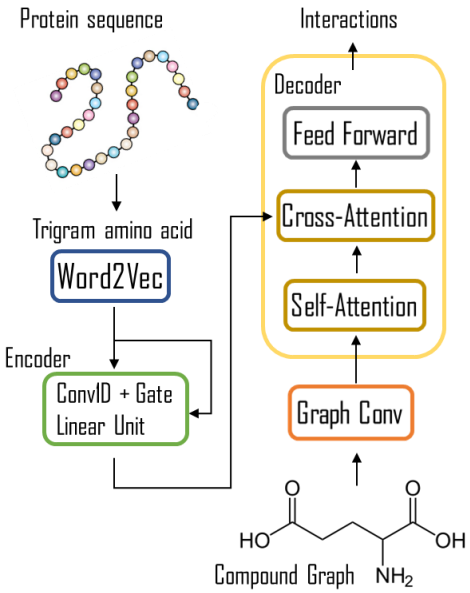
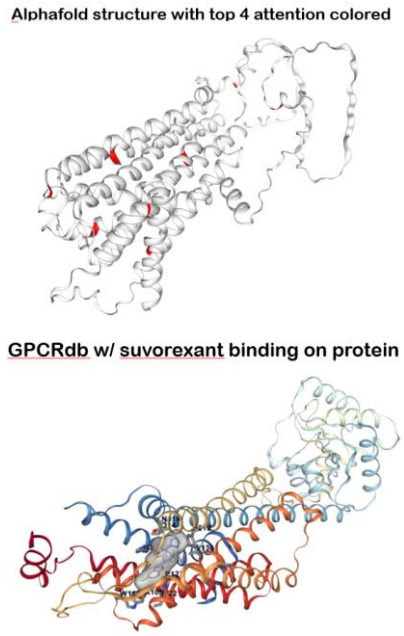
Interpreting Transformer Attention Weights in 2D information Compound–Protein Interaction Models

Zong Rong Ye,[1] Sheng-Hsuan Hung,[1] and erlin Chen[2*], and Ming-Kang Tsai[1*]

[1] *Department of Chemistry, National Taiwan Normal University, Taiwan;*
[2] *Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan*
Email:rocklee2578@gmail.com

We investigate attention weight interpretation in Transformer-based compound–protein interaction models to enhance predictive accuracy using 2D representations. Building on TransformerCPI[1], we trained three variants—bn_gcn, bn_gcn_cls, and TylorCE_cls (using a TaylorSoftmax-based loss[2]). We extracted tri-gram level attention weights and combined them with solvent accessible surface area (SASA) data, including a variant focusing on α-helical regions, to probe their role in binding site prediction. While high attention did not always correlate with high SASA, certain moderately exposed regions exhibited strong attention, suggesting a multifactorial influence on model focus. Classical ML models (Random Forest and Linear Regression) applied to these features revealed that Random Forests generally provided better predictive performance, emphasizing the non-linear relationships in the data. Our findings highlight the dual role of attention weights as both predictive features and interpretable signals in CPI modeling.





References

1. Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Bioinformatics, **2020**, 36, 4406–4414
2. Feng, L.; Shu, S.; Lin, Z.; Lv, F.; Li, L.; An, B. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, **2020**, 29, 2206–2212