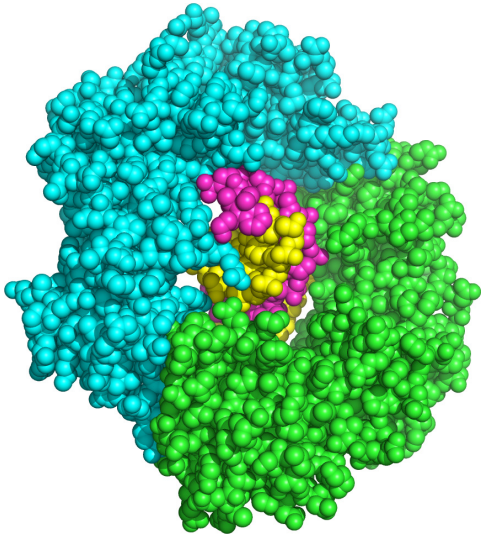
Analysis of interfaces in protein complexes using Voronoi tessellations and graph neural networks

Dr. Kliment Olechnovič

CNRS Laboratoire Jean Kuntzmann, Grenoble, France

Vilnius University Life Sciences Center, Vilnius, Lithuania

2024-02-12



Common problems:

- analyzing how different parts in a molecule interact
- selecting the best prediction of a multimeric complex

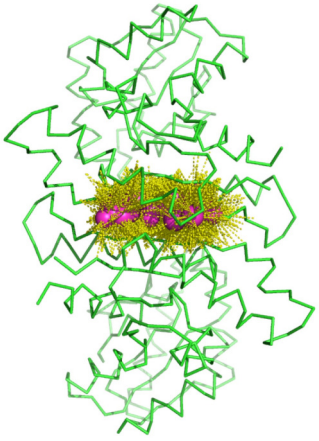
Our solutions involve:

- computational geometry
- machine learning
- developing free software

The prevalent approach to define interactions

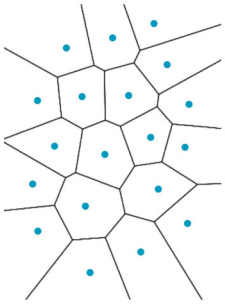
Interactions defined by distance cutoff of 5 angstroms

Interactions defined by distance cutoff of 7 angstroms

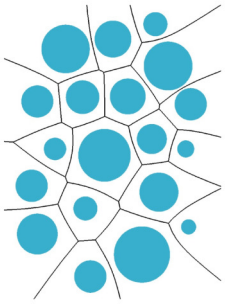


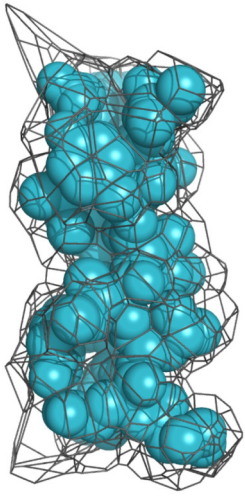
Voronoi diagram of points and balls

"Classic" Voronoi diagram of points



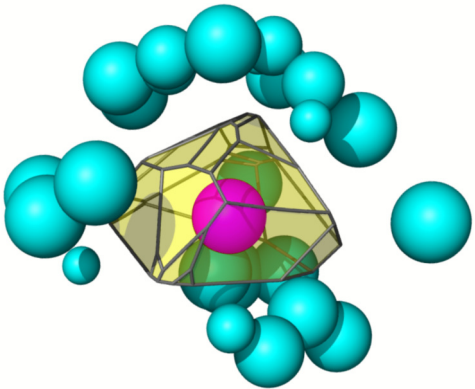
Voronoi diagram of balls



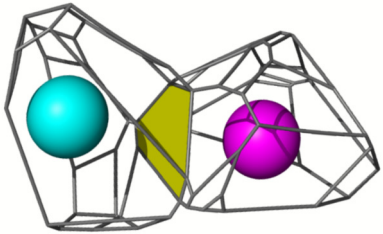


Voronoi tessellation-based analysis of structures

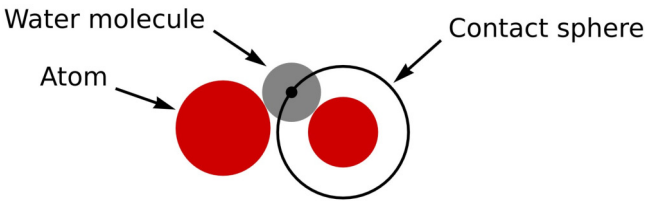
Voronoi cell of an atom surrounded by its neighbors

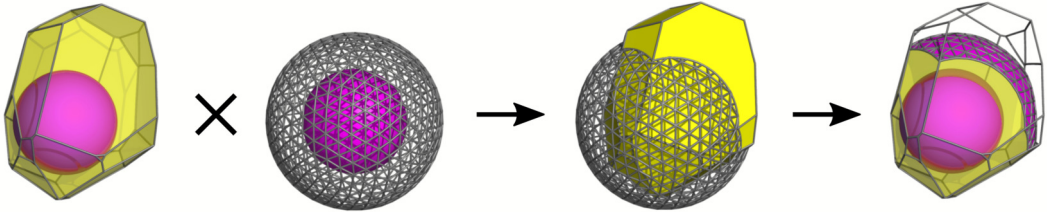


Atom-atom contact surface defined as the face shared by two adjacent Voronoi cells.



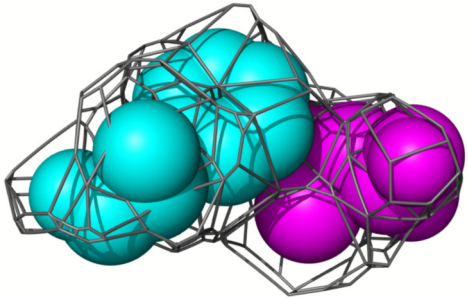
Constrained contacts



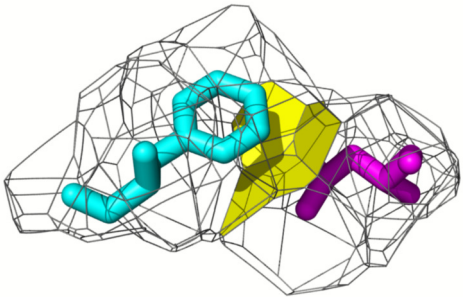


Deriving residue-residue contacts

Voronoi cells of two neighboring residues

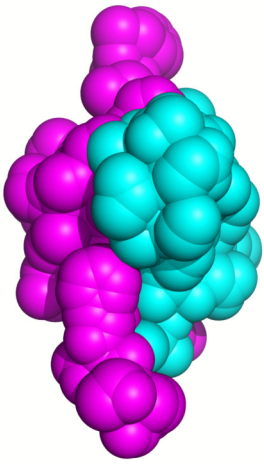


Residue-residue contact surface defined as a union of atom-atom contact surfaces

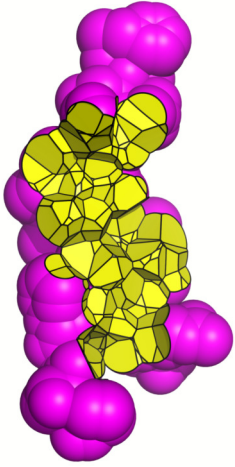


Inter-chain contacts

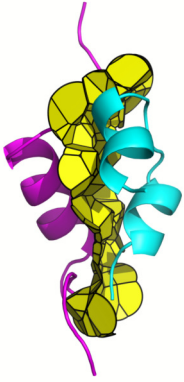
Solvent-accessible surface of an insulin heterodimer PDB:4UNG colored by subunit



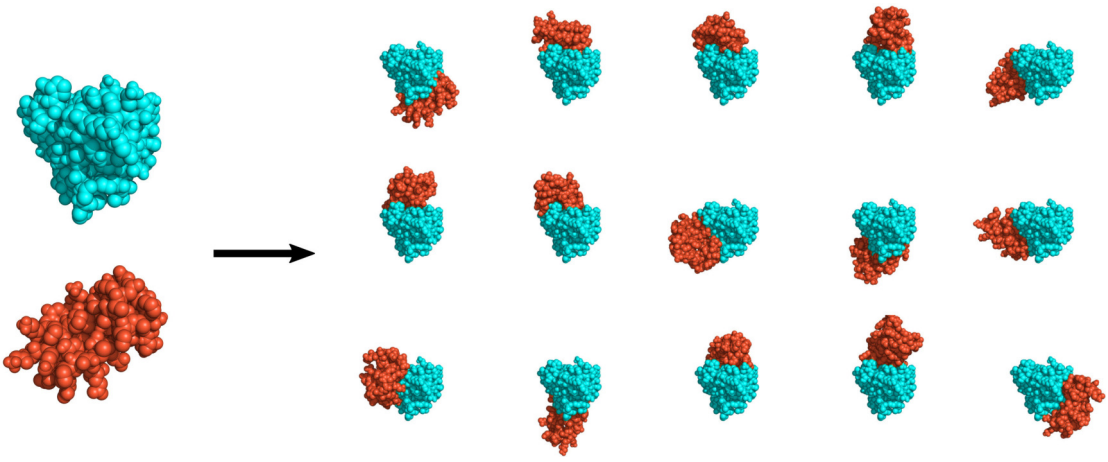
The intersubunit interface shown together with the SAS of one subunit



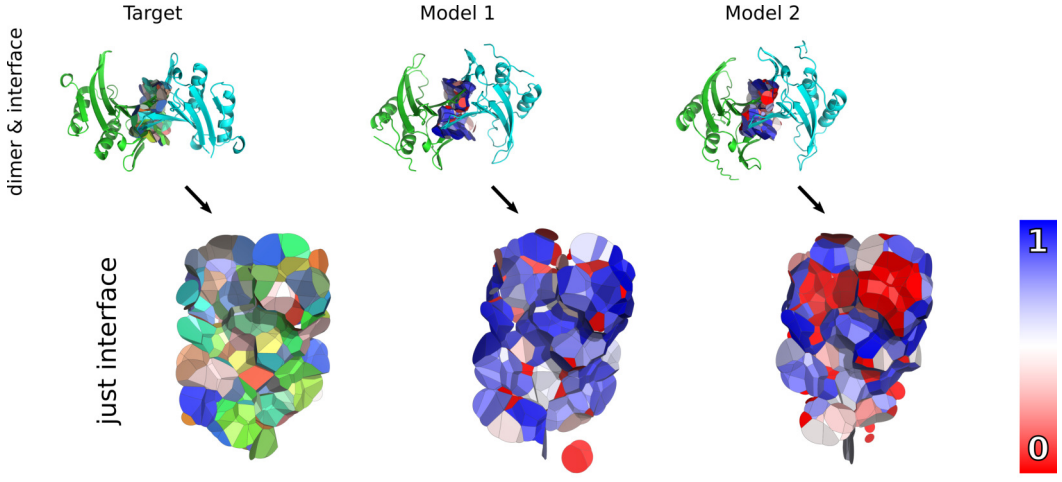
The intersubunit interface shown together with both subunits represented as cartoons



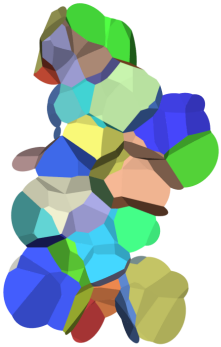
Same chains can have differently modelled interfaces



Comparing interfaces using CAD-score (Contact Area Difference score)



Evaluating interfaces with VoroMQA



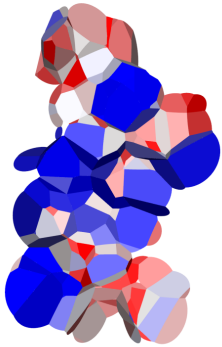
Interface
contact areas

$$\mathrm{E}(a_i, a_j, c_k) = \log \frac{\mathrm{P}_{\mathrm{exp}}(a_i, a_j, c_k)}{\mathrm{P}_{\mathrm{obs}}(a_i, a_j, c_k)} =$$

$$= \log \frac{\mathrm{F}_{\mathrm{exp}}(\mathrm{area}(a_i), \mathrm{area}(a_j), \mathrm{area}(c_k))}{\mathrm{F}_{\mathrm{obs}}(\mathrm{area}(a_i, a_j, c_k))}$$

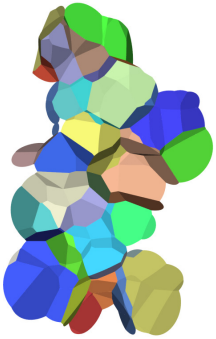
$$\mathrm{E_n}(\Omega_\phi) = \frac{\sum_{\omega \in \Omega_\phi} \mathrm{E}(\mathrm{type}_\omega) \cdot \mathrm{area}_\omega}{\sum_{\omega \in \Omega_\phi} \mathrm{area}_\omega}$$

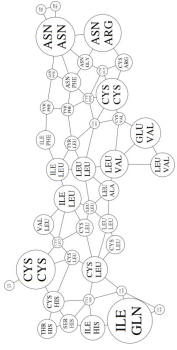
Statistical potential
for contact areas



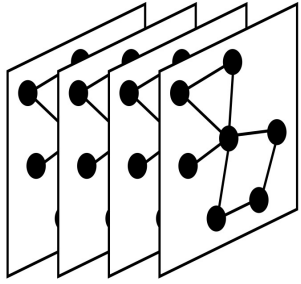
Interface
pseudo-energy

Evaluating interfaces with VoroIF-GNN

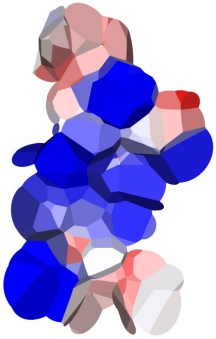




Interface contacts graph

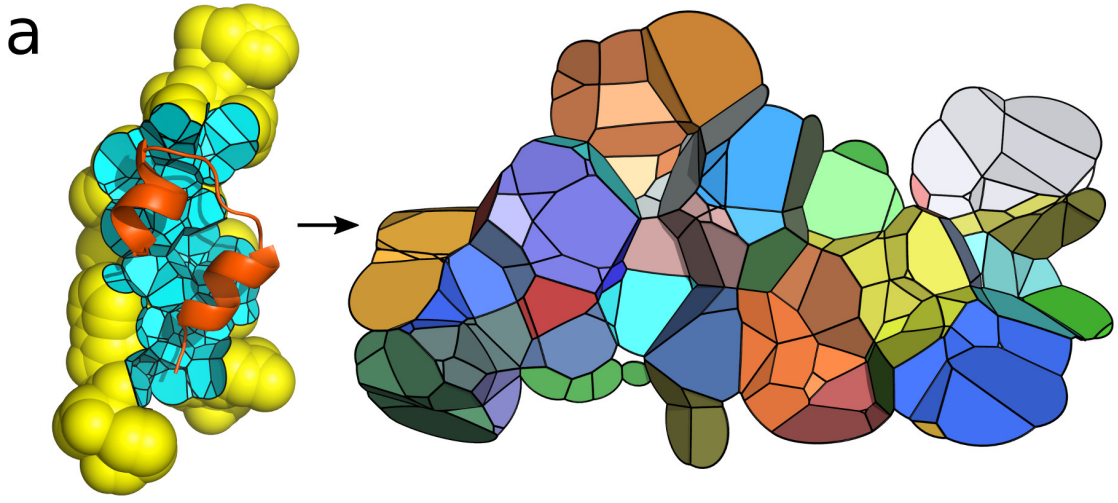


Graph attention neural network



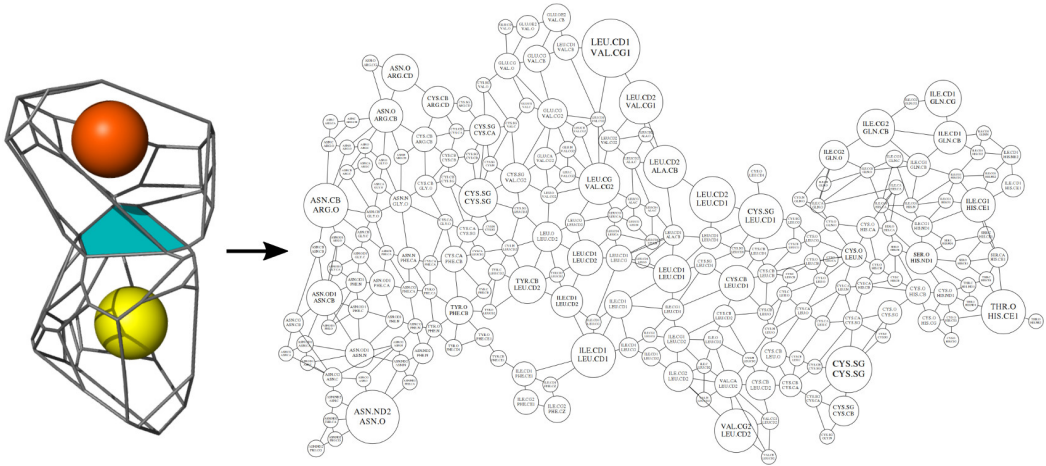
Predicted CAD-score

Interface graph — source

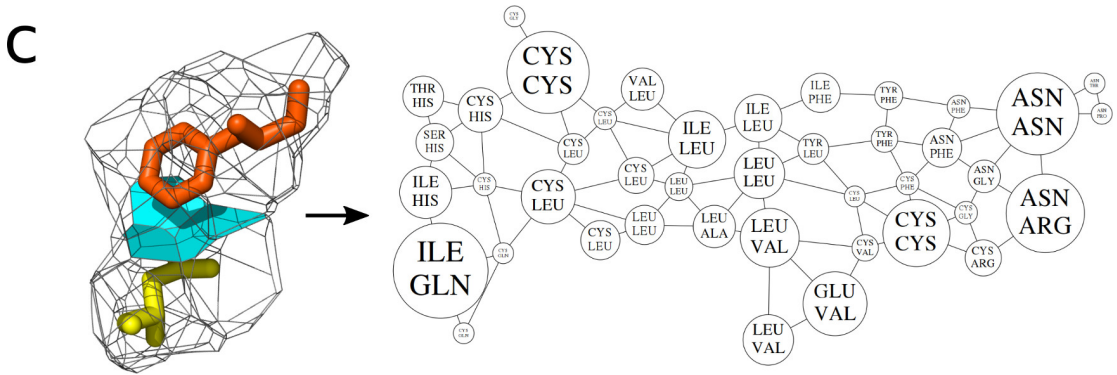


Interface graph — atom-atom level

b

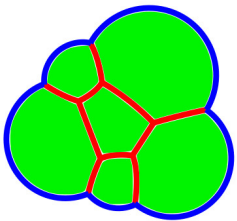


Interface graph — residue-residue level



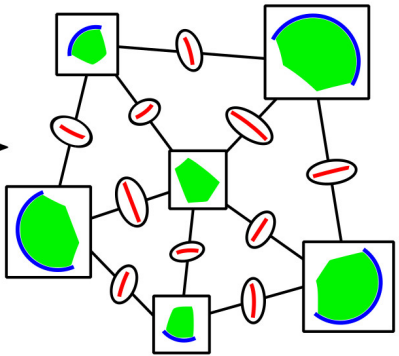
Input interface graph annotation

Tessellation-derived interface contacts



Contact surface
Contact-solvent border
Inter-contact border

Interface graph



Graph **node** attributes (15 values)

Contact surface area
Contact-solvent border length
Sum of inter-contact border lengths
Contact type-dependent descriptors (12 values)

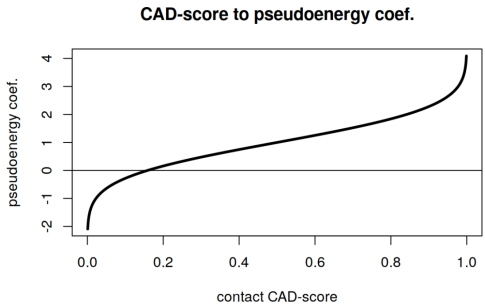
Graph **edge** attribute (1 value)

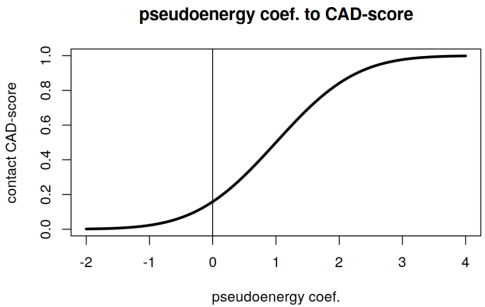
Inter-contact border length

What values to predict for graph nodes

Node level scores must behave like a "pseudoenergy":

- must be "summable", i.e. a sum of node scores should represent a more global score
- must be weighted by corresponding contact areas
- very bad scores must penalize the total sum





Generating datasets

Training/testing/validation sets were constructed as follows:

- 1567 native heterodimers were downloaded from PDB (Protein Data Bank) and split into training/validation/testing sets containing 1097/235/235 heterodimers
- each native structure (target) was redocked and a set of models of varying quality was selected (about 15-20 models for a target), e.g.:

ID	x	y	z	a1	a2	a3	cadscore	site_cadscore
1E50_2250	-7	27	4	45	153	90	0.74375	0.87635
1E50_32	-13	25	2	18	153	90	0.63728	0.75543
1E50_2735	-7	28	1	72	162	120	0.53173	0.68644
1E50_15946	-16	26	-2	45	162	120	0.38075	0.55364
1E50_10393	-16	28	5	0	153	90	0.24134	0.47034
1E50_3759	7	29	7	351	117	40	0.13939	0.51889
1E50_17192	24	22	8	315	63	0	0.0386	0.42122
1E50_15006	-13	27	13	342	18	0	0	0.40432
1E50_5533	28	-13	20	0	45	204	0	0.30295
1E50_14280	27	-22	-22	180	126	60	0	0.20266
1E50_532	34	4	-18	207	54	100	0	0.10126
1E50_20368	1	-39	10	324	117	80	0	0.00119
1E50_9297	37	5	-22	261	54	80	0	0

GNN architecture selection and application

Initial ideas for the graph neural network (GNN):

- train to predict node scores (i.e. train to minimize MSE loss between predicted and ground truth CAD-score pseudoenergies)
- use both node and edge features in an attention mechanism
- in the validation stage, judge GNN performance by assessing how a global score (equal to the sum of node predictions) is able to select the best multimeric model out of many

GATv2

A multilayer GNN based on on GATv2 convolutional operator (Brody and Yahav, 2021) was chosen, because in GATv2 the edge features are used straightforwardly when computing attention coefficients.

In GATv2, every node attends to all its neighbors:

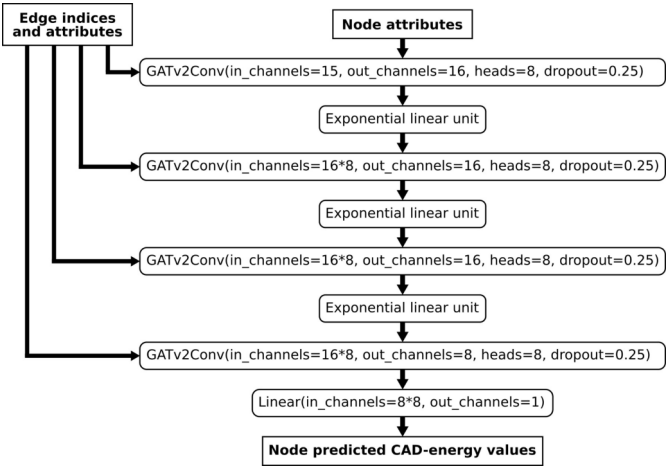
$$\mathbf{x}'_i = \alpha_{i,i}\mathbf{\Theta}_s\mathbf{x}_i + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}\mathbf{\Theta}_t\mathbf{x}_j,$$

where the attention coefficients are computed as

$$\alpha_{i,j} = \frac{\exp\left(\mathbf{a}^\top \mathrm{LeakyReLU}\left(\mathbf{\Theta}_s\mathbf{x}_i + \mathbf{\Theta}_t\mathbf{x}_j + \mathbf{\Theta}_e\mathbf{e}_{i,j}\right)\right)}{\sum_{k \in \mathcal{N}(i)\cup\{i\}} \exp\left(\mathbf{a}^\top \mathrm{LeakyReLU}\left(\mathbf{\Theta}_s\mathbf{x}_i + \mathbf{\Theta}_t\mathbf{x}_k + \mathbf{\Theta}_e\mathbf{e}_{i,k}]\right)\right)}.$$

Selected GNN architecture

ML experiments resulted in selecting the following 4-layer GATv2 architecture:



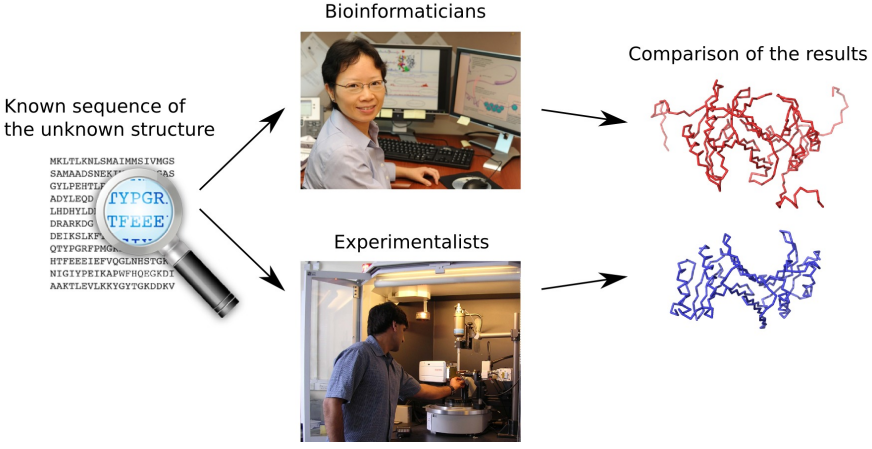
In-house testing results

Performance of the final method on a 235 sets of dimeric models generated by redocking and not used in training:

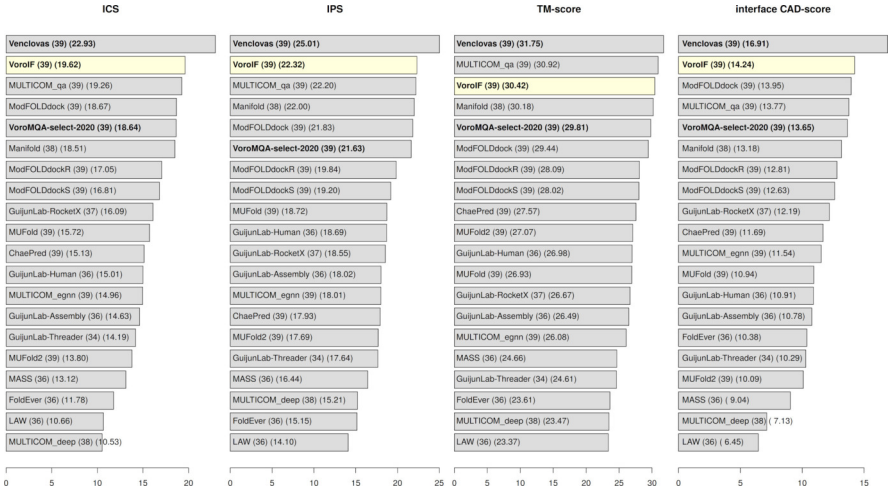
Selection method	**Rate of correct top 1**	**Mean interface CAD-score**	**Mean z-score of interface CAD-score**
Ideal	100%	0.78	1.85
VoroIF-GNN (new)	86%	0.74	1.72
VoroMQA energy (old)	53%	0.63	1.34

Double-blind testing

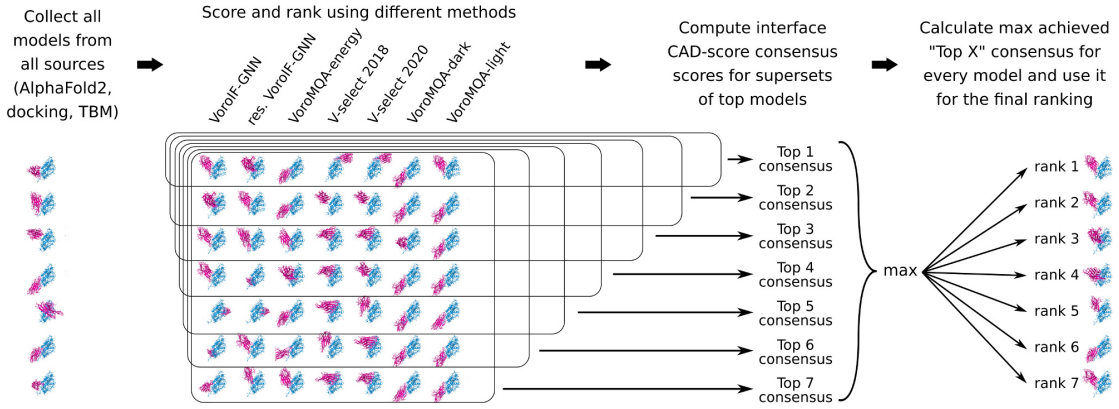
The main way to test new methods in structural bioinformatics are the community-wide double-blind testing experiments — CASP and CAPRI.



VoroIF-GNN results in CASP15 (2022)



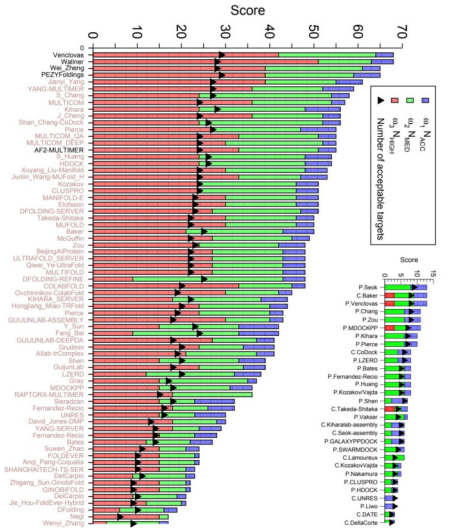
Selecting complex models using VoroIF-jury



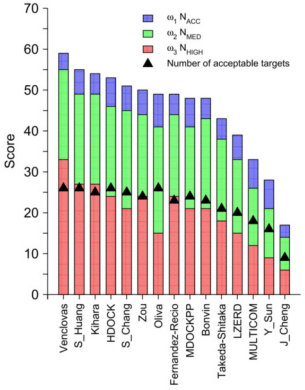
VoroIF-jury was the best-performing multimeric model selector in 2022 CASP and CAPRI double-blind challenges.

CASP15-CAPRI challenge results

Assembly prediction results



Assembly scoring results



Plots from Lensink et al. (2023) "Impact of AlphaFold on Structure Prediction of Protein Complexes: The CASP15-CAPRI Experiment". Proteins (accepted)

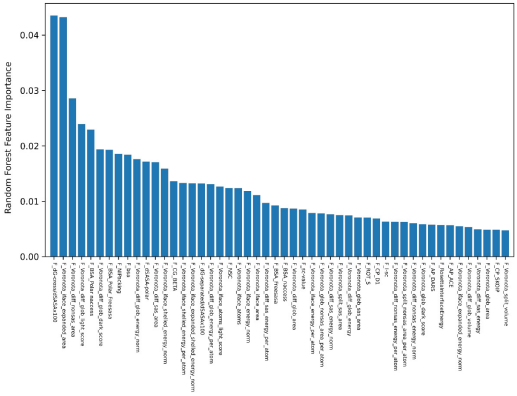
VoroIF-jury -> FTDMP

The VoroIF-jury method is the base of the FTDMP framework for protein-protein, protein-DNA and protein-RNA docking and scoring.

Rita Banciul from Vilnius University will talk about FTDMP this Thirsday at 15:00.

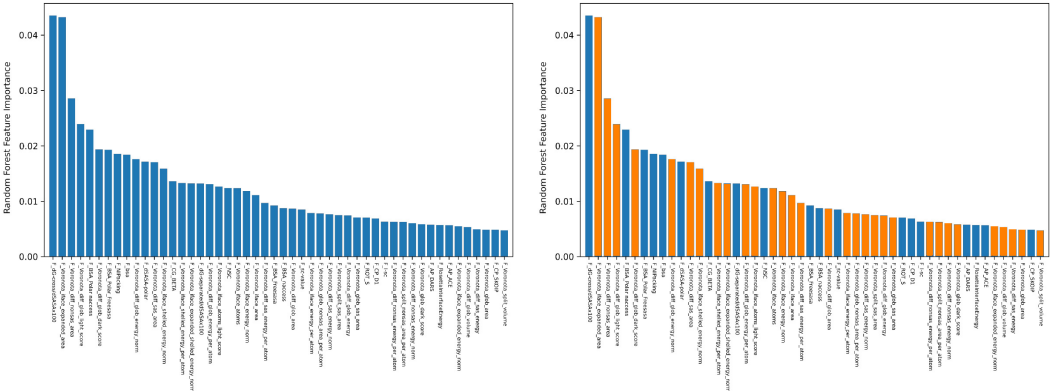
Discriminating Physiological from Non-Physiological Interfaces in Protein Complexes — top 50 useful features

- Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. Schweke H et al. *Proteomics*. **2023** Jun 27.



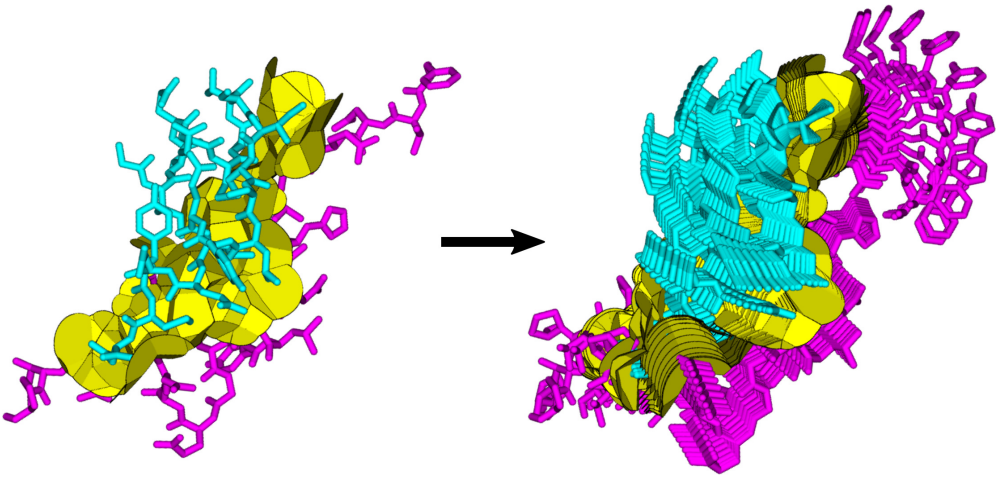
Discriminating Physiological from Non-Physiological Interfaces in Protein Complexes — top 50 useful features

- Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. Schweke H et al. *Proteomics*. **2023** Jun 27.



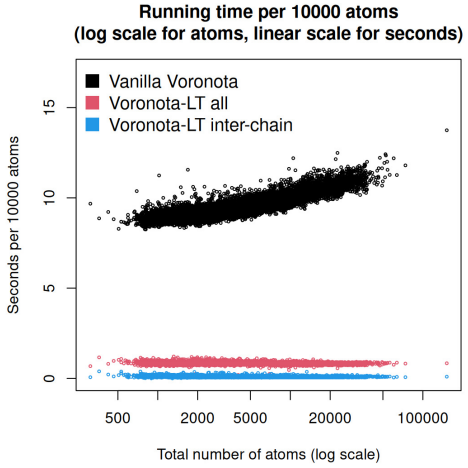
What's next

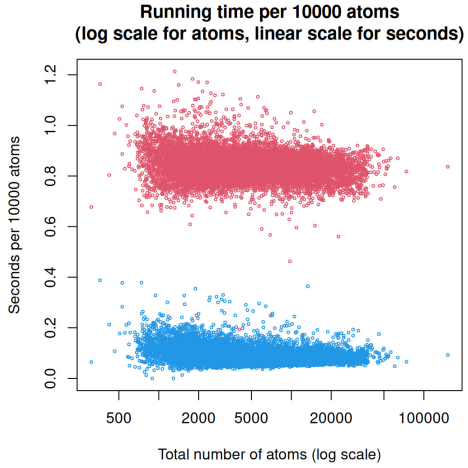
Tessellation-based analysis of dynamic protein structures and their complexes:



Voronota-LT

Voronota-LT is an alternative, significantly faster version of Voronota for constructing tessellation-derived atomic contact areas and volumes:





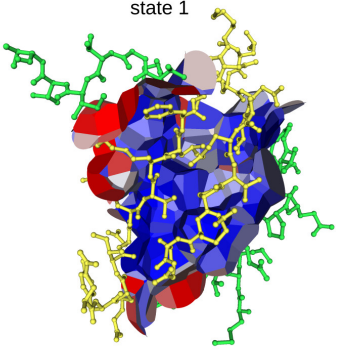
Example application of Voronota-LT

Studying inter-chain interface contact variability in an ensemble of structures:

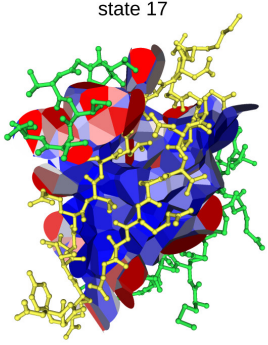
NMR ensemble (PDB 1CIR) of 20 states with inter-chain contact contours

Inter-chain interface colored by relative variance of **atom-atom** contact areas

state 1

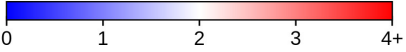


state 17



Chain A

Chain B



Relative variance (variance-to-mean ratio)

Related papers

- VoroIF-GNN: Voronoi tessellation-derived protein-protein interface assessment using a graph neural network. Olechnovič K, Venclovas Č. *Proteins*. **2023** Jul 21.

- Prediction of protein assemblies by structure sampling followed by interface-focused scoring. Olechnovič K, Valančauskas L, Dapkūnas J, Venclovas Č. *Proteins*. **2023** Aug 14.

- (preprint) Voronota-LT: efficient, flexible and solvent-aware tessellation-based analysis of atomic interactions. Olechnovič K, Grudinin S. *bioRxiv*. **2024** Feb 5.

Thank you!

Vilnius University CASP15-CAPRI team:

- Justas Dapkūnas
- Lukas Valančauskas
- Česlovas Venclovas

CNRS Laboratoire Jean Kuntzmann:

- Sergei Grudinin

Useful links:

- https://www.voronota.com
- https://www.kliment.lt






Funded by
the European Union