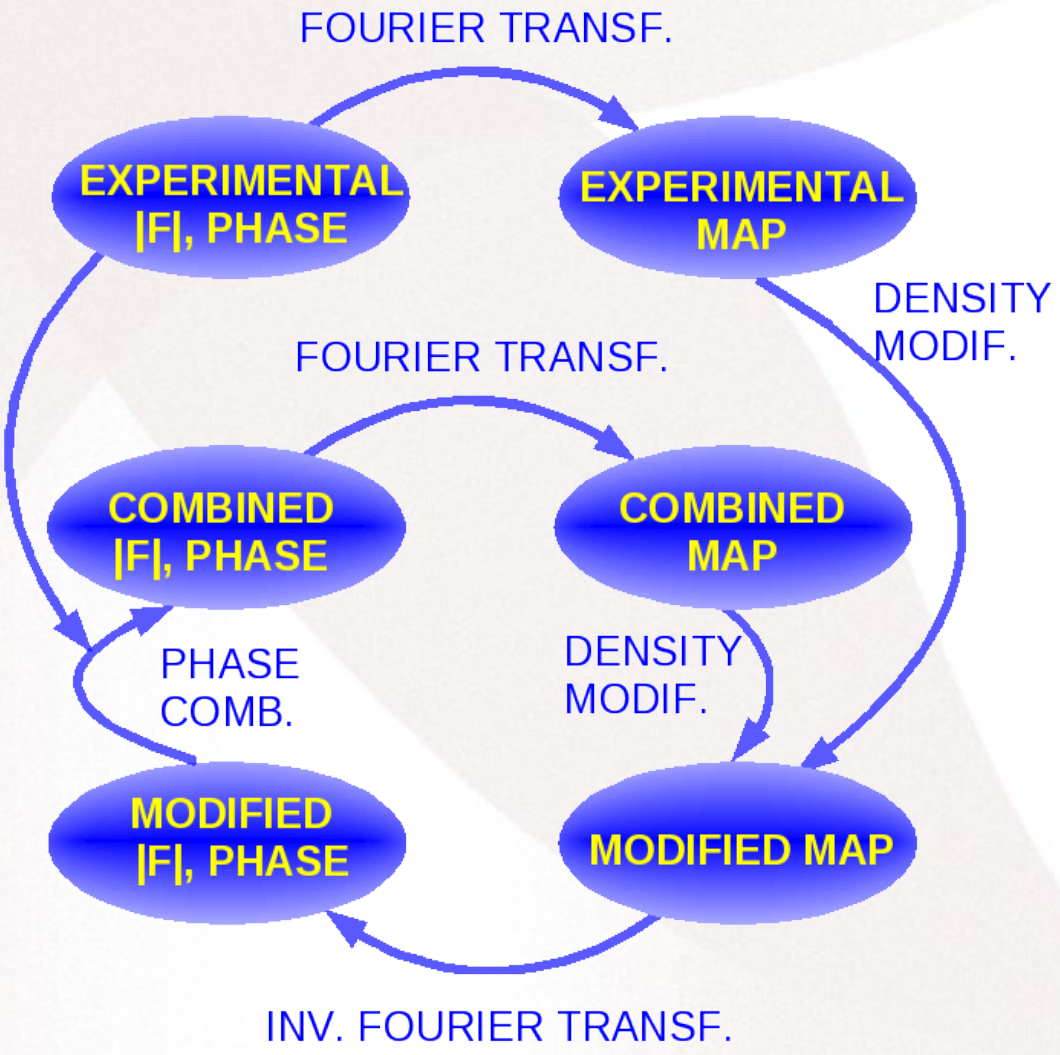
After we have initial phases: Density modification

- Improving phases of the initial map by incorporation of prior information about protein map features into it
- Space recycling with phase combination
- Available programs in Crank2:
  - Parrot
  - Solomon
  - SHELXE
- In case of SAD or SIRAS, Parrot and Solomon can use external multivariate combination with bias reduction by MULTICOMB or REFMAC

Density modification principles

- Idea: apply prior information about protein electron density maps to the experimental map
- Solvent flatness prior: the density in solvent regions is close to constant → solvent flattening
- Histogram similarity prior: the histograms of protein density maps are similar → histogram matching
- NCS prior: the density in NCS related regions should be very similar → NCS averaging

Density modification procedure



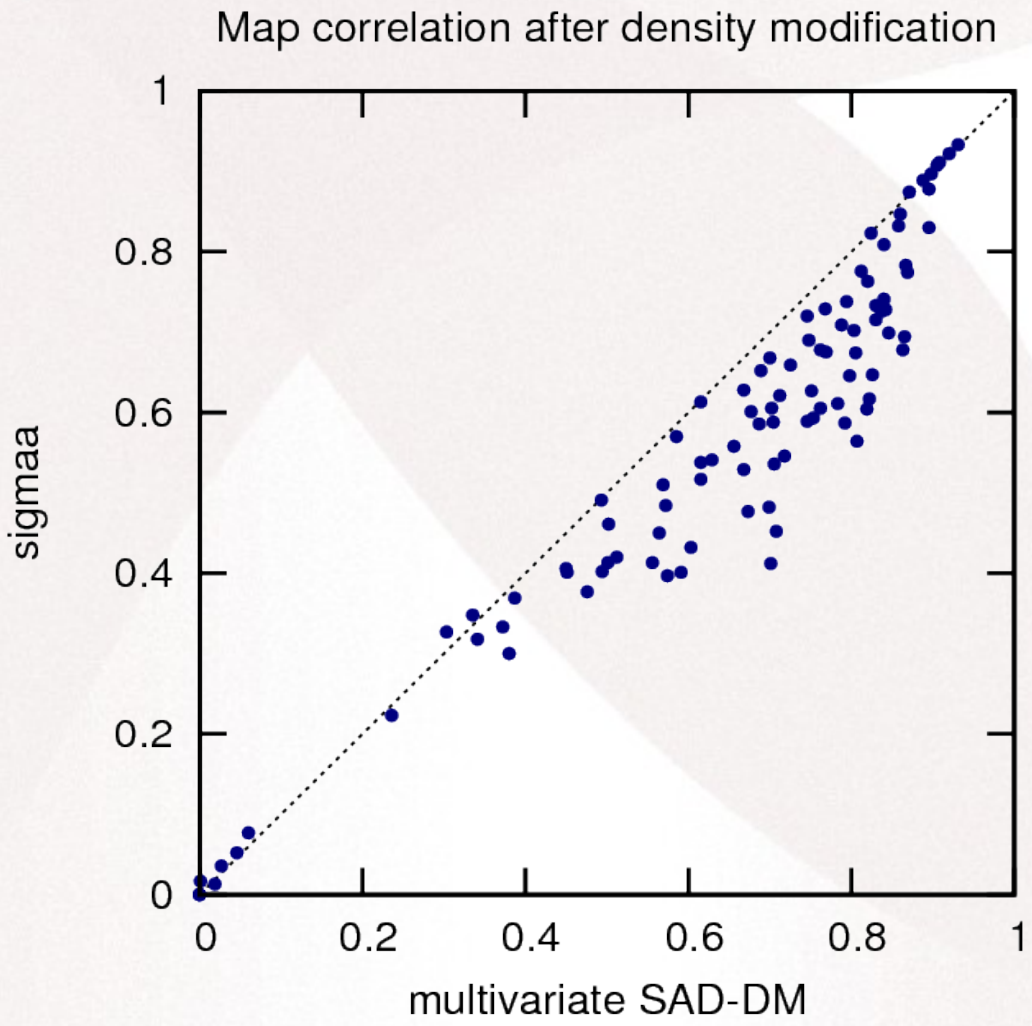
Multivariate phase combination for density modification

- Density modification procedures usually neglect the correlation between the original map and the density modified map
- Our approach: multivariate SAD-DM probability distribution

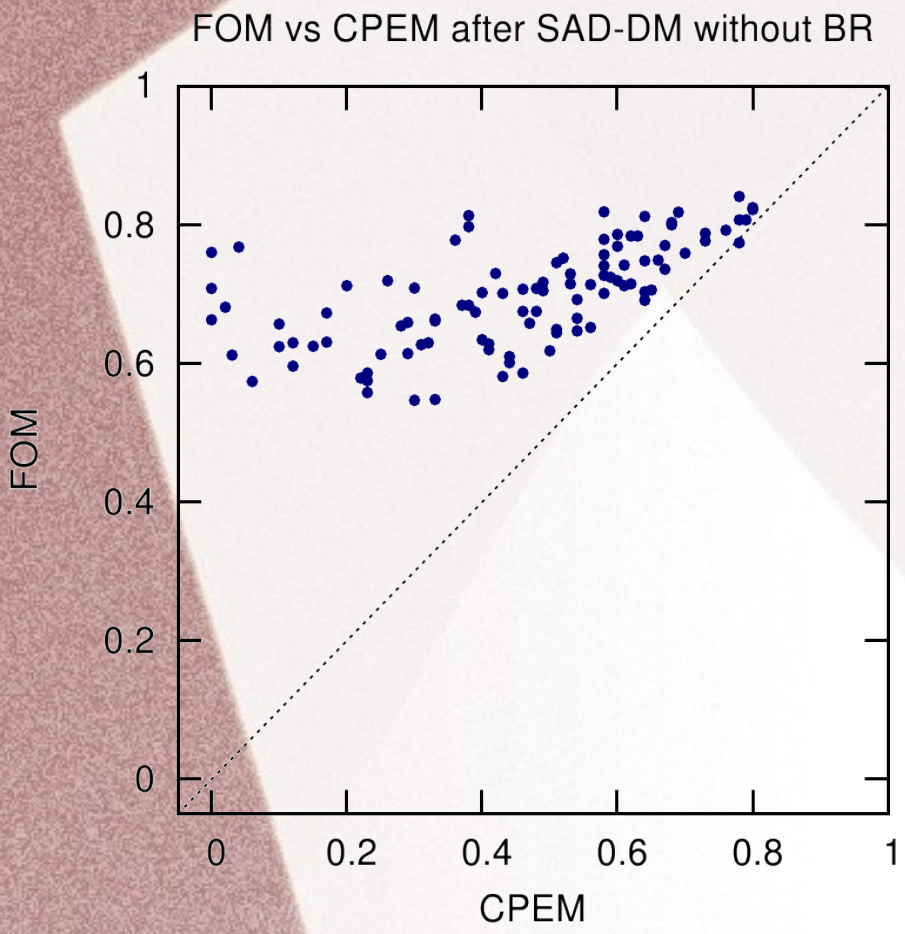
$$P_{DM}\left(F_o^+, F_o^- \middle| F_H^+, \varphi_H^+, F_H^-, \varphi_H^-, F_{DM}, \varphi_{DM}\right)$$

- Advantages: no independence assumption, dynamic construction of phase probabilities rather than static HL
- Implementation: MULTICOMB, REFMAC

Comparison of univariate vs. multivariate (SAD-DM) function



Density modification bias

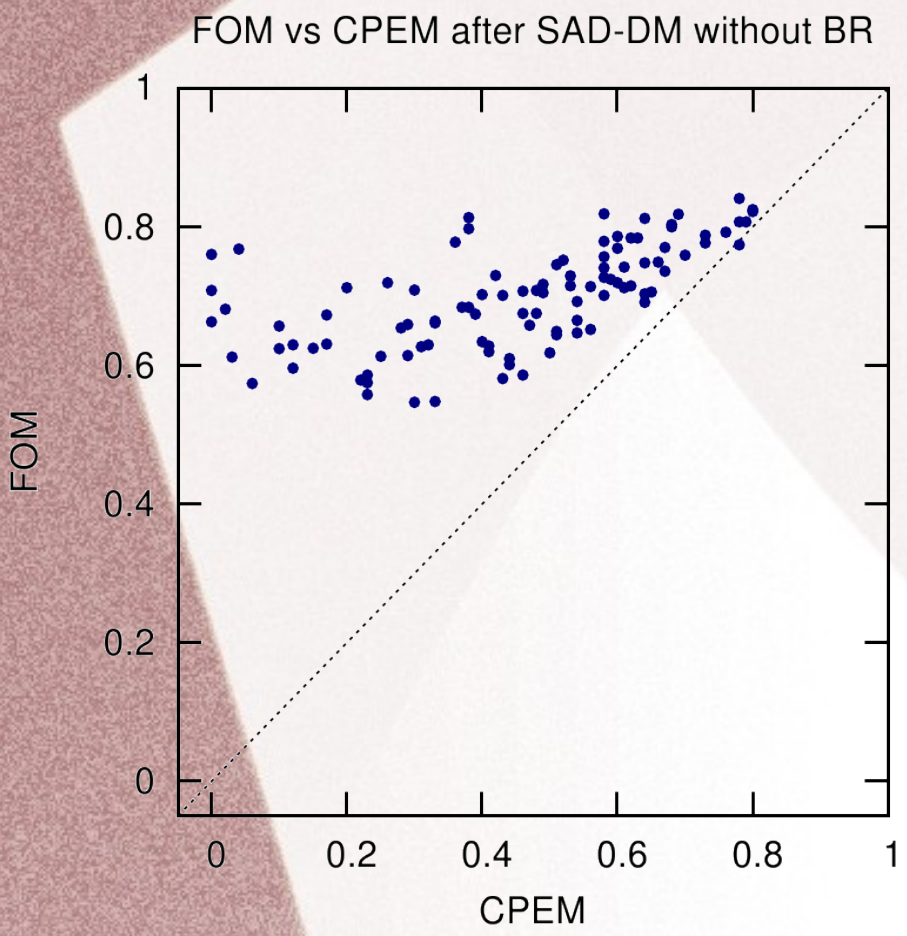


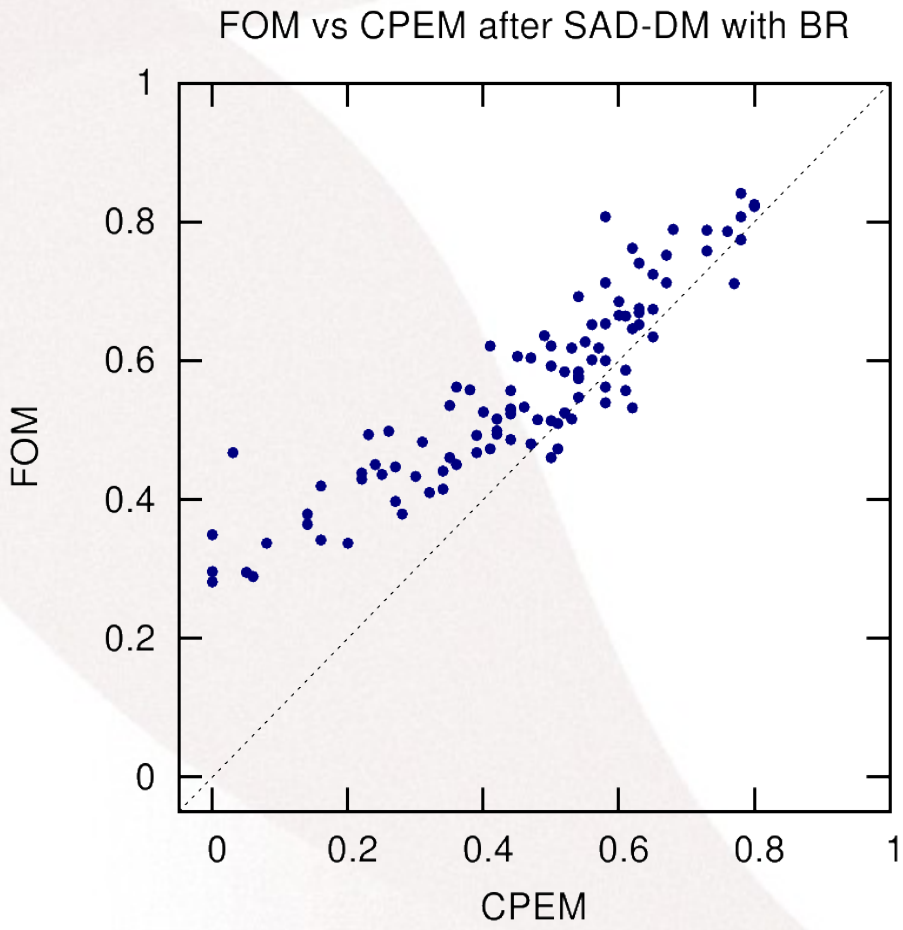
- Estimated <FOM> is much larger than <cos(ph.error)>
- Reason: the phase quality is estimated from fit between the observations and the model - which was constructed from the observations

Solution for SAD-DM: β-correction

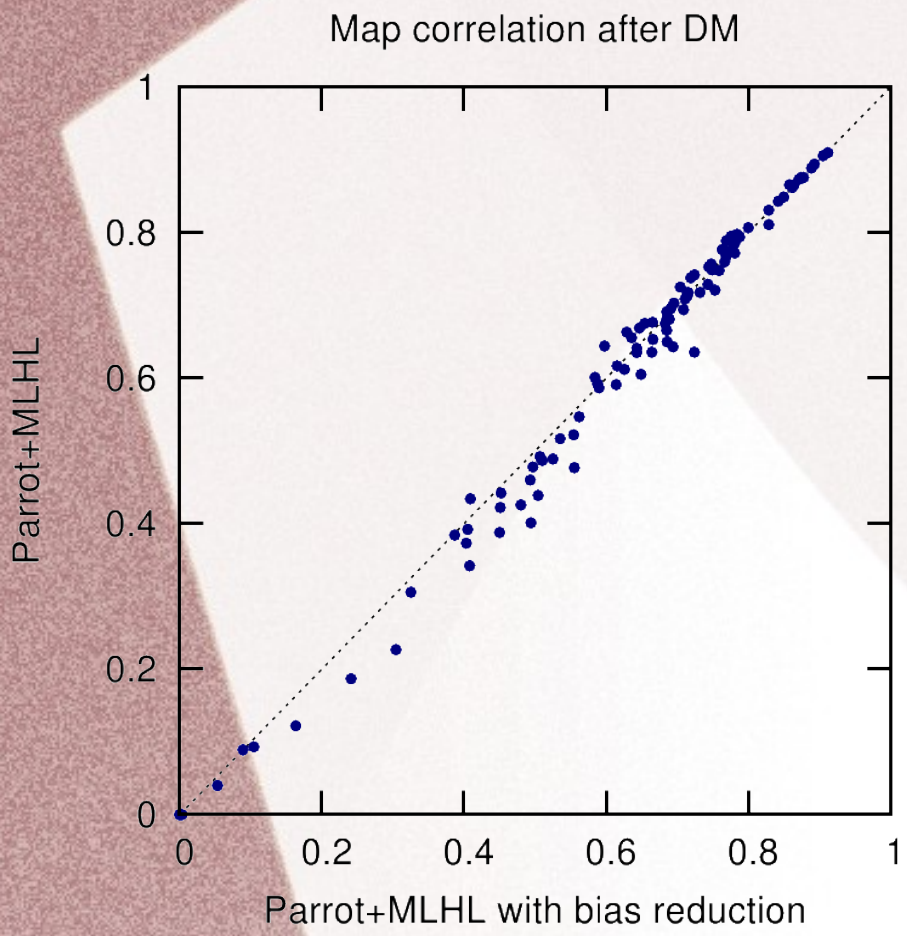
- Estimate overall bias parameter β:
  - run 5 DM cycles with 'free' set of reflections set aside and comparing the correlation between F_o and F_c in free and working set
- Run standard DM using β (and all reflections):
  - β is applied to the correlation term betwen F_o and F_c in the covariance matrix
- β is ~1 in case of no bias and can reach 0.2-0.3 in case of very strong bias

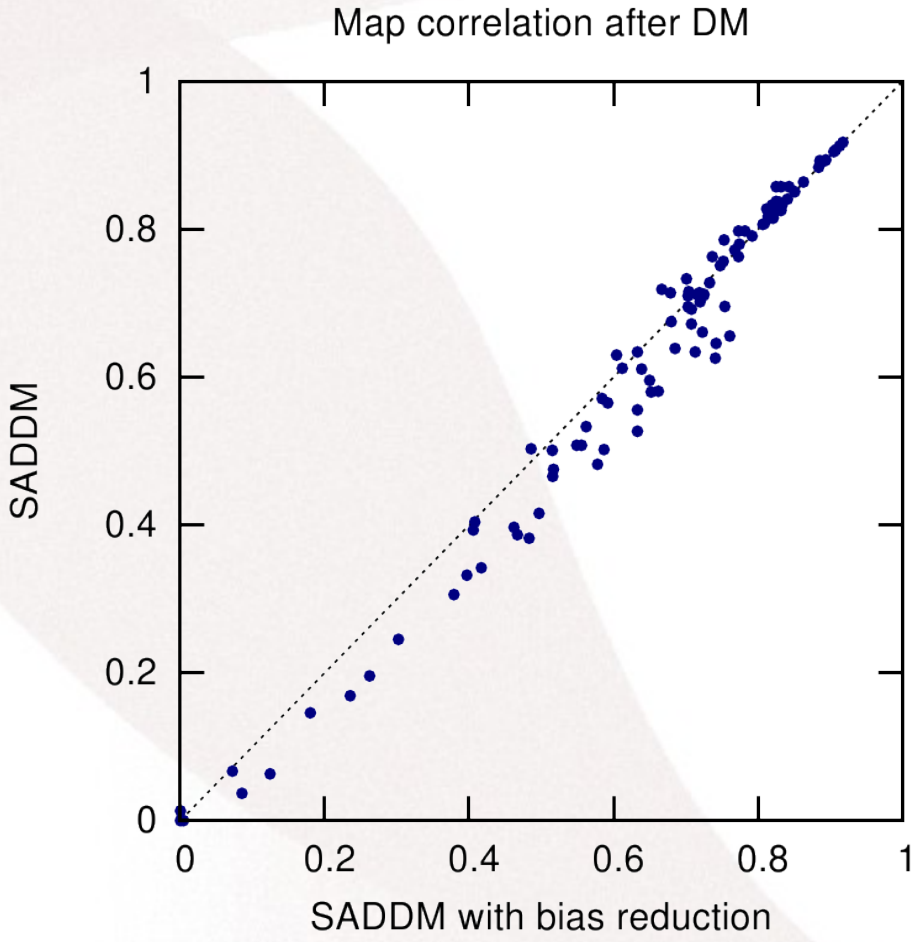
FOM and phase error after DM with/without bias reduction





Map correlation after DM with/without bias reduction





Automatic hand determination

- The hand is not known - either the found substructure or its inverse is correct
- Safe (albeit slower) approach: try building with both hands
- Crank2 chooses the hand before building, assuming that a correctly handed substructure provides a better map
- Criteria used: combination of CLD of the map after phasing with FOM after “fast” density modification
- In tests on 150 datasets, the wrong hand is chosen in 2 cases (none of which could be built due to very weak anomalous signal)

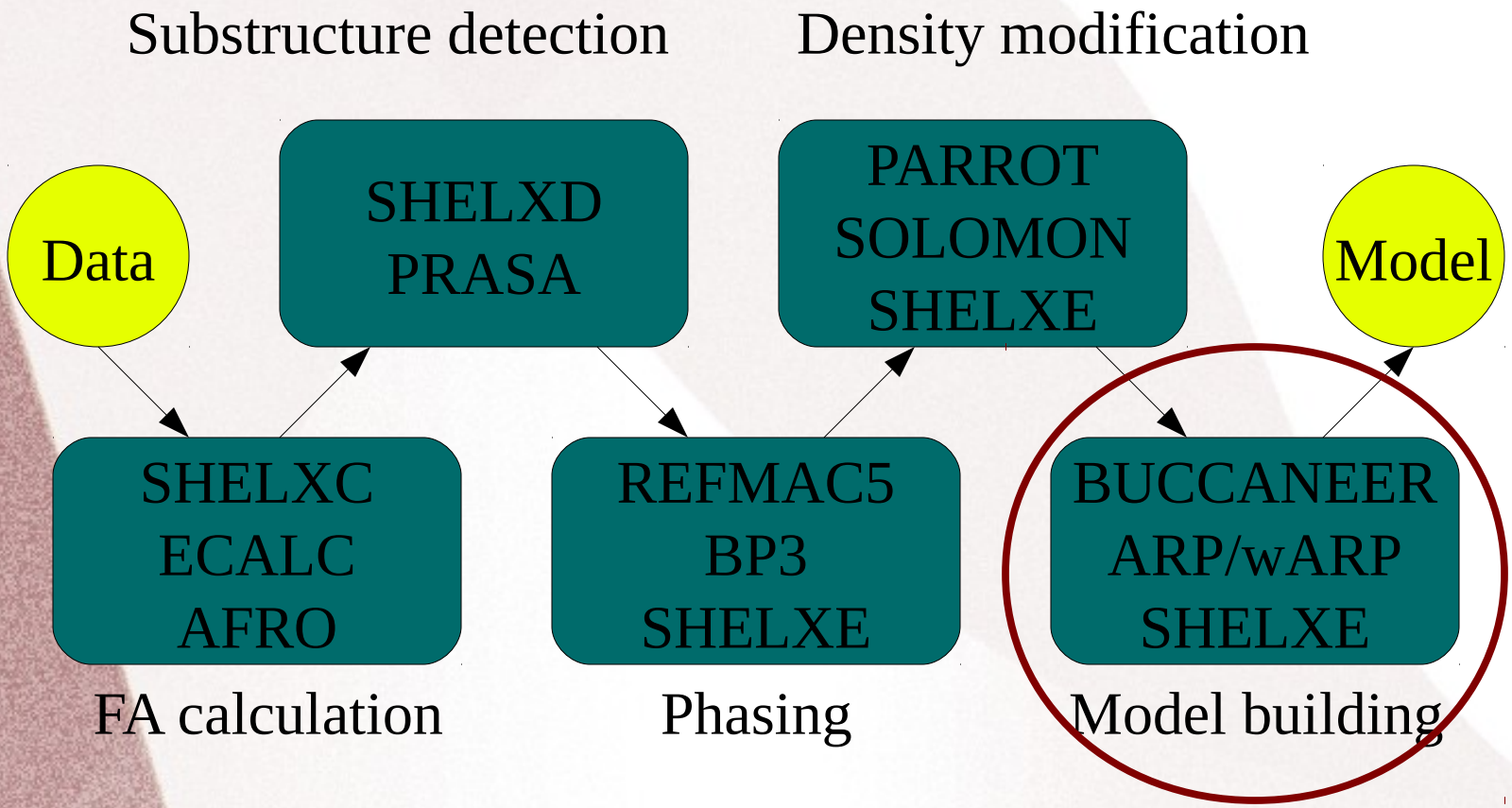
How are we doing?

- Statistics from substructure phasing:
  - FOM (>0.3 usually a solution, <0.15 usually not)
  - CLD (or skewness) of the map - outputted by MAPRO utility (>0.02 usually a solution)
- Statistics from density modification:
  - FOM (>0.5 usually a solution, <0.35 usually not - only useful if bias reduction was employed!)
- Statistics from hand determination:
  - Distinction in score between the hands
- Does it look like a protein? (visualization)

Further improving the map

- Adjusting solvent content can improve the map after density modification. (Since the number of monomers is generally not known beforehand, neither is the solvent content.)
- Try to determine NCS manually – sometimes the automatic NCS detection in Parrot does not succeed
- Try to find additional anom. scatterers or remove likely wrong atoms (the automatic peak picking and removal thresholds in Crank2 are conservative)

Structure solution from experimental phases with Crank/Crank2.



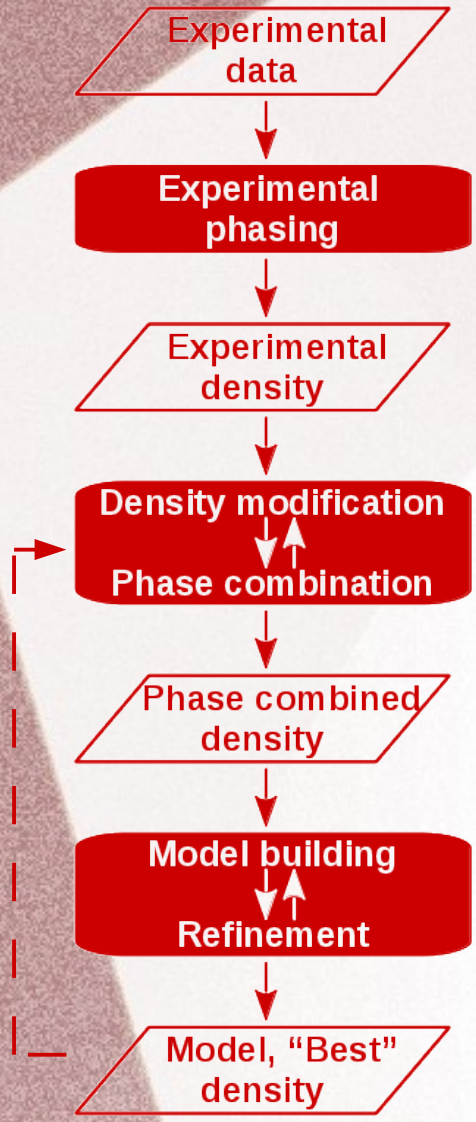
Automatic model building

- Available programs in Crank2:
  - Buccaneer
  - ARP/wARP
  - SHELXE

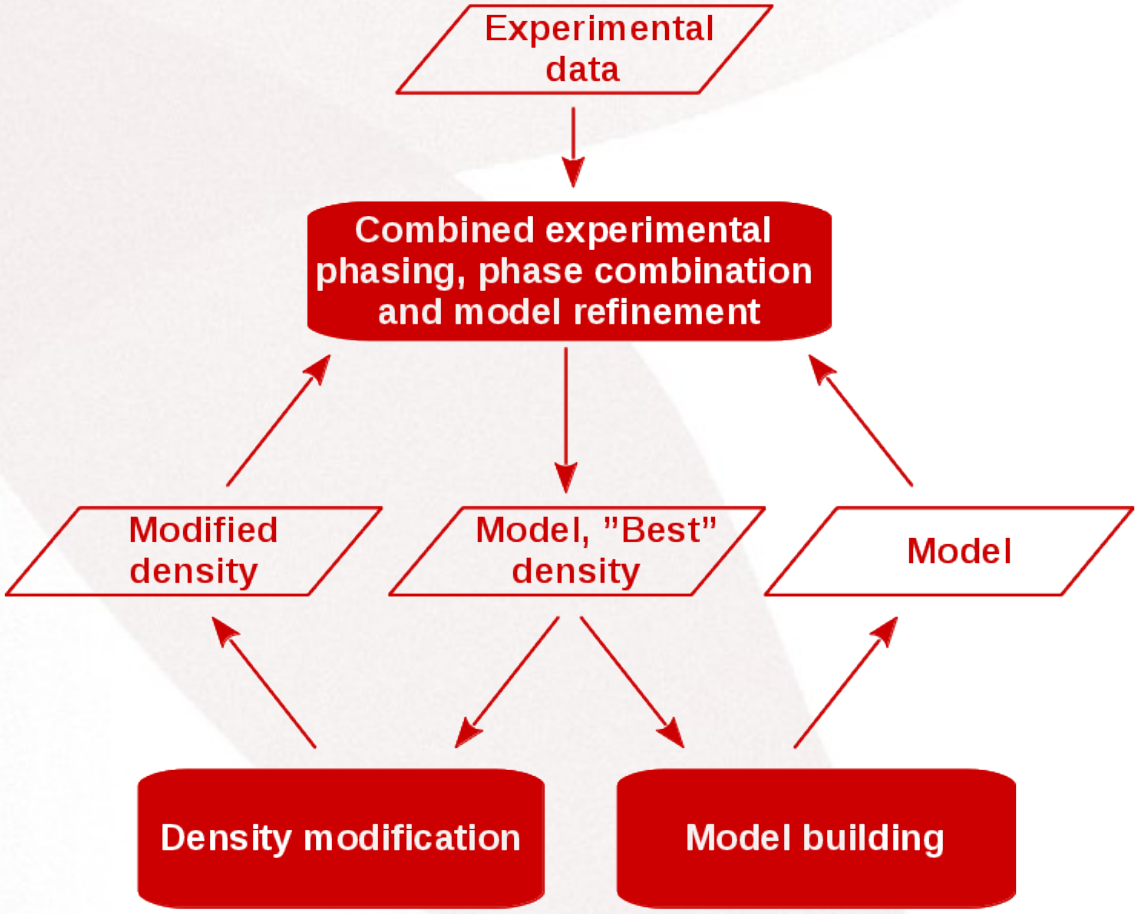
- in most pipelines, model building is either iterated with model refinement by Refmac or more complicated algorithms are used

Combined algorithm: Flowchart

step-wise



Combined (SAD only)



(Skubak&Pannu, Nature Comm., 2013)

Multivariate probability distributions

- Phasing (2003):

$$P_{ph}\left(F_o^+, F_o^- \mid F_H^+, \varphi_H^+, F_H^-, \varphi_H^-\right)$$

- Phase combination (2010):

$$P_{DM}\left(F_o^+, F_o^- \mid F_H^+, \varphi_H^+, F_H^-, \varphi_H^-, F_{DM}, \varphi_{DM}\right)$$

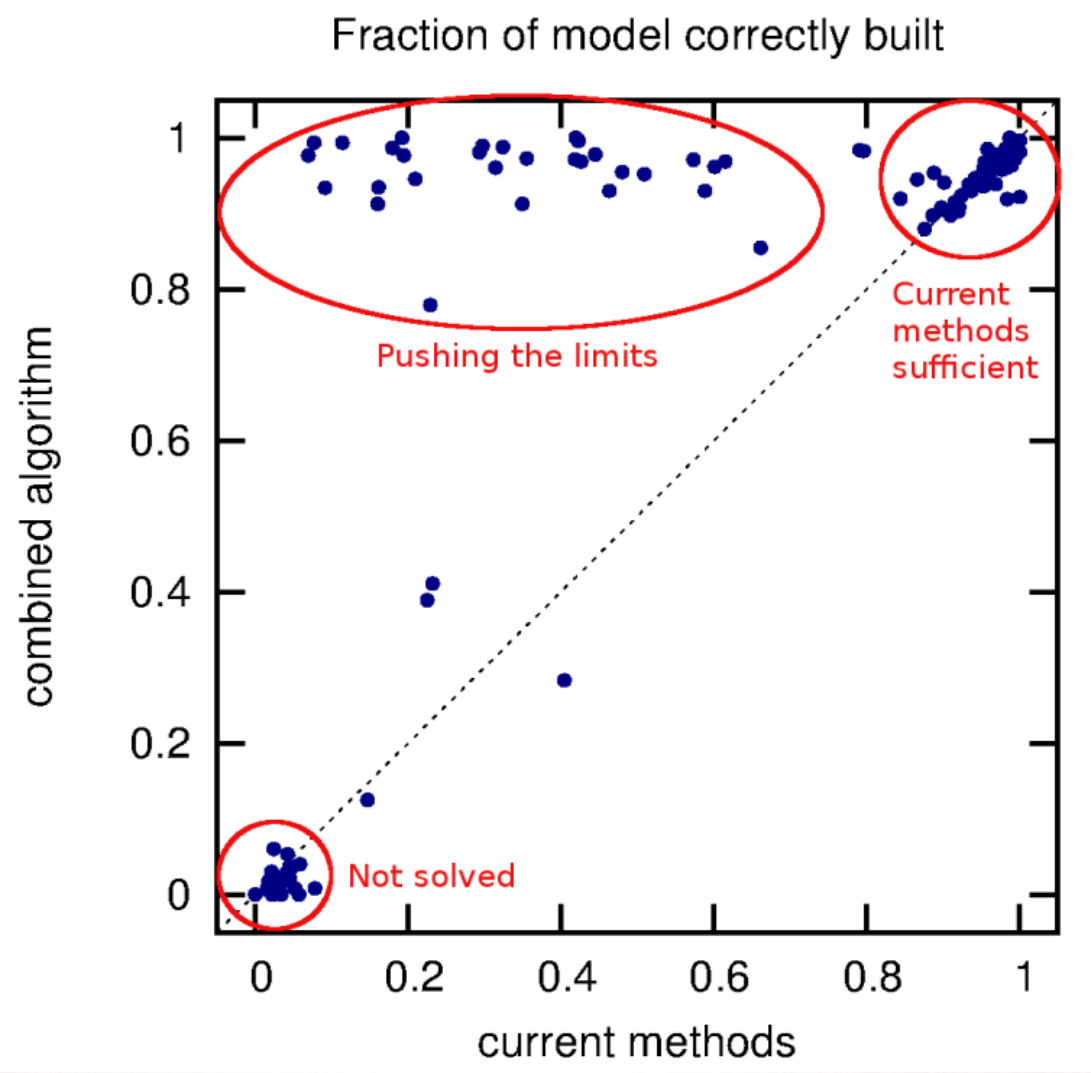
- Refinement (2005):

$$P_{ref}\left(F_o^+, F_o^+ \mid F_P^+, \varphi_P^+, F_P^-, \varphi_P^-\right)$$

- Combined (2013):

$$P_{comb}\left(F_o^+, F_o^- \mid F_P^+, \varphi_P^+, F_P^-, \varphi_P^-, F_{DM}, \varphi_{DM}\right)$$

Model building results on almost 150 real SAD data sets

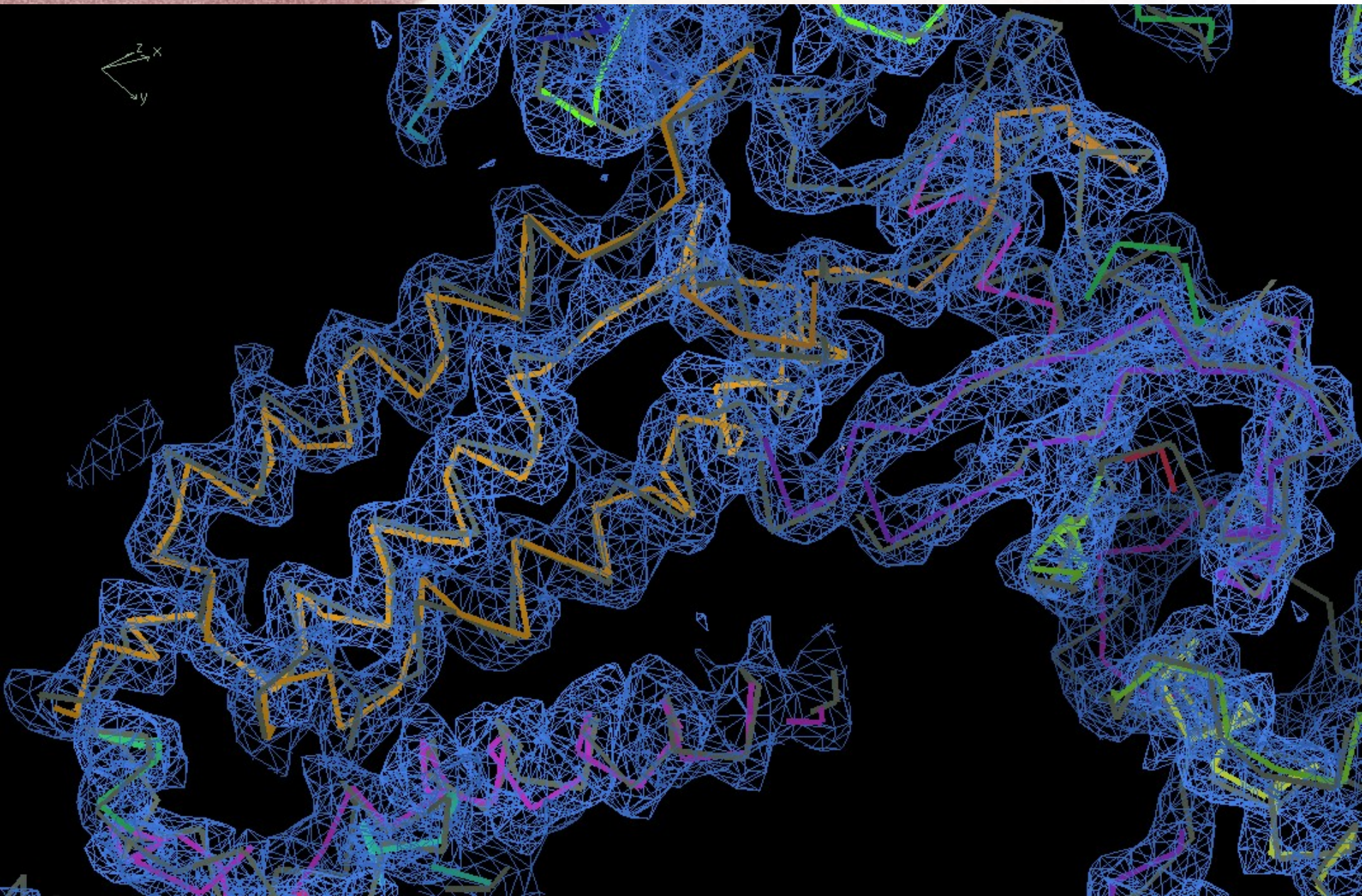


12-subunit RNA polymerase II

- 3.8A resolution dataset with anomalous signal from eight intrinsic zinc ions

  (Meyer P.A. et al., *J.Biol.Chem.*, 2009)
- ~4000 residues in the asym. unit; no NCS
- Originally solved by a partial model MR followed by multi-crystal MAD phasing, MR-MAD phase combination and manual iterative model building
- The combined approach in CRANK2 can build ~70% of the protein backbone automatically from a single SAD dataset only, with R-free of 37.5

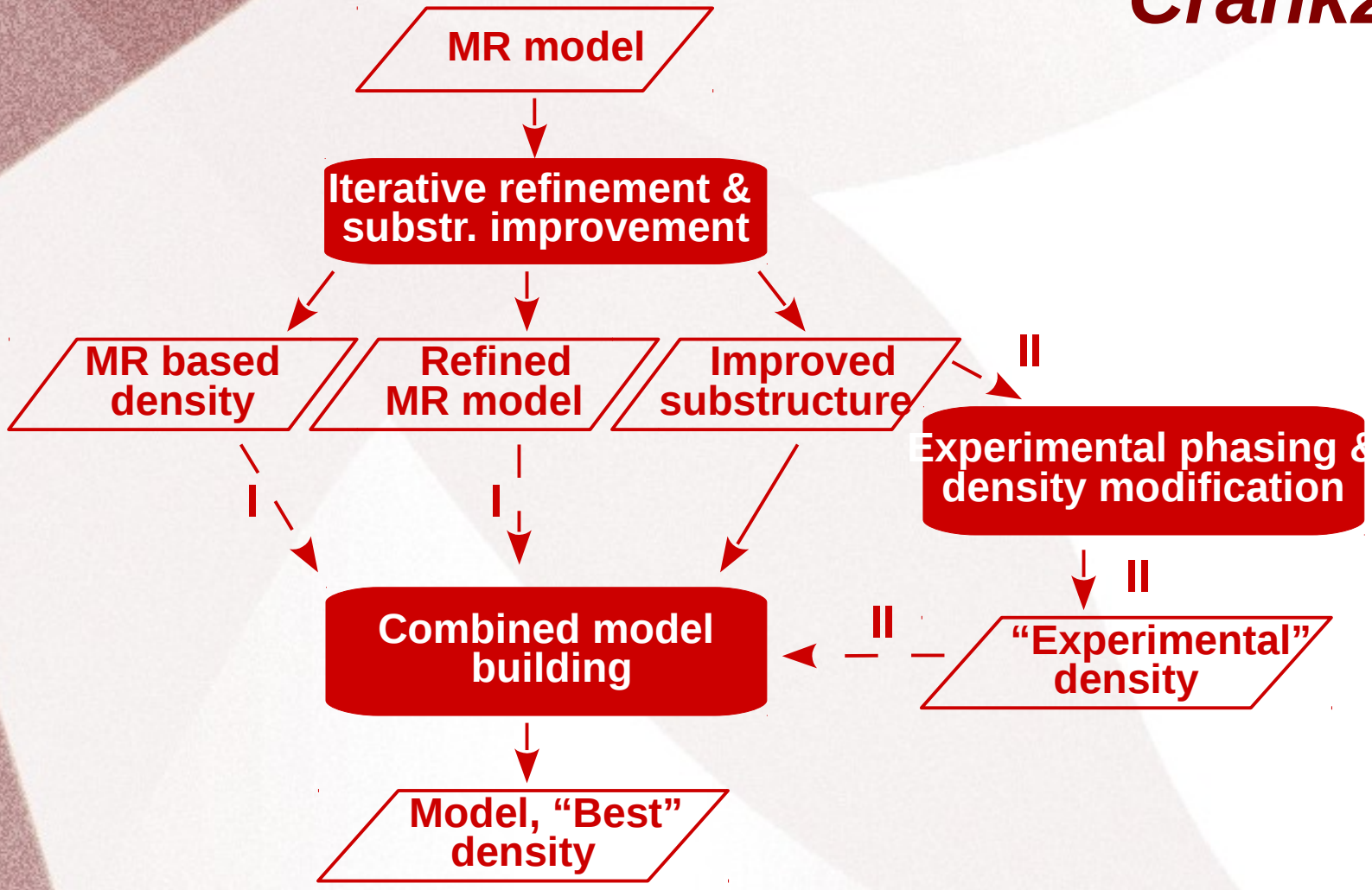
12-subunit RNA polymerase II



SAD after MR (MR-SAD)

- If initial phases are from MR but also significant anomalous information is available, it can improve the model building
- Basic steps:
  - Anomalous substructure completion
  - Model building using the “combined” algorithm
- Two approaches:
  - Rebuild from the MR phases and combine with SAD phases (MR-SAD rebuilding)
  - Rebuild from the SAD map obtained from the (completed) MR substructure

MR-SAD pipelines in Crank2

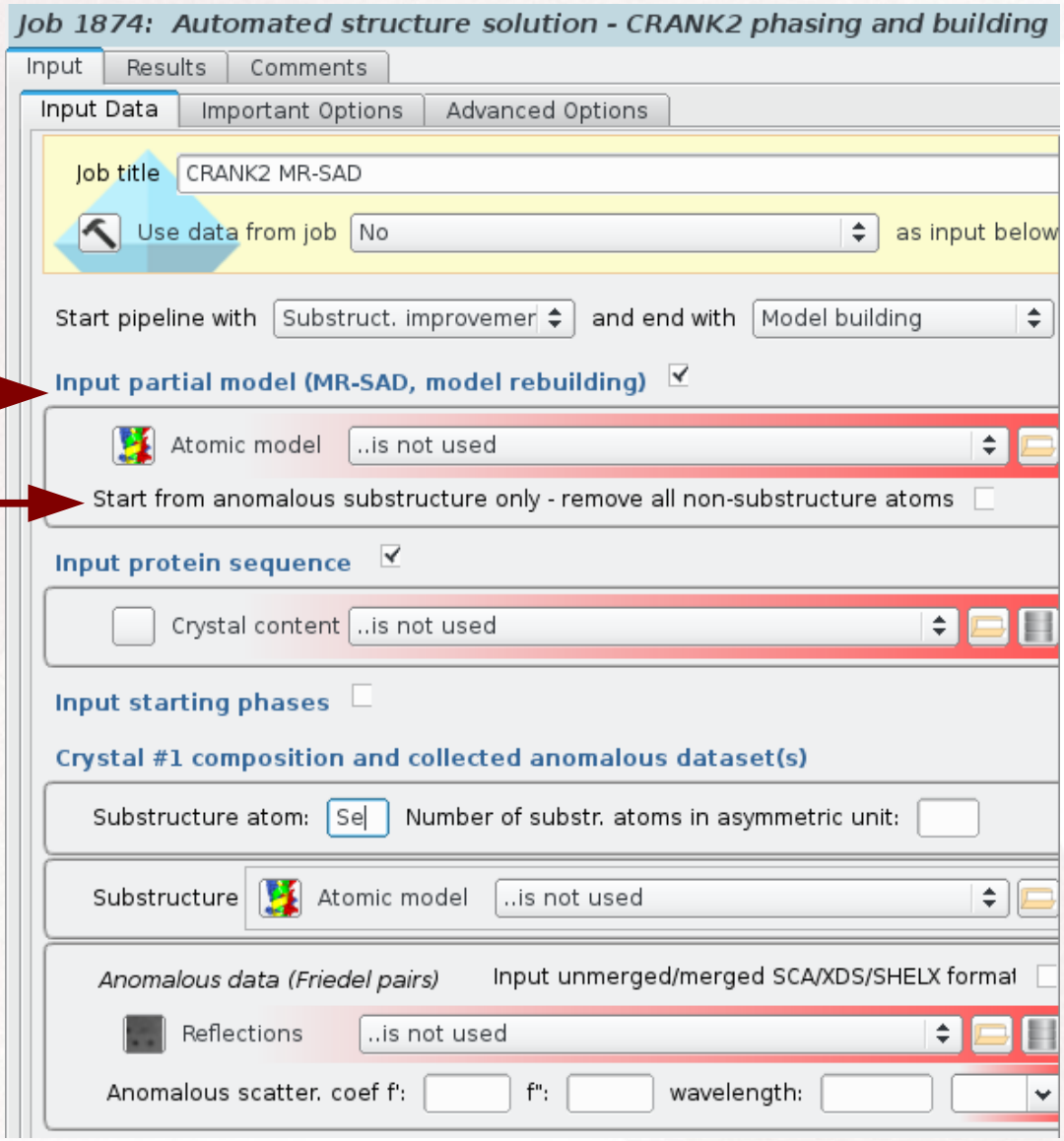


I = MR-SAD rebuilding pipeline
II = SAD-only pipeline (from the MR anom. substructure)

Low resolution MR-SAD examples

	resol. [A]	Rfree init.MR ref.	Rfree SAD-only	Rfree MR-SAD
• unpubl.1	3.6	51.2	32.6	29.8
• unpubl.2	3.2	53.7	51.4	35.8
• 5kvm	3.0	48.6	39.1	38.4
• 4d80	3.6	47.5	39.0	40.9
• 3din	4.5	51.8	39.9	39.6
• 3u5z	3.5	56.8	40.8	39.8

MR-SAD / SAD-only in ccp4i2



Conclusions, remarks

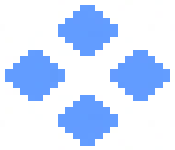
- CRANK2 aims to:
  - push the anomalous signal and resolution limits for automatic structure solution
  - provide as good models as possible - by default, better rather than fastest methods

- If structure is not automatically built (or only partially built), first determine which step has failed (or can be improved): CRANK2 attempts to make re-running steps easy.

Acknowledgements

- All dataset contributors (JCSG, SSGCID, Z. Dauter, M.Weiss, A.Sharma, ...)

- Garib Murshudov, Kevin Cowtan, George Sheldrick, Victor Lamzin, Charles Ballard, Francois Remacle, Peter Briggs, Norman Stein, Martyn Winn



Cyttron