Literature Review (CSCI2952G)

An Overview of Graph Neural Networks, Protein-Protein Interaction, and Manifold Graph Embedding

Charumathi Narayanan, Daniel Posmik, Minh Le Tran

October 10, 2025

1. Protein-Protein Interactions

Proteins are the essential workers in the body. They also rarely act alone, often interacting with many other proteins to perform their task properly. It is these web of interactions that lay the foundation for all the activities and functions ensuring survival of the body (Braun and Gingras [1]). These protein-protein interactions (PPIs) are thus a vital source of information of how the body functions at the cellular level. Moreover, since illness are often caused by mutations and disruption to PPIs, understanding more about PPIs can greatly inform drug development(Greenblatt et al. [2]). In research, PPI are often represented graphically, with nodes representing the protein, and edges representing interactions between the two connected nodes (Rao et al. [3]). These graphs are then analyzed computationally. As biological experiments uncover more proteins and their interactions, there are now many databases for PPIs, including STRING, KEGG, BioGRID, and BioRank(Rao et al. [3]). There are also many PPI datasets pertaining to a specific illness, such as OncoPPi, a cancer-focused dataset (Li et al. [4]). However, the biochemical relationships between proteins and the functional consequences of physical interactions remain extremely complicated, given the magnitude and variety of proteins. Due to this complex data, the field of PPIs is a popular area of application for computational analysis. In particular, many deep learning methods can integrate such data for a variety of tasks, including prediction of unknown protein interactions, or identification of targets for drug effectiveness (Soleymani et al. [5]). However, as typical of the deep learning field, there is no single best architecture for modeling these PPIs, and more work remains to be able to fully capture and utilize the dataset effectively.

2. Deep Learning Architectures for Protein-Protein Interactions

There are multiple Deep Learning architectural paradigms critical for robust protein feature representation,geometry modeling and evolutionary sequence comprehension. Literature focused on models like Discriminative Network Embedding (DNE) (Yan et al. [6]) portrays the protein landscape as a large, non-Euclidean network. The primary input is the network topology (who interacts with whom). The core capability is generating robust,

low-dimensional network embeddings calculated through contrastive loss that capture holistic and functional relationships, supporting tasks like PPI prediction and functional module identification. Architectures such as ProS-GNN (Wang et al. [7]) use GNNs that operate directly on the $3D$ molecular structure. Therefore, the $3D$ structure and spatial coordinates is used as input data. The main idea is the use of message-passing schemes to encode local $3D$ spatial information and non-covalent interactions, crucial for accurately predicting physical properties dependent on structural properties, such as changes in thermodynamic stability ($\Delta\Delta G$). The research further develops on these changes to predict insilico mutations. Furthermore, the pLLM paradigm (Xiao et al. [8]) adapts Transformer-based models from NLP (like BERT etc.). Databases like Uniref/UniParc consisting of billions of linear amino acid sequences are used. The model leverages this data to learn the evolutionary grammar of proteins, which generates highly contextual embeddings that encode latent information about function and structure as well as distal relations, serving as a powerful transfer learning backbone for numerous downstream tasks. As an example, state-of-the-art (SOTA) models like ProteinGPT (Xiao et al. [9]) represent the amalgamation of both sequence and structure. These systems process both sequence (via the pLLM backbone) and structural data (via GNNs or other geometric encoders). The key feature involves introducing adapter or projection layers to align the pLLMs latent space with that of the GNN's embeddings, enabling the resulting large language model to perform downstream tasks based on its comprehensive multi-modal understanding of the protein landscape.

3. Manifold Graph Embedding

Manifold estimation is motivated by the manifold hypothesis, which posits that observed data originates from a lower-dimensional data generating process corrupted by high-dimensional noise (Fefferman et al. [10], Meng and Eloyan [11]). This technique generalizes non-linear dimensionality reduction while explicitly preserving geometric properties of the data. For graph-structured data, manifold estimation leverages well-known connections between discrete graphs and non-Euclidean embedding space curvature, where characteristic graphs correspond to principle curvatures—tree-like graphs embed optimally in hyperbolic spaces while chain graphs are best represented in spherical space (Bronstein et al. [12], Weber and Nickel [13], Weber [14]). The learning problem requires estimating the manifold class $\mathcal{M}$, intrinsic dimension p, curvature κ, and projection map $\pi : \mathcal{D} \to \mathcal{M}_\kappa^p$ where $\mathcal{D} = G = \{V, E\}$, with Lubold et al. [15] and Meng and Eloyan [11] providing data-driven estimation techniques. Under correct specification, a node's projection $\pi_{\mathcal{M}}^p(X_{i,j})$ yields information about node importance (Yamada [16], Xu [17]), where highly connected nodes lie in flatter space and isolated nodes lie in highly curved space. The inverse gradient $\hat{w}_{ij} := |\nabla\ \pi_{\mathcal{M}}^p(X_{i,j})|^{-1}$ provides weights that capture geometric, topological, and relational context, making them particularly valuable for high-dimensional unsupervised settings where spectral methods may fail (Baptista et al. [18], Rubin-Delanchy [19]). For practical implementation, Giovanni et al. [20] propose "soft manifolds" using Riemannian stochastic gradient descent (Bonnabel [21]), with similar approaches offered by Marinoni et al. [22] and Jyothish and Jannesari [23]. Refer to Appendix 3 for a more detailed exposition.

Appendix A.

Manifold estimation is motivated by the manifold hypothesis [10] which intuitively states that our realized sample (i.e. our data) stems from a lower-dimensional data generating process that is subsequently corrupted by high-dimensional noise (Fefferman et al. [10], Meng and Eloyan [11]). We shall refer to the process of reconstructing and estimating this low-dimensional manifold as manifold estimation or manifold embedding (two terms we will use interchangeably). Intuitively, manifold estimation is a generalization of non-linear dimensionality reduction in which we may explicitly preserve geometric properties of our data.

For high-dimensional noisy graph-structured data, manifold estimation is particularly well-suited to well-known links between discrete graphs and the curvature of non-Euclidian embedding spaces [12]. While no choice of embedding space is perfect due to distortion trade-offs, it can be shown that certain "characteristic graphs" correspond to principle curvatures, e.g., tree-like graphs are optimally embedded in hyperbolic spaces while chain graphs are best represented in spherical space (Weber and Nickel [13], Weber [14]).

Manifold estimation is a non-trivial learning problem that requires the careful consideration of trade-offs (e.g., distortion and relational context) and parameter estimation. The parameters that need to be estimated are the class of manifold $\mathcal{M}$, choosing mixed effects structure (e.g., clique-level fixed effects), the intrinsic dimension of the manifold (p), and the projection map from the data ($\mathcal{D}$) to the manifold, i.e.

$$\pi : \mathcal{D} \to \mathcal{M}_\kappa^p \quad where \quad \mathcal{D} = G = \{V, E\}$$

Here, κ is a measure of curvature and our data $\mathcal{D}$ can be expressed as a graph $G = \{V, E\}$. Lubold et al. [15] present a data-driven, replicable alternative to ex-ante choosing manifold class ($\mathcal{M}$), dimension (p), and curvature (κ). The technique of Lubold et al. [15] demonstrate how the connection likelihood has an inverse relationship with the distance of the projected points on the latent manifold. Meng and Eloyan [11] present a regularized principled manifold estimation technique.

Under correct specification of $\pi(\cdot)$ and the manifold parameters, the key is that for any node $X_{ij} \in G = \{V_X, E_X\}$, its projection on the manifold $\pi_{\mathcal{M}}^p(X_{i,j})$ of class $\mathcal{M}$ and intrinsic dimension $\mathcal{M}$ gives us valuable information about node importance (e.g., Yamada [16] use Forman Curvature, Xu [17] offers a broader discussion). By the above embedding results, highly connected nodes will lie in "flatter" space while more isolated nodes (e.g., the terminal nodes of a tree-graph) will lie in highly curved space. Thus, the inverse of the absolute gradient at the projection $\pi_{\mathcal{M}}^p(X_{i,j})$, say $\hat{w}_{ij} := |\nabla\ \pi_{\mathcal{M}}^p(X_{i,j})|^{-1}$, will give us insight into how important this node is in the graph. Our goal is to use these inverse weights w_{ij} as an additional feature in protein-protein interaction prediction. These weights w_{ij} lend themselves particularly well to high-dimensional, unsupervised environments since they respect geometric (e.g., curvature, embedding space), topological, and relational contexts of the data. Alternative embedding techniques, e.g., well-established spectral matrix embedding methods, may not readily extend to heterogeneous geometry and high-dimensional settings (for relevant discussions, see Baptista et al. [18], Rubin-Delanchy [19]).

While we have described manifold estimation in theory, we have yet to cover practical considerations. Due to time constraints, we want flexible, out-of-the-box embedding algo-

rithms that can handle high-dimensional, heterogeneous graphs. Giovanni et al. [20] propose "soft manifolds" as a curvature-aware, flexible solution to the heterogeneity problem. In Giovanni et al. [20] Proposition 4.1, we are given a numerical embedding procedure based on Riemannian stochastic gradient descent (R-SGD) [21]. Marinoni et al. [22] offer a similar implementation. Jyothish and Jannesari [23] offer insight into how manifold embeddings work within a graph transformer framework and provide algorithmic details.

References

[1] Pascal Braun and Anne-Claude Gingras. History of protein–protein interactions: From egg-white to complex networks. *PROTEOMICS*, 12 (10):1478–1498, 2012. doi: https://doi.org/10.1002/pmic.201100563. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/pmic.20110056

[2] Jack F. Greenblatt, Bruce M. Alberts, and Nevan J. Krogan. Discovery and significance of protein-protein interactions in health and disease. *Cell*, 187(23):6501–6517, 2024. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2024.10.038. URL https://www.sciencedirect.com/science/article/pii/S0092867424012534.

[3] V. Srinivasa Rao, K. Srinivas, G. N. Sujini, and G. N. Sunand Kumar. Protein-protein interaction detection: Methods and analysis. *International Journal of Proteomics*, 2014(1):147648, 2014. doi: https://doi.org/10.1155/2014/147648. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/147648.

[4] Zenggang Li, Andrei A Ivanov, Rina Su, Valentina Gonzalez-Pecchi, Qi Qi, Songlin Liu, Philip Webber, Elizabeth McMillan, Lauren Rusnak, Cau Pham, Xiaoqian Chen, Xiulei Mo, Brian Revennaugh, Wei Zhou, Adam Marcus, Sahar Harati, Xiang Chen, Margaret A Johns, Michael A White, Carlos Moreno, Lee A D Cooper, Yuhong Du, Fadlo R Khuri, and Haian Fu. The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nat. Commun.*, 8 (1):14356, February 2017.

[5] Farzan Soleymani, Eric Paquet, Herna Viktor, Wojtek Michalowski, and Davide Spinello. Protein-protein interaction prediction with deep learning: A comprehensive review. *Comput. Struct. Biotechnol. J.*, 20:5316–5341, September 2022.

[6] Rui Yan, Md Tauhidul Islam, and Lei Xing. Deep representation learning of protein-protein interaction networks for enhanced pattern discovery. *Science Advances*, 10(51):eadq4324, 2024. doi: 10.1126/sciadv.adq4324. URL https://www.science.org/doi/abs/10.1126/sciadv.adq4324.

[7] Shuyu Wang, Hongzhou Tang, Peng Shan, Zhaoxia Wu, and Lei Zuo. Prosgnn: Predicting effects of mutations on protein stability using graph neural networks. *Computational Biology and Chemistry*, 107:107952, 2023. ISSN 1476-9271. doi: https://doi.org/10.1016/j.compbiolchem.2023.107952. URL https://www.sciencedirect.com/science/article/pii/S1476927123001433.

[8] Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, Xiao Luo, Yu Zhang, James Zou, Yizhou Sun, and Wei Wang. Protein large language models: A comprehensive survey, 2025. URL https://arxiv.org/abs/2502.17504.

[9] Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding, 2025. URL https://arxiv.org/abs/2408.11363.

[10] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis, 2013. URL https://arxiv.org/abs/1310.0425.

[11] Kun Meng and Ani Eloyan. Principal manifold estimation via model complexity selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):369–394, March 2021. ISSN 1467-9868. doi: 10.1111/rssb.12416. URL http://dx.doi.org/10.1111/rssb.12416.

[12] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017. ISSN 1558-0792. doi: 10.1109/msp.2017.2693418. URL http://dx.doi.org/10.1109/MSP.2017.2693418.

[13] Melanie Weber and Maximilian Nickel. Curvature and representation learning: Identifying embedding spaces for relational data. In *Advances in Neural Information Processing Systems*, volume 31. NeurIPS, 2018.

[14] Melanie Weber. Neighborhood growth determines geometric priors for relational representation learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 2020.

[15] Shane Lubold, Arun G. Chandrasekhar, and Tyler H. McCormick. Identifying the latent space geometry of network models through analysis of curvature, 2022. URL https://arxiv.org/abs/2012.10559.

[16] Taiki Yamada. Vertex evaluation of multiplex graphs using forman curvature, 2025. URL https://arxiv.org/abs/2504.17286.

[17] Mengjia Xu. Understanding graph embedding methods and their applications, 2020. URL https://arxiv.org/abs/2012.08019.

[18] Anthony Baptista, Rubén J. Sánchez-García, Anaïs Baudot, and Ginestra Bianconi. Zoo guide to network embedding, 2023. URL https://arxiv.org/abs/2305.03474.

[19] Patrick Rubin-Delanchy. Manifold structure in graph embeddings, 2021. URL https://arxiv.org/abs/2006.05168.

[20] Francesco Di Giovanni, Giulia Luise, and Michael Bronstein. Heterogeneous manifolds for curvature-aware graph embedding, 2022. URL https://arxiv.org/abs/2202.01185.

[21] Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[22] Andrea Marinoni, Pietro Lio', Alessandro Barp, Christian Jutten, and Mark Girolami. Improving embedding of graphs with missing data by soft manifolds, 2023. URL https://arxiv.org/abs/2311.17598.

[23] Ankit Jyothish and Ali Jannesari. Leveraging manifold embeddings for enhanced graph transformer representations and learning, 2025. URL https://arxiv.org/abs/2507.07335.