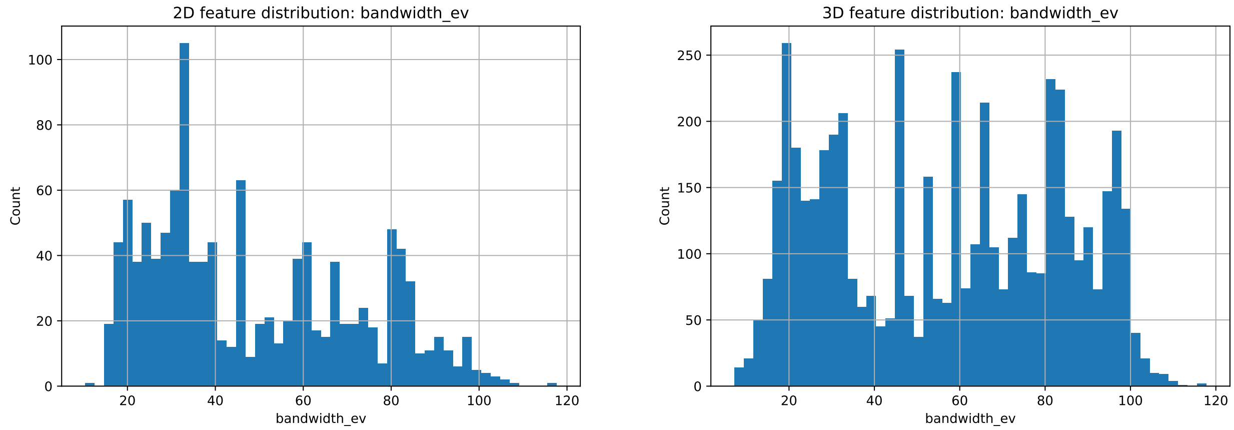
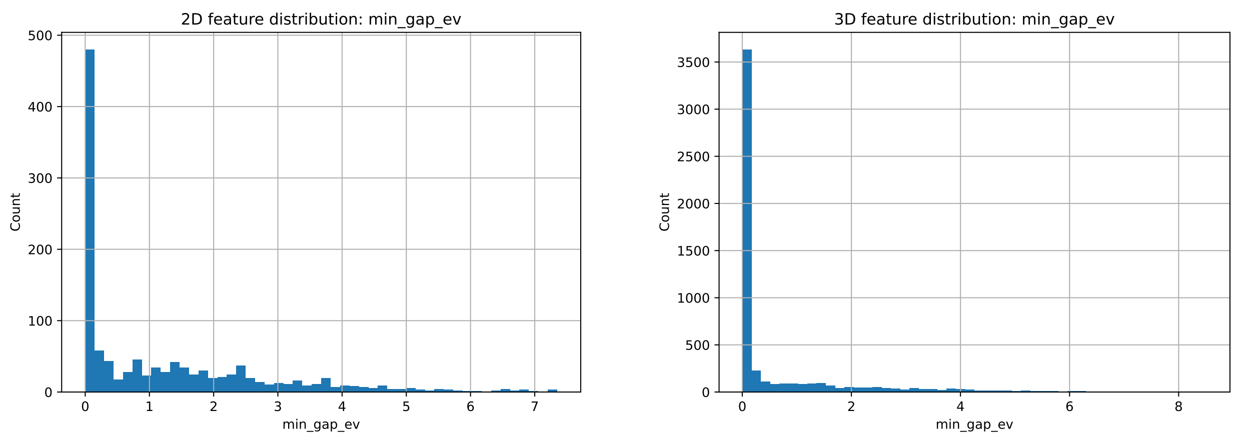
# Supplementary Information

## S1 Dataset Statistics and Symmetry Analysis



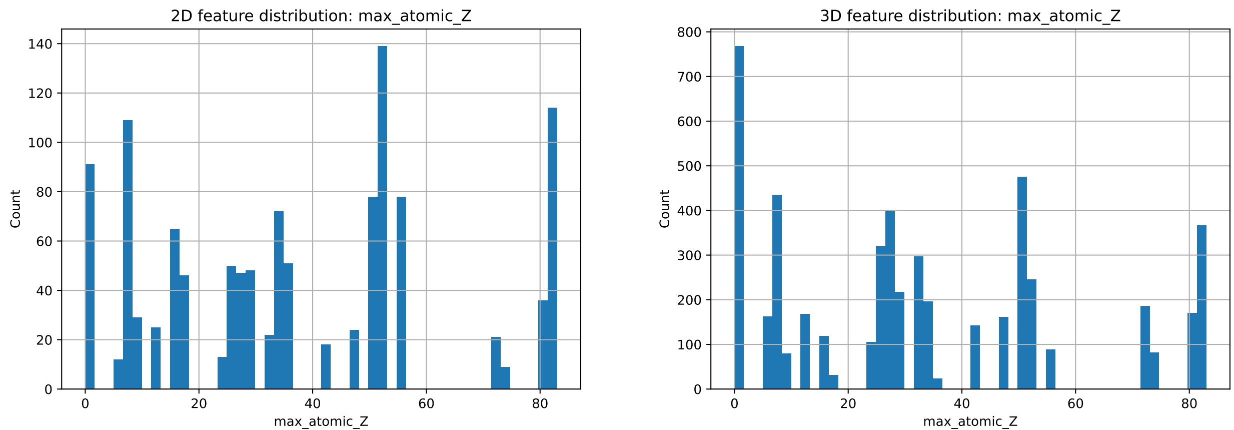

(a) 2D materials (b) 3D materials

Figure S1: Frequency distribution of bandwidth across the full (a) 2D and (b) 3D datasets.




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

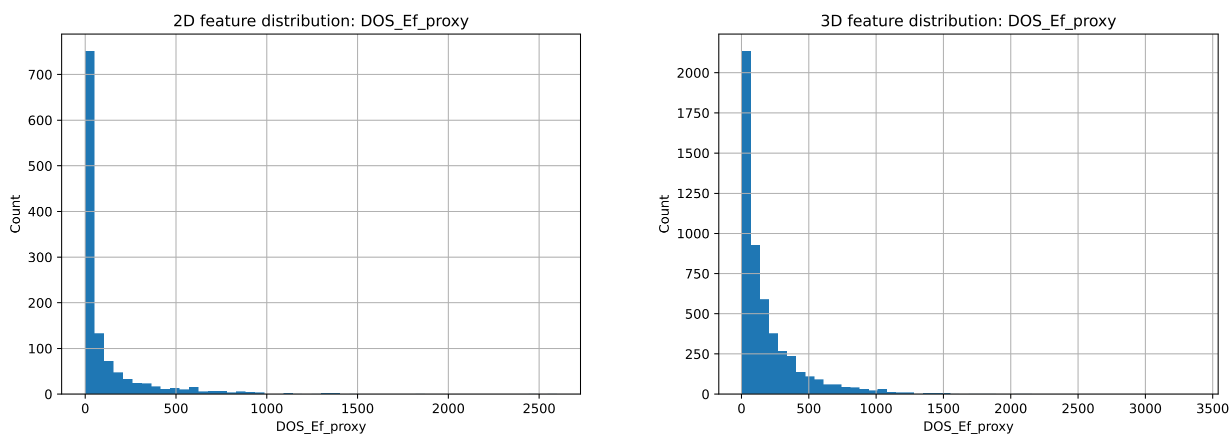
Figure S2: Frequency distribution of minimum bandgap across the full (a) 2D and (b) 3D datasets.




(a) 2D materials

(b) 3D materials

Figure S3: Frequency distribution of maximum atomic number across the full (a) 2D and (b) 3D datasets.

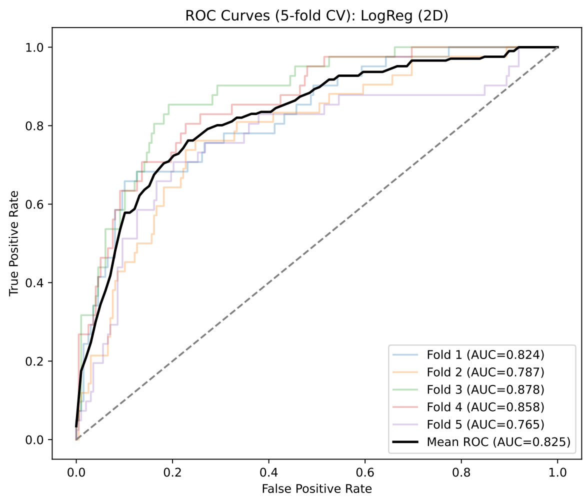



(a) 2D materials

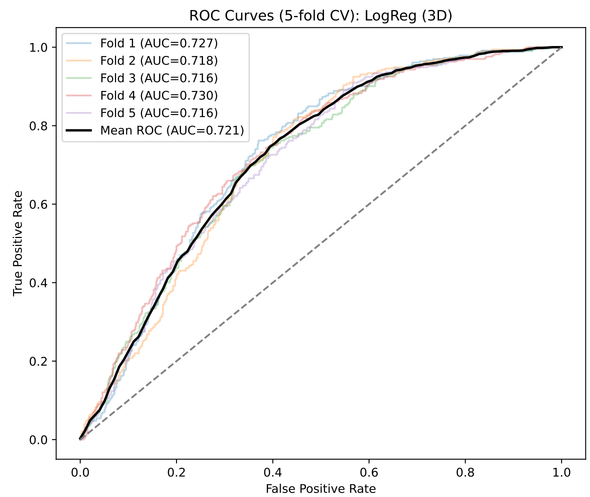
(b) 3D materials

Figure S4: Frequency distribution of DOS proxy across the full (a) 2D and (b) 3D datasets.

# S2 Baseline Tabular Models ROC Curves

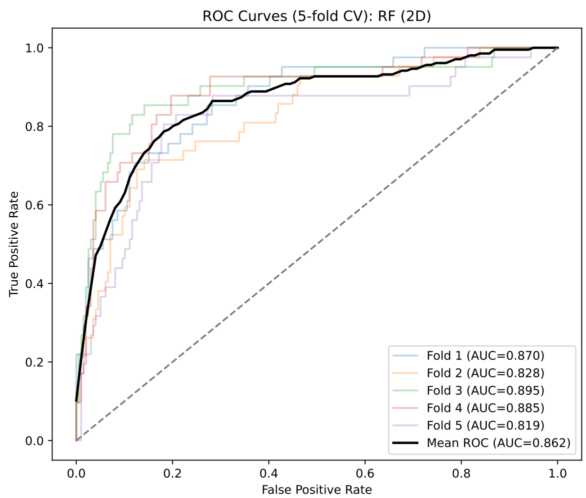



(a) 2D materials

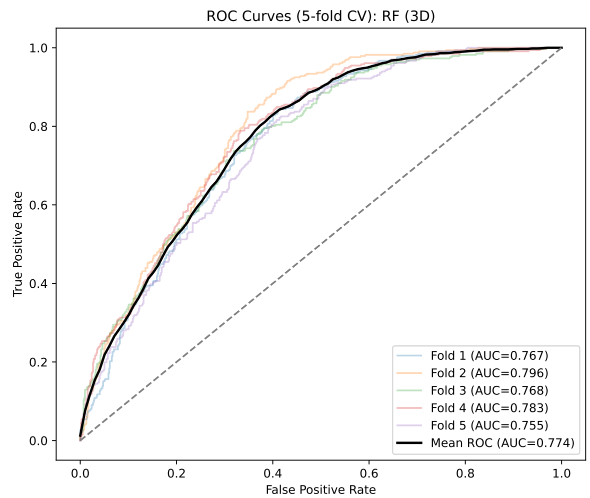



(b) 3D materials

Figure S5: ROC curves for Logistic Regression applied to (a) 2D and (b) 3D materials.

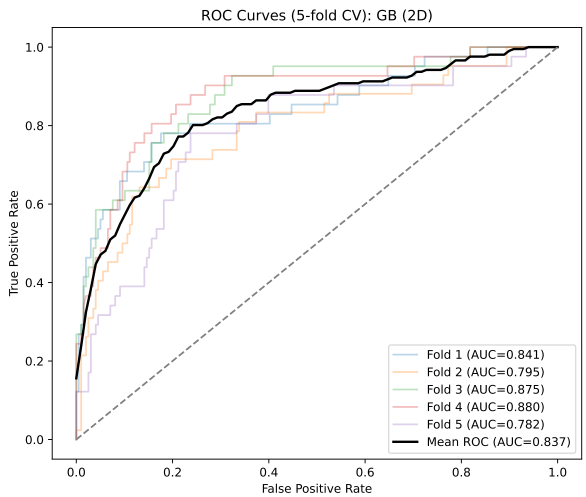



(a) 2D materials

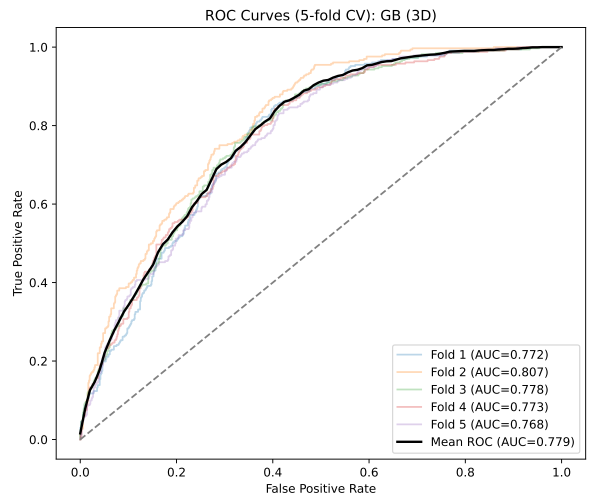



(b) 3D materials

Figure S6: ROC curves for Random Forest applied to (a) 2D and (b) 3D materials.

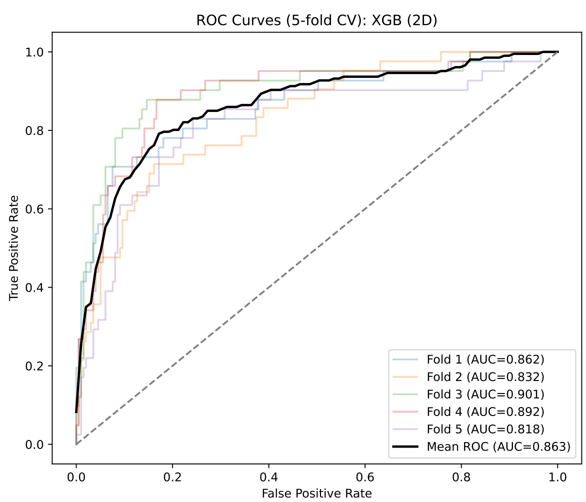



(a) 2D materials

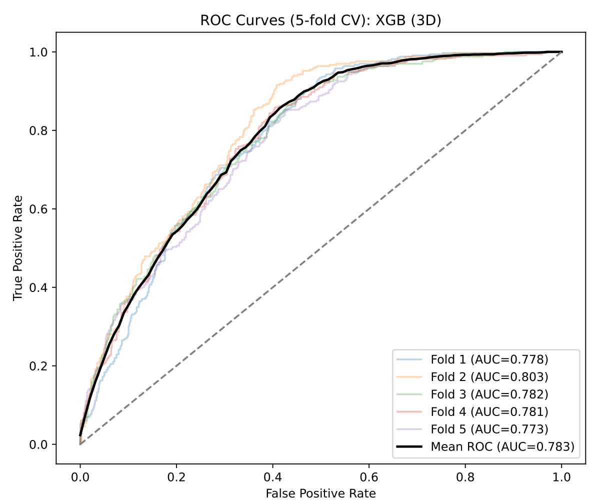



(b) 3D materials

Figure S7: ROC curves for Gradient Boosting applied to (a) 2D and (b) 3D materials.



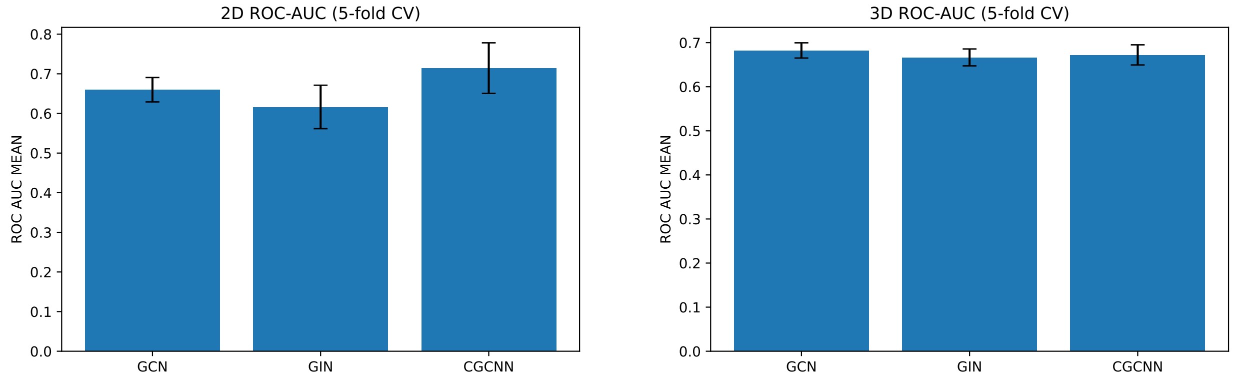

(a) 2D materials




(b) 3D materials

Figure S8: ROC curves for XGBoost applied to (a) 2D and (b) 3D materials.

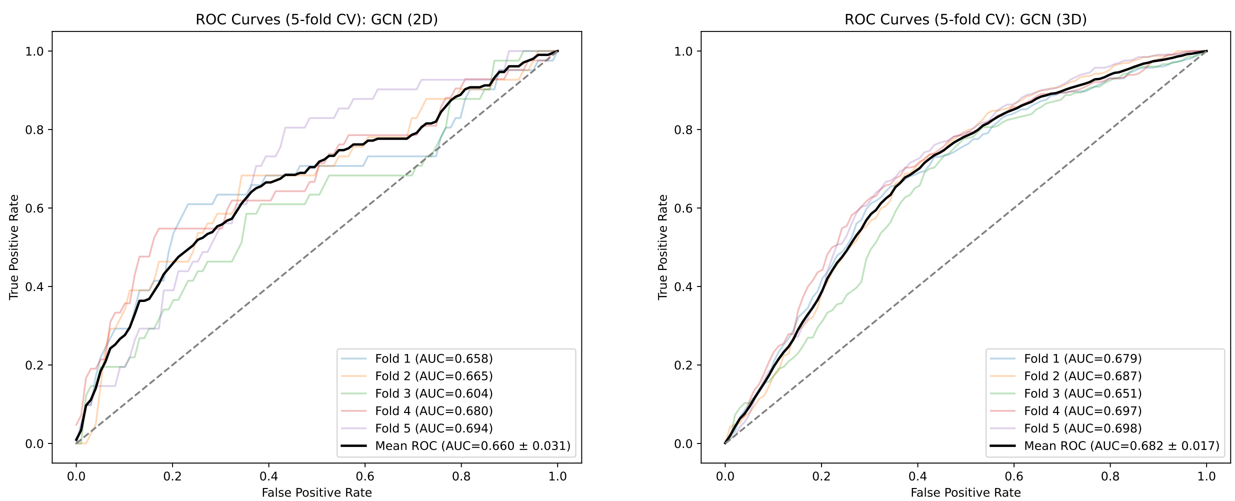
# S3 Graph Neural Network Models Performance




(a) 2D materials

(b) 3D materials

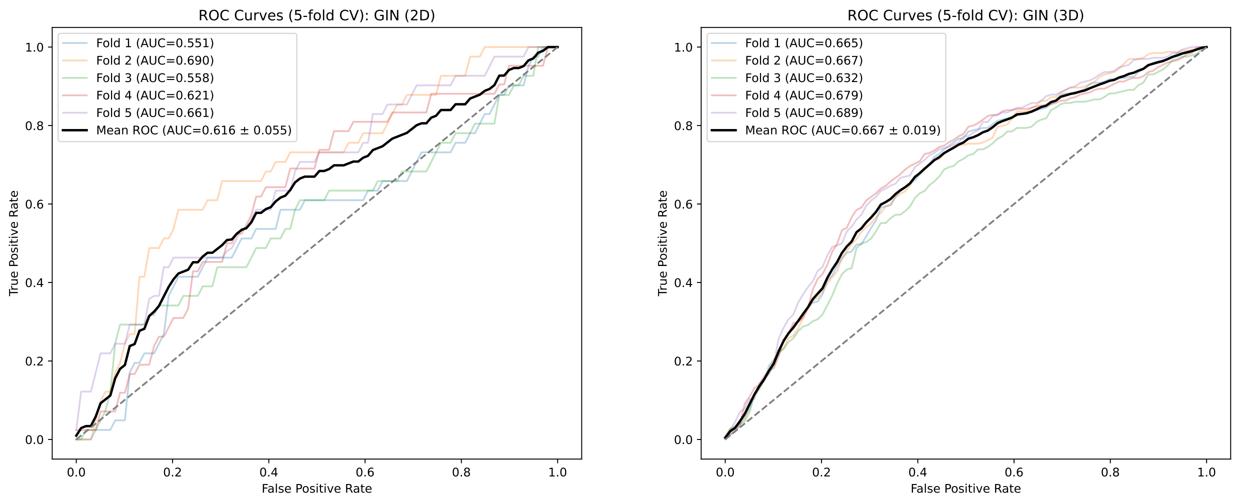
Figure S9: ROC comparison (bar plots) for models trained on (a) 2D and (b) 3D crystal graphs.




(a) 2D materials

(b) 3D materials

Figure S10: ROC curves for the GCN model trained on (a) 2D and (b) 3D crystal graphs.




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

Figure S11: ROC curves for the GIN model trained on (a) 2D and (b) 3D crystal graphs.

# S4 Definition of Inverse Design Scores

We define the probabilistic inverse design scores used to rank candidate materials and describe the spin orbit coupling proxy employed during chemical-space pruning. The framework combines composition and symmetry driven tabular models with structure aware crystal graph neural networks.

## S4.1 Tabular Inverse Probability $P_{\text{Tabular}}$

The tabular inverse probability $P_{\text{Tabular}}$ estimates the likelihood that a material hosts a non trivial topological phase based exclusively on composition- and symmetry-derived descriptors. Separate LightGBM classifiers were trained for two-dimensional (2D) and three dimensional (3D) datasets. All electronic structure-derived quantities (e.g., band gaps, Fermi-level features, explicit topological invariants) were excluded to prevent information leakage.

For a standardized feature vector $\mathbf{x} \in \mathbb{R}^d$, the model learns a nonlinear decision function

$$f_{\text{Tab}}(\mathbf{x}) = \sum_{m=1}^{M} \gamma_m h_m(\mathbf{x}), \tag{1}$$

where $h_m$ are decision trees and $\gamma_m$ are learned weights. The output is mapped to a probability using the logistic sigmoid function:

$$P_{\text{Tabular}} = \sigma(f_{\text{Tab}}(\mathbf{x})) = \frac{1}{1 + e^{-f_{\text{Tab}}(\mathbf{x})}}. \tag{2}$$

This score captures chemically and symmetry driven trends associated with topological band inversion.

## S4.2 Graph Based Inverse Probability $P_{\mathrm{CGCNN}}$

Crystal structures are represented as graphs $G = (V, E)$, where nodes correspond to atoms and edges encode interatomic connections within a fixed radial cutoff. Each atom is described by an 8-dimensional elemental feature vector. A Crystal Graph Convolutional Neural Network (CGCNN) updates node representations through successive message-passing layers, followed by global mean pooling to obtain a crystal level embedding $\mathbf{h}_{\mathrm{crystal}}$. The final logit

$$z_{\mathrm{CGCNN}} = W\mathbf{h}_{\mathrm{crystal}} + b \tag{3}$$

is converted to a probability

$$P_{\mathrm{CGCNN}} = \sigma(z_{\mathrm{CGCNN}})\,. \tag{4}$$

Unlike $P_{\mathrm{Tabular}}$, this score explicitly depends on atomic arrangement and local coordination environment, enabling refinement across crystallographic prototypes.

## S4.3 Joint Inverse Probability $P_{\mathrm{Joint}}$

To combine complementary chemical and structural information, the final ranking score is defined as a fixed weighted fusion:

$$P_{\mathrm{Joint}} = \alpha\, P_{\mathrm{Tabular}} + (1 - \alpha)\, P_{\mathrm{CGCNN}}, \tag{5}$$

with $\alpha = 0.6$ throughout this work.

Materials are ranked according to $P_{\mathrm{Joint}}$ for inverse discovery and downstream validation.

## S4.4 SOC Proxy

To prioritize materials likely to exhibit strong relativistic effects, a spin orbit coupling (SOC) proxy was employed during chemical space reduction. The proxy is defined as

$$\mathrm{SOC}_{\mathrm{proxy}} = \left\langle Z^2 \right\rangle = \frac{1}{N_{\mathrm{el}}} \sum_{i=1}^{N_{\mathrm{el}}} Z_i^2, \tag{6}$$

where $Z_i$ denotes the atomic number of element $i$ in the composition and $N_{\mathrm{el}}$ is the number of distinct constituent elements.

This proxy reflects the leading-order atomic scaling of spin–orbit interaction strength with $Z^2$ and was used exclusively for physics-informed pruning of the hypothetical chemical space prior to model scoring.

# S5 Supplementary Tables

Table S1: List of features used in tabular machine learning models.

| Feature category | Features |
|---|---|
| Electronic structure proxies | bandwidth_ev, min_gap_ev, DOS_Ef_proxy, n_crossings_Ef |
| Chemical composition | num_elements, max_atomic_Z, heavy_element_flag |
| Crystallographic | dim, centered, nat, ne, space_group_number |
| Topological invariants | nu0, nu1, nu2, nu3 |
| Derived descriptors | dispersion_anisotropy, flat_band_ratio |

Table S2: Definitions of performance metrics used to evaluate tabular and graph-based machine learning models.

| Metric | Definition and interpretation |
|---|---|
| ROC-AUC | Area under the receiver operating characteristic (ROC) curve, measuring the probability that a randomly chosen positive (topologically non-trivial) sample is ranked higher than a randomly chosen negative (trivial) sample. Values range from 0.5 (random guessing) to 1.0 (perfect discrimination). |
| PR-AUC | Area under the precision–recall (PR) curve, which emphasizes model performance on the positive class and is particularly informative for imbalanced datasets common in materials discovery tasks. |
| Accuracy | Fraction of correctly classified samples among all samples. While intuitive, accuracy can be misleading for imbalanced datasets and is therefore reported alongside complementary metrics. |
| Precision | Fraction of predicted positive samples that are truly positive, defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$. High precision indicates a low false-positive rate, important for minimizing wasted experimental or computational validation. |
| Recall | Fraction of true positive samples correctly identified, defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. High recall reflects the model's ability to recover known topological materials. |
| F1-score | Harmonic mean of precision and recall, defined as $2(\mathrm{Precision} \times \mathrm{Recall})/(\mathrm{Precision} + \mathrm{Recall})$. This metric balances false positives and false negatives in a single scalar quantity. |
| Matthews Correlation Coefficient (MCC) | Correlation coefficient between observed and predicted classifications, defined as $$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$ MCC ranges from $-1$ (total disagreement) to $+1$ (perfect prediction) and is robust to class imbalance. |
| Pearson correlation ($r$) | Linear correlation coefficient between predicted probabilities and true labels, quantifying the strength of linear association. |
| Spearman correlation ($\rho$) | Rank-based correlation coefficient assessing monotonic agreement between predicted scores and true labels, insensitive to non-linear scaling. |

Table S3: Top 100 inverse candidates ranked by joint probabilities

| **S.No.** | **Formula** | $\mathbf{P}_{\text{Tabular}}$ | $\mathbf{P}_{\text{CGCNN}}$ | $\mathbf{P}_{\text{Joint}}$ |
|---|---|---|---|---|
| 1 | $CuAg_2Sb$ | 0.944045337 | 0.738165268 | 0.861693310 |
| 2 | $Cu_2Ag_2Sb$ | 0.937373228 | 0.738165268 | 0.857690044 |
| 3 | AgBaBi | 0.898340823 | 0.794299030 | 0.856724106 |
| 4 | $Cu_2AgSb$ | 0.935550658 | 0.738165268 | 0.856596502 |
| 5 | $Cu_2Ag_2Te$ | 0.913987057 | 0.740313476 | 0.844517625 |
| 6 | $CuAg_2Te$ | 0.908472776 | 0.740313476 | 0.841209056 |
| 7 | CuAgTe | 0.906361207 | 0.740313476 | 0.839942115 |
| 8 | CuBaBi | 0.854534964 | 0.816035545 | 0.839135197 |
| 9 | CuBaPb | 0.849410067 | 0.813318741 | 0.834973537 |
| 10 | CdBaPb | 0.872728526 | 0.775959128 | 0.834020767 |
| 11 | $Ag_3Bi$ | 0.897991985 | 0.733047280 | 0.832014103 |
| 12 | $CuAg_2Bi$ | 0.889431996 | 0.742565560 | 0.830685421 |
| 13 | $Cu_2Ag_2Bi$ | 0.888769184 | 0.742565560 | 0.830287735 |
| 14 | $Cu_2Ag_2Sn$ | 0.892453591 | 0.735892421 | 0.829829123 |
| 15 | AgBaPb | 0.858660140 | 0.786076146 | 0.829626543 |
| 16 | $Cu_2AgTe$ | 0.886818023 | 0.740313476 | 0.828216204 |
| 17 | SnBaHg | 0.880459882 | 0.749591464 | 0.828112515 |
| 18 | $CuAg_2Pb$ | 0.879561948 | 0.741105962 | 0.824179553 |
| 19 | NaAuPb | 0.908942588 | 0.696784555 | 0.824079375 |
| 20 | $Cu_3Te$ | 0.885571465 | 0.728019723 | 0.822550768 |
| 21 | $NiAg_2Te$ | 0.874563928 | 0.743676102 | 0.822208797 |
| 22 | $MnPd_2Te$ | 0.865684786 | 0.755023050 | 0.821420092 |
| 23 | $Cu_2AsTe$ | 0.877601435 | 0.736717322 | 0.821247790 |
| 24 | $Cu_2Pd_2Te$ | 0.870217420 | 0.745028090 | 0.820141688 |
| 25 | SnBaAu | 0.863761034 | 0.753857899 | 0.819799780 |
| 26 | $Cu_2SbTe$ | 0.870483451 | 0.741362984 | 0.818835264 |
| 27 | $SrAu_2Pb$ | 0.864977481 | 0.749522740 | 0.818795585 |
| 28 | CuSrBi | 0.828777800 | 0.800976300 | 0.817657200 |
| 29 | GeBaPb | 0.820312925 | 0.813439214 | 0.817563441 |
| 30 | YAuPb | 0.868848784 | 0.739802378 | 0.817230222 |
| 31 | ZnBaPb | 0.817227116 | 0.816953588 | 0.817117704 |
| 32 | SrSnAu | 0.850561527 | 0.766468894 | 0.816924474 |
| 33 | $Ag_2SbAu$ | 0.879333066 | 0.722856805 | 0.816742562 |
| 34 | $Cu_2Au_2Bi$ | 0.874640091 | 0.728358257 | 0.816127357 |
| 35 | CaAuPb | 0.888472803 | 0.707581559 | 0.816116306 |
| 36 | SbBaAu | 0.859291250 | 0.750931752 | 0.815947451 |
| 37 | SrAgPb | 0.836212740 | 0.785160172 | 0.815791713 |
| 38 | $CoAg_2Te$ | 0.862474386 | 0.745411640 | 0.815649288 |
| 39 | $Ag_2SnHg$ | 0.882296590 | 0.714881235 | 0.815330448 |
| 40 | $Cu_2RhTe$ | 0.863219343 | 0.743298700 | 0.815251086 |

*Continued on next page*

| S.No. | Formula | $P_{Tabular}$ | $P_{CGCNN}$ | $P_{Joint}$ |
|---|---|---|---|---|
| 41 | $Ca_2Au_2Pb$ | 0.886744822 | 0.707581559 | 0.815079516 |
| 42 | SnBaTl | 0.855584661 | 0.754161096 | 0.815015235 |
| 43 | CuAgSn | 0.867663063 | 0.735892421 | 0.814954806 |
| 44 | $Ni_2Ag_2Sb$ | 0.862844115 | 0.741772521 | 0.814415477 |
| 45 | $Ni_2Ag_2Sn$ | 0.864040544 | 0.739733750 | 0.814317826 |
| 46 | CuSrPb | 0.823116304 | 0.800206649 | 0.813952442 |
| 47 | $Cu_2AuPb$ | 0.872224103 | 0.726058617 | 0.813757909 |
| 48 | $NiAg_2Sb$ | 0.861383792 | 0.741772521 | 0.813539283 |
| 49 | $Cu_2SrPb$ | 0.822015927 | 0.800206649 | 0.813292216 |
| 50 | PdBaPb | 0.824439130 | 0.796418130 | 0.813230730 |
| 51 | SrAuPb | 0.855303128 | 0.749522740 | 0.812990973 |
| 52 | SrSnHg | 0.845869311 | 0.762927753 | 0.812692688 |
| 53 | $Cu_2AgSn$ | 0.863842080 | 0.735892421 | 0.812662217 |
| 54 | $Ni_2Ag_2Bi$ | 0.855900648 | 0.747016281 | 0.812346901 |
| 55 | ZnSrPb | 0.823187145 | 0.796057332 | 0.812335219 |
| 56 | $MnPd_2Sb$ | 0.850708997 | 0.754028347 | 0.812036737 |
| 57 | $Na_2Au_2Pb$ | 0.888869995 | 0.696784555 | 0.812035819 |
| 58 | AgSbBa | 0.846081053 | 0.760775495 | 0.811958829 |
| 59 | $Na_2Au_2Bi$ | 0.891204632 | 0.692852057 | 0.811863602 |
| 60 | $Cu_2AgBi$ | 0.857689678 | 0.742565560 | 0.811640031 |
| 61 | $Cu_2AgPb$ | 0.857405862 | 0.741105962 | 0.810885902 |
| 62 | $Cu_2AuBi$ | 0.863978650 | 0.728358257 | 0.809730493 |
| 63 | $Cu_2HgBi$ | 0.863577618 | 0.727151802 | 0.809007292 |
| 64 | SbTeBi | 0.855351795 | 0.739316285 | 0.808937591 |
| 65 | $Cu_2Hg_2Bi$ | 0.863408461 | 0.727151802 | 0.808905798 |
| 66 | $Cu_2PdTe$ | 0.851479617 | 0.745028090 | 0.808899006 |
| 67 | $CuPd_2Te$ | 0.850981072 | 0.745028090 | 0.808599879 |
| 68 | $AgSbAu_2$ | 0.865696952 | 0.722856805 | 0.808560893 |
| 69 | $Cu_2SnTe$ | 0.854384988 | 0.739577889 | 0.808462149 |
| 70 | MgAuPb | 0.886398600 | 0.691365486 | 0.808385355 |
| 71 | $Ge_2TlBi$ | 0.861351211 | 0.728750455 | 0.808310909 |
| 72 | $Cu_2TeBi$ | 0.851848321 | 0.741852954 | 0.807850174 |
| 73 | $Fe_2Te_2Bi$ | 0.849814866 | 0.744620624 | 0.807737169 |
| 74 | $Ag_3Pb$ | 0.859479545 | 0.730026561 | 0.807698352 |
| 75 | $SrAu_2Bi$ | 0.842955251 | 0.753765124 | 0.807279200 |
| 76 | $Sb_2Bi$ | 0.856651476 | 0.733195613 | 0.807269131 |
| 77 | $Si_2TlBi$ | 0.882474788 | 0.694343844 | 0.807222411 |
| 78 | $SrAg_2Pb$ | 0.821821322 | 0.785160172 | 0.807156862 |
| 79 | $NiAg_2Bi$ | 0.845708254 | 0.747016281 | 0.806231464 |
| 80 | BaAuBi | 0.883152098 | 0.690209982 | 0.805975251 |
| 81 | RhBaPb | 0.811497840 | 0.796464872 | 0.805484653 |
| 82 | $SbBaAu_2$ | 0.841800499 | 0.750931752 | 0.805453000 |

*Continued on next page*

| S.No. | Formula | $\mathbf{P}_{\text{Tabular}}$ | $\mathbf{P}_{\text{CGCNN}}$ | $\mathbf{P}_{\text{Joint}}$ |
|---|---|---|---|---|
| 83 | $CuPdSb_2$ | 0.846941518 | 0.743079972 | 0.805396900 |
| 84 | $Cu_3Bi$ | 0.853398680 | 0.733160871 | 0.805303556 |
| 85 | SrCdPb | 0.827878617 | 0.771408379 | 0.805290522 |
| 86 | CuPdTe | 0.845315719 | 0.745028090 | 0.805200668 |
| 87 | $Ag_2SnAu$ | 0.862774908 | 0.718830547 | 0.805197164 |
| 88 | $Cu_2Rh_2Te$ | 0.846444275 | 0.743298700 | 0.805186045 |
| 89 | $Cu_2Ag_2Pb$ | 0.847902424 | 0.741105962 | 0.805183839 |
| 90 | $Ni_2AgSb$ | 0.846775079 | 0.741772521 | 0.804774056 |
| 91 | BaBi | 0.799467476 | 0.810773909 | 0.803990049 |
| 92 | $NiPd_2Sb$ | 0.841331262 | 0.746786362 | 0.803513302 |
| 93 | $Cu_2As_2Te$ | 0.847512763 | 0.736717322 | 0.803194587 |
| 94 | CuAgSb | 0.846153266 | 0.738165268 | 0.802958067 |
| 95 | $MnAg_2Te$ | 0.838434876 | 0.749329793 | 0.802792843 |
| 96 | $CuSb_2Au$ | 0.845443302 | 0.738432324 | 0.802638911 |
| 97 | $Ni_2SrPb$ | 0.798237870 | 0.809126937 | 0.802593497 |
| 98 | $Cu_2SbBi$ | 0.843910158 | 0.740481603 | 0.802538736 |
| 99 | GeBaBi | 0.792932732 | 0.816115534 | 0.802205853 |
| 100 | $Mg_2Au_2Bi$ | 0.877819810 | 0.687624042 | 0.801741503 |

Table S4: Software packages and computational tools used in forward and inverse discovery workflows.

| Tool / Library | Version / Reference |
|---|---|
| Python | 3.10 |
| NumPy | 1.26.4 |
| pandas | 2.x |
| SciPy | 1.x |
| scikit-learn | 1.x |
| XGBoost | 2.x |
| LightGBM | 4.x |
| CatBoost | 1.x |
| PyTorch | 2.1.2 |
| PyTorch Geometric | 2.7.0 |
| pymatgen | 2024.x |
| matminer | 0.9.x |
| tqdm | 4.x |
| matplotlib | 3.x |
| joblib | 1.x |
| CUDA Toolkit | 11.x (if GPU used) |
| Anaconda | 2023.x |