The *PHENIX* project

PSI Protein Structure Initiative

Crystallographic software for automated structure determination

Computational Crystallography Initiative (LBNL)
-*Paul Adams*, Ralf Grosse-Kunstleve, Peter Zwart, Nigel Moriarty, Nicholas Sauter, Pavel Afonine

Los Alamos National Lab (LANL)
-*Tom Terwilliger*, Li-Wei Hung,Thiru Radhakannan

Cambridge University
-*Randy Read*, Airlie McCoy, Laurent Storoni, -Hamsapriye

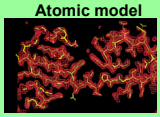
Texas A&M University
-*Tom Ioerger*, Jim Sacchettini, Kreshna Gopal, Lalji Kanbi, -Erik McKee, Tod Romo, Reetal Pai, Kevin Childs, Vinod Re

Structure Determination by MAD/SAD/MIR in PHENIX

Data files (H K L Fobs Sigma)
Crystal information (Space group, Cell)
Scattering factors (for MAD)

↓

Heavy-atom coordinates
Non-crystallographic symmetry
Electron-density map
Atomic model



***PROVIDED THAT*:**

Data are strong, accurate, < 3 Å
Strong anomalous signal
Little decay
Space group is correct
Scattering factors close (for MAD data)
You are willing to wait a little while…
(10 minutes to hours, depending on size)

↓

Model is 50-95% complete (depending on resolution)

Model is (mostly) compatible with the data…but is not completely correct

Model requires manual rebuilding

Model requires validation and error analysis

Major needs in automated structure solution

MAD/SAD/MIR

Robust structure determination procedures
Best possible electron density maps to build most complete model
Decision-making about best path for structure solution

Molecular Replacement

Use of distant models
Preventing model bias

All structures

Model completion/Ligand fitting
Error analysis
Decision-making for what data to use and what path to follow
How to incorporate vast experience of crystallographic community

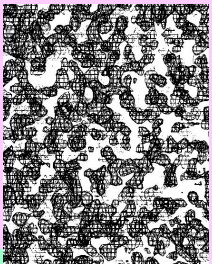
Best possible electron density maps to build the most complete model

Statistical density modification
Local patterns of density
ID of fragments
Iterative model-building and refinement
FULL-OMIT density modification and model-building

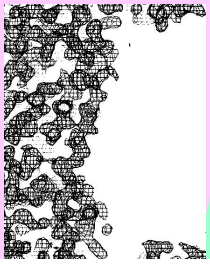
Why we need good measures of the quality of an electron-density map:

Which solution is best?

Are we on the right track?



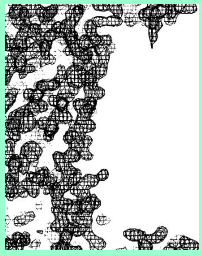
If map is good:
It is easy



Statistical density modification

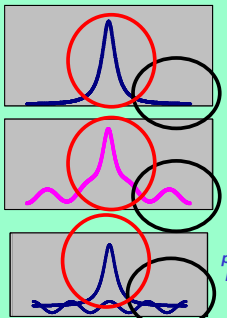
(A framework that separates map information from experimental information and builds on density modification procedures developed by Wang, Bricogne and others)

- *Principle: phase probability information from probability of the map and from experiment:*
- $P(\phi) = P_{map\ probability}(\phi)\ P_{experiment}(\phi)$
- *"Phases that lead to a believable map are more probable than those that do not"*
- *A believable map is a map that has…*
  - *a relatively flat solvent region*
  - *NCS (if appropriate)*
  - *A distribution of densities like those of model proteins*
- *Calculating map probability (ϕ) :*
  - *calculate how map probability varies with electron density ρ*
  - *Use chain rule to deduce how map probability varies with phase (equations of Bricogne, 1992).*



Map probability phasing:
Getting a new probability distribution for a single phase given estimates of all others

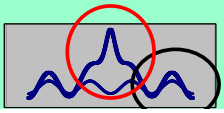
1. Identify expected features of map (flat far from center)
2. Calculate map with current estimates of all structure factors except one (k)
3. Test all possible phases ϕ for structure factor k (for each phase, calculate new map including k)
4. Probability of phase ϕ estimated from agreement of maps with expectations



A function that is (relatively) flat far from the origin

Function calculated from estimates of all structure factors but one (k)

Test each possible phase of structure factor k. P(ϕ) is high for phase that leads to flat region



A map-probability function

Log-probability of the map is sum over all points in map of local log-probability

$$LL^{MAP}(\{\mathbf{F_h}\}) \approx \frac{N_{REF}}{V}\int_{V} LL(\rho(\mathbf{x},\{\mathbf{F_h}\}))d^3\mathbf{x}$$

A map with a flat (blank) solvent region is a likely map

Local log-probability is believability of the value of electron density (ρ(x)) found at this point

$$LL(\rho(\mathbf{x},\{\mathbf{F_h}\})) = \ln[p(\rho(\mathbf{x})|PROT)p_{PROT}(\mathbf{x}) + p(\rho(\mathbf{x})|SOLV)p_{SOLV}(\mathbf{x})]$$

If the point is in the PROTEIN region, most values of electron density (ρ(x)) are believable

If the point is in the SOLVENT region, only values of electron density near zero are believable

Statistical density modification features and applications

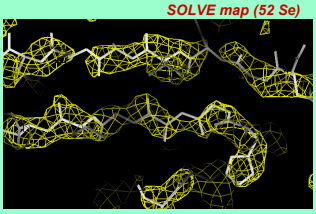
Features:

- *Can make use of any expectations about the map.*
- *A separate probability distribution for electron density can be calculated for every point in the map*

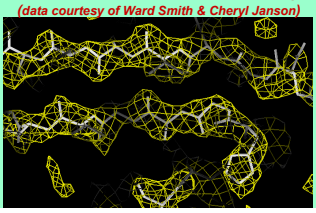
Applications:

- *Solvent flattening*
- *Non-crystallographic symmetry averaging*
- *Template matching*
- *Partial model phasing*
- *Prime-and-switch phasing*
- *General phase recovery*
- *Iterative model-building*

SOLVE map (52 Se)



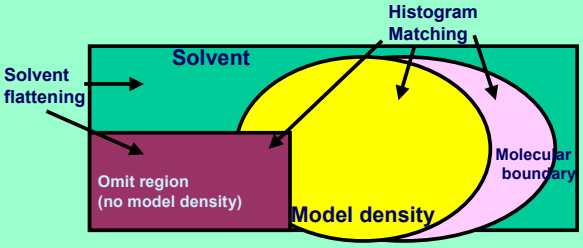
RESOLVE map
(data courtesy of Ward Smith & Cheryl Janson)



Reference: Terwilliger, T. C. (2000). Maximum-likelihood density modification. Acta Crystallographica, D55, 1863-1871.

Composite omit map with statistical density modification

Statistical density modification allows a separate probability distribution for electron density at each point in the map: can specify that "missing" density is within molecular boundary



Can be used with or without experimental phases…with or without omit

Image enhancement using local feature recognition

Electron density maps of proteins have many features in common

- *Connected density*
- *Preferred distances for spacing between regions of high density*
- *Preferred shapes of density*

Starting image in red

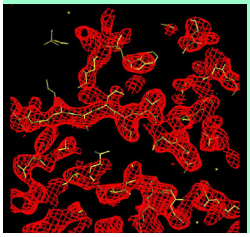


Image improved using expectations about local features

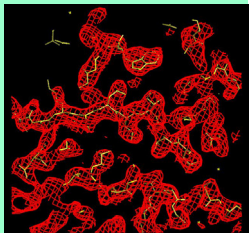


Image enhancement using local feature recognition

Approach:

- *Use the pattern of density near a point x to estimate the value of density at x*
- *Combine new estimate of density with previous one to improve the overall image*

"Local NCS averaging"

Starting image in red

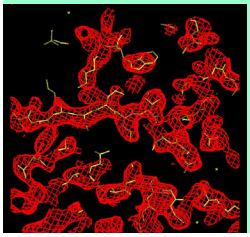


Image improved using expectations about local features

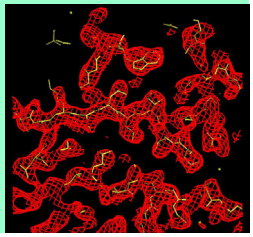


Image enhancement using local feature recognition

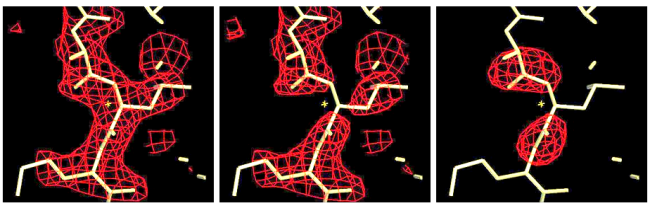
Approach:

- *Create N templates of local density using model data*
  - *Examine density near each point x in image (within 2 Å)*
  - *Exclude region very close to x (about 1 Å)*
  - *Cluster and average local patterns of density (after rotation to maximize CC)*
- *Identify relationship between finding pattern k of density near x, and density at x*
  - *Find all locations in the image where template k best matches the local density near x*
  - *Calculate average value of density at x for these cases = $\rho_{mean}(k)$*
- *Identify pattern near each point in actual map and use it to estimate density at that point*
  - *For each point x in the image, identify which template k best matches the local density near x*
  - *Use $\rho_{mean}(k)$ as estimate of density at x*

Image enhancement using local feature recognition

Remove all information about density at *x* from $\rho(x+\Delta x)$
-> $g(x+\Delta x)$, unbiased estimate of local pattern at x
Select most similar template *k* from library of unbiased patterns
Generate new estimate of density at *x* from average value at center of template *k*

A template associated with positive density…



$\rho_{current}(x+\Delta x)$	$g(x+\Delta x)$	$t(\Delta x)$
Local density in current map	Bias-removed local density…subtract $\rho_{current}(x)$ convoluted with origin of Patterson from all nearby points	Closest template in library (after testing 168 rotations) $\langle\rho\rangle$ for this template : 0.8 +/- 0.9

Image enhancement using local feature recognition

Templates associated with low density (top rows) and high density (bottom rows)
RED=positive contours BLUE=negative contours for the same template

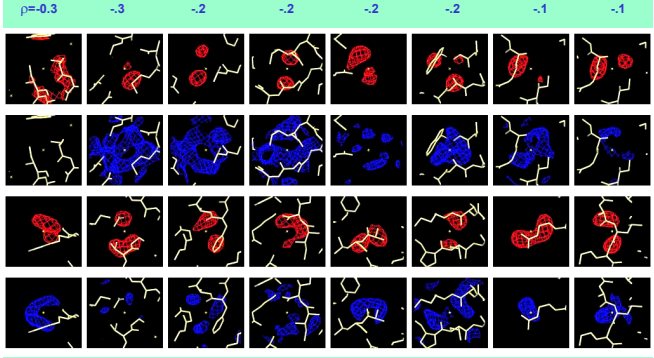
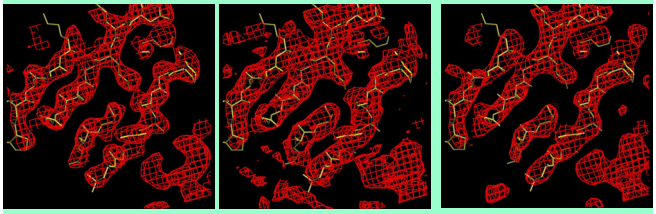


Image enhancement using local feature recognition

Image recovery from a good map…

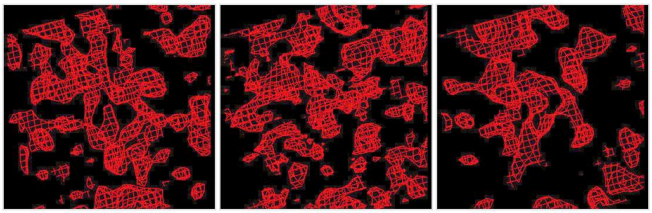
- gives an image that has (mostly) correct features
- errors are (almost) uncorrelated with original errors



RESOLVE map gene 5 protein at 2.6 A CC to perfect map = 0.8	Recovered image derived from RESOLVE map CC to perfect map = 0.36	Map phased using only using information from recovered image CC to perfect map = 0.64 CC of errors with errors in RESOLVE map = 0.11

Image enhancement using local feature recognition

Image recovery from a random map…gives an uncorrelated image



Random map at 2.6 A	Recovered image derived from random map CC to original random map=-0.01	Map phased using only recovered image CC to original random map=-0.04

Iterative procedure for image enhancement using local feature recognition

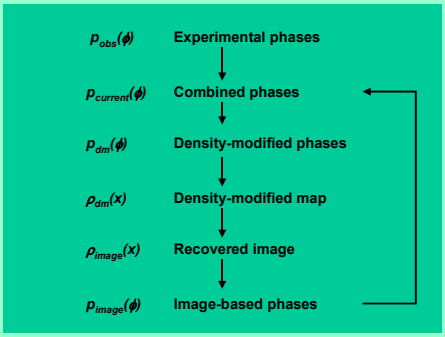
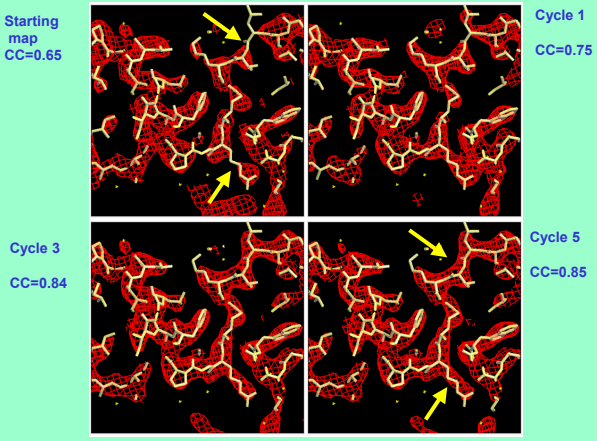


Image enhancement using local feature recognition (nusA protein structure)

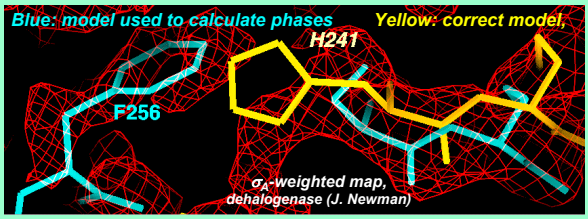


Removing model bias with prime-and-switch phasing

The problem:

Atomic model used to calculate phases -> map looks like the model

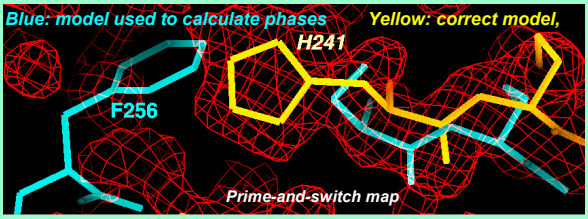
Best current solution: σ_A-weighted phases



Prime-and-switch phasing

A solution:

Start with σ_A-weighted map
Identify solvent region (or other features of map)
Adjust the phases to maximize the probability of the map – without biasing towards the model phases

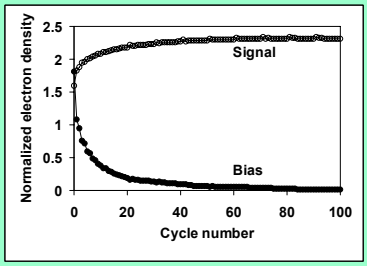


Prime-and-switch phasing

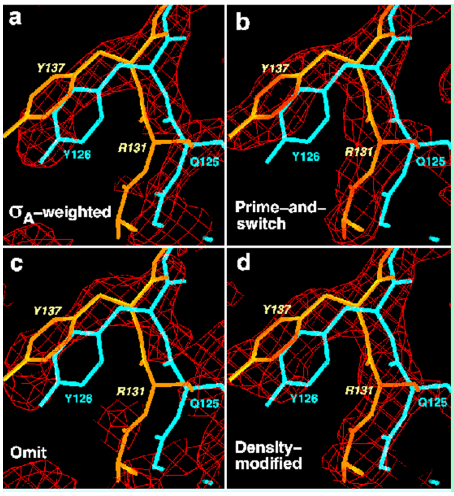
Why it should work...

Priming: Starting phases are close to correct ones...but have bias towards misplaced atoms

Switching: Map-probability phase information comes from a different source...which reinforces just the correct phase information



Signal: peak height at correct atomic positions
Bias: peak height at incorrect atoms in starting model



Prime-and-switch example

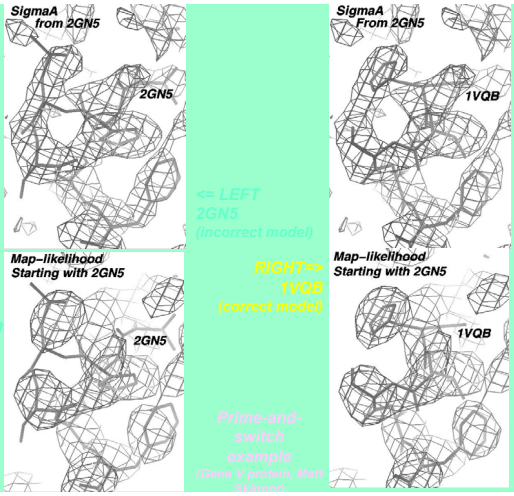
(IF5A, T. Peat)

Orange: correct model

Blue: model used to calculate phases

Top:
SigmaA phases from incorrect model

Bottom:
Prime-and-switch phases starting from incorrect model



<= LEFT 2GN5 (incorrect model)

RIGHT=> 1VQB (correct model)

Prime-and-switch example

(Gene V protein, Matt Skinner)

PHENIX AutoBuild wizard standard sequence

(Following ideas from Lamzin & Perrakis)

Fp, phases, HL coefficients

↓

Density modify (with NCS, density histograms, solvent flattening, fragment ID, local pattern ID)

↓

Build and score models
Refine with phenix.refine

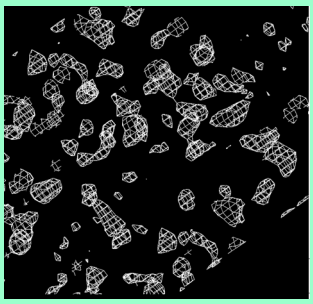
↓

Density modify including model information

↓

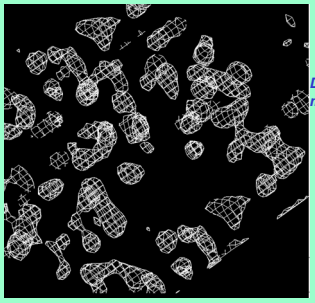
Evaluate final model

SAD data at 2.6 A
gene 5 protein



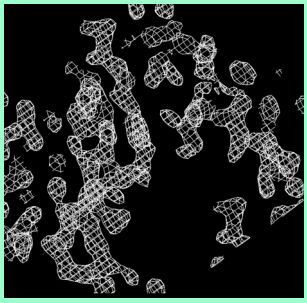
SOLVE SAD map

SAD data at 2.6 A
gene 5 protein



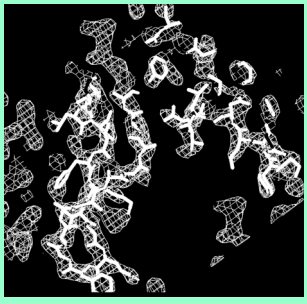
Density-modified SAD map

SAD data at 2.6 A
gene 5 protein



Cycle 50 of iterative model-building, density modification and refinement

SAD data at 2.6 A
gene 5 protein



Cycle 50 of iterative model-building, density modification and refinement

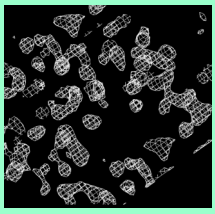
(with model built from this map)

Why iterative model building, density modification, and refinement can improve a map (following ideas of Perrakis & Lamzin):

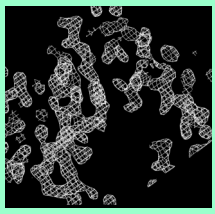
1. New information is introduced: flat solvent, density distributions, stereochemically reasonable geometry and atomic shapes

2. Model rebuilding removes correlations of errors in atomic positions introduced by refinement

3. Improvement of density in one part of map improves density everywhere.



Density-modified map

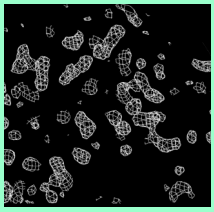


Iterative model-building map

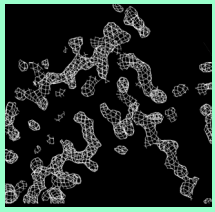
Iterative model-building and refinement is very powerful but isn't perfect…

Model-based information is introduced in exactly the same place that we will want to look for details of electron density

→ How can we be sure that the density is not biased due to our model information?
(Will density be higher just because we put an atom there?)
(Will solvent region be flatter than it really is because we flattened it?)
(Will we underestimate errors in electron density from a density-modified map?)
(Are we losing some types of information by requiring the map to match partially incorrect prior knowledge?)



Density-modified map



Iterative model-building map

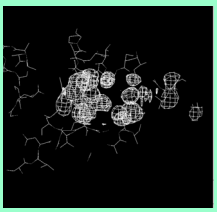
A FULL-OMIT iterative-model-building map: everywhere improved, everywhere unbiased

→ Use prior knowledge about one part of a map to improve density in another

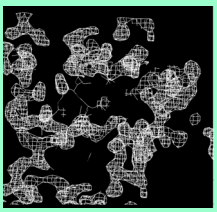
Related methods: "Omit map", "SA-composite omit map", density-modification OMIT methods, "Ping-pong refinement"

Principal new feature:
The benefits of iterative model-building are obtained yet the entire map is unbiased

Requires:
Statistical density modification so that separate probability distributions can be specified for omit regions (allow anything) and modified regions (apply prior knowledge)



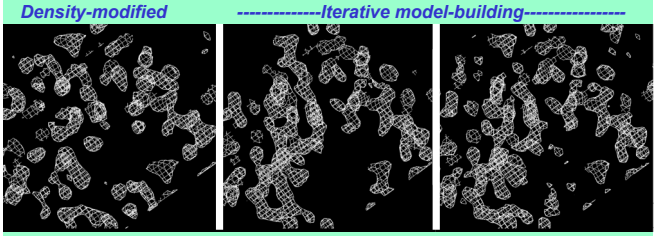
OMIT region –no model, no NCS, solvent flattening optional



Outside OMIT region – full density modification

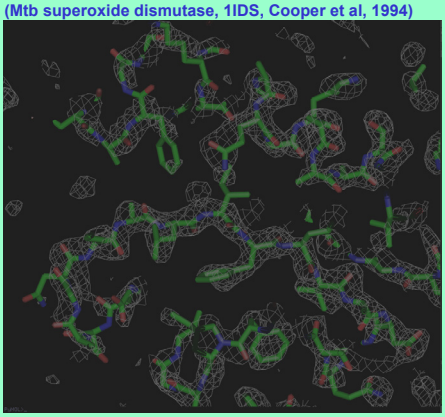
Including all regions in density modification comparison with FULL-OMIT

Density-modified *----------Iterative model-building----------*

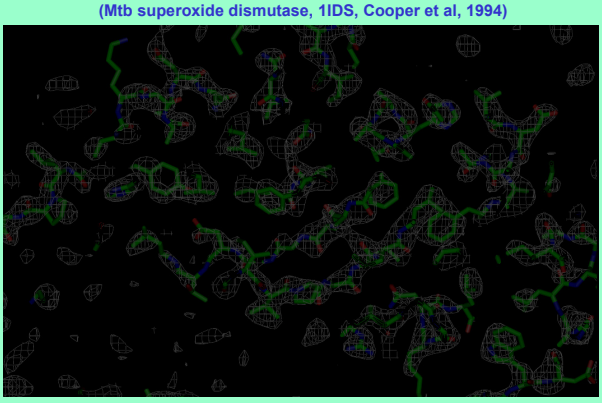


All included *FULL-OMIT*

FULL-OMIT iterative-model-building maps
Molecular Replacement:
(Mtb superoxide dismutase, 1IDS, Cooper et al, 1994)



FULL-OMIT iterative-model-building maps
Molecular Replacement:
(Mtb superoxide dismutase, 1IDS, Cooper et al, 1994)

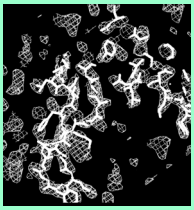


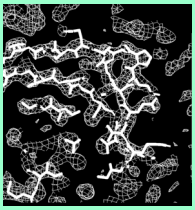
FULL-OMIT iterative-model-building maps

Uses:

Unbiased high-quality electron density from experimental phases
High-quality molecular replacement maps with no model bias
Model evaluation

Computation required:
~24 x the computation for standard iterative model-building



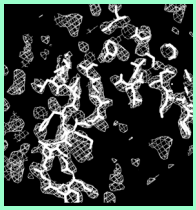


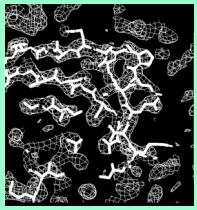
FULL-OMIT iterative-model-building maps

Requirement for preventing bias:

Density information must have no long-range correlated errors
(the position of one atom must not have been adjusted to compensate for errors in another)

→ **Starting model (if MR) must be unrefined in this cell**





Major needs in automated structure solution

MAD/SAD/MIR

Robust structure determination procedures
Best possible electron density maps to build most complete model

Molecular Replacement

Use of distant models

Preventing model bias

All structures

Model completion/Ligand fitting

Error analysis

Decision-making for what data to use and what path to follow

How to incorporate vast experience of crystallographic community

Acknowledgements

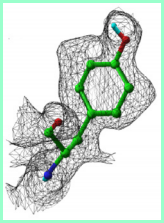
PHENIX: www.phenix-online.org

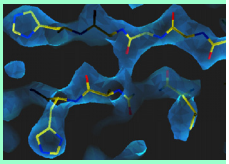
Computational Crystallography Initiative (LBNL): Paul Adams, Ralf Grosse-Kunstleve, Nigel Moriarty, Nick Sauter, Pavel Afonine, Peter Zwart

Randy Read, Airlie McCoy, Laurent Storoni, Hamsaprie (Cambridge)

Tom Ioerger, Jim Sacchettini, Kresna Gopal, Lalji Kanbi, Erik McKee Tod Romo, Reetal Pai, Kevin Childs, Vinod Reddy (Texas A&M)

Li-wei Hung,Thiru Radhakannan (Los Alamos)





PSI Protein Structure Initiative

Generous support for PHENIX from the NIGMS Protein Structure Initiative

PHENIX web site:
http://phenixonline.org

SOLVE/RESOLVE web site:
http://solve.LANL.gov

SOLVE/RESOLVE user's group:
solve@LANL.gov