KATEDRA FYZIKÁLNÍ CHEMIE
UNIVERZITY PALACKÉHO V OLOMOUCI

INSTITUTE OF MOLECULAR AND TRANSLATIONAL MEDICINE

6th Advanced *in silico* Drug Design

KFC/ADD

Structure Bioinformatics Tools for Drug Design

Karel Berka

UP Olomouc, 30.1.-3.2. 2023

EMBL-EBI

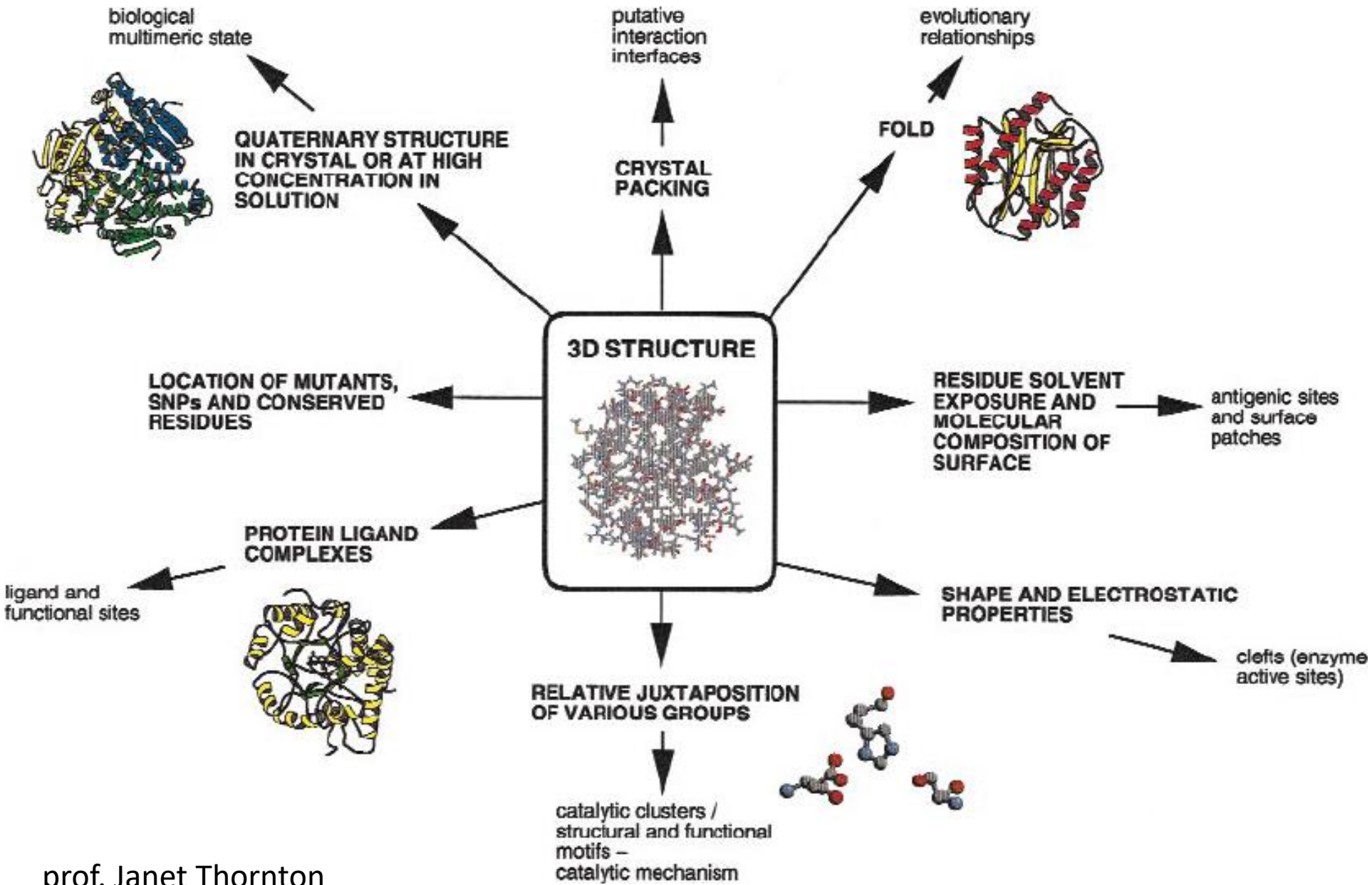
INSTITUTE OF PHYSICS
National academy of Sciences of Ukraine

ÚOCHB AV ČR
IOCB PRAGUE

Motto

Function of protein is given by its structure.

Use of Protein 3D Structure



prof. Janet Thornton

STRUCTURAL DATABASES

“Century of Life Science Data”

Nucleic acid sequences

European Nucleotide Archive
ArrayExpress
Expression Atlas
…

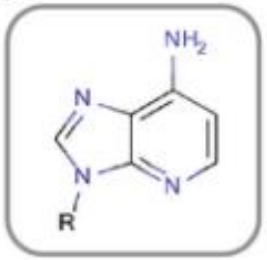
Protein sequences

Uniprot

Protein (biomacromolecule) structures

Protein Data Bank

Drug-like molecules



PubChem
Zinc
DrugBank
PDB CCD
ChEBI
ChEMBL
…

https://www.elixir-czech.cz/services

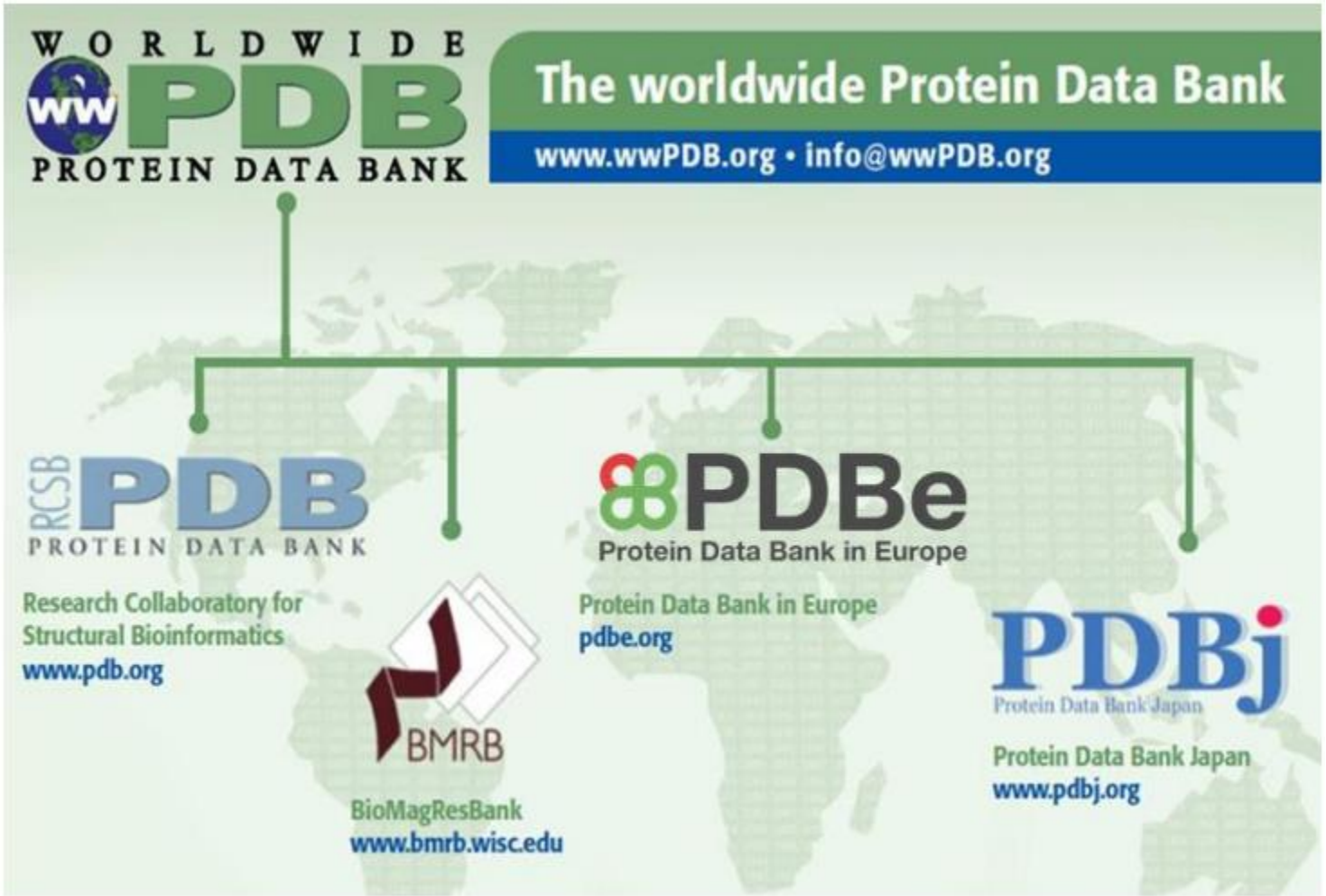
R. Svobodová-Vařeková – habilitation lecture 2016

Databases, you ask?…

CATH	http://www.biochem.ucl.ac.uk/bsm/cath/
FSSP	http://www2.ebi.ac.uk/dali/fssp/
Molecular Modeling Database	http://www.ncbi.nlm.nih.gov/Structure/
CAMPASS	http://www-cryst.bioc.cam.ac.uk/~campass/
ISSD	http://www.protein.bio.msu.su/issd/
Library of Protein Family Cores (LPFC)	http://WWW-SMI.Stanford.EDU/projects/helix/LPFC/
3D_ALI (a database of aligned protein structures and related sequences)	http://www.embl-heidelberg.de/argos/ali/ali_info.html
IDITIS (relational database and query tool for proteins)	http://www.oxmol.co.uk/prods/iditis/
HSSP	http://www.sander.embl-heidelberg.de/hssp/
Speciality databases	
HIV Protease Database	http://www-lbsc.ncifcrf.gov/HIVdb/
Nucleic Acid Database	http://ndbserver.rutgers.edu/
Prolysis (protease and protease inhibitor Web server)	http://delphi.phys.univ-tours.fr/Prolysis/
International Immunogenetics Database (IMGT)	http://imgt.cnusc.fr:8104/
Enzyme Structures Database	http://www.biochem.ucl.ac.uk/bsm/enzymes/
Features databases	
Molecular Movements Database	http://bioinfo.mbb.yale.edu/MolMovDB/
OLDERADO	http://neon.chem.le.ac.uk/olderado/
PROCAT	http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html
Protein Quaternary Structures (PQS)	http://pqs.ebi.ac.uk/
ReLIBase (receptor–ligand complexes database)	http://www2.ebi.ac.uk:8081/home.html
PROMISE	http://bioinf.leeds.ac.uk/promise/
PDBSum	http://www.biochem.ucl.ac.uk/bsm/pdbsum/
Biological Macromolecule Crystallization Database (BMCD)	http://h178133.nist.gov:4400/bmcd/bmcd.html
Resources	
Protein Data Bank	**htttp://www.rcsb.org/pdb/**

For more complete list of structural databases please refer to
http://www.oxfordjournals.org/our_journals/nar/database/subcat/4/14.

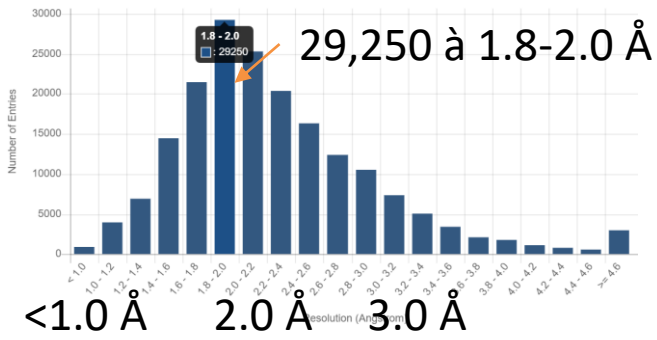
PDB – Protein Data Bank



> 120k biomacromolecular structures

PDB Statistics (30.1.2023)

- Organism:
  - Homo Sapiens 60,931
  - Mus musculus 8,426
- Enzymes:
  - Hydrolases (44,563)
  - Transferases (37,626)
- Residues
  - 100-300 (60138)
- Journal
  - To be published (31,250)
  - J. Biol. Chem. (13,460)
  - J. Mol. Biol. (11,385)
  - Biochemistry (10,883)
- R:



29,250 à 1.8-2.0 Å

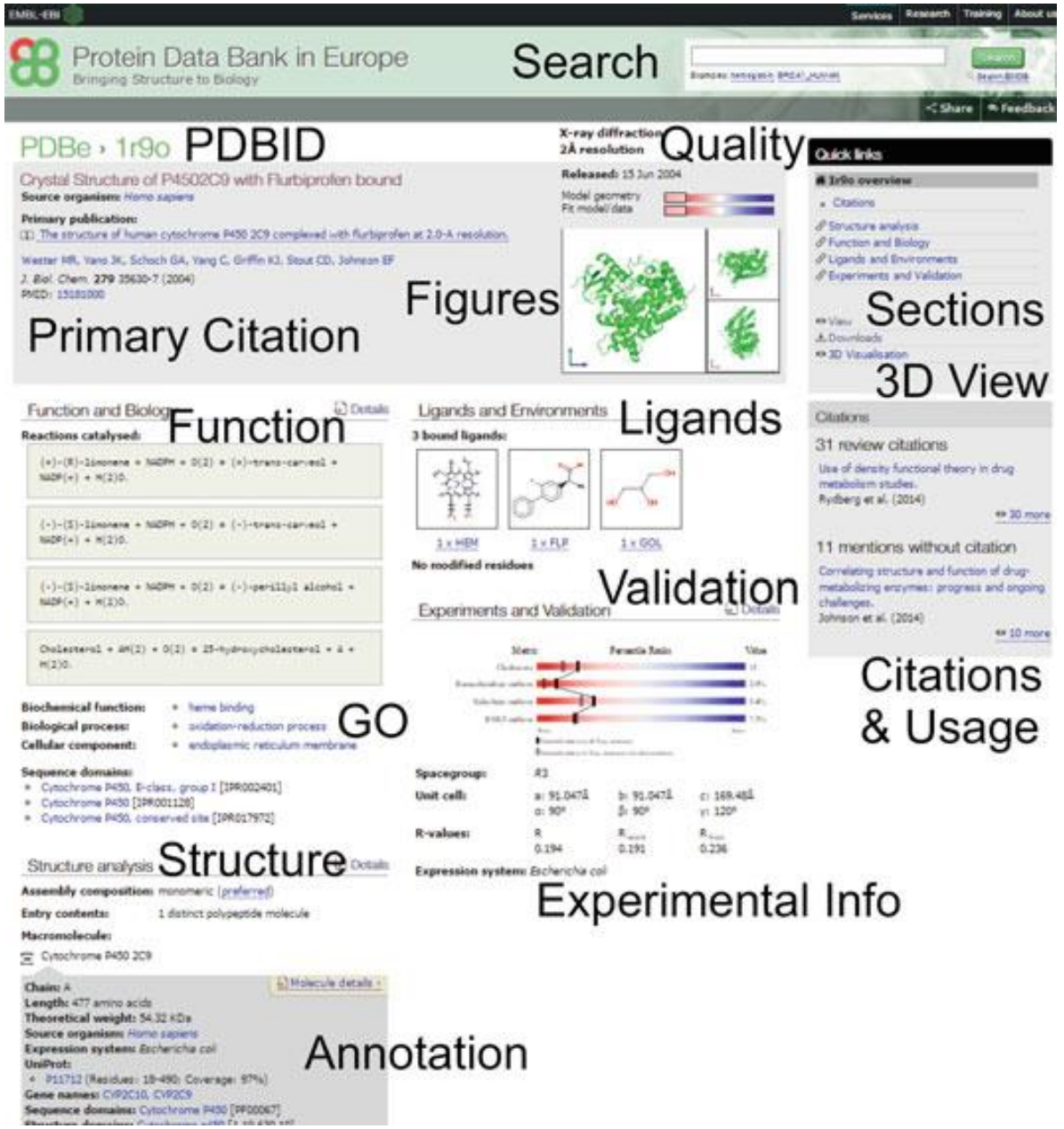
<1.0 Å 2.0 Å 3.0 Å

- http://www.pdb.org/pdb/statistics/

- Structures **200,708**
  - ❑ X-ray 172,458 (+5.7%)
  - ❑ NMR 13,899 (+2.2%)
  - ❑ EM 14,022 (**+29.3%**)
    - ➢ Proteins 174,424
    - ➢ NA 2,284D + 1,691R
    - ➢ Protein/NA 11,253
    - ➢ +Ligands 39,079

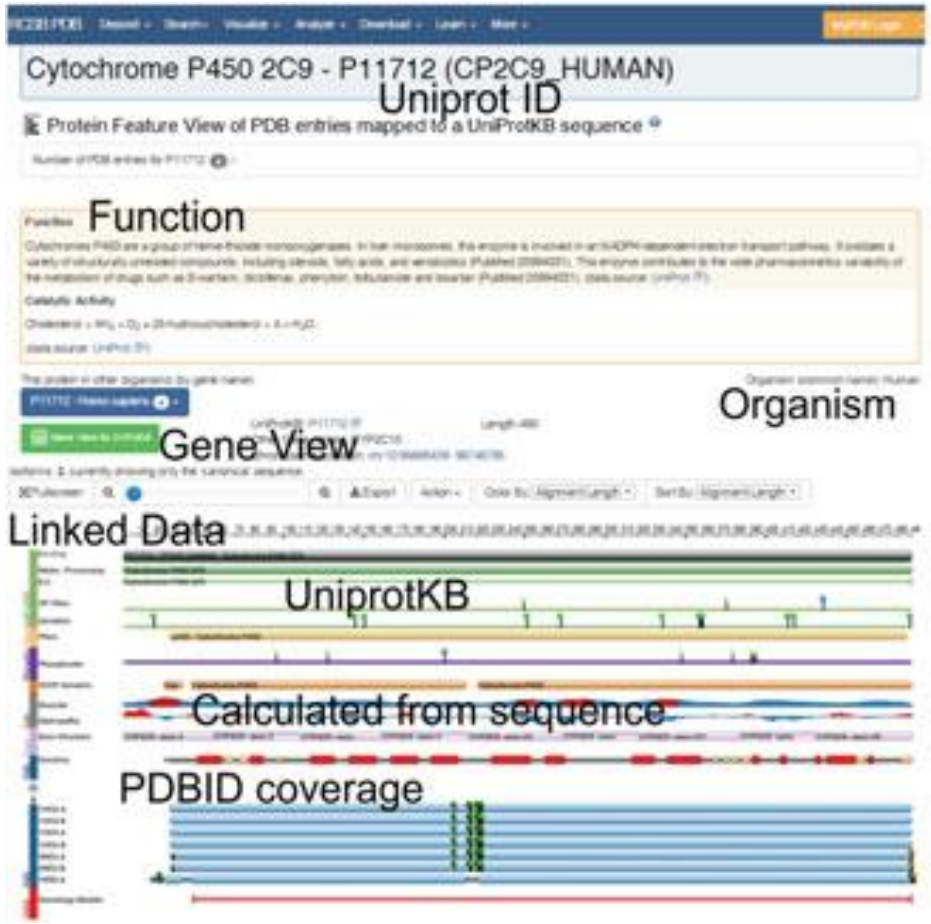
PDBe

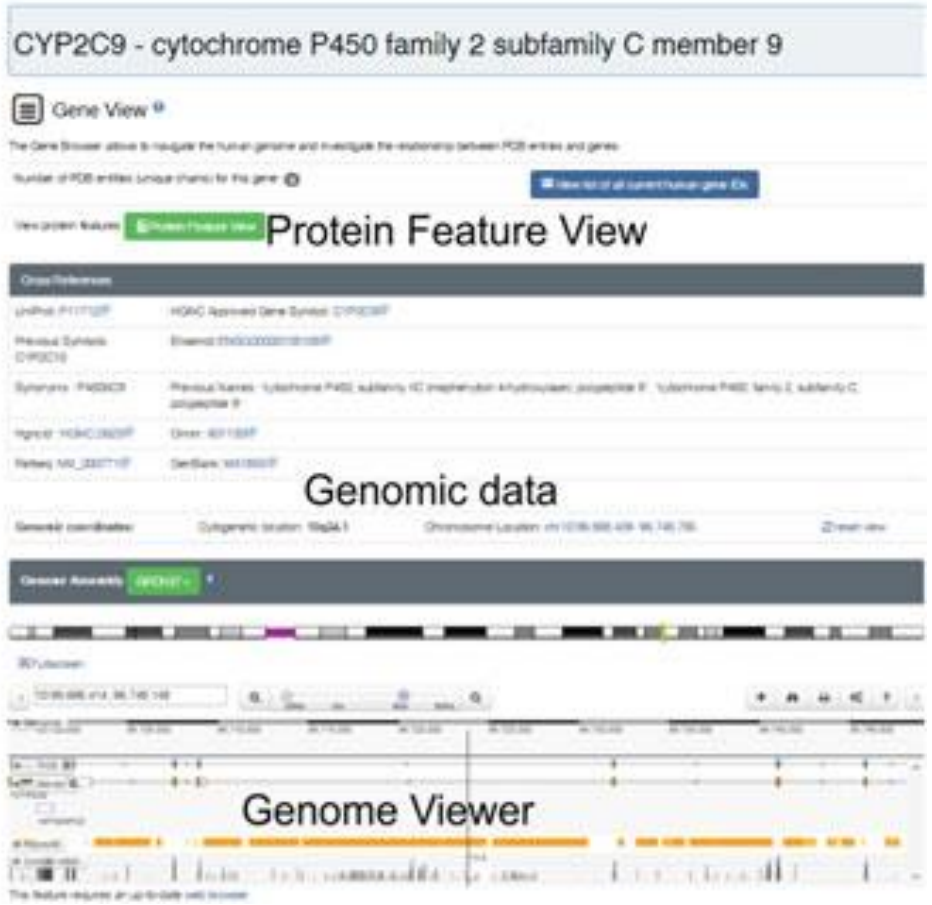
- PDB = Primary database
- Structures of proteins, nucleic acids, sugars, ligands,…
- Links to other databases (Uniprot)



J. Koča et al., *Structural Bioinformatics Tools for Drug Design*,
SpringerBriefs in Biochemistry and Molecular Biology,
DOI 10.1007/978-3-319-47388-8_3

RCSB PDB





J. Koča et al., *Structural Bioinformatics Tools for Drug Design*, SpringerBriefs in Biochemistry and Molecular Biology, DOI 10.1007/978-3-319-47388-8_3

Other Structural DBs

	Classification	
CATH	Domain classification of structures	http://www.cathdb.info/
Pfam	Classification of sequence families	http://pfam.xfam.org/
	Flexibility and disorder	
PDB FLEX	Intrinsic flexibility in proteins	http://pdbflex.org/
DisProt	Database of Protein Disorder	http://www.disprot.org/
Pocketome	Encyclopedia of ensembles of druggable binding sites	http://www.pocketome.org/
	Detailed biomacromolecules	
OPM	Orientations of proteins in membranes	http://opm.phar.umich.edu/
NDB	Nucleic Acids Database	http://ndbserver.rutgers.edu
GFDB	Glycan Fragment Database	http://www.glycanstructure.org/

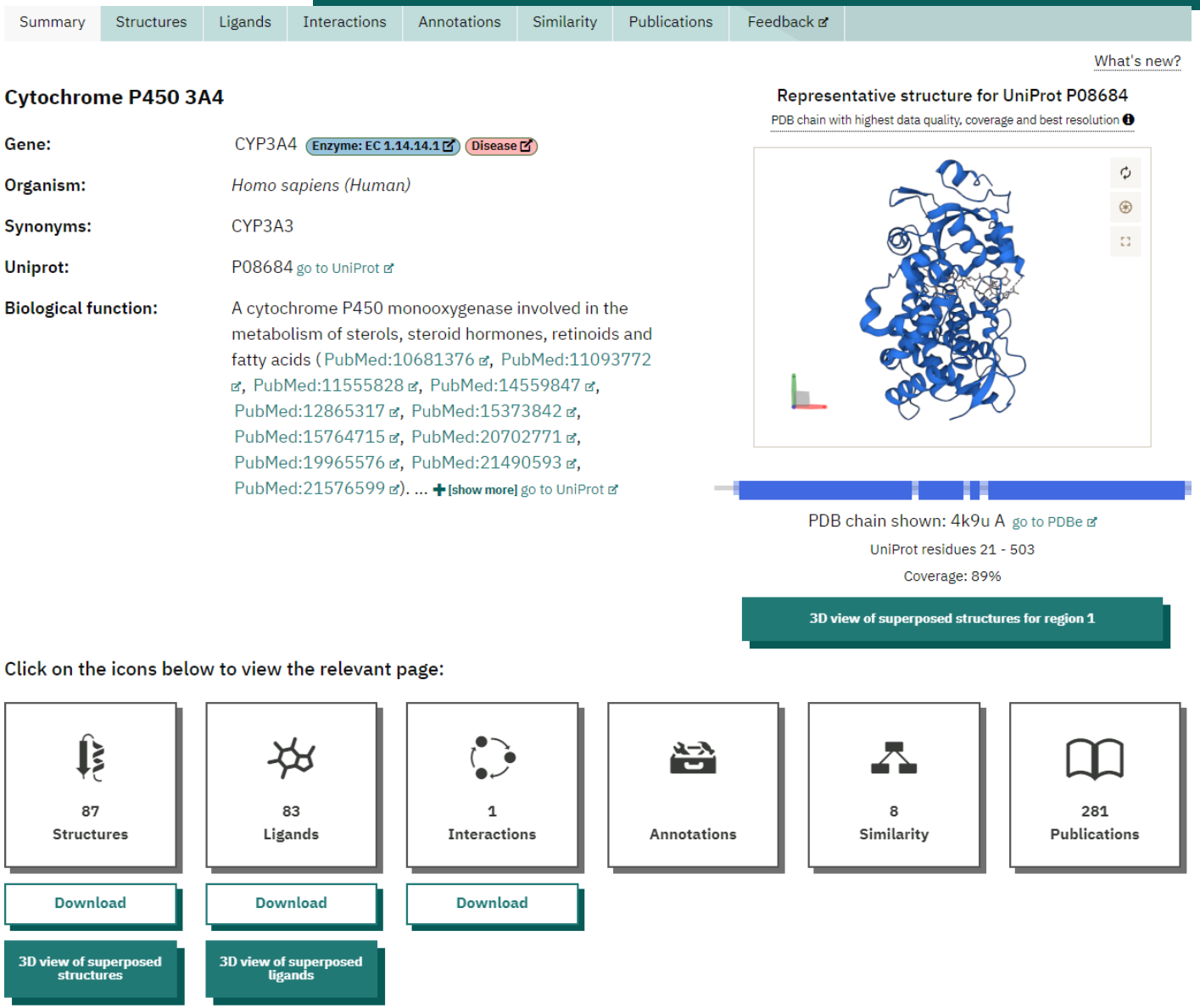
For more complete list of structural databases please refer to **http://www.oxfordjournals.org/our_journals/nar/database/subcat/4/14.**

J. Koča et al., *Structural Bioinformatics Tools for Drug Design*, SpringerBriefs in Biochemistry and Molecular Biology, DOI 10.1007/978-3-319-47388-8_3

PDBe-KB

Protein Data Bank in Europe Knowledge Base

Aggregated Views of Proteins



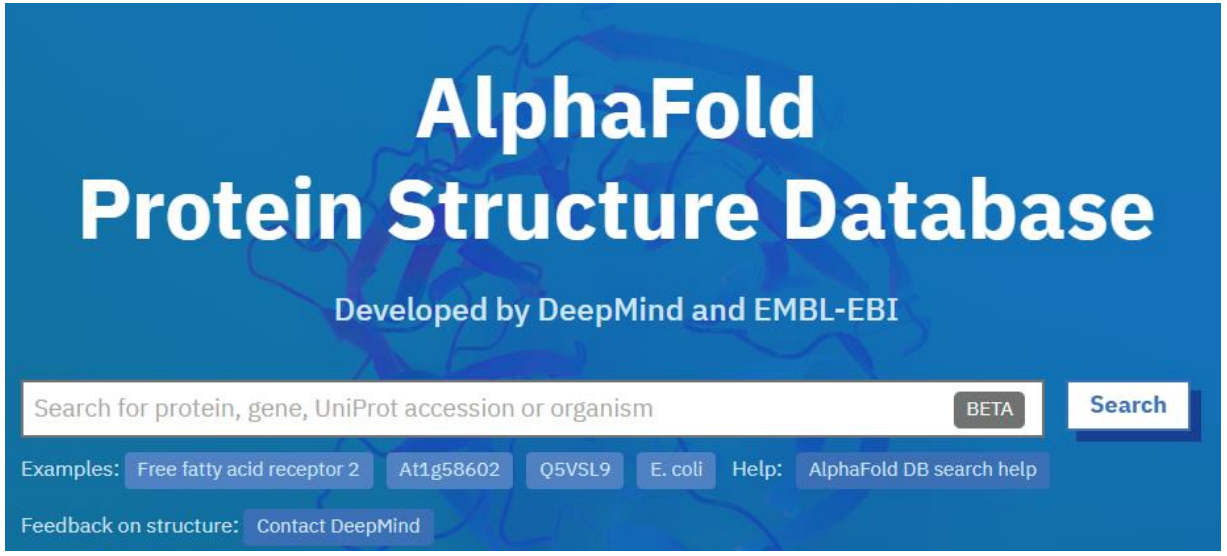
- Clusters by UNIPROT IDs
  - Structures
  - Ligands
  - Interactions
  - Annotations
  - Similarity
  - Publications

https://www.ebi.ac.uk/pdbe/pdbe-kb

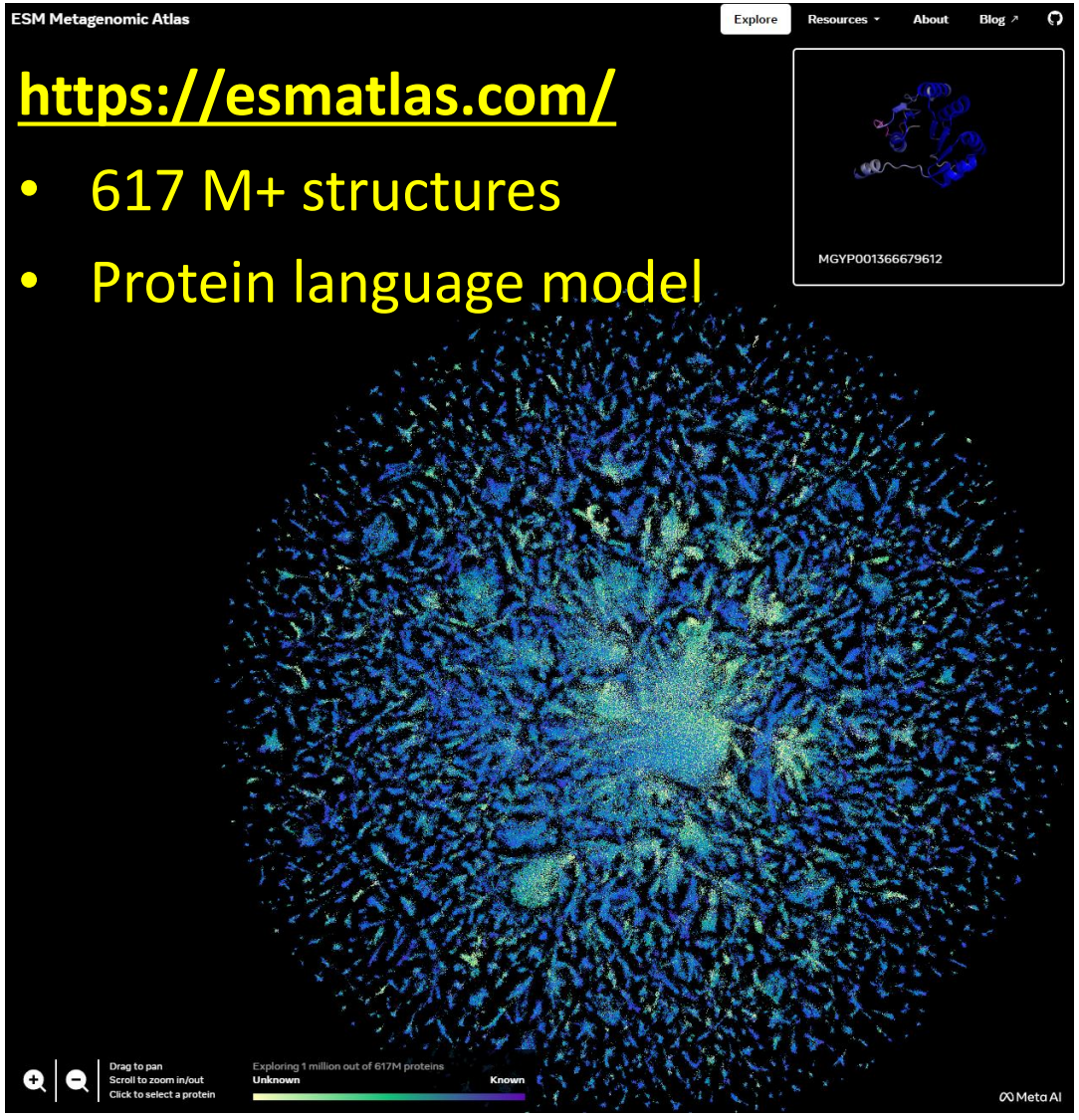
AlphaFoldDB + ESM Metagenomic Atlas

https://www.alphafold.ebi.ac.uk/

- 200 M+ computed structures



- 16 model organisms (e.g. Homo Sapiens)
- 33 infective organisms (e.g. M. tuberculosis)
- 500k SwissProt, 17k MANE

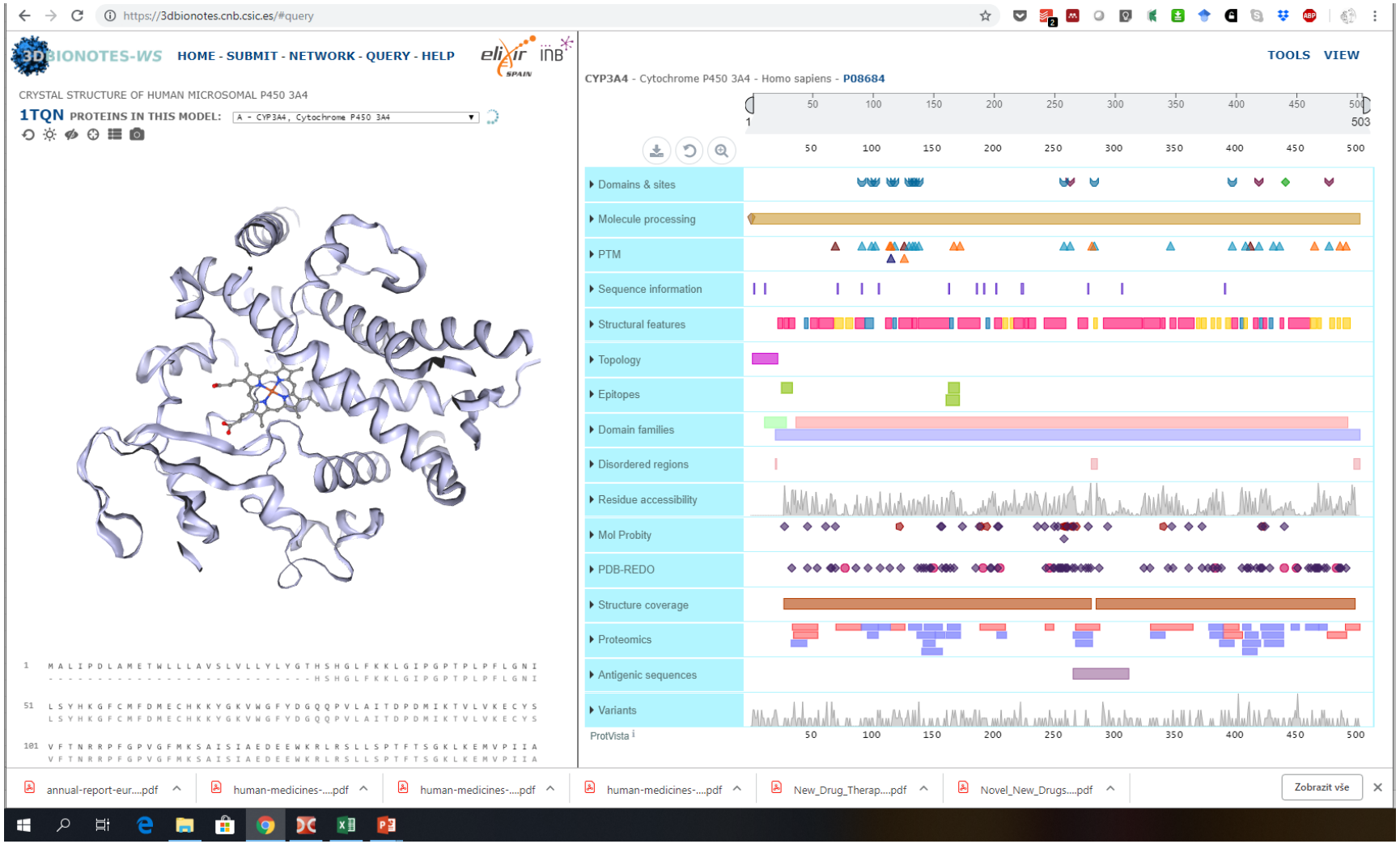


- 617 M+ structures
- Protein language model

3DBIONOTES-WS

West-Life elixir SPAIN INB

see annotations on structure



https://3dbionotes.cnb.csic.es/ws/database

PROTEIN FUNCTION DETECTION

Function Definition

EC – enzyme classification

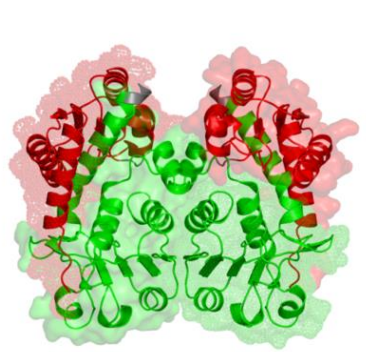
(4 levels of detail)

- Oxidoreductases (EC 1)
  - substrate is a donor of hydrogen or electrons
    - what substrate
      - acceptor-type
- only enzymes
- 1 enzyme – more functions
  - methylenetetrahydrofolate dehydrogenase/cyclohydrolase (EC 1.5.1.5 and EC 3.5.4.9)

GO - Gene Ontology

genome annotation – more generic

- **Biological process**
  - generic (cell growth)
  - specific (glycolysis)
- **Molecular function**
  - generic (enzyme)
  - specific (hexokinase)
- **Cell compartmentalisation**
  - where it works (ER, cytosol)



catalytic domain

NADPH-binding domain

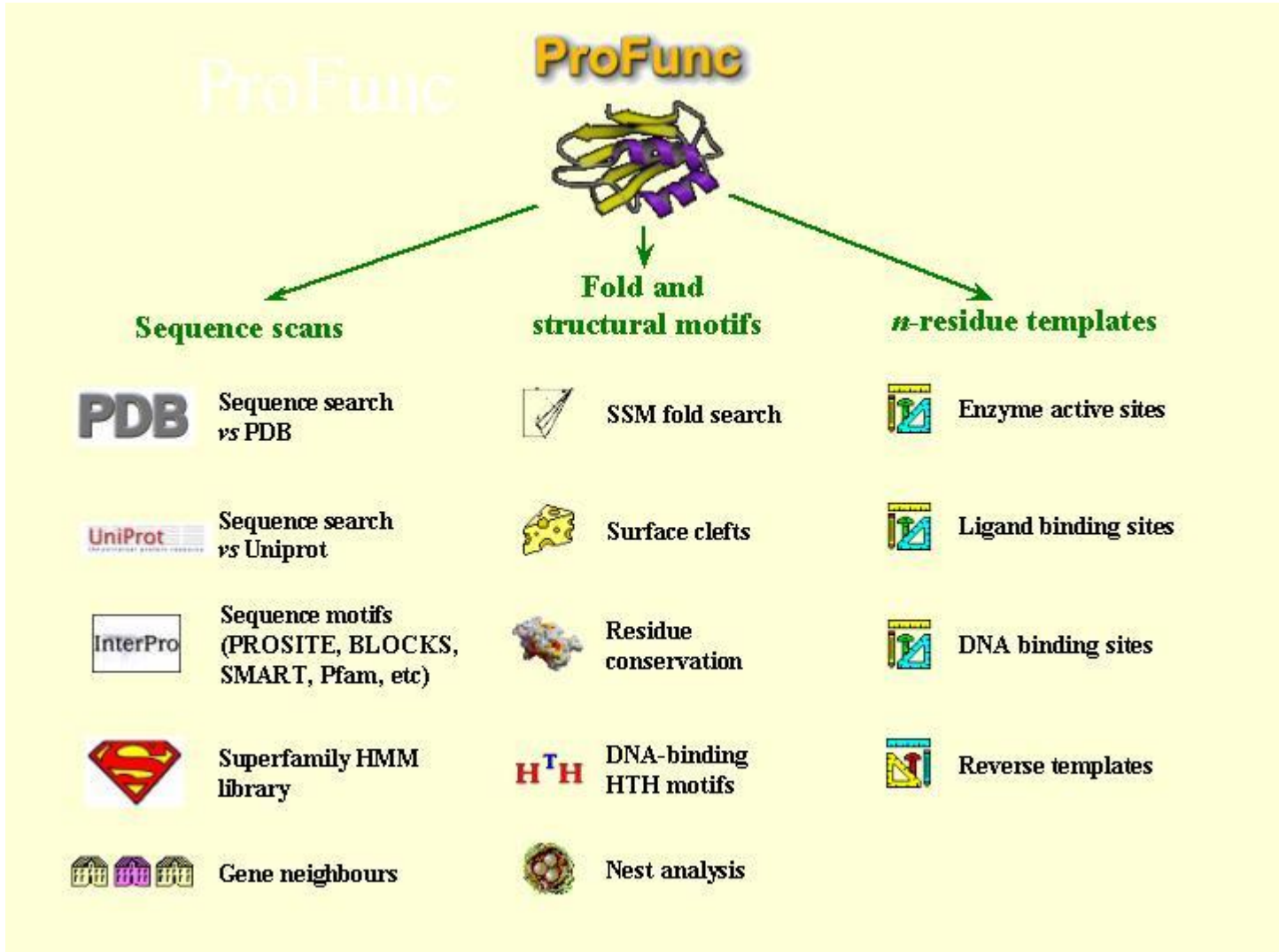
Function Search - Possibilities

- Sequence
  - below 40% identity is function less conserved (exception CYP)
- Fold matching
  - structure alignment (DALI, FATCAT) against annotated structural DBs (CATH, SCOP)
- 3D motives
  - active site position (TESS, PDBSiteScan, PatternQuery)
- Active sites
  - geometric similarity (SURFNET)
  - physical properties (SURF'S UP!, SiteEngine)
  - ! same substrate can have multiple sites – ATP

Function Search – Possibilities II

- Phylogenetic analysis
  - search for conserved residues by multiple alignement (ConSurf)
- DNA binding
  - Helix-turn-helix motiv (HTHquery)
- Ligands
  - similar sites as in other crystals in PDB (PDBsum, MSDsite)
- Channel anotation
  - Towards active site – substrate channelling (Caver, MOLE)

Function Search - ProFunc

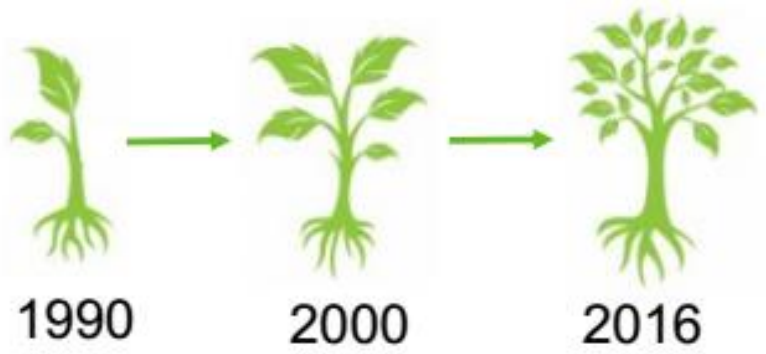


http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/

PATTERNS

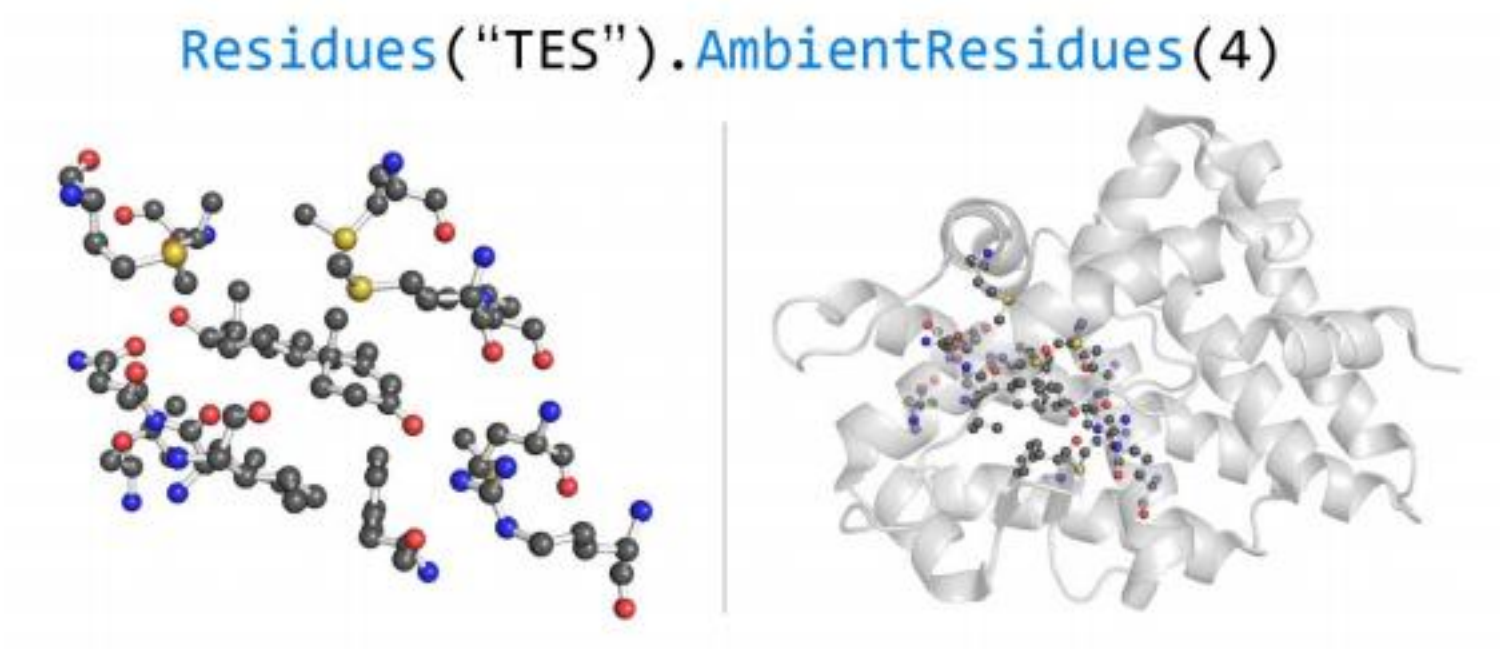
Pattern Detection: Why?

Year	1990	2000	2016
Testosterone binding proteins	0	4	72
Cytochrome P450	1	22	121
Lectine LecB	0	9	43
Apoptosis regulators BAK a BAX	0	3	190
Methemoglobine	4	11	37
Can we perform a comparison / analysis?	No	?	Yes



R. Svobodová-Vařeková – habilitation lecture 2016

Pattern Detection: Approaches



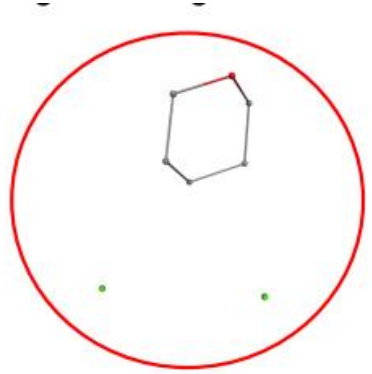
- Software: OpenStructure, Protein segment finder, PatternQuery*

https://webchem.ncbr.muni.cz/Platform/PatternQuery

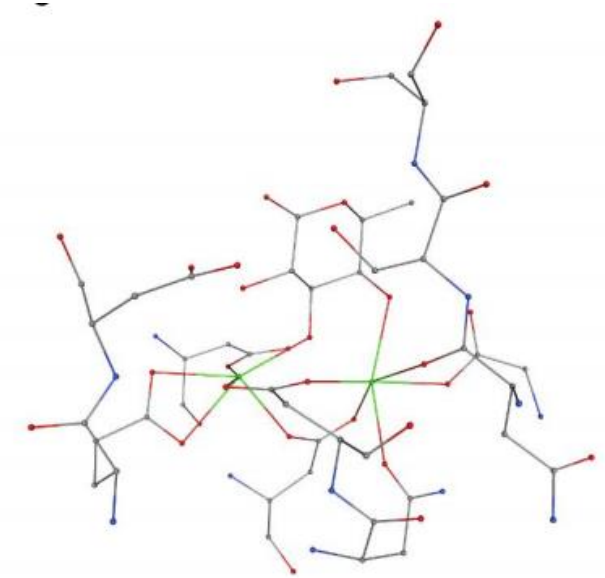
Sehnal et al., **Nucl. Acids Res.** (2015)

Pattern Detection Example

- sugar binding site from Pseudomonas aeruginosa lectin



```
Near(4,
   Rings(5 * ["C"] + ["O"]),
   Near(4, Atoms("Ca"), Atoms("Ca"))
   )
   .AmbientResidues(4)
```

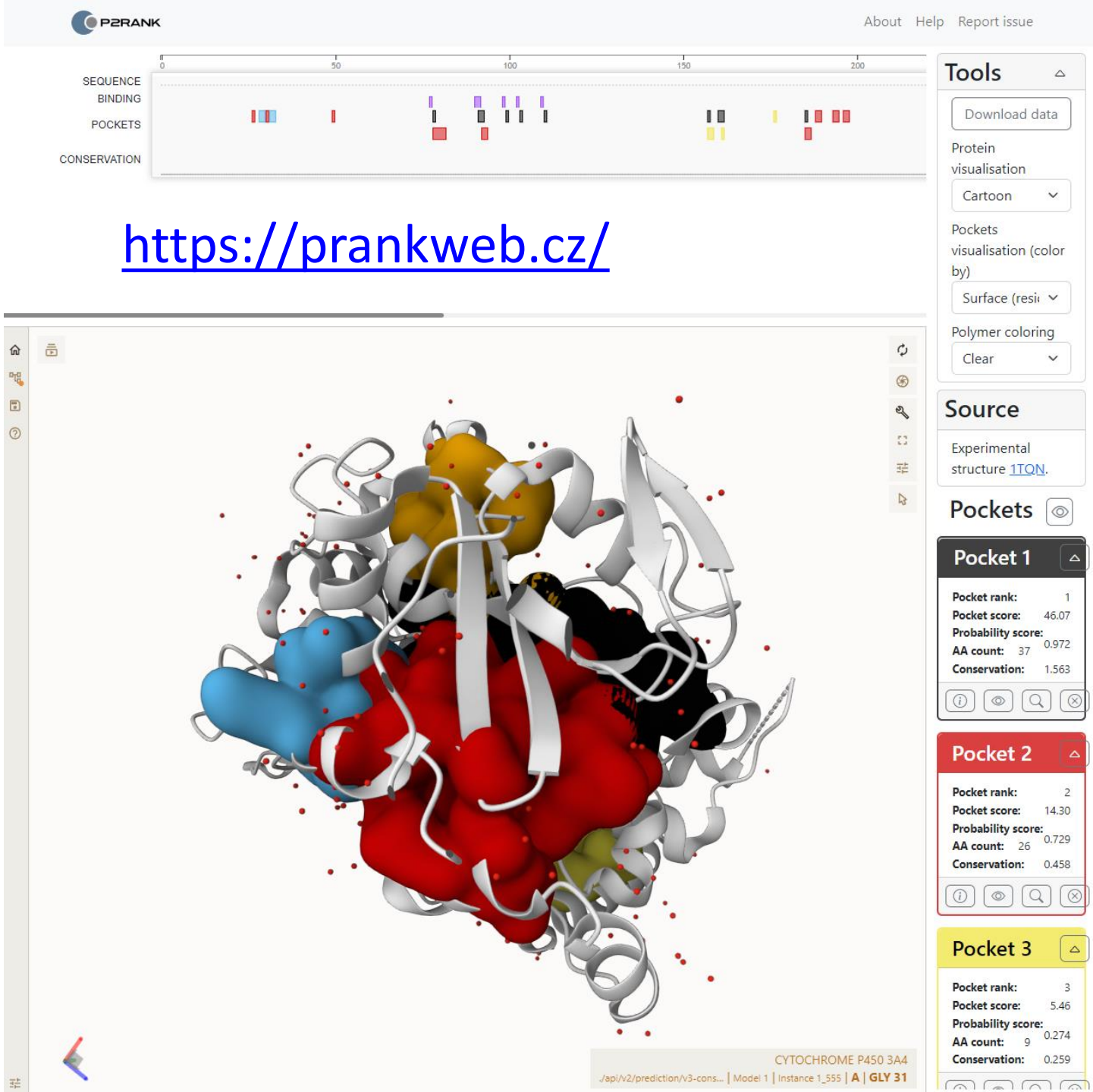


- Results (38 entries from PDB):
  - Pseudomonas aeruginosa (27 entries) or its synthetic (2 entries)
  - Burkholderia cenocepacia (4 entries)
  - Ralstonia solanacearum (2 entries)
  - Chromobacterium violaceum (2 entries)
  - Bacillus subtilis (1 entry)

ACTIVE SITE, CLEFTS, CHANNELS

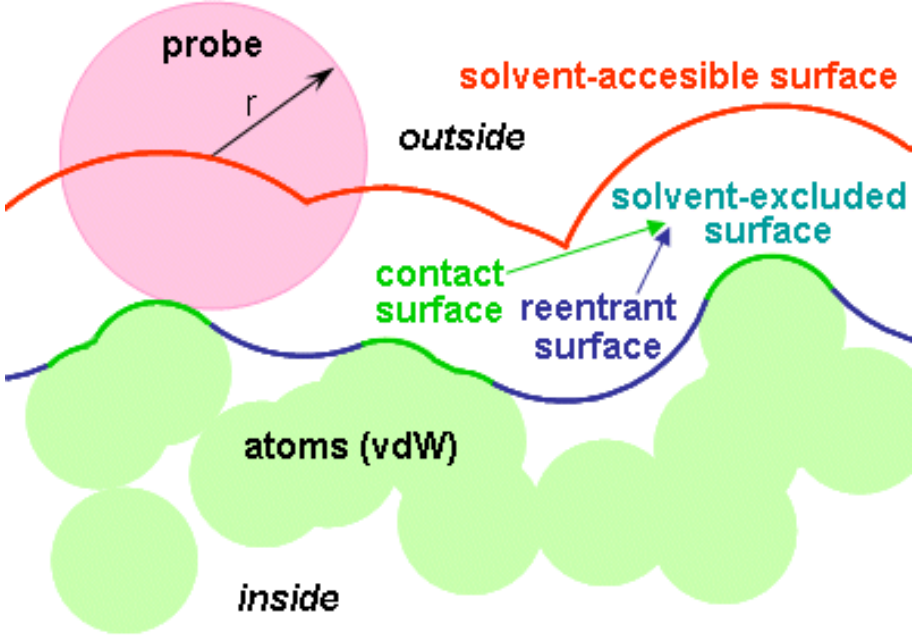
Active Site

- Determines function
  - Size and shape
    - Large (protein-protein, ev. protein-DNA)
      - PPI-PRED
    - small (clefts, ligands) - Surface, or Deep-inside (channels)
      - SURFNET, PocketFinder, Q-SiteFinder
  - Phys chem properties
    - hydrophobic patches, charged surfaces
  - Residue conservation – evolution trace method
    - ConSurf
  - meta methods
    - aka ProFunc
  - Machine learning
    - E.g. Blast2GO, **P2Rank**



https://prankweb.cz/

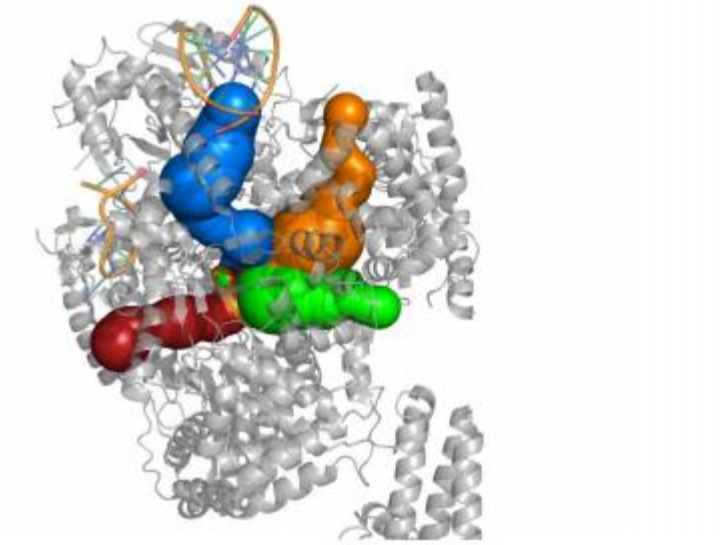
Cleft search

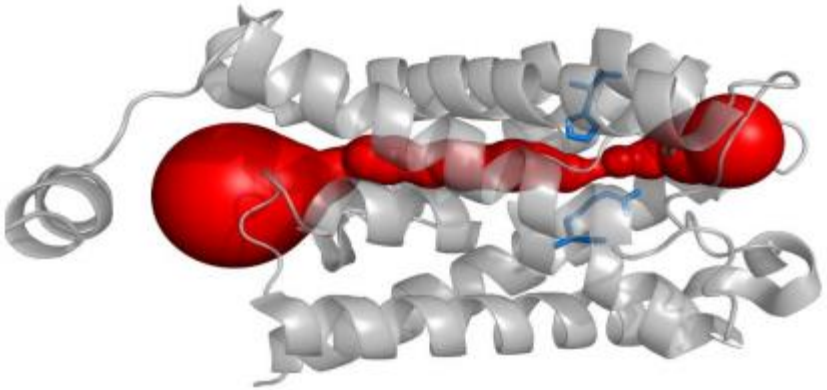


- SAS a SES
  - solvent accessible surface
  - solvent excluded surface
- Cleft search
  - Pocket-Finder
    - Slicing of receptor in several orientations
      - Closed spaces – cavities
      - Partially closed – pockets
  - QSiteFinder
    - grid search with generic probe (methan)

Channel Detection

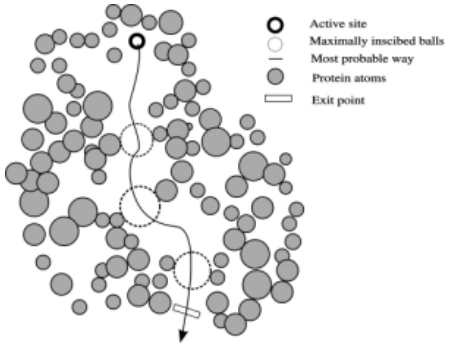
- important role in biomolecular function – co-determine specificity



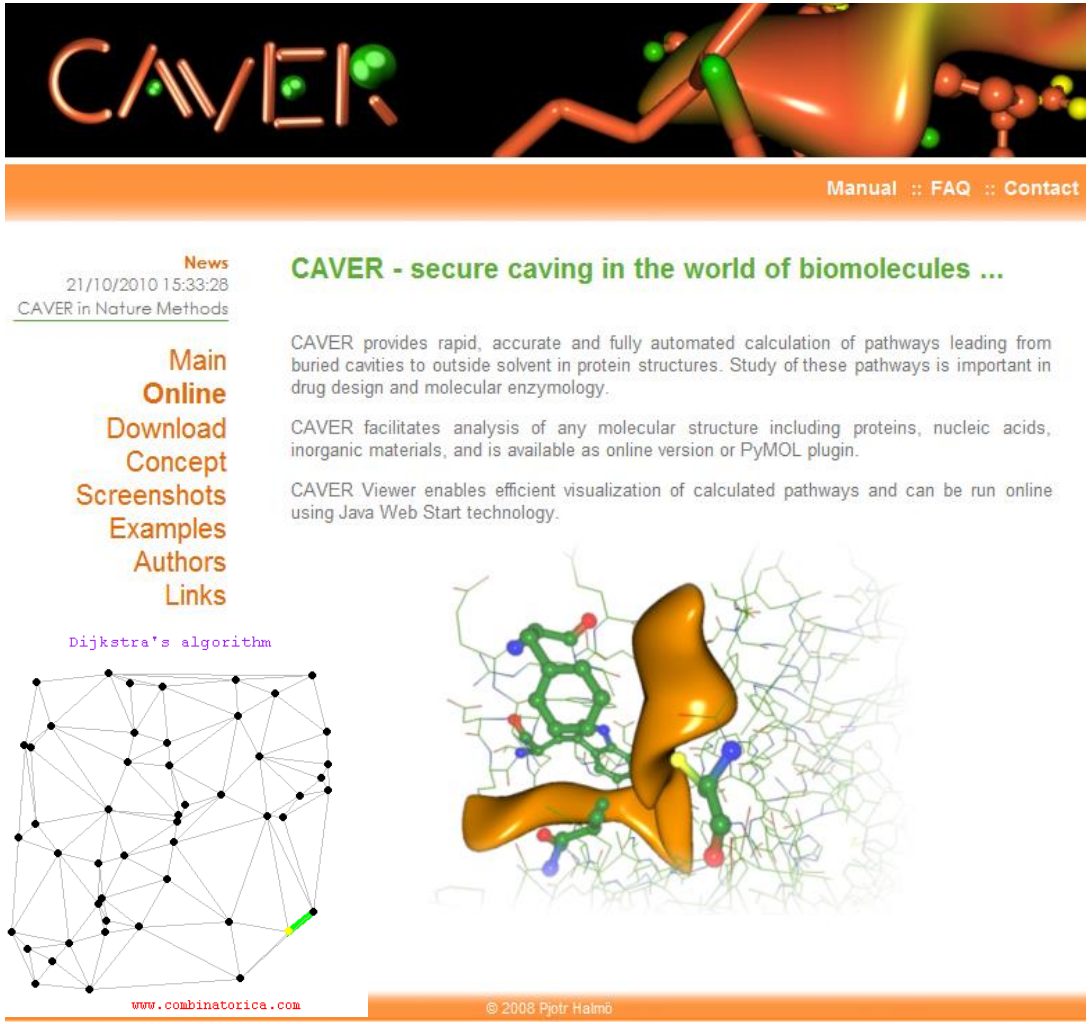


Channel Detection: Caver

- Caver



Dijkstra's algorithm
for optimal pathway search
on grid
or on Voronoi mesh

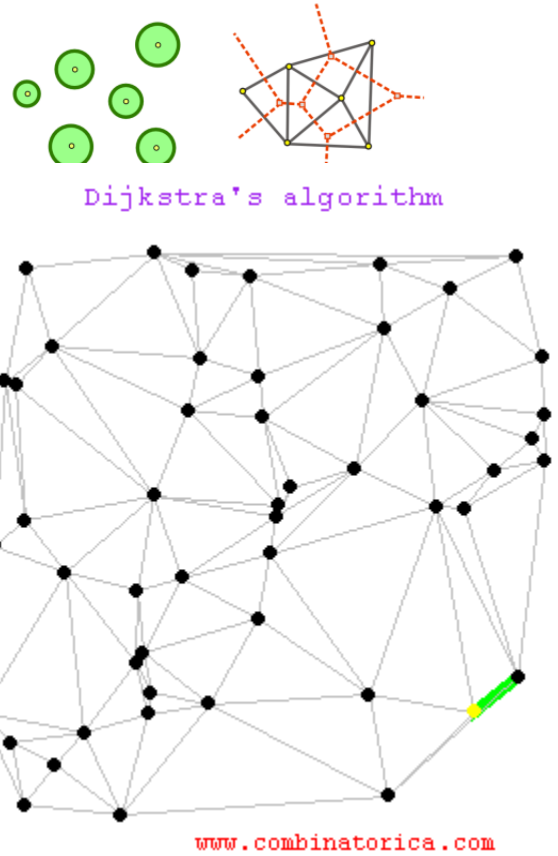


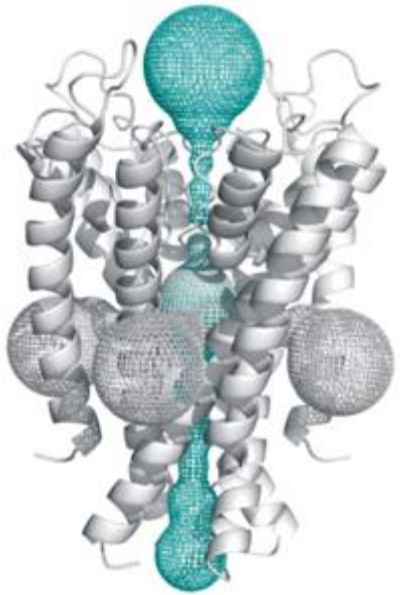
http://loschmidt.sci.muni.cz/caver/index.php

Petrek, M. et al - **CAVER: a new tool to explore routes from protein clefts, pockets and cavities,** *BMC Bioinformatics* 2006, **7**:316
Beneš, P. et al. - **CAVER 2.1 software**, 2010.

Channel Detection: MOLE

MOLE Atoms maps => Voronoi diagram

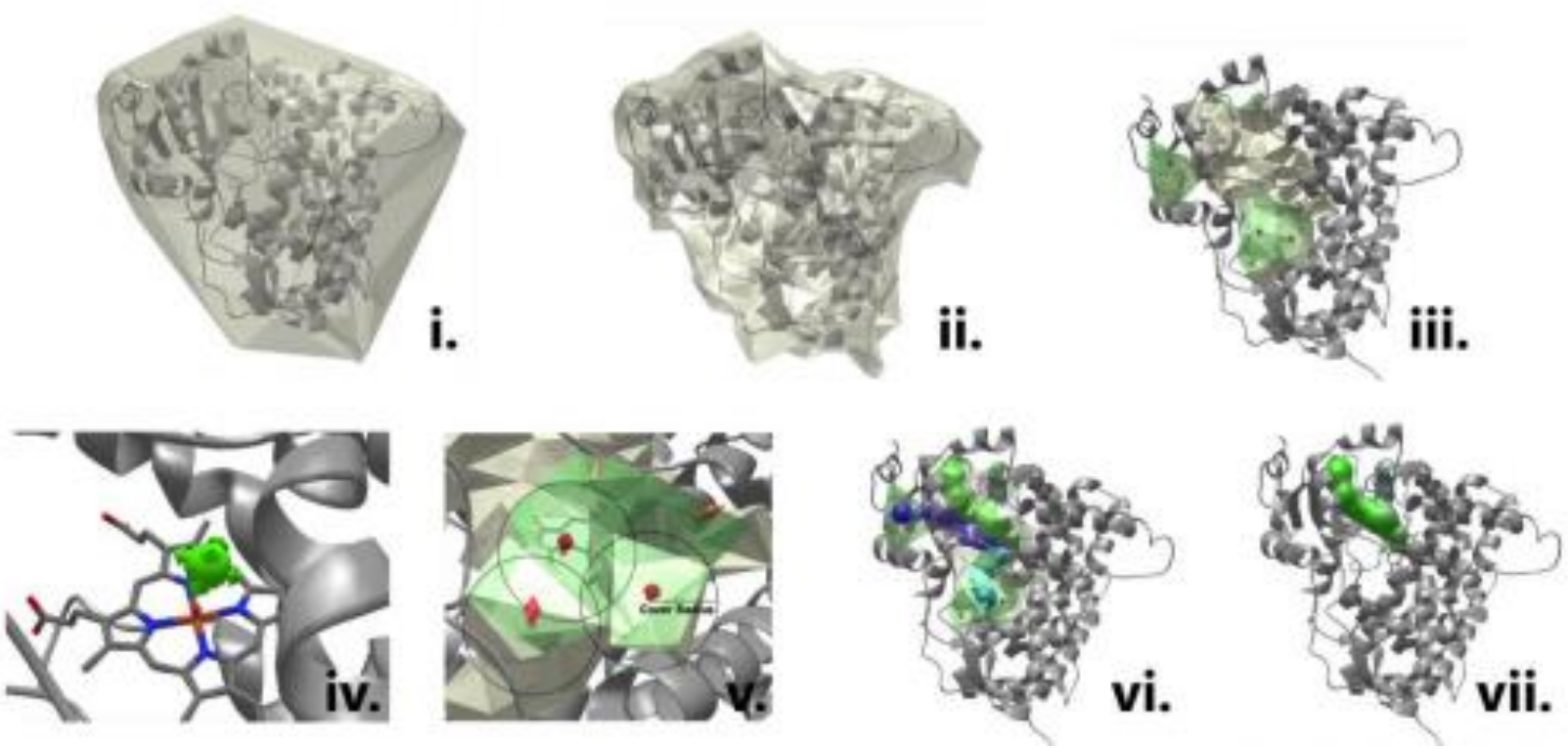




http://mole.chemi.muni.cz/online/current/
http://mole.upol.cz

Petrek M., Kosinova P., Koca J., Otyepka M.: MOLE: A Voronoi Diagram-Based Explorer of Molecular Channels, Pores, and Tunnels. Structure (2007) 15, 1357-1363.

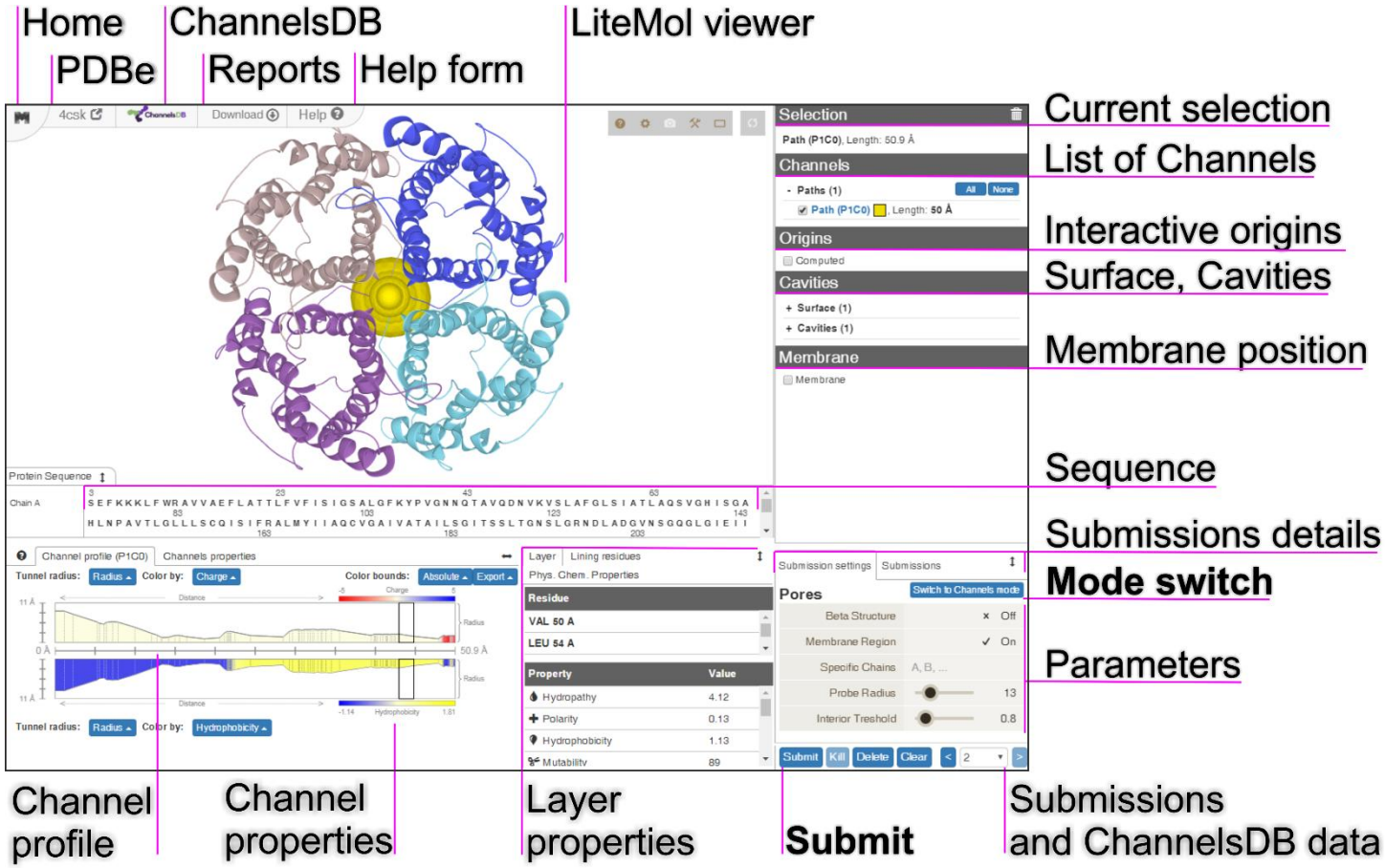
Channel Detection: MOLE 2.0



* Sehnal, Svobodová, et al., **J. Cheminf.** (2013)
** Berka, ..., Svobodová, et al., **Nucl. Acids Res.** (2012)

MOLEonline – mole.upol.cz

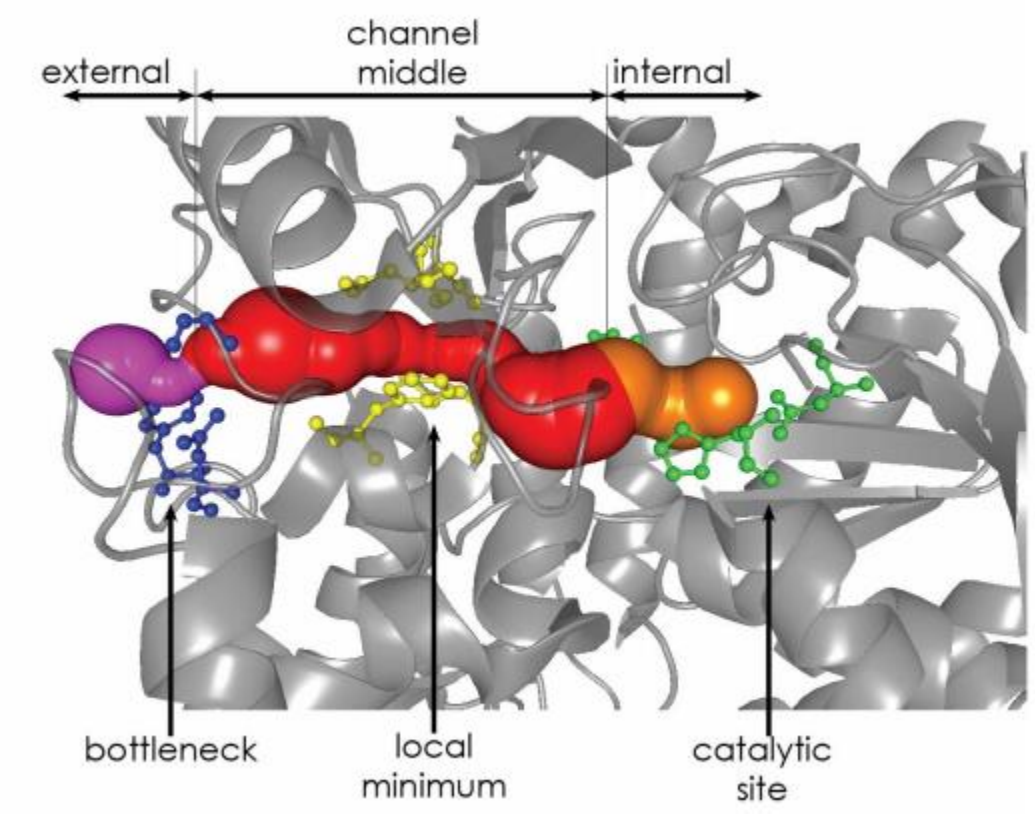
- Detection and analysis of channels, tunnels and pores



Pravda L, et al *Nuc. Acids Res,* 2018, doi:10.1093/nar/gky309

Anatomy of Channels

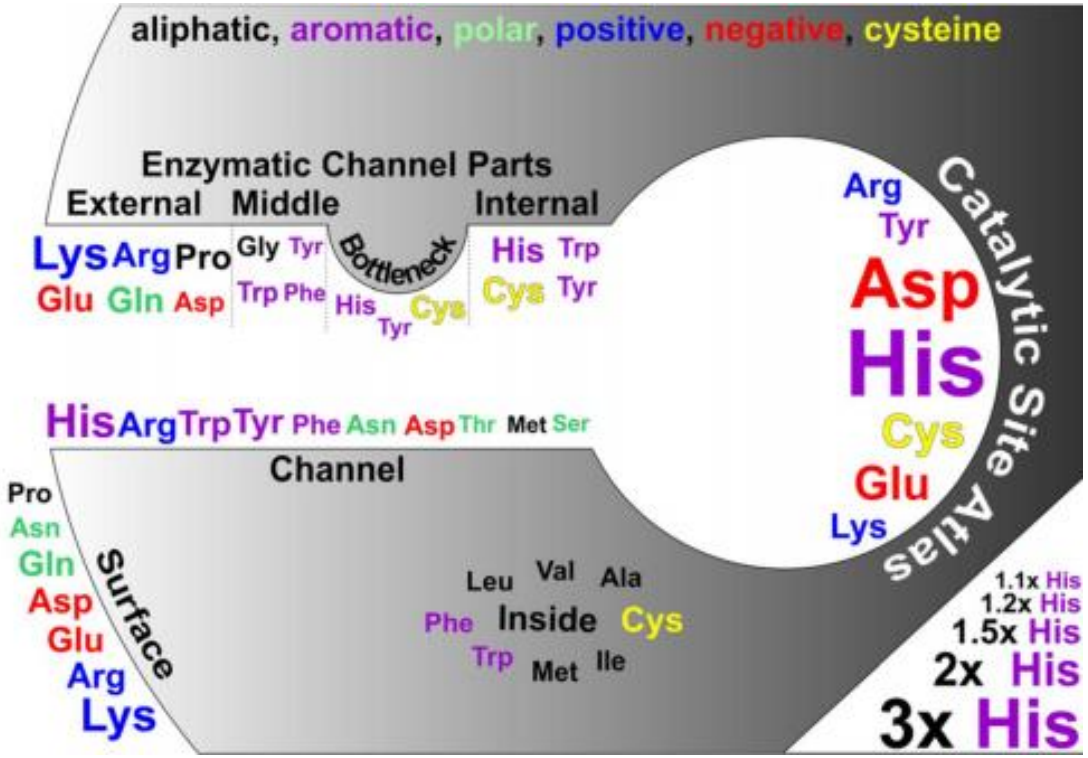
- **Methodology:**
  - Analysis of all the enzyme structures from PDB (4.306 structures)
  - Calculation of their properties via MOLE
  - Statistical analysis



Pravda L, Berka K, Svobodová Vařeková R, Sehnal D, Banáš P, Laskowski RA, Koča J, Otyepka M: Anatomy of enzyme channels. *BMC Bioinf., 15(1), 379, 2014*.

Anatomy of Channels

- **Results*:**
  - 64% of enzymes have at least one channel
  - Median channel length was 27.7 Å
  - Most of the channels are polar
  - An average channel is lined by 2 negative and 2 positive AAs
  - gating AAs enriched:

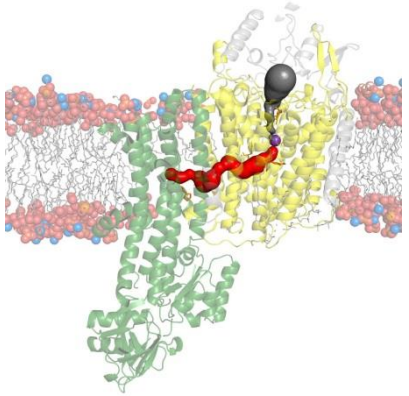


Pravda L, Berka K, Svobodová Vařeková R, Sehnal D, Banáš P, Laskowski RA, Koča J, Otyepka M: Anatomy of enzyme channels. *BMC Bioinf., 15(1), 379, 2014*.

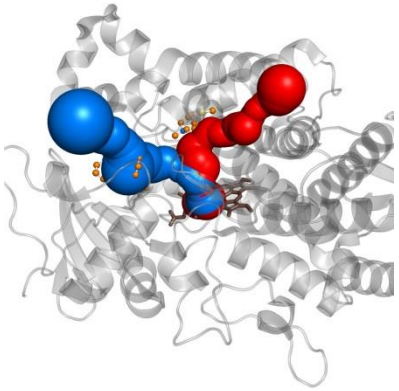
ChannelsDB

- Channel annotations for PDB
  - Reviewed from literature
  - Tunnels from Catalytic Site Atlas
  - Tunnels from enzyme cofactors (e.g. HEM or FAD)
  - Transmembrane pores
- Calculated with MOLE[1,2]

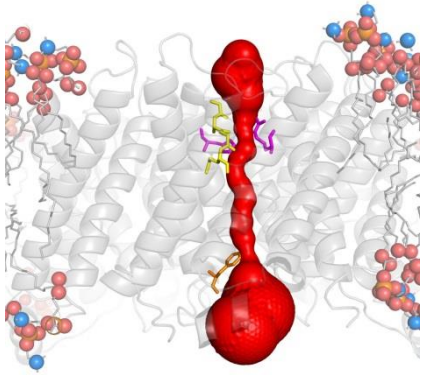
http://ncbr.muni.cz/ChannelsDB



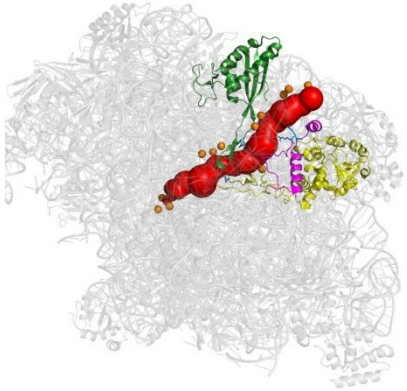
KdpFABC complex pore (5mrw)



Cytochrome P450 2D6 tunnel (3tbg)



Aquaporin (1ymg)

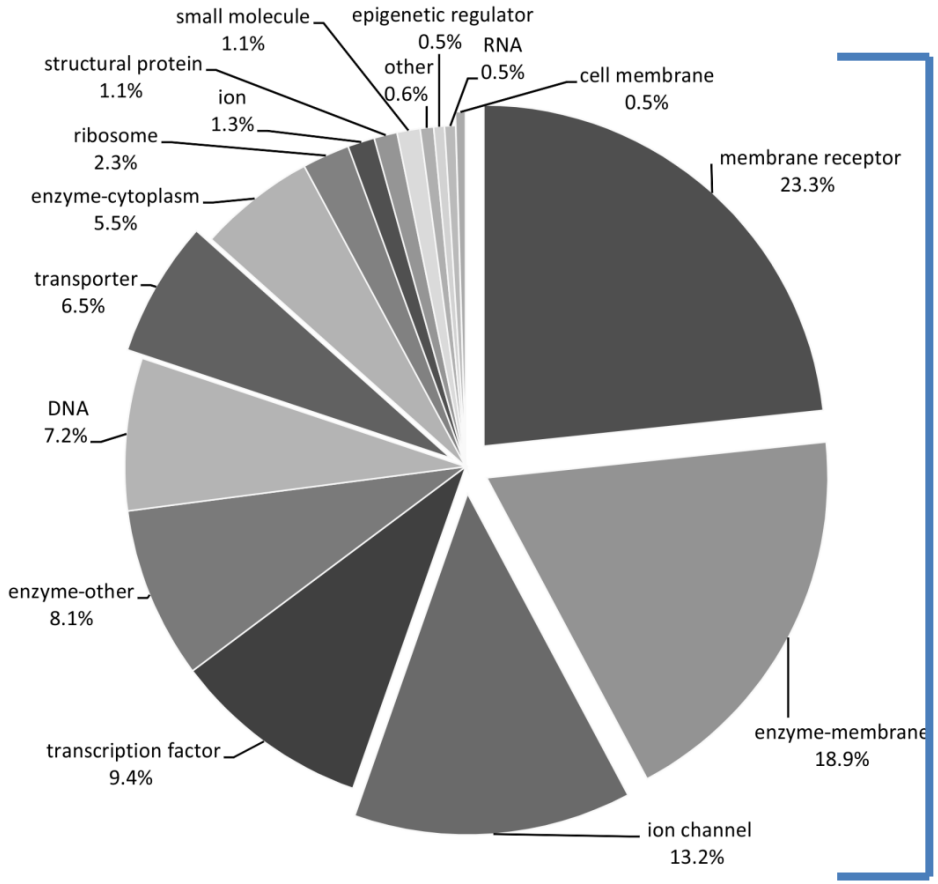


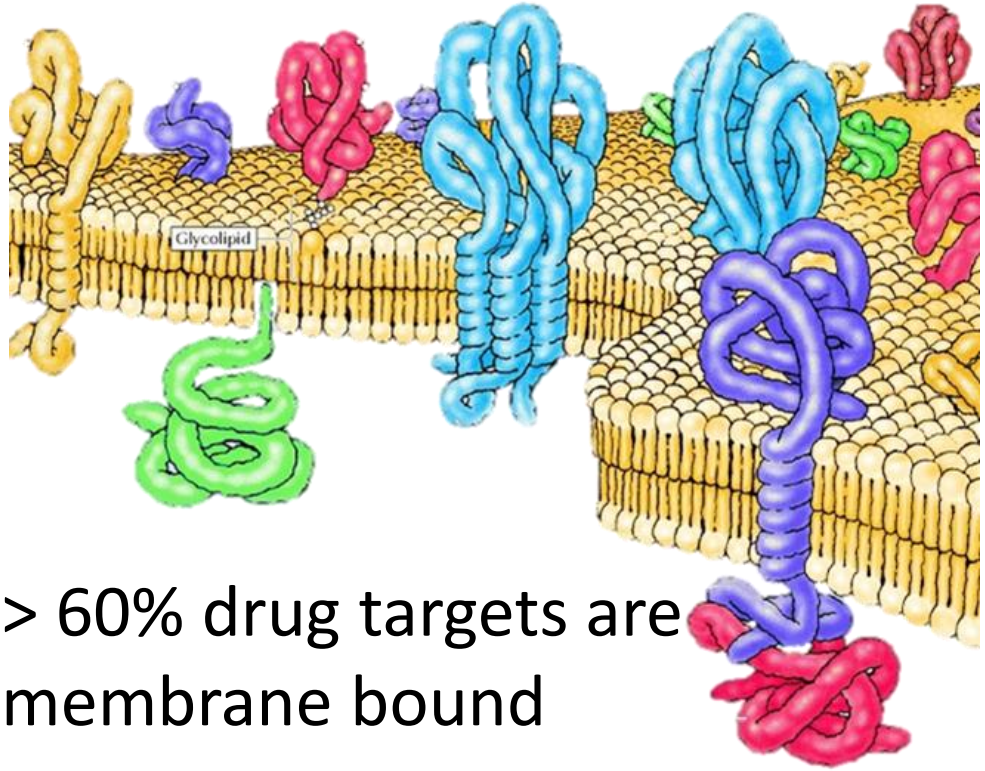
Ribosomal exit tunnel (1jj2)

1. Berka K et al, *Nuc. Acids Res., 40(W1), W222-W227, 2012*
2. Sehnal D et al., *J. Cheminform*, 5(1), 39, 2013
3. Pravda L et al. *Nuc. Acids Res*. *46(D1), D399–D405, 2018*. doi: 10.1093/nar/gkx868

SPECIALTY SERVERS – MEMBRANE PROTEINS, DISORDER

Membrane Proteins



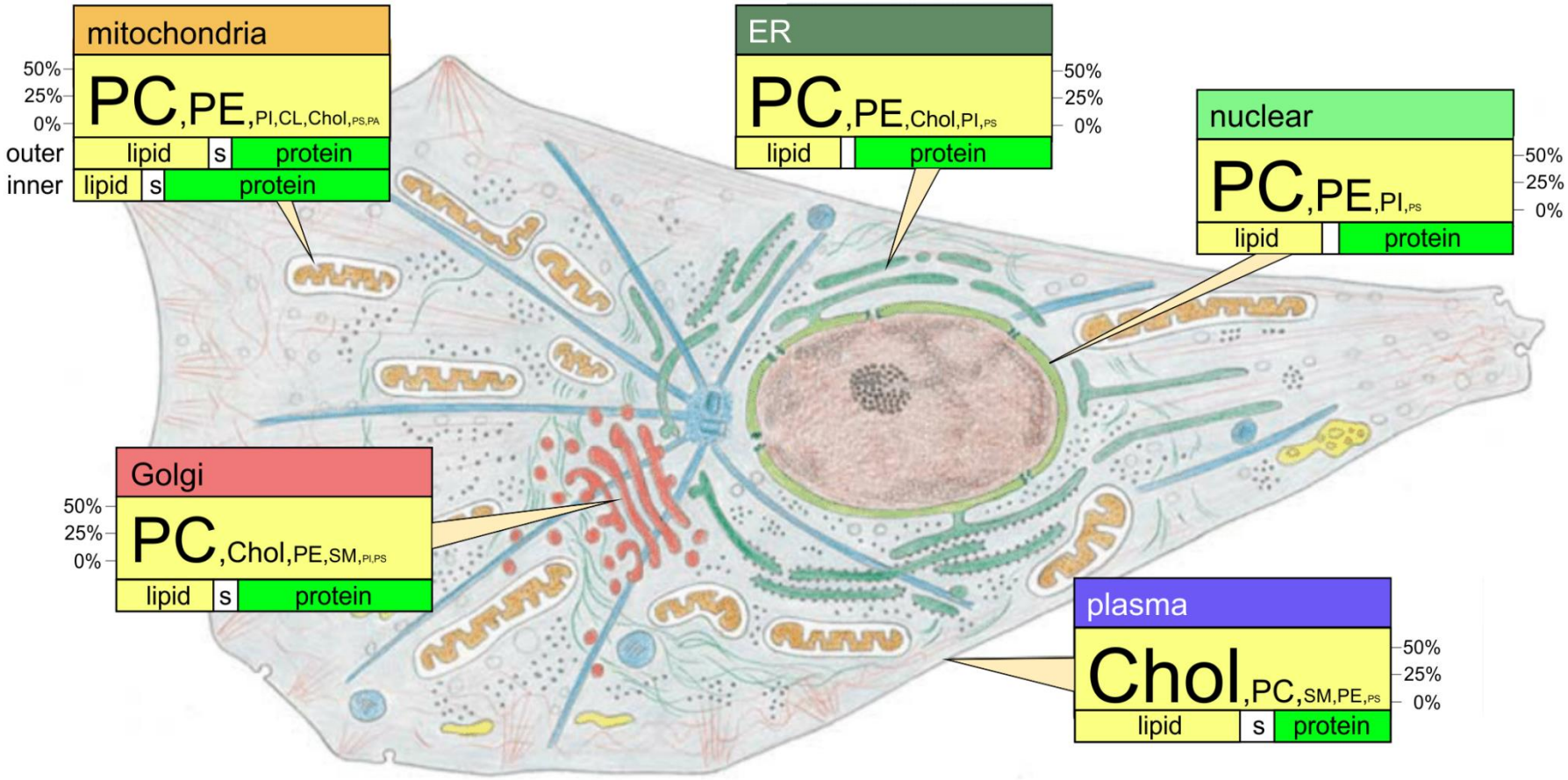


> 60% drug targets are membrane bound

Alas – not much membrane protein structures are known

Di Meo F, Fabre G, Berka K, Ossman T, Chantemargue B, Paloncýová M, Marquet P, Otyepka M, Trouillas P: In Silico Pharmacology: Drug Membrane Partitioning and Crossing. *Pharmacol. Res., 111, 471–486, 2016*.

Cell Membranes

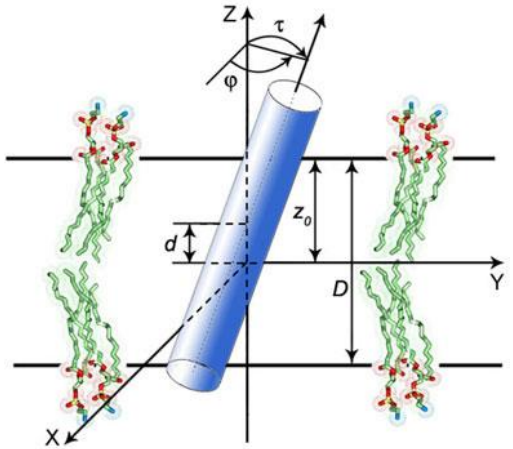


Data from Van Meer, G.; Voelker, D. R.; Feigenson, G. W.; Manuscript, A. Membrane Lipids: Where They Are and How They Behave. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 112–124..

Orientations of Proteins in Membranes OPM server

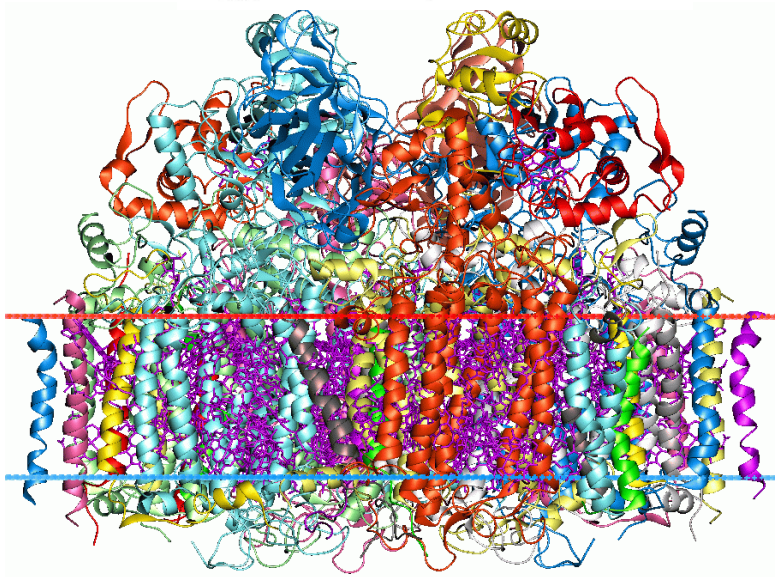
- protein = rigid body

$$\Delta G_{transf} = \sum_{i=1}^{N} \sigma_i^{wat \to bil}(z_i)\, ASA_i + \sum_{j=1}^{M} \eta_j^{wat \to bil}(z_j)\, \mu_j + \sum_{k=1}^{L} \min\{\Delta E_k^{ion}, \Delta E_k^{neutr}\}$$



- Optimization of protein orientation in anisotropic implicit membrane model

+ membrane composition

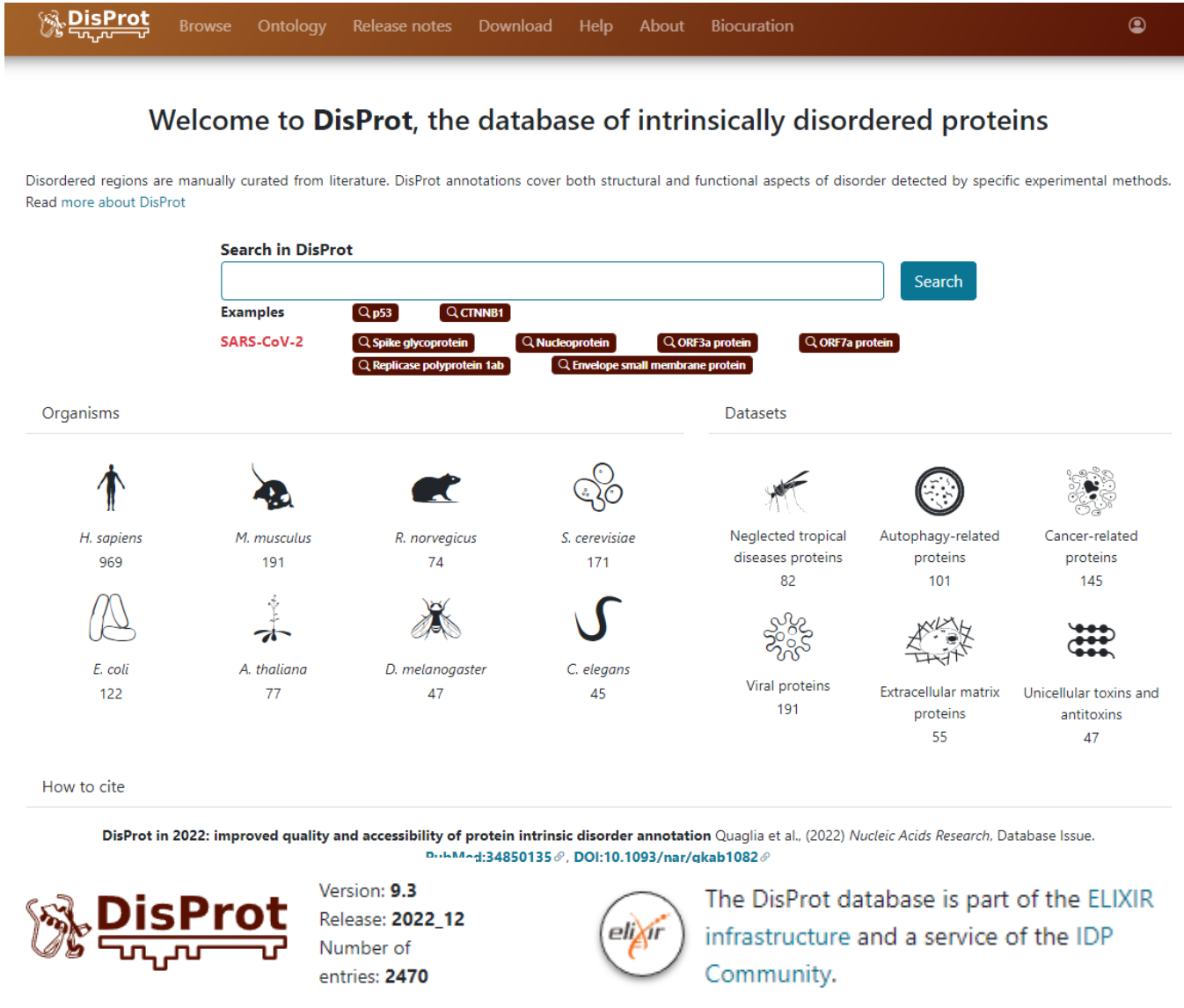


photosystem II

http://opm.phar.umich.edu

Disorder – Intrinsically Disordered Proteins (IDP)

- **(Invisible)** – no structures
- estimates – B-faktor in Xray, NMR ensemble, DISOPRED, AlphaFold
- intristically disordered proteins – **no need of structure to function**
- Regulation functions:
  - molecular recognition (promiscuitous)
  - molecular assembly (viral kapsids)
  - protein modification
  - entropic chain activities (entropic hourglasses)



TAKE HOME MESSAGE

Take Home Message

- There is an enormous amount of information available
- Primary databases - **PDB** and **Uniprot**
- Structure defines function
- Active site can have quite often channels
- Membrane proteins are one of the most important groups
- Disorder is also important especially in regulatory proteins (protein-protein interactions)