Computers in Biology and Medicine 169 (2024) 107969

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# Novel dimensionality reduction method, Taelcore, enhances lung transplantation risk prediction

Fatma Gouiaa [a], Kelly L. Vomo-Donfack [a], Alexy Tran-Dinh [b,c], Ian Morilla [a,d,*]

[a] Université Sorbonne Paris Nord, LAGA, CNRS, UMR 7539, Laboratoire d'excellence Inflamex, Villetaneuse, France
[b] Université Paris Cité, AP-HP, Hôpital Bichat Claude Bernard, Département d'anesthésie-Réanimation, INSERM, Paris, France
[c] Universié Paris Cité, LVTS, Inserm U1148, F-75018 Paris, France
[d] University of Malaga, Department of Genetics, MLiMO, 29010, Málaga, Spain

ARTICLE INFO

*Keywords:*
Acute cellular rejection
Topological data analysis
Machine learning
Predictor of ACR risk

ABSTRACT

In this work, we present a new approach to predict the risk of acute cellular rejection (ACR) after lung transplantation by using machine learning algorithms, such as Multilayer Perceptron (MLP) or Autoencoder (AE), and combining them with topological data analysis (TDA) tools. Our proposed method, named topological autoencoder with best linear combination for optimal reduction of embeddings (Taelcore), effectively reduces the dimensionality of high-dimensional datasets and yields better results compared to other models. We validate the effectiveness of Taelcore in reducing the prediction error rate on four datasets. Furthermore, we demonstrate that Taelcore's topological improvements have a positive effect on the majority of the machine learning algorithms used. By providing a new way to diagnose patients and detect complications early, this work contributes to improved clinical outcomes in lung transplantation.

## 1. Introduction

Lung transplantation is a challenging medical procedure that is often performed on patients with severe lung disease who have significantly reduced quality of life [1–4]. Unfortunately, the risk of mortality is high, with only two-thirds of patients surviving and a mere 50% surviving after three years, mainly due to infections and acute rejection [5]. Graft rejection is a significant concern after surgery and throughout a patient's life, and immunosuppressive drugs are necessary to mitigate this risk [6]. However, these drugs can increase the risk of infection and subsequent rejection [7]. The occurrence of complications can be identified by medical symptoms such as fever, changes in body weight, or deteriorating lung function [8].

In recent years, machine learning and topological data analysis have shown tremendous potential to impact the healthcare field and improve patient outcomes [9–14]. In this context, the prediction of acute cellular rejection (ACR) after lung transplantation is a crucial medical challenge that requires innovative approaches to manage complications and mitigate risks.

Traditional statistical methods have been used to analyse patient data, but machine learning and topological data analysis represent a paradigm shift in the field of medical research. These techniques offer a more comprehensive and systematic approach to understanding complex medical phenomena, such as ACR, by integrating high-dimensional data from multiple sources. By leveraging these approaches, we can improve our ability to diagnose and predict ACR, providing clinicians with more precise and effective treatments.

To address these challenges, we propose to use topological autoencoder with best linear combination for optimal reduction of embeddings (Taelcore). Taelcore is a precise and flexible classifier that utilises a novel technique of dimensionality reduction. By combining topological data analysis [15] and machine learning autoencoders [16], we have transformed information obtained from all parameter values into a comprehensible and easy-to-represent form that provides hospital clinicians with an effective starting point to manage complications after a transplant.

In this study, we recycle the dataset from "Personalised Risk Predictor for Acute Cellular Rejection in Lung Transplant Using Soluble CD31" [17] to improve its results by combining topological data analysis and machine learning. We harness this combined classifier to introduce a novel dimensionality reduction method named Taelcore that makes the improvement possible.

We systematically examine and evaluate our dataset with other studies, using topological data analysis and machine learning to visualise and simplify the input structure of our learning model. Our clinical dataset comprised 40 patients who underwent lung transplant surgery, and 7 of them had acute cellular rejection after one year. We

* Corresponding author at: University of Malaga, Department of Genetics, MLiMO, 29010, Málaga, Spain.
*E-mail address:* morilla@math.univ-paris13.fr (I. Morilla).

https://doi.org/10.1016/j.compbiomed.2024.107969
Received 22 August 2023; Received in revised form 23 December 2023; Accepted 2 January 2024
Available online 4 January 2024

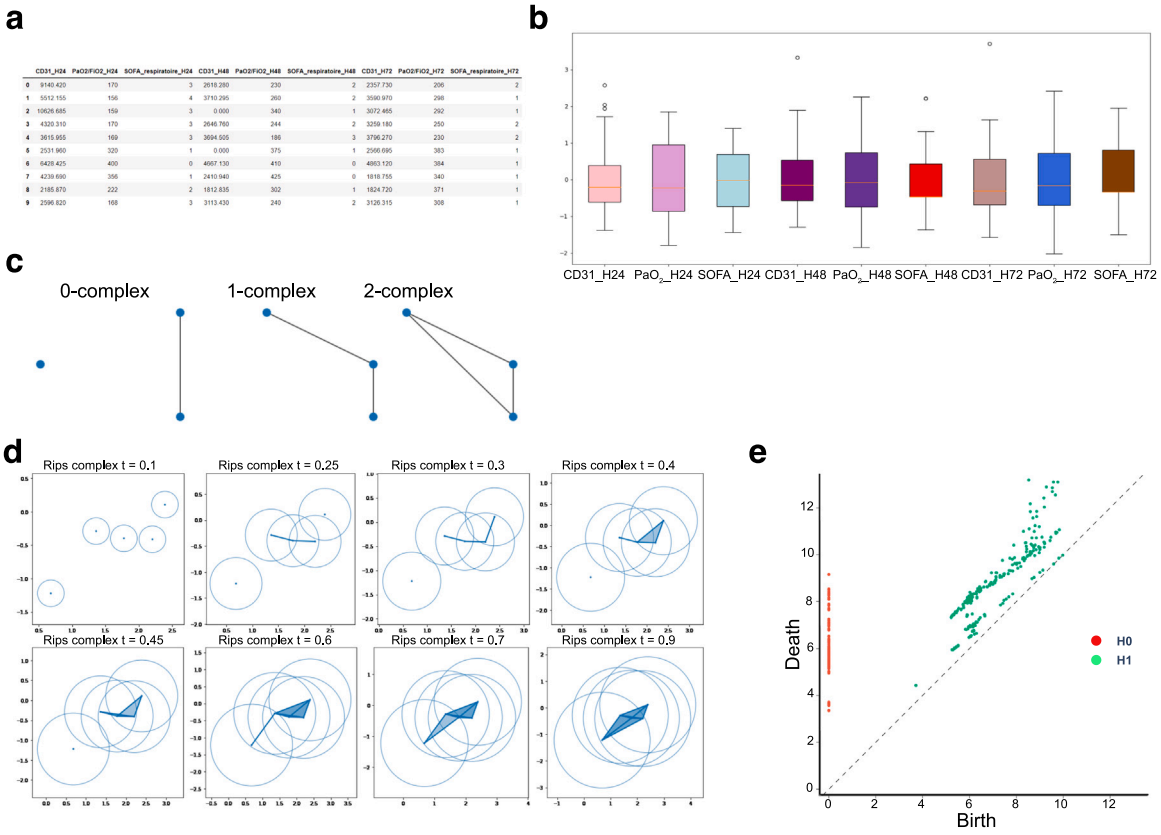


**Fig. 1. Data Denoising and Topological Transformers.** This figure provides an overview of the data preprocessing and topological transformation steps employed in our study. (a) The dataset required standardisation to address variations in feature dimensions and scales. (b) Outlier Detection and Denoising: The box plot method was utilised to visualise the variables and detect outliers. Moreover, the relationship between sCD31 levels and the risk of acute cellular rejection (ACR) was leveraged. (c) The creation of simplicial complexes by connecting vertices. (d) Rips Filtration: A Rips filtration was applied to capture the dataset's topological characteristics. (e) Persistence Diagram Formation: Persistence diagrams were constructed to classify patients into two groups based on their ACR risk. The persistence diagrams provided insights into the topological characteristics relevant to the presence or absence of ACR risk.

measured nine characteristics over the first three days after transplantation for each patient, which indicate complications after surgery. These characteristics include $CD31$ H24, $CD31$ H48, and $CD31$ H72, which predict acute rejection, and $PaO_2/FiO_2$ H24, $PaO_2/FiO_2$ H48, and $PaO_2/FiO_2$ H72, which measure the percentage of oxygen in the gas mixture that the person breathes after 24, 48, and 72 h of the operation. Additionally, we used the $SOFA$ respiratory H24, $SOFA$ respiratory H48, and $SOFA$ respiratory H72 scores to assess the risk of ACR in patients who cited CD31 [17,18].

In summary, the prediction of ACR after lung transplantation is a critical medical challenge that requires innovative approaches to manage complications and mitigate risks. Machine learning and topological data analysis represent powerful tools for achieving this goal, providing clinicians with the ability to diagnose and predict ACR more accurately, ultimately improving patient outcomes. Our study presents a novel approach to addressing this challenge by introducing a combined classifier and a new dimensionality reduction method that can enhance the accuracy and effectiveness of ACR prediction.

## Results

Thus, we outline the approach taken to predict the ACR risk score of patients. Firstly, the available data will be thoroughly examined and standardised to ensure a fair and effective modelling process. Next, a Rips filtration will be used to extract topological characteristics, and entropy will be utilised to vectorise the persistence diagram. With this information, the power of TDA tools will be leveraged to improve the performance of machine learning models. The evaluation of the models will then be compared before and after implementing the TDA approach, with reference to studies [17,19]. The aim is to demonstrate the effectiveness of the approach in accurately predicting ACR risk scores for patients, with potential applications in clinical settings.

### 1.1. Data denoising and topological transformers

To begin the data preprocessing step, exploratory data analysis was performed to gain insights into the dataset and identify any potential outliers and interesting relationships between variables. As seen in Fig. 1a, we examined the distribution of the dataset and determined that it required standardisation to scale the features with different dimensions and scales. The box plot method was then used to visualise the variables through their quartiles, and detect outliers as seen in Fig. 1b. This step allowed us to minimise the presence of outliers in the dataset and denoise the data projection using the relationship between sCD31 levels and the risk of ACR. As stated in [17], higher levels of sCD31 were associated with a lower risk of ACR, indicating that sCD31 can be a useful biomarker for predicting ACR risk. Therefore, we used a threshold based on the sCD31 level to filter out data points that are likely to have a higher risk of ACR. By doing so, the resulting dataset projection had a reduced noise level, which improved the performance of downstream analyses of classification, as shown in Fig. 1b. These preprocessing steps were necessary to ensure the accuracy and reliability of our machine learning models and topological data analysis.

Next, a Rips filtration [20] was created to extract the topological characteristics of the dataset (Fig. 1c). The Rips filtration (see methods) involves creating balls around each point and observing the moment

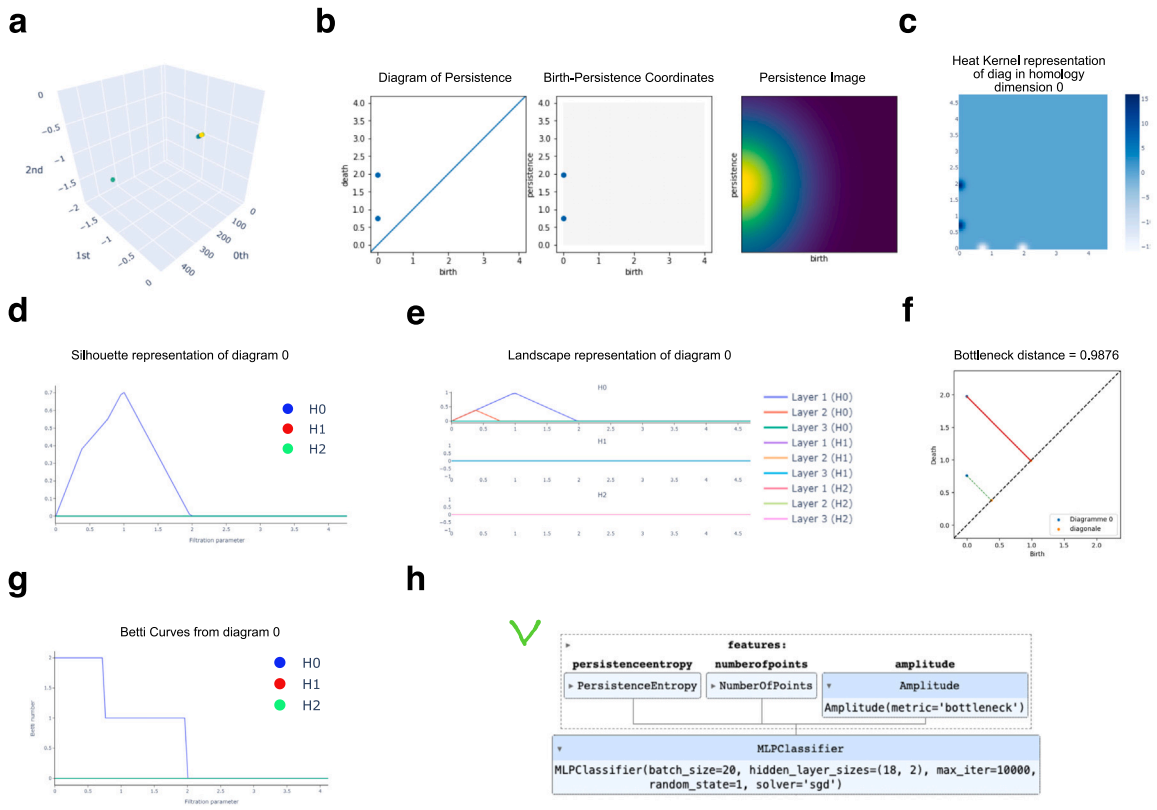


**Fig. 2. Vectorisation, Distance Calculation, and Classification.** (a) Vectorisation Using Entropy: To facilitate the application of machine learning algorithms, the persistence diagram was vectorised using entropy. (b–g) $L^p$ Distance Calculation: Various methods were employed to calculate the $L^p$ distance between the smoothed Gaussian diagrams. (h) Combination of Steps and Classification: The final step involved integrating the previous stages, which included extracting topological characteristics, entropy calculation, and topological enhancement, with a classifier.

**Table 1**
Evaluation of different classification algorithms before and upon topological preservation metric application.

| | RF | MLP | KNN | GNB |
|---|---|---|---|---|
| Percentage of success before topological improvement | 50 | 80 | 50 | 90 |
| Best combinations | Wasserstein, Heat, Persistence image | Bottleneck, Wasserstein, betti, Landscape, Persistence image | Persistence image | Persistence image |
| Percentage of success after topological improvement | 80 | 90 | 50 | 90 |

when nascent or dying topological characteristics appear (see Fig. 1d). Persistence diagrams were then formed to classify patients into two groups — those who have a risk of ACR and those who do not have a risk of ACR (see Fig. 1e). To vectorise the persistence diagram and apply machine learning algorithms, entropy was used (see Fig. 2a).

Different methods were used to calculate the $L^p$ distance between the smoothed Gaussian diagrams, such as silhouette [21], heat kernel, and persistence image [22]. These methods allowed for the comparison of persistence diagrams and the application of machine learning algorithms to predict the ACR risk score of patients accurately (see Fig. 2b–g).

The final step involved combining the previous steps, including extracting topological characteristics, entropy, and topological improvement, with a classifier (see Fig. 2f). By making combinations between the different metrics of the amplitude function, the results obtained from the best combination were compared with the results of the classification without topological features as shown in Table 1.

The table provides a comparison of the performance of several machine learning algorithms (see methods) that share similar assumptions and feature engineering models, including Random Forest (RF) [23], Multi-Layer Perceptron (MLP) [24], k-Nearest Neighbours (KNN) [25], and Gaussian Naive Bayes (GNB) [26]. The comparison was conducted on our lung transplantation dataset before and after applying a topological improvement technique, which is described in the methods section. By evaluating the performance of these algorithms before and after applying the topological improvement technique, we can assess the effectiveness of this technique in enhancing the accuracy of the predictions.

Before topological improvement, the percentage of success varies between the algorithms. GNB achieved the highest success rate at 90%, followed by MLP at 80%, while RF and KNN had a success rate of 50%.

After applying the topological improvement technique, MLP and GNB remained the top-performing algorithms, with success rates of 90% and 90%, respectively. RF and KNN did not improve, with RF remaining at 50% and KNN dropping to 50%. Possible reasons for the limited improvement in KNN and GNB could be attributed to the inherent characteristics of these algorithms. KNN heavily relies on local

proximity, and if the topological features introduced do not align optimally with the underlying patterns in the data, it may not experience significant enhancement. Similarly, the probabilistic assumptions made by GNB may not benefit substantially from topological improvements when the data distribution fails to align with the model's assumptions. Further investigation into the specific characteristics of the dataset and the algorithms themselves may provide insights into these observations.

In terms of the topological improvement techniques used, the best combinations varied for each algorithm. MLP achieved its highest success rate using bottleneck, Wasserstein, Betti, landscape [27], and persistence image techniques. GNB had its best success rate with only persistence image techniques. RF and KNN had their best results with Wasserstein, heat, and persistence image techniques.

Given the higher success rate of the MLP model and its greater interpretability, it seems like a good choice for downstream calculations, which involve dimensionality reduction using a topological autoencoder [28]. In that case, the MLP model may be a good choice as it could provide more insight into the learned representations of the data, which could be useful in understanding the performance of the autoencoder.

MLP models have been used in various ways for dimensionality reduction, including as part of autoencoders. The learned representations of an autoencoder can often be difficult to interpret, but having an MLP model as part of the architecture could potentially make the learned representations more interpretable. Of course, it is important to carefully design and train the autoencoder to ensure that it performs well on the downstream task. But having a good starting point, such as a high-performing MLP model, could be beneficial.

### 1.2. Dimensionality reduction: Topology-preserving autoencoder

This chapter presents a new generic framework that we developed during this project to reduce the dimension of a dataset using topological preservation by combining topological structures (Taelcore), measured via persistent homology, of the data space in autoencoder latent encodings. Fig. 3 illustrates an overview of our proposed method, and subsequent sections will provide further details about the individual steps involved.

*Motivation*

Starting from the topological autoencoder introduced in [29], we created a new method for dimensionality reduction using the tools presented in methods (see Fig. 3a, b). The idea is to add a regularisation term to the loss function of an autoencoder (see Fig. 3c). However, instead of using the bottleneck distance [30] between the input space and the latent space as indicated in [29], we used the amplitude with the best topological combination found in Table 1 to improve machine learning methods (see Fig. 3d, e). We worked with the best combination of MLP: $\{bottleneck, Wasserstein, Betti, landscape, persistence\ image\}$ (see Fig. 3c). Let $X$ be the input set, $Z = h(X)$ the latent space, and $Y = g(h(X))$ the output of the autoencoder. Then we define the loss function as:

$$l = l(X, Y) + \lambda l_t, \tag{1}$$

where $\lambda$ is a parameter to control the strength of regularisation and $l_t$ is the topological loss function defined as

$$l_t = l_{X\to Z} + l_{Z\to X}$$

with

$$l_{X\to Z} = \frac{1}{2}|a-b|^2$$

and

$$l_{Z\to X} = \frac{1}{2}|b-c|^2$$

$$a = \mathrm{lct}_1(D^X) + \cdots + \mathrm{lct}_5(D^X), \quad b = \mathrm{lct}_1(D^Z) + \cdots + \mathrm{lct}_5(D^Z),$$

$$c = \mathrm{lct}_1(D^Y) + \cdots + \mathrm{lct}_8(D^Y)$$

$D^X$, $D^Z$, and $D^Y$ are respectively the persistence diagrams of $X$, $Z$, and $Y$. $l_{ct_i}(D^X)$ refers to the amplitude of $D^X$ with the $i$th topological feature of the best MLP combination as a metric (see Fig. 3d). As a consequence of [29], the loss of Taelcore is differentiable for each update step during training. The code associated with this study is publicly available (https://github.com/MorillaLab/Taelcore/).

Although persistence diagrams are known to be stable with respect to small perturbations of the underlying space [29], we still need to analyse our topological approximation on the level of mini-batches. To ensure the accuracy of our subsampled persistence diagrams, we rely on the following theorem, which guarantees that they are close to the persistence diagrams of the original point cloud.

**Theorem 1.** *Let $X$ be a point cloud of cardinality $N$, and let $X^{(r)} \subseteq X$ be any $r$-batch. For any $\epsilon, \widetilde{\epsilon} > 0$ with $\widetilde{\epsilon} > \epsilon$, there exists a finite constant $C \sim P\left[\sum_{i=1}^m \alpha_i \cdot d_H(X_i, X_i^{(r)}) > \widetilde{\epsilon}\right]$ such that $P\left[\left(\sum_{i=1}^m \alpha_i \cdot lct_i(D^{X_i}, D^{X_i^{(r)}})\right) > \epsilon\right] \le C$, where $lct_i(D^X)$ denotes the persistence amplitude of the $i$th topological characteristic in the persistence diagram $D^X$ exceeding a threshold, and the coefficients $\alpha_i$ are real numbers. Furthermore, $m$ is the number of topological characteristics used in the linear combination, and $C$ may depend on the choice of coefficients $\alpha_i$ and the dimension of $X$.*

**Proof.** Let $lct_i(D^{X_i})$ be the persistence amplitude of the $i$th topological characteristic in the persistence diagram $D^{X_i}$. Consider the stability result $db(D^{X_i}, D^{X_i^{(r)}}) \le 2dH(X_i, X_i^{(r)})$ established by Chazal et al. (2014) [29,31] for finite metric spaces.

By the definition of Hausdorff distance:

$$dH(X_i, X_i^{(r)}) = \max\left\{\sup_{x\in X_i}\inf_{y\in X_i^{(r)}} \mathrm{dist}(x,y), \sup_{y\in X_i^{(r)}}\inf_{x\in X_i} \mathrm{dist}(x,y)\right\}.$$

We establish $lct_i(D^{X_i}) \le dH(X_i, X_i^{(r)})$ for each $i$.

For any point $p \in X_i$, let $q$ be its closest point in $X_i^{(r)}$ (according to the metric used for $dH$). By the definition of the persistence amplitude, $lct_i(D^{X_i})$ measures the "width" of the persistence diagram, indicating how long a topological feature persists. Since $q$ is the closest point in $X_i^{(r)}$ to $p$, it follows that the persistence diagram $D^{X_i}$ must include the persistence of the corresponding topological feature up to the distance between $p$ and $q$.

Summing over $m$ topological characteristics:

$$\sum_{i=1}^m \alpha_i \cdot lct_i(D^{X_i}) \le \sum_{i=1}^m \alpha_i \cdot d_H(X_i, X_i^{(r)}).$$

Taking probabilities:

$$P\left[\left(\sum_{i=1}^m \alpha_i \cdot lct_i(D^{X_i}, D^{X_i^{(r)}})\right) > \epsilon\right] \le P\left[\sum_{i=1}^m \alpha_i \cdot d_H(X_i, X_i^{(r)}) > 2\varepsilon\right]. \square$$

Table 2 shows the results obtained on several datasets. For the Iris dataset, Taelcore consistently achieves remarkable performance across most metrics. It outperforms other methods with significantly lower values of KL divergence (−37.86, −3.98, −0.40) for $KL_{0.01}$, $KL_{0.1}$, and $KL_1$, respectively, indicating better preservation of the high-dimensional structure. Although PCA had the best performance in terms of RMSE with a value of 0.0036, and TRUST with a value of 0.993, Taelcore also demonstrates a competitive RMSE (0.007) and TRUST score (0.973), highlighting its ability to accurately represent the dataset while preserving its topological characteristics. Furthermore, Taelcore achieves the lowest MSE (0.052), indicating superior overall performance compared to other methods.

Similarly, on the Shapes dataset, Taelcore excels in preserving the high-dimensional structure, as evidenced by the lowest KL divergence values $(0.099, 0.010, 0.001)$. It achieves a relatively low RMSE (0.002) and a respectable TRUST score (0.865). However, PCA had the best

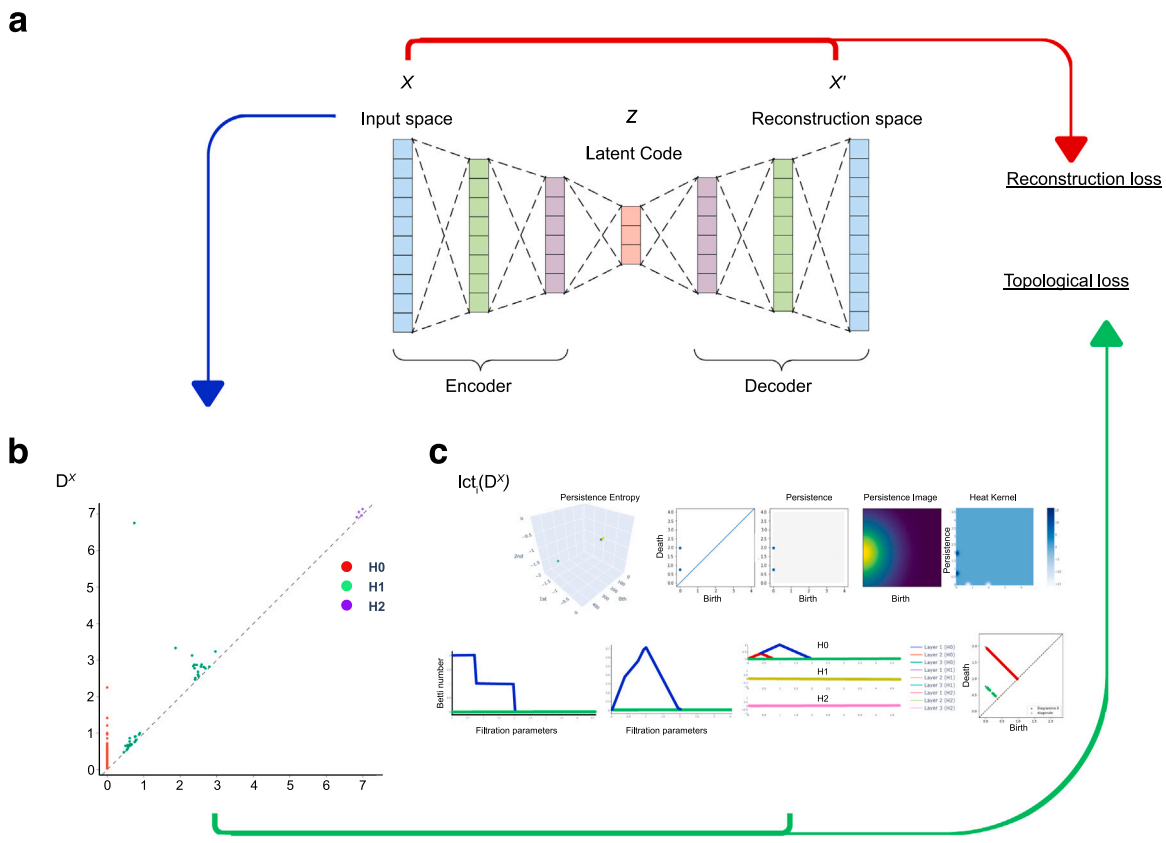


**Fig. 3. Overview of the Proposed Method for Topological Dimensionality Reduction.** Along with the usual reconstruction loss, we calculate our topological loss based on the topological differences between the best linear combination for optimal reduction of embeddings (Taelcore) from persistence diagrams, i.e., topological feature descriptors, calculated on the mini-batch $X$ and its corresponding latent code $Z$. The topological loss term is designed to constrain the autoencoder to preserve topological features in the data space within the latent representations. (a) and (b) Illustration of the topological structures and methods used in the proposed framework. (c) Autoencoder architecture with the addition of a regularisation term and the best topological combination. The computation of the topological loss function, denoted as $l_t$, is derived from (a), utilising the persistence diagrams $D^X$, $D^Z$, and $D^Y$ of $X$, $Z$, and $Y$, respectively. The amplitudes correspond to the $i$th topological feature of the best combination used in MLP. From (b) and (c), we derived the training process, showing the different steps involved in optimising the loss function and preserving topological information.

performance in terms of RMSE with a value of 0.001, while T-sne had the best TRUST score with a value of 0.908. In terms of MSE, Taelcore performs favourably with a value of 0.014.

Regarding the ACR dataset, Taelcore once again demonstrates its strength in preserving the high-dimensional structure. It achieves the lowest KL divergence values $(-138.563, -15.247, -0.208)$, signifying its ability to capture and retain the essential topological information. PCA had the best performance in terms of RMSE with a value of 0.035, and TRUST with a value of 0.939. Additionally, Taelcore achieves a relatively low RMSE (0.038) and a competitive TRUST score (0.946). It also achieves the best MSE value among the methods (0.307).

In the dataset related to Chronic Pain, Taelcore demonstrated a remarkable $KL_{0.01}$ value of $-126.053$ at the 0.01 quantile, indicating its ability to capture fine-grained details in the data distribution. The performance in $KL_{0.1}$ was 2.194 at 0.1 quantile, reflecting a reasonable preservation of the underlying structure at a broader scale. However, the $KL_1$ value of 0.147 at the 1 quantile, while suggesting a slight deviation from the true data distribution, remains the best performance in its category. In terms of RMSE, Taelcore achieved a value of 0.032, indicating its capability to reconstruct the input data with reasonable accuracy. The TRUST score of 0.869 signifies a good preservation of topological features. Additionally, the MSE value of 0.317 indicates an overall effective performance.

It is noteworthy that T-sne and Umap methods generated embeddings with non-Euclidean distance metrics, rendering the application of MSE not meaningful in these cases, as denoted by "Non" in the corresponding cells.

While Taelcore consistently demonstrates strong performance in terms of preserving high-dimensional structure across multiple datasets, it is important to consider the specific goals and characteristics of the data when selecting the most appropriate method. However, if high-dimensional structure preservation is a priority, Taelcore proves to be a promising choice based on its consistent performance across different metrics and datasets.

In summary, our model has demonstrated its ability to preserve the high-dimensional structure of the data density across multiple scales ($\sigma$), as evidenced by the results in the $(KL_{0.01}, KL_{0.1}, KL_1)$ columns of Table 2. Additionally, our model has been successful in accurately reconstructing the input, as demonstrated in the MSE column. For our dataset on evaluating the risk of ACR after lung transplantation, our method stands out as the only one capable of preserving the relevant structural information. In the case of Shapes, while the autoencoder exhibits the lowest reconstruction error, our method possesses the additional objective of capturing the underlying data structure. As for Iris, PCA appears to be preferred by other evaluation measures, such as RMSE and TRUST, despite not being able to capture the data structure. Similar trends are observed in our analysis of other datasets, wherein conventional evaluation measures favour basic methods (T-sne, Umap, PCA, AE) that fail to detect the relevant structural information (see Fig. 4).

*Visualising the latent space*

Based on the analysis of Fig. 4, our method showcases an intriguing and interpretable representation in the latent space. It successfully organises the classes in a spatially meaningful manner, offering valuable insights into the data. For the Shapes dataset, our method not only effectively preserves the underlying structure but also outperforms the

**Table 2**
Empirical validation of different methods (Taelcore, Autoencoder, PCA, T-sne, and Umap) on four datasets (Iris, Shapes, ACR, and Pain). The validation is based on six evaluation metrics: KL divergence at three different quantiles (0.01, 0.1, and 1), root mean square error (RMSE), TRUST, and mean squared error (MSE). The table shows the values of these evaluation metrics for each method on each dataset. The best performing method for each metric on each dataset is highlighted in bold, and the second-best performing method is underlined. "None" indicates that the corresponding method did not provide results for that particular evaluation metric: namely, T-sne and Umap generated embeddings with non Euclidean distance metrics. Hence, applying MSE is not meaningful.

| | Method | $KL_{0.01}$ | $KL_{0.1}$ | $KL_{1}$ | RMSE | TRUST | MSE |
|---|---|---|---|---|---|---|---|
| **Iris** | Taelcore | **-22.644** | **-2.381** | **-0.239** | 0.0077 | 0.9729 | **0.052** |
| | Autoencoder | 24.839 | 2.722 | 0.274 | 0.0075 | 0.9690 | **0.054** |
| | PCA | -5.933 | -0.611 | -0.061 | **0.0035** | **0.9932** | 0.125 |
| | T-sne | **-37.860** | **-3.983** | **-0.400** | **0.0072** | **0.9934** | None |
| | Umap | 223.774 | 25.769 | 2.614 | 0.0427 | 0.9546 | None |
| **Shapes** | Taelcore | **0.099** | **0.010** | **0.001** | 0.0021 | 0.8657 | 0.014 |
| | Autoencoder | **0.258** | **0.026** | **0.002** | **0.0020** | 0.9071 | **0.011** |
| | PCA | 1.423 | 0.143 | 0.014 | **0.0016** | **0.9078** | **0.013** |
| | T-sne | 3.365 | 0.338 | 0.033 | 0.0022 | **0.9995** | None |
| | Umap | 2.609 | 0.262 | 0.026 | 0.0043 | 0.5661 | None |
| **ACR** | Taelcore | **-138.563** | **-15.247** | **-0.208** | 0.0385 | 0.9464 | **0.307** |
| | Autoencoder | -116.481 | -13.422 | -0.189 | **0.0347** | **0.9467** | 0.309 |
| | PCA | -120.752 | **-13.718** | **-0.191** | **0.0352** | 0.9392 | **0.299** |
| | T-sne | -8.144 | -3.398 | -0.060 | 0.0702 | 0.5743 | None |
| | Umap | **-134.192** | -13.001 | -0.171 | 0.0436 | **0.9504** | None |
| **Pain** | Taelcore | **-126.053** | -2.194 | **0.147** | 0.0320 | 0.8697 | **0.317** |
| | Autoencoder | **-118.876** | -4.625 | -0.222 | 0.0306 | **0.8563** | 0.319 |
| | PCA | -118.028 | -3.329 | -0.051 | 0.0304 | 0.8693 | **0.303** |
| | T-sne | -99.575 | **-6.511** | -0.558 | **0.0280** | **0.9321** | None |
| | Umap | -114.117 | **-10.783** | **-1.034** | **0.0297** | **0.8728** | None |

autoencoder in terms of separation. Similarly, when applied to the Iris dataset, our method achieves a notably more linear separation compared to the other methods. Moreover, in the context of the lung transplantation dataset, our method demonstrates a significant linear separation between the ACR and non-ACR groups of patients, with a relatively lower degree of sparsity. In the case of Pain dataset, our method is able to capture fine-grained structural details between patient's classes. These findings emphasise the efficacy of our method in preserving the high-dimensional structure while concurrently enhancing separability across different datasets.

## Discussion

Taelcore's novel combination of persistent homology computation and autoencoder back-propagation offers a unique way to preserve topological information in high-dimensional datasets. While traditional methods for reducing the dimensionality of such datasets often lose critical topological information, Taelcore's approach allows for the identification and preservation of topological features that may be important for accurate data analysis and prediction. Taelcore uses machine learning algorithms, such as Multilayer Perceptron (MLP) [24] and Autoencoder (AE), and combines them with tools of topological data analysis to yield improved results.

In addition to its application in predicting the risk of ACR following lung transplantation (i.e., an accuracy achieved of 90%), Taelcore has potential applications in a variety of fields, such as image analysis and natural language processing. By reducing the dimensionality of complex data while preserving its topological features, Taelcore offers a valuable tool for researchers and data analysts who seek to better understand complex datasets and make accurate predictions.

One of the key advantages of Taelcore is its ability to reduce dimensionality more effectively than other models while still preserving topological information. This can lead to more accurate predictions and insights from high-dimensional data, which can be particularly valuable in medical applications. Taelcore's effectiveness in predicting the risk of ACR following lung transplantation can contribute to improved clinical outcomes, and future research can continue to explore its potential to make a difference in other medical applications. Furthermore, we have shown that our method is theoretically sound. On several datasets, we have observed that the regularisation term added to the loss function of the autoencoder improved the performance in terms of various quality indicators, such as density preservation, without affecting the reconstruction error.

Overall, Taelcore represents a significant contribution to the field of topological data analysis, and its potential applications are wide-ranging. As more researchers begin to adopt Taelcore and explore its potential, we can expect to see further advancements in the field of data analysis and prediction.

## Methods

In this section, we provide an overview of the methodologies employed in this work, namely topological data analysis (TDA) and machine learning. We introduce cutting-edge techniques for analysing high-dimensional data, including unsupervised learning through autoencoders, linear dimensionality reduction via PCA [32], and non-linear dimensionality reduction using t-SNE and Umap [33]. These powerful techniques are leveraged and compared to our topological autoencoder (Taelcore) to explore and visualise complex data structures.

### *1.3. Datasets*

*Lung transplantation*

We analysed a clinical dataset of 40 patients who underwent lung transplantation surgery, with the goal of identifying ACR a potential complication. Among the patients, seven (17.5%) experienced acute cellular rejection within one year of the operation.

The following nine characteristics were measured for each patient during the first three days following lung transplantation surgery, to assess the occurrence of post-surgery complications:

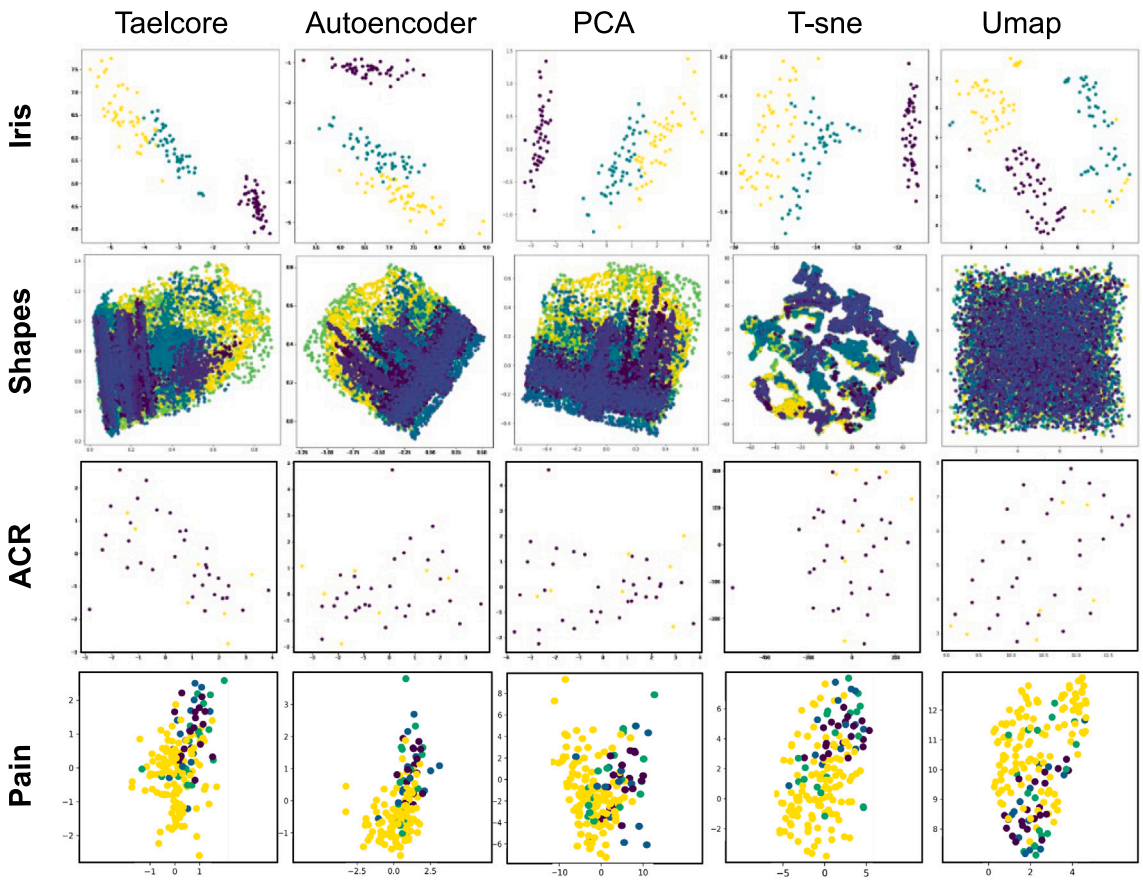


**Fig. 4. Comparison of Latent Space Representations.** A comprehensive comparison of the latent space representations obtained by our proposed method and several baseline techniques, namely Autoencoder, PCA, t-SNE, and UMAP, across different datasets. (a) Shapes Dataset: The latent space representations of the Shapes dataset are depicted for our method, Autoencoder, PCA, t-SNE, and UMAP. Our method not only effectively preserves the intrinsic structure but also outperforms the other techniques in terms of separation. This is evident from the distinct and well-separated clusters observed in our method's representation. (b) Iris Dataset: The latent space representations of the Iris dataset are presented for our method, Autoencoder, PCA, t-SNE, and UMAP. Our method demonstrates a notably more linear separation between the classes compared to the other methods. This indicates its efficacy in capturing the essential discriminative information present in the dataset. (c) Lung Transplantation Dataset: The latent space representations of the lung transplantation dataset are illustrated for our method, Autoencoder, PCA, t-SNE, and UMAP. Our method exhibits a significant linear separation between the ACR and non-ACR groups of patients while maintaining a lower degree of sparsity compared to the alternative techniques. (d) Pain Dataset: The latent space representations of the Pain dataset are depicted for our method, Autoencoder, PCA, t-SNE, and UMAP. Our method captures fine-grained details with distinct and well-defined clusters, showcasing its effectiveness.

- $CD31$ H24, H48, H72: These measurements were taken to assess the predictive role of CD31 protein in acute rejection. CD31 is a protein that controls the balance between blood and vessels from within the endothelial cells. When this balance does not work, the function of CD31 as a "peacemaker" is lost, and the cells begin to activate inappropriately, potentially leading to unsuccessful transplantation.
- $PaO_2/FiO_2$ H24, H48, H72: These measurements indicate the percentage of oxygen present in the gas mixture that the patient breathes after 24, 48, and 72 h of surgery.
- $SOFA$ respiratory H24, H48, H72: A score greater than or equal to two on the SOFA respiratory measurement indicates a 10% risk of mortality in the patient.

The recipients of lung transplants had a median age of 60 years (IQR: 52–64) and predominantly belonged to the male gender (70%). The primary indications for transplantation were emphysema (33%), interstitial lung disease (50%), or other etiologies (17.5%). Single lung transplantation was performed in 45% of cases, while double lung transplantation was performed in 55%. Following lung transplantation, patients experienced a median (IQR) length of stay in the intensive care unit (ICU) of 19 days (IQR: 13–39), and mechanical ventilation was administered for a median (IQR) duration of 2.5 days (IQR: 1–6). The mortality rates in the ICU and at 1 year post-transplantation were 5% and 15%, respectively. Within the first year post-transplantation, 17.5% of patients (n = 7) developed at least one episode of acute cellular rejection (ACR), with a median (IQR) onset time of 18 days (IQR: 13-221). Among these episodes, five were classified as A1 and two as A2. The median (IQR) postoperative plasma sCD31 levels were 4240 pg/ml (IQR: 2753-6114) at H24, 4251 pg/ml (IQR: 2860-6197) at H48, and 4285 pg/ml (IQR: 2950-6414) at H72. Comparison of plasma sCD31 levels between patients with and without ACR revealed the following medians (IQR) at H24: 4280 pg/ml (IQR: 3137-4646) and 4160 pg/ml (IQR: 2738-6428), at H48: 3757 pg/ml (IQR: 2570-4173) and 4618 pg/ml (IQR: 3184-7105), and at H72: 3259 pg/ml (IQR: 2753-6154) and 4773 pg/ml (IQR: 3099-6871), respectively. Additionally, the mean sCD31 levels for patients with ACR grades A1 and A2 were as follows: at H24 – 3973 pg/ml and 12,688 pg/ml, at H48 – 2826 pg/ml and 5799 pg/ml, and at H72 – 3597 pg/ml and 2596 pg/ml, respectively.

All the aforementioned characteristics exhibited an intra-correlation below 0.25 and slightly exceeding 0.4 in the inter-correlation calculation. Concerning the discrimination of underrepresented classes, namely the ACR category, and with a primary focus on optimising balanced accuracy, we advocated for maintaining equilibrium in class distribution within mini-batches, each comprising 32 samples. In the context of binary classification involving 2 classes, approximately 5% of each class's samples are included in every training batch. Acknowledging the imbalanced nature of our dataset, we employed k-fold stratified cross-validation, ensuring a consistent class distribution in each subset. We adhered to a default testing setup, allocating 0.75 of the data for

training and 0.25 for testing. This implementation of stratified sampling underscores the significance of allowing the actual classes or targets to guide the sampling procedure. For example, applying a default fivefold cross-validation task with around 30% of the data reserved for validation to our model perpetuates the balanced distribution of each individual data batch. This strategic approach ensures alignment with the overall distribution of the training task. In summary, amalgamating all components, our method achieved a reduction of approximately 70% in data dimensions.

*Iris*

This is a multivariate dataset containing measurements of the sepal length, sepal width, petal length, and petal width of 150 iris flowers. These measurements were taken from three different species of iris flowers: setosa, versicolour, and virginica. The dataset is commonly used in machine learning as a classic example of supervised learning, particularly for classification tasks.

*Shapes*

The "Shapes" dataset is a widely-used synthetic dataset in machine learning studies for classification tasks. It comprises 60,000 grayscale images of medium complexity geometric shapes, such as planes, chairs, mugs, and humans, with a resolution of 28 × 28 pixels. The images are artificially generated, introducing random noise and rotations, and equally divided among the four classes. This dataset serves as a standard benchmark to assess the effectiveness of various machine learning algorithms, mainly neural networks, in image recognition tasks.

*Pain*

In addition to the Iris and Shapes datasets, we have included a new cohort from our collaborators in the Avicenne Hospital, the "Pain" database, consisting of 168 patients with 130 variables under study. This cohort represents a diverse patient population experiencing chronic pain and undergoing surgical procedures. The collected data encompasses various aspects, including medical history, the medical condition causing the pain, its duration and intensity, the type of chronic pain (nociceptive, neuropathic, inflammatory, or mixed), cognitive status, and functional status of patients prior to surgery. Postoperatively, patients were followed for up to 2.5 years, during which cognitive and functional outcomes were assessed. Outcomes were categorised into healthy, cognitive decline, functional impairment, or mortality. Cognitive evaluation was conducted using the telephone version of the Montreal Cognitive Assessment (T-MoCA) tool. Functional status was assessed using a composite of the Instrumental Activity of Daily Living Scale (IADL) and the Metabolic Equivalent of Task (METs).

Concerning the testing setup and training procedures, we adopted the same approach as employed for the Lung Transplantation dataset. Specifically, we allocated 75% of the data for training and 25% for testing. Additionally, we utilised k-fold stratified cross-validation in mini-batches to ensure robustness in the evaluation process.

### 1.4. Topological data analysis

Topological data analysis (TDA) is an approach for analysing complex datasets by focusing on their topological structure. This approach is particularly useful for visualising high-dimensional datasets and extracting descriptive features from data that are often represented by point clouds in Euclidean or metric spaces. TDA uses classical tools of algebraic topology to measure the importance of topological features according to the lifetime of a connected component. The main tool of topological data analysis is persistence homology [34]. In this work, we use different TDA tools, including persistence homology and the Rips filtration. We provide a brief introduction to these tools below.

*Persistence homology*

Persistence homology is a fundamental concept in topological data analysis. It is a method for identifying the connected components, loops, and cavities in a dataset, as well as their persistence over different scales. To understand persistence homology, we first introduce some basic definitions:

- Simplex: An $n$-simplex is a convex hull of $n+1$ affine independent points in a Euclidean space $\mathbb{R}^d$, where $n \leq d$. The vertices of a simplex are denoted by $s_0, \ldots, s_n$.
- Simplicial complex: A simplicial complex $K$ in $\mathbb{R}^d$ is a set of simplexes in $\mathbb{R}^d$ that satisfies certain conditions. Specifically, every face of a simplex in $K$ must also be in $K$, and the intersection of any two simplexes in $K$ must be a simplex in $K$.
- Subcomplex: A simplicial complex $L$ that is a subset of $K$ is called a subcomplex of $K$.
- Homology group: The homology of a topological space $X$ is the set of topological invariants of $X$ represented by its homology groups $H_0$, $H_1$, $H_2$, etc. The $k$th homology group $H_k$ describes the $k$-dimensional holes in $X$.
- Betti number: The $i$th Betti number, denoted by $b_i(K)$, is the cardinality of the $i$th homology group $H_i(K)$. Specifically, $b_0$ is the number of connected components, $b_1$ is the number of one-dimensional holes or loops, and $b_2$ is the number of voids or two-dimensional cavities.

A filtration of a simplicial complex is a nested sequence of subcomplexes, where the last subcomplex is the entire complex. In this work, we use the Rips filtration, which constructs a simplicial complex for each scale parameter $t$ by including all simplexes whose vertices are within a certain distance of each other. This results in a nested sequence of simplicial complexes, where the subcomplexes at each scale parameter $t$ are included in the subcomplexes at larger scale parameters.

*Rips filtration*

The Rips filtration is a type of filtration that constructs a simplicial complex for each scale parameter $t$ by including all simplexes whose vertices are within a certain distance of each other. Specifically, for any $t \geq 0$ and $i \geq 0$, the Rips complex is defined by:

$$R_i(X_t) = \{(x_0, \ldots, x_i) \mid d(x_j, x_k) < 2t \text{ for all } j, k \in 0, \ldots, i$$

Here, $X_t$ denotes a set of points in $\mathbb{R}^d$, and $d(x_j, x_k)$ denotes the distance between the points $x_j$ and $x_k$. The Rips complex at scale parameter $t$ includes all simplices up to dimension $i$ whose vertices are within a distance of $2t$ of each other.

The Rips filtration is a nested sequence of simplicial complexes, where the subcomplexes at each scale parameter $t$ are included in the subcomplexes at larger scale parameters (i.e., $\forall s \leq t, R(X_s) \subseteq R(X_t)$). This allows for the computation of persistence diagrams, which summarise the evolution of homology classes as the scale parameter $t$ increases. Persistence diagrams are a powerful tool for identifying topological features that are persistent across multiple scales, as well as for measuring the robustness of these features to noise and perturbations in the data.

The persistence diagram, denoted as $D$, is a collection of points and lines that serve to represent topological characteristics over time [35]. Specifically, the points in $D$ are given by pairs $(b, d)$, where $b$ and $d$ represent the birth and death time, respectively, of a topological feature. Additionally, vertical lines that pass through points of the form $(b, b)$ denote topological features that appear at time $b$ and never die.

*Persistence entropy*

To facilitate the application of machine learning techniques, this project employs a tool from TDA to transform the persistence diagram. This transformation involves using entropy, a mathematical function that separates the persistence diagram into sub-diagrams based on

the homology dimension. The entropy of each sub-diagram is then calculated using the formula

$$E(D) = -\sum_{i=0}^{m} p_i \log p_i,$$

where $p_i = \prod_{j=1}^{m}(d_j - b_j)/(m)$ represents the probability of a point in $D$ belonging to the $i$th sub-diagram, and $m$ denotes the total number of points in the persistence diagram. Finally, the result of this process is a three-dimensional vector $(E(D_1), E(D_2), E(D_3))$, where each component corresponds to the entropy of a particular sub-diagram.

### 1.5. Machine learning models

The provided statement describes an overview of the learning approaches used to predict the risk score of ACR after lung transplantation. Additionally, it discusses how this knowledge can be leveraged to create a method to reduce dimensionality by exploring the current bias and limits of machine learning techniques commonly used in medical analysis. In our study, we have indeed conducted a comprehensive analysis by employing a grid search for ML model parameters. This exploration encompasses various parameter choices, starting from default settings, to ensure a thorough examination of the model's performance landscape. However, for clarity and brevity in presentation, we chose to showcase the best-performing configurations in the results section.

*Random forest (RF)*

It is a popular supervised learning algorithm used for classification or regression tasks [23,36]. To understand RF, it is necessary to define the decision tree. A decision tree consists of a root node, which represents the main node of the tree, internal nodes that represent attributes, and leaves that indicate the decisions. The decision tree algorithm works by recursively dividing the training dataset into two or more subsets until a leaf node is reached.

Random Forest builds on decision tree algorithms by creating multiple decision trees and combining their outputs to make predictions. Each tree is constructed using a different subset of the training data and a random subset of the features. This randomness helps to prevent overfitting and improve the accuracy of the model.

Briefly, RF algorithm operates as follows [37]:

1. Create a bootstrap dataset by randomly selecting samples from the original dataset. Both the bootstrap dataset and the original dataset must have the same size.
2. Build a decision tree for the bootstrap dataset.
3. Repeat steps 1 and 2 several times to obtain a variety of trees.
4. Each decision tree generates an output.
5. The most frequent class by the decision trees is returned.

*Multilayer perceptron (MLP)*

It is an artificial neural network organised into several layers, including the input layer, hidden layers, and output layer [24].

To train the multilayer perceptron, the steps are as follows:

**Forward propagation:** Let $x = (x_1, \ldots, x_{n_0})$ be the input vector, $W \in \mathbb{R}^{n_0 \times n_1}$ be the weight matrix, with $n_1$ being the number of neurons in a layer, and $b = (b_1, \ldots, b_{n_1})$ be the bias vector. The output of a layer is given by the following formula: $y = \varphi(Wx+b)$, where $\varphi$ is an activation function. The activation functions applied in this work were:

- ReLU: $g(x) = \max(0, x)$
- tanh: $g(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
- sigmoid: $g(x) = \sigma(x) = \frac{1}{1+e^{-x}}$

Calculate the error between the model output $y$ and the desired output $d$ using a cost function $L$, often using the mean squared error:

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - d_i)^2, \tag{2}$$

where $n$ is the number of data.

**Backpropagation:** involves calculating a chain of gradients to determine how the cost function $L$ varies with respect to the parameters $(W, b)$ of each layer.

**Update the parameters $(W, b)$ of each layer to minimise the error:**

$$W_{t+1} = W_t - \alpha \frac{\partial L}{\partial W} \tag{3}$$

where $W_t$ is the parameter $W$ at time $t$, $W_{t+1}$ is the parameter $W$ at time $t+1$, and $\alpha$ is the positive learning rate. And $\frac{\partial L}{\partial w_i}$ is the gradient of $L$ with respect to $w_i$ at time $t$.

*The K-nearest neighbours (KNN)*

It is a supervised learning method that can be used for classification or regression [25]. The classification algorithm works as follows [12]:

1. Calculate the distance between the point to be classified and the training example.
2. Sort the training examples in ascending order of distance.
3. Choose the $k$ closest points with the smallest distance.
4. Return the most frequent class among the $k$ closest points.

The different distances used are:

- Euclidean distance: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Manhattan distance: $d(x, y) = \sum_{i=0}^{n} |x_i - y_i|$
- Chebyshev distance: $d(x, y) = \max_{0 \le i \le n} |x_i - y_i|$

*Naive Bayes classifier (GNB)*

It is a supervised learning algorithm used for classification. It is based on the application of Bayes' theorem with strong (naive) independence assumptions between the features [26].

Bayes' Theorem: Let $A$ and $B$ be two events such that $P(B) \neq 0$, then we have: $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$ where $P(A|B)$ is the posterior probability of $A$ given $B$, and $P(A)$ and $P(B)$ are the marginal or prior probabilities.

We apply Bayes' theorem on the training set as follows:

$$P(y \mid X) = \frac{P(X \mid y)P(y)}{P(X)}$$

where $X = (x_1, .., x_n)$ is a feature vector and $y$ is a class variable. Then, we have:

$$P(y \mid (x_1, \ldots, x_n)) = \frac{P((x_1, \ldots, x_n) \mid y)P(y)}{P((x_1, \ldots, x_n))}$$

According to the naive assumption of Bayes' theorem, we have:

$$P(y \mid (x_1, \ldots, x_n)) = \frac{P((x_1, \ldots, x_n) \mid y)P(y)}{P((x_1, \ldots, x_n))} = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P((x_1, \ldots, x_n))}$$

Regarding the formula, the denominator is a constant, so it can be removed. Thus,

$$P(y \mid x_1, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

Finally, for $x = (x_1, \ldots, x_n)$, which is the feature vector of the test set, the predicted class can be determined by

$$y = \underset{y}{\operatorname{argmax}}\, P(y)\prod_{i=1}^{n} P(x_i \mid y)$$

### 1.6. State-of-the-art techniques for analysing high-dimensional data

Some of the dimensional reduction algorithms used during this work were:

*An autoencoder (AE)*

It is an artificial neural network that learns to reconstruct input data through unsupervised learning [28]. The architecture of an AE is referred to as a bottleneck because the hidden layers are smaller than the input and output layers [30]. The AE is composed of two parts: the encoder, which compresses the input into a smaller representation, and the decoder, which reconstructs the input from the new representation. The objective of the AE is to learn the most important information from the input, achieve the best possible compression, and allow the decoder to reconstruct an output that is faithful to the input. The AE is trained by defining two functions, $\varphi : X \to Z$ and $\psi : Z \to X$, where $Z$ is the bottleneck space that corresponds to the compressed representation. The reconstruction error $L$ defined for $x \in X$ as $L = |x - \psi \circ \varphi(x)|^2$ is then backpropagated through the network to update the weights.

*Principal component analysis (PCA)*

It is an unsupervised linear dimensionality reduction technique that converts a set of observations of possibly correlated variables into a set of linearly uncorrelated variable values called principal components [32]. The PCA algorithm proceeds by centring and reducing the input data, computing the variance and covariance matrix of the transformed data, and calculating the eigenvectors and eigenvalues. The eigenvectors are sorted in descending order based on their highest associated eigenvalues, as they capture the most significant variance in the data. This criterion ensures that the selected components contribute the most to the overall information retained in the reduced-dimensional space. The first 2 or 3 eigenvectors are selected to form the matrix $P$. The transformed data is then computed as $Z^* = ZP^*$, where $Z$ is the original data matrix, $P$ the matrix containing the selected top 2 or 3 eigenvectors, $P^*$ the transpose of matrix $P$, and $Z^*$ the matrix of transformed data, obtained by multiplying $Z$ with $P^*$.

*The t-SNE*

This is a nonlinear dimensionality reduction technique that preserves both global and local structure [33]. It maps high-dimensional data to two dimensions by using the local relationship between the data. The algorithm calculates the perplexity parameter $\sigma^2$ and the conditional probability of $(x_i, x_j)$ with $i \neq j$. It then calculates the symmetric probability $p_{ij}$ and initialises $Y$ using the multivariate normal distribution $N(0, 10^{-4}I)$. The algorithm then calculates the probability $q_{ij}$ and minimises the Kullback–Leibler divergence with gradient descent, which is equivalent to minimising the gap between the probability distributions between the original space $p_{ij}$ and the lower-dimensional space $q_{ij}$.

*Uniform approximation and projection of manifolds (Umap)*

It is an improvement on t-SNE that is faster and provides a better representation of the global structure [33,38]. The Umap algorithm consists of three steps: building a graph in high dimensions, constructing the Cěch complex with a variable radius determined for each point based on the distance to its $k$th nearest neighbour [39], and projecting the data into lower dimensions via a force-directed graph layout algorithm.

### 1.7. Topological amplitude

The metric amplitude [40] is a measure of the discrepancy between the observed and expected values of a topological summary statistic. Specifically, let $\mathcal{X}$ be a topological space and $f : \mathcal{X} \to \mathbb{R}$ be a real-valued function on $\mathcal{X}$. Given a finite point cloud $X \subset \mathcal{X}$, we can compute the empirical measure $\mu_X$ on $X$, which assigns a probability mass to each point in $X$. The metric amplitude of $f$ on $X$ is then defined as

$$\mathrm{MA}_f(X) = \sup_{\nu \in P(\mathcal{X})} \left| \int_{\mathcal{X}} f \, d\mu_X - \int_{\mathcal{X}} f \, d\nu \right|, \tag{4}$$

where $\mathcal{P}(\mathcal{X})$ is the space of probability measures on $\mathcal{X}$. In other words, the metric amplitude measures the largest possible difference between the expected value of $f$ under the empirical measure $\mu_X$ and the expected value of $f$ under any probability measure on $\mathcal{X}$.

This metric is used to assess the stability of topological summaries, such as persistent homology, under perturbations of the point cloud $X$. Intuitively, if the metric amplitude of a topological summary is small, then the summary is stable and robust to small perturbations of $X$.

### 1.8. Evaluation

To evaluate our model, we calculated:

*Kullback–Leibler (KL)*

We use this divergence metric to measure the similarity between the input space density distribution and the latent space density distribution. Let $X$ be the input space and $Z$ be the latent space. For $x \in X$ and $\sigma \in R_+^*$, we define the density estimator of $X$ as

$$f_\sigma^X(x) = \sum_{y \in X} \exp(-\sigma^{-1} \mathrm{dist}(x, y)^2),$$

where $\mathrm{dist}(x, y)$ is the normalised Euclidean distance between 0 and 1. Then we calculate $\mathrm{KL}_\sigma = KL(f_\sigma^X / f_\sigma^Z)$ to determine how similar the two distributions are.

*Root Mean Square Error (RMSE)*

It uses the pairwise distance matrices of the input and latent spaces to measure how often the two distance distributions coincide. Let $D^X$ be the distance matrix associated with the input space $X$ and $D^Z$ be the distance matrix associated with the latent space $Z$. Then we calculate $\mathrm{RMSE} = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (x_i - y_j)^2}$, where $x_i \in D^X$ for $i \in 1, ..n$, $y_i \in D^Z$ for $i \in 1, ..n$, and $n$ is the number of points. RMSE is not related to the reconstruction error.

*Reliability (trUSt, transitive unbiased similarity matrix)*

It uses the pairwise distance matrices of the input and latent spaces. Let $D^X$ be the distance matrix associated with the input space $X$ and $D^Z$ be the distance matrix associated with the latent space $Z$. Then we calculate TRUST as

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in N_i^k} max(0, (r(i, j) - k)),$$

where $N_i^k$ are the $k$ nearest neighbours of $i$ in the output space for $i \in 1, ..n$ and $r(i, j)$ is the closest neighbour of $j$ in the input space for $j \in N_{i,k}$. The goal is to penalise any unexpected nearest neighbour in the output space proportional to its rank in the input space.

Finally, we computed the:

*Mean squared error (MSE)*

Let $X$ be the set of input data and $Y$ be the set of reconstructed data. We define the reconstruction error as $\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2$.

## Code availability

The code for the implementation of the proposed method, Taelcore, is available on GitHub at https://github.com/MorillaLab/Taelcore/Dimensionality reduction/. Additionally, the code used for the analysis and experiments presented in this paper is also available on GitHub at https://github.com/MorillaLab/Taelcore/Topological improvement.

## CRediT authorship contribution statement

**Fatma Gouiaa:** Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Kelly L. Vomo-Donfack:** Formal analysis. **Alexy Tran-Dinh:** Resources, Methodology, Data curation, Conceptualization. **Ian Morilla:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

## Acknowledgements

We would like to thank the funding from the National Research Association (ANR) (Inflamex renewal 10-LABX-0017 to I Morilla), Consejería de Universidades, Ciencias *y* Desarrollo, fondos FEDER de la Junta de Andalucía (ProyExec_0499 to I Morilla), DHU FIRE Emergence 4, and the l'Agence de la Biomedecine.

## References

[1] A. Kumar, F. Anjum, Lung transplantation, 2023, StatPearls [Internet]. Treasure Island (FL): StatPearls: https://www.ncbi.nlm.nih.gov/books/NBK565849/ (Accessed 27 April 2023).

[2] J.P. Singer, L.G. Singer, Quality of life in lung transplantation, Semin Respir. Crit. Care Med. 34 (3) (2013) 421–430.

[3] P.J. McShane, L.G. Ruiz, E.R. Garrity, Chapter 75 - lung transplantation, in: S.G. Spiro, G.A. Silvestri, A. Agustí (Eds.), Clinical Respiratory Medicine, fourth ed., W.B. Saunders, Philadelphia, 2012, pp. 882–903, URL https://www.sciencedirect.com/science/article/pii/B9781455707928000751.

[4] A. Uluer, F.M. Marty, 73 - Cystic fibrosis, in: J.E. Bennett, R. Dolin, M.J. Blaser (Eds.), Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases, eighth ed., W.B. Saunders, Philadelphia, 2015, pp. 874–885.e3, URL https://www.sciencedirect.com/science/article/pii/B9781455748013000734.

[5] J. Raskin, et al., Mortality after lung transplantation: a single-centre cohort analysis, Transpl. Int. 33 (2) (2020) 130–141, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tri.13540, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/tri.13540.

[6] M. Field, R. Lawrence, L. Zwanziger, 5 – immunosuppressive drugs for transplant patients, in: M. Jones (Ed.), Extending Medicare Coverage for Preventive and Other Services, National Academies Press (US), Washington (DC), 2000, URL https://www.ncbi.nlm.nih.gov/books/NBK225251/.

[7] S. Kotecha, S. Ivulich, G. Snell, Review: immunosuppression for the lung transplant patient, J. Thorac. Dis. 13 (11) (2021) URL https://jtd.amegroups.org/article/view/53209.

[8] P. Trachuk, R. Bartash, M. Abbasi, A. Keene, Infectious complications in lung transplant recipients, Lung 198 (2020) 879–887.

[9] J. Reininghaus, S. Huber, U. Bauer, R. Kwitt, A stable multi-scale kernel for topological machine learning, in: CVPR, IEEE Computer Society, 2015, pp. 4741–4748, URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#ReininghausHBK15.

[10] M. Nicolau, A. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, Proc. Natl. Acad. Sci. USA 108 (17) (2011) 7265–7270.

[11] C. Wu, C. Hargreaves, Topological machine learning for mixed numeric and categorical data. International Journal on Artificial Intelligence Tools, J. Mach. Learn. Res. 30 (5) (2021) 215–225.

[12] T. Hinks, et al., Multidimensional endotypes of asthma: topological data analysis of cross-sectional clinical, pathological, and immunological data., Lancet 385 (2015).

[13] A. Dagliati, et al., Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records, AIM 108 (2020) 101930, URL https://www.sciencedirect.com/science/article/pii/S0933365719311844.

[14] Y. Singh, et al., Topological data analysis in medical imaging: current state of the art, Insights Imaging 14 (1) (2023) 1–10.

[15] I. Morilla, et al., Deep models of integrated multiscale molecular data decipher the endothelial cell response to ionizing radiation, iScience 25 (1) (2022) 103685, URL https://www.sciencedirect.com/science/article/pii/S2589004221016552.

[16] I. Morilla, A deep learning approach to evaluate intestinal fibrosis in magnetic resonance imaging models, Neural Comput. Appl. 32 (18) (2020) 14865–14874, URL http://dblp.uni-trier.de/db/journals/nca/nca32.html#Morilla20.

[17] A. Tran-Dinh, et al., Personalized risk predictor for acute cellular rejection in lung transplant using soluble CD31, Sci. Rep. (2022).

[18] S. Lambden, P.F. Laterre, M.M. Levy, B. Francois, The SOFA score-development, utility and challenges of accurate assessment in clinical trials, Crit. Care 23 (1) (2019) 1–9, URL https://ccforum.biomedcentral.com/track/pdf/10.1186/s13054-019-2663-7.pdf.

[19] S. Gauthier, A. Tran-Dinh, I. Morilla, Plasma proteome dynamics of COVID-19 severity learnt by a graph convolutional network of multi-scale topology, LSA 6 (5) (2023) arXiv:https://www.life-science-alliance.org/content/6/5/e202201624.full.pdf.

[20] G. Tauzin, et al., giotto-tda:: A topological data analysis toolkit for machine learning and data exploration, J. Mach. Learn. Res. 22 (39) (2021) 1–6, URL https://www.jmlr.org/papers/volume22/20-325/20-325.pdf.

[21] F. Chazal, B.T. Fasy, F. Lecci, A. Rinaldo, L.A. Wasserman, Stochastic convergence of persistence landscapes and silhouettes, J. Comput. Geom. 6 (2) (2015) 140–161, URL http://dblp.uni-trier.de/db/journals/jocg/jocg6.html#ChazalFLRW15.

[22] H. Adams, et al., Persistence images: A stable vector representation of persistent homology, J. Mach. Learn. Res. 18 (1) (2017) 218–252, URL https://www.jmlr.org/papers/volume18/16-337/16-337.pdf.

[23] Y. Brostaux, Etude du classement par forêts aléatoires d'échantillons perturbés à forte structure d'interaction, ULiège. GxABT-Liège Université. Gembloux Agro-Bio Tech [Médecine vétérinaire], 2005, URL https://orbi.uliege.be/bitstream/2268/23636/1/PhD_YBT_05.pdf.

[24] H. Wang, J. Wang, Short-term wind speed prediction based on feature extraction with multi-task lasso and multilayer perceptron, Energy Rep. 8 (2022) 191–199, URL https://www.sciencedirect.com/science/article/pii/S2352484722006497.

[25] P. Mulak, N. Talhar, Analysis of distance measures using k-nearest neighbor algorithm on kdd dataset, Int. J. Sci. Res. 4 (7) (2015) 2319–7064, URL https://scholar.googleusercontent.com/scholar?q=cache:_aV66tEzRlUJ:scholar.google.com/+k+nearest+neighbors+algorithm+les+distances&hl=fr&as_sdt=0,5.

[26] D. Berrar, Bayes' theorem and naive Bayes classifier, in: Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, Vol. 403, Elsevier Science Publisher Amsterdam, The Netherlands, 2018, URL https://scholar.google.com/scholar?hl=fr&as_sdt=0%2C5&q=+Bayes%27+theorem+and+naive+Bayes+classifier&btnG=.

[27] P. Bubenik, et al., Statistical topological data analysis using persistence landscapes, J. Mach. Learn. Res. 16 (1) (2015) 77–102, URL https://www.jmlr.org/papers/volume16/bubenik15a/bubenik15a.pdf?source=post_page.

[28] I. Gjorshoska, T. Eftimov, D. Trajanov, Missing value imputation in food composition data with denoising autoencoders, J. Food Compos. Anal. (2022) 104638, URL https://www.sciencedirect.com/science/article/pii/S0889157522002563#fig0005.

[29] M. Moor, M. Horn, B. Rieck, K. Borgwardt, Topological autoencoders, in: International Conference on Machine Learning, PMLR, 2020, pp. 7045–7054, URL http://proceedings.mlr.press/v119/moor20a.html?ref=https://githubhelp.com.

[30] C.S.N. Pathirage, et al., Structural damage identification based on autoencoder neural networks and deep learning, Eng. Struct. 172 (2018) 13–28, URL https://www.sciencedirect.com/science/article/pii/S0141029618302062?casa_token=D6w7sEvE_MoAAAAA:aIAEl4IE1AGEs4SawcLeDpw-kLTfUAO74jzWA851_9mPVajHJYIkKNEtu97XUHhySWP1P_DcgO8.

[31] F. Chazal, D. Cohen-Steiner, L.J. Guibas, F. Mémoli, S.Y. Oudot, Gromov-Hausdorff stable signatures for shapes using persistence, in: Computer Graphics Forum, proc. SGP 2009, 2009, pp. 1393–1403, URL http://geometrica.saclay.inria.fr/team/Steve.Oudot/papers/ccgmo-ghsssp-09/index.html.

[32] S.P. Mishra, et al., Multivariate statistical data analysis-principal component analysis (PCA), IJLR 7 (5) (2017) 60–78, URL https://scholar.googleusercontent.com/scholar?q=cache:6OpNbJY263cJ:scholar.google.com/+principal+component+analysis+pca&hl=fr&as_sdt=0,5.

[33] Y. Wang, H. Huang, C. Rudin, Y. Shaposhnik, Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization, J. Mach. Learn. Res. 22 (201) (2021) 1–73, URL https://www.jmlr.org/papers/volume22/20-1061/20-1061.pdf.

[34] P. Lascabettes, Homologie Persistante Appliquéea le reconnaissance de genres musicaux (Master's thesis), École Normale Supérieure Paris-Saclay, 2018, URL http://repmus.ircam.fr/_media/moreno/dissertation_paul_lascabettes.pdf.

[35] E. Munch, A user's guide to topological data analysis, JLA 4 (2) (2017) 47–61, URL https://learning-analytics.info/index.php/JLA/article/view/5196/6061.

[36] A. Palczewska, J. Palczewski, R. Marchese Robinson, D. Neagu, Interpreting random forest classification models using a feature contribution method, in: T. Bouabana-Tebibel, S.H. Rubin (Eds.), Integration of Reusable Systems, Springer International Publishing, 2014, pp. 193–218.

[37] T.A. Pham, V.Q. Tran, Developing random forest hybridization models for estimating the axial bearing capacity of pile, PLoS One 17 (3) (2022) e0265747, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0265747.

[38] Y. Hozumi, R. Wang, C. Yin, G.-W. Wei, UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets, Comput. Biol. Med. 131 (2021) 104264, URL https://www.sciencedirect.com/science/article/pii/S0010482521000585?casa_token=fF6PHDekRE8AAAAA:2DZ6eDvvChT-t0MOHUVmI0Qh7TWv7NpVOvlKia_HBSbmHAUm46SGpT970uYO0ESkAmwBnpy-Bmo.

[39] S.S. Dantchev, I.P. Ivrissimtzis, Efficient construction of the Čech complex, Comput. Graph. 36 (6) (2012) 708–713, URL http://dblp.uni-trier.de/db/journals/cg/cg36.html#DantchevI12.

[40] G. Tauzin, et al., giotto-tda: A topological data analysis toolkit for machine learning and data exploration, CoRR abs/2004.02551, 2020, URL http://dblp.uni-trier.de/db/journals/corr/corr2004.html#abs-2004-02551.