Machine Learning Detection of Majorana Zero Modes from Zero Bias Peak Measurements

Mouyang Cheng[1,2,*], Ryotaro Okabe[1,3], Abhijatmedhi Chotrattanapituk[1,4], and Mingda Li[1,5,]**

[1]Quantum Measurement Group, MIT, Cambridge, MA 02139, USA
[2]School of Physics, Peking University, Beijing 100084, China
[3]Department of Chemistry, MIT, Cambridge, MA 02139, USA
[4]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139, USA
[5]Department of Nuclear Science and Engineering, MIT, Cambridge, MA 02139, USA
[*]e-mail: vipandyc@mit.edu
[**]e-mail: mingda@mit.edu

ABSTRACT

Majorana zero modes (MZMs), emerging as exotic quasiparticles that carry non-Abelian statistics, hold great promise for achieving fault-tolerant topological quantum computation. A key signature of the presence of MZMs is the zero-bias peaks (ZBPs) from tunneling differential conductance. However, the identification of MZMs from ZBPs has faced tremendous challenges, due to the presence of topological trivial states that generate spurious ZBP signals. In this work, we introduce a machine-learning framework that can discern MZM from other signals using ZBP data. Quantum transport simulation from tight-binding models is used to generate the training data, while persistent cohomology analysis confirms the feasibility of classification via machine learning. In particular, even with added data noise, XGBoost classifier reaches 85% accuracy for 1D tunneling conductance data and 94% for 2D data incorporating Zeeman splitting. Tests on prior ZBP experiments show that some data are more likely to originate from MZM than others. Our model offers a quantitative approach to assess MZMs using ZBP data. Furthermore, our results shed light on the use of machine learning on exotic quantum systems with experimental-computational integration.

Introduction

The identification of quantum many-body phases from experimental observations is one of the central tasks in condensed matter physics[1–4]. While symmetry-breaking phases can be detected unequivocally using local order parameters, topological phases of matter pose a more complex problem. Unlike the former, the topological phases cannot be characterized by local order parameters but instead carry global topological invariants[5]. As a result, detecting topological phases often requires an indirect measurement where topology can manifest, such as examining bulk excitations or specific boundary states[6]. Successful examples include the quantum anomalous Hall effect with insulating bulk and spin-polarized chiral edge states that can be probed by electrical transport[7–9], or topological Weyl semimetals with bulk Weyl fermions and surface Fermi arcs using photoemission[10]. In other cases, probing topology can become notably more challenging. In quantum spin liquids, for instance, bulk spinon excitations and edge Majorana fermions only leave subtle experimental evidence[3,11]. An enhanced capability to detect topological phases of matter will greatly enrich our understanding of quantum phases and hold paramount importance for next-generation microelectronic and quantum computing applications.

Among the exotic topological phases of matter, Majorana Zero Modes (MZM), characterized by the non-Abelian, Ising-type anyonic statistics, have captured significant research and industrial attention over the past decade. Thanks to their unique ability to store information nonlocally, and their intrinsic zero energy that guards against hybridization, MZMs are deemed a highly promising platform to realize fault-tolerant topological quantum computation[12–15]. Theoretically, MZMs were first proposed in the Kitaev 1D chain model with p-wave superconductor, where pairs of MZMs can emerge at the ends of the chain[16]. However, the evidence of p-wave superconductors has been elusive, with an unclear pathway to lift the double degeneracy of the spin pairing. Several remedies have been proposed in previous literature[17]. Fu and Kane suggest constructing MZMs using the proximity effect at the interface between an s-wave superconductor (SC) and a topological insulator, which resembles a $p_x + ip_y$ SC with additional time reversal symmetry[18]. Candidates like 5/2 fractional quantum Hall states[19,20] and other platforms[21–26] are also potential candidates for hosting MZMs. Another milestone was reached to construct MZMs on a 1D nanowire with semiconductor (SM) coupled with proximity s-wave SC[27–29]. Under strong Rashba spin-orbit coupling and external Zeeman field, MZMs can emerge from an effective p-wave SC with the double degeneracy lifted. This SM/SC nanowire system has been considered extremely feasible to realize MZMs, with numerous experimental reports demonstrated in the past decade[30–40]. In these cases, the zero biased peaks (ZBPs) of the differential tunneling conductance under the scanning tunneling spectroscopy (STS) provide a strong experimental signature for MZMs[41]. However, there has been a long concern that there are other topologically trivial states that can also produce ZBPs, such as Andreev bound states (ABS), Yu-Shiba-Rusinov states, or simply large disorders[42–46]. To make detection of MZMs even more elusive, it is difficult to actually define what is a Majorana mode in topological superconductors, because there are several low-energy localized states, e.g., the so-called quasi-Majorana states, with intermediate properties between topological Majorana modes and trivial low energy states.[47–51]. Various practical approaches to distinguish MZMs from trivial modes have been proposed, such as topological gap protocol[52,53] and interferometry with floating Majorana islands[54]. However they both rely on more involved operations such as non-local conductance measurements or embedding a Majorana island into an A-B interferometer. Even under such intricated designs, it is still hard to eliminate false positives mimicking MZMs. Therefore, a more direct approach to identify topological MZMs based on solely the traditional experimental ZBP measurements would be highly desirable. And the power of interpretation of machine learning could facilitate such a task, which is beyond the imagination of conventional protocols based on principles of physics.

In this work, we develop a machine-learning pipeline that aims to differentiate topological MZM from other topologically trivial states using experimental ZBP signals. The primary obstacles are the scarcity of experimental data and the absence of a universally acknowledged MZM ground truth. However, thanks to the STS technique, which can provide direct access to the single-particle density-of-states and further enables quantitative comparisons between experiments and computations, we were able to generate the ZBP training data computationally. Using effective Hamiltonian and quantum transport simulations, we cover a broad spectrum of physical parameters and mechanisms and further add data noises to mimic experiments. Although distinguishing MZM has created challenges due to the spectral similarity of ZBP between topological MZM and topologically trivial states, from a machine-learning perspective, this complexity is transformed into a classification task. Persistent cohomology analysis shows that the hidden global features of different topological classes remain robust, indicating that such a classification task is fundamentally machine-classifiable. By further implementing various machine-learning methods, such as linear classifiers, convolutional neural networks, and XGBoost, excellent accuracy is finally reached even with a reasonable level of data noise. We carry out additional tests on the experimental ZBP data from existing literature and found that some ZBP data are more likely to arise from MZM, while others are not. This does not rule out the potential presence of MZM in any of the reported experimental systems, given the limitation of the effective Hamiltonian approach and other experimental complexities not considered in this work. Our model offers an attempt to solve the MZM detection problem with machine learning. The work can also shed light on the application of machine learning in other exotic many-body quantum systems with very limited training data and a lack of ground truth.

Results

Model setup

The general machine learning workflow is shown in Fig. 1. We consider the popular 1D SC/SM nanowire discussed earlier as the modeled system. The pristine nanowire system can be described by the 1D Boguliubov-de-Gennes (BdG) Hamiltonian following the s-wave pairing Oreg-Lutchyn model $H = \frac{1}{2}\int \Psi^\dagger(x) H_{\text{tot}} \Psi(x) dx$[26–28], where

$$\begin{aligned} H_{\text{tot}} &= T + H_{\text{soc}} + U + H_Z + H_{\text{couple}} \\ &= \left(-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial x^2} - i\alpha \frac{\partial}{\partial x}\sigma_y - \mu \right)\tau_z + E_Z \sigma_x + \Delta \tau_x. \end{aligned} \tag{1}$$

Here, $\hat{\Psi}(x) = \left(\hat{\psi}_\uparrow(x), \hat{\psi}_\downarrow(x), \hat{\psi}^\dagger_\downarrow(x), \hat{\psi}^\dagger_\uparrow(x) \right)^T$ spans a Nambu space with four spinors, and $\vec{\sigma}$ and $\vec{\tau}$ stand for Pauli matrices in the spin and particle-hole space, respectively. The five terms $T, H_{\text{soc}}, U, H_Z$ and H_{couple} denote the kinetic energy, spin-orbit coupling, on-site potential, Zeeman coupling, and the SC-SM coupling, respectively. Detailed information about the choice of parameters is shown in Supplementary Information 1.

The Hamiltonian Eq. 1 is the pristine Hamiltonian that leads to MZM. We further apply weak diagonal disorder $V_{\text{imp}}(x) \sim \varepsilon N(0,1)$ sampled from a normal distribution to mimic the noise but without destroying the topology. In real experiments, trivial ZBPs may arise from a non-ideal potential landscape on the nanowire. At least two scenarios can lead to topologically trivial states, including **I.** quantum dots located at ends of the nanowire and **II.** large fluctuating disorder spread on the whole nanowire[44]. Therefore, for the topological trivial classes without MZM, we construct the Hamiltonian in two ways: For scenario **I.**, we add a Gaussian potential as an incommensurate on-site perturbation to the diagonal Hamiltonian. It has been shown that such smearing potential could be the culprit to the ABS. When the two Andreev bias peaks come closer under tuned parameters, these peaks will merge and form a trivial ZBP[55,56]; For scenario **II.**, we amplify the disorder strength so that the fluctuation energy is comparable to the original chemical potential μ. This can also give rise to topologically trivial states with ZBPs, creating a challenge for the MZM identification[44]. Overall, to justify our approach we also include an extra check on whether smearing potential or large disorder will generally create trivial modes with observable ZBPs in our dataset (see Supplementary Information 1).

To generate the training data for machine learning, we cover a wide range of input Hamiltonian parameters (see Supplementary Information 1). The continuous Hamiltonian is discretized in real space to a finite tight-binding matrix. Then, we perform tight-binding simulations on this discretized system to calculate the tunneling conductance $G = \mathrm{d}I/\mathrm{d}V$ via the S matrix formalism (see Methods for more details). A total of 12,000 labeled Hamiltonians are generated, with 4,000 for topological MZM, 4,000 for trivial ABS, and 4,000 for trivial large disordered states. The tunneling conductance signal can thus be calculated under sweeping a 2D parameter space composed of bias voltage V_{bias} and Zeeman splitting E_Z, each with 28 different values. This leads to the use of 28×28 image to represent the tunneling conductance data, labeled by either topological (hosting MZM) or trivial states (either ABS or large disorders) for machine learning classification. In addition, since some experimental works focus on 1D dI/dV data without sweeping the Zeeman splitting, we single out the 1D data with zero Zeeman splitting for additional training. This can be done by searching the ZBPs while sweeping through E_Z horizontally. Lastly, to improve the training robustness and bridge the theoretical-experimental gap, we perform pre-processing on the raw data, including Gaussian smearing, additional noise, anomaly detection on the dataset. We can also refine our workflow by carrying out a post selection on the dataset to ensure the topological/trivial labels are assigned properly, and the ZBPs are generally present in our dataset. More details on the Hamiltonian model, data generation, processing and selection can be found in Methods and Supplementary Information 1.

Global pattern with topological data analysis

We first display typical tunneling conductance dI/dV data generated from the workflow above in Fig. 2(b) for topological MZM and trivial classes (see Supplementary Information 2 for more examples). It can be seen that the 2D dI/dV data from topological and trivial states have similar patterns. One earlier approach to achieve MZM pattern recognition[44] is finding the phase boundary between the topological and trivial classes. By pointing out the difference in the position of the topological phase transition compared to the pristine data, it was concluded that quantum dots and large disorder destroy the topology of the system, thus creating trivial ZBPs. However, this approach is performed with fixed Hamiltonian parameters; when the parameters are unknown, discerning the topological MZM phase is still challenging for human eyes.

To investigate the potential intrinsic separability between the topological MZM and trivial classes, we employ the persistent cohomology analysis on a portion of the training dataset for all classes. Persistent cohomology is a type of topological data analysis (TDA) that studies the global feature difference at various scales. Figure 2(a) shows an example of persistent cohomology analysis on simplified 2D data. Starting from a gray image, a threshold value is tuned from the lowest pixel value to the highest. For a given threshold value, each pixel can be masked to binary black/white (lower/higher than threshold). Then two topological features emerge: Feature 0 identifies isolated black clusters in data (partially marked with light blue); Feature 1 focuses on closed loops encircled by a black cluster (partially filled with red). By sweeping the threshold values, different patterns assigned with different features emerge and annihilate, which create a birth-death scattered plot[57]. Therefore, persistent cohomology provides insights into the robustness and significance of these topological characteristics in the data.

Our analysis involves the 3D data composed of Bias voltage, Zeeman splitting, and other Hamiltonian parameters as one dimension. As a result, there is additional Feature 2 which captures voids or cavities entirely enclosed by surfaces. The persistent cohomology analysis is performed on our datasets using the GUDHI package with cubical complex[58]. Results are shown in Fig. 2(c, d), where the difference between the topological MZM class and the trivial class can be seen clearly. Taking Feature 0 as an example; on the one hand, for the topological MZM dataset, there are very few clusters (light blue) that emerge near zero birth and annihilate early. On the other hand, for the trivial dataset, there is a continuous distribution of clusters that creates at zero birth and annihilates. Orange and Green ovals marked in Fig. 2(d) clearly highlights such distinct topological feature difference between topological and trivial data. The results indicate that though individual conductance data shown in Fig. 2(b) could be hard to classify by human eye, TDA analysis can show the relationships for more than one Hamiltonian parameters for each data class, giving us crucial information on the connectivity of varying similar Hamiltonians. Such collective information allows us to spot feature differences and feel confident that such binary classification task is machine-separable. More detailed insights brought by persistent cohomology analysis are shown in Supplementary Information 1.

Therefore, the persistent cohomology analysis implies that although the human eyes cannot readily distinguish the topological MZM states from trivial states, there exists a global topological feature difference between them. Such difference builds confidence that the MZM classification problem with ZBP is machine-separable prior to any design of machine-learning models.

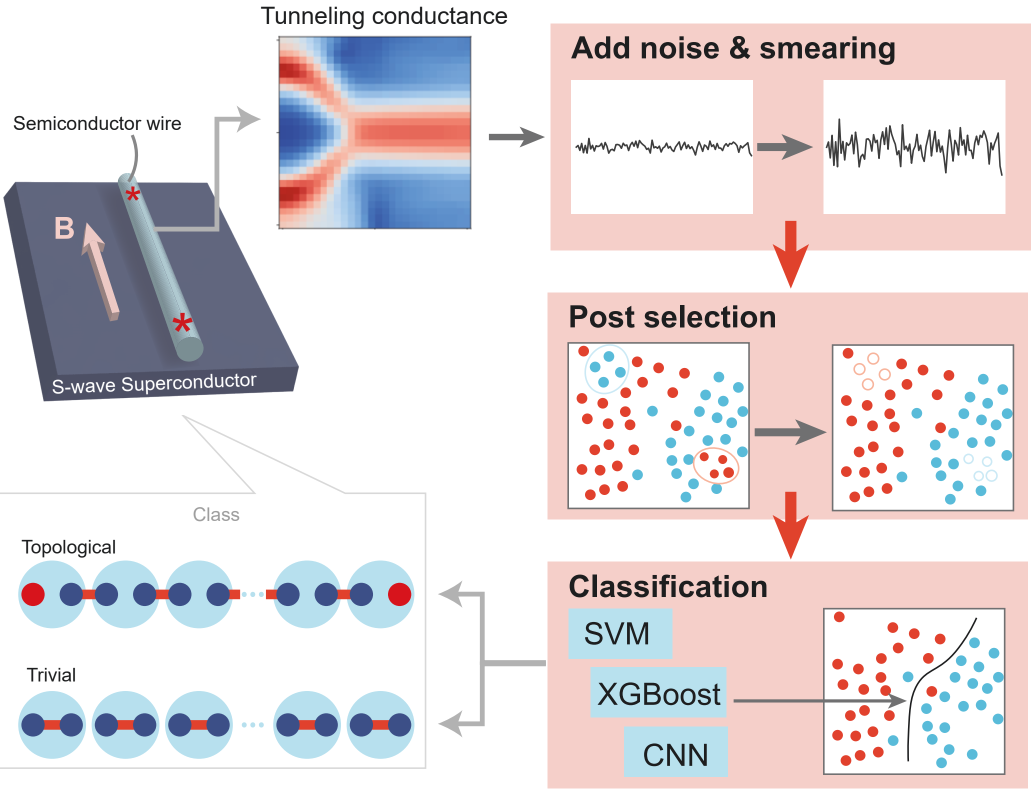

Figure 1. The overview of the machine learning workflow to detect Majorana zero modes (MZMs) from zero bias peaks (ZBPs). The system consists of a 1D semiconducting nanowire coupled in proximity with an *s*-wave superconductor, which resembles a 1D *p*-wave superconducting Hamiltonian under a parallel magnetic field B. Training data are generated by an effective Hamiltonian approach. By modifying the on-site potential landscape, states that host topological MZMs and topologically trivial states are generated and labeled by the topological class. The tunneling conductance dI/dV signals from the scanning tunneling spectroscopy are further computed using the tight-binding and quantum transport approach, which are used as input data. Optionally, noise and smearing are added to the dataset to better mimick experimental data, and a post selection is performed to ensure the topological/trivial labels are assigned appropriately. Various machine-learning models are established to achieve the MZM classification, with additional tests performed using existing experimental data.

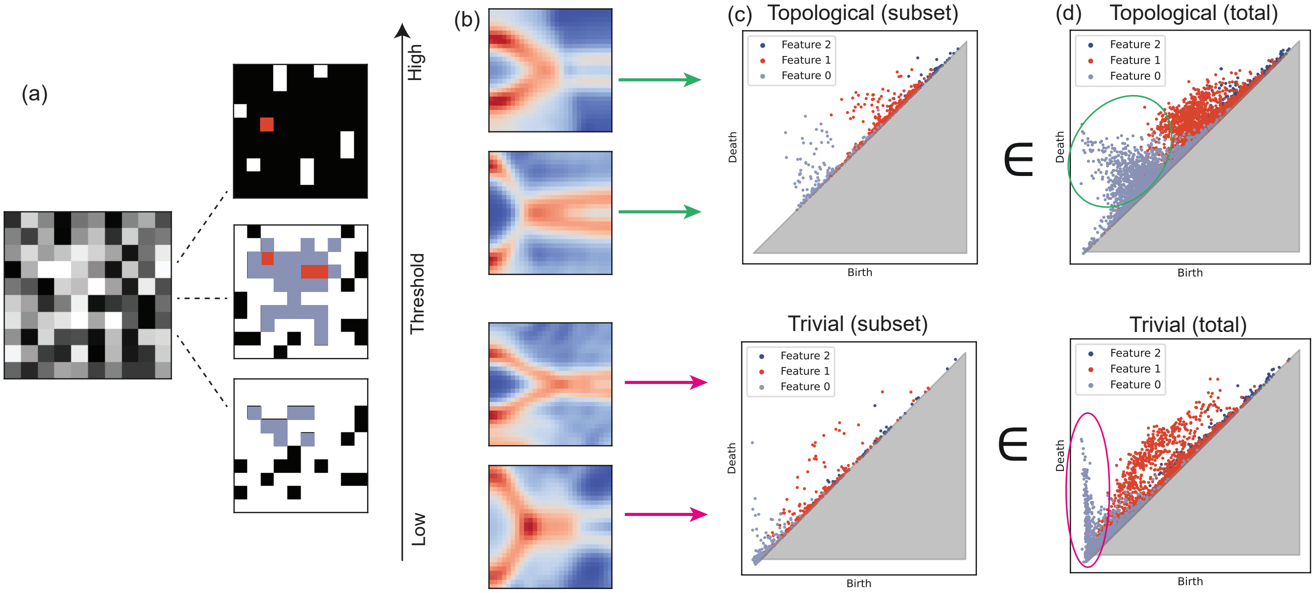

Figure 2. Persistent cohomology analysis with the training data. (a) Schematic of the principles on persistent cohomology, using simplified 2D data as an example. Given a fixed threshold value, each original pixel can be masked to binary black or white. Isolated black clusters and white loops are highlighted out with blue and red colors as key topological features. As the masking threshold is tuned continuously from minimum to maximum magnitudes, these distinct topological features emerge and die out. Only features near a centered area are colored for better visualization. **(b)** Typical computationally generated tunneling conductance data used for machine learning training, for topological MZM and trivial classes. The 2D heatmap plots are tunneling conductance dI/dV as a function of bias voltage V_{bias} and Zeeman energy E_z; **(c, d)** Topological data analysis on topological MZM and trivial classes, respectively, using persistent cohomology analysis. Panel (c) shows results for a small random subset containing conductance data shown in (b), where a distinct feature difference can be seen; Panel (d) shows the full results for the whole topological/trivial dataset. Green and orange ovals highlight the key feature difference between these two classes. Although the individual raw data in **(b)** are barely distinguishable with bare eyes, an obvious difference is shown between the topological MZM class and the other topological trivial classes through topological data analysis.

Machine learning results

We employ a few machine-learning models to perform the topological MZM classification task. For the model inputs, 2D data of tunneling conductance images with 28×28 pixels are flattened into 1D arrays, except for Convolutional Neural Network (CNN) which directly receives the 2D data. As a baseline check, we first perform linear Principal Component Analysis (PCA) analysis to compress the data dimension. We reduce the 2D and 1D datasets' complexity to 2 dimensions for better visualization, and the reduced result with labels 0 or 1 are shown in Fig. 3(a, e). On the scattered plot for the first two leading principal components, there is no clear boundary between two separated clusters with different labels. Fig. 3(a, e) shows a linear Support Vector Machine (SVM) boundary line that separates two regions (shaded blue and red). However, there is a notable portion of data points crossing the boundary, indicating the limited power of linear classification at least on the PCA dimensionality-reduced dataset (performance shown in Fig. 3(b, f)). Particularly, for 1D PCA, the prediction of data labeled topological with 0.47 accuracy is close to random guess. Further attempts to use linear methods consistently provide lower accuracy than 90% (see Supplementary Information 3), indicating the intrinsic data nonlinearity and calling on the necessity of nonlinear machine learning methods.

We carry out non-linear classification methods and ensemble methods including kernel-SVM, Random forest, CNN, and Extreme Gradient Boosting (XGBoost). The results as well as hyperparameters tuning process are described in Supplementary Information 3. Among them, XGBoost, which combines ensemble models and improved gradient boosting, gives overall better performance than other methods for both 1D and 2D tasks. The confusion matrix results for XGBoost training are shown in Fig. 3(c, d, g, h) as for 2D and 1D data, with and without data noise, respectively. It is worthwhile mentioning that binary classification with 2D tunneling conductance data for topological MZM class reaches $\sim 94\%$ accuracy, even in the presence of data noise. Additionally, although the 1D classifier gives a $\sim 28\%$ false positive for the topological MZM class, it still gives a high, 95% confidence in true positive, and the overall accuracy still reaches a 85%. While adding noise reduces the accuracy of identifying trivial classes, it significantly improves the performance of detecting trivial classes from $\sim 72\%$ to $\sim 86\%$, which may be attributed to the large data variance and better data generalization.

The success in machine learning classification agrees well with the persistent cohomology observation. Also, the introduction of the Zeeman energy sweeping in 2D data outperforms the 1D data, indicating the benefit and possible necessity to take data with sweeping Zeeman energy. Additionally, we also test the capability of our model for multi-classification. Fig. 3(i, j) shows results for ternary classification results for the most accurate XGBoost model. It shows that even for ternary classification, the model can still perform at $> 90\%$ accuracy with at most 11% false positive rate for the Andreev class.

We also evaluate the robustness of our model by testing it against untrained categories of trivial disordered data, such as nanowires with large disorder in g factor and superconducting gap Δ. We claim that the model has roughly $\sim 20\%$ false-positive rate for these trivial testing data, indicating that it is still moderately robust(See Supplementary Information 3).

Experimental tests

For the final part, we use our trained classifiers on real experimental ZBP data from recent literature. Since our classifier with 2D data input gives overall higher accuracy than the 1D classifier, we focus on the tests on the 2D ZBP data testing. Additional 1D data sets are shown in Supplementary Information 3. We extract 16 ZBP data images from 10 references during the past decade[31–40]. The images are cropped online and processed to fit properly within our model input format (see Supplementary Information 4 for more details). Since XGBoost returns the continuous probability $p \in [0,1]$ before the final binary classification, here we show the probability since it carries more information than binary value, with a cutoff value $p_{\text{crit}} = 0.5$. The positive result probability, i.e., the probability that the model suggests that the system hosts topologically MZMs, for the test set is shown in Fig. 4(a). Here we only emphasize the examples that manage to pass the trial test either with or without noise in the figure.

Four experimental samples from four prior works pass the test from the 2D model either with or without data noise. The pattern of these samples are show in Fig. 4(b) from sample 1 to sample 4 in order[31, 34–36]. We also included a prediction result for another XGBoost model with noise, but trained on a larger dataset including larger Zeeman energy window, i.e. larger max range $E_z = 4.48$meV. The results indicate that our model prediction is somewhat robust against the choice of energy window (See details in Supplementary Information 1). Among them, the most robust sample, upon which both models with and without noise imply positive MZM presence, has been retracted[34]. For the other 12 samples, the predicted probability for the existence of MZMs always lies consistently below 0.5, indicating that those systems are unlikely to host MZMs. The complete test results are shown in Supplementary Information 4. It is also worth mentioning that due to the moderate false-positive rate for untrained disordered catagories of our model, even if an example passes the filter of our model, it might still arise from disordered landscapes beyond our model's consideration. Thus, we comment that the actual number of topological MZMs is likely even less than the 4 out of 16 candidates in Fig. 4. Overall, our model predicts that a dominant portion of experimental measurements is unlikely to host MZMs on SM-SC coupling nanowires.

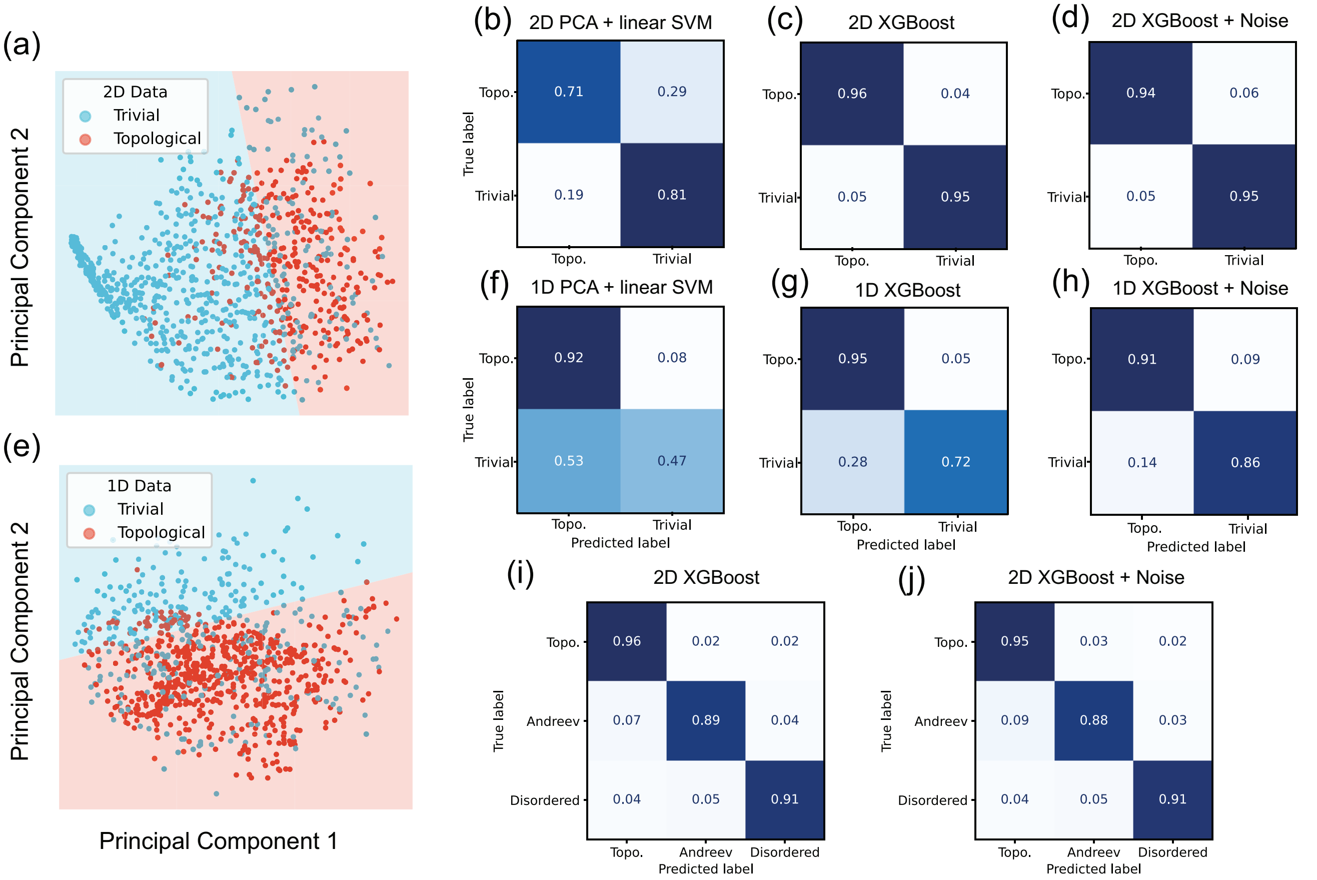

Figure 3. Machine learning classifications to identify the topological MZMs using 1D and 2D tunneling conductance data. (a, e) PCA analysis on the generated 2D (a) and 1D (e) data projected on the first two principal components. The SVM linear boundary roughly separates the topological MZM and trivial classes. **(b, f)** Confusion matrices for PCA + linear SVM learning results for 2D (b) and 1D (f) training data. **(c, g)** Confusion matrices from XGBoost for 2D (c) and 1D (g) training model without noise. **(d, h)** Confusion matrices from XGBoost for 2D (c) and 1D (g) training model with added data noise. Note that in all cases, the model with 2D data outperforms the model with 1D data, indicating the advantage of collect data with Zeeman energy sweeping. **(i, j)** Ternary classification confusion matrices from XGBoost for 2D training model with/without added data noise.

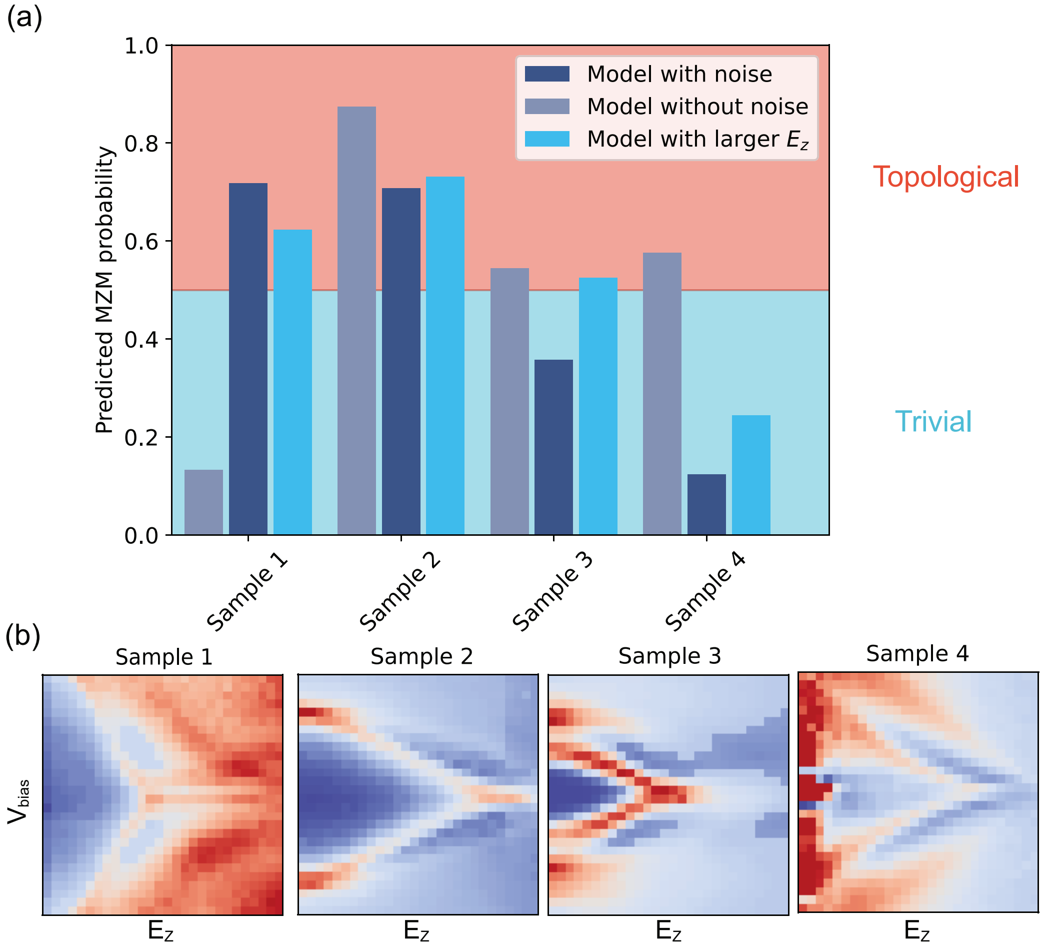

Figure 4. Tests on experimental 2D ZBP data based on our machine learning models. (a) The probability of positive prediction for the XGBoost model is plotted as a histogram. Prediction results from the 2D XGBoost model with/without noise, and the 2D XGBoost model trained on dataset including larger $E_z = 4.48$meV range are shown for more information. Only 4 results labeled from sample 1 to sample 4[31, 34–36] imply that the system possibly hosts real MZM, while all the others suggest that they are unlikely to host MZMs. **(b)** Patterns of tunneling signal for sample 1 to sample 4 that pass the test.

Discussion

In this work, we propose a machine-learning pipeline to detect MZMs in experimentally measured differential tunneling conductance signals. Our work constructively aligns with the recent efforts to identify the topological MZMs from trivial states, replacing human eyes with machine-learning-based visual aids. It offers a few potential advantages, including less bias and the possibility to quantify the performance.

It is important to note that our model is only valid under a number of assumptions. It assumes that **I.** the experimental nanowire system can be well described within the 1D *s*-wave Oreg-Lutchyn model framework; **II.** the physical mechanisms for impurity and disorder can be mimicked by modifying the diagonal potential landscape, and they are the only false positive sources for misleading ZBPs; **III.** the finite temperature effect can be modified by Gaussian smearing (see Supplementary Information 1).

To summarize, though our work is still limited by various assumptions and false positives, compared to other more complicated protocols where more sophisticated measurements are required, our work offers the first framework as a machine learning attempt to identify MZMs directly from experimentally measured ZBP signals. Our classifier model could easily be generalized to suit other quantum property predictions, as long as the system can be well captured by effective model Hamiltonians. In the context of methodology, our machine learning model uses a mean-field approximation to capture the topological MZM feature under the condition of suppressed quantum fluctuations. This approach, to a broader aspect, could inspire more machine learning works integrated with experiments to tackle strongly correlated systems as a starting point. The model could also be further generalized to conduct parameters extraction on the experimental STS data, which resembles the philosophy of other machine-learning parameter extraction from experimental data such as time-resolved diffraction[59] or neutron scattering[60]. We also note a recent work extracting disorder landscape of SM/SC nanowire with machine learning[61].

Methods

Tunneling conductance simulation

The training data of our work is generated by tight-binding simulation on transport properties using the KWANT package[62]. To calculate the scattering matrix, we attach a normal SM nanowire with a lead to the end of the nanowire. The normal SM nanowire has the same form of Hamiltonian as the SC/SM system except for the SC coupling, i.e.

$$\begin{aligned} H_{\text{normal}} &= T + H_{\text{soc}} + U + H_Z \\ &= \left(-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial x^2} - i\alpha\frac{\partial}{\partial x}\sigma_y - \mu_{\text{normal}}\right)\tau_z + E_Z\sigma_x. \end{aligned}$$

Note that there is a finite difference between the normal wire and the SC/SM nanowire in a chemical potential $\mu_{\text{normal}} - \mu = eV_{\text{gate}}$, which represents the gate voltage added to the scattering region. As for the lead, the on-site Hamiltonian is the same as the normal nanowire except for an additional potential barrier V_{barrier}:

$$H_{\text{normal}} = \left(-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial x^2} - i\alpha\frac{\partial}{\partial x}\sigma_y - \mu_{\text{normal}} + V_{\text{barrier}}\right)\tau_z + E_Z\sigma_x.$$

All relevant physical parameters in the Hamiltonian can be found listed in Supplementary Information 1. After constructing such a system, KWANT allows convenient calculation on the scattering matrix S on the defined scattering region, i.e., the connecting junction on the lead.

$$G_0(E) = 2 + \sum_{\sigma,\sigma'=\uparrow,\downarrow}\left(\left|r_{eh}^{\sigma\sigma'}\right|^2 - \left|r_{ee}^{\sigma\sigma'}\right|^2\right),$$

where r_{eh} and r_{ee} are the Andreev and normal reflection amplitudes from the S matrix, respectively. The calculated tunneling conductance is energy-dependent, and by sweeping the Zeeman energy E_Z, we can obtain a diagram with dI/dV versus E_Z and bias energy(voltage) V_{bias}, which finally gives an image of our 2D data. Such numerical method is extensively performed in the relevant area of literature, and we refer readers to references like[43,44,63,64] for more details.

Data processing before training

After generating these raw data, we add Gaussian smearing by adding Gaussian function convolution to our 2D image:

$$F(G) \sim \exp(-G^2/2\sigma^2)$$

where $\sigma = 1\text{pixel}^{-1}$. The reason for such processing are of two folds: First, such Gaussian smearing mimics the finite temperature effect of experimental measurements based on our zero-temperature simulation (see Supplementary Information 1); Secondly, our smearing also smooths out the experimental STS measurement signal, mimicking the resolution function resembling the Gaussian convolution.

In addition to such smearing, to ensure the robustness of our model and emulate the measurement noise we further add a small noise to the tunneling conductance signal subject to the normal distribution $\delta G \sim 0.2N(0,1)e^2/h$.

Machine learning

Each machine learning (ML) model, depending on the design sophistication, is more suitable for some type of problems than the others. However, especially for problems like MZM detection which are not well explored through ML perspective, it is better to start approaching the problem with multiple model architectures. This method would not only provide us with the best model for the job, but one can also later utilize multiple models for a better performance through boosting technique. With that, in our study, we perform in total five ML architectures: Principle Component Analysis (PCA), Support Vector Machine (SVM), Random Forest, eXtreme Gradient Boosting (XGBoost), and Convolutional Neural Network (CNN). All of these essentially have the same frame work in which they are trying to search for classifying criteria that divide the input data high-dimensional space (28×28) into regions of trivial, and topological labeling.

In PCA, the input got basis transform such that the first basis (principle component) represents the axis in which there is the largest variation of data among the input data, i.e., a projection of data that separates the data by greatest distance. The second component is the same as the first but for the remaining dimensions of the data space. The subsequent components follow the same idea recursively. We then keep the first two components which capture two largest main features present in the inputs. Of course, as can be seen from our result, the principle component that define the classification of the data need not be one of the first two. In fact, there might not be one if the relation is not linear between the input and the MZM class. After that, one can utilize any model to try to subdivide the regions into their respective labels. In this work, we use the linear version of SVM which will be described next.

For SVM, the model focuses on finding the hyperplane in input space that can separate trivial, and topological data points. The method is actually used in the last step of PCA. However, one can also directly apply it to the high-dimensional input space directly. Furthermore, it is also common to transform the input with a predetermine non-linear maps, called kernel method, which increase the potential of the method by allowing the non-linear hyperplane.

Both random forest, and XGBoost methods use ensemble of multiple decision tree models. Each decision tree is a collection of hyperplanes that are mostly perpendicular to the input space axes. Hence, each hyperplane is not as powerful as the one from SVM. However, by having multiple of these, the collection can separate the spaces into many regions which, if fine enough, can separate the data points into trivial, and topological accurately. The difference between these two are that random forest performs majority voting between the trees, i.e., each tree performs the same task, while XGBoost manages each tree to perform the smaller task that is the weakness of other trees.

Lastly, CNN model takes advantage of the input being images in which the defining feature of MZM should be captured with some local patterns inside the images. This means that there should be local correlation between near by pixels as well as invariant to translation. Hence, it suggests that the classifying criteria should be in the form of convolution between the input image, and a collection of patterns (kernels). From that, the model need to find the kernel that only appear in the real MZM data.

The methods and models mentioned are well implemented in Python open-source packages. We use the scikit-learn package[65] for PCA analysis, SVM, random forest, and XGBoost classification, and we implement the Pytorch[66] package for building up the simple CNN network for classification. We also include the model and hyperparameter settings in Supplementary Information 3.

Author Contributions

M.L conceived and supervised the project. M.C built the workflow, performed transport calculations to generate training data and performed machine learning and topological data analysis. R.O and A.C interpreted key results in main text figures, and helped machine learning analysis. All authors took part in preparing the manuscript.

Data and code availability

The data used in this study are numerically generated using our code implementing KWANT, and the code used in this study is available at https://github.com/vipandyc/ML_majorana.

Declaration of Interests

The authors declare no competing interests.

Acknowledgements

The authors thank J Sau, B November, and J Hoffman for the helpful discussions. RO and AC acknowledge support from the US Department of Energy (DOE), Office of Science (SC), Basic Energy Sciences (BES), Award No. DE-SC0021940, and National Science Foundation (NSF) Designing Materials to Revolutionize and Engineer our Future (DMREF) Program with Award No. DMR-2118448. AC is partially supported by the MIT-QEC Seed Fund. ML acknowledges the support from the Class of 1947 Career Development Chair and support from R Wachnik.

References

1. Wang, Y., Wu, H., McCandless, G. T., Chan, J. Y. & Ali, M. N. Quantum states and intertwining phases in kagome materials. *Nat. Rev. Phys* (2023).

2. Zhou, X. *et al.* High-temperature superconductivity. *Nat. Rev. Phys.* **3**, 462–465 (2021).

3. Wen, J., Yu, S.-L., Li, S., Yu, W. & Li, J.-X. Experimental identification of quantum spin liquids. *npj Quantum Mater.* **4**, 12 (2019).

4. von Klitzing, K. *et al.* 40 years of the quantum hall effect. *Nat. Rev. Phys.* **2**, 397–401 (2020).

5. Wen, X.-G. Colloquium: Zoo of quantum-topological phases of matter. *Rev. Mod. Phys.* **89**, 041004 (2017).

6. Qi, X.-L. & Zhang, S.-C. Topological insulators and superconductors. *Rev. Mod. Phys.* **83**, 1057 (2011).

7. Yu, R. *et al.* Quantized anomalous hall effect in magnetic topological insulators. *science* **329**, 61–64 (2010).

8. Chang, C.-Z. *et al.* Experimental observation of the quantum anomalous hall effect in a magnetic topological insulator. *Science* **340**, 167–170 (2013).

9. Deng, Y. *et al.* Quantum anomalous hall effect in intrinsic magnetic topological insulator mnbi2te4. *Science* **367**, 895–900 (2020).

10. Armitage, N., Mele, E. & Vishwanath, A. Weyl and dirac semimetals in three-dimensional solids. *Rev. Mod. Phys.* **90**, 015001 (2018).

11. Zhou, Y., Kanoda, K. & Ng, T.-K. Quantum spin liquid states. *Rev. Mod. Phys.* **89**, 025003 (2017).

12. Nayak, C., Simon, S. H., Stern, A., Freedman, M. & Sarma, S. D. Non-abelian anyons and topological quantum computation. *Rev. Mod. Phys.* **80**, 1083 (2008).

13. Kitaev, A. Anyons in an exactly solved model and beyond. *Annals Phys.* **321**, 2–111 (2006).

14. Alicea, J., Oreg, Y., Refael, G., Von Oppen, F. & Fisher, M. P. Non-abelian statistics and topological quantum information processing in 1d wire networks. *Nat. Phys.* **7**, 412–417 (2011).

15. Marra, P. Majorana nanowires for topological quantum computation. *J. Appl. Phys.* **132** (2022).

16. Kitaev, A. Y. Unpaired majorana fermions in quantum wires. *Physics-uspekhi* **44**, 131 (2001).

17. Flensberg, K., von Oppen, F. & Stern, A. Engineered platforms for topological superconductivity and majorana zero modes. *Nat. Rev. Mater.* **6**, 944–958 (2021).

18. Fu, L. & Kane, C. L. Superconducting proximity effect and majorana fermions at the surface of a topological insulator. *Phys. review letters* **100**, 096407 (2008).

19. Read, N. & Green, D. Paired states of fermions in two dimensions with breaking of parity and time-reversal symmetries and the fractional quantum hall effect. *Phys. Rev. B* **61**, 10267 (2000).

20. Moore, G. & Read, N. Nonabelions in the fractional quantum hall effect. *Nucl. Phys. B* **360**, 362–396 (1991).

21. Linder, J., Tanaka, Y., Yokoyama, T., Sudbø, A. & Nagaosa, N. Unconventional superconductivity on a topological insulator. *Phys. Rev. Lett.* **104**, 067001 (2010).

22. Ghosh, P., Sau, J. D., Tewari, S. & Sarma, S. D. Non-abelian topological order in noncentrosymmetric superconductors with broken time-reversal symmetry. *Phys. Rev. B* **82**, 184525 (2010).

23. Alicea, J. Majorana fermions in a tunable semiconductor device. *Phys. Rev. B* **81**, 125318 (2010).

24. Qi, X.-L., Hughes, T. L. & Zhang, S.-C. Chiral topological superconductor from the quantum hall state. *Phys. Rev. B* **82**, 184516 (2010).

25. Sato, M. & Fujimoto, S. Topological phases of noncentrosymmetric superconductors: Edge states, majorana fermions, and non-abelian statistics. *Phys. Rev. B* **79**, 094504 (2009).

26. Sau, J. D., Lutchyn, R. M., Tewari, S. & Sarma, S. D. Generic new platform for topological quantum computation using semiconductor heterostructures. *Phys. review letters* **104**, 040502 (2010).

27. Lutchyn, R. M., Sau, J. D. & Sarma, S. D. Majorana fermions and a topological phase transition in semiconductor-superconductor heterostructures. *Phys. review letters* **105**, 077001 (2010).

28. Oreg, Y., Refael, G. & Von Oppen, F. Helical liquids and majorana bound states in quantum wires. *Phys. review letters* **105**, 177002 (2010).

29. Lutchyn, R. M. *et al.* Majorana zero modes in superconductor–semiconductor heterostructures. *Nat. Rev. Mater.* **3**, 52–68 (2018).

30. Das, A. *et al.* Zero-bias peaks and splitting in an al–inas nanowire topological superconductor as a signature of majorana fermions. *Nat. Phys.* **8**, 887–895 (2012).

31. Gül, Ö. *et al.* Ballistic majorana nanowire devices. *Nat. nanotechnology* **13**, 192–197 (2018).

32. Yu, P. *et al.* Non-majorana states yield nearly quantized conductance in proximatized nanowires. *Nat. Phys.* **17**, 482–488 (2021).

33. Deng, M. *et al.* Parity independence of the zero-bias conductance peak in a nanowire based topological superconductor-quantum dot hybrid device. *Sci. reports* **4**, 7261 (2014).

34. Zhang, H. *et al.* Retracted article: Quantized majorana conductance. *Nature* **556**, 74–79 (2018).

35. Deng, M.-T. *et al.* Nonlocality of majorana modes in hybrid nanowires. *Phys. Rev. B* **98**, 085125 (2018).

36. Nichele, F. *et al.* Scaling of majorana zero-bias conductance peaks. *Phys. review letters* **119**, 136803 (2017).

37. Chen, J. *et al.* Ubiquitous non-majorana zero-bias conductance peaks in nanowire devices. *Phys. review letters* **123**, 107703 (2019).

38. Chen, J. *et al.* Experimental phase diagram of zero-bias conductance peaks in superconductor/semiconductor nanowire devices. *Sci. advances* **3**, e1701476 (2017).

39. Mourik, V. *et al.* Signatures of majorana fermions in hybrid superconductor-semiconductor nanowire devices. *Science* **336**, 1003–1007 (2012).

40. Deng, M. *et al.* Majorana bound state in a coupled quantum-dot hybrid-nanowire system. *Science* **354**, 1557–1562 (2016).

41. Jäck, B., Xie, Y. & Yazdani, A. Detecting and distinguishing majorana zero modes with the scanning tunnelling microscope. *Nat. Rev. Phys.* **3**, 541–554 (2021).

42. Yin, J.-X. *et al.* Observation of a robust zero-energy bound state in iron-based superconductor fe (te, se). *Nat. Phys.* **11**, 543–546 (2015).

43. Pan, H., Cole, W. S., Sau, J. D. & Sarma, S. D. Generic quantized zero-bias conductance peaks in superconductor-semiconductor hybrid structures. *Phys. Rev. B* **101**, 024506 (2020).

44. Pan, H. & Das Sarma, S. Physical mechanisms for zero-bias conductance peaks in majorana nanowires. *Phys. Rev. Res.* **2**, 013377 (2020).

45. Frolov, S., Manfra, M. & Sau, J. Topological superconductivity in hybrid devices. *Nat. Phys.* **16**, 718–724 (2020).

46. Valentini, M. *et al.* Flux-tunable andreev bound states in hybrid full-shell nanowires. In *APS March Meeting Abstracts*, vol. 2021, X58–002 (2021).

47. Prada, E. *et al.* From andreev to majorana bound states in hybrid superconductor–semiconductor nanowires. *Nat. Rev. Phys.* **2**, 575–594 (2020).

48. Vuik, A., Nijholt, B., Akhmerov, A. R. & Wimmer, M. Reproducing topological properties with quasi-majorana states. *SciPost Phys.* **7**, 061 (2019).

49. Avila, J., Peñaranda, F., Prada, E., San-Jose, P. & Aguado, R. Non-hermitian topology as a unifying framework for the andreev versus majorana states controversy. *Commun. Phys.* **2**, 133 (2019).

50. Yu, P. *et al.* Non-majorana states yield nearly quantized conductance in proximatized nanowires. *Nat. Phys.* **17**, 482–488 (2021).

51. Marra, P. & Nigro, A. Majorana/andreev crossover and the fate of the topological phase transition in inhomogeneous nanowires. *J. Physics: Condens. Matter* **34**, 124001 (2022).

52. Pikulin, D. I. *et al.* Protocol to identify a topological superconducting phase in a three-terminal device. *arXiv preprint arXiv:2103.12217* (2021).

53. Aghaee, M. *et al.* Inas-al hybrid devices passing the topological gap protocol. *Phys. Rev. B* **107**, 245423 (2023).

54. Whiticar, A. M. *et al.* Coherent transport through a majorana island in an aharonov–bohm interferometer. *Nat. communications* **11**, 3212 (2020).

55. Stanescu, T. D. & Tewari, S. Robust low-energy andreev bound states in semiconductor-superconductor structures: Importance of partial separation of component majorana bound states. *Phys. Rev. B* **100**, 155429 (2019).

56. Liu, C.-X., Sau, J. D., Stanescu, T. D. & Sarma, S. D. Andreev bound states versus majorana bound states in quantum dot-nanowire-superconductor hybrid structures: Trivial versus topological zero-bias conductance peaks. *Phys. Rev. B* **96**, 075161 (2017).

57. Carlsson, G. Topological methods for data modelling. *Nat. Rev. Phys.* **2**, 697–708 (2020).

58. The GUDHI Project. *GUDHI User and Reference Manual* (GUDHI Editorial Board, 2015).

59. Chen, Z. *et al.* Panoramic mapping of phonon transport from ultrafast electron diffraction and scientific machine learning. *Adv. Mater.* **35**, 2206997 (2023).

60. Samarakoon, A. M. *et al.* Machine-learning-assisted insight into spin ice dy2ti2o7. *Nat. Commun.* **11**, 892 (2020).

61. Taylor, J. R., Sau, J. D. & Sarma, S. D. Machine learning majorana nanowire disorder landscape. *arXiv preprint arXiv:2307.11068* (2023).

62. Groth, C. W., Wimmer, M., Akhmerov, A. R. & Waintal, X. Kwant: a software package for quantum transport. *New J. Phys.* **16**, 063065 (2014).

63. Prada, E., San-Jose, P. & Aguado, R. Transport spectroscopy of n s nanowire junctions with majorana fermions. *Phys. Rev. B* **86**, 180503 (2012).

64. Liu, C.-X., Sau, J. D. & Sarma, S. D. Role of dissipation in realistic majorana nanowires. *Phys. Rev. B* **95**, 054502 (2017).

65. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

66. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. neural information processing systems* **32** (2019).