A Crystal Knowledge-Enhanced Pre-training Framework for Crystal Property Estimation

Haomin Yu[1], Yanru Song[2], Jilin Hu[2], Chenjuan Guo[2], Bin Yang[2(✉)], and Christian S. Jensen[1]

[1] Aalborg University, Aalborg, Denmark
{haominyu,csj}@cs.aau.dk
[2] East China Normal University, Shanghai, China
songyanru@stu.ecnu.edu.cn, {jlhu,cjguo,byang}@dase.ecnu.edu.cn

Abstract. The design of new crystalline materials, or simply crystals, with desired properties relies on the ability to estimate the properties of crystals based on their structure. To advance the ability of machine learning (ML) to enable property estimation, we address two key limitations. First, creating labeled data for training entails time-consuming laboratory experiments and physical simulations, yielding a shortage of such data. To reduce the need for labeled training data, we propose a pre-training framework that adopts a mutually exclusive mask strategy, enabling models to discern underlying patterns. Second, crystal structures obey physical principles. To exploit the principle of periodic invariance, we propose multi-graph attention (MGA) and crystal knowledge-enhanced (CKE) modules. The MGA module considers different types of multi-graph edges to capture complex structural patterns. The CKE module incorporates periodic attribute learning and atom-type contrastive learning by explicitly introducing crystal knowledge to enhance crystal representation learning. We integrate these modules in a **CR**ystal kn**O**wledge-enhanced **P**re-training (CROP) framework. Experiments on eight different datasets show that CROP is capable of promising estimation performance and can outperform strong baselines.

Keywords: Crystal property · Pre-training · Knowledge-enhanced

1 Introduction

A crystalline material, or simply crystal, is a solid material with a highly regular internal structure, where atoms, molecules, and ions are arranged in a specific pattern in a unit cell. This pattern is then repeated to form a crystal lattice [30], as exemplified in Fig. 1. Traditional approaches to developing new crystals are driven by manual experiments that are often costly and time-consuming [29]. Next, calculations based on density functional theory (DFT) have emerged as valuable means of determining the properties of crystals [7]. However, despite the success of high-throughput DFT calculations, their high computational costs limit their utility for estimating crystal properties. Most

A. Bifet et al. (Eds.): ECML PKDD 2024, LNAI 14950, pp. 231–246, 2024.
https://doi.org/10.1007/978-3-031-70381-2_15

recently, machine learning (ML) methods [24] have emerged that use predominantly neural networks (GNNs) to model crystals as graphs and then mine graph-structured dependencies, thereby achieving competitive crystal property estimation accuracy. However, crystal property estimation remains challenging due to two key limitations. (1) *Limited labeled crystal data.* Although ML methods are promising for crystal property estimation, they require large amounts of labeled crystal data for model training. However, property labeling of crystals requires laboratory experiments or physical simulations that are labor-intensive and time-consuming. Thus, only limited labeled data is available. (2) *Underutilized crystal principles.* Crystal structures obey several principles [30,33], e.g., E(3) invariance and periodic invariance (see Sect. 3), which are foundational concepts in materials science. When building ML methods for crystal property estimation, it is important to exploit such principles. Yet, most existing methods focus on E(3) invariance and ignore periodic invariance, limiting their ability to estimate crystal properties.

To address these challenges, we propose a **CR**ystal kn**O**wledge-enhanced **P**re-training (CROP) framework for crystal property estimation. First, to reduce the need for labeled crystal data, we leverage large amounts of unlabeled crystal data using self-supervised learning (SSL) that employs mutually exclusive masking that provides two mutually exclusive views for achieving a pre-training framework. The pre-training framework uses not only an autoencoder that emphasizes reconstruction and feature preservation but also the Barlow Twins approach [36], which encourages the learning of discriminative features. This combination can provide a more comprehensive understanding of underlying crystal structures. Second, to better exploit crystal principles, we design a multi-graph attention (MGA) module and a crystal knowledge enhanced (CKE) module. The former builds a novel multi-graph attention network to capture complex crystal structure patterns by exploiting periodic invariance. According to the periodic invariance of crystals, we customize the attention mechanism for processing different types of multi-graph edges, i.e., inter-unit edges that connect atoms within the same unit and intra-unit edges that connect atoms across units. For example, the edges in Fig. 1 that connect atoms within the target unit are inter-unit edges, and the remaining ones are intra-unit edges. Next, the CKE module incorporates intrinsic attributes related to crystal periodic invariance, e.g., discrete direction, unit cell position, and edge distance, to enable a better representation to be learned during the self-supervised pre-training.

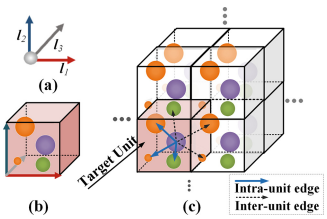

Fig. 1. Illustration of crystal structure: (a) periodic lattice, (b) a unit cell, and (c) an infinite periodic structure.

The paper makes four main contributions: (1) It proposes a novel masking strategy that exploits two mutually exclusive views for effective self-supervised pre-training, reducing the need for labeled data. (2) It proposes a multi-graph attention (MGA) module that exploits periodic invariances in crystal structures to process different types of multi-graph edges, enhancing the expressiveness of

learned representations and capturing complex structural patterns. (3) It proposes a crystal knowledge-enhanced (CKE) module that exploits periodic invariants in crystal structures to support periodic attribute learning and atom-type contrastive learning. (4) It reports experiments on eight crystal datasets, providing insights into CROP's design properties and evidence of its effectiveness.

2 Related Work

Material Property Estimation. There has been a substantial increase in studies that leverage ML techniques, specifically graph neural networks (GNN), for estimating molecular properties. GNN-based proposals use either 2D [6,27] or 3D [2,3,25,31] molecular graphs. Since crystal properties depend heavily on their 3D structures, we focus on 3D molecular graphs. Most such proposals do not explicitly consider the periodical repeating patterns in crystal structures. As one exception, Matformer [33] introduces a graph construction method that introduces edges connecting the same atoms in neighboring units according to periodic invariance. However, Matformer does not consider the complex impacts of different atoms in neighboring units. We go further and design a novel multi-graph attention network, where edges capture relationships among different atoms and further distinguish different types of multi-graph edges, i.e., inter-unit edges and intra-unit edges. Second, we enable self-supervised model training using mutually exclusive mask views, reducing the need for labeled data.

Self-supervised Learning. Self-supervised learning (SSL) is becoming popular [1,16,17,21]. SSL techniques can be divided into contrastive and generative SSL. *Contrastive SSL* aims to learn meaningful representations by maximizing the similarity between pairs of augmented samples [34,35]. As a specific type of contrastive SSL, the Barlow Twins [36] approach aims to learn valuable data representations without requiring negative samples. Magar et al. [22] propose Crystal Twins, using Barlow Twins to make graph latent embeddings of augmented instances from the same crystal system similar. However, while this approach encourages learning discriminative features, it may inadvertently neglect explicit structure patterns. *Generative SSL* focuses on reconstructing original input information [21]. Motivated by the success of masking techniques [8], recent studies, such as Hou et al. [14], have integrated masking into SSL frameworks, focusing on feature reconstruction using an autoencoder. Unlike the previous studies, we propose a masking strategy that combines both contrastive (Barlow Twins) and generative (autoencoder) SSL, leveraging a novel masking approach.

3 Preliminaries

Crystal Representation. A crystal structure can be modeled using three vectors $\mathbf{C} = (\mathbf{A}, \mathbf{X}, \mathbf{L})$ [30]. It is represented by repeated translations of a unit cell $(\mathbf{A}, \mathbf{X})$ according to the a lattice matrix $\mathbf{L}$, as illustrated in Fig. 1. The periodic lattice matrix $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3] \in \mathbb{R}^{3\times3}$ is used to capture how a unit cell repeats

itself in three dimensions. Specifically, we represent a unit cell that includes N atoms as $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N] \in \mathbb{R}^{N \times Q}$, where $\mathbf{a}_i$ is a one-hot encoded feature vector of its atom type (i.e., a chemical element), where Q is the count of all chemical elements. Then, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N] \in \mathbb{R}^{N \times 3}$ denotes the coordinates of each atom, where atom position $\mathbf{x}_i \in \mathbb{R}^3$ is represented by 3D Cartesian coordinates. For example, the unit cell of the crystal $Hf_2Si_2Te_2$ comprises 6 atoms: two of each chemical element—Hf, Si, and Te. For this crystal, the one-hot encoded feature vector is $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_6] \in \mathbb{R}^{6 \times Q}$. The atom coordinates for $Hf_2Si_2Te_2$ are represented by the matrix $\mathbf{X}$. Thus, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_6]$ falls within $\mathbb{R}^{6 \times 3}$. The lattice matrix for $Hf_2Si_2Te_2$ is $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3] \in \mathbb{R}^{3 \times 3}$ and is given by: $\begin{bmatrix} 3.66730534 & 0.0 & 0.0 \\ 0.0 & 3.66730534 & 0.0 \\ 0.0 & 0.0 & 27.311209 \end{bmatrix}$.

Given a crystal $\mathbf{C}$, its infinite periodic structure can be represented by utilizing a periodic lattice to repeat the unit cell in the 3D space as follows.

$$\mathbf{A}' = \{\mathbf{a}'_i | \mathbf{a}'_i = \mathbf{a}_i, 1 \leq i \leq N\} \tag{1}$$

$$\mathbf{X}' = \{\mathbf{x}'_i | \mathbf{x}'_i = \mathbf{x}_i + k_1\mathbf{l}_1 + k_2\mathbf{l}_2 + k_3\mathbf{l}_3, 1 \leq i \leq N, k_1, k_2, k_3 \in \mathbb{Z}\}, \tag{2}$$

where integers k_1, k_2, and k_3 serve as translation vectors, enabling the unit cell to be replicated in three dimensions by means of the periodic lattice $\mathbf{L}$.

Crystal Principles. Crystal structures satisfy fundamental chemical principles [30,31,33], including translation, rotation, and reflection invariance, as well as periodic invariance. *Translation Invariance* states that a crystal structure remains unchanged when translating the atom coordinate matrix by an arbitrary vector. *Rotation Invariance* states that a crystal structure remains unchanged when rotating the coordinate matrix and lattice matrix simultaneously. *Reflection Invariance* states that crystals remain unchanged when mirrored across a plane. The combination of translation, rotation, and reflection invariance is termed *E(3) invariance*. *Periodic Invariance* refers to the repeating nature of a crystal's structure. For example, the crystal structure in Fig. 1 (c) is the repeating structure of the crystal unit in Fig. 1 (b) according to regular intervals.

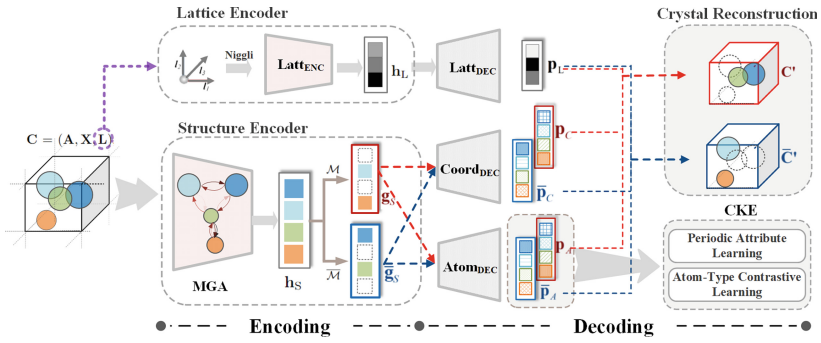

Fig. 2. The crystal knowledge-enhanced pre-training framework.

4 Methodology

4.1 Framework Overview

We propose a crystal knowledge-enhanced pre-training framework that uses mutually exclusive masked views for learning crystal representations—see Fig. 2.

The encoding phase involves two components: a lattice encoder and a structure encoder. The lattice encoder Latt_{ENC} encodes the lattice matrix $\mathbf{L}$ into a lattice representation $\mathbf{h}_L \in \mathbb{R}^{d_l}$ that emphasizes the periodic information. The lattice matrix $\mathbf{L}$ does not follow E(3) invariance, so we pass it to the Niggli Algorithm [12], which can establish a set of conditions that determine a unique choice of basis vectors for a lattice. The structure encoder encodes the crystal $\mathbf{C}$ into a structure representation $\mathbf{h}_S = [\mathbf{h}_S^1, \mathbf{h}_S^2, ..., \mathbf{h}_S^N] \in N \times \mathbb{R}^{d_s}$. To capture complex and subtle structural patterns, the structure encoder encompasses a multi-graph attention (MGA) module. The structure representation $\mathbf{h}_S$ is then masked randomly to produce $\mathbf{g}_S$ and $\overline{\mathbf{g}}_S$ under two mutually exclusive views.

During the decoding phase, the encoded representations are fed to different decoders. The lattice representation $\mathbf{h}_L$ is fed to the lattice decoder Latt_{DEC}, which yields a reconstructed lattice representation $\mathbf{p}_L$. The representations $\mathbf{g}_S$ and $\overline{\mathbf{g}}_S$, created under mutually exclusive views, are processed by the coordinate decoder $\text{Coord}_{\text{DEC}}$, resulting in $\mathbf{p}_C$ and $\bar{\mathbf{p}}_C$. Concurrently, the representations $\mathbf{g}_S$ and $\overline{\mathbf{g}}_S$ are also passed to the atom decoder Atom_{DEC}, generating reconstructed representations $\mathbf{p}_A$ and $\bar{\mathbf{p}}_A$. The atom decoder Atom_{DEC} and coordinate decoder $\text{Coord}_{\text{DEC}}$ are Graph Isomorphism Networks (GIN) [32] that capture the relationships and dependencies between different atoms and substructures.

Next, the decoding process includes the crystal reconstruction part and the crystal knowledge enhanced (CKE) module. The crystal reconstruction part reconstructs based on two inputs, i.e., $(\mathbf{p}_L, \mathbf{p}_C, \mathbf{p}_A)$ and $(\mathbf{p}_L, \bar{\mathbf{p}}_C, \bar{\mathbf{p}}_A)$, as shown by the red and blue reconstructed crystal (i.e., $\mathbf{C}'$ and $\bar{\mathbf{C}}'$) in Fig. 2. The CKE module leverages crystal knowledge to enhance the crystal reconstruction. Inspired by periodic invariance, it consists of periodic attribute learning and atom-type contrastive learning to exploit periodic invariance and improve reconstruction.

4.2 Reconstruction Under Mutually Exclusive Masked Views

Inspired by the successful masked language modeling technique used in BERT [8], we introduce masking techniques into our pre-training framework. The relationships among atoms in crystals are analogous to contextual relationships among words. To avoid favoring certain patterns and obtain more diverse representations, we design a masking strategy with two mutually exclusive views.

Mutually Exclusive Masking. The masking purposefully corrupts the representation $\mathbf{h}_S$ to enforce the model to learn representations that consider surrounding atoms and their relationships, rather than learning individual atom representations. We first define mutually exclusive masks. We consider a full

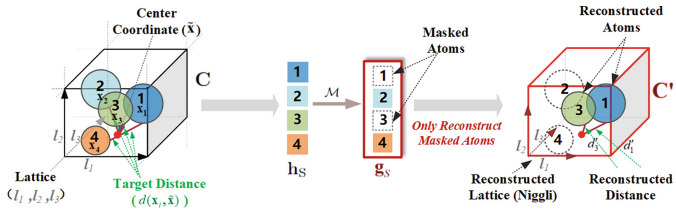

Fig. 3. The reconstruction procedure with view $\mathcal{M}$ (atoms 1 and 3 are masked).

feature set with N indices, i.e., $\{1, ..., N\}$. If a mask randomly selects some of the indices $\mathcal{M}$ with a uniform distribution then the mask processing replaces the values with these indices by [MASK]. Another mask replaces the values of the complementary set of indices, denoted as $\overline{\mathcal{M}}$ ={1,..., N}\$\mathcal{M}$. We refer to $\mathcal{M}$ and $\overline{\mathcal{M}}$ which are disjoint sets of indices, as *mutually exclusive masks*.

For the structure representation $\mathbf{h}_S = [\mathbf{h}_S^1, \mathbf{h}_S^2, ..., \mathbf{h}_S^N]$, we randomly select a subset of atoms $\mathcal{M}$ and replace their subset of features with [MASK] $\in \mathbb{R}^{d_s}$ token. Each atom's feature vector remains unchanged if the atom is not in $\mathcal{M}$; otherwise, it is replaced by the [MASK] token. This creates a new representation, $\mathbf{g}_S$, consisting of both original and masked features, as shown in Fig. 3. We use a subset of $\mathcal{M}$ to mask $\mathbf{h}_S$ to get $\mathbf{g}_S$ and its mutually exclusive subset $\overline{\mathcal{M}}$ to mask $\mathbf{h}_S$ to get $\overline{\mathbf{g}}_S$. This masking process is critical as it forces the model to learn representations based on the structural relationships between two complementary sets of atoms rather than solely depending on each atom's features.

Crystal Reconstruction. After passing the masked representations to the decoder, we use a pre-training framework without labeled data to reconstruct crystal $\mathbf{C}$ as $\mathbf{C}'$ and $\bar{\mathbf{C}}'$. Instead of reconstructing the crystal directly, we focus on reconstructing the three core components, $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{L}$, by optimizing the reconstruction loss $\mathcal{L}_{REC}$. We only calculate the loss related to the masked values, as shown in Fig. 3. The effectiveness of only reconstructing masked values has been demonstrated in the MAE model [13]. The reconstruction loss is the sum of three sub-losses, i.e., $\mathcal{L}_{REC} = L_{atom} + L_{coord} + L_{lattice}$. (1) Atom types $\mathbf{A}$ are reconstructed by minimizing the L_{atom} loss between the ground truth atom types $\mathbf{a}_i$ and $\mathbf{a}_j$ and reconstructed atom types $\mathbf{a}'_i$ and $\bar{\mathbf{a}}'_j$. (2) Directly predicting a crystal's absolute coordinates $\mathbf{X}$ cannot follow the E(3) invariance of crystals. The properties of crystals do not relate to their absolute coordinates but rather to the structural relationships among their atoms. Hence, rather than directly reconstructing the absolute coordinates, we reconstruct the distance between atom coordinate $\mathbf{x}_i$ and the center coordinate $\tilde{\mathbf{x}}$, which is calculated as the average of the atom coordinates of all atoms in a crystal in the unit cell, as shown in Fig. 3. Then, we calculate the L_{coord} loss based on the target distances (i.e., $d(\mathbf{x}_i, \tilde{\mathbf{x}})$ and $d(\mathbf{x}_j, \tilde{\mathbf{x}})$) and reconstructed distances (i.e., d'_i and $\bar{d}'_j$). (3) We optimize the lattice reconstruction process through an $L_{lattice}$ loss between the estimated lattice $\hat{\mathbf{L}}$ and the real lattice $\mathbf{L}$.

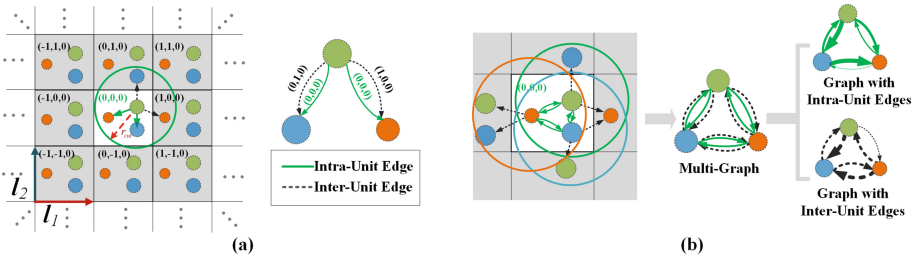

Fig. 4. Multi-graph construction for a crystal with 3 atoms.

4.3 Multi-graph Attention Module

We employ a multi-graph attention module to capture complex structural patterns adaptively. We start by detailing the multi-graph construction [31].

Multi-graph Construction. A crystal material can be represented as a multi-graph $\mathbf{G} = (\mathbf{A}, \mathbf{E})$ when learning the crystal structure. $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_N] \in \mathbb{R}^{N \times Q}$ is the set of atom nodes in the crystal (see Sect. 3). $\mathbf{E} = \{e_{ij,(k_1,k_2,k_3)} | i, j \in \{1, ..., N\}, k_1, k_2, k_3 \in \mathbb{Z}\}$ is the set of edges representing relevant atom pairs in the crystal. An edge $e_{ij,(k_1,k_2,k_3)}$ denotes a directed edge from node i in the original unit cell to node j in the cell translated by $k_1\mathbf{l}_1 + k_2\mathbf{l}_2 + k_3\mathbf{l}_3$, where $\mathbf{l}_1, \mathbf{l}_2$, and $\mathbf{l}_3$ are the lattice vectors of the crystal. To construct edges for relevant atom pairs while observing periodic invariance, the k-nearest neighbor (kNN) approach with cutoff distances r_{cut} is often used. We ensure that the multi-graph follows the periodic invariance by constructing edges between atom nodes i and j that satisfy $d_{ij} \leq r_{cut}$, where $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j + k_1\mathbf{l}_1 + k_2\mathbf{l}_2 + k_3\mathbf{l}_3\|_2$, $||\cdot||_2$ denotes Euclidean distance and $k_1, k_2, k_3 \in \mathbb{Z}$. As shown in Fig. 4 (a), taking the green atom in the target unit as an example, edges are constructed to other atoms that are within the cutoff distance r_{cut}. The (k_1, k_2, k_3) values on the edges show the translation operation according to the lattice matrix. For example, (0,0,0) indicates that the edge connection resides within the same unit cell, while (0,1,0) signifies a connection to a neighboring unit cell via translation along the lattice vector $\mathbf{l}_2$. However, the relationships among atoms in a crystal are intricate, especially for atoms spanning units. Therefore, we have tailored a multi-graph attention module guided by multi-graph edge types.

Multi-graph Attention Mechanism. The unit cell of a crystal is defined as the smallest repeating unit that shows the full symmetry of its structure. Thus, the goal of our reconstruction task is to reconstruct the unit cell. An atom within the target unit cell should have a different impact on our reconstruction target than should an atom outside the target cell. Thus, we customize a multi-graph attention module employing both intra-unit and inter-unit attention mechanisms to learn the crystal structure effectively.

To describe our method, we first define an indicator function $U_{i,j}$ that captures whether atoms i and j are in the same unit cell. When atoms i and j

satisfy $(k_1, k_2, k_3) = (0, 0, 0)$, they are connected by an *intra-unit edge*. Otherwise, they are connected by an *inter-unit edge*. For ease of illustration, we consider a 2D space by setting $k_3 = 0$ and only repeat the unit cell along the $\mathbf{l}_1$ and $\mathbf{l}_2$ directions. Figure 4 (a) shows the multi-graph construction for the green atom, including outgoing multi-edges labeled with (k_1, k_2, k_3). When an edge connects two atoms in the same unit (i.e., $U_{i,j} = 1$), it is an intra-unit edge. Otherwise, it is an inter-unit edge.

We can now introduce the proposed multi-graph attention mechanism. To capture complex and subtle structural patterns adaptively, we transform the one-hot atom node feature $\mathbf{a}_i$ into an embedding feature $\mathbf{b}_i$ and then perform self-attention on the embedding feature as the strategy of the graph attention network [28], which is a shared attention mechanism. The shared linear transformation, parameterized by a weight matrix $\mathbf{W}$, is applied to the initial feature vector of each atom node $\mathbf{b}_i$.

$$\mathbf{b}_i = Emb(\mathbf{a}_i), r_{ij} = f_a(\mathbf{W}\mathbf{b}_i, \mathbf{W}\mathbf{b}_j), \tag{3}$$

where f_a is a single-layer feed-forward neural network and $Emb(\cdot)$ denotes a one-hot atom embedding function. To further determine the influence of different types of multi-graph edges, we separate the multi-graph into a graph with intra-unit edges and a graph with inter-unit edges, as shown in Fig. 4 (b). Then, we calculate the correlation coefficients between the graph edges in the two graphs.

$$A_{ij}^{intra} = \frac{\exp(\Phi(r_{ij}))}{\sum_{k \in \mathcal{N}_i} U_{i,j} \exp(\Phi(r_{ik}))}, A_{ij}^{inter} = \frac{\exp(\Phi(r_{ij}))}{\sum_{k \in \mathcal{N}_i} (1 - U_{i,j}) \exp(\Phi(r_{ik}))} \tag{4}$$

Here, A_{ij}^{intra} is an intra-unit edge correlation coefficient that is normalized by the softmax function, and A_{ij}^{inter} is an inter-unit edge correlation coefficient. These coefficients indicate the importance of a neighbor node j to a target node i. Next, $\Phi(\cdot)$ is the LeakyReLU activation function, and $\mathcal{N}_i$ represents the neighbor node set of node i. Then, we perform attention on the atom nodes and calculate the reweighted atom node $\hat{\mathbf{b}}_i$ as follows.

$$\hat{\mathbf{b}}_i = \sum_{j \in \mathcal{N}_i} U_{ij} A_{ij}^{intra} \mathbf{b}_j \oplus \sum_{j \in \mathcal{N}_i} (1 - U_{ij}) A_{ij}^{inter} \mathbf{b}_j, \tag{5}$$

where $\oplus$ denotes concatenation followed by fully connected layers.

Message Updating. We employ DimeNet++ [11], an E(3) invariant graph neural network, to capture atom interactions within the crystal. It uses spherical functions for angles and radial functions for distances between atoms. DimeNet++ refines atom representations $\hat{\mathbf{B}} = \left[\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2, \cdots, \hat{\mathbf{b}}_N\right]$ through its network structure, capturing comprehensive crystal interactions $\mathbf{h}_s = \text{DimeNet++}(\hat{\mathbf{B}}, \mathbf{X}, \mathbf{L})$.

4.4 Crystal Knowledge Enhanced Module

To enable the use of crystal knowledge for crystal reconstruction, we design a crystal knowledge-enhanced module, encompassing periodic attribute learning and atom-type contrastive learning.

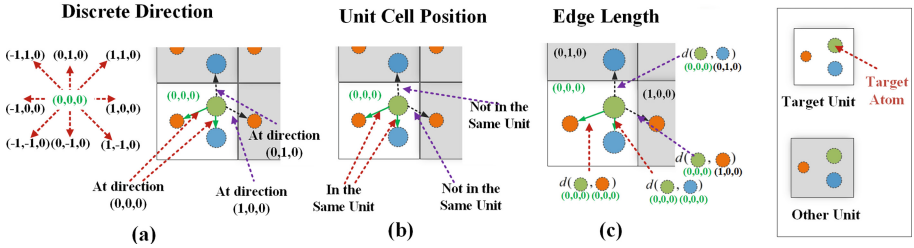

Fig. 5. The periodic attribute learning module. The white square is the target unit, and the green atom in the target unit is the target atom. (a) Illustration of discrete directions for the target atom to other atoms in neighbor units. (b) Illustration of whether the target atom and its neighbor atoms are in the same unit cell. (c) Illustration of the distances between the target atom and its neighbor atoms. (Color figure online)

Periodic Attribute Learning. We customize a periodic attribute learning module for the self-supervised learning framework. This module predicts three types of attributes, introducing learning constraints that facilitate more robust representation learning. We use the atom representation, i.e., $\mathbf{p}_A$ and $\bar{\mathbf{p}}_A$, as the node feature for the graph covered in Sect. 4.3. For each edge, we concatenate the i-th node of crystal $\mathbf{C}$ with the representations of its neighboring atom nodes and obtain $\mathbf{p}_A^{ij} = [\mathbf{p}_A^i || \mathbf{p}_A^j]$ and $\bar{\mathbf{p}}_A^{ij} = [\bar{\mathbf{p}}_A^i || \bar{\mathbf{p}}_A^j]$. These representations (i.e., $\mathbf{p}_A^{ij}$ and $\bar{\mathbf{p}}_A^{ij}$) are then fed to three types of attribute learners, which are multi-layer perceptrons [10], for predicting three periodic-specific attributes: discrete direction, unit cell position, and distance between nodes.

More specifically, given a node with a coordinate $\mathbf{x}_i$ and its neighbor $\mathbf{x}_j$ can be represented as $\mathbf{x}_j = \mathbf{x}'_j + k_1\mathbf{l}_1 + k_2\mathbf{l}_2 + k_3\mathbf{l}_3$. The crystal attributes between edges $\mathbf{x}_i$ and $\mathbf{x}_j$ are detailed as follows. (1) *Discrete Direction.* We determine the directions between each atom and its corresponding neighboring atoms. We streamline the process by discretizing directions into the 27 distinct directions in 3D space. If atom pairs are in the same unit (i.e., $k_1, k_2, k_3 = 0$), we regard this situation as the same direction. Thus, directions can be discretized by the arrangement and combination of k_1, k_2, k_3 after the modulo operation, i.e., $\frac{k_1}{|k_1|}, \frac{k_2}{|k_2|}, \frac{k_3}{|k_3|} \subseteq \{1, 0, -1\}$. To better illustrate this concept, we show it in 2D space, where there are only 9 directions—see Fig. 5 (a). We use the L_{dir} loss to optimize the discrete direction estimation task. (2) *Unit Cell Position.* The unit cell position attribute is used to determine whether two connected atoms belong to the same unit cell. As shown in Fig. 5 (b), if both atoms are in the same unit cell, they have an inter-unit edge, i.e., $k_1, k_2, k_3 = 0$. Otherwise, the atom pairs have intra-unit connections, i.e., $\{(k_1, k_2, k_3) \in \mathbb{Z}^3 \mid (k_1, k_2, k_3) \neq (0, 0, 0)\}$. The unit cell position estimation is a binary classification task optimized using the L_{pos} loss. (3) *Edge Length.* As shown in Fig. 5 (c), we predict the edge length between connected atoms by calculating the distance between atoms $\mathbf{x}_i$ and $\mathbf{x}_j$. The edge length prediction is a regression problem, which is optimized by using the L_{dis} loss. The prediction process still employs a distance learner, i.e.,

a multilayer perceptron [10], taking atom-type representations $\mathbf{p}_A$ and $\bar{\mathbf{p}}_A$ as input and outputting the edge length.

Atom-Type Contrastive Learning. The type of an atom is important when estimating crystal properties because the properties of atoms determine how they interact with one another in a crystal lattice [20]. Different atom types have distinct electron configurations and bonding characteristics, which can influence a crystal's properties markedly. Therefore, in addition to reconstructing the masked representations of crystal material, we introduce constraints for atom types. The Barlow Twins approach is a specific type of contrastive learning that can learn useful representations of data while reducing redundancy between input vectors. We use this approach to further constrain atom-type representation learning. Our objective is to learn informative and complementary representations by designing mutually exclusive masks.

We introduce the Barlow Twins loss L_{BT} [36] to measure the relationship between $\mathbf{p}_A$ and $\bar{\mathbf{p}}_A$, i.e., $L_{BT} = BT(\mathbf{p}_A, \bar{\mathbf{p}}_A)$, where $BT(\cdot)$ is the Barlow Twins loss. Although $\mathbf{p}_A$ and $\bar{\mathbf{p}}_A$ are derived from mutually exclusive views, they gather ample information from neighbor nodes after the decoder's reconstruction, resulting in reconstructed vectors that should be similar. To further ensure the reliability of this reconstructed information, the Barlow Twins loss measures the cross-correlation matrix between $\mathbf{p}_A$ and $\bar{\mathbf{p}}_A$ and makes it as close to the identity matrix as possible. It can make these two vectors similar while minimizing the redundancy between their components.

In summary, we exploit crystal knowledge to enhance the representation reconstruction by optimizing the $\mathcal{L}_{CKE}$ loss, which is the sum of the periodic attribute learning losses L_{dir}, L_{pos}, and L_{dis} and the contrastive loss L_{BT}. We refrain from introducing loss tradeoff weights for the $\mathcal{L}_{CKE}$ loss since their impact on subsequent downstream datasets is negligible within a certain range.

4.5 Optimization Objectives

We optimize CROP by minimizing the total loss $\mathcal{L}_{total}$, which consists of the reconstruction loss $\mathcal{L}_{RES}$ and the crystal knowledge-enhanced loss $\mathcal{L}_{CKE}$.

$$\mathcal{L}_{total} = \alpha \times \mathcal{L}_{RES} + (1 - \alpha) \times \mathcal{L}_{CKE}, \tag{6}$$

where hyperparameter α enables balancing the two losses. As α increases, the model places higher emphasis on the core objective of the reconstruction task over the contributions from the enhanced module.

5 Experiments

5.1 Dataset Description

To pre-train the framework, we use a subset [4] of the Open Materials Database (OQMD) [19,23]. The OQMD offers a substantial amount of unlabeled data

that is sufficient for pre-training. The subset was obtained by using JARVIS tools [4], an open-access software package for atomistic data-driven materials computation. To obtain a high-quality subset, we eliminated materials with extreme formation energy, above 4.0 or below -5.0. We also eliminated one crystal structure that could not be loaded. The resulting subset contains **817,139** material structures. Each structure is saved as a CIF file and contains an atom type, atom coordinates, and lattice features. To evaluate the performance of the pre-training framework, we subject it to rigorous testing on eight challenging downstream datasets, i.e., JDFT2D [5], Dielectric [15], Mp_Shear [2], Mp_bulk [2], and KVRH [15], Jarvis_gap [3], Jarvis_ehull [3] and Mp_gap [2]. We primarily use small and medium size datasets to assess model effectiveness. Next, we use larger datasets, e.g., Javis_ehull and Mp_gap, to assess robustness.

5.2 CROP Configurations

The experiments are conducted on a device with NVIDIA TITAN RTX 24GB GPU, and the CROP framework is implemented with Pytorch. The code will be made publicly available upon acceptance. To compare the performance of different configurations of CROP, we use the mean absolute error (MAE) of property estimations. The learning rate used for training is set to 1e-5. The random seed is set to 123. The model parameters are optimized using Adam optimizer [18]. We optimize CROP over 15 epochs, selecting parameters that achieve the lowest loss to ensure maximum model efficacy.

In the fine-tuning stage, we employ the lattice encoder Latt_{ENC} and structure encoder to obtain $\mathbf{h}_L$ and $\mathbf{h}_S$ within CROP, which allows for the sharing of weights obtained from the CROP pre-training with eight downstream datasets. Following this, both max-pooling and mean-pooling operations are applied to the vector $\mathbf{h}_S$. The resulting features, along with $\mathbf{h}_L$, are then concatenated to form a comprehensive feature vector. This vector is input into a specifically designed target head, comprising four fully connected linear layers, which is crucial for crystal property estimation and which is distinct from the CROP pre-training framework. When training models in the fine-tuning stage for the eight downstream datasets, we use two learning rates for optimal performance. For the datasets aggregated from MatBench Suite [9], including JDFT2D, Dielectric, and KVRH datasets, we use the AdamW optimizer with a learning rate of 0.001 for training models that comprise the lattice and structure encoders, effectively tuning their parameters for optimal performance. For the target head, which lacks pre-training parameters from the pre-training model, we use the higher learning rate of 0.005 to effectively train this component. For other datasets, we use a learning rate of 0.01 with the OneCycleLR scheduling strategy.

5.3 Experimental Results

To evaluate the effectiveness of CROP, we compare with both supervised and self-supervised baselines, including SchNet [25], MEGNet [2], ALIGNN [3], Matformer [33], InfoGraph [26], and Crystal Twins [22]. InfoGraph and Crystal

Table 1. Comparison between CROP and baselines in terms of test MAE.

Dataset Size	Small				Medium		Large	
Dataset	JDFT2D	Dielectric	Mp_Shear	Mp_bulk	KVRH	Jarvis_gap	Jarvis_ehull	Mp_gap
# Crystals	636	4,764	5,449	5,450	10,987	18,171	55,370	69,239
SchNet	42.6637	0.3277	0.099	0.066	0.0590	0.43	0.140	0.345
CGCNN	49.2440	0.5988	0.077	0.047	0.0712	0.41	0.170	0.292
MEGNet	54.1719	0.3391	0.099	0.060	0.0668	0.34	0.084	0.307
ALIGNN	43.4244	0.3449	0.078	0.051	0.0568	0.31	0.076	0.218
Matformer	47.6964	0.6817	0.073	0.043	0.0620	0.30	0.064	0.211
InfoGraph	48.5135	0.4684	0.075	0.046	0.0674	0.38	0.128	0.284
Crystal Twins	44.3536	0.4276	0.082	0.050	0.0665	0.39	0.140	0.291
CROP	**37.7532**	**0.3172**	**0.068**	**0.037**	**0.0553**	**0.25**	**0.062**	**0.207**
Relative Improvement	11.51%	3.20%	6.85%	13.95%	2.64%	16.67%	3.13%	1.90%
Inference Time	0.45 s	0.49 s	0.71s	0.70 s	0.87 s	1.23 s	2.68 s	3.02 s

Twins are self-supervised methods, while the others are supervised. In our experiments, we also pre-train the Crystal Twins and InfoGraph models using data from the OQMD dataset.

Table 1 reports the experimental findings for CROP and the baselines on the eight datasets. The decimal precision reported follows previous works [9,33]. To quantify CROP's performance, we report relative improvement over the best baseline approach. The best and the second-best results are in bold and underlined, respectively. Furthermore, Table 2 compares inference times with self-supervised baselines. We observe the following: (1) CROP achieves the best accuracy compared to both supervised and self-supervised baselines. The results show that CROP is effective on small and medium datasets, achieving an increase of at least 2.64% and up to 16.67%. This aligns with our initial intention, which was to address the issue of limited labeled data. For large datasets, simply utilizing their data can lead the model to converge effectively; hence, the benefit of introducing our pre-trained model is not significant. Overall, the results offer evidence of the effectiveness of CROP on datasets with limited labels. (2) CROP outperforms self-supervised baselines, including InfoGraph and Crystal Twins, which do not fully leverage the E(3) and periodic invariances of crystals. Fully considering the E(3) invariance and period invariance of the crystal is effective at improving performance. Next, considering inference times, the findings in Table 2 show that while CROP does not exhibit the fastest inference speed, it achieved very good inference speeds. Considering both performance and speed, CROP emerges as the best choice.

Table 2. The inference time cost

Models	Mp_bulk	Jarvis_gap	Mp_gap
InfoGraph	1.85 s	2.13 s	3.53 s
Crystal Twins	0.55 s	0.65 s	0.91 s
CROP	0.70 s	1.23 s	3.02 s

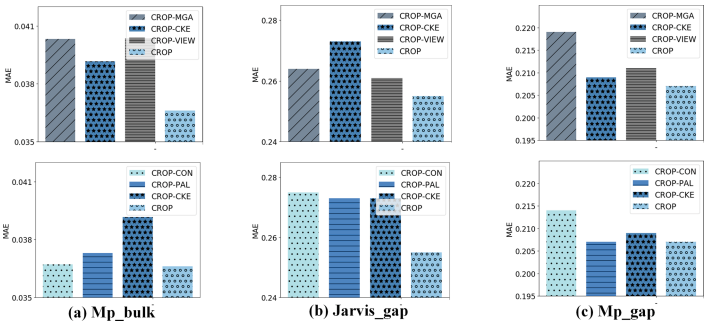

Fig. 6. Ablation studies on (a) Jarvis_gap, (b) Mp_bulk and (c) Mp_gap Datasets.

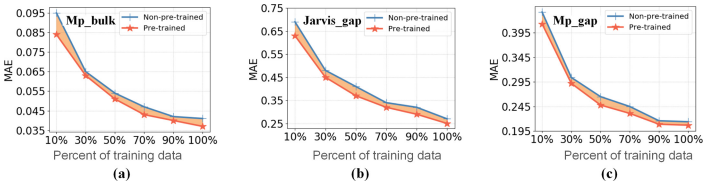

Fig. 7. Effects of Pre-training (a) Mp_bulk, (b) Jarvis_gap and (c) Mp_gap Datasets.

5.4 Ablation Studies

To explore the effectiveness of each component of CROP, we compare different variants of CROP. (1) CROP-MGA: CROP without the multi-graph attention mechanism. (2) CROP-CKE: CROP without the crystal knowledge-enhanced module. (3) CROP-VIEW: A CROP variant that uses only one masked view rather than mutually exclusive masked views. To emphasize the performance differences more distinctly, we retain one additional decimal place in the results of the ablation experiments. We employ three dataset types: small, medium, and large, i.e., Mp_bulk, Jarvis_gap, and Mp_gap, respectively. The findings on the top side of Fig. 6 cause the following observations: First, our attention mechanism and mutually exclusive views are crucial for improving performance. Without these components, performance drops across the downstream tasks. Second, removing the CKE module yields a significant performance drop on the Mp_bulk and Jarvis_gap datasets. However, the performance change is not substantial for the large dataset Mp_gap.

To understand better the role of CKE, we consider two additional CROP variants: (1) CROP-Con omits the atom-type contrastive learning in CKE. (2) CROP-PAL omits the periodic attribute learning component in the CKE. The findings reported on the bottom side of Fig. 6 cause the following observations: First, on the medium and small datasets, i.e., Mp_bulk and Jarvis_gap, CROP achieves superior performance. This indicates that both atom-type contrastive learning and periodic attribute learning contribute to model performance. Second, for the large dataset Mp_gap, the introduction of periodic attribute learn-

ing in CKE has minimal impact. In contrast, the introduction of atom-type contrastive learning significantly affects model performance. This highlights the importance of interacting with features under mutually exclusive views through atom-type contrastive learning.

Overall, these findings validate the effectiveness of CROP, particularly in enhancing performance for downstream tasks with limited data, addressing the challenges caused by limited labeled data.

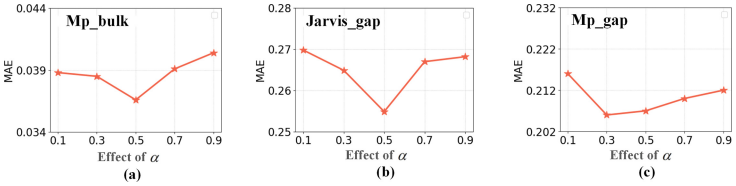

Fig. 8. Effects of α (a) Mp_bulk, (b) Jarvis_gap and (c) Mp_gap Datasets.

5.5 Parameter Sensitivity Analysis

Here, we consider the impact of varying training dataset proportions and hyperparameter settings. (1) we use different training data proportions in the downstream datasets to validate the advantages of the pre-trained model. Specifically, we use 10%, 30%, 50%, 70%, 90%, and 100% of the training samples in the training process, as shown in Fig. 7. When the training data proportion exceeds 50%, our method outperforms most of the baseline approaches, underscoring the efficacy of CROP. Moreover, we further compare the performance of training downstream tasks with our pre-training framework versus without a pre-training framework. These findings show that using the pre-trained model improves performance over not using it, especially with limited labeled data. (2) we vary the loss tradeoff coefficient α over $\{0.1, 0.3, 0.5, 0.7, 1, 2\}$. This coefficient controls the importance assigned to the crystal reconstruction and the crystal knowledge enhancement module. As shown in Fig. 8, as α increases, the performance of the model first increases and then decreases. As α increases, the model gets less guidance from the crystal knowledge-enhanced module. On Mp_bulk and Jarvis_gap, $\alpha = 0.5$ is optimal. On Mp_gap, $\alpha = 0.3$ is best. This justifies the choice of $\alpha = 0.5$ for all datasets.

6 Conclusion

We propose a crystal knowledge-enhanced pre-training framework called CROP that is capable of exploiting mutually exclusive masked views for learning crystal representations with self-supervision. CROP is designed to tackle the challenge of limited labeled data while being able to exploit better the physical principles that crystals obey. The masking strategy enables the learning of atom

representations under two mutually exclusive views that consider surrounding atoms and their relationships, rather than learning individual atom representations. CROP's multi-graph attention module can enhance the expressiveness of learned representations by leveraging the knowledge of periodic invariants. CROP's crystal knowledge-enhanced module introduces crystal principles explicitly. An experimental study offers evidence that CROP improves crystal property estimation over strong baselines.

References

1. Campos, D., et al.: Unsupervised time series outlier detection with diversity-driven convolutional ensembles. Proc. VLDB Endow **15**(3), 611–623 (2022)
2. Chen, C., Ye, W., Zuo, Y., Zheng, C., Ong, S.P.: Graph networks as a universal machine learning framework for molecules and crystals. Chem. Mater. **31**(9), 3564–3572 (2019)
3. Choudhary, K., DeCost, B.: Atomistic line graph neural network for improved materials property predictions. NPJ Comput. Mater. **7**(1), 185 (2021)
4. Choudhary, K., et al.: The joint automated repository for various integrated simulations (Jarvis) for data-driven materials design. NPJ Comput. Mater. **6**(1), 1–13 (2020)
5. Choudhary, K., Kalish, I., Beams, R., Tavazza, F.: High-throughput identification and characterization of two-dimensional materials using density functional theory. Sci. Rep. **7**(1), 1–16 (2017)
6. Coley, C.W., et al.: A graph-convolutional neural network model for the prediction of chemical reactivity. Chem. Sci. **10**(2), 370–377 (2019)
7. Curtarolo, S., Hart, G.L., Nardelli, M.B., Mingo, N., Sanvito, S., Levy, O.: The high-throughput highway to computational materials design. Nat. Mater. **12**(3), 191–201 (2013)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Dunn, A., Wang, Q., Ganose, A., Dopp, D., Jain, A.: Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. NPJ Comput. Mater. **6**(1), 138 (2020)
10. Gardner, M.W., Dorling, S.: Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. Atmos. Environ. **32**(14–15), 2627–2636 (1998)
11. Gasteiger, J., Giri, S., Margraf, J.T., Günnemann, S.: Fast and uncertainty-aware directional message passing for non-equilibrium molecules. arXiv preprint arXiv:2011.14115 (2020)
12. Grosse-Kunstleve, R.W., Sauter, N.K., Adams, P.D.: Numerically stable algorithms for the computation of reduced unit cells. Acta Crystallogr. A **60**(1), 1–6 (2004)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of CVPR, pp. 16000–16009 (2022)
14. Hou, Z., et al.: Graphmae: Self-supervised masked graph autoencoders. In: Proceedings of SIGKDD, pp. 594–604 (2022)
15. Jain, A., et al.: The materials project: a materials genome approach to accelerating materials innovation, Apl mater (2013)

16. Kieu, T., et al.: Anomaly detection in time series with robust variational quasi-recurrent autoencoders. In: Proceedings of ICDE, pp. 1342–1354. IEEE (2022)
17. Kieu, T., et al.: Robust and explainable autoencoders for unsupervised time series outlier detection. In: Proceedings of ICDE, pp. 3038–3050. IEEE (2022)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kirklin, S., et al.: The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. NPJ Comput. Mater. **1**(1), 1–15 (2015)
20. Kittel, C., McEuen, P.: Introduction to solid state physics. John Wiley & Sons (2018)
21. Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., Tang, J.: Pre-training molecular graph representation with 3D geometry. In Proc, ICLR (2022)
22. Magar, R., Wang, Y., Barati Farimani, A.: Crystal twins: self-supervised learning for crystalline material property prediction. NPJ Comput. Mater. **8**(1), 231 (2022)
23. Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., Wolverton, C.: Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). JOM **65**(11), 1501–1509 (2013)
24. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. Neural Netw. **20**(1), 61–80 (2008)
25. Schütt, K., Kindermans, P.J., Sauceda Felix, H.E., Chmiela, S., Tkatchenko, A., Müller, K.R.: Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. In: Proceedings of NeurIPS (2017)
26. Sun, F.Y., Hoffman, J., Verma, V., Tang, J.: Infograph: unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In: Proceedings of ICLR (2020)
27. Unke, O.T., Meuwly, M.: Physnet: a neural network for predicting energies, forces, dipole moments, and partial charges. J. Chem. Theory Comput. **15**(6), 3678–3693 (2019)
28. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
29. Wei, J., et al.: Machine learning in materials science. InfoMat **1**(3), 338–358 (2019)
30. Xie, T., Fu, X., Ganea, O.E., Barzilay, R., Jaakkola, T.S.: Crystal diffusion variational autoencoder for periodic material generation. In: Proceedings of ICLR (2022)
31. Xie, T., Grossman, J.C.: Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett. **120**(14), 145301 (2018)
32. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)
33. Yan, K., Liu, Y., Lin, Y., Ji, S.: Periodic graph transformers for crystal material property prediction. In: Proceedings of NeurIPS, pp. 15066–15080 (2022)
34. Yang, S.B., Guo, C., Hu, J., Tang, J., Yang, B.: Unsupervised path representation learning with curriculum negative sampling. In: Proceedings of IJCAI, pp. 3286–3292 (2021)
35. Yang, S.B., Guo, C., Yang, B.: Context-aware path ranking in road networks. IEEE Trans. Knowl. Data Eng. **34**(7), 3153–3168 (2022)
36. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: self-supervised learning via redundancy reduction. In: Proceedings of ICML, pp. 12310–12320 (2021)