

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

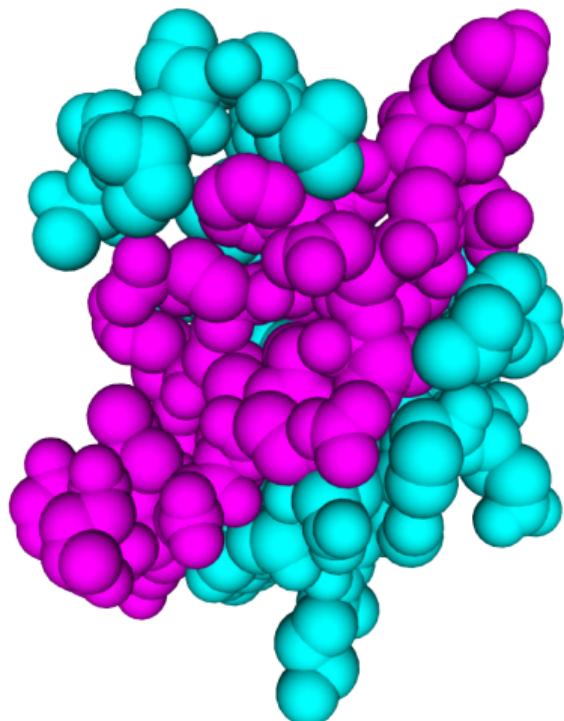
Dr. Kliment Olechnovič

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:

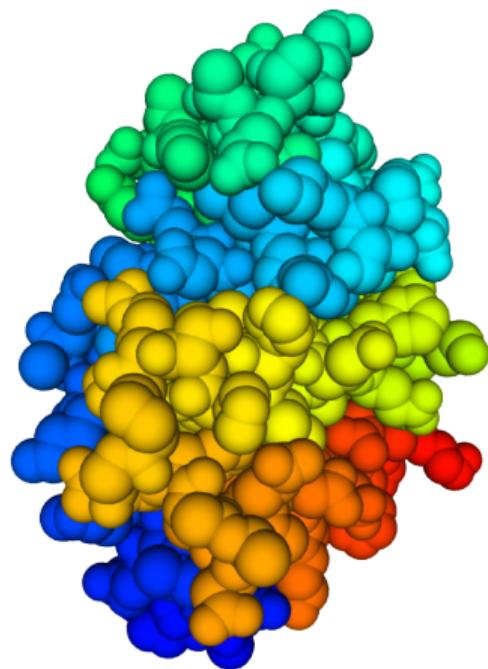
- ▶ computational geometry
- ▶ machine learning
- ▶ developing free software

Today's questions:

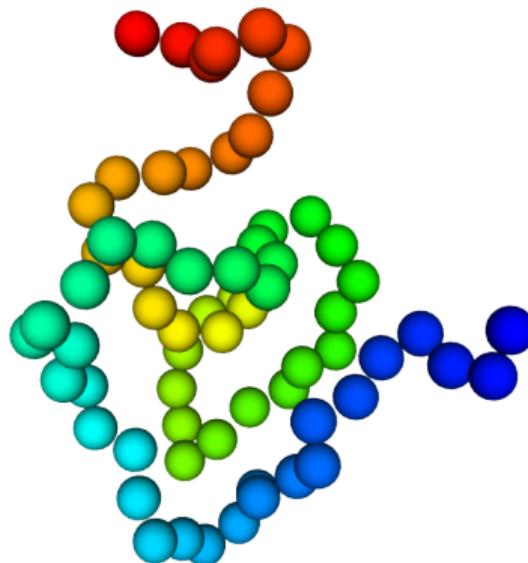
- ▶ What can we learn from disordered protein?
- ▶ And would it be useful for analyzing protein-protein interactions?

Data of molecular conformations

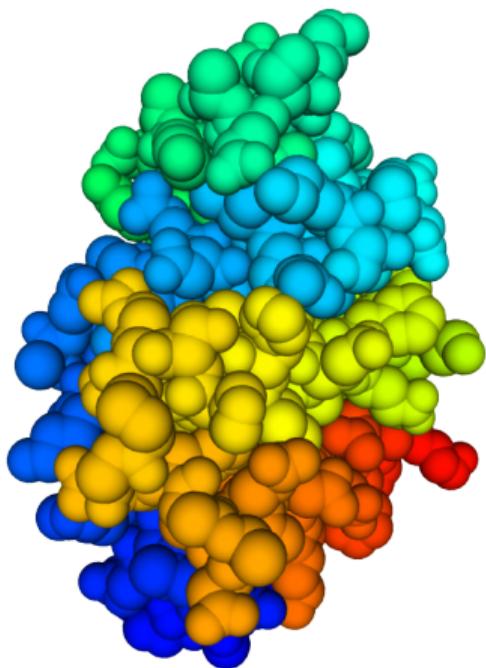
Protein Data Bank (PDB) data



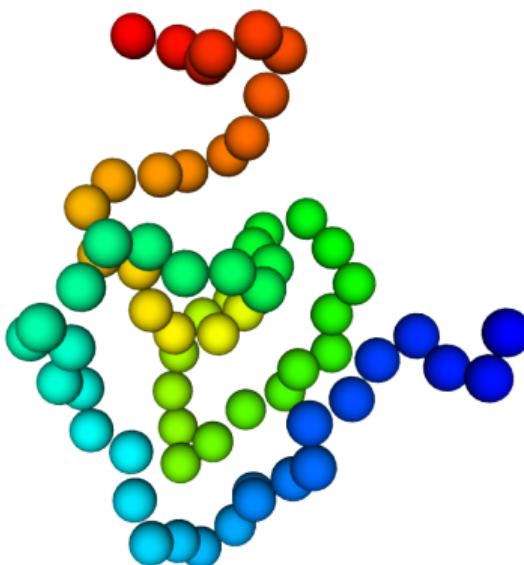
Simulated intrinsically disordered protein (IDP) data



PDB

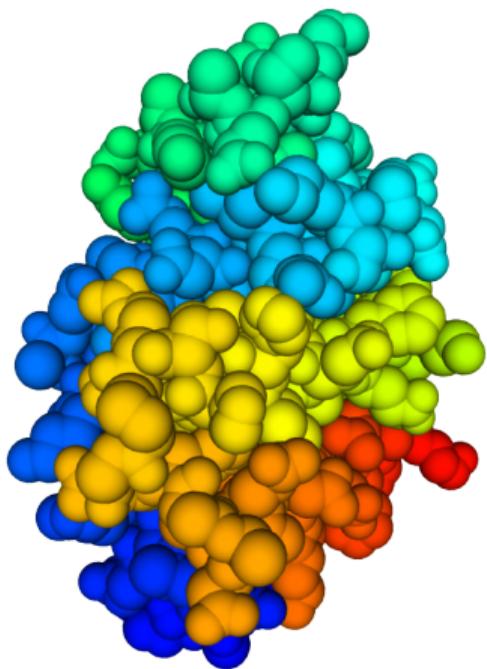


IDP

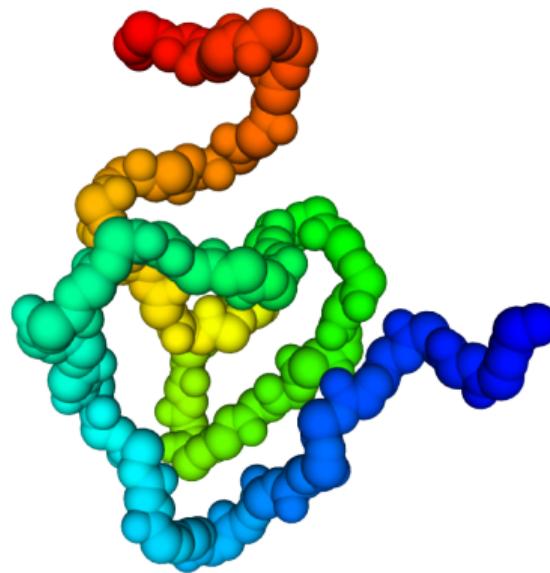


C-alpha atoms

PDB

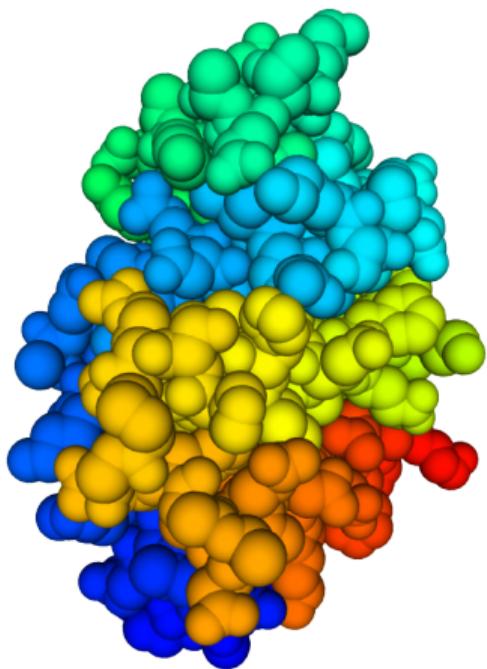


IDP

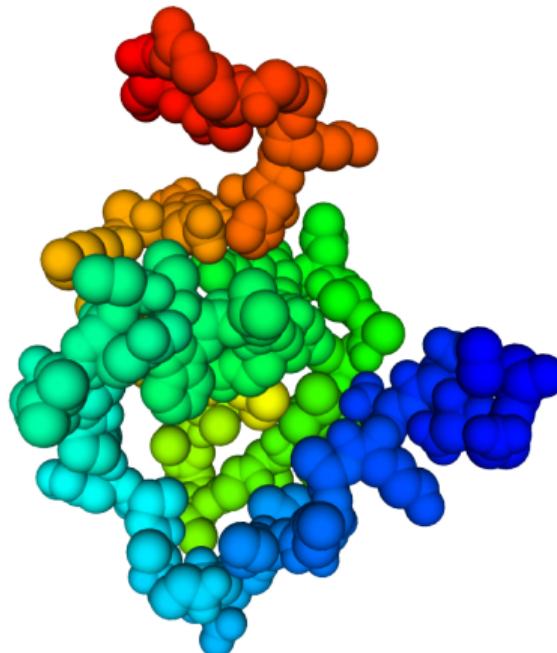


backbone atoms built by Pulchra

PDB

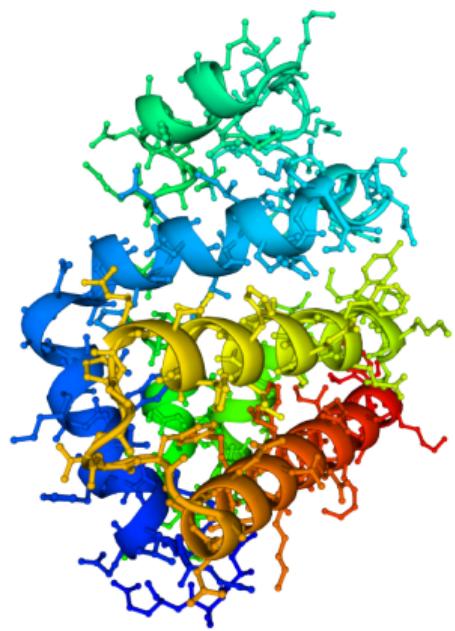


IDP

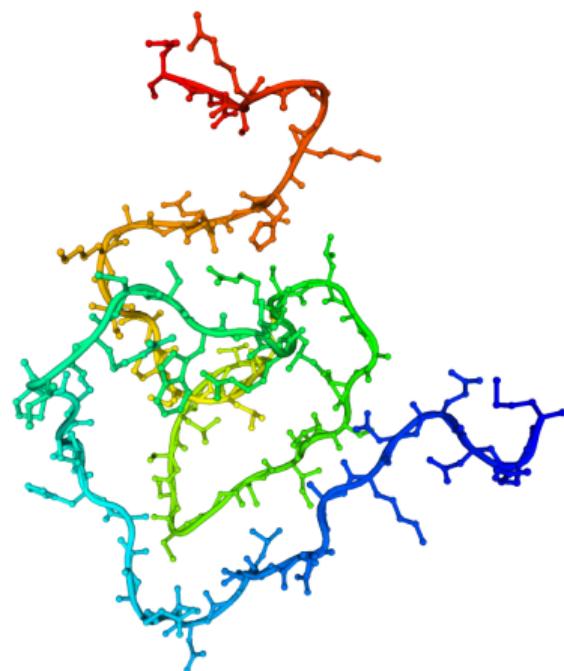


side-chain atoms built by FASPR

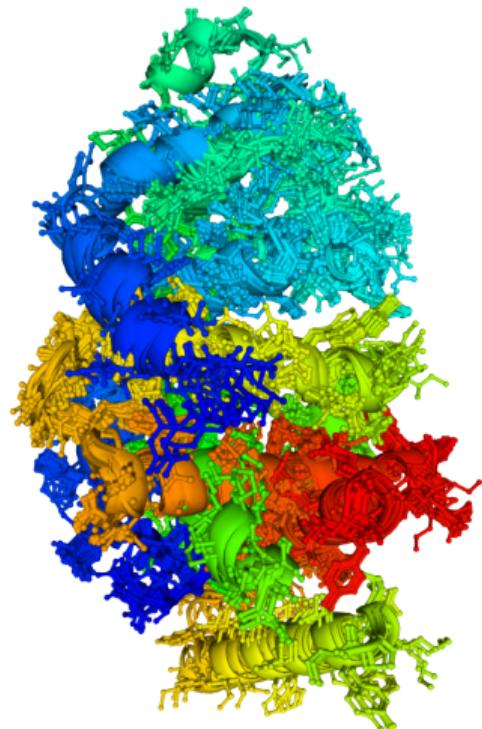
PDB



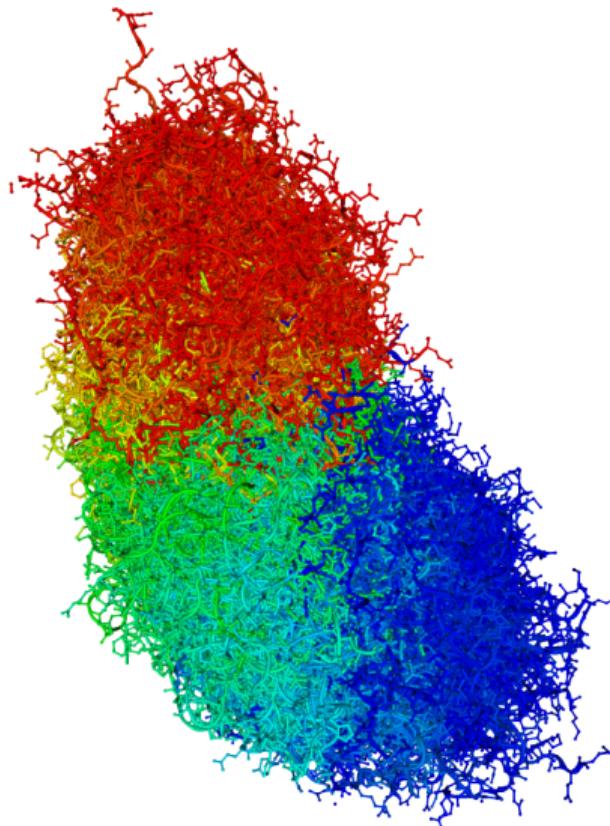
IDP



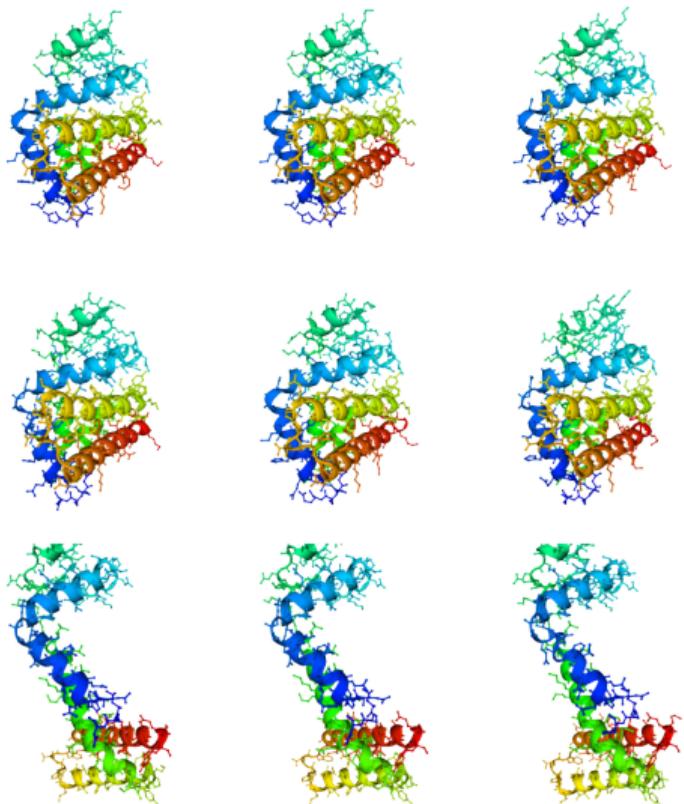
PDB ensemble



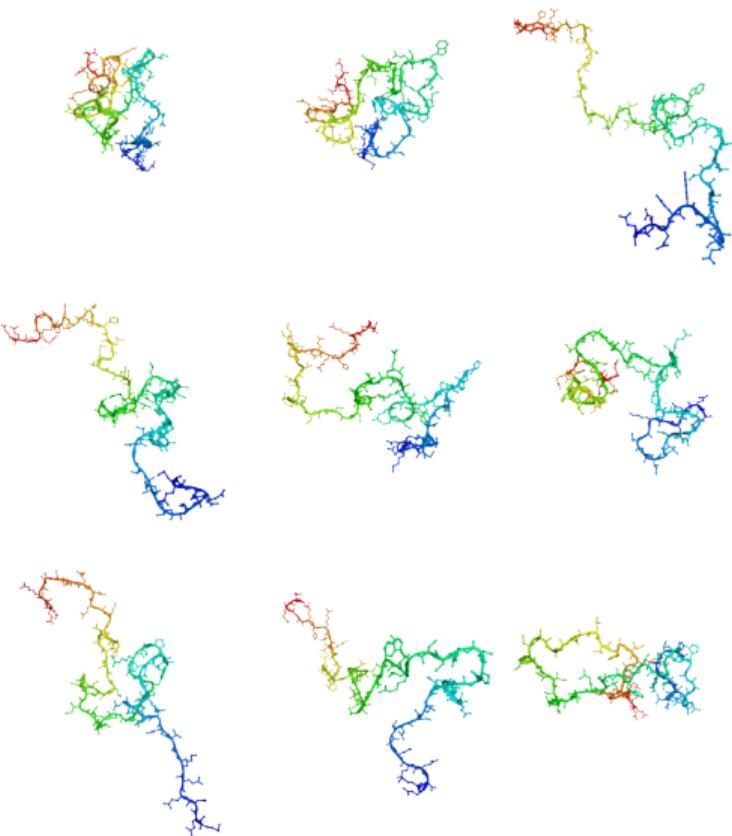
IDP ensemble



PDB ensemble



IDP ensemble



## PDB dataset

- ▶ From the Protein Data Bank, [www.pdb.org](http://www.pdb.org)
- ▶ 429945 protein chains clustered using 90% sequence identity threshold into 38807 ensembles
- ▶ Largest ensemble contains 1413 chains
- ▶ 9989 ensembles contain only two chains

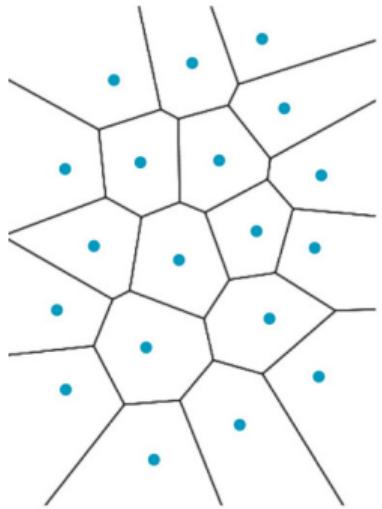
## IDP dataset “IDRome”

- ▶ From Tesei, G., Trolle, A.I., Jonsson, N. et al. *Conformational ensembles of the human intrinsically disordered proteome*. Nature (2024)
- ▶ 28058 IDP-like protein sequences
- ▶ For every sequence, there is an ensemble of 1010 conformations that were simulated with CALVADOS
- ▶ We took all the 16774 ensembles of chains that are from 60 to 600 residues in length, so we considered  $16774 * 1010 = 16941740$  conformations in total

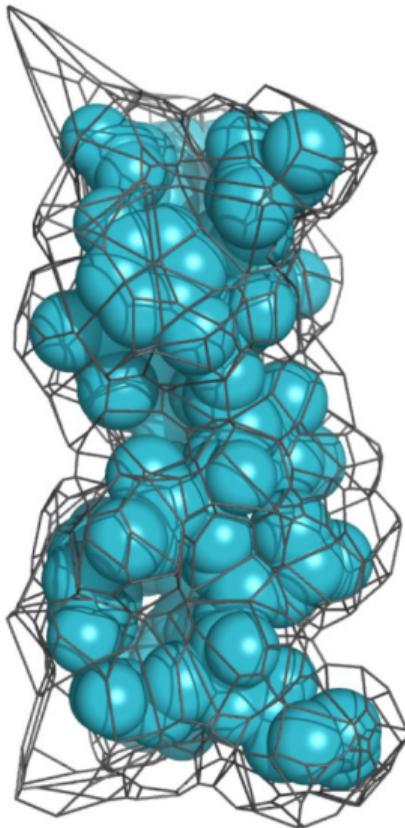
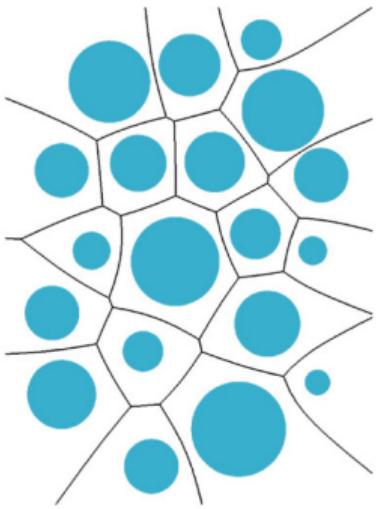
Describing interactions in molecular conformations using the  
Voronoi tessellation

# Voronoi diagram of points and balls

"Classic" Voronoi diagram  
of points

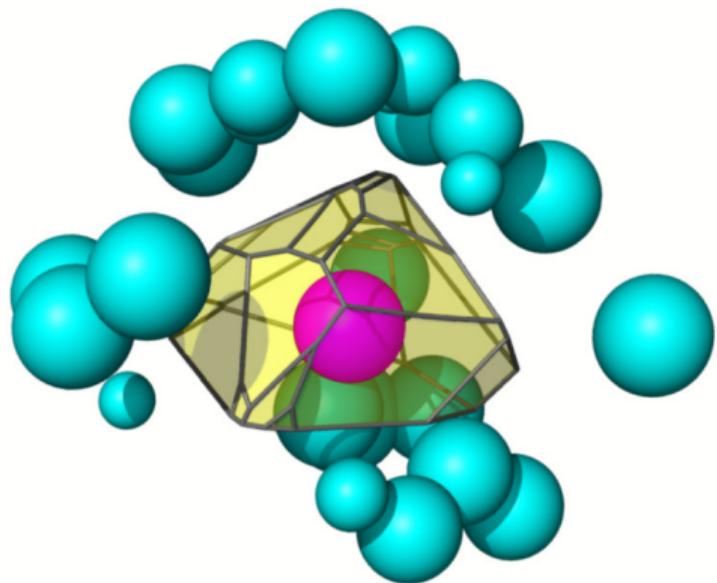


Voronoi diagram  
of balls

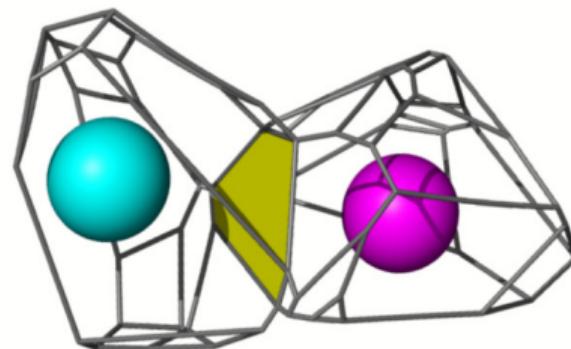


## Voronoi tessellation-based analysis of structures

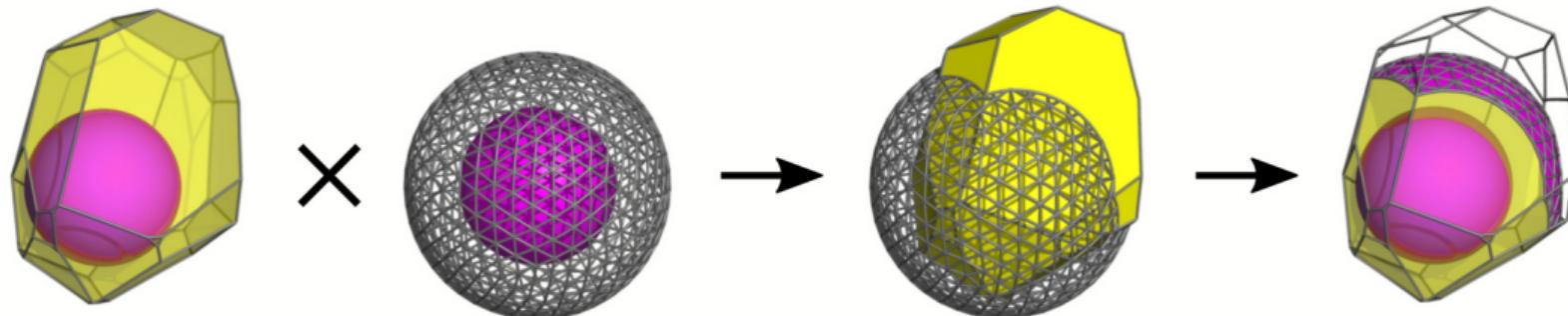
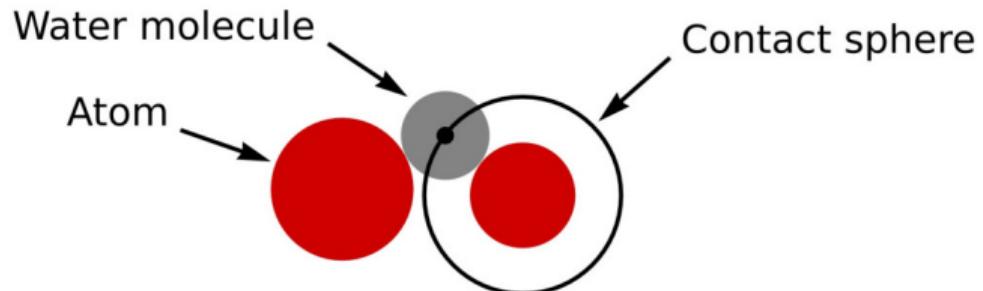
Voronoi cell of an atom surrounded by its neighbors



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

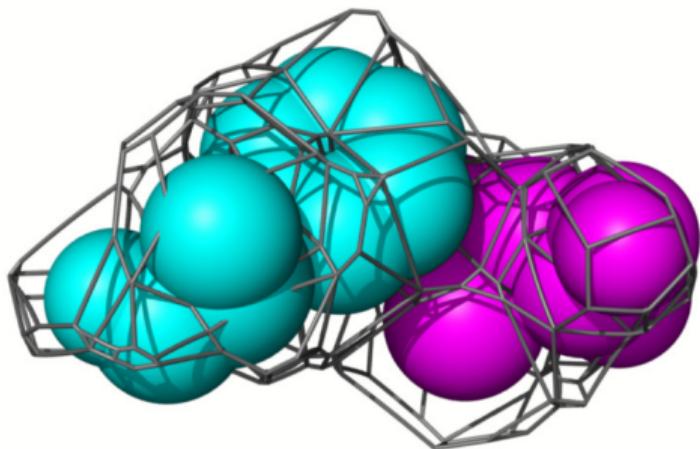


## Constrained contacts

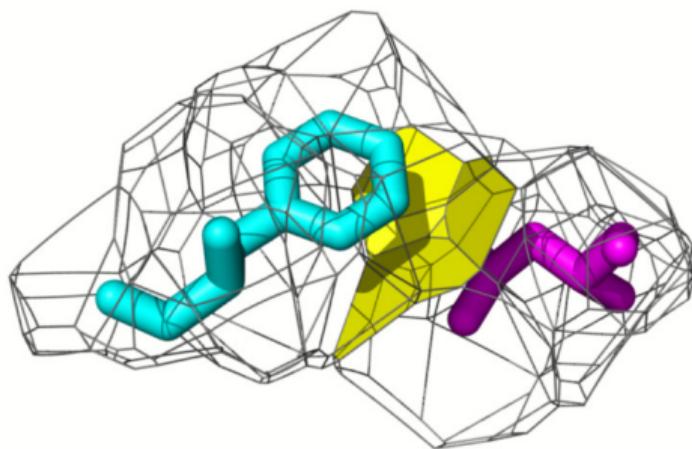


## Deriving residue-residue contacts

Voronoi cells of two neighboring residues

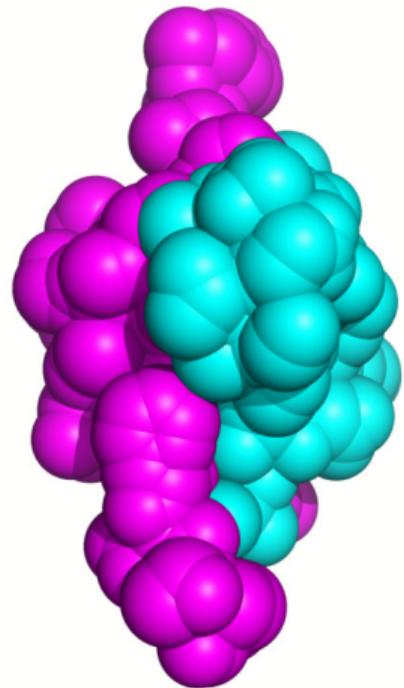


Residue-residue contact surface  
defined as a union of  
atom-atom contact surfaces

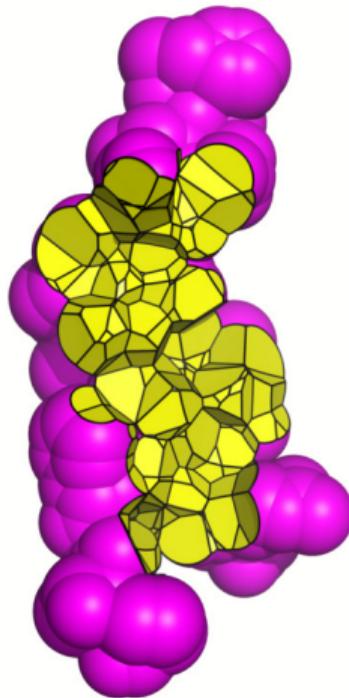


## Inter-chain contacts

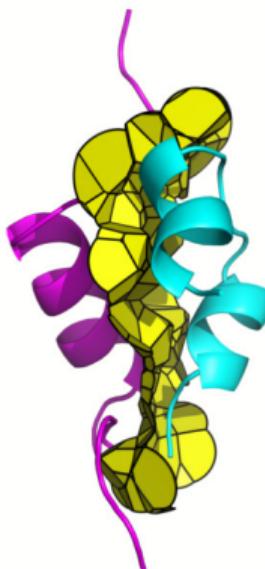
Solvent-accessible surface  
of an insulin heterodimer  
PDB:4UNG colored by subunit



The intersubunit interface  
shown together with the  
SAS of one subunit

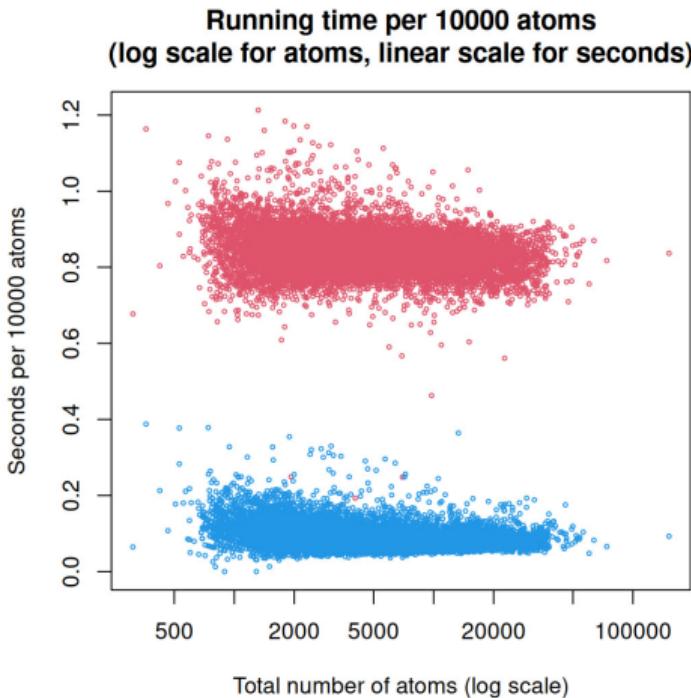
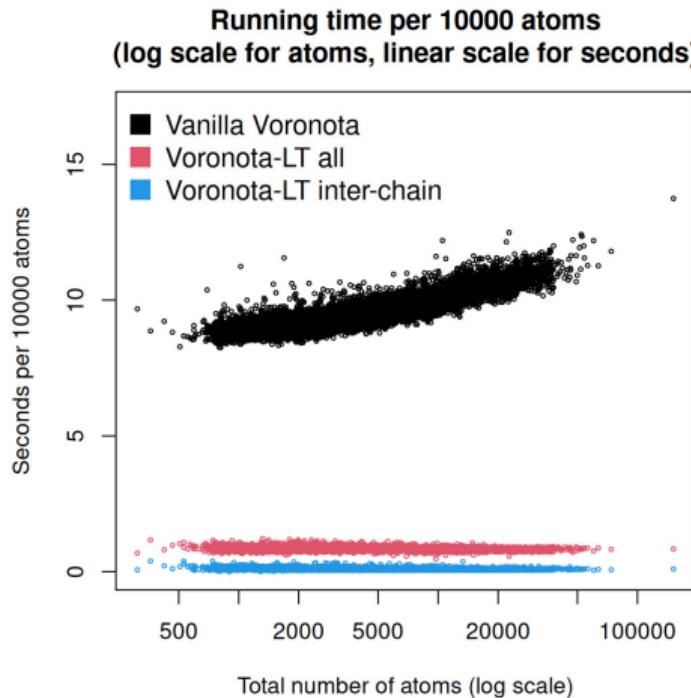


The intersubunit interface  
shown together with  
both subunits represented  
as cartoons



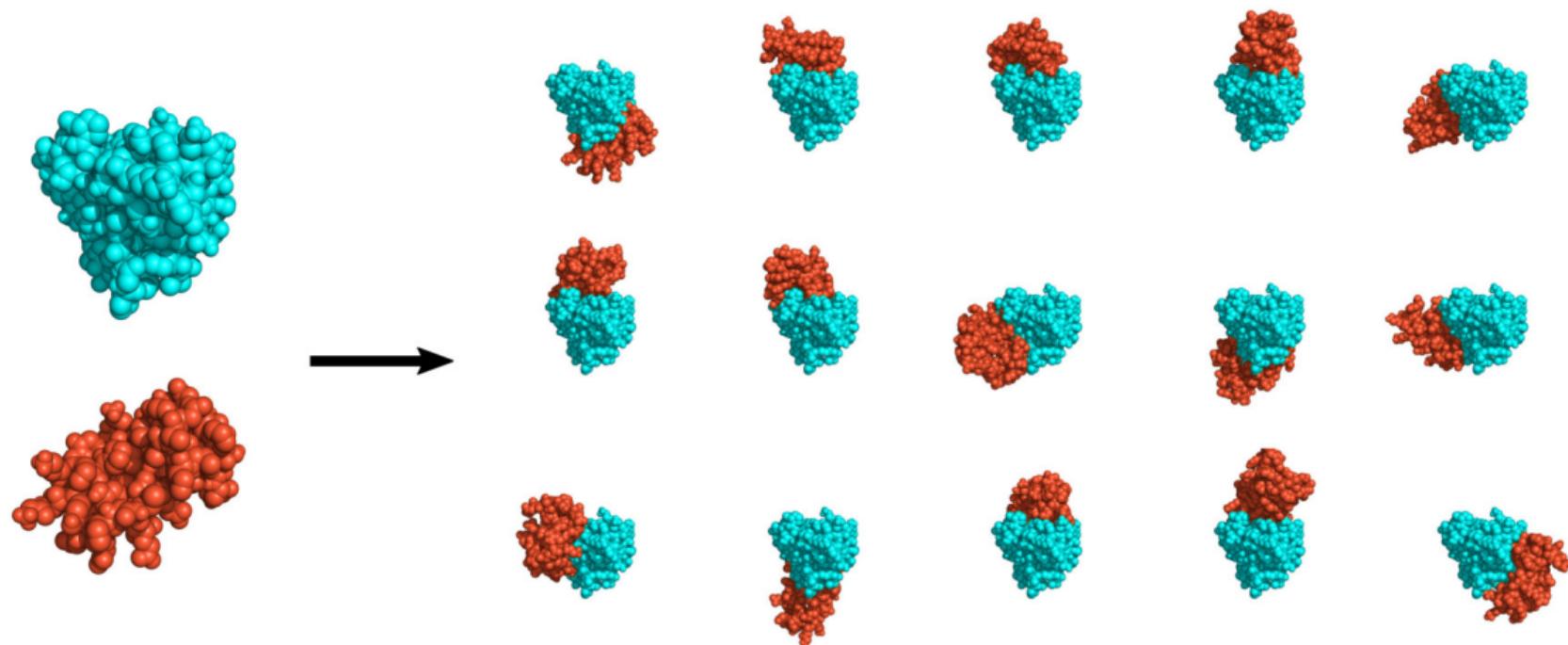
# Voronota-LT

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

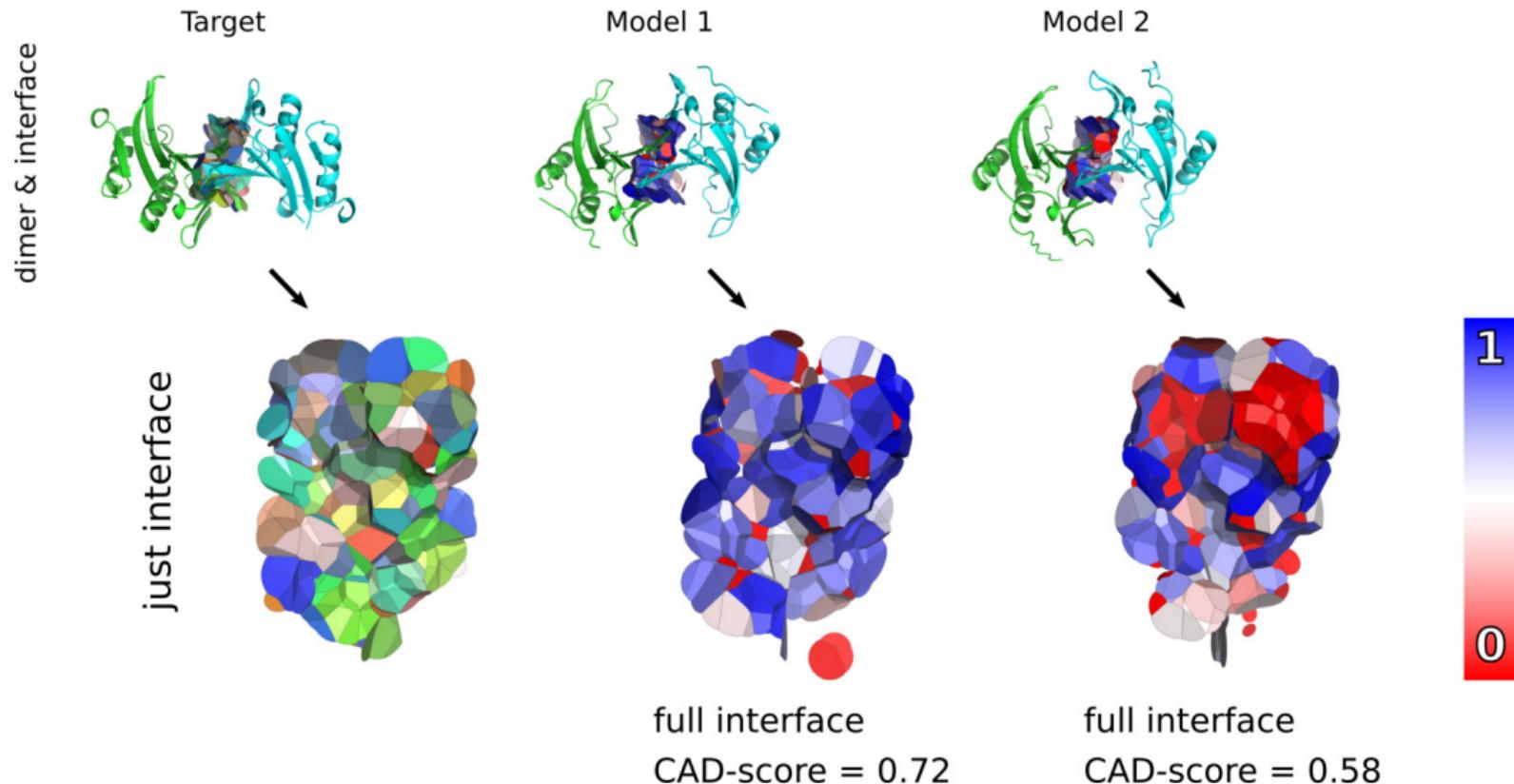


An application of tessellation-based description of interactions

Same chains can have differently modelled interfaces



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

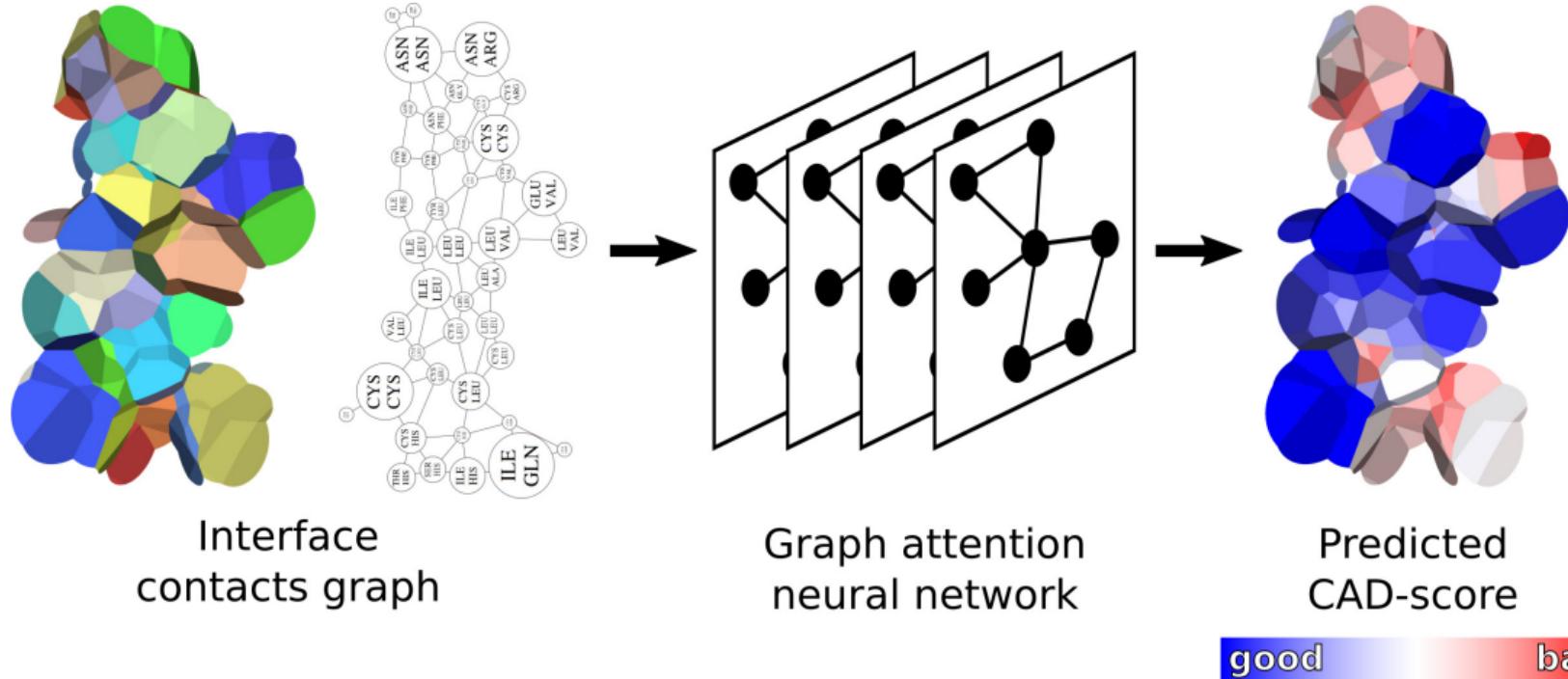


## A dataset of correct and incorrect interfaces

- ▶ 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. VorolF-GNN)



Interface  
contacts graph

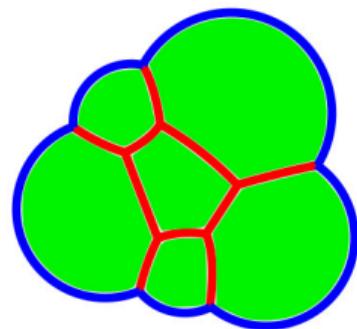
Graph attention  
neural network

Predicted  
CAD-score



# Input interface graph annotation in VoronF-GNN

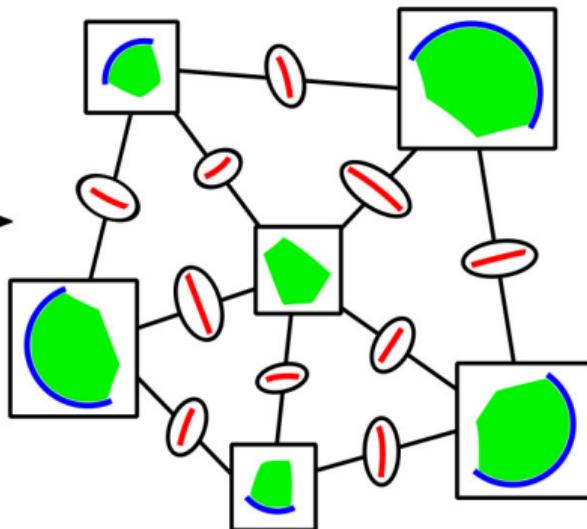
## Tessellation-derived interface contacts



Contact surface  
Contact-solvent border  
Inter-contact border



## Interface graph



Graph **node** attributes  
(15 values)

Contact surface area

Contact-solvent  
border length

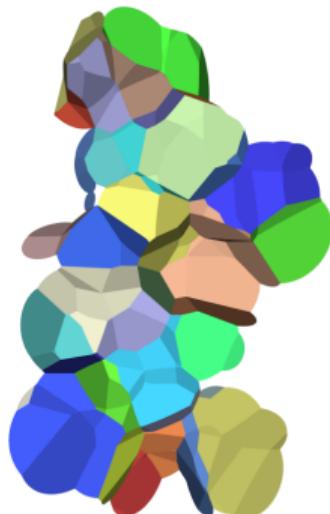
Sum of inter-contact  
border lengths

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)

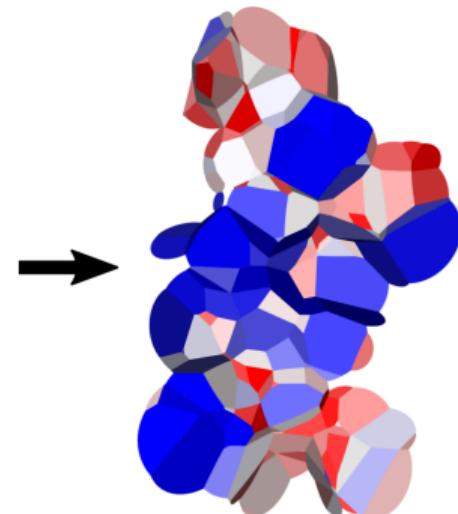


Interface  
contact areas



$$\begin{aligned} E(a_i, a_j, c_k) &= \log \frac{P_{\text{exp}}(a_i, a_j, c_k)}{P_{\text{obs}}(a_i, a_j, c_k)} = \\ &= \log \frac{F_{\text{exp}}(\text{area}(a_i), \text{area}(a_j), \text{area}(c_k))}{F_{\text{obs}}(\text{area}(a_i, a_j, c_k))} \\ E_n(\Omega_\phi) &= \frac{\sum_{\omega \in \Omega_\phi} E(\text{type}_\omega) \cdot \text{area}_\omega}{\sum_{\omega \in \Omega_\phi} \text{area}_\omega} \end{aligned}$$

Statistical potential  
for contact areas



Interface  
pseudo-energy

good

bad

Deriving and using statistics of contact areas from ensembles  
of conformations

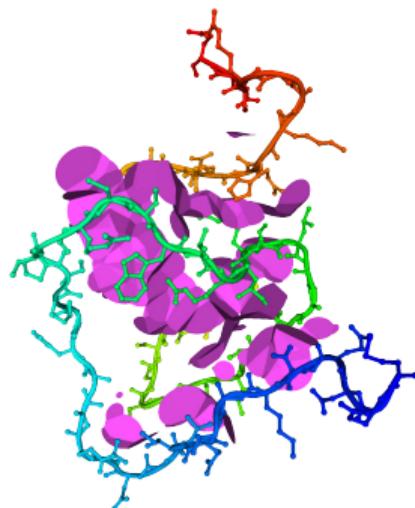
## Contacts from a single conformation

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

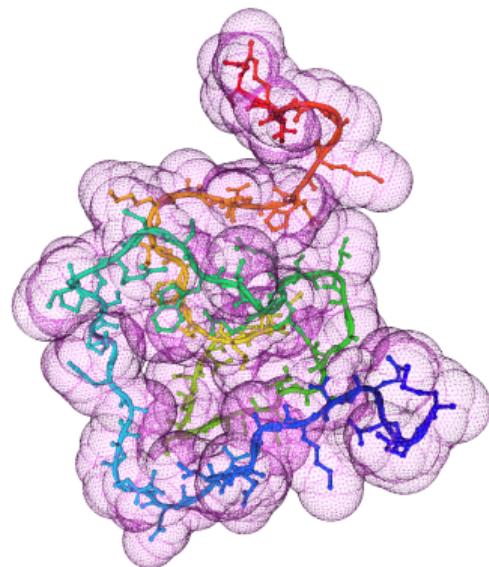
sequence separation  $\leq 5$



sequence separation  $> 5$

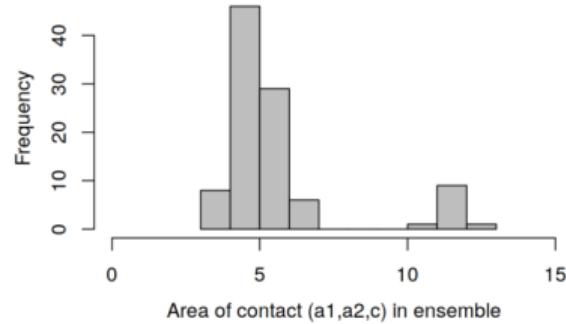
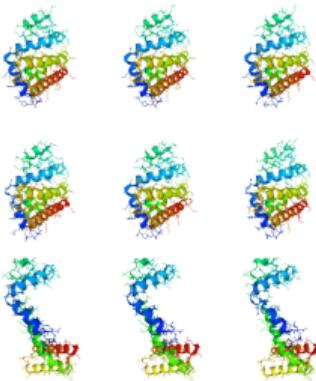


solvent-accessible surface

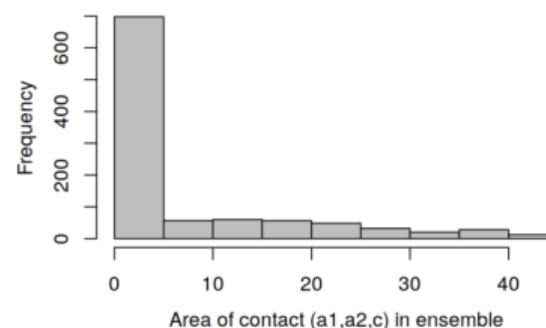
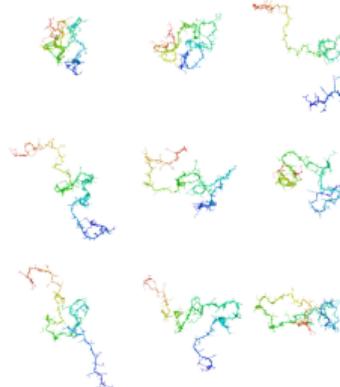


## Contact areas from a single ensemble of conformations

PDB ensemble

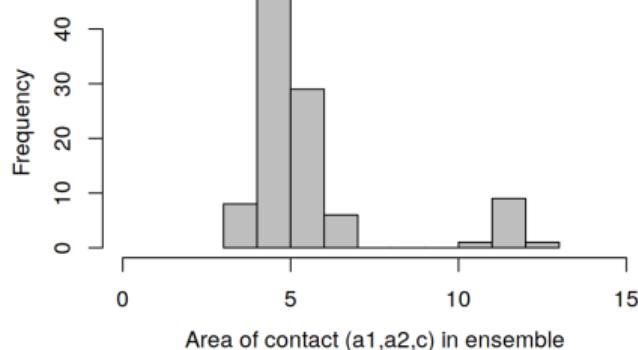


IDRome ensemble

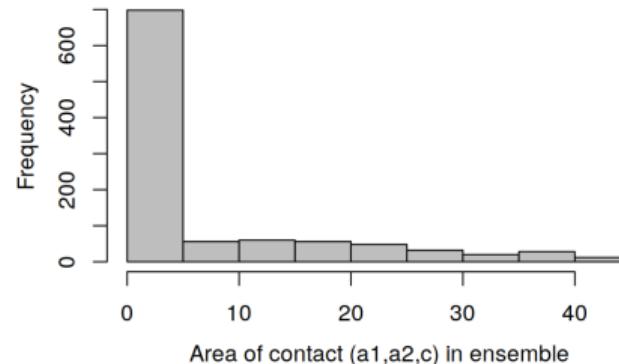


## Contact areas from a single ensemble of conformations

PDB ensemble



IDRome ensemble



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

- ▶  $v = \min(\text{observed contact areas})$
- ▶  $u = \max(\text{observed contact areas})$

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

- ▶  $v = \text{mean}(\text{observed contact areas})$
- ▶  $u = \max(\text{observed contact areas})$

## Areas of contact types from a multiple ensembles of conformations

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

For every contact type  $t = (a_1, a_2, c)$  from the set of all possible types contact  $T$ , we sum the relevant  $v^t$  and  $u^t$  values from all the ensembles to get  $V^t$  and  $U^t$  sums.

## Observed probabilities of areas of contact types

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

$$P_{\text{obs}}^t(\text{occur}) = \frac{V^t + U^t}{\sum_{s \in T} (V^s + U^s)} \quad (1)$$

Observed conditional probability estimate of contact area unit to persist:

$$P_{\text{obs}}^t(\text{persist} | \text{occur}) = \frac{2V^t}{V^s + U^s} \quad (2)$$

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

$$P_{\text{obs}}^t(\text{occur and persist}) = P_{\text{obs}}^t(\text{occur}) \cdot P_{\text{obs}}^t(\text{persist} | \text{occur}) \quad (3)$$

## 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):

$$P_{\text{exp}}^{t=(a_1, a_2, c)}(\text{occur}) \sim P_{\text{obs}}^{(a_1, *, *)}(\text{occur}) \cdot P_{\text{obs}}^{(*, a_2, *)}(\text{occur}) \cdot P_{\text{obs}}^{(*, *, c)}(\text{occur}). \quad (4)$$

Expected conditional probability estimate of contact area unit to persist:

$$P_{\text{exp}}^t(\text{persist}|\text{occur}) = \frac{2 \cdot \sum_{s \in T} V^s}{\sum_{s \in T} (V^s + U^s)} \quad (5)$$

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

$$P_{\text{exp}}^t(\text{occur and persist}) = P_{\text{exp}}^t(\text{occur}) \cdot P_{\text{exp}}^t(\text{persist}|\text{occur}) \quad (6)$$

## Deriving pseudo-energy coefficient from probability estimates

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

$$E^t \sim \log \left( \frac{P_{\text{exp}}^t(\text{occur and persist})}{P_{\text{obs}}^t(\text{occur and persist})} \right) \quad (7)$$

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

$$E^t = \alpha_1 \cdot E_{\text{obs}}^t(\text{occur}) + \alpha_2 \cdot E_{\text{exp}}^t(\text{occur}) + \alpha_3 \cdot E_{\text{obs}}^t(\text{persist|occur}) + \alpha_4 \cdot E_{\text{exp}}^t(\text{persist|occur}) + \beta \quad (8)$$

where:

$$E_{\text{obs}}^t(\text{occur}) = \log P_{\text{obs}}^t(\text{occur})$$

$$E_{\text{exp}}^t(\text{occur}) = \log P_{\text{exp}}^t(\text{occur})$$

$$E_{\text{obs}}^t(\text{persist|occur}) = \log P_{\text{obs}}^t(\text{persist|occur})$$

$$E_{\text{exp}}^t(\text{persist|occur}) = \log P_{\text{exp}}^t(\text{persist|occur})$$

## Using pseudo-energy to score inter-chain interfaces

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

$$S_{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta}(G) = \sum_{g \in G} \text{area}(g) \cdot E_{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta}^{\text{type}}(g) \quad (9)$$

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(occur)   | 0.78           |
|                |        |                 | P(persist occur)                                   | 0.78           |
|                |        |                 | $P(\text{occur}) * P(\text{persist} \text{occur})$ | <b>0.89</b>    |
| Pseudo energy  | PDB    | residue-residue | P(occur)   | 0.59           |
|                |        |                 | P(persist occur)                                   | 0.59           |
|                |        |                 | $P(\text{occur}) * P(\text{persist} \text{occur})$ | 0.62           |
| Pseudo energy  | IDRome | residue-residue | P(occur)   | 0.50           |
|                |        |                 | P(persist occur)                                   | 0.53           |
|                |        |                 | $P(\text{occur}) * P(\text{persist} \text{occur})$ | 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(occur)                  | 0.78           |
|                |        |                 | P(persist occur)          | 0.78           |
|                |        |                 | P(occur)*P(persist occur) | <b>0.89</b>    |
| Pseudo energy  | PDB    | residue-residue | P(occur)                  | 0.59           |
|                |        |                 | P(persist occur)          | 0.59           |
|                |        |                 | P(occur)*P(persist occur) | 0.62           |
| Pseudo energy  | IDRome | residue-residue | P(occur)                  | 0.50           |
|                |        |                 | P(persist occur)          | 0.53           |
|                |        |                 | P(occur)*P(persist occur) | 0.55           |
| VorolF-GNN     | all    | hybrid          | all                       | <b>0.98</b>    |

## 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(occur)                  | 0.60           |
|                |        |                 | P(persist occur)          | 0.59           |
|                |        |                 | P(occur)*P(persist occur) | <b>0.64</b>    |
| Pseudo energy  | PDB    | residue-residue | P(occur)                  | 0.54           |
|                |        |                 | P(persist occur)          | 0.52           |
|                |        |                 | P(occur)*P(persist occur) | 0.55           |
| Pseudo energy  | IDRome | residue-residue | P(occur)                  | 0.49           |
|                |        |                 | P(persist occur)          | 0.51           |
|                |        |                 | P(occur)*P(persist occur) | 0.50           |
| VorolF-GNN     | all    | hybrid          | all                       | <b>0.77</b>    |

## 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 tessellation-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

