Supplementary Information

A universal strategy for the creation of machine learning based atomistic force fields

Tran Doan Huan, Rohit Batra, James Chapman, Sridevi Krishnan, Lihua Chen, and Rampi Ramprasad

Fingerprint optimization

The fingerprints $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$ were optimized so that the root-mean-square error δ_{RMS} of the force predictions performed on the test set is minimized.

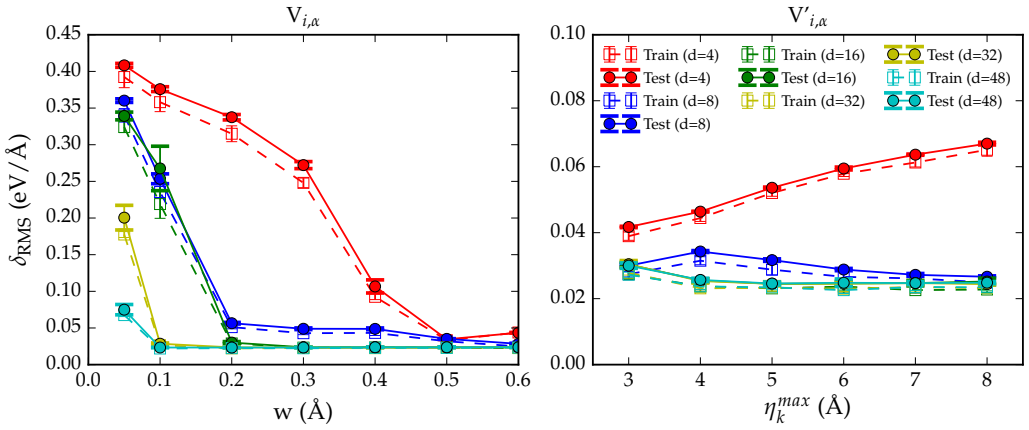

Fig. S1: Root-mean-square error δ_{RMS} of the force predictions performed on the test set. Data were averaged over 10 independently created force fields.

For $\mathbf{V}_{i,\alpha}$, the dimensionality d was chosen to be 4, 8, 16, 32, and 48 while w was scanned over 0.05Å, 0.10Å, 0.20Å, 30Å, 0.40, 0.50, and 0.60 Å. The centers a_k of the Gaussian functions were uniformly distributed from a half of the nearest-neighbor distance to $R_c = 8$Å. To optimize $\mathbf{V}'_{i,\alpha}$, d was also chosen to be 4, 8, 16, 32, and 48 while, at the same time, η_k were uniformly distributed in the log scale between a half of the nearest-neighbor distance to η_k^{max}, ranging from 3Å to $R_c = 8$Å. Results obtained for δ_{RMS} are shown in Fig. S1, indicating that the optimal parameters of $\mathbf{V}_{i,\alpha}$ are $d = 48$ and $w = 0.10$ while for $\mathbf{V}'_{i,\alpha}$, the optimal dimensionality $d = 48$ and $\eta_k^{\text{max}} = R_{\text{c}} = 8$Å.

Principal component analysis

Principal component analysis (PCA) was used to analyze the possibly correlated dimensions of $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$. In brief, this procedure uses an orthogonal transformation to convert the dimensions of each fingerprint to a set of *uncorrelated* variables, referred to as principal components. Each of them is associated with a variance, reflecting the variability (in the data) it accounts for. In other words, the variance of a principal component signals its relevance to the data.

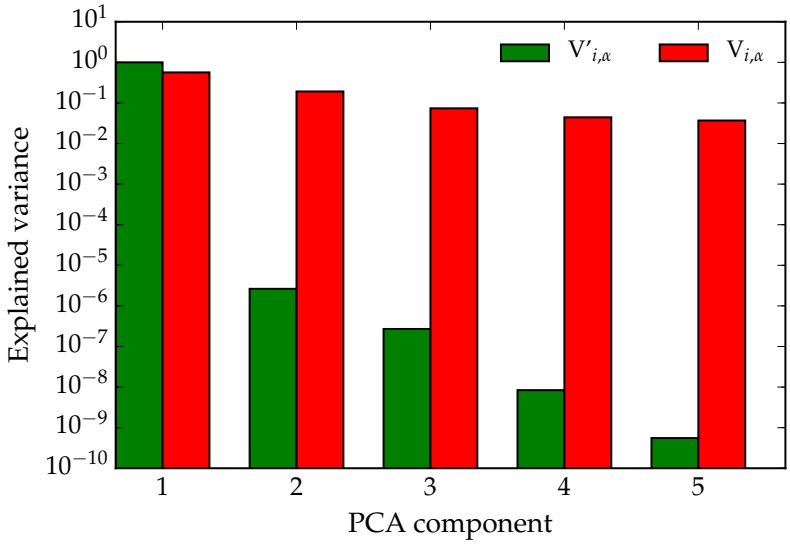

Fig. S2: Variance of the principal components of $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$.

Results of the PCA performed for $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$ are shown in Fig. S4. The first principal component of $\mathbf{V}'_{i,\alpha}$ captures most of the variance while for $\mathbf{V}_{i,\alpha}$, none of the principal components is dominant. This observation reflects the key difference between the two fingerprints. For $\mathbf{V}_{i,\alpha}$, the overlap of the basis functions $G(a_k, w; r)$s can easily be adjusted by w, thus the dimensions of the optimized $\mathbf{V}_{i,\alpha}$ are highly uncorrelated. On the other hand, any pair of $G(0, \eta_k; r)$s (in the definition of $\mathbf{V}'_{i,\alpha}$) have a significant non-vanishing overlap, being very poorly controllable by $\{\eta_k\}$, and thus, the dimensions of this fingerprint are highly correlated.

Molecular dynamics simulations using developed force fields

This section provides supplemental information on the applicability of the ML force fields we developed in MD simulation. For this purpose, we present the energy profiles of the NVE and NVT MD simulations performed for bulk Al (a 256 atoms supercell), Ti (an 128 atoms supercell), Si (a 512 atom supercell), and C (a 512 atoms supercell), employing LAMMPS and the ML force field of class (IV), combining $\mathbf{V}_{i,\alpha}$ and the force binning sampling method. More details on these runs can be found in the main text.

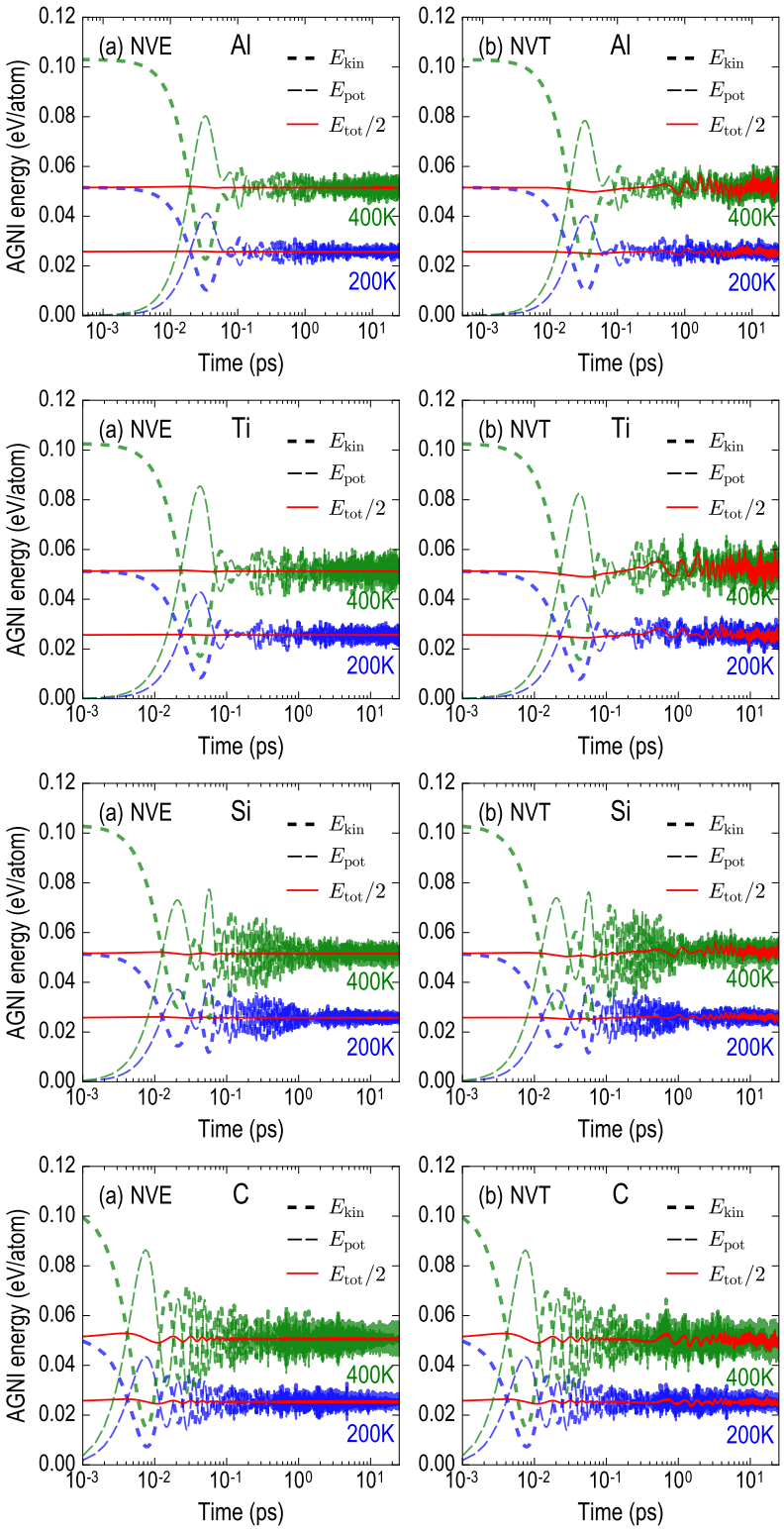

Fig. S3: NVE (a) and NVT (b) MD simulations of bulk Al, Ti, Si, and C employing LAMMPS and the ML force fields of class (IV).

Learning curves

This section provides critical assessments of the learning curves for the force fields of class (I), (II) and (V) described in the main article. In case of the *domain-specific* force fields, i.e. class (V), five fingerprint-force mappings were established each consisting of an equal training size. These curves demonstrate highest level of accuracy being reached by the class (V) force fields, especially in the case of Ti and W.

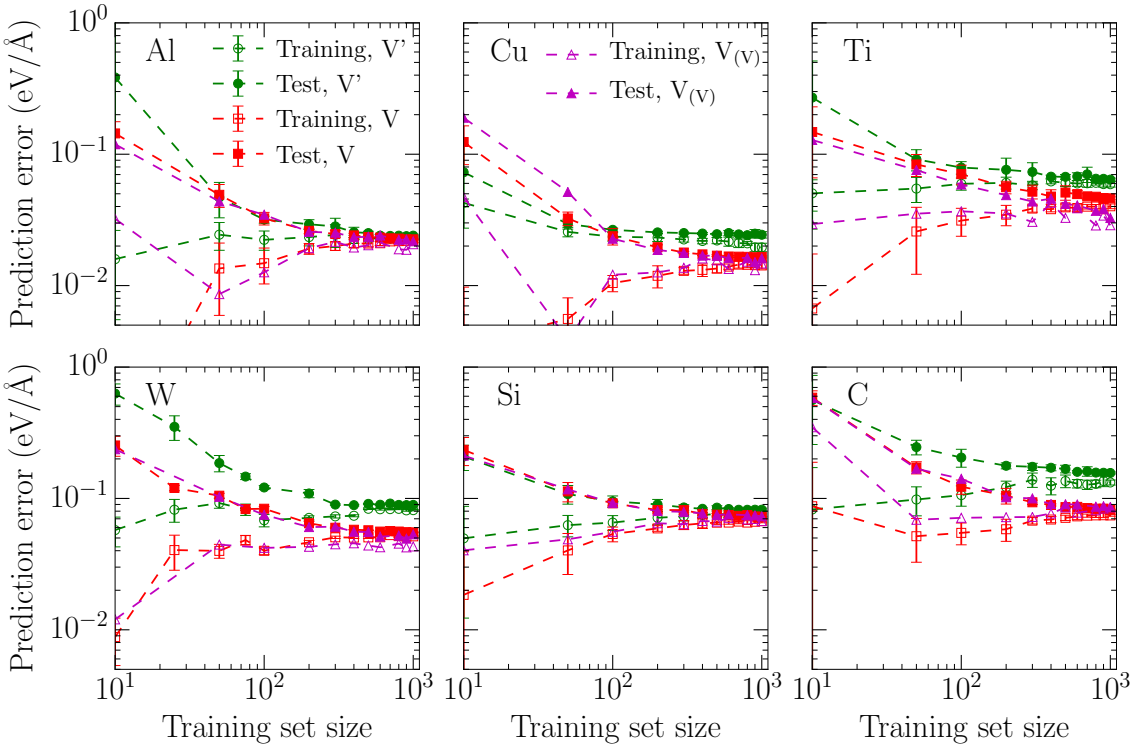

Fig. S4: Learning curves for ML force fields of class (I), (II) and (V), demonstrating highest force accuracy reached in the case of class (V) force field, especially for the case of Ti and W.