Supplementary Information (SI) for Physical Chemistry Chemical Physics.

Advancing Band Structure Simulations of Complex Systems of C, Si and SiC: A Machine Learning Driven Density Functional Tight-Binding Approach

Guozheng Fan[ab], Yu Jing[a*], and Thomas Frauenheim[cd]

a. Jiangsu Co-Innovation Centre of Efficient Processing and Utilization of Forest Resources, College of Chemical Engineering, Nanjing Forestry University, Nanjing 210037, China. E-mail: yujing@njfu.edu.cn.

b. Bremen Center for Computational Materials Science, University of Bremen, 28359 Bremen, Germany.

c. School of Science, Constructor University, Bremen 28759, Germany.

d. Institute for Advanced Study, Chengdu University, Chengdu 610106, China.

Methods:

Dataset

Table S1: Data collection (lattice, K-mesh, and atoms in the system) for machine learning band structures calculations.

system	lattice	K-mesh	size
Si	diamond	9 × 9 × 9	2
	diamond	7 × 7 × 7	8
	diamond	5 × 5 × 5	64
	diamond slab 100	7 × 7 × 5	24
	diamond slab 110	7 × 7 × 5	24
	diamond slab 111	7 × 7 × 5	24
	diamond defect	5 × 5 × 5	63
	hexagonal	9 × 9 × 5	2
	tetragonal	9 × 9 × 9	2
	tetragonal	5 × 5 × 9	4
C	diamond	9 × 9 × 9	2
	diamond	7 × 7 × 7	8
	diamond	5 × 5 × 5	64
	diamond slab 100	7 × 7 × 5	24
	diamond slab 110	7 × 7 × 5	24
	diamond slab 111	7 × 7 × 5	24
	diamond defect	5 × 5 × 5	63
	hexagonal	9 × 9 × 5	2
	hexagonal[a]	9 × 9 × 5	4
	hexagonal[b]	9 × 9 × 5	4
SiC	diamond[c]	9 × 9 × 9	2
	diamond	7 × 7 × 7	8
	diamond	5 × 5 × 5	64
	diamond slab 100	7 × 7 × 5	24
	diamond slab 110	7 × 7 × 5	24
	diamond slab 111	7 × 7 × 5	24
	diamond defect	5 × 5 × 5	63
	cubic[d]	9 × 9 × 9	2
	cubic	7 × 7 × 7	2

Hexagonal[a] and hexagonal[b] are structures of carbon that correspond to AA- and AB-stacked bilayer graphene, respectively, while hexagonal structures with two atoms represent a lattice cell containing a single layer of graphene. In the SiC diamond system[c], one silicon atom is bonded to four equivalent carbon atoms to form corner-sharing SiC_4 tetrahedra, while in cubic[d], one silicon atom is bonded to six equivalent carbon atoms to form a mixture of corner and edge-sharing SiC_6 octahedra. We obtained a total of 50 geometries for each sub-dataset; however, only 49 of these geometries were used for the carbon vacancy system. One of the geometries failed to converge in the DFT calculations with the hybrid functional, and thus was not included in our study. The K-mesh parameters described in Table S1 were consistently applied to each geometry for DFT and DFTB band structure calculations and DFTB-MD simulations. The high-symmetry K-points for band structure calculations in DFT and DFTB calculations were automatically generated using the method proposed by Setyawan and Curtarolo.[1]

Methods

We used the Tight Binding Machine Learning Toolkit (TBMaLT) to perform all the DFTB-ML training and predictions in the second parametrization step. For geometry optimization and MD simulations to generate the dataset, we employed DFTB+. TBMaLT has been extended to support electronic band structure calculations. Note that in this version, we realized Monkhorst-Pack meshes and K-point meshes along high-symmetry lines for electronic band structure calculations in addition to supporting Gamma-centered meshes in the old version. Geometry optimizations were performed before MD simulations. Geometry optimizations were performed using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)[2] optimization method. We used the NVT ensemble at 1273 K with a Nose-Hoover thermostat.[3] A time step of 1 fs with 1000 steps was used, and the geometries after every 20 steps were recorded as the structures for learning. In the case of slab models, only the surface layer was allowed to move during MD simulations, while for defect systems, the first and second neighboring atoms of the vacancy were allowed to move. The following sections' geometries for band structure calculations are based on the MD simulations. For slab models, a three-layer system with a 15-angstrom vacuum was employed. The supercell of the pristine bulk diamond system consisted of 64 atoms. Defect systems of silicon and carbon were created by removing one Si/C atom from the supercell, resulting in a structure with 63 atoms. For the defect system of silicon carbide, one silicon atom was removed for the silicon carbide system to generate a vacancy.

Parametrization (step 1)

For the traditional parametrization, the compression radii grid points of carbon *s* and *p* orbitals were set to 2.0, 2.25, 2.5, 2.75, 3.0, 3.5, 4.0, and 4.5 Bohr. For the *d* orbital, the grid points were 3.0, 3.5, 4.0, 4.5, 5.0, 6.0, 7.0, and 9.0 Bohr. We start the discussion of silicon systems by using the minimal basis sets *pbc-0-3* parameters.[4] We demonstrate the efficiency of our model without using complex basis sets, such as double zeta basis.[5] All the DFTB calculations are based on the self-consistent charge density functional tight-binding (SCC-DFTB) method.

DFTB-ML (step 2)

The Adam optimizer was employed to optimize the parameters through backpropagation of gradients. Unless otherwise specified, the mean absolute errors (MAEs) were used as the default loss function. The convergence tolerance for all training processes was set at 1×10^{-4}eV. The number of estimators in random forest regression is 100. To determine the optimal size of the training sets, we tested the training ratio of 0.1, 0.2, 0.3, 0.4, and 0.5, respectively. As illustrated in Fig. S3 (a), the training ratio of 0.4 provided the best balance between training set size and predictive accuracy. Therefore, this ratio was used for subsequent training. The testing dataset ratio was set at 0.2. Fig. S3 (b) illustrates the effect of onsite energies on the simulation accuracy of the primitive carbon diamond system. The results suggest that local onsite training and predicting perform the best, which is understandable because local training makes the onsite energies adaptable to the chemical environment. The only difference between global onsite training and no onsite training is that the former process adjusts the on-site energies for the primitive carbon diamond system. At the same time, the latter uses default values for the on-site energies. We can see that the training errors of global onsite and no onsite are very close, which suggests the onsite energies in the first step have been fine-tuned.

For feature representations, the cutoff function G_1, radial symmetry function G_2, and angular function G_4 were applied. The G_1, G_2 and G_4 functions of traditional atom-centered symmetry functions (ACSFs) [6] are:

$$G^1{}_i = \sum_j f_c(R_{ij})$$

$$G^2{}_i = \sum_j f_c(R_{ij}) e^{-\eta(R_{ij}-R_s)^2}$$

$$G^4{}_i = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda cos_{ijk})^{\zeta} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) e^{-\eta(R_{ij}+R_{ik}+R_{jk})^2}$$

where R_c is 10 Å, with η and R_s values of 1.0 for G_2, and η, ζ, and λ values of 0.02, 1.0, and -1.0 for G_4. The feature representations in this study for scaling parameters are:

$$G^1{}_{ij} = f_c(R_{ij}) \left[\sum_k f_c(R_{ik}) + \sum_k f_c(R_{jk}) \right]$$

$$G^2{}_{ij} = f_c(R_{ij}) \left[\sum_k f_c(R_{ik}) e^{-\eta(R_{ik}-R_s)^2} + \sum_k f_c(R_{jk}) e^{-\eta(R_{jk}-R_s)^2} \right]$$

$$G^4{}_{ij} = f_c(R_{ij}) \left[2^{1-\zeta} \sum_{k,m \neq i} (1 + \lambda cos_{imk})^{\zeta} f_c(R_{im}) f_c(R_{ik}) f_c(R_{km}) e^{-\eta (R_{im}+R_{ik}+R_{km})^2} \right.$$

$$\left. + 2^{1-\zeta} \sum_{k,n \neq j} (1 + \lambda cos_{knj})^{\zeta} f_c(R_{jn}) f_c(R_{jk}) f_c(R_{kn}) e^{-\eta (R_{jn}+R_{jk}+R_{kn})^2} \right]$$

The values of η, R_s, ζ, and λ are the same as the above parameters in traditional symmetric functions.

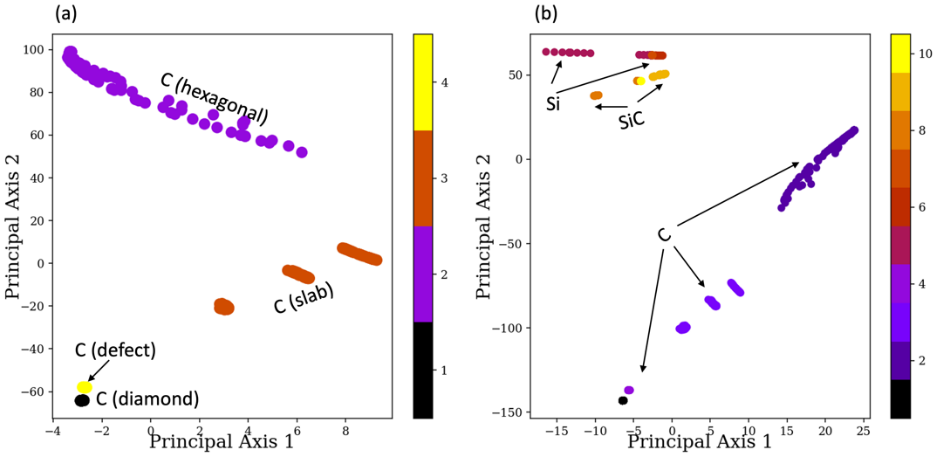

Fig. S1. Kernel principal component analysis (KPCA) projections using Smooth Overlap of Atomic Positions (SOAP) descriptors for (a) carbon systems with different bulk, slab, and defect models and (b) bulk, slab, and defect models for silicon, carbon, and silicon carbide systems.

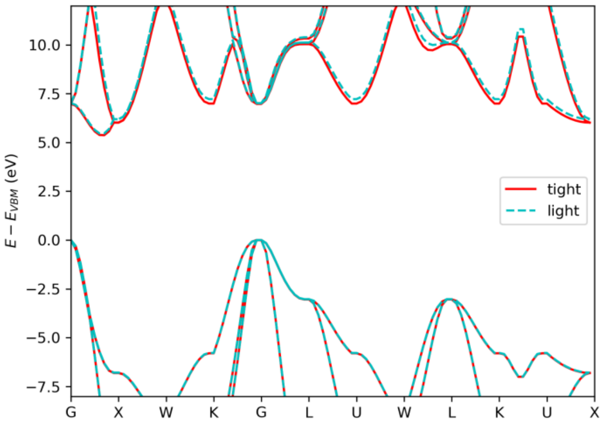

Fig. S2. Silicon diamond band structures performed using the HSE hybrid functional method with the primitive cell. We compare the results obtained using two different basis sets, i.e. light and tight basis sets. Therefore, the band structures of reference data were all calculated using the hybrid functional based on light basis levels. The hierarchy of predefined basis sets includes "light," "intermediate", "tight," and "really-tight." As the basis set from "light" to "really-tight," the accuracy and convergence of the calculation increases. A higher-level basis set, such as the "tight" set, includes all the functions of the lower-level sets, like the "light" set. The basis sets, ranging from light to really-tight, are pre-defined and consist of hydrogen-like, cation-like, or atom-like functions. The "light" basis sets are suitable for fast pre-relaxations and structure searches, while the "tight" sets offer meV-level accuracy in energy differences, even for large structures.

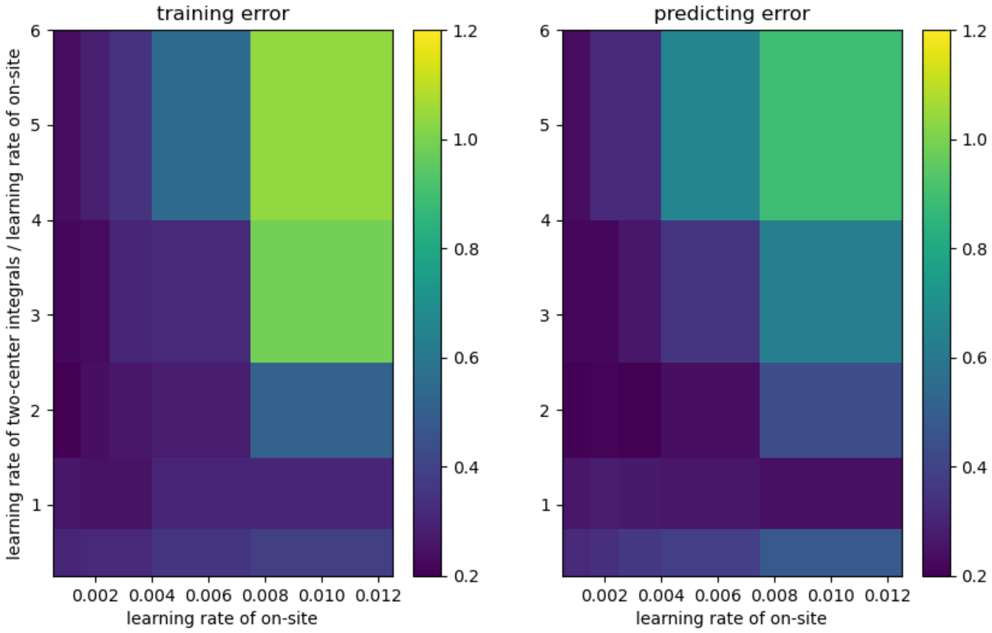

Fig. S3. Effect of learning rates on the training errors and predicting errors using the carbon diamond systems with 2 atoms. The final learning rate for the Hamiltonian integral scaling factors was set to 3×10^{-3},while the learning rate for on-site energies was 1×10^{-3}. Specifically, for carbon systems, the learning rate for the Hamiltonian integral scaling parameters was set to 3×10^{-3}, while the learning rate for on-site energies was 1×10^{-3}. In silicon carbide systems, these rates were 3×10^{-3} and 2×10^{-3} for the Hamiltonian integrals and on-site energies, respectively. For silicon systems, the learning rates were 5×10^{-4} for the Hamiltonian integrals and 4×10^{-4} for on-site energies.

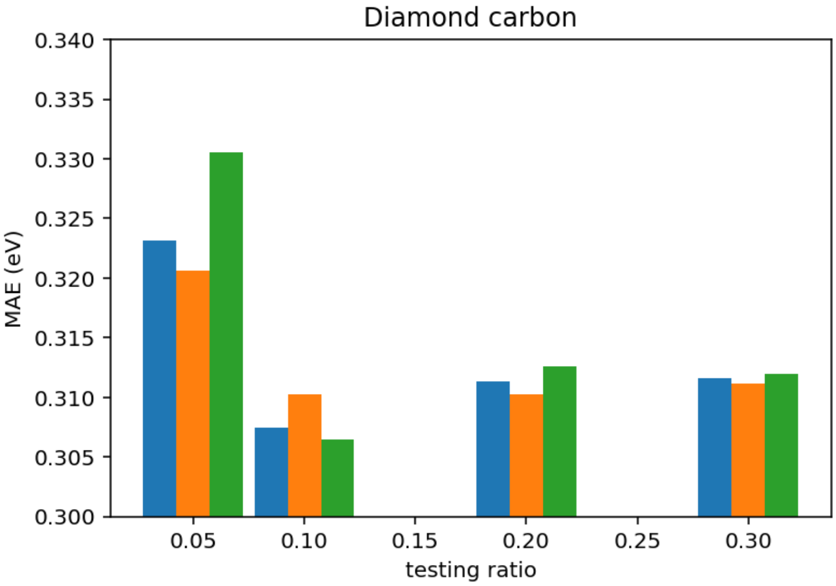

Fig. S4 The effect of different testing ratios on the MAEs for band structure predictions based on three independent runs, using diamond carbon as an example. The results suggest that when the testing ratio is 0.2, the testing results converge, and the differences between the three independent runs become negligible.

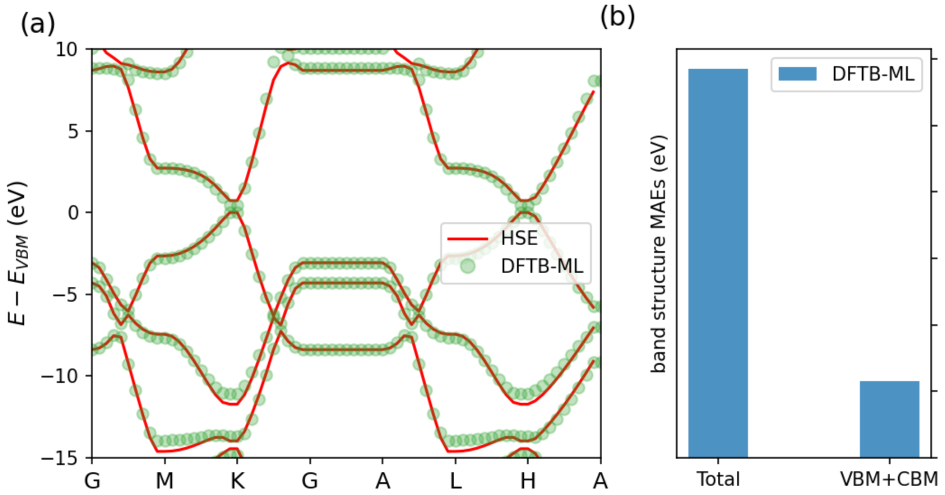

Fig. S5 (a) The band structures of the carbon hexagonal system performed using the DFTB-ML model in comparison with that obtained by DFT-HSE. The geometry was taken from a snapshot of MD step 780. (b) The resulting MAEs for band structure predictions of all trained energy states and energy states near Fermi energy (including VBM and CBM) by comparing the results with that of DFT-HSE calculations.

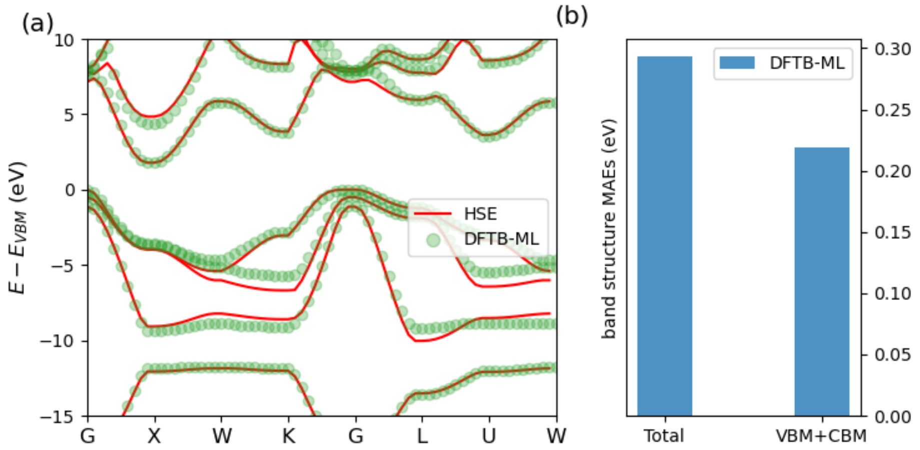

Fig. S6 (a) Band structures of silicon carbide diamond system performed using the DFTB-ML model compared with that of DFT-HSE, with the geometry taken from a snapshot of MD step 80. (b) The resulting MAEs for band structure predictions of all trained energy states and energy states near Fermi energy (including VBM and CBM) by comparing the results with that of DFT-HSE calculations.

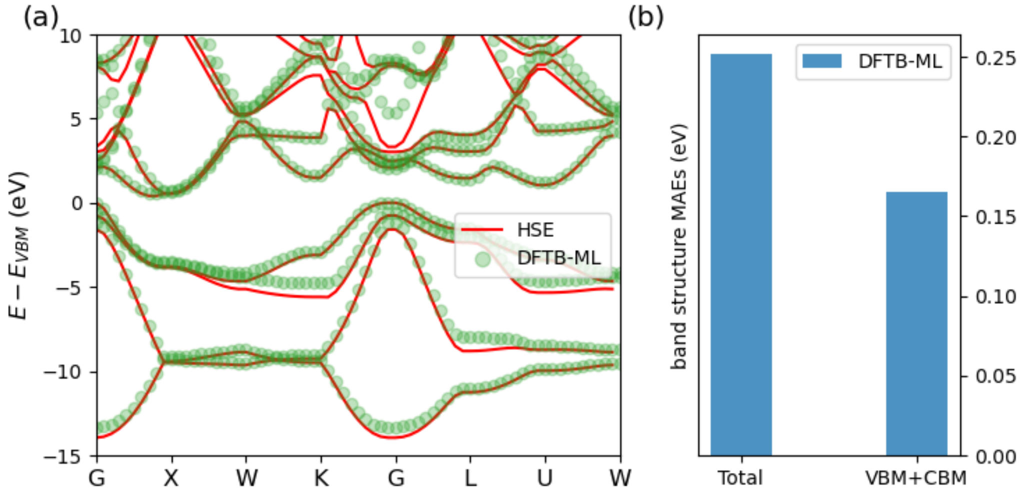

Fig. S7 (a) Band structures of the silicon diamond system performed using the DFTB-ML model compared to that of DFT-HSE, with the geometry taken from a snapshot of MD step 220. (b) The resulting MAEs for band structure predictions of all trained energy states and energy states near Fermi energy (including VBM and CBM) by comparing the results with that of DFT-HSE calculations.

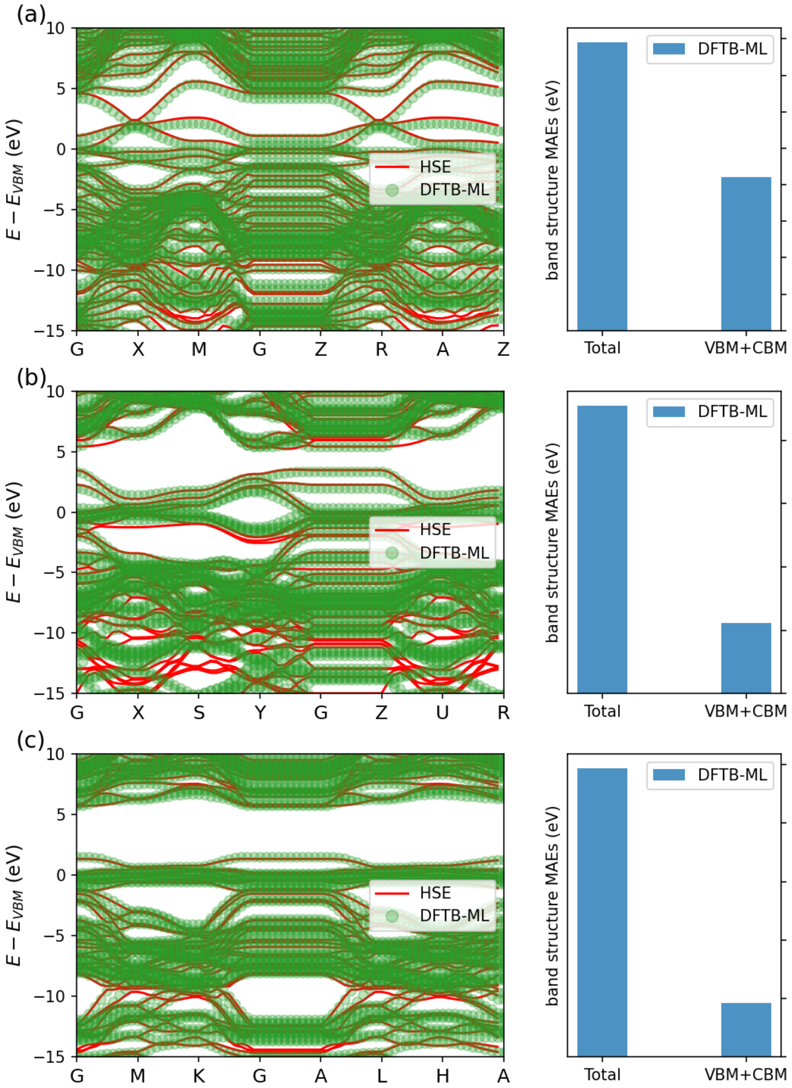

Fig. S8 Band structures of the carbon slab (a) (100) system, (b) (110) system, and (c) (111) system performed using the DFTB-ML model in comparison with that of DFT HSE, with the geometry taken from a snapshot of MD step 440, 920 and 940, respectively. Aside are the resulting MAEs for band structure predictions of all trained energy states and energy states near Fermi energy (including VBM and CBM) by comparing the results with that of DFT-HSE calculations.

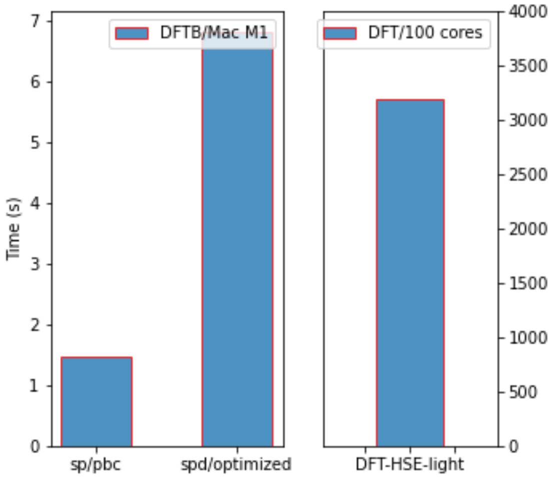

Fig. S9 The DFT-HSE of a 64-atom silicon system using a cluster with 100 cores is 3146.406 s. At the same time, the DFTB calculations were performed using an Apple M1 Pro laptop. The time was 1.5 seconds when only the s and p orbital were applied, and when the extra d orbital was applied in our DFTB-ML approach, the time increased to 6.8 seconds. We can conclude that the DFTB-ML is several orders of magnitude faster than DFT-HSE.

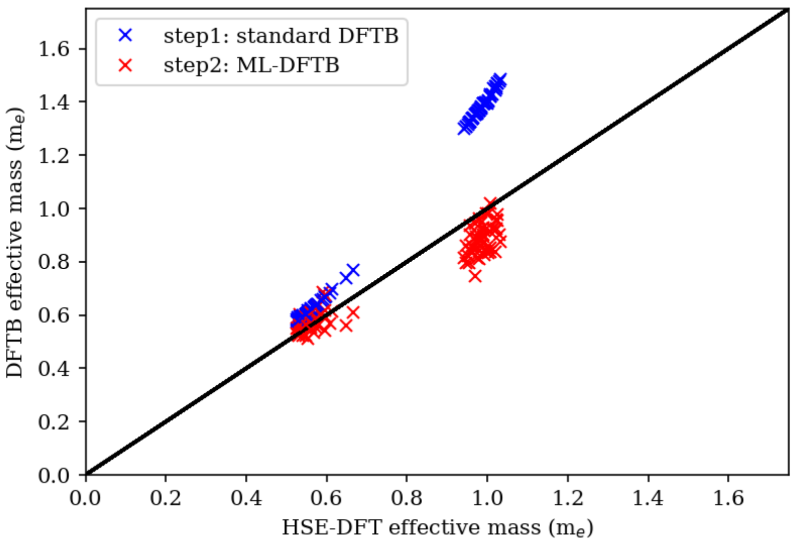

Fig. S10 The effective mass of carbon diamond systems: The red points represent predictions from the DFTB-ML model, while the blue points correspond to traditional SCC-DFTB calculations using parameters from the first-step parametrization. The m_e is the electron mass. The ***k*** vectors for effective VBM and CBM mass calculations correspond to the ***k*** vectors in Figure 3 (c).

Reference:

1 W. Setyawan and S. Curtarolo, *Computational Materials Science*, 2010, **49**, 299–312.
2 J. Nocedal and S. J. Wright, *Numerical optimization*, Springer, New York, 2nd ed., 2006.
3 G. J. Martyna, M. E. Tuckerman, D. J. Tobias and M. L. Klein, *Molecular Physics*, 1996, **87**, 1117–1157.
4 E. Rauls, J. Elsner, R. Gutierrez and Th. Frauenheim, *Solid State Communications*, 1999, **111**, 459–464.
5 M. Elstner, *Theor Chem Acc*, 2006, **116**, 316–325.
6 J. Behler, *J. Chem. Phys.*, 2011, **134**, 074106.