Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 1365

InterMat: accelerating band offset prediction in semiconductor interfaces with DFT and deep learning†

Kamal Choudhary * and Kevin F. Garrity

We introduce a computational framework (InterMat) to predict band offsets of semiconductor interfaces using density functional theory (DFT) and graph neural networks (GNN). As a first step, we benchmark OptB88vdW generalized gradient approximation (GGA) work functions and electron affinities for surfaces against experimental data with accuracies of 0.29 eV and 0.39 eV, respectively. Similarly, we evaluate band offset values using independent unit (IU) and alternate slab junction (ASJ) models leading to accuracies of 0.45 eV and 0.22 eV, respectively. We use bulk band structure calculations with the TBmBJ meta-GGA functional to correct for band gap underestimation when predicting conduction band properties. During ASJ structure generation, we use Zur's algorithm along with a unified GNN force-field to tackle the conformation challenges of interface design. At present, we have 607 surface work functions calculated with DFT, from which we can compute 183 921 IU band offsets as well as 593 directly calculated ASJ band offsets. Finally, as the space of all possible heterojunctions is too large to simulate with DFT, we develop generalized GNN models to quickly predict bulk band edges with an accuracy of 0.26 eV. We show how these models can be used to predict relevant quantities including ionization potentials, electron affinities, and IU-based band offsets. We establish simple rules using the above models to pre-screen potential semiconductor devices from a vast pool of nearly 1.4 trillion candidate interfaces. InterMat is available at website: https://github.com/usnistgov/intermat.

Received 26th January 2024
Accepted 21st May 2024

DOI: 10.1039/d4dd00031e

rsc.li/digitaldiscovery

Introduction

Interfaces are critical for a variety of technological applications including semiconductor transistors and diodes, solid-state lighting devices, solar-cells, data-storage and battery applications.[1–8] In particular, the continued scaling of semiconductor devices towards the atomic limit[9] makes interface properties even more important and a focus area of recent investments in research and development including the Creating Helpful Incentives to Produce Semiconductors (CHIPS) act.[10] While interfaces are ubiquitous, predicting even basic interface properties from bulk data or chemical models remains challenging. There have been numerous scientific efforts to model interfaces with a variety of techniques including density functional theory (DFT),[11–20] force-field,[21–24] tight-binding[25–28] and machine learning techniques.[19,29–31] However, to the best of our knowledge, there is no systematic investigation of interfaces for a large class of structural variety and chemical compositions. Most of the previous efforts focus on a limited number of interfaces, and hence there is a need for a dedicated infrastructure for data-driven interface materials design.

Some of the key quantities for determining interface properties are: equilibrium geometries, energetics, work functions, ionization potentials, electron affinities, band offsets, carrier effective masses, mobilities, and thermal conductivities. Calculations of band offsets and band-alignment at semiconductor heterojunctions are of special interest for device design. Semiconductor device transport and performance depend critically on valence band offsets (ΔE_v) and conduction band offsets (ΔE_c), as well as interfacial roughness and defects.[32–34] Based on the band-alignment, heterostructures can be categorized into three classes: (i) type-I (straddling gap), (ii) type-II (staggered gap), and (iii) type-III (broken gap). The type-I heterostructures are used for transistors, lasers and light-emitting diode (LED) applications, type-II are used for photo-absorbers and photocatalysts, and type-III are used for tunneling field effect transistors.

Experimentally, band offsets can be measured using optical spectroscopy, X-ray photoelectron spectroscopy (XPS), ultraviolet photoelectron spectroscopy (UPS), and electrical measurements.[5] However, these experiments can be quite time and resource consuming. Additionally, the variability across multiple reported measurements can be reasonably high. For example, the reported AlN/GaN interface ΔE_v varies from 0.57 eV

Material Measurement Laboratory, National Institute of Standards and Technology, Maryland, 20899, USA. E-mail: kamal.choudhary@nist.gov

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00031e

to 1.36 eV with reported uncertainties of up to 0.24 eV.[34] In this respect, the computation of band offsets can serve as a complementary tool to experimental analysis. Nevertheless, the calculation of band offsets is rather challenging[35] and has been an area of research for about a century.[36–38] Density functional theory (DFT) calculations are one of the most widely used techniques for predicting band offsets, as they can describe the electronic and atomic structures at the interface in a self-consistent manner. There are two main approaches to predicting band offsets using DFT. The first is to directly simulate the interface using either an alternating slab-junction (ASJ)/superlattice or surface terminated junction (STJ)/slab vacuum geometry, either of which requires a computationally expensive calculation for each pair of materials. Alternatively, the independent unit (IU)/electron affinity/Anderson's model[16,39] requires only independent surface calculations of each material, greatly reducing computational cost but ignoring specific interface effects. ASJ models were shown to be most accurate in ref. 16, but IU models are surprisingly competitive.

Importantly, the generation of an atomistic interface geometry is a challenging task due to the high number of possible conformations and configurations. There are several important factors determining an interface such as: the selection of the lattice alignment, the relative orientation/displacement between surfaces, the separation distance, point/line defects, and the presence of interfacial charge transfer. Several previous tools have attempted to address this challenge, including MPInterfaces,[18] TribChem[40] and QuantumATK[41] packages.

Moreover, DFT calculations of interfaces require initial pre-relaxed bulk structures which in this work are obtained from the Joint Automated Repository for Various Integrated Simulations (JARVIS)-DFT[42,43] database containing nearly 80 000 bulk 3D and 1100 2D materials. The JARVIS-DFT originated about 5 years ago and contains millions of properties materials and has carefully converged atomic structures with tight convergence parameters, various exchange–correlation functionals such as OptB88vdW,[44] TBmBJ,[45] R2SCAN[46] and HSE06.[47] JARVIS-DFT contains metallic, semiconducting, insulator, superconductor, high-strength, topological, solar, thermoelectric, piezoelectric, dielectric, two-dimensional, magnetic, porous, defect and various other classes of bulk materials.[48,49] We have also previously looked at the band alignment of layered two dimensional materials using JARVIS-DFT.[19] However, three dimensional systems with chemical bonding between the materials require much greater effort, as the interfacial bonding has a much greater effect on the interface properties, and the determination of even a single interface structure is a challenging task. Out of the above material class combinations, semiconductor–semiconductor are of special interest for this work. As DFT calculations can be time-consuming for surfaces and interfaces machine-learning (ML)/deep learning (DL) techniques based on DFT data can be used to accelerate atomistic predictions.[50,51] Such models have often been applied for bulk property predictions and their applicability for defects and interfaces remains an open question. Several machine learning tools available in JARVIS such as classical force-field inspired descriptors (CFID),[43] atomistic line graph neural network (ALIGNN),[52,53] computer vision for atomistic images (AtomVision)[54] and natural language processing for chemistry (ChemNLP)[55,56] can be used in this regards to accelerate the interface design tasks. In particular, ALIGNN has been used to develop several fast surrogate models for property predictions as well as a unified force-field for fast structure optimizations.

Most importantly, for all the above predictions, it is important to benchmark and quantify error with respect to experimental data to gain confidence in the prediction methodology. This work addresses the above challenges and provides a streamlined framework for semiconductor interface design (InterMat). Although focusing on semiconductors, this work has relevance to other applications such as battery, data-storage, and solar-cell devices. We believe that this work will be a precursor to more thorough theoretical and experimental investigations of semiconductor interfaces.

Results and discussion

A schematic overview of InterMat along with the combinatorial problem of interfaces is shown in Fig. 1. InterMat can be used to generate surface and interface structures, perform multi-fidelity calculations to predict properties, analyze and benchmark data against experiments, and train and utilize machine learning models based on the resulting data. Initial atomic structures can be obtained from the JARVIS-DFT repository. As an example of the combinatorial challenge of interfaces, we can consider starting with the 20 901 semiconductors in the JARVIS-DFT database with OptB88vdW band gaps between 0.1 eV and 6 eV. Using a maximum Miller index (M) of 1, 2, and 3, the number of symmetrically distinct surface slabs are 186 847, 591 639 and 1 642 584, respectively. Using these surfaces, the number of binary interface systems that can be generated are 17.5 billion, 175 billion and 1.4 trillion. Including the possibility of different atomic surface terminations, reconstructions, and defects further complicates matters.

It is unrealistic to analyze such a large search space to find combinations for device applications by using conventional experimental or computational techniques. We will instead use ALIGNN models trained on bulk materials to guide and prioritize DFT calculations among the vast pool of candidate surfaces and interfaces. We develop machine learning models for fast predictions of valence band maxima (VBM) and conduction band minima (CBM) using ALIGNN on JARVIS-DFT bulk material dataset. These predictions can be further used for fast IU based band alignment using electron affinity/Anderson's rule.[39] To assess the strengths and limitations of such models, we develop a surface dataset for independent unit (IU) models and an interface dataset for alternate slab junction (ASJ)/superlattice models using DFT. We particularly focus on industrially relevant semiconductors including group IV (C, Si, Ge *etc.*), III–IV (AlN, GaN, GaAs, GaP, InSb *etc.*), II–VI (CdS, CdSe, ZnO, ZnS *etc.*). We also assess the strengths and limitations of IU and ASJ models against experimental measurements. This DFT dataset can then be fed back into the ALIGNN models to further improve accuracy.

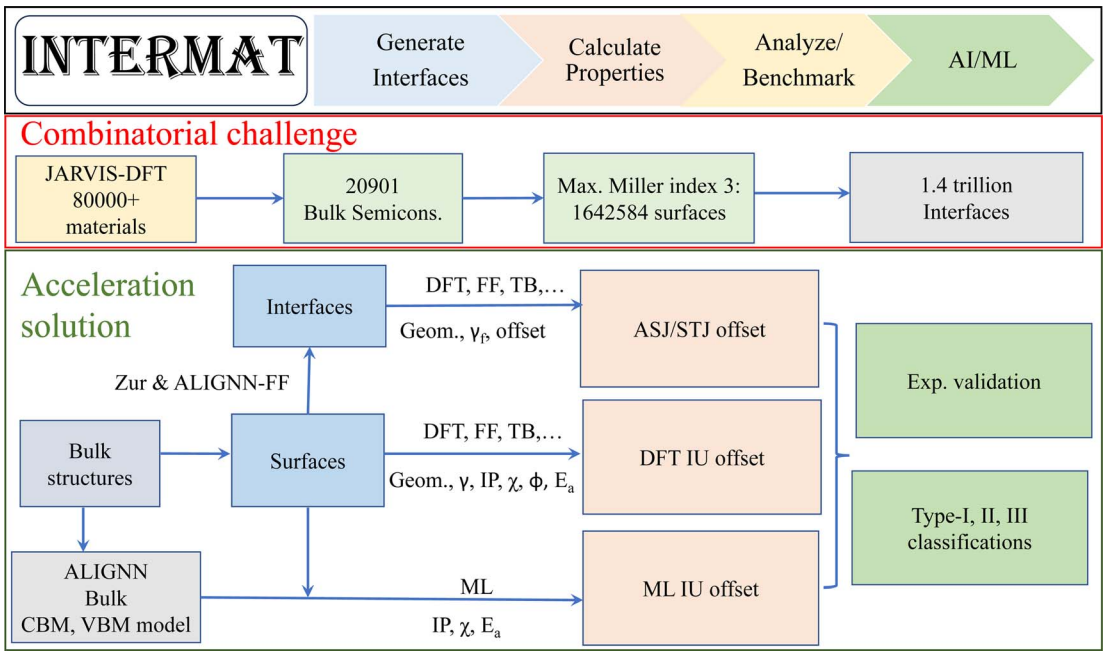

Fig. 1 Schematic overview of the workflow. InterMat can be used to generate surface and interface geometric structures (Geom.), perform multi-fidelity calculations (such as density functional theory, force-field, tight-binding and machine learning) to predict properties (such as surface energy, interface formation energy, band offset, work function, ionization potential, electron affinity), analyze and benchmark data against experiments, and utilize machine learning models for such data. The number of possible semiconductor–semiconductor interfaces is exceedingly large. The workflow aims to provide a toolkit to generate interface structures and use multi-fidelity methods to accelerate interface/heterostructure design.

DFT surface dataset: work function, electron affinity, ionization potential and surface energy

We develop a dataset of non-polar unreconstructed slab surfaces using the JARVIS-DFT workflow and bulk material dataset. Examples of silicon and gallium arsenide bulk atomic structures are shown in Fig. 2a and b respectively. Next, we generate surface slab structures with a thickness of 1.6 nm and vacuum padding of 1.2 nm, as shown in Fig. 2c. Recently, it was shown that vacuum and slab thicknesses of at least 10 Å are sufficient for surface models.[77] During the DFT calculations, the converged k-point[78] values from the relevant bulk calculations are used for surfaces. We optimize the internal coordinates of these surfaces keeping the cell volume constant.

We carefully benchmark surface energy (γ), ionization potential (IP), electron affinity (χ), and work function (ϕ) values against experimental measurements from the literature. The surface energy (γ) can be calculated using the formula:

$$\gamma = \frac{E_{\text{slab}} - N_{\text{bulk}} \cdot E_{\text{bulk}}}{2A} \quad (1)$$

where (E_{slab}) is the total energy of the relaxed slab model, (N_{bulk}) is the number of bulk-like atoms in the slab model, (E_{bulk}) is the energy per atom in the bulk material, and (A) is the surface area of the slab model. The factor of 2 accounts for the fact that there are two surfaces in the non-polar slab model (top and bottom).

We obtain the valence band maximum (VBM) and vacuum level (E_{vac}) of surface slabs from DFT calculations using the OptB88vdW functional. The work function is obtained by subtracting the vacuum level from the Fermi level ($\phi = E_{\text{vac}} - E_{\text{F}}$). Similarly, the ionization potential is the difference between the VBM and E_{vac}. Then, we add the electronic bandgap (E_{g}) of the bulk material to the ionization potential to get the electron affinity (EA, χ). Semi-local DFT has proven quite effective in describing the valence bands of materials but is known to underestimate band gaps. For accurate prediction of both valence and conduction bands, particularly in materials with complex electronic interactions, higher-level theories like many-body perturbation theory (*e.g.* GW calculations) might be necessary, but are computationally very expensive. In order to address this problem in a more computationally efficient manner, we make use of bulk band gaps from the JARVIS-DFT database computed using the TBmBJ metaGGA functional. TBmBJ predictions can provide band gap descriptions with accuracy close to more expensive methods but at an order of magnitude less computational cost,[79] which is important for high-throughput studies. We calculate surface conduction band quantities by first calculating the valence band at the GGA-level using OptB88vdW and then add to that the TBmBJ bulk gap to get the conduction band minimum (CBM). Performing full surface calculations using hybrid functionals or GW is beyond the scope of the present work because of excessive computational cost, but we plan to provide further tests of those approaches in the future.[14,15]

In order to benchmark our surface dataset, we compare work functions, electron affinities, and surface energies of several

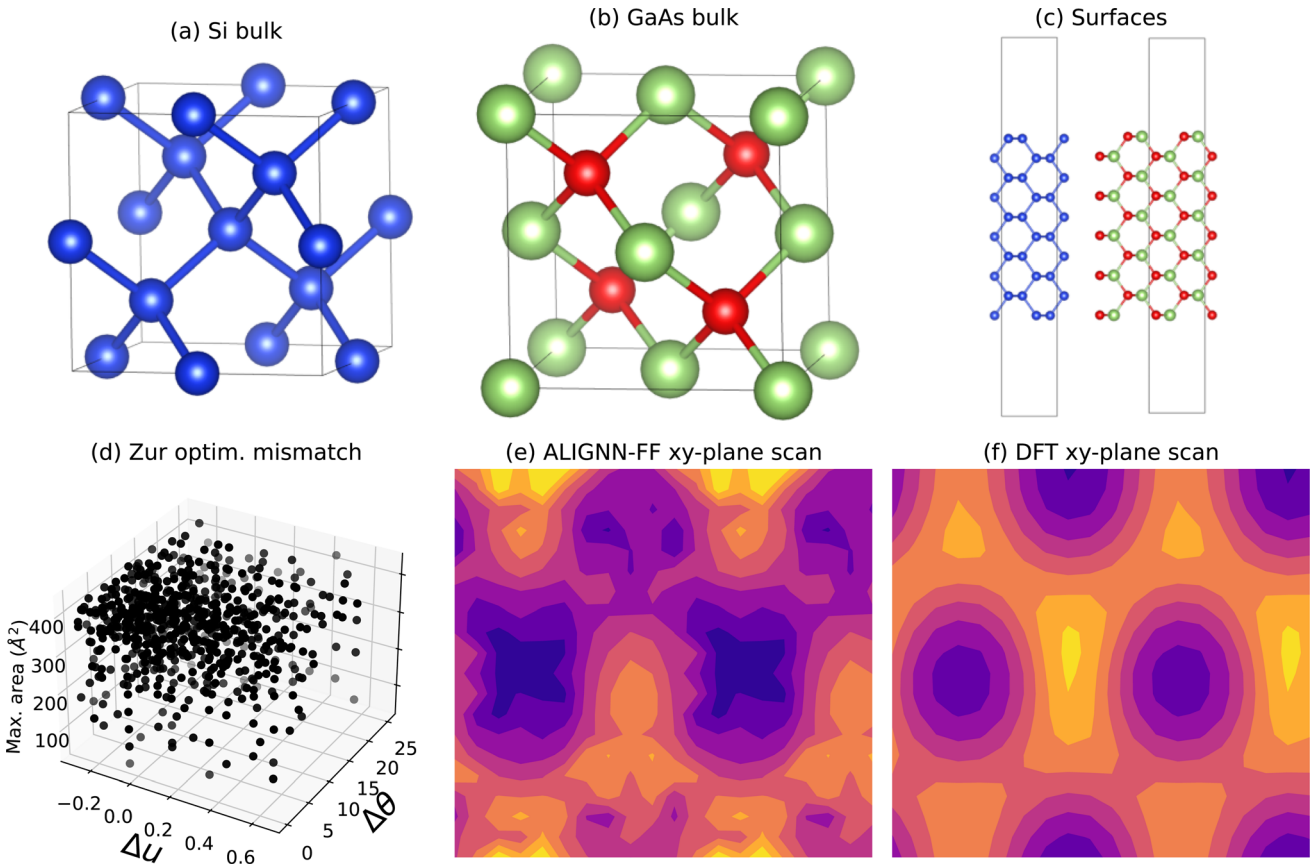

Fig. 2 Structure generation criteria selection and initial *xy*-plane scan with ALIGNN-FF for Si(110)/GaAs(110) interface. (a) atomic structure of silicon (Si), (b) atomic structure of gallium arsenide (GaAs), (c) surfaces (110) generated from the bulk structures of Si (left) and GaAs (right), (d) candidate interface parameters from Zur algorithm: mismatch in *x*-direction (u), mismatch in *y*-direction (θ) and maximum allowed area to generate suitable structures. (e) ALIGN-FF and (f) DFT energy as a function of displacement in *xy*-plane of interface.

dozen surfaces with experimental data in Table 1. We find excellent agreement for the work functions, with a mean absolute error value of 0.29 eV, consistent with previous benchmarking efforts.[80] Similarly, we obtain a mean absolute error of 0.39 eV and 0.34 Jm^{-2} for the electron affinity and surface energy, respectively. Currently, we have performed calculations on 607 surfaces using the workflow, and the dataset is still growing. Using, these 607 surfaces, 183 921 IU-band offsets can be predicted. Also, we plan to include reconstructed and polar surfaces in the future.

DFT interface dataset: alternate slab junction (ASJ) band alignment

We next consider explicit DFT calculations of interfaces, which first require generating candidate interface structures. This can be done in either of the following two ways: (1) by attaching the two surface slabs together without vacuum padding, creating a superlattice or alternating slab junction (ASJ) structure, or (2) by attaching the two surface slabs with vacuum padding, creating a surface terminated junction (SJT) structure. We have focused on the ASJ approach.[16,81] After obtaining surface slab structures as discussed in the previous section, we generate the interfaces following the Zur *et. al.* algorithm.[82] The Zur algorithm generates a number of superlattice transformations within a specified maximum surface area and also evaluates the length and angle between film and substrate superlattice vectors to determine if they can match within a tolerance. This algorithm is applicable to different crystal structures and their surface orientations. We use a maximum lattice mismatch of 8%, maximum area of 300 $Å^2$, and maximum angle tolerance of 1°. Note that in previous studies, lattice mismatch of 20% has been reported.[14,83] An example of the application of the algorithm to the Si(110)/GaAs(110) interface is shown in Fig. 2d with several lattice length and angle mismatches (Δu and $\Delta\theta$) as well as maximum area. After eliminating structures with area higher than max-area tolerance and structures with mismatch angle more than the specified angle threshold, we then choose the remaining structure (if any) with the minimum mismatch lattice vector lengths.

The Zur algorithm determines a candidate unit cell, but the relative alignment of the structures in the in-plane, as well as the slab terminations still need to be decided. For the in-plane alignment, we perform a grid search of possible options with a spacing interval of 0.05 fractional coordinates to determine the initial structure for further relaxation. Doing such a large number of calculations with DFT would be prohibitive, so we

Table 1 Work function (ϕ, eV), electron affinity (χ, eV) and surface energy (γ, Jm^{-2}) of a few unreconstructed non-polar surface slabs from OptB88vdW (OPT) against experimental data. The IDs represent JARVIS-DFT identifiers

System	IDs	Miller	ϕ (OPT)	ϕ (Exp)	χ (OPT)	χ (Exp)	γ (OPT)	γ (Exp)
Si	1002	111	5.00	4.77 (ref. 57)	4.10	4.05 (ref. 58)	1.60	1.14 (ref. 59)
Si	1002	110	5.30	4.89 (ref. 57)	4.10	—	1.66	1.9 (ref. 59)
Si	1002	001	5.64	4.92 (ref. 57)	3.60	—	2.22	2.13 (ref. 60)
C	91	111	4.67	5.0 (ref. 61)	−2.9	—	5.27	5.50 (ref. 62)
Ge	890	111	4.87	4.80 (ref. 63)	5.2	4.13 (ref. 32)	0.99	1.30 (ref. 60)
SiGe	105 410	111	4.93	4.08 (ref. 64)	4.5	—	1.36	—
SiC	8118	001	5.26	4.85 (ref. 65)	1.3	—	3.51	—
GaAs	1174	110	4.89	4.71 (ref. 66)	4.40	4.07 (ref. 58)	0.67	0.86 (ref. 59)
InAs	1186	110	4.85	4.90 (ref. 66)	4.9	4.9 (ref. 32)	0.57	—
AlSb	1408	110	5.11	4.86 (ref. 66)	3.70	3.65 (ref. 32)	0.77	—
GaSb	1177	110	4.48	4.76 (ref. 66)	3.70	4.06 (ref. 32)	0.71	—
AlN	39	100	5.56	5.35 (ref. 65)	1.3	2.1 (ref. 67)	2.27	—
GaN	30	100	5.74	5.90 (ref. 68)	2.8	3.3 (ref. 69)	1.67	—
BN	79 204	110	6.84	7.0 (ref. 70)	1.4	—	2.41	—
GaP	1393	110	5.31	6.0 (ref. 65)	4.0	4.3 (ref. 58)	0.88	1.9 (ref. 59)
BP	1312	110	5.61	5.05 (ref. 71)	2.8	—	2.08	—
InP	1183	110	5.17	4.65 (ref. 66)	4.10	4.35 (ref. 58)	0.73	—
CdSe	1192	110	5.70	5.35 (ref. 72)	6.4	—	0.38	—
ZnSe	96	110	5.67	6.00 (ref. 73)	5.4	—	0.44	—
ZnTe	1198	110	5.17	5.30 (ref. 74)	4.10	3.5 (ref. 32)	0.36	—
Al	816	111	4.36	4.26 (ref. 61)	—	—	0.82	—
Au	825	111	5.5	5.31 (ref. 61)	—	—	0.90	—
Ni	943	111	5.35	5.34 (ref. 61)	—	—	2.02	2.34 (ref. 75)
Ag	813	001	4.5	4.2 (ref. 61)	—	—	0.99	—
Cu	867	001	4.7	5.1 (ref. 61)	—	—	1.47	—
Pd	963	111	5.54	5.6 (ref. 61)	—	—	1.57	—
Pt	972	001	5.97	5.93 (ref. 61)	—	—	1.94	—
Ti	1029	100	3.84	4.33 (ref. 61)	—	—	2.27	—
Mg	919	100	3.76	3.66 (ref. 61)	—	—	0.35	—
Na	931	001	2.97	2.36 (ref. 61)	—	—	0.10	—
Hf	802	111	3.7	3.9 (ref. 61)	—	—	2.02	—
Co	858	001	5.22	5.0 (ref. 61)	—	—	3.49	—
Rh	984	001	5.4	4.98 (ref. 61)	—	—	2.46	—
Ir	901	100	5.85	5.67 (ref. 61)	—	—	2.77	—
Nb	934	100	3.87	4.02 (ref. 61)	—	—	2.41	—
Re	981	100	4.96	4.72 (ref. 61)	—	—	2.87	—
Mo	21 195	100	4.17	4.53 (ref. 61)	—	—	3.30	
Zn	1056	001	4.27	4.24 (ref. 76)	—	—	0.36	—
Bi	837	001	4.31	4.34 (ref. 61)	—	—	0.65	0.43 (ref. 77)
Cr	861	110	5.04	4.5 (ref. 61)	—	—	3.31	—
Sb	993	001	4.64	4.7 (ref. 61)	—	—	0.67	—
Sn	1008	110	4.82	4.42 (ref. 61)	—	—	0.91	—
MAE	—	—	0.29	—	0.39	—	0.34	—

use ALIGNN-FF[53] to identify the starting in-plane alignment. ALIGNN-FF is a universal force field ML model developed using JARVIS-DFT data with 307 113 structures and can be used to model combinations of 89 elements from the periodic table. An example of ALIGNN-FF predictions of an in-plane grid search is shown in Fig. 2e. For the Si/GaAs(110) case, we also perform corresponding DFT calculations as shown in Fig. 2f. Here high-peaks (yellow color using magma colormap) usually represent too close atoms during the translation operations, which should be avoided. Clearly, the DFT contours are smoother than ALIGNN-FF because of its relatively rough potential energy surface (PES). Nevertheless, the minimums of the contours, which are of interest for in-plane alignments, closely resemble each other. As the ALIGNN-FF accuracy increases with more data, we expect to get much smoother PES in future. After the ALIGNN-FF calculations to select the initial alignment, a full DFT relaxation is performed.

For computational purposes, it is important to have a unique identifier for an interface. While generating the interfaces, we use a naming convention to include (a) material IDs (such as JVASP-1002 for Si and JVASP-1174 for GaAs), (b) film and substrate Miller indices (such as 110 for each), (c) film and substrate thickness values (such as 16 Å each), (d) separation between these two surface slabs (such as 2.5 Å for an ASJ model, 18 Å for STJ interface models), (e) relative displacement in *xy*-plane (such as a displacement vector of [0.5, 0.2]), (f) calculator method (such as DFT (VASP), ALIGNN-FF *etc.*) giving rise to an interface with a name such as: interface-JID1_JID2_film_miller_M1_sub_miller_M2_film_thickness_T1_subs_-thickness_T2_separation_S_disp_X_Y_vasp (where JID1 is

JVASP-1002, JID2 is JVASP-1174, M1 and M2 are both 1_1_0, T1 and T2 are 16, S is 2.5, X is 0.5, Y is 0.2). Such a scheme helps to reproduce the unique interfaces. Of course, realistically, more complex parameters for an interface can be important such as terminations, reconstructions, misfit-dislocation, vacancies on the interface *etc.*, but they can be easily included in the naming scheme as well later.

After selecting a good guess of the interface using the above approach, DFT calculations are performed to calculate quantities such as the interface formation energy and the band offset value. During the DFT calculations, the more converged k-point grids and energy cutoffs of the two constituent bulk materials[78] is used. An example of the Si(110) and GaAs(110) interface is shown in Fig. 3a. Furthermore, we can project the electron density of states across the cell dimension in Fig. 4 to show how electronic states are distributed along the interface region. We observe the GaAs gap decrease near the silicon region. Such analysis can help to understand the local band alignment and atomic character, which are important for device modeling.

After determining the optimized geometric structure for the interface using DFT, we obtain the interface formation energy and valence band offset data using the formalism detailed in ref. 84 and 11 respectively. As an example, we show a detailed analysis of Si(110)/GaAs(110) and AlN(001)/GaN(001) in Fig. 3. The interface formation energy (γ_f) is calculated using the formula:

$$\gamma_f A = E_{tot} - \sum_i n_i \mu_i \quad (2)$$

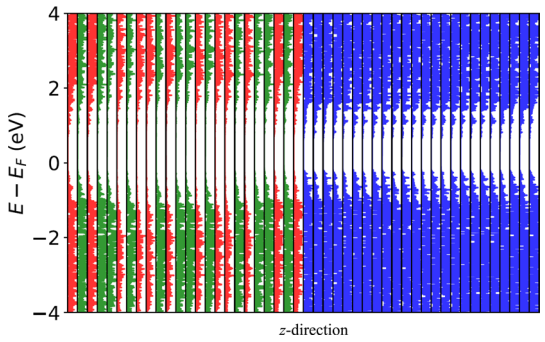

Fig. 4 Atom projected density of states for the system, separated along the z direction, normal to the interface. Red, green and blue colors represent gallium, arsenic and silicon atom contributions respectively.

where γ interface formation energy, E_{tot} is the total energy of the superlattice, μ_i is the chemical potential of the specie i, n_i is the number of atoms of the specie i, and A is the interface unit cell area. Using the bulk materials energy per atom in its most stable form in JARVIS-DFT and OptB88vdW functional, we obtain an interface formation energy of -0.056 Jm^{-2} for the Si(110)/GaAs(110) system. A negative formation energy suggests a feasible formation of the interface. Moreover, such interface formation energies with varying chemical potentials of the constituent elements can provide information about the thermodynamic stability of the interface in different growth

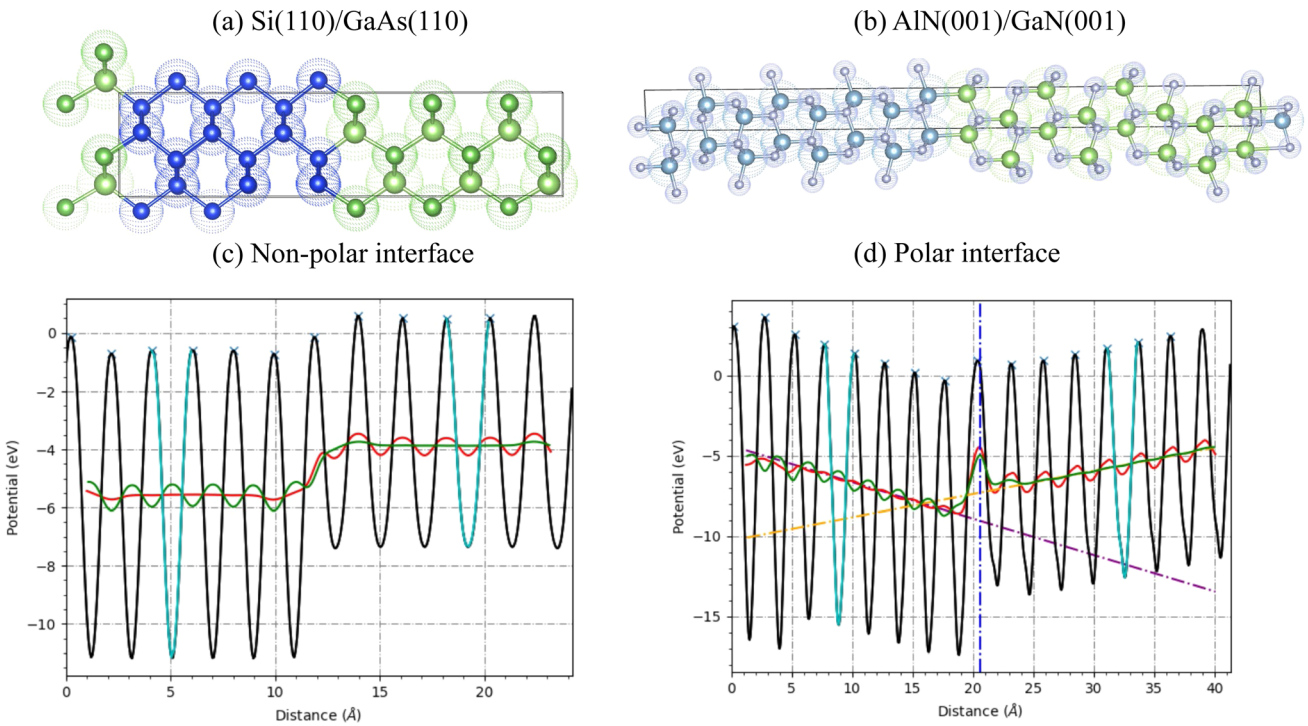

Fig. 3 Atomic structures and band-alignment using the average electrostatic potential of semiconductor interfaces. (a) atomic structure view of Si/GaAs(110), (b) atomic structure view of polar interface AlN/GaN(001), (c) electrostatic potential profile for non-polar interface Si/GaAs(110), (d) average electrostatic potential profile for polar interface of AlN/GaN (001). The cyan lines are used to find the repeat unit layers for the left and right parts. The red and green lines show the potential averaged over the repeat distances of the left and right slabs, respectively. The dotted vertical blue line marks the interface.

conditions. Such detailed tasks for individual interfaces will be carried in future.

In Fig. 3a, we show the atomic structure of the ASJ based heterostructure of Si(110)/GaAs(110). The left side (with blue atoms) represents the Si and the right side is the GaAs region. In Fig. 3c, we show the electrostatic potential profile, averaged in-plane, of the interface. The approximately sinusoidal profile on both regions represents the presence of atomic layers. The cyan lines show the region used to define the repeat distance, L, used for averaging in each material (see below). The red and green lines show the average potential profiles for the left and right parts using the repeat distance. The valence band offset (ΔE_v) of an interface between semiconductor A and B, ΔE_v is obtained using eqn (4). The difference in the averages for the left and right parts gives the ΔV term. Now the bulk VBMs of the left and right parts are also calculated to determine the ΔE. The sum of these two quantities gives the valence band offset that can be compared to experiments.

$$\Delta E_v(\mathrm{A/B}) = (E_v^{\mathrm{B}} - E_v^{\mathrm{A}}) + \Delta V \quad (3)$$

$$\Delta V = \overline{\overline{V}}_{\mathrm{A}} - \overline{\overline{V}}_{\mathrm{B}} \quad (4)$$

Here, E_v^{A} (E_v^{B}) represents the position of the VBM with respect to the average electrostatic potential in the bulk material A (B), and ΔV represents dipole potential or the difference between the macroscopic-averaged electrostatic potential between A and B. Moreover, $\overline{\overline{V}}$ is the average along the repeat unit L of $\bar{V}$, which is the planar averaged electrostatic potential:

$$\overline{\overline{V}}(z) = \frac{1}{L}\int_{-L/2}^{L/2} \overline{V}\left(z + z'\right) \mathrm{d}z' \quad (5)$$

$\bar{V}$ is given by:

$$\overline{V}(z) = \frac{1}{S}\int_S V(x, y, z)\mathrm{d}x\mathrm{d}y \quad (6)$$

where, L is the distance between repeat units and S is the area which is parallel to the interface. The corresponding conduction-band offset can be determined by using TBmBJ band-gap values from the respective bulk calculations or experimental data. We will use the convention that a positive value of the valence-band offset at an interface A/B indicates that the VBM is higher in material B. For the GaAs/Si interface we obtain ΔE_v of 0.31 eV and 0.39 eV using OptB88vdW and R2SCAN functionals, respectively, which is in close agreement with the experimental value of 0.23 eV.

Next, we show a polar semiconductor heterojunction example for AlN (001)/GAN (001) interface in Fig. 3b. The electrostatic potential profile is shown in Fig. 3d. In contrast to flat average potential values in Fig. 3c, we observe inclined profiles for this system indicating the presence of a constant electric field. We fit lines for both sides and extrapolate to the interface. The difference of the lines at the interface gives ΔV. The calculation of ΔE remains the same as the non-polar case. These calculations are automated in the workflow, however, it is important to check that the slab is thick enough to define a bulk-like region where $\overline{\overline{V}}(z)$ is linear. Now, in the Table 2 we compare some of the ASJ based valence band offsets (ΔE_v) with experimental measurements. We find a mean absolute error of 0.22 eV and 0.23 eV for OptB88vdW and R2SCAN respectively, which is comparable a value of 0.16 eV from to Liberto *et. al.*[16] for a smaller number of systems using the HSE06 functional. In the future, we plan to carry out HSE06 calculations for surfaces

Table 2 Valence band offsets (in eV) of a few independent unit (IU)/Anderson's model and alternating slab-junction (ASJ) based semiconductor/semiconductor interfaces with OptB88vdW (OPT) and R2SCAN functionals in comparison to previously reported experiments. Here ID, Miller and P represent a JARVIS-DFT identifier, Miller index and polar surface interfaces respectively

System	ID	Miller	IU (OPT)	ASJ (OPT)	ASJ (R2SCAN)	Exp
AlP/Si	1327/1002	110/110	1.24	0.88	1.04	1.35 (ref. 16)
GaAs/Si	1174/1002	110/110	0.30	0.31	0.39	0.23 (ref. 85)
CdS/Si	8003/1002	110/110	3.22	1.48	1.70	1.6 (ref. 86)
AlAs/GaAs	1372/1174	110/110	0.60	0.48	0.50	0.55 (ref. 87)
CdS/CdSe	8003/1192	110/110	0.35	0.10	0.11	0.55 (ref. 88)
InP/GaAs	1183/1174	110/110	0.25	0.72	0.75	0.19 (ref. 89)
ZnTe/AlSb	1198/1408	110/110	0.8	0.25	0.33	0.35 (ref. 90)
CdSe/ZnTe	1192/1198	110/110	1.8	0.58	0.67	0.64 (ref. 91)
InAs/AlAs	1186/1372	110/110	—	0.46	0.39	0.5 (ref. 92)
InAs/AlSb	1186/1408	110/110	—	0.05	0.16	0.09 (ref. 93)
ZnSe/InP	96/1183	110/110	—	0.13	0.18	0.41 (ref. 94)
InAs/InP	1186/1183	110/110	—	0.11	0.09	0.31 (ref. 89)
ZnSe/AlAs	96/1372	110/110	—	0.38	0.45	0.4 (ref. 95)
GaAs/ZnSe	1174/96	110/110	—	0.72	0.80	0.98 (ref. 96)
ZnS/Si	10591/1002	001/001	—	0.92	1.16	1.52 (ref. 97)
Si/SiC	1002/8118	001/001	—	0.51	0.47	0.5 (ref. 98)
GaN/SiC (P)	30/8118	001/001	—	1.12	1.37	0.70 (ref. 99)
Si/AlN (P)	1002/30	001/001	—	3.51	3.60	3.5 (ref. 100)
GaN/AlN (P)	30/39	001/001	—	0.80	0.86	0.73 (ref. 101)
AlN/InN (P)	39/1180	001/001	—	1.24	1.07	1.81 (ref. 102)
GaN/ZnO (P)	30/1195	001/001	—	0.51	0.46	0.7 (ref. 103)
MAE	—	—	0.45	0.22	0.23	—

as well as interfaces to further improve the quality of predictions. These benchmarks will also be available in the JARVIS-Leaderboard platform[104] as well. Out of numerous possible combinations, only 593 DFT calculations of ASJ-based interfaces are available right now and the database is still growing.

DFT-based independent unit (IU) band alignment

IU band alignment, also known as Anderson's rule,[39] predicts semiconductor band offsets at interfaces using only the IP and EA data from independent surface calculations. For a semiconductor heterojunction between A and B, the conduction band offset is given by:

$$\Delta E_c = \chi_B - \chi_A \quad (7)$$

Similarly, the valence band offset is given by:

$$\Delta E_v = (\chi_A + E_{gA}) - (\chi_B + E_{gB}) \quad (8)$$

In Fig. 6a, we show the DFT-based IU band alignments for a set of well-known semiconductor surfaces. We also include dotted lines for the energy levels of H_2 and H_2O, which are relevant for photo-catalyst applications. We compare the DFT based IU band offsets for 8 interfaces in Table 2 against experiments. We find a mean absolute error of 0.45 eV which is similar to a value of 0.32 eV as found in ref. 16 for different systems.

ALIGNN-based IU alignment from bulk data

We seek to accelerate the prediction of band edges using ML models, but the absolute prediction of band edges relative to vacuum requires DFT calculations with surfaces, which are too computationally expensive to create a robust dataset. JARVIS-DFT contains a much larger dataset of three dimensional materials with CBM and VBM values. Here, the CBMs and VBMs are simply the band edges written out by VASP for the bulk materials dataset using OptB88vdW. These band edges use the VASP convention that the average electrostatic potential of a unit cell is set to zero, and are not directly comparable to experimental values. A surface calculation with explicit vacuum is necessary to align the VBM/CBM to vacuum. We first train an ML model using ALIGNN based on these bulk quantities, but we will then show that this model is somewhat surprisingly also useful for predictions of band edges relative to vacuum.

To train the ALIGNN model, we split each bulk VBM/CBM dataset into 90 : 5 : 5 train : validation : test parts. We train on 90% train data and evaluate the validation and test data using ALIGNN. We find a mean absolute error (MAE) of 0.28 ev for CBM and VBM. We note that the CBM/VBM data can vary from −10 to 10 eV (as shown in Fig. 5a and b) suggesting that the model should be reasonable for predictions. The mean absolute deviation (MAD) for the CBM and VBM are 2.08 eV and 2.67 eV, respectively, so the MAD : MAE is nearly 10 relative to a trivial baseline model. Out of several other material properties trained using ALIGNN,[52] CBM/VBM models has one of the highest MAD : MAE ratios, especially compared to other electronic properties like the band gap. We show the CBM and VBM

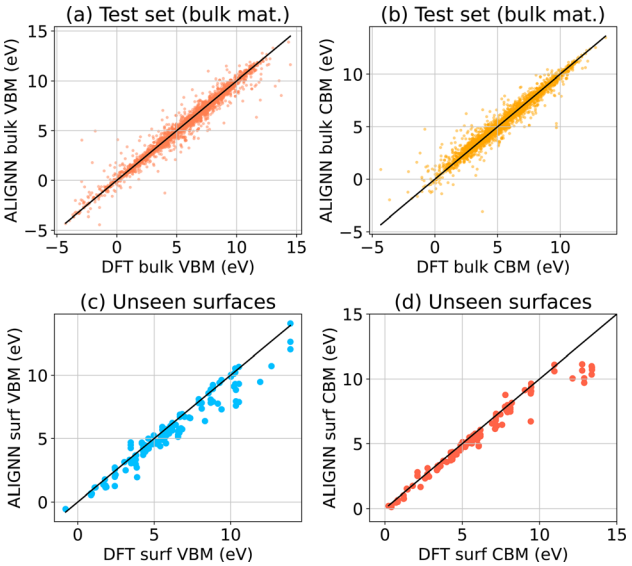

Fig. 5 ALIGNN based regression models for (a) VBM for JARVIS-DFT 3D/bulk materials test set data, (b) CBM for bulk materials test set data, (c) slab surface VBMs not part of the training set to evaluate extrapolation strength, (d) slab surface CBMs not part of the training set to evaluate extrapolation strength.

prediction models in Fig. 5a and b respectively. We believe with more data using the active learning loop we can further increase the MAD : MAE in future.[50]

Next, we evaluate the bulk-trained ALIGNN models on the DFT surface dataset and show results in Fig. 5c and c for CBM and VBM respectively. We note that the training data does not include any surfaces. Nevertheless, most of the data points are on $x = y$ line suggesting excellent agreement. We find MAE values of 0.55 eV and 0.96 eV for VBM and CBM respectively. This level of agreement is surprising because the value of the averaged electrostatic potential will change as the ratio of vacuum thickness to slab thickness changes. We also note that the error in CBM is higher than that of VBM, perhaps an explanation for why predicting band gaps of materials using machine learning is ever-standing difficult problem.

Up to this point, our model can only predict quantities relative to a cell-averaged electrostatic potential, which cannot be directly compared to experiment. However, we observe that our ALIGNN model predictions of the bulk VBM/CBM are in fact strongly correlated with the VBM/CBM values calculated with respect to vacuum using DFT calculations of surface slabs. We can get useful predictions of the vacuum-aligned VBM by subtracting a heuristic constant value of 10 eV from the ALIGNN predictions, and adding the bulk TBmBJ band gap to get the corresponding CBM. We visualize this IU based band alignment using ALIGNN model in Fig. 6b. We observe that the overall trends of DFT and ALIGNN closely resemble each other. For these surfaces, we calculate the classification accuracy of the heterostructures in type-I, type-II and type-III heterostructures. We find precision scores of 66.7%, 66.1% and 58.2% respectively. Precision is defined as the fraction of relevant instances among all of the retrieved instances. The classification

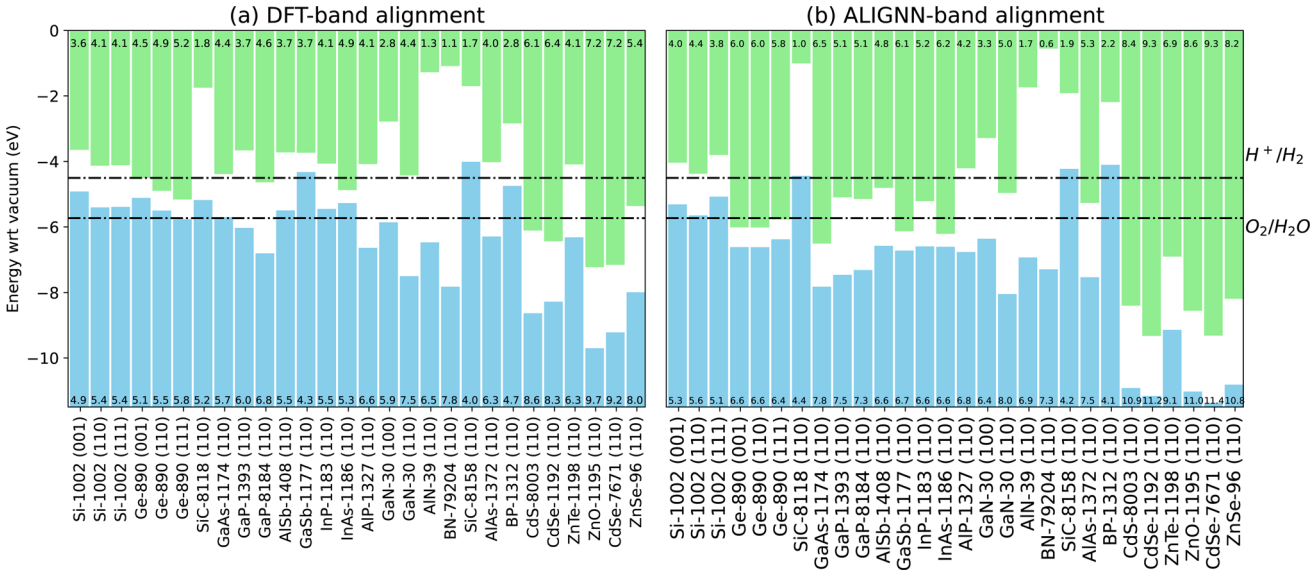

Fig. 6 A few examples of band alignment based on the ionization potential (IP) and electron affinity (EA) for Anderson's/independent unit (IU) model. (a) Density functional theory based alignments can be used to obtain band offsets. (b) Fast ALIGNN based band alignment predictions. The numbers in the green bars represent electron affinity while that in the blue bars represent ionization potentials. The trained models based band alignment will be available at JARVIS-Heterostructure website soon.

precision scores are based on DFT optimized surfaces, which are not available for all the materials in the database. So, we generate structures directly from the bulk counterparts, relax them using ALIGNN-FF and then predict the electron affinity and ionization potentials using the procedure mentioned above. In this way, we find precision scores of 55.0%, 63.4% and 60.0% respectively suggesting that structure optimization of surfaces has an impact on the ALIGNN predictions. The random baseline is $1/3 = 33\%$, which is more than 2 times lower than what we achieve.

As an example of the type of analysis that can be done with this data, we analyze all heterostructures where the film is silicon. As shown in Fig. 7, we identify which elements in the second semiconductor make it most likely that the heterostructure will have a type-1/straddling band alignment appropriate for diode applications. We find these elements to be Al, P, S, N, O, Li which is consistent with known silicon devices. Many other analyses are possible, we provide this data in hopes that it will be useful to the community.

Sufficiently high precision scores suggest that such models can be used for pre-screening applications followed by density functional theory calculations and experiments. Also, as the DFT bulk, surface and interface dataset is growing continuously, there is plenty of scope to improve the model performance in the future. For the 1.4 trillion semiconductor interfaces, we find 294 billion as type-I, 322 billion as type-II and rest as type-III heterostructures using the ALIGNN + IU model. The results suggest that finding type-I interfaces for transistor applications is more challenging than other heterostructure applications. Having such a large number of options and further screened for desirable properties such as effective masses, dielectric, piezoelectric, thermoelectric properties *etc.* can be helpful for technological applications. We emphasize the point that AI models should be considered as a pre-screening step only and would require thorough DFT and/or experimental validation.

Given the lack of surface-specific training data, the level of agreement with the unseen surface data is surprising, however, there are discrepancies in some cases. In other words, the model has never seen bonding environments that occur on slab surfaces, and extrapolating to such environments is challenging. The goal here is to obtain a fast model that can be used for quick screening of surfaces, with subsequent DFT calculations for confirmation. It may be possible to finetune this ALIGNN model with a surface dataset to further improve the accuracy of the model, but we leave that for future work. Nevertheless, the close resemblance in alignment predictions is promising and suggests that our models can be useful. Similar successful extrapolations for bulk-trained models were observed in ref. 105, which demonstrates that vacancy energies can be predicted from a ML model fit to bulk crystal data only. We clarify that we are not using a ALIGNN model to predict either the OptB88vdW or TBmBJ band gaps in this work. We are simply looking up the bulk band gaps in the JARVIS-DFT database, so this does not contribute to the error. However, we do have models for these quantities with MAEs of 0.14 eV and 0.31 eV respectively (see ref. 52 and 104). For materials not in the JARVIS-DFT database, we could use these models to predict the gaps.

In summary, we have provided a computational framework and dataset for investigating interface systems using multi-fidelity computational approaches. We have developed one of the largest datasets, containing 607 surfaces, 183 921 IU-band offsets, and 593 ASJ interface band offset using DFT. Using

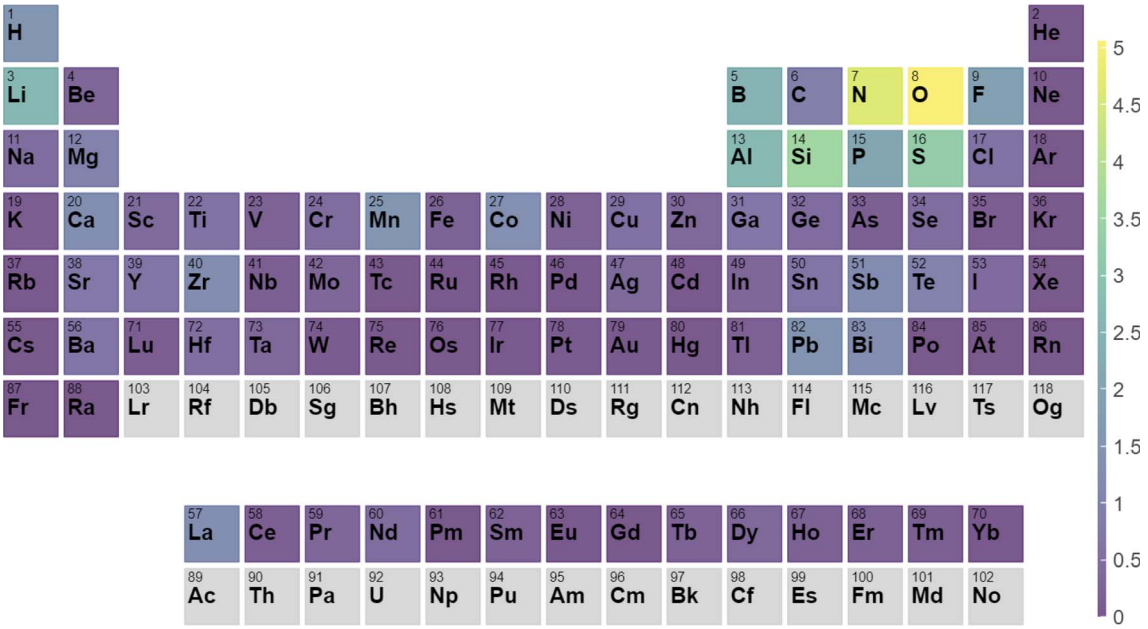

Fig. 7 Percentage chance that a heterostructure with a given element will form a type-I heterostructure with silicon as the second semiconductor.

universal graph neural network models, we have quickly screened potential semiconductor device candidates as transistors from a pool of 1.4 trillion possible interfaces, which would not have been possible using conventional computational or experimental techniques. Although we have applied this framework for semiconductors, it can be useful for other technological applications as well. After pre-screening, we have shown and benchmarked this streamlined workflow for band offset predictions using the independent unit and alternate slab junction models. This work paves the way for the application of materials design approach to interface systems. All of the tools and datasets developed in this work will be distributed publicly in the spirit of open-science.

Methods

Graph neural networks are trained using Atomistic Line Graph Neural Network (ALIGNN) framework[52] which uses PyTorch and deep graph library (DGL). Such GNN models can be used for graph level prediction (such as total energy of the system, bandgap *etc.*) or node level predictions (such forces, charges, atomic magnetic moments *etc.*) In ALIGNN, a crystal structure is represented as a graph using atomic elements as nodes and atomic bonds as edges. Each node in the atomistic graph is assigned 9 input node features based on its atomic species: electronegativity, group number, covalent radius, valence electrons, first ionization energy, electron affinity, block and atomic volume. The inter-atomic bond distances are used as edge features with radial basis function up to 8 Å cut-off and a 12-nearest-neighbor (N). This atomistic graph is then used for constructing the corresponding line graph using interatomic bond-distances as nodes and bond-angles as edge features. ALIGNN uses edge-gated graph convolution for updating nodes as well as edge features using a propagation function (f) for layer (l), atom features (h), and node (i), details of which can be found in ref. 52 and 53:

$$h_i^{(l+1)} = f(h_i^l\{h_j^l\}_i) \quad (9)$$

ALIGNN is trained for 500 epochs and with default parameters in the package. We use a 90 : 5 : 5 training : validation : testing randomly distributed data split for the CBM and VBM of the bulk materials dataset. The data splits and corresponding identifiers used during the training are made available in the figshare repository. While ALIGNN was used as surrogate/property prediction model at a graph level, in the later version, we also included atomwise/nodewise property predictions such as forces. These forces are directly derived from the energies hence leading to force-field development (ALIGNN-FF). ALIGNN-FF was trained on the JARVIS-DFT dataset. Note that, we did not need to modify the ALIGNN model for surfaces because these GNN are based on the local environment around each atom only. We have used the same ALIGNN model in molecules[52] and metal–organic frameworks[106] also without changing any architecture and still leading to accurate results. Next, JARVIS-DFT is a collection of 80 000 diverse materials primarily using OptB88vdW in VASP software following strict protocols for convergence *etc.* In addition to the datasets, JARVIS-DFT is seamlessly integrated with the JARVIS-tools package for setting up calculations and performing analysis using a variety of multi-fidelity and multi-scale simulation approaches. ALIGNN-FF was trained on 307 811 bulk structures with 1 million forces obtained from SCF relaxation step for materials in the JARVIS-DFT. ALIGNN-FF was shown to capture both structural and chemical diversity with reasonable accuracy

especially for structure optimization. DFT calculations were carried out using the Vienna *Ab initio* Simulation Package (VASP) software[107,108] with OptB88vdW,[44] TBmBJ[45] and R2SCAN[46] functionals using the workflow given on our 'jarvis-tools' GitHub page (https://github.com/usnistgov/jarvis). We use the OptB88vdW functional, which gives accurate lattice parameters for both vdW and non-vdW (3D-bulk) solids. The crystal structure was optimized until the forces on the ions were less than 0.01 eV $Å^{-1}$ and energy less than 10^{-6} eV. Also, we calculate the local potential containing ionic plus Hartree contributions to determine the vacuum potential (VAC) of surface slabs. The VAC is subtracted from the valence band maxima (VBM) and conduction band minima (CBM) to enable the comparison of band-diagrams of individual slabs in band-alignment diagrams. The converged *k*-points and cut-off for the bulk materials were also used for the corresponding surface slab models. The ASJ based interface structures were generated using Zur algorithm[82] available in the JARVIS-Tools. For a quick scan of *xy*-displacements for surfaces in the interfaces, ALIGNN-FF was used.

Code availability

The code used in this work, InterMat is made publicly available at: https://github.com/usnistgov/intermat. It depends on closely related codes available at https://github.com/usnistgov/jarvis and https://github.com/usnistgov/alignn.

Data availability

The data generated by this work will be made publicly available at JARVIS websites: https://pages.nist.gov/jarvis/databases, https://jarvis.nist.gov/jarvisdft/ and Figshare (https://doi.org/10.6084/m9.figshare.25514719, https://doi.org/10.6084/m9.figshare.25832614). A webapp will also be made available at the JAVRIS-Heterostructure website (https://jarvis.nist.gov/jarvish/).

Author contributions

K. C. conceived the high-throughput workflow and conducted all calculations, K. F. G helped in setting up calculations and analysing the results. All authors reviewed the manuscript.

Conflicts of interest

The authors declare no competing interests.

Acknowledgements

We thank computational resource from National Institute of Standards and Technology (NIST). This work was performed with funding from the CHIPS Metrology Program, part of CHIPS for America, National Institute of Standards and Technology, U.S. Department of Commerce. Please note commercial software is identified to specify procedures. Such identification does not imply recommendation by National Institute of Standards and Technology (NIST).

References

1 K. T. Butler, G. Sai Gautam and P. Canepa, *npj Comput. Mater.*, 2019, **5**, 19.
2 A. P. Sutton, *Monographs on the Physics and Chemistry of Materials*, 1995, pp. 414–423.
3 H. Kroemer, *Proc. IEEE*, 1982, **70**, 13–25.
4 W. Monch, *Electronic properties of semiconductor interfaces*, Springer Science & Business Media, 2013, vol. 43.
5 T. Y. Edward, J. O. McCaldin and T. C. McGill, *Solid State Physics*, Elsevier, 1992, vol. 46, pp. 1–146.
6 J. Robertson, *J. Vac. Sci. Technol., A*, 2013, **31**, 050821.
7 M. Smeu and K. Leung, *Phys. Chem. Chem. Phys.*, 2021, **23**, 3214–3218.
8 G. Agostini and C. Lamberti, *Characterization of semiconductor heterostructures and nanostructures*, Elsevier, 2011.
9 Y. Taur, *IEEE Spectrum*, 1999, **36**, 25–29.
10 CHIPS.Gov—nist.gov, https://www.nist.gov/chips, accessed 13-11-2023.
11 C. G. Van de Walle and R. M. Martin, *J. Vac. Sci. Technol., B: Microelectron. Process. Phenom.*, 1985, **3**, 1256–1259.
12 A. Franciosi and C. G. Van de Walle, *Surf. Sci. Rep.*, 1996, **25**, 1–140.
13 L. Weston, H. Tailor, K. Krishnaswamy, L. Bjaalie and C. Van de Walle, *Comput. Mater. Sci.*, 2018, **151**, 174–180.
14 Y. Hinuma, A. Grüneis, G. Kresse and F. Oba, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **90**, 155405.
15 A. Ghosh, S. Jana, T. Rauch, F. Tran, M. A. Marques, S. Botti, L. A. Constantin, M. K. Niranjan and P. Samal, *J. Chem. Phys.*, 2022, **157**, 124108.
16 G. Di Liberto and G. Pacchioni, *J. Phys.: Condens. Matter*, 2021, **33**, 415002.
17 D. Dardzinski, M. Yu, S. Moayedpour and N. Marom, *J. Phys.: Condens. Matter*, 2022, **34**, 233002.
18 K. Mathew, A. K. Singh, J. J. Gabriel, K. Choudhary, S. B. Sinnott, A. V. Davydov, F. Tavazza and R. G. Hennig, *Comput. Mater. Sci.*, 2016, **122**, 183–190.
19 K. Choudhary, K. F. Garrity, S. T. Hartman, G. Pilania and F. Tavazza, *Phys. Rev. Mater.*, 2023, **7**, 014009.
20 P. Restuccia, G. Losi, O. Chehaimi, M. Marsili and M. C. Righi, *ACS Appl. Mater. Interfaces*, 2023, **15**, 19624–19633.
21 K. Choudhary, T. Liang, A. Chernatynskiy, S. R. Phillpot and S. B. Sinnott, *J. Phys.: Condens. Matter*, 2015, **27**, 305004.
22 J. Yu, S. B. Sinnott and S. R. Phillpot, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2007, **75**, 085311.
23 T.-R. Shan, B. D. Devine, T. W. Kemper, S. B. Sinnott, S. R. Phillpot, *et al.*, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2010, **81**, 125328.
24 K. R. Hahn, M. Puligheddu and L. Colombo, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2015, **91**, 195313.
25 W. Harrison and J. Tersoff, *J. Vac. Sci. Technol., B: Microelectron. Process. Phenom.*, 1986, **4**, 1068–1073.

26 N. Bernstein, M. J. Aziz and E. Kaxiras, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **58**, 4579.
27 G. K. Schenter and J. P. LaFemina, *J. Vac. Sci. Technol., A*, 1992, **10**, 2429–2435.
28 M. Munoz, V. Velasco and F. Garcia-Moliner, *Prog. Surf. Sci.*, 1987, **26**, 117–133.
29 D. Willhelm, N. Wilson, R. Arroyave, X. Qian, T. Cagin, R. Pachter and X. Qian, *ACS Appl. Mater. Interfaces*, 2022, **14**, 25907–25919.
30 Y. Huang, C. Yu, W. Chen, Y. Liu, C. Li, C. Niu, F. Wang and Y. Jia, *J. Mater. Chem. C*, 2019, **7**, 3238–3245.
31 Z. Zhu, B. Dong, H. Guo, T. Yang and Z. Zhang, *Chin. Phys. B*, 2020, **29**, 046101.
32 A. G. Milnes, *Heterojunctions and metal semiconductor junctions*, Elsevier, 2012.
33 J. Robertson, *Rep. Prog. Phys.*, 2005, **69**, 327.
34 B. Roul, M. Kumar, M. K. Rajpalke, T. N. Bhat and S. Krupanidhi, *J. Phys. D: Appl. Phys.*, 2015, **48**, 423001.
35 C. Ohler, C. Daniels, A. Förster and H. Lüth, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **58**, 7864.
36 W. Schottky, *Phys. Rev.*, 1926, **28**, 74.
37 J. Bardeen, *Phys. Rev.*, 1947, **71**, 717.
38 L. N. Oliveira and J. W. Wilkins, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1985, **32**, 696.
39 R. Anderson, *IBM J. Res. Dev.*, 1960, **4**, 283–287.
40 G. Losi, O. Chehaimi and M. C. Righi, *arXiv*, 2023, preprint, arXiv:2304.14367, DOI: **10.1021/acs.jctc.3c00459**.
41 S. Smidstrup, T. Markussen, P. Vancraeyveld, J. Wellendorff, J. Schneider, T. Gunst, B. Verstichel, D. Stradi, P. A. Khomyakov, U. G. Vej-Hansen, *et al.*, *J. Phys.: Condens. Matter*, 2019, **32**, 015901.
42 D. Wines, R. Gurunathan, K. F. Garrity, B. DeCost, A. J. Biacchi, F. Tavazza and K. Choudhary, *Appl. Phys. Rev.*, 2023, **10**, 041302.
43 K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, *et al.*, *npj Comput. Mater.*, 2020, **6**, 173.
44 J. Klimeš, D. R. Bowler and A. Michaelides, *J. Phys.: Condens. Matter*, 2009, **22**, 022201.
45 F. Tran and P. Blaha, *Phys. Rev. Lett.*, 2009, **102**, 226401.
46 J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew and J. Sun, *J. Phys. Chem. Lett.*, 2020, **11**, 8208–8215.
47 J. Heyd, G. E. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2003, **118**, 8207–8215.
48 K. Choudhary and K. Garrity, *npj Comput. Mater.*, 2022, **8**, 244.
49 D. Wines, K. Choudhary, A. J. Biacchi, K. F. Garrity and F. Tavazza, *Nano Lett.*, 2023, **23**, 969–978.
50 K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. Billinge, *et al.*, *npj Comput. Mater.*, 2022, **8**, 59.
51 R. K. Vasudevan, K. Choudhary, A. Mehta, R. Smith, G. Kusne, F. Tavazza, L. Vlcek, M. Ziatdinov, S. V. Kalinin and J. Hattrick-Simpers, *MRS Commun.*, 2019, **9**, 821–838.
52 K. Choudhary and B. DeCost, *npj Comput. Mater.*, 2021, **7**, 185.
53 K. Choudhary, B. DeCost, L. Major, K. Butler, J. Thiyagalingam and F. Tavazza, *Digital Discovery*, 2023, **2**, 346–355.
54 K. Choudhary, R. Gurunathan, B. DeCost and A. Biacchi, *J. Chem. Inf. Model.*, 2023, **63**, 1708–1722.
55 K. Choudhary and M. L. Kelley, *J. Phys. Chem. C*, 2023, **127**, 17545–17555.
56 K. Choudhary, *arXiv*, 2024, preprint, arXiv:2405.03680, DOI: **10.48550/arXiv.2405.03680**.
57 J. Dillon Jr and H. Farnsworth, *J. Appl. Phys.*, 1958, **29**, 1195–1202.
58 P. Bhattacharya, *Semiconductor optoelectronic devices*, Prentice-Hall, Inc., 1997.
59 C. Messmer and J. Bilello, *J. Appl. Phys.*, 1981, **52**, 4623–4629.
60 R. Jaccodine, *J. Electrochem. Soc.*, 1963, **110**, 524.
61 J. Hölzl and F. K. Schulte, *Solid Surface Physics*, 2006, pp. 1–150.
62 J. Field and C. Freeman, *Philos. Mag. A*, 1981, **43**, 595–618.
63 G. Gobeli and F. Allen, *Surf. Sci.*, 1964, **2**, 402–408.
64 S. Pouch, M. Amato, M. Bertocchi, S. Ossicini, N. Chevalier, T. Melin, J.-M. Hartmann, O. Renault, V. Delaye, D. Mariolle, *et al.*, *J. Phys. Chem. C*, 2015, **119**, 26776–26782.
65 J. Pelletier, D. Gervais and C. Pomot, *J. Appl. Phys.*, 1984, **55**, 994–1002.
66 W. Liu, W. Zheng and Q. Jiang, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2007, **75**, 235322.
67 C. Wu, A. Kahn, E. Hellman and D. Buchanan, *Appl. Phys. Lett.*, 1998, **73**, 1346–1348.
68 A. Rosa and J. Neugebauer, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2006, **73**, 205346.
69 S.-C. Lin, C.-T. Kuo, X. Liu, L.-Y. Liang, C.-H. Cheng, C.-H. Lin, S.-J. Tang, L.-Y. Chang, C.-H. Chen and S. Gwo, *Appl. Phys. Express*, 2012, **5**, 031003.
70 S. Lu, P. Shen, H. Zhang, G. Liu, B. Guo, Y. Cai, H. Chen, F. Xu, T. Zheng, F. Xu, *et al.*, *Nat. Commun.*, 2022, **13**, 3109.
71 A. Crovetto, J. M. Adamczyk, R. R. Schnepf, C. L. Perkins, H. Hempel, S. R. Bauers, E. S. Toberer, A. C. Tamboli, T. Unold and A. Zakutayev, *Adv. Mater. Interfaces*, 2022, **9**, 2200031.
72 I. Csik, S. P. Russo and P. Mulvaney, *Chem. Phys. Lett.*, 2005, **414**, 322–325.
73 M. Haase, H. Cheng, J. DePuydt and J. Potts, *J. Appl. Phys.*, 1990, **67**, 448–452.
74 K. Shen, X. Wang, Y. Zhang, H. Zhu, Z. Chen, C. Huang and Y. Mai, *Sol. Energy*, 2020, **201**, 55–62.
75 E. Clark, R. Yeske and H. Birnbaum, *Metall. Trans. A*, 1980, **11**, 1903–1908.
76 N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Cengage Learning, 2022.
77 R. Tran, Z. Xu, B. Radhakrishnan, D. Winston, W. Sun, K. A. Persson and S. P. Ong, *Sci. Data*, 2016, **3**, 1–13.
78 K. Choudhary and F. Tavazza, *Comput. Mater. Sci.*, 2019, **161**, 300–308.
79 K. Choudhary, Q. Zhang, A. C. Reid, S. Chowdhury, N. Van Nguyen, Z. Trautt, M. W. Newrock, F. Y. Congo and F. Tavazza, *Sci. Data*, 2018, **5**, 1–12.

80 S. De Waele, K. Lejaeghere, M. Sluydts and S. Cottenier, *Phys. Rev. B*, 2016, **94**, 235418.
81 J. C. Conesa, *Nanomaterials*, 2021, **11**, 317–330.
82 A. Zur and T. McGill, *J. Appl. Phys.*, 1984, **55**, 378–386.
83 P. Goodhew and K. Giannakopoulos, *Micron*, 1999, **30**, 59–64.
84 O. Romanyuk, O. Supplie, T. Susi, M. May and T. Hannappel, *Phys. Rev. B*, 2016, **94**, 155309.
85 R. List, J. Woicik, I. Lindau and W. Spicer, *J. Vac. Sci. Technol., B: Microelectron. Process. Phenom.*, 1987, **5**, 1279–1283.
86 M. Kundu, S. Mahamuni, S. Gokhale and S. Kulkarni, *Appl. Surf. Sci.*, 1993, **68**, 95–102.
87 J. Batey and S. Wright, *J. Appl. Phys.*, 1986, **59**, 200–209.
88 D. V. Talapin, R. Koeppe, S. Götzinger, A. Kornowski, J. M. Lupton, A. L. Rogach, O. Benson, J. Feldmann and H. Weller, *Nano Lett.*, 2003, **3**, 1677–1681.
89 J. Waldrop, R. Grant and E. Kraut, *Appl. Phys. Lett.*, 1989, **54**, 1878–1880.
90 G. Schwartz, G. Gualtieri, R. Feldman, R. Austin and R. Nuzzo, *J. Vac. Sci. Technol., B: Microelectron. Process. Phenom.*, 1990, **8**, 747–750.
91 E. Yu, M. Phillips, J. McCaldin and T. McGill, *J. Vac. Sci. Technol., B: Microelectron. Nanometer Struct.–Process., Meas., Phenom.*, 1991, **9**, 2233–2237.
92 J. Arriaga, G. Armelles, M. Muoz, J. Rodrguez, P. Castrillo, M. Recio, V. Velasco, F. Briones and F. Garca-Moliner, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1991, **43**, 2050.
93 A. Nakagawa, H. Kroemer and J. H. English, *Appl. Phys. Lett.*, 1989, **54**, 1893–1895.
94 H. Lange and D. F. Kelley, *J. Phys. Chem. C*, 2020, **124**, 22839–22844.
95 S. Rubini, E. Milocco, L. Sorba and A. Franciosi, *J. Cryst. Growth*, 1998, **184**, 178–182.
96 S. P. Kowalczyk, E. Kraut, J. Waldrop and R. Grant, *J. Vac. Sci. Technol.*, 1982, **21**, 482–485.
97 L. Lew Yan Voon, L. Ram-Mohan and R. Soref, *Appl. Phys. Lett.*, 1997, **70**, 1837–1839.
98 G. Dufour, F. Rochet, F. Stedile, C. Poncey, M. De Crescenzi, R. Gunnella and M. Froment, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1997, **56**, 4266.
99 A. Rizzi, R. Lantier, F. Monti, H. Lüth, F. D. Sala, A. Di Carlo and P. Lugli, *J. Vac. Sci. Technol., B: Microelectron. Nanometer Struct.–Process., Meas., Phenom.*, 1999, **17**, 1674–1681.
100 S. W. King, R. J. Nemanich and R. F. Davis, *J. Appl. Phys.*, 2015, **118**, 045304.
101 L. Sang, Q. S. Zhu, S. Y. Yang, G. P. Liu, H. J. Li, H. Y. Wei, C. M. Jiao, S. M. Liu, Z. G. Wang, X. W. Zhou, *et al.*, *Nanoscale Res. Lett.*, 2014, **9**, 1–5.
102 J. Waldrop and R. Grant, *Appl. Phys. Lett.*, 1996, **68**, 2879–2881.
103 J. Liu, A. Kobayashi, S. Toyoda, H. Kamada, A. Kikuchi, J. Ohta, H. Fujioka, H. Kumigashira and M. Oshima, *Phys. Status Solidi B*, 2011, **248**, 956–959.
104 K. Choudhary, D. Wines, K. Li, K. F. Garrity, V. Gupta, A. H. Romero, J. T. Krogel, K. Saritas, A. Fuhr, P. Ganesh, *et al.*, *arXiv*, 2023, preprint, arXiv:2306.11688, DOI: **10.1038/s41524-024-01259-w**.
105 K. Choudhary and B. G. Sumpter, *AIP Adv.*, 2023, **13**, 095109.
106 K. Choudhary, T. Yildirim, D. W. Siderius, A. G. Kusne, A. McDannald and D. L. Ortiz-Montalvo, *Comput. Mater. Sci.*, 2022, **210**, 111388.
107 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169.
108 G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.