CCP4 Cloud: Potential for a European-Wide Resource in Computational Crystallography

Eugene Krissinel

Research Complex at Harwell, Rutherford-Appleton Laboratory,
Science & Technology Facilities Council, Harwell, UK

eugene.krissinel@stfc.ac.uk

Science and Technology Facilities Council

Research Complex at Harwell

CCP4

Science and Data Today

- In many fields, research becomes progressively data-driven
  - *taking on complex systems (e.g., climate, molecular biology, medicine)*
  - *puristic models not fitting real-world situations*
  - *imperfect metrics and rules for decision making*
  - *incomplete understanding and knowledge base*
  - *“I need to get things working first, and understand how they work second”*

- Combination of data, AI and computing power starts giving a break-through in many directions

  Recent example: predicting protein structures with AlphaFold 2 from DeepMind Technologies
  - *unprecedented accuracy in predicting protein structures in 3d*
  - *expanding proteomics to genomics scales (from 180K known to 130M new structures)*
  - *implications for structural biology research are breathtaking and still not fully understood*

Data and Computing Infrastructures

PSDS
PHYSICAL SCIENCES DATA-SCIENCE SERVICE

ESCAPE
DIOS | Data Infrastructure for Open Science

Materials data platform
DICE

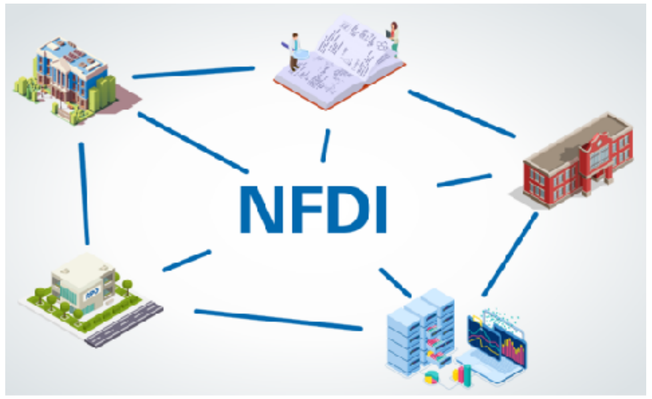
MatNavi
NIMS Materials Database

ARDC
Australian Research Data Commons

EUROPEAN OPEN SCIENCE CLOUD

DAFNI

D4SCIENCE
INFRASTRUCTURE



OpenAIRE

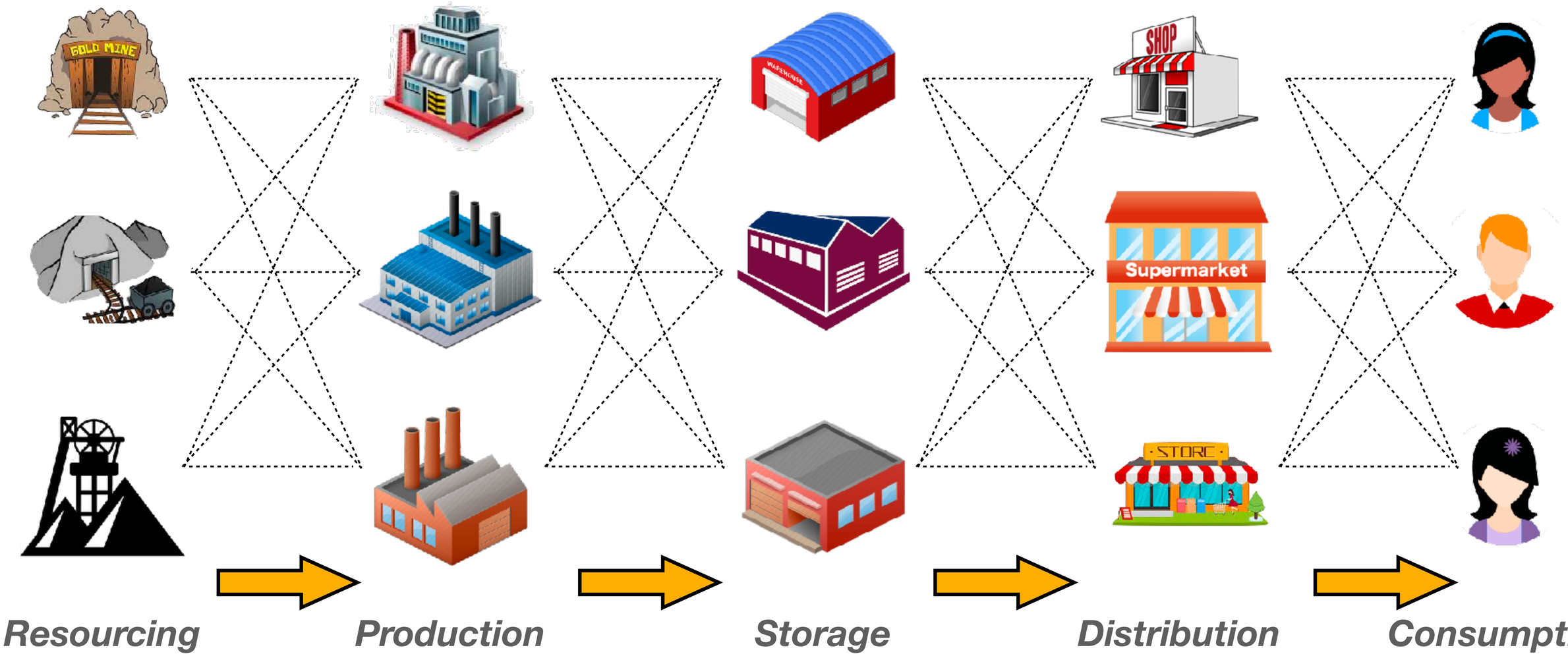
UKRI Science and Technology Facilities Council

Research Complex at Harwell

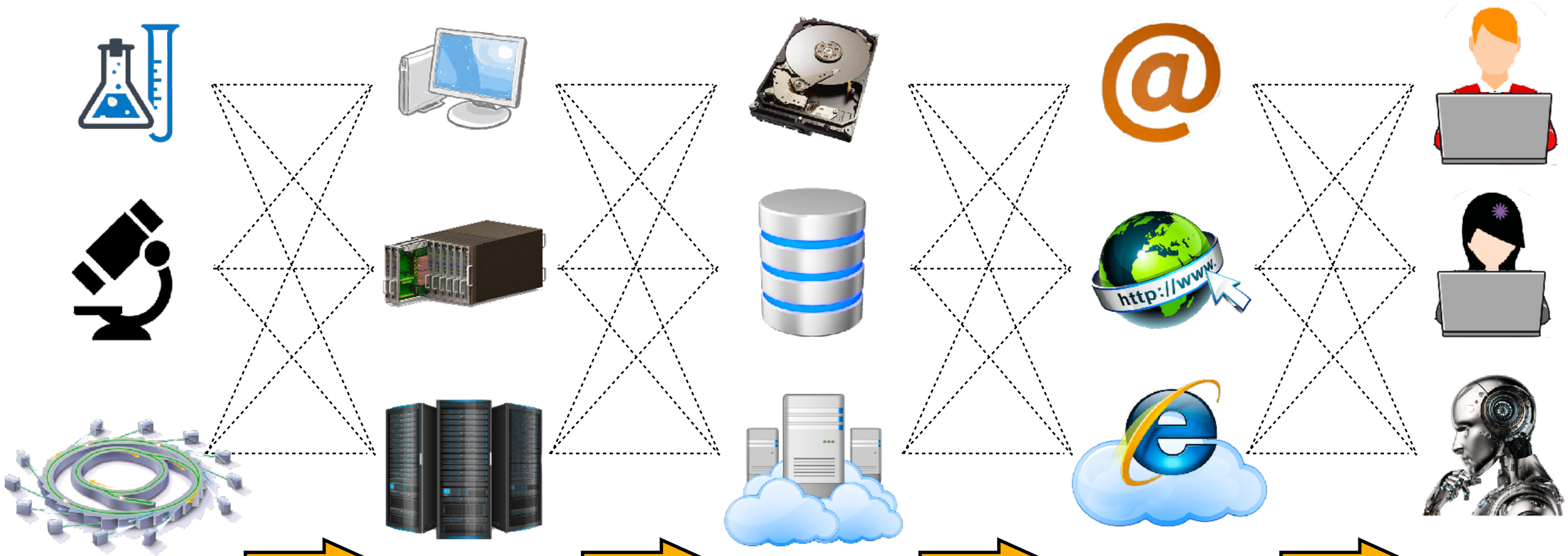
CCP4

EGI and Research Infrastructures

Gross Idea: Distributed Production with Unified Access



Gross Idea: Distributed Production with Unified Access



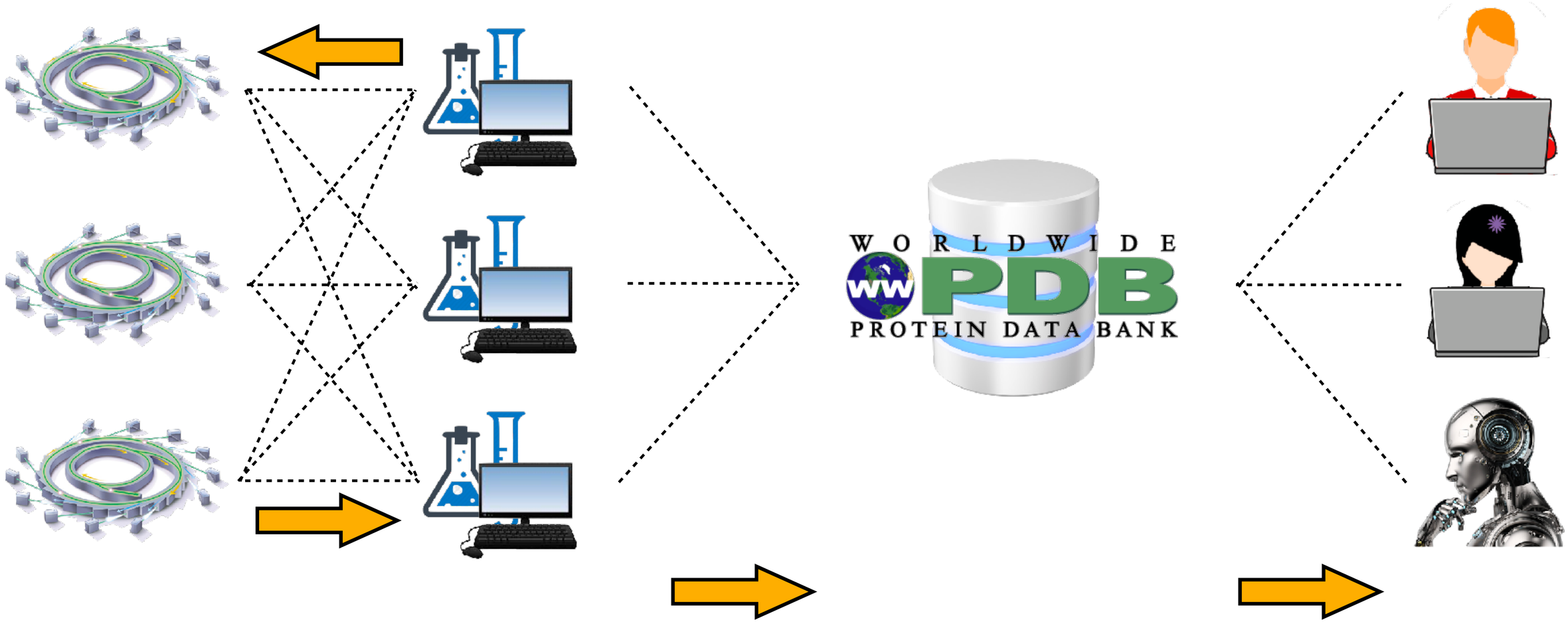
Experiment *Data Processing* *Data Storage* *Data Servers* *Users*

Science and Technology Facilities Council

Research Complex at Harwell

CCP4

Data Production in Structural Biology



Imaging Facilities *Research Labs* *Data Bank and Server* *Users*

Data Maintenance in Structural Biology

- Single repository
  - *Protein Data Bank established in 1971*
  - *Underpins academic and industrial research on global scale*
- Does not keep all data
  - *originally kept only atomic coordinates of biological macromolecules*
  - *reduced experimental data accepted from 1999 and made mandatory in 2008*
  - *processed experimental data accepted from 2020 but not mandatory*
  - *raw experimental data are not deposited*
- No framework for keeping structure solution projects
  - *valuable information about how the structure was solved, is not retained*
  - *structure solution cost varies between $100K and $2M*
  - *the total cost of re-creating the PDB, if lost, is estimated at over $4B*
  - *life-time expectancy of data kept in local labs may be as short as few years*

CCP4 Cloud Initiative

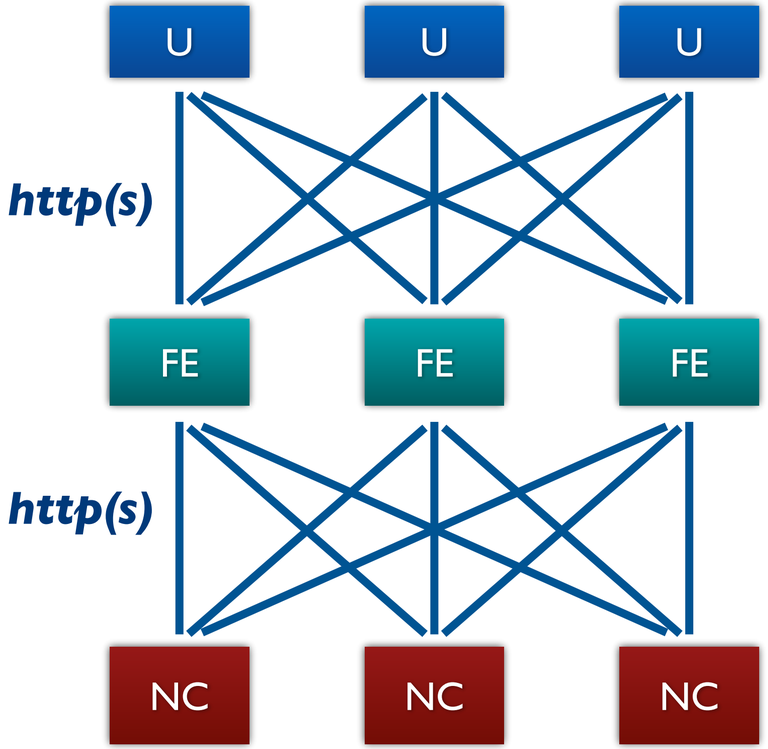
- Conceived in 2016
  - *Funded by BBSRC UK and CCP4*
- Response to demands and trends rapidly emerging in the field
  - *CPU power (due to increased automation)*
  - *Centralised database support (due to expansion of methods based on data templates)*
  - *Software as a service (due to increased size and complexity of software setups)*
  - *Supporting distributed projects for team work*
  - *Cloud model for geographically-agnostic access and project data safety*
  - *Supporting personal mobile platforms (tablets and smartphones)*
  - *Communication with data facilities (synchrotrons, PDB, AFDB, etc)*
- Released in 2018
  - *Add-on in CCP4 7.0 (September 2018), fully integrated in CCP4 7.1 (April 2020)*
  - *Timely in pandemic situation*
  - *2,000 registered users*

CCP4 Cloud Architecture

User Nodes (Integration)

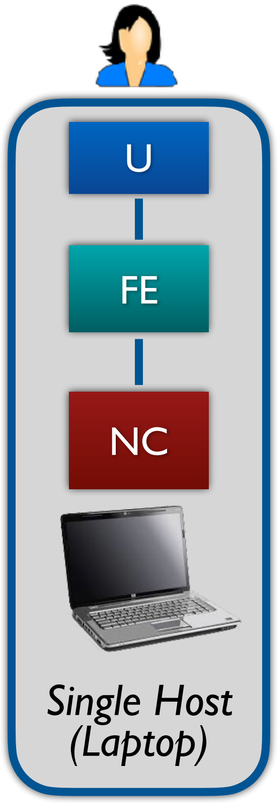
Front End Nodes (Data and communication)

Number Cruncher Nodes (Computations)

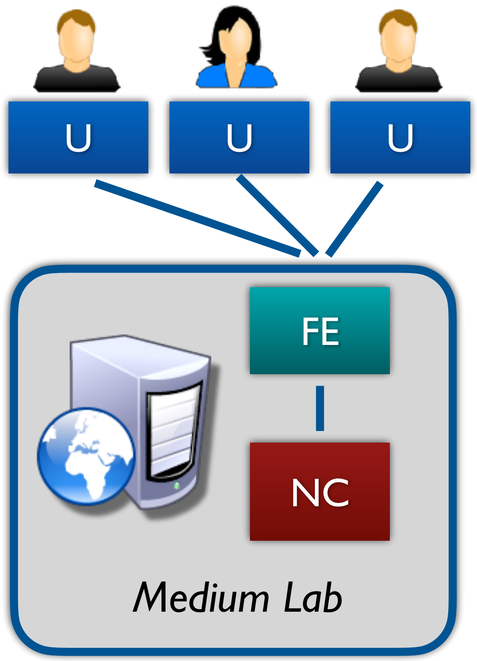


- Not an application, but rather a configurable and expandable framework
  - *All nodes are functionally independent entities*
  - *Configuration done with light-weight JSON files*
  - *Based on Node JS*
  - *Python adapters for executables*
  - *No principal restrictions on the location and number of components*

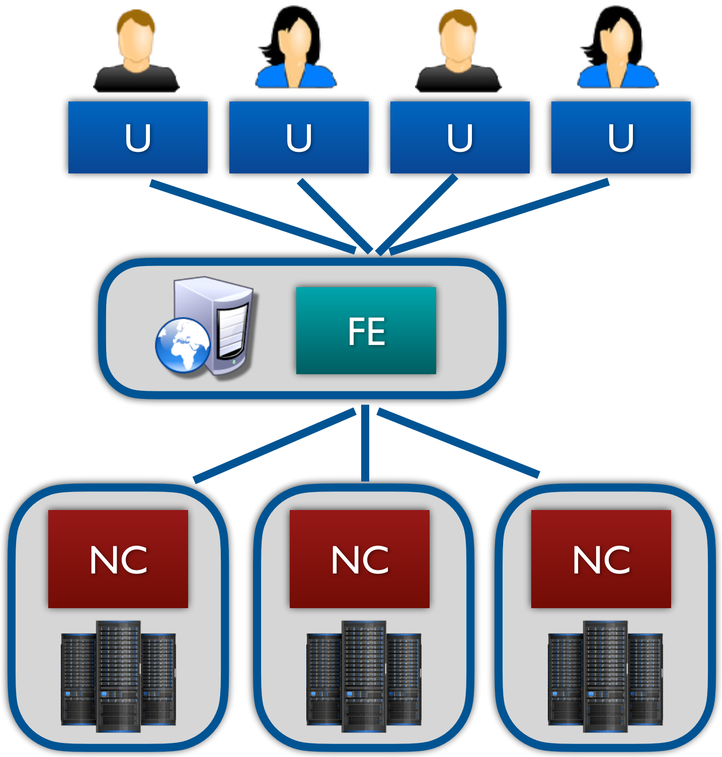
Typical CCP4 Cloud Configurations



out-of-the-box



almost out-of-the-box



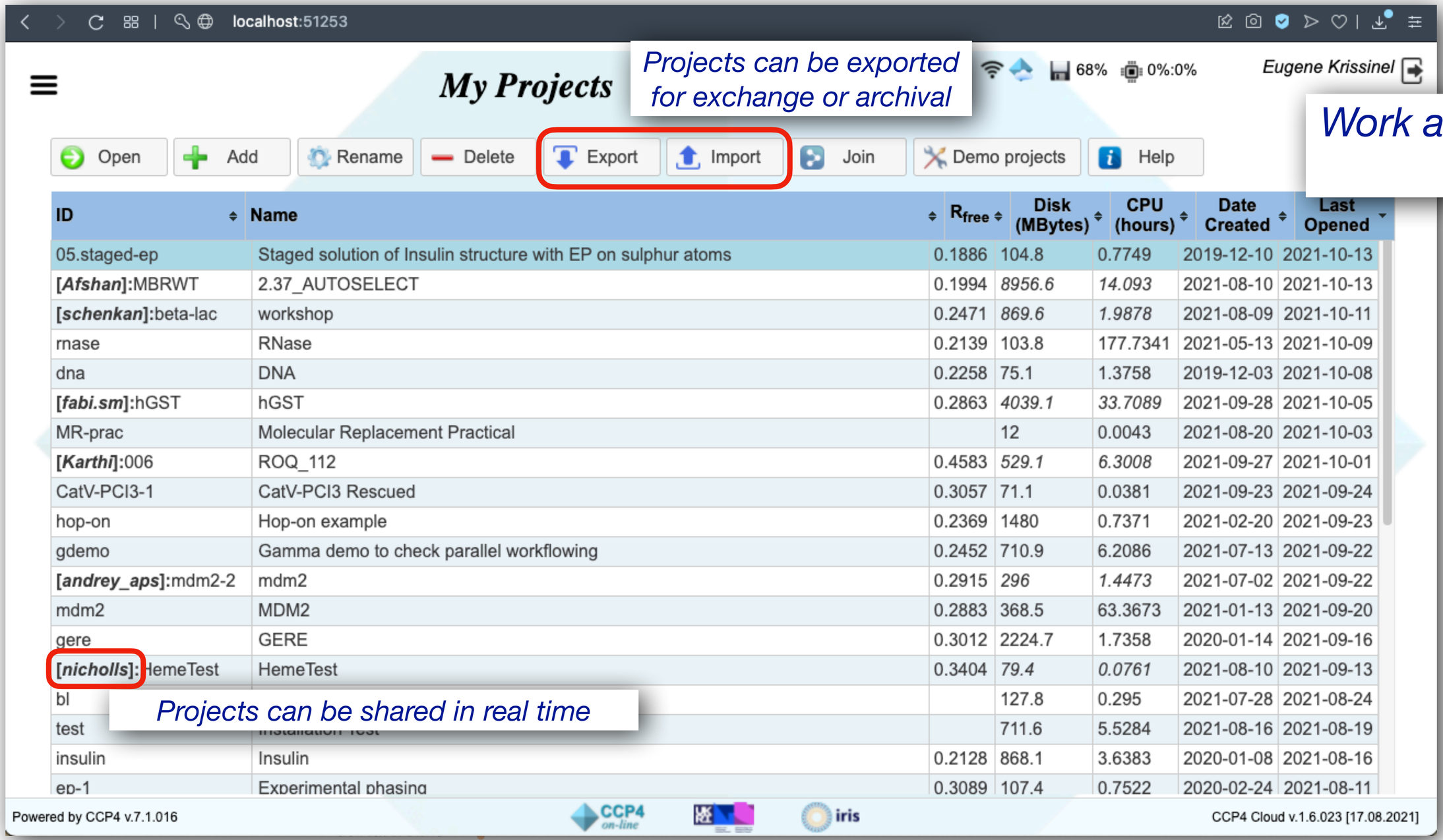
Major Hub (CCP4-Harwell)

- From CCP4 7.1, any CCP4 setup may be used as one or few CCP4 Cloud node(s)
  - *All nodes have identical codebase*
  - *Single Host comes as a standard (effectively a desktop GUI)*
  - *User Node comes pre-configured for working with CCP4-Harwell instance*

On User Side

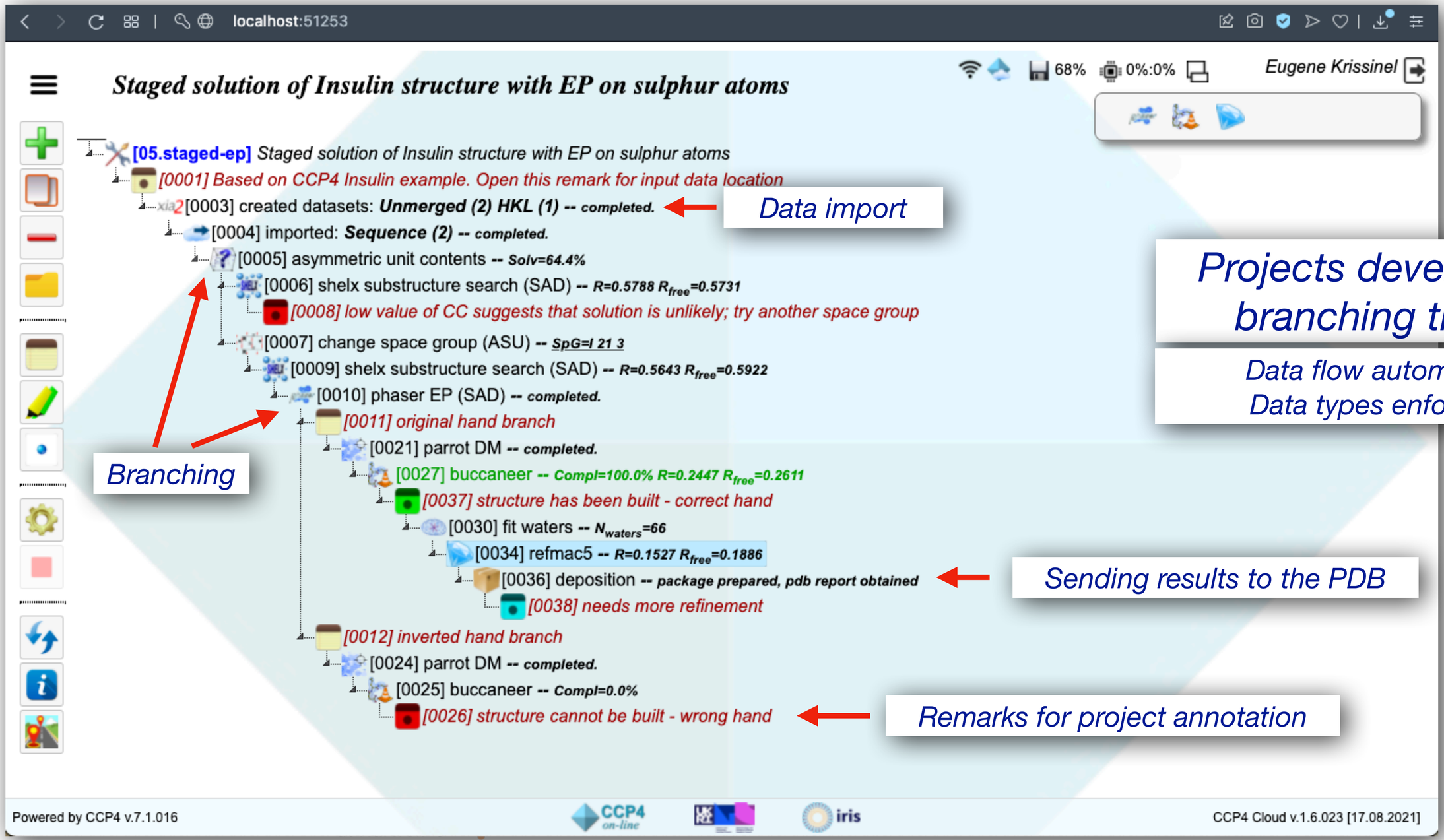
A Web-Application with rich graphical front-end

On User Side

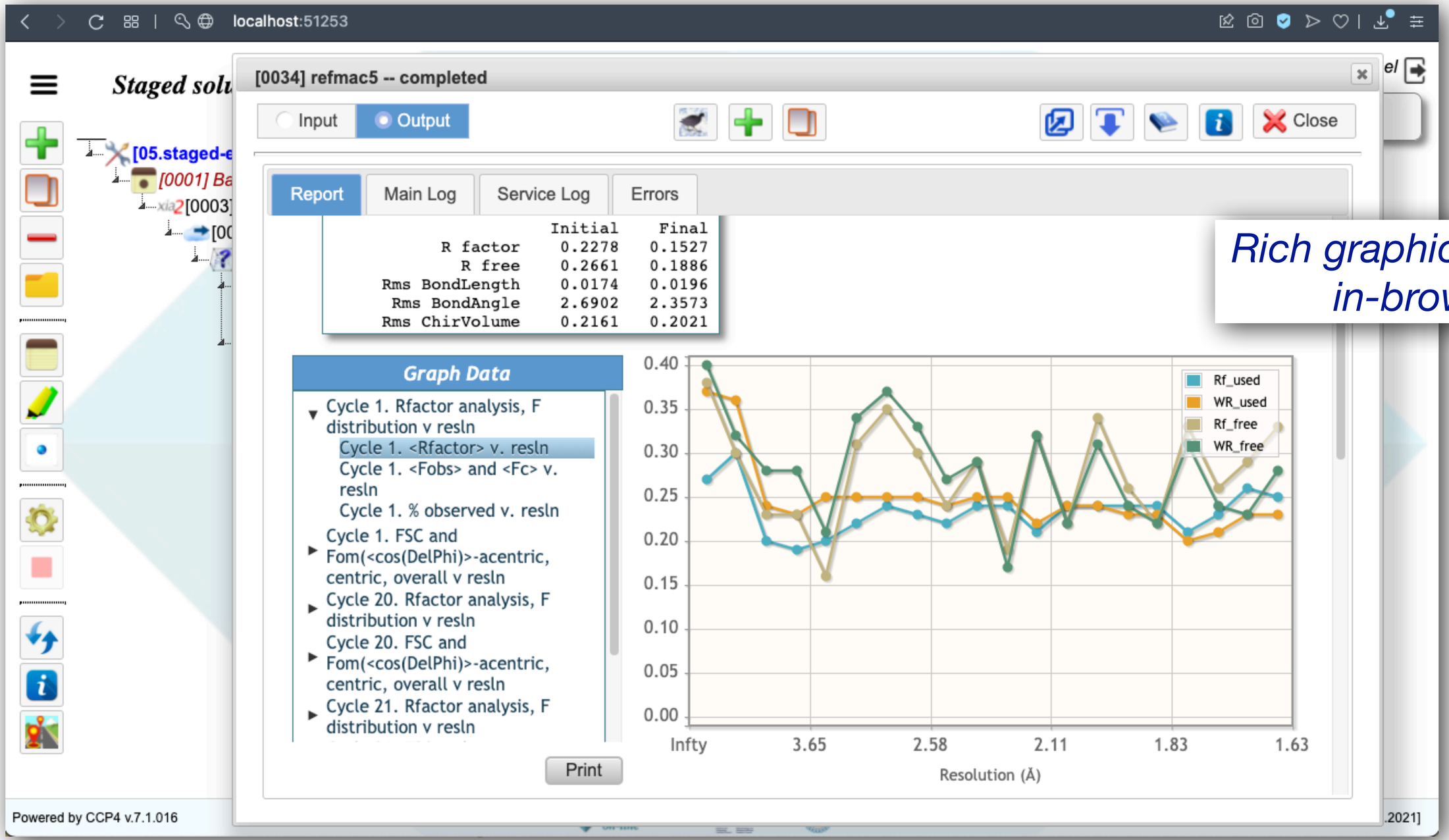


ID	Name	R_{free}	Disk (MBytes)	CPU (hours)	Date Created	Last Opened
05.staged-ep	Staged solution of Insulin structure with EP on sulphur atoms	0.1886	104.8	0.7749	2019-12-10	2021-10-13
[*Afshan*]:MBRWT	2.37_AUTOSELECT	0.1994	*8956.6*	*14.093*	2021-08-10	2021-10-13
[*schenkan*]:beta-lac	workshop	0.2471	*869.6*	*1.9878*	2021-08-09	2021-10-11
rnase	RNase	0.2139	103.8	177.7341	2021-05-13	2021-10-09
dna	DNA	0.2258	75.1	1.3758	2019-12-03	2021-10-08
[*fabi.sm*]:hGST	hGST	0.2863	*4039.1*	*33.7089*	2021-09-28	2021-10-05
MR-prac	Molecular Replacement Practical		12	0.0043	2021-08-20	2021-10-03
[*Karthi*]:006	ROQ_112	0.4583	*529.1*	*6.3008*	2021-09-27	2021-10-01
CatV-PCI3-1	CatV-PCI3 Rescued	0.3057	71.1	0.0381	2021-09-23	2021-09-24
hop-on	Hop-on example	0.2369	1480	0.7371	2021-02-20	2021-09-23
gdemo	Gamma demo to check parallel workflowing	0.2452	710.9	6.2086	2021-07-13	2021-09-22
[*andrey_aps*]:mdm2-2	mdm2	0.2915	*296*	*1.4473*	2021-07-02	2021-09-22
mdm2	MDM2	0.2883	368.5	63.3673	2021-01-13	2021-09-20
gere	GERE	0.3012	2224.7	1.7358	2020-01-14	2021-09-16
[*nicholls*]:HemeTest	HemeTest	0.3404	*79.4*	*0.0761*	2021-08-10	2021-09-13
bl			127.8	0.295	2021-07-28	2021-08-24
test			711.6	5.5284	2021-08-16	2021-08-19
insulin	Insulin	0.2128	868.1	3.6383	2020-01-08	2021-08-16
ep-1	Experimental phasing	0.3089	107.4	0.7522	2020-02-24	2021-08-11

On User Side

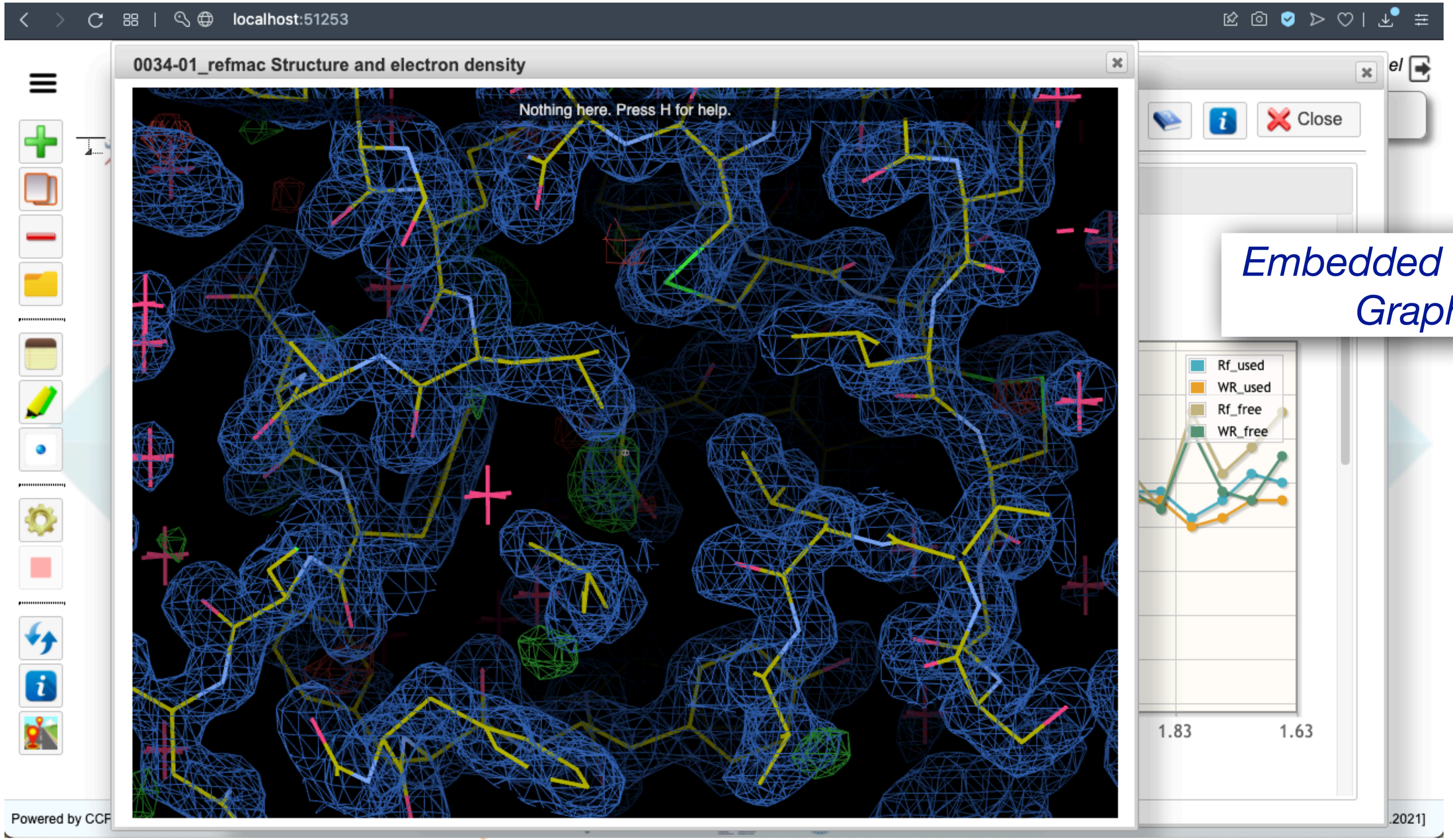

localhost:51253
Staged solution of Insulin structure with EP on sulphur atoms
68%
0%:0%
Eugene Krissinel
[05.staged-ep] Staged solution of Insulin structure with EP on sulphur atoms
[0001] Based on CCP4 Insulin example. Open this remark for input data location
[0003] created datasets: Unmerged (2) HKL (1) -- completed.
[0004] imported: Sequence (2) -- completed.
[0005] asymmetric unit contents -- Solv=64.4%
[0006] shelx substructure search (SAD) -- R=0.5788 Rfree=0.5731
[0008] low value of CC suggests that solution is unlikely; try another space group
[0007] change space group (ASU) -- SpG=I 21 3
[0009] shelx substructure search (SAD) -- R=0.5643 Rfree=0.5922
[0010] phaser EP (SAD) -- completed.
[0011] original hand branch
[0021] parrot DM -- completed.
[0027] buccaneer -- Compl=100.0% R=0.2447 Rfree=0.2611
[0037] structure has been built - correct hand
[0030] fit waters -- Nwaters=66
[0034] refmac5 -- R=0.1527 Rfree=0.1886
[0036] deposition -- package prepared, pdb report obtained
[0038] needs more refinement
[0012] inverted hand branch
[0024] parrot DM -- completed.
[0025] buccaneer -- Compl=0.0%
[0026] structure cannot be built - wrong hand
Data import
Branching
Projects develop as branching trees
Data flow automated
Data types enforced
Sending results to the PDB
Remarks for project annotation
Powered by CCP4 v.7.1.016
CCP4 on-line
iris
CCP4 Cloud v.1.6.023 [17.08.2021]

On User Side

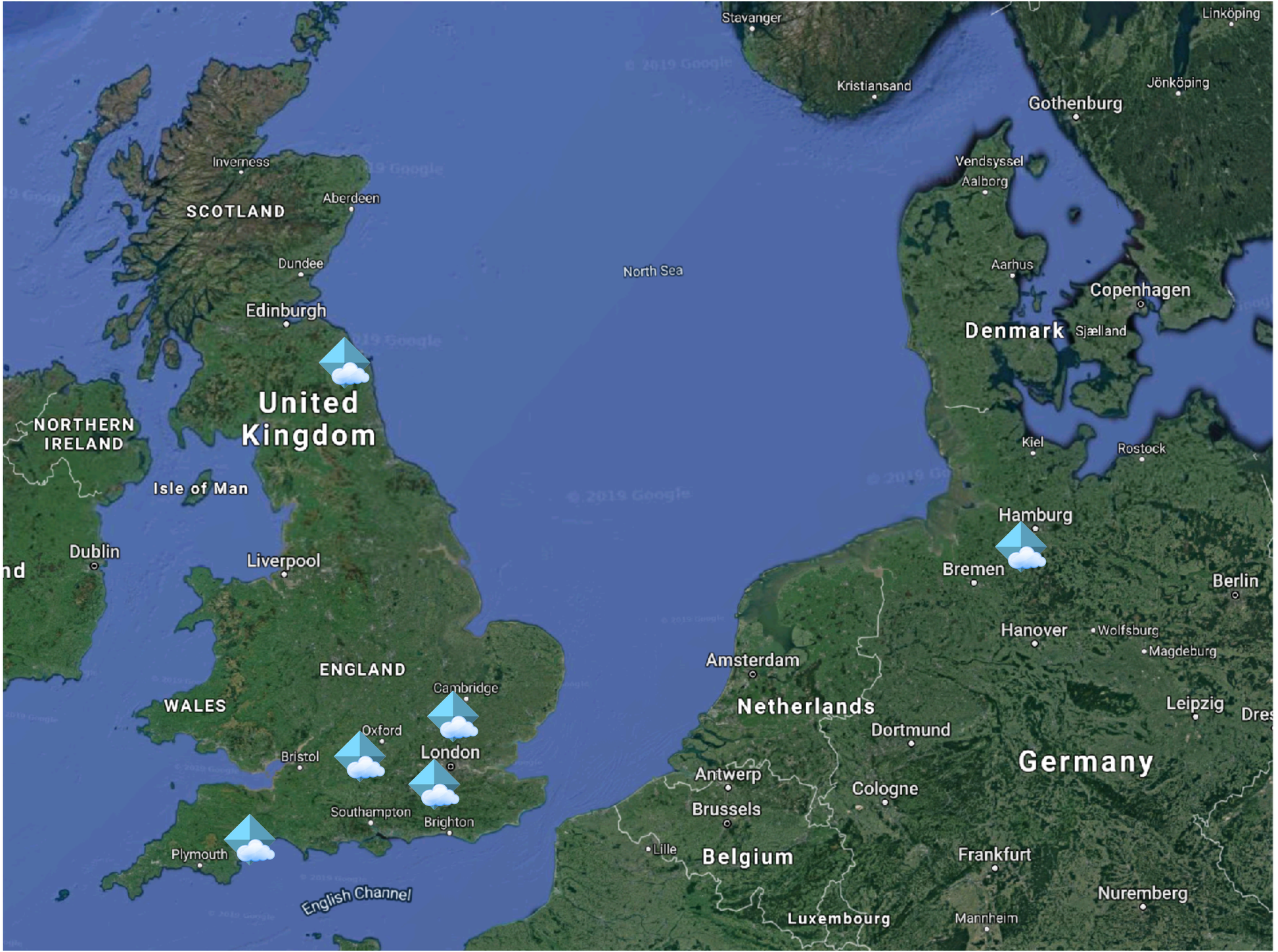


Rich graphical reports in-browser

On User Side



Embedded Molecular Graphics



Active CCP4 Cloud Instances

Public (global)

- CCP4-Harwell

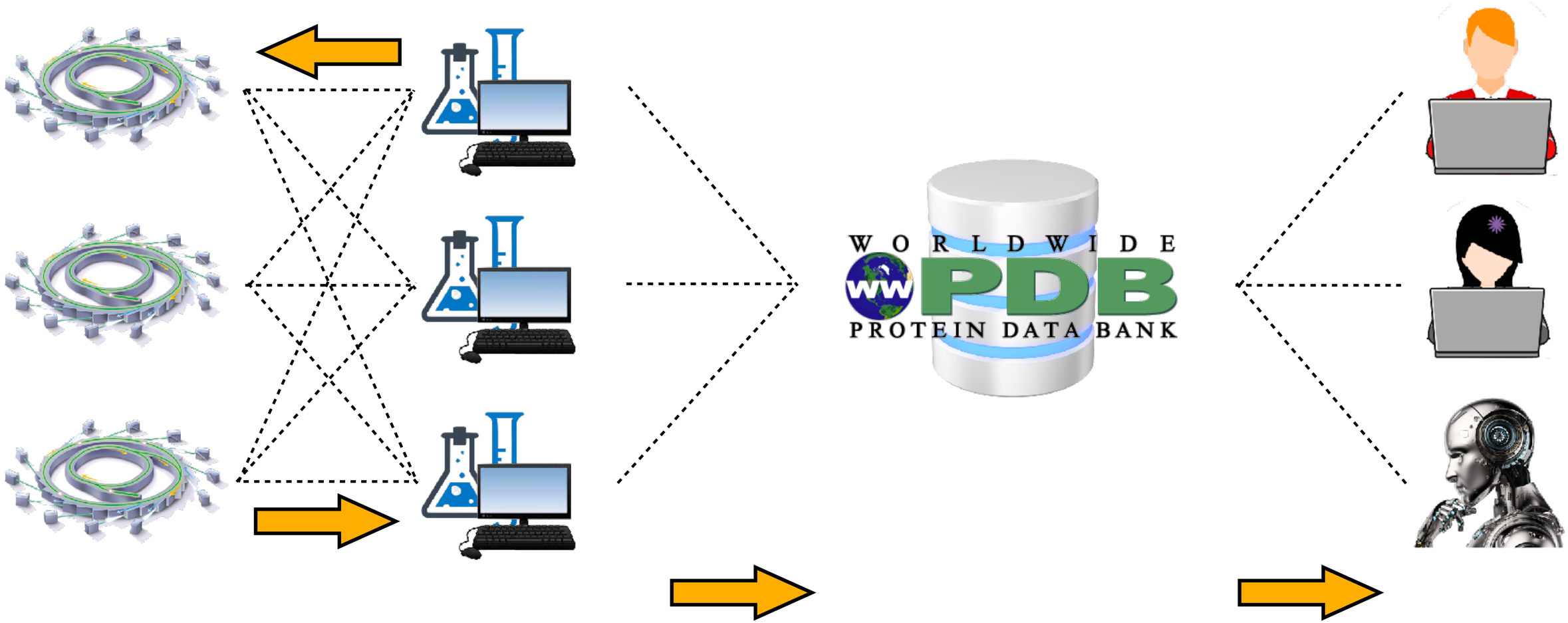
Institutional (local)

- Uni Exeter
- Francis Crick
- Uni Newcastle
- LMB Cambridge
- EMBL-Hamburg
- Incyte Inc. (USA)

Main Features

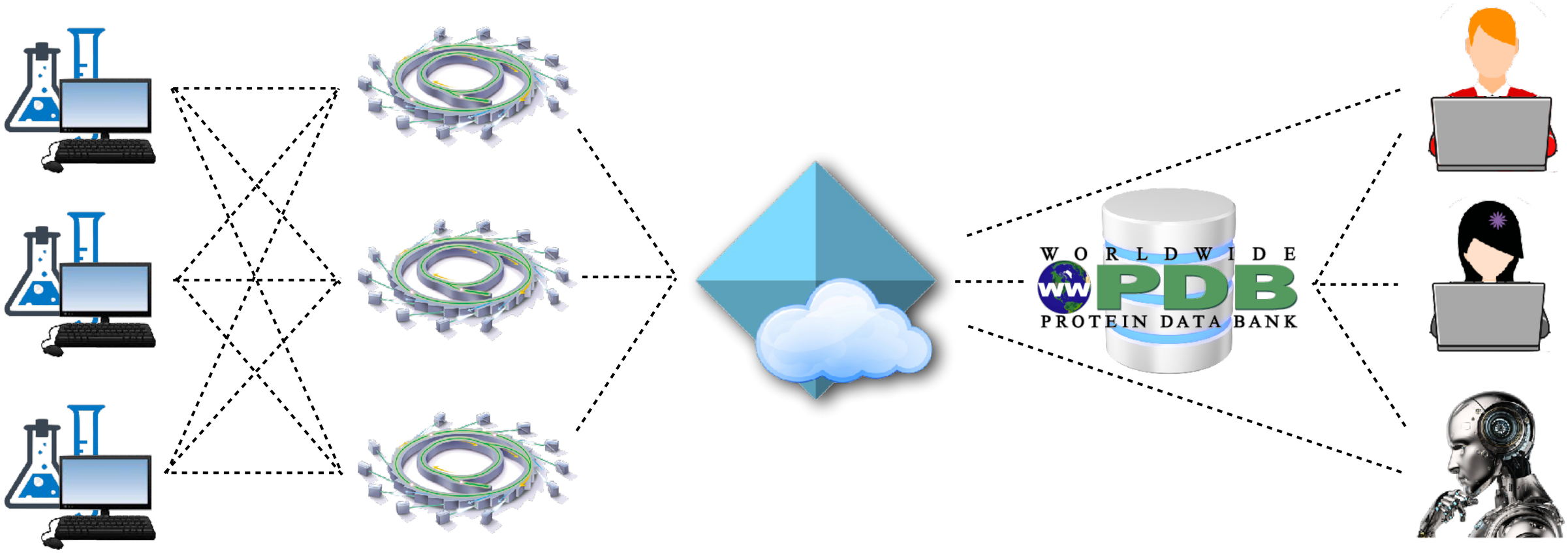
- Functionally complete
  - *includes all stages of structure solution*
- Designed to serve a number of users
  - *2,000 registered users*
  - *70,000 jobs/year at CCP4-Harwell*
- Designed to work with external data sources and services
  - *import from data facilities (currently iCAT, other in progress)*
  - *import from PDB and AFDB*
  - *integrated PDB deposition line*
- Designed to maintain structure solution projects
  - *backward compatible, versioned, metadating to ensure projects' integrity in future*
  - *all data is kept with projects*
  - *sharing projects in real time*

Data Production in Structural Biology



Imaging Facilities *Research Labs* *Data Bank and Server* *Users*

Data Production in Structural Biology



Research Labs → *Imaging Facilities* → *CCP4 Cloud* → *Data Bank and Server* → *Users*

Summary

- Data production in Science, in general, and Structure Biology, in particular, is taking industrial scales and approaches
  - *Data revolution, rise of AI*
  - *Automation of research and engineering*
  - *Approach to complex and fuzzy problems*
- Developing research infrastructures in form of centrally maintained data, software and computing services, linked to data producing and research centres, is a desirable way forward
  - *Efficiency through concentration and standardisation*
  - *Long-term data and software maintenance, knowledge keeping*
- Structural Biology infrastructure can be complemented with CCP4 Cloud
  - *Uniform approach to data management and structure solution across sites*
  - *Efficient data logistics*
  - *Retaining all data and metadata related to costly experiments*
- EGI is invited

Acknowledgements

CCP4, STFC & RCaH — *Fantastic work environment, support and dissemination*

CCP4 Collaboration, CCP4 School hosts and CCP4 developers — *Contribution of task reports, general support and valuable feedback*

~300 test users *Worldwide* — *Trial use and feedback on development versions of CCP4 Cloud*

Andy Purkiss *Francis Crick Institute, London*
Grzegorz Chojnowski *EMBL-Hamburg*
Arnaud Basle *Newcastle University*
Michael Isupov *University of Exeter*
— *Setup and maintenance of CCP4 Cloud instances in their home labs*

Biotechnology and Biological Sciences Research Council *(BBSRC) UK* — *Research grant BB/L0070317/1 (2014-2019)*

Developers and contributors

STFC, CCP4, Harwell, UK:
Andrey Lebedev, Oleg Kovalevskyi, Charles Ballard, Ville Uski, Ronan Keegan

RAS, Puschino, Russia:
Maria Fando

MRC/LMB, Cambridge, UK:
Robert Nicholls

EMBL-EBI, Hinxton, UK:
John Berrisford

Uni Leiden, The Netherlands:
Navraj Pannu, Pavol Skubak

Global Phasing Ltd, Cambridge, UK:
Marcin Wojdyr, Clemens Vonrhein

Uni York, UK:
Stuart McNicholas

Uni Liverpool, UK:
Adam Simpkin, Jens Thomas

Uni Birmingham, UK:
Christopher Oliver