

# Model comparison via simplicial complexes and persistent homology

Sean T. Vittadello<sup>1,\*</sup> and Michael P. H. Stumpf<sup>1,2</sup>

## Abstract

In many scientific and technological contexts we have only a poor understanding of the structure and details of appropriate mathematical models. We often, therefore, need to compare different models. With available data we can use formal statistical model selection to compare and contrast the ability of different mathematical models to describe such data. There is, however, a lack of rigorous methods to compare different models *a priori*. Here we develop and illustrate two such approaches that allow us to compare model structures in a systematic way by representing models in terms of simplicial complexes. Using well-developed concepts from simplicial algebraic topology, we define a distance between models based on their simplicial representations. Employing persistent homology with a flat filtration provides for alternative representations of the models as persistence intervals, which represent the structure of the models, from which we can also obtain the distances between models. We then expand on this measure of model distance to study the concept of model equivalence in order to determine the conceptual similarity of models. We apply our methodology for model comparison to demonstrate an equivalence between a positional-information model and a Turing-pattern model from developmental biology, constituting a novel observation for two classes of models that were previously regarded as unrelated.

---

**Key words and phrases:** Turing pattern, positional information, algebraic topology, algebraic biology, persistent homology, persistence barcode, model distance, model equivalence

<sup>1</sup> School of BioSciences, The University of Melbourne, Parkville VIC 3010 Australia

\* Corresponding author: sean.vittadello@unimelb.edu.au

<sup>2</sup> School of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3010 Australia.

# 1 Introduction

Scientific models are representations of the physical world that isolate features of interest through various levels of abstraction [1]. The complexity of biological phenomena, exemplified in systems biology, necessitates the development of models to further the understanding of these systems [2–5].

There are many approaches to modelling biological systems, including continuum models [6–8], rule-based models [9–11], network models [12–14], and mechanistic models [5, 15], which are not necessarily mutually exclusive classifications. While a model should represent the corresponding phenomenon as faithfully as possible, it is also a requirement that the model is manageable. It is neither practicable to develop a complete model of a biological system, due to *combinatorial complexity* [3, 10, 16], nor is it necessary, since many components of a given system have relatively small effects on the system behaviour [4, 17].

A biological system may be represented by various alternative models, which can differ with respect to the specific features isolated by each model and the interconnections between these features. Contrarily, similar models may represent ostensibly divergent biological systems, where the difference between the models may be only conceptual or, indeed, simply different parameterisations [18]. The ability to rigorously and efficiently compare models is therefore imperative for understanding complex biological systems, and can elucidate the relationship between apparently distinct phenomena [19–23]. Model comparison is, however, a nontrivial task for multiple reasons: the various representational forms of models; the different levels of abstraction and granularity between models; and the need for a systematic formalism that provides elucidation of conceptual similarities between models. In particular, we require a way of obtaining distances between models: with a meaningful distance between models we can start to cluster ‘similar’ models and look for shared characteristics; we can explore how ‘different’ proposed alternative models for a given process really are; and we can use them to distill design principles from large sets of models that have been proposed to model a given biological process or problem.

There are very few attempts to establish distances between models [24, 25], and many of the candidates for a distance between models have obvious shortcomings: graph-based distances on model structures, for example, fail because many dynamical systems cannot be described in terms of simple graphs [18]; other well-defined distances used in functional analysis are too limited in scope, as are distances among e.g. stoichiometric matrices, which only apply for certain types of models; semantic distances between model formulations in e.g. CellML or SMBL, can depend too strongly on inconsistencies in modelling terminology, and also can fail to distinguish between structural and semantic differences; distances among model outputs such as information-theoretical distances, distances used in control theory, or correlation, (dis)similarity measures require assessment of model outputs and generally do not allow a direct structural comparison between models; finally, coding distances

or the minimum description length would be promising candidates for distances between models. We could consider these, but they, too, have shortcomings, especially for different modelling frameworks (say ordinary differential equations or stochastic Petri-nets used to model the same scenario [26]).

Therefore, in this article we present a novel formalism for model comparison that is flexible, systematic, quantitative, and visual. Since models consist of a finite collection of components and their interconnections, we represent a given model as a combinatorial object where the components of the model are represented as labelled vertices. Further, since the interconnections between model components may have dimension higher than the dyadic connections in combinatorial graphs, we employ simplicial complexes to represent the models which allow for interconnections of any finite dimension. Representing models as simplicial complexes ensures that any model can be represented within our framework, irrespective of form, granularity, and level of abstraction. We have to define model components – for this we may draw on domain expertise relevant to the scientific modelling problem at hand – and model comparison is then performed by comparison of the simplicial representations of the models given these components. Our framework is both flexible, incorporating different models and modelling approaches, and rigorous.

Given the model components, the simplicial representations provide a graphical form for the models; however we can also obtain visual representations of the models as collections of persistence intervals by applying persistent homology to the simplicial complexes. Persistent homology identifies the global topological structure of a space that persists across multiple scales [27,28], and is often employed in topological data analysis to study high-dimensional data sets [29]. Persistent homology is robust with respect to small perturbations in the input data, and represents the qualitative structure of the input data in a compact manner [29]. While we use tools that are also employed in topological data analysis, our aims here are different: we are trying to represent models in a framework that allows for meaningful comparisons, rather than looking for topological structure in high-dimensional data. Our application of persistent homology is novel since our simplicial complexes correspond to models rather than data sets, and we utilise the persistence intervals associated with a simplicial complex as an alternative representation of the model and hence the corresponding biological system. The persistence-interval representations tend to facilitate the visual comparison between various models, which may be more difficult with the simplicial representations and particularly with the models in their original forms. Indeed, since three- and higher-dimensional simplices cannot be drawn in two dimensions, only projected, representations as persistence barcodes provide immediate information for all dimensions in an often clearer visual form.

We provide two general methodologies for model comparison. The first is comparison by distance, which measures the difference between models in terms of the differences between the corresponding labelled simplicial representations, and can be calculated either directly using the simplicial complexes or indirectly with the cor-

responding persistence intervals. The second is comparison by equivalence, where we identify any equivalences between the components of the simplicial representations of different models, and then employ operations on the complexes to transform one into the other. Where this is possible we can learn about shared underlying characteristics between models.

As a particular application of our methodology we compare the two main classes of models for developmental-pattern formation, namely Turing-pattern and positional-information models. While there is an extensive literature for models of both Turing-pattern and positional-information systems, the relationship between these models, and between the corresponding biological systems, has long been unclear [30]. One main outcome of our model-comparison methodology is the demonstration that the Turing-pattern activator-inhibitor model is equivalent to the positional-information annihilation model from a significant conceptual perspective, since the fundamental difference between the two models is the location of the source of the gradient-forming morphogen. This novel observation suggests that the location of the morphogen source may influence the particular mechanism, namely Turing-pattern or positional information, for gradient formation.

The remainder of this article is organised as follows. In Section 2 we introduce our methodology for model comparison, which consists of five subsections: Subsection 2.1 provides a brief discussion of the required background in algebraic topology; in Subsection 2.2 we define our notion of a simplicial representation of a model; in Subsection 2.3 we develop the persistent homology of the simplicial representations of models; Subsection 2.4 provides the definition of distance between two simplicial representations of models; and, in Subsection 2.5 we establish our notion of model equivalence. We then apply our model-comparison methodology to two examples in Section 3, firstly a comparison of bisubstrate enzyme reactions, and secondly a comparison of Turing-pattern and positional-information models. Finally, in Section 4 we summarise the utility of our methodology for comparing models.

## 2 Methodology

### 2.1 Background

We begin with a brief, informal, and self-contained discussion of the background in simplicial algebraic topology that is relevant for our work. A more detailed overview of simplicial complexes and homology is provided in the Electronic Supplementary Material document. There is also a growing literature on related issues in topological data analysis [29], but there the aims and the details of implementation differ considerably from what we set out to achieve here.

A *p-simplex* is a generalisation of a filled triangle in the plane to an arbitrary dimension  $p$ , whereby a point

is a 0-simplex, a line segment is a 1-simplex, a filled triangle is a 2-simplex, a filled tetrahedron is a 3-simplex, and so forth. We can think of 0-simplices as vertices and 1-simplices as edges, so that a simple graph is therefore a set of 0-simplices and 1-simplices. A  $k$ -*face* of a simplex is a  $k$ -dimensional subsimplex, and the 0-faces of a simplex are said to *span* the simplex. A face  $\tau$  of a simplex  $\sigma$  is *proper* if  $\tau \neq \sigma$ .

A *simplicial complex* is a generalisation of a simple graph, allowing for simplicies of dimension higher than one which represent higher-dimensional interactions. Specifically, a simplicial complex  $K$  consists of a set of simplices such that: if  $\sigma$  is a simplex in  $K$  then every face of  $\sigma$  is also in  $K$ ; and, the nonempty intersection of any two simplices in  $K$  is a simplex in  $K$ . A *simplicial subcomplex*  $L$  of a simplicial complex  $K$  is a collection  $L \subseteq K$  that is also a simplicial complex, and a *simplicial supercomplex*  $M$  of  $K$  is a collection  $M \supseteq K$  that is also a simplicial complex. The  $k$ -*skeleton* of a simplicial complex  $K$  is the subcomplex  $K^{(k)}$  consisting of the simplices in  $K$  with dimension at most  $k$ . In particular, the 0-skeleton  $K^{(0)}$  is the set of vertices and the 1-skeleton  $K^{(1)}$  is the *underlying graph* of the simplicial complex  $K$ .

A simplicial complex  $K$  is a topological space, formed from the union of its simplices, and simplicial homology characterises the complex using algebraic techniques to compute the number of  $k$ -dimensional holes in the complex. So, for example, 0-dimensional holes are connected components and 1-dimensional holes are nonbounding cycles of edges. The homology of  $K$  depends on the underlying space of simplices and their intersections in  $K$ . Persistent homology studies the topological features, namely the  $k$ -dimensional holes, of a complex  $K$  across multiple scales, based on a *filtration* of a simplicial complex which is an increasing sequence of subcomplexes. Each  $k$ -dimensional hole in  $K$  is created at some index in the filtration, and either persists through to the full complex  $K$  or is annihilated at some intermediate index. Persistent homology therefore gives a multiset of *persistence intervals*  $\mathcal{P}(K)$  describing the creation and annihilation of all  $k$ -dimensional holes in  $K$ . We denote by  $\mathcal{P}_k(K)$  the submultiset of  $\mathcal{P}(K)$  consisting of the intervals corresponding to  $k$ -dimensional holes. The persistence intervals can be visualised as a *persistence barcode*, displaying the intervals as horizontal line segments.

## 2.2 Simplicial representations of models

By a *model* we mean an abstraction of an observable phenomenon [1]. A specific detail of a given model is a *component* of the model, and all models consist of a finite number of components and their interconnections, which represent direct relationships between specific components. For example, a Turing-pattern system of reaction-diffusion equations, which is a model of a developmental-patterning process, includes components such as morphogens, diffusion, boundary conditions, reactions, and a morphogen gradient, along with the interconnections between particular components such as a morphogen and its diffusion.

In comparing models, including when defining a distance between models, it is important to ensure that model components are defined consistently across all models. Beyond this, the definition of model components typically arises naturally from the scientific-modelling context. It is essential to keep in mind that model comparisons are always with regard to the chosen level of conceptual detail used to represent the models. The intention of a simplicial representation of a model is to identify the concepts underlying the model, and their interconnections, thereby removing the formality of the mathematical representations to allow for the comparison of models within the same framework.

We begin by defining the set of model components:

**Definition 2.1 (Model components, generated model).** Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison. We say that each such model is *generated* by a subset of components from  $\mathcal{C}$ .

We represent model components as 0-simplices, and label the 0-simplices in a flexible manner to allow for more efficient or detailed labelling as required:

**Notation 2.2 (Representations of model components).** Let  $\text{Ord}: \mathcal{C} \rightarrow \{1, 2, \dots, |\mathcal{C}|\}$  be a bijection that specifies an arbitrary order for the categorical data elements in  $\mathcal{C}$ . We represent each model component from  $\mathcal{C}$  as a 0-simplex which is labelled variously as  $i$ ,  $v_i$  in recognition that 0-simplices are vertices, or the name of the model component given by  $\text{Ord}^{-1}(i)$ , where  $i$  is the position in the total order specified by  $\text{Ord}$ .

The ordering function  $\text{Ord}$  provides for computationally-efficient labelling of simplices. For example, a 2-simplex can be labelled as  $\{1, 2, 3\}$ , where the 0-simplices are  $\{1\}$ ,  $\{2\}$ , and  $\{3\}$ , and the 1-simplices are  $\{1, 2\}$ ,  $\{1, 3\}$ , and  $\{2, 3\}$ . Note that we have no need for the orientation of the simplicial complexes that is provided by the total ordering of the vertices by  $\text{Ord}$ , since we only consider simplicial homology over  $\mathbb{Z}/2\mathbb{Z}$ .

To a given model we can associate various related simplicial complexes depending on the required levels of abstraction and component detail. Each associated simplicial complex consists of the relevant labelled 0-simplices that represent the model components, along with the one- and higher-dimensional simplices that represent the interconnections between the components, where the dyadic interactions are 1-simplices, the triadic interactions are 2-simplices, and so forth as required. The labelling of the 0-simplices induces a labelling of the higher-dimensional simplices through their spanning 0-simplices. Since the models of interest are associated with the same set of components  $\mathcal{C}$ , we can identify a particular labelled simplex, representing specific components and interconnections, within the different simplicial complexes associated with the models.

A general algorithm to associate a labelled simplicial complex  $K$  with a model is as follows:

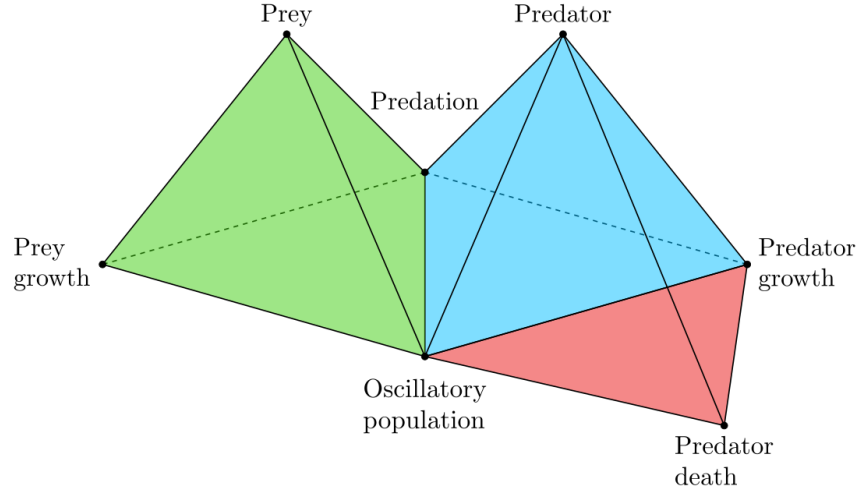
1. Identify the agents in the model.
2. Identify the reactions and interactions of the agents in the model, which may provide for various levels of conceptual detail.
3. Identify any boundary conditions for the model.
4. Identify, if desired, the parameters in the model, which may allow for various levels of detail.
5. Define the set  $\mathcal{C}$  as consisting of the components identified in steps 1–4.
6. The simplicial representation  $K$  has  $|\mathcal{C}|$  vertices with the labels from  $\mathcal{C}$ .
7. The 1-simplices of  $K$  correspond to direct connections between the relevant concepts, which are immediate from the model. To determine whether there should be a 1-simplex linking two vertices it may be easier to consider whether the absence of the 1-simplex is consistent with the model structure.
8. Higher-dimensional simplices may or may not be required, depending on the particular model and the desired level of representation. Any higher-dimensional simplices must represent mutual interconnections between all model concepts involved.

With regard to steps 7–8 we note that, while all 1- and higher-dimensional simplices of  $K$  are completely determined by the model, such interconnections between the model components may not be obvious and may require a deep understanding of the model.

As an example we consider the Lotka-Volterra model [31, Chapter 3, page 79], which is described by

$$\frac{dx}{dt} = x(a - by) \quad \text{and} \quad \frac{dy}{dt} = y(cx - d), \quad (1)$$

where  $x$  is the prey density,  $y$  is the predator density, and  $a$ ,  $b$ ,  $c$ , and  $d$  are positive real constants. The underlying assumptions of the model are: in the absence of predation the prey population grows without bound; the reduction of the prey's per capita growth rate due to predation is proportional to both the prey and predator populations; in the absence of prey the predator's death rate decays exponentially; and, the prey's contribution to the predator's growth rate is proportional to both the available prey and the size of the predator population. In Figure 1 we provide a simplicial representation of the Lotka-Volterra model. The simplicial representation accounts for the underlying concepts of the model rather than the particular mathematical form of the model, as we want to avoid the specific details associated with any particular mathematical representation to allow for



**Figure 1: Simplicial representation of the Lotka-Volterra model.** The representation contains higher-order simplices which capture the interactions between the components of the mathematical model.

direct comparison with other models.

To arrive at the simplicial representation in Figure 1 we identify the underlying concepts of Equation 1: the agents are ‘Prey’ and ‘Predator’, corresponding to the population variables  $x$  and  $y$ , respectively; the terms  $ax$  and  $-bxy$  on the right side of the prey equation correspond to ‘Prey growth’ and ‘Predation’, respectively; the terms  $cxy$  and  $-dy$  on the right side of the predator equation correspond to ‘Predator growth’ and ‘Predator death’, respectively; finally, the solutions of Equation 1 exhibit oscillatory dynamics for both populations, which we identify as ‘Oscillatory population’ in each case. The vertices of the simplicial representation are labelled with these concepts. To form the higher-dimensional simplices we identify the interconnections of these model concepts, based on Equation 1, as indicated in Table 1. For example, ‘Prey’ is directly connected in a conceptual

**Table 1:** Conceptual interconnections for Equation 1. A ‘Y’ indicates that there is an interconnection between the two concepts.

	Prey	Prey growth	Predation	Oscillatory population	Predator	Predator growth	Predator death
Prey		Y	Y	Y			
Prey growth	Y		Y	Y			
Predation	Y	Y		Y			
Oscillatory population	Y	Y	Y		Y	Y	Y
Predator			Y	Y		Y	Y
Predator growth			Y	Y	Y		Y
Predator death				Y	Y	Y	



manner with ‘Prey growth’, ‘Predation’, and ‘Oscillatory population’, where each pair of these concepts forms a 1-simplex, each subset of three of these four concepts forms a 2-simplex, and all four of these concepts form a 3-simplex.

Consideration of the conceptual basis of models allows for constructive comparison between models irrespective of their mathematical form. Since we generally need to compare models with different forms, we construct simplicial representations that contain the essential components of the models and their interconnections — components here naturally depend on the scientific problem studied. For instance, we may be interested in comparing a continuum model and a discrete model of the same system in terms of the components of the system that are identified in each model, in order to understand how the models are related. The simplicial representation in Figure 1 is such a conceptual representation of the Lotka-Volterra model. The simplicial representation provides the essential concepts and interactions underlying the model. The vertices correspond to the underlying concepts of the model, the edges link concepts that are interconnected, and the 2- and 3-simplices indicate higher-dimensional interconnections between related concepts.

### 2.3 Homology of simplicial representations

Given a simplicial representation of a model we employ persistent homology to provide a unique and visual characterisation of the model. Note that simplicial homology of the full simplicial complex, while simpler than persistent homology, is not applicable here since it does not in general provide a unique characterisation of a model. Indeed, simplicial homology calculates the number of  $k$ -dimensional holes of the full simplicial complex, for nonnegative integers  $k$ , and the numbers of  $k$ -dimensional holes can be equal for distinctly-labelled complexes. A simple example is given by two simplicial complexes that are each just a 0-simplex, where the 0-simplex has a different label in each complex. Then the two labelled complexes are distinct, however they have the same numbers of  $k$ -dimensional holes since they both consist of a single component, or 0-dimensional hole, and have no higher-dimensional holes. We therefore require the filtration in persistent homology to uniquely characterise distinct simplicial complexes. This is different from the situation typically encountered in topological data analysis.

To compare a collection of simplicial representations of models using persistent homology we require filtrations for each complex that are compatible, so that the same topological structures in different complexes give the same persistence intervals. For this we utilise a reference complex such that the simplicial complexes associated to all of the models for comparison are subcomplexes.

*Notation 2.3 (Reference complex).* Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $m$  be the maximum dimension of the simplicial complexes associated

to the models. Denote by  $\mathcal{R}$  the  $(|\mathcal{C}| - 1)$ -simplex spanned by the labelled 0-simplices  $\{v_1, v_2, \dots, v_{|\mathcal{C}|}\}$ . The *reference complex* is the  $m$ -skeleton of  $\mathcal{R}$ , therefore  $\mathcal{R}^{(m)}$ . Note that each simplicial representation of a model under consideration is a subcomplex of the reference complex.

Therefore, to obtain filtrations for the simplicial complexes that allow for direct comparison of the persistence barcodes we first obtain a filtration for the reference complex  $\mathcal{R}^{(m)}$ . The filtration of each simplicial representation is then an induced filtration of the filtration for  $\mathcal{R}^{(m)}$ . In this way we can directly compare the persistence intervals between different models, since identical persistence intervals will arise from identical topological features associated with the same subset of model components.

Note that we could always use the simplex  $\mathcal{R}$  as the reference complex, however it is computationally efficient to use the  $m$ -skeleton  $\mathcal{R}^{(m)}$  where possible. If we then need to compare an additional model generated by a subset of  $\mathcal{C}$  whose simplicial representation has dimension higher than  $m$ , we can simply extend the filtration of the reference simplex to a filtration of a higher-dimensional skeleton of  $\mathcal{R}$  (see Proposition 2.12).

We want the persistent homology of a particular simplicial complex to be unique amongst all of the simplicial complexes of the models of interest, in order to reflect all differences between the complexes and hence models. We therefore employ a *flat filtration* [32, Chapter 11, Page 83] whereby only one simplex is added at each filtration index, and each simplex is added after all of its proper faces.

**Definition 2.4 (Flat filtration).** Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex. Further, let  $w : \mathcal{R}^{(m)} \rightarrow \{1, 2, \dots, |\mathcal{R}^{(m)}|\}$  be a bijective weight function such that if  $\tau, \sigma \in \mathcal{R}^{(m)}$  and  $\tau$  is a proper face of  $\sigma$  then  $w(\tau) < w(\sigma)$ . The *flat filtration* of  $(\mathcal{R}^{(m)}, w)$  is the nested sequence of subcomplexes  $\{R_i\}_{i=1}^{|\mathcal{R}^{(m)}|}$ , where  $R_i = w^{-1}((-\infty, i])$  for  $1 \leq i \leq |\mathcal{R}^{(m)}|$ .

The flat filtration on  $\mathcal{R}^{(m)}$  induces a filtration on each subcomplex of  $\mathcal{R}^{(m)}$ .

**Definition 2.5 (Induced filtration of a subcomplex of the reference complex).** Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $(\mathcal{R}^{(m)}, w)$  be the associated reference complex with a flat filtration  $\{R_i\}_{i=1}^{|\mathcal{R}^{(m)}|}$ . Further, let  $L$  be a simplicial representation of a model generated by a subset of components from  $\mathcal{C}$ . Then the *induced filtration* for the subcomplex  $L \subseteq \mathcal{R}^{(m)}$  is the nested sequence of subcomplexes  $\{F_i\}_{i \in I}$  where, for each  $i \in I$ ,  $F_i = w|_L^{-1}((-\infty, i]) = R_i \cap L \neq \emptyset$ , and  $i \leq j$  for all  $j$  such that  $1 \leq j \leq |\mathcal{R}^{(m)}|$  and  $R_i \cap L = R_j \cap L$ .

Note that for the induced filtration we exclude filtration indices  $i$  for which  $R_i \cap L = \emptyset$ , and since we may have  $R_i \cap L = R_j \cap L$  for some filtration indices  $i \neq j$  we take the smallest such index as the corresponding filtration index. Note further that the induced filtration adds one simplex at each index of the filtration, which

allows us to use it as a basis for a distance between models (given the set of model components  $\mathcal{C}$ ) as we show below (see Definition 2.15 and Theorem 2.17).

We now establish that the simplicial representations and associated filtrations provide the desired descriptions of the models. The following theorem shows that the simplicial complex associated with a given model has a unique multiset of persistence intervals with respect to the induced filtration. Therefore the persistence intervals, visualised as a persistence barcode, provide a unique ‘fingerprint’ of the model.

**Theorem 2.6.** *Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex. Let  $K$  and  $L$  be two labelled simplicial complexes associated with two models generated by subsets of  $\mathcal{C}$ , and let  $K$  and  $L$  have the induced filtrations from a flat filtration on  $\mathcal{R}^{(m)}$ . Then  $K = L$  if and only if  $\mathcal{P}_k(K) = \mathcal{P}_k(L)$  for all  $k \geq 0$ .*

*Proof.* The forward implication is trivial so consider the backward implication, for which we provide a contrapositive proof. Suppose that  $K \neq L$ , and let  $\sigma^p$  be a simplex of smallest dimension  $p$  in the symmetric difference  $K \triangle L$ . Without loss of generality we may assume that  $\sigma^p \in K$ . We show that there exists  $k \geq 0$  such that  $\mathcal{P}_k(K) \neq \mathcal{P}_k(L)$ .

If  $p = 0$  then  $\mathcal{P}_0(K)$  contains an interval with left endpoint equal to the filtration index  $i$  at which  $\sigma^p$  is added to the filtered complex, corresponding to the creation of a 0-dimensional homology class. Since  $\sigma^p \notin L$  there is no simplex added at index  $i$  in the filtration of  $L$ , hence no interval with left endpoint  $i$  in  $\mathcal{P}_0(L)$ .

For  $p \geq 1$  the boundary of  $\sigma^p$  is a  $(p-1)$ -cycle  $\tau^{p-1}$ , which is a nonbounding cycle in the filtered complex prior to the addition of  $\sigma^p$ . So either  $\tau^{p-1}$  corresponds to a linearly independent  $(p-1)$ -dimensional homology class, or the homology class for  $\tau^{p-1}$  is a linear combination of the  $(p-1)$ -dimensional homology classes. If  $\tau^{p-1}$  is in a linearly independent homology class then  $\mathcal{P}_{p-1}(K)$  contains an interval with right endpoint equal to the index  $j$  of the filtration where the homology class is annihilated by the addition of  $\sigma^p$ . Since  $\sigma^p \notin L$  there is no interval in  $\mathcal{P}_{p-1}(L)$  with the same right endpoint  $j$ . Alternatively, suppose the homology class for  $\tau^{p-1}$  is a linear combination of the  $(p-1)$ -dimensional homology classes. If the addition of  $\sigma^p$  at some index  $i$  in the filtration creates a  $p$ -cycle, and hence a new  $p$ -dimensional homology class, containing  $\sigma^p$  then there is a corresponding interval in  $\mathcal{P}_p(K)$  with left endpoint  $i$  and no interval in  $\mathcal{P}_p(L)$  with left endpoint  $i$ . Otherwise, if the addition of  $\sigma^p$  at some index  $i$  in the filtration does not create a  $p$ -cycle then, since  $\tau^{p-1}$  is now a bounding cycle, it follows that one of the previously linearly independent  $(p-1)$ -dimensional homology classes is now linearly dependent and therefore annihilated at index  $i$ . Therefore, there is a corresponding interval in  $\mathcal{P}_{p-1}(K)$  with right endpoint  $i$ , and no interval in  $\mathcal{P}_{p-1}(L)$  with right endpoint  $i$ .  $\square$

We can rephrase Theorem 2.6 in terms of the full multisets of persistence intervals:

**Corollary 2.7.** *Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex. Let  $K$  and  $L$  be two labelled simplicial complexes associated with two models generated by subsets of  $\mathcal{C}$ , and let  $K$  and  $L$  have the induced filtrations from a flat filtration on the reference simplex  $\mathcal{R}^{(m)}$ . Then  $K = L$  if and only if  $\mathcal{P}(K) = \mathcal{P}(L)$ .*

*Proof.* The forward implication follows from Theorem 2.6. The backward implication will also follow from Theorem 2.6 if we establish that  $\mathcal{P}(K) = \mathcal{P}(L)$  if and only if  $\mathcal{P}_k(K) = \mathcal{P}_k(L)$  for all  $k \geq 0$ . Note that  $\mathcal{F}(K) = \{ \mathcal{P}_k(K) \mid \mathcal{P}_k(K) \neq \emptyset \text{ and } k \geq 0 \}$  is a partition of  $\mathcal{P}(K)$ , and similarly for  $\mathcal{F}(L)$ . So  $(a, b) \in \mathcal{P}(K)$  implies  $(a, b) \in \mathcal{P}_k(K)$  for some  $k \geq 0$ , and if  $(a, b) \in \mathcal{P}(L)$  then  $(a, b) \in \mathcal{P}_j(L)$  for some  $j \geq 0$ . Since both  $K$  and  $L$  are filtered with induced filtrations, whereby at most one simplex is added at each filtration index, the dimension of the simplex added at index  $a$  uniquely determines the dimension of the homology class created at index  $a$  corresponding to the interval  $(a, b)$ . Therefore  $k = j$ , and it follows that  $\mathcal{P}(K) = \mathcal{P}(L)$  is equivalent to  $\mathcal{P}_k(K) = \mathcal{P}_k(L)$  for all  $k \geq 0$ .  $\square$

For a general filtration the multiplicity of each persistence interval may be greater than one. We now show that with the induced filtration, however, the multiplicity of each persistence interval is equal to one, so that the multiset of persistence intervals is a set.

**Proposition 2.8.** *Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex. Let  $K$  be a labelled simplicial complex associated with a model generated by a subset of  $\mathcal{C}$ , and let  $K$  have the induced filtration from a flat filtration on the reference complex  $\mathcal{R}^{(m)}$ . Then every interval in  $\mathcal{P}(K)$  has multiplicity one.*

*Proof.* It suffices to show that no two intervals in  $\mathcal{P}(K)$  have the same left endpoint. So suppose to the contrary that the two intervals  $I_1$  and  $I_2$  in the multiset  $\mathcal{P}(K)$  have the same left endpoint.

Note that  $I_1$  and  $I_2$  do not correspond to homology classes in different dimensions. Indeed, one  $p$ -cycle and one  $q$ -cycle with  $p < q$  cannot be created at the same index as this would require the simultaneous addition of a  $(p - 1)$ -simplex and a  $(q - 1)$ -simplex, respectively, which is not possible with the induced filtration.

So  $I_1$  and  $I_2$  must correspond to linearly-independent homology classes in the same dimension, and therefore to the creation of two linearly-independent nonbounding  $p$ -cycles  $\sigma_1$  and  $\sigma_2$  at the same filtration index  $i$  due to the addition of a  $p$ -simplex  $\tau$ . Then  $\sigma_1 + \sigma_2$  under  $\mathbb{Z}/2\mathbb{Z}$  addition is a nonbounding cycle at index  $i - 1$ , so the addition of  $\tau$  at index  $i$  yields one new homology class, contradicting the creation of two homology classes. Therefore, there is at most one interval in  $\mathcal{P}(K)$  with a particular left endpoint.  $\square$

While we have established that the multiset of persistence intervals is a set, we continue to describe it as a multiset for clarity of reference.

Since a submodel consists of a subset of the components and interconnections of a given model, we expect a similar relationship to hold between the simplicial representations and the corresponding persistent homology, which we now confirm.

**Proposition 2.9.** *Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex. Let  $\mathbf{M}$  and  $\mathbf{N}$  be two models for comparison, where  $\mathbf{M}$  is a submodel of  $\mathbf{N}$ , and let  $K$  and  $L$  be the associated labelled simplicial complexes of  $\mathbf{M}$  and  $\mathbf{N}$ , respectively, generated by subsets of  $\mathcal{C}$ . Let  $K$  and  $L$  have the induced filtrations from a flat filtration on  $\mathcal{R}^{(m)}$ . Then  $K$  is a subcomplex of  $L$ . Further, for each  $k \geq 0$ , the submultiset of finite intervals from  $\mathcal{P}_k(K)$  is a submultiset of  $\mathcal{P}_k(L)$ , and each infinite interval in  $\mathcal{P}_k(K)$  is either in  $\mathcal{P}_k(L)$  or corresponds to a finite interval in  $\mathcal{P}_k(L)$  with the same left endpoint.*

*Proof.* Since  $\mathbf{M}$  is a submodel of  $\mathbf{N}$ , all of the model components and interconnections in  $\mathbf{M}$  are also in  $\mathbf{N}$ , so the corresponding simplicial complexes satisfy  $K \subseteq L$ .

Now, fix  $k \geq 0$  and let  $(a, b)$  be a finite interval in  $\mathcal{P}_k(K)$  corresponding to the creation and subsequent annihilation of a  $k$ -homology class at indices  $a$  and  $b$ , respectively, in the filtration. Since  $K \subseteq L$ , the subcomplex of  $K$  corresponding to the  $k$ -homology class is also a subcomplex of  $L$ , so  $(a, b) \in \mathcal{P}_k(L)$ .

Suppose now that the infinite interval  $(a, \infty)$  is in  $\mathcal{P}_k(K)$ , corresponding to the creation of a  $k$ -homology class at index  $a$  in the filtration which is not annihilated and persists in the full complex  $K$ . Suppose further, without loss of generality, that  $(a, \infty) \notin \mathcal{P}_k(L)$ . Since  $K \subseteq L$ , the subcomplex of  $K$  corresponding to the  $k$ -homology class is also a subcomplex of  $L$ , so the  $k$ -homology class is also created in the filtration of  $L$ . Since  $(a, \infty) \notin \mathcal{P}_k(L)$ ,  $L$  contains the  $(k+1)$ -simplex for which the  $k$ -homology class is the bounding cycle, so the  $k$ -homology class is annihilated at some index  $b$  in the filtration of  $L$  due to the addition of the  $(k+1)$ -simplex. Thus,  $(a, b) \in \mathcal{P}_k(L)$ .  $\square$

We visualise the persistent homology of a simplicial complex, summarised by the multisets of persistence intervals, as a persistence barcode. The barcodes have the distinct advantage of allowing for a visual comparison of models.

**Definition 2.10 (Persistence barcode).** Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex. Let  $K$  be a labelled simplicial complex associated with a model generated by a subset of  $\mathcal{C}$ , and let  $K$  have the induced filtration from a flat filtration on the reference complex  $\mathcal{R}^{(m)}$ . For each dimension  $p \geq 0$  the  $p$ th persistence barcode  $\text{BAR}_p(K)$  for  $K$  is a graphical representation of the multiset of  $p$ -dimensional persistence intervals  $\mathcal{P}_p(K)$ . Specifically, let  $f: \mathcal{P}_p(K) \rightarrow \{1, 2, \dots, |\mathcal{P}_p(K)|\}$  be a bijection, then  $\text{BAR}_p(K)$  is given by the set of

points  $\{ (x, f(a, b)) \mid (a, b) \in \mathcal{P}_p(K) \text{ and } x \in [a, b] \}$ , which is unique up to the bijection specifying the order with respect to the  $y$ -axis.

While two different flat filtrations of the reference complex  $\mathcal{R}^{(m)}$  can produce two different multisets of persistence intervals for a subcomplex  $K$ , the numbers of finite and infinite intervals in each submultiset  $\mathcal{P}_k(K)$  are unchanged.

**Proposition 2.11.** *Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex. Let  $K$  be a labelled simplicial complex associated with a model generated by a subset of components in  $\mathcal{C}$ , let  $\mathcal{F}$  and  $\mathcal{F}'$  be two induced filtrations of  $K$  corresponding to two different flat filtrations of the reference complex  $\mathcal{R}^{(m)}$ , and let  $\mathcal{P}(K)$  and  $\mathcal{P}'(K)$  be the corresponding multisets of persistence intervals. Then, for all  $k \geq 0$ ,  $|\mathcal{P}_k(K)| = |\mathcal{P}'_k(K)|$ , and in particular the numbers of finite and infinite intervals are the same in  $\mathcal{P}_k(K)$  and  $\mathcal{P}'_k(K)$ .*

*Proof.* Fix  $k \geq 0$ . We first establish a bijection  $f: \mathcal{P}_k(K) \rightarrow \mathcal{P}'_k(K)$ . Let  $I \in \mathcal{P}_k(K)$ , with left endpoint  $i$  corresponding to the creation of a nonbounding  $k$ -cycle  $\sigma^k$  at index  $i$  of filtration  $\mathcal{F}$ . Since the  $k$ -cycle is also created at some index  $j$  of filtration  $\mathcal{F}'$ , there exists  $I' \in \mathcal{P}'_k(K)$  with left endpoint  $j$ , so we define  $f(I) = I'$ . The mapping  $f$  is well defined, since each topological feature appears at a unique filtration index.

We need to show that  $f$  is injective and surjective. For injectivity, suppose that  $I_1, I_2 \in \mathcal{P}_k(K)$  with  $f(I_1) = f(I_2)$ . In particular, the intervals  $f(I_1)$  and  $f(I_2)$  have the same left endpoint corresponding to the creation of a nonbounding  $k$ -cycle in the filtration  $\mathcal{F}'$ , and since this nonbounding  $k$ -cycle is created at the left endpoints of both  $I_1$  and  $I_2$  we have  $I_1 = I_2$ . For surjectivity, let  $I' \in \mathcal{P}'_k(K)$ . Then the left endpoint of  $I'$  corresponds to the creation of a nonbounding  $k$ -cycle in the filtration  $\mathcal{F}'$ , and this nonbounding  $k$ -cycle is also created at some index  $i$  of the filtration  $\mathcal{F}$ , so there exists  $I \in \mathcal{P}_k(K)$  with left endpoint  $i$ , so that  $f(I) = I'$ .

Finally, the two induced filtrations correspond to the same simplicial complex  $K$ , so  $\mathcal{P}_k(K)$  and  $\mathcal{P}'_k(K)$  have the same number of infinite intervals, and therefore the same number of finite intervals.  $\square$

We can extend the model comparison to include new models for which not all components are in  $\mathcal{C}$  by appending the new components to  $\mathcal{C}$  to obtain a superset  $\bar{\mathcal{C}}$ , and then extending the ordering function on  $\mathcal{C}$  to  $\bar{\mathcal{C}}$ . This may be necessary as our set of candidate models grows, or if we want to increase the level of detail in which we describe models. This extension method preserves the simplicial representations, and associated persistence intervals, of the models generated by  $\mathcal{C}$ . Notationally, we use an overline to denote concepts in the extended system.

**Proposition 2.12.** *Let  $\mathcal{C}$  and  $\bar{\mathcal{C}}$  be two sets of model components with  $\bar{\mathcal{C}} \supseteq \mathcal{C}$ . Let  $\bar{\text{Ord}}: \bar{\mathcal{C}} \rightarrow \{1, 2, \dots, |\bar{\mathcal{C}}|\}$  be a bijection specifying an order for  $\bar{\mathcal{C}}$  such that  $\bar{\text{Ord}}|_{\mathcal{C}} = \text{Ord}$ . Further, let  $\mathcal{R}^{(m)}$  and  $\bar{\mathcal{R}}^{(m)}$  be the reference complexes associated with  $\mathcal{C}$  and  $\bar{\mathcal{C}}$ , respectively. Then we have the following:*

1. *The reference complexes satisfy  $\mathcal{R}^{(m)} \subseteq \bar{\mathcal{R}}^{(m)}$ .*
2. *Each flat filtration of  $\mathcal{R}^{(m)}$  can be extended to a flat filtration of  $\bar{\mathcal{R}}^{(m)}$ .*
3. *Suppose  $\mathcal{R}^{(m)}$  has a flat filtration, which is extended to a flat filtration on  $\bar{\mathcal{R}}^{(m)}$ . For a simplicial representation  $K$  of a model generated by a subset of components in  $\mathcal{C}$  we have  $\mathcal{P}_k(K) = \bar{\mathcal{P}}_k(K)$  for all  $k \geq 0$ . In other words, the persistent homology for  $K$  corresponding to the induced filtration from the flat filtration on  $\mathcal{R}^{(m)}$  is the same as the persistent homology for  $K$  corresponding to the induced filtration from the extended flat filtration on  $\bar{\mathcal{R}}^{(m)}$ .*

*Proof.* For 1, note that  $\mathcal{R}^{(m)}$  is spanned by a subset of the 0-simplicies that span  $\bar{\mathcal{R}}^{(m)}$ , so  $\mathcal{R}^{(m)}$  is a subcomplex of  $\bar{\mathcal{R}}^{(m)}$ .

For 2, let  $w: \mathcal{R}^{(m)} \rightarrow \{1, 2, \dots, |\mathcal{R}^{(m)}|\}$  be a bijective weight function such that if  $\tau$  is a proper face of  $\sigma$  then  $w(\tau) < w(\sigma)$ , yielding a flat filtration  $\{R_i\}_{i=1}^{|\mathcal{R}^{(m)}|}$  where each  $R_i = w^{-1}((-\infty, i])$ . Extend  $w$  to a bijective weight function  $\bar{w}: \bar{\mathcal{R}}^{(m)} \rightarrow \{1, 2, \dots, |\bar{\mathcal{R}}^{(m)}|\}$  such that if  $\tau$  is a proper face of  $\sigma$  then  $\bar{w}(\tau) < \bar{w}(\sigma)$ . Then the flat filtration  $\{\bar{R}_i\}_{i=1}^{|\bar{\mathcal{R}}^{(m)}|}$  of  $\bar{\mathcal{R}}^{(m)}$ , where  $\bar{R}_i = \bar{w}^{-1}((-\infty, i])$ , extends the flat filtration of  $\mathcal{R}^{(m)}$ , that is,  $R_i = \bar{R}_i$  for  $i \in \{1, 2, \dots, |\mathcal{R}^{(m)}|\}$ .

For 3, let  $K$  be a simplicial representation of a model generated by a subset of components in  $\mathcal{C}$ . Since the flat filtration of  $\mathcal{R}^{(m)}$  and the extended flat filtration of  $\bar{\mathcal{R}}^{(m)}$  are equal over the first  $|\mathcal{R}^{(m)}|$  filtration indices, the corresponding induced filtrations are equal, and it follows that  $\mathcal{P}_k(K) = \bar{\mathcal{P}}_k(K)$  for all  $k \geq 0$ .  $\square$

## 2.4 Model comparison by distance

To compare models of interest given a set of model components  $\mathcal{C}$  in a quantitative manner we introduce a measure of distance between the simplicial representations. We provide a direct measure between the simplicial complexes, along with a measure based on the multisets of persistence intervals, and then show that the two measures yield the same distance.

**Definition 2.13 (Distance between simplicial complexes).** Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison. Let  $\mathcal{S}$  be a collection of labelled simplicial complexes corresponding to models generated by subsets from  $\mathcal{C}$ . Define the function

$$d_{\mathcal{C}}: \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R} \quad \text{such that} \quad d_{\mathcal{C}}(K, L) = |K \Delta L|, \quad (2)$$

so that  $d_{\mathcal{C}}$  gives the cardinality of the symmetric difference of the two labelled simplicial complexes.

Some important remarks regarding the distance function between simplicial complexes:

- The symmetric difference of two labelled simplicial complexes accounts for the labelling of the simplices, so that two labelled simplices from different simplicial complexes are considered identical when they are spanned by the same set of labelled 0-simplices.
- The function  $d_{\mathcal{C}}$  calculates the number of labelled simplices that must be added to or subtracted from one of the simplicial complexes to obtain the other simplicial complex. In particular, this measure of distance applies for any two simplicial representations, whether or not they have labelled simplices in common.
- The function  $d_{\mathcal{C}}$  is invariant with respect to a reordering of the model components in  $\mathcal{C}$ , and therefore a relabelling of the vertices of the reference complex, since the original and relabelled simplicial complexes are isomorphic.
- We label the distance function  $d_{\mathcal{C}}$  with the subscript  $\mathcal{C}$  to emphasise that the distance between two simplicial representations is dependent on the level of conceptual detail employed according to the chosen model components in  $\mathcal{C}$ .
- Since we add simplices under the field  $\mathbb{Z}/2\mathbb{Z}$  we have that  $K \triangle L = K + L$ .

*Notation 2.14.* Denote by  $\overline{\mathbb{R}}$  the set of extended real numbers. For a multiset  $S$  of intervals  $(a, b) \in \mathbb{R} \times \overline{\mathbb{R}}$ , we denote the projection onto the first coordinates as  $\text{proj}_1(S) = \{a \in \mathbb{R} \mid (a, b) \in S \text{ for some } b \in \overline{\mathbb{R}}\}$ , and the projection onto the second coordinates as  $\text{proj}_2(S) = \{b \in \overline{\mathbb{R}} \mid (a, b) \in S \text{ for some } a \in \mathbb{R}\}$ .

*Definition 2.15 (Distance between multisets of persistence intervals).* Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex with a flat filtration. Let  $\mathcal{M}$  be a collection of multisets of persistence intervals  $\mathcal{P}(K)$  corresponding to labelled simplicial complexes  $K$  with labels from  $\mathcal{C}$  and induced filtrations from the flat filtration. Further, let

$$\Theta: \mathcal{M} \rightarrow \mathbb{R} \quad \text{be such that} \quad \Theta(\mathcal{P}(K)) = \text{proj}_1(\mathcal{P}(K)) \cup (\text{proj}_2(\mathcal{P}(K)) \setminus \{+\infty\}), \quad (3)$$

which is the set of all left endpoints and finite right-endpoints of the persistence intervals. Define the function

$$\hat{d}_{\mathcal{C}}: \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R} \quad \text{by} \quad \hat{d}_{\mathcal{C}}(\mathcal{P}(K), \mathcal{P}(L)) = |\Theta(\mathcal{P}(K)) \triangle \Theta(\mathcal{P}(L))|. \quad (4)$$

We now show that  $d_{\mathcal{C}}$  is a distance function on  $\mathcal{S}$ , that  $\hat{d}_{\mathcal{C}}$  is a distance function on  $\mathcal{M}$ , and that the metric



spaces  $(\mathcal{S}, d)$  and  $(\mathcal{M}, \hat{d}_{\mathcal{C}})$  are isometric, so we can use either metric to determine the distances between simplicial complexes. To prove these claims we use the triangle inequality for the symmetric difference: for sets  $X$ ,  $Y$ , and  $Z$ , the cardinality of the symmetric difference satisfies subadditivity, that is,  $|X \Delta Z| \leq |X \Delta Y| + |Y \Delta Z|$ . This relation follows by observing that  $X \Delta Z \subseteq (X \Delta Y) \cup (Y \Delta Z)$ .

We first need a lemma to show that the function  $\hat{d}_{\mathcal{C}}$  is invariant with respect to a change in the flat filtration on the reference complex.

**Lemma 2.16.** *Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let  $\mathcal{R}^{(m)}$  be the associated reference complex. The function  $\hat{d}_{\mathcal{C}}$  is invariant with respect to a change in the flat filtration of the reference complex  $\mathcal{R}^{(m)}$ .*

*Proof.* Let  $\mathcal{P}(K)$  and  $\mathcal{P}(L)$  be the multisets of persistence intervals for  $K$  and  $L$ , respectively, corresponding to the induced filtrations from the first flat filtration with associated weight function  $w$ , and let  $\mathcal{P}'(K)$  and  $\mathcal{P}'(L)$  be the multisets of persistence intervals for  $K$  and  $L$ , respectively, corresponding to the induced filtrations from the second flat filtration with associated weight function  $w'$ .

There exists a permutation  $\pi$  on  $\{1, 2, \dots, |\mathcal{R}^{(m)}|\}$  such that  $\pi \circ w = w'$ , from which it follows that there is a bijection  $f: \Theta(\mathcal{P}(K)) \cup \Theta(\mathcal{P}(L)) \rightarrow \Theta(\mathcal{P}'(K)) \cup \Theta(\mathcal{P}'(L))$  where the restriction of  $f$  to  $\Theta(\mathcal{P}(K))$  is a bijection onto  $\Theta(\mathcal{P}'(K))$  and the restriction of  $f$  to  $\Theta(\mathcal{P}(L))$  is a bijection onto  $\Theta(\mathcal{P}'(L))$ . It follows that there is a bijection between  $\Theta(\mathcal{P}(K)) \Delta \Theta(\mathcal{P}(L))$  and  $\Theta(\mathcal{P}'(K)) \Delta \Theta(\mathcal{P}'(L))$ , so  $\hat{d}_{\mathcal{C}}(\mathcal{P}(K), \mathcal{P}(L)) = \hat{d}_{\mathcal{C}}(\mathcal{P}'(K), \mathcal{P}'(L))$ .  $\square$

**Theorem 2.17.** *Let  $\mathcal{C}$  be the set of all components of models under consideration, and let  $\mathcal{R}^{(m)}$  be the associated reference complex with a flat filtration. Let  $\mathcal{S}$  be a collection of labelled simplicial complexes with labels from  $\mathcal{C}$ , and let  $\mathcal{M}$  be the collection of multisets of persistence intervals  $\mathcal{P}(K)$ , for  $K \in \mathcal{S}$  with the induced filtrations. Then the function  $d_{\mathcal{C}}$  is a metric on  $\mathcal{S}$ , the function  $\hat{d}_{\mathcal{C}}$  is a metric on  $\mathcal{M}$ , and the metric spaces  $(\mathcal{S}, d)$  and  $(\mathcal{M}, \hat{d}_{\mathcal{C}})$  are isometric.*

*Proof.* To see that  $d_{\mathcal{C}}$  is a metric on  $\mathcal{S}$ , let  $K, L \in \mathcal{S}$ . Then  $d_{\mathcal{C}}(K, L) = 0$  if and only if  $|K \Delta L| = 0$  if and only if  $K = L$ , so the identity of indiscernibles holds. Symmetry follows from the symmetric difference. It remains to show subadditivity, so let  $M \in \mathcal{S}$ . By the triangle inequality for the symmetric difference we then have  $d_{\mathcal{C}}(K, L) = |K \Delta L| \leq |K \Delta M| + |M \Delta L| = d_{\mathcal{C}}(K, M) + d_{\mathcal{C}}(M, L)$ , as required.

Next we show that  $\hat{d}_{\mathcal{C}}$  is a metric on  $\mathcal{M}$ , noting that we can consider any particular induced filtration since  $\hat{d}_{\mathcal{C}}$  is invariant with respect to a change in the flat filtration of the reference complex by Lemma 2.16. So let  $\mathcal{P}(K), \mathcal{P}(L) \in \mathcal{M}$  be two multisets of persistence intervals corresponding to the simplicial complexes  $K$  and  $L$ . For the identity of indiscernibles, note that  $\hat{d}_{\mathcal{C}}(\mathcal{P}(K), \mathcal{P}(L)) = 0$  if and only if  $\Theta(\mathcal{P}(K)) = \Theta(\mathcal{P}(L))$ , so that  $K$  and  $L$  contain the same simplices, if and only if  $K = L$  if and only if  $\mathcal{P}(K) = \mathcal{P}(L)$  by Corollary 2.7.

Symmetry follows from the symmetric difference, and subadditivity follows from the triangle inequality for the symmetric difference.

Finally, to show that the two metric spaces  $(\mathcal{S}, d_{\mathcal{C}})$  and  $(\mathcal{M}, \hat{d}_{\mathcal{C}})$  are isometric we need to establish a surjective isometry  $f: \mathcal{S} \rightarrow \mathcal{M}$ . Define  $f$  such that  $K \mapsto \mathcal{P}(K)$ , which is a surjection. Recall that  $f$  is an isometry if and only if  $\hat{d}_{\mathcal{C}}(f(K), f(L)) = d_{\mathcal{C}}(K, L)$  for all  $K, L \in \mathcal{S}$ , or equivalently the cardinality of  $K \triangle L$  equals the cardinality of  $\Theta(\mathcal{P}(K)) \triangle \Theta(\mathcal{P}(L))$  for all  $K, L \in \mathcal{S}$ . To show that  $f$  is an isometry it therefore suffices to establish a bijection  $\phi_{(K,L)}$ , for each  $K, L \in \mathcal{S}$ , from  $K \triangle L$  onto  $\Theta(\mathcal{P}(K)) \triangle \Theta(\mathcal{P}(L))$ . Let  $g: \mathcal{R}^{(m)} \rightarrow \mathbb{R}$  be the flat filtration for the reference complex. Define the function  $\phi_{(K,L)} = g|_{K \triangle L}$  by restriction. Then  $\phi_{(K,L)}$  is a bijection onto its image, since simplices are added individually in the flat filtration. It remains to show that the image of  $K \triangle L$  under  $\phi_{(K,L)}$  is  $\Theta(\mathcal{P}(K)) \triangle \Theta(\mathcal{P}(L))$ . Indeed,  $\sigma \in K \triangle L$  if and only if  $a \in \Theta(\mathcal{P}(K)) \triangle \Theta(\mathcal{P}(L))$  where  $g|_{K \triangle L}(\sigma) = a$ .  $\square$

If we know the relationships between two pairs of simplicial representations with a complex in common, say between  $J$  and  $K$  and between  $K$  and  $L$ , then we can infer the relationship between  $J$  and  $L$ . Specifically, the distance between  $J$  and  $L$  can be determined either from  $J \triangle K$  and  $K \triangle L$  or, if the complexes are all subcomplexes of  $\mathcal{R}^{(m)}$  with a flat filtration, from the multisets of persistence intervals. Formally,

**Proposition 2.18.** *Let  $\mathcal{C}$  be the set of all components of models under consideration, and let  $\mathcal{R}^{(m)}$  be the associated reference complex with a flat filtration. Let  $\mathcal{S}$  be a collection of labelled simplicial complexes corresponding to models generated by subsets from  $\mathcal{C}$ . For  $J, K, L \in \mathcal{S}$  we have*

$$d_{\mathcal{C}}(J, L) = |(J \triangle K) \triangle (K \triangle L)|,$$

and

$$\hat{d}_{\mathcal{C}}(\mathcal{P}(J), \mathcal{P}(L)) = \left| \left( \Theta(\mathcal{P}(J)) \triangle \Theta(\mathcal{P}(K)) \right) \triangle \left( \Theta(\mathcal{P}(K)) \triangle \Theta(\mathcal{P}(L)) \right) \right|.$$

*Proof.* Both equations follow immediately from the observation that the symmetric difference satisfies  $X \triangle Z = (X \triangle Y) \triangle (Y \triangle Z)$  for sets  $X, Y$ , and  $Z$ .  $\square$

## 2.5 Model comparison by equivalence

We now consider model comparison by equivalence, which accounts for particular similarities between models that we may regard as essentially identical. So rather than comparing the simplicial complexes associated to models based on the presence or absence of a particular simplex, as in model comparison by distance, here we consider the equivalence of models in terms of an equivalence of the corresponding labelled simplicial complexes. As we show below ‘equivalence’ as used here is restricted to five narrowly-defined operations.

Moreover, equivalence between two models is only possible (but not guaranteed) for models that have high similarity in terms of the associated simplicial complexes.

For this we need to specify a *predicate*,  $\text{Pr}$ , which is a statement that contains a finite number of variables with the specified domain  $\text{Dom}$ , such that  $\text{Pr}$  becomes a (Boolean) proposition when instantiated. We can consider a predicate to be a Boolean-valued function  $\text{Pr}: \text{Dom} \rightarrow \{\text{true}, \text{false}\}$ . Predicates are often described in terms of the number of their variables, so that for an integer  $n \geq 1$  an  $n$ -place predicate  $\text{Pr}(x_1, \dots, x_n)$  has  $n$  variables  $x_1, \dots, x_n$  with domain  $\text{Dom} \subseteq D_1 \times \dots \times D_n$ , where  $x_i \in D_i$  for  $i = 1, \dots, n$ .

**Definition 2.19 (Equivalent simplicial complexes).** Let  $\mathcal{C}$  be the set of all components of models under consideration, let  $\mathcal{S}$  be a collection of labelled simplicial complexes with labels from  $\mathcal{C}$ , and let  $\text{Pr}$  be a two-place predicate on  $\mathcal{S} \times \mathcal{S}$  such that

$$R_{\text{eq}} = \{ (K, L) \in \mathcal{S} \times \mathcal{S} \mid \text{Pr}(K, L) \} \quad (5)$$

is an equivalence relation on  $\mathcal{S}$ . Then the simplicial complexes  $K, L \in \mathcal{S}$  are *equivalent* if and only if  $(K, L) \in R_{\text{eq}}$ .

To determine the similarity of two models, represented as simplicial complexes, we first need to formulate a predicate on which equivalence is based. There is not a unique predicate for this purpose of comparison, and the established equivalence of models must always be considered with regard to the particular choice of predicate. Note further that equivalence is not a quantitative comparison, but rather a qualitative observation that models have similar features with respect to the specified predicate. Informally, we want to say that  $K, L \in \mathcal{S}$  are equivalent when their labelled simplices are conceptually the same, so in particular we need to identify the labelled subcomplexes in  $K \setminus L$  with those in  $L \setminus K$ . For this we employ five operations on a labelled simplicial complex  $K$ . Since the 1-skeleton of a labelled simplicial complex is a labelled undirected graph, we refer to the 0-simplices and 1-simplices as vertices and edges, respectively.

We first recall the definition of a simplicial map [33, 34]:

**Definition 2.20 (Vertex map, simplicial map).** Let  $K$  and  $L$  be two simplicial complexes. A *vertex map* is a function  $\psi: K^{(0)} \rightarrow L^{(0)}$ , and a *simplicial map*  $\psi: K \rightarrow L$  is a function such that  $\psi|_{K^{(0)}}$  is a vertex map and whenever  $W = \{w_i\}_{i=0}^n$  spans a simplex in  $K$  then  $\psi(W) = \{\psi(w_i)\}_{i=0}^n$  spans a simplex in  $L$ .

We also require the notion of a simplicial multivalued map, which we define as follows:

**Definition 2.21 (Simplicial multivalued map).** Let  $K$  and  $L$  be labelled simplicial complexes with labels from  $\mathcal{C}$ , and let  $F: K^{(0)} \rightrightarrows L^{(0)}$  be a left-total binary relation such that  $F(w)$  is a nonempty subset of  $L^{(0)}$  for each  $w \in K^{(0)}$ . Then  $F$  is a *simplicial multivalued map*, denoted  $F: K \rightrightarrows L$ , if whenever a set of vertices  $W = \{w_i\}_{i=0}^n$  spans a simplex in  $K$  then the set of vertices  $F(W) = \{F(w_i)\}_{i=0}^n$  spans a simplex in  $L$ , and for

each  $i$  we possibly have  $|F(w_i)| > 1$ .

We now define the five operations on simplicial complexes. Recall that two distinct vertices in a simplicial complex are *adjacent* if they belong to the same simplex, and for a vertex  $u$  in a simplicial complex  $K$  we denote the set of all vertices adjacent to  $u$  as  $V_K(u)$ , noting that  $u \notin V_K(u)$ .

**Definition 2.22 (Operation 1: Adjacent-vertex identification).** Let  $K$  and  $L$  be two labelled simplicial complexes with labels from  $\mathcal{C}$ . Let  $\{u, v\}$  be a pair of adjacent vertices in  $K$  such that  $V_K(u) \setminus \{v\} = V_K(v) \setminus \{u\}$ . A simplicial map  $\pi_1: K \rightarrow L$  is an *adjacent-vertex identification* if  $\pi_1$  is surjective, and is injective on every vertex except at the pair of vertices  $\{u, v\}$  that are mapped to a single vertex  $c \in L^{(0)}$ . That is, for  $z \in K^{(0)}$ ,

$$\pi_1(z) = \begin{cases} z & \text{if } z \notin \{u, v\}, \\ c & \text{if } z \in \{u, v\}. \end{cases} \quad (6)$$

**Definition 2.23 (Operation 2: Nonadjacent-vertex identification).** Let  $K$  and  $L$  be two labelled simplicial complexes with labels from  $\mathcal{C}$ . A simplicial map  $\pi_2: K \rightarrow L$  is a *nonadjacent-vertex identification* if  $\pi_2$  is surjective, and is injective on every vertex except at a pair of nonadjacent vertices  $\{u, v\}$  that are mapped to a single vertex  $c \in L^{(0)}$ . That is, for  $z \in K^{(0)}$ ,

$$\pi_2(z) = \begin{cases} z & \text{if } z \notin \{u, v\}, \\ c & \text{if } z \in \{u, v\}. \end{cases} \quad (7)$$

**Definition 2.24 (Operation 3: Vertex split).** Let  $K$  be a labelled simplicial complex with labels from  $\mathcal{C}$ , let  $u \in K^{(0)}$ , and let  $c, d \notin K^{(0)}$  be two vertices with labels in  $\mathcal{C}$ . A *vertex split* is the operation whereby the vertex  $u$  is split into the two new vertices  $c$  and  $d$  to give a new simplicial complex  $L$ , in which  $c$  and  $d$  are adjacent and all of the vertices adjacent to  $u$  are adjacent to both  $c$  and  $d$ . Then  $\sigma \in L$  if and only if there exists  $\tau \in K$  such that either  $u \notin \tau^{(0)}$  and  $\sigma = \tau$ , or  $u \in \tau^{(0)}$  and  $\sigma$  is spanned by  $(\tau^{(0)} \setminus \{u\}) \cup \{c, d\}$ . Vertex splitting is therefore given by the simplicial multivalued map  $\pi_3: K \rightrightarrows L$  where, for  $z \in K^{(0)}$ ,

$$\pi_3(z) = \begin{cases} z & \text{if } z \neq u, \\ \{c, d\} & \text{if } z = u. \end{cases} \quad (8)$$

**Definition 2.25 (Operation 4: Inclusion).** Let  $K$  and  $L$  be two labelled simplicial complexes with labels from  $\mathcal{C}$ . A simplicial map  $\pi_4: K \rightarrow L$  is an *inclusion* if  $\pi_4$  is injective. That is, for  $z \in K^{(0)}$ ,  $\pi_4(z) = z$ .

**Definition 2.26 (Operation 5: Vertex substitution).** Let  $K$  and  $L$  be two labelled simplicial complexes with

labels from  $\mathcal{C}$ . A simplicial map  $\pi_5: K \rightarrow L$  is a *vertex substitution* if  $\pi_5$  is bijective and preserves all labels except for one whereby the labelled vertex  $u \in K^{(0)}$  is mapped to the labelled vertex  $c \in L^{(0)}$ . Therefore, vertex substitution is given by the simplicial map  $\pi_5: K \rightarrow L$  where, for  $z \in K^{(0)}$ ,

$$\pi_5(z) = \begin{cases} z & \text{if } z \neq u, \\ c & \text{if } z = u. \end{cases} \quad (9)$$

Note that, for appropriate simplicial complexes, an adjacent-vertex identification is mutually inverse with a corresponding vertex split, a nonadjacent-vertex identification is mutually inverse with an inclusion, and a vertex substitution is mutually inverse with another vertex substitution.

To rigorously establish an equivalence between models through application of Operations 1–5, we must ensure that each operation preserves the representation of the general physical system. It is important to note that an equivalence between models is based on the components of the models that we allow to be identified as equivalent, and therefore the corresponding operations on the simplicial complexes that we regard as admissible. Thus, in addition to the formal requirements (Operations 1–5) the notion of equivalence also incorporates tight domain-specific constraints. The existence of an equivalence between models is therefore dependent on the level of model detail that we include in the simplicial representations, and the model components that we allow to be equivalent at this level of model detail. This approach provides great flexibility to compare models at various levels of detail, and whether or not models are equivalent or inequivalent will depend on the perspective with which we want to view the models. Model equivalence is not an absolute determination, but rather a perspective that is relative to the operations regarded as admissible.

The application of both Operations 1 and 3 is intrinsically restrictive, since they require similar interconnections between the vertices. Here we explicitly specify the form in which the five operations are admissible:

*Definition 2.27 (Admissible operations on labelled simplicial complexes).* An *admissible operation* on a labelled simplicial complex associated with a model is one of the following:

1. Operation 1 (Adjacent-vertex identification): the two identified vertices and the new vertex must have labels that represent conceptually-equivalent components of the model, which should intrinsically hold for the two identified vertices since they are adjacent so have an intrinsic interconnection.
2. Operation 2 (Nonadjacent-vertex identification): the two identified vertices and the new vertex must have labels that represent conceptually-equivalent components of the model, which do not necessarily hold for the two identified vertices since they are not adjacent.
3. Operation 3 (Vertex split): the original vertex and the two new vertices must have labels that represent

conceptually-equivalent components of the model, which should intrinsically hold for the two new vertices since they are adjacent.

4. Operation 4 (Inclusion): the inclusion already implies that the original simplicial complex is conceptually related to the supercomplex.
5. Operation 5 (Vertex substitution): the new and old vertices must have labels that represent conceptually-equivalent components of the model.

We are now in a position to define the required predicate.

*Definition 2.28* (Predicate for the equivalence of models). The two-place predicate  $\text{Pr}$  on  $\mathcal{S} \times \mathcal{S}$  is as follows: for  $(K, L) \in \mathcal{S} \times \mathcal{S}$  there exists a sequence, which may be empty, of the five admissible operations on simplicial complexes, say  $(f_i)_{i=0}^n$ , such that each  $f_i$  is invertible and  $f_n \circ \cdots \circ f_1 \circ f_0(K) = L$ .

So for  $(K, L) \in \mathcal{S} \times \mathcal{S}$ , the proposition  $\text{Pr}(K, L)$  is true when a sequence of the five admissible operations can transform  $K$  to  $L$  and the corresponding sequence of inverse operations can transform  $L$  to  $K$ , otherwise if no such sequence exists then  $\text{Pr}(K, L)$  is false. The following theorem establishes that the specified predicate gives the necessary equivalence relation.

**Theorem 2.29.** *If  $\text{Pr}$  is the two-place predicate on  $\mathcal{S} \times \mathcal{S}$  given in Definition 2.28 then the relation  $R_{\text{eq}} = \{ (x, y) \in \mathcal{S} \times \mathcal{S} \mid \text{Pr}(x, y) \}$  is an equivalence relation on  $\mathcal{S}$ .*

*Proof.*  $R_{\text{eq}}$  is reflexive, since we can apply the empty sequence of operations to any simplicial complex. Symmetry of  $R_{\text{eq}}$  follows from the required invertibility of each operation, and transitivity follows from the associativity of composition of the operations. Therefore,  $R_{\text{eq}}$  is an equivalence relation on  $\mathcal{S}$ .  $\square$

We reiterate that the model equivalence that we consider here is quite restrictive, so only models that have a high level of similarity would be identified as equivalent. There are many other possible definitions of model equivalence that may be more appropriate when comparing different collections of models. It is important to note that our notion of equivalence of models occurs infrequently, and the demonstration of an equivalence of two models reveals fundamental and nontrivial similarities between the models.

### 3 Results and discussion

We now apply our methodology for model comparison to two different collections of models: first, we examine the model equivalence of three enzyme-catalysed reaction mechanisms involving two substrates; and second, we employ our methodology for model comparison to the two main categories of models for developmental pattern formation.

### 3.1 Comparison of bisubstrate reactions

Here we consider the equivalence of three models for enzyme-catalysed reaction mechanisms involving two substrates: the ordered sequential bisubstrate reaction and the random sequential bisubstrate reaction, which are both ternary-complex mechanisms whereby both substrates bind to the enzyme simultaneously; and, the ping-pong bisubstrate reaction, which involves a chemically-modified intermediate form of the enzyme [35].

#### 3.1.1 Ordered sequential bisubstrate reaction

In an ordered sequential bisubstrate reaction, two substrates A and B first combine with the enzyme E to form a ternary complex EAB, followed by the reaction, and then the release of the products P and Q. This reaction is shown schematically in Figure 2(a), along with the corresponding simplicial representation for our chosen level of conceptual detail. Note that the substrates combine with the enzyme in a particular order, here A followed by B, and the products are released in a particular order, here P then Q, resulting in a single reaction path.

#### 3.1.2 Random sequential bisubstrate reaction

In a random sequential bisubstrate reaction, two substrates A and B first combine with the enzyme E to form a ternary complex EAB, followed by the reaction, and then the release of the products P and Q. This reaction is shown diagrammatically in Figure 2(b), along with the corresponding simplicial representation for our chosen level of conceptual detail. Note that, in this case, there is no required order for substrate combination with the enzyme and for product release, resulting in four possible reaction paths. This simplicial representation consists of four components, from a homological perspective, since it represents four different potential reaction paths.

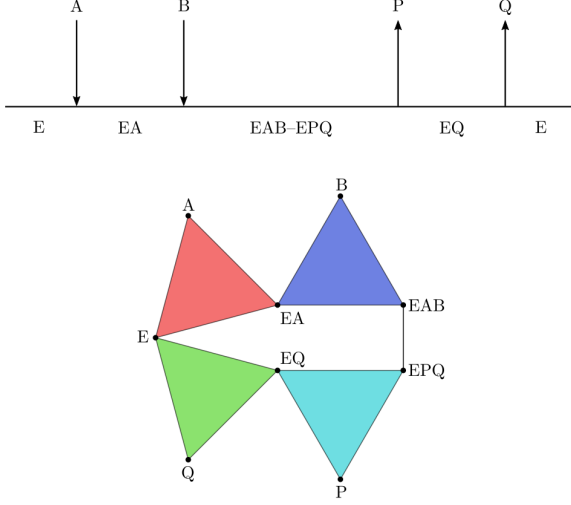
#### 3.1.3 Ping-pong bisubstrate reaction

In a ping-pong bisubstrate reaction, also called a double-displacement reaction, the substrate A combines with the enzyme resulting in the release of product P and the formation of the intermediate  $E^*$ , and then substrate B combines with  $E^*$  resulting in the release of product Q and regeneration of the enzyme  $E$ . This reaction is shown diagrammatically in Figure 2(c), along with the corresponding simplicial representation for our chosen level of conceptual detail. Note that the ping-pong mechanism requires the release of the first product before the second substrate can react.

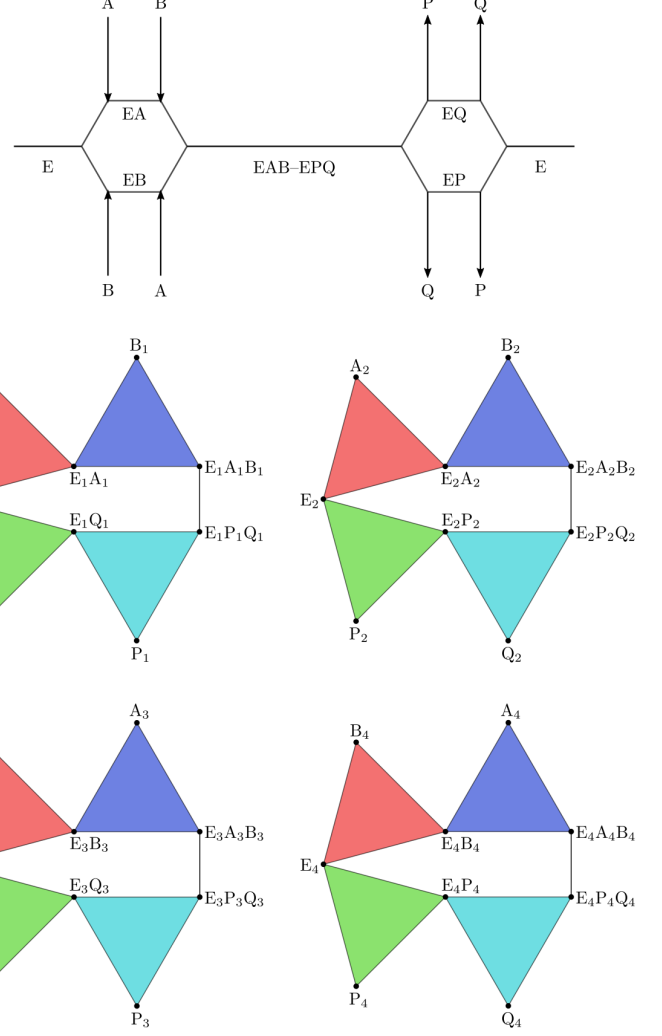
#### 3.1.4 Equivalence of bisubstrate reactions

We first demonstrate an equivalence between the ordered sequential bisubstrate reaction and the random sequential bisubstrate reaction. To transform the complex for the ordered reaction to the complex for the random

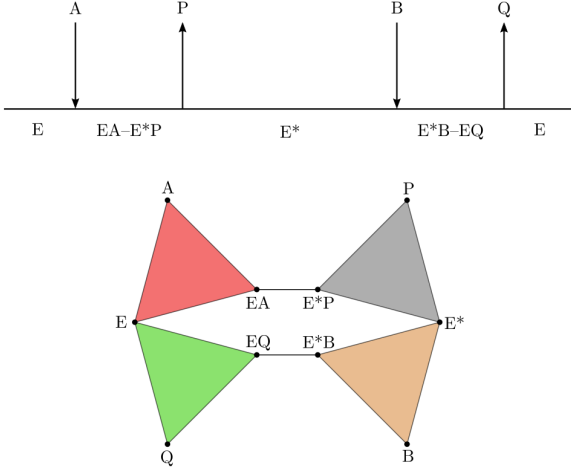
(a) Ordered sequential bisubstrate reaction



(b) Random sequential bisubstrate reaction



(c) Ping-pong bisubstrate reaction



**Figure 2: Bisubstrate reactions represented schematically and as associated labelled simplicial complexes** (a) Ordered sequential bisubstrate reaction; (b) Random sequential bisubstrate reaction; (c) Ping-pong bisubstrate reaction.

reaction, apply vertex substitutions so that  $E, A, EA, B, EAB, EPQ, P, EQ$ , and  $Q$  become the corresponding  $E_1, A_1, E_1A_1, B_1, E_1A_1B_1, E_1P_1Q_1, P_1, E_1Q_1$ , and  $Q_1$ , respectively, and then apply an inclusion operation. Conversely, to transform the complex for the random reaction to the complex for the ordered reaction, apply nonadjacent-vertex identifications so that  $\{E_i\}_{i=1}^4, \{A_i\}_{i=1}^4, \{E_iA_i\}_{i=1}^4, \{B_i\}_{i=1}^4, \{E_iA_iB_i\}_{i=1}^4, \{E_iP_iQ_i\}_{i=1}^4, \{P_i\}_{i=1}^4, \{E_iQ_i\}_{i=1}^4$ , and  $\{Q_i\}_{i=1}^4$  are identified with  $E, A, EA, B, EAB, EPQ, P, EQ$ , and  $Q$ , respectively.

Now we demonstrate that the ordered sequential bisubstrate reaction is not equivalent to the ping-pong bisubstrate reaction. We cannot change the 2-simplices with vertex sets  $\{E, A, EA\}$  and  $\{E, Q, EQ\}$  in each bisubstrate reaction, since they correspond to identical components and interactions. The ping-pong bisubstrate reaction contains a 1-simplex with vertex set  $\{EA, E^*P\}$ , which is not in the ordered sequential bisubstrate



reaction and would require a vertex split operation applied to EA in the ordered reaction. The new vertex from the vertex split operation, labelled E\*P, would however be a vertex in several new simplices, in particular the simplex with vertex set  $\{E, EA, E*P\}$ , which is not in the ping-pong reaction. Therefore, we cannot transform the ordered sequential bisubstrate reaction into the ping-pong bisubstrate reaction, so they are not equivalent.

### 3.2 Comparison of Turing-pattern and positional-information models

Here we apply our methods for model comparison to the two main categories of models for developmental pattern formation, namely Turing-pattern models and positional-information models. We consider the patterning to occur throughout a two-dimensional rectangular domain, where the boundary conditions are always zero-flux on the two opposite sides parallel to the morphogen concentration gradient. We assume that the velocities of the cytoplasm and the growing tissue are negligible, and we therefore assume no advection.

To represent the models as simplicial complexes we first determine the set of components  $\mathcal{C}$  on which the models under consideration are based. We consider five positional-information models and four Turing-pattern models. In this case the general components are: agents, namely morphogens, modulators, and substrates; reactions involving the agents, such as self-activation, activation, inhibition, and annihilation; agent degradation; influx and outflux boundary conditions; agent diffusion; profile of the morphogen gradient; and scale invariance of the morphogen gradient. The ordered set of components is shown in Figure 3. Note that the components can be given in any order since our definition of distance is invariant with respect to changes in the flat filtration (see Lemma 2.16). The distance between models is dependent on the level of conceptual detail used to represent the models, so the distances obtained must be considered with respect to the representative simplicial complexes. The most straightforward way to obtain meaningful comparisons is to use a consistent level of conceptual detail for all models. Here, we take the reference complex as the simplex spanned by the complete set of components for all nine models.

To construct the simplicial representation of each Turing-pattern and positional-information model we first specify the 0-simplices, which represent the model components, and the 1-simplices, which represent direct interconnections between the components. While the 0- and 1-simplices are specified by the model, to give a combinatorial graph, the higher-dimensional simplices are obtained by forming cliques [36], where possible, incrementally in dimensions 2 and higher. These higher-dimensional simplices indicate higher-dimensional interactions between the corresponding model components.

To illustrate the barcode representation, we use a particular flat filtration called the shortlex filtration, which we first describe. Note that while all flat filtrations can be extended (see Proposition 2.12), an extension of the shortlex filtration is not necessarily another shortlex filtration. We begin by applying the shortlex order

<b>Morphogen 1</b>	(15) Diffusion 3	(29) Inhibition of Morphogen 1 by Morphogen 2
(1) Morphogen 1	(16) Degradation 3	(30) Inhibition of Morphogen 1 by Morphogen 3
(2) Diffusion 1	(17) Production 3	(31) Inhibition of Morphogen 3 by Morphogen 1
(3) Degradation 1	<b>Substrate 1</b>	(32) Inhibition of an inhibition by Morphogen 2
(4) Production 1	(18) Substrate 1	(33) Production of Morphogen 2 by Morphogen 1
(5) Basal production 1	(19) Diffusion of Substrate 1	(34) Depletion of Substrate 1 by Morphogen 1
(6) Influx 1	(20) Degradation of Substrate 1	(35) Modulation of Diffusion 1 by Modulator 1
(7) Outflux 1	(21) Basal production of Substrate 1	(36) Modulation of Degradation 1 by Modulator 1
(8) Morphogen 1 bound	<b>Modulator 1</b>	(37) Inhibition of Modulator 1 by Morphogen 1
<b>Morphogen 2</b>	(22) Modulator 1	(38) Adsorption of Morphogen 1
(9) Morphogen 2	(23) Diffusion of Modulator 1	(39) Desorption of Morphogen 1 bound
(10) Diffusion 2	(24) Degradation of Modulator 1	<b>Morphogen 1 gradient</b>
(11) Degradation 2	(25) Production of Modulator 1	(40) Monotonic gradient
(12) Basal production 2	<b>Reactions and Interactions</b>	(41) Oscillatory gradient
(13) Influx 2	(26) Annihilation between Morphogens 1 and 2	(42) Local scale-invariance
<b>Morphogen 3</b>	(27) Self-activation of Morphogen 1	(43) Global scale-invariance
(14) Morphogen 3	(28) Activation of Morphogen 2 by Morphogen 1	

**Figure 3: Ordered set of components for the four Turing-pattern and five positional-information models.**

to the simplices of the simplicial complex with respect to the assigned vertex order, where the shortlex order is defined as follows [37, Chapter 0, page 14]:

*Definition 3.1 (Shortlex order).* Let  $\mathcal{A}$  be a totally-ordered finite set, called the *alphabet*, and let  $\mathcal{W}$  be the set of all *words* that are finite sequences of symbols from  $\mathcal{A}$ . The *shortlex order* on  $\mathcal{W}$  orders the words as follows:

1. Two different words in  $\mathcal{W}$  with equal length are ordered according to the alphabetic order of  $\mathcal{A}$ , therefore lexicographically.
2. For two words with unequal lengths, the shorter word precedes the longer word.

Note that, since  $\mathcal{A}$  is totally-ordered, the shortlex order is also a total order. Applying the shortlex order to a collection of simplices, we first order the simplices by increasing dimension, and then the simplices of the same dimension are ordered lexicographically. Since the set of vertices is totally ordered as specified by Ord,

so is the shortlex order for the simplicies.

We now define the *shortlex filtration*:

**Definition 3.2 (Shortlex filtration).** Let  $\mathcal{C}$  be the set of all components that appear in the collection of models under consideration for comparison, and let the reference complex  $\mathcal{R}^{(m)}$  have the shortlex ordering. Define the bijective weight function  $w: \mathcal{R}^{(m)} \rightarrow \{1, 2, \dots, |\mathcal{R}^{(m)}|\}$  by  $w(\sigma) = j$ , where  $j$  is the index of  $\sigma$  in the shortlex ordering for  $\mathcal{R}^{(m)}$ . The *shortlex filtration* of  $\mathcal{R}^{(m)}$  is the nested sequence of subcomplexes as defined in Definition 2.4.

With this filtration it is possible to define the distances of the models that we have described under the umbrella of the reference complex  $\mathcal{R}^{(m)}$ . Note that, since our definition of distance is invariant with respect to changes in the flat filtration (see Lemma 2.16), the components indicated in Figure 3 can be given an arbitrary ordering, resulting in an alternative shortlex filtration.

Here we describe one positional-information model and one Turing-pattern model, and provide the descriptions of the four additional positional-information models and three additional Turing-pattern models in the Electronic Supplementary Material document. We begin by describing the positional-information annihilation model.

### 3.2.1 Positional-information annihilation model

The annihilation model, and also the opposing gradients model (Electronic Supplementary Material), are mechanisms whereby two opposing morphogen gradients provide size information for developmental patterning. In the annihilation model, the target gene responds to the concentration of Morphogen 1, to which Morphogen 2 irreversibly binds and thereby inhibits the action of Morphogen 1 on activity of transcription, so that the gradient of Morphogen 2 provides size information to the concentration field of Morphogen 1 [38].

The sources of each morphogen are at opposite ends of the domain, and the morphogens interact by an annihilation reaction with rate  $k$  that results in global scale-invariant patterning [38, 39]. Mathematically, the two morphogen gradients with concentrations  $m(\mathbf{x}, \mathbf{t})$  and  $c(\mathbf{x}, \mathbf{t})$  can be modelled as

$$\frac{\partial m}{\partial t} = D_m \nabla^2 m - k_m m - kmc, \quad (10)$$

$$\frac{\partial c}{\partial t} = D_c \nabla^2 c - k_c c - kmc, \quad (11)$$

where  $D_m$  and  $D_c$  are diffusivities, and  $k_m$  and  $k_c$  are degradation rates. The boundary conditions for each morphogen are Neumann at both boundaries, with influx at the source and zero flux at the opposite boundary.

For the simplicial representation of the annihilation model, the vertices and corresponding model components

are:

- $v_1 \longleftrightarrow$  Morphogen 1
- $v_2 \longleftrightarrow$  Diffusion 1
- $v_3 \longleftrightarrow$  Degradation 1
- $v_6 \longleftrightarrow$  Influx 1
- $v_9 \longleftrightarrow$  Morphogen 2
- $v_{10} \longleftrightarrow$  Diffusion 2
- $v_{11} \longleftrightarrow$  Degradation 2
- $v_{13} \longleftrightarrow$  Influx 2
- $v_{26} \longleftrightarrow$  Annihilation between Morphogens 1 and 2
- $v_{40} \longleftrightarrow$  Monotonic gradient
- $v_{43} \longleftrightarrow$  Global scale-invariance

For the annihilation model, the simplicial complex and corresponding persistence barcode is shown in Figure 4(a)–(b). Note that the simplicial complex is 4-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.

We now describe the Turing-pattern activator-inhibitor model.

### 3.2.2 Turing-pattern activator-inhibitor model

The activator-inhibitor system [8, 40, 41] consists of two diffusible morphogens, an autocatalytic activator with concentration  $m(\mathbf{x}, \mathbf{t})$  and a rapidly diffusing inhibitor with concentration  $c(\mathbf{x}, \mathbf{t})$ . This model can be represented mathematically as

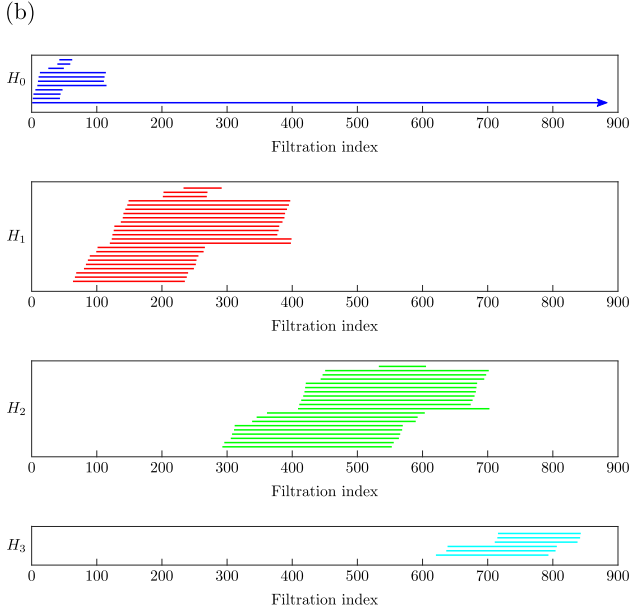
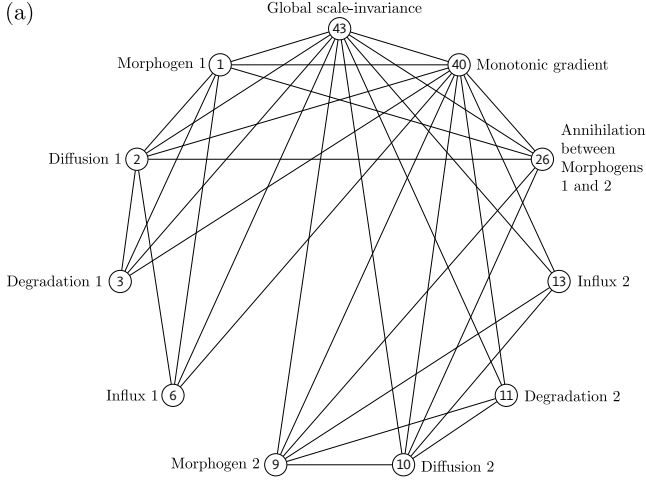
$$\frac{\partial m}{\partial t} = D_m \nabla^2 m + \frac{\rho m^2}{c(1 + \mu_m m^2)} - k_m m + \rho_m, \quad (12)$$

$$\frac{\partial c}{\partial t} = D_c \nabla^2 c + \rho m^2 - k_c c + \rho_c, \quad (13)$$

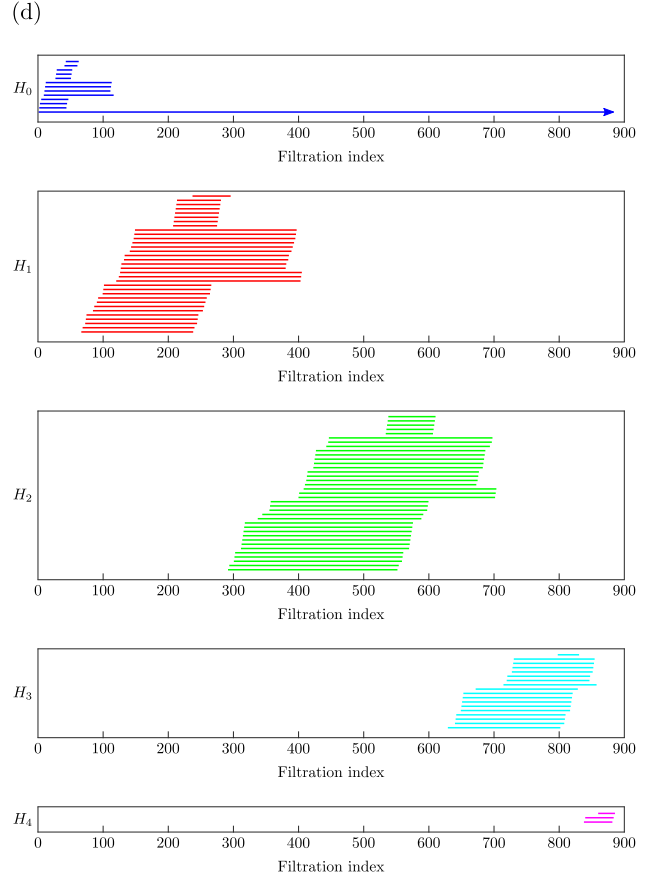
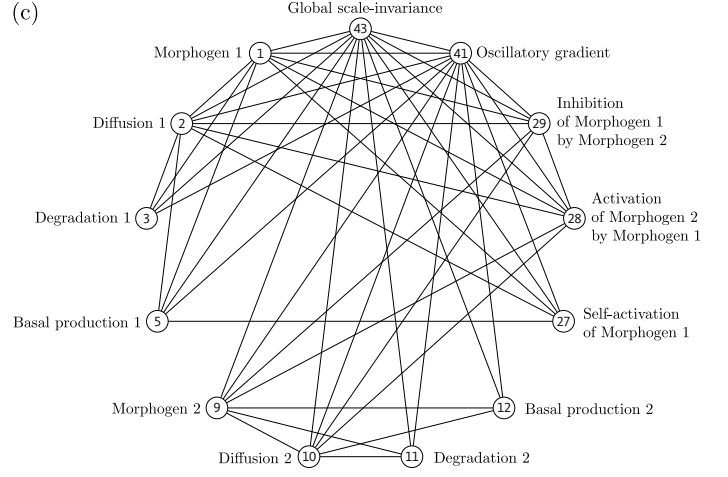
where  $D_m$  and  $D_c$  are diffusivities,  $\rho_m$  and  $\rho_c$  are basal production rates,  $k_m$  and  $k_c$  are degradation rates, and  $\mu_m$  is a saturation constant. The parameter  $\rho$  is the *source density*, which measures the general ability of the cells to perform the autocatalytic reaction. The patterning arises through local self-enhancement of the activator, activation of the inhibitor, and long-range inhibition of the activator. We assume that the boundary conditions are zero flux at both boundaries.

For the simplicial representation of the activator-inhibitor model, the vertices and corresponding model components are:

Positional-information model - annihilation



Turing-pattern model - activator-inhibitor



**Figure 4: Simplicial-complex representations and corresponding persistence barcodes** Positional-information annihilation model: (a) 1-skeleton of the 4-dimensional simplicial representation; (b) Persistence barcode. Turing-pattern activator-inhibitor model: (c) 1-skeleton of the 5-dimensional simplicial representation; (d) Persistence barcode. Note that the  $H_n$  labels on the persistence barcodes refer to the  $n$ -dimensional homology classes, and the arrows at the right endpoints of certain persistence intervals indicate that the right endpoints are  $+\infty$ , corresponding to homology classes that are not annihilated in the filtration.

- $v_1 \longleftrightarrow$  Morphogen 1
- $v_2 \longleftrightarrow$  Diffusion 1
- $v_3 \longleftrightarrow$  Degradation 1
- $v_5 \longleftrightarrow$  Basal production 1
- $v_9 \longleftrightarrow$  Morphogen 2
- $v_{10} \longleftrightarrow$  Diffusion 2
- $v_{11} \longleftrightarrow$  Degradation 2
- $v_{12} \longleftrightarrow$  Basal production 2
- $v_{27} \longleftrightarrow$  Self-activation of Morphogen 1
- $v_{28} \longleftrightarrow$  Activation of Morphogen 2 by Morphogen 1
- $v_{29} \longleftrightarrow$  Inhibition of Morphogen 1 by Morphogen 2
- $v_{41} \longleftrightarrow$  Oscillatory gradient
- $v_{43} \longleftrightarrow$  Global scale-invariance

For the activator-inhibitor model, the simplicial complex and corresponding persistence barcode is shown in Figure 4(c)–(d). Note that the simplicial complex is 5-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.

### 3.2.3 Distances between models

We now apply Theorem 2.17 to calculate the distances between the simplicial representations of the five positional-information and four Turing-pattern models. These distances, shown in Table 2, indicate the total number of labelled simplices that must be added and subtracted to transform one labelled simplicial complex associated with a model to the labelled simplicial complex of another model.

**Table 2:** Distances between the four Turing-pattern (TP) and five positional-information (PI) models

	PI 2	PI 3	PI 4	PI 5	TP 1	TP 2	TP 3	TP 4
PI 1 (linear gradient)	40	108	104	112	240	176	376	340
PI 2 (synthesis-diffusion-degradation)		68	96	104	216	152	352	316
PI 3 (opposing gradients)			120	172	270	220	406	370
PI 4 (annihilation)				152	268	232	404	368
PI 5 (active modulation)					304	240	440	404
TP 1 (activator-inhibitor)						192	432	100
TP 2 (substrate depletion)							424	292
TP 3 (inhibition of an inhibition)								532
TP 4 (modulation)								

In Table 3 we also show the number of simplices in each simplicial representation of the four Turing-pattern (TP) and five positional-information (PI) models.

We observe that the distance between two models can be relatively large even when the models have similar numbers of simplices of each dimension. For example, the Turing-pattern inhibition-of-an-inhibition model (TP 3) and the Turing-pattern modulation model (TP 4) have similar numbers of simplices in each dimension,

**Table 3:** Number of  $n$ -dimensional simplices in the labelled simplicial representations of the four Turing-pattern (TP) and five positional-information (PI) models

$n$	0	1	2	3	4	5
PI 1 (linear gradient)	6	14	16	9	2	0
PI 2 (synthesis-diffusion-degradation)	5	9	7	2	0	0
PI 3 (opposing gradients)	10	27	32	18	4	0
PI 4 (annihilation)	11	33	43	26	6	0
PI 5 (active modulation)	13	39	47	24	4	0
TP 1 (activator-inhibitor)	13	45	70	55	21	3
TP 2 (substrate depletion)	12	37	50	33	10	1
TP 3 (inhibition of an inhibition)	17	70	119	96	36	5
TP 4 (modulation)	16	65	108	84	30	4

however they have the largest distance amongst the four Turing-pattern models that we compare. This illustrates that the metric accounts for not just the numbers of simplices but also for the labelling of the simplices.

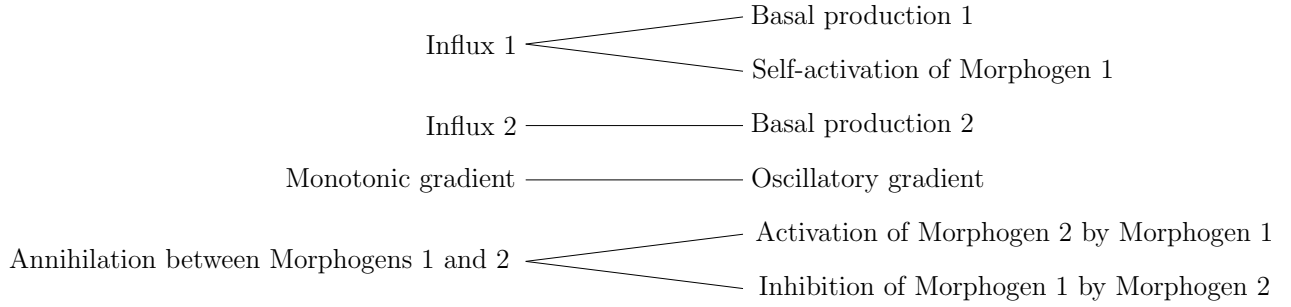
Further, our measure of distance between models is able to reveal when two models are relatively similar, for example the Turing-pattern modulation model (TP 4) is a modification of the Turing-pattern activator-inhibitor (TP 1) model, and these two models have the smallest distance amongst the four Turing-pattern models that we compare.

### 3.2.4 Equivalence of models

We now apply our notion of model equivalence to compare the positional-information annihilation model with the Turing-pattern activator-inhibitor model. The 1-skeletons of the simplicial representations of the annihilation model and the activator-inhibitor model are shown in Figure 4(a) and (c), respectively. We shall demonstrate the equivalence of the two models by applying Operations 1, 3, and 5 to the Turing-pattern activator-inhibitor model to obtain the positional-information annihilation model. Note that, since the three operations are invertible, we could similarly apply operations to the positional-information annihilation model to obtain the Turing-pattern activator-inhibitor model. In accordance with the admissibility requirements for the operations, we regard as equivalent the subsets of model components indicated in Figure 5. We consider ‘Influx 1’ to be equivalent to the combination of ‘Basal production 1’ and ‘Self-activation of Morphogen 1’ since they are all a source of ‘Morphogen 1’. Similarly, ‘Influx 2’ is equivalent to ‘Basal production 2’ since they are both a source of ‘Morphogen 2’. Since we are interested in the existence of the pattern-forming gradient and not the particular profile of the gradient, we consider ‘Monotonic gradient’ and ‘Oscillatory gradient’ as equivalent. Finally, ‘Annihilation between Morphogens 1 and 2’, which reduces the concentrations of both morphogens, is equivalent to the combination of ‘Activation of Morphogen 2 by Morphogen 1’ and ‘Inhibition of Morphogen

## Positional information - annihilation model

## Turing pattern - activator-inhibitor model



**Figure 5: Model components considered as equivalent.**

1 by Morphogen 2'. Indeed, 'Activation of Morphogen 2 by Morphogen 1' and 'Inhibition of Morphogen 1 by Morphogen 2' form an inhibitory cycle whereby 'Morphogen 2' inhibits 'Morphogen 1', which in turn reduces the activation of 'Morphogen 2' by 'Morphogen 1'. Note that these observed equivalences between model components do not suggest that equivalent components are the same, but rather that equivalent components have similar roles in the conceptual frameworks of the corresponding models, which is demonstrated formally through application of the admissible operations on simplicial complexes to ascertain that the models are in the same equivalence class.

We now show that the positional-information annihilation model and the Turing-pattern activator-inhibitor model are in the same equivalence class. While we describe the application of the operations on vertices, there are also changes to the higher-dimensional simplices in the simplicial complex. Note that it is easy to automate the process on computer to transform one simplicial complex to another by application of the operations on simplicial complexes. First consider transforming the complex for the Turing-pattern activator-inhibitor model to the complex for the positional-information annihilation model. Perform an adjacent-vertex identification of the two vertices 'Basal production 1' and 'Self-activation of Morphogen 1' to obtain the new vertex 'Influx 1'. The vertex 'Basal production 2' is substituted with the vertex 'Influx 2'. Similarly, we substitute the vertex 'Monotonic gradient' for the vertex 'Oscillatory gradient'. Finally, we perform an adjacent-vertex identification of the two vertices 'Activation of Morphogen 2 by Morphogen 1' and 'Inhibition of Morphogen 1 by Morphogen 2' to obtain the vertex 'Annihilation between Morphogens 1 and 2'. The resulting simplicial complex is that which represents the positional-information annihilation model.

Conversely, to transform the complex for the positional-information annihilation model to the complex for the Turing-pattern activator-inhibitor model we apply the inverse operations. First perform a vertex split of the vertex 'Annihilation between Morphogens 1 and 2' to obtain the two vertices 'Activation of Morphogen 2 by Morphogen 1' and 'Inhibition of Morphogen 1 by Morphogen 2'. Substitute the vertex 'Oscillatory gradient' for the vertex 'Monotonic gradient', and then substitute the vertex 'Basal production 2' for the vertex 'Influx



2'. Finally, apply a vertex split of the vertex 'Influx 1' to obtain the two vertices 'Basal production 1' and 'Self-activation of Morphogen 1'. The resulting simplicial complex is that which represents the Turing-pattern activator-inhibitor model. Since we have only used admissible and invertible operations to transform the simplicial representation of the Turing-pattern model to the simplicial representation of the positional-information model, the two models are in the same equivalence class, and therefore are equivalent. We can also compare the persistence barcodes shown in Figure 4(b) and (d), which have a high degree of similarity in accordance with the established equivalence between the positional-information annihilation model and the Turing-pattern activator-inhibitor model.

This analysis demonstrates that the Turing-pattern activator-inhibitor model and the positional-information annihilation model, each one of the main models for their patterning mechanism in developmental biology, are in fact closely related in terms of structure and mechanism. This is not obvious given the nature of the proposed models (e.g. [42] and [43]), has not previously been reported, and is unexpected given the frequent historical acrimony between proponents of the two mechanisms [30]. The main difference between the positional-information model and the Turing-pattern model is the source of the gradient-forming morphogen, whereby the positional-information mechanism takes advantage of an influx of morphogen from outside the patterning domain, whereas for the Turing-pattern mechanism the morphogen is produced uniformly within the domain. The morphogen influx naturally produces a morphogen gradient in the positional-information mechanism, so there is no requirement for specialised dynamics. In contrast, the Turing-pattern mechanism requires a Turing instability to generate an oscillating gradient from the spatially-uniform production of the morphogen, requiring very particular conditions for existence.

It is important to note here that very few models would be equivalent with respect to our strict definition of model equivalence. Indeed, with the various model components and their interconnections, along with higher-dimensional interactions, it is unlikely that two simplicial representations of two models could be transformed into each other by application of admissible operations. Our established equivalence of the positional-information annihilation model and the Turing-pattern activator-inhibitor model, with only a few admissible operations required, therefore demonstrates that the two models are conceptually very similar from our considered perspective. This is further substantiated by the ability to employ adjacent-vertex identification operations, or conversely vertex-split operations, which require identical conceptual interconnections for the model concepts involved.

## 4 Conclusion

In this article we present a new methodology for comparing models based on the relationships between various model aspects. Compared to some previous attempts at comparing model structures [42] our approach is readily automatable, and thus able to meet the demands of large-scale modelling attempts and model-curation projects [44].

Representing models as labelled simplicial complexes allows us to determine meaningful distances between models, and persistent homology provides an alternative and often simplified representation of the models. In addition to a measure of distance, we have also developed and applied the concept of equivalence to compare model features. Model equivalence, as developed here, gives more nuanced insights into the relationships between different models and allows us to detect nontrivial similarities between mathematical models. Our analysis of the Turing-pattern activator-inhibitor model and the positional-information annihilation model shows that these two models are more similar than had previously been suggested [30]. Given the conditions outlined in the discussion they are in fact equivalent, something that we had not expected and which to our knowledge has not been seen or considered before. This is one example demonstrating the potential insights that can be gained from our formalism for model comparison.

We foresee more need for this in the immediate future: modelling is increasing in importance in the life and biomedical sciences, yet remains to be fallible and often poorly grounded in reality [45, 46]. The ability to compare, contrast, reconcile, and triage potentially large sets of models will aid in making mathematical modelling more helpful in biology.

While we have applied our methodology to similar types of models, namely systems of reaction-diffusion equations, another important aspect of our formalism is that we can compare different models, irrespective of how distinct they are in form and size: by identifying the model components and their interconnections we can represent any model as a simplicial complex which allows for direct comparison with any other simplicial representation of a model. Identifying differences, but also highlighting similarities in model structures, can aid our understanding of scientific problems. For example, in the particular example of Turing-pattern versus positional-information mechanisms this approach can really resolve long-standing problems and trigger the search for a more synthetic approach [30, 42]. Because of its flexibility and rigorous grounding our methodology is universally applicable to all models.

## Code availability

The code for our model-comparison methodology which calculates the distances between both the simplicial complexes and the persistence intervals is available on GitHub at <https://github.com/DrSeanTVittadello/ModelComparison2> in both Julia and MATLAB languages.

## Author contributions

STV and MPHS conceived and planned this analysis; STV conducted the research, performed the analysis, and drafted the manuscript; all authors reviewed, edited, and approved the final version.

## Funding

The authors gratefully acknowledge funding through a “Life?” programme grant from the Volkswagen Stiftung. MPHS is funded through the University of Melbourne Driving Research Momentum program.

## Acknowledgments

We thank members of the Theoretical Systems Biology group at the University of Melbourne and Imperial College London, Heike Siebert (FU Berlin), James Briscoe (The Francis Crick Institute), and Mark Isalan (Imperial College London) for helpful discussions on Turing patterns and model discrimination.

## References

- [1] Rosenblueth A, Wiener N. The role of models in science. *Philosophy of Science*. 1945;12:316–321. doi:10.1086/286874.
- [2] Tomlin CJ, Axelrod JD. Biology by numbers: mathematical modelling in developmental biology. *Nature Reviews Genetics*. 2007;8:331–340. doi:10.1038/nrg2098.
- [3] Sneddon MW, Faeder JR, Emonet T. Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. *Nature Methods*. 2010;8(2):177–183. doi:10.1038/nmeth.1546.
- [4] Pezzulo G, Levin M. Top-down models in biology: explanation and control of complex living systems above the molecular level. *Journal of The Royal Society Interface*. 2016;13:20160555. doi:10.1098/rsif.2016.0555.
- [5] Transtrum MK, Qiu P. Bridging mechanistic and phenomenological models of complex biological systems. *PLOS Computational Biology*. 2016;12:e1004915. doi:10.1371/journal.pcbi.1004915.
- [6] Turing AM. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*. 1952;237:37–72. doi:10.1098/rstb.1952.0012.
- [7] Wolpert L. Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology*. 1969;25:1–47. doi:10.1016/S0022-5193(69)80016-0.
- [8] Gierer A, Meinhardt H. A theory of biological pattern formation. *Kybernetik*. 1972;12:30–39. doi:10.1007/BF00289234.
- [9] Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M, Fontana W. Rules for modeling signal-transduction systems. *Science Signaling*. 2006;2006:re6. doi:10.1126/stke.3442006re6.
- [10] Danos V, Feret J, Fontana W, Harmer R, Krivine J. Rule-based modelling of cellular signalling. In: Caires L, Vasconcelos VT, editors. *CONCUR 2007 – Concurrency Theory*. Springer Berlin Heidelberg; 2007. p. 17–41. doi:10.1007/978-3-540-74407-8.
- [11] Bachman JA, Sorger P. New approaches to modeling complex biochemistry. *Nature Methods*. 2011;8:130–131. doi:10.1038/nmeth0211-130.
- [12] Wang RS, Saadatpour A, Albert R. Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*. 2012;9:055001. doi:10.1088/1478-3975/9/5/055001.
- [13] Wolkenhauer O. Why model? *Frontiers in Physiology*. 2014;5:21. doi:10.3389/fphys.2014.00021.

- [14] Tavassoly I, Goldfarb J, Iyengar R. Systems biology primer: the basic methods and approaches. *Essays in Biochemistry*. 2018;62:487–500. doi:10.1042/EBC20180003.
- [15] Stalidzans E, Zanin M, Tieri P, Castiglione F, Polster A, Scheiner S, et al. Mechanistic modeling and multi-scale applications for precision medicine: theory and practice. *Network and Systems Medicine*. 2020;3:36–56. doi:10.1089/nsm.2020.0002.
- [16] Le Novère N. Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*. 2015;16:146–158. doi:10.1038/nrg3885.
- [17] Erguler K, Stumpf MPH. Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Molecular BioSystems*. 2011;7(5):1593–1602. doi:10.1039/C0MB00107D.
- [18] Ingram PJ, Stumpf MPH, Stark J. Network motifs: structure does not determine function. *BMC Genomics*. 2006 may;7(1):108. doi:10.1186/1471-2164-7-108.
- [19] Gay S, Soliman S, Fages F. A graphical method for reducing and relating models in systems biology. *Bioinformatics*. 2010;26:i575–i581. doi:10.1093/bioinformatics/btq388.
- [20] Schulz M, Krause F, Novère NL, Klipp E, Liebermeister W. Retrieval, alignment, and clustering of computational models based on semantic annotations. *Molecular Systems Biology*. 2011;7:512. doi:10.1038/msb.2011.41.
- [21] Kirk P, Thorne T, Stumpf MPH. Model selection in systems and synthetic biology. *Current Opinion in Biotechnology*. 2013;24:767–774. doi:10.1016/j.copbio.2013.03.012.
- [22] Clark C, Kalita J. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics*. 2014;30:2351–2359. doi:10.1093/bioinformatics/btu307.
- [23] Henkel R, Hoehndorf R, Kacprowski T, Knüpfer C, Liebermeister W, Waltemath D. Notions of similarity for systems biology models. *Briefings in Bioinformatics*. 2018;19:77–88. doi:10.1093/bib/bbw090.
- [24] Ettinger M. The complexity of comparing reaction systems. *Bioinformatics*. 2002;18:465–469. doi:10.1093/bioinformatics/18.3.465.
- [25] Deza MM, Deza E. Encyclopedia of Distances. 3rd ed. Springer-Verlag New York Inc; 2014.

- [26] Toni T, Jovanovic G, Huvet M, Buck M, Stumpf MPH. From qualitative data to quantitative models: analysis of the phage shock protein stress response in *Escherichia coli*. *BMC Systems Biology*. 2011;5:69. doi:10.1186/1752-0509-5-69.
- [27] Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete & Computational Geometry*. 2002;28:511–533. doi:10.1007/s00454-002-2885-2.
- [28] Zomorodian A, Carlsson G. Computing persistent homology. *Discrete & Computational Geometry*. 2005;33:249–274. doi:10.1007/s00454-004-1146-y.
- [29] Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA. A roadmap for the computation of persistent homology. *EPJ Data Science*. 2017;6:17. doi:10.1140/epjds/s13688-017-0109-5.
- [30] Green JBA, Sharpe J. Positional information and reaction-diffusion: two big ideas in developmental biology combine. *Development*. 2015;142:1203–1211. doi:10.1242/dev.114991.
- [31] Murray JD. *Mathematical Biology*. vol. I: An Introduction. Third edition ed. Springer New York; 2002. doi:10.1007/b98868.
- [32] Edelsbrunner H. *A Short Course in Computational Geometry and Topology*. Springer International Publishing; 2014. doi:10.1007/978-3-319-05957-0.
- [33] Rotman JJ. *An Introduction to Algebraic Topology*. Springer New York; 1988. doi:10.1007/978-1-4612-4576-6.
- [34] Munkres J. *Elements of Algebraic Topology*. CRC Press; 2018. doi:10.1201/9780429493911.
- [35] Cleland WW. The kinetics of enzyme-catalyzed reactions with two or more substrates or products: I. Nomenclature and rate equations. *Biochimica et Biophysica Acta (BBA) - Specialized Section on Enzymological Subjects*. 1963;67:104–137. doi:10.1016/0926-6569(63)90211-6.
- [36] Aktas ME, Akbas E, Fatmaoui AE. Persistence homology of networks: methods and applications. *Applied Network Science*. 2019;4:61. doi:10.1007/s41109-019-0179-3.
- [37] Sipser M. *Introduction to the Theory of Computation*. 3rd ed. Cengage Learning; 2013.
- [38] McHale P, Rappel WJ, Levine H. Embryonic pattern scaling achieved by oppositely directed morphogen gradients. *Physical Biology*. 2006;3:107–120. doi:10.1088/1478-3975/3/2/003.
- [39] Ben-Naim E, Redner S. Inhomogeneous two-species annihilation in the steady state. *Journal of Physics A*. 1992;25:L575–L583. doi:10.1088/0305-4470/25/9/012.

- [40] Meinhardt H. Turing's theory of morphogenesis of 1952 and the subsequent discovery of the crucial role of local self-enhancement and long-range inhibition. *Interface Focus*. 2012;2:407–416. doi:10.1098/rsfs.2011.0097.
- [41] Landge AN, Jordan BM, Diego X, Müller P. Pattern formation mechanisms of self-organizing reaction-diffusion systems. *Developmental Biology*. 2020;460:2–11. doi:10.1016/j.ydbio.2019.10.031.
- [42] Scholes NS, Schnoerr D, Isalan M, Stumpf MPH. A comprehensive network atlas reveals that Turing patterns are common but not robust. *Cell Systems*. 2019;9:243–257. doi:10.1016/j.cels.2019.07.007.
- [43] Schaerli Y, Munteanu A, Gili M, Cotterell J, Sharpe J, Isalan M. A unified design space of synthetic stripe-forming networks. *Nature Communications*. 2014;5:4905. doi:10.1038/ncomms5905.
- [44] Malik-Sheriff RS, Glont M, Nguyen TVN, Tiwari K, Roberts MG, Xavier A, et al. BioModels — 15 years of sharing computational models in life science. *Nucleic Acids Research*. 2019;48:D407–D415. doi:10.1093/nar/gkz1055.
- [45] Gunawardena J. Models in biology: ‘accurate descriptions of our pathetic thinking’. *BMC Biology*. 2014;12:29. doi:10.1186/1741-7007-12-29.
- [46] Stumpf MPH. Multi-model and network inference based on ensemble estimates: avoiding the madness of crowds. *Journal of The Royal Society Interface*. 2020;17:20200419. doi:10.1098/rsif.2020.0419.

Supplementary Material

for:

Model comparison via simplicial complexes and persistent homology

Sean T. Vittadello and Michael P. H. Stumpf



# Contents

<b>1</b>	<b>Summary</b>	<b>2</b>
<b>2</b>	<b>Simplicial representations of additional Turing-pattern and positional-information models</b>	<b>2</b>
2.1	Positional-information models . . . . .	2
2.1.1	Linear gradient . . . . .	2
2.1.2	Synthesis-diffusion-degradation (SDD) . . . . .	3
2.1.3	Opposing gradients . . . . .	3
2.1.4	Scaling by modulation . . . . .	5
2.2	Turing-pattern models . . . . .	7
2.2.1	Substrate depletion . . . . .	7
2.2.2	Inhibition of an inhibition . . . . .	8
2.2.3	Modulation . . . . .	9
<b>3</b>	<b>Persistent homology of simplicial complexes</b>	<b>10</b>
3.1	Simplicies and simplicial complexes . . . . .	10
3.2	Simplicial homology . . . . .	11
3.3	Persistent homology . . . . .	14
<b>4</b>	<b>Systems of reaction-diffusion equations</b>	<b>19</b>

# 1 Summary

In this Supplementary Material document we provide the simplicial representations for the additional Turing-pattern and positional-information models that we discuss in the main document.

To assist the reader of the main document, we also provide a concise discussion of the background in algebraic topology that is relevant to this work, namely simplicial and persistent homology. Further details on simplicial homology are available from standard references for algebraic topology [1–3], and further details on persistent homology can be found in [4–6].

Additionally, we describe the notation for reaction-diffusion equations that we employ in the main document.

## 2 Simplicial representations of additional Turing-pattern and positional-information models

### 2.1 Positional-information models

We consider four additional models of positional information that produce patterning.

#### 2.1.1 Linear gradient

A linear concentration profile of a morphogen results when the production and degradation of the morphogen occur outside of the tissue domain on opposite sides, and the morphogen passively diffuses along the domain from the side where it is produced to the side where it is degraded [7–10]. Mathematically, the steady-state morphogen concentration in this system satisfies Laplace’s equation, which yields global scale-invariant positional information. Specifically, if the tissue length is  $L$  with initial position at  $x = 0$  then the morphogen concentration  $m(x, t)$  can be modelled as

$$\frac{\partial m}{\partial t} = D \frac{\partial^2 m}{\partial x^2}, \tag{S1}$$

where  $D$  is the morphogen diffusivity. The steady-state solution of Equation S1 is the linear equation  $m(x) = ax + b$ , for arbitrary constants  $a, b \in \mathbb{R}$ . The boundary conditions must be influx at one end and outflux at the other end, and we may assume without loss of generality that influx occurs at  $x = 0$  and outflux at  $x = L$ . The original Dirichlet boundary conditions specify the constant concentrations  $m(0, t) = m_0 > 0$  and  $m(L, t) = m_L \geq 0$ , with  $m_0 > m_L$ , so the solution is  $m(x) = ((m_L - m_0)/L)x + m_0$ . Monotonically decreasing linear gradients can also be achieved with Neumann boundary conditions at one boundary and Dirichlet boundary conditions at the opposite boundary.

For the simplicial representation of the linear model, the vertices and corresponding model components are:

- $v_1 \longleftrightarrow$  Morphogen 1
- $v_2 \longleftrightarrow$  Diffusion 1
- $v_6 \longleftrightarrow$  Influx 1
- $v_7 \longleftrightarrow$  Outflux 1
- $v_{40} \longleftrightarrow$  Monotonic gradient
- $v_{43} \longleftrightarrow$  Global scale-invariance

The simplicial representation should capture the interconnections of the model components that result in the formation of the steady-state morphogen gradient, so for the linear model the simplicial complex is shown in Figure S1(a). Note that the simplicial complex is 4-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.

### 2.1.2 Synthesis-diffusion-degradation (SDD)

In the SDD model, a morphogen gradient forms by morphogen production from a localised source at the boundary combined with morphogen diffusion and uniform degradation throughout the tissue [10–12]. Mathematically, the morphogen concentration  $m(\mathbf{x}, t)$  can be modelled as

$$\frac{\partial m}{\partial t} = D\nabla^2 m - km, \quad (\text{S2})$$

where  $D$  is the morphogen diffusivity, and  $k$  is the morphogen degradation rate which is independent of morphogen concentration. The boundary conditions are Neumann for the influx and zero outflux. The steady-state morphogen gradient is an exponential function, so this system is not scale invariant unless  $D$ ,  $k$ , and the influx vary with the tissue length  $L$  in a very specific manner [12].

For the simplicial representation of the SDD model, the vertices and corresponding model components are:

- $v_1 \longleftrightarrow$  Morphogen 1
- $v_2 \longleftrightarrow$  Diffusion 1
- $v_3 \longleftrightarrow$  Degradation 1
- $v_6 \longleftrightarrow$  Influx 1
- $v_{40} \longleftrightarrow$  Monotonic gradient

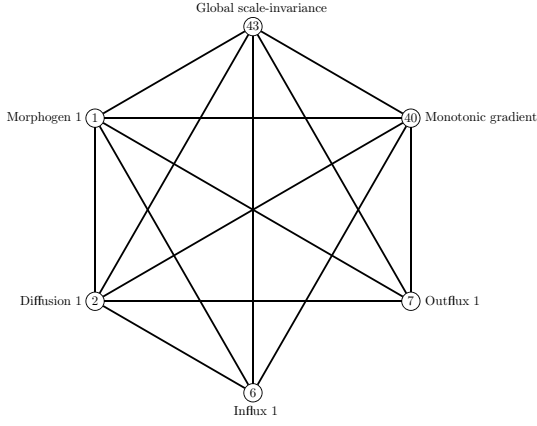
For the SDD model the simplicial complex is shown in Figure S1(b). Note that the simplicial complex is 3-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.

### 2.1.3 Opposing gradients

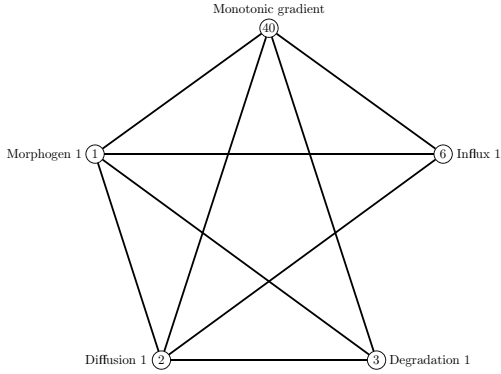
There are two mechanisms whereby two opposing morphogen gradients provide size information for developmental patterning [13]. One is the annihilation model, which is discussed in the main document. The other is the scaling by opposing gradients model, whereby the gene expression depends on the relative concentrations of the two morphogens.

For the opposing gradients model the sources of each morphogen are at opposite ends of the domain. The combination of the two morphogen gradients provides effective local scaling for the boundaries of individual target

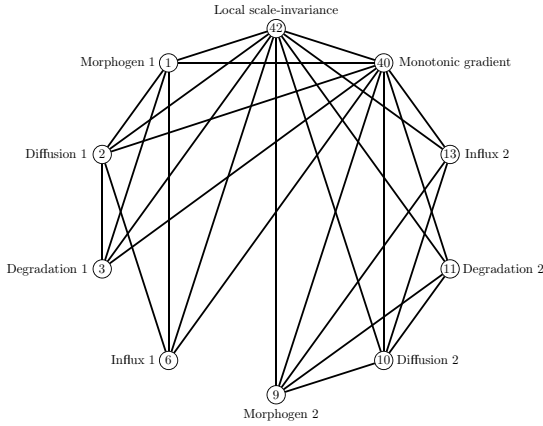
(a) Positional-information model - linear gradient



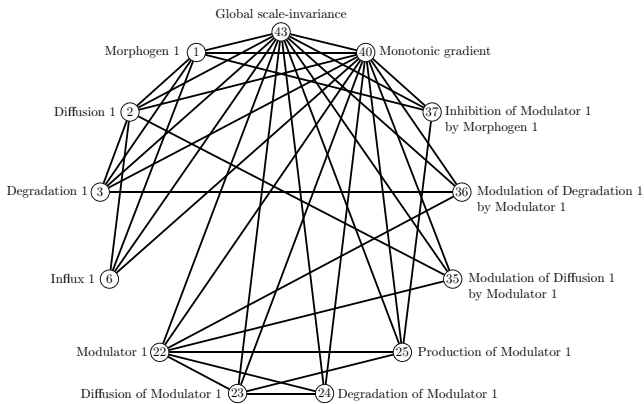
(b) Positional-information model - synthesis-diffusion-degradation



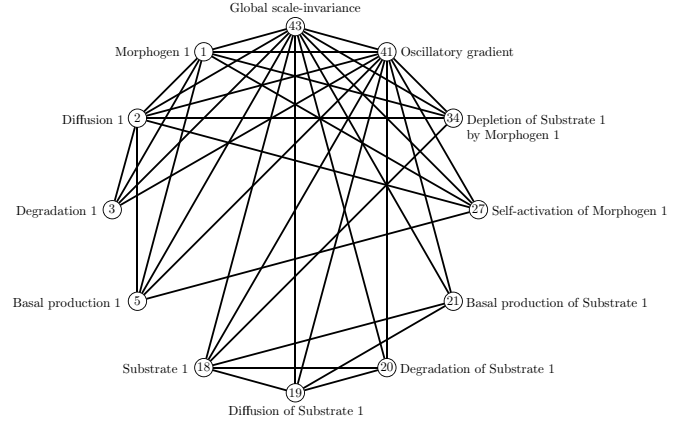
(c) Positional-information model - opposing gradients



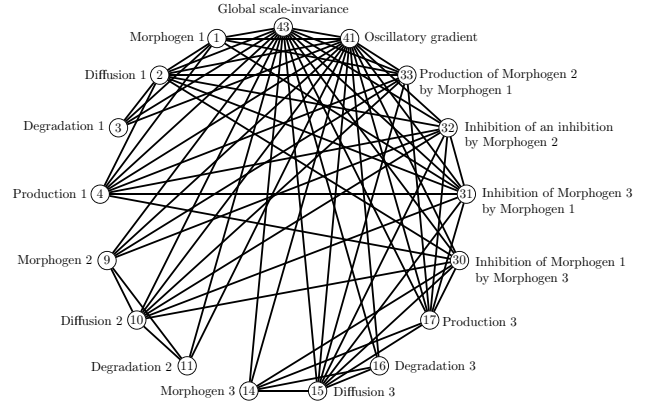
(d) Positional-information model - modulation



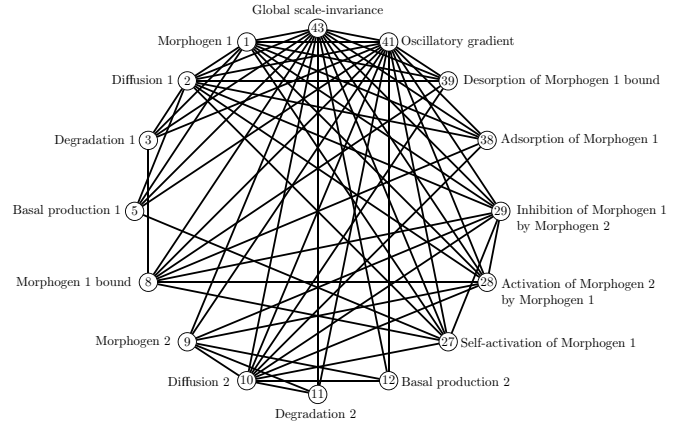
(e) Turing model - substrate depletion



(f) Turing model - inhibition of an inhibition



(g) Turing model - modulation



**Figure S1: 1-skeletons of the simplicial-complex representations for additional positional-information and Turing-pattern models.** Positional-information models: (a) Linear gradient; (b) Synthesis-diffusion-degradation; (c) Opposing gradients; (d) Modulation. Turing-pattern models: (e) Substrate depletion; (f) Inhibition of an inhibition; (g) Modulation.

genes, however it does not produce global scaling across the entire domain [8, 10, 13, 14]. The degree of scaling for the opposing gradients mechanism can be increased if the morphogens are irreversibly inactivated upon binding to each other [13]. Here we assume no direct interaction between the two morphogens. Mathematically, the two morphogen gradients with concentrations  $m(\mathbf{x}, t)$  and  $c(\mathbf{x}, t)$  can be modelled as

$$\frac{\partial m}{\partial t} = D_m \nabla^2 m - k_m m, \quad (\text{S3})$$

$$\frac{\partial c}{\partial t} = D_c \nabla^2 c - k_c c, \quad (\text{S4})$$

where  $D_m$  and  $D_c$  are diffusivities, and  $k_m$  and  $k_c$  are degradation rates. The boundary conditions for each morphogen are Neumann at both boundaries, with flux at the source and zero flux at the opposite boundary.

For the simplicial representation of the opposing gradients model, the vertices and corresponding model components are:

- $v_1 \longleftrightarrow$  Morphogen 1
- $v_2 \longleftrightarrow$  Diffusion 1
- $v_3 \longleftrightarrow$  Degradation 1
- $v_6 \longleftrightarrow$  Influx 1
- $v_9 \longleftrightarrow$  Morphogen 2
- $v_{10} \longleftrightarrow$  Diffusion 2
- $v_{11} \longleftrightarrow$  Degradation 2
- $v_{13} \longleftrightarrow$  Influx 2
- $v_{40} \longleftrightarrow$  Monotonic gradient
- $v_{42} \longleftrightarrow$  Local scale-invariance

For the opposing gradients model the simplicial complex is shown in Figure S1(c). Note that the simplicial complex is 4-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.

### 2.1.4 Scaling by modulation

The largest degree of scaling across the patterning domain is obtained when the biophysical properties of the morphogen are influenced by an accessory modulator that senses the size of the domain [10–12, 15]. The modulator molecules with concentration  $c(\mathbf{x}, t)$  may influence the diffusivity, degradation rate, or influx rate, of the morphogen with concentration  $m(\mathbf{x}, t)$ . If the kinetics and source of the modulator are independent of the morphogen concentration then the mechanism is *passive modulation*. Otherwise the mechanism is *active modulation*, and the distribution of the modulator is dependent on feedback from the morphogen. An example of active modulation is the *expansion-repression* mechanism, whereby the modulator expands the range of the morphogen gradient through an increase in the morphogen diffusivity or a decrease in the morphogen degradation rate, while the morphogen represses/inhibits production of the modulator [15]. The expander molecule, here the modulator, needs to have a high diffusivity in order to rapidly equilibrate across the domain and therefore provide information about domain size. Therefore, this

mechanism establishes global scaling by increasing morphogen range in larger domains and decreasing morphogen range in smaller domains. Correspondingly, the *induction-contraction* model also scales globally [16,17]. In this case, the morphogen induces the production of a contractor molecule, here the modulator with high diffusivity once again, which contracts the range of the morphogen gradient through a decrease in the morphogen diffusivity or an increase of the morphogen degradation rate. Note that while the amplitude and shape of the morphogen gradient is globally scale-invariant, the same is not necessarily true of the modulator since the modulator level reflects the domain size and therefore increases or decreases accordingly.

Here we consider the induction-contraction mechanism, which can be represented mathematically as

$$\frac{\partial m}{\partial t} = D_m(c)\nabla^2 m - k_m(c)m, \quad (\text{S5})$$

$$\frac{\partial c}{\partial t} = D_c\nabla^2 c - k_c c + \rho(m), \quad (\text{S6})$$

where  $D_m(c)$  and  $D_c$  are diffusivities,  $k_m(c)$  and  $k_c$  are degradation rates, and  $\rho(m)$  is the localised production source for the modulator. Note that  $D_m(c)$  is a decreasing function of  $c$ ,  $k_m(c)$  is an increasing function of  $c$ , and  $\rho(m)$  is an increasing function of  $m$ . Further note that ‘degradation’ in this context refers to not only physical destruction, but to any mechanism that effects the removal of the morphogen from the patterning system, including irreversible complex formation. Morphogen boundary conditions are Neumann for the influx and zero flux at the opposite end of the domain. For the modulator there are various possibilities for the boundary conditions, such as the same type of boundary conditions as the morphogen when the modulator source is outside the domain or, as we consider here, zero flux at both boundaries when the modulator source is within the domain.

For the simplicial representation of the induction-contraction modulation model, the vertices and corresponding model components are:

- $v_1 \longleftrightarrow$  Morphogen 1
- $v_2 \longleftrightarrow$  Diffusion 1
- $v_3 \longleftrightarrow$  Degradation 1
- $v_6 \longleftrightarrow$  Influx 1
- $v_{22} \longleftrightarrow$  Modulator 1
- $v_{23} \longleftrightarrow$  Diffusion of modulator 1
- $v_{24} \longleftrightarrow$  Degradation of modulator 1
- $v_{25} \longleftrightarrow$  Production of modulator 1
- $v_{35} \longleftrightarrow$  Modulation of diffusion 1 by modulator 1
- $v_{36} \longleftrightarrow$  Modulation of degradation 1 by modulator 1
- $v_{37} \longleftrightarrow$  Inhibition of modulator 1 by morphogen 1
- $v_{40} \longleftrightarrow$  Monotonic gradient
- $v_{43} \longleftrightarrow$  Global scale-invariance

For the induction-contraction modulation model the simplicial complex is shown in Figure S1(d). Note that the simplicial complex is 4-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.

## 2.2 Turing-pattern models

We consider three further Turing-pattern models that produce patterning. Note that the boundary conditions may be Dirichlet, Neumann, Robin, or periodic [18–20].

### 2.2.1 Substrate depletion

While long-range inhibition of an autocatalytic activator can occur with an inhibitor, as in the activator-inhibitor model, the antagonistic effect can also arise from the depletion of a substrate that is consumed during activator production. In this system, the production of an autocatalytic activator with concentration  $m(\mathbf{x}, t)$  results in either direct or indirect depletion of a substrate with concentration  $s(\mathbf{x}, t)$  [21]. This model can be represented mathematically as

$$\frac{\partial m}{\partial t} = D_m \nabla^2 m + \rho s m^2 - k_m m + \rho_m, \quad (\text{S7})$$

$$\frac{\partial s}{\partial t} = D_s \nabla^2 s - \rho s m^2 - k_s s + \rho_s, \quad (\text{S8})$$

where  $D_m$  and  $D_s$  are diffusivities,  $\rho_m$  and  $\rho_s$  are basal production rates,  $k_m$  and  $k_s$  are degradation rates, and  $\rho$  is the source density for the autocatalytic reaction of the activator, similar to the activator-inhibitor model. The diffusivity  $D_s$  of the substrate must be much faster than the diffusivity  $D_m$  of the activator, and it is assumed that the substrate is produced uniformly throughout the domain. We assume that the boundary conditions are zero flux at both boundaries.

For the simplicial representation of the substrate depletion model, the vertices and corresponding model components are:

- $v_1 \longleftrightarrow$  Morphogen 1
- $v_2 \longleftrightarrow$  Diffusion 1
- $v_3 \longleftrightarrow$  Degradation 1
- $v_5 \longleftrightarrow$  Basal production 1
- $v_{18} \longleftrightarrow$  Substrate 1
- $v_{19} \longleftrightarrow$  Diffusion of substrate 1
- $v_{20} \longleftrightarrow$  Degradation of substrate 1
- $v_{21} \longleftrightarrow$  Basal production of substrate 1
- $v_{27} \longleftrightarrow$  Self-activation of morphogen 1
- $v_{34} \longleftrightarrow$  Depletion of substrate 1 by morphogen 1
- $v_{41} \longleftrightarrow$  Oscillatory gradient
- $v_{43} \longleftrightarrow$  Global scale-invariance

For the substrate depletion model the simplicial complex is shown in Figure S1(e). Note that the simplicial complex is 5-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.

### 2.2.2 Inhibition of an inhibition

Dynamics analogous to the activator-inhibitor model can be realised through an inhibition of an inhibition mechanism [21]. In this case, two morphogens with concentrations  $a(\mathbf{x}, t)$  and  $c(\mathbf{x}, t)$  inhibit the production of each other, thereby forming a switching system in which one of the morphogens becomes fully activated similar to being autocatalytic. In order that pattern formation occur, a third morphogen with concentration  $b(\mathbf{x}, t)$  acts as a long-range signal that disrupts the indirect self-enhancement of either  $a$  or  $c$ . Morphogen  $b$  is rapidly diffusing, is produced under control of  $a$ , and inhibits the inhibition of  $c$  production by  $a$ . therefore acting as inhibitor. This model can be represented mathematically as

$$\frac{\partial a}{\partial t} = D_a \nabla^2 a + \frac{\rho_a}{\kappa_a + c^2} - k_a a, \quad (\text{S9})$$

$$\frac{\partial b}{\partial t} = D_b \nabla^2 b + k_b(a - b), \quad (\text{S10})$$

$$\frac{\partial c}{\partial t} = D_c \nabla^2 c + \frac{\rho_c}{\kappa_c + a^2/b^2} - k_c c, \quad (\text{S11})$$

where  $D_a$ ,  $D_b$ , and  $D_c$  are diffusivities,  $\rho_a$  and  $\rho_c$  are production rates,  $k_a$  and  $k_c$  are degradation rates,  $k_b$  is both the rate of production and degradation of  $b$ , and  $\kappa_a$  and  $\kappa_c$  are saturation constants that limit the production rate if the concentrations of  $a$  or  $c$  become too low. We assume that the boundary conditions are zero flux at both boundaries.

For the simplicial representation of the inhibition of an inhibition model, the vertices and corresponding model components are:

- |  |   |
|--|---|
| • $v_1 \longleftrightarrow$ Morphogen 1      | • $v_{16} \longleftrightarrow$ Degradation 3                              |
| • $v_2 \longleftrightarrow$ Diffusion 1      | • $v_{17} \longleftrightarrow$ Production 3                               |
| • $v_3 \longleftrightarrow$ Degradation 1    | • $v_{30} \longleftrightarrow$ Inhibition of morphogen 1 by morphogen 3   |
| • $v_4 \longleftrightarrow$ Production 1     | • $v_{31} \longleftrightarrow$ Inhibition of morphogen 3 by morphogen 1   |
| • $v_9 \longleftrightarrow$ Morphogen 2      | • $v_{32} \longleftrightarrow$ Inhibition of an inhibition by morphogen 2 |
| • $v_{10} \longleftrightarrow$ Diffusion 2   | • $v_{33} \longleftrightarrow$ Production of morphogen 2 by morphogen 1   |
| • $v_{11} \longleftrightarrow$ Degradation 2 | • $v_{41} \longleftrightarrow$ Oscillatory gradient                       |
| • $v_{14} \longleftrightarrow$ Morphogen 3   | • $v_{43} \longleftrightarrow$ Global scale-invariance                    |
| • $v_{15} \longleftrightarrow$ Diffusion 3   |   |

For the inhibition of an inhibition model the simplicial complex is shown in Figure S1(f). Note that the simplicial complex is 5-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.



### 2.2.3 Modulation

The diffusion of a morphogen may be inhibited by adsorption on negatively-charged extracellular matrix (ECM) components, resulting in modulated diffusion and a smaller effective diffusivity for the morphogen [22]. The corresponding modulation model [22] is an extension of the activator-inhibitor model [21, 23, 24]. Our description of the modulation model is based on the version of the activator-inhibitor model described above. The modulation model consists of two diffusible morphogens, an autocatalytic activator and an inhibitor, along with available binding sites on the ECM onto which the activator adsorbs. The inhibitor does not bind to the ECM, and the free activator and the inhibitor have equal diffusivities  $D$ . It is assumed that autocatalysis and activation of the inhibitor by the activator occurs in both the free and adsorbed states. It is also assumed that the degradation rate of the activator is equal in both the free and adsorbed states. Free binding sites on the ECM therefore appear due to both desorption and degradation of the activator. Such modulation allows for stable dissipative morphogens gradients. Mathematically, the free activator has concentration  $a(\mathbf{x}, t)$ , the bound activator has concentration  $b(\mathbf{x}, t)$ , the inhibitor has concentration  $c(\mathbf{x}, t)$ , and the concentration of available binding sites on the ECM is  $s(\mathbf{x}, t)$ :

$$\frac{\partial a}{\partial t} = D\nabla^2 a + \frac{\rho(a+b)^2}{c(1+\mu_a(a+b)^2)} - k_a a + \rho_a - k_1 s a + k_{-1} b, \quad (\text{S12})$$

$$\frac{\partial c}{\partial t} = D\nabla^2 c + \rho(a+b)^2 - k_c c + \rho_c, \quad (\text{S13})$$

$$\frac{\partial s}{\partial t} = -k_1 s a + (k_{-1} + k_a) b, \quad (\text{S14})$$

where  $D$  is the diffusivity,  $\rho_a$  and  $\rho_c$  are basal production rates,  $k_a$  and  $k_c$  are degradation rates,  $\mu_a$  is the saturation constant,  $\rho$  is the source density,  $k_1$  is the rate of adsorption of the activator onto the ECM, and  $k_{-1}$  is the rate of desorption of the activator from the ECM. We assume that the boundary conditions are zero flux at both boundaries.

For the simplicial representation of the modulation model, the vertices and corresponding model components are:

- $v_1 \longleftrightarrow$  Morphogen 1
- $v_2 \longleftrightarrow$  Diffusion 1
- $v_3 \longleftrightarrow$  Degradation 1
- $v_5 \longleftrightarrow$  Basal production 1
- $v_8 \longleftrightarrow$  Morphogen 1 bound
- $v_9 \longleftrightarrow$  Morphogen 2
- $v_{10} \longleftrightarrow$  Diffusion 2
- $v_{11} \longleftrightarrow$  Degradation 2
- $v_{12} \longleftrightarrow$  Basal production 2
- $v_{27} \longleftrightarrow$  Self-activation of morphogen 1
- $v_{28} \longleftrightarrow$  Activation of morphogen 2 by morphogen 1
- $v_{29} \longleftrightarrow$  Inhibition of morphogen 1 by morphogen 2
- $v_{38} \longleftrightarrow$  Adsorption of morphogen 1
- $v_{39} \longleftrightarrow$  Desorption of bound morphogen 1
- $v_{41} \longleftrightarrow$  Oscillatory gradient
- $v_{43} \longleftrightarrow$  Global scale-invariance

For the modulation model the simplicial complex is shown in Figure S1(g). Note that the simplicial complex is 5-dimensional, however we only show the 1-skeleton of the simplicial complex for simplicity.

### 3 Persistent homology of simplicial complexes

#### 3.1 Simplicies and simplicial complexes

Homology associates algebraic invariants, in particular abelian groups, to a topological space. These invariants determine the number of connected components, holes, and voids in the space. This homological characterisation therefore provides a method for distinguishing between topological spaces. There are various types of homology theory, however simplicial homology often allows for easier calculation of the algebraic invariants in applications, and agrees with more general theories such as singular homology on spaces that can be triangulated. Note that a topological space and its triangulation, a simplicial complex, are homeomorphic, so they have the same homology. Triangulations of data sets provide common examples of simplicial complexes [25].

We can view simplicial complexes as either combinatorial objects or as geometric objects, which we now describe.

**Definition 3.1 (Abstract simplicial complex).** An *abstract simplicial complex* is a set  $V$  along with a collection  $K$  of finite nonempty subsets of  $V$  such that:

1.  $\{v\} \in K$  for all  $v \in V$ .
2. If  $\sigma \in K$  and  $\tau$  is a nonempty subset of  $\sigma$  then  $\tau \in K$ .

The set  $V$  is the *vertex set* of the simplicial complex  $K$ , and the vertex  $v \in V$  is identified with  $\{v\} \in K$ . Each element  $\sigma \in K$  is called a *simplex* of  $K$ , and the *dimension* of  $\sigma$  is the nonnegative integer given by  $\dim(\sigma) = |\sigma| - 1$ . A simplex  $\sigma \in K$  is a *p-simplex* if the dimension of  $\sigma$  is  $p$ , which is written as  $\sigma^p$  for clarity if required. In particular, vertices are 0-simplices. Each nonempty subset  $\tau \subseteq \sigma$  is called a *face* of  $\sigma$ , and is a *proper face* when  $\tau \neq \sigma$ . The *dimension* of  $K$  is the largest dimension of its simplices, unless there is no finite upper bound for the dimensions of the simplices in which case the dimension of  $K$  is infinite. A *subcomplex* of  $K$  is a collection  $L \subseteq K$  that is also an abstract simplicial complex. While the definition of an abstract simplicial complex allows for an infinite set of vertices  $V$ , henceforth we only consider finite  $V$  and therefore finite simplicial complexes  $K$ , since infinite-dimensional complexes are not required for our work.

Abstract simplicial complexes are combinatorial objects, however simplicial complexes can also be defined as geometric objects in Euclidean space  $\mathbb{R}^n$ . In fact, an abstract simplicial complex has a *geometric realisation* as a geometric simplicial complex in  $\mathbb{R}^n$ . Correspondingly, if we just consider the subsets of vertices in a geometric simplicial complex then we can obtain a corresponding abstract simplicial complex.

**Definition 3.2 (Geometric  $p$ -simplex).** A *geometric  $p$ -simplex*  $\sigma$  in  $\mathbb{R}^n$  is the convex hull of  $p+1$  affinely independent points in  $\mathbb{R}^n$ , which are the vertices of the  $p$ -simplex. A *face* of  $\sigma$  is the convex hull of a subset of the  $p+1$  affinely independent points that determine  $\sigma$ .

Therefore, 0-simplices are points, 1-simplices are line segments, 2-simplices are triangles, 3-simplices are tetrahedra, and so forth in higher dimensions.

**Definition 3.3 (Geometric simplicial complex).** A *geometric simplicial complex* is a collection  $K$  of geometric simplices such that:

1. Every face of a simplex from  $K$  is also in  $K$ .
2. If two simplices have nonempty intersection then they intersect at a common face.

The abstract and geometric realisations of a simplicial complex are equivalent, so we may move between the two realisations as required.

## 3.2 Simplicial homology

**Definition 3.4 (Orientation of a simplex).** Let  $V = \{v_i\}_{i=0}^p$  be the vertex set of a simplex  $\sigma$ . Denote an ordering of  $V$  by  $[v_i]_{i \in I}$ , where  $I$  is a permutation of the integers  $\{0, 1, \dots, p\}$ . Two orderings of  $V$  are equivalent if they differ by an even permutation. If  $\dim(\sigma) \geq 1$  then the orderings of  $V$  form two equivalence classes, and there is one equivalence class when  $\dim(\sigma) = 0$ . Each equivalence class is an *orientation* of the simplex  $\sigma$ . An *oriented* simplex is a simplex  $\sigma$  together with an orientation of  $\sigma$ .

**Definition 3.5 (Oriented simplicial complex).** An *oriented* simplicial complex is a simplicial complex  $K$  together with a partial order on the set of vertices of  $K$ , such that the restriction of the partial order to the vertices of any simplex in  $K$  is a total order.

Note that every total ordering of the vertices of a simplicial complex  $K$  gives an oriented simplicial complex. In the following we assume that  $K$  is an oriented simplicial complex.

**Definition 3.6 ( $p$ -chain).** A  *$p$ -chain* is a finite formal sum of oriented  $p$ -simplices in  $K$ ,  $c = \sum_i a_i \sigma_i^p$ , where the coefficients  $a_i$  are elements of an abelian group.

For our purposes the coefficients are always from the finite cyclic group  $\mathbb{Z}/2\mathbb{Z}$ , which we can identify as  $\{0, 1\}$ , and we can denote by  $\mathbb{F}_2$  for simplicity since  $\mathbb{Z}/2\mathbb{Z}$  is also a finite field. Under coefficients from  $\mathbb{F}_2$  a  $p$ -chain is simply a set of  $p$ -simplices.

**Definition 3.7 ( $p$ -chain group).** Denote by  $C_p$  the set of all  $p$ -chains over  $K$ . Define a binary operation, denoted  $+$ ,

on the set of  $p$ -chains  $C_p$  by

$$c_0 + c_1 = \sum_i a_i \sigma_i^p + \sum_i b_i \sigma_i^p = \sum_i ((a_i + b_i \pmod{2})) \sigma_i^p, \quad (\text{S15})$$

so that the sum of a simplex with itself is zero. Then  $(C_p, +)$  is an abelian group with identity  $\sum_i 0 \sigma_i^p$  and inverse  $-c = c$ . The set of  $p$ -simplices  $\{\sigma_i^p\}_{i=1}^n$  in  $K$  generates  $C_p$  and is the minimal set of such generators. Therefore,  $(C_p, +)$  is a free abelian group, called the  $p$ -chain group, with basis  $\{\sigma_i^p\}_{i=1}^n$ .

For simplicity we denote the  $p$ -chain group as  $C_p$ , without reference to the group operation.

**Definition 3.8 (Boundary operator).** For  $p \geq 1$  the boundary operator  $\partial_p: C_p \rightarrow C_{p-1}$  maps a  $p$ -simplex to the sum of the  $(p-1)$ -simplices in its boundary. Specifically, for an oriented  $p$ -simplex  $\sigma^p = [v_i]_{i=0}^p$  we have

$$\partial_p(\sigma^p) = \sum_{i=0}^p (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_p], \quad (\text{S16})$$

where  $[v_0, \dots, \hat{v}_i, \dots, v_p]$  is the  $(p-1)$ -simplex obtained by deleting the vertex  $v_i$ . The boundary operator on general  $p$ -chains is given by

$$\partial_p \left( \sum_i a_i \sigma_i^p \right) = \sum_i (-1)^i a_i \partial_p(\sigma_i^p). \quad (\text{S17})$$

Note that  $\partial_p$  is a group homomorphism. For 0-chains we define  $C_{-1} = 0$ , then the boundary operator  $\partial_0$  satisfies  $\partial_0(c) = 0$  for all 0-chains  $c$ . Since we only consider chains with coefficients from  $\mathbb{F}_2$ , all of the coefficients  $(-1)^i$  equal one.

We now consider two important subgroups of  $C_p$ .

**Definition 3.9 (Cycle group,  $p$ -cycle).** Denote by  $Z_p$  the set of all  $p$ -chains in  $C_p$  with boundary zero, that is,

$$Z_p = \ker(\partial_p) = \{c \in C_p \mid \partial_p(c) = 0\}. \quad (\text{S18})$$

Since  $\partial_p$  is a homomorphism,  $Z_p$  is a subgroup of  $C_p$  called the *cycle group*. The  $p$ -chains in  $Z_p$  are called  *$p$ -cycles*.

Note that every 0-chain is a 0-cycle, so it follows that  $Z_0 = C_0$ .

**Definition 3.10 (Boundary group,  $p$ -boundary).** Denote by  $B_p$  the set of all  $p$ -chains in  $C_p$  which are the boundary of a  $(p+1)$ -chain, that is,

$$B_p = \text{im}(\partial_{p+1}) = \{c \in C_p \mid \exists b \in C_{p+1} \text{ with } \partial_{p+1}(b) = c\}. \quad (\text{S19})$$

Since  $\partial_{p+1}$  is a homomorphism,  $B_p$  is a subgroup of  $C_p$  called the *boundary group*. The  $p$ -chains in  $B_p$  are called  *$p$ -boundaries*.

Note that  $\partial_p \circ \partial_{p+1} = 0$  for all  $p \geq 0$ , hence a boundary does not have a boundary so is a cycle, or equivalently  $B_p$  is a subgroup of  $C_p$ .

The boundary operators provide a connection between the chain groups:

**Definition 3.11 (Chain complex).** Let  $K$  be an  $n$ -dimensional simplicial complex. Then the *chain complex*  $C_*$  is the sequence

$$0 \longrightarrow C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0, \quad (\text{S20})$$

where  $\partial_k \partial_{k+1} = 0$  for all  $k$ . The sequence is augmented on the right by a 0, and  $\partial_0 = 0$ . On the left we have  $C_{n+1} = 0$  as there are no  $(n+1)$ -simplices in  $K$ .

Since  $C_p$  is abelian,  $B_p$  is a normal subgroup and we can therefore construct the quotient group  $Z_p/B_p$ .

**Definition 3.12 ( $p$ -homology group).** The  $p$ -homology group  $H_p$  is defined as the quotient group  $Z_p/B_p$ .

The  $p$ -homology group  $H_p$  consists of equivalence classes of  $p$ -cycles that are not boundaries for any  $(p+1)$ -chains. Further, each equivalence class in  $H_p$  consists of homologous  $p$ -cycles that differ only in a  $p$ -boundary. So,  $H_p$  is nonzero if and only if the simplicial complex  $K$  has  $p$ -cycles that are not boundaries, or informally,  $K$  has  $p$ -dimensional holes. Computationally we are interested in the number of linearly independent  $p$ -dimensional holes, which we will see is given by the  $p$ -Betti number.

Since the oriented  $p$ -simplices form a basis for the  $p$ -chain group  $C_p$ , we can represent the boundary operator  $\partial_p: C_p \rightarrow C_{p-1}$  as a matrix with entries in  $\mathbb{F}_2$ , called the boundary matrix. We can compute the homology groups of a simplicial complex by reducing the boundary matrix to Smith normal form [3, Page 55, Theorem 11.3]. The reduction algorithm employs the following elementary column and row operations on the boundary matrix [5]:

**Matrix operation 1:** Exchange two columns or rows;

**Matrix operation 2:** Multiply any column or row by  $-1$ ;

**Matrix operation 3:** Add an integer multiple of one column or row to another.

The elementary column operations effect a change of basis for  $C_p$ , and similarly the elementary row operations effect a change of basis for  $C_{p-1}$ , so these operations preserve the rank of the matrix. Since we are computing homology over  $\mathbb{F}_2$ , for which  $-1 = 1$ , we only need the matrix operations 1 and 3, and moreover all matrix entries are either 0 or 1.

**Definition 3.13 (Boundary matrix and Smith normal form).** Let  $\{\sigma_i^p\}_{i=1}^n$  and  $\{\sigma_i^{p-1}\}_{i=1}^m$  be bases for  $C_p$  and

$C_{p-1}$ , respectively. The *boundary matrix*  $M_p$  of the boundary operator  $\partial_p: C_p \rightarrow C_{p-1}$  is the matrix

$$\begin{matrix} & \sigma_1^p & \sigma_2^p & \cdots & \sigma_n^p \\ \sigma_1^{p-1} & a_{11} & a_{12} & \cdots & a_{1n} \\ \sigma_2^{p-1} & a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_m^{p-1} & a_{m1} & a_{m2} & \cdots & a_{mn} \end{matrix} \quad (S21)$$

where  $a_{ij} = 1$  if and only if  $\sigma_i^{p-1}$  is a face of  $\sigma_j^p$ , otherwise  $a_{ij} = 0$ . By using column and row operations,  $M_p$  can be reduced to *Smith normal form*  $S_p$ , which has the same rank as  $M_p$ :

$$S_p = \left[ \begin{array}{ccc|c} d_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & d_l & \\ \hline & 0 & & 0 \end{array} \right]. \quad (S22)$$

The nonzero entries on the main diagonal,  $d_i$  for  $1 \leq i \leq l$ , satisfy  $d_i \mid d_{i+1}$  for  $1 \leq i < l$ , which holds trivially for a matrix over  $\mathbb{F}_2$  since each  $d_i$  is one.

The rank of  $S_p$  is the number of ones on the main diagonal of  $S_p$ , which we denote by  $b_{p-1}$ , and corresponds to  $p$ -chains with nonzero boundaries that generate  $B_{p-1}$ . Denote by  $z_p$  the number of right-most columns of  $S_p$  that are zero, which corresponds to  $p$ -cycles that generate  $Z_p$ . Then  $r_p = b_{p-1} + z_p$ , where  $r_p$  is the number of columns of  $S_p$ . For  $p \geq 0$  the  $p$ -Betti number  $\beta_p$  is obtained as

$$\beta_p = \text{rank}(Z_p) - \text{rank}(B_p) = z_p - b_p, \quad (S23)$$

and gives the number of linearly independent  $p$ -dimensional holes in  $K$ .

### 3.3 Persistent homology

Persistence is a measure for topological attributes [6], and persistent homology studies topological invariants that reveal the topological features of a space that persist across multiple scales [4, 5]. Our discussion of persistent homology follows [5]. We first define the notion of a weighted simplicial complex [26].

**Definition 3.14 (Weighted simplicial complex).** A *weighted simplicial complex* is a pair  $(K, w)$  where  $K$  is a simplicial complex and  $w: K \rightarrow \mathbb{R}$  is a weight function over  $K$  such that if  $\sigma, \tau \in K$  with  $\sigma \subseteq \tau$  then  $w(\sigma) \leq w(\tau)$ .

There are various methods to assign a weight function to an unweighted simplicial complex. For example, a weight

function can be constructed from a discrete Morse function [27].

**Definition 3.15 (Simplicial filtration, filtered simplicial complex, filtration index).** Given a weighted simplicial complex  $(K, w)$  and a strictly-increasing finite sequence  $\{\lambda_i\}_{i=0}^n \subseteq \mathbb{R}$ , the *simplicial filtration* is a nested sequence of subcomplexes  $K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K$  where  $K_i := w^{-1}((-\infty, \lambda_i])$ . For notational simplicity we denote the filtration as  $\{K_i\}_{i=0}^n$ . A simplicial filtration, or simply filtration, is also referred to as a *filtered simplicial complex*. The *filtration index* of a simplex  $\sigma \in K$  is equal to  $i$  if  $\sigma \in K_i \setminus K_{i-1}$ .

Sometimes  $\lambda_0$  is chosen so that  $K_0 = \emptyset$ , however this is not always necessary in practice.

**Definition 3.16 (Induced filtration).** Let  $(K, w)$  be a weighted simplicial complex, let  $\{\lambda_i\}_{i=0}^n \subseteq \mathbb{R}$  be a strictly-increasing finite sequence, and let  $\{K_i\}_{i=0}^n$  be a filtration for  $K$ . For a subcomplex  $L \subseteq K$  the induced filtration for  $L$  is obtained by the restriction of  $w$  to  $L$ . That is, the weighted simplicial complex is  $(L, w|_L)$  and the *induced filtration* is  $\{F_i\}_{i \in I}$  where, for each  $i \in I$ ,  $F_i = w|_L^{-1}((-\infty, \lambda_i]) = K_i \cap L \neq \emptyset$ , and  $i \leq j$  for all  $j$  such that  $0 \leq j \leq n$  and  $K_i \cap L = K_j \cap L$ . In this definition, since we may have  $K_i \cap L = K_j \cap L$  for some filtration indices  $i \neq j$ , we take the smallest such index as the corresponding filtration index. We also exclude filtration indices  $i$  for which  $K_i \cap L = \emptyset$ .

Given a filtration we can consider the  $p$ -homology groups of each subcomplex in the filtration.

**Definition 3.17 (Homology of filtration).** Let  $(K, w)$  be a weighted simplicial complex with filtration  $\{K_i\}_{i=0}^n$ , and fix  $0 \leq i \leq n$ . Let  $Z_p^i$  and  $B_p^i$  be the  $p$ -cycle and  $p$ -boundary groups, respectively, of  $K_i$ . The  *$p$ -homology group* of  $K_i$  is  $H_p^i = Z_p^i / B_p^i$ , and the  $p$ -Betti number  $\beta_p^i$  of  $K_i$  is the rank of  $H_p^i$ .

The  $p$ -Betti numbers of a simplicial filtration describe the topology of the filtered complex in terms of a sequence of integers. These Betti numbers, however, may not describe the underlying topological space very well, since while the Betti numbers describe the topology of the underlying space, they also describe the noise associated with the particular simplicial representation which appears as additional topological attribute. So the meaningful topological information is hidden within the topological noise.

In order to obtain meaningful information about the underlying space using a simplicial representation, we need to be able to distinguish between the noise arising from the representation and the topological features of the underlying space. We therefore require a measure of significance for the topological attributes associated with a simplicial representation, and one very effective measure is persistence. The idea behind persistence is that a significant topological attribute has a long lifetime, or persists, in a filtration. Persistence can therefore be defined solely in terms of a filtration.

In simplicial homology, nonbounding cycles from the same homology class are homologous to each other, meaning that the cycles differ only by a boundary. In persistent homology we want to determine the nonbounding cycles that persist in the filtration, that is, the cycles that are nonbounding for a significant number of filtration indices before possibly becoming boundaries. More formally, we factor the  $p$ -cycle group  $Z_p^i$  of the subcomplex  $K_i$  by the

$p$ -boundary group  $B_p^{i+j}$  of the subcomplex  $K_{i+j}$  that is  $j$  indices ahead in the filtration.

**Definition 3.18 (Persistent homology).** Let  $(K, w)$  be a weighted simplicial complex with filtration  $\{K_i\}_{i=0}^n$ , and fix  $0 \leq i \leq n$ . The  $j$ -persistent  $p$ -homology group  $H_p^{i,j}$  of  $K_i$  is

$$H_p^{i,j} = Z_p^i / (B_p^{i+j} \cap Z_p^i). \quad (\text{S24})$$

The  $j$ -persistent  $p$ -Betti number  $\beta_p^{i,j}$  of  $K_i$  is the rank of  $H_p^{i,j}$ .

Note that  $H_p^{i,j}$  is a group since it is the intersection of the two subgroups  $Z_p^i$  and  $B_p^{i+j}$  of  $C_p^{i+j}$ .

In order to understand the structure of persistent homology it is useful to approach it from a different perspective. Our intuition suggests that the computation of persistence necessitates compatible bases for the groups  $H_p^i$  and  $H_p^{i,j}$ , however we need to establish the existence of a succinct description for the compatible bases. This is achieved in [5] by combining the homology of every complex in the filtration into a single algebraic structure.

**Definition 3.19 (Persistence complex).** A persistence complex  $\mathcal{C}$  is a family of chain complexes  $\{C_*^i\}_{i \geq 0}$  over a commutative ring  $R$  with unity, together with chain maps  $f^i: C_*^i \rightarrow C_*^{i+1}$  so that we have the following diagram:

$$C_*^0 \xrightarrow{f^0} C_*^1 \xrightarrow{f^1} C_*^2 \xrightarrow{f^2} \dots \quad (\text{S25})$$

A filtered simplicial complex  $K$  with inclusion maps for the simplices is a persistence complex.

**Definition 3.20 (Persistence module).** Let  $R$  be a commutative ring with unity. A persistence module  $\mathcal{M}$  is a family of  $R$ -modules  $M^i$  together with homomorphisms  $\phi^i: M^i \rightarrow M^{i+1}$ .

The homology of a persistence complex is a persistence module, where the homomorphism  $\phi^i$  maps a homology class to the one that contains it.

**Definition 3.21 (Finite type).** Let  $R$  be a commutative ring with unity. A persistence complex  $\{C_*^i, f^i\}$  (resp. persistence module  $\{M^i, \phi^i\}$ ) is of *finite type* if each component complex (resp. module) is a finitely generated  $R$ -module and if the maps  $f^i$  (resp.  $\phi^i$ ) are isomorphisms for  $i \geq m$  for some integer  $m$ .

Since we consider only finite simplicial complexes  $K$ , each such  $K$  generates a persistence complex  $\mathcal{C}$  of finite type whose homology is a persistence module  $\mathcal{M}$  of finite type.

In the following we require the notions of graded rings and modules [28, Chapter 1, page 1; Chapter 2, page19]:

**Definition 3.22 (Graded rings and modules).** Let  $R$  be a ring with unity, and let  $G$  be a multiplicative group. Then  $R$  is a  $G$ -graded ring if there is a family  $\{R_g \mid g \in G\}$  of additive subgroups  $R_g$  of  $R$  such that  $R = \bigoplus_{g \in G} R_g$  as commutative groups and  $R_g R_h \subseteq R_{gh}$  for all  $g, h \in G$ . The set  $h(R) = \bigcup_{g \in G} R_g$  is the set of *homogeneous elements* of  $R$ , and a nonzero element  $x \in R_g$  is said to be *homogeneous of degree  $g$* .



A (left)  $G$ -graded  $R$ -module, or simply a *graded module over  $R$* , is a left  $R$ -module  $M$  such that  $M = \bigoplus_{g \in G} M_g$  where each  $M_g$  is an additive subgroup of  $M$ , and for every  $g, h \in G$  we have  $R_g M_h \subseteq M_{gh}$ . Since  $R_e M_h \subseteq M_h$ , every  $M_h$  is an  $R_e$ -submodule of  $M$ . The elements of  $\bigcup_{h \in G} M_h$  are the *homogeneous elements* of  $M$ . A nonzero element  $m \in M_h$  is *homogeneous of degree  $h$* .

An important example of a graded ring is the standard grading of a polynomial ring [29, Chapter 1, page 1]:

**Definition 3.23 (Standard grading).** Let  $S$  be the polynomial ring  $\mathbb{F}[x_1, \dots, x_n]$  over the field  $\mathbb{F}$ . Set  $\deg(x_i) = 1$  for each  $i$ . A monomial  $x^{c_1} \cdots x^{c_n}$  has *degree*  $c_1 + \cdots + c_n$ . For a nonnegative integer  $i$  we denote by  $S_i$  the  $\mathbb{F}$ -vector space spanned by all monomials of degree  $i$ . In particular,  $S_0 = \mathbb{F}$ . A polynomial  $u \in S$  is called *homogeneous* if  $u \in S_i$  for some  $i$ , and we then say that  $u$  has degree  $i$  and write  $\deg(u) = i$ . Note that 0 is a homogeneous element with arbitrary degree. The following two properties are equivalent, and hold for  $S$ :

1.  $S_i S_j \subseteq S_{i+j}$  for all nonnegative integers  $i$  and  $j$ .
2.  $\deg(uv) = \deg(u) + \deg(v)$  for every two homogeneous elements  $u, v \in S$ .

Every polynomial  $f \in S$  can be written uniquely as a finite sum  $f = \sum_i f_i$  of nonzero elements  $f_i \in S_i$ , and in this case  $f_i$  is called the *homogeneous component* of  $f$  of degree  $i$ . Therefore we have a direct-sum decomposition  $S = \bigoplus_{i \geq 0} S_i$  of  $S$  as a  $\mathbb{F}$ -vector space, where  $S_i S_j \subseteq S_{i+j}$  for all nonnegative integers  $i$  and  $j$ . We say that  $S$  has the *standard grading*.

Suppose we have a persistence module  $\mathcal{M} = \{M^i, \phi^i\}_{i \geq 0}$  over a commutative ring  $R$  with unity. Equip the polynomial ring  $R[t]$  with the standard grading, and define a graded module over  $R[t]$  by

$$\alpha(\mathcal{M}) = \bigoplus_{i=0}^{\infty} M^i, \quad (\text{S26})$$

where the  $R$ -module structure is simply the sum of the structures on the individual components, and the action of  $t$  is given by

$$t \cdot (m^0, m^1, m^2, \dots) = (0, \phi^0(m^0), \phi^1(m^1), \phi^2(m^2), \dots). \quad (\text{S27})$$

Note that  $t$  simply shifts elements of the module up in the gradation.

We now obtain the following correspondence:

**Theorem 3.24 (Correspondence).** *The correspondence  $\alpha$  defines an equivalence of categories between the category of persistence modules of finite type over  $R$  and the category of finitely generated nonnegatively graded modules over  $R[t]$ .*

Intuitively, we are constructing a single algebraic structure that contains all of the simplicial subcomplexes in the filtration. We begin by computing a direct sum of the complexes, arriving at a much larger space that is graded

according to the filtration ordering. We then remember the time each simplex enters using a polynomial coefficient. The key idea is that the filtration ordering is encoded in the coefficient polynomial ring.

The correspondence described in Theorem 3.24 suggests the nonexistence of simple classifications of persistence modules over a ground ring that is not a field, such as  $\mathbb{Z}$ . It is well known in commutative algebra that the classification of modules over  $\mathbb{Z}[t]$  is extremely complicated. While it is possible to assign interesting invariants to  $\mathbb{Z}[t]$ -modules, a simple classification is not available, nor is it ever likely to be available.

On the other hand, the correspondence gives a simple decomposition when the ground ring is a field  $\mathbb{F}$ . Here, the graded ring  $\mathbb{F}[t]$  is a principal ideal domain and its only graded ideals are homogeneous of form  $(t^n)$ , so the structure of the  $\mathbb{F}[t]$ -module is described as:

$$\left( \bigoplus_{i=1}^n \Sigma^{\alpha_i} \mathbb{F}[t] \right) \oplus \left( \bigoplus_{j=1}^m \Sigma^{\gamma_j} \mathbb{F}[t] / (t^{n_j}) \right), \quad (\text{S28})$$

where  $n_j \leq n_{j+1}$  for  $1 \leq j \leq m-1$ ,  $\alpha_i, \gamma_j \in \mathbb{Z}$ , and  $\Sigma^\alpha$  denotes an  $\alpha$ -shift upward in grading.

We wish to parametrise the isomorphism classes of  $\mathbb{F}[t]$ -modules by suitable objects.

**Definition 3.25 ( $\mathcal{P}$ -interval).** A  $\mathcal{P}$ -interval is an ordered pair  $(i, j)$  where  $0 \leq i < j$  for  $i \in \mathbb{Z}$  and  $j \in \mathbb{Z} \cup \{+\infty\}$ .

We associate a graded  $\mathbb{F}[t]$ -module to a set  $S$  of  $\mathcal{P}$ -intervals via a bijection  $Q$ , which is defined by  $Q(i, j) = \Sigma^i \mathbb{F}[t] / (t^{j-1})$  for  $\mathcal{P}$ -interval  $(i, j)$ , noting that  $Q(i, +\infty) = \Sigma^i \mathbb{F}[t]$ . For a set of  $\mathcal{P}$ -intervals  $S = \{(i_1, j_1), (i_2, j_2) \dots, (i_n, j_n)\}$ , we define

$$Q(S) = \bigoplus_{l=1}^n Q(i_l, j_l). \quad (\text{S29})$$

Our correspondence may now be restated as follows.

**Corollary 3.26.** *The correspondence  $S \rightarrow Q(S)$  defines a bijection between the finite sets of  $\mathcal{P}$ -intervals and the finitely generated graded modules over the graded ring  $\mathbb{F}[t]$ . Consequently, the isomorphism classes of persistence modules of finite type over  $\mathbb{F}$  are in bijective correspondence with the finite sets of  $\mathcal{P}$ -intervals.*

In summary, we are interested in the  $k$ th homology of a filtered simplicial complex  $K$ . In each dimension the homology of the subcomplex  $K_i$  becomes a vector space over a field, and is described fully by its rank  $\beta_k^i$ . We need to choose compatible bases across the filtration in order to compute persistent homology for the entire filtration. So, we form the persistence module corresponding to  $K$ , which is a direct sum of these vector spaces. The structure theorem states that a basis exists for this module that provides compatible bases for all of the vector spaces. In particular, each  $\mathcal{P}$ -interval  $(i, j)$  describes a basis element for the homology vector spaces starting at time  $i$  until time  $j-1$ . This element is a  $k$ -cycle that is completed at time  $i$ , forming a new homology class. It also remains nonbounding until time  $j$ , at which time it joins the boundary group  $B_k^j$ .

The persistent homology of a filtered simplicial complex is obtained by combining the homology of all subcomplexes in the simplicial filtration into a single algebraic structure, namely a graded module over a polynomial ring with

indeterminate  $t$ , for which the standard homology is the persistent homology of the complex. Informally, the direct sum of the subcomplexes in the filtration gives a larger space that is graded according to the order of the filtration. Moving through the filtration, when a simplex enters the next subcomplex it is indicated with a polynomial coefficient, or from the perspective of the graded module, the simplex is shifted along the grading by multiplying the simplex using  $t$ . The graded module therefore is a single structure that contains all of the subcomplexes in the filtration.

*Definition 3.27 ( $H_p$ -barcode).* Let  $(K, w)$  be a weighted simplicial complex. An  $H_p$ -barcode is a multiset  $L_p$  of intervals of two forms:

1.  $[i, \infty)$ , which corresponds to a topological attribute that is created at filtration index  $i \in \mathbb{N} \cup \{0\}$  and exists in the final structure;
2.  $[j, k)$ , which corresponds to a topological attribute that is created at filtration index  $j \in \mathbb{N} \cup \{0\}$  and is destroyed at index  $k \in \mathbb{N} \cup \{0\}$ .

The algorithm for computing the  $H_p$ -barcodes is detailed in [5].

## 4 Systems of reaction-diffusion equations

A *reaction-diffusion system* for  $N \geq 2$  interacting agents in a domain  $\Omega \times \mathcal{I} \subset \mathbb{R}^n \times \mathbb{R}_{\geq 0}$ ,  $n \geq 1$ , consists of a set of  $N$  coupled reaction-diffusion equations in  $n$  spatial dimensions,

$$\frac{\partial \mathbf{c}}{\partial t} = \nabla \cdot (\mathbf{D} \nabla \mathbf{c}) + \mathbf{f}(\mathbf{c}), \quad (\text{S30})$$

where  $t \in \mathcal{I}$  is time,  $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \Omega$  is position, the state variable  $\mathbf{c}: \Omega \times \mathcal{I} \rightarrow \mathbb{R}^N$ ,  $(\mathbf{x}, t) \mapsto \mathbf{c}(\mathbf{x}, t)$ , is the vector of agent concentrations,  $\mathbf{D} = \text{diag}(D_1, D_2, \dots, D_N)$  is the diagonal diffusivity matrix, and the system of reactions is described by the vector-valued function of a vector variable  $\mathbf{f}$ , which is generally nonlinear. We allow for the diffusivity of each agent to depend on the local concentrations of the other agents, and we assume there is no cross-diffusion. Solutions of the system (S30) on the domain  $\Omega \times I$  with spatial boundary  $\partial\Omega$  require functions specifying both the boundary conditions  $\mathbf{c}(\mathbf{x}, t)|_{\partial\Omega}$  and initial conditions  $\mathbf{c}(\mathbf{x}, 0)$ .

From a mathematical perspective the model and associated boundary condition define a dynamical system, which evolves from a specified initial condition. The boundary conditions employed for Turing-pattern models and positional-information models are:

**Neumann:**

$$\frac{\partial \mathbf{c}}{\partial \mathbf{n}}(\mathbf{x}) = h(\mathbf{x}) \quad \forall \mathbf{x} \in \partial\Omega, \quad (\text{S31})$$

where  $\mathbf{n}$  is the exterior normal to the boundary  $\partial\Omega$ , and  $h$  is a scalar function. The normal derivative is defined by

$$\frac{\partial \mathbf{c}}{\partial \mathbf{n}}(\mathbf{x}) = \nabla \mathbf{c}(\mathbf{x}) \cdot \hat{\mathbf{n}}(\mathbf{x}), \quad (\text{S32})$$

where  $\hat{\mathbf{n}}$  is the unit normal. In particular, zero flux boundary conditions, where  $h$  is the zero function, correspond to a boundary of the spatial domain which is insulated, reflective, or isolated from the external environment.

**Dirichlet:**

$$\mathbf{c}(\mathbf{x}) = h(\mathbf{x}) \quad \forall \mathbf{x} \in \partial\Omega, \quad (\text{S33})$$

where  $\partial\Omega$  is the boundary, and  $h$  is a scalar function.

**Periodic:** for functions and derivatives.

These boundary conditions may also be combined as in Robin and mixed boundary conditions.

## References

- [1] Spanier EH. Algebraic Topology. Springer New York; 1966. doi:10.1007/978-1-4684-9322-1.
- [2] Rotman JJ. An Introduction to Algebraic Topology. Springer New York; 1988. doi:10.1007/978-1-4612-4576-6.
- [3] Munkres J. Elements of Algebraic Topology. CRC Press; 2018. doi:10.1201/9780429493911.
- [4] Edelsbrunner H, Letscher D, Zomorodian A. Topological persistence and simplification. *Discrete & Computational Geometry*. 2002;28:511–533. doi:10.1007/s00454-002-2885-2.
- [5] Zomorodian A, Carlsson G. Computing persistent homology. *Discrete & Computational Geometry*. 2005;33:249–274. doi:10.1007/s00454-004-1146-y.
- [6] Zomorodian A. Topology for Computing. Cambridge University Press; 2005. doi:10.1017/cbo9780511546945.
- [7] Stumpf HF. Über den Verlauf eines schuppenorientierenden Gefälles bei *Galleria mellonella*. *Wilhelm Roux' Archiv für Entwicklungsmechanik der Organismen*. 1967;158:315–330. doi:10.1007/bf00573402.
- [8] Wolpert L. Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology*. 1969;25:1–47. doi:10.1016/S0022-5193(69)80016-0.
- [9] Crick F. Diffusion in embryogenesis. *Nature*. 1970;225:420–422. doi:10.1038/225420a0.
- [10] Čapek D, Müller P. Positional information and tissue scaling during development and regeneration. *Development*. 2019;146:dev177709. doi:10.1242/dev.177709.
- [11] Wartlick O, Kicheva A, González-Gaitán M. Morphogen gradient formation. *Cold Spring Harbor Perspectives in Biology*. 2009;1:a001255. doi:10.1101/cshperspect.a001255.
- [12] Umulis DM, Othmer HG. Mechanisms of scaling in pattern formation. *Development*. 2013;140:4830–4843. doi:10.1242/dev.100511.
- [13] McHale P, Rappel WJ, Levine H. Embryonic pattern scaling achieved by oppositely directed morphogen gradients. *Physical Biology*. 2006;3:107–120. doi:10.1088/1478-3975/3/2/003.
- [14] Houchmandzadeh B, Wieschaus E, Leibler S. Precise domain specification in the developing *Drosophila* embryo. *Physical Review E*. 2005;72:061920. doi:10.1103/PhysRevE.72.061920.
- [15] Ben-Zvi D, Barkai N. Scaling of morphogen gradients by an expansion-repression integral feedback control. *Proceedings of the National Academy of Sciences*. 2010;107:6924–6929. doi:10.1073/pnas.0912734107.

- [16] Rahimi N, Averbukh I, Haskel-Ittah M, Degani N, Schejter ED, Barkai N, et al. A WntD-dependent integral feedback loop attenuates variability in *Drosophila* toll signaling. *Developmental Cell*. 2016;36:401–414. doi:10.1016/j.devcel.2016.01.023.
- [17] Shilo BZ, Barkai N. Buffering global variability of morphogen gradients. *Developmental Cell*. 2017;40:429–438. doi:10.1016/j.devcel.2016.12.012.
- [18] Dillon R, Maini PK, Othmer HG. Pattern formation in generalized Turing systems. *Journal of Mathematical Biology*. 1994;32:345–393. doi:10.1007/BF00160165.
- [19] Varea C, Aragón JL, Barrio RA. Confined Turing patterns in growing systems. *Physical Review E*. 1997;56:1250–1253. doi:10.1103/PhysRevE.56.1250.
- [20] Barrio RA, Varea C, Aragón JL, Maini PK. A two-dimensional numerical study of spatial pattern formation in interacting Turing systems. *Bulletin of Mathematical Biology*. 1999;61:483–505. doi:10.1006/bulm.1998.0093.
- [21] Meinhardt H. Turing's theory of morphogenesis of 1952 and the subsequent discovery of the crucial role of local self-enhancement and long-range inhibition. *Interface Focus*. 2012;2:407–416. doi:10.1098/rsfs.2011.0097.
- [22] Nesterenko AM, Kuznetsov MB, Korotkova DD, Zarskiy AG. Morphogene adsorption as a Turing instability regulator: theoretical analysis and possible applications in multicellular embryonic systems. *PLOS ONE*. 2017;12:e0171212. doi:10.1371/journal.pone.0171212.
- [23] Gierer A, Meinhardt H. A theory of biological pattern formation. *Kybernetik*. 1972;12:30–39. doi:10.1007/BF00289234.
- [24] Landge AN, Jordan BM, Diego X, Müller P. Pattern formation mechanisms of self-organizing reaction-diffusion systems. *Developmental Biology*. 2020;460:2–11. doi:10.1016/j.ydbio.2019.10.031.
- [25] Giusti C, Ghrist R, Bassett DS. Two's company, three (or more) is a simplex. *Journal of Computational Neuroscience*. 2016;41:1–14. doi:10.1007/s10827-016-0608-6.
- [26] Espinoza JF, Hernández-Amador R, Hernández-Hernández HA, Ramonetti-Valencia B. A numerical approach for the filtered generalized Čech complex. *Algorithms*. 2020;13:11. doi:10.3390/a13010011.
- [27] Kannan H, Saucan E, Roy I, Samal A. Persistent homology of unweighted complex networks via discrete Morse theory. *Scientific Reports*. 2019;9:13817. doi:10.1038/s41598-019-50202-3.
- [28] Năstăsescu C, van Oystaeyen F. *Methods of Graded Rings*. Springer Berlin Heidelberg; 2004. doi:10.1007/b94904.
- [29] Peeva I. *Graded Syzygies*. Springer London; 2011. doi:10.1007/978-0-85729-177-6.