The shape of cancer relapse: Topological data analysis predicts recurrence in paediatric acute lymphoblastic leukaemia

Salvador Chulián[1,2,*], Bernadette J. Stolz[3,4], Álvaro Martínez-Rubio[1,2], Cristina Blázquez Goñi[2,5,6], Juan F. Rodríguez Gutiérrez[2,5], Teresa Caballero Velázquez[7], Águeda Molinos Quintana[6], Manuel Ramírez Orellana[8], Ana Castillo Robleda[8], José Luis Fuster Soler[9], Alfredo Minguela Puras[10], María V. Martínez Sánchez[10], María Rosa[1,2], Víctor M. Pérez-García[11,12,13], Helen M. Byrne[3,*]

1 Department of Mathematics, Universidad de Cádiz, 11510 Puerto Real, Cádiz, Spain
2 Biomedical Research and Innovation Institute of Cádiz (INiBICA), Hospital Universitario Puerta del Mar, 11009 Cádiz, Spain
3 Mathematical Institute, University of Oxford, OX2 6GG Oxford, United Kingdom
4 Laboratory for Topology and Neuroscience, EPFL, Station 8, 1015 Lausanne, Switzerland
5 Department of Paediatric Haematology and Oncology, Hospital de Jerez, 11407 Jerez de la Frontera, Spain
6 Department of Haematology, Hospital Universitario Vírgen del Rocío, 41013 Sevilla, Spain
7 Department of Haematology, Hospital Universitario Vírgen del Rocío / University of Sevilla, 41013 Sevilla, Spain
8 Department of Paediatric Haematology and Oncology, Hospital Infantil Universitario Niño Jesús, Instituto Investigación Sanitaria La Princesa, 28009 Madrid, Spain
9 Department of Paediatric Haematology and Oncology, Hospital Clínico Universitario Virgen de la Arrixaca, Instituto Murciano de Investigación Biosanitaria (IMIB), 30120 Murcia, Spain
10 Immunology Service, Hospital Clínico Universitario Virgen de la Arrixaca, Instituto Murciano de Investigación Biosanitaria (IMIB), 30120 Murcia, Spain
11 Department of Mathematics, Mathematical Oncology Laboratory (MOLAB), Universidad de Castilla-La Mancha, 13005 Ciudad Real, Spain
12 Instituto de Matemática Aplicada a la Ciencia y la Ingeniería (IMACI), Universidad de Castilla-La Mancha, 13005 Ciudad Real, Spain
13 ETSI Industriales, Universidad de Castilla-La Mancha, 13005 Ciudad Real, Spain
***** Corresponding authors. Emails: salvador.chulian@uca.es, helen.byrne@maths.ox.ac.uk

Abstract

Although children and adolescents with acute lymphoblastic leukaemia (ALL) have high survival rates, approximately 15-20% of patients relapse. Risk of relapse is routinely estimated at diagnosis by biological factors, including flow cytometry data. This high-dimensional data is typically manually assessed by projecting it onto a subset of biomarkers. Cell density and "empty spaces" in 2D projections of the data, i.e. regions devoid of cells, are then used for qualitative assessment. Here, we use topological data analysis (TDA), which quantifies shapes, including empty spaces, in data, to analyse pre-treatment ALL datasets with known patient outcomes. We combine these fully unsupervised analyses with Machine Learning (ML) to identify significant shape characteristics and demonstrate that they accurately predict risk of relapse, particularly

for patients previously classified as 'low risk'. We independently confirm the predictive power of CD10, CD20, CD38, and CD45 as biomarkers for ALL diagnosis. Based on our analyses, we propose three increasingly detailed prognostic pipelines for analysing flow cytometry data from ALL patients depending on technical and technological availability: 1. Visual inspection of specific biological features in biparametric projections of the data; 2. Computation of quantitative topological descriptors of such projections; 3. A combined analysis, using TDA and ML, in the four-parameter space defined by CD10, CD20, CD38 and CD45. Our analyses readily extend to other haematological malignancies.

Author summary

Acute lymphoblastic leukaemia (ALL) is a blood cancer which affects predominantly children and adolescents. Therapy typically fails in approximately 20 % of patients, who suffer from relapse. To determine disease status, clinicians assess cell types, their interactions, as well as deviations from normal behaviour. Flow cytometry (FC) is a method that quantifies the intensity of specific cell markers, giving rise to high-dimensional data. This routinely collected information is then reduced to obtain human-interpretable visualisation for prognosis. Topological Data Analysis (TDA) is a field of mathematics that studies shapes in data, considering isolated data islands and empty spaces between them. We showcase how to use TDA to extract shape characteristics in FC data of relapsing patients. We propose three pipelines, of increasing methodological complexity, to aid clinical decisions for risk stratification in ALL. In combination with Machine Learning, TDA enables high-accuracy predictions of relapse to be made at the time of diagnosis.

Introduction

Acute Lymphoblastic Leukaemia (ALL) is a haematological malignancy that affects all age groups, particularly children. Indeed, it is the most frequent type of paediatric cancer [1]. While current chemotherapy regimes have improved survival rates, as many as 15-20% of patients relapse following chemotherapy, i.e. the disease returns, even though the initial response to treatment may be good. Current protocols tailor the intensity of therapy to the risk of relapse, estimating this risk through a combination of biological information at diagnosis and the individual response to therapy. Patients at high risk of relapse receive more intensive chemotherapeutic regimes, before resistant clones emerge and become dominant [2, 3], improving survival rates. Very high risk patients are subjected to more aggressive therapies, such as bone marrow transplants or CAR-T-cell therapies [4]. Identifying patients at risk of relapse as early as possible and tailoring treatment plans accordingly is, therefore, of major clinical importance.

At diagnosis, clinicians typically collect a combination of quantitative biomolecular, and qualitative (usually binary) cell morphology and genetic data from a patient [5] to determine the type of haematological cancer, disease severity, i.e. prognosis, and likely response to treatment. Based on experience gained from multi-centre clinical trials, patient data is manually assessed to determine a patient's treatment plan, following an induction-consolidation-reinduction-maintenance chemotherapy regime, which is standard in many countries [6]. Many clinical trials also collect data about patient responses to treatment, including whether they experienced relapse. These data are ideally suited to retrospective analyses, not only to optimise treatment plans, but also to predict relapse. Exploiting existing datasets represents a promising approach to overcome many of the challenges in personalised medicine in oncology, such as patient

phenotyping for prognosis or biomarker discovery [7–9].

Flow cytometry (FC) data are routinely used for the diagnosis of haematological malignancies such as ALL [10]. These data provide quantitative information about the bone marrow status at the level of single cells. A particular challenge of FC data is its high-dimensionality: each cell from a patient sample is represented by one data point whose coordinates correspond to the intensities of a set of predetermined fluorescent biomarkers. This collection of data points is also referred to as a point cloud. The biomarkers usually comprise immunophenotypic (IPT) markers [11] such as CD19, CD10 and CD20, which are characteristic of B-lymphocytes subpopulations, more general markers such as CD45, CD38 and CD34, which indicate cell differentiation and maturation status, as well as CD3, CD33 or cyMPO, which are specific to other cell lineages, e.g. T lymphocytes or myeloid cells. The current strategy for managing the high-dimensionality in these data is to generate two-dimensional projections of pairwise combinations of biomarkers [12], and then to manually assess the shapes of the projected data. For example, the shapes can be inspected to identify subsets of points within the point cloud that correspond to cancerous cells. Indeed, previous studies have highlighted the importance of 'empty spaces' in FC data over the course of treatment [13–15].

However, manual assessment is time-consuming, subjective, and prone to errors. Therefore, more systematic and automated methods of analysis are required. In the context of FC data, Machine learning classifiers have been used to predict relapse in B-ALL patients, yet these studies focus on either minimal residual disease follow-up, i.e., the number of cancer cells after treatment [16], or on biomarker intensity levels in different B-ALL subpopulations [17]. Several methods have been developed for application in other diseases [18], mostly concentrating on dimensionality reduction and cell population identification. Others, such as CellCNN [19], CITRUS [20], and FloReMi [21] employ convolutional neural networks (CNN), LASSO, and Random Forest (RF) models, respectively, to perform classification. All these approaches, however, have one thing in common: they rely on the level of expression of cell markers, i.e. their fluorescence intensity values, for classification and feature extraction. The clinical approach, however, focusses on a higher-level description of the data: the shape, orientation and structure of the point cloud. Therefore, we propose the integration of topological biomarkers which characterise the shape of the data with machine learning methods. In this way, we can exploit the advantages of each approach to improve the value of these techniques in biomedical data analysis.

Topological data analysis (TDA) [22, 23] is a set of methods that uses ideas from the mathematical discipline topology to study the 'shape' of high-dimensional data. The most prominent method in TDA is persistent homology (PH) [22–26]. PH quantifies features, such as empty spaces, in point cloud data, at different spatial scales. Thus, in contrast to other data analysis methods, it is not necessary to predetermine a spatial scale at which relationships between data points are considered; PH takes into account the full range of possible scales. PH describes the empty spaces of data via topological features such as connected components (i.e. clusters of data points), and loops (i.e. data points surrounding an empty hole). The output from PH are barcodes which correspond to topological features in the data. Features that appear across a range of scales in these barcodes, are called 'persistent' and are often viewed as particularly characteristic of a dataset (for an example, see Fig. 1 (E) and (F)). The mathematical foundations of PH ensure that small changes in the data lead to only small changes in the barcodes. PH is readily computable [26], robust to noise [27] and its outputs are interpretable. In recent years, improved computational feasibility of PH has increased its applications to (high-dimensional) data in many contexts [28, 29], including studies of the shape of brain arteries [30], neurons [31], the neural code [32], airways [33], stenosis [34], zebrafish patterns [35], ion aggregation [36], contagion dynamics [37], proteins [38], spatial

networks [39–42, 44], and geometric anomalies [43]. In oncology, PH has been used to construct new biomarkers [40, 45, 46], to classify tumours [47, 48] and genetic alterations [49] and to quantify patterns of immune cell infiltration into solid tumours [50].

In this paper we design a novel pipeline to analyse FC data that combines PH and ML. TDA has been successfully applied to study COVID-19 in FC data [51]. We apply our pipeline to data collected at three different hospitals in Spain [52] from patients with the most common subtype of paediatric ALL [53], one that is characterised by rapid growth of early B lymphocytes (B-ALL). We demonstrate that our shape-based analysis enables us to predict relapse in ALL patients with very high accuracy (see Fig. 1) and we independently confirm the known biomarkers CD10, CD20, CD38 and/or CD45 as most relevant for relapse (R) versus non-relapse (NR) classification. We further identify specific topological features (connected components and loops) which characterise patients at risk of relapse in 2D projections. Finally, we highlight that by including vectorisation methods (persistence images [54]) our pipeline can make full use of the topological information available in a 4D projection of the FC data, and outperform the commonly used 2D projections in detecting patients at risk of relapse. Our approach can be readily generalised to many other haematological malignancies. Additionally, TDA provides an interpretative layer to ML techniques, which clinicians have traditionally identified as a bottleneck [55]. Further, this contributes to the recent advances in interpretable machine learning [56].

Based on our findings, we suggest three novel approaches of increasing complexity for assessment of risk of relapse in (B-ALL). The first approach is derived from our insights from PH and suggests shape characteristics that are specific to 2D projections of data from patients at risk of relapse which can be included in visual inspection when PH is not available. If PH is available for data analysis but not ML, we suggest a second approach where barcodes are summarised by computing how many connected components (topological features in dimensions 0) and loops (topological features in dimension 1) persist longer than a given range of thresholds. The third approach includes a combination of PH and ML and is fully automated. It achieves optimal accuracy in relapse prediction compared to manual assessments performed at diagnosis, and also in patients who with conventional methods were falsely predicted to have low risk for relapse. To our knowledge, this is the first study that quantifies (rather than qualitatively assesses) 'empty spaces' in FC data to predict relapse patients at the time of diagnosis.

Results

Machine learning classifier applied to TDA independently confirmed state-of-the-art markers for prognosis. We applied Machine Learning (ML) to Topological Data Analysis (TDA) descriptors of patient flow cytometry (FC) data. With our quantitative TDA descriptors, we independently confirmed the prognostic relevance of empty spaces in data projections of state-of-the-art biomarkers. The general pipeline used is shown in Figure 1, where we classified patients into either the non-relapsing (NR) group, i.e. patients who show no refractory values for minimal residual disease after three years, or the relapsing (R) group. Our data contained 80 NR patients and 16 R patients. We divided the datasets into discovery and validation sets, as described in 'Methods' and Table S1.

In line with standard clinical practice to assess prognosis, we constructed 2D projections of the high-dimensional point clouds, considering all pairwise combinations of the 16 immunophenotypic (IPT) markers for each patient (see Figure 2(A–C)). We then applied persistent homology (PH) (specifically, Vietoris-Rips complex, see

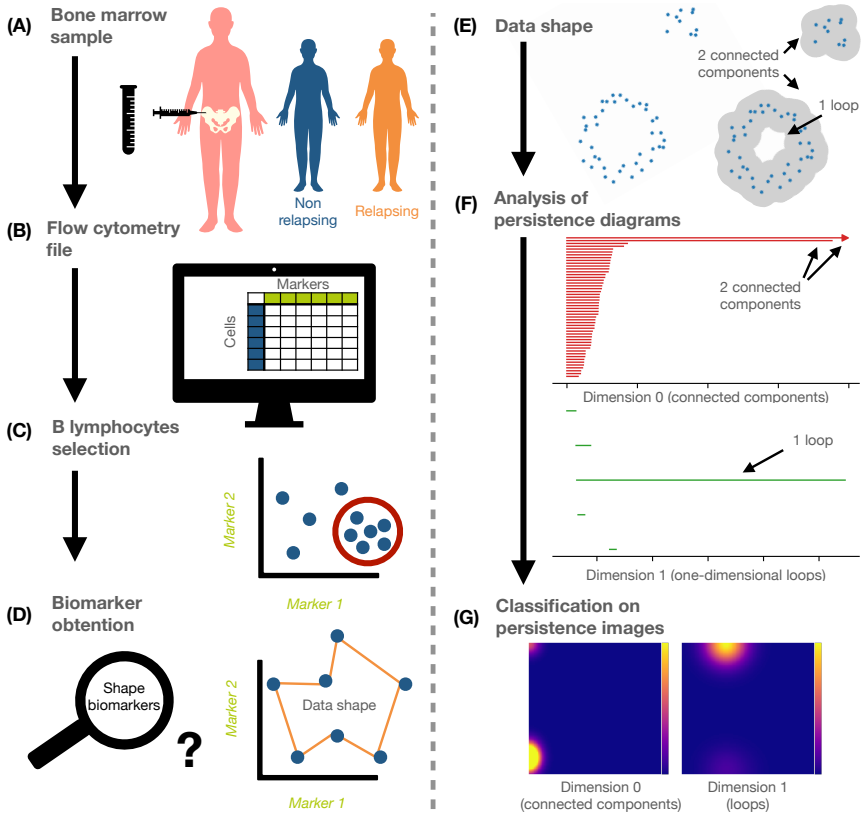

Fig 1. Flowchart of methodology. (A) We collected bone marrow data from relapsing (R) and non-relapsing (NR) B-ALL patients at the time of diagnosis. **(B)** These samples were analysed via flow cytometry (FC) which gave rise to single-cell data with information on immunophenotypic (IPT) markers. **(C)** We pre-processed these data and selected the patients' B lymphocytes. **(D)** Our aim was to obtain quantitative biomarkers from the IPT markers that can be used to identify patients at risk of relapse. **(E)** We conducted topological data analysis (TDA), which quantifies the shape of data via topological characteristics. **(F)** An important tool in TDA is persistent homology (PH), which can be used to obtain summaries in the form of persistence barcodes. Barcodes are topological signatures of the data and show topological features such as connected components and loops as well as their scale in the data. **(G)** Persistence barcodes can be transformed into vector summaries called persistence images. As a second approach, we used these images for classification.

'Methods') and obtained PH barcodes for dimension 0, i.e. connected components, and dimension 1, i.e. loops (see Figure 2(D); and, for a detailed description, see 'Methods'). In a first step, we extracted simple descriptors from each barcode to create vectors for each patient, for each of the pairwise combinations of markers (see Figure 2(E)).

We then performed a Random Forest analysis, with cross-validation, on each patient subgroup within the discovery set, assigning 60% of the patients to a training group, and 40% to a testing group. After classification, we excluded those IPT pairs with low accuracy (see 'Methods'), where the IPT markers that appeared most frequently in the remaining pairs were CD10, CD19, CD20, CD22, CD34, CD38, cyMPO and cyTdT (see Figure 2(F) and Data in SI). Only combinations including CD10 or CD20 had a mean accuracy over 70%.

Next, we considered the validation dataset to verify that the selected markers

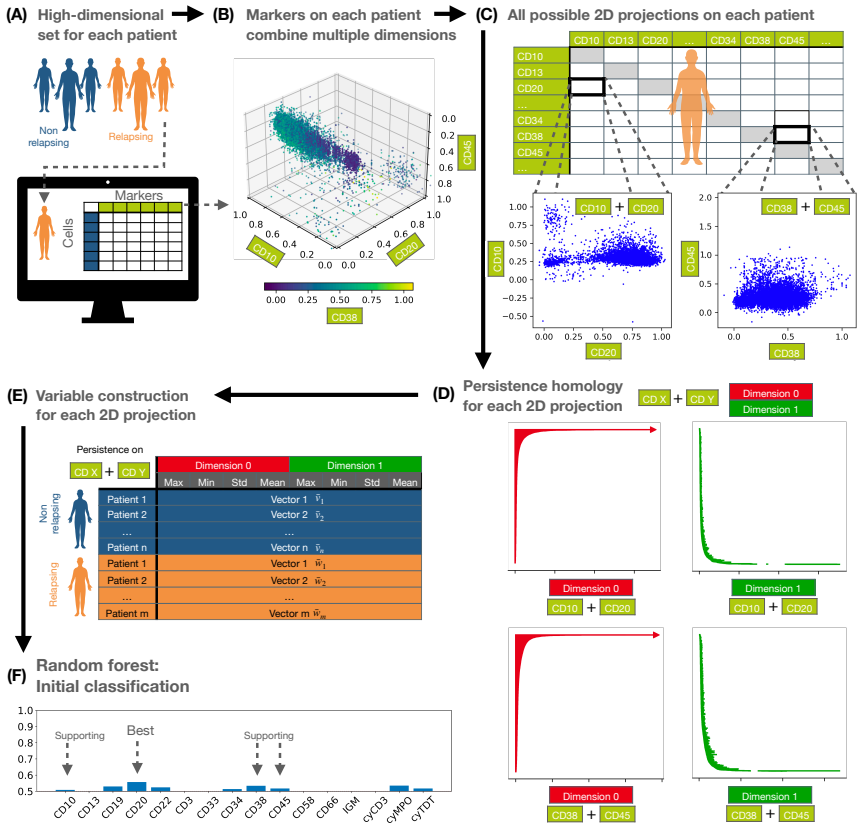

Fig 2. Workflow for confirming prognostic, state-of-the-art markers. **(A)** We collected data from non-relapsing (NR) and relapsing (R) patients containing 16 immunophenotypic (IPT) markers. **(B)** To reduce the dimensionality of the data and enable TDA computations, we first selected a subset of markers that contribute most to differences in the shape of the data between the R and NR groups. **(C)** We constructed pairwise projections of the selected IPT markers in line with standard clinical practice. **(D)** We applied PH to each patient point cloud, and quantified connected components and loops in the data. **(E)** For each pairwise combination of IPT markers associated with a particular patient, we computed an 8-dimensional feature vector summarising information from the PH analysis. **(F)** We applied Random Forest classification to the feature vectors and confirmed that the biomarkers CD10, CD20, CD8 and CD45 contain information that is most relevant for distinguishing R from NR patients based on the shape of their data.

provided relevant shape information. The set of pairwise combinations yielded an additional 5520 persistence barcodes in each dimension. As before, we studied these barcodes using Random Forest analysis, with k-fold cross-validation to check for robustness. When we used oversampling, our results showed no significant differences for different biomarkers (see Data S1). However, when we used the unbalanced data and stratified k-fold, i.e., maintaining the proportion of samples for each class, only IPT markers CD10, CD20, CD38 and CD45 produced mean AUCs>50% (see Data S2-Data S3). To establish whether these markers carried independent information, we also computed the correlations between the markers. None of the IPT markers showed strong correlations in either the NR or R groups (see Figures S6 and S7). Biologically, it is well-known that CD10, CD20 and CD45 identify subpopulations within the $CD19^{+}$ B

lymphocyte clouds [57], while CD38 is a marker of aberrant B-lymphoblasts [58, 59]. While the expression levels of these biomarkers have been used to predict B-ALL relapse [17], empty spaces in the data have not previously been quantified. Our independent Random Forest analysis, performed on simple topological descriptors, confirmed the importance of these biomarkers. We performed our subsequent analysis on a reduced dataset containing only these variables.

PH identifies connected components and loops characteristic of R patients. Following our selection of the four most informative biomarkers for prognosis (CD10, CD20, CD38 and CD45), our goal is to use PH to compute differences in the number of connected components and loops in the data point clouds from R and NR patients. First, we applied PH to all pairwise combinations of these biomarkers for each patient. We visualised the persistence of connected components (dimension 0) and loops (dimension 1) by constructing what we term *persistence threshold (PT) curves*; PT curves reveal characteristic topological features, i.e. persistent features, for each patient cohort (see 'Methods'). We averaged the PT curves for both patient cohorts and all pairwise combinations of biomarkers and found significant differences between the discovery and validation sets. The results for the mean PT curves for all pairwise combinations are presented in Figures S10-S15 for both dimensions 0 and 1, i.e. connected components and loops.

We present illustrative results for CD10-CD20 in Figure 3. In both the validation and discovery datasets, the mean values of the dimension 1 PT curves were larger in R patients than in NR patients (see Figure 3(A.2)). These differences were statistically significant in dimension 1, where τ is the persistence threshold with $\tau \in [0.04, 0.05]$ (see 'Methods' for other threshold ranges). An illustrative example for dimension 0 for CD10-CD20 is presented in Figure S9.

Intuitively, there are fewer persistent loops, i.e. empty spaces, in the NR patients than in the R patients as the loops in the NR patients tend to be 'narrower' and disappear at smaller spatial scales (radius $r = 0.03$ versus $r = 0.06$, see Figures 3(B.1) and (B.2), respectively). In the barcodes, less persistent loops correspond to shorter bars. This is captured by the PT curves. In dimension 0, we observed the same trends for CD10-CD20: the number of less persistent connected components in the data of R patients was larger than that of NR patients (see Appendix for details). Complete results for all pairwise combinations are included in Figures S14 and S15, respectively; they show the absolute and relative numbers of bars that are larger than a fixed threshold value for each patient. The number of persistent loops was higher in R patients for all combinations that included either CD10 or CD20, except for the combination CD10-CD45. We only found a higher number of persistent loops for NR patients in CD38-CD45. The number of non-persistent loops was only larger for NR patients than for R patients for the combination CD20-CD38. In dimension 0, connected components with short persistence were only found in R patients for the combination CD10-CD20. These results were statistically significant and consistent between discovery and validation datasets.

When we applied the same analysis to the 4-dimensional point clouds of each patient, the differences between the R and NR group were not significant in the whole dataset (see Appendix). As before, we first performed statistical analyses on simple barcode descriptors, i.e. the maximum, minimum and mean persistence and the standard deviation, extracted from the dimension 0 and 1 barcodes of the NR and R groups. Since the resulting p-values for all variables were larger than 0.05, we concluded that, in general, the simple barcode descriptors are unable to discriminate between the two groups (see results summary in Figure S8(A)).

Given the analogy between clusters and connected components, we performed

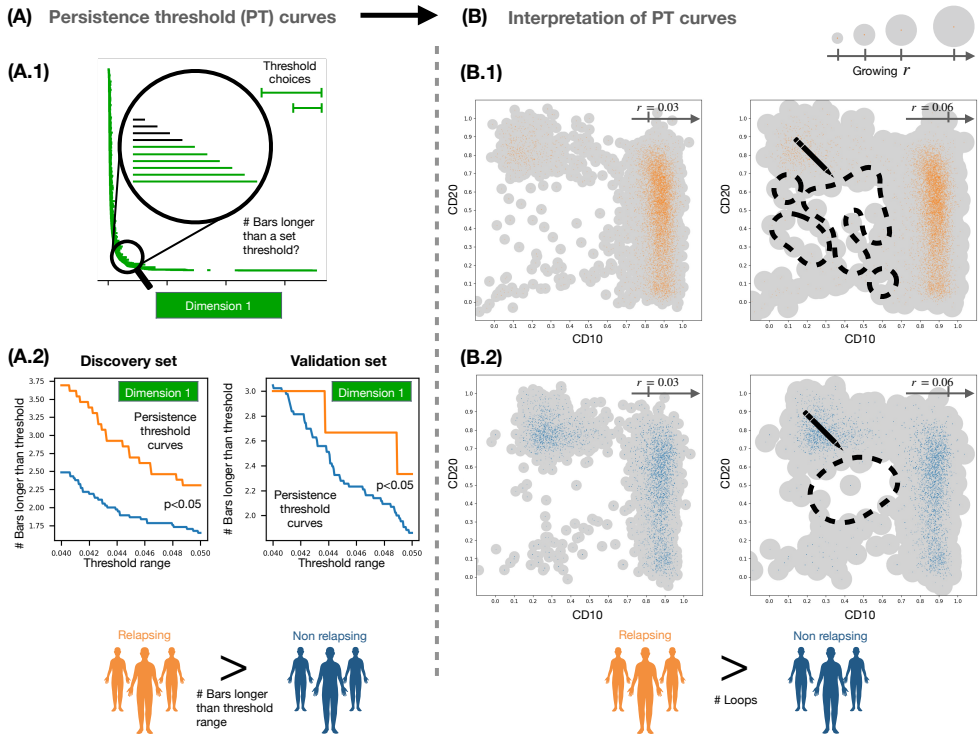

Fig 3. Analysis and interpretation of output from persistent homology. (A) We computed the number of bars in dimension 1 barcodes whose lengths exceed a set of threshold values (long persistence for $\tau \in [0.04, 0.05]$, medium persistence for $\tau \in [0.007, 0.015]$). The resulting persistence threshold curves provide information about the persistence of loops in the data. In both the discovery and validation sets, the number of bars for all pairwise combinations of biomarkers CD10, CD20, CD38 and CD45 differ statistically significantly for R and NR patients, except for the case of CD10-CD45. In this case, the persistence threshold curves for R patients contained a higher number of persistent loops than for NR patients. **(B)** Based on our quantitative analysis, we suggest a qualitative approach for analysing the 2D projection CD10-CD20 (in cases where a full quantitative approach is not possible): consider balls centred at each data point whose radii increase; our results indicate that, as the ball radii grow, individual loops merge more quickly for NR patients (B.1) than for R patients (B.2).

clustering in the data point cloud to verify the dimension 0 PH results, and thus looked for statistical differences between R and NR patients. We considered 4-dimensional patient point clouds in the discovery set and applied the FlowSom algorithm [60] (see 'Methods' and Figures S16-S21) to each patient point cloud to determine the number of clusters. We then compared results between both cohorts. The number of clusters in point clouds from NR patients (4.01 ± 1.17) was lower (t-test, p-value 0.09) than the number in point clouds from R patients (4.56 ± 1.11). When we performed the same analysis with the 2-dimensional data, the same trends were observed for the 2D projections of CD10-CD20 (t-test, p-value 0.18), with fewer clusters in NR patients (3.51 ± 0.73) than in R patients (3.81 ± 1.17). While these results were marginally significant, they are consistent with those obtained from the persistence threshold curves (see Figure 3).

Persistence images accurately identified relapsing patients making full use of 4D data. In our previous analyses, we computed simple barcode descriptors from the PH barcodes which are summaries and thus inevitably lose some of the information. To fully exploit the information on empty spaces in the 4-dimensional datasets (with biomarkers CD10-CD20-CD38-CD45; see Figure 4(A)), we transformed the barcodes into persistence images [54]. This method transforms barcodes into matrices which can be vectorised and, as such, are more amenable to analysis while also being stable to noise and maintaining an interpretable relationship to the barcodes from which they were generated (see Figure 4(B) and Section S1.4 and S2.3.c-d for methodological details).

We used persistence images in matrix form as input to Logistic Regression (LR) and Support Vector Machine (SVM) analyses. We considered three variables: datasets studied (balanced, or not, by oversampling), dimensions of the topological features analysed (0, 1 or 2) and biomarkers included (CD10, CD20, CD38, CD45). We included dimension 2 barcodes in this analysis by reducing the number of data points from 10^5 to 10^4.

	Spread	**Grid**	**Method**	**Dim. 0 Acc.**	**Dim. 1 Acc.**	**Dim. 2 Acc.**
Without oversampling	**0.05**	**100x100**	LR	0.93	0.93	0.93
			SVM	1	1	0.91
		50x50	LR	0.97	0.93	0.93
			SVM	1	1	0.89
		25x25	LR	1	0.93	0.93
			SVM	1	1	0.91
		10x10	LR	1	0.95	0.93
			SVM	1	1	0.89
		5x5	LR	1	0.95	0.93
			SVM	1	1	0.93
With oversampling	**0.05**	**100x100**	SVM	1	1	0.84
		50x50	SVM	1	1	0.79
		25x25	SVM	1	1	0.82
		10x10	SVM	1	1	0.83
		5x5	SVM	1	1	0.83

Table 1. Summary of classification scores of Persistence Images (PIs) after cross-validation on the discovery and validation datasets. We constructed classifiers using logistic Regression (LR) and Support Vector Machines (SVM) on the discovery and validation datasets, using PIs obtained from PH analysis of biomarkers CD10, CD20, CD38, and CD45. The accuracy of the SVM method depends on the dimension of the barcode, the spread of the Gaussian and grid size used to generate the PIs, as well as the parameters C and γ for the SVM method. The accuracy (Acc.) of the model predictions after 6-fold cross-validation is greatest for Dimension 0 and 1. Additional summaries are provided in Table S3-Table S5. An expanded analysis, using fixed hyperparameters and other classification scores (AUC, mean accuracy, standard deviation of the accuracy, recall, precision, F1-score and confusion matrices) can be found in Table S6 and Table S7.

We used the discovery and validation datasets for training and testing, respectively. The classification results are presented in Table 1. We performed cross-validation within the training sets to obtain representative SVM parameters as described in 'Methods'. We obtained perfect classification scores and 100% accuracy. By comparison, a random forest analysis of clinical variables which are routinely used for relapse risk assessment (such as age, sex and other molecular biology data as presence of mutations) lead to an accuracy of 86% after cross-validation, and an expert clinical analysis based on the same data lead to an accuracy of approximately 91% (for more details, see Section 2.1.c in the Appendix). These differences depend on the hyperparameters used to construct the PIs, and were obtained for the discovery and validation datasets (see Figure S23 for details). We constructed the mean PIs for R and NR patients to visually compare such

differences between the two cohorts. Additionally, we computed the coefficients of the decision function in the LR classification and showed by means of an image the discriminatory areas between both cohorts. Moreover, since R patients constituted approximately 20% of the total number of patients, we performed oversampling (see 'Methods') and reran our analysis with SVM. As before, we obtained 100% accuracy, with better classification for dimensions 0 and 1 than for dimension 2. We obtained the same trends when all three PIs (in dimensions 0,1 and 2) were used together as input for the classification methods (see Table S4). Thus, the PH methods combined with the machine learning techniques were able to identify the R patients from the diagnostic FC data with high accuracy.

	Spread	**Grid**	**Method**	**Dim. 0 Acc.**	**Dim. 1 Acc.**	**Dim. 2 Acc.**
Without oversampling	**0.05**	**100x100**	SVM	1	1	0.87
		50x50	SVM	1	1	0.87
		25x25	SVM	1	1	0.84
		10x10	SVM	1	1	0.9
		5x5	SVM	1	0.99	0.82
With oversampling	**0.05**	**100x100**	SVM	1	1	0.72
		50x50	SVM	1	1	0.81
		25x25	SVM	1	1	0.88
		10x10	SVM	1	1	0.86
		5x5	SVM	1	0.99	0.88

Table 2. Summary of classification scores of Persistence Images (PIs) for patients with low-intermediate risk. An SVM classifier was constructed for patients with low or intermediate risk at the time of diagnosis, using PIs obtained from PH analysis of biomarkers CD10, CD20, CD38, and CD45. These results depend on the dimension of the barcode, the spread of the Gaussian and the grid size used for the PIs. We fixed parameters $C = 10$ and $\gamma = 0.001$ for the SVM method, based on the previous results. The accuracy (Acc.) of the model predictions after 6-fold cross-validation is greatest for Dimensions 0 and 1. An expanded analysis with additional classification scores (AUC, mean accuracy, standard deviation of the accuracy, recall, precision, F1-score and confusion matrices) can be found in Table S8.

We also performed classification for a subgroup of patients, specifically those deemed to be of low or intermediate risk at time of diagnosis. From a clinical perspective, it would be advantageous to identify those patients whose initial risk assessment was incorrect and then to reclassify them. As shown in Table S2, it is noteworthy that 85.7% of all patients who experience relapse were initially categorized as having a low-intermediate risk profile. We repeated our TDA and ML based classification of R versus NR. We classified the full dataset with SVM, with 20% training test and 6-fold cross-validation. The results are included in Table 2, where patients with low-intermediate risk were classified with high accuracy using the dimension 0 or 1 information from the PI images, both with and without oversampling.

Discussion and conclusions

Predicting risk of relapse after treatment is one of the major challenges in cancer, particularly now that effective second and third-line treatments are available. In B acute lymphoblastic leukaemia (B-ALL), first-line chemotherapy regimes have increased overall survival by up to 90%, but a substantial fraction of patients relapse. Being able to start other treatments as soon as possible, i.e. before resistant clones emerge and become dominant [2, 3], could significantly improve treatment outcomes in these patients. In the present work we focused on predicting risk of relapse based on FC data, which is routinely collected at diagnosis of haematological malignancies. Clinicians manually assess the shape of 2D projections of this high-dimensional data for diagnosis

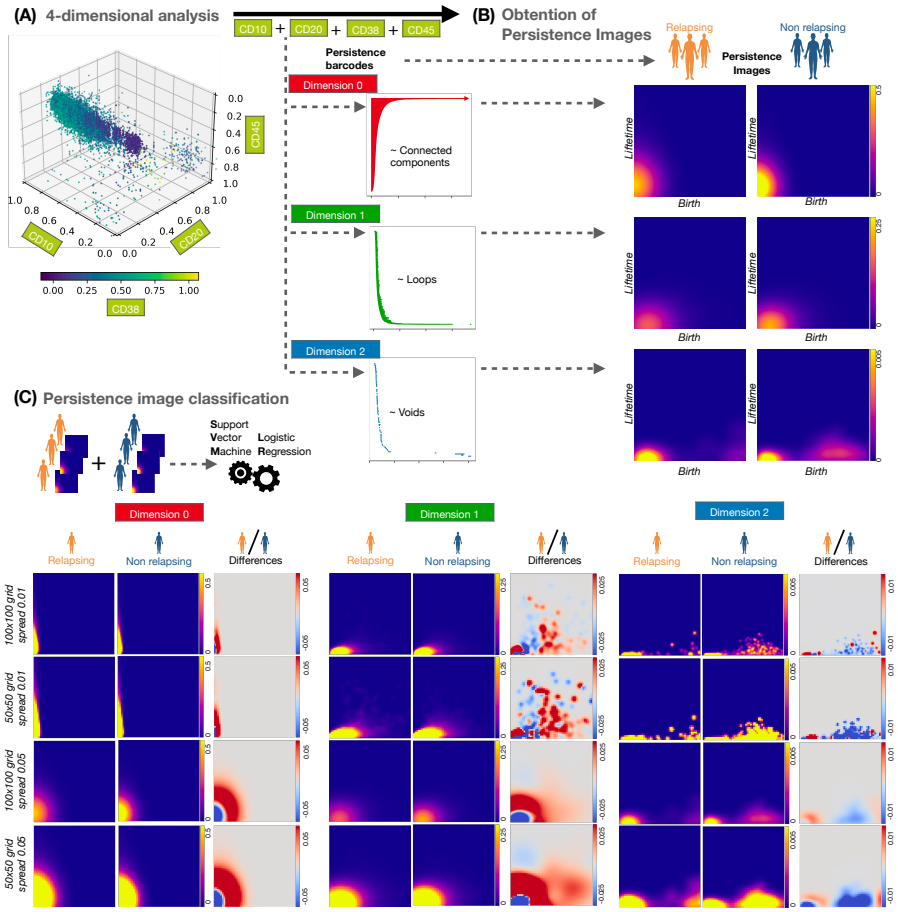

Fig 4. Analysis using Persistence Images. **(A)** We computed PH barcodes in dimensions 0 (connected components), 1 (loops) and 2 (voids in 3D) on the 4-dimensional space of biomarkers CD10-CD20-CD38-CD45. **(B)** We transformed the barcodes for each dimension into persistence images, i.e. a matrix of intensity values. This transformation depends on hyperparameters, such as the choice of grid size and the spread of the Gaussian distribution which is used to convert barcode information into an image (see 'Methods' for a precise description). **(C)** The persistence images (one for each patient in each dimension) were the input for classification using Support Vector Machine and Logistic Regression. We present the mean images for each patient cohort, as well as their differences.

and visualise its empty spaces. We combined persistent homology (PH), a method from topological data analysis, with machine learning to obtain quantitative descriptors which are directly related to empty spaces which have been qualitatively observed in FC data. Our descriptors distinguish R and NR B-ALL patients at diagnosis with high accuracy. We focused our analysis around the following topological features: connected components (dimension 0) and loops (dimension 1).

We first used PH together with machine learning to detect markers relevant for the classification of R patients in an unsupervised manner. IPT markers CD10, CD20, CD45 and CD38 were among those with the highest relevance for classification. These biomarkers are characteristic of the B-cell line; CD10, CD20 and CD45 are routinely used to identify early B cell sub-populations [57], while CD38 is known to have prognostic value for B-ALL [52, 59]. Thus, our results confirmed the importance of these

four biomarkers and, therefore, we retained only these four biomarkers for further analyses.

We analysed the 4D dataset (with the above biomarkers) and pairwise, 2D projections of the data with PH and identified those features which most influence the shape of the point clouds. Interestingly, using simple barcode descriptors extracted from the 2D data we were able to more accurately distinguish between R and NR patients than when considering the 4D data. Our ability to distinguish between R and NR patients increased further when barcodes from the 4D data were transformed into persistent images, i.e. when more of the topological information provided by PH was used for classification. This implies that while the shape of the 2D projections can give insights via both visual inspection and crude shape descriptors, PH captures valuable information that goes beyond these features in the 4D data. This additional information can only be exploited when the data is transformed into a format that can be combined with machine learning while retaining as much information from the barcodes as possible.

In the 2D data, by considering differences between the R and NR persistence threshold curves (PT curves), we found that R patients were characterised by larger numbers of persistent connected components and loops (see Figure 3(B)). In terms of biological insight, these results are consistent with the bone marrow cell population of R patients being more phenotypically diverse than those of NR patients. In the R patients, such cellular heterogeneity can be linked to non-dominant leukaemic subpopulations that might be drivers of disease relapse [61]. Increased cellular heterogeneity in R patients manifests as a higher number of isolated leukaemic point clouds (i.e. clusters of cells spread across phenotype space) in their projected FC data. In particular, the projections of marker combinations, including CD10 and CD20, show such B-cell subpopulations in the data. This may also explain why there are more loops in the data of R patients in the corresponding marker combinations; a more heterogeneous cell population leads to more and larger voids in the data which can result in more persistent loops. This is also related to the presence or absence of 'empty spaces' in FC data: in NR patients there is a lower number of 'empty' locations, where possible future cancer cells can reappear. However, the presence of cancer cells in such locations create more loops in the R patients dataset. In contrast, for NR patients we found more persistent loops in the marker combination CD38-CD45, the only combination not including either CD10 or CD20. A possible explanation is that long-time-survival patients with B-ALL express higher levels of $CD38^+$ in their FC data, which could lead to homogeneity in NR patients. These considerations may also explain why 4D analysis of PT curves did not discriminate well between R and NR groups: we observe more isolated clusters and heterogeneity in the data point clouds of R patients in CD10-CD20 projections, while the intensity of $CD38^-$ in CD38-CD45 projections (which could be related to more homogeneity in these biomarkers for R patients) points to bad prognosis [52, 58]. In general, these significant differences do not seem to be associated with a consistent threshold range in either dimension, although dimension 1 features tend to be present at lower thresholds. As such, we recommend repeating the parameter scan when analysing new data.

In the 4D data, persistence images (PIs) which use the full information content of the barcodes rather than summaries (simple barcode descriptors and PT curves) enabled us to classify the data from R and NR patients via machine learning. We computed PIs from the barcodes in dimension 0, 1 and also 2 (voids) for this analysis. The construction and classification of the PIs depended on the choice of several hyperparameters, such as the grid size, the spread of the Gaussian distribution, and the SVM classification parameters C and γ (see Figure S23 and 'Methods'); their values were optimized for each case. We note that we were able to perform high-grade

classification using a small number of pixels (PI grid size of 5x5) and with fixed SVM hyperparameters $\gamma = 10^{-3}, C = 10$. Likewise, the classification scores were higher with a Gaussian spread of 0.05. Following balancing by oversampling, the classification was excellent in every dimension. On average, classification scores were higher for dimension 0 than dimension 1, and lower for dimension 2. Concatenating PIs from all three dimensions yielded the same trends as those for dimensions 0 and 1. Regardless of classification method, dimension 0 attained a 100% classification score. Similarly, for dimension 1, the results showed a high score for almost every choice of hyperparameters in the construction of the PIs. We obtained slightly worse results in dimension 2 when Logistic Regression was used for classification. However, we note that, due to the high computational costs of PH in dimension 2, only a subset of 10^3 data points were used, instead of the full number of 10^4 points used to construct PIs in dimensions 0 and 1. Support Vector Machine (SVM) analysis applied to the PIs from dimensions 0 and 1 achieved 93% and 100% accuracy in the original and balanced data, respectively. Even when a 5x5 grid was used to genereate the PIs the accuracy of SVM was maintained. This contrasted with existing classifiers, which have an accuracy (area under the curve) of less than 92%, with similar sample size [17, 52]. Moreover, in our dataset, 85.7% of patients who finally relapsed were assessed at the moment of diagnosis as having low or intermediate risk (see Table S2 and Data S4). The results in Table 2 show how our method improves this initial classification, this last including clinical variables and other molecular biology data (for more details, see Section 2.1.c in the Appendix, as well as the previous molecular biology correlations for this dataset in Ref. [52]).

Our results show how using PH to analyse routine high-dimensional FC data can significantly improve the accuracy of predictions of relapse for patients with B-ALL. Our levels of accuracy and precision highlight the potential for topological methods to dissect phenotypic characteristics of B-ALL relapse. While other methods used to predict clinical outcomes exploit the intensity of marker expression or cell proportions (such as [62, 63]), PH measures characteristics of the shape of the point cloud in high-dimensional spaces that are not accessible with other methods. These topological features account for variation in the data due to noise or calibration of the flow cytometer. However, several variables considered during routine, manual inspections performed in the clinic (such as the range of expression levels or the mean fluorescence intensity) can be affected by data transformations. Fully automated and validated, this methodology could be readily integrated into clinical practice. Based on our observations, we propose three practical approaches to assess the risk of relapse on diagnosis, depending on the available methodology (see Figure 5). Each method requires biomarkers CD10, CD20, CD38 and CD45 to be included in the FC data.

1. When only visual inspection of 2D projections of FC data is possible, i.e. when neither PH nor ML methods are available, insights from PH applied to 2D projected data may help to identify qualitative patterns that distinguish R and NR patients. As seen in the Table from Figure 5, patients at risk of relapse tend to have large enclosed voids in most pairwise combinations of the biomarkers and smaller clusters that are far apart from each other in the CD10-CD20 projection. The NR group also exhibit smaller and larger loops in some projections including CD38.

2. If PH methods are available but not ML, then PH analysis should be applied to all 2D projections of biomarkers CD10, CD20, CD38 and CD45 and the numbers of long bars in the dimension 0 and dimension 1 barcodes computed. If, for the dimension 1 barcode, the distance of the corresponding PT curve to the mean PT curve of NR patients is larger than the distance to the PT curve of R patients, then the data is likely to correspond to a relapsing patient. The same analysis can

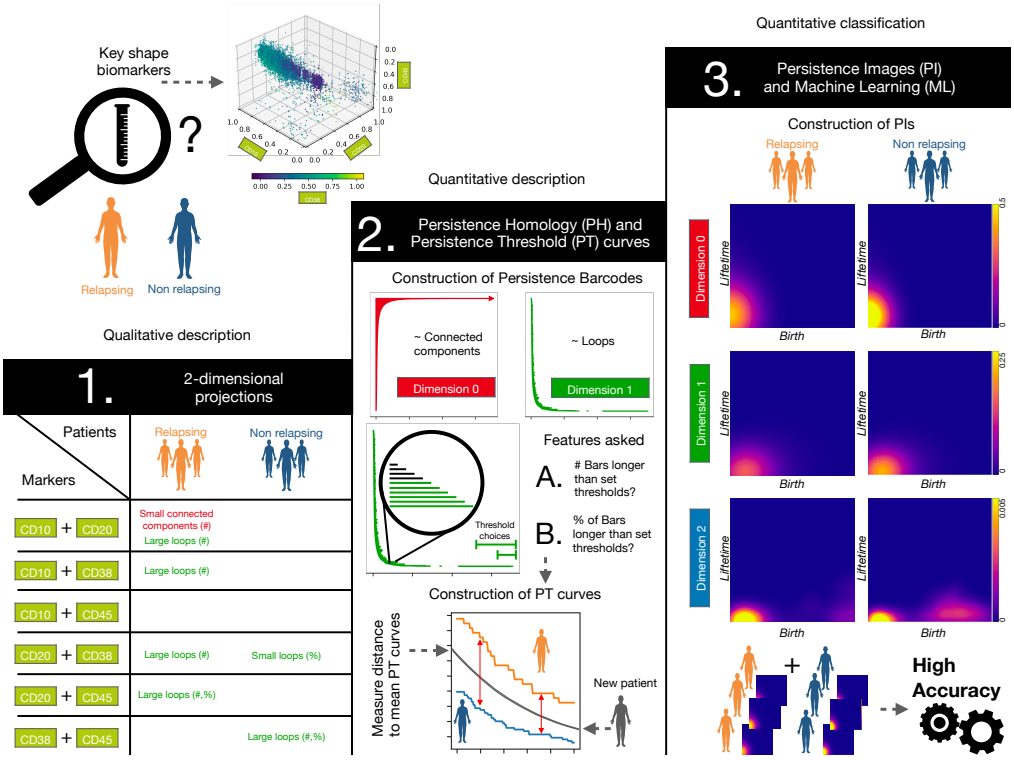

Fig 5. Summary of results and methodology used. We consider biomarkers CD10, CD20, CD38 and CD45 to compare FC data from R and NR patients. **1.** Two-dimensional projections of these biomarkers show differences between both cohorts in the number (#) and percentage (%) of both connected components and loops in the data point clouds. **2.** If Persistent Homology (PH) is available, PH barcodes should be computed for each pairwise combination of biomarkers in dimension 0 and 1. One can construct Persistence Threshold (PT) curves for each combination and then quantify differences between NR and R patients according to A.) the number (#) and B.) percentage (%) of connected components and loops. For a new patient, we can measure the distances of their PT curves to the mean PT curves of NR and R patients associated with each combination of biomarkers. We can then predict the risk of relapse by estimating an averaged probability based on the distances to the corresponding PT curves. **3.** For classification, Persistence Images (PI) can be constructed from the persistence barcodes. Given the results in Table 1, this could lead to accurate classification of R and NR patients via Machine Learning (ML) with methods such as Support Vector Machines.

also be applied to the CD10-CD20 projection in dimension 0.

3. If both PH and ML analyses are available, we suggest PH analysis of the 4D datasets, followed by transformation of the barcodes to persistent images (PIs). Using a classification algorithm on PIs, such as SVM, would allow a high degree of accuracy.

While our analysis reveals that TDA has great potential for assessing the risk of relapse on diagnosis, our study has several limitations. The size of the datasets used may render the results sensitive to overfitting. Despite the robustness to noise of topological methods, further validation of the method on different datasets is needed. While the numbers of NR and R patients were balanced with oversampling, this

difference should be reduced in future work. Additionally, a larger study using bone marrow samples with a larger number of common, standardized IPT markers could be insightful [64], particularly when data would be collected using the same hospital protocols for preprocessing. The use of a single flow cytometer would further improve the accuracy of the study, as machines can differ with respect to the data scales and compensation techniques that they use. Another limitation of our study is the high computational cost of PH: we reduced our data to 10^4 data points to enable computations in dimensions 0 and 1, and to 10^3 data points to enable computations in dimension 2. This represented 10% and 1%, respectively, of the 10^5 points selected from the whole bone marrow. Even when a representative subsample was selected using the max-min algorithm (see 'Methods'), it took on average a week to compute all 16 pairwise combinations of IPT markers, with high-performance computing machines (see 'Methods' for details). However, as PH is an active research area, these computational limitations are constantly being improved. Finally, PH does not consider information about the intensity of markers or cell proportions. It would be interesting to incorporate such information into future PH analysis.

An extension of the methods presented here could be to apply our techniques to other FC datasets, monitoring patient responses to e.g. chemotherapy, transplant or CAR-T cell infusion. This could reveal new information about the emergence of resistance to these treatments. FC data can also provide information about all bone marrow cells, not just B-lymphocytes, and is therefore routinely collected for many haematological conditions such as lymphomas, T-cell leukaemias, myeloblastic disorders. Our analysis could be adapted to study these diseases. Combining information from IPT cell markers with qualitative biomolecular or morphology features has been used in [65, 66] to perform risk stratification and ALL diagnosis. By combining such qualitative data with PH analyses of FC data it may be possible to further increase the accuracy of patient prognosis and relapse prediction, and the selection of effective personalised treatment strategies.

Our results could have been obtained using either PH or ML. However, their combination provides biological interpretability and also exploits the high-dimensional character of FC data, whose shape is currently manually assessed in the clinic by focussing on 2D projections of this routinely-collected data. The approaches outlined in this article represent a first step towards a more general approach in which topological descriptors computed from FC data inform risk prognosis with high accuracy and interpretable results. In other datasets, the same systematic approach could be used to identify informative biomarkers. We propose our approach to be tested prospectively in clinic to determine its practical use. Finally, our work could stimulate further studies which investigate the potential of methods from topological data analysis to reveal new features of cancer relapse.

Materials and Methods

Patients. A retrospective study was designed in accordance with the Declaration of Helsinki, and the protocol was approved by the institutional review board (IRB) of the four participating local institutions (LLAMAT Project, 2018). The inclusion criteria for the study were B-ALL diagnosis between February 2009 and July 2019, age less than 19 years, and availability of bone marrow FC data in Flow Cytometry Standard 3.0 (FCS 3.0). A total of 129 patients were included in the study: 2 from Jerez Hospital (HJ), 54 from Virgen de la Arrixaca Hospital (HVA), 27 from Niño Jesús Hospital (HNJ) and 45 from Virgen del Rocío Hospital (HVR). After preprocessing and marker selection, 96 patients were retained for further analysis. A discovery set was constructed with 30 patients from HVR (dataset 1) and 20 patients from HNJ (dataset 2). A validation set

was later reanalysed with 46 patients from HVA (dataset 3). In all datasets, we counted with 80 non-relapsing (NR) patients, with at least three years of follow-up with no evidence of minimal residual disease, and 16 relapsing (R) patients. Patients characteristics can be found in Table S1 and Data S4).

Flow cytometry markers and data preprocessing. Marker expression was obtained on FACSCanto II flow cytometers, in accordance with the manufacturer's specifications for sample preparation. FCS files were first imported into FlowJo (Becton Dickinson, 10.6.1) and FACSDiva (Becton Dickinson, 8.0.1) and inspected manually by standard methods (see Figure S4): Acquisition anomalies, margin events (measurements that match the maximum or minimum values for any parameter), doublets (cells that are analysed together, as a single event), debris and dead cells were removed [12]. Files were further processed in R (4.1) using FC specific packages from Bioconductor. Specifically, files were compensated and transformed with Logicle transform [67].

For each patient, we subsampled aliquots to the same number of cells and then brought them into a single file via nearest neighbour imputation [68, 69]. This resulted in one file per patient. We subsampled every patient file to 10^5 cells. Batch effects were accounted for with a rescaling transformation. While PH analyses are robust to small perturbations in the data point cloud [27], they may not be robust to outliers. To ensure that outliers do not introduce spurious topological features, we chose the 0.001 and 0.999 quantiles to transform every intensity value x to x': $x' = \frac{x - x_{q0.001}}{x_{q0.999} - x_{q0.001}}$. This was performed to avoid outliers, where $x_{q0.001}$ is the quantile 0.001 and $x_{q0.999}$ is the quantile 0.999. Subsequently, the common B-cell antigen CD19 was used to select the B-cell subpopulation.

Finally, we selected 10^4 landmarks from the lymphocyte cloud by applying the max-min algorithm [70] to each group of markers considered. Because of computational constraints in persistent homology, we reduced the number of landmarks to 10^3 for studies performed in dimension 2. The last step involved the selection of a common set of potential biomarkers for each patient; these included FSC-A, SSC-A, CD10, CD13, CD19, CD20, CD22, CD3, CD33, CD34, CD38, CD45, CD58, CD66c, IGM, cyCD3, cyMPO, cyTDT. We omitted geometric markers FSC-A and SSC-A from the persistence analysis, to prevent unconscious bias. Patients that did not contain the full set were excluded from the analysis.

TDA and Persistence Images. *Persistent homology* (PH) is the most commonly used method from topological data analysis. The technique provides information on topological features such as connected components and loops in point cloud data. To apply persistent homology, one needs to construct a filtration of simplicial complexes from the point cloud data, i.e. a sequence of nested graph-like structures that include nodes (e.g., the data points) and edges as well as higher-order connections (simplices) such as triangles and tetrahedra. A method to obtain such simplicial complexes is the Vietoris-Rips filtration which we use here: Given a point cloud such as the one shown in Figure 1(E), we place a ball of radius r centered at each point. We now connect two points by an edge, if the balls surrounding them intersect. If three points have pairwise intersecting balls, we consider the edges to form a filled-in triangle, if four such points have pairwise intersections, their edges form a filled-in tetrahedron. As the radius r increases, the simplicial complex built on the data contains more edges, triangles, and tetrahedra until at some point all points are connected to each other forming a high-dimensional simplex. PH monitors how topological invariants such as connected components, loops and 3D voids are born and later die along this filtration as the radius r grows. For example, initially for very small r all data points form their own connected components. It is only as r grows that components merge together, i.e. one component

dies while the other lives on now including two data points connected by an edge. For each radius r, the homology group of the simplicial complex captures the connected components (dimension 0), loops (dimension 1), and voids (dimension 2).

The term 'persistence', or lifetime, refers to how long these topological features live within the filtration. This can be represented by *persistence barcodes*, which are collections of intervals $[r_b, r_d)$ where each interval represents the lifetime of one topological feature in the filtration. Intuitively, r_b is the radius at which the feature is first observed in the filtration and r_d corresponds to the radius when the feature dies in the filtration. For an example of a barcode, see Figure 1(F).

To further quantify these features, here we constructed so-called *persistence threshold (PT) curves*: we counted the number and percentage of bars in each barcode whose corresponding topological features persist longer than a threshold taken from a range of thresholds. Specifically, we studied threshold ranges $\tau \in [0.03, 0.07]$ and $\tau \in [0.1, 0.18]$ for dimension 0, as well as threshold ranges $\tau \in [0.007, 0.015]$ and $\tau \in [0.04, 0.05]$ for dimension 1. We chose these threshold ranges to represent different scales of persistence ('medium' versus 'long' scales).

In our study, we further applied *persistence images* (PIs), which we generated from persistence barcodes. These images are real-valued matrices that can be used as an input into a variety of machine-learning approaches and are then useful for classification. The construction of PIs depends on the hyperparameters, such as the grid size (5x5, 10x10, 25x25, 50x50, or 100x100 in this study) and the spread of the Gaussian distribution (0.01 and 0.05 in this study) which is used when transforming the information from the barcode to the image. A more detailed summary of TDA can be found in the Appendix, as well as in further references [26, 54, 71].

Machine learning methods. Artificial intelligence methods, specifically Machine Learning, were used for two purposes: feature selection and classification. For the first task, we divided data into discovery and validation datasets as stated in Patients (see Table S1). For feature selection and prediction, we used Random Forests (RF) [72], a popular and efficient algorithm based on model aggregation ideas. The number of estimators in RF ranged between 20 and 100, with 10 iterations to account for robustness. Firstly, we computed receiver operating characteristic (ROC) curves for the classification of R versus NR patients. We used persistence barcodes of the pairwise combinations of the IPT markers to confirm their prognostic relevance. We extracted the following simple descriptors from each barcode: the maximum, minimum and mean persistence and the standard deviation in dimensions 0 and 1. We used the descriptors to create 8-dimensional topological feature vectors for each patient. IPT pairs with an area under the ROC curve of less than 50% were deemed to have very low information and excluded from subsequent analysis.

For our second task regarding classification, we applied Support Vector Machines (SVM) [73], as this method provides good decision surfaces by maximizing margins between two classes. In our case, we assigned R and NR labels to the data matrices obtained from the PIs of the patients. We used the same set division into discovery and validation for this method. We varied two parameters associated with the SVM classification kernel: γ, which represents the curvature of the decision boundary, and C, which is the trade-off between correct classification and the distance between the decision boundary and support vectors. We randomly selected the SVM hyperparameters by performing stratified k-fold cross-validation in the training set for a logarithmic range of values of the parameters $C \in [10^{-2}, 10^{13}]$ and $\gamma \in [10^{-9}, 10^{3}]$. We also used Logistic Regression to construct binary regressions. We computed the mean score of fitting the model after 6-fold cross-validation. R patients constituted approximately 20% of the total number of patients. In order to address this imbalance,

we repeated the SVM method for classifying R and NR patients with oversampling on the first ones, i.e., duplicating PIs from the R group.

Computing machines. The max-min algorithm and Persistence Analysis were run on six machines from the Oxford Mathematical Institute, each with 36 cores, with up to 3.9 Ghz speed, and 768 GB of RAM. Figures and other calculations were run on an iMac, running under Mac OS 10.15, with four i5 cores, 3.4 Ghz speed and 32 GB RAM.

Software. Python (3.1) was used for the computations. R (3.6.0) and RStudio (1.2.1335) were used for data curation. Persistence barcodes and images were constructed using Ripser (0.3.2) [74] and Persim packages (0.1.3) in the Python Scikit-TDA toolbox. Note that one could achieve faster computations using the recently introduced Ripser++ library [81]. FCS files were read using the Python CytoFlow library (1.0) and FC libraries from Bioconductor (3.11) in RStudio. FlowJo (Becton Dickinson, 10.6.1) and FACSDiva (Becton Dickinson, 8.0.1) software packages were used for manual gating of FCS data. Random Forest and Support Vector Machine methods were obtained from the Python Scikit-Learn (0.24.2) classification library. Clustering was performed using the FlowSom algorithm [60], normally used for FC data. This last is included in the Spectre package (from R, v.0.5.0), embedded in Python with the rpy2 module (v. 3.4.5).

Data and Code Availability Implementation details, code and processed PH data are freely available on `https://github.com/salvadorchulian/shapecancerrelapse`. FC files are available upon request. These datasets can be provided by the corresponding author, S.C., or H.B., pending scientific review and a completed material transfer agreement. Requests for the datasets should be submitted to: S.C., or H.B.. All other data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Author Contributions Conceptualization, S.C., B.J.S., V.M.P.-G., M.R. and H.B.; Data curation, S.C., Á.M.-R., C.B.G., J.F.R.G., Á.M.Q., T.C.-V., M.R.-O., A.C.R., J.L.F.S., A.M.P. and M.V.M.S.; Formal analysis, S.C.; Funding acquisition, M.R., V.M.P.-G. and H.B.; Investigation, S.C. and B.J.S.; Methodology, S.C. and B.J.S.; Project administration, C.B.G., M.R., V.M.P.-G. and H.B.; Resources, Á.M.Q., T.C.-V., M.R.-O., A.C.R., J.L.F.S., A.M.P. and M.V.M.S.; Software, S.C. and Á.M.-R.; Supervision, V.M.P.-G., M.R. and H.B.; Validation, S.C., C.B.G., J.F.R.G.; Visualization, S.C. and Á.M.-R.; Writing—original draft, S.C. and B.J.S.; Writing—review & editing, S.C., B.J.S., Á.M.-R., C.B.G., J.F.R.G., Á.M.Q., T.C.-V., M.R.-O., A.C.R., J.L.F.S., A.M.P., M.V.M.S., M.R., V.M.P.-G. and H.B. All authors have read and agreed to the published version of the manuscript.

Competing Interests The authors declare that they have no competing financial interests. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Supporting information

Appendix. Supplementary Materials. Content on Topological Data Analysis and Methodological structure. This last includes the data processing techniques, obtention of biomarkers, and the interpretation and classification on such biomarkers.

Data S1. Random forest classification results for the whole dataset with stratified k-fold and oversampling.

Data S2. Random forest classification results for the discovery group.

Data S3. Random forest classification results for the validation group.

Data S4. Patients clinical information.

Table S1 Patients clinical features. It can be found at the Appendix.

Table S2 Frequencies and count of patients with data on risk stratification. It can be found at the Appendix.

Table S3 Comparison of classification scores of persistence images using the training setand the discovery dataset with testing on the validation dataset. It can be found at the Appendix.

Table S4 Comparison of classification scores of persistence images with upper sampling. It can be found at the Appendix.

Table S5 Comparison of classification scores of persistence images of dimensions 0, 1 and 2 together as input for the cross-validation. It can be found at the Appendix.

Table S6 Comparison of full classification scores of persistence images using the whole dataset, without oversampling and using stratified 6-fold validation. It can be found at the Appendix.

Table S7 Comparison of full classification scores of persistence images using the whole dataset, with oversampling and stratified 6-fold validation. It can be found at the Appendix.

Table S8 Comparison of full classification scores of persistence images using the low-intermediate patients dataset, without oversampling and using stratified 6-fold validation. It can be found at the Appendix.

Acknowledgments

This work was partially supported by the Fundación Española para la Ciencia y la Tecnología (FECYT project PR214), the Asociación Pablo Ugarte (APU, Spain), Junta de Andalucía (Spain) group FQM-201, Junta de Comunidades de Castilla-La Mancha (grant number SBPLY/17/180501/000154), the Programme of Research and Transfer Promotion from the University of Cádiz (grant number EST2020-025), Ministery of Science and Technology, Spain (grant number PID2019-110895RB-I00, funded by MCIN/AEI/10.13039/501100011033). This work was also subsidized by a grant for the research and biomedical innovation in the health sciences within the framework of the Integrated Territorial Initiative (ITI) for the province of Cadiz (grant number ITI-0038-2019). ITI is 80% co-financed by the funds of the FEDER Operational Program of Andalusia 2014-2020 (Council of Health and Families). BJS and HMB are

members of the Centre for Topological Data Analysis, funded by the EPSRC grant (EP/R018472/1). The authors also thanks Heather Harrington (Centre for Topological Data Analysis, Mathematical Institute, Oxford) for helpful discussions.

References

1. Z.J. Ward, J.M. Yeh, N. Bhakta, A.L. Frazier, R. Atun, Estimating the total incidence of global childhood cancer: a simulation-based analysis. *Lancet Oncol.* **20**, 483–493 (2019).

2. R. A. Gatenby, J. Zhang, J. S. Brown. First Strike–Second Strike Strategies in Metastatic Cancer: Lessons from the Evolutionary Dynamics of Extinction. *Cancer Res.* **79**(13) 3174-3177 (2019).

3. R. A. Gatenby, Y. Artzy-Randrup, T. Epstein, D. R. Reed, J. S. Brown, Eradicating Metastatic Cancer and the Eco-Evolutionary Dynamics of Anthropocene Extinctions. *Cancer Res.* **80**(3) 613-623 (2020).

4. R. R. Sarkar, N. J. Gloude, D. Schiff, J. D. Murphy, Cost-effectiveness of chimeric antigen receptor T-cell therapy in pediatric relapsed/refractory B-cell acute lymphoblastic leukemia. *J. Natl. Cancer Inst.*, **111**(7), 719-726 (2019).

5. B.J. Bain, Ed. 5, *Leukaemia diagnosis* (Wiley Blackwell, 2017).

6. R.A Egler, S.P. Ahuja, Y. Matloub, L-asparaginase in the treatment of patients with acute lymphoblastic leukemia. *J. Pharmacol. Pharmacother.* **7**(2), 62-71 (2016).

7. P. Charoentong, F. Finotello, M. Angelova, C. Mayer, M. Efremova, D. Rieder et al, Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* **18**, 248-262 (2017).

8. M. Ghandi, F.W. Huang, J. Jané-Valbuena, G.V. Kryukov, C.C. Lo, E. Robert McDonald III et al, Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508 (2019).

9. D.J. Kerr, L. Yang, Personalising cancer medicine with prognostic markers. *EBioMedicine*, **72**, 103577 (2021).

10. F. E. Craig, K. A. Foon, Flow cytometric immunophenotyping for hematologic neoplasms. *Blood* **111**(8), 3941-3967 (2008).

11. T. Terwilliger, M. Abdul-Hay, Acute lymphoblastic leukemia: a comprehensive review and 2017 update. *Blood Cancer J.* **7**, e577 (2017).

12. Y. Saeys, S. Van Gassen, B. N. Lambrecht. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Rev. Immunol.* **16**(7), 449-462 (2016).

13. P. Piñero, M. Morillas, N. Gutierrez, E. Barragán, E. Such, J. Breña et al, Identification of Leukemia-Associated Immunophenotypes by Databaseguided Flow Cytometry Provides a Highly Sensitive and Reproducible Strategy for the Study of Measurable Residual Disease in Acute Myeloblastic Leukemia. *Cancers*, **14**(16), 4010 (2022).

14. E. Björklund, J. Mazur, S. Söderhäll, A. Porwit-MacDonald, Flow cytometric follow-up of minimal residual disease in bone marrow gives prognostic information in children with acute lymphoblastic leukemia. *Leukemia*, **17**(1), 138-148 (2003).

15. P. Lúcio, A. Parreira, M.W.M. van den Beemd, E.G. van Lochem, E.R. van Wering, E. Baars et al, Flow cytometric analysis of normal B cell differentiation: a frame of reference for the detection of minimal residual disease in precursor-B-ALL. *Leukemia*, **13**(3), 419-427 (1999).

16. A. Schumich, M. Maurer-Granofszky, A. Attarbaschi, U. Pötschger, B. Buldini, G. Gaipa et al, Flow-cytometric minimal residual disease monitoring in blood predicts relapse risk in pediatric B-cell precursor acute lymphoblastic leukemia in trial AIEOP-BFM-ALL 2000 *Pediatr Blood Cancer*, **66**(5), e27590 (2018).

17. Z. Good, J. Sarno, A. Jager, N. Samusik, N. Aghaeepour, E. F Simonds et al, Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse *Nat. Med.* **24**, 474–483 (2018).

18. Z. Hu, S. Bhattacharya, A. J. Butte, Application of machine learning for cytometry data. *Frontiers in immunology* **12**(787574), 1-8 (2022).

19. E. Arvaniti, M. Claassen, Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature communications*, **8**(1), 14825 (2017).

20. R.V. Bruggner, B. Bodenmiller, D. L. Dill, R.J. Tibshirani, G.P. Nolan, Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, **111**(26), E2770-E2777 (2014).

21. S. Van Gassen, C. Vens, T. Dhaene, B. N. Lambrecht, Y. Saeys, FloReMi: Flow density survival regression using minimal feature redundancy. *Cytometry Part A*, **89**(1), 22-29 (2015).

22. H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification *Discrete Comput. Geom.* **28**, 511-533 (2002).

23. G. Carlsson, Topology and data *Bull. Am. Math. Soc.* **46**, 255-308 (2009).

24. H. Edelsbrunner, J.L. Harer, Persistent homology - A survey. *Contemp. Math.* **453**, 257-282 (2008)

25. H. Edelsbrunner, J.L. Harer, *Computational Topology* (American Mathematical Society, 2010).

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

27. Cohen-Steiner, H. Edelsbrunner, J. Harer, Stability of persistence diagrams. *Discrete Comput. Geom.* **37**(1), 103-120 (2007)

28. Y. Skaf, R. Laubenbacher. Topological Data Analysis in Biomedicine: A Review. *J. Biomed. Inform.* **130**, 104082 (2022).

29. A. Bukkuri, N. Andor, I. K. Darcy. Applications of topological data analysis in oncology. *Front. Artif. Intell.* **4**, 659037. (2021).

30. P. Bendich, J. S. Marron, E. Miller, A. Pieloch, S. Skwerer, Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* **10**(1), 198-218 (2014).

31. L. Kanari, P. Dłotko, M. Scolamiero, R. Levi, J. Shillcock, K. Hess et al,A topological representation of branching neuronal morphologies. *Neuroinformatics* **16**(1), 3-13 (2018).

32. C. Curto, What can topology tell us about the neural code? *Bull. Am. Math. Soc.* **54**, 63-78 (2017).

33. F. Belchi, M. Pirashvili, J. Conway, M. Bennett, R. Djukanovic, J. Brodzki, Lung Topology Characteristics in patients with Chronic Obstructive Pulmonary Disease. *Sci. Rep.* **1**, 5341.

34. J. Nicponski, J. Jung, Topological data analysis of vascular disease: A theoretical framework. *Front. Appl. Math. Stat.* **6**, 34.

35. M.R. McGuirl, A. Volkening, B. Sandstede, Topological data analysis of zebrafish patterns. *Proc. Nat. Acad. Sci.* **117**(10), 5113-5124.

36. K. Xia, Persistent homology analysis of ion aggregations and hydrogen-bonding networks. *Phys. Chem. Chem. Phys.* **20**, 13448-13460 (2018).

37. D. Taylor, F. Klimm, H. A. Harrington, M. Kramár, K. Mischaikow, M. A. Porter et al, Topological data analysis of contagion maps for examining spreading processes on networks. *Nat. Commun.* **6**, 7723 (2015).

38. W.S. Tang, G. Monteiro da Silva, H. Kirveslahti, E. Skeens, B. Feng, T. Sudijono et al, A topological data analytic approach for discovering biophysical signatures in protein dynamics. *PLoS comput. biol.* **18**(5) e1010045 (2022).

39. M. Feng, M. A. Porter, Spatial Applications of Topological Data Analysis: Cities, Snowflakes, Random Structures, and Spiders Spinning Under the Influence. *Phys. Rev. Res.* **2**, 033426 (2020).

40. B. Stolz, J. Kaeppler, B. Markelc, F. Mech, F. Lipsmeier, R. J. Muschel et al, Multiscale Topology Characterises Dynamic Tumour Vascular Networks. *Science Advances* **8**(23), eabm2456 (2022).

41. C. Giusti, E. Pastalkova, C. Curto, V. Itskov, Clique topology reveals intrinsic geometric structure in neural correlations. *Proc. Nat. Acad. Sci.* **112**(44), 13455-13460 (2015).

42. J.L. Nielson, J. Paquette, A. W. Liu, C. F. Guandique, C. A. Tovar, T. Inoue et al, Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nat. Commun.* **6**, 8581 (2015).

43. B.J. Stolz, J. Tanner, H.A. Harrington, V. Nanda. Geometric anomaly detection in data. *Proc. Nat. Acad. Sci.* **117** (33), 19664-19669 (2020).

44. J. T. Nardini, B. J. Stolz, K. B. Flores, H. A. Harrington, H. M. Byrne, Topological data analysis distinguishes parameter regimes in the Anderson-Chaplain model of angiogenesis, *PLoS Comput. Biol.* **17**(6) e1009094 (2021).

45. T. Lakshmikanth, A. Olin, Y. Chen, J. Mikes, E. Fredlund, M. Remberger et al, Mass cytometry and topological data analysis reveal immune parameters associated with complications after allogeneic stem cell transplantation. *Cell Rep.* **20**, 2238–2250 (2017).

46. L. Crawford, A. Monod, A. X. Chen, S. Mukherjee, R. Rabadán. Predicting Clinical Outcomes in Glioblastoma: An Application of Topological and Functional Data Analysis. *J. American Stat. Assoc.* **115**, 1139-1150 (2020).

47. M. Nicolau, A.J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Nat. Acad. Sci.* **108**, 7265–7270 (2011).

48. P. Lawson, A. B. Sholl, J. Q. Brown, B. T. Fasy, C. Wenk, Persistent Homology for the Quantitative Evaluation of Architectural Features in Prostate Cancer Histology. *Sci. Rep.* **9**, 1139 (2019).

49. R. Rabadán, Y. Mohamedi, U. Rubin, T. Chu, A. N. Alghalith, O. Elliott et al, Identification of relevant genetic alterations in cancer using topological data analysis. *Nat. Comm.* **11**, 3808 (2020).

50. O. Vipond, J. A. Bull, P. S. Macklin, U. Tillmann, C. W. Pugh, H. M. Byrne et al, Multiparameter persistent homology landscapes identify immune cell spatial patterns in tumors. *Proc. Natl. Acad. Sci.* **118**(41) (2021).

51. S. Mukherjee, D. Wethington, T. K. Dey, J. Das, Determining clinically relevant features in cytometry data using persistent homology. *PLoS Comput. Biol.* **18**(3) e1009931 (2022).

52. S. Chulián, Á. Martínez-Rubio, V. M. Pérez-García, M. Rosa, C. Blázquez Goñi, J. F. Rodríguez Gutiérrez et al, High-Dimensional Analysis of Single-Cell Flow Cytometry Data Predicts Relapse in Childhood Acute Lymphoblastic Leukaemia. *Cancers* **13**(1), 17 (2021).

53. A. Kakaje, M. Marwan Alhalabi, A. Ghareeb, B. Karam, B. Mansour, B. Zahra et al, Rates and trends of childhood acute lymphoblastic leukaemia: an epidemiology study. *Sci. Rep.* **10**, 6756 (2020).

54. H. Adams, S. Chepushtanova, T. Emerson, E. Hanson, M. Kirby, F. Motta et al, Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18**(1),218-252 (2017).

55. H. Hakkoum, I. Abnane, A. Idri, Interpretability in the medical field: A systematic mapping and review study. *Appl. Soft Comput.* **117**, 108391 (2022).

56. C. Molnar, Interpretable machine learning. 2020. `https://leanpub.com/interpretable-machine-learning`, last visited 04/23.

57. E. Van Lochem, H.J. van der Velden, H.K. Wind, J.G. te Marvelde, N.A.C. Westerdaal, J.J.M. van Dongen, Immunophenotypic differentiation patterns of normal hematopoiesis in human bone marrow: Reference patterns for age-related changes and disease-induced shifts. *Cytom. Part B: Clin. Cytom.* **60**, 1–13 (2004).

58. S. Burgler, Role of CD38 Expression in Diagnosis and Pathogenesis of Chronic Lymphocytic Leukemia and Its Potential as Therapeutic Target. *Crit. Rev. Immunol.* **35**(5), 417-432 (2015).

59. Z. Jiang, D. Wu, S. Lin, P. Li, CD34 and CD38 are prognostic biomarkers for acute B lymphoblastic leukemia. *Biomark. Res.* **4**, 23 (2016).

60. S. Van Gassen, B. Callebaut, M. J. Van Helden, B. N. Lambrecht, P. Demeester, T. Dhaene et al, FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**(7), 636-645 (2015).

61. C. Lutz, P. S. Woll, G. Hall, A. Castor, H. Dreau, G. Cazzaniga et al, Quiescent leukaemic cells account for minimal residual disease in childhood lymphoblastic leukaemia. *Leukemia* **27** 1204-1207 (2013).

62. S. Van Gassen, C. Vens, T. Dhaene, B. N. Lambrecht, Y. Saeys, FloReMi: Flow density survival regression using minimal feature redundancy. *Cytometry A* **89**(1), 22-29 (2016).

63. R. V. Bruggner, B. Bodenmiller, D. L. Dill, R. J. Tibshirani, G. P. Nola, Automated identification of stratifying signatures in cellular subpopulations. *Proc. Nat. Acad. Sci.* **111**(26), E2770-E2777 (2014).

64. J. J. M. Van Dongen, L. Lhermitte, S. Böttcher, J. Almeida, V. H. J. van der Velden, J. Flores-Montero et al, EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia* **26**(9), 1908-1975 (2012).

65. A.V. Moorman, A. Enshaei, C. Schwab, R. Wade, L. Chilton, A. Elliott et al, A novel integrated cytogenetic and genomic classification refines risk stratification in pediatric acute lymphoblastic leukemia. *Blood* **124**, 1434–1444 (2014).

66. A.T. Sahlol, A.M. Abdeldaim, A.E. Hassanien, Automatic acute lymphoblastic leukemia classification model using social spider optimization algorithm. *Soft. Comput.* **23**, 6345–6360 (2019).

67. D. R. Parks, M. Roederer, W. A. Moore. A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytometry A* **69**(6), 541-551 (2006).

68. C. E. Pedreira, E. S. Costa, M. E. Arroyo, J. Almeida, A. Orfao, A multidimensional classification approach for the automated analysis of flow cytometry data. *IEEE Trans. Biomed. Eng.* **55**(3), 1155-1162 (2008).

69. P. Theunissen, E. Mejstrikova, L. Sedek, A. J. van der Sluijs-Gelling, G. Gaipa, M. Bartels et al, Standardized flow cytometry for highly sensitive mrd measurements in b-cell acute lymphoblastic leukemia. *Blood* **129**, 347–357 (2017).

70. V. De Silva, G. E. Carlsson, Topological estimation using witness complexes, presented at SPBG'04 Symposium on Point - Based Graphics 2004, **4**, 157-166, Zurich, Switzerland, 2-4 June 2004.

71. R. Ghrist, Barcodes: the persistent topology of data. *Bull. Am. Math. Soc.* **45**(1), 61-75 (2008).

72. L. Breiman, Random forests. *Mach Learn* **45**(1), 5-32 (2001).

73. W. S. Noble, What is a support vector machine? *Nat. Biotechnol.* **24**(12), 1565-1567 (2006).

74. C. Tralie, N. Saul, R. Bar-On. Ripser.py: A lean persistent homology library for python. *J. Open Source Softw.* **3**(29), 925 (2018).

75. C. Topaz, Self-help homology tutorial for the simple(x)-minded. 2015. `https://chadtopaz.com/publications/`, last visited 10/21.

76. B. J. Stolz, B. I. Mahler, H is for Homology. 2021. `https://www.maths.ox.ac.uk/about-us/life-oxford-mathematics/oxford-mathematics-alphabet/h-homology`, last visited 10/21.

77. A. E. Sizemore, J. E. Phillips-Cremins, R. Ghrist, D. S. Bassett, The importance of the whole: topological data analysis for the network neuroscientist. *Netw. Neurosci.* **3**, 656 - 673 (2018).

78. H. M. Byrne, H. A. Harrington, R. Muschel, G. Reinert, B.J. Stolz, U. Tillmann, Topology characterises tumour vasculature. *Math. today* **5**(55), 206 - 210 (2019).

79. B. J. Stolz-Pretzer, "Global and local persistent homology for the shape and classification of biological data", thesis, University of Oxford, United Kingdom (2019).

80. B. J. Stolz, T. Emerson, S. Nahkuri, M. A. Porter, H. A. Harrington, Topological data analysis of task-based fMRI data from experiments on Schizophrenia. *J. Phys. Complex.* **2**(3), 035006 (2021).

81. S. Zhang, M. Xiao, H. Wang, GPU-Accelerated Computation of Vietoris-Rips Persistence Barcodes, in: *36th International Symposium on Computational Geometry (SoCG 2020)*, 2020.