Questioning the question: Exploring how physical degrees of freedom are retrieved with neural networks

Joeri Lenaerts [*] and Vincent Ginis [†]
Data Lab/Applied Physics, Vrije Universiteit Brussel, Elsene 1050, Belgium

(Received 11 February 2022; accepted 3 May 2022; published 13 June 2022)

When studying a physical system, it is crucial to identify the degrees of freedom that characterize that system. Recently, specific neural networks have been designed to retrieve these underlying degrees of freedom automatically. Indeed, fed with data from a physical system, a variational autoencoder can learn a latent representation of that system that directly corresponds to its underlying degrees of freedom. However, the understanding of these neural networks is limited on two fronts. First, very little is known about the impact of the question vector, a key parameter in designing performant autoencoders. Second, there is the mystery of why the correct degrees of freedom are found in the latent representation, not an arbitrary function of these parameters. Both gaps in our understanding are addressed in this paper. To study the first question on the optimal design of the question vector, we investigate physical systems characterized by analytical expressions with a limited set of degrees of freedom. We empirically show how the type of question influences the learned latent representation. We find that the stochasticity of a random question is fundamental in learning physically meaningful representations. Furthermore, the dimensionality of the question vector should not be too large. To address the second question, we make use of a symmetry argument. We show that the learning of the degrees of freedom in the latent space is related to the symmetry group of the input data. This result holds for linear and nonlinear transformations of the degrees of freedom. In this way, in this paper, we contribute to the research on automated systems for discovery and knowledge creation.

DOI: 10.1103/PhysRevResearch.4.023206

I. INTRODUCTION

A crucial task in physics is identifying the degrees of freedom of the system under investigation based on experimental data [1]. When observing data in a time series, e.g., representing planetary motion [2], one typically wants to extract the underlying physical parameters that give rise to these dynamics. In physics, this is often accomplished using first-principles arguments [3], symmetry considerations [4], or analogies with other theoretical models [5]. In the context of machine learning, parameter extraction often goes under the name of dimensionality reduction, the action of reducing an input time series to a minimal representation that describes it.

Recently, the neural network architecture SciNet was proposed to discover the underlying degrees of freedom in physical data [6]. SciNet is based on a popular generative model in the machine learning literature called the β-variational autoencoder (β-VAE). This generative model creates a latent representation of the input data that captures all relevant information. To make the latent representation physically relevant, a so-called question was introduced in the architecture of SciNet.

For time-series input data, e.g., the position of a damped harmonic oscillator, this question is the specific time points for which we want to reconstruct the time series. The output of the network then corresponds to the time series at the time point(s) encoded by the question(s). While previous work showed good results in learning the degrees of freedom, it remained however elusive how variations of this question influence the latent representation that is learned. Moreover, it is unclear why the latent representation would store exactly the degrees of freedom and not an arbitrary function of them.

In this paper, we show that the introduction of a question is crucial to learning a physically meaningful latent representation. Without an added question, the standard architecture of a β-VAE, where the original time series input is reconstructed, does not lead to a physically interpretable representation in the latent space. A careful design of the question is thus key in giving physical meaning to the latent representation formed in neural networks.

To study what is learned in the latent representation, we take a look at a symmetry argument given in Ref. [7]. The argument proposes that we want to learn a disentagled representation. This is a representation of which the parts transform independently under the subgroups of the symmetry group of the input data. We show that, using this symmetry argument, we indeed expect to find the degrees of freedom. Our results

[*] joeri.lenaerts@vub.be
[†] Also at School of Engineering and Applied Sciences, Harvard University; ginis@seas.harvard.edu

show that a β-VAE can learn disentangled representations of physical systems.

This paper is divided as follows. Section III covers the architecture that we use and its learning dynamics. It also explains how the architecture changes for different questions. Section IV shows the results when training architectures with different questions on three datasets of physical time series. These datasets are generated from analytical expressions. The results are further discussed in Sec. V. The final Sec. VI contains our conclusions and perspectives.

II. RELATED WORK

A significant number of scientists and engineers are working on the interface of physics and artificial intelligence (AI), intending to create an AI system for knowledge discovery. Such a system could understand experimental data and formulate theories and analytical expressions describing them. Exciting work in this area includes the development of the AI physicist [8]. This learning agent can discover the dynamics of several so-called mystery worlds with various physical interactions such as gravitational or electromagnetic fields. More recently, the AI Feynman was created, an algorithm that can discover analytical expressions based on data of complex functions [9]. In both previous use cases, it is essential to find the degrees of freedom, a task sometimes referred to as symbolic pregression. Two approaches have been taken to accomplish this task. The first approach uses an autoencoder that is regularized by a loss penalizing the nonlinearity of the autoencoder [10]. The second approach regularizes the autoencoder according to the rules of a β-VAE [6]. This is the approach we elaborate on in this paper.

Since the occurrence of a VAE [11], many variations of this architecture have been proposed such as the β-VAE [12], AnnealedVAE [13], FactorVAE [14], β-TCVAE [15], and DIP-VAE-I [16]. For a comprehensive overview, see Ref. [17]. The goal of these variations is to learn more disentangled representations, where each dimension is independent of the other. In physical systems, this exactly corresponds to learning a degree of freedom in each latent dimension. This has been applied to classical and quantum systems [6] as well as dark matter research [18].

The VAE is not the only neural network architecture that can be used to uncover physical concepts. An architecture called a restricted Boltzmann machine was used to extract the relevant degrees of freedom in a classical statistical mechanics system [19]. Moreover, classical fully connected neural networks were used to discover different phases of matter [20] and to discover the concept of a quantum mechanical wave function and the Schrödinger equation it obeys [21]. In addition, fully connected neural networks were used to extract interpretable physical parameters from spatiotemporal data of a partial differential equation [22] and to find an analytical expression for dark matter dynamics [23]. In a final example, fully connected neural networks were used to directly learn the conserved quantities from a time series [24–26].

A lot of progress has also been made extracting dynamical equations from experimental data [27–31], but so far, these methods still require prior domain knowledge. In contrast, the β-VAE, discussed in this paper, does not need prior knowledge on the physics of the system.

III. METHODS

The neural network architecture in this paper consists of two parts, the encoder and the decoder, as shown in Fig. 1. Both are fully connected neural networks. The encoder maps an input $\mathbf{x}$ to a vector $\mathbf{z}$ that we call the latent representation. The aim is to store the degrees of freedom underlying the input $\mathbf{x}$ in the different nodes of the latent layer $\mathbf{z}$. Instead of learning this vector directly, we learn a probabilistic encoder $q(\mathbf{z}|\mathbf{x})$. This ensures that we have a continuous and smooth latent space, in which the interpolation between two points corresponds to a meaningful input.

To learn this probability distribution, we need to restrict it to a fixed family of functions. A popular family of functions is the set of multivariate Gaussians, for which every dimension of the latent vector $\mathbf{z}$ corresponds to a normal distribution $\mathcal{N}(\mu, \sigma)$. The encoder returns the parameters μ and σ of this distribution.

For the decoder to make a reconstruction, we need to sample a latent vector $\mathbf{z}$ from the probability distribution $q(\mathbf{z}|\mathbf{x})$. Since the sampling operation is not differentiable, we use the reparameterization trick: instead of sampling directly, we draw a random vector ϵ from $\mathcal{N}(\mathbf{0}, \mathbf{1})$. The latent vector $\mathbf{z}$ is then given by

$$\mathbf{z} = \mu + \epsilon \cdot \sigma. \tag{1}$$

Crucially, at this point, the representation $\mathbf{z}$ is concatenated with a question vector $\mathbf{q}$ that contains the time points at which we want to evaluate the time series. The decoder then maps the question $\mathbf{q}$ and the latent representation $\mathbf{z}$ to an output $\mathbf{y}$ that is of the same dimension as the question—the output vector needs to encode the answer to the question vector using the information stored in the latent representation. The introduction of the question allows us to control what is learned in the latent representation. It forces the latent representation to maximally store physically relevant information that can be used to predict the time series at different time points from the question vector.

Once the neural network is set up, we train it by minimizing a loss function called the ELBO, given by

$$\text{ELBO} = (\mathbf{y} - \hat{\mathbf{y}})^2 + \beta \cdot D_{\text{KL}}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]. \tag{2}$$

The first term is the mean squared error between the real output $\mathbf{y}$ and the output $\hat{\mathbf{y}}$ generated by the decoder. Minimizing this term leads to an accurate reconstruction of the time series at the time points of the question $\mathbf{q}$. The second term is a KL divergence between the distribution over the latent representation and a prior Gaussian $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. It regularizes the representation such that it is minimal by forcing the architecture to activate a latent dimension only when it provides useful information. If there is no useful information to be stored in an extra dimension, the values in a dimension are zero.

The last and perhaps most important question is how to design the question vector $\mathbf{q}$. In the general literature on autoencoders, the goal is usually to make an accurate reconstruction of the input. We can achieve this by constructing the

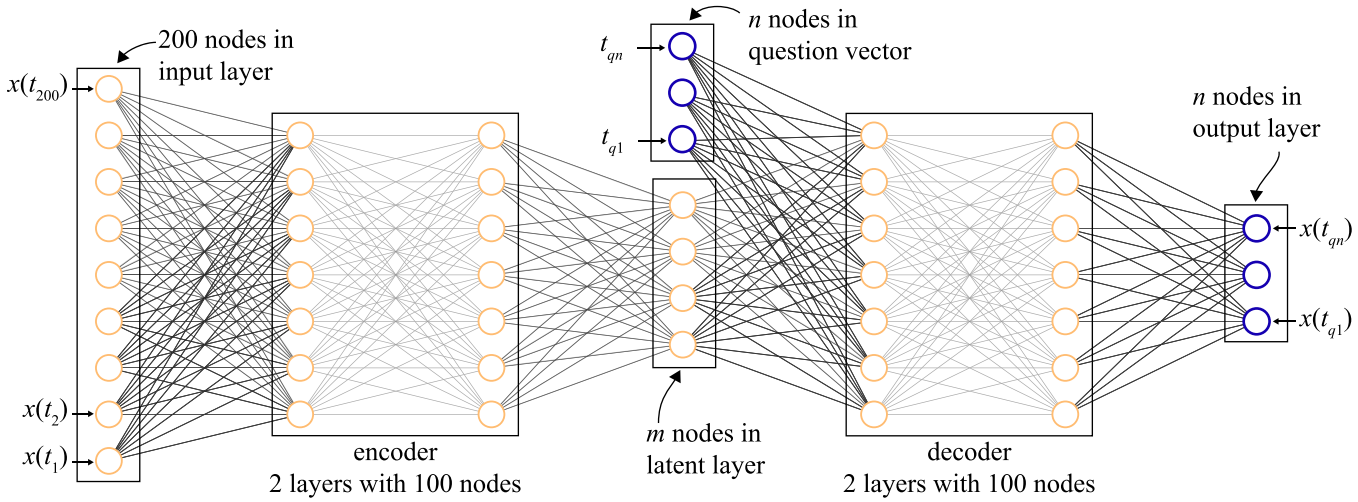

FIG. 1. The architecture of a β-VAE consists of an encoder and a decoder. The encoder maps an input to a latent representation of dimension m. The degrees of freedom appear in this latent representation. Then a question of dimension n is concatenated with the representation and fed to the decoder. This results in an output of dimension n that contains the answer to the question. The encoder and decoder in our experiments are fully connected neural networks with two hidden layers each of 100 nodes and the Swish activation function.

question vector in agreement with the fixed sampling time points of the input. The time series at these sampling points is then exactly the original input.

The other option explored in Ref. [6] is to ask a question at one random time point. This random time point is different for each training sample. We see thus that we have a choice of randomly sampling time points at which we evaluate or having fixed points that are the same for every data sample. The number of time points, or in other words the dimension of the question $\mathbf{q}$, is also something we can vary.

We explore two adaptations of the question vector. The first is to change the dimension of the question. We evaluate the time series at time points a fixed distance apart but at lower dimension than the input. The second way to adapt the question vector is to add stochasticity. Instead of asking a question at fixed time points, we allow for random time points. The question vector can thus be different for every data sample. We can also change the question dimension in this stochastic case. The adaptation explored in Ref. [6] used a stochastic question vector of dimension one. In this paper, we rigorously investigate how these choices influence the learning of the latent representation.

IV. RESULTS

To study the impact of the question that we add to the latent representation, we train several neural networks with a different question vector $\mathbf{q}$. First, we make the comparison between a question asked at fixed time intervals, i.e., the question vector contains uniformly spaced time coordinates identical for all data samples, and a question vector that contains random time points, different for every data sample. Second, we change the dimension of the question vector, where we go from the limit of asking a question containing only one time point to the case of asking a question of the same number of time points as the input vector.

The neural networks are trained on three different datasets, created from analytical expressions. These expressions include a unity amplitude sine wave with the frequency f as the only degree of freedom, a sine wave with both amplitude A and f as degrees of freedom, and a damped harmonic oscillator with two degrees of freedom k and b.

A. Sine wave with one degree of freedom

The first dataset consists of sine waves defined as

$$y = \sin(ft), \qquad \text{for } t \in [0, 2\pi]. \tag{3}$$

The frequency $f \in [2, 3.5]$ is the single degree of freedom in this dataset. We evaluate the sine functions for 200 evenly spaced values of $t \in [0, 2\pi]$, creating a dataset of 100 000 sine functions. Some example sine functions are shown in Fig. 2(a).

The encoder and decoder are both neural networks with two hidden layers with 100 neurons and the Swish activation function. We use the Adam optimizer with a learning rate of 10^{-3}, batch size of 250, and $\beta = 0.001$.

We train different architectures where the question is either evaluated on fixed time points that are the same for the full dataset or on random points in the time interval, different for every sample in the dataset. We do this for different dimensions of the question vector $\mathbf{q}$ that we vary from 1 up to 200, the size of the input vector.

To test the performance of the trained neural networks, we divide the dataset of 100 000 sine waves into a training set, a validation set, and a test set, following a 80-10-10 split. The neural network uses the training set for actual training, the validation set is used to tune hyperparameters, and the test set is used to report the final performance.

The results of the different architectures are shown in Figs. 3(a), 3(d), 3(g), and 3(h), where we plot the mean absolute error on the test set (test MAE), the final KL divergence, and the number of active dimensions. A dimension is active when μ_i is not zero for all test samples, formally when $\mathrm{Cov}(\mu_i) > \delta$, where δ is a threshold value set to 0.001. The covariance is computed over all values of μ_i that are obtained

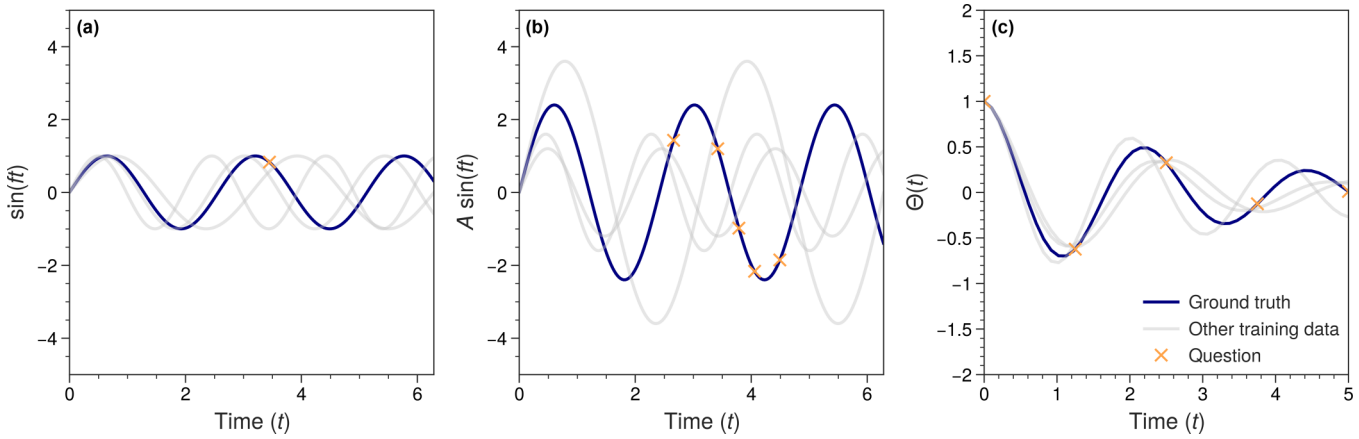

FIG. 2. Three datasets with analytical expressions parameterized by a limited number of degrees of freedom. (a) The first dataset contains sine waves with variable frequency $f \in [2, 3.5]$. Different waves are shown in gray. The gold cross indicates the location of the random question. (b) The second dataset contains sine waves with variable frequency $f \in [2, 3.5]$ and variable amplitude $A \in [1, 4]$. The gold crosses indicate the location of the random question of dimension five. (c) The third dataset contains instances of a damped harmonic oscillator with variable parameters $k \in [5, 10]$ and $b \in [0.5, 1]$. The gold crosses indicate the location of the fixed question of dimension five.

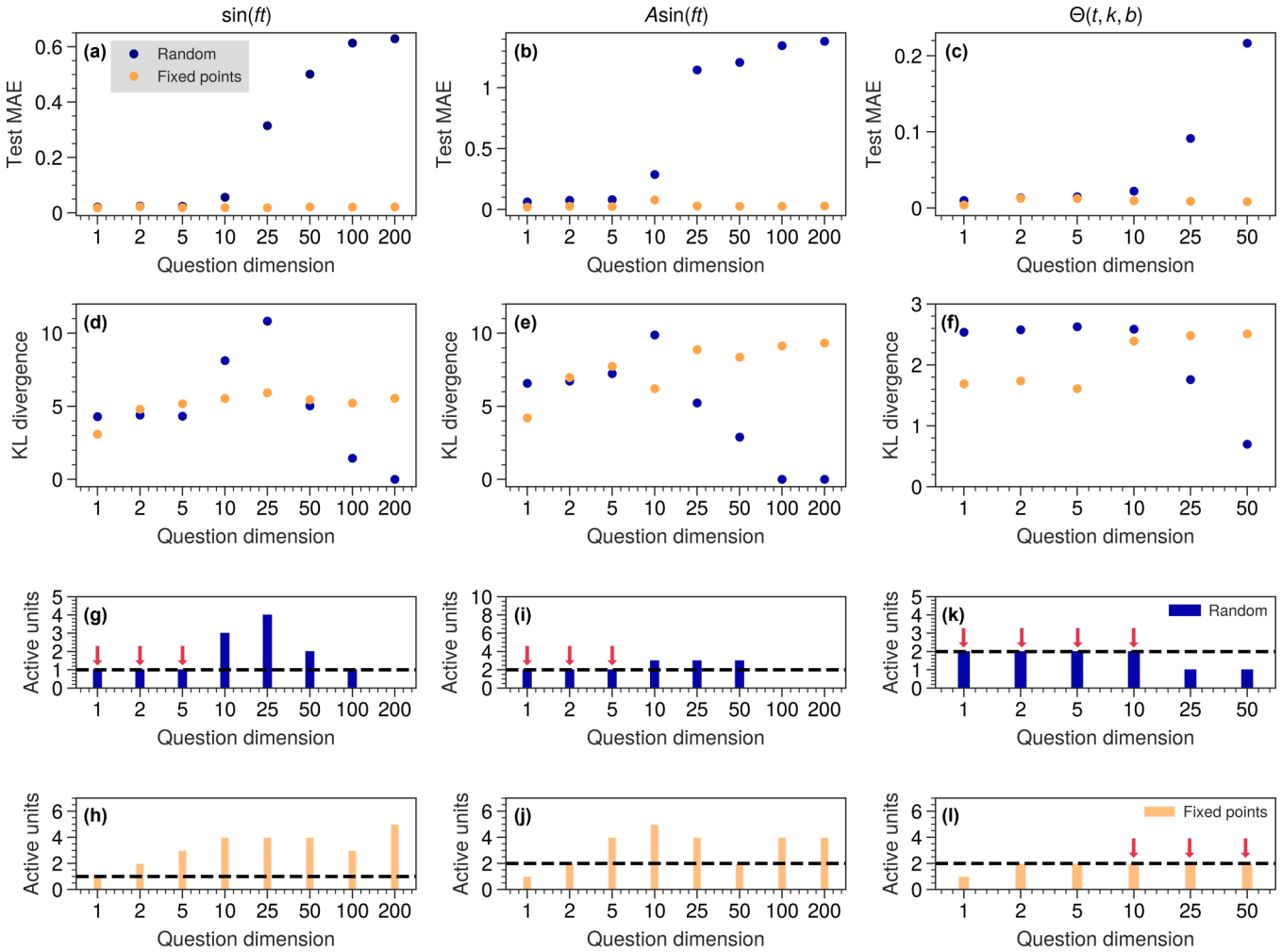

FIG. 3. Training results for different datasets. (a)–(c) Mean absolute error on the test set (test MAE) as a function of the dimension of the question asked. We compare the cases where the question consists of random points in the time interval (blue) or points that are a fixed distance apart in the time interval (yellow). (d)–(f) Final KL divergence on the training set vs the dimension of the question asked. (g)–(l) Number of active dimensions in the latent space. A dimension is activated when $\mathrm{Cov}(\mu_i) > \delta$ for the test set, where μ_i is the mean of the latent dimension and the threshold $\delta = 0.001$. A red arrow is drawn when the degrees of freedom are successfully learned in the active latent dimensions.

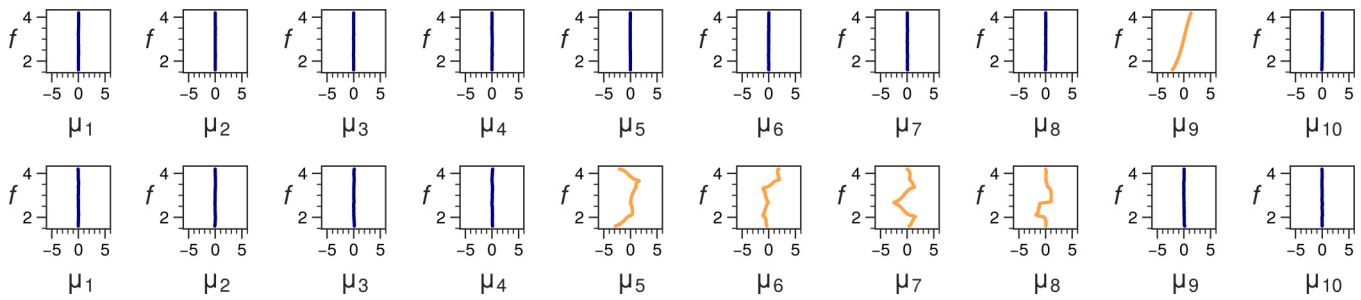

FIG. 4. Scatter plot between the dimensions of the latent space and the degree of freedom f for neural networks trained on the dataset $\sin(ft)$. The first row shows the latent representation for a network trained with a question of dimension five, evaluated at random points on the time interval. There is one active dimension, indicated in yellow. This dimension shows an approximately linear correlation with the degree of freedom f, which indicates that f is stored inside this latent variable. The second row shows the latent representation for a network trained with a question of dimension 25, evaluated at evenly spaced points on the time interval. There are four active dimensions, indicated in yellow. While all active dimensions show a relation with f, this relation is not monotonous. Therefore, we cannot say that f is stored in any of these dimensions; the information is spread out across the four dimensions.

for the test set. All architectures have 10 latent dimensions in which information can be stored. The red arrows in Fig. 3(g) mark the neural networks where the latent representation corresponds to the underlying degree of freedom f.

We see that a good latent representation can be learned for a random question of dimensions 1, 2, or 5. For these cases, we also obtain a test MAE <0.03. For higher dimensions of the random question vector, the test MAE gets progressively worse. We also see an increase in the number of latent dimensions, such that we get a nonminimal latent representation. While the latent dimensions are still correlated with f, there is no clear linear relationship.

For a question at fixed time points, the test MAE is <0.03 for all numbers of output nodes. The information about the sine wave is stored inside the latent representation, but there is no single dimension containing f. This is shown in Fig. 4. The first row shows a good latent representation for a neural network with one random question of dimension five. One active dimension, indicated in yellow, has a linear correlation with the frequency. The second row shows the latent dimensions for a neural network with a fixed question of dimension 25. While the neural network can make sound predictions of the time series, four active dimensions are formed to store some of the information on the frequency f.

B. Sine wave with two degrees of freedom

The second dataset consists of sine functions given by

$$y = A\sin(ft), \qquad \text{for } t \in [0, 2\pi]. \tag{4}$$

The frequency $f \in [2, 3.5]$ and the amplitude $A \in [1, 3]$ are the two degrees of freedom in this dataset. These sine functions are evaluated for 200 evenly spaced values of $t \in [0, 2\pi]$, leading to 100 000 sine functions. Some example sine functions are given in Fig. 2(b).

We use the same hyperparameters for the neural network as for the previous dataset. However, the training process for this dataset was not always stable, leading to predictions of not a number. To stabilize the training, we optimized the hyperparameter β. The random question of dimensions 1, 2, and 5 was trained with $\beta = 0.005$, and the random question of dimensions 10, 25, 50, 100, and 200 was trained with $\beta = 0.01$. The fixed questions were trained with $\beta = 0.001$, except for the fixed question of dimension 10, trained with $\beta = 0.01$.

Results for architectures with a different question vector are shown in Figs. 3(b), 3(e), 3(i), and 3(j). We retrieve the degrees of freedom for a random question dimension 1, 2, or 5. For these questions, we also find a low test MAE <0.03. For random questions of dimension 10, 25, 50, 100, or 200, the degrees of freedom are not found in latent space. Furthermore, the test MAE gets progressively worse, so these networks cannot give good predictions of the time series.

The degrees of freedom are also not found in the latent space for the fixed questions. We do, however, have good predictions of the time series. The active dimensions of the latent space thus capture the information in A and f but spread them over multiple dimensions in a nonlinear way. This is like what happens for the first dataset, where the frequency f is spread out over four active dimensions. A plot of the latent space for A and f can be found in the Supplemental Material [32].

C. Damped harmonic oscillator

The third dataset represents the damped harmonic oscillator. This dataset was first studied using neural networks in Iten *et al.* [6]. We use the code of this paper provided on Github for further experiments. The analytical expression of the oscillator is given by

$$\Theta(t, k, b) = \exp\left(\frac{-b}{2t}\right)\cos\left(\sqrt{\kappa - \frac{b^2}{4}}t\right). \tag{5}$$

The parameters $\kappa \in [5, 10]$ and $b \in [0.5, 1]$ are the two degrees of freedom of the system. We evaluate the time series for 50 evenly spaced values of $t \in [0, 5]$, creating a dataset of 100 000 time series. Some examples for the damped harmonic oscillator are given in Fig. 2(c).

The architecture and hyperparameters are identical to the ones used for the first dataset. In the scenario where we construct the question at random time points, we retrieve the degrees of freedom for a question dimension of 1, 2, 5, or 10, as shown in Figs. 3(c), 3(f), 3(k), and 3(l). The MAE on

the test set is $\lesssim 0.02$, which makes for a good reconstruction. Upon closer inspection of the two active dimensions, we find that one of these dimensions has a linear correlation with k, while the other has a linear correlation with b. Note that this reproduces and generalizes the results of SciNet, which has a random question of dimension one [6].

The results get increasingly worse for a question dimension of 25 or 50. The test MAE increases, while the KL divergence decreases, corresponding to a decrease in information captured in the latent representation. There is only one active dimension for this case. This dimension correlates linearly with k. We thus only retrieved one degree of freedom.

As revealed by a low test MAE, we have a good reconstruction for a question vector at fixed time points of dimensions 1, 2, or 5. However, looking at the 1 or 2 active dimensions, we find that they are correlated nonlinearly with both k and b. The information is contained as a combination of both degrees of freedom, in the same way as we saw that the info on f was spread out over four active dimensions for the first dataset.

For a question dimension of 10, 25, or 50, we find k and b in separate active dimensions. The test MAE is again very low. We also see that the KL divergence is higher than for question dimensions 1, 2, and 5, showing an increase in information contained in the latent representation. A plot of the latent spaces for the harmonic oscillator is provided in the Supplemental Material [32].

The claim that b and κ are found in the latent space is supported by the linear correlation between the mean of the latent dimension and these parameters. An important question is why exactly these parameters are found and no others. We can define the frequency of the damped harmonic oscillator $\omega = \sqrt{\kappa - \frac{b^2}{4}}$. It turns out that, if we plot this parameter ω together with the dimension in which κ is found, we also find a linear relation. We can thus equally well say that ω was found. According to the formal definition of a disentangled representation, given in Ref. [7], one would also expect to retrieve this parameter ω based on a symmetry argument.

To check whether κ or ω is really learned, we can compare the Pearson correlation coefficients. Taking the average over the networks that found the degrees of freedom, we find a Pearson correlation between the active dimension and k of 0.998 and between the active dimension and ω of 0.999. While both Pearson correlations are very high, we conclude that not κ but ω is found in this latent space. This insight refines earlier results [6].

V. DISCUSSION

Both the stochasticity of the question and being of a low dimension are important to learn a good latent representation. We can get intuition as to why this is so by considering the other extreme: a fixed (nonstochastic) question vector with the same dimensionality as the input layer. This architecture can hardly add information compared with the original β-VAE, an architecture that fails to structurally recover degrees of freedom. One can say that constructing a question vector at fixed points of lower dimension does not make a large difference. Correct predictions can still be achieved with a traditional β-VAE and only outputting a specific part of the input.

Asking a random question thus improves the learning of the latent representation because it adds stochasticity to the output. This is an essential feature beyond a typical β-VAE operation: the neural network needs to capture the full information hidden in the data in its latent space. When the question vector has a lower dimension, the stochasticity is more pronounced because the stochasticity is averaged out when the dimension is too high. We thus need to add a question of low dimension to benefit from the stochasticity fully.

An interesting question is why we find precisely the degrees of freedom we are looking for. In principle, the latent space could store an arbitrary function of the degrees of freedom since neural networks can learn this function in the encoder and the inverse function in the decoder. The answer can be found in the fact that the degrees of freedom form a disentangled representation. A formal definition of a disentangled representation is given in Ref. [7]: "A vector representation is called a disentangled representation with respect to a particular decomposition of a symmetry group into subgroups if it decomposes into independent subspaces, where each subspace is affected by the action of a single subgroup, and the actions of all other subgroups leave the subspace unaffected."

This definition allows us to describe what the disentangled representation is for the datasets described. The dataset of sine functions is parameterized by an amplitude A and frequency f. All sine functions in this dataset are related by the symmetry group of horizontal and vertical scalings. It is important to note that this is a symmetry group of the underlying two-dimensional Euclidean space on which the graphs are defined. However, it is not a proper symmetry of the dataset. Since the sampled data were given on a large enough region of input space, an approximate symmetry is found by the neural network. The symmetry group decomposes in scalings along the x and y axes. The scalings in the x direction only affect the frequency f, and the scalings in the y direction only affect the amplitude A. The degrees of freedom transform independently under the subgroups of the scaling symmetry; hence, they are the unique disentangled representation we want to find.

However, there is no particular reason why a β-VAE should learn this disentangled representation. To verify if this architecture is preferably finding disentangled representation instead of others, we perform an experiment on a similar dataset of sine functions with a different parametrization by C_1 and C_2. We construct a dataset using the following function:

$$y = \frac{C_1 + C_2}{2} \sin\left(\frac{C_1 - C_2}{2} t\right), \qquad \text{for } t \in [0, 2\pi]. \tag{6}$$

We have that $C_1 \in [3.5, 4.5]$ and $C_2 \in [-1.5, -0.5]$, leading to a dataset of 100 000 sine waves. Looking at this function, one would expect to retrieve C_1 and C_2 in the latent space. However, we find, for an architecture of a random question with dimension 1, 2, and 5, that C_1 and C_2 are not retrieved in the latent space. Results for a question of dimension five are shown in Fig. 5. We can explain this by making the following identification:

$$\frac{C_1 + C_2}{2} = A, \qquad \frac{C_1 - C_2}{2} = f. \tag{7}$$

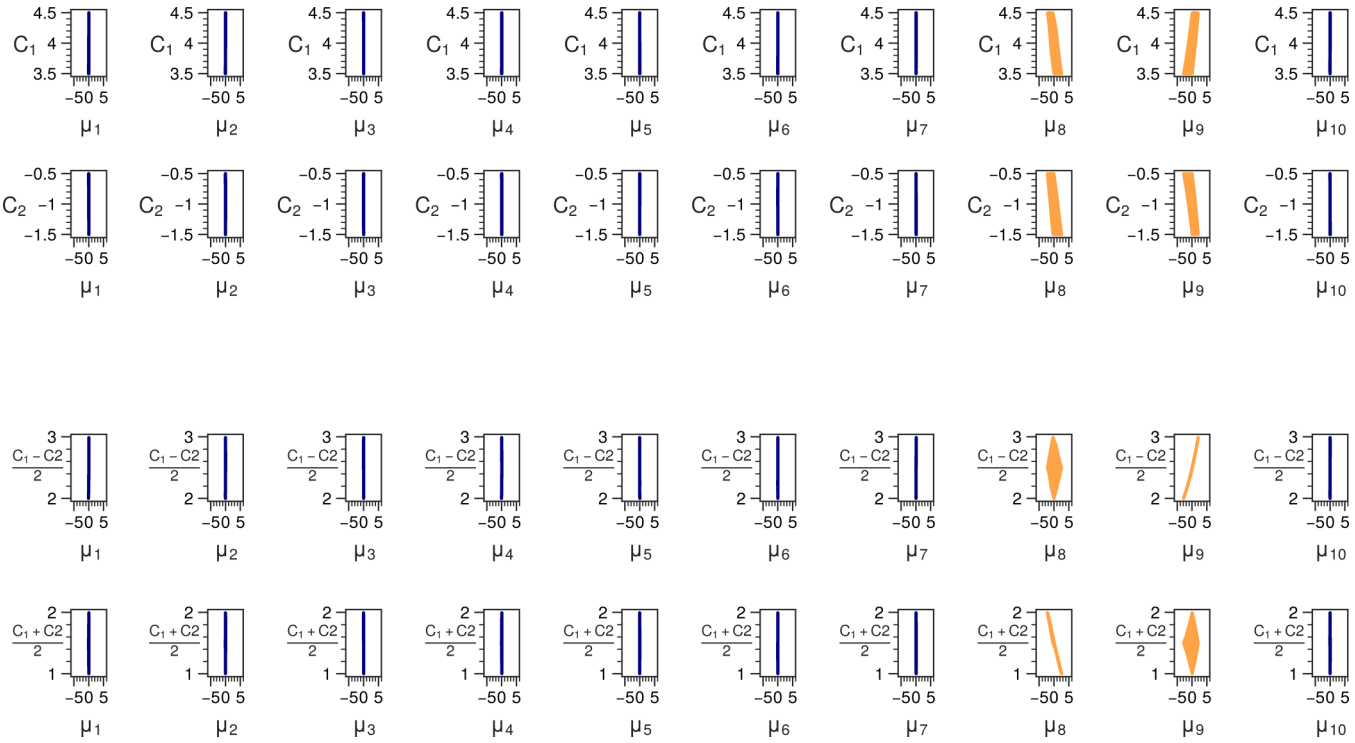

FIG. 5. Scatter plot between the dimensions of the latent space and the degrees of freedom C_1 and C_2 for the neural network trained on a modification of the second dataset. Results are shown for a network trained with a question at random time points of dimension five. The first two rows show the correlation with C_1 and C_2. There are two active dimensions, indicated in yellow. The absence of a clear correlation indicates that the degrees of freedom are not found in the latent representation. The last two rows show the correlation with $\frac{C_1-C_2}{2}$ and $\frac{C_1+C2}{2}$. There are two active dimensions indicated in yellow. A linear relation can be observed here. This shows that the latent representation stores a linear combination of C_1 and C_2. It is exactly this linear representation that forms a disentangled representation under the symmetry group of the dataset.

These two combinations then form a disentangled representation of the sine functions. When we plot the correlations with these linear combinations of C_1 and C_2, we see that they are retrieved in the latent space. This shows that, even though the dataset was created by varying C_1 and C_2, we retrieve the disentangled representation of $\frac{C_1+C_2}{2}$ and $\frac{C_1-C_2}{2}$.

The interesting question now arises whether the disentangled representation can be learned after being nonlinearly transformed. Therefore, we introduce two new datasets with

$$\begin{aligned} y &= \sqrt{D_1}\sin(D_2 t), \qquad \text{for } t \in [0, 2\pi]. \\ y &= (E_1)^2 \sin(E_2 t), \qquad \text{for } t \in [0, 2\pi]. \end{aligned} \tag{8}$$

The latent spaces for these datasets are shown in Fig. 6. We plot correlations between the latent dimensions and both the original variable D_1, D_2, E_1, and E_2 and the disentangled representations $\sqrt{D_1}$, D_2, $(E_1)^2$, and E_2. While there is a good linear correlation with both sets of variables, closer inspection using the Pearson correlation coefficient reveals that the correlation is higher for the disentangled representations, shown in Table I. Results show that a disentangled representation is retrieved, even when starting with a dataset where the degrees of freedom are nonlinearly transformed.

We now turn to the third dataset of the harmonic oscillator. The formula is given by

$$\Theta(t, k, b) = \exp\left(\frac{-b}{2t}\right)\cos\left(\sqrt{\kappa - \frac{b^2}{4}}\,t\right). \tag{9}$$

Comparing this to the previous sine dataset, we see that the functions of the damped harmonic oscillator can also be related by horizontal and vertical scalings. The vertical scalings correspond to the parameter b, and the horizontal scalings correspond to the frequency $\sqrt{\kappa - \frac{b^2}{4}} = \omega$ of the cosine.

We showed earlier that b and ω were found in the latent representation, even though we expected to find b and κ. The disentangled representation is thus retrieved by the β-VAE in the case of the damped harmonic oscillator. It provides an

TABLE I. Pearson correlation between latent dimensions and degrees of freedom. A high correlation means that the degree of freedom is stored in the latent dimension.

Dimension	C_2	$\frac{C_1+C_2}{2}$	Dimension	C_1	$\frac{C_1-C_2}{2}$
μ_8	−0.701	−0.997	μ_9	0.700	0.997
Dimension	D_1	$\sqrt{D_1}$	Dimension	E_1	$(E_1)^2$
μ_2	0.992	0.995	μ_4	−0.995	−0.997

FIG. 6. Scatter plot between the dimensions of the latent space and the degrees of freedom D_1, D_2, E_1, and E_2 for the neural network trained on a nonlinear modification of the second dataset. The first four rows show results for network trained with a question at a random time point of dimension one. The first two rows show the correlation with D_1 and D_2, and the following two rows correlation with the parameters $\sqrt{D_1}$ and D_2 of the disentangled representation. There are two active dimensions, indicated in yellow. The following four rows show results for network trained with a question at a random time points of dimension five. The first two rows show the correlation with E_1 and E_2, and the following two rows correlation with the parameters $(E_1)^2$ and E_2 of the disentangled representation. There are two active dimensions, indicated in yellow.

example of how the symmetry argument can give insight into the latent parameters that are learned.

VI. CONCLUSIONS

Neural networks can retrieve the underlying physical parameters driving the dynamics of a time series. The architecture consists of an encoder that maps the time series to a latent representation and a decoder that maps the representation to an output. For the physical degrees of freedom to be learned in the latent representation, it is crucial to introduce a question in the neural network architecture. The question consists of different time points at which the time series is evaluated. The output of the decoder is then the answer to this question.

Applying this to three datasets, we have shown that this question needs to be asked at random time points that are different for every data sample. The degrees of freedom in both sine datasets can only be retrieved when the question is asked at a low number of random time points for each data sample. A higher number of time points does not lead to a good representation nor to a good prediction of the sine. The added stochasticity of the random question is fundamental in learning physically meaningful representations.

In this paper, we also test a symmetry argument to better understand what is learned in the latent space. Our results support the claim that a disentangled representation is learned in the latent space. The results are valid for linear as well as nonlinear transformations. In our case, the disentangled representation consists of exactly the degrees of freedom of the physical systems.

Looking forward, we have shown that asking a stochastic question to a latent representation can give it a physical meaning. These efforts lead the way to better and more interpretable AI systems that can be used for autonomous knowledge generation of physical systems.

ACKNOWLEDGMENTS

J.L. acknowledges a fellowship from the Research Foundation Flanders (FWO-Vlaanderen) under Grant No. 11G1621N. Work at VUB was partially supported by the Research Foundation Flanders under Grant No. G032822N.

[1] L. Infeld, Leonardo da Vinci and the fundamental laws of science, Sci. Soc. **17**, 26 (1953).

[2] A. Koyré, *The Astronomical Revolution: Copernicus-Kepler-Borelli* (Routledge, London, 2013).

[3] A. Einstein, Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies], Ann. Phys. **322**, 891 (1905).

[4] E. Noether, Invariante variationsprobleme, Nachr. Ges. Wiss. Goettingen, Math. Phys. Kl. 235 (1918).

[5] C. Wang, Y. Chong, J. D. Joannopoulos, and M. Soljacic, Observation of unidirectional backscattering-immune topological electromagnetic states, Nature (London) **461**, 772 (2009).

[6] R. Iten, T. Metger, H. Wilming, L. Del Rio, and R. Renner, Discovering Physical Concepts with Neural Networks, Phys. Rev. Lett. **124**, 010508 (2020).

[7] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, Towards a definition of disentangled representations, arXiv:1812.02230.

[8] T. Wu and M. Tegmark, Toward an artificial intelligence physicist for unsupervised learning, Phys. Rev. E **100**, 033311 (2019).

[9] S. M. Udrescu, A. Tan, J. Feng, O. Neto, T. Wu, and M. Tegmark, AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity, in *Advances in Neural Information Processing Systems*, Vol. 33 (Curran Associates, Inc., 2020), pp. 4860–4871.

[10] S.-M. Udrescu and M. Tegmark, Symbolic pregression: Discovering physical laws from distorted video, Phys. Rev. E **103**, 043307 (2021).

[11] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, arXiv:1312.6114.

[12] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, β-VAE: Learning basic visual concepts with a constrained variational framework, in 5th International Conference on Learning Representations, ICLR 2017—Conference Track Proceedings.

[13] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, Understanding disentangling in β-vae, arXiv:1804.03599.

[14] H. Kim and A. Mnih, Disentangling by factorising, PMLR **80**, 2649 (2018).

[15] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, Isolating sources of disentanglement in variational autoencoders, in *Advances in Neural Information Processing Systems*, Vol. 31 (Curran Associates, Inc., 2018).

[16] A. Kumar, P. Sattigeri, and A. Balakrishnan, Variational inference of disentangled latent concepts from unlabeled observations, arXiv:1711.00848.

[17] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, PMLR **97**, 4114 (2018).

[18] L. Lucie-Smith, H. V. Peiris, A. Pontzen, B. Nord, J. Thiyagalingam, and D. Piras, Discovering the building blocks of dark matter halo density profiles with neural networks, arXiv:2203.08827.

[19] M. Koch-Janusz and Z. Ringel, Mutual information, neural networks and the renormalization group, Nat. Phys. **14**, 578 (2018).

[20] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).

[21] C. Wang, H. Zhai, and Y.-Z. You, Emergent Schrödinger equation in an introspective machine learning architecture, Sci. Bull. **64**, 1228 (2019).

[22] P. Y. Lu, S. Kim, and M. Soljačić, Extracting Interpretable Physical Parameters from Spatiotemporal Systems Using Unsupervised Learning, Phys. Rev. X **10**, 031056 (2020).

[23] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho, Discovering symbolic models from deep learning with inductive biases, in *Advances in Neural Information Processing Systems*, Vol. 33 (Curran Associates, Inc., 2020), pp. 17429–17442.

[24] S. Ha and H. Jeong, Discovering invariants via machine learning, Phys. Rev. Research **3**, L042035 (2021).

[25] S. Greydanus, M. Dzamba, and J. Yosinski, Hamiltonian neural networks, in *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., 2019).

[26] M. Cranmer, S. Greydanus, S. Hoyer, P. Battaglia, D. Spergel, and S. Ho, Lagrangian neural networks, arXiv:2003.04630.

[27] B. C. Daniels and I. Nemenman, Automated adaptive inference of phenomenological dynamical models, Nat. Commun. **6**, 8133 (2015).

[28] S. L. Brunton, J. L. Proctor, J. N. Kutz, and W. Bialek, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Nat. Acad. Sci. USA **113**, 3932 (2016).

[29] D. Zhang, L. Guo, and G. E. Karniadakis, Learning in modal space: Solving time-dependent stochastic PDEs using physics-informed neural networks, SIAM J. Sci. Comput. **42**, A639 (2020).

[30] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. **378**, 686 (2019).

[31] M. Schmidt and H. Lipson, Distilling free-form natural laws from experimental data, Science **324**, 81 (2009).

[32] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.4.023206 for figures of additional latent spaces.