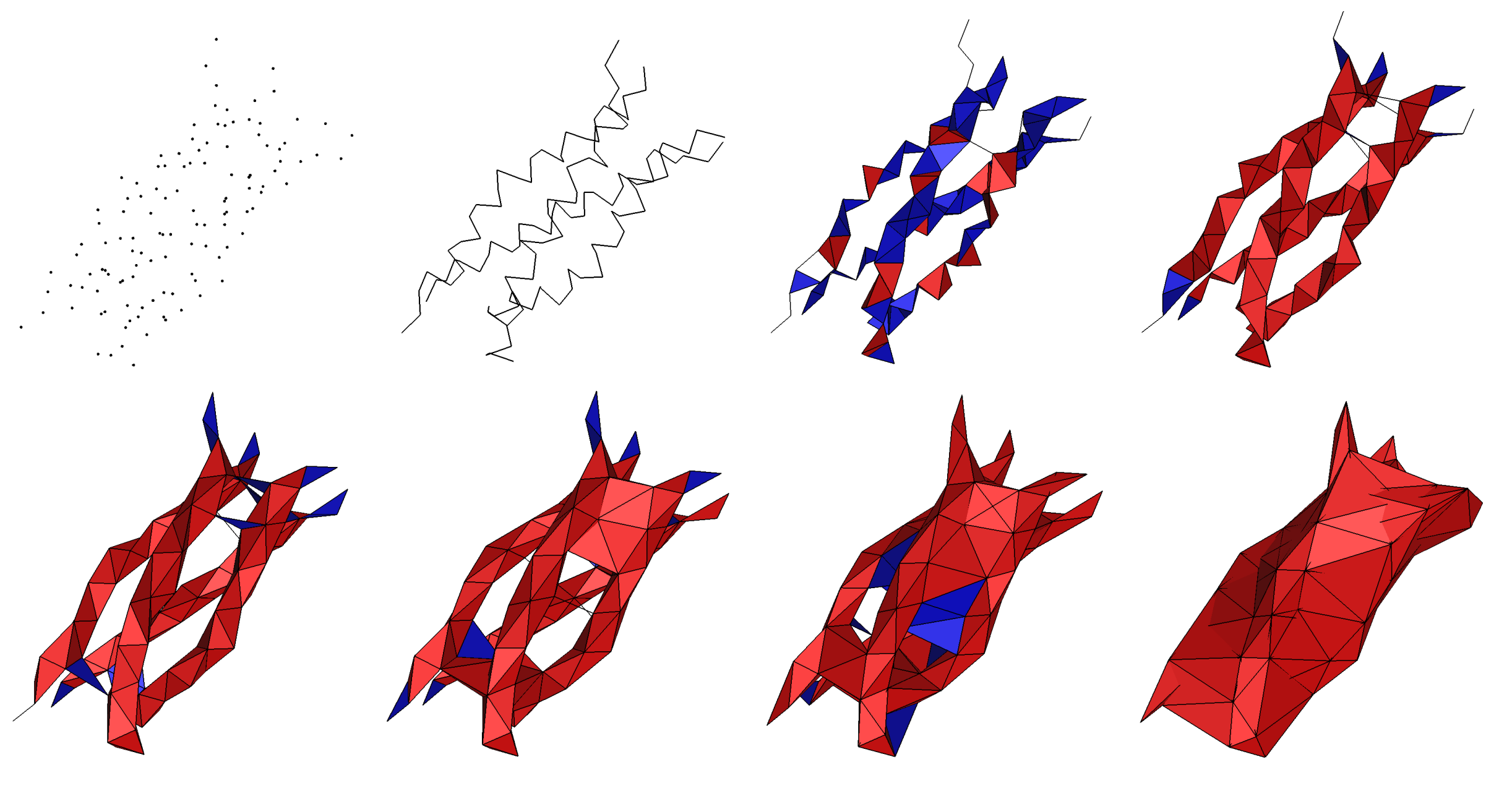
Topological Analysis and Prediction of Biomolecular Data

Zixuan Cang[1], Lin Mu[3], Kedi Wu[1], Kristopher Opron[2], Kelin Xia[1], and Guo-wei Wei[1,2]

[1]*Department of Mathematics, Michigan State University, MI 48824, USA*
[2]*Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA*
[3]*Oak Ridge National Laboratory, TN 37831, USA*

Introduction

Protein function and dynamics are closely related to its sequence and structure. However, the prediction of protein function and dynamics from its sequence and structure is still a fundamental challenge in molecular biology. Prediction of protein related observables provides advices for experiments and sheds light on how protein functions. Persistent homology is a new branch of algebraic topology that has found its success in the topological data analysis in a variety of disciplines, including molecular biophysics.



Filtration of simplicial complex

Filtration of Vietoris-Rips complex built on α-carbon point cloud of protein (ID:2LJC)

We explore the potential of using persistent homology as an independent tool for protein structure classification and protein-ligand/drug binding affinity prediction. From persistent homology computations, we extract protein topological fingerprints which are topological invariants generated during a filtration process. We develop topological machine learning in which feature vectors are generated solely based on the output of topological fingerprints.

Persistent Homology

In the past decade, persistent homology has been developed as a new multiscale representation of topological features.

Simplex A k-simplex denoted by σ^k is a convex hull of $k+1$ vertices which is represented by a set of points

$$\sigma^k = \{\lambda_0 u_0 + \lambda_1 u_1 + ... + \lambda_k u_k | \sum \lambda_i = 1, \lambda_i \geqslant 0, i = 0, 1, ..., k\},$$

where $\{u_0, u_1, ..., u_k\} \subset \mathbb{R}^n$ is a set of affinely independent points.

Simplicial complex A simplicial complex $\mathcal{K}$ is a finite collection of simplices satisfying two conditions. First, faces of a simplex in $\mathcal{K}$ are also in $\mathcal{K}$; Secondly, intersection of any two simplices in $\mathcal{K}$ is a face of both the simplices. The highest dimension of simplices in $\mathcal{K}$ determines dimension of $\mathcal{K}$.

Homology For a simplicial complex $\mathcal{K}$, a k-chain is a formal sum of the form $\sum_{i=1}^{N} c_i[\sigma_i^k]$, where $[\sigma_i^k]$ is oriented k-simplex from $\mathcal{K}$. A boundary operator ∂_k over a k-simplex σ^k is defined as,

$$\partial_k \sigma^k = \sum_{i=0}^{k} (-1)^i [u_0, u_1, ..., \widehat{u_i}, ..., u_k],$$

where $[u_0, u_1, ..., \widehat{u_i}, ..., u_k]$ denotes the face obtained by deleting the ith vertex in the simplex. The boundary operator induces a boundary homomorphism $\partial_k : C_k(\mathcal{K}) \to C_{k-1}(\mathcal{K})$. The composition operator $\partial_{k-1} \circ \partial_k$ is a zero map,

$$\begin{aligned}\partial_{k-1}\partial_k(\sigma^k) &= \sum_{j<i} (-1)^i(-1)^j [u_0, ..., \widehat{u_i}, ...\widehat{u_j}, ...u_k] + \sum_{j>i} (-1)^i(-1)^{j-1} [u_0, ..., \widehat{u_j}, ...\widehat{u_i}, ...u_k] \\ &= 0\end{aligned}$$

A sequence of chain groups connected by boundary operation form a chain complex,

$$\cdots \longrightarrow C_n(\mathcal{K}) \xrightarrow{\partial_n} C_{n-1}(\mathcal{K}) \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_1} C_0(\mathcal{K}) \xrightarrow{\partial_0} 0.$$

The equation $\partial_k \circ \partial_{k+1} = 0$ is equivalent to the inclusion $\mathrm{Im}\partial_{k+1} \subset \mathrm{Ker}\ \partial_k$, where Im and Ker denote image and kernel. Elements of $\mathrm{Ker}\partial_k$ are called kth cycle group, and denoted as Z_k=$\mathrm{Ker}\partial_k$. Elements of $\mathrm{Im}\partial_{k+1}$ are called kth boundary group, and denoted as B_k=$\mathrm{Im}\partial_{k+1}$. A kth homology group is defined as the quotient group of Z_k and B_k.

$$H_k = Z_k / B_k.$$

The kth Betti number of simplicial complex $\mathcal{K}$ is the rank of H_k,

$$\beta_k = \mathrm{rank}(H_k) = \mathrm{rank}(Z_k) - \mathrm{rank}(B_k).$$

Betti number β_k is finite number, since $\mathrm{rank}(B_p) \leqslant \mathrm{rank}(Z_p) < \infty$. Betti numbers computed from a homology group are used to describe the corresponding space.

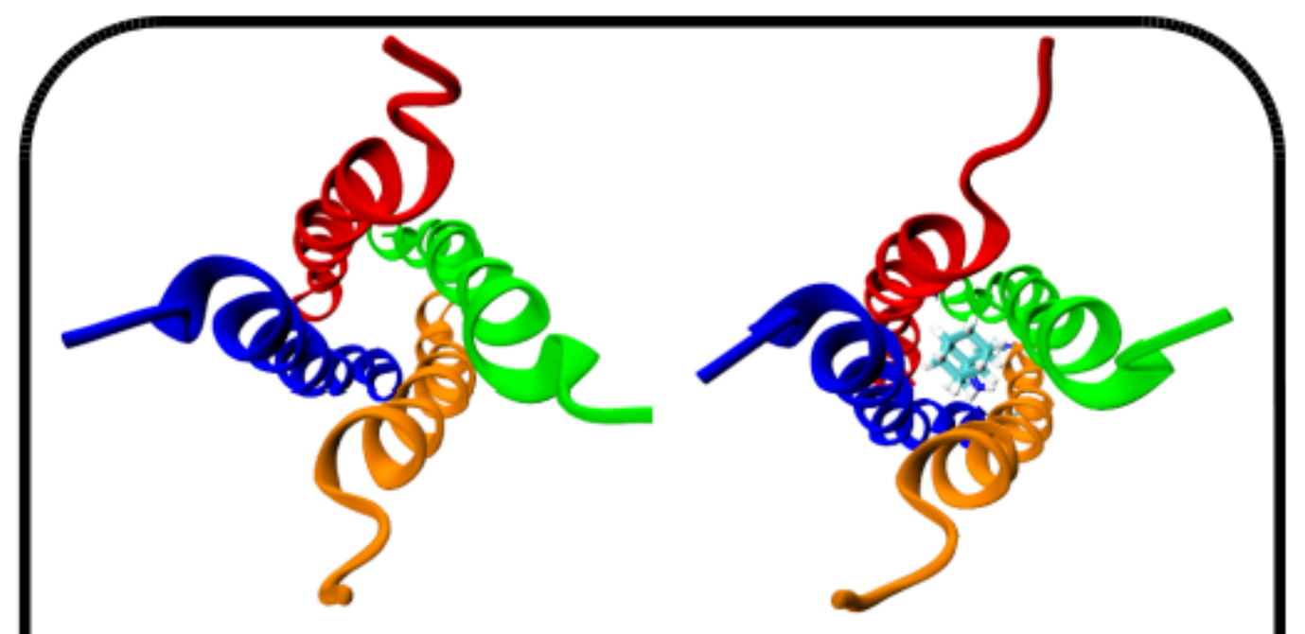
Feature construction for structure classification tasks

A number of features describing properties of the samples in different scales are extracted from persistent homology bar codes. Illustrated below are some examples of features used in machine learning.

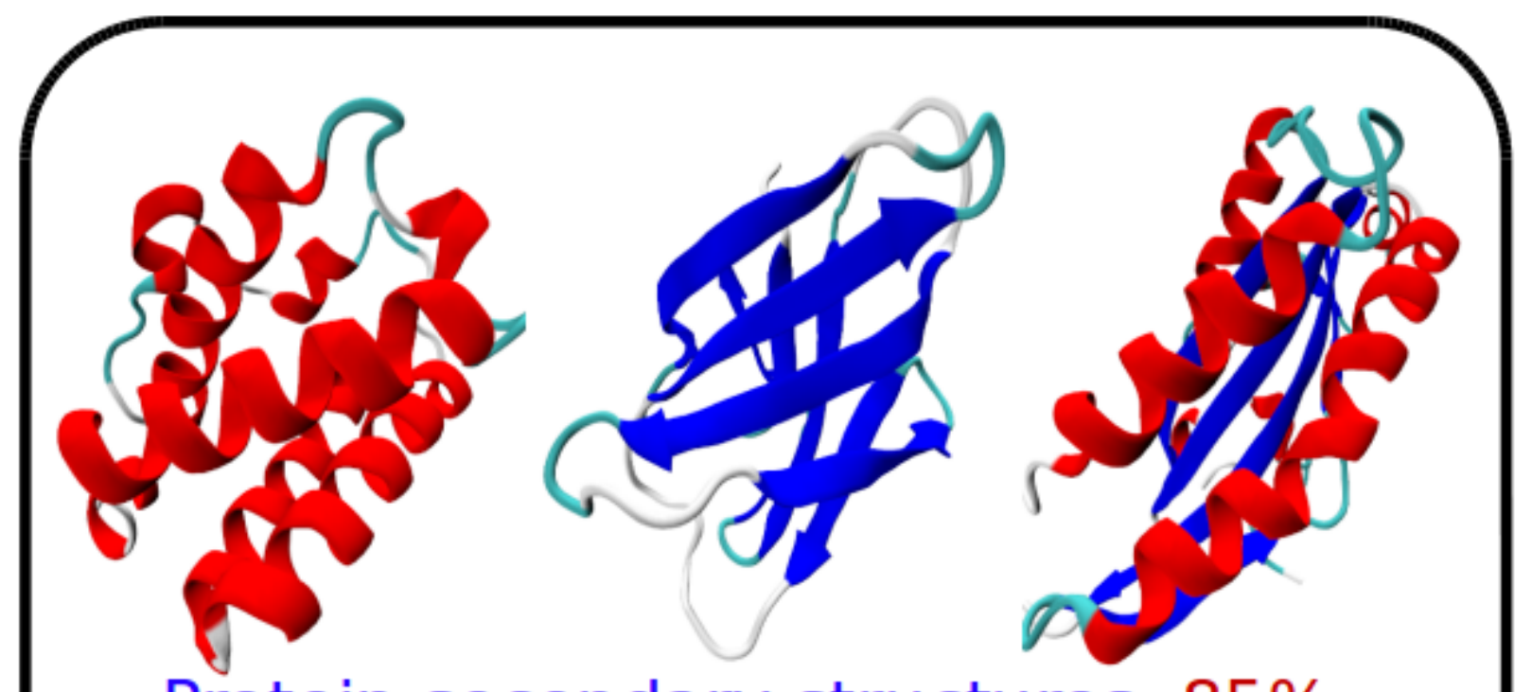
- The length of the second longest Betti 0 bar.
- The summation of lengths of all Betti 0 bars except for those exceed the max filtration value.
- The average length of Betti 0 bars except for those exceed the max filtration value.
- The onset value of the longest Betti 1 bar.
- The number of Betti 1 bars that locate at [4.5, 5.5], divided by the number of atoms.
- The onset value of the first Betti 2 bar that ends after a given number.

Bar codes in different dimension with different lifespan, birth time, and death time show characteristics of the sample from different scales. With persistent homology, we are able to describe local properties like alpha helices or beta sheets and global properties like size of the cavity of a spherical structure or size of tunnel of a cylindrical structure.

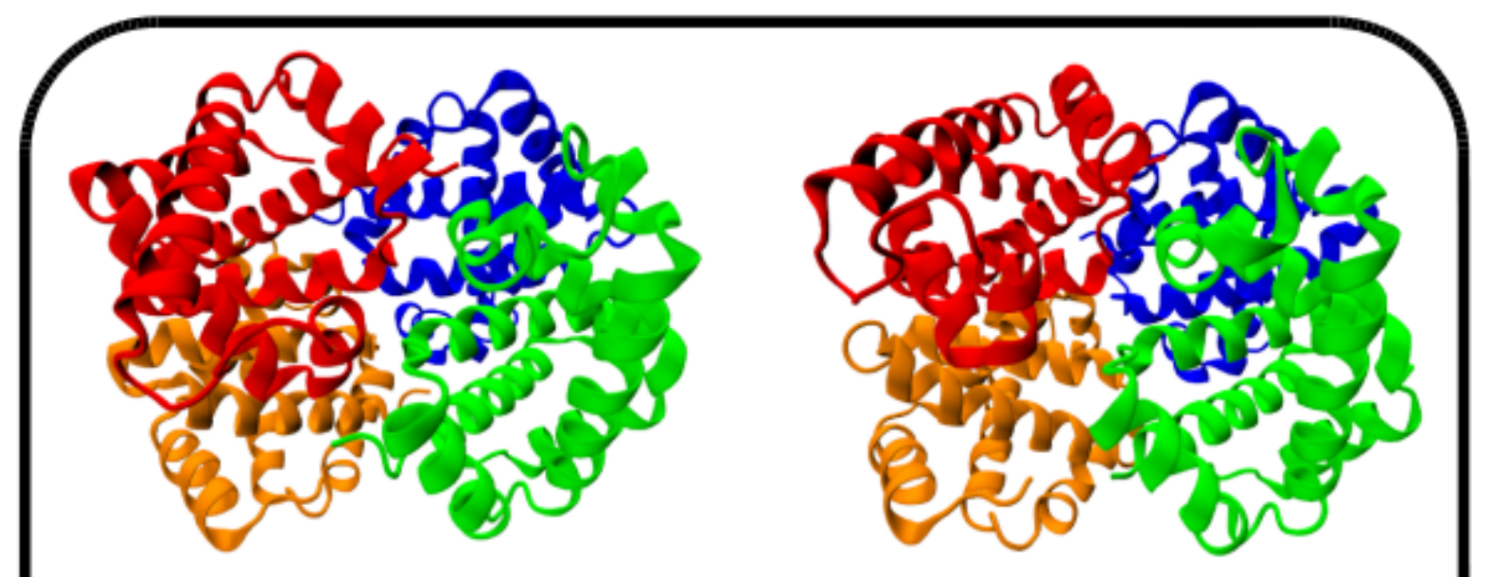
Performance on Classification Tasks



Influenza A virus drug inhibition: 96% Accuracy



Protein secondary structures: 85% Accuracy

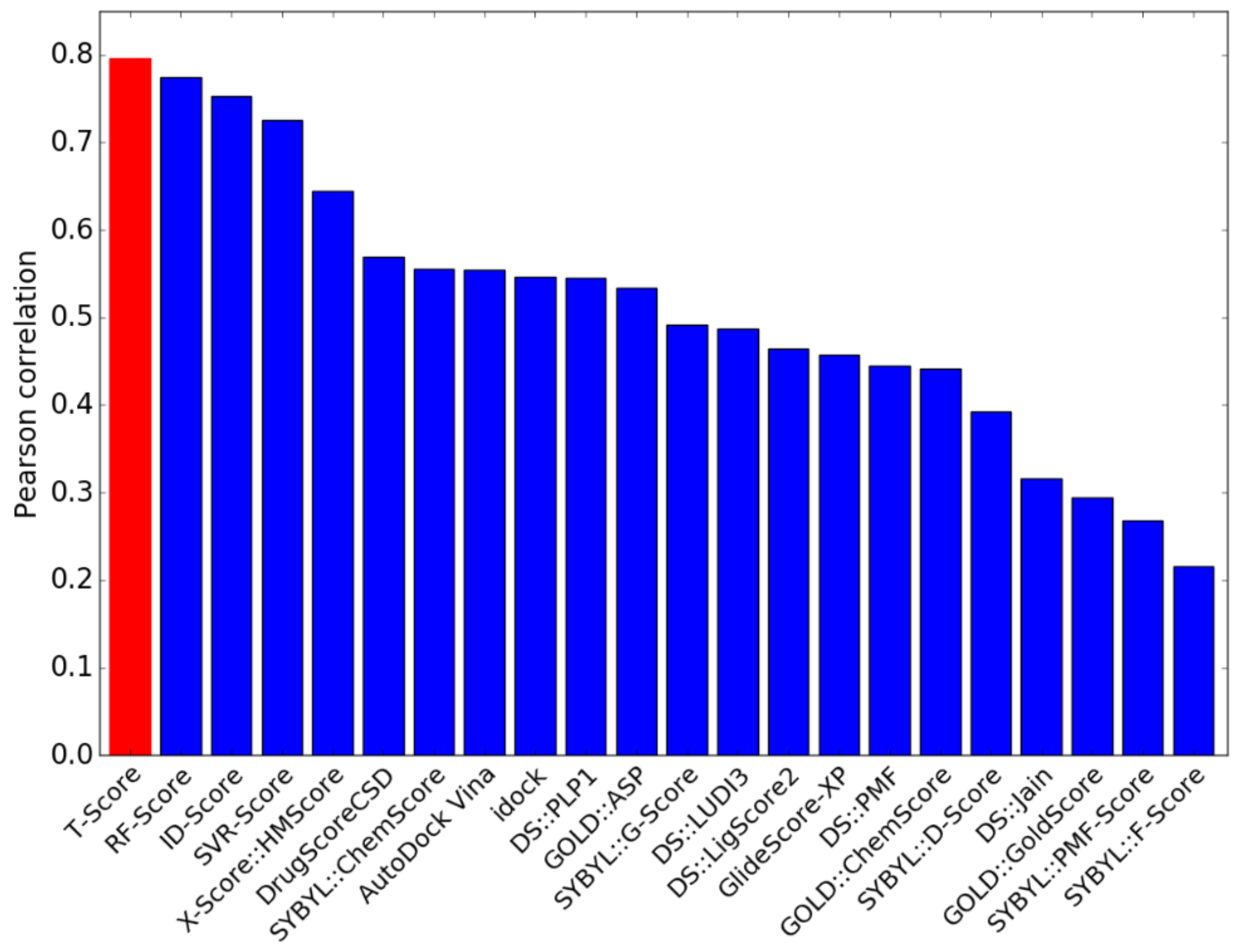


Hemoglobins in their relaxed and taut forms: 80% accuracy

SCOPe

55 classification tasks of protein superfamilies over 1357 proteins from *Protein Classification Benchmark Collection*: 82% Accuracy

Performance on Binding Affinity Prediction

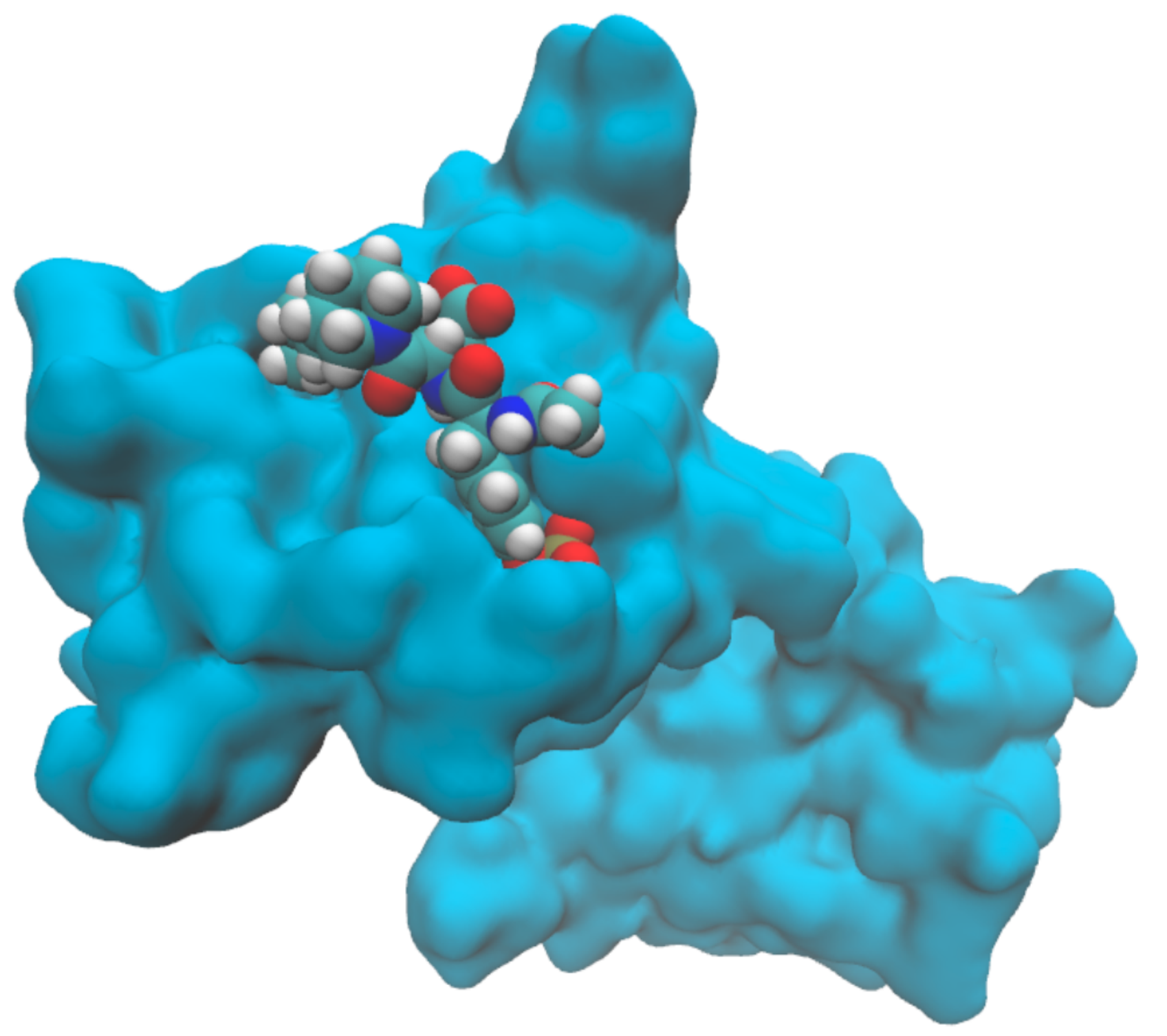


The persistent homology based protein-ligand/drug binding affinity predictor named T-Score is tested on the PDBBind v2007 core set with the v2007 refined set as a training set where the testing set has been excluded from the training set. A high Pearson correlation of 0.80 is achieved and our T-Score outperforms all the other eminent methods in computational biophysics.

References

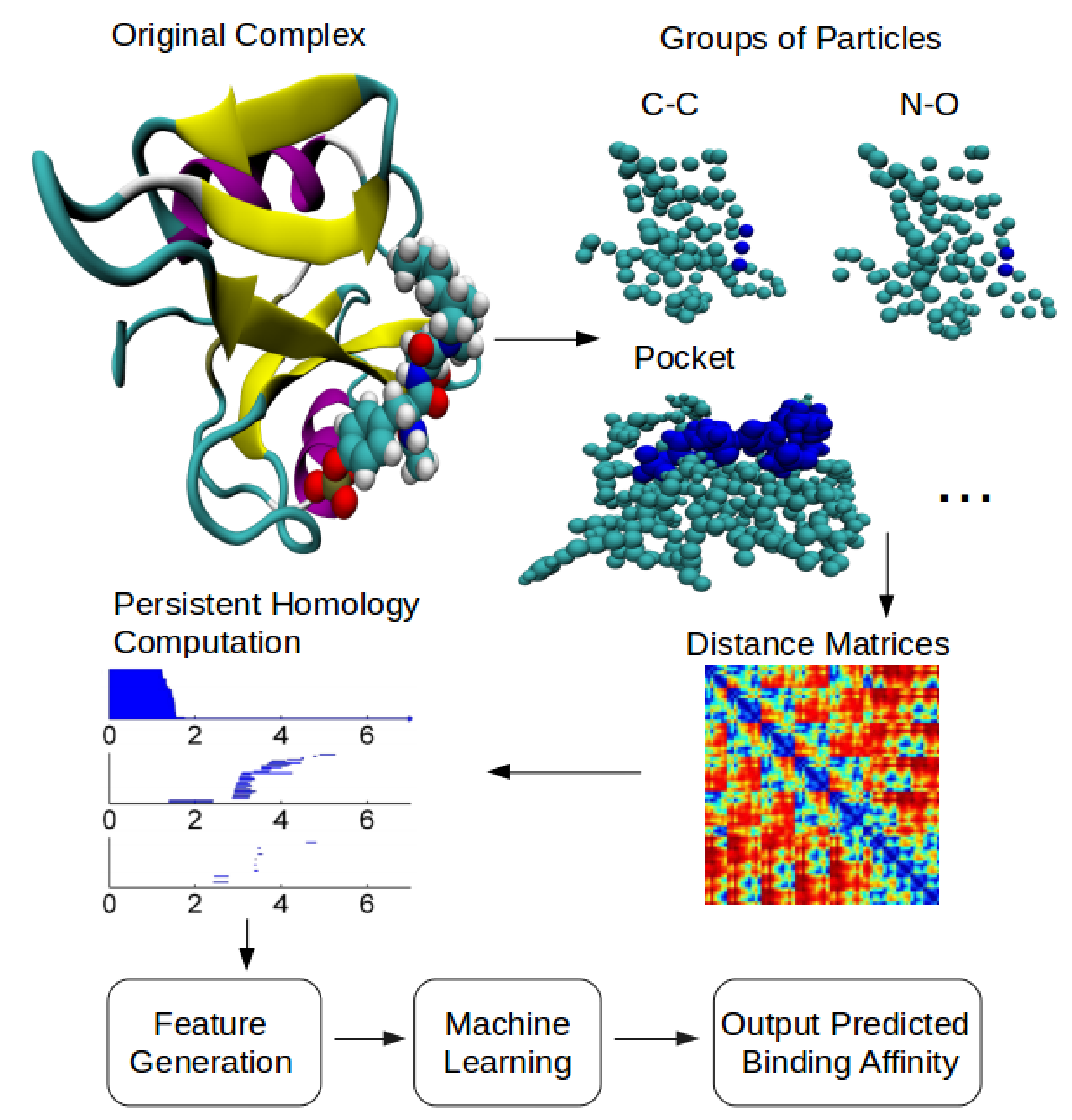
- Zixuan Cang, Lin Mu, Kedi Wu, Kristopher Opron, Kelin Xia and Guo-Wei Wei, "A topological approach to protein classification", Molecular Based Mathematical Biology, 3, 140-62 (2015).
- Kelin Xia and Guo-Wei Wei, "Persistent homology analysis of protein structure, flexibility and folding", International Journal for Numerical Methods in Biomedical Engineering, 30(8):814-844 (2014).
- Kelin Xia, Zhixiong Zhao and Guo-Wei Wei,"Multiresolution persistent homology for excessively large biomolecular datasets", Journal of Chemical Physics, 143, 134103 (2015).

Protein-Ligand/Drug Binding Free Energy



Structure-based drug design relies on computational methods to identify and optimize potential drugs. Molecule docking is the most widely used approach which predicts the location and orientation (pose) of a ligand bound to a protein to form a stable complex. In the process of search for a pose for the ligand, a scoring function which measures the binding affinity between the two molecules is needed to distinguish the favorable poses from the unfavorable ones. An accurate and efficient protein-ligand/drug binding affinity predictor is therefore the key of molecule docking process. As the dominating forces that regulates protein-ligand binding are mainly weak forces which heavily depends on spacial arrangements, persistent homology becomes a competitive candidate for this job.

Flowchart of Persistent Homology Based Protein-Ligand Binding Affinity Prediction



Conclusion

We test the performance of persistent homology in various protein structure classification tasks as well as protein-ligand binding/drug affinity prediction. It is found that persistent homology is able to offer a representation of proteins and capture their intrinsic interactions. Our persistent homology based T-Score outperforms all the other eminent methods in computational biophysics on the blind prediction of protein-ligand/drug binding affinities.

Acknowledgment

This work was supported in part by NSF grants IIS-1302285, and DMS-1160352, and NIH Grant R01GM-090208.

http://www.math.msu.edu/~wei/ cangzixu@math.msu.edu mul1@ornl.gov wukedi@msu.edu opronkri@msu.edu xiakelin@msu.edu wei@math.msu.edu