

# Crystal Hypergraph Convolutional Neural Networks

Alexander J. Heilman, Weiyi Gong, and Qimin Yan

February 28, 2025

[twocolumn]

## Abstract

Graph representations of solid state materials that encode only interatomic-distance information lack geometrical resolution, resulting in degenerate representations that may map distinct structures to equivalent graphs. Here we propose a hypergraph representation scheme for materials that allows for the association of higher-order geometrical information with hyperedges. Hyperedges generalize edges to connected sets of more than two nodes, and may be used to represent triplets and local environments of atoms in materials. This generalization of edges requires a different approach in graph convolution, three of which are developed in this work. These crystal hypergraph convolutional networks are trained based on various property prediction tasks for a vast set of solid-state materials in the Materials Project and MatBench databases. Results presented here focus on the improved performance of models based on both pair-wise edges and local environment hyperedges. These results demonstrate that hypergraphs are an effective and efficient method for incorporating geometrical information in material representations.

## 1 Introduction

Machine learning has proven to be a computationally cost-effective and powerful predictive tool in the screening of large sets of material systems for certain material properties<sup>1–4</sup>. Some of the most effective state-of-the-art models applied to invariant target predictions represent material systems as graphs<sup>5–10</sup>. These graphs encode physical properties in feature vectors associated with graph components, and update or ‘learn’ these features with a trained graph neural network or message passing network<sup>11</sup>.

One problem with such graphical representations, however, is the lack of representation of higher-order geometrical structure, since the constructed crystal graphs can only include pair-wise descriptors. This may make it difficult or even impossible for models to distinguish between compositionally similar but structurally distinct systems with unique material properties<sup>12</sup>. Other works have approached this problem by including higher-order geometrical features such as overlapping bonds’ angles<sup>12–14</sup>. However, these approaches come with an increase in the total number of messages exchanged through convolution that is quadratic with respect to  $\bar{N}_{\text{edges}}$ , the average number of edges per atom.

Here, we propose the concept of *crystal hypergraphs* to address this lack of geometrical information in the more restrictive graph representations. In a crystal hypergraph, we may define larger (than strictly pair-wise) hyperedges that correspond to higher-order geometrical structures of material systems explicitly, such as triplets of neighboring atoms, or coordination polyhedra/motifs<sup>15–20</sup>. These different structures then may have different coordinate invariant features associated with them, such as

angles and local order parameters<sup>21–23</sup>, respectively. Note that in that regard, crystal hypergraphs are naturally heterogeneous in their hyperedges, since there are different feature sets for different types of hyperedges.

Of course, the definition of a more general hypergraph representation requires the generalization of the message passing framework mentioned above. Here, we propose three possible approaches to such a generalization that handle the now-variable size of hyperedges. In a certain sense, these allow for the learning of a certain type of ‘cluster-correlation expansion’<sup>24–26</sup> by the model, where clusters of interest correspond to the hyperedges defined.

As a proof of concept, we propose and implement a crystal hypergraph convolutional model (CHGCNN) that incorporates invariant geometric features for bonds, triplets, and motifs in crystals as hyperedge features. This provides us a unique opportunity to demonstrate the importance of different order structures for these different material property prediction tasks. Namely, we compare the performance of models based on atom, bond, and triplet information against those incorporating atom, bond, and motif information (i.e. first shell hyperedges) on various predictive tasks with varying data sizes.

Results presented here indicate that first-shell (motif) hyperedges may be sufficient, if not more informative, than triplet hyperedges for many common predictive tasks. This comes at a substantially lower computational cost, in terms of the total number of messages exchanged through graph convolution.

The structure of this work is as follows. We give a brief overview of crystal graph construction and message passing networks. A motivating representation problem is then identified with our definitions and the concept of crystal hypergraphs is introduced, with a particular focus on different types of hyperedges and their corresponding feature sets. Three generalized message passing frameworks are then considered, and a specific model architecture is presented. Finally, this specific architecture is used on various datasets to evaluate the performance of different sets of hyperedge types.

## 2 Background

### 2.1 Crystal Graphs

A common representation of crystalline systems in machine learning is via graphs (that is, collections of nodes and binary connections between them). We may define a crystal graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$  as a set of vertices  $v_i \in \mathcal{V}$ , corresponding to each atom  $i$ , and edges  $e_{ij} \in \mathcal{E}$ , where edges are determined by some physical criteria. Physical information is then associated with the objects in these graphs by way of feature vectors. These are vectors with components describing the physical characteristics of their corresponding graph component, and which may be further ‘learned’ or updated through a graph neural network.

A commonly applied criterion for the formation of edges between atoms is a combination of a maximum distance cutoff  $r_{max}$  and a maximum number of neighbors for each node  $N_{max}$ . That

is, for each atom, edges are constructed between its node and its  $\leq N_{max}$ -th closest neighbors in the crystalline structure within a shell of radius  $r_{max}$ .

The nodes' feature vectors encode the atomic information of the sites they describe. Two usual techniques include: explicitly engineered feature vectors (as in<sup>6</sup>); and the learning of encodings for atomic sites based only on their atomic number (as in<sup>7</sup>), beginning with some random initialization. Edge features are often derived exclusively from their distance.

Crystal graphs are usually constructed solely for use in graph convolutional neural networks. Perhaps the most general framework in which we may define graph convolution is the message passing framework, defined by Gilmore<sup>11</sup>. A message passing network updates nodes based on 'messages' generated by the features of, and passed through, neighboring nodes (that is, nodes sharing an edge).

However, the construction above lacks higher-order geometrical information within the crystals. That is, distinct local geometrical environments of atoms (motifs) may be mapped degenerately to the same crystal graph. As a simple example of the low resolution manifest in crystal graphs, consider two atomic systems below: one with a local cubic symmetry, and another with a square anti-prism local environment; but both with the same bonding atoms. As demonstrated in Figure 1 both structures would map to the same crystal graph, but could be easily distinguished with an additional descriptor describing the local geometry of each central atom.

Alternatively, one could include atomic position in the node features or a vector direction in edge features. However, this generally requires a unique treatment of such coordinate-system dependent information through convolution if the output is to maintain *invariance* with respect to changes in the coordinate system. As such, often only coordinate system invariant features are included in crystal graph representations, such as distance and atomic properties.

## 3 Methods

### 3.1 Crystal Hypergraph Construction

The method proposed here reduces the intrinsic limitations of invariant crystal graph features by allowing the explicit incorporation of higher-order geometrical information in the form of hyperedges, which can be used to directly represent these higher-order structures.

A crystal hypergraph  $\mathcal{H} = \{\mathcal{V}, \mathcal{H}\}$  is a collection of nodes  $v_i \in \mathcal{V}$  and hyperedges  $h_j \in \mathcal{H}$  (containing an arbitrary number of nodes), where the hyperedges are most generally heterogeneous. That is, we may wish to describe different types of hyperedges (e.g. bonds, triplets, and motifs) in the same hypergraph. These objects then have associated feature vectors encoding relevant physical information, which we also refer to as  $v$  and  $h$ .

For the purpose of modeling material systems, we need to identify what different order structures are most important in their representation. Of course, atomic and bond level information is particularly important. However, higher-order structures may also be of interest, such as triplets of atoms and local environments of atoms, which we refer to as motifs in crystals.

Each of the aforementioned structures also has a natural set of distinct, coordinate-system invariant features that may be associated with them. At the triplet level (where two bonds share some common node), there is always a corresponding angle. While at the motif level, order parameters<sup>21,22</sup> or continuous symmetry

measures<sup>27,28</sup> may be used to describe 3-dimensional coordination environments quantitatively. These different order structures may all be represented in a single crystal hypergraph.

Below, we discuss the generation of, and association of features with, all of the above-mentioned structures in crystalline solids.

#### 3.1.1 Bond Edges

Bonds, or pair-wise atomic connections, are determined in the same manner as in a crystal graph. In the results below, we choose edges from a maximum number of neighbors  $N_{max} = 12$  found within a shell of radius  $r_{max} = 6$ .

#### 3.1.2 Triplet Hyperedges

Triplet hyperedges are then formed from the set of bonds. For each set of bonds connected by one node, a triplet hyperedge is formed. The feature of these triplet hyperedges is also a Gaussian expansion, though now of the angle formed by the unit vectors of the two bonds<sup>13</sup>. Triplet hyperedges give us a way to incorporate some angular resolution into our representation scheme in a coordinate-system-invariant manner. For a node with  $N$  bonds then, there will be  $N(N - 1)/2$  triplets. Thus, the price we pay for complete angular resolution of any two bonds is a quadratic increase in the number of hyperedges.

#### 3.1.3 Motif Hyperedges

Motif determination may be achieved by a wide range of functions, and is akin to an algorithmic determination of coordination number<sup>29</sup>. Here, we use a modified Voronoi algorithm with a cut-off radius implemented as CrystalNN in pymatgen. Note this is a much stricter algorithm than that used to determine edges and triplets, since the motifs features depend heavily on the selected first-shell.

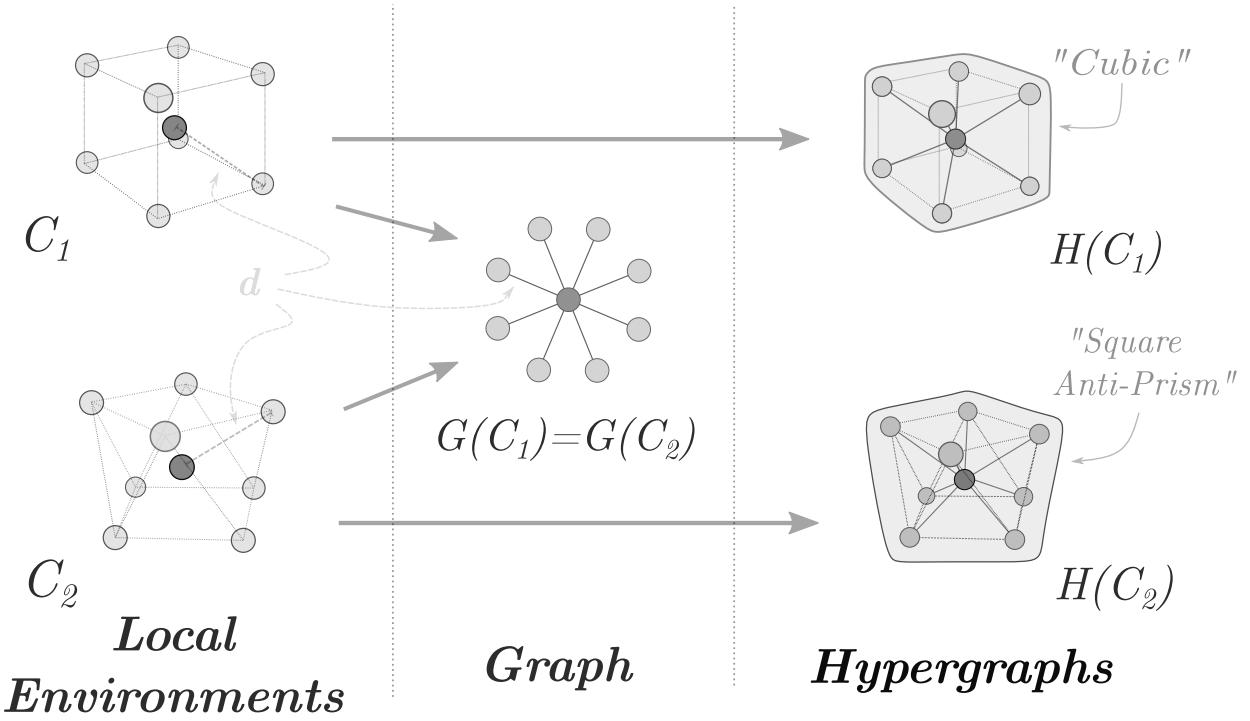
The features of these motifs are a concatenation of Zimmerman's 34 local structure order parameters<sup>21,22</sup>, and continuous symmetry measures<sup>27</sup> (e.g. 'distance to a perfect shape') for 59 common coordination environments. In essence, both are just sets of quantitative measures designed to describe 3 dimensional physical shape. Motifs give us a way to describe the local geometry of sites in material systems with much fewer hyperedges. Since each node will contribute one motif hyperedge, for a crystal with  $N_{nodes}$  nodes, we just have  $N_{nodes}$  motifs.

## 4 Crystal Hypergraph Convolution

We now must consider a message passing framework analogous to Gilmore, *et al*<sup>11</sup> but applying to hypergraph structures. That is, we now have:

$$\begin{aligned} m_v^{t+1} &= \sum_{h_j \in \mathcal{H}(v)} M_t(n_v^t, h_j^t, \{n_w^t | n_w \in h_j\}), \\ n_v^{t+1} &= U_t(n_v^t, m_v^{t+1}), \\ \hat{y} &= R(\{n_v^T\}), \end{aligned}$$

so that each node is still updated according to some layer-wise update function  $U_t$ , aggregating messages  $m_v^{t+1}$  formed from origin node features, hyperedge features  $h_j$ , and hyperedge neighborhood features  $n_w \in h_j$ . This update occurs node-wise and then after  $T$  layers, some readout function  $R$  is used to output the corresponding predicted value  $\hat{y}$ , which utilizes the set of learned node features.



**Figure 1** An example of two distinct geometries that are mapped to the same distance-based crystal graph. With inclusion of a first-shell feature vector encoding local geometry however, these structures are mapped to two distinct crystal hypergraphs. Note these are two possible coordination environments in oxides, determined statistically in<sup>17</sup>.

The biggest difference here is that we now need a message forming function  $M_t$  that accounts for a set of node features  $\{n_w^t | n_w \in h_j\}$  which may vary in size between different hyperedges (even of the same type). This stands in opposition to the case of regular edges, where we are assured a fixed size of two nodes per edge.

One approach would be to fix the dimensionality of each type of hyperedge, or have a different convolutional operator for each different size hyperedge (as is effectively the approach taken with line graph networks<sup>30</sup>). Here, however, we wish to maintain generality in edge size so we need not fix hyperedge sizes for hyperedge type, since structures of different sizes may be described by similar metrics. For example, we may wish that motifs resembling polyhedra with different numbers of vertices are described by common sets of features.

Of course, there should be different message and update functions for each different order structure (bonds, triplets, motifs, etc.) with different features. This is accounted for by treating the data as a heterogeneous graph, with different hyperedge types. Below, we consider three strategies that allow us to apply our convolutional operator to hyperedges of arbitrary size.

#### 4.1 Three Possible Approaches to Hyperedge Convolution

Three general approaches for message passing that account for this multi-order nature have been considered in this work: 1. the construction of a hyperedge relatives graph, upon which regular graph convolution may be applied; 2. total exchange hyperedge message convolution, which completely generalizes the CGCNN<sup>6</sup> and ALIGNN<sup>13</sup> models to hypergraphs; and 3. neighborhood aggregation, which balances performance of the former approach by forming a single neighborhood feature for each hyperedge.

Each approach has a different computational cost in terms of the total number of messages, along with a potentially different practical definition of a hypergraph. These considerations are presented below, with a specific convolutional structure and empirical results on common test datasets presented after.

##### 4.1.1 Hyperedge Relatives Graph

We may define a dual graph  $\mathcal{D}(h)$  to a hypergraph  $h$  to be a graph in which nodes represent the hyperedges of the hypergraph, and connections represent the overlap of respective hyperedge neighborhoods. In the case of a crystal hypergraph with heterogeneous hyperedges, this dual graph is a graph with heterogeneous nodes. We term this heterogeneous dual graph of a crystal hypergraph the relatives graph for simplicity. Atomic features may be included in this framework by adding a singleton hyperedge for each node.

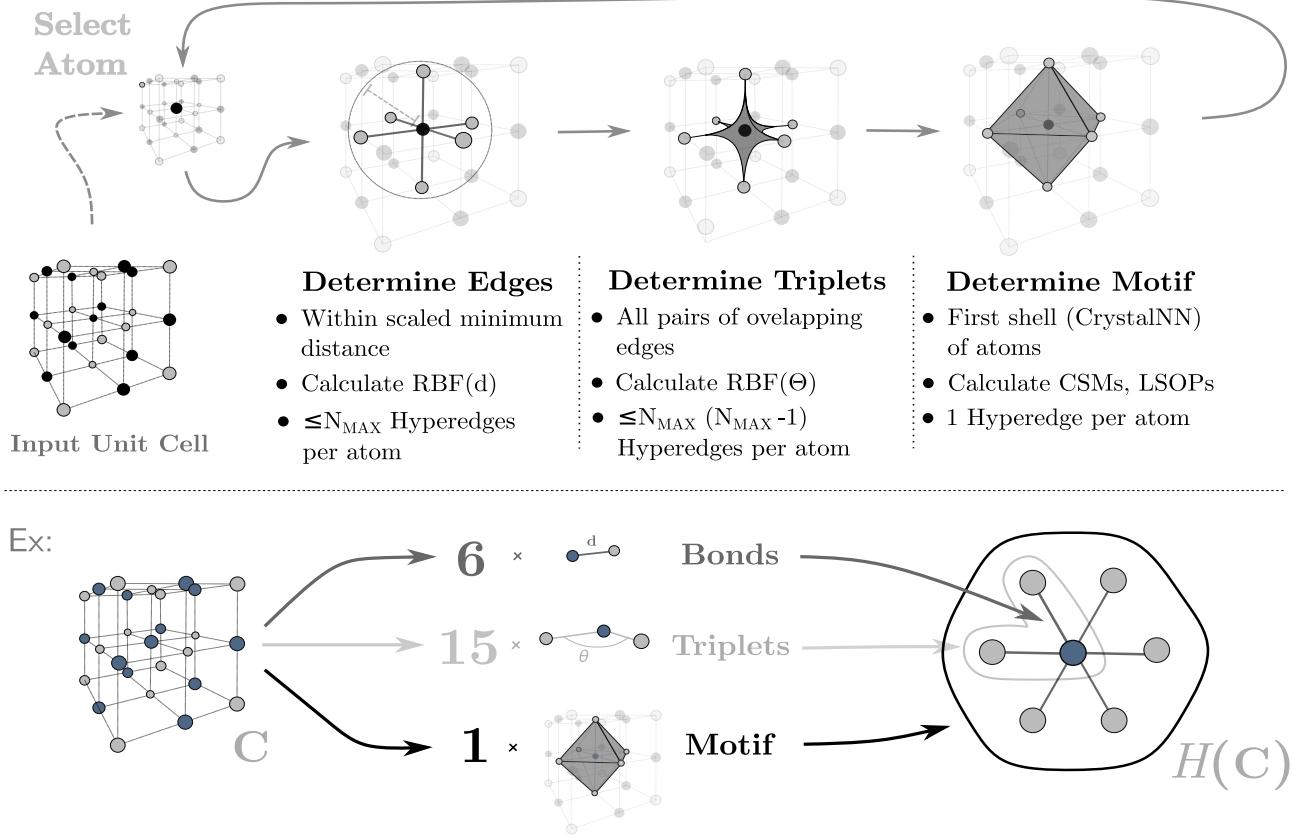
The definition of the relatives graph allows us to perform the usual methods of graph convolution on hyperedge features. Such an approach also allows us to define our relatives graph as we would a graph, with just a standard edge index.

However, this approach lacks the interaction of neighboring features in convolution via the connecting hyperedge. That is, without a clear definition of the edge attribute, messages are generally of the form below:

$$m_v^t = \sum_{h_j \in \mathcal{N}(v)} M_t(n_v^t, h_j^t)$$

in which we simply discard the neighborhood of other node features contained in the hyperedge.

Computationally, this approach has a total number of messages that scales linearly with average hyperedge size, since each hyperedge only contributes one message to each node it contains. Accounting only for node-hyperedge connections in a relatives graph



**Figure 2** Typical construction loop for a crystal hypergraph. First, pair-wise bonds/edges are determined, then triplets are derived from overlapping pairs of bonds, and finally motifs are determined as first-shells of neighbors by some (generally more restrictive) criteria. Features for each and upper bounds on numbers of hyperedges for each type are also listed.

derived from a hypergraph with  $m$  hyperedges of average order  $n$ , the total number of messages per convolution will scale as  $\mathcal{O}(nm)$ .

#### 4.1.2 Total Exchange Message Passing

Of course, we may wish to incorporate the neighboring features of some representation via their connecting hyperedge. This may be accomplished by simply forming a message for every pair of connected representations along with their connecting hyperedge's representation.

$$m_v^t = \sum_{h_j \in \mathcal{H}(v)} \sum_{n_w \in h_j} M_t(n_v^t, h_j^t, n_w^t),$$

Here, though, we have introduced a new summation which may drastically increase the number of messages for larger hyperedges. In this scheme, if each hyperedge contains an average of  $n$  nodes and there are  $m$  hyperedges, the total number of messages exchanged per node-wise convolution will scale as  $\mathcal{O}(n^2m)$ .

#### 4.1.3 Neighborhood Aggregation

Since the number of messages will scale tremendously with larger hyperedges in the framework described above, we may seek a way to incorporate the neighborhood of features of a hyperedge into a single message. In this case, we may essentially form a 'neighborhood feature' representative of all contained nodes of a hyperedge.

Typical aggregation methods may be used and trained to perform this neighborhood feature generation. Here then, we use the message functions of the form:

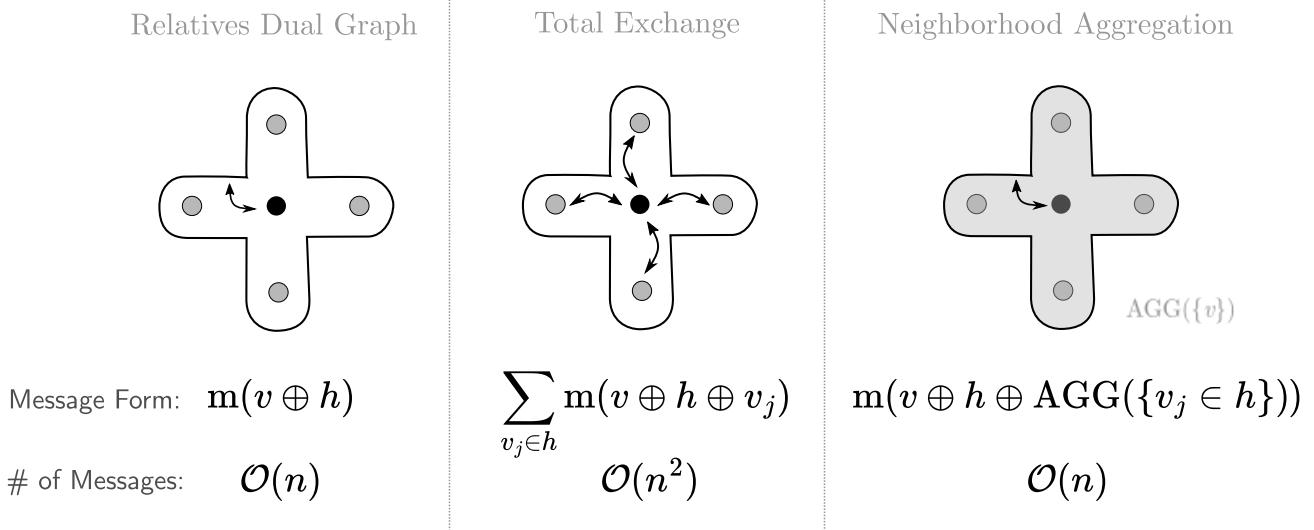
$$m_v^{t+1} = \sum_{h_j \in \mathcal{H}(v)} M_t(n_v^t, h_j^t, \text{AGG}(\{n_w^t | n_w \in h_j\}))$$

This results in a set of node-wise messages that scales linearly with the average size of hyperedges, so that we now have a relationship of order  $\mathcal{O}(nm)$  again, while still incorporating the features of neighboring nodes.

## 4.2 Model Architecture

In our CHGCNN model, initial atomic features were those used in CGCNN<sup>6</sup>, consisting of a concatenated set of one-hot encoded atomic properties. For bond features, we used a Gaussian expansion of interatomic distance of dimension 40 ranging from 0 to 6 Å, triplet features were a Gaussian expansion of the cosine of bond angle, also of dimension 40, and motif features were a concatenation of 93 scalar order parameters and continuous symmetry measures. In the model considered in this work, initial node and hyperedge features were first passed into a linear embedding layer (with no activation function) with an output dimension of 64.

These embedded features were then fed into a set of Crystal HyperGraph Convolutional (CHGConv) layers which utilize the neighborhood aggregation method presented above. In CHGConv,



**Figure 3** Overview of three possible message functions  $m$  for nodes  $v$  that generalize the message function in<sup>11</sup> to hyperedges  $h$  with more (or less) than two nodes. Here,  $n$  represents the average number of nodes in a hyperedge and the scaling relation is then for the total number of messages for one layer, for each hyperedge and each approach.

we use a set of CGConv<sup>6</sup> layers applied to consecutively larger hyperedge types, taking as input the origin node of the hyperedge, the connecting hyperedge feature as the edge feature, and an aggregated set of neighborhood features as the connected node feature. So, for every CHGConv layer, the atoms are updated by each hyperedge type chosen once (see Fig. 4). Note that each CGConv for different hyperedge types have independent trainable parameters.

These learned node features are then mean pooled to form a crystal vector, which is passed to a fully connected layer and then projected down to a one-dimensional (scalar) output for regression. In the case of classification tasks, the fully connected layer, after mean pooling, utilized a dropout mechanism and output a probability distribution of classes by way of a softmax activation function.

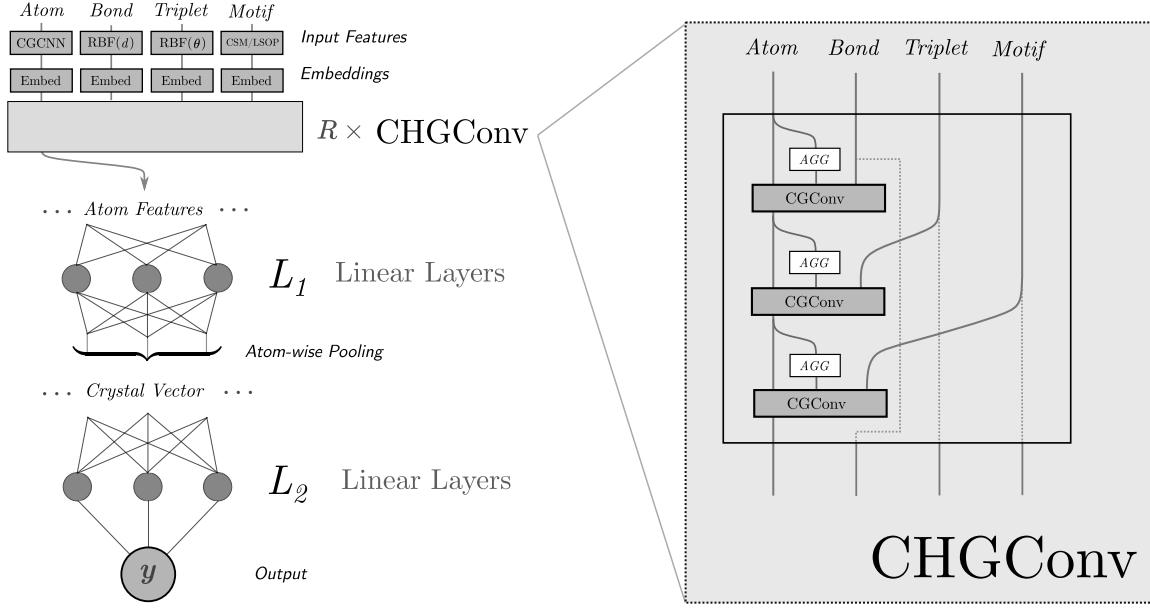
## 5 Results & Discussion

Crystal hypergraph networks provide a unique opportunity to investigate the importance of different order structures in the prediction of various material properties. Specifically, we may compare performance between models based on different types of hyperedges to probe the relevance of certain structures (e.g. motifs vs. triplets) in material property prediction. From the different hyperedge types considered here, we build three different models based on the architecture given in Fig. 4. We first test a basic bond-only network equivalent to CGCNN, and compare its performance against two models incorporating two types of hyperedges: bond-and-triplet and bond-and-motif models. For compound models (including more than one hyperedge type) each CHGConv layer performs convolution over the hyperedges in ascending order of hyperedge size. These models were each trained on sets of training data from two different databases of material properties: the Materials Project<sup>31</sup> and MatBench<sup>32</sup>. Details on the hyperparameters and training protocol can be found in the supplementary materials. Note that models using both motif and triplet-level edges, in general, diverged through training or did not perform any better

than models using just motifs or triplets. As such, we only compare models using one or the other here.

We first focus on the comparative performance of models incorporating only bonds, models incorporating both bond-and-motif hyperedges and models with both bond-and-triplet hyperedges on five MatBench target datasets<sup>32</sup>. Mean absolute error (MAE) on validation sets for these tasks are reported in Table 1. These five datasets consist of the following targets: the highest frequency phonon peak  $\omega_p$  for 1,265 materials, refractive indices  $n$  for 4,764 materials, formation energies  $E_f$  for a set of 18,829 perovskite materials, and 10,987 bulk and shear moduli,  $K_{vrh}$  and  $G_{vrh}$ , respectively. On all tasks of this set, the larger models incorporating both bonds and higher-order hyperedges performed best. Bond-and-triplet models performed best overall in predicting refractive indices and perovskite formation energies. For refractive index prediction, the triplet-based models had an average MAE of 0.398 across the test sets compared to an MAE of 0.424 for the bond-and-motif model, with both showing improvement over the bond-only model with an MAE of 0.454. Perovskite formation energy prediction results were very close for the three models, while the bond-and-triplet model again had the best performance with an MAE of 0.0598 eV/atom, the bond-and-motif model was close behind at an MAE of 0.0599 eV/atom, both showing only a slight improvement over the bond-only model's performance with an MAE of 0.604 eV/atom. However, since perovskites are a class of materials with a relatively standard structure, it should be unsurprising that the inclusion of additional structural information has little to no effect on performance.

Bond-and-motif models performed best in the remaining tasks. In the prediction of highest frequency phonon peak, the bond-and-motif model had the best performance with an average MAE of  $74.3 \text{ cm}^{-1}$  across test sets compared to  $77.5 \text{ cm}^{-1}$  for the triplet based model and  $82.1 \text{ cm}^{-1}$  for the bond-only model. For elastic targets, motif based models boasted average MAEs of 0.0943  $\log_{10}(\text{GPa})$  and 0.0730  $\log_{10}(\text{GPa})$  on the test sets of bulk moduli  $G_{vrh}$  and shear moduli  $K_{vrh}$ , respectively. This is compared to the performance of the bond-and-triplet models with MAEs of 0.0966



**Figure 4** Example architecture for the crystal hypergraph convolutional network implemented in this work. Essentially, the model is a generalization of CGCNN’s<sup>6</sup> model architecture with CGConv being replaced by  $R$  hypergraph convolutional layers (CHGConv). Here, CHGConv updates nodes first according to edges (bonds), and then according to triplets or motifs.

| Hyperedge Types | $\omega_p$<br>(1,265)<br>MAE ( $\text{cm}^{-1}$ ) | $n$<br>(4,764)<br>MAE | $E_f$<br>(18,829)<br>MAE (eV/Atom) | $\text{Log}_{10}(G_{vrh})$<br>(10,987)<br>MAE ( $\text{Log}_{10}\text{GPa}$ ) | $\text{Log}_{10}(K_{vrh})$<br>(10,987)<br>MAE ( $\text{Log}_{10}\text{GPa}$ ) |
|-----------------|---|-----------------------|------------------------------------|---|---|
| Bond-only       | 82.0  | 0.454                 | 0.0604                             | 0.1005  | 0.0816  |
| Bond & Triplet  | 77.5  | <b>0.398</b>          | <b>0.0598</b>                      | 0.0966  | 0.0757  |
| Bond & Motif    | <b>74.3</b>                                       | 0.424                 | 0.0599                             | <b>0.0943</b>   | <b>0.0730</b>   |

**Table 1** Validation dataset results for five MatBench target sets: highest frequency phonon peak  $\omega_p$ , refractive index  $n$ , perovskite formation energy  $E_f$ , and bulk moduli  $K_{vrh}$  and shear moduli  $G_{vrh}$ . Note that the italicized numbers below the target name correspond to the total size of each dataset. Best results are indicated in bold.

$\text{Log}_{10}(\text{GPa})$  and  $0.0757 \text{ Log}_{10}(\text{GPa})$  on  $G_{vrh}$  and  $K_{vrh}$ , with both being an improvement over the bond-only models with MAEs of  $0.1005 \text{ Log}_{10}(\text{GPa})$  and  $0.0816 \text{ Log}_{10}(\text{GPa})$ . This may be indicative of the importance of such information in the relation of stress to infinitesimal strain, since the local environments of atoms would be of particular importance in considerations of shear response (that is, calculations of  $G_{vrh}$ ) though, perhaps, less so in considerations of bulk response ( $K_{vrh}$ ).

We now consider these models performance on three more target properties for a much larger set of 152,605 materials from the Materials Project database<sup>31</sup>, with targets including formation energy  $E_f$ , band gap  $E_g$ , and metallicity. Results for this set of tests are reported in Table 2 with MAE reported for regression tasks and area under curve (AUC) for classification tasks. For the Materials Project dataset, the bond-and-triplet model performed best for all tasks. In the prediction of formation energy, the bond-and-motif model performed better than the bond-only model with an MAE of 0.0506 eV/atom vs. an MAE of 0.0512 eV/atom, while the motif-and-triplet model performed best with an MAE of 0.0484 eV/atom on the test set. This trend also held for the metal/nonmetal classification task, with the best performance on the test set again by the bond-and-triplet model with an AUC of 0.933. The bond-and-motif model had the next best performance on the test set, with an AUC of 0.932 compared to an AUC of 0.927 for the bond-only model. In the prediction of band gaps, the bond-and-motif model and bond-only model had prediction MAEs of 0.356 eV and 0.366 eV, respectively. The bond-and-triplet model again performed best with an MAE of 0.325 eV on the test set.

Perhaps the strongest point to be made in regard to these results is that for most tasks, motif information contributed to comparable or better performance than triplet-level results. This is at a significantly lower computational cost, in terms of the total number of messages exchanged through convolution, since the number of motifs is simply the number of atoms  $n$ , whereas the number of triplets scales with the average number of bonds per atom  $N$  as  $N(N - 1)/2$ .

A similar observation was made in the AMDNet architecture<sup>33</sup>, where motif information (included via an additional 'motif graph' for each material) also improved performance on most tasks, but here we compare results directly to the inclusion of bond angles via triplets. Our results indicate that one local neighborhood feature per atom may be sufficient to describe the local geometries of atoms for many predictive tasks, as opposed to the more data-intensive triplet representation scheme usually employed (often by way of line graphs). Taking this as a learned guiding principle, future crystal representations may benefit from reduced size while being assured similar geometric resolution. However, it should be noted that the greatest improvements seen here were on smaller datasets, suggesting the inductive bias of the motif-level hyperedges yields diminishing returns with larger datasets. This suggests that with large enough training data sets, continuous convolution filters for hyperedges encoding distance or angle alone may be sufficient for larger targets. However, given the intrinsic difficulty of the production of most target data in the materials science, the specific methods presented here may still prove beneficial in many applications.

## 6 Conclusions

State-of-the-art GNN models applied to material property prediction often represent material systems as graphs with relatively low geometrical resolution. This low resolution is often increased by

associating bond angles with auxiliary line graphs derived from the graph itself. The primary argument of our work is that hypergraphs are a more natural representation of material structures that allow us to explicitly incorporate geometrical information with different substructures of choice in one unified representation. The results suggest that such an approach allows for a substantial decrease in computational cost by incorporating such geometrical information with single local environment hyperedges for each node as opposed to triplets of atoms for each pair of overlapping bonds. This is shown within one unified framework to have comparable performance on a number of common predictive tasks.

Future works may investigate more powerful hypergraph convolutional operators that automatically detect motifs<sup>18,20,34</sup>; or apply this framework to molecular systems<sup>35,36</sup> with functional groups. Inter-order convolution may also be of interest for certain tasks, where different hyperedge types may update each other's representations as opposed to just atom representations. Note that inter-order convolution would allow for a complete generalization of previous line-graph convolution schemes, where triplets effectively update their respective bonds' representations through convolution, as in<sup>12,13</sup>. Other order structures (beyond motif-level) may also be of interest, such as hyperedges representing defect complexes or entire unit cells. Equivariant features and convolution<sup>37–39</sup> may also be incorporated for the prediction of coordinate-system dependent properties of materials from hypergraph representations, with the present work being focused on coordinate-system invariant features and targets.

## Author contributions

A.H. and Q.Y. conceived the research and wrote the manuscript. W.G. provided partial software support and was involved in discussions. Q.Y. supervised the project.

## Conflicts of interest

There are no conflicts to declare.

## Data availability

The code used in this paper's results were built on pytorch-geometric and can be found in the following Github repository: <https://github.com/qmatyanlab/CHGCNN>. The processed data including bond, triplet, and motif features is available in a Zenodo repository with DOI: 10.5281/zenodo.14756640.

## Acknowledgements

This work is supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award No. DE-SC0023664. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0029544.

## References

- [1] L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, *npj Comput. Mater.*, 2016, 2, 1–7.

| Hyperedge<br>Types | $E_f$<br>Best MAE (eV/Atom) | $E_g$<br>Best MAE (eV) | Metal/Non-metal<br>Classification<br>Best AUC |
|--------------------|-----------------------------|------------------------|---|
| Bond-only          | 0.0512                      | 0.356                  | .927  |
| Bond & Triplet     | <b>0.0484</b>               | <b>0.325</b>           | <b>.933</b>                                   |
| Bond & Motif       | 0.0506                      | 0.366                  | .932  |

**Table 2** Validation dataset results for three Materials Project target sets: formation energy  $E_f$ , band gap  $E_g$ , and metalicity. Here, each dataset included a total of 152,605 materials after 300 epochs of training with 5-fold nested cross validation. Best results are indicated in bold.

- [2] J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei and M. Lei, *InfoMat*, 2019, **1**, 338–358.
- [3] Y. Liu, C. Niu, Z. Wang, Y. Gan, Y. Zhu, S. Sun and T. Shen, *J. Mater. Sci. Tech.*, 2020, **57**, 113–122.
- [4] M. H. Mobarak, M. A. Mimona, M. A. Islam, N. Hossain, F. T. Zohura, I. Imtiaz and M. I. H. Rimon, *Appl. Surf. Sci. Adv.*, 2023, **18**, 100523.
- [5] K. T. Schütt, H. E. Saucedo, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.
- [6] T. Xie and J. C. Grossman, *PRL*, 2018, **120**, 145301.
- [7] C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564–3572.
- [8] C. Chen, E. Xu, D. Yang, C. Yan, T. Wei, H. Chen, Y. Wei and M. Chen, *Neural Comput. Appl.*, 2024, **37**, 3287–3301.
- [9] J. Cheng, C. Zhang and L. Dong, *Commun. Mater.*, 2021, **2**, 92.
- [10] C. W. Park and C. Wolverton, *Phys. Rev. Mater.*, 2020, **4**, 063801.
- [11] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, *34th Int. Conf. Mach. Learn.*, 2017, **70**, 1263–1272.
- [12] R. Ruff, P. Reiser, J. Stühmer and P. Friederich, *Digit. Discov.*, 2024, **3**, 594–601.
- [13] K. Choudhary and B. DeCost, *npj Comput. Mater.*, 2021, **7**, 1–8.
- [14] C. Chen and S. P. Ong, *Nat. Comput. Sci.*, 2022, **2**, 718–728.
- [15] L. Pauling, *J. Chem. Soc.*, 1929, **51**, 1010–1026.
- [16] R. King, *J. Chem. Ed.*, 1996, **73**, 993.
- [17] D. Waroquiers, X. Gonze, G.-M. Rignanese, C. Welker-Nieuwoudt, F. Rosowski, M. Gobel, S. Schenk, P. Degelmann, R. André, R. Glaum *et al.*, *Chem. Mater.*, 2017, **29**, 8346–8360.
- [18] J. Dan, X. Zhao, Q. He, N. D. Loh and S. J. Pennycook, *Microsc. Microanal.*, 2022, **28**, 3002–3003.
- [19] Z. Yang and L.-H. Tang, *Phys. Rev. B*, 2009, **79**, 045402.
- [20] K. Sheriff, Y. Cao and R. Freitas, *npj Comput. Mater.*, 2024, **10**, 215.
- [21] N. E. R. Zimmermann, M. K. Horton, A. Jain and M. Haranczyk, *Front. Mater.*, 2017, **4**, 34.
- [22] N. E. Zimmermann and A. Jain, *RSC Adv.*, 2020, **10**, 6063–6081.
- [23] E. E. Santiso and B. L. Trout, *J. Chem. Phys.*, 2011, **134**, 6.
- [24] J. H. Chang, D. Kleiven, M. Melander, J. Akola, J. M. Garcia-Lastra and T. Vegge, *J. Phys. Condens. Matter*, 2019, **31**, 325901.
- [25] A. van de Walle, *Nat. Mater.*, 2008, **7**, 455–458.
- [26] J. Sanchez, F. Ducastelle and D. Gratias, *Physica A Stat. Mech. Appl.*, 1984, **128**, 334–350.
- [27] M. Pinsky and D. Avnir, *Inorg. Chem.*, 1998, **37**, 5575–5582.
- [28] D. Waroquiers, J. George, M. Horton, S. Schenk, K. A. Persson, G.-M. Rignanese, X. Gonze and G. Hautier, *Acta Crystallogr. B*, 2020, **76**, 683–695.
- [29] H. Pan, A. M. Ganose, M. Horton, M. Aykol, K. A. Persson, N. E. R. Zimmermann and A. Jain, *Inorg. Chem.*, 2021, **60**, 1590–1603.
- [30] Z. Chen, X. Li and J. Bruna, *arXiv:1705.08415*, 2017.
- [31] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
- [32] A. Dunn, Q. Wang, A. Ganose, D. Dopp and A. Jain, *npj Comput. Mater.*, 2020, **6**, 138.
- [33] H. R. Banjade, S. Hauri, S. Zhang, F. Ricci, W. Gong, G. Hautier, S. Vucetic and Q. Yan, *Sci. Adv.*, 2021, **7**, eabf1754.
- [34] S. Zhang, Z. Hu, A. Subramonian and Y. Sun, *arXiv:2012.12533*, 2020.
- [35] C. R. Collins, G. J. Gordon, O. A. Von Lilienfeld and D. J. Yaron, *J. Chem. Phys.*, 2018, **148**, 24.
- [36] N. Fedik, R. Zubatyuk, M. Kulichenko, N. Lubbers, J. S. Smith, B. Nebgen, R. Messerly, Y. W. Li, A. I. Boldyrev, K. Barros *et al.*, *Nat. Rev. Chem.*, 2022, **6**, 653–672.
- [37] M. Geiger and T. Smidt, *arXiv:2207.09453*, 2022.
- [38] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff and P. Riley, *arXiv:1802.08219*, 2018.
- [39] K. Yan, C. Fu, X. Qian, X. Qian and S. Ji, *arXiv:2403.11857*, 2024.