npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01655-w

Faithful novel machine learning for predicting quantum properties

Gavin Nop[1], Micah Mundy[2], Jonathan D. H. Smith[1] & Durga Paudyal[3] ✉

Machine learning (ML) has accelerated the process of materials classification, particularly with crystal graph neural network (CGNN) architectures. However, advanced deep networks have hitherto proved challenging to build and train for quantum materials classification and property prediction. We show that *faithful representations*, which directly represent crystal structure and symmetry, both refine current ML and effectively implement advanced deep networks to accurately predict these materials and optimize their properties. Our new models reveal the previously hidden power of novel convolutional and pure attentional approaches to represent atomic connectivity and achieve strong performance in predicting topological properties, magnetic properties, and formation energies. With faithful representations, the state-of-the-art CGNN accurately predicts quantum chemistry materials and properties, accelerating the design and discovery and improving the implicit understanding of complex crystal structures and symmetries. On two separate benchmarks, our non-graphical neural networks achieve near parity with the CGNN architecture, making them viable alternatives.

Quantum materials classification and regression occupy a crucial space in the identification of new materials and the optimization of their properties to facilitate novel energy and quantum technology solutions. The primary tools for modern materials exploration, for example density functional theory (DFT), often require days, weeks, or even months to compute properties of complex materials such as topological indices, micromagnetic inputs, and electronic structure parameters[1]. The techniques of machine learning (ML) are becoming the predominant tool for materials research, given the availability of the Materials Project and other online datasets, with an influx of publications on ML designs and applications[2–4]. Many of these methods are highly dependent on the materials properties being examined. They may be difficult to implement and bring to convergence. Thus, there is a clear need for the development of a reliable, fast approach to model and correlate diverse materials properties.

As most foundational ML models require fixed tensor dimensions for input, early uses of ML algorithms for materials research typically hashed properties of the atoms in the primitive cell to produce predictions with random forests, simple multilayer perceptrons, and other techniques[5]. Recent developments build on this with CGNN and convolutional networks to achieve better results[6,7]. ML algorithms are even capable of predicting non-trivial topological indices[8]. Combined approaches with attentional graph layers[9–12], as well as more advanced augmentations to represent the symmetry properties of the material implicitly, have shown great promise[13]. In this paper, basic architectural innovations on ML are compared with the classic graphical architecture given by Xie et al. to demonstrate the validity of alternative architectures, implicitly capturing atomic locality rather than explicitly specifying connectivity with an adjacency matrix[14].

Here we develop four fully general ML algorithms which can predict and categorize arbitrary material properties. The key feature of our approach is the use of *faithful representations* of the underlying materials, representing the crystal structure and symmetry directly. Each model is fully capable of distinguishing any pair of unique materials, side-stepping the representational reduction employed by current models. Our models are tested primarily on the topological data enabled by topological quantum chemistry (topological quantum chemistry (TQC))[15]. Further, all models achieve exemplary performance with purely structural information about the materials involved, without recourse to additional experimental data. Formation energy and magnetic ordering are tested to demonstrate the ability of the models to adapt to arbitrary settings. State-of-the-art is achieved for TQC classification. The crystal convolution neural network (CCNN) acheives state-of-the-art on the important ML benchmark of point and space group classification. The crystal attention neural network (CANN) achieves near state-of-the-art performance on multiple benchmarks, demonstrating an unexpectedly high capability to capture atomic connectivity[16]. The CANN operates without a graphical layer, thereby avoiding input of an adjacency matrix. Large-scale architectures, such as CEGAN and graph attention layers, have achieved state-of-the-art (SOTA) on a number of benchmarks[9–12]. Implementations and pre-trained models are provided in `GitHub`. As the majority of time in ML is spent on dataset

[1]Department of Mathematics, Iowa State University, Ames, IA, USA. [2]Department of Mechanical Engineering, Iowa State University, Ames, IA, USA. [3]Department of Physics and Astronomy, University of Iowa, Iowa City, IA, USA. ✉e-mail: durga.quantum@gmail.com

verification, further tools are provided to automatically extract and purify new materials datasets.

Model performance is sufficiently strong to augment DFT application, acting as an initial filter for further investigation. An additional test of physically interpretable model knowledge is introduced, using the atomic limit concept of TQC to determine the impact of scaling factors on material topology. This development transforms ML from a demonstrative technology in materials science to a tool that is readily available for experimental and theoretical materials researchers. Specifically, our CGNN model generates state-of-the-art predictions for TQC materials and their properties, impacting quantum materials science by enabling accurate and interpretable prediction of properties, accelerating the design and discovery of new materials, and improving our understanding of complex crystal structures and symmetries.

Gadolinium (III) sesquioxide (Gd_2O_3) with space group 164 is taken as the basic example to illustrate both the physics and the algorithmic processes in the paper. This is a lower symmetry phase than the cubic phase of Gd_2O_3, and is referenced as material 20470 in the provided `GitHub` dataset. The trigonal symmetry and smaller number of distinct atoms in the primitive cell are pedagogically and representationally more useful. The relevant formulaic aspects of Gd_2O_3 for ML are the space group, the formation energy, the magnetic classification, and the topological designation, which are 164, -3.723 eV/atom, ferromagnetic, and a split elementary band represented topological insulator, respectively.

Crystalline materials defined by a real space primitive cell were taken as inputs to the ML models. Four characteristics are considered for each crystal: the formation energy per atom, the space group (219 labels), magnetic (non-magnetic, ferromagnetic, ferrimagnetic, and antiferromagnetic), and topological classifications. Topological indices are non-local over the Brillioun zone and are defined in TQC by the following categories and subcategories:

- trivial material (tM), which is a linear combination of elementary band representations (LCEBR);
- topological insulator (TI), labeled as having no linear combination (NLC), or as a split elementary band representation (SEBR), and
- topological semimetal (TSM), labeled as an enforced semimetal (ES), or as an enforced semimetal with Fermi degeneracy (ESFD).

Relative to formation energy and magnetic classification, topological state classification is more complex in terms of mathematical and physical origins[15], involving symmetry-enforced electronic states.

A general program for classification of TI's by symmetry introduced by Zak[17,18] relied on band representations. This program culminated in the enumeration of all possible trivial band representations, resulting in a predicted 2D and 3D 26, 938 topological materials (TM)s via TQC[15,19,20]. The resulting dataset was curated to train an ML model achieving an accuracy of 86% (as compared to the baseline accuracy of 50% by simply marking every material as non-topological)[17,21–24]. To understand the issues underlying ML predictions of these materials, a brief overview of the theory is provided: first formally defining topological insulators, and then providing a framework for understanding TQC.

Three primary categories of the TI concept are distinguished:

TI_e — determined by a topological index directly on a gapped electronic band structure;

TI_b — necessarily possessing a conductive boundary bordering a trivially insulating material such as the vacuum;

TI_x — an insulating material C for which an expansion rC is conductive[25].

An *expansion* is defined within the third category, for a real scalar $r \geq 1$, as a modified crystal rC, where for each atom at position p in the original crystal C, the modified crystal rC has a corresponding atom at rp. Thus, expansion increases the inter-atomic distances in the material. For $r \gg 1$, an expansion rC is regarded as forming an *approximate vacuum*.

The relationships between the categories of TI are displayed in Fig. 1. Under certain circumstances, a TI lying in one category may by implication

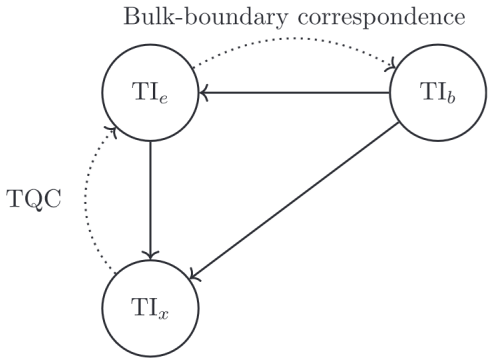

Fig. 1 | Logical relationships between different notions of topological insulators. The dotted arrows represent relationships between different notions of topological insulators that may require additional assumptions to establish, reflecting the fact that, while bulk definitions of crystals have relatively simple translational symmetry, boundaries can be extremely complex.

(⇒) also lie in a second category. For example, the implication $TI_e \Rightarrow TI_b$ forms a class of results known as *bulk-boundary theorems* for specific topological indices[23]. Further, $TI_b \Rightarrow TI_x$ and $TI_e \Rightarrow TI_x$ (Fig. 2). Finally, $TI_b \Rightarrow TI_e$ is a trivial consequence of the fact that materials are specified by their electronic structure. This gives a simple test of the interpretability of the requisite models. In the atomic limit, all models are expected to predict materials as trivial.

The prerequisites for the TQC theory are group theory[26], representation theory[27,28], electronic structure[29], and graph theory[30]. Comprehensive reviews exist[31–33]. We generally follow the notation of the latter. TQC utilizes the notion of an atomic limit with $r \gg 1$ to establish a class of non-topological materials. If a given band structure has symmetry indicators respecting the non-topological band representations, it is trivial. Otherwise, it is topological. The TQC algorithm may be extended to distinguish semimetals as well. However, it distinguishes topology for separated groups of bands, and is not exhaustive[34,35].

Results

Materials embeddings, machine learning architectures, and theoretical implementations

Previous research has demonstrated success with partial data models such as gradient boosted trees (GBT), random forests, k-nearest neighbor classifiers, support vector classifiers, and neural networks[36]. Earlier GBTs were successfully trained by data from[22,37].

Following training of the GBT algorithm for topological data, a subsequent analysis demonstrated that electron counts and space groups were the primary distinguishing decision factors to determine material topology[22]. Model performance was excellent, peaking at 90% for the full GBT model. When the GBT was coupled with ab initio calculations that neglected spin-orbit coupling, accuracy peaked at 92% on the materials with strong confidence in the predicted topological state. As full spin-orbit ab initio calculations enable the direct prediction of material topology, these calculations were not used to supplement the ML models. The primary benefit of using purely structure-based predictions is the encompassing generality, granting an easy method of retraining the models to new situations. Since the original dataset was not accessible, the GBT algorithm without DFT was exactly reconstructed, and applied to the current dataset. On the advanced TQC dataset, it achieved an accuracy of 76% as in Table 3. All algorithms considered are compared to the 76% benchmark, as no additional ab initio calculations were included. In ref. 22, the CGNN was tested, but failed to converge to a reasonable accuracy for topological prediction. Now, it will be seen to have excellent predictive capability.

Four faithful embeddings of the underlying materials are tested. For each embedding, the data format is standardized as follows. Take A to be the set of atoms in the primitive cell. Each atom $a \in A$ is associated with two

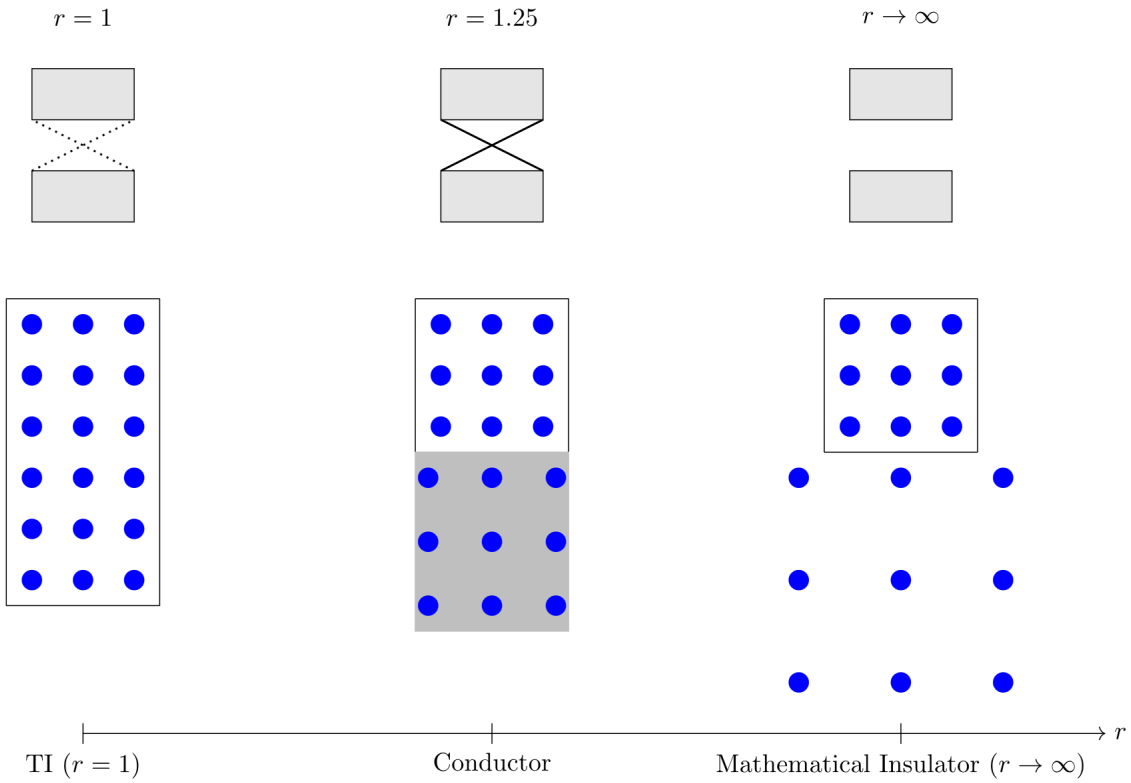

Fig. 2 | Illustration of the atomic limit. For three different value ranges of the scalar r, crystal samples C and rC share a common interface. Above each respective physical picture (TI, conductor, and insulator), a schematic of the corresponding band structure for the rC crystal is presented (valence bands below, conduction bands above). For $r = 1$, where $rC = C$, the conductive boundary surrounds both samples. For r approaching infinity, the expansion rC forms an approximate vacuum, so the conductive border is around C. We locate a scalar r' midway between the supremum of the set of r such that rC is a TI and the infimum of the set of r such that rC is an approximate vacuum, marking this point with a notch in the diagram. If C is a TI_b, consider the boundary states of the material $r'C$. Suppose that $r'C$ was insulating, i.e., the density of states falling within a certain energy range $[E_a, E_b]$ is 0. Then, since expansions of insulating materials are insulating, the border of $r'C$ with both C and the vacuum would be insulating. However, as the bordering C is a TI_b, at least one of the borders must be conductive, and so $r'C$ itself is conductive. Note the informality of this general argument, due to the difficulty of defining a TI_b directly. Nevertheless, for a TI_e, the topological insulator status varies according to a continuous function of the electronic states, and therefore of r. In this case, $r'C$ is necessarily conductive. Thus, an ML approach to TQC classification must be sensitive to r[25].

types of information: the *atomic identifier* v_a and the *atomic position* p_a. Finally, the *global vector* g is a vector containing primitive cell dimensions and symmetries. Different embeddings are considered for each of the input vectors, and tested over all ML frameworks to determine the best representation.

For a classification with n categories, recall that the *one-hot* encoding of the i-th category is $0^{\oplus(i-1)} \oplus 1 \oplus 0^{\oplus(n-i)}$. To enhance generalization over merely using a one-hot embedding of atomic number, the embedding was chosen as $h(r) \oplus h(c \,(\mathrm{mod}\ 2)) \oplus \lfloor c/2 \rfloor$, using the left-step periodic table in Fig. 3 to supply r and c[38]. This allows generalization over the rows and columns of the periodic table with 7 (rows) + 16 (spinless columns) + 1 (spin slot) = 24 positions per atom. Embedding additional atomic properties was tested, but no additional performance gains were found.

The position embedding p_a is network-dependent, but is stored using fractional units relative to the primitive cell basis. There are two major components of the global data vector g. The first gives the primitive cell dimensions using a sinusoidal encoding[39], while the second records the space group with a one-hot embedding. Hyperparameter tuning was used solely for the TQC dataset to demonstrate maximal network performance, and neglected for the remaining tests to demonstrate the ability to immediately generalize.

The full models as described in the methodology are capable of overfitting on any coherent set of training data to an arbitrary extent. Thus, training accuracy is not emphasized. For some tables, previous papers are used as approximate benchmarks for comparison. Since these papers may not use the same dataset, the comparisons are at best indicative.

One implication of the faithfulness of the models is that limits had to be introduced to speed training time. To compare a model to the GBT algorithm, a penalty was assigned to the primitive cells unable to fit in the representation as follows: without knowledge of the underlying input variables, the best predictor of an element in the validation set V is a single label, p, for each element of V. So, this optimal element was used as a default prediction for when materials were too large to use with the ML models. Note that p is extracted from the training set to prevent data contamination. For classification, p is the most common label. For regression, if the loss is root mean squared error (RMSE) or mean absolute error (MAE), then the p which optimizes each of these measures of model error is the mean and median, respectively. This gives a well-defined methodology to compare dissimilar models over an underlying dataset. It also gives a simple baseline model for comparisons, as represented in Tables 1–3.

General quantum materials property predictions

The material representations are sufficient to determine the symmetry group. Thus, as a first test of the global power of the ML algorithms, 151,000 materials were taken from the Materials Project and ICSD datasets[40,41]. The POSCAR file format[1] was used as input to supply the ML with atomic types and positions, and the primitive cell basis. The target variable for each material was the space group classification. The symmetry of a material is derived easily from the POSCAR description using structural geometry. Thus, symmetry group classification is perfectly accurate, enabling a verification of the models' practical implementation.

Two primary implementations for the symmetry groups were tested: the one-hot encodings of the space and point groups. The space groups

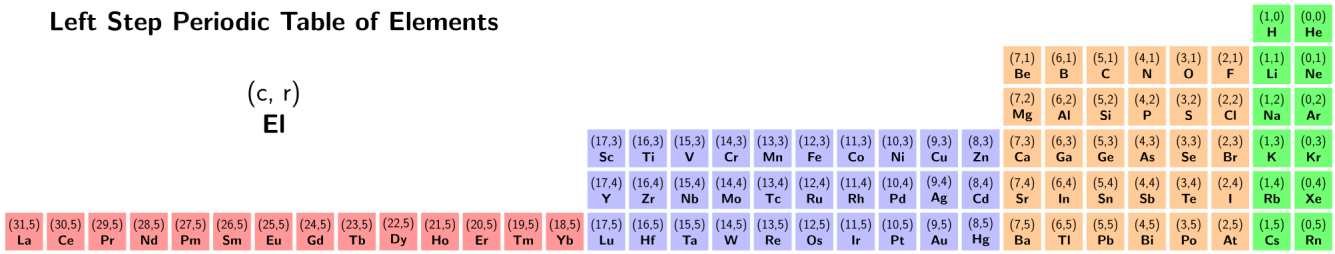

Fig. 3 | The left-step periodic table of elements. Each atom in the periodic table is labeled with the column and row annotations used for ML.

Table 1 | Comparison of ML models for space group and point group one-hot classification problems

Model	Point Group Accuracy	Space Group Accuracy
Baseline	0.15	0.10
naive Neural Network (NNN)	0.14	0.14
CGNN	0.78	0.78
CCNN	0.81	0.79
CANN	0.73	0.62

CCNN achieved the highest performance, indicating an alternative way forward for structural classification.

Table 2 | Comparison of ML models for the categorization problem

Model	Formation energy MAE	Magnetic classification
Baseline comparison	1.0	0.54
NNN	0.26	0.75
CGNN	0.10	0.84
CCNN	0.19	0.79
CANN	0.11	0.81

For MAE, a smaller number is better.

Table 3 | Comparison of ML models for TQC

Model	Basic accuracy	Advanced accuracy
baseline	0.49	0.49
GBT baseline[22]	0.81	0.76
NNN	0.72	0.65
CGNN	0.83	0.80
CCNN	0.76	0.71
CANN	0.80	0.75

Importantly, the CANN and CCNN architectures perform well in comparison to an optimized CGNN architecture. In tests without internal skip connections, these alternative architectures exceeded CGNN performance.

comprise 230 labels, and the point groups comprise 32 labels. As can be seen from Table 1, ML performance was low compared to analytic techniques. Indeed, this is a known weakness of ML, and is an ongoing area of research in the ML community. The CCNN algorithm did manage to capture the majority of the space symmetries, indicating that spatial relationships are handled best with this direct approach, by comparison with the other three methods.

Formation energy per atom and the magnetic classification were both indexed from ref. 40 for 151,000 materials. Natural errors were expected, due to temperature dependence for experimental results and limited DFT accuracy. Performance on the magnetic dataset was strong, compared to the 81% accuracy on a smaller dataset[42]. This illustrates the universality of model design as implemented for formation energy (Table 2). However, as expected, classification model performance for regression tasks without modification was weak, which will be improved in the subsequent development.

Topological classification

Three primary sources were used to train the model. The first dataset[21] contains a comprehensive list of topological indices for materials. The material information was extracted in the form of POSCAR files from the two largest materials datasets available[40,41]. For each material, two sets of topological labels were extracted: T_s, a simplified labeling, and T_r, a refinement of T_s. Here, T_s consists of three labels: LCEBR, TI, and SM, while T_r consists of five augmented labels: LCEBR, NLC, SEBR, ES, and ESFD. There are 75, 000 materials with this labeling.

Two requirements were placed on the data. As a first criterion, primitive cells were required to have fewer than 60 atoms. The second criterion arose from the issue that materials are often duplicated by stoichiometric label and symmetry group with minor variations in the POSCAR file. Thus, in cases where the topological labels agreed, the entries were condensed. In cases where there was a discrepancy in the topological data, the material was simply eliminated from the dataset, due to the high probability of a mistaken calculation, or of unusual ambient factors such as temperature and pressure. As an example of this type of situation, 39 tuples of materials were merely minor distortions of each other, distinguished in the ICSD database, but identified in the Materials Project. After the filtration process, 36, 580 materials remained, with 455 datapoints removed. The original dataset evidently contained thousands of duplicate materials. It is worth noting that an ML process based on the original dataset would score artificially higher due to cross-contamination between the training and testing datasets. The topological composition of the dataset is ES, TI, SM, NLC, ESFD with 0.10, 0.27, 0.07, 0.07, and 0.49 as fractions of the whole dataset, respectively.

The majority of model experiments were performed on the TQC dataset. This enabled the diagnosis of specific model issues based on accuracy. Unless otherwise stated, all comments specifically pertain to the full 5 TQC classifications. At the 49% threshold, the model does not necessarily have information transfer between the input and the output, since the most common material type, non-topological, comprises 49% of the dataset. An additional apparent plateau occurs near the 75% accuracy range, after which training is diminished. The CGNN model notably exceeds this threshold. Models were trained on the whole available dataset 20–60 times (epochs) to achieve maximal accuracy on the testing set. All four tested models exhibited an initial fast growth, then an apparent plateau that lasted for approximately one epoch before a more subtle long-term increase in accuracy became apparent. To account for dataset differences, an alternative GBT algorithm was trained for Table 2, based exactly on the specification provided in ref. 22 to compare approaches directly. All the models are either comparable to the GBT baseline, or exceed it, as seen from Table 3.

The optimized implementation for each network is provided in `GitHub` with notes on optimization. Additional correlation effects were

Fig. 4 | Illustration of the CANN model. Global data is simply appended to each token atom, and successive layers of attention and simply 3-layer feed-forward networks are applied in succession.

examined, showing weak correspondences between formation energy, magnetic classification, and topological effects in the supplementary material. Ensembles were created to test systemic model errors. Material misclassification was found to happen most frequently for less common elements, especially Pt. Materials with multiple symmetries corresponding to the same stoichiometric formula were more frequently misclassified, unless the topological label was identical for all symmetry phases.

Due to the vast differences in model architecture, ensemble approaches offer a method of enhancing model predictions. As none of the models achieved perfect performance on the testing set, cases where all the models failed to categorize a material's topological classification properly may be taken as an indication of two potential situations:

- The material is accurately represented by DFT, but is misclassified by the neural network (NN)'s due to violating their internal heuristics;
- The material itself is miscatalogued due to a deficiency in the DFT computation of the band structure.

As an extension of the model classification, a filtration process is performed. Since the four model archetypes (NNN, CANN, CCNN, CGNN) are capable of achieving greater than 95% accuracy on the training dataset, all four models are trained over several epochs to 95% accuracy on the entire dataset. If the mistakes among the models are uncorrelated, the misclassifications will be uncorrelated, as $36,580(0.05)^4 \sim 0$. Any deviation from this scenario demonstrates an interdependence between the models, and allows a model-agnostic method of diagnosing similarities between sources. There were 54 such misclassified materials. Of those materials, $CeIn_2Ni_9$, Fe_2SnU_2, B_4Fe (space group 58), and $InNi_4Tm$ are positively identified as being topological and likely misclassified due to an insufficient DFT calculations. Additionally, 1:3 and 1:5 compounds are frequent in the misclassifications, corresponding to compounds $PtNi_3$, $MoPt_3$, $PdFe_3$, $HfPd_3$, $CrNi_3$, $AlCu_3$, $HgTi_3$ and $HoCu_5$, $GdZn_5$, $EuAg_5$, $CePt_5$, $ThNi_5$, $CeNi_5$.

Discussion

This work presents significant advances in the application of ML to predict, classify, and optimize the properties of quantum crystalline materials, including topological properties, magnetic properties, formation energies, and symmetry groups. By adopting faithful representations, with their direct connection to crystal structure and symmetry, we have enhanced both current graphical ML networks and advanced deep networks. The strong performance of the CANN and CCNN networks in parallel with the CGNN network on a variety of crucial quantum materials prediction problems demonstrates the predictive power of novel convolutional and pure attentional approaches with intrinsically mapped atomic connectivity. In these models, the full representation of the crystal diagnoses difficult-to-predict materials with potentially novel quantum properties and physics. Additionally, the relative strengths and weaknesses of each model are cataloged for practical use and impact. Specifically, our enhanced CGNN generates state-of-the-art predictions for TQC materials and their properties, while the CCNN surpasses the CGNN on the task of crystalline symmetry reconstruction, improving our understanding of complex crystal structures and symmetries.

The tools and models developed here are indexed online for public use and the simple development of new avenues for quantum materials prediction. All models presented were trained within hours and are capable of extremely rapid prediction relative to both DFT and composite model designs, such as CEGAN[12] and GATGNN[16] for initial materials exploration. The automated methods for data preprocessing, along with the full implementations and pretrained models provided on `GitHub`, can be efficiently applied both to novel and pre-existing datasets. This enables the rapid classification and correlation of all crystalline materials, including quantum, magnetic, semiconducting, and topological properties.

Methods

Contemporary naive neural network

This approach employs a fully-connected feedforward neural network. To classify materials, the NN maps a material's properties to a one-hot encoding of its classification. Common to all materials in the datasets explored, there were fewer than 6 of each type of atom in A. Thus, the atoms were partitioned by type into at most 6 subsets and ordered from most common to least common as $A_1, A_2, \cdots, A_6 \subseteq A$, respectively. Each subset A_i has a corresponding maximum size n_i, and, as all $a \in A_i$ share the same v_a, the common atomic vector may be designated v_i. To account for when $|A_i| < n_i$, the empty position $p_\emptyset$ is set to $0^{\oplus|p_a|}$, the 0-vector in the same vector space as p_a. Then, input to the NN is organized into bins as $b_i = v_i \oplus \bigoplus_{a \in A_i} p_a^{\oplus|A_i|} \bigoplus p_\emptyset^{\oplus(n_i-|A_i|)}$, ensuring a fixed-size bin $b_i \in \mathbb{R}^{|v_i|+n_i}$ and therefore a constant-size input tensor for the NN. Finally, all bins are concatenated as $g \oplus \bigoplus_{i=1}^{6} b_i$.

Current crystal graph neural networks

CGNNs are instances of convolutional graph neural networks applied to solid state materials[14]. Information is embedded in each part of the graph. Here the global vector g is considered as a vector separate from the graph. Each node is associated to v_a, and each edge $e = (a, b)$ has the information $v_e = (\overrightarrow{p_a - p_b}, |p_a - p_b|)$. During the graphical passes, the shape of each vector associated with the edges, vertices, and global data is maintained, allowing skip connections. In order to increase the descriptive capacity of the network, v_a, v_e and g are first embedded into the graph using networks NN_e^a, NN_e^e, NN_e^g to larger embedded vectors v_e^v, v_e^e, v_e^g. The final categorization is read from the last components of v_e^g. Thus, v_e^g is at least the sum of the sizes of g and the label vector. This follows the work of Xie et al.[14] with the use of deep skip layers internal to each of the NN_e^a, NN_e^e, NN_e^g layers. Additionally, connectivity is determined by including atoms within a specified distance, not taking the nearest h atoms, resulting in increased training accuracy with a variable number of edges.

Novel crystal attention neural network

While graphical attention layers incorporate an adjacency matrix, we now demonstrate the effectiveness of pure attentional layers for materials property prediction. This eliminates the hyperparameter choices that would have been incurred by the adjacency matrix. A CANN is attention applied to encoded atoms. This generalizes deep set networks, which were previously found to exhibit extremely poor inference on materials datasets. Attention layers (notated as $MultiHead(Q, K, V)$ for query, key, and value matrices, respectively) frequently operate on ordered structures[43]. However, attention naturally treats inputs as elements of a set. The equational status of the network is described by the input $Z = \bigoplus_{a \in A}(p_a \oplus v_a \oplus g)$ supplied to alternating layers of feed-forward networks and attentional layers, with skip connections past each attentional layer. An additional architectural modification based on the commonly known *set transformer* framework was tested to increase training speed with similar results[44]: (rFF) as $rFF(SAB(rFF(SAB(rFF(SAB(Z))))))$ with skip connections between every layer except the last, as illustrated in Fig. 4.

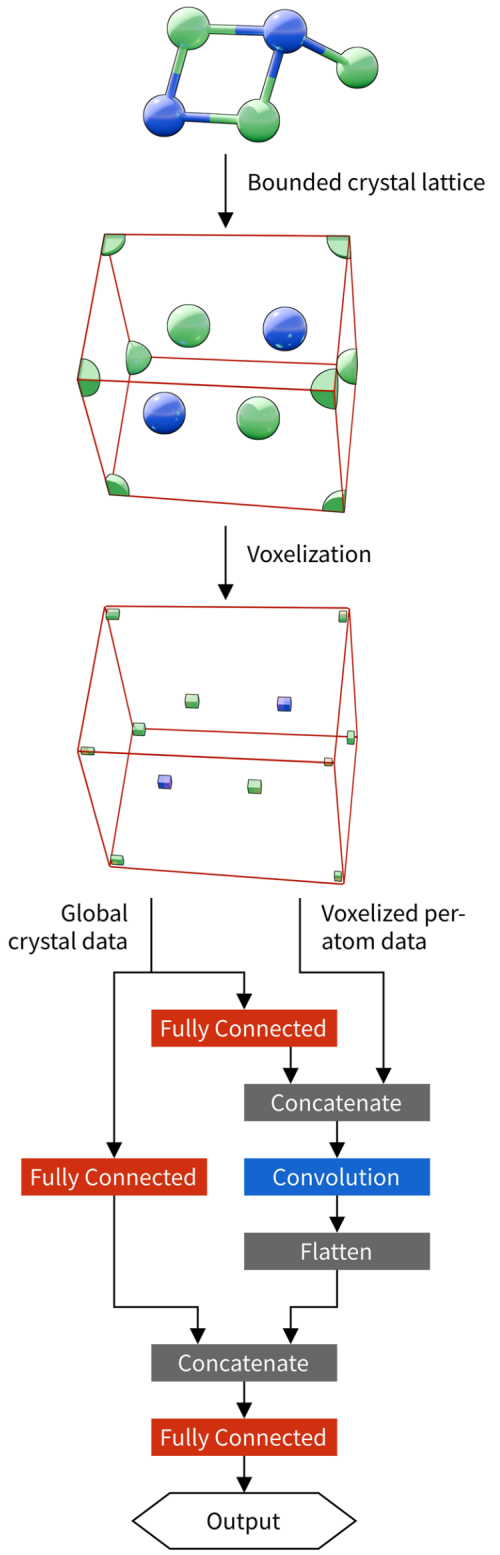

Fig. 5 | The CCNN architecture design flow. A small cubic region surrounding one molecule of the crystal (Gd_2O_3 as an example) is converted into an antialiased voxel lattice. Each voxel encodes a user-configurable representation of an atom whose center is less than one voxel unit away from the voxel's center. Classification is performed by augmenting a fully connected network (red) with a series of convolution layers (blue) that process per-voxel atomic embeddings.

This architecture allows for modeling pairwise and higher-order interactions among elements in the input set, while maintaining permutation invariance. If the set transformer architecture is used, the computational complexity of the attention layers reduces from $O(n^2)$ to $O(nm)$, where m is the number of inducing points, allowing the model to scale to large input sets while maintaining full connectivity. For non-topological classification, this method performed on par with full attention. However, full attention was necessary for the topological dataset. This demonstrates stronger performance for a non-graphical network design than previously expected[16].

Innovative crystal convolutional neural network

The final network examined is the CCNN. This network uses a spatial representation of the atoms[45]. CCNNs are instances of convolutional neural networks (CNN's) applied to solid state materials. Convolutional networks have been used extensively in both voxel and video domains, exploiting spatial and spatio-temporal uniformity by applying a kernel to a 2-, 3- or 4-dimensional representation.

As visualized in Fig. 5, the tensorial embedding for the network is $N^3 \times (|v_{atom}| + 1 + v_g)$ dimensional. The first three indices of the tensor are spatial indices, with the N^3 cube corresponding to the $[0, 1)^3$ space consisting of the atoms' positions relative to the primitive cell spanning vectors. Note that this explicitly violates spatial isotropy. However, network performance was improved compared to isotropy-respecting models. We note that the isotropic expansion and compression were in fact considered early on, and it was found that breaking the symmetry on the input representation level, while maintaining the faithful material representation in the global symmetry, gave the strongest performance. We therefore speculate that this improved performance is due to the easier correlation of symmetry with the voxelization in our approach. The addition of v_g corresponds to generating an $N^3 \times v_g$ tensor directly from the global features via a multiperceptron network as $v_g = NN'(g)$. Tests demonstrated that concatenating $N^3 \times g$ to the voxel crystal cell was both computationally expensive, and failed to perform. To embed the atoms in the first $|v_{atom}| + 1$ spots in the tensor, the atoms from the crystal are represented relative to the bounds of the 3D tensor using the relative coordinates in the crystal cell. Anti-aliasing is used to encode the atomic representations v_a with a filling term directly into the voxel mesh[46].

Data availability

All research data is available with instructions at the GitHub repository at https://github.com/gnnop/Faithful-novel-machine-learning-for-predicting-quantum-properties.

Received: 13 January 2025; Accepted: 14 May 2025;
Published online: 26 July 2025

References

1. Hafner, J. Ab-initio simulations of materials using VASP: Density-functional theory and beyond. *J. Comput. Chem.* **29**, 2044–2078 (2008).
2. Chan, C., Sun, M. & Huang, B. Application of machine learning for advanced material prediction and design. *Eco. Mat.* **4**, e12194 (2022).
3. Wei, J. et al. Machine learning in materials science. *Info Mat.* **1**, 338–358 (2019).
4. Liu, Y., Zhao, T., Ju, W. & Shi, S. Materials discovery and design using machine learning. *J. Materiomics* **3**, 159–177 (2017).
5. Schmidt, J., Marques, M., Botti, S. & Marques, M. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**, 83 (2019).
6. Xie, T. & Grossman, J. Crystal graph convolutional neural networks for accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2017).

7. Zheng, X., Zheng, P. & Zhang, R. Machine learning material properties from the periodic table using convolutional neural networks. *J. Chem. Sci.* **9**, 8426–8432 (2018).
8. Sun, N., Yi, J., Zhang, P., Shen, H. & Zhai, H. Deep learning topological invariants of band insulators. *Phys. Rev. B* **98**, 085402 (2018).
9. Louis, S.-Y. et al. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* **22**, 18141–18148 (2020).
10. Veličković, P. et al. Graph attention networks. *ICLR conference paper arXiv:1710.10903* (2017).
11. Liao, Y.-L., Wood, B., Das, A. & Smidt, T. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations https://arxiv.org/abs/2306.12059 2306.12059 (2024).
12. Banik, S. et al. Cegann: Crystal edge graph attention neural network for multiscale classification of materials environment. *npj Comput. Mater.* **9**, 23 (2023).
13. Rasul, A. et al. A machine learning based classifier for topological quantum materials. *Sci. Rep.* **14**, 31564 (2024).
14. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
15. Vergniory, M. et al. A complete catalogue of high-quality topological materials. *Nature* **566**, 480–485 (2019).
16. Fung, V., Zhang, J., Juarez, E. & Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* **7**, 84 (2021).
17. Moore, J. The birth of topological insulators. *Nature* **464**, 194–8 (2010).
18. Zak, J. Band representations and symmetry types of bands in solids. *Phys. Rev.* **23**, 2824 (1981).
19. Slager, R., Mesaros, A., Juričić, V. & Zaanen, J. The space group classification of topological band-insulators. *Nat. Phys.* **9**, 98–102 (2013).
20. Po, H., Vishwanath, A. & Watanabe, H. Symmetry-based indicators of band topology in the 230 space groups. *Nat. Comm.* **8**, 50 (2017).
21. Topological materials database. https://www.topologicalquantumchemistry.com/. Accessed: 2023-02-18.
22. Claussen, N., Bernevig, B. A. & Regnault, N. Detection of topological materials with machine learning. *Phys. Rev. B* **101**, 245117 (2020).
23. Asbóth, J., Oroszlány, L. & Pályi, A. *A Short Course on Topological Insulators* Vol. 919 of *LNP* (Springer, 2016).
24. Chiu, C., Teo, J., Schnyder, A. & Ryu, S. Classification of topological quantum matter with symmetries. *Rev. Mod. Phys.* **88**, 035005 (2016).
25. Bradlyn, B. et al. Topological quantum chemistry. *Nature* **547**, 298–305 (2017).
26. Rotman, J. *An Introduction to the Theory of Groups* Vol. 148 (Springer Science & Business Media, 2012).
27. Fulton, W. & Harris, J. *Representation Theory: A First Course* Vol. 129 (Springer Science & Business Media, 2013).
28. Paxton, A. et al. An introduction to the tight binding approximation–implementation by diagonalisation. *NIC Ser.* **42**, 145–176 (2009).
29. Kittel, C. & McEuen, P. *Introduction to Solid State Physics* (John Wiley & Sons, 2018).
30. Bollobás, B. *Modern Graph Theory* Vol. 184 (Springer Science & Business Media, 2013).
31. Cano, J. & Bradlyn, B. Band representations and topological quantum chemistry. *Annu. Rev. Condens. Matter Phys.* **12**, 225–246 (2021).
32. Cano, J. et al. Building blocks of topological quantum chemistry: Elementary band representations. *Phys. Rev. B* **97**, 035139 (2018).
33. Van Mechelen, T., Bharadwaj, S., Jacob, Z. & Slager, R. Optical n-insulators: Topological obstructions to optical Wannier functions in the atomistic susceptibility tensor. *Phys. Rev. Res.* **4**, 023011 (2022).
34. Bouhon, A., Bzdušek, T. & Slager, R. Geometric approach to fragile topology beyond symmetry indicators. *Phys. Rev. B* **102**, 115135 (2020).
35. Lange, G., Bouhon, A. & Slager, R. Subdimensional topologies, indicators and higher order boundary effects. *Phys. Rev. B* **103**, 195145 (2021).
36. Boateng, E., Otoo, J. & Abaye, D. Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review. *JDAIP* **8**, 341–357 (2020).
37. Myles, A., Feudale, R., Liu, Y., Woody, N. & Brown, S. An introduction to decision tree modeling. *J. Chemom.* **18**, 275–285 (2004).
38. Scerri, E. Various forms of the periodic table including the left-step table, the regularization of atomic number triads and first-member anomalies. *ChemTexts* **8**, 1–13 (2022).
39. Tancik, M. et al. Fourier features let networks learn high frequency functions in low dimensional domains. *Adv. Neural Inf. Process Syst.* **33**, 7537–7547 (2020).
40. The materials project. https://materialsproject.org/. Accessed: 2023-02-18.
41. National Institute of Standards and Technology. Nist inorganic crystal structure database. NIST Standard Reference Database Number 3. https://doi.org/10.18434/M32147, (Accessed: 2023-02-18).
42. Merker, H. et al. Machine learning magnetism classifiers from atomic coordinates. *iScience* **25**, 105192 (2022).
43. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process Syst.* **30**, 5999–6009 (2017).
44. Lee, J. et al. Set transformer: A framework for attention-based permutation-invariant neural networks. *ICML* **36**, 3744–3753 (2019).
45. Davariashtiyani, A. & Kadkhodaei, S. Formation energy prediction of crystalline compounds using deep convolutional network learning on voxel image representation. *Commun. Mater*. **4**, 105 (2023).
46. Zhang, R. et al. Making convolutional networks shift-invariant again. *PLMR* **97**, 7324–7334 (2019).

Acknowledgements

This work was supported as part of the Center for Energy Efficient Magnonics, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award number DE-AC02-76SF00515.

Author contributions

Conceptualization & Project Administration: G.N., J.D.H.S. and D.P. Investigation and methodology: All authors Supervision: D.P. and J.D.H.S. Writing—original draft: G.N. and D.P. Writing—review & editing: All authors. Resources and funding acquisition: D.P. These author contributions are defined according to the CRediT contributor roles taxonomy.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Durga Paudyal.