REPLY TO THE FIRST REFEREE

REPLY TO THE SECOND REFEREE

Dear referee,

we appreciate the valuable insights on how to improve our article that you have provided us with. Below we address each of the emphasized points separately, as they were written in the reply that was sent back to us. We quote the issues raised in italic, and highlight in boldface the points that we thought were particularly relevant.

Major

I am not sure about the novelty and generality of their method. [1] *For example, their pipeline cannot work for complex experimental physical systems, where the Hamiltonian could be unknown. Moreover, the authors state that it is sufficient to use the input data in real space to predict the topological phase with high accuracy.* ***It is not clear to me what "sufficient" and "high accuracy" mean [2]*** *, and the advantages of using the input data in real space instead of the Hamiltonian in wavevector space.* ***I expect the evidence to indicate that their method is superior or comparable to any other method [3]*** *. It is worth mentioning that the author already discussed* ***the motivations for developing a data-driven approach based on real space [4]*** *in the "Learning topological phases from real space data" section. However, these arguments should be pointed carefully in the "Introduction" section.*

[1] We can think of several use cases for the eigenvector ensembling procedure outlined in the paper. Among these, we emphasize its applicability in dealing with the so-called "curse of dimensionality" that plagues many problems in condensed matter and statistical physics. In the paper, we have demonstrated an application in which the dimensionality of an SSH lattice can be greatly reduced in so far as its topological properties are concerned, simply by extracting the most relevant lattice sites in its information entropy signature. Similar feature extraction and dimensionality reduction procedures are often employed in machine learning applications and are surely to become the norm in data-driven physics. The eigenvector ensembling algorithm can thus be viewed as a general framework in this direction. Similarly, by thoughtfully attaching a probability distribution to the eigenvectors, we end up with a powerful sampling technique that may ultimately lead to great reduction in the dimensionality of a Monte Carlo simulation, while still capturing essentially all the relevant physics pertaining to the problem. We have added a paragraph to the section **Discussion** outlining these possibilities. As for the novelty of the method, the article binds together contemporary concepts in physics (topological states of matter and information theory) to a leading research theme in machine learning (model explainability). Our work pioneers the use of model explainability in the discovery of new concepts in physics, namely, the Shannon entropy signature and the information entropy uncertainty relations for topological phase transitions presented here for the first time. We have added a paragraph to the **Introduction** emphasizing this point.

[2] , **[3]** We use the term "sufficient accuracy" to denote a system that performs considerably better at predicting the phases of Hamiltonians than the baseline system that simply guesses the most frequently occurring class for all Hamiltonians. This gauge is needed because it ensures that the resulting entropy signatures are meaningful in the sense that they encode a summary of where useful information about the topological phase of a system can be found. We have added a paragraph to the section **Numerical experiments** to emphasize the use of this baseline. As to "high accuracy", we note that the accuracy scores obtained by us are comparable to the highest scores reported in Phys. Rev. Lett. 120, 066401, where neural networks were trained on wavevector space data in a similar supervised learning task. This second gauge further reinforces that the entropy signatures presented in the article indeed express a realistic summary of the availability of topological information from each lattice site. A small paragraph was added to the section **Discussion** to explicitly note this point. In another paragraph added to the section **Discussion**, we further draw comparisons between our paper and the just recently published article Phys. Rev. Research 2, 023283, whose theme is model interpretability of single-layer feedforward neural networks trained to recognize phase transitions of some topological condensed matter systems. When contrasting our paper to the latter, we emphasize the important distinction between "model interpretability" and "model explainability". While both techniques rely on model introspection, the first is aimed at interpreting the model itself vis-à-vis previously known concepts and properties of the training data, while in the latter the goal is to use these model introspection tools to arrive at new concepts and ideas about the physical phenomenon underlying the data, which is the main purpose of our work.

[4] We agree that the motivations for a data driven approach based on real space data should be in the Introduction. We have therefore moved the first paragraph of the section **Learning topological phases from real space data** in the Supplementary Material that addresses this issue to the Introduction.

I am not sure which crucial problems their algorithm can solve that was not possible before with machine learning. Furthermore, along with the emerging research of unsupervised learning methods in realizing the topological phases, ***why should the supervised learning method be focused in this context? [5]*** *In the present stage of this manuscript, where I do not see the proposal's advantages, I would prefer the* ***unsupervised approaches that could grasp information from a given system without knowing their phases [6]*** *. In my opinion, a proper intrinsic route for understanding physical systems without much information on the dynamics of the system will lead to significant future developments and a better understanding of physic systems.*

[5] , **[6]** Our main goal in the article is to investigate the localizability of topological information along lattices of SSH systems via model explainability tools in machine learning. We developed a pipeline that accomplishes this task in a very straightforward way by performing supervised learning with decision tree-based algorithms, since their use of Shannon entropy as loss function provides a clear way to measure feature importance. This is compounded by the fact that Shannon entropy plays a very important role in information theory and physics, as is evidenced by the entropic uncertainty relations included in this new version of the article. It should be noted, however, that the eigenvector ensembling algorithm can equally be combined with unsupervised learning algorithms as well. This point is now explicitly mentioned in this new version of the paper, in another paragraph added to the section **Discussion**. Finally, as we mentioned in our reply to **[1]** , our procedure can be used for example as a powerful dimensionality reduction tool in data-driven physics, which by itself is an unsupervised learning task.

The details of ***the eigenvector ensembling algorithm should be addressed in the main text [7]*** *instead of supplemental material. What is the specific algorithm used in the paper to train on eigenvectors? I think it is important even if the readers are not familiar with some ML techniques. Some readers in physics may find it difficult to understand some ML technical terms such as bootstrapping, training and validation, test sets, etc. In addition,* ***I expect proof of what "physics" their method captures and why the eigenvector can play an important role in characterizing the topological phase [8]*** *. If this physical interpretability is not mentioned, it is very difficult to see the contribution of the method in physics.*

[7] We have moved the section **The eigenvector ensembling algorithm** from the Supplementary Material to the main paper. The issue of specifying the particular learning algorithms used to train on eigenvectors is addressed in the subsections for each SSH experiment (see subsections **Experiment 1: Learning a first-neighbor hopping SSH model with decision trees** and **Experiment 2: Learning a first- and second-neighbor hoppings SSH model with random forests**). In these subsections, we explicitly cite the papers by Breiman *et al.* that were relevant to the implementation of decision trees and random forests in the Python scikit-learn module.

[8] As we argue in a paragraph added to the new section **The eigenvector ensembling algorithm**, all information from a Hamiltonian can be recovered from its eigendecomposition. Therefore, by expressing a data set of eigenvectors on a suitable choice of basis (which in this paper is the real space basis), we are able to investigate properties of a family of Hamiltonians using the coordinates of eigenvectors in the chosen basis as features. Thus the use of eigenvectors greatly facilitates the exploration of model explainability techniques in data-driven physics applications while at the same time not leading to a loss of information from the systems being investigated. As to the physics being captured by our procedure, we emphasize that the Shannon entropy signatures of topological phase transitions demonstrated in our paper are physically interpretable in terms of entropic uncertainty relations which ultimately lead to the development of an uncertainty principle limiting the localizability of topological phase transitions. This topic is explored in the new section **Entropic uncertainty relations**.

The authors made efforts to analyze how the algorithm was able to recover the Hamiltonians' global property in the "Information Entropy Signatures" section. The authors state that learning topological phases from local real-space data in bulk is still possible even for small subsets of lattice sites, then refer us to the section "Learning topological phases from real space data" in the Supplementary Material. The authors mention on page 4 of the Supplementary Material that ***"key topological information can be said to be localized on a few lattice sites," which is a particularly interesting statement to me [9]*** *. However,* ***I fail to understand the physical insights behind it [10]*** *. Without the proper explanation, it is difficult to see the effectiveness of their method or verify the method with a more complex model instead of the simple SSH form.*

[9] We agree that this sentence greatly synthesizes the results presented in the paper about the possibility of localizing topological information in SSH lattices. We have therefore moved it to the final patragraph of the section "Information entropy signatures" in the main text.

[10] Add FFT of entropy signatures? Mention entropic uncertainty? Talk about information gain? Mention that interpretability in ML is often challenging and a fast evolving field. Furthermore, eigenvector ensembling can always be used as a preprocessing step to other ML models, such as in semi supervised learning tasks.

Minor

The authors state in the abstract that "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." ***Could you explain a bit more about properties that may power future quantum computing? [11]***

[11] Insert comments about topological qubits?

On page 4, the authors state that "Figures 2 and 3 illustrate single iterations of experiments 1 and 2 as seen from parameter space". ***What are the experiments 1 and 2? [12]***

[12] Experiments 1 and 2 were mentioned as subtitles to the sections.

The authors should explain the evaluation metric in the main text [13] *, such as "the accuracy" and "the probability heatmap," "eigenvector accuracy," and "Hamiltonian accuracy."*

[13]

I prefer to ***put the definition of information entropy signatures in the main [14]*** *text to help the readers understand what the method tries to do.*

[14]

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 Ciências Naturais, Universidade Federal de São João Del Rei,
Praça Dom Helvécio 74, 36301-160, São João Del Rei, MG, Brazil
(Dated: June 27, 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 [1]. 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 [2]$^-$[5], ultracold atoms [6]$^-$[10], photonic crystals [11]$^-$[13], photonic quantum walks [14]$^-$[16] and Weyl semimetals [17, 18]. Among these, the Su-Schrieffer-Heeger (SSH) model [19] 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 [20], polymeric materials [21] and topological mechanisms [22, 23].

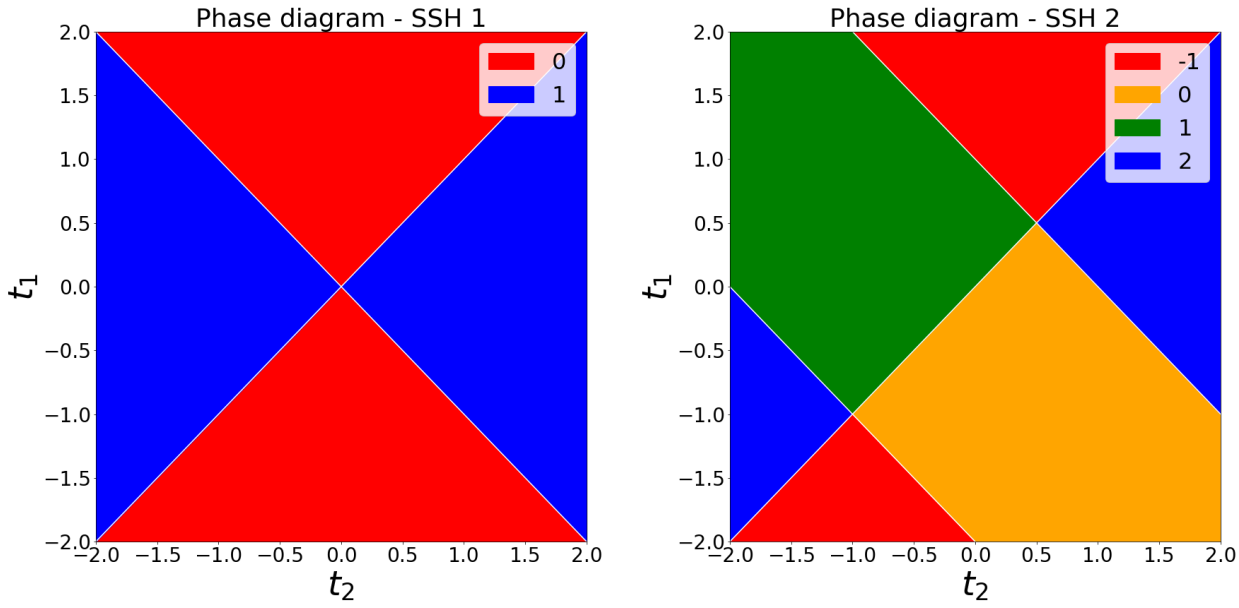
Many recent papers have explored the possibility of treating the general problem of determining phase transition boundaries of physical systems as machine learning tasks [24]$^-$[40]. In the particular case of topological phase transitions, the usual approach for supervised learning is to generate a 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.

The main motivations for developing a data-driven approach based on real space are that wavevector space computations of topological invariants 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. The data-driven approach described in this article addresses these issues. In particular, we employ it to scrutinize the issue of localizability of information in topological condensed matter systems **[4]**

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 patterns 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 system as a whole 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 topological signals recovered from the analysis of real space eigenvector data are then further theoretically explored as information entropy mass functions along the lattices in real space. By computing their Fourier transform, we retrieve similar information entropy mass functions in wavevector space. These two topological signals, the real space Shannon entropy mass function and its Fourier counterpart in the reciprocal lattice, are knit together by entropic uncertainty relations. This allows us to establish the Bialynicki-Birula-Mycielski inequality as an uncertainty principle for the localizability of Shannon information entropy in topological phase transitions. **[8]** , **[10]**



(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.

The Shannon information entropy signals and the corresponding entropic uncertainty relations associated with topological phase transitions 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.**[1]**

As of yet model explainability[**? ? ?**] is one of the topics at the edge of machine learning research that has been little explored by the physics community working at the interface between the two disciplines (this is also emphasized in [**?**]). This raises important questions as to whether machine learning can in fact help to advance theoretical investigation in physics, since the majority of physics papers published on the subject are proofs of concept aimed at showing that modern machine learning techniques are capable of recognizing the relevant patterns in data from physical systems whose properties were already known. By proposing new concepts from the analysis of data from physical systems of contemporary interest and knitting together ideas from topological phase transitions and information theory by dint of model explainability, we expect to draw the physics community's attention to this essential machine learning tool. **[1]**

THE EIGENVECTOR ENSEMBLING ALGORITHM

[7]

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. We describe here in detail each of these steps as they were implemented in this work.

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 hopping parameters vector $\mathbf{t}$ 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 provide N eigenvectors $\mathbf{v}_i^{(1)}, \mathbf{v}_i^{(2)}, ..., \mathbf{v}_i^{(N)}$ to our data set. 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.

Before continuing to the analyses of the SSH systems with the eigenvector ensembling algorithm, it will be timely to digress a moment and peek into the algorithm itself. The focus on eigenvectors (and hence the algorithm's name) as the input data to a machine learning algorithm of choice is a hallmark of the procedure as it differentiates it from related applications of machine learning to the study of phase transitions. The intuition that eigenvectors can be used in replacement to raw Hamiltonians can be grasped when we consider the spectral decomposition of a Hamiltonian H,

$$H = \sum_{i=1}^{N} \lambda^{(i)} \left|\mathbf{v}^{(i)}\right\rangle \left\langle\mathbf{v}^{(i)}\right| \tag{2}$$

where $\lambda^{(i)}$ is the eigenenergy corresponding to the eigenstate $|\mathbf{v}^{(i)}\rangle$. It is therefore clear that all information available from a Hamiltonian can be recovered from its spectral decomposition. By expressing the eigenvectors in a basis suitable to a particular problem (e.g. the real space basis chosen in this article), it becomes possible to investigate the properties of a set of Hamiltonians using the coordinates of eigenvectors in the chosen basis as features. Thus the eigenvector ensembling procedure described above provides a broad framework for the implementation of model explainability in applications to data-driven physics.

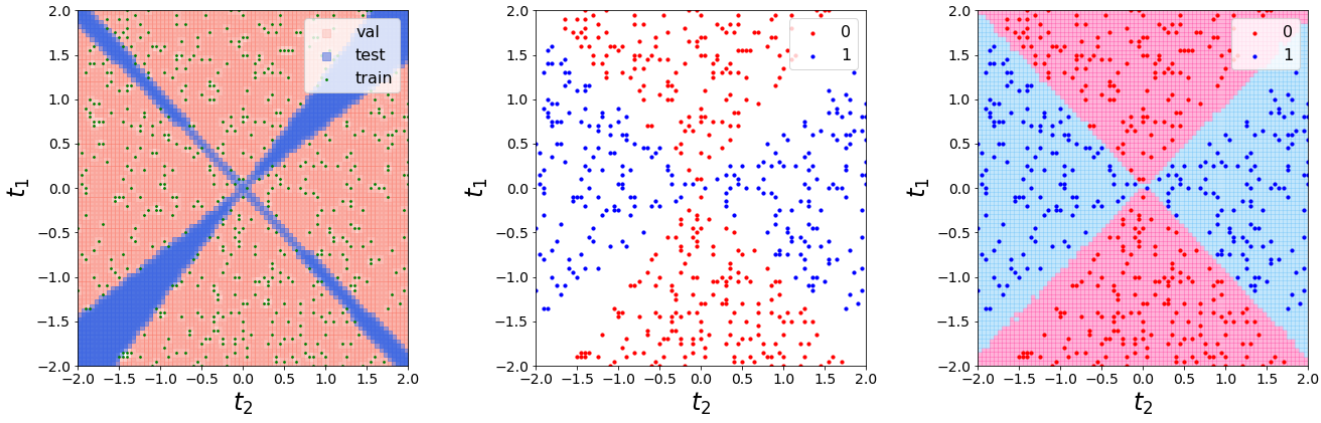
NUMERICAL EXPERIMENTS

We performed two numerical experiments with the eigenvector ensembling algorithm. The first experiment deals with the simplest case, the SSH model with nearest-neighbor hopping (here called SSH 1, figure 1(a)), while the second experiment uses 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 correct 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.

As performance metrics, we report here both accuracy of predicted classes for eigenvectors as well as accuracy of predicted classes for Hamiltonians obtained from eigenvector ensembling. These accuracy scores are to be gauged against the baseline of a system that simply guesses the most frequent class for all eigenvectors (or Hamiltonians).



(a) (b) (c)

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.

Checking against this baseline is important because it indicates whether the decision trees are in fact learning the underlying patterns that relate real space coordinates to winding numbers, and therefore whether the associated information entropy signature is meaningful or not.**[2] [13]**

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.

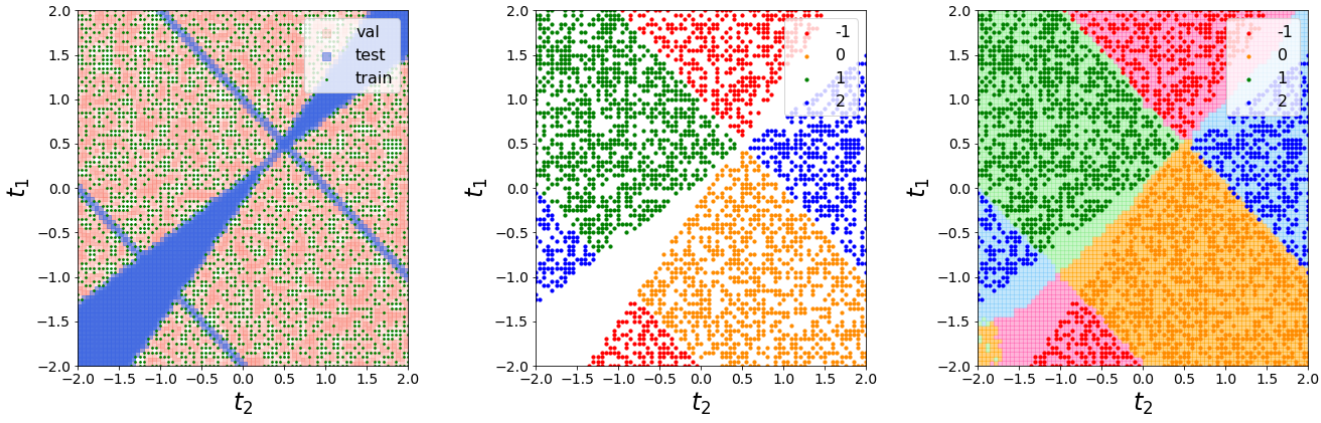
blueThe numerical experiments with the eigenvector ensembling algorithm described in the next subsections were implemented in Python using the scikit-learn module [**?**].

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 learning algorithm of choice for this experiment was a simple decision tree model [41].**[7]**

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. When compared with the baseline test accuracy of 55.2% of a system that predicts the whole test set as having winding number $W = 1$, the accuracy of 99.1% on test Hamiltonians indicates that the decision trees indeed learned the patterns that relate real space coordinates to winding numbers. **[2]**

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.



(a) (b) (c)

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.

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 [42] 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.

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 [43]). 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,

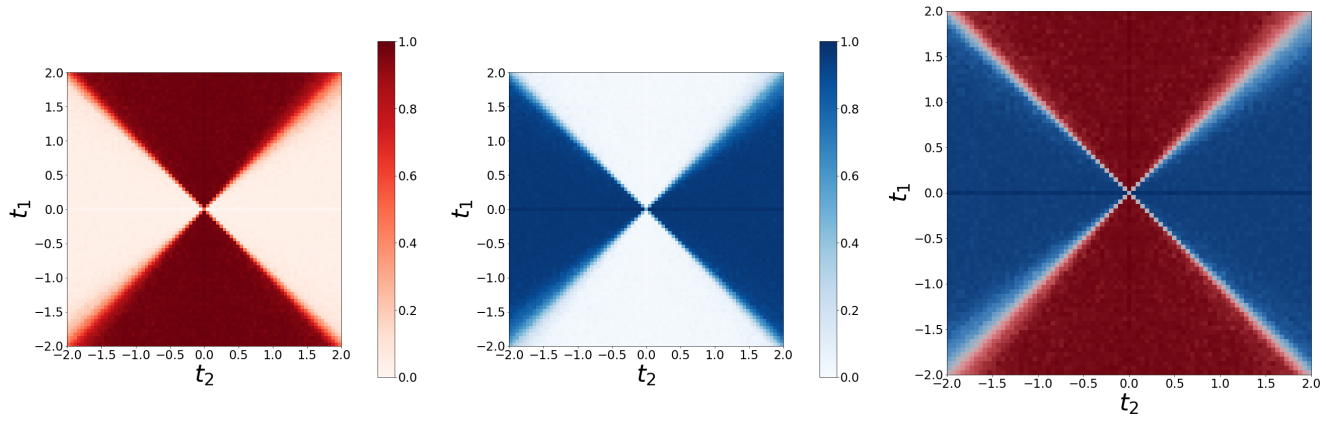


FIG. 4. Probability heatmaps learned by a combination of decision trees with eigenvector ensembling from bulk real space eigenvector data in experiment 1. Heatmaps were averaged across all 100 iterations of the experiment. (a) Probability heatmap showing the probability that a Hamiltonian in the grid 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).

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. In this sense, key topological information can be said to be localized on few sites in the lattice.**[9]** We refer the reader to the section **Learning topological phases from real 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

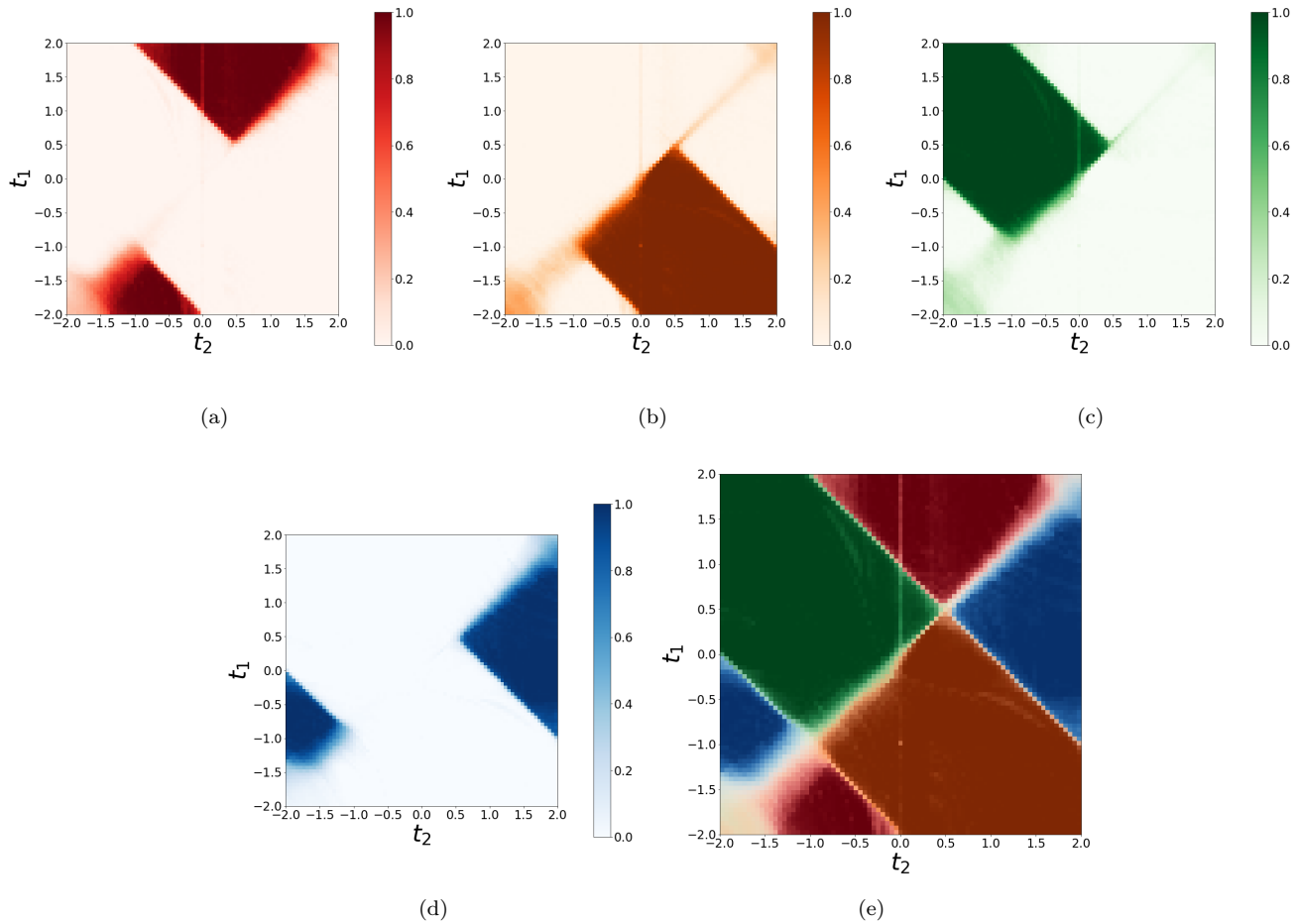


FIG. 5. Probability heatmaps learned by a combination of random forests with eigenvector ensembling from bulk real space eigenvector data in experiment 2. Heatmaps were averaged across all 100 iterations of the experiment. (a) Probability heatmap showing the probability that a Hamiltonian in the grid 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).

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.

The accuracy scores obtained in this paper were comparable to those reported in [29], where dense and convolutional neural networks were trained on wavevector space data to predict the winding numbers of SSH Hamiltonians via supervised learning. This high accuracy level serves as a strong evidence that the entropy signatures presented here indeed express where topological information can most readily be recovered from in the SSH lattices investigated.**[3]**

We should also compare our results to those obtained in [**?**], where the subject of investigation is the interpretability of neural network models trained to recognize topological phase transitions in some physical systems. In that paper, interesting visualizations are shown demonstrating that the patterns captured by a single-layer feedforward neural network indeed correspond to known physical quantities that are relevant to the problems at hand. This should indeed be called "model interpretability", as in that case the authors introspect into their models to make sure that they are learning patterns of physical pertinence to the systems at hand. In our paper we prefer the term "model explainability", as we are using similar model introspection tools to propose previously unknown concepts and properties of the physical systems being investigated. While the nuances in the meanings of these two terms are the subject of often heated philosophical debates in the artificial intelligence community, this choice of nomenclature suits the practical application of these model introspection techniques to physics well.**[3]**

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.

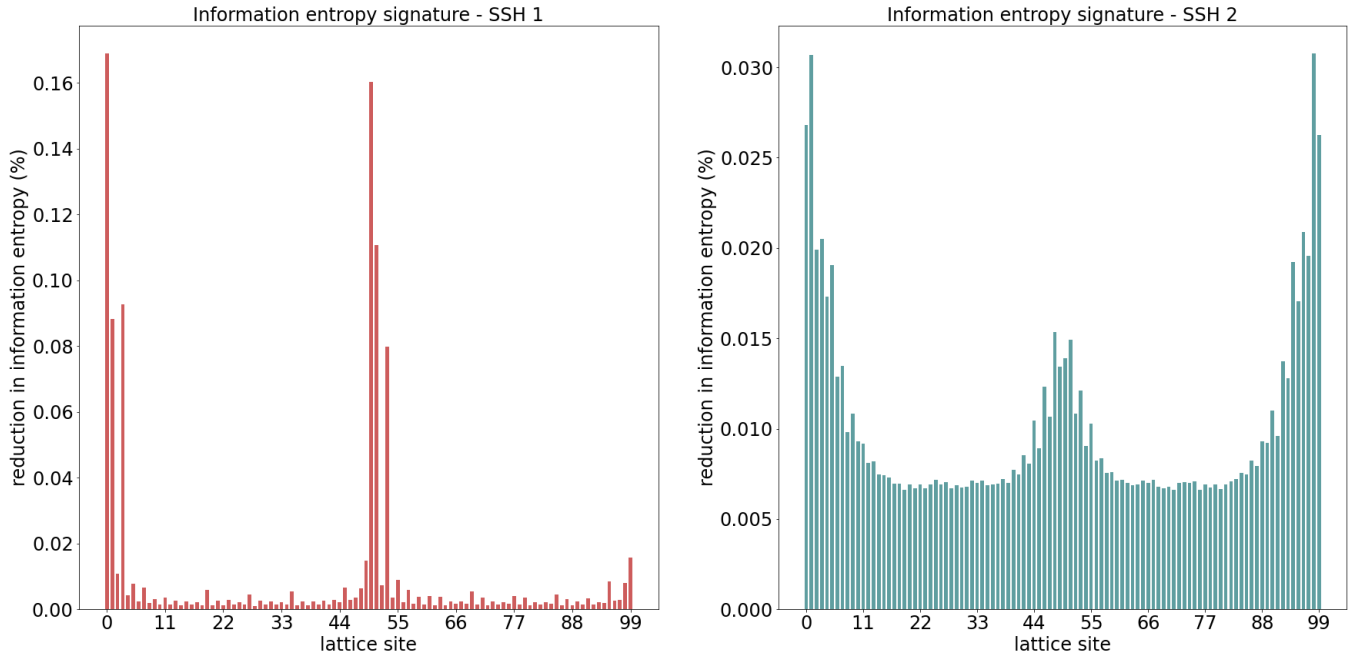
Recent works have demonstrated the existence of local topological markers in real space that carry important information on the topological state of a system [44, 45]. Given that topological signals such as the ones shown in figures 6(a) and 6(b) are measured in terms of quantities that have actual physical 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.

As a concrete example, the study of several many-body systems of current interest in condensed matter physics is hindered by their large dimensionality. This problem, known as *the curse of dimensionality* in the scientific computing community, arises from the necessity of collecting or processing exponentially larger amounts of data as the feature space dimensionality of a problem grows. An approach based on eigenvector ensembling can be of use in such situations both as a dimensionality reduction tool and as a sampling strategy. The first case was illustrated in this work, where it was shown that relevant topological information of SSH systems can be retrieved from few sites in a lattice, which can be exploited as a dimensionality reduction strategy. The latter case, which was not explored here, also poses interesting possibilities, such as sampling eigenstates according to a desired distribution in Monte Carlo simulations of condensed matter systems. Indeed, sampling eigenvectors from a carefully designed probability distribution can ultimately lead to a great reduction in the dimensionality of a problem while still capturing all the relevant physics of a system.**[1]** We therefore expect that a much broader class of data-driven physics problems could benefit from the techniques described in this paper.

Another interesting prospect is the combination of eigenvector ensembling with unsupervised learning algorithms. In the paper, our preference for decision trees and random forests was based on their powerful and accessible model explainability aptitudes. This choice was made in conformity with our main purpose, which was to exploit model explainability tools to investigate how topological information is distributed along a spatial lattice in SSH systems. Nevertheless, the eigenvector ensembling procedure we described here is flexible and can easily be repurposed for other supervised or unsupervised learning tasks.**[5]**

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 [46]$^{-}$[50]. 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 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.



(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.

[4]

[1] M. Z. Hasan and C. L. Kane, Rev. Mod. Phys. **82**, 3045 (2010).
[2] M. A. Continentino, Physica B: Condensed Matter **505**, A1 (2017).
[3] T. O. Puel, P. D. Sacramento, and M. A. Continentino, Phys. Rev. B **95**, 094509 (2017).
[4] M. A. Griffith and M. A. Continentino, Phys. Rev. E **97**, 012107 (2018).
[5] S. Ryu, A. P. Schnyder, A. Furusaki, and A. W. Ludwig, New Journal of Physics **12**, 065010 (2010).
[6] M. Atala, M. Aidelsburger, J. T. Barreiro, D. Abanin, T. Kitagawa, E. Demler, and I. Bloch, Nature Physics **9**, 795 (2013).
[7] B. K. Stuhl, H.-I. Lu, L. M. Aycock, D. Genkina, and I. B. Spielman, Science **349**, 1514 (2015).
[8] M. Leder, C. Grossert, L. Sitta, M. Genske, A. Rosch, and M. Weitz, Nature communications **7**, 13112 (2016).
[9] N. Goldman, J. Budich, and P. Zoller, Nature Physics **12**, 639 (2016).
[10] E. J. Meier, F. A. An, and B. Gadway, Nature communications **7**, 13986 (2016).
[11] M. Hafezi, S. Mittal, J. Fan, A. Migdall, and J. Taylor, Nature Photonics **7**, 1001 (2013).
[12] L. Lu, J. D. Joannopoulos, and M. Soljačić, Nature Physics **12**, 626 (2016).
[13] V. Peano, C. Brendel, M. Schmidt, and F. Marquardt, Phys. Rev. X **5**, 031011 (2015).
[14] T. Kitagawa, M. A. Broome, A. Fedrizzi, M. S. Rudner, E. Berg, I. Kassal, A. Aspuru-Guzik, E. Demler, and A. G. White, Nature communications **3**, 882 (2012).
[15] F. Cardano, M. Maffei, F. Massa, B. Piccirillo, C. De Lisio, G. De Filippis, V. Cataudella, E. Santamato, and L. Marrucci, Nature communications **7**, 11439 (2016).
[16] E. Flurin, V. V. Ramasesh, S. Hacohen-Gourgy, L. S. Martin, N. Y. Yao, and I. Siddiqi, Phys. Rev. X **7**, 031023 (2017).
[17] A. A. Soluyanov, D. Gresch, Z. Wang, Q. Wu, M. Troyer, X. Dai, and B. A. Bernevig, Nature **527**, 495 (2015).
[18] B. Q. Lv, H. M. Weng, B. B. Fu, X. P. Wang, H. Miao, J. Ma, P. Richard, X. C. Huang, L. X. Zhao, G. F. Chen, Z. Fang, X. Dai, T. Qian, and H. Ding, Phys. Rev. X **5**, 031013 (2015).
[19] W. P. Su, J. R. Schrieffer, and A. J. Heeger, Phys. Rev. Lett. **42**, 1698 (1979).
[20] M. Maffei, A. Dauphin, F. Cardano, M. Lewenstein, and P. Massignan, New Journal of Physics **20**, 013023 (2018).
[21] A. J. Heeger, Rev. Mod. Phys. **73**, 681 (2001).

[22] C. Kane and T. Lubensky, Nature Physics **10**, 39 (2014).
[23] B. G.-g. Chen, N. Upadhyaya, and V. Vitelli, Proceedings of the National Academy of Sciences **111**, 13004 (2014).
[24] J. Carrasquilla and R. G. Melko, Nature Physics **13**, 431 (2017).
[25] K. Ch'ng, J. Carrasquilla, R. G. Melko, and E. Khatami, Phys. Rev. X **7**, 031038 (2017).
[26] L. Wang, Phys. Rev. B **94**, 195105 (2016).
[27] P. Broecker, J. Carrasquilla, R. G. Melko, and S. Trebst, Scientific reports **7**, 8823 (2017).
[28] E. P. Van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Nature Physics **13**, 435 (2017).
[29] P. Zhang, H. Shen, and H. Zhai, Phys. Rev. Lett. **120**, 066401 (2018).
[30] N. Sun, J. Yi, P. Zhang, H. Shen, and H. Zhai, Phys. Rev. B **98**, 085402 (2018).
[31] P. Suchsland and S. Wessel, Phys. Rev. B **97**, 174435 (2018).
[32] Y. Zhang and E.-A. Kim, Phys. Rev. Lett. **118**, 216401 (2017).
[33] J. Venderley, V. Khemani, and E.-A. Kim, Phys. Rev. Lett. **120**, 257204 (2018).
[34] T. Ohtsuki and T. Ohtsuki, Journal of the Physical Society of Japan **86**, 044708 (2017).
[35] N. Yoshioka, Y. Akagi, and H. Katsura, Phys. Rev. B **97**, 205110 (2018).
[36] D.-L. Deng, X. Li, and S. Das Sarma, Phys. Rev. B **96**, 195145 (2017).
[37] P. Huembeli, A. Dauphin, and P. Wittek, Phys. Rev. B **97**, 134109 (2018).
[38] D. Carvalho, N. A. García-Martínez, J. L. Lado, and J. Fernández-Rossier, Phys. Rev. B **97**, 115453 (2018).
[39] Y. Zhang, R. G. Melko, and E.-A. Kim, Phys. Rev. B **96**, 245119 (2017).
[40] J. F. Rodriguez-Nieva and M. S. Scheurer, arXiv preprint arXiv:1805.05961 (2018).
[41] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees* (Chapman and Hall/CRC, London, UK, 1984).
[42] L. Breiman, Machine Learning **45**, 5 (2001).
[43] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning* (Springer New York, NY, USA, 2001).
[44] R. Bianco and R. Resta, Phys. Rev. B **84**, 241106 (2011).
[45] M. D. Caio, G. Möller, N. R. Cooper, and M. Bhaseen, Nature Physics , 1 (2019).
[46] G. Carlsson, Bulletin of the American Mathematical Society **46**, 255 (2009).
[47] L. Wasserman, Annual Review of Statistics and Its Application **5**, 501 (2018).
[48] J. Wang, Z. Zhang, and H. Zha, in *Advances in neural information processing systems* (2005) pp. 1473–1480.
[49] T. Lin and H. Zha, IEEE Transactions on Pattern Analysis and Machine Intelligence **30**, 796 (2008).
[50] M. Belkin, *Problems of learning on manifolds* (The University of Chicago, Chicago, IL, USA, 2003).
[51] J. K. Asbóth, L. Oroszlány, and A. Pályi, Lecture notes in physics **919** (2016).
[52] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, Cambridge, MA, USA, 2016).
[53] C. M. Bishop, *Pattern recognition and machine learning* (Springer New York, NY, USA, 2006).
[54] L. Cayton, Univ. of California at San Diego Tech. Rep **12**, 1 (2005).
[55] H. Narayanan and S. Mitter, in *Advances in Neural Information Processing Systems* (2010) pp. 1786–1794.
[56] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Muller, in *Advances in Neural Information Processing Systems* (2011) pp. 2294–2302

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.