"This is the peer reviewed version of the following article: [**ADVANCED THEORY AND SIMULATIONS, 2019, 2 (1), pp. ? - ? (11)], which has been published in final form at [*https://onlinelibrary.wiley.com/doi/full/10.1002/adts.201800128*]. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving."***

DOI: 10.1002/ ((please add manuscript number))
Article type: Full Paper

Efficient prediction of structural and electronic properties of hybrid 2D materials using complementary DFT and machine learning approaches

Sherif Abdulkader Tawik#, Olexandr Isayev, Catherine Stampfl, Joe Shapter, David A. Winkler and Michael J. Ford

Dr. Sherif Abdulkader Tawfik, Prof. Michael J. Ford

School of Mathematical and Physical Sciences, University of Technology Sydney, Ultimo, New South Wales 2007, Australia

E-mail: sherif.tawfic@gmail.com, mike.ford@uts.edu.au

Prof. Olexandr Isayev

Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, USA

E-mail: olexandr@olexandrisayev.com

Prof. Catherine Stampfl

School of Physics, The University of Sydney, New South Wales, 2006, Australia

Prof. Joe Shapter

Australian Institute for Bioengineering and Nanotechnology, University of Queensland, St. Lucia, Queensland, 4072, Australia

School of Chemical and Physical Sciences, Flinders University, Bedford Park, South Australia 5042, Australia

E-mail: j.shapter@uq.edu.au

Prof. David A. Winkler

Manufacturing, Commonwealth Scientific and Industrial Research Organization, Bag 10,
Clayton South MDC, Victoria 3169, Australia

Monash Institute of Pharmaceutical Sciences, Monash University, 381 Royal Parade,
Parkville, Victoria 3052, Australia

Latrobe Institute for Molecular Science, La Trobe University, Kingsbury Drive, Bundoora,
Victoria 3086, Australia

E-mail: d.winkler@latrobe.edu.au

#Current address: School of Science, RMIT University, Melbourne, 3001, Australia

Keywords: machine learning, van der Waals materials

There are now, in principle, a limitless number of hybrid van der Waals heterostructures that can be built from the rapidly growing number of two-dimensional layers. The key question is how to explore this vast parameter space in a practical way. Computational methods can guide experimental work. However, even the most efficient electronic structure methods such as density functional theory, are too time consuming to explore more than a tiny fraction of all possible hybrid 2D materials. Here we demonstrate that a combination of DFT and machine learning techniques provide a practical method for exploring this parameter space much more efficiently than by DFT or experiment. As a proof of concept, we applied this methodology to predict the interlayer distance and band gap of bilayer heterostructures. Our methods quickly and accurately predicted these important properties for a large number of hybrid 2D materials. This work paves the way for rapid computational screening of the vast parameter space of van der Waals heterostructures to identify new hybrid materials with useful and interesting properties.

1. Introduction

Materials constructed from the large number of 2D materials now identified[1] offer enormous possibilities for fundamental research, promising improved or even novel electronic or optical technologies.[2] Recently, there has been increased interest in the LEGO-like creation of few-layer hybrid 2D heterostructures[3] for photovoltaic and photonic applications.[4,5,6,7,8] For example hybrid 2D heterostructures, in which the charge carriers move in the direction perpendicular to the plane of the 2D layers, have recently attracted attention as potential vertical p-n junctions. The 2D materials that have been investigated as p and n dopants include MoS_2, $MoSe_2$, WSe_2. With the growing number of semiconducting van der Waals (vdW) materials, the number of possible hybrid bilayers (that is, heterostructure bilayers formed from two different monolayer species) that achieve p-n band alignment is increasing, and the prediction of the band gap of these bilayers using ML models would greatly support the search for new atomically thin p-n junction materials for optoelectronics applications.

Bulk materials constructed from vdW 2D heterostructures offer an even larger range of properties than homostructures as they can be arranged in multiple layers consisting of different 2D materials in each layer. They could provide almost infinitely tuneable bespoke materials for almost any application. For example, a recent data mining study[1] reported the existence of 1,825 2D materials that could be exfoliated from known experimental inorganic compounds. This set can make ~1.7 million bilayer structures, around 10^9 trilayers. This quickly becomes an intractable problem for accurate but CPU-intensive computational methods like DFT to explore. Clearly, a method for rapidly predicting the properties of these structures without having to perform many expensive and time-consuming quantum chemical calculations is needed to efficient explore the properties these materials spaces offer. We addressed this by applying several ML models, trained on data from a small number of DFT calculations, to predict structural and electronic properties of layered vdW materials, specifically bilayer materials constructed from different 2D materials. The success of the ML approach using a small training set could potentially save a significant amount of experimental and computational time and cost, while retaining acceptable accuracy.

In 2D materials, the interlayer vdW forces are essential to maintain the equilibrium structure. The key structural quantity that indicates the strength of the vdW force in these materials is the interlayer distance, and its related quantity, the layer binding energy. For DFT to accurately predict these two quantities, it needs to be corrected by incorporating a vdW correlation potential to the DFT correlation potential.[9,10] To this end, various forms of the vdW correlation have been proposed and applied to 2D materials, such as the Tkatchenko-Scheffler (TS) method[11,12] and the SCAN+rVV10 method.[13], many of which displayed impressive accuracy. DFT databases, such as AFLOW and MaterialsProject, contain structures calculated using non-vdW corrected functionals. These non-dispersive correlation potentials can result in overestimates of the interlayer spacing, for example, MoS_2-WS_2 in which $c = 22.37$ Å.[14] The current work uses a dispersion-supported DFT method with the aim of constructing a data set of hybrid 2D vdW materials with realistic interlayer distances.

However, a critical problem in the application of DFT to hybrid vdW structures is creating lattice-matching interfaces between noncommensurate 2D materials. Using DFT to study hybrid 2D materials requires the supercell describing the interface between two materials (whether parallel or perpendicular to the plane of the 2D materials) to have commensurate supercells for the two materials. One commonly adopted solution to this problem is searching for supercells that minimise the strain in each of the incommensurate monolayers. This is a non-trivial problem that often requires strains of a few percent to keep the size of the supercell reasonable. The use of ML largely alleviates this problem since DFT calculations are only performed on a small subset of bilayers to generate the training and test sets for the ML models.

Here, we describe a proof of concept study showing how parsimonious use of carefully corrected DFT calculations of the properties of hybrid bilayer vdW structures can be used to train ML models that predict the interlayer distance and the band gap of a larger set of hybrid materials with reasonable accuracy.

2. Computational details

2.1. The first principles approach

2D structures were selected from the large collection of 2D materials in the "2D atlas".[15] All of the 2D hetero-structures in this work are bilayers consisting of a combination of any two of the 2D monolayers in our data set. A 12 Å vacuum gap is used to separate periodic images of the bilayer and to minimize interaction between these repeated images. We used VASP[16] to calculate atomic and electronic structures using the generalized gradient approximation based on the PBE parameterization.[17] We accounted for the vdW interaction by adding the Tkatchenko-Scheffler vdW correlation potential.[12] We applied a **k**-point space of 8 × 8 × 1 for unit cells, and 3 × 3 × 1 for supercells, and an energy cut-off of 400 eV. For both the unit cells and for the smallest supercells, that is a 2x2 supercell, total energies are converged to better than 1 meV with the above k-point meshes. The energy minimization tolerance is 10^{-6} eV, and the force tolerance is 10^{-2} eV/Å. For the 267 bilayers in the data set, we calculate the interlayer distance *d* as the distance separating the two layers (that is, the minimum distance between the two layers), and the band gap. Bilayers formed from one or more monolayers with zero bandgap also have a zero bandgap and are excluded from the data set used to train and test the ML algorithms. Note that the application of VASP for the geometries provided in the "2D atlas" by Miro *et al.* [15] induces a small strain on the individual layers. However, the presence of small planar strain has a small effect on *d* and the band gap.[18]

For our training set, we perform DFT calculations for the bilayers assembled from 53 monolayer structures, as shown in Table 1. In selecting these monolayers, we focused on structures that satisfy two criteria: (1) possess trigonal symmetry, and (2) do not suffer from lattice distortions arising from covalent interaction with the adjacent layers. For example, we remove the CdX and ZnX monolayers (X= S, Se, Te) because of the significant layer distortions they exhibit when stacked with other layered materials.

2.2. The bilayer data set

For two different 2D materials, their unit cells are generally different (that is, they have different values for the in-plane lattice parameters). Provided the ratio of lattice constants is not irrational the bilayer cell can, in principle be constructed by a suitable supercell, although in many cases this would be computationally prohibitive and the monolayers must be strained. We searched over all possible combinations of the 53 monolayers to find instances where the bilayer cell could be constructed from a 5x5 or smaller supercell of either monolayer with less than 2 % strain in either monolayer.[18] The number of bilayers we choose from this set is 267. These bilayers will be used for the prediction of the interlayer distance.

For the prediction of the band gap, we took into consideration that 33 out of the 53 monolayers are metallic (zero band gap) and excluded them from our data set. Within the set of 267 bilayers selected above, the number of bilayers formed from these 20 non-metallic monolayers is 49.

Table 1: The 53 2D monolayers used in the study.

BN	$NbSe_2$	$NiTe_2$	Silicene	TaS_2	TiS_2
CdO	$NbTe_2$	PdS_2	SiC	$TaSe_2$	$TiSe_2$
GaS	NiS_2	$PdSe_2$	†1T-HfS_2	$TaTe_2$	$TiTe_2$
GaSe	$NiSe_2$	$PdTe_2$	1T-$HfSe_2$	1T-NbS_2	WCl_2
Graphene	InSe	PtS_2	1T-$HfTe_2$	1T-$NbSe_2$	WS_2
HfS_2	MoS_2	$PtSe_2$	1T-MoS_2	1T-$NbTe_2$	WSe_2
$HfSe_2$	$MoSe_2$	$PtTe_2$	1T-$MoSe_2$	1T-ReS_2	WTe_2
$HfTe_2$	$MoTe_2$	ReS_2	1T-$MoTe_2$	1T-WS_2	ZnO
InS	NbS_2	$ReSe_2$	$ReTe_2$	$TaCl_2$	

†1T prefix denotes the 1T polymorph of transition metal dichalcogenides (TMDCs). TMDCs without this prefix are of the 2H polymorph.

The application of DFT methods with vdW correlation correction to layered materials has been demonstrated to yield accurate results for the interlayer distances and the binding energies.[9,19] We use the TS method which accurately predicts the values of interlayer distances compared with the available experimental values as accurately as the benchmark Random Phase Approximation (RPA) method.[19]

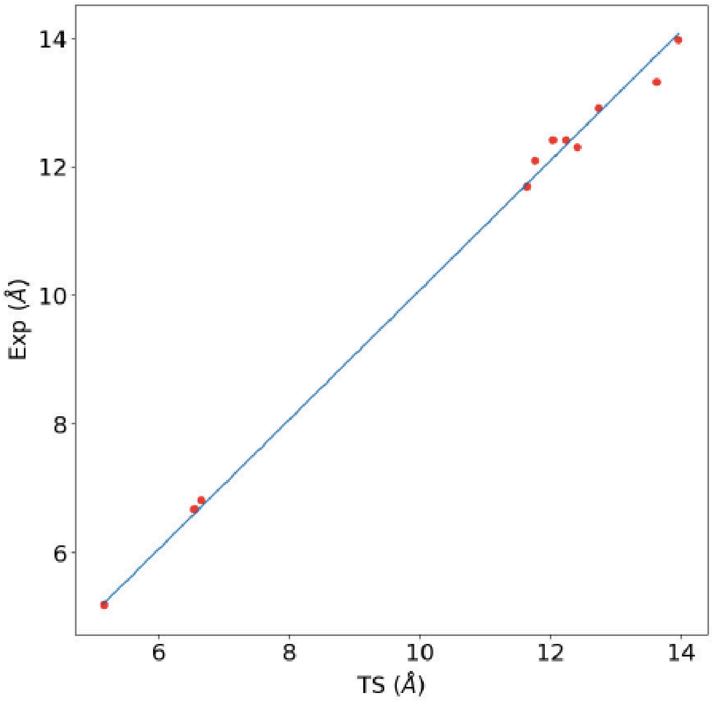

Figure 1: Comparison between the *c* lattice parameter for 11 2D materials[19] predicted using the Tkatchenko-Scheffler (TS) vdW functional and experimental values.[9]

Figure 1 compares the values of the *c* lattice parameter for the 11 2D materials predicted using the TS vdW correlation RPA,[19] and the experimental values.[9] The DFT method is clearly able to accurately recapitulate the experimental lattice parameters. These values are for pristine bulk materials, not hybrid materials, however. The choice of the best method for hybrid multilayered materials is still an open problem, and is currently the subject of an ongoing theoretical investigation.

To perform the DFT calculations on bilayers, the optimal stacking configuration for the bilayer must be found. There are two categories of stacked bilayers: those where the simulation cell is constructed from the unit cells of the two monolayers; and those where the cell needs to be constructed from a larger supercell of each monolayer. For the former, we obtain the stacking equilibria by performing a geometry relaxation for three stacking configurations AA, AB and AB'. These configurations are displayed in Figure 2. The structure with lowest energy is then taken as the equilibrium stacking configuration. For the incommensurate unit cells such as the boron nitride|silicon carbide bilayer (formed from 5 × 5 boron nitride unit cells and 4 × 4 silicon carbide unit cells), sliding one monolayer over the other in such large bilayers does not significantly affect the binding energy. Hence, we do not search for equilibrium stacking configurations in these bilayers.

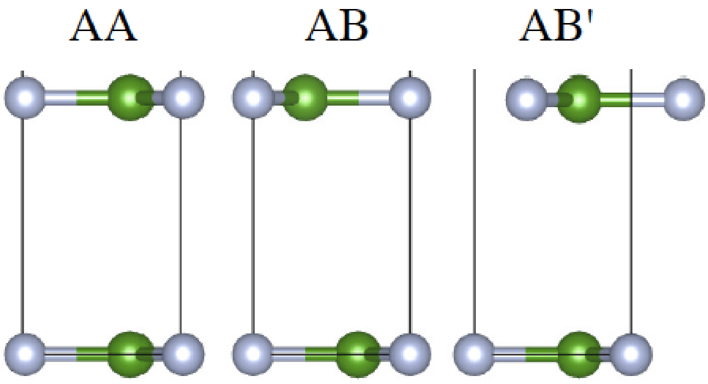

Figure 2: The three stacking configurations, AA, AB and AB' used for obtaining the lowest energy configuration. For boron nitride, we also added the AA' stacking sequene, which is similar to AA but has a nitrogen atom on one layer faced by a boron on the other.

Figure 3 summarizes the number of bilayers in which a particular monolayer is a component. The figure identified monolayers that are over- or under-represented in the data set. Here, the InSe monolayer is only present as the bilayer InSe|InSe, and $MoTe_2$ is the most connected (exists in 27 bilayers).

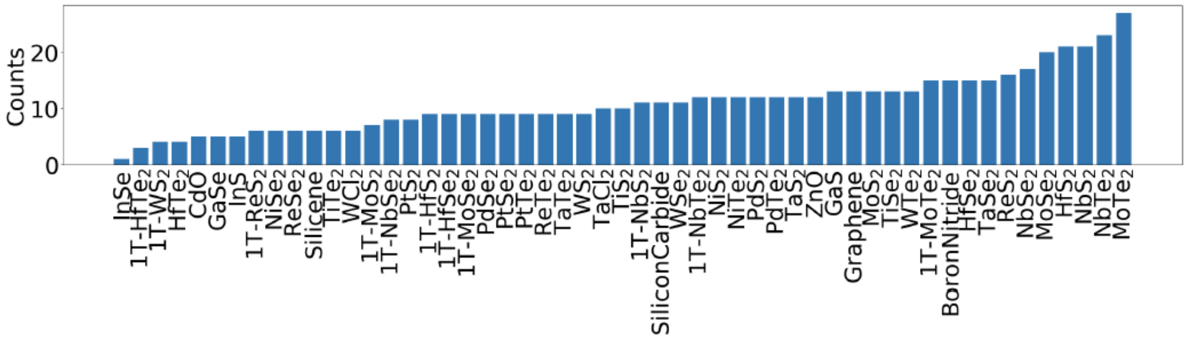

Figure 3: Representation of monolayers within the bilayer data set. The y-axis gives the number of bilayers in which each monolayer (along the *x*-axis) is a component.

A useful method for visualizing such high dimensionality data is the t-distributed Stochastic Neighbour Embedding (t-SNE)[20] algorithm that generates a 2D plot that clusters the data into groups labelled by values in the output vector. t-SNE is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It maps multi-dimensional data to two dimensions suitable for human observation. Figure 4 shows a two-dimensional t-SNE projection mapping of chemical space of 267 hybrid 2D materials. Due to the selection of large number elements as building blocks, materials distribute in chemical space without forming well-defined clusters. However, we identified three broad regions with specific character: sulphides, tellurides, and CNSi (graphene, carbides, nitrides, and silicenes). On average sulphide materials have a low interlayer distance of 2.1−2.7 Å. Tellurides typically have the highest at 3.2−3.8 Å. Two extreme materials PtS_2-PtS_2 and $PdTe_2$-WTe_2 are marked accordingly.

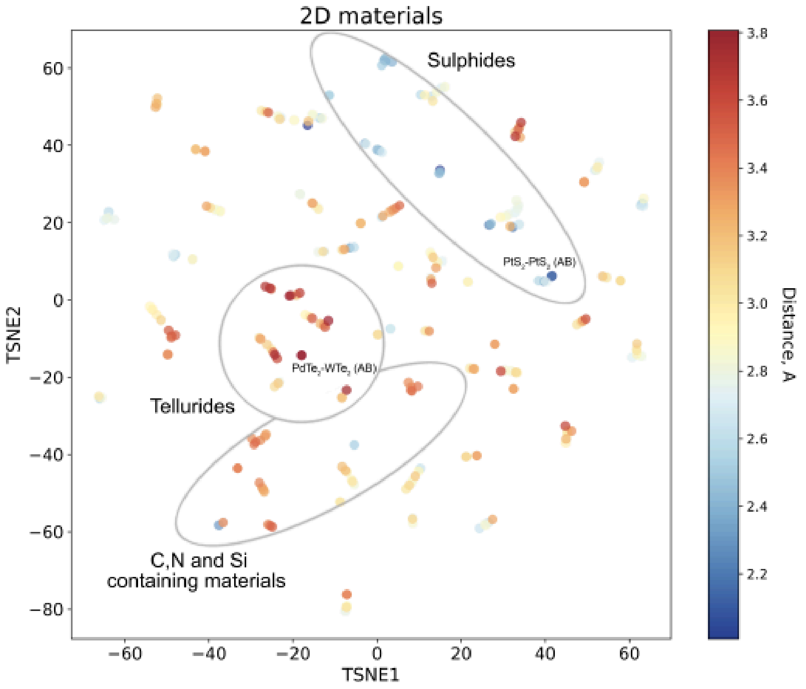

Figure 4: The t-distributed Stochastic Neighbour Embedding (t-SNE) plot for the data set of 267 bilayer structures based on the value of their interlayer distance.

2.3. The descriptor vector

The key to implementing successful machine learning model are descriptors, relevant mathematical representations of the structure and properties of materials. Fragment-based descriptors have demonstrated superior performance for ML models of molecules[21] and crystals.[22] In this work we adopt the method of Isayev *et al.*, [22] the Property-Labelled Materials Fragments (PLMF), modified for 2D materials, composed of 1529 descriptors. In the PLMF approach a crystal structure is represented as a graph, with vertices labelled according to the reference properties of the atoms they represent and nodes are connecting topological neighbours using Voronoi tessellation. The adjacency matrix of this graph determines the global topology for a given system, including interatomic bonds and contacts within a crystal. The final descriptor vector used to train the Machine Learning (ML) model is obtained by partitioning a full graph into smaller subgraphs, which we call fragments by the analogy with fragment-based descriptors in cheminformatics. Every fragment starts from a node (an atom and its properties) and captures a path in the graph through a collection of bonded atoms.[22]

Given the monolayer descriptors, the next critical problem is how to represent each bilayer using the monolayer data. As the interaction between the monolayers in a bilayer is dispersive, it does not affect the structure of either of the constituting monolayers. Therefore, it is possible to use the monolayer descriptors to describe the monolayers in the bilayer. The next issue is how to construct the bilayer descriptor vector. An intuitive choice is to create a larger descriptor vector composed of PLMF vector descriptors for the two monolayers. The problem in this approach is that it is sensitive to the swapping of the bilayers; that is, the descriptor vector for the bilayer A-B, made from monolayers A and B, will be a different vector from that of the bilayer B-A, even though the two bilayers are physically identical. We addressed this problem by generating the descriptor vectors in different two ways.

Bilayer representation 1 (BR1). For each bilayer A-B we created two data records. One record has layer A descriptors concatenated to layer B descriptors and the other has layer B descriptors concatenated to layer A descriptors. This method of representing multiple representations of the same data item has been previously applied to organic molecules.[23] Using this representation for the prediction of d generates a data set with $482 = 267 \times 2 - 52$ records (52 records are subtracted, instead of 53, because we have removed the $PdS_2|PdS_2$ bilayer from the data set due to the appearance of covalent interaction between the S atoms of the adjacent layers). For the prediction of the band gap it generates $78 = 49 \times 2 - 20$ records.

Bilayer representation 2 (BR2). Instead of creating a descriptors vector that is double the size of the monolayer descriptor vector, we add the values of the descriptors in both monolayers, effectively averaging them. This method is intrinsically invariant to changes in the order of the monolayers and can be applied to supercells with more than two layers. This generates a data set of 267 records for the prediction of *d* and 49 records for the prediction of the band gap. Due to the small number of records in BR2 for the band gap prediction, we do not perform machine learning on this set.

With 1529 elements in a descriptor vector it is crucial to use a dimensionality reduction algorithm to avoid overfitting the models. We applied the least absolute shrinkage and selection operator (LASSO) algorithm to our datasets.[24] Since LASSO is a supervised dimensionality reduction algorithm it cannot be applied to the set of monolayer descriptors. We used LASSO on the BR2 descriptors in the bilayer data set for interlayer distance *d* model. To obtain the optimal number of descriptors we varied the value of the sparsity parameter, and used the value of α and the sparse set of descriptors that yielded the highest R^2 value in the LASSO models. We then used these descriptors to predict properties using several ML methods (discussed below).

2.4. The machine learning models

As the relationships between the bilayer descriptors and the interlayer distance and band gap are potentially nonlinear we used four ML algorithms to contract the property models.

Feedforward Neural Network (NN)[25] These can generate nonlinear relationships between input and output variables and adaptively learn highly complex relationships. The input layer receives the descriptor vector, the hidden layers composed of a number of neurons perform nonlinear computation, and the output layer generate the response variable. We used a fully connected network where each neuron in a layer is connected to all neurons in the previous layer. Each neuron operates on a weighted sum of the data it receives from the elements of the previous layer using an activation function i.e.

$$p_j(t) = \sum_i o_i(t) w_{ij}$$

where w_{ij} is the weight connecting neurons *i* and *j*, and $o_i(t)$ is the output of neuron *i*. The values of w_{ij} are updated by a backpropagation learning algorithm.

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} + \xi(t)$$

where η is the learning rate, *C* is the loss function, $\xi(t)$ is a stochastic term, and *t* is the propagation step. There are multiple choices available for the activation function associated with each neuron. We used the logistic sigmoid which is given by $a(z) = \frac{1}{1+e^{-z}}$, where *z* is the quantity received by the neuron. We use the Keras[26] python platform to implement the NN model. The network we have used for BR1 representation has 35 × 2 = 70 input nodes, 5 nodes in a single hidden layer, and one output node. In the BR2 representation, the input layer has 35 input nodes. The sigmoid activation function is used in the hidden layer, while the linear activation function is used for the input and output layers. The learning rate was 0.03.

Support Vector Machine (SVM).[27,28] This is a supervised learning method first introduced for classification models[27] and then modified for regression problems. [28] The SVM classifier performs the classification of the data set from selected subsets of samples, called support vectors, in which the characteristic information on class distinction is compressed. In the linear support vector regression problem which we utilize in the present work, the aim is to find the linear function $f(\mathbf{x}) = \mathbf{wx} + \mathbf{b}$ that approximates the output vector $\mathbf{y}$ with weights vector $\mathbf{W}$such that the *primal function J* is minimized, which is given by

$$J = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}(\eta_n + \beta_n^*),$$

with the constraints $\left|y_n - (\beta x + b)\right| < \varepsilon + \eta_n$ and $\left|(\beta x + b) - y_n\right| < \varepsilon + \eta_n^*$ for some small ε and positive variables η_n and η_n^* (known as *slack variables*), for all *n*, where *n* is the data record, and *N* is the total number of records. Both *C* and ε are input parameters to the model.

Relevance Vector Machine (RVM).[29,30] This is a sparse version of the support vector machine which attempts to amend several of the shortcomings of SVM, [30] such as non-probabilistic predictions, low sparsity causing a tendency to overfit data, and the presence of the two fitting parameters *C* and ε which require cross validation. The RVM increases the sparsity of SVM and introduces a probabilistic weighting of the model weights based on Bayes' rule, assuming a Gaussian distribution of weights.

Random Forest (RF).[31] This is an ensemble learning method for classification, regression, and other tasks that constructs an ensemble of decision trees from the training data and outputs a class membership that is the mean prediction of the individual trees. The training in RFs is based on the feature aggregating[32] method. Given a training set **x** with output **y**, bagging repeatedly selects a random sample from **x** and **y** with replacement of the training set and fits decision trees to these samples. Once the training is complete, the prediction function operates by averaging the predictions from all the individual regression trees. The number of trees and the maximum tree depth are input parameters to the model.

The objective of ML methods is to build accurate prediction models. The quality of the model is determined by the ability of the model to predict the properties of new materials that the model has never encountered; – that is, how accurately can the model generalize to new outcomes based on its learning. This can be measured by splitting the data set into two parts: the training set used to build the ML model, and the test set used to test the quality of the model. The accuracy of the prediction is best judged by loss functions or measures of dispersion. We use the following statistical measures to assess the accuracy of the training:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{N}(Y_i - \widehat{Y_\iota})^2,$$

$$\mathrm{MARE} = \frac{1}{n}\sum_{i=1}^{N}\frac{|Y_i - \widehat{Y_\iota}|}{Y_i} \times 100,$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \widehat{Y_\iota})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2},$$

where MSE is the mean square error, MARE is the mean absolute relative error (%), R^2 is the coefficient of determination, Y_i are the original test set outcomes (in our case, the DFT-calculated interlayer distances for the bilayers), $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ are the predicted test set outcomes, and $\bar{\boldsymbol{Y}}$ is the average of the original test set outcomes. The significance of R^2 is that it expresses the proportion of the variance in $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$ that can be predicted from the descriptor vector, and is an important measure of the ML model quality. However, its values are dependent on the size of the data set, and therefore we adopt the three quantities together, R^2, MARE and MSE, to gauge the accuracy. For the case of the band gap prediction, we do not use the MARE because some of the values obtained are zero. Thus, for each of the four models in this work, we train the model on 80% of the data set and use the remaining 20% as the test set. We construct the test set by applying the k-means clustering[33] to the data set (the set of bilayers in BR2), which yields 6 clusters by the Silhouettes analysis.[34] Then, we randomly choose 20% of each cluster and build our test set. In the NN model, the learning iteration stops when the MSE is below 0.03 (for BR1) and 0.05 (for BR2). Then we compare the accuracy of the models based on the R^2, the MSE and MARE values.

Each of the four models involved in this work requires the tuning of a number of parameters to ensure optimal performance. We tune the NN parameters manually, and we use the GridSearch algorithm provided by the Python scikit-learn library (GridSearchCV) to tune the parameters of the SVM and RF models. GridSearchCV calculates the best parameter combination by performing a cross-validated grid-search over a parameter grid. The parameters which we optimize for the SVM are: the C and γ, while those of the RF are: number of estimators and maximum depth.

3. Results and Discussion

3.1. Prediction of the interlayer distance

Using the LASSO algorithm, the optimal number of descriptors for predicting d is 35 per bilayer. We summarize these descriptors in the Supporting Information. We summarize the MSE, MARE and R^2 values for the predictions of each ML model in Tables 2 for both bilayer representations BR1 and BR2. Comparing the MSE and MARE values of the test sets, it is clear that all ML models exhibit very similar accuracy in predicting the properties of materials in the test set not used to train the models. The MSE and MARE are similar for most models and for both classes of descriptors, BR1 and BR2. The RVM model using the BR1 descriptors is slightly less accurate in predicting the properties of the test set than the other three models. The SVM model using BR2 descriptors gives the lowest prediction accuracy, probably because it was overfitted. The NN model makes slightly better predictions, although the differences between all models are small. The relatively large difference between the training and test set accuracies suggest that the SVM model is overfitted, a phenomenon that has been noted by others.[30]

Table 2: The R^2, mean square error (MSE), in Å^2, and mean absolute relative error (MARE) (%) for each of the four ML models applied to the BR1 and BR2 bilayer representation.

Descriptors		R^2		MSE		MARE %	
		Train	Test	Train	Test	Train	Test
BR1	RF	0.96	0.82	0.005	0.024	1.8	3.9
	SVM	0.90	0.83	0.012	0.023	1.8	4.1
	RVM	0.87	0.79	0.017	0.028	3.4	4.5
	NN	0.84	0.82	0.020	0.025	3.5	4.0
BR2	RF	0.73	0.83	0.035	0.021	4.5	3.7
	SVM	0.99	0.67	0.001	0.041	0.6	5.0
	RVM	0.84	0.73	0.021	0.034	3.8	4.7
	NN	0.88	0.90	0.016	0.012	0.7	2.7

Figure 5 Correlation between the predicted and the DFT-calculated interlayer spacing (*d)* for the bilayers in the test set using the four ML models trained on BR1 descriptors.

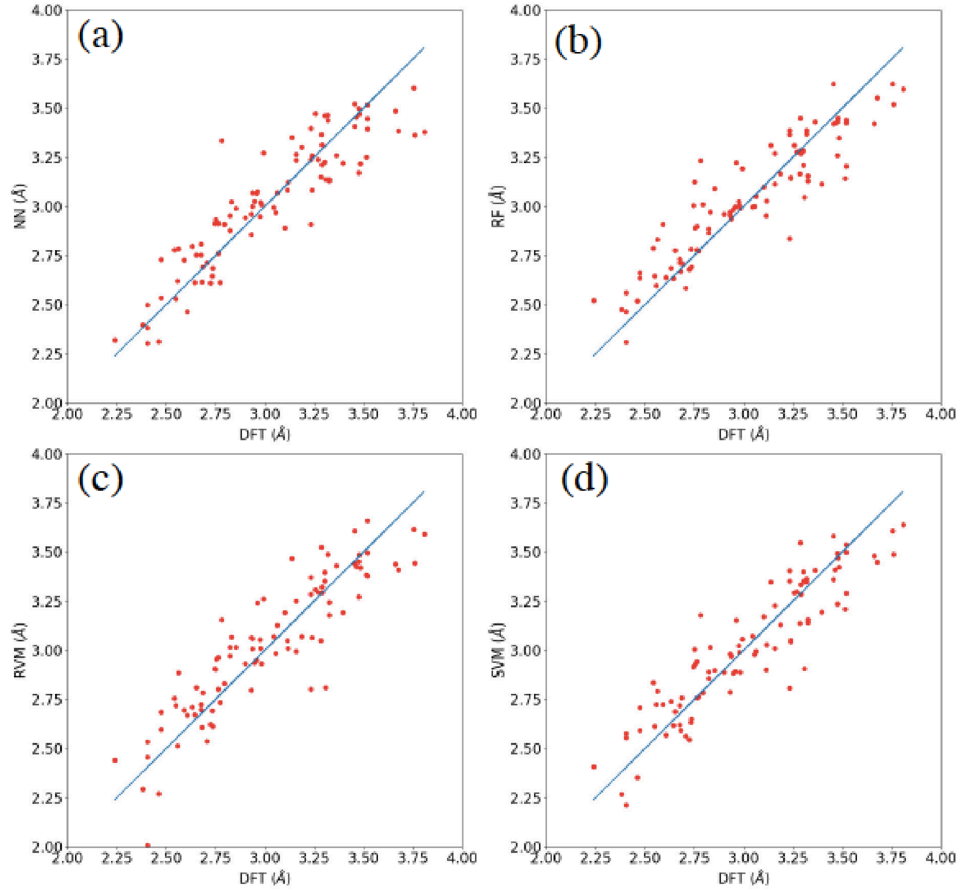

Figure 5: Comparison between the interlayer distances in the bilayers obtained using DFT, and those using (a) NN, (b) RF, (c) RVM and (d) SVM.

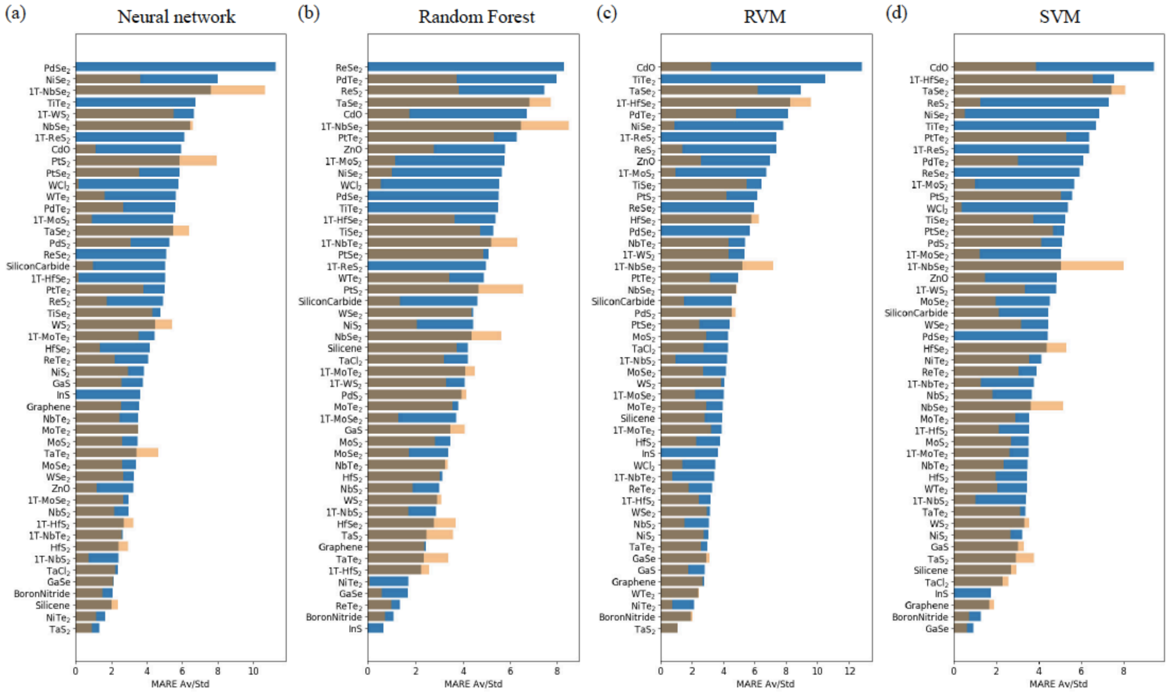

Figure 6: For each monolayer (vertical axis), the average (blue) and standard deviation (orange) of the MARE for the interlayer spacing for all bilayers containing that particular monolayer. The four panels represent the different machine learning algorithms used to predict the interlayer spacing: (a) NN, (b) RF, (c) SVM and (d) RVM. All MARE values are percentages.

An interesting feature in Figure 6 is that the presence of some component monolayers in bilayers leads to high *d* prediction errors, irrespective of the ML model used. For example, CdO and $TiTe_2$ result in the largest prediction errors in all four ML models. Similarly, ReS_2 and $PdTe_2$ are common higher prediction error components in the RF, SVM and RVM models. In addition, in all four models, the non-metallic monolayers graphene, boron nitride and silicon carbide have relatively low prediction errors. GaSe, TaS_2, and boron nitride also have some of the lowest prediction errors in the four models. The distribution of monolayers within the bilayers shown in Figure 3 may contribute to the accuracy of the bilayer properties predictions (such as in the case of under-represented or over-represented monolayers). However, the trends of errors displayed in Figure 6 show that the accuracy of the models is relatively independent these factors.

We used the four ML models, trained using the BR1 descriptors, to predict the interlayer distances of all 1431 possible bilayers (Table 3). Clearly the mean and standard deviation of the interspacing distance are the same for all models. The smallest minimum *d* is predicted by the RVM model (1.753 Å). The SVM gives the most accurate prediction for pristine bilayers (2.0%), followed by RF (2.9%).

Table 3: The summary statistics of the interlayer distances predicted for all 1431 bilayers constructed from the 53 monolayers (values in Å). The last row gives the MARE for

predicting interlayer spacing in bilayers made from identical monolayers (pristine bilayers). Note that virtually all the differences between the four ML models are smaller than the standard deviation.

	NN	RF	RVM	SVM
Mean	3.00	3.03	3.04	3.02
Standard deviation	0.28	0.24	0.27	0.26
Minimum	2.10	2.12	1.75	2.02
Maximum	3.68	3.72	3.80	3.74
5% percentile	2.54	2.62	2.58	2.59
50% percentile	3.02	3.02	3.04	3.02
95% percentile	3.46	3.40	3.44	3.42
Pristine bilayer MARE (%)	4.0	2.9	3.8	2.0

Table 4: The bilayers with the smallest 5 and largest 5 predicted interlayer spacings. All values are in Å.

	NN		RF		RVM		SVM	
Smallest 5	PtS_2|PtS_2	2.10	PtS_2|PtS_2	2.12	PdS_2|PdS_2	1.75	PtS_2|PtS_2	2.02
	PdS_2|PdS_2	2.12	PdS_2|PtS_2	2.16	PdS_2|NiS_2	2.01	PdS_2|PdS_2	2.12
	PtS_2|$PtSe_2$	2.16	PdS_2|PdS_2	2.18	PtS_2|PtS_2	2.06	PdS_2|NiS_2	2.21
	PdS_2|PtS_2	2.22	$PdSe_2$|PtS_2	2.22	PdS_2|$NiTe_2$	2.09	PdS_2|NbS_2	2.27
	$PdSe_2$|PtS_2	2.22	$PdSe_2$|PdS_2	2.24	PdS_2|NbS_2	2.15	PdS_2|1T-NbS_2	2.28
Largest 5	GaSe|WTe_2	3.63	$MoTe_2$|$MoTe_2$	3.63	$MoTe_2$|$TaTe_2$	3.66	WTe_2|1T-$MoTe_2$	3.65
	WTe_2|WTe_2	3.64	$MoTe_2$|$PdTe_2$	3.67	$PdTe_2$|$TaTe_2$	3.70	$ReTe_2$|WTe_2	3.66
	$PdTe_2$|BoronNitride	3.65	$MoTe_2$|WTe_2	3.69	WTe_2|WTe_2	3.76	$MoTe_2$|$PdTe_2$	3.68
	$PdTe_2$|Graphene	3.66	$ReTe_2$|WTe_2	3.70	$MoTe_2$|$PdTe_2$	3.77	WTe_2|WTe_2	3.73
	GaSe|WCl_2	3.68	WTe_2|WTe_2	3.72	$MoTe_2$|WTe_2	3.80	$MoTe_2$|WTe_2	3.74

Table 4 shows that the bilayers with lowest *d* contain Pt and Pd atoms that have very similar vdW radii and are associated with the group VI elements S or Se. The WTe_2|WTe_2 bilayer was predicted to have one of the largest *d* values by all models. As Te has a larger vdW radius than S or Se, this is intuitively sensible. Likewise, materials containing Se (vdW radius between Te and S) have intermediate values of *d* in 35, 26, 34 and 24 bilayers predicted by NN, RF, SVM and RVM respectively in the 5% percentile, compared with 12, 14, 16 and 16 bilayers predicted by NN, RF, SVM and RVM respectively in the 95% percentile. The main difference between the smallest and largest predicted *d* values is in the group VI atoms, where there is a significant trend in size from S < Se < Te, while the metals have similar vdW radii.

3.1. Prediction of the band gap

Given the success in predicting the interlayer distances for hybrid 2D materials using our combined DFT/ML approach, we also conducted a brief proof-of-concept experiment on whether we could predict properties more relevant to electrical or optical applications of 2D materials. We used DFT methods to calculate a relatively small number of band gaps for hybrid 2D materials. For the band gap prediction, we applied the BR1 bilayer representation. Using the LASSO algorithm, we obtained 11 significant descriptors per bilayer (far fewer than the 35 descriptors used per bilayer for the *d* prediction). Those descriptors are listed in the Supporting Information. In calculating the band gaps, some DFT functionals are known to considerably underestimate the band gap.[35] For example, some functionals predict a bandgap of ~4.5 eV for hexagonal boron nitride, while its experimental band gap is ~6 eV.[36] To overcome this problem, hybrid functionals that mix the DFT exchange with an exact exchange component have been developed that offer impressive agreement with experimental band gaps.[37] However, the present implementations of hybrid functionals are very expensive. For the initial proof of concept work, we used cheaper, non-corrected DFT for the prediction of the band gap, to see whether the ML models could predict the property. If so, it follows that ML methods would be capable of predicting band gaps calculated using more accurate but expensive DFT hybrid functionals, with similar accuracy. The DFT band gap value for the boron nitride bilayer is 3.94 eV for the AB' stacking configuration, which is close to the value of 4.01 eV obtained for same stacking (though using a different vdW method) reported in a recent work.[38] These authors also reported the equilibrium stacking as AA', with a band gap of 4.34 eV. However, using the TS vdW method, the lowest energy stacking configuration is

AB', and is only 3 meV lower in energy than AA'. To ensure consistency, we adopt the minimum energy stacking of the TS method. The band gap of each of the bilayers corresponds to a specific band gap alignment of the two constituting monolayers. While the alignments yield a band gap that is smaller than that of the monolayers, as is typically the case in semiconducting interfaces, the following 8 bilayers have a zero band gap: HfS_2|$MoTe_2$, HfS_2|WTe_2, $MoSe_2$|TiS_2, $MoTe_2$|1T-HfS_2, 1T-HfS_2|WTe_2, TiS_2|WSe_2, TiS_2|ZnO, and $TiSe_2$|WTe_2. These bilayers are interesting because they exhibit a special kind of type III band alignment,[39] where two semiconducting interfaces form a metallic structure across the vdW vacuum. We will explore these structures in detail in a future contribution. We display the band alignment for two of these bilayers, namely HfS_2|$MoTe_2$ and $MoSe_2$|TiS_2, in Figure 7(a,b). Here, the conduction band minimum (CBM) of one layer (HfS_2 in Figure 7(a) and TiS_2 in Figure 7(b)) and the valence band maximum (VBM) of the other layer ($MoTe_2$ in Figure 7(a) and $MoSe_2$ in Figure 7(b)) are aligned, leading to a zero band gap in an interface between two semiconductors.

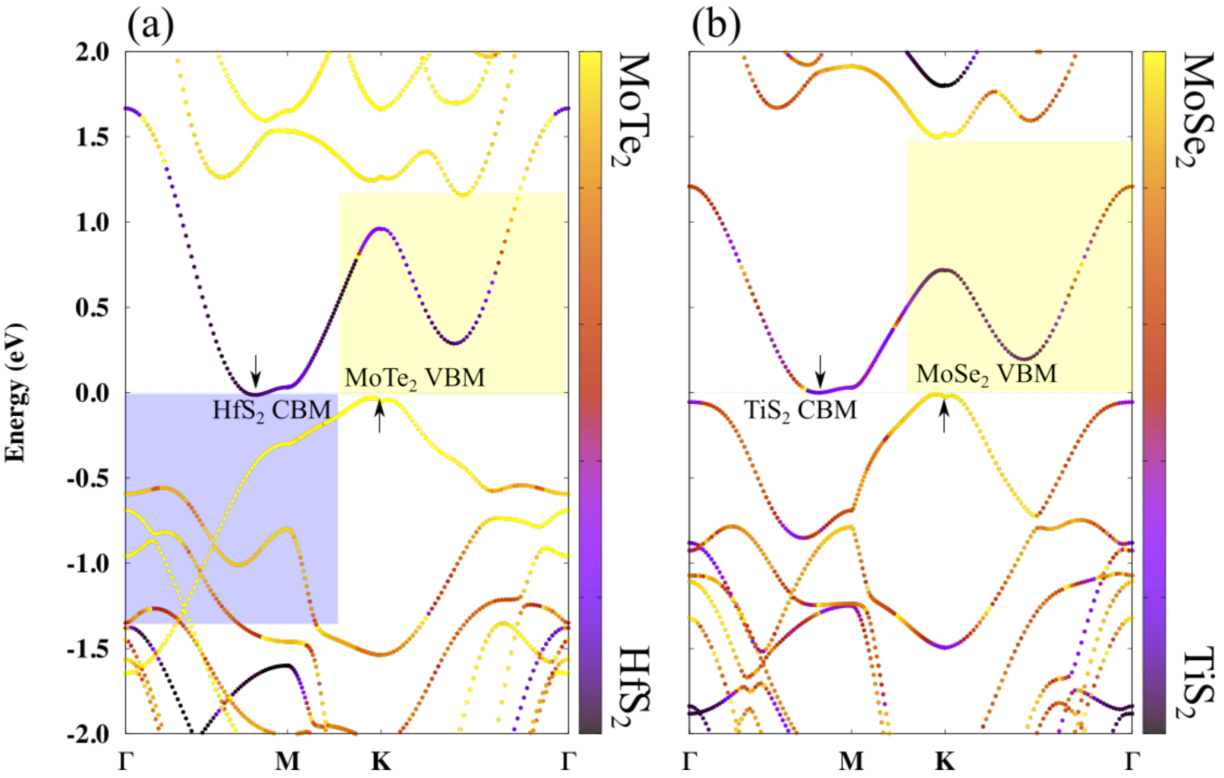

Figure 7: Band alignment in (a) HfS_2|$MoTe_2$ and (b) $MoSe_2$|TiS_2 bilayers. These bilayers exhibit type III alignment in which the CBM of one layer (HfS_2 in (a) and TiS_2 in (b)), the VBM of the other layer ($MoTe_2$ in (a) and $MoSe_2$ in (b)) and the Fermi energy are all aligned, leading to a zero band gap in an interface between two semiconductors. Note that the VBM of TiS_2 hybridizes with the $MoSe_2$ states. The energy is shifted such that the Fermi energy corresponds to zero.

The results of the ML predictions of the band gap are summarized in Table 5 and Figure 8. Even with a small data set, the predictive power of all four ML models, assessed by the MSE values for the test sets not used to generate the band gaps models, are similar to those predicting the interlayer distances (Tables 2 and 3). For band gaps models, the RVM model was less accurate than the other three models, which showed similar accuracies to each other. The high training set accuracies of the SVM model suggest that it could be overfitted like the interlayer distance model, although the test set predictions are similar to those of the RF and NN models. However, we must still be cautious using the band gap models that are based on a small data set, as they will not perform as well for bilayers that are well outside the domain of applicability of the model (defined by the training data). Figure 8 displays the correlation between the values of the band gap obtained by DFT and those obtained by each of the four ML models. The SVM and RVM greatly underestimate the band gap of boron nitride|MoS_2, which is 1.867 eV by 0.6 eV and 0.7 eV, respectively. RF and NN, on the contrary, predict it within an error of 0.3 eV and 0.2 eV, respectively.

Table 5: The R^2 and mean square error (MSE), in eV, for prediction of the band gap by each of the four ML models applied to the BR1 bilayer representation.

	R^2		MSE	
	Train	Test	Train	Test

RF	0.96	0.90	0.028	0.040
SVM	0.99	0.90	0.001	0.040
RVM	0.98	0.83	0.014	0.070
NN	0.90	0.88	0.063	0.047

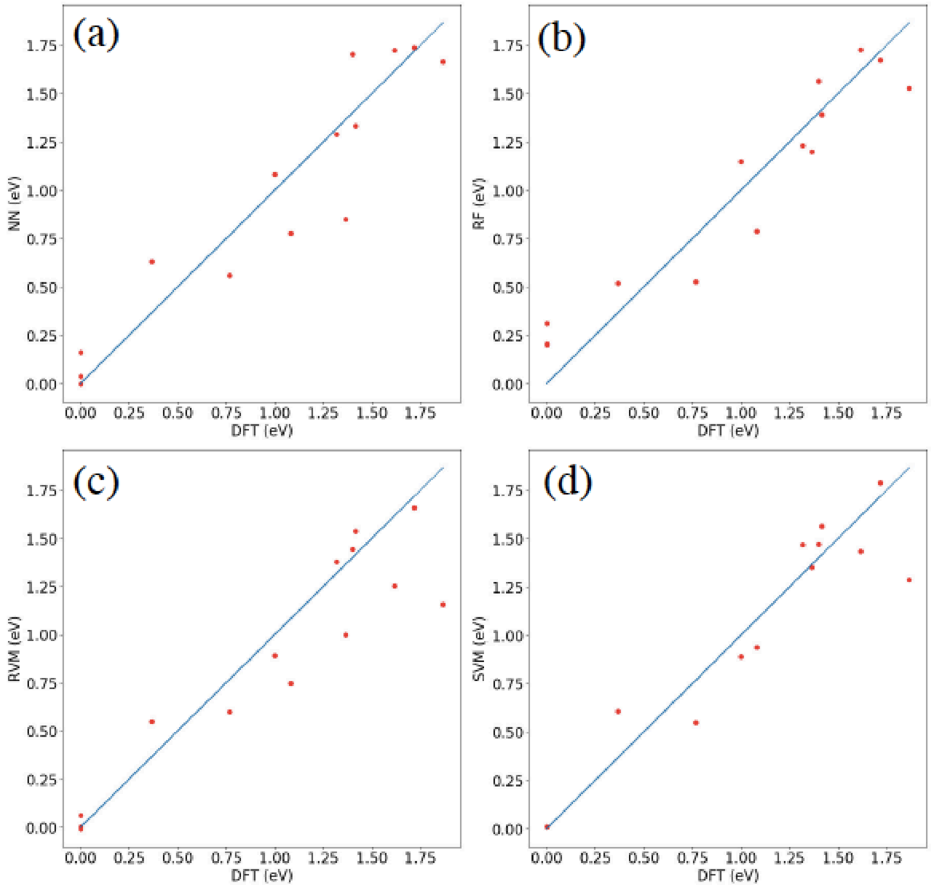

Figure 8: Comparison between the band gaps for the bilayers in the test set obtained using DFT, and those predicted from (a) NN, (b) RF, (c) RVM and (d) SVM. All values are in eV.

Using ML models trained using the BR1 representation, we predicted the band gaps for all of the possible 20 × 21/2 = 210 bilayers, based (summarized in Table 7). The NN, RVM and SVM all predict small negative values (−0.15 eV) for the minimum band gap, which is unphysical. These probably represent metallic bilayers and are due to extrapolation errors in the models. The BN bilayer is consistently predicted by the NN, RVM and SVM models to have the largest band gap, but the RF model predicts a substantially lower band gap. The distribution of the predicted band gaps varies significantly across the ML model, as can be seen in the percentile values. Unlike the interlayer distance predictions in Table 3, where very similar values were seen across the models, the larger standard deviation in Table 6 compared can be attributed to the small size of the data set that was used to train the models for band gap prediction.

Table 6: The summary statistics of the band gaps predicted for all 210 bilayers constructed from the 20 semiconducting monolayers (values in eV).

	NN	RF	RVM	SVM
Mean	0.56	0.88	0.55	0.88
Standard deviation	0.40	0.38	0.40	0.40
Minimum	−0.15	0.07	−0.15	−0.05
Maximum	3.92	3.34	3.92	3.93
5% percentile	0.09	0.32	0.09	0.27
50% percentile	0.47	0.89	0.47	0.93
95% percentile	1.33	1.49	1.33	1.43

4. Conclusions

Hybrid materials built from 2D monolayers are gaining attention as novel materials with potentially more easily tuneable properties. The current bottle-neck in exploiting these materials is the vast space of possible materials combinations, difficulty in predicting which will be best for a given application, and the real-world experimental difficulties in synthesizing them. While electronic structure calculations can make these predictions accurately, even the most efficient methods DFT are too time consuming to calculate the properties of every possible multilayer hybrid 2D material.

In the present work we have demonstrated how machine learning approaches can very effectively augment properties predicted by accurate but expensive DFT calculations. A selection of ML models could effectively predict structural and electronic properties of van der Waals heterostructures. The use of property labelled materials fragments[22] as descriptors for the monolayers proved to be effective, yielding relatively high and practically useful prediction accuracies for the interlayer spacing and bandgap. Fast prediction of these properties should also improve the synthesis bottleneck by predicting which materials are likely to function effectively and which are not worth synthesizing.

The current results show substantial promise for a combined DFT/machine learning approach to solving the problem of designing bespoke materials for a new generation of electronic devices and technologies. Here we have been able to predict the properties of nearly 1500 bilayer structures based on only 267 DFT calculations. Given that the time spent in the ML calculations is negligible compared with the DFT calculations, this represents a speed-up by a factor of about 5 compared to using DFT calculations alone. Moreover, it should now be possible to predict the properties of all 1.7 million bilayers built from the 1800 2D building blocks reported by Mounet *et al.*[1] using this same model. This represents a speed-up of nearly 4 orders of magnitude.

Supporting Information
Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements
This work was supported by resources provided by the National Computational Infrastructure (NCI), and Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. This work is also supported by the Australian Research Council, ARC DP160101301.

Received: ((will be filled in by the editorial staff))
Revised: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial staff))

References

[1] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I.E. Castelli, A. Cepellotti, G. Pizzi, N. Marzari, *Nat. Nano.* **2018**, *13*, 246.

[2] Jariwala, Deep, Tobin Marks, and Mark Hersam, *Nature Materials* **2017**, *16*, 155.

[3] A.K. Geim, I.V. Grigorieva, *Nature* **2013**, *499*, 419.

[4] Y. Jin, D.H. Keum, S.J. An, J. Kim, H.S. Lee, Y.H. Lee, *Adv. Mater.* **2015**, *27*, 5534.

[5] N. Flory, A. Jain, P. Bharadwaj, M. Parzefall, T. Taniguchi, K.Watanabe, L. Novotny, *Appl. Phys. Lett.* **2015**, *107*, 123106.

[6] J.S. Ross, P. Rivera, J. Schaibley, E. Lee-Wong, H. Yu, T. Taniguchi, K. Watanabe, J. Yan, D. Mandrus, D. Cobden, W. Yao, *Nano Lett.* **2017**, *17*, 638.

[7] C.H. Lee, G.H. Lee, A.M. van der Zande, W. Chen, Y. Li, M. Han, X. Cui, G. Arefe, C. Nuckolls, T.F. Heinz, J. Guo, *Nat. Nanotech.* **2014**, *9*, 676.

[8] M.S. Choi, D. Qu, D. Lee, X. Liu, K. Watanabe, T. Taniguchi, W.J. Yoo, *ACS Nano* **2014**, *8*, 9332.

[9] T. Bjorkman, A. Gulans, A.V. Krasheninnikov, R.M. Nieminen, *Phys. Rev. Lett*. **2012**, *108*, 235502.

[10] J. Paul, A.K. Singh, Z. Dong, H. Zhuang, B.C. Revard, B. Rijal, M. Ashton, A. Linscheid, M. N. Blonsky, D. Gluhovic, J. Guo, R.G. Hennig, *J. Phys.: Condens. Matt.* **2017**, *29*, 473001.

[11] V.V. Gobre, A. Tkatchenko, *Nat. Comm.* **2013**, *4*, 2341.

[12] A. Tkatchenko, M. Scheffler, *Phys. Rev. Lett.* **2009**, *102*, 073005.

[13] H. Peng, Z.-H. Yang, J.P. Perdew, J. Sun, *Phys. Rev. X* **2016**, *6*, 041005.

[14] I. Leven, T. Maaravi, I. Azuri, L. Kronik, O. Hod, *J. Chem. Theory Comput*. **2016**, *12*, 2896.

[15] P. Miro, M. Audiffred, T. Heine, *Chem. Soc. Rev.* **2014**, *43*, 6537.

[16] G. Kresse, J. Furthmuller, *Phys. Rev. B* **1996**, *54*, 11169.

[17] J.P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.

[18] Y. Cai, G. Zhang, Y. Zhang, *J. Phys. Chem. C* **2015**, *119*, 13929.

[19] S.A. Tawfik, T. Gould, C. Stampfl, M.J. Ford, *Phys. Rev. Mater.* **2018**, *2*, 034005.

[20] L. van der Maaten, G. Hinton, *J Mach. Learn. Res.* **2008**, *9*, 2579.

[21] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268.

[22] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Comm.* **2017**, *8*, 15679.

[23] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, A. von Lilienfeld, A. Tkatchenko, K. Muller, *J. Chem. Theory. Comput*. **2013**, *9*, 3404.

[24] R. Tibshirani, *J. Roy. Stat. Soc. Ser. B* **1996**, *51*, 267.

[25] C. M. Bishop, Pattern Recognition and Machine Learning, Springer **2006**.

[26] Francois Chollet, Keras, **2015**, https://github.com/keras-team/keras.

[27] V. Vapnik, The nature of statistical learning theory, Springer New York, **1995**.

[28] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, V. Vapnik, *in Advances in neural information processing systems*, **1997**, (pp. 155-161).

[29] M.E. Tipping, *J. Mach. Learn. Res.* **2001**, *1*, 211.

[30] D.A. Winkler, F.R. Burden, *J. Chem. Inf. Model.* **2015**, *55*, 1529.

[31] T.K. Ho, Random Decision Forests, *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC* **1995**, *14*, 278.

[3] T.K. Ho, *IEEE Trans. Patt. Anal. Mach. Intell.* **1998**, *20*, 832.

[33] K. M. Faraoun, A. Boukelif, *INFOCOMP* **2006**, *3*, 28.

[34] P.J. Rousseeuw, *Comput. Appl. Math.* **1987**, *20*, 53.

[35] A.J. Cohen, P. Mori-Sanchez, W. Yang, *Chem. Rev.* **2011**, *112*, 289.

[36] G. Cassabois, P. Valvin B. Gil, *Nat. Phot.* **2016**, *10*, 262.

[37] J.P. Perdew, *MRS Bull*. **2013**, *38*, 743.

[38] Y. Fujimoto, S. Saito, *Phys. Rev. B* **2016**, *94*, 245427.

[39] H. Kroemer, *Rev. Mod. Phys.* **2001**, *73*, 783.