

## EXPRESS LETTER

## Universal Machine Learning Kohn–Sham Hamiltonian for Materials

To cite this article: Yang Zhong *et al* 2024 *Chinese Phys. Lett.* **41** 077103

View the [article online](#) for updates and enhancements.

## You may also like

- [Band Structures of Ultrathin Bi\(110\) Films on Black Phosphorus Substrates Using Angle-Resolved Photoemission Spectroscopy](#)  
Sailong Ju, , Maokun Wu et al.
- [Normal and Superconducting Properties of  \$\text{La}\_2\text{Ni}\_2\text{O}\_7\$](#)   
Meng Wang, , Hai-Hu Wen et al.
- [Magneto-Transport and Shubnikov–de Haas Oscillations in the Type-II Weyl Semimetal Candidate  \$\text{NbIrTe}\_4\$  Flake](#)  
Xiang-Wei Huang, , Xiao-Xiong Liu et al.

## Universal Machine Learning Kohn–Sham Hamiltonian for Materials

Yang Zhong(钟阳)<sup>1,2</sup>, Hongyu Yu(于宏宇)<sup>1,2</sup>, Jihui Yang(杨吉辉)<sup>1,2</sup>, Xingyu Guo(郭星宇)<sup>1,2</sup>,  
Hongjun Xiang(向红军)<sup>1,2\*</sup>, and Xingao Gong(龚新高)<sup>1,2</sup>

<sup>1</sup>Key Laboratory of Computational Physical Sciences (Ministry of Education), Institute of Computational Physical Sciences, State Key Laboratory of Surface Physics, and Department of Physics,  
Fudan University, Shanghai 200433, China

<sup>2</sup>Shanghai Qi Zhi Institute, Shanghai 200030, China

(Received 4 June 2024; accepted manuscript online 15 June 2024)

While density functional theory (DFT) serves as a prevalent computational approach in electronic structure calculations, its computational demands and scalability limitations persist. Recently, leveraging neural networks to parameterize the Kohn–Sham DFT Hamiltonian has emerged as a promising avenue for accelerating electronic structure computations. Despite advancements, challenges such as the necessity for computing extensive DFT training data to explore each new system and the complexity of establishing accurate machine learning models for multi-elemental materials still exist. Addressing these hurdles, this study introduces a universal electronic Hamiltonian model trained on Hamiltonian matrices obtained from first-principles DFT calculations of nearly all crystal structures on the Materials Project. We demonstrate its generality in predicting electronic structures across the whole periodic table, including complex multi-elemental systems, solid-state electrolytes, Moiré twisted bilayer heterostructure, and metal-organic frameworks. Moreover, we utilize the universal model to conduct high-throughput calculations of electronic structures for crystals in GNoME datasets, identifying 3940 crystals with direct band gaps and 5109 crystals with flat bands. By offering a reliable efficient framework for computing electronic properties, this universal Hamiltonian model lays the groundwork for advancements in diverse fields, such as easily providing a huge data set of electronic structures and also making the materials design across the whole periodic table possible.

DOI: 10.1088/0256-307X/41/7/077103

Electronic structures<sup>[1–5]</sup> of materials are crucial in understanding and predicting a wide range of physical properties, including electrical conductivity, optical behavior, mechanical strength, chemical reactivity, and magnetic characteristics. Electronic structure calculations provide insights into the electronic band structures, bonding, and reactivity, enabling designs of new materials and research of chemical reactions. Among diverse quantum mechanics approaches, density functional theory (DFT)<sup>[6–9]</sup> has become a widely used computational method in electronic structure calculations since DFT has drastically reduced the computational cost by employing the electron density instead of the many-body wave function as the fundamental variable. The price to pay for using the electron density as the fundamental variable is that the Kohn–Sham DFT Hamiltonian is no longer a simple explicit function of the atomic structure, but should be obtained by solving a self-consistent Kohn–Sham equation. However, the computational cost of DFT self-consistent field (SCF) cycles remains expensive for large systems.

In recent years, the use of neural networks to parameterize the DFT Hamiltonian has emerged as an important and effective method to accelerate electronic structure calculations.<sup>[10–22]</sup> The machine learning (ML) Hamiltonian models offer a significant advantage by

providing a direct mapping from the structure to the self-consistent Hamiltonian matrix, eliminating the need for time-consuming self-consistent iterations typically required in Kohn–Sham DFT. Initially, Hegde and Bowen proposed a kernel ridge regression (KRR) model and successfully applied it to fit the empirical Hamiltonian matrix of the cubic Cu crystal.<sup>[10]</sup> Subsequently, Unke<sup>[14]</sup> proposed the PhiSNet model, and Schütt *et al.*<sup>[12]</sup> introduced the SchNOrb model, both of which demonstrated remarkable performance in accurately fitting the Hamiltonian matrices of various small organic molecules such as water, ethanol, and uracil. In addition, some other researchers have also made important contributions. For instance, Nigam *et al.*<sup>[18]</sup> used the Gaussian process regression (GPR) model to fit the Hamiltonian matrices of water and benzene molecules, while Zhang *et al.*<sup>[20]</sup> successfully fitted the Hamiltonian matrix of aluminum using the atomic cluster expansion (ACE) model. The authors of Refs. [17,21] proposed the graph neural network (GNN)-based DeepH model, which can be used to predict the Hamiltonian matrices of crystals like graphene and MoS<sub>2</sub>.

Despite these advancements, there still exist several challenges in utilizing machine learning for electronic Hamiltonian prediction. Firstly, exploring new systems necessitates retraining a completely new model specifically

\*Corresponding author. Email: hxiang@fudan.edu.cn

© 2024 Chinese Physical Society and IOP Publishing Ltd

tailored for that system, which currently lacks automation and can prove time-consuming and computationally expensive, thereby constraining the practical applicability of this approach. Secondly, many practical materials, such as high-entropy alloys<sup>[23–25]</sup> and ceramics,<sup>[26,27]</sup> are composed of a large number of elements, posing challenges in establishing accurate machine-learning models for their electronic Hamiltonians due to the requirement of a substantial amount of DFT training data. Current research endeavors have predominantly concentrated on systems comprising no more than three elements, and surmounting this obstacle to encompass more intricate systems remains a significant challenge in the field. Overcoming these challenges is crucial to enable the broader application of machine learning in predicting electronic Hamiltonians for diverse and multi-elemental materials. Recently, several groups reported developments of universal machine learning interatomic potentials (MLIPs),<sup>[28–33]</sup> which can handle almost all elements across the whole periodic table. Naturally, one may wonder whether it is possible to develop a universal machine-learning model for electronic Hamiltonians. Such a model would enable efficient and accurate calculations of electronic structures for a wide range of materials and large-scale systems. Recently, we have proposed a transferable Hamiltonian graph neural network (HamGNN)<sup>[22]</sup> that enables the prediction of Hamiltonian matrices for various structures within a chemical space comprising several specific elements, such as SiO<sub>2</sub> isomers and Bi<sub>x</sub>Se<sub>y</sub> compounds with varying stoichiometric ratios. Compared to the above-mentioned models, the excellent transferability of the equivariant GNN-based HamGNN framework makes it the most feasible candidate for constructing a universal Hamiltonian model across the entire periodic table. Unlike the construction of universal MLIPs, achieving a universal Hamiltonian model is not as straightforward and cannot simply rely on training with large-scale datasets due to the high dimensionality and inherent complexity of the Hamiltonian matrix. The training process also plays a crucial role in achieving the universality of the Hamiltonian model.

In this work, we propose a universal Kohn–Sham Hamiltonian model in the sense that this single model is applicable to all elements of the entire periodic table and all structures of a given chemical composition. By employing the HamGNN model, we develop a universal Hamiltonian model via a “two-step training procedure” utilizing the Hamiltonian matrices of 55000 structures obtained from the Materials Project<sup>[34,35]</sup> as our training dataset. The universality of our Hamiltonian model is demonstrated by its successful prediction of the electronic structures for various bulk or low-dimensional materials with different combinations of chemical elements in the periodic table. The universal model successfully captures not only common systems but also those containing uncommon or rare transition metal elements. Furthermore, it proves high accuracy even in complex multi-element systems comprising more than five elements. By providing a reliable framework for understanding electronic properties across the periodic table, this research paves the way

for advancements in material design, catalysis, electronics, and other fields that heavily rely on efficient predictions of electronic structures.

*Framework of the Universal Electronic Hamiltonian Model.* Constructing an ML model for the electronic Hamiltonian is more complicated than constructing machine learning interatomic potentials (MLIPs).<sup>[36–41]</sup> MLIPs provide a mapping from the two degrees of freedom, atomic types  $\{Z_i\}$  and atomic positions  $\{\tau_i\}$ , to the scalar potential energy  $E$ . In addition to the two degrees of freedom, the mapping from a crystal structure to the electronic Hamiltonian matrix also necessitates handling the supplementary degree of freedom arising from distinct atomic orbital bases  $\{\phi_{i\alpha}\}$  associated with each atom. As the number of elements in the system increases, the relevant degrees of freedom and interactions in the electronic Hamiltonian matrix also increase dramatically, leading to a significant increase in complexity when fitting the electronic Hamiltonian matrix. Therefore, a universal electronic Hamiltonian model across the periodic table requires much more network capacity than the MLIPs to capture all degrees of freedom accurately. For large molecules or crystals, especially in the cases where the basis set is large and the system is complex, the dimension of the electronic Hamiltonian matrix can be extremely large. This increases the difficulty of model training, as it requires handling large-scale matrix and storing a significant amount of parameters. In addition, different sub-blocks of a Hamiltonian matrix are subject to different equivariant constraints under a rotation operation, so the Hamiltonian matrix predicted by the model must also adhere to these constraints.

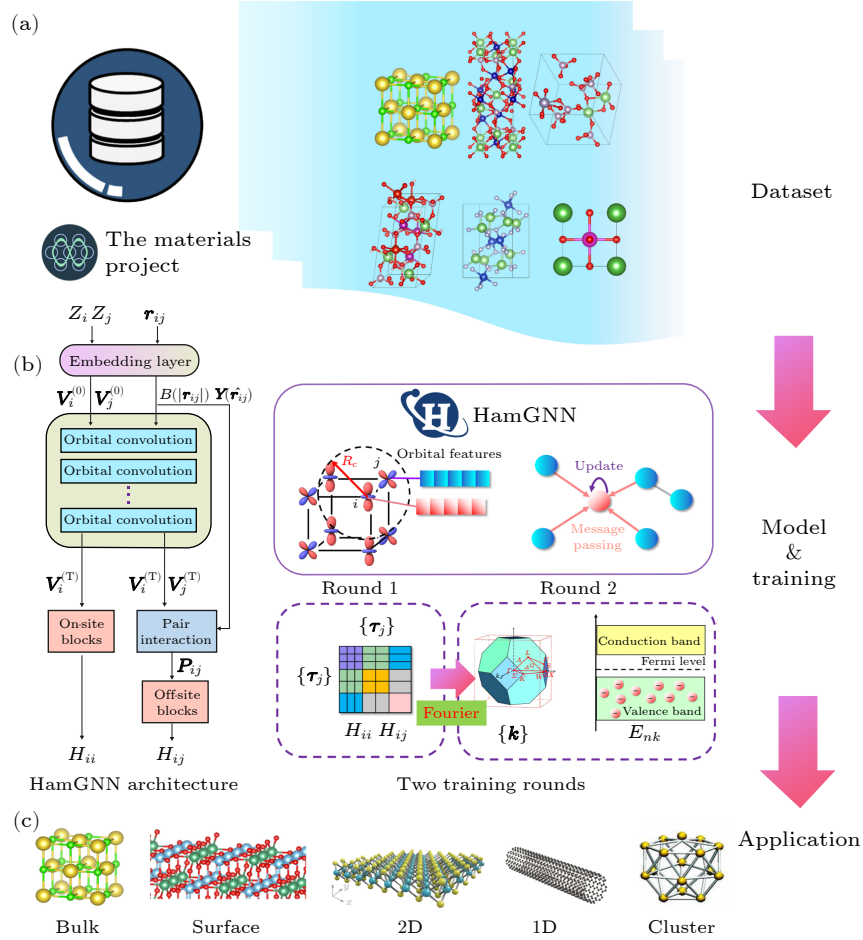
We have trained a universal Hamiltonian model following the process shown in Fig. 1. To develop a universal Hamiltonian model for the whole periodic table, we utilized one of the world’s largest open databases for DFT-relaxed crystal structures, namely the Materials Project.<sup>[34,35]</sup> We used OpenMX,<sup>[42,43]</sup> a DFT software package based on norm-conserving pseudopotentials and pseudo-atomic localized basis functions, to calculate the Hamiltonian matrices of approximately 55000 structures on the Materials Project. Among them, approximately 44000 structures’ Hamiltonian matrices were used as a training set, while validation and test sets were created using the Hamiltonian matrices of around 5500 structures each. Then, we use these datasets to train a universal HamGNN<sup>[22]</sup> model. HamGNN is a deep learning model based on equivariant graph neural networks, which can automatically learn the features of each element on the entire periodic table without any prior physical or chemical properties of elements. The architecture of the HamGNN model and the training process are shown in Fig. 1(b). We will briefly introduce the principle of HamGNN for predicting the Hamiltonian matrix in the following paragraph, and for more network details see Ref. [22].

Since the irreducible representations with rotation orders  $l = 0, 1, 2, \dots$  of the  $O(3)$  group possess the same rotational equivariance and parity symmetry as the atomic orbitals  $s, p, d, \dots$ , HamGNN uses a direct sum of equiv-

ariant atomic features  $\mathbf{V}_l$  with different rotation orders  $l$  to characterize each atom:  $\mathbf{V} = \mathbf{V}_0 \oplus \mathbf{V}_1 \oplus \cdots \oplus \mathbf{V}_{l_{\max}}$ . This feature tensor satisfies the rotational equivariance under the rotation operation  $\hat{R}$ :  $\mathbf{V}_l(\hat{R} \cdot (\mathbf{r}_1, \dots, \mathbf{r}_N)) = \mathbf{D}_l(\hat{R})\mathbf{V}_l(\mathbf{r}_1, \dots, \mathbf{r}_N)$ , where  $\mathbf{D}_l$  ( $l < l_{\max}$ ) is a Wigner  $D$  matrix<sup>[44,45]</sup> of order  $l$ . HamGNN updates the equivariant atomic features through an equivariant message-passing function in the orbital convolution layer. After  $T$  orbital convolution layers, the atomic features are transformed

into on-site Hamiltonian matrices by an “on-site layer”. HamGNN merges the features of atom pairs  $ij$  into the edge features  $\mathbf{P}_{ij}$  in the “pair interaction layer”. The edge features  $\mathbf{P}_{ij}$  is later transformed into an off-site Hamiltonian matrix through an “off-site layer”. Each subblock of the Hamiltonian matrix can be decomposed into a set of  $O(3)$  equivariant irreducible spherical tensors (ISTs) according to the equation<sup>[44–47]</sup>

$$l_i \otimes l_j = |l_i - l_j| \oplus |l_i - l_j| + 1 \oplus \cdots \oplus l_i + l_j. \quad (1)$$



**Fig. 1.** Framework of a universal Hamiltonian model based on HamGNN. (a) Training dataset preparation. The training dataset is generated by calculating the real-space Hamiltonian matrices of crystal structures available on the Materials Project using an *ab initio* tight-binding software based on numerical atomic orbitals. (b) Model architecture and the training procedure. This dataset is utilized for training the HamGNN model, a deep learning approach that employs equivariant graph neural networks to predict the Hamiltonian matrix. The model can automatically learn the intrinsic features of each element on the periodic table solely based on their atomic numbers, without relying on any prior physical or chemical properties. In HamGNN, the orbital features of the central atom are updated by considering interactions between the neighboring atoms within a cutoff radius of  $R_c$ . For atomic pairs beyond this cutoff radius, multi-layer message passing is employed to exchange orbital features. To achieve universality, HamGNN requires two rounds of training. In the first round, the loss function for network training solely considers the error in the real-space Hamiltonian matrix. After this initial round of training, this model can accurately predict the real-space Hamiltonian matrices with high precision. In the second round of training, the real-space Hamiltonian matrices are transformed into the reciprocal Hamiltonian matrices at randomly selected  $\mathbf{k}$  points in the Brillouin zone, and the errors of the orbital energies near the Fermi level obtained by diagonalizing the reciprocal Hamiltonian matrices are incorporated into the total loss function. (c) Applications of the universal HamGNN model. After two rounds of training, the universality of the HamGNN model has significantly improved, enabling accurate prediction of electronic structures in crystals with arbitrary periodic boundary conditions and any components.

The on-site and off-site layers output each sub-block of the Hamiltonian matrix by the following equation:

$$H_{l_i m_i, l_j m_j} = \sum_{l=|l_i-l_j|}^{l_i+l_j} \sum_{m=-l}^l C_{m_i, m_j, m}^{l_i, l_j, l} T_m^l, \quad (2)$$

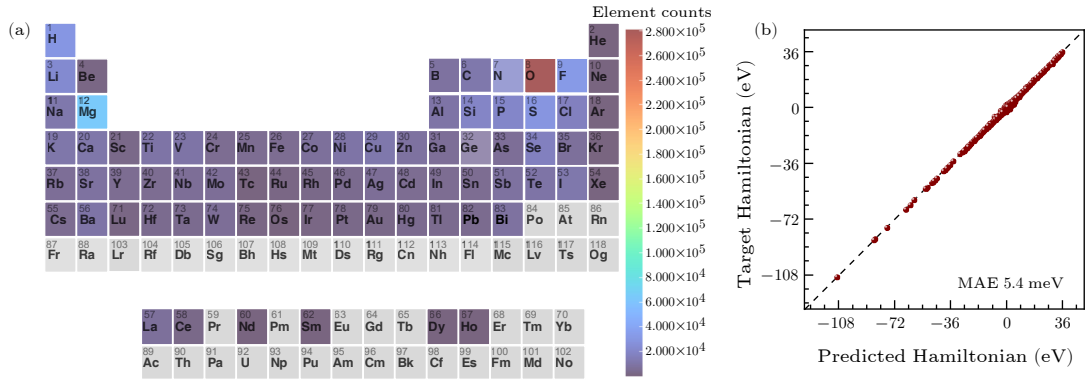
where  $T_m^l$  is an equivariant IST with rotation order  $l$  in  $\mathbf{V}_i^{(T)}$  or  $\mathbf{P}_{ij}$ ,  $C_{m_i, m_j, m}^{l_i, l_j, l}$  is the Clebsch–Gordan coefficient.

To achieve a universal model, HamGNN needs to undergo two rounds of training. During the first round of training, only the error of the real-space Hamiltonian matrix is considered as the loss function for the network's training. After the first round of training, the model is capable of reasonably predicting the real-space Hamiltonian matrix, which lays a good foundation for obtaining accurate band structures in subsequent tasks. To improve the accuracy of the model in predicting the eigenvalues of Bloch states, which constitutes the primary objective of our Hamiltonian model, we further applied fine-tuning techniques for a second round of training. Fine tuning in

neural networks has shown a crucial role in training recently developed large language models to optimize their adaptation to specific tasks or domains.<sup>[48,49]</sup> The results of our tests indicate that the two training rounds are necessary to obtain a truly universal Hamiltonian model. During each training step in the second round, Fourier transformations are performed on the predicted and target real-space Hamiltonian matrices at randomly selected  $\mathbf{k}$  points in the Brillouin zone. The orbital energies  $\varepsilon_{nk}$  are obtained by diagonalizing the reciprocal Hamiltonian matrices, and the error of the orbital energies near the Fermi level is incorporated into the loss function as follows:

$$L = \|\tilde{H} - H\| + \frac{\lambda}{N_{\text{orb}} \times N_k} \sum_{k=1}^{N_k} \sum_{n=1}^{N_{\text{orb}}} \|\tilde{\varepsilon}_{nk} - \varepsilon_{nk}\|, \quad (3)$$

where the variables marked with a tilde refer to the corresponding predictions and  $\lambda$  denotes the loss weight of the orbital energy error.  $N_{\text{orb}}$  is the number of orbits selected near the Fermi level,  $N_k$  is the number of the random  $\mathbf{k}$  points generated in each training step.



**Fig. 2.** (a) Element distribution of the training dataset. (b) Comparison of the Hamiltonian calculated by HamGNN and OpenMX for the test dataset.

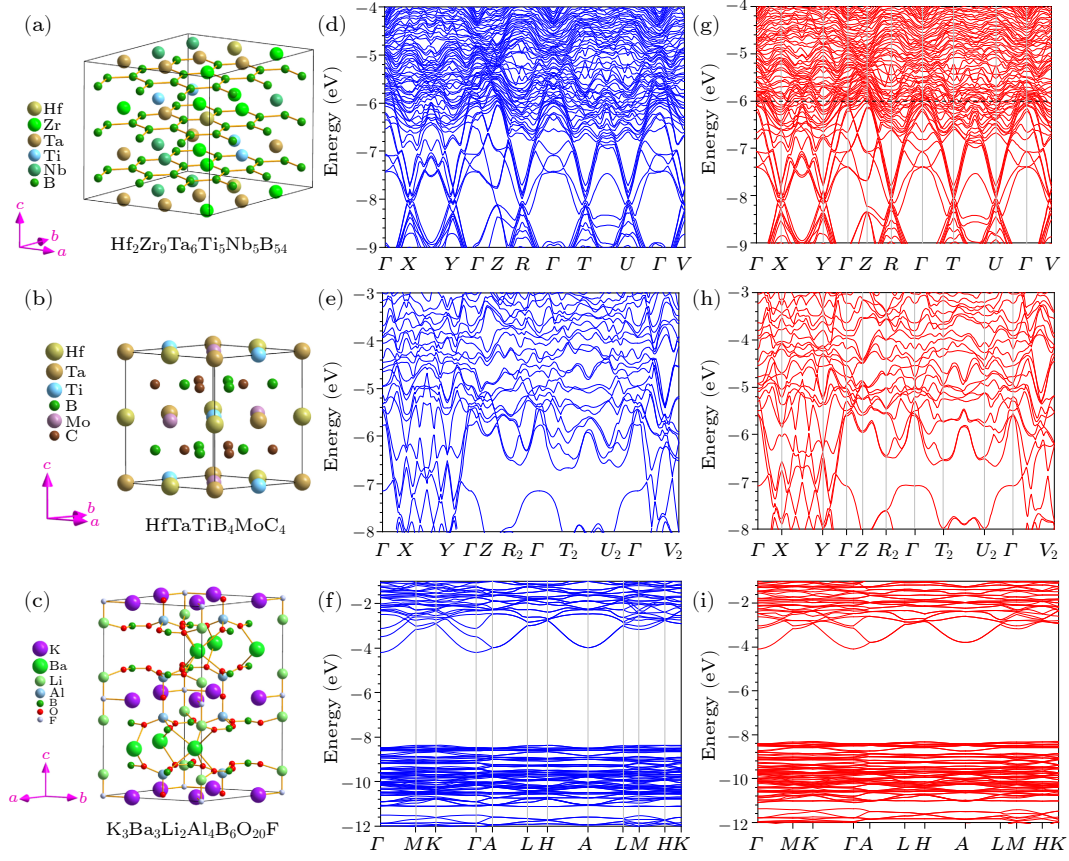
Figure 2(a) displays the count of each chemical element in the training set for the Hamiltonian. The metallic elements of groups IA and IIA, as well as the non-metallic elements of groups IVA, VA, VIA, and VIIA, have the highest proportion in the training set. Except for some less common transition metal elements that are not included in the training set, this dataset includes all elements supported by OpenMX's PBE pseudopotential library. After the first round of training, the HamGNN model achieved a mean absolute error (MAE) of only 5.4 meV for the real-space Hamiltonian matrix on the test set. The accuracy of the model for the real-space Hamiltonian matrix is shown in Fig. 2(b). In the second round of training, we incorporate the error of orbital energies at five randomly selected  $\mathbf{k}$ -points in the Brillouin zone into the loss function to restart the training, with a  $\lambda$  value set at 0.01. After the second round of model fine-tuning, the generalization ability of the HamGNN model has been significantly improved. The model's high accuracy on the test set is evident from its predictions of energy bands and Fermi surfaces for several crystal structures in the test set (see Discussion 1 in the Supplementary Material). By addressing potential over-

fitting issues, this model demonstrates better adaptability to different datasets and real-world scenarios, showcasing enhanced universality and stability. In the following discussions, we will evaluate HamGNN's prediction accuracy for more complex crystals across the entire periodic table.

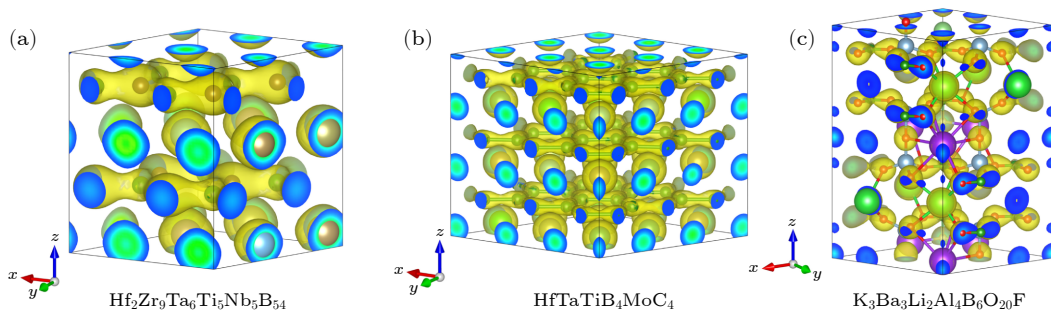
**Tests on Multi-Element Materials.** We are now applying our universal Hamiltonian model to systems that are not included in the Materials Project to verify the generality and accuracy of our model. Previous ML Hamiltonian models<sup>[10,15,20]</sup> typically handle crystal structures composed of only 1–3 elements, and training an accurate model that can deal with complex crystal structures containing more elements is challenging. The training datasets commonly used by these models are built from structures perturbed using molecular dynamics. However, as the number of atomic species in the crystal increases, the degrees of freedom of the Hamiltonian matrix increase sharply. Consequently, the training samples generated by perturbing a seed structure cannot fully cover the entire configuration space of the crystal. In these cases, the perturbed structures often contain many similar and repetitive patterns, which can cause the Hamiltonian model to be trapped in

local minima and fail to accurately fit the Hamiltonian of crystals with multiple elements and complex configurations. However, the universal Hamiltonian model can effectively address such concerns. Through extensive train-

ing on a comprehensive and diverse dataset, the universal Hamiltonian model develops a profound understanding of the intricate interactions among atoms in various configurations.



**Fig. 3.** Crystal structures of (a)  $\text{Hf}_2\text{Zr}_9\text{Ta}_6\text{Ti}_5\text{Nb}_5\text{B}_{54}$ , (b)  $\text{HfTaTiB}_4\text{MoC}_4$ , and (c)  $\text{K}_3\text{Ba}_3\text{Li}_2\text{Al}_4\text{B}_6\text{O}_{20}\text{F}$ . Predicted energy bands of (d)  $\text{Hf}_2\text{Zr}_9\text{Ta}_6\text{Ti}_5\text{Nb}_5\text{B}_{54}$ , (e)  $\text{HfTaTiB}_4\text{MoC}_4$ , and (f)  $\text{K}_3\text{Ba}_3\text{Li}_2\text{Al}_4\text{B}_6\text{O}_{20}\text{F}$ . DFT calculated energy bands of (g)  $\text{Hf}_2\text{Zr}_9\text{Ta}_6\text{Ti}_5\text{Nb}_5\text{B}_{54}$ , (h)  $\text{HfTaTiB}_4\text{MoC}_4$ , and (i)  $\text{K}_3\text{Ba}_3\text{Li}_2\text{Al}_4\text{B}_6\text{O}_{20}\text{F}$ .



**Fig. 4.** Predicted charge density of (a)  $\text{Hf}_2\text{Zr}_9\text{Ta}_6\text{Ti}_5\text{Nb}_5\text{B}_{54}$ , (b)  $\text{HfTaTiB}_4\text{MoC}_4$ , and (c)  $\text{K}_3\text{Ba}_3\text{Li}_2\text{Al}_4\text{B}_6\text{O}_{20}\text{F}$ .

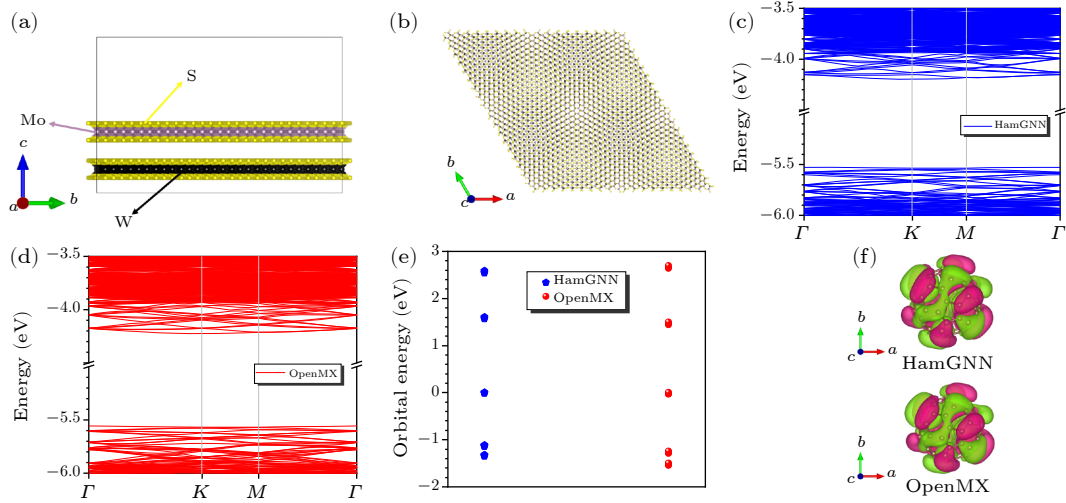
The universal HamGNN's generalization performance and accuracy were evaluated by conducting tests on three different crystal structures:  $\text{Hf}_2\text{Zr}_9\text{Ta}_6\text{Ti}_5\text{Nb}_5\text{B}_{54}$ ,  $\text{HfTaTiB}_4\text{MoC}_4$ , and  $\text{K}_3\text{Ba}_3\text{Li}_2\text{Al}_4\text{B}_6\text{O}_{20}\text{F}$ . These crystals were specifically chosen to represent a diverse range of compositions and structural complexities.<sup>[50–52]</sup> The compound  $\text{Hf}_2\text{Zr}_9\text{Ta}_6\text{Ti}_5\text{Nb}_5\text{B}_{54}$  exhibits a hexagonal  $\omega$ -phase derived structure and belongs to the  $P1$  space group of the triclinic crystal system. It consists of five

distinct metal ions, which are inserted randomly into the gaps between the hexagonal monolayers of boron.  $\text{HfTaTiB}_4\text{MoC}_4$  crystallizes in the triclinic  $P1$  space group, with the metal elements inserted into the gaps between the hexagonal monolayers composed of boron and carbon.  $\text{K}_3\text{Ba}_3\text{Li}_2\text{Al}_4\text{B}_6\text{O}_{20}\text{F}$  is a deep ultraviolet transparent non-linear optical crystal, composed of seven elements and possessing a large bandgap.<sup>[53]</sup>

To assess the generalization performance of HamGNN,

a comparison was made between the predicted energy bands generated by the model and the calculated energy bands by OpenMX for each of these crystals, as shown in Fig. 3. By examining Fig. 3, it becomes evident that HamGNN demonstrates remarkable generalizability in predicting the energy bands for the three crystal structures. The charge densities of these three crystals were also obtained using the predicted Hamiltonian matrix, as shown in Fig. 4. The predicted charge densities exhibit excellent agreement with the corresponding DFT-calculated charge densities (shown in Fig. S6 of the Supplementary Material). This further confirms that HamGNN is capable of accurately capturing not only energy bands but also charge distribution within these materials. Furthermore, to test its versatility and applicability on practi-

cal materials, we extended our analysis to a sulfide solid electrolyte with a composition of  $\text{Li}_{10}\text{Si}_{1.5}\text{P}_{1.5}\text{S}_{11.5}\text{Cl}_{0.5}$ , a highly disordered system consisting of 7200 atoms. Our universal model shows that this disordered system is insulating with a band gap of 1.4 eV, suggesting that it indeed has excellent electric insulating properties and may serve as a promising solid electrolyte (see Supplementary Discussion 2 for details). These successful evaluations highlight the effectiveness and reliability of HamGNN as a universal model for predicting electronic structures across various complex crystal systems. Its ability to generalize well across different compositions and structural complexities makes it an invaluable tool in high-throughput electronic structure calculations for the periodic table.



**Fig. 5.** Prediction results of the universal HamGNN model for the bilayer  $\text{MoS}_2/\text{WS}_2$  heterostructure with a twist angle of  $3.5^\circ$  and the  $\text{C}_{60}$  cluster. (a) Side view of the bilayer  $\text{MoS}_2/\text{WS}_2$  heterostructure. (b) Top view of the bilayer  $\text{MoS}_2/\text{WS}_2$  heterostructure. (c) Predicted band structure of the bilayer  $\text{MoS}_2/\text{WS}_2$  heterostructure. (d) Band structure of the bilayer  $\text{MoS}_2/\text{WS}_2$  heterostructure calculated by OpenMX. (e) The HamGNN-predicted and OpenMX-calculated orbital energies for the  $\text{C}_{60}$  cluster. (f) The HamGNN-predicted and OpenMX-calculated wavefunction for the highest occupied molecular orbital (HOMO) of the  $\text{C}_{60}$  cluster.

**Tests on Low-Dimensional Materials.** The above discussions have demonstrated the accuracy of the universal HamGNN model on bulk materials. Now, we further explore its generalizability and prediction accuracy in the field of low-dimensional materials. We constructed a two-dimensional heterostructure consisting of  $\text{MoS}_2/\text{WS}_2$  with a twist angle of  $3.5^\circ$ , as shown in Figs. 5(a) and 5(b). This structure comprises 1625 atoms and exhibits a higher level of complexity compared to the bulk structures typically included in our training set. The universal HamGNN model effectively captures the interatomic interactions within the bilayer  $\text{MoS}_2/\text{WS}_2$  heterostructure and demonstrates excellent agreement between the energy bands predicted by the universal HamGNN model [Fig. 5(c)] and those calculated by OpenMX [Fig. 5(d)]. The universal HamGNN model was further tested on the  $\text{C}_{60}$  cluster. Figure 5(e) demonstrates a good alignment between the predicted energy level of the  $\text{C}_{60}$  cluster near the band gap and the results obtained through DFT calculations, while Fig. 5(f)

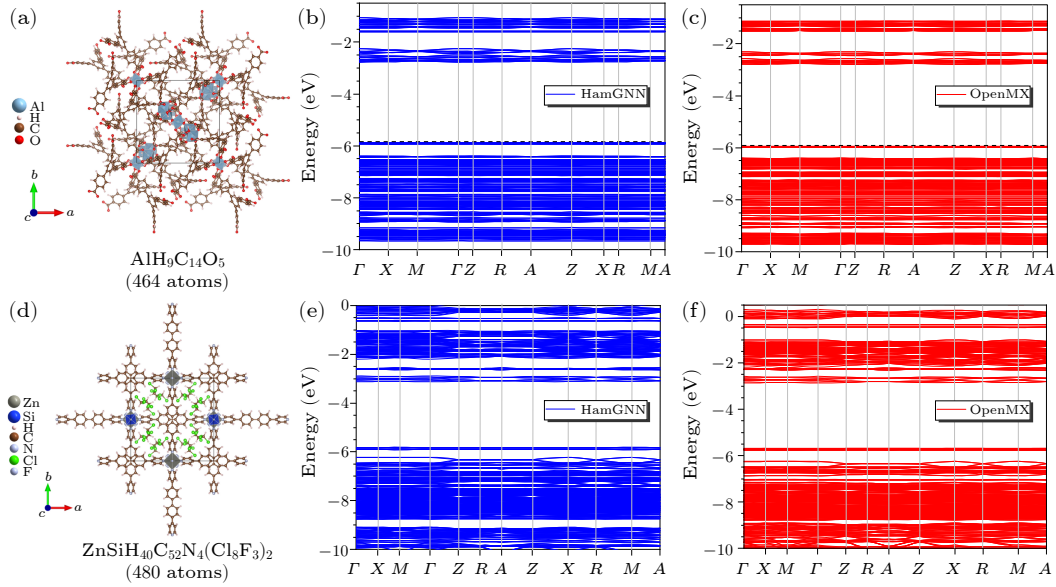
illustrates a close match between the predicted wave function and the wave function calculated by DFT. These tests demonstrate the powerful and wide applicability of the universal HamGNN model in various material systems, ranging from bulk crystals to low-dimensional materials. By utilizing this universal model, researchers can explore a vast array of low-dimensional materials with tailored functionalities and properties.

**Application of the Universal Hamiltonian Model to Large-Scale Hybrid Inorganic-Organic Crystals.** Now, we utilize the universal Hamiltonian model to predict the electronic structures of metal-organic frameworks (MOFs)<sup>[54–56]</sup> and demonstrate its efficiency and broad applicability in studying hybrid inorganic-organic crystals. MOFs<sup>[54–56]</sup> are a type of porous material composed of metal ions and organic ligands, forming intricate three-dimensional network structures. The electronic structures<sup>[57,58]</sup> of MOF materials, such as band structure, electron density of states, and electron orbital distribution,

directly influence the conductivity, optical properties, and potential applications in fields like photocatalysis and photovoltaics. However, due to the complex structure and large size of MOFs, accurately predicting their electronic structures using DFT can be challenging and computationally expensive.

By training on inorganic crystal structures available on the Materials Project, we have successfully developed a universal Hamiltonian neural network model that demonstrates high accuracy across various types of inorganic materials. However, this model lacks training on organic crystal structures which encompass a significant number of covalent bonds. Consequently, when predicting some complex organic or hybrid inorganic-organic crystal structures, the model may encounter certain challenges. In order to further improve the accuracy of the model in predicting the Hamiltonian matrix of organic crystal struc-

tures, we employed incremental training to further train and fine-tune the model. We selected approximately only 1800 small MOF crystal structures from the QMOF<sup>[59]</sup> database, with a maximum number of atoms per unit cell not exceeding 50, as the training set. Subsequently, we further restart the training of the previous universal HamGNN model using this dataset. The mean absolute error of the Hamiltonian matrix predicted by the further trained model on this MOF dataset is about 3.95 meV. The comparison between the Hamiltonian matrix elements predicted by HamGNN and those calculated by OpenMX on the small MOF dataset is shown in Supplementary Fig. S8. After undergoing fine-tuning, the Hamiltonian model demonstrates enhanced precision in inferring the interactions among diverse covalent bonds within organic crystalline materials.



**Fig. 6.** (a) Crystal structure, (b) predicted energy bands, and (c) calculated energy bands of  $\text{AlH}_9\text{C}_{14}\text{O}_5$ . (d) Crystal structure, (e) predicted energy bands, and (f) calculated energy bands of  $\text{ZnSiH}_{40}\text{C}_{52}\text{N}_4(\text{Cl}_8\text{F}_3)_2$ .

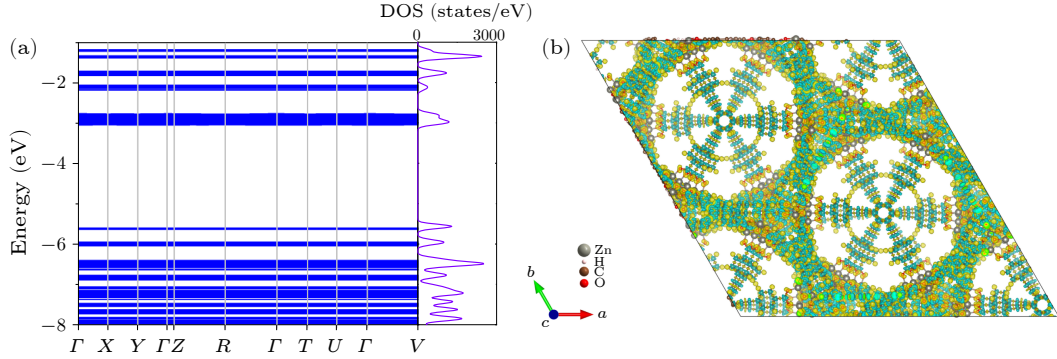
As a test, we used the model to predict the electronic structures of two MOF materials,  $\text{AlH}_9\text{C}_{14}\text{O}_5$  and  $\text{ZnSiH}_{40}\text{C}_{52}\text{N}_4(\text{Cl}_8\text{F}_3)_2$ , and compared them with the results obtained from DFT calculations. Both crystal structures have complex topological configurations, and the latter even contains up to seven elements, posing great challenges to the model. Despite these complexities, the energy bands predicted by the model for both structures are in good agreement with those obtained from DFT calculations, as shown in Fig. 6. This indicates that our developed model has high accuracy and reliability when dealing with complex organic crystal structures. Furthermore, the difference charge densities predicted for these two materials achieved excellent consistency with those obtained from DFT calculations, as shown in Supplementary Fig. S9. Moreover, the speed of DFT calculation can be improved by several orders of magnitude via using machine learning methods. Taking  $\text{ZnSiH}_{40}\text{C}_{52}\text{N}_4(\text{Cl}_8\text{F}_3)_2$  as an example, the time cost of using DFT methods is as high

as 181 core-h. However, the HamGNN model only takes 0.33 core-h to complete the computational task.

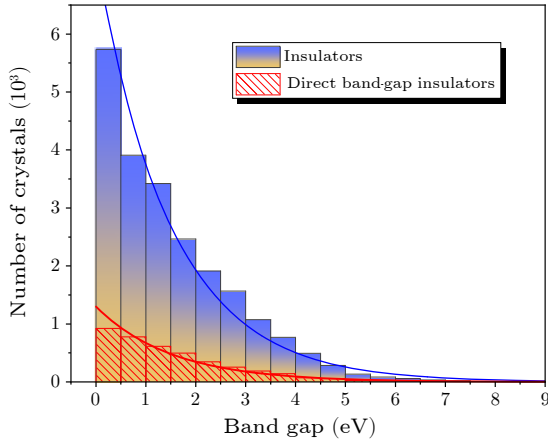
With further research, some more complex and larger MOF materials show wide application prospects in catalyst design, photovoltaics, solar cells, light-emitting diodes (LEDs), etc.<sup>[60]</sup> In this case, the advantage of using machine learning Hamiltonian models for computing large MOF materials becomes apparent. To demonstrate the applicability of our model, we conducted electronic structure calculations on a specific MOF crystal known as  $\text{Zn}_4\text{H}_{28}\text{C}_{58}\text{O}_{13}$  (labeled as c6ce00407e\_c6ce00407e6\_clean) from the CoRE MOF database.<sup>[61,62]</sup> This particular crystal is one of the largest MOFs in the CoRE MOF database and possesses a large unit cell containing 5562 atoms. Using the universal HamGNN model, we obtain the energy bands and density of states for  $\text{Zn}_4\text{H}_{28}\text{C}_{58}\text{O}_{13}$ , as shown in Fig. 7. It can be seen from the energy bands that the structure has a bandgap value of about 2.6 eV, which makes it a suitable candidate for various applications in the field

of energy harvesting and emission, such as photovoltaics and light-emitting diodes. Furthermore, we also predicted the difference charge density of this structure and observed that electrons primarily transfer from Zn and C atoms to

the oxygen and hydrogen atoms. The difference charge density can visually display the distribution of static electric potential and possible catalytic sites.



**Fig. 7.** (a) Predicted energy bands and (b) the difference charge density of  $\text{Zn}_4\text{H}_{28}\text{C}_{58}\text{O}_{13}$ . The yellow and blue colors represent charge accumulation and depletion, respectively.



**Fig. 8.** Histogram of the bandgap distribution for insulators and direct bandgap insulators.

*High-Throughput Predictions of Electronic Structures for Materials in the GNoME Dataset.* The GNoME<sup>[63]</sup> database is a remarkable collection of 2.2 million stable crystal structures that have been discovered using large-scale active learning techniques. This dataset has expanded our knowledge of stable materials by almost an order of magnitude, providing an unprecedented opportunity for in-depth investigation of material properties and the development of novel machine learning models. Despite the stability of the GNoME materials, their electronic structures remain largely unknown. While first-principles calculations could in principle determine the electronic structures, performing such computationally expensive calculations for the entire database would be prohibitively costly. Therefore, we employ the universal HamGNN model, which enables fast, accurate, and high-throughput predictions of electronic structures across the whole dataset, offering valuable insights for future theoretical predictions and experimental synthesis.

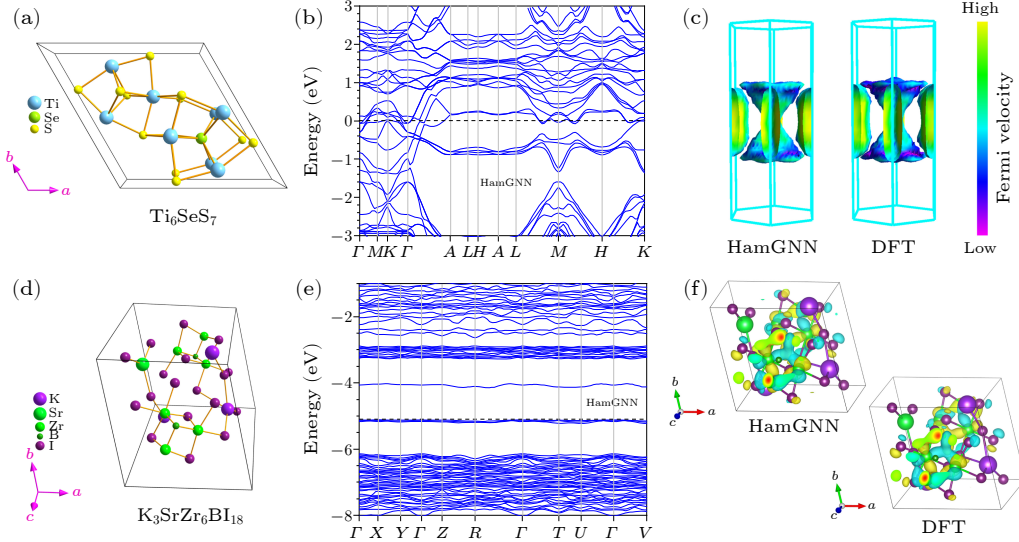
We employ the universal HamGNN model to compute the electronic structures of a total of 188722 structures from the GNoME database. Our analysis reveals

that 21973 of these structures exhibit insulating properties, with 3940 direct bandgap insulators and 18033 indirect band gap insulators. Direct bandgap materials have broad applications in photovoltaics and light-emitting devices due to their efficient electron-hole recombination processes. The histogram in Fig. 8 illustrates the distribution of bandgaps for both insulators and direct bandgap insulators. The figure shows an exponential decline in the number of insulators as their bandgap values increase. Utilizing this universal Hamiltonian model not only enables the rapid selection of materials with direct band gaps but also facilitates the identification of crystals exhibiting flat bands near the Fermi level. This unique electronic structure holds great promise as it can give rise to intriguing properties and phenomena, such as fractional quantum Hall effects, Bose-Einstein condensation, unconventional superconductivity, and strong correlation effects.<sup>[64–66]</sup> After conducting a search using HamGNN, we have successfully identified 5109 crystals with flat bands. This number is nearly double compared to the 2379 flat-band crystals listed in the Materials Flatband Database. In order to demonstrate the accuracy of this approach for the GNoME dataset, we take a gapless material and a semiconductor material with flat bands,  $\text{Ti}_6\text{Se}_7$  and  $\text{K}_3\text{SrZr}_6\text{BI}_{18}$ , from the GNoME database as examples. Notably, these two crystals are currently not included in the Materials Project.

The prediction results for the electronic structures of  $\text{Ti}_6\text{Se}_7$  and  $\text{K}_3\text{SrZr}_6\text{BI}_{18}$  are presented in Fig. 9. In Fig. 9(b), the predicted energy bands of  $\text{Ti}_6\text{Se}_7$  exhibit excellent agreement with the results obtained from DFT calculations, as shown in Supplementary Fig. S10(a). In addition, we compare the Fermi surfaces obtained from our predicted Hamiltonian matrices with those computed using DFT in Fig. 9(c). The comparison reveals remarkable similarities between the two Fermi surfaces, further validating the accuracy and reliability of our predictions. As shown in Fig. 9(e), the presence of flat bands at both HOMO and lowest-unoccupied-molecular-orbital (LUMO)

levels in the predicted energy bands of  $\text{K}_3\text{SrZr}_6\text{BI}_{18}$  is apparent. The predicted flat band electronic structure is validated by the DFT calculation (see Supplementary Fig. S10(b)). Figure 9(f) reveals that the flat band at the LUMO level is primarily formed by atomic orbitals from Zr and I atoms. This observation suggests a significant

contribution from these elements to the unique electronic behavior exhibited by this compound. Furthermore, it is worth noting that our predicted wave function aligns well with actual calculated results, reinforcing the accuracy of our computational approach.



**Fig. 9.** (a) Crystal structure of  $\text{Ti}_6\text{Se}_7$ . (b) Predicted energy bands of  $\text{Ti}_6\text{Se}_7$  by the universal HamGNN model. (c) Comparison between the predicted Fermi surface of  $\text{Ti}_6\text{Se}_7$  and those calculated using DFT. (d) Predicted energy bands of  $\text{K}_3\text{SrZr}_6\text{BI}_{18}$  by the universal HamGNN model. (e) Predicted energy bands of  $\text{K}_3\text{SrZr}_6\text{BI}_{18}$  by the universal HamGNN model. (f) Comparison of the HamGNN predicted LUMO wave function at the  $\Gamma$  point and the DFT calculated LUMO wave function at the  $\Gamma$  point for  $\text{K}_3\text{SrZr}_6\text{BI}_{18}$ .

**Discussion.** This work not only achieves a universal electronic Hamiltonian model for the entire periodic table but also demonstrates its practicality and reliability through successful predictions of electronic structures in various materials. We propose a framework to achieve a universal Hamiltonian model by training HamGNN on the crystal's Hamiltonian matrices obtained from the Materials Project or other large datasets. We have found that incorporating energy eigenvalues in the second training step, in addition to Hamiltonian matrices, is crucial for achieving a truly universal Hamiltonian model.

The universal model can accurately capture not only simple systems but also those containing uncommon or rare transition metal elements. One notable advantage of this universal model is its ability to handle complex multi-element systems comprising more than five elements. This capability opens up new possibilities for studying and understanding the electronic properties of advanced materials that often involve intricate combinations of chemical elements. The reliable framework provided by the universal electronic Hamiltonian model allows for efficient predictions of electronic structures across the periodic table. While the current publicly available GNoME dataset contains 380000 stable crystal structures, the complete dataset encompasses an unprecedented 2.2 million stable materials discovered through large-scale active learning.

We anticipate that with more structures being made publicly accessible, the Universal Hamiltonian model can

be leveraged to rapidly and accurately compute the electronic band structures of this vast collection through high-throughput calculations with much less computational cost compared to first-principle calculations. This would enable the identification of a significantly larger number of crystals with desirable electronic properties, driving further advancements in materials science. This breakthrough has significant implications for material design, catalysis, electronics, and other fields that heavily rely on accurate knowledge and understanding of electronic properties.

**Network Details.** The equivariant node features are  $32 \times 0\text{o} + 128 \times 0\text{e} + 128 \times 1\text{o} + 64 \times 1\text{e} + 128 \times 2\text{e} + 32 \times 2\text{o} + 64 \times 3\text{o} + 32 \times 3\text{e} + 32 \times 4\text{o} + 32 \times 4\text{e} + 16 \times 5\text{o} + 8 \times 5\text{e} + 8 \times 6\text{e}$ , where  $32 \times 0\text{o}$  means that there are 32 channels in this feature part, and the features in each channel are  $O(3)$  irreducible representations with  $l = 0$  and odd parity. The node features utilized in this universal HamGNN model surpass those employed in our previous work<sup>[22]</sup> to enhance the network capacity for describing the entire periodic table. The universal HamGNN model has five orbital convolution layers. The spherical harmonic basis functions used to expand the interatomic directions are  $0\text{e} + 1\text{o} + 2\text{e} + 3\text{o} + 4\text{e} + 5\text{o} + 6\text{e}$ . The interatomic distance between atom  $i$  and its neighboring atom  $j$ , which falls within the cutoff radius  $r_c$ , is expanded utilizing the Bessel basis function:

$$B(|\tau_{ij}|) = \sqrt{\frac{2}{r_c}} \frac{\sin(n\pi|\tau_{ij}|/r_c)}{|\tau_{ij}|}. \quad (4)$$

The atomic neighbors are determined based on the cutoff radius of each atom's orbital basis. The interatomic distance is expanded using a series of Bessel functions with  $n = 1, 2, \dots, N_b$ , where  $N_b$  represents the number of Bessel basis functions. In this study,  $N_b$  is set to 64.

**DFT Details.** All the Hamiltonian matrices in the training set were computed using the PBE functional, with a Monkhorst-pack grid of  $6 \times 6 \times 6$ , and a convergence criterion of  $1.0 \times 10^{-8}$  Hartree. The energy cutoff employed for discretizing the real space is set at 200 Rydberg.

**Data and Code Availability.** The HamGNN code is publicly available at <https://github.com/QuantumLab-ZY/HamGNN>. The trained network weights for the universal model, the predicted energy bands for the test set, and the identified crystals with flat bands are available on Zenodo (<https://zenodo.org/records/10827117>).

**Acknowledgments.** This work was supported by the National Key R&D Program of China (Grant No. 2022YFA1402901), the National Natural Science Foundation of China (Grant Nos. 11825403, 11991061, and 12188101), and the Guangdong Major Project of the Basic and Applied Basic Research (Future Functional Materials Under Extreme Conditions) (Grant No. 2021B0301030005).

## References

- [1] Marzari N, Ferretti A, and Wolverton C 2021 *Nat. Mater.* **20** 736
- [2] McCardle K 2023 *Nat. Comput. Sci.* **3** 915
- [3] Chen Z X, Li W Q, Sabuj M A *et al.* 2021 *Nat. Commun.* **12** 5889
- [4] Dzade N Y 2021 *Sci. Rep.* **11** 4755
- [5] Reidy K, Varnavides G, Thomsen J D *et al.* 2021 *Nat. Commun.* **12** 1290
- [6] Pederson R, Kalita B, and Burke K 2022 *Nat. Rev. Phys.* **4** 357
- [7] Schleider G R, Padilha A C M, Acosta C M, Costa M, and Fazzio A 2019 *J. Phys.: Mater.* **2** 032001
- [8] Makkar P and Ghosh N N 2021 *RSC Adv.* **11** 27897
- [9] Jones R O 2015 *Rev. Mod. Phys.* **87** 897
- [10] Hegde G and Bowen R C 2017 *Sci. Rep.* **7** 42669
- [11] Li H C, Collins C, Tanha M, Gordon G J, and Yaron D J 2018 *J. Chem. Theory. Comput.* **14** 5764
- [12] Schütt K T, Gastegger M, Tkatchenko A, Müller K R, and Maurer R J 2019 *Nat. Commun.* **10** 5024
- [13] Gastegger M, McSloy A, Luya M, Schütt K T, and Maurer R J 2020 *J. Chem. Phys.* **153** 044123
- [14] Unke O T, Bogojeski M, Gastegger M *et al.* 2021 *arXiv:2106.02347 [physics.chem-ph]*
- [15] Wang Z F, Ye S Z, Wang H *et al.* 2021 *npj Comput. Mater.* **7** 11
- [16] Westermayr J and Maurer R J 2021 *Chem. Sci.* **12** 10755
- [17] Li H, Wang Z, Zou N *et al.* 2022 *Nat. Comput. Sci.* **2** 367
- [18] Nigam J, Willatt M J, and Ceriotti M 2022 *J. Chem. Phys.* **156** 014115
- [19] Schattauer C, Todorović M, Ghosh K, Rinke P, and Libisch F 2022 *npj Comput. Mater.* **8** 116
- [20] Zhang L W, Onat B, Dusson G *et al.* 2022 *npj Comput. Mater.* **8** 158
- [21] Gong X X, Li H, Zou N L *et al.* 2023 *Nat. Commun.* **14** 2848
- [22] Zhong Y, Yu H, Su M, Gong X, and Xiang H 2023 *npj Comput. Mater.* **9** 182
- [23] Ye Y F, Wang Q, Lu J, Liu C T, and Yang Y 2016 *Mater. Today* **19** 349
- [24] Feng R, Zhang C, Gao M C *et al.* 2021 *Nat. Commun.* **12** 4329
- [25] George E P, Raabe D, and Ritchie R O 2019 *Nat. Rev. Mater.* **4** 515
- [26] Zhang R Z and Reece M J 2019 *J. Mater. Chem. A* **7** 22148
- [27] Oses C, Toher C, and Curtarolo S 2020 *Nat. Rev. Mater.* **5** 295
- [28] Zhao P, Xiao C, and Yao W 2021 *npj 2D Mater. Appl.* **5** 38
- [29] Chen C 2022 *Nat. Comput. Sci.* **2** 703
- [30] Chen C and Ong S P 2022 *Nat. Comput. Sci.* **2** 718
- [31] Takamoto S, Shinagawa C, Motoki D *et al.* 2022 *Nat. Commun.* **13** 2991
- [32] Deng B W, Zhong P C, Jun K *et al.* 2023 *Nat. Mach. Intell.* **5** 1031
- [33] Batatia I, Benner P, Chiang Y *et al.* 2023 *arXiv:2401.00096 [physics.chem-ph]*
- [34] Jain A, Ong S P, Hautier G *et al.* 2013 *APL Mater.* **1** 011002
- [35] Jain A, Montoya J, Dwaraknath S *et al.* 2018 *Handbook of Materials Modeling* (Berlin: Springer) p 1
- [36] Wang H, Zhang L F, Han J Q, and E W N 2018 *Comput. Phys. Commun.* **228** 178
- [37] Behler J 2011 *J. Chem. Phys.* **134** 074106
- [38] Pinheiro M, Ge F C, Ferré N, Dral P O, and Barbatti M 2021 *Chem. Sci.* **12** 14396
- [39] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
- [40] Chmiela S, Tkatchenko A, Sauceda H E *et al.* 2017 *Sci. Adv.* **3** e1603015
- [41] Glielmo A, Sollich P, and De Vita A 2017 *Phys. Rev. B* **95** 214302
- [42] Ozaki T and Kino H 2004 *Phys. Rev. B* **69** 195113
- [43] Behler J 2003 *Phys. Rev. B* **67** 155108
- [44] Weinert U 1980 *Arch. Ration. Mech. Anal.* **74** 165
- [45] Morrison M A and Parker G A 1987 *Aust. J. Phys.* **40** 465
- [46] Grisafi A, Wilkins D M, Csányi G, and Ceriotti M 2018 *Phys. Rev. Lett.* **120** 036002
- [47] Thomas N, Smidt T, Kearnes S *et al.* 2018 *arXiv:1802.08219 [cs.LG]*
- [48] Naveed H, Khan A U, Qiu S *et al.* 2023 *arXiv:2307.06435 [cs.CL]*
- [49] Zhao W X, Zhou K, Li J Y *et al.* 2023 *arXiv:2303.18223 [cs.CL]*
- [50] Zeng L Y, Hu X W, Zhou Y Z *et al.* 2023 *Innov. Mater.* **1** 100042
- [51] Zeng L Y, Hu X W, Zhou Y Z *et al.* 2024 *Adv. Sci.* **11** 2305054
- [52] Zeng L Y, Wang Z Q, Song J *et al.* 2023 *Adv. Funct. Mater.* **33** 2301929
- [53] Zhao S G, Kang L, Shen Y G *et al.* 2016 *J. Am. Chem. Soc.* **138** 2961
- [54] Zhang J W, Wang W H, Xie M X *et al.* 2023 *Electrochim. Acta* **461** 142691
- [55] Baumann A E, Burns D A, Liu B, and Thoi V S 2019 *Commun. Chem.* **2** 86
- [56] Felix Sahayaraj A, Joy Prabu H, Maniraj J *et al.* 2023 *J. Inorg. Organomet. Polym. Mater.* **33** 1757
- [57] Mancuso J L, Mroz A M, Le K N, and Hendon C H 2020 *Chem. Rev.* **120** 8641
- [58] Rosen A S, Fung V, Huck P *et al.* 2022 *npj Comput. Mater.* **8** 112
- [59] Rosen A S, Iyer S M, Ray D *et al.* 2021 *Matter* **4** 1578
- [60] Stassen I, Burtch N, Talin A *et al.* 2017 *Chem. Soc. Rev.* **46** 3185
- [61] Chung Y G, Camp J, Haranczyk M *et al.* 2014 *Chem. Mater.* **26** 6185
- [62] Chung Y G, Haldoupis E, Bucior B J *et al.* 2019 *J. Chem. Eng. Data* **64** 5985
- [63] Merchant A, Batzner S, Schoenholz S S *et al.* 2023 *Nature* **624** 80
- [64] He C Y, Liao Y J, Ouyang T *et al.* 2023 *Fundam. Res.*
- [65] Bhattacharya A, Timokhin I, Chatterjee R, Yang Q, and Mishchenko A 2023 *npj Comput. Mater.* **9** 101
- [66] Huber S D and Altman E 2010 *Phys. Rev. B* **82** 184502