Hyperbolic Manifold Learning on Differential Expression Signatures

Domonkos Pogány ¹ and Péter Antal ²

¹Department of Measurement and Information Systems
²Affiliation not available

December 7, 2023

Abstract

Our paper contributes to understanding differential expression gene (DEG) signatures, a crucial element in the study of gene expression patterns. One of the key findings of our research, which has not been previously published, is the assertion that the DEG signature space significantly manifests hyperbolic properties. This discovery has far-reaching implications for both the fields of bioinformatics and machine learning, such as drug-target interaction prediction and visualizations for drug discovery.
Abstract—In computational biology, understanding gene expression patterns is pivotal. Our study explores several unsupervised dimensionality reduction approaches for supporting the analysis of transcriptomic signatures, shedding light on hidden clusters and relations. We investigate the application of both Euclidean and hyperbolic manifold learning to differential expression signatures, a domain where hyperbolic methods have yet to be widely explored. In our methodology, comprehensive evaluations were applied, utilizing diverse metrics to assess both local and global structure preservation. The findings of our comparative analysis reveal the applicability of hyperbolic embeddings to differentially expressed gene signatures. Beyond assisting researchers in choosing suitable dimensionality reduction techniques for signature visualization, our results provide evidence for the non-Euclidean nature of the differential expression space, providing novel prospects for leveraging hyperbolic methods in downstream data analysis tasks.

Index Terms—Differentially expressed gene signatures, LINCS, manifold learning, dimensionality reduction, hyperbolic geometry.

I. INTRODUCTION

DIFFERENTIALLY expressed gene (DEG) signatures are integral components in computational biology, offering insights into the complex realm of transcriptomics. Two distinct approaches are frequently employed in utilizing the underlying correlations within these signatures. The first involves manual feature selection, leading to the curation of landmark genes [1]. The second approach leverages unsupervised machine learning techniques, such as dimensionality reduction [2] and clustering [3], to extract valuable information from high-dimensional gene expression data. Manifold learning has emerged as a powerful method to transform high-dimensional data into lower-dimensional embeddings, providing insights into the underlying data structure. Another recently popular field of study is learning in non-Euclidean manifolds. Hyperbolic spaces have emerged as powerful tools for capturing hierarchical and tree-like structures inherent in various datasets. Applying hyperbolic representations to gene expression data introduces a novel paradigm, potentially unveiling latent hierarchical structures that traditional Euclidean representations might overlook [4]–[7].

Notably, existing research on hyperbolic representation learning has not explored its application to carefully curated landmark genes. Furthermore, previous hyperbolic manifold learning studies have primarily focused on basic gene expression data, even though DEG signatures hold substantial relevance across various scientific tasks. To address this gap, we conduct a comparative analysis, applying hyperbolic dimensionality reduction to different DEG datasets, including the entire gene set and the meticulously selected landmark genes. This comprehensive approach enables us to thoroughly explore manifold learning techniques and assess the suitability of Euclidean and hyperbolic methods for embedding DEG signatures. Our analysis encompasses an array of evaluation measures to consider both local and global structure preservation.

Our findings underscore the significant advance of hyperbolic embeddings for DEG data, providing compelling evidence in favor of their utilization. Beyond aiding in selecting dimensionality reduction techniques for signature visualization, our results affirm the hyperbolic nature of the differential expression space, opening new directions for future drug discovery. Among potential future applications lie the non-Euclidean versions of data visualization, clustering, and predictive models such as matrix factorization or neural networks. These models operate on low-dimensional internal representations of input signatures, potentially benefiting from a hyperbolic representation space as an alternative to the prevailing Euclidean approaches.

Figure 1 serves as a graphical abstract summarizing the key aspects of our research. The rest of this paper is organized as follows: we begin with a brief overview of relevant existing approaches in Section II. In Section III, we detail the datasets, manifold learning methods, and evaluation metrics employed in our research. Section IV presents the outcomes of our comparative study, detailing the hyperparameter optimization process, quantitative evaluations, and qualitative assessments. Finally, in Section V, after a summary, we outline potential avenues for future development.

II. BACKGROUND

Before diving into the methodology, we summarize the most important literature, deciphering all three parts of the title.

A. Differential Expression Signatures

Differential Expression Signatures play a pivotal role in computational biology, particularly in studying the intricate landscape of gene expression patterns. Gene expression encompasses the process of transcribing genetic information into functional molecules, such as RNAs, within a cell. However, differential gene expression distinguishes itself by spotlighting the variations in gene expression levels between distinct experimental conditions or biological states. With their help, we can
analyze differentially expressed genes that exhibit significant alterations in expression, shedding light on their potential roles in various biological processes and diseases.

The Library of Integrated Network-Based Cellular Signatures (LINCS) consortium stands as a pioneering effort in the realm of systems biology. LINCS aims to provide a universal language for biology-related tasks by amassing a comprehensive repository of gene expression signatures relying on the Connectivity Map (CMap). CMap exploits the transcriptome and utilizes gene expression profiling to connect biology, chemistry, and clinical conditions for drug discovery, i.e., to discover disease–gene-drug connections regardless of the microarray platforms used [8].

One of the seminal contributions from LINCS lies in the meticulous selection of the 978 landmark genes, which capture ~80% of the information in the whole transcriptome due to its correlated nature. These landmark genes are pivotal in unraveling complex biological relationships. Later analyses showed that the non-measured transcripts can be accurately inferred from the ~1000 landmark genes with deep learning-based approaches [10]–[12]. Moreover, the utilization of landmark genes instead of randomly selected non-landmark genes yields superior results in both unsupervised clustering [3] and supervised classification [13].

The consortium also developed a cost-effective high-throughput microarray platform, the L1000, to measure only the landmark genes [1]. The L1000 Luminex bead technology made it possible to simultaneously scale up the available data and minimize information loss due to the non-measured transcripts. In total, there are currently over three million available L1000 microarray profiles, which can be converted to RNA-seq-like gene expression profiles as well [14]. The availability of large data sets has led to the emergence of machine learning benchmarks and baseline models that can help exploit the potential inherent in gene expression signatures [15].

A potential application, particularly in the context of drug repositioning, involves the prediction of various attributes of biological entities, such as phenotypes [16], drug-induced side-effects, and mechanisms of action (MOA) [17]–[20]. Additionally, leveraging the LINCS landmark genes or their any low-dimensional embeddings, we can use machine-learning approaches to predict interactions between entities, such as drug-drug interactions [21] and drug-target interactions [22]–[24].

LINCS signatures traverse five distinct preprocessing levels: (1) raw, (2) deconvoluted, (3) normalized, (4) differentially expressed, and (5) consensus signatures. Our research was focused on DEG signatures above level 4. Originally, the MODZ method was used to calculate them, but later, the consortium incorporated a new geometrical multivariate approach called Characteristic Discretion (CD) [25]. CD significantly improves the signal-to-noise ratio compared with the previous z-score-based method [26] and better extracts and ranks the most relevant DEGs [27].

In our paper, due to the advantages and relevance mentioned above, we decided to work with LINCS level 4 and 5 CD DEG signatures.

### B. Manifold Learning

Dimensionality reduction techniques project data from high to low dimensions, empowering us to distill the essence of high-dimensional data, ultimately paving the way for insightful analyses [28]. While linear Principal Component Analysis (PCA) serves as an excellent starting point, the intricate nature of gene expression data often demands nonlinear dimensionality reduction approaches, also known as manifold learning. These methods offer a more nuanced perspective by capturing complex relationships that linear methods may overlook. Techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) [29], Uniform Manifold Approximation and Projection (UMAP) [30], and Pairwise Controlled Manifold Approximation Projection (PaCMAP) [31] are at the forefront of nonlinear dimensionality reduction. Previously, the LINCS consortium used another method to visualize the DEG landscape, namely the fireworks display (FWD), which applies the Allegro edge-repulsive strong clustering algorithm on k-nearest neighbor (KNN) graphs. Researchers created interactive two-dimensional FWD maps for signatures measured...
with the L1000 technique, such as the Breast Cancer Network Browser [32] or the L1000FWD [33].

The evaluation of dimensionality reduction techniques demands investigating both local and global structure preservation aspects (for earlier reviews, see [2], [28], [34]). Local metrics, such as neighborhood preservation, focus on assessing the preservation of pairwise relationships between data points after dimensionality reduction. Global metrics, on the other hand, gauge the overall structure preservation, ensuring that the broader patterns and clusters remain intact. For instance, it has been shown that t-SNE emphasizes preserving pairwise similarities, while PCA keeps the global structure intact, UMAP performs well in both aspects, and PaCMAP may be the best in preserving both local and global structures [2], [31].

Within the LINCS consortium, dimensionality reduction and visualization techniques are instrumental in navigating the vast landscape of gene expression signatures. Selecting only the ~1,000 landmark genes is also a form of dimensionality reduction, a.k.a. feature selection, but only aims to ease the measurement process. Further reduction is needed for visualization and other analyses. Researchers harness these techniques to gain insights into the intricate biology encoded within LINCS datasets. By employing dimensionality reduction, they create maps that reveal hidden relationships and patterns among signatures, facilitating the discovery of novel associations.

Several works aimed to visualize the CD DEG landscape as well [27], [32], [33], [35]. However, the applicability of different dimensionality reduction techniques on DEG signatures has not yet been investigated thoroughly. Thus, we aimed to fill this gap with our research. Although there are existing works evaluating manifold learning methods for transcriptomic data visualization [2], analyzing DEG signatures may lead to different observations.

C. Hyperbolic Representations

In recent years, representation learning has expanded its horizons to embrace non-Euclidean spaces, unraveling new opportunities to model complex data structures. Traditionally, machine learning has been predominantly grounded in flat Euclidean spaces. Nonetheless, the significance of tree representations in modeling biological sequence similarities, particularly ultrametric trees, was quickly recognized in phylogenetics [36]. Beyond genetics, numerous real-world datasets exhibit inherent hierarchies and non-linear structures that are not effectively captured within Euclidean space. Hyperbolic machine learning seeks to address this disparity by accommodating geometric spaces characterized by a constant negative curvature. In such spaces, the surface area increases exponentially in relation to the radius, offering the capability to capture intricate data patterns, represent hierarchies, and embed trees within a low-dimensional continuous space without distortion.

Mathematical models of non-Euclidean spaces were originally formulated in the context of hyperbolic geometry and later found their way into machine learning. A fundamental representation of hyperbolic space is the Poincaré ball model or its two-dimensional version, the Poincaré disk, often used for visualization purposes. The model represents the hyperbolic space as an open d-dimensional unit ball, denoted as $B^d = \{ x \in \mathbb{R}^d | 1 > \|x\| \}$, equipped with a Riemannian metric. In this model, the Poincaré distance $d_P$ between two embedding vectors $x$ and $y$ in the unit ball is defined as:

$$d_P(x, y) = \text{arcosh} \left( 1 + 2 \frac{\|x - y\|^2}{(1 - \|x\|^2)(1 - \|y\|^2)} \right),$$

where $\|x\|$ and $\|y\|$ are the Euclidean norms of points $x$ and $y$ in $\mathbb{R}^d$ respectively, and $\text{arcosh}$ denotes the inverse hyperbolic cosine function. It can be seen that the Poincaré distance within $B^d$ exhibits a gradual increase concerning the norm of $x$ and $y$. This locality property of the Poincaré distance makes the model suitable for embedding trees and data with inherent hierarchies. For instance, if the root node of a tree is positioned at the origin of $B^d$, it maintains relatively short distances from all other nodes due to its Euclidean norm being zero. Conversely, leaf nodes can be situated near the boundary of the Poincaré ball as the distance escalates rapidly between embeddings with norms close to one. This characteristic enables us to construct embeddings that capture both the hierarchical nature of objects, as evident from their norms, while also encapsulating their similarities with the Poincaré distance.

Multiple equivalent models of the hyperbolic space exist, such as the Lorentz model, which is particularly well-suited for machine learning applications. Initially, these models were employed to create shallow word representations that capture inherent hierarchies with greater efficiency [37], [38]. Later, utilizing differential geometry, researchers incorporated hyperbolic spaces into deep learning methods, constructing hyperbolic versions of known neural architectures such as feed-forward neural networks [39], attention mechanisms [40], variational autoencoders (VAE) [41], and even graph convolutional neural networks [42].

A wide variety of biomedical domains have an underlying hierarchical structure. Therefore, applying hyperbolic embeddings to them could be a reasonable choice. As an illustration, the utilization of hyperbolic matrix factorization in predicting drug-target interactions demonstrates superior performance compared to previous Euclidean matrix factorization methods [43]. In another case, K. Yu. et al. applied a VAE for drug molecule generation, while at the same time, they regularized it to align its hyperbolic latent space with the encoding of the Anatomical Therapeutic Chemical (ATC) hierarchy of drugs [44]. Another prominent example is the Poincaré maps (PMAP), a manifold learning method, which produced state-of-the-art two-dimensional representations of cell trajectories on single-cell RNA sequencing (scRNA-seq) datasets [45]. Later, A. K. Susmelj et al. used the same method to embed proteins based on multiple sequence alignment data [46].

Before preprocessing, PMAP first estimates the local similarities and produces a KNN graph, then approximates the global manifold structure based on the KNN graph using the relative forest accessibility index. By optimizing the KL divergence, these global proximities are preserved through the Poincaré embeddings. The resulting two-dimensional representations
have proved to be effective in a wide variety of downstream data analysis tasks, such as visualization, clustering, lineage detection, and pseudotime inference. Following PMAP, several approaches were proposed to harness the hyperbolic nature of gene expression data and provide non-Euclidean embeddings for both scRNA-seq and microarray measurements, some of them relying on manifold learning techniques such as t-SNE [4], while others using deep learning-based discriminative [6], [7], or even generative models [5].

Using hyperbolic representations on gene expression data for compound screening is a novel and promising approach. Following the idea, we utilized and evaluated different hyperbolic manifold learning techniques on the LINCS screenings defined by the corresponding differentially expressed landmark signatures. In the following, we use the LINCS experimental setup and its corresponding differentially expressed landmark signatures interchangeably. To the best of our knowledge, non-Euclidean embeddings have not yet been applied neither to the LINCS landmark genes nor to any other DEG signatures. Although hyperbolic effects become more potent when more genes are considered, it has been shown that even in the case of signatures with only 1,000 genes, applying hyperbolic embeddings leads to superior results compared to the Euclidean methods [4]. Considering that the carefully selected landmark genes represent the space spanned by the whole genome with less distortion than any randomly selected 1,000 genes, hyperbolic embeddings are expected to perform well on the landmark genes as well. Therefore, the main novelty and relevance of our work lie in applying hyperbolic representations to DEG signatures regardless of whether they contain only LINCS landmark genes or not. The global structure of the level 3 gene expression data is strongly characterized by the hierarchical cell line information, making it well suited for hyperbolic methods. However, level 4 DEG signatures emphasize the alterations between the expressed genes with different experimental conditions, thereby surpassing the cell line information. Thus, it is not a foregone conclusion that techniques performing well on gene expression data will yield the same results when applied to DEGs. Our study aims to address the research question concerning whether DEG signatures exhibit an inherent hyperbolic structure or are more suitably represented within Euclidean space.

### III. Experimental Setup

In our research, we employed two distinct datasets to facilitate a comparative analysis of five dimensionality reduction methods across nine metrics.

#### A. Datasets

We leveraged two LINCS datasets, both of which provide differential expression signatures using the CD method [2].

The L1000FWD method [33] serves as one of our primary data sources. This publicly available dataset comprises drug and small-molecule induced CD DEG signatures with the 978 landmark genes obtained through the L1000 microarray technique. Initially, it contained a substantial 42,809 samples (called LINCS experiment subsequently). However, after preprocessing and eliminating non-significant signatures, the authors arrived at a refined dataset containing 16,848 samples. For our analysis, we utilized the same filtered signatures, which encapsulate a rich diversity, representing 68 distinct cell lines, 3,237 drugs, three perturbation time points, and 132 different dosages. Alongside each landmark signature, the dataset also provides the complete signature encompassing all 22,268 genes. Additionally, metadata such as cell type, time point, concentration, and attributes of the drugs, including MOA and clinical phase, are available. The dataset comes with an interactive two-dimensional visualization, providing a valuable tool for domain experts to explore the signature landscape and identify drug MOA. Beyond the metadata and visualizations, the authors provide the coordinates of the FWD embeddings as well, thereby facilitating the reproducibility of their work.

For hyperparameter optimization, we primarily utilized the L1000FWD dataset. As an additional resource, to complement our evaluation and validate our findings, we incorporated the SigCom LINCS dataset as well [27]. SigCom LINCS also offers level 5 CD DEG signatures, but it slightly differs from the L1000FWD. It does not use L1000 measurements, nor the 978 landmark genes, instead, it encompasses a unique set of 12,327 genes. The other difference is that the authors utilized UMAP for visualizing the signature space, in contrast to the FWD method used by L1000FWD. Being compatible with L1000FWD, we chose to work only with the signatures of chemical perturbations provided. To manage the substantial volume of data effectively, we segmented the SigCom LINCS dataset into sub-datasets based on distinct cell lines. This approach mirrors the one employed by the authors of the original paper to create the UMAP visualizations. Specifically, we selected four different cell types — NEU, NPC, HELA, and A375 — each characterized by varying sample sizes.

As for labeling, we used the available MOA, cell type, and time point information in both datasets. Utilizing the labels, we have created sub-datasets suitable for classification tasks. We focused on the top 20 MOAs and top 2 time points for classification tasks in both datasets to address label imbalance and align with prior works. For the simplicity of the MOA task evaluation, we opted to filter out signatures associated with more than one known MOA, following a practice consistent with previous benchmarks [15]. Since the L1000FWD data is not separated by cell type, we constructed a top 20 cell line prediction task on it as well. The characteristics of the resulting classification tasks are summarized in Table I, providing a comprehensive overview of the datasets we used.

#### B. Methods

In our experimental setup, we employed and evaluated a range of dimensionality reduction techniques. The simplest baseline method we used is the PCA, known for its ability to capture global structure and variance within the data. For Euclidean
TABLE I

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Number of Samples</th>
<th>Most Frequent Class</th>
<th>Least Frequent Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1000FWD all cell lines</td>
<td>Top 20 MOA</td>
<td>2,979</td>
<td>topoisomerase inhibitor (280)</td>
<td>calcineurin inhibitor (74)</td>
</tr>
<tr>
<td></td>
<td>Top 20 cell line</td>
<td>16,130</td>
<td>VCAP (2,677)</td>
<td>NEU (51)</td>
</tr>
<tr>
<td></td>
<td>Top 2 perturbation time</td>
<td>16,831</td>
<td>24 (7,380)</td>
<td>6 (9,451)</td>
</tr>
<tr>
<td>SigCom LINCS NEU</td>
<td>Top 20 MOA</td>
<td>561</td>
<td>HDAC inhibitor (92)</td>
<td>Cyclooxygenase inhibitor (17)</td>
</tr>
<tr>
<td></td>
<td>Top 2 perturbation time</td>
<td>4,452</td>
<td>24 h (3,773)</td>
<td>Ephrin inhibitor (56)</td>
</tr>
<tr>
<td></td>
<td>Top 20 MOA</td>
<td>1,777</td>
<td>HDAC inhibitor (288)</td>
<td>6 h (679)</td>
</tr>
<tr>
<td></td>
<td>Top 2 perturbation time</td>
<td>11,052</td>
<td>24 h (9,656)</td>
<td>6 h (1,396)</td>
</tr>
<tr>
<td>SigCom LINCS NPC</td>
<td>Top 20 MOA</td>
<td>5,018</td>
<td>Proteasome inhibitor (816)</td>
<td>Potassium channel blocker (108)</td>
</tr>
<tr>
<td></td>
<td>Top 2 perturbation time</td>
<td>24,763</td>
<td>24 h (21,154)</td>
<td>3 h (3,609)</td>
</tr>
<tr>
<td>SigCom LINCS HELA</td>
<td>Top 20 MOA</td>
<td>6,027</td>
<td>Proteasome inhibitor (627)</td>
<td>Sodium channel blocker (179)</td>
</tr>
<tr>
<td></td>
<td>Top 2 perturbation time</td>
<td>46,194</td>
<td>24 h (39,679)</td>
<td>6 h (6,515)</td>
</tr>
</tbody>
</table>

In this manner, we obtained solely the two-dimensional FWD embeddings, exclusively on the L1000FWD dataset. However, due to its suboptimal performance, we chose to exclude it from subsequent analyses.

features, and with the latter, we can control how tightly it is allowed to pack points together. HUMAP underwent a similar analysis, considering the same two hyperparameters while maintaining default values for the others. PMAP demonstrated robustness to its hyperparameters. Thus, we only compared different neighbor numbers and learning rates while keeping other hyperparameters intact.

C. Evaluation Metrics

Beyond the qualitative visual analysis, we applied a well-rounded evaluation strategy considering both local and global perspectives to provide a comprehensive assessment of the techniques.

To evaluate the preservation of the local structure, we conducted KNN classification on the tasks detailed in table I. We examined the hyperparameters of the KNN classifier, such as the neighbor numbers, and whether to use a weighted distance during prediction. For the best results, we chose to use weighted distances and five neighbors, a decision supported by experiments with varying neighbor numbers (5, 10, and 50). The performance of our methods was assessed using an array of classification metrics, including accuracy, top 3 accuracy, top 5 accuracy (only on tasks where there are at least 3 or 5 different classes), precision, recall, macro-F1 score, and the Area Under the Receiver Operating Characteristic Curve (ROCAUC). Notably, we primarily report the F1 score, which is found to be a robust metric and especially suited for addressing class imbalances. Throughout our study, we adopted a 5-fold repeated cross-validation (CV) methodology, dividing the data into five equal segments, and repeatedly training the classifier five times, each time employing a distinct segment as the test dataset. This approach enabled us to obtain the mean and standard deviation for the abovementioned metrics, providing a robust performance measure. We employed a stratified splitting technique during each CV iteration to provide balanced class distributions across splits, ensuring the presence of at least one sample from each class in the test set, even in the smallest datasets. Given the inherent non-deterministic nature of manifold learning techniques, apart from the deterministic PCA and the precomputed FWD measurements, we repeated not only the CV evaluation but also the dimensionality reduction three times, capturing a more comprehensive view of their performance. All mean and standard deviation values reported in the results section were obtained accordingly.
Following the work of H. Huang et al. [2], we utilized Spearman correlation (SC) and Random triplet accuracy (RTA) to measure the global structure preservation. SC score is the Spearman rank-order correlation between two vectors: distances between pairs of points in the high-dimensional signature space and distances between those same points in the low-dimensional reduced space. On the other hand, obtaining the RTA relies on a batch of triplets, considering the orders of similarities between all three possible pairs within each triplet in both the signature and reduced spaces. The accuracy is then determined by the percentage of triplets where the obtained orders match. We sampled 5,000 pairs and triplets to calculate the SC and RTA scores and repeated the measurements five times to ensure robustness, recording mean and standard deviation values. Moreover, in the case of the nondeterministic manifold learning techniques, we did the same as mentioned with KNN classification, repeating the entire process, including dimensionality reduction three times and calculating the overall mean and standard deviation. An example of all the metrics used in our study is provided in Table [1].

In addition to local and global structure preservation, we sought to assess the preservation of the underlying hierarchy within the data. The structure of the DEG space is mainly determined by the perturbations. Thus, we decided to use the hierarchy of the drugs, namely the ATC hierarchy. Our approach closely followed the methodology outlined by K. Yu et al. [44]. We used the same data and dendrogram purity evaluation technique. However, without applying a hierarchy regularization, none of the manifold learning approaches produced representations that adequately preserved the ATC hierarchy. Dendrogram purity values were consistently low and noisy, showing no significant differences between the various methods. We also evaluated the hierarchy preservation on the input signatures and got the same performance, indicating the absence of the ATC hierarchy in the DEG signatures. As a result, we chose to exclude this metric from our comparative analysis.

IV. RESULTS
We executed the various models on a 32GB NVIDIA Tesla V100 GPU. In the subsequent section, we present the results obtained from hyperparameter optimization, as well as different visualizations and a comprehensive quantitative comparison of the optimized methods across various latent dimensions.

A. Hyperparameter Optimization
To reduce the complexity of our analysis, we primarily conducted hyperparameter optimization for two-dimensional representations. Our evaluation encompassed both local and global metrics, with a particular emphasis on the former. Given that differences between global metrics were much lower in general, and MOA prediction constitutes a primary objective in most cases, we accorded greater significance to local metrics when configuring hyperparameters.

In exploring similarity metrics for signatures, we considered various options, including the KS statistic, cosine similarity, Canberra distance, and Euclidean distance, both with and without prior normalization. Notably, cosine similarity without normalization emerged as the most effective choice for both the landmark and the complete signatures consisting of all the 22,268 genes. This aligns with standard practices in CD signature analysis, which predominantly employs cosine similarity for similarity-based searches. Lately, the utilization of the Canberra distance on LINCS signatures exhibited enhanced KNN classification performance [15]. We confirmed that it performs slightly better in cell type and time prediction but displayed reduced performance in the global metrics and MOA prediction. Moreover, the computational cost associated with the Canberra distance was significantly higher. We also explored the utility of these similarity metrics as input to dimensionality reduction techniques, assessing them with and without data normalization, as well as with and without preliminary PCA dimensionality reduction to 100 dimensions. Ultimately, we determined that PCA preprocessing and data normalization were not advantageous. Raw cosine similarity yielded the best results across both datasets. Regarding the signatures used as inputs, landmark signatures consistently outperformed their unfiltered counterparts, underscoring the benefits of expertly curated gene selection. Consequently, we opted to employ landmark signatures as input for the dimensionality reduction models.

For UMAP and HUMAP, we arrived at the optimal configuration of 5 neighbors and a minimum distance of 0.01. While higher neighbor counts between 50 and 250 exhibited marginal improvements in global metrics, they yielded inferior results in the local ones. We experimented with minimum distance values of 0.001, 0.01, 0.05, 0.1, and 0.5, with 0.01 demonstrating superior performance. PMAP was most effective with five neighbors as well, as higher neighbor counts (25, 50, and 250) did not significantly impact global metrics but considerably degraded KNN classification. More than that, the higher the neighbor number, the higher the computational cost. We explored learning rates ranging from 0.01 to 1 and determined that rates below 0.1 yielded suboptimal results, while rates above 0.1 exhibited no significant differences. Thus, we settled on a learning rate of 0.1. Additional hyperparameters were examined, but no substantial variations were observed, prompting us to maintain their default values. Figure [2] illustrates the comparative analysis of various PMAP hyperparameters concerning different local metrics.

To ensure the robustness of our findings, we conducted parallel evaluations on the SigCom LINCS dataset, with results mirroring those obtained with the L1000FWD data, except for one notable discrepancy with the UMAP and HUMAP methods. SigCom initially employed UMAP with 50 neighbors and a 0.05 minimum distance following a z-score normalization and using Euclidean distance. We tried the different similarity metrics with and without normalization on the signatures, and just as with the L1000FWD data, cosine distance without prior normalization performed best. We also tried different minimum distances but found no significant differences, so we kept the initial 0.05 value. However, when evaluating UMAP and HUMAP with neighbor counts of 5 and 50, in contrast to our findings on the L1000FWD data, we determined that the
latter configuration performed better for the SigCom LINCS dataset.

B. Quantitative Comparison

After optimizing hyperparameters, a comprehensive evaluation of different manifold learning techniques was conducted, assessing their performance using various metrics as mentioned previously. Table II presents a detailed comparison of two-dimensional representations acquired through different methods on the L1000FWD dataset. It is evident that, across all KNN classification tasks and metrics, PMAP consistently outperforms other approaches. UMAP ranks as the second-best method regarding local structure preservation, with the primary discrepancy observed in the cell line prediction task. Surprisingly, HUMAP exhibits inferior performance compared to its Euclidean counterpart, although it still surpasses FWD. While PCA ranks as the poorest performer in terms of local structure preservation, it excels in the two global preservation tasks, followed by PMAP, which scores highest among the nonlinear manifold learning methods. The comprehensive evaluation in two dimensions proves valuable for selecting an appropriate visualization method. Given the significance of MOA in DEG signature visualizations and the second-best performance of PMAP in global metrics, PMAP is highly recommended for visualization purposes.

Beyond two-dimensional representations, higher dimensions may also have utility in various machine-learning tasks. To better understand the performance and applicability of the methods, we conducted comparisons across increasing embedding dimensions while maintaining previously optimized hyperparameters. A broad range of dimensions was evaluated, and as an extrapolation limit, we also assessed performance with the non-reduced signatures. Figure 3 illustrates the comparison of various methods across different dimensions using L1000FWD data. Landmark genes consistently perform the best across all tasks, not only in the global metrics, where landmark signatures served as the reference with a score of 1.0 by definition. This underscores the meticulous selection of landmark genes by domain experts. FWD consistently ranks as the poorest performer among the nonlinear manifold learning methods, while PMAP excels across all metrics. The performance curve of the PMAP exhibits a logarithmic growth, reaching saturation values close to those of the landmark genes for local metrics. UMAP also exhibits logarithmic growth, followed by a rapid decline after 64 dimensions, which is not shown in the figures. Furthermore, UMAP consistently lags behind PMAP across all metrics, with notable disparities in the cell line and time prediction tasks. HUMAP mirrors the performance of UMAP, apart from a poor performance at only two dimensions and a sooner performance deterioration after 16 dimensions. PCA shows a steeper increase in performance and higher starting points for global metrics, establishing its superiority in this regard. However, in KNN classification, it performs poorly at lower dimensions. Nevertheless, as dimensionality increases, PCA may outperform other methods even in the local metrics. For instance, PCA surpasses PMAP in the MOA prediction task at 16 dimensions, which can be attributed to the global structure formation around drugs in the DEG space, with MOA serving as an attribute of these drugs. Although KNN MOA prediction is considered a local task, PCA, which better preserves the global structure, can perform well at higher dimensions.

Similar evaluations were conducted with the SigCom data, where only the top 20 MOA and the top 2 perturbation time KNN prediction tasks were available due to cell-based splitting. These datasets provided an opportunity to examine differences across different data volumes. In smaller datasets (NEU and NPC), no significant differences were observed between methods, and the variance was too large to conclude which technique to use. However, with larger datasets (HELA and A375), performance gaps emerged between methods, yielding results similar to those obtained with L1000FWD.
measurements. Figure 4 illustrates the comparison of various methods across different dimensions using A375 data. PMAP consistently performs the best among nonlinear techniques. UMAP and HUMAP exhibit similar performance trends, with HUMAP displaying poor performance at two dimensions and both methods showing declining performance above 256 dimensions. PCA, as before, excels in global tasks but lags behind in local metrics. The MOA gap with SigCom LINCS is more pronounced than the differences according to the perturbation time task, primarily due to the ease of the time prediction task, where all methods readily achieve an F1 score above 0.95.

In summary, no significant differences were observed with small data volumes (below 5,000 samples), and the choice of technique appeared inconsequential. Method selection only became relevant with larger datasets. For lower dimensions (below 16), PMAP seems to be the best choice, while PCA may excel in higher dimensions. Notably, PMAP exhibits robustness to hyperparameters. Although HUMAP offers computational efficiency, it fails to outperform its Euclidean counterpart, suggesting that a hyperbolic embedding space alone does not suffice. The manifold approximation method of PMAP appears to be better suited for capturing the non-Euclidean structure of DEG signatures.

C. Qualitative Evaluation

We further analyzed the two-dimensional representations obtained through various methods and compared them visually on both datasets. In the case of the L1000FWD data, we utilized the same coordinates and colors as the available FWD maps on the L1000FWD web application for improved comparability. SigCom UMAP visualizations are also accessible through the SigCom LINCS portal, but UMAP is non-deterministic, and the coordinates are not given. Thus, it differs from our representations.

Figure 5 displays the embeddings colored according to the most frequently occurring MOAs, colored by cell line information. A comparative analysis of the methods reveals distinct patterns. PCA, for instance, exhibits noisy and disordered representations that align with neither cell line nor MOA categories, making meaningful cluster identification challenging. While showing some well-defined and separable clusters associated with MOAs, FWD scatters signatures from the same cell line across the space. Apart from some distinct, meaningful clusters, most of the signatures are centered in a high-density, noisy region. UMAP, on the other hand, distributes signatures more evenly throughout the space, forming numerous small clusters with distinguishable MOA and cell line properties. Nevertheless, signatures sharing MOA or cell line characteristics are dispersed into multiple distinct clusters, indicating a lack of global organization based on these properties. HUMAP exhibits a similar pattern, with the only difference that embeddings are being hyperbolic. For the hyperbolic spaces, we illustrate both the center and the border of the Poincaré disk. Notably, HUMAP fails to efficiently utilize the available space, with sparse regions around the center, an absence of signature embeddings close to infinity, and the top half of the Poincaré disk being more densely populated compared to the bottom. In contrast, PMAP representations are uniformly spread across the entire space, devoid of high-density, noisy regions observed in PCA and FWD. Furthermore, compared to UMAP and HUMAP, PMAP effectively organizes signatures with shared MOA or cell lines into large coherent clusters.

Figure 7 provides an example using the SigCom A375 dataset, where signatures are color-coded according to perturbation time. Almost all the embeddings result in distinct and well-separable clusters, which supports why estimating...
the perturbation time was an easy task. In this case, it is even more profound that PCA yields excessively noisy representations, HUMAP inefficiently utilizes the available space, while PMAP evenly distributes the signatures, effectively covering the entire Poincaré disk.

Considering these landscape features, PMAP emerges as a promising choice for DEG visualization, yielding an evenly distributed space and improved separation of cell line and MOA information.

V. CONCLUSION

In summary, our investigation has delved into the realm of Differential Expression Gene signatures, exploring their underlying geometric structure and the efficacy of manifold learning methods in unveiling their intricate features. In our study, we utilized two separate datasets provided by the LINCS consortium to facilitate a comparative analysis of both Euclidean and hyperbolic dimensionality reduction approaches.

The results of our study underscore the prominence of Poincaré maps as a compelling choice for handling CD DEG...
Fig. 5. Two-dimensional signature embeddings colored according to the most frequent MOAs in the L1000FWD dataset. The visualizations showcase the distinctive patterns obtained using various manifold learning methods.

Fig. 6. Two-dimensional signature embeddings obtained through different dimensionality reduction techniques colored with the most frequent cell lines in the L1000FWD dataset.

signatures. The resilience of PMAP to hyperparameter variations, coupled with its superior performance across the evaluation metrics, positions it as a reasonable option, particularly in lower dimensions, such as two-dimensional visualization. This reinforces the notion that the DEG space possesses hyperbolic characteristics and retains cell line and time-related information, requiring non-Euclidean manifolds for its effective exploration. In summary, as a general guideline, we
recommend employing hyperbolic embeddings when reducing over 5,000 DEG signatures to fewer than 16 dimensions. Otherwise, a more efficient Euclidean method might be a better choice.

Despite these advancements, our study unveils certain limitations that necessitate exploring novel approaches. The observed differences between manifold learning methods occur primarily in the context of sufficiently large datasets. Unfortunately, the computational demands of PMAP limit its application to a significantly larger data volume, and its inability to embed new entities into the learned signature space is also a notable constraint. An intriguing alternative, the Lorentz model UMAP, while offering computational efficiency, falls short of unveiling the full extent of the hyperbolic nature of the DEG signature space. Consequently, there emerges a need for the development of dimensionality reduction techniques that strike a balance between computational efficiency, hyperbolic space utilization, and adaptability to novel inputs.

Furthermore, our findings present a promising direction for leveraging DEG signatures in predictive tasks for drug discovery. While PMAP exhibits exemplary performance in KNN classification tasks among reduced representations, it falls short compared to non-reduced landmark genes. This observation makes dimensionality reduction useless when considering only the KNN classification performance. However, to enhance predictive performance, utilizing fully connected neural networks or graph convolutional models instead of KNN classification holds promise [15]. In this context, leveraging L1000 signatures as input, the incorporation of hyperbolic variants of existing machine learning-based methods may yield improved performance across various supervised tasks. These tasks encompass predicting side effects and MOA of drugs or the expression levels for the non-landmark genes using hyperbolic neural networks, as well as predicting drug-drug or drug-target interactions via hyperbolic matrix factorization. Both of these approaches utilize low-dimensional internal representations of the input, where using non-Euclidean spaces could prove advantageous, potentially capturing additional information within the DEG signatures, as demonstrated by our findings with unsupervised manifold learning methods.

We hope that our study contributes valuable insights that aid the selection of optimal methods for visualizing DEG spaces and catalyze further research endeavors aimed at harnessing hyperbolic techniques for downstream tasks involving LINCS or other DEG data.

**ACKNOWLEDGMENTS**

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 RF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory.

**REFERENCES**


