# Precise Identification of Topological Phase Transitions with Eigensystem-Based Clustering

**Xianquan Yan**[b,a], **Jian-Song Pan**[b]

[a] *Department of {Physics, Computer Science}, National University of Singapore, Singapore 117551* yanx@u.nus.edu

[b] *Department of Physics and Key Laboratory of High Energy Density Physics and Technology of Ministry of Education, Sichuan University, Chengdu, 610065, China* panjsong@scu.edu.cn

* Presenting author

Recent advances in machine learning have spurred new ways to explore and classify quantum phases of matter. We propose an *Eigensystem-based* representation, combined with a Gaussian Mixture Model (GMM), to unsupervisedly cluster Hamiltonians into distinct topological phases with minimal feature engineering. The method identifies different topological phases without any prior knowledge, pinpoints phase boundaries with remarkable precision $\sim \mathcal{O}(10^{-5})$, remains robust under moderate noise, and scales efficiently via a simple dimensionality-reduction step. The success of GMM offers a novel physical insight — each phase forms a well-separated multivariate Gaussian in a high-dimensional "Eigensystem space." We illustrate the approach on several 1D lattice models, all achieving near 100% accuracies.

## 1. Introduction and Motivation

Topological phases of matter—recognized by the 2016 Nobel Prize in Physics—are characterized by non-local invariants rather than conventional symmetry breaking. They enable fault-tolerant quantum computing proposals and guide novel material discoveries.

Automated discovery of such phases is a compelling goal. Early works employed supervised learning that requires labeled data [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. More recent approaches increasingly favor unsupervised methods [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38], paving the way for automatic discovery of new or unexpected phases.

However, existing unsupervised approaches often need hyperparameter tuning or sophisticated physics-biased features (e.g., correlation functions, entanglement spectra, partially known symmetries, pseudo-spin configurations).

Here, we introduce a direct **Eigensystem vector** representation of Hamiltonians, involving only eigen-decomposition. Combined with a Gaussian Mixture Model (GMM) for clustering, the approach:

1. Identifies distinct topological phases with near 100% accuracy without any prior knowledge.
2. Determines topological phase boundaries with high precision (on the order of $10^{-5}$).
3. Eliminates specialized feature engineering and tuning of free hyperparameters.
4. Scales efficiently by applying a simple linear projection to the full Eigensystem vectors.
5. Demonstrates success across several benchmark models, suggesting its broad applicability.

## 2. Method

*Eigensystem vector —* Consider an $N \times N$ Hamiltonian $\boldsymbol{H}$ with eigenvalues $\{E_i\}$ and normalized eigenvectors $\{|\psi_i\rangle := (\psi_{i1}, \psi_{i2}, \ldots, \psi_{iN})^T\}$[1]. We concatenate all eigenvector components and the eigenvalues into a single, long *Eigensystem vector*:

$$|\mathrm{Eig}\rangle := \big(\psi_{11}, \psi_{12}, \ldots, \psi_{1N}, \psi_{21}, \ldots, \psi_{NN}, E_1, E_2, \ldots, E_N\big)^T \quad (1)$$

When generated for a series of Hamiltonians $\boldsymbol{H}(\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, we obtain a collection of Hamiltonians distributed over all phases, which then are transformed into $|\mathrm{Eig}\rangle$, i.e., points in "**Eigensystem space**."

*Gaussian Mixture Model (GMM) —* Given a dataset of Eigensystem vectors $\{|\mathrm{Eig}\rangle_1, |\mathrm{Eig}\rangle_2, \ldots\}$, we fit a GMM with $k$ components:

$$p(\mathrm{data}) = \sum_{i=1}^{k} \pi_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (2)$$

where each component is a Gaussian specified by mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$, and $\pi_i$ are mixture weights.

The number of components $k$ is chosen via *silhouette analysis*, which equates to the number of distinct phases in the dataset.

After fitting, each sample is assigned to the cluster (phase) with the highest responsibility, i.e., *the dataset is automatically partitioned into different topological phases*.

## 3. Example: 1D Spin-Orbit Coupled Fermi Chain

*Model and Dataset —* We illustrate our method on a 1D lattice model with spin-orbit coupling [39] (details in appendix B). In momentum space, its Hamiltonian reads

$$h(k) = \big[m_z - 2t_s \cos(ka)\big]\sigma_z - 2\,t_{so} \sin(ka)\,\sigma_y. \quad (3)$$

[1] Following the convention in quantum mechanics, we adopt the Dirac notation to represent the eigenvector.

Here, $t_s, t_{so}, m_z$ are Hamiltonian parameters. For $t_{so} \neq 0$, the system is in a topological phase when $m_z < 2t_s$ and in a trivial phase when $m_z > 2t_s$. We fix $t_s = 1$, $t_{so} = 0.3$, and generate a dataset of Eigensystem vectors across $m_z \in [0, 4]$.

*Clustering results —* Table 1 compares several clustering methods on the unlabeled Eigensystem vectors from this model. GMM attains a perfect 100% accuracy in distinguishing topological ($m_z < 2$) and trivial ($m_z > 2$) phases. In contrast, popular baseline methods including diffusion map [14] all perform worse.

| Method | Accuracy (%) | ARI | AMI |
|---|---|---|---|
| Gaussian Mixture | 100.0 | 1.00 | 1.00 |
| $k$-means | 95.4 | 0.82 | 0.77 |
| Spectral Clustering | 88.6 | 0.60 | 0.59 |
| Diffusion Map | 61.4 | 0.05 | 0.16 |

Table 1: Clustering results on Eigensystem vectors of the 1D spin-orbit Fermi chain.

The GMM also precisely localizes the phase transition with $\mathcal{O}(10^{-5})$ precision by identifying the $|\mathrm{Eig}\rangle_{(m_z)}$ whose responsibilities for the two Gaussians are almost equal. Details of the clustering result metrics and GMM algorithm are in appendix A and appendix C respectively.

*Noise robustness —* To assess stability, we add random Gaussian noise to each feature of the Eigensystem vector. As summarized in table 2, even noise levels comparable to a few times the feature standard deviations degrade GMM performance only modestly.

| Noise Level | Accuracy (%) | ARI | AMI |
|---|---|---|---|
| $0.5\,\sigma$ | 97.5 | 0.90 | 0.85 |
| $1\,\sigma$ | 96.5 | 0.87 | 0.79 |
| $2\,\sigma$ | 93.6 | 0.76 | 0.66 |
| $3\,\sigma$ | 89.9 | 0.64 | 0.53 |

Table 2: Noise robustness. $\sigma$ is the feature-wise standard deviation, serving as a reference scale.

## 4. New Physics Insights

The success of GMM offers an intuitive geometric interpretation: **each topological phase corresponds to a distinct multivariate Gaussian in the high-dimensional Eigensystem space**.

This perspective enriches our understanding of topological phases, and worth further exploration — e.g., in figure 1, the topological phase cluster forms multiple sectors, whereas the trivial phase "trivially" forms a single cluster.

Another insight is: as suggested in recent theoretical results [40, 41, 42], local Hamiltonians typically have $\mathcal{O}(N)$ degrees of freedom, $|Eig\rangle$ as the full set of eigenvectors and eigenvalues *over-parameterizes* the system, analogous to *data augmentation* – **each eigenvector and the set of eigenvalues provide different "views" of the same Hamiltonian**.

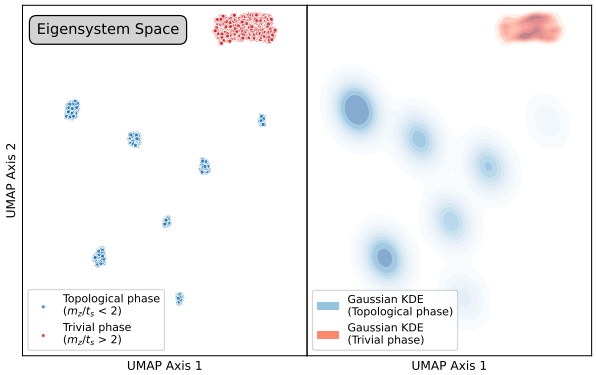


Fig. 1: UMAP visualization of Eigensystem vectors for the spin-orbit Fermi chain. **(Left)** Points are colored by their true phase (topological or trivial). **(Right)** Fitted GMM components projected into the same space.

## 5. Scalability and Generalization

Since many local Hamiltonians effectively have $\mathcal{O}(N)$ degrees of freedom, despite the raw Eigensystem vector dimension being $\mathcal{O}(N^2)$, we can use standard Principal Component Analysis (PCA) to reduce dimensionality to $\mathcal{O}(N)$, without losing accuracy in this example (table 3).

| Method | Accuracy (%) | ARI | AMI |
|---|---|---|---|
| Gaussian Mixture | 100.0 | 1.00 | 1.00 |
| $k$-means | 77.5 | 0.82 | 0.95 |
| Spectral Clustering | 33.7 | 0.24 | 0.75 |
| Diffusion Map | 16.4 | 0.05 | 0.61 |

Table 3: Clustering results after PCA reduces each Eigensystem vector to $\mathcal{O}(N)$ dimensions.

To test broader applicability, we examined several other 1D models, including the Su-Schrieffer-Heeger chain (SSH), a non-Hermitian SSH variant (nH-SSH), and the Kitaev $p$-wave superconductor (Kitaev). Table 4 shows that GMM again reliably classifies phases, often achieving near-perfect accuracy.

| | SoC Fermi | SSH | nH-SSH | Kitaev |
|---|---|---|---|---|
| Acc. (%) | 100.0 | 99.9 | 100.0 | 100.0 |

Table 4: GMM clustering performance on various 1D topological models.

## 6. Conclusion

We introduced an *Eigensystem-based representation* that, when combined with a Gaussian Mixture Model, provides a simple and effective way to discover topological phases from Hamiltonian data. The representation requires minimal domain knowledge, yields precise phase boundaries, and generalizes to multiple models. The success of GMM further implies a transparent interpretation: topological phases form well-separated Gaussian clusters in Eigensystem space.

**References**


[1] Evert P. L. Van Nieuwenburg, Ye-Hua Liu, and Sebastian D. Huber. Learning phase transitions by confusion. *Nature Phys*, 13(5):435–439, May 2017.

[2] Juan Carrasquilla and Roger G. Melko. Machine learning phases of matter. *Nature Phys*, 13(5):431–434, May 2017.

[3] Gino Cassella, Halvard Sutterud, Sam Azadi, N. D. Drummond, David Pfau, James S. Spencer, and W. M. C. Foulkes. Discovering Quantum Phase Transitions with Fermionic Neural Networks. *Phys. Rev. Lett.*, 130(3):036401, January 2023.

[4] Zhuo Cheng and Zhenhua Yu. Supervised Machine Learning Topological States of One-Dimensional Non-Hermitian Systems. *Chinese Phys. Lett.*, 38(7):070302, July 2021.

[5] Kelvin Ch'ng, Juan Carrasquilla, Roger G. Melko, and Ehsan Khatami. Machine Learning Phases of Strongly Correlated Fermions. *Phys. Rev. X*, 7(3):031038, August 2017.

[6] N. L. Holanda and M. A. R. Griffith. Machine learning topological phases in real space. *Phys. Rev. B*, 102(5):054107, August 2020.

[7] Patrick Huembeli, Alexandre Dauphin, and Peter Wittek. Identifying quantum phase transitions with adversarial neural networks. *Physical Review B*, 97(13):134109, April 2018.

[8] Si Jiang, Sirui Lu, and Dong-Ling Deng. Adversarial machine learning phases of matter. *Quantum Front*, 2(1):15, November 2023.

[9] Monika Richter-Laskowska, Marcin Kurpas, and Maciej Maśka. A learning by confusion approach to characterize phase transitions. *Physical Review E*, 108(2):024113, August 2023.

[10] Akinori Tanaka and Akio Tomiya. Detection of Phase Transition via Convolutional Neural Networks. *J. Phys. Soc. Jpn.*, 86(6):063001, June 2017.

[11] Yuan-Hong Tsai, Meng-Zhe Yu, Yu-Hao Hsu, and Ming-Chiang Chung. Deep learning of topological phase transitions from entanglement aspects. *Phys. Rev. B*, 102(5):054512, August 2020.

[12] Pengfei Zhang, Huitao Shen, and Hui Zhai. Machine Learning Topological Invariants with Neural Networks. *Phys. Rev. Lett.*, 120(6):066401, February 2018.

[13] Ling-Feng Zhang, Ling-Zhi Tang, Zhi-Hao Huang, Guo-Qing Zhang, Wei Huang, and Dan-Wei Zhang. Machine learning topological invariants of non-Hermitian systems. *Phys. Rev. A*, 103(1):012419, January 2021.

[14] Joaquin F. Rodriguez-Nieva and Mathias S. Scheurer. Identifying topological order through unsupervised machine learning. *Nat. Phys.*, 15(8):790–795, August 2019.

[15] Li-Wei Yu and Dong-Ling Deng. Unsupervised Learning of Non-Hermitian Topological Phases. *Phys. Rev. Lett.*, 126(24):240402, June 2021.

[16] Mathias S. Scheurer and Robert-Jan Slager. Unsupervised machine learning and band topology. *Phys. Rev. Lett.*, 124(22):226401, June 2020.

[17] Alexander Lidiak and Zhexuan Gong. Unsupervised Machine Learning of Quantum Phase Transitions Using Diffusion Maps. *Phys. Rev. Lett.*, 125(22):225701, November 2020.

[18] Yefei Yu, Li-Wei Yu, Wengang Zhang, Huili Zhang, Xiaolong Ouyang, Yanqing Liu, Dong-Ling Deng, and L.-M. Duan. Experimental unsupervised learning of non-Hermitian knotted phases with solid-state spins. *npj Quantum Inf*, 8(1):116, September 2022.

[19] Nicolas Sadoune, Giuliano Giudici, Ke Liu, and Lode Pollet. Unsupervised interpretable learning of phases from many-qubit systems. *Phys. Rev. Research*, 5(1):013082, February 2023.

[20] Yang Long, Jie Ren, and Hong Chen. Unsupervised Manifold Clustering of Topological Phononics. *Phys. Rev. Lett.*, 124(18):185501, May 2020.

[21] Yang Long and Baile Zhang. Unsupervised Data-Driven Classification of Topological Gapped Systems with Symmetries. *Phys. Rev. Lett.*, 130(3):036601, January 2023.

[22] Yang Long, Haoran Xue, and Baile Zhang. Unsupervised Learning of Topological Non-Abelian Braiding in Non-Hermitian Bands, January 2024.

[23] En-Jui Kuo and Hossein Dehghani. Unsupervised learning of interacting topological and symmetry-breaking phase transitions. *Phys. Rev. B*, 105(23):235136, June 2022.

[24] Niklas Käming, Anna Dawid, Korbinian Kottmann, Maciej Lewenstein, Klaus Sengstock, Alexandre Dauphin, and Christof Weitenberg. Unsupervised machine learning of topological phase transitions from experimental data. *Mach. Learn.: Sci. Technol.*, 2(3):035037, September 2021.

[25] Eliska Greplova, Agnes Valenti, Gregor Boschung, Frank Schäfer, Niels Lörch, and Sebastian D Huber. Unsupervised identification of topological phase transitions using predictive models. *New J. Phys.*, 22(4):045003, April 2020.

[26] Jiangzhi Chen, Zi Wang, Yu-Tao Tan, Ce Wang, and Jie Ren. Machine Learning of Knot Topology in Non-Hermitian Band Braids, January 2024.

[27] Yanming Che, Clemens Gneiting, Tao Liu, and Franco Nori. Topological quantum phase transitions retrieved through unsupervised machine learning. *Phys. Rev. B*, 102(13):134213, October 2020.

[28] Ming-Chiang Chung, Guang-Yu Huang, Ian P. McCulloch, and Yuan-Hong Tsai. Deep Learning of Phase Transitions for Quantum Spin Chains from Correlation Aspects. *Phys. Rev. B*, 107(21):214451, June 2023.

[29] Eran Lustig, Or Yair, Ronen Talmon, and Mordechai Segev. Identifying Topological Phase Transitions in Experiments Using Manifold Learning. *Phys. Rev. Lett.*, 125(12):127401, September 2020.

[30] Nannan Ma and Jiangbin Gong. Unsupervised identification of Floquet topological phase boundaries. *Phys. Rev. Research*, 4(1):013234, March 2022.

[31] T. Mendes-Santos, X. Turkeshi, M. Dalmonte, and Alex Rodriguez. Unsupervised Learning Universal Critical Behavior via the Intrinsic Dimension. *Phys. Rev. X*, 11(1):011040, February 2021.

[32] Sungjoon Park, Yoonseok Hwang, and Bohm-Jung Yang. Unsupervised learning of topological phase diagram using topological data analysis. *Phys. Rev. B*, 105(19):195115, May 2022.

[33] Yuan-Hong Tsai, Kuo-Feng Chiu, Yong-Cheng Lai, Kuan-Jung Su, Tzu-Pei Yang, Tsung-Pao Cheng, Guang-Yu Huang, and Ming-Chiang Chung. Deep learning of topological phase transitions from entanglement aspects: An unsupervised way. *Phys. Rev. B*, 104(16):165108, October 2021.

[34] Lei Wang. Discovering phase transitions with unsupervised learning. *Phys. Rev. B*, 94(19):195105, November 2016.

[35] Jielin Wang, Wanzhou Zhang, Tian Hua, and Tzu-Chieh Wei. Unsupervised learning of topological phase transitions using the Calinski-Harabaz index. *Phys. Rev. Research*, 3(1):013074, January 2021.

[36] Yuan Yang, Zheng-Zhi Sun, Shi-Ju Ran, and Gang Su. Visualizing Quantum Phases And Identifying Quantum Phase Transitions By Nonlinear Dimensionality Reduction. *Phys. Rev. B*, 103(7):075106, February 2021.

[37] Li-Wei Yu, Shun-Yao Zhang, Pei-Xin Shen, and Dong-Ling Deng. Unsupervised learning of interacting topological phases from experimental observables. *Fundamental Research*, page S2667325823000067, January 2023.

[38] Daria Zvyagintseva, Helgi Sigurdsson, Valerii K. Kozin, Ivan Iorsh, Ivan A. Shelykh, Vladimir Ulyantsev, and Oleksandr Kyriienko. Machine learning of phase transitions in nonlinear polariton lattices. *Communications Physics*, 5(1):8, January 2022.

[39] Jian-Song Pan, Xiong-Jun Liu, Wei Zhang, Wei Yi, and Guang-Can Guo. Topological superradiant states in a degenerate Fermi gas. *Physical review letters*, 115(4):045303, 2015.

[40] Xiao-Liang Qi and Daniel Ranard. Determining a local Hamiltonian from a single eigenstate. *Quantum*, 3:159, July 2019.

[41] James R. Garrison and Tarun Grover. Does a Single Eigenstate Encode the Full Hamiltonian? *Physical Review X*, 8(2):021026, April 2018.

[42] Eyal Bairey, Itai Arad, and Netanel H. Lindner. Learning a Local Hamiltonian from Local Measurements. *Physical Review Letters*, 122(2):020504, January 2019.

## Appendix A. Clustering Metrics

### *1.1 Adjusted Rand Index (ARI):*

The Rand Index computes the similarity between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. The raw Rand Index is then "adjusted for chance" into the ARI score using the following formula:

$$\text{RI} = \frac{a+b}{\binom{N_{samples}}{2}} \tag{A1}$$

$$\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]} \tag{A2}$$

Where RI is the Rand Index, $E[\text{RI}]$ is the expected value of the Rand Index. The ARI ranges from -1 to 1. A score close to 1 indicates that the clusterings are almost identical, a score close to 0 indicates that the clusterings are random, and a score close to -1 indicates that the clusterings are dissimilar.

### *1.2 Adjusted Mutual Information (AMI):*

Mutual Information of two variables is a measure of the amount of information obtained about one variable through observing the other variable. AMI is an adjustment of the Mutual Information (MI) score to account for chance. It accounts for the

fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared. The AMI removes this bias:

$$\mathrm{MI} = \sum_{u\in U}\sum_{v\in V} p(u,v)\log\left(\frac{p(u,v)}{p(u)p(v)}\right) \tag{A3}$$

$$\mathrm{AMI} = \frac{\mathrm{MI} - E[\mathrm{MI}]}{\max(H(U),H(V)) - E[\mathrm{MI}]} \tag{A4}$$

Where MI is the Mutual Information, $E[\mathrm{MI}]$ is the expected value of the Mutual Information, $H(U)$ and $H(V)$ are the entropies of the two clusterings. AMI ranges from 0 to 1. A score of 0 indicates two clusterings are independent, and 1 indicates that they are identical.

*1.3 Accuracy:*

Simply the percentage of correctly identified data points.

## Appendix B. 1D Spin-Orbit Coupled Fermi Chain

We study a 1D optical lattice with Raman-induced artificial spin-orbit coupling. The optical lattice is of great significance for quantum simulation based on ultra-cold atoms because most condensed matter exists in the form of crystals. Firstly, the lattice depth, periodic length, and atom-atom interaction of the optical lattice can be adjusted. Secondly, compared with solid materials, the optical lattice itself does not have problems such as disorder, lattice vibration, and structural defects, but these factors can be realized in the optical lattice through quantum control. In addition, because it does not carry a net charge, the ultra-cold atomic gas in the optical lattice has almost no coupling with the outside world, and the environment is very clean. Compared to the electron motion in solid materials, the time scale of ultra-cold atomic motion is in the millisecond or even second range, so there is no need for fine ultrafast optical means for detection in the experiment. The above factors make the ultra-cold atomic gas in the optical lattice very suitable for simulating the electron motion in crystals.

Spin-orbit coupling is a physical effect caused by the coupling between the internal spin degree of freedom and the external motion degree of freedom of particles. Spin-orbit coupling is widespread in nature: in atomic physics, the coupling of electron spin and orbital motion leads to the fine structure of atomic energy levels; in condensed matter physics, spin-orbit coupling is the basis of many novel physical phenomena, such as the spin Hall effect, topological insulators, etc. At present, the main method to create artificial spin-orbit coupling experimentally is to couple the internal states of atoms through a two-photon Raman process. Using the single-band tight-binding approximation, we can obtain the following Hamiltonian:

$$\begin{aligned}\mathcal{H} =& m_z\sum_{\sigma,j}\xi_\sigma\hat{c}^\dagger_{j\sigma}\hat{c}_{j\sigma} - t_s\sum_{\sigma,\langle i,j\rangle}\hat{c}^\dagger_{i\sigma}\hat{c}_{j\sigma}\\ &+ t_{so}\sum_j\left[(-1)^j\,\hat{c}^\dagger_{j\uparrow}\hat{c}_{j+1\downarrow} + H.c.\right]\end{aligned} \tag{A5}$$

Further performing a unitary transformation $\hat{c}_{j\downarrow}\to(-1)^j\,\hat{c}_{j\downarrow}$, and then a Fourier transformation $\hat{c}_{j\sigma} = \frac{1}{\sqrt{N}}\sum_{k_x}\hat{c}_{k_x\sigma}e^{ik_xja}$, where $N$ is the number of lattice points, and $a$ is the lattice constant, we can get the momentum space representation of the Hamiltonian:

$$\mathcal{H} = \sum_{k_x}\psi^\dagger_{k_x,\sigma}\begin{pmatrix} m_z - 2t_s\cos(ka) & 2it_{so}\sin(ka)\\ -2it_{so}\sin(ka) & 2t_s\cos(ka) - m_z\end{pmatrix}\psi_{k_x,\sigma}$$

$$\Longleftrightarrow \mathcal{H} = \sum_{k_x}\begin{pmatrix}\hat{c}^\dagger_{k_x\uparrow} & \hat{c}^\dagger_{k_x\downarrow}\end{pmatrix}h\left(k_x\right)\begin{pmatrix}\hat{c}_{k_x\uparrow}\\ \hat{c}_{k_x\downarrow}\end{pmatrix} \tag{A6}$$

$$h\left(k_x\right) = \left[m_z - 2t_s\cos\left(k_xa\right)\right]\sigma_z - 2t_{so}\sigma_y\sin\left(k_xa\right) \tag{A7}$$

By exactly diagonalizing this Hamiltonian, we obtain the dispersion relation:

$$E_\pm(k) = \pm\sqrt{(2t_s\cos(ka) - m_z)^2 + (2t_{so}\sin(ka))^2} \tag{A8}$$

When $t_{so}\neq 0$, the system undergoes a topological phase transition at the critical point $m_z = 2t_s$, characterized by a change in the topological invariant, that is, the change in the winding number (also called Zak phase). When $m_z < 2t_s$, the system is in a topological phase, and the winding number is 1; when $m_z > 2t_s$, the system is in a trivial phase, and the winding number is 0. This is a Su-Schrieffer-Heeger-type phase transition.

To generate the input dataset $\mathbf{X} := \{|Eig\rangle\}$, we first set $t_{so} = 0.3$ and $t_s = 1$ and then randomly generate $M = 40000$ values of $m_z$ in the range $[0, 2t_s)$ for topological phases, and $M = 40000$ values in the range $(2t_s, 4t_s]$ for trivial phases. Next, for each $m_z$, we diagonalize the corresponding Hamiltonian $\mathcal{H}(m_z)$, whereupon splice the obtained eigenstates and eigenvalues into a Eigensystem vector $|Eig(m_z)\rangle$ as defined in Eq.1. Now we have $\mathbf{X} = \{|Eig\rangle\} = \{|Eig^t\rangle\}\cup\{|Eig^n\rangle\}$.

## Appendix C. Gaussian Mixture Model

The basic idea of GMM is to approximate the data distribution as a weighted sum of several Gaussian distributions. In other words, GMM assumes that the data is generated from a mixture of Gaussian distributions with unknown parameters—the mean $\mu_i$ and covariance $\Sigma_i$ of each Gaussian, and $\pi_i$ that determine the weight of each Gaussian in the mixture. It is a generalized version of the KMeans clustering algorithm, with the advantage that it allows for soft clustering and performs well even when small overlaps

between clusters exist as long as they are Gaussian-shaped. The procedure is as follows:

**Initialization:** Select the number of clusters $k$ you want to identify in your data. Randomly initialize the parameters of the Gaussians - the mean $\mu_\mathbf{i}$, the covariance matrix $\mathbf{\Sigma_i}$, and the mixing coefficients $\pi_i$ for each cluster.

**Expectation-Maximization (EM):** *Expectation (E) Step:* Calculate the posterior probability of each data point belonging to each cluster, also known as the "responsibility" that cluster $j$ takes for data point $x_i$. This is computed using the current parameter values.

$$w_{ij} = \frac{\pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_i|\mu_l, \Sigma_l)} \tag{A9}$$

Where $w_{ij}$ is the responsibility that the $j^{th}$ Gaussian takes for the $i^{th}$ data point $x_i$; $\pi_j$ is the mixing coefficient of the $j^{th}$ Gaussian; $\mathcal{N}(x_i|\mu_j, \Sigma_j)$ is the probability density function of the $j^{th}$ Gaussian at $x_i$.

*Maximization (M) Step:* Update the parameters of the Gaussians using the current responsibilities:

$$\mu_j = \frac{\sum_{i=1}^{N} w_{ij} x_i}{\sum_{i=1}^{N} w_{ij}} \tag{A10}$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} w_{ij}(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{N} w_{ij}} \tag{A11}$$

$$\pi_j = \frac{1}{N} \sum_{i=1}^{N} w_{ij} \tag{A12}$$

Where $N$ is the total number of data points.

*Convergence:* Repeat the E and M steps until the parameters do not change significantly, or a maximum number of iterations is reached.

## Appendix D. Silhouette Analysis

Note that before applying GMM we need to specify the number of clusters $k$ we want to identify in our data. In order to find clusters corresponding to actual physical phases, we need to choose the “right” number that best reflects the underlying global structure of the data. A common approach is to use the silhouette analysis, which is a graphical tool to evaluate the performance of clustering algorithms, and thus help us determine the optimal number of clusters.

Fig.A1 shows the distributions of silhouette scores as $k$ varies from 2 to 5. From which we can see that $k_c = 2$ yields the highest / peak of average silhouette score $SC(\mathbf{X}, k_c)$. Moreover, we see a balanced, single peak distribution of silhouette scores, instead of having unbalanced multiple peaks in other cases, which indicate the appropriateness of set the number of phases to $k_c = 2$.

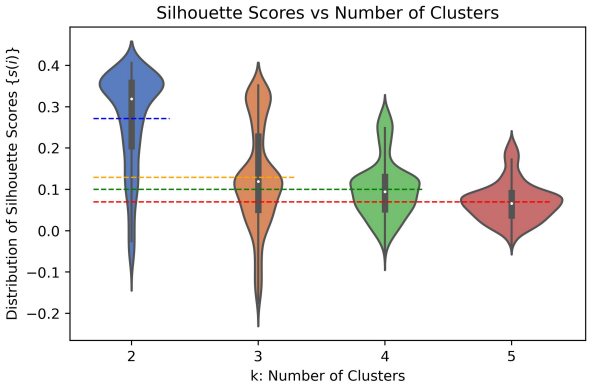


Fig. A1: The violin plots of the distributions of silhouette scores as $k$ varies from 2 to 5.