Persistent homology applied to protein stability

Carlos Ronchi
Marcio Gameiro

University of São Paulo

Part I - the scary part

- Topology
- Simplices
- Simplicial complexes
- Filtration
- Persistent homology
- Persistence diagrams

What is topology? Why should I care?

- Study of shapes under **continuous deformations**.
- No tearing or gluing

What is topology? Why should I care?

- Study of shapes under **continuous deformations**
- No tearing or gluing

- Abstract nonsense for mathematicians sometimes

What can you extract from topology?

What can you extract from topology?

- **Topological invariants**

What can you extract from topology?

- **Topological invariants**
  - Connected components

What can you extract from topology?

- **Topological invariants**
  - Connected components
  - Holes

What can you extract from topology?

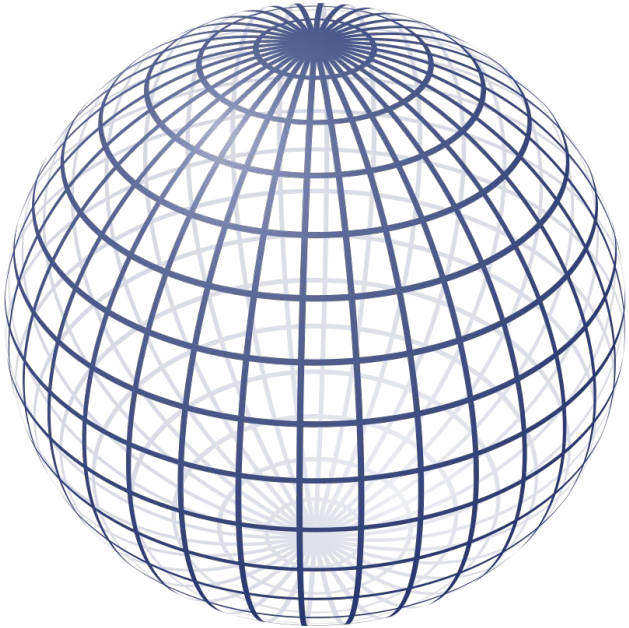
- **Topological invariants**
  - Connected components
  - Holes
  - Cavities

What can you extract from topology?

- **Topological invariants**
  - Connected components
  - Holes
  - Cavities

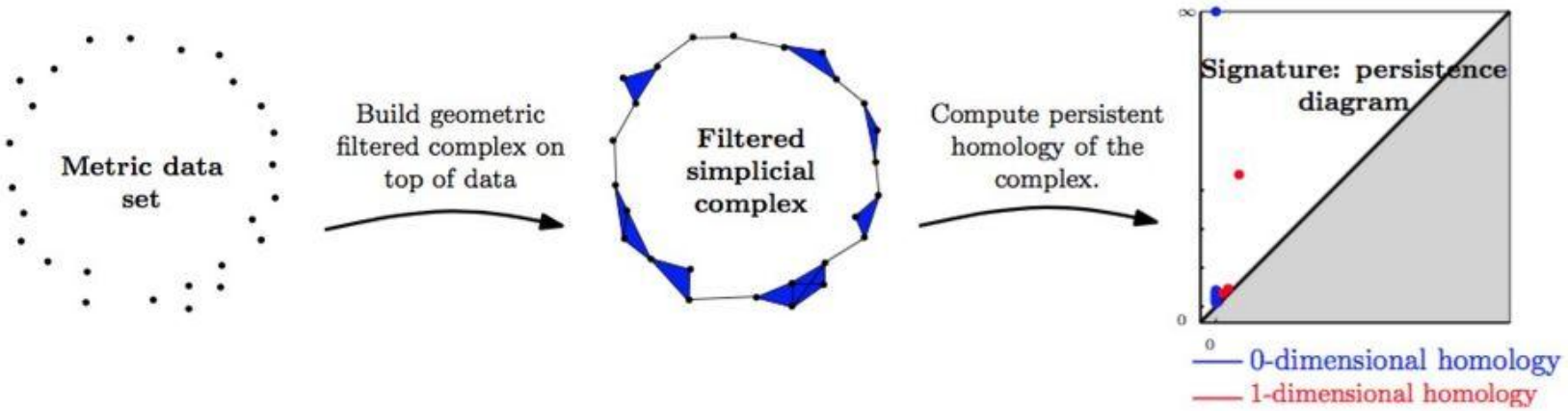
What can you extract from topology?

- **Topological invariants**
  - Connected components
  - Holes
  - Cavities



Everything is beautiful, but...

From topology to computational topology

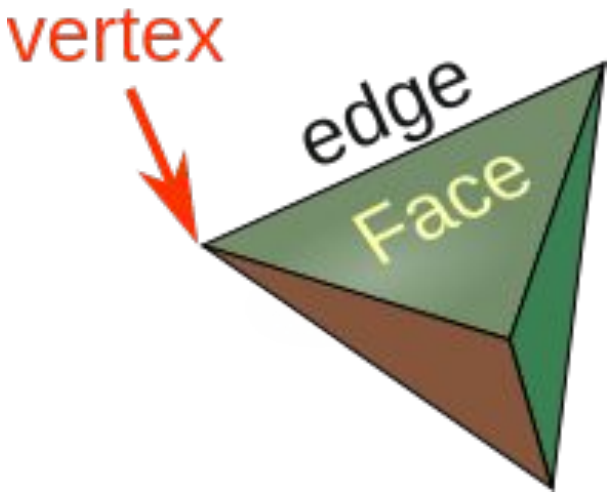


Chazal, Frédéric; Glisse, Marc; Labruère, Catherine; Michel, Bertrand (2013-05-27). "Optimal rates of convergence for persistence diagrams in Topological Data Analysis". arXiv:1305.6239 [math.ST].

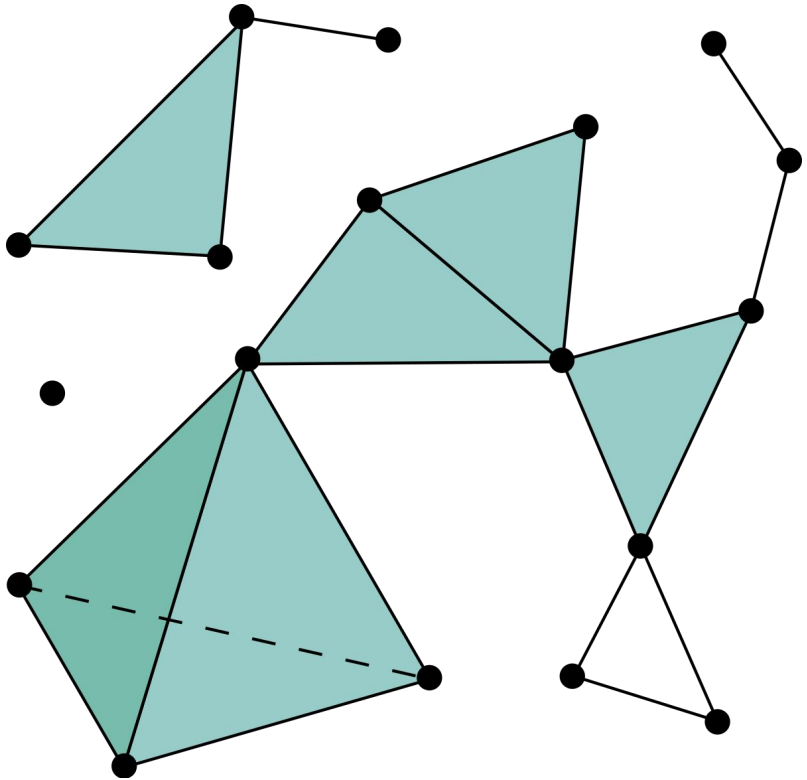
The basic structure - Simplices

A simplex is just an element from the following list:

Vertice, edge, face, tetrahedron, …

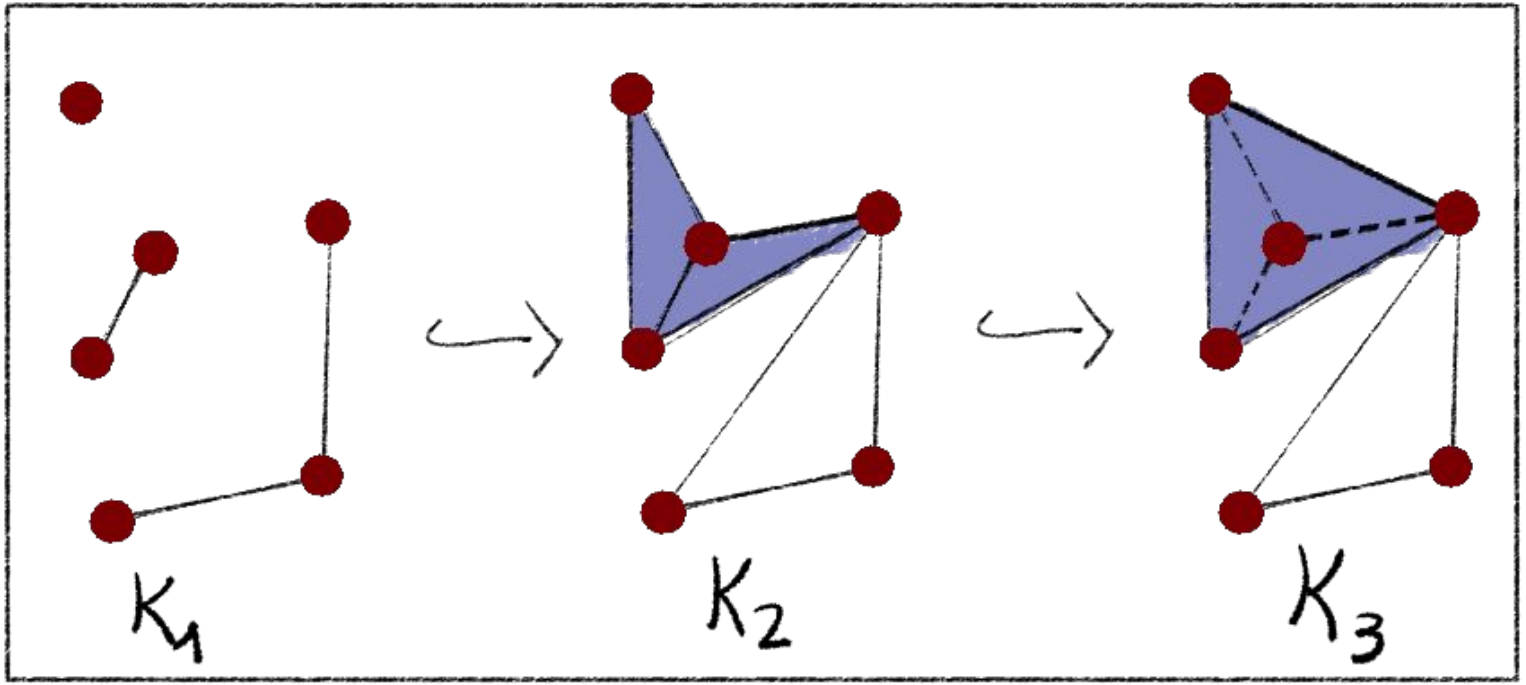


From simplices to a simplicial complex

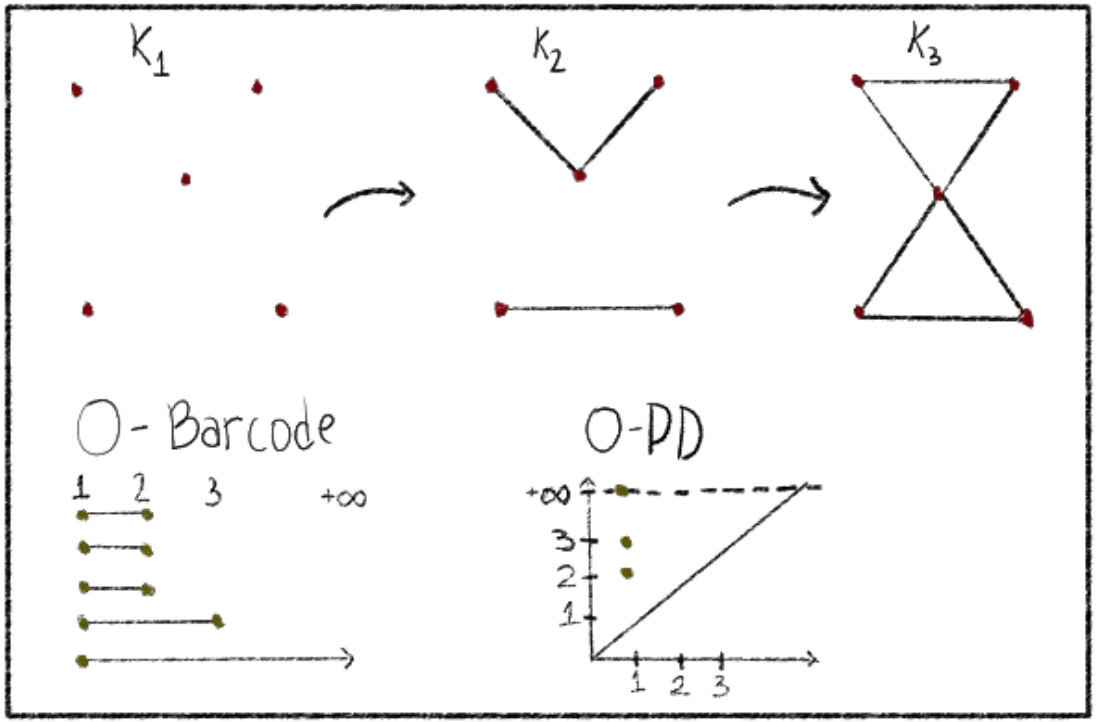


Now computers can work, but how do you interpret the simplicial complexes?

Extracting topological invariants - Filtration



Filtration and persistence diagrams - the duo.



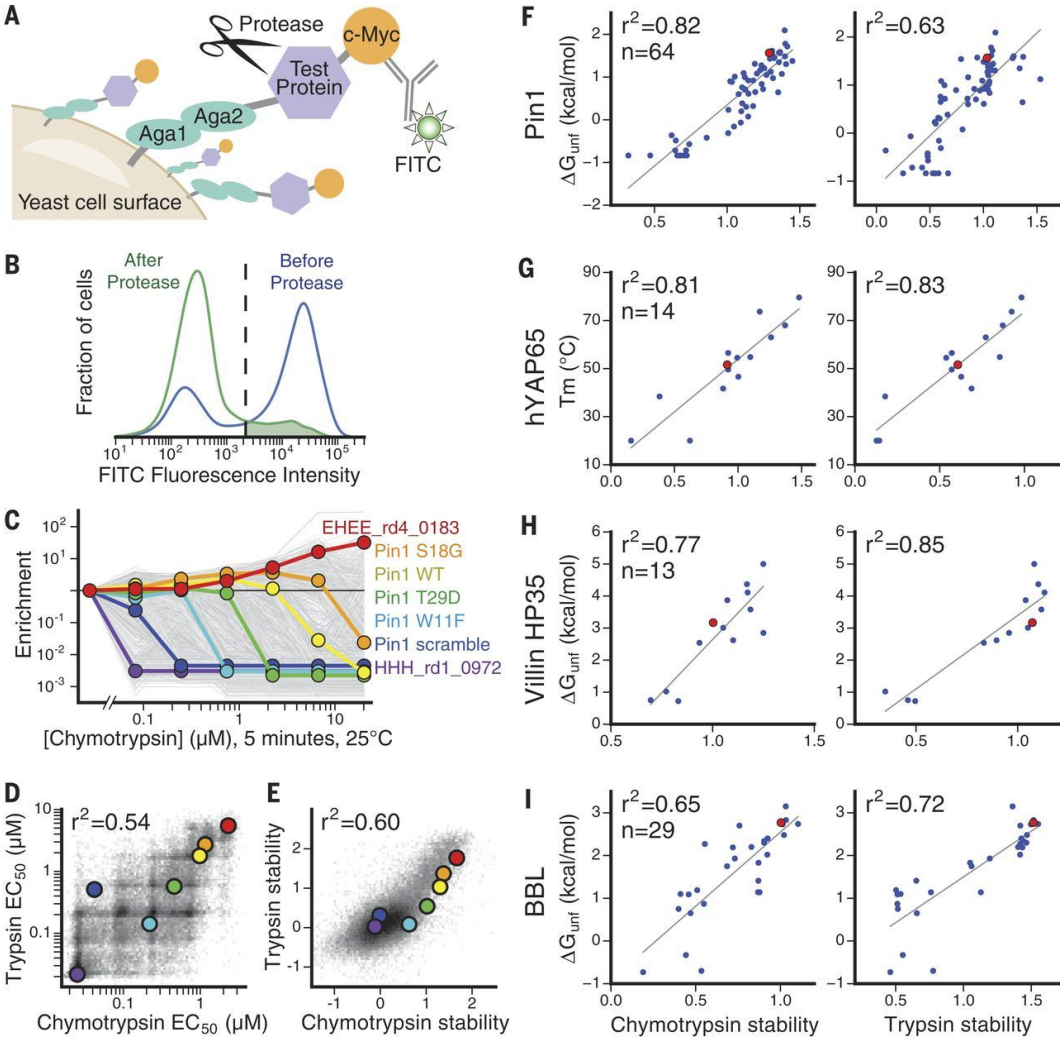
Some interesting applications

- Protein-ligand binding affinity prediction [1]
- Prediction of protein folding stability change upon mutation [2]
- Topological Data Analysis of Single-cell Hi-C Contact Maps [3]

[1] Cang Z, Wei G (2017) https://doi.org/10.1371/journal.pcbi.1005690
[2] Cang Z, Wei G (2017) https://doi.org/10.1093/bioinformatics/btx460
[3] Carriere M., Rabadan R. (2018) arXiv:1812.01360

Part II - persistent homology and protein stability

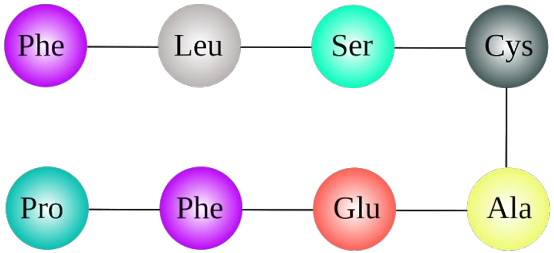
- Measurements of stability
- Computational development of proteins
- Feature extraction using TDA



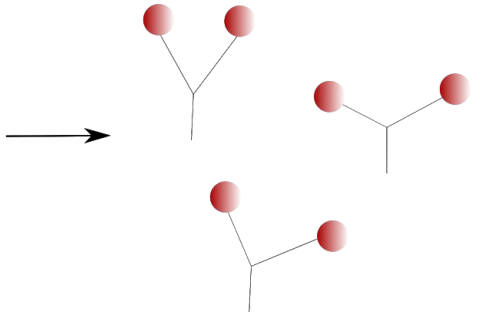
Science 14 Jul 2017:
Vol. 357, Issue 6347, pp. 168-175
DOI: 10.1126/science.aan0693

Computational steps to obtain a protein

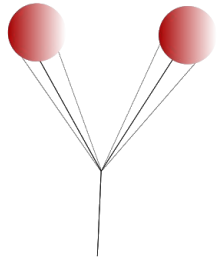
Amino acid sequence



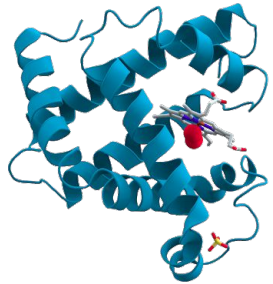
Structure Selection



Refinement



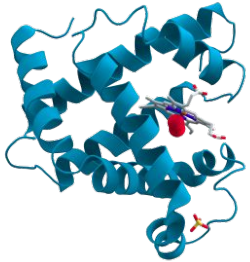
Final Protein



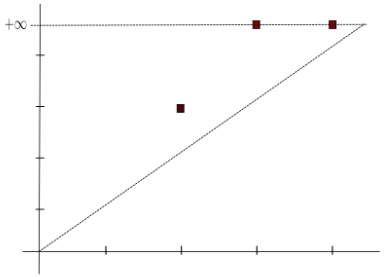
Predicting stability using physical and statistical terms of protein

Model	RMSE	Percent Error (%)
Rocklin model	0.419	11.381

Using TDA to extract features from proteins



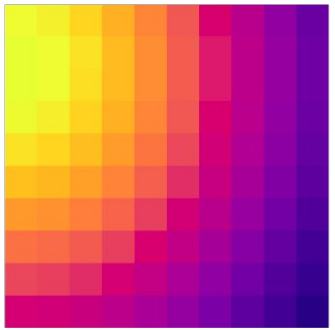
$\simeq 16000$

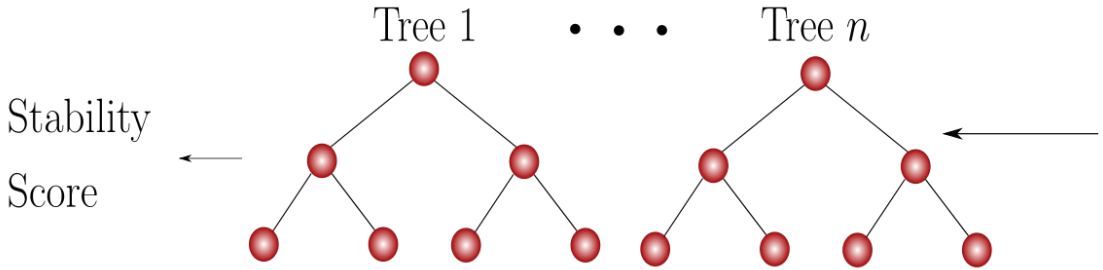


for each atom set in

$\{\{C\},\{O\},\{N\},\{C,O\},$
$\{C,N\},\{N,O\},\{C,N,O\}\}$

Persistence Images



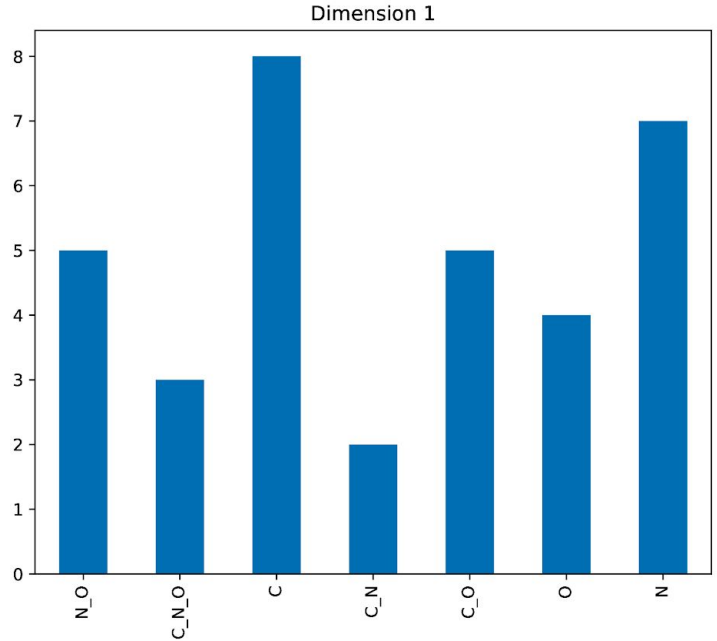


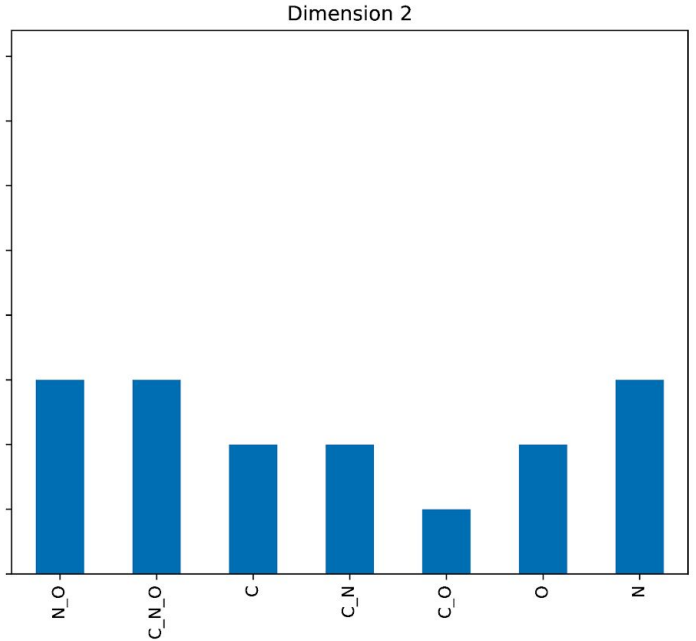
PH can be used to predict protein stability

Model	RMSE	Percent Error (%)
Linear Regression	0.5046	13.69
Random Forest I	0.4877	13.24
Random Forest II	0.4874	13.23
GBoost Optimal	0.4770	12.95

Unidimensional cycles of low persistence are important

Carbon and nitrogen atoms are related to stability scores

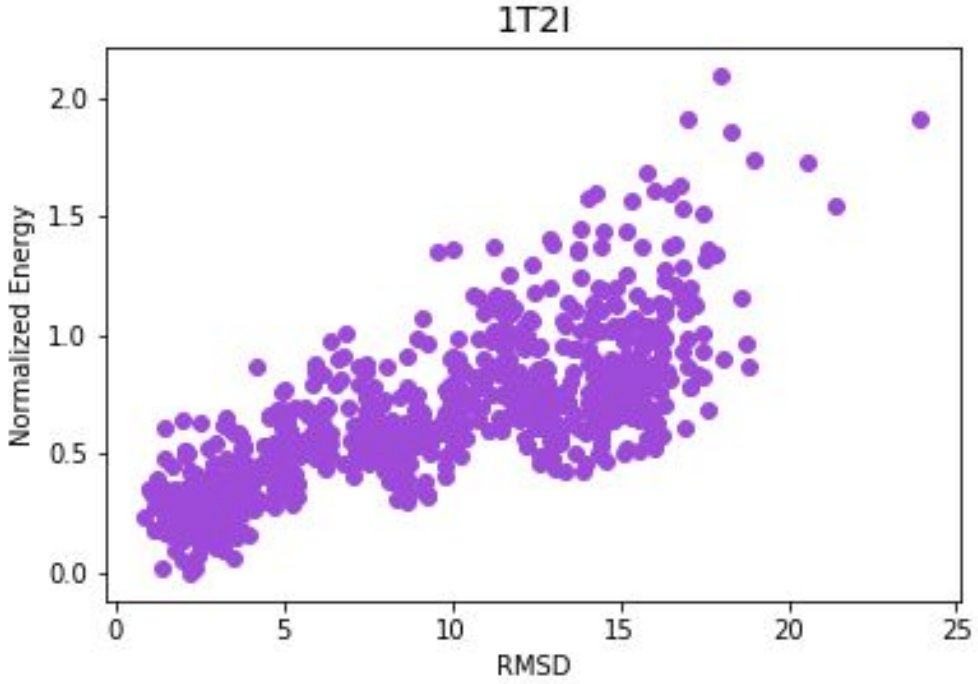




Part III - predicting RMSD

- Rosetta
- Template and simulated proteins
- RMSD: Root Mean Squared Deviation
- Prediction of RMSD: protein features x TDA

The energy landscape



The ranking of normalized energy

Rank	Normalized Energy	RMSD
1	0.000	2.233
2	0.023	1.37
3	0.025	2.395
4	0.057	2.004
5	0.061	2.356

TDA predicts RMSD better than protein terms

Metric	**Regressor**	**Pixel Size**	**Spread**	**Atom List**[1]	Mean Score
R^2	Neural Network	100	1.0	C	−5.780
MSE	Neural Network	100	1.0	C	8.299
RMSE	Ridge Regression	10	1.2	whole	2.599
Binary Accuracy	GBoost	10	0.6	N,O	0.657

TDA predicts RMSD better than protein terms

Metric	Regressor	Pixel Size	Spread	Atom List[1]	Mean Score
R^2	Neural Network	100	1.0	C	−5.780
MSE	Neural Network	100	1.0	C	8.299
RMSE	Ridge Regression	10	1.2	whole	2.599
Binary Accuracy	GBoost	10	0.6	N,O	0.657

Metric	Regressor	Score
R^2	Random Forest II	-13.706
MSE	Random Forest II	10.113
RMSE	Random Forest II	2.707
Binary Accuracy	Ridge Regression	0.586

Comparison between binary accuracy of regressors