Persistent Spectral Theory

Nick Sale - Swansea TDA Seminar - 14/04/20

- Zhenyu Meng and Kelin Xia, **Persistent spectral based machine learning (PerSpect ML) for drug design**, 2020, https://arxiv.org/abs/2002.00582
- Rui Wang, Duc Duy Nguyen and Guo-Wei Wei, **Persistent spectral graph**, 2019, https://arxiv.org/abs/1912.04135

Spectral Graph Theory

Consider a graph $G = (V, E)$

Definition: Adjacency matrix $$A_{i,j} = \begin{cases} 1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

Definition: Graph Laplacian $$L_{i,j} = \begin{cases} deg(v_i) & \text{if } i = j \\ -1 & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

Note that $deg(v_i) = \sum_j A_{i,j}$, and we can also write $L = D - A$ where $D = diag(\{v\}_{v \in V})$

Properties of the Laplacian

Consider $f : V \to \mathbb{R}$ as a vector in $\mathbb{R}^n$ with components f_i

Proposition: $f^T L f = \sum_{(v_i, v_j) \in E} (f_i - f_j)^2$

Corollary: L is positive semi-definite, so all its eigenvalues are non-negative

Since L is symmetric, the spectral theorem also tells us it has n linearly-independent eigenvectors with eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$

Proposition: $dim(ker L) = \beta_0$ (# of connected components)

Hence G is connected iff $\lambda_2 \geq 0$. In fact, λ_2 gives a “measure of connectedness”

Example: What can the other eigenvalues tell us?

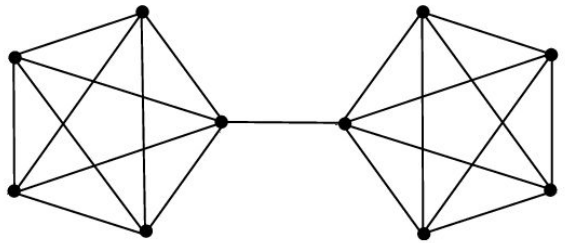
Definition: Cheeger constant
(also Isoperimetric ratio)

$$h(G) = \min_{A \subseteq V,\ |A| \leq \frac{1}{2}|V|} \frac{|\partial A|}{|A|}$$

where $\partial A = \{(v, w) \in E \mid v \in A,\ w \notin A\}$

Theorem: (Cheeger-Alon-Milman) Let d_{max} be the maximum degree of any vertex. Then

$$\frac{\lambda_2}{2} \leq h(G) \leq \sqrt{2 d_{max} \lambda_2}$$



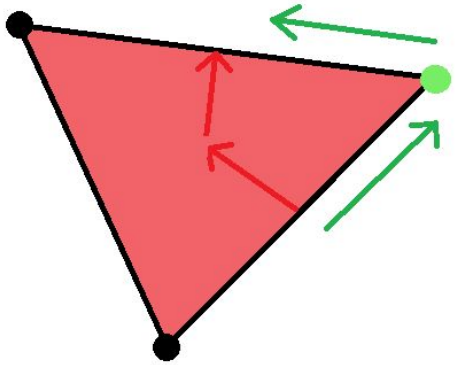
Extension to Simplicial Complexes

Consider a chain complex (C_i, ∂_i)

Definition: k-th Combinatorial Laplacian $L_k = \partial_k^T \partial_k + \partial_{k+1} \partial_{k+1}^T$

Note that when $k = 0$ we obtain the same definition as before

Proposition: $dim(ker L_k) = \beta_k$



We can also relate the combinatorial Laplacian to random walks on simplicial complexes (see *Random walks on simplicial complexes and the normalized Hodge 1-Laplacian* by Shaub et al.)

Persistent Spectral Theory

Idea: PerSpect (*Persistent spectral based machine learning (PerSpect ML) for drug design*)

$$K^1 \subseteq K^2 \subseteq \ldots \subseteq K^n$$

filtration of a simplicial complex

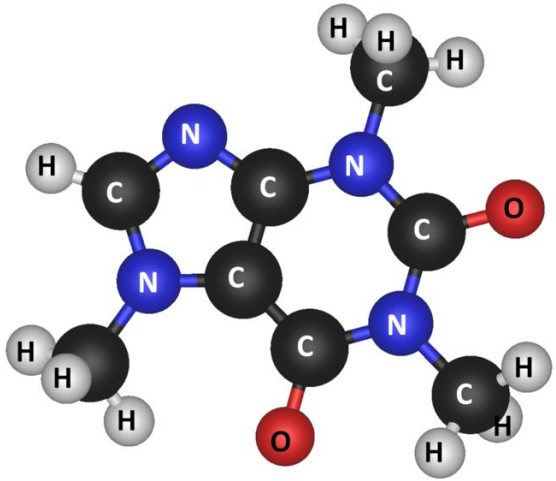
$$L_k^1, L_k^2, \ldots, L_k^n$$

k-th combinatorial Laplacians

$$\{\lambda_2^1,\ \textstyle\sum_i \lambda_i^1,\ \sum_i |\lambda_i^1 - \overline{\lambda^1}|,\ \ldots\},\ \ldots,\{\lambda_2^n,\ \sum_i \lambda_i^n,\ \sum_i |\lambda_i^n - \overline{\lambda^n}|,\ \ldots\}$$

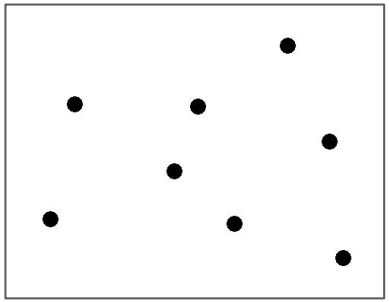
spectral features

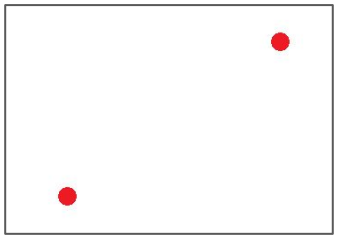
Biomolecular Topological Modeling

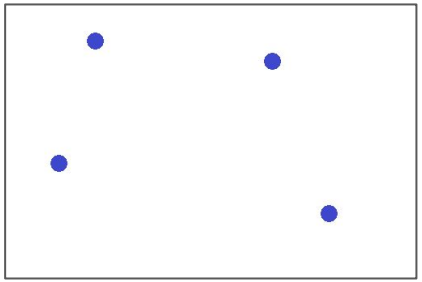


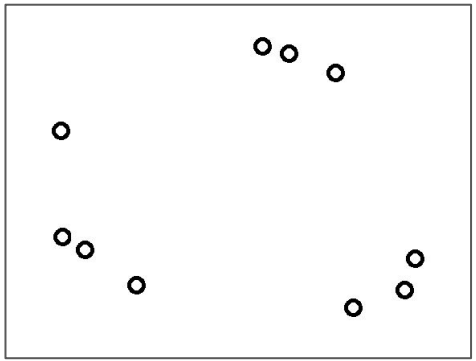
Model: Element-specific (ES) modelling

Given the spatial configuration of a molecule, consider separate point clouds for each element

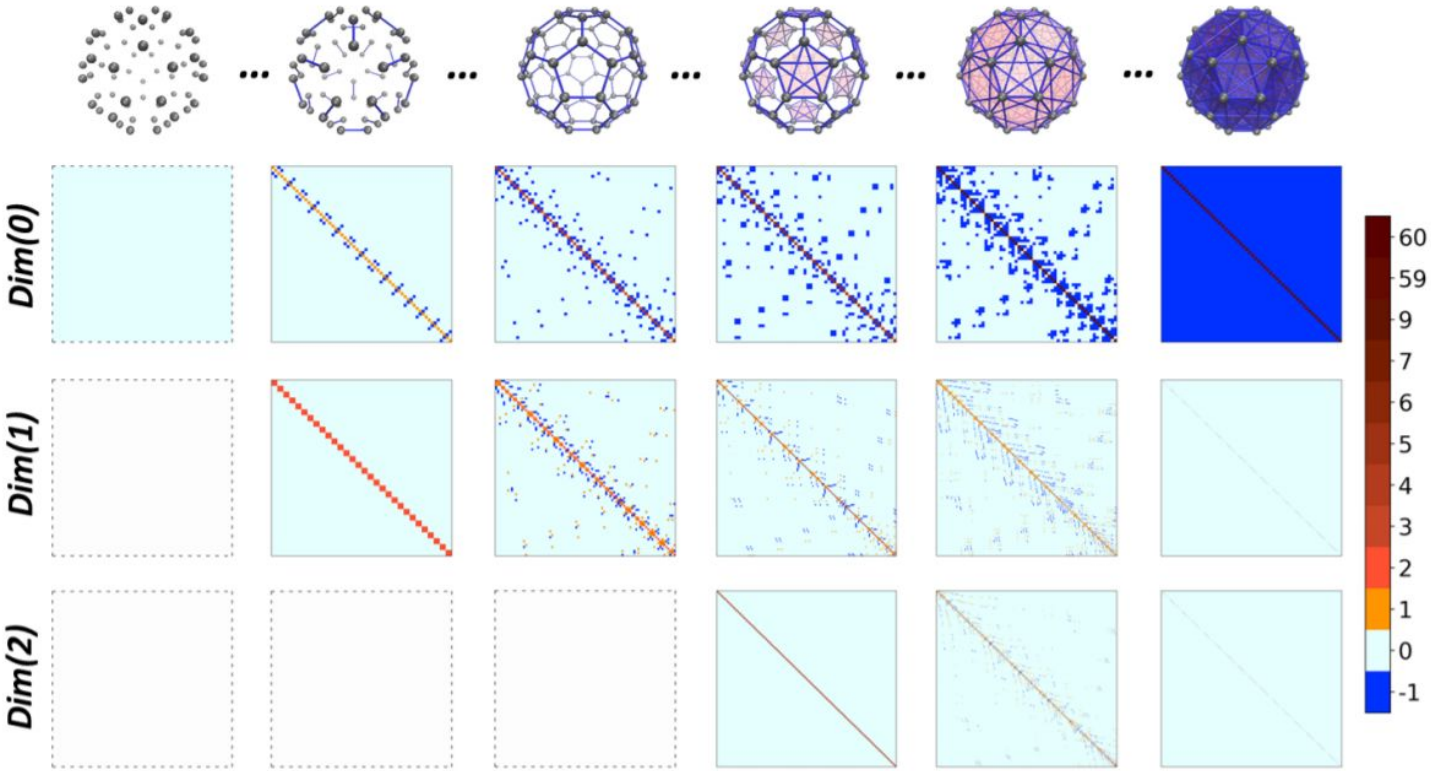




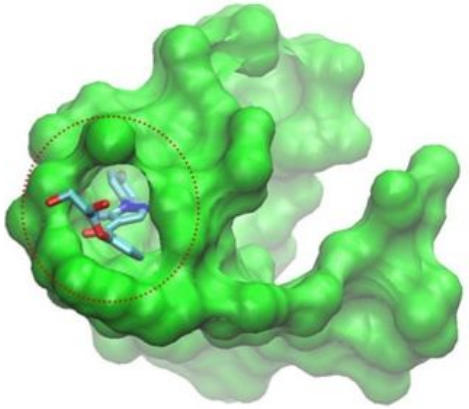




Example: Buckyball C_{60}

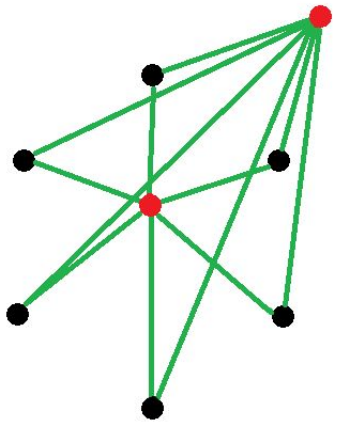


Modelling Interactions



In drug design we want to look at the interaction between two molecules: a protein and a ligand. In particular, we might want to predict the binding affinity

Take element point clouds R_P, R_L from the protein and ligand



Definition: ES interactive distance

$$d(x,y) = \begin{cases} \|x-y\| & \text{if } x \in R_P, y \in R_L \text{ or } y \in R_P, x \in R_L \\ \infty & \text{otherwise} \end{cases}$$

Definition: ES interactive electrostatic distance

$$d_E(x,y) = \begin{cases} (1 + exp(\frac{cq_xq_y}{\|x-y\|}))^{-1} & \text{if } x \in R_P, y \in R_L \text{ or } y \in R_P, x \in R_L \\ \infty & \text{otherwise} \end{cases}$$

Predicting Binding Affinity

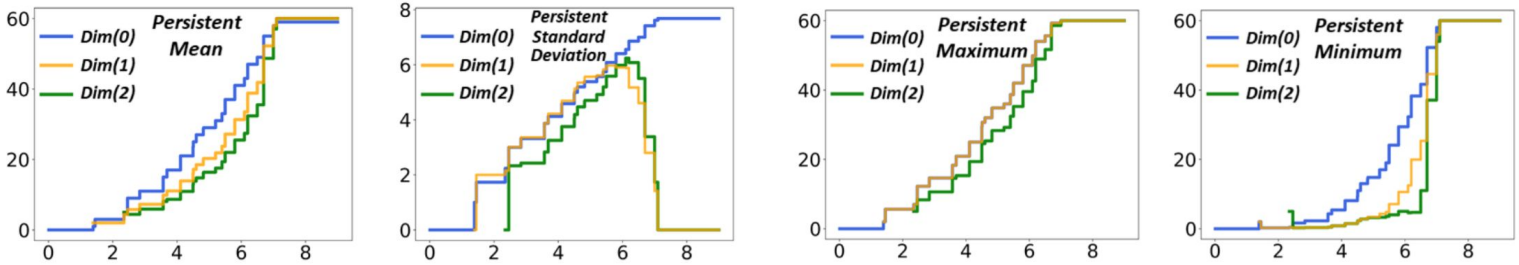
Given protein P and ligand L

We have 4 protein element point clouds
C, N, O, S

And 9 ligand element point clouds
C, N, O, S, P, F, Cl, Br, I

Giving 4*9 = 36 Vietoris-Rips filtrations $VR_d(C_P \cup C_L),\ VR_d(C_P \cup N_L),\ \ldots$

11*250*36 = 99,000 features computed (11 features at 250 filtration values)



Similarly, 11*100*50 = 55,000 features are computed using d_E

Spectral Features

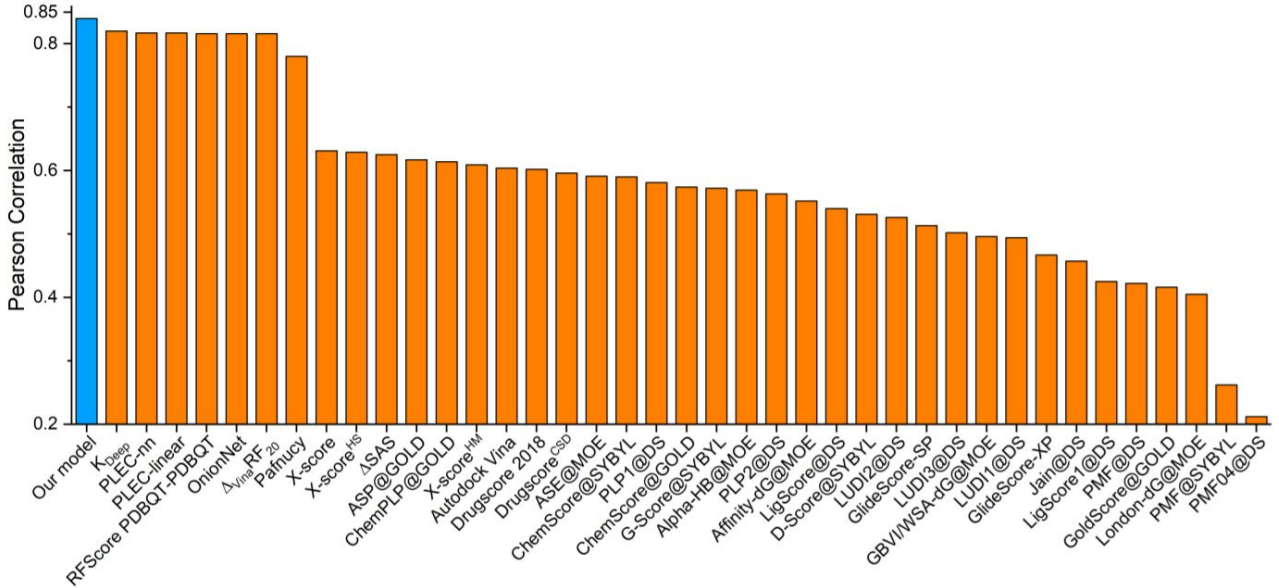
1. Betti 0
2. Betti 1
3. Mean
4. Standard deviation
5. Maximum
6. Minimum
7. Laplacian graph energy
8. Generalized mean graph energy
9. Spectral moment (second order)
10. Quasi-Wiener index
11. Spanning tree number

$$E(G) = \sum_i \lambda_i \qquad GME(G) = \sum_i |\lambda_i - \bar{\lambda}| \qquad SM_k(G) = \sum_i (\lambda_i)^k$$

$$QWI(G) = \sum_{\lambda \neq 0} \frac{|\{\lambda \neq 0\}|+1}{\lambda} \qquad ST(G) = lg(\frac{1}{|\{\lambda \neq 0\}|+1} \prod_{\lambda \neq 0} \lambda)$$

Performance

Trained using Gradient Boost Trees on data from the PDBbind-2007, PDBbind-2013 and PDBbind-2016 databases



Persistent Laplacians

Similar to ideas in the presentation *Persistent harmonic forms* by André Lieutier (https://project.inria.fr/gudhi/files/2014/10/Persistent-Harmonic-Forms.pdf)

Idea: (*Persistent spectral graph*) Consider a filtration of chain complexes

$$\begin{array}{ccccccccccccccc}
\cdots & C^1_{q+1} & \underset{\partial^{1*}_{q+1}}{\overset{\partial^1_{q+1}}{\rightleftarrows}} & C^1_q & \underset{\partial^{1*}_{q}}{\overset{\partial^1_{q}}{\rightleftarrows}} & \cdots & \underset{\partial^{1*}_{3}}{\overset{\partial^1_{3}}{\rightleftarrows}} & C^1_2 & \underset{\partial^{1*}_{2}}{\overset{\partial^1_{2}}{\rightleftarrows}} & C^1_1 & \underset{\partial^{1*}_{1}}{\overset{\partial^1_{1}}{\rightleftarrows}} & C^1_0 & \underset{\partial^{1*}_{0}}{\overset{\partial^1_{0}}{\rightleftarrows}} & C^1_{-1} \\
 & \cap & & \cap & & & & \cap & & \cap & & \cap & & \\
\cdots & C^2_{q+1} & \underset{\partial^{2*}_{q+1}}{\overset{\partial^2_{q+1}}{\rightleftarrows}} & C^2_q & \underset{\partial^{2*}_{q}}{\overset{\partial^2_{q}}{\rightleftarrows}} & \cdots & \underset{\partial^{2*}_{3}}{\overset{\partial^2_{3}}{\rightleftarrows}} & C^2_2 & \underset{\partial^{2*}_{2}}{\overset{\partial^2_{2}}{\rightleftarrows}} & C^2_1 & \underset{\partial^{2*}_{1}}{\overset{\partial^2_{1}}{\rightleftarrows}} & C^2_0 & \underset{\partial^{2*}_{0}}{\overset{\partial^2_{0}}{\rightleftarrows}} & C^1_{-1}
\end{array}$$

Definition: p-persistent k-th Laplacian $L_k^{t+p} = \bar{\partial}_{k+1}^{t+p}(\bar{\partial}_{k+1}^{t+p})^T + (\partial_k^t)^T \partial_k^t$

where $A_k^{t+p} = \{\sigma \in C_k^{t+p} \mid \partial\sigma \in C_{k-1}^t\}$ $\quad \bar{\partial}_k^{t+p} = \partial_k^{t+p}|_{A_k^{t+p}}$

Proposition: $dim(ker L_k^{t+p}) = \beta_k^{t+p}$

Questions

- What level of stability do we have for these spectral features?
- If none, how much do we actually miss it in applications like these?
- Aren’t these spectral features very expensive to compute lots of times? What methods for computing these faster for filtrations might there be>

Image URLs

- https://courses.lumenlearning.com/introchem/chapter/molecules/
- https://www.creative-proteomics.com/services/protein-ligand-binding-site-prediction-service.htm