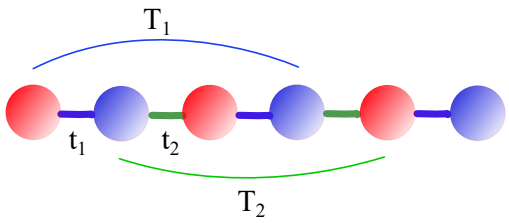


FIG. 1. Schematic set-up of a 1D SSH system. Here t_1 and t_2 are nearest-neighbor hoppings while T_1 and T_2 are second nearest-neighbor hoppings.

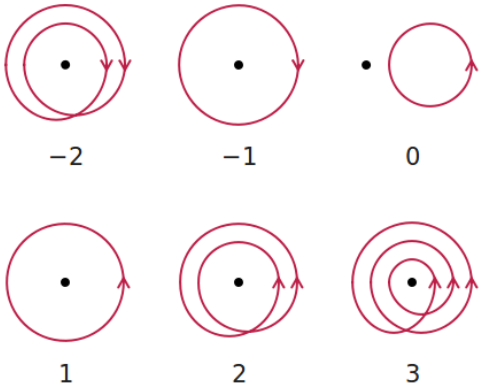


FIG. 2. Winding number. The winding number of a closed, oriented curve with respect to a reference point is a topological invariant that counts how many times the curve winds around the point. Picture credits: Jim Belk, public domain.

SUPPLEMENTARY MATERIAL

The SSH model

The SSH model [1] describes the movement of free electrons along a dimerized chain whose basic units consist of two distinct atoms. This movement, usually called "hopping" in the literature, can be made either between atoms in a unit cell or between unit cells, and the allowed hopping rules for a given system completely determine its Hamiltonian. This is because the kinetic energies of the electrons are parameterized by a vector of real numbers $\mathbf{t}$ that also encodes hopping terms, thus allowing for a compact mathematical description of a Hamiltonian in terms of creation/annihilation operators as

$$\hat{H}(\mathbf{t}) = \mathbf{c}^\dagger H(\mathbf{t}) \mathbf{c} \tag{1}$$

where the column vector

$$\mathbf{c} = \left(c_1^A, c_1^B, \cdots, c_{\frac{N}{2}}^A, c_{\frac{N}{2}}^B\right)^T$$

contains annihilation operators $c_p^{A(B)}$ that erase electrons at atom A (B) and lattice site p and similarly the row vector

$$\mathbf{c}^\dagger = \left(c_1^{A\dagger}, c_1^{B\dagger}, \cdots, c_{\frac{N}{2}}^{A\dagger}, c_{\frac{N}{2}}^{B\dagger}\right)$$

contains creation operators $c_p^{A(B)\dagger}$ that produce electrons at atom A (B) and lattice site p. Please note that N is twice the number of unit cells in the chain and therefore an even integer.

The convenience of equation (1) is that all information about a system such as its eigenstates and eigenenergies can be recovered from the $N \times N$ matrix $H(\mathbf{t})$. We can thus think of the vectors $\mathbf{t}$ in parameter space as very compact representations of SSH models: each point in $\mathbf{t}$-space can be mapped to a $N \times N$ matrix $H(\mathbf{t})$ whose eigenvectors and eigenvalues can then be computed, as is usually done in quantum mechanics. As an example, a general matrix $H(\mathbf{t})$

describing a SSH system with hoppings between nearest and second nearest neighbors is given by

$$H(t_1,t_2,T_1,T_2)=\begin{pmatrix} 0 & t_1 & 0 & T_1 & 0 & \cdots \\ t_1 & 0 & t_2 & 0 & T_2 & \cdots \\ 0 & t_2 & 0 & t_1 & 0 & \cdots \\ T_1 & 0 & t_1 & 0 & t_2 & \cdots \\ 0 & T_2 & 0 & t_2 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}_{N\times N}. \tag{2}$$

Knowing the vector $\mathbf{t}=(t_1,t_2,T_1,T_2)$ corresponding to a particular system described by the Hamiltonian in equation (2) should suffice to compute any of its physical properties, including its topological phase. Figure 1 depicts a SSH system described by equation (2).

The reason why topological materials garnered so much interest in recent years is that their physical properties are topologically robust. This means that these properties are stable under continuous (i.e., adiabatic) mathematical operations performed on the system's underlying wave functions. This topological robustness is expressed theoretically in terms of a topological invariant that characterizes different phases of a system. In the particular case of the SSH model, the topological invariant used to classify the topological phases is the winding number.

The winding number is a topological property of any closed, oriented curve that measures how many times the curve winds around a point that does not belong to itself. It can be any integer and is usually chosen to be positive when the curve winds in counterclockwise motion with respect to the reference point (equivalently, when a closed, oriented curve winds in clockwise motion around a reference point its winding number is negative). Figure 2 shows several closed, oriented curves and their winding numbers computed with respect to a given point.

An interesting property of topological invariants like the winding number is that they are a global feature of geometric objects: for each of the curves in figure 2 for example the winding number is a property of the whole curve that cannot be defined locally for each of its points. This fundamental characteristic of topological invariants makes the study of topological phases of matter from local lattice data a very challenging task.

For SSH systems with translational symmetry like the finite systems with periodic boundary conditions investigated in the article, the winding number is usually computed in wavevector space via

$$W=\frac{1}{4\pi i}\int_0^{2\pi} dk Tr(\sigma_3 H(k)^{-1}\partial_k H(k)), \tag{3}$$

where $H(k)$ is the kernel in wavevector space of a Hamiltonian $\hat{H}(\mathbf{t})$ and σ_3 is the chiral operator. Equation (3) can be evaluated for several Hamiltonians by varying the parameter $\mathbf{t}$, resulting in phase diagrams in parameter space like the ones shown in figure 1 in the article.

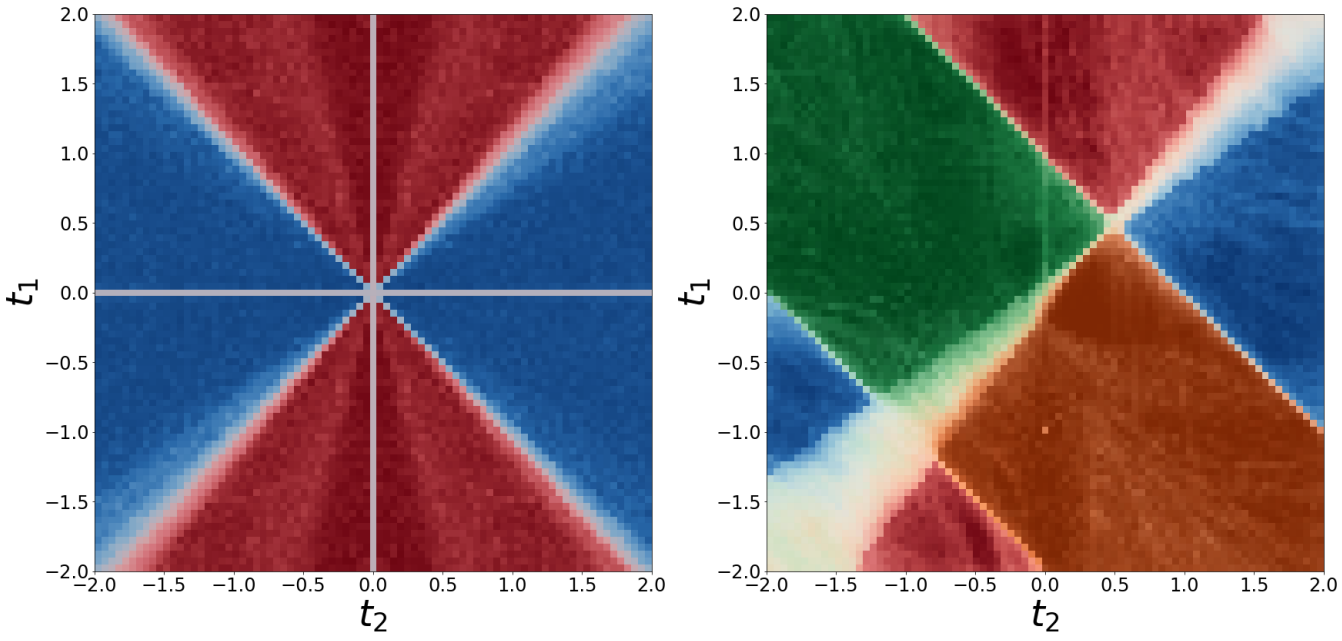
Learning topological phases from real space data

The main motivations for developing a data-driven approach based on real space are that wavevector space computations such as eq. (3) are only possible for systems with translational symmetry, which many physical systems of current interest (e.g. disordered systems in condensed matter) do not have. Moreover, since real space and wavevector space eigenvectors are related by Fourier transforms, the latter are essentially delocalized and therefore so is any information recovered from them. We argue here that a data-driven approach based on real space offers a viable alternative to wavevector space computations that addresses these shortcomings.

Our reasoning goes as follows. First, while the topological invariants that characterize distinct topological phases are usually computed in wavevector space, the topological properties of a Hamiltonian are the same regardless of the basis in Hilbert space used to represent it. Thus, information on the topological phase of a Hamiltonian should still be available when it is represented in real lattice space. Second, even though the topological properties of a Hamiltonian are global, meaning that in general they cannot be said to be localized at a particular lattice site, in parameter space topology is indeed a local property: knowing the vector $\mathbf{t}$ associated with a Hamiltonian completely determines its topological phase.

In a data-driven approach, locality is often exploited by means of a local constancy hypothesis [2]. Mathematically, this policy prescribes the value of a function $W(\mathbf{t}')$ at points where it is unknown in a vicinity of a data point $\mathbf{t}$ as approximately equal to its known value $W(\mathbf{t})$,

$$W(\mathbf{t}+\boldsymbol{\delta})\approx W(\mathbf{t}). \tag{4}$$



(a) (b)

FIG. 3. Phase diagrams learned from most relevant lattice sites. (a) Phase diagram learned by a combination of decision trees with eigenvector ensembling in experiment 1 using only lattice sites $S_1' = (0, 50, 51, 99)$. The accuracy achieved for Hamiltonians in the test set using only the four lattice sites in S_1' was approximately 95%. (b) Phase diagram learned by a combination of random forests with eigenvector ensembling in experiment 2 using only lattice sites $S_2' = (0, 1, 3, 48, 50, 51, 96, 98, 99)$. The accuracy achieved for Hamiltonians in the test set using only the nine lattice sites in S_2' was approximately 74%.

That such a policy will be successful in classifying topological phases in parameter space can be visualized in figure 1 in the article where we draw phase diagrams of SSH models with first-neighbor (figure 1(a)) and first- and second-neighbor (figure 1(b)) hoppings. In figure 1(a) for example, it is clear that knowing a particular Hamiltonian $H(t_1, t_2)$ with winding number $W = 0$ (that is, in one of the red regions) means that there is a small neighborhood around (t_1, t_2) in which all Hamiltonians belong to the same topological phase. Were we able to collect data on the topological phases of several Hamiltonians in parameter space, the problem of learning phase boundaries in a supervised setting would reduce to a standard problem of curve estimation which could be tackled with conventional machine learning algorithms.

It does not immediately follow, however, that the same strategy will be successful in real space. Indeed, at first sight it may appear that a local constancy policy should be able to easily exploit locality in real space through the diagonalization maps $v^{(j,l)} : \mathbb{R}^h \to \mathbb{R}^N$,

$$\mathbf{t} = (t_1, .., t_h) \to \left(v^{(j,1)}(t_1, ..., t_h), ..., v^{(j,N)}(t_1, ..., t_h)\right) = \mathbf{v}^{(j)}(\mathbf{t}) \tag{5}$$

where $j = 1, ..., N$ and $\mathbf{v}^{(j)}(\mathbf{t})$ is an eigenvector of the Hamiltonian $H(\mathbf{t})$. The trouble with this reasoning is that it disregards the high dimensionality of real space, i.e., the fact that $h \ll N$.

This fact is well illustrated by the numerical experiments discussed in the article. Although it may seem from figures 2(b) and 3(b) (from the article) in 2D parameter space that we have used a large number of data points for this learning task, it is important to note that in the numerical experiments the decision trees have taken as inputs 100D eigenvectors in real lattice space. In such high-dimensional spaces, the data should be much sparser.

The difficulty arising from machine learning problems in high-dimensional spaces is commonly referred to as the curse of dimensionality [3]. It essentially expresses the fact that the amount of data needed to ensure a machine learning algorithm will generalize well out of its training set grows exponentially with the dimensionality of feature space.

These apparently conflicting facets of our learning problem are harmonized by the manifold hypothesis [4, 5]: even though the eigenvectors exist in a high-dimensional space (100D in the numerical experiments), they are actually

much lower-dimensional surfaces (2D in the numerical experiments) embedded in this space. Furthermore, the different classes in our problem correspond to different submanifolds as can easily be seen in parameter space (this is often referred to as the manifold hypothesis for classification [6]). As we have demonstrated in the article, only a small fraction of the 2D surfaces (i.e., eigenvector lattice coordinates $v^{(j,l)}(t_1, t_2)$) were needed to retrieve the 2D parameter space phase diagrams from the 100D real space data (see figure 3). In this sense, key topological information can be said to be localized on few lattice sites.

The eigenvector ensembling algorithm

We describe here in detail each step of the eigenvector ensembling algorithm.

1) **Generating Hamiltonians and winding numbers:** we start generating a number of paremeterized Hamiltonians $H(\mathbf{t})$ in real space and their corresponding winding numbers $W(\mathbf{t})$, where $\mathbf{t} = (t_1, t_2, ..., t_h)$ is a vector of h hopping parameters (in the simplest case of the SSH model $h = 2$). These Hamiltonians are $N \times N$ matrices, where N is twice the number of unit cells in the chain.

2) **Creating training, validation and test sets:** we split our set of parameterized Hamiltonians and winding numbers into training, validation and test sets, as is usualy done in machine learning. More explicitly, assume our parameter $\mathbf{t}$ takes on the values $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_n$ corresponding to the Hamiltonian-winding number pairs (H_1, W_1), ..., (H_n, W_n). We partition the set $\{(H_i,\, W_i)|\ i = 1, ..., n\}$ in three disjoint subsets: the training set, the validation set and the test set.

3) **Training on eigenvectors in real space:** since each Hamiltonian H_i is represented by an $N \times N$ matrix, each one will generate N eigenvectors $\mathbf{v}_i^{(1)}, \mathbf{v}_i^{(2)}, ..., \mathbf{v}_i^{(N)}$. Our supervised learning algorithm of choice will take as inputs the real space eigenvectors $\mathbf{v}_i^{(j)}$ of each Hamiltonian H_i in the training set and be trained to learn the winding number W_i of their parent Hamiltonian H_i. Therefore, our dataset will consist of eigenvector-winding number pairs $(\mathbf{v}_i^{(j)}, W_i)$.

4) **Eigenvector ensembling:** in order to predict the phase of a system described by a particular Hamiltonian we need to take into account how each of its eigenvectors were classified. This amounts to performing ensemble learning on the eigenvectors of each Hamiltonian. In this work we estimate the phase probabilities for each Hamiltonian as the fraction of its eigenvectors that were classified in each phase.

5) **Bootstrapping:** We refine the phase probabilities for each Hamiltonian using a bootstrapping procedure, i.e., we repeat steps (1)-(4) n_{exp} times, at each round sampling randomly a new training set from our grid in $\mathbf{t}$-space. The final estimated probabilities are then arrived at by averaging the probabilities obtained in each experiment.

In this work we implemented the eigenvector ensembling algorithm with Python's Scikit tools.

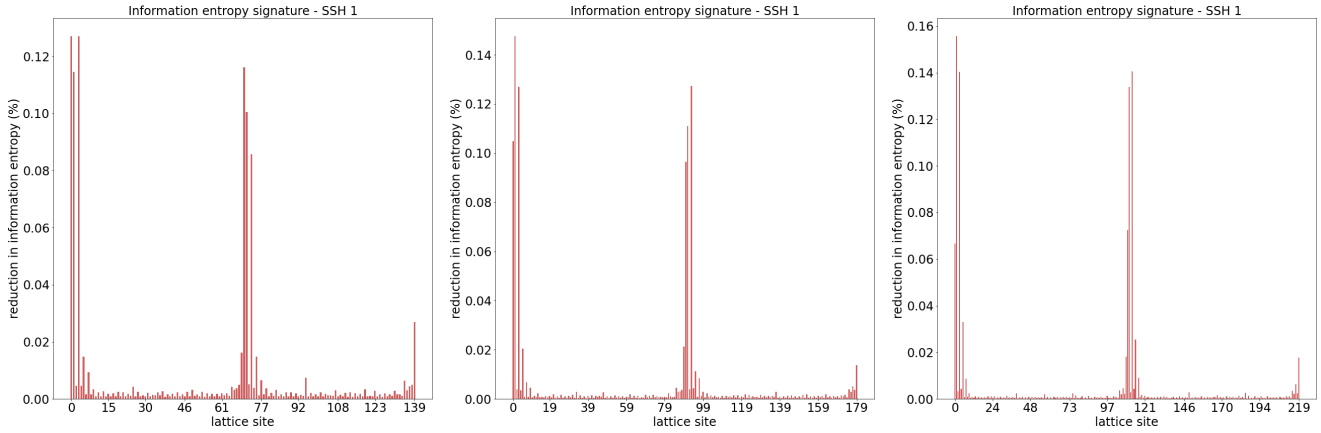
Information entropy signatures in the macroscopic limit

A natural question that comes to mind regarding the information entropy signatures presented here is whether these signals are artifacts of the numerical procedure used to generate them.

The eigenvector ensembling algorithm used in this work contains steps for generating data (step 1), sampling training data (step 2) and training a supervised learning algorithm on the sampled training data (step 3). Each of these steps can generate misleading artifacts that do not represent real properties of the physical systems we are investigating.

Artifacts resulting from randomization in the eigenvector ensembling algorithm can be traced to sampling (step 2) or any random components in the supervised learning algorithm used (step 3). As an example, random forests allocate subsets of features stochastically to each of its decision trees, thus generating a randomization effect. The bootstrapping (step 5) is designed to remove artifacts originating from randomization in the eigenvector ensembling algorithm.

There are still artifacts that might arise from the hyperparameters used to generate the data. These hyperparameters include lattice size and grid specifications (i.e., choices regarding the discretization of parameter space). Such artifacts can only be controlled for by bootstrapping over different hyperparameter settings, which may be computationally



(a) (b) (c)

FIG. 4. Information entropy signatures obtained for experiment 1 with higher lattice sizes. (a) Information entropy signature for 70 unit cells. (b) Information entropy signature for 90 unit cells. (c) Information entropy signature for 110 unit cells.

prohibitive. Intuitively, the hyperparameter we identify as likely having the most noticeable effect in the information entropy signatures is lattice size.

All results presented in the article were obtained for lattices with 50 unit cells. Each cell contains two different atoms, thus leading to 100×100 Hamiltonian matrices and their corresponding eigenvectors in $\mathbb{R}^{100}$. Here we present the information entropy signatures obtained for lattices with 70, 90 and 110 unit cells for both experiments (figures 4 and 5). These signals were generated in exactly the same way as the information entropy signatures obtained for 50 unit cells (i.e., after bootstrapping $n_{exp} = 100$ times and averaging lattice site relevances across all iterations).

It is clear from figures 4 and 5 that the patterns seen with 50 unit cells (figures 6(a) and 6(b)) are stable across higher lattice sizes, with finer details emerging in the signals as the lattice size increases.

Figures 4 and 5 suggest that the information entropy signatures presented here can be viewed as entropy mass functions along the lattices. In the limit of an infinite chain, the cumulative entropy distribution $F_S(x)$ associated with a Shannon information entropy signal will be given by the integral of an entropy density function $\rho_S(x)$,

$$F_S(x) = \int_0^x \rho_S(x')dx' \tag{6}$$

where $x \in$ [0,1] is a spatial coordinate along the lattice (being strict, our use of periodic boundary conditions implies that the coordinate x should be defined in the quotient space $[0,1]/R$, where R is the equivalence relation in [0,1] defining the unit circle S^1, i.e. $x\ R\ x'$ if and only if $x = x'$ or (x,x') $\in \{(0,1), (1,0)\}$. For open boundary conditions, the spatial coordinate x is defined in the closed interval [0,1]). In this continuum limit, an information entropy signature will be given by the entropy density function $\rho_S(x)$. Figures 6 and 7 show the cumulative entropy distributions corresponding to the information entropy signatures in figures 4 and 5.

The results presented in this article suggest that the macroscopic limit $\rho_S(x)$ of the information entropy signatures may have strong theoretical interest and pave new roads to the investigation of topological materials.

[1] J. K. Asbóth, L. Oroszlány, and A. Pályi, "A short course on topological insulators," *Lecture notes in physics*, vol. 919, 2016.

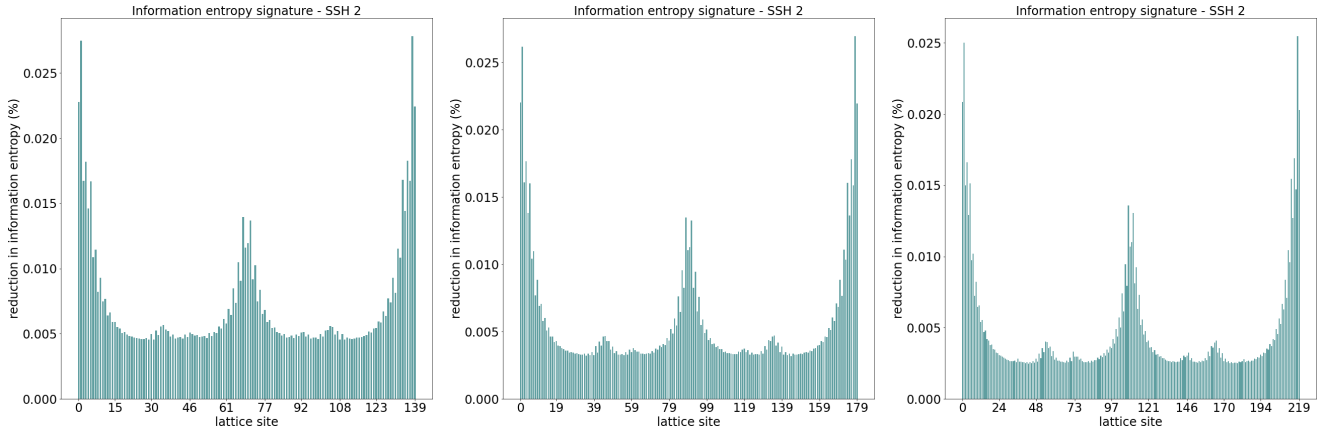
[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, Cambridge, MA, USA, 2016.

[3] C. M. Bishop, *Pattern recognition and machine learning*. Springer New York, NY, USA, 2006.

[4] L. Cayton, "Algorithms for manifold learning," *Univ. of California at San Diego Tech. Rep*, vol. 12, no. 1-17, p. 1, 2005.

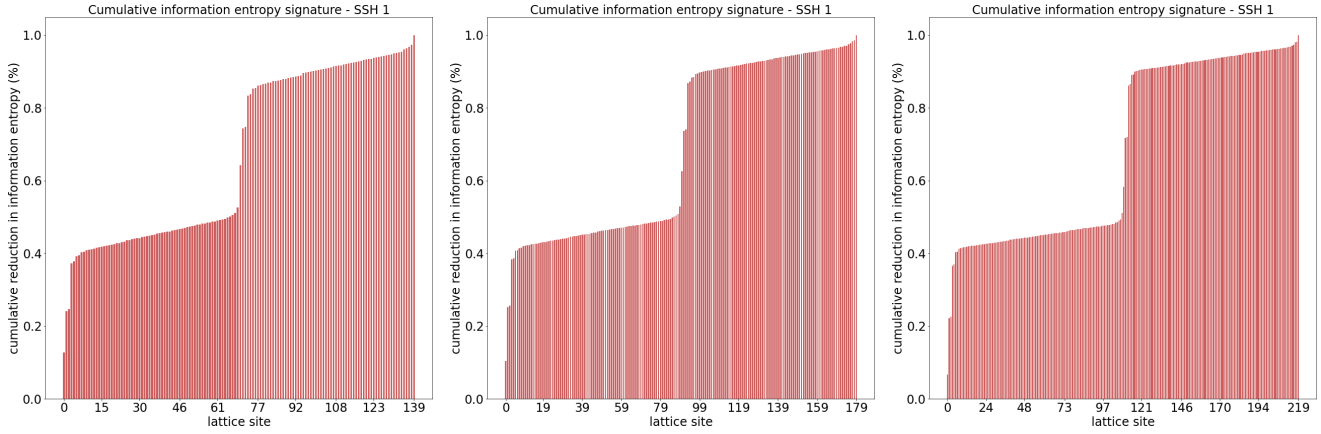
[5] H. Narayanan and S. Mitter, "Sample complexity of testing the manifold hypothesis," in *Advances in Neural Information Processing Systems*, pp. 1786–1794, 2010.

[6] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller, "The manifold tangent classifier," in *Advances in Neural Information Processing Systems*, pp. 2294–2302, 2011.



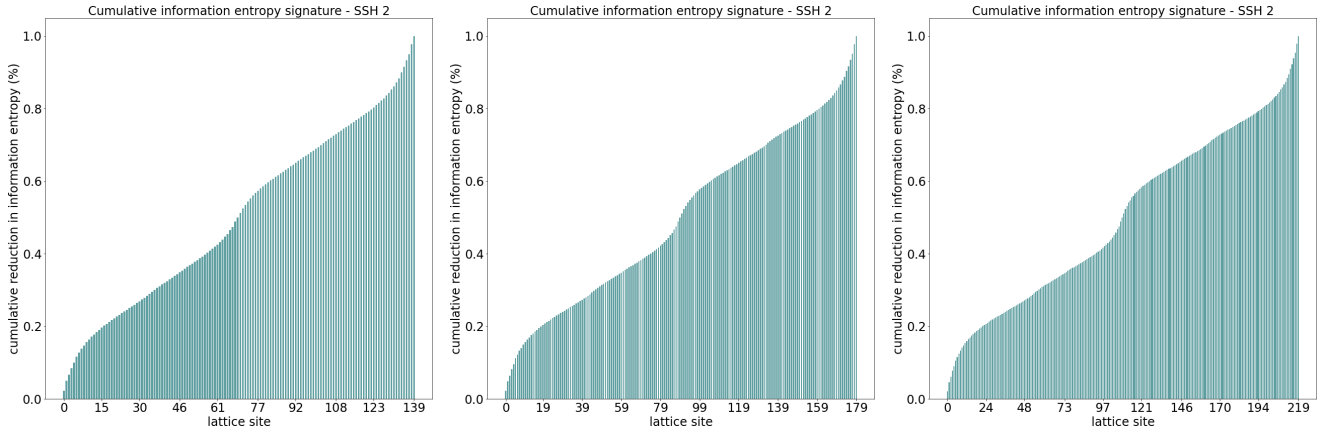
(a) (b) (c)

FIG. 5. Information entropy signatures obtained for experiment 2 with higher lattice sizes. (a) Information entropy signature for 70 unit cells. (b) Information entropy signature for 90 unit cells. (c) Information entropy signature for 110 unit cells.



(a) (b) (c)

FIG. 6. Cumulative entropy distributions for higher lattice sizes in experiment 1. (a) Cumulative entropy distribution for 70 unit cells. (b) Cumulative entropy distribution for 90 unit cells. (c) Cumulative entropy distribution for 110 unit cells. The two visible leaps in the cumulative entropy distributions shown above occur at regions corresponding to the lattice sites S_1 in the case with 50 unit cells.



(a) (b) (c)

FIG. 7. Cumulative entropy distributions for higher lattice sizes in experiment 2. (a) Cumulative entropy distributions for 70 unit cells. (b) Cumulative entropy distributions for 90 unit cells. (c) Cumulative entropy distributions for 110 unit cells. Here the leaps in the cumulative entropy distributions are less pronounced than in experiment 1, but still noticeable. The three visible leaps occur at regions corresponding to the lattice sites S_2 in the case of 50 unit cells.