The use of RDKit in the Protein Data Bank in Europe

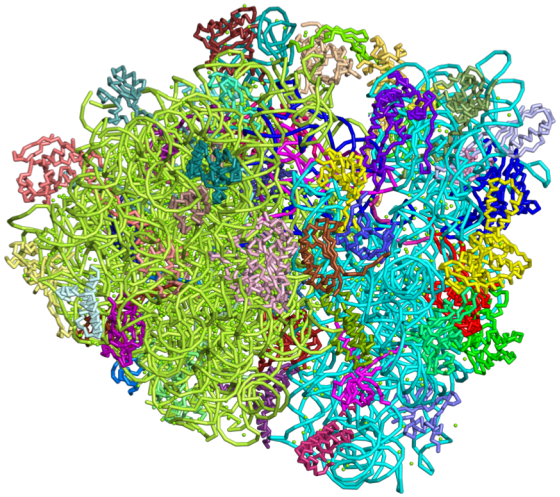
Lukas Pravda

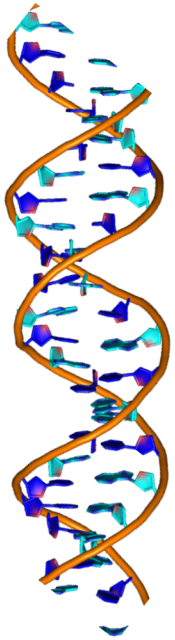
lpravda@ebi.ac.uk

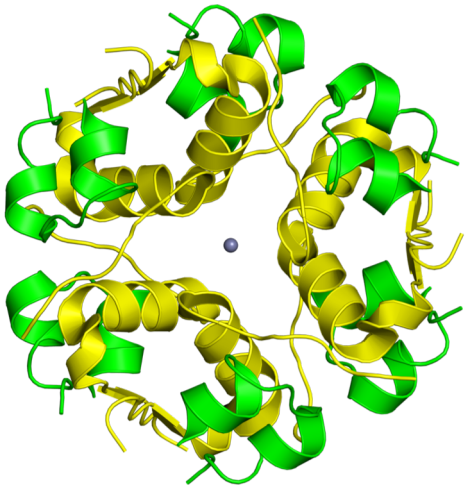
EMBL-EBI

What is the Protein Data Bank

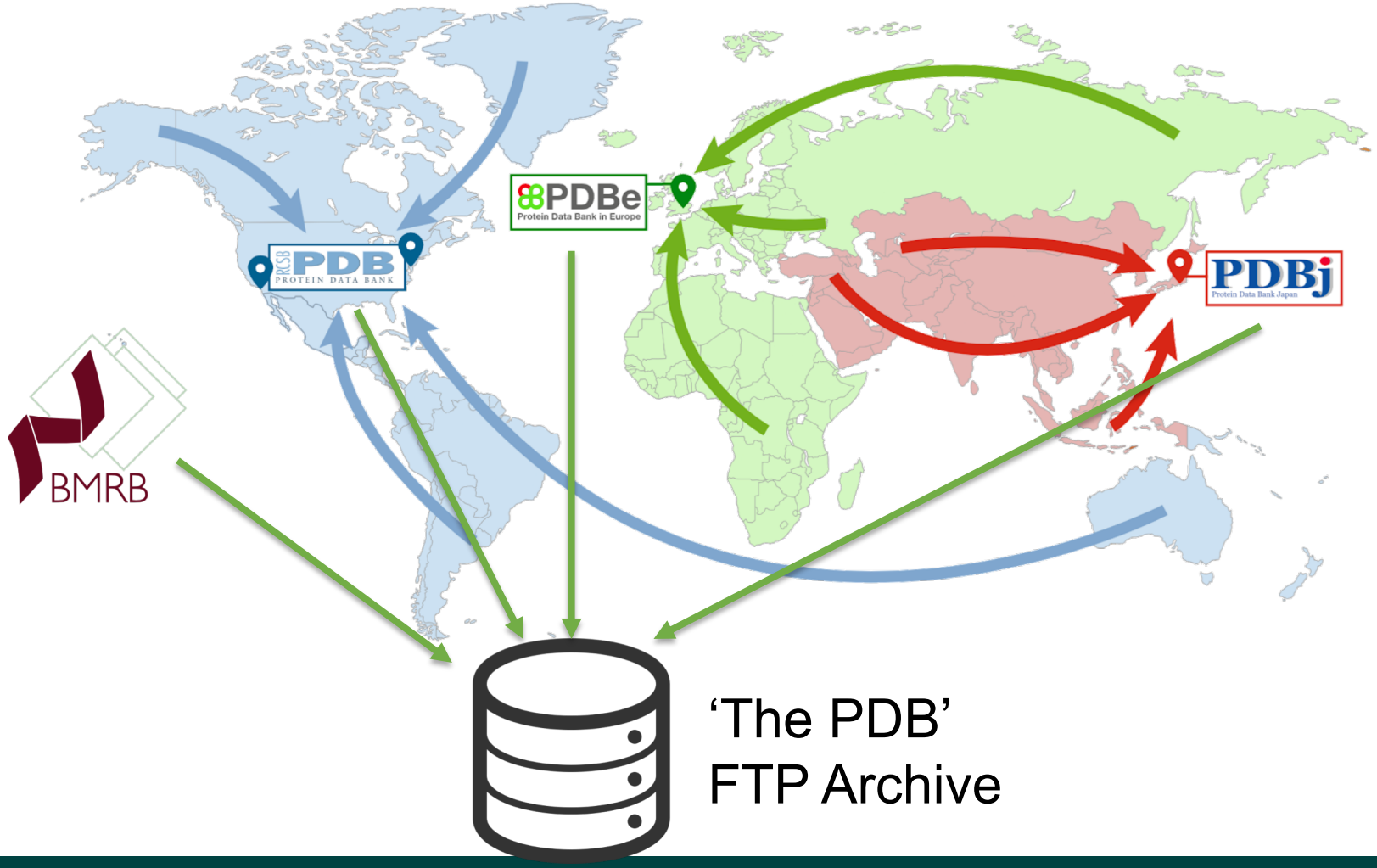
- An archive of experimentally determined 3-dimensional structures of biological macromolecules
- Protein, nucleic acids, sugars, complexes







WORLDWIDE PDB
PROTEIN DATA BANK



PDBe entry page

EMBL-EBI | Services | Research | Training | About us

Protein Data Bank in Europe
Bringing Structure to Biology

Search
Examples: hemoglobin, BRCA1_HUMAN
Advanced search
Feedback

PDBe › 1tbf

X-ray diffraction
1.3Å resolution

Catalytic Domain Of Human Phosphodiesterase 5A in Complex with Sildenafil

Source organism: *Homo sapiens*

Primary publication:

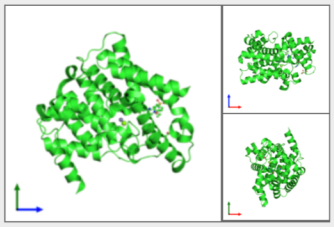
A glutamine switch mechanism for nucleotide selectivity by phosphodiesterases.

Zhang KY, Card GL, Suzuki Y, Artis DR, Fong D, Gillette S, Hsieh D, Neiman J, West BL, Zhang C, Milburn MV, Kim SH, Schlessinger J, Bollag G

Mol. Cell **15** 279-86 (2004)
PMID: 15260978

Released: 03 Aug 2004

Model geometry
Fit model/data



Quick links

- **1tbf overview**
  - Citations
- Structure analysis
- Function and Biology
- Ligands and Environments
- Experiments and Validation

View
Downloads
3D Visualisation

Function and Biology

Details

Reaction catalysed:

```
Guanosine 3',5'-cyclic phosphate + H(2)O = guanosine
5'-phosphate
```

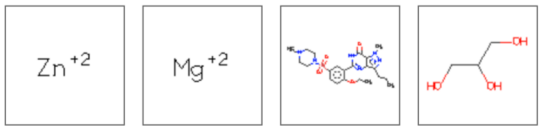
Biochemical function: 3',5'-cyclic-nucleotide phosphodiesterase activity
Biological process: signal transduction
Cellular component: not assigned

Sequence domains:

- 3'5'-cyclic nucleotide phosphodiesterase
- HD/PDEase domain
- 3'5'-cyclic nucleotide phosphodiesterase, catalytic domain
- Acetylglutamate kinase-like superfamily
- 3'5'-cyclic nucleotide phosphodiesterase, catalytic domain superfamily
- 3'5'-cyclic nucleotide phosphodiesterase, conserved site

Ligands and Environments

4 bound ligands:

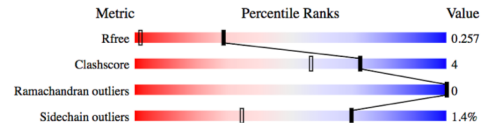


1 x ZN | 1 x MG | 1 x VIA | 1 x GOL

No modified residues

Experiments and Validation

Details



Citations

15 review citations

Guidelines for the successful generation of protein-ligand complex crystals.
Müller I. (2017)

14 more

3 mentions without citation

Surflex-Dock: Docking benchmarks and real-world application.
Spitzer et al. (2012)

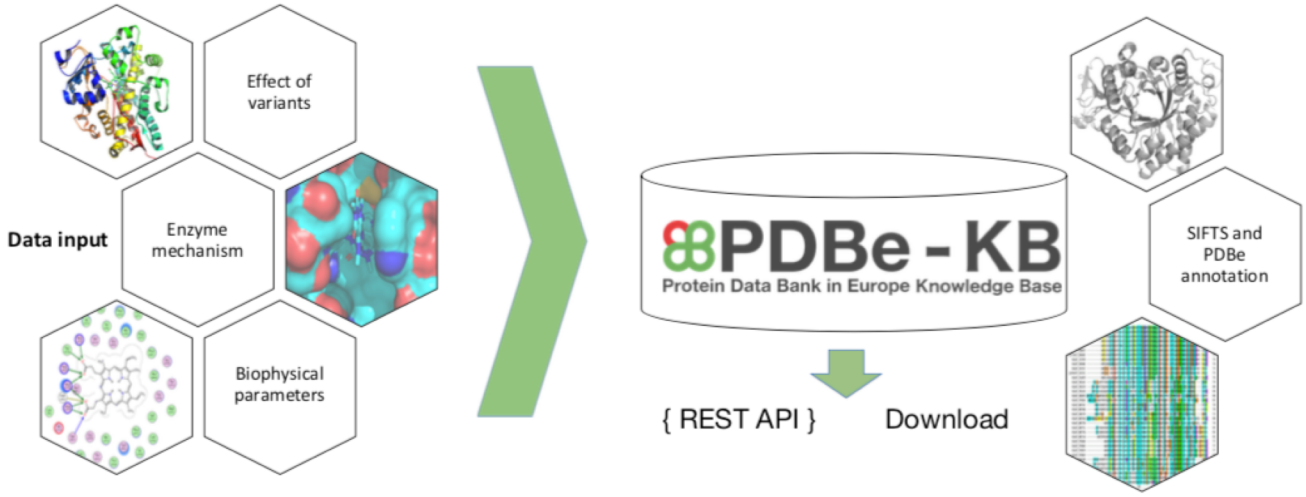
2 more

PDB_REDO

The sliders below show the change in model quality between original PDB entry and the

PDBe-Knowledge Base

- PDBe-KB (Protein Data Bank in Europe - Knowledge Base) is a community-driven resource managed by the PDBe.
- We collate functional annotations and predictions for structural data in the PDB archive to provide biological context.
- Collaborative effort with diverse research teams.

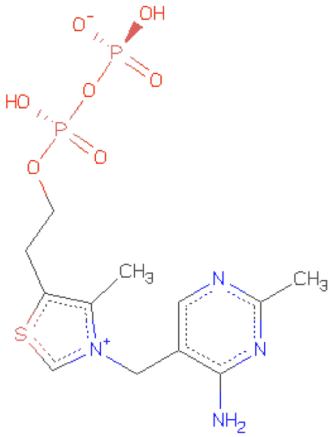


PDBe REST API

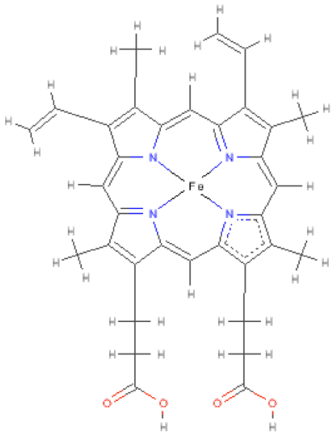
- A programmatic way to obtain information from the PDB archive.
- Details from the PDB archive and other resources are exposed. Data from the graph database from the PDBe-KB project are going to be public soon.
- Consumed by the PDBe webpages and services, so it ‘is’ fast and stable.
- Maintained, new calls added regularly.

Small molecules in the PDB

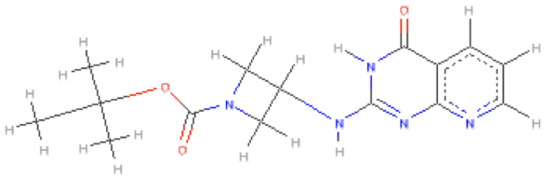
- Curated chemical component dictionary of compounds found in PDB entries (PDB-CCD).
- PDBeChem features:
  - Different chemical drawings
  - A few details (name, formula, inchi's, etc.)



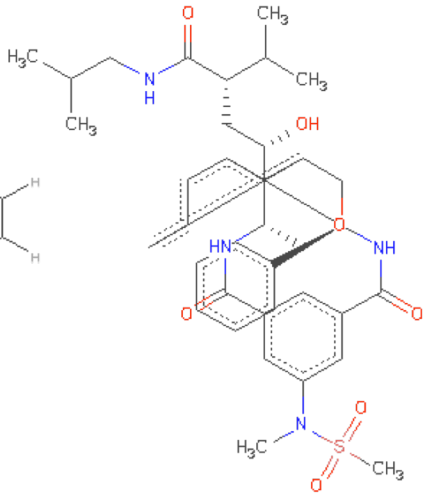
TDP



HEM



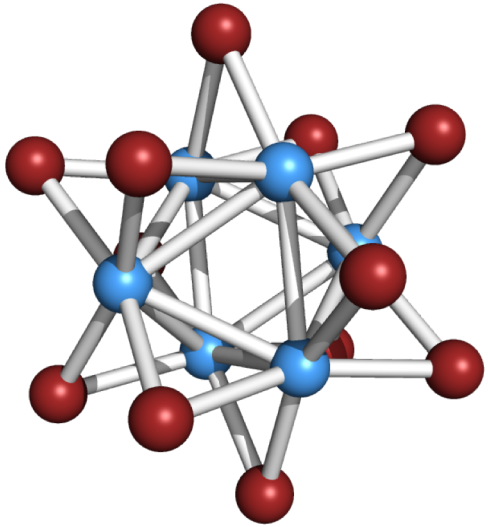
TB7



3OW

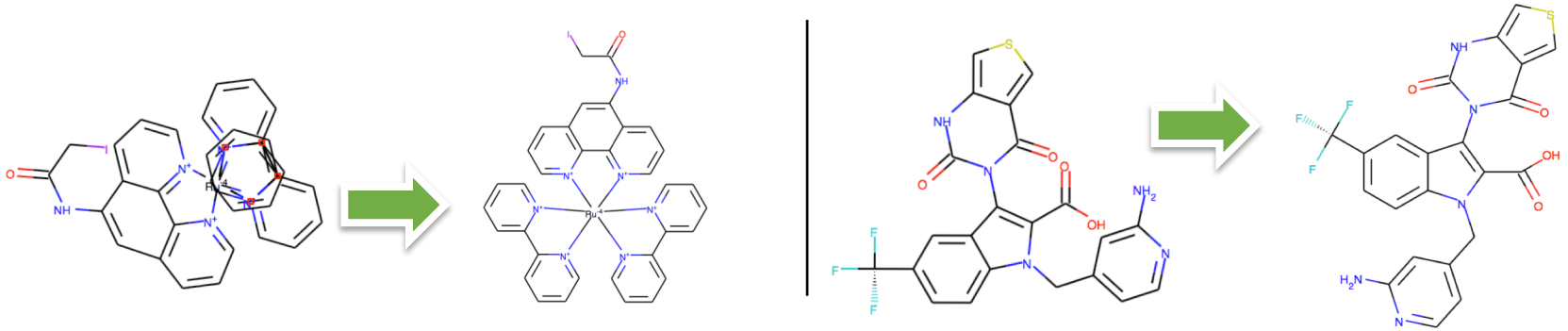
Sanitization process

- There are all sorts of ‘twisted’ chemistry in the CCD. e.g.
  - 5 valence carbons
  - 4 valence nitrogens
- RDKit standard error is read and parsed.
- Problematic atoms are attempted to get fixed.
  - E.g. 4 valence nitrogen gets positive charge and the bound metal gets negative charge
- Does not work all the time
- Modifies the original structure!



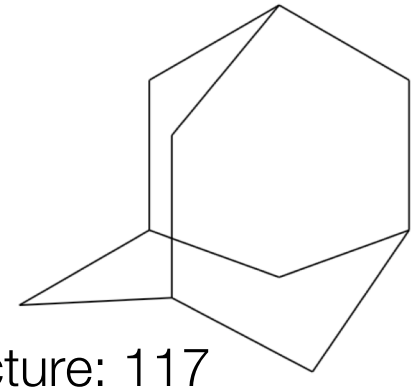
Images

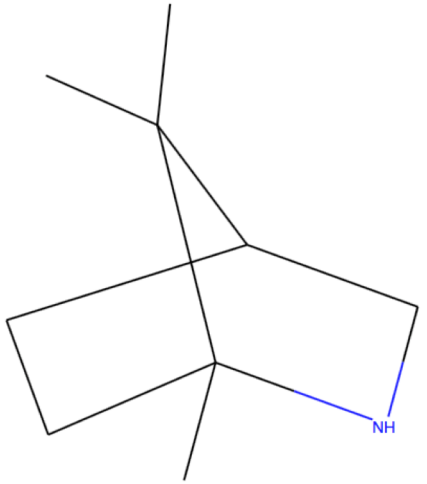
- Images are generated from 3 resources
  - 2D layouts from PubChem
  - Custom templates (porphyrin ring, ruthenium complexes)
  - Default RDKit functionality
- A simple scoring function was implemented to pick 'the prettiest' image (fewest collisions).



Statistics (September 2018)

- Ligands: 27,137
- Collision free depictions: 26,495
- Collisions
  - Total: 640 (< 2.5%)
  - Cannot be improved
    - Adamantane as substructure: 117
    - Nonborane as substructure: 59

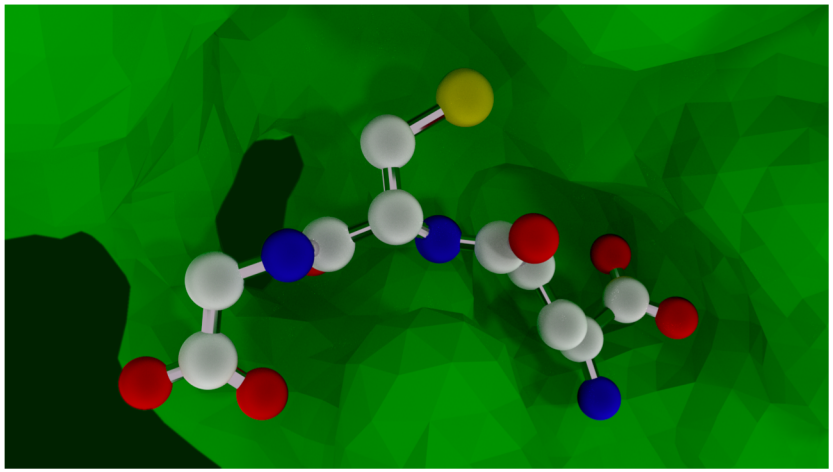




- Failures: 14 (*-decaboranes)
- A handful of ligands (< 10) were fixed and re-released based on the sanitization information from RDKit.

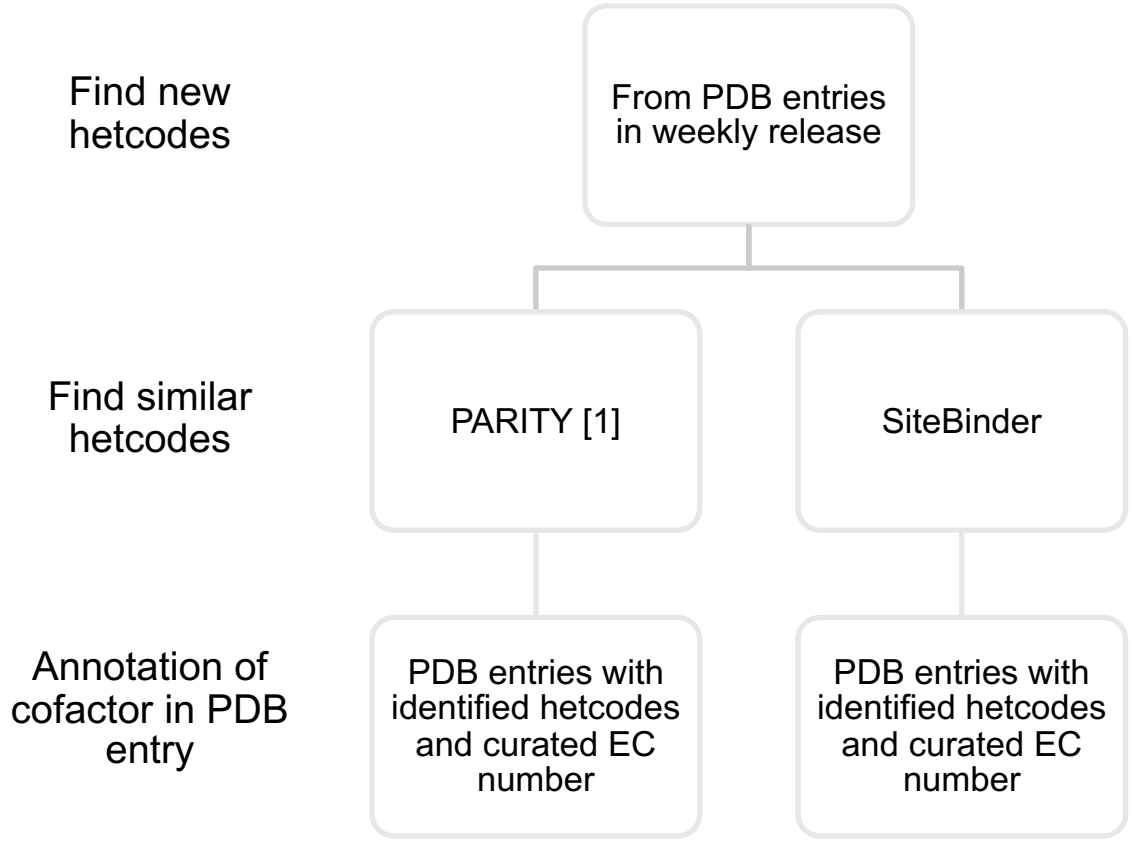
Cofactors in the PDBe

- Cofactors are small chemical species bound to protein structures and are required for a large body of enzyme activities.
- Basic annotation taken from the CoFactor database [1].



[1] Fischer, Julia D., Gemma L. Holliday, and Janet M. Thornton. "The CoFactor database: organic cofactors in enzyme catalysis." Bioinformatics 26.19 (2010): 2496-2497.

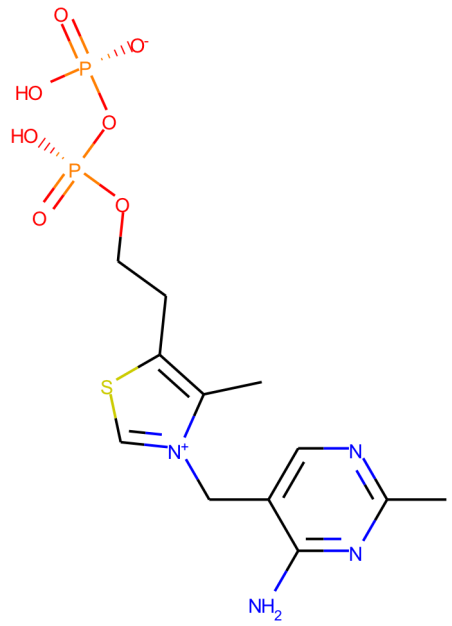
Keeping cofactors in the PDBe up-to-date

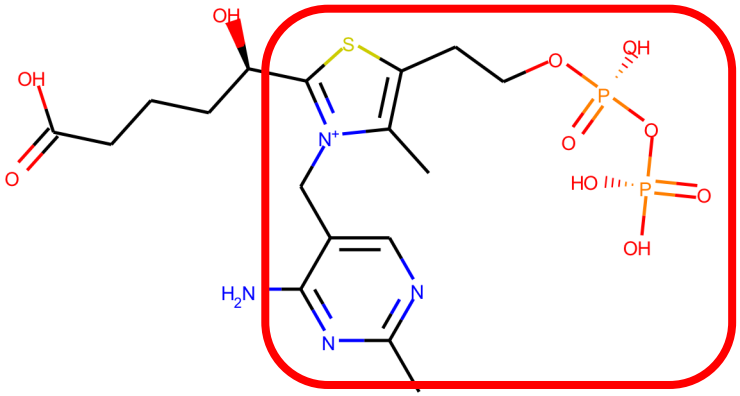


- Weekly updated and exposed over the PDBe REST API.

[1] Tyzack, Jonathan D., et al. "Ranking enzyme structures in the PDB by bound ligand similarity to biological substrates." Structure 26.4 (2018): 565-571.

Cofactors – example





TDP - thiamine diphosphate

TD8 – 0.8 similarity

Cofactors – statistics

- Original hand-curated dataset:
  - 27 classes
  - 54 cofactor mol.
  - 1575 enzymes
- Extended dataset with the use of RDKit:
  - 27 classes
  - 364 cofactor mol.
  - 3305 enzymes

3zhu › TD8

(5R)-5-{3-[(4-amino-2-methylpyrimidin-5-yl)methyl]-4-methyl-5-(2-{[(phosphonatooxy)phosphinato]oxy}ethyl)-1,3-thiazol-3-ium-2-yl}-5-hydroxypentanoate

Formula: $C_{17}\ H_{27}\ N_4\ O_{10}\ P_2\ S$

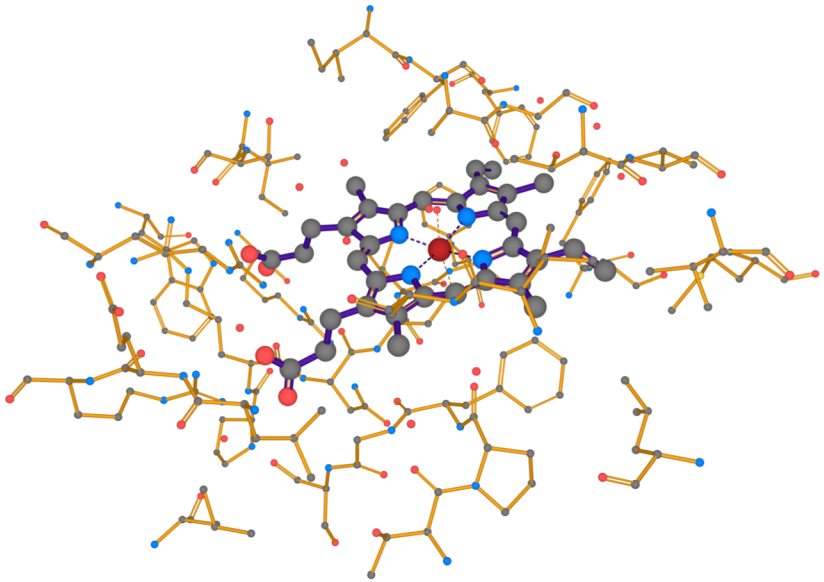
Molecular weight: 541 Da

Charge: 1

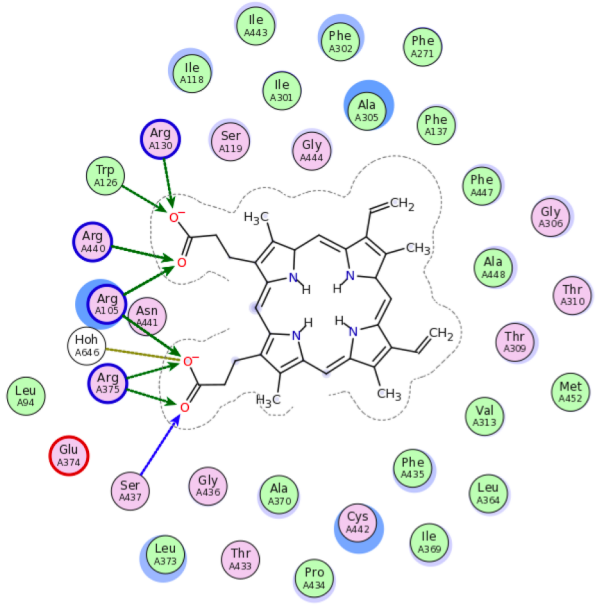
Putative function: Cofactor
Cofactor class: *Thiamine diphosphate*
Similarity to cofactor template (TDP): 0.8

Interactive viewer

- Interactive display of protein – ligand interactions in 2D.



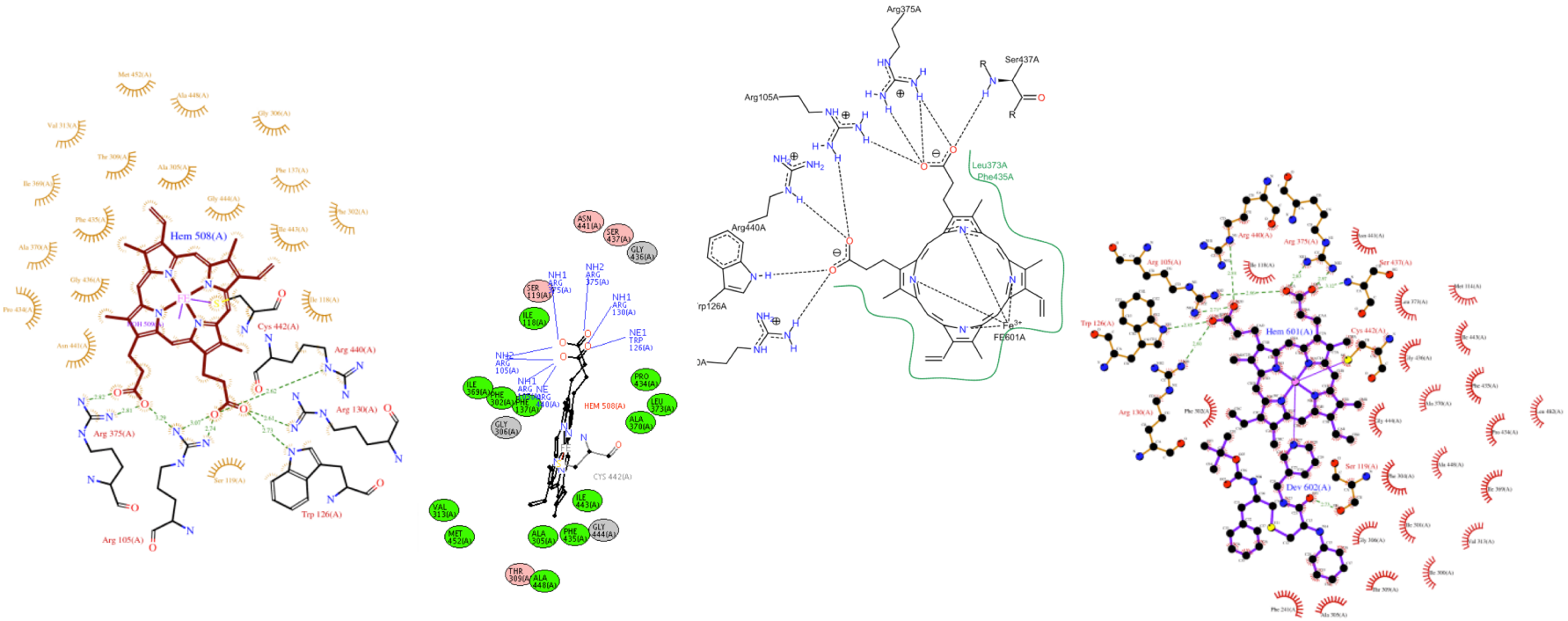
LiteMol | 3D



2D

Present tools

- Existing tools have limitations (cluttered visualization, cannot read mmCIF, license needed, closed source)
- Views cannot be compared



Procedure

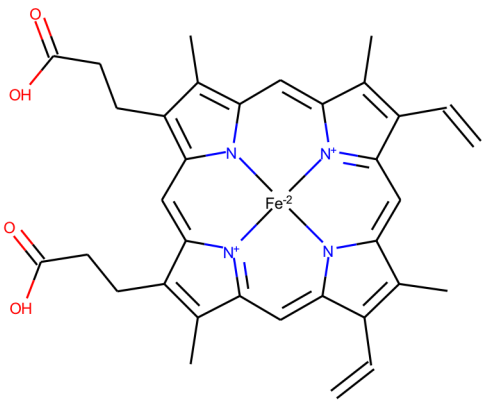
1. Display of preprocessed 2D depictions of ligands
2. Annotate it with the interactions by a modified version of Arpeggio [1] to get CREDO-style [2] annotations.
    1. Added support of mmCIF
    2. Fixed a couple of issues
    3. Refactored, modular, PIP installable.
    4. https://github.com/lpravda/arpeggio/tree/refactored
3. Interactively display with the use of D3.js according to the Clark and Labute 2007 [3].

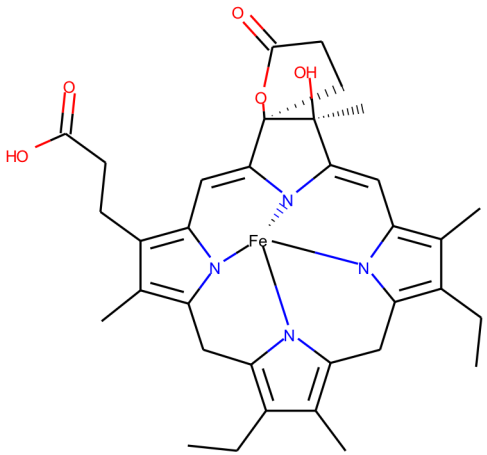
---

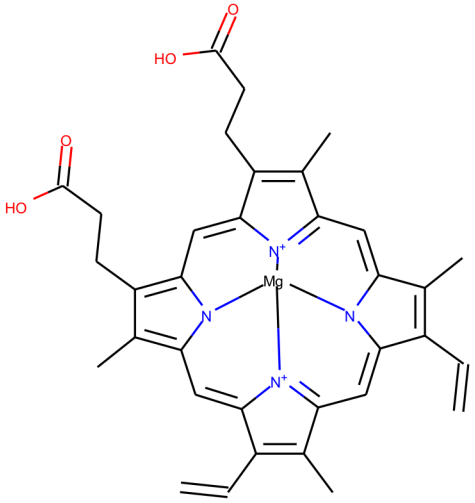
[1] Jubb, Harry C., et al. "Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures." Journal of molecular biology 429.3 (2017): 365-371.
[2] Schreyer, Adrian, and Tom Blundell. "CREDO: a protein–ligand interaction database for drug discovery." Chemical biology & drug design 73.2 (2009): 157-167.
[3] Clark, Alex M., and Paul Labute. "2D depiction of protein– ligand complexes." Journal of chemical information and modeling 47.5 (2007): 1933-1944.

Present problems to be addressed

- Sanitization still not ‘perfect’
  - e.g. carboranes.
- Depiction process does not sometimes pick the deemed template at all. Hopefully you can help me with that ☺.







pdbeccdutils

- All the pipelines and code is implemented in the toolkit called pdbeccdutils.
  - PDBeChem pipeline
  - protein-ligand interactions
  - pre-release ligand check
  - cofactors
- Python 3, open-source, PIP installable, documented, unit tested.
- mmCIF reader/writer

Conclusion

- RDKit is becoming increasingly important for the PDBe for our small molecule pipelines.
- A wrapper for RDKit called pdbeccdutils is developed.
- All the data calculated with the use of RDKit are/will be exposed either as flat files or via REST API.
- Web components to consume some of the data are being developed.

Acknowledgements

Sameer Velankar

Aleksandras Gutmanas

Oliver S. Smart

Abhik Mukhopadhyay

@PDBeurope

proteindatabank

EMBL

wellcome trust

BBSRC bioscience for the future