Topologically-based parameter inference for agent-based model selection from spatiotemporal cellular data

Alyssa R. Wenzel[1], Patrick M. Haughey[1], Kyle C. Nguyen[2,5], John T. Nardini[3], Jason M. Haugh[4], Kevin B. Flores[1,5*]

[1]Department of Mathematics, North Carolina State University, Raleigh, 27607, NC, USA.

[2]Sandia National Laboratories, Albuquerque, 87123, NM, USA.

[3]Department of Mathematics and Statistics, The College of New Jersey, Ewing, 08628, NJ, USA.

[4]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, 27695, NC, USA.

[5]Center for Research in Scientific Computation, North Carolina State University, Raleigh, 27607, NC, USA.

*Corresponding author(s). E-mail(s): kbflores@ncsu.edu;

Abstract

Advances in spatiotemporal single-cell imaging have enabled detailed observations of cell population dynamics and intercellular interactions. However, translating these rich data sets into mechanistic insight remains a significant challenge. Agent-based models (ABMs) are a bottom-up computational framework for investigating the emergent behavior of cell populations that can arise from rules defining the interactions between individual neighboring cells, while topological data analysis (TDA) provides robust descriptors of spatial organization. We present TOPAZ (TOpologically-based Parameter inference for Agent-based model optimiZation), a computational pipeline that integrates TDA with approximate Bayesian computation (ABC) and Bayesian model selection to identify biologically plausible ABMs from spatiotemporal cellular data. TOPAZ uses persistent homology to quantify spatial features of cell trajectories and combines this topological information with parameter inference via ABC and model comparison using the Bayesian Information Criterion. We validate TOPAZ using simulations of collective fibroblast movement, demonstrating its ability to accurately recover model parameters and distinguish between a baseline ABM and an extended model that incorporates alignment interactions. Our results and open-source code demonstrate the utility of TOPAZ as a generalizable framework for mechanistic inference and model discrimination in spatial single-cell analysis.

Keywords: topological data analysis, approximate Bayesian computation, agent based modeling, spatiotemporal cellular data, model selection

1 Introduction

Advances in microscopy imaging and live-cell tracking have enabled the high-throughput collection of single-cell spatiotemporal dynamics across a wide range of biological systems [1, 2]. A central challenge in computational biology is to develop integrative methods that extract mechanistic insight from multi-scale datasets, particularly linking subcellular signaling dynamics to population-level behavior, such as collective cell migration at the tissue level. Connecting these scales with computational inference and mechanistic modeling has the potential to provide data-driven insight into processes that are fundamental to morphogenesis, wound healing, and cancer progression [3, 4]. Agent-based models (ABMs) have recently been used to computationally investigate the mechanisms through which cell-to-cell heterogeneity, derived from single-cell transcriptomics and proteomics data, gives rise to population-level behaviors, e.g., by leveraging single-cell data for ABM initialization and calibration [5]. Topological data analysis (TDA) is another computational framework that has been utilized in bioinformatics studies for analyzing single-cell omics data. For example, TDA has been used to infer the geometric structure of cellular niches and spatial expression gradients by using persistent homology to identify spatially coherent gene expression domains and to characterize tissue architecture in high-dimensional spatial omics datasets [6, 7]. Yet, while TDA excels at quantifying structural organization features in spatial single-cell data, such as gradients, boundaries, and connectivity, it is not inherently mechanistic and does not capture causal or dynamical processes. Conversely, agent-based models (ABMs) provide a bottom-up framework for simulating individual cell behaviors and interactions, but are computationally intensive, sensitive to parameter choices, and often difficult to fit or validate directly against data. This motivates the development of hybrid approaches that integrate TDA and ABMs to bridge intracellular, intercellular, and population-level scales to enable mechanistic inference grounded in topological summaries of spatial omics data.

Previous work exemplified that TDA can be used for extracting multi-scale structural features from high-dimensional data, which can be used to inform dynamical modeling frameworks such as ABMs [8, 9]. While these studies demonstrate the utility of TDA in extracting structural features, they primarily offer descriptive insights without establishing direct links between topological summaries and specific mechanistic parameters within dynamical models. This gap was affiled in part by Nguyen et al., who developed a framework combining TDA with Approximate Bayesian Computation (ABC) for parameter inference in an ABM of collective fibroblast motion [10]. Nguyen et al. demonstrated how persistent homology can be applied to single-cell trajectory data to extract summary statistic for performing likelihood-free inference with ABC. However, the model considered in their study considered a single ABM that only included repulsion and attraction forces and did not consider additional biologically motivated mechanisms such as directional alignment or anisotropy [11, 12]. Moreover, while Nguyen et al. exemplified that parameter inference with TDA was possible, the framework lacked a model selection component that could formally distinguish between competing hypotheses of cellular interaction. These shortcomings highlight the need for a unified data-driven pipeline that couples spatiotemporal single-cell data, TDA, ABMs and statistical model comparison for developing mechanistically interpretable and data-constrained simulations of cellular population dynamics. Specifically, model comparison is a step often overlooked in ABM pipelines which is essential to ensure that model complexity is justified by explanatory power [13, 14]. We demonstrate how integrating model selection into the TDA-ABC framework yields a novel and generalizable approach for analyzing complex collective cell behaviors in biomedical contexts.

2 Agent-based models

In this work, we focus on agent-based models (ABMs) of single-cell dynamics informed by time-lapse imaging data. Specifically, we investigate how simple interaction rules can give rise to emergent collective behaviors, such as the parallel (and anti-parallel) streaming—or *fluidization*—observed in dense fibroblast monolayers. To capture these dynamics, we adopt the D'Orsogna ABM [10, 15]. We

use the D'Orsogna model, referred to here as Model_{DO}, as an example of ABM for which minimal interaction rules can give rise to many distinct population-level movement patterns, some of which recapitulate collective cell migration. Model_{DO} describes the dynamics of self-propelled agents with uniform mass, subject to linear drag and pairwise interactions governed by attractive and repulsive forces, with magnitudes C_a and C_r, and length scales ℓ_a and ℓ_r. These parameters capture biologically motivated behaviors such as contact inhibition of locomotion (repulsion) and matrix-mediated long-range attraction, as described in [16–18]. A full description of the Model_{DO} ABM can be found in the Supplementary Material Section 1.1.

To assess whether distinct mechanisms of interaction can be reliably distinguished using summary statistics derived from topological data analysis (TDA), we compare two ABMs: the baseline D'Orsogna model (Model_{DO}) and an extended model with alignment interactions (Model_{AL}). We simulate synthetic trajectory data from each model across a range of parameter settings and then fit the opposing model to that data using approximate Bayesian computation (ABC). By applying model selection via the Bayesian Information Criterion (BIC), we assess whether the inference pipeline can correctly recover the generative model. This simulation-based framework serves as a proof-of-concept for using TDA-informed ABC to distinguish between competing mechanistic hypotheses—a critical step toward future applications involving real spatial single-cell datasets that incorporate omics-level molecular information.

We nondimensionalize the D'Orsogna model and express the dynamics in terms of two key parameters: $\text{C} = C_a/C_r$ and $\text{L} = \ell_a/\ell_r$, representing relative attraction strength and interaction range, respectively. These parameters are sufficient to reproduce a variety of collective behaviors in self-propelled cell populations [10].

To capture alignment effects observed in fibroblasts, we extend the model to include a third parameter W, which governs the strength of orientation coupling (see Supplementary Material Section 1.2 for a derivation of Model_{DO}). This extension is biologically motivated by experimental observations of contact-mediated alignment [19] and theoretical models of directional organization in cell populations [20]. In our extended model (Model_{AL}), agents adjust their heading based on local alignment while retaining the original D'Orsogna speed dynamics. Thus, the baseline model estimates two parameters (C, L), while the alignment model estimates three $(\text{C}, \text{L}, \text{W})$. Although our motivating application is fibroblast migration, the models are general and relevant to a range of active biological systems governed by short- and long-range interactions.

3 TOpologically-based Parameter inference for Agent-based model optimiZation (TOPAZ)

We developed a computational pipeline named "TOPAZ" that combines topological data analysis (TDA) computed from agent-based model (ABM) simulations, approximate Bayesian computation (ABC), and Bayesian information criteria (BIC) for ABM model selection derived from spatiotemporal cellular data. Fig 1 displays the computational steps in the pipeline. Step 1: TDA attributes in the form of Betti-Crocker diagrams are computed from spatiotemporal point clouds, in the form (x, y, θ), derived from experimental data or ABM simulations. Step 1.5: Visualization with dimensionality reduction techniques, such as t-distributed stochastic neighbor embedding (t-SNE) or principal component analysis, are recommended. Step 2: TDA diagrams are used with ABC as summary statistics for parameter inference as in [10]. Step 3: The posterior distribution from ABC, the number of parameters, and the number of data points are used to calculate the BIC score for the ABM. Step 4: The ABM with the lowest BIC score is selected as the optimal model.

4 Dimensionality reduction for visualization

To aid in affiling the inverse problem's identifiability, t-distributed stochastic neighbor embedding (t-SNE) is employed as a nonlinear dimensional reduction method, which is particularly effective when

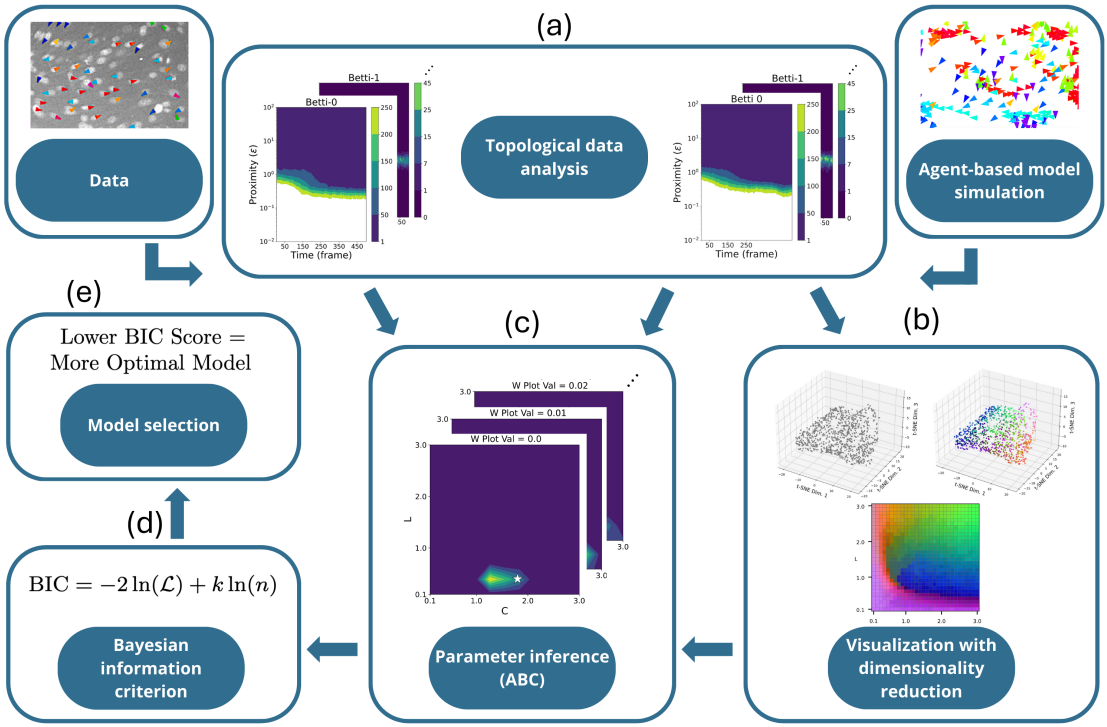

Fig. 1 A computational pipeline for agent-based model selection. Our pipeline called **TO**pologically-based **P**arameter inference for **A**gent-based model optimi**Z**ation (TOPAZ) has the following steps: (a) From data or an agent-based model (ABM) simulation, topological data analysis in the form of Betti-Crocker plots. (b) (Optional) Visualization with dimensionality reduction techniques such as t-SNE or PCA. (c) Parameter inference techniques such as approximate Bayesian computation. (d) Bayesian information criterion score calculation. (e) Model selection based on lowest BIC score.

computationally feasible. By reducing high-dimensional data to a three-dimensional plot, t-SNE helps visualize complex relationships more clearly. We simulated the new Alignment model for 30 equally spaced values of C and L ranging from 0.1 to 3.0, and 11 equally spaced values of W ranging from 0.0 to 0.1, resulting in 9,900 total combinations. The simulation involved 40,000 iterations, 100 time points $\times$ 200 approximations $\times$ 2 Betti numbers (zero and one). We applied t-SNE to reduce the 40,000 Betti-Crocker plots (example shown in (a) of Fig 2) to a three-dimensional representation. A random 10% subset of the 9,900 points is visible in the t-SNE reduced space in (b) of Fig 2, with each point colored according to its 3D coordinates, converted into an RGB value (shown in (c) of Fig 2). This color data was then extracted to create eleven 30 $\times$ 30 2D t-SNE plots, one for each W value (examples shown in (d) and (e) of Fig 2; all eleven plots can be found in Supplementary Figure S1).

5 Results

5.1 Simulation Study Design and Hypotheses

We developed a broadly applicable computational pipeline **TO**pologically-based **P**arameter inference for **A**gent-based model optimi**Z**ation (TOPAZ) for selecting which agent-based model (ABM) is the most optimal fit for topological attributes derived from spatiotemporal cellular data. We tested this pipeline in a simulation study comparing fits of the ABM Model_{AL} to data simulated by the ABM Model_{DO} and vice versa. Our hypothesis is that because Model_{AL} is a one-parameter extension of the Model_{DO} model, Model_{AL} will be selected as the best model when the data are simulated

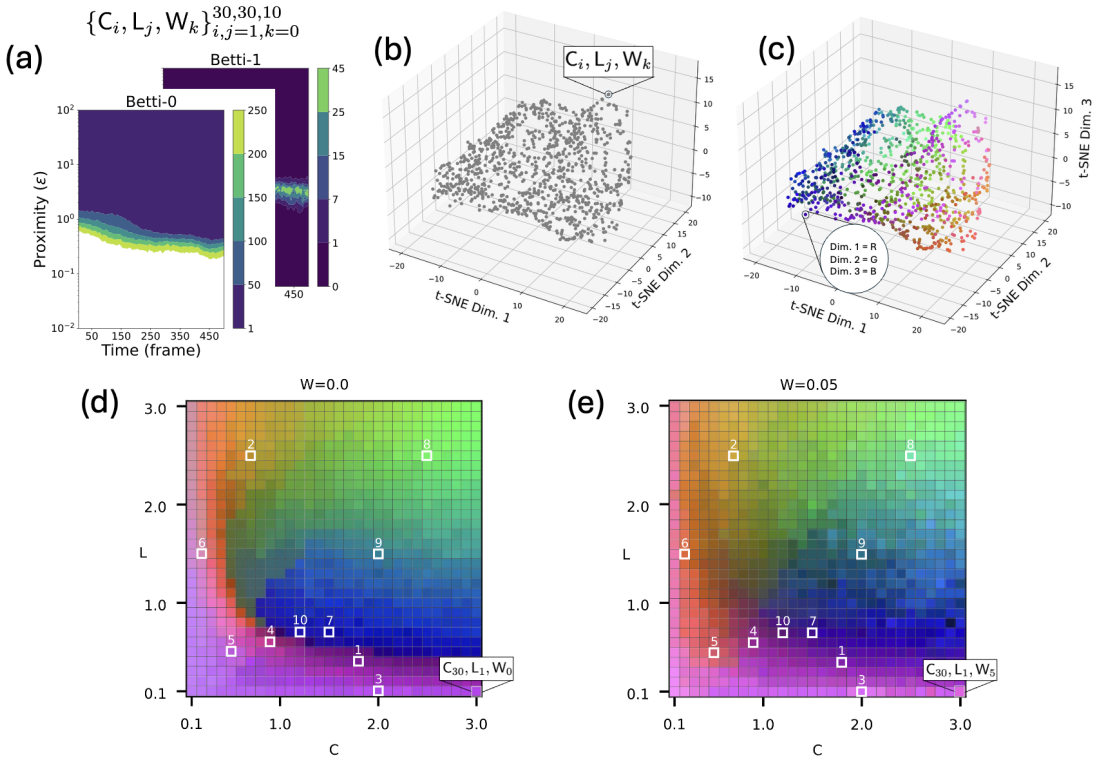

Fig. 2 Visualization with dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE). (a) Betti-Crocker plots of the ABM simulation. (b) 40,000 dimensions reduced to 3 dimensions. Only 10% of the 9,900 simulations are shown. (c) Each point is colored based on their 3D position where (Dim. 1, Dim. 2, Dim. 3) = (R, G, B). (d and e) A 2D extraction from the 3D plot. Each point in the 3D plot has a C, L, and W value. These are plotted in 2D grid of C and L for each value of W. Left (d) is the W=0 slice and right (e) is the W=0.05 slice. The ten selection points were chosen for further analysis to cover a diverse range of behaviors (see Table 1).

from Model_{AL} but not from Model_{DO}. By leveraging Bayesian information criterion (BIC) in addition to Topological Data Analysis (TDA), our TOPAZ pipeline penalizes for models with additional parameters and therefore only selects more complex ABMs when that complexity is represented in the topology of the data. Similarly, we also hypothesized that when data are simulated from the Model_{DO} model, TOPAZ will select Model_{DO} instead of Model_{AL} as the optimal model because the additional parameter in Model_{AL} does not provide a significantly better fit to the topological information in the Model_{DO} simulation, as determined by BIC (see step 4, Fig 1).

5.2 Parameter Selection and Synthetic Data Generation

We selected a set of 10 ground-truth parameter vectors for each ABM corresponding to different values of (C,L,W) for Model_{DO} and Model_{AL} where Model_{DO} corresponds to $W = 0$ (Table 1, columns 1-3). Our ground-truth parameter choices were based on previous study showing a wide representation of topologically distinct spatiotemporal features in simulations of the Model_{DO} model. The same values of C and L were used for both Model_{DO} and Model_{AL}, and we chose $W = 0.05$ for simulating data for Model_{AL} (see parameter choices 1-10 in (d) and (e) of Fig 2). Representative snapshots of ABM simulations from Model_{DO} and Model_{AL} are shown for parameter choices 1 and 5 in Fig 3, visually exemplifying topological differences between the cases when $W = 0$ (No Alignment) and $W = 0.05$ (Alignment). See Supplementary Material for the corresponding ABM simulation movies.

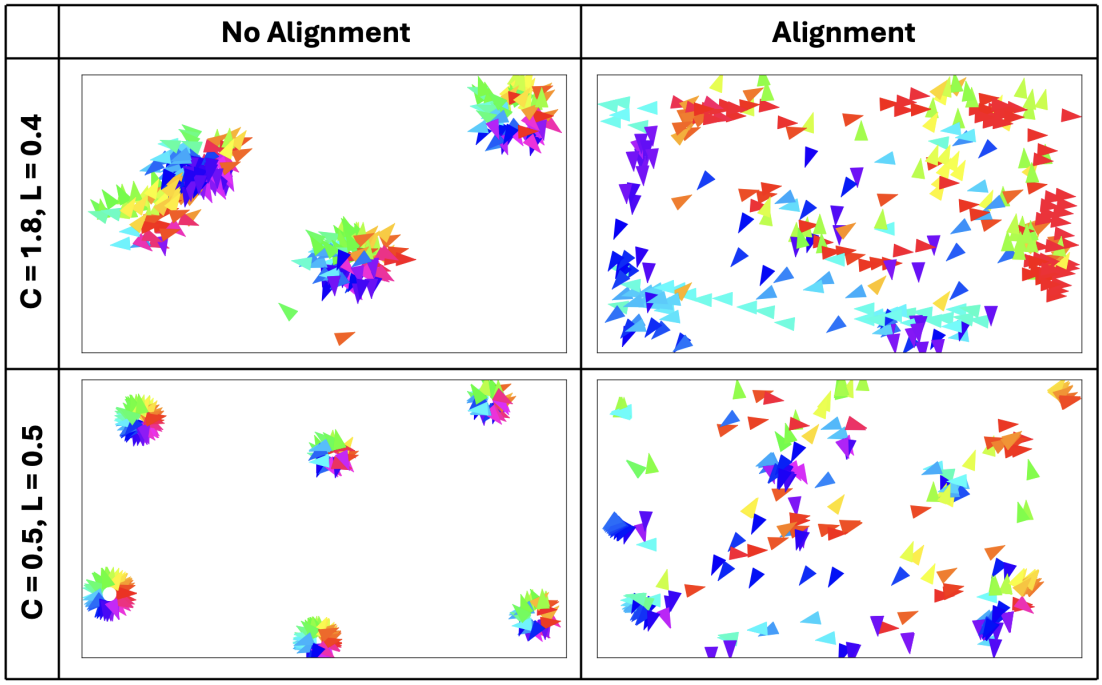

Fig. 3 Visualization of Model_{AL} output for two different C and L combinations. Simulation snapshots are shown without alignment case (W=0.0, left) and with alignment case (W=0.05, right). The arrows and colors represent the direction the cells are moving. These snapshots are at the end of the simulation (t=120, last frame).

5.3 Parameter Inference Using ABC

Using simulations generated from Model_{DO} or Model_{AL}, we computed TDA summaries from the generate point clouds and then used them for parameter inference within ABC (steps 1 and 2, Fig 1). Visualizing the ABC posterior distributions exemplifies that the median of the posterior was close to the ground-truth parameter value when the ground-truth case was for W = 0 or W = 0.05 (Fig 4). The complete set of posterior plots for all W values is displayed in Supplementary Figure S2. Concordantly, the Betti-Crocker plots at the ground-truth and estimated parameter values (median of posterior) exemplify that the corresponding topological diagrams are visually similar (Fig 5).

5.4 Model Selection via BIC

To quantitatively and comprehensively summarize the visual observations we made, BIC was calculated for Model_{AL} and Model_{DO} against simulated data generated at each of the 10 chosen parameter vectors. Table 1 displays BIC scores as well as the difference between the Model_{AL} and Model_{DO} BIC scores for comparison (right-most column). Except for one chosen parameter vector (see Supplementary Material Section 1.3), we found that the BIC for Model_{AL} was lowest for all the Model_{AL} simulated data and the BIC for Model_{DO} was lowest for all the Model_{DO} simulated data. These results confirm that our TOPAZ pipeline can be used to quantitatively compare ABMs and select an optimal model for spatiotemporal cellular data.

6 Discussion

We introduced TOPAZ, a novel methodology to facilitate model selection for agent-based models (ABMs). This algorithm integrates concepts from topological data analysis (TDA), approximate

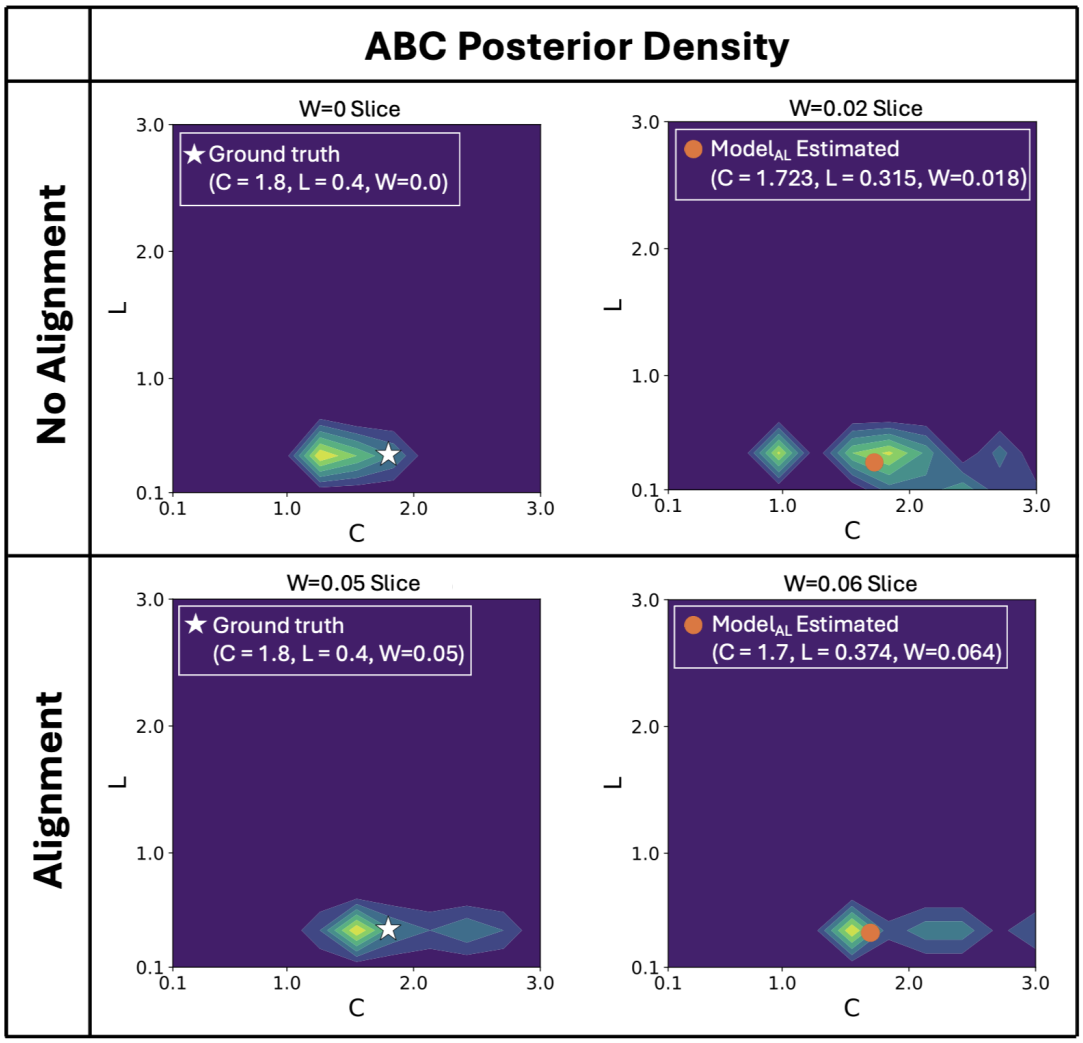

Fig. 4 Examples of ABC posterior density plots. The W-slices that contain the ground truth values (white stars) and ABC-estimated parameter values (orange dots) are displayed on top of the ABC posterior densities for Model_{AL}, in the cases where the data are generated from a model without alignment (Top) and with alignment (Bottom). The complete selection of ABC posterior density plots can be found in Supplementary Figure S2.

Bayesian computation (ABC), and statistical modeling to achieve this. We apply the Vietoris-Rips filtration to summarize the topological signature of an ABM simulation using agents' output data. For a given dataset and library of candidate models, we use ABC to infer the posterior distribution of each considered model's parameter values. The Bayesian information criterion (BIC) then identifies which model best describes the data with a reasonable amount of complexity. Model selection facilitates the identification of relevant biological mechanisms in the presence of noisy data [21].

To showcase this novel methodology, we developed an extension of the D'Orsogna Model that includes an agent alignment force. This model incorporates an alignment term that promotes directional coordination among neighboring cells. We found that simulations from the new model are more likely to exhibit fluidization behavior, i.e., the formation of groups of cells moving in coherent directions antiparallel to adjacent groups. This recapitulates experimental observations of confluent fibroblast populations, where cells self-organize into dynamic, parallel/anti-parallel streams [10]. We then considered a suite of ABM simulations that were simulated from different parameter values.

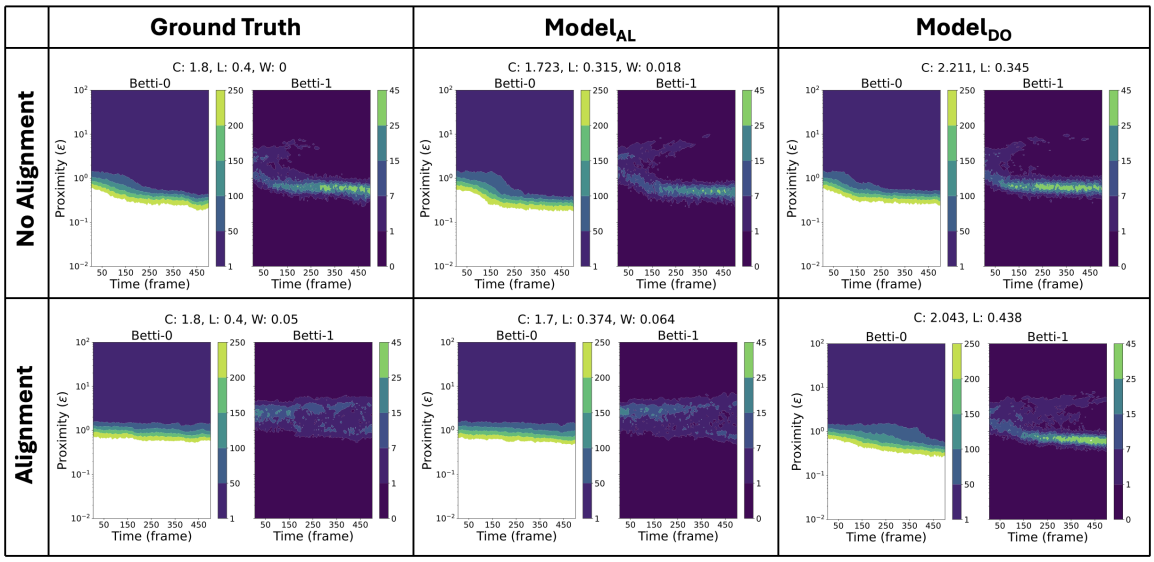

Fig. 5 Crocker plots summarizing the topology of the ground truth and ABC simulations. For the ground truth values and ABC-estimated values for Model_{AL} and Model_{DO}, we constructed Crocker plots for Betti-0 and Betti-1 (left, middle, and right, respectively) for the case where there is no alignment (W=0.0, top) and alignment (W=0.05, bottom). The color legend represents the contour levels corresponding to the number of simplicial complexes at each proximity value and frame number; counts exceeding 250 are shown in white for visualization purposes.

Ground-Truth Parameters			Model$_{AL}$			Model$_{DO}$		BIC		SSE		BIC
C	L	W	$C_{estimate}$	$L_{estimate}$	$W_{estimate}$	$C_{estimate}$	$L_{estimate}$	Model_{AL}	Model_{DO}	Model_{AL}	Model_{DO}	Model_{AL} - Model_{DO}
1.8	0.4	0	1.723	0.315	0.018	2.211	0.345	4.08e+06	1.99e+06	4.00e+06	1.91e+06	2.09e+06
0.7	2.5	0	0.744	2.646	0.021	0.705	2.517	2.00e+06	1.93e+06	1.93e+06	1.86e+06	7.47e+04
2	0.1	0	0.945	0.165	0.012	1.416	0.121	3.31e+06	5.46e+05	3.24e+06	4.72e+05	2.77e+06
0.9	0.6	0	0.986	0.217	0.034	0.934	0.556	3.62e+06	2.20e+06	3.55e+06	2.12e+06	1.43e+06
0.5	0.5	0	0.735	0.189	0.013	0.551	0.406	2.83e+07	4.27e+06	2.82e+07	4.20e+06	2.40e+07
0.2	1.5	0	0.216	0.838	0.017	0.182	1.472	5.70e+06	1.60e+06	5.63e+06	1.53e+06	4.10e+06
1.5	0.7	0	2.152	0.732	0.007	1.708	0.647	9.33e+05	3.83e+05	8.60e+05	3.10e+05	5.50e+05
2.5	2.5	0	2.668	2.707	0.013	2.583	2.568	5.35e+05	3.75e+05	4.61e+05	3.01e+05	1.60e+05
2	1.5	0	2.303	1.48	0.006	2.111	1.488	3.84e+05	3.32e+05	3.10e+05	2.59e+05	5.14e+04
1.2	0.7	0	1.971	0.745	0.007	1.313	0.702	1.91e+06	5.01e+05	1.84e+06	4.28e+05	1.41e+06
1.8	0.4	0.05	1.7	0.374	0.064	2.043	0.438	7.16e+05	2.20e+07	6.43e+05	2.19e+07	-2.13e+07
0.7	2.5	0.05	0.607	2.77	0.059	0.706	2.61	5.28e+05	5.83e+06	4.54e+05	5.76e+06	-5.30e+06
2	0.1	0.05	0.612	0.197	0.045	0.518	0.914	5.47e+06	7.08e+07	5.39e+06	7.07e+07	-6.54e+07
0.9	0.6	0.05	1.125	0.431	0.069	1.417	0.528	6.97e+05	1.81e+07	6.23e+05	1.80e+07	-1.74e+07
0.5	0.5	0.05	0.472	0.418	0.068	0.421	1.816	3.69e+06	2.58e+06	3.62e+06	2.51e+06	1.11e+06
0.2	1.5	0.05	0.173	1.57	0.069	0.54	2.409	1.08e+06	4.74e+06	1.01e+06	4.67e+06	-3.66e+06
1.5	0.7	0.05	1.625	0.69	0.064	1.556	0.579	4.51e+05	2.05e+06	3.78e+05	1.98e+06	-1.60e+06
2.5	2.5	0.05	2.559	2.464	0.047	2.321	2.144	4.29e+05	4.99e+05	3.55e+05	4.25e+05	-7.02e+04
2	1.5	0.05	1.969	1.499	0.049	1.409	0.64	2.99e+05	1.59e+06	2.26e+05	1.52e+06	-1.29e+06
1.2	0.7	0.05	1.335	0.628	0.069	1.471	0.598	3.28e+05	6.17e+05	2.54e+05	5.44e+05	-2.89e+05

Table 1 Model selection comparison results of Model_{AL} to Model_{DO}. Results are summarized using ABC-estimated parameter values for both models, the BIC score and SSE for each model, and the difference of BIC scores. This is done for the ten ground truth samples chosen in (d) and (e) of Figure 2 corresponding to the selection of C, L, and W values in the first three columns.

We showed that the TOPAZ algorithm can be used as a quantitative tool for model selection and parameter estimation utilizing topological summaries of spatiotemporal single-cell data.

We applied this flexible framework to a model of biological swarming, but it is more broadly applicable for many modeling approaches in molecular and cellular biology [22]. We designed our computational pipeline with future applications in mind, particularly spatial single-cell datasets that

incorporate transcriptomic and proteomic measurements. In particular, intracellular signaling pathways regulate dynamics of the actin cytoskeleton, thereby breaking symmetry of intracellular forces to drive both individual- and population-level cell behavior [23–25]. For *in vivo* settings, dynamic changes to the tissue microenvironment might also need to be considered [26–28]. In these scenarios, our model selection framework can be readily applied by incorporating microenvironment variables into an ABM, similar to other recently developed multiscale models [29, 30]. Since the TOPAZ framework operates on arbitrary multidimensional simulation outputs, it can be used to select among models that track both spatial trajectories and intracellular signaling states [31]. Such an analysis could reveal the relationship between biochemical pathway activation and collective cellular migration [32, 33].

We incorporated many simple algorithms into our first pass of the TOPAZ algorithm to demonstrate its utility. For example, we summarized the time-varying topological signature of ABM simulation using Crocker plots and performed the rejection algorithm for our ABC computations. We plan to implement other methods in future work to enhance the TOPAZ algorithm. Vineyards and the Crocker stack have been introduced as stable approaches to topologically represent time-varying data [34, 35]. For ABC parameter inference, we successfully used the ABC algorithm to estimate 3 parameter values; this approach may struggle to estimate more parameter values. In this scenario, we may turn to more data-efficient algorithms, including Sequential Monte Carlo methods [36].

Topological data analysis, and persistent homology in particular, provides robust, multiscale summaries that capture spatial organization, gradients, and connectivity. The ABMs we investigate here demonstrate that such topological summaries offer an effective interface between complex single-cell data and mechanistic models across biological scales. While the current study is limited to cell tracking data and a minimal ABM, TOPAZ is a generalizable model selection tool for ABMs that integrates TDA, thereby enabling the extraction of mechanistic insight from the type of noisy high-dimensional measurements that are characteristic of spatiotemporal single-cell datasets.

7 Methods

7.1 Topological data analysis

Homology is a fundamental concept in algebraic topology, quantifying the topological structure of data by identifying the number of n-dimensional holes within a given space. Persistent homology (PH) extends this notion by capturing topological features across multiple scales, providing a multi-scale summary of the shape of data. In particular, PH can be used to quantify changes in the topological structure of biological data, such as the collective motion of cells over time, by calculating Betti numbers. For a comprehensive introduction to PH and its application to biological systems, refer to [10].

To investigate the topological structure of cell motion, we compute the persistent homology of point clouds representing the time-varying locations and orientations of a population of cells. Each frame of the simulation is represented as a point cloud, denoted $X_f = \{(x_i, y_i, \theta_i)\}_{i=1}^{N_f}$, where N_f is the number of cells in frame f. In this context, each point represents a 0-simplex, with coordinates x_i, y_i, and θ_i corresponding to the spatial location and orientation of the i-th cell at time f.

To analyze the topology of these point clouds, we use the Vietoris-Rips filtration, which generates simplicial complexes from point clouds by progressively adding higher-dimensional simplices (edges, triangles, etc.) based on a proximity parameter ε. The Vietoris-Rips complex $V_R^{\varepsilon}(X_f)$ is constructed by connecting any set of $k+1$ points within distance ε of each other. The distance between two points is computed as the Euclidean distance in the x and y dimensions, plus the angular distance in the θ dimension, where angular distances are calculated as the shortest path on a circle, i.e., $\theta_1 - \theta_2 = \min(\theta_1, \theta_2) - \max(\theta_1, \theta_2) + 360$.

By taking an increasing sequence of proximity parameters $T = (\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_m)$, we generate a nested family of simplicial complexes, known as a filtration. For each ε, we compute the k-th Betti

number of the associated complex $V_R^\varepsilon(X_f)$, which counts the number of k-dimensional holes (or topological features) in the complex. Specifically, the k-th Betti number is defined as the rank of the k-th homology group, which quantifies the number of k-dimensional holes enclosed by $V_R^\varepsilon(X_f)$ [37].

To summarize the persistent homology of dynamic point clouds, we compute Betti curves for each time step and then concatenate them into a single matrix, which summarizes the homology of the entire time series. This matrix is known as a *Crocker plot* [38]. Crocker plots are a powerful tool for visualizing the evolution of Betti numbers over time by plotting the level sets or contours of concatenated Betti curves. These plots have been used to study the topology of biological systems, such as cell trajectories, by encoding topological changes at multiple spatial scales [10].

For all persistent homology computations, we used the Ripser package, a highly optimized C++ library for computing persistent homology [39]. Ripser provides state-of-the-art performance by reducing both memory consumption and computation time. We used the Python port of Ripser, available through the Scikit-TDA library, to implement all PH calculations in this analysis [40].

7.2 Approximate Bayesian computation

In contrast to frequentist methods, which provide point estimates for model parameters, Bayesian inference estimates a probability distribution over parameters conditioned on observed data. Specifically, the goal is to estimate the posterior distribution $p(\boldsymbol{q}|\mathbf{Y}^o)$, where $\boldsymbol{q}$ is the vector of parameters and $\mathbf{Y}^o$ represents the observed data. According to Bayes' theorem, this posterior is given by:

$$p(\boldsymbol{q}|\mathbf{Y}^o) = \frac{p(\mathbf{Y}^o|\boldsymbol{q})\, p(\boldsymbol{q})}{p(\mathbf{Y}^o)}, \tag{1}$$

where $p(\mathbf{Y}^o|\boldsymbol{q})$ is the likelihood function, $p(\boldsymbol{q})$ is the prior distribution, and $p(\mathbf{Y}^o)$ is the marginal likelihood:

$$p(\mathbf{Y}^o) = \int p(\mathbf{Y}^o|\boldsymbol{q})\, p(\boldsymbol{q})\, d\boldsymbol{q}. \tag{2}$$

While the marginal likelihood plays a crucial role in model comparison, it is often computationally intractable due to the high-dimensional integration involved. Fortunately, since it does not depend on the parameters $\boldsymbol{q}$, it can be treated as a constant during parameter estimation and safely omitted when comparing posterior densities across parameter values. However, computing the marginal likelihood is still necessary for formal model selection tasks.

A major challenge in Bayesian inference arises when the likelihood function $p(\mathbf{Y}^o|\boldsymbol{q})$ cannot be evaluated directly. This situation is common in complex or stochastic models [13, 41, 42]. In such cases, Approximate Bayesian Computation (ABC) provides a powerful alternative. ABC was originally introduced in Tavaré et al. as a likelihood-free, rejection-based method [43]. Since then, ABC has been extended and refined in various studies [41, 44, 45]. Comprehensive overviews of ABC methods are available in [46, 47]. In this study, we use the standard rejection-based ABC algorithm, also known as the prototype rejection-ABC algorithm [46], which we refer to simply as the ABC algorithm.

ABC approximates the likelihood by generating simulated datasets $\hat{\mathbf{Y}}$ from the model using parameters drawn from the prior distribution. These simulated datasets are then compared to the observed data $\mathbf{Y}^o$ using a distance function $\rho(\mathbf{Y}^o, \hat{\mathbf{Y}})$. Parameter samples for which the simulated output is sufficiently close to the observed data (i.e., within a specified tolerance τ) are accepted. As described in [10], the posterior can be approximated as:

$$p(\boldsymbol{q}|\mathbf{Y}^o) = \int p(\boldsymbol{q}, \hat{\mathbf{Y}}|\mathbf{Y}^o)\, d\hat{\mathbf{Y}}, \tag{3}$$

with the joint distribution given by:

$$p(\boldsymbol{q}, \hat{\mathbf{Y}}|\mathbf{Y}^o) \approx \frac{1\left[\rho(\mathbf{Y}^o, \hat{\mathbf{Y}}) < \tau\right]\, p(\hat{\mathbf{Y}}|\boldsymbol{q})\, p(\boldsymbol{q})}{p(\mathbf{Y}^o)}. \tag{4}$$

In our approach, we use Crocker matrices as summary statistics to characterize model behavior. Specifically, the observed data $\mathbf{Y}^o$ is replaced by the ground-truth Crocker matrix $\mathbf{C}^o_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon})$, and the simulated data $\hat{\mathbf{Y}}$ is represented by $\hat{\mathbf{C}}_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon}; \boldsymbol{q})$, which is generated using candidate parameter values. To quantify similarity, we employ the sum of squared errors (SSE) as our distance metric. A sample $\boldsymbol{q}$ is accepted if it satisfies:

$$SSE\left(\hat{\mathbf{C}}_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon}; \boldsymbol{q}); \mathbf{C}^o_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon})\right) \leq \tau. \tag{5}$$

We assume that only the upper and lower bounds of the parameters C and L are known, and we use a joint uniform prior over the interval $[0.1, 3.0]$. A total of $N_s = 10{,}000$ samples are drawn from this prior.

Choosing the tolerance τ is crucial for balancing accuracy and acceptance rate. However, in this work, we choose to accept only the top 1% samples with best SSE value. The accepted samples form the so-called ABC-posterior density. To compare the ground-truth and the estimated, we compute the median for each parameter from the accepted samples.

7.3 BIC

The Bayesian Information Criterion (BIC), also known as the Schwarz Information Criterion (SIC), is a widely used criterion for model selection and comparison, particularly in the context of statistical modeling and machine learning. It is employed to assess the trade-off between model complexity and goodness of fit, providing a way to avoid overfitting by penalizing the inclusion of additional parameters in the model.

The BIC is defined as:

$$\text{BIC} = -2\ln(\mathcal{L}) + k\ln(n)$$

where $\mathcal{L}$ is the likelihood of the model given the data, k is the number of parameters in the model, and n is the number of data points. The $\ln(\mathcal{L})$ is defined as:

$$\ln(\mathcal{L}) = -(n/2)\ln(2\pi) - (n/2)\ln(\sigma^2) - SSE/(2\sigma^2)$$

where $\sigma^2 = 1$.

The penalty term $k\ln(n)$ discourages overly complex models by penalizing models with more parameters. The ABC rejection procedure and BIC computation are summarized in Algorithm 1.

Algorithm 1 Approximate Bayesian computation rejection and Bayesian information criterion algorithm

Compute ground-truth Crocker matrix $\mathbf{C}^o_{0,1}$ from the observed data $\mathbf{Y}^o$
Choose the number of samples N_s
for *i=1,2, . . . ,* N_s **do**
  Sample $\boldsymbol{q}$ from the prior $p(\boldsymbol{q})$
  Numerically solve the D'Orsogna model for $\hat{\mathbf{Y}}(t; \boldsymbol{q})$
  Compute the estimated Crocker matrix $\hat{\mathbf{C}}_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon}; \boldsymbol{q})$
  Compute the distance $SSE\left(\boldsymbol{q}; \mathbf{C}^o_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon})\right)$
end
Accept samples where $SSE < SSE_{1\%}$
Compute the log-likelihood $\ln(\mathcal{L})$
Compute the BIC score

8 Funding

This work was supported by a Comparative Medicine Institute Ideation Award and the Data Science and AI Academy Seed Grants program from NC State University (to K.B.F. and J.M.H.); the National Science Foundation under the UQ4Life Research Training Group (DMS-2342344, to K.B.F.); and the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (R01-GM141691, to J.M.H.).

9 Data and Code Availability

All code used to implement the TOPAZ framework, including simulation scripts, topological data analysis modules, and example notebooks, is publicly available at:
https://github.com/biomathlab/TOPAZ

References

[1] Meijering, E., Dzyubachyk, O., Smal, I., Cappellen, W.A.: Tracking in cell and developmental biology. Seminars in cell & developmental biology **20**(8), 894–902 (2012)

[2] Ulman, V., Maška, M., Magnusson, K.E.G., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., *et al.*: An objective comparison of cell-tracking algorithms. Nature methods **14**(12), 1141–1152 (2017)

[3] Mi, H., Sivagnanam, S., Ho, W.J., Zhang, S., Bergman, D., Deshpande, A., Baras, A.S., Jaffee, E.M., Coussens, L.M., Fertig, E.J., Popel, A.S.: Computational methods and biomarker discovery strategies for spatial proteomics: a review in immuno-oncology. Briefings in Bioinformatics **25**(5), 421 (2024)

[4] Rappel, W.-J., Edelstein-Keshet, L.: Mechanisms of collective cell migration in development, regeneration and cancer. Nature Reviews Molecular Cell Biology **18**(12), 743–757 (2017)

[5] Hickey, J.W., Agmon, E., Horowitz, N., Tan, T.-K., Lamore, M., Sunwoo, J.B., Covert, M.W., Nolan, G.P.: Integrating multiplexed imaging and multiscale modeling identifies tumor phenotype conversion as a critical component of therapeutic t cell efficacy. Cell Systems **15**(4), 322–3385 (2024)

[6] Hartsock, I., Park, E., Toppen, J., Bubenik, P., S., D.E., Kemp, M.L., Cruz, D.A.: Topological data analysis of pattern formation of human induced pluripotent stem cell colonies. Scientific Reports **15**(1), 1234–1245 (2025)

[7] Huynh, T., Cang, Z.: Topological and geometric analysis of cell states in single-cell transcriptomic data. Briefings in Bioinformatics **25**(3), 176 (2024)

[8] Sizemore, A.E., Phillips-Cremins, J.E., Giusti, C., Bassett, D.S.: Importance of the whole: topological data analysis for the network neuroscientist. Network Neuroscience **3**(3), 656–673 (2019)

[9] Ulmer, M., Ziegelmeier, L., Topaz, C.M.: A topological approach to selecting models of biological experiments. PLOS ONE **14**(3), 0213679 (2019)

[10] Nguyen, K.C., Jameson, C.D., Baldwin, S.A., Nardini, J.T., Smith, R.C., Haugh, J.M., Flores, K.B.: Quantifying collective motion patterns in mesenchymal cell populations using topological data analysis and agent-based modeling. Mathematical Biosciences **370**, 109158 (2024)

[11] Ermentrout, G.B., Edelstein-Keshet, L.: Cellular automata approaches to biological modeling. Journal of theoretical Biology **160**(1), 97–133 (1993)

[12] Ray, A., Lee, O., Win, Z., Edwards, R.M., Alford, P.W., Kim, D.-H., Provenzano, P.P.: Anisotropic forces from spatially constrained focal adhesions mediate contact guidance directed cell migration. Nature Communications **16**, 1029–1037 (2017)

[13] Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C.P., Stumpf, M.P.: A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation. Nature protocols **9**(2), 439–456 (2014)

[14] Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.: Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface **6**(31), 187–202 (2009)

[15] D'Orsogna, M.R., Chuang, Y.-L., Bertozzi, A.L., Chayes, L.S.: Self-propelled particles with soft-core interactions: patterns, stability, and collapse. Physical review letters **96**(10), 104302 (2006)

[16] Abercrombie, M.: Contact inhibition in tissue culture. In vitro **6**(2), 128–142 (1970)

[17] Davis, J.R., Luchici, A., Mosis, F., Thackery, J., Salazar, J.A., Mao, Y., Dunn, G.A., Betz, T., Miodownik, M., Stramer, B.M.: Inter-cellular forces orchestrate contact inhibition of locomotion. Cell **161**(2), 361–373 (2015)

[18] Natan, S., Koren, Y., Shelah, O., Goren, S., Lesman, A.: Long-range mechanical coupling of cells in 3d fibrin gels. Molecular biology of the cell **31**(14), 1474–1485 (2020)

[19] Elsdale, T., Bard, J.: Cellular interactions in mass cultures of human diploid fibroblasts. Nature **236**(5343), 152–155 (1972)

[20] Edelstein-Keshet, L., Ermentrout, G.B.: Contact response of cells can mediate morphogenetic pattern formation. Differentiation **45**(3), 147–159 (1990)

[21] Warne, D.J., Baker, R.E., Simpson, M.J.: Using Experimental Data and Information Criteria to Guide Model Selection for Reaction–Diffusion Problems in Mathematical Biology. Bulletin of Mathematical Biology **81**(6), 1760–1804 (2019)

[22] Metzcar, J., Jutzeler, C.R., Macklin, P., Köhn-Luque, A., Brüningk, S.C.: A review of mechanistic learning in mathematical oncology. Frontiers in Immunology **15**, 1363144 (2024)

[23] Roycroft, A., Mayor, R.: Molecular basis of contact inhibition of locomotion. Cellular and Molecular Life Sciences **73**(6), 1119–1130 (2016)

[24] Schaks, M., Giannone, G., Rottner, K.: Actin dynamics in cell migration. Essays in Biochemistry **63**(5), 483–495 (2019)

[25] SenGupta, S., Parent, C.A., Bear, J.E.: The principles of directed cell migration. Nature Reviews Molecular Cell Biology **22**(8), 529–547 (2021)

[26] Baldwin, S.A., Haugh, J.M.: Semi-autonomous wound invasion via matrix-deposited, haptotactic cues. Journal of Theoretical Biology **568**, 111506 (2023)

[27] Bubna-Litic, M., Mayor, R.: Beyond mechanosensing: How cells sense and shape their physical environment during development. Current Opinion in Cell Biology **94**, 102514 (2025)

[28] Insall, R.H., Paschke, P., Tweedy, L.: Steering yourself by the bootstraps: how cells create their own gradients for chemotaxis. Trends in Cell Biology **32**(7), 585–596 (2022)

[29] Metzcar, J., Duggan, B.S., Fischer, B., Murphy, M., Heiland, R., Macklin, P.: A simple framework for agent-based modeling with extracellular matrix. Bulletin of Mathematical Biology (2025)

[30] Borau, C., Chisholm, R., Richmond, P., Walker, D.: An agent-based model for cell microenvironment simulation using flamegpu2. Computers in Biology and Medicine **172**, 107916 (2024)

[31] Crossley, R.M., Maini, P.K., Baker, R.E.: Modelling the impact of phenotypic heterogeneity on cell migration: a continuum framework derived from individual-based principles. arXiv. Version Number: 1 (2025)

[32] Bui, J., Conway, D.E., Heise, R.L., Weinberg, S.H.: Mechanochemical Coupling and Junctional Forces during Collective Cell Migration. Biophysical Journal **117**(1), 170–183 (2019)

[33] Nardini, J., Bortz, D.: Investigation of a Structured Fisher's Equation with Applications in Biochemistry. SIAM Journal on Applied Mathematics **78**(3), 1712–1736 (2018)

[34] Li, Y., Wang, D., Ascoli, G.A., Mitra, P., Wang, Y.: Metrics for comparing neuronal tree shapes based on persistent homology. PLOS ONE **12**(8), 0182184 (2017)

[35] Xian, L., Adams, H., Topaz, C.M., Ziegelmeier, L.: Capturing dynamics of time-varying data via topology. Foundations of Data Science **4**(1), 1 (2022)

[36] Sisson, S.A., Fan, Y., Tanaka, M.M.: Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences **104**(6), 1760–1765 (2007)

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

[38] Crocker, J.C., Grier, D.G.: Methods of digital video microscopy for colloidal studies. Journal of Colloid and Interface Science **179**(1), 298–310 (2000)

[39] Bauer, U.: Ripser: efficient computation of vietoris–rips persistence barcodes. Journal of Applied and Computational Topology **5**(1), 391–423 (2021)

[40] Scikit-TDA: Scikit-TDA: Persistent homology tools in Python. https://scikit-tda.github.io/ (2024)

[41] Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate Bayesian Computation in Population Genetics. Genetics **162**(4), 2025–2035 (2002)

[42] Thorne, T., Stumpf, M.P.H.: Graph spectral analysis of protein interaction network evolution. Journal of The Royal Society Interface **9**(75), 2653–2666 (2012)

[43] Tavaré, S., Balding, D.J., Griffiths, R.C., Donnelly, P.: Inferring Coalescence Times From DNA Sequence Data. Genetics **145**(2), 505–518 (1997)

[44] Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S.: Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences **100**(26), 15324–15328 (2003)

[45] Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W.: Population growth of human

Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution **16**(12), 1791–1798 (1999)

[46] Beaumont, M.A.: Approximate Bayesian Computation in Evolution and Ecology. Annual Review of Ecology, Evolution, and Systematics **41**(1), 379–406 (2010)

[47] Marin, J.-M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate Bayesian computational methods. Statistics and Computing **22**(6), 1167–1180 (2012)