The Persistence of Probability Landscapes of Stochastic Gene Regulatory Networks

Ali M. Farhat[1,*], Hammad Farooq[1], Jiun Cho[2], Wei Tian[1], Calin C. Guet[3], Herbert Edelsbrunner[3], Farid Manuchehrfar[1], Jie Liang[1,†], Hubert Wagner[2,†]

[1] Center for Bioinformatics and Quantitative Biology, Department of Biomedical Engineering, University of Illinois at Chicago, Chicago, IL 60607, USA
[2] Department of Mathematics, University of Florida, Gainesville, FL, USA
[3] Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria
[*] To whom correspondence should be addressed.

[†] Co–senior authors.

Classification: Biological Sciences; Computational Biology

Keywords: chemical master equation; probability landscape; stochastic reaction networks; topological data analysis; persistent homology; high-dimensional voids

Abstract

Biological systems at the level of interacting molecules are inherently stochastic, and noise is ubiquitous in their dynamics. Probability landscapes therefore provide a natural description of system behavior; however, their high dimensionality presents a fundamental obstacle to both quantitative and qualitative analysis. Existing approaches typically rely on low-dimensional projections or dimensionality reduction, which can obscure peaks, distort connectivity, and eliminate higher-dimensional structure. To enable analysis of large, high-dimensional landscapes, we introduce a new specialized persistent homology-based algorithm for characterizing probability landscapes directly in their original high-dimensional state space. Using superlevel-set filtrations of landscapes computed from the discrete Chemical Master Equation, we identify prominent features across scales and quantify their persistence, providing a principled measure of feature significance. Because persistent homology is stable to noise and captures multiscale topological features, our framework reveals previously inaccessible structural organization in canonical stochastic reaction networks, including the feedback activation–inhibition network and three- and four-species repressilators. More broadly, this approach provides a topology-preserving framework to compare stochastic network architectures and to map the structural organization of complex landscapes without projection artifacts. This capability supports systems-level interpretation and provides a useful characterization of phenotypic organization that is relevant to understanding naturally occurring biological networks as well as to synthetic gene circuit design and analysis.

Significance Statement

Stochasticity in molecular events governs how biological reaction networks move between functional states. We introduce a computational method that enables the interpretation of these systems by analyzing the full, high-dimensional probability landscape computed from the Chemical Master Equation—a map that assigns occupancy probabilities to every possible molecular state configuration at a given time. Using persistent homology—a tool from topological data analysis—our method quantifies multiscale structure in this landscape by identifying (i) distinct probability regions that correspond to phenotypes, (ii) loop-like high-probability routes that connect phenotypes and indicate likely switching paths, and (iii) a void of lower-probability states encased within a closed shell of higher-probability states. Together, these signatures quantify the topological structure of the phenotypic landscape.

1 Introduction

Biomolecules such as protein, DNA, and RNA interact in cells to carry out cellular functions. These molecular interactions are often described as networks of chemical reactions. When the quantities of individual molecular species are small, there are inherently pronounced stochastic effects in these interactions. Stochasticity in molecular networks plays important roles in a variety of biological phenomena, ranging from signal transduction [1, 2], metabolic regulations [3, 4], to cellular fate determination [5, 6]. It is therefore fundamental to understand the stochastic behavior of reaction networks.

The stochastic processes of a network of chemical reactions, also termed the *Stochastic Chemical Kinetics (SCK) processes* [7, 8, 9, 10, 11, 12], can be described by its time-evolving probability density function, also referred to as *probability landscape* over the microstates of the network. Here the microstates are integer vectors of copy numbers of the different combinations of the individual molecular species. This probability landscape is governed by the Chemical Master Equation over discrete state space [13, 8, 10, 14, 15, 16]. An alternative description of the stochastic reactions and the SCK process is by trajectories of reaction paths, which are governed by random-time changed integral equations of a Poisson process [17, 18, 9, 10].

The complexity of biological processes dictates that their probability landscapes are often high-dimensional. There have been important advances in computational techniques aiming to directly solve the discrete Chemical Master Equation (dCME) [19, 20, 15, 16, 21, 22, 23, 24, 25, 26]. Among these, the finite buffer n-simplex algorithm provides an optimal method for enumerating the number of different states [20, 27, 15]. Theoretical error bounds have also been obtained that can specify simple conditions guaranteeing acceptable accuracy when full state-space enumeration is not possible and truncations are applied [16]. As a result, the probability landscapes for a large class of high-dimensional stochastic reaction networks have now been constructed [15, 16, 27, 28, 29, 30, 31, 32, 33].

Examples include a 16-dimensional MAPK network [15], the Tat circuit of HIV-latency [27], and the construction of the phase diagram of multi-modality of the feed-forward networks through parameter sweeps [33]. The ability to construct probability landscapes of networks has significant advantages over analysis of trajectory-based simulations [9, 10], as an exorbitantly large number of trajectories is required to ensure sufficient sampling for estimating probability densities.

With the availability of accurately computed probability landscapes, a number of important challenges arise. **First**, there is a need for methods that can navigate high-dimensional probability landscapes in a manner that reliably identifies all major probability peaks, quantifies their relative persistence, and establishes the structural connections between them. **Second**, because the notion of reactivity in complex stochastic networks is not well defined, it remains unclear how to identify reactive regions of the high-dimensional landscape in which the system evolves with higher probability. In particular, it is unknown whether such regions are organized as one-dimensional pathways, closed cycles, or higher-dimensional manifolds. **Third**, a fundamental challenge is to develop frameworks that enable explanation of abnormal behavior in stochastic networks observed in nature, as well as the rational

design of synthetic networks with prescribed stochastic properties. Addressing these challenges requires an exhaustive characterization of the underlying probability landscape, from which network behavior can be fully understood.

There are fundamental shortcomings in all current approaches JC: could this be too harshly worded?for analyzing the behavior of stochastic networks. A common practice is to investigate the probability landscape after projection on an *a priori* selected subspace [34, 35]. Conclusions drawn using this approach may be misleading, as spurious probability peaks may be identified and incorrect high-probability transition pathways may be inferred. This is due to the fact that the topology of 2- or 3- dimensional projections rarely reflects the actual topology of the high-dimensional space, as shown in a recent study [36]. This is in addition to potential difficulties that arise when landscapes can only be approximated from trajectories obtained from stochastic simulations [37], with the accuracy deteriorating rapidly with increasing dimension. Furthermore, dimensionality reduction methods such as linear principal component analysis (PCA) [38] are often used to analyze probability landscapes by transforming the data into a low-dimensional space. However, PCA can distort the topological structure of a high-dimensional surface and introduce artificial features [39]. Similarly, none of the available non-linear methods is guaranteed to preserve the topological structure of the original data, making them unsuitable for the accurate topological analysis of high-dimensional probability landscapes. Another approach is through the analysis on the Freidlin-Wentzell or similarly constructed quasi-potential surfaces for multi-stable systems [40, 41, 42, 43]. However, this approach is not suitable as perturbations around local extrema of constructed quasi-potential surfaces do not guarantee a full characterization of the global features of the probability landscape. Furthermore, the requirement of small noise often cannot be assumed when copy numbers are small and binding is weak.

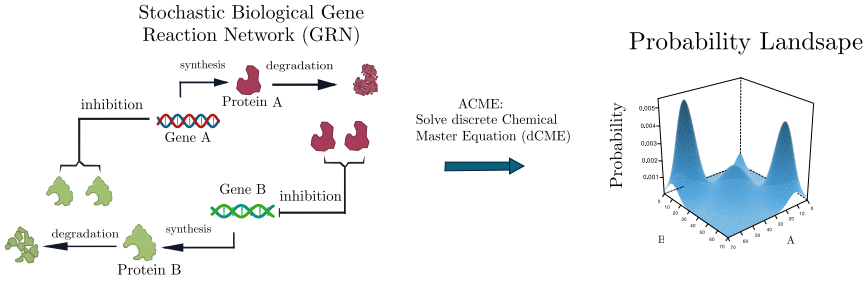
In this work, we introduce a new approach for quantifying the topological structure of probability landscapes. Based on the techniques of persistent homology [44], our method identifies peaks, cycles, and shell-enclosed voids in high-dimensional probability landscapes, and quantifies their prominence. Our method works well in high-dimensional space (currently up to dimension 7), making it feasible to investigate the topological structures of a number of networks of higher dimension. With this approach, we obtain a topological view of high-dimensional probability landscapes which allows us to uncover structural properties of the underlying stochastic networks that were not previously detectable (See Box 1).

2 Methods

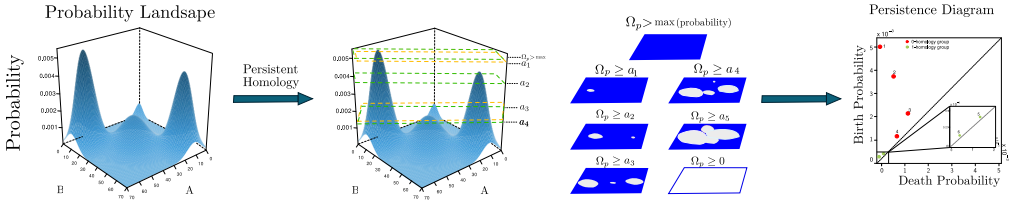
The main mathematical concept in this work is the probability density function of a chemical or biological reaction network, which evolves in time. Section 2.1 formally introduces the time-series of density functions, together with the discrete chemical master equations used to compute them. The novel aspect of this work, however, is the quantification of its topological features, described in Section 2.2 and computed in Section 2.3. Such features reveal phenotypic structure and the multiscale topological organization of high-probability regions in state space.

Box 1: Overview of the method

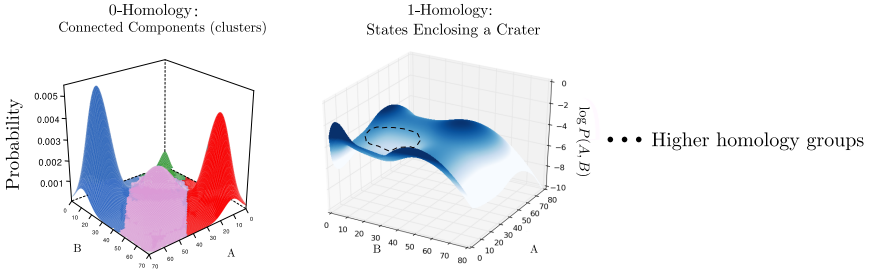
From a stochastic gene regulatory network (GRN), we calculate the probability of occupying each state in the state space by obtaining the exact solution of the discrete Chemical Master Equation (dCME) using ACME.



Next, we apply persistent homology by lowering a level set from the highest-probability state to the lowest probability state. This is analogous to mountains submerged under water: as we lower the water level, we track which mountains emerge (birth) and the level at which they meet at a valley (death). Death follows the elder rule, meaning the younger mountain (revealed later) dies for the older mountain (revealed earlier).



Using persistent homology, we uncover the topological properties of GRNs, including connected regions (0-homology), loops that enclose low-probability "craters" (1-homology), and higher-dimensional homology classes.



2.1 The Probability Landscape

We refer to the probability density function of a reaction network as its *probability landscape*. In this work, we assume a well-mixed reaction system with constant temperature and volume, and a total of n *molecular species*, denoted X_i for $i = 1, 2, \ldots, n$. Let $x_i(t) \geq 0$ be the copy number of X_i at time t, and call the vector $\boldsymbol{x}(t) = [x_1(t), x_2(t), \ldots, x_n(t)]$ the *microstate* of the system at time t. The species participate in m reactions, denoted R_j for $j = 1, 2, \ldots, m$, taking the form

$$c_{j,1}X_1 + c_{j,2}X_2 + \ldots + c_{k,n}X_n \xrightarrow{r_j} c'_{j.1}X_1 + c'_{j,2}X_2 + \ldots + c'_{j,n}X_n, \tag{1}$$

which brings the system from the microstate $\boldsymbol{x}$ to the new microstate $\boldsymbol{x} + \boldsymbol{s}_j$, where $\boldsymbol{s}_j$ is the *stoichiometry vector* of R_j, defined as $\boldsymbol{s}_j = [c'_{j,1} - c_{j,1}, c'_{j,2} - c_{j,2}, \ldots, c'_{j,n} - c_{j,n}]$, and r_j is the *reaction rate*. The reaction preserves mass, so the total mass before and after the reaction is the same. Exceptions are reactions that synthesize species from the outside, in which case the left-hand side is empty, and reactions that degrade, in which case the right-hand side is empty. We follow how the system evolves from time $t = 0$ to infinity, and write Ω for the *state space*, which is the collection of all possible microstates, whether or not they arise:

$$\Omega = \{[x_1, x_2, \ldots, x_n] \mid x_i \geq 0 \text{ for } 1 \leq i \leq n\}. \tag{2}$$

A single evolution of the system may thus be understood as a path through state space, but note that this path can cross itself because it may go through the same microstate more than once. The shape of such a path is influenced by the likelihood of a reaction to happen. This motivates the main definition in this work.

Definition 1 *The* probability landscape *of the reaction network,* $\boldsymbol{p}\colon \Omega \times [0, \infty) \to [0, 1]$*, maps every microstate,* $\boldsymbol{x} \in \Omega$*, and moment in time,* $t \in [0, \infty)$*, to the probability that* $\boldsymbol{x}$ *is the microstate of the system at time* t*. For fixed* t*, we write* $\boldsymbol{p}_t\colon \Omega \to [0, 1]$ *for the* time-slice *at* t*, defined by* $\boldsymbol{p}_t(\boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x}, t)$*.*

With the assumption of a well-mixed system, the *propensity* of the j-th reaction depends on the availability of the species needed for the reaction:

$$A_j(\boldsymbol{x}) = r_j \prod_{i=1}^{n} \binom{x_i}{c_{j,i}}, \tag{3}$$

The propensity $A_j(x)$ admits a natural interpretation as the instantaneous *transition rate of probability mass* out of microstate x through reaction channel j. While the probability landscape $p_t(x)$ specifies the amount of probability mass residing at microstate x at time t, the propensity determines how rapidly this mass is redistributed to neighboring microstates due to reaction events. The intrinsic rate constant r_j encodes the kinetic timescale of reaction j, whereas the combinatorial factor $\prod_{i=1}^{n} \binom{x_i}{c_{j,i}}$ counts the number of distinct molecular configurations in microstate x that are capable of firing reaction j. Consequently, microstates with larger copy numbers admit more reaction-ready configurations and therefore exhibit higher transition rates. In the discrete chemical master equation, each term $A_j(x)\, p_t(x)$ represents an outgoing transition rate of probability mass from x, while terms of the form $A_j(x - s_j)\, p_t(x - s_j)$ represent incoming probability from adjacent microstates. The dCME thus enforces local conservation of probability mass on the discrete state space by balancing these inflow and outflow contributions.

We compute the probability landscape using the *discrete Chemical Master Equation* (dCME) of the reaction network, which consists of a collection of ordinary differential equa-

tions (ODEs):

$$\frac{\mathrm{d}\boldsymbol{p}(\boldsymbol{x},t)}{\mathrm{d}t} = \sum_{j=1}^{m} \left[A_j(\boldsymbol{x}-\boldsymbol{s}_j)p_t(\boldsymbol{x}-\boldsymbol{s}_j) - A_j(\boldsymbol{x})p_t(\boldsymbol{x})\right], \tag{4}$$

in which $\boldsymbol{x}$ and $\boldsymbol{x}-\boldsymbol{s}_j$ both belong to Ω. The time-evolving and steady-state probability landscape can be computed by solving these differential equations using methods such as the n-simplex finite buffer ACME algorithm [20, 29, 15, 16].

Remark 1 *While the restriction of the probability landscape to a fixed microstate is continuous, the restriction to a fixed time is not. Indeed, the state space is a countable collection of microstates, each mapped to a probability. It makes sense to model a time-slice as a continuous function, obtained by piecewise linear interpolation between the integer points in $\mathbb{R}^n$. We will take this point of view and pretend that each time-slice is indeed a smooth function on $\mathbb{R}^n$. While it is convenient to assume smoothness, it will not be necessary to compute such functions because we will work with concepts that are defined for piecewise linear continuous functions as well. Indeed, the algorithm we use to compute features of the probability landscape does so directly from the values at the points in Ω.*

2.2 Quantifying the Prominence of Critical Points

The traditional approach to identify features of a real-valued function is through the definition of critical points, and we refer to [45, 46] for background reading. Here we assume a smooth function on a manifold, $f\colon M \to \mathbb{R}$, but we hasten to mention that smoothness is convenient but neither necessary nor realistically achievable. We call $\boldsymbol{x} \in M$ a *critical point* of f if all its partial derivatives vanish; that is: the gradient of f at $\boldsymbol{x}$ is the zero vector, and $\boldsymbol{x}$ is *non-degenerate* if the matrix of second derivatives is invertible; that is: the Hessian of f at $\boldsymbol{x}$ has non-zero determinant. As an example, consider Figure **1B**, where we see a non-degenerate critical point (a maximum) surrounded by a circle of degenerate critical points forming the top of a crater. Unlike the points of this circle, a non-degenerate critical point is necessarily isolated and thus cannot be arbitrarily close to another critical point.

Letting n be the dimension of M, the Hessian of f at a non-degenerate critical point, $\boldsymbol{x}$, has n non-zero eigenvalues, and the *index* of $\boldsymbol{x}$ is the number of eigenvalues that are negative. Importantly, this number is independent of the local parametrization that defines the Hessian, so the index of a critical point is a property of the function and not of its parametrization. Intuitively, the index is the number of pairwise orthogonal directions such that f decreases when we leave $\boldsymbol{x}$ along such a direction or its opposite. For example, in Figure **1A** , we see three non-degenerate critical points, two with index 2 and one with index 1. For comparison, a degenerate critical point in Figure **1B** has one direction along which f is decreasing while f is flat in the orthogonal direction. Correspondingly, the Hessian has one negative and one zero eigenvalue.

A critical point of index n is a *maximum* since f decreases in every direction, and a critical point of index 0 is a *minimum* since f increases in every direction. All other critical

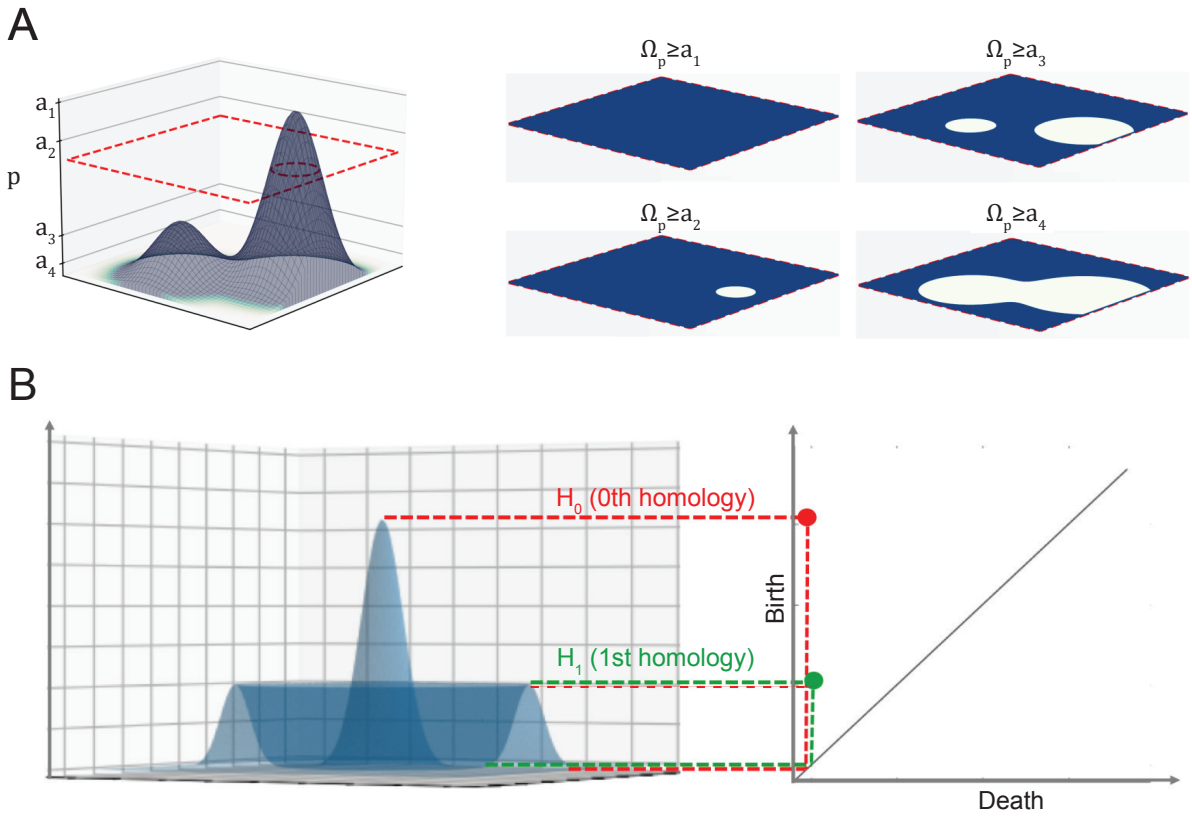


Figure 1: Two probability landscapes with illustrations of their persistent critical points. **(A)** a 2-dimensional probability landscape with two peaks separated by a saddle on the *left*, and four of its superlevel sets on the *right*. Note that the superlevel sets at different values can have different topologies, with different numbers of connected components (in *white*) and differently connected background (in *color*). **(B)** a two-dimensional probability landscape with a peak and a crater on the *left*, and the corresponding persistence diagram on the *right*. Each point in the persistence diagram represents a topological feature, with coordinates corresponding to the probability values at which the associated homology class is born and dies. This landscape has two 0-homology groups and one 1-homology group. Birth values are plotted on the vertical axis and death values on the horizontal axis, so all points lie above the diagonal.

points are *saddles*, and there are $n-1$ kinds, one for each index between 1 and $n-1$.

Remark 2 *The critical points have obvious shortcomings as representatives of the features of a real-valued function, namely that they come at different scales and thus have different but yet undefined importance. In particular, small numerical errors may introduce spurious critical points that can hardly be considered features of the function. We therefore introduce a measure of scale, or prominence of a critical point, which will allow us to focus on the important features and ignore spurious ones altogether.*

The technical term for the prominence is the *persistence* of a critical point. To define it, we introduce the *superlevel set* of f at a threshold $s \in \mathbb{R}$ as the set of points at which the

function exceeds the threshold:

$$M^s = f^{-1}[s, \infty) = \{\boldsymbol{x} \in M \mid f(\boldsymbol{x}) \geq s\}. \tag{5}$$

Note that $s \leq t$ implies $M^t \subseteq M^s$. A decreasing sequence of thresholds thus defines an increasing sequence of sets, each a superlevel set of the manifold:

$$\emptyset = M^\infty \subseteq \ldots \subseteq M^t \subseteq \ldots \subseteq M^s \subseteq \ldots \subseteq M^{-\infty} = M. \tag{6}$$

We call this sequence a *filtration* of M. Think of it as an incremental construction of the manifold. According to classic Morse theory, each critical point of index i adds an $(n-i)$-dimensional handle to the growing superlevel set [45, 46]. For $i = n$, this is a 0-handle; that is: a new connected component. For $i = n-1$, this is a 1-handle; that is: a bridge between two connected components or between different locations of the same component. In the former case, the number of components decreases by one, and we write $\beta_0^{\text{after}} = \beta_0^{\text{before}} - 1$, in which β_0 counts the components after and before the addition of the bridge. In the latter case, the number of 1-cycles (closed loops going around holes) increases by one, and we write $\beta_1^{\text{after}} = \beta_1^{\text{before}} + 1$, in which β_1 counts the loops. In the former case, we say the bridge gives *death* to a component, while in the latter case, it gives *birth* to a loop. Similarly, each critical point of index $n-i$ either gives death to an $(i-1)$-cycle or birth to an i-cycle. Indeed, there are two cases for each index, except for critical points of index n, which all give birth, and critical points of index 0, which all give death.

Using the superlevel sets in the filtration and, more specifically, their homology groups, we can pair each death-giving critical point with a birth-giving critical point, such that the latter gives birth to the cycle the former gives death to. We refer to [44] for the algebraic background needed to see that this pairing exists and is unique. Letting $(\boldsymbol{x}, \boldsymbol{y})$ be such a pair, we call $f(\boldsymbol{x}) - f(\boldsymbol{y})$ the *persistence* of the pair, or of the two critical points. This information is conveniently collected in the *persistence diagram* of f, which for each pair $(\boldsymbol{x}, \boldsymbol{y})$ contains the point $(f(\boldsymbol{x}), f(\boldsymbol{y}))$. This is a point below the diagonal because the values decrease from left to right in the filtration, so $f(\boldsymbol{x}) \geq f(\boldsymbol{y})$.

Example 1 *The two functions visualized in Figure **1A and B** are good examples. In (A), we have three critical points, two maxima and one saddle of index* 1*. The lower maximum is paired with the saddle, so there is only one finite point in the persistence diagram. The higher maximum is not paired and represented by the point whose second coordinate is* $-\infty$*, since the component does not die until the end of time.*

*In (B), there is a circle of degenerate critical points at the top of the crater surrounding the maximum in the middle. A small perturbation suffices to turn this circle of degenerate critical points into a cyclic sequence of maxima and index-*1 *saddles. Letting* $2k$ *be their number, we get* $k-1$ *maxima paired with* $k-1$ *saddles, and the corresponding* $k-1$ *points in the persistence diagram are close to the diagonal. In other words, their persistence is negligible, and we could eliminate these points altogether if we perturbed the top of the crater so it has a single maximum (its highest point) and a single saddle (its lowest point). The maximum gives birth to a component in the filtration of superlevel sets, and the saddle gives birth to a* 1*-cycle tracing out the top of the crater.*

2.3 Computing Prominence

We have used the language of smooth functions and critical points, but this was only to ease the exposition, and the probability landscapes we work with are indeed discrete. In this section, we describe computations that are based on the discrete Morse theory of cubical complexes and work for such discrete data. After introducing the necessary concepts, we will mention challenges related to computational efficiency and sketch how we overcame them.

The input to the computation is an n-dimensional array whose entries store the probabilities in the landscape. We represent this data as a filtered cubical complex, which we now define. We consider cubes of the form

$$\sigma = I_1 \times I_2 \times \ldots \times I_n, \text{ in which } I_i = \begin{cases} [a, a+1] \text{ or} \\ a, \end{cases} \tag{7}$$

with $a \in \mathbb{Z}$, for each $1 \leq i \leq n$. We call σ a *k-cube* if k of the I_i are intervals while the other $n-k$ of the I_i are integers. For example, the 0-cubes are vertices, the 1-cubes are edges, the 2-cubes are squares, and the n-cubes are full-dimensional. We call a cube, σ, a *face* of another cube, τ, if σ is contained in τ, which is only possible if the dimension of σ is at most the dimension of τ. A *cubical complex* is a collection of cubes such that all faces of every cube in the collection also belong to the collection. For example, whenever a square belongs to the complex, so do the four edges and four vertices of the square.

The n-cubes correspond to the entries of the array that contains the probabilities, and we write $f(\tau)$ for the probability stored in the entry that corresponds to the n-cube τ. We extend the function to less than full-dimensional cubes by setting $f(\sigma)$ to the maximum $f(\tau)$ over all n-cubes τ such that σ is a face of τ. With this definition, the faces of a cube have values that are at least the value of the cube, which implies that each superlevel set of f is a cubical complex, namely a subcomplex of the complex that represents the entire probability landscape. This property is essential to be able to define homology and compute persistence. Suppose we refine this filtration, so that each step adds a single cube to the superlevel set. If the dimension of the cube is k, its addition either gives birth to a k-cycle or it gives death to a $(k-1)$-cycle, and we pair up the births with the deaths. We call the cube *critical* if its value is different from that of its paired cube, and *non-critical* otherwise. The critical cubes are more important than the others, and their prominence is further quantified by persistence. The pairs of non-critical cubes satisfy the standard definitions in discrete Morse theorem [47], so f is indeed a discrete Morse function. This is the majority of cubes, and they will be ignored in our analysis of the data.

Algorithm and practical considerations. To compute persistence, we first get the matching as described in [48], identify the critical cubes, connect the pairs of critical cubes via paths that alternate between cubes of adjacent dimension, and collect the parity of these paths in a matrix, whose columns and rows correspond to the critical cubes and are sorted by their values. Persistence is extracted from the matrix using a reduction algorithm similar to Gaussian elimination [49]. Specifically, if after reduction σ corresponds to the row that contains the last non-zero entry in the column of τ, then the persistence of the pair (σ, τ) is $f(\tau) - f(\sigma)$.

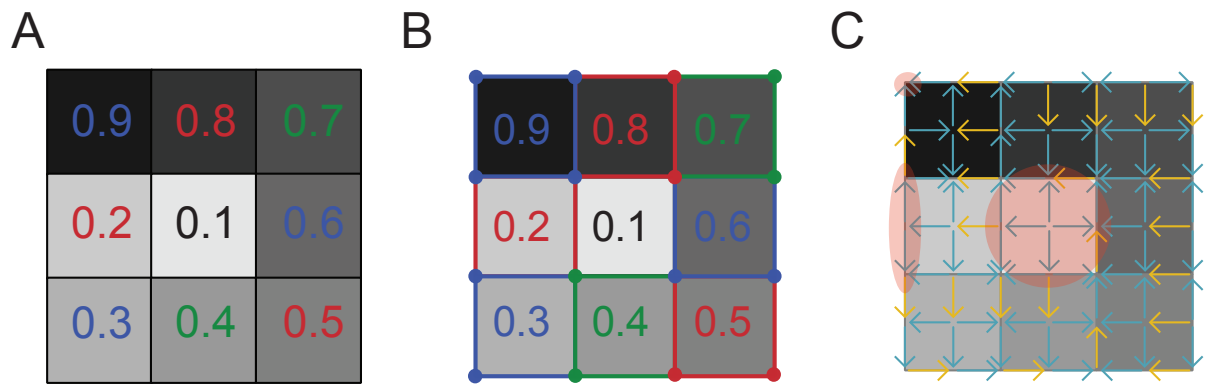


Figure 2: An Illustration of discrete Morse theory. **(A)** is a 2D probability landscape f. **(B)** turns the probability landscape into a (filtered) cubical complex, color coded to show how f extends to 1-cubes and 0-cubes. **(C)** depicts the matching between pairs of non-critical cubes using yellow arrows and boundary relationships between cubes using blue arrows. There are only 3 critical cubes (highlighted in red), one in each dimension. This number is significantly smaller than the 47 cubes in the entire cubical complex.

It is worth noting that in practice the boundary matrix is sparse and has few non-zero entries relative to the size of the entire cubical filtration. However, storing the cubical filtration in memory tends to be prohibitive for large data or in higher dimensions. Indeed, a cubical complex for a probability landscape with w^n values contains w^n n-cubes and therefore $(2w+1)^n$ cubes of dimension between 0 and n. For example, storing a cubical filtration for a dataset of size 2000^3 requires 32GB of memory, and several times more would be necessary for the whole computation. To circumvent this difficulty, the data is cut into slices, which are processed one by one and eventually stitched back together as described in [50]. Within each slice, the computations are local, and the information about critical cells and connections between them are stored. In the end, a global boundary matrix is assembled and reduced, yielding the correct persistence diagram. This setup allows for the straightforward parallelization of the computations, as each slice can be handled independently. We implemented this way of processing the data by modifying the *cubicle* software package.

Remark 3 *The standard way of computing persistence of cubical filtrations [51] uses a matrix that encodes relationships between all cubes and not just the critical ones, which tends to be prohibitively large. For input of size $m = w^n$, this matrix contains more than $m\sum_{d=0}^{n} 2d\binom{n}{d} = mn2^n$ non-zero elements. For example for an input of size 2000^3, it requires more than $1.5\,TB$ of memory, while the approach described above needs a hundred times less memory. Overall, the preprocessing approach described above gives substantial savings in both memory usage and computation time, which enables the persistence computation for large datasets on commodity hardware.*

Our ACME code can be accessed at `https://github.com/alifarhat40/bioacme`. For larger reaction networks and high-performance applications, we also provide compiled C++ executables packaged as a downloadable archive at `https://github.com/alifarhat40/bioacme/blob/master/acme_ali.zip`. The persistent homology code can be found at

https://bitbucket.org/hubwag/cubicle/ and https://github.com/alifarhat40/Topoland.git.

2.4 Interpreting Persistence Diagrams

The persistence diagram of a superlevel-set filtration is a summary of its topological features across probability values. We demonstrate how these features can be interpreted in the biological context. Later in the next section, we provide examples of how this interpretation can be used to infer pathways in the original state space that could not be detected by dimensionality reduction methods. In particular, the H_0 features and births of H_1 features provide the most striking interpretations. H_0 features identify the local maxima (at birth) and the probability thresholds (at death) at which distinct phenotype-associated regions first become connected, while the birth of H_1 features identifies the highest probability thresholds at which these connections first enable cyclic routes among two or more phenotypes. Furthermore, given a predetermined prominence threshold, since the persistence of H_0 quantifies the prominence of peaks, we may effectively filter out points below this prominence as noise.

An H_0 feature is born at a local maximum, and as the threshold is lowered, microstates in its vicinity are included in the same connected component. After filtering out noise as described above, we interpret all the remaining components as a region of microstates associated with a particular phenotype in the system. In the language of dynamical systems, a persistent H_0 component may be interpreted as a basin of attraction with the local maximum (peak) as an attractor [52]. Furthermore, the death of an H_0 feature, occurring when two previously separate components merge, corresponds to the filtration lowering to the probability value of a saddle point (of index $n-1$, for an n-dimensional probability landscape) lying on the separatrix between two basins. The separatrix acts as the boundary surface partitioning the state space into distinct basins [53, 54]. Geometrically, the saddle point represents the local minimum along this high-potential ridgeline, effectively the 'mountain pass' that offers a path for transitions between the two phenotypes. This saddle represents the microstate the system must pass through, and the probability level (equal to the probability value of the saddle) it must attain, in order to cross into a neighboring phenotype.

Topologically, the death of an H_0 feature and the birth of an H_1 feature are both triggered by the addition of a 1-handle in the superlevel set. This entails the addition of saddle points which enable new paths between pairs of peaks. While H_0 deaths record the first connection between corresponding attraction basins, the birth of an H_1 feature occurs when such connections first close up to form a non-trivial loop. We therefore emphasize the following interpretation: the birth value of H_1 marks the highest probability threshold at which cycle-mediated accessibility is established among the phenotype-associated regions. Above this threshold, this cyclic accessibility does not yet exist. In this way, H_0 features identify the local maxima (at birth) and the probability thresholds (at death) at which distinct phenotype-associated regions first become connected, while the birth of H_1 features identifies the highest probability thresholds at which these connections first enable cyclic routes among two or more phenotypes. Such routes allow movement without detouring through the attraction basins of additional phenotypes. Compared to higher-degree homology, H_0 and

H_1 therefore provide the most directly interpretable probability thresholds for phenotype-to-phenotype accessibility in the filtered landscape.

Higher-degree persistent homology likewise identifies probability thresholds at which higher-dimensional topological features emerge in the filtered landscape. For instance, the birth of an H_2 feature also corresponds to a probability threshold where non-trivial loops between phenotype associated regions exist in the superlevel set, since these loops may exist on nontrivial 2-dimensional cycles. This is, however, met with the slight nuance that this probability threshold might not be the highest that enables this accessibility. These features are still useful, for example, in cases where no H_1 feature appears on the persistence diagram.

3 Biological Examples

The probability landscape of a biochemical reaction network evolves in time before reaching a steady state [55]. Differences in reaction rates such as synthesis, degradation, and binding, can tilt the probability landscape, leading to dramatic changes in the dynamic behavior of the individual components of the network. Such changes may be exploited for desired behavior, with implications for a more fundamental understanding of biological systems and thereby treatment of diseases [27]. Here we present three examples, where topological features can be identified and their properties quantified using persistent homology. These include the presence of phenotypical regions (which show up as components measured by degree-0 homology), and the probability values above which cycle-mediated accessibility exist between phenotypical regions (indicated by higher degree homology).

3.1 Feedback Activation-Inhibition Networks

The interaction between two genes and their products can be broadly classified as either *activation* or *inhibition*. An activating gene product promotes the activity of another gene, whereas an inhibiting gene product suppresses it. When two genes regulate each other, their interactions may be mutually activating, mutually inhibiting, or asymmetric, with one activating and the other inhibiting. The latter configuration defines a *feedback activation-inhibition network*, a motif that is widespread in biological systems and capable of generating rich nonlinear dynamics.

From a theoretical perspective, feedback activation-inhibition networks are a central object of study in chemical reaction network theory, where their structure has been shown to support multistability and robust dynamical behavior independent of specific parameter choices [56].

A classic example is the bistable autophosphorylating kinase described by Lisman [57], where positive feedback from kinase activation is counterbalanced by inhibitory phosphatase activity, producing two stable steady states. This illustrates how feedback activation-inhibition motifs can generate robust, switch-like behavior in biochemical regulatory systems. More broadly, such feedback-driven bistability and switch-like responses have been extensively studied in signaling and regulatory networks, where they underlie decisive cellular transi-

tions and persistent cellular states [58].

We list several additional examples of such networks below:

- In the metabolic network, AMP-activated protein kinase (AMPK)—which is known as a cellular energy tensor—inhibits the activity of noxROS—which plays an important role in metabolic and inflammatory signaling [59, 60, 61, 62, 63]. In addition, noxROS promotes the activity of AMPK. Therefore, AMPK and noxROS engage in a feedback activation-inhibition sub-network.

- In the epithelial-mesenchymal transition (EMT) network, feedback activation-inhibition sub-networks are present between some of the genes. For example, MDM2—which plays and important role in cell growth—inhibits P53—which plays important role in cell growth and cell division[64, 65, 66]. P53, on the other hand, activates MDM2 [67]. Therefore, MDM2 and P53 together form a feedback activation-inhibition sub-network.

- In circadian rhythm, there are inhibitor as well as activator clock genes [68, 69]. Researchers suggest a feedback loop for the circadian rhythm, in which activating genes, such as BMAL1, NPAS2, and CLOCK interact with inhibiting genes, such as PER, CRY, NR1D, and ROR [70, 71]. They form a feedback activation-inhibition sub-network, which controls the circadian rhythm. It is known that the malfunction of the activating or inhibiting genes may associate with mood disorders, such as major depressive disorder (MDD) and biopolar disorder (BD) [68, 69, 71, 70].

At the molecular level, activation and inhibition in gene regulatory networks are most commonly mediated by transcription factors that bind to regulatory regions of DNA. Many transcription factors function as homodimers or higher-order oligomers, a property that enhances binding specificity, cooperativity, and regulatory control over target genes. Upon binding, transcription factor dimers can either promote transcription by recruiting the transcriptional machinery or repress transcription by blocking access or recruiting inhibitory cofactors. Such dimerization-dependent regulation provides a natural biochemical mechanism for implementing activation–inhibition interactions and introduces intrinsic nonlinearity into gene regulatory dynamics [72, 73, 74]. Because transcription factor copy numbers and DNA-binding events are often sparse and discrete, stochastic fluctuations can dominate system dynamics, enabling noise-induced switching between regulatory states. Such effects are inherently probabilistic and are therefore not represented in deterministic ordinary differential equation descriptions, motivating a stochastic modeling framework.

Consider a feedback activation-inhibition network that consists of the inhibiting gene, $GeneX$, the activating gene, $GeneY$, and the transcription factors, X and Y, they encode. The homodimer of transcription factor X can bind to $GeneY$ and suppress $GeneY$'s actions, while the homodimer of transcription factor Y binds to $GeneX$ and promotes its activities. All together, the network consists of six molecular species as shown in Figure **3A**. In addition,

there are nine reactions:

$$GeneX \xrightarrow{ks_x} GeneX + X; \qquad GeneY \xrightarrow{ks_y} GeneY + Y; \tag{8}$$

$$2Y + GeneX \xrightarrow{b_1} BGeneY; \qquad 2X + GeneY \xrightarrow{b_2} BGeneY; \tag{9}$$

$$BGeneX \xrightarrow{u} 2Y + GeneX; \qquad BGeneY \xrightarrow{u} 2X + GeneY; \tag{10}$$

$$X \xrightarrow{kd_x} \emptyset; \quad Y \xrightarrow{kd_y} \emptyset; \quad BGeneX \xrightarrow{ks} BGeneX + X, \tag{11}$$

in which we set $ks_x = 5$, $ks_y = k_s = 50$, $b_1 = 2.5 \cdot 10^{-4}$, $b_2 = 10^{-4}$, $u = 0.1$, $kd_x = kd_y = 1$. All the reaction rate constants are per second. "BGeneX" means GeneX is bound by two activating Y transcription factor proteins. "BGeneY" means GeneY is bound by two inhibiting X transcription factors.

Because there is a single copy of $GeneX$ and a single copy of $GeneY$ at the DNA level, we represent gene binding states using a binary encoding, with 0 denoting a bound gene and 1 denoting an unbound gene. The state space of the system is four-dimensional, consisting of the variables X, Y, $GeneX$, and $GeneY$. The corresponding 4D probability landscape, illustrated in Figure **3B**, is computed using the ACME method [15, 16].

The analysis of the topological structure of the probability landscape is important as it characterizes the phenotypical behavior of the network and may provide the pathways of dynamic switching between different behaviors. The peaks represent the centers of phenotypical behaviors in the state space, and cycles represent the high probability pathways where the dynamic switching may occur. Here, we identify the most prominent topological features with persistent homology and visualize them geometrically.

We examine the probability landscape of the feedback activation-inhibition network obtained by solving its corresponding dCME; see Figure **3C**. There are four peaks in the landscape of the network; see the red dots in Figure **3C**, which mark the peaks, and the blue dots, which mark the saddles.

- The most prominent peak is b_1, which gives birth to the component marked as point 1 in the persistence diagram shown in Figure **3D**. It is most prominent because the persistence of point 1 (the distance of the point from the diagonal) is larger than that of any other point in the diagram. This peak corresponds to the phenotypic behavior in which the inhibiting $GeneX$ is on and the activating $GeneY$ is off, indicating that the system most frequently occupies this on-off configuration (Figure **3B**).

- The second most prominent peak is b_2, which gives birth to the component marked as point 2 in the persistence diagram. It corresponds to both genes being off.

- Two additional prominent peaks, b_3 and b_4, correspond to the phenotypic behaviors in which both genes are on and in which the inhibiting gene is off and the activating gene is on, respectively. These two peaks exhibit comparable probability mass and persistence, indicating that the system has approximately equal likelihood of occupying either state (Figure **3B**).

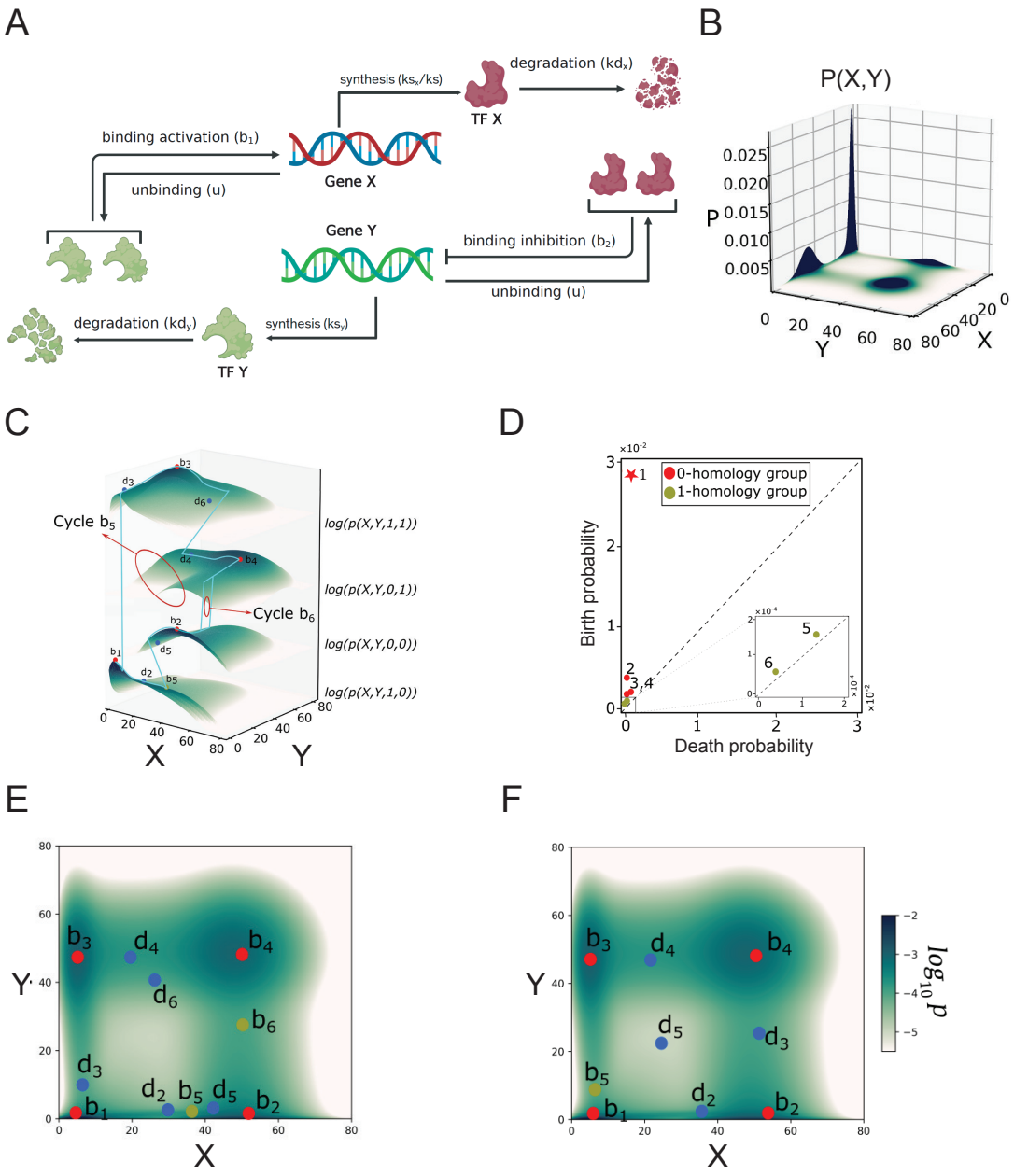


Figure 3: Feedback activation-inhibition network and its probability landscape. **(A)** Schematic of the network, consisting of the inhibiting *GeneX*, the activating *GeneY*, and the transcription factors X and Y that they encode. Homodimers of X inhibit *GeneY*, while homodimers of Y activate *GeneX*. **(B)** The full four-dimensional probability landscape associated with the feedback activation-inhibition network. the x- and y-axes show the counts for species X and Y, respectively. **(C)** Stratified 3D surface visualization of the probability landscape. Because the landscape resides in a 4D state space, it is stratified into four 2D subspaces to enable faithful visualization. **(D)** Persistence diagram of the landscape, showing four prominent probability peaks (labeled 1-4) and two 1-cycles (labeled 5 and 6). Each topological feature is mapped to its corresponding 2D subspace in **(C)**. **(E)** Topological features computed in the original 4D space. For visualization, these features are overlaid on a 2D representation obtained by stacking the four stratified probability landscapes shown in (C). **(F)** Topological features computed after projecting the probability landscape into a 2D space.

Our results show that there are also two loops in the probability landscape, corresponding to the green dots in the persistence diagram of Figure **3D**. The saddles that give birth to these loops are labeled b_5 and b_6 in Figure **3C**. The most prominent loop corresponds to b_5 and connect the four prominent peaks. It shows that there is a high probability pathway between the phenotypes of the system. The loop given birth to by the saddle b_5 connects peak b_1 to b_2, b_2 to b_4, b_4 to b_3, and b_3 to b_1; see the cyan-colored pathway in Figure **3C**. Our results suggest that in the dynamic switching of the states through this loop, the pathway is such that only one gene can change its status at any one time. They also suggest that there is active dynamic switching between the activating *GeneY* Off and On, when the inhibiting *GeneX* is in the Off state. This is represented by the loop given birth to by the saddle b_6.

To compare the commonly used projected probability landscape with the original 4D landscape, we compute the persistence diagrams for both representations. In the original 4D setting, four prominent probability peaks are present, corresponding to b_1, b_2, b_3, and b_4 in Figure **3E**. These peaks are also visible in the projected 2D landscape shown in Figure **3F**. However, the saddles where these peaks merge differ between the original 4D landscape and its 2D projection. This discrepancy indicates that projection substantially distorts the topology of the underlying probability landscape. Moreover, the projected landscape in Figure **3F** suggests the presence of only a single dynamical switching pathway between the four peaks, whereas it fails to capture the loop passing through the saddle associated with b_6 in the original 4D landscape. Existing projection methods do not guarantee preservation of topological structure [39].

3.2 Repressilator Network

Oscillatory behavior is ubiquitous in biological networks, which include many-gene interaction networks ranging from circadian rhythms to cell cycle and embryonic stem cell networks [59, 60, 61, 62, 63, 75, 76]. Gene interaction networks reveal essential information about biological behavior of various living cells [77]. Despite vigorous efforts to quantitatively analyze different gene regulatory networks, the enumeration and characterization of their topological features remains poorly understood. On the other hand, living cells often include species with very low copy numbers, including transcription factors, mRNA, and regulatory molecules, so stochasticity plays an important role in the modeling of such networks.

Among the work on stochastic networks, there have been efforts to model oscillation in the copy numbers by synthetic oscillatory networks, such as repressilator network [78, 79]. Their modeling helps in the reverse engineering of the behavior of biological systems, but also in the modeling of new behaviors in synthetic biology. The high-dimensionality of such networks makes them difficult to analyze by traditional methods of dimension reduction or projection to low dimensions.

A repressilator network consists of three or more species, each represented as a node [79, 80, 81]. Each species may activate or inhibit other species, and the combination of these influences may cause oscillation in the copy number of the species. In addition, there are different numbers of phenotypical behaviors, based on the number of nodes in the net-

work. There may also be different kinds of high probability pathways between these behaviors. By stochastic modeling of such network and analyzing the topological features of the high-dimensional probability landscape, we may reveal the behavior, such as the centers of phenotypical behavior, and the loops connecting these centers. In this section, we model a 3-species and a 4-species repressilator network stochastically, and analyze the topological features of these two networks using our new approach.

3.2.1 Three-Species Repressilator

We consider a network with three transcription factors (TFs), A, B, and C. Each transcription factor is synthesized into the system with the synthesis rate θ from an outside source, and degraded with the degradation rate k_d. In addition, transcription factor A is inhibited by C, B is inhibited by A, and C is inhibited by B. The reaction rates of these inhibitions, r_A, r_B, and r_C, are modeled as Hill functions, which depend on the copy number of the inhibiting transcription factor (i.e. r_A depends on the copy number of C, etc.), such that the inhibition happens only if the copy number of the inhibiting transcription factor is more than a threshold, γ. See Figure **4A** for a schematic of the reaction network. There are nine reactions in the 3D state space:

$$\emptyset \xrightarrow{\theta} A; \qquad \emptyset \xrightarrow{\theta} B; \qquad \emptyset \xrightarrow{\theta} C; \tag{12}$$

$$A \xrightarrow{k_d} \emptyset; \qquad B \xrightarrow{k_d} \emptyset; \qquad C \xrightarrow{k_d} \emptyset; \tag{13}$$

$$A + C \xrightarrow{r_A} \emptyset + C; \quad B + A \xrightarrow{r_B} \emptyset + A; \quad C + B \xrightarrow{r_C} \emptyset + B, \tag{14}$$

where we define $h_X = n_X^\eta/(\gamma^\eta + n_X^\eta)$ and set $r_A = h_C$, $r_B = h_A$, and $r_C = h_B$. Here, n_X denotes the copy number (molecule count) of species X in the current microstate (n_A, n_B, n_C). We used the ACME method to compute the 3D probability landscape for the specific parameters $k_d = 10$, $\theta = 35$, $\gamma = 20$, and $\eta = 50$. Figure 4 (D) shows the 3D probability landscape for different counts of transcription factors A, B, and C = 3, 17, 27, and 35 and the projected probability landscape onto the A-B plane. We examine the topological features of this 3D probability landscape with persistent homology. There are three peaks giving birth to connected components marked as points 1, 2, and 3 in the persistent diagram of Figure **4B**. Each peak corresponds to the activated state of one transcription factor. Red points in Figure **4D** show the peaks in the A-B plane, and the green point shows the saddle that gives birth to the loop projected to this plane. Peak 1 is at $(A, B, C) = (3, 17, 27)$, which corresponds to C being active, peak 2 is at $(A, B, C) = (17, 27, 3)$, which corresponds to B being active, and peak 3 is at $(A, B, C) = (27, 3, 17)$, which corresponds to A being active. The use of persistent homology was crucial in finding these phenotypical behaviors.

In addition, there is a loop in the probability surface, given birth to by the saddle labeled 4 and located at $(A, B, C) = (5, 21, 14)$ in Figure **4B**. This loop indicates the existence of a high probability pathway in the landscape; that is: a non-equilibrium circuit in the state space. To identify this pathway, we used the method mentioned in Section 2 and find all microstates that belong to the loop. The black closed curve in Figure **4C** shows the loop in

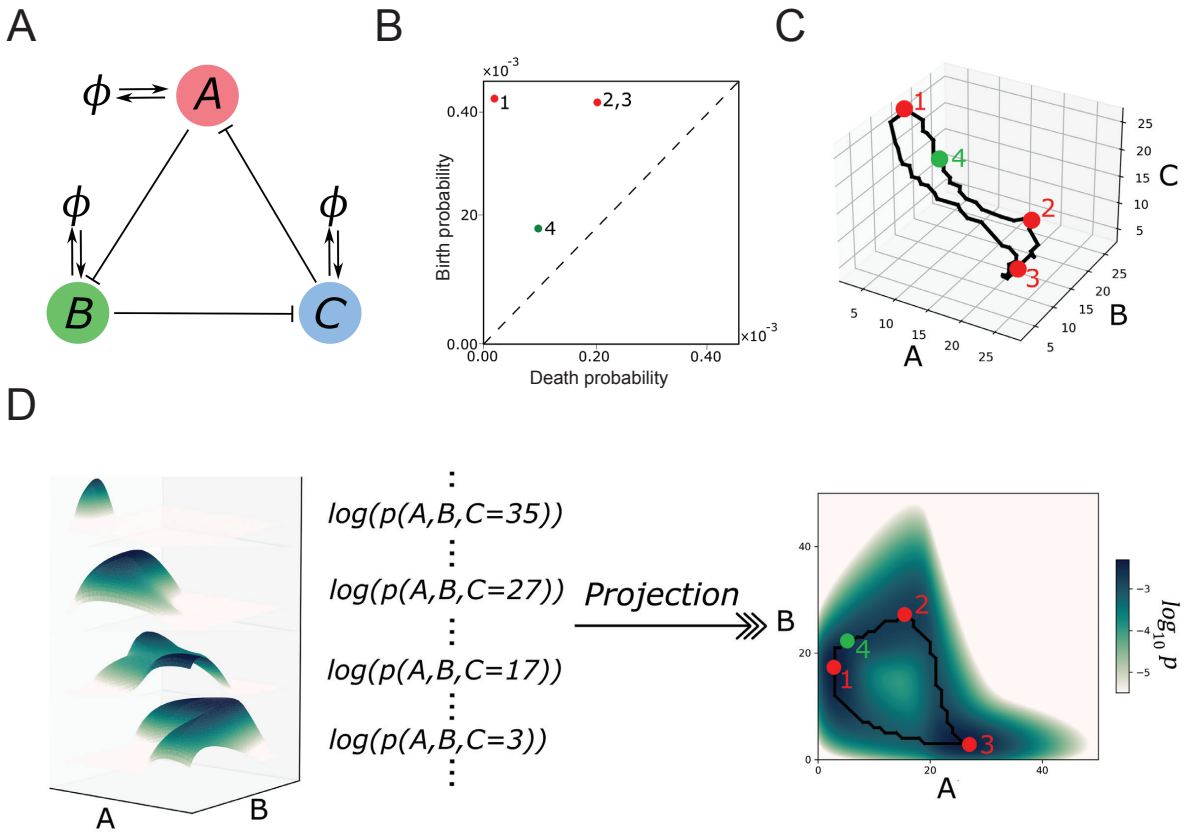


Figure 4: Persistent homology analysis of the 3-species repressilator. **(A)** Reaction network diagram consisting of the three transcription factors, A, B, and C. **(B)** Persistence diagram of the probability landscape, showing three prominent probability peaks (labeled 1, 2, and 3 in red) and a single 1-cycle born at the saddle point labeled 4 (in green). **(C)** The corresponding 1-cycle represented as a high-probability pathway in A–B–C space, connecting the three peaks. This cycle is projected onto the probability landscape shown in **(D)**. **(D)** Probability landscape of the 3-species repressilator for parameters $k_d = 10.0$, $\theta = 35$, $\gamma = 20$, and $\eta = 50$. The three-dimensional probability landscape is shown in the left panel, and for visualization purposes it is projected onto the A–B plane in the right panel. The x- and y-axes in panels C and D represent the molecule counts for each species.

the 3D A-B-C space, together with the locations where the peaks and the loop are born. Our results show that in the 3-species repressilator, there is a high probability pathway connecting the three phenotypical behaviors to each other. Most likely, the dynamic switching between these behaviors happens along this loop.

3.2.2 Four-Species Repressilator

Repressilator networks may consist of more than three nodes, in particular when there are more than three species interacting with each other. With the number of nodes the dimension of the state space increases, so the probability landscape is even more difficult to be analyzed with traditional methods, such as projection or other dimension reduction techniques. Here we design a 4-species repressilator network and show how persistent homology can help understanding the homological features of this more complicated system.

Assume a system with four transcription factors (TFs), where each TF can be synthesized with the synthesize rate θ from an outside source, and each TF is degraded with the degradation rate k_d. In addition, each gene is inhibited by two other transcription factors when both have a sufficiently high copy number. The reaction rates of the inhibitions, r_A, r_B, r_C, and r_D, depend on the copy numbers of two inhibiting species (i.e. r_A depends on the copy numbers of B and C, etc.). We therefore model the reaction rates as products of Hill functions. Network with small copy numbers of species can be highly stochastic, so the stochastic analysis of this network is vitally important. See Figure **5A** for the reaction network schematic with twelve reactions as given here:

$$\emptyset \xrightarrow{\theta} A; \quad \emptyset \xrightarrow{\theta} B; \quad \emptyset \xrightarrow{\theta} C; \quad \emptyset \xrightarrow{\theta} D; \tag{15}$$

$$A \xrightarrow{k_d} \emptyset; \quad B \xrightarrow{k_d} \emptyset; \quad C \xrightarrow{k_d} \emptyset; \quad D \xrightarrow{k_d} \emptyset; \tag{16}$$

$$A + B + C \xrightarrow{r_A} \emptyset + B + C; \qquad B + C + D \xrightarrow{r_B} \emptyset + C + D; \tag{17}$$

$$C + A + D \xrightarrow{r_C} \emptyset + A + D; \qquad D + A + B \xrightarrow{r_D} \emptyset + A + B, \tag{18}$$

where we recall $h_X = n_X^\eta/(\gamma^\eta + n_X^\eta)$ and set $r_A = h_C h_B$, $r_B = h_C h_D$, $r_C = h_A h_D$, $r_D = h_A h_B$. For the parameter values $k_d = 10.0$, $\theta = 35$, $\gamma = 20$, and $\eta = 50$, we compute the probability landscape $\boldsymbol{p}(A, B, C, D)$ over the 4D state space using the ACME method; see Figure **5D**. We then identify the most prominent topological features of this landscape using persistent homology (Figure **5B**) and map these features back to the underlying microstates that give rise to their birth and death events, indicated by the marked points in Figure **5D**.

The results in Figure **5B** show that the network has four major peaks, labeled $1, 2, 3, 4$, and located at $(A, B, C, D) = (16, 1, 31, 27)$, $(2, 31, 27, 16)$, $(32, 27, 16, 7)$, and $(28, 16, 3, 22)$. The locations of the peaks projected to the A-B plane are shown in the right panel of Figure **5D**. In addition, there is one 2-cycle, born at point 5 at location $(29, 22, 4, 13)$ (Figure **5B**), which is drawn as a blue point in Figure **5D**.

4-dimensional 2-Cycle illustration. As shown in Figures **5B** and **5D**, the 4D probability landscape of the repressilator network contains a single 2-cycle that connects all four phenotypic behaviors. Figure **5C** illustrates the 2-cycle in 4D A-B-C-D space on the left, with color coded copy number of D. It shows that there is a 2-cycle that passes through all four peaks. Our results show that the 2-cycle is a high-dimensional pathway between the four phenotypical behaviors of the probability landscape.

From the projection of the probability landscape to the A-B plane, neither the peaks nor the cycle connecting them can be identified. Figure **5C** shows the projected 2-cycle in the A-B plane on the right. The red points mark the location of peaks, and the blue point marks the location that gives birth to the 2-cycle in the original 4-dimensional space. Furthermore, the black lines indicate the states in the A-B plane surrounding this 2-homology void. There are no 1-homology groups or 1-cycles in this network.

This analysis illustrates again the difficulties of gaining valid topological information from projections to planes, and stresses the importance of developing tools that are able to do the

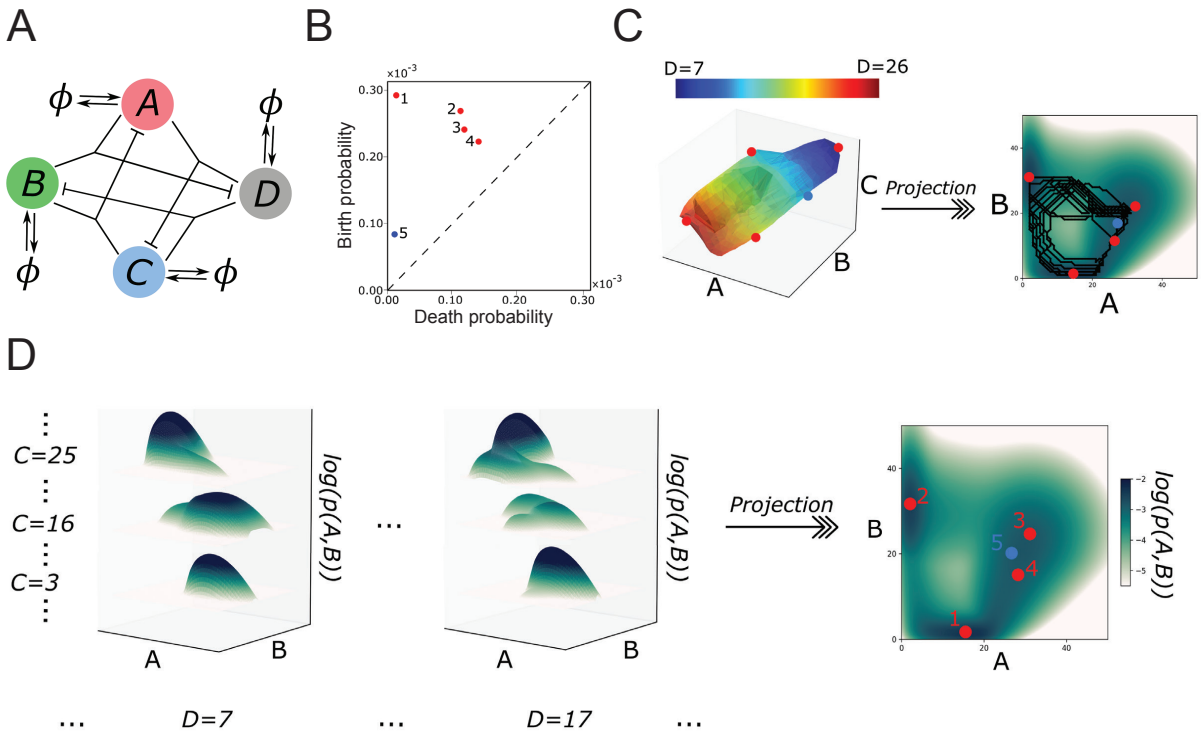


Figure 5: Persistent homology analysis of the 4-species repressilator network. **(A)** Reaction network diagram of the 4-species repressilator. **(B)** Persistence diagram recording the birth and death probability values of connected components and higher-dimensional cycles. The diagram reveals four prominent probability peaks (labeled 1–4 in red) and a single 2-cycle (void), labeled 5 in blue, in the full 4D probability landscape. **(C)** Illustration of the 2-cycle (void) traversing the four phenotypic behaviors (shown as red points), with the critical location at which the 2-cycle is born indicated in blue. Projecting the probability landscape and the cycle onto the A–B plane obscures the underlying topology and pathway structure. **(D)** Probability landscape of the 4-species repressilator for parameters $k_d = 10.0$, $\theta = 35$, $\gamma = 20$, and $\eta = 50$. Red and blue markers indicate the locations of the probability peaks and the critical location associated with the birth of the 2-cycle, respectively. The x- and y-axes in panels C and D represent the molecule counts for each species.

analysis in the original, higher-dimensional space.

4 Discussion

In the above we introduced a new framework for characterizing the topological structure of high-dimensional probability landscapes arising from biological networks. Starting with the critical points, as defined in classic Morse theory [46], we use persistent homology [44] to quantify their prominence, which allows us to focus on the important topography of the landscape. The approach does not replace traditional methods, but provides structural information that traditional methods are unable to detect.

Using persistent homology, we identify the locations of birth and death of features within high-dimensional state spaces and compute their global persistence. The current implementation scales well with dimension, which makes it a practical tool for investigating the

topological features of high-dimensional networks without dimensionality reduction. For the first time, we introduce a computational approach to characterize loops and higher-dimensional analogs in high-dimensional probability landscapes, which can be interpreted as high-dimensional pathways in non-equilibrium steady-state probability density landscapes.

We apply the methodology to three types of biological networks: feedback activation-inhibition networks, three-node repressilators, and four-node repressilators. In each case, we identify phenotypic behaviors and the most likely pathways connecting them. The state spaces in these examples are three- and four-dimensional, and we detect 0-, 1-, and 2-dimensional features with meaningful biological interpretations. More broadly, this framework provides a practical way to compare gene-circuit designs by revealing which stable behaviors exist and how the system most likely switches between them.

We stress that most biological networks of interest are higher-dimensional, and we hope the work reported here is a step toward similar analyses of their probability landscapes. Current bottlenecks include computing the landscape by solving the discrete Chemical Master Equation and scaling persistent homology software. We hope that this work will motivate further development of both components.

In closing, we would like to emphasize that the persistent homology framework developed here is not restricted to biological networks, but is applicable to a wide class of high-dimensional real-valued functions, such as probability distributions, energy landscapes, and frequency spectra. Thus, this methodology should find wide applicability across domains where identifying prominent features and their organization in high-dimensional landscapes is critical for interpretation.

ACKNOWLEDGMENT

This work was supported by NIH Grant R35-GM127084 awarded to J. L and A. F. was supported by NIH Grant F30HL176078.

References

[1] Adam Arkin, John Ross, and Harley H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. *Genetics*, 1998.

[2] M. Samoilov, S. Plyasunov, and A. P. Arkin. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proceedings of the National Academy of Sciences*, 2005.

[3] P Ao. Metabolic network modelling: Including stochastic effects. *Computers & chemical engineering*, 29(11-12):2297–2303, 2005.

[4] Erel Levine and Terence Hwa. Stochastic fluctuations in metabolic pathways. *Proceedings of the National Academy of Sciences*, 104(22):9224–9229, 2007.

[5] J. J. Tyson, K. C. Chen, and B. Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current Opinion in Cell Biology*, 15(2):221–231, 2003.

[6] Makio Ogawa. Hemopoietic stem cells: stochastic differentiation and humoral control of proliferation., 1989.

[7] Max Delbrück. Statistical fluctuations in autocatalytic reactions. *The Journal of Chemical Physics*, 8(1):120–124, 1940.

[8] Donald A McQuarrie. Stochastic approach to chemical kinetics. *Journal of applied probability*, 4(3):413–478, 1967.

[9] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 1976.

[10] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[11] Hong Qian and Daniel A Beard. Thermodynamics of stoichiometric biochemical networks in living systems far from equilibrium. *Biophysical chemistry*, 114(2-3):213–220, 2005.

[12] Paola Lecca. Stochastic chemical kinetics. *Biophysical reviews*, 5(4):323–345, 2013.

[13] AG Fredrickson. Stochastic triangular reactions. *Chemical engineering science*, 21(8):687–691, 1966.

[14] Youfang Cao, Hsiao-Mei Lu, and Jie Liang. Probability landscape of heritable and robust epigenetic state of lysogeny in phage lambda. *Proceedings of the National Academy of Sciences*, 107(43):18445–18450, 2010.

[15] Youfang Cao, Anna Terebus, and Jie Liang. Accurate Chemical Master Equation solution using multi-finite buffers. *Multiscale Modeling & Simulation*, 2016.

[16] Youfang Cao, Anna Terebus, and Jie Liang. State space truncation with quantified errors for accurate solutions to discrete Chemical Master Equation. *Bulletin of Mathematical Biology*, 2016.

[17] LJ Schaad. The Monte Carlo integration of rate equations. *Journal of the American Chemical Society*, 85(22):3588–3592, 1963.

[18] H DEGN. A compiler for digital computation in chemical kinetics and its application to oscillatory reaction schemes. *Acta Chem. Scand*, 21(3), 1967.

[19] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, 2006.

[20] Youfang Cao and Jie Liang. Optimal enumeration of state space of finitely buffered stochastic molecular networks and exact computation of steady state landscape probability. *BMC Systems Biology*, 2(1):1–13, 2008.

[21] Roger B Sidje. Expokit: A software package for computing matrix exponentials. *ACM Transactions on Mathematical Software (TOMS)*, 24(1):130–156, 1998.

[22] Trang Dinh and Roger B Sidje. An adaptive solution to the chemical master equation using quantized tensor trains with sliding windows. *Physical Biology*, 17(6):065014, 2020.

[23] Thomas A Henzinger, Maria Mateescu, and Verena Wolf. Sliding window abstraction for infinite markov chains. In *International Conference on Computer Aided Verification*, pages 337–352. Springer, 2009.

[24] Shev MacNamara, Kevin Burrage, and Roger B Sidje. Multiscale modeling of chemical kinetics via the master equation. *Multiscale Modeling & Simulation*, 6(4):1146–1168, 2008.

[25] Vladimir Kazeev, Mustafa Khammash, Michael Nip, and Christoph Schwab. Direct solution of the chemical master equation using quantized tensor trains. *PLoS computational biology*, 10(3):e1003359, 2014.

[26] Ion Gabriel Ion, Christian Wildner, Dimitrios Loukrezis, Heinz Koeppl, and Herbert De Gersem. Tensor-train approximation of the chemical master equation and its application for parameter inference. *The Journal of Chemical Physics*, 155(3):034102, 2021.

[27] Youfang Cao, Xue Lei, Ruy M Ribeiro, Alan S Perelson, and Jie Liang. Probabilistic control of hiv latency and transactivation by the tat gene circuit. *Proceedings of the National Academy of Sciences*, 115(49):12453–12458, 2018.

[28] Youfang Cao, Hsiao-Mei Lu, and Jie Liang. Stochastic probability landscape model for switching efficiency, robustness, and differential threshold for induction of genetic circuit in phage λ. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 611–614. IEEE, 2008.

[29] Youfang Cao, Hsiao-Mei Lu, and Jie Liang. Probability landscape of heritable and robust epigenetic state of lysogeny in phage lambda. *Proceedings of the National Academy of Sciences*, 107(43):18445–18450, 2010.

[30] Anna Terebus, Chun Liu, and Jie Liang. Discrete flux and velocity fields of probability and their global maps in reaction systems. *The Journal of Chemical Physics*, 149(18):185101, 2018.

[31] Anna Terebus, Chun Liu, and Jie Liang. Discrete and continuous models of probability flux of switching dynamics: Uncovering stochastic oscillations in a toggle-switch system. *The Journal of Chemical Physics*, 151(18):185104, 2019.

[32] Farid Manuchehrfar, Wei Tian, Tom Chou, and Jie Liang. Evolution of coagulation-fragmentation stochastic processes using accurate chemical master equation approach. *Communications in information and systems*, 19(1):37, 2019.

[33] Anna Terebus, Farid Manuchehrfar, Youfang Cao, and Jie Liang. Exact probability landscapes of stochastic phenotype switching in feed-forward loops: Phase diagrams of multimodality. *Frontiers in Genetics*, 12:645640, 2021.

[34] Pengcheng Bu, Lihua Wang, Kai-Yuan Chen, Tara Srinivasan, Preetish Kadur Lakshminarasimha Murthy, Kuei-Ling Tung, Anastasia Kristine Varanko, Huanhuan Joyce Chen, Yiwei Ai, Sarah King, Steven M. Lipkin, and Xiling Shen. A mir-34a-numb feedforward loop triggered by inflammation regulates asymmetric stem cell division in intestine and colon cancer. *Cell Stem Cell*, 18(2):189–202, 2016.

[35] Anupam Dey and Debashis Barik. Potential landscapes, bifurcations, and robustness of tristable networks. *ACS Synthetic Biology*, 10(2):391–401, 2021. PMID: 33533596.

[36] Farid Manuchehrfar, Huiyu Li, Wei Tian, Ao Ma, and Jie Liang. Exact topology of the dynamic probability surface of an activated process by persistent homology. *The Journal of Physical Chemistry B*, 125(18):4667–4680, 2021.

[37] Daniel T Gillespie et al. Stochastic simulation of chemical kinetics. *Annual review of physical chemistry*, 58(1):35–55, 2007.

[38] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[39] Farid Manuchehrfar, Huiyu Li, Wei Tian, Ao Ma, and Jie Liang. Exact topology of the dynamic probability surface of an activated process by persistent homology. *The Journal of Physical Chemistry B*, 125(18):4667–4680, 2021.

[40] P Ao. Potential in stochastic differential equations: novel construction. *Journal of Physics A: Mathematical and General*, 37(3):L25–L30, jan 2004.

[41] J. Wang, L. Xu, and E. Wang. Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations. *Proc. Natl. Acad. Sci. U. S. A.*, 105(34):12271—12276, 2008.

[42] J. Szücs M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer Science & Business Media, London, 2012.

[43] Peijie Zhou and Tiejun Li. Construction of the landscape for multi-stable systems: Potential landscape, quasi-potential, a-type integral and beyond. *The Journal of Chemical Physics*, 144(9):094109, 2016.

[44] Herbert Edelsbrunner and John L Harer. *Computational Topology: an Introduction*. American Mathematical Society, Providence, Rhode Island, 2010.

[45] Yukio Matsumoto. *An Introduction to Morse Theory*. American Mathematical Society, Providence, Rhode Island, 2002.

[46] John Milnor. *Morse Theory*. Princeton University Press, Princeton, New Jersey, 1963.

[47] Robin Forman. A user's guide to discrete morse theory. *Séminaire Lotharingien de Combinatoire [electronic only]*, 48:B48c–35, 2002.

[48] Vanessa Robins, Peter John Wood, and Adrian P Sheppard. Theory and algorithms for constructing discrete morse complexes from grayscale digital images. *IEEE Transactions on pattern analysis and machine intelligence*, 33(8):1646–1658, 2011.

[49] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 2002.

[50] Hubert Wagner. Slice, simplify and stitch: topology-preserving simplification scheme for massive voxel data. In Erin W. Chambers and Joachim Gudmundsson, editors, *39th International Symposium on Computational Geometry (SoCG 2023)*, volume 258 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 60:1–60:16, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[51] Hubert Wagner, Chao Chen, and Erald Vuçini. Efficient computation of persistent homology for cubical data. In *Topological methods in data analysis and visualization II*, pages 91–106. Springer, 2012.

[52] Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Chapman and Hall/CRC, New York, 3 edition, January 2024.

[53] Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260 of *Grundlehren Der Mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2012.

[54] Ben C. Nolting and Karen C. Abbott. Balls, cups, and quasi-potentials: Quantifying stability in stochastic systems. *Ecology*, 97(4):850–864, 2016.

[55] Martin Feinberg. The existence and uniqueness of steady states for a class of chemical reaction networks. *Archive for Rational Mechanics and Analysis*, 132(4):311–370, December 1995.

[56] Martin Feinberg. *Foundations of Chemical Reaction Network Theory*, volume 202 of *Applied Mathematical Sciences*. Springer International Publishing, Cham, 2019.

[57] J E Lisman. A mechanism for memory storage insensitive to molecular turnover: a bistable autophosphorylating kinase. *Proceedings of the National Academy of Sciences*, 82(9):3055–3057, 1985.

[58] James E. Ferrell, Jr., Tony Yu-Chen Tsai, and Qiong Yang. Modeling the Cell Cycle: Why Do Certain Circuits Oscillate? *Cell*, 144(6):874–885, March 2011.

[59] Annette C Dolphin. Calcium channel auxiliary $\alpha 2\delta$ and β subunits: trafficking and one step beyond. *Nature Reviews Neuroscience*, 13(8):542–555, 2012.

[60] D Grahame Hardie. Amp-activated protein kinase: maintaining energy homeostasis at the cellular and whole body levels. *Annual review of nutrition*, 34:31, 2014.

[61] Steven J Forrester, Daniel S Kikuchi, Marina S Hernandes, Qian Xu, and Kathy K Griendling. Reactive oxygen species in metabolic and inflammatory signaling. *Circulation research*, 122(6):877–902, 2018.

[62] Linglin Yu. *Gene Network Modeling of Cancer Metabolism*. PhD thesis, Rice University, 2016.

[63] Xin Kang and Chunhe Li. A dimension reduction approach for energy landscape: Identifying intermediate states in metabolism-emt network. *Advanced Science*, 8(10):2003133, 2021.

[64] Ygal Haupt, Ruth Maya, Anat Kazaz, and Moshe Oren. Mdm2 promotes the rapid degradation of p53. *Nature*, 387(6630):296–299, 1997.

[65] Jacques Piette, Henry Neel, and Vincent Maréchal. Mdm2: keeping p53 under control. *Oncogene*, 15(9):1001–1010, 1997.

[66] M Reza Saadatzadeh, Adily N Elmi, Pankita H Pandya, Khadijeh Bijangi-Vishehsaraei, Jixin Ding, Christopher W Stamatkin, Aaron A Cohen-Gadol, and Karen E Pollok. The role of mdm2 in promoting genome stability versus instability. *International Journal of Molecular Sciences*, 18(10):2216, 2017.

[67] Ute M Moll and Oleksi Petrenko. The mdm2-p53 interaction. *Molecular cancer research*, 1(14):1001–1008, 2003.

[68] Mira Park, Soon Ae Kim, Jaeyong Yee, Jieun Shin, Kyu Young Lee, and Eun-Jeong Joo. Significant role of gene–gene interactions of clock genes in mood disorder. *Journal of Affective Disorders*, 257:510–517, 2019.

[69] Lindsay Melhuish Beaupre, Gregory M Brown, and James L Kennedy. Circadian genes in major depressive disorder. *The World Journal of Biological Psychiatry*, 21(2):80–90, 2020.

[70] Jiajun Shi, Jacqueline K Wittke-Thompson, Judith A Badner, Eiji Hattori, James B Potash, Virginia L Willour, Francis J McMahon, Elliot S Gershon, and Chunyu Liu. Clock genes may influence bipolar disorder susceptibility and dysfunctional circadian rhythm. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(7):1047–1055, 2008.

[71] Annaëlle Charrier, Bertrand Olliac, Pierre Roubertoux, and Sylvie Tordjman. Clock genes and altered sleep–wake rhythms: their role in the development of psychiatric disorders. *International journal of molecular sciences*, 18(5):938, 2017.

[72] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002.

[73] Mark. Ptashne. *A Genetic Switch : Phage Lambda Revisited*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y, 3rd ed. edition, 2004.

[74] Mark Ptashne and Alexander Gann. Imposing specificity by localization: Mechanism and evolvability. *Current Biology*, 8(22):R812–R822, November 1998.

[75] Jay C Dunlap. Molecular bases for circadian clocks. *Cell*, 96(2):271–290, 1999.

[76] Satchidananda Panda, Marina P Antoch, Brooke H Miller, Andrew I Su, Andrew B Schook, Marty Straume, Peter G Schultz, Steve A Kay, Joseph S Takahashi, and John B Hogenesch. Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, 109(3):307–320, 2002.

[77] Dennis Bray. Protein molecules as computational elements in living cells. *Nature*, 376(6538):307–312, 1995.

[78] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[79] Michael B Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[80] Jordi Garcia-Ojalvo, Michael B Elowitz, and Steven H Strogatz. Modeling a synthetic multicellular clock: repressilators coupled by quorum sensing. *Proceedings of the National Academy of Sciences*, 101(30):10955–10960, 2004.

[81] Lingxia Qiao, Zhi-Bo Zhang, Wei Zhao, Ping Wei, and Lei Zhang. Network design principle for robust oscillatory behaviors with respect to biological noise. *bioRxiv*, 2021.

A Notation

$X_i; i = 1, \ldots, n$	molecular species
$\boldsymbol{x}(t) = (x_1(t), \ldots, x_n(t))$	microstate of copy numbers at time t
$R_j, r_j; j = 1, \ldots, m$	reactions, reaction rates
$c_{j,i}, c'_{j,i}$	copy numbers of X_i before and after R_j
$\boldsymbol{s}_j = (\ldots, c'_{j,i} - c_{j,i}, \ldots)$	stoichiometry vector of R_j
$\Omega = \{[x_1, \ldots, x_n]\}$	state space
$\boldsymbol{p}\colon \Omega \times [0, \infty) \to [0, 1]$	probability landscape
$\boldsymbol{p}_t\colon \Omega \to [0, 1]$	time-slice of landscape
$A_j(\boldsymbol{x})$	propensity of R_j
$f\colon M \to \mathbb{R}$	smooth function on n-dimensional manifold
$M^s = f^{-1}[s, \infty)$	superlevel set
$\beta_k; k = 0, 1, \ldots$	k-th Betti number
$I_1 = [a, a + 1]$ or a	interval between consecutive integers or integer
$\sigma, \tau = I_1 \times \ldots \times I_n$	cubes

Table 1: Notation used in the paper.

B Glossary

ACME (Accurate Chemical Master Equation) A numerical framework for solving the Chemical Master Equation with controlled truncation error, enabling accurate computation of high-dimensional probability landscapes for stochastic reaction networks.

Persistent homology A method from topological data analysis that characterizes multiscale topological structure by tracking the appearance and disappearance of homological features across a filtration.

Persistence The lifetime of a homological feature across a filtration, typically measured as the difference between its birth and death parameter values, and used as a measure of feature significance. In this work, the filtration is defined by decreasing probability values, effectively revealing topological features as the probability landscape is progressively thresholded.

Prominence A geometric notion describing how pronounced a topological feature is relative to its surrounding landscape. In this work, prominence is quantified by the persistence of the corresponding homological feature, measured as the difference between its birth and death values in the superlevel-set filtration of the probability landscape.

0**-homology** Homology corresponding to connected components; in probability landscapes, it corresponds to distinct probability peaks or basins. Each 0-homology class represents a group of microstates that form a single connected region.

1**-homology** Homology corresponding to one-dimensional cycles (loops); in probability landscapes, a 1-homology class corresponds to a loop of high-probability microstates enclosing a lower-probability region, such as a crater-like structure.

2**-homology** Homology corresponding to two-dimensional voids or cavities inside a three-dimensional object. For three-dimensional probability landscapes, these topological features are generated by 2-cycles (closed, shell-like or spherical surfaces of higher-probability states) that wrap around and fully enclose a region of low probability states. More generally, in dimension n, a void is captured by $n-1$-homology.