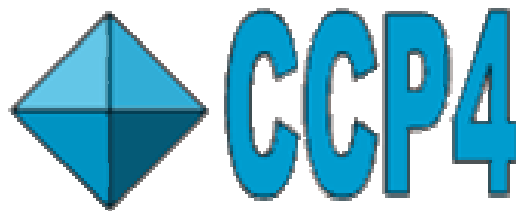




and



Making Deposition Easier

Shuchismita Dutta, Ph.D.

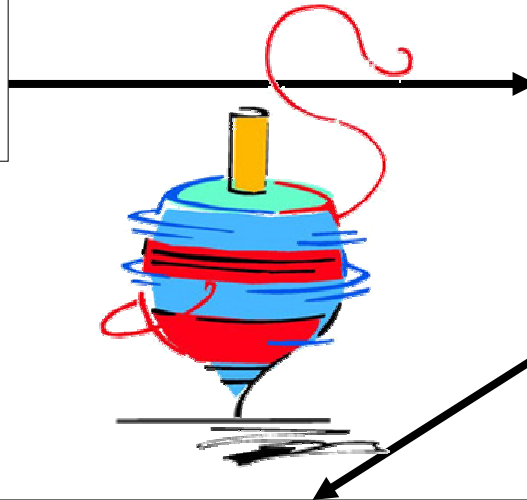
ACA 2004 Chicago

July 17th 2004

Motivation for this workshop:

Change your spin about structural data deposition

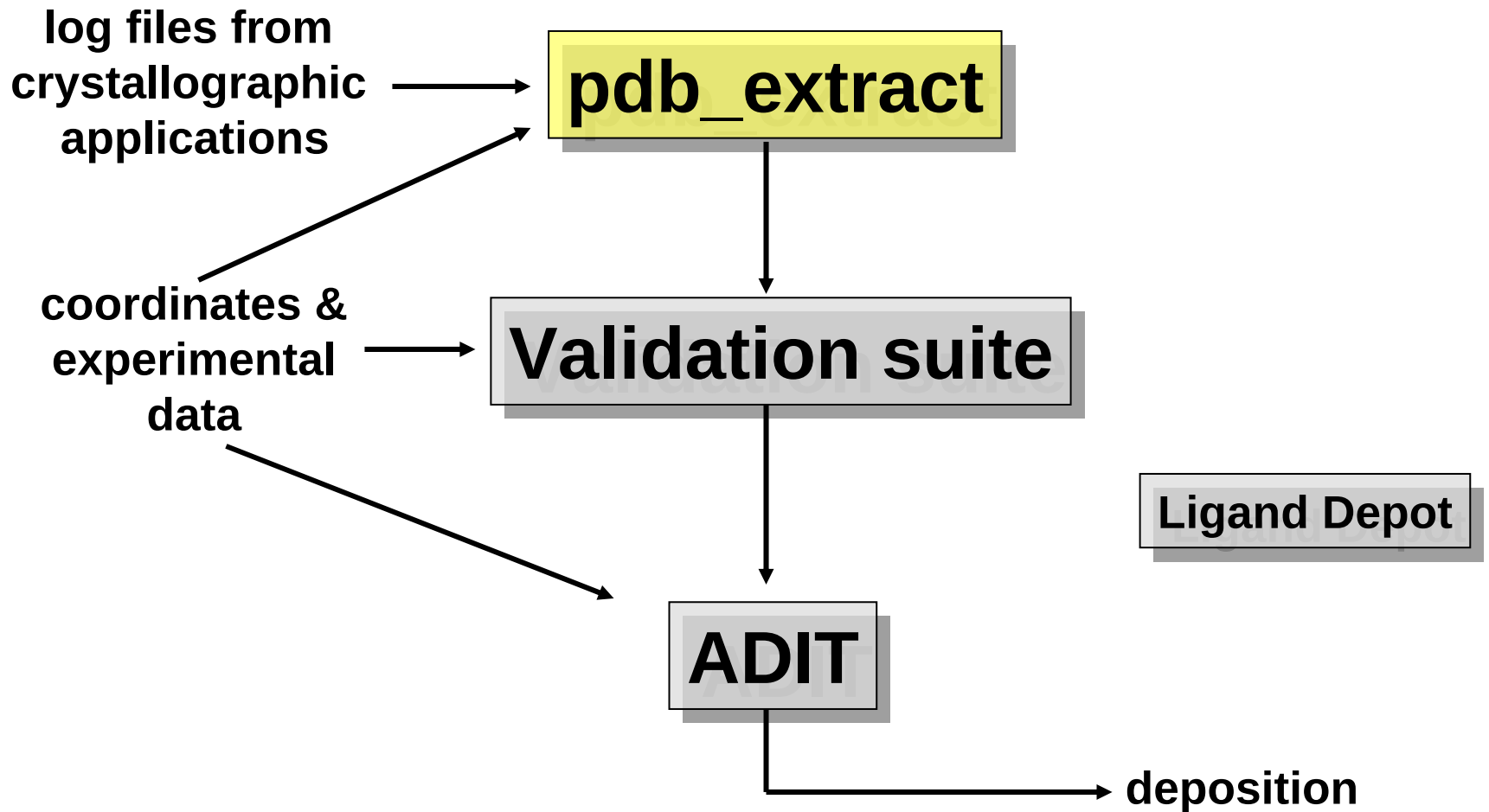
Data deposition is
a chore



Data deposition is
a chore no more

I can't wait to use the cool deposition tools
at the RCSB-PDB
to deposit some more (structures)

Overview of Data Deposition Tools



Structural data deposition today

The why, when, how, where and
what of deposition

Why do you deposit your structural data to the PDB

- **“Compulsory”** reasons
 - Primary citation journal policies requires it
 - Funding agency requires it
- **“Voluntary”** reasons
 - For safe-keeping of structural data
 - For the benefit of the entire scientific community

When do you deposit?

- **Immediately after structure determination**
- **Just prior to or after submission of manuscript**
- **After the manuscript has been accepted – urgent request for PDB ID**
- **Just before the researcher is leaving the lab**
- **Several years after the initial data collection**

How and Where do you deposit?

- Using the **ADIT** tool
 - <http://deposit.pdb.org/adit/> (RCSB-PDB) or
 - <http://pdbdep.protein.osaka-u.ac.jp/adit/> (PDBj).
- Using **AutoDep**
 - <http://autodep.ebi.ac.uk/> (MSD/EBI).

What do you deposit?

- The **coordinates**
- The **structure factor file(s)**
- and **more** ...
 - Information that only you can provide
 - Information that you should complete and verify
 - about the molecule(s) or complex
 - about the crystallization and data collection
 - Information that can be extracted from log files of crystallographic applications.

Information - only you can provide

- **Contact information**: author names, e-mail, postal address, phone, fax, including PI
- **Release instructions**: for coordinates, structure factors & sequence(s)
- **Title** for the deposited structure
- **Related entries**: name of database, ID, description
- **Citation information**: authors, title, journal details if available

Information about the molecule(s) - complete and verify

- **Molecule Name**, ligand name if appropriate
- **Molecule details**: Fragment name, mutations, EC #
- **Sequence information**: sequence, chain identifiers, appropriate database references
- **Source information**: genetically manipulated, natural or synthetic
- **Keywords**: To describe and search for the structure
- **Biological assembly** description

Information about crystallization and data collection - complete and verify

- **Crystallization details**: method, pH, temperature, crystallization solution components, solvent content, Matthews coefficient
- **Crystal data**: cell dimensions and space group
- **Data collection information**: number of crystals, type of diffraction experiment, radiation source, wavelength(s) used, detector type, data collection date, collection temperature

Information - extract from log files

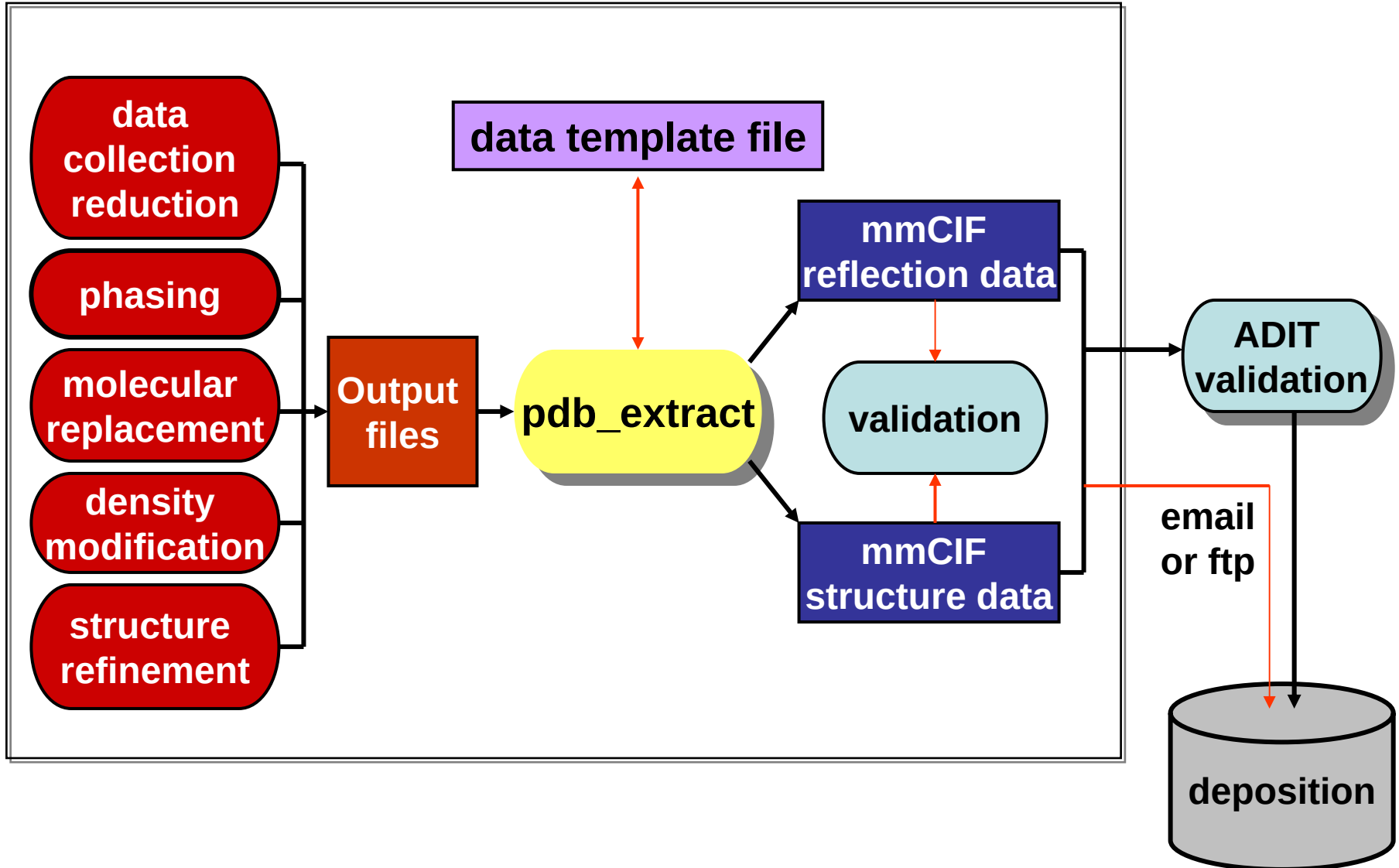
- **Data collection information**: resolution limits, observed criterion for sigma (F) or sigma (I), number of unique reflections (all and observed), percentage of possible reflections observed, R-merge I or R-sym I, details about the highest resolution shell
- **Refinement statistics**: resolution limits for refinement, cut-off on sigma(F), number of unique reflections (all and observed) used in refinement, R-factors for all reflections, R-factor for observed reflections, R-factor for working set reflections, associated R-free for the cross-validation set, structure determination method, cross-validation reflection selection details, stereochemistry target values
- **Software used**: for data collection, data reduction, structure solution, and refinement

Structural data deposition in the future

pdb_extract:

**an automated data extraction tool
to prepare your structural data
for deposition.**

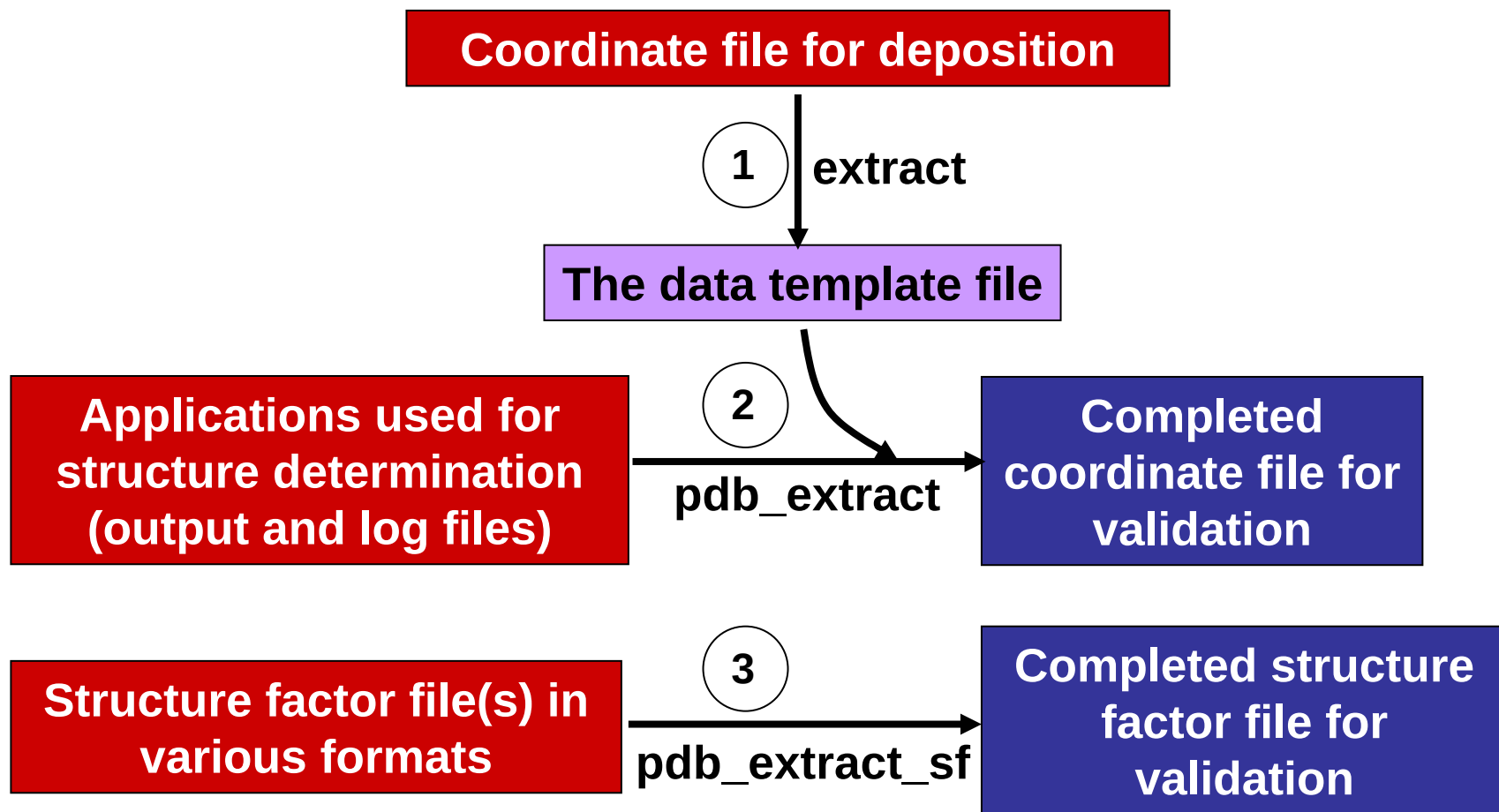
What does pdb_extract do?



Advantages of using pdb_extract

- **Automated data capture**
- **Creates more detailed deposition in files (phasing statistics)**
- **Output files can be directly validated and deposited**
- **Makes it easier for us to annotate**
- **Allows you to keep an electronic notebook for structures that are solved over a long period of time.**

Logic for running pdb_extract



File flavors



mmCIF

PDB

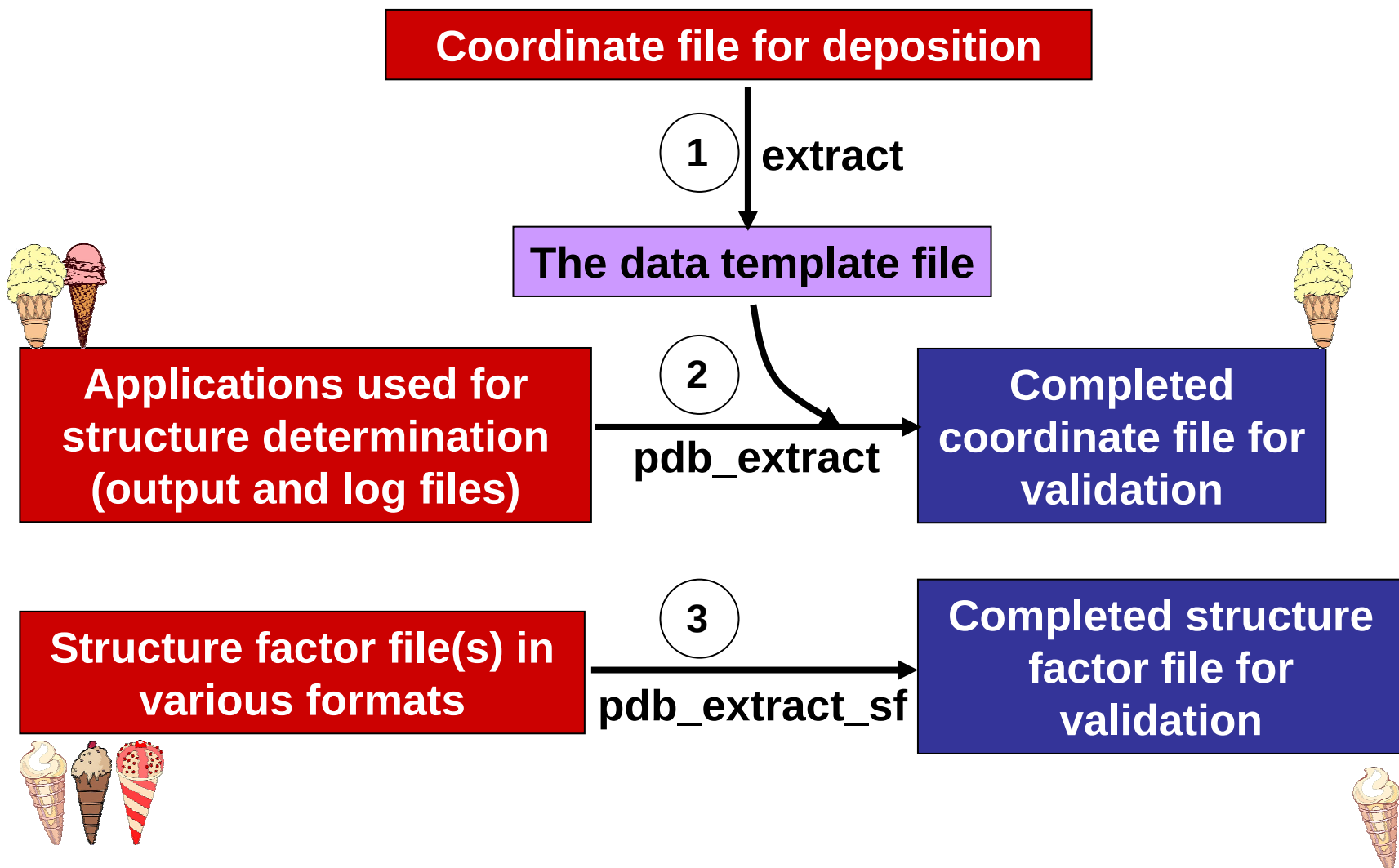
mmCIF
SF

ASCII
SF

mtz
SF

XML

Logic for running pdb_extract



Getting the sequence right

in the data template file

- **Missing residues**: marked as question marks '????' in the one-letter-code sequence. Complete the sequence at all these locations
- **Missing side chains**: Correct the sequence of any residue modeled as Ala or Gly due to missing side chain density
- **Missing N- and/or C-termini**: complete the sequence of the termini (include the sequence of cloning artifacts, expression tags etc. if present)
- **Non-standard residues**: extracted according to their 3 letter code (e.g. (MSE))

Additional data

in the data template file

- **contact authors**
- **release status**
- **citation and author list**
- **molecule name and details**
- **source information**
- **keywords**
- **biological assembly**
- **crystallization and data collection details**

How to use pdb_extract?

1. The CCP4i interface (CCP4)

- Intuitive and easy interface

2. The command line interface (CCP4, pdb_extract)

- Flexible interface
- Need to use specific arguments

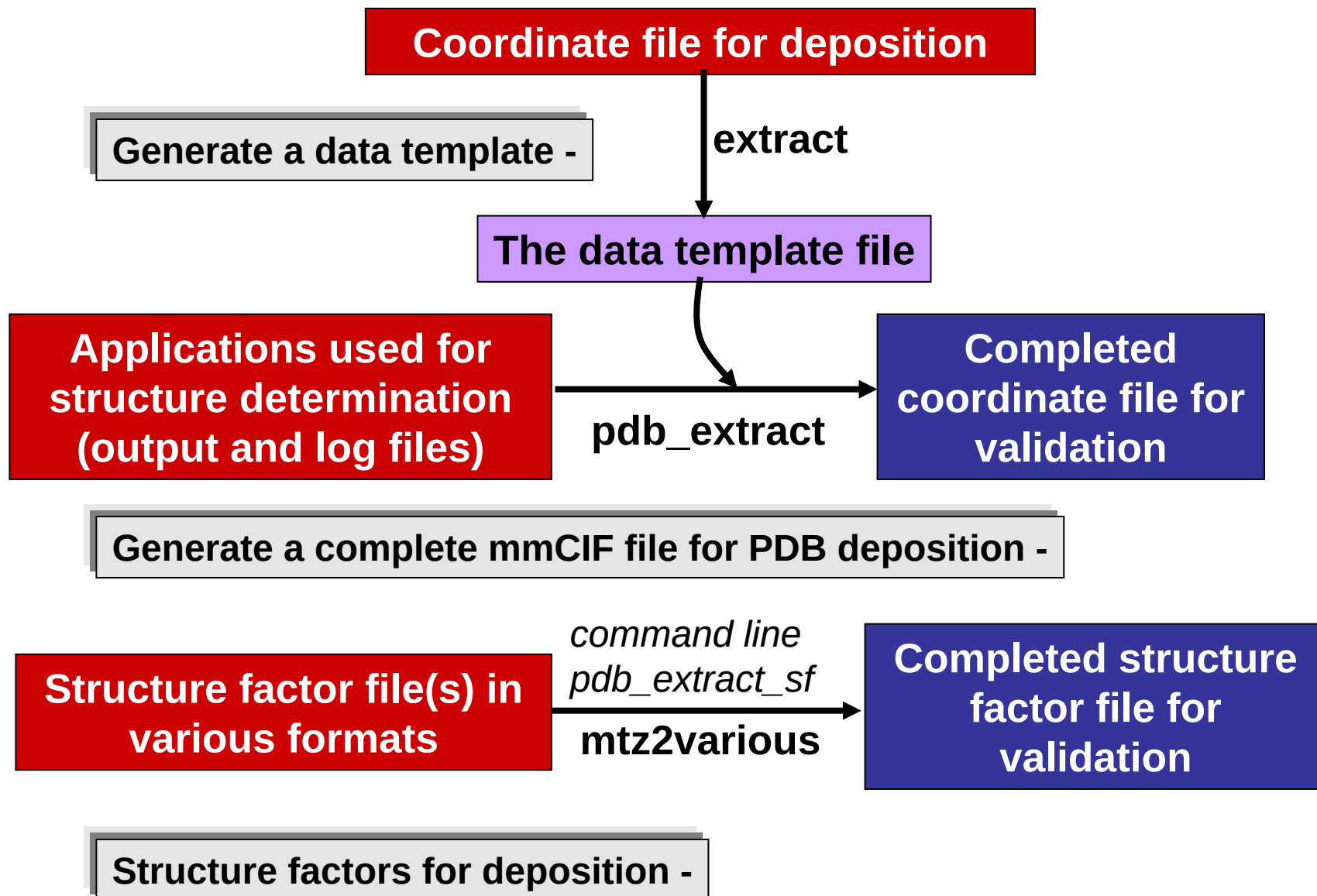
3. The script interface (CCP4, pdb_extract)

- User friendly interface
- Script input file

4. The Web interface (<http://pdb-extract.rutgers.edu/>)

- Can be run online from the RCSB-PDB

The CCP4i interface



Harvesting Manager

Help

Job title

Open File Browser to select Harvesting Files

List of harvest files selected

Programs

Run Program to Extract additional information for deposition

Extract information from data scaling, MR, phasing, DM, and refinement for PDB deposition

Extract information from Generate a data template step

Generate a data_template file from the coordinate file (PDB or mmCIF format)

PDB File test-data refmac.pdb Browse View

OR

mmCIF File test-data Browse View

Output File

File Name test-data data_template.text Browse View

Output

Run Save or Restore Close

data
scaling

phasing

density
modifi-
cation

Harvesting Manager

Help

Extract information from data scaling, MR, phasing, DM, and refinement for PDB deposition

Extract information from **Generate a complete mmCIF file for PDB deposition** step

Extract and merge all the information below to obtain a complete mmCIF file for PDB deposition

Extract information from output file of data scaling: Select Program **HKL2000**

LOG File **test-data** **for_refine.log** Browse View

LOG File **test-data** Browse View

mmCIF File **test-data** Browse View

Extract information from output file of phasing: Method **MAD** Select Program **SOLVE**

LOG File **test-data** **solve.prt** Browse View

LOG File **test-data** Browse View

PDB File **test-data** Browse View

mmCIF File **test-data** Browse View

Extract information from output file of molecular replacement: Select Program

LOG File **test-data** Browse View

LOG File **test-data** Browse View

mmCIF File **test-data** Browse View

Extract information from output file of density modification: Select Program **RESOLVE**

LOG File **test-data** **resolve.log** Browse View

mmCIF File **test-data** Browse View

Run Save or Restore Close

Harvesting Manager

Help

PDB File test-data Browse View

mmCIF File test-data Browse View

Extract information from output file of molecular replacement: Select Program

LOG File test-data Browse View

LOG File test-data Browse View

mmCIF File test-data Browse View

Extract information from output file of density modification: Select Program RESOLVE

LOG File test-data resolve.log Browse View

mmCIF File test-data Browse View

Extract information from output file of structure refinement: Select Program REFMAC5

PDB File test-data refmac.pdb Browse View

mmCIF File test-data peak.refmac Browse View

LOG File test-data Browse View

Extract information from data_template file (generated from the previous step)

Data_template File Name test-data data_template.txt Browse View

Output File

Output mmCIF File test-data output.mmci Browse View

Output

Run Save or Restore Close

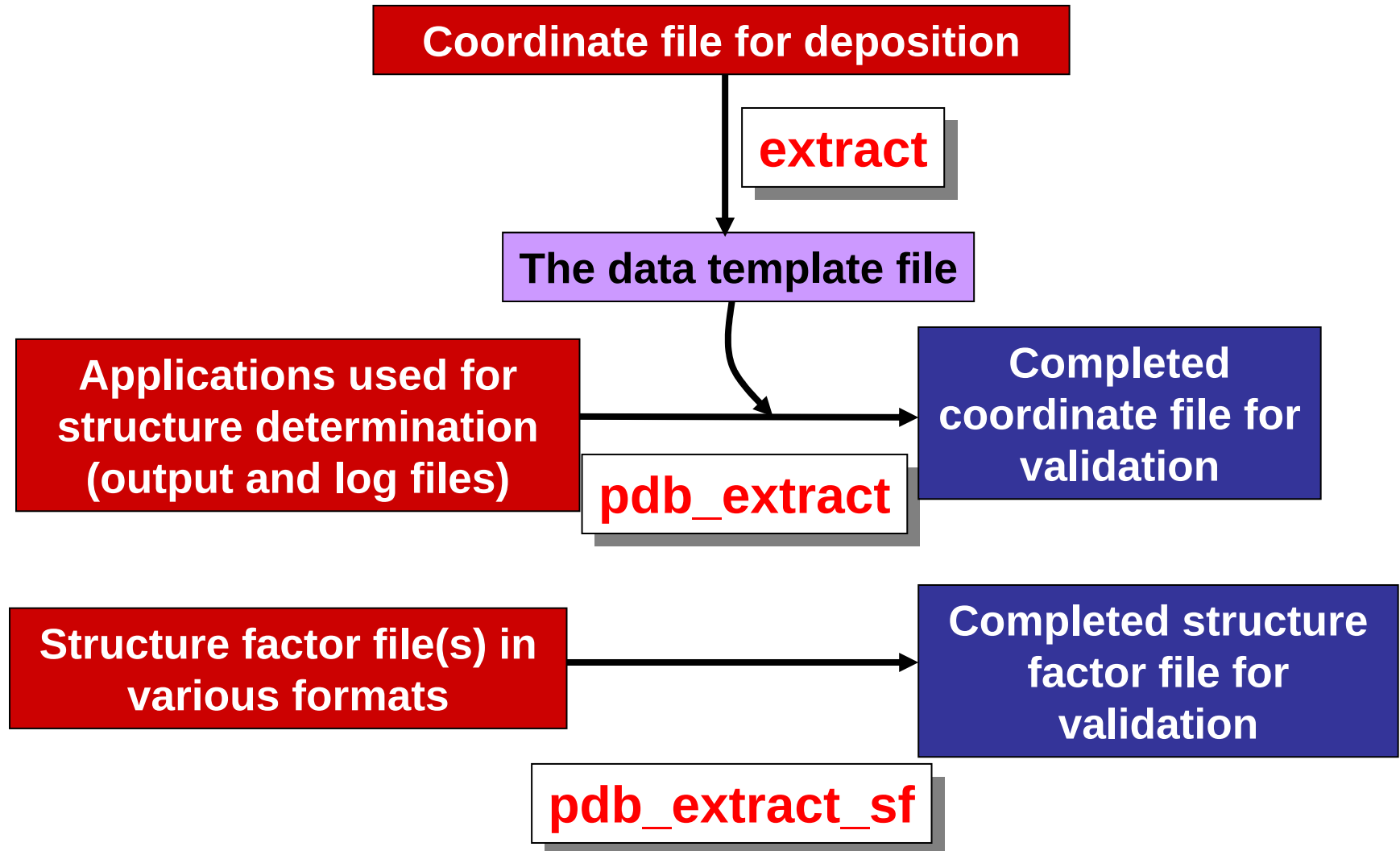
density
modifi-
cation

refine-
ment

Data
template



The command line interface



extract -pdb coordinate_PDB_file_name
or

extract -cif coordinate_CIF_file_name

pdb_extract -e MAD \

-p SOLVE -iLOG solve.prt \

-d RESOLVE -iLOG resolve.log \

-r refmac5 -icif peak.refmac -ipdb refmac.pdb\

-s HKL -iLOG scale-refine.log \

-sp HKL scale1.log scale2.log scale3.log \

-iENT date_template.text \

-o output.cif

pdb_extract_sf -rt F -rp refmac5 -idat refmac_sf.mmcif \ (for refinement)

-dt I -dp HKL \ (for phasing)

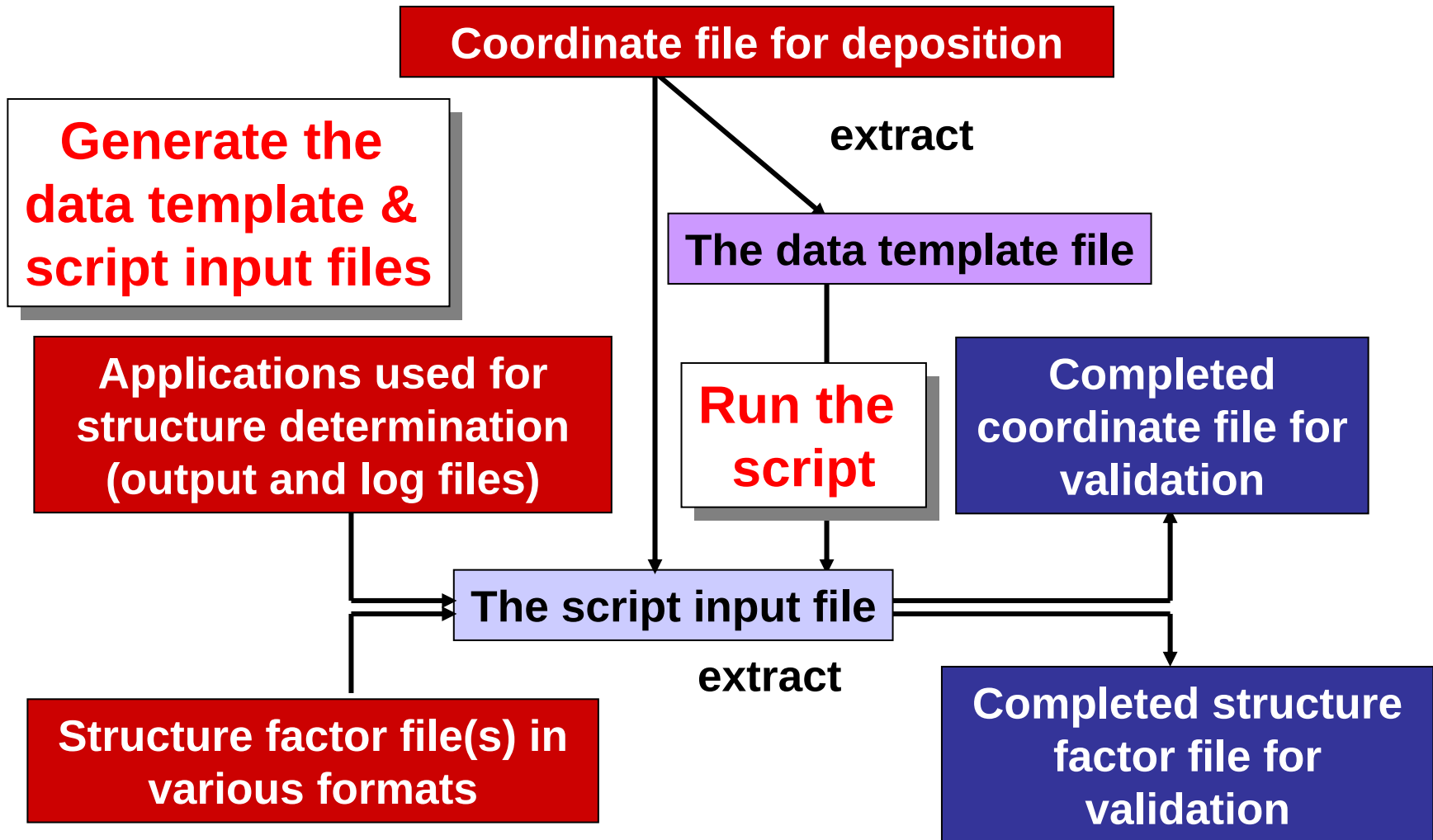
-c 1 -w 1 -idat scale1.sca \

-c 1 -w 2 -idat scale2.sca \

-c 1 -w 3 -idat scale3.sca \

-o output_sf.cif

The script interface



=====PART 1: Structure Factor for Final Refinement=====

Enter reflection data file used for final structure refinement

<reflection_data_type = "F" > (enter I (intensity) or F (amplitude))

<reflection_data_format = "CCP4" >

<reflection_data_file_name = " " >

=====PART 2: Structure Factors for Protein Phasing=====

Enter reflection data files used for heavy atom or MAD phasing

<scale_data_type = "I" > (enter I (intensity) or F (amplitude))

<scale_program_name = "HKL" >

For data set 1:

<crystal_number = "1" >

<diffract_number = "1" >

<scale_data_file_name_1 = " " >

<scale_log_file_name_1 = " " >

=====PART 4: Statistics for Molecular Replacement=====

Enter log files and software name for molecular replacement

<mr_software = "AMORE" >

<mr_log_file_LOG_1 = " " >

<mr_log_file_LOG_2 = " " >

The web interface (from RCSB-PDB)

Coordinate file for deposition

Upload the coordinate file

extract

Sequence of polymers in the structure

**Applications used for
structure determination
(output and log files)**

pdb_extract

**Coordinate file for
ADIT (editing &
validation)**

Press submit button

**Structure factor file(s) in
various formats**

pdb_extract_sf

**Completed structure
factor file for
validation**

Add additional details in ADIT

DATA EXTRACTION FOR STRUCTURE FACTORS

[details](#)

Structure Factors Used for Final Refinement		Help
REFMAC5	Intensity (I)	Help
Data file name		Browse... Help

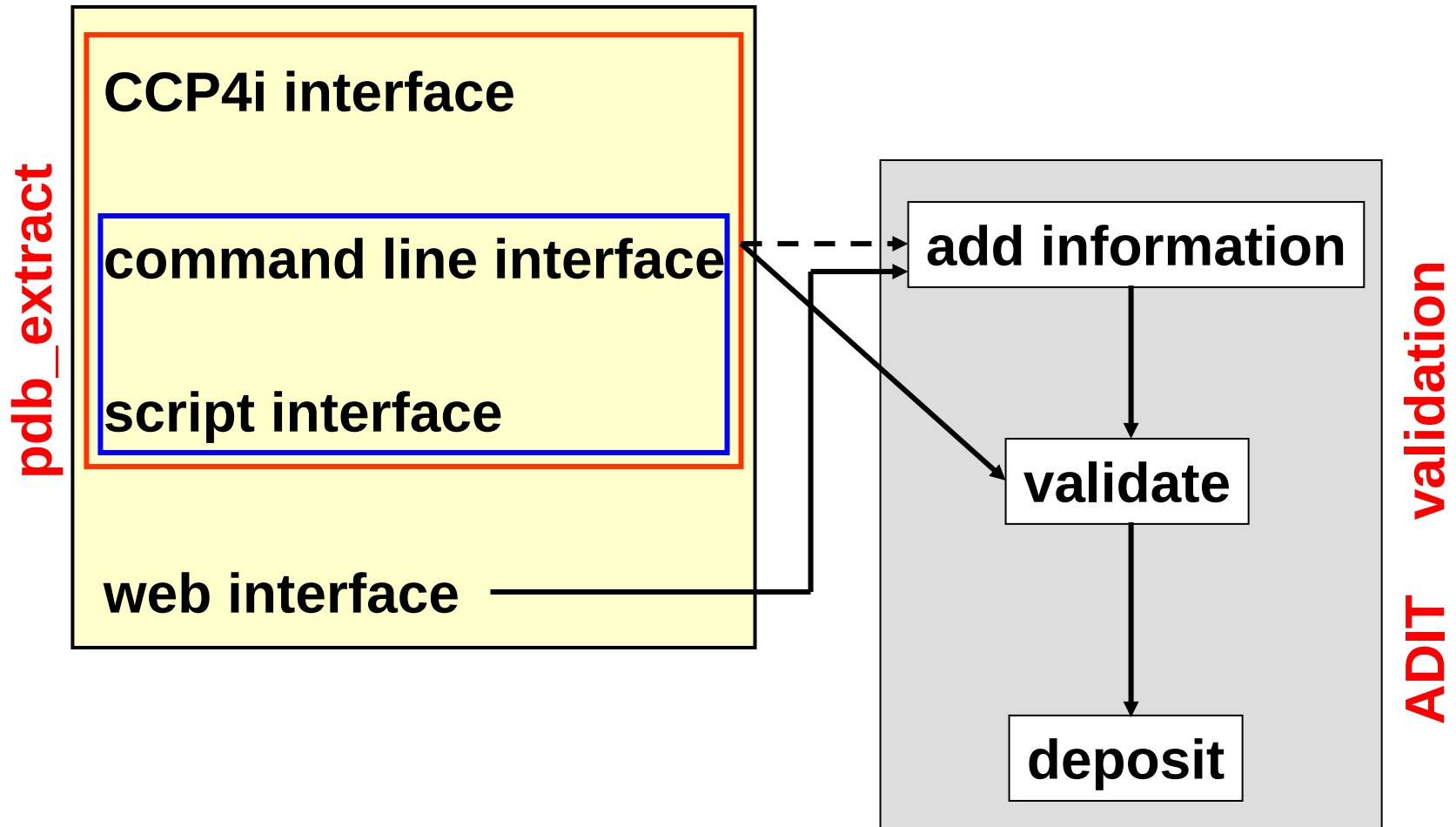
Structure Factors Used for Phasing		Help
Data Format	Data Type	Help
Crystal number		Diffraction Number
Data file name		Browse... Help
Log file name		Browse... Help

Molecular Entity

[details](#)

Entity identifier 1	
One-letter sequence	E????IRKLKNIYINGEWVESKTDQYEDVWNPATKEVLCQVPISTKEDIDYAA QTAAEAFKTSKVAVPRRARILFNFOQLLSQHKELAH????LITIENGKNT KEALGEVGRGIENVEFAAGAPSLMMGDSLASITDVEAANYRYPIGVVGGIA PFNFPMVPCWMPMAIALGNIFILKPSERTPLLTEKLVELFEKAGLPKGVF Help
Chain identifiers	A,B,C,D Help
Type of polymer	polypeptide(L) Help
Entity identifier	
One-letter sequence	Help
Chain identifiers	Help

Multiple paths to data deposition

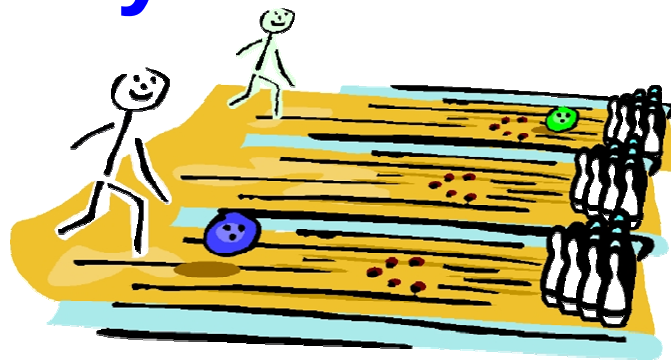


In summary

- Use **pdb_extract** to prepare your data
- **Validate** your files before deposition
- Use **ADIT** to deposit your files

Please Visit the RCSB PDB Booth #325 in “Data Alley”

- **Demonstrations**
 - `pdb_extract`
 - validation
 - ADIT
 - reengineered PDB site demos during coffee breaks
- **Questions answered**
- **Tattoos, posters and literature**



You can always write to us at
deposit@rcsb.rutgers.edu

All information is available from deposit.pdb.org

Acknowledgements

- The Protein Data Bank (PDB) is operated by
 - Rutgers, The State University of New Jersey
 - San Diego Supercomputer Center at the University of California, San Diego
 - Center for Advanced Research in Biotechnology/UMBI/NIST
- The RCSB PDB is supported by funds from
 - National Science Foundation (NSF)
 - National Institute of General Medical Sciences (NIGMS)
 - Office of Science, Department of Energy (DOE)
 - National Library of Medicine (NLM)
 - National Cancer Institute (NCI)
 - National Center for Research Resources (NCRR)
 - National Institute of Biomedical Imaging and Bioengineering (NIBIB)
 - National Institute of Neurological Disorders and Stroke (NINDS)
- The worldwide PDB (wwPDB) is a collaboration between
 - RCSB
 - MSD/EBI
 - PDBj



RCSB-PDB Data Deposition Services

- **pdb_extract**
 - Web- <http://pdb-extract.rutgers.edu/>
 - Standalone - http://deposit.pdb.org/mmcif/PDB_EXTRACT/index.html
- **Validation Server**
 - Web - <http://deposit.pdb.org/validate/>
 - Standalone - <http://deposit.pdb.org/mmcif/VAL/index.html>
- **ADIT**
 - Web – <http://deposit.pdb.org/adit/>
 - Standalone - <http://deposit.pdb.org/mmcif/ADIT/index.html>
- **Ligand Depot** - <http://ligand-depot.rutgers.edu/>
- **Overview and tutorials** for all RCSB-PDB data deposition services – <http://deposit.pdb.org>