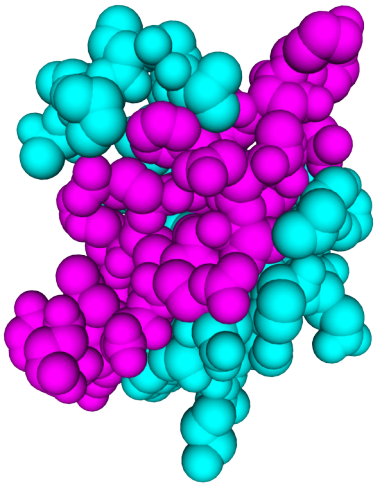
Evaluating models of protein-protein complexes using tessellation-based contact area persistence descriptors derived from conformational ensembles of proteins

Dr. Kliment Olechnovič

CNRS Laboratoire Jean Kuntzmann, Grenoble, France

Vilnius University Life Sciences Center, Vilnius, Lithuania

2025-03-20



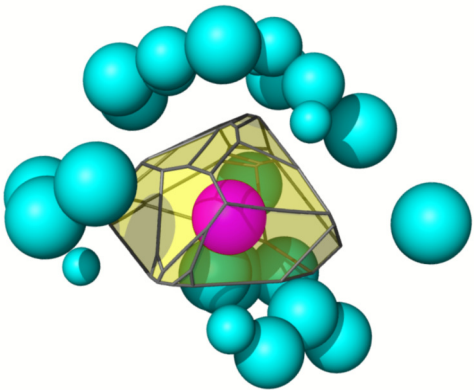
Common problems:

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

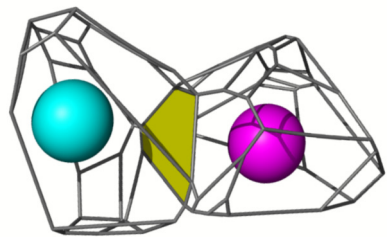
Describing interactions in molecular conformations using the Voronoi tessellation

Voronoi tessellation-based analysis of structures

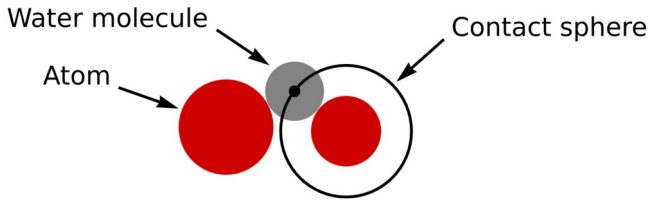
Voronoi cell of an atom surrounded by its neighbors

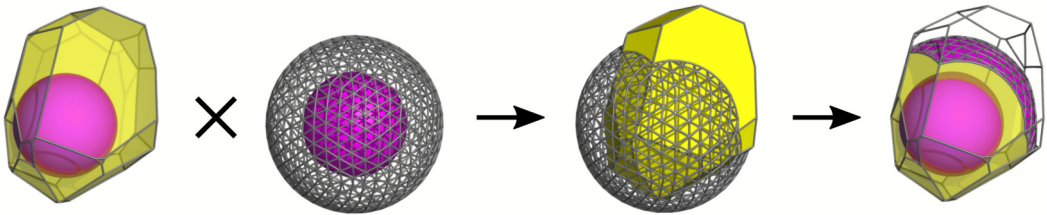


Atom-atom contact surface defined as the face shared by two adjacent Voronoi cells.



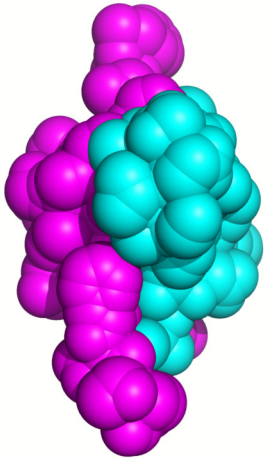
Constrained contacts



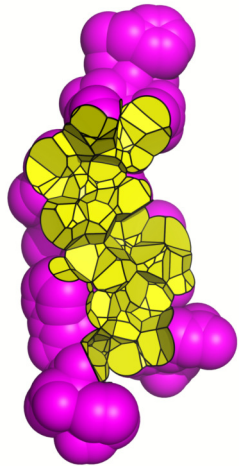


Inter-chain contacts

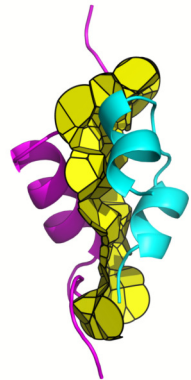
Solvent-accessible surface of an insulin heterodimer PDB:4UNG colored by subunit



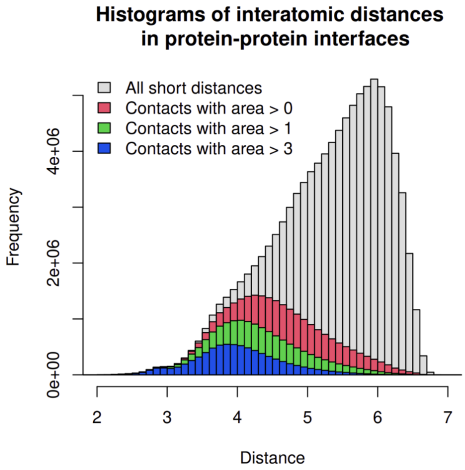
The intersubunit interface shown together with the SAS of one subunit

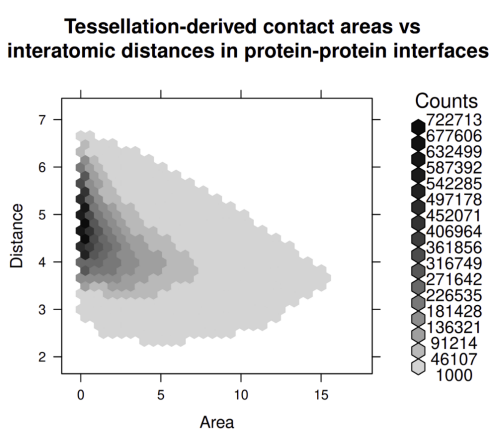


The intersubunit interface shown together with both subunits represented as cartoons



Inter-chain contact areas vs distances, PDB-based statistics

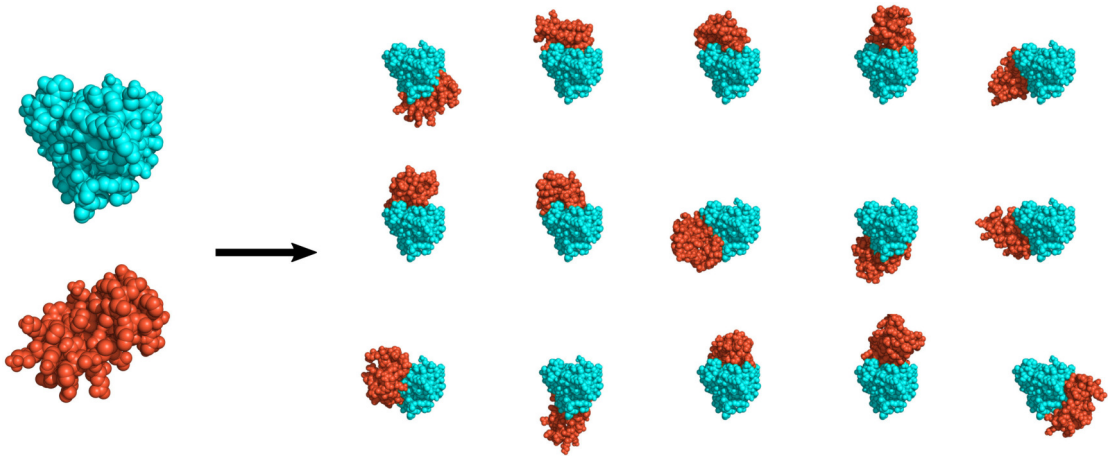




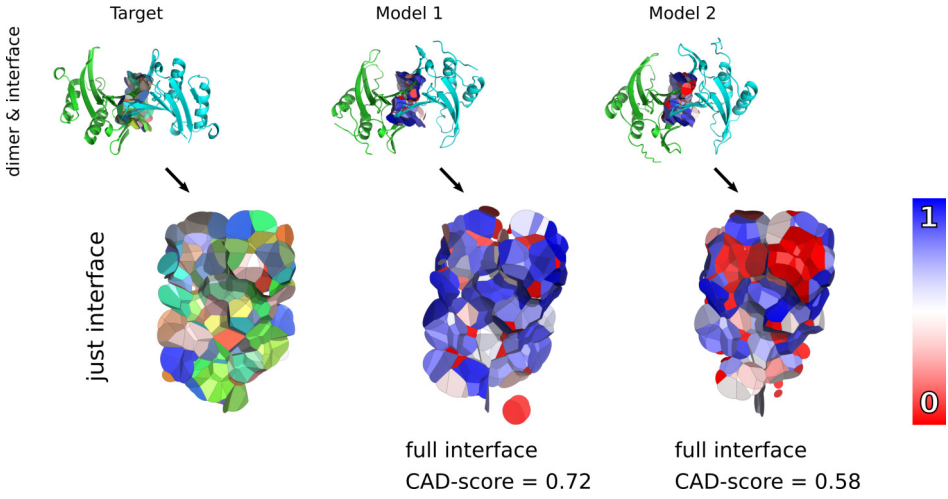
corr(area, distance) ≈ -0.43

Some applications 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)

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

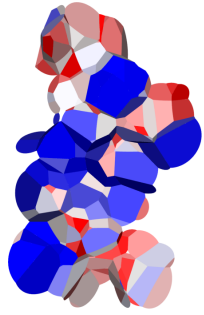
Interface
contact areas

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

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

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

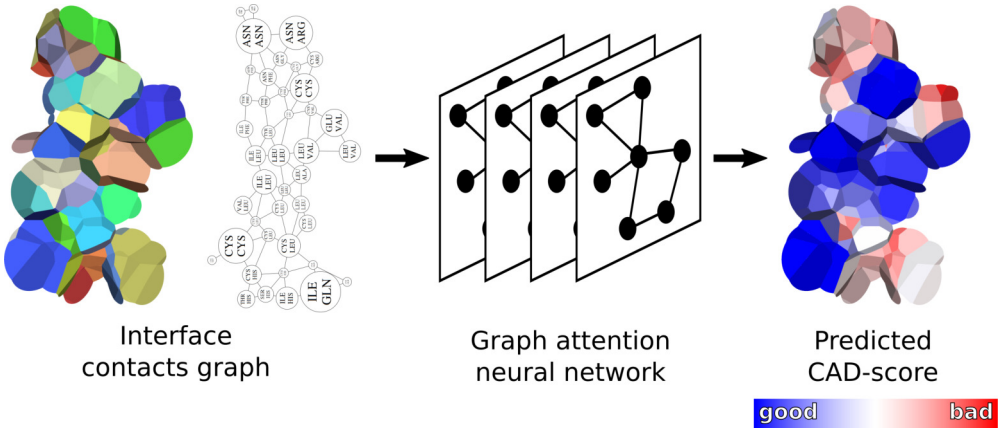
Statistical potential
for contact areas



Interface
pseudo-energy

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

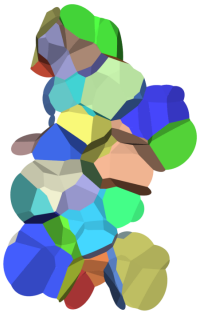
Evaluating interfaces with a graph neural network (e.g. VoroIF-GNN)



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

Area-based potential may still be used for testing new ideas

Area-based pairwise interaction potential alone is not the best scoring method, but it may still serve as simple a tool to explore benefits of newer data and descriptors.



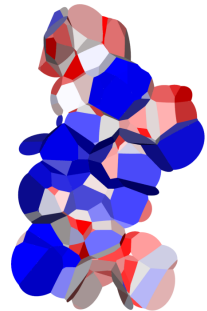
Interface
contact areas

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

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

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

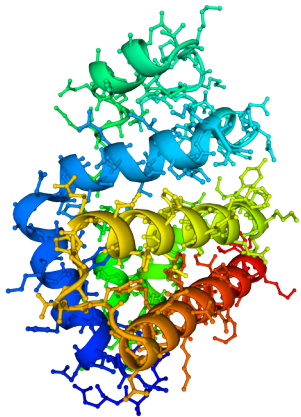
Statistical potential
for contact areas



Interface
pseudo-energy

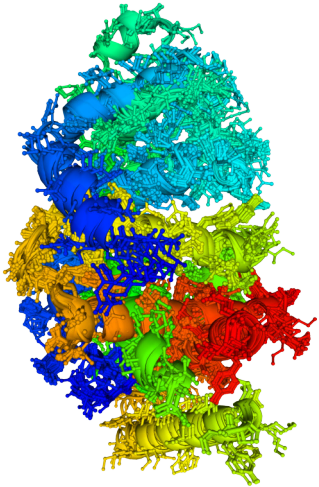
Deriving and using statistics of contact areas from ensembles of conformations

a single conformation

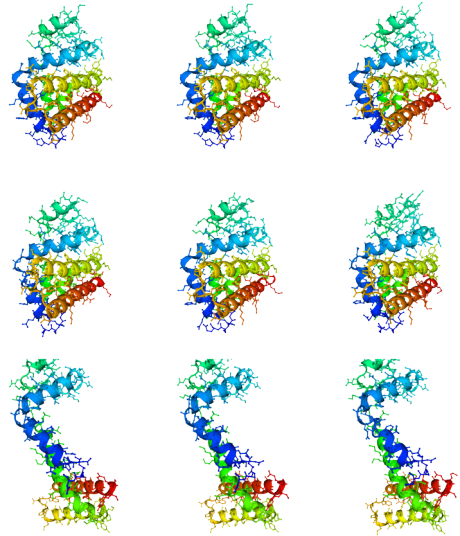


an ensemble of conformations

an ensemble of conformations



the same ensemble of conformations



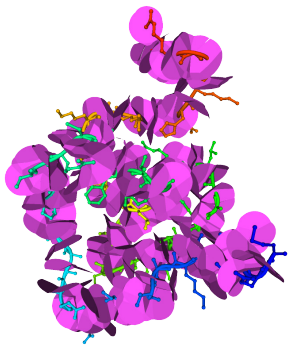
A dataset of ensembles of conformations from PDB

- Collected from the **Protein Data Bank** (PDB), https://www.wwpdb.org/.
- Ensembles formed by clustering chain sequences using **90%** identity.
- We used all **38'807** ensembles that were available.
- Ensembles have very different numbers of chains:
  - the largest ensemble contains **1413** chains
  - **9989** ensembles contain **only two** chains.

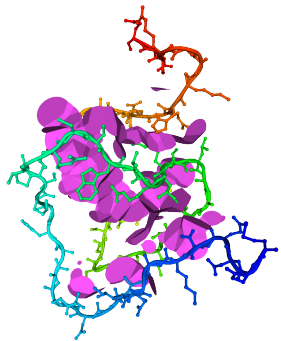
Contacts from a single conformation

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

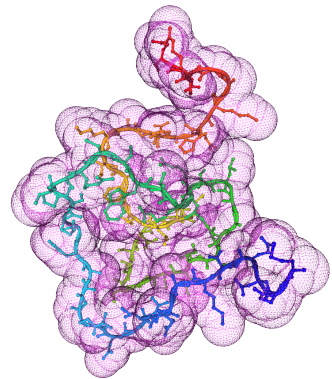
sequence separation <=5



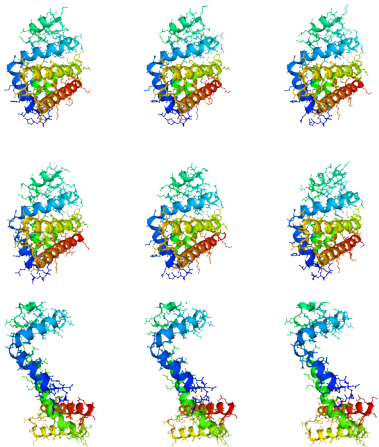
sequence separation >5



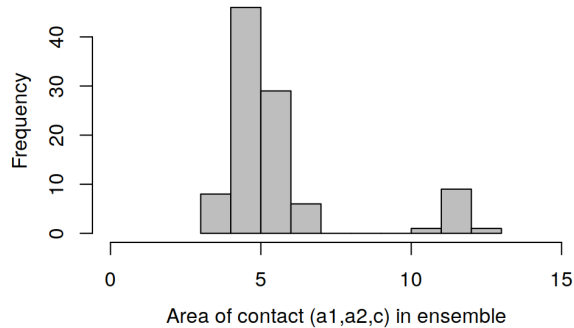
solvent-accessible surface



An ensemble



Distribution of areas of some contact in the ensemble



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

- $v^t = \min(\text{observed } t \text{ areas})$
- $u^t = \max(\text{observed } t \text{ areas})$

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:

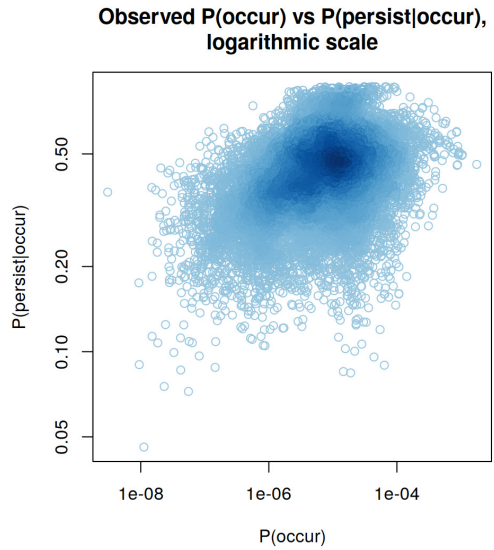
$$P^t_{\text{obs}}(\text{persist}|\text{occur}) = \frac{2V^t}{V^t + U^t} \tag{4}$$

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

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

Low correlation between occurence and persistence probabilities

$$\text{corr}(P_{\text{obs}}(\text{occur}), P_{\text{obs}}(\text{persist}|\text{occur})) = 0.11$$



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:

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

Using pseudo-energy to score inter-chain interfaces

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

$$E^t_{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\beta} = \alpha_1 \cdot \log P^t_{\text{obs}}(\text{occur}) + \alpha_2 \cdot \log P^t_{\text{exp}}(\text{occur}) + \\ + \alpha_3 \cdot \log P^t_{\text{obs}}(\text{persist}|\text{occur}) + \alpha_4 \cdot \log P^t_{\text{exp}}(\text{persist}|\text{occur}) + \beta \tag{9}$$

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

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

We optimize the weights $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta)$ for the task of selecting best-modelled interfaces.

A dataset of correct and incorrect interfaces

- A non-redundant set of 1549 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	CAD-score	binding_site_CAD-score
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

5-fold cross-validation results of selection performance

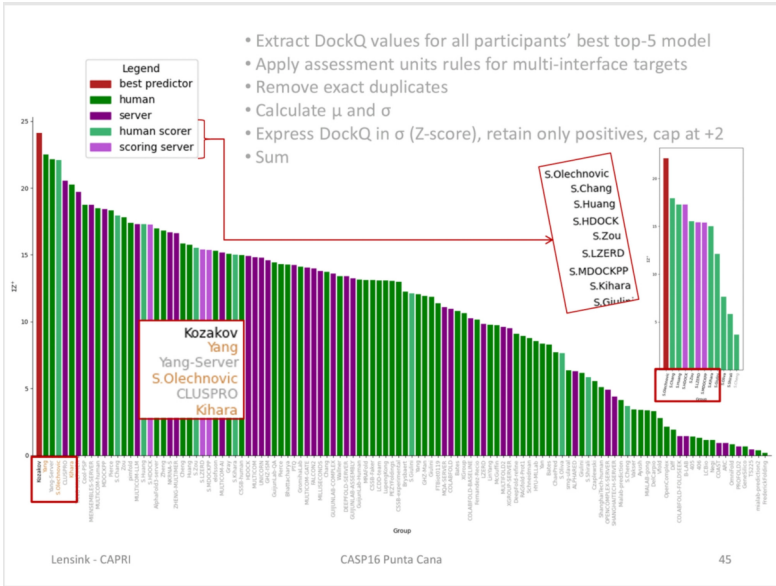
Correct selection rate = (number of good cases) / (number of all cases = 1549)

A "good case" is when both the native structural model and the most accurate prediction have lower total pseudo-energy than all the other predictions.

Method	**Components**	**Correct selection rate**	
		mean	**std. deviation**
Ideal selector		1	0
Random		0.06	0.016
Total area		0.06	0.017
Simple pseudo-energy	$P(\text{occur})$	0.60	0.042
	$P(\text{persist}\|\text{occur})$	0.64	0.028
	$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.72	0.034

Progress: from **0.60** to **0.72**

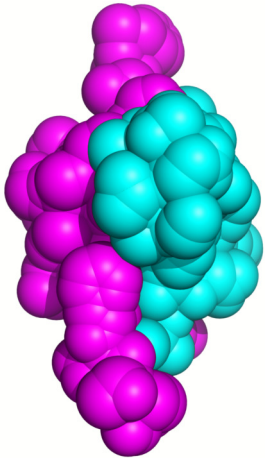
We used persistence descriptors in CASP16-CAPRI scoring experiment



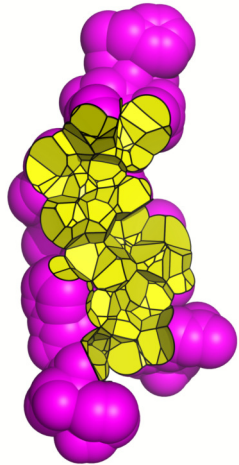
Extending geometry of of contacts

Inter-chain contacts

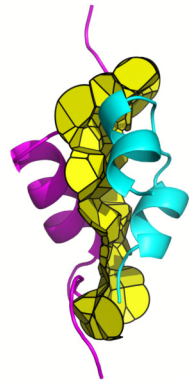
Solvent-accessible surface of an insulin heterodimer PDB:4UNG colored by subunit



The intersubunit interface shown together with the SAS of one subunit

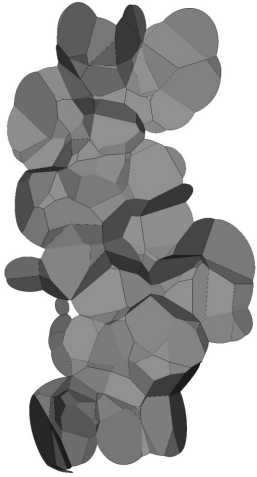


The intersubunit interface shown together with both subunits represented as cartoons

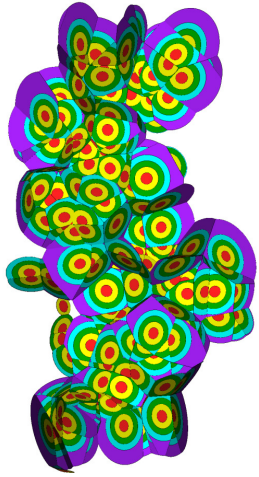


Introducing contact layers

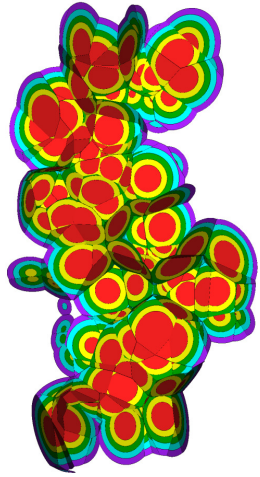
Simple contacts



Contact layers v1

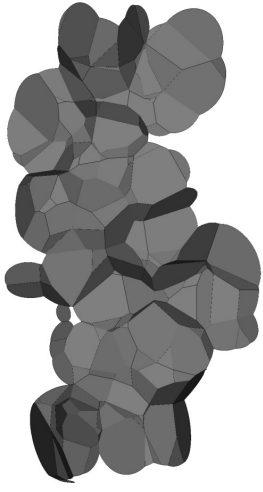


Contact layers v2

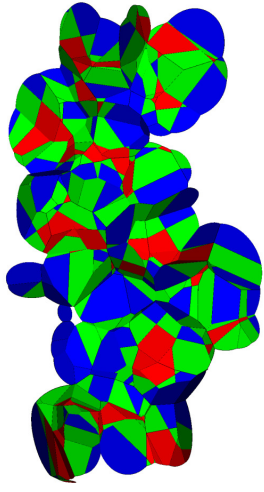


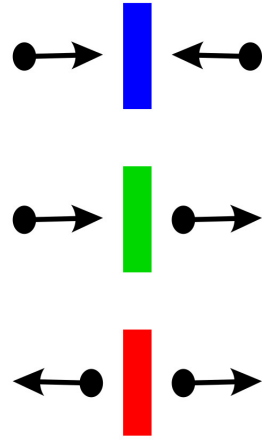
Introducing contact sectors

Simple contacts

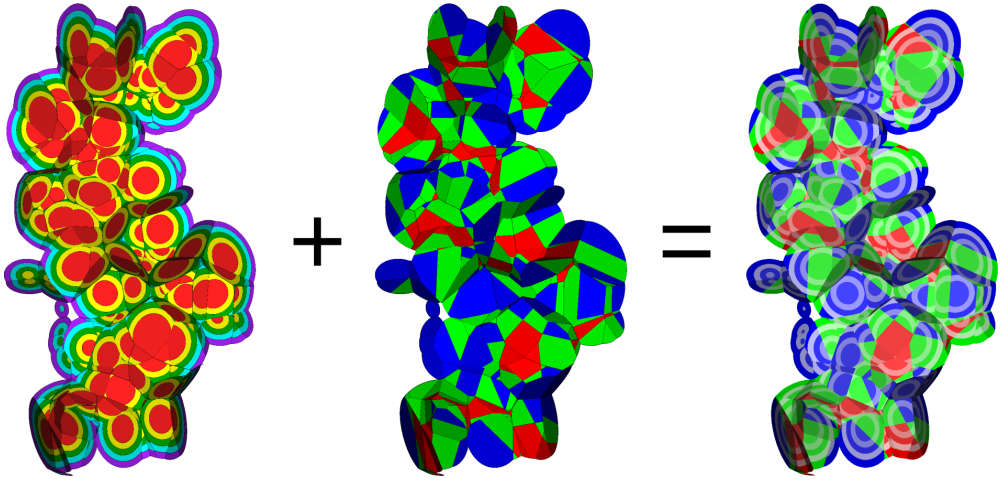


Contact sectors





Combining contact layers and sectors



5-fold cross-validation results of selection performance

Method	**Components**	**Correct selection rate**	
		mean	**std. deviation**
Ideal selector		1	0
Random		0.06	0.016
Total area		0.06	0.017
Simple pseudo-energy	$P(\text{occur})$	0.60	0.041
	$P(\text{persist}\|\text{occur})$	0.64	0.028
	$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.72	0.034
Extended geometry pseudo-energy	$P(\text{occur})$	0.92	0.008
	$P(\text{persist}\|\text{occur})$	0.91	0.023
	$P(\text{occur}) \cdot P(\text{persist}\|\text{occur})$	0.95	0.009

Progress: from **0.72** to **0.95**

Conclusions

- Tessellation-derived contact area descriptors can be used to collect information about contact stability from ensembles of conformations in PDB.
- The information about contact stability can improve scoring of protein-protein interfaces.
- Subdivinding tessellation-derived contacts into layers and sectors provides an improved description of protein-protein interfaces, making the contact stability information even more useful.
- The presented new contact descriptors can be used in more sophisticated, NN-based methods, e.g. VoroIF-GNN.

Thanks

Thank you!

CNRS Laboratoire Jean Kuntzmann:

- Sergei Grudinin
- The GruLab Team (https://grulab.imag.fr)

Useful links:

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

Funded by
the European Union