Supporting Information

Discrepancies and Error Evaluation Metrics for Machine Learning Interatomic Potentials

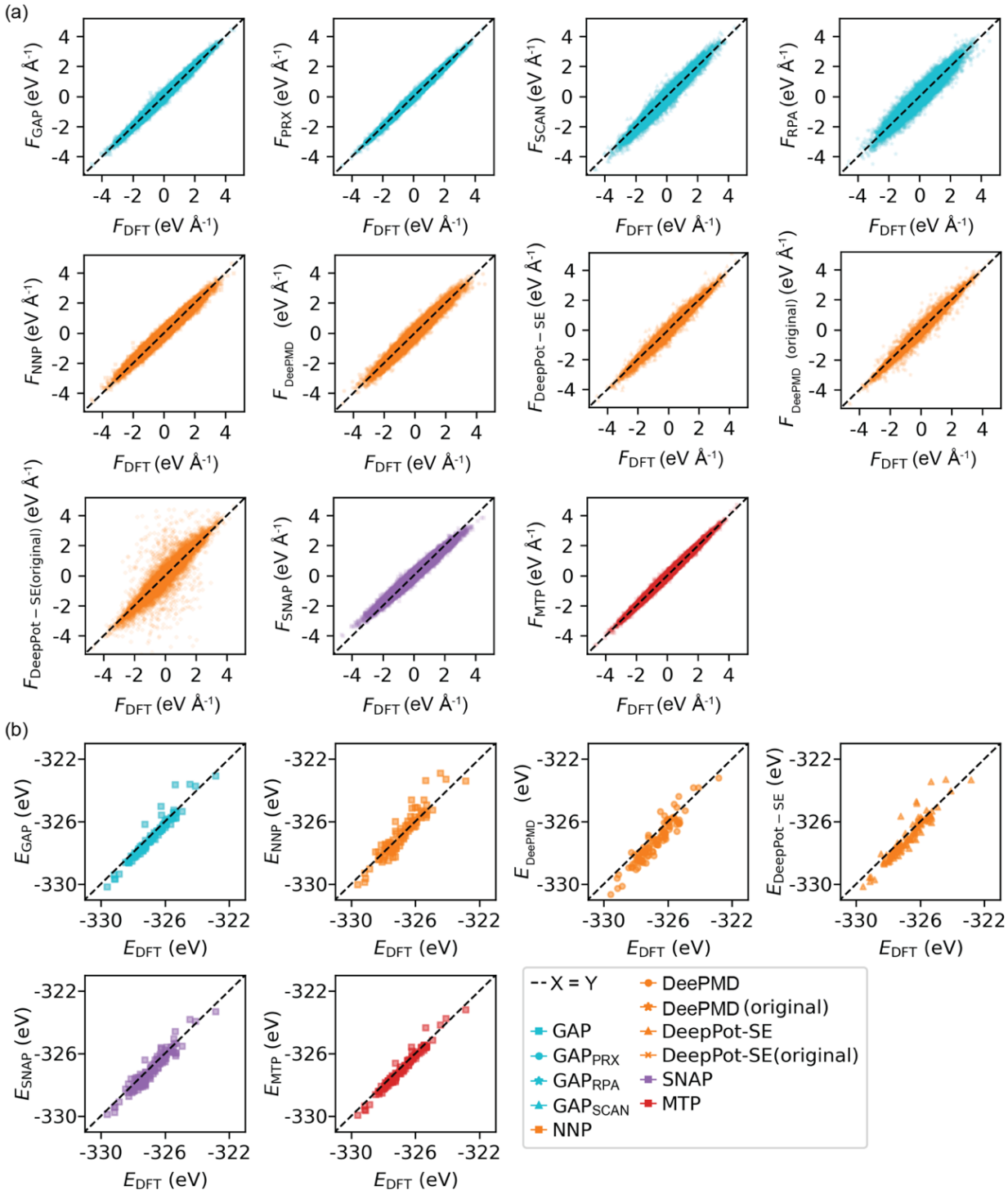
Yunsheng Liu, Xingfeng He, and Yifei Mo*

Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA.

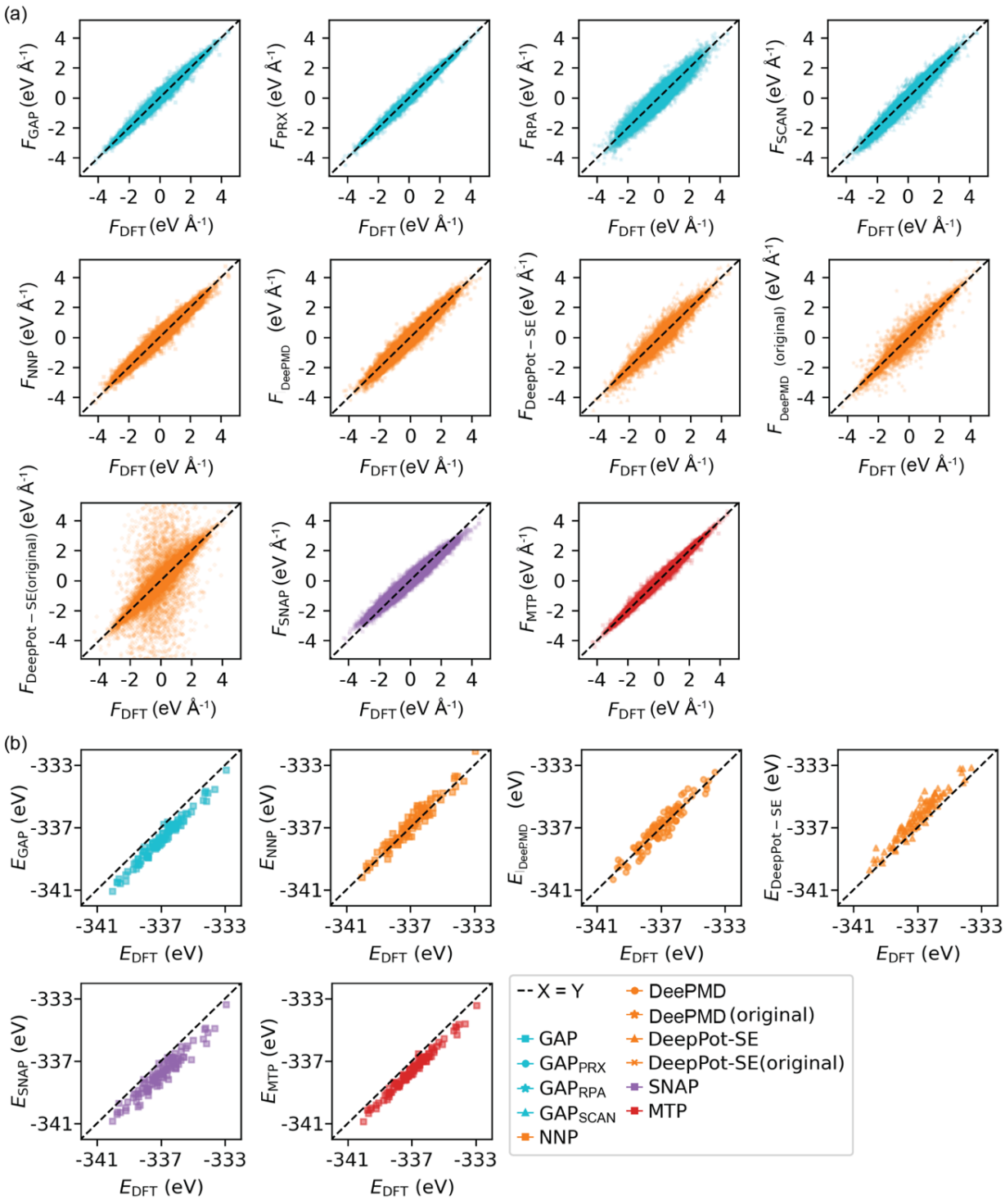
* Email: yfmo@umd.edu

Supplementary Note 1. *Errors of MLIPs.* Here, as shown in **Supplementary Figure 1 and 2,** we perform additional testing for the energy and force errors of machine-learning interatomic potentials (MLIPs) and density functional theory (DFT) calculations using vacancy and interstitial rare-event (RE) testing sets, where the construction of these testing sets is described in the *Error Evaluation of Atomic Forces* in Methods. Here we analyze a total of 11 known MLIPs, including six MLIPs from previous studies as shown in main text (GAP, GAP_{PRX}, NNP, DeePMD, SNAP, and MTP), and GAP_{RPA} and GAP_{SCAN},[1] DeePMD (original) and DeepPot-SE (original) directly retrieved from Ref.[2], from Zuo et al.,[3] Bartók et al.,[4] Deringer et al,[1] and Zhang et al.,[2] and the reproduced DeepPot-SE retrained from the same training data and optimization scheme, as described in Methods.

Supplementary Table 1 shows the energy and force performances of 12 RE-enhanced MLIPs on vacancy- and interstitial- RE testing sets. The six interstitial-enhanced MLIPs are the representative MLIPs obtained through our RE-enhanced workflow in *Optimizing Enhanced MLIPs* in Methods and in main text. We obtain the root-mean-squared errors (RMSEs) of the energies and forces, σ_E^{Test} and σ_F^{Test}, predicted by MLIPs (Supplementary Table 1).[2–4] RMSEs of the energies and forces on vacancy-RE testing set are labelled $\sigma_E^{Vac\ Test}$ and $\sigma_F^{Vac\ Test}$, while those on interstitial-RE testing set are labelled $\sigma_E^{Int\ Test}$ and $\sigma_F^{Int\ Test}$, respectively. Most MLIPs show small averaged errors for atomic forces of all atoms in the dataset where σ_F^{Test} are mostly within the range between 0.1 to 0.4 eV $Å^{-1}$, consistent with the force RMSEs range 0.1 – 0.35 eV $Å^{-1}$ in original publications[1–4], except for DeepPot-SE (original), which may be because its original training data lacks of similar configurations in the testing dataset. All RMSEs of energies and forces on training data, σ_E^{Train} and σ_F^{Train}, are comparable with the values reported in previous publications.[1–4]



Supplementary Figure 1. Comparison of the (a) atomic forces and (b) energies predicted by the known MLIPs to DFT calculations for the vacancy-RE testing set.



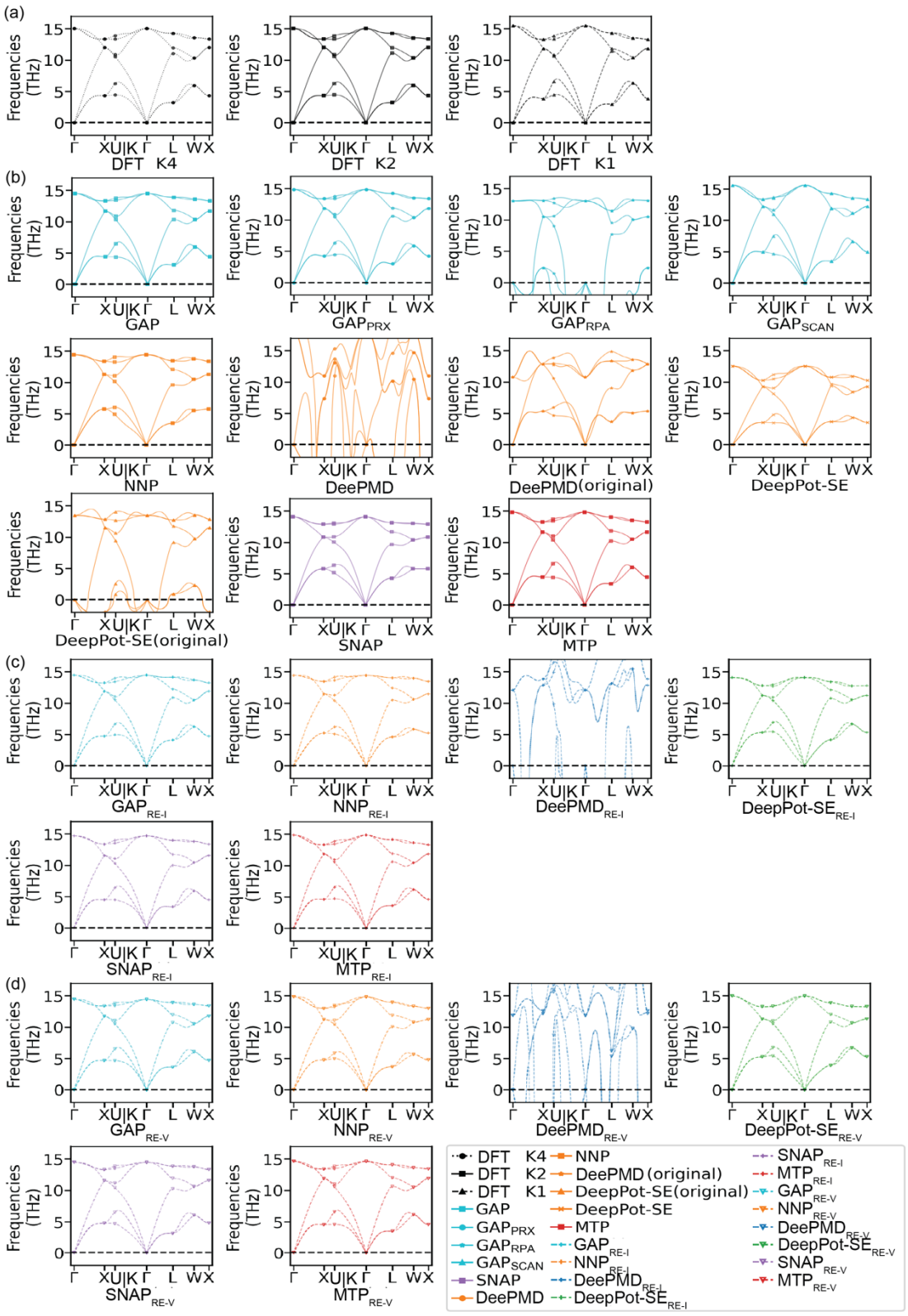
Supplementary Figure 2. Comparison of the (a) atomic forces and (b) energies predicted by known MLIPs to DFT calculations for the interstitial-RE testing set.

Supplementary Table 1. Summary of MLIP errors. RMSEs of energies on the training data, on the testing data, and of atomic forces on the testing data, using the vacancy-RE testing set and interstitial-RE testing set. The training energy RMSEs, σ_E^{Train}, are based on the original training dataset in previous studies.[1–4]

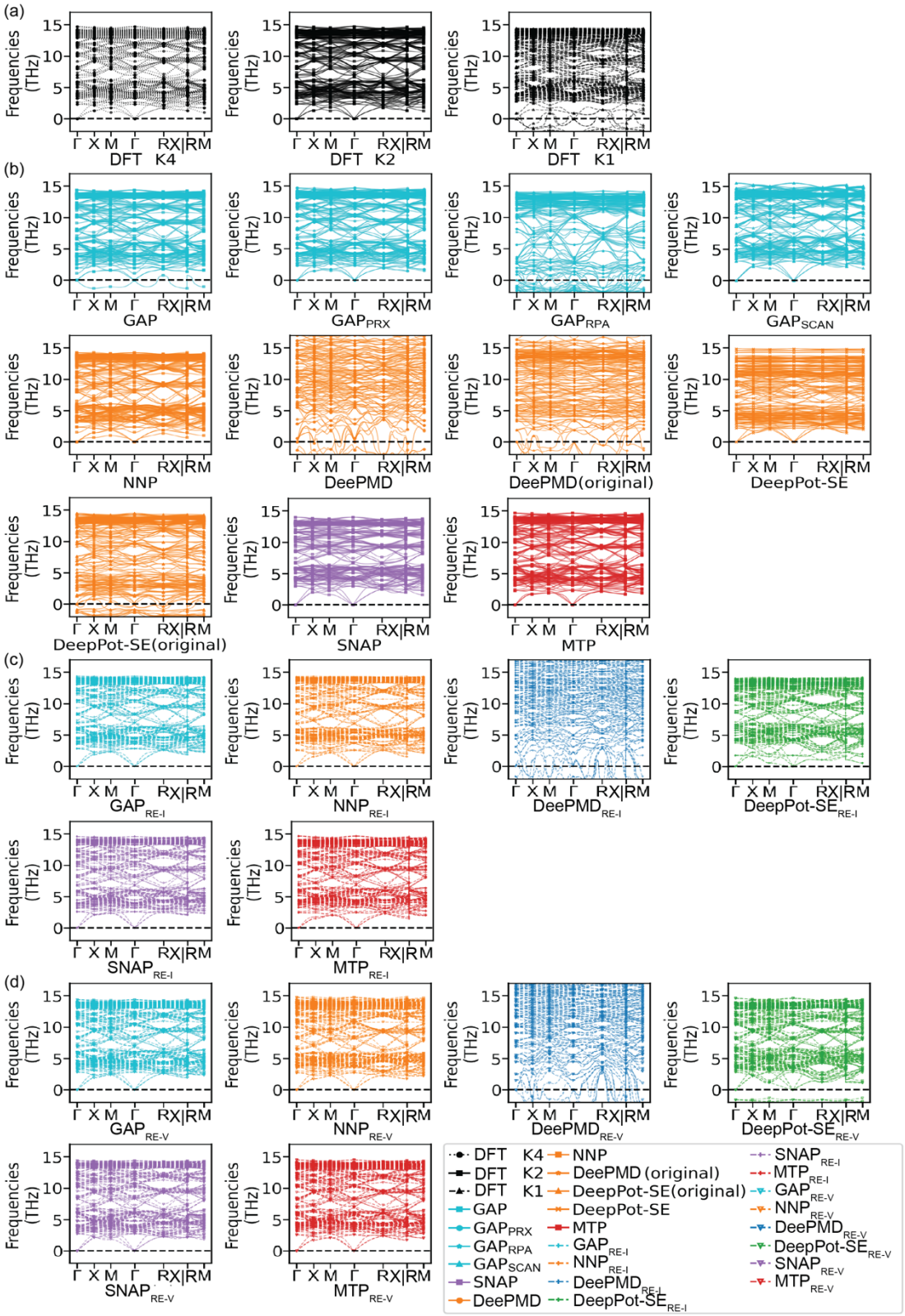
	Models	RMSE σ_E^{Train} (meV atom^{-1})	RMSE $\sigma_E^{Vac\ Test}$ (meV atom^{-1})	RMSE $\sigma_F^{Vac\ Test}$ (eV Å^{-1})	RMSE $\sigma_E^{Int\ Test}$ (meV atom^{-1})	RMSE $\sigma_F^{Int\ Test}$ (eV Å^{-1})	Ref.
Known MLIPs trained in previous studies	GAP	5	6	0.12	13	0.15	3
	GAP_{PRX}	1	N/A	0.10	N/A	0.12	4
	GAP_{RPA}	46	N/A	0.38	N/A	0.38	1
	GAP_{SCAN}	3	N/A	0.28	N/A	0.27	1
	NNP	11	8	0.20	6	0.22	3
	DeePMD	5	10	0.24	7	0.26	
	DeepPot-SE	5	8	0.19	12	0.26	
	DeePMD (original)	1	N/A	0.21	N/A	0.29	2
	DeepPot-SE (original)	0.3	N/A	0.51	N/A	1.43	2
	MTP	3	5	0.10	10	0.12	3
	SNAP	8	7	0.25	14	0.26	3
Vacancy-enhanced MLIPs	GAP_{RE-V}	3	3	0.09	8	0.13	
	NNP_{RE-V}	10	6	0.16	12	0.21	
	$DeePMD_{RE-V}$	11	6	0.14	12	0.24	
	$DeepPot\text{-}SE_{RE-V}$	7	5	0.14	14	0.19	
	$SNAP_{RE-V}$	12	5	0.15	5	0.18	
	MTP_{RE-V}	1	1	0.07	12	0.12	
Interstitial-enhanced MLIPs	GAP_{RE-I}	3	4	0.10	3	0.10	
	NNP_{RE-I}	10	6	0.17	5	0.17	
	$DeePMD_{RE-I}$	14	6	0.22	6	0.19	
	$DeepPot\text{-}SE_{RE-I}$	9	7	0.16	4	0.16	
	$SNAP_{RE-I}$	13	6	0.15	5	0.16	
	MTP_{RE-I}	2	4	0.09	3	0.09	
Reported MLIP errors from previous studies	GAP	5	4	0.12			3
	GAP_{PRX}	/	/	0.1[a]			4
	NNP	11	10	0.17			3
	MTP	3	3	0.09			3
	SNAP	8	8	0.34			3
	DeePMD	1	/	0.03			2
	DeepPot-SE	0.2	/	0.04			2

[a] We used the reported estimation of force RMSE for GAP_{PRX} from Bartók et. al.[4] here.

** Energy errors of testing data on GAP_{PRX}, GAP_{RPA}, GAP_{SCAN}, DeePMD (original), and DeepPot-SE (original) are not evaluated (N/A) since they were trained using different DFT functionals or **k**-point meshes in previous publications, which cannot be directly compared with our DFT calculations.



Supplementary Figure 3. Comparison of the phonon dispersion spectrums of the bulk Si calculated by (a) DFT calculations, (b) original MLIPs, (c) interstitial-enhanced MLIPs, and (d) vacancy-enhanced MLIPs.

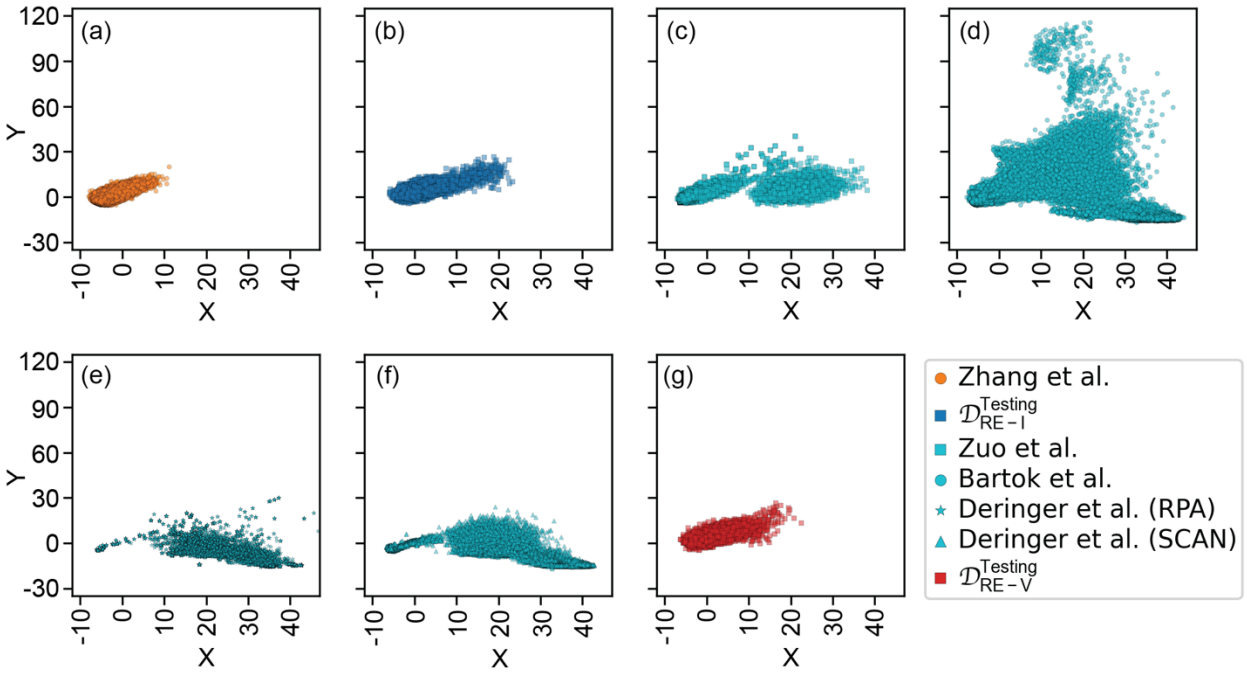


Supplementary Figure 4. Comparison of the phonon dispersion spectrums of the Si supercell with a single vacancy calculated from (a) DFT calculations, (b) the known MLIPs, (c) the interstitial-enhanced MLIPs, and (d) the vacancy-enhanced MLIPs.

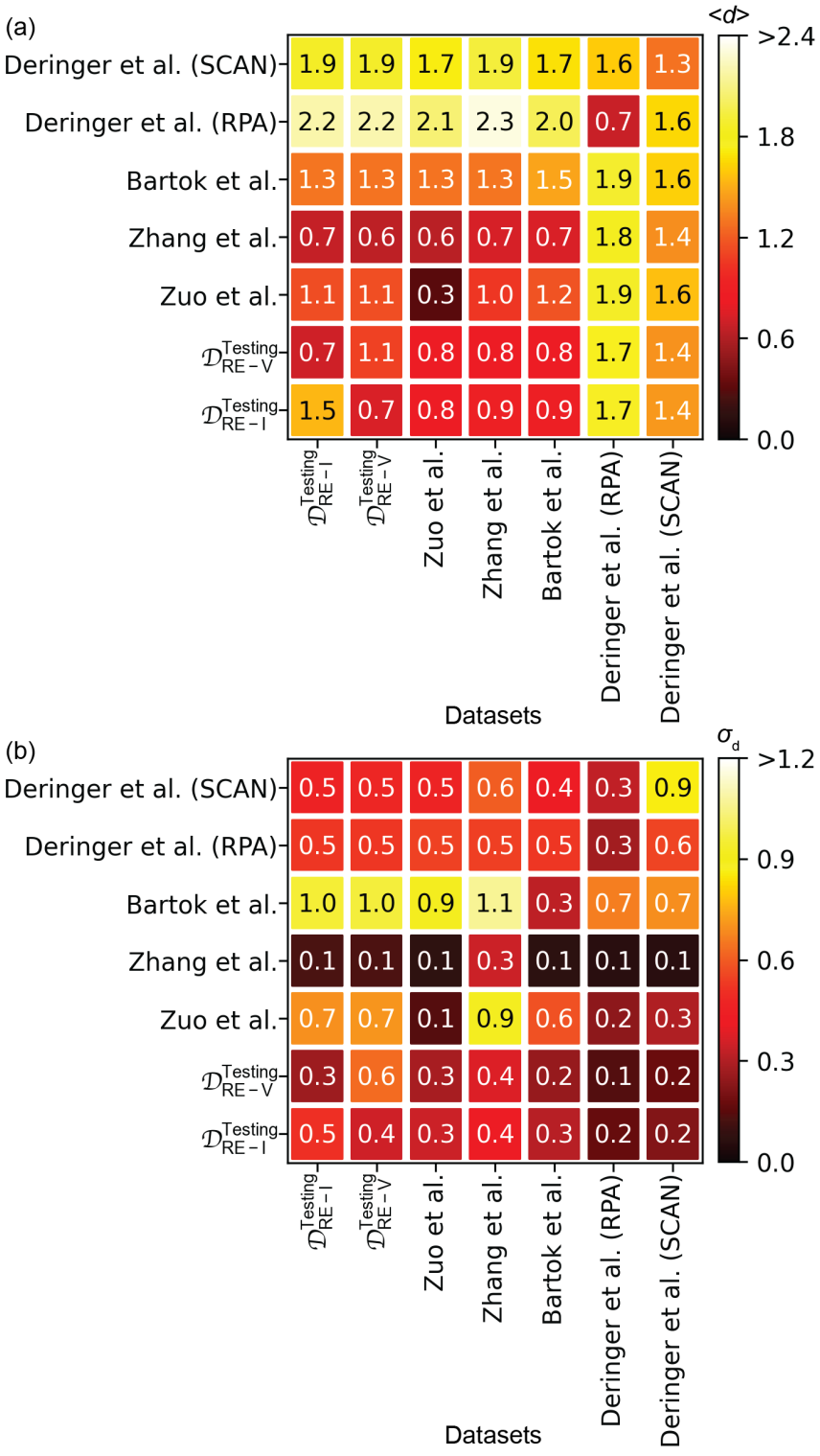
Supplementary Note 2. *Comparing atomic environments of training datasets.* Here we quantify and compare the atomic environments in different testing datasets and training datasets from previous studies and this study. The Smooth Overlap of Atomic Positions (SOAP) descriptor is employed using the same hyperparameter set and following the same approach as described in the Methods and Ref. [3], and has a feature vector of 325 elements for each atomic environment. To visualize the atomic environments and keep the computation process consistent for different datasets, we use all 657451 SOAP descriptors (17785 for GAP_{SCAN} training data in Deringer et al. (SCAN),[1] 9498 for GAP_{RPA} training data in Deringer et al. (RPA),[1] 171815 for GAP_{PRX} training data in Bartok et al.,[4] 13553 for dataset in Zuo et al.,[3] 432000 for DeePMD (orginal) training data in Zhang et al.,[2] 6500 for $\mathcal{D}_{\text{RE-I}}^{\text{Testing}}$, and 6300 for $\mathcal{D}_{\text{RE-V}}^{\text{Testing}}$) here and employ the following steps for dimension reduction and visualization. Firstly, we standardize all SOAP vectors by subtracting the mean and then scaling them by the standard deviation. Then, to visualize the distributions of atomic environments, we perform principal component analysis (PCA) on all standardized descriptors and plot the data distributions using the 1st and the 2nd principal components. We use the two principal components generated here to plot the atomic environments in Fig. 1a, b, and c in the main text (Supplementary Figure 5).

To numerically quantify the differences of all atomic environments across the datasets, we first calculate a central average SOAP vector for each dataset by averaging the SOAP vector of all atom environments in the corresponding dataset. We then calculate the Euclidean distances d between the atomic environments in the given dataset (y axis) and the central SOAP vectors of the target datasets (x axis) as shown in **Supplementary Figure 6**. Thus, the differences of the atomic environments within the same dataset are quantified by the average and the standard deviation of the distances d to its central SOAP vector (the diagonal elements in Supplementary Figure 6). The

differences between two different datasets are calculated using the same method shown as the non-diagonal elements in Supplementary Figure 6.



Supplementary Figure 5. Comparison of atomic environments using the first two principal components of PCA based on the SOAP descriptors for the training datasets adopted in (a) Zhang et al. (orange dots), (c) Zuo et al. (cyan squares), (d) Bartok et al. (cyan dots), (e) RPA in Deringer et al. (cyan stars), (f) SCAN in Deringer et al. (cyan triangles), and the testing datasets (b) $\mathcal{D}_{\mathrm{RE\text{-}I}}^{\mathrm{Testing}}$ (blue squares) and (g) $\mathcal{D}_{\mathrm{RE\text{-}V}}^{\mathrm{Testing}}$ (red squares).



Supplementary Figure 6. The differences of atomic environments within and between the datasets. (a) The average Euclidean distances, <*d*>, and (b) the standard deviation of Euclidean distances, σ_d, for each pair of datasets (non-diagonal elements) and within the datasets (diagonal elements). The target datasets (*x*-axis) use their central SOAP vectors to calculate *d* with all the atomic environments in the given datasets (*y*-axis).

Supplementary Note 3. *Vacancy-enhanced MLIPs.* In addition to interstitial-enhanced MLIPs shown in main text, we perform the RE-enhanced workflow (Methods) for vacancies, and we construct vacancy-enhanced training dataset and select six representative vacancy-enhanced MLIPs. Following the same procedure of constructing interstitial-enhanced dataset in *Generating RE-Enhanced Training Dataset*, Methods, vacancy-enhanced training dataset is constructed by replacing 120 structures in the original dataset in Ref.[3] by the snapshots containing vacancy REs that are selected by the vacancy *k*-means clustering model described in main text and Methods. All processes and protocols, including the replaced structures and the sampling approach for vacancy snapshots, remain the same as described in Methods.

Trained by the vacancy-enhanced training data, we select 54 good vacancy-enhanced MLIPs, using the enhanced validation set, the optimization processes, and the eight selection criteria same as in the *Optimizing Enhanced MLIPs* in Methods. Finally, we calculate the joint force error score $P(\mathcal{D}_{\text{RE-I}}^{\text{EVS}})^2+P(\mathcal{D}_{\text{RE-V}}^{\text{EVS}})^2$ of these 54 selected vacancy-enhanced MLIPs and pick six of them with the highest scores (*Optimizing Enhanced MLIPs* in the Methods) for each model as the representative vacancy-enhanced MLIPs.

Supplementary Note 4. *Diffusional properties.* The diffusional properties evaluated from ab initio molecular dynamics (AIMD) and MLIP-MD simulations for all MLIPs (including known and RE-enhanced MLIPs) are summarized in **Supplementary Table 2 and 3**. For NNP (with interstitial) and DeepPot-SE (with either vacancy or interstitial), the crystal structures of Si melt at all temperatures tested (730 – 1600 K). For GAP_{SCAN}, MD simulations with a vacancy show no diffusion at all temperatures tested, and for GAP_{RPA}, the Si structures melt at temperatures 1230 K to 1600 K and show no diffusion below 1230 K. For $NNP_{RE\text{-}I}$, $DeePMD_{RE\text{-}I}$, and $SNAP_{RE\text{-}I}$, MD simulations with a single vacancy show no diffusion at 1000 K.

Supplementary Table 2. The diffusivities at 1000 K, the pre-exponential factors D_0 and activation energies E_a from fitted Arrhenius relations for single-interstitial (superscript I) from AIMD and MLIP-MD simulations.

	Models	D_{1000K}^{I} ($cm^2\ s^{-1}$)	Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	D_0^{I} ($cm^2\ s^{-1}$)	Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	E_a^{I} (eV)
AIMD	K2	6.6×10^{-7}	[8.9×10^{-7}, 4.4×10^{-7}]	2.3×10^{-5}	[4.1×10^{-5}, 1.3×10^{-5}]	0.30 ± 0.06
	K1	4.5×10^{-7}	[5.7×10^{-7}, 3.4×10^{-7}]	7.7×10^{-6}	[1.3×10^{-5}, 4.5×10^{-6}]	0.25 ± 0.05
Known MLIPs trained from previous studies	GAP	4.6×10^{-7}	[5.8×10^{-7}, 3.4×10^{-7}]	5.9×10^{-5}	[9.4×10^{-5}, 3.7×10^{-5}]	0.42 ± 0.05
	GAP_{PRX}	1.8×10^{-6}	[2.2×10^{-6}, 1.4×10^{-6}]	8.3×10^{-6}	[1.4×10^{-5}, 4.9×10^{-6}]	0.12 ± 0.06
	GAP_{RPA}	1.8×10^{-6}	[2.1×10^{-6}, 1.5×10^{-6}]	6.1×10^{-5}	[1.4×10^{-4}, 2.6×10^{-5}]	0.29 ± 0.06
	GAP_{SCAN}	6.7×10^{-7}	[8.7×10^{-7}, 4.7×10^{-7}]	1.6×10^{-5}	[2.2×10^{-5}, 1.1×10^{-5}]	0.23 ± 0.04
	NNP	/	/	/	/	/
	DeePMD	2.3×10^{-6}	[2.5×10^{-6}, 2.2×10^{-6}]	1.0×10^{-4}	[1.1×10^{-4}, 9.2×10^{-5}]	0.24 ± 0.01
	DeepPot-SE	7.7×10^{-7}	[8.7×10^{-7}, 6.8×10^{-7}]	2.6×10^{-5}	[3.1×10^{-5}, 2.2×10^{-5}]	0.31 ± 0.02
	DeePMD (original)	3.1×10^{-6}	[3.4×10^{-6}, 2.9×10^{-6}]	5.5×10^{-6}	[6.6×10^{-6}, 4.5×10^{-6}]	0.04 ± 0.01
	DeepPot-SE (original)	/	/	/	/	/
	MTP	5.5×10^{-7}	[6.3×10^{-7}, 4.7×10^{-7}]	5.7×10^{-5}	[7.1×10^{-5}, 4.6×10^{-5}]	0.42 ± 0.02
	SNAP	1.3×10^{-7}	[1.5×10^{-7}, 1.1×10^{-7}]	6.4×10^{-4}	[1.0×10^{-3}, 4.1×10^{-4}]	0.74 ± 0.05
Vacancy-enhanced MLIPs	$GAP_{RE\text{-}V}$	3.7×10^{-7}	[4.5×10^{-7}, 2.9×10^{-7}]	5.9×10^{-5}	[7.7×10^{-5}, 4.5×10^{-5}]	0.42 ± 0.03
	$NNP_{RE\text{-}V}$	1.4×10^{-6}	[1.6×10^{-6}, 1.2×10^{-6}]	3.6×10^{-6}	[4.1×10^{-6}, 3.2×10^{-6}]	0.08 ± 0.01
	$DeePMD_{RE\text{-}V}$	5.1×10^{-6}	[5.4×10^{-6}, 4.8×10^{-6}]	2.2×10^{-5}	[2.4×10^{-5}, 2.1×10^{-5}]	0.13 ± 0.01
	$DeepPot\text{-}SE_{RE\text{-}V}$	1.0×10^{-6}	[1.1×10^{-6}, 9.1×10^{-7}]	1.5×10^{-5}	[1.7×10^{-5}, 1.3×10^{-5}]	0.22 ± 0.01
	$SNAP_{RE\text{-}V}$	2.3×10^{-6}	[2.5×10^{-6}, 2.0×10^{-6}]	1.4×10^{-5}	[1.7×10^{-5}, 1.2×10^{-5}]	0.17 ± 0.02
	$MTP_{RE\text{-}V}$	4.3×10^{-8}	[6.1×10^{-8}, 2.6×10^{-8}]	4.1×10^{-4}	[7.1×10^{-4}, 2.3×10^{-4}]	0.77 ± 0.07
Interstitial-enhanced MLIPs	$GAP_{RE\text{-}I}$	1.0×10^{-6}	[1.2×10^{-6}, 8.3×10^{-7}]	1.1×10^{-5}	[1.4×10^{-5}, 8.2×10^{-6}]	0.20 ± 0.02
	$NNP_{RE\text{-}I}$	7.8×10^{-7}	[9.2×10^{-7}, 6.4×10^{-7}]	3.2×10^{-5}	[3.9×10^{-5}, 2.7×10^{-5}]	0.33 ± 0.02
	$DeePMD_{RE\text{-}I}$	1.1×10^{-5}	[1.1×10^{-5}, 1.0×10^{-5}]	4.1×10^{-5}	[4.4×10^{-5}, 3.8×10^{-5}]	0.15 ± 0.01
	$DeepPot\text{-}SE_{RE\text{-}I}$	1.9×10^{-6}	[2.0×10^{-6}, 1.7×10^{-6}]	8.5×10^{-6}	[9.5×10^{-6}, 7.6×10^{-6}]	0.14 ± 0.01
	$SNAP_{RE\text{-}I}$	1.2×10^{-6}	[1.3×10^{-6}, 9.7×10^{-7}]	3.1×10^{-5}	[3.7×10^{-5}, 2.6×10^{-5}]	0.29 ± 0.02
	$MTP_{RE\text{-}I}$	1.1×10^{-6}	[1.2×10^{-6}, 9.5×10^{-7}]	1.9×10^{-5}	[2.2×10^{-5}, 1.6×10^{-5}]	0.26 ± 0.02

Supplementary Table 3. The diffusivities at 1000 K, the pre-exponential factors D_0 and activation energies E_a from fitted Arrhenius relations for single-vacancy (superscript V) from AIMD and MLIP-MD simulations.

	Models	D_{1000K}^{V} ($cm^2\ s^{-1}$)	Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	D_0^{V} ($cm^2\ s^{-1}$)	Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	E_a^{V} (eV)
AIMD	K2	5.2×10^{-7}	$[6.9\times10^{-7}, 3.6\times10^{-7}]$	6.0×10^{-6}	$[1.1\times10^{-5}, 3.1\times10^{-6}]$	0.20 ± 0.07
	K1	2.2×10^{-7}	$[2.8\times10^{-7}, 1.6\times10^{-7}]$	7.4×10^{-6}	$[1.5\times10^{-5}, 3.7\times10^{-6}]$	0.31 ± 0.07
Known MLIPs trained from previous studies	GAP	5.8×10^{-7}	$[7.1\times10^{-7}, 4.5\times10^{-7}]$	6.8×10^{-6}	$[1.0\times10^{-5}, 4.6\times10^{-6}]$	0.21 ± 0.04
	GAP_{PRX}	5.6×10^{-7}	$[7.8\times10^{-7}, 3.3\times10^{-7}]$	5.1×10^{-6}	$[1.2\times10^{-5}, 2.1\times10^{-6}]$	0.18 ± 0.10
	GAP_{RPA}	/	/	/	/	/
	GAP_{SCAN}	/	/	/	/	/
	NNP	6.7×10^{-7}	$[7.3\times10^{-7}, 6.1\times10^{-7}]$	5.8×10^{-6}	$[8.0\times10^{-6}, 4.2\times10^{-6}]$	0.19 ± 0.03
	DeePMD	2.8×10^{-6}	$[3.0\times10^{-6}, 2.5\times10^{-6}]$	3.6×10^{-5}	$[4.0\times10^{-5}, 3.3\times10^{-5}]$	0.17 ± 0.01
	DeepPot-SE	9.1×10^{-7}	$[1.0\times10^{-6}, 8.0\times10^{-7}]$	6.3×10^{-6}	$[7.4\times10^{-6}, 5.4\times10^{-6}]$	0.17 ± 0.02
	DeePMD (original)	1.1×10^{-5}	$[1.1\times10^{-5}, 9.9\times10^{-6}]$	4.6×10^{-5}	$[5.1\times10^{-5}, 4.2\times10^{-5}]$	0.16 ± 0.01
	DeepPot-SE (original)	/	/	/	/	/
	MTP	6.0×10^{-7}	$[6.8\times10^{-7}, 5.2\times10^{-7}]$	1.1×10^{-5}	$[1.4\times10^{-5}, 9.3\times10^{-6}]$	0.25 ± 0.02
	SNAP	7.8×10^{-7}	$[8.5\times10^{-7}, 7.1\times10^{-7}]$	5.4×10^{-6}	$[7.0\times10^{-6}, 4.1\times10^{-6}]$	0.16 ± 0.02
Vacancy-enhanced MLIPs	GAP_{RE-V}	3.9×10^{-7}	$[4.7\times10^{-7}, 3.0\times10^{-7}]$	8.7×10^{-6}	$[1.1\times10^{-5}, 6.7\times10^{-6}]$	0.24 ± 0.03
	NNP_{RE-V}	9.4×10^{-7}	$[1.1\times10^{-6}, 7.7\times10^{-7}]$	9.4×10^{-6}	$[1.2\times10^{-5}, 7.3\times10^{-6}]$	0.21 ± 0.03
	$DeePMD_{RE-V}$	3.5×10^{-6}	$[3.7\times10^{-6}, 3.2\times10^{-6}]$	1.2×10^{-5}	$[1.3\times10^{-5}, 1.1\times10^{-5}]$	0.12 ± 0.01
	$DeepPot\text{-}SE_{RE-V}$	7.5×10^{-7}	$[8.3\times10^{-7}, 6.7\times10^{-7}]$	3.6×10^{-6}	$[4.0\times10^{-6}, 3.2\times10^{-6}]$	0.13 ± 0.01
	$SNAP_{RE-V}$	5.7×10^{-7}	$[7.0\times10^{-7}, 4.4\times10^{-7}]$	8.5×10^{-6}	$[1.0\times10^{-5}, 6.9\times10^{-6}]$	0.23 ± 0.02
	MTP_{RE-V}	8.0×10^{-7}	$[9.0\times10^{-7}, 7.0\times10^{-7}]$	6.8×10^{-6}	$[8.1\times10^{-6}, 5.7\times10^{-6}]$	0.19 ± 0.02
Interstitial-enhanced MLIPs	GAP_{RE-I}	9.7×10^{-7}	$[1.2\times10^{-6}, 7.9\times10^{-7}]$	6.4×10^{-6}	$[8.2\times10^{-6}, 4.9\times10^{-6}]$	0.17 ± 0.02
	NNP_{RE-I}	/	/	1.8×10^{-6}	$[2.3\times10^{-6}, 1.4\times10^{-6}]$	0.08 ± 0.03
	$DeePMD_{RE-I}$	/	/	2.7×10^{-5}	$[2.9\times10^{-5}, 2.5\times10^{-5}]$	0.12 ± 0.01
	$DeepPot\text{-}SE_{RE-I}$	6.2×10^{-7}	$[6.9\times10^{-7}, 5.4\times10^{-7}]$	7.1×10^{-6}	$[8.1\times10^{-6}, 6.3\times10^{-6}]$	0.21 ± 0.01
	$SNAP_{RE-I}$	/	/	2.6×10^{-6}	$[3.2\times10^{-6}, 2.1\times10^{-6}]$	0.14 ± 0.03
	MTP_{RE-I}	1.2×10^{-6}	$[1.3\times10^{-6}, 1.1\times10^{-6}]$	4.6×10^{-6}	$[5.4\times10^{-6}, 4.0\times10^{-6}]$	0.11 ± 0.01

Supplementary Table 4. The diffusivities at 1600 K, 1500K, 1230 K, 1000 K, 840 K, and 730 K for interstitial (superscript I) from AIMD and MLIP-MD simulations. The MD simulations with melting structures (noted as melt) and not enough hopping events (noted as N/A) are marked.

	Models	D_{1600K}^{I} ($cm^2\ s^{-1}$)	D_{1500K}^{I} ($cm^2\ s^{-1}$)	D_{1230K}^{I} ($cm^2\ s^{-1}$)	D_{1000K}^{I} ($cm^2\ s^{-1}$)	D_{840K}^{I} ($cm^2\ s^{-1}$)	D_{730K}^{I} ($cm^2\ s^{-1}$)
AIMD	K2	2.3×10^{-6}	3.0×10^{-6}	1.3×10^{-6}	6.6×10^{-7}	N/A	N/A
	Error bound	[2.7×10^{-6},	[3.6×10^{-6},	[1.6×10^{-6},	[8.9×10^{-7},		
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.9×10^{-6}]	2.4×10^{-6}]	9.5×10^{-7}]	4.4×10^{-7}]		
	K1	Melt	Melt	6.4×10^{-7}	4.5×10^{-7}	1.9×10^{-7}	1.6×10^{-7}
	Error bound			[8.0×10^{-7},	[5.7×10^{-7},	[2.6×10^{-7},	[2.2×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)			4.8×10^{-7}]	3.4×10^{-7}]	1.2×10^{-7}]	8.7×10^{-8}]
Known MLIPs trained from previous studies	GAP	2.8×10^{-6}	Melt	1.1×10^{-6}	4.6×10^{-7}	1.7×10^{-7}	N/A
	Error bound	[3.2×10^{-6},		[1.3×10^{-6},	[5.8×10^{-7},	[2.3×10^{-7},	
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.4×10^{-6}]		8.2×10^{-7}]	3.4×10^{-7}]	1.1×10^{-7}]	
	GAP_{PRX}	3.4×10^{-6}	2.9×10^{-6}	3.4×10^{-6}	1.8×10^{-6}	N/A	N/A
	Error bound	[4.0×10^{-6},	[3.4×10^{-6},	[4.1×10^{-6},	[2.2×10^{-6},		
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.8×10^{-6}]	2.4×10^{-6}]	2.8×10^{-6}]	1.4×10^{-6}]		
	GAP_{RPA}	Melt	Melt	Melt	1.8×10^{-6}	1.5×10^{-6}	4.5×10^{-7}
	Error bound				[2.1×10^{-6},	[1.8×10^{-6},	[5.5×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)				1.5×10^{-6}]	1.3×10^{-6}]	3.5×10^{-7}]
	GAP_{SCAN}	2.9×10^{-6}	2.8×10^{-6}	1.6×10^{-6}	6.7×10^{-7}	7.7×10^{-7}	4.7×10^{-7}
	Error bound	[3.5×10^{-6},	[3.3×10^{-6},	[2.0×10^{-6},	[8.7×10^{-7},	[1.0×10^{-6},	[6.4×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.4×10^{-6}]	2.3×10^{-6}]	1.3×10^{-6}]	4.7×10^{-7}]	5.3×10^{-7}]	2.9×10^{-7}]
	NNP	Melt	Melt	Melt	Melt	Melt	Melt
	Error bound						
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)						
	DeePMD	Melt	2.1×10^{-5}	5.7×10^{-6}	2.3×10^{-6}	3.8×10^{-6}	2.5×10^{-6}
	Error bound		[2.2×10^{-5},	[6.1×10^{-6},	[2.5×10^{-6},	[4.1×10^{-6},	[2.6×10^{-6},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)		2.0×10^{-5}]	5.3×10^{-6}]	2.2×10^{-6}]	3.5×10^{-6}]	2.3×10^{-6}]
	DeepPot-SE	2.9×10^{-6}	2.1×10^{-6}	1.7×10^{-6}	7.7×10^{-7}	3.2×10^{-7}	2.0×10^{-7}
	Error bound	[3.1×10^{-6},	[2.3×10^{-6},	[1.8×10^{-6},	[8.7×10^{-7},	[3.7×10^{-7},	[2.4×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.7×10^{-6}]	1.9×10^{-6}]	1.5×10^{-6}]	6.8×10^{-7}]	2.7×10^{-7}]	1.6×10^{-7}]
	DeePMD (original)	Melt	Melt	4.0×10^{-6}	3.1×10^{-6}	3.1×10^{-6}	3.1×10^{-6}
	Error bound			[4.3×10^{-6},	[3.4×10^{-6},	[3.3×10^{-6},	[3.3×10^{-6},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)			3.8×10^{-6}]	2.9×10^{-6}]	2.9×10^{-6}]	2.8×10^{-6}]
	DeepPot-SE (original)	Melt	Melt	Melt	Melt	Melt	Melt
	Error bound						
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)						
	MTP	2.9×10^{-6}	2.1×10^{-6}	1.1×10^{-6}	5.5×10^{-7}	1.4×10^{-7}	7.4×10^{-8}
	Error bound	[3.2×10^{-6},	[2.3×10^{-6},	[1.3×10^{-6},	[6.3×10^{-7},	[1.8×10^{-7},	[1.0×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.7×10^{-6}]	1.9×10^{-6}]	1.0×10^{-6}]	4.7×10^{-7}]	1.1×10^{-7}]	4.8×10^{-8}]
	SNAP	3.0×10^{-6}	Melt	4.7×10^{-7}	1.3×10^{-7}	2.0×10^{-8}	N/A
	Error bound	[3.5×10^{-6},		[6.4×10^{-7},	[1.5×10^{-7},	[2.9×10^{-8},	
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.6×10^{-6}]		3.0×10^{-7}]	1.1×10^{-7}]	9.9×10^{-9}]	
Vacancy-enhanced MLIPs	GAP_{RE-V}	2.9×10^{-6}	2.2×10^{-6}	1.1×10^{-6}	3.7×10^{-7}	1.6×10^{-7}	1.1×10^{-7}
	Error bound	[3.3×10^{-6},	[2.5×10^{-6},	[1.2×10^{-6},	[4.5×10^{-7},	[2.1×10^{-7},	[1.5×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.6×10^{-6}]	2.0×10^{-6}]	9.1×10^{-7}]	2.9×10^{-7}]	1.1×10^{-7}]	7.1×10^{-8}]
	NNP_{RE-V}	Melt	Melt	Melt	1.4×10^{-6}	1.5×10^{-6}	9.4×10^{-7}
	Error bound				[1.6×10^{-6},	[1.6×10^{-6},	[1.1×10^{-6},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)				1.2×10^{-6}]	1.3×10^{-6}]	7.8×10^{-7}]
	$DeePMD_{RE-V}$	Melt	Melt	7.3×10^{-6}	5.1×10^{-6}	3.3×10^{-6}	3.4×10^{-6}
	Error bound			[7.8×10^{-6},	[5.4×10^{-6},	[3.6×10^{-6},	[3.6×10^{-6},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)			6.8×10^{-6}]	4.8×10^{-6}]	3.1×10^{-6}]	3.2×10^{-6}]
	$DeepPot\text{-}SE_{RE-V}$	3.1×10^{-6}	2.7×10^{-6}	1.6×10^{-6}	1.0×10^{-6}	6.2×10^{-7}	4.9×10^{-7}
	Error bound	[3.3×10^{-6},	[2.9×10^{-6},	[1.7×10^{-6},	[1.1×10^{-6},	[7.0×10^{-7},	[5.5×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.8×10^{-6}]	2.5×10^{-6}]	1.4×10^{-6}]	9.1×10^{-7}]	5.5×10^{-7}]	4.3×10^{-7}]
	$SNAP_{RE-V}$	4.0×10^{-6}	3.7×10^{-6}	3.1×10^{-6}	2.3×10^{-6}	1.1×10^{-6}	1.0×10^{-6}
	Error bound	[4.4×10^{-6},	[4.1×10^{-6},	[3.5×10^{-6},	[2.5×10^{-6},	[1.3×10^{-6},	[1.2×10^{-6},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	3.6×10^{-6}]	3.4×10^{-6}]	2.7×10^{-6}]	2.0×10^{-6}]	9.6×10^{-7}]	8.4×10^{-7}]
	MTP_{RE-V}	1.6×10^{-6}	9.6×10^{-7}	3.1×10^{-7}	4.3×10^{-8}	N/A	N/A
	Error bound	[1.7×10^{-6},	[1.1×10^{-6},	[3.6×10^{-7},	[6.1×10^{-8},		
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.4×10^{-6}]	8.6×10^{-7}]	2.6×10^{-7}]	2.6×10^{-8}]		

Interstitial-enhanced MLIPs	GAP_{RE-I} Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.4×10^{-6} [2.7×10^{-6}, 2.1×10^{-6}]	2.3×10^{-6} [2.7×10^{-6}, 2.0×10^{-6}]	1.6×10^{-6} [1.9×10^{-6}, 1.4×10^{-6}]	1.0×10^{-6} [1.2×10^{-6}, 8.3×10^{-7}]	6.5×10^{-7} [7.9×10^{-7}, 5.1×10^{-7}]	4.6×10^{-7} [5.8×10^{-7}, 3.5×10^{-7}]
	NNP_{RE-I} Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	Melt	Melt	1.3×10^{-6} [1.6×10^{-6}, 1.1×10^{-6}]	7.8×10^{-7} [9.2×10^{-7}, 6.4×10^{-7}]	2.7×10^{-7} [3.5×10^{-7}, 2.0×10^{-7}]	1.7×10^{-7} [2.3×10^{-7}, 1.1×10^{-7}]
	$DeePMD_{RE-I}$ Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	Melt	Melt	7.4×10^{-6} [7.9×10^{-6}, 6.9×10^{-6}]	1.1×10^{-5} [1.1×10^{-5}, 1.0×10^{-5}]	4.3×10^{-6} [4.5×10^{-6}, 4.0×10^{-6}]	3.5×10^{-6} [3.7×10^{-6}, 3.2×10^{-6}]
	$DeepPot\text{-}SE_{RE-I}$ Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	6.9×10^{-6} [7.3×10^{-6}, 6.5×10^{-6}]	2.8×10^{-6} [3.1×10^{-6}, 2.6×10^{-6}]	2.2×10^{-6} [2.4×10^{-6}, 2.1×10^{-6}]	1.8×10^{-6} [2.0×10^{-6}, 1.7×10^{-6}]	1.0×10^{-6} [1.1×10^{-6}, 9.3×10^{-7}]	9.7×10^{-7} [1.1×10^{-6}, 8.8×10^{-7}]
	$SNAP_{RE-I}$ Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	Melt	Melt	2.1×10^{-6} [2.4×10^{-6}, 1.8×10^{-6}]	1.2×10^{-6} [1.3×10^{-6}, 9.7×10^{-7}]	3.6×10^{-7} [4.5×10^{-7}, 2.7×10^{-7}]	4.3×10^{-7} [5.4×10^{-7}, 3.2×10^{-7}]
	MTP_{RE-I} Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	3.1×10^{-6} [3.4×10^{-6}, 2.9×10^{-6}]	2.4×10^{-6} [2.6×10^{-6}, 2.2×10^{-6}]	1.6×10^{-6} [1.7×10^{-6}, 1.4×10^{-6}]	1.1×10^{-6} [1.2×10^{-6}, 9.5×10^{-7}]	4.9×10^{-7} [5.6×10^{-7}, 4.2×10^{-7}]	3.2×10^{-7} [3.8×10^{-7}, 2.7×10^{-7}]

Supplementary Table 5. The diffusivities at 1600 K, 1500K, 1230 K, 1000 K, 840 K, and 730 K for vacancy (superscript V) from AIMD and MLIP-MD simulations. The MD simulations with melting structures (noted as melt) and not enough hopping events (noted as N/A) are marked.

	Models	D_{1600K}^{V} ($cm^2\ s^{-1}$)	D_{1500K}^{V} ($cm^2\ s^{-1}$)	D_{1230K}^{V} ($cm^2\ s^{-1}$)	D_{1000K}^{V} ($cm^2\ s^{-1}$)	D_{840K}^{V} ($cm^2\ s^{-1}$)	D_{730K}^{V} ($cm^2\ s^{-1}$)
AIMD	K2	1.3×10^{-6}	1.4×10^{-6}	1.2×10^{-6}	5.2×10^{-7}	N/A	N/A
	Error bound	[1.6×10^{-6},	[1.7×10^{-6},	[1.5×10^{-6},	[6.9×10^{-7},		
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	9.9×10^{-7}]	1.0×10^{-6}]	8.2×10^{-7}]	3.6×10^{-7}]		
	K1	1.3×10^{-6}	Melt	4.5×10^{-7}	2.2×10^{-7}	1.1×10^{-7}	N/A
	Error bound	[1.8×10^{-6},		[5.7×10^{-7},	[2.8×10^{-7},	[1.6×10^{-7},	
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	8.4×10^{-7}]		3.3×10^{-7}]	1.6×10^{-7}]	6.4×10^{-8}]	
Known MLIPs trained from previous studies	GAP	1.5×10^{-6}	Melt	8.1×10^{-7}	5.8×10^{-7}	3.3×10^{-7}	2.4×10^{-7}
	Error bound	[1.8×10^{-6},		[1.0×10^{-6},	[7.1×10^{-7},	[4.1×10^{-7},	[3.2×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.2×10^{-6}]		6.1×10^{-7}]	4.5×10^{-7}]	2.4×10^{-7}]	1.7×10^{-7}]
	GAP_{PRX}	1.5×10^{-6}	1.0×10^{-6}	1.2×10^{-6}	5.6×10^{-7}	N/A	N/A
	Error bound	[1.9×10^{-6},	[1.3×10^{-6},	[1.5×10^{-6},	[7.8×10^{-7},		
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.1×10^{-6}]	7.2×10^{-7}]	8.0×10^{-7}]	3.3×10^{-7}]		
	GAP_{RPA}	Melt	Melt	Melt	N/A	N/A	N/A
	Error bound						
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)						
	GAP_{SCAN}	N/A	N/A	N/A	N/A	N/A	N/A
	Error bound						
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)						
	NNP	1.5×10^{-6}	Melt	9.2×10^{-7}	6.7×10^{-7}	4.4×10^{-7}	2.9×10^{-7}
	Error bound	[1.8×10^{-6},		[1.1×10^{-6},	[7.3×10^{-7},	[5.0×10^{-7},	[3.4×10^{-7},
	[D_{min}, D_{max}] (cm^2/s)	1.2×10^{-6}]		7.0×10^{-7}]	6.1×10^{-7}]	3.8×10^{-7}]	2.5×10^{-7}]
	DeePMD	Melt	1.2×10^{-5}	5.7×10^{-6}	2.8×10^{-6}	4.9×10^{-6}	2.5×10^{-6}
	Error bound		[1.3×10^{-5},	[6.1×10^{-6},	[3.0×10^{-6},	[5.3×10^{-6},	[2.6×10^{-6},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)		1.2×10^{-5}]	5.4×10^{-6}]	2.5×10^{-6}]	4.5×10^{-6}]	2.3×10^{-6}]
	DeepPot-SE	1.8×10^{-6}	1.8×10^{-6}	1.4×10^{-6}	9.1×10^{-7}	5.4×10^{-7}	5.0×10^{-7}
	Error bound	[2.0×10^{-6},	[1.9×10^{-6},	[1.6×10^{-6},	[1.0×10^{-6},	[6.1×10^{-7},	[5.7×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.7×10^{-6}]	1.6×10^{-6}]	1.3×10^{-6}]	8.0×10^{-7}]	4.7×10^{-7}]	4.3×10^{-7}]
	DeePMD (original)	Melt	Melt	9.5×10^{-6}	1.1×10^{-5}	3.8×10^{-6}	4.3×10^{-6}
	Error bound			[1.0×10^{-5},	[1.1×10^{-5},	[4.0×10^{-6},	[4.5×10^{-6},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)			8.9×10^{-6}]	9.9×10^{-6}]	3.5×10^{-6}]	4.0×10^{-6}]
	DeepPot-SE (original)	Melt	Melt	Melt	Melt	Melt	Melt
	Error bound						
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)						
	MTP	1.8×10^{-6}	1.7×10^{-6}	9.7×10^{-7}	6.0×10^{-7}	3.6×10^{-7}	2.3×10^{-7}
	Error bound	[2.0×10^{6},	[1.9×10^{-6},	[1.1×10^{-6},	[6.8×10^{-7},	[4.2×10^{-7},	[2.8×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.7×10^{-6}]	1.6×10^{-6}]	8.6×10^{-7}]	5.2×10^{-7}]	3.0×10^{-7}]	1.8×10^{-7}]
	SNAP	2.1×10^{-6}	Melt	1.2×10^{-6}	7.8×10^{-7}	6.3×10^{-7}	5.0×10^{-7}
	Error bound	[2.5×10^{-6},		[1.4×10^{-6},	[8.5×10^{-7},	[7.0×10^{-7},	[5.5×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.8×10^{-6}]		9.2×10^{-7}]	7.1×10^{-7}]	5.6×10^{-7}]	4.3×10^{-7}]
Vacancy-enhanced MLIPs	GAP_{RE-V}	1.6×10^{-6}	1.4×10^{-6}	9.1×10^{-7}	3.9×10^{-7}	3.0×10^{-7}	2.4×10^{-7}
	Error bound	[1.8×10^{-6},	[1.6×10^{-6},	[1.1×10^{-6},	[4.7×10^{-7},	[3.7×10^{-7},	[3.0×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.4×10^{-6}]	1.2×10^{-6}]	7.6×10^{-7}]	3.0×10^{-7}]	2.3×10^{-7}]	1.7×10^{-7}]
	NNP_{RE-V}	2.4×10^{-6}	1.6×10^{-6}	1.2×10^{-6}	9.4×10^{-7}	5.3×10^{-7}	3.4×10^{-7}
	Error bound	[2.7×10^{-6},	[1.8×10^{-6},	[1.3×10^{-6},	[1.1×10^{-6},	[6.5×10^{-7},	[4.3×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.1×10^{-6}]	1.3×10^{-6}]	9.8×10^{-7}]	7.7×10^{-7}]	4.1×10^{-7}]	2.5×10^{-7}]
	$DeePMD_{RE-V}$	Melt	Melt	4.8×10^{-6}	3.5×10^{-6}	1.3×10^{-6}	5.7×10^{-6}
	Error bound			[5.1×10^{-6},	[3.7×10^{-6},	[1.4×10^{-6},	[6.1×10^{-6},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)			4.5×10^{-6}]	3.2×10^{-6}]	1.2×10^{-6}]	5.3×10^{-6}]
	$DeepPot\text{-}SE_{RE-V}$	Melt	1.5×10^{-6}	1.0×10^{-6}	7.5×10^{-7}	6.7×10^{-7}	5.1×10^{-7}
	Error bound		[1.6×10^{-6},	[1.1×10^{-6},	[8.3×10^{-7},	[7.4×10^{-7},	[5.7×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)		1.3×10^{-6}]	9.4×10^{-7}]	6.7×10^{-7}]	6.0×10^{-7}]	4.5×10^{-7}]
	$SNAP_{RE-V}$	Melt	1.4×10^{-6}	1.1×10^{-6}	5.7×10^{-7}	3.1×10^{-7}	2.7×10^{-7}
	Error bound		[1.6×10^{-6},	[1.2×10^{-6},	[7.0×10^{-7},	[3.9×10^{-7},	[3.5×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)		1.2×10^{-6}]	9.1×10^{-7}]	4.4×10^{-7}]	2.3×10^{-7}]	1.9×10^{-7}]
	MTP_{RE-V}	1.7×10^{-6}	1.7×10^{-6}	1.1×10^{-6}	8.0×10^{-7}	4.7×10^{-7}	3.5×10^{-7}
	Error bound	[1.9×10^{-6},	[1.8×10^{-6},	[1.2×10^{-6},	[9.0×10^{-7},	[5.4×10^{-7},	[4.1×10^{-7},
	[D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.5×10^{-6}]	1.5×10^{-6}]	9.4×10^{-7}]	7.0×10^{-7}]	4.0×10^{-7}]	2.9×10^{-7}]

Interstitial-enhanced MLIPs	GAP_{RE-I} Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	1.9×10^{-6} [2.2×10^{-6}, 1.6×10^{-6}]	1.8×10^{-6} [2.1×10^{-6}, 1.5×10^{-6}]	1.3×10^{-6} [1.6×10^{-6}, 1.1×10^{-6}]	9.7×10^{-7} [1.2×10^{-6}, 7.9×10^{-7}]	5.5×10^{-7} [6.7×10^{-7}, 4.2×10^{-7}]	5.0×10^{-7} [6.2×10^{-7}, 3.9×10^{-7}]
	NNP_{RE-I} Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	Melt	Melt	8.4×10^{-7} [9.9×10^{-7}, 6.8×10^{-7}]	N/A	5.9×10^{-7} [7.2×10^{-7}, 4.5×10^{-7}]	5.0×10^{-7} [6.2×10^{-7}, 3.9×10^{-7}]
	$DeePMD_{RE-I}$ Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	Melt	Melt	7.0×10^{-6} [7.4×10^{-6}, 6.6×10^{-6}]	Melt	6.4×10^{-6} [6.9×10^{-6}, 6.0×10^{-6}]	2.9×10^{-6} [3.1×10^{-6}, 2.7×10^{-6}]
	$DeepPot\text{-}SE_{RE-I}$ Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	Melt	1.6×10^{-6} [1.8×10^{-6}, 1.5×10^{-6}]	8.3×10^{-7} [9.2×10^{-7}, 7.5×10^{-7}]	6.2×10^{-7} [6.9×10^{-7}, 5.4×10^{-7}]	4.5×10^{-7} [5.1×10^{-7}, 4.0×10^{-7}]	2.8×10^{-7} [3.2×10^{-7}, 2.3×10^{-7}]
	$SNAP_{RE-I}$ Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	Melt	Melt	6.7×10^{-7} [8.0×10^{-7}, 5.3×10^{-7}]	Melt	4.0×10^{-7} [5.0×10^{-7}, 3.0×10^{-7}]	2.5×10^{-7} [3.3×10^{-7}, 1.7×10^{-7}]
	MTP_{RE-I} Error bound [D_{min}, D_{max}] ($cm^2\ s^{-1}$)	2.2×10^{-6} [2.4×10^{-6}, 2.0×10^{-6}]	1.8×10^{-6} [2.0×10^{-6}, 1.6×10^{-6}]	1.5×10^{-6} [1.6×10^{-6}, 1.3×10^{-6}]	1.2×10^{-6} [1.3×10^{-6}, 1.1×10^{-6}]	1.0×10^{-6} [1.1×10^{-6}, 9.0×10^{-7}]	7.4×10^{-7} [8.3×10^{-7}, 6.4×10^{-7}]

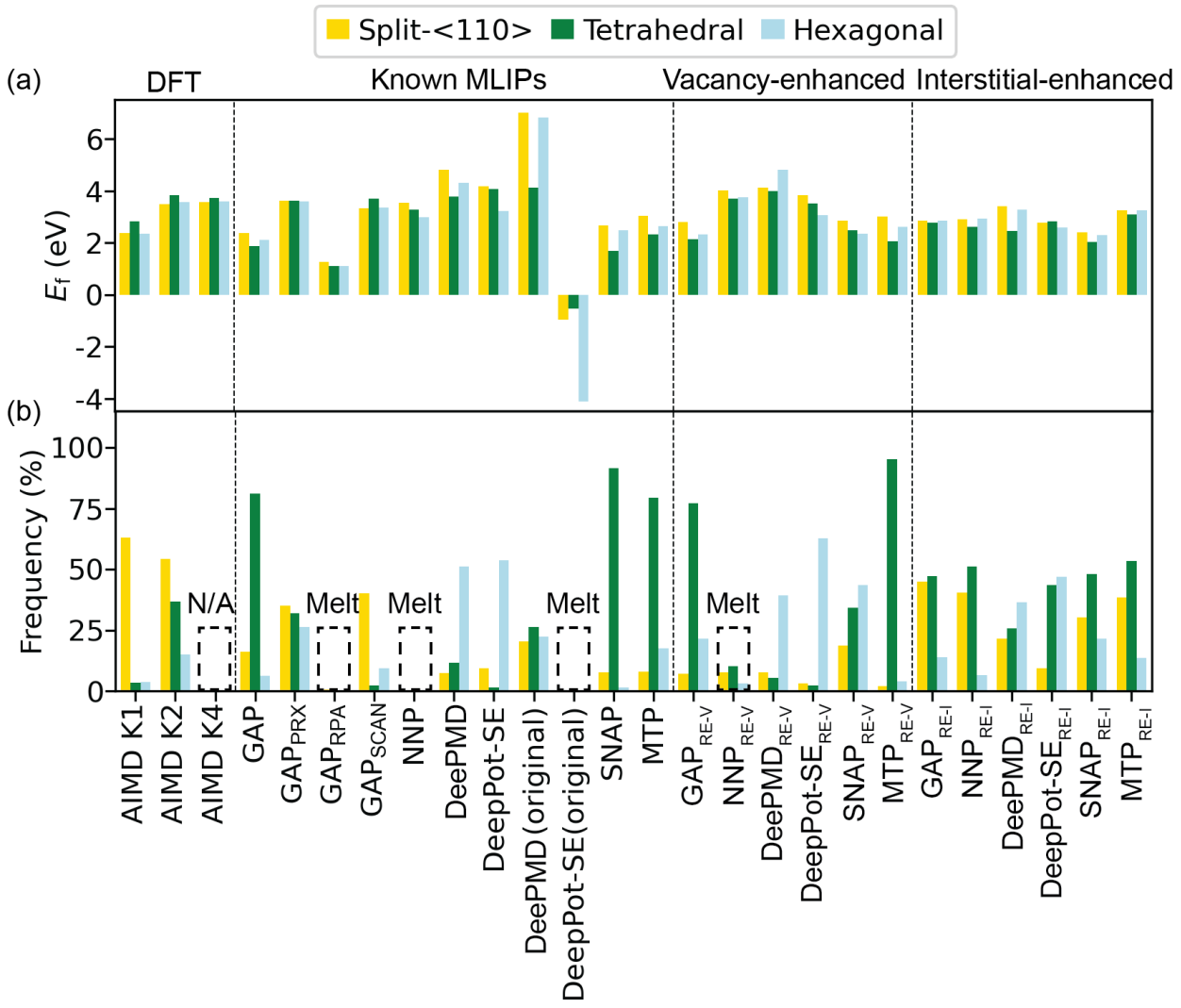
Supplementary Table 6. The additional diffusivities at 1333 K and 1080 K for single-interstitials and single-vacancies from AIMD simulations.

	AIMD	K2	K1
Interstitial	D_{1333K}^{I} ($cm^2\ s^{-1}$)	1.5×10^{-6}	8.7×10^{-7}
	Error bound	$[1.7\times10^{-6}$,	$[1.0\times10^{-6}$,
	$[D_{min}, D_{max}]$ ($cm^2\ s^{-1}$)	$1.2\times10^{-6}]$	$7.4\times10^{-7}]$
	D_{1080K}^{I} (cm^2/s)	1.0×10^{-6}	5.4×10^{-7}
	Error bound	$[1.2\times10^{-6}$,	$[6.4\times10^{-7}$,
	$[D_{min}, D_{max}]$ ($cm^2\ s^{-1}$)	$8.0\times10^{-7}]$	$4.4\times10^{-7}]$
Vacancy	D_{1333K}^{V} (cm^2/s)	1.2×10^{-6}	3.7×10^{-7}
	Error bound	$[1.4\times10^{-6}$,	$[4.5\times10^{-7}$,
	$[D_{min}, D_{max}]$ ($cm^2\ s^{-1}$)	$9.4\times10^{-7}]$	$2.9\times10^{-7}]$
	D_{1080K}^{V} ($cm^2\ s^{-1}$)	7.9×10^{-7}	2.5×10^{-7}
	Error bound	$[9.7\times10^{-7}$,	$[3.2\times10^{-7}$,
	$[D_{min}, D_{max}]$ ($cm^2\ s^{-1}$)	$6.1\times10^{-7}]$	$1.8\times10^{-7}]$

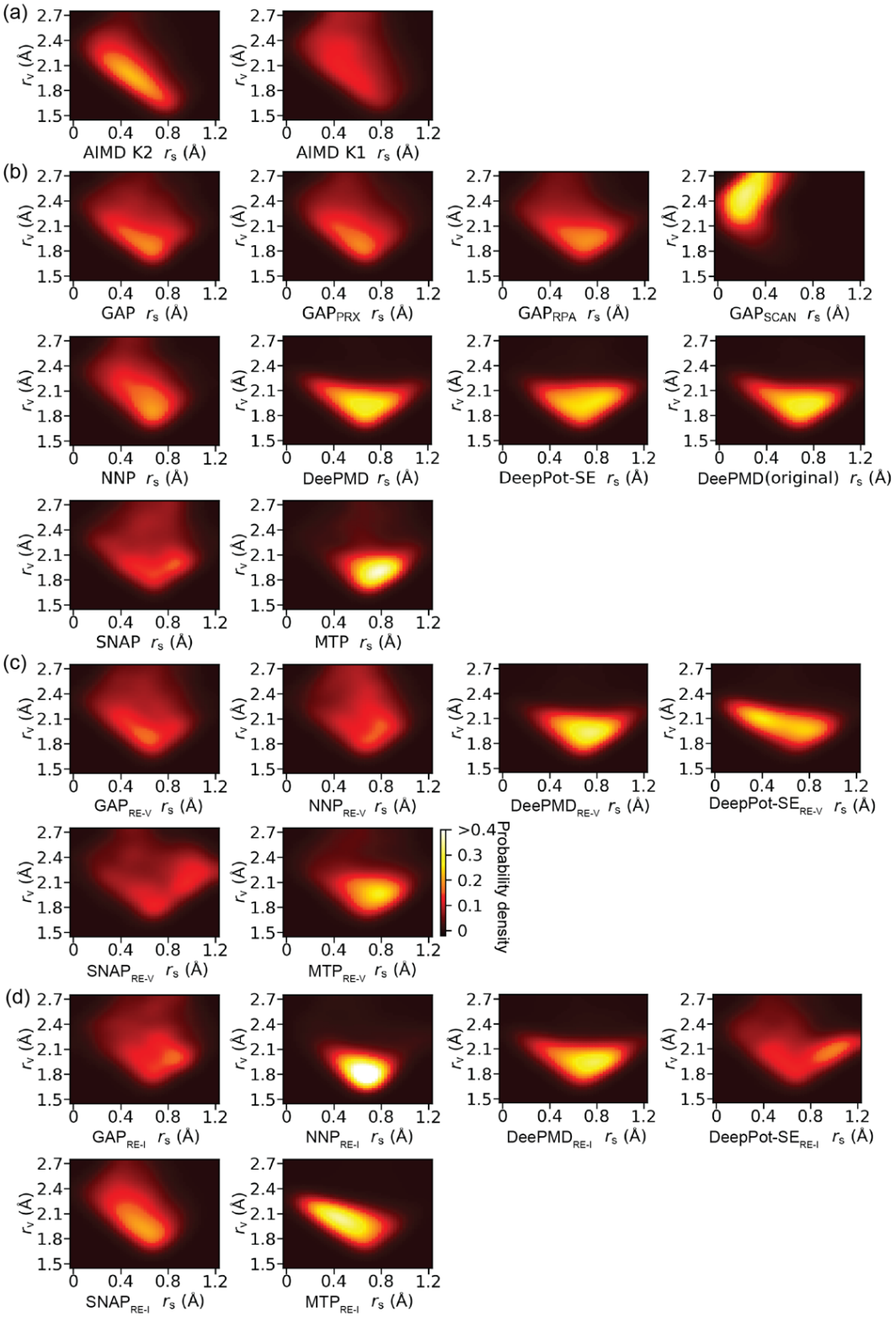
Supplementary Note 5. *Analysis of Si interstitials.* The comparison of E_f predicted by all MLIPs and DFT calculations are summarized in **Supplementary Figure 7a**. As shown in the main text, existing MLIPs poorly predict the formation energies E_f (and relative trends) of different interstitial configurations and their occurrence frequencies in MD simulations (Fig. 3d). The interstitial-enhanced MLIPs show significantly improved prediction of the occurrence frequencies of different interstitials as shown in Supplementary Figure 7b. However, RE-enhanced MLIPs still cannot fully predict of E_f in enhanced MLIPs. Most RE-enhanced MLIPs still give higher E_f of DFT-ground-state split-<110> than those of tetrahedral, except for DeepPot-SE_{RE-I}. In addition, the original and RE-enhanced MLIPs generally underestimate E_f by around 1 eV. GAP, GAP_{RPA}, SNAP, MTP, GAP_{RE-V}, $SNAP_{RE-V}$, MTP_{RE-V}, GAP_{RE-I}, NNP_{RE-I}, DeepPot-SE_{RE-I}, and $SNAP_{RE-I}$ predict E_f of all three interstitial configurations to be less than 3 eV, compared to the E_f range of 3.5 – 3.8 eV calculated by DFT K4. As a comparison to MLIPs, DFT K1 gives lower E_f range of DFT K1 between 2.3 – 2.8 eV but reproduces the correct trend of E_f among three configurations, with the split-<110> as the lowest and the tetrahedral as the highest. The result suggests low accuracy DFT still performs better than MLIPs on predicting complex defect configurations.

In most of our MD simulations, we also note that the lower formation energies don't always lead to higher occurrence frequencies. For example, the occurrence frequency of split-<110> in the MD simulation by GAP is higher than that of hexagonal (Supplementary Figure 7b), while the predicted E_f of hexagonal (2.1 eV) is lower than that of split-<110> (2.3 eV) (Supplementary Figure 7a). Similarly, the predicted E_f of split-<110> and hexagonal are the same (2.85 eV) for GAP_{RE-I}, while the occurrence frequency of split-<110> is higher than that of hexagonal in GAP_{RE-I}-MD simulations. This may be caused by one or multiple of the following reasons. First, in addition to the formation energy E_f, the occurrence frequency may also be affected by the entropy

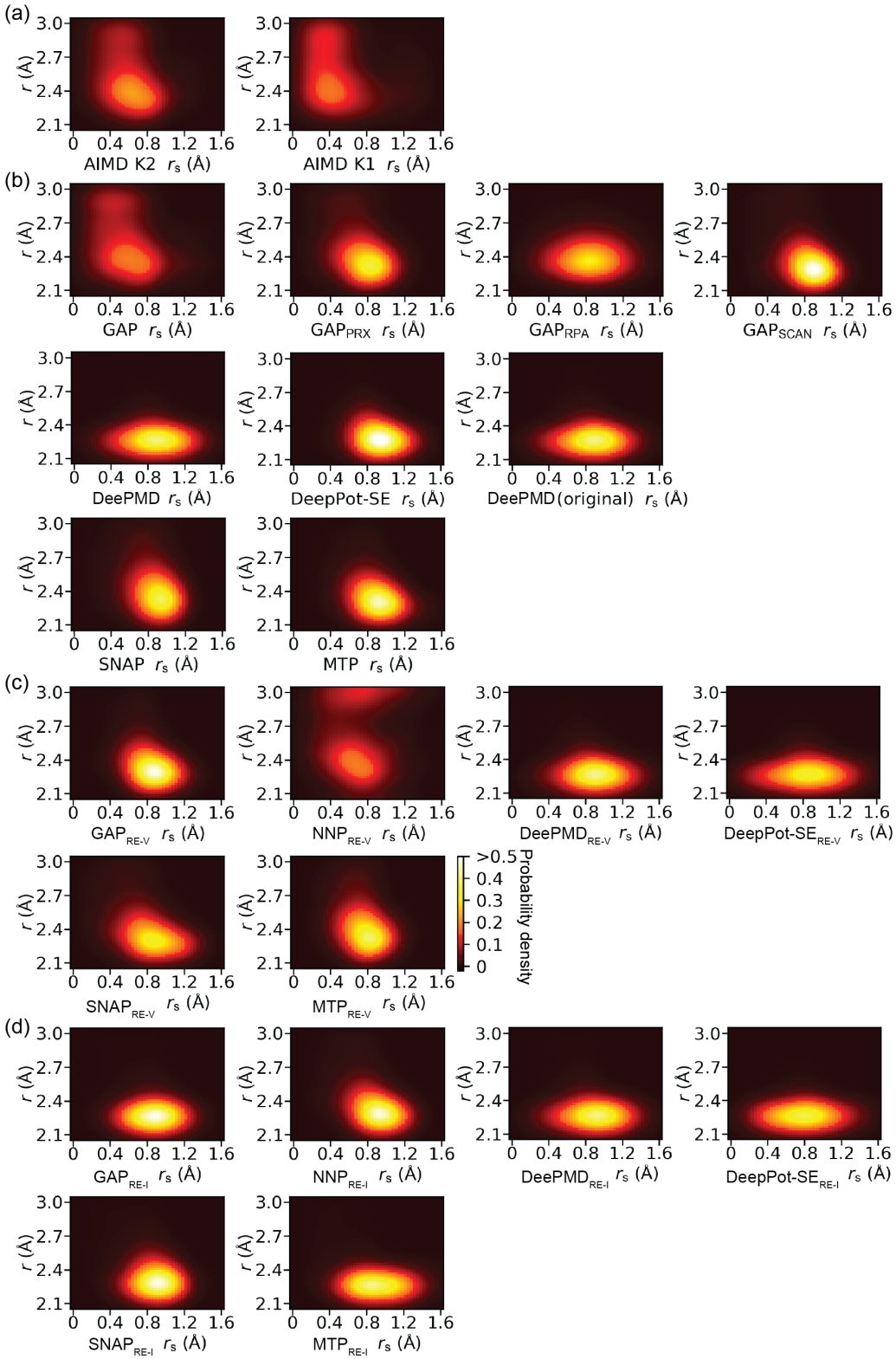
of different interstitials. Second, given the different interstitials are formed by the transition among them during MD simulations, the occurrence frequencies are influenced by their energy barriers, which is not identical to the formation energy E_f. Nonetheless, lower formation energy E_f is still highly correlated with higher occurrence frequencies. We compare the E_f and occurrence frequencies for each pair of interstitial configurations and count the number of pairs that have a consistent trend of increasing E_f and decreasing occurrence frequencies, and 73% of the interstitial pairs have the correct trend for all MLIPs in Fig. 3, and 61% for existing MLIPs and 87% for RE-enhanced MLIPs. Thus, the formation energy E_f may largely account for the occurrence frequencies of different interstitials in MD simulations.



Supplementary Figure 7. Comparison of DFT, known MLIPs, and RE-enhanced MLIPs, for (a) the formation energies E_f and (b) occurrence frequencies in MD simulations of different interstitial configurations. The occurrence frequency for AIMD K4 is unavailable (N/A) due to the high computation cost to run AIMD simulations with 4×4×4 **k**-point mesh.

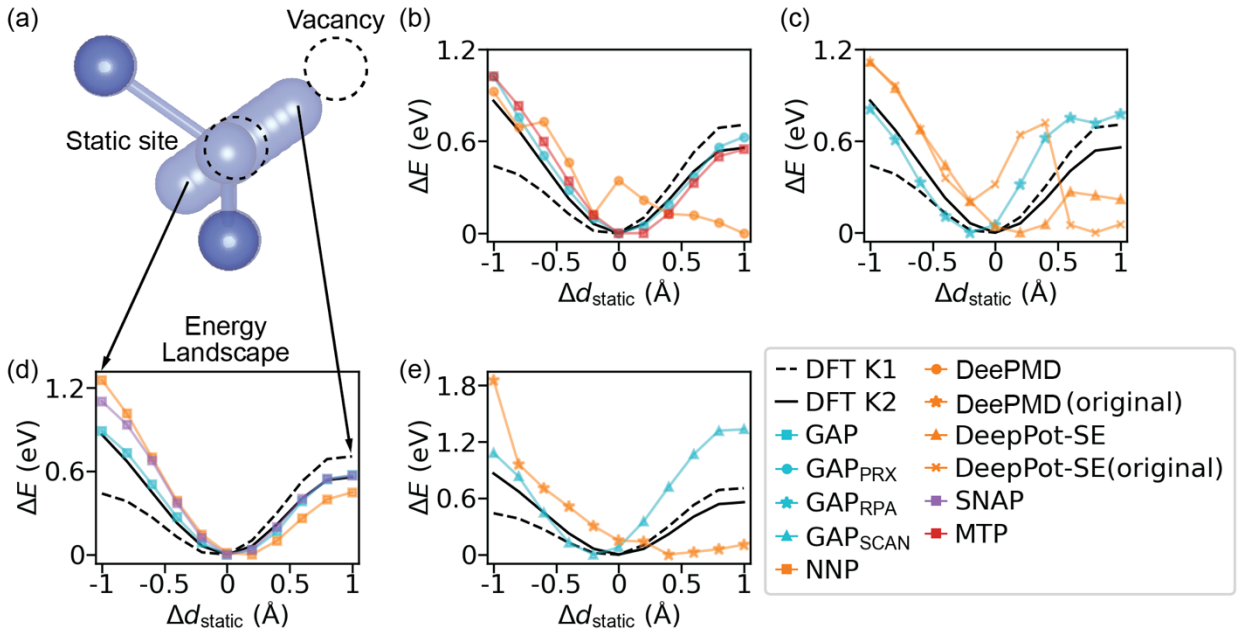


Supplementary Figure 8. The distributions of r_v versus r_s of the vacancy nearest neighbor (vacancy-NN) atoms from the MD simulations with single vacancy by (a) AIMD simulations, (b) original MLIPs, (c) vacancy-enhanced MLIPs, and (d) interstitial-enhanced MLIPs, in addition to Fig. 3.

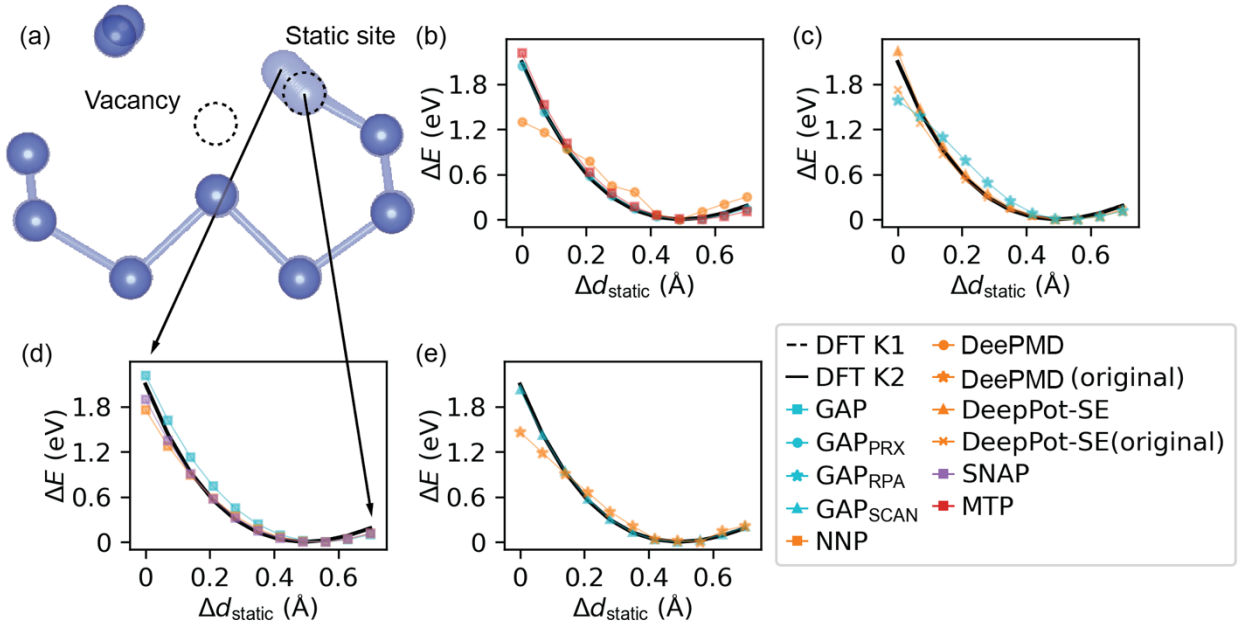


Supplementary Figure 9. The distributions of r versus r_s of the interstitial nearest neighbor (interstitial-NN) atoms from the MD simulations with single interstitial defect by (a) AIMD simulations, (b) original MLIPs, (c) vacancy-enhanced MLIPs, and (d) interstitial-enhanced MLIPs, following the same plotting protocols (Methods) as in Fig. 3.

Supplementary Note 6. *Analysis of potential energy landscapes.* **Supplementary Figure 10** shows the MLIP-predicted potential energy surfaces (PESs) of a Si atom moving toward the nearest-neighbor vacancy same as Fig. 4a and f. The predicted PESs of GAP_{RPA}, GAP_{SCAN}, DeePMD, DeepPot-SE, DeePMD (original), and DeepPot-SE (original) deviate significantly from DFT calculations (Supplementary Figure 10b, c, and e), suggesting poorly reproduced PESs by MLIPs. In addition, we also test the PES of a Si atom moving from its initial position in the snapshot towards the static site position in a relaxed cell, around a vacancy in a snapshot randomly taken from AIMD simulation with single vacancy defect (**Supplementary Figure 11**). Interestingly, these PESs agree well with DFT calculations (Supplementary Figure 11b, c, and e). The results of different PESs suggests that MLIPs may potentially show large discrepancies when predicting PESs with similar but slightly distorted atomic configurations.



Supplementary Figure 10. Comparison of the PESs predicted by known MLIPs and DFT calculations on (a) a single Si atom around a static site moving towards the vacancy in a static relaxed configuration, the same as in the main text Fig. 3a.



Supplementary Figure 11. Comparison of the PESs predicted by known MLIPs and DFT calculations on (a) a vacancy NN atom moving towards the static relaxed position, in a snapshot taken from AIMD simulations with single vacancy.

Supplementary Note 7. *Analysis of forces with large errors.* Following the same methods illustrated in main text, here we analyze the errors of atomic forces on all atoms in interstitial-RE and vacancy-RE testing sets and compare the force errors among three categories of atoms, such as RE migrating atoms (r = 0 Å), nearest neighbor (NN) atoms of RE migrating atoms (r = 0 - 3 Å), and other non-NN atoms (r > 3 Å) (**Supplementary Table 7**). We count the fraction of those atoms having large force errors with either the magnitude errors of atomic forces δ_F > 0.5 eV Å^{-1} or directional errors of atomic forces δ_θ > 15º. The fraction of atoms with large force errors are substantially higher in known MLIPs than DFT calculations (below 2% for DFT K2) in all three categories. For known MLIPs, 42% - 97% of RE migrating atoms have large force errors in interstitial-RE testing set. For vacancy-RE testing set, 31% to 64% of RE migrating atoms have large force errors by most MLIPs, except for GAP, GAP$_{PRX}$, and MTP, which are similar to or better than DFT K1. Therefore, atomic force predictions of most MLIPs are prone to show large errors than even low-accuracy DFT calculations.

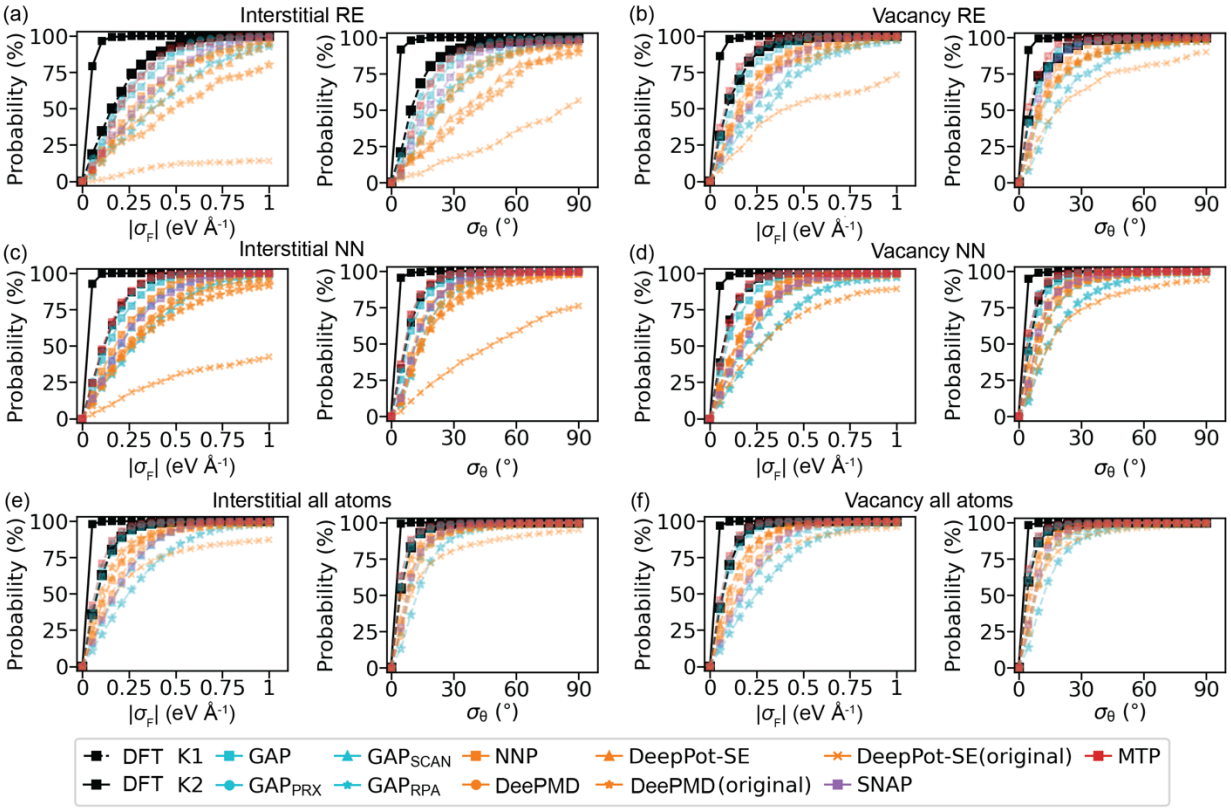
We compare known MLIPs for the cumulative density functions (CDFs) of force errors on RE migrating atoms (**Supplementary Figure 12a and b**) using the same approach as in the main text (Fig. 5d, e, i, and j), in interstitial-RE and vacancy-RE testing sets. We also perform the same analyses on NN atoms around RE (Supplementary Figure 12c and d) and on all atoms (Supplementary Figure 12e and f). Known MLIPs from previous studies clearly underperform DFT calculations for force predictions on RE migrating atoms. The CDFs of most MLIPs on RE-NN atoms are below the CDF curves of DFT K1, except for GAP$_{PRX}$ and MTP (Supplementary Figure 12c and d).

In contrast to RE or RE-NN atoms, the CDFs of force errors on all atoms show better agreement with DFT calculations (Supplementary Figure 12e and f). Particularly, the CDFs of

GAP, GAP_{PRX}, and MTP on all atoms are similar to or higher than the DFT K1 CDF curve in interstitial-RE testing set. The similar CDFs of force errors for all atoms are in consistent with the low RMSEs of force errors given by conventional error testing, because averaged total errors are evaluated on the dataset with a large fraction of near equilibrium atomistic configurations with small errors.

Supplementary Table 7. Fraction of atoms with large errors on atomic forces for three categories of atoms (RE migrating atoms, NN atoms of RE migrating atoms, and other non-NN atoms) in interstitial-RE testing set and in vacancy-RE testing set. Those atoms having large force errors with force magnitude errors $\delta_F > 0.5$ eV $Å^{-1}$ or force direction errors $\delta_\theta > 15^o$ are counted as large errors.

	Interstitial-RE Testing Set			Vacancy-RE Testing Set		
Models	RE r = 0 Å (%)	RE-NN 0 - 3 Å (%)	Non-NN r > 3 Å (%)	RE r = 0 Å (%)	RE-NN 0 - 3 Å (%)	Non-NN r > 3 Å (%)
DFT K1	33	17	5	21	8	5
DFT K2	1	0.4	0	0	0.6	0
GAP	58	23	4	21	12	4
GAP_{PRX}	42	16	2.5	13	6	3
GAP_{RPA}	97	59	47	64	60	47
GAP_{SCAN}	74	40	22	48	36	25
NNP	73	36	14	31	25	15
DeePMD	86	53	21	45	35	23
DeepPot-SE	93	59	14	39	27	11
DeePMD (original)	86	58	7	43	23	11
DeepPot-SE (original)	93	89	30	67	59	29
MTP	49	15	2	11	6	3
SNAP	68	44	19	34	26	19



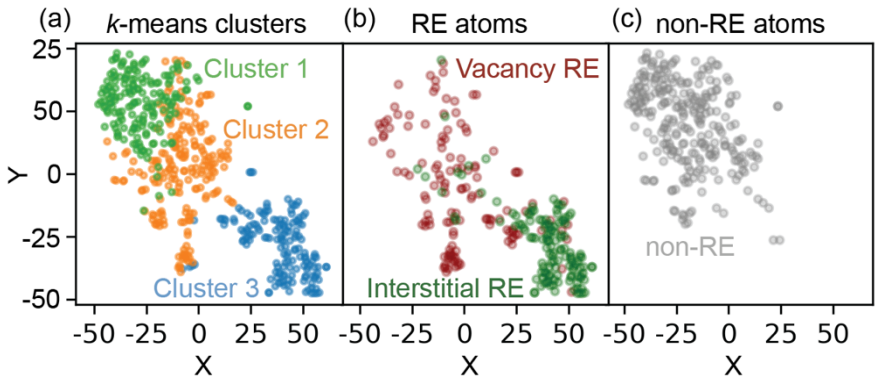
Supplementary Figure 12. The CDFs of force errors δ_F and δ_θ on (a) interstitial RE, (b) vacancy RE, (c) interstitial RE-NN, (d) vacancy RE-NN, (e) all atoms in interstitial-RE testing set, and (f) all atoms in vacancy-RE testing set of DFT calculations and known MLIPs.

Supplementary Note 8. *K-means clustering analysis.* We here employ the clustering algorithms to automate the process of identifying RE atoms from the dataset encompassing snapshots from MD simulations. We apply the *k*-means clustering technique using the SOAP descriptor to identify RE atoms (Methods). In the $\mathcal{D}_{\text{RE-I}}{}^{\text{Testing}}$ and $\mathcal{D}_{\text{RE-V}}{}^{\text{Testing}}$ datasets, the projections of SOAP show three clusters (**Supplementary Figure 13**), clearly distinguishing RE atoms (migrating vacancies and interstitials) from non-RE atoms (**Supplementary Table 8**). In addition, for a range of hyperparameters tested, this clustering approach is effective in identifying RE atoms. We trained three *k*-means clustering models, for all RE atoms, interstitial RE atoms, and vacancy RE atoms, respectively, using different datasets to identify RE atoms in simulations with single vacancy or interstitial, described as follows.

The all-RE *k*-means model is employed to distinguish RE atoms from non-RE atoms in a dataset of 12800 atomic environments from snapshots in the $\mathcal{D}_{\text{RE-I}}{}^{\text{Testing}}$ and $\mathcal{D}_{\text{RE-V}}{}^{\text{Testing}}$ dataset. The SOAP descriptors of atomic environments, which are calculated using QUIP package,[5] had 325 features and the same hyperparameters as Ref [3]. The *k*-means clustering is performed using *scikit-learn* package, with the maximal iteration of 1000, a convergence tolerance of 10^{-4}, and a number of clusters between three to eight (Supplementary Table 8). The results of the *k*-means models as shown in Supplementary Figure 13 are visualized by projecting onto a two-dimensional plane using the *t*-distributed stochastic neighbor embedding (TSNE) model in *scikit-learn* package[6]. The hyperparameter perplexity in TSNE model was set to be 50. The visualization in Supplementary Figure 13 is for a subset of data points consisting of 520 atomistic environments from four categories (130 each), such as RE atoms in $\mathcal{D}_{\text{RE-V}}{}^{\text{Testing}}$, RE atoms in $\mathcal{D}_{\text{RE-I}}{}^{\text{Testing}}$, non-RE atoms in $\mathcal{D}_{\text{RE-I}}{}^{\text{Testing}}$ and $\mathcal{D}_{\text{RE-V}}{}^{\text{Testing}}$.

The interstitial-RE *k*-means clustering model to distinguish interstitials is trained by using 6500 atomic environments from $\mathcal{D}_{\text{RE-I}}^{\text{Testing}}$ set, following the same protocols described above (the cluster number was set to 3). The 3rd *k*-means clustering model (vacancy-RE *k*-means clustering model) is trained from 6300 atomic environments taken from $\mathcal{D}_{\text{RE-V}}^{\text{Testing}}$ following the same training procedure of interstitial *k*-means model.

Here we show the precisions and recalls of the *k*-means clustering models, including the general all-RE *k*-means models trained by the $\mathcal{D}_{\text{RE-I}}^{\text{Testing}}$ and $\mathcal{D}_{\text{RE-V}}^{\text{Testing}}$ datasets (Supplementary Table 8), the interstitial-RE *k*-means models (**Supplementary Table 9**), and the vacancy-RE *k*-means models (**Supplementary Table 10**) described above. Our results show that the efficiencies of most *k*-means models deviate less than 5% for a range of cluster number parameters. In addition, the interstitial-RE *k*-means models, trained by atomistic environments in interstitial-RE testing set, is highly effective for identifying interstitial REs, with around 90% of real REs identified and 70% of identified REs being real REs.



Supplementary Figure 13. The two-dimensional *t*-SNE projections of the SOAP descriptor of 520 atomic environments, showing (left) the clusters by all-RE *k*-means model compared with (middle) migrating vacancies (red) or interstitials (green) and (right) non-migrating atoms (grey), from $\mathcal{D}_{\text{RE-V}}^{\text{Testing}}$ and $\mathcal{D}_{\text{RE-I}}^{\text{Testing}}$.

Supplementary Table 8. The precisions and recalls of the all-RE *k*-means models, with different cluster number parameters, for identifying both interstitial and vacancy RE migrating atoms (Methods).

	All-REs	
Cluster number	Precision (%)	Recall (%)
3	62	61
4	62	59
5	64	59
6	65	57
7	66	57
8	66	57

Supplementary Table 9. The precisions and recalls of interstitial-RE *k*-means models with different cluster numbers, which is trained using interstitial-RE testing set and for identifying interstitial RE migrating atoms for RE-enhanced workflow (*Generating RE Data and Constructing the Enhanced Training Dataset* in Methods).

	Interstitial-RE *k*-means Models	
Cluster number	Precision (%)	Recall (%)
3	65	90
4	67	90
5	69	89
6	68	89
7	69	89
8	70	88

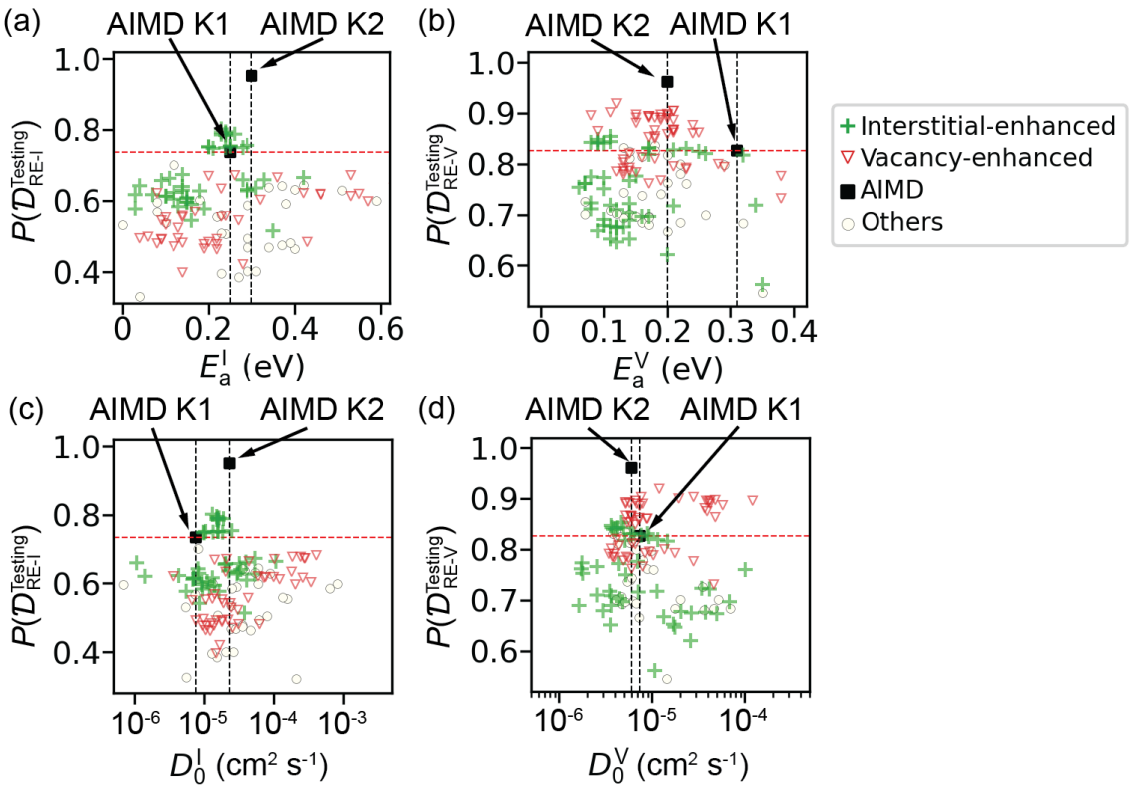
Supplementary Table 10. The precisions and recalls of vacancy-RE *k*-means models with different cluster numbers, which is trained using vacancy-RE testing set and for identifying vacancy RE migrating atoms (*Optimizing Enhanced MLIPs* in Methods).

	Vacancy-RE *k*-means Models	
Cluster number	Precision (%)	Recall (%)
3	10	75
4	8	80
5	10	74
6	20	60
7	23	51
8	24	47

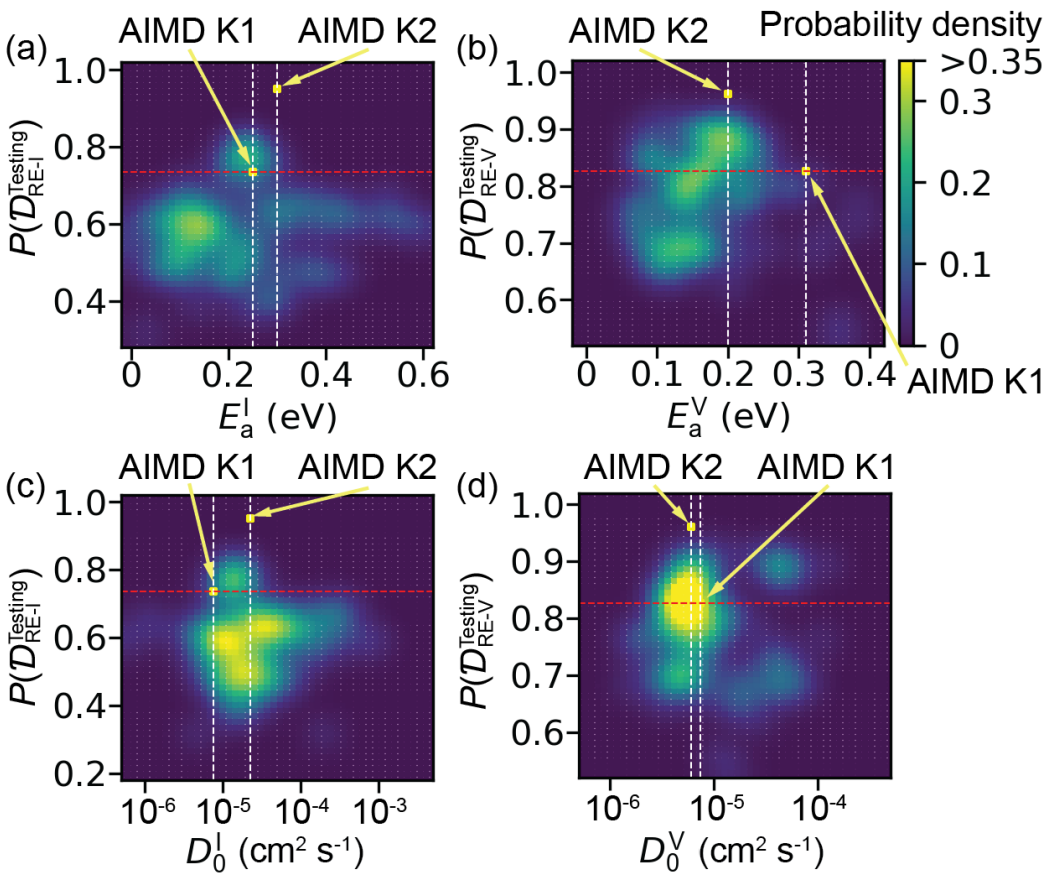
Supplementary Note 9. *Force accuracy as an evaluation metrics for MLIP performance.* Here, we compare the 11 known MLIPs and a large number of selected enhanced MLIPs for the force performance scores, $P(\mathcal{D}_{\text{RE-I}}{}^{\text{Testing}})$ and $P(\mathcal{D}_{\text{RE-V}}{}^{\text{Testing}})$, (see Methods) and their activation energy, $E_a{}^{\text{I}}$ and $E_a{}^{\text{V}}$ respectively, from the fitted Arrhenius relation of the diffusivities from their MLIP-MD simulations (**Supplementary Figure 14**). The studied MLIPs include 11 known MLIPs, 46 interstitial-enhanced MLIPs (*Optimizing Enhanced MLIPs* in Methods), 54 selected vacancy-enhanced MLIPs (*Analysis of Errors of MLIPs* above), and 24 additional MLIPs produced using the original training dataset in Ref,[3] but with different hyperparameters, which were selected by the optimization processes using the 8 selection criterion (*Optimizing Enhanced MLIPs* in Methods). The force performance scores, $P(\mathcal{D}_{\text{RE-I}}{}^{\text{Testing}})$ and $P(\mathcal{D}_{\text{RE-V}}{}^{\text{Testing}})$, were evaluated on RE migrating atoms in interstitial-RE testing and vacancy-RE testing datasets, respectively (*Optimizing Enhanced MLIPs* in Methods). The interstitial-enhanced MLIPs generally outperform other MLIPs on $P(\mathcal{D}_{\text{RE-I}}{}^{\text{Testing}})$ as shown in main text (Fig. 6a and Supplementary Figure 14a), and similarly MLIPs trained by vacancy-enhanced dataset have higher $P(\mathcal{D}_{\text{RE-V}}{}^{\text{Testing}})$ than other MLIPs (Supplementary Figure 14b). By comparing a range of MLIPs with different force performance scores versus the activation energy from MLIP-MD simulations with corresponding defect, the MLIPs generally give better predicted values of E_a if their force performance scores are higher than DFT K1 calculations.

To quantitatively illustrate this, in Supplementary Table S11 and S12, we compare the MLIPs with $P(\mathcal{D}_{\text{RE}}{}^{\text{Testing}})$ higher than DFT K1 benchmarks (referred as MLIPs with high $P(\mathcal{D}_{\text{RE}}{}^{\text{Testing}})$ in **Supplementary Table S11 and S12**) with the MLIPs with $P(\mathcal{D}_{\text{RE}}{}^{\text{Testing}})$ lower than DFT K1 benchmarks (referred as MLIPs with low $P(\mathcal{D}_{\text{RE}}{}^{\text{Testing}})$) for the mean absolute errors (MAEs) of diffusional properties E_a and D_0 (Supplementary Table 11) and the fractions of MLIPs

with high $P(\mathcal{D}_{\mathrm{RE}}{}^{\mathrm{Testing}})$ having diffusional properties within AIMD K1 and AIMD K2 range (Supplementary Table 12). Taking the diffusional properties from AIMD K2 simulations as the true values, the MAE of E_a in interstitial diffusions for MLIPs with high $P(\mathcal{D}_{\mathrm{RE\text{-}I}}{}^{\mathrm{Testing}})$ is 0.06 eV, compared to 0.18 eV MAE for MLIPs having low $P(\mathcal{D}_{\mathrm{RE\text{-}I}}{}^{\mathrm{Testing}})$ (Supplementary Figure 14a and Supplementary Table 11). 33.3% of MLIPs with high $P(\mathcal{D}_{\mathrm{RE\text{-}I}}{}^{\mathrm{Testing}})$ have their interstitial E_a between the range of 0.25 and 0.30 eV (corresponding to AIMD K1 and K2), while only 9.7% of MLIPs with low $P(\mathcal{D}_{\mathrm{RE\text{-}I}}{}^{\mathrm{Testing}})$ are within this range (Supplementary Figure 14a and Supplementary Table 12). Thus, there are significant improvements in the diffusional properties, such as E_a and D_0, in interstitial diffusions for MLIPs with high $P(\mathcal{D}_{\mathrm{RE\text{-}I}}{}^{\mathrm{Testing}})$ compared to those with low $P(\mathcal{D}_{\mathrm{RE\text{-}I}}{}^{\mathrm{Testing}})$ (Supplementary Table 11 and 12). For vacancy diffusion, the activation energies in vacancy diffusions, $E_a{}^{\mathrm{V}}$, are slightly improved for MLIPs with high $P(\mathcal{D}_{\mathrm{RE\text{-}V}}{}^{\mathrm{Testing}})$. The MAE of $E_a{}^{\mathrm{V}}$ is 0.04 eV and 35.6% of MLIPs are within the range of AIMD K1 and K2, compared to 0.07 eV MAE and 21.2% for MLIPs with low $P(\mathcal{D}_{\mathrm{RE\text{-}V}}{}^{\mathrm{Testing}})$. **Supplementary Figure 15** plotting the probability density distributions of $P(\mathcal{D}_{\mathrm{RE}}{}^{\mathrm{Testing}})$ and these properties shows the correlations between force performance scores and the diffusional properties. The correlations of $E_a{}^{\mathrm{I}}$ (Supplementary Figure 15a) and $E_a{}^{\mathrm{V}}$ (Supplementary Figure 15b) on force performance scores can be clearly observed for the yellow areas with high probability density (> 0.25). Such correlations are less clear on the pre-factor D_0 of both interstitial and vacancy diffusion (Supplementary Figure 15c and d). Overall, as shown by the quantitative correlations of MAEs and fractions for MLIPs with high and low $P(\mathcal{D}_{\mathrm{RE}}{}^{\mathrm{Testing}})$, $P(\mathcal{D}_{\mathrm{RE}}{}^{\mathrm{Testing}})$ are effective evaluation metrics for diffusional properties. While further studies are still needed for improving their effectiveness, force performance scores are in general good metrics to evaluate MLIPs for their performance in MD simulations.



Supplementary Figure 14. The force performance scores vs. E_a for known, interstitial-enhanced, vacancy-enhanced, and additional MLIPs on (a) interstitial RE migrating atoms and (b) vacancy RE migrating atoms. The force performance scores vs. D_0 for known, interstitial-enhanced, vacancy-enhanced, and additional MLIPs on (c) interstitial RE migrating atoms and (d) vacancy RE migrating atoms.



Supplementary Figure 15. The probability density distributions for all MLIPs shown in Supplementary Figure 14 on (a) force performance scores $P(\mathcal{D}_{\text{RE-I}}^{\text{Testing}})$ vs. E_a^I, (b) force performance scores $P(\mathcal{D}_{\text{RE-V}}^{\text{Testing}})$ vs. E_a^V, (c) $P(\mathcal{D}_{\text{RE-I}}^{\text{Testing}})$ vs. D_0^I, and (d) $P(\mathcal{D}_{\text{RE-V}}^{\text{Testing}})$ vs. D_0^V. The probability density distributions are calculated by applying a Gaussian smearing filter on the scattering plots of all MLIPs in Supplementary Figure 14 using a smearing radius of 2.5 (Supplementary Figure 15), implemented in the *ndimage* function in *scipy* package, on the plots use 50×50 grids.

Supplementary Table 11. The MAE of diffusional properties for the MLIPs with $P(\mathcal{D}_{\text{RE-I}}^{\text{Testing}})$ or $P(\mathcal{D}_{\text{RE-V}}^{\text{Testing}})$ higher or lower than DFT K1 calculations.

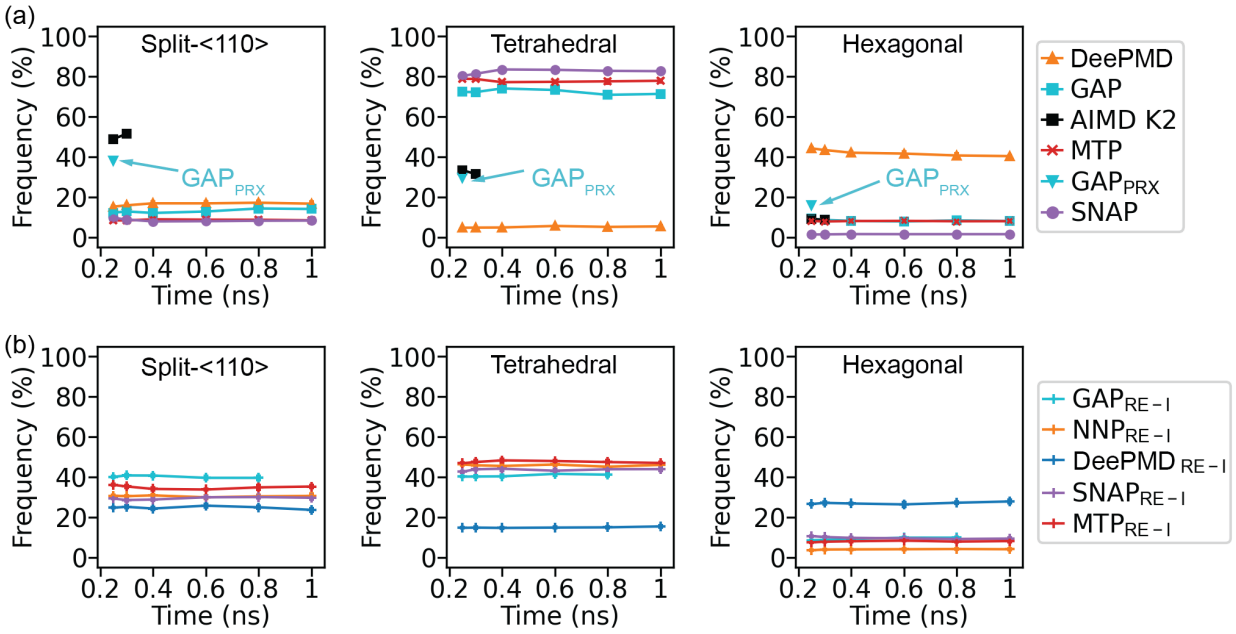
Diffusional property	Force performance scores of DFT K1	Properties of AIMD K2	MAE of MLIPs with $P(\mathcal{D}_{\text{RE}}^{\text{Testing}})$ > DFT K1	MAE of MLIPs with $P(\mathcal{D}_{\text{RE}}^{\text{Testing}})$ < DFT K1
E_a^I	0.73, $P(\mathcal{D}_{\text{RE-I}}^{\text{Testing}})$	0.30 eV	0.06 eV	0.18 eV
D_0^I	0.73, $P(\mathcal{D}_{\text{RE-I}}^{\text{Testing}})$	2.3×10^{-5} cm^2 s^{-1}	7.97×10^{-6} cm^2 s^{-1}	3.27×10^{-4} cm^2 s^{-1}
E_a^V	0.83, $P(\mathcal{D}_{\text{RE-V}}^{\text{Testing}})$	0.20 eV	0.04 eV	0.07 eV
D_0^V	0.83, $P(\mathcal{D}_{\text{RE-V}}^{\text{Testing}})$	6.0×10^{-6} cm^2 s^{-1}	1.20×10^{-5} cm^2 s^{-1}	9.88×10^{-6} cm^2 s^{-1}

Supplementary Table 12. The fraction of MLIPs having the diffusional properties within the range of AIMD K1 and K2 results.

Diffusional property	Force performance scores of DFT K1	Range between AIMD K1 and K2	Fraction of MLIPs with $P(\mathcal{D}_{\text{RE}}^{\text{Testing}})$ > DFT K1	Fraction of MLIPs with $P(\mathcal{D}_{\text{RE}}^{\text{Testing}})$ < DFT K1
E_a^I	0.73, $P(\mathcal{D}_{\text{RE-I}}^{\text{Testing}})$	[0.25, 0.30] eV	33.3%	9.7%
D_0^I	0.73, $P(\mathcal{D}_{\text{RE-I}}^{\text{Testing}})$	[7.7×10^{-6}, 2.3×10^{-5}] cm^2 s^{-1}	93.3%	36.3%
E_a^V	0.83, $P(\mathcal{D}_{\text{RE-V}}^{\text{Testing}})$	[0.20, 0.31] eV	35.6%	21.2%
D_0^V	0.83, $P(\mathcal{D}_{\text{RE-V}}^{\text{Testing}})$	[6.0×10^{-6}, 7.4×10^{-6}] cm^2 s^{-1}	17.8%	15.3%

Supplementary Note 10. *Convergence of interstitial occurrence frequencies.* Here we examine whether the occurrence frequencies of different interstitials have converged in the MD simulations. We calculate the average and standard deviation of occurrence frequencies for simulations conducted over 0.25, 0.3, 0.4, 0.6, 0.8, and 1 ns shown in **Supplementary Figure 16**. The average values and standard deviations of occurrence frequencies are calculated for 10 configuration sets obtained from the MD simulations with different durations. Every set of configurations consists of 2000 configurations that are randomly selected during MD simulations, with a minimum separation time of 100 fs, following the same scheme in Methods. The occurrence frequencies of three interstitials for all MLIP-MD simulations in Fig. 3 are well converged over different durations of the MD simulations, with standard deviations of occurrence frequencies lower than 1.3% for all three interstitials for all MLIPs. Therefore, the MD simulations have a long duration and have sampled a large number of configurations, and thus the obtained occurrence frequencies are well converged.

We also note that the occurrence frequencies of three interstitials do not always sum up to 1, because some interstitial configurations, e.g., migrating interstitials or concerted migrations of multiple interstitials, happened in MD simulations cannot be classified as either of three interstitials.



Supplementary Figure 16. The average occurrence frequencies and their standard deviations for MD simulations with different durations. The average and standard deviation of occurrence frequencies for split-<110> (left), tetrahedral (middle), and hexagonal (right) interstitials, in (a) AIMD K2 (black squares) simulations and MD simulations performed by known MLIPs: GAP (cyan squares), GAP_{PRX} (cyan triangles), DeePMD (orange triangles), SNAP (purple circles), and MTP (red crosses), and (b) MD simulation performed by RE-enhanced MLIPs: GAP_{RE-I} (cyan crosses), NNP_{RE-I} (orange crosses), $DeePMD_{RE-I}$ (blue crosses), $SNAP_{RE-I}$ (purple crosses), and MTP_{RE-I} (red crosses). The error bars of occurrence frequencies are smaller than 1.3% in all MD simulations.

Supplementary References.

1. Deringer, V. L. *et al.* Origins of structural and electronic transitions in disordered silicon. *Nature* **589**, 59–64 (2021).
2. Zhang, L. *et al.* End-to-end Symmetry Preserving Inter-atomic Potential Energy Model for Finite and Extended Systems. *Adv. Neural Inf. Process. Syst.* **32**, 4441–4451 (2018).
3. Zuo, Y. *et al.* Performance and Cost Assessment of Machine Learning Interatomic Potentials. *J. Phys. Chem. A* **124**, 731–745 (2020).
4. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Phys. Rev. X* **8**, 041048 (2018).
5. QUIP and quippy documentation. https://github.com/libAtoms/QUIP.
6. Varoquaux, G. *et al.* Scikit-learn. *GetMobile Mob. Comput. Commun.* **19**, 29–33 (2015).