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], John T. Nardini[3], Jason M. Haugh[4], Kevin B. Flores[1,2*]

1 Department of Mathematics, North Carolina State University, Raleigh, NC, USA
2 Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, USA
3 Department of Mathematics and Statistics, The College of New Jersey, Ewing, NJ, USA
4 Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC, USA

* 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), approximate approximate Bayesian computation (AABC), 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 AABC 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 extensible framework for mechanistic inference and model discrimination in spatial single-cell analysis.

Author summary

Understanding how individual cells coordinate to produce complex collective behaviors is a major challenge in computational biology, especially with the increasing availability of high-resolution, spatiotemporal single-cell data. While agent-based models (ABMs) offer a flexible framework for simulating cell behaviors and interactions, they are often difficult to calibrate and compare. Topological data analysis (TDA), on the other hand, captures spatial organization in a robust and scale-invariant way but lacks mechanistic interpretability. In this work, we present TOPAZ (TOpologically-based Parameter

inference for Agent-based model optimiZation), a novel computational pipeline that integrates TDA with approximate Bayesian computation, approximate approximate Bayesian computation, and Bayesian model selection to infer biologically meaningful parameters and identify the most plausible ABM from spatiotemporal cellular data. We benchmark TOPAZ using synthetic data from ABMs of collective cell movement in dense fibroblast populations. Our results show that TOPAZ can distinguish between competing mechanistic hypotheses, namely the presence or absence of alignment interactions among neighboring cells. This approach provides a powerful and extensible framework for model inference and selection with the potential to enable deeper insights into the mechanisms driving complex emergent behaviors in cell populations.

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]. Agent-based models (ABMs) are computational experiments that model individual agents and how they interact with other agents and the environment [4, 5]. They 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 [6]. Topological data analysis (TDA) is another computational framework that has been utilized in bioinformatics studies for analyzing single-cell omics data. It uses concepts from topology to describe the shape of data and can be used to detect changes between distinct simulations that vary over time and space [7]. 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 [8, 9]. 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, 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 [10, 11]. 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 addressed 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 [12]. 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 [13, 14]. Moreover, while Nguyen et al. exemplified that parameter inference with TDA was possible, the biologically-motivated mechanisms 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 [15, 16]. We demonstrate how integrating model selection into the TDA-ABC framework yields a novel and extensible approach for analyzing complex collective cell behaviors in biomedical contexts and how adding a directional alignment mechanism may lead to more biologically-accurate results.

Agent-based models

In this work, we focus on 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 [12, 17]. We use the D'Orsogna model, referred to here as Model_{DO}, as an example of an 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 [18–20]. A full description of the Model_{DO} ABM can be found in S1 Appendix.

To assess whether distinct mechanisms of interaction can be reliably distinguished using summary statistics derived from TDA, we compare two ABMs: the baseline D'Orsogna model (Model_{DO}) and an extended model with alignment interactions (Model_{AL}). We created Model_{AL} to determine if the addition of alignment into Model_{DO} leads to the emergence of fluidization, as seen in experiments of fibroblast migration in [12]. We simulate synthetic trajectory data from each model across a range of parameter settings and then fit the alternative model to that data using ABC and AABC. By applying model selection via the Bayesian information criterion (BIC), we assess whether the inference pipeline can correctly recover the true generative model. This simulation-based framework serves as a proof-of-concept for using TDA-informed ABC and AABC to distinguish between competing mechanistic hypotheses—a critical step toward future applications involving real spatial single-cell datasets that incorporate omics-level molecular information. For example, Johnson et al. incorporate spatial transcriptomic sequencing of resected pancreatic lesions to initialize agent-based models, noting that these high-resolution omics datasets may also enable systematic model selection among competing mechanistic hypotheses [21].

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 [12].

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 S1 Appendix for a derivation from Model_{DO}). This extension is biologically motivated by experimental observations of contact-mediated alignment [22] and theoretical models of directional organization in cell populations [23]. In our extended model (Model_{AL}), agents adjust their velocity angle 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 (C, L, 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.

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

We developed a computational pipeline named "TOPAZ" that combines TDA computed from ABM simulations, ABC , approximate approximate Bayesian computation (AABC), and 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 "Contour Realization Of Computed K-dimensional hole Evolution in the Rips complex" (Crocker) [24] diagrams are computed using the Vietoris-Rips filtration 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 and posterior density generation as in [12]. **Step 2.5:** Use of AABC to generate more samples for parameter inference if needed. **Step 3:** Assess whether posterior predictive summary statistic distributions are statistically distinguishable across candidate models using Permutational Multivariate Analysis of Variance (PERMANOVA), Energy Distance, and/or Maximum Mean Discrepancy (MMD) prior to BIC-based model selection. **Step 4:** The posterior distribution from the top 1% of ABC samples, the number of parameters, and the number of data points are used to calculate the BIC score for the ABM. **Step 5:** The ABM with the lowest BIC score is selected as the optimal model. We use BIC instead of other information criterion, such as Akaike information criterion, due to the log-likelihood of the sum of squared errors being so large and wanting to emphasize the penalty for the number of parameters.

Dimensionality reduction for visualization

To aid in addressing the inverse problem's identifiability, t-SNE is employed as a nonlinear dimensional reduction method, which is particularly effective when 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. For every simulation, we computed a Crocker matrix with size 100 time points $\times$ 200 proximity parameters $\times$ 2 Betti numbers (zero and one). We applied t-SNE to reduce the 9,900 Crocker plots (example shown in (a) of Fig 2) with size $100 \times 200 \times 2$ matrix to a 3×1 vector. 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 30×30 2D t-SNE plots, one for each W value

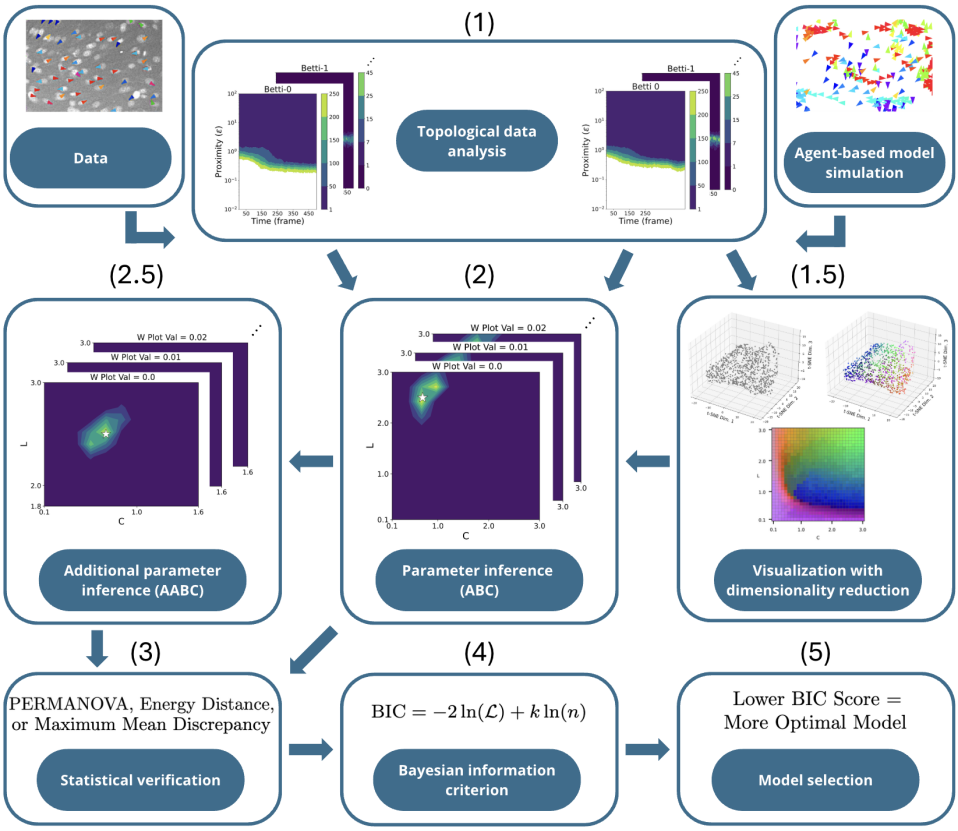

Fig 1. A computational pipeline for agent-based model selection. Our pipeline called TOPAZ has the following steps: (1) From data or an ABM simulation, topological data analysis in the form of Crocker plots. (1.5) (Optional) Visualization with dimensionality reduction techniques such as t-SNE or principal component analysis. (2) Parameter inference techniques such as ABC. (2.5) (Optional) Generation of additional samples using AABC for parameter inference if needed. (3) Assessment of distributional differences in posterior predictive summary statistics across models using PERMANOVA, Energy Distance, and/or MMD.
(4) Bayesian information criterion score calculation. (5) Model selection based on lowest BIC score.

(examples shown in (d) and (e) of Fig 2; all eleven plots can be found in S2 Fig). For more visualization using t-SNE, see S3 Appendix and S1 Fig.

Results

Simulation Study Design and Hypotheses

We developed a computational pipeline TOPAZ for selecting which 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 also fits of the ABM Model_{DO} to data simulated by the ABM Model_{AL}. Our hypothesis is that by analyzing the outputs of our TOPAZ framework, including t-SNE, *SSE*, BIC, nonparametric statistical tests (PERMANOVA, Energy Distance, and MMD), and posterior density plots, that the user will be able to

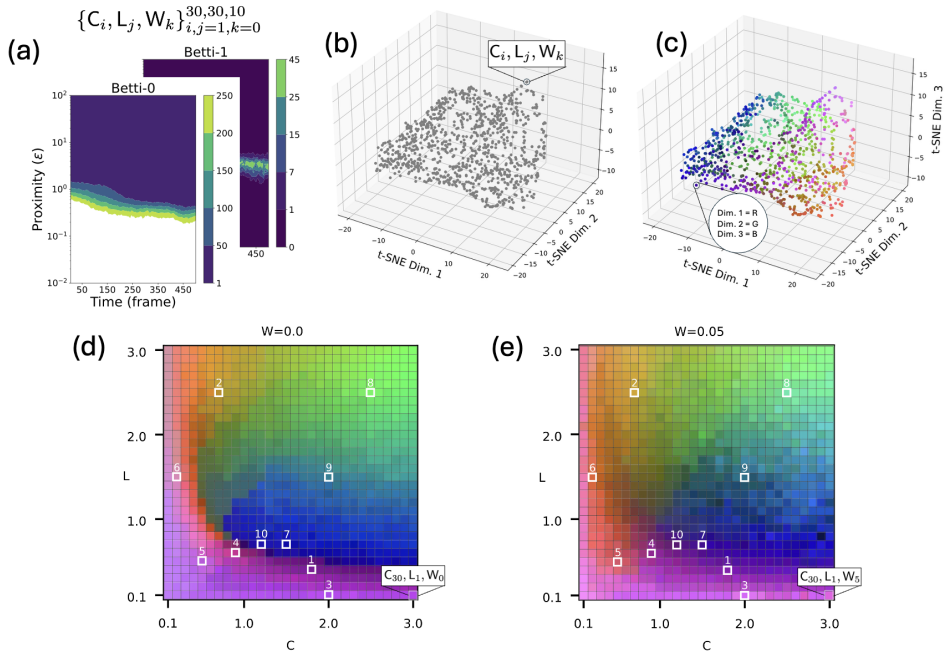

Fig 2. Visualization with dimensionality reduction using t-SNE. (a) Crocker plots of the ABM simulation. (b) Size $100 \times 200 \times 2$ matrix reduced to a 3×1 vector. Only 10% of the 9,900 simulations are shown. (c) Each point is colored based on their 3D position where (Dim. 1, Dim. 2, Dim. 3) = (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).

conclude when Model_{AL} can be selected as the better model. By leveraging BIC in addition to 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).

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 by Nguyen et al. showing a wide representation of topologically distinct spatiotemporal features in simulations of the Model_{DO} model [12]. The same values of C and L were used for both Model_{DO} and Model_{AL} to isolate the effects of the added alignment framework. If we randomized all C, L, and W values, it would be harder to compare the effects of the added alignment parameter. Thus, by keeping C and L the same between models and only changing W, we can see the effect that our new alignment parameter has on the system. 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 S1 Video-S4 Video for the corresponding ABM simulation movies.

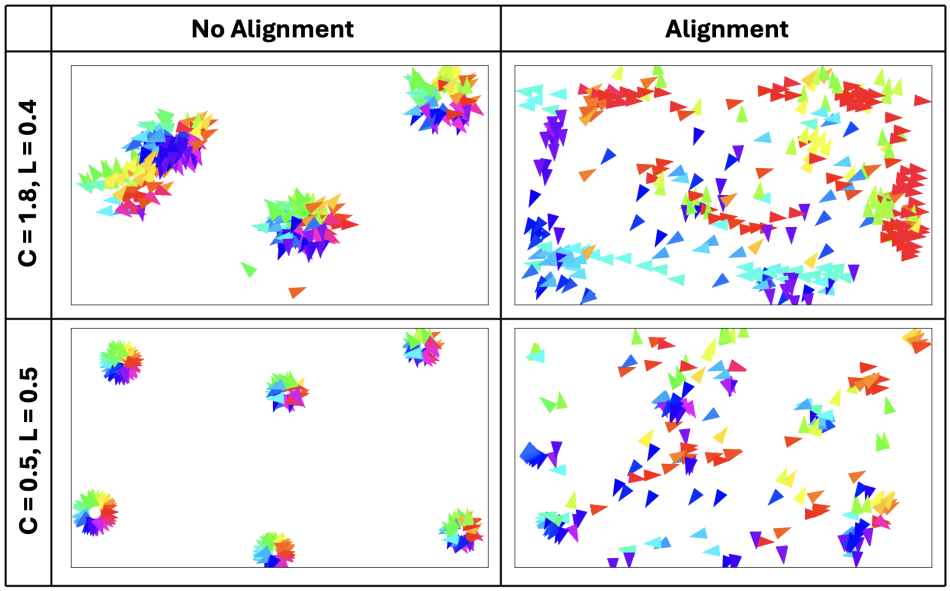

Fig 3. Visualization of Model_{AL} output for two different C and L combinations. Simulation snapshots are shown without alignment case (W=0.0, left) and with alignment case (W=0.05, right). The arrows and colors represent the direction the cells are moving. These snapshots are at the end of the simulation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

Parameter Inference Using ABC and AABC

Using simulations generated from Model_{DO} or Model_{AL}, we computed TDA summaries from the generated point clouds and then used them for parameter inference within ABC and AABC (steps 1, 2, and 2.5 in Fig 1). Visualizing the AABC 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 S3 Fig. Concordantly, the Crocker plots at the ground-truth and estimated parameter values (median of posterior) exemplify that the corresponding topological diagrams are visually similar (Fig 5).

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). 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 except two cases (see S4 Appendix). These results confirm that our TOPAZ pipeline can be used to quantitatively compare ABMs and select an optimal model for spatiotemporal cellular data.

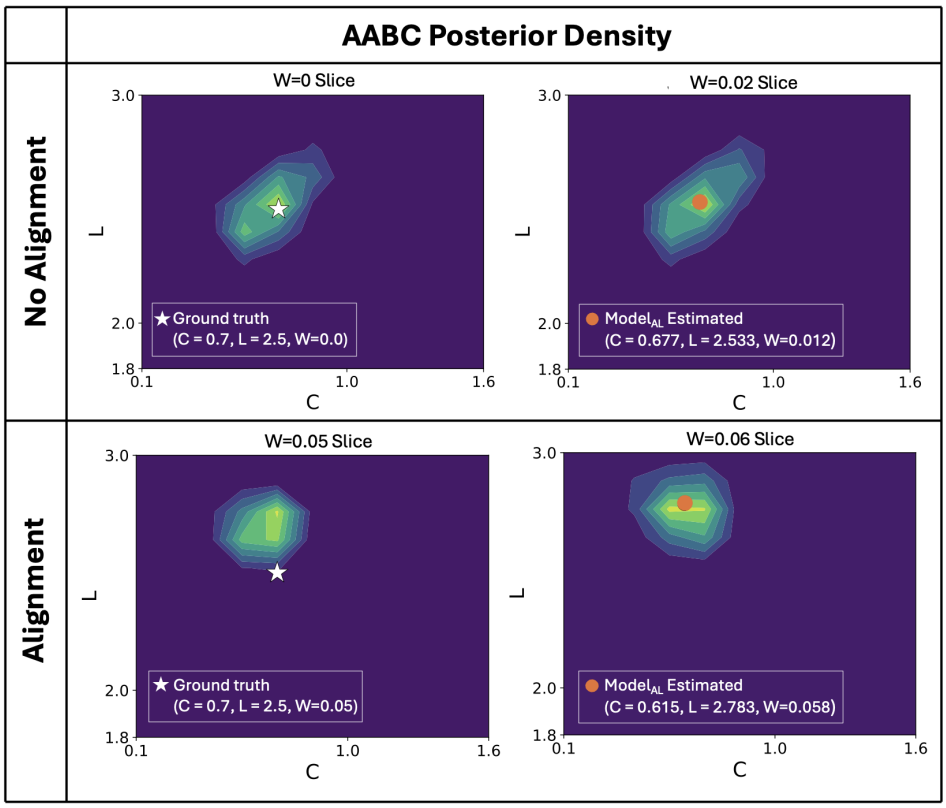

Fig 4. Examples of AABC posterior density plots for Model$_{\mathbf{AL}}$. The ground truth (white stars) and sample median AABC-estimated parameter values for Model$_{\mathrm{AL}}$ (orange dots) are displayed as well as the AABC posteriors for Model$_{\mathrm{AL}}$. Left and right are different W-slices and top and bottom use data generated with no alignment and alignment, respectively. For the sample median AABC-estimated parameter values for Model$_{\mathrm{DO}}$, see Table 1. The complete selection of AABC posterior density plots for Model$_{\mathrm{AL}}$ can be found in S3 Fig.

Discussion

We introduced TOPAZ, a novel methodology to facilitate model selection for ABMs. This algorithm integrates concepts from TDA, ABC, AABC, 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 and AABC to infer the posterior distribution of each considered model's parameter values. The 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 [25].

We developed an extension of the D'Orsogna Model that includes an agent alignment force. The goal was to incorporate a new biophysical mechanism that is biologically grounded to create a more realistic model. This new 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 parallel or antiparallel to adjacent groups. This recapitulates experimental observations of confluent fibroblast populations, where cells self-organize into dynamic,

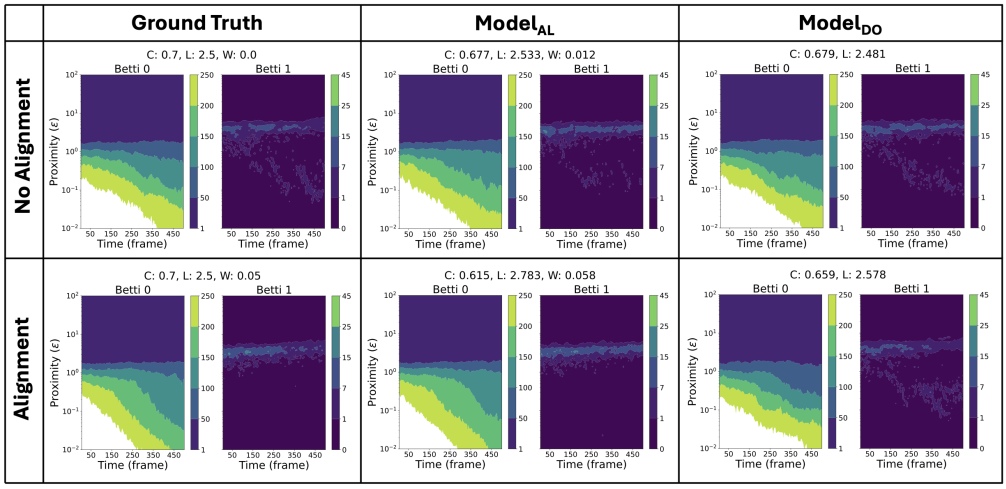

Fig 5. Crocker plots summarizing the topology of the ground truth and AABC simulations. For the ground truth values and AABC-estimated values for Model_{AL} and Model_{DO}, we constructed Crocker plots for Betti-0 and Betti-1 (left, middle, and right, respectively) for the case where there is no alignment (W=0.0, top) and alignment (W=0.05, bottom). The color legend represents the contour levels corresponding to the number of simplices of degree 0 and 1 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}	$\text{Model}_{\text{AL}} - \text{Model}_{\text{DO}}$
1.8	0.4	0	2.270	0.302	0.009	2.044	0.361	1.12e+06	1.70e+06	1.05e+06	1.63e+06	-5.84e+05
0.7	2.5	0	0.677	2.533	0.012	0.679	2.481	1.66e+06	8.54e+05	1.59e+06	7.80e+05	8.09e+05
2.0	0.1	0	1.135	0.151	0.009	1.215	0.133	2.22e+06	8.79e+05	2.15e+06	8.05e+05	1.35e+06
0.9	0.6	0	0.950	0.205	0.035	0.926	0.584	3.92e+06	6.38e+05	3.85e+06	5.64e+05	3.28e+06
0.5	0.5	0	0.646	0.186	0.010	0.520	0.471	1.92e+07	8.02e+06	1.91e+07	7.95e+06	1.11e+07
0.2	1.5	0	0.181	0.941	0.014	0.185	1.510	7.34e+06	1.37e+06	7.27e+06	1.30e+06	5.97e+06
1.5	0.7	0	2.420	0.631	0.007	1.606	0.655	1.21e+06	3.46e+05	1.13e+06	2.72e+05	8.60e+05
2.5	2.5	0	2.572	2.642	0.009	2.506	2.530	3.72e+05	3.51e+05	2.98e+05	2.78e+05	2.06e+04
2.0	1.5	0	2.371	1.481	0.006	2.113	1.487	3.50e+05	3.74e+05	2.76e+05	3.00e+05	-2.41e+04
1.2	0.7	0	2.209	0.656	0.007	1.236	0.703	2.25e+06	3.81e+05	2.18e+06	3.07e+05	1.87e+06
1.8	0.4	0.05	1.709	0.381	0.056	1.593	0.524	6.48e+05	2.12e+07	5.74e+05	2.12e+07	-2.06e+07
0.7	2.5	0.05	0.615	2.783	0.058	0.659	2.578	9.56e+05	7.74e+06	8.82e+05	7.67e+06	-6.78e+06
2.0	0.1	0.05	0.720	0.165	0.045	0.612	0.770	1.35e+06	6.20e+07	1.28e+06	6.19e+07	-6.06e+07
0.9	0.6	0.05	1.010	0.483	0.066	1.309	0.567	6.29e+05	1.84e+07	5.55e+05	1.83e+07	-1.77e+07
0.5	0.5	0.05	0.449	0.469	0.065	0.441	1.699	1.45e+06	1.82e+06	1.37e+06	1.74e+06	-3.68e+05
0.2	1.5	0.05	0.167	1.606	0.067	0.538	2.398	6.78e+05	5.81e+06	6.05e+05	5.74e+06	-5.13e+06
1.5	0.7	0.05	1.572	0.702	0.062	1.496	0.592	4.03e+05	1.50e+06	3.29e+05	1.43e+06	-1.10e+06
2.5	2.5	0.05	2.518	2.477	0.046	2.432	2.203	4.27e+05	4.40e+05	3.54e+05	3.67e+05	-1.29e+04
2.0	1.5	0.05	2.080	1.501	0.052	1.194	0.673	3.20e+05	3.10e+06	2.47e+05	3.03e+06	-2.78e+06
1.2	0.7	0.05	1.238	0.674	0.067	1.201	0.644	4.16e+05	2.16e+06	3.43e+05	2.08e+06	-1.74e+06

Table 1. Model selection comparison results of Model_{AL} to Model_{DO}. Results are summarized using AABC-estimated parameter values for both models, the BIC score and sum of squared errors (*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.

parallel/anti-parallel streams [12]. We then considered a suite of ABM simulations that were simulated from different parameter values. 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. This helped answer the question of can we use quantitative or topological metrics to tell if our extended model is actually a better model.

While we applied this flexible framework to a model of biological swarming, but it is broadly applicable for many modeling approaches in molecular and cellular biology [26].

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 [27–29]. For *in vivo* settings, dynamic changes to the tissue microenvironment might also need to be considered [30–32]. 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 [33, 34]. 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 [35]. Such an analysis could reveal the relationship between biochemical pathway activation and collective cellular migration [36, 37].

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 and AABC 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 [38, 39]. We can also use the principal component analysis or t-distributed stochastic neighbor embedding dimensionality results in place of our full Crocker plots as a new summary statistic instead of only for visualization purposes. However, this is beyond the scope of this paper and will be investigated in future studies. For ABC and AABC parameter inference, we successfully used the ABC and AABC algorithms to estimate 3 parameter values. For future work, we will investigate if these approaches may be able to estimate more parameter values. In this scenario, we may turn to more data-efficient algorithms, including Sequential Monte Carlo methods [40]. Since the TOPAZ method has so far been validated only on simulated data, an important direction for future work is assessing its robustness in the presence of experimental noise, as demonstrated in related workflows such as Nguyen et al. [12].

Although many alternative model families could have been considered for model selection, in this initial study we focused on the D'Orsogna model used in Nguyen et al. [12] and a minimally extended version that incorporates an additional alignment term. Because the baseline D'Orsogna model is already known to reproduce key qualitative features of our experimental system, we viewed the alignment extension as a biologically motivated next step toward improving its ability to capture experimental observations. Our current work establishes this extension and evaluates the TOPAZ framework under controlled conditions using simulated data. Importantly, comparing two closely related models provides a stringent test of the model-selection component of TOPAZ. Distinguishing between models that differ only subtly in their mechanistic structure is inherently challenging; demonstrating that the framework can resolve these small differences increases confidence in its ability to discriminate among more disparate or non-nested models. In future work we will explore the application of the TOPAZ framework to a broader set of candidate models, including further extensions of the D'Orsogna family as well as models that incorporate qualitatively different interaction rules.

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 extensible 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.

Methods

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 [12].

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 $V_R^\varepsilon(X_f)$ [41].

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* [42]. 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 [12].

For all persistent homology computations, we used the Ripser package, a highly optimized C++ library for computing persistent homology [43]. 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 [44].

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 [15, 45, 46]. In such cases, ABC provides a powerful alternative. ABC was originally introduced in Tavaré et al. as a likelihood-free, rejection-based method [47]. Since then, ABC has been extended and refined in various studies [45, 48, 49]. Comprehensive overviews of ABC methods are available in [50, 51]. In this study, we use the standard rejection-based ABC algorithm, also known as the prototype rejection-ABC algorithm [50], 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 [12], 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 of Betti numbers 0 and 1 as summary statistics to characterize model behavior, as done in previous work [12]. 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.

Approximate approximate Bayesian computation

As an extension of ABC, we employ AABC to efficiently generate additional samples for parameter inference. AABC follows the same initial and final steps as ABC, beginning with a finite set of forward simulations and concluding with a rejection-based approximation to the posterior distribution. Its primary advantage lies in its ability to produce a substantially larger number of approximate samples by leveraging an initial set of simulated data, thereby reducing the computational burden associated with repeated forward simulations [52].

We first performed ABC using $N_s = 10{,}000$ simulations to obtain posterior density estimates for each model. To concentrate subsequent sampling in regions of non-negligible posterior support, we restricted the original $30 \times 30 \times 11$ parameter grid to a smaller grid containing all parameter combinations for which the posterior density was nonzero for either Model_{DO} or Model_{AL} in each row of Table 1. This restriction allows computational effort to be focused on regions of interest rather than distributed uniformly across the full parameter space.

Following this grid refinement, we applied the first stage of the AABC algorithm by generating an expanded set of original simulations within the reduced parameter space. The total number of simulations was scaled proportionally to the size of the reduced grid and then multiplied by a factor of ten relative to the original ABC sample size; for example, if the restricted grid occupied half of the original parameter space, we generated 50,000 simulations. These simulations served as the basis for constructing additional approximate samples without further evaluations of the underlying agent-based model.

New samples were generated using the AABC procedure described in Buzbas et al. [52]. For each parameter proposal, the $k = 5$ nearest neighbors were identified using Euclidean distance in summary statistic space, where each Crocker plot was treated as a single data point. Parameter values were then averaged using Epanechnikov kernel weights to produce approximate posterior samples. This process was repeated iteratively, with parameter estimates, error metrics, and BIC values recomputed after each batch of newly generated samples. The procedure was terminated upon convergence, defined as a change of less than 1% in the error values for each parameter combination in Table 1. Convergence behavior is illustrated in S4 Fig.

Permutational Multivariate Analysis of Variance

PERMANOVA is a nonparametric statistical test used to assess differences between groups based on multivariate data. Unlike traditional analysis of variance (ANOVA), which relies on assumptions of multivariate normality and homogeneity of variances, PERMANOVA operates directly on a distance or dissimilarity matrix, making it well suited for complex, high-dimensional, or non-Euclidean data.

PERMANOVA partitions the total variation in the distance matrix into within-group and between-group components and computes a pseudo-F statistic analogous to the classical ANOVA F-statistic. Given a distance matrix D, the pseudo-F statistic is defined as

$$F = \frac{SS_{\text{between}}/(g-1)}{SS_{\text{within}}/(n-g)},$$

where SS_{between} and SS_{within} are the sums of squares associated with between-group and within-group variation, respectively, g is the number of groups, and n is the total number of observations.

Statistical significance is assessed through a permutation procedure rather than reliance on a known parametric distribution. Group labels are randomly permuted a large number of times, and the pseudo-F statistic is recomputed for each permutation. The resulting empirical distribution is used to estimate a p-value as

$$p = \frac{1 + \sum_{i=1}^{N} \mathbb{I}(F_i \geq F_{\text{obs}})}{1 + N},$$

where F_{obs} is the observed pseudo-F statistic, F_i denotes the statistic from the i-th permutation, N is the total number of permutations, and $\mathbb{I}(\cdot)$ is the indicator function.

Because PERMANOVA is based on distances, its results depend on the choice of distance metric, and it is sensitive to differences in group dispersion. Nonetheless, it provides a flexible and robust framework for testing group-level differences in multivariate settings and is commonly used in ecological, biological, and other data-intensive applications. For similar statistical tests and results, please see S2 Appendix and S1 Table.

Bayesian information criterion

BIC, also known as the Schwarz information criterion, 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 and the AABC rejection procedure and BIC computation are summarized in Algorithm 2.

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;

Algorithm 2: Approximate 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 initial number of forward simulations N_s (ABC stage) and the number of AABC-generated samples per batch N_a;
Choose k (number of nearest neighbors) and an Epanechnikov kernel $K(\cdot)$;

Stage 1: ABC library construction;
for $i = 1, 2, \ldots, N_s$ **do**
  Sample $\boldsymbol{q}^{(i)} \sim p(\boldsymbol{q})$;
  Numerically solve the D'Orsogna model for $\hat{\mathbf{Y}}\big(t; \boldsymbol{q}^{(i)}\big)$;
  Compute summary statistic (Crocker matrix) $\hat{\mathbf{C}}^{(i)}_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon})$;
  Compute distance $\mathrm{SSE}^{(i)} = \mathrm{SSE}\big(\boldsymbol{q}^{(i)}; \mathbf{C}^{o}_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon})\big)$;
end
Set the ABC acceptance threshold $\mathrm{SSE}_{1\%}$ (e.g., the 1st percentile of $\{\mathrm{SSE}^{(i)}\}_{i=1}^{N_s}$);
Let $\mathcal{D} = \{(\boldsymbol{q}^{(i)}, \hat{\mathbf{C}}^{(i)}_{0,1}, \mathrm{SSE}^{(i)})\}_{i=1}^{N_s}$ denote the original simulation library;
(Optional) Restrict the parameter space to the reduced grid containing all $\boldsymbol{q}$ with nonzero posterior density under ABC;
(Optional) Augment $\mathcal{D}$ with additional forward simulations within the reduced grid;

Stage 2: AABC sample generation (no new forward simulations);
repeat
  for $j = 1, 2, \ldots, N_a$ **do**
    Propose $\tilde{\boldsymbol{q}}^{(j)} \sim p(\boldsymbol{q})$ (or within the reduced grid);
    Identify the k nearest neighbors of $\tilde{\boldsymbol{q}}^{(j)}$ in summary-statistic space using $\mathcal{D}$;
    Compute kernel weights $w_m \propto K\big(\frac{d_m}{h}\big)$ for neighbor distances d_m, normalized so $\sum_{m=1}^{k} w_m = 1$;
    Form an approximate sample (e.g., weighted average) $\boldsymbol{q}^{*(j)} = \sum_{m=1}^{k} w_m \, \boldsymbol{q}^{(m)}$;
    (Optional) Form an approximate summary statistic $\hat{\mathbf{C}}^{*(j)}_{0,1} = \sum_{m=1}^{k} w_m \, \hat{\mathbf{C}}^{(m)}_{0,1}$;
    Compute an approximate distance $\mathrm{SSE}^{*(j)} = \mathrm{SSE}\big(\boldsymbol{q}^{*(j)}; \mathbf{C}^{o}_{0,1}(\boldsymbol{t}, \boldsymbol{\varepsilon})\big)$;
  end
  Accept AABC samples where $\mathrm{SSE}^{*} < \mathrm{SSE}_{1\%}$;
  Update posterior approximation using accepted $\{\boldsymbol{q}^{*}\}$;
  Compute the log-likelihood $\ln(\mathcal{L})$ and BIC score;
until convergence (e.g., $< 1\%$ change in SSE);

Supporting information

S1 Appendix. D'Orsogna and Alignment models.

S2 Appendix. Statistical evaluation for model comparison.

S3 Appendix. Nearest neighbor t-SNE visualizations.

S4 Appendix. Case Analysis: When $\mathbf{Model_{AL}}$ outperforms $\mathbf{Model_{DO}}$.

S1 Table. $\mathbf{Model_{AL}}$ versus $\mathbf{Model_{DO}}$ Crocker plot distribution statistical difference results.

S1 Fig. T-SNE nearest-neighbor visualizations. T-SNE nearest-neighbor visualizations for the selected (C,L) pairs from Table 1 when W=0.0 and W=0.05. All points where W=0 are in red and similarly all points where W=0.05 are in blue. Each plot has one selected (C,L) pair at W=0.0 and its nearest neighbor where W=0.05 for the left two columns and the reverse on the right.

S2 Fig. Full selection of 2d t-sne extractions. The full selection of the Betti-0 and Betti-1 t-SNE colormaps, showcasing an example of no alignment (W=0.0, top left) and different levels of alignment (W=0.01-W=.01). The colors for each (C, L, W) combination are the colors that were generated in the 3D t-SNE plot ((c) of Figure 2) for the corresponding (C, L, W) point.

S3 Fig. Full selection of AABC posterior density plots for $\mathbf{Model_{AL}}$. Top and bottom use data generated with no alignment (W=0.0) and alignment (W=0.05), respectively. The white star represents the ground-truth values whereas the orange dots represent the sample median AABC-estimated values for $\mathrm{Model_{AL}}$.

S4 Fig. Convergence plots corresponding to . The full selection of convergence plots for each row of Table 1 for both $\mathrm{Model_{DO}}$ and $\mathrm{Model_{AL}}$ when the ground truth value for W is 0 and 0.05 for a total of four convergence plots per row. The AABC method was used repeatedly to generate more samples until the error had converged for each row of Table 1.

S1 Video. Full simulation gif with (C, L, W)=(1.8, 0.4, 0.0). Video simulation from t=20-120. he arrows and colors represent the direction the cells are moving. Since W=0, this represents a case without alignment.

S2 Video. Full simulation gif with (C, L, W)=(1.8, 0.4, 0.05). Video simulation from t=20-120. he arrows and colors represent the direction the cells are moving. Since W=0.05, this represents a case with alignment.

S3 Video. Full simulation gif with (C, L, W)=(0.5, 0.5, 0.0). Video simulation from t=20-120. he arrows and colors represent the direction the cells are moving. Since W=0, this represents a case without alignment.

S4 Video. Full simulation gif with (C, L, W)=(0.5, 0.5, 0.05). Video simulation from t=20-120. he arrows and colors represent the direction the cells are moving. Since W=0.05, this represents a case with alignment.

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.); the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (1U54AI191253-01, 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.).

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/alyswenz31/TOPAZ`

References

1. Meijering E, Dzyubachyk O, Smal I, van Cappellen WA. Tracking in cell and developmental biology. Seminars in cell & developmental biology. 2012;20(8):894–902.

2. Ulman V, Maška M, Magnusson KEG, Ronneberger O, Haubold C, Harder N, et al. An objective comparison of cell-tracking algorithms. Nature methods. 2017;14(12):1141–1152.

3. Mi H, Sivagnanam S, Ho WJ, Zhang S, Bergman D, Deshpande A, et al. Computational methods and biomarker discovery strategies for spatial proteomics: a review in immuno-oncology. Briefings in Bioinformatics. 2024;25(5):bbae421.

4. Montagud A, Ponce-de Leon M, Valencia A. Systems biology at the giga-scale: Large multiscale models of complex, heterogeneous multicellular systems. Current Opinion in Systems Biology. 2021;28:100385.

5. Pleyer J, Fleck C. Agent-based models in cellular systems. Frontiers in Physics. 2023;10-2022.

6. Hickey JW, Agmon E, Horowitz N, Tan TK, Lamore M, Sunwoo JB, et al. Integrating multiplexed imaging and multiscale modeling identifies tumor phenotype conversion as a critical component of therapeutic T cell efficacy. Cell Systems. 2024;15(4):322–338.e5.

7. Munch E. A User's Guide to Topological Data Analysis. Journal of Learning Analytics. 2017;4:47–61.

8. Hartsock I, Park E, Toppen J, Bubenik P, Dimitrova ES, Kemp ML, et al. Topological data analysis of pattern formation of human induced pluripotent stem cell colonies. Scientific Reports. 2025;15(1):1234–1245.

9. Huynh T, Cang Z. Topological and geometric analysis of cell states in single-cell transcriptomic data. Briefings in Bioinformatics. 2024;25(3):bbae176.

10. Sizemore AE, Phillips-Cremins JE, Ghrist R, Bassett DS. Importance of the whole: topological data analysis for the network neuroscientist. Network Neuroscience. 2019;3(3):656–673.

11. Ulmer M, Ziegelmeier L, Topaz CM. A topological approach to selecting models of biological experiments. PLoS ONE. 2019;14(3):e0213679.

12. Nguyen KC, Jameson CD, Baldwin SA, Nardini JT, Smith RC, Haugh JM, et al. Quantifying collective motion patterns in mesenchymal cell populations using topological data analysis and agent-based modeling. Mathematical Biosciences. 2024;370:109158.

13. Ermentrout GB, Edelstein-Keshet L. Cellular automata approaches to biological modeling. Journal of Theoretical Biology. 1993;160(1):97–133.

14. Ray A, Lee O, Win Z, Edwards RM, Alford PW, Kim DH, et al. Anisotropic forces from spatially constrained focal adhesions mediate contact guidance directed cell migration. Nature Communications. 2017;8:14923.

15. Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MPH. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. Nature protocols. 2014;9(2):439–456.

16. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface. 2009;6(31):187–202.

17. D'Orsogna MR, Chuang YL, Bertozzi AL, Chayes LS. Self-propelled particles with soft-core interactions: patterns, stability, and collapse. Physical review letters. 2006;96(10):104302.

18. Abercrombie M. Contact inhibition in tissue culture. In vitro. 1970;6(2):128–142.

19. Davis JR, Luchici A, Mosis F, Thackery J, Salazar JA, Mao Y, et al. Inter-cellular forces orchestrate contact inhibition of locomotion. Cell. 2015;161(2):361–373.

20. 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. 2020;31(14):1474–1485.

21. Johnson JAI, Bergman DR, Rocha HL, Zhou DL, Cramer E, Mclean IC, et al. Human interpretable grammar encodes multicellular systems biology models to democratize virtual cell laboratories. Cell. 2025;188:4711–4733.e37.

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

23. Edelstein-Keshet L, Ermentrout GB. Contact response of cells can mediate morphogenetic pattern formation. Differentiation. 1990;45(3):147–159.

24. Topaz CM, Ziegelmeier L, Halverson T. Topological Data Analysis of Biological Aggregation Models. PLoS ONE. 2015;10:1–26.

25. Warne DJ, Baker RE, Simpson MJ. Using Experimental Data and Information Criteria to Guide Model Selection for Reaction–Diffusion Problems in Mathematical Biology. Bulletin of Mathematical Biology. 2019;81(6):1760–1804.

26. Metzcar J, Jutzeler CR, Macklin P, Köhn-Luque A, Brüningk SC. A review of mechanistic learning in mathematical oncology. Frontiers in Immunology. 2024;15:1363144.

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

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

29. SenGupta S, Parent CA, Bear JE. The principles of directed cell migration. Nature Reviews Molecular Cell Biology. 2021;22(8):529–547.

30. Baldwin SA, Haugh JM. Semi-autonomous wound invasion via matrix-deposited, haptotactic cues. Journal of Theoretical Biology. 2023;568:111506.

31. Bubna-Litic M, Mayor R. Beyond mechanosensing: How cells sense and shape their physical environment during development. Current Opinion in Cell Biology. 2025;94:102514.

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

33. Metzcar J, Duggan BS, Fischer B, Murphy M, Heiland R, Macklin P. A Simple Framework for Agent-Based Modeling with Extracellular Matrix. Bulletin of Mathematical Biology. 2025;87(3):43.

34. Borau C, Chisholm R, Richmond P, Walker D. An agent-based model for cell microenvironment simulation using FLAMEGPU2. Computers in Biology and Medicine. 2024;179:108831.

35. Crossley RM, Maini PK, Baker RE. Modelling the Impact of Phenotypic Heterogeneity on Cell Migration: A Continuum Framework Derived from Individual-Based Principles. Bulletin of Mathematical Biology. 2025;87(9):123.

36. Bui J, Conway DE, Heise RL, Weinberg SH. Mechanochemical Coupling and Junctional Forces during Collective Cell Migration. Biophysical Journal. 2019;117(1):170–183.

37. Nardini JT, Bortz DM. Investigation of a Structured Fisher's Equation with Applications in Biochemistry. SIAM Journal on Applied Mathematics. 2018;78(3):1712–1736.

38. Li Y, Wang D, Ascoli GA, Mitra P, Wang Y. Metrics for comparing neuronal tree shapes based on persistent homology. PLoS ONE. 2017;12(8):e0182184.

39. Xian L, Adams H, Topaz CM, Ziegelmeier L. Capturing dynamics of time-varying data via topology. Foundations of Data Science. 2022;4(1):1–36.

40. Sisson SA, Fan Y, Tanaka MM. Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences. 2007;104(6):1760–1765.

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

42. Crocker JC, Grier DG. Methods of Digital Video Microscopy for Colloidal Studies. Journal of Colloid and Interface Science. 1996;179(1):298–310.

43. Bauer U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. Journal of Applied and Computational Topology. 2021;5(1):391–423.

44. Scikit-TDA. Scikit-TDA: Persistent homology tools in Python; 2024. `https://scikit-tda.github.io/`.

45. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in Population Genetics. Genetics. 2002;162(4):2025–2035.

46. Thorne T, Stumpf MPH. Graph spectral analysis of protein interaction network evolution. Journal of The Royal Society Interface. 2012;9(75):2653–2666.

47. Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring Coalescence Times From DNA Sequence Data. Genetics. 1997;145(2):505–518.

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

49. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution. 1999;16(12):1791–1798.

50. Beaumont MA. Approximate Bayesian Computation in Evolution and Ecology. Annual Review of Ecology, Evolution, and Systematics. 2010;41(1):379–406.

51. Marin JM, Pudlo P, Robert CP, Ryder RJ. Approximate Bayesian computational methods. Statistics and Computing. 2012;22:1167–1180.

52. Buzbas EO, Rosenberg NA. AABC: Approximate approximate Bayesian computation for inference in population-genetic models. Theoretical Population Biology. 2015;00:31–42.