Machine learning topological phases in real space

N. L. Holanda[*]

Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, United Kingdom and
Centro Brasileiro de Pesquisas Físicas,
Rua Dr. Xavier Sigaud, 150 - Urca,
22290-180, Rio de Janeiro, RJ, Brazil

M. A. S. Griffith[†]

Centro Brasileiro de Pesquisas Físicas,
Rua Dr. Xavier Sigaud, 150 - Urca,
22290-180, Rio de Janeiro, RJ, Brazil and
Departamento de Cincias Naturais, Universidade Federal de São João Del Rei,
Praa Dom Helvécio 74, 36301-160, São João Del Rei, MG, Brazil

(Dated: May 7, 2020)

We develop a supervised machine learning algorithm that is able to learn topological phases for finite condensed matter systems from bulk data in real lattice space. The algorithm employs diagonalization in real space together with any supervised learning algorithm to learn topological phases through an eigenvector ensembling procedure. We combine our algorithm with decision trees and random forests to successfully recover topological phase diagrams of Su-Schrieffer-Heeger (SSH) models from bulk lattice data in real space and show how the Shannon information entropy of ensembles of lattice eigenvectors can be used to retrieve a signal detailing how topological information is distributed in the bulk. The discovery of Shannon information entropy signals associated with topological phase transitions from the analysis of data from several thousand SSH systems illustrates how model explainability in machine learning can advance the research of exotic quantum materials with properties that may power future technological applications such as qubit engineering for quantum computing.

[*] linneuholanda@gmail.com, linneu@cbpf.br
[†] griffithphys@gmail.com

INTRODUCTION

The quest for innovative materials that harness exotic quantum properties has lured physicists into the realm of topological insulators and topological states of matter [?]. These materials feature previously unthought-of traits like bulk insulation coupled with metallic conductance at the surface and the splitting of currents according to spin orientation. Adding to that, these properties are protected by non-trivial topology that renders them robust to many sources of perturbation like thermal noise. Such characteristics make them promising candidates to being the cornerstone of 21st century technologies like spintronics and quantum computing.

These new topological states of matter have been studied in several contexts in condensed matter physics including superconductors [?]−[?], ultracold atoms [?]−[?], photonic crystals [?]−[?], photonic quantum walks [?]−[?] and Weyl semimetals [? ?]. Among these, the Su-Schrieffer-Heeger (SSH) model [?] has attracted particular theoretical interest due to its simplicity and generality.

The SSH model is the simplest tight-binding model that exhibits a topological phase transition. As such, it can be viewed as the *Drosophila* of the field, providing a simple framework for testing new techniques. The model can be expressed in terms of creation and annihilation operators by the Hamiltonian

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

and describes e.g. the hopping of electrons along a one-dimensional chain comprising two atoms per unit cell (a brief discussion of the SSH model and its topological properties can be found in the section **The SSH model** in the Supplementary Material). The SSH model has found several interesting applications in the modelling of diverse systems with non-trivial topology like optical lattices [?], polymeric materials [?] and topological mechanisms [? ?].

Many recent papers have explored the possibility of treating the general problem of determining phase transition boundaries of physical systems as machine learning tasks [?]−[?]. In the particular case of topological phase transitions, the usual approach for supervised learning is to generate a data set $\big(H_1(k), W_1\big)$, ..., $\big(H_n(k), W_n\big)$ whose inputs are representations of Hamiltonians in wavevector space $H_i(k)$ and targets are their corresponding topological invariants W_i (for the SSH model the topological invariant is the winding number). Our paper extends this task to the case of learning topological phase diagrams from input data in real space. Strikingly, we find that information localized on a few lattice sites in the bulk is sufficient to predict with high accuracy which topological phase a particular Hamiltonian belongs to.

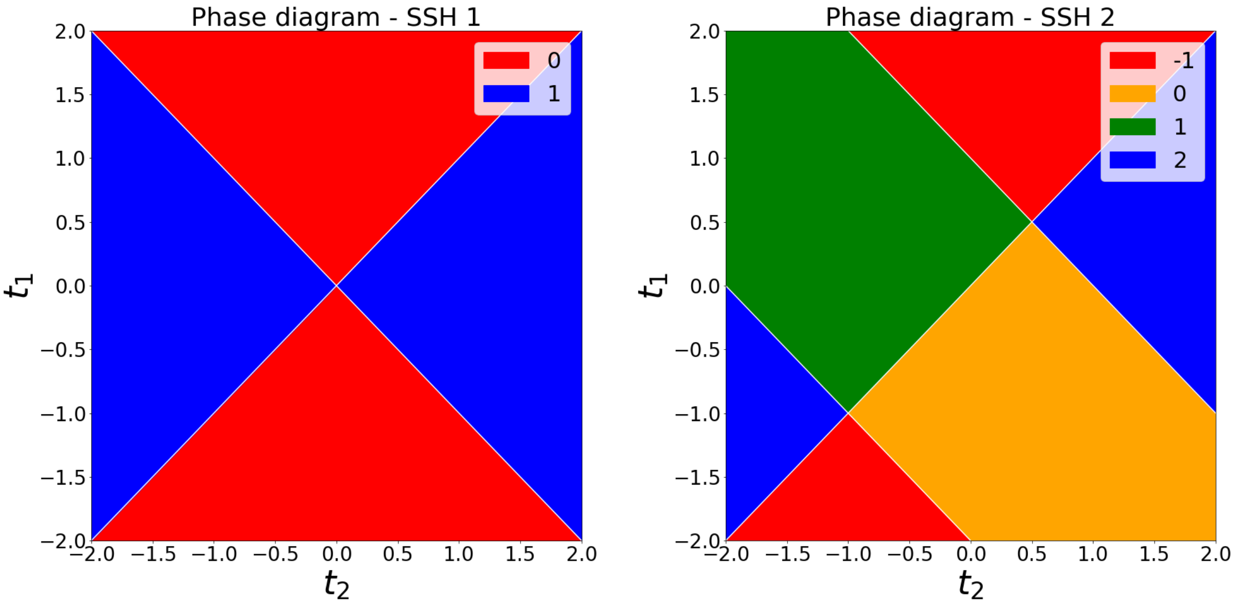
To investigate topological phases of matter in real space we have designed a novel supervised learning algorithm (here called eigenvector ensembling algorithm) tailored for the task of learning phase transition boundaries from local features. The algorithm is based on eigenvector decomposition and eigenvector ensembling and therefore will require minimal changes to be applicable to a broader class of data-driven physics problems. We demonstrate its effectiveness by combining it with decision trees and random forests to recover the topological phase diagrams of SSH systems from local coordinates of eigenstates in real space.

The advantage of using decision tree-based algorithms to learn topological phases from local eigenvector data is that their use of entropy-based cost functions (such as Shannon information entropy or Gini impurity) furnishes them with an intrinsinc model explanaibility tool that summarizes how important each feature was to learn the desired pattern in the data. This makes it much easier to trace the localization of relevant information along the features of a data set. Here we use the Shannon information entropy of ensembles of real space eigenvectors to recover a signal quantifying the amount of topological information available from each lattice site. This is a highly non-trivial proposition since the topological phase of a system is a global property of the whole system emerging from complex interactions between its components, and therefore even defining a local topological signal is a daunting theoretical task. To our knowledge this is the first time that a signal describing the localization of topological information in the bulk of topological condensed matter systems is presented in the literature.

The Shannon information entropy signals presented for the first time in this work provide a clear illustration of how model explainability in machine learning can guide new discoveries in condensed matter and quantum materials physics, since the existence of these signals was established by analyzing data from several thousand SSH systems which, taken individually, could not have provided any concrete hint of their existence.

NUMERICAL EXPERIMENTS

The eigenvector ensembling algorithm consists of five steps: 1) Generating Hamiltonians in real space and their corresponding winding numbers; 2) Creating training, validation and test sets; 3) Training on real space eigenvectors of Hamiltonians in the training set; 4) Eigenvector ensembling and 5) Bootstrapping. A detailed description of the



(a)

(b)

FIG. 1. Phase diagrams in parameter space. a) SSH model with first-neighbor hoppings t_1 and t_2. The (red) regions with winding number $W = 0$ are trivial, while the (blue) regions with winding number $W = 1$ are topologically non-trivial. b) SSH model with first (t_1 and t_2) and second (T_1 and T_2) nearest-neighbor hoppings. In this article we set $t_1 = t_2 = 1$ and renamed the variables $T_1 \to t_1$, $T_2 \to t_2$ for convenience. The (orange) region with winding number $W = 0$ is trivial while the others with winding numbers $W =$ -1, $W = 1$ and $W = 2$ (red, green and blue respectively) are topologically non-trivial.

algorithm is found in the section **The eigenvector ensembling algorithm** in the Supplementary Material. Here we present results of the numerical experiments we performed with it. We start with the results from the simplest case, the SSH model with nearest-neighbor hopping (here called SSH 1, figure 1(a)), then we analize the SSH model with first and second nearest-neighbor hoppings (here called SSH 2, figure 1(b)).

In each experiment our grid consisted of 6561 Hamiltonians uniformly distributed in the closed square $[-2, 2] \times [-2, 2]$ in the t_1-t_2 plane in parameter space. The goal in each experiment is to recover the corresponding phase diagram in 2D (two-dimensional) parameter space, figures 1(a) and 1(b), from local lattice data in the much higher-dimensional real space (100D - in both experiments lattices have 50 unit cells, yielding 100×100 Hamiltonian matrices).

This task is particularly hard near phase transition boundaries, where numerical computation of winding numbers become less stable. For this reason, when sampling the training set we only consider those Hamiltonians in the grid whose numerically computed winding numbers lie in a range $\epsilon = 0.01$ around the allowed winding number values. Therefore, a good performance metric is the accuracy measured at those Hamiltonians near phase transitions that are never used for training, and thus we assign them to the test set. The remaining Hamiltonians in the grid are split into training and validation sets as detailed in the subsections below.

When generating the Hamiltonians we applied periodic boundary conditions to eliminate border effects. This should make recovering a topological signal from local eigenvector coordinates even harder, since in this case the translational symmetry of the systems should allow for no obvious way to distinguish between unit cells. The choice of periodic boundary conditions also implies that the information recovered from real space data comes from the bulk of the topological systems considered and therefore provides strong evidence for the existence of topological signatures in the bulk of such systems.

Figures 2 and 3 respectively illustrate single iterations of experiments 1 and 2 as seen from parameter space. The accuracy statistics presented in the following subsections and probability heatmaps shown in figures 4 and 5 were obtained after bootstrapping each experiment $n_{exp} = 100$ times. The recovered probability heatmaps faithfully portray the phase diagrams in figure 1, with clear phase transition lines appearing in the regions of highest uncertainty.

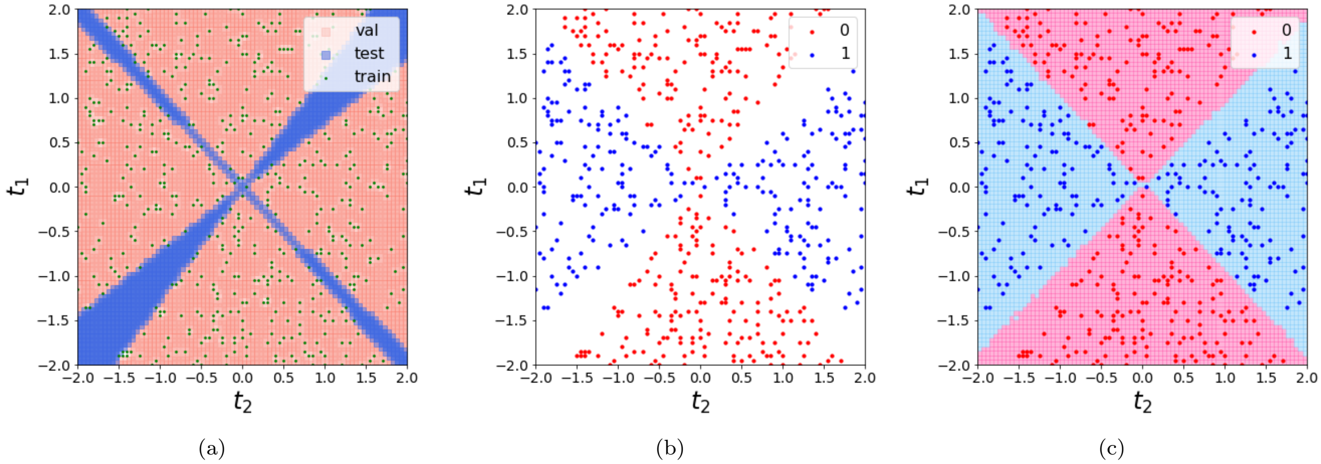


FIG. 2. Visualization of a single iteration of experiment 1 as seen from 2D parameter space. (a) Train/validation/test split. (b) Distribution of winding numbers in the training set. (c) Phase diagram learned from real space lattice data by a combination of decision tree and eigenvector ensembling.

Experiment 1: Learning a first-neighbor hopping SSH model with decision trees

Our test set in this experiment contained 1005 Hamiltonians (approx. 15.3% of all data). Of the remaining 5556 Hamiltonians, 556 were randomly assigned to the training set (approx. 8.5%) and 5000 (approx. 76.2%) were used to compute validation scores at each iteration. These proportions between training and validation sets are such that approximately 10% of Hamiltonians from outside of the test set were used for training at each iteration. The composition of the train + validation set for this experiment was 50.8% of Hamiltonians with winding number $W =$ 0 and 49.2% with winding number $W = 1$. The composition of the test set was 44.8% of Hamiltonians with winding number $W = 0$ and 55.2% with winding number $W = 1$. Our algorithm of choice for this experiment was a simple decision tree model [**?**].

The bootstrap allows us to collect several statistics to evaluate performance. In particular, we report mean accuracies on training eigenvectors (98.2%), validation eigenvectors (96.4%) and test eigenvectors (78.8%). Eigenvector ensembling substantially improved mean accuracies for Hamiltonians. These were 100% for training Hamiltonians, 100% for validation Hamiltonians and 99.1% for test Hamiltonians.

The probability heatmaps and phase diagram learned by the combination of decision trees with eigenvector ensembling used in experiment 1 are shown in figure 4.

Experiment 2: Learning a first- and second-neighbor hoppings SSH model with random forests

This task is considerably more difficult than the previous one due to the higher number of classes and the fact that some of the labels encompass disconnected regions. For this reason, instead of using a single decision tree, we upgraded our model to a random forest [**?**] with 25 decision trees. Our data set consisted of 1040 (15.8%) test Hamiltonians. The remaining Hamiltonians are randomly split in half between training and validation sets at each iteration, giving 2761 (42.1%) training Hamiltonians and 2760 (42.1%) validation Hamiltonians. The distribution of winding numbers for the Hamiltonians in the train + validation set for this experiment was $W =$ -1 (17.9%), $W = 0$ (32.5%), $W = 1$ (32.3%) and $W = 2$ (17.3%). The distribution of winding numbers for the Hamiltonians in the test set was $W =$ -1 (36.3%), $W = 0$ (11.1%), $W = 1$ (12.7%) and $W = 2$ (39.9%).

Mean accuracies across 100 repetitions of experiment 2 were 99.9% for training eigenvectors, 97.1% for validation eigenvectors and 66.4% for test eigenvectors. Mean accuracies resulting from eigenvector ensembling were 100% for training Hamiltonians, 99.7% for validation Hamiltonians and 88.2% for test Hamiltonians. The large accuracy gain achieved by eigenvector ensembling in the test set (going from 66.4% eigenvector accuracy to 88.2% Hamiltonian accuracy) attests to its power.

The probability heatmaps and phase diagram learned by the combination of random forests with eigenvector ensembling used in experiment 2 are shown in figure 5.

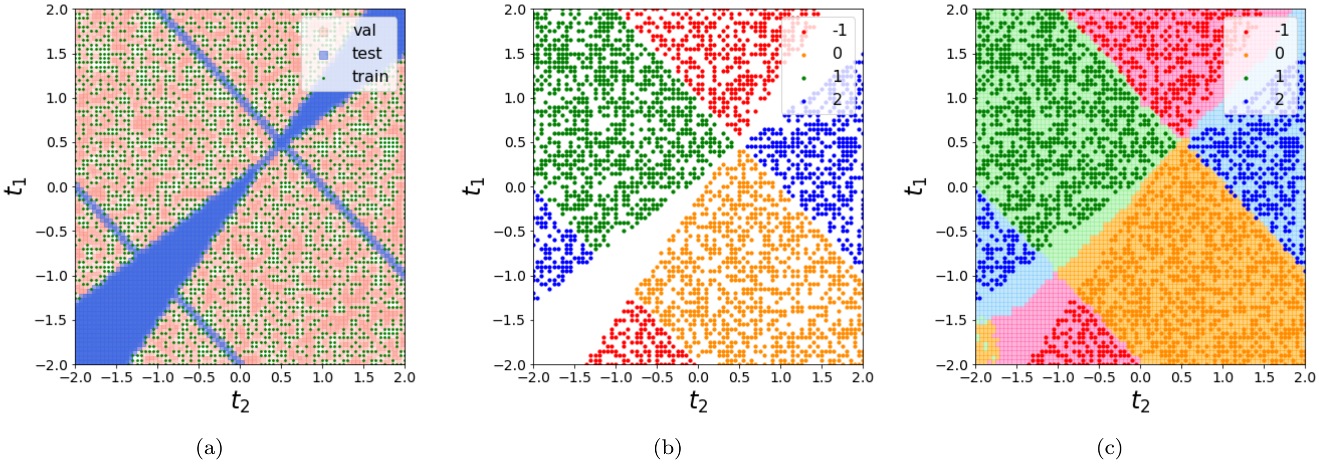


FIG. 3. Visualization of a single iteration of experiment 2 as seen from 2D parameter space. (a) Train/validation/test split. (b) Distribution of winding numbers in the training set. (c) Phase diagram learned from real space lattice data by a combination of random forest and eigenvector ensembling.

INFORMATION ENTROPY SIGNATURES

We now analyze how the algorithm was able to recover a global property of the Hamiltonians (their topological phase) from bulk local features (real space eigenvector coordinates on each lattice site). Alongside the fact that decision trees and random forests are very easy to train and visualize, the other reason that led us to test our algorithm with them was that they allow us to check which features (and thus which lattice sites) were most informative in training.

The (normalized) relevance of a feature is given by how much it reduces a loss function (e.g. Shannon information entropy or Gini impurity [**?**]). By averaging normalized relevances as measured by reduction in the information entropy of ensembles of real space eigenvectors across $n_{exp} = 100$ iterations of both experiment 1 and experiment 2 we recovered Shannon entropy signals that reveal which lattice sites were consistently more relevant in learning topological phases from data in real space for each experiment. These signals are the information entropy signatures of each topological phase transition.

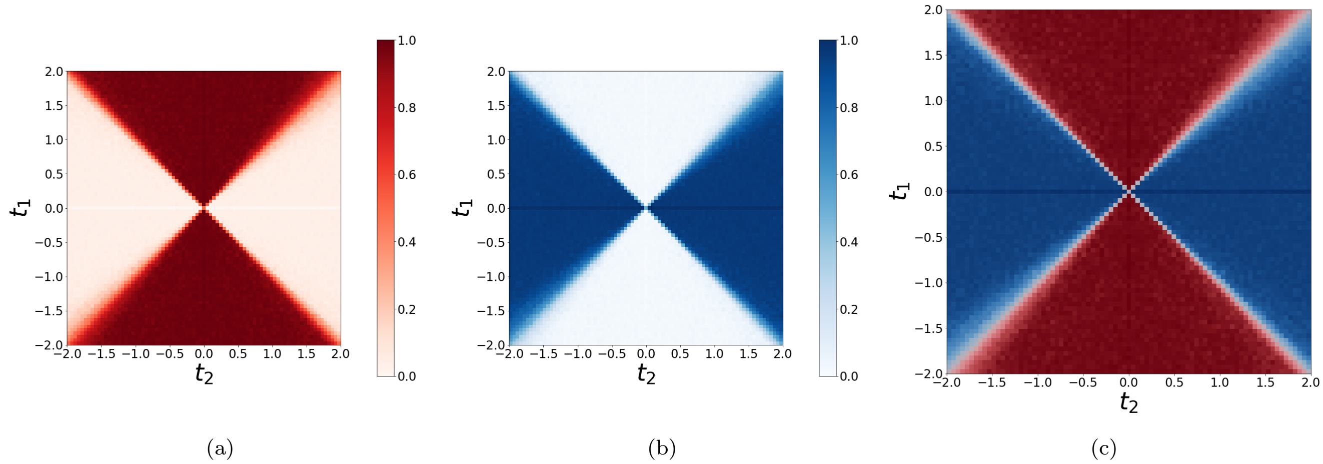
The bar plots in figure 6 show how informative each lattice site was in learning topological phases for experiments 1 and 2. They represent the information entropy signatures along the lattices in each SSH system. For experiment 1, only six lattice sites $S_1 = (0, 1, 3, 50, 51, 53)$ corresponding to the two sharp peaks seen in figure 6(a) contributed approximately 70% of total reduction in Shannon entropy. Similarly, approximately 30% of total reduction in the Shannon information entropy of eigenvector data from experiment 2 was achieved by eighteen lattice sites $S_2 = (0,$ $1, 2, 3, 4, 5, 46, 48, 49, 50, 51, 53, 94, 95, 96, 97, 98, 99)$ distributed along the three peaks in figure 6(b).

Each of the information entropy signatures shown in figure 6 captures a general pattern that persists regardless of the length of the lattice (i.e, number of unit cells) used to compute them. They are not, therefore, artifacts of particular choices of hyperparameters used to run the eigenvector ensembling algorithm. We present the information entropy signatures for longer lattices in the section **Information entropy signatures in the macroscopic limit** in the Supplementary Material.

To see if learning the topological phases can be achieved efficiently by employing simpler models we reran experiments 1 and 2 using only a small subset of most relevant lattice sites. In our rerun of experiment 1 using only lattice sites $S_1' = (0, 50, 51, 99)$ (which contributed approximately 45 % of total reduction in Shannon information entropy in experiment 1), mean accuracies were 97.0% for training eigenvectors, 91.5% for validation eigenvectors and 72.8% for test eigenvectors. Mean accuracies obtained from eigenvector ensembling were 99.1% for training Hamiltonians, 99.5% for validation Hamiltonians and 94.5% for test Hamiltonians.

Mean accuracies for our rerun of experiment 2 using only lattice sites $S_2' = (0, 1, 3, 48, 50, 51, 96, 98, 99)$ (which contributed approximately 20 % of total reduction in Shannon information entropy in experiment 2) were 99.9% for training eigenvectors, 87.7% for validation eigenvectors and 47.3% for test eigenvectors. Eigenvector ensembling yields mean accuracies of 100% for training Hamiltonians, 99.5% for validation Hamiltonians and 74.5% for test Hamiltonians.

These results demonstrate that learning topological phases from local real space data in the bulk is still possible even for small subsets of lattice sites. We refer the reader to the section **Learning topological phases from real**



(a)

(b)

(c)

FIG. 4. Probability heatmaps learned by a combination of decision trees with eigenvector ensembling from bulk real space eigenvector data in experiment 1. Heatmaps were averaged across all 100 iterations of the experiment. (a) Probability heatmap showing the probability that a Hamiltonian in the grid has winding number equal to 0. (b) Probability heatmap showing the probability that a Hamiltonian in the grid has winding number equal to 1. (c) The phase diagram resulting from heatmaps (a) and (b).

space data in the Supplementary Material for a discussion of how this is possible.

DISCUSSION

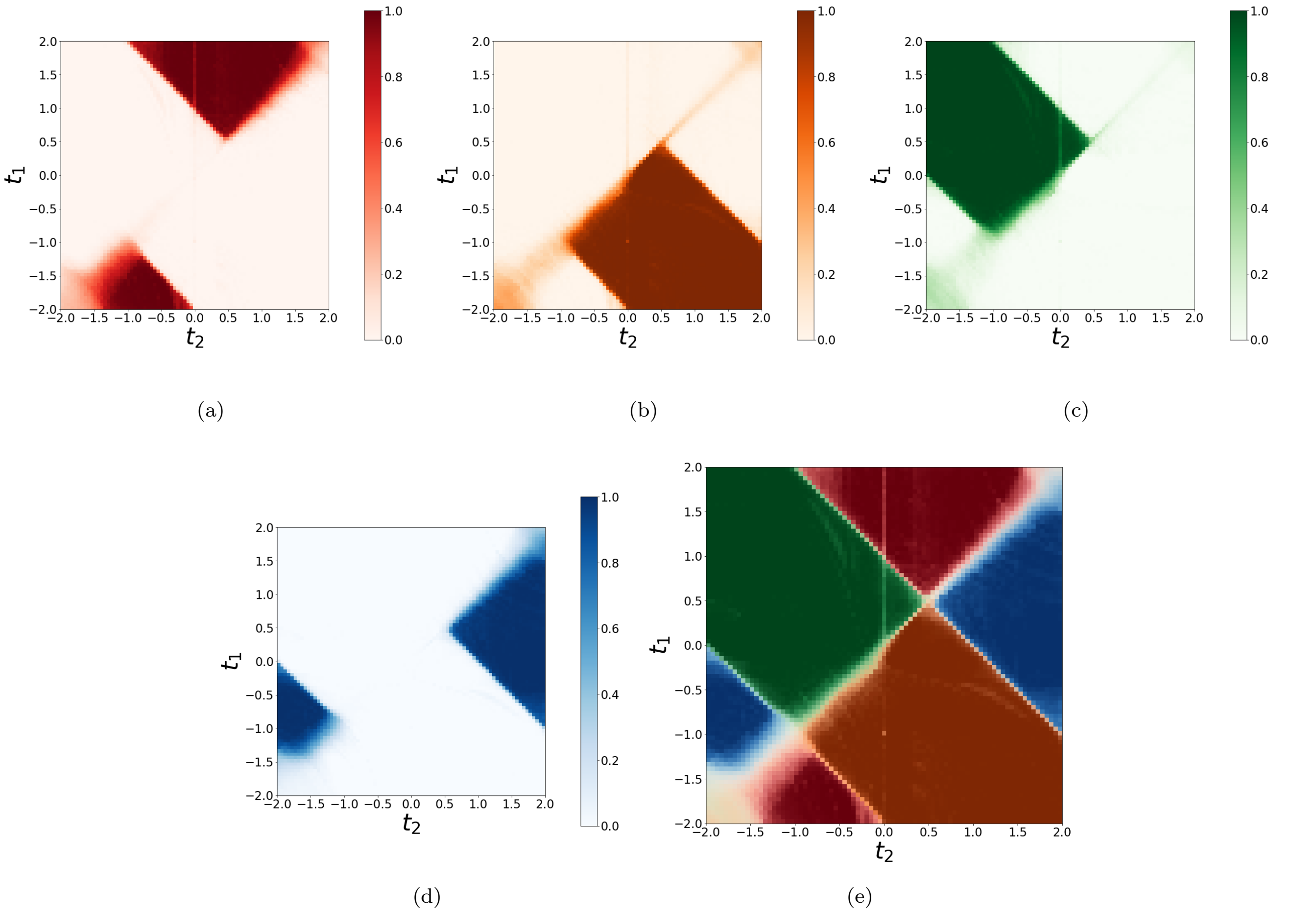
Given the increasing complexity of systems studied in condensed matter physics and the rising demand for materials with exotic and robust properties to power future technological progress, it is only expected that data-driven approaches to physics will grow in demand. Our work represents a step in this direction, as we have implemented a data-driven approach to the search for new topological materials bypassing the use of wavevector space data.

The development of data-driven methods based on real space lattice data will be particularly relevant to the study of disordered systems in condensed matter. Such systems usually break translational symmetry and therefore are not amenable to wavevector space methods.

An advantage of using data from real space is that it enables us to investigate how topological information is distributed in the system. This was demonstrated by the information entropy signatures recovered from the Shannon entropy of ensembles of eigenvectors in each experiment. The existence of such signals that can be recovered from data from many distinct physical systems but are hard to conceptualize from sheer theoretical reasoning provides a clear example of how machine learning can be an important tool in the investigation and discovery of new quantum materials.

We should remark on the subtleties of the information entropy signatures presented here. Although they give us a visualization of how important each lattice site was in determining the topological phases of Hamiltonians, these importances actually express a global property of the whole lattice. Therefore, a lattice site that appears unimportant in an information entropy signature plot may not be unimportant or void of topological information by itself. To give a concrete example, reduction in Shannon entropy tends to be distributed among highly correlated variables. This implies that if only a single lattice site in a highly correlated subset is used, it will likely inherit most of the reduction in Shannon entropy from the other correlated lattice sites in the subset. In this regard the information entropy signatures presented here express a summary of relations between lattice sites and are therefore intrinsically global. Furthermore, it should be emphasized that these signals were recovered from the analysis of data from several thousand SSH systems in each experiment and therefore they are not a property of a single SSH lattice. They are rather a pattern that emerges from correlating topological phase with lattice eigenvector data for several SSH systems.

We performed several tests on the information entropy signatures. By rerunning each experiment with longer lattices (i.e. increasing the number of unit cells) we have verified that the signals in figures 6(a) and 6(b) appear to converge to well defined continuous density functions in the macroscopic limit. The macroscopic limit of these information entropy signals may be an important signature of topological systems and thus merits further theoretical investigation. A detailed discussion of this point is presented in the section **Information entropy signatures in the macroscopic limit** in the Supplementary Material.



(a)

(b)

(c)

(d)

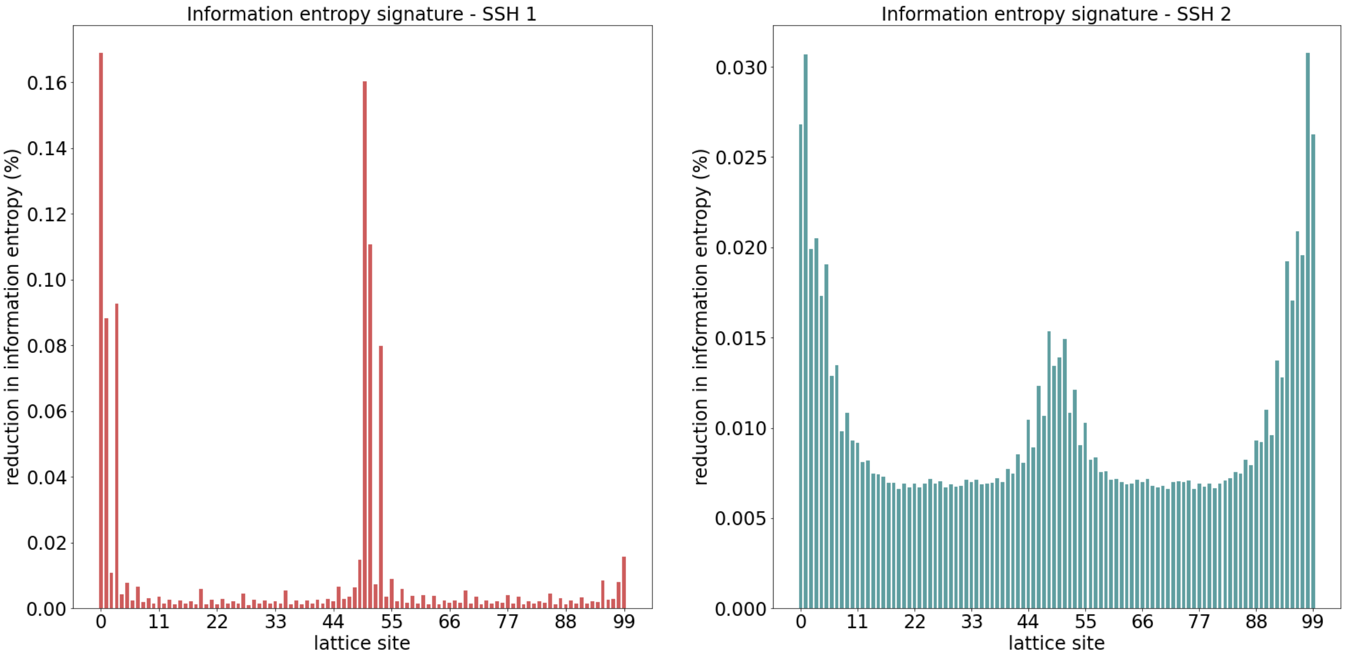
(e)

FIG. 5. Probability heatmaps learned by a combination of random forests with eigenvector ensembling from bulk real space eigenvector data in experiment 2. Heatmaps were averaged across all 100 iterations of the experiment. (a) Probability heatmap showing the probability that a Hamiltonian in the grid has winding number equal to -1. (b) Probability heatmap showing the probability that a Hamiltonian in the grid has winding number equal to 0. (c) Probability heatmap showing the probability that a Hamiltonian in the grid has winding number equal to 1. (d) Probability heatmap showing the probability that a Hamiltonian in the grid has winding number equal to 2. (e) The phase diagram resulting from heatmaps (a)-(d).

Recent works have demonstrated the existence of local topological markers in real space that carry important information on the topological state of a system [? ?]. Given that topological signals such as the ones shown in figures 6(a) and 6(b) are measured in terms of quantities that have actual physical meaning such as Shannon information entropy or Gini impurity, the results presented here suggest a new road for theoretical investigation. Whether there is any relationship between local topological markers and the information entropy of the ensemble of eigenvectors is left for speculation.

The eigenvector ensembling algorithm employed in this work is likely to have further applications in data-driven physics. This is because most of physics is based on eigenvector decomposition, and statistical physics itself can be seen as an application of similar ensembling principles. It is therefore expected that a much broader class of data-driven physics problems could benefit from the techniques described in this paper.

One final comment should be made about the flourishing relationship between physics and machine learning. In this work we have demonstrated how a machine learning approach can provide new insights into complex physical phenomena of current interest. The other direction of this relationship (physics enhancing understanding in machine learning) is equally important. As the need for ever more powerful machine learning algorithms continues to grow, the development of mathematical frameworks for understanding general data spaces (i.e., a physics of data) will be of crucial relevance. This pursuit is seen in many theoretical works investigating the intriguing connections between geometry, topology and data [?]–[?]. The detailed study of data generated by physical models with non-trivial geometrical and topological properties such as the SSH model may provide invaluable insights into the structure



(a)

(b)

FIG. 6. Information entropy signatures of the topological phase transitions from experiments 1 and 2. (a) In experiment 1, the two sharp peaks in the Shannon entropy signal account for approximately 70% of reduction in information entropy. These two peaks correspond to the lattice sites S_1. (b) In experiment 2, the three visible peaks account for approximately 30% of reduction in information entropy. These three peaks are located along lattice sites S_2.

and shape of real world high-dimensional data, since these models usually underscore well known mathematical frameworks behind the data generating process, a feature that is often absent from machine learning applications. Thus, far from being restricted to applications in physics, the study of the topological and geometrical properties of data sets generated by physical models will also be of great value to the machine learning and artificial intelligence communities.

ACKNOWLEDGEMENTS

We thank S. E. Rowley, J. F. de Oliveira, T. Micklitz and M. A. Continentino for insightful discussions and S. E. Rowley for carefully reading the manuscript and suggesting improvements. N. L. Holanda acknowledges financial support from CENPES/Petrobrás/CBPF. M. A. R. Griffith acknowledges financial support from Capes. N. L. Holanda is grateful to the Theory of Condensed Matter and Quantum Materials groups at the Cavendish Laboratory and the Quantum Information Group at CBPF.

AUTHOR CONTRIBUTIONS

Both authors of this work contributed equally to its realization at all stages.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial or non-financial interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to N. L. Holanda. The source code used to run simulations is available on Github.

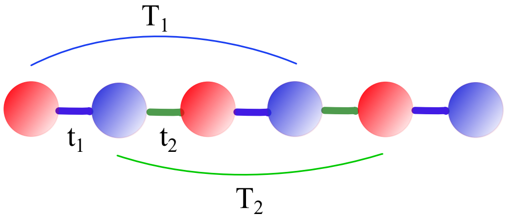


FIG. 7. Schematic set-up of a 1D SSH system. Here t_1 and t_2 are nearest-neighbor hoppings while T_1 and T_2 are second nearest-neighbor hoppings.

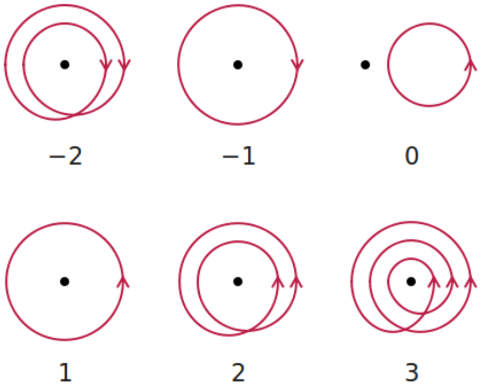


FIG. 8. Winding number. The winding number of a closed, oriented curve with respect to a reference point is a topological invariant that counts how many times the curve winds around the point. Picture credits: Jim Belk, public domain.

SUPPLEMENTARY MATERIAL

The SSH model

The SSH model [**?**] 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 **t** that also encodes hopping terms, thus allowing for a compact mathematical description of a Hamiltonian in terms of creation/annihilation operators as

$$\hat{H}(\mathbf{t}) = \mathbf{c}^\dagger H(\mathbf{t})\mathbf{c} \tag{2}$$

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 (2) 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 **t** in parameter space as very compact representations of SSH models: each point in **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

$$H(t_1, t_2, T_1, T_2) = \begin{pmatrix} 0 & t_1 & 0 & T_1 & 0 & \cdots \\ t_1 & 0 & t_2 & 0 & T_2 & \cdots \\ 0 & t_2 & 0 & t_1 & 0 & \cdots \\ T_1 & 0 & t_1 & 0 & t_2 & \cdots \\ 0 & T_2 & 0 & t_2 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}_{N\times N} . \tag{3}$$

Knowing the vector $\mathbf{t} = (t_1, t_2, T_1, T_2)$ corresponding to a particular system described by the Hamiltonian in equation (3) should suffice to compute any of its physical properties, including its topological phase. Figure 7 depicts a SSH system described by equation (3).

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 8 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 8 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

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

where $H(k)$ is the kernel in wavevector space of a Hamiltonian $\hat{H}(\mathbf{t})$ and σ_3 is the chiral operator. Equation (4) 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.

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. (4) 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 [**?**]. 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 $W(\mathbf{t})$,

$$W(\mathbf{t} + \boldsymbol{\delta}) \approx W(\mathbf{t}). \tag{5}$$

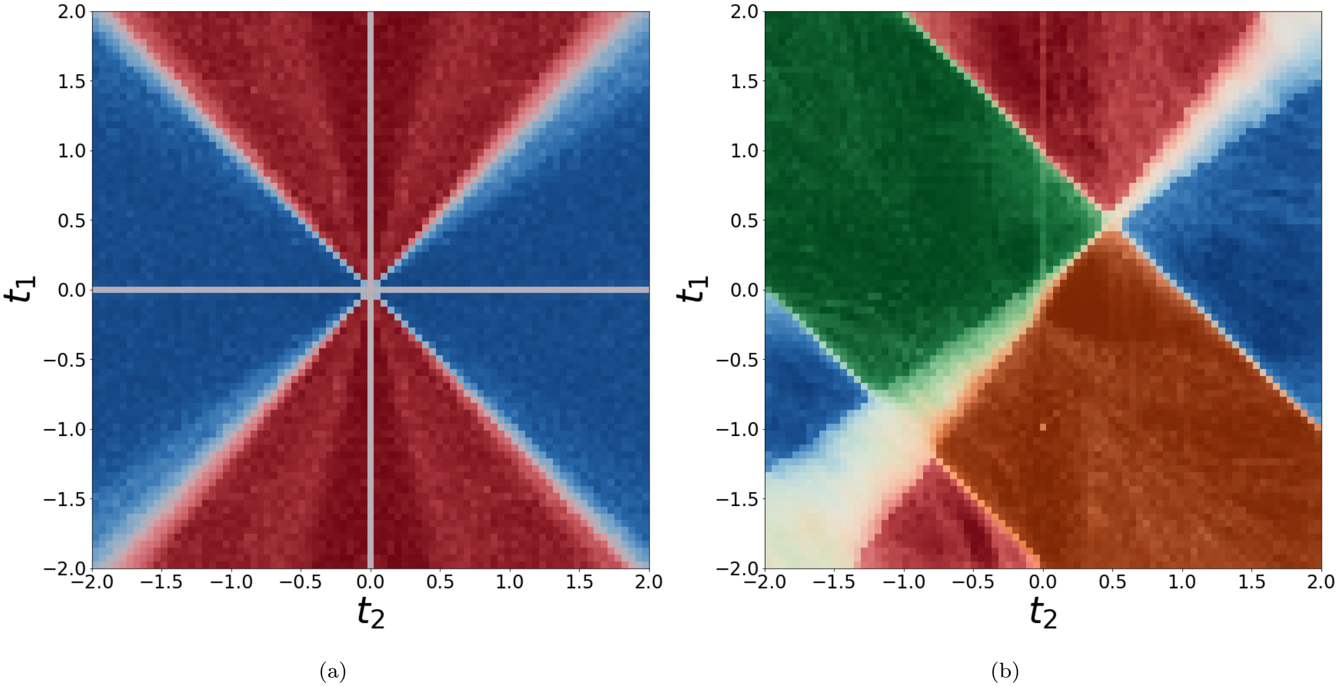


FIG. 9. Phase diagrams learned from most relevant lattice sites. (a) Phase diagram learned by a combination of decision trees with eigenvector ensembling in experiment 1 using only lattice sites $S_1' = (0,\ 50,\ 51,\ 99)$. The accuracy achieved for Hamiltonians in the test set using only the four lattice 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, 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 $v^{(j,l)} : \mathbb{R}^h \to \mathbb{R}^N$,

$$\mathbf{t} = (t_1, .., t_h) \to \left(v^{(j,1)}(t_1, ..., t_h), ..., v^{(j,N)}(t_1, ..., t_h)\right) = \mathbf{v}^{(j)}(\mathbf{t}) \tag{6}$$

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) 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 [**?**]. 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 [**?** **?**]: 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 [**?**]). 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 9). 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.

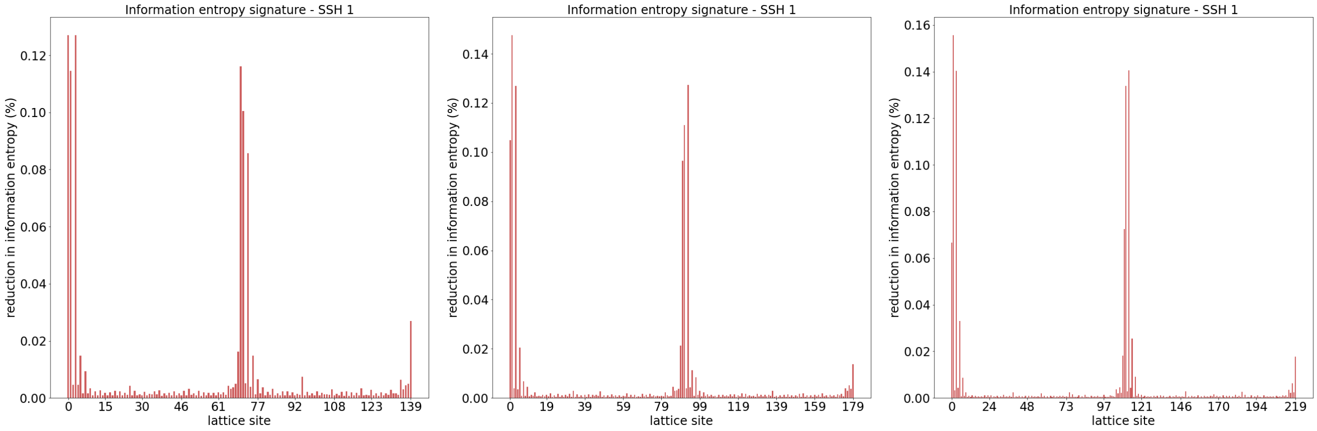
Information entropy signatures in the macroscopic limit

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



(a) (b) (c)

FIG. 10. 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.

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 10 and 11). 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).

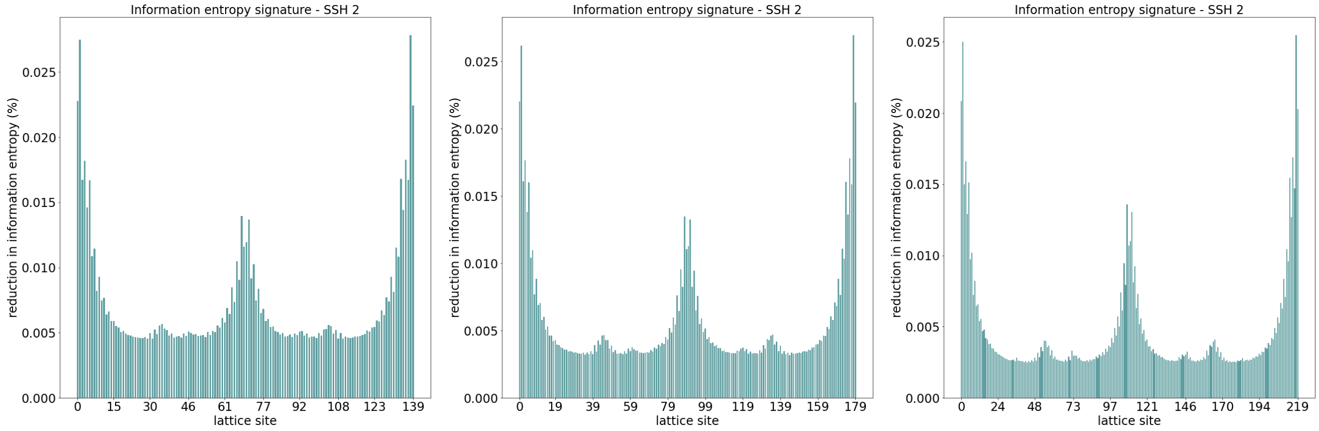
It is clear from figures 10 and 11 that the patterns seen with 50 unit cells (figures 6(a) and 6(b)) are stable across higher lattice sizes, with finer details emerging in the signals as the lattice size increases.

Figures 10 and 11 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 $\rho_S(x)$,

$$F_S(x) = \int_0^x \rho_S(x')dx' \tag{7}$$

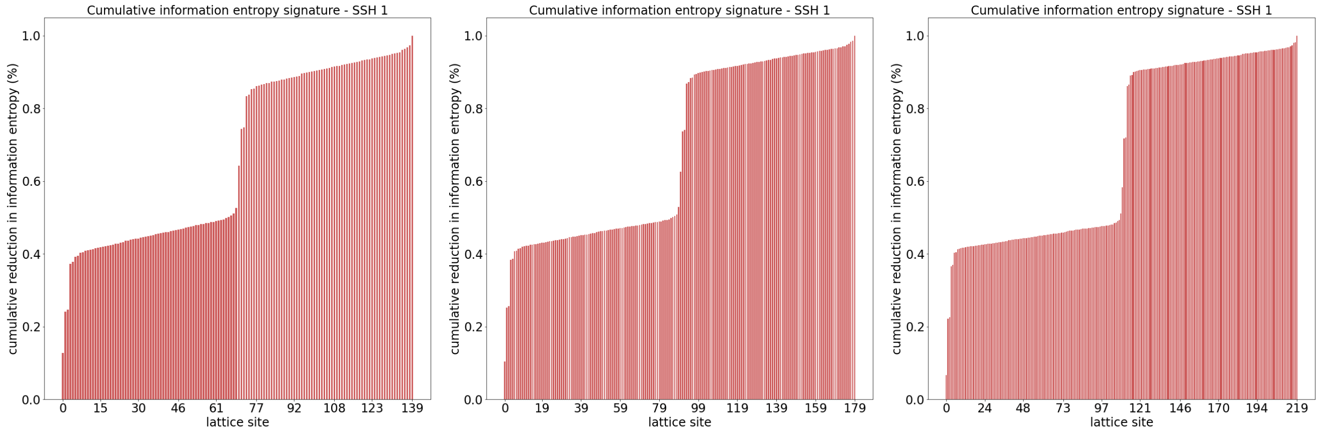
where $x \in$ [0,1] is a spatial coordinate along the lattice (being strict, our use of periodic boundary conditions implies that the coordinate x should be defined in the quotient space $[0,1]/R$, where R is the equivalence relation in [0,1] defining the unit circle S^1, i.e. $x\,R\,x'$ if and only 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 12 and 13 show the cumulative entropy distributions corresponding to the information entropy signatures in figures 10 and 11.

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.



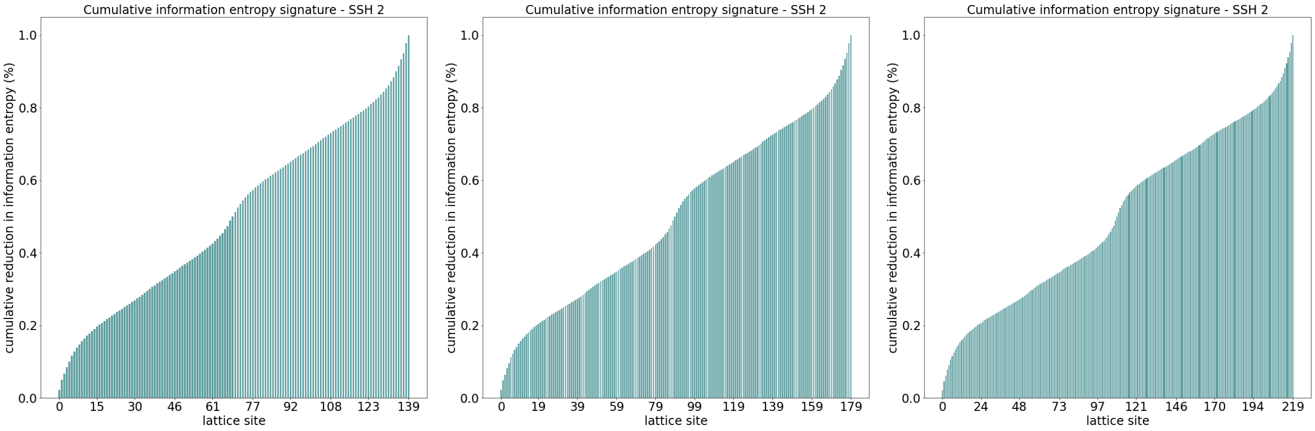
(a) (b) (c)

FIG. 11. 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. 12. 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. 13. 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.