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 23, 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.

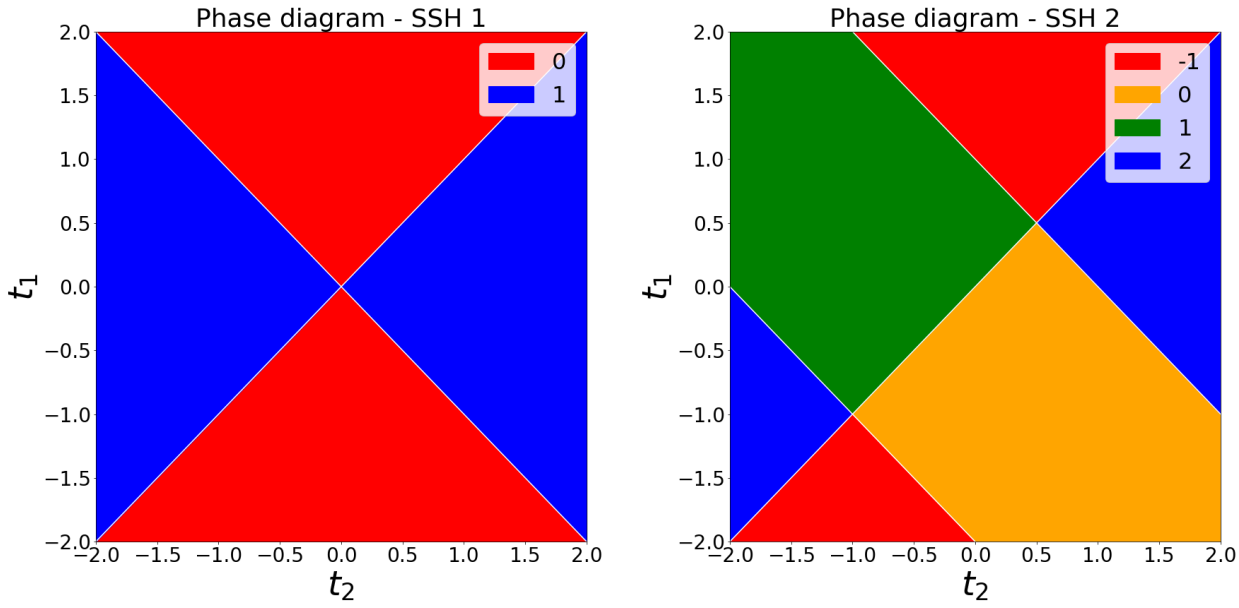
To investigate topological phases of matter in real space we have designed a novel supervised learning algorithm (here called eigenvector ensembling algorithm) tailored for the task of learning phase transition boundaries from local features. The algorithm is based on eigenvector decomposition and eigenvector 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 algorithm is found in the section **The eigenvector ensembling algorithm** in the Supplementary Material. Here



(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.

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).

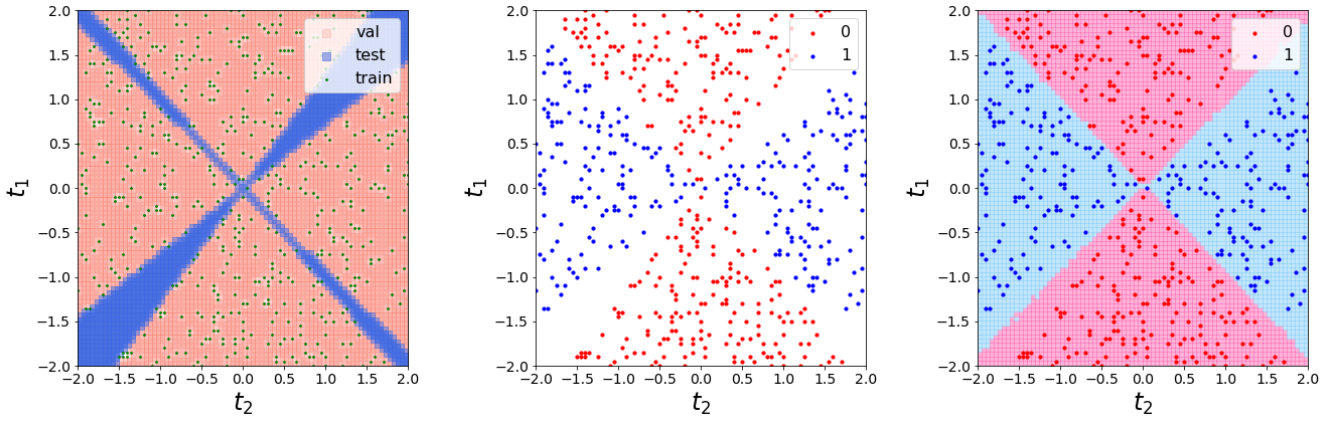
This task is particularly hard near phase transition boundaries, where numerical computation of winding numbers become less stable. For this reason, when sampling the training set we only consider those Hamiltonians in the grid whose numerically computed winding numbers lie in a range $\epsilon = 0.01$ 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.

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



(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.

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 [41].

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.

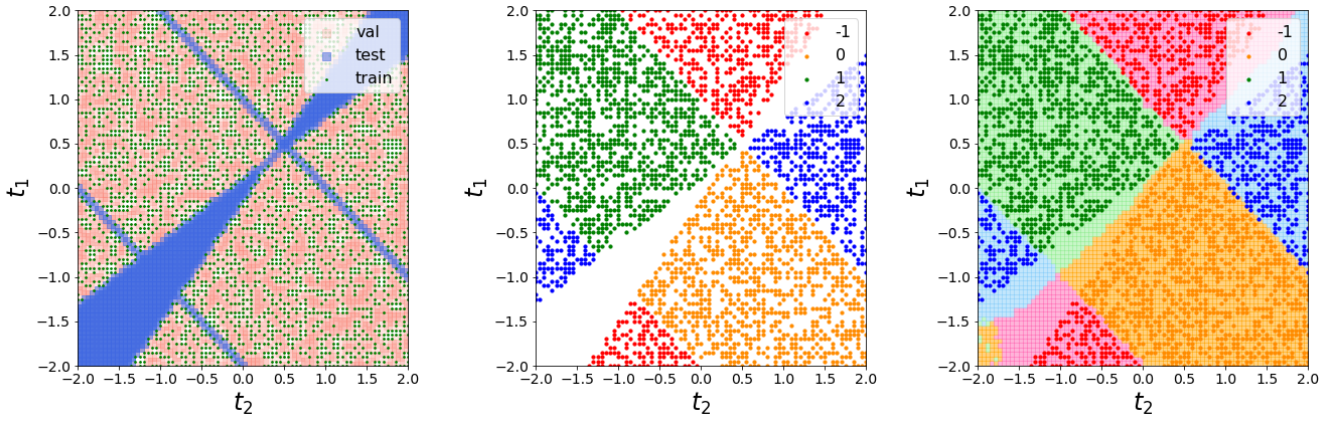
The probability heatmaps and phase diagram learned by the combination of decision trees with eigenvector ensembling used in experiment 1 are shown in figure 4.

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 [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.



(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.

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.

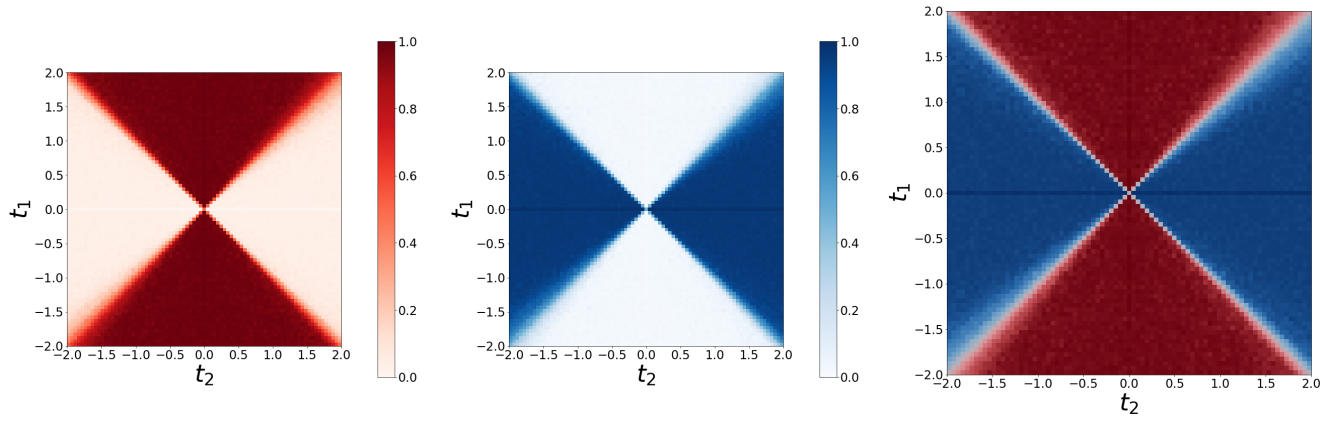
The bar plots in figure 6 show how informative each lattice site was in learning topological phases for experiments 1 and 2. They represent the information entropy signatures along the lattices in each SSH system. For experiment 1, only six lattice sites $S_1 = (0, 1, 3, 50, 51, 53)$ corresponding to 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 $S_2 = (0, 1, 2, 3, 4, 5, 46, 48, 49, 50, 51, 53, 94, 95, 96, 97, 98, 99)$ 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.

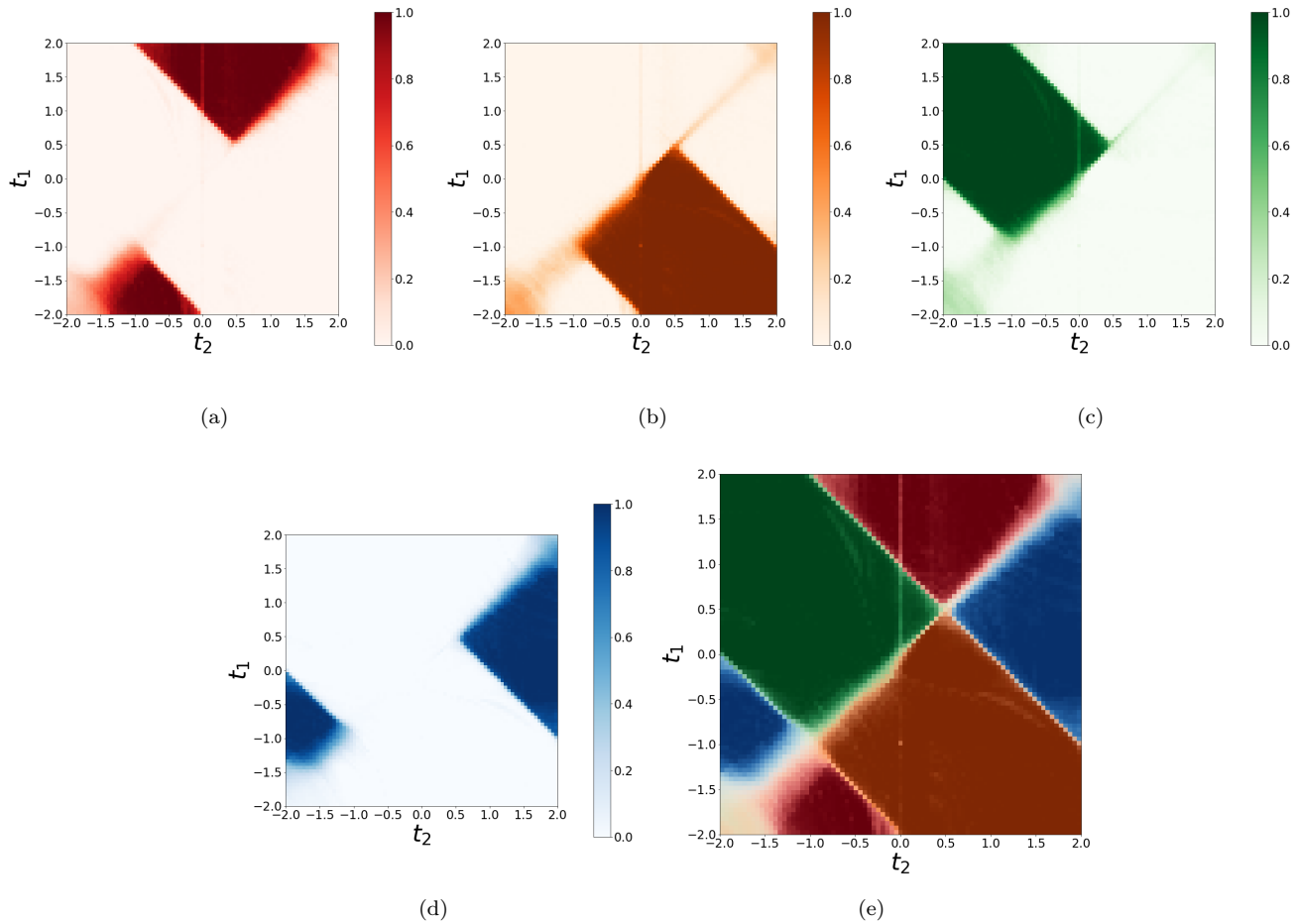
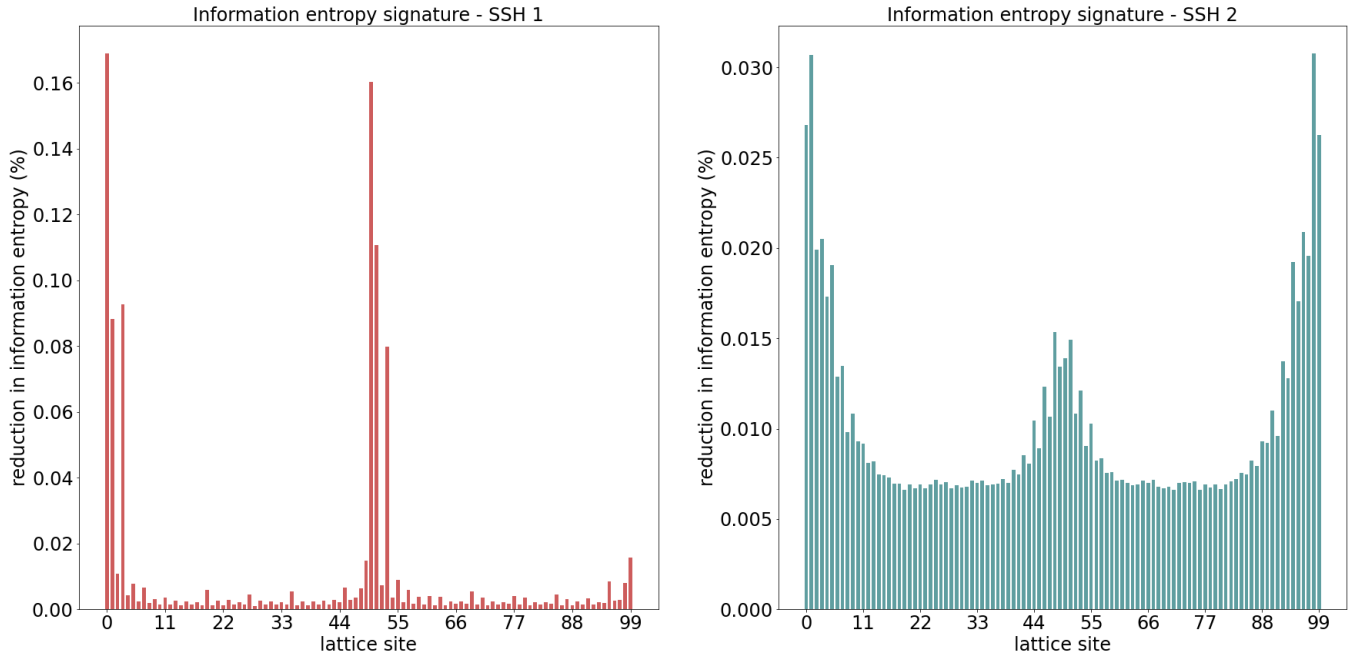


FIG. 5. Probability heatmaps learned by a combination of random forests with eigenvector ensembling from bulk real space eigenvector data in experiment 2. Heatmaps were averaged across all 100 iterations of the experiment. (a) Probability heatmap showing the probability that a Hamiltonian in the grid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



(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.

[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.