

<https://doi.org/10.1038/s41524-025-01634-1>

A machine learning approach to predict tight-binding parameters for point defects via the projected density of states

Check for updates

Henry Phillip Fried ¹✉, Daniel Barragan-Yani¹, Florian Libisch² & Ludger Wirtz ¹

Calculating the impact of point defects on the macroscopic properties of technologically relevant semiconductors remains a considerable challenge. Semi-empirical approaches, such as the tight-binding method, are very efficient in calculating the electronic structure of large supercells containing one or several defects. However, the accuracy of these calculations depends on the quality of the parameters. Obtaining reliable parameters by fitting to the large number of entangled bands in defective supercells is a demanding task. We therefore present an alternative way by fitting to the atom and orbital projected densities of states. Starting with a tight-binding fit of the pristine material, we only need a few physically motivated parameters for the fitting of defects. The training is done on data sets generated purely with parameter variations of tight-binding Hamiltonians. We demonstrate the efficiency of our approach for the calculation of the carbon monomer and the carbon dimer substitutions in hexagonal boron nitride. The method opens a path towards understanding complicated defect landscapes using a computationally affordable semi-empirical approach without sacrificing accuracy.

The calculation of electronic properties of defects in solids using density-functional theory (DFT) in a supercell approach is now a well-established and frequently used method^{1,2}. Total energy calculations give access to formation energies, charge transition levels or even optical transitions^{3–6}. For relatively simple and localized defect states, it is even feasible to investigate the optical properties through many-body perturbation theory^{7–15}. However, studying the impact of defects on transport, mechanical, or excitonic properties of complex materials requires a large number of atoms. This challenges pure ab-initio approaches and suggests the use of complementary methods.

Recently, there has been a growing interest in applying machine learning techniques to the scalability of electronic structure calculations. The predictive power of neural networks can utilize information obtained by DFT on smaller systems to construct Hamiltonians for larger systems (e.g., by obeying the system symmetries)^{16–18}.

Another use of machine learning is in the context of semi-empirical tight-binding, where it has emerged as a powerful regression tool. For example, Wang et al.¹⁹ used machine learning for a tight-binding parameterization where the neurons of the network act as matrix elements. Nakhaee et al.²⁰ fitted Slater-Koster parameters for BiTeCl by optimizing the parameters through machine learning. Soccodato et al.²¹ and Gu et al.²² obtain distance-dependent tight-binding (TB) parameters by applying local

changes to the structures. However, all of the mentioned methods either require training data sets based on many DFT calculations, or they emphasize that a small cell is necessary to achieve a reasonable TB fit.

These efforts are complemented by the use of machine learning within the self-consistent charge density-functional tight-binding method which aim to accurately perform geometry relaxations^{23,24}. The goal of such general tight-binding approaches is the ability to predict, for a given set of atomic positions, the corresponding tight-binding Hamiltonian and the resulting energy and forces.

For the description of point defects in their dilute limit, large super cells in real space are necessary. This results in a large density of intertwined bands in reciprocal space. Disentangling the bands (in the sense of making sure that the right tight-binding band is fitted to the right abinitio band) becomes challenging. In recent years, quantitative parameterizations of the tight-binding model for defects, fitting to ab-initio calculations, have been attempted. One way is a transformation from the (delocalized) DFT Bloch states to localized Wannier functions²⁵. This approach remains challenging in the presence of many entangled bands, although some progress has been made by Fontana et al.²⁶. By enforcing the localization of the orbitals with the help of additional “spread functionals”, they were able to obtain a Wannier fit of the electronic bands of the NV⁻ center in diamond in a 64 atom supercell.

¹Department of Physics and Materials Science, University of Luxembourg, Luxembourg, EU, Luxembourg. ²Institute for Theoretical Physics, TU Wien, Vienna, EU, Austria. ✉ e-mail: henry.fried@uni.lu

One way to overcome the problem of band disentanglement in large supercells is by evaluating the Green's function $G(\mathbf{r}, \mathbf{r}'; E)$ of the perturbed system (i.e., the defective material). The imaginary part \Im of its diagonal elements is proportional to the local density of states $\rho(\mathbf{r}; E)$:

$$-\frac{1}{\pi} \Im G(\mathbf{r}, \mathbf{r}; E) = \rho(\mathbf{r}; E). \quad (1)$$

This enables the qualitative study of defective systems and has been widely applied for defects in graphene^{27–32}. In particular, nitrogen substitution in graphene has been studied by Lambin et al.³² by introducing a dependence of the onsite energies on the distance from the defect. They optimize the tight-binding parameters with the local density of states.

In this contribution, we follow the idea of using the local (i.e., atom and orbital projected) density of states for the description of defects. In contrast to earlier work, we make it quantitative by optimizing a small (but physically relevant) set of defect-related parameters through machine learning. We note that a fitting by machine learning of (total) electronic densities of states for various periodic systems (including defects) has been performed within the TBMaLT toolkit of DFTB^{33,34}. In our work, we introduce the fit to the atom and orbital projected density of states as the essential ingredient to obtain not only a reliable result for the (global and local) densities of states but also for the electronic band structure, a topic that was mentioned in ref. 34 to be addressed in future studies.

Considering the balance between accuracy and transferability, we choose an approach different from training on a large set of DFT calculations. We rather create a large training data set directly by using our lean tight-binding model and varying its parameters in a certain range around the values for the pristine material. With the trained neural network, we then efficiently parameterize a defect using only *one single DFT calculation* for this defect.

A similar approach in this direction has been made by Schattauer et al. who used hopping parameters obtained from a pristine Wannierization as a starting point to describe different vacancy defects in graphene³⁵. They have trained a machine learning model with band structures obtained by a distant-dependent Wannier parameters and showed the capability of the model to describe a vacancy in a transition metal dichalcogenide monolayer via the variation of pristine Slater-Koster parameters found in the literature. However, this approach is still based on the band structure, which becomes increasingly challenging to interpret for larger supercells due to the expected complex band folding (Fig. 1). For larger cells, a model based on real-space instead of reciprocal-space data seems useful.

In this paper, we describe a systematic way to obtain an intuitive, accurate, and transferable set of tight-binding parameters for the electronic structure of defects. We present a procedure to obtain such a model with the help of machine learning. We start with the tight-binding parameters of the pristine material. The defect (in this work we consider only substitutional atoms) is then treated as a perturbation of the pristine material and enters the TB model in four ways: (i) it has a different onsite energy and hopping parameters to its nearest neighbors, mimicking the difference in chemical properties of the substitutional atom. (ii) In the vicinity of the defect, the shift of onsite energies is modeled by a Gaussian dependence on the distance from the defect. (iii) An extra hopping parameters between the nearest neighbor of the defect is introduced to capture the impact on the pristine parameters in the close vicinity. (iv) The defect may lead to a distortion of the surrounding lattice. This distortion is taken into account through the distance dependence of the tight-binding parameters of the pristine material.

Our approach keeps the number of fitting parameters for the defective cell to a minimum. As standard fitting tools fail to consistently predict a good parameter set, we employ a machine learning algorithm for this task. The machine learning model is trained with a set of local densities of states calculated in the tight-binding approximation. Consequently, no expensive DFT calculations are involved in the training. Once trained, the models are given the projected density of states from DFT and provide the

corresponding tight-binding parameters that properly reproduce both the projected density of states and band structures (obtained with DFT) of supercells with defects.

To demonstrate our method, we study three cases of substitutional defects in monolayer hexagonal boron nitride (hBN): the carbon substitution of a boron atom (C_B), the carbon substitution of a nitrogen atom (C_N) and the carbon dimer substitution ($C_B C_N$). For single carbon substitutions, there are many-body calculations of the band structures and optical absorption levels of the defects^{14,15}. The carbon dimer is the main candidate for the 4.1 eV peak observed in the PL-spectra, as shown by constrained hybrid DFT calculations³⁶. Semi-empirical models, even at a tight-binding level, have not yet been made available for these defects. Thus, in addition to serving as benchmark calculations for the development of our method, our results will enable the exploration of defects in hBN, which is a promising host for defect-related single-photon emission³⁷.

This paper is structured as follows: We first discuss the problems of fitting the defect parameters in “The band structure problem”, where we also explain the general workflow. This is followed by motivating the machine learning approach and explaining its components, after which we introduce our approximations for the defect tight-binding Hamiltonian. We then benchmark our results with carbon defects in hBN by means of the projected density of states (PDOS) and the band structure, before concluding in “Discussion”.

Results

The band structure problem

To properly describe a defect in an otherwise perfect crystal, using periodic boundary conditions, the defect is usually embedded within a sufficiently large supercell of the host material. The construction of a supercell leads to the back-folding of the bands of the primitive unit cell into the correspondingly reduced 1st Brillouin zone of the supercell (Fig. 1). The introduction of a defect (here a carbon dimer substitution) then leads to a perturbation of the valence and conduction bands and, depending on the defect, to the appearance of one or more localized (i.e., weakly dispersing) states in the band gap. The description of such a system using a tight-binding Hamiltonian obtained by jointly fitting the defect bands and the perturbed valence/conduction bands becomes a formidable task. Directly fitting the TB bands to the respective DFT bands (i.e., minimizing $|\epsilon_{nk}^{DFT} - \epsilon_{nk}^{TB}|^2$ for all n and k) requires their disentanglement: the respective eigenvalues have to be from the same band n at the same k . Tweaking tight-binding parameters to shift the “wrong” eigenstate to match a DFT energy will actually worsen the overall fit. The disentanglement of the bands becomes cumbersome quite quickly with increasing supercell size, and Wannierization reaches its limits already for defective 2D materials³⁵.

Using instead the PDOS offers several advantages. Most importantly, the key problem of the disentanglement of the bands is overcome by projecting the density of states onto the orbitals and atomic sites. Furthermore, the PDOS at the defect site and its close neighbors converges quickly as a function of supercell size. Even in the limit of large and complex defect supercells (where DFT calculations are restricted to a single k -point), the PDOS is still accessible and reliable.

The PDOS has significant advantages for the fitting process, but the perturbation of the host material by the defect also influences the tight-binding Hamiltonian which presents another challenge. The reduction of the symmetry from the primitive cell to the defect supercell leads to many new parameters. In principle, all site energies and the parameters for all possible hoppings between the first, second, and third nearest neighbors within the supercell would need to be refitted. In the case of a defective hBN supercell of 162 atoms, we would be faced with more than 1000 parameters. Although this number can be reduced by symmetry³⁵, in the present work we aim for a minimal parameter set. We therefore employ two distance dependencies for the pristine parameters, resulting in a simpler fitting problem.

Specifically, we used a neural network that predicts these tight-binding parameters for defects from the PDOS. The work flow is shown in Fig. 2. The

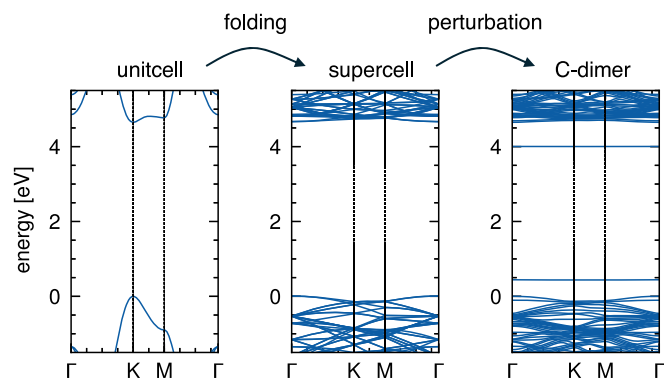


Fig. 1 | Illustration of the problems associated with a defective band structure. DFT band structures along high symmetry points of a hBN unit (2 atoms), of a $9 \times 9 \times 1$ pristine supercell (162 atoms), and of a $9 \times 9 \times 1$ supercell containing a carbon dimer. Increasing the unit cell leads to folding of the band structure. The presence of the defect then perturbs the bands and introduces states within the band gap.

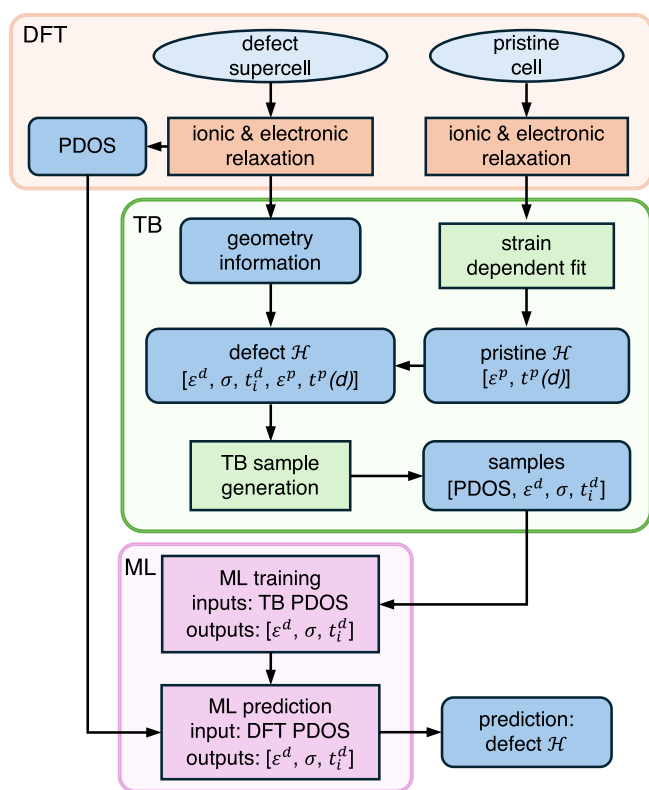


Fig. 2 | Workflow to obtain tight-binding parameters for defects. Oval nodes indicate the starting points, cornered nodes represent steps including calculations and rounded nodes represent data. DFT related calculations are displayed in orange, TB in green and ML in purple. Although some similarities are present to Schattauer et al.³⁵, instead of refitting the pristine parameters, we treat the defect as a perturbation to reduce the number of fitting parameters. Furthermore, we introduce an additional step for the generation of the pristine model, where we calculate a distance (strain) dependence of the TB parameters for the long-range description. Last but not least, instead of using the band structure, we use the PDOS to avoid the cumbersome disentanglement of the bands.

starting points are optimized geometries of the pristine cell and the defect supercell, calculated using DFT. The pristine tight-binding model is obtained by fitting to the ab-initio band structure in the primitive cell. A distance dependence of the hopping parameters can be obtained by refitting to ab-initio calculations of strained unit cells. Once the pristine tight-binding

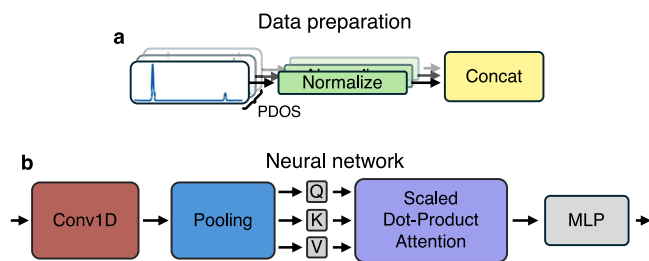


Fig. 3 | Scheme of the data preparation and the neural network. Upper part: scheme of the data preparation. PDOS $D(\mathbf{r}_i; E)$ at different sites \mathbf{r}_i is normalized separately. Lower part: Scheme of the neural network, displaying the order of the different parts of the network.

parameters have been established, we used the defect geometry obtained from a DFT relaxation to build the defect Hamiltonian as a perturbation of the strain-dependent pristine Hamiltonian.

Random sampling of the tight-binding parameters that define this Hamiltonian generates the data sets on which we train, and the neural network learns the mapping between the tight-binding PDOS (input) and the tight-binding parameters (output). A key benefit is that the training data is solely acquired within the tight-binding framework, making it computationally inexpensive. After the training, a DFT PDOS is used to predict tight-binding parameters that accurately describe the Hamiltonian defect. Before discussing the details of the tight-binding model, we will introduce the machine learning approach to obtain the parameters.

Machine learning

The key task in our approach is to find tight-binding parameters that give rise to localized defect states within the band gap and simultaneously describe the bands properly. Using standard fitting tools to directly fit the tight-binding PDOS to DFT provides a first hint on the complexity of this task (Supplementary Information). Without prior knowledge of the defect parameters, standard fitting tools struggle to find the localized peak and simultaneously describe the bands properly. Another problem is the scalability of this procedure. While with a small number of parameters, one can still manually fit the parameters, this becomes increasingly difficult the more parameters need to be fitted.

In recent years, state-of-the-art fitting tools have been complemented by machine learning algorithms, enabling the exploration of a complex parameter space with many parameters. It is the goal of this article to make this connection between machine learning and fitting TB parameters via the PDOS: The neural network learns the direct correspondence between the PDOS and the parameters used for the generation, minimizing the chance of being stuck at a local minimum and making sure that the defect state and the bands are properly described.

The PDOS $D(\mathbf{r}; E)$ as a fitting property has multiple advantages, but it also inherits some difficulties. The considered defects are deep defects within hBN, a material with a large band gap. For such a material, $D(\mathbf{r}; E)$ is mostly zero within the band gap, with narrow defect peaks with large contributions at the defect site, resulting in few activations in the first layer.

To prepare the data for the training, we first simplify the problem by considering $D(\mathbf{r}; E)$ only at a small set of points \mathbf{r}_i and only within a restricted energy range around the band gap to train only on the 1D PDOS data. In addition, we include the sum $\sum_i D(\mathbf{r}_i; E)$ for training, since we get this additional information free. We then normalize each $D_i(E) = D(\mathbf{r}_i; E)$ (between 0 and 1) separately as normalization of the data is essential for training and allows access of all features equally at every site (Fig. 3).

The separately normalized $D_i(E)$ are then stacked before entering a 1D convolution (conv1D) layer. The aim is to capture the correlated nature of the PDOS, but also to compress the data and simultaneously reduce the fraction of zeros (Supplementary Fig. 1). Here, the stride defines how much the data is squished while the kernel size determines the broadening of the PDOS features. The output of the conv1D layer is pooled along the channel

direction and averages the outputs for different kernels to maintain the features of the input.

The pooling of the conv1D is followed by a scaled dot-product attention as proposed by Vaswani et al.³⁸. The aim is to include long-range dependencies between the defect states and bands. This is achieved by separating the input in query (Q), key (K), and value (V). A matrix-multiplication of the weighted Q and K results in a dependency between each of the entries of the respective inputs. Finally, the matrix is multiplied by the weighted value matrix to incorporate the dependencies into the output. The short formulation of the process reads as³⁸

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (2)$$

here the softmax function associates the matrix entries with a probability and $1/\sqrt{d_k}$ stabilizes the training.

The final component of the network architecture is a multi-layer perceptron, which consists of fully connected linear layers. For each layer the ReLU function is used as an activation function. The architecture of the network is depicted in Fig. 3. The final output is evaluated by the standard mean-square error (MSE)

$$\text{loss} = \frac{1}{N} \sum_n (y_n - \hat{y}_n)^2, \quad (3)$$

where \hat{y}_n is the generated and y_n the expected output of the network. We tested including the Gaussian relation for ϵ^{defect} and σ for a given d , but we found no improvement for the final results.

This network is trained using the data produced with the tight-binding approximation. The data set consists of PDOSs, serving as input, and the corresponding labels $[\epsilon^d, \sigma, t_n^d]$ as output. After training, the network can predict the tight-binding parameters from DFT PDOS obtained from a single calculation.

The network has been developed in parallel with the tight-binding model. Since such architectures solved the parametrization problem at hand, we did not increase its complexity further. For more complex materials and more parameters to be fitted, this can be adjusted. Possible adjustments include further separation of the training at different levels of the neural network. For example, a separation of the conv1D layer with padding in reflection mode enables more data from the bands. One can also use multi-head attention or include more scaled dot-product attention mechanism similar to a transformer neural network.

Defect tight-binding model

The tight-binding Hamiltonian is an approximation of the full many-body Hamiltonian projected onto the localized atomic orbital basis $|n\rangle$ located at site n . The Hamiltonian can be formulated as follows

$$\hat{\mathcal{H}}_0 = \sum_n \epsilon_n^{\text{prist}} |n\rangle\langle n| + \sum_{n,m} |n\rangle t_{nm} \langle m|, \quad (4)$$

where diagonal matrix entries $\langle n|\hat{\mathcal{H}}_0|n\rangle$ are described with the onsite energies $\epsilon_n^{\text{prist}}$, whereas the off diagonal elements, $\langle n|\hat{\mathcal{H}}_0|m\rangle$, are referred to as hopping matrix elements, t_{nm} , from site m to n . Due to the localized nature of the orbitals (fast decay), the hopping parameters can be restricted to the n -th nearest neighbor (usually $n \leq 3$), limiting at the same time the number of hopping parameters and the accuracy of the model. The Hamiltonian can be obtained by a fit of the parameters to experiments or ab-initio calculations.

In the case of a monolayer of hBN, we employ a pristine model including the p_z orbitals of both N and B with nonzero hopping parameters up to the 3rd nearest neighbors. Furthermore, we use different 2nd nearest neighbor parameters for N-N and B-B hoppings, t_{2nd}^{NN} and t_{2nd}^{BB} , in order to reproduce the asymmetry in the highest valence and lowest conduction

Table 1 | Strain dependent tight-binding parameters

| | ϵ_N | ϵ_B | t_{1st}^{NB} | t_{2nd}^{NN} | t_{2nd}^{BB} | t_{3rd}^{NB} |
|----------|--------------|--------------|----------------|----------------|----------------|----------------|
| | 0.0 | 2.37 | -3.12 | 0.002 | -0.77 | -0.35 |
| α | | | -2.66 | 0.18 | 0.05 | -0.19 |
| β | | | 5.86 | -0.08 | 0.97 | 0.55 |

The table contains the pristine tight-binding parameter, including the two onsite energies and four hopping parameters. The second nearest neighbor hopping is regarded separately for nitrogen and boron. The parameters α and β refer to the quadratic fit for the distance dependent hopping parameters.

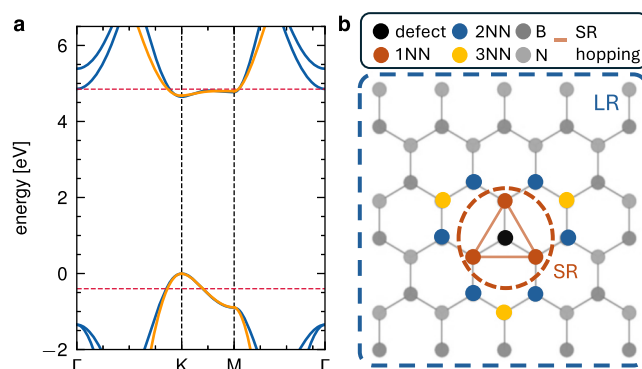


Fig. 4 | Pristine band structures and hopping parameter models in the presence of the defect. **a** DFT (blue) and the TB (orange) band structure fitted to reproduce the DFT band structure along \overline{MK} . Red dotted line indicates range for later PDOS calculations. **b** Defect and its nearest neighbors and the adaptations used for the pristine hopping parameters. The short-range (SR) impact on the hopping parameters is described with an additional parameter between the nearest neighbors (1NN) of the defect. The long-range impact effects all other hopping parameters via a distance dependence between the hopping neighbors.

band. Together with the two onsite energies, a total of six tight-binding parameters have been obtained and are shown in table 1. To best describe the valence band maximum and the conduction band minimum, these have been fitted to an ab-initio band structure along \overline{KM} (Fig. 4a).

In the more general case of a defective monolayer of hBN, the perturbation introduced by the defect is naturally expected to significantly influence the physics of the system, including changes in the positions of neighboring atoms. The defect has different chemical properties which we account for by new parameters. A new onsite parameter ϵ^{defect} is complemented by three hopping parameters to its nearest neighbors similar to the pristine model, all having the same distance to the defect, as there is no Jahn-Teller distortion for the investigated defects.

Although this introduction of four additional TB parameters for the defect site is straightforward, it is not sufficient because the defect also has a subtle and delicate influence on the chemical properties of its surroundings. Instead of refitting all the hoppings and site energies in the defect supercell, we limit the number of additional fitting parameters to two by shifting the onsite energy of each atom as a function of its distance from the defect site. Additionally, the hopping between pristine neighbors may vary due to geometric deformation. This is taken into account by making the corresponding hopping parameters dependent on the hopping distance.

For the onsite energies, we employ a dependence on the distance from the defect. We use a Gaussian function to describe the perturbation potential introduced by the defect. The function is defined as follows

$$\epsilon(\epsilon^{\text{defect}}, \sigma, d) = \epsilon^{\text{prist}} + [\epsilon^{\text{defect}} - \epsilon^{\text{prist}}] \cdot e^{-\frac{d^2}{2\sigma^2}}, \quad (5)$$

where d is the distance to the defect site and σ is the variance, which is directly proportional to the width of the potential. The dependency is

modeled such that the height of the potential is represented by the difference of the onsite energy of the defect (ϵ^{defect}) and the host atom (ϵ^{prist}). This approach is similar to the approach of Lambin et al.³² who used a Gaussian potential to investigate the N_C substitution defect in graphene in the TB approximation. Accounting for the defect and its perturbation on the onsite parameters (height and width of the Gaussian function), this results in a total of five parameters for the defect itself.

For the defect's influence on the hopping parameters, we distinguish between short-range and long-range effects. The first model contribution accounts for the local change of the chemical environment introduced by the defect. This is done by means of an additional hopping parameter between its nearest neighbors and is referred to as the short-range (SR) model (Fig. 4b). Because no Jahn-Teller distortion is present for the studied monomers, only one extra parameter needs to be fitted for the defective supercell.

The second contribution to the model accounts for the long-range (LR) component of the perturbation, and it is an expansion of the pristine tight-binding model by a distance dependence between the hopping neighbors. With it we aim at capturing the perturbation on the geometry of the host crystal (Fig. 4b). The four hopping parameters were fitted to strain-dependent band structures, while the onsite energies were kept fixed for later adjustments. For the studied defects, at the DFT level, the maximum change in the distances of the host atoms is 3.7% for the C_N and 3.5% for the C_B . A biaxial strain within a range of 4% is therefore sufficient to capture the distance dependence for which the band structure was still well described. The strain-dependent fit of the hopping parameters can be found in Supplementary Fig. 4.

The best description of the long-range distance dependence is found to be a quadratic dependence

$$t(d) = \alpha \cdot (d - d_{\text{prist}})^2 + \beta \cdot (d - d_{\text{prist}}) + t_{\text{prist}}, \quad (6)$$

which is constructed to reproduce the pristine hopping parameters for the pristine distance between atoms. It adds two more parameters to the respective pristine hopping terms. An advantage of our method is that this distance dependency is naturally well suited for applications where strain plays a role.

To benchmark our methods, we investigated the carbon monomers and the carbon dimer. Both show deep defect states respectively and are perfect candidates to test the workflow. The defect states within the band gap are also p_z orbitals and therefore match the p_z model for the pristine crystal. In the following, we will benchmark the tight-binding method by fitting to the PDOS and comparing the band structures. We will investigate, in particular the importance of including SR and/or LR contributions to the TB model.

Benchmark results

For both carbon monomers, we want to answer the question of how important adjustments of the pristine hopping parameters are. To do so, we analyze the individual contributions of all components of our model. Therefore, we use the complete model (SR + LR), along with models obtained by inclusion of either short or long range (SR, LR) or the exclusion of both (w/o).

We first check the performance of the neural network with respect to the number of PDOS used as an input for the neural network. To identify the trends, we take into account the density of states projected to the defect site, its first, second and, finally, its third nearest neighbors. Our analysis shows that the usage of the PDOS at three sites is sufficient to describe the monomers (Supplementary Information). We can use these results and generate new data sets with a denser sampling of the PDOS to obtain the best parameters (Table 2).

These parameters are then used for the evaluation of the hopping contributions. We compare the sum of the PDOS values and their cosine similarities up to the 6th nearest neighbor (defect site and 3 atom sites of each element). Since the defect peak outweighs the differences in the bands,

Table 2 | Defect tight-binding parameters

| C_N | ϵ^{defect} | σ | t_{1st}^{CB} | t_{2nd}^{CN} | t_{3rd}^{CB} | t_{SR}^{BB} |
|----------|------------------------------|------------|------------------------------|----------------|----------------|---------------|
| SR + LR | 2.611 | 0.110 | -2.772 | -0.044 | -0.415 | -0.830 |
| SR | 2.813 | 0.106 | -2.835 | -0.015 | -0.471 | -0.904 |
| LR | 3.208 | 0.094 | -2.932 | 0.028 | -0.400 | |
| w/o | 3.092 | 0.088 | -2.917 | -0.031 | -0.380 | |
| C_B | ϵ^{defect} | σ | t_{1st}^{CN} | t_{2nd}^{CB} | t_{3rd}^{CN} | t_{SR}^{NN} |
| SR + LR | 0.142 | 0.091 | -3.164 | -0.758 | -0.735 | -0.090 |
| SR | -0.268 | 0.092 | -3.414 | -0.801 | -0.698 | -0.323 |
| LR | 0.053 | 0.086 | -3.106 | -0.830 | -0.837 | |
| w/o | 0.178 | 0.092 | -3.102 | -0.761 | 0.683 | |
| C_{BN} | $\epsilon_N^{\text{defect}}$ | σ_N | $\epsilon_B^{\text{defect}}$ | σ_B | t_{1st}^{CC} | |
| LR | 4.363 | 0.077 | 0.303 | 0.124 | -3.003 | |

The table contains the predicted parameters for the respective hopping models used for the comparison of the monomers and the parameters for the carbon dimer. Superscript of the hopping terms denote the elements of the hopping neighbors. Subscripts for the dimer denote the species of the substituted host atom. The hopping terms for the dimer are the hopping terms of the respective w/o model. The energies are in eV.

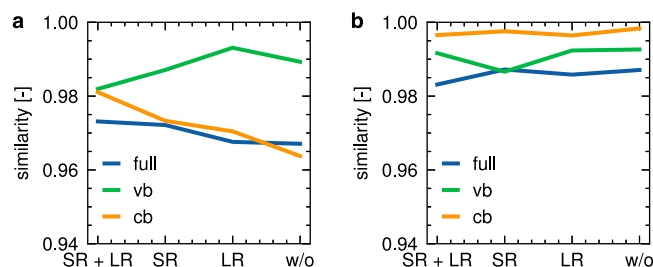


Fig. 5 | Cosine similarities for different hopping descriptions. Cosine similarity of the different hopping description, including the complete model (SR+LR), either just the short-range (SR) or the long-range (LR) model, and without any adaptation. The cosine similarity is calculated for the valence (vb), conduction band (cb), and the full PDOS of C_N in (a) and C_B in (b). The corresponding PDOS of the respective bands can be found in Fig. 6.

we separate the evaluation into different parts of the PDOS, namely the valence band (-0.4–0.1 eV) and the conduction band (4.6–4.85 eV). The cosine similarities of the respective contributions are shown in Fig. 5 and reveal no major differences with high similarities of above 0.96. The valence band is described slightly worse for the models including an additional parameter within the SR contribution. The opposite is observed for the conduction band, where the additional parameter improves the cosine similarity. For all contributions, the cosine similarity of the full PDOS does not change considerably. The C_B , is described equally well for all different hopping models.

The good agreement of similarities is also reflected in the direct comparison of the summed PDOS of the bands (Fig. 6). Again, we observe a slightly worse description for the valence band when including the SR contribution for the C_N , which is reflected in a shift of the valence band maximum. In general, we find that the differences for all contributions are negligible.

This indicates that the perturbation of both monomers to the host material is sufficiently accounted for without any hopping contribution. Therefore, considering new defect tight-binding parameters and their impact on the pristine onsite energies is enough to describe the carbon monomers. This is a key advantage for the monomers, as it enables a tight-binding model with few fitting parameters, but also for more complex defects like the carbon dimer.

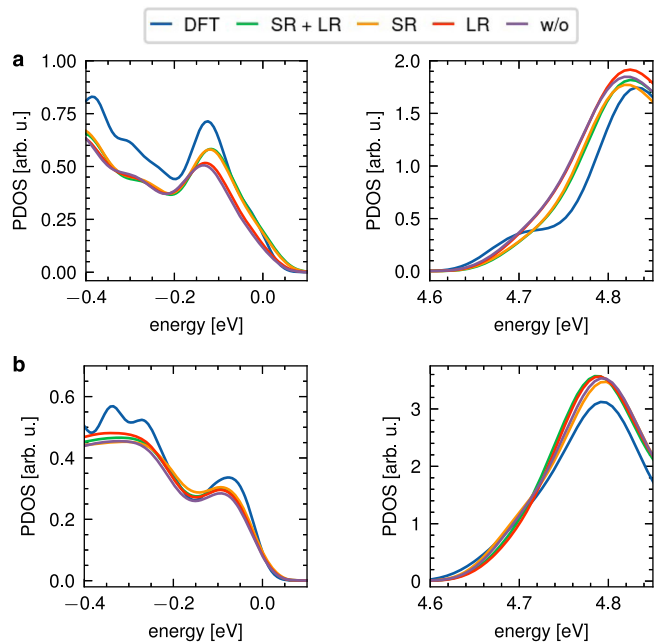


Fig. 6 | Summed projected densities of states of the valence and conduction bands. Comparison of the summed PDOS at the defect and its 6 first nearest neighbors of the CN in (a) and CB in (b) for DFT and the different hopping approximations.

The carbon dimer consists of two carbon atoms that substitute for one pair of neighboring nitrogen and boron atoms. This results in a shift of the two respective monomer states towards the band edges, an effect that is similar to that of two atoms forming a molecule with a bonding and an antibonding state. This requires the introduction of an additional hopping parameter t_{1st}^{CC} . However, our fitting attempts show that this is not sufficient to obtain a good fit for the positions of both defect peaks at the same time. In order to achieve this, one needs to take into account the additional symmetry breaking for the carbon dimer that leads to new distances in the vicinity of the defect. To properly account for all changes, more hopping parameters than those already used for the respective monomers would be necessary. However, we have obtained a distance dependence for the carbon hopping matrix elements that describe the evolution of the respective peaks under compressive and tensile strain (Supplementary Information). These can be used to properly account for the different distances of the carbon atoms from its neighbors. Our calculations for varying distances between carbon monomers indicate that for distances where the Gaussian perturbations of the respective defects overlap, the parameters are not transferable (Supplementary Information). Thus, we have to fit the respective Gaussian dependencies (ϵ_N^{defect} , σ_N , ϵ_B^{defect} , σ_B) and the new hopping parameter (t_{1st}^{CC}).

In conclusion, the analysis of the long- and short-range contributions showed no major improvements for the monomers. However, to accurately describe strained conditions, the long-range part should be included, since the distance dependence also impacts the peak position. Therefore, further benchmark calculations for the two monomers are performed with the LR contribution. For the fitting of the carbon dimer, we use the model described above, in which we only need to fit five parameters.

The PDOS on the full energy range that we trained provides additional insight into the descriptive power of the predicted tight-binding parameters (Fig. 7). Both carbon monomers show a very good agreement between their respective tight-binding and DFT PDOS. In particular, for C_N , the second nearest neighbor has a larger contribution to the defect peak than the first, whereas the opposite is observed for C_B . Both are properly described with the respective tight-binding Hamiltonians.

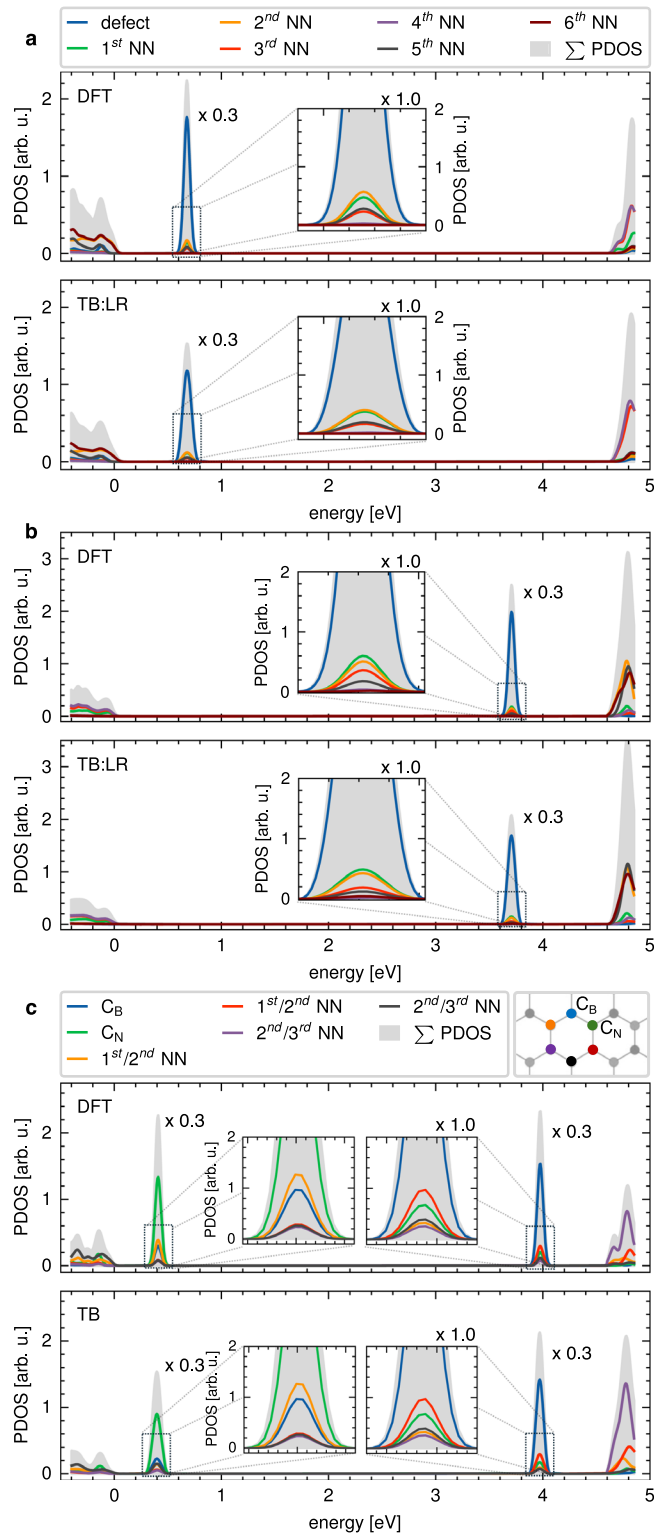


Fig. 7 | Projected densities of states of the carbon monomers and the carbon dimer. Projected densities of states obtained from DFT (upper subplots) and from the fitted TB parameters (lower subplots) of the (a) C_N and (b) of the C_B for the defect and its first 6 nearest neighbors including the sum of all. c Projected densities of states of the C_{BN} for the defects and the neighbors depicted in the upper-right sub-panel. The defect peaks are scaled by 0.3 for better visibility in comparison to the band edge contributions to the PDOS. The inset shows the defect peak without rescaling to assess the different contributions of the neighbors.

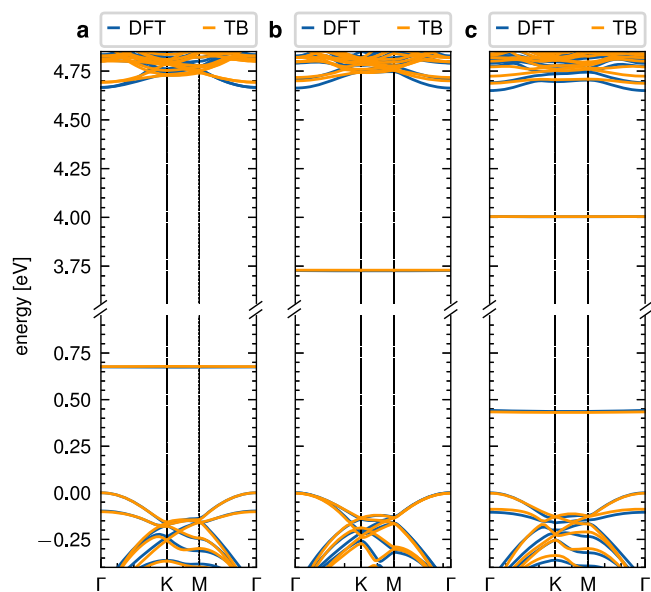


Fig. 8 | Band structures of the carbon monomers and the carbon dimer. DFT (blue) and tight-binding (orange) band structures of **a** C_N , **b** C_B and **c** C_{BN} . The tight-binding band structures are calculated with the parameters obtained from the fit to the PDOS (LR). The energy axis has been cut.

For the carbon dimer, we use the simplified model described in the previous section, and we use PDOS at both defect sites to obtain the parameters. We observe some differences for the contributions of different atomic sites for the peaks, but the results are satisfactory as we were able to properly describe the main features by only fitting five parameters.

In summary, our results show that the machine learning algorithm is capable of predicting tight-binding models for carbon defects in hBN via the PDOS. Although the tight-binding descriptions are relatively simple, they are able to reproduce the DFT PDOS fairly well. The remaining question is whether the PDOS as a fitting observable is enough to also reproduce the band structure of the defective supercell.

We compare the tight-binding and DFT band structures in Fig. 8, where we use the same tight-binding parameter as for the previous PDOS. As expected, the narrow defect peaks observed in the PDOS result in a correct description of the undispersive defect states in the band gap. The introduction of the defect results in the splitting of the bands from which the defect states emerge. This feature is captured by all tight-binding Hamiltonians as well. However, for C_B , we observe an underestimation of this effect which is enhanced by the difference of the conduction band minimum already observed in the pristine fit. This difference is related to the simple pristine model with six parameters. It is not able to capture simultaneously the shape of the conduction band and the position of its minimum (Supplementary Fig. 5). More hopping parameters might fix this, but the aim of this work is to use a pristine model with its limitations to describe a defect. An improvement of the pristine model should also result in a better description of the defective model.

The conduction band of the carbon dimer shows a split off band which is captured with the tight-binding model. Although the splitting itself is similar, we observe a difference between the conduction band minima, similar to the C_B . Nevertheless, our very simple model reproduces the main features of the band structure, including the valence band, the defect states and the split off conduction band.

Overall, we have demonstrated that the use of the PDOS to obtain tight-binding parameters also shows good agreement for the band structures, thus enabling a novel method to obtain a tight-binding model for defective systems. The observed differences are related to the simplicity of tight-binding models (pristine and defective), rather than to the fitting method.

Discussion

We propose a new way to construct semi-empirical models, at a tight-binding level, of technologically relevant defects by means of a machine learning approach. The embedding of the defect perturbs the pristine Hamiltonian, and the refit of all the parameters quickly gets out of hand. We suggest using a perturbed version of a pristine tight-binding model along with additional defect parameters to overcome this problem and to enable a description with only a few parameters.

We make use of the atom and orbital projected density of states to obtain the parameters, instead of the common way of fitting parameters to the band structure. This overcomes the known problem of disentanglement of the bands. We demonstrate that a neural network is capable of predicting a tight-binding model from the projected density of states (PDOS) for defects with localized peaks within a large band gap. We show that these parameters are also reproducing the band structure, thus confirming the PDOS as a well suited observable for fitting.

We investigate different adjustments of the pristine hopping terms and find that neither a short-range (additional parameter) nor long-range contribution (distance dependent hopping parameters) show a significant improvement of the results. However, we propose to include the long range contribution, as it allows to describe strain dependencies. Therefore, a very simple tight-binding model is able to deliver an adequate description of carbon defects in hBN.

Our tight-binding approach is able to describe defects under strained conditions. The distance dependence for the hopping parameters obtained for the pristine crystal under strained conditions is transferable, and we only need to introduce a distance dependence for the carbon hopping terms. Both distance dependencies have been obtained for biaxial strain. A comparison of uniaxial strain suggests that they accurately describe the behavior. We used these results to build a tight-binding model for the well known carbon dimer, expected to be responsible for the single photon emission at around 4.1 eV.

Calculations for varying distances between two monomers indicate that the parameters are also transferable to systems where the defects are closer to each other. Comparisons for different distances show that the DFT results are well reproduced up to the point where the respective Gaussian functions for the onsite perturbations overlap.

We emphasize that while this has been done for relatively simple defects, the use of the projected density of states is expected to be transferable to more complex systems. The projection on the orbitals enables a simple access to the description of multi-orbital states. The machine learning architecture might need to be adapted to describe a more complex tight-binding description, but this is not expected to be a barrier.

In conclusion, we show that a tight-binding model consisting of defect parameters and a perturbation of the pristine onsite energies is sufficient to describe the carbon monomers in monolayer hBN. We propose a simple model for the carbon dimer that makes use of the respective results for the monomers, thus enabling a description with few parameters for more complex carbon defects. We demonstrate that the tight-binding Hamiltonian can be fitted via the projected density of states with a neural network. This opens a novel pathway for obtaining tight-binding parameters for defective systems.

Methods

Computational details

The ab-initio calculations have been performed with the VASP software package^{39–42}. For all density functional theory calculations, we used the Perdew-Burke-Ernzerhof (PBE) exchange correlation functional and the projector augmented-wave method. The pristine unit cell has been optimized prior to the defect calculations using a plane wave cutoff at 580 eV and a $9 \times 9 \times 1$ Monkhorst-Pack grid for the sampling of the first Brillouin zone. The total energy did not change more than 1×10^{-4} eV after a vacuum distance of 10 Å.

The defects were embedded in a supercell of $9 \times 9 \times 1$ to include all high-symmetry points in the folding process. The forces have been

converged to be below 1×10^{-3} eV/Å, whereas the electronic optimization converged for 1×10^{-5} eV. The projected density of states was post processed from the VASP calculations with the VASPKIT package⁴³. The projected density of states converged for a Monkhorst-Pack grid of $9 \times 9 \times 1$ in the 1st BZ of the supercell (corresponding to $81 \times 81 \times 1$ in the 1st BZ of the primitive unit cell) with a Gaussian smearing of 0.04 eV.

The tight-binding models have been implemented using the pybinding package⁴⁴. The projected density of states has been calculated with the kernel polynomial method and a Gaussian broadening of 0.04 eV as implemented in pybinding.

For the training of the neural network, we used the Adam optimizer with a learning rate of $\eta = 1.0 \times 10^{-3}$ and an exponential learning rate decay with $\lambda = 0.99$. Before the training we split the data into train, validation and test data (80/10/10) to monitor the training and possible over-fitting. The training converges for 700 epochs (Supplementary Fig. 3) and takes around 1.5–3 h on a NVIDIA Tesla V100 SXM2 16G/32G.

Data availability

The VASP input files and the plot data of this study are available via the Zenodo repository at <https://doi.org/10.5281/zenodo.15359254>. Tight-binding and machine-learning scripts are available from the corresponding author upon reasonable request.

Received: 14 September 2024; Accepted: 31 March 2025;

Published online: 11 June 2025

References

- Alkauskas, A., Deák, P., Neugebauer, J., Pasquarello, A. & Walle, C. G. *Advanced Calculations for Defects in Materials*, <https://doi.org/10.1002/9783527638529> (Wiley, 2011).
- Freyssoldt, C. et al. First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **86**, 253–305 (2014).
- Alkauskas, A., McCluskey, M. D. & Van de Walle, C. G. Tutorial: defects in semiconductors—combining experiment and theory. *J. Appl. Phys.* **119**, 181101 (2016).
- Alkauskas, A., Buckley, B. B., Awschalom, D. D. & Walle, C. G. V. First-principles theory of the luminescence lineshape for the triplet transition in diamond nv centres. *N. J. Phys.* **16**, 073026 (2014).
- Alkauskas, A., Lyons, J. L., Steiauf, D. & Walle, C. G. First-principles calculations of luminescence spectrum line shapes for defects in semiconductors: the example of gan and zno. *Phys. Rev. Lett.* **109**, 267401 (2012).
- Razinkovas, L., Doherty, M. W., Manson, N. B., Walle, C. G. & Alkauskas, A. Vibrational and vibronic structure of isolated point defects: the nitrogen-vacancy center in diamond. *Phys. Rev. B* **104**, 045303 (2021).
- Rinke, P., Janotti, A., Scheffler, M. & Walle, C. G. Defect formation energies without the band-gap problem: combining density-functional theory and the *gw* approach for the silicon self-interstitial. *Phys. Rev. Lett.* **102**, 026402 (2009).
- Chen, W. & Pasquarello, A. Accuracy of *gw* for calculating defect energy levels in solids. *Phys. Rev. B* **96**, 020101 (2017).
- Wu, F., Galatas, A., Sundararaman, R., Rocca, D. & Ping, Y. First-principles engineering of charged defects for two-dimensional quantum technologies. *Phys. Rev. Mater.* **1**, 071001 (2017).
- Smart, T. J., Wu, F., Govoni, M. & Ping, Y. Fundamental principles for calculating charged defect ionization energies in ultrathin two-dimensional materials. *Phys. Rev. Mater.* **2**, 124002 (2018).
- Li, D., Liu, Z.-F. & Yang, L. Accelerating GW calculations of point defects with the defect-patched screening approximation. *J. Chem. Theory Comput.* **19**, 9435–9444 (2023).
- Karsai, F. et al. *F* center in lithium fluoride revisited: comparison of solid-state physics and quantum-chemistry approaches. *Phys. Rev. B* **89**, 125429 (2014).
- Tiwald, P. et al. Ab initio perspective on the Mollwo-Ivey relation for *F* centers in alkali halides. *Phys. Rev. B* **92**, 144107 (2015).
- Attaccalite, C., Bockstedte, M., Marini, A., Rubio, A. & Wirtz, L. Coupling of excitons and defect states in boron-nitride nanostructures. *Phys. Rev. B* **83**, 144115 (2011).
- Kirchhoff, A., Deilmann, T., Krüger, P. & Rohlfing, M. Electronic and optical properties of a hexagonal boron nitride monolayer in its pristine form and with point defects from first principles. *Phys. Rev. B* **106**, 045118 (2022).
- Wang, Z. et al. Graph representation-based machine learning framework for predicting electronic band structures of quantum-confined nanostructures. *Sci. Chin. Mater.* **65**, 3157–3170 (2022).
- Gong, X. et al. General framework for E(3)-equivariant neural network representation of density functional theory Hamiltonian. *Nat. Commun.* **14**, 1–10 (2023).
- Zhong, Y., Yu, H., Su, M., Gong, X. & Xiang, H. Transferable equivariant graph neural networks for the Hamiltonians of molecules and solids. *npj Comput. Mater.* **9**, 1–13 (2023).
- Wang, Z. et al. Machine learning method for tight-binding Hamiltonian parameterization from ab-initio band structure. *npj Comput. Mater.* **7**, 11 (2021).
- Nakhaee, M., Ketabi, S. A. & Peeters, F. M. Machine learning approach to constructing tight binding models for solids with application to BiTeCl. *J. Appl. Phys.* **128**, 215107 (2020).
- Socodato, D., Penazzi, G., Pecchia, A., Phan, A.-L. & Maur, M. A. D. Machine learned environment-dependent corrections for a *spds*^{*} empirical tight-binding basis. *Mach. Learn.: Sci. Technol.* **5**, 025034 (2024).
- Gu, Q. et al. Deep learning tight-binding approach for large-scale electronic simulations at finite temperatures with ab initio accuracy. *Nat. Commun.* **15**, 1–12 (2024).
- Li, H., Collins, C., Tanha, M., Gordon, G. J. & Yaron, D. J. A density functional tight binding layer for deep learning of chemical Hamiltonians. *J. Chem. Theory Comput.* **14**, 5764–5776 (2018).
- Liu, C., Aguirre, N. F., Cawkwell, M. J., Batista, E. R. & Yang, P. Efficient parameterization of density functional tight-binding for 5f-elements: a Th–O case study. *J. Chem. Theory Comput.* **20**, 5923–5936 (2024).
- Marzari, N., Mostofi, A. A., Yates, J. R., Souza, I. & Vanderbilt, D. Maximally localized Wannier functions: theory and applications. *Rev. Mod. Phys.* **84**, 1419–1475 (2012).
- Fontana, P. F., Larsen, A. H., Olsen, T. & Thygesen, K. S. Spread-balanced Wannier functions: robust and automatable orbital localization. *Phys. Rev. B* **104**, 125140 (2021).
- Skrypnik, Yu. V. & Loktev, V. M. Impurity effects in a two-dimensional system with the Dirac spectrum. *Phys. Rev. B* **73**, 241402 (2006).
- Wehling, T. O. et al. Local electronic signatures of impurity states in graphene. *Phys. Rev. B* **75**, 125425 (2007).
- Lherbier, A., Blase, X., Niquet, Y.-M., Triozon, F. & Roche, S. Charge transport in chemically doped 2D graphene. *Phys. Rev. Lett.* **101**, 036808 (2008).
- Wu, S. et al. Average density of states in disordered graphene systems. *Phys. Rev. B* **77**, 195411 (2008).
- Pereira, V. M., Santos, J. M. B. & Castro Neto, A. H. Modeling disorder in graphene. *Phys. Rev. B* **77**, 115109 (2008).
- Lambin, P., Amara, H., Ducastelle, F. & Henrard, L. Long-range interactions between substitutional nitrogen dopants in graphene: electronic properties calculations. *Phys. Rev. B* **86**, 045448 (2012).
- McSloy, A. et al. TBMaLT, a flexible toolkit for combining tight-binding and machine learning. *J. Chem. Phys.* **158** <https://doi.org/10.1063/5.0132892> (2023).
- Sun, W. et al. Machine learning enhanced dftb method for periodic systems: learning from electronic density of states. *J. Chem. Theory Comput.* **19**, 3877–3888 (2023).

35. Schattauer, C., Todorović, M., Ghosh, K., Rinke, P. & Libisch, F. Machine learning sparse tight-binding parameters for defects. *npj Computational Mater.* **8**, 116 (2022).
36. Mackoīt-Sinkevičienė, M., Maciaszek, M., Walle, C. G. & Alkauskas, A. Carbon dimer defect as a source of the 4.1 eV luminescence in hexagonal boron nitride. *Appl. Phys. Lett.* **115** <https://doi.org/10.1063/1.5124153> (2019).
37. Tran, T. T., Bray, K., Ford, M. J., Toth, M. & Aharonovich, I. Quantum emission from hexagonal boron nitride monolayers. *Nat. Nanotechnol.* **11**, 37–41 (2016).
38. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).
39. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
40. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Mater. Sci.* **6**, 15–50 (1996).
41. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
42. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
43. Wang, V., Xu, N., Liu, J.-C., Tang, G. & Geng, W.-T. Vaspkit: a user-friendly interface facilitating high-throughput computing and analysis using vasp code. *Computer Phys. Commun.* **267**, 108033 (2021).
44. Moldova, D., Anđelković, M. & Peeters, F. pybinding v0.9.5: a python package for tight-binding calculations (v0.9.5) <https://doi.org/10.5281/zenodo.4010216> (2020).

Acknowledgements

This research was in part funded by the Luxembourg National Research Fund (FNR), grant reference PRIDE17/12246511/PACE and in part by the Austrian Science Fund (FWF) 10.55776/COE5. We would like to acknowledge Christoph Schattauer and Mohamed Ali Abdulmalik for fruitful discussions.

Author contributions

All authors contributed to the idea of the project. H.P.F. developed the tight-binding theory under supervision of D.B., F.L., and L.W. He also developed the machine learning architecture and did the coding. All authors contributed to the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-025-01634-1>.

Correspondence and requests for materials should be addressed to Henry Phillip Fried.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025