Editors' Suggestion | Featured in Physics

Discovering Physical Concepts with Neural Networks

Raban Iten,[*,†] Tony Metger,[*,‡] Henrik Wilming, Lídia del Rio, and Renato Renner
ETH Zürich, Wolfgang-Pauli-Strasse 27, 8093 Zürich, Switzerland

(Received 17 July 2019; published 8 January 2020)

Despite the success of neural networks at solving concrete physics problems, their use as a general-purpose tool for scientific discovery is still in its infancy. Here, we approach this problem by modeling a neural network architecture after the human physical reasoning process, which has similarities to representation learning. This allows us to make progress towards the long-term goal of machine-assisted scientific discovery from experimental data without making prior assumptions about the system. We apply this method to toy examples and show that the network finds the physically relevant parameters, exploits conservation laws to make predictions, and can help to gain conceptual insights, e.g., Copernicus' conclusion that the solar system is heliocentric.

DOI: 10.1103/PhysRevLett.124.010508

Theoretical physics, like all fields of human activity, is influenced by the schools of thought prevalent at the time of development. As such, the physical theories we know may not necessarily be the simplest ones to explain experimental data, but rather the ones that most naturally followed from a previous theory at the time. Both general relativity and quantum theory were built upon classical mechanics—they have been impressively successful in the restricted regimes of the very large and very small, respectively, but are fundamentally incompatible, as reflected by paradoxes such as the black hole information loss paradox [1,2]. This raises an interesting question: Are the laws of quantum physics, and other physical theories more generally, the most natural ones to explain data from experiments if we assume no prior knowledge of physics? While this question will likely not be answered in the near future, recent advances in artificial intelligence allow us to make a first step in this direction. Here, we investigate whether neural networks can be used to discover physical concepts from experimental data.

Previous work.—The goal of using machines to help with discovering the physical laws underlying experimental data has been pursued in several contexts (see the Supplemental Material (SM) [3] for a more detailed overview and Refs. [30–33] for recent reviews). A lot of early work focused on finding mathematical expressions describing a given dataset (see, e.g., Refs. [34–36]). For example, in Ref. [35] an algorithm recovers the laws of motion of simple mechanical systems, like a double pendulum, by searching over a space of mathematical expressions on given input variables. More recently, significant progress was made in extracting dynamical equations from experimental data [37–45]. These methods are highly practical and they were successfully applied to complex physical systems, but require prior knowledge on the systems of interest, for example in the form of knowing what the relevant variables are or that dynamics should be described by differential equations. In certain situations one might not have such prior knowledge or does not want to impose it to allow the machine to find entirely different representations of the physical system.

Over the last few years, neural networks have become the dominant method in machine learning and they have successfully been used to tackle complex problems in classical as well as quantum physics (see the SM [3] including Refs. [46–64] for further discussions). Conversely, neural networks may also lead to new insights into how the human brain develops physical intuition from observations [65–71]. Very recently, physical variables were extracted in an unsupervised way from time series data of dynamical systems in Ref. [72].

Our goal in this work is to minimize the extent to which prior assumptions about physical systems impose structure on the machine learning system. Eliminating assumptions that may not be satisfied for all physical systems, such as assuming that particles only interact in a pairwise manner, is necessary for the long-term goal of an artificial intelligence physicist (see Ref. [73] for recent progress in this direction) that can be applied to any system without a need for adaptions and might eventually contribute to progress in the foundations of physics. Very recently, neural networks were used in this spirit to detect differences between observed data and a reference model [74,75]. However, there is a tradeoff between generality and performance, and the performance of the machine learning system proposed here—based on autoencoders [76–78]—is not yet comparable to more established approaches that are adapted to specific physical systems.

Modeling the physical reasoning process.—This work makes progress towards an interpretable artificial intelligence agent that is unbiased by prior knowledge about physics by proposing to focus on the human physical

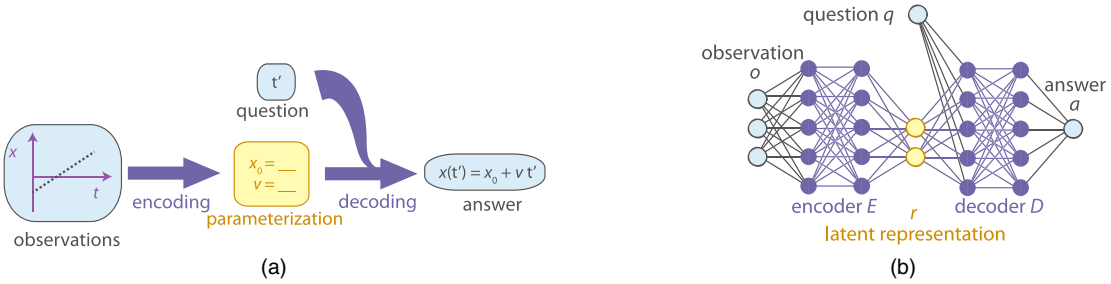

FIG. 1. Learning physical representations. (a) Human learning. A physicist compresses experimental observations into a simple representation (encoding). When later asked any question about the physical setting, the physicist should be able to produce a correct answer using only the representation and not the original data. We call the process of producing the answer from the representation "decoding." For example, the observations may be the first few seconds of the trajectory of a particle moving with constant speed; the representation could be the parameters "speed v" and "initial position x_0" and the question could be "where will the particle be at a later time t'?" (b) Neural network structure for SciNet. Observations are encoded as real parameters fed to an encoder (a feed-forward neural network, see SM [3]), which compresses the data into a representation (latent representation). The question is also encoded in a number of real parameters, which, together with the representation, are fed to the decoder network to produce an answer. (The number of neurons depicted is not representative.)

modeling process itself, rather than on specific physical systems. We formalize a simplified physical modeling process, which we then translate into a neural network architecture. This neural network architecture can be applied to a wide variety of physical systems, both classical and quantum, and is flexible enough to accommodate different additional desiderata on representations of the system that we may wish to impose.

We start by considering a simplified version of the physical modeling process, pictured in Fig. 1(a). Physicists' interactions with the physical world take the form of experimental observations [e.g., a time series $(t_i, x(t_i))_{i\in\{1,\ldots,N\}}$ describing the motion of a particle at constant speed]. The models physicists build do not deal with these observations directly, but rather with a representation of the underlying physical state of the observed system [e.g., the two parameters initial position and speed, (x_0, v)]. Which parameters are used is an important part of the model, and we will give suggestions about what makes a good representation below. Finally, the model specifies how to make predictions (i.e., answer questions) based on the knowledge of the physical state of the system (e.g., "where is the particle at time t'?"). More formally, this physical modeling process can be regarded as an "encoder" $E\colon \mathcal{O} \to \mathcal{R}$ mapping the set of possible observations $\mathcal{O}$ to representations $\mathcal{R}$, followed by a "decoder" $D\colon \mathcal{R} \times \mathcal{Q} \to \mathcal{A}$ mapping the sets of all possible representations $\mathcal{R}$ and questions $\mathcal{Q}$ to answers $\mathcal{A}$.

Network structure.—This modeling process can be translated directly into a neural network architecture, which we refer to as SciNet in the following [Fig. 1(b)]. The encoder and decoder are both implemented as feed-forward neural networks. The resulting architecture, except for the question input, resembles an autoencoder in representation learning [76,77], and more specifically the architecture in Ref. [79]. During the training, we provide triples of the form $(o, q, a_{\mathrm{corr}}(o, q))$ to the network, where $a_{\mathrm{corr}}(o, q) \in \mathcal{A}$ is the correct reply to question $q \in \mathcal{Q}$ given the observation $o \in \mathcal{O}$. The learned parametrization is typically called latent representation [76,77]. To feed the questions into the neural network, they are encoded into a sequence of real parameters. Thereby, the actual representation of a single question is irrelevant as long as it allows the network to distinguish questions that require different answers.

It is crucial that the encoder is completely free to choose a latent representation itself, instead of us imposing a specific one. Because neural networks with at least one hidden layer composed of sufficiently many neurons can approximate any continuous function arbitrarily well [80], the fact that the functions E and D are implemented as neural networks does not significantly restrict their generality. However, unlike in an autoencoder, the latent representation need not describe the observations completely; instead, it only needs to contain the information necessary to answer the questions posed.

This architecture allows us to extract knowledge from the neural network: all of the useful information is stored in the representation, and the size of this representation is small compared to the total number of degrees of freedom (d.o.f.) of the network. This allows us to interpret the learned representation. Specifically, we can compare SciNet's latent representation to a hypothesized parameterization to obtain a simple map from one to the other. If we do not even have any hypotheses about the system at hand, we may still gain some insights solely from the number of required parameters or from studying the change in the representation when manually changing the input, and the change in output when manually changing the representation (as in, e.g., Ref. [78]).

Desired properties for a representation.—For SciNet to produce physically useful representations, we need to formalize what makes a good parameterization of a

physical system, i.e., a good latent representation. We stress that this is not a property of a physical system, but a choice we have to make. We will give two possible choices below.

Generally, the latent representation should only store the minimal amount of information that is sufficient to correctly answer all questions in $\mathcal{Q}$. For minimal sufficient uncorrelated representations, we additionally require that the latent neurons be statistically independent from each other for an input sampled at random from the training data, reflecting the idea that physically relevant parameters describe aspects of a system that can be varied independently and are therefore uncorrelated in the experimental data. Under this independence assumption, the network is then motivated to choose a representation that stores different physical parameters in different latent neurons. We formalize these demands in the SM [3] and show, using techniques from differential geometry, that the number of latent neurons equals the number of underlying d.o.f. in the training data that are needed to answer all questions $\mathcal{Q}$. To implement these requirements in a neural network, we use well-established methods from representation learning, specifically disentangling variational autoencoders [78,81] (see SM [3] for details).

Alternatively, for situations where the physically relevant parameters can change, either over time or by some time-independent update rule, we might prefer a representation with a simple such update rule. We explain below how this requirement can be enforced.

Results.—To demonstrate that SciNet helps to recover relevant concepts in physics by providing the relevant physical variables, both in quantum- and classical-mechanical settings, we consider four toy examples from different areas of physics. In summary, we find (i) given a time series of the positions of a damped pendulum, SciNet can predict future positions with high accuracy and it uses the relevant parameters, namely, frequency and damping factor, separately in two of the latent neurons (and sets the activation of unnecessary latent neurons to zero), (ii) SciNet finds and exploits conservation laws: it uses the total angular momentum to predict the motion of two colliding particles, (iii) given measurement data from a simple quantum experiment, SciNet can be used to determine the dimension of the underlying unknown quantum system and to decide whether a set of measurements is tomographically complete, i.e., whether it provides full information about the quantum state, and (iv) given a time series of the positions of the Sun and Mars as observed from Earth, SciNet switches to a heliocentric representation—that is, it encodes the data into the angles of the two planets as seen from the Sun. The results show that SciNet finds, without having been given any prior information about the specific physical systems, the same quantities that we use in physics textbooks to describe the different settings. We also show that our results are robust against noise in the experimental data. To illustrate our approach, we will now describe two of these settings in some depth. For detailed descriptions of the four different settings, the data generation, interpretation, and additional background information, we refer to the SM [3].

In all our examples, the training data we use are operational and could be generated from experiments; i.e., the correct answer is the one observed experimentally. Here, we use simulations instead because we only deal with classical and quantum mechanics, theories whose predictions are experimentally well tested in the relevant regimes. One might think that using simulated data would restrict SciNet to rediscovering the theory used for data generation. However, in particular for quantum mechanics, we are interested in finding conceptually different formulations of the theory with the same predictions.

Quantum state tomography.—In quantum mechanics, it is not trivial to construct a simple representation of the state of a quantum system from measurement data, a task called quantum tomography [82]. In the following, we will show that SciNet finds representations of arbitrary (pure) one- and two-qubit states. To ensure that no prior knowledge about quantum physics is required to collect the measurement data, we assume an operational setting in which we have access to two devices in a lab, where one device can create (many copies of) a quantum system in a certain state depending on the chosen parameters of the device. The other device performs binary measurements on the quantum system. The input to SciNet consists of the outcome probabilities of a random fixed set of "reference measurements" on quantum systems in the unknown quantum state ψ. As a question input, we provide a parametrization of a measurement ω (one may think of the setting of the dials and buttons of the measurement device). SciNet has to predict the outcome probability of the measurement ω on a quantum system in the state ψ. We train SciNet with different pairs (ω, ψ) for one and two qubits. The results are shown in Fig. 2. Training different networks with different numbers of latent neurons, we can observe how the quality of the predictions (after training has been completed) improves as we allow for more parameters in the representation of ψ. This allows us to gain relevant information, without previous hypotheses about the nature of this representation (for example, whether it is a vector in a Hilbert space).

If the reference measurements are tomographically complete, meaning that they are sufficient to reconstruct a complete representation of the underlying quantum system, the plots in Fig. 2 show a drop in prediction error when the number of latent neurons is increased up to two and six for the cases of one and two qubits, respectively [83]. This is in accordance with the number of d.o.f. required to describe a one- or a two-qubit state in our current theory of quantum mechanics. For the case where the set of measurements is tomographically incomplete, it is not possible for SciNet to predict the outcome of the final

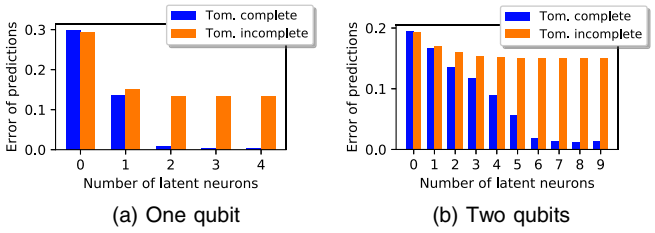

FIG. 2. Quantum tomography. SciNet is given tomographic data for one or two qubits, as shown in part (a) and (b) of the figure, respectively, and an operational description of a measurement as a question input and has to predict the probabilities of outcomes for this measurement. The plots show the root mean square error of SciNet's measurement predictions for test data as a function of the number of latent neurons. In the tomographically complete case, SciNet recovers the number of (real) d.o.f. required to describe a one and a two qubit state (which are two and six, respectively). Tomographically incomplete data can be recognized, since the prediction error remains high as one increases the number of latent neurons.

measurement perfectly regardless of the number of latent neurons. This means that purely from operational data, we can make a statement about the tomographic completeness of measurements and about the number of d.o.f. of the underlying unknown quantum system.

Enforcing a simple time evolution.—As mentioned above, if the physically relevant parameters can change, we can enforce a representation that has a simple update rule. For illustration, we will consider time evolution here, but more general update rules are possible. To accommodate changing physical parameters, we need to extend the latent representation as shown in Fig. 3(a). Instead of a single latent representation with a decoder attached to it, we now have many latent representations that are generated from the initial representation by a time evolution network. Each representation has a decoder attached to it to produce an answer to a question. Because we only want the parameters, but not the physical model, to change in time, all time evolution steps and decoders are identical; i.e., they implement the same function. The encoder, time evolution network, and decoder are trained simultaneously. To enforce parameters with a simple time evolution, we restrict the time evolution network to implementing very simple functions, such as addition of a constant [84].

Heliocentric solar system.—In the 16th century, Copernicus used observations of the positions of different planets in the night sky [Fig. 3(b)] to hypothesize that the Sun, and not the Earth, is at the center of our solar system. This heliocentric view was confirmed by Kepler at the start of the 17th century based on astronomic data collected by Brahe, showing that the planets move around the Sun in simple orbits. Here, we show that SciNet similarly uses heliocentric angles when forced to find a representation for which the time evolution of the variables takes a very simple form, a typical requirement for time-dependent variables in physics.

The observations given to SciNet are angles $\theta_M(t_0)$ of Mars and $\theta_S(t_0)$ of the Sun as seen from Earth at a starting time t_0 (which is varied during training). The time evolution network is restricted to addition of a constant (the value of which is learned during training). At each time step i, SciNet is asked to predict the angles as seen from Earth at the time t_i using only its representation $r(t_i)$. Because this question is constant, we do not need to feed it to the decoder explicitly.

We train SciNet with randomly chosen subsequences of weekly (simulated) observations of the angles θ_M and θ_S within Copernicus' lifetime (3665 observations in total). For our simulation, we assume circular orbits of Mars and Earth around the Sun. Figure 3(c) shows the learned representation and confirms that SciNet indeed stores a linear combination of heliocentric angles. We stress that the training data only contains angles observed from Earth, but SciNet nonetheless switches to a heliocentric representation.

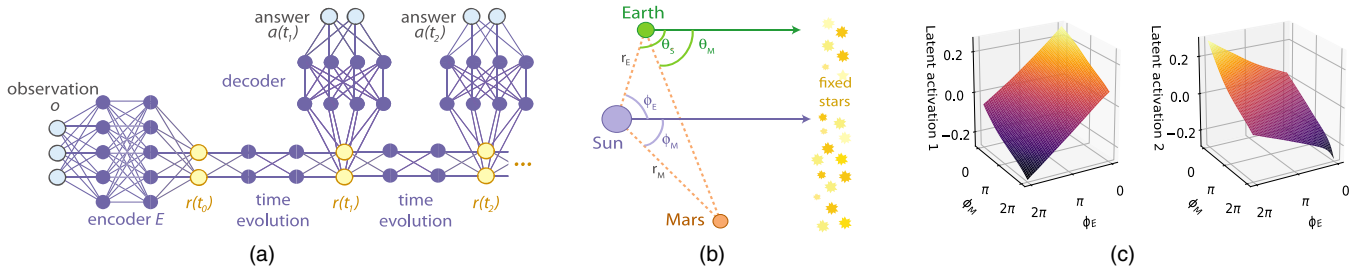

FIG. 3. Heliocentric model of the solar system. SciNet is given the angles of the Sun and Mars as seen from Earth at an initial time t_0 and has to predict these angles for later times. (a) Recurrent version of SciNet for time-dependent variables. Observations are encoded into a simple representation $r(t_0)$ at time t_0. Then, the representation is evolved in time to $r(t_1)$ and a decoder is used to predict $a(t_1)$, and so on. In each (equally spaced) time step, the same time evolution network and decoder network are applied. (b) Physical setting. The heliocentric angles ϕ_E and ϕ_M of the Earth and Mars are observed from the Sun; the angles θ_S and θ_M of the Sun and Mars are observed from Earth. All angles are measured relative to the fixed star background. (c) Representation learned by SciNet. The activations $r_{1,2}(t_0)$ of the two latent neurons at time t_0 [see Fig. 3(a)] are plotted as a function of the heliocentric angles ϕ_E and ϕ_M. The plots show that the network stores and evolves parameters that are linear combinations of the heliocentric angles.

Conclusion.—In this work, we have shown that SciNet can be used to recover physical variables from experimental data in various physical toy settings. The learned representations turned out to be the ones commonly used in physics textbooks, under the assumption of uncorrelated sampling. In future work we plan to extend our approach to data where the natural underlying parameters are correlated in the training distribution. The separation of these parameters in the representation found by SciNet requires the development of further operational criteria for disentangling latent variables. In more complex scenarios, the methods introduced here may lead to entirely novel representations, and extracting human physical insight from such representations remains challenging. This could be addressed using methods from symbolic regression [85] to obtain analytical expressions for the encoder and decoder maps, or for a map between a hypothesized and the actual representation. Alternatively, methods such as the ones presented in Refs. [86,87] could help to improve the interpretability of the representation. Following this direction, it might eventually become possible for neural networks to produce insights expressed in our mathematical language.

The source code and the training data are available at the web address in Ref. [88]. See also the Supplemental Material [3] for the implementation details. SciNet worked well on all tested examples; i.e., we did not postselect examples based on whether SciNet worked or not.

We would like to thank Alessandro Achille, Serguei Beloussov, Ulrich Eberle, Thomas Frerix, Viktor Gal, Thomas Häner, Maciej Koch-Janusz, Aurelien Lucchi, Ilya Nemenman, Joseph M. Renes, Andrea Rocchetto, Norman Sieroka, Ernest Y.-Z. Tan, Jinzhao Wang, and Leonard Wossnig for helpful discussions. We acknowledge support from the Swiss National Science Foundation through SNSF Project No. 200020_165843, the Swiss National Supercomputing Centre (CSCS) under project ID da04, and through the National Centre of Competence in Research *Quantum Science and Technology* (QSIT). L. d.R. and R. R. furthermore acknowledge support from the FQXi grant *Physics of the observer*. T. M. acknowledges support from ETH Zürich and the ETH Foundation through the *Excellence Scholarship & Opportunity Programme*.

R. I. and T. M. contributed equally to this work.

*These authors equally contributed to this work.
†itenr@itp.phys.ethz.ch
‡tmetger@ethz.ch

[1] S. W. Hawking, Phys. Rev. D **14**, 2460 (1976).
[2] J. Preskill, in International Symposium on Black Holes, Membranes, Wormholes, and Superstrings, 1993.
[3] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.124.010508 including Refs. [4–29], for more detailed discussion of previous work, detailed discussions of all examples, implementation details and additional background information on neural networks and variational autoencoders.
[4] M. Abadi *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems (2015).
[5] https://github.com/eth-nn-physics/nn_physical_concepts/blob/copernicus/analysis/copernicus_analysis.ipynb.
[6] M. A. Nielsen, Neural networks and deep learning (2018).
[7] Y. LeCun, Y. Bengio, and G. Hinton, Nature (London) **521**, 436 (2015).
[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Nature (London) **529**, 484 (2016).
[9] H. J. Briegel and G. D. l. Cuevas, Sci. Rep. **2**, 400 (2012).
[10] J. B. Hamrick, P. W. Battaglia, and J. B. Tenenbaum, Internal physics models guide probabilistic judgments about object dynamics, in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society Austin, TX*, edited by L. Carlson, C. Holscher, and T. F. Shipley (Cognitive Science Society, Austin, 2011), pp. 1545–1550, http://palm.mindmodeling.org/cogsci2011/papers/0350/paper0350.pdf.
[11] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, Proc. Natl. Acad. Sci. U.S.A. **110**, 18327 (2013).
[12] T. D. Ullman, A. Stuhlmüller, N. D. Goodman, and J. B. Tenenbaum, Cogn. Psychol. **104**, 57 (2018).
[13] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. Kavukcuoglu, in *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, 2016), p. 4509.
[14] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum, arXiv:1612.00341.
[15] D. Raposo, A. Santoro, D. Barrett, R. Pascanu, T. Lillicrap, and P. Battaglia, arXiv:1702.05068.
[16] T. S. Cubitt, J. Eisert, and M. M. Wolf, Phys. Rev. Lett. **108**, 120503 (2012).
[17] M. Fraccaro, S. Kamronn, U. Paquet, and O. Winther, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, 2017), Vol. 30, pp. 3601–3610.
[18] Y. Bengio, arXiv:1305.0445.
[19] A. Achille and S. Soatto, IEEE Trans. Pattern Anal. Mach. Intell. **40**, 2897 (2018).
[20] N. Tishby, F. C. Pereira, and W. Bialek, arXiv:physics/0004057.
[21] N. Tishby and N. Zaslavsky, in *2015 IEEE Information Theory Workshop (ITW)* (2015), p. 1.
[22] R. Shwartz-Ziv and N. Tishby, arXiv:1703.00810.
[23] J. Lee, *Introduction to Smooth Manifolds*, Graduate Texts in Mathematics, 2nd ed. (Springer-Verlag, New York, 2012).
[24] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, arXiv:1511.07289.
[25] G. Cybenko, Math. Control Signals Syst. **2**, 303 (1989).
[26] K. Hornik, M. Stinchcombe, and H. White, Neural Netw. **2**, 359 (1989).

[27] N. Pitelis, C. Russell, and L. Agapito, in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington, DC, 2013), p. 1642.
[28] E. O. Korman, arXiv:1803.00156.
[29] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, Cambridge, England, 2010).
[30] V. Dunjko and H. J. Briegel, arXiv:1709.02779.
[31] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, arXiv:1905.08883.
[32] I. Alhousseini, W. Chemissany, F. Kleit, and A. Nasrallah, arXiv:1905.01023.
[33] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, arXiv:1903.10563.
[34] J. P. Crutchfield and B. S. McNamara, Complex Syst. **1**, 417 (1987).
[35] M. Schmidt and H. Lipson, Science **324**, 81 (2009).
[36] M. Schmidt, R. R. Vallabhajosyula, J. W. Jenkins, J. E. Hood, A. S. Soni, J. P. Wikswo, and H. Lipson, Phys. Biol. **8**, 055011 (2011).
[37] B. C. Daniels and I. Nemenman, Nat. Commun. **6**, 8133 (2015).
[38] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Proc. Natl. Acad. Sci. U.S.A. **113**, 3932 (2016).
[39] B. Lusch, J. N. Kutz, and S. L. Brunton, arXiv:1712.09707.
[40] N. Takeishi, Y. Kawahara, and T. Yairi, arXiv:1710.04340.
[41] S. E. Otto and C. W. Rowley, arXiv:1712.01378.
[42] M. Raissi, arXiv:1801.06637.
[43] D. Zhang, L. Guo, and G. E. Karniadakis, arXiv:1905.01205.
[44] M. Raissi and G. E. Karniadakis, J. Comput. Phys. **357**, 125 (2018).
[45] M. Raissi, P. Perdikaris, and G. E. Karniadakis, J. Comput. Phys. **378**, 686 (2019).
[46] G. Carleo and M. Troyer, Science **355**, 602 (2017).
[47] A. Rocchetto, E. Grant, S. Strelchuk, G. Carleo, and S. Severini, npj Quantum Inf. **4**, 28 (2018).
[48] I. Glasser, N. Pancotti, M. August, I. D. Rodriguez, and J. I. Cirac, Phys. Rev. X **8**, 011006 (2018).
[49] G. Carleo, Y. Nomura, and M. Imada, Nat. Commun. **9**, 5322 (2018).
[50] Z. Cai and J. Liu, Phys. Rev. B **97**, 035116 (2018).
[51] Y. Huang and J. E. Moore, arXiv:1701.06246.
[52] D.-L. Deng, X. Li, and S. D. Sarma, Phys. Rev. B **96**, 195145 (2017).
[53] M. Schmitt and M. Heyl, SciPost Phys. **4**, 013 (2018).
[54] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Nat. Phys. **14**, 447 (2018).
[55] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, Phys. Rev. B **96**, 205152 (2017).
[56] D.-L. Deng, X. Li, and S. D. Sarma, Phys. Rev. X **7**, 021021 (2017).
[57] X. Gao and L.-M. Duan, Nat. Commun. **8**, 662 (2017).
[58] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, and A. Zeilinger, Phys. Rev. Lett. **116**, 090405 (2016).
[59] A. A. Melnikov, H. P. Nautrup, M. Krenn, V. Dunjko, M. Tiersch, A. Zeilinger, and H. J. Briegel, Proc. Natl. Acad. Sci. U.S.A. **115**, 1221 (2018).
[60] M. Koch-Janusz and Z. Ringel, Nat. Phys. **14**, 578 (2018).
[61] A. Decelle, V. Martin-Mayor, and B. Seoane, arXiv:1904.07637.
[62] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita, Nat. Mach. Intell. **1**, 155 (2019).
[63] M. J. S. Beach, I. De Vlugt, A. Golubeva, P. Huembeli, B. Kulchytskyy, X. Luo, R. G. Melko, E. Merali, and G. Torlai, SciPost Phys. **7**, 009 (2019).
[64] G. Torlai and R. G. Melko, arXiv:1905.04312.
[65] C. Bates, I. Yildirim, J. B Tenenbaum, and P. W Battaglia, in *Proceedings of the 37th Annual Conference of the Cognitive Science Society, Pasadena, CA* (Cognitive Science Society, Hoboken, 2015), p. 172.
[66] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum, in *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc., Red Hook, 2015), pp. 127–135.
[67] N. R. Bramley, T. Gerstenberg, J. B. Tenenbaum, and T. M. Gureckis, Cogn. Psychol. **105**, 9 (2018).
[68] D. Rempe, S. Sridhar, H. Wang, and L. J. Guibas, arXiv:1901.00466.
[69] M. Kissner and H. Mayer, arXiv:1905.09891.
[70] S. Ehrhardt, A. Monszpart, N. Mitra, and A. Vedaldi, arXiv:1805.05086.
[71] T. Ye, X. Wang, J. Davidson, and A. Gupta, arXiv:1808.10002.
[72] D. Zheng, V. Luo, J. Wu, and J. B. Tenenbaum, arXiv:1807.09244.
[73] T. Wu and M. Tegmark, Phys. Rev. E **100**, 033311 (2019).
[74] A. De Simone and T. Jacques, Eur. Phys. J. C **79**, 289 (2019).
[75] R. T. D'Agnolo and A. Wulzer, Phys. Rev. D **99**, 015014 (2019).
[76] Y. Bengio, A. Courville, and P. Vincent, IEEE Trans. Pattern Anal. Mach. Intell. **35**, 1798 (2013).
[77] G. E. Hinton and R. R. Salakhutdinov, Science **313**, 504 (2006).
[78] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, ICLR **2**, 1 (2017).
[79] S. M. A. Eslami *et al.*, Science **360**, 1204 (2018).
[80] K. Hornik, Neural Netw. **4**, 251 (1991).
[81] D. P. Kingma and M. Welling, arXiv:1312.6114.
[82] *Quantum State Estimation*, edited by M. Paris and J. Reháček, *Lecture Notes in Physics* (Springer, Berlin, Heidelberg, 2004).
[83] In the case of a single qubit, there is an additional small improvement in going from two to three latent neurons: this is a technical issue caused by the fact that (finite size) neural networks cannot represent discontinuous functions (see SM [3]). The same likely applies in the case of two qubits, going from 6 to 7 latent neurons.
[84] For a general system, there might not exist a representation that admits such a simple time evolution. In such a case, one may have to define a complexity measure for the time update rule and search over different rules successively increasing the complexity until SciNet can achieve good prediction accuracy.
[85] J. R. Koza, Stat. Comput. **4**, 87 (1994).
[86] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, arXiv:1905.13741.
[87] C. Bény, arXiv:1904.10387.
[88] https://github.com/eth-nn-physics/nn_physical_concepts.