Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physics Informed Inverse Discovery of Potential Topological Materials via Descriptor-Graph Ensemble Learning

Tasneem Rehman
University of Kashmir https://orcid.org/0009-0005-6308-4146
Nasreena Yaqoob
University of Kashmir https://orcid.org/0009-0001-1029-8221
Aadil Ahmad Ganaie
University of Kashmir
Sheikh Irfan
University of Kashmir
Ghulam Dar
gulnoor.dar@gmail.com
University of Kashmir https://orcid.org/0000-0002-5066-0141

Article

Keywords: Topological Materials, Machine Learning, Crystal GNN, Inverse Materials Discovery, Topological Quantum Chemistry

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9475817/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is no conflict of interest

Physics Informed Inverse Discovery of Potential Topological Materials via Descriptor-Graph Ensemble Learning

Tasneem U Rehman[1], Nasreena Yaqoob[1], Aadil Ahmad Ganaie[1], Sheikh Irfan Ganaie[1], Ghulam Nabi Dar[1*]

[1]Department of Physics, University of Kashmir, Hazratbal, Srinagar, 190006, Jammu and Kashmir, India.

*Corresponding author(s). E-mail(s): gulnoor.dar@gmail.com;
Contributing authors: hamitasneem@gmail.com; nasreenayaqoob7@gmail.com; aadilaganaie@gmail.com; sirfan009@gmail.com;

Abstract

Topological materials (TMs) constitute a fundamentally new class of quantum matter, hosting symmetry-protected electronic states with robust transport and spin–momentum locking that are central to next-generation electronic, spintronic, and quantum technologies. However, their identification is conventionally driven by first-principles electronic structure calculations and explicit evaluation of topological invariants, which limits scalability across large chemical spaces. Here, we present a physics-informed machine learning (ML) framework for discovering topologically non-trivial materials without directly computing wavefunction-based invariants, integrating electronic structure–derived proxy descriptors, symmetry-resolved topological databases, tabular ML models, and crystal graph neural networks (GNNs) into a unified prediction and inverse discovery pipeline. Using symmetry-indicator-derived $\mathbb{Z}_2$ invariants ($\boldsymbol{\nu_0, \nu_1\nu_2\nu_3}$) as supervised labels, we curate a dataset of 9,142 two- and three-dimensional crystalline materials and construct physically motivated descriptors capturing band gaps, bandwidths, Fermi-level crossings, and dispersion characteristics alongside composition- and structure-based representations. Tabular models achieve ROC–AUC values $> \mathbf{0.8}$, while GNNs trained solely on atomic connectivity achieve $> \mathbf{0.65}$, demonstrating that crystal structure alone encodes statistically meaningful but incomplete information about topological propensity. In an inverse discovery setting, the models assign probabilistic scores and enable consensus ranking that filters approximately 45,000 candidates to about 1,200 high-confidence materials, revealing complementary sensitivities between composition- and structure-based representations. Overall, this work establishes a scalable and physics-guided pathway for inverse discovery of topological materials beyond brute-force electronic structure calculations.

Keywords: Topological Materials, Machine Learning, Crystal GNN, Inverse Materials Discovery, Topological Quantum Chemistry

1 Introduction

Topological electronic phases including topological insulators, topological crystalline insulators, and topological semimetals host robust boundary states and unconventional bulk responses that are promising for low dissipation electronics, quantum information, and spintronic applications[1–5]. The discovery and classification of TMs has therefore become a major thrust in condensed matter and materials science. Early theoretical frameworks such as Topological Quantum Chemistry (TQC) and symmetry indicator methods enabled systematic identification of topological phases from band representations, dramatically accelerating database scale searches for topological compounds [6, 7].

High throughput first principles scans and catalogues have since produced extensive atlases of candidate TMs, demonstrating that a large fraction of experimentally characterised inorganic crystals host nontrivial topology and providing curated labels that serve as ground truth for data driven studies [8, 9]. Despite these successes, direct evaluation of topological invariants requires electronic wavefunction information and (often expensive) post processing (Wilson loop / Wannier charge center or symmetry indicator calculations). This limits the throughput of purely first principles discovery workflows and motivates predictive surrogates that can infer topological propensity from easier inputs. ML provides a promising route, that is composition or structure based ML models can rapidly screen large chemical spaces and produce candidate lists for focused DFT verification [10, 11].

Two complementary ML paradigms have emerged in materials informatics. Baseline Tabular ML models trained on engineered descriptors (stoichiometric statistics, elemental attributes, DFT-derived electronic proxies) deliver fast and interpretable classifiers for many properties. GNNs, on the other hand, operate directly on crystal graphs and can learn subtle geometry and coordination effects without manual featurization. Distance aware message passing (e.g., CGCNN) and more recent equivariant architectures (e.g., NequIP, E(3)-equivariant GNNs) have set new state of the art results for property prediction tasks [12–14].

Prior works have applied ML to detect topological phases by learning from bandstructure features or database labels, but many approaches either rely on band derived inputs (which removes the advantage of bypassing DFT) or have limited ability to generalize across chemistries and dimensionalities. To enable practical inverse discovery i.e., find candidate materials likely to be topologically nontrivial without performing full wavefunction analysis for every candidate. We therefore need (i) robust, physics guided proxy descriptors that capture electronic fingerprints of topology (ii) structure aware learning architectures that exploit geometry and symmetry while remaining scalable and (iii) reproducible training and cross validation protocols with rigorous uncertainty and performance reporting [9, 15].

In this work we present a unified pipeline that integrates curated DFT outputs, symmetry based topological labels, a set of physically motivated electronic proxies, composition only and structure aware ML models (tabular ensembles and GNNs), and an inverse scoring procedure to rank candidate materials for topological non triviality. Our contributions are threefold- (i) we assemble and clean a large dataset of materials with SCF converged DFT outputs and database topological labels $(\nu_0, \nu_1\nu_2\nu_3)$ and derive robust electronic and structural proxies (ii) we benchmark discriptors based tabular baselines against structure based GNNs using stratified cross validation and comprehensive metrics (ROC-AUC, PR-AUC, F1, MCC, correlation measures) and (iii) we demonstrate an inverse discovery workflow that explores materials from vast chemical space and fuses tabular and graph model scores to produce a candidate list those are potentially posses topological phases [11, 16, 17].

2 Methodology

This work presents a physics informed ML framework for identifying non trivial TMs without explicitly computing topological invariants from first principles wavefunctions. The methodology integrates electronic structure data, symmetry based topological databases, tabular baseline ML, and crystal GNN into a unified forward prediction and inverse discovery pipeline. The workflow consists of four major stages: (i) data acquisition and preprocessing, (ii) construction of physically motivated electronic and structural descriptors, (iii) supervised learning using both tabular baseline and graph based models, and (iv) inverse ranking of candidate materials for topological non triviality.

2.1 Data Collection & Feature Engineering

Electronic structure data were collected from open access repositories containing density functional theory (DFT) calculations of crystalline materials. Only calculations that achieved self consistent field (SCF) convergence were retained. Key quantities extracted include total energy, Fermi energy, band dispersions, and k-point sampling density. Supervised labels were assigned using curated TMs databases that provide symmetry indicators and $\mathbb{Z}_2$ topological invariants derived from band representations and topological quantum chemistry theory [6, 7]. The strong topological index ν_0 was selected as the primary classification target and treated as a binary label distinguishing topologically trivial and non trivial phases. Materials were matched across datasets using normalized chemical formulas. In cases where multiple entries existed for a given formula, conservative labeling was applied to minimize false positive assignments.

Direct learning of topological invariants from raw electronic wavefunctions is computationally expensive and data intensive. Instead, this work adopts a set of physically motivated proxy descriptors that capture electronic signatures commonly associated with topological phases, inspired by prior studies on band inversion, nodal crossings, and flat band physics [15, 18].

The extracted proxies include:

- **Band gap (E_g):** Minimum direct or indirect band gap near the Fermi level.
- **Bandwidth (W):** Energy spread of occupied and near Fermi bands.
- **Number of Fermi level crossings ($N_{\mathbf{cross}}$):** Count of band crossings within a small energy window around E_F.
- **Density of states proxy at E_F (D_{E_F}):** Approximated from band slopes near the Fermi level.
- **Dispersion anisotropy (A_d):** Ratio of maximal to minimal band dispersion across reciprocal space directions.
- **Flat band ratio (F_b):** Fraction of bands with near zero dispersion, indicative of enhanced correlation or topology driven localization.

These descriptors provide continuous, noise tolerant representations suitable for ML, while remaining grounded in electronic structure physics (See Section Table S1 in Supplementary Information for detailed description of proxy discriptors).

After merging electronic descriptors with topological labels, extensive data cleaning was performed. Non numeric identifiers were removed, binary flags were consistently encoded, and columns containing exclusively missing values were discarded. The final dataset comprises 9,142 unique materials described by 175 numerical features. Materials were separated into two dimensional (2D) and three dimensional (3D) subsets based on crystallographic dimensionality. This separation avoids conflating fundamentally distinct bonding environments and electronic behavior [19].

2.2 Model Selection, Training and Evaluation

A diverse set of baseline classifiers was trained and evaluated, including Logistic Regression [20], Random Forests [21], Gradient Boosting [22], XGBoost [23] and LightGBM [24]. Models were trained using stratified cross validation. GridSearch was performed to tune and select best hyperparameters for all the models. Performance was evaluated using Receiver Operating Characteristic - Area Under the Curve (ROC-AUC), precision recall AUC (PR-AUC), F1 score and Matthews correlation coefficient (MCC), which is particularly robust under class imbalance [25, 26]. See Supplementary Table S2 for detailed definitions of the metrics used.

In addition to the descriptor based learning, crystal structures were retrieved programmatically from the `Materials Project` database [27] for the said materials. Structures were stored as CIF files and parsed using the `pymatgen` library [28]. Each crystal was converted into a periodic graph representation, where nodes correspond to atomic sites and edges represent interatomic bonds defined via a radial cutoff. Node features encode elemental properties such as atomic number, electronegativity, and valence electron count, while edge attributes represent interatomic distances. Structure aware learning was then performed using GNNs implemented in PyTorch Geometric [17]. The evaluated architectures were **Graph Convolutional Networks (GCN)** [29], **Graph Isomorphism Networks (GIN)** [30] & **Crystal Graph Convolutional Neural Networks (CGCNN)** [12].

Models were trained separately for 2D and 3D datasets using stratified cross validation. Binary classification was optimized using the sigmoid cross entropy loss:

$$\mathcal{L} = -\left[y \log(\sigma(z)) + (1-y) \log(1-\sigma(z))\right], \tag{1}$$

where z is the model output and σ denotes the sigmoid function.

2.3 Inverse discovery and candidate ranking

We implemented a multi stage inverse discovery pipeline to prioritize candidate TMs from a large enumerated chemical space as summarized in the Table 1. First, binary and ternary stoichiometries were systematically enumerated and subjected to physics informed pruning, including an atomic number based spin orbit coupling (SOC) proxy, dimensional heuristics, and elemental cardinality constraints. The surviving compositions were featurized using Magpie style elemental descriptors together with valence orbital features as implemented in `matminer`. These composition level features matched the schema used to train tabular classifiers during forward learning. LightGBM trained on nominal 2D and 3D datasets using stratified cross validation and hyperparameter optimisation was subsequently

applied to these hypothetical compositions to assign a probabilistic score P_{Tabular} interpreted as the likelihood of topological non triviality (See Section S4 in Supplementary Information for detailed description on the definition of scoring parameters). This composition driven screening step enables large scale exploration of chemical space without explicit calculation of topological invariants, consistent with prior TMs cataloguing efforts [8, 9].

High confidence compositions were then expanded into multiple crystallographic prototypes spanning representative space groups, and corresponding CIF structures were generated. These structures were converted into crystal graphs using radius based neighbour construction (6 Å cutoff) and 8 dimensional atomic feature vectors, consistent with the Crystal Graph Convolutional Neural Network (CGCNN) formalism [12]. An ensemble of five cross validated CGCNN folds (trained on the labeled dataset) was used to predict structure sensitive probabilities P_{CGCNN} for each prototype. For each composition, the maximum scoring prototype was retained. The final consensus probability was defined as a fixed weighted fusion,

$$P_{\text{Joint}} = \alpha P_{\text{Tabular}} + (1 - \alpha) P_{\text{CGCNN}}, \tag{2}$$

with $\alpha = 0.6$ to balance composition driven trends and structure aware refinement. Candidates were ranked by P_{Joint}, de-duplicated across stoichiometric variants and prototypes, and exported for downstream exploration. This hybrid composition structure inverse framework combines baseline tabular screening with GNN refinement, providing an interpretable and computationally efficient route for TMs discovery.

Table 1: Multi-stage inverse design and screening pipeline

Step	Stage Name	Method (Reduction Strategy)	Count
1	Combinatorial Enumeration	Ultra broad chemical enumeration across allowed compositions	2,505,840
2	SOC Pre Screening	SOC based filtering for heavy element systems	2,080,701
3	Physics Informed Pruning	Descriptor based physical constraints and symmetry rules	695,315
4	Advanced Feature Pruning	Chemical heuristics and stability based elimination	531,580
5	Formula Deduplication	Removal of compositional redundancies (unique formulas retained)	43,695
6	Tabular ML Screening	Probability scoring using tabular ensemble models	43,695
7	High Confidence Selection	Threshold filtering ($P_{\text{Tabular}} > 0.75$)	1,280
8	Prototype Expansion	CIF generation (10 symmetry prototypes per composition)	12,800
9	CGCNN Structural Screening	Structure aware graph neural network evaluation	1,280
10	Consensus Ranking	Joint probability fusion and final candidate ranking	1,280

2.4 Computational Details

All data processing and ML workflows were implemented in Python. Electronic structure parsing was performed using custom scripts, while structural handling relied on `pymatgen`. Tabular models were implemented using `scikit-learn` [31], `XGBoost`, `LightGBM`, and `CatBoost`. GNN were implemented using PyTorch [16] and PyTorch Geometric. All experiments were conducted using fixed random seeds. Intermediate datasets, trained models, cross-validation metrics, and evaluation plots were saved to disk. Calculations were performed on CPU based systems, and library versions were explicitly pinned to ensure numerical stability and reproducibility. See Supplementary Table S4

3 Results and Discussion

3.1 Dataset Overview

After automated parsing, cleaning, and deduplication, the final dataset comprises 9,142 unique crystalline materials, each annotated with symmetry resolved topological invariants (Table 2 & 3). The dataset spans a wide chemical and structural diversity, with materials distributed across effective dimensionalities 2D & 3D. Three dimensional materials dominate the dataset, while two dimensional systems form a smaller but chemically distinct subset.

Table 2: Summary statistics of the final curated dataset used in this study.

Statistic	**Value**
Total materials	9142
2D materials	1198
3D materials	5238
Topological (positive class) fraction	2150
Non-topological fraction	6992
Centrosymmetric materials	7807
Non-centrosymmetric materials	1335

Table 3: Statistics of crystal graph construction from CIF files.

Statistic	**Value**
Total CIFs attempted	9142
Successfully parsed structures	9106
Graphs successfully built	9100
Failed graph constructions	06
Common failure reasons	Valence ambiguity, malformed CIFs

The prediction target is the strong topological invariant ν_0, treated as a binary classification problem. The resulting class distribution is moderately imbalanced, with topologically non trivial materials forming a minority class in both 2D and 3D subsets as summarised in Table 2. Supplementary Figures S1-S4 provide some important distributions of the dataset used in this work.

3.2 Model Performance Metrics

Under five-fold stratified cross validation, Tabular baseline trained on the complete physics guided feature set, including electronic structure proxies derived from band dispersions, achieve near perfect predictive performance as summarised in Table 4. These models achieve ROC-AUC values in the range 0.82-0.87 for 2D materials, with LightGBM providing the strongest overall performance across ROC-AUC, PR-AUC, and MCC. For 3D materials, performance decreases to ROC-AUC values of 0.71-0.78, reflecting the increased complexity and reduced predictability of three dimensional electronic topology from composition alone. Tabular models Ensemble methods such as Random Forests, ExtraTrees, XGBoost, and LightGBM consistently yield ROC-AUC values approaching unity, as summarized in Fig.

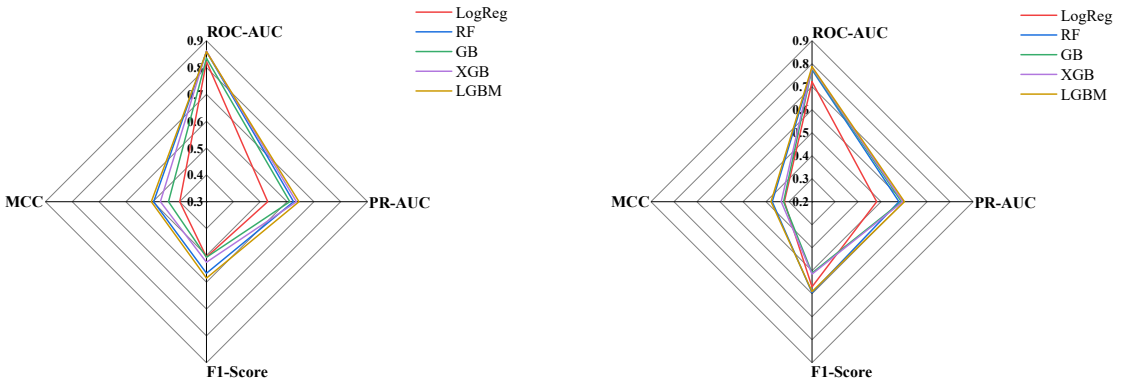
Table 4: Performance of baseline ML models with values corresponding to mean ± standard deviation over all folds.

Model	**ROC–AUC**	**PR–AUC**	**F1-score**	**MCC**
		2D Materials		
LogReg	0.822 ± 0.042	0.528 ± 0.078	0.506 ± 0.037	0.400 ± 0.051
RF	0.859 ± 0.030	0.624 ± 0.075	0.567 ± 0.068	0.497 ± 0.082
GB	0.835 ± 0.040	0.611 ± 0.088	0.507 ± 0.073	0.442 ± 0.102
XGB	0.861 ± 0.032	0.635 ± 0.076	0.525 ± 0.049	0.472 ± 0.071
LGBM	0.859 ± 0.025	0.645 ± 0.084	0.585 ± 0.082	0.506 ± 0.098
		3D Materials		
LogReg	0.721 ± 0.006	0.481 ± 0.007	0.570 ± 0.011	0.320 ± 0.019
RF	0.774 ± 0.014	0.578 ± 0.020	0.596 ± 0.018	0.373 ± 0.029
GB	0.779 ± 0.014	0.588 ± 0.021	0.510 ± 0.019	0.324 ± 0.026
XGB	0.783 ± 0.010	0.594 ± 0.014	0.515 ± 0.010	0.334 ± 0.016
LGBM	0.787 ± 0.018	0.601 ± 0.023	0.591 ± 0.028	0.378 ± 0.042

Figure 2 represent the ROC curves for LGBM model trained on 2D and 3D datasets. See Supplementary Figures S5- S8 for for other baseline tabular models.

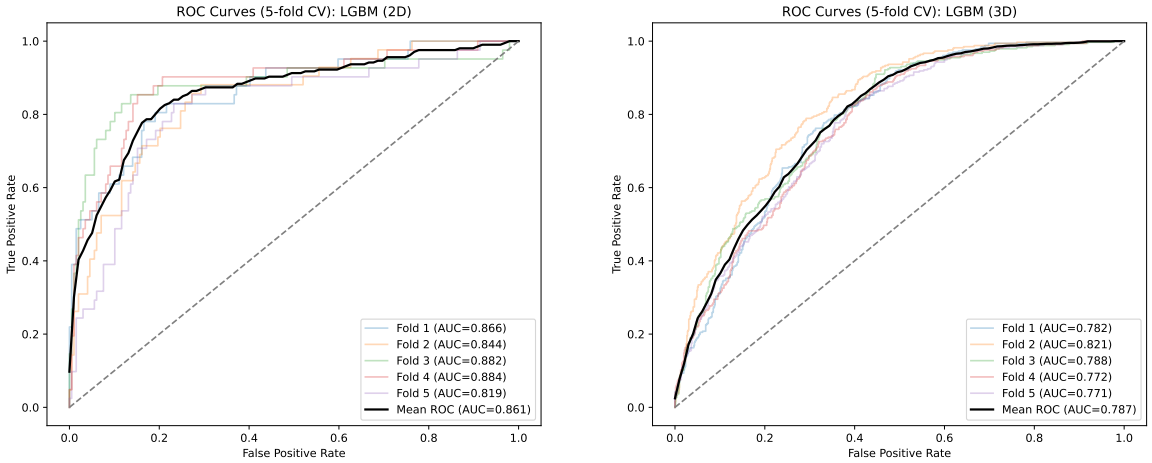
GNN were trained using only crystal structure information, without access to electronic structure descriptors. The metrics of these models are detailed in Table 5.

Across five-fold cross-validation, GNNs achieve ROC–AUC values ranging from 0.62 to 0.71 for 2D materials, with CGCNN consistently outperforming message-passing baselines. For 3D materials, performance converges across architectures, with ROC–AUC values clustered around 0.66–0.68. The



(a) 2D materials (b) 3D materials

Fig. 1: Overall performance comparison of tabular machine learning models (a) 2D and (b) 3D materials.



(a) 2D materials (b) 3D materials

Fig. 2: ROC curves for LightGBM applied to (a) 2D and (b) 3D materials.

superior performance of CGCNN highlights the importance of explicit distance-dependent edge features in capturing coordination environments and bonding geometry. However, CGCNN also exhibits increased sensitivity to graph construction parameters and edge feature dimensionality, reducing robustness in large-scale automated workflows. Overall, GNNs perform below composition-only tabular models in absolute metrics, but operate under strictly weaker information constraints, relying solely on structural geometry.

Figure 4 represent the ROC curves for CGCNN model trained on 2D and 3D datasets. See Supplementary Figures S9- S11 for for other Graph based models.

A direct comparison between baseline tabular and GNN models reveals a clear but physically interpretable hierarchy (Figures 5 and 6). Composition-only tabular models outperform graph models in both ROC–AUC and PR–AUC, while the performance gap narrows for 3D materials, where compositional descriptors alone are less predictive.

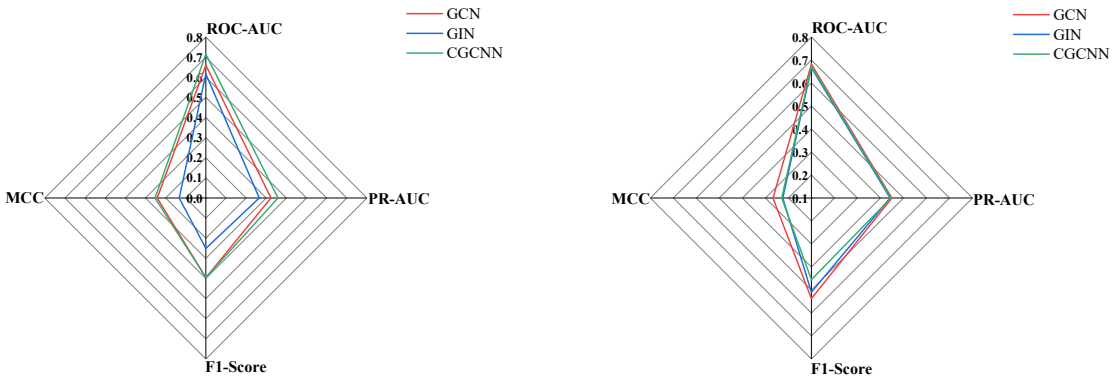
3.3 Inverse discovery and candidate ranking

Application of the trained models to the inverse designed candidate space yielded probabilistic topological scores for 43,695 chemically unique compositions and 12,800 structure instantiated graphs. For each candidate, three complementary confidence metrics were evaluated:

$$P_{\text{Tabular}}, \quad P_{\text{CGCNN}}, \quad P_{\text{Joint}}, \tag{3}$$

Table 5: Performance of GNN models with values corresponding to mean ± standard deviation over all folds.

Model	ROC–AUC	PR–AUC	F1-score	MCC
2D Materials				
GCN	0.660 ± 0.031	0.323 ± 0.048	0.396 ± 0.040	0.242 ± 0.055
GIN	0.616 ± 0.055	0.265 ± 0.051	0.251 ± 0.126	0.132 ± 0.070
CGCNN	0.714 ± 0.064	0.358 ± 0.060	0.395 ± 0.061	0.252 ± 0.099
3D Materials				
GCN	0.682 ± 0.017	0.450 ± 0.015	0.538 ± 0.015	0.267 ± 0.019
GIN	0.667 ± 0.019	0.442 ± 0.017	0.509 ± 0.054	0.224 ± 0.040
CGCNN	0.672 ± 0.023	0.446 ± 0.026	0.455 ± 0.174	0.229 ± 0.092



(a) 2D materials

(b) 3D materials

Fig. 3: Overall performance comparison of tabular machine learning models (a) 2D and (b) 3D materials.

where P_{Joint} provides a consensus ranking combining composition and structure based predictions.

Using P_{Tabular} as the primary ranking metric, 1,280 high confidence candidates were identified ($P_{\text{Tabular}} \geq 0.75$), with 375 materials exceeding $P_{\text{Tabular}} \geq 0.80$. Amomg these, 66 compounds exhibit strong consensus behavior with simultaneously high P_{Tabular} and P_{CGCNN} therefore high $P_{\text{Joint}} \geq 0.85$, indicating robust agreement between chemical and structural representations. The highest ranked materials are predominantly heavy element systems (Bi, Sb, Pb, Te based chemistries), consistent with enhanced SOC effects. Notably, the CGCNN refinement reorders several candidates relative to tabular only screening, demonstrating that structural information contributes non trivial discrimination beyond composition derived trends. Top 20 candidates are reported in Table 6. Top 100 can be seen in Supplementary Table S3. The distribution of inverse design confidence scores (P_{Tabular}, P_{CGCNN}, and P_{Joint}) is shown in Fig. 7a, highlighting the sharpening effect of consensus aggregation at high probability values. The agreement between composition based and structure based predictions is illustrated in Fig. 7b, where correlated yet non identical trends confirm complementary model sensitivity.

Table 6: Predicted probabilities for top 20 candidate materials.

S.No.	Formula	$\mathbf{P}_{\text{Tabular}}$	$\mathbf{P}_{\text{CGCNN}}$	$\mathbf{P}_{\text{Joint}}$
1	$CuAg_2Sb$	0.944045	0.738165	0.861693
2	AgBaBi	0.898341	0.794299	0.856724
3	Cu_2Ag_2Te	0.913987	0.740313	0.844518
4	CuBaBi	0.854535	0.816036	0.839135
5	CuBaPb	0.849410	0.813319	0.834974
6	CdBaPb	0.872729	0.775959	0.834021
7	Ag_3Bi	0.897992	0.733047	0.832014

Continued on next page

S.No.	Formula	$P_{Tabular}$	P_{CGCNN}	P_{Joint}
8	$CuAg_2Bi$	0.889432	0.742566	0.830685
9	Cu_2Ag_2Sn	0.892454	0.735892	0.829829
10	AgBaPb	0.858660	0.786076	0.829627
11	SnBaHg	0.880460	0.749591	0.828113
12	$CuAg_2Pb$	0.879562	0.741106	0.824180
13	NaAuPb	0.908943	0.696785	0.824079
14	Cu_3Te	0.885571	0.728020	0.822551
15	$NiAg_2Te$	0.874564	0.743676	0.822209
16	$MnPd_2Te$	0.865685	0.755023	0.821420
17	Cu_2AsTe	0.877601	0.736717	0.821248
18	Cu_2Pd_2Te	0.870217	0.745028	0.820142
19	SnBaAu	0.863761	0.753858	0.819800
20	Cu_2SbTe	0.870483	0.741363	0.818835

3.4 Physical interpretation of model behavior

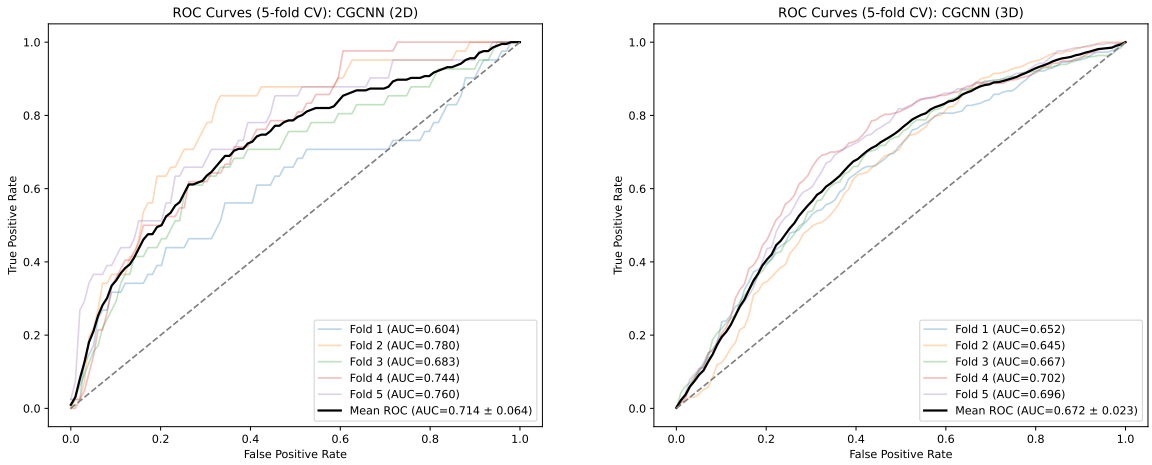
Composition-only tabular models succeed by learning statistical correlations between chemistry and topology, including the prevalence of heavy elements, valence electron counts, and electronegativity contrasts. These correlations are physically meaningful but indirect. GNN, by contrast, operate entirely on crystal geometry, learning coordination patterns, connectivity, and spatial symmetry. Their lower absolute performance reflects the fact that topological order is not purely structural, yet their success demonstrates that structural motifs alone can bias materials toward non-trivial electronic phases. CGCNN achieves the strongest graph-based performance due to its explicit use of interatomic distances. However, its sensitivity to edge feature dimensionality and graph construction highlights a broader challenge: distance-aware message passing is powerful but fragile in large, heterogeneous datasets. This observation motivates future work on more robust, symmetry aware graph architectures, potentially incorporating equivariant formulations or learned radial basis representations.

4 Limitations and Outlook

Several limitations in the study should be acknowledged. Topological labels are inherited from database annotations rather than recomputed for each structure. Spin-orbit coupling, magnetic ordering, and many-body effects are not explicitly modeled. Additionally, current graph models do not incorporate explicit symmetry constraints. The present framework provides a scalable foundation for data driven spintronic materials discovery and can be extended in several important directions. Incorporating crystal structure aware representations and graph based neural networks may further improve predictive accuracy and enable structure resolved inverse design. Integration with targeted first principles calculations or high throughput experiments would allow validation and refinement of top ranked candidates. Future extensions will integrate equivariant GNN, spin-orbit-aware descriptors, and active-learning loops to iteratively refine candidate selection. Future work may also expand the proxy framework to include topology aware descriptors, magnetic anisotropy indicators, and spin orbit coupling effects, enabling exploration of multifunctional materials such as magnetic topological semimetals.

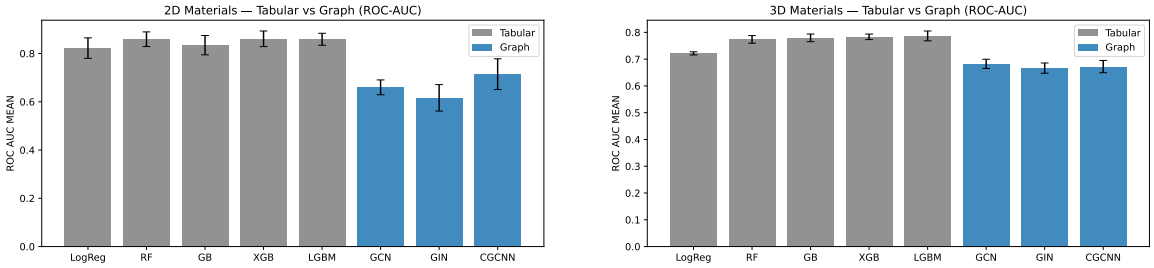
5 Conclusion

In this work, we established a physics-informed and scalable machine learning framework for the identification and inverse discovery of topological materials without explicit computation of wavefunction-based invariants. By systematically integrating electronic proxy descriptors, composition-driven tabular models, and structure-aware graph neural networks, we demonstrated that topological propensity can be learned across multiple levels of physical abstraction, with each representation capturing complementary aspects of the underlying physics. Benchmarking reveals a clear hierarchy in predictive power, where electronic proxies provide an upper bound, composition encodes statistically meaningful trends, and crystal structure alone offers partial yet non-trivial predictive capability. Leveraging this complementary behavior, the proposed inverse discovery pipeline efficiently screens



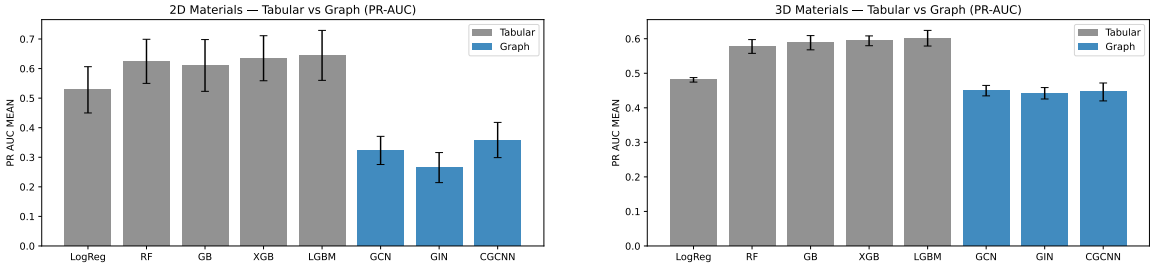
(a) 2D crystal graphs (b) 3D crystal graphs

Fig. 4: ROC curves for the CGCNN model trained on (a) 2D and (b) 3D crystal graphs.



(a) 2D materials (b) 3D materials

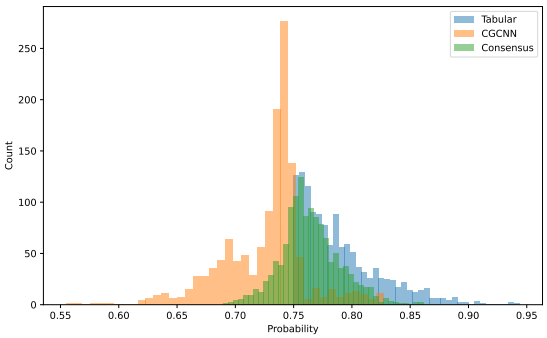
Fig. 5: ROC–AUC comparison between Tabular-B and GNN for (a) 2D and (b) 3D materials.



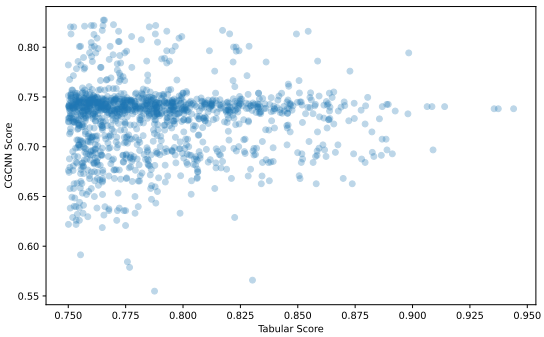
(a) 2D materials (b) 3D materials

Fig. 6: Precision–recall comparison between Tabular-B and GNN for (a) 2D and (b) 3D materials.

large chemical spaces and enables consensus-driven ranking of candidates, culminating in the identification of a focused set of high-confidence materials, including the top 50 candidates for targeted validation. Beyond predictive performance, this work highlights the role of physically grounded descriptors and hybrid learning strategies in bridging accuracy, interpretability, and scalability. The presented framework thus provides a reproducible and extensible pathway for accelerated topological materials discovery, with direct implications for data-driven exploration of quantum materials beyond conventional first-principles workflows.



(a) Distribution of inverse design confidence scores



(b) The agreement between Tabular and CGCNN

Fig. 7: Distribution plots

6 Data Availability

The data and code supporting the findings of this study shall be shared on reasonable request to the corresponding author.

7 Author Contributions

Tasneem U Rehman: Writing - original draft, Methodology, Conceptualization, Software, Formal analysis **Nasreena Yaqoob:**, Conceptualization, Visualization, Formal analysis **Aadil Ahmad Ganaie:** Writing - review & editing, Data Curation, **Ghulam Nabi Dar:** Supervision, Conceptualization.

8 Declaration of conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the author(s) used ChatGPT (OpenAI) and Quillbot in order to improve the language, grammar and readability of the manuscript. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article. No AI Tool was used in the core research work or analysis.

References

[1] Qi, X.-L., Zhang, S.-C.: Topological insulators and superconductors. Rev. Mod. Phys. **83**, 1057–1110 (2011) https://doi.org/10.1103/RevModPhys.83.1057

[2] Hasan, M.Z., Kane, C.L.: Colloquium: Topological insulators. Rev. Mod. Phys. **82**, 3045–3067 (2010) https://doi.org/10.1103/RevModPhys.82.3045

[3] Armitage, N.P., Mele, E.J., Vishwanath, A.: Weyl and dirac semimetals in three-dimensional solids. Rev. Mod. Phys. **90**, 015001 (2018) https://doi.org/10.1103/RevModPhys.90.015001

[4] Mellnik, A.R., Lee, J.S., Richardella, A., Grab, J.L., Mintun, P.J., Fischer, M.H., Vaezi, A., Manchon, A., Kim, E.-A., Samarth, N., Ralph, D.C.: Spin-transfer torque generated by a topological insulator. Nature **511**(7510), 449–451 (2014) https://doi.org/10.1038/nature13534

[5] Nayak, C., Simon, S.H., Stern, A., Freedman, M., Das Sarma, S.: Non-abelian anyons and topological quantum computation. Rev. Mod. Phys. **80**, 1083–1159 (2008) https://doi.org/10.1103/RevModPhys.80.1083

[6] Bradlyn, B., Elcoro, L., Cano, J., Vergniory, M.G., Wang, Z., Felser, C., Aroyo, M.I., Bernevig, B.A.: Topological quantum chemistry. Nature **547**(7663), 298–305 (2017) https://doi.org/10.1038/nature23268

[7] Po, H.C., Vishwanath, A., Watanabe, H.: Symmetry-based indicators of band topology in the 230 space groups. Nature Communications **8**, 50 (2017) https://doi.org/10.1038/s41467-017-00133-2

[8] Zhang, T., Jiang, Y., Song, Z., Huang, H., He, Y., Fang, Z., Weng, H., Fang, C.: Catalogue of topological electronic materials. Nature **566**, 475–479 (2019) https://doi.org/10.1038/s41586-019-0944-6

[9] Vergniory, M.G., Elcoro, L., Felser, C., Regnault, N., Bernevig, B.A., Wang, Z.: A complete catalogue of high-quality topological materials. Nature **566**(7745), 480–485 (2019) https://doi.org/10.1038/s41586-019-0954-4

[10] Ward, L., Agrawal, A., Choudhary, A., Wolverton, C.: Matminer: An open source toolkit for materials data mining. Computational Materials Science **152**, 60–69 (2018) https://doi.org/10.1016/j.commatsci.2018.03.037

[11] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., Persson, K.A.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Materials **1**(1), 011002 (2013) https://doi.org/10.1063/1.4812323

[12] Xie, T., Grossman, J.C.: Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters **120**(14), 145301 (2018) https://doi.org/10.1103/PhysRevLett.120.145301

[13] Choudhary, K., DeCost, B.: Atomistic line graph neural network for improved materials property predictions. npj Computational Materials **7**(1), 1–9 (2021) https://doi.org/10.1038/s41524-021-00670-3

[14] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J.P., Kornbluth, M., Molinari, N., Smidt, T.E., Kozinsky, B.: E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature Communications **13**, 2453 (2022) https://doi.org/10.1038/s41467-022-29939-5

[15] Yan, B., Felser, C.: Topological materials discovery by large-order symmetry indicators. Annual Review of Condensed Matter Physics **8**, 337–354 (2017) https://doi.org/10.1146/annurev-conmatphys-031016-025458

[16] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019)

[17] Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric. arXiv preprint arXiv:1903.02428 (2019)

[18] Vergniory, M.G., Elcoro, L., Felser, C., Bernevig, B.A., Regnault, N.: A complete catalogue of high-quality topological materials. Nature **566**(7745), 480–485 (2019) https://doi.org/10.1038/s41586-019-0954-4

[19] Mounet, N., Gibertini, M., Schwaller, P., Campi, D., Merkys, A., Marrazzo, A., Sohier, T., Castelli, I.E., Cepellotti, A., Pizzi, G., Marzari, N.: Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. Nature Nanotechnology **13**(3), 246–252 (2018) https://doi.org/10.1038/s41565-017-0035-5

[20] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 3rd edn. Wiley, ??? (2013)

[21] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001) https://doi.org/10.1023/A:

1010933404324

[22] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of Statistics **29**(5), 1189–1232 (2001) https://doi.org/10.1214/aos/1013203451

[23] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[24] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems **30** (2017)

[25] Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics **21**, 6 (2020) https://doi.org/10.1186/s12864-019-6413-7

[26] Ward, L., Agrawal, A., Choudhary, A., Wolverton, C.: A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials **2**, 16028 (2016) https://doi.org/10.1038/npjcompumats.2016.28

[27] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., Persson, K.A.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Materials **1**(1), 011002 (2013) https://doi.org/10.1063/1.4812323

[28] Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., Ceder, G.: Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. Computational Materials Science **68**, 314–319 (2013) https://doi.org/10.1016/j.commatsci.2012.10.028

[29] Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (2017)

[30] Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? International Conference on Learning Representations (2019)

[31] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.pdf