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

Salvador Chulián[1,2,*], Bernadette J. Stolz[3],
Álvaro Martínez-Rubio[1,2], Cristina Blázquez Goñi[2,4,5],
Juan F. Rodríguez Gutiérrez[4], Teresa Caballero Velázquez[6],
Águeda Molinos Quintana[5], Manuel Ramírez Orellana[7],
Ana Castillo Robleda[7], José Luis Fuster Soler[8],
Alfredo Minguela Puras[9], María Victoria Martínez Sánchez[9],
María Rosa[1,2], Víctor M. Pérez-García[10,11,12], Helen 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]Department of Paediatric Hematology and Oncology,
Hospital de Jerez, 11407 Jerez de la Frontera, Spain
[5]Department of Hematology,
Hospital Universitario Vírgen del Rocío, 41013 Sevilla, Spain
[6]Department of Haematology,
Hospital Universitario Vírgen del Rocío / University of Sevilla, 41013 Sevilla, Spain
[7]Department of Paediatric Hematology and Oncology,
Hospital Infantil Universitario Niño Jesús,
Instituto Investigación Sanitaria La Princesa, 28009 Madrid, Spain
[8]Department of Paediatric Hematology and Oncology,
Hospital Clínico Universitario Virgen de la Arrixaca,
Instituto Murciano de Investigación Biosanitaria (IMIB), 30120 Murcia, Spain
[9]Immunology Service,
Hospital Clínico Universitario Virgen de la Arrixaca,
Instituto Murciano de Investigación Biosanitaria (IMIB), 30120 Murcia, Spain
[10]Department of Mathematics, Mathematical Oncology Laboratory (MOLAB),
Universidad de Castilla-La Mancha, 13005 Ciudad Real, Spain

[11]Instituto de Matemática Aplicada a la Ciencia y la Ingeniería (IMACI),
Universidad de Castilla-La Mancha, 13005 Ciudad Real, Spain
[12]ETSI Industriales, Universidad de Castilla-La Mancha, 13005 Ciudad Real, Spain

* Corresponding authors. Emails: salvador.chulian@uca.es, helen.byrne@maths.ox.ac.uk

Acute Lymphoblastic Leukaemia (ALL) is the most frequent paediatric cancer. Modern therapies have improved survival rates, but approximately 15-20 % of patients relapse. At present, patients' risk of relapse are assessed by projecting high-dimensional flow cytometry data onto a subset of biomarkers and manually estimating the shape of this reduced data. Here, we apply methods from topological data analysis (TDA), which quantify shape in data via features such as connected components and loops, to pre-treatment ALL datasets with known outcomes. We combine these fully unsupervised analyses with machine learning to identify features in the pre-treatment data that are prognostic for risk of relapse. We find significant topological differences between relapsing and non-relapsing patients and confirm the predictive power of CD10, CD20, CD38, and CD45. Further, we are able to use the TDA descriptors to predict patients who relapsed. We propose three prognostic pipelines that readily extend to other haematological malignancies.

Teaser: *Topology reveals features in flow cytometry data which predict relapse of patients with acute lymphoblastic leukemia*

Short title: *The shape of cancer relapse*

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. Identifying patients, specifically at the paediatric age, at risk of relapse as early as possible - ideally at diagnosis - and tailoring treatment plans accordingly is, therefore, of major clinical importance. Rather than receiving toxic and ineffective first-line chemotherapy regimes, patients at risk of relapse could, for example, immediately receive alternative treatments, such as more intensive chemotherapeutic regimes, before resistant clones emerge and become dominant (*2, 3*), saving valuable time and improving survival rates. Alternatively, it may be better to immediately use more aggressive or expensive therapies, such as bone marrow transplants or CAR-T-cell therapies as a first line treatment (*4*).

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 hematological 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 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 flow cytometry 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 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. However, manual assessment is time-consuming, subjective, and prone to errors. Therefore, more systematic and automated methods of analysis are required. Conventional data analysis methods, such as principal component analysis or machine learning (ML) methods (neural networks, support vector machines, etc), which focus on linear relationships between biomarkers rather than characteristics of shape, could be applied for classification. However, they are not ideal in this context because the results are typically difficult to interpret.

Topological data analysis (TDA) (*13, 14*) 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) (*13–17*). PH provides a quantification of the shape of point cloud data at different spatial scales. Thus, in contrast to other data analysis methods, it is not necessary to predetermine a spatial scale up to which relationships between data points are considered; PH takes into account the full range of possible scales. PH studies the shape of data via topological features such as connected components, i.e. data point clusters, and loops, which are formed by data points surrounding an empty hole. One can imagine PH as a process in which data is viewed at different resolution scales. The output from PH are topological signatures of the data called barcodes which visualise topological features and their scales 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 (*17*), robust to noise (*18*) and its outputs are interpretable. In order to compare barcodes, perform statistical analyses, and/or apply classification methods from ML, barcodes can further be transformed into persistent images (*55*) which are a vectorised representation of the output of PH. In recent years, improved computational feasibility of PH has increased its applications to (high-dimensional) data in many different contexts, including studies of the shape of brain arteries (*20*), neurons (*21*), the neural code (*22*), airways (*23*), stenosis (*24*), zebrafish patterns (*25*), ion aggregation (*26*), contagion dynamics (*27*), spatial networks (*28–31*), and geometric anomalies (*32*). In oncology, PH has been used to construct new biomarkers (*29, 33, 34*), to classify tumours (*35, 36*) and genetic alterations (*37*) and to quantify patterns of immune cell infiltration into solid tumours (*38*).

In this paper we design a pipeline that combines PH and ML methods to predict relapse in ALL with very high accuracy (see Fig. 1). We combine flow cytometry data collected at three different hospitals in Spain (*39*) from patients with the most common subtype of paediatric ALL (*40*), one that is characterised by rapid growth of early B lymphocytes (B-ALL). We apply PH to these datasets, obtain their topological signatures and analyse them to identify those biomarkers with the strongest discriminatory power to classify relapsing (R) and non-relapsing (NR) patients. Our approach confirms that the known biomarkers CD10, CD20, CD38 and/or CD45 are most relevant for this classification. We identify specific topological features (connected components and loops) which appear predominantly in patients at risk of relapse. We further transform the barcodes into persistent images which allow for the topological classification of patients at risk of relapse. Finally, by applying ML to persistence images, we accurately classify persistent images of relapsing (R) versus non-relapsing (NR) patients. Based on our analyses, we suggest three novel and practical approaches for prognosis. The first approach relies only on PH and can be applied when ML is not available. With this approach, we identify interpretable shape characteristics in the data that are specific to patients at risk of relapse. The second 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. If only PH is available, we suggest a third 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 scale thresholds. Our work can be readily extended to many other haematological malignancies.

Results

Machine learning classifier applied to TDA confirmed state-of-the-art markers for prognosis. We applied PH to our datasets and then applied machine learning to the output descriptors of PH to identify those immunophenotypic (IPT) markers that contain the information most relevant for classifying 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

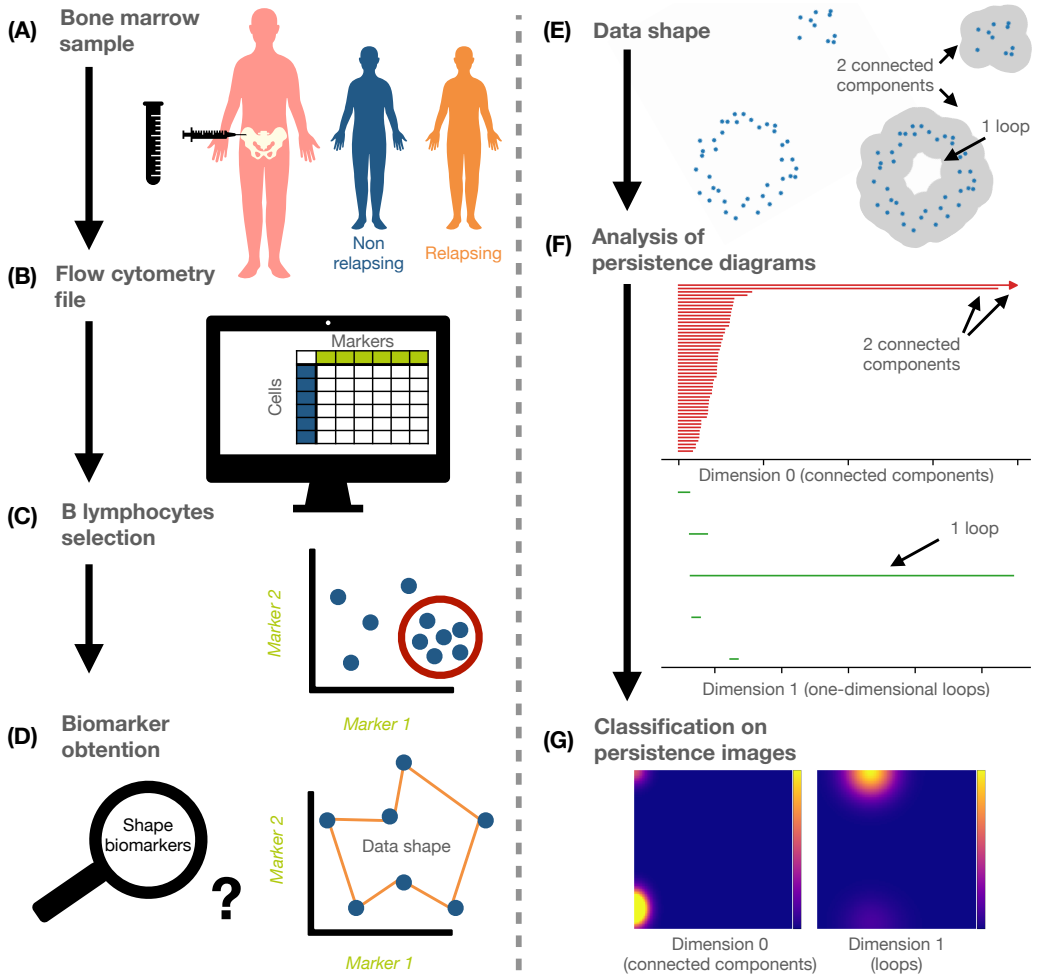

Figure 1: **Flowchart of methodology.** **(A)** We collected bone marrow data from relapsing (R) and non-relapsing (NR) B-ALL patients at the time of diagnosis. **(B)** These samples were analysed via flow cytometry (FC) which gave rise to single-cell data with information on immunophenotypic (IPT) markers. **(C)** We pre-processed these data and selected the patients' B lymphocytes. **(D)** Our aim was to obtain biomarkers from the IPT markers which provide quantitative information about the shape of the data 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.

(see Figure 2). Our data contained 80 NR patients and 16 R patients. We merged data from all bone marrow samples for each patient into a single file (*41*) which included values for the IPT parameters CD10, CD13, CD19, CD20, CD22, CD3, CD33, CD34, CD38, CD45, CD58, CD66c, IGM, cyCD3, cyMPO, cyTDT (see Figure 1(B)). As we were focussing on B-ALL, we also manually selected data corresponding to B-lymphocytes in each patient (see Figure 1(C)). Since this data was too large for efficient computation of PH, we considered subsamples of markers. We identified the IPT markers which hold the most information for classification between R and NR patients. In line with standard clinical practice to assess prognosis, we constructed 2D projections of these high-dimensional point clouds, considering all pairwise combinations of the 16 IPT markers for each patient (see Figure 2(A–C)). We then applied PH (specifically, Vietoris-Rips complex, see 'Methods') to each of the 120 pairwise combinations of the 16 markers 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 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, and for each of the 120 pairwise combinations of IPT markers (see Figure 2(E)).

To classify our patients, we applied Machine Learning methods. We divided the datasets into a discovery set and a validation set, as described in 'Methods'. As a preliminary study, we downsampled the number of NR patients in the discovery set to match the number of R patients, i.e. 13 R and 13 NR patients (we held back 3 R patients for the validation set). We obtained a total of 3120 persistence barcodes for each dimension analysed. 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. We computed receiver operating characteristic (ROC) curves for the classification of R versus NR patients. 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. 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 S1). Only combinations including CD10 or CD20 had a mean AUC>70%.

Next, we considered the validation dataset. The complete set of pairwise combinations resulted in an additional 5520 persistence barcodes in each dimension. As before, these were analysed via Random Forest analysis, with cross-validation. Only IPT markers CD10, CD20, CD38 and CD45 produced mean AUCs>50% (see Data S1-S2). To establish whether these markers carried independent information, we 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 characterise the shape of $CD19^+$ B lymphocyte clouds (*42*), while CD38 is a marker of aberrant B-lymphoblasts (*43, 44*). 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.

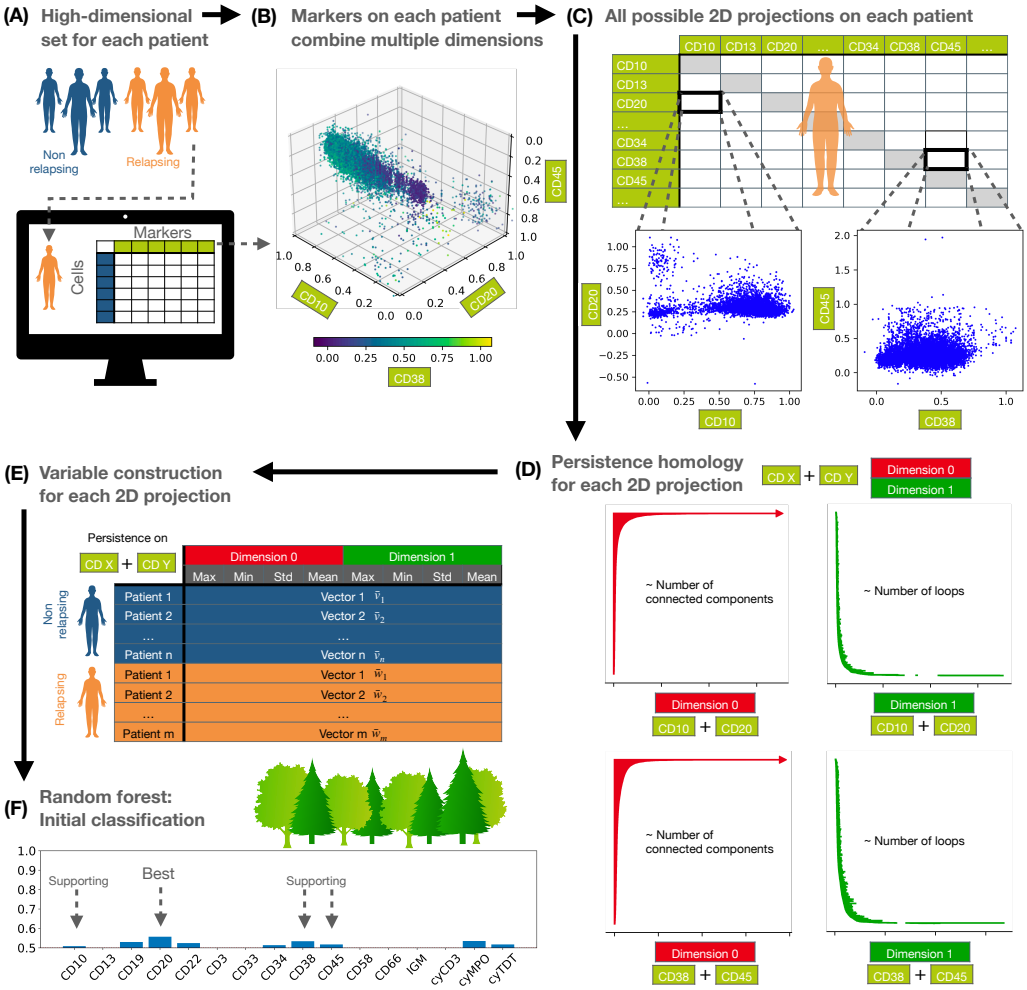

Figure 2: **Workflow for confirming prognostic, state-of-the-art markers**. **(A)** We collected data from non-relapsing (NR) and relapsing (R) patients containing 16 immunophenotypic (IPT) markers. **(B)** To reduce the dimensionality of the data and enable TDA computations, we first selected a subset 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.

PH found that the connected components and loops were more persistent in R patients than in NR patients. Following our selection of the four most informative biomarkers for prognosis (CD10, CD20, CD38 and CD45), 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. We counted the number of topological features in each barcode which persist over scales that exceed a fixed threshold. To obtain PT curves, we performed these computations for a range of threshold values (see Figure 3(A)). We averaged the PT curves for both patient cohorts and all pairwise combinations of biomarkers and found 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. We considered different threshold τ ranges to capture bars with long persistence, e.g. $\tau \in [0.1, 0.18]$ for dimension 0 and $\tau \in [0.04, 0.05]$ for dimension 1, versus bars with medium scale persistence, e.g. $\tau \in [0.03, 0.07]$ for dimension 0 and $\tau \in [0.007, 0.015]$ for dimension 1.

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 $\tau \in [0.04, 0.05]$. An illustrative example for dimension 0 for CD10-CD20 is presented in Figure S9.

Intuitively, there are fewer persistent loops 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 (see Figures 3(B.1) and (B.2), respectively). In the barcodes these 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 SI 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 and the discovery and validation datasets yielded different results (see SI). 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

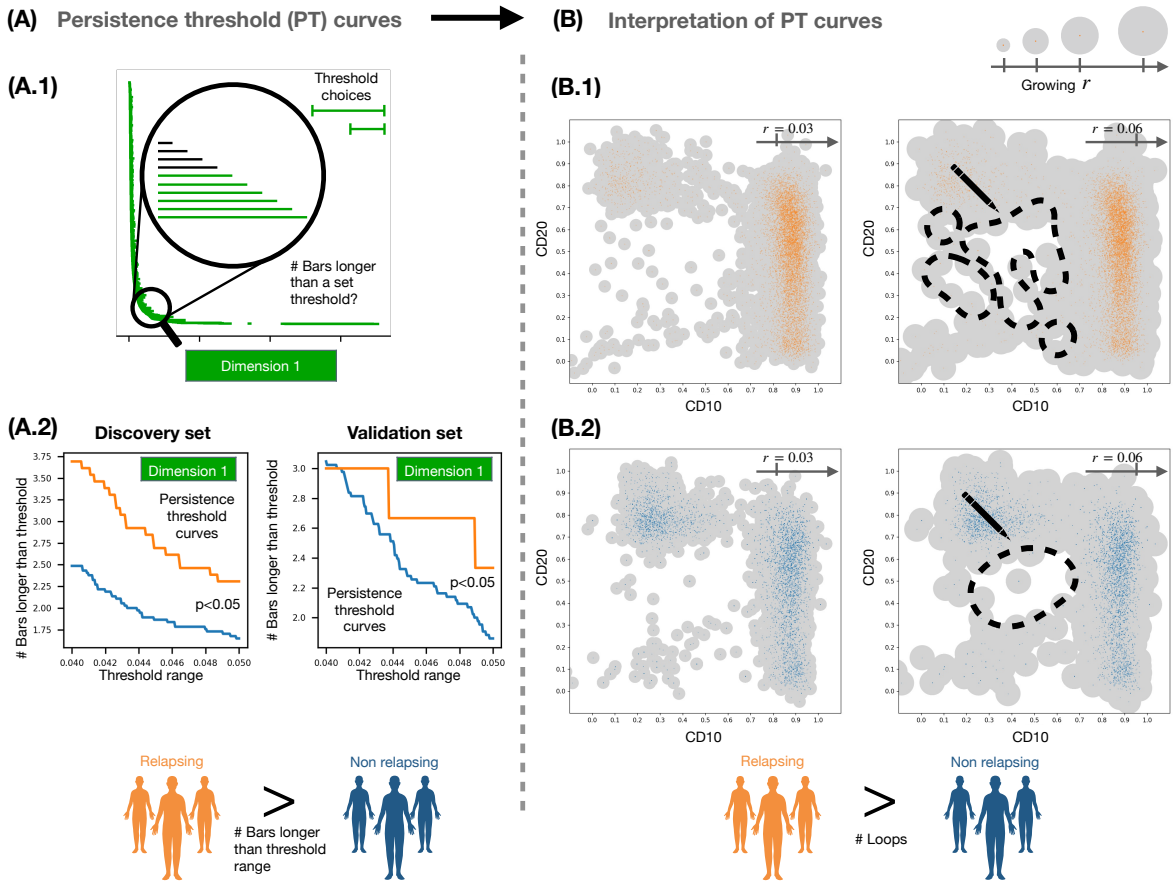

Figure 3: **Analysis and interpretation of output from persistent homology.** **(A)** We computed the number of bars in dimension 1 barcodes whose lengths exceed a set of threshold values (long persistence for $\tau \in [0.04, 0.05]$, medium persistence for $\tau \in [0.007, 0.015]$). The resulting persistence 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).

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)).

We performed clustering to verify the dimension 0 PH results. We considered 4-dimensional

patient point clouds in the discovery set and applied the FlowSom algorithm (*45*) (see 'Methods' and Figures S16-S19) 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. In our previous analyses, we computed simple barcode descriptors from the PH barcodes. To fully exploit the information contained in the barcodes from the 4-dimensional datasets (with biomarkers CD10-CD20-CD38-CD45; see Figure 4(A)), we transformed the barcodes into persistence images (*55*). 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 uppersampling), 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.

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 (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 SI).

Since R patients constituted approximately 20% of the total number of patients, we performed upper-sampling and reran our analysis. These results are presented in Table 2. 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.

We constructed the mean PIs for R and NR patients to compare differences between the two cohorts. Additionally, we computed the coefficients of the decision function in the LR classification and showed by means of a an image the discriminatory areas between both cohorts. These differences depended on the hyperparameters used for the construction of PIs, and were

obtained for the discovery and validation datasets (see Figure S21 for details).

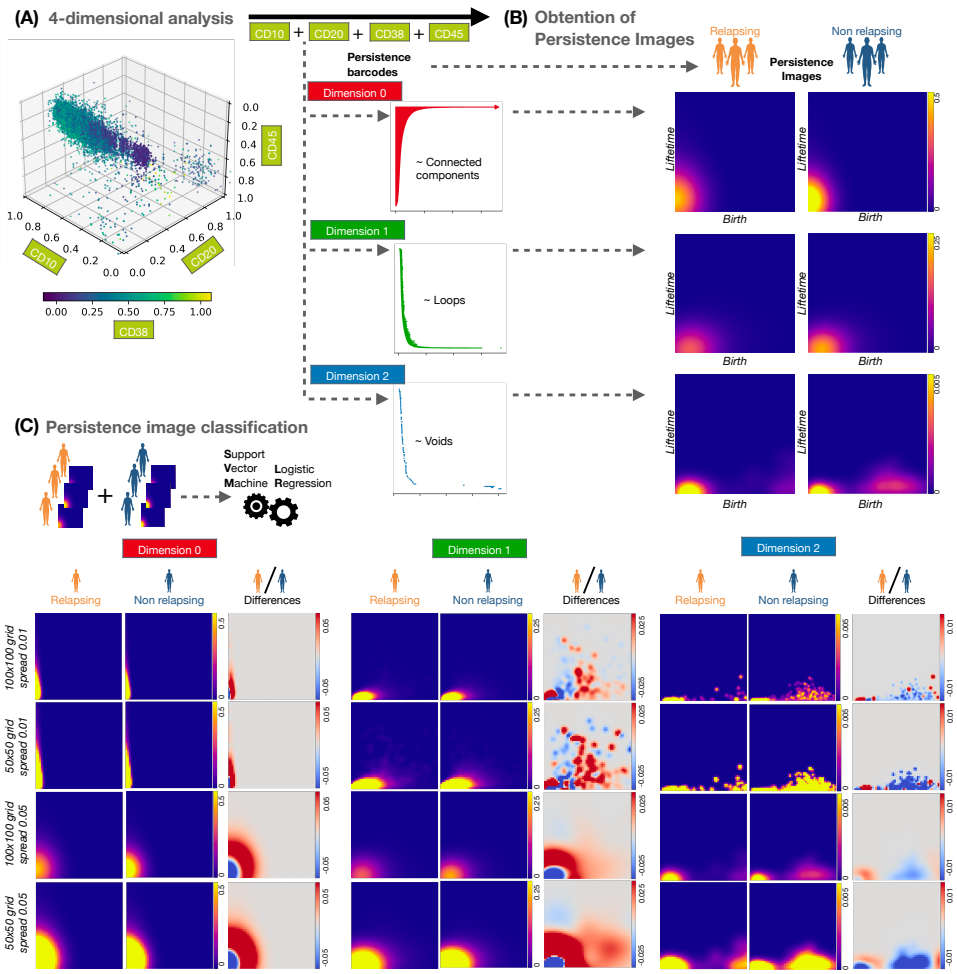

Figure 4: **Analysis using Persistence Images**. **(A)** We computed PH barcodes in dimensions 0 (connected components), 1 (loops) and 2 (voids in 3D) on the 4-dimensional space of biomarkers CD10-CD20-CD38-CD45. **(B)** We transformed the barcodes for each dimension into persistence images, i.e. a matrix of intensity values. This transformation depends on hyperparameters, such as the choice of grid size and the spread of a 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.

Spread	Grid	Method	Dim. 0 Acc.	Dim. 1 Acc.	Dim. 2 Acc.
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

Table 1: **Comparison 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) by performing cross-validation on the discovery and validation datasets, 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 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 data are provided in Table S1.

Spread	Grid	Method	Dim. 0 Acc.	Dim. 1 Acc.	Dim. 2 Acc.
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 2: **Comparison of classification scores of Persistence Images (PIs) with upper sampling**. A classifier based on Support Vector Machines (SVM) was constructed by performing cross-validation on the discovery and validation datasets, using PIs obtained from PH analysis of biomarkers CD10, CD20, CD38, and CD45. These results depend on the dimension of the barcodem 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 data are provided in Table S2.

Discussion and conclusions

Predicting risk of relapse after treatment is one of the major challenges in cancer, particularly when effective second and third-line treatments are available. In B acute lymphoblastic leukemia (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 our present work we focused on predicting risk of relapse based on flow cytometry 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. We combined persistent homology (PH), a method from topological data analysis, with machine learning to obtain quantitative descriptors of the data which distinguish R and NR B-ALL patients at diagnosis with high accuracy. Regarding PH, we based our analysis on 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 (*42*), while CD38 is known to have prognostic value for B-ALL (*39, 44*). Thus, our results confirmed the importance of these four biomarkers and we retained only these four biomarkers for further analyses.

We analysed both the 4D dataset (with the above biomarkers) and pairwise, 2D projections of the data with PH and identified the features which most influence the shape of the point clouds. Interestingly, using simple barcode descriptors extracted from the 2D data we were able to better 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), R patients were characterised by larger numbers of persistent connected components and loops (see Figure 3(B)). These results are consistent with the bone marrow cell population of R patients being more phenotypically diverse than those who do not relapse. In the R patients, such cellular heterogeneity can be linked to non-dominant leukaemic subpopulations that might be the drivers of disease relapse (*46*). 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 flow cytometry data. In particular, the pro-

jections 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. 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 (*39, 43*).

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. Following balancing by uppersampling, 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.

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 (*47, 48*)), PH measures characteristics of the shape of the point cloud in high-dimensional spaces that are not accessible with other methods. These topological features obtained do 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 flow cytometry data.

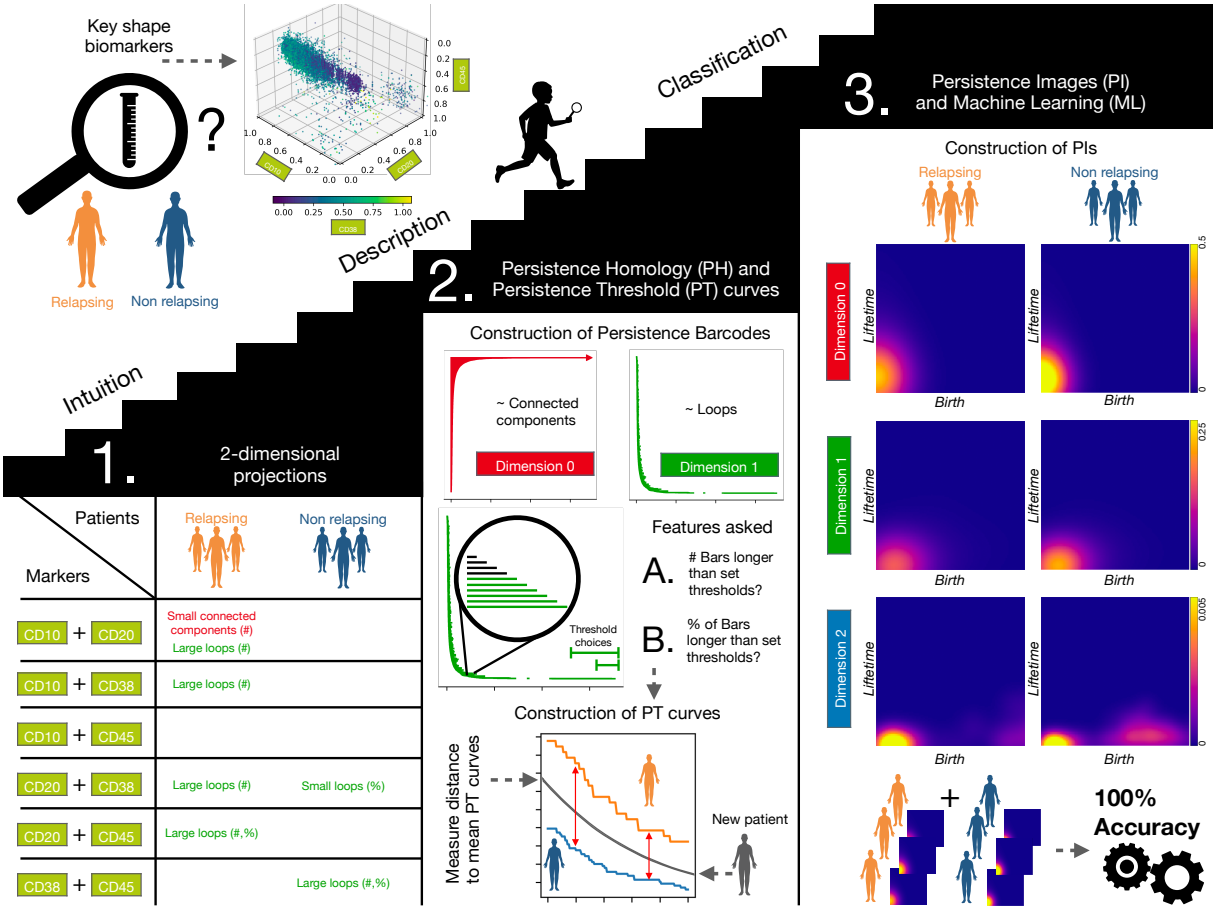

Figure 5: **Summary of results and methodology used**. We consider biomarkers CD10, CD20, CD38 and CD45 to compare FC data from R and NR patients. **1.** Two-dimensional projections of these biomarkers show differences between both cohorts in the number (#) and percentage (%) of both connected components 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 Tables 1 and 2, this could lead to accurate classification of R and NR patients via Machine Learning (ML) with methods such as Support Vector Machines.

1. When only manual inspection of 2D projections of flow cytometry 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 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.

3. If only PH methods are available, 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 also be applied to the CD10-CD20 projection in dimension 0.

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. Additionally, a larger study using bone marrow samples with a larger number of common, standardized IPT markers could be insightful (*49*), 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.

Flow cytometry 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 (*50, 51*) to perform risk stratification and ALL diagnosis. By combining such qualitative data with PH analyses of flow cytometry 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 while also exploiting the high-dimensional character of FC data, whose shape is currently manually assessed in the clinic by focussing on 2D projections of the 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 patient classification and treatment. In other datasets, the same systematic approach could be used to identify informative biomarkers.

We have demonstrated the potential of topological methods to reveal prognostic information in data that is routinely collected from B-ALL patients. Our work is a first step towards automating risk prognosis with high accuracy and interpretable results. 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 BM flow cytometry 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 characterstics can be found in Table S3).

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 FACS-Diva (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) by means of flow cytometry specific packages from Bioconductor. Specifically, files were compensated and transformed with Logicle transform (*52*).

For each patient, we subsampled aliquots to the same number of cells and then brought them into a single file via nearest neighbour imputation (*53*). 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. 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 (*54*) 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 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 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 PT curves: we counted the number 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 SI, as well as in further references (*17, 55, 56*).

Machine learning methods. We divided data into discovery and validation datasets. For classification and prediction, we used Random Forests, a popular and efficient algorithm based on model aggregation ideas (*57*). We also applied Support Vector Machines (SVM) (*58*) for classification by assigning R and NR labels to the data matrices obtained from the PIs of the patients. 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 selected these parameters by randomly splitting the training set and performing internal cross-validations 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 upper-sampling on the first ones.

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) (*59*) and Persim packages (0.1.3) in the Python Scikit-TDA toolbox. FCS files were read using the Python CytoFlow library (1.0) and Flow Cytometry 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 by means of the FlowSom algorithm (*45*), normally used for Flow Cytometry 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`. Flow cytometry 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.

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, and 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, H. Hackl, Z. Trajanoski, 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, J. Barretina, E.T. Gelfand, C.M. Bielski, H. Li, K. Hu, A.Y. Andreev-Drakhlin, J. Kim, J.M. Hess, B.J. Haas, F. Aguet, B.A. Weir, M.V. Rothberg, B.R. Paolella, M.S. Lawrence, R. Akbani, Y. Lu, H.L. Tiv, P.C. Gokhale, A. de Weck, A. Amin Mansour, C. Oh, J. Shih, K. Hadi, Y. Rosen, J. Bistline, K. Venkatesan, A. Reddy, D. Sonkin, M. Liu, J. Lehar, J.M. Korn, D.A. Porter, M.D. Jones, J. Golji, G. Caponigro, J.E. Taylor, C.M. Dunning, A.L. Creech, A.C. Warren, J.M. McFarland, M. Zamanighomi, A. Kauffmann, N. Stransky, M. Imielinski, Y.E. Maruvka, A.D. Cherniack, A. Tsherniak, F. Vazquez, J.D. Jaffe, A.A. Lane, D.M. Weinstock, C.M. Johannessen, M.P. Morrissey, F. Stegmeier, R. Schlegel, W.C. Hahn, G. Getz, G.B. Mills, J.S. Boehm, T.R. Golub, L.A. Garraway, W.R. Sellers, 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. H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification *Discrete Comput. Geom.* **28**, 511-533 (2002).

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

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

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

17. 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).

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

19. L. Wasserman, Topological data analysis. *Ann. Rev. Stats. Appl.* **5**, 501-532 (2018).

20. 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).

21. L. Kanari, Paweł Dłotko, Martina Scolamiero, Ran Levi, Julian Shillcock, Kathryn Hess, Henry Markram, A topological representation of branching neuronal morphologies. *Neuroinformatics* **16**(1), 3-13 (2018).

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

23. 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.

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

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

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

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

28. M. Feng, M. A. Porter, Spatial Applications of Topological Data Analysis: Cities, Snowflakes, Random Structures, and Spiders Spinning Under the Influence. *arXiv:2001.01872* (2020).

29. B. Stolz, J. Kaeppler, B. Markelc, F. Mech, F. Lipsmeier, R. J. Muschel, H. M. Byrne, H. A. Harrington, Multiscale Topology Characterises Dynamic Tumour Vascular Networks. *arXiv:2008.08667* (2020).

30. 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).

31. J.L. Nielson, J. Paquette, A. W. Liu, C. F. Guandique, C. A. Tovar, T. Inoue, K.-A. Irvine, J. C. Gensel, J. Kloke, T. C. Petrossian, P. Y. Lum, G. E. Carlsson, G. T. Manley, W. Young, M. S. Beattie, J. C. Bresnahan, A. R. Ferguson, Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nat. Commun.* **6**, 8581 (2015).

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

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

34. 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).

35. 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).

36. 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).

37. R. Rabadán, Y. Mohamedi, U. Rubin, T. Chu, A. N. Alghalith, O. Elliott, L. Arnés, S. Cal, A. J. Obaya, A. J. Levine, P. G. Cámara, Identification of relevant genetic alterations in cancer using topological data analysis. *Nat. Comm.* **11**, 3808 (2020).

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

39. S. Chulián, Á. Martínez-Rubio, V. M. Pérez-García, M. Rosa, C. Blázquez Goñi, J. F. Rodríguez Gutiérrez, L. Hermosín-Ramos, Á. Molinos Quintana, T. Caballero-Velázquez, M. Ramírez-Orellana, A. Castillo Robleda, J. L. Fernández-Martínez, High-Dimensional Analysis of Single-Cell Flow Cytometry Data Predicts Relapse in Childhood Acute Lymphoblastic Leukaemia. *Cancers* **13**(1), 17 (2021).

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

41. P. Theunissen, E. Mejstrikova, L. Sedek, A. J. van der Sluijs-Gelling, G. Gaipa, M. Bartels, E. Sobral da Costa, M. Kotrová, M. Novakova, E. Sonneveld, C. Buracchi, P. Bonaccorso, E. Oliveira, J. G. te Marvelde, T. Szczepanski, L. Lhermitte, O. Hrusak, Q. Lecrevisse, G. E. Grigore, E. Froňková, J. Trka, M. Brüggemann, A. Orfao, J. J. M. van Dongen, V. H. J. van der Velden, Standardized flow cytometry for highly sensitive mrd measurements in b-cell acute lymphoblastic leukemia. *Blood* **129**, 347–357 (2017).

42. 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).

43. 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).

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

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

46. C. Lutz, P. S. Woll, G. Hall, A. Castor, H. Dreau, G. Cazzaniga, J. Zuna, C. Jensen, S. A. Clark, A. Biondi, C. Mitchell, H. Ferry, A. Schuh, V. Buckle, S.-E. W. Jacobsen, T. Enver, Quiescent leukaemic cells account for minimal residual disease in childhood lymphoblastic leukaemia. *Leukemia* **27** 1204-1207 (2013).

47. 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).

48. 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).

49. J. J. M. Van Dongen, L. Lhermitte, S. Böttcher, J. Almeida, V. H. J. van der Velden, J. Flores-Montero, A. Rawstron, V. Asnafi, Q. Lécrevisse, P. Lucio, E. Mejstrikova, T. Szczepański, T. Kalina, R. de Tute, M. Brüggemann, L. Sedek, M. Cullen, A. W. Langerak, A. Mendonça, E. Macintyre, M. Martin-Ayuso, O. Hrusak, M. B. Vidriales, A. Orfao, EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia* **26**(9), 1908-1975 (2012).

50. A.V. Moorman, A. Enshaei, C. Schwab, R. Wade, L. Chilton, A. Elliott, S. Richardson, J. Hancock, S. E. Kinsey, C. D. Mitchell, N. Goulden, A. Vora, C. J. Harrison, A novel integrated cytogenetic and genomic classification refines risk stratification in pediatric acute lymphoblastic leukemia. *Blood* **124**, 1434–1444 (2014).

51. 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).

52. 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).

53. 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).

54. 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.

55. H. Adams, S. Chepushtanova, T. Emerson, E. Hanson, M. Kirby, F. Motta, R. Neville, C. Peterson, P. Shipman, L. Ziegelmeier, Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18**(1),218-252 (2017).

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

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

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

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

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

61. 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.

62. 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).

63. 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).

64. 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, *arXiv:2101.00523* (2021).

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

66. 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).

Acknowledgments

This work has been 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). We would also like to acknowledge Heather Harrington (Centre for Topological Data Analysis, Mathematical Institute, Oxford) for helpful discussions.