Supplementary information of

Predefined attention-focused mechanism using center-environment features: A machine learning study of alloying effects on stability of Nb_5Si_3 alloys

Yuchao Tang[1, 2], Bin Xiao[1], Shuizhou Chen[3], Quan Qian[3], and Yi Liu*[1]

[1]Materials Genome Institute, Shanghai Engineering Research Center for Integrated Circuits and Advanced Display Materials, Shanghai University, Shanghai 200444, China

[2]State Key Laboratory of Functional Materials for Informatics, Shanghai Institute of Micro-system and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China

[3]School of Computer Engineering & Science, Shanghai University, Shanghai, 200444, China

*Corresponding author: yiliu@shu.edu.cn (Yi Liu)

Text S1

In this work, the E_{SS}, $E_{DS,}$ and <Δd> were calculated respectively. The calculation equations are shown as follows (1-3):

$$E_{SS} = E_{pure+X+Y} + E_{S1} - E_{pure+Y} - E_X \quad (1)$$

$$E_{DS} = E_{pure+X+Y} + E_{S1} + E_{S2} - E_{pure} - E_X - E_Y \quad (2)$$

where ***X*** and ***Y*** are double-site substitution elements in α-Nb_5Si_3. $E_{pure+X+Y}$ is the total energy of the α-Nb_5Si_3 system after double-site substitution calculated by DFT; E_{pure+Y} is the energy of pristine α-Nb_5Si_3 with ***Y*** site substitution elements; E_{S1} and E_{S2} are the total energy per atom of pure Nb or Si crystal, respectively. E_X and E_Y are the crystalline single-atom energy of each alloying element.

$$<\Delta d> = \frac{1}{n} * \sum_{i=1}^{n} (|r_i^{'} - r_0^{'}| - |r_i - r_0|) \quad (3)$$

where <Δd> is the average bond length change of the local structure, the r_i and $r_i^{'}$ are positions of substitution elements before and after ion relaxation in the crystal structure, and r_0 and $r_0^{'}$ are the positions of the *i*-th atom in the substitution element's nearest neighbor *n* atoms before and after ion relaxation in the crystal structure, respectively.

All computational simulations were conducted utilizing Density Functional Theory (DFT) as integrated within the Vienna Ab-initio Simulation Package (VASP) software[1,2]. We employed the Perdew-Burke-Ernzerhof (PBE) parametrization for the generalized gradient approximation (GGA-PBE) [3,4] to calculate the exchange-correlation energy. The interactions between core and valence electrons were approximated using Projector-Augmented Wave (PAW) potentials. The kinetic energy cutoff for the wave functions was established at 450 eV. For α-Nb_5Si_3, a Gamma-centered k-mesh of 5 × 5 × 3 was applied to discretize the Brillouin zone (BZ). The total energy calculations were refined to a convergence threshold of 1×10^{-5} eV, while the forces were converged with an accuracy of less than 0.02 eV/Å. Taking into account the dilute concentration of the alloying elements, lattice relaxation was initially carried out within the α-Nb_5Si_3 systems. Subsequently, ionic relaxation was executed under the constraint of maintaining a constant cell volume and shape.

Table S1 Non-equivalent substitution pairs with the corresponding site distances in α-Nb_5Si_3 studied in this work.

System	Sites	Non-equivalent site pair	Site distance (Å)	Number of systems studied in this work
α-Nb_5Si_3 more closely packed layer	Nb_I site	$\boldsymbol{X}_{NbI}\boldsymbol{Y}_{SiI}$	2.59	196
		$\boldsymbol{X}_{NbI}\boldsymbol{Y}_{SiII}$	2.99	196
		$\boldsymbol{X}_{NbI}\boldsymbol{Y}_{NbII}$	3.05	196
	Si_I site	$\boldsymbol{X}_{SiI}\boldsymbol{Y}_{SiI}$	2.42	196
		$\boldsymbol{X}_{SiI}\boldsymbol{Y}_{NbI}$	2.59	196
		$\boldsymbol{X}_{SiI}\boldsymbol{Y}_{NbII}$	2.62	196
		$\boldsymbol{X}_{SiI}\boldsymbol{Y}_{NbII}$	2.66	196
α-Nb_5Si_3 less closely packed layer	Nb_{II} site	$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{SiI}$	2.62	196
		$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{SiI}$	2.66	196
		$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{SiII}$	2.75	196
		$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{NbII}$	2.88	196
		$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{NbI}$	3.05	196
		$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{NbII}$	3.11	196
		$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{NbII}$	3.26	196
		$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{NbII}$	3.49	196
		$\boldsymbol{X}_{NbII}\boldsymbol{Y}_{NbII}$	3.57	196
	Si_{II} site	$\boldsymbol{X}_{SiII}\boldsymbol{Y}_{NbII}$	2.75	196
		$\boldsymbol{X}_{SiII}\boldsymbol{Y}_{NbI}$	2.99	196

Table S2 Elementary properties used in the feature construction.

#	Features	Abbreviate	
		Center	Environment
1	Atomic number, start counting left top, left-right sequence	AN_C	AN_E
2	Atomic weight	AW_C	AW_E
3	Charge, nuclear effective (Clementi)	CH_C	CH_E
4	Density	DS_C	DS_E
5	Distance, core electron (Schubert) Molar	DC_C	DC_E
6	Distance, valence electron (Schubert) Moment	DV_C	DV_E
7	Electronegativity (Martynov and Batsanov)	EN_C	EN_E
8	Energy, cohesive Brewer	EC_C	EC_E
9	Energy of ionization, first	EI_C	EI_E
10	Enthalpy of melting	EM_C	EM_E
11	Mendeleev Pettifor	MP_C	MP_E
12	Nuclear charge, effective	NC_C	NC_E
13	Quantum number	QN_C	QN_E
14	Radius, covalent	RC_C	RC_E
15	Radius, metal (Waber)	RM_C	RM_E
16	Temperature, melting	TM_C	TM_E
17	Thermal neutron capture cross-section	TN_C	TN_E
18	Valence electron number	VE_C	VE_E
19	Bulk modulus	BM_C	BM_E
20	Periodic number	PN_C	PN_E

Text S2 Chemical composition (CC) feature model

The Chemical Composition (CC) feature model is constructed by the local chemical composition information, excluding the information of crystal structural[5]. This approach is also applied to model the chemical composition of double-site substitutions in α-Nb_5Si_3.

Similar to the construction of the Center-Environment (CE) features, the CC feature model can be expressed as follows.

$$D = [D_1, \ldots, D_i, \ldots, D_n, T], (e.g., \mathrm{n} = 20) \tag{4}$$

$$D_i = [d_{C,i}, d_{E,i}], i = 1, 2, \ldots, n \tag{5}$$

$$d_{C,i} = p_{C,i} \tag{6}$$

$$d_{E,i} = \sum_{j=1}^{N} \omega_{E,j} p_{E,j,i} \tag{7}$$

$$\omega_{E,j} = \frac{r_j^m}{\sum_{j=1}^{N} r_j^m} (m = 0) \tag{8}$$

where $p_{E,j,i}$ is the i-th property of the j-th environment atom. When m= 0, $\omega_{E,j}$ is essentially a normalized factor so that the local chemical composition is naturally taken into account after the summation over all local environment atoms.

Text S3

In this study, after establishing modeling algorithms for both supporting vector regression and random forest regression, the next step was to enhance model performance through hyper-parameter optimization. Given that each method possesses a multitude of tunable hyper-parameters, a comprehensive exploration of each would demand substantial computational resources. To address this, we conducted a series of iterative experiments, ultimately deciding to employ a grid search optimization approach for a select set of critical parameters. This strategy effectively identified the optimal configuration. The parameters that underwent optimization are detailed in the list that follows.

Table S3 Modeling hyper-parameter optimization candidates

Regression algorithm	Parameter List
Support Vector Regression (SVR)	C: 0.1, 1, 10, 100,1000
	kernel: RBF
	gamma: 0.001, 0.01, 0.1, 0.5
Random Forest (RF)	n_estimators: 20, 50, 70
	max_depth: 3, 4, 5, 7, 10
	min_samples_split: 2, 4, 6, 10

Table S4 Prediction performance of α-Nb_5Si_3 by the CE_{NN} model with different weights r_j^m ($m = -1, -\frac{1}{2}$).

Target property	Performance metric	CE_{NN} ($\omega_j = 1/r$)				CE_{NN}($\omega_j = 1/\sqrt{r}$)			
		SVR		RF		SVR		RF	
		Train	Test	Train	Test	Train	Test	Train	Test
E_{SS} (meV)	R^2	0.90	0.89	0.92	0.82	0.94	0.92	0.94	0.85
	MAE	383.27	421.79	397.37	612.24	323.82	345.37	351.95	528.40
	RMSE	586.46	641.44	513.06	808.46	481.81	487.05	451.69	692.18
E_{DS} (meV)	R^2	0.91	0.90	0.87	0.81	0.93	0.92	0.91	0.82
	MAE	352.27	384.36	493.55	638.43	327.84	372.95	417.26	573.34
	RMSE	531.66	560.77	632.19	803.57	486.23	524.68	533.91	750.93
<Δd> (10^{-2}Å)	R^2	0.77	0.72	0.88	0.74	0.81	0.75	0.86	0.75
	MAE	6.29	7.99	6.04	8.0	6.83	10.90	6.53	9.62
	RMSE	11.59	13.23	8.49	11.90	11.44	11.99	9.33	11.60

Table S5 Prediction performance of α-Nb_5Si_3 by the CE_{AET} model with different weights r_j^m ($m = -1, -\frac{1}{2}$).

Target property	Performance metric	CE_{AET} ($\omega_j = 1/r$)				CE_{AET}($\omega_j = 1/\sqrt{r}$)			
		SVR		RF		SVR		RF	
		Train	Test	Train	Test	Train	Test	Train	Test
E_{SS} (meV)	R^2	0.96	0.95	0.95	0.89	0.95	0.94	0.95	0.90
	MAE	258.01	278.65	328.10	480.17	277.99	324.68	320.46	471.96
	RMSE	397.44	409.86	425.08	627.95	414.46	459.52	415.82	621.26
E_{DS} (meV)	R^2	0.95	0.93	0.86	0.81	0.94	0.93	0.95	0.89
	MAE	276.74	329.43	486.73	578.16	317.95	339.29	309.98	506.63
	RMSE	416.49	465.83	675.40	780.35	476.85	498.75	393.20	661.31
<Δd> (10^{-2}Å)	R^2	0.80	0.72	0.85	0.74	0.85	0.74	0.84	0.74
	MAE	6.37	7.50	6.29	7.81	7.04	8.20	7.76	7.80
	RMSE	10.80	12.01	11.19	13.92	11.19	13.92	10.96	15.16

Table S6 Prediction performance of α-Nb_5Si_3 by the CC model.

Target property	Performance metric	CC model			
		SVR		RF	
		Train	Test	Train	Test
E_{SS} (meV)	R^2	0.32	0.29	0.33	0.32
	MAE	1228.03	1266.92	1210.74	1298.51
	RMSE	1545.38	1624.56	1535.84	1545.96
E_{DS} (meV)	R^2	0.67	0.65	0.71	0.68
	MAE	667.86	810.91	684.61	829.79
	RMSE	1029.65	1084.57	970.28	1020.40
<Δd> (10^{-2}Å)	R^2	0.48	0.46	0.49	0.35
	MAE	11.43	13.33	11.57	12.69
	RMSE	16.93	20.79	16.90	19.11

Table S7 GCN Model Parameter Settings

Parameter	Value	Description	Scope
nfeat	32	Input node feature dimension (*din*)	All graphs
nhid	64	Hidden layer dimension of the first GCN layer (*dhid*)	All graphs
nclass	16	Output dimension of the second GCN layer (*dmid*)	Intermediate layer
node_num	100	Fixed number of nodes per graph (NN)	All graphs
nout	1	Final output dimension of the model (*dout*)	Fully connected layer
readout	"linear"	Graph-level feature aggregation strategy (options: "mean" or "linear")	Entire graph
dropout	0.5	Dropout rate for regularization (applied during training)	Training phase

Table S8 GAT Model Parameter Settings

Parameter	Value	Description	Scope
in_features	64	Input feature dimension (din)	Node-level
out_features	32	Output feature dimension(dout)	Node-level
dropout	0.6	Dropout rate for randomly masking attention weights	Training phase
node_num	0.2	Negative slope coefficient for LeakyReLU activation	Attention coefficient computation
alpha	True	Whether to apply ELU activation and concatenate multi-head outputs	Node-level

ALIGNN model Parameter

Table S9 RBFExpansion Parameter Settings

Parameter	Value	Description	Type
vmin	0	Minimum value of RBF centers (normalized input range, e.g., bond lengths)	float
vmax	8	Maximum value of RBF centers (normalized input range, e.g., bond lengths)	float
bins	40	Number of RBF centers (controls feature resolution)	int
lengthscale	None	Width parameter (σ) of Gaussian kernels. If None, auto-calculated as (vmax - vmin)/bins	float

Table S10 ALIGNN Key Parameter Settings

Parameter	Value	Description	Type
triplet_input_features	32	Dimension of angle features after RBF embedding (matches bins in RBFExpansion)	int
hidden_dim	64	Hidden layer dimension of graph convolution layers (controls model capacity)	int
num_conv_layers	5	Number of alternating graph convolution iterations (atom-line graph updates)	int

Table S11 Prediction results of substitution energies in α-Nb_5Si_3 using different machine learning algorithms.

Sites	Target	Performance metric	GBR	LGBM	XGB
Nb_I	E_{SS}	R^2	0.88	0.922	0.923
		RMSE	529.69	419.72	418.04
		MAE	359.52	328.22	338.33
	E_{DS}	R^2	0.933	0.893	0.934
		RMSE	387.08	508.52	383.85
		MAE	282.70	345.70	287.98
Nb_{II}	E_{SS}	R^2	0.879	0.854	0.884
		RMSE	430.90	498.96	467.76
		MAE	307.34	384.92	340.12
	E_{DS}	R^2	0.928	0.919	0.900
		RMSE	397.78	421.95	481.78
		MAE	291.22	308.54	331.23
Si_I	E_{SS}	R^2	0.962	0.956	0.969
		RMSE	508.38	543.84	459.77
		MAE	341.02	373.56	326.41
	E_{DS}	R^2	0.963	0.928	0.960
		RMSE	421.91	584.23	434.43
		MAE	306.75	434.21	317.17
Si_{II}	E_{SS}	R^2	0.874	0.855	0.877
		RMSE	445.99	597.58	557.02
		MAE	279.73	448.18	383.81
	E_{DS}	R^2	0.906	0.911	0.923
		RMSE	520.96	591.88	542.28
		MAE	333.85	445.23	378.25
Nb_5Si_3	E_{SS}	R^2	0.954	0.941	0.949
		RMSE	451.12	500.80	461.58
		MAE	311.43	362.87	324.63
	E_{DS}	R^2	0.934	0.925	0.937
		RMSE	521.98	529.91	512.91
		MAE	348.53	386.75	372.60

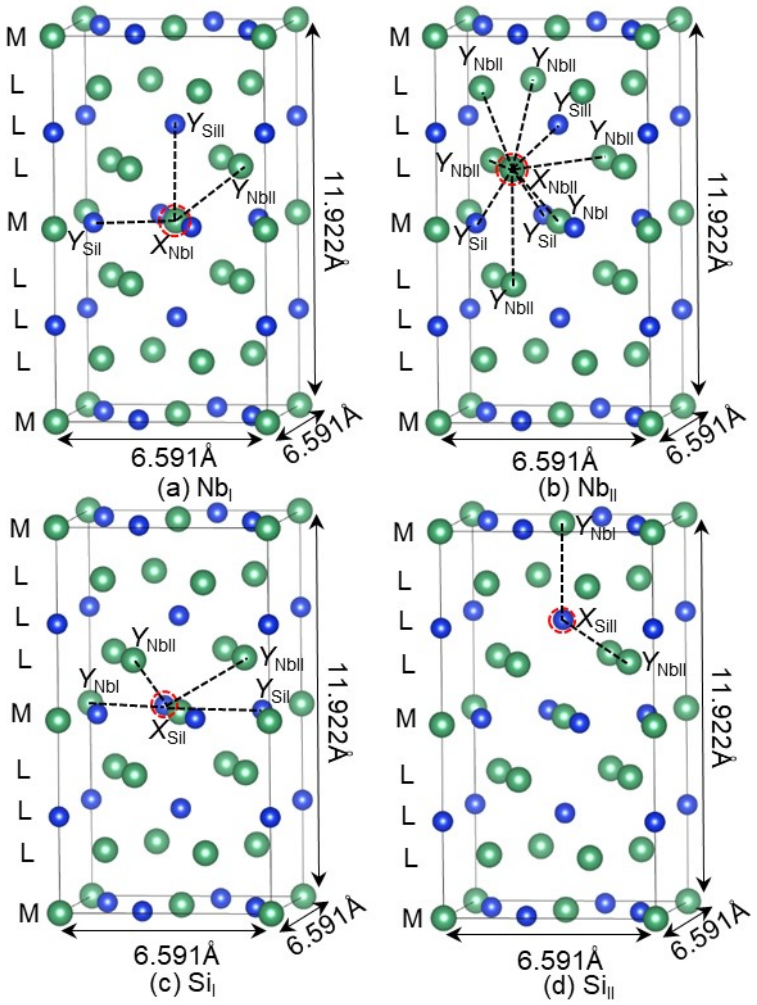

Fig. S1 The non-equivalent substitution sites of α-Nb_5Si_3 (a) Nb_I, (b) Si_I, (c) Nb_{II}, and (d) Si_{II}.

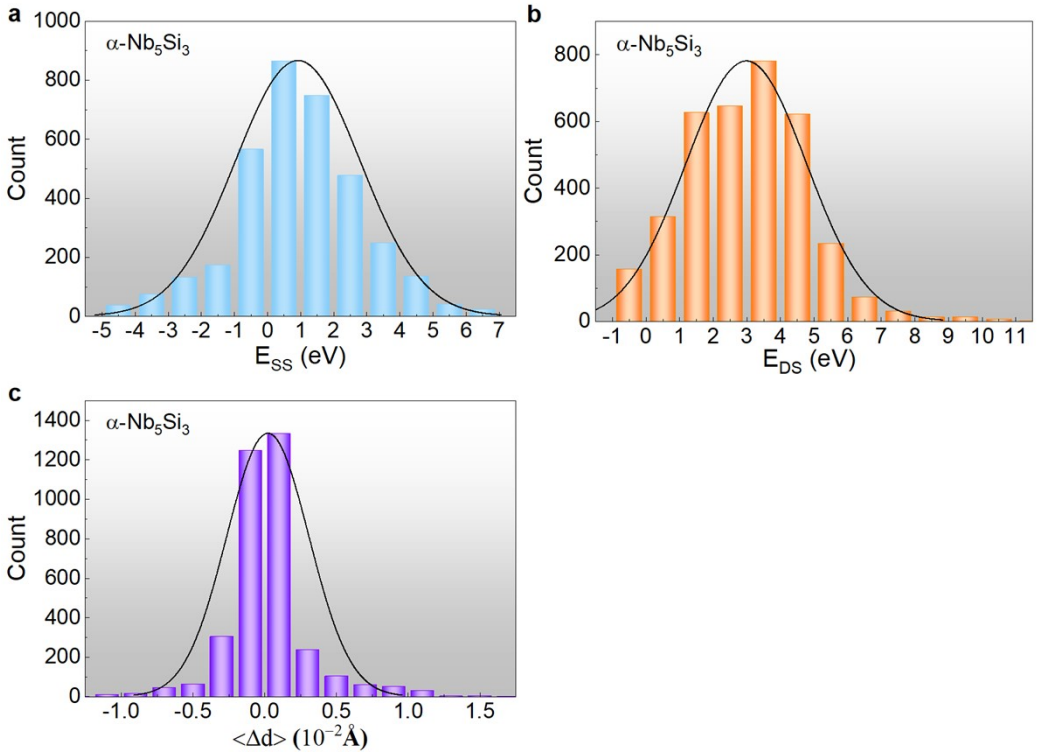

Fig. S2 Statistical plots of target data (E_{SS}, E_{DS}, and <Δd>), where (a)-(c) were for the substitution of alloying elements in α-Nb_5Si_3.

H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	*	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn

(g)

Fig. S3 The distribution of the 14 substituted alloying elements in the periodic table for the α-Nb_5Si_3 structures, with dark green representing the substitution elements.

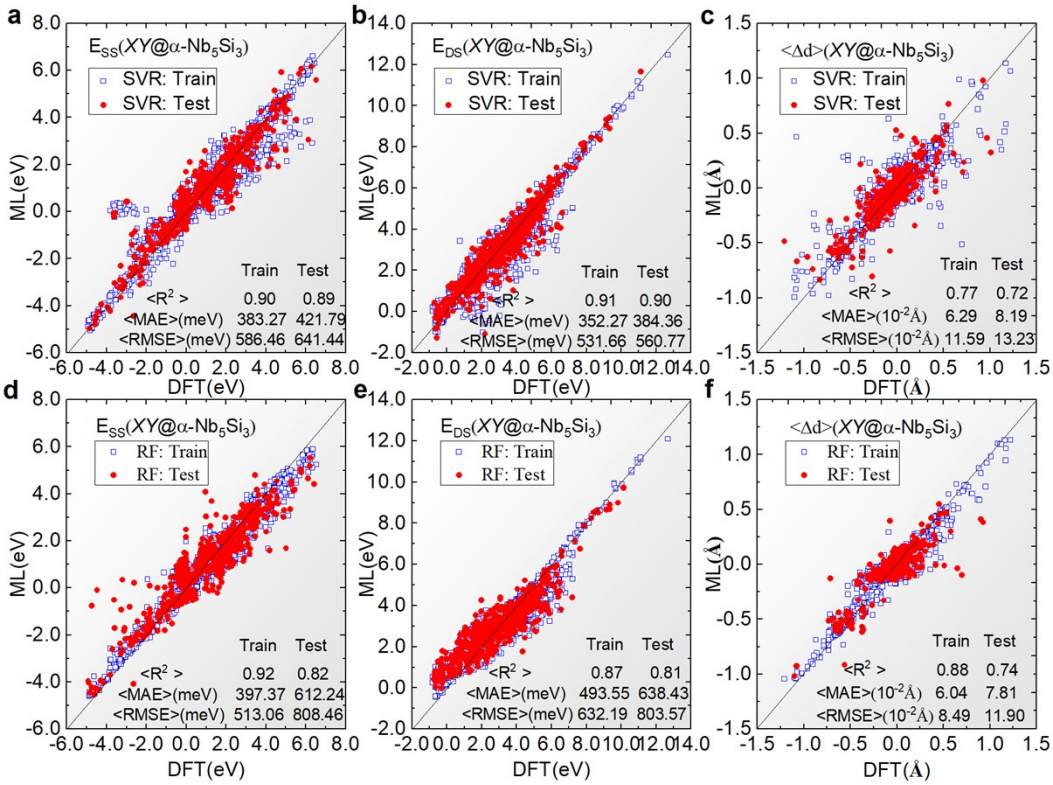

Fig. S4 (a, d) E_{SS}, (b, e) E_{DS}, (c, f) <Δd> of α-Nb_5Si_3 predicted by the CE_{NN} models with SVR and RF algorithms with $\omega_j = 1/r$.

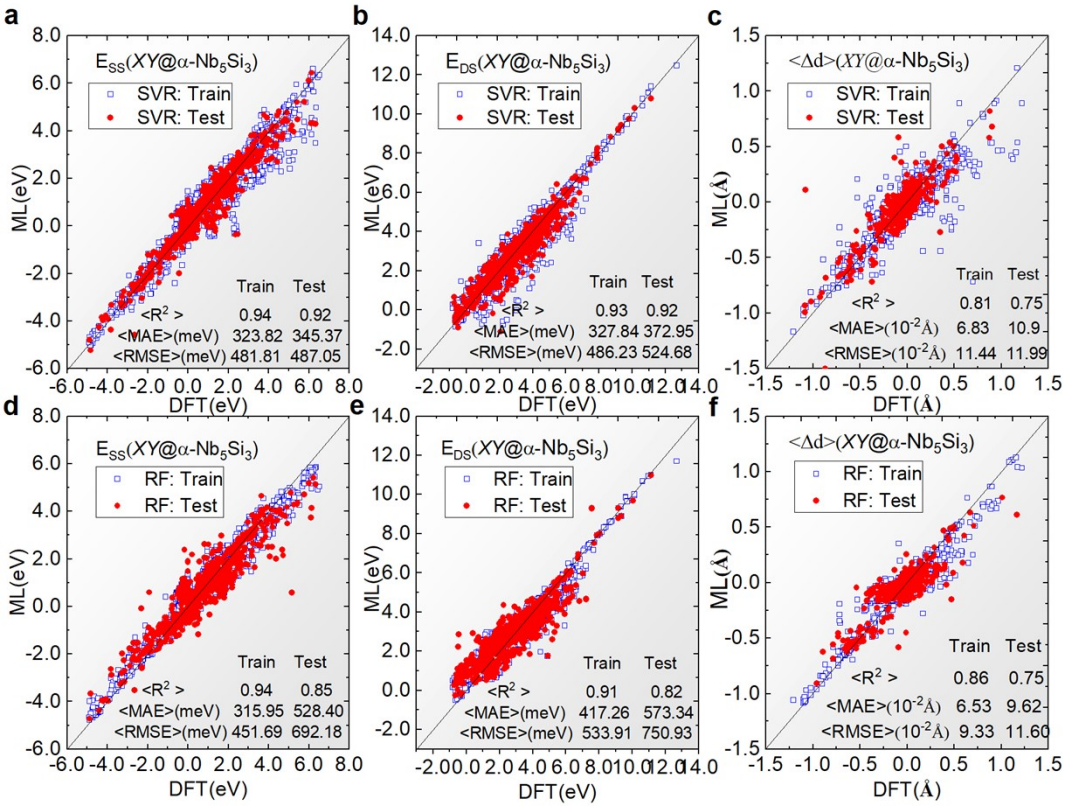

Fig. S5 (a, d) E_{SS}, (b, e) E_{DS}, (c, f) <Δd> of α-Nb_5Si_3 predicted by the CE_{NN} models with SVR and RF algorithms with $^{\omega_j}= 1/\sqrt{r}$.

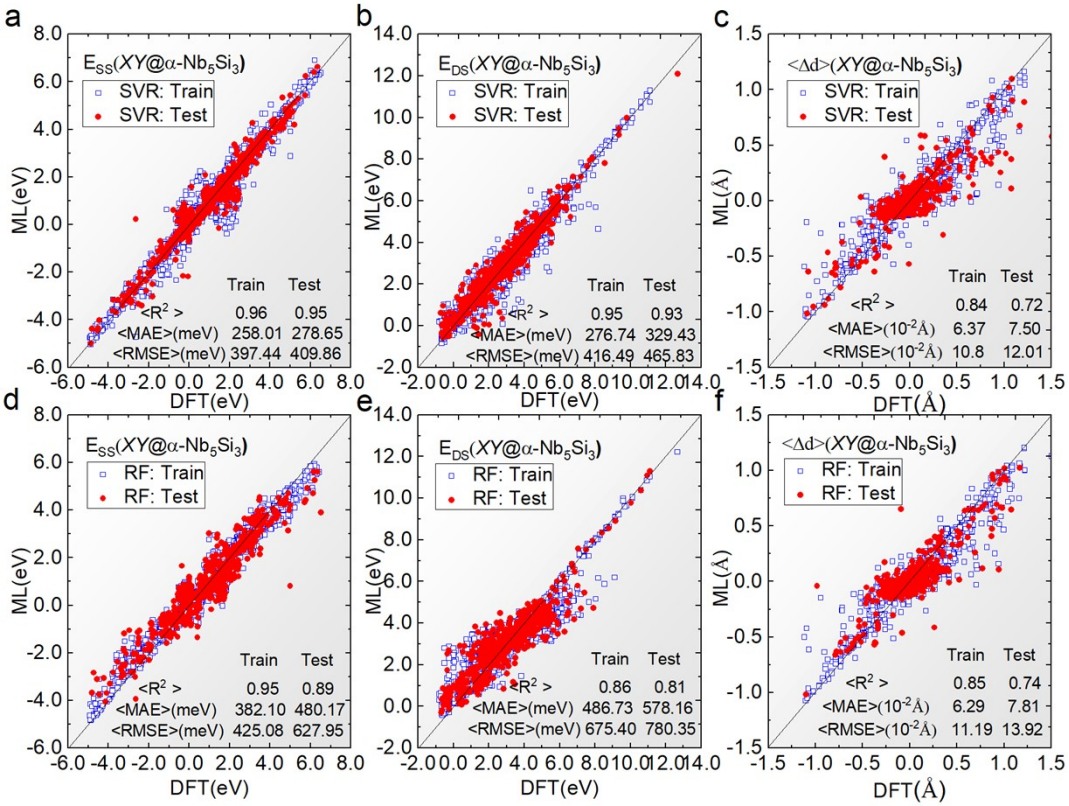

Fig. S6 (a, d) E_{SS}, (b, e) E_{DS}, (c, f) <Δd> of α-Nb_5Si_3 predicted by the CE_{AET} models with SVR and RF algorithms with $^{\omega_j}$= 1/*r*.

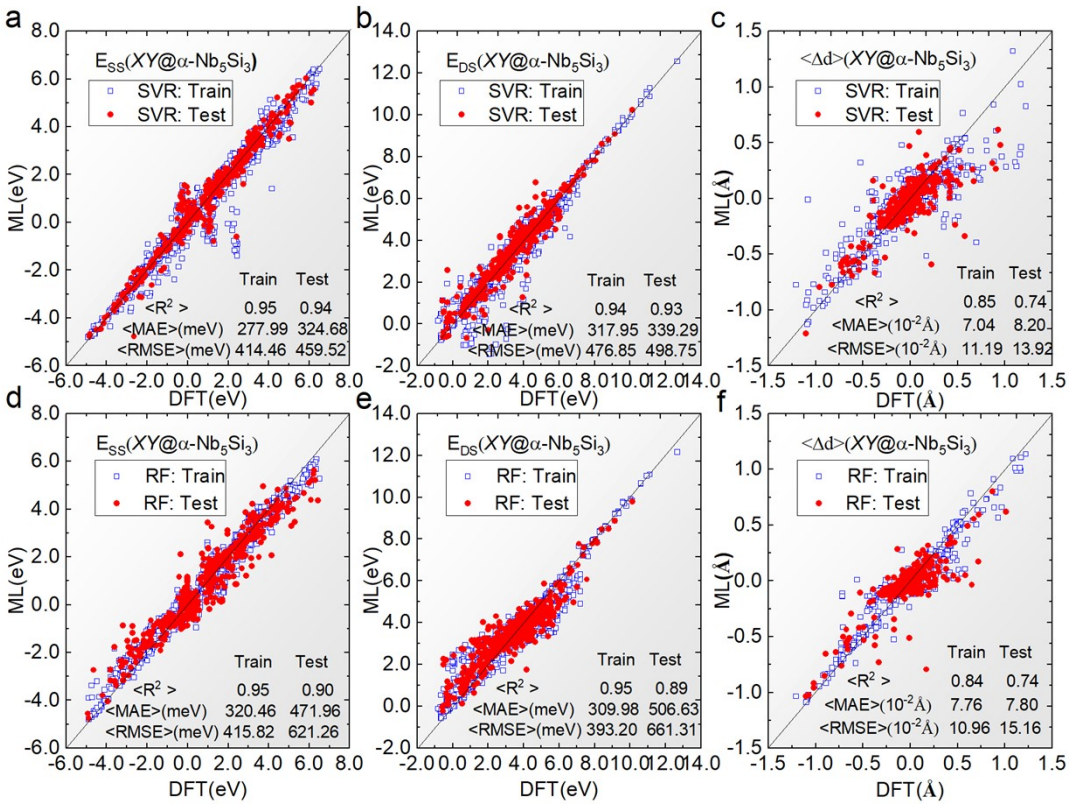

Fig. S7 (a, d) E_{SS}, (b, e) E_{DS}, (c, f) <Δd> of α-Nb_5Si_3 predicted by the CE_{AET} models with SVR and RF algorithms with $^{\omega_j}= 1/\sqrt{r}$.

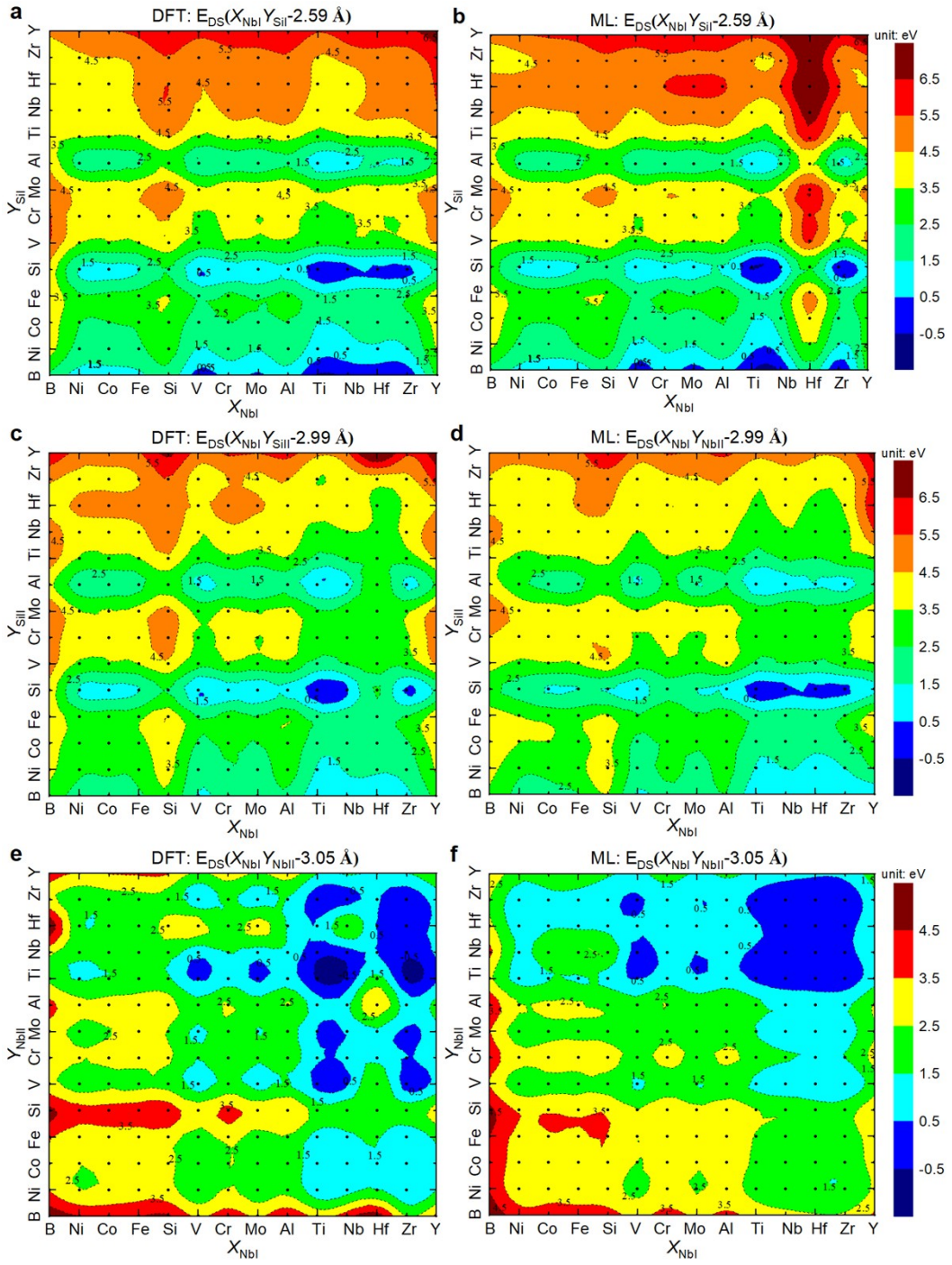

Fig. S8 Heat maps of E_{DS} projected on the substitution pair sites containing Nb_I in α-Nb_5Si_3 predicted by (a, c, e) DFT and (b, d, f) ML, respectively.

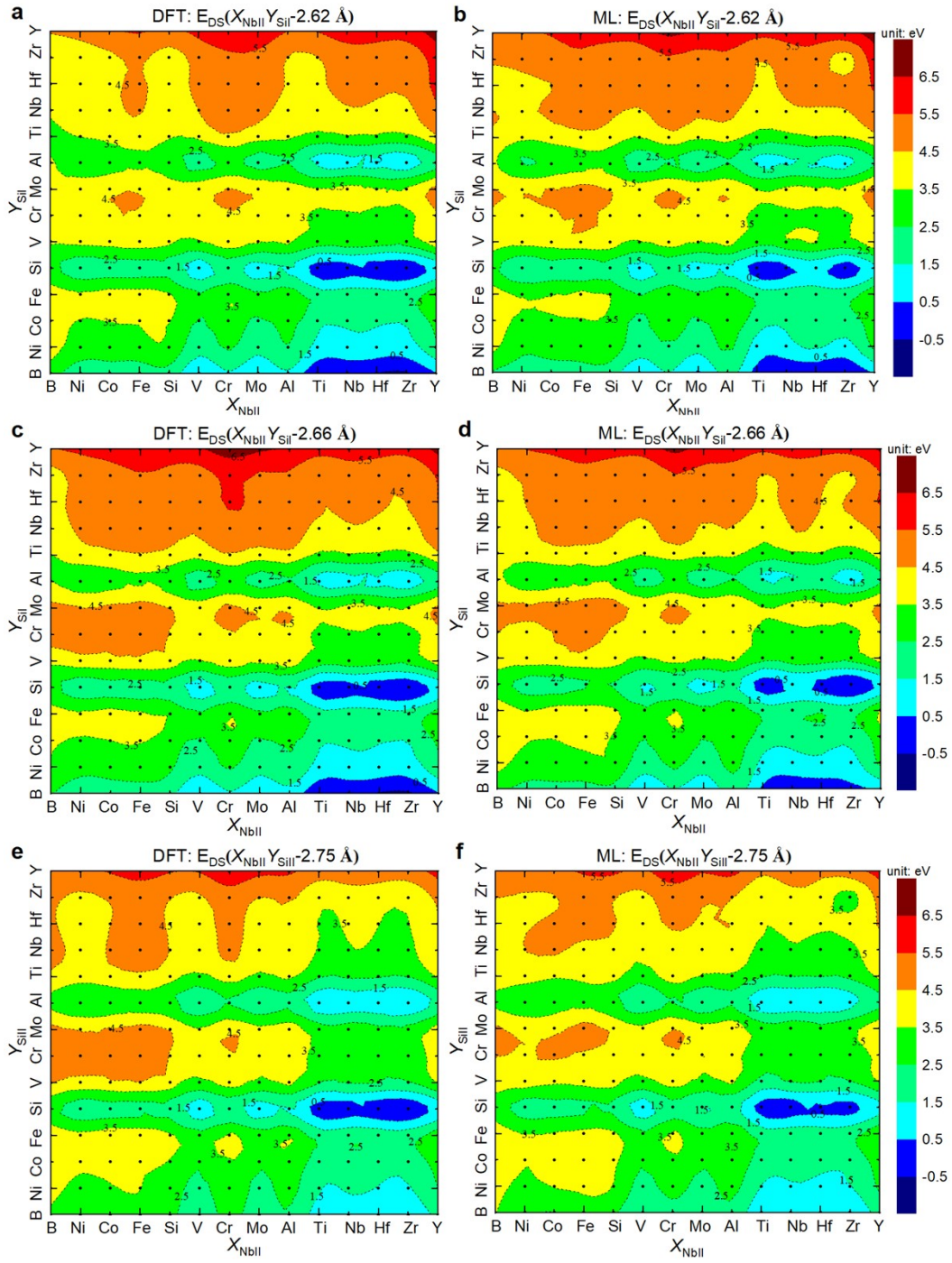

Fig. S9 Heat maps of E_{DS} projected on the substitution pair sites containing Nb_{II} in α-Nb_5Si_3 predicted by (a, c, e) DFT and (b, d, f) ML, respectively.

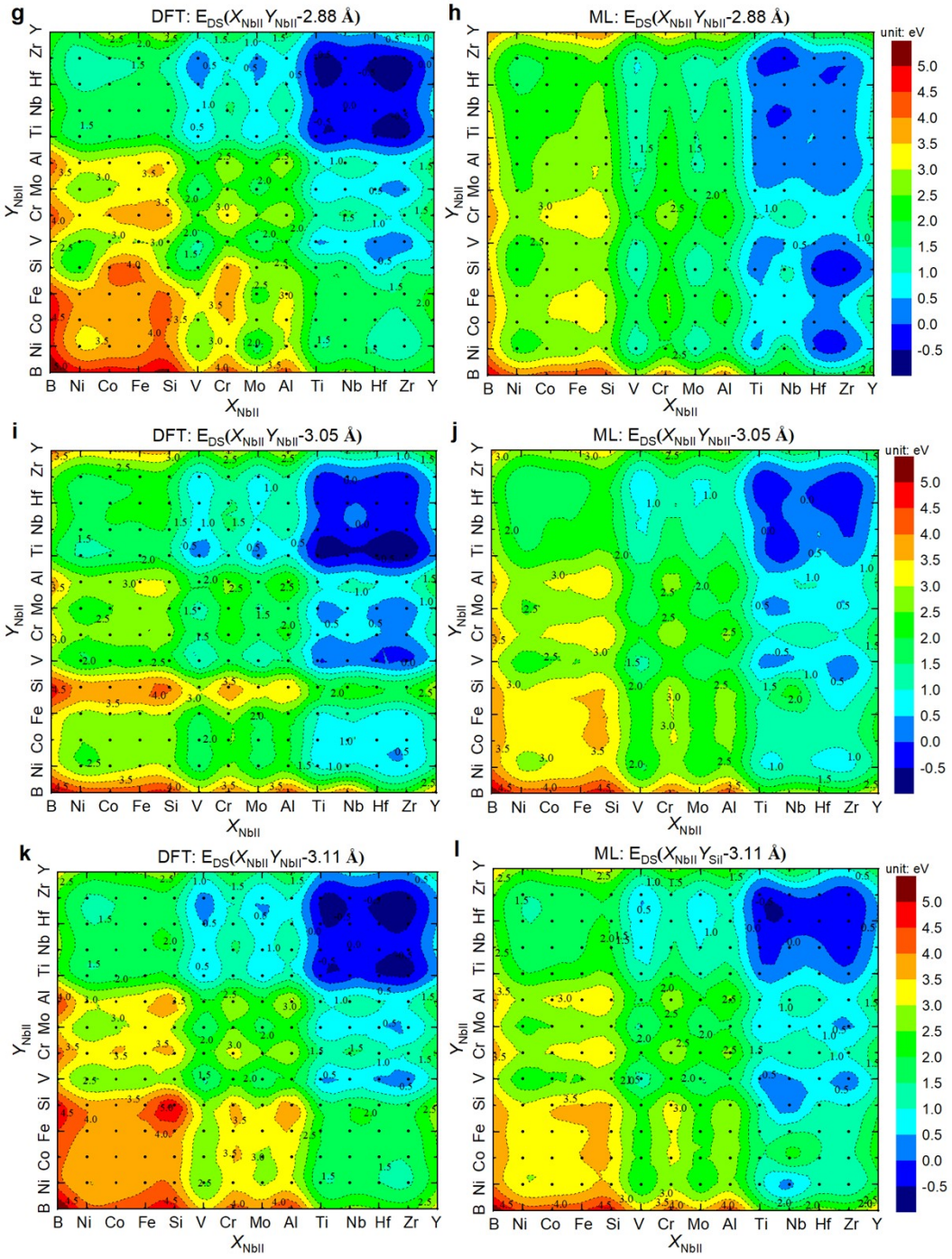

Fig. S9 (continued) Heat maps of E_{DS} projected on the substitution pair sites containing Nb_{II} in α-Nb_5Si_3 predicted by (g, i, k) DFT and (h, j, l) ML,

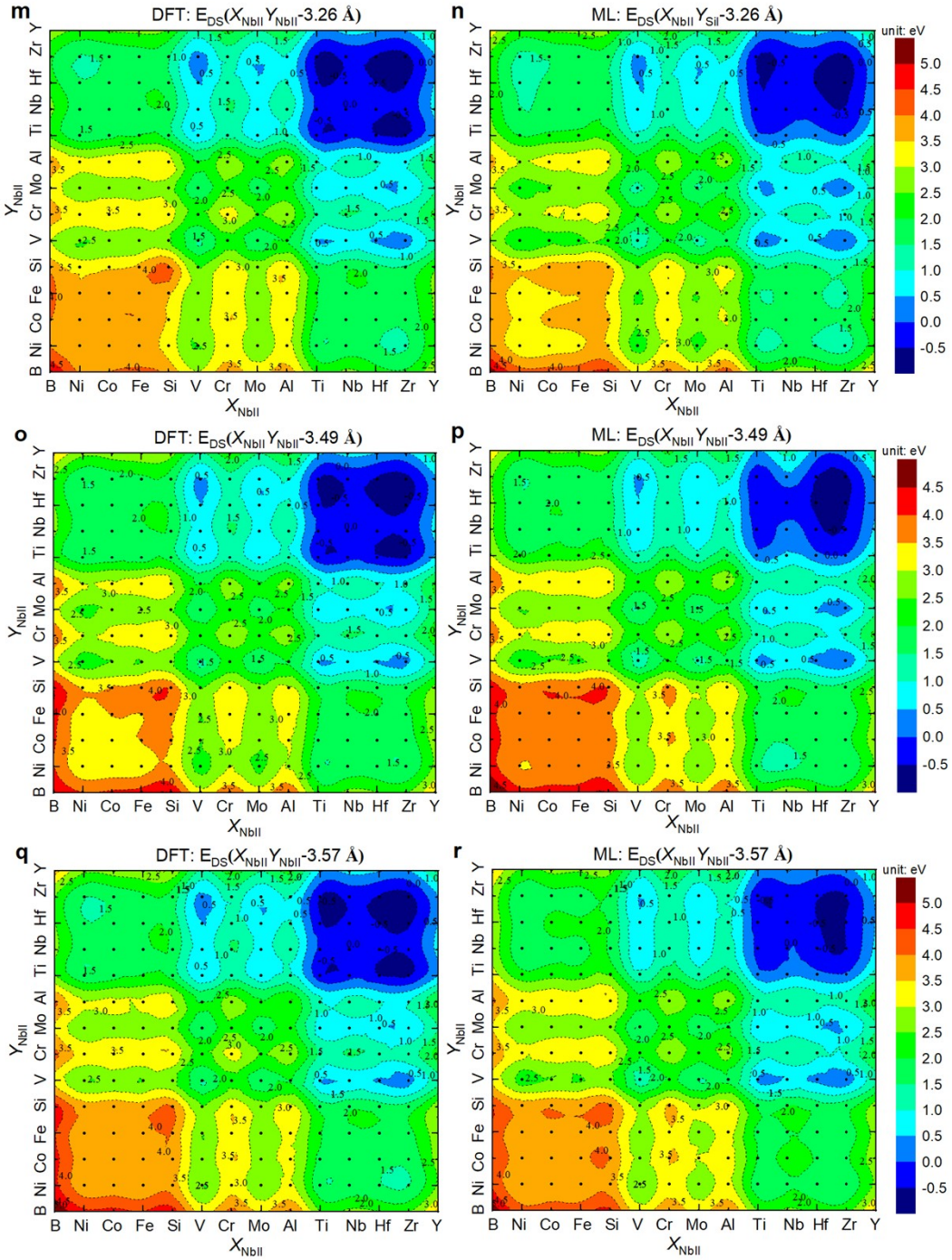

Fig S9 (continued) Heat maps of E_{DS} projected on the substitution pair sites containing Nb_{II} in α-Nb_5Si_3 predicted by (m, o, q) DFT and (n, p, r) ML, respectively.

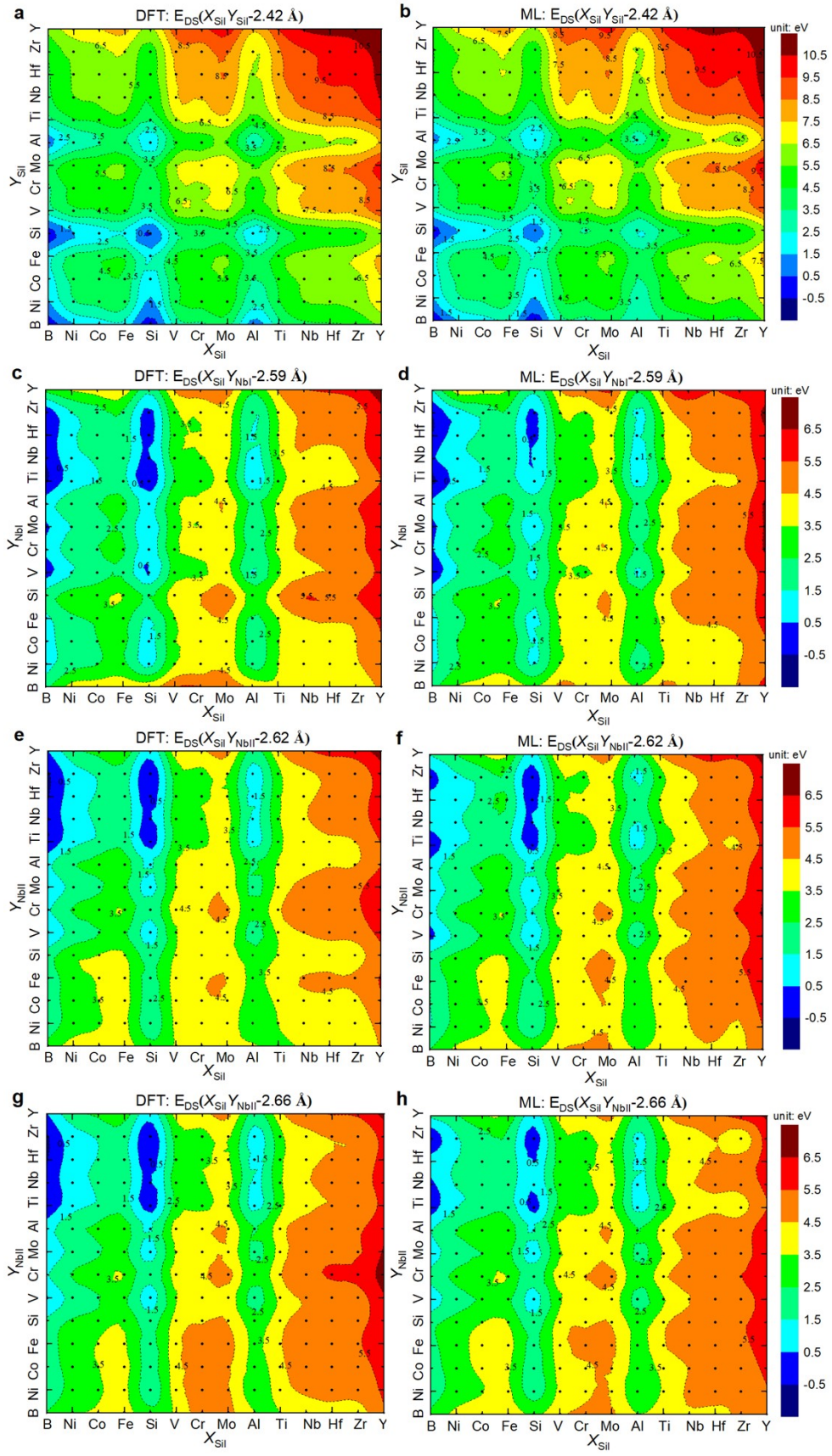

Fig. S10 Heat maps of E_{DS} projected on the substitution pair sites containing Si_I in α-Nb_5Si_3 predicted by (a, c, e, g) DFT and (b, d, f, h) ML, respectively.

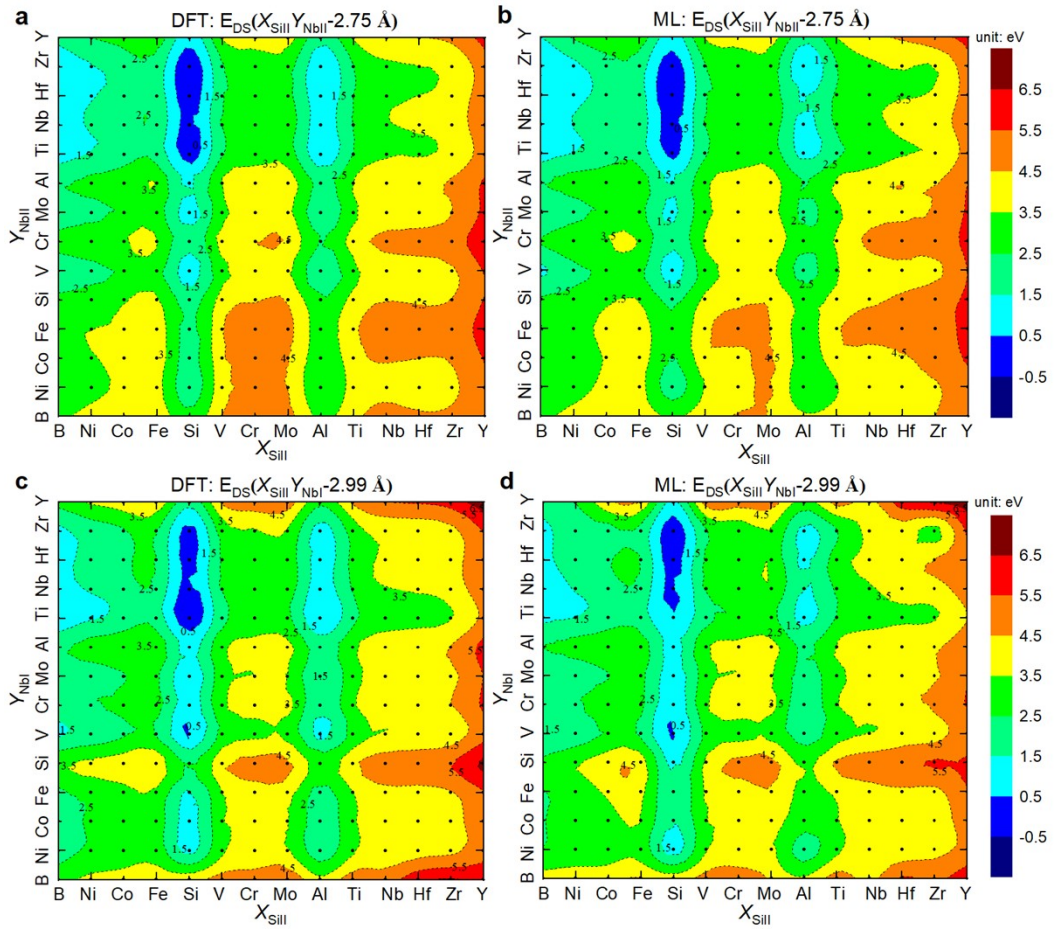

Fig. S11 Heat maps of E_{DS} projected on the substitution pair sites containing Si_{II} in α-Nb_5Si_3 predicted by (a, c) DFT and (b, d) ML, respectively.

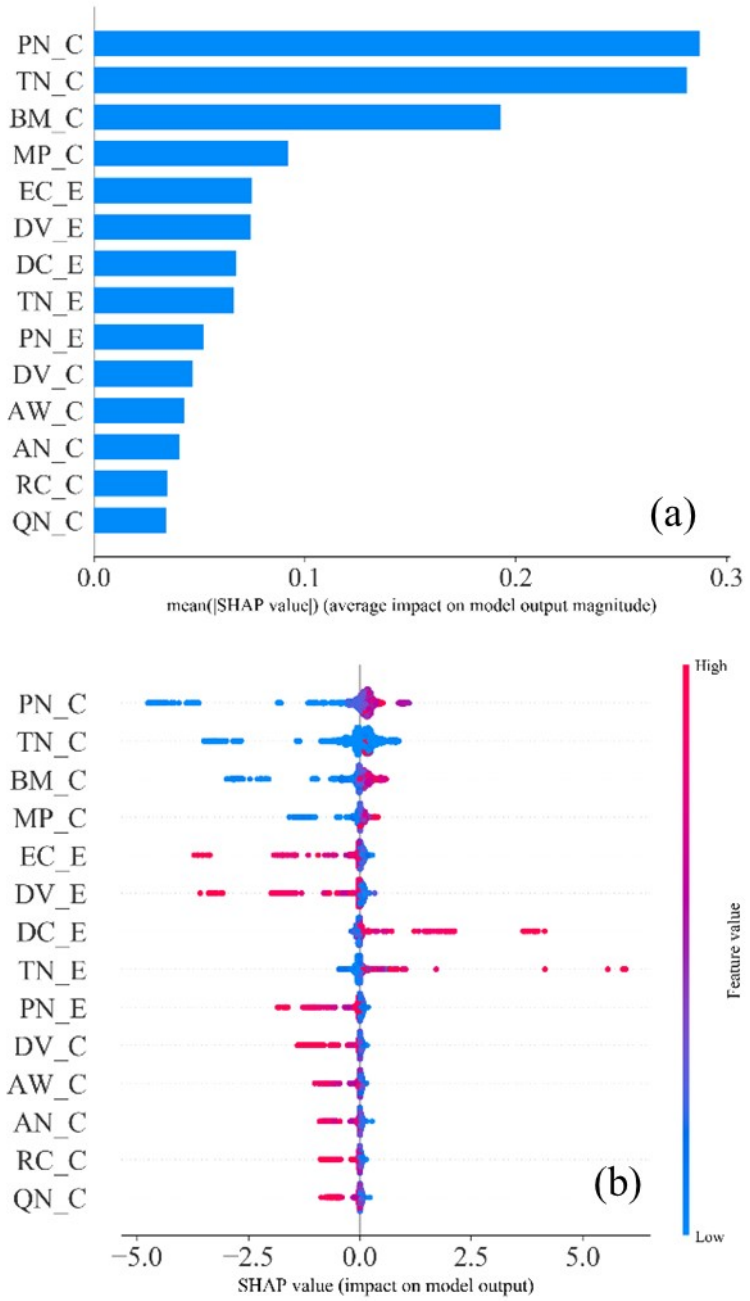

Fig. S12 Feature analysis of E_{DS} in α-Nb_5Si_3 based on SHAP method. (a) Feature importance ranked by SHAP. (b) SHAP value distribution of different samples.

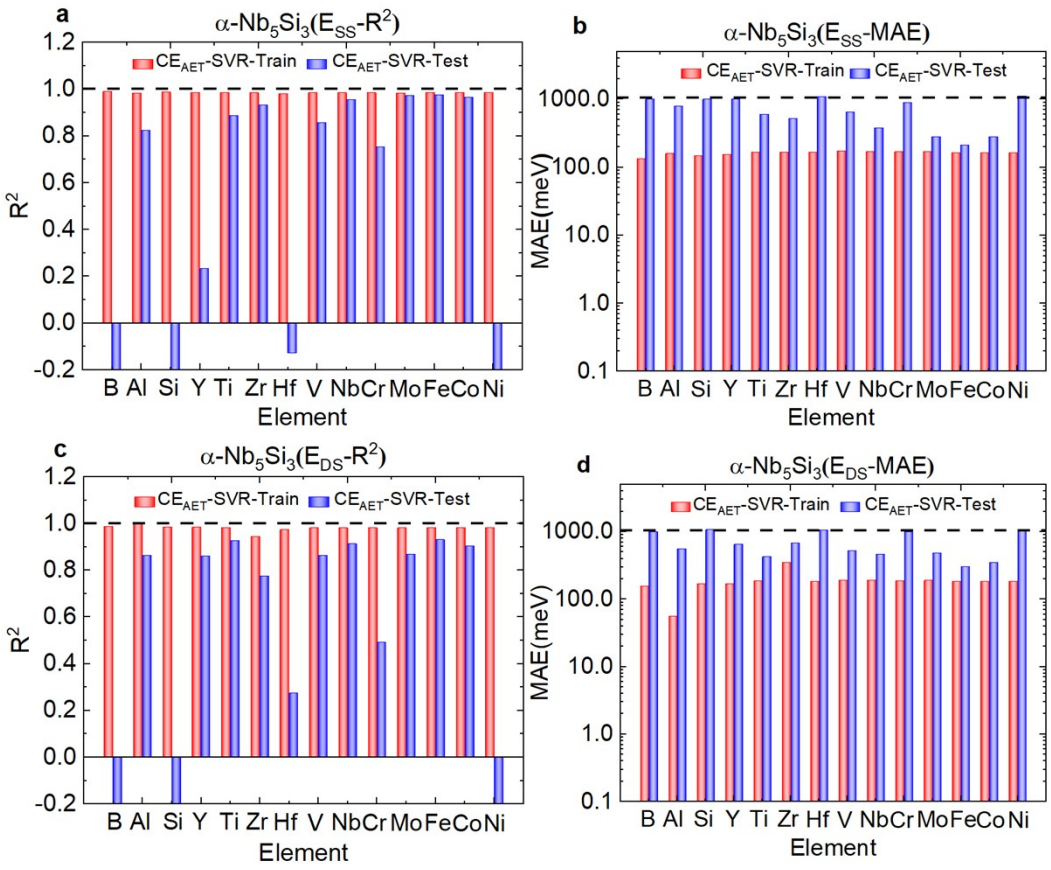

Fig S13 Performance metrics of CE_{AET}-SVR ML models in the leave-p-out cross validation for each of the 14 elements in α-Nb_5Si_3 phase: (a, c) R^2 and (b, d) MAE of E_{SS} and E_{DS}, respectively. The R^2 plot were truncated to -0.2. Both the training and test dataset results are shown for comparison.

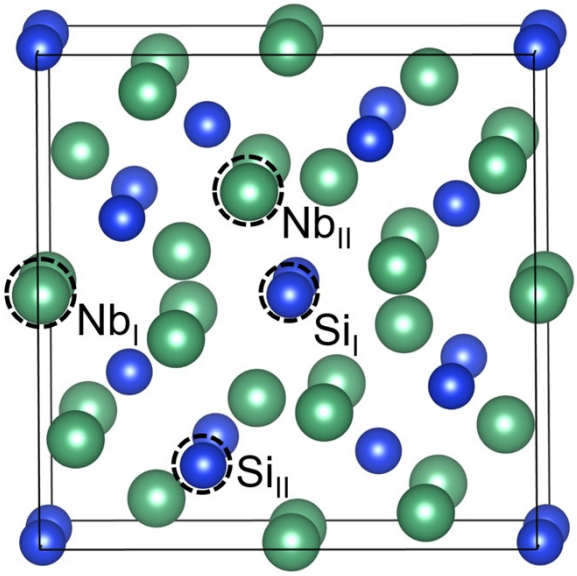

Fig. S14 Conventional cell of β-Nb_5Si_3 crystal structure with four non-equivalent sites Nb_I, Nb_{II}, Si_I, and Si_{II}.

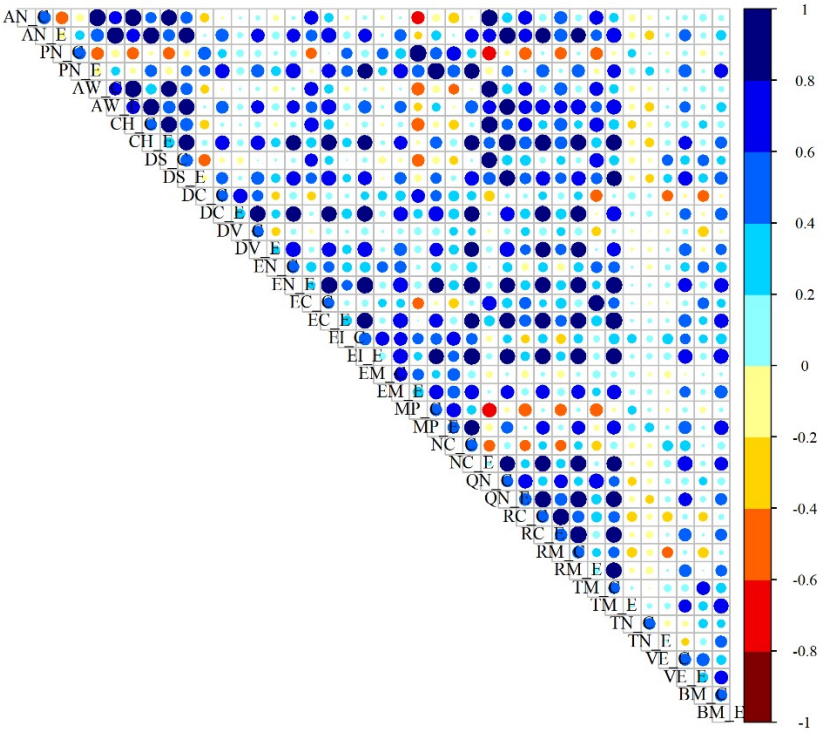

Fig. S15 Correlation matrix of features related to the machine learning prediction of double-site substitution energy in Nb_5Si_3.

References

[1] Kresse G, Furthmuller J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Comp Mater Sci 1996;6:15–50. https://doi.org/10.1016/0927-0256(96)00008-0.

[2] Kresse G, Furthmuller J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. Phys Rev B 1996;54:11169–86. https://doi.org/10.1103/PhysRevB.54.11169.

[3] Kresse G, Joubert D. From ultrasoft pseudopotentials to the projector augmented-wave method. Phys Rev B 1999;59:1758–75. https://doi.org/10.1103/PhysRevB.59.1758.

[4] Blochl PE. Projector augmented-wave method. Phys Rev B 1994;50:17953–79. https://doi.org/10.1103/PhysRevB.50.17953.

[5] Legrain F, Carrete J, Roekeghem AV, Curtarolo S, Mingo N. How the chemical composition alone can predict vibrational free energies and entropies of solids. Chem Mater 2017;29.