ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Hyperbolic Metabolite-Disease Association Prediction

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. In biomarker research, there is a growing demand for computational methods to efficiently identify novel metabolite-disease associations (MDAs). Current approaches, however, do not take into account the underlying geometry of the MDA space. Here, we show that classifiers leveraging hyperbolic embeddings achieve comparable results to their Euclidean counterparts with significantly lower dimensionality, aligning better with the association network's scale-free nature. Finally, through a case study, we provide an interpretation of the model embeddings and investigate newly predicted associations. Our results demonstrate the intrinsic non-Euclidean geometry of the MDA space, providing direction for further research. A Pytorch-based implementation is available at https://github.com/PDomonkos/hyperbolic-MDA-prediction.

1 Introduction

Concentrations of certain metabolites in patients with specific diseases differ from those in healthy individuals. Identifying these specific metabolites can significantly contribute to disease diagnosis. Nonetheless, conventional biological experiments are often resource-intensive in terms of time and cost. Thus, there is a growing demand for computational methods that efficiently identify new relationships between metabolites and diseases. One possible solution is to utilize machine learning models such as matrix factorization or graph neural networks [1, 2]. However, current approaches do not consider the underlying geometry of the MDA space.

Recently, research on hyperbolic embeddings has been gaining attention for their ability to effectively represent datasets with intrinsic hierarchies and complex networks with heterogeneous degree distributions [3]. Incorporating non-Euclidean geometry into machine learning approaches enables learning continuous, hierarchy-preserving representations automatically [4]. Among various biological applications, hyperbolic models have already been applied to predict

*The project 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.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

drug-target interactions [5], yet, to the best of our knowledge, no prior study has explored non-Euclidean embeddings for MDA prediction.

To address this issue, we utilized matrix-factorization-based machine learning models and investigated the applicability of non-Euclidean embedding spaces. Through our experiments, we showed that models utilizing hyperbolic distances are better suited for predicting MDAs. Later, we also gave an explanation relying on the scale-free nature of the association network rather than underlying biological hierarchies.

The paper proceeds as follows: First, we briefly introduce the used data, models, and experimental settings. Then, we present the results of our comparative study. Finally, we investigate the hyperbolic latent representations and validate the model by showcasing that newly predicted associations have already been verified in recent literature.

2 Materials and Methods

2.1 Metabolite-disease associations

To obtain MDAs, we utilized the Human Metabolomics Database (HMDB 5.0), currently the most comprehensive dataset on human metabolites [6]. After extracting all associations, an outlier disease linked to over 20,000 metabolites was excluded. Finally, we represented the MDA network with a binary adjacency matrix comprising 2,583 metabolites in rows, 656 diseases in columns, and 7,650 positive associations between them, resulting in a density of 0.45%.

2.2 Models and manifolds

Our objective was to compare different models equipped with Euclidean and non-Euclidean embeddings. Following the work of A. Poleksic [5], we applied shallow representation learning modules to learn a d-dimensional latent embedding for each metabolite and disease. The trainable parameters comprise these latent vectors stored in the weight matrices, initialized with standard normal distributions in the corresponding manifolds. Predictions for the binary association output are made based on similarities between metabolite and disease representations, adopting a generalized matrix factorization approach. Keeping the emphasis on the manifold distances, we chose a simple activation function, $1/(1+D)$, to convert distances into predicted similarity scores.

Baseline models operate with representations on the Euclidean manifold $\mathcal{E}^d = \mathbb{R}^d$, utilizing the Euclidean distance, $D_{\mathcal{E}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$.
For the non-Euclidean version, we employed the Lorentz (or hyperboloid) model, favored in machine learning applications for its numerical stability [4, 5]. The d-dimensional representations reside on the hyperboloid manifold $\mathcal{H}^{d,\beta}$ embedded in a (d+1)-dimensional Euclidean ambient space. $\mathcal{H}^{d,\beta}$ is defined as follows:

$$\mathcal{H}^{d,\beta} = \{\mathbf{x} = (x_0, ..., x_d) \in \mathbb{R}^{d+1} | \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -\beta, x_0 > 0\},$$

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

where $-1/\beta$ is the constant negative curvature of the space, 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, $D_{\mathcal{H}}(\mathbf{x}, \mathbf{y}) = \text{arcosh}(-\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}})$.

2.3 Implementation details

We utilized the Pytorch and geoopt [7] libraries to implement the models and trained them on a 32GB NVIDIA Tesla V100 GPU utilizing Adaptive Moment Estimation (Adam) and Riemannian Adam [8] optimizers for the Euclidean and hyperbolic versions, respectively.

We found that the obtained results are robust across the various model configurations and applied the following hyperparameters: For the hyperbolic models, we chose a constant curvature of -1, i.e., $\beta = 1$, and used one extra latent dimension compared to their Euclidean counterparts since $\mathcal{H}^{d,\beta}$ is one dimension smaller than the Euclidean ambient space used for the representations. Models were trained for 32 epochs with a batch size of 256 and a learning rate of $\frac{0.001}{log_2(d)}$, where d is the latent dimension, as we found that the optimal learning rate depends on the model complexity. To obtain the negative samples, for each known positive pair of the MDA network, we took five times as many negative associations from all the possible pairs, increasing the density of the MDA matrix to 2.71%. Addressing class imbalance and the uncertainty in the negative samples, we applied weighted binary cross-entropy as the objective function with a 5 to 1 positive sample weight ratio.

3 Experiments and results

3.1 Cross-validation

To compare different models, we conducted cross-validation using five folds and ten repeats, with new negative samples in each repeat. We utilized the area under the receiver operating characteristic and precision-recall curves (ROCAUC and PRAUC) for evaluating the binary classification task, as they rely solely on prediction scores rather than a predefined threshold and are widely used to measure predictive performance in unbalanced scenarios.

Figure 1 presents the results. Classifiers with a hyperboloid manifold achieve comparable AUC scores to Euclidean models with ≈ 16 times larger latent dimensionality. Both reach similar predictive performance above 2,048 dimensions, yet hyperbolic versions notably outperform in lower dimensions. We can also see that Euclidean models show higher variance across the 50 runs. Besides, non-Euclidean versions require fewer epochs overall. Nevertheless, this does not necessarily translate to faster convergence in time, as the exponential and logarithmic maps slow down the Riemannian optimization.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

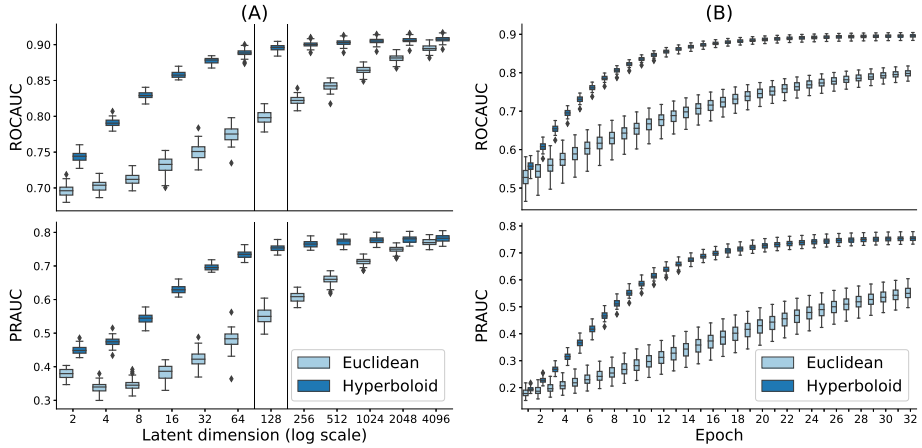

Fig. 1: Cross-validation results, with box plots summarizing the outcomes across the five folds and ten repeats. (A) ROCAUC and PRAUC scores reached in the final epoch for models with various latent manifolds and dimensions. (B) Lateral view of a slice from A, corresponding to models with 128 latent dimensions, showcasing training curves for a Euclidean and a hyperbolic model.

3.2 Case study

After fitting a classifier equipped with a 256-dimensional hyperboloid latent manifold utilizing all the available MDAs, we investigated the resulting latent embeddings and predicted associations.

We found no indication that metabolite or disease hierarchies are reflected in the embeddings. Therefore, we looked for an alternative explanation for the superior performance of hyperbolic models. Figure 2 illustrates the connection between the degree of the nodes and their latent hyperbolic embeddings. Straight lines in the log-log scale scatterplot suggest that the degree distribution follows a power law, meaning that the MDA network is scale-free [9], further indicating that a hyperbolic space might be more suitable than flat Euclidean embeddings [3]. To test this hypothesis, we investigated the embedding norms and observed a strong negative correlation between norms and degrees. Indeed, as hyperbolic space grows exponentially, it is suitable for embedding a few large degree nodes and many small degree nodes, even with smaller dimensionality.

Finally, we investigated the predictions for Alzheimer's disease to validate the classifier. Figure 3 depicts the similarity scores for all the metabolites. The predictions follow a positively skewed normal distribution, with some metabolites scoring notably high. Plotting the known associations against the prediction percentiles confirms that associated metabolites and diseases are close to each other in the resulting latent space. Table 1 shows the top 10 scoring metabolites for Alzheimer's disease. The majority are already in the HMDB database, while the others can be verified with external references.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

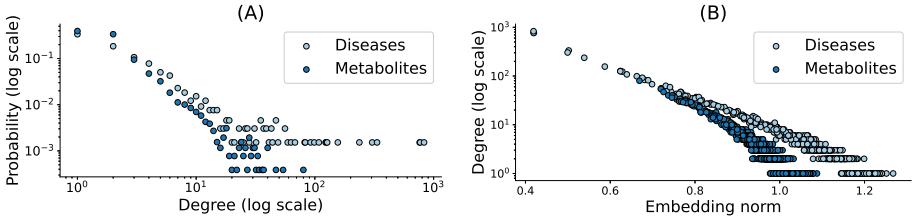

Fig. 2: Relationship between degree distribution and hyperbolic embedding norms. (A) Degree distribution of different modalities in the bipartite association graph. (B) Node degrees plotted against their corresponding embedding norms given by a model employing a 256-dimensional hyperboloid manifold.

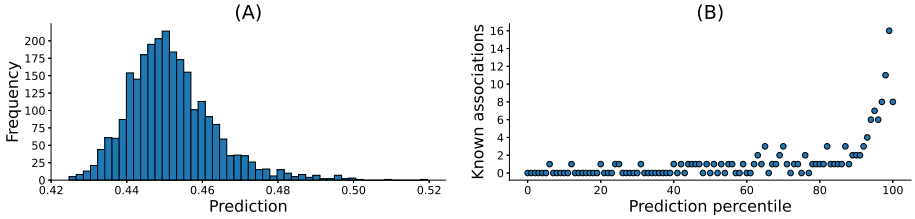

Fig. 3: Model outputs for Alzheimer's disease based on distances in the 256-dimensional hyperboloid manifold. (A) Histogram of predictions given to each metabolite. (B) Observed association frequencies for each prediction percentile.

ID	name	score	in HMD	external ref
HMDB0000190	L-Lactic acid	0.5196	✓	
HMDB0000182	L-Lysine	0.5090	✓	
HMDB0000562	Creatinine	0.5011	✓	
HMDB0000925	Trimethylamine N-oxide	0.5005	×	[10]
HMDB0000122	D-Glucose	0.5001	✓	
HMDB0000062	L-Carnitine	0.4993	×	[11]
HMDB0000696	L-Methionine	0.4980	×	[12]
HMDB0000159	L-Phenylalanine	0.4970	✓	
HMDB0000161	L-Alanine	0.4966	✓	
HMDB0000067	Cholesterol	0.4966	×	[13]

Table 1: Top 10 predicted metabolites for Alzheimer's disease, ranked by prediction score. Out-of-dataset associations are verified through external references.

4 Conclusion and future work

Through cross-validation and subsequent analysis, this paper demonstrates the suitability of hyperbolic embeddings for MDA prediction. Contrary to previous assumptions, we attribute the non-Euclidean nature of the MDA space to its

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

heterogeneous degree distribution rather than metabolite and disease hierarchies.

Consequently, further research should aim at integrating hyperbolic representations into existing, more complex model architectures, such as those based on graph neural networks [1] or matrix factorization [2], to enhance biomarker identification. Another theoretical question arises regarding the scale-free nature of the MDA network. It should be further investigated whether this phenomenon stems from biological mechanisms or aligns with the Barabási model's concept of continuous growth and preferential attachment [9] driven by the discovery process as biomedical scientists often select well-studied molecules and diseases to anchor their findings [14].

References

[1] Feiyue Sun, Jianqiang Sun, and Qi Zhao. A deep learning method for predicting metabolite–disease associations via graph neural network. *Briefings in bioinformatics*, 23(4):bbac266, 2022.

[2] Hongyan Gao, Jianqiang Sun, Yukun Wang, Yuer Lu, Liyu Liu, Qi Zhao, and Jianwei Shuai. Predicting metabolite–disease associations based on auto-encoder and non-negative matrix factorization. *Briefings in bioinformatics*, 24(5):bbad259, 2023.

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

[4] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018.

[5] Aleksandar Poleksic. Hyperbolic matrix factorization improves prediction of drug-target associations. *Scientific Reports*, 13(1):959, 2023.

[6] David S Wishart, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L Lee, et al. Hmdb 5.0: the human metabolome database for 2022. *Nucleic acids research*, 50(D1):D622–D631, 2022.

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

[8] Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018.

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

[10] Nicholas M Vogt, Kymberleigh A Romano, Burcu F Darst, Corinne D Engelman, Sterling C Johnson, Cynthia M Carlsson, Sanjay Asthana, Kaj Blennow, Henrik Zetterberg, Barbara B Bendlin, et al. The gut microbiota-derived metabolite trimethylamine n-oxide is elevated in alzheimerâs disease. *Alzheimer's research & therapy*, 10:1–8, 2018.

[11] Alina Kepka, Agnieszka Ochocinska, Małgorzata Borzym-Kluczyk, Ewa Skorupa, Beata Stasiewicz-Jarocka, Sylwia Chojnowska, and Napoleon Waszkiewicz. Preventive role of l-carnitine and balanced diet in alzheimerâs disease. *Nutrients*, 12(7):1987, 2020.

[12] Amal Alachkar, Sudhanshu Agrawal, Melica Baboldashtian, Khawla Nuseir, Jon Salazar, and Anshu Agrawal. L-methionine enhances neuroinflammation and impairs neurogenesis: implication for alzheimer's disease. *Journal of Neuroimmunology*, 366:577843, 2022.

[13] Femke M Feringa and Rik Van der Kant. Cholesterol and alzheimerâs disease; from risk genes to pathological effects. *Frontiers in Aging Neuroscience*, 13:690372, 2021.

[14] Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569–14574, 2015.