

# Unreasonable effectiveness of unsupervised learning in identifying Majorana topology

Jacob Taylor,<sup>1</sup> Haining Pan,<sup>2</sup> and Sankar Das Sarma<sup>1</sup>

<sup>1</sup>*Condensed Matter Theory Center and Joint Quantum Institute,  
Department of Physics, University of Maryland, College Park, Maryland 20742, USA*  
<sup>2</sup>*Department of Physics and Astronomy, Center for Materials Theory,  
Rutgers University, Piscataway, NJ 08854, USA*

In unsupervised learning, the training data for deep learning does not come with any labels, thus forcing the algorithm to discover hidden patterns in the data for discerning useful information. This, in principle, could be a powerful tool in identifying topological order since topology does not always manifest in obvious physical ways (e.g., topological superconductivity) for its decisive confirmation. The problem, however, is that unsupervised learning is a difficult challenge, necessitating huge computing resources, which may not always work. In the current work, we combine unsupervised and supervised learning using an autoencoder to establish that unlabeled data in the Majorana splitting in realistic short disordered nanowires may enable not only a distinction between ‘topological’ and ‘trivial’, but also where their crossover happens in the relevant parameter space. This may be a useful tool in identifying topology in Majorana nanowires.

*Introduction.* Topological quantum computing (TQC) is a paradigm where appropriate manipulations of non-Abelian anyons, such as localized Majorana zero modes (MZM) in a topological superconductor (TSC), could lead to error-free fault-tolerant quantum computation [1–6]. Since the realistic theoretical predictions of the possible existence of MZMs in superconductor-semiconductor hybrid nanowire platforms in 2010 [7–10], a great deal of theoretical and experimental activities [11, 12] have focused on TQC in nanowires, with Microsoft dedicating a huge industrial effort on MZM in nanowires [13–15].

In spite of this enormous effort, a clear case for MZM is still elusive, mainly because of the inherent difficulties in identifying topological indicators in realistic systems in the presence of disorder (suppressing the TSC gap) and the relatively short wire length (thus, possibly creating MZM overlap suppressing their anyonic nature) [16–24]. A detailed topological gap protocol (TGP) has been introduced by Microsoft, but TGP only incorporates the necessary conditions for topology, and the sufficient conditions remain elusive [25–27]. Developing decisive tools for identifying topological MZM in realistic systems remains the key open problem in TQC.

In this context, modern AI-based Machine Learning (ML) techniques in the search for topological indicators in nanowires could be useful [28–30]. Although powerful from an abstract computer science viewpoint, one problem in these supervised ML techniques is that the training data sets must be labelled, identifying the individual topology for the training to lead to successful identification of MZM topology in the test data sets. Such supervised ML is relatively easy to do theoretically since every theoretical nanowire simulation used in training can be appropriately labeled as ‘topological’ or ‘trivial’ through explicit calculations of the topological visibility (TV) [31].

Obviously, the experimental training data, by definition, cannot come with any labels, and as such, doing

supervised ML in the experimental context is a challenge as it would necessitate extensive theoretical simulations of the data to determine the appropriate label, thus partially negating the usefulness of the technique. What could be extremely useful is an unsupervised or self-supervised ML algorithm that learns by itself how to label the data by discerning hidden patterns without the input training data carrying explicit topology labels. Of course, such unsupervised learning may not work, and also in the end, one must provide insight into the patterns discerned by unsupervised ML so that the patterns can have a one-to-one correspondence to MZM topological labels, i.e., ‘topological’ or ‘trivial’. Unsupervised learning is also technically much more demanding than supervised ML.

In the current work, we introduce a comprehensive method to do unsupervised ML (UML) for MZM topology identification, by first using UML to identify underlying patterns in the nanowire data, and then utilizing supervised ML (SML) as an input to label the topology. The combining of UML/SML in a seamless manner playing complementary roles may very well solve the conundrum of identifying MZM topology. We explicitly check the fidelity of the method on test data, finding excellent fidelity, establishing this as a possible breakthrough technique in MZM TQC. Furthermore, the specific supervised learning model we use here provides a unique method for predicting the topological scattering invariant that is more robust and less parameter-specific than the previous transport-based machine learning methods [29].

The specific quantity used in our UML/SML is the MZM splitting which must be exponentially small [2] in the topological regime so that the MZM are actual zero energy anyonic Majorana bound states [32, 33], and not just low-energy trivial fermionic Andreev bound states which occur generically in the nanowire platform, hugely complicating the MZM identification [34, 35]. Therefore, whether the MZM splitting is exponentially small or not

is a direct measure of the topology or not in nanowires. We use MZM splitting with no labels for our UML, finding, rather amazingly, precisely 2 or 3 clusters coming out of UML. The 2-clusters happen when the wire length is large [compared with superconducting (SC) coherence length] and the disorder is small (compared with SC gap), indicating the explicit presence only of topological or trivial phases (depending on the system parameters). By contrast, 3-clusters emerge as the UML-discerned hidden pattern only when the wire length is not necessarily large and disorder is not necessarily small, and the UML then finds three distinct patterns in the unlabeled data: topological and trivial as well as an intermediate regime in between where the topology is ill-defined and may be dominated by ABS contamination. (Much of the current experimental data seems to be in the intermediate regime.) The fidelity of the UML drops precipitously if we force the algorithm to find more than three patterns in the data, clearly establishing that UML has been effective (with no prompting/labeling) in identifying the correct topological patterns occurring in realistic samples.

*Model and data.* We start with the standard semiconductor-superconductor single-band model

$$\hat{H} = \frac{1}{2} \int dx \Psi^\dagger(x) (H_{\text{SM}} + H_Z + H_{\text{SC}} + H_{\text{dis}}) \Psi(x), \quad (1)$$

where  $\Psi(x) = (\psi_\uparrow(x), \psi_\downarrow(x), \psi_\downarrow^\dagger(x), -\psi_\uparrow^\dagger(x))^\top$  is the Nambu spinor defined such that the single-particle Hamiltonian of the SM part  $H_{\text{SM}}$ , Zeeman field part  $H_Z$ , proximity-induced superconducting part  $H_{\text{SC}}$ , and disorder part  $H_{\text{dis}}$  are given by [7–11]:

$$H_{\text{SM}} = \left( -\frac{\partial_x^2}{2m^*} - i\alpha\partial_x\sigma_y - \mu \right) \tau_z, \quad H_Z = V_z\sigma_x, \quad (2)$$

$$H_{\text{SC}} = \Delta\tau_x, \quad H_{\text{dis}} = V(x)\tau_z$$

where the definitions and the values of the parameters are shown in Table I. We are using realistic nanowire parameters, and zero temperature is justified by the low ( $\sim 20$  mK) experimental temperatures along with the fact that the TSC gap is  $\sim 400$  mK or more. [13–15]

Our dataset consists of two ingredients: the MZM energy splitting  $E_s(V_z)$  [2, 32] and the topological visibility  $\text{TV}(V_z)$  [31], both as a function of the Zeeman field  $V_z$ . The Majorana energy splitting  $E_s$  is defined as the energy of the lowest-lying state in the nanowire, which is obtained by numerically discretizing Hamiltonian (1) with a fictitious lattice constant  $a = 10$  nm and then diagonalizing the resulting tight-binding model. (The lattice constant ‘ $a$ ’ is essentially a measure of the disorder correlation length in the system.) The topological visibility (TV) is defined as the determinant of the reflection matrix at zero energy [31], computed using the KWANT package [36], which takes values between  $-1$  (topological) and  $+1$  (trivial), and zero at the topological quantum phase transition point. Examples of both  $E_s(V_z)$

TABLE I. Parameters in the Majorana nanowire simulations adapted from [37].

effective mass $m^*$	$0.01519 m_e$
chemical potential $\mu$	1 meV
spin-orbit coupling $\alpha$	0.5 eVÅ
constant SC gap $\Delta$	0.2 meV
wire length $L$	0.6–15 $\mu\text{m}$
Gaussian disorder strength $\sigma$	0.1–6 meV
temperature	0
dissipation	0
SC coherence length $\xi$	0.78 $\mu\text{m}$
localization length $\ell_{\text{loc}} = \frac{v_F^2}{\sigma^2 a}$	$2.31 \mu\text{m} \frac{1 \text{ meV}^2}{\sigma^2}$

and  $\text{TV}(V_z)$  are shown in the left column in Fig. 1. We emphasize a crucial point: TV is continuous between  $\pm 1$  and is not binary, as it would become for an infinite system with no disorder [2]—the topological quantum phase transition is still defined by the vanishing of TV [31].

*Unsupervised learning.* We first unveil hidden patterns in the unlabeled Majorana data using UML, which contains two main steps: (i) dimensionality reduction using an autoencoder neural network, and (ii) clustering in the latent space using the  $k$ -means algorithm. The autoencoder architecture is shown in Fig. 1, where we employ 1D convolutional layers to compress the input data of  $E_s(V_z)$  and  $\text{TV}(V_z)$  into a 15-dimensional latent space, and then reconstruct the original input from the latent vector. (See Sec. I in the Supplemental Material for details of the autoencoder architecture.) Here, the middle latent space is expected to capture the most important features of the input data.

At this point, we effectively compress the original high-dimensional data into a low-dimensional latent space, where we can then perform clustering using the  $k$ -means algorithm (see Sec. V in the Supplemental Material for details of the  $k$ -means algorithm). Figure 2 shows the clustering results in the latent space for (a) two-cluster and (b) three-cluster classifications, visualized using principal component analysis (PCA) or (more precisely the latent space vector) to show the first three principal components (PC1, PC2, and PC3).

To understand the physical meaning of the clusters identified in the latent space by the  $k$ -means algorithm, we color each data point according to its associated disorder strength  $\sigma$  [Fig. 2(c)] and system size  $L$  [Fig. 2(d)]. We find that ‘Cluster 0’ (blue) in both two- and three-cluster classifications corresponds to the strong disorder regime (large  $\sigma$ ), while ‘Cluster 1’ (green) corresponds to the weak disorder regime (small  $\sigma$ ).

This heuristic observation allows us to map the clustering results in the latent space back to the original physical parameter space spanned by  $\sigma$  and  $L$ , and thus obtain unsupervised phase diagrams as shown in the top row of

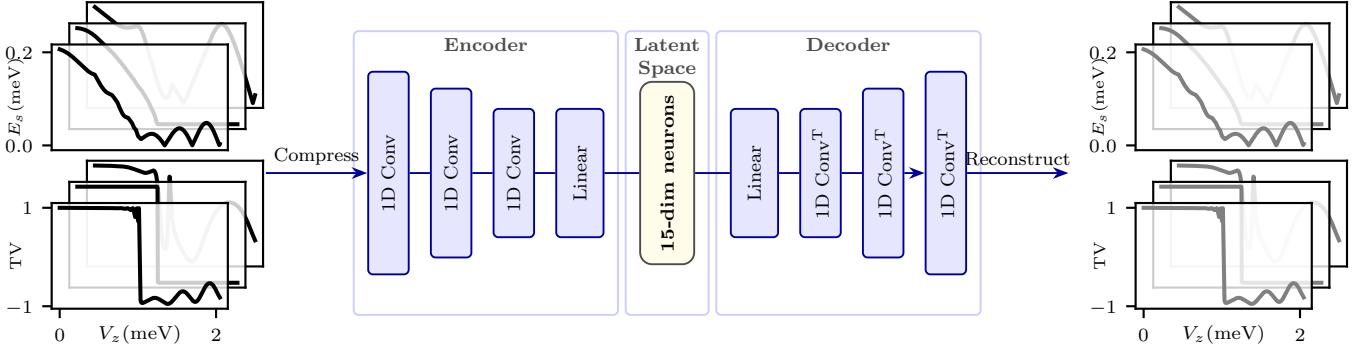


FIG. 1. Autoencoder architecture used for dimensionality reduction in the unsupervised learning framework. The network consists of a 1D convolutional encoder (with stride size of 2) and decoder with a symmetric architecture in reverse order. Left column shows examples of input data: Majorana energy splittings  $E_s(V_z)$  and topological visibility  $TV(V_z)$ . The encoder compresses these into a 15-dimensional latent vector, which the decoder reconstructs back to the original input format. Training minimizes the mean-squared error between input and reconstructed output.

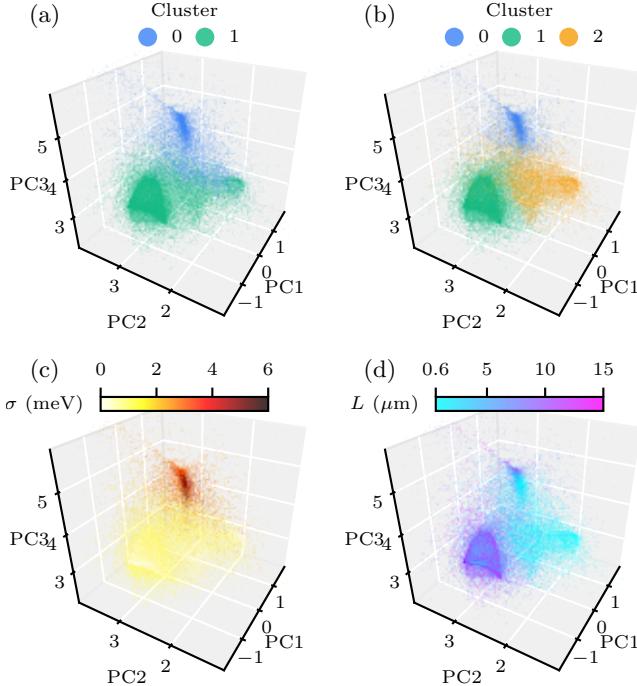


FIG. 2. Top row: Latent space visualization from the autoencoder principal component analysis (PCA), displaying the first three principal components (PC1, PC2, and PC3) for (a)  $k = 2$  and (b)  $k = 3$  cluster classifications obtained from  $k$ -means clustering. Each data point represents one disorder realization, compressed from the complete  $E_s(V_z)$  and  $TV(V_z)$  curves via the autoencoder in Fig. 1. Bottom row: Physical parameters associated with each latent space point: (c) disorder strength  $\sigma$  and (d) wire length  $L$ .

Fig. 3. Here, without any *a priori* knowledge, we find that the clusters in the latent space naturally correspond to distinct phases in the physical parameter space. For the two-cluster classification [Fig. 3(a)], we find a phase

boundary separating the weak-disorder topological phase (Cluster 1 in green) and the strong-disorder trivial phase (Cluster 0 in blue). Here, each data point in the parameter space is assigned a label according to the majority cluster among its 100 disorder realizations. The phase boundary (crossover) shifts to larger disorder strength for longer wire lengths, consistent with the expectation that longer wires are more robust against disorder. For the three-cluster classification [Fig. 3(b)], we find an additional cluster (Cluster 2 in orange) that corresponds to an intermediate crossover disorder regime.

*Stability of clustering.* We find that the clustering results are robust against different runs of the dimensionality reduction in the autoencoder, as well as different initializations of the  $k$ -means algorithm (see Sec. V in the Supplemental Material for details). We also perform a heuristic elbow-method analysis to determine the optimal number of clusters  $k$  by evaluating a standard clustering-quality metric, the Silhouette score (see Sec. IV in the Supplemental Material for details), in Fig. 3(c). Here, we find that  $k = 2$  and  $k = 3$  yield comparable separation quality, but the score drops sharply from  $k = 3$  to  $k = 4$ , indicating that having more than three clusters does not improve the clustering quality.

We further restrict the disorder range to  $\sigma \in [0, 1]$  meV (which is the estimated low-disorder range in the current MSFT experiments), and repeat the  $k$ -means clustering in the latent space, obtaining the phase diagrams shown in the bottom row of Fig. 3. Here, we find that the two-cluster classification in Fig. 3(d) is consistent with the phase diagram in Fig. 3(b) within its zoomed-in disorder range. The three-cluster classification in Fig. 3(e) reveals an additional intermediate cluster which is a clear artifact from the Silhouette score analysis in Fig. 3(f), indicating that the clustering quality is best for  $k = 2$  in the low-disorder regime.

*Supervised learning.* In the previous unsupervised

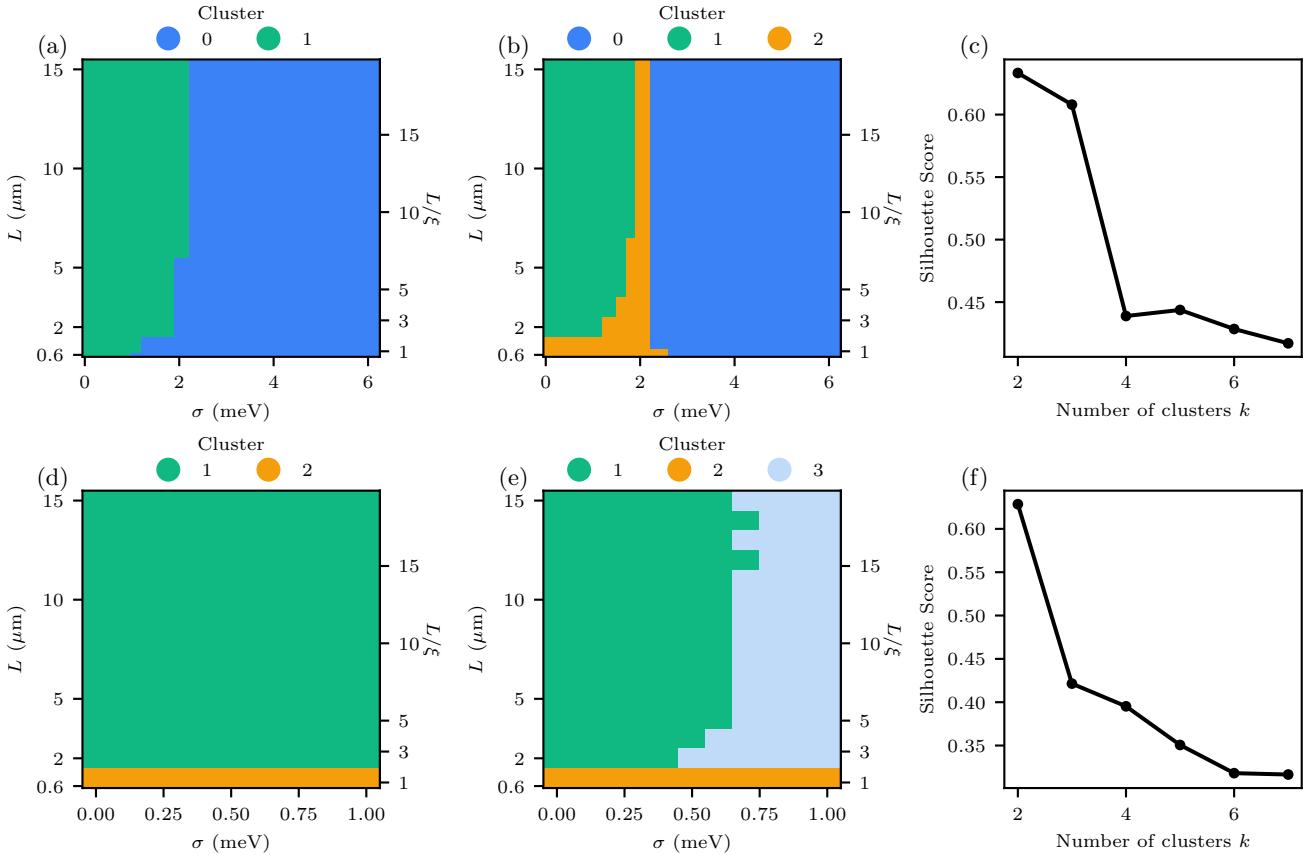


FIG. 3. Top row: Unsupervised phase diagrams in the disorder-length ( $\sigma$ - $L$ ) parameter space, obtained by mapping latent space clusters back to physical parameters. (a) Two-cluster classification reveals a phase boundary between weak-disorder (Cluster 1, green) and strong-disorder (Cluster 0, blue) regimes. (b) Three-cluster classification identifies an additional intermediate regime (Cluster 2, orange) at the crossover. Each point is labeled by the majority cluster among 100 disorder realizations. (c) Silhouette score as a function of cluster number  $k$ , showing that  $k = 2$  and  $k = 3$  have comparable quality, but the score drops sharply beyond  $k = 3$ . Bottom row (d-f): Same clustering analysis for restricted disorder range  $\sigma \in [0, 1]$  meV, where (f) indicates optimal clustering at  $k = 2$ .

learning analysis, we have already identified the internal patterns in the unlabeled Majorana data, which naturally correspond to distinct phases induced by disorder, but we do not know the topology associated with the distinct phases since UML only discerns patterns, and cannot provide labels for the patterns.

However, we find that the success of the unsupervised learning crucially depends on including the topological visibility  $TV(V_z)$  in the input data; otherwise, the clustering is trivially based on the system size  $L$  alone (see Fig.S2 in the Supplemental Material for details). This necessitates a supervised learning framework to predict the topological visibility  $TV(V_z)$ , which is accessible in theory, from the experimentally measurable Majorana energy splitting  $E_s(V_z)$  alone. Therefore, in actual experiments, the supervised learning model can be the upstream component before the unsupervised learning.

To this end, we design a supervised learning neural network (see Fig. 4(a)) based on a 1D convolutional encoder-decoder architecture, where the inputs are two exper-

imentally measurable quantities: the Majorana energy splitting  $E_s(V_z)$  and the wire length  $L$ , while the outputs are the topological visibility  $TV(V_z)$  and the disorder strength  $\sigma$ , which are not directly measurable in experiments. We split the dataset into a training 95% and a test on 5% withheld data. We find generally good prediction accuracy with  $R^2$  of 0.861 and 0.922 errors of  $\pm 0.654$  and  $\pm 0.242$  for  $\sigma(V_z)$  and  $TV(V_z)$  respectively. In Fig. 4(b), we compare the predicted topological visibility (vertical axis) with the ground-truth values (horizontal axis) for the full disorder range  $\sigma \in [0, 6]$  meV, finding a good correlation along the diagonal line (red dashed line). In Fig. 4(c), the predicted disorder strength (vertical axis) is compared with the ground-truth values (horizontal axis) for the full disorder range  $\sigma \in [0, 6]$  meV, where we find that the model performs well at weak disorder but its accuracy decreases at strong disorder. (This is not a problem for experiments, which must avoid the very strong disorder regime any way.)

The prediction accuracy improves significantly when

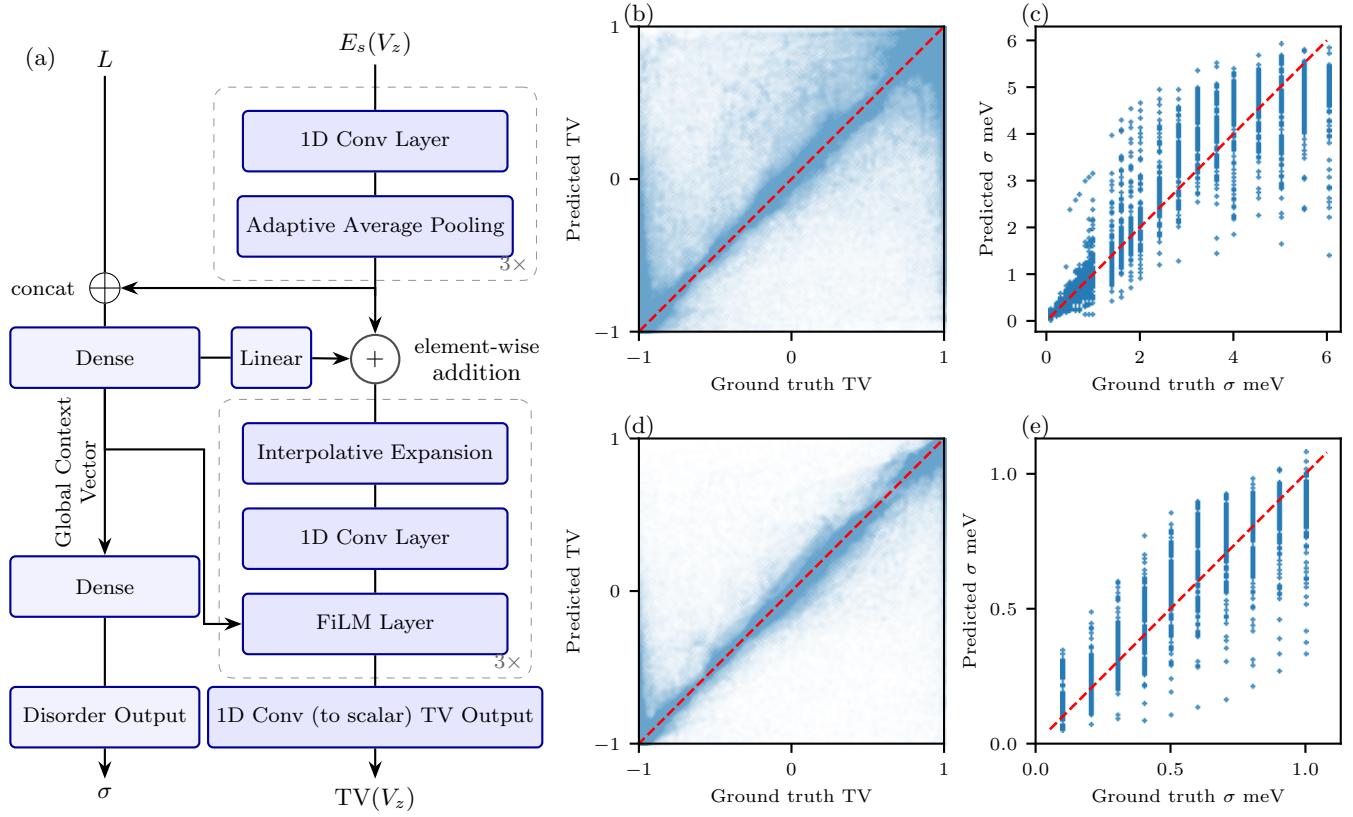


FIG. 4. (a) Architecture of the supervised learning neural network based on 1D convolutional encoder-decoder. Inputs are experimentally measurable quantities ( $E_s(V_z)$  and wire length  $L$ ); outputs are theoretically accessible quantities (topological visibility  $TV(V_z)$  and disorder strength  $\sigma$ ) that are not directly measurable in experiments. (b) Predicted versus ground-truth topological visibility for the full disorder range  $\sigma \in [0, 6]$  meV, showing good correlation along the diagonal (red dashed line indicates perfect prediction). (c) Predicted versus ground-truth disorder strength for  $\sigma \in [0, 6]$  meV; accuracy is higher at weak disorder and decreases at strong disorder. (d, e) Same predictions for the restricted weak-disorder range  $\sigma \in [0, 1]$  meV, demonstrating significantly improved accuracy for both  $TV$  and  $\sigma$ .

we restrict the disorder range to  $\sigma \in [0, 1]$  meV, as shown in Figs. 4(d) and (e) for the topological visibility and disorder strength, respectively, where we find a much better correlation along the diagonal line and report higher accuracy with  $R^2$  of 0.722 and 0.971 errors of  $\pm 0.152$  and  $\pm 0.157$  for  $\sigma(V_z)$  and  $TV(V_z)$  respectively. For the detailed accuracy profile as a function of disorder strength and wire length, see Fig. S4 and Fig. S3, respectively, for weak and strong disorder in the Supplemental Material.

*Conclusion.* We have carried out an unsupervised machine learning of unlabeled Majorana mode splitting in the current experimentally active TQC nanowire platform, finding to our pleasant surprise that the splitting naturally falls into machine-identified 2- and 3-clusters respectively for weak and strong disorder, where we can identify the clusters through supervised learning analysis to be topological or trivial in the 2-clusters, with the 3-clusters for higher disorder also containing an intermediate phase of ill-defined topology. Our work should be useful in the classification of Majorana experimental data, particularly in the context of ruling out the param-

eter regimes where topology may not manifest in realistic nanowires.

*Acknowledgements.* This work is supported by the Laboratory for Physical Sciences through the Condensed Matter Theory Center at Maryland. JRT thanks the Joint Quantum Institute for additional support. HP is supported by US-ONR grant No. N00014-23-1-2357.

- 
- [1] A. Yu. Kitaev, Fault-tolerant quantum computation by anyons, *Annals of Physics* **303**, 2 (2003).
  - [2] A. Y. Kitaev, Unpaired Majorana fermions in quantum wires, *Phys.-Usp.* **44**, 131 (2001).
  - [3] M. H. Freedman, A. Kitaev, M. J. Larsen, and Z. Wang, *Topological Quantum Computation* (2002).
  - [4] S. Das Sarma, M. Freedman, and C. Nayak, Topologically Protected Qubits from a Possible Non-Abelian Fractional Quantum Hall State, *Phys. Rev. Lett.* **94**, 166802 (2005).
  - [5] C. Nayak, S. H. Simon, A. Stern, M. Freedman, and S. Das Sarma, Non-Abelian anyons and topological quantum computation, *Reviews of Modern Physics* **80**, 1083

- (2008).
- [6] S. D. Sarma, M. Freedman, and C. Nayak, Majorana zero modes and topological quantum computation, *npj Quantum Information* **1**, 15001 (2015).
  - [7] R. M. Lutchyn, J. D. Sau, and S. Das Sarma, Majorana Fermions and a Topological Phase Transition in Semiconductor-Superconductor Heterostructures, *Phys. Rev. Lett.* **105**, 077001 (2010).
  - [8] J. D. Sau, R. M. Lutchyn, S. Tewari, and S. Das Sarma, Generic New Platform for Topological Quantum Computation Using Semiconductor Heterostructures, *Phys. Rev. Lett.* **104**, 040502 (2010).
  - [9] J. D. Sau, R. M. Lutchyn, S. Tewari, and S. Das Sarma, Robustness of Majorana fermions in proximity-induced superconductors, *Phys. Rev. B* **82**, 094522 (2010).
  - [10] Y. Oreg, G. Refael, and F. von Oppen, Helical Liquids and Majorana Bound States in Quantum Wires, *Phys. Rev. Lett.* **105**, 177002 (2010).
  - [11] S. Das Sarma, In search of Majorana, *Nat. Phys.* **19**, 165 (2023).
  - [12] L. Kouwenhoven, Perspective on Majorana bound-states in hybrid superconductor-semiconductor nanowires (2024).
  - [13] Microsoft Quantum, M. Aghaei, A. Akkala, Z. Alam, R. Ali, A. Alcaraz Ramirez, M. Andrzejczuk, A. E. Antipov, P. Aseev, M. Astafev, B. Bauer, J. Becker, S. Boddapati, F. Boekhout, J. Bommer, T. Bosma, L. Bourdet, S. Boutin, P. Caroff, L. Casparis, M. Cassidy, S. Chatoor, A. W. Christensen, N. Clay, W. S. Cole, F. Corsetti, A. Cui, P. Dalampiras, A. Dokania, G. de Lange, M. de Moor, J. C. Estrada Saldaña, S. Fallahi, Z. H. Fathabad, J. Gamble, G. Gardner, D. Govender, F. Griggio, R. Grigoryan, S. Gronin, J. Gukelberger, E. B. Hansen, S. Heedt, J. Herranz Zamorano, S. Ho, U. L. Holgaard, H. Ingerslev, L. Johansson, J. Jones, R. Kallaher, F. Karimi, T. Karzig, C. King, M. E. Kloster, C. Knapp, D. Kocon, J. Koski, P. Kostamo, P. Krogstrup, M. Kumar, T. Laeven, T. Larsen, K. Li, T. Lindemann, J. Love, R. Lutchyn, M. H. Madsen, M. Manfra, S. Markussen, E. Martinez, R. McNeil, E. Memisevic, T. Morgan, A. Mullally, C. Nayak, J. Nielsen, W. H. P. Nielsen, B. Nijholt, A. Nurmohamed, E. O'Farrell, K. Otani, S. Pauka, K. Petersson, L. Petit, D. I. Pikulin, F. Preiss, M. Quintero-Perez, M. Rajpalke, K. Rasmussen, D. Razmadze, O. Reentila, D. Reilly, R. Rouse, I. Sadovskyy, L. Sainiemi, S. Schreppler, V. Sidorkin, A. Singh, S. Singh, S. Sinha, P. Sohr, T. Stankevič, L. Stek, H. Suominen, J. Suter, V. Svidenko, S. Teicher, M. Temuerhan, N. Thiyagarajah, R. Tholapi, M. Thomas, E. Toomey, S. Upadhyay, I. Urban, S. Vaitiekėnas, K. Van Hoogdalem, D. Van Woerkom, D. V. Viazmitinov, D. Vogel, S. Waddy, J. Watson, J. Weston, G. W. Winkler, C. K. Yang, S. Yau, D. Yi, E. Yucelen, A. Webster, R. Zeisel, and R. Zhao, InAs-Al hybrid devices passing the topological gap protocol, *Phys. Rev. B* **107**, 245423 (2023).
  - [14] M. Aghaei, A. Alcaraz Ramirez, Z. Alam, R. Ali, M. Andrzejczuk, A. Antipov, M. Astafev, A. Barzegar, B. Bauer, J. Becker, U. K. Bhaskar, A. Bocharov, S. Boddapati, D. Bohn, J. Bommer, L. Bourdet, A. Bousquet, S. Boutin, L. Casparis, B. J. Chapman, S. Chatoor, A. W. Christensen, C. Chua, P. Codd, W. Cole, P. Cooper, F. Corsetti, A. Cui, P. Dalpasso, J. P. Dehollain, G. de Lange, M. de Moor, A. Ekefjärd, T. El Dandachi, J. C. Estrada Saldaña, S. Fallahi, L. Galletti, G. Gardner, D. Govender, F. Griggio, R. Grigoryan, S. Grijalva, S. Gronin, J. Gukelberger, M. Hamdast, F. Hamze, E. B. Hansen, S. Heedt, Z. Heidarnia, J. Herranz Zamorano, S. Ho, L. Holgaard, J. Hornibrook, J. Indrapiromkul, H. Ingerslev, L. Ivancevic, T. Jensen, J. Jhoja, J. Jones, K. V. Kalashnikov, R. Kallaher, R. Kalra, F. Karimi, T. Karzig, E. King, M. E. Kloster, C. Knapp, D. Kocon, J. V. Koski, P. Kostamo, M. Kumar, T. Laeven, T. Larsen, J. Lee, K. Lee, G. Leum, K. Li, T. Lindemann, M. Looij, J. Love, M. Lucas, R. Lutchyn, M. H. Madsen, N. Madulid, A. Malmros, M. Manfra, D. Mantri, S. B. Markussen, E. Martinez, M. Mattila, R. McNeil, A. B. Mei, R. V. Mishmash, G. Mohandas, C. Mølgaard, T. Morgan, G. Moussa, C. Nayak, J. H. Nielsen, J. M. Nielsen, W. H. P. Nielsen, B. Nijholt, M. Nystrom, E. O'Farrell, T. Ohki, K. Otani, B. Paquelet Wütz, S. Pauka, K. Petersson, L. Petit, D. Pikulin, G. Prawiroatmodjo, F. Preiss, E. Puchol Morejon, M. Rajpalke, C. Ranta, K. Rasmussen, D. Razmadze, O. Reentila, D. J. Reilly, Y. Ren, K. Reneris, R. Rouse, I. Sadovskyy, L. Sainiemi, I. Sanlorenzo, E. Schmidgall, C. Sfiligoj, M. B. Shah, K. Simoes, S. Singh, S. Sinha, T. Soerensen, P. Sohr, T. Stankevič, L. Stek, E. Stuppard, H. Suominen, J. Suter, S. Teicher, N. Thiyagarajah, R. Tholapi, M. Thomas, E. Toomey, J. Tracy, M. Turley, S. Upadhyay, I. Urban, K. Van Hoogdalem, D. J. Van Woerkom, D. V. Viazmitinov, D. Vogel, J. Watson, A. Webster, J. Weston, G. W. Winkler, D. Xu, C. K. Yang, E. Yucelen, R. Zeisel, G. Zheng, and J. Zilke, Interferometric single-shot parity measurement in InAs-Al hybrid devices, *Nature* **638**, 651 (2025).
  - [15] M. Aghaei, Z. Alam, R. Andersen, M. Andrzejczuk, A. Antipov, M. Astafev, L. Avilovas, A. Azizimanesh, E. Banek, B. Bauer, J. Becker, U. K. Bhaskar, A. G. Boa, S. Boddapati, N. Bohac, J. D. S. Bommer, J. Borovsky, L. Bourdet, S. Boutin, L. Casparis, S. Chakravarthi, H. Chalabi, B. J. Chapman, N. Chatzaras, T.-C. Chien, J. Cho, P. Codd, W. Cole, P. W. Cooper, F. Corsetti, A. Cui, T. E. Dandachi, C. Dinesen, A. Ekefjärd, S. Fallahi, L. Galletti, G. C. Gardner, G. L. Gonzalez, D. Govender, F. Griggio, R. Grigoryan, S. Grijalva, S. Gronin, J. Gukelberger, M. Hamdast, A. B. Hamida, E. B. Hansen, C. T. Hansen, S. Heedt, S. Ho, L. Holgaard, K. van Hoogdalem, J. Hornibrook, H. Ingerslev, L. Ivancevic, S. Jamo, M. Jantos, T. Jensen, J. S. Jhoja, J. C. Jones, V. Joshi, K. V. Kalashnikov, R. Kallaher, R. Kalra, F. Karimi, T. Karzig, S. Kimes, E. King, M. E. Kloster, C. Knapp, J. V. Koski, P. Kostamo, T. Laeven, J. Lai, G. de Lange, T. W. Larsen, K. Lee, K. Li, G. Li, S. Liang, T. Lindemann, M. Looij, M. Lucas, R. Lutchyn, M. H. Madsen, N. Madulid, M. J. Manfra, L. Manjunath, S. Markussen, E. Martinez, M. Mattila, J. R. Mattinson, R. P. G. McNeil, A. P. Millan, R. V. Mishmash, S. Mittal, C. Mølgaard, M. W. A. de Moor, E. P. Morejon, T. Morgan, G. Moussa, B. P. Nabar, A. Narla, C. Nayak, J. H. Nielsen, W. H. P. Nielsen, F. Nolet, M. J. Nystrom, E. O'Farrell, T. A. Ohki, K. Otani, C. Papon, K. D. Petersson, L. Petit, D. Pikulin, M. Rajpalke, A. A. Ramirez, D. Razmadze, Y. Ren, I. Sadovskyy, L. Sainiemi, J. C. E. Saldaña, I. Sanlorenzo, T. P. dos Santos, S. Schaaf, J. Schack, E. R. Schmidgall, C. Sfiligoj, S. Sinha, P. Sohr,

- T. L. Sørensen, K. Spiegelhauer, T. Stankević, L. J. Stek, P. Strøm-Hansen, H. J. Suominen, J. Suter, S. M. L. Teicher, R. Tholapi, M. Thomas, D. W. Tom, E. Toomey, J. Tracy, M. Turley, M. D. Turner, S. Upadhyay, I. Urban, D. V. Viazmitinov, A. W. Viazmitinova, B. Viegas, D. J. Vogel, J. Watson, A. Webster, J. Weston, T. Williamson, G. W. Winkler, D. J. van Woerkom, B. P. Wuett, C.-K. Yang, Shang-Jyun, Yu, E. Yucelen, J. H. Zamorano, R. Zeisel, G. Zheng, and A. M. Zimmerman, *Distinct Lifetimes for X and Z Loop Measurements in a Majorana Tetron Device* (2025).
- [16] J. Liu, A. C. Potter, K. T. Law, and P. A. Lee, Zero-Bias Peaks in the Tunneling Conductance of Spin-Orbit-Coupled Superconducting Wires with and without Majorana End-States, *Phys. Rev. Lett.* **109**, 267002 (2012).
- [17] P. W. Brouwer, M. Duckheim, A. Romito, and F. von Oppen, Topological superconducting phases in disordered quantum wires with strong spin-orbit coupling, *Phys. Rev. B* **84**, 144526 (2011).
- [18] D. Bagrets and A. Altland, Class D Spectral Peak in Majorana Quantum Wires, *Phys. Rev. Lett.* **109**, 227005 (2012).
- [19] A. R. Akhmerov, J. P. Dahlhaus, F. Hassler, M. Wimmer, and C. W. J. Beenakker, Quantized Conductance at the Majorana Phase Transition in a Disordered Superconducting Wire, *Phys. Rev. Lett.* **106**, 057001 (2011).
- [20] J. D. Sau and S. Das Sarma, Density of states of disordered topological superconductor-semiconductor hybrid nanowires, *Phys. Rev. B* **88**, 064506 (2013).
- [21] S. D. Sarma and H. Pan, Density of states, transport, and topology in disordered Majorana nanowires, *arXiv:2305.06837* (2023).
- [22] H. Pan and S. D. Sarma, Majorana zero modes in semiconductor-superconductor hybrid structures: Defining topology in short and disordered nanowires through Majorana splitting, *arxiv:2507.00128* (2025).
- [23] S. D. Sarma, J. D. Sau, and T. D. Stanescu, Spectral properties, topological patches, and effective phase diagrams of finite disordered Majorana nanowires, *arXiv:2305.07007* (2023).
- [24] J. D. Sau and S. D. Sarma, Capacitance-based Fermion parity read-out and predicted Rabi oscillations in a Majorana nanowire, *Phys. Rev. B* **111**, 224509 (2025).
- [25] D. I. Pikulin, B. van Heck, T. Karzig, E. A. Martinez, B. Nijholt, T. Laeven, G. W. Winkler, J. D. Watson, S. Heedt, M. Temurhan, V. Svidenko, R. M. Lutchyn, M. Thomas, G. de Lange, L. Casparis, and C. Nayak, *Protocol to identify a topological superconducting phase in a three-terminal device* (2021).
- [26] H. Pan, J. D. Sau, and S. Das Sarma, Three-terminal nonlocal conductance in Majorana nanowires: Distinguishing topological and trivial in realistic systems with disorder and inhomogeneous potential, *Phys. Rev. B* **103**, 014513 (2021).
- [27] T. Ö. Rosdahl, A. Vuik, M. Kjaergaard, and A. R. Akhmerov, Andreev rectifier: A nonlocal conductance signature of topological phase transitions, *Phys. Rev. B* **97**, 045421 (2018).
- [28] J. R. Taylor, J. D. Sau, and S. D. Sarma, Machine learning Majorana nanowire disorder landscape, *arXiv:2307.11068* (2023).
- [29] J. R. Taylor and S. Das Sarma, Vision transformer based deep learning of topological indicators in Majorana nanowires, *Phys. Rev. B* **111**, 104208 (2025).
- [30] J. R. Taylor and S. Das Sarma, Mitigating disorder and optimizing topological indicators with vision-transformer-based neural networks in Majorana nanowires, *Phys. Rev. B* **112**, L041110 (2025).
- [31] S. Das Sarma, A. Nag, and J. D. Sau, How to infer non-Abelian statistics and topological visibility from tunneling conductance properties of realistic Majorana nanowires, *Phys. Rev. B* **94**, 035143 (2016).
- [32] M. Cheng, R. M. Lutchyn, V. Galitski, and S. Das Sarma, Splitting of Majorana-Fermion Modes due to Intervortex Tunneling in a  $p_x+ip_y$  Superconductor, *Phys. Rev. Lett.* **103**, 107001 (2009).
- [33] P. W. Brouwer, M. Duckheim, A. Romito, and F. von Oppen, Probability Distribution of Majorana End-State Energies in Disordered Wires, *Phys. Rev. Lett.* **107**, 196804 (2011).
- [34] H. Pan and S. Das Sarma, Physical mechanisms for zero-bias conductance peaks in Majorana nanowires, *Phys. Rev. Research* **2**, 013377 (2020).
- [35] H. Pan, W. S. Cole, J. D. Sau, and S. Das Sarma, Generic quantized zero-bias conductance peaks in superconductor-semiconductor hybrid structures, *Phys. Rev. B* **101**, 024506 (2020).
- [36] C. W. Groth, M. Wimmer, A. R. Akhmerov, and X. Waintal, Kwant: A software package for quantum transport, *New Journal of Physics* **16**, 063065 (2014).
- [37] R. M. Lutchyn, E. P. A. M. Bakkers, L. P. Kouwenhoven, P. Krogstrup, C. M. Marcus, and Y. Oreg, Majorana zero modes in superconductor–semiconductor heterostructures, *Nature Reviews Materials* **3**, 52 (2018).
- [38] H. Hotelling, Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417 (1933).
- [39] T. Hofmann, B. Schölkopf, and A. J. Smola, Kernel methods in machine learning, *Ann. Statist.* **36** (2008).
- [40] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, FiLM: Visual Reasoning with a General Conditioning Layer, *Proceedings of the AAAI Conference on Artificial Intelligence* **32** (2018).
- [41] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, *Information Sciences* **622**, 178 (2023).

# Supplemental Materials for “The unreasonable effectiveness of unsupervised learning in identifying Majorana topology”

## Appendix I: Neural Network Architectures

There are two separate neural networks used in this paper. The first is employed for the unsupervised learning task and consists of a standard autoencoder, shown in Fig. 1. The second is used for the supervised learning task and is a more novel architecture that incorporates global context vectors into an encoder-decoder domain conversion setup, as shown in Fig. 4a.

### 1. Unsupervised autoencoder network

The auto-encoder neural network (shown in Fig. 1) works by taking in input in the form of two-row array

$$X_{in} = \begin{bmatrix} E_s(V_z^{(1)}) \dots E_s(V_z^{(n)}) \\ TV(V_z^{(1)}) \dots TV(V_z^{(n)}) \end{bmatrix} \quad (\text{S-I.1})$$

The encoder consists of three 1D convolutional layers with [32, 64, 128] filters, respectively. These layers convolve over the  $V_z$  axis, where the rows  $E_s(V_z)$  and  $TV(V_z)$  are treated as separate input channels. Each convolutional layer uses a stride of 2, halving the size of the  $V_z$  axis at each step. The encoder concludes by applying a linear layer that maps to a bottleneck latent space of size 15. This dimensionality was selected because 15 components capture approximately 95% of the explained variance in a linear PCA analysis [38]. We find that our results are qualitatively robust for latent vector sizes in the range [5, 20], although smaller latent dimensions lead to less well-defined phase boundaries. This compressed latent space vector is used by the clustering algorithm to perform unsupervised classification. The decoder is nearly identical to the encoder, but operates in reverse, using transposed convolutional layers to progressively upscale (with stride 2) the representation until the output vector  $X_{out}$  matches the original input size  $X_{in}$ . The training process consists of optimizing the network to achieve  $X_{out} = X_{in}$ , that is, to learn a compression into a latent space vector  $\tilde{L}$  from which the original input can be reconstructed through the decoder. The autoencoder network is intentionally simple in order to allow the compression to be natural for the data, and thus to prevent overfitting. The autoencoder is also more stable and less parameter-dependent than alternative kernel methods [39], as it tends to work by naturally finding an encoding of the data in an optimal manner inherent to the training process.

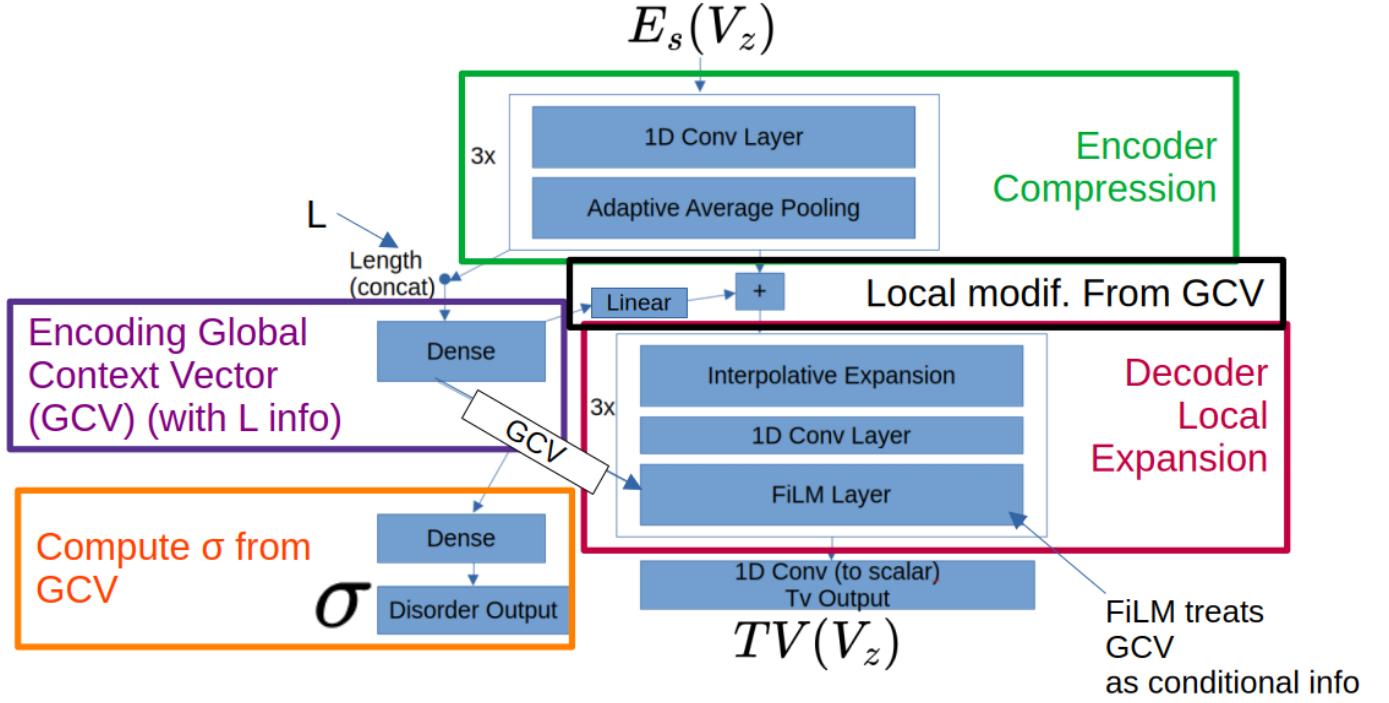
## 2. Supervised network

The supervised learning neural network, shown in Fig. S1, is more novel than the unsupervised architecture. The network input consists of a vector  $[E_s(V_z^{(1)}) \dots E_s(V_z^{(n)})]$  of length 512, with  $V_z \in [0, 2]$  meV, while the output consists of an identically sized vector  $[TV(V_z^{(1)}) \dots TV(V_z^{(n)})]$ , along with a single scalar component representing the disorder magnitude  $\sigma$ . The network is composed of five distinct parts, as outlined in the figure. The encoder consists of three blocks of 1D convolutional layers, each followed by adaptive average pooling that halves the size of the  $V_z$  axis.

The encoder convolutional layers have 128 filters. The encoder compresses the input data and, when combined with the pooling operations, allows different data  $V_z$  scales to be assessed. The network’s global context vector encoding, shown in purple, consists of taking the full output of the encoder, concatenating it with the wire length, and feeding this combined representation into a dense layer with 64 neurons to form a small global context vector (GCV) that is used as a conditional input to the decoder. This GCV is also independently fed into another dense layer with 32 neurons, from which the scalar disorder magnitude is computed and output. The GCV is additionally passed through a linear layer and then additively combined with the encoder output to form the input, shown in black, to the decoder. This design allows locally relevant information from the GCV to be incorporated into the input to the decoder, and performance is worse without this inclusion. The decoder, shown in red, consists of three blocks, with each block comprising an interpolative expansion that doubles the size of the  $V_z$  axis, followed by a 1D convolutional layer and a FiLM layer [40] that incorporates the GCV as conditional information into the convolutional process. The FiLM layer operates by applying channel-wise scaling and shifting to each convolutional output based on the GCV, thereby allowing the convolutional network to effectively utilize the global context information. These convolutional layers have 128 filters. The network concludes with a single convolutional layer that combines the multiple decoder channels into a single output vector  $TV(V_z)$ .

## Appendix II: Clustering results using Majorana splitting energy only

The cluster result in the main text uses both the topological visibility  $TV(V_z)$  and the Majorana splitting en-



Supplementary Figure S1. Explanatory diagram of Supervised learning Neural Network. The neural network consists of 5 parts: the  $E_s(V_z)$  encoder (in green) that converts the splitting into a smaller embedding space, the global context vector (GCV) (in purple) that converts the embedding along with the length into a small context vector containing global information able to be used as a conditional, the local modification from GCV (in black) that takes the GCV and modifies the embedding space in the locally relevant manner, the decoder (in red) which expands and maps the embedding along with the GCV conditional into the  $TV(V_z)$  output, and finally the small dense neural network in (orange) that converts the GCV to a global disorder magnitude.

ergy  $E_s(V_z)$  as input to the autoencoder. We find that the inclusion of  $TV(V_z)$  is crucial for a meaningful clustering result. If we only use  $E_s(V_z)$  as input to the autoencoder, the clustering result becomes trivial, as shown in Fig. S2.

### Appendix III: Accuracy of the supervised learning

The prediction accuracy as a function of the disorder magnitude  $\sigma$  and wire length  $L$  of the supervised, corresponding to Fig. 4 for the strong disorder regime  $\sigma \in [0, 6]$  meV and weak disorder regime  $\sigma \in [0, 1]$  meV are shown in Fig. S3 and Fig. S4 respectively.

### Appendix IV: Silhouette Score

Let  $\{x_i\}_{i=1}^N$  be points (e.g., latent vectors) partitioned by  $k$ -means into  $k \geq 2$  clusters  $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$  with index map  $k(i)$  for point  $x_i$ . Let  $d(\cdot, \cdot)$  denote the distance in latent space (Euclidean unless stated otherwise). For a

point  $x_i$  define the mean intra-cluster distance

$$a(i) = \frac{1}{|\mathcal{C}_{k(i)}| - 1} \sum_{\substack{x_j \in \mathcal{C}_{k(i)} \\ j \neq i}} d(x_i, x_j),$$

(use  $a(i) = 0$  if  $|\mathcal{C}_{k(i)}| = 1$ ). (S-IV.1)

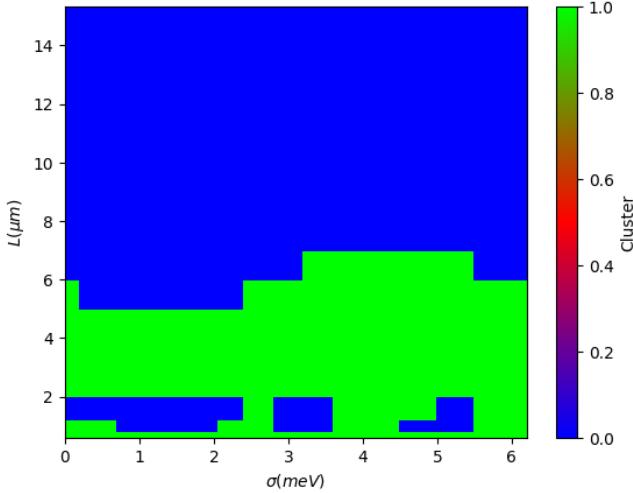
and the smallest mean distance to another cluster

$$b(i) = \min_{\ell \neq k(i)} \frac{1}{|\mathcal{C}_\ell|} \sum_{x_j \in \mathcal{C}_\ell} d(x_i, x_j). \quad (\text{S-IV.2})$$

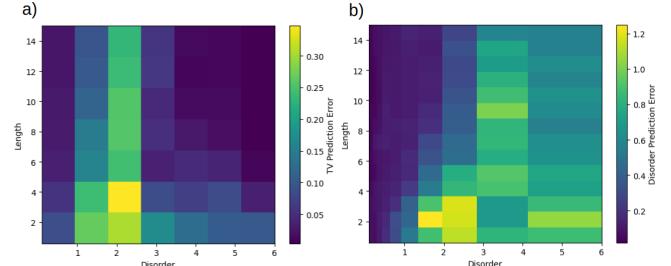
The silhouette of  $x_i$  and the overall silhouette score are

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1] \quad \text{and} \quad S = \frac{1}{N} \sum_{i=1}^N s(i). \quad (\text{S-IV.3})$$

Interpretation:  $s(i) \approx 1$  means  $x_i$  is well matched to its cluster;  $s(i) \approx 0$  indicates boundary/overlap;  $s(i) < 0$  suggests misassignment. Larger  $S$  indicates better-defined clusters (often  $S \gtrsim 0.5$  is considered good, but this is data dependent).



Supplementary Figure S2. Unsupervised phase diagram for input data lacking TV in the disorder-length ( $\sigma$ - $L$ ) parameter space, obtained by mapping latent space clusters back to physical parameters. Two-cluster classification reveals trivial clustering only between short and long systems, with some noise yielding to unclean boundaries. Forcing more than 2 clusters yields unpredictable results.

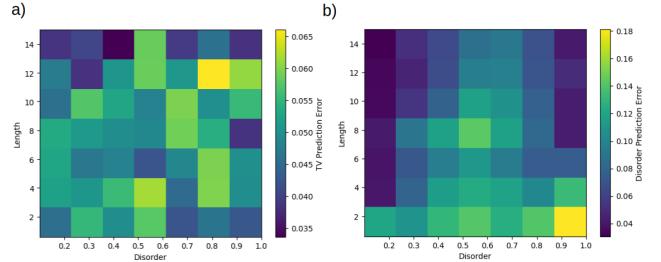


Supplementary Figure S3. Prediction error heatmaps in the  $(\sigma, L)$  parameter space for the full disorder regime  $\sigma \in [0, 6]$  meV: (a) TV prediction error, (b) disorder magnitude prediction error. Color indicates the magnitude of prediction error (lighter/yellow = higher error). Interestingly, errors appear larger near the phase transition resulting from the unsupervised learning.

#### Appendix V: $k$ -means

The  $k$ -means algorithm is employed to cluster the latent space representations obtained from the autoencoder. The goal is to group similar disorder realizations together based on their latent space features.

The  $k$ -means algorithm [41] works by initializing  $k$  centroids in the latent space, and iteratively refining their positions based on the data points assigned to each cluster. The assignment step involves computing the distance between each data point and the centroids, while the update step recalculates the centroids as the mean of the assigned points.



Supplementary Figure S4. Prediction error heatmaps in the  $(\sigma, L)$  parameter space for weak disorder regime  $\sigma \in [0, 1]$  meV: (a) TV prediction error, (b) disorder magnitude prediction error. Color indicates the magnitude of prediction error (lighter/yellow = higher error). Errors are generally larger in the high disorder, short wire length regime.

To determine the optimal number of clusters  $k$ , we utilize the silhouette score discussed above, which measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters. This process should ideally yield well-separated clusters of points within the latent space.