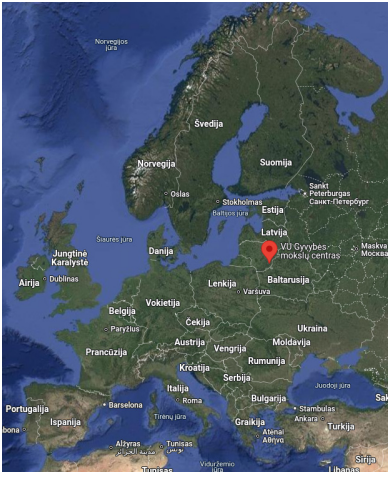
Data-driven research in structural bioinformatics: from proteins to biomolecular interactions

Justas Dapkūnas

Institute of Biotechnology, Life Sciences Center, Vilnius University, Lithuania

2026-11-03

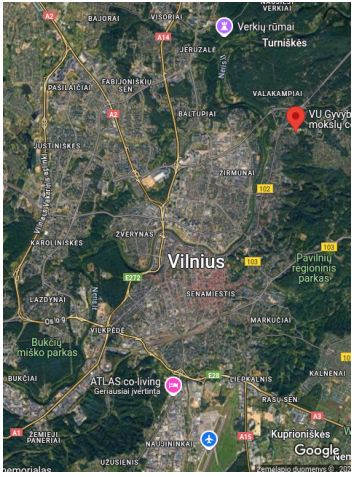
Vilnius University Life Sciences Center (1)



Vilnius University Life Sciences Center (2)



Vilnius University Life Sciences Center (3)



Vilnius University Life Sciences Center (4)



Vilnius University and Sorbonne University



- Communication at administration level
- VU visits Sorbonne on December 10-11
- My current visit: Lithuanian-French collaboration program „Gilibert“

Outline

1. Introduction: structural bioinformatics
2. The PPI3D resource:
   - Analysis of interaction interfaces using Voronoi tessellation
   - Clustered data on diverse protein interactions in PPI3D
3. Applications of PPI3D data
4. Other data resources for protein interactions
   - SKEMPI and other databases with data on interaction energetics
   - Some notes on PDBBind
5. Concluding remarks

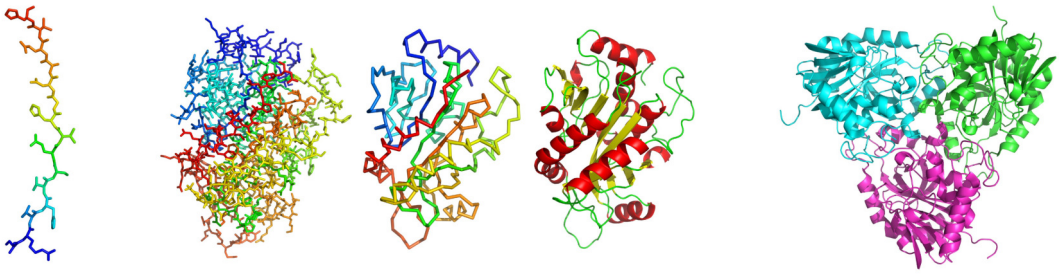
Bioinformatics

Wikipedia:
Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data.

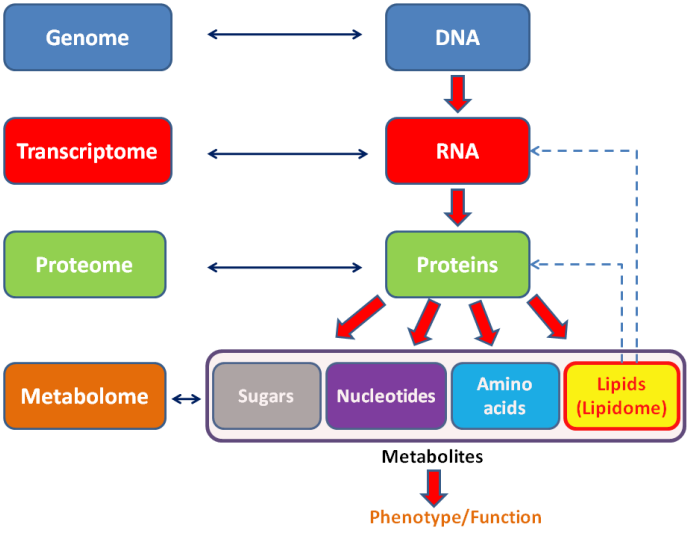
As an interdisciplinary field of science, bioinformatics combines **computer science, statistics, mathematics, and engineering** to analyze and interpret **biological data**.

Structural bioinformatics is the branch of bioinformatics which is related to the **analysis and prediction** of the three-dimensional **structure of biological macromolecules**.

Proteins



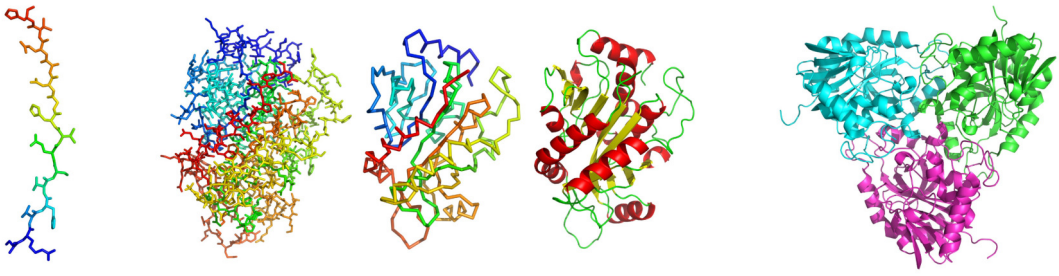
Molecular biology of the cell



Protein databases

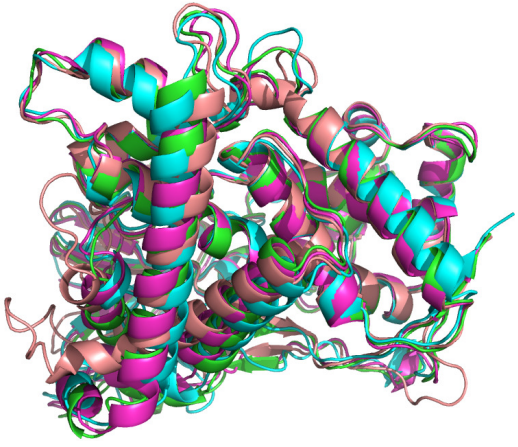
- Protein sequence databases
  - UniProt (150M)
  - UniParc
  - BFD (Big Fantastic Database) (2500M)
  - MGnify (1600M)
  - ...
- SwissProt: annotated proteins (576k)
- Protein structure databases
  - Protein Data Bank (experimental structures, 260k)
  - AlphaFold Database (structure models)
  - Other databases with predicted structures
- Protein interaction databases (e.g., String, CORUM, Complex Portal)
- Databases of protein properties

Proteins

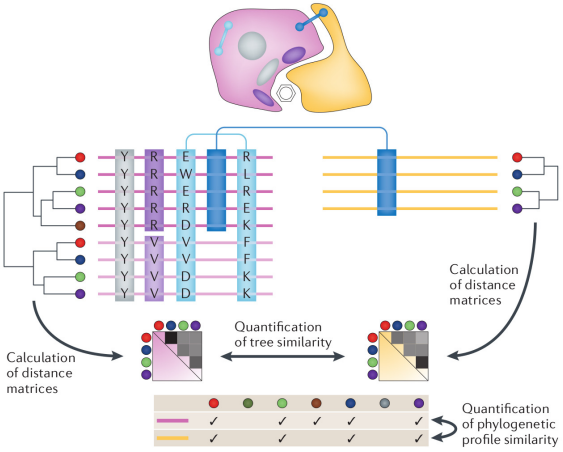


Relationship between protein sequence and structure





Coevolution and protein structures



De Juan et al., Emerging methods protein coevolution, **2013**, *Nat Rev Genet*, 14:249

AlphaFold: headlines

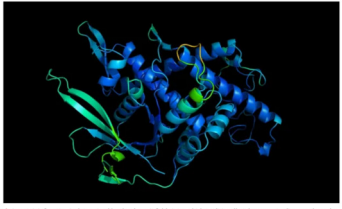
NEWS | 30 November 2020

‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures

Google’s deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

DeepMind AI cracks 50-year-old problem of protein folding

Program solves scientific problem in ‘stunning advance’ for understanding machinery of life



▲ A protein’s function is determined by the shape it folds into, which traditionally takes years to discover through lab work. Photograph: DeepMind/PA

DeepMind's AI biologist can decipher secrets of the machinery of life

TECHNOLOGY 30 November 2020

By Michael Le Page



Artificial intelligence

DeepMind’s protein-folding AI has solved a 50-year-old grand challenge of biology

AlphaFold can predict the shape of proteins to within the width of an atom. The breakthrough will help scientists design drugs and understand disease.

One of biology's biggest mysteries 'largely solved' by AI

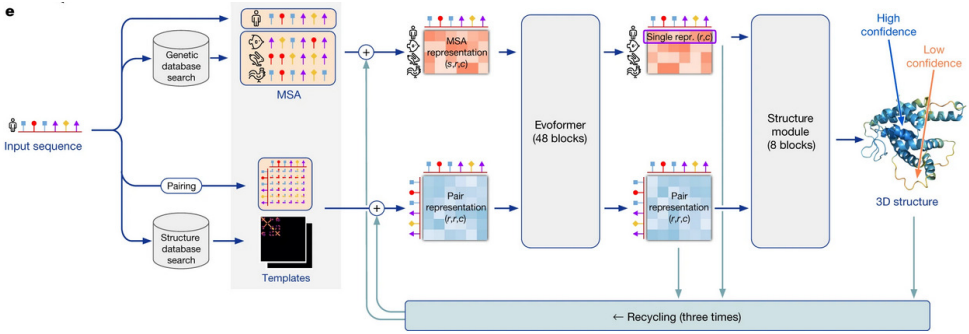
NEWS | BIOLOGY

‘The game has changed.’ AI triumphs at solving protein structures

In milestone, software predictions finally match structures calculated from experimental data

30 NOV. 2020 · BY ROBERT F. SERVICE

AlphaFold: how?

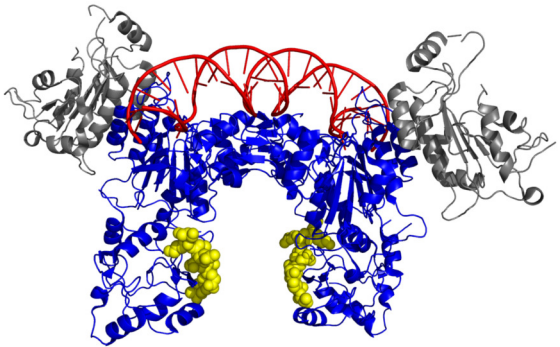


Jumper et al., **2021**, *Nature*, 596:583

Databases used:

- Sequences: UniProt, BFD, MGnify
- Structures: Protein Data Bank

Protein-protein, protein-peptide and protein-nucleic acid interactions



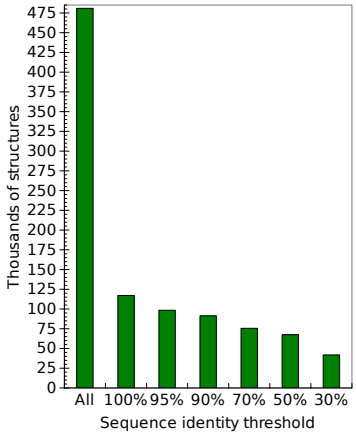
PDB: 4U7T, 5YX2

Current state of structural bioinformatics on biomolecular interactions

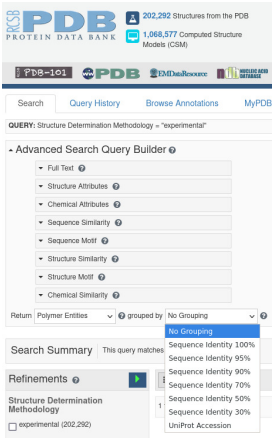
- Protein-protein interactions: AlphaFold predicts almost everything
- Protein-nucleic acid interactions: still a hard problem
- Protein-ligand interactions: very interesting
- High quality datasets are necessary in all cases

Data redundancy in the Protein Data Bank

Highly redundant data

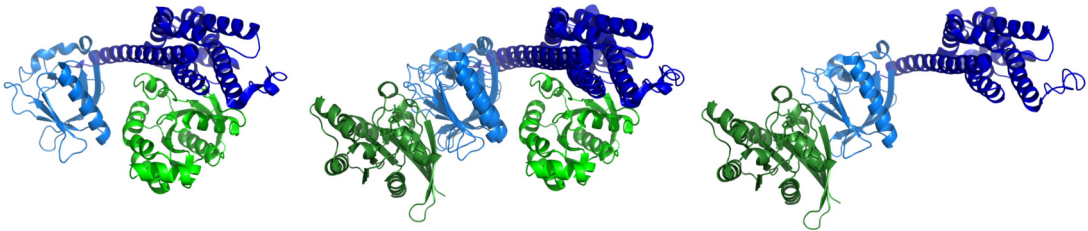


Clustering by protein sequences



`https://rcsb.org`

Alternative protein interactions are frequent exceptions

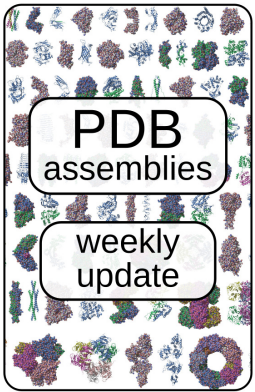


PDB: 1XCG, 3KZ1
Derewenda et al., *Structure*, **2004**, 12:1955
Chen et al., *J Biol Chem*, **2010**, 285:21070

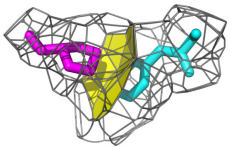
Hamp and Rost, *PLoS Comput Biol*, **2012**, 8:e1002623

PPI3D: search, analyze and model protein interactions

https://bioinformatics.lt/ppi3d/



Interfaces from
Voronoi diagram



Clustering

by **sequences**

by **structures**



PPI3D

Search by protein sequences

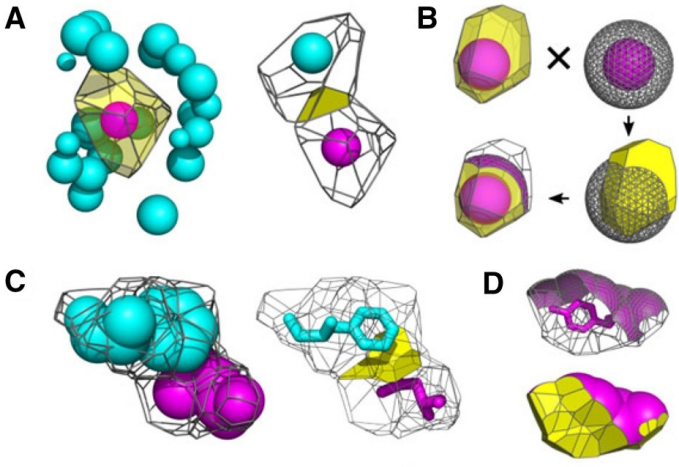
Clustered interactions with **proteins**, **peptides**, **DNA**, **RNA**

Interaction interface details

Download clusters by criteria

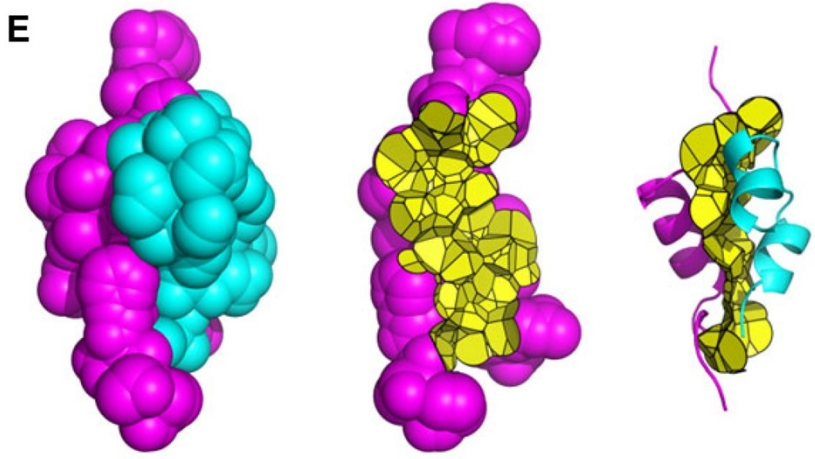
Dapkūnas et al., *Nucleic Acids Res*, **2024**, 52:W264

Analysis of macromolecular structures using the Voronoi tessellation



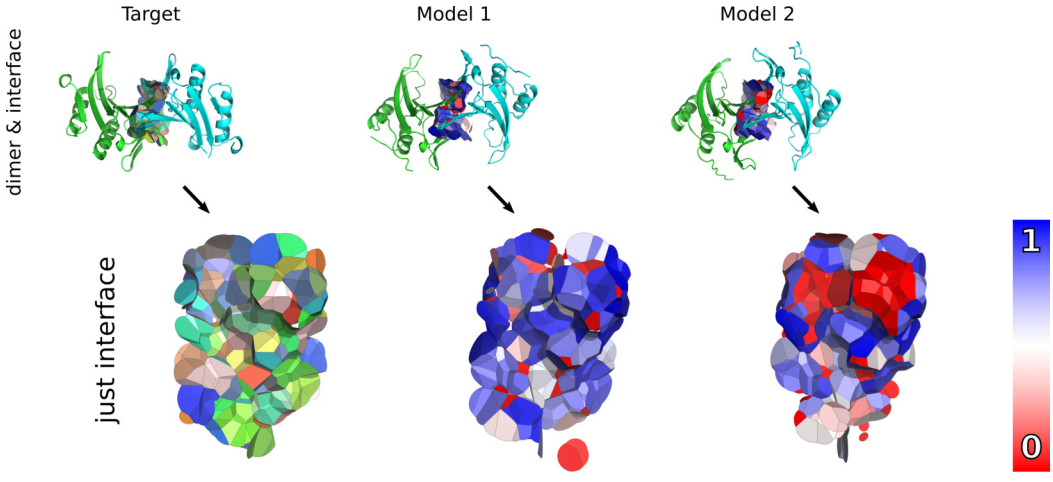
Olechnovič & Venclovas, *Bioinformatics*, **2021**, 37:4873

Definition of interaction interfaces



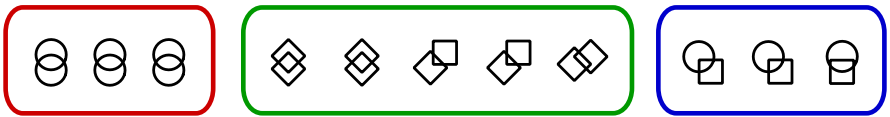
Olechnovič & Venclovas, *Bioinformatics*, **2021**, 37:4873

CAD-score – Contact Area Difference score

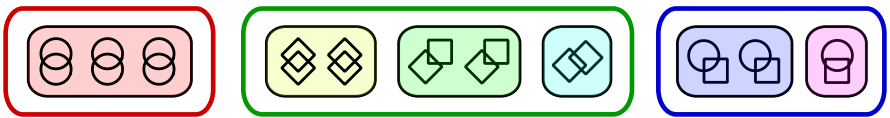


Clustering interfaces and binding sites

by sequence similarity:

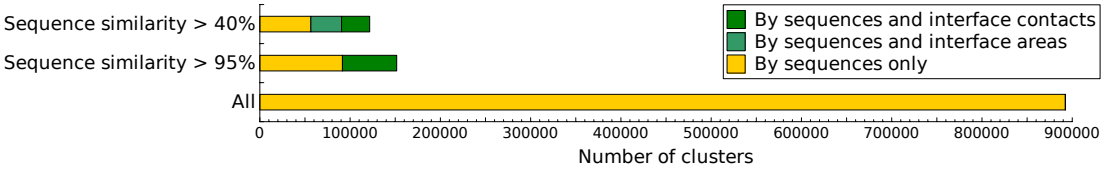


by structural similarity:

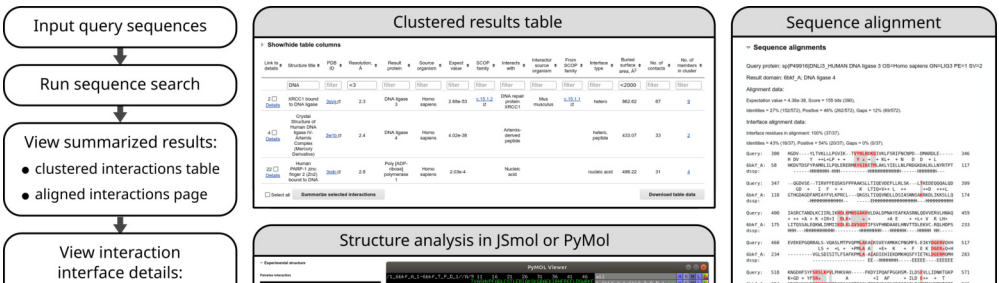


PPI3D clustering keeps alternative interaction interfaces

Clustering of protein-protein interaction interfaces by sequence and interface similarity:



PPI3D web server



Dapkūnas et al., *Nucleic Acids Res*, **2024**, 52:W264

Retrieve PPI3D clustered data according to user-defined criteria

https://bioinformatics.lt/ppi3d/clusters/

Enter dataset criteria for interaction interfaces

Interaction types

Protein-protein interactions: ☑

Protein-peptide interactions: ☑

Protein-nucleic acid interactions: ☑

PDB data

Resolution better than: 4

Release date between 01 / 01 / 1973 and 03 / 15 / 2023

Complexes criteria

Number of subunits between 1 and 10000

Number of protein subunits between 1 and 10000

Number of different subunits between 1 and 10000

Total number of residues between 1 and 1000000000

Number of visible residues between 1 and 1000000000

Subunits criteria

Total number of residues between 1 and 100000

Number of visible residues between 1 and 100000

Interfaces criteria

Interface area between 100 and 100000

Number of interface contacts between 1 and 10000

Allow ligands in interface: ☑

Homo-interactions: ☑

Hetero-interactions: ☑

Clustering

Clustering of interfaces: Sequence similarity > 95%, similarity of interface contacts > 50%

Include clustering data in table: ☐

Submit

Retrieved data

Query the PPI3D database



Download the structures, if necessary

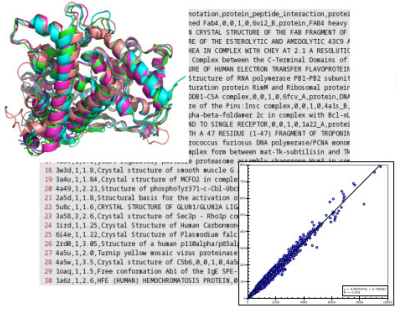
```
$ wget https://bioinformatics.lt/ppi3d/download/interface_coordinates/protein_nucleic-3l2p-1-3l2p_A-1-3l2p_D_B_C-1.pdb
```

More downloads

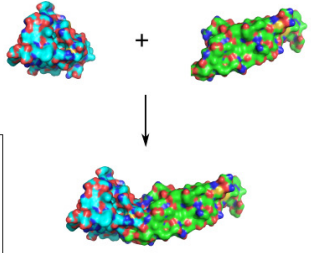
- Interface residues and contacts
- Intermediate clustering data
- Any requests?

Applications

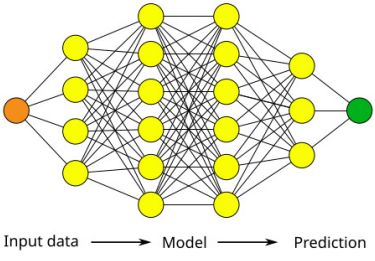
Large scale analysis of diverse protein interactions



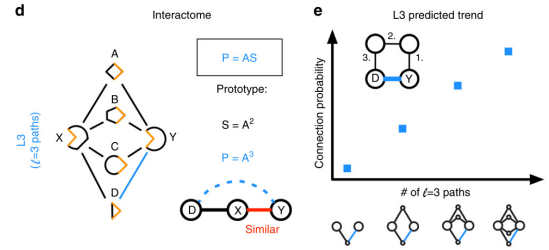
Development of new docking methods



Training of machine learning models



Application 1: structural perspective of the L3 principle





(d) Biological Motivation of the L3 Principle

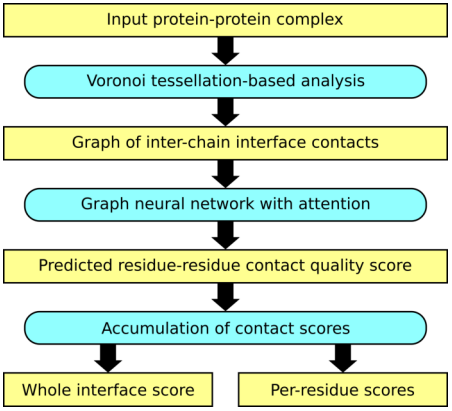
Yuen & Jannson, *BMC Bioinformatics*, **2023**, 24:59

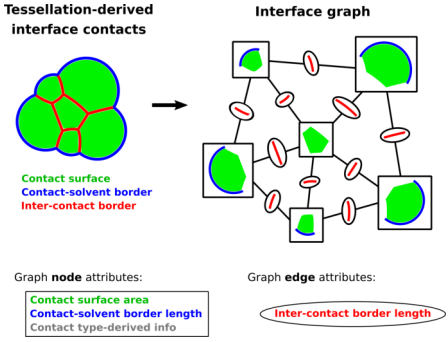
interact with each other (Fig. 1a). From a structural perspective, the common partners simply reflect the fact that proteins X and Y have a similar or identical interaction interface, that recognizes the same binding sites in proteins A, B, and C (Fig. 1d). Yet, a

Kovacs et al., *Nat Commun*, **2019**, 10:1240

Lionel Tabourier, Elodie Laine (Sorbonne University, France)

Application 2: VoroIF-GNN, protein-protein interface assessment trained using docking models

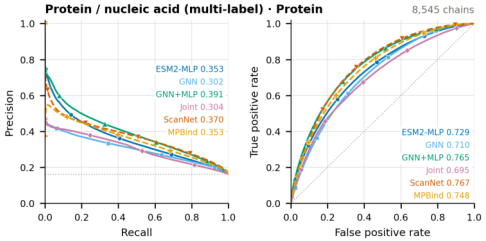


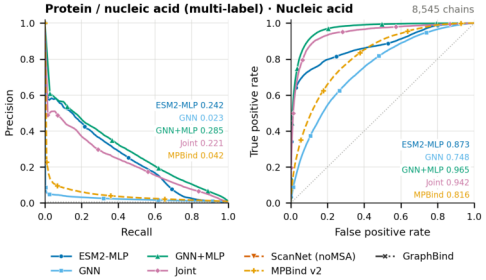


Olechnovič and Venclovas, *Proteins*, **2023**, 91:1879

Application 3: Prediction of binding residues given protein structure

- Predict which residues will bind proteins, DNA or RNA
- Use pre-processed structural data from PPI3D (PDB)
- Use only interactions from same organism (Taxonomy DB)
- Do not use antibody-antigen interactions (SabDab)





SKEMPI 2.0

SKEMPI v2.0: Structural database of Kinetics and Energetics of Mutant Protein Interactions

Jankauskaitė J, Jiménez-García B, Dapkūnas J, Fernández-Recio J, Moal IH. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation *Bioinformatics* **2019** 35:462

- More than 500 citations
- 7085 datapoints (mutations)
- 345 structures
- 295 articles
- Manually curated dataset

Other databases on thermodynamics and kinetics of biomolecular interactions

- ProNAB: Database for binding affinities of protein-nucleic acid complexes and their mutants
- NABE: Protein-Nucleic Acid Binding Energetic Database
- PDBBind: large binding database, but paywalled since recently

Concluding remarks

- Currently, structural bioinformatics is data science
- Data are available for protein sequences and structures
- PPI3D offers clustered structural data on binary interactions
- Data on biomolecular interaction affinities and other properties are hard to get

Acknowledgements



Head of the Department of Bioinformatics: Česlovas Venclovas

Voronoi tessellation tools: Kliment Olechnovič

PPI3D developers: Albertas Timinskas, Mindaugas Margelevičius, Rytis Dičiūnas

Ideas on protein-nucleic acid interaction analysis: Miglė Tomkuvienė

Binding site prediction: Jokūbas Bucevičius, Rita Banciul, Simonas Vytautas Ašmontas

L3 links in protein interaction networks: Lionel Tabourier, Elodie Laine

Funding: Research Council of Lithuania