Data-driven geometric phase in biological locomotion

Pyae Hein Htet and Kenta Ishimoto*
Department of Mathematics, Kyoto University, Kyoto 606-8502, Japan

(Received 23 June 2026; accepted 2 October 2026; published 30 October 2026)

Geometric phase quantifies net locomotion in dissipative media via gauge theory, but linking this theoretical quantity to noisy, sparse, and weakly periodic biological shape data is challenging. We develop a theory-guided, data-driven Koopman autoencoder to recover the limit cycle embedded in imperfect cyclic data and extract shape gaits and geometric phase from sperm and nematode data. We introduce a geometric phase sensitivity function that quantifies responses to shape perturbations and reveals mechanical information using only gauge-theoretic structure, without assuming mechanical laws.

DOI: 10.1103/qlj7-1kvn

I. INTRODUCTION

Geometric phase, which encodes the transformation history of a physical system, has become an influential concept across a broad range of modern physics [1], extending well beyond its original quantum-mechanical context, with examples in acoustics [2,3], robotics [4–6], geophysics [7], and nonequilibrium thermodynamics [8,9]. The hydrodynamics of swimming microorganisms, formulated in terms of gauge theory, has also offered a playground for geometric phase, which in this context corresponds to displacement and rotation over one period of deformation [10–13]. This mathematical structure arises from rapid viscous dissipation in low-Reynolds-number flow, where the swimming dynamics are solely determined by the swimmer's shape gait. More recently, experimental studies [14–17] have found that the gauge-theoretic structure ubiquitously holds in the class of kinematic locomotion, which encompasses a wider range of biological locomotion in dissipative media, including snakes crawling on sand, insects walking on the ground, and worms burrowing in mud.

With models of body-environment interaction such as the Stokes equation of low-Reynolds-number flow and empirical resistive force theory (RFT), one may theoretically compute locomotion dynamics, and thus geometric phase, from a period of shape gait [18]. Real biological data, however, contain internal and environmental fluctuations and often suffer from limited quality with low spatial and temporal resolutions. Linear dimensionality reduction via principal component analysis (PCA) has successfully represented shape gait data in (typically two-dimensional) PCA space across systems ranging from cells to animals and artificial robots [17,19–22], while finer details of shape information are often captured only by higher PCA modes [23–26].

In extracting periodic gait as a limit cycle embedded in noisy data, linear methods such as PCA only provide protophase and do not define a proper phase, termed the isochron phase or asymptotic phase [27]. Hence, nonlinear methods are essential for theoretical consistency. In the last decade, the method of phase reduction has been reformulated via Koopman operator theory, which reframes a nonlinear system as an infinite-dimensional linear system and is widely used in data-driven modeling of dynamics [28–30]. The isochron phase of a limit cycle, in this context, can be obtained via the principal Koopman eigenfunction [31–33]. Although a useful and fundamental theoretical construct, Koopman eigenfunctions are notoriously difficult to determine in practice, spurring the development of machine learning approaches. Recently, a deep-learning architecture has been proposed to construct the isochron phase [34], and it has also succeeded in extracting limit cycles in fluid and reaction-diffusion systems with large dimensions [35,36]. However, machine learning models usually require large, clean datasets, posing a challenge for data-driven approaches to small-sized biological data.

In this paper, we develop a theoretically grounded, data-driven approach to the geometric phase of biological locomotion, using real data from slender locomotors, including sperm cells and crawling nematodes, as prototypical small-sized, low-quality shape data only spanning several beat cycles and exhibiting fluctuations in shape and period. After learning the embedded periodic gait from noisy data, we extract the geometric phase via the learned isochron phase, imposing only the gauge-theoretic structure, without assuming any mechanical relations such as fluid equations or a resistive-force-type empirical theory. Further, as a natural extension of phase reduction theory for geometric phase, we introduce the *geometric phase sensitivity function (PSF)*, which measures the impact of shape perturbations on locomotion. We show that this new physical quantity characterizes the underlying mechanics directly from the data and is useful in understanding the locomotor's steering and maneuverability.

*Contact author: kenta.ishimoto@math.kyoto-u.ac.jp

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

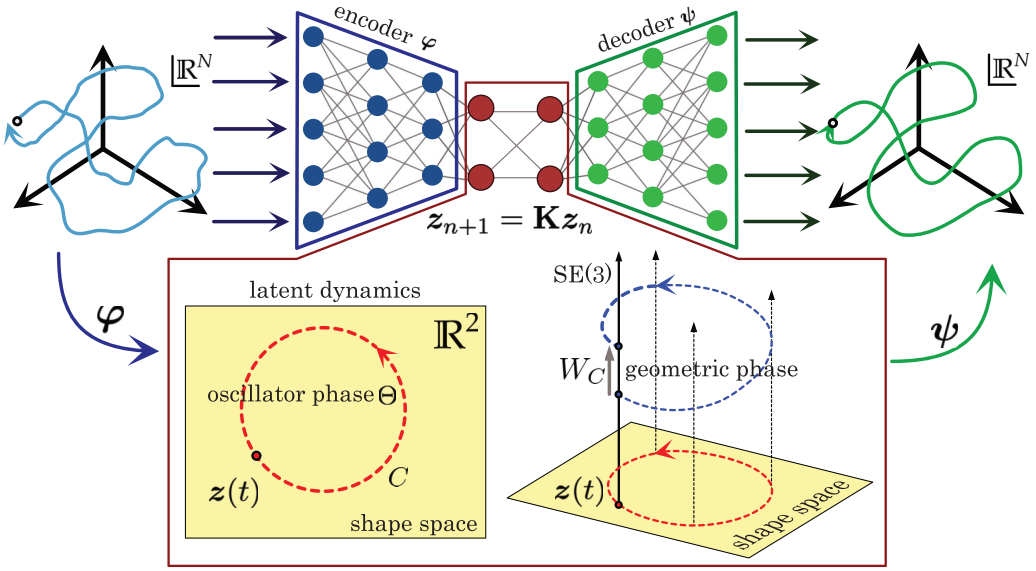


FIG. 1. Schematic of theoretical method and numerical architecture. We extract a stable limit cycle embedded in large dimensional shape data, $\boldsymbol{q} \in \mathbb{R}^N$, via the principle Koopman eigenfunction $z = \boldsymbol{\varphi}(\boldsymbol{q}) \in \mathbb{R}^2$, which maps the shape data into two-dimensional latent space, where the time evolution follows linear dynamics and oscillator phase Θ is defined. The geometric phase W_C is defined as the spatial and orientational shift after one period of deformation. The principal Koopman eigenfunction is numerically constructed by a deep autoencoder, where the map $\boldsymbol{\varphi}$ is the encoder and the decoder $\boldsymbol{\psi}(z) \in \mathbb{R}^N$ provides the denoised orbit.

The paper is organized as follows. We introduce the background theory and methodologies in Sec. II. After presenting the results on synthetic waveform data in Sec. III, we analyze experimental waveform data in Sec. IV, where we discuss the geometric phase sensitivity function. Section V concludes the study.

II. THEORY AND METHODS

A. Geometric phase in biolocomotion

The motion of a self-propelled locomotor is described by its position $\boldsymbol{x}(t) \in \mathbb{R}^3$ and orientation $\mathbf{R}(t) \in \mathrm{SO}(3)$, represented by $g = (\mathbf{R}, \boldsymbol{x}) \in \mathrm{SE}(3)$. The shape gait of the locomotor is parametrized by $\boldsymbol{q}(t) \in \mathbb{R}^N$, where N denotes the number of shape parameters. In kinematic locomotion, as in microswimmer hydrodynamics, the motion of a locomotor is proportional to the deformation velocity. Hence, the linear and angular velocities in the body-fixed frame are provided by

$$g^{-1}\dot{g} = A(\boldsymbol{q})\dot{\boldsymbol{q}}. \tag{1}$$

Here, the dot symbol denotes the time derivative and $A(\boldsymbol{q})$ is determined only by the instantaneous shape and is mathematically equivalent to a gauge field [10], or so-called mechanical connection, which encodes the physics of body-environment interactions.

The shape gait of a locomotor typically consists of the repetition of a time-periodic shape deformation, corresponding to a stable limit cycle C as $\boldsymbol{q}(t) \in \mathbb{R}^N$. The displacement and rotation after one deformation cycle correspond to the *geometric phase* of the loop C, which is obtained by integrating Eq. (1) over the loop as

$$W_C = \overline{\mathrm{P}} \exp\left[\oint_C A(\boldsymbol{q})\, d\boldsymbol{q}\right]. \tag{2}$$

Here, the symbol $\overline{\mathrm{P}}$ denotes the anti-path-ordering operator. The geometric phase $W_C \in \mathrm{SE}(3)$ emerges from an adiabatic change of shape in dissipative systems, illustrated in Fig. 1 as a mismatch of the blue trajectory in the vertical direction after one cycle of the deformation in the shape space. The geometric phase W_C contains translation and rotation and is not a scalar value but represented by a 4×4 matrix [10].

B. Phase reduction and Koopman autoencoder

The essential shape dynamics is therefore represented by a circular orbit in a low-dimensional manifold, and Koopman operator theory guarantees [31,37] that for a stable limit cycle the principal Koopman eigenfunction constructs a change of coordinates, $z = \boldsymbol{\varphi}(\boldsymbol{q}) \in \mathbb{R}^2$, that follows $\dot{z} = i\omega z$ with $z = z_1 + iz_2 \in \mathbb{C}$, where ω is the frequency of the periodic gait. The phase of the limit cycle oscillator, $\Theta(t)$, is introduced as $\dot{\Theta} = \omega$ on the loop C, and we call this the *oscillator phase* to distinguish from the geometric phase. Oscillator phase can be extended to the entire attracting region of the limit cycle via the isochron function, defined as $\Theta(\boldsymbol{z}) = \arg(z_1 + iz_2)$.

Biological locomotion is often contaminated by both internal activity noise and external noise from the environment and the observation process. To approximate the principal Koopman eigenfunction $\boldsymbol{\varphi}$ that extracts the limit cycle orbit embedded in noisy biological data, we set up a deep autoencoder neural network that enforces linear dynamics in latent space [38–41], known as a Koopman autoencoder (KAE). As schematically illustrated in Fig. 1, the encoder $z = \boldsymbol{\varphi}(\boldsymbol{q})$ is designed to approximate the Koopman eigenfunction, which follows linear dynamics in latent space. The decoder $\boldsymbol{\psi}(z)$ provides a two-dimensional manifold embedded in the shape space, functioning as a denoiser for the stochastic time series.

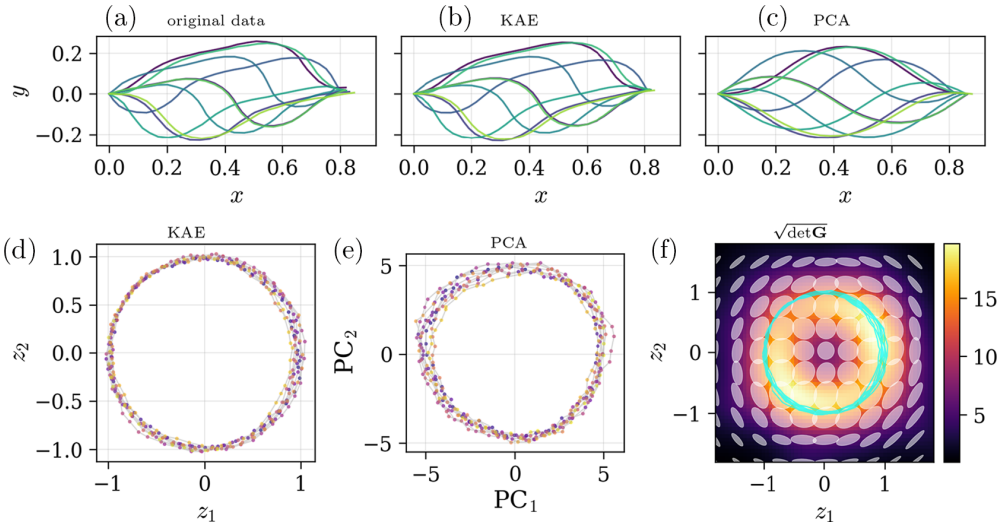


FIG. 2. (a)–(c) Snapshots of the synthesized waveform, reconstruction by KAE and two-dimensional PCA. (d), (e) Data points projected onto the two-dimensional latent space and two-dimensional PCA space. (f) Visualization of pull-back metric by $\sqrt{\det \mathbf{G}}$ and metric ellipse. Parameters of the synthesized data are set as $A_1 = 0.8$, $A_2 = 0.4$, $\delta = \pi/2$, $c = 1.0$, $b = 0.5$, $D_r = 0.1$, D_ϕ=0.1, $D_{\rm obs} = 0.05$, and $L = 1$.

Given the shape data time series $\boldsymbol{q}_n = \boldsymbol{q}(t_n)$, we enforce linear dynamics $\boldsymbol{z}_{n+1} = \mathbf{K}\boldsymbol{z}_n$ for the latent variable $\boldsymbol{z}_n = \boldsymbol{\varphi}(\boldsymbol{q}_n)$ and train our deep neural network to minimize the total cost function $L_{\rm tot} = L_{\rm rec} + L_{\rm pred} + L_{\rm lin} + L_{\rm rad}$, where

$$L_{\rm rec} = ||\boldsymbol{q}(t_n) - \boldsymbol{\psi}(\boldsymbol{\varphi}(\boldsymbol{q}(t_n)))||^2, \tag{3}$$

$$L_{\rm pred} = ||\boldsymbol{q}(t_{n+1}) - \boldsymbol{\psi}(\mathbf{K}\boldsymbol{\varphi}(\boldsymbol{q}(t_n)))||^2, \tag{4}$$

$$L_{\rm lin} = \lambda_{\rm lin}||\boldsymbol{q}(t_{n+1}) - \mathbf{K}\boldsymbol{\varphi}(\boldsymbol{q}(t_n))||^2, \tag{5}$$

$$L_{\rm rad} = \lambda_{\rm rad}||\boldsymbol{\varphi}(\boldsymbol{q}(t_n)) - 1||^2 \tag{6}$$

are reconstruction loss, prediction loss, linearity constraint, and radius constraint, respectively. $\lambda_{\rm lin}$ and $\lambda_{\rm rad}$ are regularization coefficients, and $||\bullet||$ denotes the ℓ_2 norm. We set

$$\mathbf{K} = \begin{pmatrix} \gamma\Delta t & -\omega\Delta t \\ \omega\Delta t & \gamma\Delta t \end{pmatrix}, \tag{7}$$

where Δt is the time discretization and $\gamma > 0$ accommodates the Floquet contraction of stochastic trajectories onto the limit cycle [42]. Further details of the network architecture and learning implementations are provided in Appendix A.

III. LEARNING SYNTHETIC WAVEFORM DATA

To validate our methodology, we start with synthetic waveform data of a slender object confined in a two-dimensional space. We let $s \in [0, L]$ be the arclength of the slender locomotor of total length L and take the local tangent angle $q(s,t)$ as our shape variable. Following observations on sperm cells [43], we assume that the waveform $q(s,t)$ is generated by a noisy limit cycle represented as a normal form of the supercritical Hopf bifurcation, written in polar coordinates (r, ϕ) as

$$dr = \dot{r}_{\rm det}dt + rD_r dW_r, \quad d\phi = \dot{\phi}_{\rm det}dt + D_\phi dW_\phi, \tag{8}$$

with deterministic velocity parts,

$$\dot{r}_{\rm det} = cr(1 - r^2), \quad \dot{\phi}_{\rm det} = \omega + br^2. \tag{9}$$

Here, W_r and W_ϕ denote the Wiener process, and D_r and D_θ are the noise intensities. We then synthesize the tangent angle from this oscillator, with $\zeta = \phi - ks$, as

$$q(s,t) = rA_1\sin(\zeta) + r^2A_2\sin(2\zeta + \delta) + \xi(t), \tag{10}$$

where $k = 2\pi/L$ is the wave number and $\xi(t)$ is Gaussian noise from observation with noise strength $D_{\rm obs}$. It is readily found by Fourier decomposition that this waveform contains four PCA modes, $\sin(ks)$, $\cos(ks)$, $\sin(2ks)$, and $\cos(2ks)$.

In Fig. 2(a), we show snapshots of the synthesized waveform $q(s,t)$, which is spatially discretized by 30 segments ($N = 29$) and contains approximately 6 beat cycles with 40 time points per cycle. We also illustrate the waveforms reconstructed by the KAE and two-dimensional PCA [Figs. 2(b) and 2(c)], demonstrating that KAE successfully reproduces the finer details of the waveforms using two latent variables. Further details on validation of the KAE computations are discussed in Appendix B.

As shown in Fig. 2(d), the trajectory in latent space accumulates near a unit circle, whereas the PCA space trajectory is noticeably more scattered. Using the Jacobian of the decoder $\boldsymbol{\psi}(\boldsymbol{z})$, $\mathbf{J} = \partial\boldsymbol{\psi}/\partial\boldsymbol{z}$, we can introduce a 2×2 pull-back Riemannian metric $\mathbf{G}(\boldsymbol{z}) = \mathbf{J}^{\rm T}\mathbf{J}$ on the latent space [44–46], characterizing the nonlinear transformation in extracting the limit cycle. In Fig. 2(f), we show in color contour $\sqrt{\det \mathbf{G}}$, which represents the local expansion. The higher values, shown as brighter colors, reflect nonlinear folding of the orbit in the shape space to fit onto the unit circle in latent space. The pull-back metric $\mathbf{G}(\boldsymbol{z})$ is also visualized by ellipses with semiaxes proportional to the two eigenvalues and oriented along the eigenvectors, showing nonuniform embedding in shape space.

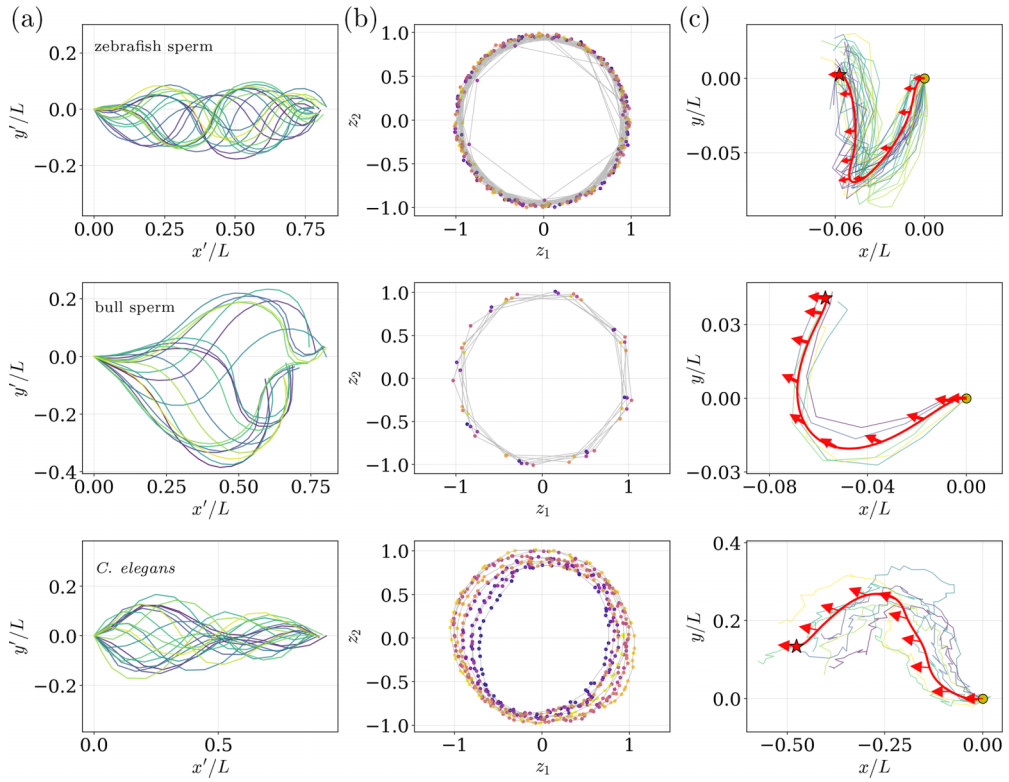


FIG. 3. (a) Shape plots in body-fixed frame, (b) scatter plots in Koopman latent space, and (c) superimposed trajectories with averaged trajectory obtained by the current method, for zebrafish sperm (top), bull sperm (middle), and *C. elegans* (bottom).

IV. LEARNING EXPERIMENTAL WAVEFORM DATA

A. Extracting limit cycle in experimental biological data

Having demonstrated our method using synthesized data with controlled noise, we now apply our method to extract geometric phase from biological trajectory data, obtained as a movie file recording the organism's shape and motion. We do not assume any mechanical equations such as the Stokes equations or resistive force theory, imposing only the gauge field structure in Eq. (1).

To illustrate the applicability of our data-driven method, we use sperm flagellar waveform data from zebrafish (*Danio rerio*) [47] and bull (*Bos taurus*) [47,48], and body-shape data from the nematode *Caenorhabditis elegans* crawling on agar gel from the Movie in the Supplemental Material of Ref. [49] (see Appendix C). Snapshots of the waveforms are shown in Fig. 3(a). These data are chosen as representatives of external fertilizers, internal fertilizers, and nonswimmer organisms, respectively. Although zebrafish and bull sperm dynamics are well described by the Stokes equation, we do not use this knowledge *a priori*. Also, the crawling nematode is known to follow resistive force theory with a large anisotropic drag ratio [15], which, however, requires a data-fitting process to determine.

As in the synthesized data, we analyze the local tangent angle along the centerline and our shape dimension N is the number of discretized local angles. The zebrafish sperm flagellar centerline is discretized into $N = 29$ shape dimensions and contains approximately 30 beat periods in $N_t = 300$ time samples and is characterized by sinusoidal waves with spatially uniform amplitude. The bull sperm data also have $N = 25$ shape dimensions, but are sparser in time, containing approximately 6 beat periods in $N_t = 50$ time points, and have larger waveform amplitudes in the distal region. The nematode data have $N = 13$ shape dimensions and 8 beat cycles in $N_t = 515$ time samples, and exhibit a distinct crawling motion with larger fluctuations compared to the flagellar data. For all three qualitatively different datasets, the KAE successfully extracts the limit cycles embedded in the waveforms, as illustrated by the unit-circular latent space dynamics shown in Fig. 3(b).

B. Extracting geometric phase from experimental biological data

Now we extract the averaged trajectory as geometric phase on a stable limit cycle only from the shape and trajectory data, without assuming any mechanical laws. The limit cycle is extracted as the unit circle in latent space via KAE. Hence, at each time step t_n, the locomotor's trajectory data are labeled by the oscillator phase $\Theta(t_n)$. On the limit cycle, the gauge field $A(\Theta)$ is expanded in a Fourier series as

$$A(\Theta) = \sum_{k=0}^{\infty} \tilde{A}_k e^{ik\Theta}, \tag{11}$$

which corresponds to a Koopman mode decomposition [50] of the gauge field as an observable. We truncate the series to K modes, with $K = 2, 3, 4$ for the zebrafish sperm, bull sperm, and nematode waveform, respectively, chosen to minimize the Bayesian information criterion [51].

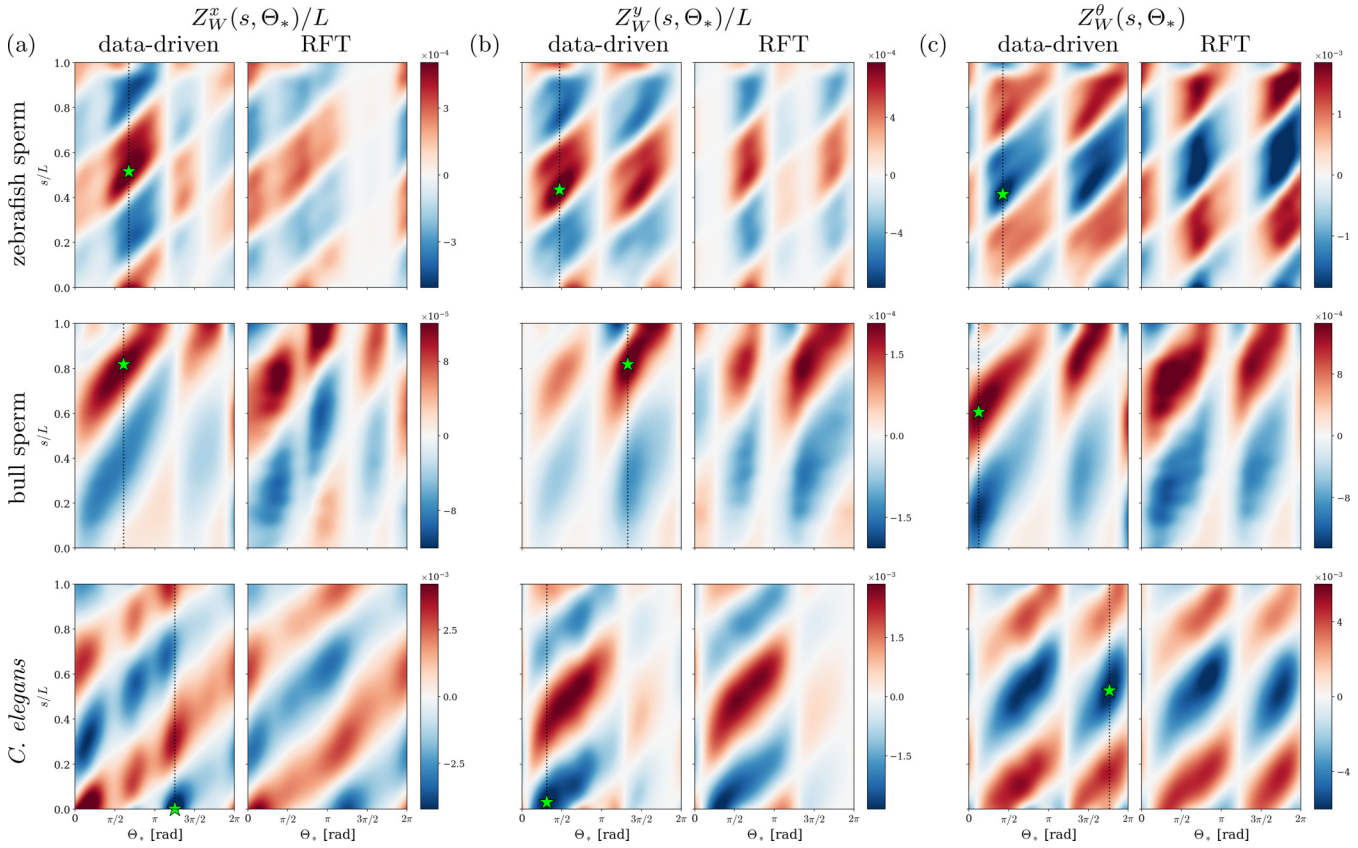


FIG. 4. Geometric PSF (a) $Z^x_W(s,\Theta_*)/L$, (b) $Z^y_W(s,\Theta_*)/L$, and (c) $Z^\theta_W(s,\Theta_*)$ for zebrafish sperm, bull sperm, and nematode data, computed by the data-driven method (left) and RFT calculations (right). The maximal absolute value in the data-driven result is indicated by a star symbol in each panel.

Biological trajectory data $g(t) \in \mathrm{SE}(3)$ are shown as shaded curves in Fig. 3(c), translated and rotated so that the trajectory starts at the origin oriented in the negative x direction at the beginning of each cycle, i.e., whenever the oscillator phase $\Theta(t) \equiv 0 \pmod{2\pi}$. The zebrafish and bull sperm data are sparse in time (8–10 samples per period), while the nematode data vigorously fluctuates. Despite these sparse and noisy samples, we successfully extract the locomotor's trajectories, which are overlaid as smooth red curves. Further methodological validations are discussed in Appendix B.

C. Geometric phase sensitivity function

The phase sensitivity function Z_1 of an oscillator, which we hereafter call the *oscillator PSF*, is defined by

$$\delta\Theta_* = \int_0^L Z_1(s,\Theta_*)\delta q(s)\,ds. \tag{12}$$

$Z_1(s,\Theta_*)$ characterizes the shift in oscillator phase upon a perturbation at the arclength s and the phase Θ_*. This phase shift then perturbs the locomotor's position and orientation via $\delta g(\Theta_*) = A(\Theta_*)\delta\Theta_*$, yielding the shift in the geometric phase W_C. Writing the motion between $\Theta = 0$ and $\Theta = \Theta_*$ as

$$g_* = g^{-1}(0)g(\Theta_*), \tag{13}$$

the shift in the geometric phase is written as

$$\delta W_C = g_* \exp\left[A(\Theta_*)\delta\Theta_*\right]g_*^{-1}W_C - W_C, \tag{14}$$

which becomes, at leading order in $\delta\Theta_*$,

$$\delta W_C = Z_2(\Theta_0)W_C\delta\Theta_* + O(\delta\Theta_*^2), \tag{15}$$

with $Z_2(\Theta_*) = g_*A(\Theta_*)g_*^{-1}$. This naturally motivates the definition of a geometric PSF as

$$Z_W := Z_2(\Theta_*)Z_1(s,\Theta_*), \tag{16}$$

which characterizes the geometric phase shift upon a perturbation of the shape variable via the relation

$$\delta W_C = \int_0^L Z_W(s,\Theta_*)\delta q(s)\,ds + O(\delta q^2). \tag{17}$$

The geometric PSF for planar locomotion possesses three components $(Z^x_W, Z^y_W, Z^\theta_W)$, which represent the progressive and lateral displacements and the rotational angle upon a perturbation of the local tangent angle. Biologically, a large geometric PSF is beneficial for locomotion and maneuvering, as the locomotor gains a large change of motion from a small shape change.

With the data-driven gauge field in Eq. (11) [Fig. 3(c)], we numerically compute the geometric PSF and show in Fig. 4 (data-driven) the kymograph of each component, $Z^\alpha_W(s,\Theta_*)$ $(\alpha = x, y, \theta)$, for the three different biological datasets. Stripe patterns in the figure reflect the traveling waveform. Vertical white stripes in the panels correspond to the phase with $Z_2(\Theta_*) = 0$. Further discussions on the robustness of the learning process and data-driven calculations of geometric PSF are provided in Appendix D.

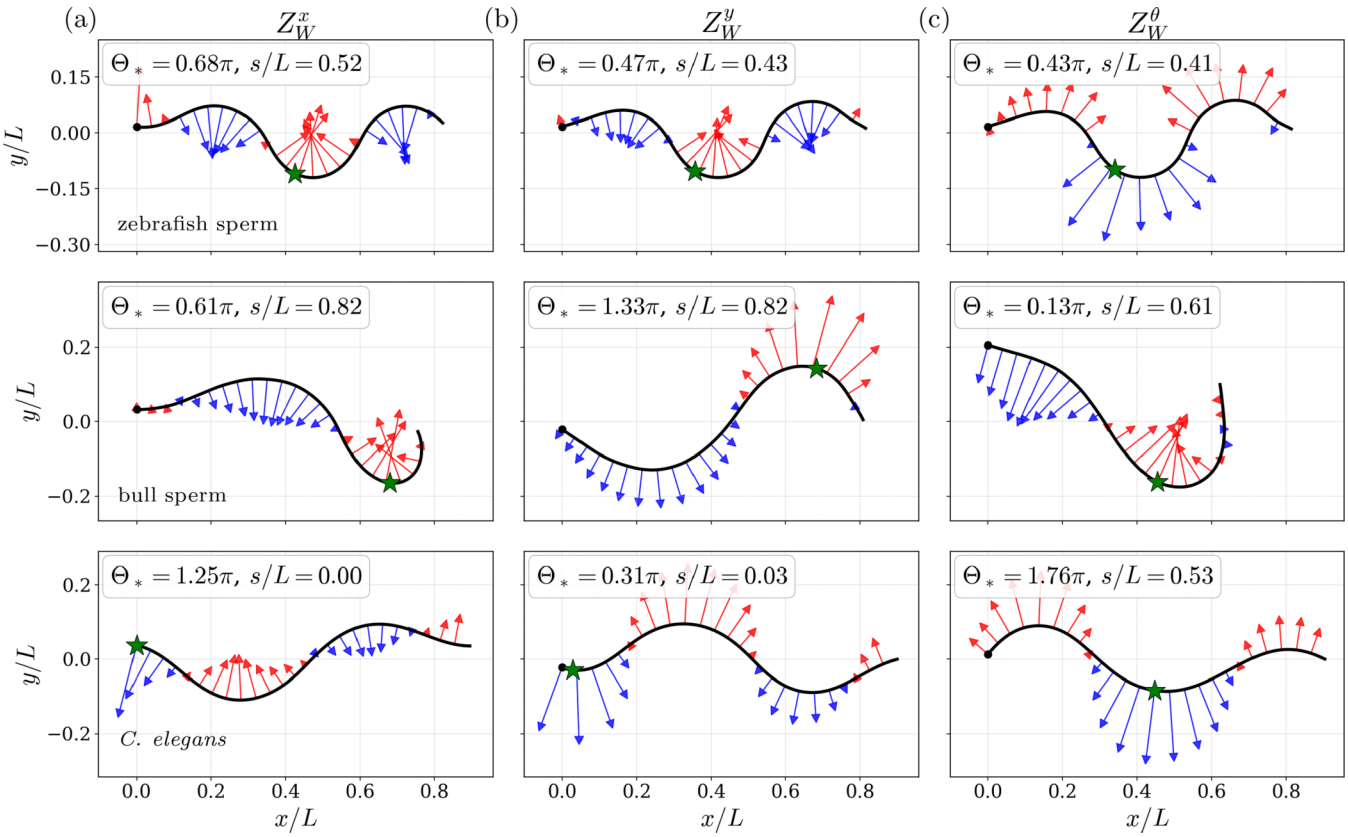


FIG. 5. Sensitivity arrows showing the local normal vector with length proportional to the geometric PSF for the zebrafish sperm (top), bull sperm (middle), and *C.elegans* (bottom) shape at the phase of maximum (a) $|Z_W^x|$, (b) $|Z_W^y|$, and (c) $|Z_W^\theta|$. Green stars mark the position of maximum $|Z_w^\alpha|$ ($\alpha = x, y, \theta$) and positive/negative values of Z_W^α are shown in red/blue.

In each component, the maximum of $|Z_W^\alpha(s, \Theta_*)|$ is marked by the star symbol and the phase Θ_* of maximum $Z_W^\alpha(s, \Theta_*)$ is indicated by dotted line. To physically interpret the obtained geometric PSF, we then show in Figs. 5(a)–5(c) the shape at the phase of maximum $|Z_W^\alpha|$ for each component. The arrow indicates the normal direction to the shape, and the magnitude reflects the size of Z_W^α.

In the zebrafish sperm, the geometric PSF is approximately uniform along the flagellum, with a maximum at the middle of the flagellum. In contrast, the bull sperm has its maximum geometric PSF in the distal region, because the flagellar amplitude is large here. This larger distal undulation of flagella is a characteristic of mammalian sperm and it has been reported that the distal region contributes most to propulsion [52]. In the nematode case, the maxima of the geometric PSF are located at the head end for both the progressive and lateral translation. This reflects the much smaller tangential drag in crawling on agar relative to normal drag [15], resulting in body motions that tend to follow the head trajectory [53]. These physical interpretations are consistent with the geometric PSF obtained purely from the data.

As seen in Fig. 5, the maximum of $|Z_W^\alpha|$ corresponds to a large curvature region in the waveform. As the phase Θ_* increases, the regions with larger values of the geometric PSF travel along the locomotor's body, which can also be seen as the traveling wave structure in the kymograph of Fig. 4.

D. Comparison with RFT reconstruction

To further validate our data-driven method, we construct the geometric PSF by assuming mechanical relations of resistive force theory with an anisotropic ratio γ. Here, we treat γ as a free parameter fitted to the biological raw data. After nondimensionalizing the raw data and normalizing each component of the swimming velocity with its variance, we numerically compute the best-fit anisotropic ratio γ by the least squares method.

In principle, to reconstruct the instantaneous velocity from resistive force theory, we require additional assumptions for the cell morphology. Here, we assumed that the zebrafish sperm possesses a spherical head of radius $b/L = 0.036$ and obtained the best-fit anisotropic drag ratio as $\gamma \approx 1.62$. For the bull sperm case, we followed Ref. [48] and assumed a spheroidal head of aspect ratio 2, with a major axis of $b/L = 0.083$, and we obtained $\gamma \approx 1.73$. In the *C. elegans* data, we assumed a uniform slender structure and the best-fit value of γ is calculated as $\gamma \approx 62.5$. The estimated values of γ in the three data are consistent with previous studies [15,48].

With the estimated anisotropic drag ratio in each dataset, we compute the gauge field $A(\Theta)$ via resistive force theory, instead of using Eq. (11), to obtain the geometric PSF, which is plotted in Fig. 4 (RFT).

Figure 4 provides a side-by-side comparison of the geometric PSF between the fully data-driven method and reconstruction by the empirical resistive force theory, showing

good qualitative agreement, although the peak magnitudes and positions do not always coincide. We anticipate that the mismatch reflects the limitations of the RFT reconstruction, because resistive force theory only provides the lowest-order asymptotics and neglects nonlocal hydrodynamics, as well as requiring additional morphological assumptions.

In contrast, our data-driven approach does not rely on morphological details or empirical mechanical laws to compute the geometric PSF. The comparison with the RFT reconstruction provides an independent verification of our data-driven methods.

V. CONCLUDING REMARKS

We have presented a data-driven approach to extract geometric phase in biological locomotion, enforcing only gauge-theoretic structure on the dataset without assuming any mechanical laws. Crucially, our approach combines the broad applicability of data-driven methods with a solid foundation in Koopman operator theory. Further, we introduced a geometric PSF that encodes the underlying mechanics directly from the sparse and noisy observations characteristic of biological measurements.

While only two-dimensional motion was presented in the current study, our data-driven method readily extends to three-dimensional deformation and locomotion [54]. Since the Koopman phase reduction is applicable to a stable limit torus [55], our method is also potentially extendable to more complex dynamics such as the multiple flagella of *Chlamydomonas* [56] and multilegged locomotion [57], although a higher-dimensional latent manifold may require a larger amount of data for reconstruction.

Biological locomotion often contains multiple gaits and switches from one to another as seen in forward and backward locomotion of *C. elegans* [53]. Since the current method relies on the existence of a stable limit cycle, it is not applicable to such transient behaviors, although we can extract periodic dynamics from a short time window of repetitive motion.

In the current study, we analyzed a representative dataset in each species to demonstrate the applicability of our method. Applied across more datasets, our method could be helpful in analyzing the variations within individuals, samples, and species [22].

Whereas data-driven approaches to learning gauge fields have been actively studied in robotic locomotion [58–60], where clean and large datasets are easily available, our methods successfully operate within the severe constraints of biological data. More generally, our study has broad implications beyond biological locomotion, demonstrating how geometric structure can be exploited to infer physical structure from limited and imperfect observations when detailed governing laws are inaccessible.

ACKNOWLEDGMENTS

P.H.H. and K.I. acknowledge the Japan Society for the Promotion of Science (JSPS) KAKENHI for JSPS International Research Fellow (Grant No. 25KF0222). K.I. acknowledges JSPS KAKENHI (Grants No. 21H05309 and No. 24K21517) and the Japan Science and Technology Agency (JST), FOREST (Grant No. JPMJFR212N), and CREST (Grant No. JPMJCR25Q1).

DATA AVAILABILITY

The code and data needed to reproduce all figures are publicly available [61].

APPENDIX A: KAE ARCHITECTURE AND TRAINING DETAILS

In this Appendix, we provide further details of our KAE architecture and learning procedures.

We implemented encoder and decoder functions using a deep neural network with two hidden layers. The number of nodes is set to $N_1 = 16$ and $N_2 = 32$ for learning the synthesized data and the bull sperm data. For the longer time series in the zebrafish sperm and nematode data, we set $(N_1, N_2) = (16, 64)$. The overall number of nodes of the neural network is set, from encoder to decoder, as $N \to N_2 \to N_1 \to 2 \to 2 \to N_1 \to N_2 \to N$, with internal layers with two nodes for the latent dynamics (see also Fig. 1). We used the tanh function as an activation function and implemented these deep neural computations via PyTorch.

For training, we used the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-5}, and the regularization parameters in the loss function were set to $\lambda_{\rm lin} = 5.5$ and $\lambda_{\rm rad} = 0.2$ in most cases, except for the zebrafish sperm data, for which we used $\lambda_{\rm rad} = 2.5$. The initial value of ω was guessed from the dataset and fixed for the first 40 epochs. In the current datasets, 350 total epochs were sufficient for learning convergence.

APPENDIX B: FURTHER ANALYSIS ON SYNTHESIZED WAVEFORMS

1. Stuart-Landau oscillator and oscillator PSF

The synthesized data are generated by a stochastic oscillator in Eq. (8), whose deterministic part is known as the Stuart-Landau oscillator. Its principal Koopman eigenfunction is analytically obtained as

$$\psi(r, \phi) = \frac{1}{r} \exp\left(i\Theta(r, \phi)\right), \tag{B1}$$

with eigenvalue $\mu = -b + i\omega_{\rm eff}$. Here, $\omega_{\rm eff} = \omega + b$ presents the phase velocity at the limit cycle $r = 1$, and the function

$$\Theta(r, \phi) = \phi + \frac{b}{c} \ln r \tag{B2}$$

is the isochron phase, which properly defines the phase of an oscillator even outside of the limit cycle [62]. Oscillator PSF $Z_1(s, \Theta)$ introduced in Eq. (12) is analytically calculated for the Stuart-Landau oscillator. On the limit cycle, in particular, the oscillator PSF from the deterministic part is written, after normalization, as

$$Z_1(s, \Theta) = \frac{\partial q}{\partial \Theta}(s, \Theta) \Big/ \int_0^L \left| \frac{\partial q}{\partial \Theta}(s, \Theta) \right|^2 ds. \tag{B3}$$

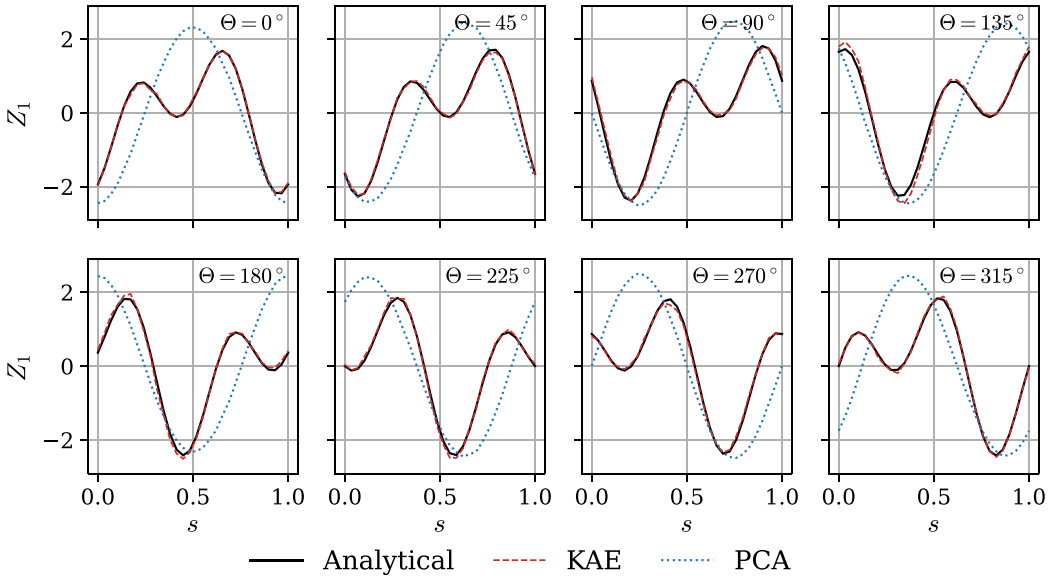


FIG. 6. Comparisons of calculated phase response function. The parameters are the same as in Fig. 2.

The oscillator PSF is also obtained from the learned decoder, with $\boldsymbol{q} = \boldsymbol{\psi}(z)$, as

$$Z_{\mathrm{KAE}} = \frac{\partial \boldsymbol{q}}{\partial \Theta} = \mathbf{J}\frac{dz}{d\Theta}, \tag{B4}$$

where $\mathbf{J}$ is the Jacobian of the decoder and numerically computed via automated differentiation. Comparisons between analytical representation and the obtained KAE computation, plotted in Fig. 6, show excellent agreement and validate the KAE computation. We also plot the oscillator PSF Z_{PCA} from two-dimensional PCA, which fails to capture the higher-frequency modes.

2. Extracting geometric phase from synthetic trajectory data

We then extract the geometric phase from fluctuating swimming trajectories. To synthesize the trajectory data, we apply resistive force theory to the shape gait generated by the stochastic dynamics of Eq. (8). Here, we assume that at each time t, the instantaneous swimming velocity is determined by the current shape and the deterministic shape change rate, which is calculated from Eq. (8) as

$$\dot{q}_{\mathrm{det}}(s,t) = \dot{r}_{\mathrm{det}}\left(\frac{\partial q}{\partial r}\right) + \dot{\phi}_{\mathrm{det}}\left(\frac{\partial q}{\partial \phi}\right). \tag{B5}$$

The stochastic increments dW_r and dW_ϕ contribute to the evolution of (r,ϕ) across time steps but do not enter the instantaneous shape velocity seen by the fluid, considering that the swimmer's internal forcing (e.g., molecular motors) produces deterministic shape changes. This yields a velocity signal $g^{-1}\dot{g} \in \mathfrak{se}(3)$ that is smooth at each instant but modulated by the fluctuating amplitude $r(t)$.

To validate the KAE-based extraction of geometric phase, we compared different methods and took averages using the analytic isochron phase [Eq. (B2)]. Denoting the linear and angular velocity components of the gauge field by the hat symbol as $\hat{A}(\phi) = (A_x, A_y, A_\theta)$, we calculate the L^2 relative error defined as

$$\epsilon = ||\hat{A} - \hat{A}_{\mathrm{ref}}||/||\hat{A}_{\mathrm{ref}}||, \tag{B6}$$

where A_{ref} is the gauge field estimated by averaging with the analytic expression of the isochron function [Eq. (B2)]. We varied the values of sampling Δt, the ratio A_2/A_1 of the second to the first harmonic, radial diffusion D_r, and nonuniform phase speed b. We tested 20 random sequences of the synthesized waveform and summarize our results in Fig. 7, with error bars representing standard deviation.

We confirm that our KAE-based extraction agrees well with the reference values, with relative errors of $\epsilon = O(10^{-2})$ for the parameter set used in Fig. 2. We also compare results from different methods. A protophase may be obtained as the angle in the two-dimensional PCA plane. Although averaging with this protophase performs similarly well to KAE-based averaging for this parameter set [see the curve labeled "PCA-2" in Fig. 7(a)], the error of PCA-based averaging increases substantially with the second harmonics ratio [Fig. 7(b)].

In the panel with different D_r [Fig. 7(c)], the KAE-based averaging shows a similar relative error to the PCA result in the upper end of the range of D_r considered. However, this has no theoretical basis in the PCA algorithm and is likely coincidental. Panel (d) examines the impact of the parameter b for the nonuniform phase velocity, showing only minor effects on the relative error.

APPENDIX C: DATA ACQUISITION

Sperm flagellum centerlines in the laboratory frame were obtained from processed centerline data provided in Ref. [47]. We used the single available bull sperm dataset, and file number 6, run number 3 from the zebrafish datasets. Nematode centerlines in the laboratory frame were extracted by thresholding each frame of Movie 3 in the Supplemental Material of Ref. [49], skeletonizing the worm body, and resampling each centerline to 14 equally spaced points along

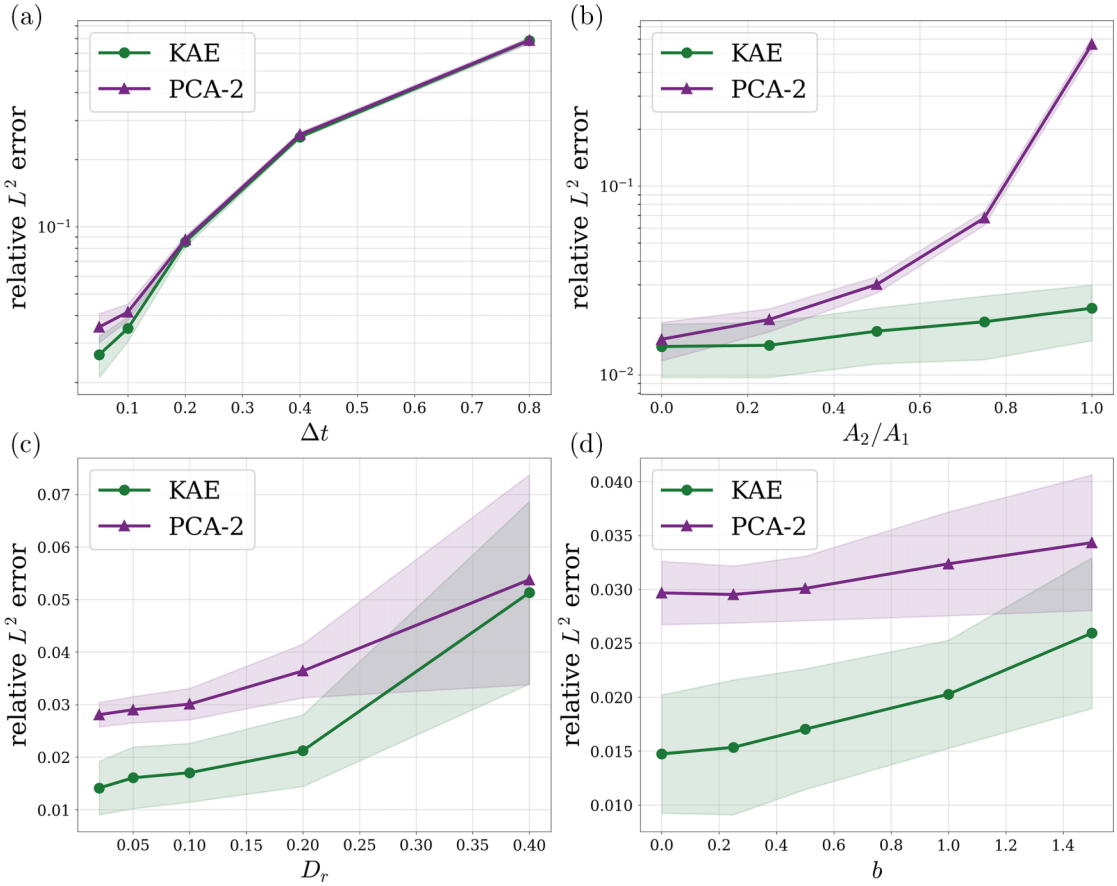


FIG. 7. Parameter dependence of the trajectory averaging for the synthetic waveform. (a) Sampling time, (b) ratio of second harmonics, (c) strength of radial diffusion, and (d) nonuniformity of phase velocity. The relative error is plotted with error bars indicating standard deviation over different sequences of stochastic waveforms.

arclength. The body-fixed frame tangent angles were obtained by computing a global rotation rate Ω, assumed constant in time, which minimizes

$$\mathrm{Var}_k\left[\frac{1}{N}\sum_{j=1}^{N} q^{(\mathrm{lab})}(s_j,t_k)-\Omega t_k\right], \tag{C1}$$

where $q^{(\mathrm{lab})}(s_j,t_k)$ denotes the laboratory-frame tangent angle at arclength s_j and time t_k.

APPENDIX D: ROBUSTNESS OF THE LEARNING PROCESS IN REAL DATA ANALYSIS

To examine the robustness of the learning process and the calculations of geometric PSF in Sec. IV, we tested the sensitivity by doubling and halving the network sizes, N_1 and N_2, and the regularization coefficients, λ_{lin} and λ_{rad}.

We confirmed for zebrafish sperm and bull sperm data that the KAE successfully learns the limit cycle with unit radius and isochron phase, regardless of changes in the network size and regularization parameters. The robustness of the obtained geometric PSF is measured by the L^2 relative error defined as

$$\epsilon_\alpha = ||Z^\alpha_W - Z^\alpha_{W,\mathrm{ref}}||/||Z^\alpha_{W,\mathrm{ref}}||, \tag{D1}$$

where $\alpha=\{x,y,\theta\}$ and $Z^\alpha_{W,\mathrm{ref}}$ is the data shown in Fig. 4. In the zebrafish sperm and bull sperm cases, the relative error was confirmed to be sufficiently small ($\epsilon_\alpha \leqslant 0.02$). The position of the peak of $|Z^\alpha_W|$ and the physical interpretations discussed in Sec. IV C are also confirmed to be robust under changes in learning processes.

In the *C. elegans* data, our KAE learning sometimes failed to extract the limit cycle in the latent space for particular changes in the network size and parameters. Noticeable declines in success rate were found when the network sized is doubled, when λ_{lin} is halved, and when λ_{rad} is doubled. The choice of network size should depend on the data size, and we do not need a large network as our latent manifold is a simple unit circle. The size of regularization parameters should be properly chosen to prevent the learned manifold from shrinking to a point, which constructs trivial linear dynamics in latent space.

Although the choice of network size and regularization parameters are common practical issues in machine learning architectures, in the current problem, failed learning was easily detected by checking the radius of the learned manifold and the value of the learned phase velocity, ω. More importantly, once the KAE passes this intuitive test, the obtained geometric PSF is found to be robust under changes in learning parameters. We confirmed that the relative L^2 error of the geometric PSF was $\epsilon_\alpha \leqslant 0.07$ and that the physical arguments along Figs. 4 and 5 are also robust.

[1] E. Cohen, H. Larocque, F. Bouchard, F. Nejadsattari, Y. Gefen, and E. Karimi, Geometric phase from Aharonov–Bohm to Pancharatnam–Berry and beyond, Nat. Rev. Phys. **1**, 437 (2019).

[2] G. Ma, M. Xiao, and C. T. Chan, Topological phases in acoustic and mechanical systems, Nat. Rev. Phys. **1**, 281 (2019).

[3] H. Xue, Y. Yang, and B. Zhang, Topological acoustics, Nat. Rev. Mater. **7**, 974 (2022).

[4] S. D. Kelly and R. M. Murray, Geometric phases and robotic locomotion, J. Rob. Syst. **12**, 417 (1995).

[5] J. Aguilar, T. Zhang, F. Qian, M. Kingsbury, B. McInroe, N. Mazouchova, C. Li, R. Maladen, C. Gong, M. Travers, R. L. Hatton, H. Choset, P. B. Umbanhowar, and D. I. Goldman, A review on locomotion robophysics: The study of movement at the intersection of robotics, soft matter and dynamical systems, Rep. Prog. Phys. **79**, 110001 (2016).

[6] S. Li, T. Wang, V. H. Kojouharov, J. McInerney, E. Aydin, Y. Ozkan-Aydin, D. I. Goldman, and D. Z. Rocklin, Robotic swimming in curved space via geometric phase, Proc. Natl. Acad. Sci. USA **119**, e2200924119 (2022).

[7] P. Delplace, J. B. Marston, and A. Venaille, Topological origin of equatorial waves, Science **358**, 1075 (2017).

[8] X. Fang, K. Kruse, T. Lu, and J. Wang, Nonequilibrium physics in biology, Rev. Mod. Phys. **91**, 045004 (2019).

[9] Z. Wang, J. Chen, and J. Ren, Geometric heat pump and no-go restrictions of nonreciprocity in modulated thermal diffusion, Phys. Rev. E **106**, L032102 (2022).

[10] A. Shapere and F. Wilczek, Geometry of self-propulsion at low Reynolds number, J. Fluid Mech. **198**, 557 (1989).

[11] R. Hatton and H. Choset, Nonconservativity and noncommutativity in locomotion, Eur. Phys. J. Spec. Top. **224**, 3141 (2015).

[12] L. Koens and E. Lauga, Geometric phase methods with Stokes theorem for a general viscous swimmer, J. Fluid Mech. **916**, A17 (2021).

[13] K. Ishimoto, J. Herault, and C. Moreau, Bending–compression coupling in extensible slender microswimmers, J. Fluid Mech. **1020**, A1 (2025).

[14] K. M. Dorgan, C. J. Law, and G. W. Rouse, Meandering worms: Mechanics of undulatory burrowing in muds, Proc. R. Soc. B **280**, 20122948 (2013).

[15] E. E. Keaveny and A. E. X. Brown, Predicting path from undulations for *C. elegans* using linear and nonlinear resistive force theory, Phys. Biol. **14**, 025001 (2017).

[16] D. Zhao, B. Bittner, G. Clifton, N. Gravish, and S. Revzen, Walking is like slithering: A unifying, data-driven view of locomotion, Proc. Natl. Acad. Sci. USA **119**, e2113222119 (2022).

[17] J. M. Rieser, B. Chong, C. Gong, H. C. Astley, P. E. Schiebel, K. Diaz, C. J. Pierce, H. Lu, R. L. Hatton, H. Choset, and D. I. Goldman, Geometric phase predicts locomotion performance in undulating living systems across scales, Proc. Natl. Acad. Sci. USA **121**, e2320517121 (2024).

[18] E. Lauga, *The Fluid Dynamics of Cell Motility*, Cambridge Texts in Applied Mathematics (Cambridge University Press, Cambridge, 2020).

[19] G. J. Stephens, B. Johnson-Kerner, W. Bialek, and W. S. Ryu, Dimensionality and dynamics in the behavior of *C. elegans*, PLoS Comput. Biol. **4**, e1000028 (2008).

[20] S. Werner, J. C. Rink, I. H. Riedel-Kruse, and B. M. Friedrich, Shape mode analysis exposes movement patterns in biology: Flagella and flatworms as case studies, PLoS ONE **9**, e113083 (2014).

[21] D. Nishiguchi, J. Iwasawa, H.-R. Jiang, and M. Sano, Flagellar dynamics of chains of active Janus particles fueled by an AC electric field, New J. Phys. **20**, 015002 (2018).

[22] B. J. Walker, S. Phuyal, K. Ishimoto, C.-K. Tung, and E. A. Gaffney, Computer-assisted beat-pattern analysis and the flagellar waveforms of bovine spermatozoa, R. Soc. Open Sci. **7**, 200769 (2020).

[23] A. Gholami, R. Ahmad, A. J. Bae, A. Pumir, and E. Bodenschatz, Waveform of free, hinged and clamped axonemes isolated from *C. reinhardtii*: Influence of calcium, New J. Phys. **24**, 053025 (2022).

[24] R. Ahmad, A. J. Bae, Y.-J. Su, S. G. Pozveh, E. Bodenschatz, A. Pumir, and A. Gholami, Bio-hybrid micro-swimmers propelled by flagella isolated from *C. reinhardtii*, Soft Matter **18**, 4767 (2022).

[25] K. E. Gutierrez, B. Thomases, P. E. Arratia, and R. D. Guy, Quantification of flagellar gait changes with combined shape mode analysis and swimming simulations, Philos. Trans. R. Soc. A **383**, 20240505 (2025).

[26] K. Ishimoto, H. Gadêlha, E. A. Gaffney, D. J. Smith, and J. Kirkman-Brown, Coarse-graining the fluid flow around a human sperm, Phys. Rev. Lett. **118**, 124501 (2017).

[27] B. Kralemann, L. Cimponeriu, M. Rosenblum, A. Pikovsky, and R. Mrowka, Phase dynamics of coupled oscillators reconstructed from data, Phys. Rev. E **77**, 066205 (2008).

[28] *The Koopman Operator in Systems and Control: Concepts, Methodologies, and Applications*, edited by A. Mauroy, I. Mezić, and Y. Susuki, Lecture Notes in Control and Information Sciences Vol. 4 (Springer International Publishing, Cham, 2020).

[29] I. Mezić, Koopman operator, geometry, and learning of dynamical systems, Not. Am. Math. Soc. **68**, 1087 (2021).

[30] S. L. Brunton, M. Budišić, E. Kaiser, and J. N. Kutz, Modern koopman theory for dynamical systems, SIAM Rev. **64**, 229 (2022).

[31] A. Mauroy, I. Mezić, and J. Moehlis, Isostables, isochrons, and Koopman spectrum for the action–angle representation of stable fixed point dynamics, Phys. D: Nonlinear Phenom. **261**, 19 (2013).

[32] D. Wilson and J. Moehlis, Isostable reduction of periodic orbits, Phys. Rev. E **94**, 052213 (2016).

[33] S. Shirasaka, W. Kurebayashi, and H. Nakao, Phase-amplitude reduction of transient dynamics far from attractors for limit-cycling systems, Chaos **27**, 023119 (2017).

[34] K. Yawata, K. Fukami, K. Taira, and H. Nakao, Phase autoencoder for limit-cycle oscillators, Chaos **34**, 063111 (2024).

[35] Y. Hiruta and K. Ishimoto, Autoencoder for limit cycle in Kolmogorov flow, J. Phys. Soc. Jpn. **94**, 064401 (2025).

[36] K. Yawata, R. Sakuma, K. Fukami, K. Taira, and H. Nakao, Phase autoencoder for rapid data-driven synchronization of rhythmic spatiotemporal patterns, Phys. Rev. E **112**, 064211 (2025).

[37] S. Shirasaka, W. Kurebayashi, and H. Nakao, Phase reduction theory for hybrid nonlinear oscillators, Phys. Rev. E **95**, 012212 (2017).

[38] N. Takeishi, Y. Kawahara, and T. Yairi, Learning Koopman invariant subspaces for dynamic mode decomposition, in *Proceedings of the 31st International Conference on Neural*

Information Processing Systems, *NIPS'17* (Curran Associates, Red Hook, NY, 2017), pp. 1130–1140.

[39] B. Lusch, J. N. Kutz, and S. L. Brunton, Deep learning for universal linear embeddings of nonlinear dynamics, Nat. Commun. **9**, 4950 (2018).

[40] Y. Li, H. He, J. Wu, D. Katabi, and A. Torralba, Learning compositional Koopman operators for model-based control, arXiv:1910.08264.

[41] I. Nayak, A. Chakrabarti, M. Kumar, F. L. Teixeira, and D. Goswami, Temporally consistent Koopman autoencoders for forecasting dynamical systems, Sci. Rep. **15**, 22127 (2025).

[42] Y. Kato, J. Zhu, W. Kurebayashi, and H. Nakao, Asymptotic phase and amplitude for classical and semiclassical stochastic oscillators via Koopman operator theory, Mathematics **9**, 2188 (2021).

[43] R. Ma, G. S. Klindt, I. H. Riedel-Kruse, F. Jülicher, and B. M. Friedrich, Active phase and amplitude fluctuations of flagellar beating, Phys. Rev. Lett. **113**, 048101 (2014).

[44] G. Arvanitidis, L. K. Hansen, and S. Hauberg, Latent space oddity: On the curvature of deep generative models, arXiv:1710.11379.

[45] H. Shao, A. Kumar, and P. T. Fletcher, The Riemannian geometry of deep generative models, arXiv:1711.08014.

[46] Y. Lee, A geometric perspective on autoencoders, arXiv:2309.08247.

[47] J. S. Guasto, J. B. Estrada, F. Menolascina, L. J. Burton, M. Patel, C. Franck, A. E. Hosoi, R. K. Zimmer, and R. Stocker, Flagellar kinematics reveals the role of environment in shaping sperm motility, J. R. Soc. Interface **17**, 20200525 (2020).

[48] B. M. Friedrich, I. H. Riedel-Kruse, J. Howard, and F. Jülicher, High-precision tracking of sperm swimming fine structure provides strong test of resistive force theory, J. Exp. Biol. **213**, 1226 (2010).

[49] J. T. Pierce-Shimomura, B. L. Chen, J. J. Mun, R. Ho, R. Sarkis, and S. L. McIntire, Genetic analysis of crawling and swimming locomotory patterns in *C. elegans*, Proc. Natl. Acad. Sci. USA **105**, 20982 (2008).

[50] A. Mauroy and I. Mezić, On the use of Fourier averages to compute the global isochrons of (quasi)periodic dynamics, Chaos **22**, 033112 (2012).

[51] K. Egan, W. Li, and R. Carvalho, Automatically discovering ordinary differential equations from data with sparse regression, Commun. Phys. **7**, 20 (2024).

[52] C. V. Neal, A. L. Hall-McNair, J. Kirkman-Brown, D. J. Smith, and M. T. Gallagher, Doing more with less: The flagellar end piece enhances the propulsive effectiveness of human spermatozoa, Phys. Rev. Fluids **5**, 073101 (2020).

[53] K. Ishimoto, C. Moreau, and J. Herault, Robust undulatory locomotion through neuromechanical adjustments in a dissipative medium, J. R. Soc. Interface **22**, 20240688 (2025).

[54] P. Hernández-Herrera, H. O. Hernández, A. Bribiesca-Sanchez, F. Montoya, D. S. Díaz-Guerrero, A. Darszon, and G. Corkidi, 3D + t human sperm flagellum centerline dataset, Sci. Data **13**, 505 (2026).

[55] I. Mezić, Spectrum of the Koopman operator, spectral expansions in functional spaces, and state-space geometry, J. Nonlinear Sci. **30**, 2091 (2020).

[56] K. Y. Wan, K. C. Leptos, and R. E. Goldstein, Lag, lock, sync, slip: The many 'phases' of coupled flagella, J. R. Soc. Interface **11**, 20131160 (2014).

[57] B. Chong, J. He, S. Li, E. Erickson, K. Diaz, T. Wang, D. Soto, and D. I. Goldman, Self-propulsion via slipping: Frictional swimming in multilegged locomotors, Proc. Natl. Acad. Sci. USA **120**, e2213698120 (2023).

[58] S. Deng, J. Liu, B. Datta, A. Pantula, D. H. Gracias, T. D. Nguyen, B. A. Bittner, and N. J. Cowan, in *2024 IEEE International Conference on Robotics and Automation (ICRA)* (2024), pp. 14988–14994.

[59] R. Hu and S. Revzen, Learning the geometric mechanics of robot motion using Gaussian mixtures, arXiv:2502.05309.

[60] Y. Yang, N. L. Hecht, Y. Salaman-Maclara, N. Justus, Z. A. Thomas, F. Rozaidi, and R. L. Hatton, Geometric data-driven multi-jet locomotion inspired by salps, in *IEEE Transactions on Robotics*, doi: 10.1109/TRO.2026.3719094.

[61] P. H. Htet and K. Ishimoto, 2026, 501d817, github, https://github.com/kentaishimoto/kae_biolocomotion.git.

[62] H. Nakao, Phase reduction approach to synchronisation of nonlinear oscillators, Contemp. Phys. **57**, 188 (2016).