Supplementary Material: Universal materials model of deep-learning density functional theory Hamiltonian

Yuxiang Wang,[1,*] Yang Li,[1,*] Zechen Tang,[1,*] He Li,[1,2] Zilong Yuan,[1] Honggeng Tao,[1] Nianlong Zou,[1] Ting Bao,[1] Xinghao Liang,[1] Zezhou Chen,[1] Shanghua Xu,[1] Ce Bian,[1] Zhiming Xu,[1] Chong Wang,[1] Chen Si,[5] Wenhui Duan,[1,2,3,†] and Yong Xu[1,3,4,‡]

[1] *State Key Laboratory of Low Dimensional Quantum Physics and Department of Physics, Tsinghua University, Beijing 100084, China*
[2] *Institute for Advanced Study, Tsinghua University, Beijing 100084, China*
[3] *Frontier Science Center for Quantum Information, Beijing 100084, China*
[4] *RIKEN Center for Emergent Matter Science (CEMS), Wako, Saitama 351-0198, Japan*
[5] *School of Materials Science and Engineering, Beihang University, Beijing 100191, China*

This PDF file includes:

- Methods
- Supplementary Figs. S1-S6

We developed a materials database named MIND (Material Intelligent Database) based on the automated interactive infrastructure and database (AiiDA) framework [1, 2]. AiiDA is an open-source Python library designed to assist researchers in the managing, automating and exploring their computational workflows. A key advantage of AiiDA is its provenance graph, which records all actions and processes applied to the data, ensuring that the entire history of a computational workflow is tracked and reproducible. This not only enhances research reproducibility but also facilitates data and workflow sharing within the community. AiiDA supports a wide range of computational codes, such as Quantum ESPRESSO, VASP, LAMMPS, among others, through its plugin system, and is flexible for interfacing with other codes. Our MIND database is built on the AiiDA framework, providing a streamlined process from initial setup of calculations to final result analysis. With AiiDA, the entire MIND workflow is trackable, reproducible, and shareable.

A key feature of the MIND database is its inclusion of density functional theory (DFT) Hamiltonian matrices for DeepH training, a component rarely found in other materials databases. We have implemented a format for storing these Hamiltonian matrices based on the HDF5 file format, which was introduced in our previous DeepH works [3]. This format allows for the storage of Hamiltonian matrices along with materials structure and basis set information. Through this robust and efficient system, we have developed a streamlined workflow that enables scalability and future expansion of the universal materials dataset.

High-throughput DFT computation is carried out with the AiiDA workflow to yield the universal dataset. The "Standard"-size basis set defined by OpenMX is applied [4, 5], with the exclusion of f-orbital basis functions for the elements P, S, and Cl. The pseudopotentials and basis sets used for each element are summarized in supplementary Table S1. A real-space grid with a mesh cutoff of 300 Ry is applied. A Monkhorst-Pack k-point sampling of $n_1 \times n_2 \times n_3$ satisfying $n_i \cdot a_i \geq 80$ Å $(i = 1, 2, 3)$ is used, where a_i refers to the length of the i-th lattice vector. The DFT output files are then parsed to the aforementioned Hamiltonian matrix format for DeepH-training process. Probability distribution functions displayed in the main text and supplementary materials are computed with a Gaussian kernel smoothing.

Upon further analysis of the test set, we observed that the material structures with the highest test errors in the DeepH model usually have large atomic forces predicted by DFT. We also noticed that when the unit cell of materials has a small angle between lattice vectors, the DFT code sometimes encounters problems due to an unreasonable selection of real-space grids, leading to artificially large atomic forces. Therefore, we excluded structures with the largest component of

* These authors contributed equally to this work.
† duanw@tsinghua.edu.cn
‡ yongxu@mail.tsinghua.edu.cn

TABLE S1: Summary of basis set applied in high-throughput computation

Element	Basis functions	Notes
H	H6.0-s2p1	
Li	Li8.0-s3p2	
Be	Be7.0-s2p2	
B	B7.0-s2p2d1	
C	C6.0-s2p2d1	
N	N6.0-s2p2d1	
O	O6.0-s2p2d1	
F	F6.0-s2p2d1	
Na	Na9.0-s3p2d1	
Mg	Mg9.0-s3p2d1	
Al	Al7.0-s2p2d1	
Si	Si7.0-s2p2d1	
P	P7.0-s2p2d1	Exclude f-orbital basis function
S	S7.0-s2p2d1	Exclude f-orbital basis function
Cl	Cl7.0-s2p2d1	Exclude f-orbital basis function
K	K10.0-s3p2d1	
Ca	Ca9.0-s3p2d1	
Cu	Cu6.0S-s3p2d1	Soft pseudopotential
Zn	Zn6.0S-s3p2d1	Soft pseudopotential
Ga	Ga7.0-s3p2d2	
Ge	Ge7.0-s3p2d2	
As	As7.0-s3p2d2	
Se	Se7.0-s3p2d2	
Br	Br7.0-s3p2d2	

atomic forces exceeding 5 eV/Å or having unit cells with the smallest lattice angle less than 30° or the largest lattice angle more than 150° from the test set, excluding a total of 113 structures. This refined test set shows a reduced MAE of 1.9 meV. The updated MAE for each element and the distribution of MAE across structures are presented in Fig. S1. The element-wise MAE is significantly reduced after using the refined test set structures.

To showcase the capacity of the universal model to be refined through fine-tuning, a specialized dataset focusing on carbon allotropes is curated in addition. This carbon allotrope dataset comprises a total of 427 allotrope structures from the SACADA database up to February 2023 [6]. The selection of structures in the carbon dataset is based on the unit cell size, with a cutoff of 60 atoms per unit cell. The dataset still includes relatively complex phases, as exemplified in Fig. 4a. In contrast, the universal dataset contains only 55 carbon allotropes, with only 34 in training set. Therefore, it is reasonable to anticipate that the carbon allotrope dataset would encompass more intricate structures, and serves as a typical case study for fine-tuning the universal model. The DFT workflow of carbon allotrope dataset is identical to that of the universal dataset, encompassing the same setup for basis set and other computation parameters.

The electric susceptibility is computed with the HopTB package [7], via the formula:

$$\chi^{ab}(\omega) = \frac{e^2}{\epsilon_0 \hbar} \int \frac{d^3\mathbf{k}}{(2\pi)^3} \sum_{n,m} f_{nm} \frac{r^a_{nm} r^b_{mn}}{\omega_{mn}(\mathbf{k}) - \omega - i\eta}, \tag{1}$$

Here, a and b are Cartesian directions, with $a, b \in \{x, y, z\}$, while ϵ_0, $\hbar$, and e denote the vacuum permittivity, the reduced Planck's constant, and the elementary charge, respectively. $\omega_{mn}(\mathbf{k}) = \frac{E_{m\mathbf{k}} - E_{n\mathbf{k}}}{\hbar}$ and $f_{nm} = f_n(\mathbf{k}) - f_m(\mathbf{k})$ represent the energy difference between bands n and m at $\mathbf{k}$ and the difference in Fermi-Dirac occupations, respectively. The Berry connection r^a_{nm} is defined as zero when $n = m$. The k-grid follows the Monkhorst-Pack sampling convention, with a higher sampling density, approximately threefold denser along each axis than that used in corresponding DFT computation. For data display consideration, the displayed data are trace of the suceptibility tensor ($\chi^{xx} + \chi^{yy} + \chi^{zz}$), which is a physical quantity invariant with the selection of Cartesian axes. The zero-frequency susceptibility $\chi^{ab}(0)$ is referred to as "static susceptibility". Furthermore, additional (non-)linear optical properties may be computed using the HopTB package based on the predicted Hamiltonians, facilitating functional materials discovery based on the universal DeepH model.

In the DeepH-2 framework, an inherent property of the Hamiltonian matrix, namely its Hermitian property $H_{\alpha\beta} = H^*_{\beta\alpha}$, is not explicitly enforced due to the treatment of atom pairs with directional edges. Consequently, $H_{\alpha\beta}$ and $H_{\beta\alpha}$ are derived from different sets of edge features, violating the Hermitian property. To address this, for testing purposes, we evaluate the loss based on a modified output of the predicted Hamiltonian, denoted as $\tilde{H} = (H + H^\dagger)/2$, where $H^\dagger$ represents the Hermitian conjugate of the output Hamiltonian. This modification typically reduces the mean absolute error (MAE) of the predicted Hamiltonian and ensures the numerical stability of subsequent evaluations of physical quantities, as these evaluations generally assume the Hamiltonian to be Hermitian.

The universal materials model trained in this work comprises 17,275,936 parameters, with 3 equivariant transformer blocks. Each node and edge in intermediate layers carry 80 channels with $l_{\max} = 4$, where l denotes the angular quantum number. In transformer blocks, the number of attention heads is 2. The dimension of the irreps feature in graph attention is 80 channels. The dimension of the scalar feature is 80 channels. The dimension of the value vector is 40 channels. The dimension of all linear layers in transformer blocks is 160 channels. The dropout rate of attention weight is 0.1. The drop path rate is 0.05. There is no dropout for outputs of attention and feed forward network. The embedding of the material structure consists of embedding of atomic numbers and inter-atomic distances utilizing Gaussian smearing of 600 bases, with centers ranging from 0.0 to 9.0 Å, and a width twice the distance between adjacent centers. The initial learning rate is 4×10^{-4}, and the minimal learning rate is set as 1×10^{-5}. The scheduler applied is "ReduceLROnPlateau", with a decay rate of 0.5, a cooldown of 15, a decay patience of 20, and a threshold of 0.05. The optimizer applied is "Adam" with betas of (0.9, 0.999) and a weight decay of 1×10^{-3}. The batch size is fixed as 1. The loss function is gauge equivariant Huber loss with a Huber delta of 0.01 eV.

In fine-tuning, most of model parameters including numbers of channels, dropout rates, and optimizer keep same as the pre-training stage. As for algorithm parameters, the batch size is still 1, but the initial learning rate is set as 2×10^{-4}, with a decay rate of 0.5, a cooldown of 40, and a decay patience of 120. The minimal learning rate is 1×10^{-5}. The training process was also carried out on an NVIDIA A100 GPU and stopped after extra 1069 epochs.

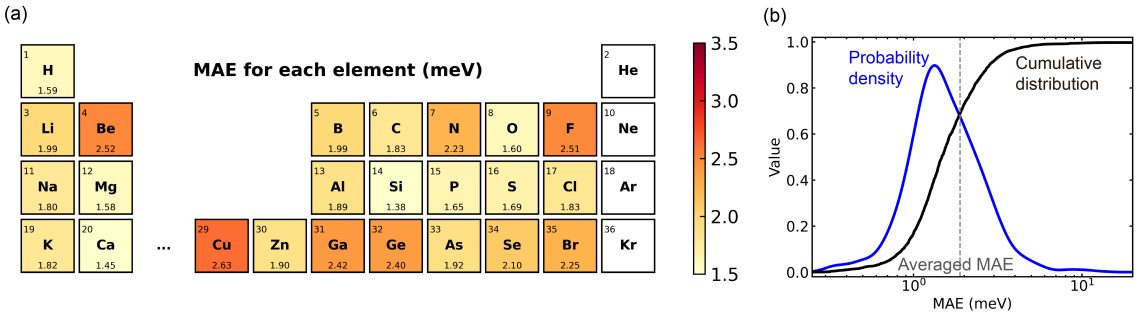


FIG. S1: Performance evaluation of the universal materials model on test set excluding 113 structures. (a) MAEs of the predicted DFT Hamiltonian averaged for each element. (b) Cumulative distribution and probability density of MAEs across test structures, giving an averaged MAE of 1.9 meV.

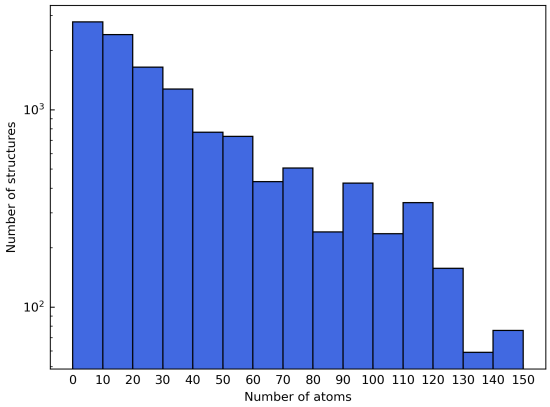


FIG. S2: The distribution of the number of atoms in each structure within the universal dataset

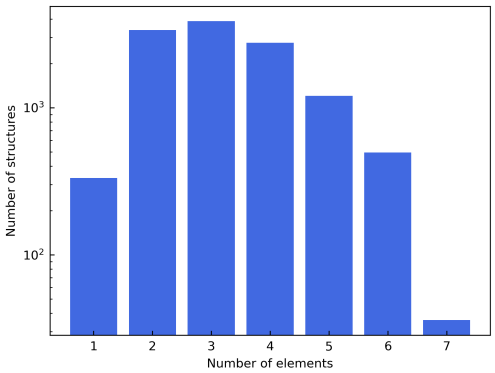


FIG. S3: The distribution of the number of elements in each structure within the universal dataset

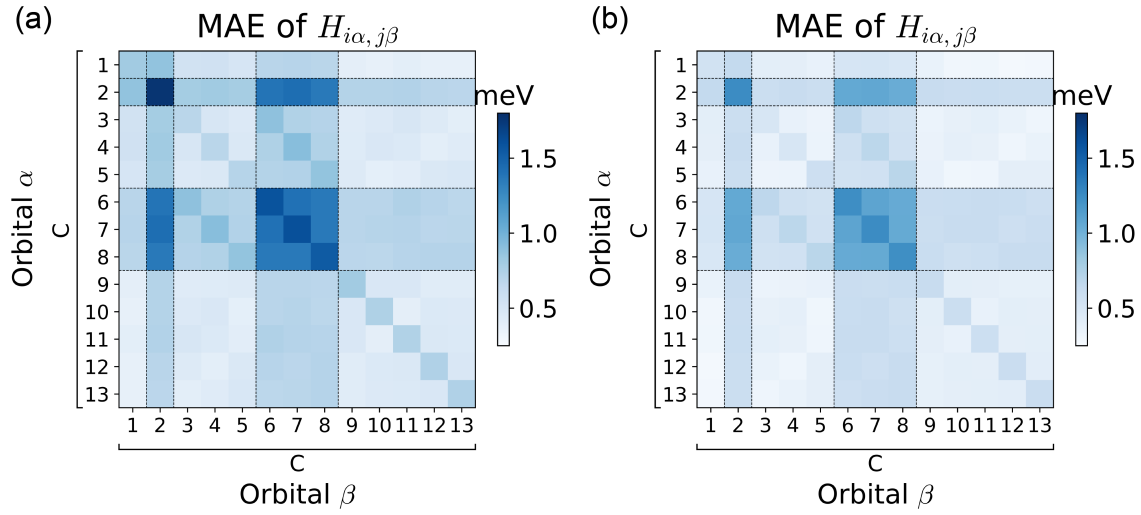


FIG. S4: Comparison of orbital-averaged error of fine-tuned and train-from-scratch DeepH model on the carbon allotrope dataset. (a) orbital-averaged error of model trained from scratch, with an overall MAE of 0.68 meV. (b) orbital-averaged error of model fine-tuned on the universal model, with an overall MAE of 0.54 meV. Here, each carbon atom has 13 basis sets, and the error is evaluated by averaging error of all Hamiltonian matrix element between two specific basis sets, yielding a 13×13 orbital-averaged error matrix. An overall accuracy improvement is seen in the fine-tuned model.

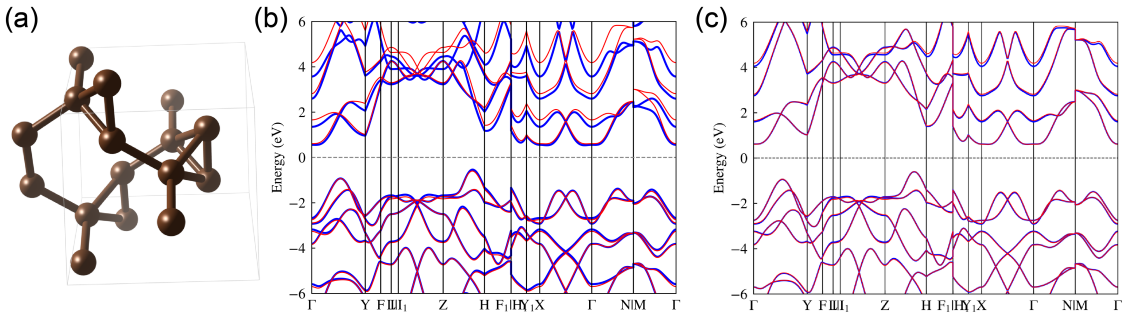
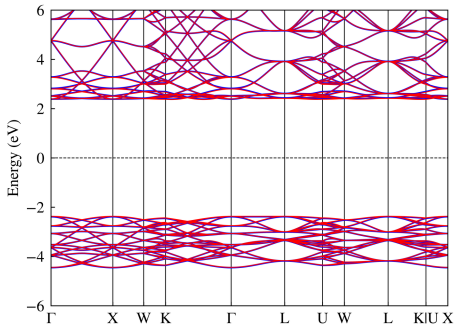
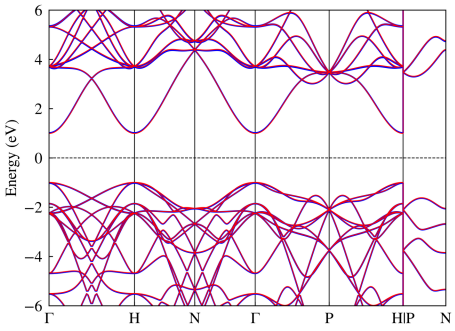


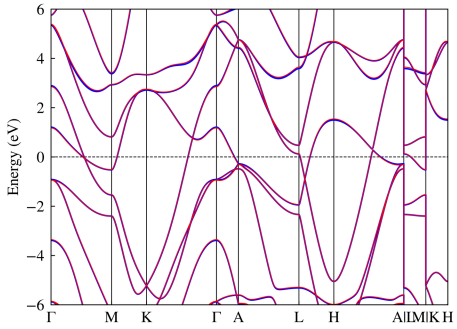
FIG. S5: Comparison of the band structure of a selected material structure (SACADA structure ID: 135) in test set of the carbon allotrope dataset. (a) Comparison of DeepH-predicted (blue) and DFT-calculated (red) band structure of the DeepH model trained from scratch. (b) Comparison of band structure of the DeepH model fine-tuned from the universal model. A substantial more precise matching is seen, indicating the improved performance from fine-tuning.



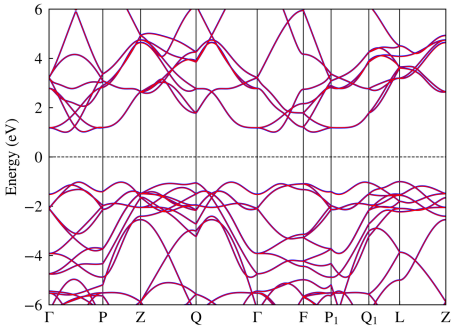
SACADA structure ID: 3



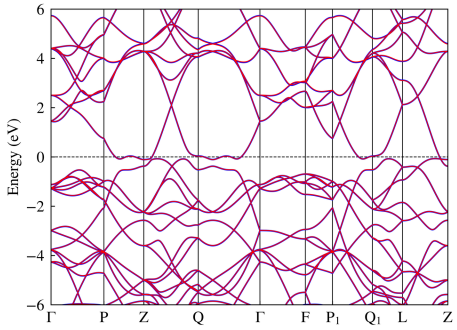
SACADA structure ID: 13



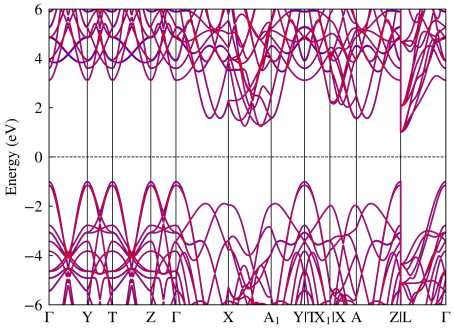
SACADA structure ID: 22



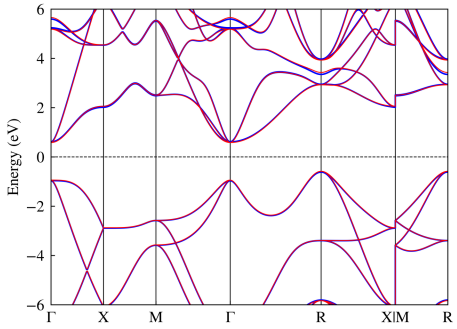
SACADA structure ID: 24



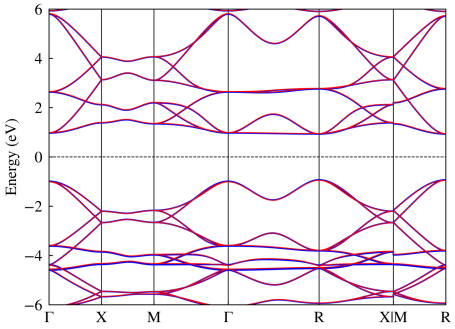
SACADA structure ID: 26



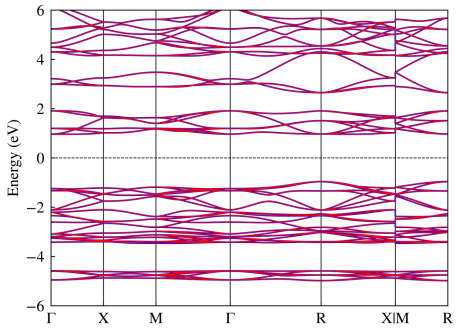
SACADA structure ID: 30



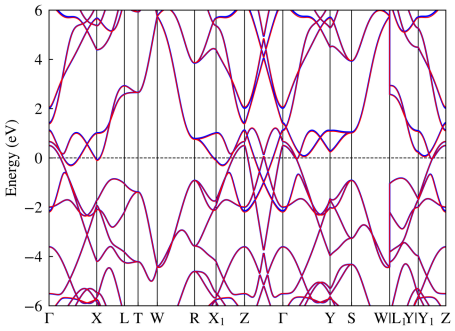
SACADA structure ID: 35



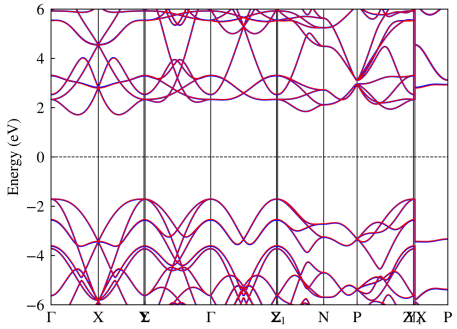
SACADA structure ID: 36



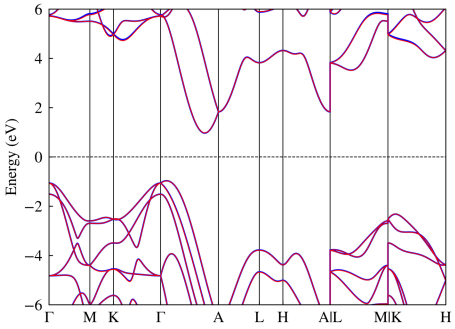
SACADA structure ID: 48



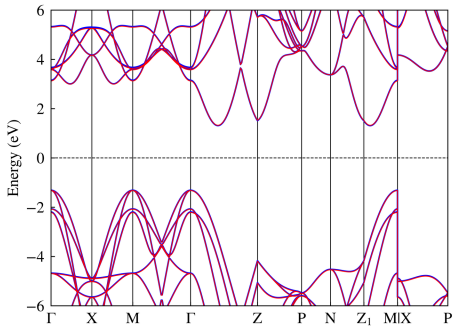
SACADA structure ID: 51



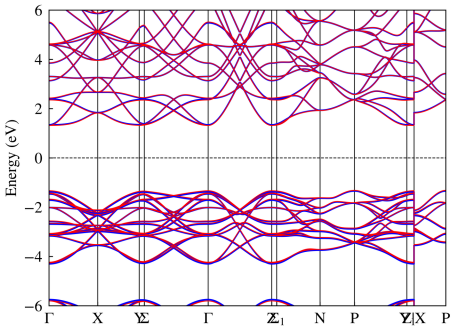
SACADA structure ID: 52



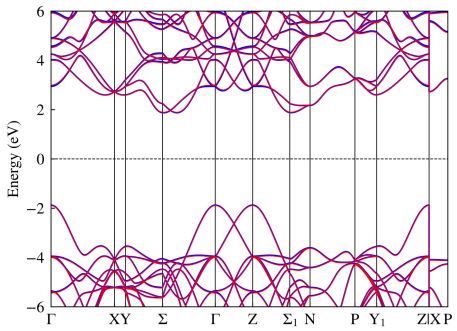
SACADA structure ID: 56



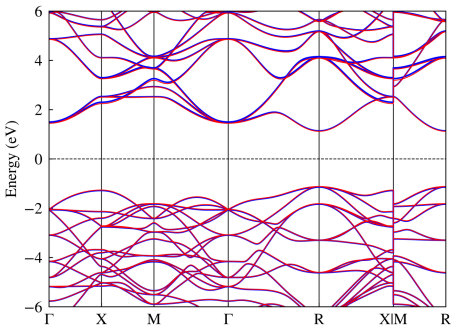
SACADA structure ID: 60



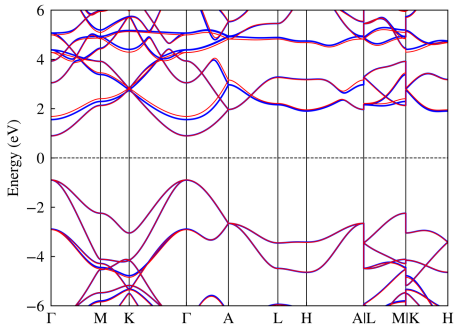
SACADA structure ID: 61



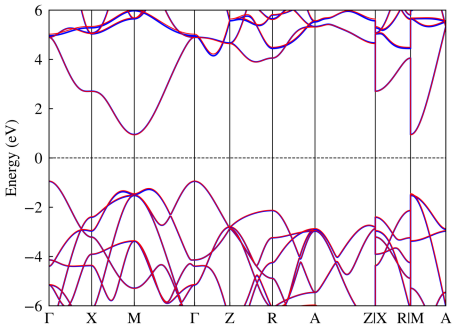
SACADA structure ID: 73



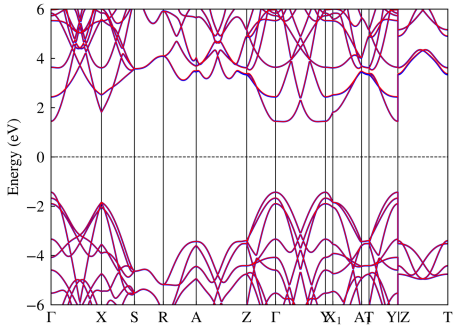
SACADA structure ID: 77



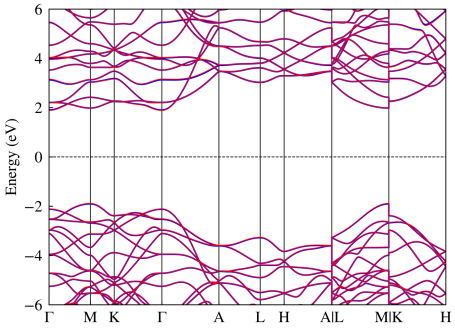
SACADA structure ID: 79



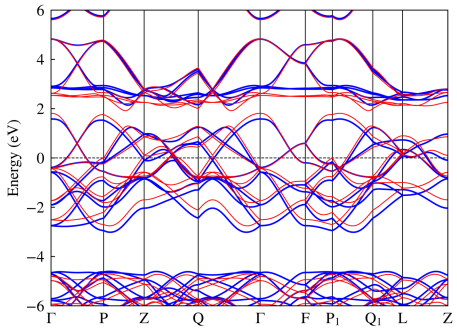
SACADA structure ID: 80



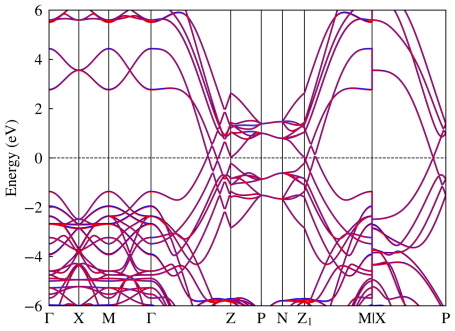
SACADA structure ID: 81



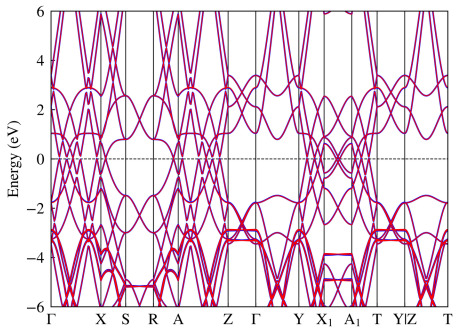
SACADA structure ID: 84



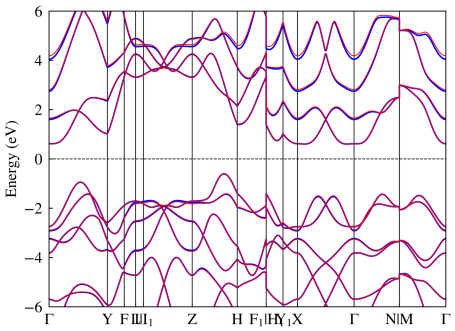
SACADA structure ID: 106



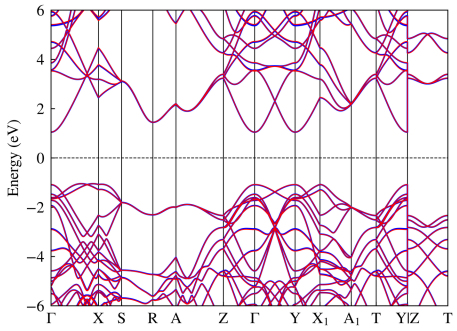
SACADA structure ID: 109



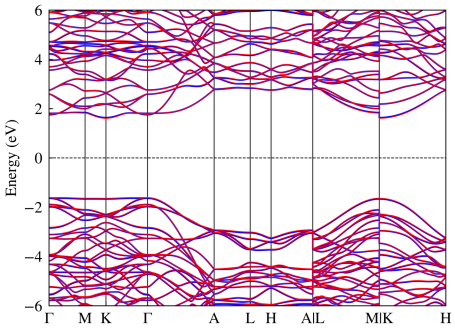
SACADA structure ID: 125



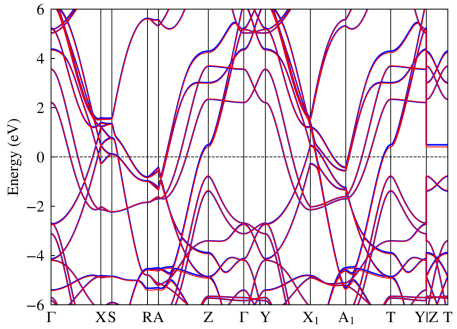
SACADA structure ID: 135



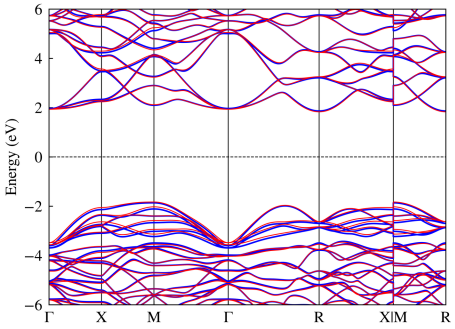
SACADA structure ID: 136



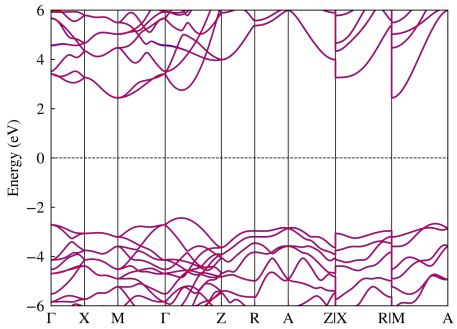
SACADA structure ID: 145



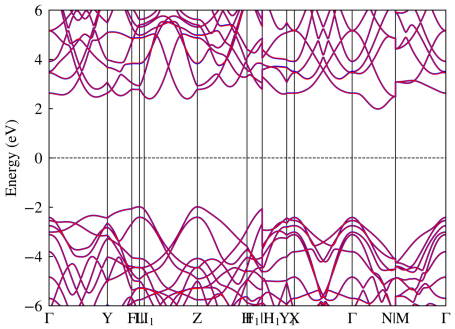
SACADA structure ID: 157



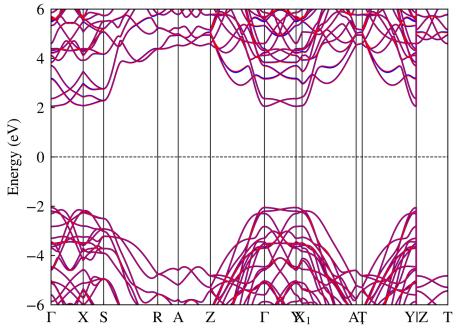
SACADA structure ID: 172



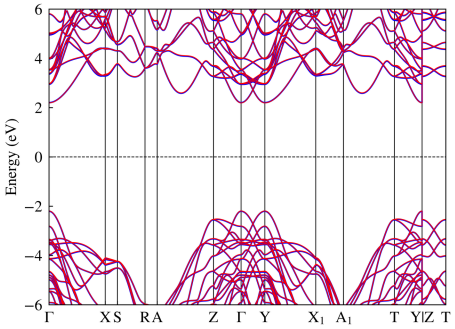
SACADA structure ID: 182



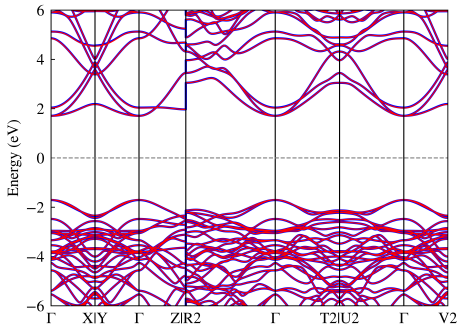
SACADA structure ID: 186



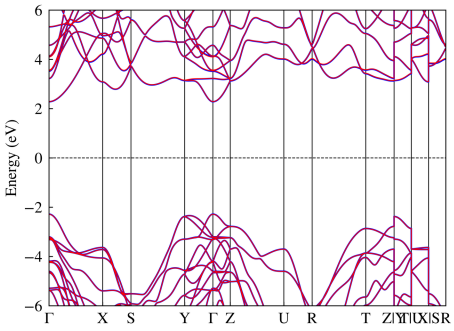
SACADA structure ID: 212



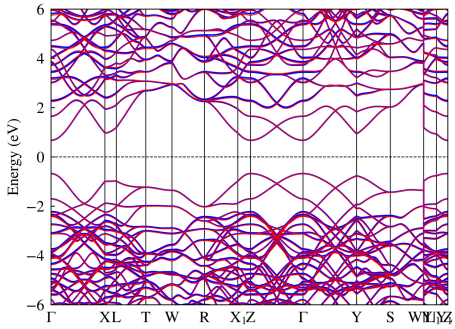
SACADA structure ID: 213



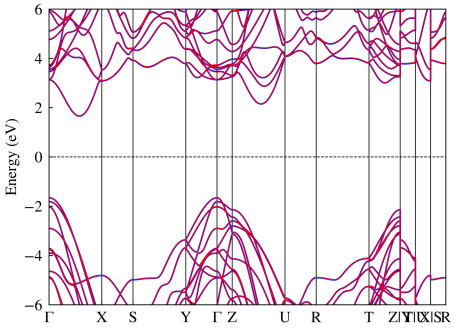
SACADA structure ID: 220



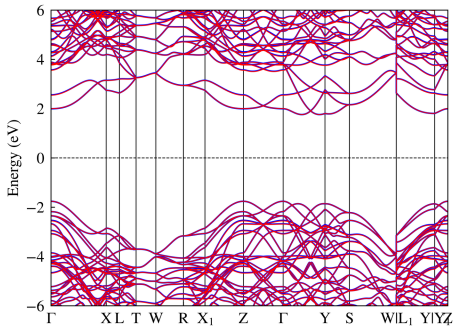
SACADA structure ID: 221



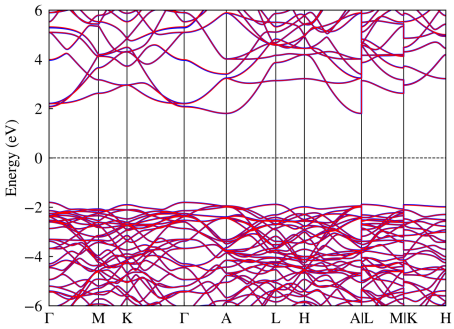
SACADA structure ID: 228



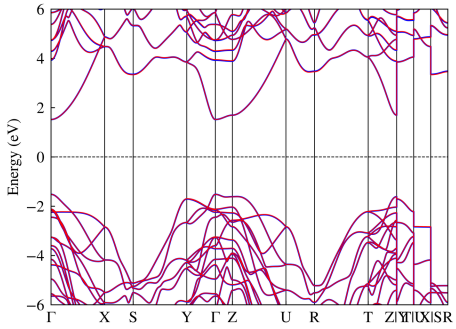
SACADA structure ID: 230



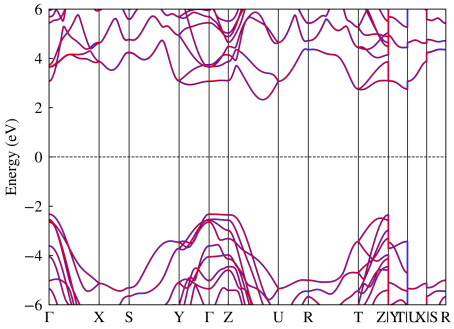
SACADA structure ID: 233



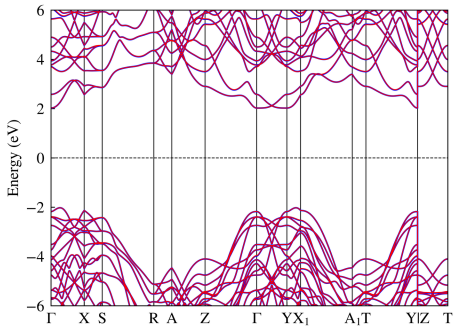
SACADA structure ID: 242



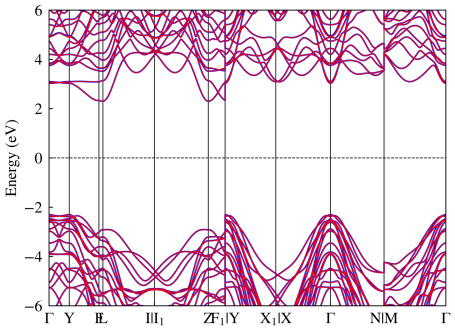
SACADA structure ID: 243



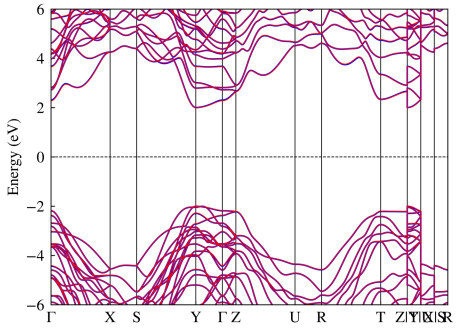
SACADA structure ID: 256



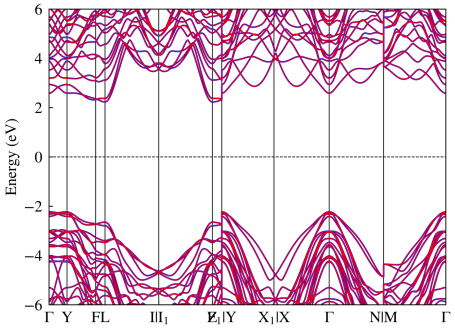
SACADA structure ID: 259



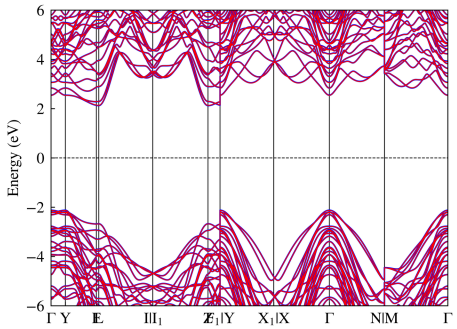
SACADA structure ID: 262



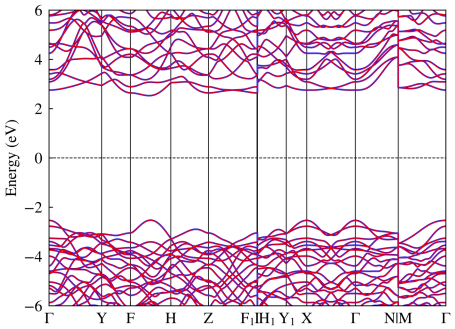
SACADA structure ID: 273



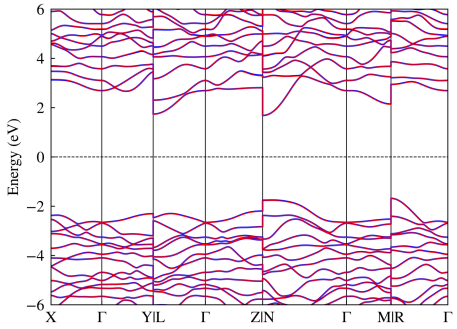
SACADA structure ID: 282



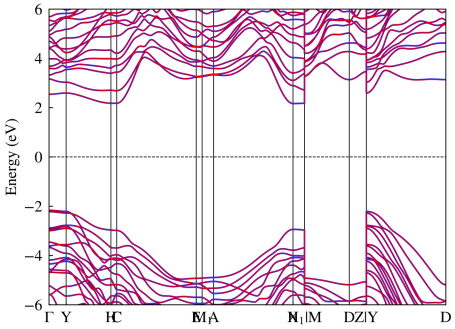
SACADA structure ID: 304



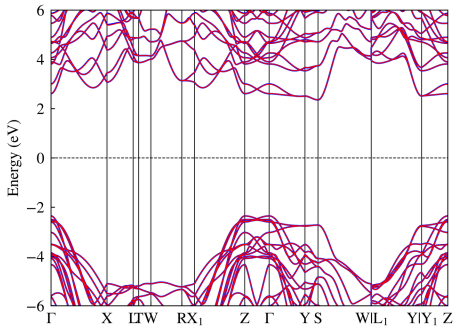
SACADA structure ID: 317



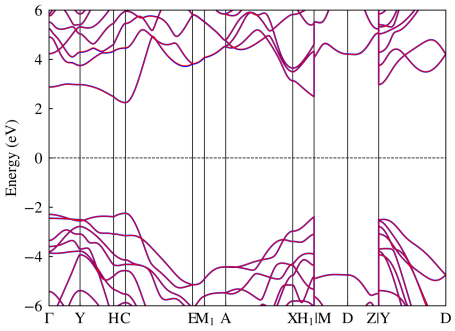
SACADA structure ID: 319



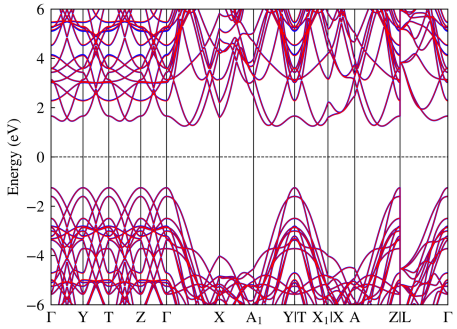
SACADA structure ID: 350



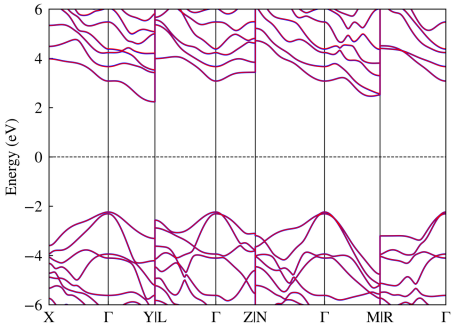
SACADA structure ID: 358



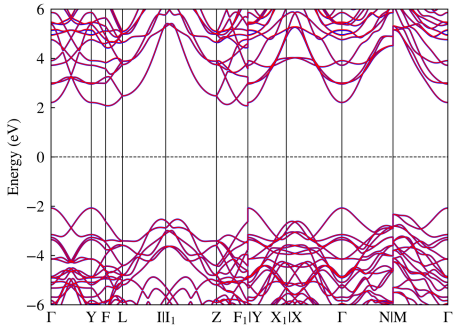
SACADA structure ID: 365



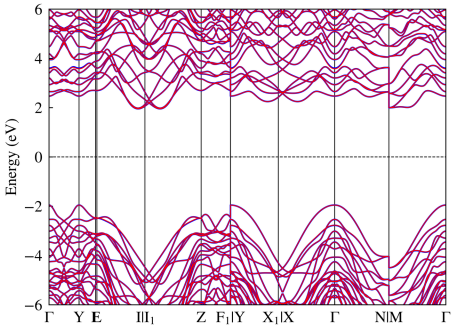
SACADA structure ID: 370



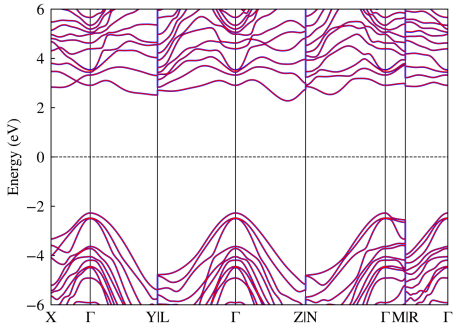
SACADA structure ID: 373



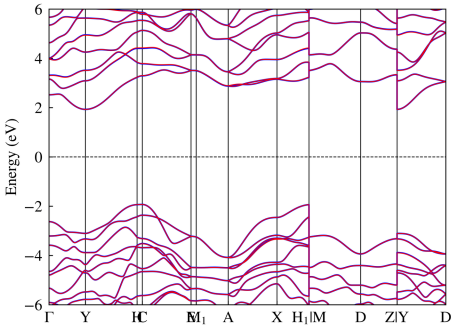
SACADA structure ID: 379



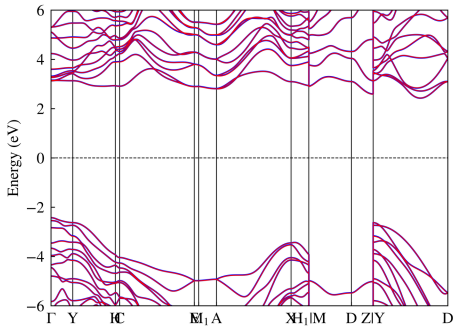
SACADA structure ID: 384



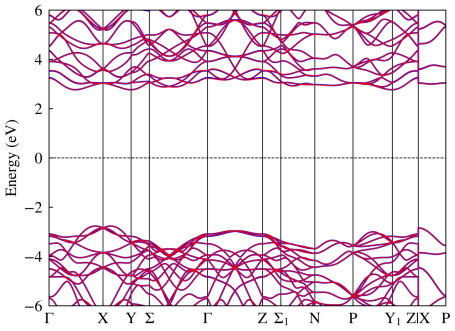
SACADA structure ID: 392



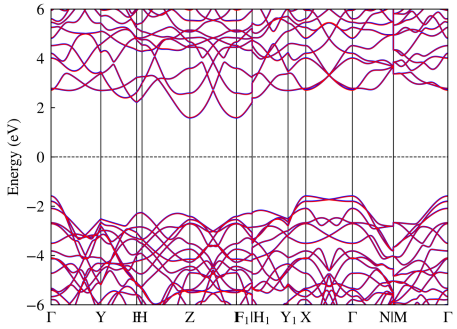
SACADA structure ID: 396



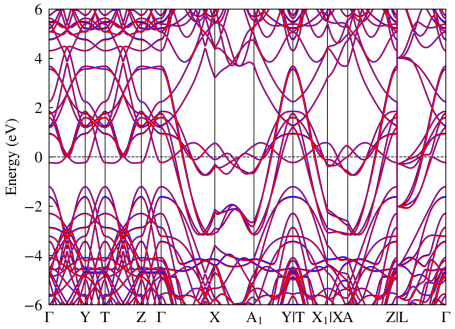
SACADA structure ID: 401



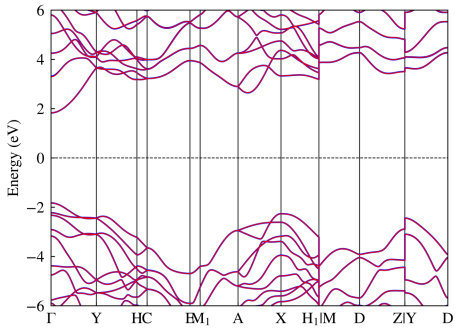
SACADA structure ID: 405



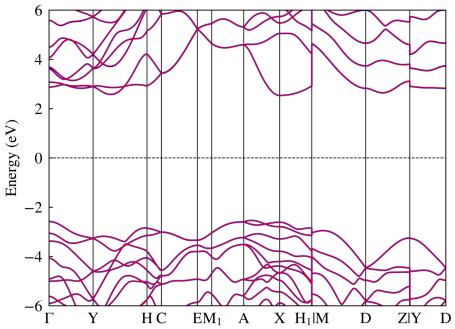
SACADA structure ID: 410



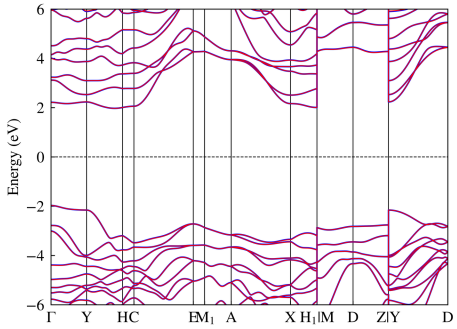
SACADA structure ID: 413



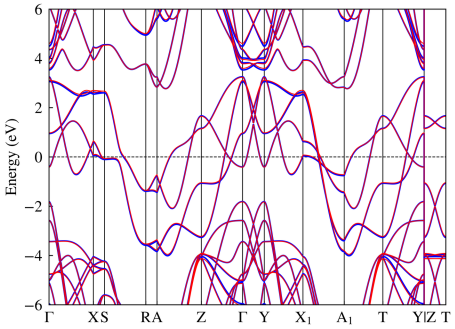
SACADA structure ID: 417



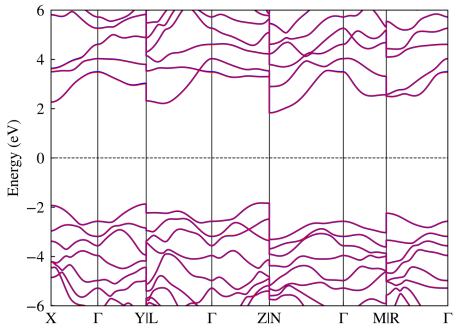
SACADA structure ID: 420



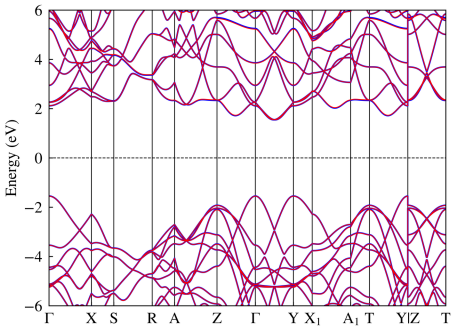
SACADA structure ID: 434



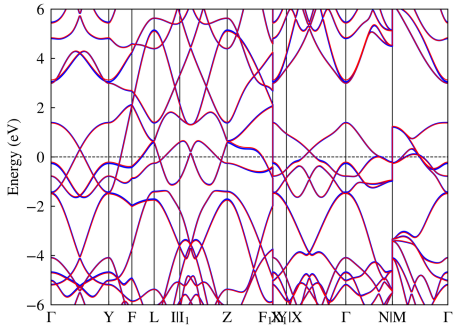
SACADA structure ID: 438



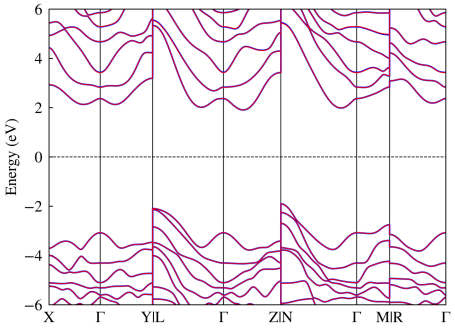
SACADA structure ID: 439



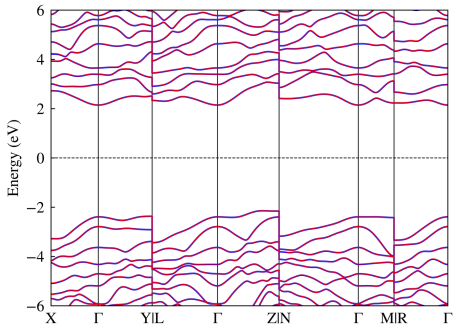
SACADA structure ID: 440



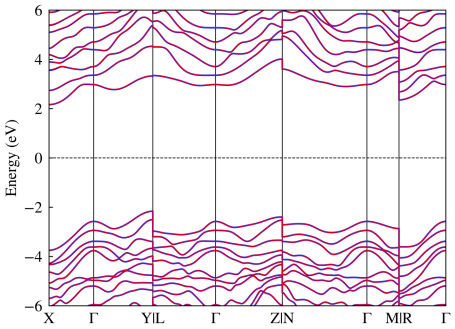
SACADA structure ID: 448



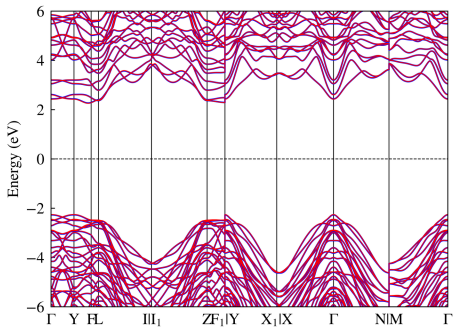
SACADA structure ID: 450



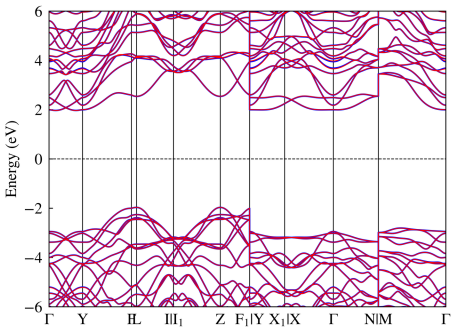
SACADA structure ID: 453



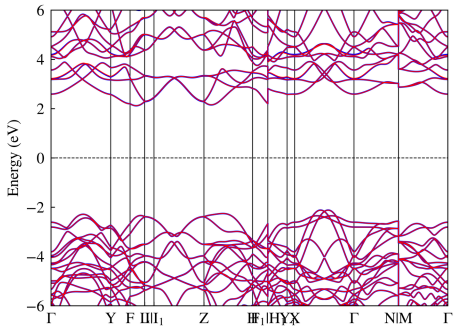
SACADA structure ID: 455



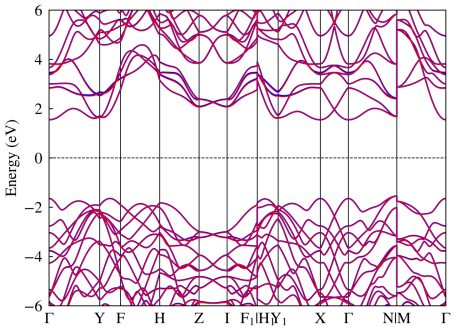
SACADA structure ID: 461



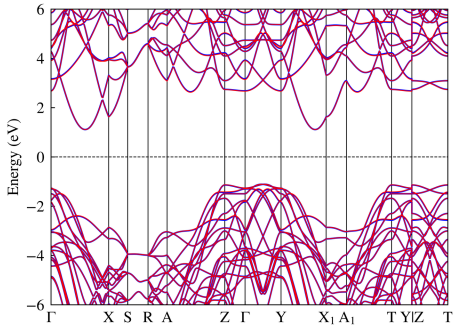
SACADA structure ID: 475



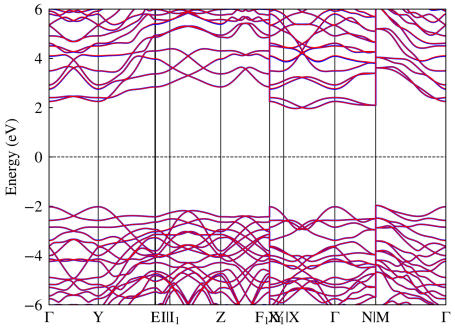
SACADA structure ID: 476



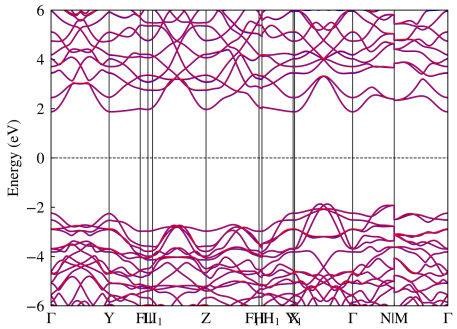
SACADA structure ID: 481



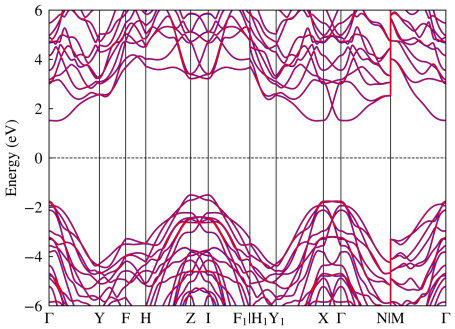
SACADA structure ID: 482



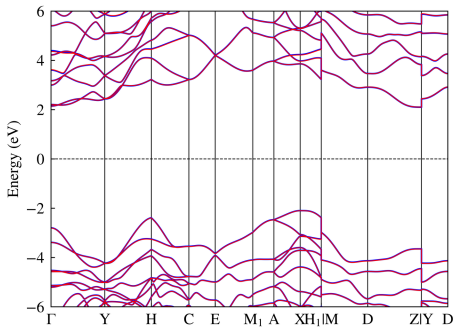
SACADA structure ID: 484



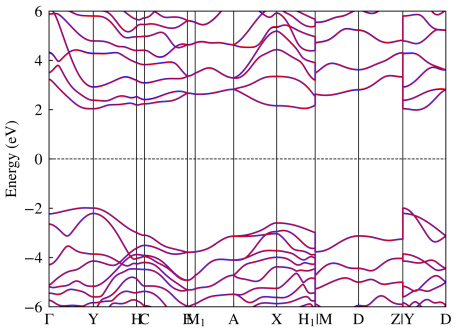
SACADA structure ID: 487



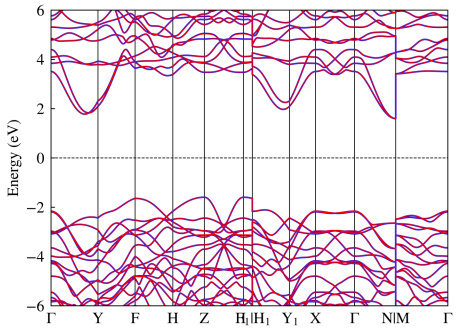
SACADA structure ID: 489



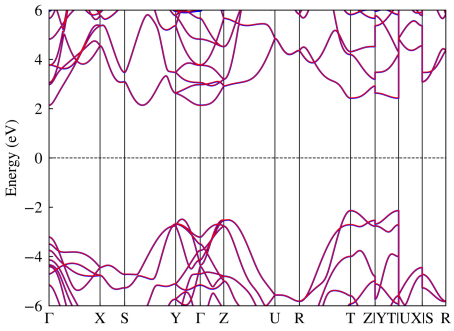
SACADA structure ID: 496



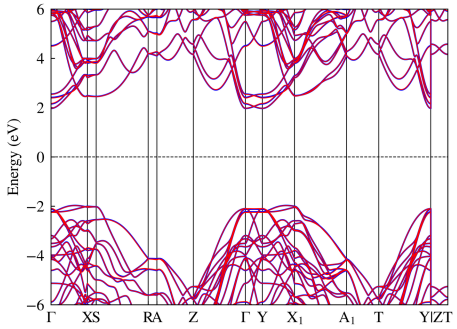
SACADA structure ID: 499



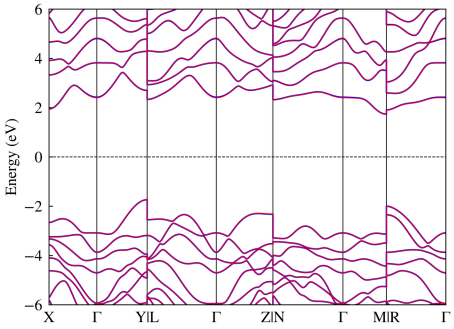
SACADA structure ID: 507



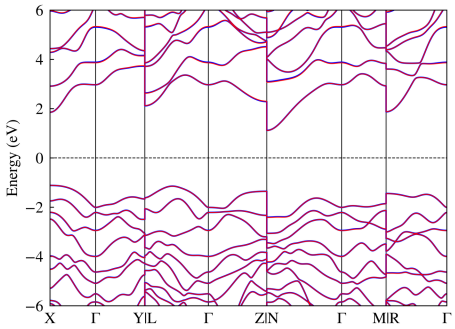
SACADA structure ID: 512



SACADA structure ID: 514



SACADA structure ID: 521



SACADA structure ID: 522

FIG. S6: The band structures of a total of 85 carbon allotropes in test set are displayed. DFT-computed and DeepH-predicted band structures are illustrated with red and blue solid line, respectively. Band structures are labelled by the SACADA structure ID. All these structures are accessible via the SACADA database [6].

[1] S. P. Huber, S. Zoupanos, M. Uhrin, L. Talirz, L. Kahle, R. Häuselmann, D. Gresch, T. Müller, A. V. Yakutovich, C. W. Andersen, F. F. Ramirez, C. S. Adorf, F. Gargiulo, S. Kumbhar, E. Passaro, C. Johnston, A. Merkys, A. Cepellotti, N. Mounet, N. Marzari, B. Kozinsky, and G. Pizzi, AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance, Sci Data **7**, 300 (2020).

[2] M. Uhrin, S. P. Huber, J. Yu, N. Marzari, and G. Pizzi, Workflows in AiiDA: Engineering a high-throughput, event-based engine for robust and modular computational workflows, Comput Mater Sci **187**, 110086 (2021).

[3] H. Li, Z. Wang, N. Zou, M. Ye, R. Xu, X. Gong, W. Duan, and Y. Xu, Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation, Nat Comput Sci **2**, 367 (2022).

[4] T. Ozaki, Variationally optimized atomic orbitals for large-scale electronic structures, Phys Rev B **67**, 155108 (2003).

[5] T. Ozaki and H. Kino, Numerical atomic basis orbitals from H to Kr, Phys Rev B **69**, 195113 (2004).

[6] R. Hoffmann, A. A. Kabanov, A. A. Golov, and D. M. Proserpio, Homo citans and carbon allotropes: For an ethics of citation, Angew Chem Int Edit **55**, 10962 (2016).

[7] C. Wang, S. Zhao, X. Guo, X. Ren, B.-L. Gu, Y. Xu, and W. Duan, First-principles calculation of optical responses based on nonorthogonal localized orbitals, New J Phys **21**, 093001 (2019).