Topological Data Analysis-Deep Learning Framework for Predicting Cancer Phenotypes

Lebohang Mashatola[1], Ismail Akhalwaya, PhD[1], and Stephanie J. Müller, PhD[1]

[1]IBM Research Africa, Johannesburg, South Africa

Abstract. Classification of patient cancer phenotypes from gene expression profiles remains a challenge in the field of transcriptomics. Gene expression data typically suffers from extreme noise and performs poorly on deep learning models.

We build on previous work by incorporating the concept of differential gene expression analysis to pre-select genes that are necessary but not sufficient individually for disease association in our topological data analysis approach. Gene pre-selection reduces the computational cost for the calculation of persistent homology. Furthermore, multiple topological representations are used as input for classifying three cancer phenotypes.

Deep learning with topological features improved cancer type prediction compared to its use on raw data. Furthermore, the use of persistent landscapes performed best for the different gene expression datasets compared to other topological representations. Thus, topological features offers a new perspective on deciphering the non-linear connection between genotype and phenotype.

Keywords. Topological data analysis, Deep learning, Gene expression, Cancer Phenotype prediction

1 Introduction

1.1 Topology Overview

Topological data analysis (TDA) is a powerful method for extracting a set of refined, robust quantitative features on the structure of data by translating the data and encoding it into shape [1]. For this study, (the combined use of TDA and Deep Learning) TDA was explored for sample-specific disease prediction using transcriptome data. Transcriptome or gene expression data utilises ribonucleic acid (RNA) sequencing for the quantitative measurement of messenger RNA transcripts [2]. Messenger RNA is translated into protein subsequently determining the level of expression which influences downstream regulation of

Summary of the TDA workflow

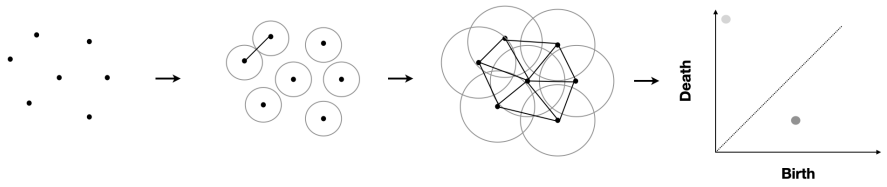


Figure 1: Iterative growth of the simplicial complex achieved by increasing the radius around each point subsequently forming the connection of edges. Followed by a persistent diagram summarising the birth and death of features.

biological pathways and molecular functions [3]. Gene expression data however, suffers from extreme noise and the aim of this study is to identify the subtle signals embedded in the data for the classification of cancer types. The use of Deep Learning (DL) alone for the classification of gene expression data performs relatively poorly as reported in [4, 5]. Thus, we propose augmenting DL with TDA to improve disease-prediction.

The TDA pipeline involves an input of finite data points represented as correlation, distance or coordinates within an N-dimensional space [6]. Geometry is inferred from data points to form a simplicial complex. This process involves continually increasing the radius around each data point. Upon the intersection of the space around each data point, an edge is connected (see Figure 1). Sample-specific topological fingerprints were attained by the use of weighted Vietoris-Rips complexes. A weighted Vietoris-Rips complex is built from the distance matrix and the weights on the vertices. Weighted Vietoris-Rips complexes are particularly useful for determining the topology amongst outliers and noise [7].

1.2 Persistent Homology

PH presents itself as a central approach for TDA by summarising the topological features in a continuously growing simplicial complex. Persistence is defined as invariant topological features that encode intrinsic information on the topology of the data [8]. Topological features exist in multiple scales and these can be categorised into Betti numbers. For example Betti-0 represents connected edges in the simplicial complex and Betti-1 represents one-dimensional loops [1].

Persistent diagrams (PD) are useful representations of the PH represented by $D = b_i, b_j \in R^2 | i \in 1, b_i < d_i$. Every point is an invariant topological feature represented by a birth and death coordinate (b_i, d_i) (see Figure 1). PD's are therefore represented as multisets and are required to be vectorised for their use

in ML and DL tasks [9].

1.3 Topological Representations

Numerous finite-dimensional vector-representations exist and we focus on topological representations that have already been used in ML/DL frameworks. The simplest form of topological feature vectors result in scalar-valued summary statistics for example the total persistent and persistent entropy [10]. These representations oversimplify PD's and are not applicable for complex ML and DL tasks [11]. More expressive and stable representations such as persistent landscapes (PL) and persistent images (PI) are tested in our study [12]. PL's are sequences of decreasing sequence of functions $(\lambda_1, \lambda_2, ...\lambda_k)$ [13]. The sequence of functions are subsequently converted into a suitable ML/DL usable vector. PI's take the persistent module G and concatenates multiple PD's from all homology dimensions into a single vector [14]. Thus, we hypothesise that PL's and PI's are robust metrics to yield optimal model performance.

2 Methods

2.1 Data Processing

Gene expression data was obtained from The Cancer Genome Atlas (TCGA; https://www.cancer.gov /tcga). Three of the most commonly occurring cancer types including lung (LUAD), breast (BRCA) and colorectal cancer (COAD-READ) gene expression data were obtained. A total of 160 LUAD, 133 BRCA and 158 COAD-READ samples were used. For each dataset 19958 protein-coding genes were selected, removing non-coding RNAs. A pairwise cancer-type comparison (LUAD vs COAD-READ, COAD-READ vs BRCA and LUAD vs BRCA) by merging the appropriate datasets into a single matrix (X) was performed. Matrix X_{ij} contains rows as patients (i) and columns as genes (j). Matrix X was randomly shuffled and split into 70% for model training (X_{train}) and 30% for model testing (X_{test}).

2.2 TDA Implementation

To determine per-sample topological features, the work-flow shown in figure 2 was followed. This involved first constructing a distance correlation matrix (D) using [15]. The distance correlation matrix was determined from X_{train} by calculating all pairwise distances between genes $(X_{s,i}; X_{s,j})$ $F \in [0, 1]$. Distances tending to zero, indicate inter-gene independence and were more correlated and those tending to one indicated inter-gene dependence and are less correlated. To construct $M_{i,j}$ differential gene expression analysis was used to identify significantly up- and down-regulated genes. Since correlation is defined by pairwise distances between the expression levels of genes, selecting genes associated with

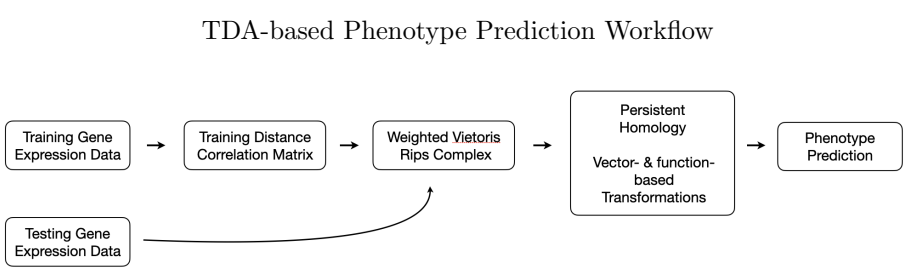


Figure 2: TDA workflow for patient-level disease-prediction.

a particular phenotype, removes confounding genes that are typically required for the maintenance of normal cellular functions regardless of the phenotype.

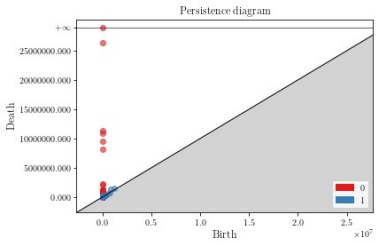
The R package edgeR [16] was used to perform differential gene expression between cancer types. To decrease the false discovery rate, the Benjamini-Hochberg adjusted probability value (BH-adjusted p-value) was applied to identify significant differential expressed genes (BH-adjusted p-value < 0.05) [17]. Significant differential expressed genes greater than a log fold change of 5 and less than -5 were then pre-selected to construct $M_{i,j}$. The removal of confounding genes not only removes redundancy but also reduces the computational cost for downstream TDA experiments. This bypasses the use of sub-sampling techniques required for the calculation of PH for high dimensional data [18].

Using the Python package Gudhi [19] (https://gudhi.inria.fr), per sample simplicial complexes were calculated using a weighted Vietoris-Rips complex which introduces weights to vertices of the growing complex. The filtration value of the vertex i is $2 * F_i$ and the filtration value of the edge (g_i, g_j) is $D_{ij} + F_i + F_j$ [20]. A collection of edges and filtrations (σ) were then generated and PH was determined. Topological birth and death coordinates for the zeroth and first homology were determined for each sample and topological vector-based transformations using PL and PI were performed.

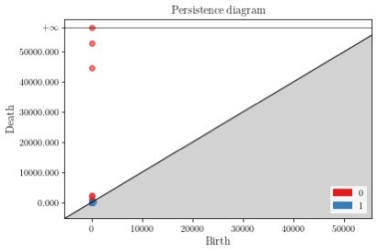
Per-patient topological fingerprints were represented as PD's. Topological differences in Betti-0 and Betti-1 features were used to discriminate between subjects (see figure 3). However, these required vectorisation for them to be translated into ML/DL models.

2.3 Vector-Based Representations

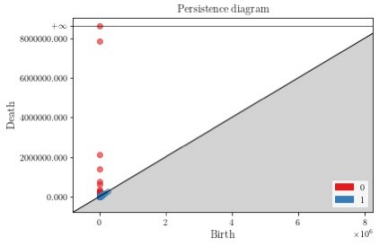
Topological birth and death coordinates were transformed for downstream DL disease-prediction using PL's that are a collection of one-dimensional piecewise-linear functions computed from the PD rank functions. These piecewise-linear



(a) LUAD Patient



(b) BRCA Patient



(c) COAD-READ Patient

Figure 3: Per patient PD's for Betti-0 and Betti-1 features used to create fixed-dimensional vectors for downstream DL modelling.

functions are sampled evenly on a given range $F = [0, \sigma_{\max}]$ (where $\sigma_{\max}$ is the filtration at which the last edge is added) and the corresponding vectors are concatenated, returning landscapes across multiple resolutions [21]. Furthermore, PI's which represent two-dimensional functions computed from a PD by embedding multisets were used. The PD is subsequently sinuated into a weighted Gaussian kernel [14]. The multisets are then dispersed into an image with pixels and flattened as a finite-dimensional vector [19].

2.4 Phenotype Prediction

A multilayer perceptron neural network (MLP) classifier was fitted on per-patient PL's and PI's classifying subjects in X by their cancer type (LUAD, BRCA and COAD-READ). The MLP classifier was also fitted on raw gene expression data representing the baseline non-TDA method for phenotype prediction. The MLP model architecture included three layers, a relu activation function with a regularisation step added to the loss function. Forward and back-propagation to adjust neural weights was performed with 50 epochs. The model learning process was repeated five times using a reshuffled X_{train} and X_{test} and the mean and standard deviation were reported on unseen data. Model training and testing was performed using the Python package scikit-learn [22] (https://scikit-learn.org/stable/).

3 Results and Discussion

To determine the generalised accuracy, model testing on unseen data was performed using multiple random splitting of training and testing datasets (Monte Carlo cross-validation). For each random training data split, the accuracy is determined. The mean accuracy is reported in table 1, 2 and 3.

Table 1: Model performance metrics for the prediction of cancer phenotype using PI and PL for LUAD and COAD-READ patients.

Representation	Accuracy	F1 (macro)	F1 (micro)	TPR	Precision
Non-TDA	46.15	46.18	44.70	70.37	46.10
Persistent Landscape	**92.30**	**92.23**	**92.31**	**88.13**	**96.29**
Persistent Image	88.03	87.83	88.34	81.48	91.67

Table 2: Model performance metrics for the prediction of cancer phenotype using PI and PL for COAD-READ and BRCA patients.

Representation	Accuracy	F1 (macro)	F1 (micro)	TPR	Precision
Non-TDA	49.10	40.31	49.49	87.50	49.11
Persistent Landscape	**91.67**	**91.61**	**92.68**	**90.46**	**91.60**
Persistent Image	85.22	84.99	82.05	84.21	85.30

Table 3: Model performance metrics for the prediction of cancer phenotype using PI and PL for LUAD and BRCA patients.

Representation	Accuracy	F1 (macro)	F1 (micro)	TPR	Precision
Non-TDA	50.91	33.73	51.09	54.44	45.12
Persistent Landscape	65.06	**64.74**	60.78	**77.5**	65.10
Persistent Image	**66.67**	62.50	**62.64**	63.20	**67.00**

Table 1, 2 and 3 shows the model accuracy, F1 scores, true positivity rate (TPR) and precision for the baseline non-TDA as well as the TDA-based method with two topological representations for the prediction of cancer type. Both micro and macro-F1 scores are included to assess the class imbalance by giving equal weight to each class [23]. The TDA approach showed an improvement compared to deep learning used on raw gene expression data in all three comparisons. Both PL and PI performed similarly in all three comparisons supporting that PL and PI provide expressive and stable vector-representations of the PD.

Improved prediction performance is attained by using a pre-selected gene set, resulting in lower computational cost. We recommend that differentially expressed genes be used for the construction of distance correlations to speed up computation of the PH without compromising the effectiveness of topology to unravel the complexities of the transcriptome. The work performed here

demonstrates that the developed framework involving deep learning and TDA can be translated to predict other phenotypes from related gene expression data.

Further work involves using different approaches for the construction of distance correlation. Prioritising genes that are associated with a specific phenotype is crucial in providing useful topological fingerprints. Furthermore, the addition of biological knowledge to prioritise phenotype-associated genes into our TDA framework will be assessed. Nonetheless, we have demonstrated that TDA is able to identify crucial signals embedded in the transcriptome and we recommend that the use of topological features be employed for further elucidation of the relationship between genotype and phenotype.

References

[1] L. Wasserman, "Topological data analysis," *Annual Review of Statistics and Its Application*, vol. 5, pp. 501–532, 2018.

[2] W. V. Gilbert, T. A. Bell, and C. Schaening, "Messenger rna modifications: form, distribution, and function," *Science*, vol. 352, no. 6292, pp. 1408–1412, 2016.

[3] K. V. Morris and J. S. Mattick, "The rise of regulatory rna," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 423–437, 2014.

[4] D. Urda, J. Montes-Torres, F. Moreno, L. Franco, and J. M. Jerez, "Deep learning to analyze rna-seq gene expression data," in *International work-conference on artificial neural networks*. Springer, 2017, pp. 50–59.

[5] R. Shahane, M. Ismail, and C. Prabhu, "A survey on deep learning techniques for prognosis and diagnosis of cancer from microarray gene expression data," *Journal of computational and theoretical Nanoscience*, vol. 16, no. 12, pp. 5078–5088, 2019.

[6] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, "A roadmap for the computation of persistent homology," *EPJ Data Science*, vol. 6, pp. 1–38, 2017.

[7] H. Anai, F. Chazal, M. Glisse, Y. Ike, H. Inakoshi, R. Tinarrage, and Y. Umeda, "Dtm-based filtrations," in *Topological Data Analysis*. Springer, 2020, pp. 33–66.

[8] K. Xia, X. Feng, Y. Tong, and G. W. Wei, "Persistent homology for the quantitative prediction of fullerene stability," *Journal of computational chemistry*, vol. 36, no. 6, pp. 408–422, 2015.

[9] C. Hofer, R. Kwitt, M. Niethammer, and A. Uhl, "Deep learning with topological signatures," *Advances in neural information processing systems*, vol. 30, 2017.

[10] M. Buchet, F. Chazal, S. Y. Oudot, and D. R. Sheehy, "Efficient and robust persistent homology for measures," *Computational Geometry*, vol. 58, pp. 70–96, 2016.

[11] F. Hensel, M. Moor, and B. Rieck, "A survey of topological machine learning methods," *Frontiers in Artificial Intelligence*, vol. 4, p. 681108, 2021.

[12] X. Glorot and Y. Bengio, "Proceedings of machine learning research." JMLR Workshop and Conference Proceedings, 2010.

[13] P. Bubenik *et al.*, "Statistical topological data analysis using persistence landscapes." *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 77–102, 2015.

[14] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, "Persistence images: A stable vector representation of persistent homology," *Journal of Machine Learning Research*, vol. 18, 2017.

[15] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The annals of statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[16] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[17] M. Bogdan, J. K. Ghosh, and S. T. Tokdar, "A comparison of the benjamini-hochberg procedure with some bayesian rules for multiple testing," *arXiv preprint arXiv:0805.2479*, 2008.

[18] F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman, "Subsampling methods for persistent homology," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2143–2151.

[19] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec, "The gudhi library: Simplicial complexes and persistent homology," in *International congress on mathematical software*. Springer, 2014, pp. 167–174.

[20] S. Mandal, A. Guzmán-Sáenz, N. Haiminen, S. Basu, and L. Parida, "A topological data analysis approach on predicting phenotypes from gene expression data," in *International Conference on Algorithms for Computational Biology*. Springer, 2020, pp. 178–187.

[21] P. Bubenik, "The persistence landscape and some of its properties," in *Topological Data Analysis*. Springer, 2020, pp. 97–117.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[23] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 42–49.