

Systems biology

# Topological approximate Bayesian computation for parameter inference of an angiogenesis model

Thomas Thorne <sup>1,\*</sup>, Paul D. W. Kirk <sup>2,3,4</sup> and Heather A. Harrington <sup>5,6,\*</sup>

<sup>1</sup>Department of Computer Science, University of Surrey, Guildford GU2 7XH, UK, <sup>2</sup>MRC Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, UK, <sup>3</sup>Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), University of Cambridge, Cambridge CB2 0AW, UK, <sup>4</sup>Cancer Research UK Cambridge Centre, Ovarian Cancer Programme, Cambridge CB2 0RE, UK, <sup>5</sup>Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK and <sup>6</sup>Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK

\*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on October 29, 2021; revised on February 7, 2022; editorial decision on February 14, 2022; accepted on February 18, 2022

## Abstract

**Motivation:** Inferring the parameters of models describing biological systems is an important problem in the reverse engineering of the mechanisms underlying these systems. Much work has focused on parameter inference of stochastic and ordinary differential equation models using Approximate Bayesian Computation (ABC). While there is some recent work on inference in spatial models, this remains an open problem. Simultaneously, advances in topological data analysis (TDA), a field of computational mathematics, have enabled spatial patterns in data to be characterized.

**Results:** Here, we focus on recent work using TDA to study different regimes of parameter space for a well-studied model of angiogenesis. We propose a method for combining TDA with ABC to infer parameters in the Anderson–Chaplain model of angiogenesis. We demonstrate that this topological approach outperforms ABC approaches that use simpler statistics based on spatial features of the data. This is a first step toward a general framework of spatial parameter inference for biological systems, for which there may be a variety of filtrations, vectorizations and summary statistics to be considered.

**Availability and implementation:** All code used to produce our results is available as a Snakemake workflow from [github.com/tt104/tabc\\_angio](https://github.com/tt104/tabc_angio).

**Contact:** [tom.thorne@surrey.ac.uk](mailto:tom.thorne@surrey.ac.uk) or [harrington@maths.ox.ac.uk](mailto:harrington@maths.ox.ac.uk)

## 1 Introduction

When analyzing mathematical models of biological systems, we often aim to reverse engineer the parameters of the model by fitting to observed data. The Bayesian formalism provides a principled way to perform parameter inference that quantifies our uncertainty in the model parameters (see, e.g. Kirk *et al.*, 2015), but traditionally requires us to be able to write down an analytical function (the likelihood function) that returns the likelihood of a parameter vector given the observed data.

However, for many models of interest, there is no straightforward way to write down the likelihood function associated with the model. This is often due to the intractability of deriving a closed form expression for the model likelihood. In such situations, it may nevertheless be possible to apply a simulation-based inference approach termed *Approximate Bayesian Computation* (ABC; see, for example, Sisson *et al.*, 2018), that substitutes a kernel on some statistics of the data for

the model likelihood, and evaluates the fit of the model at a given set of parameter values through simulations. For given parameter realizations, the model is simulated, and the statistics of the simulated data compared with the same statistics of the observed data. Informally, regions of parameter space that correspond to simulated datasets whose statistics are ‘more similar’ to those of the observed data will be associated with higher posterior probability than regions corresponding to simulated datasets with statistics that are ‘less similar’ (where ‘similarity’ is quantified using a pre-specified distance function).

Applying ABC, we can derive an approximate posterior distribution over the model parameters using standard sampling techniques such as rejection sampling. This approximate posterior distribution expresses our uncertainty in the model parameters, given the model and the observed dataset. Recently, ABC parameter inference and model selection has been successfully developed for reaction-diffusion models (Warne *et al.*, 2019). However, performing parameter inference for more general spatial models has been largely unexplored.

Topological data analysis (TDA) is a relatively new area of computational mathematics that quantifies the shape of data by computing topological properties of the data. The appeal of TDA lies in its systematic and principled tools to quantify the shape of data across multiple scales of resolution (i.e. no threshold value). The mathematical theory underlying persistence guarantees that the topological summary is stable with respect to small perturbations to the data. There are various approaches of topological inference, for example level sets or mode clusters (Wasserman, 2018). The most prominent algorithm in TDA is persistent homology (PH; Carlsson, 2009; Edelsbrunner and Harer, 2010). PH takes in data and a metric, and outputs topological features (e.g. connected components and loops) and their persistence across different scales of the data. The computation crucially depends on the choice of filtration, which is a nested sequence of spaces built on the data, that is indexed by a scale parameter (Edelsbrunner and Harer, 2010; Ghrist, 2018). There are many software implementations for persistent homology (Otter et al., 2017); however, the software used is often selected based on the types of filtrations available within it. The choice of filtration for applications is an active area of research, and there is no one-size-fits-all filtration for biological applications (Stolz-Pretzer, 2019). The persistence of the topological features as well as where topological features appear and die in the filtration may provide insight into biological processes and models.

In previous work with spatial models of biological processes (Murray, 2003), TDA has been applied to test for spatial randomness (Robins and Turner, 2016), automatically detect zebra-fish patterns (McGuire et al., 2020), characterize immune cell infiltration by changes in a chemotaxis parameter (Vipond et al., 2021) and cluster parameter regimes for angiogenesis (Nardini et al., 2021). Now we wish to address the inverse problem of recovering model parameters given some observed data, in the Bayesian formalism. ABC enables us to perform parameter inference in a statistical model on the basis of data summaries, even when there is no clear way to define a likelihood function for the model. One key challenge in ABC is the choice of summary statistic, as the statistic must capture the relevant information about the model parameters in the data to allow the parameters to be learnt. Here, we show that TDA provides informative data summaries that enable parameter inference to be performed successfully in a spatial model. In particular, we consider as a case study the Anderson–Chaplain model of angiogenesis (Anderson and Chaplain, 1998).

In previous work in the literature, Maroulas et al. (2020) model persistence diagrams as Poisson point processes and use this to allow a posterior to be inferred on a persistence diagram given some observed data and a suitable prior. This allows a posterior on topological features to be defined, and a scheme for performing Bayesian classification is developed, but it does not consider the case of performing inference on a parametric model, given an observed set of topological features. In Sgouralis et al. (2017), Bayesian inference is applied in the processing of the data, but not in a topological context or for parameter inference in the model of interest. Instead various performance measures are evaluated for a small set of selected parameter combinations, not considering a distribution over parameters or a Bayesian posterior.

In this article, we first describe the model and data generation process applied, before describing TDA and ABC in general terms, and their specific application to the Anderson–Chaplain model. We demonstrate our suggested approach for parameter inference on simulated data from the Anderson–Chaplain model and compare the outputs to the results produced by other non-topological statistics.

## 2 Model data

The Anderson–Chaplain model (Anderson and Chaplain, 1998) is a well-studied spatio-temporal model of angiogenesis. Angiogenesis is the growth of new blood vessels from pre-existing vasculature. The model combines a system of partial differential reaction equations with discrete dynamics to study the spatio-temporal evolution of three physical variables: endothelial tip cells, tumour angiogenesis

factor (TAF) and fibronectin. To set up the angiogenesis model, the right boundary of the square domain is initialized by a tumour that secretes tumour angiogenic factors (TAFs) and the left boundary of the domain is initialized with endothelial tip cells. The tip cells are embedded in a tissue matrix, which is bound to another factor, fibronectin. Tip cells can move either via chemotaxis up spatial gradients of TAF (leaving behind them new blood vessel segments) or via haptotaxis up spatial gradients of fibronectin. As the tip cells migrate, they may branch to create two tip cells, or collide with another vessel segment and join together to form a loop. The changes in vessel structure and connectivity of tumour-blood vessel network makes topology, the study of shapes or holes in different dimensions (e.g. connected components and loops), useful here. TDA can quantify the changes in the number of tip cells and the emergence of loops in experimental data of tumour vasculature (Stolz et al., 2020). Furthermore, topological approaches to analyze structure in data generated from models may be useful in other data applications (see previous section).

The model considers production and consumption of fibronectin, the secretion of tumour angiogenic factors (TAF) from a tumour, and new vasculature forms from endothelial tip cells in response to gradients of fibronectin and TAF; therefore, we focus on the two key parameters,  $\rho$  and  $\chi$ , coefficients for haptotaxis and chemotaxis, respectively. These determine the relative contribution of fibronectin-driven haptotaxis and TAF-driven chemotaxis to the movement of tip cells in the model. Other parameters determine the dynamics of the distribution of fibronectin and TAF, and we keep these fixed as in Nardini et al. (2021). Previous analysis of angiogenesis models relied on visual inspection or spatially averaged statistics such as number of vessel branches (Vilanova et al., 2017); these have been compared with TDA descriptors (Stolz et al., 2020). Previous work showed that TDA stratified the parameter space dominated by either haptotaxis or chemotaxis or both (Nardini et al., 2021). However, the inverse problem requires additional machinery, which we address here.

Data were generated by simulating the Anderson–Chaplain model on a 2D square lattice of resolution 201 by 201 (as in Anderson and Chaplain, 1998) using the implementation provided in Nardini et al. (2021), with a linear chemoattractant distribution that increases with the coordinate along the  $x$  axis. This produces sets of binary images (see Fig. 2) which are then further processed using the methods described below.

## 3 Materials and methods

### 3.1 Topological data analysis

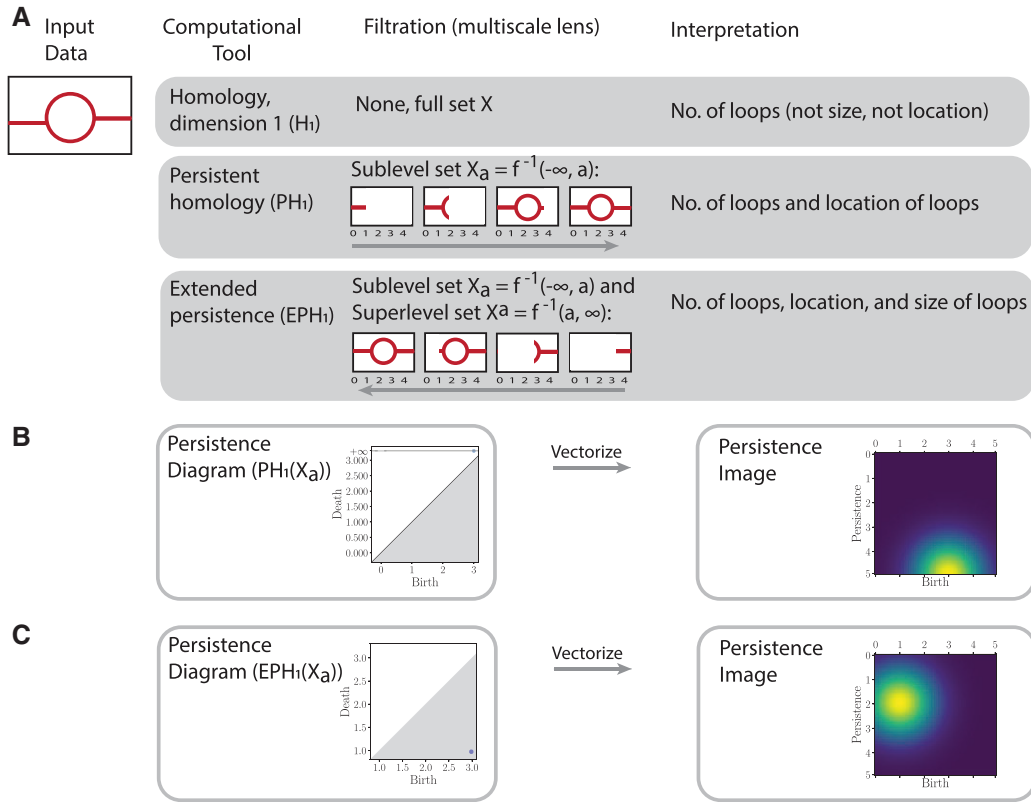
We illustrate the TDA pipeline starting from input data, homology, interpretation and visualization through to topological statistics in Figure 1.

To characterize the  $k$ -dimensional features of a topological space  $X$  we can consider the homology group in dimension  $k$ ,  $H_k(X)$ , composed of elements that intuitively correspond to equivalence classes of cycles that can be continuously deformed into one another on  $X$ . In dimension one, the generators of the homology group correspond to 1D holes in  $X$ , or loops, while in dimension zero the generators of the homology group correspond to the connected components of  $X$ .

The topological spaces we are interested in can be represented using finite sets of simplices known as simplicial complexes  $K$  that are constructed by joining together individual simplices, potentially of different dimensions, and are closed under the operation of taking faces. A 0D simplex corresponds to a single vertex, a 1D simplex an edge, and a 2D simplex a triangle. Given a real valued function on  $K$ , we can define a filtration as a sequence of homology groups in a given dimension  $k$ , with homomorphisms induced by inclusion

$$0 = H_k(K_{a_0}) \rightarrow H_k(K_{a_1}) \rightarrow \dots \rightarrow H_k(K_{a_n}) = H_k(K) \quad (1)$$

where  $K_a = f^{-1}(-\infty, a]$  and  $a_0 < a_1 < \dots < a_n$ , and  $K_{a_i} \subseteq K_{a_j}$  for  $i < j$ . Persistent homology then tracks the birth and death of elements of the homology groups as  $a$  varies. By choosing an appropriate



**Fig. 1.** Topological data analysis pipeline. (A) Illustration of topological features captured by persistence. Take data  $X$  as the image on the left. Homology is an invariant from algebraic topology that captures shape, but ignores geometry. Dimension 0 homology describes connected components whereas dimension 1 homology ( $H_1(X)$ ) describes 1D loops. Persistent homology (PH) quantifies the shape of data through a multiscale lens called a filtration. Here, we use a sublevel set filtration of the data  $X_a = \{0, 1, \dots, 4\}$ , which only includes data to the left of the index, forming a nested sequence of data spaces. PH provides additional information than homology; for this filtration of the data, PH gives the number and location of loops. Extended persistent homology (EPH) requires three computations (ordinary persistence, relative persistence and extended persistence). For this dataset, EPH provides information on the number of loops, size and location. (B, C) The output of persistence computations is summarized by a multi-set of intervals given by birth, death pairs  $(b, d)$ , where  $b$  is when a loop forms and  $d$  is when a loop ends and can be visualized as a persistence diagram. This persistence diagram is then converted into birth, persistence pairs, where persistence is given by  $(d - b)$ , and then vectorized using kernels into persistence images (Adams et al., 2017). Persistence images generate topological statistics of the data that can then be applied in statistical inference. The persistent homology (in B) captures only the birth of the loop with the death at  $\infty$ , whereas the extended component of the extended persistence (in C) also captures the death of the loop

definition of the simplicial complex and filtration built from the data, persistent homology can provide information about the topological features in data.

We build the simplicial complex and filtration from the final timepoint of model simulation data following Nardini et al. (2021). All cells in the 2D square lattice that have vasculature present are assigned a value of one, and zero elsewhere. The centroid of each non-zero cell is a 0-simplex. The simplicial complex is built on these 0-simplices based on so-called Moore neighbourhoods: if any of the eight cells surrounding a vertex are also non-zero, then we connect them via 1-simplices (edges) for two points pairwise connected, or 2-simplices for three points pairwise connected by an edge. The union of these simplices form a *simplicial complex*. There are different ways to study vascular data at multiple scales using filtrations (Bendich et al., 2016; Stolz et al., 2020). Here, we construct sequences of filtered simplicial complexes using a sweeping plane filtration (Bendich et al., 2016; Nardini et al., 2021). In the sweeping plane filtration, we move a vertical line from left to right across the 2D lattice domain and include simplices in the filtration only to the left of this line. This filtration can be considered a sublevel set filtration corresponding to a height function  $b: X \rightarrow \mathbb{R}$  on this simplicial complex.

### 3.2 Approximate Bayesian computation

In Bayesian inference, we aim to derive the posterior distribution of the parameters of a model given some observed data. To do so we first define a prior distribution on the model parameters, treating them as random variables. This describes our belief in the

distribution of the parameters before having observed any data. We then perform a so-called *Bayesian update* of the model having observed some data. This is done using the likelihood of the observed data given the model and parameters. From this, we arrive at a posterior distribution that describes the conditional distribution of the parameters given the observed data. If we denote the model parameters by  $\theta$ , and the data by  $x$ , we can first write the prior as  $p(\theta)$ , and the likelihood of the data as  $p(x|\theta)$ . In the Bayesian framework, we apply Bayes rule to update the prior distribution having observed the data, giving us the posterior distribution as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}, \quad (2)$$

where  $p(x)$  is known as the evidence or marginal likelihood, and plays a key role in Bayesian model selection. Evaluation of the marginal likelihood is often computationally expensive or intractable. However, in many settings (e.g. when sampling from the posterior using Markov chain Monte Carlo techniques), it is sufficient to be able to write down the posterior up to proportionality

$$p(\theta|x) \propto p(x|\theta)p(\theta). \quad (3)$$

This approach relies on the ability to calculate both the prior of the parameters  $p(\theta)$ , which is generally tractable, and the likelihood  $p(x|\theta)$ . However in many models of interest it is not tractable or not possible to directly evaluate  $p(x|\theta)$ , for example in population genetics (Beaumont et al., 2002), random graph models (Thorne and Stumpf, 2012) and some models of dynamical systems (Liepe et al.,

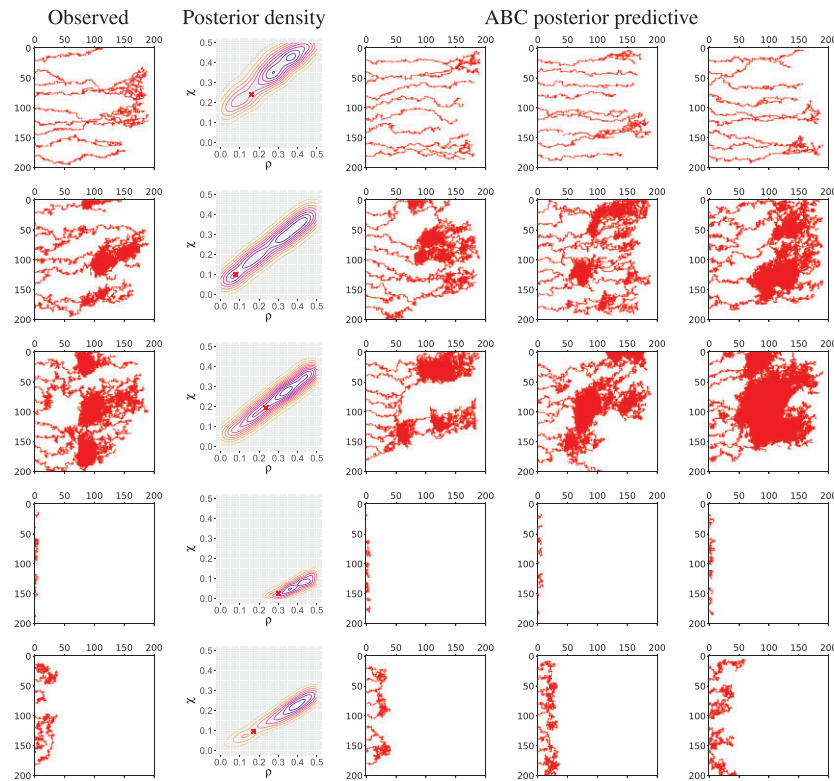


Fig. 2. Visualizations of simulation output from the Anderson–Chaplain model for five parameter sets sampled from a uniform prior on the model parameters. The first column shows the observed data, while the second shows a contour plot of the posterior density inferred by applying the TABC methodology, with the red cross indicating the known parameter values used to generate the observed data. The remaining three columns show simulations of parameter values drawn from the ABC posterior predictive distribution

2014; Toni *et al.*, 2009). To allow us to perform Bayesian inference in these situations, an approach named ABC was developed, based on initial work in Fu and Li (1997) and Tavaré *et al.* (1997), developed further in Beaumont *et al.* (2002) and Marjoram *et al.* (2003), and expanded in many works, see for example Sisson *et al.* (2007); Toni *et al.* (2009); Beaumont *et al.* (2009); Del Moral *et al.* (2012); Prangle *et al.* (2018).

In an ABC framework, we rely on the observation that given the ability to sample realizations  $y$  from  $p(x|\theta)$ , we can rewrite the posterior as

$$p(\theta|x) = \int p(\theta, y|x) dy, \quad (4)$$

where

$$p(\theta, y|x) = \frac{1(x=y)p(y|\theta)p(\theta)}{p(x)}, \quad (5)$$

and by relaxing this to

$$p(\theta, y|x) \approx \frac{1(D(x, y) < \epsilon)p(y|\theta)p(\theta)}{p(x)}, \quad (6)$$

we can generate samples from an approximate posterior (which we shall refer to as the *ABC posterior*) by using a suitably small  $\epsilon$  in Algorithm 1. Often when applying the rejection algorithm, we fix the number of samples  $S$  and select  $\epsilon$  such that the set of samples  $\hat{\theta}_s$  with  $d_s < \epsilon$  is some fraction  $\alpha S$ . The ABC rejection sampler algorithm requires us to define a distance on the data,  $D(x, y)$ , and in some cases this may itself be intractable. It is then possible to substitute a summary statistic of the data,  $g(x)$  in place of the data itself, leading to a distance on these summary statistics  $D(g(x), g(y))$  being considered. In the case where  $g$  is a *sufficient statistic* for the model, as  $\epsilon \rightarrow 0$  this will be equivalent to applying a distance on the  $x$  and  $y$  themselves. Often this is not the case, and this is another avenue

#### Algorithm 1 ABC rejection sampler algorithm

```

1: for  $s \in 1, \dots, S$  do
2:   Sample  $\hat{\theta}_s \sim p(\theta)$ 
3:   Simulate  $y \sim p(y|\hat{\theta}_s)$ 
4:   Calculate  $d_s \leftarrow D(g(y), g(x))$ 
5: end for
6: Return samples  $\hat{\theta}_s$  where  $d_s < \epsilon$ 

```

through which ABC produces an approximation to the posterior rather than a true evaluation of the posterior itself.

### 3.3 Topological statistics for approximate Bayesian computation

In previous work, Nardini *et al.* (2021) applied topological statistics of simulated data (2D binary images) to quantify different regimes in the parameter space of the Anderson–Chaplain model of angiogenesis. By constructing simplicial complexes from the output data of a spatial model, and using the same filtration as Nardini *et al.* (2021), PH can be applied to describe the presence of topological features in the simulated data.

In some cases when calculating the persistence of the topological features of a filtration, it is possible for some features to persist indefinitely, so that their death in the filtration is represented as  $+\infty$ . In our application, this causes information about certain topological features to be lost, for example loops and some connected components, as although we know when they are born in the filtration, we have no measure of their extent. For this reason, Nardini *et al.* (2021) computed persistence of a left to right sweeping plane filtration and right to left sweeping plane filtration of the simplicial



complex built from the simulated model data [see [Nardini et al. \(2021\)](#) for details]. By viewing the left to right filtration as a sublevel set filtration and the right to left filtration as a superlevel set filtration, more information (e.g. only finite bars that capture the extent of topological features) can be extracted as a consequence of duality and symmetry theorems ([Cohen-Steiner et al., 2009](#)).

### 3.4 Extended persistence

Here, we propose a more elegant solution that applies the extended persistence of [Cohen-Steiner et al. \(2009\)](#), which forces all topological features to be of finite length. Extended persistence was developed to study cavities and protrusions in protein docking ([Agarwal et al., 2006](#); [Cohen-Steiner et al., 2009](#)). Since then, [Yim and Leygonie \(2021\)](#) optimized spectral wavelets for graph classification using extended persistence, and extended a differentiability result for ordinary persistence to extended persistence.

In standard persistence, the sublevel sets  $X_a = f^{-1}(-\infty, a]$  of the manifold  $X$  are nested and PH is defined through the corresponding linear sequence of homology groups. In extended persistence, we compute the homology of the sublevel sets, as well as the relative homology with respect to the superlevel sets  $X^a = f^{-1}[a, \infty)$ . For a set of values  $a_0, \dots, a_n$  that bound and fit between the critical points of  $f$ , the extended persistence in dimension  $k$  is defined as the persistence of the homology groups and relative homology groups as

$$\begin{aligned} 0 &= H_k(X_{a_0}) \rightarrow H_k(X_{a_1}) \rightarrow \dots \rightarrow H_k(X_{a_n}) = H_k(X) \\ H_k(X) &= H_k(X, X^{a_n}) \rightarrow \dots \rightarrow H_k(X, X^{a_0}) = 0 \end{aligned} \quad (7)$$

where  $H_k(X, X^a)$  denotes the relative homology group of  $X$  and  $X^a$  in dimension  $k$  ([Edelsbrunner and Harer, 2010](#)).

This extended persistence can be broken down into multiple components ([Cohen-Steiner et al., 2009](#)), the ordinary part, formed of topological features that are both born and die within the homology groups of the sublevel sets of  $X$ , the relative part of features that are born and die in the relative homology groups, and the extended part of features that are born in the ordinary homology groups and die in the relative homology groups in the filtration. The birth time  $b$  of a feature may be larger than its death time  $d$  due to the possibility that the feature dies in the relative homology group  $H(X, X^d)$  with  $d < b$ . The extended part can be further divided into topological features that have  $b < d$ , termed extended+, and those with  $d < b$ , termed extended-.

### 3.5 Persistence images

The output of applying PH to a dataset is often represented as a persistence diagram, that for a given dimension  $k$  consists of a plot of points  $(b, d)$ , where  $b$  is the time of birth and  $d$  is the time of death  $d$  of each dimension  $k$  topological feature in the filtration. To allow for the straightforward application of methods from machine learning to these diagrams, [Adams et al. \(2017\)](#) developed the concept of a persistence image. This allows a persistence diagram to be represented as a vector in  $R^n$ , so that for example it can be used in methods such as K-means clustering, as in [Nardini et al. \(2021\)](#).

To generate the persistence image corresponding to a persistence diagram represented as a multiset of points  $(b, d)$ , the points are first transformed to give a multiset  $B$  of birth and persistence coordinates  $(b, d - b)$  (for extended persistence, we require a slightly different formulation—see below). We note that the persistent image formulation of [Adams et al. \(2017\)](#) ignores all infinite persistent features. A persistence surface in  $R^2 \rightarrow R$  is then defined as the weighted sum of kernels applied to each birth/persistence coordinate

$$f(x, y) = \sum_{(b,p) \in B} g(b, p) b(x, y; b, p), \quad (8)$$

where  $g(b, p)$  is the weight of the feature and  $b$  is a suitable kernel. From the persistence surface defined in Equation (8), an  $m \times m$  array of values is created by discretizing  $f(x, y)$  into an  $m$  by  $m$  grid in a suitable range. This array can then be flattened to give a vector in  $R^{m^2}$ . As in [Adams et al. \(2017\)](#), we apply a Gaussian kernel for  $b$  with mean  $\mu = (b, p)$  and fixed standard deviation  $\sigma$ .

We remark that extended persistence only has finite persistence; therefore, no information (i.e. the infinite bars in ordinary persistence) is lost in the persistence images for extended persistent homology.

### 3.6 TABC

We use a set of topological statistics derived from the extended persistence of a filtration over the simplicial complex representing the data as the summary statistics in an ABC framework, in a method we title TABC, to perform topological posterior inference on the Anderson–Chaplain model of angiogenesis. In the TABC methodology, the summary statistics used in ABC are the persistence images in each dimension produced by the by the four components of the extended persistence of a filtration. To allow persistence images to be generated for the extended persistence, in components of the extended persistence with points in the persistence diagram  $(b, d)$  with  $d < b$ , we flip the coordinates to consider instead  $(d, b)$ , which when transformed into a birth/persistence coordinate then represents the duration of persistence of the feature in the relative part, or the gap between birth in the ordinary homology and death in the relative homology of the feature in the extended-part. We generate persistence images of dimension 50 by 50 with a constant weight function for the persistence surface and the kernel of the persistence images set as a multivariate Gaussian distribution with standard deviation  $\sigma = 1$ , as we found this to work well. As the distance metric in the ABC algorithm, we applied the Euclidean distance between the statistics. In our implementation we use the GUDHI library (<http://gudhi.gforge.inria.fr/>) to construct simplicial complexes, generate extended persistence diagrams and produce persistence images (with standard weighting  $g = 1$ ).

### 3.7 Image-based statistics

For comparison, we also consider four statistics based on the binary image data produced by the simulations, that were chosen with the aim of differentiating the different classes of behaviours observed in [Nardini et al. \(2021\)](#), without overlapping with features that could be considered as topological descriptors (e.g. numbers of connected components). These statistics are:

- **Mean X coordinate:** The mean X value of occupied pixels.
- **Mean Y coordinate:** The mean Y value of occupied pixels.
- **Maximum X coordinate:** The maximum X value of an occupied pixel.
- **Mass:** The fraction of occupied pixels.

As with the topological statistics, we applied the Euclidean distance between vectors of statistics as the distance in the ABC rejection algorithm.

## 4 Results

We apply the TABC approach described above to parameter inference in the Anderson–Chaplain model. Taking 10 000 samples from the prior on the two model parameters, we simulated the Anderson–Chaplain model of angiogenesis for each sampled parameter pair.

To validate our approach, we drew a further 100 parameter sets from the model prior and simulated data from each to take on the role of the observed data. A representative subset of these simulated datasets can be seen in [Figure 2](#), and cover a range of different behaviours.

Given these data, we applied the TABC approach described above to derive samples of 500 parameter values from the ABC posterior. To investigate the ability of our topological approach to accurately capture the relevant behaviour of the model, we generated ABC posterior predictive samples by simulating the model using parameter values drawn at random from the ABC posterior. These are shown in [Figure 2](#), and demonstrate that TABC enables the effective recovery of parameters that replicate the qualitative behaviour of the observed data.

**Table 1.** Mean of the root sum of squared errors and entropy of the posterior distribution inferred from simulated data for 100 parameter sets drawn from a uniform prior

Statistics	Mean RSSE	$2\sigma_{\bar{x}}$ RSSE	Mean entropy	$2\sigma_{\bar{x}}$ entropy
Image	4.30	0.25	−2.86	0.12
Topological	3.61	0.27	−3.31	0.12

Note: Values for both the TABC-based posterior and ABC on the image-based statistics are shown.

It can be seen that the ABC posterior distributions for the two parameters demonstrate a degree of unidentifiability, in that in most cases the posterior follows a ridge shape with a strong correlation between the two parameters. This aligns with the results found in Nardini *et al.* (2021), where it was discovered that there were distinct classes of behaviour that occupied diagonal sections of the parameter space, as do our posterior distributions. Being able to identify such uncertainty in our parameter estimates is one of the key benefits of a Bayesian analysis, and it also provides insights into the behaviour of the model. For example we can see that ABC posterior predictive samples in Figure 2 are representative of a given class of model behaviour, and that draws from across the potentially wide distribution of parameters indicated by the posterior will follow this behaviour.

The known parameter values used to generate the data on which the posterior distributions are based are marked in Figure 2, and can be seen to be within the bulk of the ABC posterior mass.

To further quantify the efficacy of our approach, we compared statistics of the posterior distributions obtained from TABC with those generated by an ABC approach using only the image-based statistics described in Section 3.7. We quantified the accuracy of the inferred parameters by taking the mean root sum of squared errors (RSSE) between the posterior samples and the ‘true’ parameters used to generate the data, as shown in Table 1. Here, the mean RSSE achieved by the topological posterior over the 100 simulated datasets is below that of the posterior generated using image-based statistics. We also calculated the mean entropy of the posterior distributions produced for each observed data point using both TABC, and ABC with image-based statistics. As can be seen in Table 1, the entropy for the posterior derived from the topological features is lower than that derived from the image-based statistics. Taken together, the RSSE and entropy results suggest that the topological statistics used in TABC retain more of the information in the original dataset, and hence that TABC is able to more accurately infer the parameters used to generate the data, than ABC using image-based statistics alone.

5 Conclusions

We have developed an approach for performing ABC in a topological context that is able to derive posterior distributions over model parameters that can accurately reproduce multiple different classes of behaviour and structure observed within the data. We applied extended persistence, which strictly quantifies more topological features than ordinary persistence. Other topological shape statistics have focussed on sweeping across data in multiple different directions (Crawford *et al.*, 2020; Curry *et al.*, 2018; Turner *et al.*, 2014). Their utility for parameter inference and model selection will be explored in future studies.

Evaluating the ABC posterior distributions we obtain, we find that by considering topological features in the data through the TABC approach we are able to reduce the posterior uncertainty in the parameter values, and to infer posterior distributions that are more closely focused around the parameters used to generate the data.

While we use persistence images here, there are other potential approaches to summarizing TDA for use in parameter inference. For example it is possible to directly derive distances between persistence diagrams in a number of ways (Atienza *et al.*, 2020; Bubenik, 2015; Carrière *et al.*, 2017, 2015; Chazal *et al.*, 2014; Di Fabio and

Ferri, 2015; Kerber *et al.*, 2017; Lacombe *et al.*, 2018; Royer *et al.*, 2021), and these could be substituted for the Euclidean distance between the vectors of persistence images that we apply. In future work, we will investigate the possibility of applying a distance function on persistence diagrams in the ABC likelihood and how this influences the efficiency of the algorithm.

For simplicity, we have also only considered the simplest form of the ABC algorithm—many other increasingly sophisticated approaches exist, including Markov Chain Monte Carlo algorithms, Sequential Monte Carlo methods (Sisson *et al.*, 2007) and rare event schemes (Prangle *et al.*, 2018). It would be expected that for models with larger numbers of parameters, significant improvements in efficiency could be obtained by applying one of these approaches rather than a rejection sampler-based ABC approach. Doing so would not require any changes to the topological aspects of TABC, only the encompassing sampling mechanism. More precisely, since TABC can be considered as a conventional ABC approach in which the ABC summary statistic is constructed using TDA, we would anticipate that extending to SMC would follow the standard approach of propagating particles representing points in the parameter space through a sequence of  $\epsilon$  thresholds, with adaptive methods based on effective sample sizes being possible to define a suitable threshold sequence (e.g. Del Moral *et al.*, 2012; Silk *et al.*, 2013).

A further direction of study would be to consider applications of TABC in the context of model choice (Kirk *et al.*, 2013). While concepts from TDA have been successfully used to perform model comparison (Vittadello and Stumpf, 2021), we note that TABC inherits the same formal challenges regarding model selection as other ABC algorithms, due to the loss of information arising from the use of an insufficient summary statistic (Robert *et al.*, 2011). As with other ABC algorithms, model criticism (Ratmann *et al.*, 2009) and approaches that rephrase model selection as a classification problem (Pudlo *et al.*, 2016) are likely to provide fruitful avenues for future research.

As with some other applications of ABC (e.g. Russell-Buckland *et al.*, 2019), a potential strength of our approach is that it enables a form of *qualitative* inference to be performed; in our case by allowing combinations of parameters that result in model behaviour that is topologically similar to the observed data to be identified. Although we consider a specific application, to parameter inference in the Anderson–Chaplain model of angiogenesis, the TABC approach may be adapted to be widely applicable to parametric models having topological features in the data that are informative about model parameters, including in situations where a mixture of topological statistics and other complementary statistics could be used.

Acknowledgements

H.A.H. thanks PG Kevrekidis and N Whitaker for first presenting the challenge of inferring parameters in models of angiogenesis. H.A.H. also thanks members of the Centre for TDA, specifically A. Barbensi, H. Byrne, L. Marsh and U. Tillmann for many stimulating discussions and helpful comments. P.D.W.K. is grateful to L. Reali for useful conversations.

Funding

This work was supported by the Medical Research Council [MC\_UU\_00002/13] and the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust) to P.D.W.K. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. H.A.H. gratefully acknowledges funding from EPSRC EP/R018472/1, EP/R005125/1 and EP/T001968/1, the Royal Society RGF/EA\201074 and UF150238, and Emerson Collective. Partly funded by the RESCUER project. RESCUER has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No. 847912.

Conflict of Interest: none declared.

## Data availability

All code used to produce our results is available as a Snakemake (Köster and Rahmann, 2012) workflow from [github.com/tt104/tabc\\_angio](https://github.com/tt104/tabc_angio). It is also stored as an archive on Zenodo with doi: 10.5281/zenodo.5562670.

## References

- Adams, H. *et al.* (2017) Persistence images: a stable vector representation of persistent homology. *J. Mach. Learn. Res.*, **18**, 1–35.
- Agarwal, P.K. *et al.* (2006) Extreme elevation on a 2-manifold. *Discrete Comput. Geometry*, **36**, 553–572.
- Anderson, A.R. and Chaplain, M.A. (1998) Continuous and discrete mathematical models of tumor-induced angiogenesis. *Bull. Math. Biol.*, **60**, 857–899.
- Atienza, N. *et al.* (2020) On the stability of persistent entropy and new summary functions for topological data analysis. *Pattern Recognit.*, **107**, 107509.
- Beaumont, M.A. *et al.* (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
- Beaumont, M.A. *et al.* (2009) Adaptive approximate Bayesian computation. *Biometrika*, **96**, 983–990.
- Bendich, P. *et al.* (2016) Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.*, **10**, 198–218.
- Bubenik, P. (2015) Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, **16**, 77–102.
- Carlsson, G. (2009) Topology and data. *Bull. Am. Math. Soc.*, **46**, 255–308.
- Carrière, M. *et al.* (2015) Stable Topological Signatures for Points on 3D Shapes. *Computer Graphics Forum*, **34**, 1–12. <https://doi.org/10.1111/cgf.12692>.
- Carrière, M. *et al.* (2017) Sliced Wasserstein Kernel for persistence diagrams. In: *International Conference on Machine Learning, Sydney, Australia*. PMLR, pp. 664–673.
- Chazal, F. *et al.* (2014) Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the Thirtieth Annual Symposium on Computational Geometry, SOCG'14*. Association for Computing Machinery, New York, NY, USA, pp. 474–483.
- Cohen-Steiner, D. *et al.* (2009) Extending persistence using Poincaré and Lefschetz duality. *Found. Comput. Math.*, **9**, 79–103.
- Crawford, L. *et al.* (2020) Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. *J. Am. Stat. Assoc.*, **115**, 1139–1150.
- Curry, J. *et al.* (2018) How many directions determine a shape and other sufficiency results for two topological transforms. *arXiv, preprint arXiv:1805.09782*.
- Del Moral, P. *et al.* (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.*, **22**, 1009–1020.
- Di Fabio, B. and Ferri, M. (2015). Comparing persistence diagrams through complex vectors. In: Murino, V. and Puppo, E. (eds.) *Image Analysis and Processing – ICIAP 2015, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 294–305.
- Edelsbrunner, H. and Harer, J. (2010) *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI.
- Fu, Y.X. and Li, W.H. (1997) Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.*, **14**, 195–199.
- Ghrist, R. (2018) Homological algebra and data. In: Mahoney, M.W., Duchy, J.C. and Anna C. Gilbert, A.C. (eds.) *The Mathematics of Data, Volume 25 of IAS/Park City Mathematics Series*. American Mathematical Society, Providence, RI, pp. 273–325.
- Kerber, M. *et al.* (2017) Geometry helps to compare persistence diagrams. *ACM J. Exp. Algorithmics*, **22**, 1–14:20.
- Kirk, P. *et al.* (2013) Model selection in systems and synthetic biology. *Curr. Opin. Biotechnol.*, **24**, 767–774.
- Kirk, P. *et al.* (2015) Systems biology (un)certainities. *Science*, **350**, 386–388.
- Köster, J. and Rahmann, S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
- Lacombe, T. *et al.* (2018) Large scale computation of means and clusters for persistence diagrams using optimal transport. In: *Advances in Neural Information Processing Systems*, Vol. 31.
- Liepe, J. *et al.* (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.*, **9**, 439–456.
- Marjoram, P. *et al.* (2003) Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, **100**, 15324–15328.
- Maroulas, V. *et al.* (2020) A Bayesian framework for persistent homology. *SIAM J. Math. Data Sci.*, **2**, 48–74.
- McGuirl, M.R. *et al.* (2020) Topological data analysis of zebrafish patterns. *Proc. Natl. Acad. Sci. USA*, **117**, 5113–5124.
- Murray, J.D. (2003) *Mathematical Biology II: spatial Models and Biomedical Applications*. Interdisciplinary Applied Mathematics, Mathematical Biology, 3rd edn. Springer-Verlag, New York.
- Nardini, J.T. *et al.* (2021) Topological data analysis distinguishes parameter regimes in the Anderson-Chaplain model of angiogenesis. *PLoS Comput. Biol.*, **17**, e1009094.
- Otter, N. *et al.* (2017) A roadmap for the computation of persistent homology. *Eur. Phys. J. Data Sci.*, **6**, 1–38.
- Prangle, D. *et al.* (2018) A rare event approach to high-dimensional approximate Bayesian computation. *Stat. Comput.*, **28**, 819–834.
- Pudlo, P. *et al.* (2016) Reliable abc model choice via random forests. *Bioinformatics*, **32**, 859–866.
- Ratmann, O. *et al.* (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natl. Acad. Sci. USA*, **106**, 10576–10581.
- Robert, C.P. *et al.* (2011) Lack of confidence in approximate Bayesian computation model choice. *Proc. Natl. Acad. Sci. USA*, **108**, 15112–15117.
- Robins, V. and Turner, K. (2016) Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Phys. D Nonlinear Phenomena*, **334**, 99–117.
- Royer, M. *et al.* (2021) ATOL: measure vectorization for automatic topologically-oriented learning. In: *International Conference on Artificial Intelligence and Statistics, Virtual conference*. PMLR, pp. 1000–1008.
- Russell-Buckland, J. *et al.* (2019) A Bayesian framework for the analysis of systems biology models of the brain. *PLoS Comput. Biol.*, **15**, e1006631.
- Sgouralis, I. *et al.* (2017) A Bayesian topological framework for the identification and reconstruction of subcellular motion. *SIAM J. Imaging Sci.*, **10**, 871–899.
- Silk, D. *et al.* (2013) Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Stat. Appl. Genet. Mol. Biol.*, **12**, 603–618.
- Sisson, S.A. *et al.* (2007) Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, **104**, 1760–1765.
- Sisson, S.A. *et al.* (eds.) (2018) *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, Boca Raton.
- Stolz-Pretzer, B. (2019) *Global and Local Persistent Homology for the Shape and Classification of Biological Data*. Ph.D. Thesis. University of Oxford. <http://purl.org/dc/dcmitype/Text>.
- Stolz, B.J. *et al.* (2020) Multiscale topology characterises dynamic tumour vascular networks. *arXiv, preprint arXiv:2008.08667*.
- Tavaré, S. *et al.* (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.
- Thorne, T. and Stumpf, M.P.H. (2012) Graph spectral analysis of protein interaction network evolution. *J. R. Soc. Interface*, **9**, 2653–2666.
- Toni, T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.
- Turner, K. *et al.* (2014) Persistent homology transform for modeling shapes and surfaces. *Inf. Inference J. IMA*, **3**, 310–344.
- Vilanova, G. *et al.* (2017) A mathematical model of tumour angiogenesis: growth, regression and regrowth. *J. R. Soc. Interface*, **14**, 20160918.
- Vipond, O. *et al.* (2021) Multiparameter persistent homology landscapes identify spatial patterns of immune cells in tumors. *Proc. Natl. Acad. Sci. USA*, **118**, e2102166118.
- Vittadello, S.T. and Stumpf, M.P.H. (2021) Model comparison via simplicial complexes and persistent homology. *R. Soc. Open Sci.*, **8**, 211361.
- Warne, D.J. *et al.* (2019) Using experimental data and information criteria to guide model selection for reaction–diffusion problems in mathematical biology. *Bull. Math. Biol.*, **81**, 1760–1804.
- Wasserman, L. (2018) *Topological Data Analysis*. SSRN Scholarly Paper ID 3156968, Social Science Research Network, Rochester, NY.
- Yim, K.M. and Leygonie, J. (2021) Optimization of spectral wavelets for persistence-based graph classification. *Front. Appl. Math. Stat.*, **7**, 16.