Molecular replacement

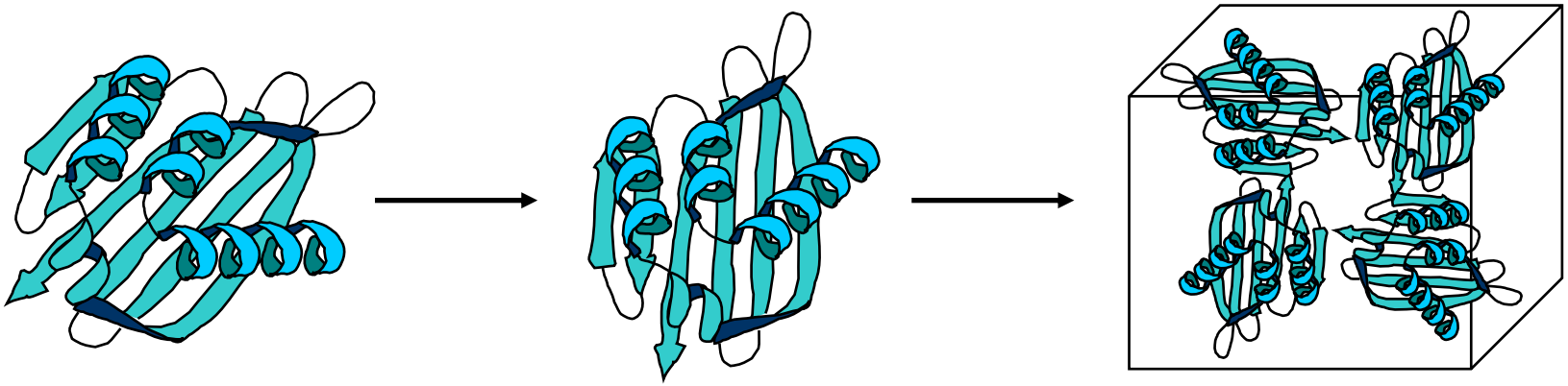
Attacking difficult problems

UNIVERSITY OF CAMBRIDGE

R J Read, Department of Haematology
Cambridge Institute for Medical Research

Phasing by molecular replacement

- Phases can be calculated from atomic model
- Rotate and translate related structure
- Only one data set required!



What makes MR difficult?

- Poor model
  - low sequence identity
  - altered conformation
- Incomplete model, or many copies
  - high non-crystallographic symmetry (NCS)
  - part of complex
  - protein with domain(s) of unknown structure
- Poor data
  - low resolution
  - data pathologies (*e.g.* anisotropy, twinning, tNCS)

Solving the MR problem *vs.* solving the structure

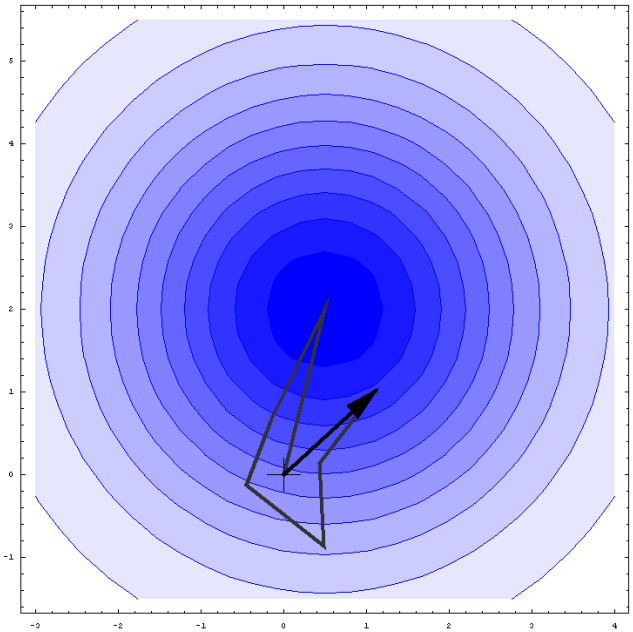
- Solution may be unambiguous but map may be too poor to allow model improvement
  - particularly with lower resolution data
- Model completion is an integral part of structure solution by MR

Why likelihood?

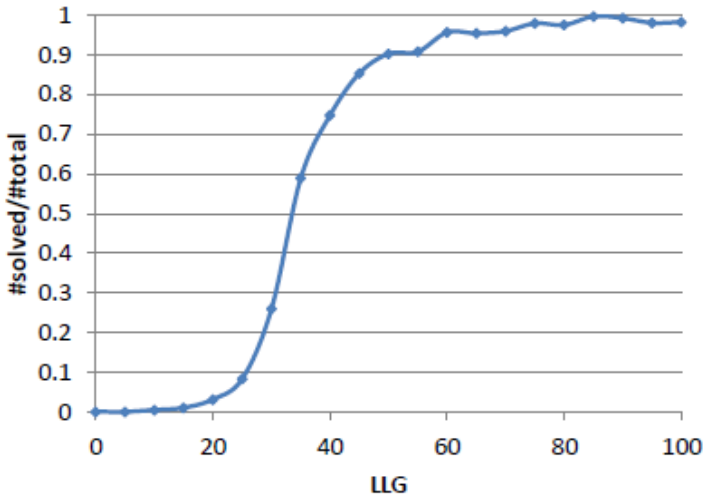
- Accounts explicitly for effects of different sources of error
  - model error
  - measurement error
- More sensitive than other methods
  - especially for multiple copies or small fragments
- Exploits information from partial solutions
- Natural framework for ensemble models
- Absolute score gives good basis for automation
  - choose among different possibilities

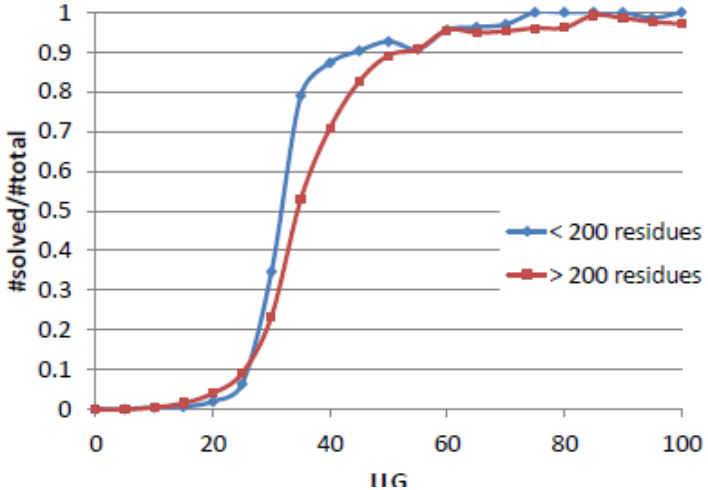
Likelihood-based molecular replacement in *Phaser*

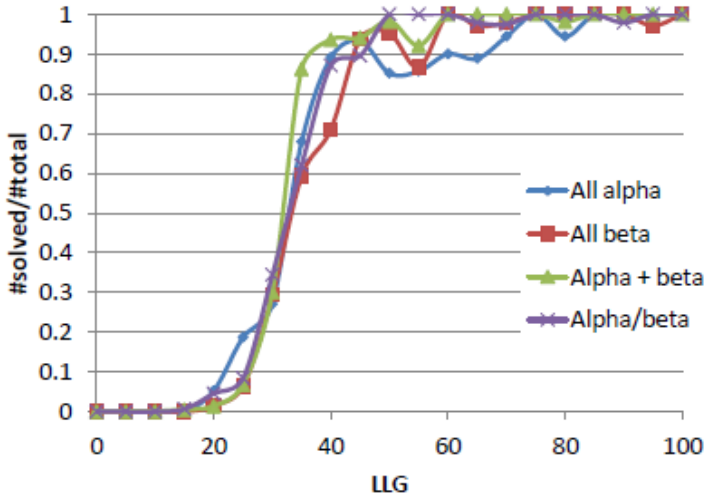
- Likelihood target:
  - probability of observed intensity given structure factor contributions from model(s)
- Log-likelihood-gain (LLG)
  - difference between the logarithm of the likelihood for the model and of the likelihood for the data given a random atom model

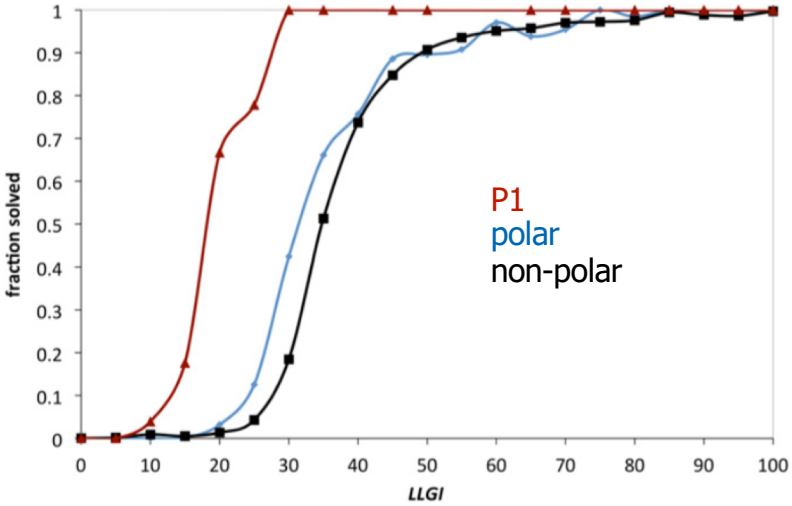


LLG: measure of confidence in solution (Rob Oeffner)









Can I solve it?

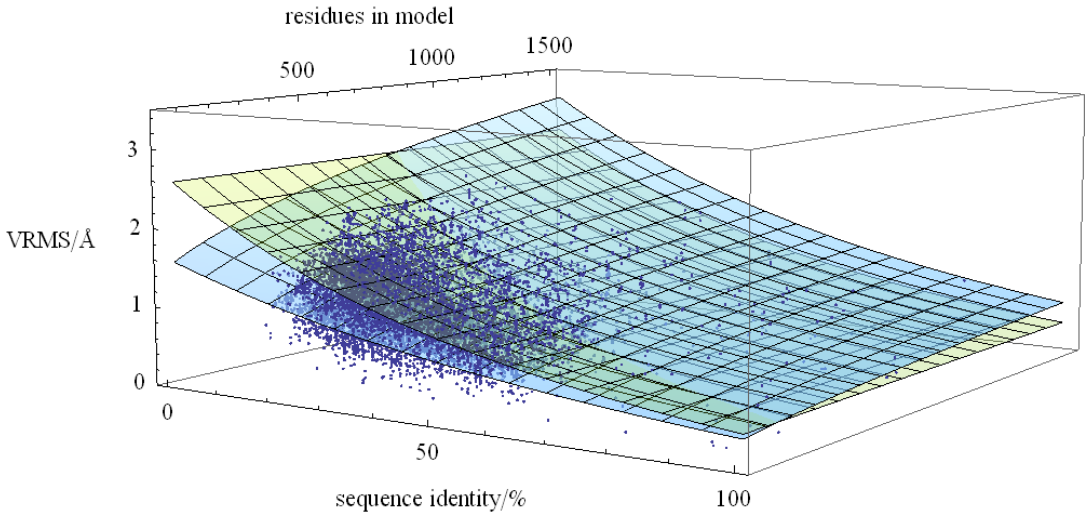
- What is the lowest sequence identity template that I can get away with?
  - depends on fold, can be improved by using ensembles or sophisticated homology modelling
    - further improvement from weighting by expected error
  - some structures with <15% identity can be solved
- How small can a fragment be?

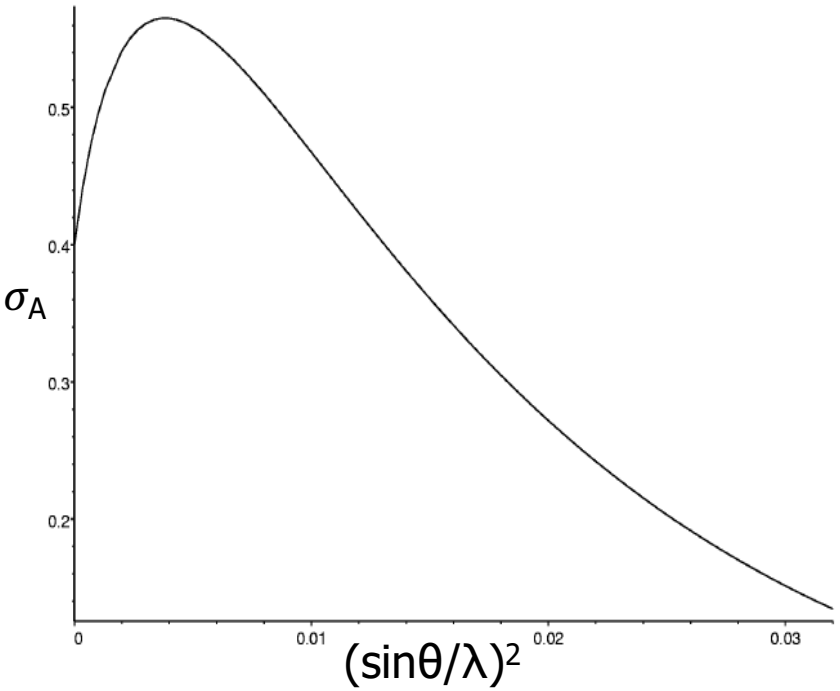
eLLG: assessing difficulty of MR

- Likelihood provides the most sensitive score for MR searches: *Phaser*
  - log-likelihood gain: LLG
    - how much better does model explain data than random atoms?
- LLG score can be estimated in advance of the search: expected LLG, <LLG>, or eLLG
  - LLG/reflection depends on σ_A:
    - function of estimated RMS error and completeness of model
  - total number of reflections, resolution of data
  - no simple rules of thumb!

A priori σ_A curve

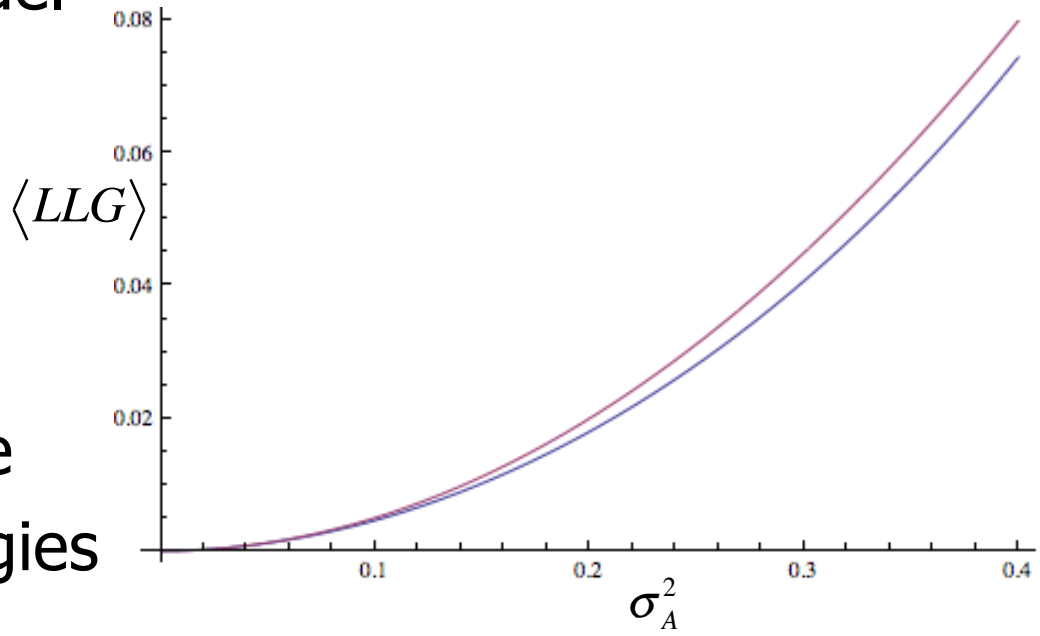
- σ_A^2: fraction of scattering explained by model
  - RMS errors and completeness of model, effects of disordered solvent
  - function of sequence identity and size of structure





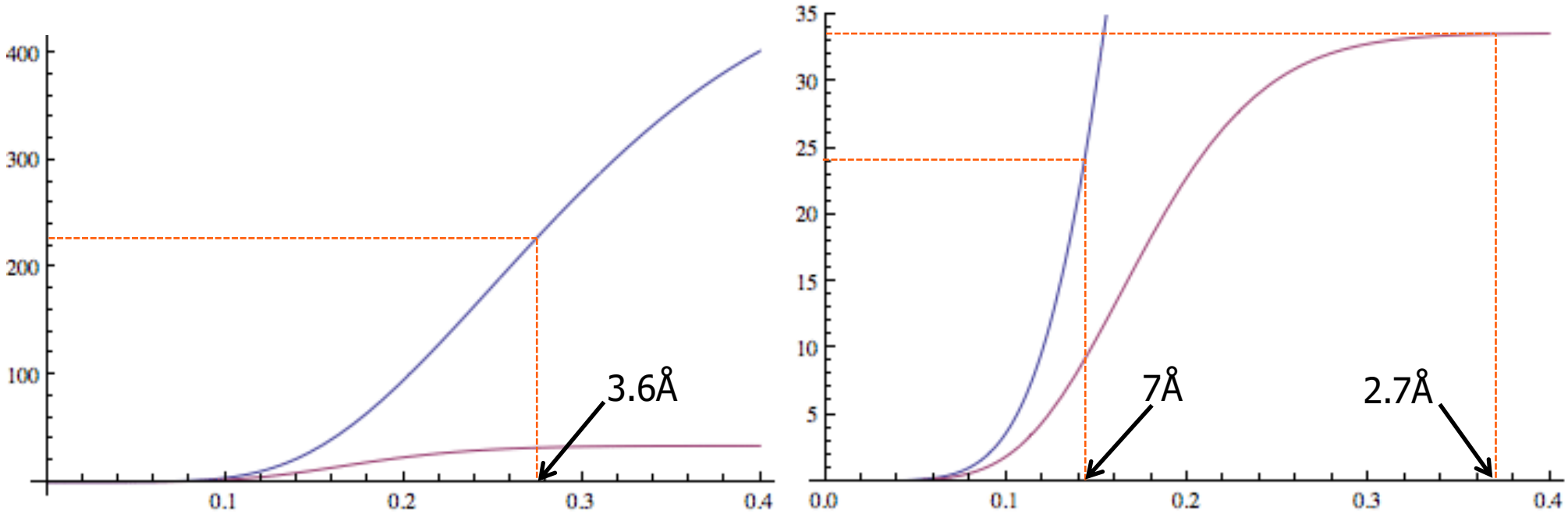
Predicting LLG signal

- Signal-to-noise depends on:
  - the fraction of scattering accounted for by the model
  - expected RMSD of model
  - number of reflections (size and resolution)
- Use this to:
  - predict what is possible
  - develop optimal strategies



Predicting course of MR from <LLG>

- Consider case of good model (0.8Å rms) *vs* bad model (1.5Å rms), both 60% complete, 10000 reflections to 2.5Å resolution



Attacking the ribosome by MR

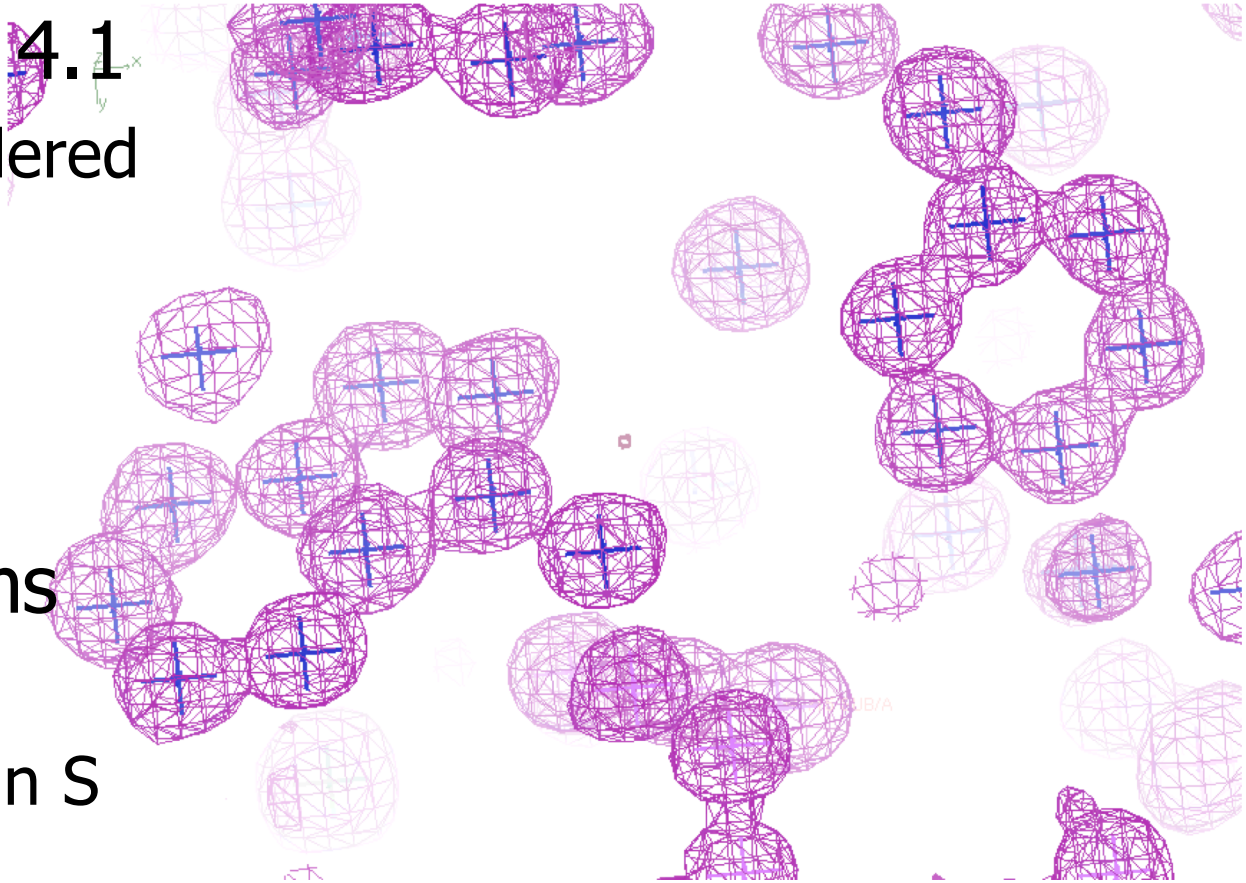
- 2j00: *Thermus thermophilus* 70S ribosome
  - two copies in a.u.
  - 1.3M reflections to 2.8Å resolution
- Models:
  - 1j5e: *Thermus thermophilus* 30S small subunit
  - 1ffk: *Haloarcula marismortui* 50S large subunit
- *Phaser* chooses limit of 7.5Å (79K reflections)
  - sufficient to use data to 12Å (19K reflections)

Arcimboldo

- Isabel Usón
- Place common fragments (*e.g.* helices) with *Phaser*
- Density-modify and trace with *SHELXE*
- High success rate if resolution better than about 2 Å

Aldose reductase

- 36 kDa, 0.78Å resolution (3bcj)
- <LLG> for 1 S is 4.1
  - higher if well-ordered
- Find 4 S (<3h)
- Complete with N atoms (3h)
- 2525 non-H atoms in structure
  - none heavier than S



How to attack a difficult MR problem

- Collect the best data possible
  - higher resolution helps
    - more signal with good models
    - more power for model completion algorithms
  - anomalous differences are very useful!
  - pathologies hinder progress
    - anisotropy reduces signal, makes maps harder to interpret
    - translational non-crystallographic symmetry (tNCS) must be accounted for
- Prepare the best possible model
  - consider possible domain movements
- Use likelihood as a target

Getting the best model before MR

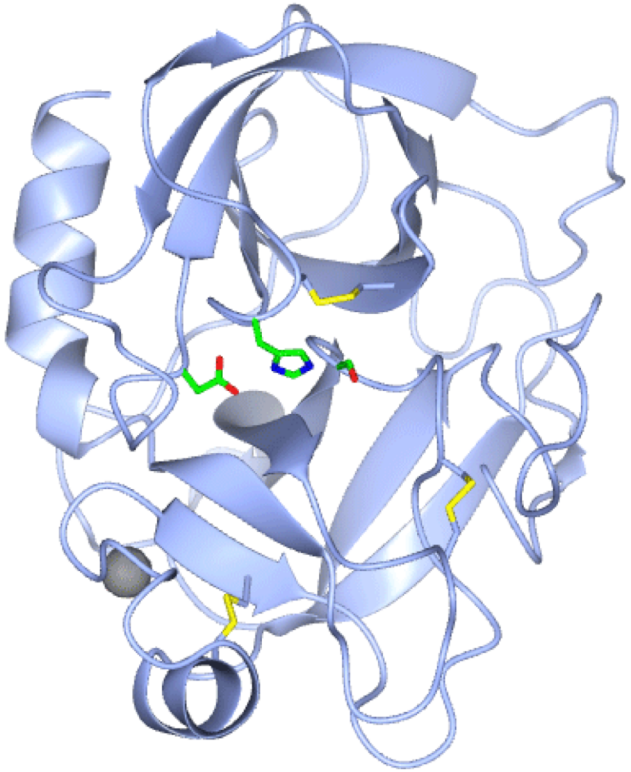
- Use sensitive algorithm to find and align
  - HHpred works well for distant homologues
- Try many alternatives
  - correlation between sequence identity and quality is approximate
  - conformational change
  - easier in a pipeline: phaser.MRage, Balbes, MrBUMP
- Improve the model
  - use an ensemble
  - edit the model to remove parts that don't belong
  - use sophisticated homology modelling
- Did you crystallise the right thing?
  - Search database of common contaminants (ContaMiner, SIMBAD) or entire PDB (WSMR, SIMBAD)

Model manipulation

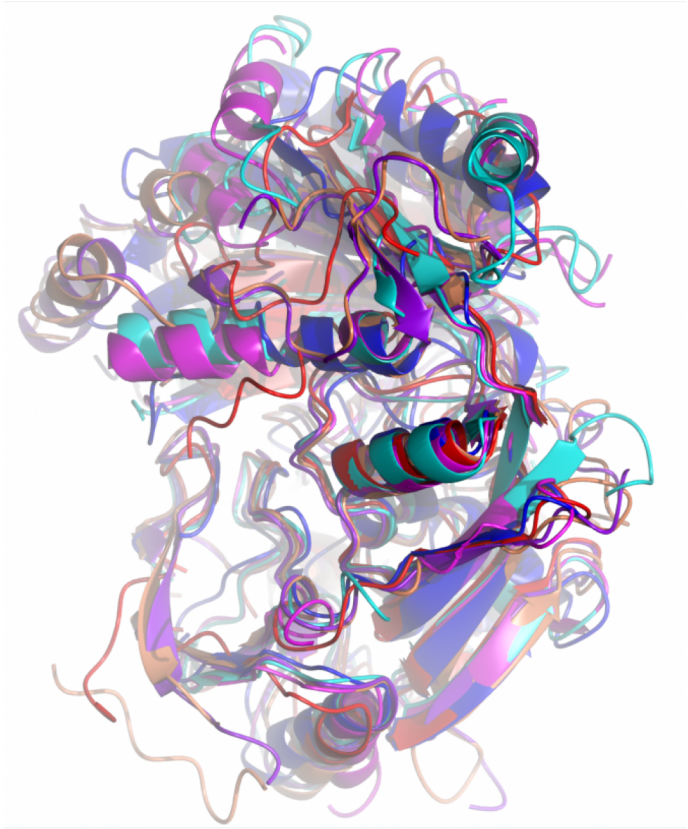
- Sculptor (also Chainsaw, Molrep)
  - use sequence alignment to:
    - trim parts of template not in target
    - adjust B-factors of poorly-conserved regions
  - use surface accessibility to:
    - adjust B-factors of surface regions
- Ensembler
  - multiple structure superposition to make ensemble of possible models
  - optionally trim non-conserved surface loops
- Divide into domains, if appropriate

Streptomyces griseus trypsin (1980-84)

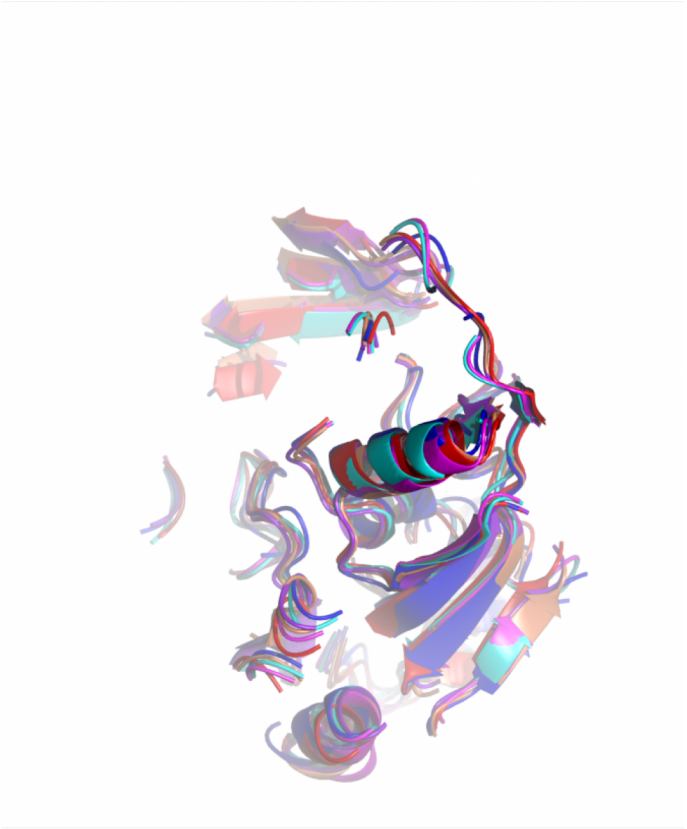
- MR using bovine trypsin (34% identical)
  - Sculptor: trim according to sequence alignment
    - Gábor Bunkóczi
  - Phaser: clear solution in < 1 minute
  - ARP/wARP: R_{free}<25% in 15 minutes
- Effect of Sculptor
  - LLG increases: 117 to 172
  - CPU decreases: 109s to 22s



DprE1 (Andrea Mattevi & Claudia Binda)
Ensemble of 6 models, 14-19% identical



Ensemble

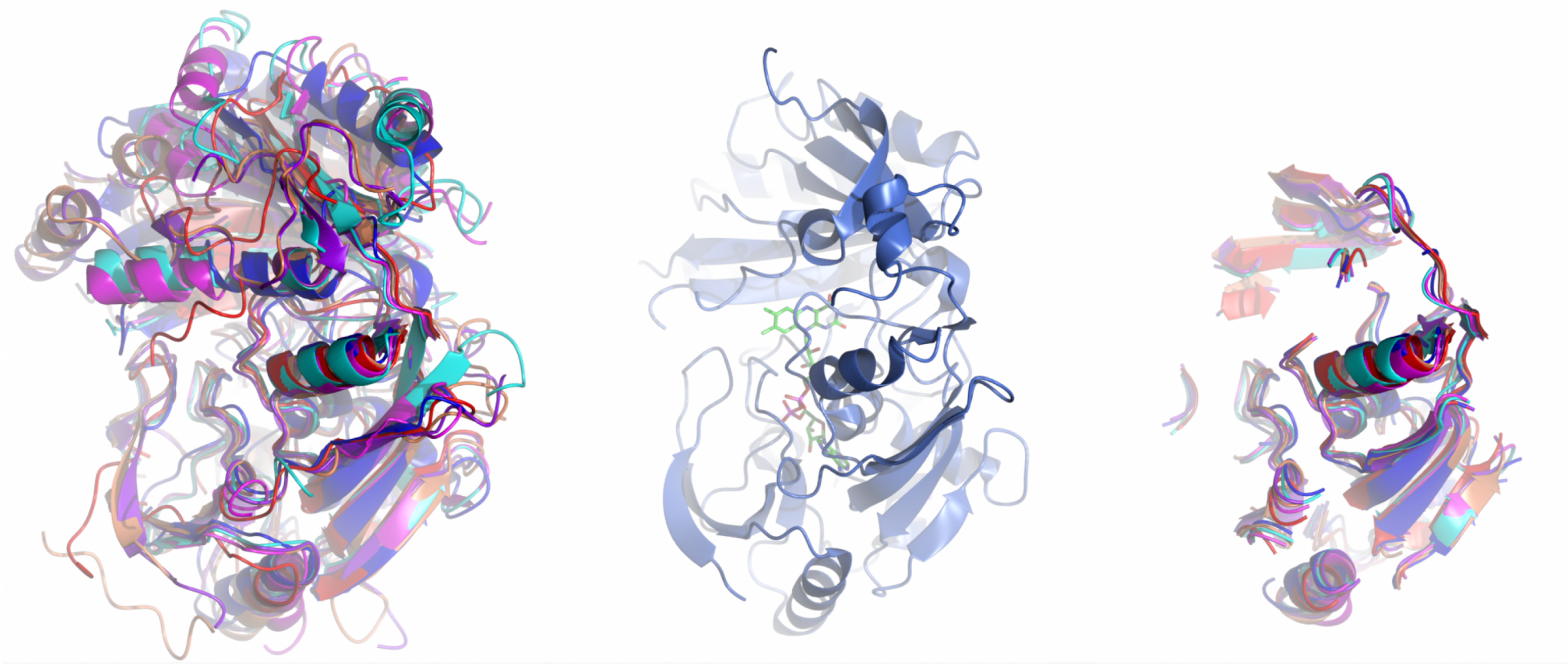


Trimmed

Homology modeling and MR

- *Rosetta*: sophisticated modeling program from David Baker's group
  - computationally intensive (Rosetta@home)
- Templates from NMR structures and distant homologues can be improved for MR
  - Bin Qian, Rhiju Das *et al.* (2007)
- Complete (possibly ambiguous) solution from poor model: phenix.mr_rosetta
  - Frank diMaio, Tom Terwilliger *et al.* (2011)
- Can get away with less extensive modelling
  - AMPLE

DprE1 (Andrea Mattevi & Claudia Binda)
Ensemble of 6 models, 14-19% identical



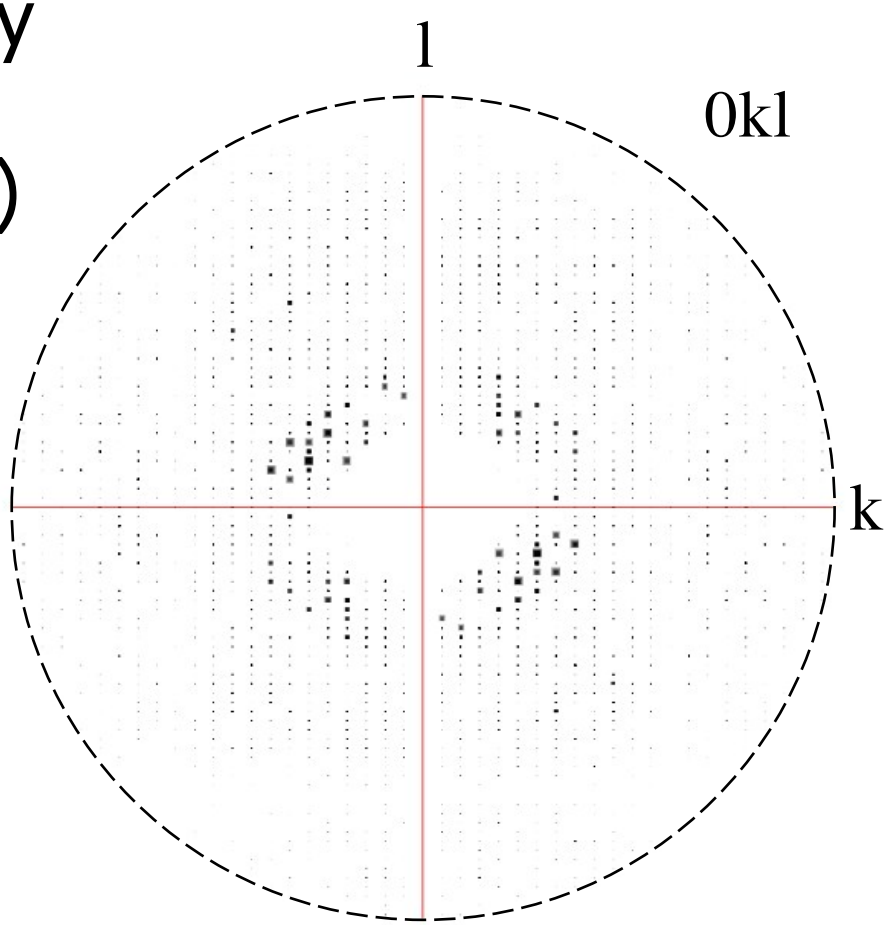
Ensemble DprE1 Trimmed

Likelihood is sensitive…

- …to correct orientation and position of molecular replacement model
  - successful in solving structures with distant relatives, small fragments, or many copies in asymmetric unit
- …to violations of assumptions
  - data implicitly assumed to be isotropic
    - important to account for anisotropy
  - components may not be equally well-ordered
    - important to correct for differences in overall B-factors

β-lactamase:BLIP complex

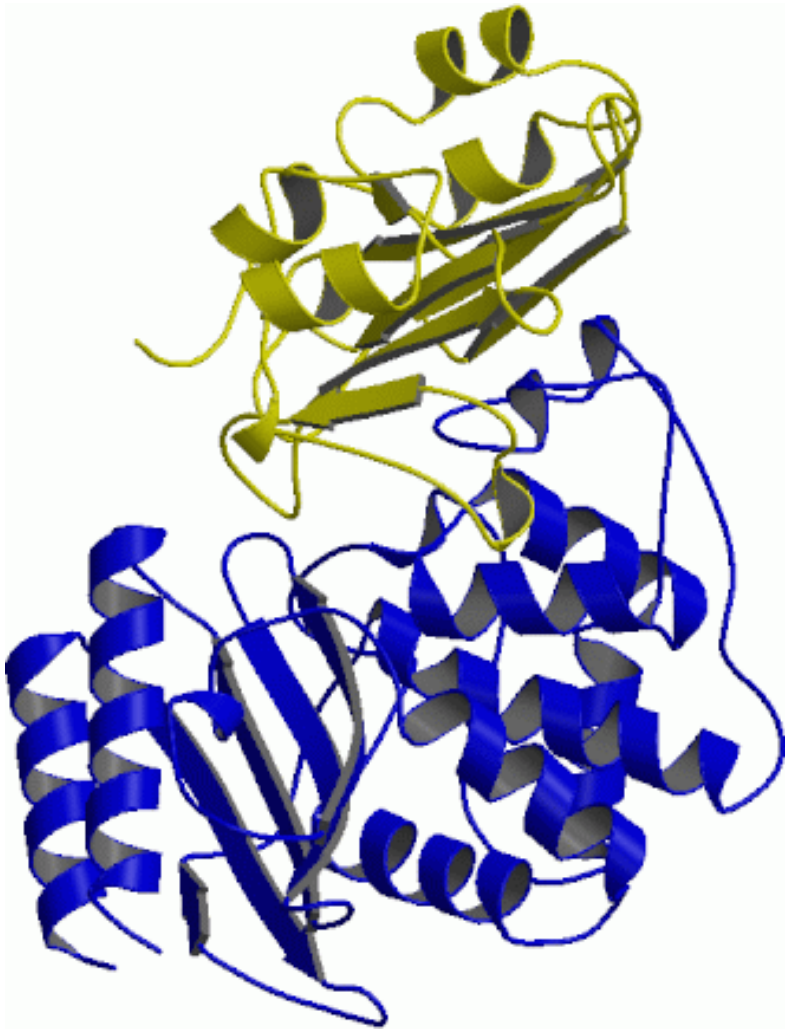
- Solved with great difficulty using AMoRe (Strynadka, James, Alzari)
- β-lactamase
  - 62% of the structure
  - easy to find
- BLIP
  - 38% of the structure
  - hard to find
- Anisotropic diffraction



β-lactamase:BLIP complex with *Phaser*

- Likelihood-based target
- fix β-lactamase
- Anisotropy corrected

- Clear peak
- Result in minutes
- Even solve with BLIP component first



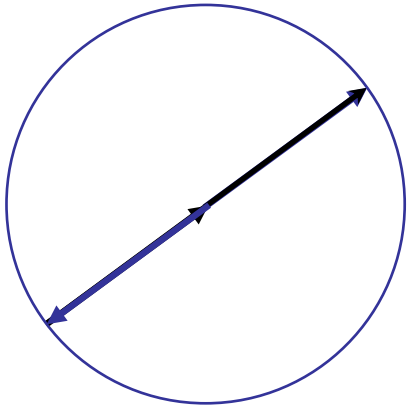
New pathologies become bottlenecks: translational NCS (tNCS)

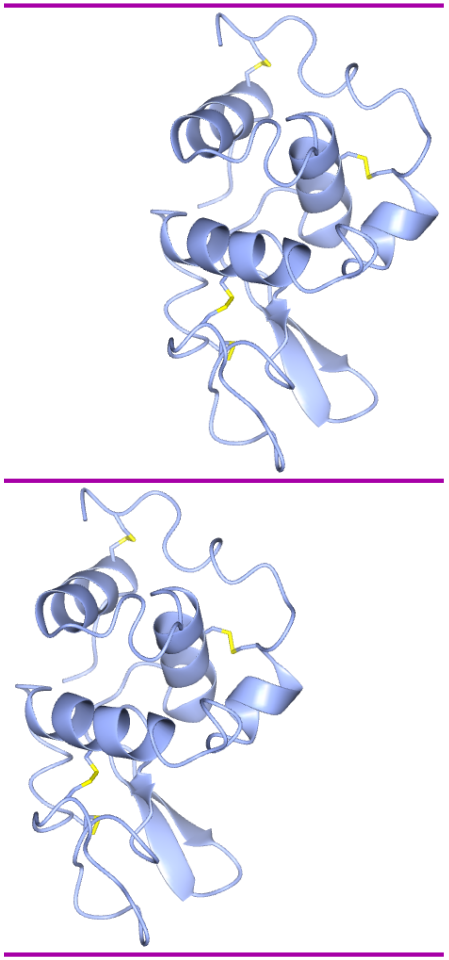
- Found in about 8% of PDB entries

Photo courtesy of Laurie Betts

Effect of tNCS on diffraction

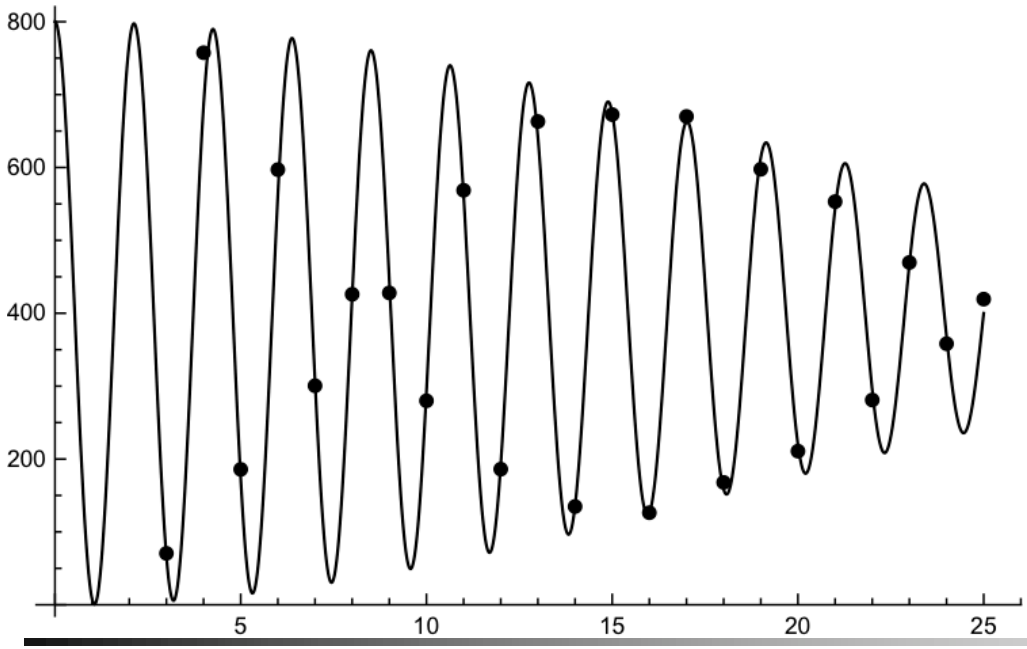
- Diffraction from copies in different orientations is uncorrelated
- Diffraction from copies in the same orientation is correlated

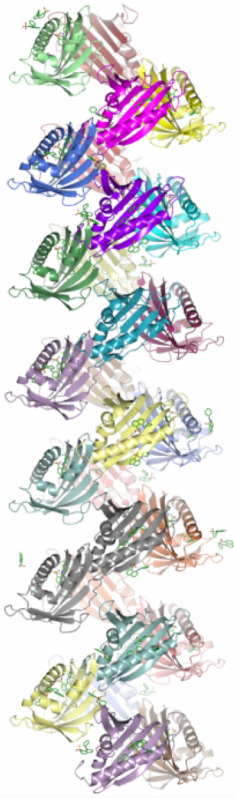




Accounting for translational NCS

- Model effect of translation combined with small rotation and random differences between copies





Hyp-1:
Sliwiak, Jaskolski, Dauter, McCoy, Read (2014)

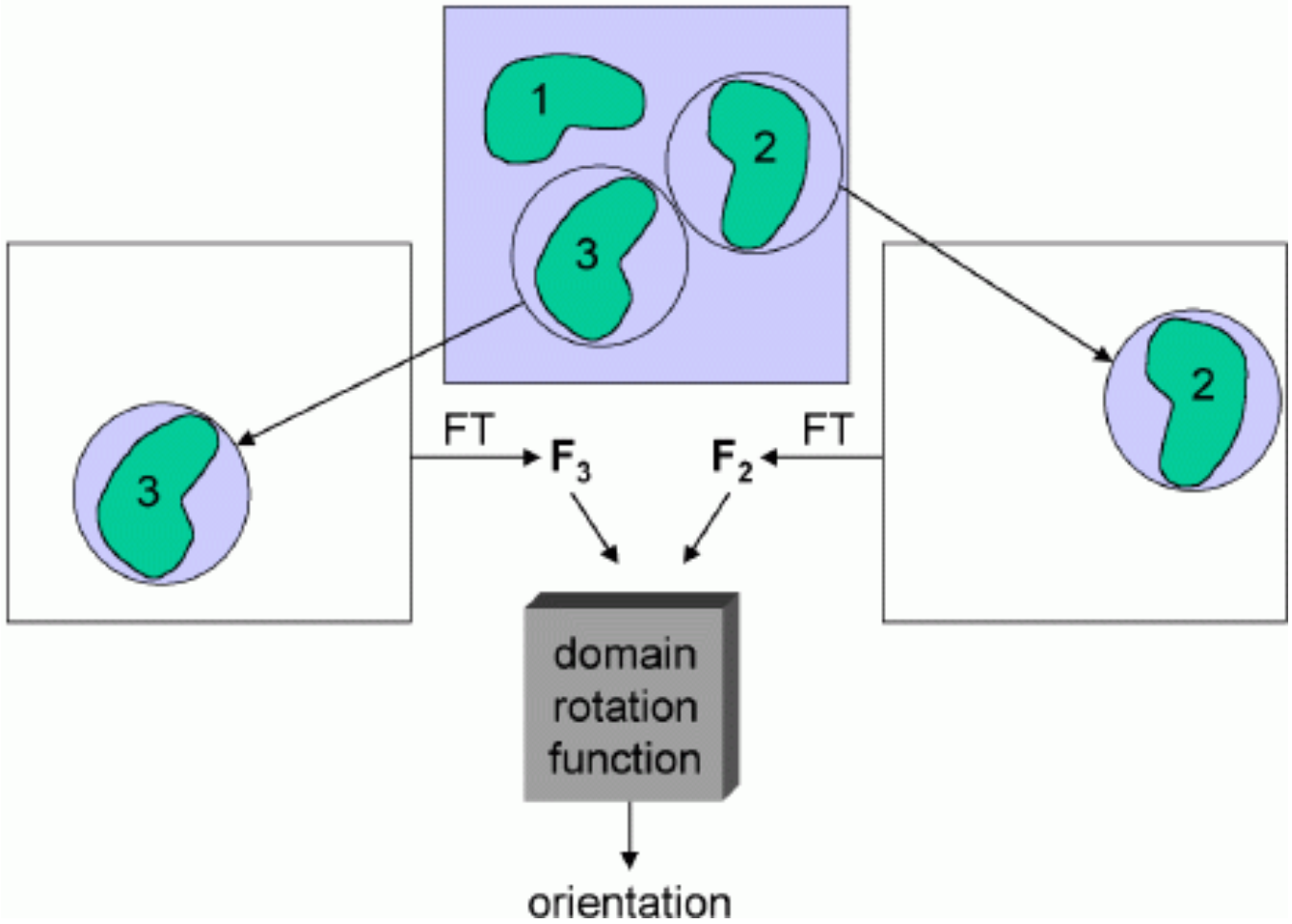
Pulling out the stops: combining sources of information

- Electron density as a model
- NCS and multi-crystal averaging
- MR-SAD
  - use MR solution to extract (even weak) experimental phase information
  - prime SIRAS or MIRAS phasing by using model to determine heavy-atom sites

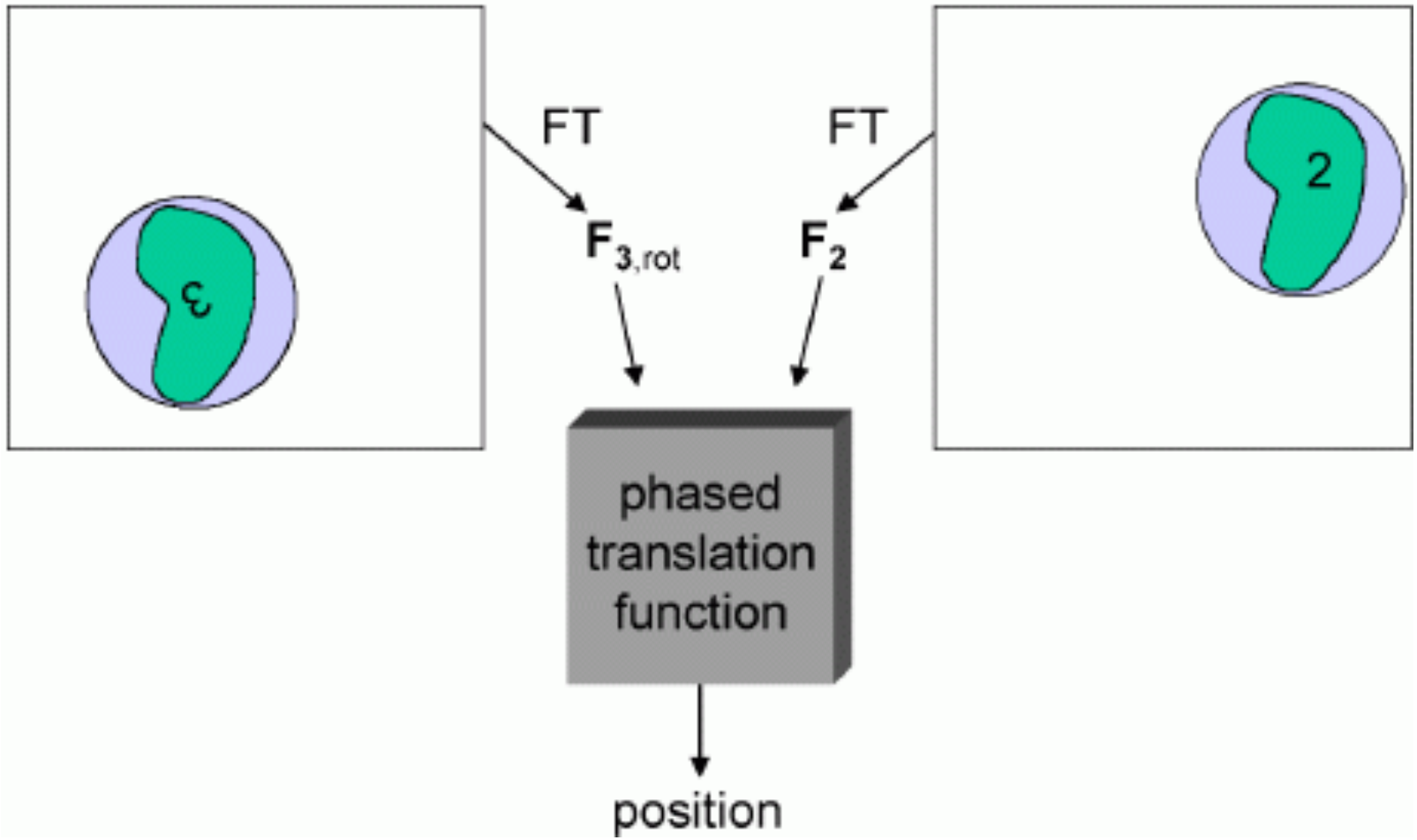
Real-space molecular replacement

- Use phase information in two ways:
  - use electron density as model
    - calculate structure factors from isolated density, then proceed as with atomic model
    - also works with cryoEM image reconstruction
      - *e.g.* Cascade structure (Jackson *et al.*, 2014)
  - fit model into electron density
    - "domain rotation function"
    - "phased translation function"

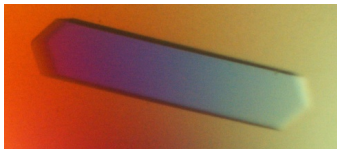
Domain rotation function



Phased translation function

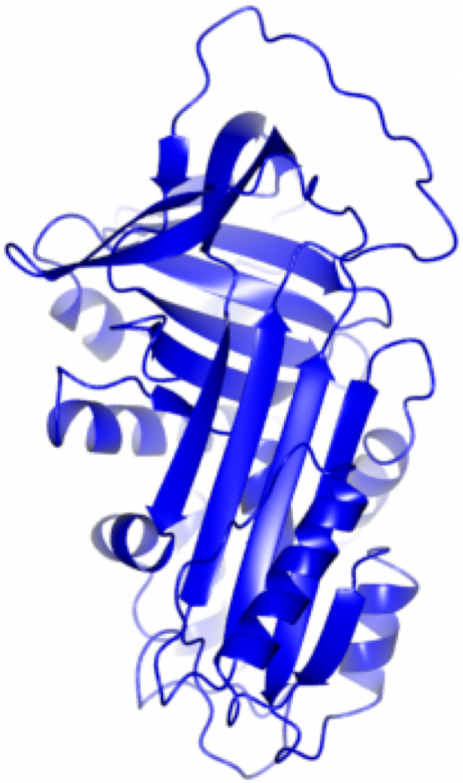


Human angiotensinogen: molecular replacement

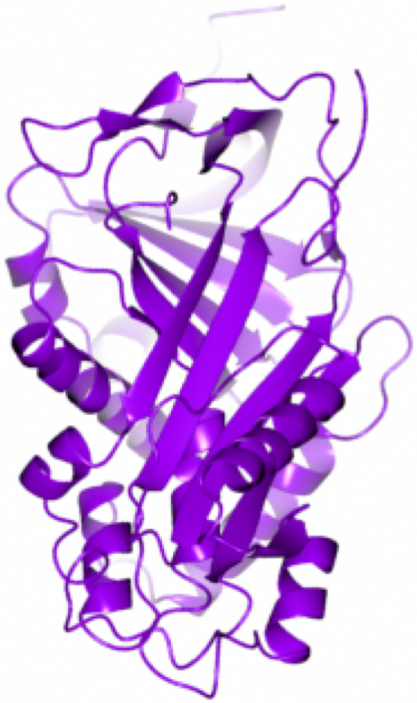


human

heparin cofactor II
(20% identical)

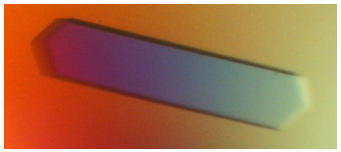


α_1-antitrypsin
(21% identical)

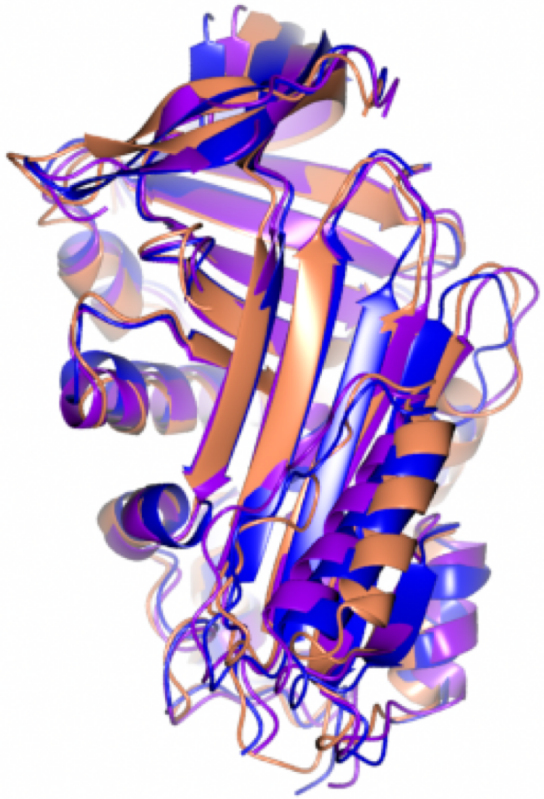


thyroxine-binding globulin
(20% identical)

Human angiotensinogen: molecular replacement

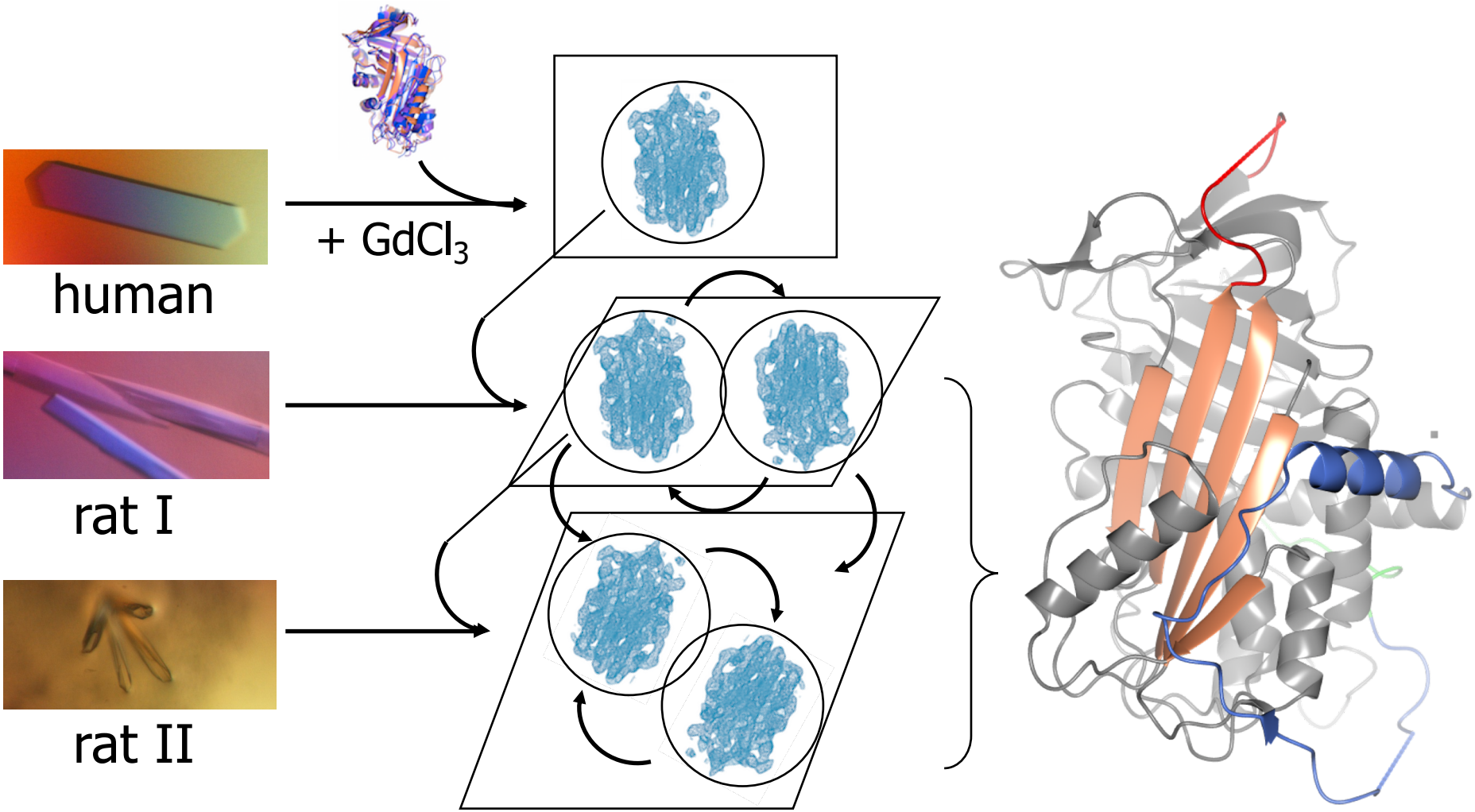


human



Trimmed ensemble

Solving angiotensinogen structures



Acknowledgements

- Phaser: Airlie McCoy, Gábor Bunkóczi, Rob Oeffner
- Arcimboldo: Isabel Usón, Claudia Millan, Massimo Sammito
- Angiotensinogen: Penny Stein, Robin Carrell, Aiwu Zhou, Yahui Yan
- Hyp-1: Mariusz Jaskolski, Joanna Sliwiak, Zbyszek Dauter