Unsupervised and supervised learning of interacting topological phases from single-particle correlation functions

Response to the Referee reports

Referee 1

We thank the Referees for their careful reading of our manuscript. We appreciate that it was found to be a clear reading with exhaustive analysis and that they found good promises in our work. We address the Referee's comments in the following:

> **Weaknesses**
>
> 1) The authors claim, that the topological invariant (winding number) is predicted "with a high degree of accuracy". However, in Fig. 7(b) the winding number for the CDW is estimated lower than 1, but still very visibly nonzero. Also, the transition to the CDW is not very sharp or clear. This lets me hesitate in readily believing in a successful application to more challenging models
>
> 2) The benefit of using supervised learning does not become completely clear to me. The transition is also predicted via unsupervised learning (transition to CDW). More accurate results could probably also be obtained by using one of the developed unsupervised techniques (see e.g. A. Dawid et al 2020 New J. Phys. 22 or E. van Nieuwenburg et al Nature Phys 13, 435–439 (2017)) The only additional property that is obtained is the winding number - but as stated above, the applicability beyond this model is a bit questionable to me, if I am not misinterpreting the data. Maybe the transition to the CDW is a very special case?
>
> 3) Novelty: Training on exactly solvable models and evaluating on not exactly solvable models is, in contrast to the authors' claim that this goes 'beyond the common scope of machine learning', to my knowledge a commonly used trick, see e.g. Valenti et al, PRR 1(3) (2019). In addition, the method is not applied beyond DMRG results and thus does not yield new physical insights. Furthermore, the authors claim that they use experimentally accessible data in contrast to previous work identifying phase transitions - however e.g. here: Käming, Niklas, et al. "Unsupervised machine learning of topological phase transitions from experimental data." Machine Learning: Science and Technology 2.3 (2021) experimentally accessible data is clearly also used

1) We thank the Referee for the comment: we agree that using the phrase "high degree of accuracy" could be misleading so we changed the text deleting the sentence.

2) The goal of our work was not to prove the benefits of the supervised method versus the unsupervised ones. Our aim was to understand if also supervised models can be trained and tested on different datasets in order to recognize the phase diagram of the interacting model.

3) We apologize for the possible misunderstanding we might have caused and we thank the Referee for giving the opportunity to clarify this point. We had no intention to state that nobody applied machine learning to experimental data (we do in fact mention explicitly [Torlai2018] as an example of experimental approach). In the Introduction we were referring to the use of synthetic data coming from correlation functions and not from Hamiltonian terms as in [Zhang2018] and [Sun2018].

   In addition, we rephrased "This goes beyond the common scope of machine learning" (beginning of Sect. 4) with "This is an approach that has been recently exploited in the context of machine learning applied to systems without analytical solution" adding a reference to Valenti et al, PRR(2019). We also included the reference Käming et al (2021).

> **Requested changes**

1) Is it possible to verify that the learned quantity is really the winding number and not another accompanying trait of this specific model? In particular the nonzero value in the CDW regime and non-sharp transition make me question it. If there is no way of verifying it, an application to a different model could bring evidence and in addition demonstrate, that the manuscript indeed meets the scipost acceptance criteria by 'opening a new pathway in an existing research direction'.

2) I would appreciate a description of the results that is more faithful to the actual results - if I'm not misinterpreting something, the passages 'All the points of the non-topological phases are correctly associated to a zero winding number' and '[...] calculate the winding number [...] with a high degree of accuracy' are just not correct. This request also includes the presentation of 'novelty' (issues addressed in point 3 of weaknesses)

3) An explanation, how the data used for training is experimentally accessible would be a nice addition

4) A minor comment: In Fig. 7, it is a bit confusing that the x-axis of the two subplots is shifted. Would it be possible to align it?

1) We are using the techniques proposed in the manuscript to study the phase diagram of other interesting interacting models, that will be the subject of next papers. In order to convince the Referee that the algorithm can be generalized to different models, we report here the results we obtained by applying it to the SSH model. We generated a training dataset of correlators from the non-interacting Hamiltonian

$$H_{SSH} = -t_1 \sum_{i=0}^{N/2-1} c^{\dagger}_{Ai} c_{Bi} - t_2 \sum_{i=1}^{N/2-1} c^{\dagger}_{Bi} c_{Ai+1} + h.c. \tag{1}$$

where $N = 100$ sites, $c_{A(B)}$ is the destruction operator of the sublattice $A(B)$-type spinless fermion. We tested an ensemble of networks like the ones in the paper on the training set and then created a test set of correlators adding an interaction term $V = \sum_i n_i n_{i+1}$ to the hamiltonian (1).

We plot the results at two different values of the potential in figure 1. We see that one gets a good agreement with the real values of $\tilde{\omega}$, obtained by looking ad the density of edge state.

Let us also notice that, in order to answer the second Referee's comments and to avoid any confusion, we renamed the quantity $\tilde{\omega}$ for the interacting models as *topological indicator*, instead of winding number.

2) We explained more carefully which points are classified with high degree of accuracy addressing the problem of the points close to the phase transition between CDW and TOP in the interacting case. Moreover we removed the sentence "high degree of accuracy" in the abstract and the conclusions.

3) We added two recent and more specific references describing two possible experimental protocols [Naldesi et al. arxiv 2205.00981],[Gluza, Eisert PRL 127, 090503] allowing one to measure one-particle correlators.

4) The image was adjusted as requested.

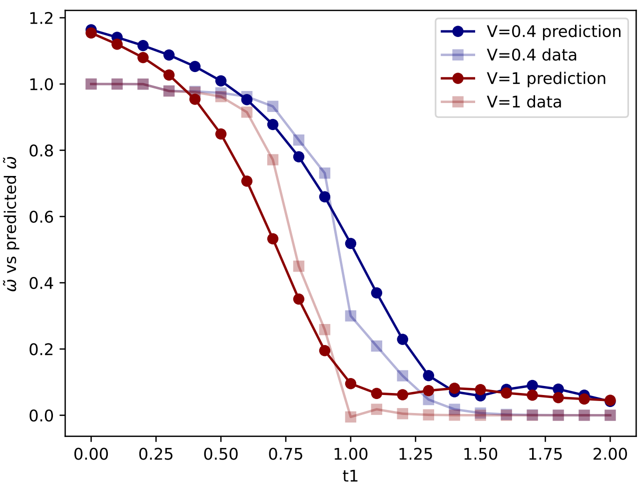

Figure 1: Predictions of the ensemble off networks trained on the non-interacting data of the SSH model and tested on data obtained by turning on an interaction $V = \sum_i n_i n_{i+1}$.