Single-Cell Topological Simplicial Analysis Reveals Higher-Order Cellular Complexity

Baihan Lin

Single-Cell Topological Simplicial Analysis Reveals Higher-Order Cellular Complexity

Baihan Lin[1,2,3*]

[1]Department of Systems Biology, Columbia University, New York, NY, USA
[2]Department of Neuroscience, Columbia University, New York, NY, USA
[3]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA

[*]To whom correspondence should be addressed; E-mail: baihan.lin@columbia.edu.

The absence of a conventional association between the cell-cell cohabitation and its emergent dynamics into cliques during development has hindered our understanding of how cell populations proliferate, differentiate, and compete, i.e. the cell ecology. With the recent advancement of the single-cell RNA-sequencing (RNA-seq), we can potentially describe such a link by constructing network graphs that characterize the similarity of the gene expression profiles of the cell-specific transcriptional programs, and analyzing these graphs systematically using the summary statistics informed by the algebraic topology. We propose the single-cell topological simplicial analysis (scTSA). Applying this approach to the single-cell gene expression profiles from local networks of cells in different developmental stages with different outcomes reveals a previously unseen topology of cellular ecology. These networks contain an abundance of cliques of single-cell profiles bound into cavities that guide the emergence of more complicated habitation forms. We visualize these ecologi-

cal patterns with topological simplicial architectures of these networks, compared with the null models. Benchmarked on the single-cell RNA-seq data of zebrafish embryogenesis spanning 38,731 cells, 25 cell types and 12 time steps, our approach highlights the gastrulation as the most critical stage, consistent with consensus in developmental biology. As a nonlinear, model-independent and unsupervised framework, our method can also be applied to tracing multi-scale cell lineage, identifying critical stages, or creating pseudo-time series.

Introduction

In recent years, technological developments in data visualizations, especially the subfield of topological data analysis (TDA), has illuminated the structure of biological data with features like clusters, holes, and skeletons across a range of scales [1]. The TDA approach has proven to be especially useful with recent advancements in experimental techniques at the single cell resolution, both in genomics and neuroscience, such as radiomics [2] and brain imaging [3, 4]. The utility of topology comes from the idea of persistence, which extract the underlying structures within data while discarding noisy elements in the single cell data collection. Unlike graph-based data like human connectomes, in most time, the high-dimensional data collected from single cell techniques are similiarity-based. Under the assumption that these data was sampled from underlying space $\mathcal{X}$, the goal is to first approximate $\mathcal{X}$ with a combinatorial representation, and then compute some sort of invariant features to recover the topology of $\mathcal{X}$.

The single-cell topological data analysis (scTDA) is one of the first attempts to apply topology-based computational analyses to study temporal, unbiased transcriptional regulation given the single-cell RNA sequencing data [5]. In order to visualize the most invariant features of the entire gene expression data, scTDA clusters low-dispersion genes with significant gene connectivity according to their centroid in the topological representation, and visualize them in low-dimension space with the Mapper algorithm [6]. Computing the cell complexity as the number of genes whose expression is detected in a cell, scTDA observes a mild dependence of library complexity over the timescale of the single cell data of 1,529 cells collected at 5 time points. This is expected because the number of genes expressed by cells in early stages of a developmental process is larger than in the adult case, as pointed out in [7]. As a result, in scTDA the library complexity is not used for any purpose at the topological data analysis and not related to any topological properties.

Intuitively thinking, if we were to introduce a definition for "cell complexity", that characterizes the behaviors of cell-cell coexpression or interactions, the quantities of cell complexity should be agnostic to the number of genes expressed by the cells, and should be different across differentiated cells and across the developmental process. Can we introduce a better summary statistic for the cell complexity that can capture the developmental trajectory with more distinctions between time points? To clarify, unlike the previous definition of "library complexity", which simply quantifies the number of genes expressed in a cell, we wish to define a cell complexity measure to better model higher-order networks and dynamic interactions in single-cell data. Understanding the cell-cell interactions can help identify intercellular signaling pathways and previous analytical studies have focused on computing a communication score between the ligand–receptor pair of interacting proteins [8]. For instance, [9] and [10] infer the intercellular signaling pathways of cell-cell communications by computing the coexpression of all genes or other cell markers. The alternative would be to compute the similarity between gene expression profiles as in [11]. In this work, we aim to focus directly on the cell level, and use the similarity between each cell's gene expression profiles as a graph to compute a topological descriptor of the complexity. The more connected a group of cells are in this similarity graph, the higher the complexity of this group of cell is. There are two major quests in this line of research:

Quest from topological data analysis.

Existing TDA applications usually focus on the low-dimensional graph visualization and the persistent homology of the data (i.e. computing the Betti numbers or barcodes up to dimension 2), because interpreting the biophysical meaning of the geometry and higher dimensional persistent modules is a conceptual challenge. Others have proposed hybrid approaches to combine the merits of data geometry and topology by adaptively selecting the proper thresholds in the pairwise distance matrix of the data points [12, 13]. Another alternative to these low-dimensional TDA methods is the simplicial analysis. Simplicial architecture was first introduced in biological data through the application on human brain connectomes [14], where each connected pairs of neurons are considered an edge to create a graph and the numbers of Rips-Vietoris simplices in dimensions up to 7 are computed at that static graphs comparing with the random graphs. Likewise in our inquiry, we are interested in the intercellular interaction within the same type of cells, the cell complexity [15], rather than the relationships between different groups of cell, as in scTDA. However, the filtration challenge of deriving a graph from the distance-based data by choosing the best threshold, hinders the practical application of such simplicial analysis in these point cloud data.

Quest from single-cell-resolution data.

With the increasingly popular usage of single-cell genomic techniques, it might be possible to infer such cell-cell interaction (or cellular ecology) in a fine resolution. However, as far as we are aware, there are only a few literature exploring the cellular ecology from single-cell RNA sequencing data. For instance, [16] and [17] apply the ecology and multi-agent models to model single-cell systems. We wish to complement this line of work by connecting it to the topological data analysis, where the focus is to model the shape or manifold of the data from the similarity of data points. One challenge of this hybrid direction, is to conceptually

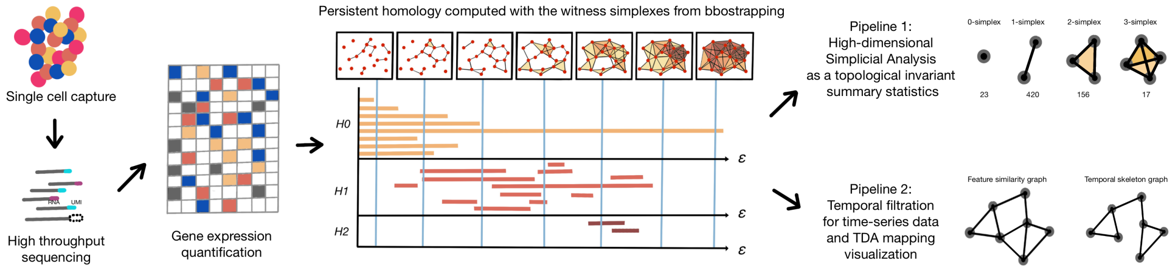
Single cell capture
High throughput sequencing
RNA
UMI
Gene expression quantification
Persistent homology computed with the witness simplexes from bbostrapping
H0
H1
H2
ε
ε
ε
Pipeline 1: High-dimensional Simplicial Analysis as a topological invariant summary statistics
0-simplex
1-simplex
2-simplex
3-simplex
23
420
156
17
Pipeline 2: Temporal filtration for time-series data and TDA mapping visualization
Feature similarity graph
Temporal skeleton graph

Figure 1: **The analytical framework of the single-cell topological simplicial analysis (scTSA)**.

understand the biological meaning behind the dissimilarity of the omic data. For instance, what does it mean if two cells have similar gene expression profiles from each other? Does that indicate a homogeneity if the two cells are from the same tissues, or is it an artifact that the manual labeling or classifications are not perfect? Can we measure the "complexity" of the cell populations based on the heterogeneity or diversity within populations? If we can, how to we evaluate and interpret lower-order versus higher-order "complexity"? The other challenge is the scalability and compariablity of the single-cell data. With the advances of multi-channel high-throughput data collection techniques in biological fields, how to compute the pairwise distances of the point clouds efficiently? In different trials of single-cell experiments, how to make sure that the persistent modules are comparable to one another?

Framework: single-cell topological simplicial analysis (scTSA)

In this study, we propose a topological simplicial analysis (TSA) pipeline (Figure 1) as an exploratory inquiry to solve these three challenges: (1) with the algebraic geometry's definitions of forming higher-order simplices, we can potentially interpret that cliques of higher orders indicates operational units of higher order; (2) with the bootstrapping techniques to sample from the data points collected at each sub-level, we can scale the analysis to large single cell datasets and compare groups of cells quantitatively; (3) with a time delay constraint on the filtration process, we can sort the projected data points of cells into distinct groups of cells collected from the same time stamps. The framework first takes the measurements of the single-cell RNA sequencing data which generates a similarity matrix among the cells based on their gene expression profiles. Other than performing the persistent homology to obtain lower-order topological descriptors of the data, we compute additional higher-order topological descriptors by counting the number of the simplices emerged from the filtration process. In addition, we introduce a technique to extract the temporal skeleton of the developmental processes, called temporally filtrated TDA, and show that the developmental trajectories of cells can be better revealed in this approach comparing to existing TDA mapping techniques.

We begin our presentation in section , with a short overview of mathematical definitions of the single cell data visualization problem and introduction of necessary concepts and definitions in the language of computational topology. Section formulates the topological simplicial analysis pipeline we are proposing as well as numerical tricks applied in the implementation to ensure the scalablity. We apply this single cell Topological Simplicial Analysis (scTSA) to the zebrafish single-cell RNA sequencing data with 38,731 cells, 25 cell types, over 12 time steps [18]. We select the top 103 genes based on the scTDA pipeline from the high-dimensional high-throughput transcriptomic data. In section , we introduce the dataset used to benchmark the method and present the analysis results with their mathematical interpretations to the biological insights. In the last section, we discuss the validity of using our framework to understand the higher-order cellular complexity, and conclude our methods by pointing out several future work directions as the next step of this line of research.

Materials and Methods

Single-cell data in the point cloud space

Genomic measurement and analysis at single-cell resolution has enabled new understandings of complex biological phenomena, such as revealing cellular composition of complex tissues and organisms [19]. Single-cell RNA sequencing (scRNA-seq) techniques measure the gene expression profiles of individual cells through mechanisms like microfludics. For instance, the benchmark dataset of zebrafish embryogenesis [18] that we use in this study, applied Drop-seq, a massively parallel scRNA-seq method to profile the transcriptomes of tens of thousands of embryonic cells [20]. These single cell data are usually point clouds in a finite metric space, a finite point set $S \subseteq \mathbb{R}^d$. Let $d(\cdot, \cdot)$ denote the distance between two points in metric space $\mathcal{Z}$. The assumption is that data was sampled from underlying space $\mathcal{X}$. The goal is to recover topology of $\mathcal{X}$. To accomplish the goal, one needs to first approximate X with a combinatorial representation (e.g. with the simplicial complex), and then compute a topological invariant summary statistics (e.g. with the persistent homology).

Definition of the simplicial and temporal filtration

Given the point cloud data, we then construct a continuous shape on top of the data to highlight the underlying topology and geometry. The process to build such a shape is through a mathematical filtration, which is often a simplicial complex or a nested family of simplicial complexes, that reflects the innate structure of the point cloud data at different scales [21]. Simplicial complexes are high dimensional objects or generalizations of neighboring graphs to represent the cliques of data points, and in another word, a notion of ecology. If we consider all the points in the point cloud data each with a coordinate of their locations in certain embedding, they each occupy a spherical space with the same radius ϵ around them, which are called nerve balls. If the two nerve balls overlap or contact each other, we consider an edge to be formed between them in this graph. The filtration is a process to tune the parameter ϵ from 0 to ∞ and record the families of simplicial complexes generated through the increasingly connected (or "complex") graph.

Usually, the challenge is to extract relevant and useful information about the shape of the data through defining such simplicial complexes from the graph (generated through the filtration process). Rips-Vietoris complex is one of the common choices in practice to compute topological invariants of point clouds, defined as follows: given the vertex set $\mathcal{Z}$, for each pair of vertices a and b edge a-b is included in Rips-Vietoris complex $C(\mathcal{Z}, t)$ if $d(a, b) \leq t$, and a higher dimensional simplex is included in $C(\mathcal{Z}, t)$ if all of its edges are included. Since $C(\mathcal{Z}, t) \in C(\mathcal{Z}, t')$ whenever $t \leq t'$, the filtered Rips-Vietoris complex is a filtered simplicial complex, and also the maximal simplicial complex that can be built on top of its $1-$skeleton, thus a clique complex or a flag complex. Unlike conventional low-dimensional topological data analysis, we computed simplices into high dimension (up to 7) during the entire filtration process. To record the number of cliques, we compute the filtered simplicial complexes and record their cumulative counts across the entire filtration process.

Since the topological data analysis usually only consider the graph constructed by the spatial proximity (i.e. the distance matrix) between the data points in the low-dimensional embedding, it is not clear how to incorporate timestamp information for meaningful inference and visualization when facing the time-series data streams [22]. One approach would be to simply consider the time stamp as the meta data for posthoc labeling of the topological representations. Another alternative would be to consider time as an additional dimension in the filtration process. We present the Temporal Filtration as the following: alongside the conventional sweeping of the parameter ϵ from 0 to ∞, we set another parameter τ to indicate a hard constraint in edge forming between two points. In another word, only if the time stamp difference between the two data points is within the time delay limit τ, can two nerve balls, if spatially proximal enough (less than ϵ), form an edge in between. On the other hand, if the time stamp difference between the two data points is larger than τ, even if they are spatially proximal enough (less than ϵ), they cannot form an edge. Given the problem settings, one can either set a reasonable time delay limit τ given the domain knowledge, or tune τ from 0 to ∞, similar to the filtration process on the spatial filtration parameter ϵ. The later approach can potentially extract temporally invariant topological summary statistics.

Topological data analysis with persistent homology

Following the definition above, an abstract simplicial complex is given by a set $\mathcal{Z}$ of vertices or $0-$simplices, for each $k \leq 1$ a set of $k-$simplices $\sigma = [z_0, z_1, \ldots, z_k]$ where $z_i \in \mathcal{Z}$, and for each $k-$simplex a set of $k+1$ faces obtained by deleting one of the vertices. A filtered simplicial complex is given by the filtration on a simplicial complex $\mathcal{Y}$, a collection of subcomplexes $\{\mathcal{Y}(t) | t \in \mathbb{R}\}$ of $\mathcal{Y}$ such that $\mathcal{Y}(t) \subset \mathcal{Y}(t')$ whenever $t \leq t'$. The filtration value of a simplex $\sigma \in \mathcal{Y}$ is the smallest t such that $\sigma \in \mathcal{Y}(t)$.

Topological data analysis methods usually involve computing the persistent homology [23]. The Betti numbers help describe the homology of a simplicial complex $\mathcal{Y}$. The Betti number value BN_k, where $k \in \mathbb{N}$, is equal to the rank of the $k-$th homology group of $\mathcal{Y}$. The Betti intervals over the filtration process help describe how the homology of $\mathcal{Y}(t)$ changes with t. A $k-$dimensional Betti interval, with endpoints $[t_{\text{start}}, t_{\text{end}})$, corresponds to a $k-$dimensional hole that appears at filtration value t_{start}, remains open for $t_{\text{start}} \leq t < t_{\text{end}}$, and closes at value t_{end}. Figure 2 is a schematic diagram outlining how to perform a filtration process (by sweeping the ϵ), document the "birth" and "death" of each complexes (the colored lines of various length in the chart), and generate this as a "barcode" representation [24] for the downstream analyses.

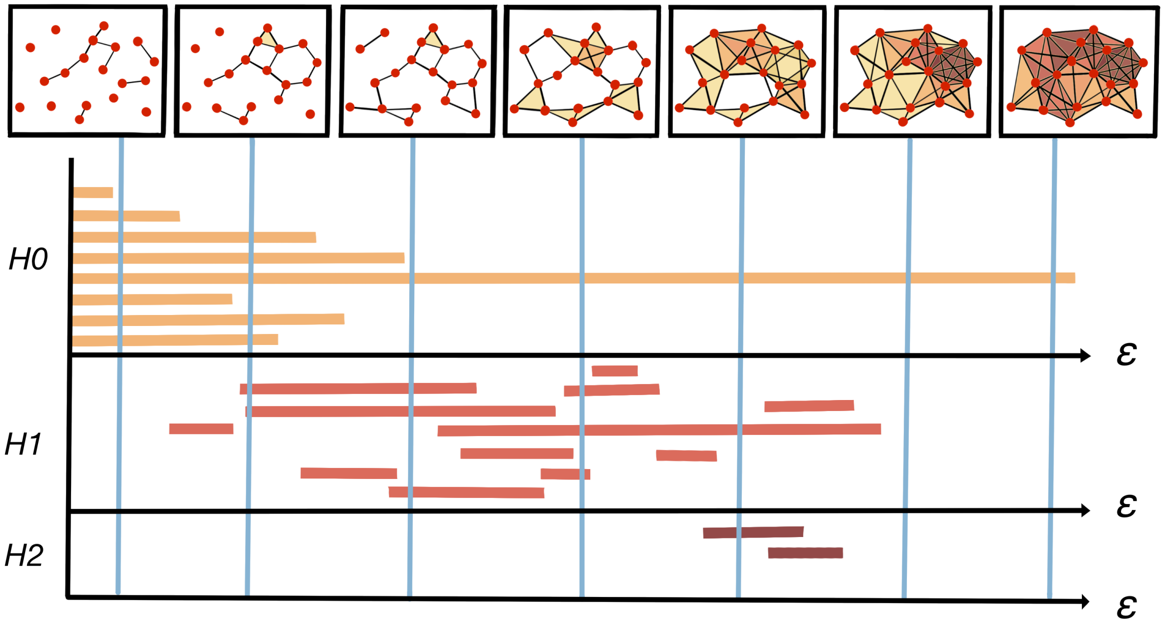

Figure 2: **Persistent homology via mathematical filtration.** In this schematic diagram, a point cloud of 19 data points are presented in a low-dimensional embedding space. In the filtration process, a parameter ϵ is swept from 0 to the maximum pairwise distance within the point cloud, indicating a distance threshold under which the two points can form an edge to become one connected component in the graph. In another word, a nerve ball of radius ϵ grows around each point cloud, and an edge will form if two nerve balls touch. H_n indicates the n-th homology group, i.e. the formation of the simplex complexes of order n, with 0-simplex to be the nodes (or clusters), 1-simplex to be the edges between two nodes, 2-simplex to be the loops (or triangles in this case), 3-simplex to be the tetrahedrons and so on. Each colored line indicates the "lifespan" of a simplex, with its starting point to be its "birth" (or first appearance) and ending point to be its "death" (or disappearance due to the two nerve balls fully overlapping). In this example, the persistent homology of the data cloud is presented in the form of a "barcode" representation. The birth and death of the simplicial complexes up to the order 2 are recorded when the filtration process gradually sweeps the distance threshold.

Topological data visualization with low-dimensional mapping

To build and visualize the topological representation of the point cloud data, we use the Mapper algorithm [25] through the implementations provided by Kepler-Mapper [1] with modifications

[1] https://github.com/scikit-tda/kepler-mapper

for temporal filtration at https://github.com/doerlbh/tkMapper. In brief, a dissimiliarity matrix is computed from the preprocessed RNA-seq data by taking the pairwise correlation distance. This metric space was then reduced to a low-dimensional embeddings with the multi-dimensional scaling [26]. Given this embedding, the point cloud data are chopped into coverings of hypercubes with a 50% percentage of overlapping between the cubes[2]. Then for each hypercube, the data points within the cube are then clustered with single-linkage rule. This step further aggregates all the points into a network in which each vertex corresponds to a cluster and each edge corresponds to a nonvanishing intersection between the clusters. As defined in section , If temporal filtration is applied, then edge forming is also controlled by the additional time delay constraint that the clusters are formed with both spatial and temporal proximity, and the edges would only exist between two clusters if all points in the two clusters are within the time delay limit τ. Once we reach a network representation, the network can eventually be visualized with force-directed algorithms for insights.

Empirical simplicial computation with lazy witness complex

As single cell data has different noise granularity across cell types and data collection procedures [27], the number of cells collected in each time points and different cell types (as in the analyzed developmental study [18]) can vary in different magnitude, making direct simplicial computation incomparable. For these larger datasets, if we include every data point as a vertex, the filtrated simplicial complexes can quickly contain too many simplices for efficient computation. To solve this numerical inconsistency issue, we instead extract the lazy witness complexes by sampling m data points [23] with a sequential maxmin procedure [28], setting a nearest neighbor inclusion of 2 (as in the term "lazy")[3]. The computation of the witness com-

[2] The choice of 50% is empirically determined by our dataset. We vary the overlap parameter among 25%, 50% and 75%, and 50% gives the best clustering effect.

[3] The selection of m depends on the scale of the dataset. The bigger the sample size m is, the better the estimate. However, since different partitions of the data points have varying sizes. For instance, if there are only 50 data

plex in high dimensions is implemented with the JPlex software [29] and Matlab. The codes to reproduce the empirical results can be accessed at https://github.com/doerlbh/scTSA.

Control models for the topological simplicial analysis

Usually for binary connectivity data (like brain connectome), Erdős-Rényi random graph [30] can be used as control models. However, in similarity-based data, the average connectivity probability is entirely dependent on the filtration factor. To avoid this caveat, we take a different approach by permuting the pairwise distances of the data points, which is equivalent to a weighted version of the Erdős-Rényi random graph. In this way, the low-dimensional embeddings computed by the multidimensional scaling can form different connectivity profiles while maintaining the same distance distribution. Then we apply the same topological data analysis pipelines to the embeddings computed from the pairwise distance matrices from both the actual data and the control models.

To this point, we propose a formal definition of cellular complexity, as the *normalized n-simplicial complexity*, SC_n, a family of summary statistics with an increasing order n:

$$SC_n = \frac{\#simplex_n^{data}}{\#simplex_n^{null}} \tag{1}$$

where SC_n is computed by taking the ratio between the number of the simplicial complexes for a certain order n computed from the actual data, and the number of those computed from the control models. Empirically, we compute the SC_n with the order *n* from 1 to 7, as the summary statistics characterizing the ecology among the data points with cliques and cavities of increasing modularities.

points collected in time step 1, while there are more than 100 points in other times steps, then the maximum of m that can be picked is 50.

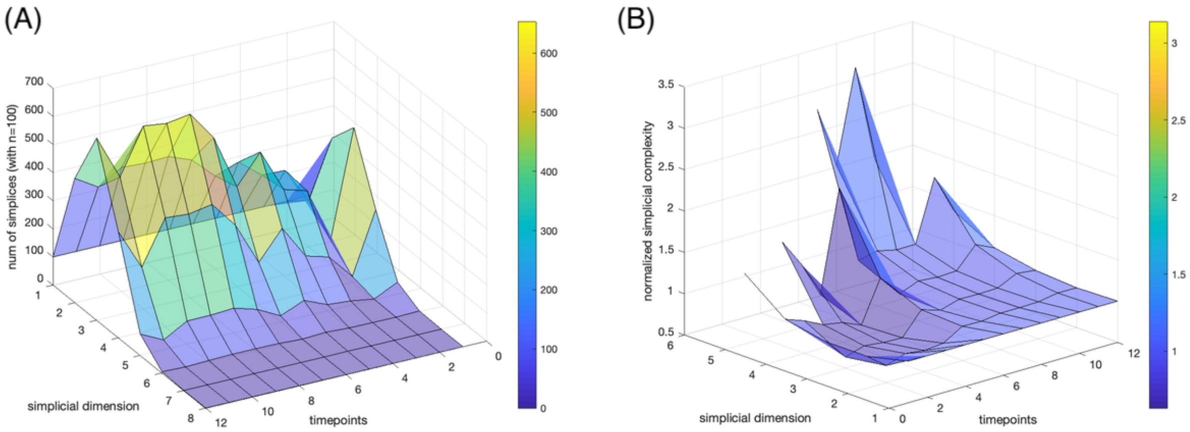

Figure 3: **Simplicial dynamics across developmental stages.** In (A) and (B), we sample 100 data points in each time point of the single cell data, apply the multidimensional scaling (MDS) to reduce its dimension to 2, and compute the simplicial complexes up to dimension 7. The color and the surface height in the z-axis indicates of the size of the computed topological summary statistics. (A) The number of n-simplices is computed from the sampled data points in each time points. (B) The normalized n-simplicial complexity, i.e. the normalized number of n-simplices, is computed as the ratio of the number of the n-th order simplicial complexes from the data over the number of those from the null models. The normalized simplicial complexity of higher order appears to be well above 1 in certain developmental stages with a distinctive separation between the 5th and 6th time points.

Results

We benchmark the scTSA method on the zebrafish single-cell RNA sequencing data with 38,731 cells, 25 cell types, over 12 time steps [18]. The data has dimension of 103 corresponding to the expression levels of 103 significant genes selected by the scTDA pipeline [5]. For each time points, we sample 100 data points, and embed them with multidimensional scaling (MDS) of dimension 2 to preserve their distance information[4]. Upon the MDS embedding, we compute the filtrated simplicial complexes up to the dimension of 7.

The TSA pipeline identifies the simplicial complexity to vary over the time, suggesting a

[4] The choice of two dimensions is an empirical consideration. The computation of mathematical filtration can be expensive, while MDS is known to preserve the geometric information well even with two dimensions.

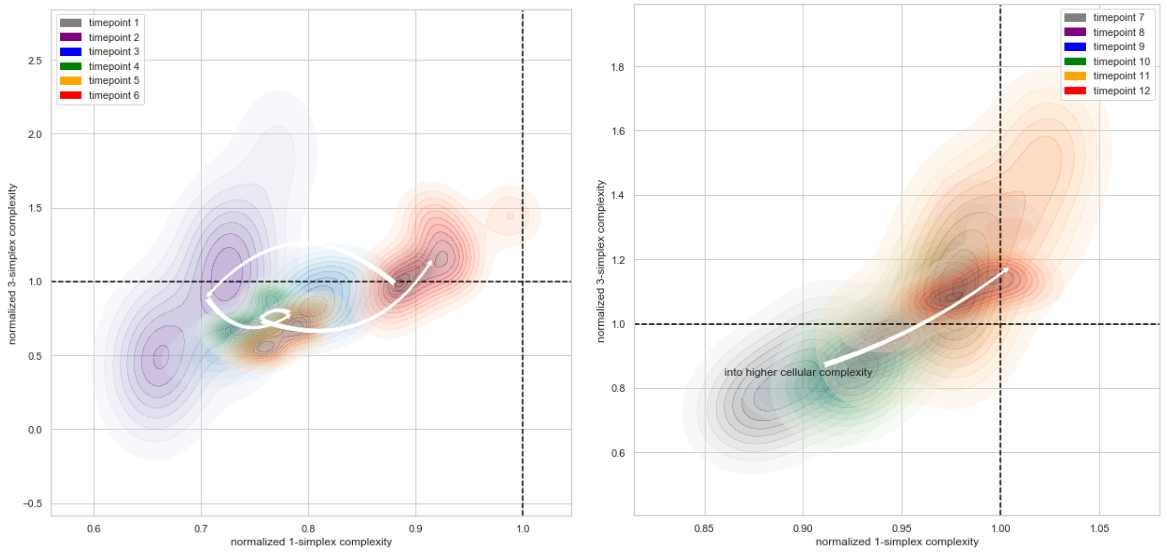

Figure 4: **Simplicial dynamics across developmental stages.** To investigate the tradeoff between the higher-order and the lower-order simplicial complexity in the developmental stages, the normalized 3-simplicial complexity is mapped against the normalized 1-simplicial complexity. The color indicates different time points. The arrow indicates the transition between the centroids in each groups of time points. A transition of lower-order and higher-order normalized cell complexity is marked with the white trajectories across sequential time points.

potential better summary statistic with better distinction (Figure 3). The normalized simplicial complexity (computed as the ratio of the number of simplicial complexes discovered within the data over the number of those discovered within the null model) suggests an abundance of high-dimensional simplices over the null models. The existence of a significant number of high-dimensional simplices is observed for the first time in the single cell level. In all time points, the number of simplices of dimensions larger than 1 in the null model was far smaller than those found in the actual data. In addition, we observe this relative differences between what we discover in null models and the actual data increase drastically when the dimensions are higher. Furthermore, the number of low-dimensional simplices (up to dimension 3) of the data appears to be equal or smaller than the null models (with normalized complexity less than 1), suggesting a possible transfer from lower order clique structure to a higher-order structure.

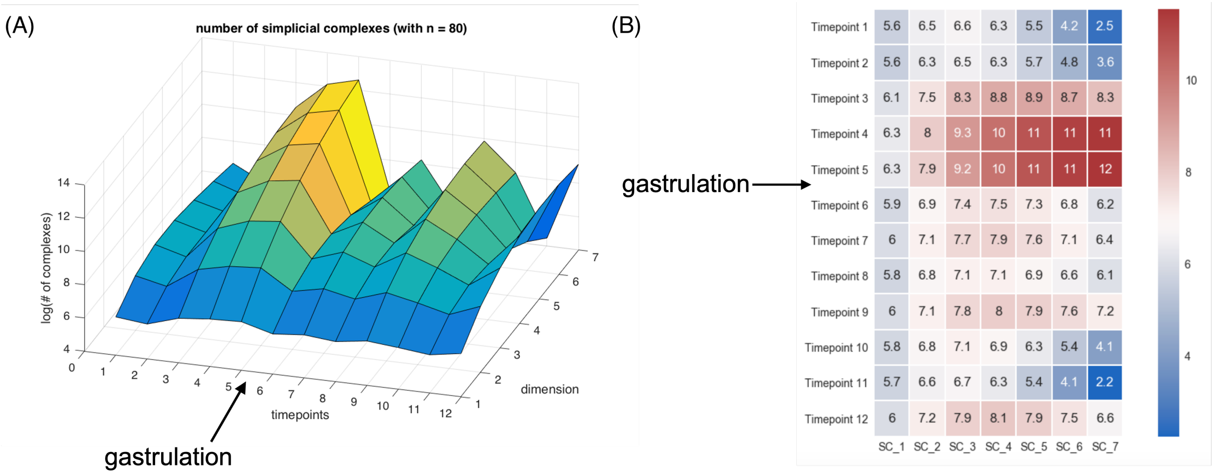

Figure 5: **scTSA identifies the critical stage of cellular complexity change.** To showcase the flexibility of the scTSA approach with different low-dimensional embeddings and sample sizes, we sample 80 data points in each time point of the single cell data and apply principal component analysis (PCA) to extract the first two component before apply the scTSA. (A) The number of n-simplices in the log scale to highlight the drastic change of cellular complexity between the 5th and 6th time points. (B) The heatmap of the normalized n-simplicial complexity across the time points supports the observation. To draw insights on the developmental trajectories, we perform a visualization of the network extracted from the topological data analysis (TDA) with the Mapper algorithm. This type of visualization aims to identify subpopulations of cells that form modular clusters and sparse connections between the clusters.

In order to investigate the tradeoff between the higher-order and the lower-order simplicial complexity in the developmental stages, we map the normalized 3-simplicial complexity against the normalized 1-simplicial complexity. Figure 4 suggests an overall above-null higher-order complexity starting from the 5th time point, and an overall below-null lower-order complexity in a monotonically increasing direction since the 2nd time point. Comparing to the null model, the presence of a much larger numbers of cliques across a range of dimensions in the single cell data suggests that the connectivity between these cells might be highly organized into numerous fundamental building blocks with increasing complexity.

The scTSA approach has the flexibility to different low-dimensional embeddings and sample

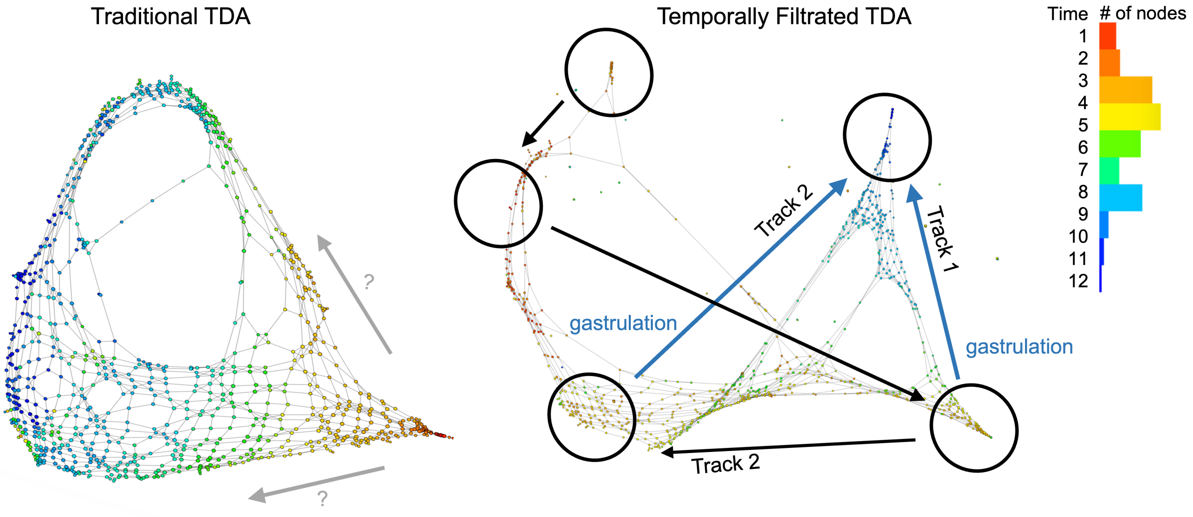

Figure 6: **Temporal filtration identifies the critical stage of cellular complexity change.** The color indicates the time points and each node corresponds to a small cluster of cells collected at the same time points. The conventional TDA mapping (the left panel) identifies a bifuraction structure, but there are spatial locations that has a mixture of clusters that belong to non-consecutive time points. This makes the identifications of a developmental pathway challenging. When applying the temporal filtration (the right panel), the mapping identifies a cleaning separation of two tracks, or two subpopulations of cells that evolves in the gastrulation stage, matching the observation in our summary statistics from the algebraic topology.

sizes. To demonstrate, we sample 80 data points in each time point and apply the principal component analysis (PCA) to extract the first two component. Figure 5 demonstrate the log scale of the number of $n-$simplices. It shows that the gastrulation stage is a very critical stage in vertebrate development, matching the established understanding in the developmental biology that it is a process where the embryo begins the differentiation process to develop into different cell lineages [31]. Before gastrulation, the embryo is a continuous epithelial sheet of cells. After the gastrulation stage, organogensis starts where individual organs develop within the newly formed germ layers.

This observation is further supported by the visualization of topological data analysis mapping. Figure 6 compares the network visualizations with and without the temporal filtration. We

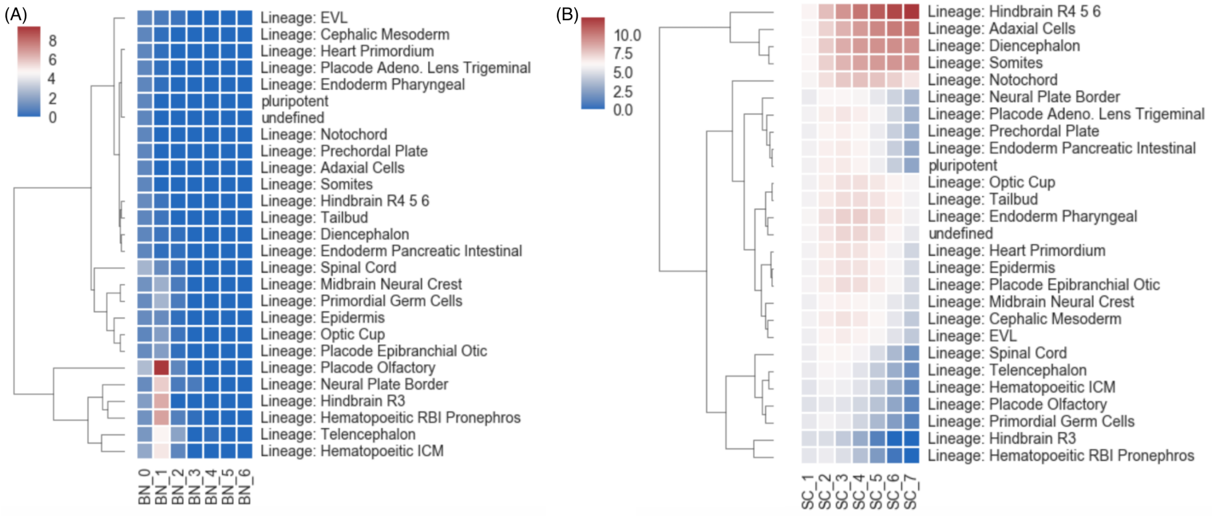

Figure 7: **Cell lineage tracing with the simplicial statistics.** In this analysis, the hierarchical clustering is performed on the summary statistics of transcriptomic data of different cell types. (A) The heatmap and clustering result using the Betti numbers as the clustering features. (B) The heatmap and clustering result using the normalized simplicial complexity as the features for the hierarchical clustering.

observe that, when color-labelled with the time points, the conventional topological data analysis outlines a progression of cellular development, but there are many subsequent time points in the middle of earlier timesteps. For instance, we see there are many dark blue nodes from the 11th or 12th time points in the middle of web where the majority of the nodes are earlier stages from the 5th to 7th. When using the temporal filtration (with τ set to be just 1 time step), we observe that the network has much more skeleton and branches, where each branching nodes consist only of points of the same time stamp. The gastrulation stage, which happens between the 5th and 6th time points, appears to belong to two separate tracks, supporting the hypothesis that after the notochord and prechordal plate territories become transcriptionally distinct, the gastrulation process refines the boundary between the two cellular populations [18].

These filtrated simplicial architectures may also offer insights in cell lineage tracing. As in the previous analyses, we sample 50 cells from each cell types and apply scTSA over the PCA embedding. We perform the hierarchical clustering of the summary statistics computed from the transcriptome data of different cell types. We compare the result using the proposed normalized simplicial complexity versus the one using the Betti numbers (which is more conventionally used in many downstream topological data analyses). As shown in Figure 7, the normalized simplicial complexity offers a more reasonable clustering performance as a more distinctive summary statistics than the Betti numbers by themselves.

Discussion

What is cellular complexity and what does the higher-order complexity mean? As an inquiry to this question, we explore the possibility of introducing the mathematical notion of higher-order simplicial complexes into analyzing distance-based single cell data. Benchmarked on a single cell gene expression data with multiple developmental stages, we propose the single-cell Topological Simplicial Analysis, and demonstrate that the simplicial complexity can be a well-defined summary statistic for celluar complexity.

This investigation provides a scalable, parameter-free, expressive and unambiguous mathematical framework to represent the cellular complexity with its underlying structure. Locally, these structures are characterized in terms of the simplicial complexes. Globally, these structures are characterized in terms of the cavities formed by these simplices. This framework reveals an intricate topology of cellular similarity which includes a vast number of cliques of cells and of the cavities that bind these cliques together. These topological summary statistics that captures the relationships among the high-dimensional cliques uncover the transcriptional differences in the connectivity of cells of different types during graph reconstruction.

From the scTSA visualization, we discover, for the first time in any single cell data, an abundant number and variety of higher-order cliques and cavities. Comparing to the control models, the framework measures a much higher number of high-dimensional cliques and cavities in the graph construction filtration process. The critical stage identified by the framework matches the current understanding in the developmental biology. Comparing with the statistics of Betti numbers, the normalized simplicial complexity demonstrates better distinctions between time points and cell types.

There are potentially different questions we can explore: Can we determine developmental stages without physiological features? Can we generate pseudo-time series based on single cell sequencing data? And most importantly, does the vast presence of high-dimensional cliques suggest that the interaction between these cells is organized into fundamental building blocks of increasing complexity? Through this inquiry with topological simplicial analysis, we can form such hypothesis that the cells organize themselves into high-dimensional cliques for certain functional or developmental reasons. Further research includes developing mechanistic theories behind the emergence of such high-dimensional cellular cliques and experimentally testing these hypotheses to reveal the missing link between functions and cellular complexity.

Conclusions

In summary, our work describes a novel scalable and unsupervised machine learning method that tackles several technical challenges in bioinformatics: (1) a lack of time-series analytical methods in quantifying the underlying temporal skeleton within the manifold of the similarities among data points; (2) a lack of scalable computational methods to characterize single-cell sequence signals in the scale of 10k+ data points, while the single-cell sequencing data are dominating the bioinformatics in recent few years; (3) a lack of insight and interpretation that connects the mathematical language of algebraic topology to the physical references to the biological phenomena. The time-series problem is especially a topic that is applicable beyond the application proposed in our work, and thus a major interest in the unsupervised machine learning communities dealing with high-dimensional time series signals.

References

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

[2] Lorin Crawford, Anthea Monod, Andrew X Chen, Sayan Mukherjee, and Raúl Rabadán. Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. *Journal of the American Statistical Association*, 115(531):1139–1150, 2020.

[3] Manish Saggar, Olaf Sporns, Javier Gonzalez-Castillo, Peter A Bandettini, Gunnar Carlsson, Gary Glover, and Allan L Reiss. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nature communications*, 9(1):1–14, 2018.

[4] Angkoon Phinyomark, Esther Ibanez-Marcelo, and Giovanni Petri. Resting-state fmri functional connectivity: Big data preprocessing pipelines and topological data analysis. *IEEE Transactions on Big Data*, 3(4):415–428, 2017.

[5] Abbas H Rizvi, Pablo G Camara, Elena K Kandror, Thomas J Roberts, Ira Schieren, Tom Maniatis, and Raul Rabadan. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature biotechnology*, 35(6):551, 2017.

[6] Gunnar Carlsson. Topological pattern recognition for point cloud data. *Acta Numerica*, 23:289–368, 2014.

[7] Gunsagar S Gulati, Shaheen S Sikandar, Daniel J Wesche, Anoop Manjunath, Anjan Bharadwaj, Mark J Berger, Francisco Ilagan, Angera H Kuo, Robert W Hsieh, Shang Cai, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, 367(6476):405–411, 2020.

[8] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

[9] Douglas Arneson, Guanglin Zhang, Zhe Ying, Yumei Zhuang, Hyae Ran Byun, In Sook Ahn, Fernando Gomez-Pinilla, and Xia Yang. Single cell molecular alterations reveal target cells and pathways of concussive brain injury. *Nature communications*, 9(1):1–18, 2018.

[10] Eun-Yeong Oh, Stephen M Christensen, Sindhu Ghanta, Jong Cheol Jeong, Octavian Bucur, Benjamin Glass, Laleh Montaser-Kouhsari, Nicholas W Knoblauch, Nicholas Bertos, Sadiq MI Saleh, et al. Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. *Genome biology*, 16(1):1–22, 2015.

[11] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107, 2018.

[12] Baihan Lin and Nikolaus Kriegeskorte. Adaptive geo-topological independence criterion. *arXiv preprint arXiv:1810.02923*, 2018.

[13] Baihan Lin. Geometric and topological inference for deep representations of complex networks. In *Proceedings of the Web Conference 2022*, 2022.

[14] Michael W Reimann, Max Nolte, Martina Scolamiero, Katharine Turner, Rodrigo Perin, Giuseppe Chindemi, Paweł Dłotko, Ran Levi, Kathryn Hess, and Henry Markram. Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in computational neuroscience*, 11:48, 2017.

[15] Baihan Lin. Cliques of single-cell RNA-seq profiles reveal insights into cell ecology during development and differentiation. In *ISMB*, Basel, Switzerland, July 2019.

[16] Jill A Gallaher, Susan C Massey, Andrea Hawkins-Daarud, Sonal S Noticewala, Russell C Rockne, Sandra K Johnston, Luis Gonzalez-Cuyar, Joseph Juliano, Orlando Gil, Kristin R Swanson, et al. From cells to tissue: How cell scale heterogeneity impacts glioblastoma growth and treatment response. *PLoS computational biology*, 16(2):e1007672, 2020.

[17] Sarah R Amend, Sounak Roy, Joel S Brown, and Kenneth J Pienta. Ecological paradigms to understand the dynamics of metastasis. *Cancer letters*, 380(1):237–242, 2016.

[18] Jeffrey A Farrell, Yiqun Wang, Samantha J Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, 2018.

[19] Tomer Kalisky and Stephen R Quake. Single-cell genomics. *Nature methods*, 8(4):311–314, 2011.

[20] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[21] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*, 2017.

[22] Baihan Lin. Topological data analysis in time series: Temporal filtration and application to single-cell genomics. *arXiv preprint arXiv:2204.14048*, 2022.

[23] Vin De Silva and Gunnar E Carlsson. Topological estimation using witness complexes. *SPBG*, 4:157–166, 2004.

[24] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[25] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2, 2007.

[26] Al Mead. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(1):27–39, 1992.

[27] Andre J Faure, Jörn M Schmiedel, and Ben Lehner. Systematic analysis of the determinants of gene expression noise in embryonic stem cells. *Cell systems*, 5(5):471–484, 2017.

[28] Henry Adams and Gunnar Carlsson. On the nonlinear statistics of range image patches. *SIAM Journal on Imaging Sciences*, 2(1):110–117, 2009.

[29] Harlan Sexton and MV Johansson. Jplex. *url: http://comptop. stanford. edu/programs/j*, 2008.

[30] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[31] Scott F Gilbert and MJF Barresi. Developmental biology, 2016.

Acknowledgments

The authors thank members of the Rabadan, Cecchi and Kriegeskorte laboratories for helpful discussion, especially Dr. Raul Rabadan, Dr. Ioan Filip, Dr. Luis Aparicio and Dr. Guillermo Cecchi for their helpful advice.

Funding:

This work was financially supported in part by the Systems Biology Fellowship awarded by Columbia University and the research training grants awarded by the National Science Foundation and the National Institutes of Health.

Author contributions:

Following the CRediT model,

Conceptualization: BL

Methodology: BL

Investigation: BL

Visualization: BL

Supervision: BL

Writing—original draft: BL

Writing—review & editing: BL

Competing interests:

Authors declare that they have no competing interests.

Data and materials availability:

The codes and data to reproduce all analytical and empirical results can be accessed and reproduced at the repository https://github.com/doerlbh/scTSA.