

Statistical Physics through the Lens of Real-Space Mutual Information


Doruk Efe Gökmen¹,[✉] Zohar Ringel,² Sebastian D. Huber,¹ and Maciej Koch-Janusz^{1,3,4,*}

¹*Institute for Theoretical Physics, ETH Zurich, 8093 Zurich, Switzerland*

²*Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel*

³*Department of Physics, University of Zurich, 8057 Zurich, Switzerland*

⁴*James Franck Institute, The University of Chicago, Chicagoo, Illinois 60637, USA*

 (Received 1 April 2021; revised 10 August 2021; accepted 13 October 2021; published 6 December 2021)

Identifying the relevant degrees of freedom in a complex physical system is a key stage in developing effective theories in and out of equilibrium. The celebrated renormalization group provides a framework for this, but its practical execution in unfamiliar systems is fraught with *ad hoc* choices, whereas machine learning approaches, though promising, lack formal interpretability. Here we present an algorithm employing state-of-the-art results in machine-learning-based estimation of information-theoretic quantities, overcoming these challenges, and use this advance to develop a new paradigm in identifying the most relevant operators describing properties of the system. We demonstrate this on an interacting model, where the emergent degrees of freedom are qualitatively different from the microscopic constituents. Our results push the boundary of formally interpretable applications of machine learning, conceptually paving the way toward automated theory building.

DOI: [10.1103/PhysRevLett.127.240603](https://doi.org/10.1103/PhysRevLett.127.240603)

Fundamental physical theories, in a reductionist spirit, are often formulated at the smallest scales, describing the interactions of elementary constituents. Nonetheless, the experimentally accessible features typically arise from their collective behavior. Indeed, there exist profound examples of *effective* theories, e.g., classical hydrodynamics and thermodynamics, consistently describing complex phenomena in terms of a few macroscopic variables, without making any reference to individual particles.

Bridging this scale gap to *derive* the emergent macroscopic properties from microscopic models is a perpetual challenge. The renormalization group (RG) [1–4] provides a powerful framework for this, associating physical theories at different length scales by iteratively coarse-graining configurations of local degrees of freedom (d.o.f.). The induced RG transformation acts as a telescope in the space of models, generating the RG flow, whose structure around the fixed point eventually reveals the relevant d.o.f. They are the scaling operators, which determine the correlations, and thus the physical properties, at large scales.

In practice, executing this program in the real-space RG is very difficult. The accuracy of the procedure is improved by optimizing the coarse graining to retain the highest real-space mutual information (RSMI) [5,6], quantifying correlations to distant parts of the system. However, this still misses a crucial insight: any long-range information is due to the scaling operators and thus its optimal compression not only can serve as a better RG transformation, but should allow one to extract all the operators themselves, without ever explicitly executing the RG flow. This was recently proven in part: in critical systems, the formal solutions to

the RSMI compression problem are determined by the most relevant operators [7], *in theory* allowing one to access them directly. Unfortunately, solving this mathematical problem is notoriously hard in a general setting [8].

Here, using state-of-the-art machine learning results in estimating mutual information [8], we overcome this challenge to develop a highly efficient algorithm extracting relevant operators of the theory from real-space configurations. In contrast to standard approaches, no RG maps are iterated: scaling operators are *not* constructed from the RG flow, but instead using their definition as dominant contributions to RSMI, in a single step. The RSMI neural estimator (RSMI-NE) returns them parametrized as neural networks, which can be assigned scaling dimensions and used in computations [see Fig. 1(a)]. Moreover, we empirically demonstrate the power of the method across the whole phase diagram, also away from criticality.

In particular, the algorithm can, unsupervised, construct order parameters, locate phase transitions, and identify spatial correlations and symmetries for complex and large-dimensional real-space data. Our findings, elevating the coarse-graining transformations to formal operators, give a new paradigm in investigating statistical systems and a numerical toolbox to do so.

An often raised criticism of the use of machine learning in physics is the lack of interpretability of the results [9]. Particularly, the extent to which architecture- and training-dependent conclusions from machine learning relate to formal concepts in physical theories is unclear. RSMI-NE overcomes this challenge: its outputs are explicitly identified with the scaling operators [7] on the lattice. Thus, in

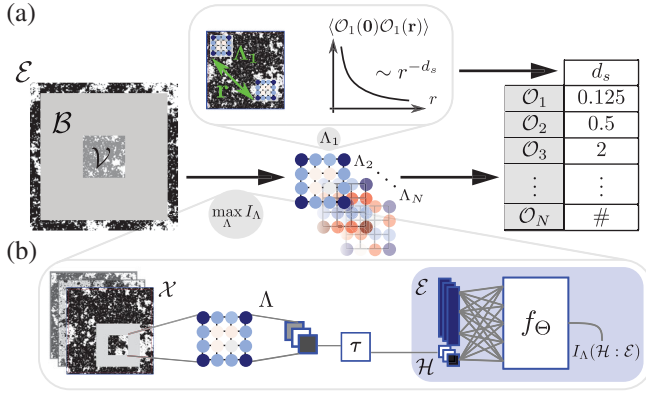


FIG. 1. Extracting the relevant operators with RSMI NE. (a) The most relevant operators are learnt as the optimal compressions of long-range information $I(\mathcal{H}:\mathcal{E})$ at each point in the phase diagram. The learnt maps can be associated with the physical operators by computing the correlators and extracting the scaling dimensions. (b) The architecture of RSMI-NE: the relevant operators are extracted via the transformations Λ and discretizing step τ . The long-range information that Λ maximizes is estimated by f_Θ , all of which are parametrized by neural networks and cotrained together.

contrast to generic data-driven approaches, RSMI-NE executes a *physical* principle using machine-learning tools to produce theoretically interpretable results.

Below we give an overview of the general RSMI setup, introducing the probabilistic language of the coarse-graining optimization. We then present the RSMI-NE algorithm and the theoretical and numerical results in machine learning underlying its efficiency. We validate its capabilities on an interacting model, whose nontrivial RG flow was a subject of a detailed theoretical analysis [10,11]. We investigate the physical data contained in the ensemble of RSMI filters. We conclude with a discussion of further applications, most notably to nonequilibrium problems.

Consider a system of classical d.o.f. in any dimension denoted by a collective random variable \mathcal{X} , whose physics is specified by a probability measure $p(x)$, either Gibbsian dictated by the energy of the realization x of \mathcal{X} , or a generic nonequilibrium distribution. A coarse-graining (CG) rule $\mathcal{X} \rightarrow \mathcal{X}'$ is defined as a conditional distribution $p_\Lambda(x'|x)$, determined by a set of parameters Λ to be optimized. The coarse graining is typically carried out on disjoint spatial blocks $\mathcal{V}_i \subset \mathcal{X}$, and it factorizes $p(x'|x) = \prod_i p_{\Lambda_i}(h_i|v_i)$, such that $\mathcal{X} = \cup_i \mathcal{V}_i$ and $\mathcal{X}' = \cup_i \mathcal{H}_i$, with $p_{\Lambda_i}(h_i|v_i)$ as the CG rule applied to block i . In translation-invariant systems, a fixed $\Lambda_i \equiv \Lambda$ suffices; with disorder each block can be individually optimized.

The RSMI principle identifies CG rules extracting the most relevant long-range features as the ones retaining the most information shared by a block $\mathcal{V} \subset \mathcal{X}$ to be coarse grained and its distant environment \mathcal{E} [5,6], i.e., those that optimally *compress* this information. The environment is

separated from \mathcal{V} by a shell of nonzero thickness constituting the buffer \mathcal{B} and forms the remainder of the system [see Fig. 1(a)]. The “shared information” between the random variables \mathcal{H} and \mathcal{E} is given by the mutual information

$$I_\Lambda(\mathcal{H}:\mathcal{E}) = \sum_{h,e} p_\Lambda(e,h) \log \left(\frac{p_\Lambda(e,h)}{p_\Lambda(h)p(e)} \right), \quad (1)$$

where $p_\Lambda(e,h)$ and $p(h)$ are the marginal probability distributions of $p_\Lambda(h,x) = p_\Lambda(h|v)p(x)$ obtained by summing over the d.o.f. in $\{\mathcal{V},\mathcal{B}\}$ and $\{\mathcal{V},\mathcal{B},\mathcal{E}\}$, respectively. Finding such optimal coarse graining requires thus maximizing I_Λ as a function of parameters Λ .

The conceptual importance of the buffer \mathcal{B} cannot be overstated: it sets the length scale, filtering out contributions of short-range correlations between \mathcal{V} and \mathcal{E} . Increasing its thickness L_B corresponds to growing the RG scale, preserving only the long-range physics. With an arbitrary fixed CG rule, this can only be achieved in the RG by iterating the coarse graining, with all the ensuing difficulties, particularly amplifying the errors in the formulation of the rule. In our approach, the CG rules themselves contain long-range information and are obtained in a single shot, by solving the I_Λ optimization problem directly at large L_B , at different points in the phase diagram.

The optimization problem of Eq. (1) is, however, difficult, as estimating or maximizing mutual information is notoriously hard [8]. This was a major weakness of the RSMI proposal [5], hindering numerical and theoretical progress. We can now overcome this challenge. At the heart of our approach, encapsulated in the RSMI-NE algorithm, is a series of recent results combining mathematically rigorous variational bounds on mutual information [12–14] with deep learning [8,15]. A family of *differentiable* lower bounds to I_Λ is introduced, parametrized by neural networks f_Θ [see Fig. 1(b)], which in the course of gradient descent training on the joint samples of \mathcal{H} and \mathcal{E} become accurate, and in the limit exact estimators of I_Λ , see the Supplemental Material [16]. The transformation $p_\Lambda(h|v)$ feeding the coarse-grained variables into the estimator is also expressed by a neural network ansatz. We use the following composite architecture [see Fig. 1(b)]:

$$h = \tau \circ (\Lambda \cdot v). \quad (2)$$

Here Λ become parameters of a convolutional neural network (CNN) applied to the configurations, and τ differentially maps $\Lambda \cdot v$ into states of variable h of predetermined type (e.g., pseudo-binary spins). This last embedding step is both crucial [24] and algorithmically nontrivial [25]. We emphasize that, while the CNN choice is motivated by convenience, any sufficiently expressive ansatz can be used.

We have thus cast *both* the CG rule and the lower bound to the cost function it optimizes as differentiable neural networks. Next, we can chain them together [see Fig. 1(b)] and *simultaneously* optimize via stochastic gradient descent, improving the RSME estimator and the CG rule in each pass. Note that it is this numerical breakthrough that enables the exploration of new theoretical ideas and renders the RSME algorithm a promising new approach to tackle open challenges in complex domains.

We demonstrate this on the example of an interacting dimer model. This is an optimal test bed for the illustration of our algorithm. First, a large class of classical statistical physics problems can be mapped to interacting dimer models [26–35]. Moreover, aspects of the quantum dimer model [36,37] leave their footprint on the phase diagram [10,11]. Second, in the dimer model, the relevant low-energy degrees of freedom are profoundly different from the microscopic building blocks of the theory and change qualitatively throughout the phase diagram. Hence, the algorithm is presented a nontrivial task.

The model is defined by the partition function $Z(T) = \sum_{\{C\}} \exp(-E_C/T)$ at a given temperature T and the configurations C involve binary-valued microscopic degrees

of freedom, dimers, that sit on the edges of the square lattice. They obey the constraint of exactly one dimer being connected to every vertex. The energy $E_C = N_C(\parallel) + N_C(=)$ counts plaquettes covered by parallel dimers favored by the interaction, see Fig. 2(a).

The essence of this system is in the interplay of aligning interaction energy and entropic effects due to the nonlocal cooperation of local dimer covering constraints. At low T the former facilitates a long-range order crystallizing the system into one of the four translation symmetry breaking *columnar* states, see Fig. 2(a). With increasing T the system undergoes a Berezinskii-Kosterlitz-Thouless (BKT) transition at $T_{\text{BKT}} = 0.65(1)$ [11], entering a critical phase characterized by algebraic decay of correlations (also at $T \rightarrow \infty$) with exponents continuously changing with T . This is reflected in the effective continuum field theory, which, via the mapping of dimer configurations to height field $\varphi(\mathbf{r})$ [38] (see also the Supplemental Material for the definition [16]) is given by a sine-Gordon action [11]

$$S[\varphi(\mathbf{r})] = \int d^2\mathbf{r} \left[\frac{g(T)}{2} |\nabla\varphi(\mathbf{r})|^2 + V \cos[4\varphi(\mathbf{r})] \right]. \quad (3)$$

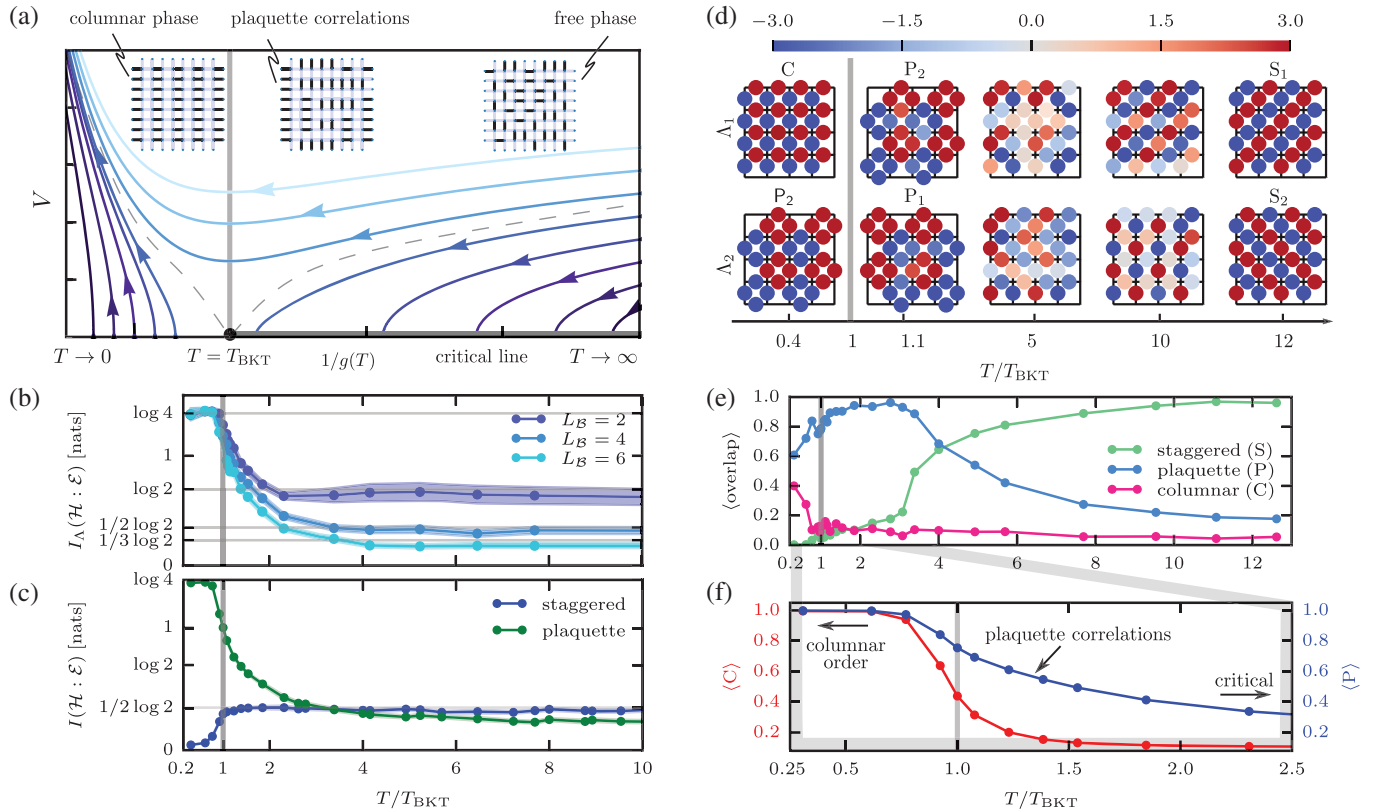


FIG. 2. RSME analysis of the interacting dimer model. (a) RG flow of the model [see Eq. (3)] and representative configurations (top). (b) Total RSME extracted with the optimal filters as a function of T and its scaling with the buffer size. (c) Information extracted by the pristine staggered and plaquette filters at different T . (d) Samples of optimal filters obtained with RSME-NE for different T [columnar (C), plaquette (P1, P2), and staggered (S1, S2)]. (e) The average overlap of the optimal filters at T with the pristine components. (f) The dimer symmetry breaking and plaquette order parameters extracted using the low- T pristine filters.

The potential V locks $\varphi(\mathbf{r})$ into four values corresponding to the columnar states. The stiffness $g(T)$ controls fluctuations of $\varphi(\mathbf{r})$: large $g(T)$ favors “flat” fields of high entropy; low $g(T)$ allows large gradients corresponding to the staggered configurations, which are not suppressed in the algebraic phase. The RG flow is shown in Fig. 2(a): the $T < T_{\text{BKT}}$ fixed point with finite $g(T)$ and $V \rightarrow \infty$ leaves energy minimization as the sole relevant constraint; the *line* of fixed points at $V = 0$ at $T > T_{\text{BKT}}$ indicates that the energetic interactions are irrelevant and exponents vary with T . The flow reveals the physical nature of the algebraic correlations: $\nabla\varphi(\mathbf{r})$ obeys Gauss’s law and so the fixed point theory is that of electrical fields.

To showcase RSMI-NE, we input Monte Carlo samples of the model across the whole temperature range to the algorithm. For concreteness, we restrict the coarse-grained variables \mathcal{H} to a two-component binary vector $\{\pm 1, \pm 1\}$ (the optimal dimensionality can be found systematically [39]). Hence, we are looking for a two-component vector of filters Λ_1, Λ_2 determining how the visible region \mathcal{V} is mapped onto \mathcal{H} . Optimizing the filters Λ_1, Λ_2 for all T separately gives a comprehensive picture of the long-wavelength physics, culminating in the construction of the relevant operators on the lattice, as we now show.

First, we find that already the curve $I_\Lambda(T)$, i.e., the amount of long-range information attained with the optimal Λ , reveals the structure of the phase diagram [see Fig. 2(b)]. To wit, for $T < T_{\text{BKT}}$ its value is constant and equal to $\log 4$. The information shared between distant parts of the system in the ordered phase is precisely which of the four columnar states they are in.

Phase transitions are reflected by nonanalyticities in $I_\Lambda(T)$ (cf. [40,41]). Moreover, the algebraic decay of $I_\Lambda(T)$ with the buffer size for $T > T_{\text{BKT}}$ is indicative of a critical phase, see Fig. 2(b) and Fig. 6(c) in Ref. [39]. This behavior should also be contrasted with the exponential decay for the paramagnetic phase of 2D Ising model in Ref. [39].

Going beyond the mutual information and examining the filters $\Lambda(T)$ yields further insight about spatial correlations. As conjectured, the optimal CG rules depend on the tuning parameters of the system. In the high- and low-temperature limits, three classes of filters emerge: independent optimizations return exclusively sets of $\Lambda_{1,2}$ that correspond to columnar and plaquette at low temperatures, and staggered ones at high temperatures, see Fig. 2(d). We call these filters “pristine” as they reflect simple limiting cases. They are orthogonal to each other and represent independent degrees of freedom. The filters for intermediate temperatures and their overlap with the pristine ones is shown in Figs. 2(d) and 2(e), respectively. We first discuss in detail the individual filters $\Lambda_{1,2}$ in the different temperature regimes $T \rightarrow 0$, $T \sim T_{\text{BKT}}$, and $T \gg T_{\text{BKT}}$, and then explicitly match them with the RG-relevant operators of the continuum sine-Gordon theory.

The pristine plaquette and columnar filters at $T \rightarrow 0$ break translation or rotation symmetry, respectively. Any pair of $\Lambda_{1,2}$ drawn out of these filters is a *bijection* between the four ordered columnar states and the four states $(\pm 1, \pm 1)$ taken by the compressed degrees of freedom \mathcal{H} . This degeneracy of plaquette and columnar filters is lifted when the rotation symmetry is restored: the pristine columnar filter is not found above T_{BKT} . Strikingly, its modulus acquires an expectation value for $T < T_{\text{BKT}}$ [see Fig. 2(f)]. This filter is thus an order parameter discovered by RSMI-NE and is, in fact, equal to the dimer symmetry breaking (DSB) order parameter identified in Ref. [10].

The optimal CG rules around T_{BKT} hold yet further insights. Particularly, the plaquette filters give rise to a *putative* plaquette order parameter [see Fig. 2(f)]. The corresponding regime where it attains a nonvanishing value does not survive in the thermodynamic limit. However, the nonzero expectation value at finite system sizes [see Fig. 2(f)] reveals the importance of such plaquette correlations, which are stabilized in the quantum dimer model (QDM) [36]. RSMI-NE indicates this without any prior insights about QDM, which inspired previous studies [10,11].

Finally, the critical phase $T > T_{\text{BKT}}$ interpolates between pristine plaquette and staggered filters, due to the competition between the electric field operator and plaquette correlations in the finite system, as per Eq. (3). The value of RSMI attained with *fixed* rules reflects this competition: the plaquette filter retains more information until well above T_{BKT} , where the staggered one takes over as plaquette correlations dwindle [see Fig. 2(c)]. The staggered filters are the electric fields, viz. they define coarse-grained variables $E_{1,2}(\mathbf{r}) := \tau \circ \Lambda_{1,2}[\mathcal{V}(\mathbf{r})]$, which precisely target the operator $\nabla\varphi(\mathbf{r})$ (see the Supplemental Material [16]).

The RSMI-NE finding the order parameters or the electric fields is no accident: the pristine Λ filters define the relevant operators on the lattice. The considerable technical machinery behind this is the subject of Ref. [7]; here we show it using the field theory of the dimer model, also away from criticality. To wit, the columnar and the DSB order parameters in Fig. 2(f) correspond to the relevant electric charge operators $\mathcal{O}_n(\varphi) = (\cos(n\varphi), \sin(n\varphi))$ [42] for $n = \pm 1$ and $n = \pm 2$, respectively. This is seen explicitly, using the height-field map in Table I as a dictionary,









$$(\Lambda_{P1}, \Lambda_{P2}) \circ \varphi = (\cos(\varphi + 3\pi/4), \sin(\varphi + 3\pi/4)), \quad (4)$$

$$\Lambda_C \circ \varphi = \cos(2\varphi), \quad (5)$$

where on the left dimer configurations (on which the Λ act) mapped to height-field value φ are denoted by φ itself.

Though competing correlations, especially in finite-size systems, may result in mixing of the pristine components, they can be identified by applying standard machine-learning

TABLE I. Pairs of filters drawn from the columnar and plaquette coarse-graining rules unambiguously label each of the four columnar ground states. The mapping is given by the sign of the scalar product of the filter (blue, -1 , red, $+1$) with a dimer configuration (1 for occupied, 0 for unoccupied link). As the columnar configurations correspond to uniform height field, the electric charge operators $\mathcal{O}_{n=1,2}$ acting on the height field φ also serve as order parameters for the columnar phase and they directly correspond to the RSMI-optimal filters at low T .

$\mathcal{V}(\mathbf{r})$				
 	$(-1, -1)$	$(-1, +1)$	$(+1, -1)$	$(+1, +1)$
 	$(-1, -1)$	$(+1, +1)$	$(-1, +1)$	$(+1, -1)$
$\varphi(\mathbf{r})$	$\frac{\pi}{2}$	$\frac{3\pi}{2}$	π	0
$\mathcal{O}_1(\varphi) = (\cos \varphi, \sin \varphi)$	$(0, 1)$	$(0, -1)$	$(-1, 0)$	$(+1, 0)$
$\mathcal{O}_2(\varphi) = \cos(2\varphi)$	-1	-1	$+1$	$+1$

tools to the ensemble of filters. Note that RSMI-NE is a stochastic algorithm and through independent runs produces a distribution of optimal CG rules. Thus, Fig. 2(d) shows a sample of filters at each T , and in Fig. 2(e) the overlap is averaged over the filter ensemble at each temperature. The distribution contains crucial information, e.g., the disappearance of the columnar filter above T_{BKT} signals the lifting of the columnar and plaquette degeneracy (consistent with the scaling dimensions of \mathcal{O}_n , which go as n^2) and restoration of the rotation symmetry. More concretely, representations of the broken symmetries can be identified in the distribution, whereas at high T it can be used to retrieve even the *emergent* $U(1)$ symmetry of the electrical field. See Ref. [39] for a more detailed discussion.

We thus managed to automatically sequence the operators of the theory, returning their lattice representations, which are modular, reusable, and may be formally labeled by their scaling dimensions. Indeed, evaluating a correlator of two neural networks parametrized by the plaquette filters, we fit a scaling dimension of 1.00037 at $T \rightarrow \infty$ (see the Supplemental Material [16]), in excellent agreement with 1.0 predicted for \mathcal{O}_1 [42]. This raises the remarkable prospect of building a complete effective theory from raw data using machine learning.

Though the discussion centred around an equilibrium example in two dimensions, our procedure works in any dimension, can be adapted to disorder [24], and does not require the existence of a Hamiltonian, as it only uses probability distributions. While a formal understanding of this approach for nonequilibrium distributions, extending

the results of Ref. [7], is missing, in the companion paper [39] we validated the concept on the example of lattice model with aggregation and chipping [43] for which RSMI-NE locates precisely the nonequilibrium phase transition. We believe complex systems, such as realized in, e.g., active matter [44,45] or atmospheric phenomena [46], to be a natural arena where information-theoretic methods can be applied [47], and our conceptual and numerical advancements may provide new theoretical insights (see also Ref. [48]). The understanding of challenging higher-dimensional interacting and quasiperiodic statistical systems [49–52] may also benefit from this new method.

The source code for the RSMI-NE is available online [53].

M. K.-J. is grateful to F. Alet for his comments on the physics of the interacting dimer model. D. E. G., S. D. H., and M. K.-J. gratefully acknowledge financial support from the Swiss National Science Foundation and the NCCR QSIT, and the European Research Council under the Grant Agreement No. 771503 (TopMechMat), as well as from European Union’s Horizon 2020 Programme under Marie Skłodowska-Curie Grant Agreement No. 896004 (COMPLEX ML). Z. R. acknowledges support from ISF Grant No. 2250/19. Some of the computations were performed using the Leonhard cluster at ETH Zurich. This work was supported by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID eth5b.

*Corresponding author.

maciej.koch-janusz@uzh.ch

- [1] L. P. Kadanoff, Scaling laws for Ising models near T_c , *Phys. Phys. Fiz.* **2**, 263 (1966).
- [2] K. G. Wilson and J. Kogut, The renormalization group and the ϵ expansion, *Phys. Rep.* **12**, 75 (1974).
- [3] K. G. Wilson, The renormalization group: Critical phenomena and the Kondo problem, *Rev. Mod. Phys.* **47**, 773 (1975).
- [4] M. E. Fisher, Renormalization group theory: Its basis and formulation in statistical physics, *Rev. Mod. Phys.* **70**, 653 (1998).
- [5] M. Koch-Janusz and Z. Ringel, Mutual information, neural networks and the renormalization group, *Nat. Phys.* **14**, 578 (2018).
- [6] P. M. Lenggenhager, D. E. Gökmen, Z. Ringel, S. D. Huber, and M. Koch-Janusz, Optimal Renormalization Group Transformation from Information Theory, *Phys. Rev. X* **10**, 011037 (2020).
- [7] A. Gordon, A. Banerjee, M. Koch-Janusz, and Z. Ringel, Relevance in the Renormalization Group and in Information Theory, *Phys. Rev. Lett.* **126**, 240601 (2021).
- [8] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, On variational bounds of mutual information, [arXiv:1905.06922](https://arxiv.org/abs/1905.06922).

- [9] P. V. Coveney, E. R. Dougherty, and R. R. Highfield, Big data need big theory too, *Phil. Trans. R. Soc. A* **374**, 20160153 (2016).
- [10] F. Alet, J. L. Jacobsen, G. Misguich, V. Pasquier, F. Mila, and M. Troyer, Interacting Classical Dimers on the Square Lattice, *Phys. Rev. Lett.* **94**, 235702 (2005).
- [11] F. Alet, Y. Ikhlef, J. L. Jacobsen, G. Misguich, and V. Pasquier, Classical dimers with aligning interactions on the square lattice, *Phys. Rev. E* **74**, 041124 (2006).
- [12] M. D. Donsker and S. R. S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. IV, *Commun. Pure Appl. Math.* **36**, 183 (1983).
- [13] D. Barber and F. V. Agakov, Information maximization in noisy channels: A variational approach, in *Advances in Neural Information Processing Systems*, edited by S. Thrun, L. K. Saul, and B. Schölkopf (MIT Press, Cambridge, MA, 2004), Vol. 16, p. 201.
- [14] X. Nguyen, M. J. Wainwright, and M. I. Jordan, Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization, in *Advances in Neural Information Processing Systems*, edited by J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (MIT Press, Cambridge, MA, 2009), Vol. 20, pp. 1089–1096.
- [15] M. I. Belghazi *et al.*, MINE: Mutual information neural estimation, [arXiv:1801.04062](https://arxiv.org/abs/1801.04062).
- [16] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevLett.127.240603> for details about the dimer model and more technical details of the RSMI-NE algorithm and its dependence on length scales, where also Refs. [17–23] are included.
- [17] A. van den Oord, Y. Li, and O. Vinyals, Representation learning with contrastive predictive coding, [arXiv:1807.03748](https://arxiv.org/abs/1807.03748).
- [18] M. Gutmann and A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Vol. 9 of Proceedings of Machine Learning Research*, edited by Y. W. Teh and M. Titterton (PMLR, 2010), pp. 297–304.
- [19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016), <http://www.deeplearningbook.org>.
- [20] C. J. Maddison, D. Tarlow, and T. Minka, A* sampling, in *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., Red Hook, 2014), Vol. 27, pp. 3086–3094.
- [21] E. J. Gumbel, The maxima of the mean largest value and of the range, *Ann. Math. Stat.* **25**, 76 (1954).
- [22] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [23] L. P. Kadanoff, A. Houghton, and M. C. Yalabik, Variational approximations for renormalization group transformations, *J. Stat. Phys.* **14**, 171 (1976).
- [24] P. Y. Lu, S. Kim, and M. Soljačić, Extracting Interpretable Physical Parameters from Spatiotemporal Systems Using Unsupervised Learning, *Phys. Rev. X* **10**, 031056 (2020).
- [25] E. Jang, S. Gu, and B. Poole, Categorical reparameterization with Gumbel-Softmax, [arXiv:1611.01144](https://arxiv.org/abs/1611.01144).
- [26] P. W. Kasteleyn, Dimer statistics and phase transitions, *J. Math. Phys. (N.Y.)* **4**, 287 (1963).
- [27] H. W. J. Blote and H. J. Hilhorst, Roughening transitions and the zero-temperature triangular Ising antiferromagnet, *J. Phys. A* **15**, L631 (1982).
- [28] C. L. Henley, Relaxation time for a dimer covering with height representation, *J. Stat. Phys.* **89**, 483 (1997).
- [29] R. Kenyon, Dominos and the Gaussian free field, *Ann. Probab.* **29**, 1128 (2001).
- [30] R. Kenyon, The Laplacian and Dirac operators on critical planar graphs, *Inventiones Mathematicae* **150**, 409 (2002).
- [31] D. Cimasoni and N. Reshetikhin, Dimers on surface graphs and spin structures. I, *Commun. Math. Phys.* **275**, 187 (2007).
- [32] S. Powell and J. T. Chalker, SU(2)-Invariant Continuum Theory for an Unconventional Phase Transition in a Three-Dimensional Classical Dimer Model, *Phys. Rev. Lett.* **101**, 155702 (2008).
- [33] G. Chen, J. Gukelberger, S. Trebst, F. Alet, and L. Balents, Coulomb gas transitions in three-dimensional classical dimer models, *Phys. Rev. B* **80**, 045112 (2009).
- [34] S. Powell and J. T. Chalker, Classical to quantum mapping for an unconventional phase transition in a three-dimensional classical dimer model, *Phys. Rev. B* **80**, 134413 (2009).
- [35] G. J. Sreejith, S. Powell, and A. Nahum, Emergent SO(5) Symmetry at the Columnar Ordering Transition in the Classical Cubic Dimer Model, *Phys. Rev. Lett.* **122**, 080601 (2019).
- [36] D. S. Rokhsar and S. A. Kivelson, Superconductivity and the Quantum Hard-Core Dimer Gas, *Phys. Rev. Lett.* **61**, 2376 (1988).
- [37] E. Fradkin, D. A. Huse, R. Moessner, V. Oganesyan, and S. L. Sondhi, Bipartite Rokhsar–Kivelson points and Cantor deconfinement, *Phys. Rev. B* **69**, 224415 (2004).
- [38] E. Fradkin, *Field Theories of Condensed Matter Physics*, 2nd ed. (Cambridge University Press, Cambridge, England, 2013).
- [39] D. E. Gökmen, Z. Ringel, S. D. Huber, and M. Koch-Janusz, companion paper, Symmetries and phase diagrams with real-space mutual information neural estimation, *Phys. Rev. E* **104**, 064106 (2021).
- [40] J. Wilms, M. Troyer, and F. Verstraete, Mutual information in classical spin models, *J. Stat. Mech.* (2011) P10011.
- [41] H. W. Lau and P. Grassberger, Information theoretic aspects of the two-dimensional Ising model, *Phys. Rev. E* **87**, 022128 (2013).
- [42] S. Papanikolaou, E. Luijten, and E. Fradkin, Quantum criticality, lines of fixed points, and phase separation in doped two-dimensional quantum dimer models, *Phys. Rev. B* **76**, 134514 (2007).
- [43] R. Rajesh and S. N. Majumdar, Exact phase diagram of a model with aggregation and chipping, *Phys. Rev. E* **63**, 036114 (2001).
- [44] J. Toner and Y. Tu, Flocks, herds, and schools: A quantitative theory of flocking, *Phys. Rev. E* **58**, 4828 (1998).
- [45] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Statistical mechanics for natural flocks of birds, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4786 (2012).

- [46] O. Peters and J. D. Neelin, Critical phenomena in atmospheric precipitation, *Nat. Phys.* **2**, 393 (2006).
- [47] R. Dewar, Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states, *J. Phys. A* **36**, 631 (2003).
- [48] A. Nir, E. Sela, R. Beck, and Y. Bar-Sinai, Machine-learning iterative calculation of entropy for physical systems, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30234 (2020).
- [49] I. S. Tupitsyn, A. Kitaev, N. V. Prokof'ev, and P. C. E. Stamp, Topological multicritical point in the phase diagram of the toric code model and three-dimensional lattice gauge Higgs model, *Phys. Rev. B* **82**, 085114 (2010).
- [50] E.-G. Moon and C. Xu, Exotic continuous quantum phase transition between F_2 topological spin liquid and Néel order, *Phys. Rev. B* **86**, 214414 (2012).
- [51] U. Agrawal, S. Gopalakrishnan, and R. Vasseur, Universality and quantum criticality in quasiperiodic spin chains, *Nat. Commun.* **11**, 2225 (2020).
- [52] F. Flicker, S. H. Simon, and S. A. Parameswaran, Classical Dimers on Penrose Tilings, *Phys. Rev. X* **10**, 011005 (2020).
- [53] D. E. Gökmen, Z. Ringel, S. D. Huber, and M. Koch-Janusz, RSMI-NE/RSMI-NE, 2021, <https://github.com/RSMI-NE/RSMI-NE>.