Optimal Renormalization Group Transformation from Information Theory

Author(s):
Lenggenhager, Patrick M.; Gökmen, Doruk Efe; Ringel, Zohar; Huber, Sebastian D.; Koch-Janusz, Maciej

Publication date:
2020

Permanent link:
https://doi.org/10.3929/ethz-b-000400030

Originally published in:
Physical Review X 10(1), https://doi.org/10.1103/physrevx.10.011037

Optimal Renormalization Group Transformation from Information Theory

Patrick M. Lenggenhager,[1] Doruk Efe Gökmen,[1] Zohar Ringel,[2] Sebastian D. Huber,[1] and Maciej Koch-Janusz[1]
[1]*Institute for Theoretical Physics, ETH Zurich, 8093 Zurich, Switzerland*
[2]*Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel*

(Received 30 October 2018; revised manuscript received 4 October 2019; accepted 19 December 2019; published 14 February 2020)

Recently, a novel real-space renormalization group (RG) algorithm was introduced. By maximizing an information-theoretic quantity, the real-space mutual information, the algorithm identifies the relevant low-energy degrees of freedom. Motivated by this insight, we investigate the information-theoretic properties of coarse-graining procedures for both translationally invariant and disordered systems. We prove that a perfect real-space mutual information coarse graining does not increase the range of interactions in the renormalized Hamiltonian, and, for disordered systems, it suppresses the generation of correlations in the renormalized disorder distribution, being in this sense *optimal*. We empirically verify decay of those measures of complexity as a function of information retained by the RG, on the examples of arbitrary coarse grainings of the clean and random Ising chain. The results establish a direct and quantifiable connection between properties of RG viewed as a compression scheme and those of physical objects, i.e., Hamiltonians and disorder distributions. We also study the effect of constraints on the number and type of coarse-grained degrees of freedom on a generic RG procedure.

DOI: 10.1103/PhysRevX.10.011037

Subject Areas: Complex Systems, Computational Physics, Statistical Physics

I. INTRODUCTION

The conceptual relations between physics and information theory date back to the very earliest days of statistical mechanics; they include the pioneering work of Boltzmann [1] and Gibbs [2] on entropy, finding its direct counterpart in Shannon's information entropy [3], and investigations of Szilard [4] and Landauer [5]. In the quantum regime, research initially focused on foundational challenges posed by the notion of entanglement, but it soon gave rise to the wide discipline of quantum-information theory [6], whose more practical aspects include quantum algorithms and computation.

In recent years, there has been renewed interest in applying the formalism and tools of information theory to fundamental problems of theoretical physics. The motivation comes mainly from two, not entirely unrelated, directions. On the one hand, the high-energy community is actively investigating the idea of holography in quantum field theories [7–9] originally inspired by black-hole thermodynamics. On the other hand, in condensed matter theory there is a growing appreciation of the role of the entanglement *structure* of quantum wave functions in determining the physical properties of the system, exemplified by the short- and long-range entanglement distinguishing the symmetry-protected topological phases [10–12] (e.g., topological insulators) from genuine, fractionalized topological orders (e.g., fractional quantum Hall states). The conceptual advances led also to *constructive* developments in the form of new *Ansätze* for wave functions (matrix-product states [13], multi-scale entanglement renormalization *Ansatz* [14]) and numerical algorithms (density-matrix renormalization group [15], neural quantum states [16]).

The focus of this work is on the renormalization group (RG). One of the conceptually most profound developments in theoretical physics, in particular, condensed matter theory, it provides—beyond more direct applications—a theoretical foundation for the notion of universality [17–21]. The connections of RG to information theory have been explored in a number of works [22–28] in both classical and quantum settings. In particular, some of the present authors introduced a numerical algorithm for real-space RG of classical statistical systems [28] based on the characterization of relevant degrees of freedom (d.o.f.) supported in a spatial block as the ones sharing the most real-space mutual information (RSMI) with the environment of the block. The algorithm employs machine-learning techniques to extract those d.o.f. and combines it with an iterative sampling scheme of Monte Carlo RG [29,30], though, in a crucial difference, the form of the RG coarse-graining rule is not given but rather *learned*. Strikingly, the coarse-graining rules discovered by the algorithm for the test systems were in an operational

sense optimal [31]: They ignored irrelevant short-scale noise, and they result in simple effective Hamiltonians or match nontrivial analytical results.[1]

The above suggests that real-space RG can be universally defined in terms of information theory rather than based on problem-specific physical intuition. Here we develop a theoretical foundation inspired by and underlying those numerical results. We show they were not accidental but rather a consequence of general principles. To this end, we prove that a *perfect* full-RSMI-retaining coarse graining of a finite-range Hamiltonian does not increase the range of interactions in the renormalized Hamiltonian in any dimension. We then study analytically generic coarse grainings and the effective Hamiltonian they define as a function of the real-space mutual information with the environment (RSMI) retained. For the example of the Ising chain, we perturbatively derive all the couplings in the renormalized Hamiltonian resulting from, and RSMI captured by, an *arbitrary* coarse graining and show monotonic decay of the higher-order and/or long-range terms with increased RSMI.

Those properties also hold in the presence of disorder. We prove that perfect RSMI-maximizing coarse graining is stable to local changes in disorder realization and suppresses generation of correlations in the renormalized disorder distribution. Using the solvable example of the random dilute Ising chain, we study the properties of the renormalized disorder distribution induced by an arbitrary RG procedure and show the decay of statistical measures of correlation in disorder as a function of the RSMI retained.

We also theoretically investigate the effects imposed by constraints on the number and type of coarse-grained variables, which can make the loss of part of the relevant information inevitable. We construct simple toy models providing intuitive understanding of our results.

Our results establish a direct link between compression theory intuition behind the introduction of RSMI [32] and physical properties of the renormalized Hamiltonian or disorder distribution. They strongly support RSMI maximization as a model-independent variational principle defining the *optimal* RG coarse graining. In contrast to fixed schemes, this RG approach is, by construction, informed by the physics of the system considered, including the position in the phase diagram. This generality could allow application of RG schemes to systems, for which they are currently not known, avoiding many of the pitfalls befalling fixed RG transformations [33,34].

The paper is organized as follows: In Sec. II, the information-theoretic formalism and the RSMI algorithm are reviewed. In Sec. III, we prove that a RSMI-perfect RG does not generate longer-range interactions. In Sec. IV, we investigate the renormalized Hamiltonian as a function of the information retained on the example of arbitrary coarse grainings of the 1D Ising model. In Sec. V, we study the effect of constraints on the number and type of coarse-grained variables on a RG procedure. We introduce toy models explaining the differences in optimal coarse grainings in 1D and 2D. In Sec. VI, we extend the analysis to disordered systems. We prove RSMI-perfect RG does not generate correlations in disorder. We study properties of the renormalized disorder as a function of the information retained for arbitrary coarse grainings of the random dilute Ising chain. In Sec. VII, we discuss implications of the results, generalizations, and open questions. Appendices give proof details, derivations of the statements in the main text, and additional information.

[1]For a brief discussion of related concepts in coarse graining of differential equations, we refer to the Appendix F.

II. THE RSMI ALGORITHM

The RSMI algorithm is defined in the context of real-space RG originally introduced by Kadanoff for lattice models [17]. The goal of real-space RG [21] is to coarse grain a given set of d.o.f. $\mathcal{X}$ in position space in order to integrate out short-range fluctuations and retain only long-range correlations, and in so doing, to construct an effective theory. An iterative application of this procedure should result in recursive relations between coupling constants of the Hamiltonian at successive RG steps—those are the RG flow equations formalizing the relationship between effective theories at different length scales.

Consider a generic system with a set of real-space d.o.f. $\mathcal{X}$, e.g., spins, described by the Hamiltonian $H[\mathcal{X}]$ and a canonical partition function:

$$Z = \sum_{\mathcal{X}} e^{-\beta H[\mathcal{X}]} \equiv \sum_{\mathcal{X}} e^{-\mathcal{K}[\mathcal{X}]} \tag{1}$$

with the inverse temperature $\beta = 1/k_B T$ and the reduced Hamiltonian $\mathcal{K} \coloneqq -\beta H$. Equivalently, the system is specified by a probability measure:

$$P(\mathcal{X}) = \frac{1}{Z} e^{\mathcal{K}[\mathcal{X}]}. \tag{2}$$

The coarse-graining transformation $\mathcal{X} \to \mathcal{X}'$ between the set of the original d.o.f. and a (smaller) set of new d.o.f. is given by a conditional probability distribution $P_\Lambda(\mathcal{X}'|\mathcal{X})$, where Λ is a set of parameters completely specifying the rule (the rule can be totally deterministic, in which case P_Λ is a (δ) function). The probability measure of the coarse-grained system is then

$$P(\mathcal{X}') = \sum_{\mathcal{X}} P_\Lambda(\mathcal{X}'|\mathcal{X}) P(\mathcal{X}). \tag{3}$$

If $P(\mathcal{X}')$ is (or can be approximated by) a Gibbs measure, then the requirement to correctly reproduce thermodynamics enforces $Z' = Z$, and a renormalized Hamiltonian $H'[\mathcal{X}']$ in the new variables $\mathcal{X}'$ is defined implicitly via

$$e^{\mathcal{K}'[\mathcal{X}']} = \sum_{\mathcal{X}} P_\Lambda(\mathcal{X}'|\mathcal{X}) e^{\mathcal{K}[\mathcal{X}]}. \tag{4}$$

The procedure is often implemented in the form of a block RG [21,35] corresponding to a factorization of the conditional probability distribution into independent contributions from equivalent (assuming translation invariance) blocks $\mathcal{V} \subset \mathcal{X}$:

$$P(\mathcal{X}'|\mathcal{X}) = \prod_{j=1}^{n} P_\Lambda(\mathcal{H}_j|\mathcal{V}_j), \tag{5}$$

where $\{\mathcal{V}_j\}_{j=1}^n$ and $\{\mathcal{H}_j\}_{j=1}^n$ are disjoint partitions of $\mathcal{X}$ and $\mathcal{X}'$, respectively, and P_Λ now defines the coarse graining of a single block (and therefore, Λ contains substantially fewer parameters). Concrete examples of such a P_Λ include the standard "decimation" or "majority-rule" transformations [see Eqs. (13) and (14)].

Not every choice of P_Λ is *physically* meaningful. It should at least be consistent with the symmetries of the system, for instance. This is, however, not sufficient in practice. While it may be difficult to formulate a concise criterion for the choice of the coarse-graining transformation, it is clear that in order to derive the recursive RG equations the effective Hamiltonian cannot proliferate new couplings at each step. If there is to be a chance of analytical control over the procedure, the interactions in the effective Hamiltonian should be tractable (short ranged, for instance). That is to say, if one chooses the "correct" d.o.f. to describe the system, the resulting theory should be "simple." Numerous examples of failure to achieve this can be found in the literature [33,34] and include cases as simple as decimation of the Ising model in 2D. Implicit in this discussion is the notion that there does not exist a single RG transformation which does the job, but rather the transformation should be designed for the problem at hand [36].

Recently, some of us proposed the maximization of RSMI (introduced below) as a criterion for a physically meaningful RG transformation [28]. The idea behind it is that the effective block d.o.f., in whose terms the long-wavelength theory is simple, are those which retain most of the information (already present in the block) about long-wavelength properties of the system. This informally introduced "information" can be formalized by the following construction. Consider a single block $\mathcal{V}$ at a time and divide the system into four disjoint regions $\mathcal{X} = \mathcal{V} \cup \mathcal{B} \cup \mathcal{E} \cup \mathcal{O}$: the visibles (i.e., the block) $\mathcal{V}$, the buffer $\mathcal{B}$, the environment $\mathcal{E}$, and the remaining outer part of the system $\mathcal{O}$ (which is introduced only for algorithmic reasons, conceptually the environment $\mathcal{E}$ could also contain this part). Figure 2 depicts this decomposition in the case of a 1D spin model, but it trivially generalizes to any dimension. The real-space mutual information between the new (coarse-grained) d.o.f. $\mathcal{H}$ and the environment $\mathcal{E}$ of the original ones (i.e., of the block) is then defined as

$$I_\Lambda(\mathcal{H}:\mathcal{E}) = \sum_{\mathcal{H},\mathcal{E}} P_\Lambda(\mathcal{E},\mathcal{H}) \log\left(\frac{P_\Lambda(\mathcal{E},\mathcal{H})}{P_\Lambda(\mathcal{H})P(\mathcal{E})}\right), \tag{6}$$

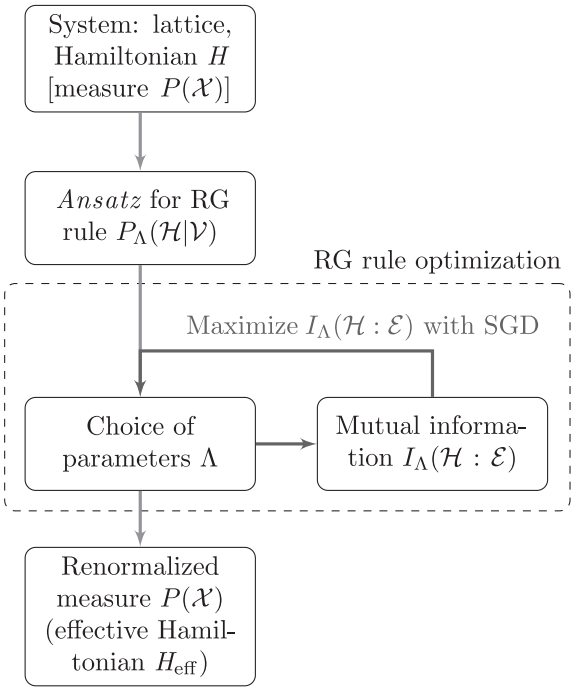

FIG. 1. Flow diagram of the RSMI algorithm [28]. Given a lattice and Hamiltonian H (or, in practice, given Monte Carlo samples) a RBM *Ansatz* for the RG rule is optimized by maximizing the mutual information between the new d.o.f. $\mathcal{H}$ and the environment $\mathcal{E}$ of the original ones using stochastic gradient descent. The trained $P_\Lambda(\mathcal{H}|\mathcal{V})$ is used to define a new effective measure and Hamiltonian H_{eff}.

where $P_\Lambda(\mathcal{E},\mathcal{H})$ and $P_\Lambda(\mathcal{H})$ are marginal distributions of $P_\Lambda(\mathcal{H},\mathcal{X}) = P_\Lambda(\mathcal{H}|\mathcal{V})P(\mathcal{X})$. Thus, $I_\Lambda(\mathcal{H}:\mathcal{E})$ is the standard mutual information between the random variables $\mathcal{H}$ and $\mathcal{E}$. Exclusion of the buffer $\mathcal{B}$ (in contrast to other adaptive schemes; see, for instance, Ref. [37]), generally of linear extent comparable to $\mathcal{V}$, is of fundamental importance: It filters out short-range correlations, leaving only the long-range contributions to $I_\Lambda(\mathcal{H}:\mathcal{E})$.

The RSMI satisfies the bounds (see also Appendix A):

$$0 \le I_\Lambda(\mathcal{H}:\mathcal{E}) \le H(\mathcal{H}), \tag{7}$$

$$I_\Lambda(\mathcal{H}:\mathcal{E}) \le I(\mathcal{V}:\mathcal{E}), \tag{8}$$

where $H(\mathcal{H})$ denotes the information entropy of $\mathcal{H}$, and $I(\mathcal{V}:\mathcal{E})$ is the mutual information of the visibles with the environment. The optimization algorithm starts with a set of samples drawn from $P(\mathcal{X})$ and a differentiable *Ansatz* for $P_\Lambda(\mathcal{H}|\mathcal{V})$, which in Ref. [28] takes the form of a restricted Boltzmann machine (RBM) parametrized by Λ (see Appendix C 3) and updates the parameters using a (stochastic) gradient descent procedure. The cost function to be maximized is precisely $I_\Lambda(\mathcal{H}:\mathcal{E})$, which in the course of the training is increased toward the value of $I(\mathcal{V}:\mathcal{E})$. The iterative procedure is shown in Fig. 1. Using the trained $P_\Lambda(\mathcal{H}|\mathcal{V})$, the original set of samples drawn from $P(\mathcal{X})$ can

be coarse grained and the full procedure recomputed for a subsequent RG step.

III. OPTIMALITY: THE MEASURE AND THE EFFECTIVE HAMILTONIAN

In what sense is the RSMI coarse graining *optimal*? By construction, the scheme preserves as much information as possible about the long-range properties of the system, and thus, when viewed as a compression of the *relevant* information in $\mathcal{V}$ into $\mathcal{H}$, it is information-theoretically optimal [38]. We show that this well-defined but abstract notion implies *physical* "simplicity" of the renormalized Hamiltonians. The latter property, though intuitively clear and operationally useful, may be difficult to define unambiguously. We therefore examine natural measures of Hamiltonian complexity and show they *all* decay with increased MI, also for disordered systems. It proves useful to approach this problem at the level of properties of the probability measure (which is the fundamental object the RSMI algorithm works with).

Consider first the following setup: Given a 1D system with a short-ranged Hamiltonian, introduce a coarse graining $\{\mathcal{V}_j\}$ with a block size chosen so that the Hamiltonian is the nearest neighbor with respect to the blocks. Let us choose an arbitrary block $\mathcal{V}_0$, denote its immediate neighbors $\mathcal{V}_{\pm 1}$ as the buffer $\mathcal{B}$, and all the remaining blocks $\{\mathcal{V}_{j<-1}\}$ and $\{\mathcal{V}_{j>1}\}$ as the environment $\mathcal{E}_0$, or in more detail, as the left and right environment $\mathcal{E}_{L/R}(\mathcal{V}_0)$, respectively. Assume now that $\mathcal{H}_0$, the coarse-grained variable for $\mathcal{V}_0$, is constructed so that $I(\mathcal{H}_0{:}\mathcal{E}_0) = I(\mathcal{V}_0{:}\mathcal{E}_0)$; i.e., the coarse-grained variable retains all of the information which the original block $\mathcal{V}_0$ contained about the environment and thus about any long-wavelength physics. The following then holds true (proof in Appendix B):

Proposition 1. Let $I(\mathcal{H}_0{:}\mathcal{E}_0) = I(\mathcal{V}_0{:}\mathcal{E}_0)$. Then the probability measure on the coarse-grained variables $P(\{\mathcal{H}_j\})$ obeys the factorization property:

$$P(\mathcal{H}_{j\leq -2}, \mathcal{H}_{j\geq 2}|\mathcal{H}_0) = P(\mathcal{H}_{j\leq -2}|\mathcal{H}_0)P(\mathcal{H}_{j\geq 2}|\mathcal{H}_0), \quad (9)$$

where in the conditional probabilities, the buffer variables (i.e., the neighbors $\mathcal{H}_{\pm 1}$ of $\mathcal{H}_0$) are integrated out. In other words, for a fixed $\mathcal{H}_0$, the probabilities of its environments $\mathcal{E}_{L/R}(\mathcal{H}_0)$ are independent of each other.

An immediate consequence of the above is the following corollary:

Corollary 1. The effective Hamiltonian does not contain terms directly coupling $\mathcal{E}_L(\mathcal{H}_0)$ and $\mathcal{E}_R(\mathcal{H}_0)$.

This is because the factorization Eq. (9) implies

$$\begin{aligned} E(\mathcal{H}_{j\leq -2}, \mathcal{H}_0, \mathcal{H}_{j\geq 2}) &\propto \log[P(\mathcal{H}_{j\leq -2}, \mathcal{H}_0, \mathcal{H}_{j\geq 2})] \\ &= E(\mathcal{E}_L, \mathcal{H}_0) + E(\mathcal{E}_R, \mathcal{H}_0) + E(\mathcal{H}_0). \end{aligned} \quad (10)$$

Since the variables $\mathcal{E}_L(\mathcal{H}_0) \coloneqq \{\mathcal{H}\}_{j<-1}$ and $\mathcal{E}_R(\mathcal{H}_0) \coloneqq \{\mathcal{H}\}_{j>1}$ are decoupled *after* integrating out the buffer $\mathcal{H}_{\pm 1}$, there generally would not have been any longer-range interaction (in particular, next-nearest neighbor) involving $\mathcal{H}_{\pm 1}$ in the renormalized Hamiltonian, or the measure would not factorize. Since the choice of $\mathcal{V}_0$, $\mathcal{E}_0$ is arbitrary in the first place, we have the following corollary:

Corollary 2. For a finite-range Hamiltonian, if $I(\mathcal{H}_0{:}\mathcal{E}_0) = I(\mathcal{V}_0{:}\mathcal{E}_0)$, the RSMI coarse graining is guaranteed not to increase the range of interactions.

These observations generalize, under very mild additional assumptions, to any dimension D. Taking a coarse graining
with blocks sufficiently large to make the short-ranged Hamiltonian nearest neighbor, and under the assumption of full information capture, we repeat the above reasoning, conditioning on—instead of a single arbitrary variable $\mathcal{H}_0$—a hyperplane of dimension $D-1$ separating the coarse-grained variables $\{\mathcal{H}_j\}$ into two disconnected sets to show that no longer-ranged interactions across the hyperplane can exist. Since the choice of hyperplane is arbitrary, the effective Hamiltonian is nearest neighbor, as the original one was (see Appendix B). A *perfect* RSMI scheme does not, therefore, increase the range of a short-ranged Hamiltonian, i.e., its complexity.

While a strong property, the above results appear to have one serious shortcoming: For a generic physical system and coarse-graining scheme, it may not be possible to satisfy the assumption $I(\mathcal{H}_0{:}\mathcal{E}_0) = I(\mathcal{V}_0{:}\mathcal{E}_0)$, which is a strict upper bound on $I(\mathcal{H}_0{:}\mathcal{E}_0)$, for any RG rule. This is due to the fact that the block size, as well as the number and character (Ising spin, Potts spin, etc.) of coarse-grained variables are usually chosen *a priori*, and given those constraints a solution $P_\Lambda(\mathcal{H}|\mathcal{V})$ satisfying $I(\mathcal{H}_0{:}\mathcal{E}_0) = I(\mathcal{V}_0{:}\mathcal{E}_0)$ is not mathematically guaranteed to exist (see Sec. V for examples). This limitation, however, is only a superficial problem. First, Proposition 1 is a sufficient, not a necessary, condition. Much more importantly, the RSMI prescription is a variational principle. If the physics of the problem and constraints imposed preclude the existence of a "perfect" solution, as is usually the case, the maximization of RSMI still yields the best possible one, given the conditions. A mathematical proof of this statement requires establishing the decay of some measures of the effective Hamiltonian complexity (such as range and the ones we consider below) as a function of the mutual information. In the absence of such a rigorous result, in what follows we instead study, analytically and numerically, tractable models and verify that this decay indeed holds empirically; i.e., the more mutual information the RG rule captures, the smaller the complexity of the effective Hamiltonian. Furthermore, we show this also holds in the presence of disorder (see Sec. VI).

We now investigate a realistic setup, in which the RSMI is maximized under the constraint of number and type of coarse-grained d.o.f. Additionally, since the RG rule is optimized iteratively, we study the approach to the optimal solution via the properties of the renormalized Hamiltonian

defined by the RG rule at any stage of the procedure. We briefly review how the effective Hamiltonian can be expressed by appropriate cumulant expansion [35] (though the RSMI algorithm deals with a probability measure as the basic object, and at no point computes the Hamiltonian, the Hamiltonian is more interpretable physically), and we apply this machinery to the Ising chain with and without disorder.

IV. SOLUBLE EXAMPLE: ARBITRARY RG TRANSFORMATIONS OF THE CLEAN 1D ISING MODEL

To investigate the relationship between the renormalized Hamiltonian and the real-space mutual information for practical coarse-graining procedures, we consider the example of the one-dimensional Ising model with nearest-neighbor interactions and periodic boundary conditions. We deliberately use this simple model, since it allows us to analytically derive properties not only of the optimal RG procedure (which we do first) but also those of *arbitrary* coarse grainings: Both the effective Hamiltonian and the amount of RSMI captured can be calculated explicitly and without any arbitrary truncations to establish the relation between them. The Ising Hamiltonian reads

$$\mathcal{K}[\mathcal{X}] = K \sum_{i=1}^{N} x_i x_{i+1} \tag{11}$$

with $x_i = \pm 1$ collectively denoted by $\mathcal{X} = \{x_i\}_{i=1}^N$ and with $K \coloneqq -\beta J$. The sizes of the block, buffer, and environment regions introduced in Sec. II are given by $L_\mathcal{V}$, $L_\mathcal{B}$, and $L_\mathcal{E}$. Accordingly, there are $n = N/L_\mathcal{V}$ blocks.

To best illustrate the results, we now specialize to the (typical) case of blocks of two visible spins $\mathcal{V} = \{v_1, v_2\}$ coarse grained into a single hidden spin h (computations for general $L_\mathcal{V}$ are analogous). The RG rule is parametrized by a RBM *Ansatz*:

$$P_\Lambda(\mathcal{H}|\mathcal{V}) = \frac{1}{1 + e^{-2h\sum_i \lambda_i v_i}} \tag{12}$$

with $\Lambda = (\lambda_1, \lambda_2)$ describing the quadratic coupling of visible to hidden spins (see Appendix C 3 for a discussion of the *Ansatz*). In Fig. 2, the decomposition of the system and the RG rule are schematically shown.

The standard decimation and the majority-rule coarse-graining schemes are given in our language by

$$P_{\mathrm{dec}}(h|\{v_1, v_2\}) = \begin{cases} 1, & h = v_1, \\ 0, & h \neq v_1, \end{cases} \tag{13}$$

and by

$$P_{\mathrm{maj}}(h|\{v_1, v_2\}) = \begin{cases} 1, & v_1 = v_2 = h, \\ 0, & v_1 = v_2 \neq h, \\ \frac{1}{2}, & v_1 \neq v_2, \end{cases} \tag{14}$$

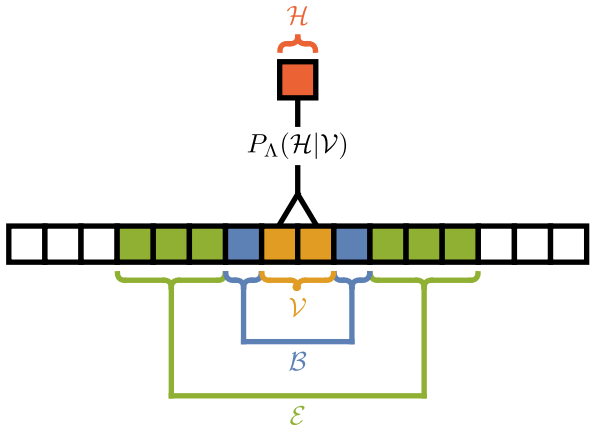

FIG. 2. Schematic decomposition of the system for the purpose of defining the mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$ (in 1D, for concreteness). The full system is partitioned into blocks of visibles $\mathcal{V}$ (yellow) embedded into a buffer $\mathcal{B}$ (blue) and surrounded by the environment $\mathcal{E}$ (green). The remaining part of the system is denoted by $\mathcal{O}$ in the main text. The conditional probability distribution $P_\Lambda(\mathcal{H}|\mathcal{V})$ couples $\mathcal{V}$ to the hiddens $\mathcal{H}$ (red).

respectively. They correspond to the choice of $\Lambda_{\mathrm{dec}} = (\lambda, 0)$ and $\Lambda_{\mathrm{maj}} = (\lambda, \lambda)$ in the limit $\lambda \to \infty$.

For the case of decimation, an exact calculation using the transfer-matrix approach yields an effective Hamiltonian of the same nearest-neighbor form, albeit with a renormalized coupling constant [39,40]:

$$K' = \frac{1}{2}\log[\cosh(2K)]. \tag{15}$$

For the majority rule, and any other choice of parameters Λ, the renormalized Hamiltonian cannot be obtained in a closed form, but it can still be derived analytically. To this end, we split it into two parts [35]:

$$\mathcal{K}[\mathcal{X}] = \mathcal{K}_0[\mathcal{X}] + \mathcal{K}_1[\mathcal{X}], \tag{16}$$

where $\mathcal{K}_0$ contains *intrablock* and $\mathcal{K}_1$ *interblock* terms. Using the cumulant expansion, the new Hamiltonian is given perturbatively:

$$\mathcal{K}'[\mathcal{X}'] = \log[Z_0 P_{\Lambda,0}(\mathcal{X}')] + \sum_{k=0}^{\infty} \frac{1}{k!} C_k[\mathcal{X}'], \tag{17}$$

where the cumulants C_k can be expressed in terms of averages of the form $\langle \mathcal{K}_1[\mathcal{X}]^k \rangle_{\Lambda,0}$, which factorize into averages of operators from a single block (see Appendix C 1 for details). The renormalized coupling constants are not apparent in Eq. (17). In order to identify them, we introduce the following canonical form of the Hamiltonian:

$$\mathcal{K}'[\mathcal{X}'] = K'_0 + \sum_{\{\alpha_\ell\}_{\ell=1}^n} K'_{\alpha_1,\alpha_2,\ldots,\alpha_n} \left(\sum_{j=1}^{n} \prod_{\ell=1}^{n} (x'_{j+\ell})^{\alpha_\ell} \right), \tag{18}$$

with $\alpha_1 = 1$ and $\alpha_\ell \in \{0, 1\}$ for all $\ell > 1$. Here, the addition of the indices is to be understood modulo n

(i.e., with periodic boundary conditions). Note that arbitrary orders k of the cumulant expansion C_k contribute to each coupling constant $K'_{\alpha_1,\alpha_2,\ldots,\alpha_n}$.

In the example of the Ising model, the only nonvanishing averages contributing to the cumulants are

$$\langle v_1\rangle_{\Lambda,b}[h] =: a_1 h, \tag{19a}$$

$$\langle v_2\rangle_{\Lambda,b}[h] =: a_2 h, \tag{19b}$$

$$\langle v_1 v_2\rangle_{\Lambda,b}[h] =: b, \tag{19c}$$

with the effective block parameters a_1, a_2, b independent of the coarse-grained variable h and functions of Λ and K only, whose closed form expressions can easily be found (see Appendix D). Consequently, the averages $\langle \mathcal{K}_1^k\rangle_{\Lambda,0}$, and thus, also the Hamiltonian $\mathcal{K}'$, are polynomials in the new d.o.f. $\mathcal{X}'$, the reduced temperature K, and the block parameters, which gives rise to Eq. (18). In practice, the cumulant expansion is terminated at a finite order M, which results in an expansion of $\mathcal{K}'$ and thus of each coupling constant $K'_{\alpha_1,\alpha_2,\ldots,\alpha_n}$ up to that order in K. All the information about the RG rule (except for the size of $\mathcal{H}$, which is fixed at the outset) is contained in the dependence of the effective block parameters on Λ (and on N, K).

Expressing the moments $\langle \mathcal{K}_1^k\rangle_{\Lambda,b}$ appearing in the cumulant expansion in terms of the new variables $\mathcal{X}'$ is a combinatorial problem. Each term in $\mathcal{K}_1$ couples spins from neighboring blocks j and $j+1$ so that

$$\mathcal{K}_1[\mathcal{X}]^k = K^k \sum_{j_1,\ldots,j_k=1}^{n} \prod_{\ell=1}^{k} x_{2j_\ell} x_{2j_\ell+1}. \tag{20}$$

The average of each summand factorizes into contributions from each block, whose value [see Eq. (19)] is determined by the arrangement of $j_1,\ldots,j_k$. Thus, the calculation is reduced to finding and grouping all equivalent (under the fact that for Ising variables $x_j^2 = 1$) configurations $(j_1,\ldots,j_k)$. Bringing the resulting polynomial in canonical form (18) is an inverse problem and is solved by recursively eliminating noncanonical terms. For a given M, we can thus finally arrive at an expression of coupling constants $K'_{\alpha_1,\alpha_2,\ldots,\alpha_n}$ as functions of Λ and K (see Appendix D for details).

We are now in a position to examine the effective Hamiltonian obtained by applying the RSMI-maximization procedure Fig. 1 to the model Eq. (11). Anticipating the results in Fig. 4, in Fig. 3 we compare, for varying K and order of cumulant expansion M, the renormalized nearest-neighbor (NN) coupling obtained in the RSMI-favored solution with the exact nonperturbative one Eq. (15) (which we refer to as "exact decimation"). The two results converge with increasing M, and the convergence is faster for weak coupling or higher temperatures, which is unsurprising since the cumulant expansion is in powers of K. We emphasize again that the RSMI algorithm itself works on the level of the probability measure, and at no point does it compute the effective Hamiltonian. It is only when we want to examine the renormalized Hamiltonian which the converged—in the sense of saturating the mutual information during optimization of the Λ parameters—RSMI solution corresponds to, that we are performing the cumulant expansion.

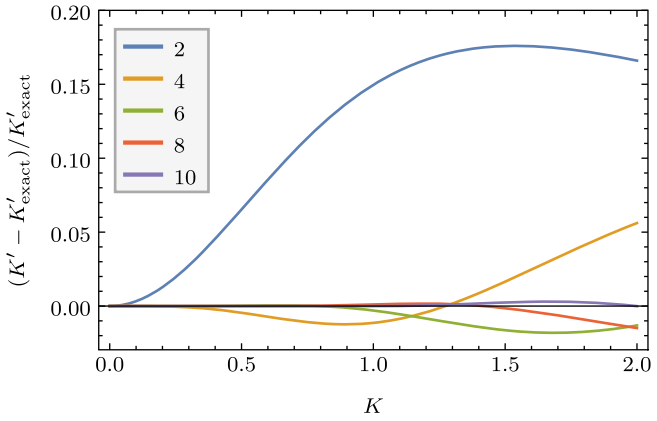

FIG. 3. The relative difference between the renormalized NN coupling obtained from the cumulant expansion of the RSMI-favored solution (decimation) and the nonperturbative result Eq. (15). The convergence improves with increasing order of cumulant expansion M and lower K.

Since exact decimation leads to a strictly NN effective Hamiltonian in the 1D Ising case, and since perturbatively the RSMI-favored solution converges to the decimation value for the NN coupling, it is instructive to inspect the behavior of the m-body couplings in the effective Hamiltonian for larger order m. Denoting the m-spin coupling with distances $\ell_1,\ell_2,\ldots,\ell_m$ between the spins by $K_m(\ell_1,\ell_2,\ldots,\ell_m)$, with $K_m(\ell)$ short for $K_m(\ell,\ell,\ldots,\ell)$, we observe that, in the limit of weak coupling (small K), both $\ell \mapsto K_2(\ell)$, i.e., arbitrary range two-body interactions, as well as $m \mapsto |K_m(1)|$, i.e., arbitrary order NN interactions, decay exponentially. This behavior is shown in Figs. 10 and 11 in Appendix D. The decay length is characterized by $K_2(2)/K_2(1)$ and $K_m(1)/K_2(1)$, respectively. Thus, the RSMI approach indeed converges to the exact decimation in this case, which is known to be the optimal choice.

To further strengthen the link between the amount of RSMI retained and the resulting properties of the effective Hamiltonian, we now consider a *generic* coarse graining, suboptimal from the RSMI perspective (i.e., away from the maximum the RSMI algorithm strives for). To this end, we compute the mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$ captured for the Ising model by a general coarse-graining rule Eq. (12) with parameters $\Lambda = (\lambda_1,\lambda_2)$. This calculation can be performed exactly using the transfer-matrix method (see Appendix D 4) and yields a closed-form expression Eq. (D30).

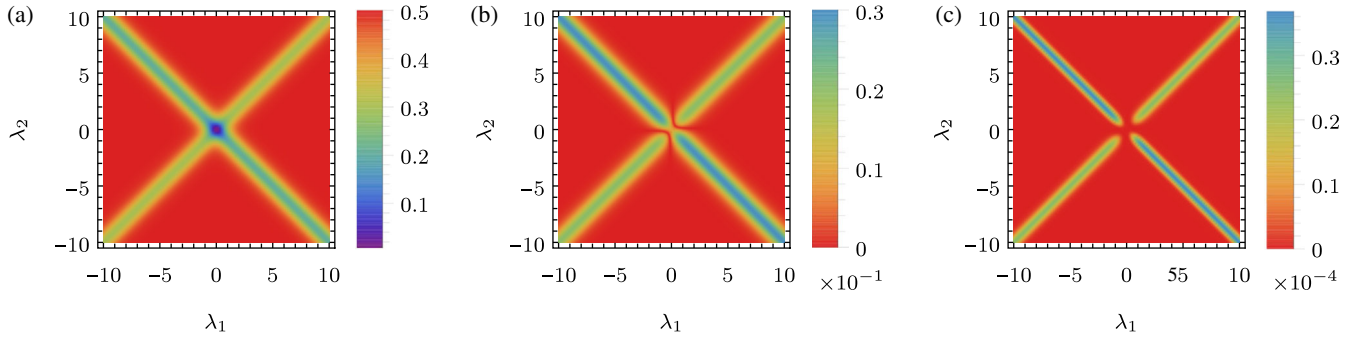

FIG. 4. Density plots of (a) the mutual information of the hiddens with the environment scaled to the mutual information of the visibles with the environment $I_\Lambda(\mathcal{H}:\mathcal{E})/I(\mathcal{V}:\mathcal{E})$, (b) the ratio of the NNN to the NN coupling constants $|K_2'(2)/K_2'(1)|$, and (c) the ratio of the NN four-point to two-point coupling constants $|K_4'(1,1,1)/K_2'(1)|$. All three quantities are shown as a function of the parameters of the RG rule $\Lambda = (\lambda_1, \lambda_2)$. Note the inverted color scale in (b) and (c). For large enough $||\Lambda||^2$, a maximum of mutual information corresponds to a minimum of rangeness and m-bodyness, and vice versa. See the main text and Appendix D 3 for details.

Equipped with these results, for an arbitrary coarse graining defined by a choice of Λ, we can now compute both the amount of mutual information with the environment retained (RSMI), as well as the effective Hamiltonian generated. In Fig. 4(a), the amount of information captured is shown as a function of (λ_1, λ_2) in units of $I(\mathcal{V}:\mathcal{E})$ (for concreteness, all plots are for $K = 0.1$ and a single site buffer $L_\mathcal{B} = 1$). A few observations can be made: The choices of Λ retaining more RSMI are not symmetric in $|\lambda_1|$ and $|\lambda_2|$ but instead tend to $(\pm\lambda, 0)$ and $(0, \pm\lambda)$ for large enough $|\lambda|$; i.e., they resemble decimation Eq. (13) (the four plateaux in Fig. 4 are not exactly flat, as also examined in Fig. 5), as opposed to majority rule Eq. (14) which, in fact, captures the least information. The symmetries of the plot are due to global $\mathbb{Z}_2$ Ising symmetry as well as an additional $\mathbb{Z}_2$ symmetry of the mutual information: Correlation and anticorrelation for random variables are equivalent from the point of view of information. Furthermore, the lack of information retained for small $||\Lambda||^2$ is due to the fact that in this case the coarse-graining Eq. (12) only weakly depends only on the visible spins and is essentially randomly assigning the value of the hidden spin (i.e., it is dominated by random noise). In other words, it only makes sense to think of Eq. (12) as a coarse graining if it strongly depends on the original spins, i.e., for large $||\Lambda||^2$.

The properties of the corresponding effective Hamiltonians can be understood with the help of Figs. 4(b) and 4(c), where the ratio of next-nearest-neighbor (NNN) to NN terms as well as the ratio of NN four-body to two-body terms in the effective Hamiltonian are plotted as a function of Λ (note the inverted color scale). It is apparent that decimationlike choices, which maximize RSMI, result also in vanishing NNN and four-body terms (and more generally, long-range or high-order terms, as discussed previously and shown in Figs. 10 and 11 in Appendix D). This is examined in more detail in Fig. 5: Trajectories in the parameter space Λ are chosen according to $\lambda[\cos(\theta), \sin(\theta)]$ with $\theta \in [0, \pi]$, for different magnitudes $|\lambda|$. The ratios in Figs. 4(b) and 4(c), which we dub "rangeness" and "m-bodyness" for brevity, are plotted against the mutual information along the trajectories. The mutual information is maximized for $\theta = 0$ and $\theta = \pi$, and the maximum increases with λ (though it saturates, there is little difference between $\lambda = 3$ and $\lambda = 1000$). Simultaneously, for large enough $|\lambda|$, both ratios in Figs. 5(b) and 5(c) vanish, rendering the effective Hamiltonian two body and nearest neighbor. It is now clear how the RSMI maximization results in a decimation coarse graining for the 1D Ising model. A more detailed discussion of Figs. 4 and 5 [including asymmetries in Fig. 5(a) and accidental vanishings in Fig. 5(b)] can be found in the Appendix D, but it does not change the general picture: Maximizing RSMI results in the decay of longer-ranged and higher-order terms in the Hamiltonian.

The superiority of decimation over majority rule in our example can be understood intuitively from a physical perspective by considering fluctuations of the original (visible) spins for a fixed (clamped) configuration of the new variables $\mathcal{X}'$. In 1D, decimation fixes every other spin in $\mathcal{X}$, which prevents all but isolated fluctuations of the remaining d.o.f., which are being integrated out in the clamped averages of Eqs. (C3) and (C4). Consequently, only nearest neighbors in $\mathcal{X}'$ are coupled in the effective Hamiltonian. In contrast, the majority rule fixes a linear combination of the visibles (the average), thereby allowing fluctuations of orthogonal linear combinations. These fluctuations can span multiple blocks and thus generate higher-order coupling terms. In the following section, an alternative information-based intuition is offered, which also explains the difference between the optimal coarse-graining procedures in 1D and 2D.

Finally, we note that the results that we describe above from a static perspective, i.e., considering properties of arbitrary coarse graining, for a fixed, potentially suboptimal, choice of Λ, can also be interpreted dynamically. In this sense, they characterize the convergence of the

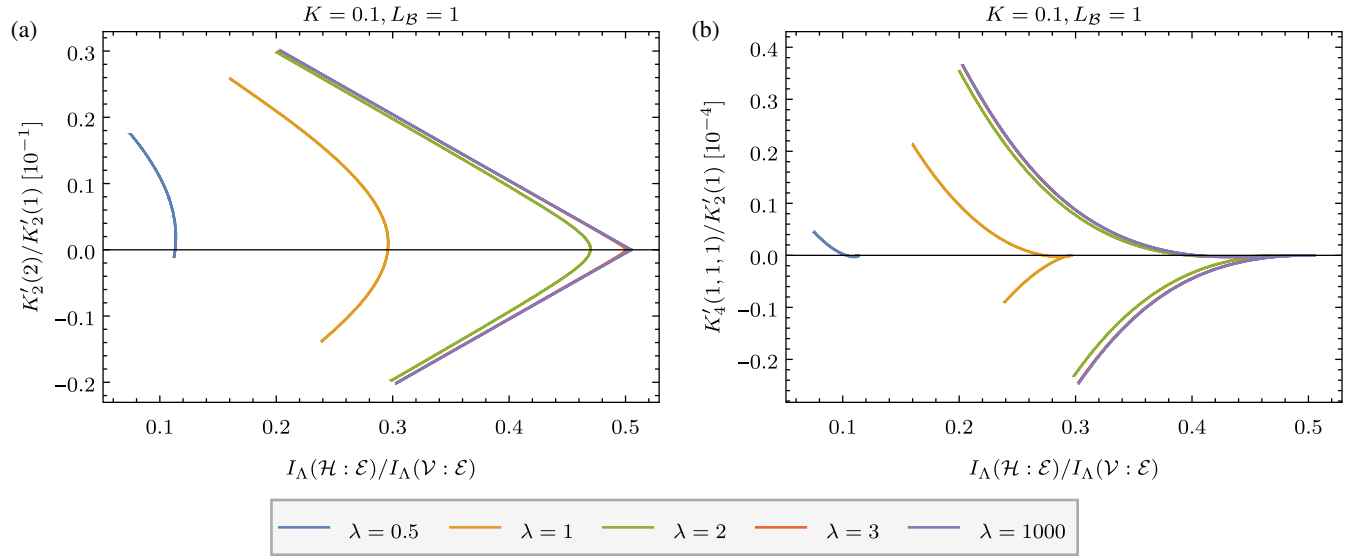

FIG. 5. The two proxy measures of complexity of the renormalized Hamiltonian discussed in the text are shown against mutual information retained: the rangeness, i.e., the ratio of the NNN to the NN coupling constants, and the m-bodyness, i.e., the ratio of the NN four-point to two-point coupling constants. The mutual information is scaled to the total mutual information the block $\mathcal{V}$ shares with the environment. The curves are obtained by parametrizing the RG rule as $\lambda[\cos(\theta), \sin(\theta)]$ and varying $\theta \in [0, \pi]$ for different magnitudes of λ. In the physically relevant limit of large λ, the maximum of mutual information corresponds to a minimum of rangeness and m-bodyness. The plots are discussed in more detail in Appendix D 3.

RSMI algorithm of Ref. [28] as the Λ parameters are iteratively optimized during the training (see Fig. 1).

V. THE "SHAPE" OF THE COARSE-GRAINED VARIABLES

So far, we motivated on physical grounds (the properties of the effective Hamiltonian) why maximizing RSMI generally provides a guiding principle for constructing a real-space RG procedure. We then investigated the example of the 1D Ising system the properties of such a scheme in a typical situation, when the RSMI-maximization problem is additionally constrained by the number and type of d.o.f. the system is coarse grained into. In particular, we gave physical intuitions which explain the solution RSMI converges to in the 1D case, i.e., decimation. This outcome is to be contrasted with the situation in 2D, when the decimation procedure is known to immediately generate long-range and many-spin interactions and can be shown not to posses a nontrivial fixed point at all [33]. For the square-lattice Ising model in two dimensions, the majority-rule transformation is preferable: Numerical evidence, at least, points to the existence of a fixed point [41]. Remarkably, the RSMI solution in 2D converges (numerically) toward a majority-rule block transformation (for two-by-two blocks) [28]. In this section, we provide an information-theory-based explanation of these observations. In doing so, we also elucidate and quantify the nontrivial influence on the RG scheme of the constraints imposed by the properties (type and number) of the new coarse-grained variables for the general case. Finally, we exemplify our findings using simple and intuitive toy models.

To this end, let us revisit the inequality Eq. (8). We refine it by explicitly introducing the random variables $\mathcal{V}_\Lambda$, which the hidden d.o.f. $h_i \in \mathcal{H}$ couple to in a RG scheme parametrized by $\Lambda = \{\lambda_{ij}\}$. For instance, in the RBM parametrization discussed previously, while generically $\mathcal{H}$ depends on the full $\mathcal{V}$, the coarse graining defined by the conditional probability $P_\Lambda(\mathcal{H}|\mathcal{V})$ makes each $h_i \in \mathcal{H}$ dependent only on the combination

$$\mathcal{V}_{\Lambda_i} = \frac{1}{||\Lambda_i||}\sum_j \lambda_{ij} v_j. \tag{21}$$

Note that the overall normalization in the definition is not important but only the relative strengths of λ_{ij} which define the linear combination of d.o.f. in the block. The following now holds:

$$I_\Lambda(\mathcal{H}:\mathcal{E}) \leq I(\mathcal{V}_\Lambda:\mathcal{E}) \leq I(\mathcal{V}:\mathcal{E}). \tag{22}$$

That is, the information about the environment carried by the particular chosen variables $\mathcal{V}_\Lambda$ is potentially smaller that the overall information about the environment contained in the block $I(\mathcal{V}:\mathcal{E})$. Still less of the information may ultimately be encoded in the d.o.f. $\mathcal{H}$.

Where do the inequalities Eq. (22) originate from? Formally, this is because we have a Markov chain

$$\mathcal{E} \to \mathcal{V} \to \mathcal{V}_\Lambda \to \mathcal{H}. \tag{23}$$

But the more pertinent question is, what can make those inequalities sharp? The second one is rather trivial: If we decide only to keep a few (one, as is often the case) variables $\mathcal{V}_{\Lambda_i}$, then their entropy may be simply too small to even store the full information $I(\mathcal{V}:\mathcal{E})$. Still, for the same entropy, there may be choices of Λ which result in bigger or smaller $I(\mathcal{V}_{\Lambda_i}:\mathcal{E})$. Crucially though, $I(\mathcal{V}_{\Lambda_i}:\mathcal{E})$ does not depend on the nature of $h_i \in \mathcal{H}$ (i.e., on whether h_i is a binary variable or not, for instance). It characterizes only how good the particular set of physical d.o.f. $\mathcal{V}_\Lambda$ is at describing fluctuations in the environment $\mathcal{E}$.

Whether this information can be efficiently encoded in $\mathcal{H}$ is a different question entirely. The answer, and the origin of the first inequality Eq. (22), is revealed by

$$I_\Lambda(\mathcal{H}:\mathcal{E}) = I(\mathcal{V}_\Lambda:\mathcal{E}) - I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H}), \tag{24}$$

where $I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$ is the conditional mutual information, and we use the chain rule and the Markov property Eq. (23). Since $I(\mathcal{V}_\Lambda:\mathcal{E})$ is independent of $\mathcal{H}$ in the sense described above, $I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$ quantifies the failure of the encoding into $\mathcal{H}$ due to the properties of the $\mathcal{H}$ itself (conditional mutual information being always non-negative). We thus manage to identify the contributions to $I_\Lambda(\mathcal{H}:\mathcal{E})$ resulting from coupling to a certain choice of physical modes in $\mathcal{V}$ and to isolate them from the losses incurred due to the impossibility of encoding this information perfectly in a particular type of $\mathcal{H}$.

The conditional probability distribution $I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$ can be thought of as describing the mismatch of the probability spaces of the random variables $\mathcal{H}$ and $\mathcal{V}_\Lambda$; it tells us how much information is *still* shared between $\mathcal{V}_\Lambda$ and $\mathcal{E}$ after $\mathcal{V}_\Lambda$ is restricted to only values compatible with a given outcome of $\mathcal{H}$. For example, in the 1D Ising case we examined previously, the majority rule defines $\mathcal{V}_\Lambda = v_1 + v_2$, for which the set of possible outcomes is equivalent to $\{-1, 0, 1\}$. The entropy of $\mathcal{V}_\Lambda$ is bounded by and possibly equal to $\log_2(3)$. Since the system is $\mathbb{Z}_2$ symmetric, then unless $\mathrm{Prob}[\mathcal{V}_\Lambda = 0] = 0$, this distribution cannot be faithfully encoded into *any* probability distribution of a single binary variable $\mathcal{H}$. Below, we construct simple toy models to provide more examples and intuition for the somewhat abstract notions we introduce here.

First, let us stress though, that the RSMI prescription maximizes $I_\Lambda(\mathcal{H}:\mathcal{E})$ as a whole, and that, for a type of $\mathcal{H}$ fixed at the outset, the procedure *cannot* be split into maximization of $I(\mathcal{V}_\Lambda:\mathcal{E})$ followed by a linear coupling of $\mathcal{H}$ to the $\mathcal{V}_\Lambda$ found. Such a naive greedy approach does not necessarily lead to an optimal solution; the toy models below provide an explicit counterexample. The RSMI-based solution of $P_\Lambda(\mathcal{H}|\mathcal{V})$ thus converges to the optimal trade-off between finding the best modes in $\mathcal{V}$ to describe $\mathcal{E}$ and finding those whose description can be faithfully written in $\mathcal{H}$ of a given type.

To illustrate the above considerations, we construct minimal toy models. In 1D this consists of four coupled Ising spins: v_1, v_2 in the block $\mathcal{V}$ and e_1, e_2 representing the left and right environment (in 1D the environment is not simply connected) with the Hamiltonian

$$\mathcal{K} = K_{\mathcal{V}\mathcal{E}}(e_1 v_1 + v_2 e_2) + K_{\mathcal{V}} v_1 v_2, \tag{25}$$

where as before, the coupling constants contain a factor of $\beta = 1/k_B T$. The two spins in $\mathcal{V}$ are coupled to a single hidden spin $\mathcal{H}$ using a RBM *Ansatz* Eq. (12), and the random variable $\mathcal{V}_\Lambda$ is defined as in Eq. (21). In Fig. 6, the results of the calculation of the mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$ and $I(\mathcal{V}_\Lambda:\mathcal{E})$ for decimation and the majority rule are shown. In the regime of strong coupling to the environment $K_{\mathcal{V}\mathcal{E}}$ [see Fig. 6(a)], for small $K_{\mathcal{V}}$ both visible spins are nearly independent and almost copy the state of the left and right environments, respectively. Consequently, $\mathcal{V}_\Lambda$ for the majority rule carries almost $\log_2(3)$ bits of information about the environment, while $\mathcal{V}_\Lambda$ for decimation, being a binary variable, at most one bit. However, when $I_\Lambda(\mathcal{H}:\mathcal{E})$ is examined, it becomes apparent that for decimation it is exactly equal to $I(\mathcal{V}_\Lambda:\mathcal{E})$, while for majority rule, it is significantly lower, so much so, that overall decimation is better across the whole parameter regime. The difference between the solid and dashed curves in Fig. 6(a) is precisely the mismatch of Eq. (24), and the above provides a counterexample to a greedy maximization of $I(\mathcal{V}_\Lambda:\mathcal{E})$ instead of $I_\Lambda(\mathcal{H}:\mathcal{E})$, which was mentioned previously. In the large-$K_{\mathcal{V}}$ limit, both spins in $\mathcal{V}$ become bound into an effective single binary variable, and the distinction between the two rules vanishes. In Fig. 6(b), we show the same in the regime when the spins in $\mathcal{V}$ are only weakly coupled to the environment (or the temperature is high). Again, decimation perfectly encodes information $I(\mathcal{V}_\Lambda:\mathcal{E})$ into $\mathcal{H}$ and is overall better.

Let us contrast this picture with the situation in higher (in particular, two) dimensions, when the environment is simply connected. Based on the discussion above, we may anticipate that the optimal solution could be different and that majority rule may instead be preferable. This is because, on the one hand, for the same coupling strength to the environment and the same linear dimensions $L_{\mathcal{V}} = 2$ of the block, the ratio of $I(\mathcal{V}_\Lambda:\mathcal{E})$ for the majority rule to the one for decimation increases with increasing dimension (a consequence of all visible spins interacting with the same environment). On the other hand, the mismatch $I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$ for the majority rule decreases compared to 1D, since the probability of $\mathcal{V}_\Lambda = \sum_i v_i$ being zero is smaller. This fact is due to both dimensional considerations as well as (again) the environment being simply connected, the importance of which, even in 1D, we illustrate in Appendix E.

We verify those expectations using a simple toy model of the 2D setting: The environment is represented by a single random variable E with a large number of states, to which all the spins in $\mathcal{V}$ couple. These states should be thought of intuitively as fluctuations of some large environment at wavelengths longer than the size of the coarse-graining cell. The Hamiltonian is

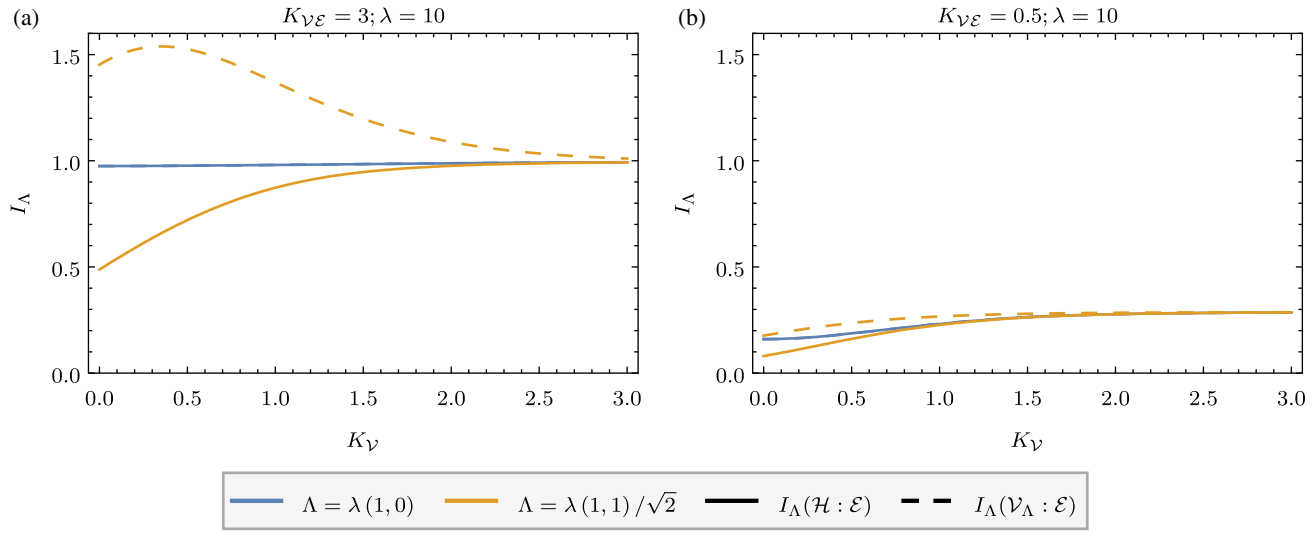

FIG. 6. The mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$ and $I(\mathcal{V}_\Lambda:\mathcal{E})$ for decimation (blue) and majority-rule (yellow) procedures in the 1D toy model Eq. (25). Two parameter regimes are shown: (a) strong coupling to the environment or low-temperature $K_{\mathcal{VE}}$ (recall that the coupling constants contain a factor of $\beta = 1/k_B T$) and (b) weak coupling $K_{\mathcal{VE}}$; note that the absolute value of all mutual information is lower in this limit. The solid lines differ from the dashed lines of the same color by the mismatch $I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$ (see main text). In both parameter regimes, the dashed blue line exactly coincides with the solid blue line: For the decimation procedure, the information $I(\mathcal{V}_\Lambda:\mathcal{E})$ is perfectly encoded into $\mathcal{H}$. Majority rule $I_\Lambda(\mathcal{H}:\mathcal{E})$ is inferior to decimation, even though $I(\mathcal{V}_\Lambda:\mathcal{E})$ is significantly larger: All that information is lost in encoding. The distinction between the two rules vanishes in the $K_\mathcal{V} \to \infty$ limit when both visible spins are effectively bound into a single binary variable.

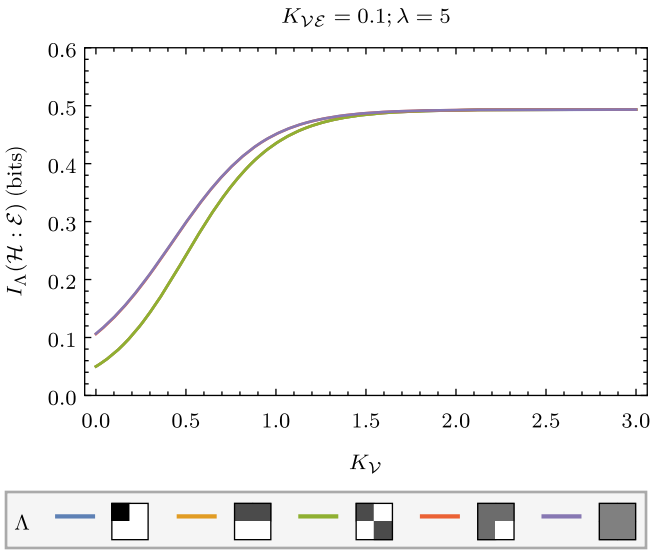

FIG. 7. Mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$ in the toy model of a 2D system Eq. (26) as a function of the coupling $K_\mathcal{V}$ between the visibles. The coupling pattern to the visibles in different RG rules Λ is shown schematically in the (physically relevant) large-$||\Lambda||^2$ limit. The majority rule (and interestingly, also coupling to three spins depicted by a red line coinciding with the purple one) consistently retains more information than decimation or any coupling to two spins (blue and yellow coinciding with green). Again, the distinction vanishes for large $K_\mathcal{V}$ when all visible spins are bound into a single one.

$$\mathcal{K} = K_{\mathcal{VE}} E(v_1 + v_2 + v_3 + v_4) + K_\mathcal{V}(v_1 v_2 + v_1 v_3 + v_2 v_4 + v_3 v_4). \quad (26)$$

As before, the spins in block $\mathcal{V}$ are coupled to a single hidden spin $\mathcal{H}$ with a RBM *Ansatz* parametrized by Λ.

In Fig. 7, the mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$ is computed for the model Eq. (26) for different course-graining rules given by Λ. Indeed, the decimation is now inferior to the majority rule across the full parameter range. This is also consistent with the known properties of decimation and majority rule for the 2D Ising model and suggests their information-theoretic origin.

VI. DISORDERED SYSTEMS

While investigations of clean higher-dimensional models (to which RSMI can be applied without any restriction), such as, e.g., the 3D Ising model, are still relevant, of much more interest are disordered systems. We show that RSMI naturally generalizes to this case and that the information-theoretic approach provides important insights, particularly concerning disorder correlations.

The proper object of study in the disordered setting is not the individual Hamiltonian H but rather the disordered Hamiltonian *distribution* $\mathcal{P}(H)$ [42–44], which equivalently can be thought of as a distribution over the vector space spanned by all the possible coupling constants $\{K_{i_1,i_2,\ldots,i_M}\}$.

Denoting the (potentially infinite-dimensional) vector of couplings by $\mathbf{K}$, the RG transformation induces a mapping

$$\mathcal{P}(\mathbf{K}) \to \mathcal{P}'(\mathbf{K}) \tag{27}$$

generating RG flows of $\mathcal{P}$ with fixed-point distributions $\mathcal{P}^*$. The formalism subsumes the clean case: Any fixed Hamiltonian is a trivial (δ)-like distribution with all probability mass concentrated in one point.

Let us examine the mapping Eq. (27). The probabilistic framework of Sec. II can also be used in this case. For any fixed disorder realization $\mathbf{K}$, the RG transformation (conditional probability distribution) is applied to the Gibbsian probability measure defined by the Hamiltonian $H(\mathbf{K})$, and the new effective Hamiltonian $H(\mathbf{K}')$ is implicitly defined exactly as in Eq. (4). The new coupling constants are in this way the functions of the old ones $\mathbf{K}' = \mathbf{K}'(\mathbf{K})$, and they can be recovered by solving the inverse problem. Their distribution is obtained by integrating over $\mathcal{P}(\mathbf{K})$:

$$\mathcal{P}'(\mathbf{K}') = \int \delta[\mathbf{K}' - \mathbf{K}'(\mathbf{K})]\mathcal{P}(\mathbf{K})\mathbf{dK}. \tag{28}$$

Equation (28) appears trivial, but, of course, all the complexity of the problem is concealed in the functional dependence of $\mathbf{K}'$ on $\mathbf{K}$. The distribution $\mathcal{P}$ is usually assumed to be factorized into a product of independent distributions, over, say, bond strengths [42–47]. The flow of the distribution is then analyzed either analytically or numerically in terms of a variable characterizing the strength of disorder, i.e., the variance of the individual distribution factor in $\mathcal{P}$ [47], by forcing a factorized parametrization at each stage. It is clear, however, that even if this (often unrealistic, since one can expect disorder in nearby areas to be correlated [48]) assumption holds initially, the renormalized distribution need not necessarily obey it, except in special cases. Generically, coarse graining the system introduces correlations in $\mathcal{P}$. Additionally, as in the clean case, higher-order and longer-range couplings are generated, in effect shifting the disorder distribution away from the hyperplane defined by only nearest-neighbor couplings. Both effects depicted in Fig. 8 increase the complexity of distribution and render the problem of computing and analyzing RG flows for disordered systems very challenging.

The real-space RG transformations applied to disordered systems are either similar to those used in the clean case, i.e., various decimation or Migdal-Kadanoff prescriptions, or based on the strong-disorder RG [49,50]. We focus on block transformations, which have the advantage of maintaining a regular topology in higher dimensions [51] (though the arguments below apply also when coarse-graining cells are chosen in a sequential, greedy fashion). The very same questions as in the translation-invariant setting need to be answered: Is there a more fundamental

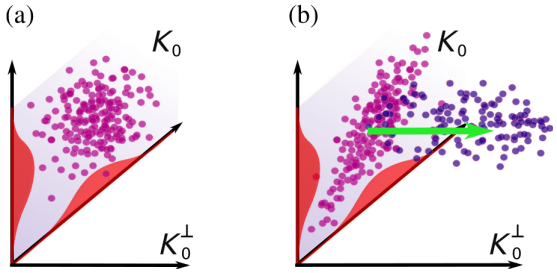

FIG. 8. Generic behavior of disorder probability under RG. Every point is a disorder realization (Hamiltonian). The shaded plane K_0 spanned by axes perpendicular to $K_0^\perp$ represents the subspace of two-body nearest-neighbor couplings; $K_0^\perp$ is the complement space of all other couplings. (a) Initially, a *factorized* distribution is usually assumed, schematically shown as a product of independent Gaussians. (b) After the RG step, the distribution can develop correlations depicted for simplicity in the K_0 plane, and additional couplings can be generated, resulting in probability mass shift out of K_0. The former effect is quantified by KL divergence or distance correlation, the latter by the shift of the center of mass of the distribution depicted with a green arrow (see also Fig. 9).

reason—beyond a simple algebraic coincidence—why certain RG transformations should work better in particular cases? Is there a *constructive* way to find the best such transformation within a certain class for a given physical system?

Our results suggest that the answer to both questions is affirmative: Beyond controlling the range of the interactions, RSMI maximization also suppresses generation of correlations in the renormalized distribution $\mathcal{P}$. As in Secs. III and IV, we first prove that factorizability properties of $\mathcal{P}$ under the full-information-capture assumption are stable to local changes in disorder, at least in (quasi)-1D systems and imply suppression of disorder correlations. Subsequently, we study the effect of arbitrary RG rules on the renormalized disorder distribution using a model system where the optimal solution is known, and the distribution can be computed (perturbatively) for arbitrary transformation. The following counterpart to Proposition 1 holds true (proof in Appendix B):

Proposition 2. Consider a disordered 1D system, with a factorizable (product) disorder distribution over, without loss of generality, nearest-neighbor couplings. The choice Λ^* of the optimal coarse graining of a block $\mathcal{X}_0$ satisfying $I(\mathcal{X}_0' : \mathcal{E}_0) = I(\mathcal{X}_0 : \mathcal{E}_0)$, and thus the factorization property of Proposition 1, are stable to local changes in disorder, provided those do not directly affect the block or the buffer, i.e., are fully confined to the environment.

Proposition 2 has two important consequences: (i) As seen from Eq. (B22) of the proof, in the explicit factorization of the conditional probability distribution of the coarse-grained d.o.f. in the left and right environments (cf. Proposition 1), changes to the disorder realization in one of the coarse-grained environments do not affect the

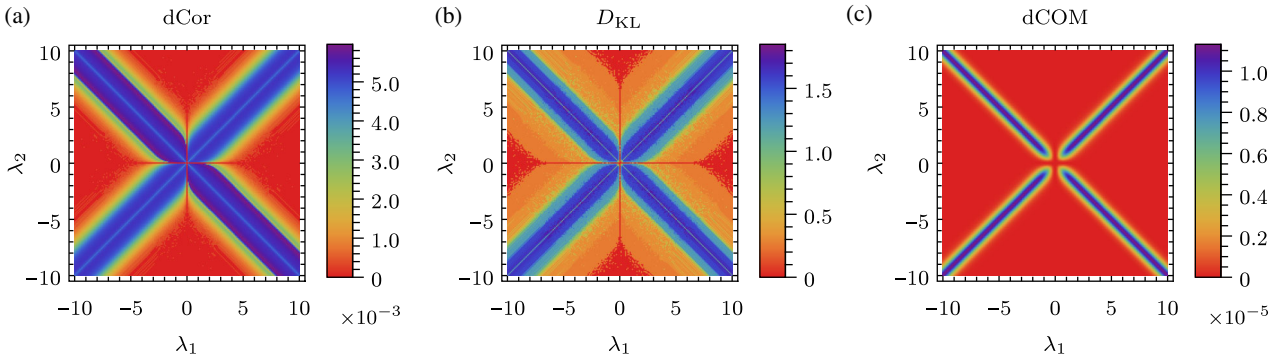

FIG. 9. Properties of the renormalized disorder distribution as a function of the coarse-graining rule for the case of the random dilute Ising chain. The RG rules are parametrized by λ_1, λ_2, as before. (a) The distance correlation (see the main text) dCor between renormalized distributions of two neighboring NN couplings. The couplings are uncorrelated if and only if dCor vanishes; compare with Fig. 4(a). (b) An alternative measure of correlations is the Kullback-Leibler divergence $\mathrm{D_{KL}}$ computed between the product of marginal distributions of neighboring couplings and their joint distribution (c) The $K_0^{\perp}$ center of mass of the disorder distribution dCOM (see the main text and Fig. 8). Note that all those quantities vanish when MI is maximized.

distribution of d.o.f. in the other. This invariance implies the following corollary:

Corollary 3. The probability distribution of d.o.f. in $\mathcal{E}_R(\mathcal{X}_0')$ being completely insensitive to the choice of disorder realization in $\mathcal{E}_L(\mathcal{X}_0')$, there cannot exist any correlations in the renormalized disorder distribution $\mathcal{P}$ between the regions $\mathcal{E}_L(\mathcal{X}_0')$ and $\mathcal{E}_R(\mathcal{X}_0')$ (i.e., no such correlations across $\mathcal{X}_0'$ are generated by the optimal course graining).

Since this corollary holds for every block, we conclude that with the above assumptions, disorder correlations remain suppressed under coarse graining. Note that this can be generalized to higher dimensions similar to Sec. III. (ii) Proposition 2 also implies that for the purpose of finding the optimal course graining of a block, which in general should depend on the disorder realization (as is also the case in strong-disorder RG), only the disorder configuration in the local neighborhood of the block can be considered. Though this is strictly true under the full-information-capture assumption, it provides motivation for constructing adaptive coarse grainings in more complicated systems, with the RG rule optimized for the *local* disorder realization (or, more practically, the equivalence classes thereof).

We turn to a solvable model system to empirically demonstrate decay of disorder correlations as a function of MI, especially when the stringent requirements of Proposition 2 are not satisfied. In the random Ising chain [43], the nearest-neighbor couplings are independent random variables $K_{i,i+1}$ distributed according to a probability $\mathcal{P}(K)$. For a generic $\mathcal{P}(K)$, the recursive RG Eq. (28) is intractable, but for the special case when it is Bernoulli distributed

$$\mathcal{P}(K) = p\delta(K - k_0) + (1 - p)\delta(K - k_1), \qquad (29)$$

the decimation transformation allows to solve Eq. (28) analytically, since the factorizability is then preserved exactly along the flow. The model exhibits much richer phenomenology than the clean case: For $k_0 = -k_1$, in addition to the usual (unstable) ferromagnetic and (stable) paramagnetic fixed points, the spin-glass fixed point is reached for any $0 < p < 1$ if starting exactly at $T = 0$. For $k_1 = 0$, i.e., in the random *dilute* Ising chain, additional Griffiths singularities appear in the limit $k_0 \to \infty$ and $h_0 \to 0$, where $h_0 \to 0$ are the on-site (uniform) magnetic fields [43,53]. This effect is associated with the existence of rare but arbitrarily large coupled clusters of spins [54].

We focus on the random dilute Ising chain but allow instead any arbitrary RG transformation (without loss of generality, for a block of two sites) parametrized by $\Lambda = (\lambda_1, \lambda_2)$, as in Sec. IV. For a finite periodic system, the renormalized couplings can be computed perturbatively using the cumulant expansion for any quenched disorder realization and any Λ. For illustration, we consider a system of 16 spins and *all* possible disorder realizations. For each realization, we compute the Hamiltonian after the RG step, for arbitrary Λ, by summing up to ninth order in the cumulants, obtaining the full renormalized disorder distribution $\mathcal{P}^1_\Lambda(\mathbf{K})$, where $\mathbf{K}$ is a vector of all possible couplings between the block spins.

To quantify the generated disorder correlations in $\mathcal{P}^1_\Lambda(\mathbf{K})$, we examine the joint probability distribution of two neighboring NN couplings $\mathcal{P}^1_\Lambda(K_{i,i+1}, K_{i+1,i+2})$ obtained by marginalization, as a function of Λ (it is chosen since those correlations develop the fastest). We use two statistical measures of dependence for this distribution: the distance correlation dCor [55] and the information-theoretic Kullback-Leibler divergence D_{KL} [56]. Both are sensitive also to nonlinear correlations and share the essential property that two random variables are statistically independent if and only if $\mathrm{dCor} = D_{\mathrm{KL}} = 0$, though distance correlation is generally better suited for continuous variables. In Fig. 9(a), we plot $\mathrm{dCor}(K_{i,i+1}, K_{i+1,i+2})$ as a

function of Λ, while in Fig. 9(b), $D_{\mathrm{KL}}[\mathcal{P}^1_\Lambda(K_{i,i+1},K_{i+1,i+2})||\mathcal{P}^1_\Lambda(K_{i,i+1})\mathcal{P}^1_\Lambda(K_{i+1,i+2})]$. Both measures coincide: The disorder distributions at neighboring bonds are the more independent, the more RSMI is retained by the coarse-graining rule, as seen by comparing with Fig. 4(a) (which is valid up to rescaling for every quenched disorder realization in the model). The couplings are statistically independent, rendering the renormalized disorder distribution factorizable, precisely where RSMI is maximized, i.e., for decimation. These results empirically establish the decay of correlations.

We also investigate another measure of complexity, i.e., how non-nearest-neighbor terms are generated as a function of Λ. Denote by K_0 the subspace of two-body NN couplings, and by $K_0^\perp$ the orthogonal space of all other couplings (see Fig. 8). For any (renormalized) disorder realization $\mathbf{K}$, let $\mathbf{K}^\perp$ be its restriction to $K^\perp$ obtained by truncation of all couplings in K_0. One measure of the shift of the renormalized disorder distribution away from K_0 is the $K_0^\perp$ center-of-mass dCOM:

$$\mathrm{dCOM} = \left\| \sum_{\mathbf{K}} \mathcal{P}(\mathbf{K})\mathbf{K}^\perp \right\|, \tag{30}$$

where $\|\cdot\|$ is the Euclidean norm. It is shown in Fig. 9(c) and exhibits the same qualitative behavior, as a function of Λ, as the correlation measures; i.e., it vanishes as a function of increasing RSMI.

We thus observe that empirically, RSMI maximization suppresses the generation of both spurious correlations and of higher-order and long-range couplings in the renormalized disorder distribution, which is also supported by Proposition 2 (under the appropriate assumptions).

VII. CONCLUSIONS AND OUTLOOK

We investigate information-theoretic properties of real-space RG procedures and particularly of one based on variational maximization of RSMI [28], both for clean and disordered systems. We demonstrate suppression of longer-range interactions in the renormalized Hamiltonian as a function of the RSMI retained: formally, proving this statement under explicitly stated assumptions, and empirically, by considering arbitrary coarse grainings in the solvable example of the Ising chain. For the case of disordered systems, again using formal proofs and the example of the dilute random Ising chain, we show that in addition to longer-range or higher-order terms, also correlations in the renormalized disorder distribution are suppressed. We also examine the effect of constraints on the type of coarse-grained variables on the RG procedure.

Our results provide a formal underpinning for the *physical* intuition behind the RSMI maximization: The effective long-wavelength description of the system is simple in terms of d.o.f. which carry the most information about its large-scale behavior. While the notion of simplicity may be ambiguous—despite the clear practical consequences of its absence—as there exist multiple measures of Hamiltonian complexity, the long-range information and its retention can be defined rigorously, similar to the information bottleneck approach of compression theory [32]. The different measures of complexity we compute for both clean (range, amount of n-body interactions) and disordered systems (correlations in the renormalized disorder distribution measured using KL divergence and distance correlation) are *all* suppressed as more RSMI is preserved by the coarse graining. This observation strongly indicates that the model-independent RSMI coarse graining, optimal by construction from the point of view of compression theory, is also optimal physically, resulting in operationally desirable properties (a tractable Hamiltonian). We thus establish direct and quantifiable connections between the information-theoretic properties of RG transformation and the actual physical properties of the renormalized Hamiltonian and the disorder distribution.

Beyond conceptual significance, the results can be useful practically, inspiring new numerical approaches to RG for disordered or complex systems, as we briefly discuss in Sec. VI. This is especially interesting given the progress in machine learning, numerical techniques for MI estimation [57,58], and the inverse problem [59].

A number of distinct directions are possible for further research. On the formal part of the spectrum, a mathematically rigorous investigation of the probability measure defined by the RSMI coarse graining in the spirit of Refs. [33,34,60] is desirable. Conceptually, an interesting question is whether the type and number of coarse-grained variables can also be variationally optimized, as opposed to being chosen at the outset, as is usually the case. This will have the interpretation of "discovering" whether the best variables to describe a system, originally given in terms of, say, Ising spins, are the same, or rather some emergent d.o.f. are preferable (see also Refs. [61–63]). More practically, the results invite the application of the RSMI method to the study of disordered systems, both using synthetic as well as experimental data. Finally, a quantum version of the procedure is an open question. The information bottleneck has recently been extended to the case of compression of quantum data [64]: In this setting, the conditional probability of classical systems is replaced by a quantum channel. It would be interesting to explore how the physics of the system manifests itself in properties of these optimal channels and to compare it with energy-based approaches [65–68] and the recently introduced graph-independent local truncation version of the tensor renormalization group method [69].

ACKNOWLEDGMENTS

We thank Professor Gianni Blatter for his insightful comments. S. D. H. and M. K.-J. gratefully acknowledge the support of Swiss National Science Foundationand the

European Research Council under the Grant Agreement No. 771503 (TopMechMat).

APPENDIX A: MUTUAL INFORMATION

The mutual information Eq. (6) can equivalently be defined by

$$I_\Lambda(\mathcal{H}:\mathcal{E}) = H(\mathcal{H}) - H(\mathcal{H}|\mathcal{E}), \tag{A1}$$

where

$$H(\mathcal{H}) = -\sum_{\mathcal{H}} P_\Lambda(\mathcal{H}) \log[P_\Lambda(\mathcal{H})], \tag{A2}$$

$$H(\mathcal{H}|\mathcal{E}) = -\sum_{\mathcal{H},\mathcal{E}} P_\Lambda(\mathcal{H},\mathcal{E}) \log\left(\frac{P_\Lambda(\mathcal{H},\mathcal{E})}{P(\mathcal{E})}\right) \tag{A3}$$

are the Shannon entropy and conditional entropy, respectively. It is a symmetric quantity. Positivity of mutual information and of the conditional entropy, together with the bound on entropy, immediately imply the following inequalities:

$$0 \le I_\Lambda(\mathcal{H}:\mathcal{E}) \le H(\mathcal{H}), \tag{A4}$$

where $H(\mathcal{H})$ is the entropy of $\mathcal{H}$. The mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$ is also bounded by the mutual information of the visibles and the environment:

$$I_\Lambda(\mathcal{H}:\mathcal{E}) \le I(\mathcal{V}:\mathcal{E}), \tag{A5}$$

which is obvious, since the hidden d.o.f. only couple to the environment via the visibles.

Throughout the text, we also use the notion of conditional mutual information, which, for any random variable $\mathcal{E}$, $\mathcal{H}$, $\mathcal{V}$, generically can be defined via the so-called chain rule:

$$I(\mathcal{E}:\mathcal{H},\mathcal{V}) = I(\mathcal{E}:\mathcal{H}) + I(\mathcal{E}:\mathcal{V}|\mathcal{H}). \tag{A6}$$

APPENDIX B: RSMI DOES NOT INCREASE THE RANGE OF INTERACTIONS AND MAINTAINS FACTORIZABILITY OF DISTRIBUTIONS

Here we give the details of the argument in Sec. III. We work directly in D dimensions, and we spell out explicitly the additional (reasonable) assumptions required compared to the 1D case.

Consider a generic finite-ranged Hamiltonian with d.o.f. $\mathcal{X}$ in D dimensions. For concreteness, let us assume a hypercubic lattice. We partition $\mathcal{X}$ into hypercubic coarse-graining blocks $\mathcal{V}_j$ large enough, so that only nearest-neighbor blocks interact. For the purpose of this argument, we arrange the blocks into parallel $(D-1)$-dimensional hyperplanes index by l so that $\mathcal{X} = \cup_\ell \mathcal{X}_\ell$ with $\mathcal{X}_\ell = \cup_{j\in J_\ell} \mathcal{V}_j$. Thus, in terms of the hyperplanes we end up with a quasi-one-dimensional structure. Let us choose an arbitrary hyperplane $\mathcal{X}_0$, denote its immediate neighbors $\mathcal{X}_{\pm 1}$ as the buffer $\mathcal{B}$, and the union of the remaining hyperplanes $\mathcal{X}_{\ell<-1}$ and $\mathcal{X}_{\ell>1}$ as the environment $\mathcal{E}_0(\mathcal{X}_0)$, or, in more detail, as the left and right environment $\mathcal{E}_{L/R}(\mathcal{X}_0)$, respectively.

Assume now that the coarse-grained variables $\mathcal{H}_j$ for the blocks $\mathcal{V}_j$ in $\mathcal{X}_0$ are constructed in such a way that $I(\mathcal{X}_0{}':\mathcal{E}_0) = I(\mathcal{X}_0:\mathcal{E}_0)$, where $\mathcal{X}_0{}' = \cup_{j\in J_0} \mathcal{H}_j$. This is the full-information-capture condition for the hyperplane generalizing the condition for the single block in 1D (note though, that we still optimize variables $\mathcal{H}_j$ for each block and not some new collective hidden variables for the entire hyperplanes). Strictly speaking, this extension requires an additional assumption (compared to 1D) that it is equivalent to assuming $I(\mathcal{H}_j:\mathcal{E}_0) = I(\mathcal{V}_j:\mathcal{E}_0)$ separately for each individual block in the hyperplane $\mathcal{X}_0$. This seems reasonable for a short-ranged Hamiltonian, at least in the isotropic case. Under those assumptions, we show that the probability measure on the coarse-grained variables $P(\mathcal{X}')$ obeys a D-dimensional analog of the factorization property of Proposition 1 in Sec. III:

$$P(\mathcal{X}'_{j\le -2}, \mathcal{X}'_{j\ge 2}|\mathcal{X}'_0) = P(\mathcal{X}'_{j\le -2}|\mathcal{X}'_0) P(\mathcal{X}'_{j\ge 2}|\mathcal{X}'_0). \tag{B1}$$

To prove this, we begin with a crucial separability lemma, which is a technical condition enabling Proposition 1:

Lemma. Let $\mathcal{E}_L$, $\mathcal{E}_R$ be the left and right environments, respectively, of $\mathcal{X}_0$ and let $I(\mathcal{X}_0{}':\mathcal{E}_0) = I(\mathcal{X}_0:\mathcal{E}_0)$. Then, the following factorization property with respect to the coarse-grained variable $\mathcal{X}'_0$ holds: $P(\mathcal{E}_L,\mathcal{E}_R|\mathcal{X}'_0) = P(\mathcal{E}_L|\mathcal{X}'_0)P(\mathcal{E}_R|\mathcal{X}'_0)$.

Proof. To show that, first note that from the full-information-capture assumption, it follows that

$$\begin{aligned} I(\mathcal{E}_0:\mathcal{X}_0|\mathcal{X}'_0) &= I(\mathcal{E}_0:\mathcal{X}_0,\mathcal{X}'_0) - I(\mathcal{E}_0:\mathcal{X}'_0) \\ &= I(\mathcal{E}_0:\mathcal{X}_0) - I(\mathcal{E}_0:\mathcal{X}'_0) = 0, \end{aligned} \tag{B2}$$

where the first equality is the chain rule for mutual information, and the second is due to the fact that the coarse-grained variables $\mathcal{X}'_0$ are a function of $\mathcal{X}_0$ only. Vanishing of this mutual information is equivalent to the (conditional) probability distribution factorizing, and therefore,

$$P(\mathcal{E}_0,\mathcal{X}_0|\mathcal{X}'_0) = P(\mathcal{E}_0|\mathcal{X}'_0)P(\mathcal{X}_0|\mathcal{X}'_0). \tag{B3}$$

Furthermore, the locality of the interactions assumption implies that $I(\mathcal{E}_L:\mathcal{E}_R|\mathcal{X}_0) = 0$, and thus,

$$\begin{aligned} P(\mathcal{E}_L,\mathcal{E}_R,\mathcal{X}_0|\mathcal{X}'_0) &= P(\mathcal{E}_L,\mathcal{E}_R|\mathcal{X}_0,\mathcal{X}'_0)P(\mathcal{X}_0|\mathcal{X}'_0) \\ &= P(\mathcal{E}_L,\mathcal{E}_R|\mathcal{X}_0)P(\mathcal{X}_0|\mathcal{X}'_0) \\ &= P(\mathcal{E}_L|\mathcal{X}_0)P(\mathcal{E}_R|\mathcal{X}_0)P(\mathcal{X}_0|\mathcal{X}'_0). \end{aligned} \tag{B4}$$

Comparing Eqs. (B3) and (B4), we find that

$$P(\mathcal{E}_0|\mathcal{X}_0')P(\mathcal{X}_0|\mathcal{X}_0') = P(\mathcal{E}_L|\mathcal{X}_0)P(\mathcal{E}_R|\mathcal{X}_0)P(\mathcal{X}_0|\mathcal{X}_0'). \tag{B5}$$

For a given $\mathcal{X}_0'$, let us denote the set of $\mathcal{X}_0$ such that $P(\mathcal{X}_0|\mathcal{X}_0') \neq 0$ by $\{\mathcal{X}_0(\mathcal{X}_0')\}$. For all such "compatible" $\mathcal{X}_0 \in \{\mathcal{X}_0(\mathcal{H}_0)\}$, we can divide by $P(\mathcal{X}_0|\mathcal{X}_0')$ and obtain

$$P(\mathcal{E}_0|\mathcal{X}_0') = P(\mathcal{E}_L|\mathcal{X}_0)P(\mathcal{E}_R|\mathcal{X}_0). \tag{B6}$$

Crucially, the left-hand side does not depend on $\mathcal{X}_0$, and so as long as $\mathcal{X}_0 \in \{\mathcal{X}_0(\mathcal{X}_0')\}$ the equality holds, and the conditional probability factorizes independently of a particular $\mathcal{X}_0$. In fact, the factorization holds generally, and the case $P(\mathcal{X}_0|\mathcal{X}_0') = 0$ is not a problem:

$$\begin{aligned} P(\mathcal{E}_0|\mathcal{X}_0') &= \sum_{\mathcal{X}_0} P(\mathcal{E}_L, \mathcal{E}_R, \mathcal{X}_0|\mathcal{X}_0') \\ &= \sum_{\mathcal{X}_0} P(\mathcal{E}_L|\mathcal{X}_0)P(\mathcal{E}_R|\mathcal{X}_0)P(\mathcal{X}_0|\mathcal{X}_0') \\ &= \sum_{\mathcal{X}_0 \in \{\mathcal{X}_0(\mathcal{X}_0')\}} P(\mathcal{E}_L|\mathcal{X}_0)P(\mathcal{E}_R|\mathcal{X}_0)P(\mathcal{X}_0|\mathcal{X}_0') \\ &= P[\mathcal{E}_L|\mathcal{X}_0(\mathcal{X}_0')]P[\mathcal{E}_R|\mathcal{X}_0(\mathcal{X}_0')], \end{aligned} \tag{B7}$$

where we use Eq. (B4) in the second equality, explicitly remove vanishing [by virtue of $P(\mathcal{X}_0|\mathcal{X}_0') = 0$] terms in the sum in the third, and use Eq. (B6) to take the $\mathcal{X}_0$-independent product from under the restricted summation in the third. We thus construct an explicit factorization of $P(\mathcal{E}_L, \mathcal{E}_R|\mathcal{X}_0')$ in Eq. (B7), which implies

$$I(\mathcal{E}_L : \mathcal{E}_R|\mathcal{X}_0') = 0, \tag{B8}$$

and hence, we can simply write

$$P(\mathcal{E}_L, \mathcal{E}_R|\mathcal{X}_0') = P(\mathcal{E}_L|\mathcal{X}_0')P(\mathcal{E}_R|\mathcal{X}_0'). \tag{B9}$$

■

The lemma has a very nice physical interpretation, which provides useful intuition of more general validity. It states that with a short-range Hamiltonian, an area of finite width has to mediate *all* correlations between its neighborhoods, and if the information that area has about them is accurately retained in a new variable, no correlations can exist between the neighborhoods beyond those mediated by the new variable. Note we do not rely on translation invariance at all. This will be useful in deriving a corresponding statement for disordered systems, which explains the results in Sec. VI. More immediately, it is the key element in showing Eq. (B1), giving the D-dimensional version of Proposition 1:

Proposition 1. Let $I(\mathcal{X}_0' : \mathcal{E}_0) = I(\mathcal{X}_0 : \mathcal{E}_0)$. Then, the probability measure on the coarse-grained variables $P(\mathcal{X}')$ obeys the factorization property

$$P(\mathcal{X}'_{j\leq -2}, \mathcal{X}'_{j\geq 2}|\mathcal{X}_0') = P(\mathcal{X}'_{j\leq -2}|\mathcal{X}_0')P(\mathcal{X}'_{j\geq 2}|\mathcal{X}_0'), \tag{B10}$$

where in the conditional probabilities, the buffer (i.e., the neighbors $\mathcal{X}'_{\pm 1}$ of $\mathcal{X}_0'$) is integrated out. In other words, for fixed $\mathcal{X}_0'$, the probabilities of its left and right environments $\mathcal{E}_{L/R}(\mathcal{X}_0')$ are independent of each other.

Proof. Consider the coarse-grained probability measure defined by Eqs. (3) and (5):

$$P(\mathcal{X}') = \sum_{\mathcal{X}} P(\mathcal{X}) \prod_j P(\mathcal{H}_j|\mathcal{V}_j). \tag{B11}$$

Denoting the product of the block conditional probability distributions in the hyperplanes by $\prod_\ell P(\mathcal{X}'_\ell|\mathcal{X}_\ell)$ and integrating out $\mathcal{X}'_{\pm 1}$, we have

$$P(\mathcal{X}'_{j\leq -2}, \mathcal{X}_0', \mathcal{X}'_{j\geq 2}) = \sum_{\mathcal{X}_{|\ell|\neq 1}} P(\{\mathcal{X}_\ell\}_{|\ell|\neq 1}) \prod_{|\ell|\neq 1} P(\mathcal{X}'_\ell|\mathcal{X}_\ell). \tag{B12}$$

Using the definition of conditional probability and the fact that $\mathcal{X}_0'$ directly depends only on $\mathcal{X}_0$, we have

$$\begin{aligned} P(\{\mathcal{X}_\ell\}_{|\ell|\neq 1})P(\mathcal{X}_0'|\mathcal{X}_0) &\equiv P(\mathcal{E}_0, \mathcal{X}_0, \mathcal{X}_0') \\ &= P(\mathcal{E}_0, \mathcal{X}_0|\mathcal{X}_0')P(\mathcal{X}_0'), \end{aligned} \tag{B13}$$

which allows us to write

$$P(\mathcal{X}'_{j\leq -2}, \mathcal{X}_0', \mathcal{X}'_{j\geq 2})$$

$$= \sum_{\mathcal{X}_{|\ell|\neq 1}} P(\mathcal{E}_0, \mathcal{X}_0|\mathcal{X}_0')P(\mathcal{X}_0') \prod_{|\ell|\neq 0,1} P(\mathcal{X}'_\ell|\mathcal{X}_\ell) \tag{B14}$$

$$= \sum_{\mathcal{X}_{|\ell|\neq 1}} P(\mathcal{E}_L, \mathcal{E}_R|\mathcal{X}_0')P(\mathcal{X}_0|\mathcal{X}_0')P(\mathcal{X}_0') \prod_{|\ell|\neq 0,1} P(\mathcal{X}'_\ell|\mathcal{X}_\ell) \tag{B15}$$

$$= \sum_{\mathcal{X}_{|\ell|\neq 1}} P(\mathcal{E}_L|\mathcal{X}_0')P(\mathcal{E}_R|\mathcal{X}_0')P(\mathcal{X}_0|\mathcal{X}_0')P(\mathcal{X}_0') \times \prod_{|\ell|\neq 0,1} P(\mathcal{X}'_\ell|\mathcal{X}_\ell) \tag{B16}$$

$$= \sum_{\mathcal{X}_{|\ell|\neq 0,1}} P(\mathcal{E}_L|\mathcal{X}_0')P(\mathcal{E}_R|\mathcal{X}_0')P(\mathcal{X}_0') \prod_{|\ell|\neq 0,1} P(\mathcal{X}'_\ell|\mathcal{X}_\ell) \tag{B17}$$

$$= P(\mathcal{X}_0') \left(\sum_{\mathcal{X}_{\ell\leq -2}} P(\mathcal{E}_L|\mathcal{X}_0') \prod_{\ell\leq -2} P(\mathcal{X}'_\ell|\mathcal{X}_\ell) \right) \times \left(\sum_{\mathcal{X}_{\ell\geq 2}} P(\mathcal{E}_R|\mathcal{X}_0') \prod_{\ell\geq 2} P(\mathcal{X}'_\ell|\mathcal{X}_\ell) \right) \tag{B18}$$

$$= P(\mathcal{X}_0')P(\mathcal{X}'_{j\leq -2}|\mathcal{X}_0')P(\mathcal{X}'_{j\geq 2}|\mathcal{X}_0'), \tag{B19}$$

where to obtain Eq. (B15), we condition on $\mathcal{X}_0$ and use the full-information-capture assumption to write $P(\mathcal{E}_L,\mathcal{E}_R|\mathcal{X}_0,\mathcal{X}_0')=P(\mathcal{E}_L,\mathcal{E}_R|\mathcal{X}_0')$; to obtain Eq. (B16), we use the factorization Eq. (B9) proved in the lemma; to obtain Eq. (B17) we perform the summation over $\mathcal{X}_0$; to obtain Eq. (B18), we rearrange the sums taking expressions independent of summation variables out of them; in the last line, we use Bayes's law. Dividing both sides by $P(\mathcal{X}_0')$, we obtain Eq. (B10). ∎

Proposition 1 shows that for a fixed $\mathcal{X}_0'$, the probability $P(\mathcal{X}_{j\le -2}',\mathcal{X}_0',\mathcal{X}_{j\ge 2}')$ factorizes into a product over left and right environments. As we describe in the main text, together with the arbitrariness of the choice of the hyperplane, this implies (barring a pathological fine-tuned scenario in which integration over $\mathcal{X}_{\pm 1}$ *exactly* cancels all preexisting NNN couplings) that the effective Hamiltonian in terms of new variables is still nearest neighbor (in all directions).

Furthermore, under the same assumptions of the finite-ranged Hamiltonian, we can also derive an important result about the properties of the renormalized disorder distribution. Assume, without loss of generality, that the blocks are chosen sufficiently large to render interactions nearest neighbor with respect to the blocks. Then:

Proposition 2. Consider a disordered 1D system with a factorizable (product) disorder distribution over, without loss of generality, nearest-neighbor couplings. The choice Λ^* of the optimal coarse graining of a block $\mathcal{X}_0$ satisfying $I(\mathcal{X}_0{}':\mathcal{E}_0)=I(\mathcal{X}_0:\mathcal{E}_0)$, and thus, the factorization property of Proposition 1, are stable to local changes in disorder, provided those do not affect directly the block or the buffer; i.e., they are fully confined to the environment.

Proof. For a fixed quenched disorder realization, denote the probability distribution of the d.o.f. under this Hamiltonian by $P(\mathcal{X})$. Let $P_{\Lambda^*}(\mathcal{X}_0'|\mathcal{X}_0)$ be the optimal coarse graining for the block $\mathcal{X}_0$ determined by Λ^* saturating mutual information and consequently ensuring the factorization property of the lemma is obeyed. Consider now a localized change to the disorder realization, affecting only terms acting entirely within an area $\mathcal{X}_D\subset\mathcal{E}_L$, resulting in a modified probability distribution $\tilde{P}(\mathcal{X})$. One can then show that the factorization property still holds with the very same choice of $P_{\Lambda^*}(\mathcal{X}_0'|\mathcal{X}_0)$.

To this end, denote by $\tilde{K}(\mathcal{X}_D)$ the local terms in the reduced Hamiltonian affected by the disorder change and by $K(\mathcal{X}_D)$ the original ones (the change to the Hamiltonian K is also localized since it is a NN in the blocks and the change to the disorder is confined to $\mathcal{X}_D$). Then,

$$\begin{aligned}
&\tilde{P}(\mathcal{E}_L,\mathcal{E}_R,\mathcal{X}_0)P_{\Lambda^*}(\mathcal{X}_0'|\mathcal{X}_0)\\
&=\frac{e^{\tilde{K}(\mathcal{E}_L,\mathcal{E}_R,\mathcal{X}_0)}}{\tilde{Z}}P_{\Lambda^*}(\mathcal{X}_0'|\mathcal{X}_0)\\
&=\frac{e^{K(\mathcal{E}_L,\mathcal{E}_R,\mathcal{X}_0)}}{Z}\frac{Z}{\tilde{Z}}\frac{e^{\tilde{K}(\mathcal{X}_D)}}{e^{K(\mathcal{X}_D)}}P_{\Lambda^*}(\mathcal{X}_0'|\mathcal{X}_0)\\
&=\frac{Z}{\tilde{Z}}\frac{e^{\tilde{K}(\mathcal{X}_D)}}{e^{K(\mathcal{X}_D)}}P(\mathcal{E}_L,\mathcal{E}_R,\mathcal{X}_0)P_{\Lambda^*}(\mathcal{X}_0'|\mathcal{X}_0)\\
&=\frac{Z}{\tilde{Z}}\frac{e^{\tilde{K}(\mathcal{X}_D)}}{e^{K(\mathcal{X}_D)}}P(\mathcal{E}_L,\mathcal{E}_R,\mathcal{X}_0,\mathcal{X}_0')\\
&=\frac{Z}{\tilde{Z}}\frac{e^{\tilde{K}(\mathcal{X}_D)}}{e^{K(\mathcal{X}_D)}}P(\mathcal{E}_L,\mathcal{E}_R,\mathcal{X}_0|\mathcal{X}_0')P(\mathcal{X}_0')\\
&=\frac{Z}{\tilde{Z}}\frac{e^{\tilde{K}(\mathcal{X}_D)}}{e^{K(\mathcal{X}_D)}}P(\mathcal{E}_L|\mathcal{X}_0')P(\mathcal{E}_R|\mathcal{X}_0')P(\mathcal{X}_0|\mathcal{X}_0')P(\mathcal{X}_0')\\
&=\tilde{P}(\mathcal{E}_L|\mathcal{X}_0')P(\mathcal{E}_R|\mathcal{X}_0')P(\mathcal{X}_0|\mathcal{X}_0')P(\mathcal{X}_0'),
\end{aligned}\tag{B20}$$

where Z and $\tilde{Z}$ are the original and modified partition functions. In the penultimate line, we use the lemma for the initial distribution $P(\mathcal{E}_L,\mathcal{E}_R,\mathcal{X}_0)$, and in the last, we absorb all additional factors, which are local, into the definition of $\tilde{P}(\mathcal{E}_L|\mathcal{X}_0')$. Dividing both sides by $P(\mathcal{X}_0')$ and marginalizing over $\mathcal{X}_0$, we arrive at

$$\tilde{P}(\mathcal{E}_L,\mathcal{E}_R|X_0')=\tilde{P}(\mathcal{E}_L|\mathcal{X}_0')P(\mathcal{E}_R|\mathcal{X}_0'),\tag{B21}$$

where the right factor is as for the original distribution.

Since the change to the disorder (other than being confined to $\mathcal{E}_L$) is completely arbitrary, Eq. (B21) shows that any such localized changes do not affect the choice of optimal coarse graining of the block. Thus, they do not break the factorization property and, in particular, they do not affect the other environment: Note that in Eq. (B21), we still have the *original* $P(\mathcal{E}_R|\mathcal{X}_0')$. Consequently, Proposition 1 immediately holds for both the original and modified disorder realization with the same coarse graining $P_{\Lambda^*}(\mathcal{X}_0'|\mathcal{X}_0)$ and the same probability distribution of the renormalized right environment $\mathcal{E}_R(\mathcal{X}_0')$:

$$\tilde{P}(\mathcal{X}_{j\le -2}',\mathcal{X}_{j\ge 2}'|\mathcal{X}_0')=\tilde{P}(\mathcal{X}_{j\le -2}'|\mathcal{X}_0')P(\mathcal{X}_{j\ge 2}'|\mathcal{X}_0').\tag{B22}$$

Hence, the coarse graining is stable. ∎

Proposition 2 has an important consequence for the renormalized disorder distribution $\mathcal{P}$: With the probability distribution of d.o.f. in $\mathcal{E}_R(\mathcal{X}_0')$ being completely insensitive to the choice of disorder realization in $\mathcal{E}_L(\mathcal{X}_0')$, we conclude that there cannot exist any correlations in the disorder distribution between in the regions $\mathcal{E}_L(\mathcal{X}_0')$ and

$\mathcal{E}_R(\mathcal{X}'_0)$ (i.e., no such correlations across $\mathcal{X}'_0$ are generated by the optimal coarse graining).

APPENDIX C: THE EFFECTIVE HAMILTONIAN

1. The cumulant expansion

Consider a generic Hamiltonian $\mathcal{K}[\mathcal{X}]$. We split it into two parts [35]:

$$\mathcal{K}[\mathcal{X}] = \mathcal{K}_0[\mathcal{X}] + \mathcal{K}_1[\mathcal{X}], \tag{C1}$$

where $\mathcal{K}_0$ contains *intrablock* terms, i.e., those which couple only spins within a single block, and $\mathcal{K}_1$ contains *interblock* terms, i.e., those that couple spins from different blocks. Such a decomposition simplifies the calculations significantly. For translationally invariant systems, the intrablock terms are all of the same form:

$$\mathcal{K}_0[\mathcal{X}] = \sum_{j=1}^{n} \mathcal{K}_b[\mathcal{V}_j]. \tag{C2}$$

Using the decomposition Eqs. (C1) and (C2), the definition of the renormalized Hamiltonian in Eq. (4) can be rewritten as an intrablock average of the interblock part of the Hamiltonian:

$$\begin{aligned} e^{\mathcal{K}'[\mathcal{X}']} &= Z_0 \sum_{\mathcal{X}} e^{\mathcal{K}_1[\mathcal{X}]} \prod_{j=1}^{n} \underbrace{\frac{e^{\mathcal{K}_b[\mathcal{V}_j]}}{Z_b} P_\Lambda(\mathcal{H}_j|\mathcal{V}_j)}_{=: P_{\Lambda,b}(\mathcal{H}_j,\mathcal{V}_j)} \\ &= Z_0 \sum_{\mathcal{X}} e^{\mathcal{K}_1[\mathcal{X}]} \prod_{j=1}^{n} P_{\Lambda,b}(\mathcal{V}_j|\mathcal{H}_j) P_{\Lambda,b}(\mathcal{H}_j) \\ &= Z_0 \underbrace{\prod_{j=1}^{n} P_{\Lambda,b}(\mathcal{H}_j)}_{=: P_{\Lambda,0}(\mathcal{X}')} \sum_{\mathcal{X}} e^{\mathcal{K}_1[\mathcal{X}]} \underbrace{\prod_{j=1}^{n} P_{\Lambda,b}(\mathcal{V}_j|\mathcal{H}_j)}_{=: P_{\Lambda,0}(\mathcal{X}|\mathcal{X}')} \\ &= Z_0 P_{\Lambda,0}(\mathcal{X}') \langle e^{\mathcal{K}_1[\mathcal{X}]} \rangle_{\Lambda,0}[\mathcal{X}'], \end{aligned} \tag{C3}$$

where the average $\langle\cdot\rangle_{\Lambda,0}$ is over $P_{\Lambda,0}(\mathcal{X}|\mathcal{X}')$ as a probability distribution in $\mathcal{X}$ and thus introduces a dependence on the new spin variables $\mathcal{X}'$. We indicate this dependence by square brackets [.] after the average.

Equation (C3) lends itself to a cumulant expansion:

$$\langle e^{\mathcal{K}_1[\mathcal{X}]} \rangle_{\Lambda,0}[\mathcal{X}'] = e^{\sum_{k=0}^{\infty} \frac{1}{k!} C_k[\mathcal{X}']} \tag{C4}$$

with the standard expressions for the cumulants in terms of moments, the first few of which are given by

$$C_1 = \langle \mathcal{K}_1 \rangle_{\Lambda,0}, \tag{C5a}$$

$$C_2 = \langle \mathcal{K}_1^2 \rangle_{\Lambda,0} - \langle \mathcal{K}_1 \rangle_{\Lambda,0}^2, \tag{C5b}$$

$$C_3 = \langle \mathcal{K}_1^3 \rangle_{\Lambda,0} - 3\langle \mathcal{K}_1^2 \rangle_{\Lambda,0} \langle \mathcal{K}_1 \rangle_{\Lambda,0} + 2\langle \mathcal{K}_1 \rangle_{\Lambda,0}^3, \tag{C5c}$$

where for brevity we do not indicate the dependence on $\mathcal{X}'$. The powers of $\mathcal{K}_1$ inside the averages induce couplings between multiple blocks and naturally lead to new coupling terms in the effective Hamiltonian.

The cumulant expansion Eq. (C4) allows us to determine the new Hamiltonian by taking the logarithm of Eq. (C3):

$$\mathcal{K}'[\mathcal{X}'] = \log[Z_0 P_{\Lambda,0}(\mathcal{X}')] + \sum_{k=0}^{\infty} \frac{1}{k!} C_k[\mathcal{X}']. \tag{C6}$$

The renormalized coupling constants are not apparent in Eq. (C6). In order to identify them, we introduce the following canonical form of the Hamiltonian:

$$\mathcal{K}'[\mathcal{X}'] = K'_0 + \sum_{\{\alpha_\ell\}_{\ell=1}^n} K'_{\alpha_1,\alpha_2,\ldots,\alpha_n} \left(\sum_{j=1}^{n} \prod_{\ell=1}^{n} (x'_{j+\ell})^{\alpha_\ell} \right) \tag{C7}$$

with $\alpha_1 = 1$ and $\alpha_\ell \in \{0,1\}$ for all $\ell > 1$. Here, the addition of the indices is to be understood modulo n (i.e., with periodic boundary conditions). Note that arbitrary orders k of the cumulant expansion C_k contribute to each coupling constant $K'_{\alpha_1,\alpha_2,\ldots,\alpha_n}$.

2. Factorization of quenched averages

Factorization of the conditional probability distribution results in the factorization of expectations $\langle \mathcal{O}[\mathcal{X}] \rangle_{\Lambda,0}$ for any operator, which is a product of operators o_j acting on separate blocks, i.e., $\mathcal{O}[\mathcal{X}] = \prod_{j=1}^{n} o_j[\mathcal{V}_j]$,

$$\begin{aligned} \langle \mathcal{O}[\mathcal{X}] \rangle_{\Lambda,0}[\mathcal{X}'] &= \prod_{j=1}^{n} \sum_{\mathcal{V}_j} o_j[\mathcal{V}_j] P_{\Lambda,b}(\mathcal{V}_j|\mathcal{H}_j) \\ &= \prod_{j=1}^{n} \langle o_j[\mathcal{V}] \rangle_{\Lambda,b}[\mathcal{H}_j], \end{aligned} \tag{C8}$$

where the probability over which we average is

$$P_{\Lambda,b}(\mathcal{V}|\mathcal{H}) = \frac{e^{\mathcal{K}_b[\mathcal{V}]}}{Z_b P_{\Lambda,b}(\mathcal{H})} P_\Lambda(\mathcal{H}|\mathcal{V}). \tag{C9}$$

In particular, the factorization holds for the operators $\mathcal{K}_1^k$, which appear in the expressions for the cumulants.

3. Parametrization of the RG rule using a RBM *Ansatz*

The conditional probability distribution $P_\Lambda(\mathcal{H}|\mathcal{V})$ is parametrized using a RBM *Ansatz* [70–72]. The RBMs belong to a family of energy-based models, whose main purpose is to efficiently approximate probability distributions, and, more generally, they are an example of a growing class of machine-learning techniques recently employed in a statistical physics or condensed matter setting [16,28,73–81].

In the RBM *Ansatz,* the joint probability of the visible and hidden d.o.f. is approximated by a Boltzmann distribution

$$P(\mathcal{V},\mathcal{H}) = \frac{1}{Z} e^{-E_\Lambda(\mathcal{V},\mathcal{H})} \tag{C10}$$

with a quadratic energy function

$$E_\Lambda(\mathcal{V},\mathcal{H}) = -\sum_{i,j} \lambda_i^j v_i h_j - \sum_i \alpha_i v_i - \sum_j \beta_j h_j, \tag{C11}$$

where $v_i \in \mathcal{V}$, $h_j \in \mathcal{H}$, and Λ collectively denote the set of parameters $\{\lambda_i^j\}_{i,j}$, $\{\alpha_i\}_i$, and $\{\beta_j\}_j$, which are to be variationally optimized so that the $P_\Lambda(\mathcal{V},\mathcal{H})$ they define is as close as possible to the target distribution $P(\mathcal{V},\mathcal{H})$. Note that the energy function couples only the visible to the hidden d.o.f. and includes no couplings *within* the visible or the hidden sets. This peculiarity (which the word "restricted" in the RBM refers to) is crucial to the existence of fast algorithms [82] for training and sampling from the trained distribution $P_\Lambda(\mathcal{V},\mathcal{H})$.

The conditional probability is then given by

$$P_\Lambda(\mathcal{H}|\mathcal{V}) = \frac{e^{-E_\Lambda(\mathcal{V},\mathcal{H})}}{\sum_{\mathcal{H}} e^{-E_\Lambda(\mathcal{V},\mathcal{H})}}. \tag{C12}$$

It is easy to see that the parameters $\{\alpha_i\}_i$ drop out in $P_\Lambda(\mathcal{H}|\mathcal{V})$. Additionally, because of the Ising $\mathbb{Z}_2$ symmetry, the bias (magnetic field) term for h_j is not allowed: $\beta_i = 0$ for all i. Because of the absence of interactions between hiddens, the expression factorizes and the summation over $\mathcal{H}$ is trivial. In the case of a 1D system and a single hidden spin $\mathcal{H} = \{h\}$, the conditional probability is then given explicitly by

$$P_\Lambda(\mathcal{H}|\mathcal{V}) = \frac{1}{1 + e^{-2h\sum_{i=1}^{L_\mathcal{V}} \lambda_i v_i}} \tag{C13}$$

with $\Lambda = \{\lambda_i\}_i$. The choice of the parameters *defines* the RG rule. It is intuitively clear that while one could, in principle, consider any choice of Λ, the physically meaningful choices would correspond to the limit $||\Lambda||^2 \to \infty$, i.e., when the value of h actually strongly depends on v. In that limit, Eq. (C13) becomes a Heaviside function. This is also what happens in practice during the RSMI training (see the Supplemental Material in Ref. [28]).

Thus, the virtue of the RBM *Ansatz* is twofold: First, it provides an efficient tool from the algorithmic perspective of RSMI implementation, and second, it also provides a well-behaved, differentiable analytical *Ansatz*, which we use to explicitly calculate the quantities of interest. We emphasize though, that conceptually the RBM *Ansatz* is not essential to the RSMI approach. Any other parametrization of $P_\Lambda(\mathcal{H},\mathcal{V})$ can also be used at the expense of having to devise efficient algorithms to fix the parameters of this new *Ansatz*.

APPENDIX D: THE 1D ISING MODEL

For the 1D Ising model, Eq. (11), and a single hidden spin, we define

$$\mathcal{V}_j = \{x_{(j-1)L_\mathcal{V}+1}, x_{(j-1)L_\mathcal{V}+2}, \ldots, x_{jL_\mathcal{V}}\}, \tag{D1a}$$

$$\mathcal{H}_j = \{h_j\}. \tag{D1b}$$

The Hamiltonian decomposition Eq. (16) gives

$$\mathcal{K}_b[\mathcal{V}] = K \sum_{i=1}^{L_\mathcal{V}-1} v_i v_{i+1}, \tag{D2}$$

$$\mathcal{K}_1[\mathcal{X}] = K \sum_{j=1}^{n} x_{jL_\mathcal{V}} x_{jL_\mathcal{V}+1} \tag{D3}$$

with the partition functions $Z_0 = \prod_{j=1}^n Z_b$, where $Z_b = \sum_\mathcal{V} e^{\mathcal{K}_b[\mathcal{V}]}$.

The 1D Ising model with nearest-neighbor interactions can be solved exactly using the method of transfer matrices. To this end, we define the transfer matrix T with components $\langle x_1|T|x_2\rangle := e^{Kx_1x_2}$. The matrix elements of arbitrary integer powers of T can be computed by diagonalization:

$$\langle x_1|T^m|x_2\rangle = \frac{1}{2}[2\cosh(K)]^m[1 + x_1x_2\tanh(K)^m]. \tag{D4}$$

1. Exact decimation

For the purpose of numerical comparison with the RSMI solution, we perform one step of the *exact* decimation RG transformation Eq. (13). Following Eq. (4):

$$\begin{aligned} e^{\mathcal{K}'[\mathcal{X}']} &= \sum_{\mathcal{X}} \prod_{j=1}^n P_\Lambda(\mathcal{H}_j|\mathcal{V}_j) e^{K\sum_{i=1}^N x_i x_{i+1}} \\ &= \sum_{\mathcal{X}} \prod_{j=1}^n P_\Lambda(x'_j|\{x_{2j-1}, x_{2j}\}) \\ &\quad \times \langle x_{2j-1}|T|x_{2j}\rangle\langle x_{2j}|T|x_{2(j+1)-1}\rangle \\ &= \prod_{j=1}^n \langle x'_j|T^2|x'_{j+1}\rangle, \end{aligned} \tag{D5}$$

because for every block j the (δ)-like conditional probability $P_\Lambda(x'_i|\{x_{2j-1}, x_{2j}\})$ strictly enforces $x'_j = x_{2j-1}$ and does not involve x_{2j}. Thus, x_{2j} can simply be integrated out. The above has, up to a multiplicative constant $e^{c'}$, the same form as $e^{\mathcal{K}[\mathcal{X}]}$ with a new coupling constant K', such that we can set $e^{c'}T' = T^2$. From that, we obtain

$$c' = \frac{1}{2}\log[4\cosh(2K)], \tag{D5a}$$

$$K' = \frac{1}{2}\log[\cosh(2K)], \tag{D5b}$$

such that the renormalized Hamiltonian is

$$\mathcal{K}'[\mathcal{X}'] = \frac{n}{2}\log[4\cosh(2K)] + K'\sum_{i=1}^{n} x'_i x'_{i+1}. \tag{D6}$$

2. The effective Hamiltonian

Here we compute the effective block parameters Eq. (19) of the 1D Ising model for general block size $L_\mathcal{V}$.

Using Eq. (D4), the partition function of the intrablock contribution to the Hamiltonian is given by

$$\begin{aligned} Z_b &= \sum_{\mathcal{V}} \prod_{i=1}^{L_\mathcal{V}-1} \langle v_i|T|v_{i+1}\rangle = \sum_{v_1, v_{L_\mathcal{V}}} \langle v_1|T^{L_\mathcal{V}-1}|v_{L_\mathcal{V}}\rangle \\ &= 2[2\cosh(K)]^{L_\mathcal{V}-1}. \end{aligned} \tag{D7}$$

The expectations of powers of interblock couplings $\mathcal{K}_1[\mathcal{X}]^k$ appearing in the cumulant expansion can be written as a sum of products of operators acting on single blocks (see Appendix C 2). We have

$$\begin{aligned} \mathcal{K}_1[\mathcal{X}]^k &= \left(K\sum_{j=1}^{n} x_{jL_\mathcal{V}} x_{jL_\mathcal{V}+1}\right)^k \\ &= K^k \sum_{\sum_{j=1}^n k_j = k} \frac{k!}{\prod_{j=1}^n k_j!} \underbrace{\prod_{j=1}^{n} (x_{jL_\mathcal{V}} x_{jL_\mathcal{V}+1})^{k_j}}_{=:\mathcal{O}}. \end{aligned} \tag{D8}$$

We now consider one term in the above sum and rearrange the factors according to blocks:

$$\mathcal{O} = \prod_{j=1}^{n} \underbrace{x_{(j-1)L_\mathcal{V}+1}^{k_{j-1}} x_{jL_\mathcal{V}}^{k_j}}_{=:o_j}. \tag{D9}$$

Depending on the values of k_{j-1} and k_j, the block operator o_j is one of the following three operators: $x_{(j-1)L_\mathcal{V}+1}$, $x_{jL_\mathcal{V}}$, or $x_{(j-1)L_\mathcal{V}+1}x_{jL_\mathcal{V}}$. Hence, the average $\langle\mathcal{O}\rangle_{\Lambda,b}$ factorizes into

$$\langle x_{(j-1)L_\mathcal{V}+1}\rangle_{\Lambda,b}[\mathcal{H}_j], \tag{D9a}$$

$$\langle x_{jL_\mathcal{V}}\rangle_{\Lambda,b}[\mathcal{H}_j], \tag{D9b}$$

$$\langle x_{(j-1)L_\mathcal{V}+1} x_{jL_\mathcal{V}}\rangle_{\Lambda,b}[\mathcal{H}_j]. \tag{D9c}$$

The $\mathbb{Z}_2$ symmetry of the 1D Ising model can be used to extract the dependence of $P_{\Lambda,b}(h)$ and the above three quantities on the single hidden spin h:

$$\begin{aligned} P_{\Lambda,b}(h) &= \sum_{\mathcal{V}} \frac{e^{\mathcal{K}_b[\mathcal{V}]}}{Z_b} \underbrace{P_\Lambda(h|\mathcal{V})}_{=P_\Lambda(-h|-\mathcal{V})} \\ &= \sum_{\mathcal{V}} \frac{e^{\mathcal{K}_b[\mathcal{V}]}}{Z_b} P_\Lambda(-h|\mathcal{V}) = P_{\Lambda,b}(-h), \end{aligned} \tag{D10}$$

where we use the fact that for the $\mathbb{Z}_2$-symmetric system, the coarse graining satisfies $P_\Lambda(\mathcal{H}|\mathcal{V}) = P_\Lambda(-\mathcal{H}|-\mathcal{V})$. Since $P_{\Lambda,b}(h)$ is normalized, we have

$$P_{\Lambda,b}(h) = \frac{1}{2}. \tag{D11}$$

For any operator $\mathcal{O}_p[\mathcal{V}]$ with definite $\mathcal{V}$ parity $p = \pm 1$ given by $\mathcal{O}_p[-\mathcal{V}] = p\mathcal{O}_p[\mathcal{V}]$, we find using similar arguments that

$$\langle\mathcal{O}_p[\mathcal{V}]\rangle_{\Lambda,b}[h] = p\langle\mathcal{O}_p[\mathcal{V}]\rangle_{\Lambda,b}[-h], \tag{D12}$$

since $P_{\Lambda,b}(-\mathcal{V}|-h) = P_{\Lambda,b}(\mathcal{V}|h)$. Hence, $\langle\mathcal{O}_p[\mathcal{V}]\rangle_{\Lambda,b}[h]$ also has definite h-parity p.

Since h assumes only values ± 1, then $p = +1$ implies that the average is actually independent of h, while $p = -1$ implies it is linear in h. Thus,

$$\langle x_{(j-1)L_\mathcal{V}+1}\rangle_{\Lambda,b}[h] = \langle v_1\rangle_{\Lambda,b}[1]h, \tag{D12a}$$

$$\langle x_{jL_\mathcal{V}}\rangle_{\Lambda,b}[h] = \langle v_{L_\mathcal{V}}\rangle_{\Lambda,b}[1]h, \tag{D12b}$$

$$\langle x_{(j-1)L_\mathcal{V}+1} x_{jL_\mathcal{V}}\rangle_{\Lambda,b}[h] = \langle v_1 v_{L_\mathcal{V}}\rangle_{\Lambda,b}[1]. \tag{D12c}$$

The last expression can actually be explicitly calculated independently of the choice of the RG rule:

$$\begin{aligned} \langle v_1 v_{L_\mathcal{V}}\rangle_{\Lambda,b}[1] &= [2\cosh(K)]^{-(L_\mathcal{V}-1)} \\ &\quad\times \sum_{\mathcal{V}} v_1 v_{L_\mathcal{V}} e^{\mathcal{K}_b[\mathcal{V}]} P_\Lambda(1|\mathcal{V}). \end{aligned} \tag{D13}$$

Since $v_i = \pm 1$, we also have

$$\begin{aligned} e^{\mathcal{K}_b[\mathcal{V}]} &= \prod_{i=1}^{L_\mathcal{V}-1} e^{v_i v_{i+1}} \\ &= \cosh(K)^{L_\mathcal{V}-1} \prod_{i=1}^{L_\mathcal{V}-1} [1 + v_i v_{i+1}\tanh(K)]. \end{aligned} \tag{D14}$$

Every term in the expanded expression is of the form $\mathcal{O}[\mathcal{V}]\tanh(K)^m$ for an operator $\mathcal{O}$, which is a product of several consecutive pairs $v_i v_{i+1}$. If $\mathcal{O}$ has even $\mathcal{V}$ parity and $v_1 v_{L_\mathcal{V}}\mathcal{O}[\mathcal{V}]$ is not independent of $\mathcal{V}$, then

$$\sum_{\mathcal{V}} v_1 v_{L_\mathcal{V}} \mathcal{O}[\mathcal{V}] P_\Lambda(1|\mathcal{V}) = \sum_{\mathcal{V}} v_1 v_{L_\mathcal{V}} \mathcal{O}[\mathcal{V}] \underbrace{P_\Lambda(-1|-\mathcal{V})}_{=1-P_\Lambda(1|-\mathcal{V})}$$
$$= -\sum_{\mathcal{V}} v_1 v_{L_\mathcal{V}} \mathcal{O}[\mathcal{V}] P_\Lambda(1|\mathcal{V})$$
$$= 0. \tag{D15}$$

Thus, only $\mathcal{O}$ of odd $\mathcal{V}$ parity and those for which $v_1 v_{L_\mathcal{V}} \mathcal{O}[\mathcal{V}]$ is independent of $\mathcal{V}$ can contribute to Eq. (D13). However, $e^{\mathcal{K}_b[\mathcal{V}]}$ contains only two such contributions: 1 and $v_1 v_{L_\mathcal{V}} \tanh(K)^{L_\mathcal{V}-1}$. It follows that

$$\langle v_1 v_{L_\mathcal{V}} \rangle_{\Lambda,b}[1] = \tanh(K)^{L_\mathcal{V}-1} =: b; \tag{D16}$$

i.e., it is a Λ-independent constant. The remaining two averages depend on the choice of Λ, and closed expressions for them are given below for the case of block size $L_\mathcal{V} = 2$.

As discussed previously, the cumulants can be expressed in terms of the effective block parameters Eqs. (19). The actual computations can be done by brute-force summation of all possible terms in Eq. (20). This, however, is rather impractical for obtaining higher-order cumulants. We instead implement a simple algorithm based on the combinatorial considerations discussed in the main text.

3. The case of $L_\mathcal{V} = 2$ blocks: Discussion of the numerical results

Specializing to blocks of two visible spins results in

$$\mathcal{K}'[\mathcal{X}'] = \frac{N}{2} \log[2\cosh(K)] + \sum_{n=0}^{\infty} \frac{1}{n!} C_n(\mathcal{X}'), \tag{D17}$$

and the effective block parameters are found to be

$$a_1 = \frac{2[\cosh(\lambda_1)\sinh(\lambda_1) + \cosh(\lambda_2)\sinh(\lambda_2)\tanh(K)]}{\cosh(2\lambda_1) + \cosh(2\lambda_2)},$$
$$a_2 = \frac{2[\cosh(\lambda_2)\sinh(\lambda_2) + \cosh(\lambda_1)\sinh(\lambda_1)\tanh(K)]}{\cosh(2\lambda_1) + \cosh(2\lambda_2)},$$
$$b = \tanh(K). \tag{D18}$$

As we discuss in the main text, both the two-point correlator as a function of the distance between the spins (Fig. 10) and the m-point correlator as a function of the number of consecutive spins m (Fig. 11) decay exponentially for small K for the RSMI-favored solution (i.e., decimation). This solution, unsurprisingly, is decimation, which can be seen from Figs. 4 and 5. Additionally, in Fig. 12 we show the convergence to large-λ results shown in Fig. 5(a) with increasing order of the cumulant expansion.

We also comment on the asymmetry (around 0) of the curves in Figs. 5(a) and 5(b). The curves result from traversing the path $\lambda(\cos\theta, \sin\theta)$ in Fig. 4, which is *not* fourfold symmetric (instead, there are two reflection symmetries with respect to the diagonals). Starting from $\theta = 0$ at the peak, the trajectory traces out the lower branch of the curves in Figs. 5(a) and 5(b) reaching the lowest point at $\theta = \pi/4$, before turning around and exactly retracing the trajectory toward the peak at $\theta = \pi/2$. The trajectory then moves on the upper branch reaching the uppermost point at $\theta = 3\pi/4$ and retracing toward the peak again at $\theta = \pi/2$. This exact retracing is due to two independent $\mathbb{Z}_2$ symmetries: that of the Ising model and that of the mutual information. Since $\mathbb{Z}_2 \times \mathbb{Z}_2$ is not isomorphic to $\mathbb{Z}_4$, we do not have a fourfold symmetry

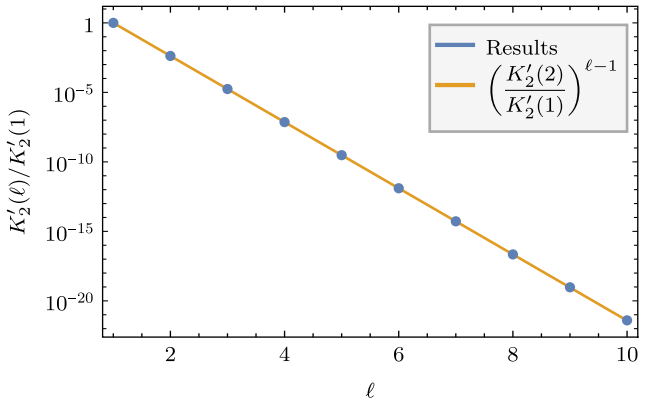

FIG. 10. Logarithmic plot showing the exponential decay of the two-point correlator $K_2'(\ell)$ with distance ℓ at $K = 0.1$. The blue data points represent the results obtained from the cumulant expansion of the RSMI-favored solution up to tenth order, while the yellow line shows the exponential decay with decay length obtained from the first two points. For small K, where the cumulant expansion is expected to be accurate, the two-point correlator decays exponentially.

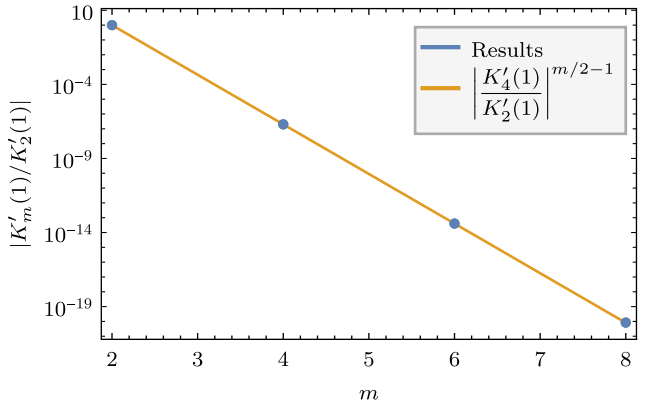

FIG. 11. Logarithmic plot showing the exponential decay of the nearest-neighbor m-point correlator $K_m'(1)$ with m at $K = 0.1$. The blue data points represent the results obtained from the cumulant expansion of the RSMI-favored solution up to tenth order, while the yellow line shows the exponential decay with decay length obtained from the first two points. Only points for even m are present, as $K_m(1) = 0$ for odd m due to reasons of symmetry. Again, at small K, the two-point correlator decays exponentially.

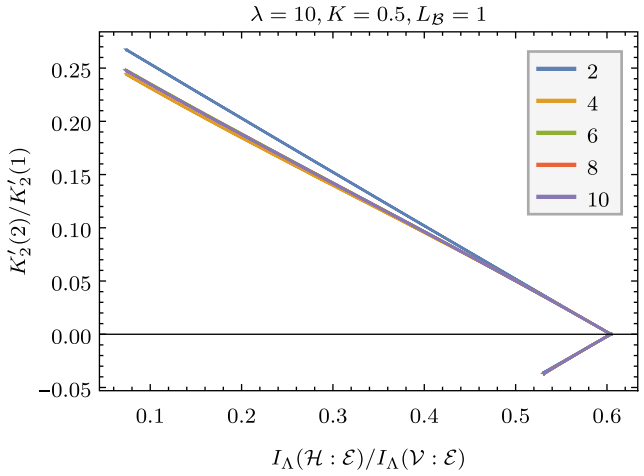

FIG. 12. The ratio between next-nearest-neighbor and nearest-neighbour coupling constants is plotted versus the mutual information for different orders of the cumulant expansion (see inset legend). The curves are obtained by parametrizing the RG rule by $(\lambda_1, \lambda_2) = \lambda[\cos(\theta), \sin(\theta)]$ and varying $\theta \in [0, \pi]$.

in Fig. 4, and consequently, we do not have a symmetry around 0 in Figs. 5(a) and 5(b). Physically, this is easily understood: The mutual information in Fig. 4(a) on the $\lambda_1 = -\lambda_2$ diagonal is lower than on the $\lambda_1 = \lambda_2$ one, since for the ferromagnetic Ising model we simulated that the neighboring spins are more likely to be aligned than not. Then, for the majority of the spin configurations, we have $\lambda_1 v_1 - \lambda_2 v_2 = 0$ on the $\lambda_1 = -\lambda_2$ diagonal, and hence, the coarse-graining rule decides the orientation of the effective spin at random, reducing the mutual information.

We emphasized before that the physically relevant coarse-graining rules are in the limit of large $||\Lambda||^2$. For small values of $||\Lambda||^2$, the coarse-graining rule is essentially independent of the underlying variables $\mathcal{V}$ (or equivalently, the rule can be thought of as having a large white-noise component). This independence manifests itself in Fig. 4(a) by low mutual information in the center. Nevertheless, Figs. 4(b) and 4(c) seem to have some (different looking) areas of vanishing rangeness and m-bodyness ratios in the center. Those are entirely accidental and nonuniversal. It is important to understand that since the central area corresponds to entirely randomly deciding the coarse-grained spin, the effective Hamiltonian (which would therefore have hardly *anything* to do with the physics of the underlying system) would not even contain nearest-neighbor terms. The central areas in Figs. 4(b) and 4(c) thus correspond to ratios of two vanishing quantities. Similarly, in Fig. 5(a) the position of the peak not being exactly at 0 for small $||\Lambda||^2$ is exactly due to the accidental features in the center of Fig. 4(b).

A slightly more practical lesson can be taken from Fig. 4(c), where even for larger λ, multiple crossings of the zero axis can be observed (i.e., the m-bodyness ratio vanishes also for some smaller value of mutual information, compared to the value at the peak, when the rangeness ratio is still large). This behavior is also accidental, but it teaches us that the proper metric to observe is the saturation of the mutual information (corresponding to the peak) and not the vanishing of some particular coefficient in the Hamiltonian (which may be accidental).

4. Mutual information

Here we explicitly calculate the information-theoretic quantities studied in the main text for the case of the NN Ising model in 1D given by Eq. (11) with a visible region of size $L_{\mathcal{V}}$ coupled to a single hidden spin $\mathcal{H} = \{h\}$. The system is split into four regions (see Fig. 2) with their respective sizes satisfying $N = L_{\mathcal{V}} + 2L_{\mathcal{B}} + 2L_{\mathcal{E}} + L_{\mathcal{O}}$. We denote the spin variables in the three inner regions of the system by

$$\mathcal{V} = \{v_1, v_2, \ldots, v_{L_{\mathcal{V}}}\}, \tag{D18a}$$

$$\mathcal{B} = \{b_{-L_{\mathcal{B}}}, b_{-L_{\mathcal{B}}+1}, \ldots, b_{-1}, b_1, b_2, \ldots, b_{L_{\mathcal{B}}}\}, \tag{D18b}$$

$$\mathcal{E} = \{e_{-L_{\mathcal{E}}}, e_{-L_{\mathcal{E}}+1}, \ldots, e_{-1}, e_1, e_2, \ldots, e_{L_{\mathcal{E}}}\}. \tag{D18c}$$

a. Mutual information between the hidden d.o.f. and the environment

The mutual information can be calculated from Eq. (A1). Since $\mathcal{H}$ is a binary variable, the two entropies appearing in Eq. (A1) can be rewritten in terms of the binary entropy $h_2(p) := -p\log(p) - (1-p)\log(1-p)$:

$$H(\mathcal{H}) = h_2[P_\Lambda(h=1)], \tag{D19}$$

$$H(\mathcal{H}|\mathcal{E}) = \langle h_2[P_\Lambda(h=1|\mathcal{E})]\rangle_{\mathcal{E}} \tag{D20}$$

with the conditional probability distribution

$$P_\Lambda(\mathcal{H}|\mathcal{E}) = \frac{\sum_{\mathcal{X}\setminus\mathcal{E}} P_\Lambda(\mathcal{H}|\mathcal{V})P(\mathcal{X})}{P(\mathcal{E})}. \tag{D21}$$

Thus, the mutual information is given by

$$I_\Lambda(\mathcal{H}:\mathcal{E}) = h_2[P_\Lambda(h=1)] - \langle h_2[P_\Lambda(h=1|\mathcal{E})]\rangle_{\mathcal{E}}. \tag{D22}$$

The relevant probability distributions $P_\Lambda(h)$ and $P_\Lambda(h|\mathcal{E})$ can be computed using transfer matrices (the result is always given in the limit $L_{\mathcal{O}} \to \infty$). For the former distribution, we observe that

$$P(\mathcal{V}) = \sum_{\mathcal{B},\mathcal{E},\mathcal{O}} P(\mathcal{X}) = \frac{1}{Z}\sum_{\mathcal{B},\mathcal{E},\mathcal{O}}\sum_{i=1}^{N}\langle x_i|T|x_i\rangle = \frac{e^{\mathcal{K}_b[\mathcal{V}]}}{Z_b},$$

which implies that $P_\Lambda(h) = \sum_{\mathcal{V}} P(h|\mathcal{V})P(\mathcal{V}) = P_{\Lambda,b}(h)$, in the thermodynamic limit. We have already found $P_{\Lambda,b}(h)$

in Eq. (D11) to be $\frac{1}{2}$, such that the first term in Eq. (D22) gives $h_2(1/2) = \log(2)$. The other relevant probability distribution is

$$P_\Lambda(h|\mathcal{E}) = \sum_{\mathcal{V}} P(h|\mathcal{V})P(\mathcal{V}|\mathcal{E}), \tag{D23}$$

where $P(h|\mathcal{V})$ is given by the RBM *Ansatz* Eq. (C13), and to obtain $P(\mathcal{V}|\mathcal{E})$, the two distributions $P(\mathcal{V},\mathcal{E})$ and $P(\mathcal{E})$ need to be computed. In the thermodynamic limit $L_\mathcal{O} \to \infty$, we obtain by Eq. (D4):

$$\begin{aligned} P(\mathcal{V},\mathcal{E}) &= \sum_{\mathcal{B},\mathcal{O}} P(\mathcal{X}) = \frac{1}{Z}\sum_{\mathcal{B},\mathcal{O}}\sum_{i=1}^{N}\langle x_i|T|x_i\rangle \\ &= \frac{1}{4}[1+v_1 e_{-1}G(L_\mathcal{B}+1)][1+v_2 e_1 G(L_\mathcal{B}+1)] \\ &\quad\times \frac{e^{K\sum_{\langle e,e'\rangle} ee'}}{[2\cosh(K)]^{2(L_\mathcal{E}-1)}}\frac{e^{\mathcal{K}_b[\mathcal{V}]}}{Z_b}, \end{aligned} \tag{D24}$$

since $\tanh(K)^m \to 0$ for $m \to \infty$ and finite K, and $Z = [2\cosh(K)]^N$ in the thermodynamic limit. Similarly,

$$\begin{aligned} P(\mathcal{E}) &= \sum_{\mathcal{V},\mathcal{B},\mathcal{O}} P(\mathcal{X}) = \frac{1}{Z}\sum_{\mathcal{V},\mathcal{B},\mathcal{O}}\sum_{i=1}^{N}\langle x_i|T|x_i\rangle \\ &= \frac{1}{4}[1+e_{-1}e_1 G(L_\mathcal{V}+2L_\mathcal{B}+1)] \\ &\quad\times \frac{e^{K\sum_{\langle e,e'\rangle} ee'}}{[2\cosh(K)]^{2(L_\mathcal{E}-1)}}, \end{aligned} \tag{D25}$$

$$\begin{aligned} P(\mathcal{V}|\mathcal{E}) &= \frac{[1+e_{-1}v_1 G(L_\mathcal{B}+1)][1+v_{L_\mathcal{V}}e_1 G(L_\mathcal{B}+1)]}{1+e_{-1}e_1 G(L_\mathcal{V}+2L_\mathcal{B}+1)} \\ &\quad\times \frac{e^{\mathcal{K}_b[\mathcal{V}]}}{Z_b}, \end{aligned} \tag{D26}$$

$$\begin{aligned} P_\Lambda(h|\mathcal{E}) &= \frac{1}{2}\sum_{\mathcal{V}} P_{\Lambda,b}(\mathcal{V}|h)\frac{1+e_{-1}v_1 G(L_\mathcal{B}+1)}{1+be_{-1}e_1 G[2(L_\mathcal{B}+1)]} \\ &\quad\times [1+v_{L_\mathcal{V}}e_1 G(L_\mathcal{B}+1)], \end{aligned} \tag{D27}$$

where we recognize $P_{\Lambda,b}(\mathcal{V}|h)$ from Eq. (C9) and use the fact that

$$\begin{aligned} G(L_\mathcal{V}+2L_\mathcal{B}+1) &= \tanh(K)^{L_\mathcal{V}-1}G[2(L_\mathcal{B}+1)] \\ &= bG[2(L_\mathcal{B}+1)]. \end{aligned} \tag{D28}$$

By expanding the numerator, we can rewrite the above in terms of averages $\langle\cdot\rangle_{\Lambda,b}$, and using Eq. (19) we obtain

$$\begin{aligned} P_\Lambda(h|\mathcal{E}) &= \frac{1+h(a_1e_{-1}+a_2e_1)G(L_\mathcal{B}+1)}{2\{1+be_{-1}e_1G[2(L_\mathcal{B}+1)]\}} \\ &\quad\times\frac{be_{-1}e_1G(2L_\mathcal{B}+2)}{2\{1+be_{-1}e_1G[2(L_\mathcal{B}+1)]\}}. \end{aligned} \tag{D29}$$

$P_\Lambda(h|\mathcal{E})$ depends only on the environment through $\{e_{-1},e_1\}$, so the sum over the remaining environment spins in the average over $P(\mathcal{E})$ can be performed explicitly, and we are left with an average over the marginal distribution: $P(e_{-1},e_1)=\frac{1}{4}[1+e_{-1}e_1G(2L_\mathcal{B}+3)]$. Finally, we can gather the results and obtain

$$\begin{aligned} &I_\Lambda(\mathcal{H}:\mathcal{E}) \\ &= \log(2) - \sum_{e_{-1},e_1} P(e_{-1},e_1)h_2 \\ &\quad\times\left(\frac{1+(a_1e_{-1}+a_2e_1)G(L_\mathcal{B}+1)+be_{-1}e_1G(2L_\mathcal{B}+2)}{2[1+be_{-1}e_1G(2L_\mathcal{B}+2)]}\right). \end{aligned} \tag{D30}$$

All dependence on Λ is in the block parameters a_1, a_2 (b is Λ independent) calculated in Eq. (D18).

b. Mutual information between the visibles and the environment

Equation (8) states that the mutual information between the hiddens and the environment $I_\Lambda(\mathcal{H}:\mathcal{E})$ is bounded from above by the mutual information between the visibles and the environment $I(\mathcal{V}:\mathcal{E})$. We now compute the latter quantity explicitly. By definition:

$$I(\mathcal{V}:\mathcal{E}) = \sum_{\mathcal{V},\mathcal{E}} P(\mathcal{V},\mathcal{E})\log\left(\frac{P(\mathcal{V},\mathcal{E})}{P(\mathcal{V})P(\mathcal{E})}\right), \tag{D31}$$

where all the probability distributions involved are already known; see Eqs. (D24) and (D25). Observe that the expression inside the logarithm depends only on the four spins e_{-1}, v_1, $v_{L_\mathcal{V}}$, and e_1, such that the sum over all other spins can be performed explicitly. We obtain

$$\begin{aligned} &I(\mathcal{V}:\mathcal{E}) \\ &= \sum_{\substack{e_{-1},v_1,\\ v_{L_\mathcal{V}},e_1}} P(e_{-1},v_1,v_{L_\mathcal{V}},e_1) \\ &\quad\times\log\left(\frac{[1+e_{-1}v_1G(L_\mathcal{B}+1)][1+v_{L_\mathcal{V}}e_1G(L_\mathcal{B}+1)]}{1+e_{-1}e_1G(L_\mathcal{V}+2L_\mathcal{B}+1)}\right), \end{aligned} \tag{D32}$$

$$\begin{aligned} &P(e_{-1},v_1,v_{L_\mathcal{V}},e_1) \\ &= \frac{1}{16}[1+e_{-1}v_1G(L_\mathcal{B}+1)] \\ &\quad\times[1+v_1v_{L_\mathcal{V}}G(L_\mathcal{V}-1)][1+v_{L_\mathcal{V}}e_1G(L_\mathcal{B}+1)]. \end{aligned} \tag{D33}$$

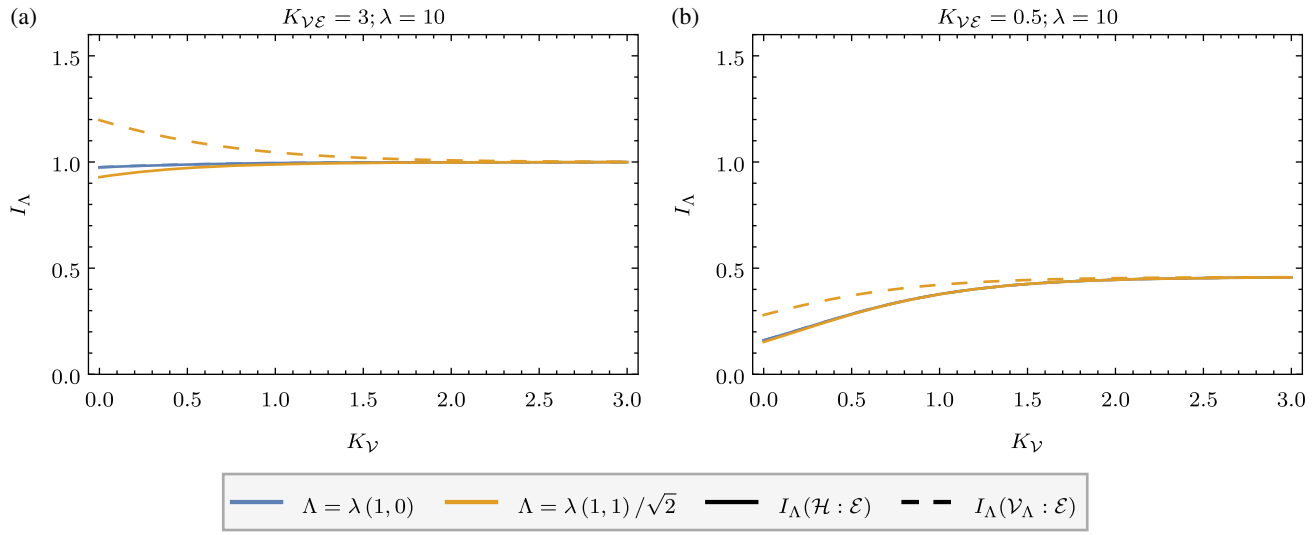

FIG. 13. The mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$ and $I(\mathcal{V}_\Lambda:\mathcal{E})$ for decimation (blue) and majority-rule (yellow) procedures in the 1D periodic toy model Eq. (E1) obtained by coupling the environment spins $e_{1,2}$ in the model Eq. (25) with a coupling $K_\mathcal{E} = 1.5$ (the value shown). Compare with Fig. 6, for which all other parameters are the same. Similarly, two parameter regimes are shown: (a) strong coupling to the environment or low-temperature $K_{\mathcal{VE}}$ (recall that the coupling constants contain a factor of $\beta = 1/k_B T$) and (b) weak coupling $K_{\mathcal{VE}}$. The solid lines differ from the dashed lines of the same color by the mismatch $I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$ (see the main text). Note that the introduction of $K_\mathcal{E}$ coupling rendering the environment simply connected also in this 1D case greatly reduces the difference between the two RG rules.

5. The case of larger blocks

For the case of $L_\mathcal{V} > 2$, additional subtleties are present. These subtleties can be attributed to differently broken symmetries in the mutual information and in the effective Hamiltonian.

On the level of interactions, the translation symmetry is explicitly broken by the Hamiltonian decomposition in Eq. (16) and subsequent cumulant expansion. This effect is not merely a feature of the method of evaluation, but rather a consequence of using a block-spin RG scheme: Interactions of the spins in the same block are inherently treated differently from interactions of the spins from different blocks. However, the full translational symmetry may sometimes be effectively restored. This happens, for instance, in the case of a decimation, when for any block size $L_\mathcal{V}$ it does not matter which single spin exactly is chosen in the block—the same effective Hamiltonian results.

When computing the mutual information, on the other hand, the full symmetry is not restored for $L_\mathcal{V} > 2$. The spins in the interior of the block are always coupled to the environment more weakly that the ones on the edges. Thus, we end up with two quantities, the renormalized Hamiltonian $\mathcal{K}'$ and the mutual information $I_\Lambda(\mathcal{H}:\mathcal{E})$, which have different symmetry properties. For example, for $L_\mathcal{V} = 3$ in the 1D Ising case, from the point of view of mutual information, we have two equivalent optimal solutions (coupling to leftmost and rightmost spins in the block), but it is intuitively clear that coupling to the center spin is equally good.

One important consequence is that the rangeness, for instance, is not necessarily a monotonic function of mutual information in the full parameter space (globally) but it is locally. Crucially though, any global maximum of mutual information corresponds to a global minimum of rangeness (but there could be additional equivalent solutions, just as the center spin in the $L_\mathcal{V} = 3$ decimation). The RSMI maximization is thus a *sufficient* criterion for a good RG transformation, establishing it as a variational principle. Further investigation of these effects for larger coarse-graining blocks might prove useful (see also the numerical results for the 2D Ising model case in the Supplemental Material of Ref. [28]).

APPENDIX E: TOY MODELS

1. 1D system

To illustrate the influence of the environment $\mathcal{E}$ being simply connected or not, we modify the 1D toy model Eq. (6) by introducing additional coupling $K_\mathcal{E}$ between the environment spins $e_{1,2}$, effectively making the system periodic (and thus, the environment simply connected):

$$\mathcal{K} = K_{\mathcal{VE}}(e_1 v_1 + v_2 e_2) + K_\mathcal{V} v_1 v_2 + K_\mathcal{E} e_1 e_2. \quad \text{(E1)}$$

This new coupling changes two things: On the one hand, the visibles become more strongly coupled to each other. On the other hand, since the environment for fixed $\mathcal{V}$ can no longer be thought of as being composed of two independent random variables $\mathcal{E}_1$ and $\mathcal{E}_2$, but rather a single one, the

information about the environment copied into the visible spins $v_{1,2}$ is much more correlated. This has the effect of reducing the mismatch $I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$. Indeed, as seen in Fig. 13, for the same values of all other parameters as in the nonperiodic case of Fig. 6, the discrepancy between the mutual information retained by the two coarse-graining rules is significantly decreased. Note though, that decimation still is (marginally) better.

2. 2D system

As we discuss in the main text, the situation in a two-dimensional system is qualitatively different. We consider the toy model with the Hamiltonian given by Eq. (26). Since all visibles couple to the same environment $\mathcal{E}$, which is now a single variable $E \in \{-4, -3, \ldots, 4\}$, in an identical fashion, each copies the same amount of information (at $K_\mathcal{V} = 0$). Similar to the 1D case, $\mathcal{V}_\Lambda$ captures more information about $\mathcal{E}$ if the coupling is more evenly distributed among the visibles. Additionally, with the connected environment, this coupling pattern has the effect of amplifying the shared information about $\mathcal{E}$ in each visible spin by averaging out the independent noise. While coupling to $\mathcal{V}_\Lambda$ always leads to more compression loss $I(\mathcal{V}_\Lambda:\mathcal{E}|\mathcal{H})$ compared to decimation, the scale of the two effects is different such that in the 2D (and presumably also in higher-dimensional) case, the information gain when coupling to more visibles outweighs the compression loss as seen in Fig. 7.

APPENDIX F: COMPARISON TO OTHER DEFINITIONS OF RG OPTIMALITY

The perfect action approach of Ref. [83] as well as those of Refs. [84,85] define renormalization schemes dubbed *optimal*. The goal, however, is fundamentally different from ours: The starting point is a continuum problem replaced by a coarse-grained version to numerically solve it. Optimality is defined as minimizing the error of the solution with respect to the solution of the continuum problem. In contrast, RSMI captures the long-range physics while discarding short-range fluctuations, in a very general information-theoretic sense. It is, by construction, optimally compressing long-range information. We show that reduced complexity of the effective theory (i.e., tractable Hamiltonian) is a consequence, even though it is not explicitly optimized for. Note that there is no reference problem, such as the continuum theory.

Perfect actions in field theories give cutoff-independent results on coarse-grained lattices [83]. They are not unique; in particular, the range and the order of interactions depend on the details of the coarse graining. This necessitates a second optimization to obtain an action approximated by as few couplings as possible. It is this second optimization that bears similarity to our problem. While, however, in RSMI tractability is a consequence of general principles, i.e., a result, in Ref. [83] it is explicitly optimized for by computing the actions for several different rules and tuning the RG procedure with respect to the range. The transformations studied were block averages optimized on a single parameter analogous to the magnitude of Λ, exploring only a small subset of transformations. While justified in examples, removing such choices altogether is the point of RSMI.

Coarse graining was also applied to numerically solving partial and stochastic differential equations [84–87]. The basic requirement is minimizing the difference between the solution of the coarse grained equation and the coarse-grained solution of the continuous equation. Fixed geometric coarse grainings [84] and approximations yielding short-range operators were discussed [86]. Geometric coarse grainings cannot be fully satisfactory, since dynamics can change the relevance of the d.o.f. Reference [85] proposed minimizing the error at later times, more specifically, by optimizing the coarse graining for the error incurred due to the noncommutativity of coarse graining and time evolution. We do not consider dynamical problems; however, a generalization of our approach would not involve a direct application of RSMI to coarse grain spatial d.o.f. This is inherently based on the notion that the important information to be compressed and preserved is the one pertaining to long spatial length scales (which is correct in deriving effective theory in equilibrium). Correctly generalizing the compression intuition, in the spirit of the information bottleneck method [32], would involve a procedure where the definition of the relevant information to be preserved also includes behavior at longer timescales. This is a very interesting question and another potential future research direction.

[1] L. E. Boltzmann, *Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht* (K.k. Hof- und Staatsdruckerei, Vienna, 1877).

[2] J. W. Gibbs, *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundations of Thermodynamics* (Charles Scribner's Sons, New York, 1902).

[3] C. E. Shannon, *A Mathematical Theory of Communication*, Bell Syst. Tech. J. **27**, 379 (1948).

[4] L. Szilard, *Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen*, Z. Phys. **53**, 840 (1929).

[5] R. W. Landauer, *Irreversibility and Heat Generation in the Computing Process*, IBM J. Res. Dev. **5**, 183 (1961).

[6] C. H. Bennett and P. W. Shor, *Quantum Information Theory*, IEEE Trans. Inf. Theory **44**, 2724 (1998).

[7] G. 't Hooft, *Dimensional Reduction in Quantum Gravity*, arXiv:gr-qc/9310026.

[8] L. Susskind, *The World as a Hologram*, J. Math. Phys. (N.Y.) **36**, 6377 (1995).

[9] R. Bousso, *The Holographic Principle*, Rev. Mod. Phys. **74**, 825 (2002).

[10] A. Kitaev and J. Preskill, *Topological Entanglement Entropy*, Phys. Rev. Lett. **96**, 110404 (2006).

[11] M. Levin and X.-G. Wen, *Detecting Topological Order in a Ground State Wave Function*, Phys. Rev. Lett. **96**, 110405 (2006).

[12] X. Chen, Z.-C. Gu, and X.-G. Wen, *Local Unitary Transformation, Long-Range Quantum Entanglement, Wave Function Renormalization, and Topological Order*, Phys. Rev. B **82**, 155138 (2010).

[13] S. Östlund and S. Rommer, *Thermodynamic Limit of Density Matrix Renormalization*, Phys. Rev. Lett. **75**, 3537 (1995).

[14] G. Vidal, *Class of Quantum Many-Body States That Can Be Efficiently Simulated*, Phys. Rev. Lett. **101**, 110501 (2008).

[15] S. R. White, *Density Matrix Formulation for Quantum Renormalization Groups*, Phys. Rev. Lett. **69**, 2863 (1992).

[16] G. Carleo and M. Troyer, *Solving the Quantum Many-Body Problem with Artificial Neural Networks*, Science **355**, 602 (2017).

[17] L. P. Kadanoff, *Scaling Laws for Ising Models near T_c*, Physics (Long Island City, N.Y.) **2**, 263 (1966).

[18] K. G. Wilson and J. Kogut, *The Renormalization Group and the ϵ Expansion*, Phys. Rep. **12**, 75 (1974).

[19] K. G. Wilson, *The Renormalization Group: Critical Phenomena and the Kondo Problem*, Rev. Mod. Phys. **47**, 773 (1975).

[20] M. E. Fisher, *Renormalization Group Theory: Its Basis and Formulation in Statistical Physics*, Rev. Mod. Phys. **70**, 653 (1998).

[21] E. Efrati, Z. Wang, A. Kolan, and L. P. Kadanoff, *Real-Space Renormalization in Statistical Mechanics*, Rev. Mod. Phys. **86**, 647 (2014).

[22] J. Gaite and D. O'Connor, *Field Theory Entropy, the h Theorem, and the Renormalization Group*, Phys. Rev. D **54**, 5163 (1996).

[23] J. Gaite, *Relative Entropy in 2D Quantum Field Theory, Finite-Size Corrections, and Irreversibility of the Renormalization Group*, Phys. Rev. Lett. **81**, 3587 (1998).

[24] S. M. Apenko, *Information Theory and Renormalization Group Flows*, Physica (Amsterdam) **391A**, 62 (2012).

[25] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, *Parameter Space Compression Underlies Emergent Theories and Predictive Models*, Science **342**, 604 (2013).

[26] C. Bény and T. J. Osborne, *The Renormalization Group via Statistical Inference*, New J. Phys. **17**, 083005 (2015).

[27] C. Bény and T. J. Osborne, *Information-Geometric Approach to the Renormalization Group*, Phys. Rev. A **92**, 022330 (2015).

[28] M. Koch-Janusz and Z. Ringel, *Mutual Information, Neural Networks and the Renormalization Group*, Nat. Phys. **14**, 578 (2018).

[29] S.-k. Ma, *Renormalization Group by Monte Carlo Methods*, Phys. Rev. Lett. **37**, 461 (1976).

[30] R. H. Swendsen, *Monte Carlo Renormalization-Group Studies of the $d = 2$ Ising Model*, Phys. Rev. B **20**, 2080 (1979).

[31] For a discussion of other works defining RG procedures which are "optimal" in a specific context, we refer to Appendix F.

[32] N. Tishby, F. C. Pereira, and W. Bialek, in *Proceedings of the 37th Allerton Conference on Communication, Control and Computation* (University of Illinois Press, Chicago, 2001).

[33] A. C. D. van Enter, R. Fernández, and A. D. Sokal, *Regularity Properties and Pathologies of Position-Space Renormalization-Group Transformations: Scope and Limitations of Gibbsian Theory*, J. Stat. Phys. **72**, 879 (1993).

[34] T. Kennedy, *Majority Rule at Low Temperatures on the Square and Triangular Lattices*, J. Stat. Phys. **86**, 1089 (1997).

[35] T. Niemeyer and J. M. J. Van Leeuwen, *Wilson Theory for 2-Dimensional Ising Spin Systems*, Physica (Utrecht) **71**, 17 (1974).

[36] M. E. Fisher, in *Proceedings of Stellenbosch 1982, Critical Phenomena* (Springer, Berlin, 1983), pp. 1–139.

[37] A. Brandt and D. Ron, *Renormalization Multigrid (RMG): Statistically Optimal Renormalization Group Flow and Coarse-to-Fine Monte Carlo Acceleration*, J. Stat. Phys. **102**, 231 (2001).

[38] The RSMI scheme can be understood as a realization of the famous information bottleneck compression of relevant information [32], under the constraint that the compressed variable is of predetermined type and size—a single block spin, for instance—as opposed to introducing a continuous Lagrange multiplier for its entropy (as originally). In this setting, the signal $\mathcal{V}$ is being (lossy) compressed to $\mathcal{H}$ and the "relevancy" variable determining the features of the signal to be preserved is the environment $\mathcal{E}$.

[39] H. A. Kramers and G. H. Wannier, *Statistics of the Two-Dimensional Ferromagnet. Part I*, Phys. Rev. **60**, 252 (1941).

[40] L. Onsager, *Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition*, Phys. Rev. **65**, 117 (1944).

[41] R. H. Swendsen, *Monte Carlo Calculation of Renormalized Coupling Parameters. I. $d = 2$ Ising Model*, Phys. Rev. B **30**, 3866 (1984).

[42] A. B. Harris, T. C. Lubensky, and J.-H. Chen, *Critical Properties of Spin-Glasses*, Phys. Rev. Lett. **36**, 415 (1976).

[43] G. Grinstein, A. N. Berker, J. Chalupa, and M. Wortis, *Exact Renormalization Group with Griffiths Singularities and Spin-Glass Behavior: The Random Ising Chain*, Phys. Rev. Lett. **36**, 1508 (1976).

[44] M. Wortis, C. Jayaprakash, and E. K. Riedel, *Thermodynamic Behavior of Quenched Random Magnets from a Position-Space Renormalization Group*, J. Appl. Phys. **49**, 1335 (1978).

[45] W. Kinzel and E. Domany, *Critical Properties of Random Potts Models*, Phys. Rev. B **23**, 3421 (1981).

[46] D. S. Fisher, P. Le Doussal, and C. Monthus, *Nonequilibrium Dynamics of Random Field Ising Spin Chains: Exact Results via Real Space Renormalization Group*, Phys. Rev. E **64**, 066107 (2001).

[47] M. C. Angelini and G. Biroli, *Real Space Renormalization Group Theory of Disordered Models of Glasses*, Proc. Natl. Acad. Sci. U.S.A. **114**, 3328 (2017).

[48] D. A. Keen and A. L. Goodwin, *The Crystallography of Correlated Disorder*, Nature (London) **521**, 303 (2015).

[49] S.-k. Ma, C. Dasgupta, and C.-k. Hu, *Random Antiferromagnetic Chain*, Phys. Rev. Lett. **43**, 1434 (1979).

[50] C. Dasgupta and S.-k. Ma, *Low-Temperature Properties of the Random Heisenberg Antiferromagnetic Chain*, Phys. Rev. B **22**, 1305 (1980).

[51] A block method inspired by strong-disorder RG has also been proposed in Ref. [52].

[52] C. Monthus and T. Garel, *Strong Disorder RG Principles within a Fixed Cell-Size Real Space Renormalization: Application to the Random Transverse Field Ising Model on Various Fractal Lattices*, J. Stat. Mech. (2012) P05002.

[53] M. Wortis, *Griffiths Singularities in the Randomly Dilute One-Dimensional Ising Model*, Phys. Rev. B **10**, 4665 (1974).

[54] R. B. Griffiths, *Nonanalytic Behavior above the Critical Point in a Random Ising Ferromagnet*, Phys. Rev. Lett. **23**, 17 (1969).

[55] G. J. Székely, M. L. Rizzo, and N. K. Bakirov*Measuring and Testing Dependence by Correlation of Distances*, Ann. Stat. **35**, 2769 (2007).

[56] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, Ann. Math. Stat. **22**, 79 (1951).

[57] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, in *Proceedings of the 35th International Conference on Machine Learning*, edited by J. Dy and A. Krause (PMLR, Stockholm, 2018), Vol. 80, pp. 531–540.

[58] D. G. Hernández and I. Samengo, *Estimating the Mutual Information between Two Discrete, Asymmetric Variables with Limited Samples*, Entropy **21**, 623 (2019).

[59] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov, *Optimal Structure and Parameter Learning of Ising Models*, Sci. Adv. **4** (2018).

[60] T. Kennedy, *Renormalization Group Maps for Ising Models in Lattice-Gas Variables*, J. Stat. Phys. **140**, 409 (2010).

[61] P. Ronhovde, S. Chakrabarty, D. Hu, M. Sahu, K. K. Sahu, K. F. Kelton, N. A. Mauro, and Z. Nussinov, *Detecting Hidden Spatial and Spatio-Temporal Structures in Glasses and Complex Physical Systems by Multiresolution Network Clustering*, Eur. Phys. J. E **34**, 105 (2011).

[62] P. Ronhovde, S. Chakrabarty, D. Hu, M. Sahu, K. K. Sahu, K. F. Kelton, N. A. Mauro, and Z. Nussinov, *Detection of Hidden Structures for Arbitrary Scales in Complex Physical Systems*, Sci. Rep. **2**, 329 (2012).

[63] D. Mendels, G. M. Piccini, and M. Parrinello, *Collective Variables from Local Fluctuations*, J. Phys. Chem. Lett. **9**, 2776 (2018).

[64] S. Salek, D. Cadamuro, P. Kammerlander, and K. Wiesner, *Quantum Rate-Distortion Coding of Relevant Information*, IEEE Trans. Inf. Theory **65**, 2603 (2019).

[65] C. J. Morningstar and M. Weinstein, *Contractor Renormalization Group Technology and Exact Hamiltonian Real-Space Renormalization Group Transformations*, Phys. Rev. D **54**, 4131 (1996).

[66] E. Altman and A. Auerbach, *Plaquette Boson-Fermion Model of Cuprates*, Phys. Rev. B **65**, 104508 (2002).

[67] E. Berg, E. Altman, and A. Auerbach, *Singlet Excitations in Pyrochlore: A Study of Quantum Frustration*, Phys. Rev. Lett. **90**, 147204 (2003).

[68] R. Budnik and A. Auerbach, *Low-Energy Singlets in the Heisenberg Antiferromagnet on the Kagome Lattice*, Phys. Rev. Lett. **93**, 187205 (2004).

[69] M. Hauru, C. Delcamp, and S. Mizera, *Renormalization of Tensor Networks Using Graph-Independent Local Truncations*, Phys. Rev. B **97**, 045111 (2018).

[70] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, *A Learning Algorithm for Boltzmann Machines*, Cogn. Sci. **9**, 147 (1985).

[71] G. E. Hinton and T. J. Sejnowski, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, edited by D. E. Rumelhart and J. L. Mcclelland (MIT Press, Cambridge, MA, 1986), Vol. 1, Chap. 4, pp. 282–317.

[72] R. Salakhutdinov and G. E. Hinton, *An Efficient Learning Procedure for Deep Boltzmann Machines*, Neural Comput. **24**, 1967 (2012).

[73] L. Wang, *Discovering Phase Transitions with Unsupervised Learning*, Phys. Rev. B **94**, 195105 (2016).

[74] G. Torlai and R. G. Melko, *Learning Thermodynamics with Boltzmann Machines*, Phys. Rev. B **94**, 165134 (2016).

[75] E. Stoudenmire and D. J. Schwab, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 4799–4807.

[76] J. Carrasquilla and R. G. Melko, *Machine Learning Phases of Matter*, Nat. Phys. **13**, 431 (2017).

[77] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber, *Learning Phase Transitions by Confusion*, Nat. Phys. **13**, 435 (2017).

[78] S.-H. Li and L. Wang, *Neural Network Renormalization Group*, Phys. Rev. Lett. **121**, 260601 (2018).

[79] Y.-Z. You, Z. Yang, and X.-L. Qi, *Machine Learning Spatial Geometry from Entanglement Features*, Phys. Rev. B **97**, 045153 (2018).

[80] J. Chen, S. Cheng, H. Xie, L. Wang, and T. Xiang, *Equivalence of Restricted Boltzmann Machines and Tensor Network States*, Phys. Rev. B **97**, 085104 (2018).

[81] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, *Reinforcement Learning with Neural Networks for Quantum Feedback*, Phys. Rev. X **8**, 031084 (2018).

[82] G. E. Hinton, *Training Products of Experts by Minimizing Contrastive Divergence*, Neural Comput. **14**, 1771 (2002).

[83] P. Hasenfratz and F. Niedermayer, *Perfect Lattice Action for Asymptotically Free Theories*, Nucl. Phys. **B414**, 785 (1994).

[84] N. Goldenfeld, A. McKane, and Q. Hou, *Block Spins for Partial Differential Equations*, J. Stat. Phys. **93**, 699 (1998).

[85] A. Degenhard and J. Rodriguez-Laguma, *Towards the Evaluation of the Relevant Degrees of Freedom in Nonlinear Partial Differential Equations*, J. Stat. Phys. **106**, 1093 (2002).

[86] Q. Hou, N. Goldenfeld, and A. McKane, *Renormalization Group and Perfect Operators for Stochastic Differential Equations*, Phys. Rev. E **63**, 036125 (2001).

[87] A. Degenhard and J. Rodriguez-Laguma, *Real-Space Renormalization-Group Approach to Field Evolution Equations*, Phys. Rev. E **65**, 036703 (2002).