SORT 46 (1) January-June 2022, 115-136 DOI: 10.2436/20.8080.02.120

Topological Data Analysis and its usefulness for precision medicine studies

Raquel Iniesta*,[1], Ewan Carr[1], Mathieu Carrière[2], Naya Yerolemou[3], Bertrand Michel[4] and Frédéric Chazal[5]

Abstract

Precision medicine allows the extraction of information from complex datasets to facilitate clinical decision-making at the individual level. Topological Data Analysis (TDA) offers promising tools that complement current analytical methods in precision medicine studies. We introduce the fundamental concepts of the TDA corpus (the simplicial complex, the Mapper graph, the persistence diagram and persistence landscape). We show how these can be used to enhance the prediction of clinical outcomes and to identify novel subpopulations of interest, particularly applied to understand remission of depression in data from the GENDEP clinical trial.

MSC: *Statistical aspects of big data and data science (62R07) and Topological data analysis (62R40)*

Keywords: *Precision medicine, data shape, topology, topological data analysis, persistence diagram, Mapper, persistence landscapes, machine learning.*

1. Precision medicine: what are the current needs?

The field of precision medicine is focused on the development of sophisticated algorithms that, by exploiting patient data – on clinical measurements, genomics, proteomics, medical imaging, etc. – can guide clinicians to make more accurate diagnoses, prognoses and treatment choices tailored to individual patients. The datasets used to develop these

[1] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. *Corresponding author: raquel.iniesta@kcl.ac.uk

[2] Inria Sophia-Antipolis, DataShape Team, Biot, France.

[3] The University of Oxford and The Alan Turing Institute. UK.

[4] Ecole Centrale de Nantes,LMJL – UMR CNRS 6629, Nantes, France.

[5] Inria Saclay - Ile-de-France, Alan Turing Bldg, Palaiseau, France.

Received: September 2021.
Accepted: April 2022.

models present multiple complexities. They routinely include information for thousands of subjects, and the number of included variables can easily exceed millions (i.e., these datasets are high-dimensional), variables tend to be highly correlated, and may interact in complex ways that may not be immediately obvious. These factors combined often limit the utility of classical statistical procedures in the analysis of these data. In recent years, machine learning (ML) (Mitchell, 1997, 2006), a set of tools at the interface between computer sciences and statistics, has been used in precision medicine to overcome some of these limitations. The use of ML has led to the development of interesting predictive models built from complex data sets (Ekins et al., 2019; Ho et al., 2019; Rajkomar, Dean and Kohane, 2019; Iniesta, Stahl and McGuffin, 2017). However, the success of ML for these datasets has varied across medical areas – performing moderately well in some diseases but very poorly in others (Adamson and Welch, 2019; Iniesta et al., 2016, 2017, 2018), leaving considerable room for improvement. Recently, several works including studies on COVID-19 research, have emphasised the increasing demand of novel methods that can better deal with such complexity (Khan et al., 2019, 2021).

One of the key challenges in building models that can accurately predict outcomes for new patients is correctly identifying sources of heterogeneity among patients (i.e., sources that could contribute to observed differences in patient outcomes) and including these in the model in the form of predictor variables. When tailoring the choice of medical treatment to patients' pre-treatment characteristics, methods to identify subgroups in terms of treatment effectiveness – for example, where patients respond similarly to treatment within the group, and differently between groups – constitute one of the most prominent challenges currently for medical statisticians (Sies, Demyttenaere and Mechelen, 2019).

In addition to developing predictive models, methods for visualising data in high dimensions can facilitate decision-making for diagnosis and treatments targeting. Most classical tools, such as scatter plots or heat maps, are often restricted to two dimensions (Qu et al., 2019). Although new technologies have been used to create visualisation tools applicable to complex data, fields like genomics research are rapidly evolving and continuous advancement in visualisation techniques is needed (Nusrat, Harbig and Gehlenborg, 2019).

In recent years a growing literature has highlighted the benefits of applying topological techniques in precision medicine studies. For example, to identify genetic influences on patient survival in breast cancer (Nicolau, Levine and Carlsson, 2011), to improve treatment targeting for patients with spinal cord or traumatic brain injury by uncovering previously hidden data relationships in 20-year old data (Nielson et al., 2017), or to identify disease trajectories in type 2 diabetes data (Dagliati et al., 2020).

This paper aims to provide a first introduction to some of the basic topological concepts that form the field of Topological Data Analysis (TDA): the simplicial complex, the Mapper graph, the persistence diagram and the persistence landscape. We show how these techniques offer promising tools to reveal data structures not readily accessible using other statistical techniques, which may subsequently help machine learning models

in predicting clinical outcomes. We show an application of these methods to investigate remission of depression in data from the GENDEP clinical trial. We also summarise the software implementations of these techniques.

2. Introducing Topological Data Analysis

TDA is a promising field that has emerged from different works in applied algebraic topology (Edelsbrunner, Letscher and Zomorodian, 2000; Zomorodian and Carlsson, 2005; Ghrist, 2018). It aims to provide well-founded mathematical, statistical, and algorithmic methods to infer, analyse, visualise and exploit the complex topological and geometric structure of data (Chazal, 2016). The field is based on topology, the branch of mathematics born in response to Riemann's request in 1867 for "a good foundation of the concept of space" (Riemann and Clifford, 1998). In contrast with the more familiar field of geometry – the study of the shape of the space, that is, what the space *looks like* – topology can be broadly defined as the study of only those shape properties that are unaffected by continuous transformations such as stretching, shrinking, bending and twisting (examples of non-continuous transformations are cutting or gluing) (Kosniowski, 1980). For example, if a torus (a surface like a ring doughnut, as shown in Figure 1) is stretched horizontally, it does not change the fact that there is only one 'hole' on the inside; thus, this property is preserved despite transformation. Moreover, topological techniques assume *coordinate invariance*, the property that topological features are defined not in terms of their position on a coordinate system, but rather, in terms of their shape. Therefore, TDA can identify a torus regardless of whether the torus is compressed or stretched; the torus and its transformations are said to be *topologically equivalent*. Topological invariants like the number of holes and cavities are properties of a topological space that are shared by the space and all its topological equivalents (Henle, 1994). Properties such as these characterise the *invariant shape* of a space.

If we now move to the world of data, as Prof Gunnar Carlson reminds in his landmark paper (Carlsson, 2009) *data have shape* and this shape has a meaning. This idea is not new: linear regression, for example, is a well-established statistical technique based on the idea that the shape of data is linear – a line in two dimensions and a hyperplane in higher dimensions. Understanding the linear shape is key to understanding the relationship between dependent and independent variables. However, data may resemble many

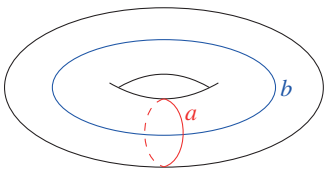


Figure 1. *The torus: has one connected component; two loops, since loops a and b are 'distinct' i.e. one cannot be transformed to the other along the torus' surface; and one void, since there is one void in the centre of the doughnut.*

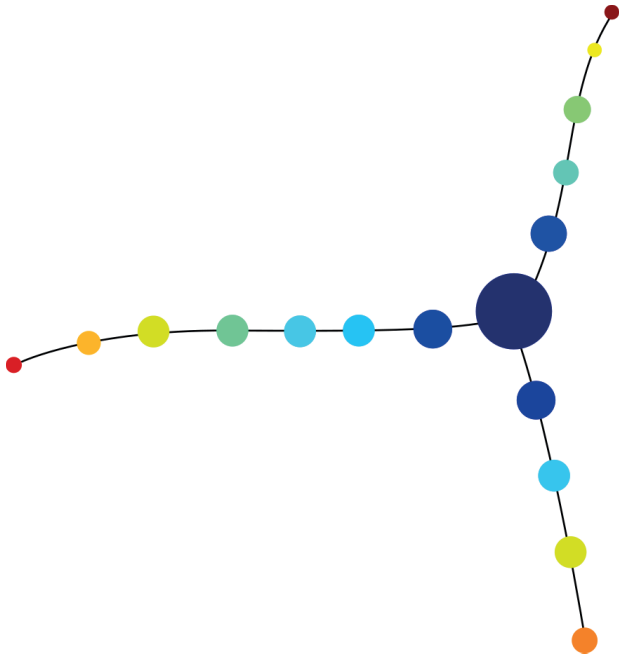


Figure 2. *Example of a flare, a single connected component consisting of three distinct groups of data*

other shapes which are harder to understand. Imagine, for example, data points that split into three distinct lines at a single point, forming a flare, i.e. a Y shape (Figure 2). Besides the flare, data may take on more complex shapes and unexpected behaviours, especially as the number of dimensions increases (e.g., the trefoil knot is knotted in three dimensions, but falls apart and becomes a trivial loop in four dimensions; see Figure 3).

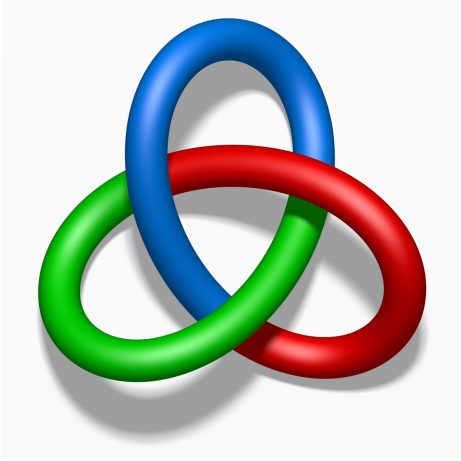


Figure 3. *A trefoil knot*

TDA provides techniques to describe these shapes by listing their topological invariants (such as holes or cavities), and to investigate the meaning of these topological features in terms of the specific data problem or clinical context.

2.1. Using TDA to help understand data structure

The question to answer is how we can 'build a bridge' from the collected patient data to a space in which topological invariants can be computed. This can be achieved in three steps.

Consider a dataset with m rows and n columns, where m is the number of observations in our sample (the number of patients, for example) and n is the number of measures collected for each patient.

1. Firstly, we need to define a measure to assess the proximity between any two data points – that is, to be able to measure how similar two patients are given the information we have for them. Interestingly, our data do not necessarily need to lie in the Euclidean space. As long as a distance can be computed between data points, we will be able to apply TDA tools. For ease of understanding, let us consider our data is made of numerical variables and let us represent them in a n-dimensional *point cloud* living in R^d, such that each patient becomes a point in the space, with each variable represented on a different coordinate axis. For the straightforward example of only two variables, values are drawn on the X and Y axes, but this can be extended to any number of dimensions, for three, four or more variables. In this way, we convert the patient data into a point cloud. For this particular case, we would assume that the point cloud is a finite sample drawn from an existing topological space. In case of a circle (see Figure 4a), we would assume that our data are a finite sample drawn from a 3D representation of a circle. In the Euclidean space R^d the natural choice of distance to assess similarity would be the Euclidean distance. There are also other distances that can be defined on numeric data as for example the Variance Normalised Euclidean or the Minkowski distance. When data are categorical rather than numeric, we can also define many different distances as for example the Gower distance.
2. Second, to highlight the underlying topology of the data we consider the construction of continuous shapes on top of the point cloud. These continuous shapes very commonly will be graphs. A graph is a finite, discrete representation of the set of points that encodes a one (or higher) dimensional skeleton of the data (Chartrand, 1985). Graphs are used in many data analysis applications and are much easier to visualise than the high-dimensional data used to construct them (see Figure 2).
3. Lastly, having built graphs based on the point cloud, we are able to compute the persistence diagram (and the extended-persistence diagram). These are topological signatures representing our data shape summary (Edelsbrunner et al., 2000; Zomorodian and Carlsson, 2005; Cohen-Steiner, Edelsbrunner and Harer, 2007; Carrière and Oudot, 2018).

Starting with a finite point cloud (Step 1), the following sections will introduce two approaches to constructing graphs on top of the point cloud (Step 2): the simplicial complex[1] and the Mapper graph. We will present the concept of persistence diagram, and will see (Step 3) how a persistence diagram can be derived from a family of simplicial complexes, and how a extended-persistence diagram can be derived from a single Mapper graph.

2.2. The simplicial complex

One way to construct a graph on top of point cloud data is by drawing a circle of radius ρ around each point in the cloud (Figure 4). If the corresponding circles for two points intersect, we connect the points with a line. If three circles intersect, we connect the three points to form a triangle, and so on. This particular graph is called a Čech complex and is a type of *simplicial complex*. A simplicial complex is a graph formed by a set of points, lines, triangles, etc. Simplicial complexes generalise the concept of one-dimensional graphs (formed only of edges and nodes) to allow other dimensional blocks like triangles (dimension 2), tetrahedrons (dimension 3) and so on. Besides the Čech complex, there are other types of simplicial complexes that can be constructed on top of point cloud data such as the Vietoris-Rips and the Alpha complex. In a Vietoris-Rips complex (Zomorodian, 2010) when 3 balls intersect (it can be a pairwise intersection, not all balls need to intersect), a triangle of dimension 2 is built. When 4 balls have a none empty intersection, a tetrahedron of dimension 3 is build, and so on. An Alpha complex is a simplicial complex constructed from the finite cells of a Delaunay Triangulation (Devillers, Hornus and Jamin, 2022). In terms of the topology of the Alpha complex (and its relationship with persistence theory) the Alpha complex is equivalent to the Čech complex and much smaller if one does not bound the radii.

2.3. The Mapper graph

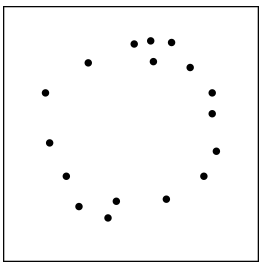
A second approach to constructing a graph on top of a point cloud is by using the Mapper algorithm (Singh, Mémoli and Carlsson, 2007). The Mapper algorithm reduces complex data to produce a one-dimensional graph – the Mapper graph. This consists of nodes (sets of clustered subjects) and edges connecting those nodes and edges connecting those nodes with non-empty intersections (that is, subjects can appear in more than one node).

The Mapper graph is built as follows. Suppose we have a finite point cloud and can compute all distances between pairs of points within the cloud. Suppose also, that we have a function called the *filter* that assigns a real value to each point in the data set. Then, the Mapper algorithm proceeds in the following steps (Figure 5):

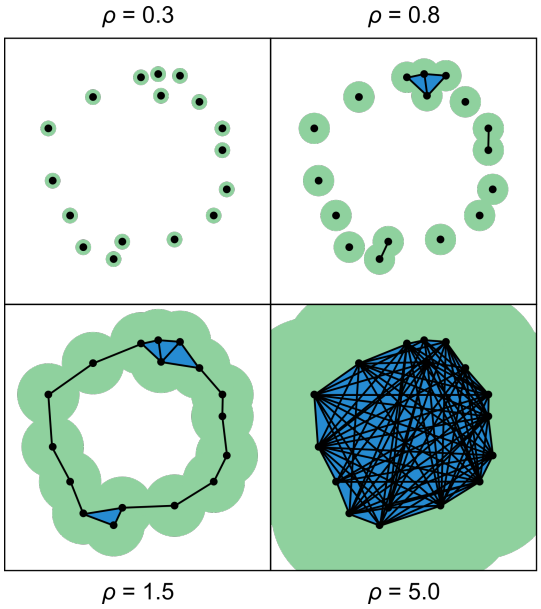
1. Find the range of the filter function (i.e., the interval of all values that the function takes);

[1] As simplicial complexes can be seen as higher dimensional generalizations of neighbouring graphs, we will make an abuse of notation and we will refer them as "graphs" throughout the paper

(a) *A point cloud sampled from a circle*



(b) *The sampled circle, now with smaller circles of an increasing radius ρ on top of each point*



(c) *The resulting persistence diagram, for a single topological feature in dimension 1.*

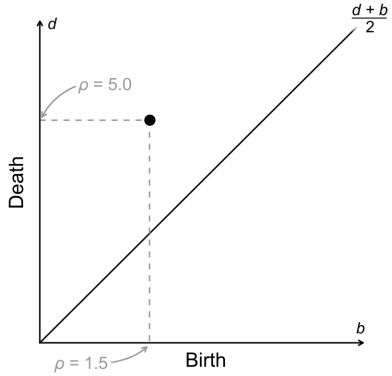


Figure 4. *Constructing the Čech complex of points sampled from a circle*

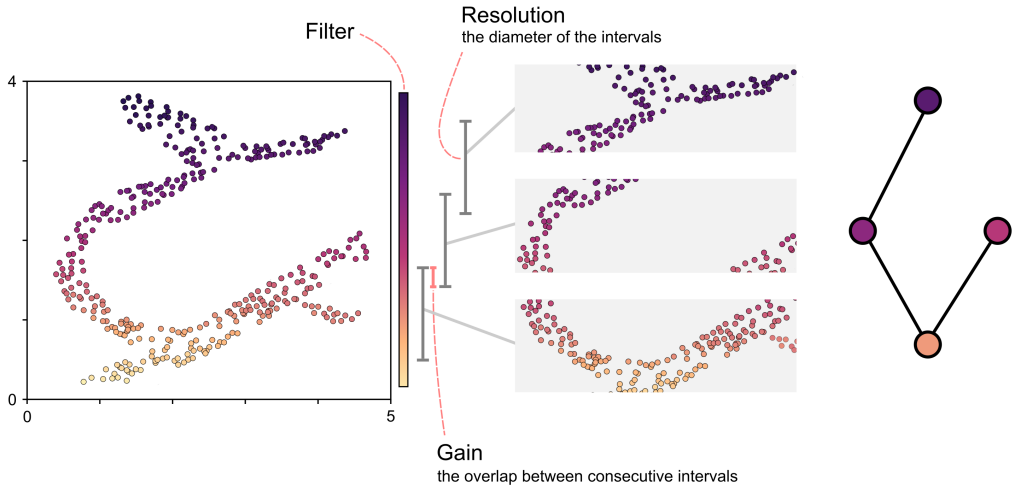


Figure 5. *The mapper graph: This shows the point cloud separated into intervals with diameter set by 'resolution' and overlap by 'gain'. Figure adapted from Munch (2017)*

2. Divide the range into smaller, overlapping intervals;

3. For each interval, find the set of data points whose values assigned by the filter function lie in the interval;

4. Decompose each of these sets into clusters based on a chosen clustering algorithm[2];

5. Represent each cluster by a node and connect nodes by an edge if the clusters intersect non-trivially, that is, they share data points.

The algorithm leaves various important choices to the user: the choice of the filter, the number of intervals and their percentage of overlap, and the clustering algorithm. See Chazal (2016) and Carrière, Michel and Oudot (2018) for a formal and complete discussion on parameter selection for Mapper. Several past studies have suggested the approach of selecting Mapper parameters based on exploration of a grid of possible values — selecting values that produce interesting or stable graphs (Carrière et al., 2018). However, as emphasised by Carrière et al. (2019), while useful for a data-driven exploratory phase, in many situations this approach may produce sub-optimal results, especially for non-trivial datasets. An alternative approach (Carrière, 2019) is to perform automatic tuning of Mapper parameters based on the rate of convergence of the Mapper graph to its continuous analogue, the Reeb graph. We return to this below.

[2]Any suitable clustering algorithm can be used; Refined analysis on the influence of the clustering method on the Mapper has been recently investigated (Belchí et al., 2019).

2.4. The persistence diagram

Recall our original goal is to obtain topological summaries that can describe complex structure in our data. Having constructed graphs based on the point cloud as described above, we can now use these graphs to compute topological invariants in our data (Chazal and Michel, 2021). One way to extract topological information is considering a family of simplicial complexes and coding the topological invariants in a two-dimensional diagram called the *persistence diagram* (Edelsbrunner et al., 2000). One can also identify the topological variants in a Mapper graph and represent them by means of the *extended-persistence diagram* (Carrière, 2019). Let us introduce both approaches below.

2.5. Persistence diagram for simplicial complexes

Let us consider a family of complexes constructed over an increasing range of values of the radius ρ (see Figure 4). This gives a *filtered complex*, a sequence of complexes such that each one is contained in the next. For each complex, we can deduce its topological invariants and trace them through the filtration as ρ increases, thus identifying their 'birth' (the radius at which they first appear) and 'death' (the radius at which they disappear). In Figure 4, the initial Čech complex for radius $\rho = 0.3$ and 0.8 shows no hole. A hole appears at $\rho = 1.5$ (the birth time of the hole) but disappears at $\rho = 5.0$ (the death time of the hole). So birth and death times represent radiuses at which the hole appears and disappears across the range of ρ values. The persistence diagram is a two dimensional plot where the X axis represents the birth time of a topological feature (a hole, in the example) and the Y axis represents its death time. The diagram includes a diagonal that represents the features that are born and die at equal time. The closer a point in the diagram is from the diagonal, the shorter was the life of that feature across the range of ρ values, i.e. the less persistent was the feature.

Intuitively, *persistent homology* captures how topological features of a space persist through the filtration, for some given time-span. The term homology refers to a mathematical (vector) space that represent the topological invariants in different dimensions. The homology group in dimension 0 represents the connectedness of the data space – a topological space is connected if it cannot be represented as the union of two or more disjoint non-empty open subsets. In dimension 1 the homology group represents the space of holes. In dimension 2 it represents the space of cavities, like the one we see in the torus or the 'bubble' inside the sphere, and so on (See Hatcher (2002) for a comprehensible introduction to homology). By identifying persistent features across a range of radiuses one avoids the need to choose a single radius ρ that would reveal the 'essential' topological features of the space. This ρ exists, and is mathematically proven, thanks to the combination of the nerve theorem and the reconstruction theorem (see Chazal and Michel (2021) for a formal formulation of both theorems). From a practical perspective, computing ρ rises many practical issues; a multiscale strategy has been introduced in (Chazal and Oudot, 2008).

Persistence diagrams of filtrations built on top of datasets are very stable with respect to some perturbations of the data. Thus, even for a dataset with some noise, the

persistence diagram obtained from this data is approximately correct because it is close to the diagram we would have obtained from the noise-free data (because the Gromov-Hausdorff distance between both datasets is assumed to be small, see Chazal and Michel, 2021).

2.6. Persistence diagram for the Mapper graph

The Mapper graph built under an optimal selection of the parameters involved in the algorithm (i.e. the filter function, intervals covering the range of the image of the filter function, and their overlap) is a discrete and computable optimal estimator of its continuous counterpart, the Reeb graph (the Mapper graph is said to 'converge' onto the Reeb graph) (Carrière and Oudot, 2018). A Reeb graph is a mathematical object reflecting the evolution of the level sets of a real-valued function on a topological space that locally resembles Euclidean space (see a Reeb graph in Figure 6 (iii)). From the Mapper graph, we can derive the extended persistence diagram (Figure 6) by tracing up and down the Reeb graph to identify pairs of critical points that mark the beginning ('birth' time) and the end ('death' time) of a topological feature in the associated Reeb graph. For example, Figure 6 shows the birth and death times for trunks, branches, and holes.

2.7. Statistical stability of points in the persistence diagram

As mentioned above, points on a persistence diagram with very short time spans, i.e. those points located close to the diagonal (the line representing points with equal birth and death), indicate features that appear and disappear quickly and which are more likely to be noise. We therefore may wish to discard 'non-significant' points that are close to the diagonal. One approach to assessing 'closeness' is to use the bootstrap to estimate and draw confidence bands on the persistence diagram, along the diagonal. Significant topological features will lie outside the confidence bands, whilst non-significant features will lie close to the diagonal, within the confidence bands, helping to distinguish between signal and noise (see Figure 7b) (Chazal, 2016). The bootstrap is a popular re-sampling method to quantify uncertainty around sample statistics (e.g. to estimate confidence intervals around a mean). To derive the confidence interval for a persistence diagram we:

1. Generate B bootstrap samples by re-sampling with replacement from the original source data, and construct a persistence diagram for each sample.

2. We then derive the 'distance' between the original persistence diagram (built using the source data) and each bootstrapped persistence diagram using the *Bottleneck distance*: two persistence diagrams are superimposed and each dot in the first diagram is assigned to its closest counterpart on the second. The Bottleneck distance is then defined as the maximum distance between any pair of matching dots. This way we get a distribution of distances for which a central 95% of values can be computed and a confidence interval (D) derived.

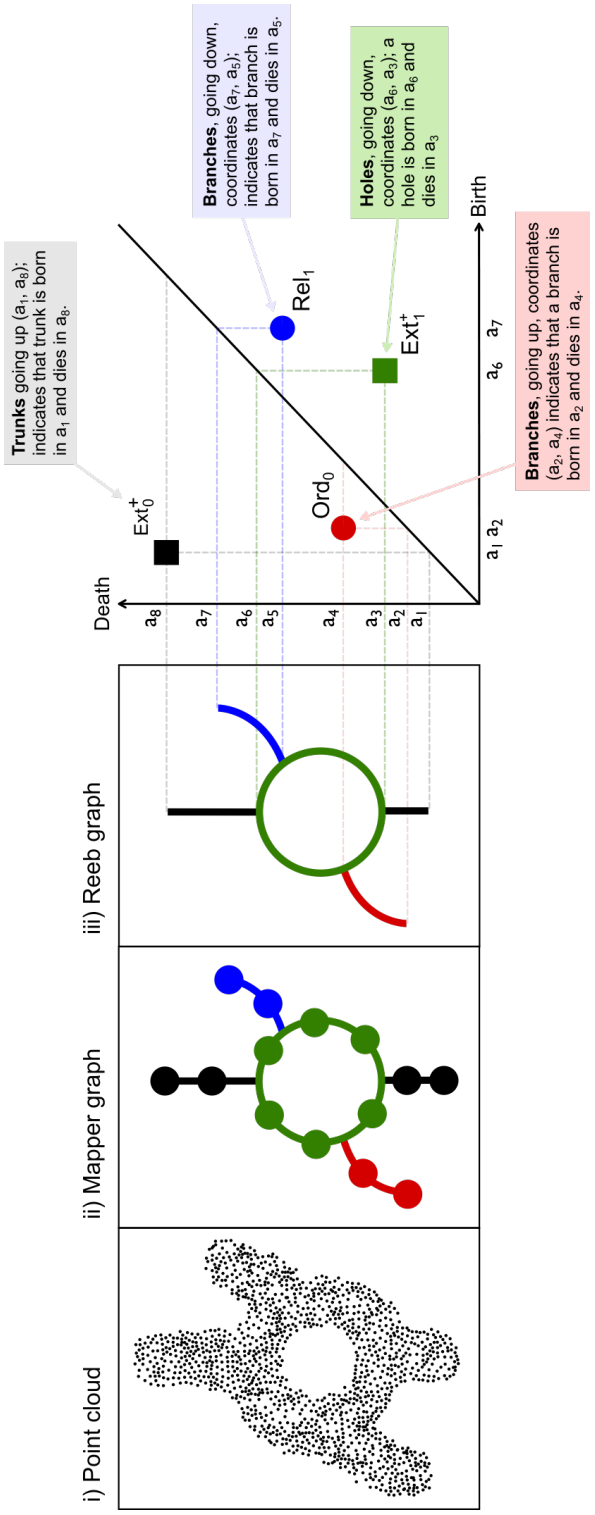


Figure 6. *Extended persistence diagram*

3. Finally, this confidence interval is drawn on the graph as a band spreading away from the diagonal (in both directions, each with width D), or as boxes around each point (of radius D).

The points outside the confidence band are considered as significant topological structures in the data, whereas those lying within the band's limits around the diagonal represent *insignificant* structures in the data set, and therefore are considered as noise and should not be interpreted nor processed for further analysis. This is a developing field, and while the validity of the use of the bottleneck bootstrap has been proven for the persistence diagram computed for a filtration of simplicial complexes, its use still remains as an open problem for the extended persistence diagrams computed for Mapper graphs (Carrière et al., 2018).

2.8. Use of Persistence Landscapes for outcomes prediction

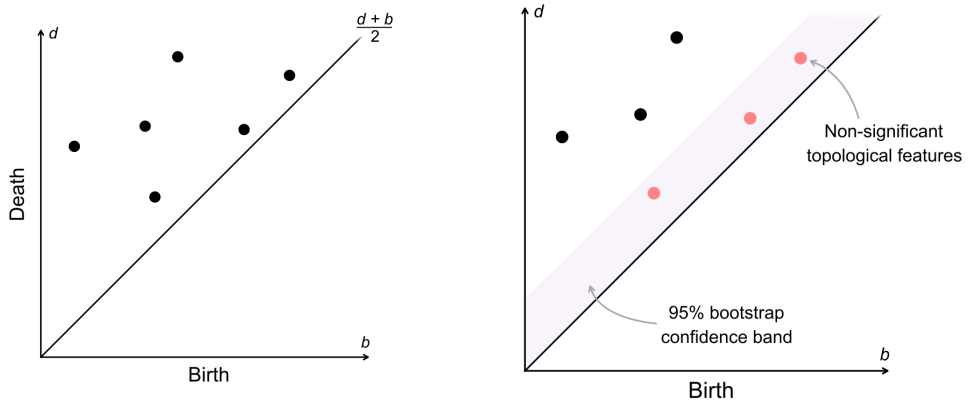
Persistence Landscapes (Bubenik, 2015) can be used to convert a persistence diagram (built from a filtration of simplicial complexes) into a vector space suitable for inclusion in ML models. Suppose we have a persistence diagram where each point represents the birth and death of a hole in our data. The corresponding persistence landscapes are constructed by 'tenting' each point in the diagram as shown in Figure 7c, to produce a collection of continuous *piecewise* linear functions, i.e. functions whose graph is composed of straight-line sections. Discretising the landscapes in a number of points produce a set of variables that encode the topological structure of data and can be included as predictors in a ML model. Interestingly, persistence landscapes share the same stability properties as persistence diagrams, described above.

2.9. Use of Mapper for subgroups detection, variable selection and data visualisation

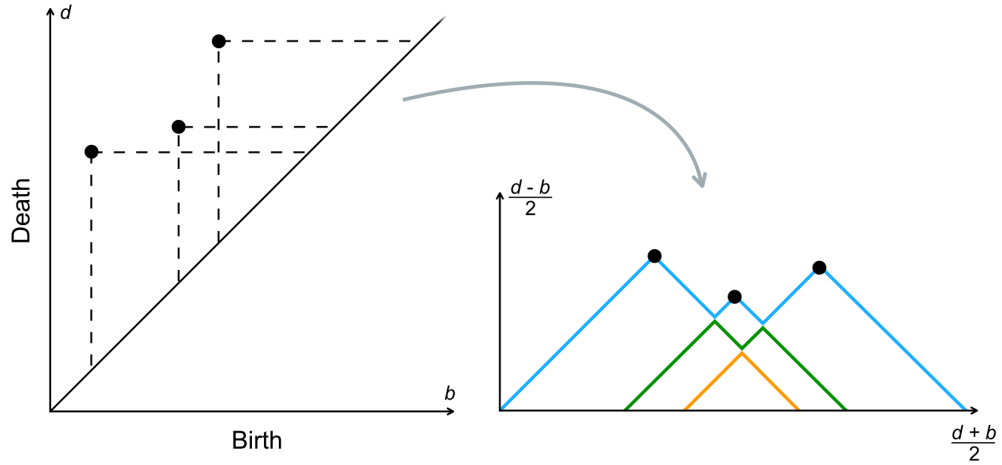
The Mapper graph can be useful to identify homogeneous subgroups of patients with regards of a characteristic of interest (Carr et al., 2021). The Mapper algorithm can highlight interesting clusters in data that might not be recoverable with traditional statistical clustering methods. Consider a data-based Mapper graph following the *flare* shape (the Y shape mentioned earlier; Figure 2). This could be interpreted as a single cluster of data. However, each arm could potentially represent a distinct data sub-population. The characterisation of topological features in a graph, like particular flares or loops, can help identify clinically relevant groups of nodes comprising subjects that experience particular prognostic outcomes or levels of treatment response.
Mapper can also be used to perform variable selection. One can build a Mapper graph from data, identify interesting structures as flares, loops or distinguished groups of coloured nodes, and then select the variables that best discriminate the data in these structures. Variables can then be assessed one-by-one for their ability to discriminate the potential sub-populations from the rest of the data using classical tests, as Kolmogorov-Smirnov. Interestingly, one can also consider a multivariate feature selection for which

(a) *The persistence diagram for a single point cloud*

(b) *Persistence diagram, showing the bootstrapped confidence interval. Points outside the interval are considered statistically significant.*



(c) *The computed persistence landscape, formed by 'tenting' the significant points on the persistence diagram. The first landscape is in blue, the second in green, and the last in orange.*



(d) *'Discretising' each landscape on a number of points, by selecting a discrete grid of values on the X-axis, and computing their corresponding Y-value on each persistence landscape.*

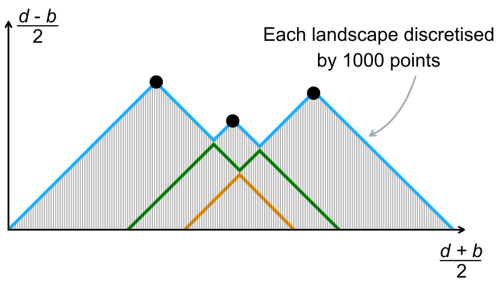


Figure 7. *Constructing the persistence landscape based on significant topological features*

Mapper can be used in conjunction with ML. Those detected flares and loops are given class labels, and a ML model including the desired set of predictors is tuned to solve the classification task of distinguishing one class from the other. This way, Mapper achieves two goals: identifying new sub-populations and selecting the combination of features that best differentiate them. We have implemented this ML based procedure to identify interesting subgroups and features in a pipeline that we will use below (https://github.com/kcl-bhi/mapper-pipeline).

The Mapper graph is also useful as a visualisation tool. If we select a set of intervals where no more than two intervals can intersect at once, Mapper becomes a visualisation tool that reflects the topology of the data. Mapper has a *multi-resolution structure*, i.e. by choosing the number of intervals and the percentage overlap between them, the user can adjust the level of the detail at which to view their data.

3. Application to a data case: using TDA to characterise depression remission in the GENDEP study

The Genome-based Therapeutic Drugs for Depression (GENDEP) is a pharmacogenetic study of antidepressant treatment response (Uher et al., 2010a). The GENDEP study aims to find a way to use clinical and genetic information about patients to help doctors decide which antidepressant treatment will work best for each patient, and with the least side-effects. A total of 220 patients were randomly allocated to be treated with escitalopram drug, a standard drug that is commonly prescribed to treat depressive symptoms. Over 12 weeks the study collected sociodemographic and clinical data including depressive symptoms. For each participant there were available sociodemographic variables (at baseline only) as age, age at onset, gender, smoking (yes/no), occupation (yes/no), partner (yes/no), years of education, number of children and body mass index. There were also available weekly repeated measures (from baseline to week 12) of depression severity by means of several standard scales: MADRS (Montgomery and Åsberg, 1979), Hamilton-17 (Hamilton, 1967), BDI (Beck et al., 1961), SCAN (Wing et al., 1990) and suicidal ideation (Perroud et al., 2012). Each scale assessed several individual items and was coded as a number (between 4 and 6, depending on the scale) of possible answers to a statement or question that allows respondents to indicate their positive-to-negative strength of agreement or strength of feeling regarding the question or statement. For example, the MADRS included 10 items assessing aspects such (1) apparent sadness; (2) reported sadness; (3) inner tension; (4) reduced sleep; (5) reduced appetite; (6) concentration difficulties; (7) lassitude; (8) inability to feel; (9) pessimistic thoughts; and (10) suicidal thoughts. Then each of these items was measured following a numerical codification ranging from 0 to 6 depending on the patient's strength of agreement. The ten resulting scores were then added to build a total numerical score. The rest of the scales were defined similarly. Data also included six symptoms dimensions (mood, anxiety, pessimism, interest-activity, sleep, appetite) from a published factor analysis (Uher et al., 2008, 2012). Remission was assessed for each patient at the last available mea-

surement after 4 – 12 weeks of treatment. Remission was defined as scoring $\leqslant 7$ on the Hamilton-17 scale (Hamilton, 1967), a commonly used definition for remission of depressive symptoms. A total of 94 patients remitted.

3.1. An analytical pipeline to predict remission depression

We aimed to use sociodemographic and clinical repeated measures in GENDEP to predict remission of depression. We implemented an analytical pipeline to compute persistence landscapes (and thus summaries of topological features) of our data, and including them in a ML model to predict remission. The pipeline requires Python 3.6 or higher (van Rossum, 1995) and R 4.1.2 or higher (R Core Team, 2020). It uses Scikit learn (Pedregosa et al., 2011) and Gudhi (The GUDHI Project, 2015) Python packages to derive the topological features based on the construction of persistence landscapes, and caret (Kuhn, 2008) and glmnet (Friedman, Hastie and Tibshirani, 2010) R packages to fit an elastic net logistic regression model that includes the topological features as predictors of a binary outcome. The pipeline can be freely downloaded at http://github.com/kcl-bhi/topological-review.

We used the pipeline to compute topological summaries on longitudinal measures of depression severity from baseline up to week 4 (a total of 5 time points). We included weekly total scores for MADRS, Hamilton-17, BDI and suicidal ideation, and a composite score for suicidal ideation[3]. We additionally included observed mood, cognitive and neurovegetative symptoms measured by means of the SCAN interview and the six symptoms dimensions from Uher et al. (2008, 2012) giving a total of 14 items measured on 5 occasions (a 14×5 matrix for each participant).

The detailed analytical pipeline we used was:

1. For each patient, compute the persistence diagram for a complex filtration based on the available data matrix (Figure 7a). For this case we computed an Alpha Complex filtration based on a 14×5 matrix (14 points in dimension 5). We considered the *connected components* and *holes* from the complex and created the associated persistence diagrams.

2. Compute the persistence landscape for each persistence diagram (Figure 7c). For this example, we computed the first three landscapes. The choice of how many landscapes to include can be guided by the predictive performance of the model (i.e. select the number of landscapes that maximises the predictive ability).

3. Discretise each landscape on a number of points (i.e. consider a discrete grid of values on the X-axis, and their corresponding Y-value on each persistence landscapes) (see Figure 7d). In our example, we considered the values of each persistence landscape on a grid of 1000 equidistant points, so that each patient was

[3]Composite scores are combinations of items that are highly related. They are computed from data in multiple variables in order to form reliable and valid measures of latent, theoretical constructs. These can be tested through factor analysis and reliability analysis (Ioannidis, Klavans and Boyack, 2016).

described by 6000 topological variables (3 landscapes $\times$ 1000 points $\times$ 2 dimensions). As one increases the number of discretisation points, the discretisation error will decrease but the resulting number of variables will increase. This decision should be based on the sample size available, although variable selection can help.

4. Include the topological variables together with the baseline variables as predictors in an elastic net logistic regression model to predict remission (yes or not). We chose a regularised regression model as this is efficient in preventing the risk of overfitting in complex data (i.e., when the model predicts well in known data, but generalises poorly to new cases) and performs variable selection, which helps to remove from the model topological variables that are not adding relevant information.

Parameter tuning for the elastic net regression model was performed with repeated (100 repetitions) 10-fold cross-validation. We compared (1) a model including sociodemographic and clinical variables only at baseline, and (2) a model including baseline sociodemographic and clinical variables and the topological variables derived from longitudinal measures on depression severity up to week 4, as described in the pipeline.

3.2. Results

In this preliminary analysis, the area under the ROC curve (AUC) for predicting remission was 0.746 for the model only including baseline measurements, which compared to an AUC of 0.799 when topological variables were added. This represented a promising improvement in predictive performance resulting from the inclusion of topological variables. Interestingly, the automated feature selection by the elastic net selected topological variables for both dimensions, that is, connected components and holes, as relevant variables for the prediction. The first landscapes tended to capture the most topological information, with subsequent landscapes bringing diminishing returns.

3.3. Using Mapper for subgroups detection in GENDEP

We used the Mapper algorithm to derive interesting clusters of patients in GENDEP based on their clinical and genetic baseline characteristics (full results are presented at Carr et al., 2021). We implemented a pipeline to tune the parameters of the Mapper graph seeking to maximise the purity of a given outcome variable within derived clusters of patients. A cluster of patients was defined as those patients with data belonging to a topological feature identified in the Mapper graph (i.e., a flare, a loop...). Mapper parameters were tuned to maximise the within clusters' level of purity with regards of depression remission (purity was computed by means of the Gini coefficient). Our pipeline allows predicting membership to a cluster using gradient boosted trees (XGBoost). This way it allows selecting the combination of variables that best differentiate a cluster of patients. The protocol also allows to consider both categorical and continuous variables (recent research in COVID-19 indicated high demand of such type of algorithms that are

suitable for mixed data types, see Khan et al., 2021). The pipeline can be freely downloaded under the GNU GPLv3 license at https://github.com/kcl-bhi/mapper-pipeline.

As it is shown in detail in Carr et al. (2021) when we applied our pipeline to the GENDEP dataset remission purity increased in the resulting clusters in comparison with the whole sample. We ranked the resulting clusters according to their remission purity, and, interestingly, the top five clusters from our pipeline outperformed the five-cluster solution from k-means clustering in terms of remission purity. Gini index in our clusters ranged from 0.30 to 0.38, whilst in clusters from k-means ranged from 0.33 to 0.50. A combination of clinical and genetic baseline measurements was able to discriminate patients in one of our top clusters with excellent discrimination.

4. TDA software

In practice, there are various algorithms implementing methods to produce simplicial complexes (the Čech complex and others) and compute topological invariants such as persistence diagrams and persistence landscapes. A good summary of software to implement persistent homology is given in Otter et al. (2017). There exist several general purpose libraries for topological data analysis including GUDHI (The GUDHI Project, 2020), Dionysus (Morozov, 2007), and PHAT (Bauer et al., 2017). All are written in `C++` and provide fast and efficient implementations of common topological invariants, with interfaces available for R and Python. Several packages have built upon these libraries to facilitate the application of common topological algorithms. The TDA package for `R` (Fasy et al., 2014) provides a user-friendly interface for `R` users. The `statmapper` (Carrière, 2020) Python package functions to derive extended persistence diagrams, to compute topological features in a Mapper graph and evaluate their statistical significance, using the bootstrap.

We have presented a pipeline that allows including summaries of topological features in a ML predictive model using persistence landscapes (http://github.com/kcl-bhi/topological-review) and a pipeline to identify sub-populations and perform multivariable selection using Mapper (https://github.com/kcl-bhi/mapper-pipeline).

5. Conclusion

TDA is a rapidly growing field that offers a unique set of tools with considerable potential for precision medicine. Topological summaries derived from persistence diagrams and landscapes have shown promising results in specific examples when included in machine learning predictive models, resulting in improved model performance, as we show in an application to a clinical trial on major depression. The Mapper algorithm makes it possible to identify homogeneous sub-populations of interest in complex data and deriving features that can be used to discriminate these groups. This paper provides a basis for the promising role that TDA can play in precision medicine using large biomedical datasets.

Acknowledgements

This work was supported by a 2017 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation granted to Dr Raquel Iniesta.
This work has been funded by the European Commission Framework 6 grant, EC Contract LSHB-CT-2003-503428 and an Innovative Medicine Initiative Joint Undertaking (IMI-JU) grant n-115008 of which resources are composed of European Union and the European Federation of Pharmaceutical Industries and Associations (EFPIA) in-kind contribution and financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013).
This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The authors further acknowledge use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London & Maudsley and Guy's & St. Thomas' NHS Foundation Trusts, and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's & St. Thomas' Charity (TR130505). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

References

Adamson, A. S. and Welch, H. G. (2019). Machine Learning and the Cancer-Diagnosis Problem — No Gold Standard. *New England Journal of Medicine*, 381(24):2285–2287.

Bauer, U., Kerber, M., Reininghaus, J., and Wagner, H. (2017). Phat – Persistent Homology Algorithms Toolbox. *Journal of Symbolic Computation*, 78:76–90.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6):561–571.

Belchí, F., Brodzki, J., Burfitt, M., and Niranjan, M. (2019). A numerical measure of the instability of mapper-type algorithms. *Journal of Machine Learning Research*, 21:1–45.

Bubenik, P. (2015). Statistical Topological Data Analysis using Persistence Landscapes. page 26.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Carr, E., Carrière, M., Michel, B., Chazal, F., and Iniesta, R. (2021). Identifying homogeneous subgroups of patients and important features: a topological machine learning approach.

Carrière, M. (2019). MathieuCarriere/Sklearn-Tda.

Carrière, M. (2020). MathieuCarriere/Statmapper.

Carrière, M., Michel, B., and Oudot, S. (2018). Statistical Analysis and Parameter Selection for Mapper. *Journal of Machine Learning Research*, 19(1):1–39.

Carrière, M. and Oudot, S. (2017). Local Equivalence and Induced Metrics for Reeb Graphs. In *Proceedings of the 33rd Symposium on Computational Geometry*.

Carrière, M. and Oudot, S. (2018). Structure and Stability of the 1-Dimensional Mapper. *Foundations of Computational Mathematics*, 18(6):1333–1396.

Chartrand, G. (1985). *Introductory Graph Theory*. Dover Publications Inc., New York, abridged edition edition edition.

Chazal, F. (2016). High-Dimensional Topological Data Analysis. In *3rd Handbook of Discrete and Computational Geometry*. CRC Press.

Chazal, F., Massart, P., and Michel, B. (2016). Rates of convergence for robust geometric inference. *Electronic Journal of Statistics*, 10(2):2243–2286.

Chazal, F. and Michel, B. (2021). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4.

Chazal, F. and Oudot, S. Y. (2008). Towards persistence-based reconstruction in euclidean spaces. In *Proceedings of the Twenty-Fourth Annual Symposium on Computational Geometry - SCG '08*, page 232, College Park, MD, USA. ACM Press.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., and Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, 3(3):243–250.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of Persistence Diagrams. *Discrete & Computational Geometry*, 37(1):103–120.

Dagliati, A., Geifman, N., Peek, N., Holmes, J. H., Sacchi, L., Bellazzi, R., Sajjadi, S. E., and Tucker, A. (2020). Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records. *Artificial Intelligence in Medicine*, 108:101930.

Devillers, O., Hornus, S., and Jamin, C. (2022). dD triangulations. In *CGAL User and Reference Manual*. CGAL Editorial Board, 5.4 edition.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2000). Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463.

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., Hickey, A. J., and Clark, A. M. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials*, 18(5):435–441.

Fasy, B. T., Kim, J., Lecci, F., and Maria, C. (2014). Introduction to the R package TDA. *arXiv:1411.1830 [cs, stat]*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Ghrist, R. (2018). Homological algebra and data. In Mahoney, M., Duchi, J., and Gilbert, A., editors, *IAS/Park City Mathematics Series*, volume 25, pages 273–325. American Mathematical Society, Providence, Rhode Island.

Hamilton, M. (1967). Development of a Rating Scale for Primary Depressive Illness. *British Journal of Social and Clinical Psychology*, 6(4):278–296.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Number 1 in Springer Series in Statistics. Springer, New York.

Hatcher, A. (2002). *Algebraic Topology*. Cambridge University Press.

Henle, M. (1994). *A Combinatorial Introduction to Topology*. Dover Books, Inc, New York.

Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., and O'Sullivan, J. (2019). Machine Learning SNP Based Prediction for Precision Medicine. *Frontiers in Genetics*, 10:267.

Iniesta, R., Hodgson, K., Stahl, D., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M. Z., Souery, D., Dobson, R., Aitchison, K. J., Farmer, A., McGuffin, P., Lewis, C. M., and Uher, R. (2018). Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Scientific Reports*, 8(1):1–9.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M. Z., Souery, D., Stahl, D., Dobson, R., Aitchison, K. J., Farmer, A., Lewis, C. M., McGuffin, P., and Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research*, 78:94–102.

Iniesta, R., Stahl, D., and McGuffin, P. (2017). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, 46(12):2455–2465.

Ioannidis, J., Klavans, R., and Boyack, K. (2016). Multiple citation indicators and their composite across scientific disciplines. *PLoS Biol*, 14(7).

Khan, W., Crockett, K., O'Shea, J., Hussain, A., and Khan, B. M. (2021). Deception in the eyes of deceiver: A computer vision and machine learning based automated deception detection. *Expert Systems with Applications*, 169.

Khan, W., Hussain, A., Ahmed Khan, S., Al-Jumailey, M., Raheel, N., and Liatsis, P. (2019). Analysing the impact of global demographic characteristics over the covid-19 spread using class rule mining and pattern matching. *Royal Society Open Science*, 8.

Kosniowski, C. (1980). *A First Course in Algebraic Topology*. CUP Archive.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw-Hill, New York.

Mitchell, T. M. (2006). The Discipline of Machine Learning. page 9.

Montgomery, S. A. and Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, 134:382–389.

Morozov, D. (2007). Dionysus, a C++ library for computing persistent homology.

Müllner, D. and Babu, A. (2013). Python Mapper: An open-source toolchain for data exploration.

Munch, E. (2017). A User's Guide to Topological Data Analysis. *Journal of Learning Analytics*, 4(2):47–61.

Nature (2019). Ascent of machine learning in medicine. *Nature Materials*, 18(5):407–407.

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.

Nielson, J. L., Cooper, S. R., Yue, J. K., Sorani, M. D., Inoue, T., Yuh, E. L., Mukherjee, P., Petrossian, T. C., Paquette, J., Lum, P. Y., Carlsson, G. E., Vassar, M. J., Lingsma, H. F., Gordon, W. A., Valadka, A. B., Okonkwo, D. O., Manley, G. T., Ferguson, A. R., and TRACK-TBI Investigators (2017). Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis. *PLOS ONE*, 12(3):e0169490.

Nusrat, S., Harbig, T., and Gehlenborg, N. (2019). Tasks, Techniques, and Tools for Genomic Data Visualization. *Computer Graphics Forum*, 38(3):781–805.

Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perroud, N., Uher, R., Ng, M. Y. M., Guipponi, M., Hauser, J., Henigsberg, N., Maier, W., Mors, O., Gennarelli, M., Rietschel, M., Souery, D., Dernovsek, M. Z., Stamp, A. S., Lathrop, M., Farmer, A., Breen, G., Aitchison, K. J., Lewis, C. M., Craig, I. W., and McGuffin, P. (2012). Genome-wide association study of increasing suicidal ideation during antidepressant treatment in the GENDEP project. *The Pharmacogenomics Journal*, 12(1):68–77.

Qu, Z., Lau, C. W., Nguyen, Q. V., Zhou, Y., and Catchpoole, D. R. (2019). Visual Analytics of Genomic and Cancer Data: A Systematic Review. *Cancer Informatics*, 18:117693511983554.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Riemann, B. and Clifford, T. W. K. (1998). On the Hypotheses which lie at the Bases of Geometry. page 15.

Sies, A., Demyttenaere, K., and Mechelen, I. V. (2019). Studying treatment-effect heterogeneity in precision medicine through induced subgroups. *Journal of Biopharmaceutical Statistics*, 29(3):491–507.

Singh, G., Mémoli, F., and Carlsson, G. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *PBG@Eurographics*.

The GUDHI Project (2015). *GUDHI User and Reference Manual*. GUDHI Editorial Board.

The GUDHI Project (2020). *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.1.1 edition.

Traylor, M., Markus, H., and Lewis, C. M. (2015). Homogeneous case subgroups increase power in genetic association studies. *European Journal of Human Genetics*, 23(6):863–869.

Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., Mors, O., Elkin, A., Williamson, R. J., Schmael, C., Henigsberg, N., Perez, J., Mendlewicz, J., Janzing, J. G. E., Zobel, A., Skibinska, M., Kozel, D., Stamp, A. S., Bajs, M., Placentino, A., Barreto, M., McGuffin, P., and Aitchison, K. J. (2008). Measuring depression: Comparison and integration of three scales in the GENDEP study. *Psychological Medicine*, 38(2):289–300.

Uher, R., Muthén, B., Souery, D., Mors, O., Jaracz, J., Placentino, A., Petrovic, A., Zobel, A., Henigsberg, N., Rietschel, M., Aitchison, K. J., Farmer, A., and McGuffin, P. (2010a). Trajectories of change in depression severity during treatment with antidepressants. *Psychological Medicine*, 40(8):1367–1377.

Uher, R., Perlis, R. H., Henigsberg, N., Zobel, A., Rietschel, M., Mors, O., Hauser, J., Dernovsek, M. Z., Souery, D., Bajs, M., Maier, W., Aitchison, K. J., Farmer, A., and McGuffin, P. (2012). Depression symptom dimensions as predictors of antidepressant treatment outcome: Replicable evidence for interest-activity symptoms. *Psychological medicine*, 42(5):967–980.

Uher, R., Perroud, N., Ng, M. Y., Hauser, J., Henigsberg, N., Maier, W., Mors, O., Placentino, A., Rietschel, M., Souery, D., Žagar, T., Czerski, P. M., Jerman, B., Larsen, E. R., Schulze, T. G., Zobel, A., Cohen-Woods, S., Pirlo, K., Butler, A. W., Muglia, P., Barnes, M. R., Lathrop<, M., Farmer, A., Breen, G., Aitchison, K. J., Craig, I., Lewis, C. M., and McGuffin, P. (2010b). Genome-Wide Pharmacogenetics of Antidepressant Response in the GENDEP Project. *American Journal of Psychiatry*, 167(5):555–564.

van Rossum, G. (1995). Python reference manual.

Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R., Jablenski, A., Regier, D., and Sartorius, N. (1990). SCAN: Schedules fonr Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry*, 47(6):589–593.

Zomorodian, A. (2010). Fast construction of the Vietoris-Rips complex. *Computers & Graphics*, 34(3):263–271.

Zomorodian, A. and Carlsson, G. (2005). Computing Persistent Homology. *Discrete & Computational Geometry*, 33(2):249–274.