A Topological Data Analysis Approach on Predicting Phenotypes from Gene Expression Data

Sayan Mandal[1], Aldo Guzmán-Sáenz[2], Niina Haiminen[2], Saugata Basu[3], and Laxmi Parida[2](✉)

[1] The Ohio State University, Columbus, OH, USA
[2] IBM Research, T. J. Watson Research Center, Yorktown Heights, NY, USA
parida@us.ibm.com
[3] Purdue University, West Lafayette, IN, USA

Abstract. The goal of this study was to investigate if gene expression measured from RNA sequencing contains enough signal to separate healthy and afflicted individuals in the context of phenotype prediction. We observed that standard machine learning methods alone performed somewhat poorly on the disease phenotype prediction task; therefore we devised an approach augmenting machine learning with topological data analysis.

We describe a framework for predicting phenotype values by utilizing gene expression data transformed into sample-specific topological signatures by employing feature subsampling and persistent homology. The topological data analysis approach developed in this work yielded improved results on Parkinson's disease phenotype prediction when measured against standard machine learning methods.

This study confirms that gene expression can be a useful indicator of the presence or absence of a condition, and the subtle signal contained in this high dimensional data reveals itself when considering the intricate topological connections between expressed genes.

Keywords: Topological data analysis · Gene expression · Phenotype prediction · Parkinson's disease

1 Introduction

Our aim in this work was to investigate if the expression of protein-coding genes provide sufficient information to classify an individual with respect to a phenotype, in particular Parkinson's disease (PD). There is an urgent need for developing biomarkers for diagnosing and monitoring the progression of Parkinson's disease. The combination of multiple cerebrospinal fluid biomarkers has emerged as an accurate diagnostic and prognostic model, while blood-based markers have

S. Mandal—Work was done while interning at IBM T. J. Watson Research Center.

C. Martín-Vide et al. (Eds.): AlCoB 2020, LNBI 12099, pp. 178–187, 2020.
https://doi.org/10.1007/978-3-030-42266-0_14

also been explored [10,22]. Recently, also gene expression and methylation signatures from blood samples have been examined for this purpose [30].

In the current work we examined the use of sequencing-based gene expression values from blood samples as features for predicting disease diagnosis. We found standard machine learning methods ineffective and instead looked into the possibilities of understanding the shape of the high-dimensional gene expression data by taking into account its topological features. More specifically, we examined persistent homology emerging from the gene expression data.

In life sciences, topological data analysis (TDA) has previously been applied in medical imaging [13,29], protein characterization [8,17], describing molecular architecture [24,26], and cancer genomics [3,21]. There have been several studies exploring TDA in genomics [7]. Gene expression from peripheral blood data has been used to build a model based on TDA network model and discrete Morse theory to look into routes of disease progression [27]. Persistent homology has also been employed for comparison of several weighted gene coexpression networks [18].

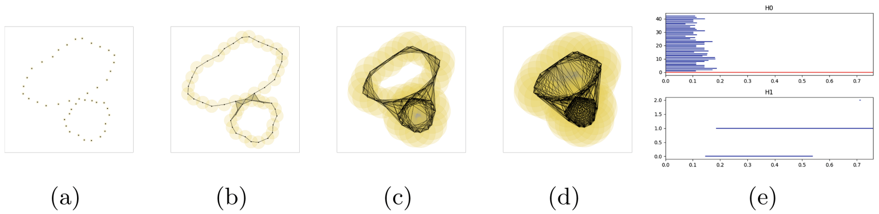
Here we describe a method that translates gene expression measurements for an individual sample into a weighted point cloud. We further hypothesize that this weighted point cloud has topological information relevant to classification tasks. Since the point cloud generated by directly mapping gene expressions can be large and in very high dimension, standard topological data analysis (TDA) algorithms suffer from a combinatorial explosion. We therefore employ subsampling and averaging methods using much fewer points. This subsampling method is also robust in terms of noise present in the original point cloud.

Ultimately, we use the gene expressions of subjects with and without Parkinson's disease to generate topological summaries per subject. These summaries essentially act as unique fingerprints that describe the topology of the gene expression in a sample. We use these fingerprints to enhance the feature vector that is used for disease phenotype prediction, and in turn achieve improved results compared to standard machine learning methods (support vector machines, random forests, neural networks). Our study also implies that gene expression measured from blood samples is a useful indicator of the presence or absence of Parkinson's disease.

2 Methods

2.1 Setting up the Problem

We work under the hypothesis that the set X of all subjects' samples, each encoded as a collection of gene expression values, can provide us with enough topological information to discern between healthy subjects and subjects with Parkinson's disease. We denote by X a matrix of size $n_{rows} \times n_{cols}$ where each row corresponds to a subject and each column corresponds to a gene. Each entry $X_{i,j}$ then corresponds to the j-th gene expression of the i-th subject.



(a) (b) (c) (d) (e)

Fig. 1. Illustration of persistent homology. Here (e) represents the barcodes associated with increasing radius from (a) to (d).

The co-expression patterns between genes can reveal functional connections between them, e.g. the expression of a set of genes belonging to the same biological pathway may be co-ordinated (see [16] for applications of co-expression analysis). In particular, pathways perturbed in disease may contribute to gene expression differences between healthy and afflicted subjects.

Co-expression can be examined by computing pairwise correlations between gene expression measurements. Therefore, we construct a new matrix $\bar{X}$ from X, consisting of all pairwise distance correlations between *genes* (columns), rather than samples (rows), as defined in [28]. Distance correlation conveys different information about the relations between genes than the standard Pearson's correlation as it measures both linear and nonlinear association, whereas the former can only detect linear association.

Later, we show how to use the theory of persistent homology to determine the persistent topological landscapes present in the gene expression data of a sample, by first transforming it into a weighted point cloud. We do this transformation by utilizing the gene correlations across all available samples (matrix $\bar{X}$). The topological summaries of the weighted point clouds (persistence landscapes) are then used to construct a machine learning model to predict the phenotype (healthy or PD) for each sample.

2.2 Topological Background

In this section, we give a brief intuition as to how persistent homology is used to describe the shape of data. For a full account of basic constructions in this area, we refer the reader to [9]. We now give a brief intuitive description of a filtered Čech complex of a specific covering.

Consider the toy example of a set points sampled uniformly from a "figure eight" shape in $\mathbf{R}^2$ as shown in Fig. 1a. We can start growing disks around each point in the sample and consider the union of these disks. Initially when the radius ρ of the disks are close to zero, we get a set of disconnected sets (Fig. 1a). As we continue, we notice that at some particular radius $\rho = \rho_1$, a set with the same topology as the figure eight is obtained (Fig. 1b). Upon increasing the radius further to $\rho = \rho_2$, the smaller of the two holes in the figure is completely filled while the other remains (Fig. 1c). Further increasing $\rho = \rho_3$, fills up the larger hole as well (Figure 1d), making the union of disks consist of a single connected component for all $\rho > \rho_3$.

The next step is to look at the combinatorial information contained in the evolving unions of disks. We achieve this by constructing a *filtered simplicial complex*, a mathematical object consisting of vertices, edges, triangles, tetrahedra, and their higher dimensional analogues, called simplices, with information on *when* these are added to the complex. First we add a vertex for each point in the point cloud at time 0, then whenever two disks intersect, we add an edge. Every time three disks intersect, we add a triangle, and similarly we add higher dimensional simplices for higher order intersections. Therefore, we get a sequence of simplices. In this sequence, the topological persistence computation is a set of *birth-death* pairs of homology cycle classes that indicate when a class is born and when it dies. In the previous example, it indicates when the loops are formed and when they are filled up. The pair (birth,death] of these homology classes can be indicated as a sets of points in $\mathbf{R}^2$ (called persistence diagrams; see [15]) or as sets of horizontal lines (called persistence barcodes; see [9]; illustrated in Fig. 1e).

Keeping track of the holes in this union of disks lets us know how long they persist, and we apply the same reasoning for more general point clouds with a notion of closeness or distance, even when they are not metrics in the mathematical sense. In particular we work with a set of points that form a *semimetric space*, a space that satisfies all but one of the metric spaces axioms: the triangle inequality does not hold in general.

In this article, we use the same construction as the one of a Weighted Vietoris-Rips complex [6, Section 5], pointing out that the properties obtained in the cited article state as hypothesis that the input space is metric. It is obvious, however, that applying the construction to a semimetric space still yields a filtered simplicial complex.

The rank of homology groups of a space, in the usual non-persistent setting, are called Betti numbers of said space, and are denoted by $\beta_k - number$. Intuitively, these numbers account for different topological features: β_0 indicates the number of connected components, β_1 indicates the number of cycles or loops, whereas β_2 counts the number of voids. In our example, Fig. 1e represents barcodes associated with Figs. 1a–d. In Fig. 1e (H_0), we have many β_0 at the beginning, as observed in Fig. 1a which die as ρ increases to connect the different components. In Fig. 1e (H_0) we have a red bar, which corresponds to the final connected component that lives forever. In Fig. 1e (H_1), we notice two long bars, representing the two loops that are formed and then eventually filled up. The fact that one bar is longer than the other suggests the size of one hole greater than the other. Also, notice that these bars in H_1 appear up after the cycles in H_0 have died. In our data analysis using topological features, we wish to use these intervals in persistent homology to better understand the shape of the data.

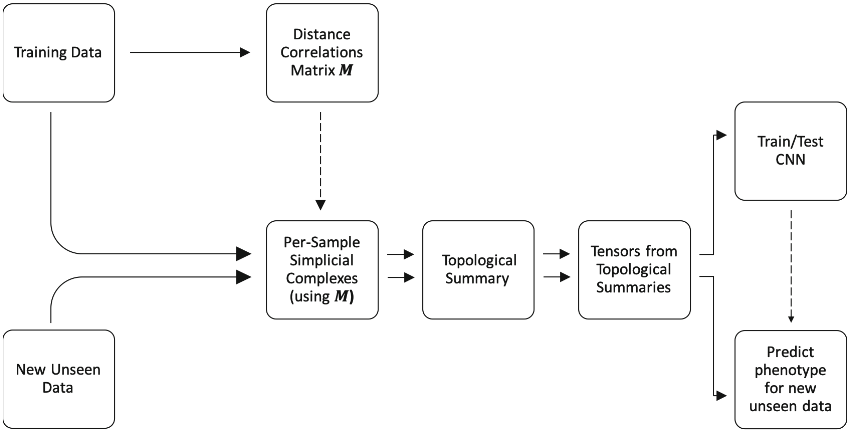

Training Data
Distance Correlations Matrix M
Train/Test CNN
Per-Sample Simplicial Complexes (using M)
Topological Summary
Tensors from Topological Summaries
New Unseen Data
Predict phenotype for new unseen data

Fig. 2. TDA pipeline flowchart.

2.3 TDA Workflow

Our overall workflow is described in Fig. 2 and the relevant steps are discussed in detail below.

Consider $Z = X_{train}$, a matrix of size $m_{rows} \times n_{columns}$ corresponding to a subset of samples in X that will be used to train a classifier later on. Recall that $Z_{i,j}$ corresponds to the j-th gene expression of the i-th subject in the gene expression matrix Z, and that the matrix $\overline{Z}$ consists of all pairwise sample distance correlations between *columns* (genes) of Z. From [28, Theorem 3], we know that the distance correlation between two vectors (samples of a random variable) is zero if and only if are they are independent, that it is non-negative, and that it is bounded above by 1. It is also symmetric. In our particular data, we additionally have that no two different columns have distance correlation of 1.

Next we define the matrix $M = \mathbf{1} - \overline{Z}$, where $\mathbf{1}$ denotes the $m_{rows} \times n_{cols}$ constant matrix with value 1. By the preceding paragraph, M defines now a semimetric on the set of genes in our dataset.

We now, for each subject s in X, construct a filtered simplicial complex K_s, as follows:

1. Its vertices are the set of genes. All added at time 0.
2. For each edge $\sigma = (g_i, g_j)$, we compute

$$t_\sigma = \frac{\sqrt{M_{i,j}^2 + \left(\frac{X_i^2 - X_j^2}{M_{i,j}}\right)^2 + 2X_i^2 + 2X_j^2}}{2},$$

   and add σ to K_s at time t_σ.
3. For each simplex σ with $|\sigma| > 2$ we add it at the maximum time of addition of its edges.

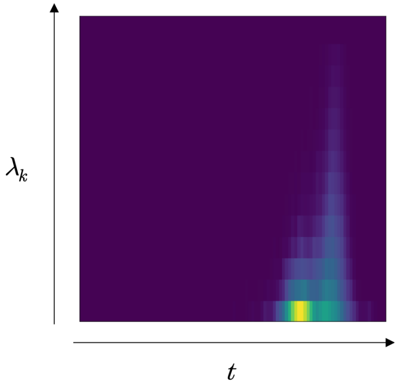


Fig. 3. Discrete sampling of a persistence landscape of a subject, brighter colors indicate higher values in the landscape.

This is an analytical solution for adding simplices to the Weighted Vietoris-Rips construction, studied in [6], for the semi metric space M (of genes) with weights on X, the original data matrix. We are essentially assigning to each gene its expression associated to the subject s. In the actual implementation, we multiplied the weights by a scaling term, since all distances are bound by 1 above but the weights themselves can be higher.

To mitigate the computational cost of our setup we used a subsampling approach, as studied in [11], so that instead of working with the entire set of genes at all times, for each subject we repeatedly subsampled smaller sets of $n_{subsample}$ genes, obtaining several filtered simplicial complexes.

For each of the simplicial complexes we obtained persistence landscapes [5] for homology dimensions 0 and 1. Such landscapes are, for each homology degree, sequences $\{\lambda_k\}$ of decreasing piecewise linear (PL) functions $\lambda_k : \mathbb{R} \to \mathbb{R}$. We elected persistence landscapes as opposed to barcodes or diagrams because of their amenability to compute statistical estimators, such as averages. After computation of all landscapes, for each subject we then obtained its average landscape.

We then quantized the resulting landscapes by sampling r_x values evenly in the interval $[0, t_{max}]$, where t_{max} is a value estimated from the data that corresponds to the last time of the filtration where there were changes in the persistent homology of the complex being processed. This results, for each persistence landscape, in a 2D array of size $r_x \times n_\lambda$, where n_λ is the number of non-zero PL functions in the landscape. See Fig. 3 for an example of one such landscape.

Note that other alternatives to performing vectorization of topological summaries, such as persistence images [2], exist and could be used in place of persistence landscapes, but to compare their performance is beyond the scope of this article.

2.4 Machine Learning Framework

For each subject, we obtained a feature vector of 19,581 gene expression measurements (see Sect. 2.6) and a known class label 0 (161 control subjects) or 1 (264 affected subjects) according to the Parkinson's disease phenotype. We split the data 80-20 into training and test sets, over 50 iterations, except for the computationally more intensive TDA-CNN where we considered 4 iterations after observing the results between iterations were nearly identical.

First we generated a basis of comparison for our TDA approach (TDA-CNN) with standard machine learning algorithms. We used several widely used binary classifiers to train a model and then test its predictions on unseen data: support vector machines with radial basis function kernel (SVM-RBF) and linear kernel (SVM-Linear), as well as random forest (RF) and a simple neural network (MLP-NN, consisting of 3 hidden layers with 20 neurons each, using relu activations). These methods were applied using the scikit-learn python library [23].

In the TDA-CNN approach, for a given resolution r_y, we fed each subject's vectorized persistence landscape as a tensor of shape $(r_x, r_y, 2)$, one channel per homology degree, into a Convolutional Neural Network (CNN) [20] implemented using the Keras library [12] with the Tensorflow [1] backend. We employed a nearest neighbors filter to scale in the y-axis when $n_\lambda \neq r_y$. The architecture we used consisted of two separate paths, one per channel, each consisting of 3 convolutional layers, each with 64 3×3 filters, with Max-Pooling layers [25] of size 2×2 after each convolutional layer. We then fed the outputs of these two paths into a dense layer with 32 neurons. Finally we used a 2-neuron layer with softmax activations [4] as output. All other activations were exponential linear units [14].

2.5 Topological Data Analysis Implementation

We used our own implementation, **maTilDA: Multi-Purpose Toolkit for TDA**, for the construction of the Weighted Vietoris-Rips complex, its persistent homology, barcode computations, persistence landscapes and discrete sampling.

2.6 Gene Expression Data Processing

We downloaded gene expression data derived from RNA sequencing, acquired from blood samples of Parkinson's disease (PD) and control subjects, from the Parkinson's Progression Markers Initiative (https://www.ppmi-info.org/, Phase 1 data). The downloaded sequencing read counts per gene (1,889 samples and 57,820 genes) were examined and outliers removed: 1,141 abundantly expressed RN7S cytoplasmic genes and 742 samples whose read count distributions were different from other samples (having reads assigned to more than 35k genes) were removed.

Table 1. F_1 scores (micro and macro), true positive rate (TPR), and true negative rate (TNR) for predicting Parkinson's disease status from gene expression data. Underlined numbers indicate best values per row.

	TDA-CNN	SVM-Linear	SVM-RBF	Random Forest	MLP-NN
F_1-micro	0.877	0.629	0.641	0.641	0.556
F_1-macro	0.871	0.580	0.477	0.549	0.400
TPR	0.870	0.773	0.960	0.874	0.672
TNR	0.890	0.391	0.112	0.256	0.362

RoDEO [19] was applied (with parameters $P = 20$ bins, $I = 10$ iterations, and $R = 10^8$ reads) to scale the samples to a normalized range of counts $[1, 20]$. For the study in this paper we focused on $20,345$ protein-coding genes per sample. We further removed uninformative genes with a constant RoDEO projected value for each sample, and duplicate genes with identical value distributions, obtaining a final set of $19,581$ genes.

For the purpose of this study, subjects with Parkinson's disease (cohorts PD, GENPD, REGPD) were denoted by phenotype label 1, and subjects without the disease (HC, GENUN, REGUN) by phenotype label 0. We thus had a total of 264 PD samples and 161 unaffected samples (if a subject had several time points sequenced, we took the first one).

3 Results and Discussion

Table 1 shows the F_1 score, true positive rate (TPR), and true negative rate (TNR) for our TDA approach as well as for baseline machine learning methods, on the task of predicting Parkinson's disease diagnosis. We included both micro-F_1 and macro-F_1 scores, the latter gives equal weight to each class to take into account class imbalance [31]. Our TDA with the convolutional neural network (TDA-CNN) approach achieved remarkable improvement of both micro- and macro-F_1 scores compared to the other methods. The TDA-CNN approach achieve scores above 0.87, while SVM (with two different kernel options) and random forest yield similar values to each other, up to 0.64 for micro-F_1 and 0.58 for macro-F_1. Note that convolutional neural networks operate on two-dimensional image data, thus we chose a multilayer perceptron neural network (MLP-NN) as a comparison on the RoDEO-processed gene expression vector data. For this data, the MLP-NN approach seems particularly ill suited, with scores 0.56 and 0.40 for micro- and macro-F_1, respectively.

In terms of true positive and true negative rates, the standard methods have very low TNR, indicating abundant false positives. The TDA-CNN approach, on the other hand achieves a balance between sensitivity and specificity with TPR and TNR having high values (0.87 and above).

The findings indicate blood-based gene expression does contain signal that is relevant for separating subjects with and without Parkinson's disease. Further work includes understanding these subtle signals in order to transform the findings into diagnostic and prognostic models. The introduced framework of topological data analysis with convolutional neural network prediction is a general approach that could be applied to gene expression data relating to other phenotypes.

Acknowledgments. Saugata Basu was partially supported by NSF Grant DMS-1620271.

References

1. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). https://www.tensorflow.org/. Software available from tensorflow.org
2. Adams, H., et al.: Persistence images: a stable vector representation of persistent homology. J. Mach. Learn. Res. **18**(8), 1–35 (2017). http://jmlr.org/papers/v18/16-337.html
3. Arsuaga, J., Borrman, T., Cavalcante, R., Gonzalez, G., Park, C.: Identification of copy number aberrations in breast cancer subtypes using persistence topology. Microarrays **4**(3), 339–369 (2015)
4. Bridle, J.S.: Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In: Soulié, F.F., Hérault, J. (eds.) Neurocomputing. NATO ASI Series (Series F: Computer and Systems Sciences), vol. 68, pp. 227–236. Springer, Heidelberg (1990). https://doi.org/10.1007/978-3-642-76153-9_28
5. Bubenik, P.: Statistical topological data analysis using persistence landscapes. J. Mach. Learn. Res. **16**(1), 77–102 (2015). http://dl.acm.org/citation.cfm?id=2789272.2789275
6. Buchet, M., Chazal, F., Oudot, S.Y., Sheehy, D.R.: Efficient and robust persistent homology for measures. Comput. Geom. Theory Appl. **58**(C), 70–96 (2016). https://doi.org/10.1016/j.comgeo.2016.07.001
7. Camara, P.: Topological methods for genomics: present and future directions. Curr. Opin. Syst. Biol., 95–101 (2017). https://doi.org/10.1016/j.coisb.2016.12.007
8. Cang, Z., Mu, L., Wu, K., Opron, K., Xia, K., Wei, G.W.: A topological approach for protein classification. Comput. Math. Biophys. **3**(1), 140–162 (2015). https://doi.org/10.1515/mlbmb-2015-0009
9. Carlsson, G., Zomorodian, A., Collins, A., Guibas, L.: Persistence barcodes for shapes. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing. SGP 2004, pp. 124–135. ACM, New York (2004). https://doi.org/10.1145/1057432.1057449
10. Chahine, L.M., Stern, M.B., Chen-Plotkin, A.: Blood-based biomarkers for Parkinson's disease. Parkinsonism Relat. Disord. **20**(S1), S99–S103 (2014)
11. Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., Wasserman, L.: Subsampling methods for persistent homology. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2143–2151. PMLR, Lille, France, 07–09 July 2015. http://proceedings.mlr.press/v37/chazal15.html
12. Chollet, F., et al.: Keras (2015). https://keras.io
13. Chung, M.K., Bubenik, P., Kim, P.T.: Persistence diagrams of cortical surface data. In: Prince, J.L., Pham, D.L., Myers, K.J. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 386–397. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02498-6_32
14. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv:1511.07289 (2015)

15. Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. In: Proceedings of the Twenty-first Annual Symposium on Computational Geometry. SCG 2005, pp. 263–271. ACM, New York (2005). https://doi.org/10.1145/1064092.1064133
16. van Dam, S., Võsa, U., van der Graaf, A., Franke, L., de Magalhães, J.P.: Gene co-expression analysis for functional classification and gene-disease predictions. Brief. Bioinform. **19**(4), 575–592 (2017). https://doi.org/10.1093/bib/bbw139
17. Dey, T., Mandal, S.: Protein classification with improved topological data analysis. In: 18th International Workshop on Algorithms in Bioinformatics (WABI 2018). Leibniz International Proceedings in Bioinformatics (2018)
18. Duman, A.N., Pirim, H.: Gene coexpression network comparison via persistent homology. Int. J. Genomics **2018**, Article ID 7329576, 1–11 (2018). https://doi.org/10.1155/2018/7329576
19. Haiminen, N., et al.: Comparative exomics of Phalaris cultivars under salt stress. BMC Genomics (Suppl 6), S18 (2014). https://doi.org/10.1186/1471-2164-15-S6-S18
20. Le Cun, Y., et al.: Handwritten digit recognition with a back-propagation network. In: Proceedings of the 2nd International Conference on Neural Information Processing Systems. NIPS 1989, pp. 396–404. MIT Press, Cambridge (1989). http://dl.acm.org/citation.cfm?id=2969830.2969879
21. Nicolau, M., Levine, A.J., Carlsson, G.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proc. Natl. Acad. Sci. **108**(17), 7265–7270 (2011). https://doi.org/10.1073/pnas.1102826108
22. Parnetti, L., et al.: CSF and blood biomarkers for Parkinson's disease. Lancet Neurol. **18**(6), 573–586 (2019)
23. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
24. Pike, J.A., et al.: Topological data analysis quantifies biological nano-structure from single molecule localization microscopy. bioRxiv (2018). https://doi.org/10.1101/400275
25. Ranzato, M., Huang, F.J., Boureau, Y., LeCun, Y.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2007. https://doi.org/10.1109/CVPR.2007.383157
26. Sauerwald, N., Shen, Y., Kingsford, C.: Topological data analysis reveals principles of chromosome structure throughout cellular differentiation. bioRxiv (2019). https://doi.org/10.1101/540716
27. Schofield, J.P.R., et al.: A topological data analysis network model of asthma based on blood gene expression profiles. bioRxiv (2019). https://doi.org/10.1101/516328
28. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. Ann. Stat. **35**(6), 2769–2794 (2007). https://doi.org/10.1214/009053607000000505
29. Turner, K., Mukherjee, S., Boyer, D.M.: Persistent homology transform for modeling shapes and surfaces. Inf. Infer. **3**(4), 310–344 (2014)
30. Wang, C., Chen, L., Yang, Y., Zhang, M., Wong, G.: Identification of potential blood biomarkers for Parkinson's disease by gene expression and DNA methylation data integration analysis. Clin. Epigenetics **11**, 24 (2019)
31. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, 15–19 August 1999, pp. 42–49. ACM (1999)