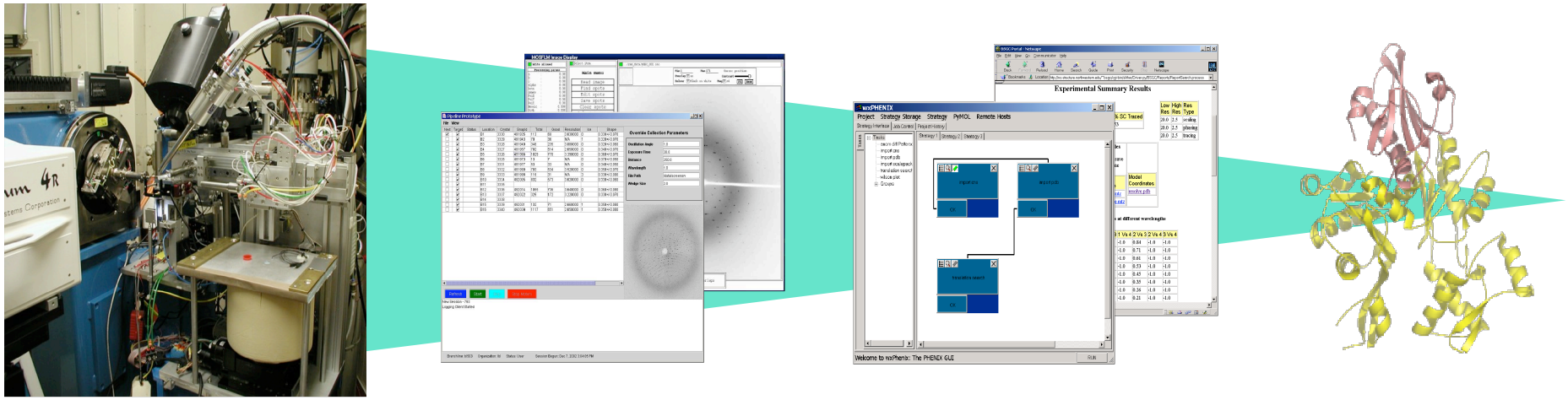
Automated Structure Solution and refinement with PHENIX

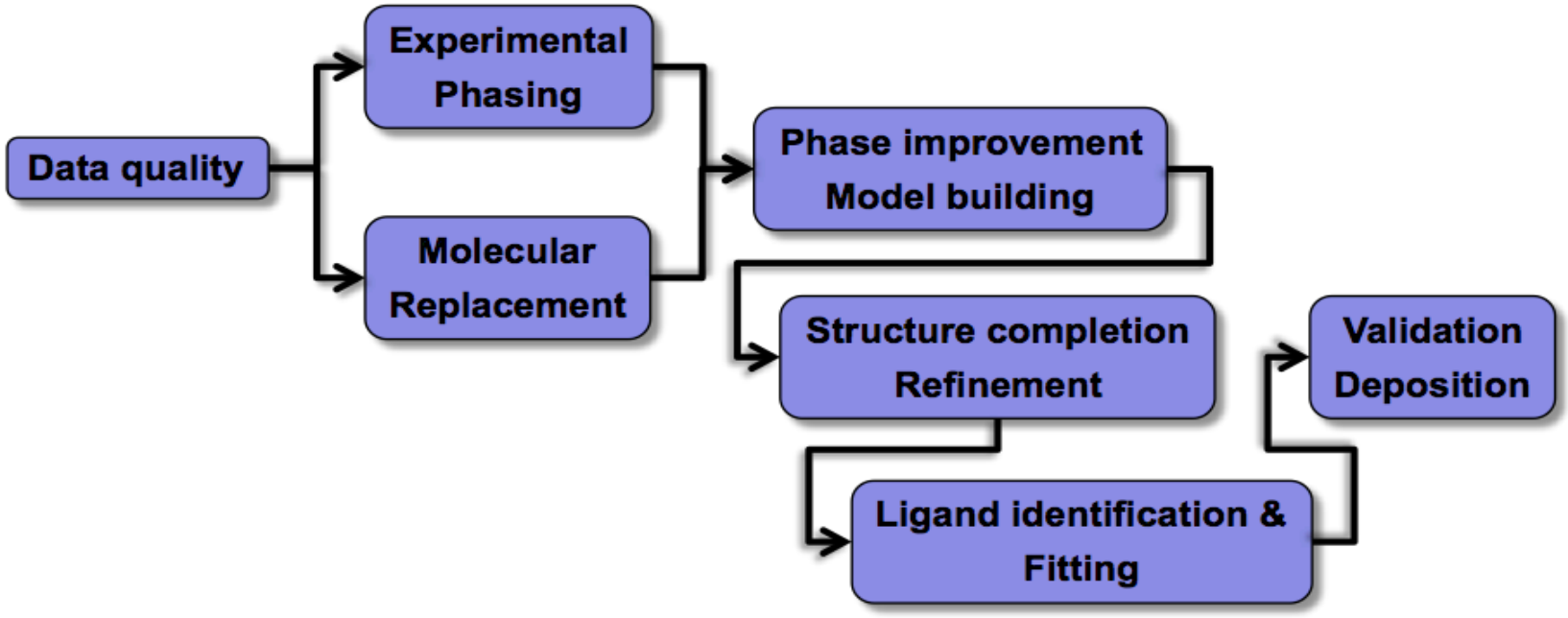
Pavel Afonine

Computation Crystallography Initiative
Physical Biosciences Division
Lawrence Berkeley National Laboratory, Berkeley CA, USA

Australasian Crystallography school
17th-24th July, 2010

Why Automation ?





- Automation can increase efficiency, and reduce human error

Why Automation ?

- Can speed up the process and can help reduce errors
- Makes difficult cases more feasible for experts
- Routine structure solution cases are accessible to a wider group of structural biologists
- Software can try more possibilities than we are typically willing to bother with
- Multiple trials or use of different parameters can be used to estimate uncertainties
- If a task is modular and automated, you can run it many times…
  - … checking different space groups, datasets to use
  - … checking if your model is biasing the map
  - … checking if you always get the same model

- What is required:
  - Software carrying out individual steps
  - Seamless connection between steps
  - A way to decide what is good
  - Strategies for structure determination and decision-making

The PHENIX Project

Lawrence Berkeley Laboratory

Paul Adams, Ralf Grosse-Kunstleve, Pavel Afonine, Nat Echols, Nigel Moriarty, Nicholas Sauter, Peter Zwart

Los Alamos National Laboratory

Tom Terwilliger, Li-Wei Hung

Cambridge University

Randy Read, Airlie McCoy, Gabor Bunkoczi, Rob Oeffner

Duke University

Jane & David Richardson, Vincent Chen, Jeff Headd, Chris Williams, Bryan Arendall, Laura Murray

An NIH/NIGMS funded Program Project

Phenix

PHENIX and Neutron crystallography

Macromolecular Neutron Crystallography Consortium (MNC)

Los Alamos National Lab
Paul Langan, Marat Mustyakimov, Benno Schoenborn

Lawrence Berkeley National Lab (LBNL)
Paul Adams, Pavel Afonine

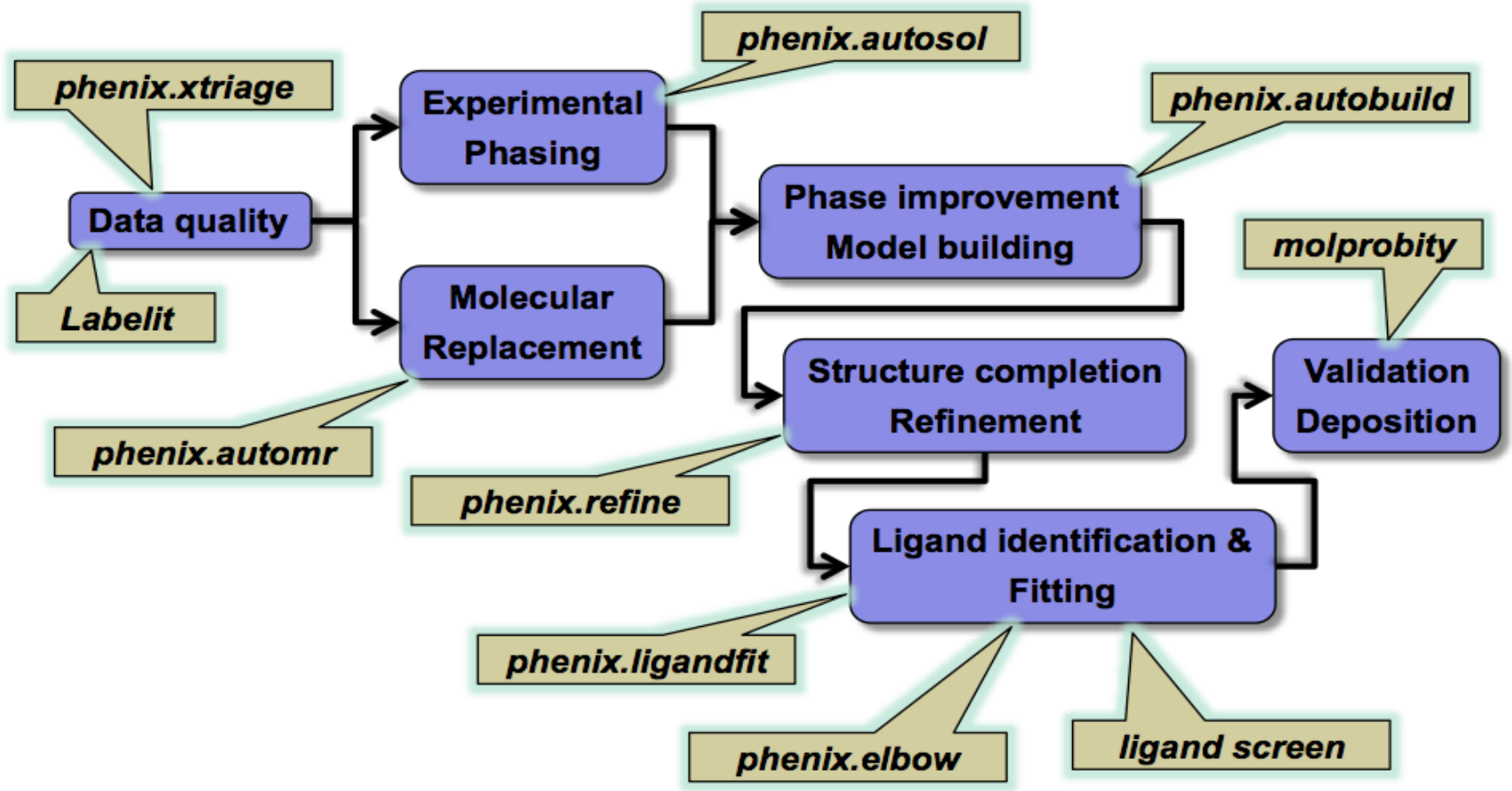
`http://mnc.lanl.gov/`

The PHENIX project

- PHENIX is a new package for automated structure solution that incorporates handling of both: X-ray and neutron data
- PHENIX is not a pipe-line made of existing programs, but a highly integrated software
- Library based development (Python, C++) and new algorithms
- Designed to be used by both novices and experienced users
- Long-term development and support

PHENIX: principal tools

Complete set of tools for crystallographic structure determination: from experimental data to PDB deposited structure



PHENIX: principal tools

- Data quality and pathology analysis (phenix.xtriage)
  Completeness, anomalous signal, outliers, tNCS, twinning, missed symmetries.
- Substructure determination (phenix.hyss)
  Automated, Patterson/Direct methods dual space recycling.
- Experimental phasing (PHASER, SOLVE)
  ML SAD phasing in Phaser, ML SIR/MIR/MAD phasing in SOLVE.
  ML SAD + MR phasing in Phaser
- Molecular Replacement phasing (PHASER)
  ML molecular replacement
- Statistical density modification (RESOLVE)
  Solvent flattening, histogram matching, automated NCS analysis.
- Automated model building (RESOLVE, Textal, phenix.refine)
  Pattern matching methods, loop fitting, NCS, protein and nucleic acids.
- Structure refinement (phenix.refine)
  Rigid body, coordinates, simulated annealing, iso/aniso-ADP, TLS, neutron data (joint refinement), twinned refinement, constrained occupancy refinement.
- Ligand coordinate and restraint generation (phenix.elbow & phenix.reel)
  Semi-empirical method for coordinate generation, CCP4 monomer library file generation.
- Model Validation (Molprobity, local electron density analysis, POLYGON)
  A set of tools for complete, local and global, structure validation.

PHENIX: principal tools

- PHENIX programs can be run using:
  - **GUI**: easy for beginners, guided process - less chance of errors
  - **Command line**: convenient for scripting of multiple and large scale tasks

- Command line tools are still easy to run:
  - Autobuild *(from starting phases to complete and refined model)*:

    ```
    phenix.autobuild data=scale.mtz model=mr.pdb seq_file=correct.seq
    ```

  - Ligandfit (automatically find and build ligands into density):

    ```
    phenix.ligandfit data=nsf.mtz model=noligand.pdb ligand=atp.pdb
    ```

  - AutoMR (molecular replacement with Phaser + Autobuild = refined model):

    ```
    phenix.refine nsf-d2.mtz nsf.pdb
    ```

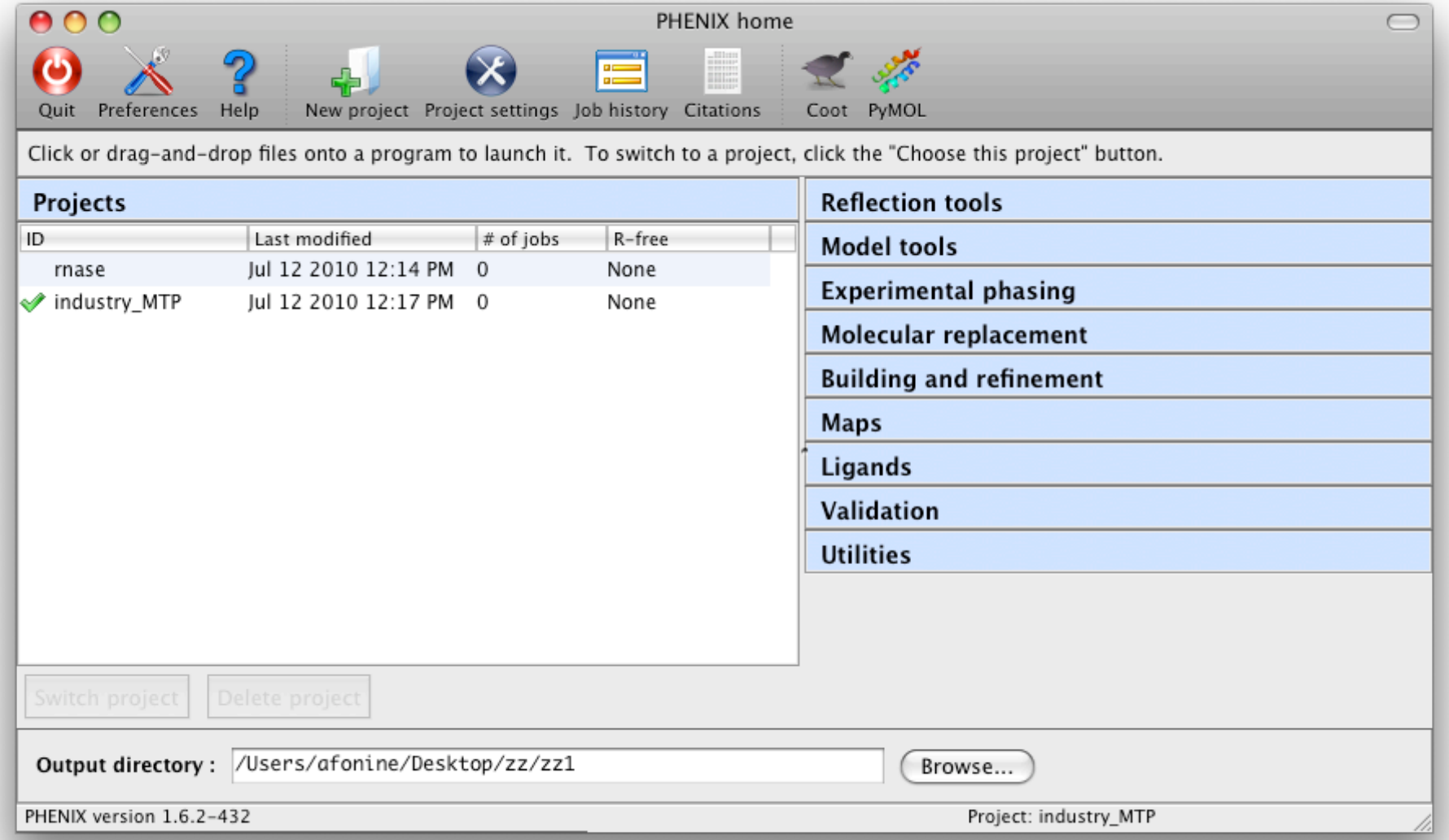
  - phenix.refine (highly automated structure refinement, X-ray, Neutron):

    ```
    phenix.refine nsf-d2.mtz nsf.pdb
    ```

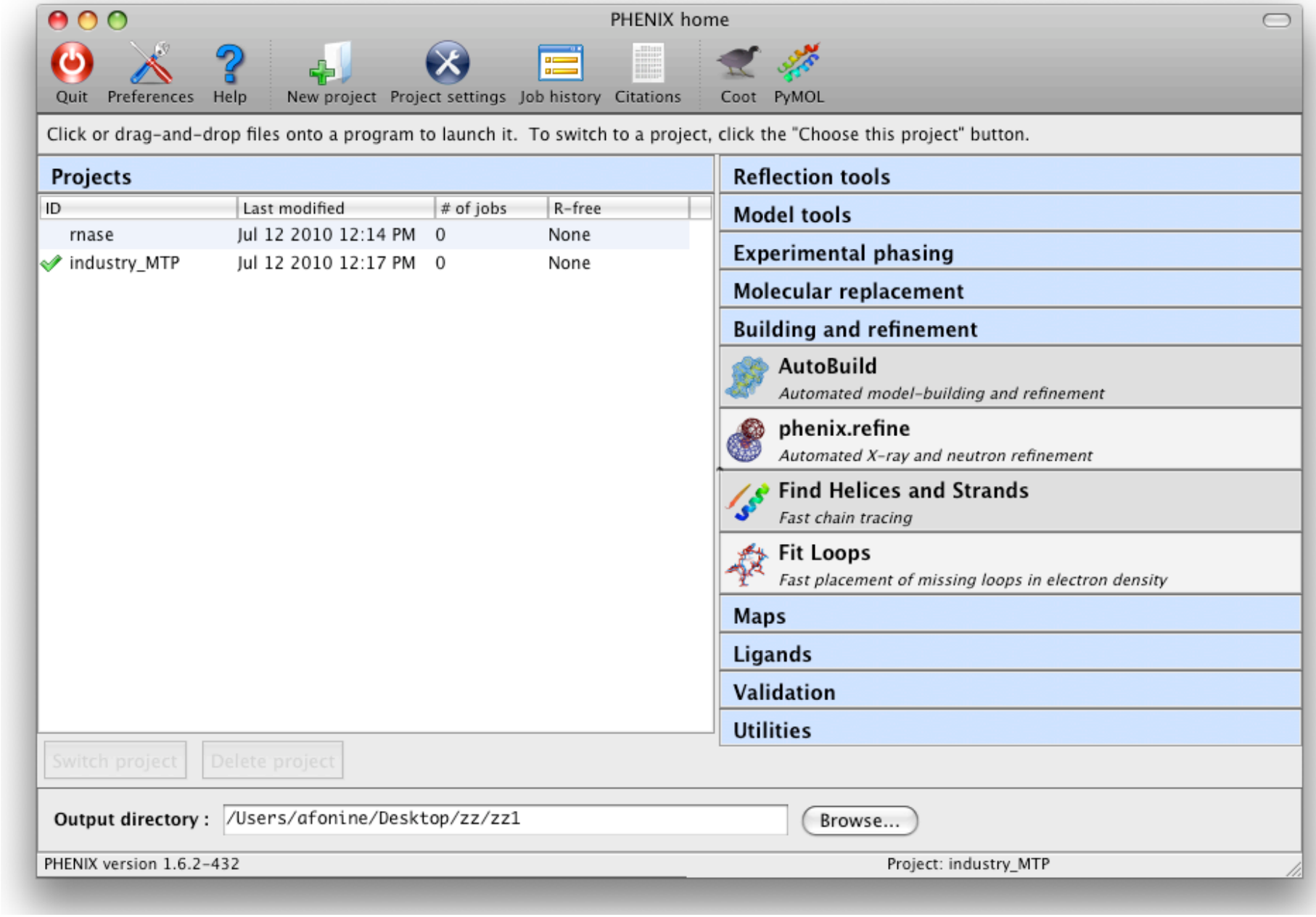
  - phenix.xtriage (complete data analisys):

    ```
    phenix.xtriage porin_fp.mtz
    ```

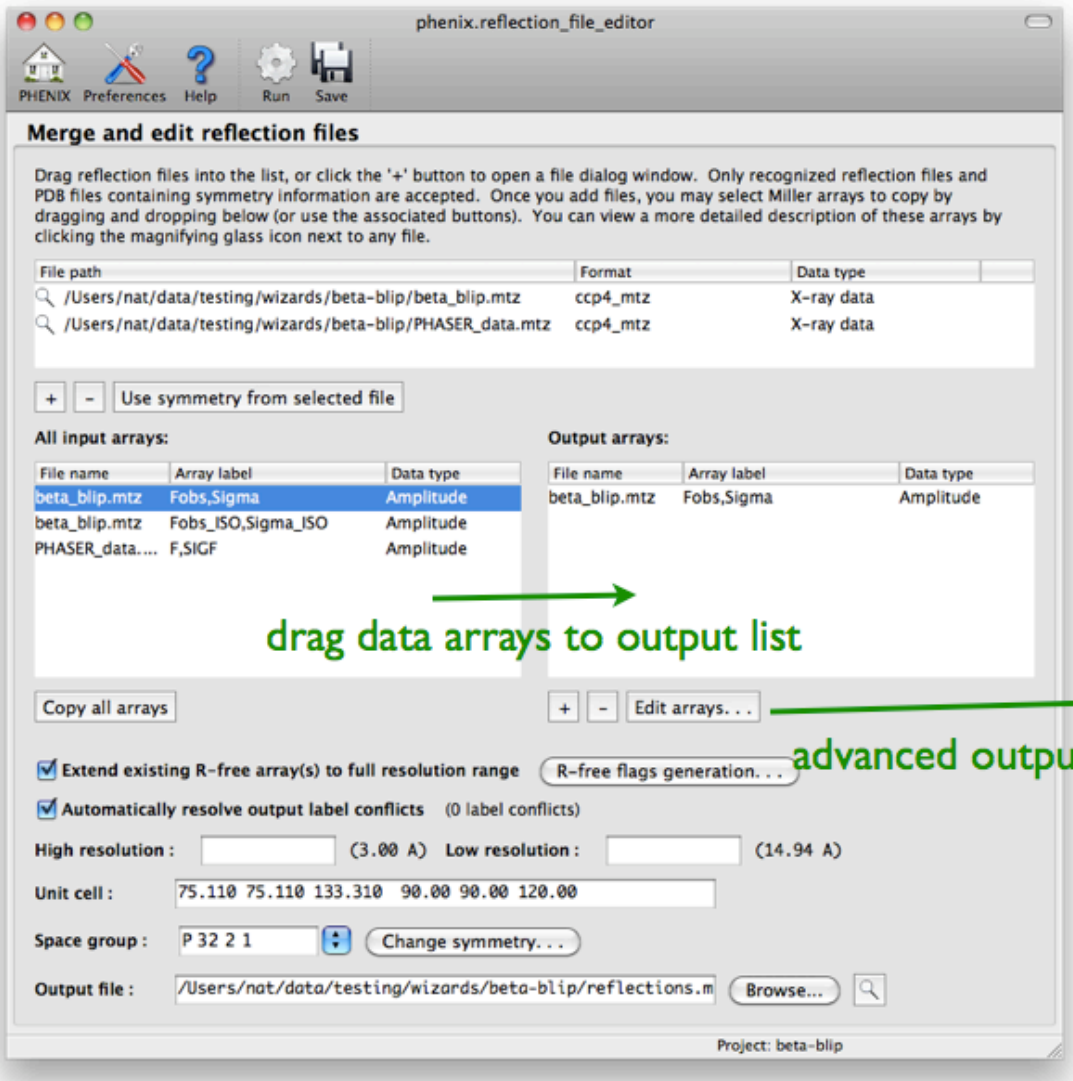
GUI examples



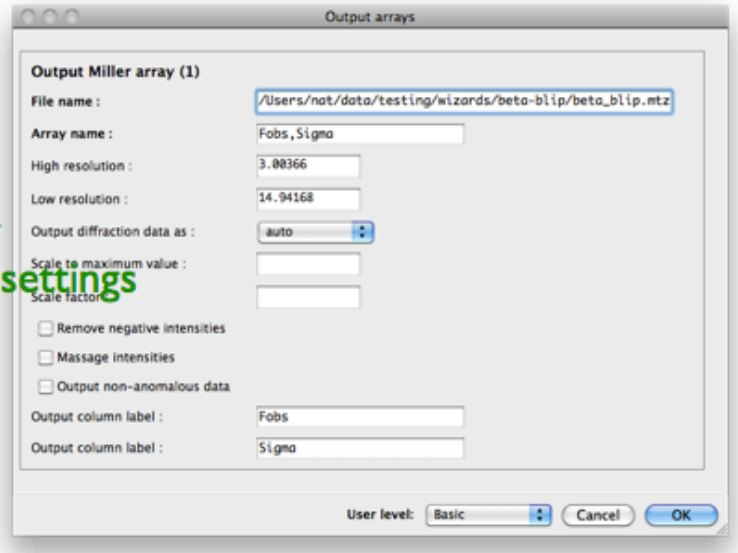
GUI examples



Reflection file editor



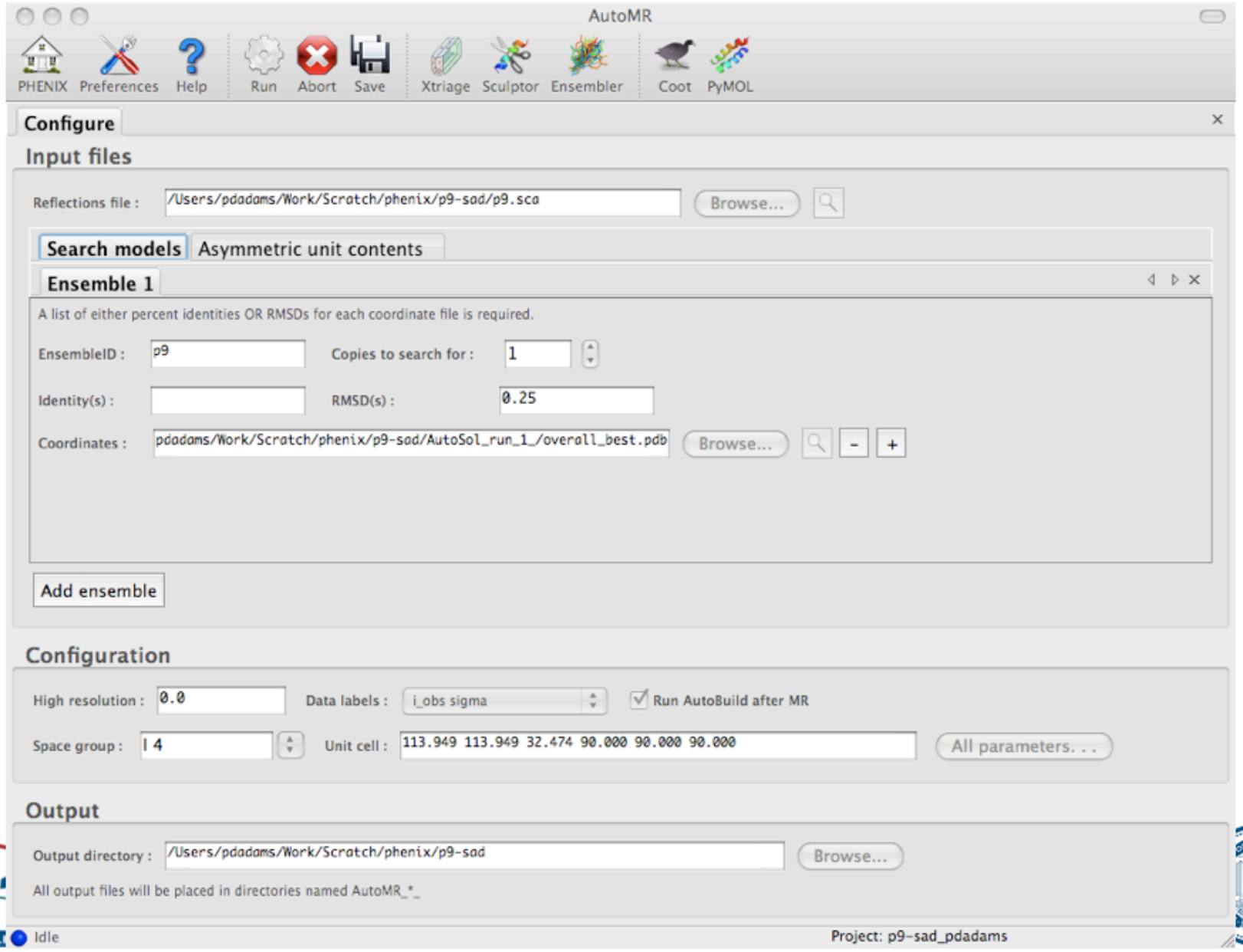
Combine and manipulate reflection files in any format, output as MTZ. Capable of extending old R-free sets, and generating new sets as thin shells (for refinement in presence of NCS). *For use with fully processed data only - reflections will be merged and h,k,l indices altered as required.*



advanced output settings

(All functionality is also available on the command line as iotbx.reflection_file_editor, but we recommend using the GUI for this unless you are scripting an automation pipeline.)

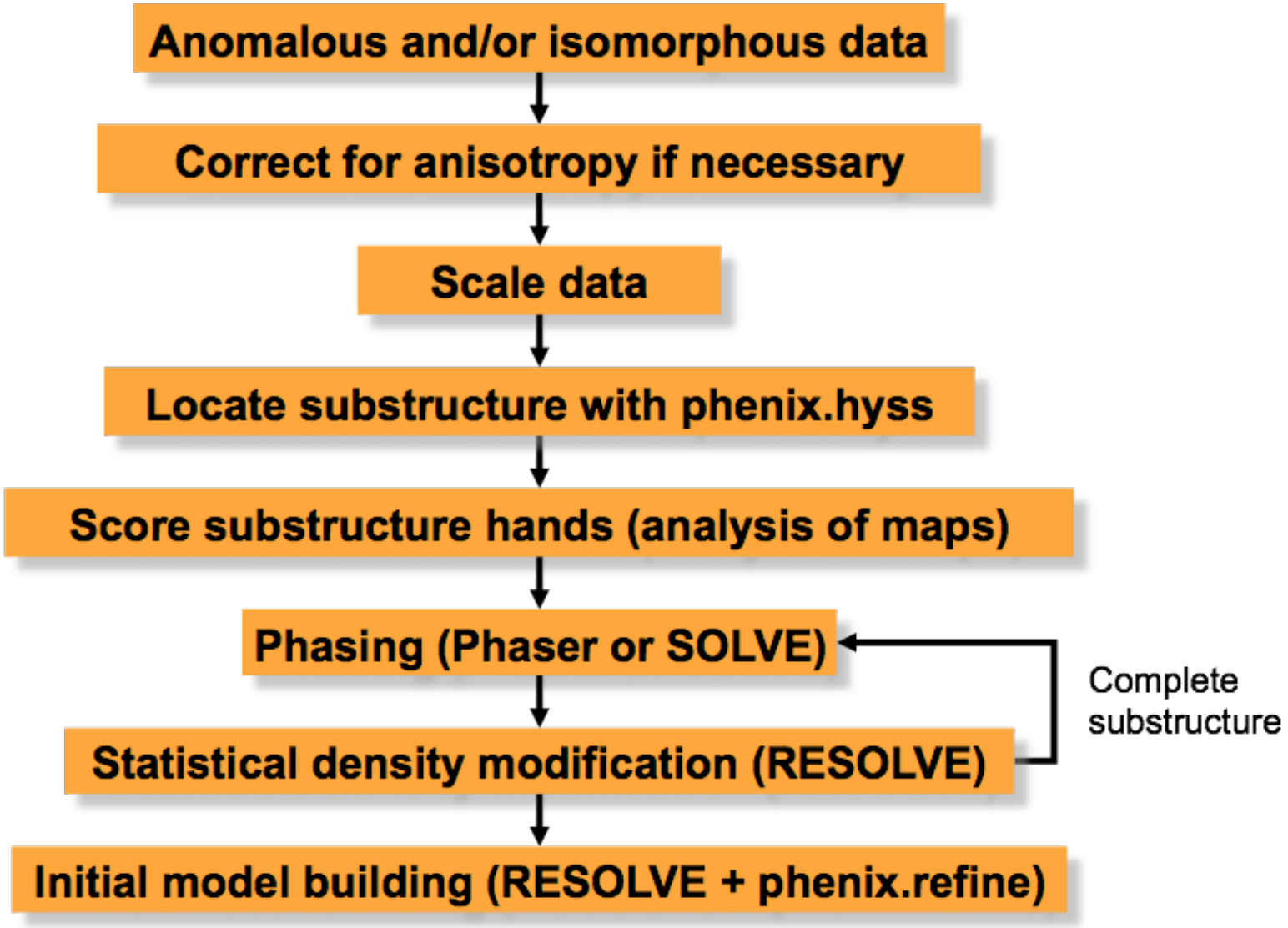
Automated Molecular Replacement - AutoMR



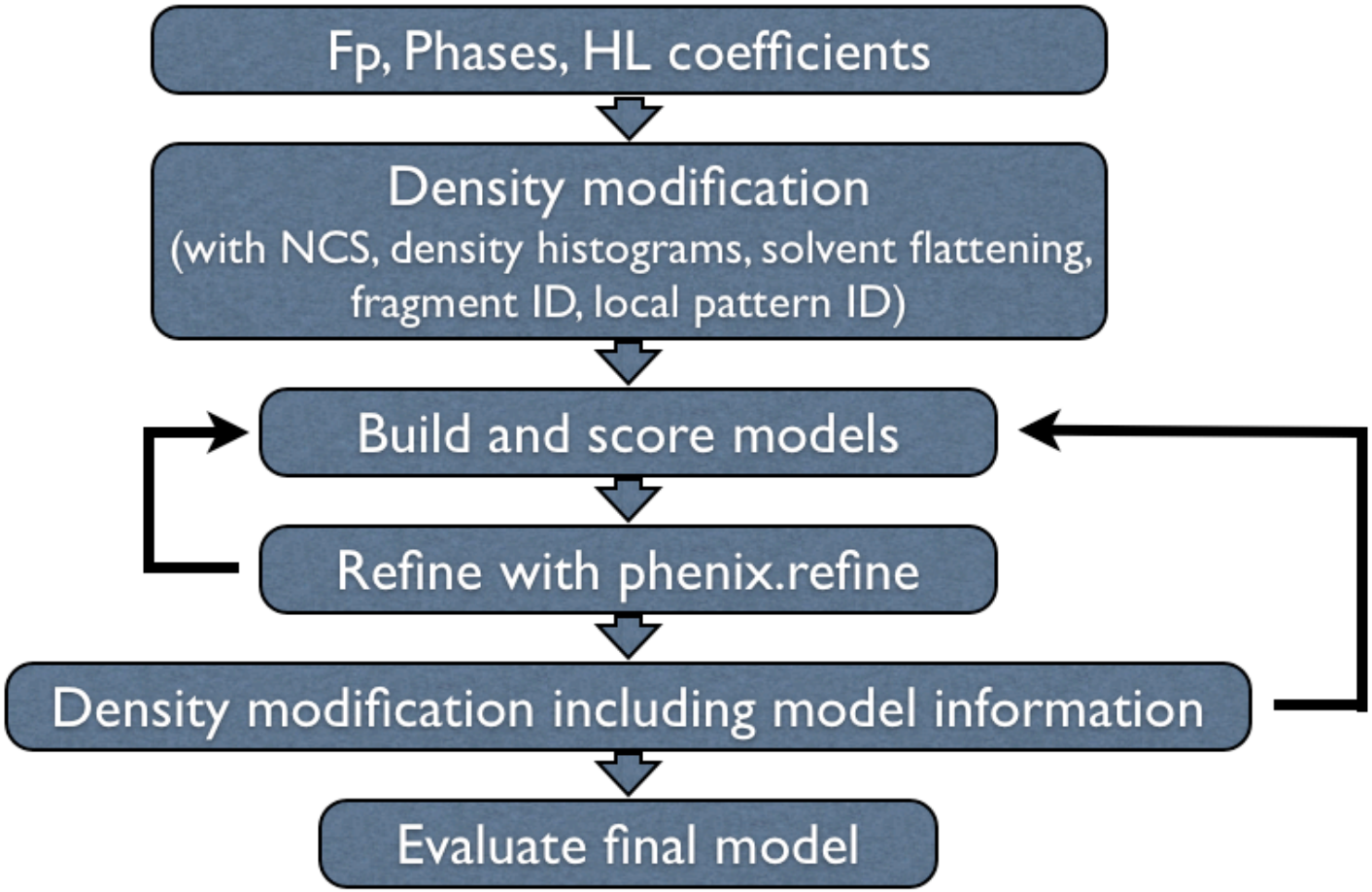
Automated structure solution

- The steps of substructure location, experimental phasing, density modification, and initial model building can be linked together and run automatically
- One of the biggest challenges is automatically deciding which results are best at each step and should be carried forward in the process
  - Choice between different substructure solutions
  - Choice of substructure hand
  - Has completing the substructure made the map better or worse?
  - Choice of phase improvement parameters

Automated structure solution: AutoSol Procedure



Automated Model Building/Rebuilding



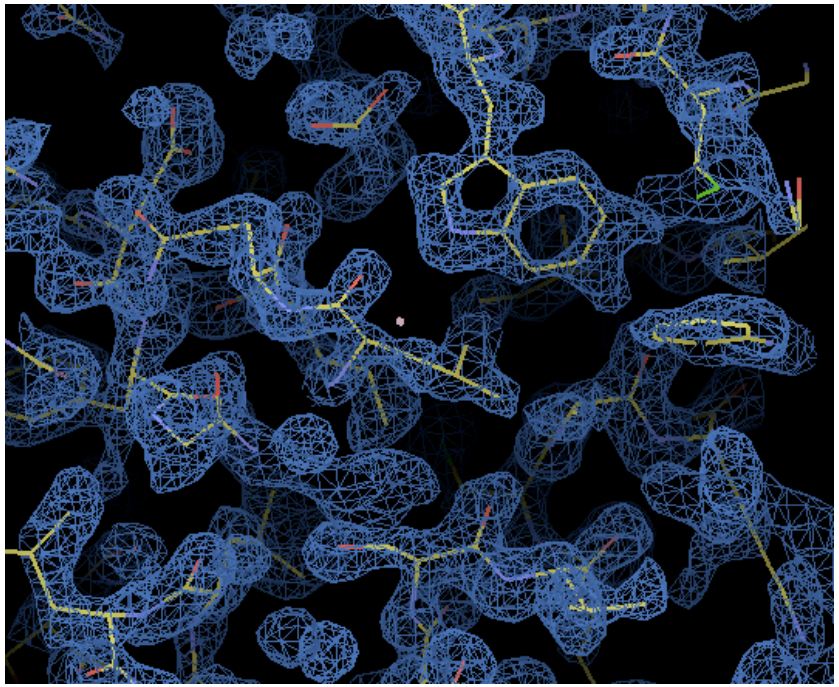
Acta Cryst. 2007, **D63**:597-610.
Acta Cryst. 2008, **D64**:61-69.
Acta Cryst. 2008, **D64**:515-524.

Tom Terwilliger, Pavel Afonine, Ralf Grosse-Kunstleve

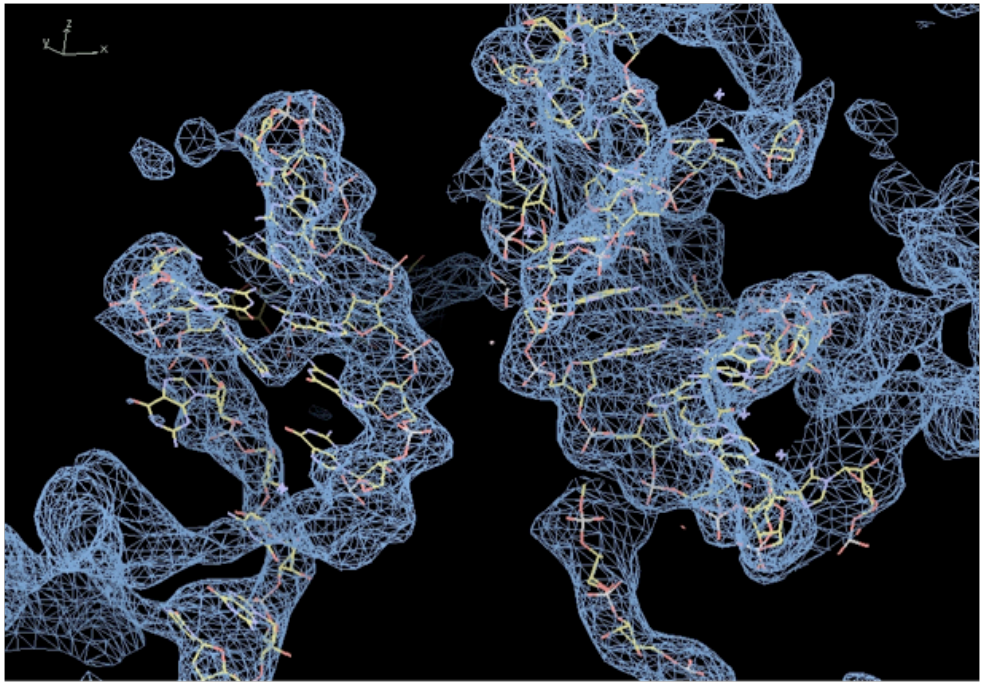
Phenix

Automated Model Building/Rebuilding

- Proteins and nucleic acids
- Low to high resolution (3.5Å or better)
- Location of secondary structure elements followed by extension, loop building and side chain docking



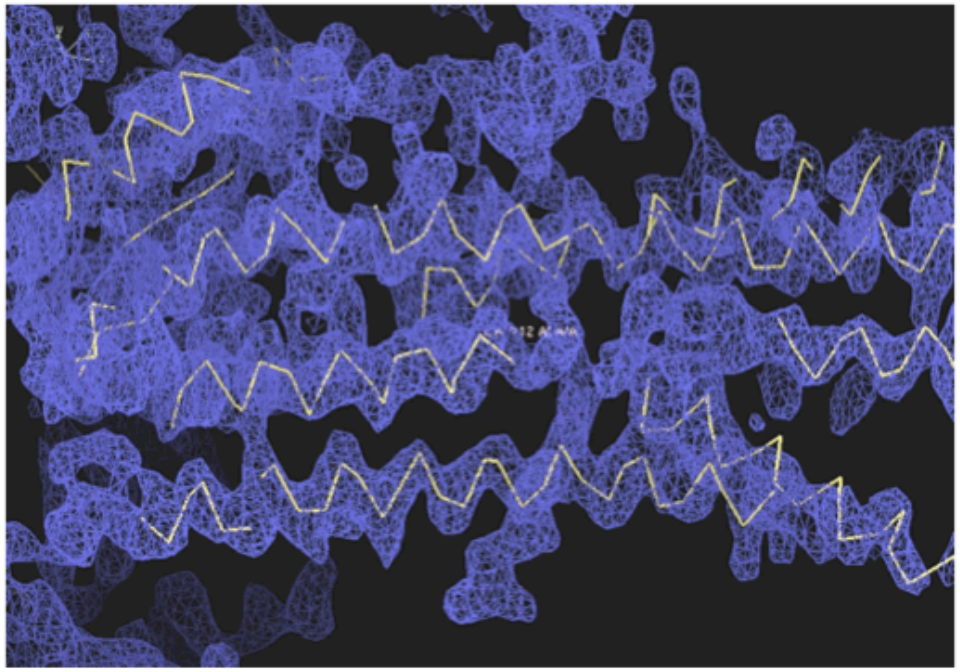
Gold Lysozyme at 1.5 Å



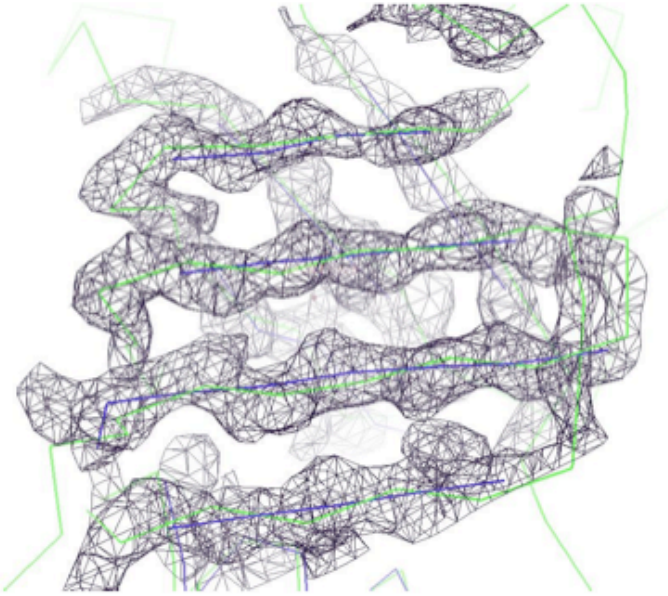
Group II intron at 3.5 Å. Data courtesy of J. Doudna

Rapid Secondary Structure Fitting

- phenix.find_helices_strands <map-file> <seq-file>
  - Rapid location of helices and strands - followed by sequence fitting
  - Fast - seconds to minutes depending on resolution and cell dimensions
  - Works at low resolution (tested at 4Å or better)



Ca^{2+} ATPase SAD map at 3.1 Å. Data courtesy of P. Nissen



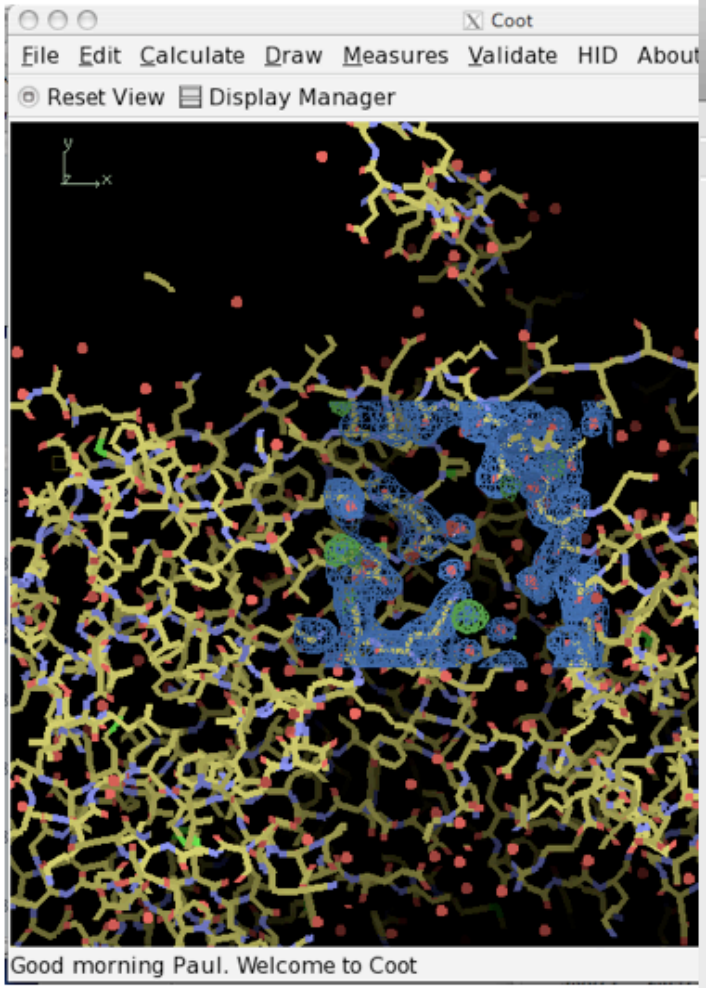
GroEL SIR averaged map at 2.7 Å

Tom Terwilliger, Los Alamos National Laboratory

Berkeley Lab

Phenix

Results - Rebuilding and Validation



phenix.refine

PHENIX Preferences Projects Help Save Graphics ReadySet NCS REEL

Configuration Run Refine_0

Results Validation

Validation results for /Users/pdadams/Work/Scratch/phenix-test/refine/bad/Refine_0/refine_001.pdb

MOLPROBITY

The validation performed by PHENIX is currently a subset of the full Molprobity analysis available on the web server. We recommend that academic groups use the server version to obtain more detailed information on structure quality, as well as files necessary to show validation results in Coot. You can start this process by clicking the Molprobity button on the left.

Ramachandran outliers:

Chain	Residue		Residue type	Score	Phi	Psi
A	LEU	45	General	0.02	68.0	106.1
A	SER	49	General	0.02	-79.0	-110.8
A	SER	76	General	0.02	-70.5	-97.4
A	LYS	78	General	0.04	-61.4	56.4
A	GLY	79	Glycine	0.04	-46.3	3.8
A	ASN	81	General	0.02	62.4	122.9
A	LYS	203	General	0.00	42.7	-77.8
A	GLU	204	General	0.00	134.3	152.6
A	LEU	207	Prepro	0.00	76.2	118.6
A	PHE	210	General	0.00	43.6	166.6
A	LEU	292	General	0.00	4.2	174.2
A	ARG	293	General	0.04	72.9	137.8
A	TYR	424	General	0.00	-68.5	-122.0
A	SER	425	General	0.02	-171.3	-47.3
A	THR	426	General	0.02	48.0	96.9

Rotamer outliers:

Chain	Residue		Score	Chi1	Chi2	Chi3	Chi4
A	ILE	6	0.39	163.3	97.6	-	-
A	THR	8	0.00	154.8	-	-	-

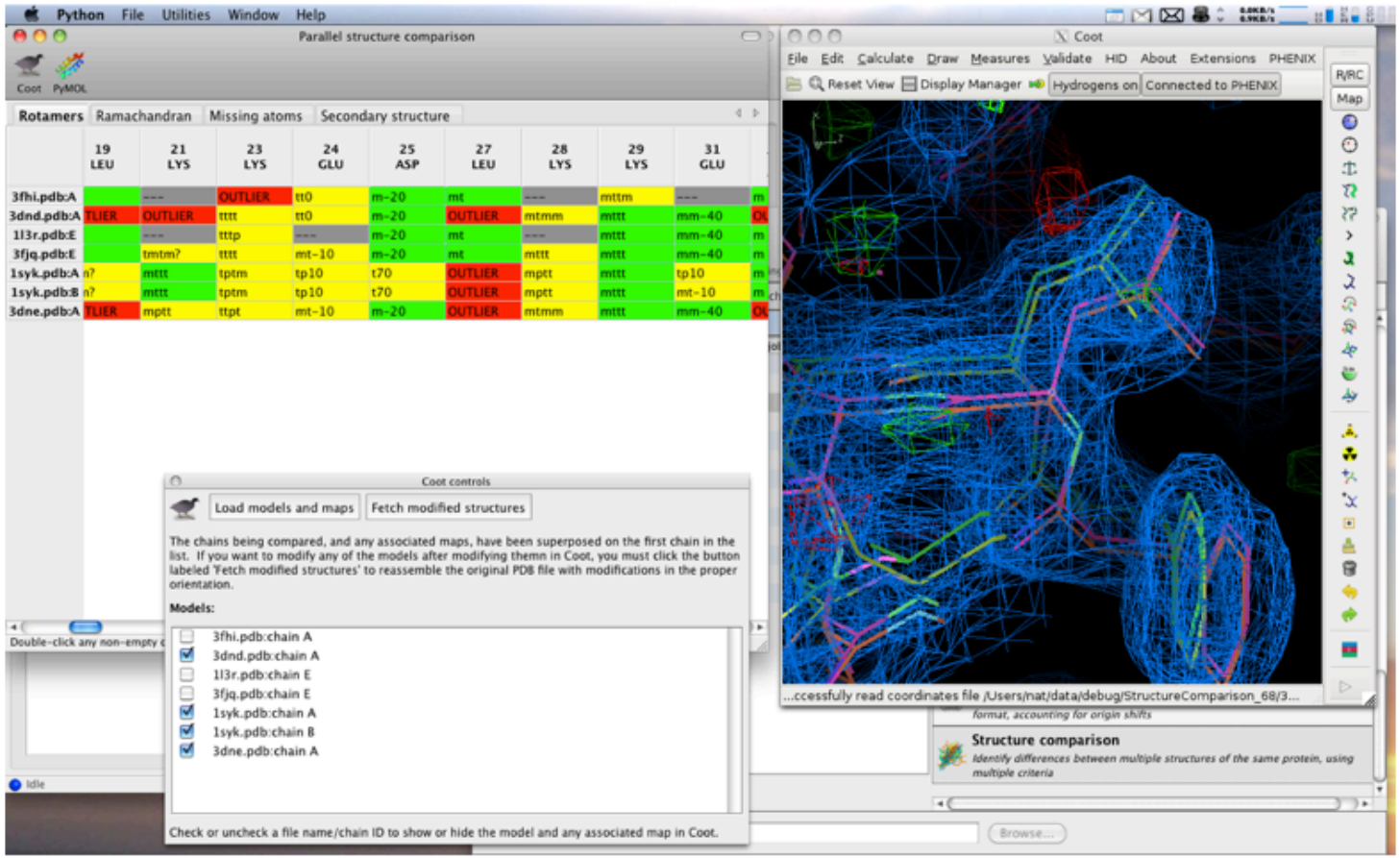
Refinement complete Project: None

BERKELEY LAB

Phenix

Parallel validation of multiple structures

- Identifies points of difference between structures of the same protein, with optional map superpositioning



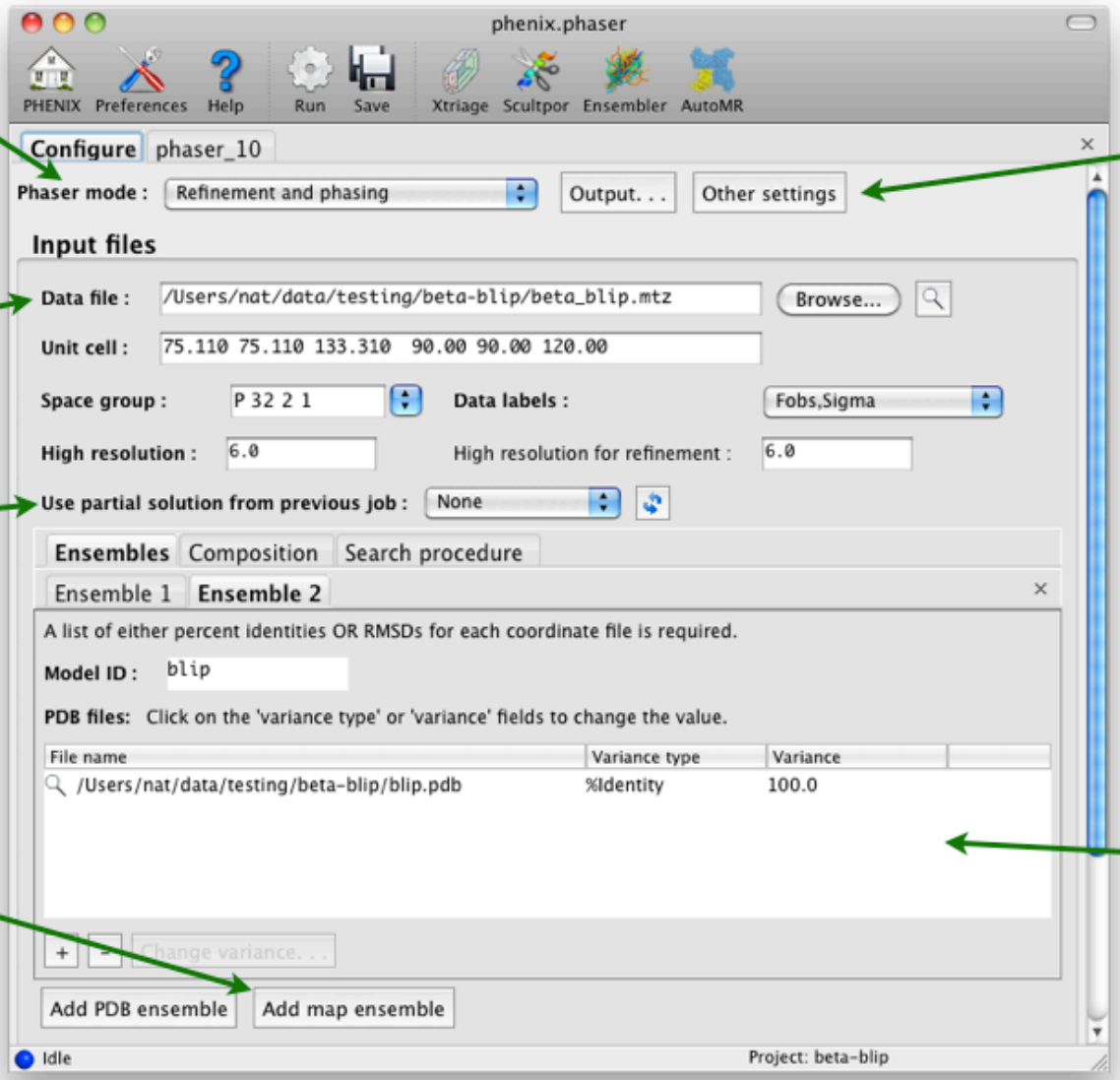
Phaser GUI

Supports modes MR_AUTO, MR_FRF, MR_BRF, MR_FTF, MR_BTF, MR_PAK, MR_RNP

Any reflection file format permitted

One-click re-use of partial solutions from past runs

Can use a low-resolution map as a search model

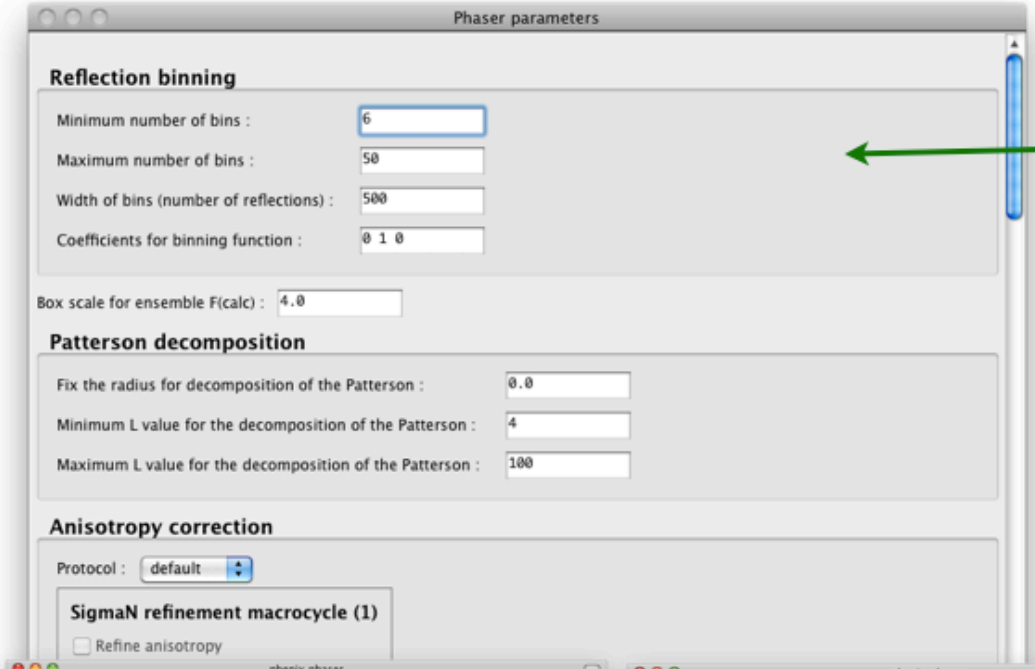


Most keywords found here

Drag-and-drop supported

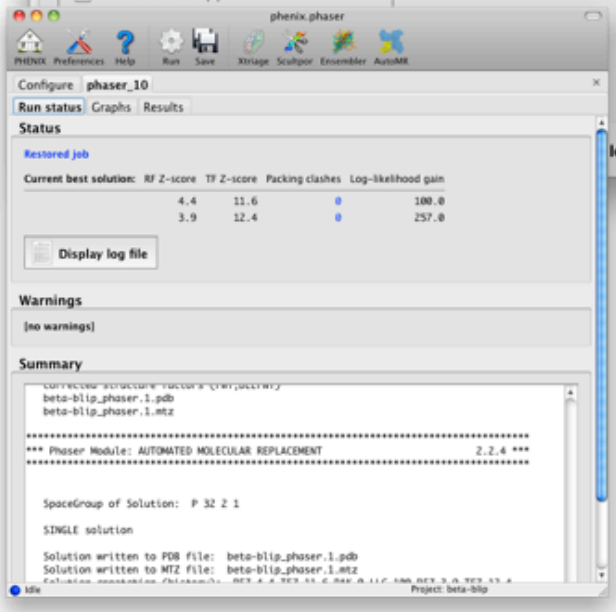
Nearly all of the features present in the CCP4i Phaser GUI should be available here, including the complete keywords. The known exceptions are a few of the less-used modes such as normal mode analysis; these may be added to a future version.

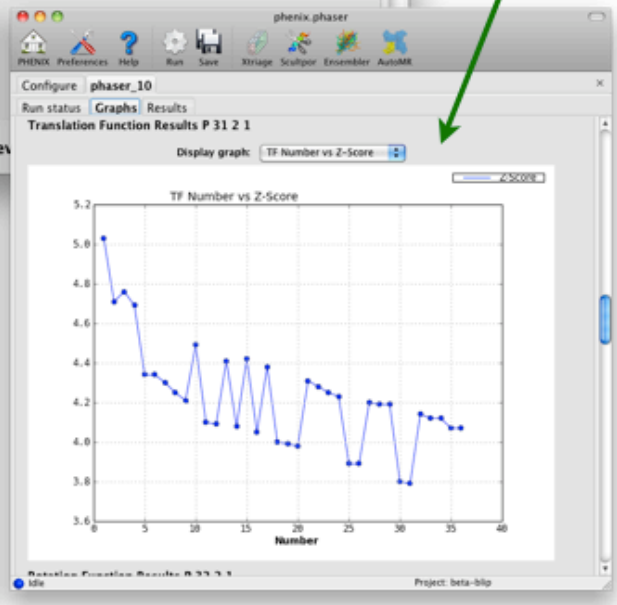
Phaser GUI

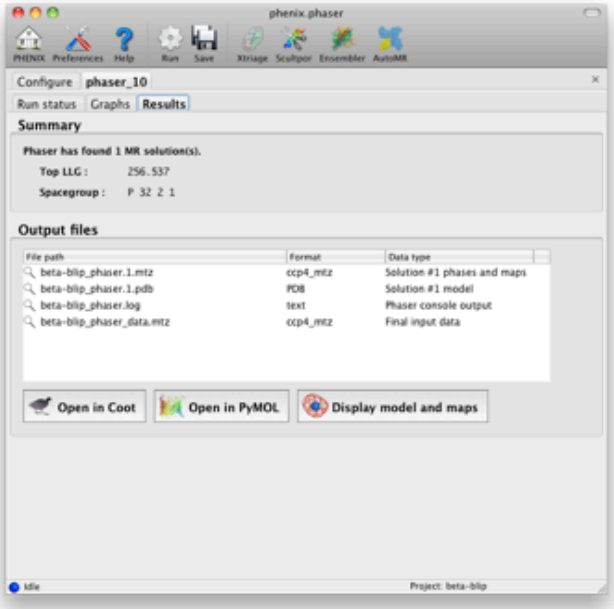


All "keywords" not shown on the main page are collected in a single dialog window, the content of which depends on the mode currently selected.

Built-in graphs.

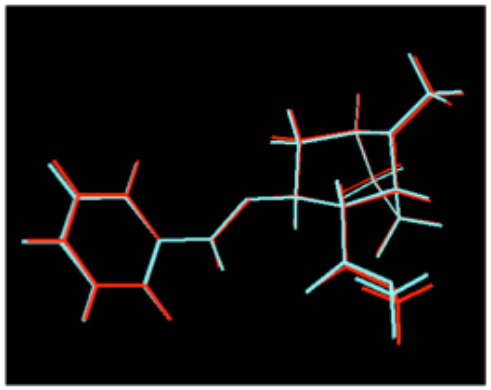


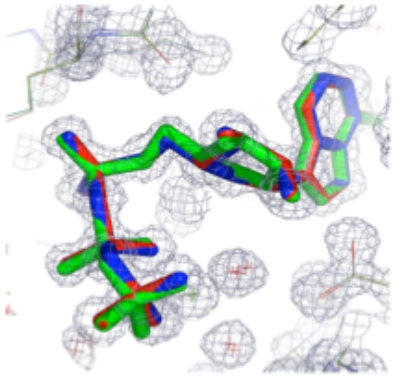


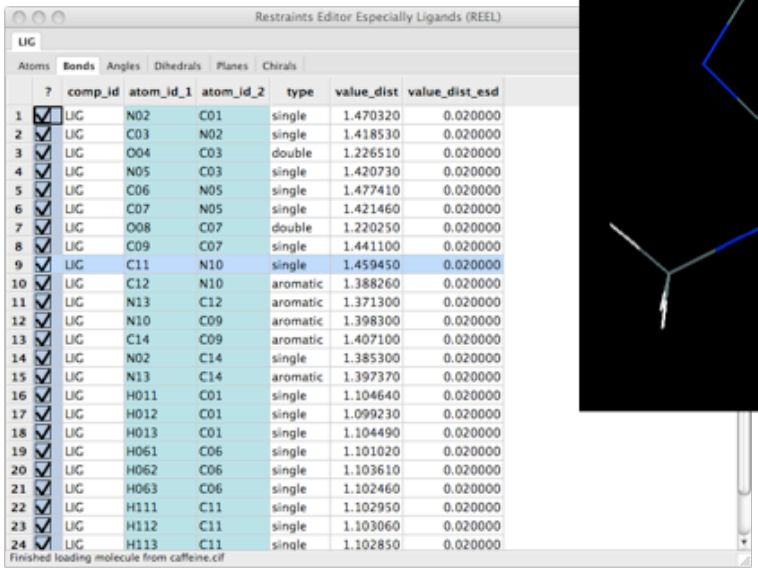


Ligands handling

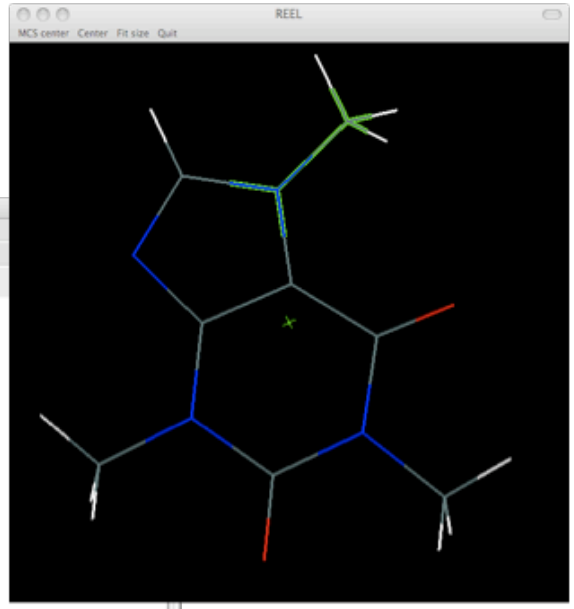
- Creation of ligand coordinates and restraints is typically a time consuming process
- phenix.elbow uses a semi-empirical method to generate atomic coordinates from a chemical topology, then calculates restraint values
- phenix.reel provides a graphical interface to manipulate restraints





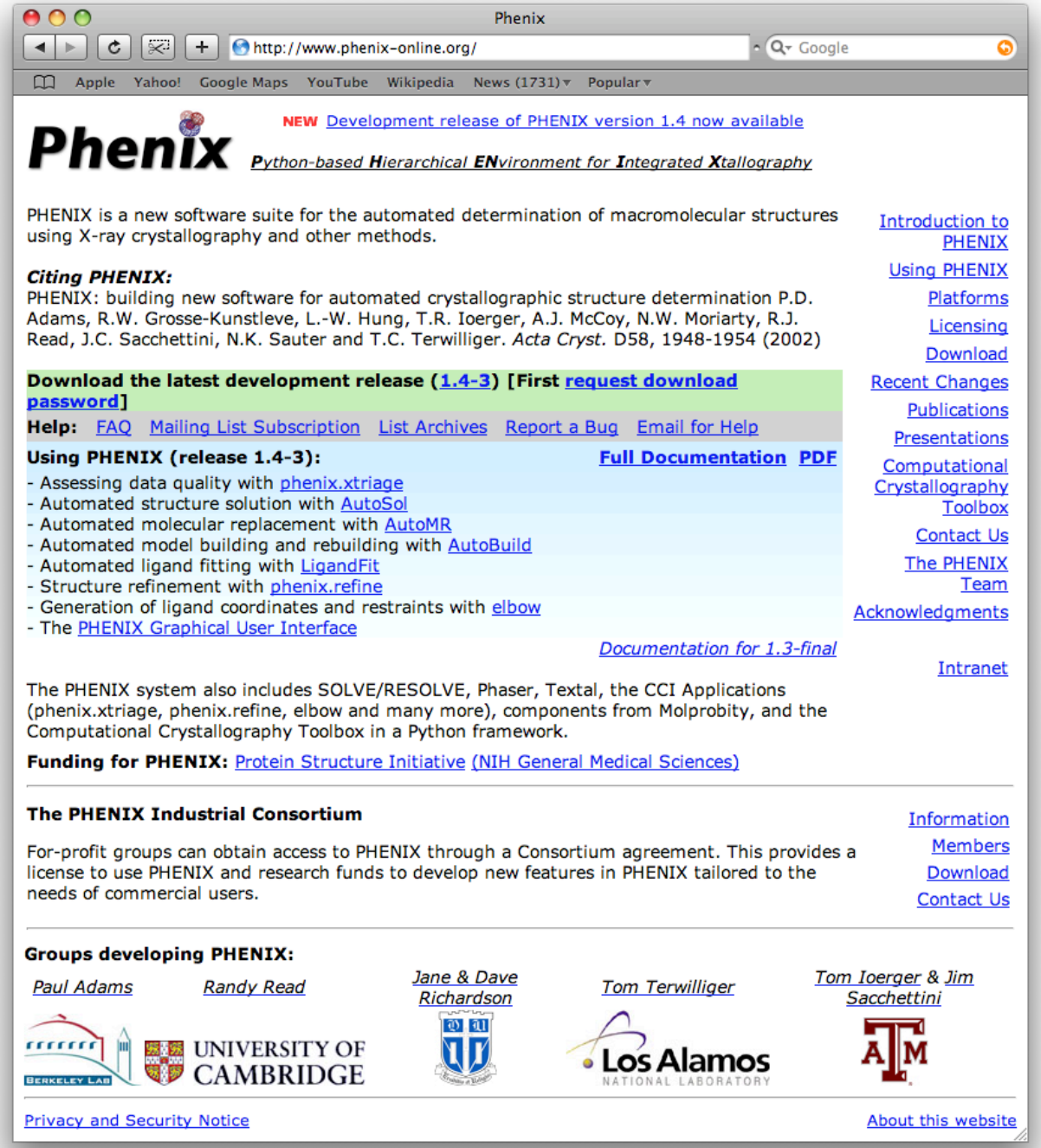


	?	comp_id	atom_id_1	atom_id_2	type	value_dist	value_dist_esd
1	☑	LIG	N02	C01	single	1.470320	0.020000
2	☑	LIG	C03	N02	single	1.418530	0.020000
3	☑	LIG	O04	C03	double	1.226510	0.020000
4	☑	LIG	N05	C03	single	1.420730	0.020000
5	☑	LIG	C06	N05	single	1.477410	0.020000
6	☑	LIG	C07	N05	single	1.421460	0.020000
7	☑	LIG	O08	C07	double	1.220250	0.020000
8	☑	LIG	C09	C07	single	1.441100	0.020000
9	☑	LIG	C11	N10	single	1.459450	0.020000
10	☑	LIG	C12	N10	aromatic	1.388260	0.020000
11	☑	LIG	N13	C12	aromatic	1.371300	0.020000
12	☑	LIG	N10	C09	aromatic	1.398300	0.020000
13	☑	LIG	C14	C09	aromatic	1.407100	0.020000
14	☑	LIG	N02	C14	single	1.385300	0.020000
15	☑	LIG	N13	C14	aromatic	1.397370	0.020000
16	☑	LIG	H011	C01	single	1.104640	0.020000
17	☑	LIG	H012	C01	single	1.099230	0.020000
18	☑	LIG	H013	C01	single	1.104490	0.020000
19	☑	LIG	H061	C06	single	1.101020	0.020000
20	☑	LIG	H062	C06	single	1.103610	0.020000
21	☑	LIG	H063	C06	single	1.102460	0.020000
22	☑	LIG	H111	C11	single	1.102950	0.020000
23	☑	LIG	H112	C11	single	1.103060	0.020000
24	☑	LIG	H113	C11	single	1.102850	0.020000



Challenges for Automation

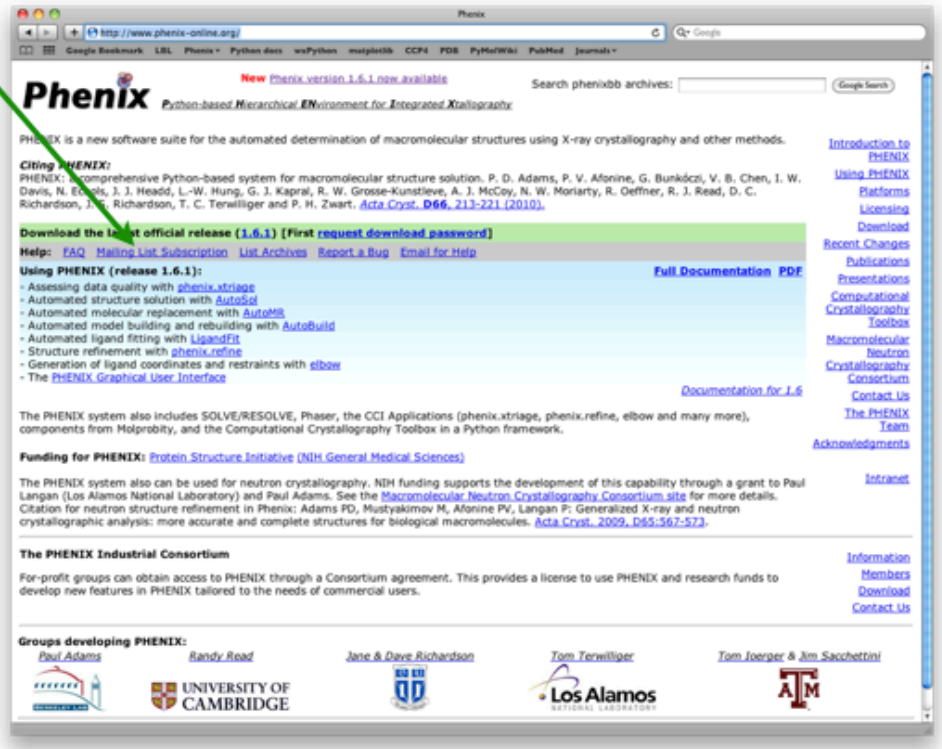
- Structure completion
  - Automated identification, fitting and refinement of ligands, metals, ions
  - Identification, fitting and refinement of discrete disorder (multiple conformations)
  - Representing other forms of disorder
  - Building poorly-defined regions
- Building complexes of protein and nucleic acid
- Identification of the best combination of datasets or fractions of datasets to use
- Handling different kinds of twinning and integrating it into the whole structure solution process
- Representations of uncertainty
- Automatic analysis of radiation damage
- Automated understanding of chemistry
- Optimal structure solution in the presence of twinning
- … and many more



www.phenix-online.org

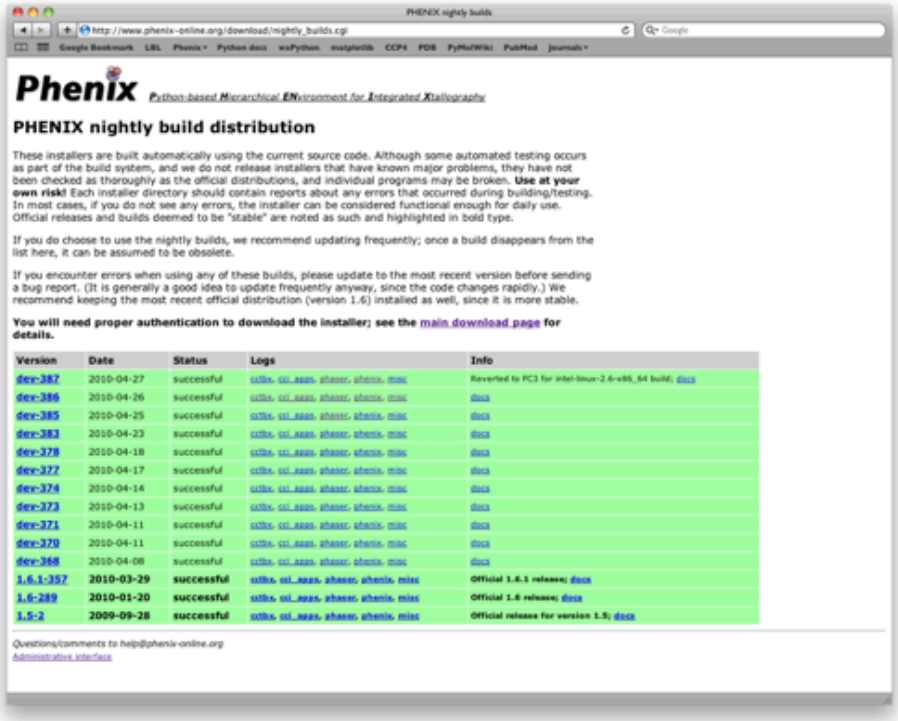
PHENIX resources online

- help@phenix-online.org: user support
- bugs@phenix-online.org: bug reports
- phenixbb@phenix-online.org: message board (subscribers only)



Obtaining PHENIX

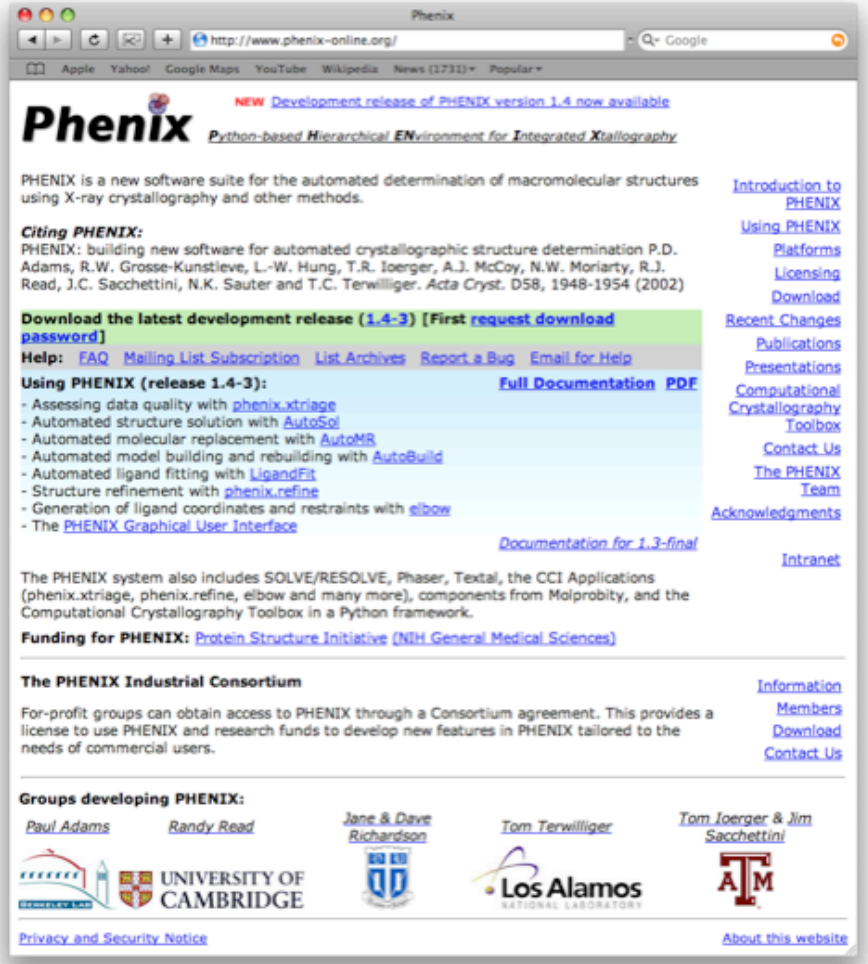
- Free to academic users; simple online registration required *(please use your academic email address!)*
- Regular official releases (typically 2-8 months)
- Nightly builds



http://www.phenix-online.org/download/nightly_builds.cgi

PHENIX Distribution

- Regular releases
- Supported on:
  - Linux (RedHat, Fedora)
  - Mac OSX
- Regular development releases:
  - Nightly builds
  - **http://www.phenix-online.org/**
- Extensive documentation



BERKELEY LAB

Phenix

Acknowledgments

- **Lawrence Berkeley Laboratory**
  - Pavel Afonine, Nat Echols, Jeff Headd, Ralf Grosse-Kunstleve, Nigel Moriarty, Nicholas Sauter, Peter Zwart
- **Los Alamos National Laboratory**
  - Tom Terwilliger, Li-Wei Hung
- **Cambridge University**
  - Randy Read, Airlie McCoy, Laurent Storoni, Gabor Bunkoczi, Robert Oeffner
- **Duke University**
  - Jane Richardson & David Richardson, Ian Davis, Vincent Chen, Jeff Headd, Chris Williams, Bryan Arendall, Laura Murray
- **Texas A&M University**
  - Tom Ioerger, Eric McKee, James Sacchettini

***Paul Adams* – project director**

- **Others**
  - Garib Murshudov & Alexi Vagin
  - Kevin Cowtan, Paul Emsley, Bernhard Lohkamp
  - CCP4 developers
  - Alexandre Urzhumtsev & Vladimir Lunin
  - David Abrahams
  - PHENIX Testers & Users: James Fraser, Herb Klei, Warren Delano, Paul Emsley, Bernhard Lokhamp, William Scott, Joel Bard, Bob Nolte, Frank von Delft, Scott Classen, Ben Eisenbraun, Phil Evans, Felix Frolow, Christine Gee, Miguel Ortiz-Lombardia, Blaine Mooers, Daniil Prigozhin, Miles Pufall, Edward Snell, Eugene Valkov, Erik Vogan, Andre White, and many more
- **Funding:**
- NIH/NIGMS:
  - *P01GM063210, P50GM062412, P01GM064692, R01GM071939*
- Lawrence Berkeley Laboratory
- PHENIX Industrial Consortium

BERKELEY LAB

Phenix

Reporting bugs, problems, asking questions

- **Something didn't work as expected?... program crashed?... missing feature?...**

  Not Good: silently give up and run away looking for alternative software.

  Good: report us a problem, ask a question, request a feature (explain why it's good to have), ask for help.

- **Reporting a bug:**

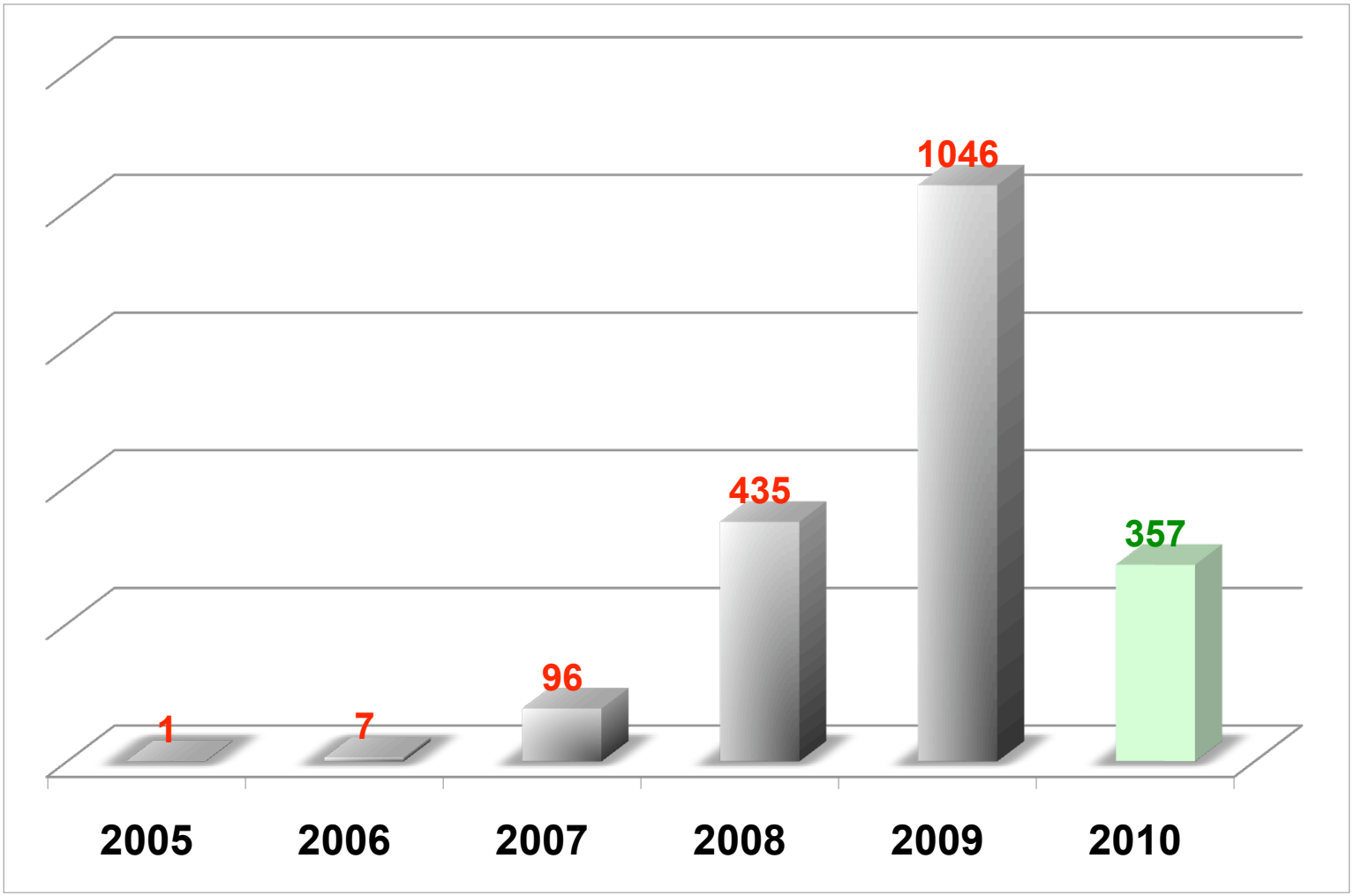
  Not good: "Hi! PHENIX crashed, I don't know what to do."

  Good: "Hi! PHENIX crashed. Here are:
  1) PHENIX version;
  2) Command and parameters I used;
  3) Input and output files (at least logs)."

Subscribe to PHENIX bulletin board: www.phenix-online.org

PHENIX use (July 12, 2010)

Number of structures in PDB with “REMARK 3 PROGRAM PHENIX”



User tested !