Deep Learning Model for Inverse Design of Semiconductor Heterostructures with Desired Electronic Band Structures

Artem Pimachev  
University of Colorado Boulder

Sanghamitra Neogi (✉ sanghamitra.neogi@colorado.edu)  
University of Colorado Boulder  https://orcid.org/0000-0003-2743-1313

Article

Keywords:

Posted Date: April 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2536716/v1

License: ☒  This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

Additional Declarations: There is NO Competing Interest.
Deep Learning Model for Inverse Design of Semiconductor Heterostructures with Desired Electronic Band Structures

Artem K Pimachev\textsuperscript{1} and Sanghamitra Neogi\textsuperscript{1,\textdagger}

\textsuperscript{1}Ann and H.J. Smead Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, Colorado 80303, USA
Email: sanghamitra.neogi@colorado.edu
(Dated: February 1, 2023)

First-principles modeling techniques have shown remarkable success in predicting electronic band structures of materials. However, the computational costs make it challenging to use them for predicting band structures of semiconductor heterostructures, that show high variability of atomic structures. We propose a machine learning-assisted first-principles framework that bypasses expensive computations and predicts band structures from the knowledge of atomic structural features. Additionally, the framework directly connects modeling results and experimental data. For example, it accepts images obtained with angle-resolved photoemission spectroscopy as input and predicts the corresponding atomic structures. The framework leverages the physical relationship between atomic environments and bands. We demonstrate the framework using silicon/germanium based superlattices and heterostructures. Once trained on silicon-based systems, the framework can even predict band structures of gallium arsenide thin films. The physics-informed framework establishes an approach to expedite the design and discovery of complex materials with desired band structures, going beyond combinatorial approaches.

Semiconductor heterostructures are important condensed matter systems, both for fundamental research as well as for device applications [1]. Exploiting the full potential of these systems requires a complete understanding of their electronic band structures. The band structures dictate the physical response of the materials to external fields and determine their electrical, magnetic and optical properties. The electronic bands of layered materials or heterostructures are strongly affected by the variations of atomic structures, caused by the fabrication processes. Even in heterostructures grown with state-of-the-art nanofabrication techniques, the presence of various imperfections becomes unavoidable [2–4]. The imperfections strongly influence the structural order. The heterostructure may inherit approximate structural order of separate layers or exhibit a mixed character. First-principles modeling techniques, such as density functional theory (DFT), have shown remarkable success in predicting electronic properties of materials. Electronic bands of heterostructures are calculated using supercells that are chosen to be as large as is feasible to represent the structural order. Due to the varied structural order, the supercell bands may resemble the Bloch states of either, or both of the layers. Different unfolding algorithms [5–10] have successfully identified the Bloch character of supercell bands, especially for alloys [11, 12]. However, the high computational costs make it challenging to analyze all possible heterostructures using first-principles techniques. A first-principles approach to predict and design band structures of heterostructures has not been established. As a result, combinatorial approaches are being pursued to design heterostructures that exhibit desired electronic properties. The key questions remain: what Bloch character do bands of fabricated heterostructures exhibit? how closely do they correspond to the bands of constituting materials? and how can one design heterostructures that exhibits bands with desired characters?

In this article, we present a first-principles modeling framework to expedite the prediction and design of band structures of semiconductor heterostructures. The framework is assisted by machine learning (ML) techniques and establishes a relationship between atomic structural environments and electronic bands. Physical principles are increasingly being incorporated in ML-based approaches to accelerate prediction of materials properties [13]. A recent study showed that awareness of structural motifs enables ML model to accurately predict bandgaps of metal oxides [14]. Our previous ML model exploited the relationship between local structures and global electronic properties of semiconductor structures. The model successfully predicted electronic transport properties of experimental heterostructures [15]. The present framework reveals the Bloch character of the heterostructure bands. Additionally, the framework directly connects first-principles modeling and experiments. For example, our framework can accept images from angle-resolved photoemission spectroscopy (ARPES) as input and predict the atomic structures of the corresponding systems. The atomic structures could be compared with structural characterization data. The comparison not only will demonstrate the model performance but identify structural features that result in desired band structures. Our framework thus facilitates the inverse design of semiconductor heterostructures with desired electronic properties.

We implement the framework using two ML models: (1) a forward learning model that accepts atomic struc-
FIG. 1. Outline of forward ML model: (a) Training strain-symmetrized and strained superlattices with various periods and compositions. Representative supercells of (i) Si₄Ge₄, (ii) Si₁₄Ge₁₄, (iii) Si₁₃Ge₁₃Si₁₃Ge₁₃ and (iv) Si₂₆Ge₂₆ superlattices. (b) Representative reference cell to obtain unfolded spectral functions (SF). Two types of descriptors are used: Element type and structural features. (c)(i)-(iv) Property value corresponding to descriptors: atomically resolved SFs (ASFs) of atoms marked with red circles in supercells shown in (a). (d) Trained ML models are tasked to predict ASFs of atoms in heterostructure, Si₂₄Ge₁₅Si₉Ge₆Si₃Ge₃. Predicted ASFs are compared with DFT results.

Forward Learning Model

Figure 1 shows the outline of our forward learning model. We consider Si/Ge superlattices with different periods and compositions as training structures. The structures are chosen such that the trained model could guide design of diverse heterostructures for various applications. We identify a set of features to describe the environments, \( CN(r_p) \), for each atom of a superlattice or heterostructure. Here, \( CN(r_p) \) refers to the local atomic configuration associated with atom \( p \), located at position \( r_p \). We compute the effective band structures (EBS) or spectral functions (SF) for each atom, using DFT. These atomically resolved band structures (AEBS) or spectral functions (ASF) vary significantly with the changes of the atomic environment. We hypothesize that a direct relationship exists between environment features and corresponding ASFs, and train the forward model accordingly.

The trained model accepts environment descriptors for atoms in test superlattices or heterostructures as input, and predicts the associated ASFs. Thus, the components of the forward learning model are as follows: (1) formulation of atomic environment descriptors, (2) generation of ASF training data, (3) implementation of ML models to identify descriptor-ASF relationships, and, (4) comparison of ML predictions for new structures with DFT results.

(1) Formulation of Atomic Environment Descriptors

In Fig. 1(a, b), we show representative configurations of training \( \text{Si}_n\text{Ge}_n \) superlattices. Here, the subscript \( n \) refers to the number of Si and Ge monolayers. We list all training and test structures investigated in this ar-
that represent superlattices grown on substrates [16, 17]. We consider strain-symmetrized and strained supercells are periodically extended in the [001] growth direction. We consider tetragonal supercells that include strained superlattices to train the ML models on band splittings due to strain. We optimize the geometry of all supercells using DFT, see Methods section.

<table>
<thead>
<tr>
<th>Structure Type</th>
<th>Training Structures</th>
<th>Input Features</th>
<th>Input Properties</th>
<th>Test Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain-symmetrized (Si$_2$pGe$_2$q) $(p = 1, 2, \ldots, 13)$ and strained (Si$_2$q−1Ge$_2$q−1)$^2$ $(q = 1, 2, \ldots, 7)$</td>
<td>Atom type: 1 feature/atom</td>
<td>Spectral weights, $A^p(k, E)$: 60 atoms</td>
<td>HS: Si$<em>2$Ge$</em>{15}$Si$_9$Ge$_6$Si$_3$Ge$_3$</td>
<td></td>
</tr>
<tr>
<td>5 applied strains: [0.00%, 0.59%, 1.16%, 1.73%, 2.31%]</td>
<td>Effective bond lengths, $b_x$ &amp; $b_z$: 2 features/atom</td>
<td>$k \times E = 64 \times 64$ per atom $(p)$</td>
<td>Input features: 9 × 60</td>
<td></td>
</tr>
<tr>
<td>Total: 120 structures</td>
<td>Order parameters, $Q_{1,2,3}^1$: 6 features/atom</td>
<td>Total: $64 \times 96 \times 3360$</td>
<td>Output weights: $64 \times 96 \times 60$</td>
<td></td>
</tr>
<tr>
<td>Atomic environments: 6 × 4 × $(\sum_{i=1}^{13} p_i)$ + $\sum_{j=1}^{7} (2q_j - 1)$ = 3360</td>
<td>Total: 9 × 3360 features</td>
<td>Total: $64 \times 96 \times 3360$</td>
<td>Input features: 9 × 56</td>
<td></td>
</tr>
<tr>
<td>(Si$_2$q−1Ge$_2$q−1)$^2$ = Si$<em>2$q−1Ge$</em>{2q-1}$Si$<em>2$q−1Ge$</em>{2q-1}$ for odd $q = 1, 2, \ldots, 7$</td>
<td>(Both strain-symmetrized)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reverse Learning Model: Convolutional Neural Network (CNN) Model

| Strain-symmetrized (Si$_2$pGe$_2$q) $(p = 1, 2, \ldots, 14)$ and strained (Si$_2$q−1Ge$_2$q−1)$^2$ $(q = 1, 2, \ldots, 7)$ | Atom type: 1 feature/atom | HS: Si$_2$Ge$_{15}$Si$_9$Ge$_6$Si$_3$Ge$_3$ |
| 5 applied strains: [0.00%, 0.59%, 1.16%, 1.73%, 2.31%] | Effective bond lengths, $b_x$ & $b_z$: 2 features/atom | 0.00% strained (on Si substrate) |
| Total: 126 structures | Order parameters, $Q_{1,2,3}^1$: 6 features/atom | Input features: 9 × 60 |
| Atomic environments: 6 × 4 × $(\sum_{i=1}^{14} p_i)$ + $\sum_{j=1}^{7} (2q_j - 1)$ = 3696 | Total: $9 \times 3696$ features | Output features: 9 × 60 |

Combined Forward-Reverse Learning Framework: NN, RF & CNN

- Relaxed and 1.73% strained bulk Si
- GaAs ARPES spectra

| CNN Model: Input pixels: 64 × 64; Output features: 9 |
| NN and RF Model: Input features: 9; Output weights: 64 × 96 |
| CNN Model: Input pixels: 64 × 64; Output features: 9 |
| NN and RF Model: Input features: 9; Output weights: 64 × 96 |
- Other bulk Si and Ge systems (see SI Table 1)}

ticle in Table 1. We consider ideal superlattices with sharp interfaces. We consider tetragonal supercells that are periodically extended in the [001] growth direction. We consider strain-symmetrized and strained supercells that represent superlattices grown on substrates [16, 17]. Strain plays a strong role in determining superlattice bands [18, 20] and electronic properties [19, 21, 24].
We choose the following two types of descriptors of atom \( X \) in a training or a test structure: elemental and structural features. The elemental feature is the atom type: 1 for Si and 0 for Ge. We compute the structural features using crystal graphs, constructed from Voronoi tessellations \cite{15}. The structural features are of two types. (1) The effective bond length, \( \bar{r}_i(X) \), of atom \( X \), located at \( \bar{r}_X \), is given by:

\[
\bar{r}_i(X) = \frac{\sum_n \omega_{i,n} A_n \times \| \bar{r}_n - \bar{r}_X \|_2}{\sum_n \omega_{i,n} A_n}.
\]

Here \( i \) refers to Cartesian directions, \( (x, y, z) \), \( \bar{r}_n \) is the location and \( A_n \) is the Voronoi cell face area of the \( n \)th neighbor atom. We only consider \( x \) and \( z \) directions, since structural environments are same along \( x \) and \( y \) directions. To describe the anisotropy of the bonding environment, we define the weights, \( \omega \). \( \omega \) represents the projection of Voronoi cell face area onto the Cartesian planes:

\[
\omega_{i,n} = (\omega_{x,n}, \omega_{y,n}, \omega_{z,n}) = (\cos^2 \phi_n \sin^2 \theta_n, \sin^2 \phi_n \sin^2 \theta_n, \cos^2 \theta_n).
\]

The angles \( \phi_n \) and \( \theta_n \) are the polar and azimuthal angles of the interatomic distance vectors between the atom \( X \) and the \( n \)th neighbor. (2) The order parameters are given by

\[
Q^{\text{order}}_{i}(X) = \sum_{\text{paths}} \prod_{\text{steps}} \frac{\omega_{i,n} A_n \delta_{n,X}}{\sum_n \omega_{i,a} A_a - \omega_{i,b} A_b},
\]

with order = 1, 2, 3. We consider crystal graphs up to a specified order (= 3) neighbor, since higher order graphs do not significantly affect the predictions. \( Q^{\text{order}}_{i}(X) \)’s are contributed by paths in the crystal graphs. As indicated by the Kronecker delta in the above equation, \( \delta_{n,X} \), the graphs have non-zero weights if atom \( X \) and neighbor \( n \) are same atom types. We provide detailed description of the order parameters in our previous publication, see Supplementary Information \cite{15}. We demonstrate here that these two descriptors are strongly correlated with the character of superlattice bands.

(2) Generation of Spectral Function Training Data

We compute the electronic band structures of the superlattices and the heterostructures in the periodic zone representation of the tetragonal supercells. The different supercell sizes result in different number of bands with varied folding. It is challenging to compare the different band structures and interpret to what extent heterostructures differ from their bulk counterparts. Additionally, the supercell bands cannot be directly compared with ARPES spectra. Hence, we unfold the supercell bands into EBS in the extended zone representation of chosen reference cells. We choose reference cells that resemble the primitive cell of FCC bulk silicon. Figure\( ^{11} \)b shows a reference cell corresponding to a Si\textsubscript{4}Ge\textsubscript{4} supercell. The reference cell and the supercell are marked with solid red and dashed blue lines, respectively. The reference cell lattice vectors are determined from the supercell lattice vectors through a matrix transformation, see Methods section. The reference cell lattice vectors vary for different supercells, however, all reference cells include two lattice sites. The similar Brillouin zones (BZs) provide a common reference for comparing the unfolded band structures.

The BZ of an example reference cell is shown in Fig.\( ^{11} \)f in the Methods section. We compute ASF of atoms along the \( X - \Gamma - K - X \) path, indicated by the green plane. Figure\( ^{11} \)c shows representative ASFs, \( A_{i}(k, E) \), for given atom \( i \). Here \( k \) values sample the chosen path and \( E = -3 \) to 3 eV. Each ASF is treated as
FIG. 3. Outline of reverse ML model: (a) Training strain-symmetrized and strained superlattices with varied periods and compositions: (i) Si₁₄Ge₄, (ii) Si₁₄Ge₁₄, (iii) Si₁₁₃Ge₁₁₃ and (iv) Si₂₈Ge₂₈. Descriptors considered: Element types, bond lengths and order parameters. (b) Property value corresponding to descriptors: Example ASFs of inner Si atoms in (i) Si₁₄Ge₁₄ and (ii) Si₂₈Ge₂₈. Trained convolutional neural network (CNN) model predicts atomic environment descriptors for input ASF of atoms in (c) heterostructure, Si₁₄Ge₁₅Si₆Ge₆Si₃Ge₃. Predictions are compared with computed descriptors.

3) Implementation of ML Models

We implement the forward learning approach using neural network (NN) and random forests (RF) models, separately. The models are trained on the relationship between atomic descriptors and respective ASFs of all atoms. We provide the details in Methods section.

4) Comparison of Predictions with DFT Results

We test the ML model on a heterostructure with multiple Si and Ge layers of randomly chosen thicknesses: Si₂₄Ge₁₅Si₆Ge₆Si₃Ge₃ (Fig. 2(a)). Such a configuration is likely to represent fabricated heterostructures with unevenly thick layers. Figure 2(b) shows the total SFs, obtained by summing over all predicted ASFs. The bottom rows show the ASFs from four regions of interest. Instead of comparing SF values and DFT results directly, we obtain the intensity, \( I(E) \), at each \( E \) by summing over SFs at different \( k \) values: \( I(E) = \sum_j A(k_j, E) \). The normalized intensities are obtained by dividing \( I(E) \) by the maximum intensity: \( I_n(E) = I(E) / \text{Max}[I(E)] \). The mean absolute errors (MAEs) are obtained from: \( MAE(I_n, \hat{I}_n) = \sum_E |I_n(E) - \hat{I}_n(E)|/64 \). The predicted total SFs display a mixture of Si-like and Ge-like characters and match the DFT results, resulting in small MAEs (NN: 0.07, RF: 0.08). The inner Si ASFs in Si₂₄ exhibit a prominent bulk Si character. The predictions are similar to the inner Si ASFs of Si₂₄Ge₂₈ superlattice, see last column of Fig. 2(c). This result indicates that inner Si atoms in Si₂₄ region and in Si₂₄Ge₂₈ superlattices are in similar bulk-like environments. The high prediction accuracy (NN: 0.09, RF: 0.07) establishes that ML models can learn and predict the local structure-ASF relationships present in varied configurations. However, the accuracy decreases for narrow regions. The NN model predicts prominent bulk Si character for inner ASFs in Si₉ and does not show Ge-like bands near the Γ point, resulting in a higher MAE: 0.14. The RF prediction shows these bands, resulting in a smaller MAE: 0.10. Larger band splittings and changes in Γ character can be noted for narrow Si regions.

The DFT results (third column) exhibit significantly more band splitting, mixing and discontinuities compared to predicted ASFs. When the supercell includes strongly broken translational order, the unfolded contributions from different regions vary, resulting in broken band structures. Additionally, the reference cells may not...
represent true irreducible cells for a given heterostructure or superlattice. On the other hand, ML results are interpolated from superlattices of similar periods. The higher translational order results in continuous band structures. The discrepancy is more pronounced for narrow Si layers. Prediction accuracy is also low for interface Si ASFs. The larger error (NN: 0.19, RF: 0.16) can be attributed to training data, that include more bulk information than interfaces. Higher training data results in higher accuracy for inner ASFs than interface ASFs. It can be concluded that the error increases with the increase of number of interfaces in the heterostructure. It is important to remember that the training set does not include such multilayer heterostructures. The error is likely to reduce if the model is trained on such structures.

**REVERSE LEARNING MODEL**

We show the outline of the reverse learning model in Fig. 3. We choose similar training structures as the forward learning model (see Table 1). We provide the DFT computed ASFs and the descriptors as input. Each ASF is treated as a greyscale image with resolution of 64 × 64 pixels, with pixel intensities scaled from 0 to 1. We only use ASF data for the energy range between -6 eV to 0 eV as input. We choose this energy range since the available ARPES images scan this energy values. The convolutional neural network (CNN) model predicts the descriptors of atomic environment associated with an input ASF image (Fig. 3(c)). We provide the model details in Methods section. We test the model performance on the heterostructure, Si$_{24}$Ge$_{15}$Si$_9$Ge$_6$Si$_3$Ge$_3$, assumed to be grown on a Si substrate. Figure 4 shows the CNN-predicted descriptors for all atoms. The MAE for each descriptor ($D$) is given by: $MAE(D, \hat{D}) = \frac{1}{n} \sum_{i=1}^{n} ||D_i - \hat{D}_i||/(p \times n)$, where $p$ is the number of atoms in the test structure and $n$ is the number of different Fermi level alignments ($n = 13$). The model remarkably predicts the variation of atomic descriptors across the heterostructure. The atom types for inner atoms are predicted with higher accuracy than interface atoms. The prediction accuracy decreases for narrow layers. The Si and Ge layers have same in-plane lattice constants, resulting in constant values of $b_z$'s. Additionally, same $b_x$ and $b_z$ values in the Si regions are imposed by the Si substrate. However, $b_z$'s in Ge layers (2.73 Å) are higher than those in Si layers (2.58 Å), as expected. The low values of the order parameters near the interfaces indicate that the interface atoms have fewer same species neighbors. As a reference, the order parameters in bulk are equal to 1. The higher order parameter plots clearly show the interface regions for wide layers. However, these plots cannot distinguish narrow layers very well. All panels of Fig. 3 show that the model accuracy is higher for bulk-like inner regions than interfaces or low-symmetry regions.

![FIG. 4. Reverse learning model prediction: (a) Supercell configuration of Si$_{24}$Ge$_{15}$Si$_9$Ge$_6$Si$_3$Ge$_3$ with epitaxial strain due to Si substrate; Predicted (b) atom type, (c) effective bond length and (d-f) spatially resolved order parameters, $Q^i_{\text{order}}$ where $i = (x, z)$ and $\text{order} = 1, 2, 3$, for all atoms. MAEs between predicted (circles) and calculated descriptors (solid lines) are shown next to legends.](image)

The complex character of interface ASFs (Fig. 3(d)) and limited training data result in low accuracy predictions. Additionally, it is important to note that the descriptors are calculated using Voronoi tessellations that is extremely sensitive to atomic environments and introduces uncertainty in the training data and predictions [25, 26].

**Proposed Forward-Reverse Learning framework**

Figure 5 shows the workflow of the combined forward and reverse learning framework. The CNN module reduces the input SF images to a set of descriptors. We pass the CNN-predicted descriptors to the forward learning models and obtain ASFs. We compare intensities computed from input data and the output of the combined model. We test the framework using DFT computed SF images as well as ARPES spectra of bulk systems. The framework is only trained on superlattices and is not aware about the properties of pristine bulk material. Figure 5 shows the DFT-predicted SF images of (a) bulk and (b) substrate-strained Si (s-Si), respectively. The s-Si model is considered to be grown on Si$_{0.7}$Ge$_{0.3}$ alloy
FIG. 5. Proposed forward-reverse learning framework. Input images: (a-b) DFT predicted total SFs for (a) relaxed and (b) substrate-strained (s-Si) bulk Si and (c) ARPES spectra of GaAs thin film, adopted from Ref. 27. CNN extracts atomic environment descriptors from input. The descriptors are provided as input to the forward learning NN and RF models. A larger valence band splitting can be noticed in predicted GaAs SFs. The conduction bands of GaAs are not likely to be accurately represented, although contributions from conduction bands near Γ-point can be noticed. Fig. 5(f) shows the comparison between normalized intensities of the input and the output SF images. We find close agreement for both bulk Si and GaAs models. This test illustrates the remarkable ability of our framework to predict electronic bands of new material from purely structural information.

**CONCLUSIONS**

We propose a first-principles modeling framework that establishes a direct relationship between atomic struc-
tural features and electronic bands of semiconductor heterostructures. The framework combines a forward and a reverse ML model. The forward model shows that atomic environments, such as neighbor bond lengths, local network of atoms etc, tune heterostructure bands. The different atomic environments determine the Bloch character of heterostructure bands. For example, an inner atom maintains bulk character, however, an interface atom displays complex character such as band splittings, avoided crossings and changes of Γ-characters. The reverse model predicts the atomic environment that is associated with an input band structure image. It is remarkable that the model can predict atom types purely from the valence band character exhibited by an experimental image. We establish that the relationship between structural features and band character is deeply fundamental in nature. We illustrate that this relationship can be leveraged to predict properties of GaAs, even though the framework is trained on Si/Ge systems. The combined model illustrates that atomic environments can be designed to manipulate heterostructure bands and achieve desired electrical, magnetic and optical properties. Our framework offers a physics-informed approach to create layered materials for new phenomena and device possibilities for diverse technologies.

**DATA AVAILABILITY**

The data supporting the findings of this study are available within the main article and the Supplementary Information document. In addition, we uploaded an example data set in the CUantamLab public GitHub repository [28]. The example data set includes geometry optimized 49 Si/Ge superlattice configurations, both strain-symmetrized and with external strain.

**CODE AVAILABILITY**

We made the Python scripts available, that can be used for extracting descriptors from the example data set [28].

**ACKNOWLEDGEMENTS**

We gratefully acknowledge funding from the Defense Advanced Research Projects Agency (Defense Sciences Office) [Agreement No.: HR0011-16-2-0043]. We acknowledge funding from the National Science Foundation Harnessing the Data Revolution NSF-HDR-OAC-1940231. This work utilized the Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.

**AUTHOR CONTRIBUTIONS**

A.K.P contributed to the acquisition and the analysis of data and the creation of new scripts used in the study. S.N. contributed to the conception and the design of the work, the interpretation of data, drafting and revision of the article.

**COMPETING INTERESTS**

The authors declare no competing interests.

**METHODS**

Training and Test Structures for All ML Models

Training Structures: We consider ideal superlattices with both even and odd number of monolayers, referred to as Si$_{2p}$Ge$_{2p}$ ($p = 1, 2, \ldots, 14$) and Si$_{2q-1}$Ge$_{2q-1}$Si$_{2q-1}$Ge$_{2q-1}$ ($q = 1, 2, \ldots, 7$), respectively. The total number of atoms in model supercells (SCs) of these superlattices are even multiples of four. The SCs of Si$_{2p}$Ge$_{2p}$ configurations include $4p$ atoms and Si$_{2q-1}$Ge$_{2q-1}$Si$_{2q-1}$Ge$_{2q-1}$ configurations have $4(2q - 1)$ atoms. We construct the SCs using a template that includes four atomic positions, as we discuss below. For the superlattices with odd number of monolayers, we double the SC sizes to account for the structure periodicity. We consider both strain symmetrized superlattices and superlattices with applied strains: 0.00%, 0.59%, 1.16%, 1.73%, and 2.31%. The strains are measured relative to bulk Si lattice constant: $((a' - a_{Si})/a_{Si}) \times 100$, where $a_{Si} = 5.47$ Å. The applied in-plane strain values correspond to alloy growth substrates, Si$_{1-x}$Ge$_x$, with varied Ge concentrations: $x = 0, 0.1, 0.2, 0.3, 0.4$.

Test Structures for Forward ML Model: Strain-symmetrized (i) Si$_{28}$Ge$_{28}$ superlattice and (ii) Si$_{24}$Ge$_{15}$Si$_9$Ge$_9$Si$_3$Ge$_3$ heterostructure.

Test Structure for Reverse Learning CNN Model: Si$_{24}$Ge$_{15}$Si$_9$Ge$_9$Si$_3$Ge$_3$ heterostructure assumed to be grown on silicon substrate.

Supercells: We generate the model SCs for all training and test superlattices and heterostructures, using a four-atom bulk Si tetragonal cell template (Si$_4$). The template is derived from a bulk Si cubic conventional cell. The template and the cubic cell are shown with blue dashed lines and black solid lines in Fig. 6(a, b), respectively. The volume of the template is half of the volume of the cubic cell. The lattice parameters of the template are given by: $a' = b' = 2.73$ Å and $c = 5.47$ Å.
following geometry optimization. \( c = 5.47 \text{Å} \) is same as lattice constant of bulk Si cubic cell. Our values agree with previous DFT results \[29\]. Although, it is known that the DFT-predicted lattice constants of bulk Si are \( \sim 1\% \) higher than experimental values \[30\]. The basis vectors of the SC template are given by \( a'[110], b'[110], \) and \( c[001] \). The template is periodically extended in the [001] growth direction. The template can be used to span a bulk Si system with cubic symmetry, e.g., [001] grown superlattices, by replicating in the [110], [110] and [001] directions. This template allows us to investigate a large variety of superlattices and heterostructures, while keeping the computational expense at a minimum.

The template includes four atomic positions, see Fig. 6. The black, red, blue and green colors, represent atomic positions residing in planes that are \( a/4 \) apart along [001]. Fig. 6(c, d) shows the atomic positions as viewed along the [001] direction. To create the superlattice models we insert Si and Ge atoms in the template atomic positions. For example, we obtain a \( \text{Si}_2\text{Ge}_2 \) superlattice model by inserting Si atoms in the blue and green positions, and Ge atoms in the black and red positions of the template, respectively (Fig. 6(e)). The red ‘+’ symbols show the two Si and the two Ge atoms, while the replicas are unmarked. We obtain strain-symmetrized or strained configurations by performing geometry optimization of the supercell models. To create superlattice models with longer periods, we start with [001] periodically replicated models of SC template. We insert Si and Ge atoms in the atomic positions of the replicated template. That way, the resulting SC includes the desired number of Si and Ge monolayers. A representative eight-atom \( \text{Si}_2\text{Ge}_2 \) superlattice model is shown in Fig. 1(b). Table 3 lists the lattice parameters of all geometry optimized supercells investigated in this article.

**Test Structures for Combined Model:** Strain-symmetrized and strained bulk Si, modeled with Si4 SC. The strained model is considered to be grown on \( \text{Si}_{1.7}\text{Ge}_{0.3} \) substrate, with 1.73% in-plane strain. Test ARCES Image for Combined Model: ARCES spectra adopted from Fig. 2 of Ref. [27]. We combine the band dispersions along \( \Gamma - X \) (Fig. 2(b)) and \( \Gamma - K - X \) paths, obtained with s-(Fig. 2(c)) and p-polarization (Fig. 2(d)) into a single image. We interpolate the experimental images from the resolution provided in Ref. [27] to 64 \times 64 pixels in size and for energy range from -6 eV to 0 eV.

**FIG. 6. Generation of training supercells (SC) and selection of reference cells (RC):** (a, b) Tetragonal SC template (blue dashed lines) generated from a bulk Si conventional cell (black solid lines). \( a \) refers to bulk Si lattice constant. SC template includes four atomic positions, with one position per monolayer stacked along [001] direction: black, red, blue and green. Representative two-atom RC chosen for unfolding (solid red lines). (c, d) Atomic positions in Si conventional cell, SC template and RC viewed along the [001] direction. (e) \( \text{Si}_2\text{Ge}_2 \) superlattice SC, with two Si and two Ge atoms marked with red. Other atoms are replicas. Pairs of marked SC atoms (‘1’ and ‘2’) are mapped to the two corresponding RC atomic positions, respectively. (f) SC (blue) and RC (red) Brillouin zone (BZ). Black dashed line represents projection of SC BZ onto the [001] plane passing through \( \Gamma \) point. Symmetry points and paths in the green plane are used to obtain effective band structures or spectral functions.

**DFT Computation Details**

We optimize the lattice constants and the atomic positions of training and test SC models with conjugate gradient algorithm \[31\]. We sample the SC BZ with \( 11 \times 11 \times 11 \) k-point mesh, generated by the Monkhorst-Pack scheme \[32\]. Although \( c \) is generally larger than \( a' \) and \( b' \), we use \( 11 \times 11 \times 11 \) k-point mesh to accommodate for ample sampling along [001]. Such sampling is particularly necessary for complex heterostructures with irregularly thick Si or Ge layers. To simulate SC under applied strain, we assign \( a' \) and \( b' \) to be equal to the substrate lattice constants and relax the cell shape in the cross-plane [001] direction. The DFT calculations reported in this article, are performed using...
### TABLE 3. Lattice Parameters of Supercells, Å

<table>
<thead>
<tr>
<th>Structures</th>
<th>Strain- Symmetrized</th>
<th>Grown on Si_{1-x}Ge_{x} Substrate</th>
<th>Ge concentration, x =</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si_{4}</td>
<td>2.73</td>
<td>5.47</td>
<td>5.04</td>
<td>5.35</td>
<td>5.29</td>
<td>5.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ge_{4}</td>
<td>2.89</td>
<td>5.79</td>
<td>5.48</td>
<td>6.03</td>
<td>6.26</td>
<td>6.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Si_{1}Ge_{3})_{2}</td>
<td>2.81</td>
<td>5.60</td>
<td>5.36</td>
<td>5.80</td>
<td>5.73</td>
<td>5.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{2}Ge_{1}</td>
<td>2.81</td>
<td>5.60</td>
<td>5.36</td>
<td>5.79</td>
<td>5.73</td>
<td>5.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Si_{1}Ge_{3})_{2}</td>
<td>2.81</td>
<td>16.86</td>
<td>17.79</td>
<td>17.39</td>
<td>17.19</td>
<td>16.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{4}Ge_{4}</td>
<td>2.81</td>
<td>11.25</td>
<td>11.87</td>
<td>11.73</td>
<td>11.60</td>
<td>11.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Si_{1}Ge_{3})_{2}</td>
<td>2.81</td>
<td>28.12</td>
<td>29.66</td>
<td>28.98</td>
<td>28.66</td>
<td>28.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{6}Ge_{5}</td>
<td>2.81</td>
<td>16.88</td>
<td>17.82</td>
<td>17.61</td>
<td>17.41</td>
<td>17.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Si_{1}Ge_{3})_{2}</td>
<td>2.81</td>
<td>23.09</td>
<td>23.73</td>
<td>23.48</td>
<td>23.21</td>
<td>22.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{6}Ge_{5}</td>
<td>2.81</td>
<td>50.63</td>
<td>53.44</td>
<td>52.83</td>
<td>52.22</td>
<td>51.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{10}Ge_{10}</td>
<td>2.81</td>
<td>28.15</td>
<td>29.72</td>
<td>29.38</td>
<td>29.04</td>
<td>28.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Si_{1}Ge_{3})_{2}</td>
<td>2.81</td>
<td>61.90</td>
<td>65.37</td>
<td>64.62</td>
<td>63.89</td>
<td>63.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{2}Ge_{12}</td>
<td>2.81</td>
<td>33.75</td>
<td>35.65</td>
<td>35.24</td>
<td>34.84</td>
<td>34.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Si_{1}Ge_{3})_{2}</td>
<td>2.81</td>
<td>73.10</td>
<td>77.14</td>
<td>76.26</td>
<td>75.39</td>
<td>74.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{14}Ge_{14}</td>
<td>2.81</td>
<td>39.39</td>
<td>41.55</td>
<td>41.07</td>
<td>40.61</td>
<td>40.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{16}Ge_{16}</td>
<td>2.81</td>
<td>45.01</td>
<td>47.51</td>
<td>46.96</td>
<td>46.43</td>
<td>45.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{18}Ge_{18}</td>
<td>2.81</td>
<td>65.05</td>
<td>69.47</td>
<td>68.65</td>
<td>67.79</td>
<td>66.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{22}Ge_{22}</td>
<td>2.81</td>
<td>56.29</td>
<td>59.46</td>
<td>58.78</td>
<td>58.11</td>
<td>57.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{24}Ge_{24}</td>
<td>2.81</td>
<td>68.88</td>
<td>75.27</td>
<td>74.52</td>
<td>73.79</td>
<td>73.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{26}Ge_{26}</td>
<td>2.81</td>
<td>75.35</td>
<td>81.26</td>
<td>80.44</td>
<td>79.65</td>
<td>78.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si_{28}Ge_{28}</td>
<td>2.81</td>
<td>78.79</td>
<td>83.20</td>
<td>82.45</td>
<td>81.31</td>
<td>80.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS</td>
<td>2.79</td>
<td>83.91</td>
<td>87.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For odd \(q = 1, 3, 5, 7, 9, 11, 13\)

---

for even \(q = 2, 4, 6, 8, 10, 12, 14\)

the OpenMX code which is based on norm-conserving pseudopotentials generated with multiple reference energies. We use the Perdew-Burke-Ernzerhof exchange-correlation formulation of the generalized gradient approximation. Self-consistent field (SCF) calculations are performed during the geometry optimization with energy convergence threshold set to \(10^{-9}\) Hartree. The SCs are optimized until the maximum force on an atom became less than \(10^{-4}\) Hartree Bohr\(^{-1}\). We use a regular mesh of 200 Ryd in real space for the numerical integrations and solution of Poisson equation. We do not include spin-orbit interaction in our calculations, since strain induced band splittings were shown to be larger than the spin-orbit splittings. For the silicon and germanium atoms, 2, 2, and 1 optimized radial functions were allocated for the s-, p- and d-orbitals, respectively, as denoted by s2p2d1. One-particle wave functions are expressed by the linear combination of pseudo-atomic orbital (PAO) basis functions centered on atomic site. A cutoff radius of 7.0 Bohr was used for all the basis functions. Following relaxation, we perform non self-consistent field (NSCF) calculations using the linear combinations of atomic orbitals (LCAO) pseudopotential method. We obtain the eigenstates, \([KJ]\) and energy, \(\epsilon_{KJ}\), for the range from -10 eV to 10 eV. Here, \([KJ]\) represents a Bloch state with crystal momentum \(K\) and band index \(J\), \(\epsilon_{KJ}\) is the corresponding eigenvalue. We use a \(7 \times 7 \times 7\) k-point mesh generated according to the Monkhorst-Pack method to sample the supercell BZ. Such k-point mesh has been used in DFT studies for calculation of electronic structure of two-atom Si lattice.

### Supercells and Reference Cells

We unfold SC electronic band structures to the BZ of chosen reference cells (RCs) via SFs. The embedding cubic conventional cell and the SC template are shown with black solid lines and blue dashed lines, respectively. The volume of the SC is 1/4 the volume of the conventional cell. The RC resembles the two-atom primitive cell of FCC Si lattice. If no symmetry breaker is present in the SC, the RC will be identical to a primitive cell. However, the RC representation is defined purely in a mathematical sense and is valid even in the presence of symmetry-breaking interfaces. The RC basis vectors are determined from SC basis vectors. In general, the SC basis vectors are given by:

\[
\vec{B} = \left( \begin{array}{c} a' \\ b' \\ c' \end{array} \right),
\]

where \(a' = b'\) and \(c'\) are the SC lattice parameters. The SC and the RC basis vectors are related: \(\vec{b} = M \vec{B}\), via a transformation matrix, \(M\), given by:

\[
M = \left( \begin{array}{ccc} -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{array} \right),
\]

where \(N\) represents number of templates stacked along [001] direction in the supercell. \(N\) is equal to the total number of atoms in the SC divided by 4. For example, \(N = 1\) for Si_{2}Ge_{2}, and \(N = 2\) for Si_{2}Ge_{4} and so on. Thus, the RC basis vectors are given by:

\[
\vec{b} = \left( \begin{array}{c} \frac{1}{N} \epsilon' - a' \\ \frac{1}{N} \epsilon' + a' \\ \frac{1}{N} c' \end{array} \right),
\]

As shown in Table 3, \(a', b'\) and \(c'\) vary for different superlattices or heterostructures, resulting in different basis...
vectors for the corresponding two-atom RCs. Fig. 6(e) shows the RC of a \( \text{Si}_2\text{Ge}_2 \) SC and Fig. 6(f) shows the corresponding BZ. We compute AEBS or ASF of atoms along the path indicated by the green plane, and generate training data. The choice of RCs is not unique. Past tight binding studies extensively discussed the identification of appropriate RCs and the advantage of unfolding SC band structures for predicting electronic properties \cite{7}. We exploit the insights to develop our ML framework. However, these studies emphasize that care must be taken when the symmetry of the SCs is different from the RCs. We discuss the consequences in this article.

**Spectral Weights and Spectral Functions**

The steps for computing SFs are as follows:

(i) Geometry optimization of chosen SCs;
(ii) Calculating SC eigenstates for a \( k \)-point mesh;
(iii) Identifying RC and a set of wave vectors \( \{ k_i \} \) along a chosen path in the respective BZ;
(iv) Computing SFs for each atom along the path;
(v) Repeating steps (i)–(iv) for a different SC;

Below we discuss the approach used for step (iv).

In the LCAO method, a Bloch state \( |KJ \rangle \) is expanded in the form of a linear combination of atomic basis functions, \( |RN\rangle \), as

\[
|KJ \rangle = \sum_{N} C_{N}^{KJ} |KN\rangle, \tag{7}
\]

while,

\[
|KN\rangle = \frac{1}{\sqrt{L}} \sum_{R} e^{iK \cdot R} |RN\rangle. \tag{8}
\]

Here, \( C_{N}^{KJ} \) are the LCAO coefficients. The atomic basis functions, \( |RN\rangle \), are placed in every unit cell and specified with a translational lattice vector \( R \). \( N \) represents a symbolic orbital index, that consists of the atomic position relative to \( R \), a multiplicity index for radial functions, an angular momentum number, and a magnetic quantum number. \( L \) is the number of unit cells included in the Born-von Karman boundary condition. We unfold the SC band structures to RC BZs via SFs, \( A(E) \), following the method proposed in Ref. \cite{9}. It can be shown that the SF expressions in the SC and the RC representations are related via

\[
A_{k_j,k_j}(E) = \sum_{m,n} S_{mn}^{-1}(k) |km\rangle A_{KJ,KJ}(E) |KJ\rangle |km\rangle. \tag{9}
\]

Here, \( |kj\rangle \) is a Bloch state and \( m \) and \( n \) represent symbolic orbital indices in the RC representations. \( S_{mn}(k) \) are the overlap matrix elements. The spectral function, \( A_{k_j,k_j}(E) \), can be represented as a linear combination of unfolded spectral weights, \( W_{KJ}^{k} \),

\[
A_{k_j,k_j}(E) = \sum_{K} W_{KJ}^{k} A_{KJ,KJ}(E) \tag{10}
\]

with

\[
W_{KJ}^{k} = \frac{L}{l} \sum_{G} \delta_{k-G,K} \times \sum_{MNr} e^{iK \cdot (r-r'(M))} C_{M}^{KJ} C_{N}^{KJ'} S_{0N,rM}(M). \tag{11}
\]

Here, \( L \) and \( l \) are the number of unit cells in SC and RC, respectively. \( r'(M) \) and \( m(M) \) refer to lattice vectors and orbital indices of the RC atomic basis, respectively, that correspond to the SC symbolic orbital index \( M \). The delta function, \( \delta_{k-G,K} \) signifies that \( W \) at \( k \) is contributed only by the folded Bloch state with \( k - G = K \), where \( G \) is a reciprocal lattice vector. The spectral weights are determined by the phase factor \( e^{iK \cdot (r-r'(M))} \), LCAO coefficients, \( C_{M}^{KJ} \) or \( N \), and overlap matrix elements, \( S_{0N,rM}(M) \). In the LCAO method, we allocate same basis functions for each atomic species in the SCs with varied periods and compositions. However, same AOs contribute differently to the electronic bands of different SCs due to structural features. The influence of atomic structures is recorded in the LCAO coefficients and the overlap integrals between basis functions. It is important to note that the spectral weights are calculated in the SC representation without relying on any RC details. Thus, unfolding is performed purely in a mathematical sense and is valid as long as a RC can be defined.

Additionally, we compute the orbitally resolved SFs that allow us to analyze how spectral weights vary based on the atomic environment. To obtain ASFs, we express the orbitally resolved SFs as:

\[
A_{k_j,k_j}(E) = \sum_{K} A_{KJ,KJ}(E) W_{KJM}^{k}, \tag{12}
\]

where, the orbitally resolved spectral weights, \( W_{KJM}^{k} \), are given by

\[
W_{KJM}^{k} = \frac{L}{T} \sum_{G} \delta_{k-G,K} C_{M}^{KJ} \times \sum_{NR} e^{iK \cdot (r-r'(M))} S_{0N,rM}(M). \tag{13}
\]

and are obtained by rearranging Eq. 11. Equation 10 can then be written as:

\[
A_{k_j,k_j}(E) = \sum_{M} A_{k_j,k_j}^{M}(E) = \sum_{M} A_{KJ,KJ}(E) \sum_{K} W_{KJM}^{k}. \tag{14}
\]

The values of index \( M \) depend on the SC size and the number of basis functions for each atom. As an example, for a supercell with \( n \) atoms and \( m \) basis function
per atom, $M$ ranges from 1 to $n \times m$. By decomposing the spectral weights and the SFs, it is possible to analyze the contribution from different localized basis functions to bands. We obtain the atomically resolved SFs from the orbitally resolved SFs. Considering that there are $m$ basis functions assigned to the $p$th atom in the supercell: $\{p_1, ..., p_m\} \subset M$, we add $p_m$ orbitally resolved spectral weights, $W^k_{KJp_m}$, to obtain the ASFs for each atom. Thus, total SF can then be written as:

$$A_{kj,kj}(E) = \sum_K A_{KJ,KJ}(E) \sum_p \sum_{m \subset M} W^k_{KJp_m}, \quad (15)$$

or, in terms of the atomically resolved SFs, $A^p_{kj,kj}(E)$,

$$A_{kj,kj}(E) = \sum_p A^p_{kj,kj}(E) = \sum_p \sum_{m \subset M} A^p_{kj,kj}(E). \quad (16)$$

Here, $A_{KJ,KJ}(E)$ is a delta function, $\delta(E - \epsilon_{KJ})$.

We obtain the SC eigenstates ($\epsilon_{KJ}$) for the energy range from $-10$ eV to 10 eV from the NSCF calculations. We compute spectral weights by unfolding the SC eigenstates, $|KJ\rangle$, on RC Bloch states $|kj\rangle$. We choose a set of 100 wave vectors $\{k_i\}$ along the $X - \Gamma - K - X$ path of the RC Brillouin zone (as depicted in Fig. 6(f)). We do not explicitly keep track of the band indices and drop the subscript $j$. The spectral weights are then convoluted with the above delta function to obtain $A^p_{k,k}(E)$ or $A^p(k,E)$. We model the delta function $\delta(E - \epsilon_{KJ})$ representing $A_{KJ,KJ}(E)$ with an exponential function with width 0.02 eV. For the forward ML model, we use a 450 point sampling for the delta function between $-6$ eV $\leq E \leq 3$ eV. We then interpolate over $300 \times 450$ values of $A^p(k,E)$ and obtain $64 \times 96$ ASF values. The ASFs, $A^p(k,E)$, are defined over $\{k\} \rightarrow X - \Gamma - K - X$ and $-6$ eV $\leq E \leq 3$ eV. The total SFs, $A(k,E)$, are obtained by summing over $A^p(k,E)$ for all atoms in the superlattice or heterostructure. $A(k,E)’s$ are also defined over the same $k$ and $E$ range, with a $64 \times 96$ sampling. In the last part of this study, while comparing with the ARPES spectra, we use a 300 point sampling for $-6$ eV $\leq E \leq 0$ eV, using 0.02 eV wide delta-function. We choose 300 $\{k_i\}$ vectors along $X - \Gamma - K - X$ to keep a square grid for $(k,E)$.  

### ML Model Implementations

#### Forward Learning Approach

We implement the forward learning approach using NN and RF model, separately.

**NN model:** Table 4 shows the layers, the number of nodes in each layer and the activation functions of the NN model. The model has four fully-connected (dense) layers represented by the four rows. The input layer with 32 nodes is followed by two hidden layers with 64 and 128 nodes, respectively. Rectified Linear Unit (ReLU) activation functions are implemented for the input and the hidden layers. The number of input parameters is equal to the number of features considered, nine in our case as shown in Table 4. The output layer has 6144 nodes and linear activation function. The number of nodes in the output layer corresponds to the $64 \times 96$ interpolated $A^p(k,E)$ values for the respective superlattice or heterostructure. The same NN model architecture can be employed even when the size of input or output data is changed.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Nodes</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense (32)</td>
<td>input dimension</td>
<td>9 features $f_a=$ReLU</td>
</tr>
<tr>
<td>Dense (64)</td>
<td>$f_a=$ReLU</td>
<td></td>
</tr>
<tr>
<td>Dense (128)</td>
<td>$f_a=$ReLU</td>
<td></td>
</tr>
<tr>
<td>Dense (6144)</td>
<td>$f_a=$Linear</td>
<td></td>
</tr>
</tbody>
</table>

$64 \times 96 A^p(k,E)$

We allocate 25% of training data for model validation. We sample random batches of size 32 sequentially from the training set (e.g., 32/3360) at each epoch during training. The last batch will be of size less than 32 if the remainder is not zero. We update the weights iteratively for 5000 epochs till MAE between predicted and validation ASF:

$$MAE(A^p, \hat{A}^p) = \frac{\sum_{k,E}|A^p(k,E) - \hat{A}^p(k,E)|}{64 \times 96}, \quad (17)$$

reaches a minimum. We employ the ADAM stochastic optimization method for gradient descent [10] with learning rate of 0.0005 to minimize the loss function, MAE in this case. We changed the number of epochs for the NN model from 4000 to 5000 and the error only changed from 0.02187 to 0.02183 (see SI Fig. 6). We are reporting the results obtained with epoch 4840 and MAE 0.02183. The high-level NNs are implemented using the Python Keras library [11]. The optimized weights are used to predict ASF values for test structures.

#### RF model

The RF model assembles results of several decision trees. Each tree is built from a random selection of training data that include both structural features and ASF of training superlattices. We use feature based decision rules to partition the training data into subsets. As an example, decision rule could be based on order parameter $Q^1$ in the range 0.5 - 0.6, representing different atomic environments (see Fig. 4). Training data with $Q^1$ in the range 0.5 - 0.6 could form a subset. The branches of the trees are constituted from the decision rules that identify features that minimize the intr subset variation of ASF. ASF that maximizes fitting over the subset data are assigned as leaves of the tree. The tree generation process is then repeated for other random subsets of training data. We average over the predicted $A^p(k,E)$ from all the trees to obtain the final predictions. We implement the RF module available in
the scikit-learn Python package [12]. We use 100 regression trees per ensemble and default values for all other parameters recommended for the package. We use input and output same as the NN model, as listed in Table [4].

Reverse Learning Approach: We implement the reverse learning approach using CNN model. CNN model is extensively used for feature extraction in digital images and is able to assemble complex patterns from small training data [43]. We employ CNN to identify patterns in the training images of ASFs. The patterns of ASF images represent effects of translational symmetry breaking on the electronic bands of heterostructures. We use the CNN model to learn the relationships between these patterns and the descriptors. The details of the model are shown in Table 5. The model includes one coordinate channel layer (CoordinateChannel2D) [44] and three convolution layers (Conv2D), with 4, 64, 64, and 64 filters each. Each convolution layers is followed by batch normalization and a max pooling layer (MaxPooling2D).

The dimension of the tensor at the input layer is (64, 64, 4), where the first two are the pixel dimensions of the image, and the third is the number of filters. The 64 \times 64 pixels of \( A^P(k, E) \) images are provided as input. Two fully connected layers (Dense) with 128 nodes with SoftPlus and Softmax activation function respectively are followed by an output layer (Dense) with linear activation function. The set of descriptors outlined above are passed through the output layer. We consider 3 different sets of descriptors in the model, namely, continuous unbounded (effective bond length), continuous bounded (order parameters), and discrete (atom type). We find that the linear activation function in the output layer is suitable for all the above predictors.

We consider 13 different Fermi level alignments. We shift the mid-gap DFT zero energy level of the SFs of each configurations by a value \( dE \) in the range from -0.5 eV to +0.5 eV with a step of 1/13 eV. The alignments serve a role of electron doping level in ARPES experiments. This training helps the model to predict the descriptors of ARPES images with different Fermi level alignments. We allocate 30% of training data for model validation. We sample random batches of size 64 sequentially from the training set at each epoch during training. We employ the ADAM stochastic optimization method for gradient descent [40] with learning rate of 0.0005 to minimize the loss function. We update the weights iteratively for 2000 epochs till the MAE between predicted and validation descriptors reaches a minimum.

Combined Forward-Reverse Learning Model: Finally, we merge the forward and reverse trained ML models shown in Tables 5 and 4, respectively, into one combined model.

Table 5. Reverse CNN block

<table>
<thead>
<tr>
<th>Layer</th>
<th>Shape</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoordinateChannel2D</td>
<td>(64, 64, 4)</td>
<td></td>
</tr>
<tr>
<td>Conv2D</td>
<td>(32, 32, 64)</td>
<td>k=3, s=2, ( f_a: ) Softmax</td>
</tr>
<tr>
<td>BatchNormalization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxPooling2D</td>
<td>(16, 16, 64)</td>
<td>k=2, s=2, ( f_a: ) Softmax</td>
</tr>
<tr>
<td>Conv2D</td>
<td>(16, 16, 64)</td>
<td>k=3, s=1, ( f_a: ) Softmax</td>
</tr>
<tr>
<td>BatchNormalization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxPooling2D</td>
<td>(8, 8, 64)</td>
<td>k=2, s=2, ( f_a: ) Softmax</td>
</tr>
<tr>
<td>Conv2D</td>
<td>(8, 8, 64)</td>
<td>k=3, s=1, ( f_a: ) Softmax</td>
</tr>
<tr>
<td>BatchNormalization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxPooling2D</td>
<td>(4, 4, 64)</td>
<td>k=2, s=2, ( f_a: ) Softmax</td>
</tr>
<tr>
<td>Flatten</td>
<td>(1024)</td>
<td></td>
</tr>
<tr>
<td>Dense</td>
<td>(128)</td>
<td>( f_a: ) Softplus</td>
</tr>
<tr>
<td>Dense</td>
<td>(128)</td>
<td>( f_a: ) Softmax</td>
</tr>
<tr>
<td>Dense</td>
<td>(n = 9 features)</td>
<td>( f_a: ) Linear</td>
</tr>
</tbody>
</table>

- k: kernel size, s: stride size, \( f_a: \) activation function

---

* sanghamitra.neogi@colorado.edu


[13] Vasudevan, R., Pilania, G. & Balachandran, P. V. Ma...
machine learning for materials design and discovery (2021).


[28] Cuquant lab - github page. [https://github.com/](https://github.com/)


[30] Semiconductor. V. *General Properties of Si, Ge, SiGe, SiO2 and Si3N4* (2002).


Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SMSpectralFunctionsviaMLmodel.pdf