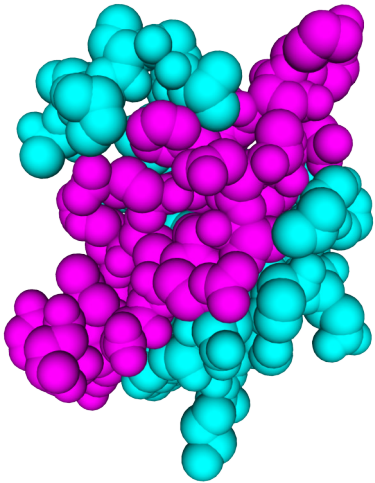
Voronoi tessellation-based analysis of 3D conformations of non-globular proteins

Dr. Kliment Olechnovič

CNRS Laboratoire Jean Kuntzmann, Grenoble, France

Vilnius University Life Sciences Center, Vilnius, Lithuania

2024-05-17



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

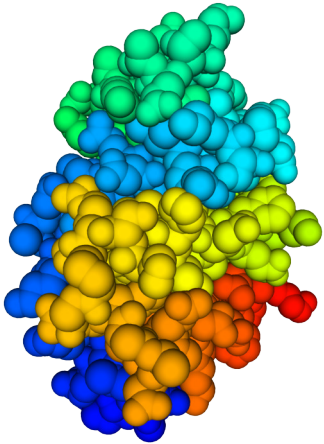
Today's question

Can we describe and score protein-protein interactions using descriptors derived from

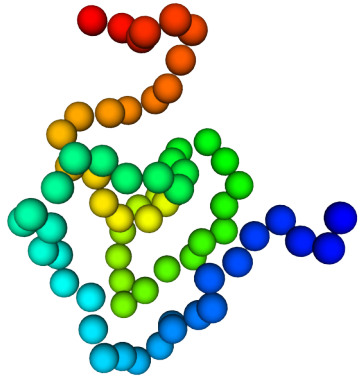
- ensembles of experimentally determined protein conformations?
- ensembles of simulated conformations of disordered proteins?

Data of molecular conformations

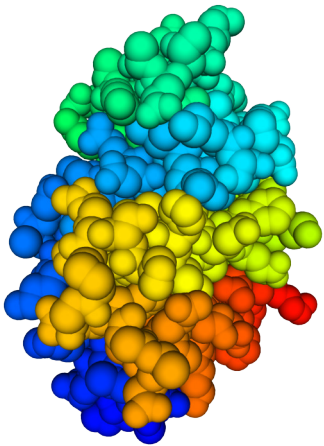
Protein Data Bank (PDB) data



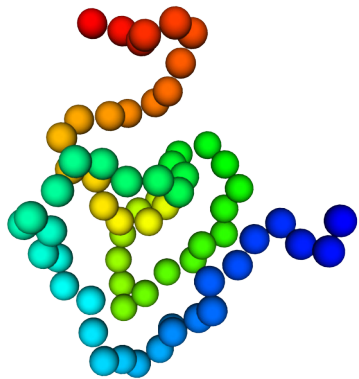
Simulated intrinsically disordered protein (IDP) data



PDB

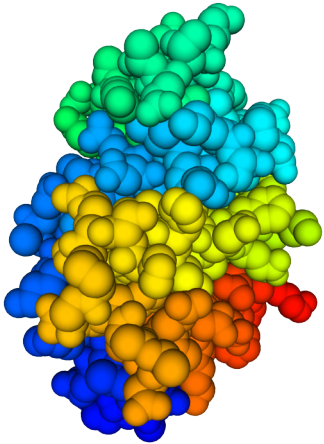


IDP

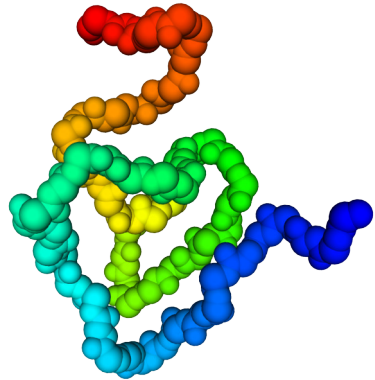


C-alpha atoms

PDB

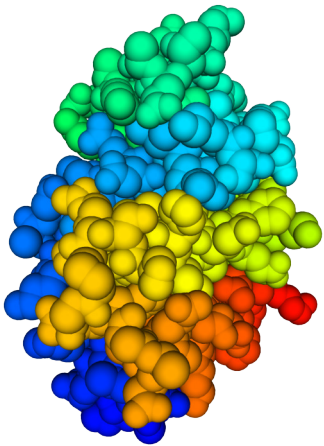


IDP

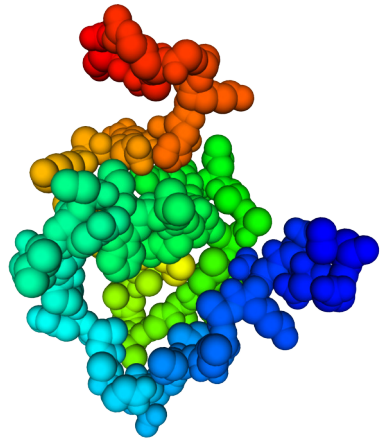


backbone atoms built by Pulchra

PDB

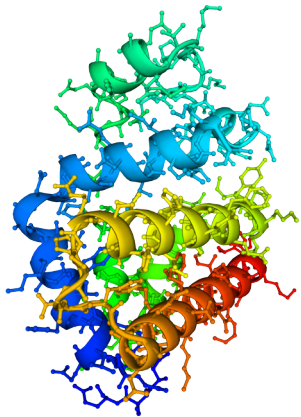


IDP

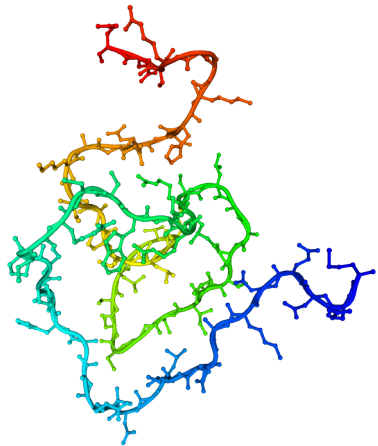


side-chain atoms built by FASPR

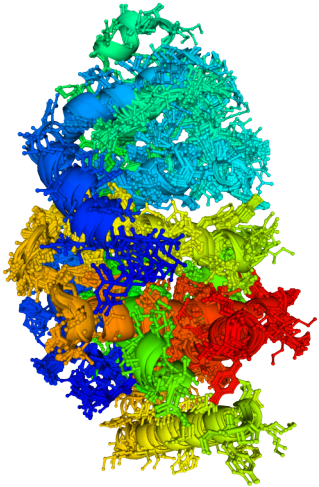
PDB



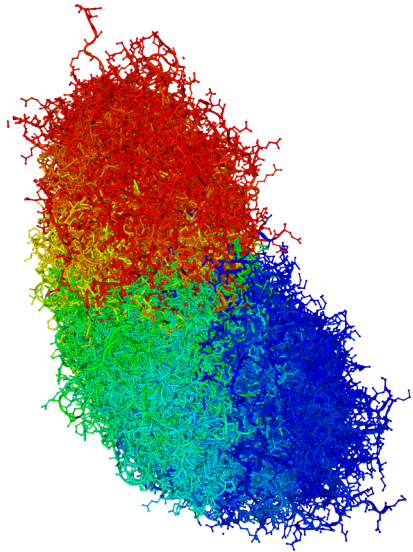
IDP



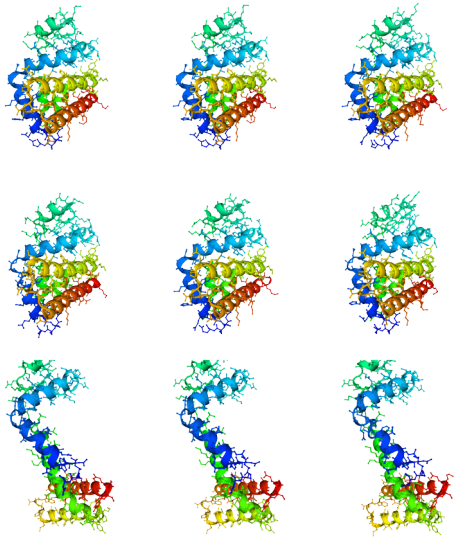
PDB ensemble



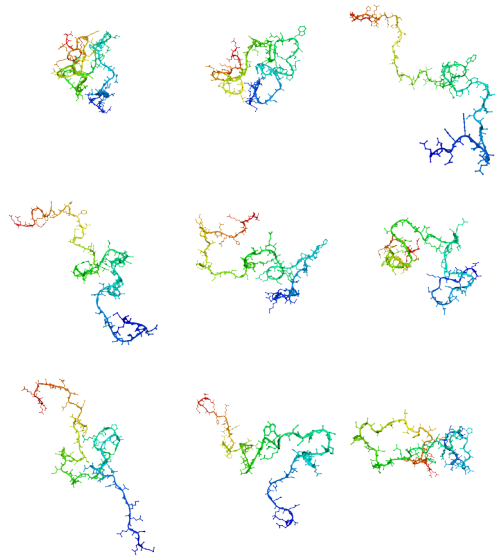
IDP ensemble



PDB ensemble



IDP ensemble



PDB dataset

- From the Protein Data Bank, wwwpdb.org
- Ensembles formed by clustering chain sequences using 90% identity
- We used all **38'807** ensembles
- Ensembles have very different numbers of chains, largest ensemble contains 1413 chains, 9989 ensembles contain only two chains, there are **429'945** protein chains in total.

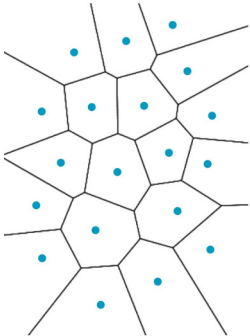
IDP dataset **"IDRome"**

- From Tesei, Trolle, Jonsson et al. *Conformational ensembles of the human intrinsically disordered proteome*. Nature (2024)
- Ensembles generated by running coarse-grained simulations using CALVADOS for 28'058 IDP-like protein sequences
- We used all **16'774** ensembles for chains of 60 to 600 residues in length
- Every ensemble has 1010 conformations, so there **16'941'740** conformations in total

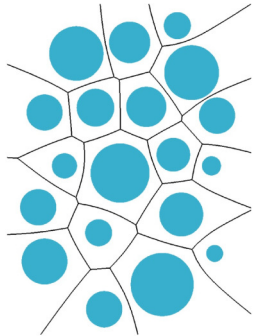
Describing interactions in molecular conformations using the Voronoi tessellation

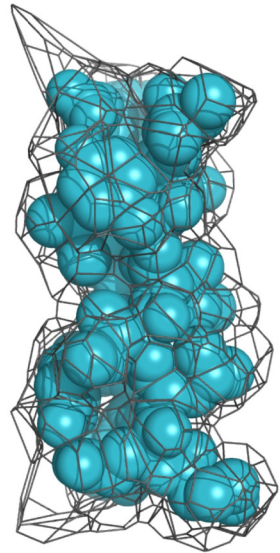
Voronoi diagram of points and balls

"Classic" Voronoi diagram of points



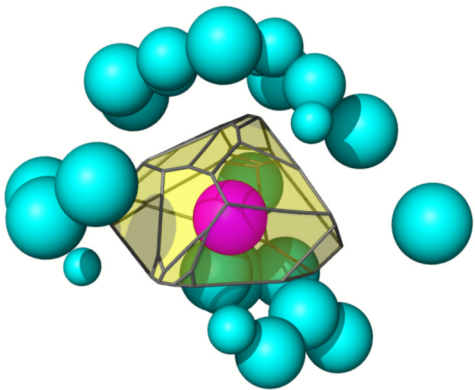
Voronoi diagram of balls



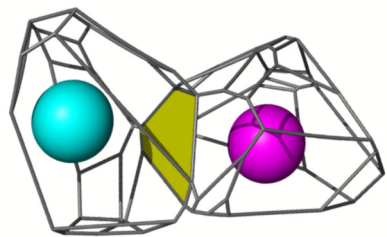


Voronoi tessellation-based analysis of structures

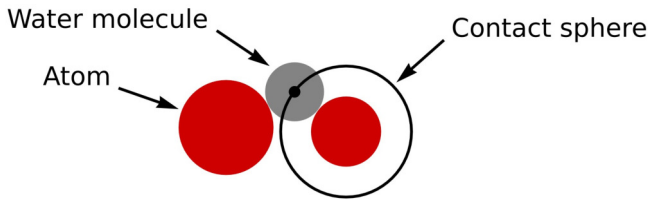
Voronoi cell of an atom surrounded by its neighbors

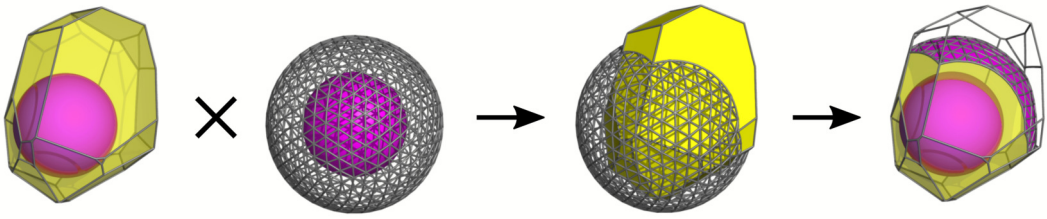


Atom-atom contact surface defined as the face shared by two adjacent Voronoi cells.



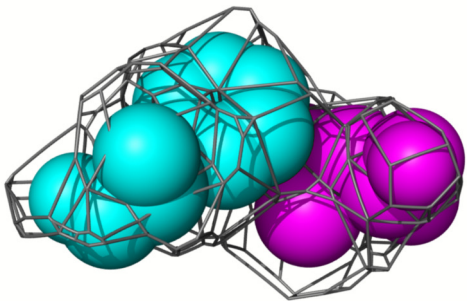
Constrained contacts



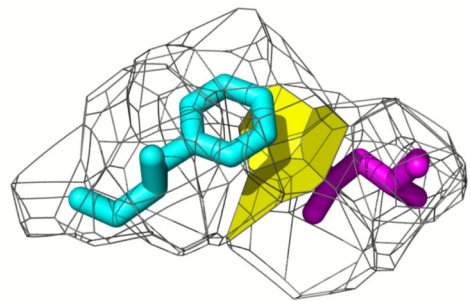


Deriving residue-residue contacts

Voronoi cells of two neighboring residues

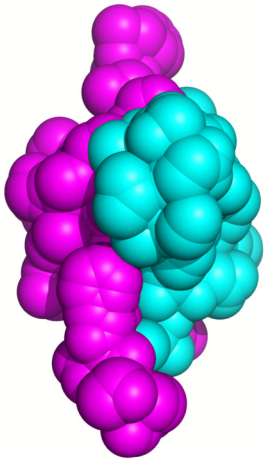


Residue-residue contact surface defined as a union of atom-atom contact surfaces

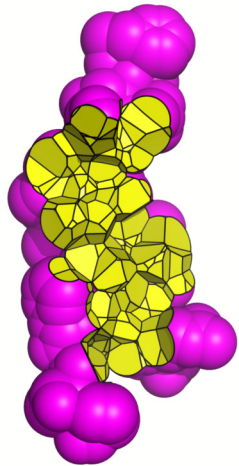


Inter-chain contacts

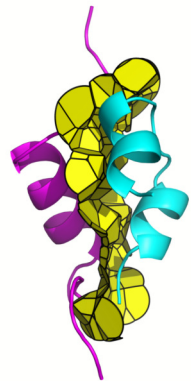
Solvent-accessible surface of an insulin heterodimer PDB:4UNG colored by subunit



The intersubunit interface shown together with the SAS of one subunit

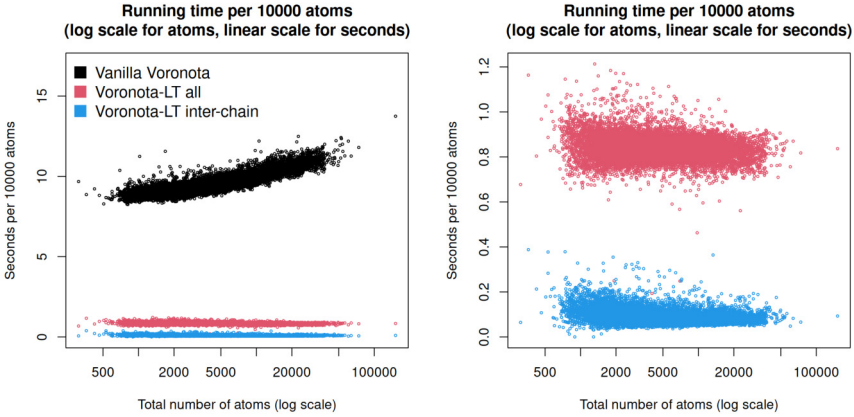


The intersubunit interface shown together with both subunits represented as cartoons



Voronota-LT

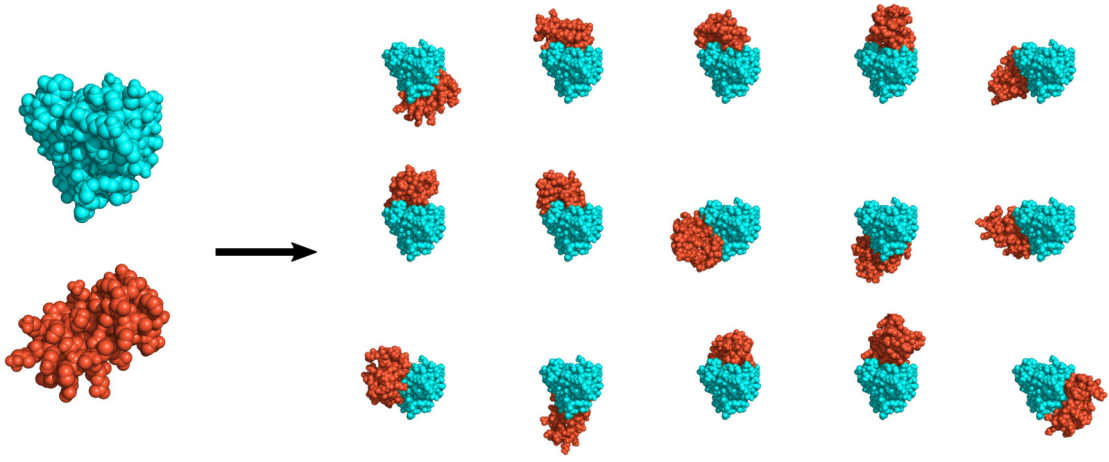
Voronota-LT is a new fast software for constructing tessellation-derived atomic contact areas and volumes. It is significantly faster than its predecessor, Voronota:



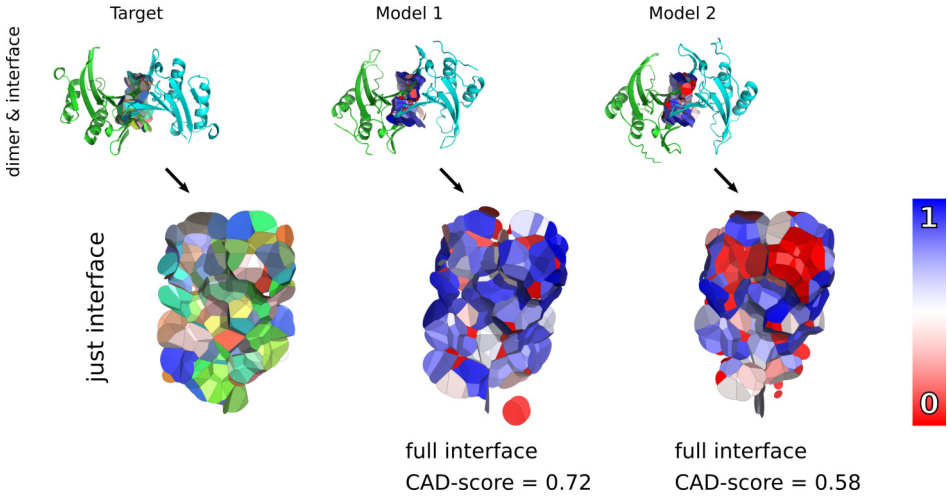
Olechnovic and Grudinin. *Voronota-LT: efficient, flexible and solvent-aware tessellation-based analysis of atomic interactions*. bioRxiv (2024)

An application of tessellation-based description of interactions

Same chains can have differently modelled interfaces



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



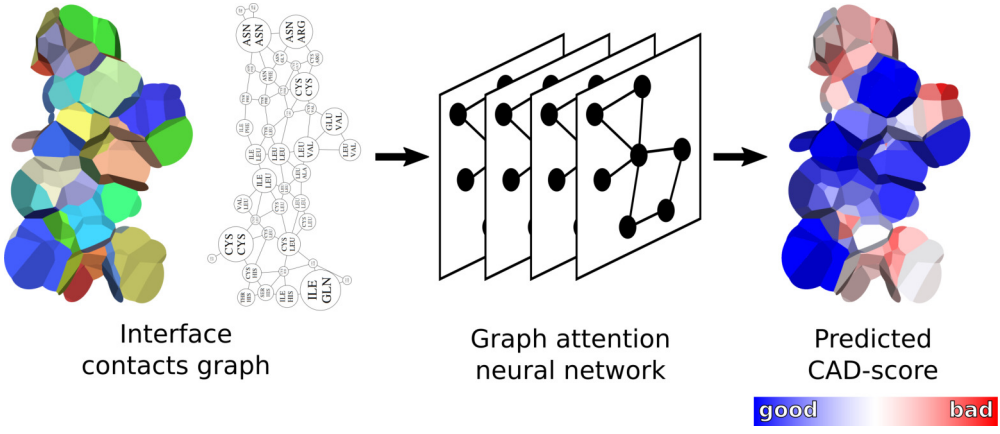
Olechnovic and Venclovas. *Contact Area-Based Structural Analysis of Proteins and Their Complexes Using CAD-Score*. Methods in Molecular Biology (2020)

A dataset of correct and incorrect interfaces

- A non-redundant set of 1567 native heterodimers, selected using PPI3D and downloaded from PDB.
- Each native structure (target) was redocked and a set of models of varying quality was selected (about 15-20 models for a target), for example:

ID	x	y	z	a1	a2	a3	cadscore	site_cadscore
1E50_nat	0	0	0	0	0	0	1	1
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

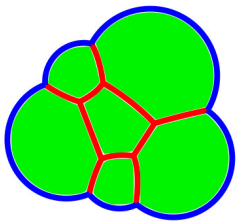
Evaluating interfaces with a graph neural network (e.g. VoroIF-GNN)



Olechnovic and Venclovas. *VoroIF-GNN: Voronoi tessellation-derived protein-protein interface assessment using a graph neural network*. Proteins (2023)

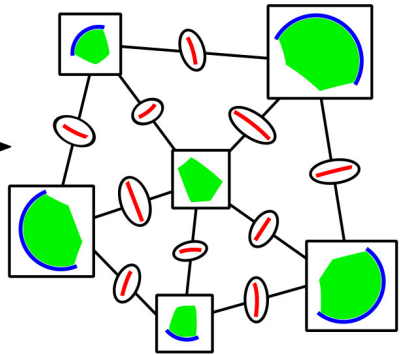
Input interface graph annotation in VoroIF-GNN

Tessellation-derived interface contacts



Contact surface
Contact-solvent border
Inter-contact border

Interface graph



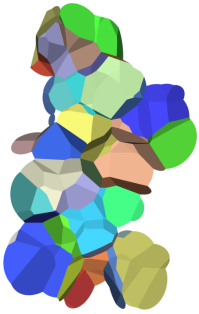
Graph **node** attributes (15 values)

- **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

Evaluating interfaces with an area-based potential (e.g. VoroMQA)



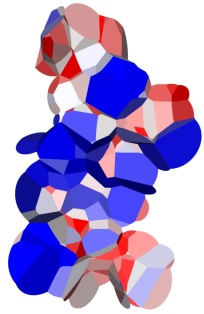
Interface
contact areas

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

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

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

Statistical potential
for contact areas



Interface
pseudo-energy

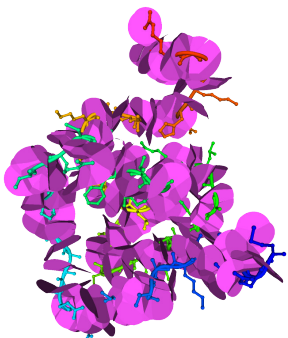
Olechnovic and Venclovas. *VoroMQA: Assessment of protein structure quality using interatomic contact areas*. Proteins (2017)

Deriving and using statistics of contact areas from ensembles of conformations

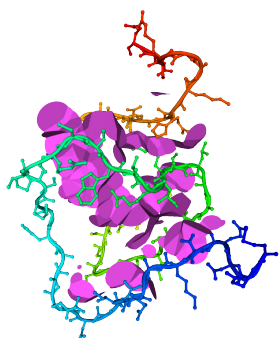
Contacts from a single conformation

A contact type is a tuple *(first atom type, second atom type, contact category)* = (a_1, a_2, c).

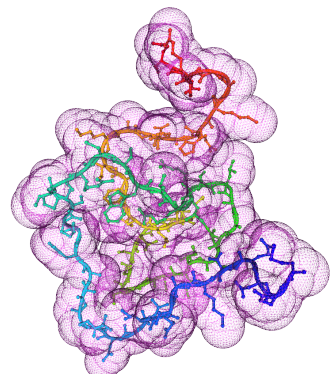
sequence separation <=5



sequence separation >5

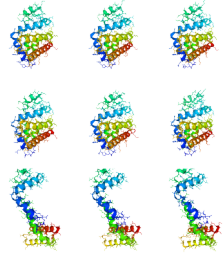


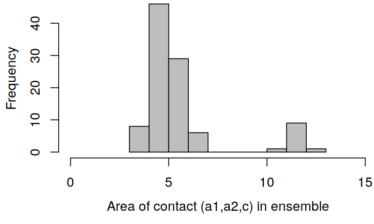
solvent-accessible surface



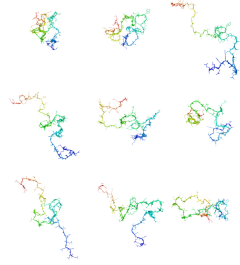
Contact areas from a single ensemble of conformations

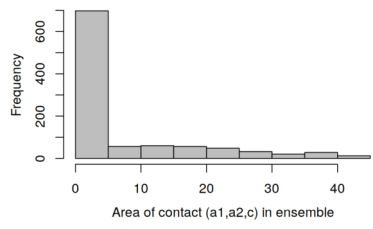
PDB ensemble





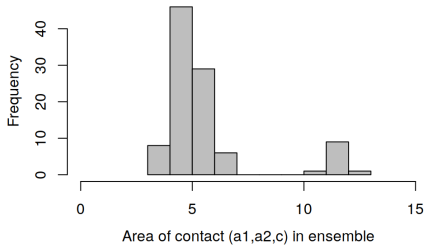
IDRome ensemble



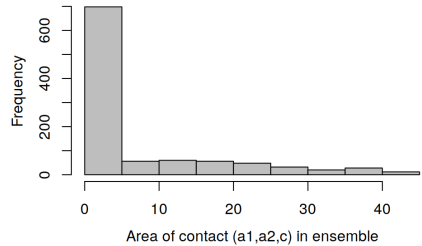


Contact areas from a single ensemble of conformations

PDB ensemble



IDRome ensemble



We summarize a contact type $t = (a_1, a_2, c)$ area distribution in a PDB ensemble with:

- v^t = min(observed t areas)
- u^t = max(observed t areas)

We summarize a contact type $t = (a_1, a_2, c)$ area distribution in an IDP ensemble with:

- v^t = mean(observed t areas)
- u^t = max(observed t areas)

Areas of contact types from a multiple ensembles of conformations

v^t and u^t values are areas, therefore we can sum them.

For a contact type $t = (a_1, a_2, c)$ we sum the relevant v^t and u^t values from all the available ensembles G to get V^t and U^t sums:

$$V^t = \sum_{g \in G} v^t(g) \tag{1}$$

$$U^t = \sum_{g \in G} u^t(g) \tag{2}$$

We do it for every contact type t from the set of all possible contact types T.

Observed probabilities of areas of contact types

Observed probability estimate of contact area unit of type $t = (a_1, a_2, c)$ to occur:

$$P^t_{\text{obs}}(\text{occur}) = \frac{V^t + U^t}{\sum_{s \in T}(V^s + U^s)} \tag{3}$$

Observed conditional probability estimate of contact area unit to persist:

$$P^t_{\text{obs}}(\text{persist}|\text{occur}) = \frac{2V^t}{V^t + U^t} \tag{4}$$

Observed probability estimate of contact area unit to occur and persist:

$$P^t_{\text{obs}}(\text{occur and persist}) = P^t_{\text{obs}}(\text{occur}) \cdot P^t_{\text{obs}}(\text{persist}|\text{occur}) \tag{5}$$

Expected probabilities of areas of contact types

Expected probability estimate of contact area unit of type $t = (a_1, a_2, c)$ to occur (modeling the situation where there are no atom type-dependent or contact category-dependent effects):

$$P_{\text{exp}}^{t=(a_1,a_2,c)}(\text{occur}) \sim P_{\text{obs}}^{(a_1,*,*)}(\text{occur}) \cdot P_{\text{obs}}^{(*,a_2,*)}(\text{occur}) \cdot P_{\text{obs}}^{(*,*,c)}(\text{occur}) \cdot \tag{6}$$

Expected conditional probability estimate of contact area unit to persist:

$$P_{\text{exp}}^{t}(\text{persist}|\text{occur}) = \frac{2 \cdot \sum_{s \in T} V^s}{\sum_{s \in T}(V^s + U^s)} \tag{7}$$

Expected probability estimate of contact area unit to occur and persist:

$$P_{\text{exp}}^{t}(\text{occur and persist}) = P_{\text{exp}}^{t}(\text{occur}) \cdot P_{\text{exp}}^{t}(\text{persist}|\text{occur}) \tag{8}$$

Deriving pseudo-energy coefficient from probality estimates

Pseudo-energy coefficient for a contact area unit of type $t = (a_1, a_2, c)$:

$$E^t \sim \log\left(\frac{P^t_{\text{exp}}(\text{occur and persist})}{P^t_{\text{obs}}(\text{occur and persist})}\right) \tag{9}$$

E^t can be written as a weighted sum (weights to be optimized later):

$$E^t = \alpha_1 \cdot \log P^t_{\text{obs}}(\text{occur}) + \alpha_2 \cdot \log P^t_{\text{exp}}(\text{occur}) +$$
$$+\alpha_3 \cdot \log P^t_{\text{obs}}(\text{persist}|\text{occur}) + \alpha_4 \cdot \log P^t_{\text{exp}}(\text{persist}|\text{occur}) + \beta$$

Using pseudo-energy to score inter-chain interfaces

A total pseudo-energy score for a set of contacts K is:

$$S_{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\beta}(G) = \sum_{k \in K} \text{area}(k) \cdot E^{\text{type}(k)}_{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\beta} \tag{10}$$

We used 70% of the docking model sets from our interface decoys dataset to grid-search (primitively, but exhaustively, using a step of 0.1) for the best combination of weighting coefficients $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta)$ for the task of selecting well-modelled interfaces.

We used the remaining 30% of the docking model sets for testing.

Results of inter-chain interfaces scoring

Method	Data	Granularity	Components	Mean CAD-score
Ideal selector				1
Random				0.25
Pseudo-energy	PDB	atom-atom	$P(\text{occur})$	0.78
			$P(\text{persist}\|\text{occur})$	0.78
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	**0.89**
Pseudo-energy	PDB	residue-residue	$P(\text{occur})$	0.59
			$P(\text{persist}\|\text{occur})$	0.59
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.62
Pseudo-energy	IDRome	residue-residue	$P(\text{occur})$	0.50
			$P(\text{persist}\|\text{occur})$	0.53
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.55

Results of inter-chain interfaces scoring

Method	Data	Granularity	Components	Mean CAD-score
Ideal selector				1.00
Random				0.25
Pseudo-energy	PDB	atom-atom	$P(\text{occur})$	0.78
			$P(\text{persist}\|\text{occur})$	0.78
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	**0.89**
Pseudo-energy	PDB	residue-residue	$P(\text{occur})$	0.59
			$P(\text{persist}\|\text{occur})$	0.59
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.62
Pseudo-energy	IDRome	residue-residue	$P(\text{occur})$	0.50
			$P(\text{persist}\|\text{occur})$	0.53
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.55
VoroIF-GNN	all	hybrid	all	**0.98**

Conclusion

- Ensembles of conformations from PDB provide useful information about contact stability, it can improve scoring protein-protein interfaces.
- Ensembles of simulated conformations of IDPs can also be useful as a source of statistics about tesselation-derived amino acid interactions.
- Different statistical descriptors can be efficiently employed using tessellation-based graph neural network.

Thank you!

CNRS Laboratoire Jean Kuntzmann:

- Sergei Grudinin

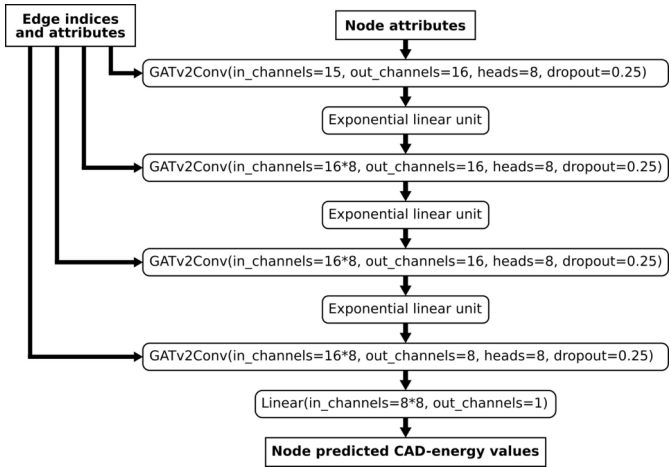
Useful links:

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



Funded by
the European Union


VoroIF-GNN architecture



GATv2 convolutional operator

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.

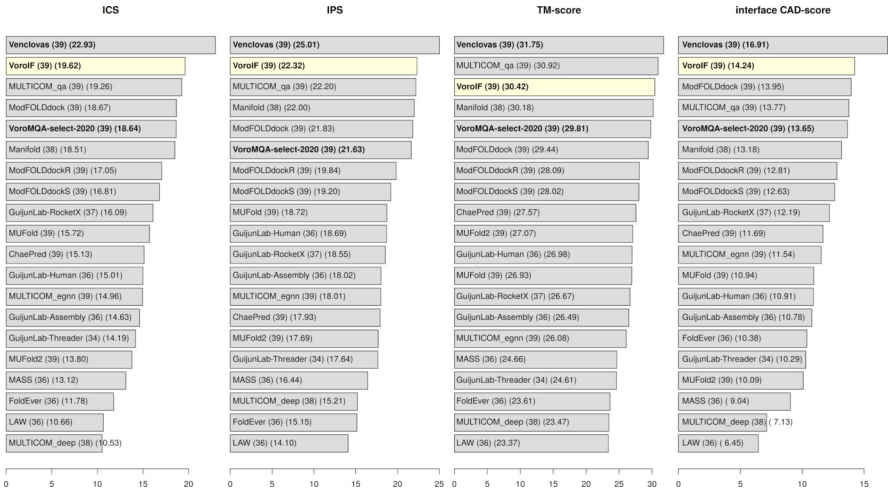
In GATv2, every node attends to all its neighbors:

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

where the attention coefficients are computed as

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

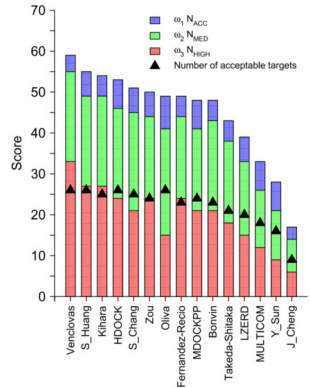
VoroIF-GNN results in CASP15 (2022)



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)

More results of inter-chain interfaces scoring

When there are no ideal models:

Method	Data	Granularity	Components	Mean CAD-score
Ideal selector				0.78
Random				0.23
Pseudo-energy	PDB	atom-atom	$P(\text{occur})$	0.60
			$P(\text{persist}\|\text{occur})$	0.59
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	**0.64**
Pseudo-energy	PDB	residue-residue	$P(\text{occur})$	0.54
			$P(\text{persist}\|\text{occur})$	0.52
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.55
Pseudo-energy	IDRome	residue-residue	$P(\text{occur})$	0.49
			$P(\text{persist}\|\text{occur})$	0.51
			$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.50
VoroIF-GNN	all	hybrid	all	**0.77**