Interpretable conservation law estimation by deriving the symmetries of dynamics from trained deep neural networks

Yoh-ichi Mototake*
The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan

(Received 19 April 2020; accepted 8 February 2021; published 18 March 2021)

Understanding complex systems with their reduced model is one of the central roles in scientific activities. Although physics has greatly been developed with the physical insights of physicists, it is sometimes challenging to build a reduced model of such complex systems on the basis of insight alone. We propose a framework that can infer hidden conservation laws of a complex system from deep neural networks (DNNs) that have been trained with physical data of the system. The purpose of the proposed framework is not to analyze physical data with deep learning but to extract interpretable physical information from trained DNNs. With Noether's theorem and by an efficient sampling method, the proposed framework infers conservation laws by extracting the symmetries of dynamics from trained DNNs. The proposed framework is developed by deriving the relationship between a manifold structure of a time-series data set and the necessary conditions for Noether's theorem. The feasibility of the proposed framework has been verified in some primitive cases in which the conservation law is well known. We also apply the proposed framework to conservation law estimation for a more practical case, that is, a large-scale collective motion system in the metastable state, and we obtain a result consistent with that of a previous study.

DOI: 10.1103/PhysRevE.103.033303

I. INTRODUCTION

Understanding complex systems with reduced models is one of the central roles in scientific activities. Some complex systems are modeled as low-dimensional canonical dynamical systems. For example, reduced models have been developed for large-scale collective motion systems, which are a type of large-scale complex system with order (e.g., plasma, acoustic waves, or vortex systems) [1–5]. To develop reduced models, collective coordinates, such as the Fourier basis of a density or charge distribution [1–4], or a vortex feature space [5], have been introduced. Then, a Hamiltonian that describes the coarse-grained properties of a dynamical system has been derived. Thus, to develop a reduced model, it is necessary to introduce collective coordinates and derive the Hamiltonian in the coordinates. The obtained Hamiltonian is verified by confirming that it can reconstruct the properties of the phenomena analyzed. This approach relies heavily on the physical insights of physicists; it would not work for modeling a dynamical system that features a more complicated structure. One example is the collective motion of living things such as fish or birds; such systems frequently have stable but very complicated patterns in a metastable state [6,7].

The problem we consider here is how to infer the reduced model using machine-learning methods. As mentioned above, this involves the solution of two problems: estimation of a coordinate system and construction of a reduced model in the coordinate system. One way to solve these problems is to construct a Hamiltonian on the basis of a given coordinate system and search for a coordinate system that improves the model. Several machine-learning methods for inferring the Hamiltonian from a time-series data set have been developed [8–11]. These methods can be broadly divided into two types. In one type, the Hamiltonian is inferred by regressing the data with an explicit function, such as the linear sum of multiple basis functions [8]. However, in the case of inferring a reduced model that consists of complicated unknown basis functions, the method only infers the approximated reduced model using an approximated function, such as a polynomial function. In the second type, a Hamiltonian is modeled by a deep-learning technique [9–11]. In this case, an explicit function used in the first one is not required. On the basis of these machine-learning methods, the coordinate system could be searched using statistical criteria such as the prediction or generalization error of the inferred Hamiltonian.

There are inherent difficulties in building a reduced model by the machine-learning approach. Such an approach finds a Hamiltonian that has properties that only hold for the given data. Historically, physicists have achieved great success in constructing reduced models by abstracting knowledge obtained from observational data and building universal models that can explain various physical phenomena, not just the given data. For example, in thermodynamics, a reduced model that describes the molecular motion of a gas was linked to chemical reaction theory by Gibbs [12,13]. This is one of

*mototake@ism.ac.jp

the most successful uses of a reduced model. That is, a good reduced model and a good coordinate system mean that the performance is high not only for the given data.

To realize such a successful reduced model, it is important to interpret the knowledge obtained during data analysis and develop a model that can be applied to different phenomena by combining the explicit and implicit knowledge of physics. In general, an inferred Hamiltonian modeled using deep neural networks (DNNs) is hardly interpretable because DNNs are models with enormous degrees of freedom. If all physical knowledge is quantified, it will be possible to construct a reduced model with a DNN, but this is an impractical assumption at present. Therefore, it is difficult for a machine-learning approach to realize the same function as a physicist could because a physicist can flexibly interpret phenomena by utilizing explicit or implicit physical knowledge and construct a reduced model.

To overcome this problem, here we attempt to extract abstract information directly from physical data without constructing a reduced model. A given coordinate system can be evaluated on the basis of the information. Furthermore, the obtained information can also help physicists construct a reduced model. The purpose of this paper is to develop a machine-learning framework that extracts interpretable abstract information from physical data and assist physicists in building reduced models.

The proposed method is developed using knowledge about DNNs. Results of several studies [14–19] suggest that DNNs can model the distribution of data sets as manifolds, which can be embedded in a low-dimensional Euclidean space. Studies applying DNNs to physical data have employed a time-series data set from the phase space (comprising position and momentum) [20–24] or a spin system data set from the configuration space [25–33]. Such data sets have a low–dimensional manifold structure, which implies that the system has a small number of degrees of freedom. Such a low-dimensional manifold structure should be related to certain physical constraints, such as conservation laws. That is, a manifold structure modeled by a DNN should be related to the conservation law or order of the system.

The proposed method is derived from Noether's theorem [34], which connects the symmetry of the Hamiltonian and the conservation laws. We derive the relationship between the symmetry of the Hamiltonian system and the distribution of the time-series data set of a dynamical system. On this basis, we develop a method of inferring the symmetry of a data manifold modeled by a deep autoencoder [15] and determine the conservation laws of the system. To infer the conservation laws, we only need the tangent space around the identity element of the manifold formed by a continuous transformation group that corresponds to the symmetry of the system. Therefore, unlike Hamiltonian estimation, the conservation law estimation requires modeling the manifold by polynomials up to only a first-order accuracy. This means that the conservation laws can be inferred with arbitrary precision by polynomial approximation.

This paper is organized as follows. In Sec. II A, we show the derivation of the relationship between the symmetry of the time-series data-set distribution and the conservation law using Noether's theorem. In Sec. III A, we describe our proposed method of inferring the symmetry of the time-series data manifold. In Sec. III B, we also describe another proposed method of inferring the conservation law from the obtained symmetry. In Sec. IV, to confirm the effectiveness of the proposed methods, we apply them to three cases, one T(1) and two SO(2) systems, corresponding to constant-velocity linear motion, a central force system, and a large-scale collective motion system called the Reynolds model [35]. In Sec. V, we present a summary and discussion.

II. THEORY

A. Noether's theorem

Noether's theorem connects continuous symmetries of a Hamiltonian system with conservation laws [34]. It is often described in the $(2d+1)$-dimensional extended phase space $\Gamma \times \mathbb{R}$, $(\mathbf{q}, \mathbf{p}) \coloneqq (q_0 = t, q_1, \cdots, q_d, p_1, \cdots, p_d)$. The theorem can also be described in the $(2d+2)$-dimensional space $\Gamma \times \mathbb{R} \times \mathbb{R}$, $(q_0 = t, q_1, \cdots, q_d, p_0 = -H, p_1, \cdots, p_d)$. In this paper, we describe the theory in the $(2d+2)$-dimensional space as follows. We consider Hamiltonian systems in the $(2d+2)$-dimensional space $\Gamma \times \mathbb{R} \times \mathbb{R}$, and restrict ourselves to the case where the system's Hamiltonian belongs to a C^2 class function $H(\mathbf{q}, \mathbf{p})$. The Hamiltonian representation of Noether's theorem is described as follows [36]. Assume that $H(\mathbf{q}, \mathbf{p})$ and the canonical equations of motion $\frac{\partial H(\mathbf{q},\mathbf{p})}{\partial q_i} = -\dot{p}_i$ and $\frac{\partial H(\mathbf{q},\mathbf{p})}{\partial p_i} = \dot{q}_i$ are invariant under the infinitesimal transformation $(q_i', p_i') = (q_i + \delta q_{ij}, p_i + \delta p_{ij})$, where $i = 1, \ldots, d$, and j is the index of the direction of the infinitesimal transformation corresponding to a conservation law. Then, on the basis of Noether's theorem, the conserved value G_j satisfies the following equation:

$$(\delta q_{ij}, \delta p_{ij}) = \left(\frac{\partial G_j}{\partial p_i}, -\frac{\partial G_j}{\partial q_i} \right). \tag{1}$$

The canonical transformation that makes the Hamiltonian system invariant is given as

$$\mathbb{C}_{\mathrm{inv}}(\boldsymbol{\theta}) : \Gamma \times \mathbb{R} \times \mathbb{R} \longrightarrow \Gamma \times \mathbb{R} \times \mathbb{R}, \tag{2}$$

$$(\mathbf{q}, \mathbf{p}) \longmapsto (Q, \mathcal{P}) \coloneqq (Q(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta}), \mathcal{P}(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta})), \tag{3}$$

where $Q(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta})$ and $\mathcal{P}(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta})$ represent the invariant transformation functions of coordinate $(\mathbf{q}, \mathbf{p})$ to $(Q, \mathcal{P})$, and $\boldsymbol{\theta}$ represents a d_θ-dimensional continuous parameter characterizing the transformation that satisfies $Q(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta} = \vec{0}) = \mathbf{q}$ and $\mathcal{P}(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta} = \vec{0}) = \mathbf{p}$. We call this transformation an invariant transformation in this paper. A set of the invariant transformations characterized by the continuous parameters $\boldsymbol{\theta}$ forms a Lie group. By the first-order Taylor expansion of $Q_i(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta})$ and $\mathcal{P}_i(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta})$ around $\boldsymbol{\theta} = \vec{0}$, we have the infinitesimal transformation,

$$(\delta q_{ij}, \delta p_{ij}) = \left(\varepsilon \left. \frac{\partial Q_i(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta}=\vec{0}}, \varepsilon \left. \frac{\partial \mathcal{P}_i(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta}=\vec{0}} \right), \tag{4}$$

where $|\varepsilon| \ll 1$.

Note that the dimension of the continuous parameter d_θ corresponds to the number of conservation laws, and by our proposed methods, we estimate conservation laws including d_θ.

B. Invariance of Hamiltonian and time-series data set

We show the relationship between such an invariant transformation and the time-series data set of a dynamical system in the $(2d+2)$-dimensional space $(\mathbf{q}, \mathbf{p})$. Here, we define the N sample time-series data set D as $D \coloneqq \{\mathbf{q}^i_{t_i}, \mathbf{p}^i_{t_i}, \mathbf{q}^i_{t_i+\Delta t}, \mathbf{p}^i_{t_i+\Delta t}\}_{i=1}^N$, where $\mathbf{q}^i_{t_i}$ and $\mathbf{p}^i_{t_i}$ represent the generalized position and momentum at time t_i, and $t_i + \Delta t$ represents a time evolution of Δt.

The transformation of the $(2d+2)$-dimensional space $(\mathbf{q}, \mathbf{p})$ is defined as

$$\mathbb{c} : \Gamma \times \mathbb{R} \times \mathbb{R} \longrightarrow \Gamma \times \mathbb{R} \times \mathbb{R}, \tag{5}$$

$$(\mathbf{q}, \mathbf{p}) \longmapsto (\mathbf{Q}, \mathbf{P}) \coloneqq (\mathbf{Q}(\mathbf{q}, \mathbf{p}), \mathbf{P}(\mathbf{q}, \mathbf{p})), \tag{6}$$

where $\mathbf{Q}(\mathbf{q}, \mathbf{p})$ and $\mathbf{P}(\mathbf{q}, \mathbf{p})$ represent transformation functions of the coordinate $(\mathbf{q}, \mathbf{p})$ to $(\mathbf{Q}, \mathbf{P})$; the transformation is not limited to the invariant transformation. It is assumed that $\mathbb{c}$ has the inverse transformation:

$$\mathbb{c}^{-1} : \Gamma \times \mathbb{R} \times \mathbb{R} \longrightarrow \Gamma \times \mathbb{R} \times \mathbb{R}, \tag{7}$$

$$(\mathbf{Q}, \mathbf{P}) \longmapsto (\mathbf{q}, \mathbf{p}) \coloneqq (\mathbf{q}(\mathbf{Q}, \mathbf{P}), \mathbf{p}(\mathbf{Q}, \mathbf{P})). \tag{8}$$

The transformed Hamiltonian $H'(\mathbf{q}, \mathbf{p})$ obeying this transformation is defined as $H'(\mathbf{Q}, \mathbf{P}) \coloneqq H(\mathbf{q}(\mathbf{Q}, \mathbf{P}), \mathbf{p}(\mathbf{Q}, \mathbf{P}))$. The necessary and sufficient condition for the transformation $\mathbb{c}$ acting on $H(\mathbf{q}, \mathbf{p})$ to be identical, $H'(\mathbf{q}, \mathbf{p}) \equiv H(\mathbf{q}, \mathbf{p})$, is equivalent to

$$\forall E,\ \{\mathbf{q}, \mathbf{p} \mid H(\mathbf{q}, \mathbf{p}) = E\} = \{\mathbf{Q}, \mathbf{P} \mid H(\mathbf{q}, \mathbf{p}) = E\}. \tag{9}$$

This condition is derived in Appendix A and implies that the transformation invariance of a Hamiltonian is equivalent to that of the energy surface at each energy level in the space $\Gamma \times \mathbb{R} \times \mathbb{R}$. If the time-series data set D has all possible data points under the Hamiltonian $H(\mathbf{q}, \mathbf{p})$, the subset of D with respect to $\mathbf{q}^i_{t_i}$ and $\mathbf{p}^i_{t_i}$ is understood as this energy surface.

C. Invariance of canonical equations and time-series data set

Next, we consider the relationship between the invariance of canonical equations of motion and the time-series data set of the dynamical system. If the canonical equations of motion are discretized with respect to time differentiation, the discretized canonical equations of motion are obtained as

$$\mathbf{q}_{t+\Delta t} = \mathbf{u}(\mathbf{q}_t, \mathbf{p}_t) \coloneqq \frac{\partial H(\mathbf{q}_t, \mathbf{p}_t)}{\partial \mathbf{p}_t}\Delta t + \mathbf{q}_t, \tag{10}$$

$$\mathbf{p}_{t+\Delta t} = \mathbf{v}(\mathbf{q}_t, \mathbf{p}_t) \coloneqq -\frac{\partial H(\mathbf{q}_t, \mathbf{p}_t)}{\partial \mathbf{q}_t}\Delta t + \mathbf{p}_t, \tag{11}$$

where $\mathbf{q}_t$ and $\mathbf{p}_t$ represent the variables that evolved according to time t, and $\mathbf{u}(\mathbf{q}_t, \mathbf{p}_t)$ and $\mathbf{v}(\mathbf{q}_t, \mathbf{p}_t)$ are elements of the C^1 map $\mathbb{u}$ defined as

$$\mathbb{u} : \Gamma \times \mathbb{R} \times \mathbb{R} \longrightarrow \Gamma \times \mathbb{R} \times \mathbb{R}, \tag{12}$$

$$(\mathbf{q}_t, \mathbf{p}_t) \longmapsto (\mathbf{q}_{t+\Delta t}, \mathbf{p}_{t+\Delta t}) \coloneqq (\mathbf{u}(\mathbf{q}_t, \mathbf{p}_t), \mathbf{v}(\mathbf{q}_t, \mathbf{p}_t)). \tag{13}$$

Following the transformations $\mathbf{Q}(\mathbf{q}, \mathbf{p})$ and $\mathbf{P}(\mathbf{q}, \mathbf{p})$ in Eq. (5), these equations can be rewritten as

$$\begin{aligned}\mathbf{Q}_{T+\Delta T} &= \mathbf{Q}(\mathbf{q}_{t+\Delta t}, \mathbf{p}_{t+\Delta t}) \\ &= \mathbf{u}'(\mathbf{Q}_T, \mathbf{P}_T) \coloneqq \mathbf{Q}[\mathbf{u}(\mathbf{q}(\mathbf{Q}_T, \mathbf{P}_T), \mathbf{p}(\mathbf{Q}_T, \mathbf{P}_T)), \\ &\quad \times \mathbf{v}(\mathbf{q}(\mathbf{Q}_T, \mathbf{P}_T), \mathbf{p}(\mathbf{Q}_T, \mathbf{P}_T))],\end{aligned} \tag{14}$$

$$\begin{aligned}\mathbf{P}_{T+\Delta T} &= \mathbf{P}(\mathbf{q}_{t+\Delta t}, \mathbf{p}_{t+\Delta t}) \\ &= \mathbf{v}'(\mathbf{Q}_T, \mathbf{P}_T) \coloneqq \mathbf{P}[\mathbf{u}(\mathbf{q}(\mathbf{Q}_T, \mathbf{P}_T), \mathbf{p}(\mathbf{Q}_T, \mathbf{P}_T)), \\ &\quad \times \mathbf{v}(\mathbf{q}(\mathbf{Q}_T, \mathbf{P}_T), \mathbf{p}(\mathbf{Q}_T, \mathbf{P}_T))],\end{aligned} \tag{15}$$

where $T = Q_0$, $\Delta T = \Delta Q_0$. For the transformation $(\mathbf{Q}, \mathbf{P}) = (\mathbf{Q}(\mathbf{q}, \mathbf{p}), \mathbf{P}(\mathbf{q}, \mathbf{p}))$ to be a canonical transformation, the following conditions must be satisfied:

$$\mathbf{u}'(\mathbf{Q}_T, \mathbf{P}_T) \equiv \frac{\partial H'(\mathbf{Q}_T, \mathbf{P}_T)}{\partial \mathbf{P}_T}\Delta T + \mathbf{Q}_T, \tag{16}$$

$$\mathbf{v}'(\mathbf{Q}_T, \mathbf{P}_T) \equiv -\frac{\partial H'(\mathbf{Q}_T, \mathbf{P}_T)}{\partial \mathbf{Q}_T}\Delta T + \mathbf{P}_T. \tag{17}$$

If H and H' are identically equal, the right sides of Eqs. (16) and (17) are equivalent to

$$\frac{\partial H(\mathbf{Q}_T, \mathbf{P}_T)}{\partial \mathbf{P}_T}\Delta T + \mathbf{Q}_T \equiv \mathbf{u}(\mathbf{Q}_T, \mathbf{P}_T), \tag{18}$$

$$-\frac{\partial H(\mathbf{Q}_T, \mathbf{P}_T)}{\partial \mathbf{Q}_T}\Delta T + \mathbf{P}_T \equiv \mathbf{v}(\mathbf{Q}_T, \mathbf{P}_T). \tag{19}$$

Thus, we only need to prove that the functions $\mathrm{u}(\cdot, \cdot)$ and $\mathrm{u}'(\cdot, \cdot)$ and the functions $\mathrm{v}(\cdot, \cdot)$ and $\mathrm{v}'(\cdot, \cdot)$ are identically equal:

$$\begin{cases}\mathbf{u}'(\mathbf{Q}_t, \mathbf{P}_t) \equiv \mathbf{u}(\mathbf{Q}_t, \mathbf{P}_t) \\ \mathbf{v}'(\mathbf{Q}_t, \mathbf{P}_t) \equiv \mathbf{v}(\mathbf{Q}_t, \mathbf{P}_t),\end{cases} \tag{20}$$

$$\Leftrightarrow \quad \forall (\mathbf{Q}_t, \mathbf{P}_t) \in \Gamma \times \mathbb{R} \times \mathbb{R}, \begin{cases}\mathbf{u}'(\mathbf{Q}_t, \mathbf{P}_t) = \mathbf{u}(\mathbf{Q}_t, \mathbf{P}_t) \\ \mathbf{v}'(\mathbf{Q}_t, \mathbf{P}_t) = \mathbf{v}(\mathbf{Q}_t, \mathbf{P}_t),\end{cases} \tag{21}$$

$$\Leftrightarrow \quad \forall (\mathbf{q}_t, \mathbf{p}_t) \in \Gamma \times \mathbb{R} \times \mathbb{R}, \begin{cases}\mathbf{u}'(\mathbf{q}_t, \mathbf{p}_t) = \mathbf{u}(\mathbf{q}_t, \mathbf{p}_t) \\ \mathbf{v}'(\mathbf{q}_t, \mathbf{p}_t) = \mathbf{v}(\mathbf{q}_t, \mathbf{p}_t),\end{cases} \tag{22}$$

$$\Leftrightarrow \quad \begin{cases}\mathbf{u}'(\mathbf{q}_t, \mathbf{p}_t) \equiv \mathbf{u}(\mathbf{q}_t, \mathbf{p}_t) \\ \mathbf{v}'(\mathbf{q}_t, \mathbf{p}_t) \equiv \mathbf{v}(\mathbf{q}_t, \mathbf{p}_t).\end{cases} \tag{23}$$

Furthermore, Eq. (23) is equivalent to the following condition (see Appendix B):

$$\begin{aligned}&\{\mathbf{q}_{t+\Delta t}, \mathbf{p}_{t+\Delta t}, \mathbf{q}_t, \mathbf{p}_t \mid (\mathbf{q}_{t+\Delta t}, \mathbf{p}_{t+\Delta t}) \\ &\quad = (\mathbf{u}(\mathbf{q}_t, \mathbf{p}_t), \mathbf{v}(\mathbf{q}_t, \mathbf{p}_t))\} \\ &\quad = \{\mathbf{Q}_{T+\Delta T}, \mathbf{P}_{T+\Delta T}, \mathbf{Q}_T, \mathbf{P}_T \mid (\mathbf{q}_{t+\Delta t}, \mathbf{p}_{t+\Delta t}) \\ &\quad = (\mathbf{u}(\mathbf{q}_t, \mathbf{p}_t), \mathbf{v}(\mathbf{q}_t, \mathbf{p}_t))\}.\end{aligned} \tag{24}$$

The time-series data set D is understood as the part of the subspace given on the left side of Eq. (24).

D. Noether's theorem and time-series data set

By combining the conditions obtained in the previous two subsections, we obtain the condition that the Hamiltonian and canonical equations are simultaneously invariant under the transformation. The condition is acquired

as

$$\begin{aligned}\forall E,\ &\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \mid H(\mathbf{q}_t,\mathbf{p}_t)=E,\mathbf{p}_{t+\Delta t}\\ &=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\Big\}\\ &=\{\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T},\mathbf{Q}_T,\mathbf{P}_T \mid H(\mathbf{q}_t,\mathbf{p}_t)=E,\mathbf{p}_{t+\Delta t}\\ &=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\Big\}. \quad (25)\end{aligned}$$

If the time-series data set D has all possible data points under the Hamiltonian $H(\mathbf{q},\mathbf{p})$ and the canonical equations, D is equivalent to the subspace defined on the left side of Eq. (25). Thus, the symmetry of the Hamiltonian system is associated with the symmetry of the time-series data set D. The transformation set satisfying Eq. (25), $\{\mathbf{Q}(\mathbf{q},\mathbf{p}),\mathbf{P}(\mathbf{q},\mathbf{p}) \mid \text{satisfy Eq. (25)}\}$, is the same as the invariant transformation set $\mathbb{c}_{\text{inv}}:\ \{\boldsymbol{Q}(\mathbf{q},\mathbf{p},\boldsymbol{\theta}),\boldsymbol{\mathcal{P}}(\mathbf{q},\mathbf{p},\boldsymbol{\theta}) \mid \boldsymbol{\theta}\in\mathbb{R}^{d_\theta}\}$ under the discretized equations of motion.

The transformed data set in Eq. (25),

$$\left\{\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T},\mathbf{Q}_T,\mathbf{P}_T \;\middle|\; H(\mathbf{q}_t,\mathbf{p}_t)=E,\mathbf{p}_{t+\Delta t}=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\right\}, \quad (26)$$

is obtained by the time evolution $t\to T$ of time-series data set at t:

$$\left\{\mathbf{Q}_{t+\Delta t},\mathbf{P}_{t+\Delta t},\mathbf{Q}_t,\mathbf{P}_t \;\middle|\; H(\mathbf{q}_t,\mathbf{p}_t)=E,\mathbf{p}_{t+\Delta t}=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\right\}. \quad (27)$$

If the Hamiltonian is given, we can obtain the time-evolved data set by evolving the data set obeying the canonical equations of motion. Even when the Hamiltonian is not given, we can obtain a time-evolved data set as follows. Assume that we have a time-series data set at $(t,t+\Delta t,t+2\Delta t,\dots,t+s\Delta t,\dots)$, where s is $\mathbb{Z}_{\geqslant 0}$. The time transformation of data from t to T can be approximated by replacing T with T':

$$T'=t+s\Delta t, \quad (28)$$

$$s=\underset{s}{\operatorname{argmin}}\,|T-(t+s\Delta t)|. \quad (29)$$

There is no guarantee that all energy states in the reduced Hamiltonian are realized in the original complex system. In particular, when constructing a reduced model of a metastable state, only its energy state is realized. To overcome this difficulty, we introduce the different expressions of the condition in Eq. (25). Let E_i be a real number representing one energy state. We also define the transformation

$$\mathbb{c}_i:\ \Gamma\times\mathbb{R}\times\mathbb{R}\longrightarrow\Gamma\times\mathbb{R}\times\mathbb{R}, \quad (30)$$

$$(\mathbf{q},\mathbf{p})\longmapsto(\mathbf{Q},\mathbf{P}):=(\mathbf{Q}_i(\mathbf{q},\mathbf{p}),\mathbf{P}_i(\mathbf{q},\mathbf{p})), \quad (31)$$

which satisfy

$$\begin{aligned}&\left\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \;\middle|\; H(\mathbf{q}_t,\mathbf{p}_t)=E_i,\mathbf{p}_{t+\Delta t}=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\right\}\\ &=\left\{\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T},\mathbf{Q}_T,\mathbf{P}_T \;\middle|\; H(\mathbf{q}_t,\mathbf{p}_t)=E_i,\mathbf{p}_{t+\Delta t}=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\right\}. \quad (32)\end{aligned}$$

The condition of Eq. (25) can be approximately rewritten using the condition Eq. (32),

$$\begin{aligned}\forall i\in\Lambda_E,\ &\left\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \;\middle|\; H(\mathbf{q}_t,\mathbf{p}_t)=E_i,\mathbf{p}_{t+\Delta t}=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\right\}\\ &=\left\{\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T},\mathbf{Q}_T,\mathbf{P}_T \;\middle|\; H(\mathbf{q}_t,\mathbf{p}_t)=E_i,\mathbf{p}_{t+\Delta t}=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\right\}, \quad (33)\end{aligned}$$

where Λ_E is the index set of all discretized energy states. Therefore, the transformation set that satisfies the condition Eq. (25) can be reexpressed as a union of the divided conditions:

$$\begin{aligned}\{\mathbf{Q}(\mathbf{q},\mathbf{p}),\mathbf{P}(\mathbf{q},\mathbf{p}) \mid \text{satisfy Eq. (25)}\} &\sim \{\mathbf{Q}_i(\mathbf{q},\mathbf{p}),\mathbf{P}_i(\mathbf{q},\mathbf{p}) \mid \forall i\in\Lambda_E,\text{satisfy Eq. (32)}\} \quad (34)\\ &=\bigcap_{i\in\Lambda_E}\{\mathbf{Q}_i(\mathbf{q},\mathbf{p}),\mathbf{P}_i(\mathbf{q},\mathbf{p}) \mid \text{satisfy Eq. (32)}\}. \quad (35)\end{aligned}$$

This implies that the invariant transformation set for a certain energy E_i must include some invariant transformations for the total energy. Thus, candidate transformations that make the Hamiltonian and canonical equations invariant are obtained as the transformations that make the subspace

$$S_i:=\left\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \;\middle|\; H(\mathbf{q}_t,\mathbf{p}_t)=E_i,\mathbf{p}_{t+\Delta t}=\mathbf{p}_t-\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{q}_t},\mathbf{q}_{t+\Delta t}=\mathbf{q}_t+\frac{\partial H(\mathbf{q}_t,\mathbf{p}_t)}{\partial \mathbf{p}_t}\right\} \quad (36)$$

invariant. This expression is useful for finding the candidates of symmetries in a complex dynamical system, such as dynamics in the metastable state.

In a finite time measurement or simulation, only data D of a subset of S_i can be obtained. On the basis of the following two physical principles, we can estimate S_i from data D. The first principle is described as follows. The subspace S_i can be represented as a product space of two subspaces:

$$S_i = S_i^a \times S_i^b, \tag{37}$$

$$S_i^a = \left\{ \mathbf{q}_t, \mathbf{p}_t \;\middle|\; H(\mathbf{q}_t, \mathbf{p}_t) = E_i, \mathbf{p}_{t+\Delta t} = \mathbf{p}_t - \frac{\partial H(\mathbf{q}_t, \mathbf{p}_t)}{\partial \mathbf{q}_t}, \mathbf{q}_{t+\Delta t} = \mathbf{q}_t + \frac{\partial H(\mathbf{q}_t, \mathbf{p}_t)}{\partial \mathbf{p}_t} \right\} \tag{38}$$

$$= \{\mathbf{q}_t, \mathbf{p}_t \mid H(\mathbf{q}_t, \mathbf{p}_t) = E_i\}, \tag{39}$$

$$S_i^b = \left\{ \mathbf{q}_{t+\Delta t}, \mathbf{p}_{t+\Delta t} \;\middle|\; H(\mathbf{q}_t, \mathbf{p}_t) = E_i, \mathbf{p}_{t+\Delta t} = \mathbf{p}_t - \frac{\partial H(\mathbf{q}_t, \mathbf{p}_t)}{\partial \mathbf{q}_t}, \mathbf{q}_{t+\Delta t} = \mathbf{q}_t + \frac{\partial H(\mathbf{q}_t, \mathbf{p}_t)}{\partial \mathbf{p}_t} \right\}. \tag{40}$$

Since the Hamiltonian is a C^2 class function, S_i^a is a differentiable manifold. The canonical equation of motion is a C^1 map because the Hamiltonian is a C^2 class function. The subspace S_i^b is a subspace mapped from manifold S_i^a according to the canonical equations of motion. Therefore, the subspace S_i^b is also a differentiable manifold and S_i is the product of differentiable manifolds S_i^a and S_i^b. From a property of product manifold, S_i is understood as a differentiable manifold. Interpolation of differentiable manifolds can be realized by machine learning methods such as deep learning. In our proposed framework, S_i is estimated from a finite number of data D using a deep-learning technique. The second principle is described as follows. In a canonical dynamical system in which the energy changes with time, it is not efficient to acquire the data of S_i because S_i is a subspace of specific energy. The important cases of a complex dynamical system to be modeled as a reduced model are in the stable or metastable state. Also, one of the final goals of this study is to extract the conservation laws in a large-scale collective motion system in a metastable state. In the stable or metastable state, the energy of the system is conserved: $H(\mathbf{q}_t, \mathbf{p}_t) = H(\mathbf{q}_{t+\Delta t}, \mathbf{p}_{t+\Delta t}) = E$. Therefore, for the purpose of this study, efficient data acquisition is realized.

E. DNN and data manifold

As mentioned in Sec. II D, the subspace S_i could be modeled as a differentiable manifold using machine learning models. Some well-trained DNNs can model the distribution of a training data set as a differentiable manifold [14–19]. In this paper, we refer to such a differentiable manifold as a data manifold.

We explain how a DNN models a d_m-dimensional manifold in d_{in}-dimensional space $\mathbf{x}$ using one of the simplest DNNs: a feed forward three-layer DNN, for which the input has d_{in} dimensions, the hidden layer has $d_{\text{h}}(> d_{\text{in}})$ dimensions, and the output has $d_{\text{out}}(< d_{\text{in}}) = d_m$ dimensions. The mapping function $\mathbf{f}_{\text{DNN}}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_{d_{\text{out}}}(\mathbf{x})]$ of the DNN is defined as $\mathbf{f}_{\text{DNN}}(\mathbf{x}) = \mathbf{w}^h\mathbf{h} = \mathbf{w}^h\boldsymbol{\varphi}(\mathbf{w}^{\text{in}}\mathbf{x})$, where $\mathbf{h} = (h_1, h_2, \cdots, h_{d_h})$ is the d_h-dimensional output of the hidden layer. We define $\boldsymbol{\varphi}(\cdot)$ as $\boldsymbol{\varphi}(\mathbf{w}^{\text{in}}\mathbf{x}) = (\varphi_1, \varphi_2, \cdots, \varphi_{d_h})$, $\varphi_j = \varphi[\sum_i^{d_{\text{in}}} (w_{ij}^{\text{in}} x_i)]$, where φ is the activation function.

The element of the output of the hidden layer $h_j = w_{1j}x_1 + w_{2j}x_2 + \cdots + w_{d_{\text{in}}j}x_{d_{\text{in}}}$ is understood to be a projection of vector $(x_1, x_2, x_3, \cdots, x_d)$ to vector $(w_{1j}, w_{2j}, \cdots, w_{d_{\text{in}}}j)$. In addition, activation function $\boldsymbol{\varphi}(\cdot)$ is usually set as a sigmoid or ReLU function. These activation functions are constructed using linear and flat domains. Therefore, φ_j maps the input subspace along vector $(w_{0j}, w_{1j} \cdots, w_{d_{\text{in}}j})$ to hidden space, and the region of the subspace is related to the linear domain of the activation function. Using some elements of $\boldsymbol{\varphi}(\mathbf{w}^{\text{in}}\mathbf{x})$, we can represent a function that maps a d_{out}-dimensional subhyperplane in the input space to a d_{out}-dimensional subhyperplane in hidden space. By continuously pasting these subhyperplanes using second layer $\mathbf{w}^h$ as if they were the tangent spaces of a data manifold, the DNN can model the data distribution as a d_{out}-dimensional manifold. That is, the DNN embeds the input space in the output space by pasting the sub-hyperplanes and compresses the tangent direction of these sub-hyperplanes (Fig. 1). Deeper and more complex DNNs can be understood as a collection of such three-layer DNNs. Thus, such deeper DNNs can model more complex manifold structures as a combination of simple manifold structures modeled by a three-layer DNN [17]. Note that the output of a three-layer DNN, a part of the deeper DNN, is referred to as a hidden layer. This is only one example of how a DNN models a data manifold. However, many studies have suggested that there are similar properties in successfully trained DNNs

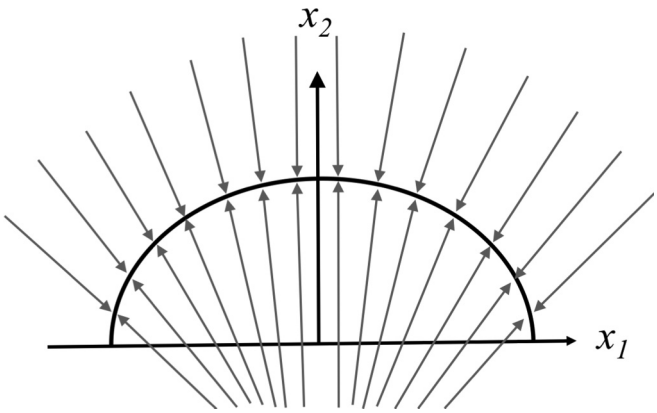

FIG. 1. Schematic diagram of the mapping structure of a DNN with $d_{\text{in}} = 2$ and $d_{\text{out}} = 1$. The DNN is trained to map two-dimensional data distributed on a black curve to one-dimensional output space. The arrows indicate the compression direction of the input space in the mapping from the input to the output.

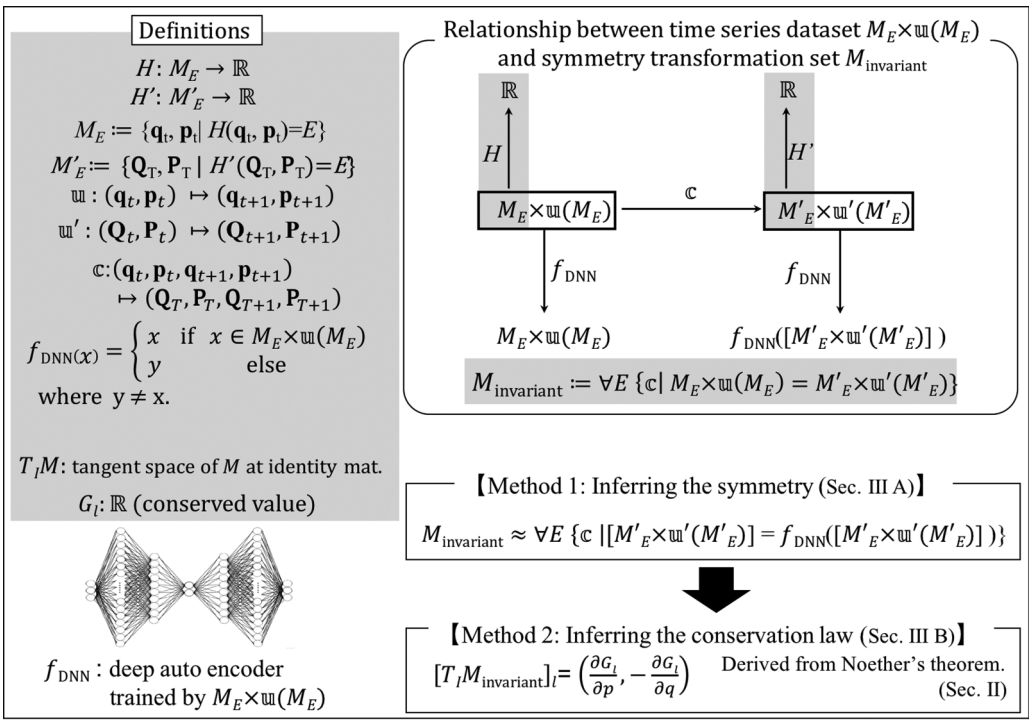

FIG. 2. Schematic diagram of the proposed framework.

[14–19]. By replacing the input space from $\mathbf{x}$ to $\Gamma \times \mathbb{R} \times \mathbb{R}$, we can also model a time-series data manifold S_i using DNN.

In this paper, using a trained DNN that models a time-series data manifold S_i, we propose a method of extracting information about the symmetry of a dynamical system. As described later in Sec. V, our proposed framework does not require special DNNs, so we can directly utilize the vast knowledge obtained from studies on physical data analysis using DNNs. This is why we select the DNN from multiple machine-learning models that can be used to model manifolds.

III. METHOD

In this section, we describe our proposed framework for estimating the conservation law from a time-series data set of dynamics. The schematic diagram of the proposed framework is shown in Fig. 2. The framework consists of two methods. In Sec. III A, on the basis of the derivation of the relationship between the symmetry of the time-series data-set distribution and the conservation law (Sec. II), we propose a method of inferring the symmetry of data manifold using the Monte Carlo sampling method. In Sec. III B, we describe the proposed method of inferring the conservation law from the obtained symmetry.

A. Method 1: Inferring the symmetry of data manifold using Monte Carlo sampling method

In this subsection, we propose a general method of inferring the symmetric property of data manifolds, which is not limited to a physical time-series data set. It can be inferred from the discussion in Sec. II E that data points that are not on the manifold in the input space are attracted to the manifold (Fig. 1). Once the data points are attracted to the manifold in the hidden layer, they continue to exist on the manifold in the output $\mathbf{f}(\mathbf{x})$. We propose a method based on this property of DNNs for extracting the symmetry of the data manifold using a deep autoencoder [15]. The deep autoencoder is a model that compresses the input space to a low-dimensional hidden layer and decompresses the layer to an output space with the same dimension as the input space. In the decompression process, only the subspace of the input space around the data manifold is recovered because of the DNN property. On the basis of this property, we can evaluate whether a transformation $\mathbf{X}(\cdot)$ causes the data-set distribution $\{\mathbf{x}_i\}_{i=1}^N$ to remain in the same subspace of the data manifold (Fig. 3). The procedure is as follows. First, we train the deep autoencoder using $\{\mathbf{x}_i\}_{i=1}^N$ as a training data set. Second, we input the transformed data set $\{\mathbf{X}(\mathbf{x}_i)\}_{i=1}^N$ into the trained deep autoencoder. Note that the deep autoencoder is not trained on the transformed data set. Third, we evaluate the transformation $\mathbf{X}(\cdot)$ using the mean-squared error between the input distribution of the data set and its mapped distribution:

$$E_{\mathrm{samp}}[\mathbf{X}(\cdot)] = \frac{1}{N}\sum_{i=1}^{N}\{\mathbf{X}(\mathbf{x}_i) - \mathbf{f}_{\mathrm{DNN}}[\mathbf{X}(\mathbf{x}_i)]\}^2. \tag{41}$$

A smaller E_{samp} value implies that $\mathbf{X}(\cdot)$ is a more invariant transformation. Using the criterion E_{samp}, we approximate the invariant transformation set as

$$\{\mathbf{X}(\cdot) | \operatorname*{argmin}_{\mathbf{X}} E_{\mathrm{samp}}[\mathbf{X}(\cdot)]\}. \tag{42}$$

To infer the conservation law, it is necessary to estimate the invariant transformation set $M_{\mathrm{invariant}}$ of the manifold S_i. The

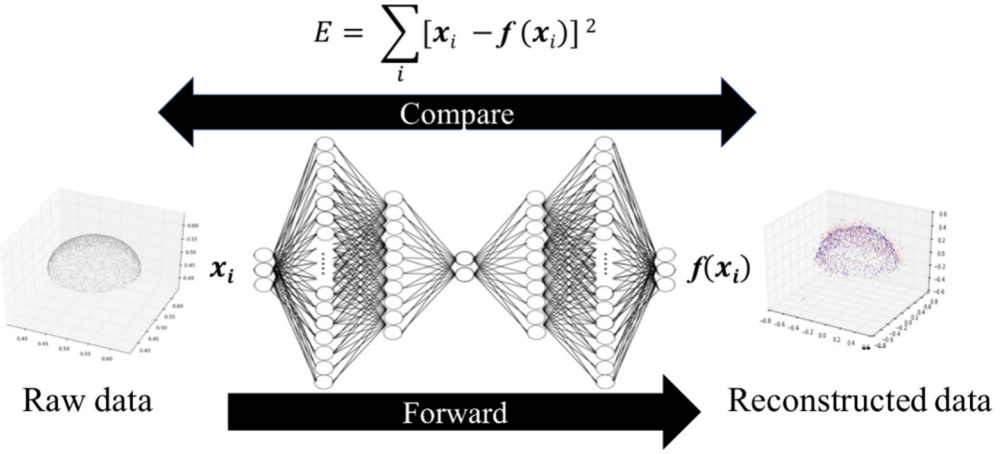

FIG. 3. Schematic diagram of method of extracting invariant transformation using autoencoder.

invariant transformation set $M_{\text{invariant}}$ is defined as

$$M_{\text{invariant}} := \left\{ Q^{S_i}(\cdot,\cdot,\boldsymbol{\theta}), \mathcal{P}^{S_i}(\cdot,\cdot,\boldsymbol{\theta}) \middle| \boldsymbol{\theta} \in \mathbb{R}^{d_\theta} \right\}. \tag{43}$$

In Eq. (41), by substituting $\{\mathbf{x}_i\}_{i=1}^N$ for $D = \{\mathbf{q}_{t_i}^i, \mathbf{p}_{t_i}^i, \mathbf{q}_{t_i+\Delta t}^i, \mathbf{p}_{t_i+\Delta t}^i\}_{i=1}^N$ and $\mathbf{X}(\cdot)$ for the transformation $\mathbb{C} : (\mathbf{Q}(\cdot,\cdot), \mathbf{P}(\cdot,\cdot))$, we can approximate $M_{\text{invariant}}$ as

$$M_{\text{invariant}} \sim \left\{ \mathbf{Q}(\cdot,\cdot), \mathbf{P}(\cdot,\cdot) \middle| \underset{\mathbf{Q}(\cdot,\cdot),\mathbf{P}(\cdot,\cdot)}{\operatorname{argmin}} E_{\text{samp}}[\mathbf{Q}(\cdot,\cdot), \mathbf{P}(\cdot,\cdot)] \right\}, \tag{44}$$

where the data set D is generated from dynamics data at energy E_i. The approximated invariant transformation set is obtained approximately by sampling from the probabilistic density:

$$\mathrm{P}(\mathbf{Q}(\cdot,\cdot), \mathbf{P}(\cdot,\cdot)) \sim \frac{1}{Z} \exp\left\{ -\frac{N}{2\sigma^2} E_{\text{samp}}[\mathbf{Q}(\cdot,\cdot), \mathbf{P}(\cdot,\cdot)] \right\}, \tag{45}$$

where σ is set as small as necessary and Z is a normalization constant. Note that to actually perform this sampling, it is necessary to first give a concrete coordinate system of $(\mathbf{q}_i, \mathbf{p}_i)$ in which physicists want to search conservation laws.

As mentioned in Sec. II A, continuous symmetries form a Lie group. If the computational cost is ignored and the transformation functions can be represented as parameterized functions, the proposed framework can be applied to a variety of conversion functions, including nonlinear conversions (Appendix I). In practice, it is important to narrow down the class of candidate transformation functions. In the proposed framework, we assume that this narrowing down is given by physicists on the basis of their physical insights. This is not necessarily a weakness of the proposed framework, but rather an advantage. By narrowing down the class of coordinate transformations and applying our methods with trial and error, the physicists would obtain clues of the nature of the physics system in the correspondence between the given candidate transformation functions and the estimated conservation laws. In this section, for simplicity of explanation, we restrict the class of the candidate transformation functions to affine transformations,

$$\begin{pmatrix} \mathbf{Q} \\ \mathbf{P} \end{pmatrix} \to A \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix} + A_0, \tag{46}$$

where A is a $2d \times 2d$ matrix and A_0 is a $2d$-dimensional vector. A more general discussion is provided in Appendix I. The invariant transformation is obtained by sampling an element a_{jk} of the matrix A and a_{0k} of the vector A_0 following the probability distribution

$$\begin{aligned} &\mathrm{P}(a_{11}, \cdots, a_{2d\,2d}, a_{01}, \cdots, a_{0\,2d}) \\ &\quad = \frac{1}{Z} \exp\left[-\frac{N}{2\sigma^2} E_{\text{samp}}(a_{11}, \cdots, a_{2d\,2d}, a_{01}, \cdots, a_{0\,2d}) \right] \\ &\qquad \times q(a_{11}, \cdots, a_{2d\,2d}, a_{01}, \cdots, a_{0\,2d}), \end{aligned} \tag{47}$$

where $q(\cdot)$ is a probability distribution that represents the constraints of transformation, resembling the prior distribution of Bayesian inference. For example, the general property of Hamiltonian systems, in which an infinitesimal volume in phase space is conserved under canonical transformations [37], is represented as a uniform distribution:

$$\begin{aligned} &q(a_{11}, \cdots, a_{2d\,2d}, a_{01}, \cdots, a_{0\,2d}) \\ &\quad = \begin{cases} \text{const.} & \text{for} \quad \det A = 1 \\ 0 & \text{for} \quad \det A \neq 1. \end{cases} \end{aligned} \tag{48}$$

Note that, in practice, we assume a uniform distribution with the range $1 - \delta < \det A < 1 + \delta$, where $1 > \delta > 0$, to account for data noise or training error of a DNN. As mentioned above, by utilizing all the knowledge to narrow down the candidate transformations, we can avoid expanding the search space and increasing the computational cost of extracting meaningless symmetries.

To perform this sampling, we need to specify σ. Ideally, σ should be set to 0. However, it is necessary to set σ to an appropriate finite value because errors are included in the time-series data set and the training results of a DNN. Such σ affected by noise cannot be set in advance. In addition, the target distributions in this study are assumed to be the global flat minima, because the same E_{samp} surface corresponding to the invariant transformation exists. Generally, such a target distribution needs an enormous amount of time to sample. Therefore, in this study, we use the replica-exchange Monte

Carlo (REMC) method [38] for sampling to overcome these problems. Such a method enables us to perform efficient sampling by parallel sampling with different noise intensities of σ while exchanging noise intensities with each other. In the state of a large noise, we can realize global sampling from the abstract distribution

$$\begin{aligned}&\mathrm{P}'(a_{11},\cdots,a_{2d\,2d},a_{01},\cdots,a_{0\,2d})\\&=\frac{1}{Z'}\exp\left[-\frac{N}{2\sigma'^2}E_{\mathrm{samp}}(a_{11},\cdots,a_{2d\,2d},a_{01},\cdots,a_{0\,2d})\right]\\&\quad\times q(a_{11},\cdots,a_{2d\,2d},a_{01},\cdots,a_{0\,2d}),\end{aligned}\tag{49}$$

where $\sigma' > \sigma$. By exchanging this sampling information with the state of a small noise, we can perform efficient sampling from the target distribution $P(a_{11},\cdots,a_{2d\,2d},a_{01},\cdots,a_{0\,2d})$. A detailed explanation of the REMC method and the setting parameters of this method are described in Appendix E. The target σ is determined by analyzing the sampling results on multiple σ. In the demonstration of the proposed framework, we set the target σ as described in Appendix F. The procedure of method 1 is summarized in Algorithm 1.

Algorithm 1 Estimation of the invariant transformation set

Input: Data set $D=\{\mathbf{q}^i_{t_i},\mathbf{p}^i_{t_i},\mathbf{q}^i_{t_i+\Delta t},\mathbf{p}^i_{t_i+\Delta t}\}_{i=1}^{N}$ in a given coordinate system.
Output:Invariant transformation set $D_a=\{(a_{11},a_{12}\cdots,a_{1d},a_{21}\cdots,a_{2d\,2d},a_{01}\cdots,a_{0\,2d})_{n_a}\}_{n_a=1}^{N_a}$.
Step 1: Train the deep autoencoder with data set D.
Step 2: Using the trained deep autoencoder and REMC method, sample transformation parameters $a_{11},\cdots,a_{2d\,2d},a_{01}\cdots,a_{0\,2d}$ from multiple probability distributions $P'(a_{11},\cdots,a_{2d\,2d},a_{01}\cdots,a_{0\,2d})$ corresponding to different noise intensities σ'.
Step 3: Select σ' from the distribution structure of the sampling results and output the sampling result of the selected σ' state as D_a.

Note that there is no description of how to train a DNN in this paper. In the training of a deep autoencoder, the number of nodes in the hidden layer is an important hyperparameter. On the other hand, since this is a quantity that determines how much a phenomenon is to be reduced, it is considered to be provided by the physicist.

B. Method 2: Inferring the conservation law from obtained symmetry

From the N_a sampling results $D_a \coloneqq \{(a_{11},a_{12}\cdots,a_{1\,2d},a_{21}\cdots,a_{2d\,2d},a_{01},\cdots,a_{0\,2d})_{n_a}\}_{n_a=1}^{N_a}$ in Sec. III A, we propose a method of estimating the infinitesimal transformation, which represents the invariance of the Hamiltonian and the equation of motion.

The set of invariant transformation $M_{\mathrm{invariant}}$ is characterized by the d_θ-dimensional continuous parameter $\boldsymbol{\theta}$. Therefore, $M_{\mathrm{invariant}}$ is a d_θ-dimensional differential manifold. Note that $M_{\mathrm{invariant}}$ forms a Lie group as we mentioned in Sec. II A. The infinitesimal transformation is estimated as the tangent vector of $M_{\mathrm{invariant}}$ at $\boldsymbol{\theta}=\mathbf{0}$. Using $A(\boldsymbol{\theta})$, we estimate $M_{\mathrm{invariant}}$ as

$$M_{\mathrm{invariant}} \sim \left\{A(\boldsymbol{\theta})\begin{pmatrix}\mathbf{q}\\\mathbf{p}\end{pmatrix}\middle|\ \boldsymbol{\theta}\in\mathbb{R}^{d_\theta}\right\}.\tag{50}$$

By serializing the transformation matrix $A(\boldsymbol{\theta})$, we define the vector

$$\begin{aligned}A'(\boldsymbol{\theta})=(a'_1(\boldsymbol{\theta}),\cdots,a'_{d'}(\boldsymbol{\theta}))\coloneqq(a_{11}(\boldsymbol{\theta}),\cdots,a_{1\,2d}(\boldsymbol{\theta}),a_{21}(\boldsymbol{\theta}),\\\times\cdots,a_{2d\,1}(\boldsymbol{\theta}),\cdots,a_{2d\,2d}(\boldsymbol{\theta}),a_{01}(\boldsymbol{\theta}),\cdots,a_{0\,2d}(\boldsymbol{\theta})),\end{aligned}\tag{51}$$

where $d'=4d^2+2d$. The implicit function representation of the manifold $M_{\mathrm{invariant}}$ is defined as

$$\begin{cases}f_1(a'_1,\cdots,a'_{d'})=0\\\quad\vdots\\f_{d'-d_\theta}(a'_1,\cdots,a'_{d'})=0.\end{cases}\tag{52}$$

In the representation of the implicit function, the infinitesimal transformation is estimated as the tangent vector of the manifold $M_{\mathrm{invariant}}$ at the position

$$e_{\mathbf{I}}=(\cdots,a_{ij}=0,\cdots,a_{i\,i}=1,\cdots),\tag{53}$$

where $i\neq j$ and $e_{\mathbf{I}}$ is the representation of the identity matrix $\mathbf{I}$ in the $A'(\boldsymbol{\theta})$ space. We estimate this tangent space $\mathrm{T}_{\mathbf{I}}M_{\mathrm{invariant}}=\mathrm{T}_{e_{\mathbf{I}}}M_{\mathrm{invariant}}$ from the sampling result D_a obtained in Sec. III A.

The Jacobian matrix of f_k for parameters of the subset A', $(b_1,b_2,\cdots,b_{d_\theta})\subset A'$, is defined as $J_{kl}=\frac{\partial f_k(a'_1,\cdots,a'_{d'})}{\partial b_l}$. If the Jacobian matrix at $A'=e_{\mathbf{I}}$ becomes nonsingular, from the implicit function theorem, variables other than $(b_1,b_2,\cdots,b_{d_\theta})$, $\{c_k\}_{k=1}^{d'-d_\theta}\coloneqq A'\setminus\{b_l\}_{l=1}^{d_\theta}$, can be expressed as $c_k=g_i(b_1,\cdots,b_{d_\theta})$. This implies that, around $e_{\mathbf{I}}$, the implicit equations in Eq. (52) representing the manifold $M_{\mathrm{invariant}}$ can be decomposed into the following $d'-d_\theta$ simultaneous equations:

$$\begin{cases}h_1(c_1,b_1,\cdots,b_{d_\theta})=0\\\quad\vdots\\h_{d'-d_\theta}(c_{d'-d_\theta},b_1,\cdots,b_{d_\theta})=0,\end{cases}\tag{54}$$

where b_l corresponds to the continuous parameter θ_l of the continuous transformation $[Q(\mathbf{q},\mathbf{p},\boldsymbol{\theta}),\mathcal{P}(\mathbf{q},\mathbf{p},\boldsymbol{\theta})]$. Differentiating these equations with respect to b_l around a point $e_{\mathbf{I}}$ yields $d'-d_\theta$ simultaneous partial differential equations:

$$\begin{cases}\frac{\partial}{\partial b_l}h_1(c_1,b_1,\cdots,b_{d_\theta})\big|_{A'=e_{\mathbf{I}}}=0\\\quad\vdots\\\frac{\partial}{\partial b_l}h_{d'-d_\theta}(c_{d'-d_\theta},b_1,\cdots,b_{d_\theta})\big|_{A'=e_{\mathbf{I}}}=0.\end{cases}\tag{55}$$

Solving these simultaneous partial differential equations gives the tangent vector $\frac{A'(b_l)}{\partial b_l}\big|_{A'=e_{\mathbf{I}}}$ of the manifold around $e_{\mathbf{I}}$. Using the tangent vector as the nonserialized representation

$\frac{A(b_l)}{\partial b_l}|_{A'=e_{\mathbf{I}}}$, we can estimate an infinitesimal transformation as

$$\begin{pmatrix}\delta\mathbf{q}_l\\ \delta\mathbf{p}_l\end{pmatrix} = \varepsilon \left.\frac{A(b_l)}{\partial b_l}\right|_{A'=e_{\mathbf{I}}}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix} = \varepsilon\begin{pmatrix}\frac{\partial a_{1\,1}}{\partial b_l}|_{A'=e_{\mathbf{I}}} & \cdots & \frac{\partial a_{2d\,1}}{\partial b_l}|_{A'=e_{\mathbf{I}}}\\ \vdots & \ddots & \vdots\\ \frac{\partial a_{1\,2d}}{\partial b_l}|_{A'=e_{\mathbf{I}}} & \cdots & \frac{\partial a_{2d\,2d}}{\partial b_l}|_{A'=e_{\mathbf{I}}}\end{pmatrix}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix}. \quad (56)$$

Thus, the invariant transformation is obtained as the tangent vector of the manifold $M_{\text{invariant}}$ at point $e_{\mathbf{I}}$. Therefore, if c_k can be regressed around $e_{\mathbf{I}}$ as the first-order polynomial of $\{b_l\}_{l=1}^{d_\theta}$, the conservation law can be inferred without approximation. Compared with the Hamiltonian estimation and conservation law estimation, this is the advantage of conservation law estimation because, in general, the Hamiltonian estimation requires infinite-order polynomial approximation. This advantage stands not only for the linear transformations used in the description but also for the general case involving nonlinear transformations [Eq. (I6) in Appendix I]. On the other hand, the estimation accuracy of the tangent space $\mathrm{T}_{e_{\mathbf{I}}}M_{\text{invariant}}$ from finite data with noise is often low. In this paper, we propose a method of estimating the infinitesimal transform with high accuracy by using all sampled transformation data, not only data around $e_{\mathbf{I}}$. Another way to avoid this problem is also discussed in Sec. V.

The simultaneous equations in Eq. (54) can be estimated by the following procedure. First, the upper limit of the dimension of the manifold $M_{\text{invariant}}$ is estimated by applying principal component analysis (PCA) and the elbow method to D_a as described in Ref. [39]. Alternatively, the approximate dimension of $M_{\text{invariant}}$ can be estimated by using the manifold dimension estimation method such as the method described in Ref. [40]. Using such an estimated dimension of $M_{\text{invariant}}$, we can prepare candidate dimension d'_θ. Second, we extract one variable set $(b_1, b_2, \cdots, b_{d'_\theta})$. By orthogonal distance regression [41], we regress $D_b \equiv \{(c_k, b_1, b_2, \cdots, b_{d'_\theta})_{n_a}\}_{n_a=1}^{N_a}$ to a d_b-order implicit polynomial function,

$$\begin{aligned}&\hat{h}_k\left(c_k, b_1, b_2, \cdots, b_{d'_\theta}; \beta, \gamma, d'_\theta\right)\\ &\coloneqq \sum_{s_0=0}^{d_b}\sum_{s_1=0}^{d_b}\cdots\sum_{s_{d'_\theta}=0}^{d_b}\gamma_{s_0 s_1 s_2\cdots s_{d'_\theta}}\beta_{s_0 s_1 s_2\cdots s_{d'_\theta}}c_k^{s_0}b_1^{s_1}b_2^{s_2}\cdots b_{d'_\theta}^{s_{d'_\theta}}\\ &= 0, \end{aligned} \quad (57)$$

where β is the regression coefficient, and γ is a binary vector indicating whether the basis is selected. The indicator vector γ and the dimension of the manifold d'_θ are determined by a model selection method, such as the Bayesian information criterion (BIC) [42]. To select the model, it is necessary to estimate the likelihood. The method of estimating the likelihood is described in Appendix G. If $d_\theta \leqslant 2$, d'_θ can be determined by visualization. Note that, unlike the estimation of the tangent space $\mathrm{T}_{e_{\mathbf{I}}}M_{\text{invariant}}$, the upper limit d_b of the order of polynomial function must be sufficiently large because all the sampling data are regressed. This regression and model selection is performed for all c_k; then, an implicit function representation of $M_{\text{invariant}}$ can be obtained.

From the obtained simultaneous equations, we obtain the simultaneous differential equations. If the Jacobian matrix J_{kl} is singular, the solution of the simultaneous equations diverges or becomes indefinite. In that case, a different variable set $\{(b_1, \cdots, b_{d'_\theta})\}$ is reextracted and the same procedure is repeated for the new variable set $\{(b_1, \cdots, b_{d'_\theta})\}$. If the Jacobian matrix J_{kl} becomes nonsingular after applying this procedure repeatedly, we can obtain the infinitesimal transformation according to Eq. (56). In this method, by narrowing down the regressing area of D_a to the neighborhood of $e_{\mathbf{I}}$, we obtain a more accurate estimation of infinitesimal transformation with a lower-order polynomial function in Eq. (57).

Algorithm 2 Estimation of infinitesimal transformation

Input: Sampling results of method 1, $D_a = \{(a_{11}, a_{12}\cdots, a_{2d\,2d}, a_{01}, \cdots, a_{0\,2d})_{n_a}\}_{n_a=1}^{N_a}$, and d'_θ.
Output:Infinitesimal transformation, $\delta\mathbf{q}_l$, $\delta\mathbf{p}_l$.
Step 1: Extract $D_b = \{(c_k, b_1, b_2, \cdots, b_{d'_\theta})_{n_a}\}_{n_a=1}^{N_a}$ from D_a.
Step 2: Fit D_b with the implicit polynomial function $\hat{h}_k(c_k, b'_1, b'_2, \cdots, b'_{d'_\theta}; \beta, \gamma, d'_\theta)$ [Eq. (57)] for each c_k.
Step 3: Estimate the likelihood [Eq. (G1)] by numerical integration of Z [Eq. (G2)].
Step 4: Select the indicator vector γ and the dimension d'_θ of $M_{\text{invariant}}$ in Eq. (57) for each c_k using the BIC.
Step 5: Determine whether the Jacobi matrix $J_{kl} = \frac{\partial h_k(c_k, b_1, \cdots, b_{d_\theta})}{\partial b_l}$ is nonsingular. If J_{kl} is singular, return to Step 1 and reextract D'_b.
Step 6: Differentiate the obtained simultaneous equations with respect to b_l around a point $e_{\mathbf{I}}$ to obtain Eq. (55).
Step 7: Solve the simultaneous equations in Eq. (55) and obtain the infinitesimal transformation, $\delta\mathbf{q}_l$, $\delta\mathbf{p}_l$.

IV. RESULTS

We evaluate the proposed method using one geometrical structure and three physical systems: (i) a half sphere, (ii) constant-velocity linear motion, (iii) a two-dimensional central force system, and (iv) a collective motion system. Case (i) has a rotational symmetry. In case (i), we confirm that method 1 can obtain a set of transformations corresponding to the symmetry. Cases (ii) and (iii) are systems that conserve the momentum and angular momentum, respectively. Using these cases, we verified method 2. Finally, we apply both proposed methods to (iv), which is a complicated collective motion system, and attempted to infer the collective coordinate and conservation law. In each case, the parameters of the DNN are set as described in Appendix H and REMC is set as described in Appendix E. In these demonstrations, we estimated invariant coordinate transformations by restricting them on the affine transformations. Since the affine transformation is

one of the simplest coordinate transformations, it would be a reasonable first assumption. Also, assuming that there are no complex symmetries due to the combination of temporal and spatial transformations, we also ignored the time transformation in these demonstrations. Note that we do not deny the existence of time translational symmetry which related to energy conservation because it can exist independently of the spatial transformation. Also, we further narrowed down the coordinate transformations using a property of time-series data distributions, as explained using cases (ii), (iii), and (vi). Moreover, we carried out PCA to estimate the dimension of manifold corresponding to invariant transformation.

(i) Half sphere

The data set of case (i) was generated by the function

$$x_1^2 + x_2^2 + x_3^2 = r, \quad (x_3 > 0), \tag{58}$$

where r was set to be 0.25. We generated 1671 samples according to Eq. (58). The data set of case (i) [shown in Fig. 4(a)] was used to verify the ability of method 1 described in Sec. III A to extract the symmetry. We set the coordinate system as (x_1, x_2, x_3) and limit the transformation on the x_1-x_2 plane. In such a case, the affine transformation is defined as

$$A\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + A_0 := \begin{pmatrix} a_{11} & a_{21} & 0 \\ a_{12} & a_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} a_{01} \\ a_{02} \\ 0 \end{pmatrix}, \tag{59}$$

where A and A_0 have constraints $q(A, A_0)$ in (47), which are described as

$$q(A, A_0) = \begin{cases} \text{const.} & \text{for} \quad 0.8 < \det A < 1.2, \quad \text{and} \\ -0.2 < a_{0j} < 0.2, & j = 1, 2 \\ 0 & \text{for} \quad \text{else.} \end{cases} \tag{60}$$

In this coordinate system, the half sphere has a rotation symmetry and a mirror symmetry. The rotation symmetry transformation is represented as

$$A_{\text{rot}}(\theta_{\text{rot}}) = \begin{pmatrix} \cos(\theta_{\text{rot}}) & \sin(\theta_{\text{rot}}) \\ -\sin(\theta_{\text{rot}}) & \cos(\theta_{\text{rot}}) \end{pmatrix}, \tag{61}$$

where θ_{rot} is a rotation angle and the mirror symmetry transformation is represented as

$$A_{\text{mirror}}(\theta_{\text{mirror}}) = \begin{pmatrix} \cos(2\theta_{\text{mirror}}) & \sin(2\theta_{\text{mirror}}) \\ \sin(2\theta_{\text{mirror}}) & -\cos(2\theta_{\text{mirror}}) \end{pmatrix}, \tag{62}$$

where θ_{mirror} is an angle of the mirror plane with the x_1 axis. The mirror symmetry is a discrete symmetry; therefore, the invariant transformation of the half sphere is represented as $A_{\text{tot}}(\theta_{\text{rot}}, \theta_{\text{mirror}}) := A_{\text{rot}}(\theta_{\text{rot}})[A_{\text{mirror}}(\theta_{\text{mirror}})]^m$, where $m := \{0, 1\}$ and

$$A_{\text{rot}}(\theta_{\text{rot}})[A_{\text{mirror}}(\theta_{\text{mirror}})]^0 := A_{\text{rot}}(\theta_{\text{rot}}), \tag{63}$$

$$\begin{aligned} &A_{\text{rot}}(\theta_{\text{rot}})[A_{\text{mirror}}(\theta_{\text{mirror}})]^1 \\ &:= \begin{pmatrix} \cos(2\theta_{\text{mirror}} - \theta_{\text{rot}}) & \sin(2\theta_{\text{mirror}} - \theta_{\text{rot}}) \\ \sin(2\theta_{\text{mirror}} - \theta_{\text{rot}}) & -\cos(2\theta_{\text{mirror}} - \theta_{\text{rot}}) \end{pmatrix} \\ &= A_{\text{mirror}}(\theta'), \end{aligned} \tag{64}$$

$$\theta' := \theta_{\text{mirror}} - \frac{\theta_{\text{rot}}}{2}. \tag{65}$$

By comparing Eq. (59) with Eqs. (63) and (64), we obtain the implicit function representation of the invariant transformation $A_{\text{tot}}(\theta_{\text{rot}}, \theta_{\text{mirror}})$ as

$$\begin{cases} a_{11}^2 + a_{21}^2 = 1 \\ a_{11}^2 + a_{12}^2 = 1 \\ (a_{11} + a_{22})(a_{11} - a_{22}) = a_{11}^2 - a_{22}^2 = 0 \\ (a_{21} - a_{12})(a_{21} + a_{12}) = a_{21}^2 - a_{12}^2 = 0 \\ a_{21}^2 + a_{22}^2 = 1 \\ a_{12}^2 + a_{22}^2 = 1. \end{cases} \tag{66}$$

Method 1 was applied to such a D_a system.

The sampling results of a_{ij} are shown in Fig. 4(b) as black dots. The results of PCA of the sampling result D_a are shown in Figs. 4(c) and 4(d). From the eigenvalues obtained by PCA, D_a is understood to be embedded in a four-dimensional space. As can be seen from the appearance of the sampling distribution [Fig. 4(b)], this is the result of embedding two intersecting circular manifolds corresponding to O(2). The principal component vectors with their eigenvalues greater than zero have zero values corresponding to the elements of a_{01} and a_{02}. This means that the distribution D_a is not spread out in the space corresponding to a_{01} and a_{02}. This is consistent with the fact that the half sphere has no translational symmetry, i.e., $a_{01} = \text{const}$ and $a_{02} = \text{const}$. We confirmed this observation by applying method 2 for the sampling results of a_{01} and a_{02} with polynomials of order up to first.

We apply method 2 for the sampling result D_a, and selected the best implicit polynomial functions from the viewpoint of the BIC. For the transformation elements of A, we regressed with polynomials of up to second order and selected a polynomial model. The red curves in Fig. 4(b) are curves fitted by the selected implicit polynomial functions. The fitting results are

$$\begin{cases} 1.00\, a_{11}^2 + 1.01\, a_{21}^2 = 1.00 \\ 1.00\, a_{11}^2 + 0.98\, a_{12}^2 = 1.00 \\ 1.00\, a_{11}^2 - 1.00\, a_{22}^2 = 0 \\ 1.00\, a_{01} = -0.01 \quad (\text{on } a_{11}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = -0.01 \quad (\text{on } a_{11}\text{–}a_{02} \text{ plane}) \\ 1.00\, a_{21}^2 - 0.97\, a_{12}^2 = 0 \\ 1.00\, a_{21}^2 + 0.99\, a_{22}^2 = 1.00 \\ 1.00\, a_{01} = -0.01 \quad (\text{on } a_{21}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = -0.01 \quad (\text{on } a_{21}\text{–}a_{02} \text{ plane}) \\ 0.99\, a_{12}^2 + 1.01\, a_{22}^2 + 0.01\, a_{12} - 0.01\, a_{22} = 1.00 \\ 1.00\, a_{01} = -0.01 \quad (\text{on } a_{12}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = -0.01 \quad (\text{on } a_{12}\text{–}a_{02} \text{ plane}) \\ 1.00\, a_{01} = -0.01 \quad (\text{on } a_{22}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = -0.01 \quad (\text{on } a_{22}\text{–}a_{02} \text{ plane}), \end{cases} \tag{67}$$

where we set d'_θ to be 1 on the basis of the knowledge derived from Eq. (66).

(ii) Constant-velocity linear motion

The data set of case (ii) was generated using the one-dimensional Hamiltonian system

$$H_2 = \frac{p^2}{2m}, \tag{68}$$

where m was set to be 1. We generated 1000 samples by solving Eq. (68). In this case, we show that the proposed

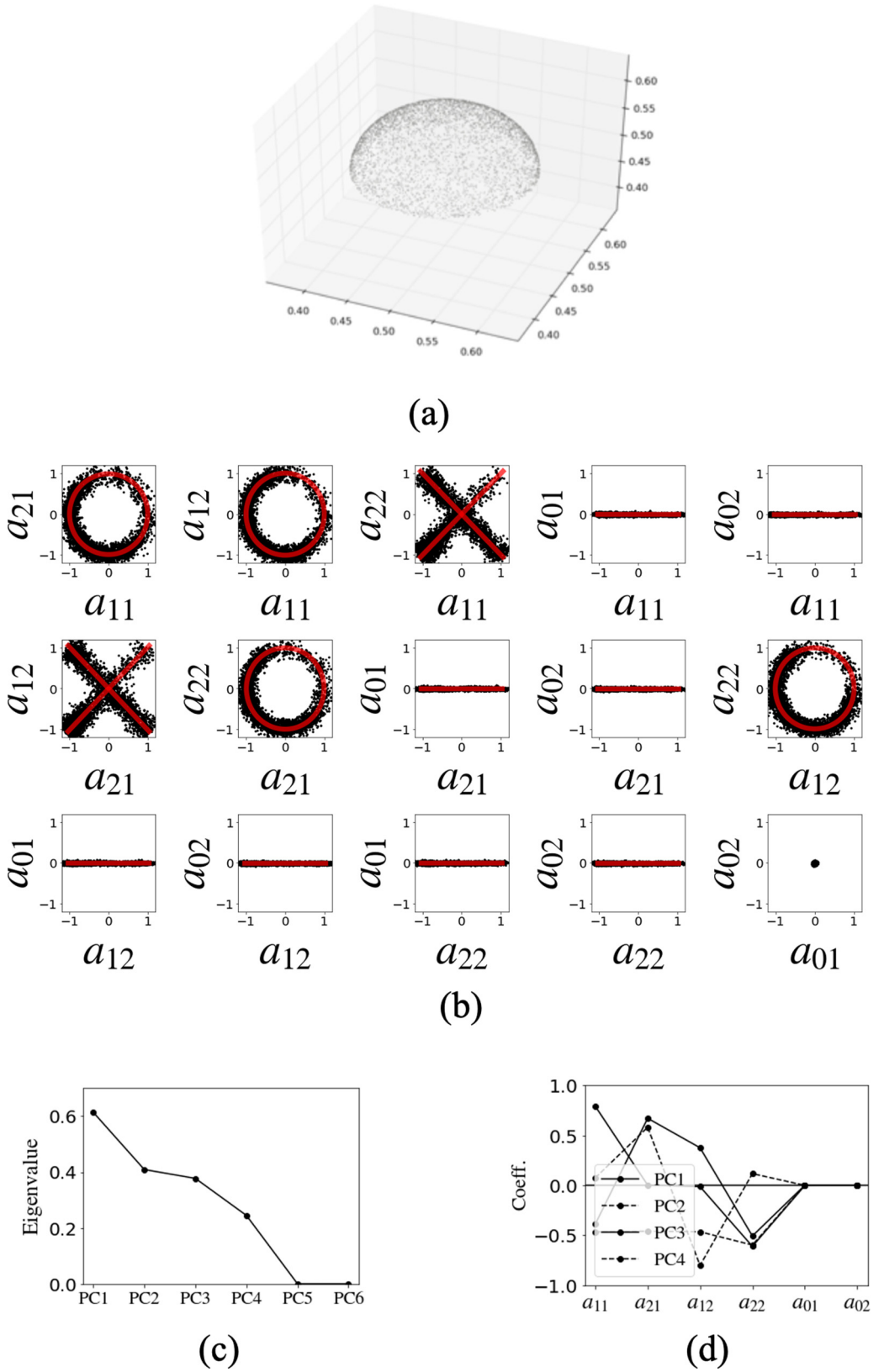

FIG. 4. Results of case (i): half sphere. (a) Data set used for the evaluation. There are 1671 samples. (b) Black dots represent sampling distributions D_a obtained by method 1 and red curves represent fitting curves estimated by method 2. Each graph shows 12 combinations of six transformation variables a_{ij}. (c) Eigenvalues of the principal components of the distribution D_a. (d) Principal component vectors of the distribution D_a with their eigenvalues greater than zero.

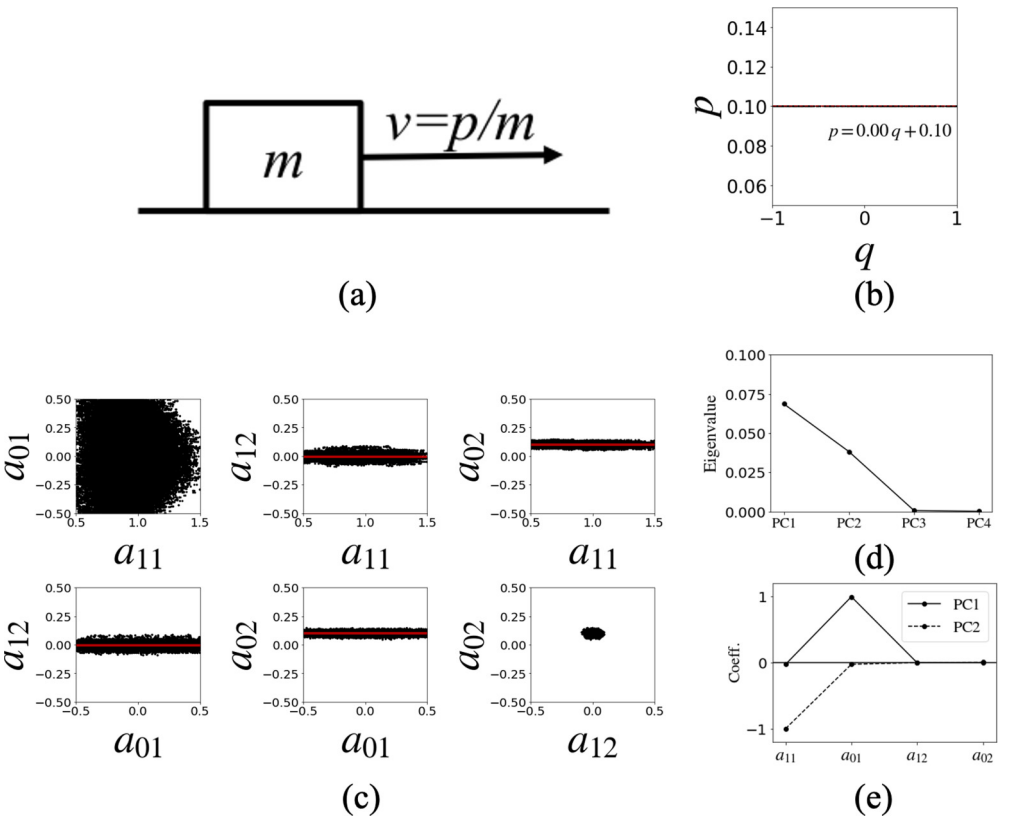

FIG. 5. Results of case (ii): Constant-velocity linear motion. (a) Conceptual diagram of constant-velocity linear motion. (b) Distribution of time-series data on q–p plane. (c) Black dots represent sampling distribution D_a obtained by method 1 and the red line represents the fitting curve estimated by method 2. (d) Eigenvalues of the principal components of the distribution D_a. (e) Principal component vectors of the distribution D_a with their eigenvalues greater than zero.

method can infer the momentum conservation law. We set the coordinate system as (q, p). In the coordinate, we found the linear relationship $p = 0.00q + 0.10$ from the distribution of time-series data [Fig. 5(b)]. On the basis of this relationship, we can reduce the affine transformation as

$$\begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} + \begin{pmatrix} a_{01} \\ a_{02} \end{pmatrix} \Leftrightarrow \begin{pmatrix} a_{11} & 0 \\ a_{12} & 0 \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} + \begin{pmatrix} a_{01} + 0.1a_{21} \\ a_{02} + 0.1a_{22} \end{pmatrix} \Leftrightarrow \begin{pmatrix} a_{11} & 0 \\ a_{12} & 0 \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} + \begin{pmatrix} a_{01} \\ a_{02} \end{pmatrix}. \quad (69)$$

Thus, we define the candidate transformation as

$$A\begin{pmatrix} q \\ p \end{pmatrix} + A_0 := \begin{pmatrix} a_{11} & 0 \\ a_{12} & 0 \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} + \begin{pmatrix} a_{01} \\ a_{02} \end{pmatrix}, \quad (70)$$

where A and A_0 have constraints $q(A, A_0)$ [see Eq. (47)] described as

$$q(A, A_0) = \begin{cases} \text{const.} & \text{for} \quad 0.5 < a_{11} < 1.5, \ \text{and} \\ -0.5 < a_{ij} < 0.5, & a_{ij} \in \{a_{12}, a_{01}, a_{02}\} \\ 0 & \text{for} \quad \text{else.} \end{cases} \quad (71)$$

The sampling results of D_a are shown in Fig. 5(c) as black dots. The eigenvalues of the principal components of D_a [Fig. 5(d)] suggest that the manifold of invariant transformation is embedded in a two-dimensional space. The two eigenvectors with eigenvalues much larger than others [Fig. 5(e)], corresponding to space embedding the manifold, do not have the components a_{12} and a_{02}. Considering that the distributions of a_{11} and a_{01} are uniformly spread out [Fig. 5(c)], the space consisting of a_{11} and a_{01} might be the manifold of the invariant transformation itself. This suggests that the sampling distribution D_a has two dimensions and that two conservation laws possibly exist. To confirm this, we set the dimension of the manifold d'_θ to be 2 and applied method 2 to the D_a for polynomial models up to first order in a three-dimensional space (a_{11}, a_{01}, a_{12}) and (a_{11}, a_{01}, a_{02}).

The regression results obtained with the selected implicit polynomial function using the BIC are plotted as red curves in Fig. 5(c). The regression result is

$$\begin{cases} a_{11} = a_{11} \\ a_{01} = a_{01} \\ a_{12} = -0.01 \quad (\text{on } a_{11}\text{–}a_{01}\text{–}a_{12} \text{ space}) \\ a_{02} = 0.10 \quad (\text{on } a_{11}\text{–}a_{01}\text{–}a_{02} \text{ space}). \end{cases} \quad (72)$$

The simultaneous partial differential equations in Eq. (55), where $b_l \in \{a_{11}, a_{01}\}$, were obtained from the fitting results. From the solution of the simultaneous partial differential equations, we obtained the two infinitesimal transformations corresponding to a_{11} or a_{01}

$$\begin{cases} \delta q = \epsilon \frac{\partial a_{11}}{\partial a_{11}} q + \epsilon \frac{\partial a_{01}}{\partial a_{11}} = \epsilon q \\ \delta p = \epsilon \frac{\partial a_{12}}{\partial a_{11}} p + \epsilon \frac{\partial a_{02}}{\partial a_{11}} = 0, \end{cases} \quad (73)$$

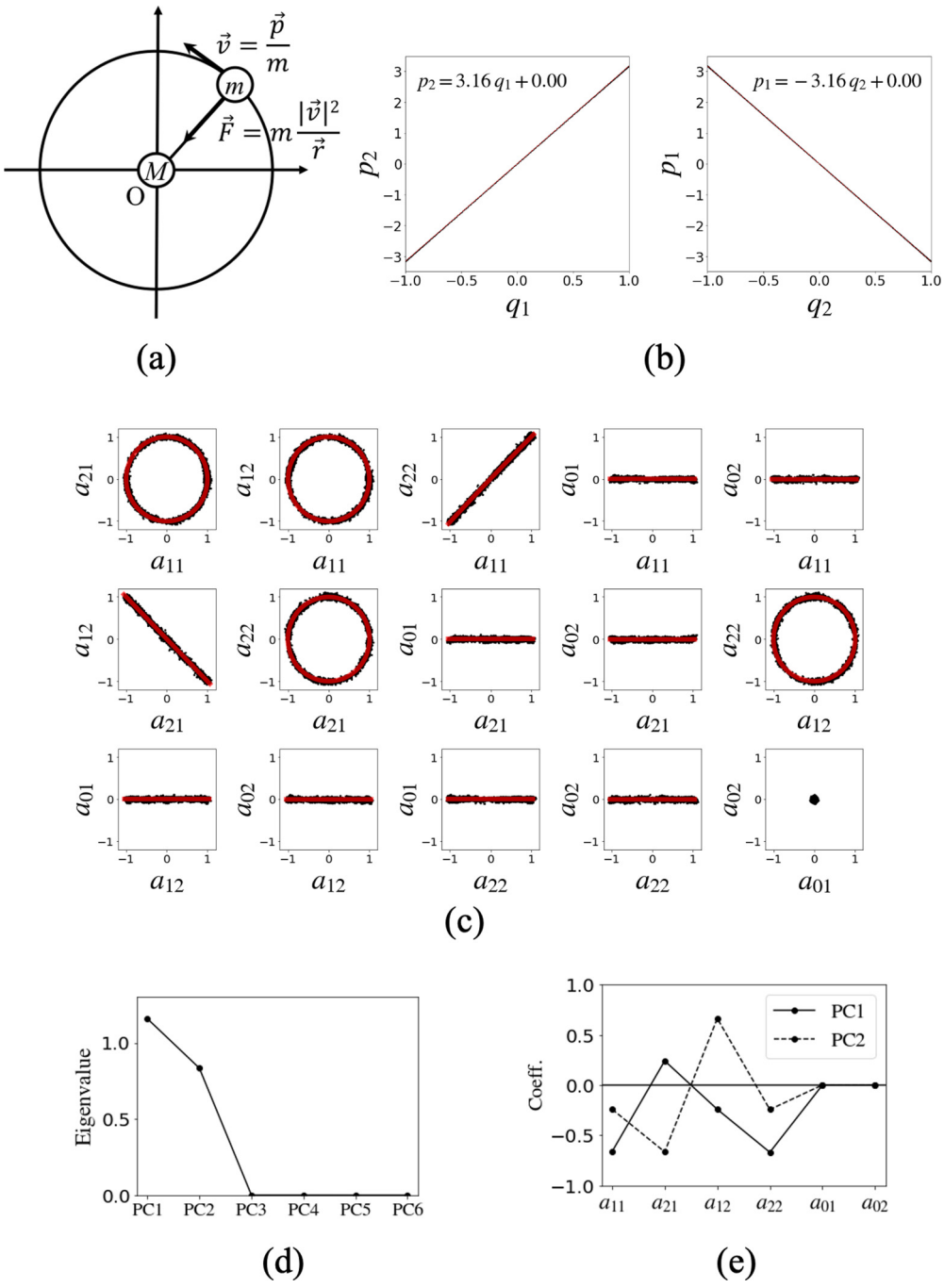

FIG. 6. Results of case (iii): Two-dimensional central force system. (a) Conceptual diagram of two-dimensional central force system. (b) Distributions of time-series data on q_1–p_2 and q_2–p_1 planes. (c) Black dots represent sampling distribution D_a obtained by method 1 and the red line represents the fitting curve estimated by method 2. (d) Eigenvalues of the principal components of the distribution D_a. (e) Principal component vectors of the distribution D_a with their eigenvalues greater than zero.

$$\begin{cases} \delta q = \epsilon \frac{\partial a_{11}}{\partial a_{01}} q + \epsilon \frac{\partial a_{01}}{\partial a_{01}} = \epsilon \\ \delta p = \epsilon \frac{\partial a_{12}}{\partial a_{01}} p + \epsilon \frac{\partial a_{02}}{\partial a_{01}} = 0. \end{cases} \tag{74}$$

By substituting Eqs. (73) and (74) into Eq. (1) and solving the simultaneous equations, we found that there is no solution for Eq. (73), whereas Eq. (74) gives us the conserved value $G_\delta = \varepsilon p$. This result shows that the momentum p was conserved.

(iii) Two-dimensional central force system

The data set of case (iii) was generated using the Hamiltonian system

$$H_3 = \frac{1}{2m}\mathbf{p}^2 + G\frac{mM}{|\mathbf{q}|}, \tag{75}$$

where $\mathbf{q} := (q_1, q_2)$, $\mathbf{p} := (p_1, p_2)$, and m, M, and G were set to be 1. In this study, the proposed framework is applied

to time-series data restricted to circular orbits to make the estimation problem of symmetries easier. The circular motion data manifold is clearly closed for the linear coordinate transformation of SO(2) corresponding to the angular momentum conservation law. Therefore, in case (iii), the effectiveness of the proposed framework can be verified by whether the angular momentum conservation law can be estimated. In addition, to simplify the problem, we excluded the symmetry of time inverse transition $t \to -t$ from the verification data, and we only focused on the counterclockwise motion. We generated 1000 samples by solving Eq. (75). We set the coordinate system as (q_1, q_2, p_1, p_2). In the coordinate, we found the linear relationships $p_2 = 3.16\, q_1 + 0.00$ and $p_1 = -3.16\, q_2 + 0.00$ from the distribution of time-series data [Fig. 6(b)]. From this relationship, the transformation is constrained on the two-dimensional plane corresponding to the normalized vectors $\frac{1}{\sqrt{1+3.16^2}}(1, 0, 0, 3.16)$ and $\frac{1}{\sqrt{1+3.16^2}}(0, 1, -3.16, 0)$. Thus, the affine transformation is reduced as

$$\frac{1}{1+3.16^2}\begin{pmatrix}1 & 0\\ 0 & 1\\ 0 & -3.16\\ 3.16 & 0\end{pmatrix}\begin{pmatrix}a_{11} & a_{21}\\ a_{12} & a_{22}\end{pmatrix}\begin{pmatrix}1 & 0 & 0 & 3.16\\ 0 & 1 & -3.16 & 0\end{pmatrix}\begin{pmatrix}q_1\\ q_2\\ p_1\\ p_2\end{pmatrix} + \frac{1}{\sqrt{1+3.16^2}}\begin{pmatrix}1 & 0\\ 0 & 1\\ 0 & -3.16\\ 3.16 & 0\end{pmatrix}\begin{pmatrix}a_{01}\\ a_{02}\end{pmatrix}, \tag{76}$$

$$= \frac{1}{1+3.16^2}\begin{pmatrix}a_{11} & a_{21} & -3.16a_{21} & 3.16a_{11}\\ a_{12} & a_{22} & -3.16a_{22} & 3.16a_{12}\\ -3.16a_{12} & -3.16a_{22} & 3.16^2a_{22} & -3.16^2a_{12}\\ 3.16a_{11} & 3.16a_{21} & -3.16^2a_{21} & 3.16^2a_{11}\end{pmatrix}\begin{pmatrix}q_1\\ q_2\\ p_1\\ p_2\end{pmatrix} + \frac{1}{\sqrt{1+3.16^2}}\begin{pmatrix}a_{01}\\ a_{02}\\ -3.16a_{02}\\ 3.16a_{01}\end{pmatrix}, \tag{77}$$

$$\Leftrightarrow \frac{1}{1+3.16^2}\begin{pmatrix}(1+3.16^2)a_{11} & (1+3.16^2)a_{21} & 0 & 0\\ (1+3.16^2)a_{12} & (1+3.16^2)a_{22} & 0 & 0\\ 0 & 0 & (1+3.16^2)a_{22} & -(1+3.16^2)a_{12}\\ 0 & 0 & -(1+3.16^2)a_{21} & (1+3.16^2)a_{11}\end{pmatrix}\begin{pmatrix}q_1\\ q_2\\ p_1\\ p_2\end{pmatrix} + \begin{pmatrix}a_{01}\\ a_{02}\\ -3.16a_{02}\\ 3.16a_{01}\end{pmatrix}. \tag{78}$$

Thus, we define the candidate transformation as

$$A\begin{pmatrix}q_1\\ q_2\\ p_1\\ p_2\end{pmatrix} + A_0 := \begin{pmatrix}a_{11} & a_{21} & 0 & 0\\ a_{12} & a_{22} & 0 & 0\\ 0 & 0 & a_{22} & -a_{12}\\ 0 & 0 & -a_{21} & a_{11}\end{pmatrix}\begin{pmatrix}q_1\\ q_2\\ p_1\\ p_2\end{pmatrix} + \begin{pmatrix}a_{01}\\ a_{02}\\ -3.16a_{02}\\ 3.16a_{01}\end{pmatrix}, \tag{79}$$

where A and A_0 have constraints $q(A, A_0)$ [see Eq. (47)], which are defined below in this demonstration:

$$q(A, A_0) = \begin{cases} \text{const} & \text{for} \quad 0.8 < \det A < 1.2, \ \text{and} \ -0.2 < a_{0j} < 0.2, \quad j = 1, 2 \\ 0 & \text{for} \quad \text{else.} \end{cases} \tag{80}$$

The sampling results of a_{ij} are shown in Fig. 6(c) as black dots. The results of PCA of the sampling result D_a are shown in Figs. 6(d) and 6(e). The eigenvalues obtained by PCA [Fig. 6(d)] suggested that D_a is embedded in a two-dimensional space. As can be seen from the appearance of the sampling distribution [Fig. 6(c)], this is the result of embedding one-dimensional circular manifolds. Thus, we set the dimension of the manifold d'_θ to be 1. Also, the principal component vectors with their eigenvalues greater than zero have zero values corresponding to the elements a_{01} and a_{02} [Fig. 6(e)]. This means that the distribution D_a is not spread out in the space corresponding to a_{01} and a_{02}. This suggests that there is no translational symmetry, i.e., $a_{01} = \text{const}$ and $a_{02} = \text{const}$. We verify this by applying method 2 for the sampling results of a_{01} and a_{02} with polynomials of up to first order and selecting a polynomial model from them. For the transformation elements of A, we regressed with polynomials of up to second order and selected a polynomial model.

The regression results obtained with the selected implicit polynomial function using the BIC are plotted as red curves in Fig. 6(c). The fitting results are

$$\begin{cases} 1.00\, a_{11}^2 + 1.00\, a_{21}^2 + 0.0\, a_{11}a_{21} = 1.00 \\ 1.00\, a_{11}^2 - 1.00\, a_{12}^2 = 1.00 \\ 1.00\, a_{11} - 0.99\, a_{22} - 0.02\, a_{22}^2 = -0.01 \\ 1.00\, a_{01} = 0.00 \quad (\text{on } a_{11}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = 0.00 \quad (\text{on } a_{11}\text{–}a_{02} \text{ plane}) \\ 1.00\, a_{21} + 1.00\, a_{12} = 0 \\ 1.00\, a_{21}^2 + 1.01\, a_{22}^2 = 1.01 \\ 1.00\, a_{01} = 0.00 \quad (\text{on } a_{21}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = 0.00 \quad (\text{on } a_{21}\text{–}a_{02} \text{ plane}) \\ 1.00\, a_{12}^2 + 1.01\, a_{22}^2 + 0.00\, a_{12}a_{22} = 1.01 \\ 1.00\, a_{01} = 0.00 \quad (\text{on } a_{12}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = 0.00 \quad (\text{on } a_{12}\text{–}a_{02} \text{ plane}) \\ 1.00\, a_{01} = 0.00 \quad (\text{on } a_{22}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = 0.00 \quad (\text{on } a_{22}\text{–}a_{02} \text{ plane}). \end{cases} \tag{81}$$

TABLE I. Parameters of the Reynolds boid model used to generate time-series data.

W_{att}	W_{sep}	W_{ali}	r_{att} [unit]	r_{sep} [unit]	r_{ali} [unit]	θ_{att} [rad]	θ_{sep} [rad]	θ_{ali} [rad]	speed [unit/s]
0.04	0.02	0.01	200	5	20	π	π	π	10–50

The simultaneous partial differential equations in Eq. (55), where $b_l = a_{21}$, were obtained from the fitting results. By solving the simultaneous partial differential equations, we obtained the infinitesimal transformation,

$$\begin{aligned}\delta\mathbf{q} &= \varepsilon\begin{pmatrix}\frac{\partial a_{11}}{\partial a_{21}} & \frac{\partial a_{21}}{\partial a_{21}}\\ \frac{\partial a_{12}}{\partial a_{21}} & \frac{\partial a_{22}}{\partial a_{21}}\end{pmatrix}\mathbf{q} + \varepsilon\begin{pmatrix}\frac{\partial a_{01}}{\partial a_{21}}\\ \frac{\partial a_{02}}{\partial a_{21}}\end{pmatrix}\\ &= \varepsilon\begin{pmatrix}\frac{-2\times a_{21}}{2a_{11}}\big|_{A'=e_{\mathbf{I}}} & 1\\ -1 & \frac{-2a_{21}}{1.01\times 2a_{22}}\big|_{A'=e_{\mathbf{I}}}\end{pmatrix}\mathbf{q} + \varepsilon\begin{pmatrix}0\\0\end{pmatrix}\\ &= \begin{pmatrix}0 & \varepsilon\\ -\varepsilon & 0\end{pmatrix}\mathbf{q}, \end{aligned} \tag{82}$$

$$\delta\mathbf{p} = \varepsilon\begin{pmatrix}\frac{\partial a_{22}}{\partial a_{21}} & -\frac{\partial a_{12}}{\partial a_{21}}\\ -\frac{\partial a_{21}}{\partial a_{21}} & \frac{\partial a_{11}}{\partial a_{21}}\end{pmatrix}\mathbf{p} + \varepsilon\begin{pmatrix}\frac{\partial a_{01}}{\partial a_{21}}\\ \frac{\partial a_{02}}{\partial a_{21}}\end{pmatrix} = \begin{pmatrix}0 & \varepsilon\\ -\varepsilon & 0\end{pmatrix}\mathbf{p}, \tag{83}$$

where the values in the final formula are up to one decimal place. By substituting Eqs. (82) and (83) into Eq. (1) and solving the equation, we estimated the conserved value G_δ as $G_\delta = \varepsilon(x_1 p_2 - x_2 p_1)$. This result shows that the angular momentum was conserved.

(iv) Collective motion system

In this case, we apply our framework to an N_{R}-body collective motion system called the Reynolds boid model [35]. In this model, each individual moves accordingly to three forces, which are the forces attracting each other, separating each other, and aligning the orientation of each other,

$$\frac{d\mathbf{p}_j}{dt} = -W_{\mathrm{ali}}\left(\mathbf{p}_j - \frac{\sum_{k\in K_{\mathrm{ali}}^{(j)}}\mathbf{p}_k}{n_{\mathrm{ali}}}\right) + W_{\mathrm{sep}}\left(\sum_{k\in K_{\mathrm{sep}}^{(j)}}\frac{(\mathbf{q}_j - \mathbf{q}_k)}{|\mathbf{q}_j - \mathbf{q}_k|}\right) + W_{\mathrm{ali}}\left(\mathbf{q}_j - \frac{\sum_{k\in K_{\mathrm{att}}^{(j)}}\mathbf{q}_k}{n_{\mathrm{att}}}\right), \tag{84}$$

$$\frac{d\mathbf{q}_j}{dt} = \mathbf{p}_j, \tag{85}$$

$$\begin{aligned}K_{\mathrm{ali}}^{(j)} &= \left\{k \;\middle|\; |\mathbf{q}_k - \mathbf{q}_j| < r_{\mathrm{ali}},\ \arccos\left(\frac{\mathbf{p}_k\cdot\mathbf{p}_j}{|\mathbf{p}_k||\mathbf{p}_j|}\right) < \theta_{\mathrm{ali}},\ k\neq j\right\},\\ K_{\mathrm{sep}}^{(j)} &= \left\{k \;\middle|\; |\mathbf{q}_k - \mathbf{q}_j| < r_{\mathrm{sep}},\ \arccos\left(\frac{\mathbf{p}_k\cdot\mathbf{p}_j}{|\mathbf{p}_k||\mathbf{p}_j|}\right) < \theta_{\mathrm{sep}},\ k\neq j\right\},\\ K_{\mathrm{att}}^{(j)} &= \left\{k \;\middle|\; |\mathbf{q}_k - \mathbf{q}_j| < r_{\mathrm{att}},\ \arccos\left(\frac{\mathbf{p}_k\cdot\mathbf{p}_j}{|\mathbf{p}_k||\mathbf{p}_j|}\right) < \theta_{\mathrm{att}},\ k\neq j\right\},\\ n_{\mathrm{ali}} &= \sum_{k\in K_{\mathrm{ali}}} 1,\quad n_{\mathrm{att}} = \sum_{k\in K_{\mathrm{att}}} 1,\end{aligned}$$

where $\mathbf{q} \coloneqq (q_1, q_2, q_3)$, $\mathbf{p} \coloneqq (p_1, p_2, p_3)$, and j, k represent the index of an individual. The attraction, separation, and alignment terms are represented by the first, second, and third terms in Eq. (84), and the forces have the interaction ranges, r_{att}, r_{sep}, and r_{ali}, and the angles of view θ_{att}, θ_{sep}, and θ_{ali}, respectively. The parameters W_{att}, W_{sep}, W_{ali}, r_{att}, r_{sep}, r_{ali}, θ_{att}, θ_{sep}, and θ_{ali} of the Reynolds boid model can be tuned to simulate the collective motion of living things such as birds or fish [35,43]. In this study, we tuned these parameters as described in Table I, and we focused on a parameter set that simulates the torus-type behavior of a school of fish in the sea. Such a torus-type collective motion can be realized in a two-dimensional space. Therefore, we set the space to have two dimensions in this study. By solving Eq. (84), we generated 2000 steps of time-series data of the torus-type collective motion by 200 individuals.

To infer the conservation law of collective motion, we need to set a candidate collective coordinate. In this study, we set the collective coordinate on the basis of the following considerations. First, from the visual symmetry of the motion, the average position $(\bar{q}_1, \bar{q}_2)$ and the average momentum $(\bar{p}_1, \bar{p}_2)$ of all particles over time are set as the origin of the coordinate system. Second, since the same behavior is observed regardless of the individual, each individual is considered to have no degree of freedom. From these considerations, we set the coordinate system as $(\tilde{\mathbf{q}}, \tilde{\mathbf{p}}) = (q_1 - \bar{q}_1, q_2 - \bar{q}_2, p_1 - \bar{p}_1, p_2 - \bar{p}_2)$, and prepared the data set as

$$D = \{\tilde{\mathbf{q}}(t_i)_i, \tilde{\mathbf{p}}(t_i)_i, \tilde{\mathbf{q}}(t_i+\Delta t)_i, \tilde{\mathbf{p}}(t_i+\Delta t)_i\}_{i=1}^{N_{\mathrm{R}}T}, \tag{86}$$

$$\coloneqq \{\tilde{\mathbf{q}}(t_{jk})_{jk}, \tilde{\mathbf{p}}(t_{jk})_{jk}, \tilde{\mathbf{q}}(t_{jk}+\Delta t)_{jk}, \tilde{\mathbf{p}}(t_{jk}+\Delta t)_{jk}\}_{\langle j,k\rangle}, \tag{87}$$

where $N_{\mathrm{R}} = 200$, $T = 2,000$, and $\langle j, k\rangle$ represents all combinations of individuals j and time steps k. We randomly selected 5000 samples from this data set for the training of the DNN. We set the coordinate system as

(q_1, q_2, p_1, p_2). In the coordinate, we found the linear relationships $p_2 = -0.161\, q_1 + 0.00$ and $p_1 = 0.167\, q_2 + 0.00$ from the distribution of time-series data [Fig. 7(b)]. From this relationship, the transformation is constrained on the two-dimensional plane corresponding to the normalized vectors $\frac{1}{\sqrt{1+0.161^2}}(1, 0, 0, -0.161)$ and $\frac{1}{\sqrt{1+0.167^2}}(0, 1, 1.67, 0)$. Thus, similarly to case (iii), we define the candidate transformation as

$$A\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \end{pmatrix} + A_0 := \begin{pmatrix} a_{11} & \frac{\sqrt{1+0.167^2}}{\sqrt{1+0.161^2}}a_{21} & 0 & 0 \\ \frac{\sqrt{1+0.161^2}}{\sqrt{1+0.167^2}}a_{12} & a_{22} & 0 & 0 \\ 0 & 0 & a_{22} & -\frac{\sqrt{1+0.161^2}}{\sqrt{1+0.167^2}}\frac{0.167}{0.161}a_{12} \\ 0 & 0 & -\frac{\sqrt{1+0.167^2}}{\sqrt{1+0.161^2}}\frac{0.161}{0.167}a_{21} & a_{11} \end{pmatrix}\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \end{pmatrix} + \begin{pmatrix} a_{01} \\ a_{02} \\ 0.167\, a_{02} \\ -0.161\, a_{01} \end{pmatrix}$$
$$\sim \begin{pmatrix} a_{11} & 1.0\, a_{21} & 0 & 0 \\ 1.0\, a_{12} & a_{22} & 0 & 0 \\ 0 & 0 & a_{22} & -1.0\, a_{12} \\ 0 & 0 & -1.0\, a_{21} & a_{11} \end{pmatrix}\begin{pmatrix} q_1 \\ q_2 \\ p_1 \\ p_2 \end{pmatrix} + \begin{pmatrix} a_{01} \\ a_{02} \\ 0.167\, a_{02} \\ -0.161\, a_{01} \end{pmatrix}, \tag{88}$$

where A and A_0 have constraints $q(A, A_0)$ [see Eq. (47)], which are described as

$$q(A, A_0) = \begin{cases} \text{const} & \text{for} \quad 0.8 < \det A < 1.2, \ \text{and} \ -0.2 < a_{0j} < 0.2, \quad j = 1, 2 \\ 0 & \text{for} \quad \text{else.} \end{cases} \tag{89}$$

The sampling results of D_a are shown in Fig. 7(c) as black dots. The results of PCA of the sampling result D_a are shown in Figs. 7(d) and 7(e). Note that, because the translation transformations a_{01} and a_{02} are affected by the scale of the time-series data distributions of q_1 and q_2, when we apply PCA to the time-series data of this case, we renormalize the sampled a_{01} and a_{02} divided by 100, which is the scale of the distributions q_1 and q_2 [Fig. 6(b)]. The eigenvalues obtained by PCA [Fig. 7(d)] suggest that D_a is embedded in a two-dimensional space. As can be seen from the appearance of the sampling distribution [Fig. 7(c)], this might be the result of embedding one-dimensional circular manifolds. Thus, we set the dimension of the manifold d'_θ to be 1. Also, the principal component vectors with their eigenvalues greater than zero have 0 values corresponding to the elements a_{01} and a_{02} [Fig. 7(e)]. This means that the distribution D_a is not spread out in the space corresponding to a_{01} and a_{02}. This suggests that there is no translational symmetry, i.e., $a_{01} = \text{const.}$ and $a_{02} = \text{const.}$. We verify it by applying method 2 for the sampling results of a_{01} and a_{02} with polynomials of up to first order and selecting a polynomial model from them. For the transformation elements of A, we regressed with polynomials of up to second order and selected a polynomial model from them.

The regression results obtained with the selected implicit polynomial function using the BIC are plotted as red curves in Fig. 7(c). The fitting results are

$$\begin{cases} 1.01\, a_{11}^2 + 1.01\, a_{21}^2 - 0.01\, a_{11} + 0.00\, a_{21} + 0.00\, a_{11}a_{21} = 1.00 \\ 1.01\, a_{11}^2 + 1.00\, a_{12}^2 - 0.01\, a_{11} = 0.99 \\ 1.00\, a_{11} - 1.00\, a_{22} + 0.00\, a_{22} = 0 \\ 1.00\, a_{01} = 0.81 \quad (\text{on } a_{11}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = 0 \quad (\text{on } a_{11}\text{–}a_{02} \text{ plane}) \\ 1.00\, a_{21} - 1.00\, a_{12} = 0 \\ 1.00\, a_{21}^2 + 1.00\, a_{22}^2 - 0.01\, a_{22} = 0.99 \\ 1.00\, a_{01} = 0.81 \quad (\text{on } a_{21}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = 0 \quad (\text{on } a_{21}\text{–}a_{02} \text{ plane}) \\ 1.02\, a_{12}^2 + 1.00\, a_{22}^2 - 0.01\, a_{12}a_{22} = 1.00 \\ 1.00\, a_{01} = 0.81 \quad (\text{on } a_{12}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = 0 \quad (\text{on } a_{12}\text{–}a_{02} \text{ plane}) \\ 1.00\, a_{01} = 0.81 \quad (\text{on } a_{22}\text{–}a_{01} \text{ plane}) \\ 1.00\, a_{02} = 0 \quad (\text{on } a_{22}\text{–}a_{02} \text{ plane}). \end{cases} \tag{90}$$

The simultaneous partial differential equations in Eq. (7), where $b_l = a_{21}$, were obtained from the fitting results. By solving the simultaneous equations, we obtained the infinitesimal transformation

$$\delta\mathbf{q} = \varepsilon\begin{pmatrix} \frac{\partial a_{11}}{\partial a_{21}} & \frac{\partial a_{21}}{\partial a_{21}} \\ \frac{\partial a_{12}}{\partial a_{21}} & \frac{\partial a_{21}}{\partial a_{21}} \end{pmatrix}\mathbf{q} + \varepsilon\begin{pmatrix} \frac{\partial a_{01}}{\partial a_{21}} \\ \frac{\partial a_{02}}{\partial a_{21}} \end{pmatrix} = \varepsilon\begin{pmatrix} \frac{-1.01\times 2a_{21}}{1.01\times 2a_{11}-0.01}\Big|_{A'=e_{\mathbf{I}}} & 1 \\ \frac{-1.00}{1.00}\Big|_{A'=e_{\mathbf{I}}} & \frac{-1.00\times 2a_{21}}{1.00\times 2a_{22}+0.01}\Big|_{A'=e_{\mathbf{I}}} \end{pmatrix}\mathbf{q} + \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{91}$$

$$= \begin{pmatrix} 0 & \varepsilon \\ -\varepsilon & 0 \end{pmatrix}\mathbf{q}, \tag{92}$$

$$\delta\mathbf{p} = \varepsilon\begin{pmatrix} \frac{\partial a_{22}}{\partial a_{21}} & -\frac{\partial a_{12}}{\partial a_{21}} \\ -\frac{\partial a_{21}}{\partial a_{21}} & \frac{\partial a_{11}}{\partial a_{21}} \end{pmatrix}\mathbf{p} + \varepsilon\begin{pmatrix} \frac{\partial a_{02}}{\partial a_{21}} \\ \frac{\partial a_{01}}{\partial a_{21}} \end{pmatrix} = \begin{pmatrix} 0 & \varepsilon \\ -\varepsilon & 0 \end{pmatrix}\mathbf{p}, \tag{93}$$

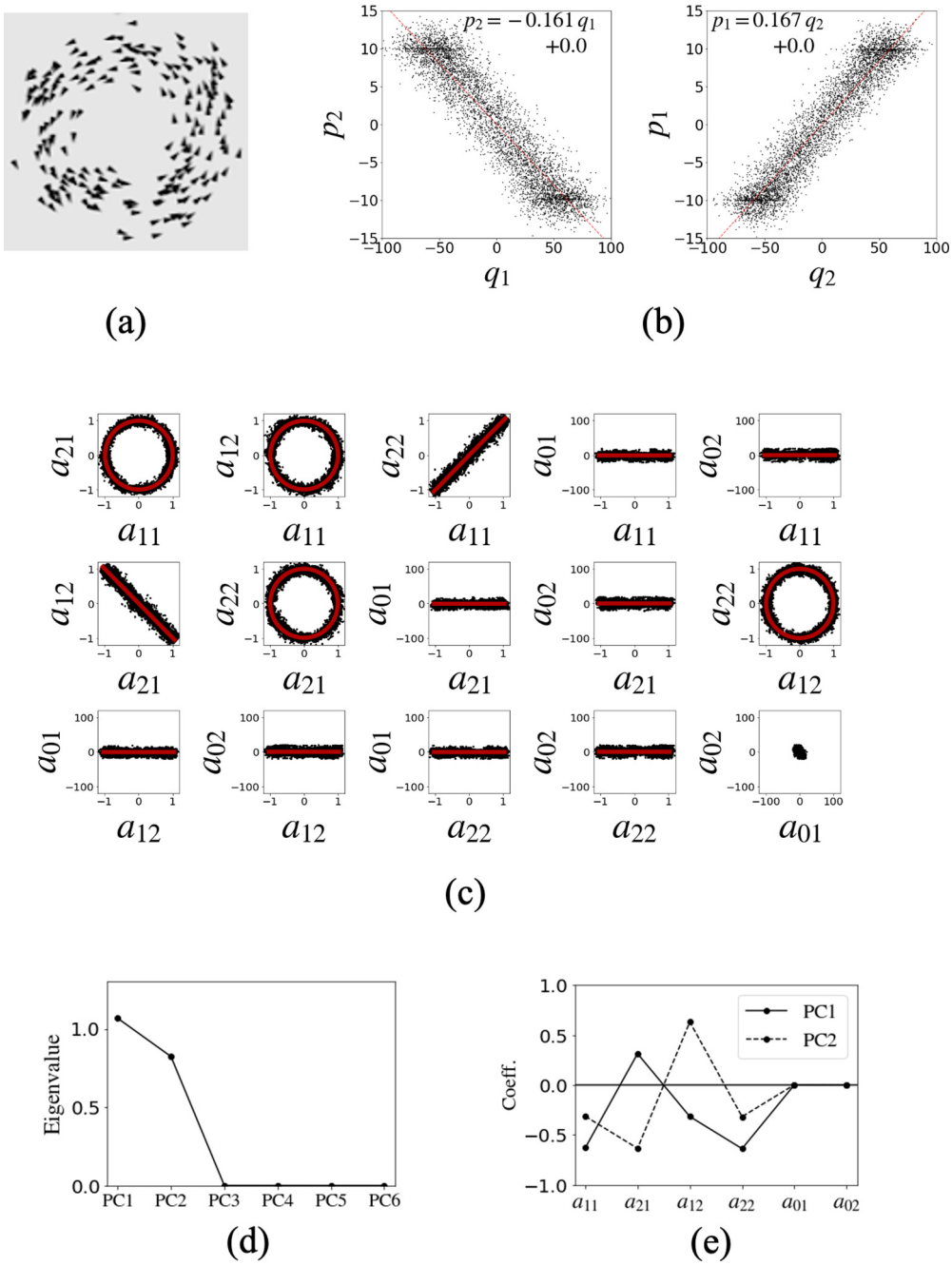

FIG. 7. Results of case (iv): Collective motion system. (a) Simulation snapshot of torus-type collective motion. The simulation data were applied to the proposed method. (b) Distributions of time-series data on q_1–p_2 and q_2–p_1 planes. (c) Black dots represent the sampling distribution D_a obtained by method 1 and the red line represents the fitting curve estimated by method 2. (d) Eigenvalues of the principal components of the distribution D_a. (e) Principal component vectors of the distribution D_a with their eigenvalues greater than zero.

where the values in the final formula are up to one decimal place. By substituting Eqs. (92) and (93) into Eq. (1) and solving the equation, we estimated the conserved value G_δ to be $\varepsilon(x_1p_2 - x_2p_1)$. This result shows that the angular momentum was conserved.

V. SUMMARY AND DISCUSSION

From the results of case (i), we confirmed that method 1 could be used to extract the symmetry. The results of cases (ii) and (iii), wherein the expected conservation laws were inferred, show that method 2 is effective. By comparing cases (i) and (iii), we observed differences in the selected implicit polynomial functions in the a_{11}-a_{22} and a_{21}-a_{12} spaces. These differences emerged from the mirror symmetry in case (i). This finding supports the assertion that the method works well in extracting the symmetry of a system. In case (ii), two translational symmetries corresponding to A_{11} and A_{01} were found, and the simultaneous differential equation corresponding to a_{11} had no solution. a_{11} maps the coordinate

system $(q_t, q_{t+\Delta t})$ to $(a_{11}q_t, a_{11}q_{t+\Delta t})$. Strictly speaking, a_{11} cannot take any value other than 1 because the distribution of the time-series data of case (iii) does not pass through the origin of this coordinate system. Therefore, the presumption that a_{11} has translational symmetry is an erroneous conclusion that ignores the intercept that the time-series data have in the coordinate system $(q_t, q_{t+\Delta t})$. Therefore, the conclusion that the simultaneous differential equations do not have a solution implies the robustness of the proposed framework for the misestimation of invariant transformation. Even if such a false symmetry has a solution, its validity can easily be evaluated by checking whether the estimated conservation law holds with original time-series data. Therefore, we assume that the proposed framework would not lead to the wrong conclusion. For a more practical collective motion system [i.e., case (iv)], we inferred the angular momentum conservation law; the results thereof are consistent with those of a previous study [43]. The previous study suggests that angular momentum is conserved in torus-type swarming patterns. Additionally, the finding of a conservation law in the collective coordinates, where the degree of freedom of an individual degenerated and the origin of the coordinates was the average position and momentum of the swarm, suggests that a dynamical system with a large degree of freedom can be reduced to a central force dynamical system.

In our present study, we dealt only with the case of a single conservation law. If there are multiple conservation laws, the number of dimensions of the manifold D_a increases, and Eq. (55) has multiple orthogonal solutions. Theoretically, the proposed method can still handle such a problem, but the number of combinations of polynomial regressions [Eq. (57)] increases exponentially, and the Jacobian matrix is more likely to be singular. Therefore, it is necessary to develop a more efficient method of estimating an infinitesimal transformation. To estimate an infinitesimal transformation, one needs only to estimate a tangent space around the identity element. Since the sample is finite, in the proposed method, the manifold formed by Lie groups is regressed over the entire space. It is expected that the tangent space can be directly estimated by orthogonal basis decomposition by introducing various constraints.

In this paper, we verified that the framework is feasible for simple cases. For example, in case (iii), the angular momentum conservation law is estimated from the circular orbit data of a central force potential system. When the coordinate transformation is limited to linear transformations of SO(2), a coordinate transformation converting circular orbits into elliptical orbits cannot be realized. Therefore, the symmetry of the system related to the angular momentum conservation law could be estimated by limiting the invariant transformations to the time-series data of circular orbits only. Similarly, there is no linear transformation of SO(2) that transforms an elliptical orbit of one long-axis radius into the elliptical orbit of another long axis radius. Therefore, by preparing all the time-series data of orbits with the same long-axis radius and energy, we may be able to estimate the angular momentum conservation law using the proposed framework in the same manner as used in the case of circular orbits. In contrast, because time-series data for all possible directions of the long-axis are required, the number of time series data required to estimate the invariant transformations is much larger than that for circular orbits, which have no such inherent orientation. In that case, the computational cost required to extract the invariant transformations from the trained DNNs by sampling would be high. Therefore, we would need to improve the method for applying the proposed framework to the time-series data of elliptical trajectories. The proposed framework can also be applied in principle to the estimation of nonlinear invariant transformations (Appendix I). If the proposed framework is applied to nonlinear invariant transformations, it should be possible to estimate the Runge–Lenz vector (Appendix J), a hidden conservation law in the central force potential system. The nonlinear transformation allows transition among all trajectories containing different long-axis radii. Therefore, it would be necessary to prepare all time-series data with a certain energy and apply the proposed framework to them. From the perspective of computational cost, estimating the invariance of such nonlinear transformations is more difficult than estimating the invariance of the matrix transformations of an elliptic orbit system. In addition, to estimate interpretable conservation laws, we would need to model nonlinear transformations of appropriate complexity as parametric functions (Appendix I). This is as difficult as setting up a reduced coordinate system. These difficulties could be addressed by incorporating constraints such as the Lie group axioms or the refinement of candidate coordinate transformations given by physicists into the proposed framework.

The proposed framework can be understood as a framework for evaluating the reduced coordinate system set by physicists directly from time-series data. The reduced coordinate system is usually evaluated by comparing a phenomenon with the model constructed in that coordinate system. Therefore, the proposed framework could allow the search for a reduced coordinate system with the desired property in an exploratory manner from a candidate coordinate system without constructing a model. This should be a great benefit in modeling complex phenomena where the reduced coordinate system is nontrivial.

In this paper, to determine the width of the sampling distribution σ, we proposed a framework for determining a reasonable σ that provides physicists with the sampling distributions of multiple σ obtained by REMC. If it is difficult to determine σ by visualizing sampling distribution D, for example, in a case with multiple conservation laws (and therefore a high-dimensional manifold $M_{\text{invariant}}$) or a case where we need to evaluate σ more quantitatively. For these cases, a possible solution is to apply method 2 of the proposed framework to estimate the conservation laws of the sampling results for all possible σ. Using the obtained conservation laws, physicists would then be able to investigate the appropriate σ using their physical intuition. Moreover, σ is not only related to the accuracy of the modeling of data manifolds by DNNs, but also the coarse-grained scale of the reduced model. Therefore, the hierarchical structure of a dynamics system should also be clarified from the estimated conservation laws at various σ values obtained by the proposed method.

In this study, we used the deep autoencoder to model the time-series data manifolds; nonetheless, there is no need to use a deep autoencoder. The only requirement for a machine learning model is that it has a mapping function that can

determine whether it is on or outside the manifold. From this perspective, the deep autoencoder can be replaced with another type of DNN model, such as a variational autoencoder [44] or a generative adversarial network [45]. Additionally, a feed-forward-type DNN, which is widely used in DNN research, can be used in our proposed method by additionally training a neural network that reconstructs the input data from the output layer of the feed forward neural network. The same method should be feasible for use with machine-learning models that have mapping functions that embed data manifolds into the output space (e.g., the kernel method). Thus, the proposed framework can potentially extract interpretable physical knowledge from a wide range of machine-learning models. Note that the structure of the extracted manifold changes depending on the DNN model and its training settings. This is because the reduced model acquired inside the DNN changes depending on the DNN model and the training settings. How to learn a time-series data set using a certain DNN model and which training settings to use are understood as the implicit construction of the reduced model.

In this paper, we have proposed a method for classical Hamiltonian systems of finite dimensions. We discuss here the extension of the method to the symmetry of the system in canonical quantum field theory. In the canonical quantum field theory, the Hamiltonian is given as

$$H(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{x}), \tag{94}$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ is the field, $\boldsymbol{\pi}(\boldsymbol{x})$ is the canonical momentum conjugate of $\boldsymbol{\phi}(\boldsymbol{x})$, and $\boldsymbol{x} = (ct, x_1, x_2, x_3)$ is the Minkowski space; $\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{\pi}(\boldsymbol{x})$ satisfy the commutation relation:

$$[\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\pi}(\boldsymbol{y})] = i\delta^{(4)}(\boldsymbol{x} - \boldsymbol{y}), \tag{95}$$

$$[\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y})] = [\boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{\pi}(\boldsymbol{y})] = 0. \tag{96}$$

The infinitesimal transformation is given as

$$\Phi^i(\boldsymbol{X}) = \phi^i(\boldsymbol{x}) + \delta\phi^i(\boldsymbol{x}), \tag{97}$$

$$\Pi^i(\boldsymbol{X}) = \pi^i(\boldsymbol{x}) + \delta\pi^i(\boldsymbol{x}), \tag{98}$$

$$X^i = x^i + \delta x^i. \tag{99}$$

Similar to the nested relations between coordinates and time in the classical system, the canonical quantum field theory states that a field and its conjugate momentum have a nested Minkowski space. Therefore, as in the discussion for classical systems, the following relation is given as a condition of the invariant transformation of a Hamiltonian system:

$$\begin{aligned}
&\forall E, \{\boldsymbol{\phi}_{t+\Delta t}, \boldsymbol{\pi}_{t+\Delta t}, \boldsymbol{\phi}_t, \boldsymbol{\pi}_t \mid H(\boldsymbol{\phi}_t, \boldsymbol{\pi}_t) \\
&\quad = E, (\boldsymbol{\phi}_{t+\Delta t}, \boldsymbol{\pi}_{t+\Delta t}) = \mathbf{u}(\boldsymbol{\phi}_t, \boldsymbol{\pi}_t)\} \\
&\quad = \{\boldsymbol{\Phi}_{T+\Delta T}, \boldsymbol{\Pi}_{T+\Delta T}, \boldsymbol{\Phi}_T, \boldsymbol{\Pi}_T \mid H(\boldsymbol{\phi}_t, \boldsymbol{\pi}_t) \\
&\quad = E, (\boldsymbol{\phi}_{t+\Delta t}, \boldsymbol{\pi}_{t+\Delta t}) = \mathbf{u}(\boldsymbol{\phi}_t, \boldsymbol{\pi}_t)\},
\end{aligned}$$

where $\mathbf{u}$ is an equation of motion such as the Klein–Gordon equation of a scalar particle. This equation is the same as Eq. (25), which suggests that the framework can be extended to canonical quantum systems. On the other hand, the Hamiltonian needs to satisfy other conditions, such as the operator order, resulting from noninterchangeability, or the renormalizability of operators. Therefore, to extend the proposed framework to the canonical quantum field theory, we need to develop a method to achieve symmetry estimation that satisfies these conditions.

In this paper, we showed that the proposed framework can infer the hidden conservation laws on a given coordinate of a complex system from DNNs that have been trained with the physical data of the system. On the basis of the obtained results, it is expected that the knowledge of physical data embedded in the trained DNNs in previous studies and the knowledge of physicists can be merged. This should accelerate the research on the construction of reduced models.

ACKNOWLEDGMENTS

I would like to thank Dr. Y. Ando, Prof. S. Goto, Dr. S. Takabe, Prof. H. Hino, Prof. K. Fukumizu, Prof. K. Hukushima, Prof. T. Ikegami, Prof. K. Ishikawa, H. Yamashita, H. Murata, Prof. Y. Yue, and K. Sakamoto for useful discussions. This work was supported by KAKENHI Grants No. JP20H04648, No. JP17H01793, and No. JP19K12111. This work was also supported by JST CREST Grant No. JPMJCR2015.

APPENDIX A: DERIVATION OF EQUIVALENT CONDITION TO MAKE HAMILTONIAN INVARIANT

The identity condition $H(\mathbf{q}, \mathbf{p}) \equiv H'(\mathbf{q}, \mathbf{p})$ has the equivalent expression

$$\forall(\mathbf{q}, \mathbf{p}), H(\mathbf{q}, \mathbf{p}) = H'(\mathbf{q}, \mathbf{p}). \tag{A1}$$

This condition can be transformed to an equivalent conditional expression represented by a set,

$$\forall E, \ \{\mathbf{q}, \mathbf{p} \mid H(\mathbf{q}, \mathbf{p}) = E\} = \{\mathbf{q}, \mathbf{p} \mid H'(\mathbf{q}, \mathbf{p}) = E\}, \tag{A2}$$

which is proved in Appendix C. Replacing $\mathbf{q}, \mathbf{p}$ with the transformed parameters $\mathbf{Q}, \mathbf{P}$ does not change the set: $\{\mathbf{q}, \mathbf{p} \mid H'(\mathbf{q}, \mathbf{p}) = E\} = \{\mathbf{Q}, \mathbf{P} \mid H'(\mathbf{Q}, \mathbf{P}) = E\}$. Therefore, Eq. (A2) is rewritten as

$$\forall E, \ \{\mathbf{q}, \mathbf{p} \mid H(\mathbf{q}, \mathbf{p}) = E\} = \{\mathbf{Q}, \mathbf{P} \mid H'(\mathbf{Q}, \mathbf{P}) = E\}. \tag{A3}$$

From the definition of the transformed Hamiltonian H', $H'(\mathbf{Q}, \mathbf{P}) \coloneqq H(\mathbf{q}(\mathbf{Q}, \mathbf{P}), \mathbf{p}(\mathbf{Q}, \mathbf{P})) = H(\mathbf{q}, \mathbf{p})$ are satisfied. By substituting these into Eq. (A3), we obtain the target condition equivalent to the identity condition $H(\mathbf{q}, \mathbf{p}) \equiv H'(\mathbf{q}, \mathbf{p})$ as

$$\forall E, \ \{\mathbf{q}, \mathbf{p} \mid H(\mathbf{q}, \mathbf{p}) = E\} = \{\mathbf{Q}, \mathbf{P} \mid H(\mathbf{q}, \mathbf{p}) = E\}. \tag{A4}$$

APPENDIX B: DERIVATION OF EQUIVALENT CONDITION TO MAKE CANONICAL EQUATIONS INVARIANT

The identity condition in Eq. (23),

$$\mathbf{u}(\mathbf{q}_t, \mathbf{p}_t) \equiv \mathbf{u}'(\mathbf{q}_t, \mathbf{p}_t) \quad \wedge \quad \mathbf{v}(\mathbf{q}_t, \mathbf{p}_t) \equiv \mathbf{v}'(\mathbf{q}_t, \mathbf{p}_t), \tag{B1}$$

has the equivalent expression:

$$\forall(\mathbf{q}_t, \mathbf{p}_t), \ (\mathbf{u}(\mathbf{q}_t, \mathbf{p}_t), \mathbf{v}(\mathbf{q}_t, \mathbf{p}_t)) = (\mathbf{u}'(\mathbf{q}_t, \mathbf{p}_t), \mathbf{v}'(\mathbf{q}_t, \mathbf{p}_t)). \tag{B2}$$

This condition can be transformed to the following equivalent conditional expression represented by a set:

$$\begin{aligned}&\forall(\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t}),\{\mathbf{q}_t,\mathbf{p}_t \mid (\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})\\&\quad=(\mathbf{u}(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}(\mathbf{q}_t,\mathbf{p}_t))\}=\{\mathbf{q}_t,\mathbf{p}_t \mid (\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})\\&\quad=(\mathbf{u}'(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}'(\mathbf{q}_t,\mathbf{p}_t))\}.\end{aligned}\tag{B3}$$

The proof of the equivalence of Eqs. (B2) and (B3) is a multivariable case of the proof described in Appendix C. By treating $\mathbf{q}_{t+\Delta t}$, $\mathbf{p}_{t+\Delta t}$ as a set of elements, we transform the condition in Eq. (B3) to the equivalent condition (see the proof in Appendix D):

$$\begin{aligned}&\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \mid (\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})\\&\quad=(\mathbf{u}(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}(\mathbf{q}_t,\mathbf{p}_t))\}\\&\quad=\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \mid (\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})\\&\quad=(\mathbf{u}'(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}'(\mathbf{q}_t,\mathbf{p}_t))\}.\end{aligned}\tag{B4}$$

Replacing $\mathbf{q}_t,\mathbf{p}_t,\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t}$ with the transformed parameters $\mathbf{Q}_T,\mathbf{P}_T,\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T}$ does not change the set:

$$\begin{aligned}&\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \mid (\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})\\&\quad=(\mathbf{u}'(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}'(\mathbf{q}_t,\mathbf{p}_t))\},\end{aligned}\tag{B5}$$

$$\begin{aligned}&=\{\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T},\mathbf{Q}_T,\mathbf{P}_T \mid (\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T})\\&=(\mathbf{u}'(\mathbf{Q}_T,\mathbf{P}_T),\mathbf{v}'(\mathbf{Q}_T,\mathbf{P}_T))\}.\end{aligned}\tag{B6}$$

Therefore, Eq. (B4) is rewritten as

$$\begin{aligned}&\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \mid (\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})\\&\quad=(\mathbf{u}(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}(\mathbf{q}_t,\mathbf{p}_t))\}\\&\quad=\{\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T},\mathbf{Q}_T,\mathbf{P}_T \mid (\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T})\\&\quad=(\mathbf{u}'(\mathbf{Q}_T,\mathbf{P}_T),\mathbf{v}'(\mathbf{Q}_T,\mathbf{P}_T))\}.\end{aligned}\tag{B7}$$

From the definition of the transformed canonical equations [Eqs. (14) and (15)], we obtain

$$(\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T})=(\mathbf{u}'(\mathbf{Q}_T,\mathbf{P}_T),\mathbf{v}'(\mathbf{Q}_T,\mathbf{P}_T)),\tag{B8}$$

$$\Leftrightarrow(\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})=(\mathbf{u}(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}(\mathbf{q}_t,\mathbf{p}_t)).\tag{B9}$$

By substituting this in Eq. (B7), we obtain the target condition equivalent to the identity condition in Eq. (23) as

$$\begin{aligned}&\{\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t},\mathbf{q}_t,\mathbf{p}_t \mid (\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})\\&\quad=(\mathbf{u}(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}(\mathbf{q}_t,\mathbf{p}_t))\}\\&\quad=\{\mathbf{Q}_{T+\Delta T},\mathbf{P}_{T+\Delta T},\mathbf{Q}_T,\mathbf{P}_T \mid (\mathbf{q}_{t+\Delta t},\mathbf{p}_{t+\Delta t})\\&\quad=(\mathbf{u}(\mathbf{q}_t,\mathbf{p}_t),\mathbf{v}(\mathbf{q}_t,\mathbf{p}_t))\}.\end{aligned}\tag{B10}$$

APPENDIX C: PROOF OF EQ. (A1)⇔EQ. (A2) AND EQ. (B2)⇔EQ. (B3)

The problem can be abstracted as the proposition below:

$$\forall x,f(x)=g(x)\Leftrightarrow\forall E,\{x|f(x)=E\}=\{x|g(x)=E\},\tag{C1}$$

where $f(x)$ and $g(x)$ are single-valued functions:

$$\begin{aligned}\text{Proof of }\forall x,f(x)=g(x)\rightarrow\forall E,\{x|f(x)\\=E\}=\{x|g(x)=E\}.\end{aligned}\tag{C2}$$

The contrapositive of (C2) is $\exists E,\{x|f(x)=E\}\neq\{x|g(x)=E\}\rightarrow\exists x,f(x)\neq g(x)$. This contrapositive is proved as follows. Since $\exists E,\{x|f(x)=E\}\neq\{x|g(x)=E\}$, there exists E' and x', which satisfy $f(x')=E'$, but $g(x')\neq E'$. Therefore, $\exists x,f(x)\neq g(x)$ is satisfied because :

□

$$\begin{aligned}\text{Proof of }\forall E,\{x|f(x)=E\}=\{x|g(x)=E\}\rightarrow\forall x,f(x)\\=g(x)\end{aligned}\tag{C3}$$

The contrapositive of (C3) is $\exists x,f(x)\neq g(x)\rightarrow\exists E,\{x|f(x)=E\}\neq\{x|g(x)=E\}$. This contrapositive is proved as follows. Select one x' from x, which satisfies $f(x')\neq g(x')$ and $f(x')=E'$. Since $f(x)$ is a single-valued function, x' is not included in the set of x that satisfies $g(x)=E'$. Thus, $\{x|f(x)=E'\}\neq\{x|g(x)=E'\}$ holds. Therefore, $\exists E,\{x|f(x)=E\}\neq\{x|g(x)=E\}$ is satisfied.

□

APPENDIX D: PROOF OF EQ. (B3)⇔EQ. (B4)

The problem can be abstracted as the proposition below:

$$\begin{aligned}\forall b,\{x|\,f(x)=b\}=\{x|\,g(x)=b\}\Leftrightarrow\{x,b|\,f(x)\\=b\}=\{x,b|\,g(x)=b\},\end{aligned}\tag{D1}$$

where $f(x)$ and $g(x)$ are single-valued functions:

$$\begin{aligned}\text{Proof of }\forall b,\{x|\,f(x)=b\}=\{x|\,g(x)=b\}\\\rightarrow\{x,b|\,f(x)=b\}\\=\{x,b|\,g(x)=b\}.\end{aligned}\tag{D2}$$

The contrapositive of (D2) is $\{x,b|\,f(x)=b\}\neq\{x,b|\,g(x)=b\}\rightarrow\exists b,\{x|\,f(x)=b\}\neq\{x|\,g(x)=b\}$. This contrapositive is proved as follows. Since $\{x,b|\,f(x)=b\}\neq\{x,b|\,g(x)=b\}$, there is a set of x' and b', which satisfies $f(x')=b'$ and $g(x')\neq b'$. Therefore, $\{x|\,f(x)=b'\}\neq\{x|\,g(x)=b'\}$ holds. It means that $\exists b,\{x|\,f(x)=b\}\neq\{x|\,g(x)=b\}$ is satisfied.

□

$$\begin{aligned}\text{Proof of }\{x,b|\,f(x)=b\}=\{x,b|\,g(x)=b\}\\\rightarrow\forall b,\{x|\,f(x)=b\}\\=\{x|\,g(x)=b\}.\end{aligned}\tag{D3}$$

The contrapositive of (D3) is $\exists b,\{x|\,f(x)=b\}\neq\{x|\,g(x)=b\}\rightarrow\{x,b|\,f(x)=b\}\neq\{x,b|\,g(x)=b\}$. This contrapositive is proved as follows. Since $\exists b,\{x|\,f(x)=b\}\neq\{x|\,g(x)=b\}$, there is a set of b' and x', which satisfies $f(x')=b'$ and $g(x')\neq b'$. Therefore, $\{x,b|\,f(x)=b\}\neq\{x,b|\,g(x)=b\}$ is satisfied.

□

APPENDIX E: REPLICA EXCHANGE MONTE CARLO (REMC) METHOD AND ITS PARAMETERS

Using $A':=(a_{11},a_{12},a_{21},\cdots,a_{2d\,2d})$, we reexpress Eq. (47) as

$$\mathrm{P}(A')=\frac{1}{Z}\exp\left[-\frac{N}{2\sigma^2}E_{\mathrm{samp}}(A')\right]q(A').\tag{E1}$$

TABLE II. Parameters of REMC method.

Parameter name	(i) Half sphere	(ii) Constant velocity	(iii) Central force	(iv) Collective motion
Sampling size, N_a	10,000	50 000	10 000	10 000
L	20	30	30	30
γ	1.4	1.9	1.4	1.4
C $[C_A/C_{A_0}]$	3.0/0.3	0.03/0.3	0.3/0.03	0.3/0.03
d	0.6	0.7	0.8	0.8
e	5.0	1.0	5.0	5.0
$\sigma_{\min}$	4.42×10^{-2}	5.41×10^{-5}	1.67×10^{-1}	8.0
Burn-in length	10.000	50,000	10,000	10,000
Selected noise intensity, σ	3.32×10^{-1}	5.80	2.92	3.83×10^{2}

The REMC method takes samples from the joint density,

$$\mathrm{P}(A'^1, \cdots, A'^l, \cdots, A'^L) = \prod_{l=1}^{L} \frac{1}{Z} \exp\left[-\frac{N}{2\sigma_l^2} E_{\mathrm{samp}}(A')\right] q(A'), \tag{E2}$$

where $\sigma_l > \sigma_{l+1}$ and $\sigma_L = \sigma$. In the REMC method, sampling from the joint density $\mathrm{P}(A'^1, \cdots, A'^l, \cdots, A'^L)$ is performed on the basis of the following updates:

(1) Sampling from each density $\mathrm{P}(A'^1, \cdots, A'^l, \cdots, A'^L)$. Sampling A'^l from $\mathrm{P}(A'^l) := \frac{1}{Z_l} \exp\left[-\frac{N}{2\sigma_l^2} E_{\mathrm{samp}}(A'^l)\right] q(A')$, where Z' is the normalization constant. The sampling is performed by a conventional Monte Carlo method, such as the Metropolis–Hastings algorithm [46].

(2) Exchange between two densities corresponding to noise intensity σ.

The exchanges between the configurations A'^l and A'^{l+1} correspond to adjacent inverse temperatures following the probability $R = \min(1, r)$, where

$$\begin{aligned} r &= \frac{\mathrm{P}(A'^1, \cdots, A'^{l+1}, A'^l, \cdots, A'^L)}{\mathrm{P}(A'^1, \cdots, A'^l, A'^{l+1}, \cdots, A'^L)} \\ &= \frac{\mathrm{P}(A'^{l+1})\mathrm{P}(A'^l)}{\mathrm{P}(A'^l)\mathrm{P}(A'^{l+1})} \\ &= \exp\left\{\frac{N}{2}\left[\sigma_{l+1}^{-2} - \sigma_l^{-2}\right]\left[E(A'^{l+1}) - E_{\mathrm{samp}}(A'^l)\right]\right\}. \end{aligned}$$

Sampling from a distribution with a larger σ_l tends not to have a local minimum. Hence, sampling from the joint density $\mathrm{P}(A'^1, A'^2 \cdots A'^L)$ overcomes the local minima in distributions with small σ_l and enables the rapid convergence of sampling.

For the execution of EMC sampling, we adopted the Metropolis–Hastings algorithm [46] to sample each state of σ_l. When we performed the Metropolis–Hastings sampling, a candidate for the next sample $a_{ij}^{l\ \mathrm{next}}$ is picked from the conditional probability distribution with the precondition $a_{ij}^{l\ \mathrm{previous}}$,

$$\mathrm{P}\left(a_{ij}^{l\ \mathrm{next}} | a_{ij}^{l\ \mathrm{previous}}\right) = \frac{1}{2U_l} \quad \left(-U_l \leqslant a_{ij}^{l\ \mathrm{next}} \leqslant U_l\right), \tag{E3}$$

where U_l is set as

$$U_l = \begin{cases} C & \left(eN\sigma_l^{-2} \geqslant 1\right) \\ \frac{C}{\left(eN\sigma_l^{-2}\right)^d} & \left(eN\sigma_l^{-2} < 1\right). \end{cases} \tag{E4}$$

C, d, and e in Eq. (E4) are set as Table II for the evaluation of the proposed method in Sec. IV. The sampling parameters C of A and A_0 were set as different values [the values are described as (C of A)/(C of A_0) in the column for constant velocity (ii) in Table II]. Each state of σ_l was determined following the exponential function [47]:

$$\sigma_l^{-2} = \begin{cases} 0.0 & (l = 1) \\ \sigma_{\min}^{-2} \gamma^{(l-1)-L} & (l \neq 1), \end{cases} \tag{E5}$$

where $\sigma_{\min}$ is set as the root-mean-square error for $A = \boldsymbol{I}$ in the trained DNN because it represents the minimum value of E_{samp}. L and γ are set as shown in Table II for each case.

APPENDIX F: NOISE INTENSITY OF SAMPLING

Depending on the difference in σ, the sampling results corresponding to low MSE and the sampling results corresponding to high MSE are obtained [Fig. 8(a)]. In the low-MSE region, the transformation matrix corresponding to the identity mapping is sampled [Fig. 8(b)]. In the high-MSE region, the transformation matrix corresponding to the rotation matrix is sampled [Fig. 8(d)]. At intermediate noise intensities, sampling between both conditions is achieved [Fig. 8(c)]. On the basis of such a structure, in this research, we select the noise intensity σ that looks like it will obtain the nonidentity transformation and a continuously connected distribution, such as $\sigma = 3.452$ of Fig. 8. Thus, in this proposed framework, we determine the σ through qualitative considerations based on the sampling results of multiple σ obtained in REMC. In the demonstration case of this study, we qualitatively determined σ by focusing on the nonidentity feature, but we could also focus on other features. In the proposed framework, we propose to provide physicists with sampling results of multiple σ instead of a specific σ determination method. For the evaluation of the proposed method in Sec. IV, we select the noise intensity σ for each evaluation case, as shown in Table II.

APPENDIX G: ESTIMATION OF LIKELIHOOD FOR THE MODEL SELECTION

Under the assumption that N_a samples of transformation are given with Gaussian noise, the following likelihood is

defined for a statistical selection of implicit function:

$$P(\vec{b}_1, \vec{b}_2, \cdots, \vec{b}_{N_a}) = \prod_{n_a=1}^{N_a} \frac{1}{Z} \exp\left\{-\frac{1}{2\sigma_b^2} D\left[\vec{b}_{n_a}, f\left(c_k, b_1, b_2, \cdots, b_{d'_\theta}; \beta, \gamma, d'_\theta\right)\right]^2\right\}, \tag{G1}$$

$$Z = \int_{-\infty}^{\infty} d\vec{b}_{n_a} \exp\left\{-\frac{1}{2\sigma_b^2} D\left[\vec{b}_{n_a}, f\left(c_k, b_1, b_2, \cdots, b_{d'_\theta}; \beta, \gamma, d'_\theta\right)\right]^2\right\} q\left(\vec{b}_{n_a}\right), \tag{G2}$$

$$\sigma_b = \left\{\frac{1}{N_a} \sum_{n_a=1}^{N_a} D\left[\vec{b}_{n_a}, f\left(c_k, b_1, b_2, \cdots, b_{d'_\theta}; \beta, \gamma, d'_\theta\right)\right]^2\right\}^{\frac{1}{2}}, \tag{G3}$$

$$\vec{b}_{n_a} = \left(c_k, b_1, b_2, \cdots, b_{d'_\theta}\right)_{n_a}, \tag{G4}$$

where $D[\vec{b}_{n_a}, f(c_k, b_1, b_2, \cdots, b_{d'_\theta}; \beta, \gamma, d'_\theta)]$ is the minimum distance from a data point $\vec{b}_{n_a}$ to a subspace defined by the implicit function $f(c_k, b_1, b_2, \cdots, b_{d'_\theta}; \beta, \gamma, d'_\theta) = 0$, and $q(\vec{b}_{n_a})$ is set corresponding to the constraints of the transformation $q(A')$, in Eq. (47). The normalized constant Z is estimated numerically as the Riemann sum.

APPENDIX H: DNN MODEL AND ITS TRAINING PARAMETERS

In this Appendix, we describe the DNN models and their training settings.

In this paper, we used deep autoencoders as DNN models. In cases (i)–(iv), the DNNs consisted of an input layer, three hidden layers, and an output layer. The number of nodes in each layer was set as shown in the network structure in Table III.

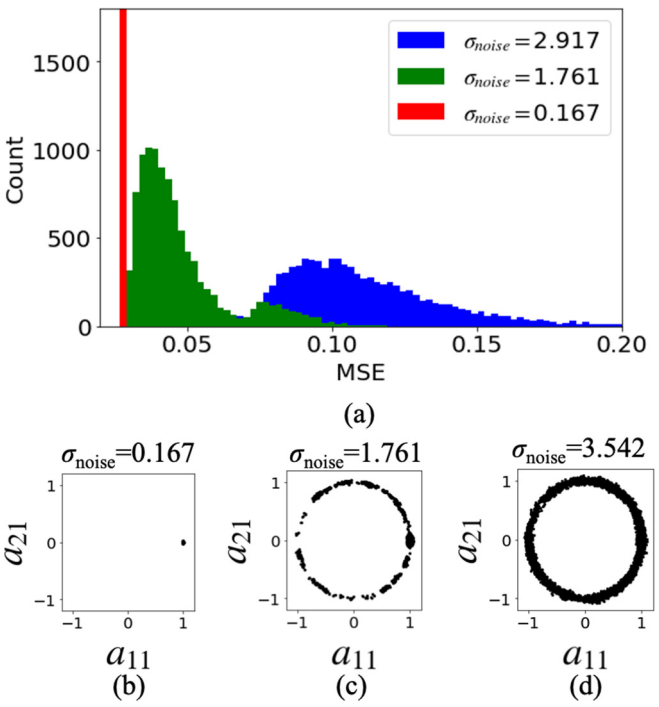

FIG. 8. Qualitative transition of sampling results due to the increase in noise intensity. The figures describe the qualitative transition of the case (iii) central force system where rotation symmetry exists. (a) Distributions of MSE with different noise intensities. (b)–(d) Sampling results of a_{11} and a_{21} at each noise intensity.

The activation functions of the deep autoencoders were set as the sigmoid or hyperbolic tangent functions, as shown in the activation function in Table III. The sigmoid function is defined as

$$\text{sigmoid}(x) = \frac{1}{1 + \exp(-x)} \tag{H1}$$

and the tanh function is defined as

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}. \tag{H2}$$

The numbers of samples used for training DNN are shown in Table III as training data size N. The Adam method [51] was used for training. The training iterations are shown in Table III. For training, the data were divided into minibatches whose sizes are shown in Table III as minibatch size.

As mentioned in the main text, the proposed framework assumes that the network structure and training parameters of the DNN are given by a physicist. In the demonstrations of this study, as a physicist, we set the network structure and learning parameters of the DNN model using the following considerations and procedures.

Step 1: We determined the number of nodes in the center layer corresponding to the dimensions of the data manifold.

Case (i): In this case, we generated the data so the dimension of the manifold is two, so we used two nodes.

Cases (ii) and (iii): For these cases, we assumed that the system has no multiple scales of motion. This means that the dimension of the manifold is clearly determined as a single value. Therefore, we expect the data manifold reconstruction error obtained by the DNN to be clearly larger when the number of nodes is lower than the dimension of the manifold. When we reduced the number of nodes under this assumption, a sufficiently small error was realized even the number of center nodes was one. Specifically, in case (ii), the reconstructed mean squared root error (MSRE) was 8.35×10^{-4} for the input data with a value range of O(1), and in case (iii), the MSRE was 1.67×10^{-1} for the input data with a value range of about O(10). Thus, we set the number of nodes in the center layer to one.

Case (iv): In this case, we set the number of nodes in the middle layer to one based on the assumption that the motion of case (iv) is analogous to that of case (iii).

TABLE III. Parameters of DNN and its training. In the network structure, the number of nodes is shown in the order from left to right: input layer–first layer–second layer–third layer–output layer.

Parameter name	(i) Half sphere	(ii) Constant velocity	(iii) Central force	(iv) Collective motion
Training data size N	1671	1000	1000	5000
Network structure	3-10-2-10-3	4-10-1-10-4	8-20-1-20-8	8-20-1-20-8
Activation function	sigmoid	tanh	sigmoid	sigmoid
Training algorithm	Adam	Adam	Adam	Adam
Training iteration	50 000	100000	50 000	50 000
Minibatch size	10	30	10	10
Library	theano [48,49]	scikit-learn [50]	theano	theano

Step 2: Based on the number of nodes in the center layer given in step 1, we tuned the number of DNN layers, number of nodes in the other intermediate layers, or learning parameters to minimize the reconstruction error.

Step 3: We determined the whole network structure and training parameters of the DNN by feeding the results of step 1 and step 2 to each other.

APPENDIX I: APPLICABILITY OF THE METHODS TO MORE GENERAL COORDINATE TRANSFORMATIONS

1. Method 1 for general coordinate transformations

In this paper, we presented an explanation of the method for a matrix representation of Lie groups. The method can be applied in principle to a wide range of Lie group realizations including nonlinear coordinate transformations. In this Appendix, we discuss method 1 for this wide range of transformation functions. The set of expected invariant transformations is defined as follows in Eq. (43):

$$M_{\text{invariant}} \coloneqq \left\{ Q^{S_i}(\cdot,\cdot,\boldsymbol{\theta}), \mathcal{P}^{S_i}(\cdot,\cdot,\boldsymbol{\theta}) |\, \boldsymbol{\theta} \in \mathbb{R}^{d_\theta} \right\}. \tag{43}$$

In the main paper, we set the transformation function $Q^{S_i}(\cdot,\cdot,\boldsymbol{\theta}), \mathcal{P}^{S_i}(\cdot,\cdot,\boldsymbol{\theta})$ as the matrix transformation function, but the transformation function can also be set as an otherwise complex transformation function, such as a nonlinear function. In such a case, because $Q^{S_i}(\cdot,\cdot;\boldsymbol{\theta})$, $\mathcal{P}^{S_i}(\cdot,\cdot;\boldsymbol{\theta})$ are usually unknown, we infer them to be a subset of a parametric function set $\{\mathbf{Q}(\cdot,\cdot;\mathbf{a}),\ \mathbf{P}(\cdot,\cdot;\mathbf{a}) |\, \mathbf{a} \in \mathbb{R}^{d_a}\}$, where $d_a \geqslant d_\theta$. This function can be complex enough to contain a true transformation function, but it will be more difficult to determine the subset from the finite data. Moreover, significant difficulties arise when applying method 2. This will be discussed further in the next Appendix section.

Similar to the matrix transformation case in the main text, the subset of the true transformation function $M_{\text{invariant}}$ is identified using the trained DNN as

$$M_{\text{invariant}} \sim \{\mathbf{Q}(\cdot,\cdot;\mathbf{a}), \mathbf{P}(\cdot,\cdot;\mathbf{a}) |\ \operatorname*{argmin}_{\mathbf{a}} E_{\text{samp}} \times [\mathbf{Q}(\cdot,\cdot;\mathbf{a}), \mathbf{P}(\cdot,\cdot;\mathbf{a})]\}, \tag{I1}$$

$$E_{\text{samp}}[\mathbf{Q}(\cdot,\cdot;\mathbf{a}), \mathbf{P}(\cdot,\cdot;\mathbf{a})] = \frac{1}{N}\sum_{i=1}^{N} \{[\mathbf{Q}(\cdot,\cdot;\mathbf{a}), \mathbf{P}(\cdot,\cdot;\mathbf{a})] - \mathbf{f}_{\text{DNN}}[\mathbf{Q}(\cdot,\cdot;\mathbf{a}), \mathbf{P}(\cdot,\cdot;\mathbf{a})]\}^2. \tag{I2}$$

Next, the invariant transformation is obtained by sampling an element a_j of the parameter vector $\mathbf{a}$ following the probability distribution, as in the matrix transformation case

$$\mathrm{P}(a_1, a_2, a_3, \cdots, a_{d_a}) = \frac{1}{Z}\exp\left\{-\frac{N}{2\sigma^2} E_{\text{samp}}[\mathbf{Q}(\cdot,\cdot;\mathbf{a}), \mathbf{P}(\cdot,\cdot;\mathbf{a})]\right\}. \tag{I3}$$

2. Method 2 for general coordinate transformations

From the N_a sampling results of Eq. (I3), $D_a \coloneqq \{(a_1, a_2 \cdots a_{d_a})_{n_a}\}_{n_a=1}^{N_a}$, the infinitesimal transformations are estimated as follows.

Assuming that $\mathbf{a}$ is a differentiable function of $\boldsymbol{\theta}$: $\mathbf{a}(\boldsymbol{\theta})$, $\mathbb{R}^{d_\theta} \to \mathbb{R}^{d_a}$, we can estimate $M_{\text{invariant}}$ as

$$M_{\text{invariant}} = \left\{\mathbf{Q}(\cdot,\cdot;\mathbf{a}(\boldsymbol{\theta})), \mathbf{P}(\cdot,\cdot;\mathbf{a}(\boldsymbol{\theta}))|\ \boldsymbol{\theta} \in \mathbb{R}^{d_\theta}\right\}. \tag{I4}$$

The set of invariant transformations $M_{\text{invariant}}$ forms a Lie group, as we mentioned in Sec. II A. Therefore, $M_{\text{invariant}}$ constructs a d_θ-dimensional differential manifold in the coordinate space of $\boldsymbol{\theta}$. The infinitesimal transformation is estimated as the tangent vector of the manifold at $\boldsymbol{\theta} = \mathbf{0}$ as follows:

$$(\delta\mathbf{q}_l, \delta\mathbf{p}_l) = \varepsilon\left(\left.\frac{\partial \mathbf{Q}(\mathbf{q},\mathbf{p};\mathbf{a}(\theta_l))}{\partial\theta_l}\right|_{\theta_l=0}, \left.\frac{\partial \mathbf{P}(\mathbf{q},\mathbf{p};\mathbf{a}(\theta_l))}{\partial\theta_l}\right|_{\theta_l=0}\right). \tag{I5}$$

Because $\mathbf{a}$ is a differentiable function of $\boldsymbol{\theta}$, the tangent vector is given as

$$(\delta\mathbf{q}_l, \delta\mathbf{p}_l) = \varepsilon\left(\sum_{k=1}^{d_a} \frac{\partial \mathbf{Q}(\mathbf{q},\mathbf{p};\mathbf{a})}{\partial a_k}\frac{\partial a_k(\boldsymbol{\theta})}{\partial\theta_l}\bigg|_{\boldsymbol{\theta}=\mathbf{0}}, \times \sum_{k=1}^{d_a} \frac{\partial \mathbf{P}(\mathbf{q},\mathbf{p};\mathbf{a})}{\partial a_k}\frac{\partial a_k(\boldsymbol{\theta})}{\partial\theta_l}\bigg|_{\boldsymbol{\theta}=\mathbf{0}}\right). \tag{I6}$$

Because functions $\mathbf{Q}$ and $\mathbf{P}$ are defined explicitly, their derivations, $\frac{\partial \mathbf{Q}(\mathbf{q},\mathbf{p};\mathbf{a})}{\partial a_k}$ and $\frac{\partial \mathbf{P}(\mathbf{q},\mathbf{p};\mathbf{a})}{\partial a_k}$, can be obtained analytically. Therefore, we should only estimate $\frac{\partial a_k(\boldsymbol{\theta})}{\partial\theta_l}|_{\boldsymbol{\theta}=\mathbf{0}}$ to obtain the infinitesimal transformation. Additionally, as for the linear-transformation case in the main text, if a_k can be regressed around $\boldsymbol{\theta} = \mathbf{0}$ as a first-order polynomial of $\{\theta_l\}_{l=1}^{d_\theta}$, the conservation law can be inferred without approximation.

Because $\mathbf{a}(\boldsymbol{\theta})$ is defined as a differentiable function, set $\{\mathbf{a}|\boldsymbol{\theta} \in \mathbb{R}^{d_\theta}\}$ constructs a d_θ-dimensional manifold structure in coordinate space $\mathbf{a}$. The implicit function representation of the

manifold is defined as

$$\begin{cases} f_1(a_1, \cdots, a_{d_a}) = 0 \\ \quad \vdots \\ f_{d_a-d_\theta}(a_1, \cdots, a_{d_a}) = 0. \end{cases} \tag{I7}$$

The Jacobian matrix of f_k for the parameters of subset $\mathbf{a}$, $(b_1, b_2, \cdots, b_{d_\theta}) \subset \mathbf{a}$, is defined as $J_{kl} = \frac{\partial f_k(a_1,\cdots,a_{d_a})}{\partial b_l}$. If the Jacobian matrix at $\mathbf{a}_{\rm id}$ becomes nonsingular, from the implicit function theorem, variables other than $(b_1, b_2, \cdots, b_{d_\theta})$, $\{c_k\}_{k=1}^{d_a-d_\theta} := A' \setminus \{b_l\}_{l=1}^{d_\theta}$, can be expressed as $c_k = g_i(b_1, \cdots, b_{d_\theta})$. This means that $\boldsymbol{\theta}$ can be replaced by $\mathbf{b}$. In this case, $\frac{\partial a_k(\boldsymbol{\theta})}{\partial \theta_l}|_{\boldsymbol{\theta}=\mathbf{0}}$ is estimated as the tangent vector $\frac{\partial a_k(\boldsymbol{b})}{\partial b_l}|_{\mathbf{a}=\mathbf{a}_{\rm id}}$ at identity map

$$\mathbf{a}_{\rm id} \in \{\mathbf{a} | \mathbf{Q}(\cdot, \cdot; \mathbf{a}) = \mathbf{q}, \mathbf{P}(\cdot, \cdot; \mathbf{a}) = \mathbf{p}\}. \tag{I8}$$

This implies that, around $e_{\mathbf{I}}$, the implicit equations in Eq. (I7) representing the manifold $M_{\rm invariant}$ can be decomposed into the following $d' - d_\theta$ simultaneous equations:

$$\begin{cases} h_1(c_1, b_1, \cdots, b_{d_\theta}) = 0 \\ \quad \vdots \\ h_{d'-d_\theta}(c_{d'-d_\theta}, b_1, \cdots, b_{d_\theta}) = 0, \end{cases} \tag{I9}$$

where b_l corresponds to the continuous parameter θ_l of continuous transformation $[\mathcal{Q}(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta}), \mathcal{P}(\mathbf{q}, \mathbf{p}, \boldsymbol{\theta})]$. Differentiating these equations with respect to b_l around a point $e_{\mathbf{I}}$ yields $d' - d_\theta$ simultaneous partial differential equations:

$$\begin{cases} \frac{\partial}{\partial b_l} h_1(c_1, b_1, \cdots, b_{d_\theta})|_{A'=e_{\mathbf{I}}} = 0 \\ \quad \vdots \\ \frac{\partial}{\partial b_l} h_{d'-d_\theta}(c_{d'-d_\theta}, b_1, \cdots, b_{d_\theta})|_{A'=e_{\mathbf{I}}} = 0. \end{cases} \tag{I10}$$

Solving these simultaneous partial differential equations gives the tangent vector $\frac{\partial \mathbf{a}(b_l)}{\partial b_l}|_{\mathbf{a}=\mathbf{a}_{\rm id}}$ of the manifold at $\mathbf{a}_{\rm id}$. Thus, if h_k can be regressed with the sampling result D_a as the polynomial of $\{b_l\}_{l=1}^{d_\theta}$, the conservation law can be inferred. Thus, we can estimate the infinitesimal transformation $(\delta\mathbf{q}_l, \delta\mathbf{p}_l)$ from the sampling result D_a.

Thus, in principle, the method can be applied to general coordinate transformations other than the matrix representation of the Lie group, which we describe in the main text. However, it is difficult to prepare a set of parametric functions $\{\mathbf{Q}(\cdot, \cdot; \mathbf{a}), \mathbf{P}(\cdot, \cdot; \mathbf{a}) | \mathbf{a} \in \mathbb{R}^{d_a}\}$ that contains the set of true invariant transformations $M_{\rm invariant}$ because the true invariant transformation set is unknown. Even if $\mathbf{Q}(\cdot, \cdot; \mathbf{a})$, $\mathbf{P}(\cdot, \cdot; \mathbf{a})$ could be prepared to include the true invariant transformations, it is not guaranteed that the function's parameters $\mathbf{a}$ satisfy the conditions to be parameters $\boldsymbol{\theta}$ of the Lie group, that is, the Jacobian $J_{kl} = \frac{\partial f_k(a_1,\cdots,a_{d_a})}{\partial b_l}$ becomes nonsingular. For example, a complex model with a high functional representation capability such as a DNN or a Gaussian process model would address the problem of including the true invariant transformations, but it is difficult to estimate the model parameters $\mathbf{a}$ of such a complex model from finite data. In addition, the number of parameters in such models can be enormous, which makes the computational cost extremely high to find the parameter set $\mathbf{b}$ at which the Jacobian J_{kl} becomes nonsingular.

APPENDIX J: LAPLACE–RUNGE–LENZ VECTOR AND SYMMETRY

Consider the motion of the central force potential in six-dimensional phase space: $(\mathbf{q}, \mathbf{p}) = (q_1, q_2, q_3, p_1, p_2, p_3)$. In this system, the Laplace–Runge–Lenz vector [52,53]

$$\vec{A} = \mathbf{p} \times L - mG\frac{\mathbf{q}}{||\mathbf{q}||_2}, \tag{J1}$$

$$L = \mathbf{q} \times \mathbf{p} \tag{J2}$$

is conserved. The Laplace–Runge–Lenz vector corresponds to the SO(4) symmetry in the coordinate space $(\tilde{\mathbf{q}}, \tilde{q}_4, \tilde{\mathbf{p}}, \tilde{p}_4) = (\tilde{q}_1, \tilde{q}_2, \tilde{q}_3, \tilde{q}_4, \tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \tilde{p}_4)$, defined as

$$\tilde{\mathbf{q}} = \tilde{\mathbf{q}}(\mathbf{q}, \mathbf{p}) := \frac{\mathbf{q}}{||\mathbf{q}||_2} - \frac{\mathbf{q} \cdot \mathbf{p}}{mG}\mathbf{p}, \quad \tilde{q}_4 = \tilde{q}_4(\mathbf{q}, \mathbf{p}) := \frac{p_0}{mG}\mathbf{q} \cdot \mathbf{p}, \tag{J3}$$

$$\tilde{\mathbf{p}} = \tilde{\mathbf{p}}(\mathbf{q}, \mathbf{p}) := \frac{2p_0\mathbf{p}}{p_0^2 + p^2}, \quad \tilde{p}_4 = \tilde{p}_4(\mathbf{q}, \mathbf{p}) := \frac{p^2 - p_0^2}{p_0^2 + p^2}, \tag{J4}$$

where $p_0 = \sqrt{-2mE}$. The transformed coordinate satisfies the conditions $\tilde{\mathbf{q}}^2 + \tilde{q}_4^2 = 1$, $\tilde{\mathbf{p}}^2 + \tilde{p}_4^2 = 1$, and $\tilde{\mathbf{q}} \cdot \tilde{\mathbf{p}} + \tilde{q}_4\tilde{p}_4 = 0$. Let us assume that the matrix representation of SO(4) is given by A. Moreover, assume the transformation is represented as $\tilde{\mathbf{q}}'^T = A\tilde{\mathbf{q}}^T$ and $\tilde{\mathbf{p}}'^T = A\tilde{\mathbf{p}}^T$.

We investigate the correspondence between the 4×4 matrix representation A of the SO(4) symmetry in $(\tilde{\mathbf{q}}, \tilde{\mathbf{p}})$ space and the coordinate transformation in $(\mathbf{q}, \mathbf{p})$ space. Because the inverse of the coordinate transformation is given by

$$\mathbf{q} = \mathbf{q}(\tilde{\mathbf{q}}, \tilde{q}_4, \tilde{\mathbf{p}}, \tilde{p}_4) = -\frac{G}{2E}[(1 - \tilde{p}_4)\tilde{\mathbf{q}} + \tilde{q}_4\tilde{\mathbf{p}}], \tag{J5}$$

$$\mathbf{p} = \mathbf{p}(\tilde{\mathbf{q}}, \tilde{q}_4, \tilde{\mathbf{p}}, \tilde{p}_4) = \sqrt{-2mE}\frac{\tilde{\mathbf{p}}}{1 - \tilde{p}_4}, \tag{J6}$$

the transformation of SO(4) in the original space becomes

$$\mathbf{Q}(\tilde{\mathbf{q}}, \tilde{q}_4, \tilde{\mathbf{p}}, \tilde{p}_4) = \mathbf{q}(\tilde{\mathbf{Q}}, \tilde{Q}_4, \tilde{\mathbf{P}}, \tilde{P}_4), \tag{J7}$$

$$\mathbf{P}(\tilde{\mathbf{q}}, \tilde{q}_4, \tilde{\mathbf{p}}, \tilde{p}_4) = \mathbf{p}(\tilde{\mathbf{Q}}, \tilde{Q}_4, \tilde{\mathbf{P}}, \tilde{P}_4), \tag{J8}$$

$$\begin{pmatrix} \tilde{\mathbf{Q}}^t \\ \tilde{Q}_4 \end{pmatrix} = A\begin{pmatrix} \tilde{\mathbf{q}}^t \\ \tilde{q}_4 \end{pmatrix}, \begin{pmatrix} \tilde{\mathbf{P}}^t \\ \tilde{P}_4 \end{pmatrix} = A\begin{pmatrix} \tilde{\mathbf{p}}^t \\ \tilde{p}_4 \end{pmatrix}. \tag{J9}$$

This is an example of symmetry realized by a nonlinear transformation function.

[1] S. Tomonaga, Prog. Theor. Phys. **5**, 544 (1950).
[2] D. Bohm and D. Pines, Phys. Rev. **82**, 625 (1951).
[3] D. Pines and D. Bohm, Phys. Rev. **85**, 338 (1952).
[4] S. Tomonaga, Prog. Theor. Phys. **13**, 467 (1955).
[5] P. G. Saffman, *Vortex Dynamics* (Cambridge University Press, Cambridge, 1992).

[6] T. Vicsek and A. Zafeiris, Phys. Rep. **517**, 71 (2012).
[7] T. Ikegami, Y. Mototake, S. Kobori, M. Oka, and Y. Hashimoto, Philos. Trans. R. Soc. A **375**, 20160351 (2017).
[8] M. Schmidt and H. Lipson, Science **324**, 81 (2009).
[9] S. Greydanus, M. Dzamba, and J. Yosinski, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., New York, 2019), pp. 15353–15363.
[10] P. Toth, D. J. Rezende, A. Jaegle, S. Racanière, A. Botev, and I. Higgins, in *International Conference on Learning Representations* (OpenReview.net, 2020).
[11] R. Bondesan and A. Lamacraft, arXiv:1906.04645.
[12] J. W. Gibbs, Trans. Conn. Acad. Arts Sci. **3**, 108 (1875–1876).
[13] J. W. Gibbs, Trans. Conn. Acad. Arts Sci. **3**, 343 (1877–1878).
[14] B. Irie and M. Kawato, Trans. Ins. Elec., Inf. Comm. Eng. D **J73-D2**, 1173 (1990).
[15] G. E. Hinton and R. R. Salakhutdinov, Science **313**, 504 (2006).
[16] P. P. Brahma, D. Wu, and Y. She, IEEE Trans. Neural Netw. Learn. Syst. **27**, 1997 (2016).
[17] R. Basri and D. W. Jacobs, in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017*, Conference Track Proceedings (OpenReview.net, 2017).
[18] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller, in *Advances in Neural Information Processing Systems 24*, edited by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Curran Associates, Inc., New York, 2011), pp. 2294–2302.
[19] Y. Mototake and T. Ikegami, in *International Symposium on Artificial Life and Robotics* (International Society of Artificial Life and Robotics, Oita, 2015).
[20] K. Yeo, arXiv:1710.01693.
[21] J. Morton, A. Jameson, M. J. Kochenderfer, and F. Witherden, in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., New York, 2018), pp. 9258–9268.
[22] S. H. Rudy, J. N. Kutz, and S. L. Brunton, J. Comp. Phys. **396**, 483 (2019).
[23] N. Takeishi, Y. Kawahara, and T. Yairi, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., New York, 2017), pp. 1130–1140.
[24] B. Lusch, J. N. Kutz, and S. L. Brunton, Nat. Commun. **9**, 4950 (2018).
[25] T. Ohtsuki and T. Ohtsuki, J. Phys. Soc. Jpn. **85**, 123706 (2016).
[26] T. Ohtsuki and T. Ohtsuki, J. Phys. Soc. Jpn. **86**, 044708 (2017).
[27] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, Sci. Rep. **7**, 8823 (2017).
[28] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, Phys. Rev. X **7**, 031038 (2017).
[29] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 431 (2017).
[30] A. Tanaka and A. Tomiya, J. Phys. Soc. Jpn. **86**, 063001 (2017).
[31] H. Saito and M. Kato, J. Phys. Soc. Jpn. **87**, 014001 (2017).
[32] E. P. Van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Nat. Phys. **13**, 435 (2017).
[33] P. Zhang, H. Shen, and H. Zhai, Phys. Rev. Lett. **120**, 066401 (2018).
[34] A. Noether, Nachr. Ges. Wiss. Goettingen Math. Phys. Kl. **235** (1918).
[35] C. W. Reynolds, in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '87* (Association for Computing Machinery, New York, 1987), pp. 25–34.
[36] J. Struckmeier and C. Riedel, Ann. Phys. **11**, 15 (2002).
[37] L. D. Landau and E. M. Lifshitz, *Mechanics: Volume 1* (Butterworth-Heinemann, Oxford, 1976).
[38] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[39] M. O. Ulfarsson and V. Solo, IEEE Trans. Signal Process. **56**, 5804 (2008).
[40] E. Levina and P. J. Bickel, in *Advances in Neural Information Processing Systems 17*, edited by L. K. Saul, Y. Weiss, and L. Bottou (MIT Press, Cambridge, 2005), pp. 777–784.
[41] P. T. Boggs and J. E. Rogers, Contemporary Mathematics **112**, 183 (1990).
[42] G. Schwarz, Ann. Stat. **6**, 461 (1978).
[43] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks, J. Theor. Biol. **218**, 1 (2002).
[44] D. P. Kingma and M. Welling, arXiv:1312.6114.
[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., New York, 2014), pp. 2672–2680.
[46] W. K. Hastings, Biometrika **57**, 97 (1970).
[47] K. Nagata and S. Watanabe, Neural Netw. **21**, 980 (2008).
[48] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio, Theano: New features and speed improvements, presented at Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012, arXiv:1211.5590.
[49] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, in *Proceedings of the Python for Scientific Computing Conference (SciPy) (2010)* (oral presentation), http://www.iro.umontreal.ca/~lisa/pointeurs/theano_scipy2010.pdf.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).
[51] D. P. Kingma and J. Ba, arXiv:1412.6980.
[52] H. H. Rogers, J. Math. Phys. **14**, 1125 (1973).
[53] A. Alemi, Laplace-runge-lenz vector, http://www.cds.caltech.edu/~marsden/wiki/uploads/projects/geomech/Alemicds205final.pdf.