Mathematical AI for Molecular Sciences

JunJie Wee[1], **Kelin Xia**[1*]

[1]*Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371*

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Physical and Mathematical Sciences
College of Science

Introduction

Artificial intelligence (AI) has taken huge leaps of advancement in molecular sciences through the massive accumulations of molecular data, computational power, and learning models. The breakthrough of AlphaFold 2 in protein folding heralds a new age of AI-based molecular data analysis in materials, chemical, and biological sciences. However, one of the main challenges in AI-based molecular sciences is to construct effective molecular descriptors and fingerprints. To tackle this main challenge, we propose several new persistent functions for molecular featurisation. By representing molecular structures such as proteins, DNA, protein-ligand complex and protein-protein complexes as graphs, simplicial complex or hypergraphs, we introduce persistent functions (PFs) such as persistent Ricci curvature and persistent spectral to track the changes in their underlying topology and geometry in a filtration process. PFs are converted into suitable input features for machine learning models to predict quantitative molecular properties. In general, our models have demonstrated a great advantage over existing models in binding affinity predictions.

Molecular data	Representation	Featurisation	Learning
Materials	Topology	Topological invariants	Machine learning
Chemistry	Geometry	Geometric invariants	Deep Learning
Biology	Combinatorics	Combinatorial invariants	Statistical Learning
	Algebra	Algebraic invariants	

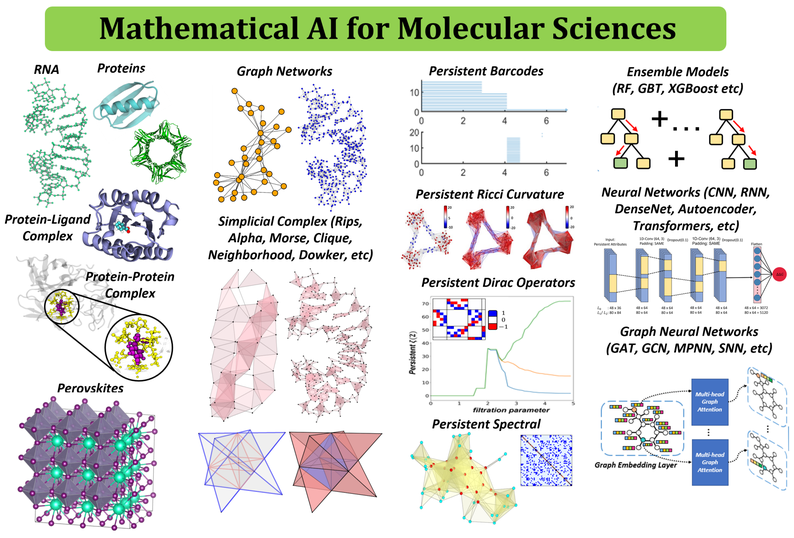

Figure: Machine learning and deep learning pipeline for mathematical AI in molecular sciences.

Ollivier Ricci curvature and Forman Ricci curvature

- **Vertex Probability Measure:**

$$m_x^\alpha(x_i) = \begin{cases} \alpha & \text{if } x_i = x. \\ (1-\alpha)/k_x & \text{if } x_i \in \Gamma_x. \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here, α is the proportion of mass that remains on vertex x.

- **Wasserstein Distance between two vertex probability measures:**

$$W_1(m_x^\alpha, m_y^\alpha) = \inf_{\xi} \sum_{x_i \in V} \sum_{y_j \in V} d(x_i, y_j)\xi(x_i, y_j). \tag{2}$$

- **Ollivier Ricci curvature along an edge:**

$$c(x,y) := 1 - \frac{W_1(m_x^\alpha, m_y^\alpha)}{d(x,y)}, \tag{3}$$

- **Forman Ricci curvature for a p-simplex σ:**

$$\mathcal{F}_p^\sharp(\sigma) = \sharp\{\beta^{(p+1)} > \sigma\} + \sharp\{\gamma^{(p-1)} < \sigma\} - \sharp\{\text{parallel neighbours of } \sigma\}, \tag{4}$$

where $\beta^{(p+1)} > \sigma$ denotes a $(p+1)$-simplex β that has σ as a face and $\gamma^{(p-1)} < \sigma$ denoting the $(p-1)$-simplex γ as a face of σ.

- **Bochner-Weitzenböck Decomposition:**

$$\mathbf{L}_p = \Delta_p + \mathrm{Ricc}_p^{\mathcal{F}}, \tag{5}$$

where Δ_p is the Bochner Laplacian, $\mathbf{L}_p = \mathbf{B}_p^T\mathbf{B}_p + \mathbf{B}_{p+1}\mathbf{B}_{p+1}^T$ $(p > 0)$ is the p-Hodge Laplacian. $\mathrm{Ricc}_p^{\mathcal{F}}$ is a matrix with diagonals each equal to Forman Ricci curvature value for each of the p-simplices.

Although both Ricci curvatures are formulated differently, both Ricci curvatures are mostly positive within clusters/communities while negative curvatures are present in "linking regions" connecting communities.

Persistent Ricci curvature based Filtration Process

Persistent Ricci curvature (PRC) is used to track the changes in curvature values of a simplicial complex along a filtration process. At each filtration parameter f, curvature distributions for 0-simplex, 1-simplex are obtained. Statistics of the curvature distributions such as persistent minimum, persistent maximum, persistent mean, persistent standard deviation, etc. are computed.

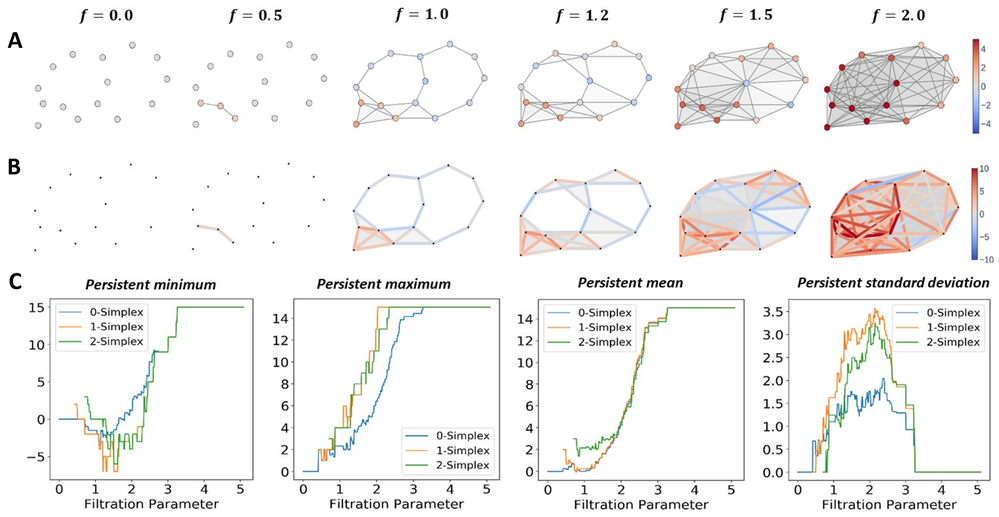

Figure: Illustration of 0, 1-simplex Forman Ricci curvatures (A–B) and four persistent attributes (C) on a nested sequence of simplicial complexes from a filtration process.

Acknowledgement

The computational work was partially done on resources of the National Supercomputing Computer, Singapore (https://www.nscc.sg). This work was supported in part by Nanyang Technological University Startup Grant M4081842.110, Singapore Ministry of Education Academic Research fund Tier 1 RG109/19, Tier 2 MOE-T2EP20120-0013 and MOE-T2EP20220-0010.

Persistent Ricci curvature based Machine Learning Models

Persistent Ricci curvature based machine learning models (PRC-ML) are proposed to predict the binding affinities of protein-ligand interactions using PDBbind databank. The PRC-ML models have outperformed over 20 existing traditional molecular based descriptor models when tested using PDBbind databanks.

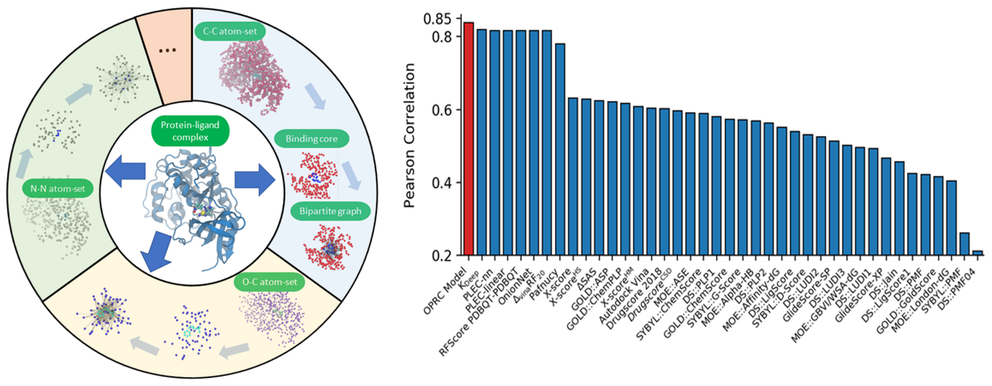

Figure: Left: Breakdown of protein-ligand complex into several graphs/simplicial complexes. Right: Comparison of PRC-ML models with state-of-the-art traditional molecular descriptor based models.

Modelling Protein-Protein Interactions

Similar to persistent Ricci curvature, persistent Hodge Laplacians are used to extract both topological and geometrical information from protein-protein interactions (PPIs) before and after mutations. Both zero and non-zero eigenvalues from persistent Hodge Laplacians reveal the intrinsic topological and geometrical information within the PPI structures. The statistics of persistent eigenspectrums serve as persistent spectral features for machine learning.

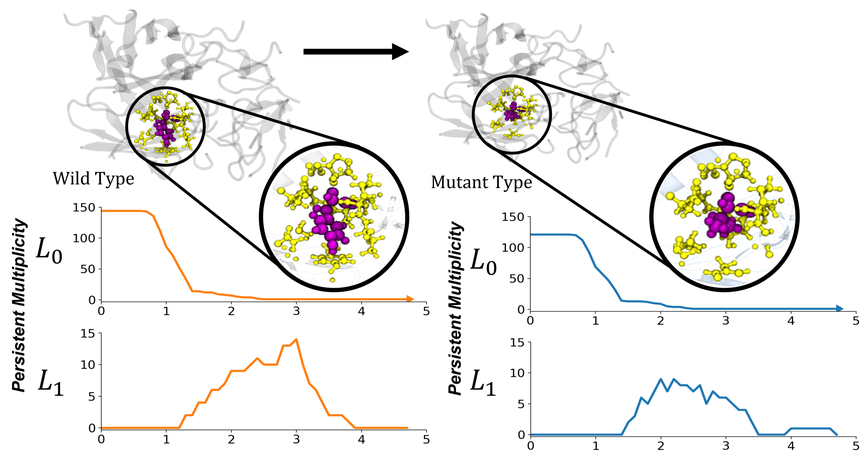

Figure: The persistent multiplicity of zero eigenvalues for L_0 and L_1 from PPI structure before and after mutation.

PerSpect-EL Models for Protein-Protein Interactions

A series of persistent spectral ensemble learning (PerSpect-EL) models are introduced to use the persistent spectral features to predict the change in binding affinity values upon mutation ($\Delta\Delta G$). Each statistical attribute of persistent spectral feature is trained with a base learner such as a convolutional neural network (CNN) or a gradient boosting tree (GBT). The trained outputs are concatenated and learned by a meta learner to produce an ensemble learning prediction. The PerSpect-EL models have surpassed existing state-of-the-art traditional molecular descriptor based models when tested with the SKEMPI-1131 dataset.

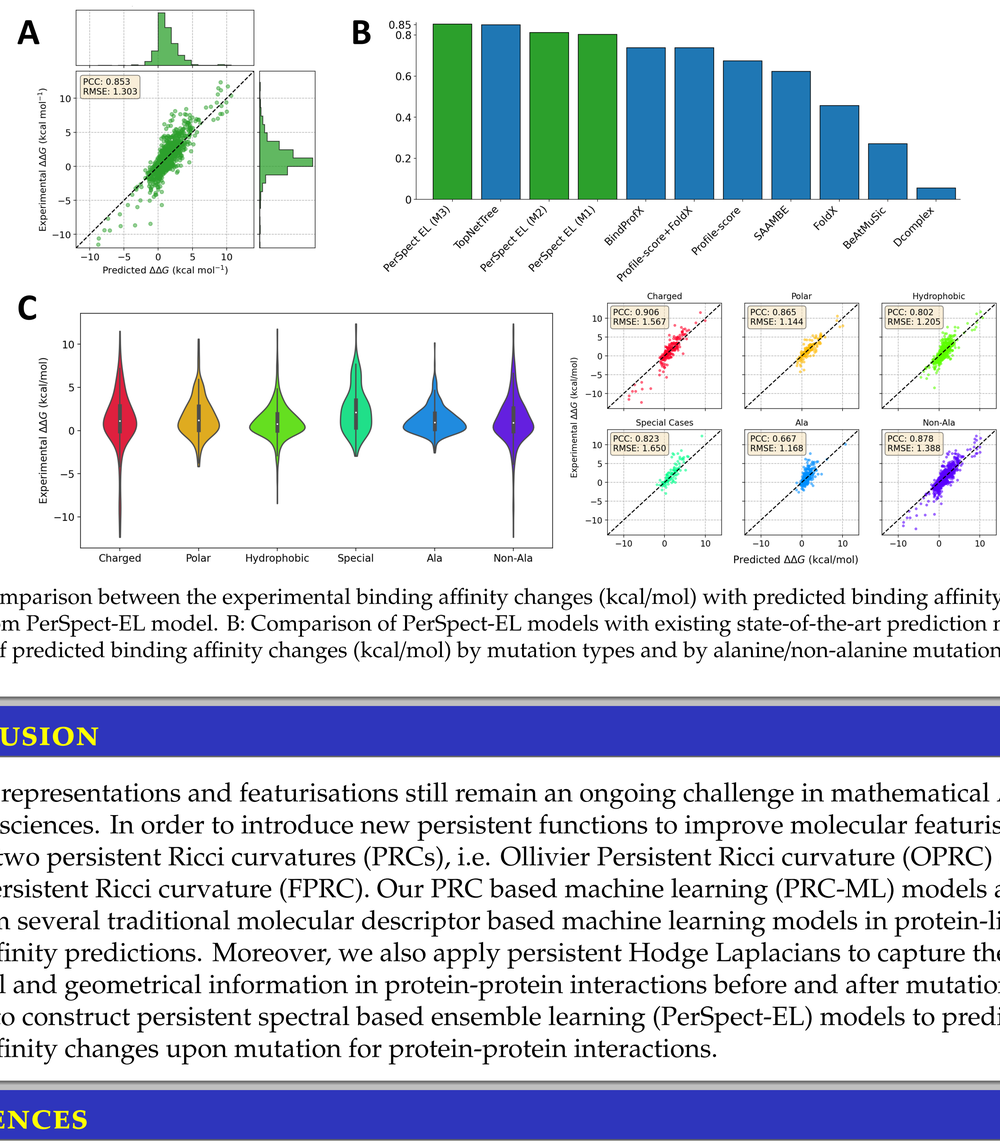

Figure: A: Comparison between the experimental binding affinity changes (kcal/mol) with predicted binding affinity changes (kcal/mol) from PerSpect-EL model. B: Comparison of PerSpect-EL models with existing state-of-the-art prediction models C: Breakdown of predicted binding affinity changes (kcal/mol) by mutation types and by alanine/non-alanine mutations.

Conclusion

Molecular representations and featurisations still remain an ongoing challenge in mathematical AI based molecular sciences. In order to introduce new persistent functions to improve molecular featurisation, we introduce two persistent Ricci curvatures (PRCs), i.e. Ollivier Persistent Ricci curvature (OPRC) and Forman Persistent Ricci curvature (FPRC). Our PRC based machine learning (PRC-ML) models are able to outperform several traditional molecular descriptor based machine learning models in protein-ligand binding affinity predictions. Moreover, we also apply persistent Hodge Laplacians to capture the topological and geometrical information in protein-protein interactions before and after mutations. This allows us to construct persistent spectral based ensemble learning (PerSpect-EL) models to predict the binding affinity changes upon mutation for protein-protein interactions.

References

- 1. Wee JunJie, Xia Kelin. Ollivier Persistent Ricci Curvature-Based Machine Learning for the Protein-Ligand Binding Affinity Prediction. Journal of Chemical Information and Modeling (2021). 61(4), 1617-1626.
- 2. Wee JunJie, Xia Kelin. Forman persistent Ricci curvature (FPRC) based machine learning for protein-ligand binding affinity prediction. Briefings in Bioinformatics (2021). Volume 22, Issue 6, November 2021, bbab136.
- 3. Wee JunJie, Xia Kelin. Persistent spectral based ensemble learning (PerSpect-EL) for protein-protein binding affinity prediction. Briefings in Bioinformatics. (2022). bbac024.

https://junjiewee.github.io https://personal.ntu.edu.sg/xiakelin/ weej0019@e.ntu.edu.sg xiakelin@ntu.edu.sg