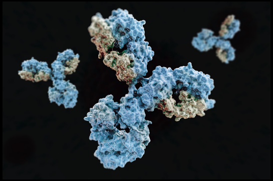
Topological Analysis and ML for biological docking

Gleb Rusyaev Letovo

Outline

1. Introduction
2. Target
3. Checking techniques
4. Dataset review
5. Naive Correlation
6. Multi-Linear Regression
7. Topological analysis
8. Gradient Boosted Decision Trees

What is it?



Antibody

- Antibody vs Bad guys (Pathogens, Viruses, Cancer Cells, etc ...)
- Biological (and physical) systems gravitate towards lowest potential energy

Why bother?

1. Check if synthetic anti-cell can glue to the virus
2. Synthesize
3. Cure virus
4. ???
5. PROFIT

How to calculate potential energy?

(i) Coulomb's law: $|F| = k_e \frac{|q_1 q_2|}{r^2}$

(ii) Van der Waals force

(iii) Energy of Statistic Potential

Dataset review: docking_data.csv

(i) E_fst: Anti-Body Total Energy

(ii) E_fst_elstat: Anti-Body Coulomb's Energy

(iii) E_fst_VdW: Anti-Body Statistical Potential Energy

(iv) group_size: size of group (in which it included)

(v) type_of_complex_in_group : center | in_between, border): place in group

(vi) E_snd_SP: SP of individual in group

Dataset review: docking_data.csv

(i) `avg_E_fst`: average energy in group

(ii) `avg_E_snd_SP`: average stat. potential energy in group

(iii) `E_third`: Minimal Energy Complex after optimisation

(iv) `alt_E_third`: Minimal Energy Complex after optimisation (alternative method)

(v) `E_third_SP`: Stat. Potential after optimisation

(vi) `avg_E_third`: Avg. `E_third` in group

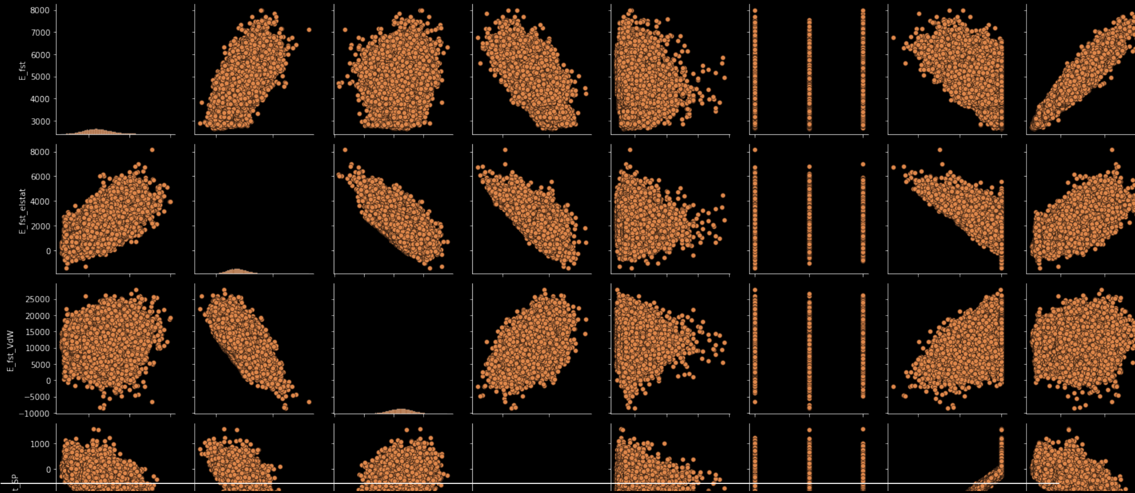
Dataset review: docking_data.csv

(i) `avg_alt_E_third`: Avg. `alt_E_third` in group

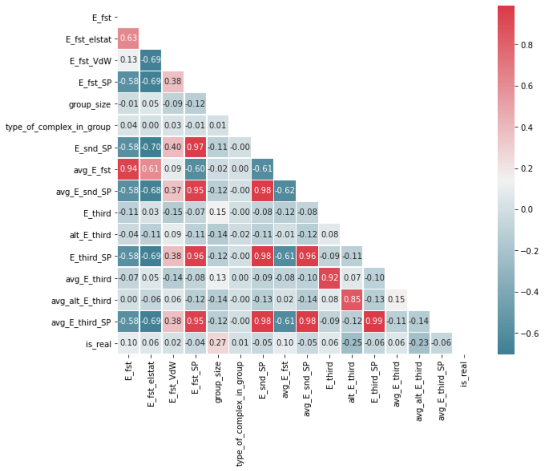
(ii) `avg_E_third_SP`: Avg. `E_third_SP` in group

(iii) `is_real`: Does it actually works this?

Naive Correlation

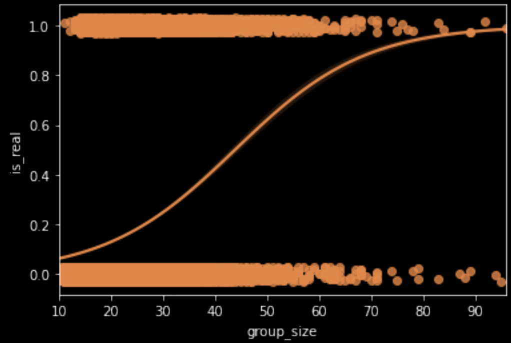


Naive Correlation



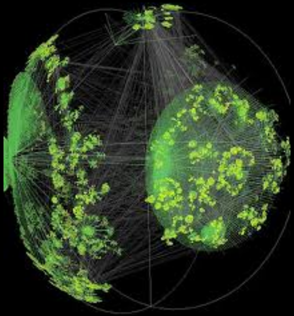
	E_fst	E_fst_elstat	E_fst_VdW	E_fst_SP	group_size	type_of_complex_in_group	E_snd_SP	avg_E_fst	avg_E_snd_SP	E_third	alt_E_third	E_third_SP	avg_E_third	avg_alt_E_third	avg_E_third_SP	is_real
E_fst																
E_fst_elstat	0.63															
E_fst_VdW	0.13	-0.69														
E_fst_SP	-0.58	-0.69	0.38													
group_size	-0.01	0.05	-0.09	-0.12												
type_of_complex_in_group	0.04	0.00	0.03	-0.01	0.01											
E_snd_SP	-0.58	-0.70	0.40	0.97	-0.11	-0.00										
avg_E_fst	0.94	0.61	0.09	-0.60	-0.02	0.00	-0.61									
avg_E_snd_SP	-0.58	-0.68	0.37	0.95	-0.12	-0.00	0.98	-0.62								
E_third	-0.11	0.03	-0.15	-0.07	0.15	-0.00	-0.08	-0.12	-0.08							
alt_E_third	-0.04	-0.11	0.09	-0.11	-0.14	-0.02	-0.11	-0.01	-0.12	0.08						
E_third_SP	-0.58	-0.69	0.38	0.96	-0.12	-0.00	0.98	-0.61	0.96	-0.09	-0.11					
avg_E_third	-0.07	0.05	-0.14	-0.08	0.13	0.00	-0.09	-0.08	-0.10	0.92	0.07	-0.10				
avg_alt_E_third	0.00	-0.06	0.06	-0.12	-0.14	-0.00	-0.13	0.02	-0.14	0.08	0.85	-0.13	0.15			
avg_E_third_SP	-0.58	-0.69	0.38	0.95	-0.12	-0.00	0.98	-0.61	0.98	-0.09	-0.12	0.99	-0.11	-0.14		
is_real	0.10	0.06	0.02	-0.04	0.27	0.01	-0.05	0.10	-0.05	0.06	-0.25	-0.06	0.06	-0.23	-0.06	

Multi-Linear Regression



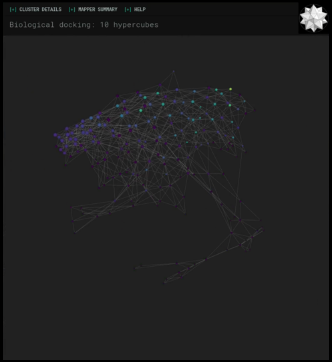
1. Complex border has 4x higher chance in being real, than centroid
2. Group size have 0.0127 correlation with `is_real` (as we early noticed, it's correlated)

Topology

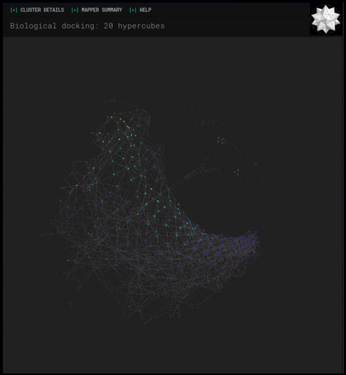


$$\text{1337-dimensions} \xrightarrow{\text{hyper-cube clusterization}} \text{nice 2d graph}$$

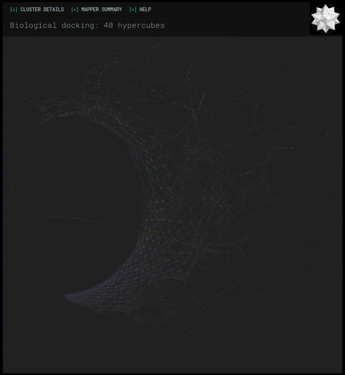
10, 20, 40 Hyper-cubes



Biological docking: 10 hypercubes

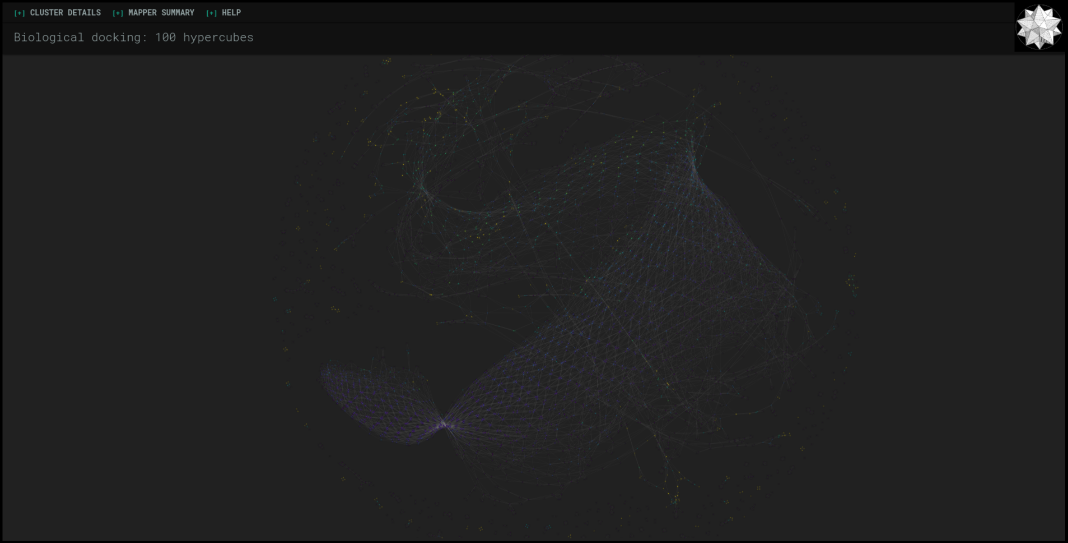


Biological docking: 20 hypercubes



Biological docking: 40 hypercubes

100 Hyper-cubes



Biological docking: 100 hypercubes

Gradient Boosted Decision Trees

- Performance on training set: 0.63 – 0.66
- Performance on validating set: 0.6503
- Can be done even better using AWS

References

Jupyter Notebook: `https://rusyaew.github.io/DockingML.ipynb`
Dataset: `https://rusyaew.github.io/docking_data.csv`
Bye!