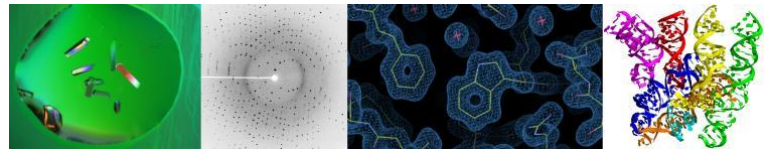
An Intelligent Data Analysis System for Biological Macromolecule Crystallography

Hao-Kai Sun

Computing Center, IHEP, CAS

May 09, 2023

Thanks to Yu Hu, Zhi Geng, Zengqiang Gao(IHEP-CAS), Wei Ding, Xin Zhang, Zengru Li(IOP-CAS)



Outline

1 Introduction

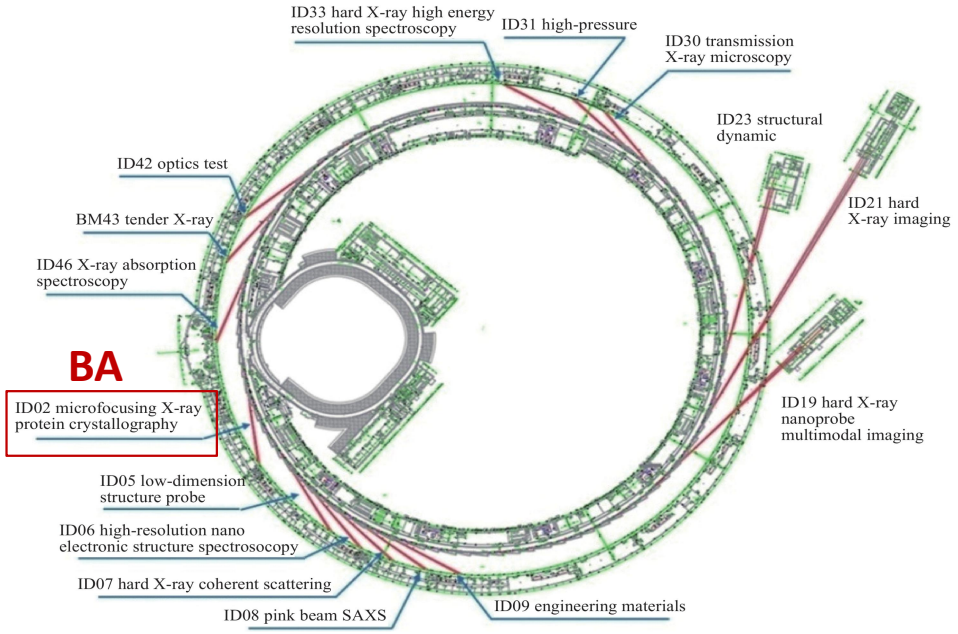
2 Design Goals & Project Architecture

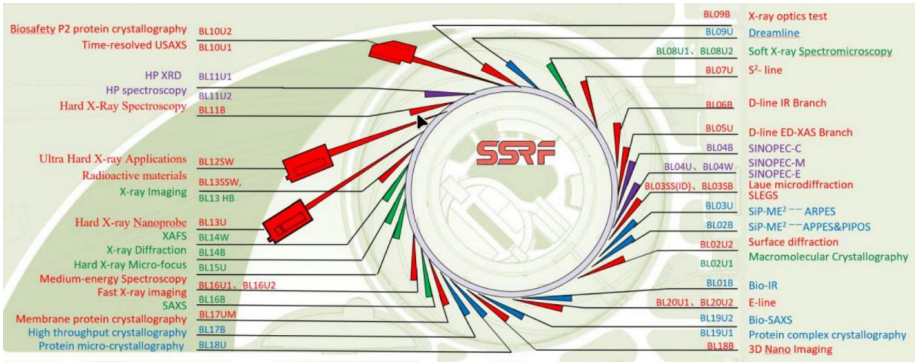
3 Current Project Status

4 Plans & Outlook

Introduction - Background

- **X-ray crystallography** : important technique for *revealing* the structures of biological macromolecules and *understanding* related biochemical processes.
- Among the 14 beamlines of the **High Energy Photon Source (HEPS) Phase I** under construction, the BA beamline is *specifically designed* for biological macromolecule crystallography (BMX).
- **Data Analysis** is the key *bridge* between "experiment" and "structure", especially for high-precision and high-throughput data generated by future advanced light sources.

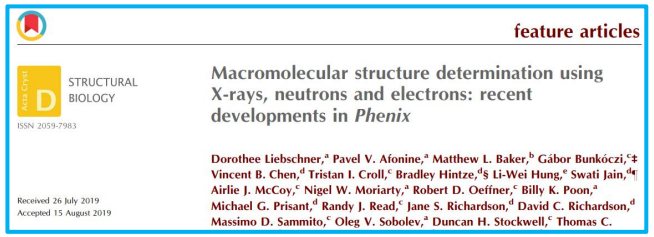




Introduction - Softwares

Given the many years of development in the methodology, *numerous* software packages are available for scaling, integration, indexing, model-building, and refinement, each excelling in its own area.

- XDS、DIALS、HKL-2000/3000、AutoPX、
- autoPROC、Phaser、Autobuild、
- Buccaneer、IPCAS、SHELX, *etc.*



feature articles

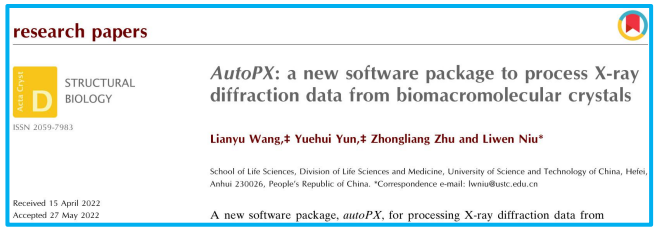
STRUCTURAL BIOLOGY

ISSN 2059-7983

Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*

Dorothee Liebschner,[a] Pavel V. Afonine,[a] Matthew L. Baker,[b] Gábor Bunkóczi,[c]‡ Vincent B. Chen,[d] Tristan I. Croll,[c] Bradley Hintze,[d]§ Li-Wei Hung,[c] Swati Jain,[d]¶ Airlie J. McCoy,[c] Nigel W. Moriarty,[a] Robert D. Oeffner,[c] Billy K. Poon,[a] Michael G. Prisant,[d] Randy J. Read,[c] Jane S. Richardson,[d] David C. Richardson,[d] Massimo D. Sammito,[c] Oleg V. Sobolev,[a] Duncan H. Stockwell,[c] Thomas C.

Received 26 July 2019
Accepted 15 August 2019



research papers

STRUCTURAL BIOLOGY

ISSN 2059-7983

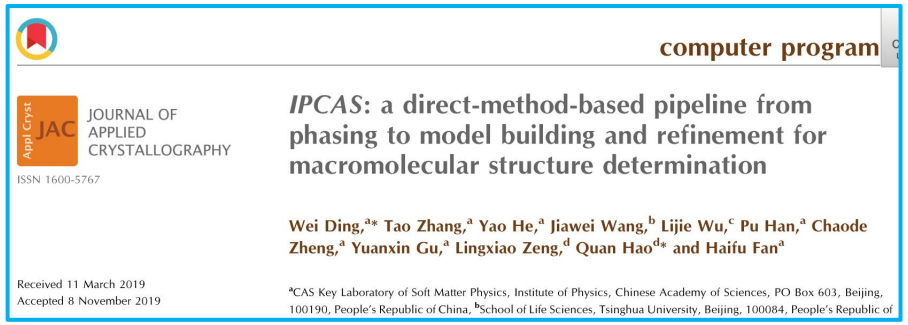
AutoPX: a new software package to process X-ray diffraction data from biomacromolecular crystals

Lianyu Wang,‡ Yuehui Yun,‡ Zhongliang Zhu and Liwen Niu*

School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China. *Correspondence e-mail: lwniu@ustc.edu.cn

Received 15 April 2022
Accepted 27 May 2022

A new software package, *autoPX*, for processing X-ray diffraction data from



computer program

JOURNAL OF APPLIED CRYSTALLOGRAPHY

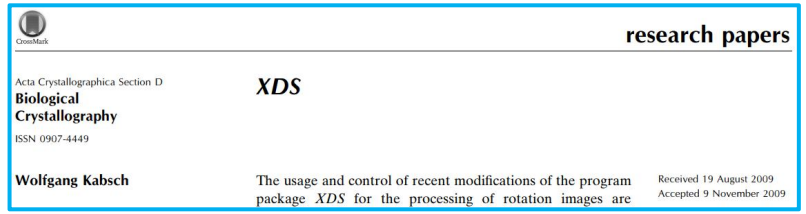
ISSN 1600-5767

IPCAS: a direct-method-based pipeline from phasing to model building and refinement for macromolecular structure determination

Wei Ding,[a]* Tao Zhang,[a] Yao He,[a] Jiawei Wang,[b] Lijie Wu,[c] Pu Han,[a] Chaode Zheng,[a] Yuanxin Gu,[a] Lingxiao Zeng,[d] Quan Hao[d]* and Haifu Fan[a]

Received 11 March 2019
Accepted 8 November 2019

[a]CAS Key Laboratory of Soft Matter Physics, Institute of Physics, Chinese Academy of Sciences, PO Box 603, Beijing, 100190, People's Republic of China, [b]School of Life Sciences, Tsinghua University, Beijing, 100084, People's Republic of



research papers

Acta Crystallographica Section D Biological Crystallography

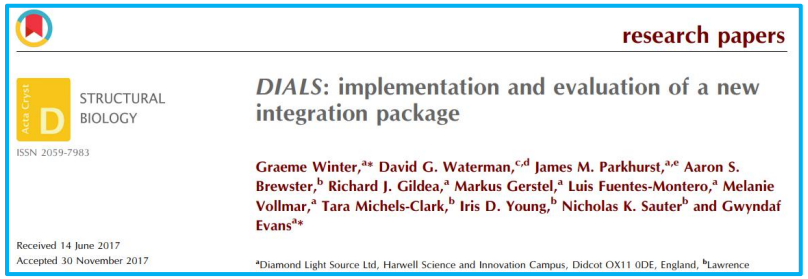
ISSN 0907-4449

XDS

Wolfgang Kabsch

The usage and control of recent modifications of the program package *XDS* for the processing of rotation images are

Received 19 August 2009
Accepted 9 November 2009



research papers

STRUCTURAL BIOLOGY

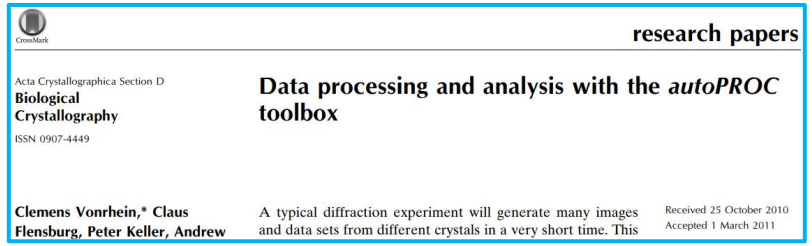
ISSN 2059-7983

DIALS: implementation and evaluation of a new integration package

Graeme Winter,[a]* David G. Waterman,[c,d] James M. Parkhurst,[a,e] Aaron S. Brewster,[b] Richard J. Gildea,[a] Markus Gerstel,[a] Luis Fuentes-Montero,[a] Melanie Vollmar,[a] Tara Michels-Clark,[b] Iris D. Young,[b] Nicholas K. Sauter[b] and Gwyndaf Evans[a]*

Received 14 June 2017
Accepted 30 November 2017

[a]Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot OX11 0DE, England, [b]Lawrence



research papers

Acta Crystallographica Section D Biological Crystallography

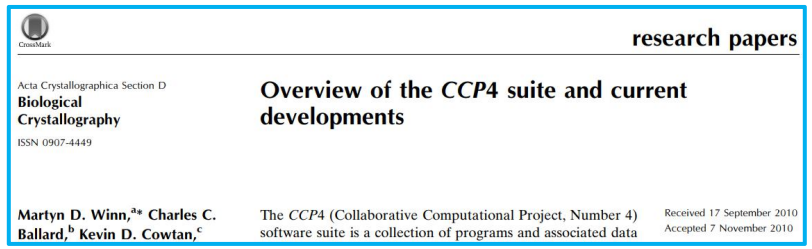
ISSN 0907-4449

Data processing and analysis with the *autoPROC* toolbox

Clemens Vonrhein,* Claus Flensburg, Peter Keller, Andrew

A typical diffraction experiment will generate many images and data sets from different crystals in a very short time. This

Received 25 October 2010
Accepted 1 March 2011



research papers

Acta Crystallographica Section D Biological Crystallography

ISSN 0907-4449

Overview of the *CCP*4 suite and current developments

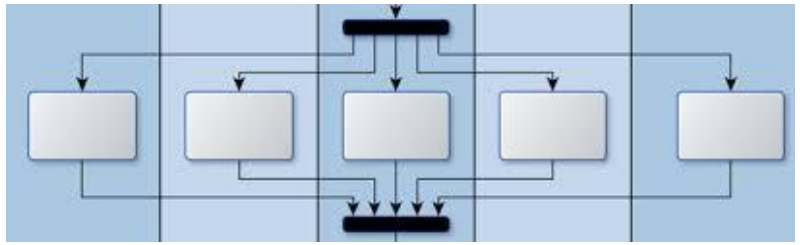
Martyn D. Winn,[a]* Charles C. Ballard,[b] Kevin D. Cowtan,[c]

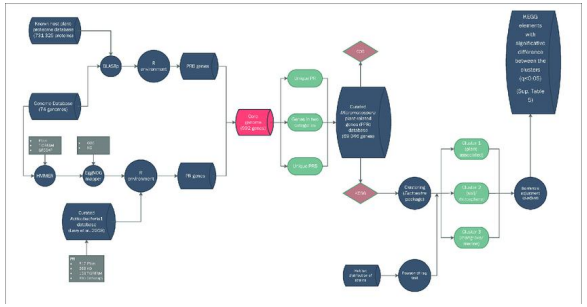
The *CCP*4 (Collaborative Computational Project, Number 4) software suite is a collection of programs and associated data

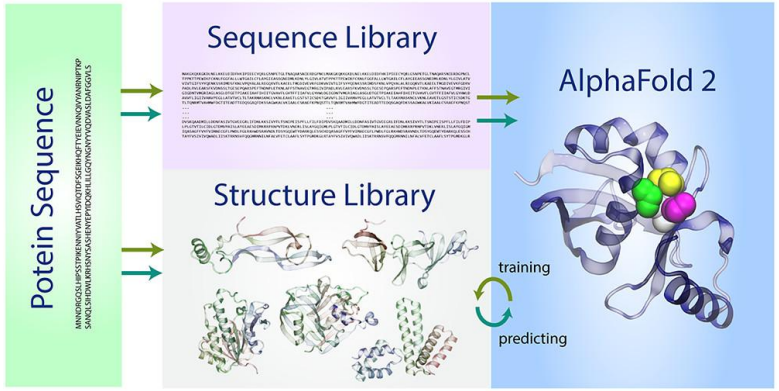
Received 17 September 2010
Accepted 7 November 2010

Design Goals

- **Automation**
  - Modules parallel-running for each step
  - Whole data processing as a pipeline
- **Intelligence**
  - Automatic/Manual selection of the best
  - AI: AlphaFold2 and Structure Refinement
- **Modularity**
  - High Performance： based on **Daisy** framework and **Docker** technology
  - Scalable: flexibly integrating new algo/sw



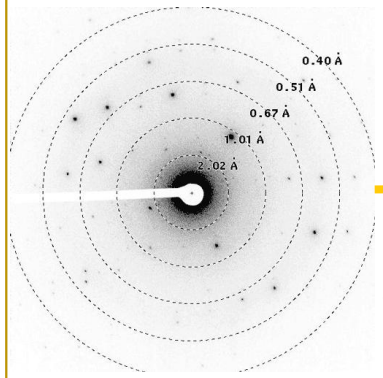




Daisy: check *https://indico.jlab.org/event/459/contributions/11400/*

Project Architecture

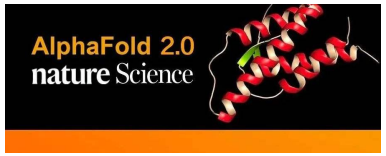
Main Project (Complete before HEPS Phase I)



XDS
DIALS
autoPROC

h	k	l	FreeR_flag	F	SIGF
0	2	16	6	344.83	2.70
0	2	18	7	310.91	2.50
0	2	20	18	1348.69	9.25
0	2	22	10	241.74	2.58
0	2	24	15	544.76	3.99
0	2	26	14	1163.75	8.00
0	2	28	17	351.58	3.45

Real-time diffraction data processing



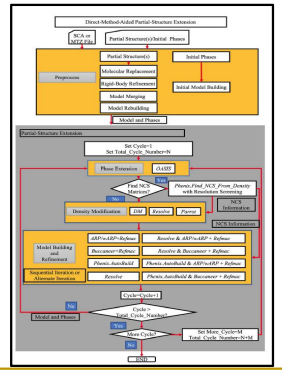
predicting

Structure Prediction from AlphaFold2

Scenario I

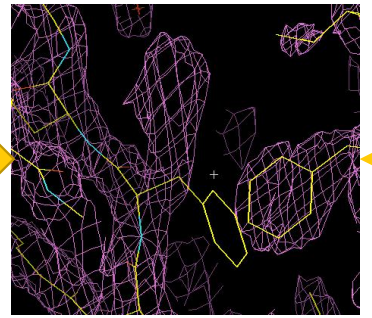
Phaser
Buccaneer
Phenix.autobuild

Scenario II: IPCAS

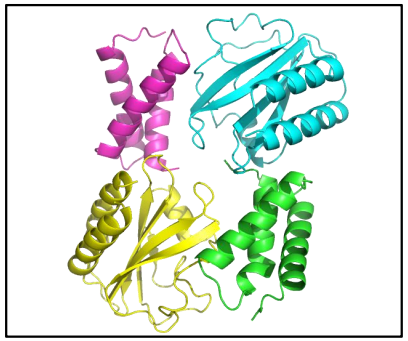


Phase Improving & Model Building

Long-Term Vision



Structure Refinement based on Machine Learning

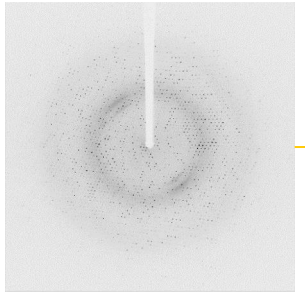


Based on direct-method, structure prediction, and AI, developing a softare system for data processing pipeline from exp. diffraction data to bio. macromolecule structures.

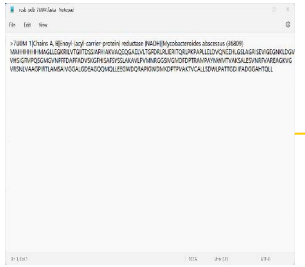
Project Architecture

Exp/User Input

Diffraction images



Protein Sequence

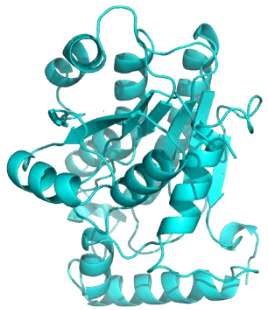


Pipeline processing

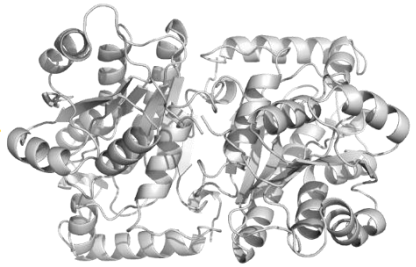
Data Reduction:
to get reflection data files (*.mtz)

	NATIVE
Wavelength (Å)	0.97854
Resolution range (Å)	4.19 - 1.44 (1.46 - 1.44)
Completeness (%)	80.76 (7.28)
Multiplicity	20.23 (1.15)
CC-half	0.9985 (0.0470)
I/sigma	0.16 (-0.03)
Rmerge(I)	1.2584 (7.0164)
Anomalous completeness (%)	72.92 (1.02)
Anomalous multiplicity	11.69 (1.03)

Structure Prediction:
to get preliminary structure (*.pdb)

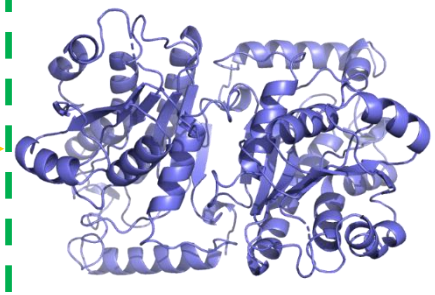


Phase Refinement:
to get final structure (.pdb)



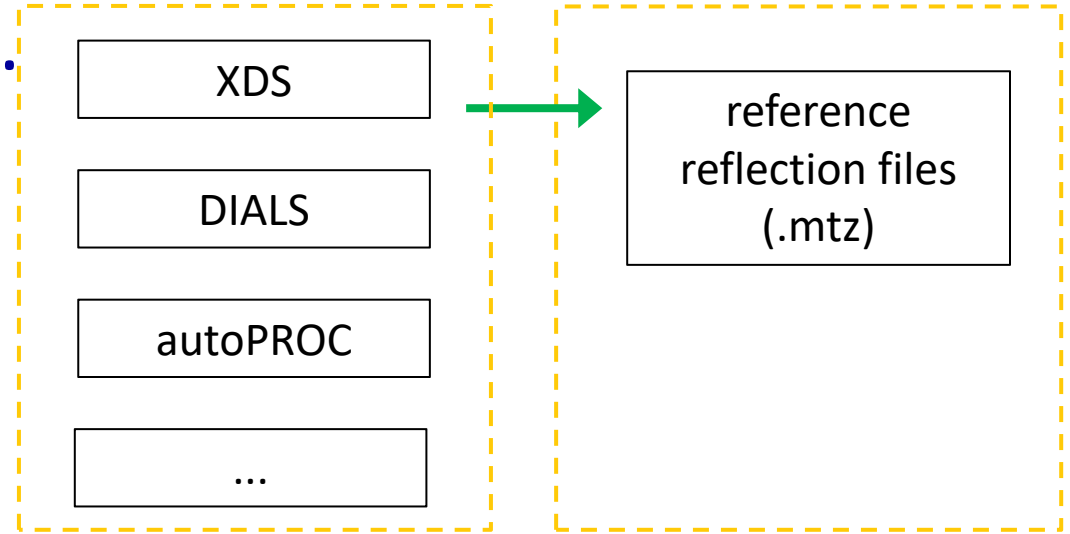
Pipeline Output

Valid & Reliable Structure

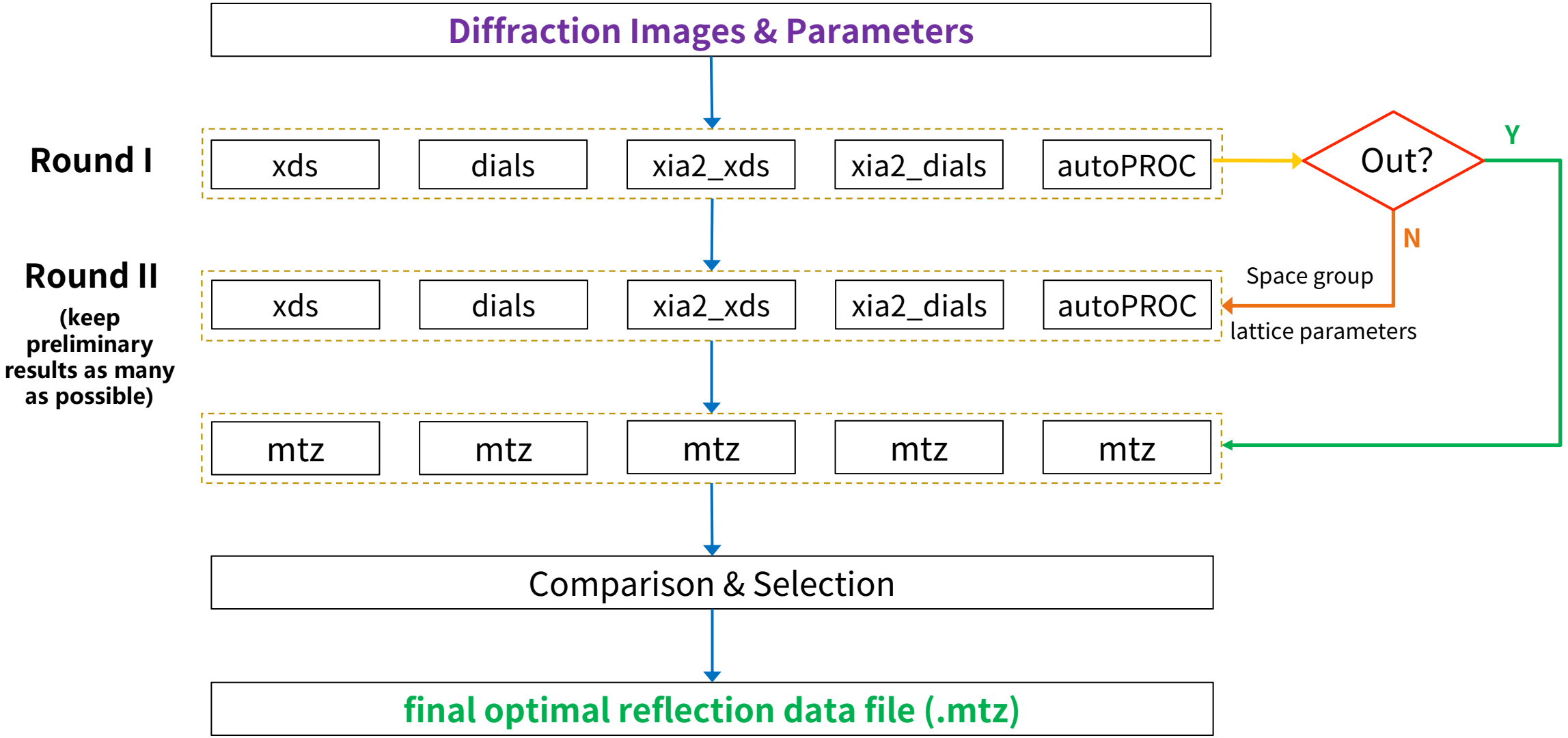


Current Status - Data Reduction

- For pipeline processing and GUI integration, each module can now be part of a *shell script* and/or *run automatically* after lots of tests & mod.
- Currently, several well-known packages, such as XDS, DIALS, and autoPROC, are integrated. more in the future
  - key parameters extraction (space group, resolution, etc.)
  - error handling
  - iteration improvements



Current Status - Data Reduction



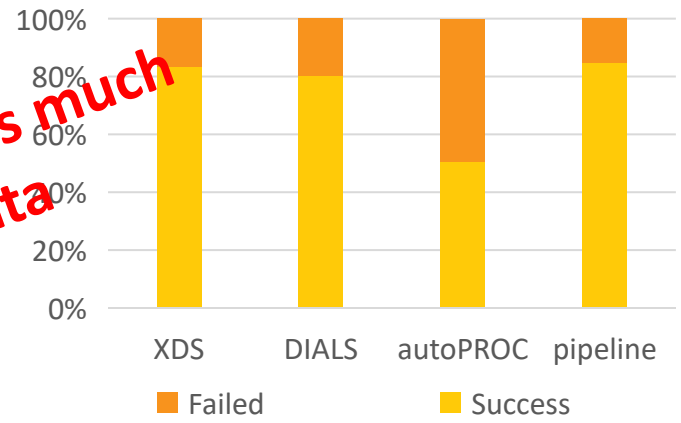
Data Reduction - Tests

Real data tests for Data Reduction

Total data sets: 156

Passed successfully: 131

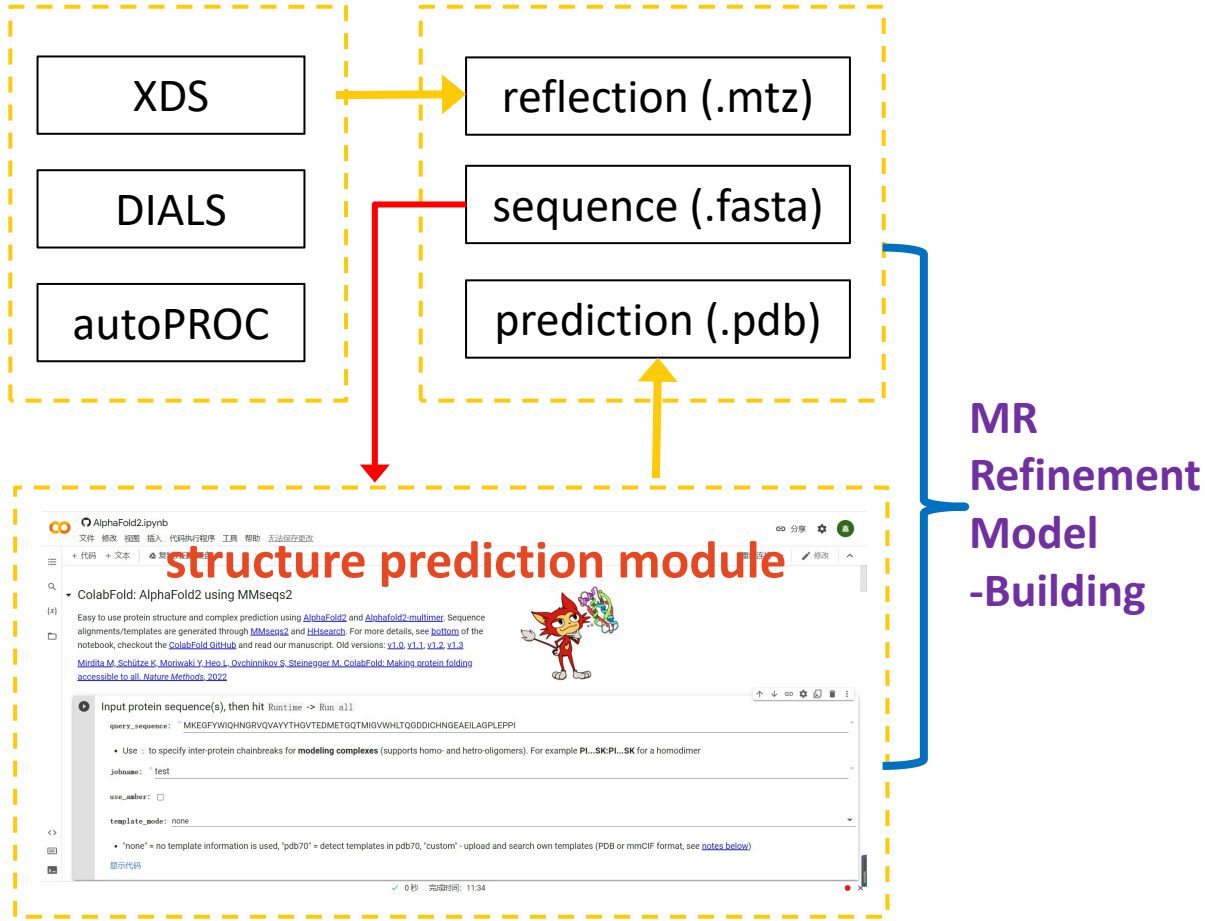
Under most cases, the output space groups are **identical** amont the three softwares



Index	Space group			Rmerge*			Resolution (Å)		
	XDS	DIALS	autoPROC	XDS	DIALS	autoPROC	XDS	DIALS	autoPROC
Case 1	C 1 2 1	C 1 2 1	C 1 2 1	**0.051**	**0.051**	**0.038**	1.54	1.69	1.93
Case 2	P 6			**0.242**			1.70		
Case 3	C 1 2 1	C 1 2 1		5.040	**0.259**		1.88	1.94	
Case 4		P 4	P 41		0.581	**0.080**		2.01	1.78
Case 5	P 61	P 61		**0.036**	**0.040**		1.47	1.42	

Current Status - Prediction

- Each data reduction module is running in parallel, and all results are kept.
- Considering resolution, **Rmerge**, I/sigma, Completeness, *etc.* key parameters, and provide rational rec. for the optimal solution. (GBDT, GA)
- When the sequence .fasta file is available, performe the structure prediction **asynchronously**

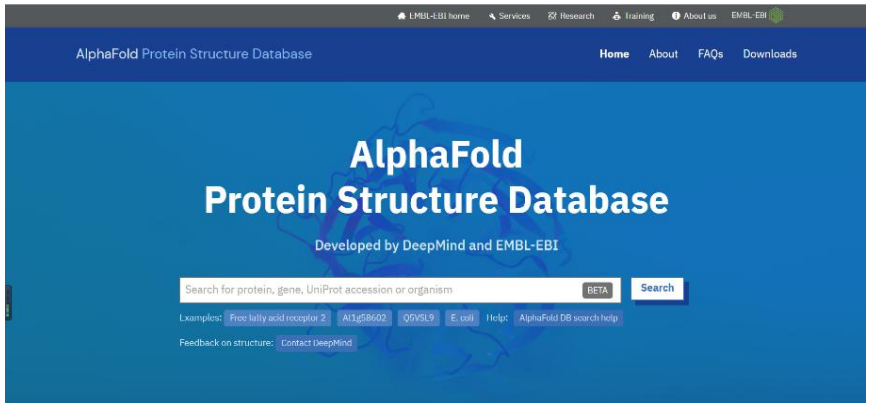


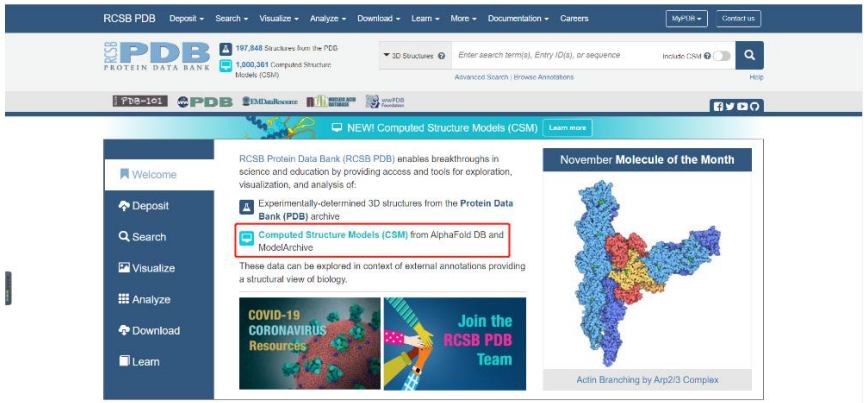
Current Status - Prediction

Preliminary Structure Prediction

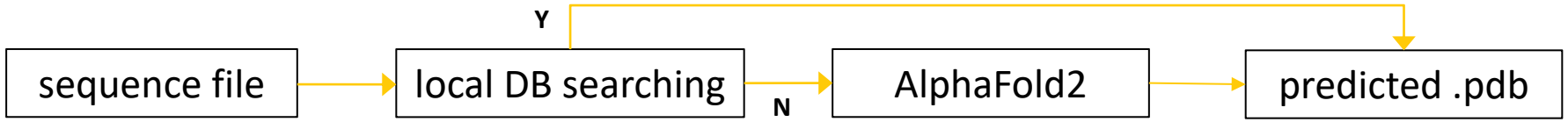
Introduce AlphaFold2 and/or .pdb database, make prediction from sequence files.

- Scenario A: searching predicted structure database (fast, but not friendly to unconventional sequences)





- Scenario B: predicted using AlphaFold2 workstation (high accuracy, but slow speed)



Current Status - Prediction

- Deploy AlphaFold2 docker, and keep it running & updating **v2.3.1**
- Using localized NVMe SSD to speed up the database-searching (bfd, uniref, uniprot, etc.) stage, **~3TB**
- By setting the Unified Memory (TF, JAX) mode to support long sequence prediction **~2800,25h**
- Wrapping them into easily-used shell scripts

steps	HEPS	Ref.
Searching Database	688.24 s	1825.35 s
Process	3.90 s	8.28 s
Predict	413.45 s	508.63 s
Relax	60.92 s	61.41 s
Total	44 m	75 m

CPU(96) MEM(512G) 2xA100(80G) => 4 users

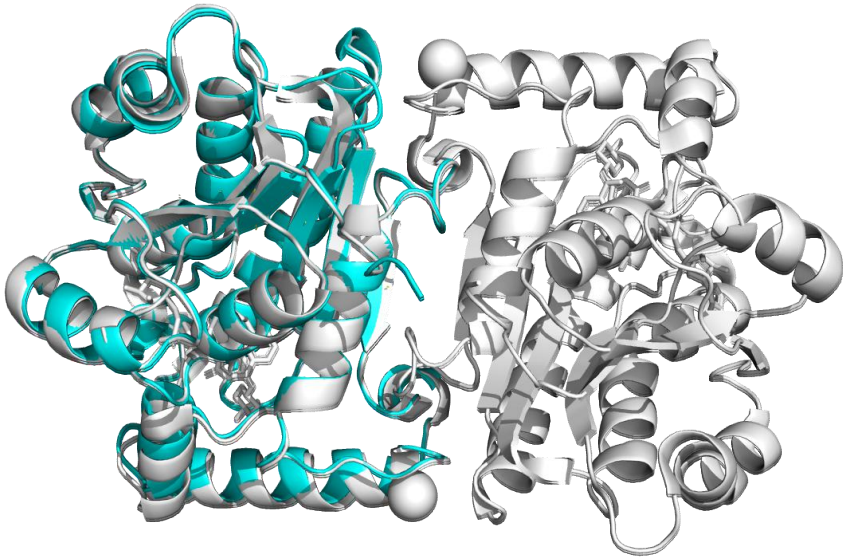
```
AF_TEMPLATE="2020-05-14"
# [monomer, monomer_casp14, monomer_ptm, multimer]
AF_MODEL="multimer"
# [full_dbs, reduced_dbs]
AF_DBPRESET="full_dbs"

queryFile="/path/to/fasta/file"
outputDir="/path/to/output/folder"
```

Current Status - Prediction

Test Case: (PDB Code: 7U0M)

- Scenario **A**: seaching PDB database

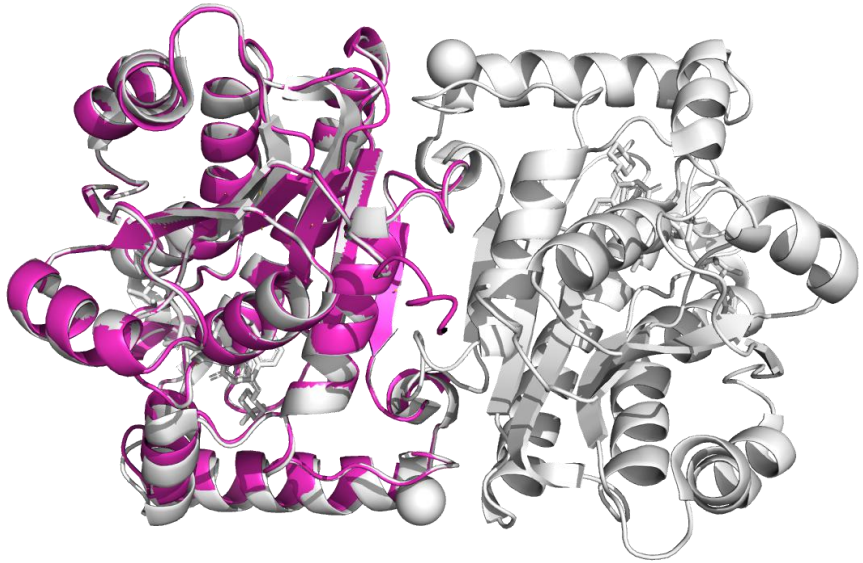


Time cost: <10s

Sequence identity: 100%

R.M.S.D: 0.4887 A

- Scenario **B**: predicted by AlphaFold2 server



Time cost: 3226s

Sequence identity: 100%

R.M.S.D: 0.4672 A

Current Status - Refinement

XDS

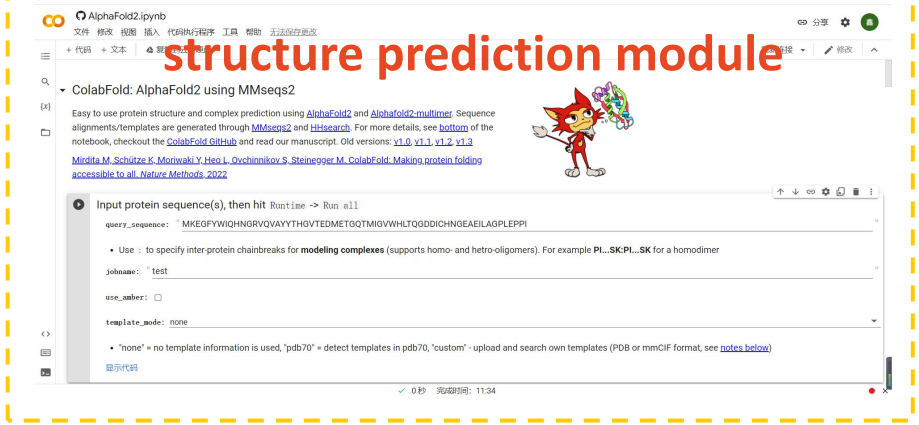
DIALS

autoPROC

reflection (.mtz)

sequence (.fasta)

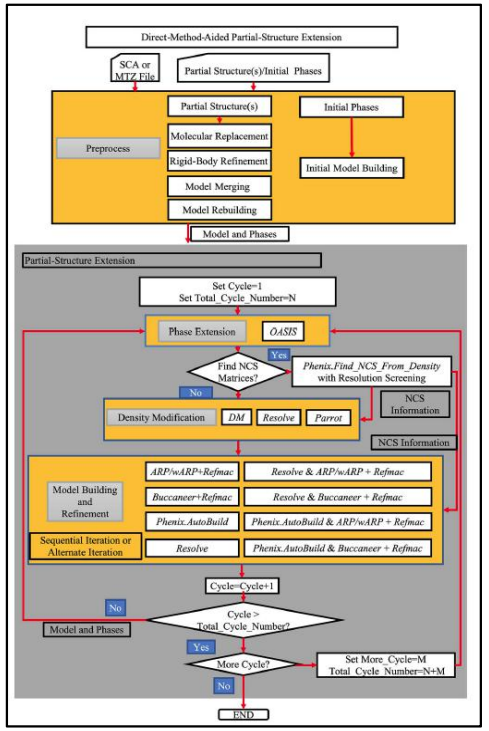
prediction (.pdb)



structure prediction module

Scenario I: IPCAS3.0

Molecular Replacement (phaser)



Higher Accuracy, lower speed

Scenario II: Traditional

Cell Content Analysis (Matthew's coefficient)

Molecular Replacement (phaser)

Refinement and Model Building (phenix.autobuild)

Lower Accuracy, Higher speed

Current Status - Refinement

Test Case: (PDB Code: 7U0M)

☐Scenario I: IPCAS3.0

Time cost: 6h24m7s

R-work: 0.2360

R-free : 0.2537

R.M.S.D: 0.2877 A

☐Scenario II: Traditional

Time cost: 15m38s

R-work: 0.3390

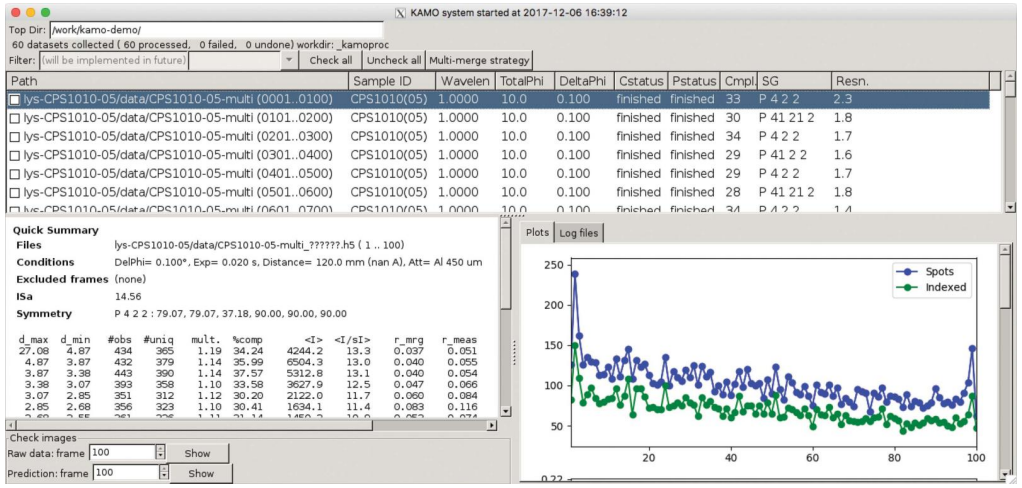
R-free: 0.3519

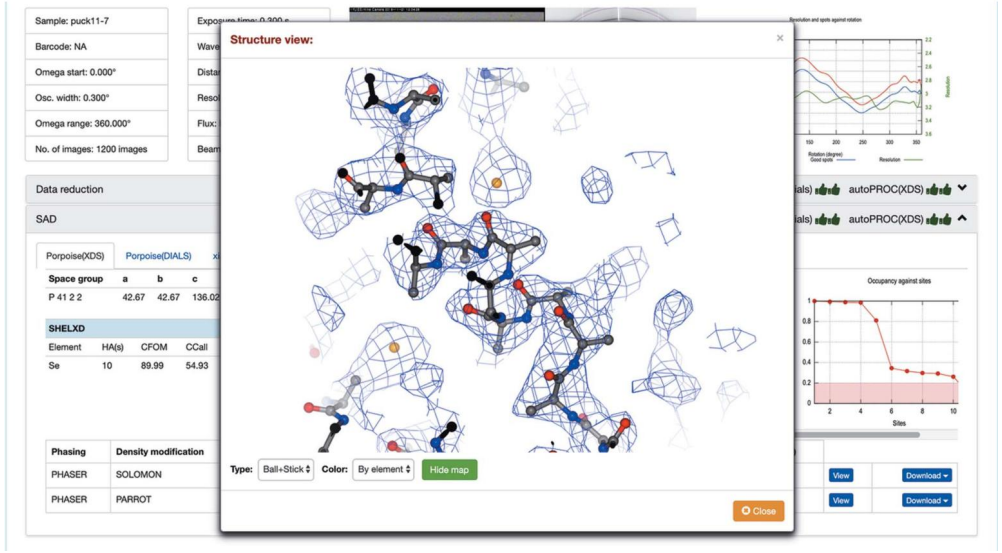
R.M.S.D: 0.2991 A

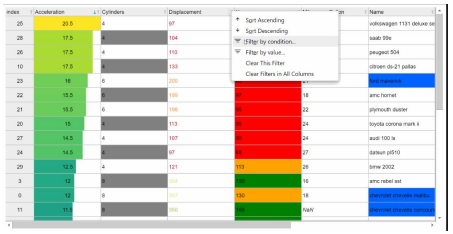
☐Standard Structure

Current Status - GUI

- Best practices: KAMO, Aquarium
  - Real operational experiences
  - User preferences and interface desgin logic
- Supporting platform/framework:
  - container+Jupyterlab
  - **Daisy** Computing, IO, Workflow modules
- Technical solutions
  - ipywidgets +
  - pandas,ipydatagrid, matplotlib
  - in-house developed widgets



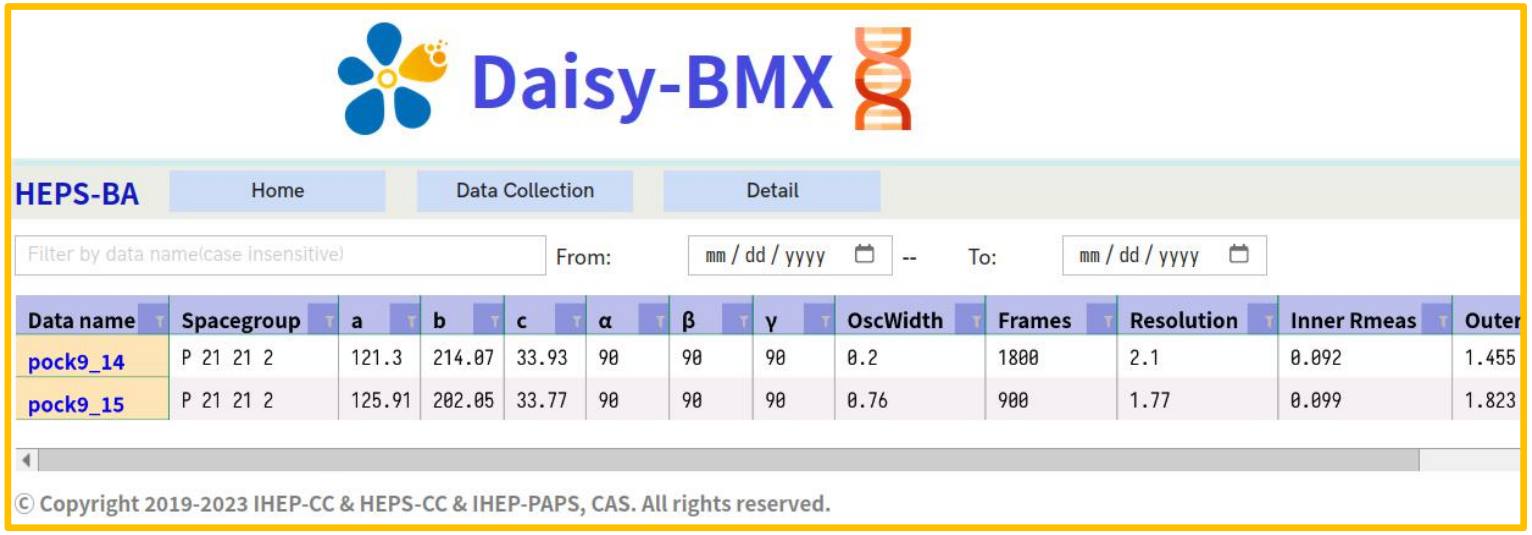


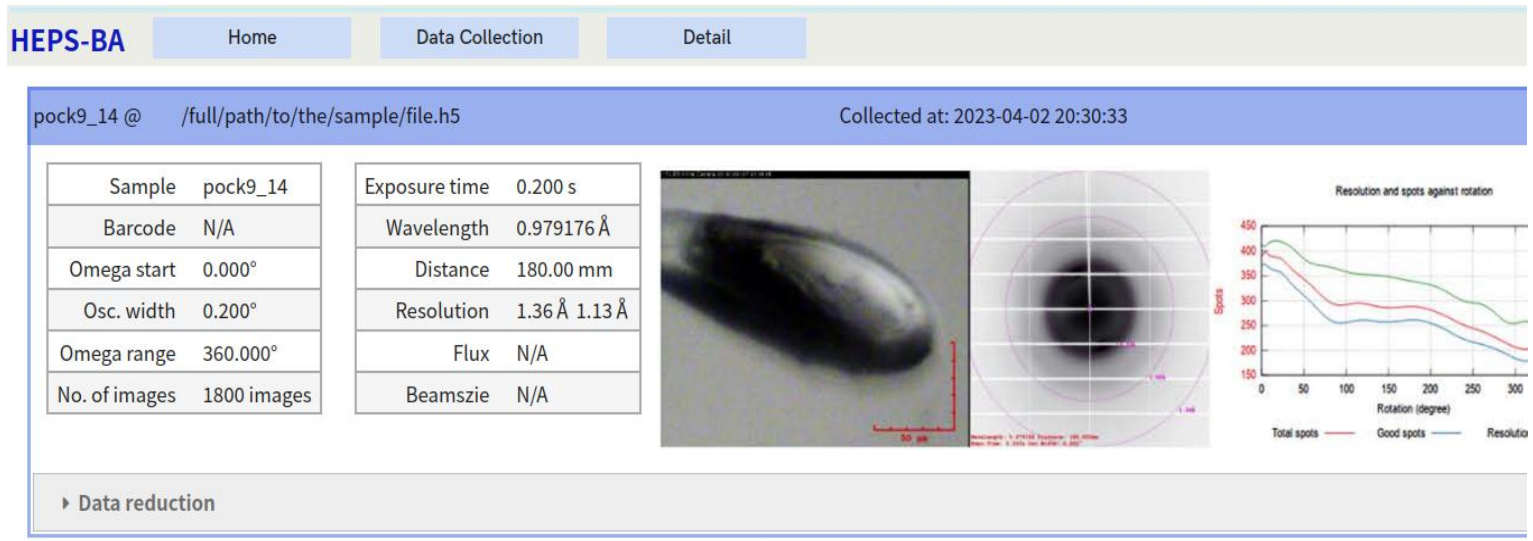


Current Status - GUI

Data Collection Tab

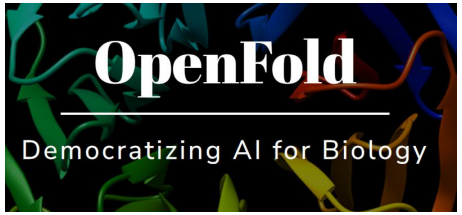
Data Processing Details Tab

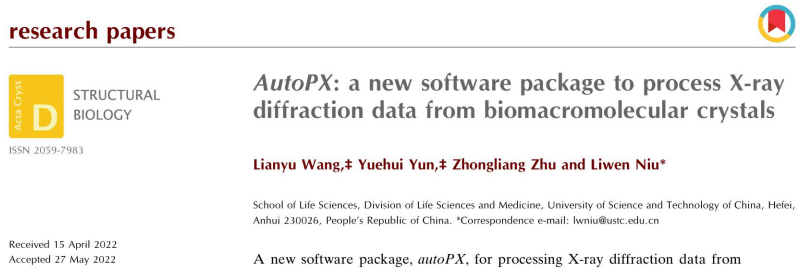




Plans & Outlook

- Modularity
  - XDS、DIALS、autoPROC + autoPX，…
  - AlphaFold2 + OpenFold，…
  - Phaser、IPCAS + …
- Performance
  - Parallelization, GPU acceleration for traditional algo/sw
  - AlphaFold2 CPU/GPU heterogeneous acceleration
  - localize predicted structures (*.pdb)
    - The ESM Metagenomic Atlas ~600M
    - AlphaFold Protein Structure Database(EMBL-EBI) >200M




research papers

STRUCTURAL BIOLOGY

ISSN 2059-7983

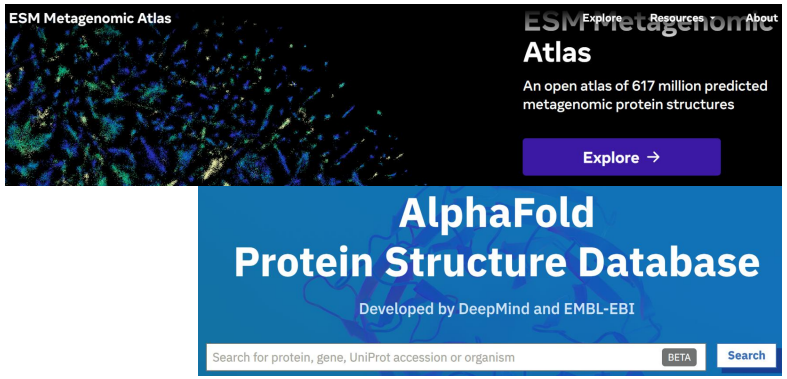
AutoPX: a new software package to process X-ray diffraction data from biomacromolecular crystals

Lianyu Wang,‡ Yuehui Yun,‡ Zhongliang Zhu and Liwen Niu*

School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China. *Correspondence e-mail: lwniu@ustc.edu.cn

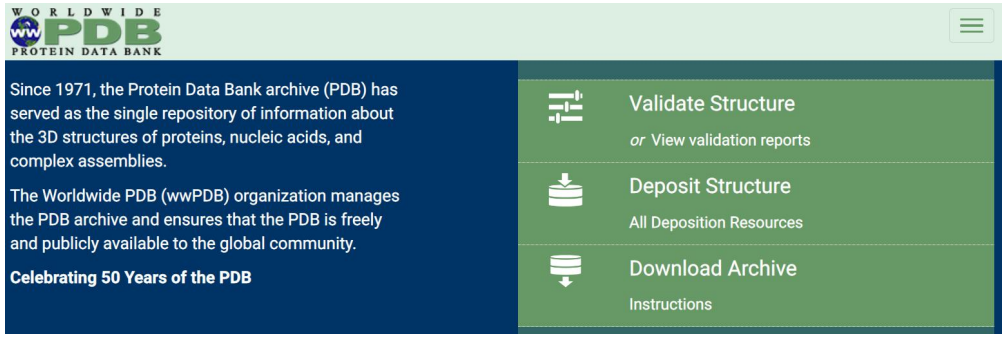
Received 15 April 2022
Accepted 27 May 2022

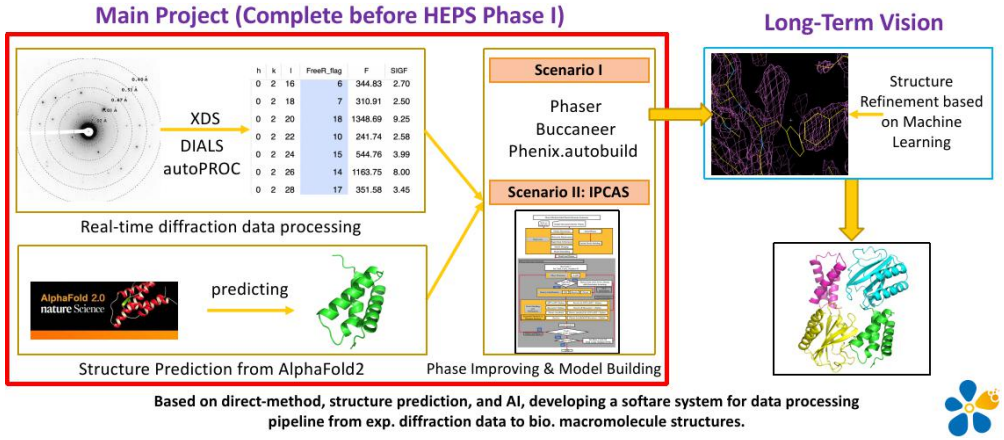
A new software package, *autoPX*, for processing X-ray diffraction data from



Plans & Outlook

- **PDB China & wwPDB**
  - Make contributions, Collaborate on division of labor, Promote PDB China jointly.
- Lots of real exp. data tests
- Data analysis pipeline customization & automation
- User-friendly and details-improved GUI





Thanks !