EOSC-Life

Cloud deployment of FAIR resources: Success stories from the EOSC-LIFE RI community

PDB-REDO Cloud:

FAIR protein structures with deep versioning for scientific reproducibility and data provenance tracking

Maarten L. Hekkelman, Ida de Vries, Hans Wienk, Anastassis Perrakis & Robbie P. Joosten

Netherlands Cancer Institute and Oncode Institute

Using experimental protein structure models

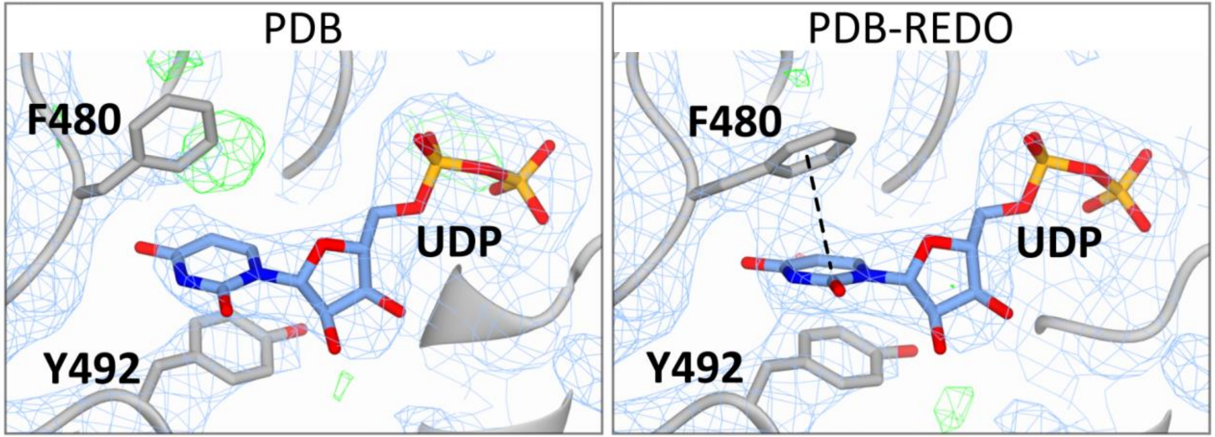
- Different types of structural biology research have different scale datasets:
  - The biology/biochemistry of a specific protein/complex (1-10 structures)
  - Drug discovery and homology studies (10-1000 structures)
  - Method development for crystallography or structure validation (1000-10,000 structures)
  - Machine learning and AI (>> 10,000 structures)
- The Protein Data Bank is a primary source of structural data, but has drawbacks
  - Created by different people, at different times – not methodological consistent
  - Are ‘never’ updated – not up-to-date with current methods
  - Too many have problems that can be solved – interpretation risk

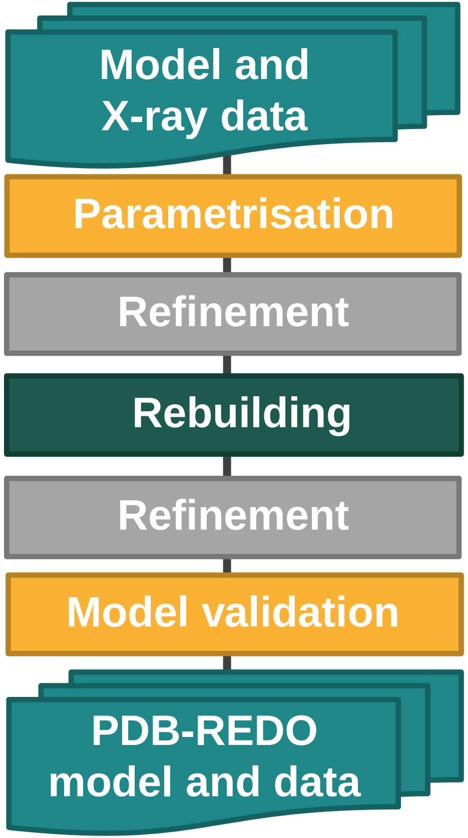
PROTEIN DATA BANK Years

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824087

Using PDB-REDO as alternative data source

- Uses PDB model + original experimental data to update PDB entries with the latest methods
- Fully automated (consistent) procedure for all models in PDB
- Generally improved model quality and fewer errors
- 155k entries with descriptive and model validation data available





Model and X-ray data

Parametrisation

Refinement

Rebuilding

Refinement

Model validation

PDB-REDO model and data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Using PDB-REDO as alternative data source

PDB-REDO is a living databank, it changes a lot:

- PDB depositions -> new PDB-REDO entries (hundreds weekly)
- PDB updates -> updated PDB-REDO entries (tens weekly, large batches sometimes)
- New PDB-REDO algorithms -> updated PDB-REDO entries (frequently)

Project challenges:

1. Make PDB-REDO databank FAIR in terms of data description
2. Improve the provenance tracking of PDB-REDO entries by ‘deep versioning’
3. Keep old PDB-REDO entries to allow scientific reproducibility
4. Allow better searching of the resource to create research data sets

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824087

Results

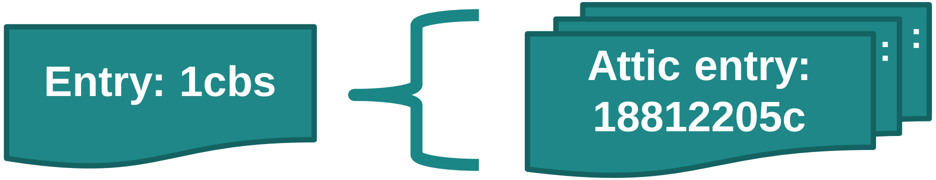
1. PDB-REDO databank is much FAIRer:
   - Performed integrity checks and cleaned up legacy data
   - Replaced proprietary (meta)data formats with JSON data (e.g. ligand validation data)
   - Added JSON schemas for all JSON data to describe contents and allow validation

2. Provenance of PDB-REDO entries is tracked in detail:
   - Document version of input data (PDB model revision, diffraction data revision)
   - Document manual edits of PDB data (needed when awaiting fix in PDB or PDB-REDO software)
   - Keep version numbers of all software in PDB-REDO pipeline (>60)
   - All version data captured in *versions.json*
   - Procedural metadata (i.e. PDB-REDO calculations with non-standard settings) is also captured

This project has received funding from the European Union's Horizon
2020 research and innovation programme under grant agreement No 824087

Results

3. Old PDB-REDO entries are rolled-over instead of overwritten:
   - Each entry has an ‘attic’ with previous versions
     - Keep final structure model, final electron density maps and versions.json
     - Other metadata (procedural, crystallographic, and validation data) stored in data.json
     - Each version has a unique, 9-character, persistent identifier based on the checksum of versions.json
   - Roll-over procedure part of main PDB-REDO pipeline to avoid accidental overwrites
     - Attic entry is also created for the current versions
   - If a PDB entry is obsoleted, PDB-REDO entry will also be obsoleted but attic kept
   - To do: make attic entries directly accessible through pdb-redo website



This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824087

Results

4. Databank searching:
   - All data.json and versions.json files indexed and stored in PostgresQL database automatically
   - Search API takes queries in JSON format
   - Queries can be combined as consecutive filters (implicit .AND.)

```
{"latest": true,
          "filters":[{"t": "sw","o": “ge","s": "pdb-redo","v": 7.21},
                     {"t": "d","o": "gt","s": "NBBFLIP","v": 0}]}
```

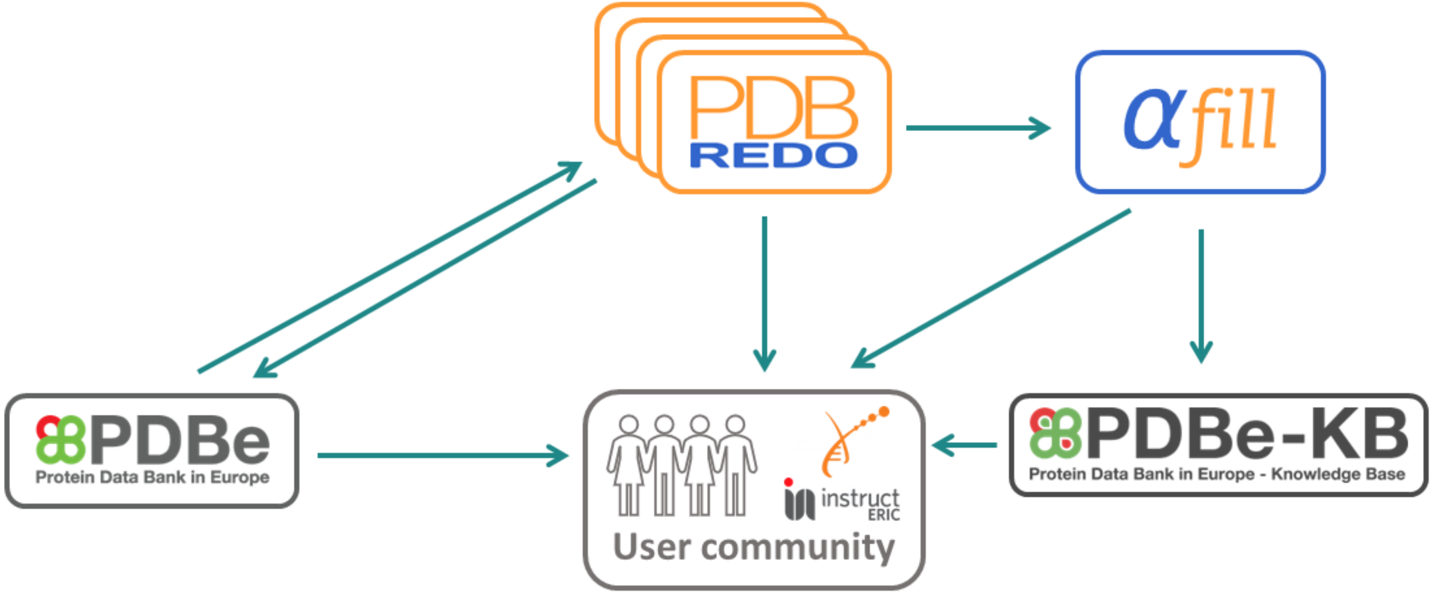
   - Return is a JSON array that lists the hits, can be used as a dataset description

```
[{"pdb-id": "1cbs", "version-hash": "18812205c" }, ... ]
```

   - GUI is being implemented on pdb-redo.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Impact of the Project on the Research Infrastructure(s)


PDB REDO
αfill
PDBe
Protein Data Bank in Europe
User community
instruct ERIC
PDBe-KB
Protein Data Bank in Europe - Knowledge Base

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Impact of the Project on EOSC

Brought access to the PDB-REDO structural databank to a new level of sophistication which will impact structural biology research in and out of EOSC for a long time to come

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Experience of working in EOSC-LIFE and technical teams

Happy with the outreach from the different work packages

FAIR hackathon was very interesting

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Future work/Sustainability of the project outcome

Future work:

- Document query API to allow easy access
- Connect PDB-REDO data to other resources (3D-Beacons)
- More outreach to reach the widest possible audience

Sustainability is well-covered:

- Updates with new data fully automated - low maintenance requirement
- Main developers on permanent contracts - manpower stays available
- Fully embedded in PDB-REDO research line - funding covered by new research

This project has received funding from the European Union's Horizon
2020 research and innovation programme under grant agreement No 824087

BBMRI-ERIC®	eatris	ECRIN EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK
elixir	EMBRC EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE	EMPHASIS
erinha European Research Infrastructure on Highly Pathogenic Agents	eu-openscreen	EURO BIOIMAGING
INFRAFRONTIER mouse disease models	instruct ERIC	ISBE Infrastructure for Systems Biology Europe
iNEXT DISCOVERY	MIRRI MICROBIAL RESOURCE RESEARCH INFRASTRUCTURE	NETHERLANDS CANCER INSTITUTE ANTONI VAN LEEUWENHOEK

Acknowledgements

Special thanks to:

- EMBL-EBI and PDBe developers
- Instruct-EOSC team

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087