Decomposing force fields as flows on graphs reconstructed from stochastic trajectories

Ramón Nartallo-Kaluarachchi
Mathematical Institute, University of Oxford
Centre for Eudaimonia and Human Flourishing, University of Oxford
nartallokalu@maths.ox.ac.uk

Paul Expert
UCL Global Business School for Health, UCL

David Beers
Department of Mathematics, UCLA

Alexander Strang
Department of Statistics, University of California, Berkeley

Morten L. Kringelbach
Centre for Eudaimonia and Human Flourishing, University of Oxford
Department of Psychiatry, University of Oxford
Center for Music in the Brain, Aarhus University

Renaud Lambiotte
Mathematical Institute, University of Oxford
The Alan Turing Institute

Alain Goriely
Mathematical Institute, University of Oxford

Abstract

Disentangling irreversible and reversible forces from random fluctuations is a challenging problem in the analysis of stochastic trajectories measured from real-world dynamical systems. We present an approach to approximate the dynamics of a stationary Langevin process as a discrete-state Markov process evolving over a graph-representation of phase-space, reconstructed from stochastic trajectories. Next, we utilise the analogy of the Helmholtz-Hodge decomposition of an edge-flow on a contractible simplicial complex with the associated decomposition of a stochastic process into its irreversible and reversible parts. This allows us to decompose our reconstructed flow and to differentiate between the irreversible currents and reversible gradient flows underlying the stochastic trajectories. We validate our approach on a range of solvable and nonlinear systems and apply it to derive insight into the dynamics of flickering red-blood cells and healthy and arrhythmic heartbeats. In particular, we capture the difference in irreversible circulating currents between healthy and passive cells and healthy and arrhythmic heartbeats. Our method breaks new ground at the interface of data-driven approaches to stochastic dynamics and graph signal processing, with the potential for further applications in the analysis of biological experiments and physiological recordings. Finally, it prompts future analysis of the convergence of the Helmholtz-Hodge decomposition in discrete and continuous spaces.

1 Introduction

The analysis of stochastic trajectories emerging from noisy, fluctuating systems, such as those found in biology, has become an important challenge in data science [1]. Such systems exhibit

R. Nartallo-Kaluarachchi et al., Decomposing force fields as flows on graphs reconstructed from stochastic trajectories. *Proceedings of the Third Learning on Graphs Conference (LoG 2024)*, PMLR 269, Virtual Event, November 26–29, 2024.

stochastic dynamics and can be well-modelled by Langevin processes, which are stochastic differential equations (SDEs) describing the time-evolution of variables subject to deterministic forces and noisy fluctuations [2]. Examples of such systems include climate [3], neural [4] and heartbeat dynamics [5] as well as fluctuations in microscopic biological processes [6], financial time-series and Brownian particles [7]. Data from such systems takes the form of noisy trajectories in which drift and diffusive fields are inaccessible prompting the development of a range of methods to infer and approximate dynamics directly from observed trajectories [1, 8–17]. A particular interest lies in the analysis of the nonequilibrium, time-irreversible nature of these processes, consistent with the consumption of energy and the dissipation of entropy to their environments [6, 18]. Existing techniques for studying nonequilibrium dynamics in data include coarse-graining for estimating probability currents [16, 17], measuring time-irreversibility [19], thermodynamic uncertainty relations [20] or the variance-sum rule for measuring entropy production [21].

Concurrently, topological and graph-theoretic methods for the analysis of large and complex datasets have emerged as a powerful paradigm within data science [22–25]. Such methods have also found applications in the analysis of time-series through techniques like attractor reconstruction [26] and, more recently, graph and topological signal processing [27–30]. Amongst these methods, the Helmholtz-Hodge decomposition (HHD), which decomposes flows into gradients and circulations, has proved to be a valuable tool in the analysis of empirical signals evolving on discrete objects [31–33]. Surprisingly, an analogous decomposition to the HHD can also be applied to stationary Langevin processes, decomposing stochastic dynamics into reversible and irreversible parts [18, 33, 34]. In an attempt to bridge the HHD of an SDE with that of an associated flow on a graph, we construct a data-driven approach for approximating a Langevin process with a Markov chain constructed entirely from stochastic trajectories. Firstly, we use a maximum-likelihood estimator to reconstruct the transition intensities of a discrete-state, continuous-time Markov chain that approximates the underlying Langevin process. By taking an appropriate ratio of the rates, we define an edge-flow on a triangulation of phase-space and decompose this flow to gain insight into the irreversible and reversible forces driving the dynamics of the trajectories. Our approach is validated with numerical simulations of both solvable and nonlinear systems and then applied to the analysis of flickering red-blood cell (RBC) experiments and healthy and arrhythmic heartbeats. In particular, our approach is able to capture the difference in the prevalence of irreversible, circulating currents between the dynamics of healthy and *adenosine triphosphate*-depleted (ATP) RBCs [21, 35, 36] as well as the increased time-irreversibility of healthy heartbeats compared to those with arrhythmia [37, 38].

2 Methods

2.1 The various faces of the Helmholtz-Hodge decomposition

The HHD is a special case of the, more general, Hodge decomposition of a flow on a manifold [39]. Three variants of the HHD include the decomposition of a real-valued vector field [40], the decomposition of the drift field of stationary Langevin process [34], or the decomposition of an edge-flow on a simplicial complex [31]. Fig. 1 illustrates each of these three faces of the HHD. The Langevin process is described by the SDE,

$$\dot{\mathbf{x}}(t) = \boldsymbol{F}(\mathbf{x}(t)) + \boldsymbol{\xi}(t), \tag{1}$$

where $\boldsymbol{F} : \mathbb{R}^N \to \mathbb{R}^N$ is a vector field on $\mathbb{R}^N$ that represents the time-constant, deterministic, driving force of the system, known as the *drift*, whilst $\boldsymbol{\xi} \in \mathbb{R}^n$ are noise in the form of Gaussian fluctuations, known as the *diffusion*, with spatially-homogeneous correlations given by $\langle \boldsymbol{\xi}(t), \boldsymbol{\xi}(t') \rangle = 2\boldsymbol{D}\delta(t-t')$ where $\boldsymbol{D} \in \mathbb{R}^N \times \mathbb{R}^N$ is the constant, positive definite, noise-covariance matrix. Alternatively, one can consider the probability density of the process where $P(\mathbf{x}, t)$ represents the probability of the process being in state $\mathbf{x}$ at time t. The density evolves according to the Fokker-Planck (FP) equation [2],

$$\frac{\partial P(\mathbf{x}, t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left[F_i P\right] + \sum_{i,j} \frac{\partial}{\partial x_i \partial x_j}\left[D_{ij} P\right]. \tag{2}$$

This can be written as a conservation of probability law,

$$\frac{\partial P}{\partial t} + \nabla \cdot \boldsymbol{J} = 0, \tag{3}$$

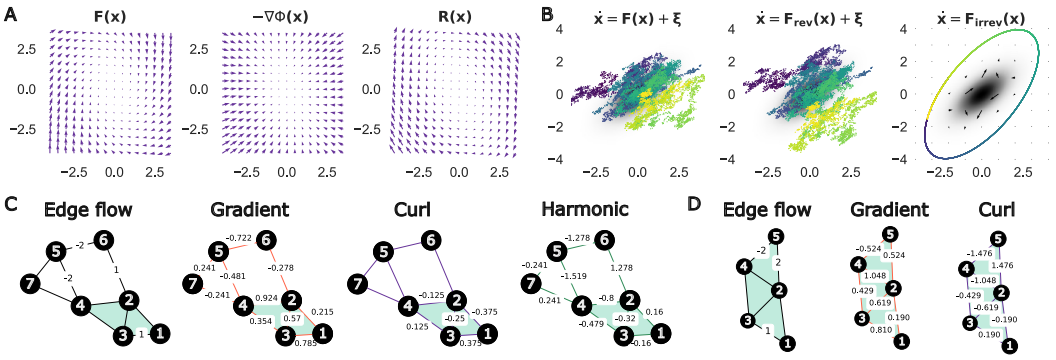

Figure 1: The Helmholtz-Hodge decomposition in various spaces. A. The HHD of a linear stable vector field $\boldsymbol{F}$ into a gradient component, $-\nabla\boldsymbol{\Phi}$, and a rotational component, $\boldsymbol{R}$. **B.** The HHD of the drift field of a stationary Langevin process into its reversible component, the portion of the drift that maintains the stationary distribution (plotted behind - black) as constant despite the noisy fluctuations, and its irreversible component, which drive deterministic rotation around the stationary distribution. The first panel shows a trajectory from the stationary OU process, whilst the second and third show the same trajectory decomposed into its reversible and irreversible dynamics respectively. The arrows in the final panel indicate the stationary probability flux $\boldsymbol{J}_{ss}$. **C.** The HHD of an edge-flow on a simplicial complex into gradient, curl and harmonic parts. Simplices are oriented naturally according to vertex labels. The gradient part sums to zero around closed loops. The curl part is only defined on triangles (example flow from Ref. [32]). **D.** The HHD of an edge-flow on a triangulation into gradient and curl parts. As the complex has no holes, there is no harmonic flow. Blank edges have zero flow.

where $\boldsymbol{J}(\mathbf{x},t) = \boldsymbol{F}(\mathbf{x})P(\mathbf{x},t) - \nabla\cdot(\boldsymbol{D}P(\mathbf{x},t))$ represents the probability flux and $\nabla\cdot$ represents the divergence operator (see App. A.1). The process is described as *stationary* when its density is constant, i.e. when $\partial P/\partial t = 0$, despite the process continuing to evolve in time. At stationarity, we can decompose the drift as follows [34],

$$\boldsymbol{F} = \boldsymbol{F}_{\text{irr}} + \boldsymbol{F}_{\text{rev}}, \tag{4}$$

$$\boldsymbol{F}_{\text{rev}} = \boldsymbol{D}\cdot\nabla\log P_{ss}, \tag{5}$$

$$\boldsymbol{F}_{\text{irr}} = \boldsymbol{J}_{ss}/P_{ss}, \tag{6}$$

where $\boldsymbol{F}_{\text{irr}}$ & $\boldsymbol{F}_{\text{rev}}$ represent the time-irreversible and time-reversible components of the drift respectively, ∇ is the gradient of a scalar potential (see App. A.1), $P_{ss}(\mathbf{x})$ is the stationary distribution, the solution to Eq. 2 when $\partial P/\partial t = 0$, and $\boldsymbol{J}_{ss}(\mathbf{x})$ is the stationary probability flux, which satisfies $\nabla\cdot\boldsymbol{J} = 0$. More specifically, $\boldsymbol{F}_{\text{irr}}$ is odd whilst $\boldsymbol{F}_{\text{rev}}$ is even under time-reversal [34]. The detailed balance condition is satisfied if $\boldsymbol{J}_{ss} \equiv \mathbf{0}$, corresponding to an equilibrium stationary distribution with time-reversible dynamics, and broken if $\boldsymbol{J}_{ss} \not\equiv \mathbf{0}$, corresponding to a nonequilibrium stationary distribution and time-irreversible dynamics [18]. Thus, information about the (ir)reversibility of the stochastic dynamics can be inferred by decomposing the drift field of the underlying stationary Langevin process. As shown in panel B of Fig. 1, the reversible component of the drift climbs the stationary distribution whilst the diffusion causes random fluctuations. The irreversible part of the drift drives deterministic rotation, contouring the stationary distribution. In the case of constant isotropic noise, $\boldsymbol{D} = \epsilon\boldsymbol{I}$, the drift can be written as the sum of a divergence-free curl flux ($\nabla\cdot\boldsymbol{J}_{ss}/P_{ss} = 0$) and the gradient of a scalar potential,

$$\boldsymbol{F} = \boldsymbol{J}_{ss}/P_{ss} + \nabla\log\epsilon P_{ss}. \tag{7}$$

As a result, the process is reversible if and only if the drift field is *conservative* i.e. it is the gradient of a scalar potential [41]. In the case of non-isotropic noise, we can perform a coordinate change $\mathbf{x} = \boldsymbol{D}^{-1/2}\mathbf{y}$ that yields a process with isotropic noise [33].

We also introduce the HHD of a flow on the edges of a simplicial complex [31]. Given a graph $G = (V, E)$, where V is a finite set of vertices and E is the set of edges, higher-order cliques such as the set of triangles (3-cliques) are automatically defined, yielding a simplicial complex. Real-valued functions defined on the $k-$cliques of a simplicial complex are denoted *flows* with the constraint that, for $k > 1$, flows must be *alternating* e.g. $X(i,j) = -X(j,i)$ is an alternating

edge-flow (see Ref. [31] for further details). Typically the HHD is formulated as a decomposition of an edge-flow into three orthogonal components, a gradient flow defined by a potential on the vertices, local circulations around triangles in the complex, and the remaining harmonic flow associated with the Hodge Laplacian [31, 32]. Loosely speaking, the dimension of the harmonic space, the nullity of the Hodge Laplacian, is given by the number of *holes* in the complex [31]. However, by loosening the restriction that circulations must take place on triangles and allowing them to take place over longer loops, the harmonic contribution becomes zero [33]. In our approach, we will consider the *Delauney triangulation* [22, 42] which automatically defines a contractible complex with no holes, thereby eliminating harmonic flow. Thus, by introducing discrete analogues of the curl, divergence and gradient operators (and their adjoints) [43], we can decompose an edge-flow, X, as the sum of two orthogonal components,

$$X = \text{curl}^* \Phi + \text{grad} f, \tag{8}$$

where Φ is an alternating function defined on the triangles, f is a function defined on the vertices, with Φ and f being analogous to a vector and scalar potential respectively (see App. A.1 for details) [31, 33]. Therefore an edge-flow on a triangulation can be uniquely decomposed into a curling component and the gradient of a potential defined on the vertices, mirroring the decomposition of the stationary Langevin process, where, informally, the term grad$f$ mirrors the gradient term in Eq. 5, thus $\exp(f/\epsilon)$ provides a natural estimate of the stationary distribution. Similarly, curl$^*\Phi$ captures the irreversible probability currents of Eq. 6.

Finally, we consider the HHD of a discrete-state Markov process which can be formulated as a flow on the edges of graph. We begin with a master equation of the form,

$$\frac{d\boldsymbol{p}}{dt} = \boldsymbol{Q}\boldsymbol{p} \tag{9}$$

where $\boldsymbol{p} = (p_1, ..., p_M)$ represents the probability of a random walker being at each of M vertices and $\boldsymbol{Q} = (q_{ij})$ is the Q-matrix (also referred to as the generator or Laplacian matrix) of the continuous-time Markov chain defined by a random walk restricted to the edges of the graph. Off diagonal entries $q_{ji} > 0$ of the $Q-$matrix represent *transition intensities* between vertices i and j whilst diagonal entries are defined to be $q_{ii} = -\sum_j q_{ji}$, thus each column sums to 0 [44]. Furthermore, we assume every transition is reciprocated i.e. if $q_{ij} > 0$ then $q_{ji} > 0$. Next, we define the edge-flow to be,

$$X_{ij} = \log \sqrt{\frac{q_{ij}}{q_{ji}}}, \tag{10}$$

which is naturally alternating i.e. $X_{ij} = -X_{ji}$, and can be motivated by both mathematical and thermodynamic arguments (see Chapter 6 of Ref. [33]). In particular, it is analogous to the excess work required to move from state i to j [33, 45]. Furthermore, this definition of the edge-flow is analogous to the drift field of the Langevin process, because the Markov process is reversible and obeys detailed balance if and only if X is a conservative flow i.e. it is the gradient of a scalar potential defined on the vertices (see Chapter 6.3 of Ref. [33] for a proof). Therefore the HHD can yield insight into the reversible and irreversible dynamics of a discrete-state, continuous-time Markov process.

2.2 Approximating force fields as flows on graphs

Our approach attempts to approximate an unknown, underlying continuous Langevin process with a discrete Markov process that is reconstructed from an ensemble of stochastic trajectories. Representing the Markov process as a flow on a triangulation, we can leverage the correspondence between the HHD of the continuous and discrete-state processes to perform the HHD in discrete-space thus decomposing the process into its reversible and irreversible parts. This will allows us to approximate the curling and gradient components of the continuous drift-field with a discrete counterpart that can be calculated in a computationally simple manner. This is due to the relative ease with which the HHD of an edge-flow can be computed through its formulation as a least-squares optimisation problem (see App. A.1 and Ref. [31]).

Our framework, outlined in Fig. 2, begins with an ensemble of trajectories assumed to be from the same stationary Langevin process. In this first case we assume that the noise is known to be isotropic with given ϵ. Subsequently, we will treat the case of non-isotropic noise. Trajectories live in phase-space but are often clustered onto a particular, potentially lower dimensional, manifold

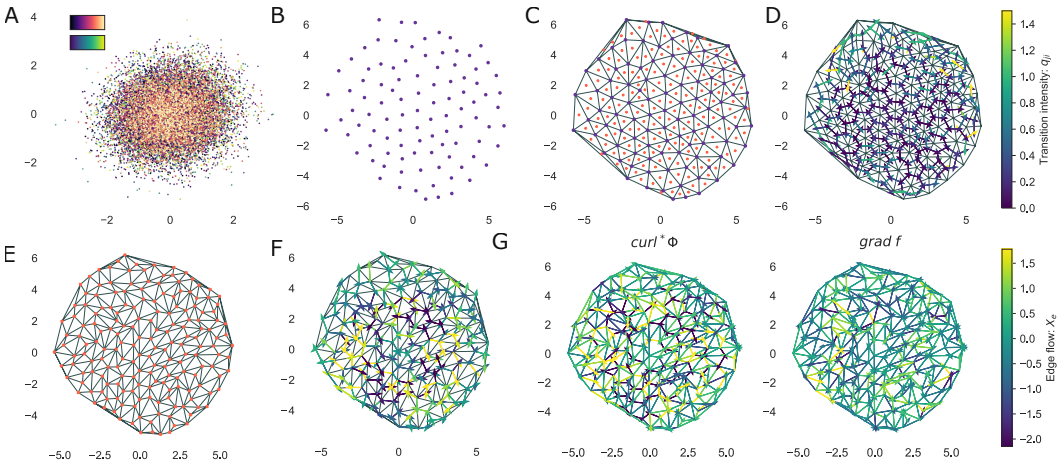

Figure 2: Reconstruction and decomposition of drift fields as flows on graphs. A. We begin with an ensemble of stochastic trajectories from the same system as a point-cloud on phase-space, illustrated here with two trajectories. **B.** We sub-sample the points using furthest point sampling to obtain a uniform density representation of phase-space. **C.** We discretise phase-space using a triangulation. Red points denote the center of each bin. **D.** We use maximum-likelihood estimation to reconstruct the transition intensities of a discrete-state Markov process with transitions between adjacent bins. **E.** Triangulating the midpoints of the bins, we obtain a simplicial complex representation of phase-space with transitions occurring over edges. **F.** Following the analysis in Sec. 2.1, we define the edge-flow by taking the logarithm of the square-root of the ratio of the transition intensities between connected bins. **G.** We perform the discrete HHD to obtain the curl and gradient components of the flow on the complex. The edge-flow colourbar refers to panels F and G.

[46]. In order to discretise phase-space optimally given our observed trajectories, we consider them to be a point-cloud and perform furthest-point sampling [47] which yields a uniformly populated representation of phase-space that is optimal for manifold learning [48]. Tessellating the sub-sampled points with the Delauney triangulation, we obtain an irregular, triangular 'grid' of M bins (for an introduction to Delauney triangulations see Ref. [42]). We aim to approximate the Langevin process with the master equation defined in Eq. 9. The discrete state of the process is the triangle/bin in which the trajectory is in at each time-point. After converting trajectories to observations of the discrete-state process, we use the maximum-likelihood estimator to estimate the transition intensities [49],

$$\hat{q}_{ji} = \frac{N_{i\to j}}{T_i}, \tag{11}$$

where $N_{i\to j}$ is the observed number of transitions between adjacent triangular bins i and j, T_i is the total 'holding' time spent in each bin and bins are considered to be adjacent if they share an edge. Additionally, we set $\hat{q}_{ii} = -\sum_j \hat{q}_{ji}$. Given sufficient time-resolution, transitions should only occur between adjacent bins, thus transitions between non-adjacent bins that arise due to insufficient sampling frequency are not included in the estimation. Moreover, we include a pseudo-count, by initialising all $N_{i\to j}$ with a single unobserved transition, which eliminates singularities due to limited observations and which corresponds to a uniform Bayesian prior for the intensities [50].

Next, we obtain the dual triangulation of the original grid by using the midpoints of the bins as vertices and performing the Delauney triangulation again. As a result, each vertex corresponds to a discrete state/bin in the Markov process and adjacent bins/reachable states are now connected by edges. The flow on the edges of this second triangulation is defined to be $X_{ij} = \log\sqrt{q_{ij}/q_{ji}}$ where we assume the natural orientation (i, j) if $i < j$. This flow is the discrete counter-part to the drift field as its curling component represents irreversible flow and its gradient component represents reversible flow. More specifically, the Markov process obeys detailed balance iff this flow is conservative [33].

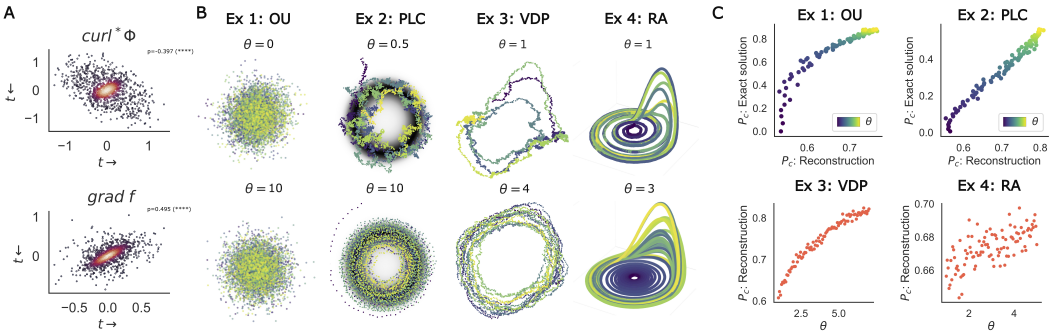

Figure 3: Numerical experiments. A. We begin with an ensemble of stochastic trajectories from a nonequilibrium OU process and apply our approach to both the original and time-reversed trajectories. As the curl part is odd under time-reversal, we see a significant negative correlation (-0.397, $p < 0.0001$) between the curl flow along each edge calculated with the original and time-reversed trajectories. Conversely, the gradient flow is even under time-reversal, thus the gradient flows are positively correlated $(0.495, p < 0.0001)$. **B.** Example trajectories from each of the four example processes with a low and high level of the rotation parameter θ. **C.** For the solvable examples (OU, PLC) we increase θ, calculate the proportion of flow in the curl space and compare to the exact HHD. We show a strong qualitative agreement with the reconstructed proportion of curl flow increasing with θ. The disparity between the overall proportions arises from finite data and the approximation of the continuous space with discrete bins. For the unsolvable models (VDP, RA), we cannot compare to the exact solution, however θ acts as a proxy parameter for the level of rotation in the drift field. For both systems, the proportion of the reconstructed flow that is in the curl space increases with θ, indicating that the edge-flow and its HHD capture the difference between irreversible and reversible currents.

2.2.1 Non-isotropic diffusion

In the case of a process with non-isotropic diffusion, we must first perform the coordinate change $\mathbf{x} = \boldsymbol{D}^{-1/2}\mathbf{y}$ to obtain a process, $\mathbf{y}(t)$, with isotropic diffusion. In order to infer the diffusion matrix from the stochastic trajectories, we utilise the original tessellation and calculate the maximum-likelihood estimator [1],

$$\hat{\boldsymbol{D}}(\Delta_{\bar{\mathbf{x}}}) = \frac{1}{N(\Delta_{\bar{\mathbf{x}}})} \sum_{i,r\,:\,\mathbf{x}_{t_i,r}\in\Delta_{\bar{\mathbf{x}}}} \frac{[\mathbf{x}_{t_{i+1},r} - \mathbf{x}_{t_i,r}] \otimes [\mathbf{x}_{t_{i+1},r} - \mathbf{x}_{t_i,r}]}{2\tau}, \tag{12}$$

where $N(\Delta_{\bar{\mathbf{x}}})$ represents the number of points in each bin $\Delta_{\bar{\mathbf{x}}}$ identified by its mid-point $\bar{\mathbf{x}}$, $\{\mathbf{x}_{t_1,r}, ..., \mathbf{x}_{t_T,r}\}_{r=1}^{R}$ is the ensemble of trajectories, $\otimes$ is the outer product and τ is the time-step between observations. As the diffusion is assumed to be spatially homogeneous, given by a constant matrix, we take the mean value of the reconstructed matrices, $\boldsymbol{D} = \sum_{\Delta} \boldsymbol{D}(\Delta)/M$, which we can then use to transform the coordinates of the stochastic trajectories and apply the original method. In some physical systems, such as the dynamics of soft-matter in a medium, spatial-dependence in the diffusion tensor is a central feature [51]. In the real-world systems considered here, the diffusion represents measurement error which we consider to be state-independent and temporally uncorrelated [5, 21]. In App. A.6.2, we briefly investigate the validity of this assumption by assessing if there is spatially-dependent diffusion in the real-world data considered in this study.

3 Results

3.1 Numerical experiments

In order to validate that our approach does, in fact, capture curl and gradient flows from stochastic trajectories, we first apply our analysis to simulations of four canonical dynamical systems with additive isotropic diffusion. We begin with two solvable models where the stationary probability distribution can be written in closed form and thus the HHD can be performed exactly. Specifically, we consider the multivariate Ornstein-Uhlenbeck (OU) process [52], a linear nonequilibrium process,

and the stochastic planar limit cycle (PLC) [53], a nonlinear, nonequilibrium process. Next, we consider two nonlinear, unsolvable models, the stochastic van der Pol oscillator (VDP) [54], a system characterised by an asymmetric limit-cycle, and the stochastic Rössler attractor (RA) [55], a 3-dimensional system with a chaotic attractor. Panel B in Fig. 3 shows example trajectories from each of the four processes with two levels of rotation in the drift field. As detailed in App. A.2, the systems are modified to be parameterised by θ, which controls the level of rotation in the drift field, and ϵ which controls the rate of diffusion.

The irreversible/reversible components of the drift are odd/even under time-reversal respectively [34]. As a first assessment of the method, we construct and decompose edge-flows from trajectories and time-reversed trajectories of the nonequilibrium OU. Panel A of Fig. 3 shows the correlation of the curl (top) and gradient (bottom) flows along each edge between the original and time-reversed trajectories, noting that both utilise the same triangulations of phase-space. The curl flows are negatively correlated (-0.397, $p < 0.0001$) whilst the gradient flows are positively correlated (0.495, $p < 0.0001$), indicating that the components of the HHD of the edge-flow qualitatively capture the odd/even behaviour of the components of the underlying Langevin process under time-reversal. Next, as a measure of the prevalence of irreversible currents in the process, we calculate the proportion of the total flow that is in the curl space i.e. $P_c = ||\boldsymbol{X}_c||/[||\boldsymbol{X}_c|| + ||\boldsymbol{X}_g||]$ where $\boldsymbol{X}_c, \boldsymbol{X}_g$ represent curl and gradient components respectively. Panel C in Fig. 3, shows the proportion of the flow in the curl space for increasing values of θ for each process. For the OU and PLC, we compare P_c for the exact HHD of the drift field, calculated using a finite bounded domain, and P_c in the reconstructed edge-flow. As θ is increased, the proportion of the flow that is irreversible/curl increases in the exact HHD, an effect that is captured with strong agreement by the discrete reconstruction. For the VDP and RA processes, the exact HHD cannot be performed, thus we use θ as a proxy measure for the level of curl in the field, showing that P_c increases with θ in the reconstructed flow. As a result, the discrete reconstruction and decomposition is able to capture the difference between reversible gradient flows and irreversible circulations which can be used to identify the prevalence of irreversible currents. Whilst the general increase in P_c agrees between the exact HHD and the discrete reconstruction, we note that the reconstructed proportion does not perfectly mirror the exact solution. This effect could stem from a combination of the discrete approximation of phase-space, the finite length of the trajectories and, potentially, from an inconsistency between the continuous and discrete forms of the HHD. A brief analysis of the effect of variation of the method parameters is presented in App. A.5. However, the method is able to differentiate between comparable systems with different proportions of irreversible curl flow, which is necessary for the following real-world applications. Details on the processes, simulation techniques, simulation parameters and variation of the parameters are given in App. A.2, A.3, A.4, A.5 and A.7.

3.2 Real-world examples

Noisy time-series from real-world systems can be modelled as stochastic trajectories from Langevin processes. In this section we consider two experiments with recordings of biophysical processes, namely flickering RBCs [21, 35] and healthy and arrhythmic heartbeats [56]. In the case of the RBCs, the dissipation of energy created by irreversible currents has been found to be lower in ATP-depleted cells compared to healthy cells [21, 57]. As a result, the flickering recorded from passive cells is closer to thermal noise, and thus has a lower level of irreversible curling. Panel A of Fig. 4 shows an example trace of the displacement of a point on the outer membrane of both a healthy and passive cell. The dynamics of the inner-membrane remains an unobserved variable, thus we assume the position trace is a scalar measurement of a 2-dimensional system [21]. Nonequilibrium dynamics have also been closely linked to the healthy functioning of biological organs [37], in particular in the heart where decreased time-irreversibility is associated with ageing and disease [37, 38]. We consider electrical signals from the heart recorded as a electrocardiogram (ECG) which captures the voltage of action potentials. We compare the dynamics of healthy heartbeats with arrhythmic ones from patients with a myocardial infarction [56]. Panel C of Fig. 4 shows a typical ECG trace of a healthy (red) and arrhythmic (blue) heartbeat. Whilst the ECG is a scalar measurement, the limit-cycle nature of human heartbeats indicates a second, unobserved variable [5].

In order to reconstruct a 2-dimensional trajectory in phase-space from a scalar measurement, we use a 2-d time-delay embedding with a single lag, τ, chosen for each experiment by the

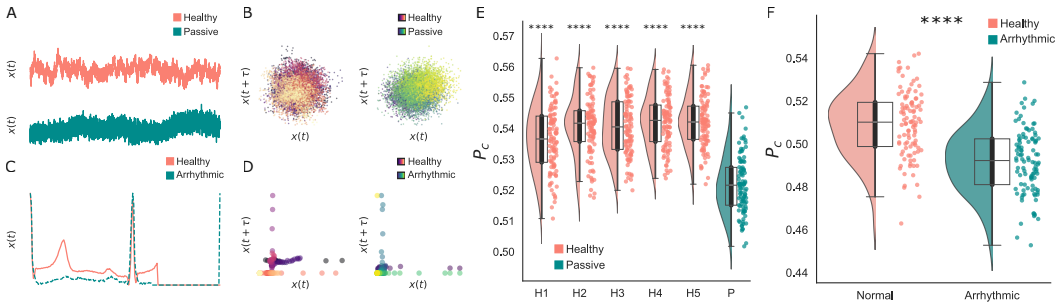

Figure 4: Analysis of real-world data. A. An example position trace for the outer membrane of a healthy (red) and passive (blue) RBC. **B.** An example trajectory in phase-space reconstructed using time-delay coordinates for both a healthy and passive RBC. **C.** An example ECG trace for a healthy participant (red) and one suffering myocardial infarction (blue). **D.** An example trajectory in phase-space reconstructed using time-delay coordinated for both a healthy and arrhythmic heartbeat. **E.** The proportion of flow in the curl space, P_c, for five healthy RBC and one passive RBC using 100 bootstrapped ensembles. We find that the prevalence of irreversible currents is significantly higher (****, $p < 0.0001$) in each healthy cell when compared to the passive one. **F.** We compare healthy and arrhythmic heartbeats and find that P_c is significantly higher (****, $p < 0.0001$) in healthy heartbeats compared to arrhythmic ones, using 100 bootstrapped ensembles.

mean of the first zero-crossing of the auto-correlation, thus we obtain a 2-d trajectory $(x(t), x(t+\tau))$ [58] (see App. A.6.1 for details and App. A.6.5 for the effect of varying this parameter.). Panels B and D of Fig. 4 show examples of reconstructed trajectories for healthy and passive RBCs and healthy and arrhythmic heartbeats. Unlike the example processes in Sec. 3.1, we cannot assume the diffusion is isotropic. As described in Sec. 2.2.1, we first sub-sample and triangulate the trajectory in the reconstructed phase-space and use Eq. 12 to estimate the diffusion matrix. Next, we change coordinates, $\mathbf{y} = \boldsymbol{D}^{1/2}\mathbf{x}$, resample and re-triangulate to then perform the analysis assuming isotropic diffusion. In order to compare if the statistical structure of the trajectories in each condition is responsible for the variation in P_c, we perform two forms of significance testing. Firstly, we use ensemble bootstrapping [16, 17] with 100 bootstraps to estimate the uncertainty in P_c in each case (see App. A.6.3). Additionally, we perform surrogate-testing with shuffled time-series where shuffling destroys the statistical structure whilst keeping the points in phase-space the same, thus generating a 'null' series [17] (see App. A.6.4). Furthermore, in App. A.6.2 we assess the validity of assuming spatially-homogeneous diffusion in these datasets. For further details on the datasets and the analysis see App. A.6.

Using our approach, we find that the proportion of irreversible, curling flow, P_c, is significantly higher (****, $p < 0.0001$) in the flickering of each of five healthy RBCs when compared directly to a passive one (independent $t-$test) as shown in panel E of Fig. 4. This result is in agreement with previous results measuring the entropy production rate, a measure of nonequilibrium, in the same data [21]. Next, we apply our approach to the healthy and arrhythmic heartbeats and find that P_c is significantly higher (****, $p < 0.0001$) in healthy heartbeats when compared to arrhythmic beats from patients with myocardial infarction, as shown in panel F of Fig. 4. Our results validate previous indications that time-irreversibility, nonequilibrium dynamics, and by extension, the prevalence of curling currents, are a key feature of healthy heart dynamics [37, 38].

4 Related work and other approaches

The analysis of nonequilibrium dynamics in biological systems has received much recent interest [6, 18], in particular in data-driven methods for analysing real-world trajectories [1, 8, 16, 59]. One typical approach is to infer approximate drift and diffusion terms directly from data [1, 9, 10, 13, 14], however only one such method 'Stochastic Force Inference' [8] allows for the natural HHD of the inferred drift as well as an estimate of the entropy production rate. Typically, methods aim to approximate a continuous field with a discrete approximation, but do not formulate a discrete-state stochastic process, as we do here. 'Probability flux analysis' [16] is another approach, designed for biological data, that attempts to approximate probability currents in a coarse-grained discrete

state-space. This approach approximates the continuous process with this discrete-state counterpart, by inferring the probability flux between system states, yet it stops short of decomposing the process into its reversible and irreversible parts or measuring the overall level of nonequilibrium. More generally, many methods have been developed to analyse nonequilibrium flows by utilising the link between nonequilibrium dynamics and the irreversibility of stationary trajectories [60, 61] with applications in biology [19] and neuroscience [62]. Similarly to the quantity P_c, such methods can quantify the overall prevalence of nonequilibrium flows in stochastic trajectories, but yield no insight into the governing dynamics of the process, unlike the reconstruction of a discrete-state process.

The discrete HHD has found many applications in the analysis of hierarchical and cyclical structures in graphs and tournaments [31, 63] with some applications to dynamics and signal processing [28–30, 32]. To this point, the discrete HHD has not been applied to Markov processes to gain insight into their thermodynamics with the exception of Ref. [33]. Nevertheless, connections can be made between the HHD and classical 'cycle decomposition' approaches that are well-studied in the thermodynamics of graphs and Markov processes [64–66]. Similarly, results in the convergence of the discrete and continuous HHD are limited to simple domains and dynamics [33].

The approximation of vector fields with discrete counter-parts is a central challenge in numerical analysis, such as finite element calculus [67]. Such approaches typically involve discretisation of continuous domains for the numerical solution of deterministic equations and have little relation to our approach. One more closely related approach is 'MARBLE' [48] which aims to learn the manifold structure of phase-space, which is approximated as a graph, and infer the vector field directly from recordings of neural signals - though the dynamics are assumed to be deterministic thus the vector field is not formulated as a discrete-state Markov process.

5 Discussion

Noisy, fluctuating signals represent the prototypical data that can be recorded from a real-world system. Whilst Langevin processes and SDEs provide a mathematical framework to describe such systems, the development of data-driven approaches is key to gaining insight into stochastic trajectories with unknown drift and diffusive fields [1]. In this paper, we focus on the HHD of a Langevin process into its irreversible, curling component and its reversible gradient component. By utilising the analogous HHD of a flow on the edges of a graph, we develop a method that is able to reconstruct and then decompose a discrete-state Markov process directly from stochastic trajectories. By focusing on the proportion of the flow in the curl space, we validate that our method captures the relative prevalence of irreversible currents between similar systems with different levels of nonequilibrium. We assess the efficacy of our approach on sampled paths from both solvable and nonlinear processes and then apply it to real-world data from biophysical experiments. Our approach is able to capture the greater prevalence of nonequilibrium flows in the dynamics of healthy RBCs when compared to ATP-depleted, 'passive' ones [21, 57] as well as the greater prevalence of nonequilibrium flows in healthy heartbeats compared to arrhythmic heartbeats from patients with myocardial infarction [37, 38, 56].

Our approach presents a new application of graph and topological signal processing to the analysis of stochastic trajectories and real-world, noisy time-series with a specific focus on biological and physiological recordings. In particular, our work builds on previous studies that highlight nonequilibrium and irreversible dynamics as central to the healthy functioning of mesoscopic [16] and macroscopic [17, 37, 38, 68] dynamics in biological systems. Finally, our method draws a novel and important link between the discrete HHD and its continuous counterpart that remains to be explored at a more theoretical and analytical level in order to investigate under what conditions this convergence is exact and both decompositions consistent [33]. As a first application, we focused on the overall prevalence of irreversible circulations, but we note that this approach can be developed, particularly in conjunction with existing techniques for analysing the discrete HHD [63, 69], to analyse the specific structure of the gradient and circular flows in phase-space to yield greater insight into underlying dynamics. Such an avenue hold great promise for the further development of graph and topological data analysis techniques for time-varying data recorded from complex systems.

Data and code availability

The code supporting the findings in this paper is available at https://github.com/rnartallo/decomposingflows.

The heartbeat data used in this study can be found at Ref. [70] as well as further details at [56, 71, 72]. The red-blood cell data used in this study can be found at Ref. [73] with further details at Ref. [21].

Author contributions

R.N.K developed the method with input from all other authors, performed research and drafted the manuscript. All authors discussed results and edited the manuscript.

Acknowledgements

The authors would like to thank Pierre Ronceray, Michael Schaub, Vincent Grande and Alexis Arnaudon for their helpful comments and conversations.

R.N.K was supported by an EPSRC doctoral scholarship from grants EP/T517811/1 and EP/R513295/1. M.L.K was supported by the Centre for Eudaimonia and Human Flourishing funded by the Pettit and Carlsberg Foundations and by the Center for Music in the Brain (MIB), funded by the Danish National Research Foundation (project number DNRF117). R.L. was supported by the EPSRC grants EP/V013068/1 and EP/V03474X/1.

References

[1] Rudolf Friedrich, Joachim Peinke, Muhammad Sahimi, and M. Reza Rahimi Tabar. Approaching complexity by stochastic methods: From biological systems to turbulence. *Physics Reports*, 506(5):87–162, 2011. 1, 2, 6, 8, 9

[2] N.G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science Publishers, Amsterdam, 1992. 2

[3] Christian L. E. Franzke and Terence J. O'Kane. *Nonlinear and Stochastic Climate Dynamics*. Cambridge University Press, 2017. 2

[4] L.M Harrison, O David, and K.J Friston. Stochastic models of neuronal dynamics. *Philosophical Transactions of the Royal Society B*, 360(1457):1075–1091, 2005. 2

[5] Z Qu, G Hu, A. Garfinkel, and J. N. Weiss. Nonlinear and stochastic dynamics in the heart. *Physics Reports*, 543(2):61–162, 2014. 2, 6, 7, 19

[6] F. S. Gnesotto, F. Mura, J. Gladrow, and C. P. Broedersz. Broken detailed balance and non-equilibrium dynamics in living systems: a review. *Reports on Progress in Physics*, 81(066601), 2018. 2, 8

[7] R. Brown. A brief account of microscopical observations made in the months of june, july and august 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philosophical Magazine*, 4(161), 1828. 2

[8] Anna Frishman and Pierre Ronceray. Learning force fields from stochastic trajectories. *Physical Review X*, 10(021009), 2020. 2, 8

[9] David B. Brückner, Pierre Ronceray, and Chase P. Broedersz. Inferring the dynamics of underdamped stochastic systems. *Physical Review Letters*, 125(058103), 2020. 8

[10] Laura Pérez García, Jaime Donlucas Pérez, Giorgio Volpe, Alejandro V. Arzola, and Giovanni Volpe. High-performance reconstruction of microscopic force fields from Brownian trajectories. *Nature Communications*, 9(5166), 2018. 8

[11] M. Genkin, O. Hughes, and T.A. Engel. Learning non-stationary Langevin dynamics from stochastic observations of latent trajectories. *Nature Communications*, 12(5986), 2021.

[12] Corneel Casert, Isaac Tamblyn, and Stephen Whitelam. Learning stochastic dynamics and predicting emergent behavior using transformers. *Nature Communications*, 15(1875), 2024.

[13] Mohamed El Beheiry, Maxime Dahan, and Jean-Baptiste Masson. InferenceMAP: mapping of single-molecule dynamics with Bayesian inference. *Nature Methods*, 12:594–595, 2015. 8

[14] Federica Ferretti, Victor Chardès, Thierry Mora, Aleksandra M. Walczak, and Irene Giardina. Building general Langevin models from discrete datasets. *Physical Review X*, 10(031018), 2020. 8

[15] David B. Brückner, Alexandra Fink, Christoph Schreiber, Peter J. F. Röttgermann, Joachim O. Rädler, and Chase P. Broedersz. Stochastic nonlinear dynamics of confined cell migration in two-state systems. *Nature Physics*, 15:595—-601, 2019.

[16] Christopher Battle, Chase P. Broedersz, Nikta Fakhri, Veikko F. Geyer, Jonathon Howard, Christoph F. Schmidt, and Fred C. Mackintosh. Broken detailed balance at mesoscopic scales in active biological systems. *Science*, 352(6285):604–607, 2016. 2, 8, 9, 20

[17] C.W. Lynn, E.J. Cornblath, L. Papadopoulos, M.A. Bertolero, and D.S. Bassett. Broken detailed balance and entropy production in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 118(47), 2021. 2, 8, 9, 20

[18] Xiaona Fang, Karsten Kruse, Ting Lu, and Jin Wang. Nonequilibrium physics in biology. *Reviews of Modern Physics*, 91(045004), 2019. 2, 3, 8, 14, 21

[19] E Roldán, Jérémie Barral, Pascal Martin, Juan M R Parrondo, and Frank Jülicher. Quantifying entropy production in active fluctuations of the hair-cell bundle from time irreversibility and uncertainty relations. *New Journal of Physics*, 23(083013), 2021. 2, 9

[20] Jordan M. Horowitz and Todd R. Gingrich. Thermodynamic uncertainty relations constrain non-equilibrium fluctuations. *Nature Physics*, 16:15–20, 2019. 2

[21] I. Di Terlizzi, M. Gironella, D. Herraez-Aguilar, T. Betz, F. Monroy, M. Baiesi, and F. Ritort. Variance sum rule for entropy production. *Science*, 383(6686):971–976, 2024. 2, 6, 7, 8, 9, 10, 19, 23

[22] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2): 255–308, 2009. 2, 4

[23] Gunnar Carlsson. Topological methods for data modelling. *Nature Reviews Physics*, 2:697–708, 2020.

[24] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34 (4):18 – 42, 2017.

[25] Herbert Edelsbrunner and John L. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010. 2

[26] M.R. Muldoon, R.S. MacKay, J.P. Hukeb, and D.S. Broomhead. Topology from time series. *Physica D: Nonlinear Phenomenon*, 65:1–16, 1993. 2

[27] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106 (5):808–828, 2018. 2

[28] Michael T. Schaub, Yu Zhu, Jean-Baptiste Seby, T. Mitchell Roddenberry, and Santiago Segarra. Signal processing on higher-order networks: Livin' on the edge... and beyond. *Signal Processing*, 187(108149), 2021. 9

[29] M.T. Schaub, JB. Seby, F. Frantzen, T.M. Roddenberry, Y. Zhu, and S Segarra. Signal processing on simplicial complexes. In *Higher Order Systems*. Springer, 2022. 15

[30] Sergio Barbarossa and Stefania Sardellitti. Topological signal processing over simplicial complexes. *IEEE Transactions on Signal Processing*, 68:2992–3007, 2020. 2, 9

[31] Lek-Heng Lim. Hodge Laplacians on graphs. *SIAM Review*, 62(3):685–715, 2020. 2, 3, 4, 9, 15, 16

[32] Michael T. Schaub, Austin R. Benson, Paul Horn, Gabor Lippner, and Ali Jadbabaie. Random walks on simplicial complexes and the normalized Hodge 1-Laplacian. *SIAM Review*, 62(2), 2020. 3, 4, 9

[33] Alexander Strang. *Applications of the Helmholtz-Hodge decomposition to networks and random processes*. PhD thesis, Case Western Reserve University, August 2020. 2, 3, 4, 5, 9

[34] Lancelot Da Costa and Grigorios A Pavliotis. The entropy production of stationary diffusions. *Journal of Physics A: Mathematical and Theoretical*, 56(36):365001, 2023. 2, 3, 7, 16, 17

[35] Hervé Turlier and Timo Betz. Unveiling the active nature of living-membrane fluctuations and mechanics. *Annual Review of Condensed Matter Physics*, 10:213–232, 2019. 2, 7, 23

[36] Timo Betz, Martin Lenz, Jean-François Joanny, and Cécile Sykes. ATP-dependent mechanics of red blood cells. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15320–15325, 2009. 2, 23

[37] C K Peng, S V Buldyrev, J M Hausdorff, S Havlin, J E Mietus, M Simons, H E Stanley, and A L Goldberger. Non-equilibrium dynamics as an indispensable characteristic of a healthy biological system. *Integrative Psychological and Behavioral Science*, 29(3):283–293, 2019. 2, 7, 8, 9, 21

[38] Madalena Costa, Ary L Goldberger, and C-K Peng. Broken asymmetry of the human heartbeat: loss of time irreversibility in aging and disease. *Physical Review Letters*, 95(198102), 2005. 2, 7, 8, 9, 21, 23

[39] Claire Voisin. *Hodge Theory and Complex Algebraic Geometry*, chapter The Hodge Decomposition. Cambridge University Press, 2010. 2

[40] Erhard Glötzl and Oliver Richters. Helmholtz decomposition and potential functions for n-dimensional analytic vector fields. *Journal of Mathematical Analysis and Applications*, 525(2): 127138, 2023. 2, 14

[41] Hong Qian, Min Qian, and Xiang Tang. Thermodynamics of the general diffusion process: Time-reversibility and entropy production. *Journal of Statistical Physics*, 107(5), 2002. 3

[42] Jesús A. Loera, Jörg Rambau, and Francisco Santos. *Triangulations: Structures for Algorithms and Applications*, chapter Configurations, Triangulations, Subdivisions, and Flips. Springer, 2010. 4, 5

[43] Leo J. Grady and Jonathan R. Polimeni. *Discrete Calculus: Applied Analysis on Graphs for Computational Science*. Springer, 2010. 4

[44] Yu M Suhov and M Kelbert. *Probability and statistics by example. 2, Markov chains : a primer in random processes and their applications*, chapter Continuous-time Markov chains. Cambridge University Press, 2008. 4

[45] Hong Qian. Nonequilibrium steady-state circulation and heat dissipation functional. *Physical Review E*, 64(022101), 2001. 4

[46] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29:983–1049, 2016. 5

[47] Carsten Moenning and Neil A. Dodgson. Fast marching farthest point sampling. Technical report, Computer Laboratory, University of Cambridge, 2003. 5

[48] Adam Gosztolai, Robert L. Peach, Alexis Arnaudon, Mauricio Barahona, and Pierre Vandergheynst. Interpretable statistical representations of neural population dynamics and geometry. *arXiv*, (2304.03376), 2023. 5, 9

[49] P Billingsley. *Statistical Inference for Markov Processes*. The University of Chicago Press, 1961. 5

[50] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. 5

[51] A. W. C. Lau and T. C. Lubensky. State-dependent diffusion: Thermodynamic consistency and its path integral formulation. *Physical Review E*, 76(011123), 2007. 6

[52] C Godrèche and JM Luck. Characterising the nonequilibrium stationary states of Ornstein-Uhlenbeck processes. *Journal of Physics A: Mathematical and Theoretical*, 52(3), 2018. 6, 16, 26

[53] Ruoshi Yuan, Ying Tang, and Ping Ao. SDE decomposition and A-type stochastic interpretation in nonequilibrium processes. *Frontiers of Physics*, 12(120201), 2017. 7, 16

[54] Steven H Strogatz. *Nonlinear Dynamics And Chaos With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press, 2008. 7, 17

[55] O.E. Rössler. An equation for continuous chaos. *Physics Letters*, 57(5):397–398, 1976. 7, 17

[56] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101:215–220, 2000. 7, 9, 10, 23

[57] Ruddi Rodríguez-García, Iván López-Montero, Michael Mell, Gustavo Egea, Nir S Gov, and Francisco Monroy. Direct cytoskeleton forces cause membrane softening in red blood cells. *Biophysical Journal*, 108(12):2794–2806, 2015. 7, 9, 23

[58] Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*, chapter Advanced embedding methods. Cambridge University Press, 2003. 8, 19, 21

[59] David B Brückner and Chase P Broedersz. Learning dynamical models of single and collective cell migration: a review. *Reports on Progress in Physics*, 87(056601), 2024. 8

[60] E Roldán and J. M. R. Parrondo. Estimating dissipation from single stationary trajectories. *Physical Review Letters*, 105(150607), 2010. 9

[61] Edward H. Feng and Gavin E. Crooks. Length of time's arrow. *Physical Review Letters*, 101 (090602), 2008. 9

[62] LA de la Fuente, F Zamberlan, H Bocaccio, ML Kringelbach, G Deco, Y Sanz Perl, C Pallavicini, and E Tagliazucchi. Temporal irreversibility of neural dynamics as a signature of consciousness. *Cerebral Cortex*, 33(5):1856–1865, 2023. 9

[63] Alexander Strang, Karen C. Abbott, and Peter J. Thomas. The network HHD: Quantifying cyclic competition in trait-performance models of tournaments. *SIAM Review*, 64(2), 2022. 9

[64] J. Schnakenberg. Network theory of microscopic and macroscopic behavior of master equation systems. *Reviews of Modern Physics*, 48(571), 1976. 9

[65] Mathematical Theory of Nonequilibrium Steady States. *Da-Quan Jiang and Min Qian and Min-Ping Qian*. Springer, 2004.

[66] B. Altaner, S. Grosskinsky, S. Herminghaus, L. Katthan, M. Timme, and J. Vollmer. Network representations of nonequilibrium steady states: Cycle decompositions, symmetries, and dominant paths. *Physical Review E*, 85(041133), 2012. 9

[67] Douglas N Arnold, Richard S Falk, and Ragnar Winther. Finite element exterior calculus: From Hodge theory to numerical stability. *Bulletin of the American Mathematical Society*, 47(2): 281–354, 2010. 9

[68] Yonatan Sanz Perl, Hernán Bocaccio, Carla Pallavicini, Ignacio Pérez-Ipiña, Steven Laureys, Helmut Laufs, Morten Kringelbach, Gustavo Deco, and Enzo Tagliazucchi. Nonequilibrium brain dynamics as a signature of consciousness. *Physical Review E*, 104(014411), 2021. 9

[69] Alexander Strang, David Sewell, Alexander Kim, Kevin Alcedo, and David Rosenbluth. Principal trade-off analysis. *Information Visualization*, 23(3), 2024. 9

[70] Shayan Fazeli. ECG heartbeat categorization dataset. *Kaggle*, 2018. URL `https://www.kaggle.com/datasets/shayanfazeli/heartbeat/data?select=ptbdb_abnormal.csv`. 10, 23

[71] Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. ECG heartbeat classification: A deep transferable representation. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 443–444, 2018. 10, 23

[72] R Bousseljot, D Kreiseler, and A. Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Biomedizinische Technik*, 40(1), 1995. 10, 23

[73] I. Di Terlizzi, M. Gironella, D. Herraez-Aguilar, T. Betz, F. Monroy, M. Baiesi, and F. Ritort. Data from: Variance sum rule for entropy production. *Zenodo*, 2024. doi: https://doi.org/10.5281/zenodo.10472505. 10, 23

[74] N. G. van Kampen. Itô versus Stratonovich. *Journal of Statistical Physics*, 24:175–187, 1981. 15

[75] Christopher C. Paige and Michael A. Saunders. LSQR: Sparse linear equations and least squares problems. *ACM Transactions on Mathematical Software*, 8(2):195—209, 1982. 16

[76] J.C. Mattingly, A.M. Stuart, and D.J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101: 185–232, 2002. 16, 17

[77] Evelyn Buckwar, Adeline Samson, Massimiliano Tamborrino, and Irene Tubikanec. A splitting method for SDEs with locally Lipschitz drift: Illustration on the FitzHugh-Nagumo model. *Applied Numerical Mathematics*, 179:191–220, 2022. 16, 17, 18

[78] Ernst Hairer, Gerhard Wanner, and Christian Lubich. *Geometric Numerical Integration, Structure-Preserving Algorithms for Ordinary Differential Equations*, chapter Numerical integrators. Springer, 2006. 16, 17, 18

[79] Martin Hutzenthaler, Arnulf Jentzen, and Peter E. Kloeden. Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients. *Annals of Applied Probability*, 22(4):1611–1641, 2012. 17, 18

A Appendix

A.1 Details on the Helmholtz-Hodge decomposition

For a vector field $\mathbf{F} \in C^1(V, \mathbb{R}^n)$ where $V \subseteq \mathbb{R}^n$ that decays sufficiently quickly, the HHD is a division of the vector field, that may not be unique, into two components,

$$\mathbf{F}(\mathbf{r}) = -\nabla\Phi(\mathbf{r}) + \mathbf{R}(\mathbf{r}), \tag{13}$$

where $\Phi \in C^2(V, \mathbb{R})$ is a scalar potential, ∇ is the gradient operator and $\mathbf{R}(\mathbf{r})$ is divergence-free i.e. $\nabla \cdot \mathbf{R} = 0$ [40]. In three dimensions, assuming the vector fields decay sufficiently quickly at infinity, the divergence-free force can be written as the curl of a vector potential $\mathbf{A}$, i.e. $\mathbf{R} = \nabla \times \mathbf{A}$.

A.1.1 The HHD of a stationary Langevin process

As detailed in Eq. 1, the Langevin process is given by,

$$\dot{\mathbf{x}}(t) = \boldsymbol{F}(\mathbf{x}(t)) + \boldsymbol{\xi}(t), \tag{14}$$

where $\boldsymbol{F}$ is the drift field, whilst $\boldsymbol{\xi}$ are Gaussian fluctuations with spatially-homogeneous correlations given by $\langle \boldsymbol{\xi}(t), \boldsymbol{\xi}(t') \rangle = 2\boldsymbol{D}\delta(t - t')$ with $\boldsymbol{D}$ the constant, positive definite, noise-covariance matrix. As shown in Eq. 2 and Eq. 3, the density dynamics can be written as a conservation of probability law,

$$\frac{\partial P}{\partial t} + \nabla \cdot \boldsymbol{J} = 0, \tag{15}$$

where $\boldsymbol{J}(\mathbf{x}, t)$ is the probability flux defined as,

$$\boldsymbol{J} = \boldsymbol{F}P - \nabla \cdot (\boldsymbol{D}P), \tag{16}$$

and $\nabla\cdot$ represents the divergence operator, which, when defined on a vector field $\boldsymbol{A} : \mathbb{R}^N \to \mathbb{R}^N$, returns a scalar quantity,

$$\nabla \cdot \boldsymbol{A}(\mathbf{x}) = \sum_i \partial \boldsymbol{A}/\partial x_i, \tag{17}$$

where $\mathbf{x} = (x_1, ..., x_N)$. When defined on a tensor field, $\boldsymbol{D} : \mathbb{R}^N \to \mathbb{R}^N \times \mathbb{R}^N$, it returns a vector in $\mathbb{R}^N$ whose i-th entry is given by,

$$(\nabla \cdot \boldsymbol{D}(\mathbf{x}))_i = \sum_j \frac{\partial D_{ij}(\mathbf{x})}{\partial x_j}. \tag{18}$$

At stationarity, i.e. $\partial P/\partial t = 0$, the density of the process remains constant whilst the trajectories still evolve in time. The probability flux at stationarity, $\boldsymbol{J}_{ss}$, is divergence-free, $\nabla \cdot \boldsymbol{J}_{ss} = 0$. The drift at stationarity can be written as [18],

$$\boldsymbol{F} = \frac{\boldsymbol{J}_{ss}}{P_{ss}} + \boldsymbol{D} \cdot \nabla \log P_{ss} + \nabla \cdot \boldsymbol{D}, \tag{19}$$

where ∇ represents the gradient operator, which takes a scalar potential, $U : \mathbb{R}^N \to \mathbb{R}$ and returns a vector field,

$$\nabla U(\mathbf{x}) = \sum_i \frac{\partial U}{\partial x_i} \mathbf{e}_i, \tag{20}$$

where $\mathbf{e}_i$ is the $i-$th unit vector i.e. $(\mathbf{e}_i)_j = \delta_{ij}$. Eq. 19 decomposes the drift field into the sum of a curl flux, $\boldsymbol{J}_{ss}/P_{ss}$, which is divergence-free, and the gradient of the scalar potential, $U = \log P_{ss}$, multiplied by the diffusion $\boldsymbol{D}$. Note that, whilst we assume spatially homogeneous diffusion, we include the term $\nabla \cdot \boldsymbol{D}(= 0)$ for completeness. This term is also the correction factor between the Îto and Stratonovich formulations of spatially-dependent diffusions [74]. As a result, the stationary distribution and diffusive field give us the unique HHD of the drift. Moreover, in the case of a isotropic diffusion, $\boldsymbol{D} = \epsilon \boldsymbol{I}$, the reversible dynamics are exactly the gradient of the scalar potential $\log \epsilon P_{ss}$, and the drift is given by,

$$\boldsymbol{F} = \boldsymbol{J}_{ss}/P_{ss} + \nabla \log \epsilon P_{ss}. \tag{21}$$

As a result, the detailed balance condition, $\boldsymbol{J}_{ss} \equiv \mathbf{0}$, which guarantees reversible dynamics and an equilibrium stationary distribution, is met if and only if the drift field is conservative, i.e. there exists a scalar potential V such that $\nabla V = \boldsymbol{F}$. This condition can be used to define a discrete-analog of the drift for the discrete-state Markov process and discrete gradient operator.

A.1.2 The HHD of a edge-flow on a simplicial complex

Given a simplicial complex with vertices, edges and triangles, V, E, T respectively, we define the set of flows on each set to be $L^2(V), L^2_\wedge(E), L^2_\wedge(T)$ respectively, where $\wedge$ indicates that the flows are alternating. Next, we define inner-products on these spaces. We take unweighted inner products, $\langle f, g\rangle = \sum_i f(i)g(i)$, $\langle X, Y\rangle = \sum_{i<j} X(i,j)Y(i,j)$ and $\langle \Phi, \Theta\rangle = \sum_{i<j<k} \Phi(i,j,k)\Theta(i,j,k)$ for the spaces of vertices, edges and triangles respectively. Using these inner products, we define the *gradient*, *curl* and *divergence* operators acting between these function spaces,

$$(\text{grad } f)(i,j) = f(j) - f(i), \tag{22}$$

$$(\text{curl } X)(i,j,k) = X(i,j) + X(j,k) + X(k,i), \tag{23}$$

$$(\text{div } X)(i) = \sum_j X(i,j). \tag{24}$$

where grad : $L^2_\wedge(V) \to L^2_\wedge(E)$, curl : $L^2_\wedge(E) \to L^2_\wedge(T)$ and div : $L^2_\wedge(E) \to L^2_\wedge(V)$ [31]. From these definitions, we are able to define the curl adjoint, curl* : $L^2_\wedge(T) \to L^2_\wedge(E)$, which is given by,

$$(\text{curl}^* \ \Phi)(i,j) = \sum_k \Phi(i,j,k). \tag{25}$$

Thus, these operators define the following relationship between the function spaces,

$$L^2_\wedge(V) \underset{\text{div}}{\overset{\text{grad}}{\rightleftarrows}} L^2_\wedge(E) \underset{\text{curl}^*}{\overset{\text{curl}}{\rightleftarrows}} L^2_\wedge(T). \tag{26}$$

Additionally, we can define the *graph-Helmholtzian*, $\Delta_1 : L^2_\wedge(E) \to L^2_\wedge(E)$, a discrete analogue of the vector Laplacian, defined by,

$$\Delta_1 = -\text{grad div} + \text{curl}^* \text{ curl}. \tag{27}$$

The HHD associated with this operator states that the space of edge-flows admits a decomposition into orthogonal subspaces,

$$L^2_\wedge(E) = \text{im}(\text{curl}^*) \bigoplus \text{im}(\text{grad}) \bigoplus \ker(\Delta_1). \tag{28}$$

As a result, the HHD states that any edge-flow X can be decomposed as follows,

$$X = \text{curl}^* \Phi + \text{grad} f + X_H, \tag{29}$$

where curl$^*\Phi$ is a divergence-free flow which is the curl-adjoint of a flow, Φ, defined on the triangles, grad$f$ is the curl-free gradient of a flow on the vertices and X_H is a harmonic flow on the edges i.e. $(\Delta_1(X_H) = \mathbf{0})$. A gradient flow is induced by assigning a potential to each of the nodes and then defining the edge-flow according to the difference of the potentials along each edge, thus gradient flows have zero net sum along any closed path in the graph. The space of non-gradient flows contains cyclical flows that are either curling, representing local circulations around triangles, or harmonic, representing global circulations around *holes* in the complex [29]. As there are no holes in the

Delauney triangulation, the harmonic flow is zero [31], thus we see the HHD that we will aim to calculate is,

$$X = \text{curl}^* \Phi + \text{grad} f, \tag{30}$$

which mirrors the decomposition of the stationary Langevin process with isotropic diffusion.

The HHD of a edge-flow can be computed efficiently as a pair of sparse least-squares problems, one for the gradient component and one for the curling component [31]. As we have no harmonic flow, we need only solve one problem and obtain the other component as the residual of the edge-flow and first component. We will solve for the gradient component. We first note that, the space of functions on vertices, edges and triangles are each finite-dimensional and permit the following isomorphisms $L^2(V) \cong \mathbb{R}^{|V|}$, $L^2_\wedge(E) \cong \mathbb{R}^{|E|}$, and $L^2_\wedge(T) \cong \mathbb{R}^{|T|}$, where $|V|, |E|, |T|$, are the number of vertices, edges and triangles respectively. As a result, flows can be uniquely expressed as a vector whilst the discrete operators that project between these spaces can be expressed as sparse matrices. Utilising that the decomposition is orthogonal, this allows us to compute the decomposition by solving the sparse least-squares problem with an algorithm such as LSQR [75],

$$f = \arg\min_{g \in \mathbb{R}^{|V|}} ||\text{grad } g - X||^2 \tag{31}$$

We can recover the curl part of the flow by taking,

$$\text{curl}^* \Phi = X - \text{grad } f. \tag{32}$$

A.2 Example stochastic processes

Example 1. The multivariate Ornstein-Uhlenbeck process

Our first example process is a linear Langevin equation,

$$\frac{d\mathbf{x}}{dt} = -\left[\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}\right]\mathbf{x} + \boldsymbol{\xi}, \tag{33}$$

with isotropic diffusion given by,

$$\boldsymbol{D} = \epsilon \boldsymbol{I}. \tag{34}$$

The parameter θ controls the level of rotation in the drift with increases in θ resulting in increased rotation and $\theta = 0$ corresponding to a reversible process with only gradient flow. The diffusion parameter ϵ controls the intensity of the noise. The stationary distribution of this process is a zero-mean multivariate Gaussian, thus it permits an exact HHD (for a general treatment of the OU process see App. A.7 and Ref. [34, 52]). The system also permits an exact simulation from its transition probability density (see App. A.3) [34].

Example 2. The stochastic planar limit cycle

Next, we consider the nonlinear process,

$$\dot{x} = -\theta y + x(1 - x^2 - y^2) + \xi_x, \tag{35}$$

$$\dot{y} = \theta x + y(1 - x^2 - y^2) + \xi_y, \tag{36}$$

with isotropic diffusion given by $\langle \xi_i(t), \xi_j(t') \rangle = 2\epsilon\delta_{ij}\delta(t - t')$. In the absence of diffusion, this system converges to a stable circular limit-cycle [53]. Additionally, it admits a 'Mexican-hat' potential of the form $\Phi(x, y) = (x^2 + y^2)(x^2 + y^2 - 2)/4\epsilon$ which is associated with the Boltzmann-Gibbs stationary distribution,

$$P_{ss}(x, y) = \frac{1}{Z_\epsilon} \exp\left(-\frac{(x^2 + y^2)(x^2 + y^2 - 2)}{4\epsilon}\right), \tag{37}$$

where the partition function is $Z_\epsilon = e^{1/4\epsilon}\sqrt{\epsilon}\pi^{3/2}(1 + \text{erf}[1/(2\sqrt{\epsilon})])$ [53]. Given the stationary distribution, we decompose the drift into,

$$\boldsymbol{F}_{\text{rev}} = \begin{pmatrix} x(1 - x^2 - y^2) \\ y(1 - x^2 - y^2) \end{pmatrix}, \quad \boldsymbol{F}_{\text{irrev}} = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}. \tag{38}$$

The parameter θ controls the degree to which the trajectories rotate and thus contour the stationary distribution. In order to sample from this model we avoid traditional finite-difference schemes, such as the Euler-Maruyama scheme, which do not preserve the ergodic, oscillatory or nonequilibrium dynamics of stochastic systems [34, 76, 77]. Instead, we employ geometric splitting integration (see A.3) [78] that preserves the structure of the stochastic oscillatory dynamics [77].

Example 3. The stochastic modified van der Pol oscillator

We consider a modified version of the stochastic VDP limit-cycle oscillator [54],

$$\dot{x} = \theta y + \xi_x, \tag{39}$$

$$\dot{y} = -\theta x + \mu y(1 - x^2) + \xi_y, \tag{40}$$

with diffusion given by $\langle \xi_i(t), \xi_j(t') \rangle = 2\epsilon\delta_{ij}\delta(t - t')$. In the absense of noise, this system converges to an asymmetric limit-cycle. Whilst we do not have access to the stationary distribution and the exact HHD in this case, we can use that increasing θ creates a stronger rotational component to the drift field as a approximate control for the curling part of the drift. Setting $\theta = 1$ corresponds to original VDP oscillator.

Example 4. The stochastic modified Rössler attractor

Finally, we consider a modified version of the Rössler system [55] given by the 3-dimensional system,

$$\dot{x} = -\theta y - z + \xi_x, \tag{41}$$

$$\dot{y} = \theta x + ay + \xi_y, \tag{42}$$

$$\dot{z} = b + z(x - c) + \xi_z, \tag{43}$$

where the diffusion is given by $\langle \xi_i(t), \xi_j(t') \rangle = 2\epsilon\delta_{ij}\delta(t - t')$. With the parameters $a = 0.2$, $b = 0.2$, $c = 5.7$, $\theta = 1$ and $\epsilon = 1$, the system exhibits attracting chaotic dynamics. The parameter θ is a modification that controls the level of rotation in the $xy-$plane. We sample from this process using the mean-square convergent *tamed Euler-Maruyama* scheme (see A.3) [79].

A.3 Sampling from Langevin processes

In Sec. 3.1, we validate our approach with trajectories sampled from a range of Langevin processes. In order for such simulations to validate the theory, we take care to use simulation techniques that preserve the underlying structure of the dynamics and stationary distribution.

A.3.1 Exact simulation of the Ornstein-Uhlenbeck

Due to its linearity, the OU process permits an exact simulation from its stationary distribution [34]. Given a Langevin process,

$$\frac{d\boldsymbol{x}}{dt} = -\boldsymbol{B}\mathbf{x}(t) + \xi(t), \tag{44}$$

with $\langle \boldsymbol{\xi}(t), \boldsymbol{\xi}(t') \rangle = 2\boldsymbol{D}\delta(t - t')$, the solution is given by,

$$\mathbf{x}(t) = e^{-\boldsymbol{B}t}\mathbf{x}(0) + \int_0^t e^{-\boldsymbol{B}(t-s)}\Sigma \, d\boldsymbol{W}(s), \tag{45}$$

where $\Sigma\Sigma^\top = 2\boldsymbol{D}$. Therefore, the transition kernel is given by,

$$\mathbf{x}(t) \sim \mathcal{N}(e^{-\boldsymbol{B}t}\mathbf{x}(0), \boldsymbol{C}(t)), \tag{46}$$

where,

$$\boldsymbol{C}(t) = \int_0^t 2e^{-\boldsymbol{B}(t-s)}\boldsymbol{D}(e^{-\boldsymbol{B}(t-s)})^\top \, ds, \tag{47}$$

thus the process can be simulated exactly.

A.3.2 Splitting integrators for oscillatory processes

In general, nonlinear processes do not admit exact simulation and we must employ numerical methods to simulate trajectories. However, standard techniques such as the Euler-Maruyama scheme do not preserve the underlying dynamics of stiff oscillatory systems [77] and are not mean-square convergent [76], thus are unsuitable. We employ a geometric splitting integrator for oscillatory SDEs [77, 78]

which preserves the underlying dynamics, in order to simulate from both the PLC and the VDP. Given a nonlinear Langevin process,

$$\frac{d\boldsymbol{x}}{dt} = F(\mathbf{x}(t)) + \xi(t), \tag{48}$$

with $\langle \boldsymbol{\xi}(t), \boldsymbol{\xi}(t') \rangle = 2\boldsymbol{D}\delta(t-t')$, we 'split' the drift term into a nonlinear, deterministic part and a linear stochastic part [77],

$$\mathbf{x}^{[1]}(t) = \boldsymbol{A}\mathbf{x}(t) + \xi(t), \tag{49}$$

$$\mathbf{x}^{[2]}(t) = \boldsymbol{G}(\mathbf{x}(t)), \tag{50}$$

where $\boldsymbol{F}(\mathbf{x}) = \boldsymbol{A}\mathbf{x} + \boldsymbol{G}(\mathbf{x})$. The linear part, $\mathbf{x}^{[1]}$, defines an OU process which can be simulated exactly. If the splitting is chosen such that the ordinary differential equation for $\mathbf{x}^{[2]}$ can be solved explicitly, then the solution of the full process can be approximated with the Strang approximation [78],

$$\mathbf{x}(t+h) \approx \varphi_1^{[h/2]} \circ \varphi_2^{[h]} \circ \varphi_1^{[h/2]}(\mathbf{x}(t)), \tag{51}$$

where $\varphi_i^{[h]}$ is a simulation from the exact h-time transition kernel of $\mathbf{x}^{[i]}$.

For the PLC, we use the decomposition,

$$\mathbf{x}^{[1]}(t) = \begin{pmatrix} 1 & -\theta \\ \theta & 1 \end{pmatrix} \mathbf{x}(t) + \xi(t), \tag{52}$$

$$\mathbf{x}^{[2]}(t) = \begin{pmatrix} -(x^2+y^2)x \\ -(x^2+y^2)y \end{pmatrix}, \tag{53}$$

where the nonlinear part can be solved easily in polar coordinates.

For the VDP, we use the decomposition ,

$$\mathbf{x}^{[1]}(t) = \begin{pmatrix} 0 & \theta \\ -\theta & \mu \end{pmatrix} \mathbf{x}(t) + \xi(t), \tag{54}$$

$$\mathbf{x}^{[2]}(t) = \begin{pmatrix} 0 \\ -\mu x^2 y \end{pmatrix}, \tag{55}$$

where the nonlinear part can be solved trivially due to the constancy of the first component x.

A.3.3 Tamed Euler-Maruyama for the Rössler attractor

Due to the difficulties in decomposing the Rössler system, we instead opt for the mean-square convergent, tamed Euler-Maruyama scheme [79],

$$\mathbf{x}(t+h) \approx \mathbf{x}(t) + h\frac{\boldsymbol{F}(\mathbf{x}(t)}{1+||\boldsymbol{F}(\mathbf{x}(t)||} + \Sigma\boldsymbol{\nu}_h, \tag{56}$$

where $\Sigma\Sigma^\top = 2\boldsymbol{D}$ and $\boldsymbol{\nu}_h$ is a Gaussian vector with independent components each with zero-mean and variance h.

A.4 Parameters for numerical simulations and data analysis

For the numerical experiments presented in panel C of Fig. 3, we utilised 500 trajectories of length 10,000 for each process. For the OU process, we utilised $h = 0.01, \epsilon = 0.5$ with 250 sub-sampled points. We swept θ between 0 and 15 with 101 equispaced values. For the PLC process, we utilised $h = 0.01, \epsilon = 0.25$ with 250 sub-sampled points. We swept θ between 0 and 10 with 100 equispaced values. For the VDP process, we utilised $h = 0.01, \epsilon = 0.25$ with 250 sub-sampled points. We swept θ between 1 and 7 with 90 equispaced values. For the RA process, we utilised $h = 0.01, \epsilon = 0.075$ with 250 sub-sampled points. We swept θ between 1 and 4 with 100 equispaced values. For the analysis of both RBCs and heartbeats, we used 250 sub-sampled points. The diffusion estimator used an arbitrary time-step of $h = 0.01$.

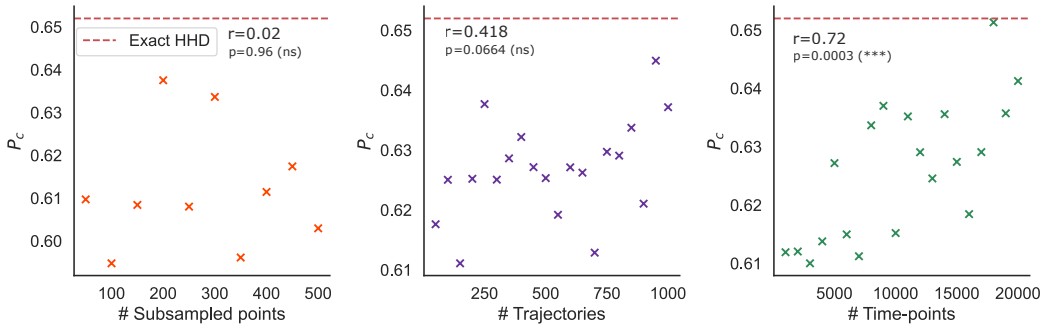

Figure 5: Variation of method parameters. Here we show the effect of varying the number of sub-sampled points used to construct the grid and graph, the number of trajectories in an ensemble and the number of time-points in a trajectory, using the OU process with $\theta = 4$ as an example. The estimate from the exact decomposition at $\theta = 4$ is 0.652 (3.s.f), as shown by the red dashed line.

A.5 Variation of method parameters

The method we propose depends on a number of key parameters such as the number of sub-sampled points used to construct the triangulations, the number of trajectories in an ensemble and the number of time-points in a trajectory. In Fig. 5 we consider the OU process fixed at a value of $\theta = 4$ and vary, independently, the number of sub-sampled points, the number of trajectories and the length of the trajectories. The relationship between these parameters is complex, unlike in typical numerical schemes. For example, simply increasing the number of sub-sampled points, thereby refining the grid, will result in fewer points per bin/triangle and a worse estimate of the transition intensity or diffusion matrix. On the other hand, using fewer points will result in coarse-grid that does not reflect state-space correctly. It is due to this added complexity, that we refrain from attempting a proof of convergence that would require taking both the infinite data and the infinite point limit. In Fig. 5, we see that there is a fluctuation of P_c with the the number of sub-sampled points, but no correlation. In the case of the number of trajectories or the length of the trajectories, we would expect increasing this parameter to result in an improved estimate of P_c. As shown in Fig. 5, the value of P_c appears to increase with both the number of time-points per trajectory ($p = 0.003$, 3.s.f) and the number of trajectories ($p = 0.0664$, 3.s.f) which implies that they are approaching the true value which can estimated to be $P_c = 0.652$ (3.s.f) from the exact decomposition at $\theta = 4$.

A.6 Details on real-world data

A.6.1 Time-delay embedding for phase-space reconstruction

Phase-space coordinates can be reconstructed from scalar measurements via a time-delay embedding [58]. For both the RBC and heartbeat analysis, we have a univariate time-series recording a single scalar measurement from the system. In the case of the RBC, this is the displacement of a single point chosen on the outer membrane. For the heartbeat, this is an electrocardiogram (ECG). However, as mentioned in Sec. 3, both systems are truly two dimensional as the RBC has an unmeasured inner membrane displacement [21] whilst heartbeat dynamics are known to obey two dimensional limit cycle dynamics [5]. Assuming we have access to coordinate $x(t)$ with unseen coordinate $y(t)$, we approximate the phase-space $(x(t), y(t))$ with $(x(t), x(t+\tau))$ where τ is known as the *lag* or *delay*. In order to select τ, we calculate the first lag at which the auto-correlation crosses 0 for each trajectory. We take the mean value over all trajectories in the dataset and fix τ for the remainder of the experiment. For the RBC experiments we obtain $\tau = 998, (T = 10,000)$, whilst for the heartbeats we obtain $\tau = 45, (T = 188)$.

A.6.2 Spatially-dependent diffusion in real-world data

The framework presented here assumes that diffusion is not spatially-dependent and thus can be represented by a single matrix $\boldsymbol{D}$. Here, we briefly assess whether this is the case in the real-world data by leveraging the spatially-dependent estimator $\hat{\boldsymbol{D}}$ of Eq. 12. Panel A of Fig. 6 shows the coefficients of the diffusion estimator in each of the triangular bins, estimated using the RBC 'Healthy

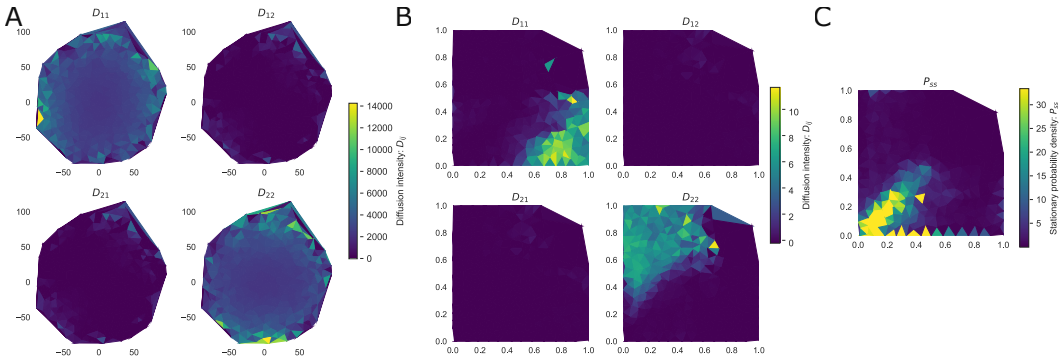

Figure 6: Diffusion estimation in real-world data. A. An example of the spatially-dependent diffusion estimator in the RBC 'Healthy 1'. **B.** An example of the spatially-dependent diffusion estimator in the healthy heartbeats. **C.** A maximum-likelihood estimation of the stationary probability distribution P_{ss} in the healthy heartbeats.

1'. The diffusion appears to be spatially-independent, with more variation at the borders where the trajectories have fewer points leading to higher estimation error. Moreover, the diffusion is close to isotropic. Panel B shows the diffusion estimator calculated on the healthy heartbeats. This appears to show more spatial dependence whilst remaining mostly isotropic. Nevertheless, by estimating the stationary distribution with maximum-likelihood estimation, shown in panel C, we can see that the trajectories spend relatively little time in the parts of phase-space that diffusion variation which indicates that this variation could arise from the numerical error associated with having only a few samples in this area of phase-space rather than a genuine spatial dependence.

A.6.3 Ensemble bootstrapping

The finite length of the trajectories limits the accuracy to which the force fields can be reconstructed and decomposed. Furthermore, as our approach ingests the ensemble of trajectories in order to reconstruct flows, we obtain a single number for each condition. In order to estimate these errors and thus to perform significance testing, we employ ensemble bootstrapping [16, 17]. Each ensemble is of the form $\{\mathbf{x}_{t_1,r}, ..., \mathbf{x}_{t_T,r}\}_{r=1}^{R}$. We construct a bootstrapped ensemble by sampling trajectories from the original ensemble with replacement. We use the same number of trajectories in each ensemble and use 100 bootstraps resulting in 101 ensembles in each condition. As shown in Fig. 4, this gives a distribution of P_c, upon which we perform independent t-tests for significance.

A.6.4 Surrogate-testing with shuffled data

In order to assess whether the proportion of curl-flow measured is due to the statistical structure of the time-series or due to a combination of chance and the finite-length of the data, we compare our time-series to surrogate 'null' data. In order to break the temporal ordering of the time-series, we shuffle randomly over the time-points. This is in contrast with other typical randomisation techniques, such as phase-randomisation, which do not disrupt the temporal ordering sufficiently [17]. For each condition in each dataset, we construct a 100 bootstrapped ensembles of delay-embedded trajectories. We then shuffle the temporal ordering of the points in phase-space randomly in time. Next we apply our approach to both the original and shuffled ensemble to calculate two measures of P_c. We then perform paired $t-$tests to identify if the statistical structure of the time-series results in higher or lower P_c than could be expected by chance. In panel A of Fig. 7, we present the results for the RBCs and in panel B we present the results for the healthy and arrhythmic heartbeats. We find that all healthy RBCs have significantly higher P_c that the shuffled time-series indicating a high prevalence of curling flow, whilst the passive cell have significantly lower P_c than the shuffled time-series indicating a very lower prevalence of curling flow. Similarly, for the healthy heartbeats, we find a significantly high level of P_c compared to shuffled data, whilst for arrhythmic heartbeats, we find that P_c is comparable between the null data and the original series.

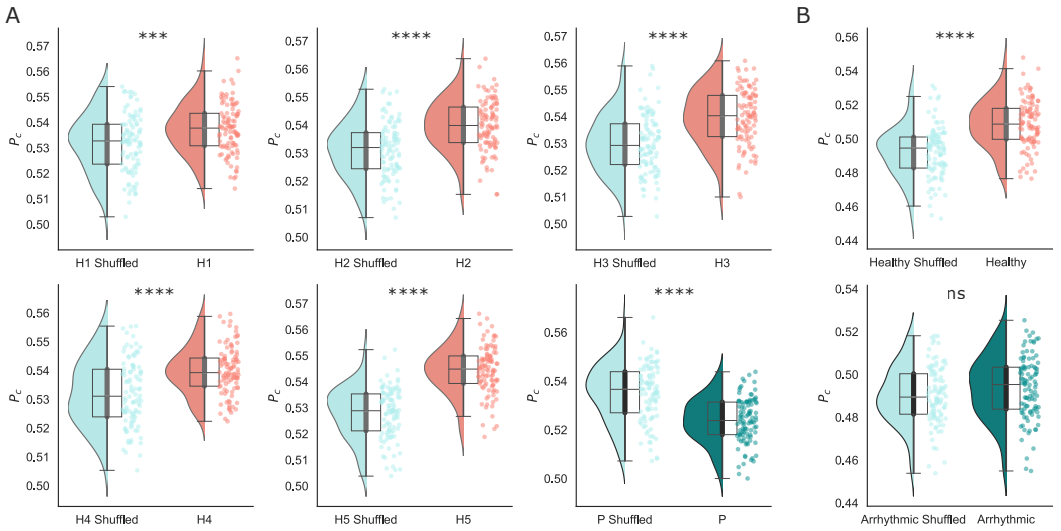

Figure 7: Surrogate-testing with shuffled time-series: For each condition in either dataset, we construct 100 bootstraps. For each bootstrap, we shuffle each of the trajectories randomly in time, apply the method, and calculate the proportion of flow in the curl space for both the original and shuffled data. We then run paired $t-$tests to check if the proportion of curl flow is significantly higher or lower than a random sample. Panel A shows the result for the five healthy and one passive RBC. Panel B shows the result for the healthy and arrhythmic heartbeats. ($p < 0.001$,***), ($p < 0.0001$, ****) and ($p > 0.05$, ns).

A.6.5 Variation of the time-delay parameter

When analysing the real-world data, we used a single lag for each experiment that was chosen to be the mean first-zero-crossing of the autocorrelation function, which is a conventional choice [58]. For the RBC experiments we obtain $\tau = 998, (T = 10,000)$, whilst for the heartbeats we obtain $\tau = 45, (T = 188)$. In Fig. 8, we consider the effect of varying this parameter. For the RBCs, we consider lags of 250, 500, 750, 1000 and 1250. For the heartbeats, we consider lags of 15, 30, 45 and 60. For the RBCs, panel A shows that variation in the lag parameter shows no significant change in the distribution of P_c over 100 boot-strapped samples, whilst the relative relationship between healthy and passive/fixed cells is kept the same. In panel B, we can see that variation in the lag parameter can cause greater variation in P_c for the heartbeat dynamics. This is because the heartbeat data has a relatively short-time series which is made even shorter by an increase in the lag. This can increase the error in the measurements and lead to greater variation. Nevertheless, the qualitative relationship between normal and arrhythmic heartbeats appears to hold. This analysis suggests that our approach is not very sensitive to changes in the lag parameter provided the lag is relatively small in comparison with the length of the time-series.

A.6.6 Classifying time-series with standard methods

Whilst we have shown that our method can differentiate between recordings from biological systems in different conditions, our main goal was to gain insight into the underlying reversible and irreversible dynamics directly from stochastic trajectories. Our approach allows for insight into the thermodynamics and nonequilibrium physics of the biological processes, and goes beyond a classification tool [18]. Measuring P_c, the proportion of the flow in the curl space, is a first application that allows us to validate that our approach does capture irreversible flows by confirming results from the literature [37, 38]. We note that differentiating between conditions using the time-series data considered here is not a challenging task in itself, as we will show in this section. Beyond the prevalence of irreversible flows, the underlying time-series have significant differences in their statistics, that allow for easy classification. To illustrate this, we consider a neural-network classifier and show that it achieves (almost) perfect classification of recordings when trained on simple time-series summary statistics. For each of recording in the dataset, we compute a battery of trajectory features in the form of summary statistics. In particular, we consider the mean, median,

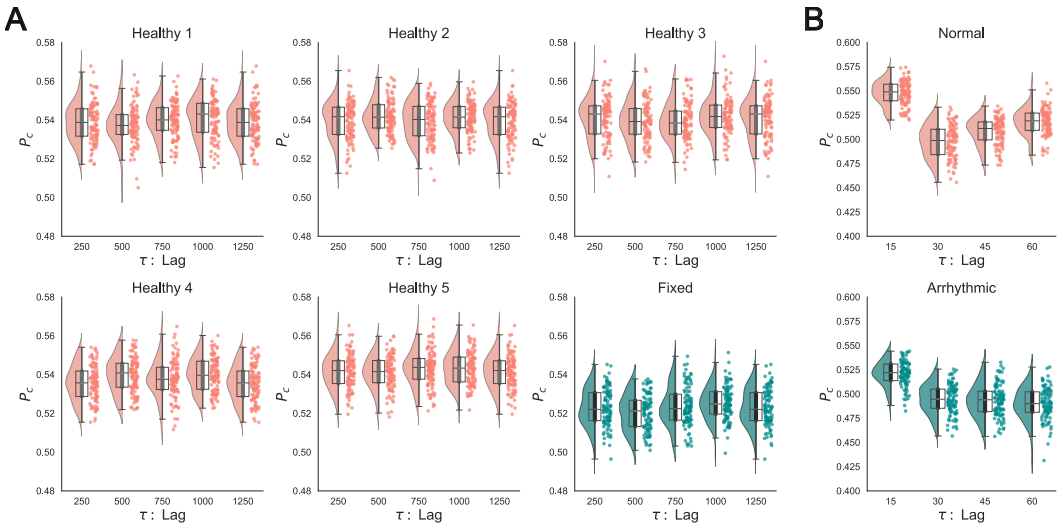

Figure 8: Variation of the lag parameter, τ, in the delay embedding. A. Varying the lag parameter in the RBC experiments shows little to no change suggesting that the phase-space dynamics are comparable for a range of different lags. B. In the case of the heartbeat dynamics, the time-series are much shorter as a result, varying the lag parameter causes variation in the quantity, but the relative relationship between normal and arrhythmic heartbeats is preserved.

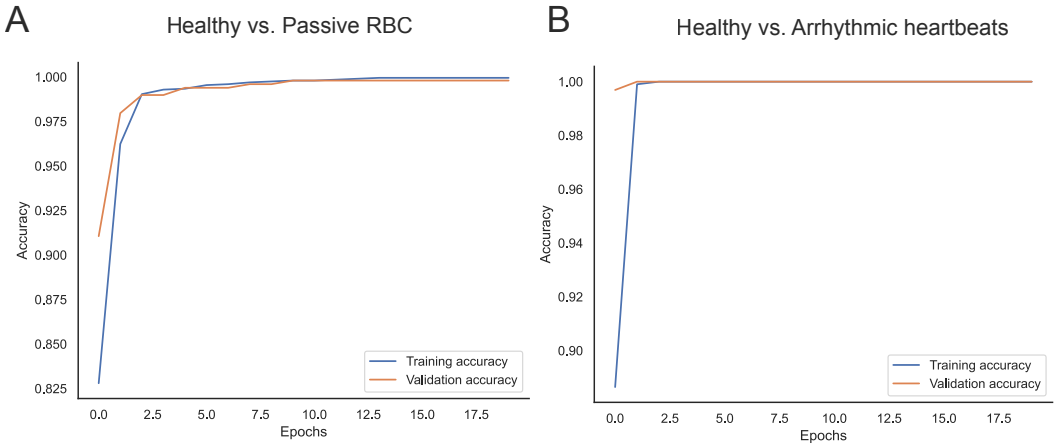

Figure 9: Classifying recordings with a neural network: We attempt to classify the time-series data considered here with a neural network trained on the summary statistics of the data. We find that for both the RBCs and the heartbeats, the neural network achieves almost perfect classification after very little training. This is due to clear differences in the summary statistics that allow for easy classification.

standard deviation (Std), minimum value (Min), maximum value (Max), sum, skew, kurtosis and last value of each signal. For the RBCs, recordings are either healthy or passive, whilst for heartbeats, recordings are either normal or arrhythmic. We then train a neural-network on the summary statistics of a subset of the recordings with the goal of classifying each recording into its respective category. Our network has 2 hidden layers with 32 and 16 'ReLu' units respectively and an output layer with sigmoidal activation. We use a binary cross-entropy loss function and the Adam optimiser. We train the network for 20 epochs with a validation split of 0.2 and a batch size of 50. In addition, we scale all recordings between [0,1] before training and use class-weighting to account for the different densities of the classes in the dataset. As shown in Fig. 9, we find that the neural network has (almost) perfect accuracy after only a few epochs of training. On the test set, we find that it classified both RBC and heartbeat recordings with 100% accuracy. Fig. 10 shows the distribution of the summary statistics, after data normalisation, for both healthy and passive cells. A subset of the statistics, namely the standard deviation, Min, Max and kurtosis are significantly different between healthy and passive cells, therefore, these features allow for simple classification of trajectories. Similarly, Fig. 11 shows the corresponding summary statistics for healthy and arrhythmic heartbeats, where the mean, median, standard deviation, sum, skew and kurtosis are all found to be significantly different.

Our approach is also able to distinguish between nonequilibrium dynamics such as those found in healthy RBCs, and the equilibrium dynamics of surrogate data generated by shuffling recordings in time. All the summary statistics considered in the previous experiment are invariant under shuffling of the data in time, which suggests that such trajectory features are insufficient for such differentiation. However, one can select summary statistics that depend on the temporal ordering, such as auto-correlation, to distinguish true and shuffled recordings. In Fig. 12, we show that the auto-correlation of data from Healthy 1, the RBC, is significantly higher than a surrogate generated by shuffling the data in time. This allows for the easy distinguishing of true dynamics from shuffled counterparts based on simple summary statistics.

A.6.7 Flickering experiments with red-blood cells

RBCs metabolise glucose causing the cell membrane to actively *flicker* with the dissipated energy and the created entropy [21, 35, 36]. By depleting the ATP of a healthy RBC, one obtains a *passive* cell whose flickering is closer to reversible thermal noise [21, 57]. The position of the RBC membrane is captured using optical microscopy to form movies at a sampling rate of 2000 frames per second. The position of a point on the membrane is extracted to give position (in nm) traces with 10,000 time-points. The dataset includes 512 position traces from each of five healthy cells and one passive cell. Further discussion of the data can be found in Ref. [21, 57] and the data is openly available at Ref. [73].

A.6.8 Healthy and arrhythmic heartbeats

Human heartbeats can be recorded using an ECG, which measures the voltage of electrical signals in the heart over time. Previous research has shown that the time-irreversibility of these signals is disrupted during aging and disease [38]. The dataset considered here is the PTB Diagnostic ECG Database [56, 72] as pre-processed in [71]. The dataset contains time-series for individual extracted beats with either 'normal' (healthy volunteers) or 'abnormal' (myocardial infarction) dynamics. The dataset contains 4046 normal heartbeats and 10,506 abnormal heartbeats each composed of 188 time-points at a sampling frequency of 125Hz. We randomly sampled 4046 of the abnormal heartbeats for our analysis. The data is openly available at Ref. [70].

A.7 The Helmholtz-Hodge decomposition of the Ornstein-Uhlenbeck process

The OU process is given by,

$$\frac{d\mathbf{x}}{dt} = -\boldsymbol{B}\mathbf{x}(t) + \xi(t), \tag{57}$$

with $\langle \boldsymbol{\xi}(t), \boldsymbol{\xi}(t') \rangle = 2\boldsymbol{D}\delta(t - t')$. If the real-part of the eigenvalues of $\boldsymbol{B}$ are all positive, then the process will converge to a Gaussian stationary distribution with mean zero and covariance satisfying

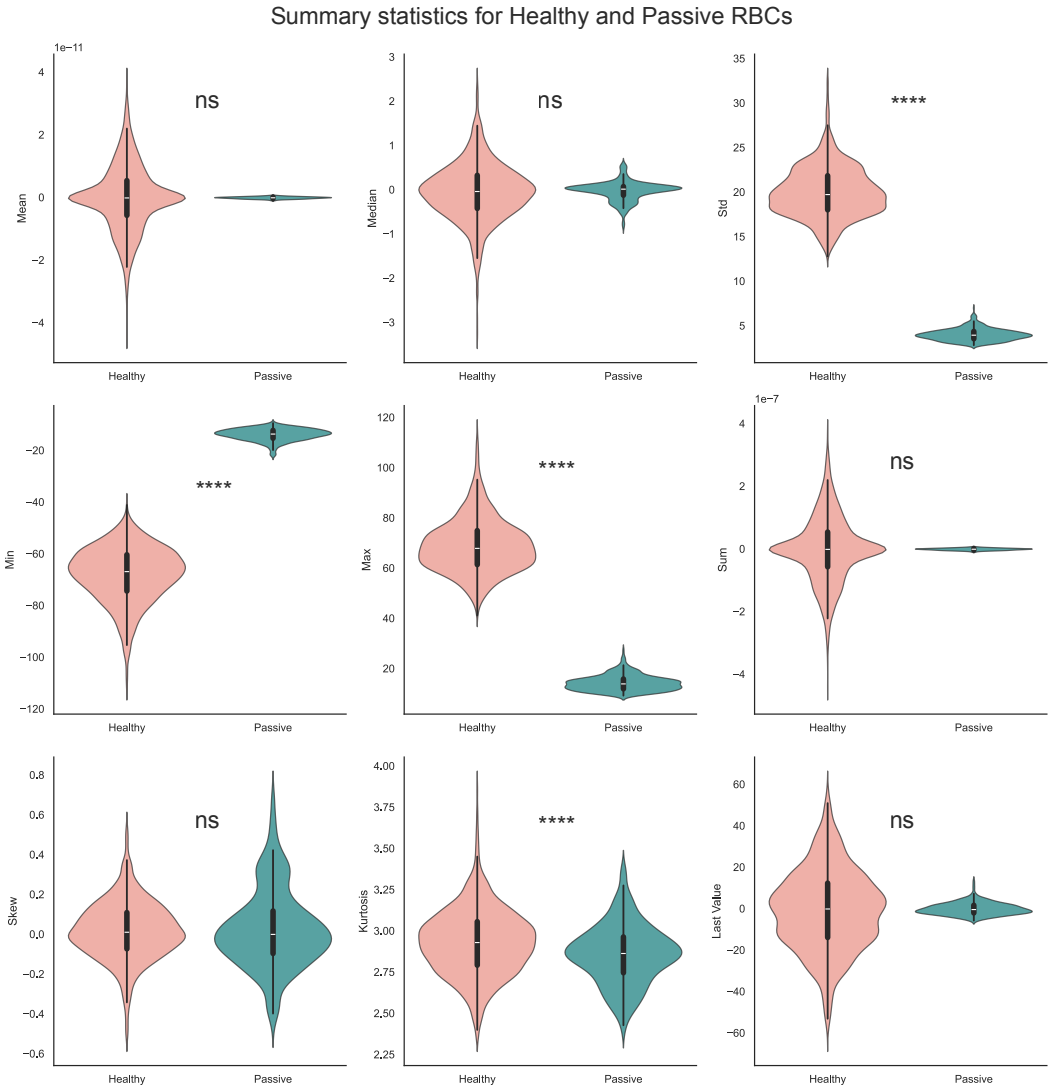

Figure 10: Distribution of summary statistics for normalised red-blood cell data: We consider 10 summary statistics as features to input into the neural network. Here we show that some statistics, including standard deviation, Min, Max and kurtosis, are significantly different between healthy and passive cells, making for easy classification of the recordings. Significance is denotes as (ns) if $p > 0.05$, (*) if $p < 0.05$, (**) if $p < 0.01$, (***) if $p < 0.001$ and (****) if $p < 0.0001$.

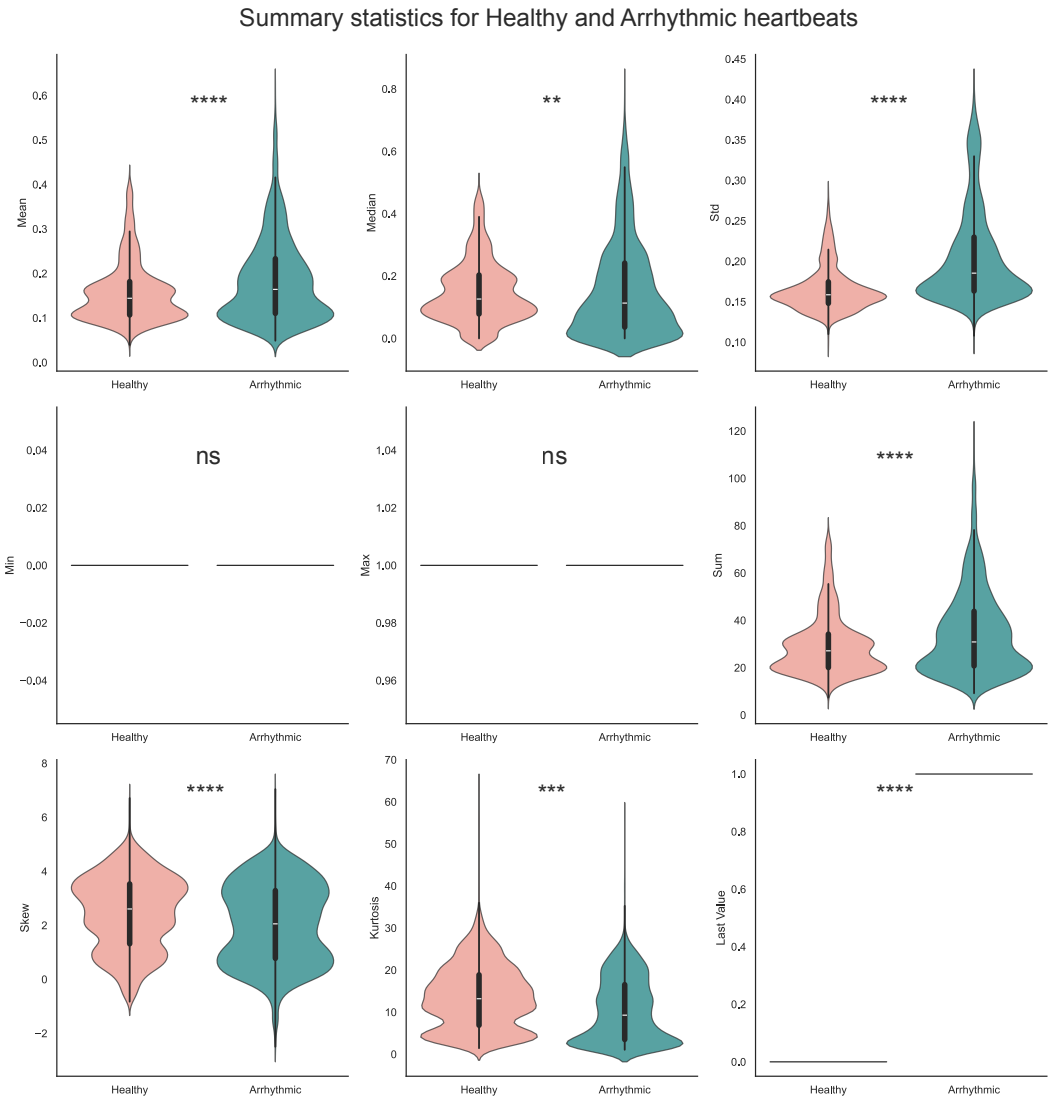

Figure 11: Distribution of summary statistics for normalised heartbeat data data: We consider 10 summary statistics as features to input into the neural network. Here we show that some statistics, including mean, median, standard deviation, sum, skew, kurtosis, and Last Value are significantly different between healthy and arrhythmic heartbeats, making for easy classification of the recordings. Significance is denotes as (ns) if $p > 0.05$, (*) if $p < 0.05$, (**) if $p < 0.01$, (***) if $p < 0.001$ and (****) if $p < 0.0001$.

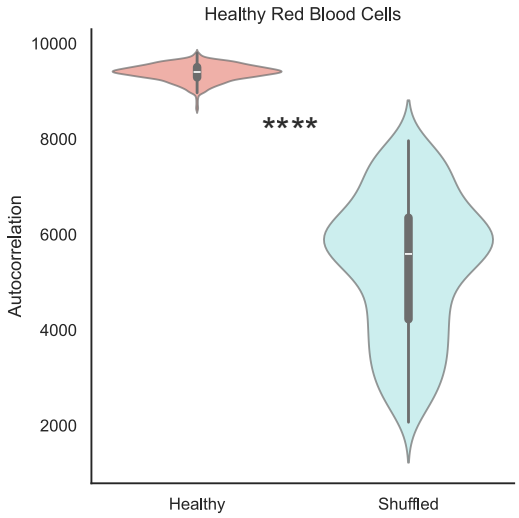

Figure 12: Distribution of auto-correlation for healthy RBC data and shuffled surrogate data: One can distinguish between true dynamics and shuffled surrogate data using temporally-dependent summary statistics. Here we show that the auto-correlation of data from the RBC Healthy 1 is significantly higher than its shuffled surrogate. Significance is denotes as (ns) if $p > 0.05$, (*) if $p < 0.05$, (**) if $p < 0.01$, (***) if $p < 0.001$ and (****) if $p < 0.0001$.

[52],

$$\boldsymbol{S} = \lim_{t\to\infty} [e^{-\boldsymbol{B}t}\boldsymbol{S}(0)e^{-\boldsymbol{B}^\top t} + 2\int_0^t e^{-\boldsymbol{B}(t-s)}\boldsymbol{D}e^{-\boldsymbol{B}^\top(t-s)}\, ds], \tag{58}$$

$$= 2\int_0^\infty e^{-\boldsymbol{B}t}\boldsymbol{D}e^{-\boldsymbol{B}^\top t}\, dt. \tag{59}$$

It can also be shown that $\boldsymbol{S}$ satisfies the following Sylvester equation,

$$\boldsymbol{B}\boldsymbol{S} + \boldsymbol{S}\boldsymbol{B}^\top = 2\boldsymbol{D}. \tag{60}$$

Next, we define the Onsager matrix, $\boldsymbol{L}$, of kinetic coefficients,

$$\boldsymbol{L} = \boldsymbol{B}\boldsymbol{S} = \boldsymbol{D} + \boldsymbol{Q}, \tag{61}$$

$$\boldsymbol{L}^\top = \boldsymbol{S}\boldsymbol{B}^\top = \boldsymbol{D} - \boldsymbol{Q}, \tag{62}$$

which parameterising the asymmetries through the anti-symmetric matrix $\boldsymbol{Q}$ which captures the rotational component of $\boldsymbol{B}$ [52]. Therefore the HHD of the OU process is given by,

$$\boldsymbol{B}\mathbf{x} = (\boldsymbol{B}_{\text{irrev}} + \boldsymbol{B}_{\text{rev}})\mathbf{x}, \tag{63}$$

$$\boldsymbol{B}_{\text{irrev}} = \boldsymbol{Q}\boldsymbol{S}^{-1}, \tag{64}$$

$$\boldsymbol{B}_{\text{rev}} = \boldsymbol{D}\boldsymbol{S}^{-1}, \tag{65}$$

where the matrix $\boldsymbol{B}_{\text{rev}}$ is symmetric. Additionally, the stationary probability fluxes are given by, $\boldsymbol{J} = -\boldsymbol{Q}\boldsymbol{S}^{-1}\mathbf{x}$ [52], used to display the arrows in panel B of Fig. 1.