

Relevance in the Renormalization Group and in Information Theory

Amit Gordon,¹ Aditya Banerjee,¹ Maciej Koch-Janusz,^{2,3} and Zohar Ringel¹

¹*Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel*

²*Department of Physics, University of Zurich, 8057 Zurich, Switzerland*

³*James Franck Institute, The University of Chicago, Chicago, Illinois 60637, USA*

The analysis of complex physical systems hinges on the ability to extract the relevant degrees of freedom from among the many others. Though much hope is placed in machine learning, it also brings challenges, chief of which is interpretability. It is often unclear what relation, if any, the architecture- and training-dependent learned “relevant” features bear to standard objects of physical theory. Here we report on theoretical results which may help to systematically address this issue: we establish equivalence between the information-theoretic notion of relevance defined in the Information Bottleneck (IB) formalism of compression theory, and the field-theoretic relevance of the Renormalization Group. We show analytically that for statistical physical systems described by a field theory the “relevant” degrees of freedom found using IB compression indeed correspond to operators with the lowest scaling dimensions. We confirm our field theoretic predictions numerically. We study dependence of the IB solutions on the physical symmetries of the data. Our findings provide a dictionary connecting two distinct theoretical toolboxes, and an example of constructively incorporating physical interpretability in applications of deep learning in physics.

The study of theoretical models is an essential part of physics. For sufficiently complex systems, however, establishing what the correct degrees of freedom are, and building a model in their terms, is a challenge in itself. The process is driven by experimental or numerical observations, but in practice physical intuition and prior knowledge are crucial to constructing a sufficiently simple model capturing the “essence” of the phenomenon, rather than abundance of raw data [1]. Still, data itself should contain sufficient information for this task, and a tantalizing prospect is to perform it in an unbiased, automatic fashion using modern computational methods, particularly deep learning (DL) [2]. A fundamental obstacle to this is the mismatch between the concepts of physics, largely formulated in the language of field theory, and the theory and engineering practice of DL, all but ensuring questions of interpretability [3]. To bridge this divide a framework is required capable of expressing, and allowing for practical computation, of quantities on both sides. Information theory, deeply connected to physics and computer science [4–6] is a natural candidate.

In its classical formulation information theory was intentionally agnostic to the contents of the information, focusing on its efficient transmission [7]. Though often only part of the information is pertinent to the problem, defining a formal notion of “relevance” in sufficient generality has proven difficult [8]. This was addressed in the seminal Information Bottleneck (IB) paper [9]: relevant information in a random variable was defined by correlations, or sharing information, with an auxiliary “relevance” variable, providing an implicit filter (an example of such correlated pairs are full frequency decomposition of a recorded speeches, and their written transcripts). Compressing data to preserve the relevant part most efficiently was cast as a Lagrangian problem, for which efficient DL methods have recently been introduced [10].

In physics, however, there already exists a fundamental and *a priori* independent notion of relevance, based on

the properties of the operators under scale transformations embodied in the celebrated renormalization group (RG) flow[11–13]. RG relevance is the most precise definition we possess of what it means for an observable to determine macroscopic physical properties of the system; it directly connects to the powerful formalism of conformal field theories (CFT)[14–17], which revolutionized the understanding of critical phenomena [18–20].

Here we show that these two notions, belonging to entirely different theoretical frameworks, are in fact equivalent in physical systems, *i.e.* the information about long-range properties “relevant” in the information-theoretic sense is formally determined by the most “relevant” operators in the sense of RG. Information loss in the context of RG has been attracting interest since the observation of irreversibility of its flow [21–29]; we introduce a formal connection to compression theory which is constructive, quantitative, and *computable*. This allows us to verify our predictions numerically. We prove that using the IB approach the most relevant operators can be extracted from the data, along with information about physical symmetries. This result is thus not only of theoretical, but also of practical importance. It provides a route towards automating theoretical tasks *e.g.* deriving Ginzburg-Landau effective descriptions, and detecting symmetries hidden or emergent, in a controlled and by construction interpretable way, using the toolbox of statistics and machine learning. To wit, while we focus on theoretical foundations, in a parallel work these results and recent DL advances [30, 31] are leveraged to construct an efficient algorithm, the real-space mutual information neural estimator (RSMI-NE)[32], extracting the physically most relevant degrees of freedom from much larger inputs, along the way characterizing spatial correlations, phase transitions and order parameters. We show here that RSMI is a limit of the IB problem, providing a theoretical underpinning for this promising numerical method.

Below we briefly review IB theory and its relation to

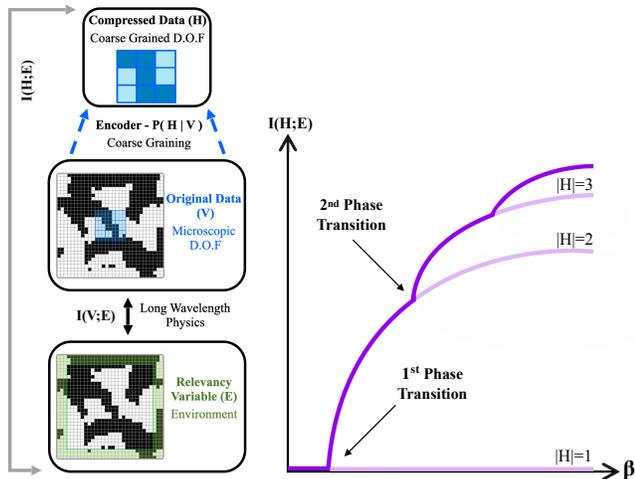


FIG. 1: **Left:** The general outline of the IB scheme, and in the physical setup of RSMI RG [33, 34]: an optimal encoder extracting information about “relevance” variable E contained in V is constructed. **Right:** IB curves depicting relevant information $I(H; E)$ retained by solutions to the IB equations (encoders), as a function of the tradeoff β (see Eq.1). At critical values of β phase transitions occur: new solutions, with compressed variable H of increased cardinality (*i.e.* tracking additional features) appear, while the old ones become unstable minima of \mathcal{L}_{IB} .

the RSMI approach to real-space RG in the context of statistical mechanical systems described by a CFT. We then present the main result: an analytical solution to the IB equations at strong compression which provides an explicit dictionary between IB relevancy, RG-relevancy, and eigenvectors of the transfer matrix in any dimension. We compare these predictions with numerics, obtaining agreement to high precision. In addition we show how symmetries are manifested in the compressed/coarse-grained degrees of freedom. Supplemental Materials give technical details and background information.

Relevant features of any data, physical or not, are only meaningfully defined relative to the task at hand, and their identification is complicated by multiple “irrelevant” (for the question asked) structures or regularities which may simultaneously exist in the data. The Information Bottleneck provides a rigorous framework for *unsupervised learning* of such most relevant features. With joint probability distribution of “data” V and an auxiliary “relevance” variable E as inputs, the IB finds the optimal (lossy) compression H of V preserving information about E (see Fig.1). The correlations with E thus *define* what is “relevant” in V , rather than arbitrary measures. IB can be posed as the following variational problem:

$$\min_{P(H|V)} \mathcal{L}_{IB}[P(H|V)] \equiv \min_{P(H|V)} I(V; H) - \beta I(H; E), \quad (1)$$

where the optimization is over conditional probability

distributions $P(H|V)$ describing the encoding of V into H . The mutual information terms I in \mathcal{L}_{IB} quantify total retained information (*i.e.* compression rate), and the relevant information thus preserved, respectively, with parameter $\beta \geq 0$ controlling the tradeoff between them.

The optimal encoder is found either by writing down a set of coupled IB equations for distributions $P(H|V)$, $P(H)$, $P(E|H)$ and solving them iteratively (see SM, and Ref.[35] for algorithms), or more practically, applying ML variational inference techniques [10]. For the formal analysis here the IB equations are used; the efficient RSMI-NE algorithm is based on ML methods [32]. Strikingly, the optimal encoders undergo a sequence of phase transitions as β is varied (see Fig.1). Particularly, the encoder is trivial (retaining zero information) until a *finite* value of $\beta_{c,1}$ at which the first IB transition occurs, when the gain due to retaining some (most) relevant aspect of data outweighs the penalty for keeping any information at all. At each subsequent transition the encoder begins to track another distinct *feature*. This discontinuous behaviour, both for discrete [36, 37] and continuous variables [38], is crucial, allowing to identify such well-defined features.

While the IB may be applied to any data, it is of fundamental interest to confront the notion of relevance it gives rise to, and the features it extracts, with *the* physical relevance, as defined by RG. The former being determined by the relevance variable, we need E ensuring the IB retains precisely the RG-relevant information. An appropriate definition for real-space RG was postulated in the context of RSMI [33, 34]: for a random variable V representing the marginal distribution of degrees of freedom in an area to be coarse-grained the variable E (the “environment”), is the remainder of the system beyond a shell of non-zero thickness around V (the “buffer”, see Fig.2). The thickness of the excluded buffer, formally taken to infinity, sets the length scale separating short-range correlations to be discarded, from information about long-range properties of the system. Despite conceptual appeal – the system itself defines relevance – and partial numerical [33] and theoretical evidence [34], the validity of this approach and its relation to RG field-theory formalism were unclear. There are also subtle differences between the IB and RSMI approaches. We now can resolve these issues.

To this end consider a statistical mechanical system on a cylinder; the subsystem to be coarse-grained V , the buffer, and the relevance variable E are its subsections as per Fig.2. We assume the system is governed by short-range interactions, and use the classic transfer-matrix (TM) method [39–41]: the partition function can be written as $\mathcal{Z} = \langle BC | \mathcal{T}^{L_\infty} | BC \rangle$, where is L_∞ the system length, and the entries of \mathcal{T} are matrix elements of the exponentiated Hamiltonian between configurations of degrees of freedom on elementary slices of the cylinder (in a lattice system; in continuum they are taken between subsequent slices of the states in the discretised path integral picture). We use bracket notation for such configurations, in particular $|BC\rangle$ are boundary conditions at the cylinder ends. The unique advantage of the

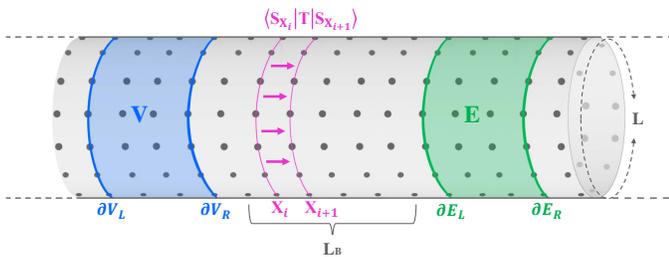


FIG. 2: The transfer matrix (TM) setup used. For a system on a cylinder the IB equations can be solved in terms of TM eigenvectors, which are related to the CFT data in the limit of large circumference L .

TM approach is that, on the one hand, all distributions entering the IB equations can be cast as matrix elements and partial traces of powers of \mathcal{T} , and on the other hand the eigenvalues λ_i and eigenvectors $|i\rangle$ of \mathcal{T} have a direct relation to the operator content of the CFT describing the system [42–44]. Specifically, $\lambda_i/\lambda_0 = e^{-\frac{2\pi}{L}\Delta_i}$ in the limit of large cylinder circumference L , where Δ_i are the total scaling dimensions of the CFT primaries (determining the RG scaling dimensions, and so the critical exponents) in ascending order. TM thus serves as a theoretical dictionary helping to establish a *quantitative* map between the field- and information-theoretic objects.

To be concrete, consider the IB equations for the optimal encoder $P(h|v)$ at fixed tradeoff β (see SM and [9]):

$$P(h|v) \propto P(h)e^{\beta \sum_e P(e|v) \log(P(e|h))} \quad (2)$$

$$P(e|h) = \sum_v p(e|v)p(v|h),$$

where e, h, v are configurations of E, H and V. Observe first that the distribution $P(X)$ of any cylindrical section X of the system (*e.g.* V or E) can be written using \mathcal{T} :

$$P(X) = \langle 0|\partial x_L\rangle \langle \partial x_L|\mathcal{T}|x_2\rangle \langle x_2|\mathcal{T} \dots \mathcal{T}|\partial x_R\rangle \langle \partial x_R|0\rangle$$

Here x_i are successive slices of X ; the configurations of the boundary slices are denoted as $\partial x_{R/L}$ and $|0\rangle = \mathcal{T}^{L_\infty}|BC\rangle$ is the CFT vacuum, on which \mathcal{T} acts as an identity. We used the eigendecomposition $\mathcal{T} = |0\rangle\langle 0| + \sum_i e^{-2\pi\Delta_i/L}|\Delta_i\rangle\langle\Delta_i|$, with $|\Delta_i\rangle = \phi_{\Delta_i}|0\rangle$ created by primary fields ϕ_{Δ_i} with conformal dimension Δ_i . All distributions in Eqs.2 can be written analogously as functions of \mathcal{T} , using the eigendecomposition and Bayes' law, phrasing the IB equations fully in TM terms.

Eqs.2 are highly non-linear and coupled. Remarkably though, the only interdependence of the conditional probabilities in the large buffer limit is, as shown in SM, via:

$$r_v = \frac{\langle 0|\phi_{\Delta_1}|\partial V_R\rangle}{\langle 0|\partial V_R\rangle} \quad r_e = \frac{\langle \partial E_L|\phi_{\Delta_1}|0\rangle}{\langle \partial E_L|0\rangle}, \quad (3)$$

i.e. the matrix elements of the CFT primary fields of lowest scaling dimensions. In particular, we prove that:

$$P(h|v) = P(h|r_v) \propto P(h)e^{\beta\epsilon^2 r_v \langle r_v \rangle_h}, \quad (4)$$

with $\epsilon = (\lambda_1/\lambda_0)^{L_B}$.

Eq.4 is one of our key results: the optimal IB encoder depends on V *only* via r_v , *i.e.* the matrix element of the most relevant operator in the sense of RG (the dependence on the variance of r_e can be absorbed into rescaling of β , and in $\langle r_v \rangle_h$ v -dependence is averaged over). The solution changes from one system (CFT) to another through the values of r_v and $\langle r_v \rangle_h$. This is the mathematical statement of the equivalence of the IB and RG relevance. In other words, the “features” the IB, and consequently the RSMI, extract are not arbitrary, but correspond to physically most relevant operators.

Though Eq.4 is implicit, as $\langle r_v \rangle_h$ depends on $P(h|v)$, it can be analytically solved around the first IB transition, *i.e.* for $\beta = \beta_{c,1} + t$. Below $\beta_{c,1}$ no information is retained: the encoder is independent of V and trivial: $P(h|r_v) = 1/|H|$, with $|H|$ the cardinality of the coarse-grained variable. Equiprobability of h reflects a structural symmetry of the encoder under permutations of h labels. Any nontrivial encoder *must* break it, introducing dependence of some h on V to preserve information: $\beta_{c,1}$ marks the first such breaking (in fact all IB transitions reflect successive breaking of permutation symmetry). Above $\beta_{c,1}$, following Refs.[36, 37], the encoder can be perturbatively expanded around the trivial solution (see SM for detailed discussion). In particular, comparing to the expansion of Eq.4 in t yields:

$$\beta_{c,1}^{-1} = \epsilon^2 + o(\epsilon^2) \xrightarrow{L \rightarrow \infty} e^{-4\pi\Delta_1 \frac{L_B}{L}} + o(\epsilon^2). \quad (5)$$

Here $o(\epsilon^2)$, containing powers of ϵ greater than two, reflects the contribution of operators of subleading relevance. As ϵ decays exponentially in L_B/L , maintaining $L_B \gg L$ suppresses these corrections exponentially.

Equation 5 is an analytical prediction for the IB phase transition, signaling emergence of nontrivial solutions to the IB equations (see Fig.1 and Fig.4 in SM), in terms of field-theoretic quantities characterizing the physical system. In SM, utilizing the structure of the Hessian of \mathcal{L}_{IB} , we also derive this solution explicitly (see also Fig.3).

The prediction is generic and verifiable: we can input the probability distribution of the system to the IB equations, and find the solutions for changing β numerically, as in a compression problem [35]. On the other hand we can use the CFT description and either compute r_v , $\langle r_v \rangle_h$ and ϵ analytically, or by a numerical transfer matrix diagonalization, and compare. In Fig.3c numerical IB solutions are plotted as a function of β in the case of critical 2D Ising model. The value $\beta_{c,1}^{IB}$ at which nontrivial encoders appear matches the predicted $\beta_{c,1}$ to high accuracy. The feature the IB extracts is indeed the most relevant local operator, *i.e.* the magnetization (see SM).

The validity of this picture is not limited to lattice models. In fact, for the continuum Gaussian field theory the entire IB curve can be computed analytically, including *all* the IB phase transitions [45], using Gaussian Information Bottleneck results [38] and Green's functions.

As mentioned, the RSMI algorithm [33, 34] is closely related to the IB. Specifically, it also maximizes the rel-

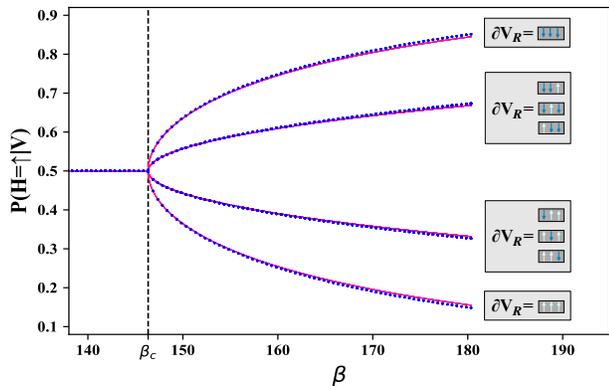


FIG. 3: Comparison of theory with numerics. For the critical 2D Ising system, the analytical prediction (solid red) for the optimal compression $P_\beta(h|v)$ (see Eq.4, and Eq.C6 in SM) is confronted with encoders obtained by numerically solving IB Eqs.2 on the probability distribution of the system (blue dots). For clarity we use a cylinder of three sites' circumference, V and E as in Fig.2. The variable H is a spin, whose probability to take value \uparrow we plot as a function of the tradeoff β (see Eq.1). The encoder is completely random and independent of V below β_c matching the prediction Eq.5, and above is determined by the magnetization on the edge, in excellent agreement with the theory.

evant information $I(H; E)$, however contains no tradeoff β , but instead a fixed cardinality $|H|$. Intuitively, the IB extracts as many features as β allows, adding them as β grows, while the RSMI from the outset optimizes exactly $|H|$ best features. RSMI is thus a $\beta \rightarrow \infty$ limit of IB under the constraint of fixed $|H|$. In practice $|H|$ is also bounded in IB, however this affects solutions only at β large enough for $|H|$ features to have already been used.

The quantitative connection between compression- and field-theoretic formalisms thus established opens the exciting possibility of applying distinct theoretical and numerical methods of either area to its counterpart. We discuss such avenues in the conclusions, here, however, we immediately demonstrate one interesting example.

Symmetries are crucial in analytical understanding of physical systems, and in RG in particular [46]. They have a direct relation to order parameters, and often effectively determine the long range properties. One thus expects IB and RSMI to reflect the relevant symmetries of the model. Let s be an element of such symmetry group \mathcal{S} acting on configurations of V and E as a permutation, denoted by multiplication, leaving the system invariant: $P(e, v) = P(se, sv)$. We expect the optimal encoder $P_\beta(h|v)$ to maintain it:

$$P(e, v) = P(se, sv) \Rightarrow P_\beta(h|v) = P_\beta(\phi_s h|sv), \quad (6)$$

so that the coarse-grained system is invariant under a representation ϕ_s of \mathcal{S} , potentially trivial. We show this indeed holds true in IB, as long as $|H|$ is large enough

to support a representation of an appropriate dimension. The argument, detailed in SM, is constructive: below $\beta_{c,1}$ the encoder is trivially invariant under all symmetries. For $\beta = \beta_{c,1} + t$ a solution can be built by an explicit symmetrization procedure, utilizing the knowledge of the perturbative structure of the encoder and the Hessian of \mathcal{L}_{IB} around the first IB transition [37]. We show this solution to be optimal. The symmetry of the encoder will hold for all $\beta < \beta_{c,2}$ by continuity; numerical experiments support validity of this picture also more generally.

Note that the symmetry \mathcal{S} may not be obvious in the microscopic formulation of the system [47] or the experimental data, or may even be emergent [48]. Eq. 6 can then be used as a constructive tool, potentially allowing to systematically learn \mathcal{S} from the symmetries of the entries of the numerically obtained $P_\beta(h|v)$ (SM, see also [49]). Moreover, the structure of the IB in the presence of physical/data symmetries shines light on the question of constructing RG transformations compatible with the symmetries of the system.

The results we presented, though requiring some level of technicality, have clear theoretical interpretations. In fact, their very point is to formalize concepts and connections which ought to be intuitive, and to give the necessary technology to make those quantitative and computable analytically and numerically. Consequently, numerous directions are now open. On a theoretical front, application of IB analysis to extract relevant quantities in the challenging case of disordered and non-equilibrium systems is extremely promising, given its non-reliance on the notion of a Hamiltonian. This may require deeper understanding of the properties of the IB equations, and their constrained version in the RSMI-NE algorithm. Numerically, given the relation to the transfer matrix, the possibility of using the IB/RSMI where TM computations are difficult (*e.g.* in 3D) is an exciting prospect, as is applying approximate numerical IB or RSMI to experimental data. Finally, we hope that the methodology of using information-theoretical formulations of physical quantities combined with the ability of deep learning to optimize them in a controlled fashion [32], can provide a blueprint for more theoretically interpretable applications of deep learning in physics.

Acknowledgements: ZR and AG acknowledge support from ISF grant 2250/19 as well as helpful discussions with Etam Benger (HUJI). MKJ gratefully acknowledges the support of Sebastian Huber, during stay in whose group at ETH Zurich part of the work was performed, the financial support from the Swiss National Science Foundation and the NCCR QSIT, and from the European Research Council under the Grant Agreement No. 771503 (TopMechMat), as well as from European Union's Horizon 2020 programme under Marie Skłodowska-Curie Grant Agreement No. 896004 (COMPLEX ML). AB acknowledges financial support from Eldad Bettelheim's ISF grant 1466/15 and Alex Retzker's MicroQC grant number 820314.

-
- [1] Peter V. Coveney, Edward R. Dougherty, and Roger R. Highfield, “Big data need big theory too,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374** (2016), 10.1098/rsta.2016.0153.
- [2] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová, “Machine learning and the physical sciences,” *Rev. Mod. Phys.* **91**, 045002 (2019).
- [3] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu, “Definitions, methods, and applications in interpretable machine learning,” *Proceedings of the National Academy of Sciences* **116**, 22071–22080 (2019).
- [4] John Preskill, “Quantum information and physics: Some future directions,” *Journal of Modern Optics* **47**, 127–137 (2000).
- [5] David J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2002).
- [6] Marc Mezard and Andrea Montanari, *Information, Physics, and Computation* (Oxford University Press, Inc., USA, 2009).
- [7] Claude E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.* **27**, 379–423 (1948).
- [8] Yoichi Blau and Tomer Michaeli, “Rethinking Lossy Compression: The Rate-Distortion-Perception Trade-off,” in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, Proceedings of Machine Learning Research, Vol. 97 (PMLR, 2019) pp. 675–685.
- [9] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, **49** (2001).
- [10] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy, “Deep variational information bottleneck,” *CoRR* [abs/1612.00410](https://arxiv.org/abs/1612.00410) (2016), [arXiv:1612.00410](https://arxiv.org/abs/1612.00410).
- [11] Kenneth G. Wilson and John Kogut, “The renormalization group and the ϵ expansion,” *Physics Reports* **12**, 75 – 199 (1974).
- [12] Kenneth G. Wilson, “The renormalization group: Critical phenomena and the Kondo problem,” *Rev. Mod. Phys.* **47**, 773–840 (1975).
- [13] Michael E. Fisher, “Renormalization group theory: Its basis and formulation in statistical physics,” *Rev. Mod. Phys.* **70**, 653–681 (1998).
- [14] A.A. Belavin, A.M. Polyakov, and A.B. Zamolodchikov, “Infinite conformal symmetry of critical fluctuations in two dimensions,” *Journal of Statistical Physics* **34**, 763–774 (1984).
- [15] A.A. Belavin, A.M. Polyakov, and A.B. Zamolodchikov, “Infinite conformal symmetry in two-dimensional quantum field theory,” *Nuclear Physics B* **241**, 333 – 380 (1984).
- [16] Daniel Friedan, Zonghan Qiu, and Stephen Shenker, “Conformal Invariance, Unitarity, and Critical Exponents in Two Dimensions,” *Phys. Rev. Lett.* **52**, 1575–1578 (1984).
- [17] Philippe Di Francesco, Pierre Mathieu, and David Sénéchal, *Conformal field theory*, Graduate texts in contemporary physics (Springer, New York, NY, 1997).
- [18] John L Cardy, *Scaling and renormalization in statistical physics*, Cambridge lectue notes in physics (Cambridge Univ. Press, Cambridge, 1996).
- [19] C. Itzykson, H. Saleur, and J.-B. Zuber, *Conformal Invariance and Applications to Statistical Mechanics* (World Scientific, 1998).
- [20] David Poland, Slava Rychkov, and Alessandro Vichi, “The conformal bootstrap: Theory, numerical techniques, and applications,” *Rev. Mod. Phys.* **91**, 015002 (2019).
- [21] A.B. Zamolodchikov, “Irreversibility of the Flux of the Renormalization Group in a 2D Field Theory,” *JETP Lett.* **43**, 730–732 (1986).
- [22] José Gaiete and Denjoe O’Connor, “Field theory entropy, the h theorem, and the renormalization group,” *Phys. Rev. D* **54**, 5163–5173 (1996).
- [23] H. Casini and M. Huerta, “A c-theorem for entanglement entropy,” *Journal of Physics A: Mathematical and Theoretical* **40**, 7031–7036 (2007).
- [24] Sergey M. Apenko, “Information theory and renormalization group flows,” *Physica A: Statistical Mechanics and its Applications* **391**, 62 – 77 (2012).
- [25] Benjamin B. Machta, Ricky Chachra, Mark K. Transtrum, and James P. Sethna, “Parameter Space Compression Underlies Emergent Theories and Predictive Models,” *Science* **342**, 604–607 (2013).
- [26] Vijay Balasubramanian, Jonathan J. Heckman, and Alexander Maloney, “Relative Entropy and Proximity of Quantum Field Theories,” *JHEP* **05**, 104 (2015), [arXiv:1410.6809 \[hep-th\]](https://arxiv.org/abs/1410.6809).
- [27] Cédric Bény and Tobias J. Osborne, “The renormalization group via statistical inference,” *New Journal of Physics* **17**, 083005 (2015).
- [28] Cédric Bény and Tobias J. Osborne, “Information-geometric approach to the renormalization group,” *Phys. Rev. A* **92**, 022330 (2015).
- [29] Cédric Bény, “Coarse-grained distinguishability of field interactions,” *Quantum* **2**, 67 (2018).
- [30] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rameshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm, “Mutual Information Neural Estimation,” (PMLR, 2018) pp. 531–540, [arXiv:1801.04062](https://arxiv.org/abs/1801.04062).
- [31] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker, “On Variational Bounds of Mutual Information,” (PMLR, 2019) pp. 5171–5180, [arXiv:1905.06922](https://arxiv.org/abs/1905.06922).
- [32] Doruk Efe Gökmen, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz, “Real-space mutual information neural estimation (in preparation),” .
- [33] Maciej Koch-Janusz and Zohar Ringel, “Mutual information, neural networks and the renormalization group,” *Nature Physics* **14**, 578–582 (2018).
- [34] Patrick M. Lenggenhager, Doruk Efe Gökmen, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz, “Optimal renormalization group transformation from information theory,” *Phys. Rev. X* **10**, 011037 (2020).
- [35] S. Hassanpour, D. Wuebben, and A. Dekorsy, “Overview and Investigation of Algorithms for the Information Bot-

- tleneck Method,” in *SCC 2017; 11th International ITG Conference on Systems, Communications and Coding* (2017) pp. 1–6.
- [36] Albert E. Parker, Tomáš Gedeon, and Alexander G. Dimitrov, “Annealing and the Rate Distortion Problem,” in *Proceedings of the 15th International Conference on Neural Information Processing Systems, NIPS’02* (MIT Press, Cambridge, MA, USA, 2002) p. 993–976.
- [37] Tomas Gedeon, Albert E. Parker, and Alexander G. Dimitrov, “The Mathematical structure of Information Bottleneck Methods,” *Entropy* **14**, 456–479 (2012).
- [38] Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss, “Information Bottleneck for Gaussian Variables,” in *Advances in Neural Information Processing Systems 16*, edited by S. Thrun, L. K. Saul, and B. Schölkopf (MIT Press, 2004) pp. 1213–1220.
- [39] H. A. Kramers and G. H. Wannier, “Statistics of the Two-Dimensional Ferromagnet. Part I,” *Physical Review* **60**, 252–262 (1941).
- [40] Lars Onsager, “Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition,” *Physical Review* **65**, 117–149 (1944).
- [41] Peter Nightingale, “Finite-size scaling and phenomenological renormalization,” *Journal of Applied Physics* **53**, 7927–7932 (1982), <https://doi.org/10.1063/1.330232>.
- [42] B. Derrida and L. De Seze, “Application of the phenomenological renormalization to percolation and lattice animals in dimension 2,” *J.Physique* **43**, 475–483 (1982).
- [43] J L Cardy, “Conformal invariance and universality in finite-size scaling,” *Journal of Physics A: Mathematical and General* **17**, L385–L387 (1984).
- [44] John L. Cardy, “Operator content of two-dimensional conformally invariant theories,” *Nuclear Physics B* **270**, 186–204 (1986).
- [45] Aditya Banerjee and Zohar Ringel, “Information bottleneck and Gaussian field theory (in preparation).”
- [46] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena*, International series of monographs on physics (Clarendon Press, 1989).
- [47] Chen Ning Yang and S.C. Zhang, “SO4 Symmetry in a Hubbard model,” *Modern Physics Letters B* **04**, 759–766 (1990).
- [48] T. Senthil, Ashvin Vishwanath, Leon Balents, Subir Sachdev, and Matthew P. A. Fisher, “Deconfined Quantum Critical Points,” *Science* **303**, 1490–1494 (2004).
- [49] Roberto Bondesan and Austen Lamacraft, “Learning Symmetries of Classical Integrable Systems,” [abs/1906.04645](https://arxiv.org/abs/1906.04645) (2019), [arXiv:1906.04645](https://arxiv.org/abs/1906.04645).
- [50] Elad Schneidman, Noam Slonim, Naftali Tishby, Rob R. de Ruyter van Steveninck, and William Bialek, “Analyzing Neural Codes Using the Information Bottleneck Method,” (2001).
- [51] Felix Creutzig and Henning Sprekeler, “Predictive Coding and the Slowness Principle: An Information-Theoretic Approach,” *Neural Computation* **20**, 1026–1041 (2008).
- [52] Lars Buesing and Wolfgang Maass, “A Spiking Neuron as Information Bottleneck,” *Neural Computation* **22**, 1961–1992 (2010), pMID: 20337537.
- [53] Noam Slonim and Naftali Tishby, “Document Clustering Using Word Clusters via the Information Bottleneck Method,” in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’00* (2000) p. 208–215.
- [54] Susanne Still, William Bialek, and Léon Bottou, “Geometric Clustering Using the Information Bottleneck Method,” in *Advances in Neural Information Processing Systems*, Vol. 16, edited by S. Thrun, L. Saul, and B. Schölkopf (MIT Press, 2004) pp. 1165–1172.
- [55] DJ Strouse and David J. Schwab, “The Information Bottleneck and Geometric Clustering,” *Neural Computation* **31**, 596–612 (2019), pMID: 30314426.
- [56] Felix Creutzig, Amir Globerson, and Naftali Tishby, “Past-future information bottleneck in dynamical systems,” *Phys. Rev. E* **79**, 041925 (2009).
- [57] Susanne Still, “Information Bottleneck Approach to Predictive Inference,” *Entropy* **16**, 968–989 (2014).
- [58] Etam Bengler and Naftali Tishby, “All Information Bottleneck Phase Transitions are Local (in preparation).”
- [59] Noam Slonim, *The information bottleneck: Theory and applications*, Ph.D. thesis (2002).

Appendix A: The IB Equations

The IB problem Ref.[9], as described in the main text, is set up as a minimization problem over the class of conditional probability distributions $P(h|v)$ of the following IB Lagrangian:

$$\mathcal{L}_{IB}[P(H|V)] \equiv I(V; H) - \beta I(H; E), \quad (\text{A1})$$

Somewhat surprisingly (since the terms in \mathcal{L}_{IB} are highly nonlinear) a formal solution can be found by performing the variation $\delta\mathcal{L}_{IB}/\delta P(h|v) = 0$. As shown in Ref.[9], the optimal solution can be written as:

$$P(h|v) = \frac{P(h)}{Z(v, \beta)} \exp(-\beta D_{KL}[P(e|x)|P(e|h)]), \quad (\text{A2})$$

where D_{KL} is the Kullback-Leibler divergence of the conditional probability distributions:

$$D_{KL}[P(e|x)|P(e|h)] = \sum_e P(e|v) \log\left(\frac{P(e|v)}{P(e|h)}\right), \quad (\text{A3})$$

Z is a normalizing factor and:

$$P(e|h) = \frac{1}{P(h)} \sum_v P(e|v)P(h|v)P(v), \quad (\text{A4})$$

Note that this is only a formal solution, which is in fact implicit. It does, however, reveal that the optimal encoder is one which results in the minimal distortion, as measured by D_{KL} , of recovery of e when using the compressed variable h in place of the original v .

The simplest way to find the solutions explicitly is to convert the self-consistent Eqs.A2 and A4 into an iterative algorithm, which can be shown to converge Ref.[9]. These are the explicit IB equations, whose shortened

form we write out in the main text as Eqs.2:

$$P(h|v) = \frac{P(h)}{Z} \exp \left(-\beta \sum_e P(e|v) \log \left[\frac{P(e|v)}{P(e|h)} \right] \right) \quad (\text{A5})$$

$$P(h) = \sum_v P(h|v)P(v) \quad (\text{A6})$$

$$P(e|h) = \sum_v P(e|v)P(v|h), \quad (\text{A7})$$

where $Z = \sum_h P(h) \exp(-\beta \sum_e P(e|v) \log(\frac{P(e|v)}{P(e|h)}))$.

In the short-hand form of the IB equations written in the main text, Eqs.2, the equation A6 was not written out explicitly, and equation A5 was slightly massaged:

$$\begin{aligned} P(h|v) &= \frac{P(h) e^{-\beta \sum_e P(e|v) \log(\frac{P(e|v)}{P(e|h)})}}{\sum_h P(h) e^{-\beta \sum_e P(e|v) \log(\frac{P(e|v)}{P(e|h)})}} \\ &= \frac{P(h) e^{\beta \sum_e P(e|v) \log(P(e|h))}}{\sum_h P(h) e^{\beta \sum_e P(e|v) \log(P(e|h))}} \end{aligned}$$

The denominator is h -independent, and the equation was given up to a proportionality constant in the main text.

The iterative IB algorithm, while useful in the theoretical investigations (and also used in the small numerical validation experiment in the appendix below) is not the only, nor necessarily the best numerical technique to solve the IB equations. For an overview of other methods we refer to Ref.[35]. Nevertheless, directly solving the IB equations for larger input distributions is generally computationally hard. Recently, an entirely different approach was developed in the context of deep learning [10]. Instead of solving the IB equations, the IB Lagrangian, or a bound on it, is taken as a cost function, and the optimal encoder is parametrised by a deep neural network and optimized using *e.g.* stochastic gradient descent. This allows to exploit the numerical efficiency of machine learning toolboxes. A similar technique is used for the optimization of $I(H; E)$ in RSMI-NE [32].

We remark here that while IB is phrased as an abstract compression theory problem, it has found applications in computational neuroscience, where the question of what is the fundamentally important information extracted from say, neuronal activity measurements, is non-trivial [50–52]. Furthermore, IB has been used in computer science problems, *e.g.* clustering analyses [53–55], but also in attempts to quantify relevant or predictive information in physics, mostly in the context of temporal correlations in non-equilibrium systems, see *e.g.* [56, 57].

Appendix B: From the transfer matrix to the reduced IB equation

In this appendix we connect the transfer matrix (TM) viewpoint to the conditional probabilities used in the main text, and derive the reduced IB equations.

In any local lattice model on the (hyper-) cylinder the partition function \mathcal{Z} can be written in terms of the trace of the transfer matrix \mathcal{T} as $\mathcal{Z} = \text{tr}(\mathcal{T}^{L_\infty})$, where L_∞ is the length of the system (here with periodic boundary conditions). For a system described by a conformal field theory (CFT), the eigenvectors and eigenvalues of \mathcal{T} correspond to operators in the CFT [43, 44]. For simplicity we consider conformal theories which have a single most relevant operator, and a corresponding microscopic lattice model with finite-range interactions. Assuming the definitions of IB quantities given in the main text, we will show that in the large buffer limit: **(i)** The description of the coarse-graining cell V and environment cell E can be reduced to only two random variables associated with certain “weak-expectation-values” r_e and r_v of most relevant primary operator on the boundary of these two regions. **(ii)** All marginal and conditional probability distributions relevant for IB can be expressed in terms of these two random variables. **(iii)** The resulting IB equations can be solved close to the first critical value of $\beta_{c,1}$, yielding the optimal encoding which, for $\beta = \beta_{c,1} + t$, amounts to tracking the above weak-expectation-values of the physically most relevant operator (which are the extracted “features”, in machine learning parlance).

Consider then a statistical-mechanical system on an infinite cylinder, with a finite cylindrical coarse-graining cell V and environment E (playing the role of the “relevance” variable) to its right, separated by a buffer, as depicted in Fig.2 in the main text. We assume a microscopic structure, for concreteness we take a square lattice. The transfer matrix \mathcal{T} , as usual, acts on the elementary slices of the cylinder, each consisting of a single (periodic) row of lattice sites. We denote by L the number of sites on the circumference of the cylinder and by L_B length of the buffer. We further denote by ∂V_R (∂E_L) the configuration of degrees of freedom on the right-most (left-most) slice of sites in V (E). These can be thought of as basis vectors of the vector space on which the transfer matrix acts, and so we shall denote them by $\langle \partial X |$ or $|\partial X \rangle = [\langle \partial X |]^T$, depending on whether \mathcal{T} acts on them from the left or from the right ($X = V$ or E as applicable).

It is well known [43, 44] that the eigenvalues λ_i and eigenvectors $|i\rangle$ of the transfer matrix have a direct relation to the CFT’s operator content. Namely, for a square lattice, $\lambda_i/\lambda_0 = e^{-\frac{2\pi}{L}\Delta_i}$, where Δ_i is the total (*i.e.* sum of the holomorphic and antiholomorphic) scaling dimension. We shall from now on normalize so that $\lambda_0 = 1$ in the limit of large circumference L . Here we consider the case where $\Delta_2 > \Delta_1 > 0$, and we take the buffer to be much larger than ratio Δ_1/L . We will see below that in this limit, the relevant degrees of freedom, or in other words the relevant random variables, in V and E are:

$$\begin{aligned} r_v &= \frac{\langle 1 | \partial V_R \rangle}{\langle 0 | \partial V_R \rangle} = \frac{\langle 0 | \phi_{\Delta_1} | \partial V_R \rangle}{\langle 0 | \partial V_R \rangle} \\ r_e &= \frac{\langle \partial E_L | 1 \rangle}{\langle \partial E_L | 0 \rangle} = \frac{\langle \partial E_L | \phi_{\Delta_1} | 0 \rangle}{\langle \partial E_L | 0 \rangle} \end{aligned} \quad (\text{B1})$$

where on both right-hand-sides we used the operator state correspondence relating the action of a primary operator ϕ_{Δ_1} on the identity state $|0\rangle$ with the state $|i\rangle$.

In a quantum mechanical setting such operator expectation values as those appearing on the r.h.s. go under the name weak-values. In cases where the primary operator turns out to be diagonal in the transfer matrix basis, the $\langle 0|\partial V_R\rangle$ and $\langle \partial E_L|0\rangle$ factors cancel and $r_{v/e}$ simply becomes the diagonal elements of that operator. As an example, for the Ising model r_v is the overall magnetization on ∂V . We note by passing that for a compact boson $\phi \in [0, R)$, $r_{v/e}$ would be the two leading electric vertex operators [45] associated with the zero transverse-momentum component of $\phi(x)$.

The conditional probabilities of the system.

Consider, for concreteness, the conditional probability $P(v|e)$, which enters the IB equations. We assume an infinite cylinder to the left of V and right of E , and work in the limit of large buffer L_B . With simple transfer matrix manipulations (using TM representation of probability distributions similar to the one below Eq.2 in the main text, and the eigendecomposition $\mathcal{T} = |0\rangle\langle 0| + \sum_i e^{-2\pi\Delta_i/L} |\Delta_i\rangle\langle \Delta_i|$), it can be written as:

$$P(v|e) = \frac{1}{N} \langle 0|\partial V_L\rangle P_{freeBC}(V) \left[\langle \partial V_R|1\rangle \langle 1|\partial E_L\rangle \lambda_1^{L_B} + \langle \partial V_R|0\rangle \langle 0|\partial E_L\rangle \lambda_0^{L_B} \right] + \mathcal{O}((\lambda_2/\lambda_0)^{L_B}) \quad (\text{B2})$$

Here ∂V_L denotes the configuration on the left boundary of V , $P_{freeBC}(V)$ is the probability of the sub-system V with free boundary conditions, which is related to the marginal probability of the sub-system V via $P(v) = N^{-1} \langle 0|\partial V_L\rangle \langle \partial V_R|0\rangle P_{freeBC}(V)$, and N is the normalization factor. In what follows we will drop the exponentially suppressed terms of order $\mathcal{O}((\lambda_2/\lambda_0)^{L_B})$. Explicitly written, $P_{freeBC}(V)$ is given by:

$$P_{freeBC}(V) \propto \langle \partial V_L|\mathcal{T}|x_2\rangle \langle x_2|\mathcal{T} \dots \mathcal{T}|\partial V_R\rangle \quad (\text{B3})$$

That is, it is simply the cumulative action of the transfer matrix on the slices of V . Note that the term in the square brackets in Eq.B2 comes from the action of the transfer matrix along the buffer, starting from the right boundary of V and ending at the left boundary of E . Taking out a factor of $P(V)$ and absorbing all normalization factors to a factor N , one obtains:

$$P(v|e) = N^{-1} P(v) \left[1 + \frac{\langle \partial V_R|1\rangle \langle 1|\partial E_L\rangle}{\langle \partial V_R|0\rangle \langle 0|\partial E_L\rangle} \left(\frac{\lambda_1}{\lambda_0} \right)^{L_B} \right] = N^{-1} P(v) [1 + \epsilon r_e r_v], \quad (\text{B4})$$

where $\epsilon = (\lambda_1/\lambda_0)^{L_B}$. The normalization N is given by:

$$\frac{1}{1 + \langle r_v \rangle r_e \epsilon}, \quad (\text{B5})$$

with:

$$\begin{aligned} \langle r_v \rangle &= \sum_v P(v) r_v = \sum_{\partial V_R} P(\partial V_R) r_v \\ &= \sum_{\partial V_R} \langle 0|\partial V_R\rangle \langle \partial V_R|0\rangle \frac{\langle 1|\partial V_R\rangle}{\langle 0|\partial V_R\rangle} = \langle 1|0\rangle = 0 \end{aligned} \quad (\text{B6})$$

The summation is over all configurations of ∂V_R . Following this we find that $\langle r_v \rangle = \langle r_e \rangle = 0$ and therefore $N = 1$. Thus, as advertised, V depends on E only through r_e i.e. $P(v|e) = P(v|r_e)$, and the same holds for $P(e|v)$. One can also show that the variances of $r_{e/v}$ obey $\langle r_e^2 \rangle = \langle r_v^2 \rangle = 1$.

The IB equation. We wish to solve the IB equation for the optimal encoder $P(h|v)$ given by [9]:

$$\begin{aligned} P(h|v) &\propto P(h) e^{\beta \sum_e P(e|v) \log(P(e|h))} \\ P(e|h) &= \sum_v P(e|v) P(v|h) \end{aligned} \quad (\text{B7})$$

Here, as before, e, h and v denote configurations of E, H, V respectively, and the symbol \propto means up to an h -independent normalization factor. In order to find the solution we next establish, generalizing the computations above, that the conditional probabilities $P(v|e), P(e|v)$ and the ‘‘decoder’’ $P(e|h)$ depend on each other only through r_v and r_e .

The reduced IB equation. The IB equation B7 for the encoder is difficult to solve, since it involves a summation over the entire configuration space of E and, furthermore, it is coupled to the equation for the decoder which involves a summation over all configurations of V . It is therefore highly beneficial to reduce these equation to ones involving only the configuration space of r_e and r_v . To this end we first note that the dependence of the encoder on V in the IB equation only appears through $P(e|v)$, and therefore can be replaced by r_v . Similarly we find:

$$P(h|r_v) \propto P(h) e^{\beta \sum_e P(e|r_v) \log(\sum_{v'} P(e|r_{v'}) P(v'|h))}. \quad (\text{B8})$$

Next we rewrite $P(e|r_{v'}) = P(e)[1 + \epsilon r_{v'} r_e]$ and expand to first order in ϵ to obtain the reduced IB equation:

$$P(h|r_v) \propto P(h) e^{\beta \epsilon^2 \langle r_e^2 \rangle r_v \langle r_v \rangle_h} = P(h) e^{\beta \epsilon^2 r_v \langle r_v \rangle_h} \quad (\text{B9})$$

where $\langle r_v \rangle_h$ is the expectation value of r_v given h , based on the joint probability $P(h, v, e) = P(v, e) P(h|v)$.

Equation B9 is the key results of this section. It shows that the optimal encoder depends on V only through r_v and in the above specific exponential manner. It changes for one CFT to another through possible values of r_v and the conditional expectation value $\langle r_v \rangle_h$. As a sanity check one finds that $P(h|r_v) = \text{const.}$ is always a solution. At small enough β it is the only one, above the IB phase transitions it is an unstable, suboptimal one (see below). Observe that Eq.B9, though simplified, is still an implicit equation, as the quantities in the exponent depend on the

left-hand side. It can, however, be solved explicitly in the vicinity of the first phase transition (which corresponds to the encoder beginning to track the first, most relevant, feature of the data).

Appendix C: Solving the reduced IB equation

Following [37] we consider the encoder at $\beta = \beta_{c,1} + t$ where $\beta_{c,1}$ marks the first breaking of the permutation symmetry of the encoder. We discuss symmetries in more detail below, here note only that below $\beta_{c,1}$ the trivial constant encoder is fully insensitive to re-labelling, or permuting, of the variables h , and without loss of generality can be written as $P(h|v) = P(h|r_v) = 1/|H|$. In order to learn *any* information whatsoever, this symmetry has to be broken. At t small enough, $\langle r_v \rangle_h$ tends to zero, and we can therefore expand:

$$\begin{aligned} P(h|r_v) &= \frac{1}{|H|} + tb_{r_v}(h) \\ \sum_h b_{r_v}(h) &= 0, \end{aligned} \quad (\text{C1})$$

where the second equation ensures proper normalization of the conditional probability. Plugging the above into the left-hand side of Eq.B9 and expanding the right-hand side in t we get, to lowest order:

$$|H|^{-1} + tb_{r_v}(h) = |H|^{-1} + \beta\epsilon^2 tr_v \sum_{v'} P(v')r_{v'}b_{r_{v'}}(h) \quad (\text{C2})$$

Examining the above one finds that a $b(h)_{r_v}$ which is constant in r_v is always a solution, for any β . This simply implies that any $P(h|r_v) = P(h)$ is a solution to the IB equation for all β . Moreover, such solutions are globally optimal before the first phase transition. This is a special case of a more general phenomenon: when some symbols h share exactly the same dependence on V , there is a manifold of equivalent solutions reflecting the freedom to re-distribute probabilities between these symbols h , in particular to assume them maximally symmetric.

In order to construct an explicit closed-form solution and to later verify it numerically, let us proceed by focusing on $|H| = 2$. Thus the compressed variable has two states and can be thought of as a single spin degree of freedom: $h = \pm 1$. The above equations now become:

$$\begin{aligned} b_{r_v}(h) &= \beta\epsilon^2 r_v \sum_{v'} P(v')r_{v'}b_{r_{v'}}(h) \\ \sum_h b_{r_v}(h) &= 0. \end{aligned} \quad (\text{C3})$$

The first equation, with a fixed h , when viewed as linear equation on the vector space spanned by the values of r_v

and equipped with an inner product weighted by $P(v)$, has two solutions: the aforementioned constant- r_v vector is a solution for any β . The vector $b_{r_v} = r_v$ is a solution for $\beta\epsilon^2 = 1$. This can be arranged into a solution for both h by taking $b_{r_v}(h) = r_v h$. The β at which this soft perturbation to the uniform encoder appears, marks the first critical β :

$$\beta_{c,1}^{-1} = \epsilon^2 + o(\epsilon^2) \quad (\text{C4})$$

where through $o(\epsilon^2)$ we re-introduced possible corrections coming from $(\lambda_{n>1}/\lambda_0)^{L_B}$, all scaling as higher powers of ϵ . These corrections exhibit a faster exponential decay as a function of L_B/L and are hence negligible for $L_B/L \gg 1$. Keeping this ratio fixed and taking $L \rightarrow \infty$, one can use the fact that:

$$\lambda_1/\lambda_0 \xrightarrow{L \rightarrow \infty} e^{-2\pi\Delta_1/L}, \quad (\text{C5})$$

where Δ_1 is the CFT scaling dimension associated with the leading primary operator. We stress though, that all the results of this section apply to any L (provided $L_B \gg L$) and do not rely on having a CFT, apart from the association between the transfer matrix eigenvalues λ_n and the scaling dimensions Δ_n in the large L limit, which is important for interpretation.

To obey normalization of $P(h|v)$, this r_v -linear vector has to be added with opposite signs to form $P(h = \pm 1|r_v)$. Examining Eq.B9 together with the assumption that $P(h)$ is constant leads to the following solution ansatz:

$$P(h = \pm 1|r_v) = \frac{e^{hm(t)r_v}}{2 \cosh(m(t)r_v)}, \quad (\text{C6})$$

where $m(t < 0) = 0$ and $m(t > 0) > 0$. To determine $m(t)$ we plug the above encoder into Eq.B9:

$$\frac{e^{hm(t)r_v}}{2 \cosh(m(t)r_v)} = \frac{1}{2} \frac{e^{\beta\epsilon^2 r_v \langle r_v \rangle_h}}{\cosh(\beta\epsilon^2 r_v |\langle r_v \rangle_h|)} \quad (\text{C7})$$

where we have used $P(h) = \text{const.}$ and the fact that $\langle r_v \rangle_h = h|\langle r_v \rangle_h|$. Clearly, if both numerators are equal, then the denominators would agree as well. Thus, we compare the logarithm of both numerators and obtain:

$$hm(t)r_v = \beta\epsilon^2 r_v \langle r_v \rangle_h. \quad (\text{C8})$$

The average on the right-hand side is evaluated with $P(r_v|h) = P(r_v)P(h|r_v)/P(h)$ and yields $\langle r_v \rangle_h = h|\langle r_v \rangle_{h=1}|$ and therefore:

$$m(t) = \beta\epsilon^2 |\langle r_v \rangle_h| \quad (\text{C9})$$

An expansion of the right-hand side in $m(t)$ up to third order yields:

$$\begin{aligned}
m(t) &= \beta\epsilon^2 \sum_{r'_v} \frac{P(r'_v)r'_v \left[1 + m(t)r'_v + \frac{1}{2}m^2(t)r'^2_v + \frac{1}{6}m^3(t)r'^3_v \right]}{1 + \frac{1}{2}m^2(t)r'^2_v} \\
&= \beta\epsilon^2 \sum_{r'_v} P(r'_v)r'_v \left[1 + m(t)r'_v + \frac{1}{2}m^2(t)r'^2_v + \frac{1}{6}m^3(t)r'^3_v \right] \left[1 - \frac{1}{2}m^2(t)r'^2_v \right] \\
&= \beta\epsilon^2 \left[m(t)\langle r^2_v \rangle + \frac{1}{6}m^3(t) (\langle r^4_v \rangle - 3\langle r^2_v \rangle^2) \right] + O(m^4)
\end{aligned} \tag{C10}$$

We thus finally obtain:

$$\begin{aligned}
(\beta - \beta_c)\epsilon^2 - \frac{\langle r^4 \rangle}{3}m(t)^2 &= 0 \\
m(t) &= \sqrt{\frac{3(\beta - \beta_c)}{\langle r^4 \rangle\beta_c}}
\end{aligned} \tag{C11}$$

We have thus an explicit analytical (and closed-form) solution for the encoder and the critical $\beta_{c,1}$ in terms of CFT quantities which can be computed in the transfer matrix formalism, in terms of TM eigenvalues and eigenvectors (either analytically, or, as is common, by numerical TM diagonalisation).

We compare this theoretical prediction for the behaviour of the IB solutions to the ones obtained numerically (*i.e.* by feeding Monte Carlo samples of the system to the IB solver as the input probability distribution, as we would do with any other data, physical or not, in a generic compression problem). As seen in Fig.3 in the main text, when presented with the data for the 2D critical Ising model on a cylinder, the analytical solution is in excellent agreement with numerics. The IB encoder $P(h|v)$ does indeed depend on ∂V only, and it assigns the value of the coarse-grained spin h based on the magnetization of the configuration of spins in ∂V , that is it depends on the physically most relevant operator, in exactly the predicted fashion.

Appendix D: IB and the physical symmetries

Here we derive several results regarding IB in the presence of physical (or model) symmetries in the data, some of which were mentioned in the main text. Along the way we also briefly review the necessary results on the internal (or structural) symmetries in IB. In this, and in the formal tools we use we follow Refs.[36, 37].

Phase transitions in IB refer to values of β where some non-analyticity appears in L_{IB} as a function of β . These transitions often come from breaking of IB structural/intrinsic symmetries although, in principle, other transitions (*e.g.* saddle-nodes [37]) are possible where the structural symmetries remain the same. The IB structural symmetries consist of permuting the classes, or elements, of H as well as re-weighting the conditional proba-

bilities: taking $P(h|v) \rightarrow P(h|v)(1 + \alpha_h)$ while maintaining the normalization constraint $\sum_h P(h|v)(1 + \alpha_h) = 1$ for all v and the value of relevant information preserved. This large continuous freedom requires some form of “gauge” fixing. We adopt the natural prescription of [36, 37], and work with encoders $P(h|v)$ which are as uniform as possible in H . Namely, we always re-weigh them so that for any symbols h, h' for which $P(h|v) = cP(h'|v) \forall v$ with $c \in \mathcal{R}_+$, the conditional probabilities are shifted to be identical. For example, this means that for $\beta < \beta_{c,1}$ the encoder is completely symmetric: $P(h|v) = 1/|H|$.

The above gauge freedom can be understood with a simple example: consider compressing information into a code with exactly three symbols $h_{1,2,3}$, with the code assigning them based on the inputs v . Imagine a code which *never* assigns the symbol h_3 to any input, and another one, in which, given an input v which should be mapped to h_2 by the previous code we instead randomly assign h_2 or h_3 with probability $p_2 + p_3 = 1$. Exactly the same information can be retrieved from both of these codes, for any p_2 , which represents the “gauge freedom”. Given the possibility of using $|H|$ symbols, we use all of them, but only a few are used nontrivially (that is $P(h|v)$ actually depends on v), the rest appear entirely randomly and independent of the inputs, with equal probability, and thus carry no information whatsoever. They can be thought of as being “unresolved”, as their probability does not depend on any feature of the data. The advantage of this formulation over simply removing unresolved symbols from the formalism, is that in IB phase transitions new symbols become “resolved”, that is the encoder starts using them nontrivially to track some additional feature of the data, and in this process their conditional probability distribution acquires dependence on v , breaking the symmetry of permuting all unresolved symbols.

Given the above re-weighting choice, IB transitions, unless fine-tuned, appear as they do in physics via first or second order symmetry-breaking transitions, where it is the permutation symmetry that is being broken at the point of transition. In IB terminology these are called subcritical and supercritical pitchfork bifurcations [36], respectively. Provided $|H|$ is taken to be $2|V| - 1$ or more, first order transitions are also excluded [58].

Despite the efforts to classify structural transitions, the

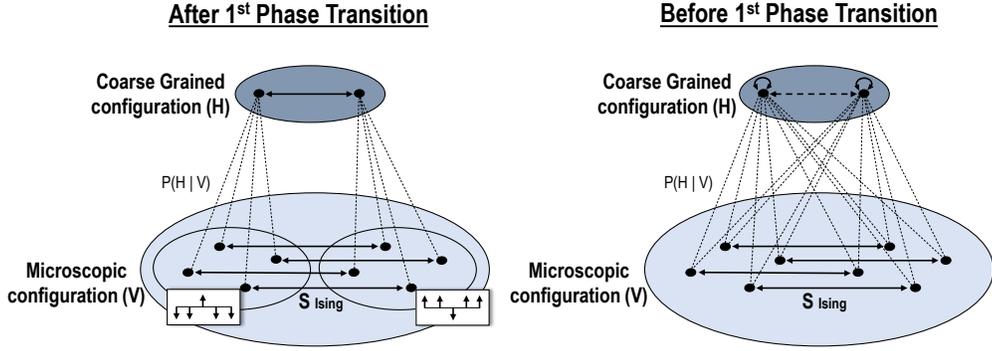


FIG. 4: **IB and the physical symmetries:** the physical system, *e.g.* an Ising model, is invariant under a symmetry. The orbits of the group action are depicted with arrows connecting the symmetry related configurations in V (here: by global Z_2 spin-flip). **Right:** Before the first IB transition the trivial encoder maps every $v \in V$ equally likely to both $h \in H$. This is consistent with a trivial action of the Z_2 symmetry on H . All symbols h are thus equivalent and connected by the action of the structural permutation symmetry (dashed line in H). **Left:** After the transition, distinct elements v would be preferentially mapped to particular symbols h . Elements v which are related by the physical (here: Ising) symmetry are mapped to h in a manner which generates a non-trivial action of the symmetry on H . Since the symbols $h \in H$ become inequivalent, the structural IB symmetry is broken.

question of how well *physical*/model symmetries present in the data are reflected or preserved in the IB transitions has, to the best of our knowledge, not been explained. Given fundamental role of symmetries in physics, this is an important point. Below we detail two contributions we make towards clarifying this issue.

Assuming that **(a)** $\beta < \beta_{c,2}$, **(b)** $|H|$ is large enough such that taking $|H| \rightarrow |H| + 1$ does not lead to a better IB solution (we are not constrained by a small code alphabet), **(c)** the first IB transition is second order (implied by [58]), and **(d)** this transition cannot be split into two separate transitions using a perturbation to $P(v, e)$ respecting the physical symmetry (*i.e.* no fine-tuning), we show that the following holds true:

$$P(e, v) = P(se, sv) \Rightarrow P_\beta(h|v) = P_\beta(\phi_s h|sv) \quad (\text{D1})$$

Here the element s of the symmetry group \mathcal{S} acts by permutation on the configurations of the system (with the action denoted by multiplication), and invariance under this symmetry is expressed by equality of their probabilities; ϕ_s is a subgroup of the permutation group on H which obeys $\phi_s \phi_{s'} = \phi_{ss'}$ (*i.e.* it is a permutation representation of the symmetry \mathcal{S} , possibly a trivial one). Eq.D1 states that under the assumptions stipulated above, the optimal IB encoder carries a representation of the physical symmetry (thus ensuring that the coarse-grained probability $P(h)$ also does). Fig.4 shows a schematic picture of the symmetry action on V and H , before and after the first IB transtion.

We also consider the case of small $|H|$, such that IB is constrained from finding an optimal solution. We construct an example with a Z_4 physical symmetry and $|H| = 2$, where Eq.D1 is violated, implying that that encoder breaks the physical symmetry.

1. IB and symmetries near $\beta_{c,1}$

For $\beta < \beta_{c,1}$, following our choice of gauge, $P(h|v) = 1/|H|$. Thus $P(h|v) = P(h|sv)$ and hence Eq. (D1) holds with $r_s = id$ being the trivial representation.

To study $\beta > \beta_{c,1}$ we follow the approach of Refs.[36, 37]. The main idea is to study the stability of the minima of the IB Lagrangian \mathcal{L}_{IB} , which correspond to the solutions for the optimal encoder $P_\beta(h|v)$. To this end the functional $\mathcal{L}(P, \lambda, \beta) = \mathcal{L}_{IB}(P, \beta) + \sum_v \lambda_v (\sum_h P(h|v) - 1)$ is introduced, with λ_v the Lagrange multipliers for the constraint enforcing normalization of the conditional probabilities P . The stable local solutions are such for which all of the eigenvalues of the Hessian $\Delta_{P,\lambda} \mathcal{L}(P, \lambda, \beta)$ have negative real parts. Here P is a vector of conditional probabilities $P(h|v)$ of dimension $|H| \cdot |V|$, and λ is a vector of λ_v of dimension $|V|$; we differentiate (twice) with respect to all of the components.

The strategy then is to consider the Hessian at $\beta = \beta_{c,1}$, analyze its kernel and how the physical symmetry manifests itself on the space spanned by the eigenvectors corresponding to eigenvalue(s) crossing from positive to negative at $\beta = \beta_{c,1}$ (the crossing eigenvalues). Following [37], we split these crossing eigenvalues into groups belonging to each element of H and show that these smaller groups generate irreducible representations of the physical symmetry, provided the transition is not fine-tuned. We then explicitly construct a globally optimal encoder at $\beta \gtrsim \beta_{c,1}$ which obeys D1. Having shown that the symmetry remains unbroken just after $\beta_{c,1}$, and given that \mathcal{L}_{IB} obeys that symmetry, by continuity we establish our claim for all $\beta < \beta_{c,2}$.

We first note that the IB Lagrangian has the property that it splits into a sum of different contributions from

FIG. 5: The structure of the Hessian $\Delta_{p,\lambda}\mathcal{L}(p, \lambda, \beta)$

distinct h , namely:

$$\begin{aligned} & \sum_h \left[\sum_v P(h|v)p(v) \log(P(h|v)) - (1 - \beta)P(h) \log(P(h)) \right. \\ & \left. - \beta \sum_{e,v} P(h|v)P(v|e)P(e) \log\left(\sum_v P(h|v)P(v|e)\right) \right] \equiv \\ & \equiv \sum_h L_h \end{aligned} \quad (\text{D2})$$

where L_h , implicitly defined above, contains only a single h . It is also symmetric under $P(h|v) \rightarrow P(h|sv)$, assuming the l.h.s. of Eq. D1. The interdependence between $P(h|v)$ with different h enters solely through the normalization requirement $\sum_h P(h|v) = 1$, which can be maintained by adding a Lagrange multiplier term $\sum_h \sum_v \lambda_v [P(h|v) - 1]$. The Hessian with respect to (w.r.t.) $P(h|v)$ and λ_v has a special structure [37]. Below $\beta_{c,1}$ it consists of $|H|$ equal block matrices B of size $|V|$ on the diagonal, along with identity matrices of size $|V|$ on the last column and row, related to the Lagrange multiplier ensuring the normalization (see Fig.5). The matrix B is simply the Hessian of L_h w.r.t. $P(h|v)$ around the trivial encoder (which is independent of h). Generically, for higher β , only the blocks B_h corresponding to unresolved symbols h are identical, their equality being the consequence of the permutation symmetry of these symbols.

An IB phase transition where new stable solutions appear, and which is second order, necessitates a non-empty set of eigenvalues of the Hessian (the aforementioned crossing eigenvalues) changing from positive to negative – at the transition the kernel of the Hessian changes. We now argue that their existence also implies the existence of a smaller set of crossing eigenvalues *within each block* B . Let us consider an eigenvector w of the Hessian, corresponding to a small eigenvalue ϵ of order $\mathcal{O}(\beta_{c,1} - \beta)$ (possibly one of many such eigenvectors). Utilising the knowledge of the structure of the full Hessian (see Fig.5) we write it in block form as $w = [w(h_1), \dots, w(h_{|H|}), \eta]^T$. The eigenvalue equation corresponds to a set of $|H| + 1$

vector equations:

$$\begin{aligned} Bw(h) + \eta &= \epsilon w(h) \\ \sum_h w(h) &= \epsilon \eta \end{aligned} \quad (\text{D3})$$

Summing the first line over all h we find $B\eta = (-|H|\epsilon^{-1} + \epsilon)\eta$, which at small enough $|\epsilon|$, and for a bounded B , has only the solution $\eta = 0$. Consequently, we find that for any h we have $Bw(h) = \epsilon w(h)$, *i.e.* the subblocks of crossing eigenvectors of the full Hessian define eigenvectors for the individual blocks, which have vanishing eigenvalues (*i.e.* which belong to their kernels). In general there may exist multiple distinct crossing eigenvectors of the full Hessian, and consequently of the blocks. We denote them by w_i .

To proceed, note that in the IB formalism the cardinality of the alphabet H is not restricted. We therefore wish to factor out the dependence of the IB Lagrangian on $|H|$. Close to the first phase transition the encoder can be written as a perturbation of the trivial (maximally symmetric) one [37]:

$$\begin{aligned} P(h|v) &= |H|^{-1} \left[1 + |H| \sum_i c_i(h) w_{i,v} \right] \\ &= |H|^{-1} \left[1 + \sum_i \tilde{c}_i(h) w_{i,v} \right] \end{aligned} \quad (\text{D4})$$

where $w_{i,v}$, viewed as vectors in the v indices for fixed h , are the crossing eigenvectors. Plugging the above r.h.s. into Eq.D2, and using $\log(|H|^{-1}[\dots]) = \log(|H|^{-1}) + \log([\dots])$, one finds that all the terms proportional to $\log(|H|^{-1})$ in L_h , vanish. In the remaining terms $|H|^{-1}$ enters explicitly, but only as an overall scaling factor. Note that Eq.C1 is a special case of the above. With this simplification, near the first transition $|H|L_h$ can be written as:

$$\tilde{L}_h \equiv |H|L_h = -\epsilon A \sum_i \tilde{c}_i(h) \tilde{c}_i(h) + \Delta \tilde{L}[\tilde{c}_1(h), \dots, \tilde{c}_N(h)] \quad (\text{D5})$$

where $\Delta \tilde{L}$ is cubic or above in $\tilde{c}_i(h)$ (consistent with second order transitions).

We proceed by analyzing the potential minima of all the terms \tilde{L}_h . At $\epsilon < 0$, all \tilde{L}_h have a single minimum at $\tilde{c}_i(h) = 0, \tilde{L}_h = 0$ and thus the encoder is uniform and trivial by Eq.D4. For a second order permutation-symmetry-breaking transitions (supercritical pitchfork transition, in the terminology of Ref.[37]), at $\epsilon \gtrsim 0$ the rescaled Lagrangian \tilde{L}_h develops $N > 0$ global minima (H-minima) of \tilde{L}_h where $\tilde{L}_h = \tilde{L}_{min} < 0$, which will later be used to construct the global minima of \mathcal{L}_{IB} . Each of these H-minima is defined by some $\tilde{c}_i(h) \neq 0$. We label the coefficients corresponding to these distinct minima by $\tilde{c}_i^{0,n}(h)$ with $n \in \{1, \dots, N\}$, which implicitly depends on ϵ . Note that around $\beta_{c,1}$ the set of H-minima (or equivalently the coefficients) are the same for different h and \tilde{L}_h , because we perturb around the fully symmetric solution.

Next we claim that any encoder which uses only the H-minima, namely one defined as $P(h|v) = |H|^{-1}[1 + \sum_i \tilde{c}_i^{0,n_h}(h)w_{i,v}]$, which is properly normalized, is globally optimal among all choices of coefficients $\tilde{c}_i(h)$ as well as sizes of $|H|$ which yield a normalized probability distribution. Indeed, $\min(\mathcal{L}_{IB}) = |H|^{-1} \sum_{h \in H} \tilde{L}_{min} = \tilde{L}_{min}$. Next due to \tilde{L}_{min} being $|H|$ independent, one has that $\tilde{L}_{min} \leq |H'|^{-1} \sum_{h \in H'} \tilde{L}_h$, since $\tilde{L}_h \geq \tilde{L}_{min}$.

We have thus shown that at the transitions, a set of H-minima appear in each \tilde{L}_h which, if composed in a way that obeys normalization, result in a globally optimal encoder. Next we discuss how the symmetry \mathcal{S} acts on the H-minima. This will be used to construct a normalized and symmetric encoder which uses only the H-minima. Given an underlying symmetry \mathcal{S} , whose elements s act as permutations of elements v of V , one finds that the blocks B obey the symmetry via $B_{v,v'} = B_{sv,sv'}$. Hence the crossing eigenvectors within each H block $w_{i,v}$ viewed as vectors in the space spanned by elements v , transform as some real representation of the symmetry, namely $w_{i,sv} = \sum_j s_{ij} w_{j,v}$ (with $\sum_j s_{ij} s_{jk} = \delta_{ik}$). It can then easily be shown that $\tilde{c}_i(h)$ transforms as $s \cdot \tilde{c}_i(h) = \sum_j s_{ji} \tilde{c}_j(h)$ and that $s \in \mathcal{S}$ acting on the n -th H-minimum $\tilde{c}_i^{0,n}(h)$ leads to an H-minimum, say $\tilde{c}_i^{0,m}(h)$, and so we write that $s(n) = m$. Consequently, we have a group action of the symmetry on the set of different H-minima. We note in passing that this action will later determine ϕ_s .

Let us next assume that s_{ij} for all $s \in \mathcal{S}$ form an irreducible representation on the crossing eigenvectors (i.e. those with eigenvalue ϵ). Notably, this is also the generic case, as one does not expect to find degenerate sets of eigenvalues beyond what is implied by symmetry. If the latter does happen it implies the transition can be split into two nearby transition using a symmetry respecting perturbation to $P(v,e)$.

Consider first the case where the representation s_{ij} is trivial. Here much of the machinery we developed is not needed since any encoder just after the transition, in particular the globally optimal one, would depend on v via $w_{i,v}$, and since the latter is invariant under the symmetry, Eq.D1 is obeyed with a trivial ϕ_s equal to the identity permutation (*id*).

We thus turn to the case of a non-trivial irreducible representation. Here we take $|H| = N$, associate each h with a specific n_h , and henceforth drop the distinction between n and h , writing $\tilde{c}_i^0(h)$ as shorthand for $\tilde{c}_i^{0,n_h}(h)$. We further split the action of \mathcal{S} on the H-minima into orbits O , each orbit understood as a set of h values corresponding to a set of H-minima. We claim that within each orbit $\sum_{h \in O} \tilde{c}_i^0(h) = 0$ for all i . Indeed, due to the orbit being closed under any $s \in \mathcal{S}$ we have :

$$\begin{aligned} \sum_{h \in O, i} \tilde{c}_i^0(h)w_{i,v} &= \sum_{h \in O, i} \tilde{c}_i^0(s(h))w_{i,v} \quad (\text{D6}) \\ &= \sum_{h \in O, i} \tilde{c}_i^0(h)w_{i,sv} \end{aligned}$$

thus if the l.h.s. is not zero, the r.h.s. implies we have found a vector w_i in the set of crossing eigenvalues, which is invariant under the action of any s . This would lead to a contradiction, as the representation was assumed to be irreducible and nontrivial.

Finally, we write our globally optimal, normalized, and symmetric encoder just after $\beta_{c,1}$:

$$P(h|v) = |H|^{-1}[1 + \sum_i \tilde{c}_i^0(h)w_{i,v}] \quad (\text{D7})$$

where the dependence on β enters implicitly via $\tilde{c}_i^0(h)$. It can be verified that it obeys Eq. (D1) with ϕ_s being the action of \mathcal{S} on the H-minima: $s(h) = h'$. Furthermore, it is normalized since:

$$\sum_h P(h|v) = 1 + |H|^{-1} \sum_O \sum_{h \in O} \sum_i \tilde{c}_i^0(h)w_{i,v} = 1 \quad (\text{D8})$$

Lastly, as this encoder uses only H-minima it is globally optimal.

Had we taken $|H| < N$ such a solution would not be possible, and in such circumstance IB can potentially break physical symmetries. We provide such an example below (though at high β). Note though, that it is always possible to increase $|H|$ until it no longer improves \mathcal{L}_{IB} , thereby avoiding these constrained settings.

We conjecture that IB, unless constrained or tuned in an adversarial fashion, respects physical/model symmetries for all values of β . In particular we have never encountered a numerical example where this does not hold for large enough $|H|$. Notably, for unrestricted $|H|$ there is no obvious competition between learning an optimal encoder and maintaining the symmetry which could encourage such physical/model symmetry breaking.

2. Potential breaking of the physical symmetry in constrained IB at large β

The results of the previous section imply that for large enough H the encoder would carry a representation of the physical symmetry. Given the fact that in practical implementations the size of alphabet H would often be fixed, and the symmetry group may possibly not be fully known, it is interesting to ask what happens in the case $|H| < N$.

The question is whether, given possible sizes of permutation representations of the symmetry and some fixed $|H| < N$, more information is necessarily retained when the encoder generates the action of one of those representation on H , or not. To make intuitive why breaking the symmetry could be favorable, consider the following example. Let $n_1 > |H| > 1$ be the size of the smallest nontrivial permutation representation of the physical symmetry, i.e. we are given more symbols than needed to “fit” the action of the trivial representation on H (which maps everything to one symbol), and not enough to fit a nontrivial one. It seems natural that not using the available H symbols in the encoder is wasteful.

Here we provide an explicit example where Eq.D1 is violated at $|H| < N$ in the limit $\beta \rightarrow \infty$. Such examples are easier to construct if $|H|$ is incompatible with the dimension of any irreducible representation of \mathcal{S} , as mentioned above. Here, however, we give a less trivial example with a Z_4 symmetry and $|H| = 2$, which intuitively at least could fit a two-dimensional representation of the symmetry.

For $\beta \rightarrow \infty$, the IB Lagrangian simplifies to maximizing $I(H; E)$. Intuitively, in this limit the solution should always be a deterministic encoder (*i.e.* one for which $P(h|v) \in \{0, 1\}$) and defines a bona fide function $f: v \rightarrow h$. This in fact follows from convex optimization arguments, as shown in [35]. Next we write the mutual information as difference of entropies:

$$I(H; E) = S(E) - S(E|H), \quad (\text{D9})$$

where:

$$S(E|H) = \sum_h P(h)S(E|h) \quad (\text{D10})$$

$$S(E|h) = - \sum_e P(e|h) \log(P(e|h)). \quad (\text{D11})$$

Note that due to the deterministic nature of $P(h|v)$:

$$P(v|h) = P(h|v)P(v)/P(h) = P(h)^{-1}P(v)\delta_{f(v),h}.$$

Therefore:

$$\begin{aligned} P(e|h) &= \sum_v P(e|v)P(v|h) \quad (\text{D12}) \\ &= \sum_{v|f(v)=h} \frac{P(v)}{P(h)} P(e|v) \equiv P(e|V(h)), \end{aligned}$$

where by $V(h)$ we denoted the pre-image of h under mapping f . By the above equations we thus seek to group the elements v into h -clusters, such that the entropy of E averaged over the different clusters is minimized. Formally, we minimize:

$$\begin{aligned} \sum_h P(h)S(E|h) &= \\ &= - \sum_h P(h) \sum_e P(e|V(h)) \log(P(e|V(h))) \\ &\equiv \sum_h S(E|V(h))P(h) \quad (\text{D13}) \end{aligned}$$

Let now $V, E = \{0, 1, 2, 3\}, H = \{0, 1\}, S = Z_4$ and let

$$P(e|v) = P(v|e) = \frac{1}{3}[\delta_{e,v} + \delta_{e,(v+1)\%4} + \delta_{e,(v+2)\%4}],$$

where $\%$ denotes the modulo operation, and let $P(v) = P(e) = 1/4$ for all e, v . We want to find the two disjoint sets V_1 and V_2 (which are mapped to distinct h) which minimize $S(E|V_1)P(h=0) + S(E|V_2)P(h=1)$, *i.e.* equation D13. Up to an action of Z_4 there are only two distinct partitions of V which are maintained by the

action of the only nontrivial subgroup of Z_4 , *i.e.* Z_2 , and thus which could be compatible with an encoder producing the action of Z_2 on H : $V_1 = \{0, 2\}, V_2 = \{1, 3\}$ and $V_1 = \{0, 1\}, V_2 = \{2, 3\}$. For these two partitions of V we have that:

$$\begin{aligned} S(E|V_1)P(h=0) + S(E|V_2)P(h=1) &= \\ &= \frac{1}{2}[S(E|V_1) + S(E|V_2)]. \end{aligned}$$

We compare these to the following non-symmetric choice $V_1 = \{0\}, V_2 = \{1, 2, 3\}$. All together these exhaust all symmetry-distinct choices of the sets. For the first symmetric choice, one has $P(e|V_1)$ being equal to e drawn uniformly from $\{0, 0, 1, 2, 2, 3\}$ leading to $S(E|V_{1/2}) = 1.3296$. For V_1 of the second symmetric choice, we get e drawn from $\{0, 1, 1, 2, 2, 3\}$ leading to the same value of $S(E|V_{1/2})$. For V_1 of the non-symmetric choice, we get e drawn uniformly from $\{0, 1, 2\}$ leading to an entropy $S(E|\{0\}) = 1.0986$. Last for V_2 of the non-symmetric choice, we get e drawn uniformly from $\{0, 0, 1, 1, 2, 2, 3, 3, 3\}$ leading to $S(E|\{1, 2, 3\}) = 1.3689$. As $1.3296 > 0.25 \cdot 1.0986 + 0.75 \cdot 1.3689 = 1.3013$ we find that the non-symmetric choice is optimal.

While the $\beta \rightarrow \infty$ example can be evaluated on the back of an envelope, we verified numerically that for this example distribution the symmetry breaking holds for β all the way down to the first IB phase transition. Interestingly, taking a larger alphabet $|H| > 2$ improves the IB Lagrangian value \mathcal{L}_{IB} of the best encoder at $\beta \geq \beta_{c,1}$, until at $|H| \geq 4$, *i.e.* at $|H| \geq N$ it reaches its optimal value and remains unchanged by further increasing $|H|$.

The lessons we take from the above example are that **(a)** for H of insufficient size, smaller than the size of the (relevant) symmetry group, the symmetry can be broken by the encoder and **(b)** in order to ensure this does not happen one should choose the minimal $|H|$ at which, for a fixed $\beta \geq \beta_{c,1}$, the value of the IB Lagrangian $\mathcal{L}_{IB}(P_\beta(h|v))$ reaches its optimal value. Recall also that $\beta_{c,1}$ can be obtained in a model-agnostic way by studying the kernel of the B block of the Hessian.

Appendix E: Extracting symmetries from a numerically obtained encoder

Here we discuss the possibility of using the symmetry-maintaining properties of the optimal encoder, *viz.* Eq.D1 to extract the physically relevant symmetries from the data.

Let us examine the situation where $P(v, e)$ possesses an unknown symmetry $s \in \mathcal{S}$, such that $P(sv, se) = P(v, e)$, and an unknown action ϕ_s on the h variables. The symmetry may be unknown since it involves a complicated combination of microscopic degrees of freedom, or because \mathcal{S} is part of a much larger symmetry group from which we wish to sift out the most relevant subgroup. In addition, even if \mathcal{S} is known, its action on h would depend on how it combines the relevant variables and

may potentially need to be extracted numerically. We discuss how both \mathcal{S} , ϕ_s , and their actions on v and h , respectively, can in principle be identified.

As an instructive example, consider the Ising model used in the main text. We deliberately split, however, each Ising spin σ_i into to product of two auxiliary spins $\sigma_i = \tau_{i1}\tau_{i2}$. The energy of the system remains the same (in terms of the original spins). By construction, this model has a huge amount of spurious symmetry: an extra Z_2 symmetry per each site i , given by $I_{i,1}I_{i,2}$ (where $I_{i,*}$ is a spin-flip operator for the variable τ_{i*}). This symmetry does not flip the Ising spin σ_i and so bears no influence on the long-range properties of the system, – the physical Ising symmetry is artificially obscured in the model phrased in τ microscopic variables.

Assume now we are given the solution to the IB problem $P_\beta(h|\tau)$ with $h = \pm 1$ for $\beta > \beta_{c,1}$, which depends on the Ising magnetization on the boundary $\sum_{i \in \partial V_R} \tau_{i1}\tau_{i2}$ (as per arguments in the main text and in the appendices above). Since for this encoder:

$$P_\beta(h|\tau) = P_\beta(h|I_{i,1}I_{i,2} \cdot \tau) \quad (\text{E1})$$

it generates a *trivial* representation of all the extra Z_2 symmetries. Since it couples to the physical magnetization, however, it generates a *faithful* Z_2 representation of the “hidden” Ising symmetry (which we can choose as $s = \prod_{i \in V} I_{i1}$), namely $P(h|\tau) = P(-h|s\tau)$. We would like to provide a prescription for identifying the relevant symmetry \mathcal{S} and its action on h .

The first step is to obtain an estimate of the $P_\beta(h) = \sum_v P_\beta(h|v)P(v)$, by numerically sampling v . Using the results on the general structure of the optimal encoder, it suffices to take $v \in \partial V_R$. The symmetry of the encoder $P(h|v) = P(\phi_s h|sv)$ implies the symmetry in the coarse-grained variables: $P_\beta(h) = P_\beta(\phi_s h)$. This allows to group the equiprobable elements h into sets, whose elements are potentially related by an action of ϕ_s for some $s \in \mathcal{S}$. These sets, forming a partition of H , are putative orbits of ϕ_s . In the above example the set simply contains both $h = \pm 1$.

Consider now configurations $v_+ \in \partial V_R$ for which $P(+1|v_+)$ is non-zero (generally this yields all configurations) and similarly so for v_- . Next reconstruct the action of the symmetry element s on V by demanding that: **1.** s maps the set of all v_+ to the set of all v_- , **2.** $P(+1|v_+) = P(-1|sv_+)$, **3.** s applies the same onsite permutation across all sites (*i.e.* s is spatially homogeneous). Notably, committing to such homogeneous s is allowed provided we focus on global symmetries. Conveniently, it also makes the space of potential permutations much smaller, in the sense that it is independent of the number of sites. Following this s can be found using a brute force scan.

In our example, provided β is finite, the set of all v_+ is simply ∂V_R and $s_1 = \prod_{i \in \partial V_R} I_{i1}$, $s_2 = \prod_{i \in \partial V_R} I_{i2}$. It is then easy to infer from the symmetry requirement $P(h|x) = P(\phi_s h|sx)$ that s_1 (and s_2) have a Z_2 action on H whereas $s_1 s_2$ has a trivial action. We have thus

exemplified how to identify qualitative information from the numerically obtained encoder: the size of the representation of the relevant symmetry was deduced from the number of equally probable symbols h , and its action on v was given by the above permutations s obeying the symmetry constraints. Using the action on v , ϕ_s acting on H can also be deduced.

Had we considered a larger set of symmetric $h \in H$'s (say $|H| = 3$, $h = \{a, b, c\}$, $P(h) = 1/3$), we would have similarly looked for a set of v_a obeying $P(a|v_a) \neq 0$, and homogeneous permutations s mapping v_a to v_b (or v_c), obeying $P(a|v_a) = P(b|sv_a)$ (or $P(a|v_a) = P(c|sv_a)$). This procedure can in principle be generalized to groups of arbitrary size, we leave however the question of how to do it efficiently to future investigations.

Observe also that focusing on $v \in \partial V_R$ rather than all of V in the above procedure comes at no loss of generality, since as we seek a spatially homogeneous action of the symmetry, the symmetry action on ∂V_R implies its action on V . Thus no information is lost and computational resources, related to matching v_+ and v_- are used more efficiently. Had we considered $v \in V$, then, from the perspective of the encoder coupling to the edge ∂V , the freedom of flipping $\tau \in V/\partial V$ behaves as a symmetry (with a trivial representation on H). However examining the probability $p(v) = \sum_e p(v, e)$ would reveal that it is not a true physical symmetry of V .

Appendix F: Numerical Experiments

To validate our prediction of the the critical temperature $\beta_{c,1}$ and the dependence of the optimal encoder after the IB phase transition on physical quantities we performed a numerical experiment. The test system was the 2D Ising model at criticality. The system was put on a cylinder of three sites' circumference, and the transfer matrix eigenvectors and eigenvalues were obtained by exact diagonalisation to obtain the numerical value of $\beta_{c,1}$ from Eq.C4 and the encoder from Eqs.C6 and C10.

This was compared with the numerical solutions to the IB equations presented with the probability distribution of the model configurations. To obtain these solutions we used the simple Iterative IB Algorithm (iIB)[59]. The input variables of the iIB algorithm are: $P(E, V)$, cardinality $|H|$, tradeoff β , the initial guess for $P(H|V)$, and ϵ , which is a convergence parameter. The outputs are $P(H|V)$, $P(E|H)$ and $P(H)$.

For a given set of input variables, the iIB algorithm iterates between the three IB Eqs.A5–A7. On every iteration the first IB equation updates the encoder $P(H|V)$ from the previous iteration. Then, $P(E|H)$ and $P(H)$ are updated according to the remaining equations. The iterations stop when the update on $P(H|V)$ becomes negligible, *i.e.* if after n iterations $JS[P_n(H|V)|P_{n-1}(H|V)] < \epsilon$, where JS is the Jensen-Shannon divergence.

To produce Fig.3 of the main-text, we applied the iIB algorithm on a range of β values. Starting from $\beta = 0$

and a maximally symmetric trivial encoder, we increased β up to maximal value greater than $\beta_{c,1}$, determined through the numerical analysis of the Hessian (see below). For every β , in order to prevent the algorithm from getting stuck in a non-optimal local minimum of the Lagrangian, a small random noise was added to the initial guess $P(H|V)$ (the outcome of the optimization for the previous value of β). The iIB algorithm was then applied on the inputs until convergence. The same steps were then repeated for the next value of β .

For this small test system the input $P(E, V)$ distribution was calculated explicitly, using the transfer matrix method. We used a setup where $|V| = 2 \times 3, |E| = 1 \times 3, L_B = 9$ and $L = 3$. We set the convergence parameter to be $\epsilon = 1e-14$, and the random noise added to the encoder was of the order of $1e-6$.

Just above $\beta_{c,1}$ the iIB algorithm tends to stay around the saddle point given by the trivial uniform $P(H|V)$. This behavior continues until at some higher value of β the algorithm converges to the optimal solution of the IB equations through a discontinuity in the IB curve. To fix such numerical artifacts, after each discontinuity we applied the algorithm again, this time by scanning β

backwards. We initialized the backward scan using the parameters and variables from the forward scan at a β value located after the discontinuity point.

To make sure that our backward scan yielded the optimal solution, we compared its IB Lagrangian values with those of the forward scan solution, confirming the former were indeed smaller or equal. Another indication of the instability of the forward scan iIB solution just after $\beta_{c,1}$ was that the eigenvalue of the Hessian which crossed 0 at $\beta_{c,1}$ stayed negative after $\beta_{c,1}$, where the equivalent eigenvalue at the backward scan was positive.

Finally, we obtain the value of the $146.33999 < \beta_{c,1}^{IB} < 146.34999$, which agrees with the $\beta_{c,1} = 146.34458$ from the analytical formula Eq.C4, and the encoder in Fig.3.

We emphasize again, that the analytical results on the IB transition we derived *do not* require the limit of large cylinder circumference L , as long as L_B is sufficiently larger than L . The prediction is for the IB transition in terms of the transfer matrix eigenvalues and eigenvectors, regardless of the circumference. This is the prediction we verify numerically. In the limit of $L \rightarrow \infty$ the extracted quantities can additionally be related to the CFT scaling operators, per the classical results on transfer matrices and CFTs [44].