Hyperbolic representation learning in multi-layer tissue networks

Domonkos Pogány[1] and Péter Antal[1] *

1- Budapest University of Technology and Economics,
Department of Artificial Intelligence and Systems Engineering,
1111 Budapest, Hungary

Abstract. Predicting tissue-specific protein functions and protein-protein interactions (PPI) is essential for understanding human biology, diseases, and potential therapeutics. Recently, as a promising direction, more and more complex unsupervised feature learning approaches have emerged in the field, but none of them consider the scale-free nature and the underlying geometry of multi-layer PPI networks. Therefore, this study proposes contextualized, tissue-specific representation learning in non-Euclidean geometries and demonstrates that hyperbolic embeddings capture the structure of multi-layer PPI networks with less distortion and achieve better performance in tissue-specific protein function prediction.

1 Introduction

Proteins, as the primary building blocks of cells, drive most biological processes through their interactions, making them essential for understanding biology, disease mechanisms, and potential therapeutics. The functions of proteins are highly context-dependent, varying based on the tissue and cell type in which they are expressed [1]. Consequently, recent research has shifted focus from universal, human-level PPI networks to tissue-specific protein interactions and functions [1, 2, 3]. Computational methods, particularly representation learning, have emerged as potential tools to model and understand multi-layer protein interaction networks while maintaining tissue specificity and to predict new interactions and multicellular functions across various human tissues [2, 3, 4].

Another relevant research area involves investigating the geometric properties of real-world networks. It has been shown that power-law degree distributions and strong clustering in scale-free networks [5] emerge naturally from underlying hyperbolic metric spaces [6]. For instance, PPI networks are scale-free, with some better-connected genes due to a preferential attachment driven by evolutionary processes [7], therefore having an inherent hyperbolic geometry [8].

*Project no. 2024-2.1.1-EKÖP financed under the EKÖP-24-3-BME-60 funding scheme, also supported by the Doctoral Excellence Fellowship Programme (DCEP), is funded by the National Research Development and Innovation Fund of the Ministry of Culture and Innovation and the Budapest University of Technology and Economics, under a grant agreement with the National Research, Development and Innovation Office. This research was also funded by the J. Heim Student Scholarship, the OTKA-K139330, the European Union (EU) Joint Program on Neurodegenerative Disease (JPND) Grant: (SOLID JPND2021-650-233), the National Research, Development, and Innovation Fund of Hungary under Grant TKP2021-EGA-02, the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory.

Building on this insight from network science, researchers started to integrate non-Euclidean geometry into machine learning models, aiming to capture the nature of real-world networks through continuous, hierarchy-preserving representations [9, 10]. Recently, hyperbolic alternatives have been developed for numerous machine learning models, including shallow graph representation learning [9, 11], matrix-factorization [10], and graph convolutional neural networks [12, 13]. These geometry-aware methods have been successfully applied on several biology-related networks such as disease taxonomy trees [11], biological pathway graphs [13], metabolite disease association [10] and PPI networks [12].

However, no studies have investigated the geometry of either multi-layer or contextualized protein networks. Therefore, in this study, we propose a non-Euclidean version of a contextualized multi-layer graph embedding method and demonstrate through experiments on tissue-specific PPI networks that hyperbolic embeddings achieve lower graph distortion and improved protein function prediction. We hope our results motivate the research community to incorporate hyperbolic geometry into multi-layer models. To this end, we made our Pytorch-based implementation with the embeddings and a hyperbolic visualization dashboard available at https://github.com/PDomonkos/hyperbolic-ohmnet.

2 Materials and Methods

2.1 Data

Our study utilized a widely used benchmark multi-layer PPI dataset compiled by Zitnik et al. [2], one of the most extensive tissue-specific human protein networks available. The backbone of the dataset is a tissue hierarchy consisting of 219 human tissues with 107 leaves and 112 internal nodes. The dataset contains tissue-specific PPI networks for all the leaves and 37 other internal nodes, with each network comprising an average of $\approx$ 1,900 proteins and 25,500 interactions. A total of 503 (less than initially reported) tissue-specific protein functions are also included, covering 48 tissues. Each contains binary labels for all proteins in the tissue, indicating whether a gene corresponds to the biological function.

As a quantitative measure of the latent PPI geometry, we computed Gromov's δ-hyperbolicity [12]. The lower the δ, the more tree-like the graph is. As expected for the scale-free PPI networks [7], we observed an average $\delta = 1.18$, a relatively low value, indicating an underlying hyperbolic geometry and supporting the potential efficiency of using non-Euclidean embeddings.

2.2 Model

2.2.1 OhmNet

We investigated the OhmNet model, a hierarchy-aware unsupervised feature learning approach proposed for the multi-layer tissue network by Zitnik et al. [2]. OhmNet utilizes the tissue hierarchy tree and the PPI networks in the leaf nodes to efficiently learn separate d-dimensional representations for proteins in each tissue by optimizing the following two objectives:

1. Proteins within a tissue that share similar PPI network neighborhoods are assigned similar features. For this objective, separate Node2Vec [14] models are trained in each leaf tissue. Based on the dot product similarities between embedding pairs, Node2Vec maximizes the likelihood of preserving neighborhoods sampled via random walks from the network.

2. Multiple representations of the same protein in adjacent layers of the hierarchy share similar features. For simplicity, this is achieved through a regularization that minimizes the Euclidean distance between embeddings corresponding to the same protein in parent and child tissues.

Starting with random feature initialization, the training algorithm alternates between updating leaf and internal node embeddings until convergence. Embeddings in the leaves are updated based on both objectives: PPI neighborhood preservation and tissue hierarchy regularization, i.e., minimizing distances between corresponding protein embeddings in the leaf and its parent. The resulting objective function is the sum of the Node2Vec loss and the regularization term weighted by a parameter λ. Each iteration includes a single epoch of gradient descent on this non-convex optimization problem. Internal nodes are updated based on only the Euclidean distances in the second objective, facilitating an efficient closed-form solution, where in each iteration, protein representations in the internal tissues are updated with the average of the corresponding embeddings in the parent and children tissues. For more details on the baseline model and data used, we refer the readers to the paper by Zitnik et al. [2].

2.2.2 Non-Euclidean OhmNet

In this paper, we propose modifications incorporating hyperbolic geometry into the OhmNet approach. More precisely, we apply the hyperboloid/Lorentz model, popular in machine learning due to its simplicity and numerical stability [11, 12, 10]. The d-dimensional hyperboloid manifold is embedded in a (d+1)-dimensional Euclidean ambient space as $\mathcal{H}^{d,\beta} = \{\mathbf{x} = (x_0, ..., x_d) \in \mathbb{R}^{d+1} | \|\mathbf{x}\|_{\mathcal{L}}^2 = -\beta, x_0 > 0\}$, where $-1/\beta$ represents the constant negative curvature of the space, $\|\mathbf{x}\|_{\mathcal{L}}^2 = \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}}$ denotes the squared Lorentzian norm, and $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i$ is the Lorentzian inner product. Distances between vectors are measured using the manifold distance, $d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \text{arcosh}(-\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}})$ or the squared Lorentzian distance, $d_{\mathcal{L}}^2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\mathcal{L}}^2 = -2\beta - 2\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}$. For the latter, there exists a closed-form centroid $\mu = \sqrt{\beta}(\sum_{i=1}^{n} \mathbf{x}_i)/|\|\sum_{i=1}^{n} \mathbf{x}_i\|_{\mathcal{L}}|$, minimizing the $\sum_{i=1}^{n} d_{\mathcal{L}}^2(\mathbf{x}_i, \mu)$ expression [11].

We kept the OhmNet algorithm the same but replaced the Euclidean vectors with embeddings on $\mathcal{H}^{d,\beta}$. To do so, we first modified the Node2Vec model. Fortunately, we only had to find an alternative to the dot product between the embeddings, as the rest of the algorithm is independent of the geometry used. There is no dot product similarity in the Lorentz model, so we converted the squared Lorentzian distance to a similarity metric in the form of $e^{-d_{\mathcal{L}}^2}$. As for the hierarchy regularization, we used $d_{\mathcal{L}}^2$ instead of the Euclidean distance $d_{\mathcal{E}}$. After

initializing the embeddings with a normal distribution on the Lorentz manifold, we ran the alternating OhmNet algorithm. Since they lie on a hyperbolic manifold, training the embeddings in leaf tissues now requires Riemannian gradient descent [15]. However, due to the choice of the Lorentzian distance function in the regularization, updates for internal tissues remain efficient, as similarly to the Euclidean average, the closed-form Lorentzian centroid μ can be used as an update for the protein representations.

3 Experiments and results

3.1 Unsupervised representation learning

Embeddings were trained in an unsupervised way on the 107 leaf PPI networks and the tissue hierarchy, as outlined above. We built upon the Pytorch and geoopt [15] libraries to implement the models and perform the Riemannian optimization. While Zitnik et al. [2] used stochastic gradient descent with a manual learning rate schedule, we found that the currently popular optimizers with an adaptive learning rate perform better. We ended up using the Root Mean Square Propagation as the alternating nature of the OhmNet updates hinders the usage of optimizers with first-order momentum. We trained 128-dimensional embeddings in both Euclidean space and in $\mathcal{H}^{d=128,\beta=1}$. The training involved 100 alternating iterations with a batch size of 64, a learning rate of 0.025, and $\lambda = 0.2$. After a grid search on the hyperparameters, we found robust performance across tested configurations, so we left most of the parameters unchanged. We refer the readers to our public repository for more details on the implementation and the hyperparameters.

3.2 Graph distortion

One aspect that we have compared the embeddings on is the graph distance distortion, i.e., how well the manifold distances between protein representations capture the shortest-path distances in the PPI networks. To account for the different magnitudes of $d_{\mathcal{E}}$ and $d_{\mathcal{L}}$, we used the scaled graph distortion proposed by McNeela et al. [13]. For each network and manifold, we scaled embedding distances by a constant, minimizing the distortion between manifold and shortest-path distances. During training, we measured distortions on the largest connected components of all 144 PPI networks. Figure 1 presents the results. Although only leaf PPIs were used during training, distortion levels in both leaf and internal tissues were almost identical. Comparing the two geometries, as the low Gromov's δ and the scale-free nature suggested, hyperbolic embeddings reach much lower distortion.

3.3 Node classification

Another aspect used to evaluate the embeddings is node classification, more precisely predicting protein functions based on the learned representations. 503

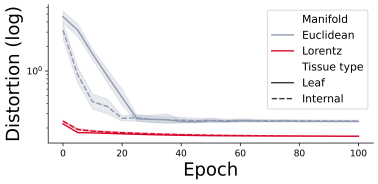

Fig. 1: Mean and 95% confidence intervals are shown for the graph distortions for $d_{\mathcal{E}}$ and $d_{\mathcal{L}}$ during training. Distortions were evaluated on 144 PPI networks corresponding to both leaf and internal tissues.

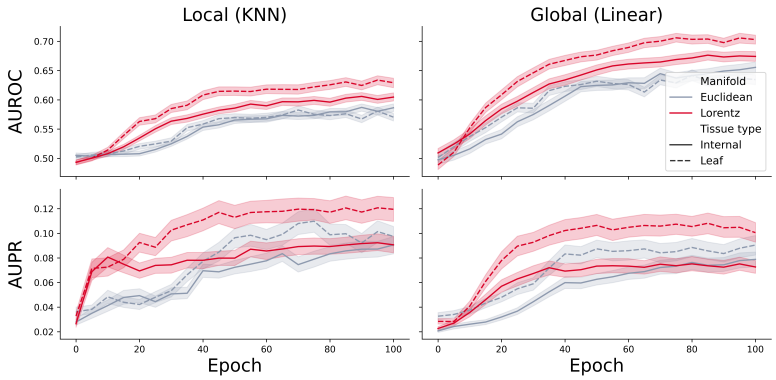

Fig. 2: Cross-validation results for protein function prediction, showing mean AUROC and AUPR metrics with 95% confidence intervals. Metrics are presented for classifiers leveraging global and local information on Euclidean and hyperbolic embeddings across internal and leaf tissues.

tissue-specific protein functions were used across 48 tissues in both internal and leaf nodes, forming 503 separate, highly imbalanced binary classification tasks with a positive-to-negative label ratio of 1:60. Following Zitnik et al. [2], we trained linear classifiers for each task using the modified Huber loss. For the hyperbolic version, we replaced the linear layer with the hyperboloid linear transform[12], which applies matrix multiplication and bias addition in the Lorentz model. Linear classifiers operate with one separating hyperplane in the embedding space, therefore assessing how well the global structure of the embeddings encodes protein functions. To get a more comprehensive view, we also applied K-Nearest Neighbors (KNN) classifiers, capturing the local structure based on $d_{\mathcal{E}}$ and $d_{\mathcal{L}}$. We performed cross-validation (CV) on each task. Instead of a 10-fold CV proposed by Zitnik et al. [2], we used a stratified 5-fold CV to get more realistic results and to ensure that every test split has at least one positive sample. Figure 2 shows the resulting area under the receiver operating characteristic and precision-recall curve metrics (AUROC and AUPR). As we can see, the hyperbolic model outperforms the Euclidean one in predicting leaf and internal tissue-specific functions based on both local and global information encoded in the embeddings.

4 Conclusion

This study investigated contextualized protein embeddings in non-Euclidean geometries, proposing a Lorentzian version of the multi-layer OhmNet model incorporating hyperbolic Node2Vec and hierarchy regularization. Our findings indicate that hyperbolic embeddings outperform the Euclidean approach in terms of both PPI graph distortion and tissue-specific protein function prediction.

Based on our results, future research should investigate the hyperbolic nature of multi-layer networks and incorporate non-Euclidean geometry into more advanced, inductive methods, such as tensor factorization [4] and graph attention networks [3]. Another promising direction is to integrate additional prior knowledge on tissues and cell types into the tissue-specific protein network [3].

References

[1] Esti Yeger-Lotem and Roded Sharan. Human protein interaction networks across tissues and diseases. *Frontiers in genetics*, 6:257, 2015.

[2] Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, 2017.

[3] Michelle M Li, Yepeng Huang, Marissa Sumathipala, Man Qing Liang, Alberto Valdeolivas, Ashwin N Ananthakrishnan, Katherine Liao, Daniel Marbach, and Marinka Zitnik. Contextual ai models for single-cell protein biology. *Nature Methods*, 21(8):1546–1557, 2024.

[4] Sameh K Mohamed. Predicting tissue-specific protein functions using multi-part tensor decomposition. *Information Sciences*, 508:343–357, 2020.

[5] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[6] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

[7] Eli Eisenberg and Erez Y Levanon. Preferential attachment in the protein network evolution. *Physical review letters*, 91(13):138701, 2003.

[8] Gregorio Alanis-Lobato, Pablo Mier, and Miguel Andrade-Navarro. The latent geometry of the human protein interaction network. *Bioinformatics*, 34(16):2826–2834, 2018.

[9] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[10] Domonkos Pogány and Péter Antal. Hyperbolic metabolite-disease association prediction. In *ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 333–338, 2024.

[11] Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2019.

[12] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

[13] Daniel McNeela, Frederic Sala, and Anthony Gitter. Product manifold representations for learning on biological pathways. *arXiv preprint arXiv:2401.15478*, 2024.

[14] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[15] Max Kochurov, Rasul Karimov, and Serge Kozlukov. Geoopt: Riemannian optimization in pytorch. *arXiv preprint arXiv:2005.02819*, 2020.