Electronic Supplementary Material (ESI) for Nanoscale.

Unveiling the layer-dependent electronic properties in transition-metal dichalcogenides heterostructures assisted by machine learning

Tao Wang [a, b], Xiaoxing Tan [b], Yadong Wei [b, a], Hao Jin[*, b]

[a] *Institute of Theoretical Physics, State Key Laboratory of Quantum Optics and Quantum Optics Devices, Shanxi University, Taiyuan 030006, People's Republic of China.*

[b] *College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, People's Republic of China*

E-mail: jh@szu.edu.cn

1. Model reliability test

Table S1 Scores of models with different training set sizes

Number of data	117	157	187
R^2	94.9%	94.7%	95.8%

Table S2 5-fold cross-validation scores of different algorithms

Algorithms	GPR	GBR	RFR	KNR	LR	SV
R^2	94.9%	95.5%	95.6%	69.9%	81.0%	77.5%

2. Importance of the features and model performance

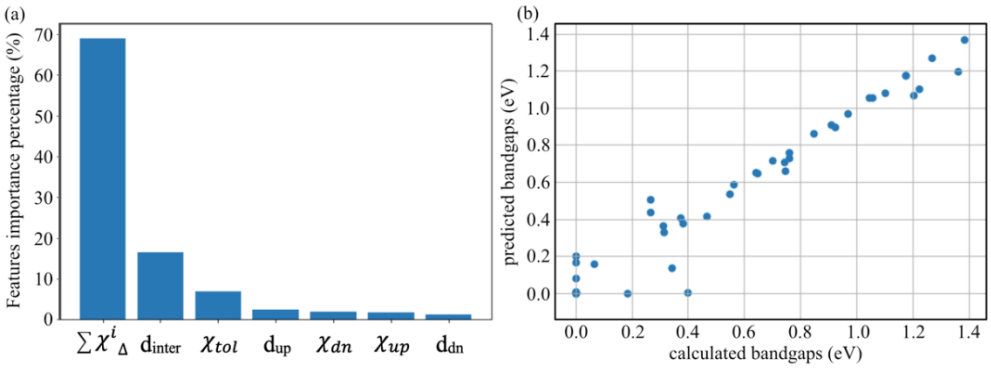

Fig. S1 (a) Importance of the features. (b) Comparison between the predicted bandgaps and the DFT-calculated bandgaps.

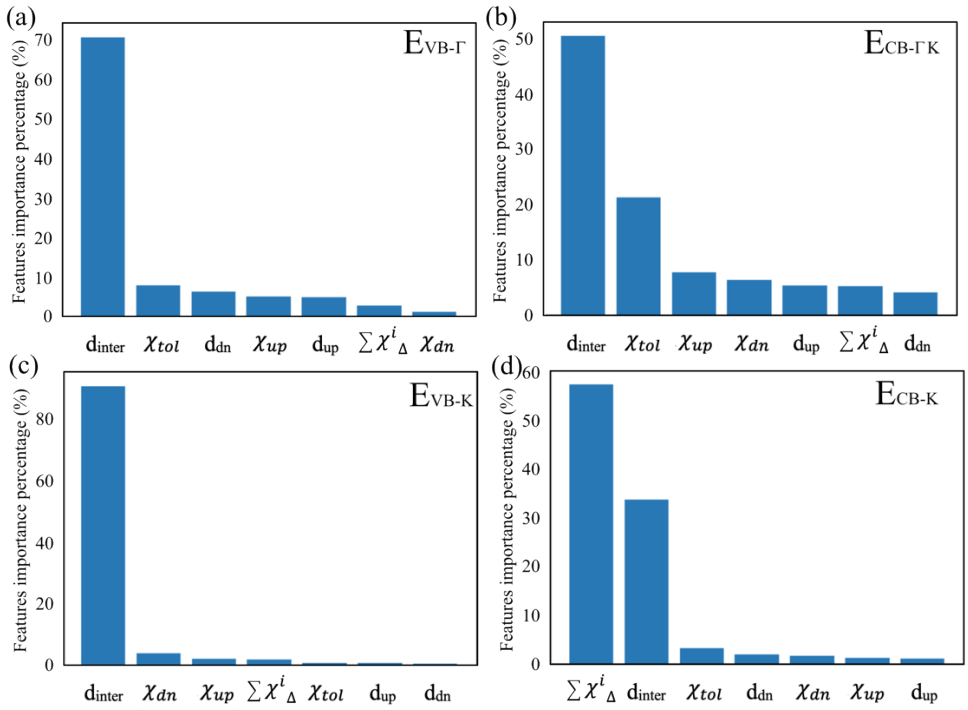

Fig. S2 Importance of the features for (a) E_{VB_Γ}, (b) $E_{CB\_G\Gamma}$, (c) E_{CB_K}, and (d) E_{VB_K}.

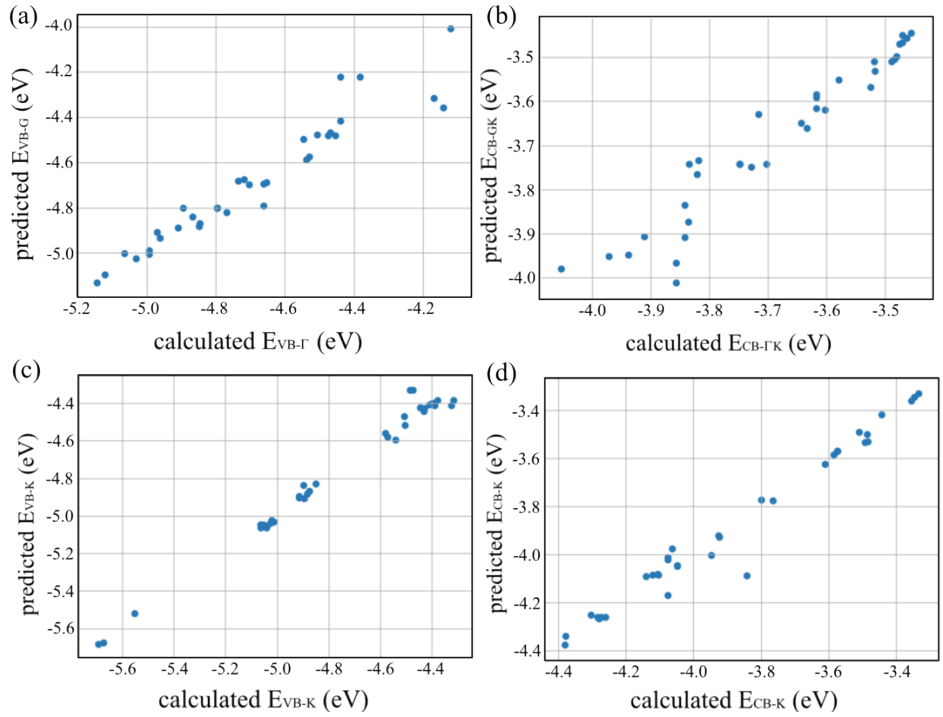

Fig. S3 Comparison between the predicted band edge positions and the DFT-calculated band edge positions for (a) E_{VB_Γ}, (b) $E_{CB\_G\Gamma}$, c) E_{CB_K}, and (d) E_{VB_K}.

3. Weight of the CBM and VBM contributed by each constituted TMD material

We predict the weight of each constituted TMD material for the CBM and VBM. If the weight of some kind of constituted TMD material exceeds 70%, it is regarded that the

band edge is dominated by this material. Otherwise, the band edge is contributed by both constituted TMD materials. Here, we use 1 (-1) to indicate that the band edge is dominated by up-layer (down-layer), and 0 to represent it is contributed by both layers. Analyzing the dataset, it is worth noting that the band edge positions at K point are almost contributed by single TMD material, especially for the CBM. While it becomes complicated at the Γ point. We find that there are around 10% of the heterostructures whose VBM is contributed by both layers. While for the CBM at the ΓK point, up to 20% of the heterostructures is jointly contributed by both layers. Here, a Gaussian process classifier is used to deal with the multi-classification problems. To comprehensively evaluate the performance of the classifier, confusion matrixes of the test dataset are illustrated in **Fig. S4**, which is used to compare the classification results with the actual measured values. The principal diagonal of the confusion matrix indicates the good ability to separate contributing layers for small sample models.

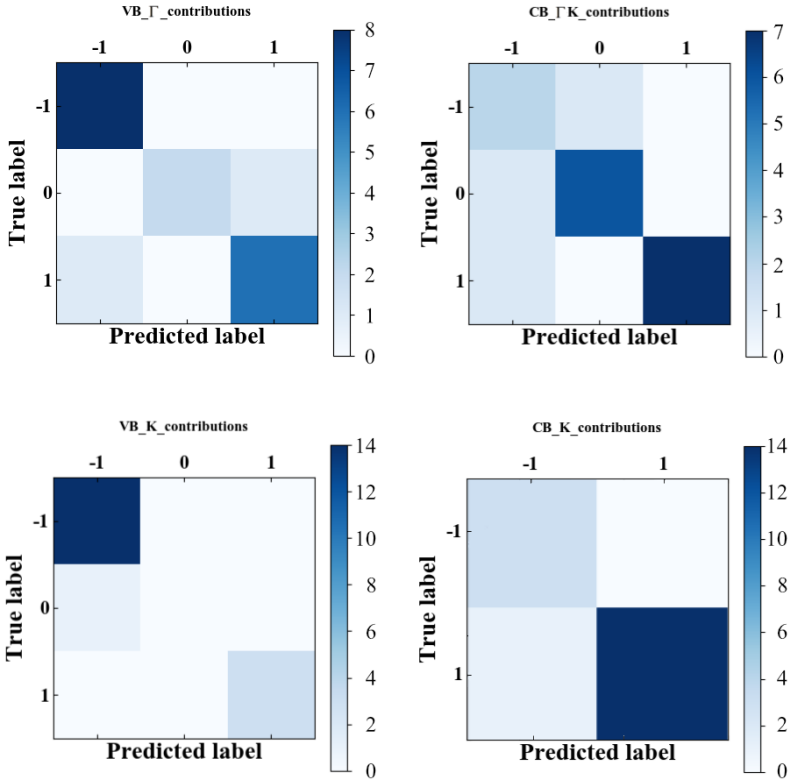

Fig. S4 Confusion matrix. The principal diagonal of the confusion matrix indicates the good ability to separate contributing layers for ML models.

4. Performance of Bayesian optimization model

Table S3. Comparison between the predicted structures obtained by Bayesian optimization method and the calculated structures in the test set.

	Predicted values (4- and 5-layer)	Calculated values (4- and 5-layer)	Predicted values (6-layer)	Calculated values (6-layer)
E_g	$5WS_2$	$5WS_2$	$6WS_2$	$6WS_2$
(max)	1.14	1.14	1.12	1.12
$E_{VB\text{-}K}$	$5MoS_2$	$5MoS_2$	$6MoS_2$	$6MoS_2$
(min)	-5.67	-5.67	-5.65	-5.65
$E_{VB\text{-}K}$	$4MoTe_2$-$MoTe_2$	$4MoTe_2$-$MoTe_2$	$4WTe_2$-$2MoTe_2$	$4WTe_2$-$2MoTe_2$
(max)	-4.05	-4.05	-4.06	-4.06
$E_{CB\text{-}K}$	$2MoTe_2$-$3MoSe_2$	$2MoTe_2$-$3MoSe_2$	$3MoTe_2$-$3MoSe_2$	$3MoTe_2$-$3MoSe_2$
(min)	-4.67	-4.67	-4.74	-4.74
$E_{CB\text{-}K}$	$4WSe_2$	$4WSe_2$	$6WSe_2$	$6WSe_2$
(max)	-3.32	-3.32	-3.35	-3.35
$E_{VB\text{-}\Gamma}$	WS_2-$3MoS_2$	WS_2-$3MoS_2$	$6MoS_2$	$6MoS_2$
(min)	-5.0	-5.0	-4.96	-4.96
$E_{VB\text{-}\Gamma}$	$3MoTe_2$-$2MoS_2$	$3MoTe_2$-$2MoS_2$	$3MoTe_2$-$3MoS_2$	$4MoTe_2$-$2MoS_2$
(max)	-3.71	-3.71	-3.68	-3.66
$E_{CB\_\Gamma K}$	WTe_2-$4MoS_2$	$3MoTe_2$-$2MoS_2$	$4MoTe_2$-$2MoS_2$	$4MoTe_2$-$2MoS_2$
(min)	-4.35	-4.56	-4.63	-4.63
$E_{CB\_\Gamma K}$	$2WTe_2$-$3WSe_2$	$2WTe_2$-$3WSe_2$	$5WTe_2$-WSe_2	$5WTe_2$-WSe_2
(max)	-3.49	-3.49	-3.43	-3.43

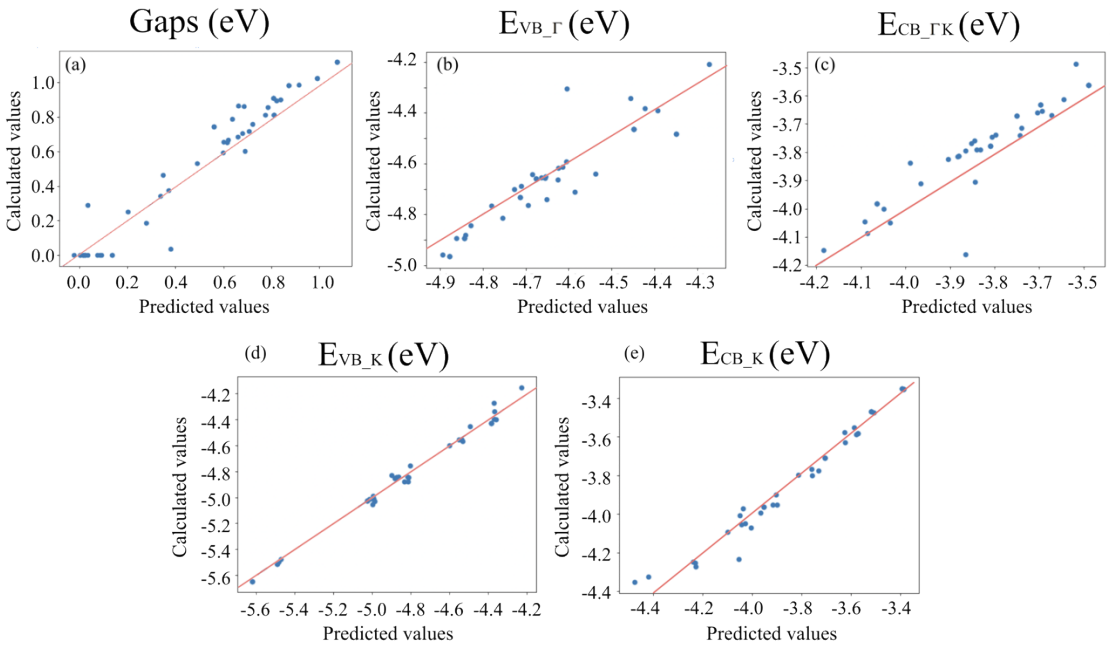

Fig. S5 Comparison between the predicted values and the calculated values for 6-layer vdW heterostructures.

5. Effect of stacking on properties of N-layer vdW heterostructures

Here, in order to understand the effect of stacking order on the properties of N-layer vdW heterostructures, we choose the 3R phase, which is very different from 2H phase, as shown in **Fig. S6**. In **Fig. S7**, we plot the proportions of N-layer (N = 2,3,4,5) vdW heterostructures with direct bandgap, indirect bandgaps and zero bandgap. The results show that with the increase of N, the proportion of direct bandgaps decreases, while the proportion of indirect bandgaps increases. And they finally tend to be stable. In addition, as shown in **Fig. S8**, the configurations with direct bandgaps are mainly present at the lower right area of the matrix, the heterojunctions with indirect bandgaps mainly occupy the upper right area, and the zero-bandgap configurations gather at the lower left and upper right corner. As shown in **Fig. S9**, for different band alignments, with the increase of N, the proportion of type-I band alignment increases gradually and finally tends to be stable. While for type-II and type-III band alignments, their proportions are relatively stable. Furthermore, as shown in **Fig. S10**, the regions of type-I and type-II are on the diagonal of the matrix. Note that the results obtained for

3R stacking order are very similar with those of 2H stacking order. In other words, for each specific TMDs the properties may be impacted by stacking order. While from the statistical level, the general trends are robust and are independent to the stacking mode.

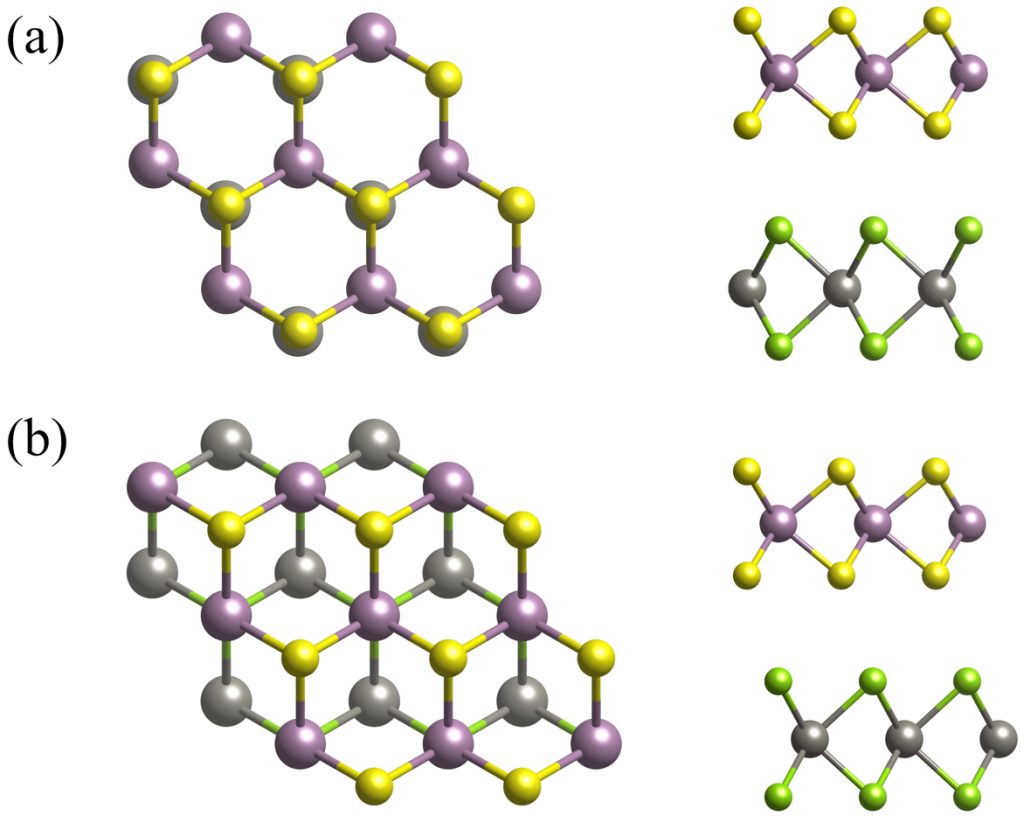

Fig. S6. Top and side views for (a) 2H and (b) 3R stacking orders.

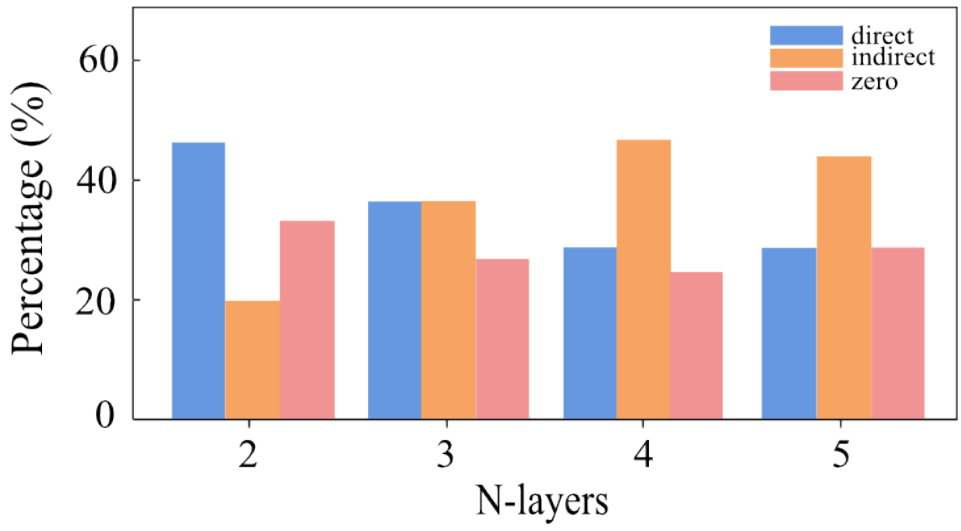

Fig. S7 Proportion of configurations with direct, indirect, and zero bandgaps in N-layers systems.

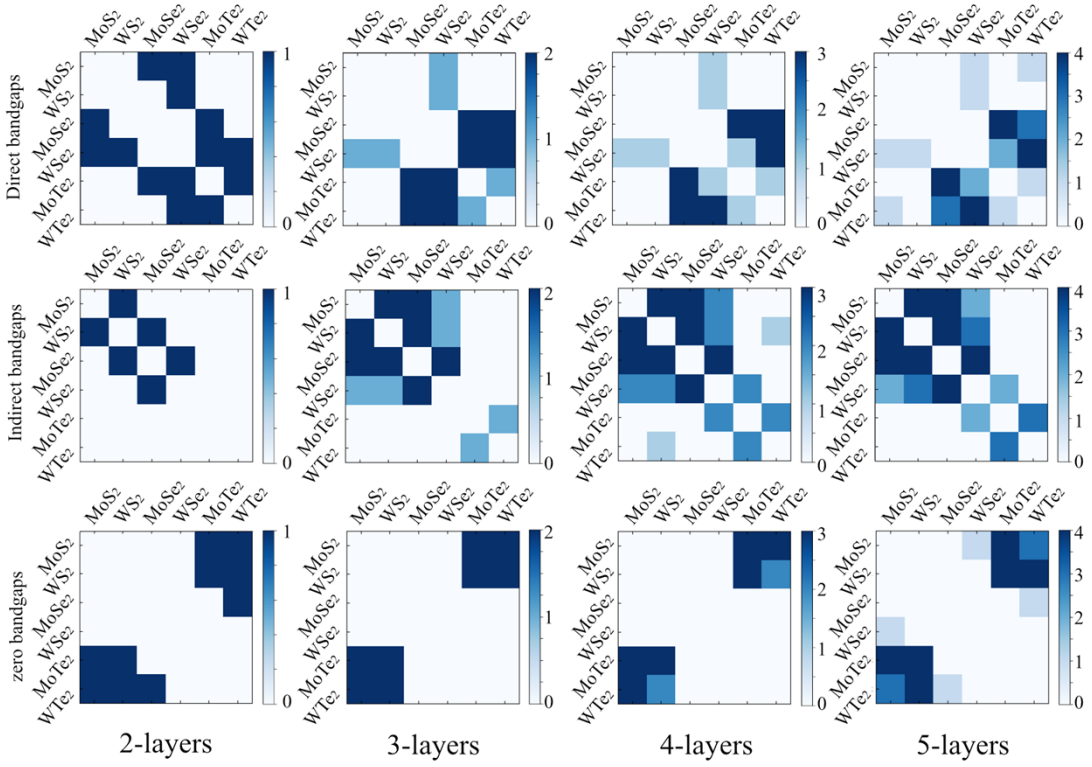

Fig. S8 Number of configurations with direct, indirect, and zero bandgaps.

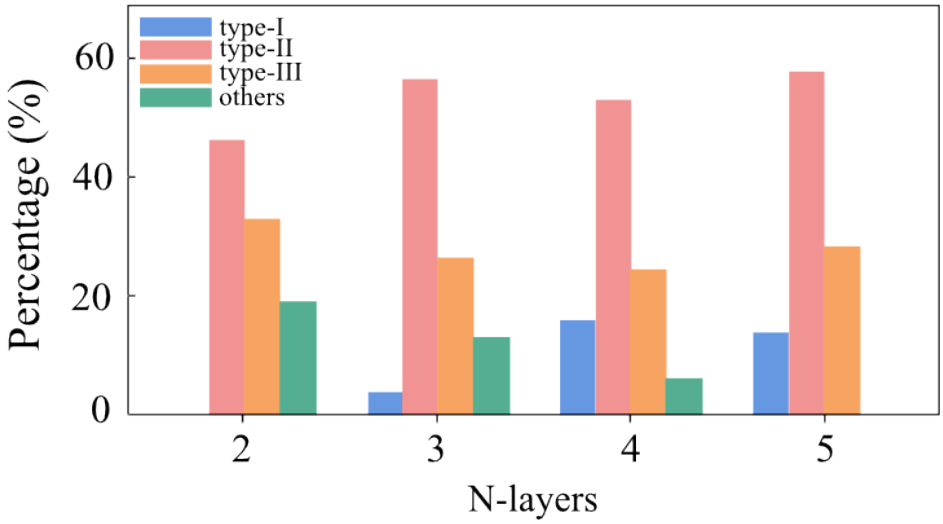

Fig. S9 Proportion of configurations of type-I, type-II, type-III and other band alignments in N-layers systems.

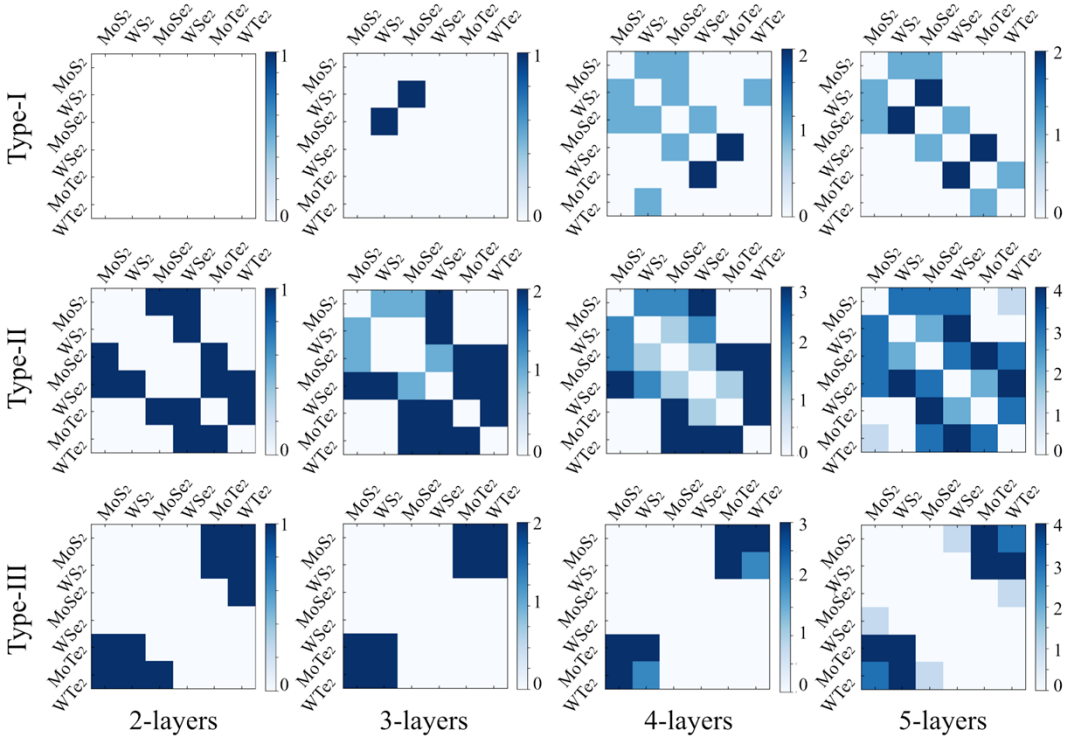

Fig. S10 Number of configurations of type-I, type-II, and type-III band alignments.