Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084503

REVIEW

Topological Materials and Machine Learning: A Comprehensive Review

Jing-Wen Gao[1,2], **Yunan He**[1,*] **and Jian Liu**[1,*]

[1]Mathematical Science Research Center, Chongqing University of Technology, Chongqing, China
[2]Mathematical Science School, Huazhong University of Science and Technology, Wuhan, China

*Corresponding Authors: Yunan He. Email: heyunan@cqut.edu.cn; Jian Liu. Email: liujian@cqut.edu.cn

Received: 23 April 2026; Accepted: 24 June 2026; Published: 23 July 2026

ABSTRACT: The unique topological properties of the electronic band structures in topological materials have increasingly attracted attention in both fundamental research and next-generation technological applications. With the rise of machine learning, the connection between topological materials and machine learning has deepened significantly. This review systematically summarizes the interaction between these two fields, tracing the history of their mutual promotion and synergistic development. We further examine the transformative impact of machine learning across multiple domains of topological materials, with a particular focus on recent progress in inverse design and generation of topological materials, topological superconductivity, and the mapping of topological phase diagrams. This paper also analyzes the data complexity arising from the topological properties of electronic structures and notes that topological deep learning is naturally suited for processing such complex data. In addition, we discuss the current limitations of existing machine learning methods and propose potential strategies to address these challenges. This review aims to provide a fundamental reference for researchers seeking to advance the bidirectional integration of machine learning and topological materials.

KEYWORDS: Topological materials; machine learning; symmetry-based theories; data-driven approaches; topological deep learning

1 Introduction

Topological materials are characterized by electronic band structures [1] that exhibit nontrivial topological arrangements in their geometric configuration. The term "topological" implies that the topological class of the electronic structure remains invariant under parameter tuning, as long as the energy gap is not closed. Such electronic band structures can give rise to robust surface states and unconventional electromagnetic responses [2–5], including the presence of surface or edge states, backscattering-free transport, topological Fermi-arc surface states, highly anisotropic negative longitudinal magnetoresistance induced by the chiral anomaly, nontrivial quantum oscillations, nonlocal transport behavior, light-induced anomalous Hall effects, non-Abelian statistics and related superconducting phenomena driven by spin-momentum locking, as well as a variety of other exotic quantum effects. Owing to these unique properties, topological materials hold great promise for applications in dissipationless electronic devices [6] and spintronic technologies [7].

However, the total number of existing topological materials, their specific types, and their relative abundance among known compounds remain unclear. A central and long-standing question in the field is how to determine whether a given material is topologically trivial or nontrivial. Thanks to advances in symmetry indicator theory [8], topological quantum chemistry (TQC) [9], and first-principles computational

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

methods [10], it is now possible to classify topologically nontrivial materials based on symmetry indicators, elementary band representations, and compatibility relations [11–15]. Building on these theoretical foundations, researchers can effectively identify topological semimetals (TSMs) [16] and topological insulators (TIs) [17–19].

By combining density functional theory (DFT) with topological quantum chemistry or symmetry indicator theory in high-throughput computations, researchers have identified tens of thousands of topological materials by analyzing the symmetry of wavefunctions in crystalline compounds from the Inorganic Crystal Structure Database (ICSD), leading to the construction of several dedicated databases [20]. However, challenges remain in discovering high-quality topological materials suitable for practical technological applications. On one hand, the sheer size and complexity of these databases make manual data analysis highly inefficient. On the other hand, it remains difficult to determine, based on experience alone, which material features most strongly influence topological properties, particularly those that provide meaningful chemical insight.

These challenges have motivated the development of novel machine learning approaches to encode various material properties, including structural and other relevant information. By learning from available data, machine learning models can identify the key descriptors that govern topological behavior and make automated predictions based on these features [21]. From this perspective, the need to explore topological materials has in turn driven the advancement and innovation of machine learning methodologies.

Conversely, machine learning has also accelerated progress in the field of topological materials. The widespread application of machine learning methods relies critically on the availability of large-scale datasets, and the establishment of these databases has enabled extensive utilization of machine learning in topological materials research. Compared to first-principles calculations, machine learning offers advantages in computational speed and ease of use. In tasks such as discovering and designing new topological materials, as well as exploring their types, properties, and structures, machine learning can achieve equivalent results several orders of magnitude faster than traditional computational approaches.

In this review, Section 2 presents the theoretical and data foundations for machine learning in topological materials. In Section 2.1, we provide an overview of TQC and symmetry indicator theories. Section 2.2 begins with the construction of large-scale topological materials databases. We then describe high-throughput screening of topological materials using computational design programs combined with first-principles calculations, highlighting how computational approaches can accelerate the discovery of new topological materials. Unlike traditional programming, in which the computer is explicitly instructed on each step, machine learning involves providing the computer with large datasets and desired outputs, allowing it to identify patterns and relationships in the data and apply these insights to new cases. Section 2.3 introduces several machine learning methods that have been developed under the guidance of data from topological materials.

Section 3 reviews the developmental stages of machine learning applications in topological materials, including the theoretical model validation, the physical-feature input, and the direct structure and composition input stage. Section 4 focuses on specific applications of machine learning in topological materials, emphasizing breakthroughs in inverse design and generation of topological materials, topological superconductivity, and mapping of topological phase diagrams. In Section 5, we first discuss the topological characteristics of electronic structures in topological materials and then highlight the advantages of topological deep learning (TDL) in handling topology-related data extracted from electronic structures. Section 6 addresses the challenges faced by machine learning in exploring new topological materials and discusses potential strategies to overcome these challenges. The final section provides a summary of the review and concluding remarks.

Through this review, we aim to contribute to the understanding of the interplay between machine learning and topological materials and to inspire further exploration at the intersection of these two rapidly evolving fields. An integrated overview of the major application areas of machine learning in topological materials and the historical evolution of this interdisciplinary field is presented in Fig. 1.

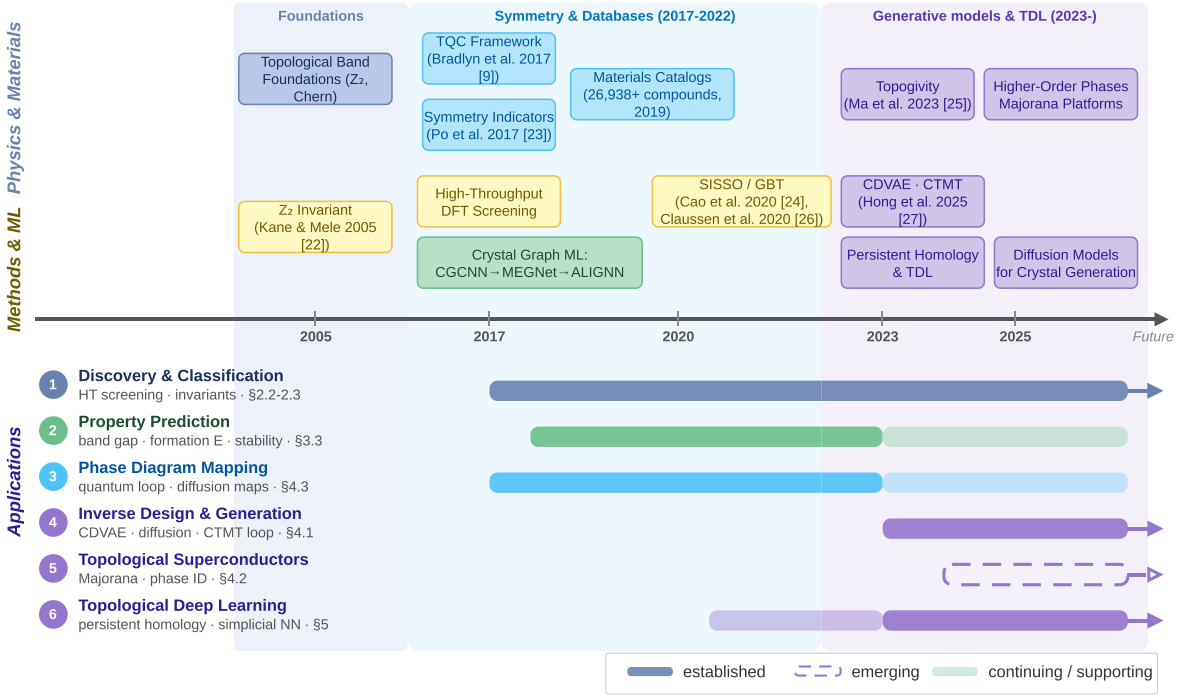

Figure 1: The coevolution of topological materials and machine learning. The upper panel presents key concepts in topological materials together with representative machine learning methods. The lower panel categorizes six major research directions in which machine learning has been applied, namely discovery and classification, inverse design and generation, topological superconductors, phase diagram construction, property prediction, and topological deep learning. The timeline illustrates how the fields of topological materials and machine learning have advanced in a mutually reinforcing manner and jointly driven scientific progress [9,22–27]. Note: Citations in the figure are shown with their reference numbers in brackets. Each citation denotes the original source of the method or concept presented.

In this review, all reported performance data are taken directly from the original publications cited. Different studies employ different datasets, data splitting strategies, evaluation metrics, and uncertainty reporting schemes. Unless explicitly stated otherwise, these results are obtained under different evaluation conditions and are therefore not directly comparable. Readers should exercise caution when comparing these performance values across different methods.

2 Theoretical and Data Foundations for Machine Learning in Topological Materials

2.1 Symmetry-Based Theories

Symmetry-based analysis is a method that utilizes the crystal symmetries of materials to diagnose topological invariants, such as the Chern number, the Chern-Simons term [4] and the $\mathbb{Z}_2$ invariant first proposed by Kane and Mele [22]. In earlier studies, only a few hundred materials were confirmed to possess nontrivial topological properties in the prediction and discovery of TIs and TSMs. Bradlyn et al. [9] proposed the TQC framework, a comprehensive electronic band theory [1] that combines a graph-theoretical description in momentum space with a dual group-theoretical description in real space, thereby revealing

the relationship between topology and localized chemical bonding. This theory systematically classifies the band structures arising from localized atomic orbitals across all 230 crystal symmetry groups, identifying nontrivial topological bands. It not only provides new interpretations for known TIs but also demonstrates a powerful capability to predict a large number of previously unknown TIs. Po et al. [23] developed a systematic approach to reveal symmetry-based TIs in all 230 space groups. This method represents band structures in terms of (elementary band representations) EBRs and then isolates topological bands by removing subsets of atomic insulators defined through localized symmetric Wannier functions. The approach uncovers previously hidden topological band symmetries, providing a systematic tool for the search for topological materials.

Subsequently, Elcoro et al. [28] extended the TQC framework to magnetic space groups, establishing a comprehensive real-space band topology theory applicable to both magnetic and nonmagnetic crystalline solids, known as Magnetic Topological Quantum Chemistry (MTQC). Using MTQC, they derived a complete set of symmetry indicators for electronic band topology, enabling the identification of symmetry-allowed bulk states, as well as anomalous surface and hinge states.

Building on this foundation, Xu et al. [29] performed the first high throughput search for magnetic topological materials using MTQC derived indicators. They developed a complete framework including magnetic indicators, co representations, and compatibility relations, along with code to calculate them from first principles. Applying this to 549 magnetic structures with DFT plus Hubbard U, they predicted several novel magnetic topological phases. The significance is clear. This work transformed the theoretical MTQC framework into a practical screening tool. Prior to this, magnetic topological materials could only be identified through manual analysis.

However, three caveats should be considered. The Hubbard U values must be chosen empirically, and sensitivity to this choice was not explored. The predictions require experimental verification. The computational cost is substantial, which may limit practical applicability. Nevertheless, this work marked an important milestone. Theoretical advances like MTQC gain real power only when paired with computational implementations that enable large scale screening. Table 1 provides a comparison between the TQC formalism and the MTQC framework.

Table 1: Comparison between the TQC framework and its magnetic extension. Based on the descriptions in [28] and [29].

Technical Achievement	TQC	MTQC
Wannier-Bloch duality	Establishes the induction relation between Wannier functions and band representations.	Extends the induction framework to magnetic orbitals and magnetic elementary band representations (MEBRs).
Graph-theoretic methods	Maps compatibility relations to graph-theoretic problems to classify allowed patterns of band connectivity.	Extends the approach to magnetic space groups, incorporating additional constraints arising from antiunitary symmetries.
Representation databases	Completes the database of small representations and EBRs for all 230 space groups.	Provides small corepresentations and MEBRs for all 1421 magnetic space groups.

(Continued)

Table 1 (continued)

Technical Achievement	TQC	MTQC
Topological classification framework	Proposes the criterion that topologically trivial bands are precisely those that can be expressed as a sum of EBRs.	Computes the symmetry indicator formulas for all 1651 double space groups and tabulates the topological phase associated with every nontrivial symmetry indicator.

2.2 High-Throughput Screening of Topological Materials in Conjunction with First-Principles Calculations

Based on the recently developed TQC framework, Vergniory et al. [30] performed the first large scale high throughput search for topological materials. Using their own code, they calculated symmetry indicators for 26,938 compounds from the ICSD and identified thousands of topological insulators and semimetals. Over 27% of the analyzed compounds exhibited nontrivial topology. The source code and data were released publicly. The significance is substantial. This work transformed the field from case by case discovery to database driven screening. Prior to this study, only a few hundred topological materials were known. However, three caveats should be noted. The study covered only non magnetic materials. Spin orbit coupling was not included for all calculations. The 27% figure is an upper bound; some identified materials may have band gaps too small for experimental observation. Despite these caveats, this work marked a turning point. The methodology became the standard, and the public code release demonstrated that large scale reproducibility is possible in computational materials science.

Topological materials exhibit unconventional linear responses in the bulk and anomalous gapless boundary states at their edges. These materials are not only of fundamental scientific interest but also hold great potential for applications in advanced low-power, high-speed spintronic, magnetoelectric, and optoelectronic devices, as well as quantum computing. However, due to the computational complexity of topological invariants or topological nodes, the identification of topological materials has historically relied heavily on materials science expertise and advanced theoretical tools, which has hindered their efficient discovery. To address this, Zhang et al. [31] developed a highly efficient, fully automated algorithm capable of diagnosing nontrivial band topology for the majority of nonmagnetic materials. The algorithm is based on the established complete correspondence between the symmetry representations of occupied bands and topological invariants. By systematically scanning nearly 40,000 materials in crystal structure databases, the algorithm identified over 8000 nontrivial topological materials. The results have been made publicly accessible through an interactive user interface within the database.

Although the diversity of spatial symmetries leads to a rapid increase in the possible topological crystalline (TC) phases of electrons, including topological crystalline insulators (TCIs), their physical realization has progressed slowly due to the practical difficulties in determining band topology computationally. The algorithm developed by Tang et al. [32] is based on the comprehensive symmetry indicators framework introduced in [23]. Analogous to the well-known Fu–Kane parity criterion for TIs, this framework enables the reliable identification of topological materials even in the absence of prior knowledge about the precise topological nature of the bands involved [33–36]. They applied their program to search for materials within a database [37] containing previously synthesized compounds, focusing on 8 of the 230 space groups. Their approach revealed topological materials with a wide range of distinct topological properties, spanning weak

and strong TIs as well as the recently proposed higher-order topological insulators (HOTIs) [12–15,35,36]. In addition, their algorithm successfully captured TSMs. The general high-throughput discovery workflow combining databases, machine learning screening, first-principles calculations, and topology diagnosis is outlined in Fig. 2.

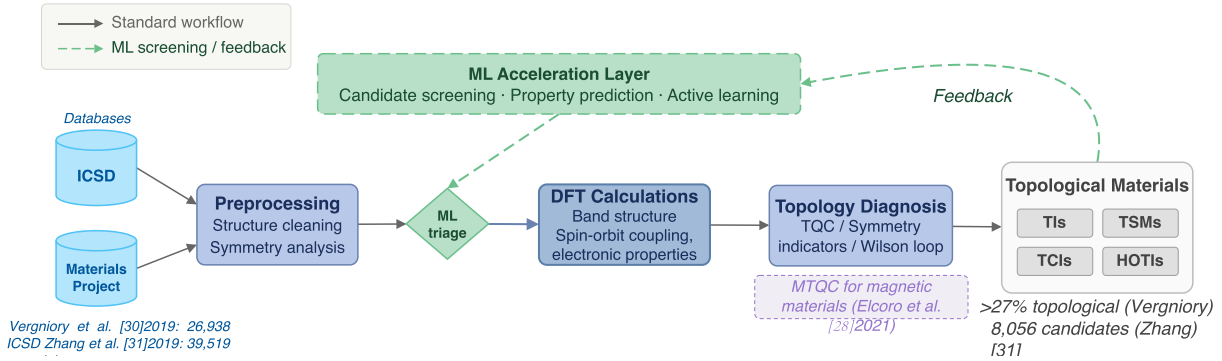

Figure 2: A typical high throughput workflow for topological materials discovery. Starting from crystallographic databases, candidate materials are first screened through a machine learning acceleration layer before undergoing computationally expensive DFT calculations and symmetry based topological diagnostics. The workflow is iterative, with feedback loops enabling the refinement of screening criteria [28,30,31].

We should note that the practical reproducibility of the aforementioned high-throughput workflows is influenced by several factors. First, the consistency between different databases. The computed properties of the same material (e.g., formation energy, band gap) can vary across databases due to differences in DFT parameters, pseudopotentials, or post-processing conventions. High-throughput screening workflows that combine multi-source data without cross-validation may introduce systematic biases. Second, the treatment of symmetry tolerances can subtly alter material classification. DFT codes typically identify crystal symmetries by applying numerical thresholds. Minor variations in these tolerances can change whether a structure is classified as cubic or tetragonal, thereby affecting symmetry-derived properties and the assignment of topological invariants. Third, numerical convergence in DFT calculations directly determines the reliability of predicted properties. Parameters such as plane-wave cutoff energy, k-point sampling density, and force convergence criteria must be systematically tested for convergence. Insufficient convergence may lead to false positives, meaning materials that appear promising in screening but fail under stricter thresholds.

Therefore, to enhance the reproducibility and reliability of high-throughput screening studies, the following practices should be adopted: (i) cross-validation when using multi-source data, (ii) explicit reporting of symmetry detection tolerances, and (iii) convergence tests for DFT calculations with well-defined numerical thresholds.

2.3 Data-Driven Approaches

Significant progress has been made in predicting novel topological materials using high-throughput, symmetry-based descriptors derived from empirical rules. To date, these methods have primarily been applied to materials within existing databases, but they are severely limited to systems with well-defined symmetries, leaving a vast portion of the materials space unexplored.

Cao et al. [24] developed an artificial intelligence based descriptor for topological insulators using the SISSO (Sure Independence Screening and Sparsifying Operator) method. The descriptor uses only atomic

numbers and electronegativities. It does not require symmetry information or detailed band structures. Applied to chalcogenide alloys, nearly half were identified as topological insulators. The significance is twofold. First, expensive band structure calculations are not always necessary. A simple elemental descriptor can be surprisingly effective. Second, the descriptor works without symmetry information, making it applicable to low symmetry or disordered systems. However, three limitations should be noted. The descriptor was only tested on chalcogenides. Its transferability to other material families is unknown. It also provides only a binary prediction without identifying the specific topological invariant. The nearly 50% hit rate is impressive but applies only within this specific chemical space. The broader implication is important. The SISSO method produces interpretable descriptors rather than black box predictions. This is valuable for materials discovery, where understanding why a material is topological is as important as knowing that it is. Future work should test transferability to other families and extend the approach to predict specific invariants.

Andrejevic et al. [38] took a different approach, training a neural network on computed X-ray absorption near-edge structure (XANES) spectra to classify materials as topological or trivial. Beyond achieving F1 scores of 89% and 93% for topological and non-topological materials, respectively, their work has three notable implications: (i) the model identified Weyl semimetals that TQC had missed, suggesting spectral features may capture topology beyond symmetry indicators; (ii) the method works on non-crystalline and non-cleavable materials, filling a gap in experimental characterization; and (iii) the simplicity of X-ray absorption spectroscopy (XAS) enables in situ observation of topological phase transitions.

However, the approach has limitations that the original study does not fully address. The model's reliance on computed spectra for training means its performance on truly novel materials—those whose XANES spectra have not been computed—remains untested. Moreover, the spectral features that correlate with topological class are not explicitly identified, leaving the method as a black-box classifier rather than a tool for physical insight. Future work could combine this spectroscopic approach with interpretable machine learning techniques to extract which spectral regions carry topological information.

The topogivity rule [25] offers a composition-only heuristic. Each element is assigned a machine-learned "topologicality" value, and the weighted average of these values predicts whether a material is topologically nontrivial (positive average) or trivial (negative average), with over 80% accuracy. The rule's importance is twofold. First, it demonstrates that topological propensity is largely encoded in elemental chemistry rather than precise atomic arrangements, corroborating findings from Claussen et al. [26]. Second, because it does not require crystal symmetry, it can screen materials that evade symmetry-based diagnosis. However, topogivity's finite accuracy, approximately 20% error, makes it suitable only as a pre-filter in high-throughput workflows, not as a definitive classifier. The trade-off between topogivity's broad applicability and structure-based models' quantitative precision suggests that hybrid workflows, which use topogivity for initial screening followed by graph-based models, may be the most practical path forward.

The topologicality framework was later extended by Xu et al. [39] to address magnetic and two-dimensional materials. Three modifications were introduced. They incorporated magnetic materials into the training set, added the Hubbard U parameter to account for electron correlation, and replaced the linear weighting scheme with a convolutional neural network (CNN). These changes allowed the method to discover six classes of previously unknown Chern insulators in 2D materials, four of which exhibit a global band gap.

However, the extension raises two methodological concerns that the original study did not fully address. First, the Hubbard U parameter introduces an additional degree of freedom that must be chosen empirically. The sensitivity of the CNN's predictions to different U values was not systematically explored, leaving uncertainty about whether the discovered Chern insulators are robust predictions or artifacts of specific U choices. Second, while the CNN improves predictive accuracy, it sacrifices interpretability. The original

Topogivity's key advantage was its simplicity and manual calculability. The CNN version loses this advantage without providing a clear mechanistic understanding of why certain compositions favor topological order.

From a broader perspective, the trade-off exemplified by this work is typical in the field. Simple heuristics like the original Topogivity offer interpretability and broad applicability but limited accuracy. Complex models like the CNN extension offer higher accuracy at the cost of transparency and generalizability across different Hubbard U regimes. For magnetic topological materials, where DFT calculations are already expensive and sensitive to U parameters, a more promising direction may be to incorporate physically informed priors rather than purely data-driven learning. The CNN approach is a step forward in predictive power, but whether it represents a step forward in physical understanding remains an open question.

3 Evolution of Machine Learning Applications in Topological Materials

3.1 Theoretical Model Validation

Early research primarily focused on validating the feasibility of machine learning using theoretical physical models. Carrasquilla and Melko [40] showed that a neural network can identify ferromagnetic and paramagnetic phases from Ising spin configurations and locate the phase transition. The network learned to recognize magnetization, the conventional order parameter. This work opened the direction of using machine learning to identify phases without predefined order parameters. It laid the foundation for topological phase recognition, where conventional order parameters often do not exist. However, the Ising model is simple. Success here does not guarantee success on more complex systems. Verifying what the network learned was possible because the correct order parameter was known. For topological phases without such a parameter, this verification is much harder. Despite these caveats, this work was a turning point. It showed that machine learning can be a tool for discovery, not just classification.

van Nieuwenburg et al. [41] proposed learning by confusion, a method that identifies phase transitions without order parameters or prior knowledge. It assigns labels based on hypothetical boundaries and trains a neural network. When the hypothetical boundary matches the true one, validation accuracy peaks. The method successfully detected topological, thermal, and many body localization transitions. The key advantage is that it works without knowing what to look for. This is valuable for topological phases where order parameters do not exist. However, the method requires scanning over many hypothetical boundaries, which is computationally costly. It also identifies where a transition occurs but does not characterize the phases themselves. Despite these limitations, this work introduced a clever and widely applicable approach.

Zhang et al. [42] showed that a neural network can predict the winding number of one-dimensional chiral topological insulators from local Hamiltonian parameters with nearly perfect accuracy. The network trained on a subset of winding numbers generalized to larger unseen numbers, indicating that it learned the winding number formula rather than memorizing samples. Analysis of the network's internal structure confirmed this interpretation.

The work's significance is twofold. It provides rare evidence that neural networks can learn a mathematical invariant from data without explicit programming. It also challenges the view that deep learning models are merely interpolating black boxes. However, two limitations exist. The study was restricted to one-dimensional toy models, and the network required Hamiltonian parameters as input, which are not experimentally accessible. Extending this approach to realistic materials remains an open challenge. Nevertheless, the work established an important proof of concept for learning topological invariants from data.

A different strategy was pursued by Cheng et al. [43], who used real-space observables rather than momentum-space Hamiltonian parameters to identify Chern numbers in the Haldane model. Unlike

Zhang et al. [42], who required Hamiltonian parameters as input, Cheng et al. designed a disk geometry with sixfold rotational symmetry and used physically measurable quantities as features. These included chiral edge states, where the number of edge branches gives the Chern number magnitude and the direction of edge currents gives its sign. They also used the real-space particle density distribution, where edge states concentrate at the boundary while bulk states are distributed in the interior. They further used the local density of states, where a topological phase exhibits finite edge-state local density of states within the energy gap while a trivial phase does not.

The key contribution of this work is not the specific choice of features but the demonstration that machine learning can extract topological information without access to the Hamiltonian or periodic boundary conditions. By tuning the hopping parameters and staggered flux, the method successfully distinguished trivial from nontrivial phases. Notably, when long-range hopping was introduced, the model captured phases with higher Chern numbers. This capability is particularly valuable for experimental settings where momentum-space measurements are difficult or where open boundaries and external potential traps prevent direct Chern number calculation.

However, the approach has practical limitations that deserve attention. First, the engineered disk geometry with sixfold rotational symmetry is not general. Applying the same method to arbitrary crystal structures would require redesigning the input representation for each new system. Second, the chosen features are physically motivated but still require prior knowledge of what to look for. A truly general method would learn which real-space features are relevant without human guidance. Despite these limitations, the work establishes an important principle. Topological information is encoded not only in momentum-space Hamiltonians but also in real-space observables, and neural networks can learn to decode it. This principle remains underexplored in later topological materials discovery efforts, which have largely returned to momentum-space or symmetry-based inputs.

3.2 Physics-Based Input

To effectively integrate topological invariants into machine learning descriptors, it is essential to construct a formal mapping

$$f : T \to F,$$

where T denotes the space of topological invariants and F denotes the feature space used by machine learning models. From the perspective of information retention, such mappings can be broadly categorized into the following three classes:

(1) Global Invariant Mapping. The entire topological distribution is compressed into one or a few global numerical quantities, such as the Chern number or Z_2 invariant. This approach is particularly suitable for topological phase classification tasks. While it incurs the greatest information loss, it offers the lowest computational cost.

(2) Statistical Distribution Mapping. The overall shape of the distribution is characterized by a set of statistical descriptors, while explicit spatial information is discarded. This representation is well suited for predicting continuous material properties that correlate with global distribution characteristics.

(3) Full-Space Sampling Mapping. The topological quantities are discretely sampled on a k-space grid or along high-symmetry paths, thereby preserving nearly all information contained in the original distribution. This representation is particularly advantageous for end-to-end deep learning frameworks, although it typically involves high-dimensional inputs and requires substantially larger datasets for effective training.

For example, to formally integrate the Berry curvature $\Omega(k)$ into the machine learning pipeline, we define the following three types of mappings $f : \Omega \to F$:

(1) Topological invariant mapping (f_{inv}). The domain is $\Omega(k)$ and the codomain is $\mathbb{Z}$. The mapping rule is

$$f_{\text{inv}}(\Omega) = \frac{1}{2\pi} \int_{\text{BZ}} \Omega(k)\, \mathrm{d}^2 k,$$

i.e., the Chern number. This mapping produces a single discrete feature and is suitable for topological classification tasks.

(2) Moment feature mapping (f_{moment}). The domain is discretely sampled $\{\Omega_i\}$ and the codomain is $\mathbb{R}^3$. The mapping rule

$$f_{\text{moment}} = [\langle\Omega\rangle, \langle\Omega^2\rangle, \langle|\nabla\Omega|\rangle],$$

is representing the mean, second moment, and average gradient, respectively. This mapping produces a continuous feature vector and is suitable for predicting continuous properties.

(3) Grid sampling mapping (f_{grid}). The domain is $\Omega(k)$ and the codomain is $\mathbb{R}^{N_x \times N_y}$. A uniform $N_x \times N_y$ k-point grid is fixed, where N_x and N_y are the numbers of sampling points along the two reciprocal lattice directions, and

$$(f_{\text{grid}})_{ij} = \Omega(k_{ij}).$$

This mapping produces image-like features and is suitable for use as input to convolutional neural networks.

With the development of symmetry based theories such as TQC, researchers have begun using physical quantities from DFT calculations as input features for predicting topology in real materials. Sun et al. [44] extended machine learning prediction of topological invariants to more complex model systems. Using local Hamiltonian inputs without manual feature engineering, their deep neural networks predicted winding numbers and Chern numbers with high accuracy, even generalizing to unseen invariant values. Hidden layer analysis revealed that the networks learned the local winding angles and Berry curvature [45,46], confirming that they had internalized the mathematical formulas of the invariants themselves.

The key insight is that neural networks can capture global topological features from purely local inputs. This is not obvious because topological invariants are defined as integrals over the entire Brillouin zone. However, two caveats apply. First, the study used model Hamiltonians with known analytical forms; real DFT Hamiltonians may be noisier. Second, the network required Hamiltonian inputs, which are not experimentally accessible. Despite these caveats, the work provided strong evidence that deep learning can learn complex mathematical invariants and remains an important reference for theory driven machine learning in this field.

Half-Heusler compounds are promising for topological insulator discovery due to their tunable electronic structures, but screening them is computationally expensive. Liu et al. [47] developed a compressed sensing descriptor that uses only atomic numbers, electronegativities, and valence electron counts to rapidly screen potential topological insulators in this family. The descriptor successfully predicted numerous candidates, including compounds with fractional stoichiometries. The practical value is clear. A simple descriptor requiring no DFT calculations can prioritize candidates for experimental validation. However, there are still three limitations that warrant attention. The descriptor was only tested on half-Heusler compounds, so its transferability to other families is unknown. It identifies correlations but provides no causal mechanism. It also predicts only binary topology (topological or not) without specifying the invariant type.

Despite these limitations, the work demonstrates that elemental properties alone carry meaningful topological signal, consistent with the Topogivity results. A promising direction would be to combine such descriptors with uncertainty quantification, so users know when predictions are reliable and when DFT validation is needed.

3.3 Structure and Composition-Based Input

CGCNN [48] pioneered the use of graph neural networks [49] for crystal property prediction. It represents crystals as graphs of atoms and bonds, uses gated convolution to learn local interactions, and achieves a mean absolute error (MAE) of 0.039 eV/atom for the formation energy on the materials project database [50,51]. Its key advantage over many later models is interpretability. By extracting contributions of local chemical environments, CGCNN can identify empirical rules, such as which elements stabilize perovskite structures. The work initiated the research direction that led to MEGNet and ALIGNN. However, two limitations are worth noting. First, CGCNN requires fully relaxed crystal structures, limiting its use for hypothetical materials. Second, it predicts ground state properties like formation energy, not topological invariants. For topological materials discovery, CGCNN is therefore a pre screening tool, not a direct classifier. Whether stability trends (like those identified for perovskites) correlate with topological nontriviality remains unknown and requires separate analysis.

Chen et al. [52] proposed MEGNet with two innovations beyond CGCNN. Global state inputs (temperature, pressure, entropy) unify internal energy, enthalpy, and Gibbs free energy into a single model. Transfer learning enables pre trained embeddings to be reused for data scarce tasks like band gap prediction. On the Materials Project dataset, MEGNet achieved a formation energy MAE of 0.028 eV/atom, outperforming CGCNN's 0.039 eV/atom. The reported accuracy is impressive. However, two considerations arise when evaluating MEGNet for topological materials discovery. First, the improvements are demonstrated on ground state properties (formation energy, band gap), not on topological invariants. Whether MEGNet's architectural advances translate to better topology prediction is unknown. Second, the transfer learning advantage depends on having a large source dataset. For topological materials, where labeled data is scarce, this precondition may not hold. The real value of MEGNet for topological materials research may lie not in direct prediction but in providing accurate band gaps and formation energies as features for downstream topology classifiers. In this sense, MEGNet is a powerful tool for pre screening, not a direct topology solver.

ALIGNN [53] was designed to capture bond angle information, a key limitation of CGCNN and MEGNet which only use atomic distances. By transforming the original graph into a line graph, ALIGNN encodes angular relationships and can distinguish tetrahedral from octahedral configurations. The performance gains are substantial. ALIGNN reduced formation energy MAE by 21.4% and band gap MAE by 33.9% compared to MEGNet. Ablation studies confirmed that angular information alone contributes a 29.14% error reduction.

The key insight is that bond angles are essential for accurate property prediction. However, three caveats apply for topological materials discovery. ALIGNN requires relaxed crystal structures, predicts ground state properties rather than topological invariants, and its generalization to out of distribution structures has not been tested. Nevertheless, ALIGNN represents the current state of the art for graph based property prediction. The clear message is that angular information is not a minor refinement but a critical component. Future work should explore whether higher order structural information yields further improvements.

Claussen et al. [26] used gradient boosting trees to predict topological properties from lightweight features. The model required only chemical composition and crystal symmetry, not relaxed structures. It achieved 89.7% accuracy with F_1 scores of 94.0% for trivial insulators, 70% for TIs, and 92.0% for TSMs. The work revealed two important insights. First, topological properties are primarily determined by coarse

grained composition and symmetry, not precise atomic positions. Second, the low TI F_1 score (70%) reflects class imbalance and the intrinsic rarity of topological insulators in known databases.

The practical value is clear. The model runs much faster than DFT, and an online tool makes it accessible to experimentalists. Despite these advantages, two limitations remain. The model misses nearly one third of TIs, and its predictions inherit noise from DFT labels near phase boundaries. For large scale screening, a fast GBT model followed by targeted DFT validation may be more practical than applying expensive graph models to every candidate. This finding aligns with the Topogivity results, reinforcing the message that composition alone carries strong topological signals. Building on these developments, recent work has further extended structure- and composition-based learning to equivariant interatomic potentials such as NequIP [54] and MACE [55] as well as large-scale generative foundation models such as DiffCSP [56] and MatterGen [57], broadening the representational capacity available for downstream topological-materials screening. Table 2 provides an overview of the stages involved in applying machine learning to topological materials, along with the types of input features used at each stage.

Table 2: Summary of the three stages in machine learning approaches for topological materials, corresponding to the subsections of Section 3.

Stage	Typical Inputs/Descriptors	Representative Goals & References	Crystal Structure Required	DFT Labels Required
Theoretical Model Validation	Spin configurations, Hamiltonian parameters, local Hamiltonians, and related quantities generated from controlled model systems.	Validate whether machine learning can identify phases, locate phase boundaries, and learn topological invariants in theoretical model systems [40–44].	No	No
Physics-Based Input	Physically motivated descriptors such as atomic numbers, electronegativities, valence electron counts, XANES spectral features, and Hubbard-U-augmented representations.	Construct interpretable and efficient predictors for screening and classifying topological materials using physics-guided features [24,38,39,47].	No (for composition)/Yes (for XANES)	Yes (DFT-calculated spectra/properties)
Structure and Composition-Based Input	Crystal graphs with atomic, bond, and angular features; global state variables; and composition-plus-symmetry descriptors such as elemental statistics, space group, and electron count per unit cell.	Learn structure-property and composition-topology relationships directly from crystal and chemical representations to support large-scale materials discovery [25,26,48,52,53].	Yes (for graph-based)/No (for composition-only)	Yes (DFT-calculated formation energy, band gap, etc.)

The three graph-based models described above, CGCNN, MEGNet, and ALIGNN, have all demonstrated strong performance on their respective benchmark tests. However, when attempting to compare these models, a critical methodological issue arises: there is currently no unified benchmarking framework that evaluates them under identical conditions. As shown in Table 3, the performance reported in the original publications is based on different datasets, different training and test splits, and different evaluation metrics MAE of formation energy per atom, but using different versions of the Materials Project database and different preprocessing pipelines. These discrepancies make it impossible to determine whether the reported performance improvements arise from genuine advances in model architecture or merely from differences in evaluation protocols.

Table 3: Incomparable evaluation protocols in graph-based materials property prediction. The data presented in this table are compiled from the literature: CGCNN [48], MEGNet [52], and ALIGNN [53].

Model	Dataset	Data Splitting Strategy	Key Metric
CGCNN	Materials Project (46,744 crystals)	Random 80/10/10 split	Formation energy MAE: 0.039 eV/atom
MEGNet	Materials Project (69,239 crystals)	Random 60/20/20 split	Formation energy MAE: 0.028 eV/atom
ALIGNN	Materials Project (69,239 crystals)	Random 60/20/20 split	Formation energy MAE: 0.022 eV/atom

At first glance, these numbers appear to suggest a clear performance hierarchy, with ALIGNN outperforming MEGNet and MEGNet outperforming CGCNN. However, a closer examination reveals that CGCNN was evaluated using an earlier version of the Materials Project database (the 2018 release version) and a different data splitting strategy, whereas MEGNet and ALIGNN were trained and tested on a later and larger dataset with different splitting protocols. In addition, the random seeds used for data splitting are typically not reported, making it impossible to reproduce the exact training conditions. The reported 21.4% reduction in formation energy MAE of ALIGNN relative to MEGNet may indeed reflect the benefit of incorporating bond angle information. However, in the absence of controlled comparisons under identical conditions, including the same dataset version, identical data splits, fixed random seeds, and consistent evaluation protocols, the quantitative magnitude of this improvement cannot be reliably assessed. Such a lack of standardization severely hinders the community's ability to assess genuine progress.

We believe that the field of machine learning for topological materials would greatly benefit from community accepted benchmarking frameworks. Drawing on best practices from materials informatics and the broader machine learning community, we propose the following recommendations.

- Standardized datasets: All models should be evaluated on one or more publicly available, version controlled datasets. Candidate datasets include the Topological Materials Database, symmetry indicator annotations from the Materials Project, or curated subsets thereof. The database version and all preprocessing procedures, such as formation energy filtering, duplicate removal, and structural relaxation status, should be explicitly documented.
- Fixed data splits: All models should use identical training, validation, and test splits to ensure fair comparison. Given the risk of data leakage arising from structurally similar materials appearing in both training and test sets, composition based or structure based splits are preferable to random splits. The random seed and splitting protocol should be reported in detail.
- Standardized metrics: Classification tasks should report common evaluation metrics, including Accuracy, Precision, Recall, and F_1 score.
- Open code and data splits: To facilitate reproducibility and fair comparison, authors are strongly encouraged to release source code, trained models, and the exact data partitions under permissive open source licenses.

Adopting these recommendations would enable the field to move beyond the current situation of numerous incomparable performance reports toward a cumulative and quantitatively rigorous body of knowledge. Several recent efforts have already moved in this direction. For example, MatBench [58] provides standardized benchmarks for general materials property prediction, while the topological materials database

offers a curated database specifically focused on topological materials. We encourage future studies in machine learning for topological materials to adopt these or similar benchmark frameworks whenever possible. Until a unified evaluation framework becomes standard practice, performance comparisons across different models should be interpreted with caution.

4 Applications of Machine Learning in Topological Materials

This section primarily presents the applications of machine learning in the areas of inverse design and generation of topological materials, topological superconductors, and the mapping of topological phase diagrams, as summarized in Fig. 1.

4.1 Inverse Design and Generation of Topological Materials

Inverse design is a data-driven strategy that combines desired material property data, domain knowledge, and artificial intelligence to discover new materials with specific functionalities. This approach is rapidly advancing in the field of novel materials research. Hong et al. [27] proposed the CTMT framework for the inverse design of topological insulators and semimetals. The framework integrates four components. The CDVAE (crystal diffusion variational autoencoder) generates new crystal structure candidates. Topogivity rules perform an initial screen based on chemical composition. The M3GNet interatomic potential [59] provides rapid energy, force, and stress evaluations for dynamical stability. Finally, TQC provides precise topological classification. Starting from 10,000 generated candidates, CTMT applied successive filters for novelty, validity, topogivity, thermodynamic stability, and dynamical stability. This process confirmed 20 new topological materials, including 16 TSMs and 4 TIs. Notably, the method discovered four chiral Kramers Weyl semimetals [60–63], a class of low symmetry materials that traditional symmetry based methods struggle to identify. The reported success rate was 62.5%, compared to less than 30% for conventional approaches.

The significance of CTMT is twofold. First, it demonstrates that generative models can produce genuinely new topological materials not present in existing databases. This addresses a key limitation of screening based approaches, which can only find what is already known. Second, the discovery of low symmetry chiral materials highlights an advantage of data driven generation over symmetry based diagnosis. The latter works well for high symmetry crystals but fails when symmetry is low.

The CTMT strategy can be interpreted as a closed discovery loop integrating generation, screening, validation, diagnosis, and database feedback, as shown in Fig. 3. Table 4 provides a summary of the machine learning methods discussed above for discovering topological materials.

Although the CTMT framework represents an important advance in the inverse design of topological materials, its validation protocol deserves critical examination. Before performing full topological classification, the CTMT framework employs the M3GNet interatomic potential as a preliminary filter for dynamical stability. While this screening step is computationally efficient, reliance on an approximate potential introduces the possibility of false positives. In other words, some materials that pass the M3GNet stability screening may later be found to be topologically trivial.

The original CTMT study provides sufficient data to quantify this risk. Among the 57 candidate materials that passed the DFT thermodynamic stability check, M3GNet phonon spectrum calculations identified 32 as dynamically stable, corresponding to a success rate of 56%. Subsequent TQC based topological classification confirmed that only 20 of these 32 candidates were topologically nontrivial. Therefore, among the candidates that passed the M3GNet dynamical stability screening, 12 were false positives with respect to topological nontriviality. This corresponds to a false positive rate of 37.5%.

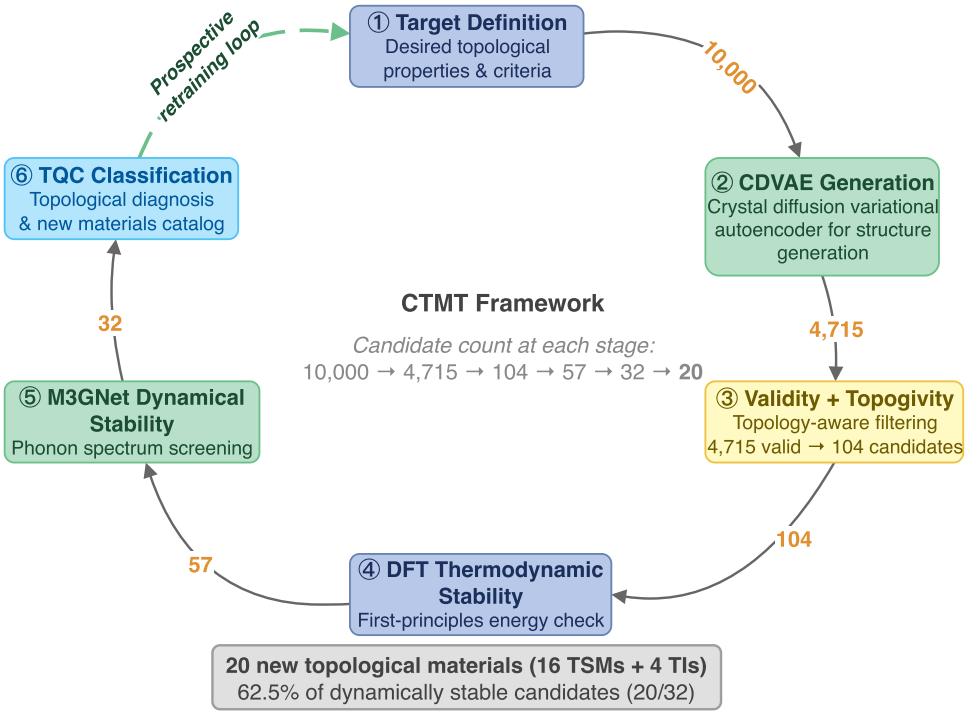

Figure 3: The closed loop CTMT framework for inverse design of topological materials. The workflow consists of six sequential stages: (1) target definition, (2) crystal generation using CDVAE with 10,000 candidate materials, (3) chemical rule screening using Topogivity with 104 candidate materials, (4) DFT validation of thermodynamic stability with 57 candidate materials, (5) dynamical stability prescreening using M3GNet with 32 candidate materials, and (6) topological classification using TQC, resulting in 20 confirmed topological materials including 4 topological insulators and 16 topological semimetals. The number of candidates at each stage, shown in parentheses, illustrates how the framework efficiently reduces the search space from 10,000 generated structures to 20 validated topological materials, achieving a discovery success rate of 62.5% among dynamically stable candidates. Adapted from [27].

Table 4: Representative machine learning methods used for topological materials discovery.

Method	Core Principle	Crystal Structure Required	DFT Labels Required
SISSO/compressed sensing	Constructs low-dimensional, interpretable descriptors from elemental properties such as atomic number, electronegativity, and valence electron count to rapidly screen candidate topological insulators without explicitly computing full band topology [24,47].	No	No
Neural Network (NN)	Learns topological classification directly from computed or experimental XANES spectra together with atomic-type information, enabling spectroscopy-based identification of topological materials [38].	No (spectra input)	Yes (DFT-computed spectra for training)

(Continued)

Table 4 (continued)

Method	Core Principle	Crystal Structure Required	DFT Labels Required
Topogivity	Assigns machine-learned topological tendencies to chemical elements and uses composition-weighted elemental scores as a simple heuristic rule for diagnosing whether a material is likely to be topologically nontrivial [25].	No	No
Convolutional Neural Network (CNN)	Extends composition-based topologicality rules by learning nonlinear weighting functions from element-fraction and Hubbard-*U* representations, improving screening of magnetic and two-dimensional topological insulators [39].	No	Yes (DFT + Hubbard U for training)
Gradient Boosting Trees	Predicts trivial insulators, topological insulators, and topological semimetals from lightweight composition and crystal-symmetry features, including elemental statistics, valence-electron counts, space group, and electron count per unit cell [26].	No (symmetry features from database)	No (labels from TQC database)
Crystal Graph Neural Networks	Represent crystal structures as graphs with atoms, bonds, global state variables, or bond-angle information, enabling data-driven prediction of materials properties that can support large-scale screening workflows [48,52,53].	Yes	Yes (DFT-calculated properties for training)
Deep Generative Models	Generate new crystal candidates and couple generation with topological screening, stability evaluation, and TQC diagnosis, as in CTMT workflows for discovering new topological insulators and semimetals [27].	Yes (generates structures)	Yes (DFT for stability verification)

This quantitative estimate indicates that the use of approximate interatomic potentials can indeed introduce a nonnegligible fraction of false positives into the validation workflow. Nevertheless, the CTMT framework remains highly effective. The M3GNet screening stage reduces the number of candidate materials from 57 to 32, thereby substantially decreasing the computational burden of subsequent TQC calculations. At the same time, the discovery of 20 new topological materials represents a significant achievement.

For future inverse design studies, we recommend the following practices to further strengthen the rigor of the validation process.

- Explicitly report the false positive rate and false negative rate of machine learning interatomic potential screening relative to DFT validation.
- Incorporate uncertainty quantification into machine learning potential predictions. Candidates associated with high uncertainty can be directly advanced to the DFT validation stage.
- Calibrate machine learning interatomic potentials using a benchmark set of known stable materials before applying them to screen newly generated candidates.

It is important to note that beyond the capability of generating structures and the topological non-triviality of candidate materials, several equally critical issues must be considered when evaluating and developing generative models for topological material design. These include first structural stability, where generated crystal structures should exhibit no imaginary phonon modes indicating dynamical stability and should remain intact under finite temperature conditions indicating thermodynamic stability. Second synthesizability, referring to whether the predicted stability can be realized under actual experimental conditions. Third diversity and novelty, namely whether the generative model can produce structures that are genuinely different from those in the training set rather than minor variations. Fourth computational efficiency, measured in comparison with the cost of brute force high throughput screening. Generative models can bypass the expensive energy evaluations required in traditional crystal structure prediction methods [64]. Conditional generation frameworks have been shown to significantly improve the efficiency of target material discovery [65]. Fifth false positive rate, defined as the proportion of materials identified as stable or topological by fast screening methods but later invalidated by more rigorous DFT and TQC calculations.

4.2 Topological Superconductors

Tsai et al. [66] compared three quantum information quantities as inputs for detecting topological phase transitions in the Kitaev chain [67]. The three inputs were the Majorana correlation matrix, the entanglement spectrum, and the entanglement feature vector. All could identify the phase transition point. However, the entanglement spectrum failed to distinguish different $U(1)$ gauge phases because it compressed the information too much. The other two inputs succeeded. Gradient-weighted class activation mapping (Grad-CAM) visualization showed that the network learned to focus on the region corresponding to edge state decay.

The key lesson is that input representation choices profoundly affect what a network can learn. The failure of the entanglement spectrum was not due to a weak network but because the necessary information had already been discarded. This principle is often overlooked in materials informatics, where most effort goes into architecture rather than input construction. However, the Kitaev chain is a one dimensional toy model. Whether the same conclusions hold for real topological superconductors remains unknown.

Chung et al. [68] extended the entanglement based method to two dimensional chiral p-wave superconductors. They compared the Majorana correlation matrix, the single particle entanglement spectrum, and the entanglement feature vector. Several findings emerged. Phase information is essential. The entanglement feature vector worked only when the phase angle was included. The magnitude alone was insufficient. Geometric structure matters. The tensor format of the Majorana correlation matrix outperformed the matrix format because it preserved spatial relationships. Rescaling inputs with dimensionless coordinates improved accuracy when the superconducting gap was large.

The key contribution is systematic. Phase information matters. Geometric structure matters. Input format matters. These principles likely generalize beyond the specific model studied. Several challenges remain, among which two are particularly noteworthy. The study used clean model Hamiltonians, not realistic materials with impurities and disorder. More importantly, the quantum information quantities used as inputs are not directly measurable in experiments. Future work should focus on bridging this gap, for example by training on simulated data with realistic noise and testing on experimental measurements.

The common machine learning pipeline for topological superconductors, from model systems and input construction to predicted topological outputs, is summarized in Fig. 4.

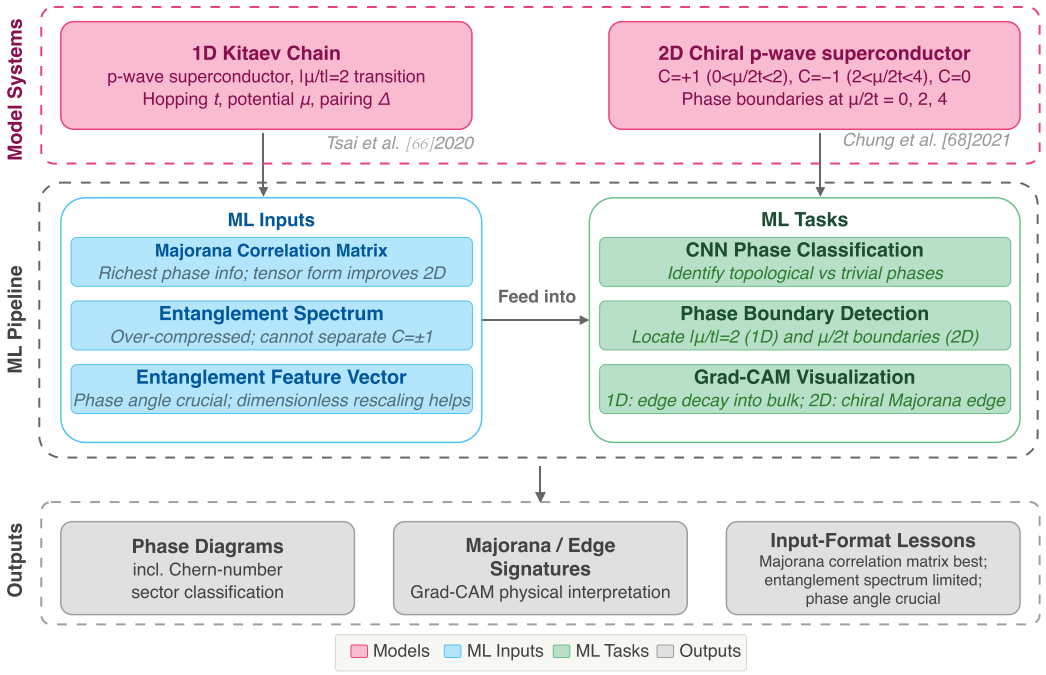

Figure 4: Machine learning pipeline for topological superconductors. Model systems (Tier 1) provide quantum-information-inspired inputs (Tier 2, left) for deep learning tasks (Tier 2, right), yielding phase diagrams, edge-mode signatures, and input-format comparisons (Tier 3) [66,68].

4.3 Mapping of Topological Phase Diagrams

Zhang and Kim [69] proposed quantum loop topography to address a key challenge. Topological phases lack local order parameters, so standard machine learning inputs cannot detect them. Their method encodes nonlocal topological information into images by constructing loop products of correlation functions. A simple neural network achieved over 99.9% accuracy on a Chern insulator model. A control experiment confirmed that raw particle configurations did not work. The method showed impressive generalization. A network trained on non interacting Chern insulators could identify fractional quantum Chern insulators without retraining. A network trained on a square lattice could identify phase transitions in a honeycomb lattice. The approach does not rely on translational symmetry or diagonalization and can handle strongly correlated and disordered systems.

The importance of this contribution is considerable. This work provides a general feature engineering strategy that works where conventional order parameters do not exist. The cross model generalization is particularly impressive. Nevertheless, two concerns arise. The method requires correlation functions from Monte Carlo samples. For experimental data with finite precision, the construction may be less robust. The method was also demonstrated only on model systems, not on real topological materials. Despite these limitations, the work established an important principle. Transform nonlocal topological information into images. Then standard neural networks can decode it. This principle has influenced later topological deep learning, though subsequent work has focused more on persistent homology than on loop based constructions.

Rodriguez Nieva and Scheurer [70] proposed an unsupervised method for identifying topological order from raw spin configurations without labeled data. The method constructs a similarity matrix, performs eigenvalue decomposition, uses eigenvalue degeneracy to count topological sectors, and applies k means clustering to assign samples. The method successfully identified the Berezinskii Kosterlitz Thouless transition

in the XY model and the topological sectors in the Ising gauge theory. The critical temperatures matched theoretical predictions. The significance is threefold. The method requires no labeled data, which is crucial for exploring unknown phases. It does not assume knowledge of the order parameter or invariant in advance. The diffusion map automatically reveals the number of sectors without user input.

Despite its strengths, two weaknesses are evident. The similarity matrix scales quadratically with sample size, making large datasets expensive. More importantly, the method has only been tested on clean Monte Carlo data. Whether it works on noisy experimental data remains unknown. The method also detects transitions and counts sectors but does not provide the topological invariant itself, such as the Chern number. Despite these limitations, this work provides a paradigm for automated phase diagram mapping. The contrast with supervised methods is stark. Supervised methods require knowing what to look for. Unsupervised methods discover what is there. For discovering entirely new topological phases, unsupervised approaches may be the only viable path forward.

The overall procedure for identifying topological phase diagrams from physical configurations and learned low-dimensional structure is illustrated in Fig. 5.

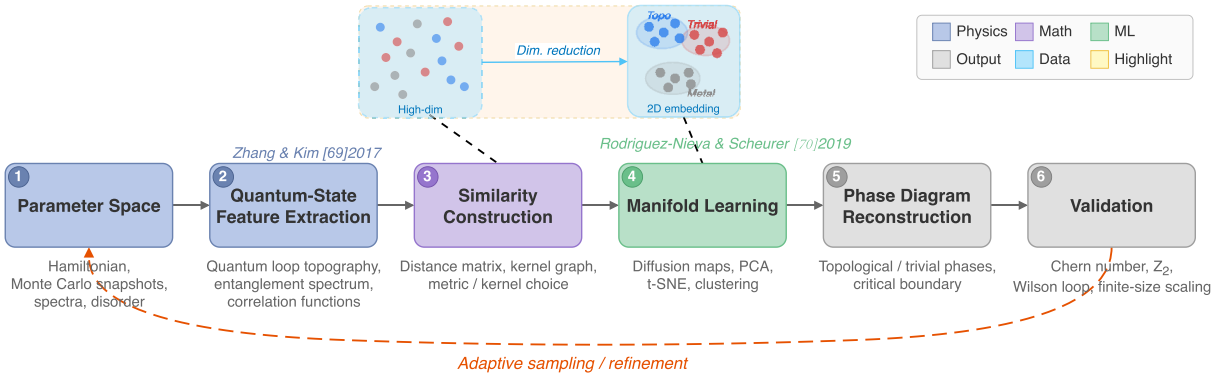

Figure 5: Unsupervised machine-learning workflow for mapping topological phase diagrams. Physical configurations are transformed into quantum-state features, projected onto low-dimensional manifolds via diffusion maps or related methods, and clustered to reconstruct phase boundaries and topological sectors [69,70].

5 Electronic Structure Features of Topological Materials and Topological Deep Learning

The electronic structure of a topological material is described by the Bloch wave functions in the Brillouin zone [71]. The Brillouin zone is the Wigner-Seitz cell in reciprocal space. Due to the periodic structure of reciprocal space, its topology is mathematically equivalent to a d-dimensional torus T^d [4]. This torus structure forms the basis for calculating topological invariants. On the Brillouin zone T^d, each wave vector k corresponds to multiple energy bands $E_n(k)$, and these band functions together form a multidimensional electronic structure defined on T^d. Since the entire electronic structure depends on the continuous variation of k and the coupling of multiple bands, it can be regarded as a high-dimensional functional object. Therefore, in some sense, the electronic structure can be treated as high-dimensional data in machine learning. It should be noted that the topological invariants of the energy bands are properties of the global structure rather than local features, for example, the Chern number is the integral of the Berry curvature over the entire Brillouin zone normalized by a factor of $1/(2\pi)$ [46,72].

The Bloch Hamiltonian $H(k)$ of a crystal satisfies the covariance relation [73] under the space group symmetry operation g as

$$U_g H(k) U_g^{-1} = H(gk).$$

Therefore, the Bloch wave functions and the energy band structure must obey the corresponding symmetry constraints throughout the Brillouin zone [74]. These symmetries determine the degeneracies of the energy bands, the irreducible representations, and the connectivity of the bands, which in turn further constrain the topological properties of the electronic structure of the material. The connection between Brillouin-zone topology, electronic-structure features, persistent homology, and machine learning prediction is schematically shown in Fig. 6.

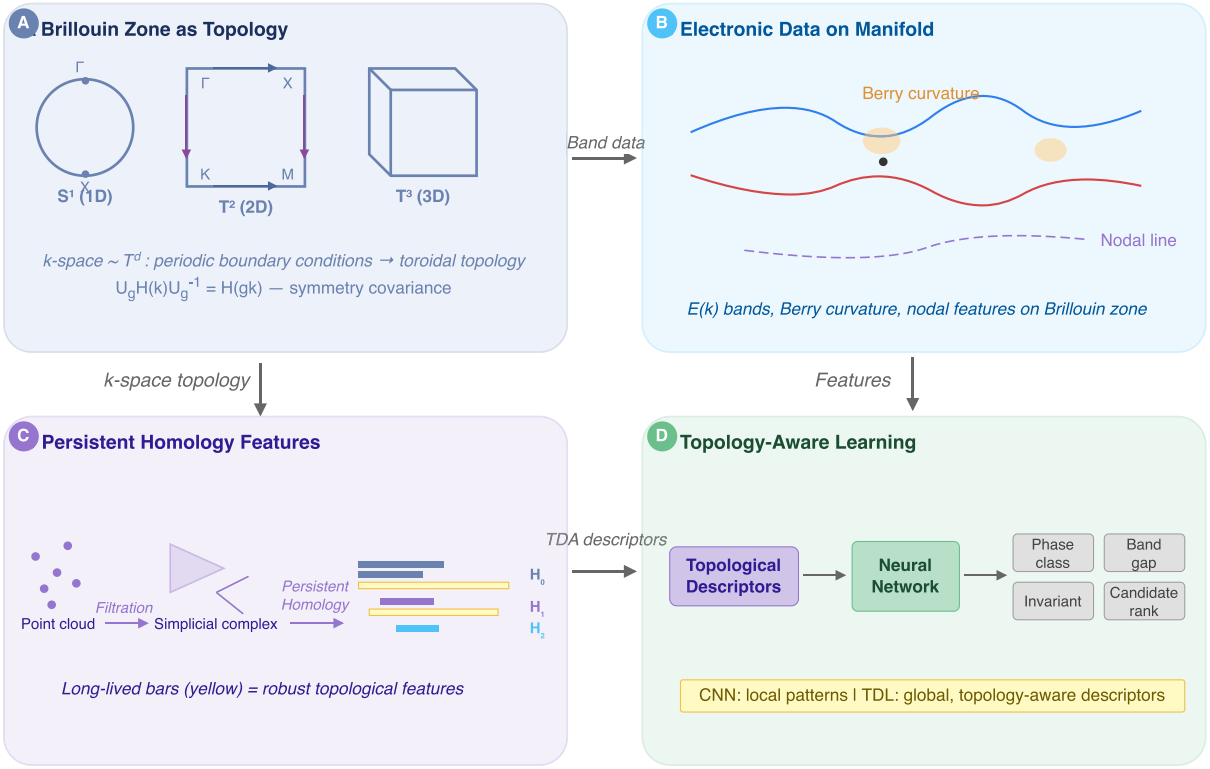

Figure 6: From electronic structure to topological deep learning: a conceptual workflow. The figure illustrates how topological information can be extracted from electronic structure data and incorporated into machine learning models. (**A**) The Brillouin zone possesses the topology of a d dimensional torus T^d, providing the geometric foundation for topological invariants. (**B**) Physical quantities defined on this manifold, including band energies $E_n(k)$, Berry curvature $\Omega(k)$, and nodal line structures, serve as the raw data for topological analysis. (**C**) Persistent homology extracts multiscale topological features from these data and produces compact descriptors in the form of persistence barcodes. (**D**) These descriptors are subsequently fed into topology aware machine learning models for tasks such as topological phase classification, property prediction, and candidate material ranking. The figure highlights how each stage reduces data dimensionality while preserving or extracting topologically relevant information.

Accordingly, modern machine-learning methods for electronic-structure data should address the topological challenges and capabilities summarized in Table 5.

Traditional machine learning methods typically assume that data are distributed in a linear vector space and can be described by feature vectors of fixed dimension. Such methods primarily capture local statistical properties but often lack the ability to represent higher-order structures, global correlations, or topological features inherent in the data. TDL is a rapidly expanding research area that uses topological features to analyze, interpret, and design deep learning models [75]. In its early usage, the term TDL referred to the integration of features generated by persistent homology [76,77] into the input pipeline of deep

neural networks [78]. Persistent homology is a tool for characterizing the evolution of homology groups of filtered complexes, and barcodes provide a common representation. For clarity, throughout this review we distinguish three layers: (i) topological data analysis (TDA) descriptors that summarize the global shape of data, (ii) persistent-homology features combined with classical or deep learning pipelines, and (iii) intrinsic TDL, which performs message passing directly on simplicial, cell, or hypergraph domains [79]. Unless otherwise noted, we use "TDL" in its broader contemporary sense that encompasses all three layers.

Table 5: Topological characteristics of electronic-structure data and corresponding TDL capabilities for topological-materials research.

Topological Challenge	TDL Capability	Practical Relevance
High-dimensional electronic-structure fields	Persistent homology and topological descriptors extract connected components, loops, cavities, and multiscale geometric signatures from scalar fields or point-cloud representations.	Provides compact representations of charge density, electron localization, and atomistic environments for learning structure–property relationships in materials systems [77].
Multiscale and hierarchical topology	Filtration-based learning captures features that persist across spatial, energetic, or density thresholds, distinguishing robust topology from local fluctuations.	Enables models to identify physically meaningful motifs across length scales, improving robustness for crystalline, molecular, and electronic-structure data [80–83].
Symmetry, periodicity, and invariance requirements	Topology-aware and graph-based learning models can encode invariant or equivariant representations while respecting crystal symmetry, periodic boundary conditions, and global connectivity.	Supports learning under translation, rotation, permutation, and lattice-periodic constraints that are central to topological-materials discovery [9,23,52,59,74].
Noisy, sparse, or discretized computational data	Topological summaries emphasize stable global structures and reduce sensitivity to mesh resolution, sampling artifacts, and numerical perturbations.	Improves reliability when learning from heterogeneous simulation outputs, finite-resolution electronic-structure calculations, or incomplete materials datasets [84–86].
Interpretable links between topology and materials behavior	TDL combines predictive models with topological features that can be inspected, ranked, or associated with physical mechanisms.	Facilitates interpretable screening and classification of topological phases, electronic states, and candidate materials by connecting learned predictions to topological signatures [82].

It is necessary to clarify the precise objects that these topological tools actually operate on. In the following, k denotes the wave vector in the Brillouin zone, and the Brillouin zone is discretized into a finite

set of k-points. The input of persistent homology is a filtration constructed from electronic structure data. Common constructions include first a point cloud of k-points sampled from the Brillouin zone, where distances between k-points define connectivity, and second a scalar field defined on the Brillouin zone, such as Berry curvature magnitude or band energy differences, from which a nested sequence of topological spaces is generated using sublevel set or superlevel set filtrations. The output of persistent homology is a persistence barcode or persistence diagram that encodes multiscale topological features. Intrinsic topological deep learning takes as input simplicial complexes or cell complexes constructed from the Brillouin zone. For example, the Brillouin zone torus can be discretized into a cell complex together with higher dimensional cell structures. Physical features such as band energies and Berry curvature are assigned to cells of corresponding dimensions. Message passing then occurs not only between vertices but also along edges, across faces, and through higher order simplices, enabling the model to directly learn topological invariants.

Many real world datasets, such as molecular structures, material lattices, social networks, and spatiotemporal dynamical systems, are embedded in complex topological spaces rather than simple Euclidean spaces. Higher order relations can capture long range or seemingly unrelated connections in a system, which supports the construction of more effective and more robust message passing mechanisms. Although graph based equivariant neural networks can already handle some forms of multivariate interactions, real data often contain richer types of higher order interactions. Topological domains provide a better representation for such higher order interaction data. TDL introduces higher order structures such as simplicial complexes, cell complexes, and Reeb graphs to build deep models that can capture diverse forms of multivariate interactions. Furthermore, the topological perspective can describe the intrinsic regularity of manifolds in a more faithful way. For example, different triangulations of a triangulable manifold do not change its topological properties. Therefore, TDL is naturally suitable for expressing such regularities. In addition, TDL emphasizes the treatment of symmetries in data. TDL can represent simple symmetries such as permutation groups or Euclidean groups, and it can also naturally express broader forms of topological equivariance. It retains more embedded information while incorporating homeomorphism groups.

Topology is often described as a tool that encodes the overall shape of data. Compared with geometric features that are local and often more rigid, topological features are better suited for capturing multiscale, global, and intrinsic properties of data. TDL, as an emerging field that integrates topological methods into machine learning and deep learning models, can extract more valuable information about topological materials from complex and large scale datasets.

Within the broader scope of TDL, it is important to distinguish between two methodological paradigms, namely TDA based approaches and intrinsically topology aware neural network architectures. These paradigms differ substantially in both their level of implementation and their practical applicability.

In the former paradigm, TDA tools are employed as a preprocessing step to compute topological descriptors from raw data. These descriptors are then transformed into feature vectors and used as inputs for conventional machine learning models such as random forests, support vector machines, or standard neural networks. The neural network itself does not operate directly on a topological domain. Instead, topological information is incorporated only during the feature engineering stage. This approach is conceptually simpler, easier to implement, and applicable to a wide variety of data types. However, it discards much of the rich structural information encoded in the topological descriptors and cannot learn task specific topological representations in an end to end manner. Representative examples in the topological materials literature include phase classification using persistent homology features and quantum loop topology methods.

In the latter paradigm, the neural network is explicitly designed to operate on topological domains such as simplicial complexes, cell complexes, or hypergraphs. Message passing occurs not only between nodes but also along edges, triangles, and higher dimensional simplices, enabling the model to learn

topological invariants directly from data without explicit feature engineering. Representative architectures include message passing simplicial networks, cell complex neural networks, and differentiable persistent homology layers. In principle, these models are more expressive and can capture higher order interactions that remain inaccessible to conventional TDA pipelines. However, they are also more complex to implement, computationally more demanding, and require data to be represented as simplicial or cell complexes. Such representations are not always straightforward to construct for noisy or incomplete datasets.

When choosing between these two paradigms, researchers should consider the specific application scenario and the available computational resources (see Table 6). For rapid screening tasks or applications where interpretability is a primary concern, TDA based feature extraction remains a practical and effective choice. For applications that require the highest possible predictive performance and have access to sufficient computational resources together with high quality topological data representations, intrinsically topology aware architectures may justify their additional complexity.

Table 6: Comparison of two topological deep learning paradigms in topological materials research.

Aspect	TDA Based Feature Extraction	Intrinsically Topology Aware Architectures
Core Idea	Extract topological features as inputs to conventional machine learning models	Neural networks operate directly on topological domains
Implementation Complexity	Low; relies on existing TDA libraries and standard neural networks	High; requires construction of simplicial or cell complexes together with specialized neural network layers
Computational Cost	Low to moderate	High; involves message passing on higher dimensional simplices
Data Requirements	Applicable to point clouds, graphs, and images; relatively tolerant to noise	Requires simplicial or cell complex representations; less robust to noisy data
End to End Learning	No; topological preprocessing is fixed	Yes; topological representations are learned from data
Interpretability	High; topological features can be directly inspected	Moderate; learned representations are less transparent
Recommended Applications	Rapid screening, interpretability focused tasks, and noisy datasets	High accuracy prediction, clean structured datasets, and representation learning studies

6 Challenges and Outlook

In the research field of topological materials, the discovery of new TIs and TSMs has remained a leading and important direction. Progress in this area has primarily relied on a series of first principles computational methods developed within the framework of topological band theory. In particular, the introduction of symmetry indicator theory together with TQC has enabled the identification of a large number of topological materials and has motivated the construction of several large scale databases based on high throughput calculations. As the volume of available data continues to grow rapidly, researchers

face an urgent need to improve the efficiency of discovering new topological materials. To accelerate the computational tasks associated with these theoretical frameworks, machine learning has begun to be applied to improve computational speed [26], determine topological properties [38], and automatically classify structures with low symmetry [25], among other related tasks.

Foundation models constitute an important branch of machine learning. Traditional machine learning primarily focuses on predicting the properties of existing materials, whereas generative models extend this paradigm by enabling the design and creation of new materials. As a result, they have been widely adopted in inverse materials design and the development of foundation models for materials science. Generative foundation models such as MatterGen [57] and CDVAE [27] have demonstrated the capability to generate millions of novel and stable crystal structures at extremely low computational cost. For topological material discovery, these models offer two major advantages. First, generative foundation models can explore regions of chemical space that are inaccessible to symmetry-based high-throughput screening. Traditional approaches rely on existing databases and are therefore restricted to known compositions and structures. In contrast, generative models are able to propose entirely new materials that do not exist in any database. For example, the CTMT framework [27] has identified 20 new topological materials. Second, recent studies have shown that foundation models can be fine-tuned to generate materials with target properties. For instance, MatterGen can be adapted through adapter-based fine-tuning to perform conditional generation based on chemical composition, symmetry, or desired mechanical, electronic, and magnetic properties [57]. This capability is particularly valuable for topological materials discovery, since target properties such as nontrivial Chern numbers and $\mathbb{Z}_2$ invariants may potentially be incorporated as generative conditions.

Although foundation models have demonstrated tremendous potential in property prediction, structure generation, and high-throughput screening in materials science, their application to topological materials research still faces two key challenges. First, foundation models are significantly more computationally expensive than task-specific models. For example, a neural network potential tailored to a specific alloy system can be trained using only a few thousand DFT calculations, whereas foundation models typically require millions of training samples and substantial GPU resources [87]. This level of cost may be acceptable for large-scale screening, but it becomes prohibitive for small-scale studies. Second, foundation models generally do not provide reliable uncertainty estimates. An evaluation of MACE-MP-0 by Bilbrey et al. [88] shows that the model uncertainty is approximately 20 times smaller than the actual error, and that the empirical coverage of the nominal 90% confidence interval is only 11%. This limitation is particularly critical in high-throughput screening, where the model may assign low uncertainty to false positive candidates, making it difficult for users to identify erroneous predictions without additional DFT or TQC validation.

Regardless of whether advanced foundation models or more conventional machine learning methods are employed, current machine learning workflows for topological materials are confronted with at least four fundamental challenges that are closely intertwined and cannot be effectively addressed in isolation:

(i) Data scarcity and class imbalance, since confirmed topological compounds remain a small minority in available databases and magnetic or higher-order topological phases are particularly under-represented. For example, the GBT model discussed in Section 3.3 achieved an F_1 score of only 70% for topological insulators, compared with an F_1 score of 94% for trivial insulators. The authors explicitly attributed this performance gap to the limited representativeness of the available topological insulator samples.

This challenge may lead to the consequence that, due to the limited number of rare topological phase samples in the training set, the model performs poorly when predicting these rare phases and is prone to false negatives, namely failing to identify true topological phases. Possible strategies for addressing this challenge include: (1) using generative models such as CDVAE or MatterGen to

create synthetic data and increase the number of samples belonging to rare topological classes; (2) applying transfer learning by pretraining on large scale general materials databases and subsequently fine tuning on smaller topological materials datasets; and (3) adopting reweighted loss functions or anomaly detection frameworks that treat minority topological phases as rare events.

(ii) Out-of-distribution generalization, since models trained on existing databases often fail on chemistries, symmetries, or dimensionalities that are under-sampled. For instance, the CTMT framework discussed in Section 4.1 provides a quantitative example. When the M3GNet interatomic potential was applied to structures generated outside its training distribution, the false positive rate with respect to topological nontriviality reached 37.5%.

This challenge may lead to unreliable predictions when the model encounters materials outside the training distribution, resulting in both false positives, where materials are predicted to be topological but fail validation, and false negatives, where true topological materials are not identified. This in turn may waste computational and experimental validation resources or cause potentially promising candidates to be overlooked. To address this challenge, several strategies can be considered: (1) adopting uncertainty quantification methods to subject low confidence predictions to additional verification; (2) performing adversarial training using synthetic out of distribution samples; and (3) employing domain adaptation techniques to align feature distributions between the training domain and the target domain.

(iii) Interpretability, since composition- or graph-based predictors rarely expose which structural or symmetry motifs drive the predicted topology. For example, among the methods reviewed in Section 3, only CGCNN provides direct interpretability by identifying the contributions of local chemical environments to global material properties. The topogivity rule offers a simple heuristic that can be evaluated manually, although its predictive accuracy remains limited. Most other models, including MEGNet, ALIGNN, and XANES based neural networks, operate largely as black boxes.

This challenge may lead to the limitation that black box models hinder the understanding of the underlying physical mechanisms and reduce the level of trust in model predictions. Several approaches may help mitigate this challenge: (1) employing attention mechanisms and Grad-CAM visualization techniques to identify the key input features that drive model predictions; (2) using symbolic regression or SISSO to extract interpretable analytical descriptors from learned representations; and (3) adopting inherently interpretable model architectures, such as linear models based on physically meaningful descriptors.

(iv) The joint representation of symmetry, geometry, and topology, since band topology is inherently defined over the Brillouin-zone torus and is constrained by space-group symmetries that local feature vectors do not automatically respect. Addressing these challenges jointly, rather than item by item, is a prerequisite for turning screening pipelines into reliable discovery engines. For instance, this challenge becomes particularly evident when comparing the methods discussed in Sections 2 and 3. Symmetry based descriptors capture topological invariants from band representations but discard geometric details. Graph based models such as CGCNN and MEGNet encode geometric information including bond lengths but do not explicitly incorporate topological invariants. ALIGNN makes further progress by introducing bond angle information, yet constructing a unified framework that consistently integrates symmetry, geometry, and topology remains an open challenge.

One implication of this challenge is that models may be unable to fully exploit symmetry, geometric structure, and topological information simultaneously, which may result in the neglect of essential symmetry constraints or incorrect predictions of nonlocal topological invariants, thereby affecting the accuracy of topological property prediction. This challenge can be addressed through several approaches: (1) employing topological deep learning architectures that operate directly

on simplicial complexes or cell complexes representing the Brillouin zone; (2) developing hybrid methods that combine symmetry indicator calculations with graph based geometric feature representations; and (3) explicitly incorporating topological invariants as auxiliary prediction tasks in the pretraining objectives.

In general, given the interplay between machine learning and topological materials, at a macroscopic level we can address the challenges discussed above from two perspectives. On one hand, generative machine learning models can be used to design new candidate materials. The developed generative models must overcome the difficulty that traditional generative models face in capturing representations that are invariant under translations and rotations when handling crystalline materials. On the other hand, efficient material representation methods and schemes for designing networks to extract useful features can be developed to enhance the automatic prediction capability of deep learning based topological materials. TDA explores qualitative features based on geometry and topology and is a powerful candidate for meeting such demands. Many tools within TDA can extract topological features that are not accessible through other statistical, physical, or mathematical methods. For example, persistent homology can detect complex topological invariants and topological patterns in data across different scales [80]. The persistent Laplacian [81,85] is used to characterize certain non topological shape evolutions. multiscale Gauss link integral [86], multiscale Jones polynomials [89], and evolutionary Khovanov homology can be applied to one dimensional curves embedded in three dimensional space. In addition, TDA also includes many other tools, such as persistent path topology [82], persistent interaction topology [83], and persistent sheaf Laplacians, which can simplify the geometric complexity of high dimensional data and provide higher order data representations, thereby extending traditional zero order algorithms.

7 Conclusion

Topological materials exhibit a variety of fascinating physical properties that often exceed the expectations of conventional materials and open possibilities for the development of devices with entirely new functionalities. Predicting and discovering new materials with desired characteristics has therefore become a leading frontier in the study of topological materials. In recent years, the rise of machine learning has stimulated numerous new applications across areas ranging from protein engineering to materials science. Compared with the computational cost and complexity associated with identifying new materials through first principles calculations, machine learning provides an effective approach for predicting topological categories. Researchers have integrated topological analyses of crystal structures into machine learning models, which strengthens model robustness and improves predictive performance. This integration enables high throughput screening and the discovery of topological materials with compelling properties.

The machine learning methods surveyed in this review have demonstrated quantitatively validated performance across a variety of topological materials prediction tasks. For classification tasks based on chemical composition and symmetry information, gradient boosting trees achieved an overall accuracy of 89.7%. The corresponding F_1 scores reached 94.0% for trivial insulators, 70.0% for topological insulators, and 92.0% for topological semimetals. For spectral recognition tasks, neural networks trained on XANEs spectra achieved F_1 scores of 89% and 93% for topological and non topological materials, respectively.

For materials property prediction based on crystal structures, graph neural network models have shown a consistent trend of performance improvement. CGCNN achieved a mean absolute error of 0.039 eV/atom for formation energy prediction on the Materials Project dataset. MEGNet reduced this error to 0.028 eV/atom. ALIGNN further reduced the error to 0.022 eV/atom. Relative to MEGNet, ALIGNN achieved a 21.4% reduction in prediction error. Relative to CGCNN, the reduction reached 43.6%. In the context of inverse design, the CTMT framework generated 10,000 candidate materials and discovered 20

new topological materials after a multistage screening process. Among the dynamically stable candidates, the final discovery success rate reached 62.5%. For topological phase diagram construction, the quantum loop topology method achieved an accuracy exceeding 99.9% when distinguishing topological phases from trivial phases in Chern insulator models. These quantitative benchmark results provide concrete and valuable reference standards for future methodological development and performance evaluation.

This review provides a systematic account of the ways in which research on topological materials and machine learning techniques have empowered one another and progressed in a mutually reinforcing manner. It also examines the many successful applications that machine learning has demonstrated across different areas and levels of topological materials research. These areas include inverse design and generation of topological materials, investigations of topological superconductors, and the mapping of topological phase diagrams. In addition, this review analyzes several topological characteristics of datasets in topological materials and points out that TDL is a well suited tool for handling these characteristics. Although machine learning has achieved remarkable advances in the study of topological materials, the continual growth in the scale and complexity of topological materials datasets presents ongoing challenges. These challenges call for persistent refinement and design of machine learning models in order to effectively reveal the patterns and relationships embedded in the data. Finally, this review proposes approaches for addressing these challenges from two complementary perspectives.

The interaction between machine learning and topological materials has driven development and innovation in both fields. On one hand, the complexity, high dimensionality, multiscale structure, higher order interactions, and nonlinear relationships present in data from topological materials place higher demands on the methods that machine learning uses to process data. On the other hand, machine learning can search for and generate new materials by learning the distribution of known materials within datasets, thereby producing a richer variety of topological materials [90]. To further strengthen interdisciplinary research between machine learning and topological materials, the research community must adopt approaches that are guided by principles of innovation, collaboration, and open science, and foster a dynamic environment for interdisciplinary cooperation. This requires researchers to actively engage in open communication and the sharing of ideas while integrating expertise from mathematics, biology, chemistry, physics, computer science, and other related disciplines.

Acknowledgement: Not applicable.

Funding Statement: This work was supported in part by the Natural Science Foundation of China (NSFC Grant No. 12401080), Scientific Research Foundation of Chongqing University of Technology and the Science and Technology Research Program of Chongqing Municipal Education Commission, grant number KJQN202501109.

Author Contributions: Jing-Wen Gao contributed to the conceptualization, methodology, investigation, formal analysis, and writing of the original draft. Yunan He was responsible for the validation, visualization, and manuscript revisions. Jian Liu supervised the overall research process, contributed to the conceptualization, and project administration. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Glossary of abbreviations

Abbreviation	Full Term
ALIGNN	Atomistic Line Graph Neural Network
CDVAE	Crystal Diffusion Variational Autoencoder
CGCNN	Crystal Graph Convolutional Neural Network
CNN	Convolutional Neural Network
CTMT	CDVAE + Topogivity + M3GNet + TQC
DFT	Density Functional Theory
DiffCSP	Diffusion for Crystal Structure Prediction
EBRs	Elementary Band Representations
GBT	Gradient Boosting Tree
GNN	Graph Neural Network
Grad-CAM	Gradient-Weighted Class Activation Mapping
HOTIs	Higher-Order Topological Insulators
ICSD	Inorganic Crystal Structure Database
MACE	Message Passing Atomic Cluster Expansion
MAE	Mean Absolute Error MatterGen: Matter Generation
MEBRs	Magnetic Elementary Band Representations
MEGNet	MatErials Graph Network ML: Machine Learning
M3GNet	Materials 3-Body Graph Network
MTQC	Magnetic Topological Quantum Chemistry
NequIP	Neural Equivariant Interatomic Potential
NN	Neural Network
PCA	Principal Component Analysis
SISSO	Sure Independence Screening and Sparsifying Operator
TC	Topological Crystalline
TCIs	Topological Crystalline Insulators
TDA	Topological Data Analysis
TDL	Topological Deep Learning
TIs	Topological Insulators
TQC	Topological Quantum Chemistry
TSMs	Topological Semimetals
t-SNE	t-Distributed Stochastic Neighbor Embedding
XANES	X-ray Absorption Near-Edge Structure XAS: X-ray Absorption Spectroscopy

References

1. Bansil A, Lin H, Das T. Colloquium: topological band theory. Rev Mod Phys. 2016;88(2):021004. doi:10.1103/RevModPhys.88.021004.
2. Luo H, Yu P, Li G, Yan K. Topological quantum materials for energy conversion and storage. Nat Rev Phys. 2022;4(9):611–24. doi:10.1038/s42254-022-00477-9.
3. Hensel F, Moor M, Rieck B. A survey of topological machine learning methods. Front Artif Intell. 2021;4:681108. doi:10.3389/frai.2021.681108.
4. Qi XL, Zhang SC. Topological insulators and superconductors. Rev Mod Phys. 2011;83(4):1057. doi:10.1103/RevModPhys.83.1057.
5. Armitage NP, Mele EJ, Vishwanath A. Weyl and Dirac semimetals in three-dimensional solids. Rev Mod Phys. 2018;90(1):015001. doi:10.1103/RevModPhys.90.015001.
6. Hu J, Xu SY, Ni N, Mao Z. Transport of topological semimetals. Annu Rev Mater Res. 2019;49(1):207–52. doi:10.1146/annurev-matsci-070218-010023.

7. Sierra JF, Fabian J, Kawakami RK, Roche S, Valenzuela SO. Van der Waals heterostructures for spintronics and opto-spintronics. Nat Nanotechnol. 2021;16(8):856–68. doi:10.1038/s41565-021-00936-x.
8. Tang F, Po HC, Vishwanath A, Wan X. Comprehensive search for topological materials using symmetry indicators. Nature. 2019;566(7745):486–9. doi:10.1038/s41586-019-0937-5.
9. Bradlyn B, Elcoro L, Cano J, Vergniory MG, Wang Z, Felser C, et al. Topological quantum chemistry. Nature. 2017;547(7663):298–305. doi:10.1038/nature23268.
10. Xiao J, Yan B. First-principles calculations for topological quantum materials. Nat Rev Phys. 2021;3(4):283–97. doi:10.1038/s42254-021-00292-8.
11. Fu L, Kane CL. Topological insulators with inversion symmetry. Phys Rev B. 2007;76(4):045302. doi:10.1103/PhysRevB.76.045302.
12. Schindler F, Cook AM, Vergniory MG, Wang Z, Parkin SS, Bernevig BA, et al. Higher-order topological insulators. Sci Adv. 2018;4(6):eaat0346. doi:10.1126/sciadv.aat0346.
13. Langbehn J, Peng Y, Trifunovic L, von Oppen F, Brouwer PW. Reflection-symmetric second-order topological insulators and superconductors. Phys Rev Lett. 2017;119(24):246401. doi:10.1103/PhysRevLett.119.246401.
14. Benalcazar WA, Bernevig BA, Hughes TL. Electric multipole moments, topological multipole moment pumping, and chiral hinge states in crystalline insulators. Phys Rev B. 2017;96(24):245115. doi:10.1103/PhysRevB.96.245115.
15. Fang C, Fu L. New classes of topological crystalline insulators having surface rotation anomaly. Sci Adv. 2019;5(12):eaat2374. doi:10.1126/sciadv.aat2374.
16. Burkov A, Hook M, Balents L. Topological nodal semimetals. Phys Rev B. 2011;84(23):235126. doi:10.1103/PhysRevB.84.235126.
17. Hsieh TH, Lin H, Liu J, Duan W, Bansil A, Fu L. Topological crystalline insulators in the SnTe material class. Nat Commun. 2012;3(1):982. doi:10.1038/ncomms1969.
18. Benalcazar WA, Bernevig BA, Hughes TL. Quantized electric multipole insulators. Science. 2017;357(6346):61–6. doi:10.1126/science.aah6442.
19. Fu L. Topological crystalline insulators. Phys Rev Lett. 2011;106(10):106802. doi:10.1103/PhysRevLett.106.106802.
20. He Y, De Breuck PP, Weng H, Giantomassi M, Rignanese GM. Machine learning on multiple topological materials datasets. npj Comput Mater. 2025;11(1):181. doi:10.1038/s41524-025-01687-2.
21. Mishchenko A, Bhattacharya A, Wang X, Pentz HK, Wei Y, Yang Q. Deep learning methods for 2D material electronic properties. Digit Discov. 2026;5(1):28–63. doi:10.1039/D5DD00155B.
22. Kane CL, Mele EJ. $\mathbb{Z}_2$ topological order and the quantum spin Hall effect. Phys Rev Lett. 2005;95(14):146802. doi:10.1103/PhysRevLett.95.146802.
23. Po HC, Vishwanath A, Watanabe H. Symmetry-based indicators of band topology in the 230 space groups. Nat Commun. 2017;8(1):50. doi:10.1038/s41467-017-00133-2.
24. Cao G, Ouyang R, Ghiringhelli LM, Scheffler M, Liu H, Carbogno C, et al. Artificial intelligence for high-throughput discovery of topological insulators: the example of alloyed tetradymites. Phys Rev Mater. 2020;4(3):034204. doi:10.1103/PhysRevMaterials.4.034204.
25. Ma A, Zhang Y, Christensen T, Po HC, Jing L, Fu L, et al. Topogivity: a machine-learned chemical rule for discovering topological materials. Nano Lett. 2023;23(3):772–8. doi:10.1021/acs.nanolett.2c03307.
26. Claussen N, Bernevig BA, Regnault N. Detection of topological materials with machine learning. Phys Rev B. 2020;101(24):245117. doi:10.1103/PhysRevB.101.245117.
27. Hong T, Chen T, Jin D, Zhu Y, Gao H, Zhao K, et al. Discovery of new topological insulators and semimetals using deep generative models. npj Quantum Mater. 2025;10(1):12. doi:10.1038/s41535-025-00731-0.
28. Elcoro L, Wieder BJ, Song Z, Xu Y, Bradlyn B, Bernevig BA. Magnetic topological quantum chemistry. Nat Commun. 2021;12(1):5965. doi:10.1038/s41467-021-26241-8.
29. Xu Y, Elcoro L, Song Z, Wieder BJ, Vergniory MG, Regnault N, et al. High-throughput calculations of magnetic topological materials. Nature. 2020;586(7831):702–7. doi:10.1038/s41586-020-2837-0.
30. Vergniory MG, Elcoro L, Felser C, Regnault N, Bernevig BA, Wang Z. A complete catalogue of high-quality topological materials. Nature. 2019;566(7745):480–5. doi:10.1038/s41586-019-0954-4.

31. Zhang T, Jiang Y, Song Z, Huang H, He H, Fang Z, et al. Catalogue of topological electronic materials. Nature. 2019;566(7745):475–9. doi:10.1038/s41586-019-0944-6.
32. Tang F, Po HC, Vishwanath A, Wan X. Efficient topological materials discovery using symmetry indicators. Nat Phys. 2019;15(5):470–6. doi:10.1038/s41567-019-0418-7.
33. Watanabe H, Po HC, Vishwanath A. Structure and topology of band structures in the 1651 magnetic space groups. Sci Adv. 2018;4(8):eaat8685. doi:10.1126/sciadv.aat8685.
34. Kruthoff J, de Boer J, van Wezel J, Kane CL, Slager RJ. Topological classification of crystalline insulators through band structure combinatorics. Phys Rev X. 2017;7(4):041069. doi:10.1103/PhysRevX.7.041069.
35. Song Z, Zhang T, Fang Z, Fang C. Quantitative mappings between symmetry and topology in solids. Nat Commun. 2018;9(1):3530. doi:10.1038/s41467-018-06010-w.
36. Khalaf E, Po HC, Vishwanath A, Watanabe H. Symmetry indicators and anomalous surface states of topological crystalline insulators. Phys Rev X. 2018;8(3):031070. doi:10.1103/PhysRevX.8.031070.
37. Hellenbrandt M. The Inorganic Crystal Structure Database (ICSD)—present and future. Crystallogr Rev. 2004;10(1):17–22. doi:10.1080/08893110410001664882.
38. Andrejevic N, Andrejevic J, Bernevig BA, Regnault N, Han F, Fabbris G, et al. Machine learning spectral indicators of topology. Adv Mater. 2022;34(49):2204113. doi:10.1002/adma.202204113.
39. Xu H, Jiang Y, Wang H, Wang J. Discovering two-dimensional magnetic topological insulators by machine learning. Phys Rev B. 2024;109(3):035122. doi:10.1103/PhysRevB.109.035122.
40. Carrasquilla J, Melko RG. Machine learning phases of matter. Nat Phys. 2017;13(5):431–4. doi:10.1038/nphys4035.
41. van Nieuwenburg E, Liu YH, Huber S. Learning phase transitions by confusion. Nat Phys. 2017;13(5):435–9. doi:10.1038/nphys4037.
42. Zhang P, Shen H, Zhai H. Machine learning topological invariants with neural networks. Phys Rev Lett. 2018;120(6):066401. doi:10.1103/PhysRevLett.120.066401.
43. Cheng QQ, Luo WW, He AL, Wang YF. Topological quantum phase transitions of Chern insulators in disk geometry. J Phys Condens Matter. 2018;30(35):355502. doi:10.1088/1361-648x/aad51f.
44. Sun N, Yi J, Zhang P, Shen H, Zhai H. Deep learning topological invariants of band insulators. Phys Rev B. 2018;98(8):085402. doi:10.1103/PhysRevB.98.085402.
45. Kapustin A, Spodyneiko L. Higher-dimensional generalizations of Berry curvature. Phys Rev B. 2020;101(23):235130. doi:10.1103/PhysRevB.101.235130.
46. Berry MV. Quantal phase factors accompanying adiabatic changes. Proc R Soc A. 1984;392(1802):45–57. doi:10.1098/rspa.1984.0023.
47. Liu J, Cao G, Zhou Z, Liu H. Screening potential topological insulators in half-Heusler compounds via compressed-sensing. J Phys Condens Matter. 2021;33(32):325501. doi:10.1088/1361-648X/abba8d.
48. Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett. 2018;120(14):145301. doi:10.1103/PhysRevLett.120.145301.
49. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans Neural Netw. 2009;20(1):61–80. doi:10.1109/TNN.2008.2005605.
50. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. 2013;1(1):011002. doi:10.1063/1.4812323.
51. Horton MK, Huck P, Yang RX, Munro JM, Dwaraknath S, Ganose AM, et al. Accelerated data-driven materials science with the Materials Project. Nat Mater. 2025;24(10):1522–32. doi:10.1038/s41563-025-02272-0.
52. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater. 2019;31(9):3564–72. doi:10.1021/acs.chemmater.9b01294.
53. Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. npj Comput Mater. 2022;8(1):221. doi:10.1038/s41524-022-00913-5.
54. Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nat Commun. 2022;13(1):2453. doi:10.1038/s41467-022-29939-5.

55. Batatia I, Kovacs DP, Simm GNC, Ortner C, Csanyi G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. Adv Neural Inf Process Syst. 2022;35:11423–36. doi:10.52202/068431-0830.
56. Jiao R, Huang W, Lin P, Han J, Chen P, Lu Y, et al. Crystal structure prediction by joint equivariant diffusion. Adv Neural Inf Process Syst. 2023;36. doi:10.52202/075280-0767.
57. Zeni C, Pinsler R, Zuegner D, Fowler A, Horton M, Fu X, et al. A generative model for inorganic materials design. Nature. 2025;639:624–32. doi:10.1038/s41586-025-08628-5.
58. Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. npj Comput Mater. 2020;6(1):138. doi:10.1038/s41524-020-00406-3.
59. Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. Nat Comput Sci. 2022;2(11):718–28. doi:10.1038/s43588-022-00349-3.
60. Lv B, Weng H, Fu B, Wang X, Miao H, Ma J, et al. Experimental discovery of Weyl semimetal TaAs. Phys Rev X. 2015;5(3):031013. doi:10.1103/PhysRevX.5.031013.
61. Bradlyn B, Cano J, Wang Z, Vergniory MG, Felser C, Cava RJ, et al. Beyond Dirac and Weyl fermions: unconventional quasiparticles in conventional crystals. Science. 2016;353(6299):aaf5037. doi:10.1126/science.aaf5037.
62. Xie YM, Gao XJ, Xu XY, Zhang CP, Hu JX, Gao JZ, et al. Kramers nodal line metals. Nat Commun. 2021;12(1):3064. doi:10.1038/s41467-021-22903-9.
63. He WY, Xu XY, Law KT. Kramers Weyl semimetals as quantum solenoids and their applications in spin-orbit torque devices. Commun Phys. 2021;4(1):66. doi:10.1038/s42005-021-00564-w.
64. Metni H, Ruple L, Walters LN, Torresi L, Teufel J, Schopmans H, et al. Generative models for crystalline materials. Adv Mater. 2026;38(18):e23620. doi:10.1002/adma.202523620.
65. Ye C, Wang Y, Xie X, Zhu T, Liu J, He Y, et al. Materials discovery acceleration by using conditional generative methodology. npj Comput Mater. 2026;12(1):63. doi:10.1038/s41524-025-01930-w.
66. Tsai YH, Yu MZ, Hsu YH, Chung MC. Deep learning of topological phase transitions from entanglement aspects. Phys Rev B. 2020;102(5):054512. doi:10.1103/PhysRevB.102.054512.
67. Kitaev AY. Unpaired Majorana fermions in quantum wires. Phys-Usp. 2001;44:131. doi:10.1070/1063-7869/44/10S/S29.
68. Chung MC, Cheng TP, Huang GY, Tsai YH. Deep learning of topological phase transitions from the point of view of entanglement for two-dimensional chiral p-wave superconductors. Phys Rev B. 2021;104(2):024506. doi:10.1103/PhysRevB.104.024506.
69. Zhang Y, Kim EA. Quantum loop topography for machine learning topological phases. Phys Rev Lett. 2017;118(21):216401. doi:10.1103/PhysRevLett.118.216401.
70. Rodriguez-Nieva JF, Scheurer MS. Identifying topological order through unsupervised machine learning. Nat Phys. 2019;15(8):790–5. doi:10.1038/s41567-019-0512-x.
71. Hasan MZ, Kane CL. Colloquium: topological insulators. Rev Mod Phys. 2010;82(4):3045. doi:10.1103/RevModPhys.82.3045.
72. Thouless DJ, Kohmoto M, Nightingale MP, den Nijs M. Quantized Hall conductance in a two-dimensional periodic potential. Phys Rev Lett. 1982;49(6):405–8. doi:10.1103/PhysRevLett.49.405.
73. Bradley CJ, Cracknell AP. The mathematical theory of symmetry in solids: representation theory for point groups and space groups. New York, NY,USA: Oxford University Press; 2009. 745 p.
74. Cano J, Bradlyn B, Wang Z, Elcoro L, Vergniory MG, Felser C, et al. Building blocks of topological quantum chemistry: elementary band representations. Phys Rev B. 2018;97(3):035139. doi:10.1103/PhysRevB.97.035139.
75. Zia A, Khamis A, Nichols J, Tayab UB, Hayder Z, Rolland V, et al. Topological deep learning: a review of an emerging paradigm. Artif Intell Rev. 2024;57(4):77. doi:10.1007/s10462-024-10710-9.
76. Edelsbrunner H, Zomorodian A. Topological persistence and simplification. Discret Comput Geom. 2002;28(4):511–33. doi:10.1007/s00454-002-2885-2.
77. Carlsson G. Topology and data. Bull Am Math Soc. 2009;46(2):255–308. doi:10.1090/S0273-0979-09-01249-X.

78. Cang Z, Wei GW. TopologyNet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLoS Comput Biol. 2017;13(7):e1005690. doi:10.1371/journal.pcbi.1005690.
79. Bodnar C, Frasca F, Wang YG, Otter N, Montúfar G, Liò P, et al. Weisfeiler and Lehman go topological: message passing simplicial networks. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 1026–37. doi:10.48550/arxiv.2103.03212.
80. Xia K, Feng X, Tong Y, Wei GW. Persistent homology for the quantitative prediction of fullerene stability. J Comput Chem. 2015;36(6):408–22. doi:10.1002/jcc.23816.
81. Mémoli F, Wan Z, Wang Y. Persistent Laplacians: properties, algorithms and implications. SIAM J Math Data Sci. 2022;4(2):858–84. doi:10.1137/21M1435471.
82. Chen D, Liu J, Wu J, Wei GW, Pan F, Yau ST. Path topology in molecular and materials sciences. J Phys Chem Lett. 2023;14(4):954–64. doi:10.1021/acs.jpclett.2c03706.
83. Liu J, Chen D, Wei GW. Persistent interaction topology in data analysis. Found Data Sci. 2025;9:34–60. doi:10.3934/fods.2025011.
84. Lyngby P, Thygesen KS. Data-driven discovery of 2D materials by deep generative models. npj Comput Mater. 2022;8(1):232. doi:10.1038/s41524-022-00923-3.
85. Liu J, Li J, Wu J. The algebraic stability for persistent Laplacians. Homol Homot Appl. 2024;26(2):3902. doi:10.4310/HHA.2024.v26.n2.a15.
86. Shen L, Feng H, Li F, Lei F, Wu J, Wei GW. Knot data analysis using multiscale Gauss link integral. Proc Natl Acad Sci U S A. 2024;121(42):e2408431121. doi:10.1073/pnas.2408431121.
87. Marchand D. Foundation models for metallurgy? MRS Bull. 2025;50(7):805–18. doi:10.1557/s43577-025-00911-0.
88. Bilbrey JA, Firoz JS, Lee MS, Choudhury S. Uncertainty quantification for neural network potential foundation models. npj Comput Mater. 2025;11(1):109. doi:10.1038/s41524-025-01572-y.
89. Song R, Li F, Wu J, Lei F, Wei GW. Multi-scale Jones polynomial and persistent Jones polynomial for knot data analysis. AIMS Math. 2025;10(1):1463. doi:10.3934/math.2025068.
90. Rasul A, Hossain MS, Dastider AG, Roy H, Hasan MZ, Khosru QDM. A machine learning based classifier for topological quantum materials. Sci Rep. 2024;14(1):31564. doi:10.1038/s41598-024-68920-8.