REPLY TO THE SECOND REFEREE

Dear referee,

we appreciate the valuable insights on how to improve our article that you have provided us with. These insights forced us into more careful theoretical reasoning about the results we had obtained and led us to expand our manuscript with new theoretical and numerical contents.

In what follows, we briefly summarize the modifications made to the paper. New results have been added to the paper in the form of two new sections: **V. Topological feature engineering and lattice compression** and **VI. Emergent information entropy wave functions**. We have also moved the description of the eigenvector ensembling algorithm from the **Supplementary material** to the main text, in which it became the section **II. The eigenvector ensembling algorithm**. The section **VII. Discussion** has been extended to include comments on the issues that were raised. The sections are now numbered for clarity.

Below we address each of the issues that have been raised separately, as they were written in the reply that was sent back to us. We quote the issues in italic, and highlight in bold the points that we thought were particularly relevant and guided us in the improvement of the manuscript. The answers to each of the major points are written in blue.

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 **VII. 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 engineering of new topological features from real space data and the Shannon information entropy signature and uncertainty relations for topological phase transitions presented here for the first time in the new sections **V. Topological feature engineering and lattice compression** and **VI. Emergent information entropy wave functions**. We have added a paragraph to the section **I. Introduction** emphasizing these new results.

[2] 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 **III. Numerical experiments** to emphasize the use of this baseline.

[3] 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--vis previously known concepts and properties related to the physics underlying the data set, while in the latter the goal is to use these model introspection tools to arrive at new concepts and ideas about the physical phenomena underlying the data set, which is the main purpose of our work.

[4] We agree that the motivations for a data-driven approach based on real space 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 section **I. Introduction** in the main paper.

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] Our main goal in the article is to investigate the distribution 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 information entropy as loss function provides a clear way to measure feature importance. This is compounded by the fact that information entropy plays a very important role in information theory and physics, as is evidenced in the new section **VI. Emergent information entropy wave functions** added to this new version of the article.

[6] The new section **V. Topological feature engineering and lattice compression** shows how unsupervised learning can be performed with eigenvector ensembling using tree based methods as learning algorithms. In fact, two of the main tasks under the broad umbrella of unsupervised learning are exactly those: dimensionality reduction and feature engineering. Furthermore, it should be noted 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 a new paragraph added to the section **VII. Discussion**. Finally, as we mentioned in our reply to **[1]** , our procedure can be used as a powerful dimensionality reduction tool in other, unrelated data-driven or numerical physics tasks, which by itself is another example of unsupervised learning and can serve as a preprocessing step to further applications of machine learning to physics.

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] The section **II. The eigenvector ensembling algorithm** has been moved 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 used to run the numerical experiments in the paper.

[8] As we argue in a paragraph added to the new section **II. 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, see the answer to the issue **[10]** .

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 **V. Topological feature engineering and lattice compression** in the main text.

[10] The upgraded version of the paper now performs a full cycle of data-driven physics, as we i) propose a method to investigate topological phase transitions using real space eigenvector data from simulations of SSH systems (the eigenvector ensembling algorithm), ii) use model explainability to extract new, previously unknown properties of the finite SSH systems (the information entropy signatures), iii) generate new topological features from real space data using the insights gained from the information entropy signatures in the added section **V. Topological feature engineering and lattice compression**, and finally iv) interpret these new properties in the light of theoretical physics in the added section **VI. Emergent information entropy wave functions**, in which we interpret the information entropy signature as a probability distribution arising from an emergent information entropy wave function in the continuum limit of an infinite lattice.

It is noteworthy that the majority of physics papers published to date on the use of machine learning in the investigation of condensed matter systems have been proofs of concept aimed at showing that the machine learning algorithms that have become ubiquitously famous during the last decade indeed are able to recognize the relevant patterns associated with previously known physical phenomena. Our paper goes beyond this task as we use machine learning and model explainability to discover new properties from the analysis of data from the physical systems under consideration and (with the addition of the new section **VI. Emergent information entropy wave functions**) map these previously unknown properties into new theoretical constructs.

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] We have added a paragraph to the section **Discussion** where we comment on the importance of information-theoretic methods to the design of prospective quantum technologies like topological qubits and quantum communication.

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] We have added references to the relevant subsections whenever each numerical experiment is mentioned outside of its subsection throughout the text. The designation "experiment 1" is used to abbreviate the subsection **III.A Experiment 1: Learning a first-neighbor hopping SSH model with decision trees**. Similarly, the designation "experiment 2" is used to abbreviate the subsection **III.B Experiment 2: Learning a first- and second-neighbor hoppings SSH model with random forests**.

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] We have added more detailed explanations about the evaluation metrics and probability heatmaps to the section **III. Numerical experiments**. Furthermore, the new section **V. Topological feature engineering and lattice compression** contains a summary of accuracy scores obtained in all numerical experiments performed in the paper.

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] The definition of information entropy signatures is included in the new section **VI. Emergent information entropy wave functions**, where an extended theoretical interpretation of the results obtained with the eigenvector ensembling algorithm is also presented.