Supplementary Information

Data Distillation for Neural Network Potentials toward Foundational Dataset

Gang Seob Jung[1†], Sangkeun Lee[2] and Jong Youl Choi[2]

[1]Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[2]Computational Sciences and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

†Email: jungg@ornl.gov

Supplementary Note 1: Data Generation

1.1 Molecular dynamics

The molecular dynamics simulations were performed *via* the LAMMPS package.[1] we utilized EAM potentials for nickel[2] and aluminum.[3] The potentials have been developed to describe liquid/amorphous and crystalline phases. To compare the conventional isobaric-isothermal and MOMT ensemble methods, we prepared a 108 nickel atoms system. We performed a short relaxation through MD simulation at T_0 = 2,000 K and P_0 = 1 bar for 50,000 steps before sampling. Then, we sampled configurations every 100 steps (0.1 ps with a 1 fs timestep) for 200 ps. The number of obtained data is 2,000 from liquid and FCC phases. Both phases are stable under the conditions based on the initial state, and there is no transition between the two phases through the conventional approach.

1.2 Multiorder-multithermal ensemble molecular dynamics

Using the Wang-Landau algorithm,[4] we performed MOMT MD simulation based on the order parameters defined by reciprocal vectors. This method can efficiently sample both liquid and solid phases of crystalline materials, e.g., silicon and MgO.[5,6] Also, thermodynamic quantities as a function of temperature from the reweighting technique were demonstrated in agreement with other estimations in the Lennard-Jones system.[7] In this study, we utilized the method for sampling configurations to evaluate the sampled data for the NNP training. The detailed values for setting nickel and aluminum are listed in Table S1.

In the isobaric-multiorder multithermal (MOMT) ensemble method, the partial enthalpy is defined as functions of H and O *as*

$$H_{momt} = H_0 + \delta H(H,O)., \quad (1)$$

where O is an order parameter function to evaluate the order of the system as a function of coordinates, $\boldsymbol{r}$, given by[5]

$$O = \frac{1}{N_{fcc}N_A}\sum_{g\in fcc}\left|\sum_i \exp(ig\cdot r_i)\right|^2, \quad (2)$$

where g, N_{fcc}, and N_{A} are the shortest reciprocal vectors of FCC, the number of reciprocal vectors, and the number of atoms, respectively. For a more detailed description of methods, we refer to the previous study.[7]

Supplementary Note 2: Neural Network Potentials

We utilized the ANI type NNP through TorchANI[8] library. The NN structure in our study is listed in Table S2. We utilized Gaussian error linear unit (GELU) activation function[9]. Atomic environment vectors (AEV), or symmetry function,[10] capture the atomic environmental feature for the NN input. We followed the parameters of the AEV from ANI-2x[11] model except for a longer radius cutoff of 6.9Å with additional Gaussian centers.

For training iteration, 20% was used for validation, and 80% of the data was used to train the model with a mini-batch size 64. The data was shuffled when they were loaded.

The loss function is defined as

$$Loss = \frac{1}{N_{data}}\sum\frac{\left(E_{NNP}-E_{ref}\right)^2}{\sqrt{N_{atom}}} + \frac{\alpha}{N_{data}}\sum\frac{\left(\vec{F}_{NNP}-\vec{F}_{ref}\right)^2}{N_{atom}}, \quad (3)$$

where α was set to 0.1, a parameter to determine the contribution of forces. The basic training conditions are the same as in the previous study.[12] The maximum epochs was set as 300 and we took the best parameters for the validation set during the training. We did not prepare an explicit test set. However, we performed structural optimization to check physical properties and evaluated the performance with other data not explicitly included in both train/validation sets.

Supplementary Note 3: Structural Optimization

To evaluate the NNPs, we performed the structural optimization for three different closed-pack crystal structures (FCC, BCC, and HCP) and compared the energy rank and bond lengths. Although we sampled the FCC-based solid phase, we did not include the energy-optimized (or at a very low temperature) structure in the training set. The perfectly aligned crystalline structure is less likely to be sampled at a finite temperature, especially near the melting temperature. Whether the trained NNPs can derive the optimized structures can be a valuable indicator of the reliability of NNP models. The optimized shortest bond length and the total energy/atom from the EAM potential are compared using a previously developed interface with LAMMPS[12], we applied the same relaxation process to each structure based on the trained NNPs.

Supplementary Note 4: Data Distillation

3.1 Active learning-based distillation

Uncertainty quantification (UQ) of the NNPs is a central part of active learning because it allows us to identify valuable data likely to be informative and worth labeling with new calculations. In the current study, we employ the ensemble-based approach, utilizing the same NNP structure but with different training and validation sets for each model.[13] We divided 20,000 data from MOMT sampling into 10 sequential data sets (each set is 2,000 data points). From the initial data set, we trained the NNP with 5 models. At each step, the atomic energy is predicted from each model, and the standard deviation from the model predictions is used as an uncertainty quantification (UQ) measure (atomic UQ). If the atomic UQ value is larger than $\mu_{UQ} + 4\sigma_{UQ}$, we consider the configuration around that atom is not included in the dataset and included in the next iteration.

3.2 Order parameter & enthalpy based distillation

For the baseline approach for distillation, we manually selected the data based on the physical properties: the order parameter and enthalpy values. Since the order parameter informs the states of structures, it can be a good indicator of configurational similarity. We utilized data reduction approach through a neighbor list as suggested in the previous study.[12] Table S3 shows the number of data (108 atoms system) selected and deselected based on the δO values.

Supplementary Note 5: Distilled vs. Non-distilled.

We checked the training speed and the performance for the structural optimization. As we did in the data distillation, we also trained 5 models for a non-distilled nickel system with 108 atoms. We picked the best model based on their MAE of force. Due to the different numbers of configurations, their training speed is different. We checked the training speed based on 10 epochs in the beginning through one Nvidia GPU in a personnel workstation. For 20,000 data points with a 108-nickel system, it takes 41s/epoch. For 3,500 data points with a 32-nickel system, it takes 2.1s/epoch. The speed-up is about 19.4x. We note that TorchANI provides a CUDA version of symmetry function calculations, but we did not utilize it. Also, the number of batches can affect the benchmark test. The comparison of the two models is shown in Table S6. Full data also results in a good model to predict the energy rank correctly. The sum of the absolute error of bond length indicates that the distilled one is better (Full data: 0.08Å vs Distilled data: 0.03Å). Also, the energy error (max residue of the energy/atom error – min residue of the energy/atom error) shows that the distilled one performed better (Full data: 50meV/atom vs Distilled data: 20meV/atom). It does not completely confirm that distilled one is better, but at least the model trained with the distilled data is comparable with the model with non-distilled data.

Although we utilized empirical potential in the current demonstration, eventually, the applications should be done with DFT calculations. It is inevitable to perform a long-time integration (currently 2 ns) to sample a wide range of configurations in the current study. Since the time-space is not as parallelized as the length-space, running MD for 2ns suddenly becomes impractical with DFT calculations. However, embarrassing parallelization is possible with the sampled configurations (we can ignore time integration). In this context, we expect a clear speed-up, even considering the number of calls to Oracle in the proposed work. Furthermore, the MOMT sampling through NNP is at least two orders of magnitude faster than MOMT sampling through DFT, while the actual speed-up depends on the system size.

Supplementary Note 6: BCC metal Niobium

Nickel and Aluminum are FCC-type metals. We further tested whether the distilled data could be translated to BCC-type metal, Niobium (Nb). The parameters for Niobium are available in the empirical potentials we utilized for nickel. We first obtained the relaxed structure of Nb. Since the BCC structure is the most stable, we determined the scaling factor based on BCC as ~1.1867 (2.86/2.41). We did not perform the MOMT MD as the aluminum case, but comparing relaxed structures is still good to show how it would work. We included the results in Table S7. We realized that FCC and HCP of Nb have a very similar energy (2meV/atom difference). Therefore, it is difficult to say that it works perfectly. However, the results are promising, showing similar accuracy with aluminum cases.

Table S1. Conditions for the MOMT ensemble molecular dynamics.

Type (# atoms)	*T* (K)	*P* (bars)	*H* (eV) (# grids)	*O* (# grids)
Ni (108)	2,000	1	-540 ~ -410 (60)	-5 ~ 108 (100)
Ni (32)	2,000	1	-165 ~ -125 (60)	-5 ~ 40 (110)
Al (108)	1,000	1	-420 ~ -290 (60)	-5 ~ 108 (100)

Table S2. Neural network structures for Ni and Al in the current study. Gaussian error linear unit (GELU) activation function[9] was utilized to add non-linearity between AEV-1st, 1st-2nd, and 2nd-3rd layers. The radius cutoff for the radial part is 6.9Å.

NN Model	1st	2nd	3rd	Output (Energy)
Ni/Al/Nb	224	192	160	1

Table S3. Data distillation from the 20,000 configurational data sampled by the MOMT ensemble molecular dynamics. Firstly, we trained the 5 NNPs with the selected data (ΔO~3.0). Then, 7 AL iterations were performed from deleted data with a sparse grid (ΔO ~ 3.0) to fine grids (ΔO ~0.02).

ΔO	Selected	Deleted	Total
0.02	17,900	2,101	20,001
0.05	15,516	2,384	17,900
0.1	12,640	2,876	15,516
0.2	9,377	3,263	12,640
0.4	6,234	3,143	9,377
0.7	4,139	2,095	6,234
3.0	1,506	2,633	4,139

Table S4. The obtained self-energy minimizes the MAE of training/validation sets of each data set sampled from different initial configurations and NPT and MOMT ensembles (108 nickel atoms).

Data Set	FCC Data	Liquid Data	MOMT Data
Self-Energy (Hartree)	-0.167824160206977	-0.1610189932429113	-0.16532494629298

Table S5. The obtained self-energy minimizes the MAE of training/validation sets of systems

Data Set	Nickel Data	Aluminum Data	Niobium Data
Self-Energy (Hartree)	-0.16440532763410765	-0.11764409019342367	-0.2502205647745999

Table S6. Results of energy minimization from the different initial structures, FCC, BCC, and HCP nickel through NNPs from distilled data and non-distilled data (Full data: 20,000 data points of 108 atoms). (blue: lowest energy, red: highest energy, green: middle, l_b: bond length).

Structure (#atoms)	Ni-FCC (32) l_b(Å); E_{tot}/atom (eV)	Ni-BCC (54) l_b(Å); E_{tot}/atom (eV)	Ni-HCP (48) l_b(Å); E_{tot}/atom (eV)
Reference (EAM)	2.49/-4.876	2.41/-4.833	2.44/-4.847
NNP (Full Data) (errors)	2.46/-4.896 (-0.03/-0.020)	2.45/-4.803 (+0.04/+0.030)	2.43/-4.848 (-0.01/-0.001)
NNP (Distilled Data) (errors)	2.50/-4.875 (+0.01/+0.001)	2.41/-4.812 (0.00/+0.021)	2.46/-4.835 (+0.02/0.012)

Table S7. Results of energy minimization from the different initial structures, FCC, BCC, and HCP Niobium (Nb) through NNPs from translated data. (blue: lowest energy, red: highest energy, green: middle, l_b: bond length)

Structure (#atoms)	Nb-FCC (32) l_b(Å); E_{tot}/atom (eV)	Nb-BCC (54) l_b(Å); E_{tot}/atom (eV)	Nb-HCP (48) l_b(Å); E_{tot}/atom (eV)
Reference (EAM)	3.05/-7.159	2.86/-7.347	2.94/-7.157
NNP (Translated Data) (errors)	3.07/-7.173 (0.02/-0.014)	2.86/-7.350 (+0.00/-0.003)	2.93/-7.166 (-0.01/-0.009)

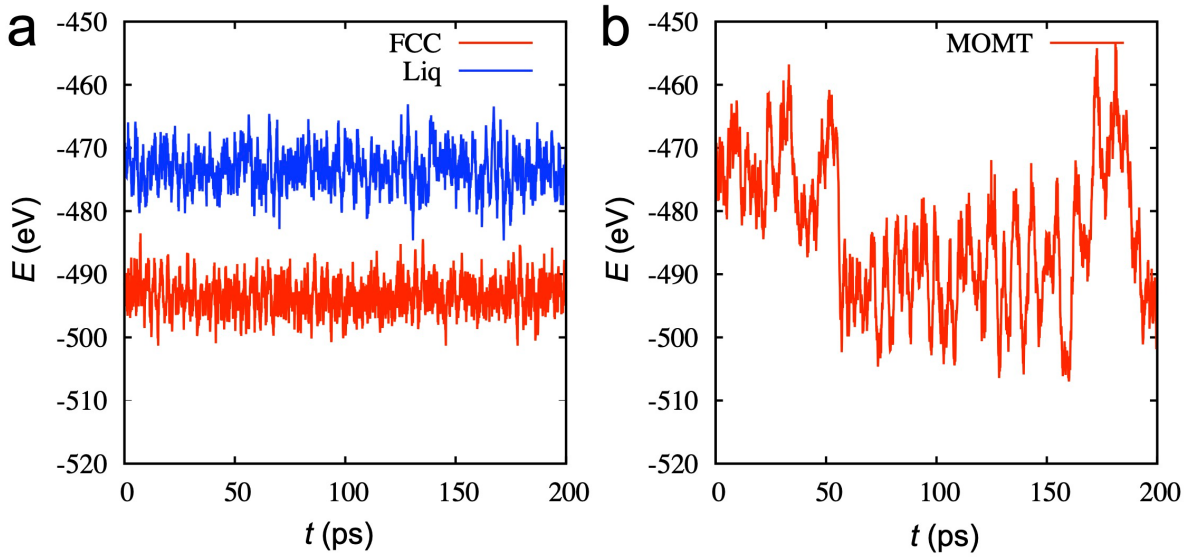

Figure S1. Time evolutions of energy (E) of nickel at T_0=2,000 K, P_0=1.0 bar for 200 ps. (**a**) Conventional NPT ensemble with different initial phases, FCC, and liquid phases. (**b**) Time evolution of E through MOMT ensemble. It shows that the transition between the two phases and the energy range where the liquid and FCC phases can be efficiently sampled.

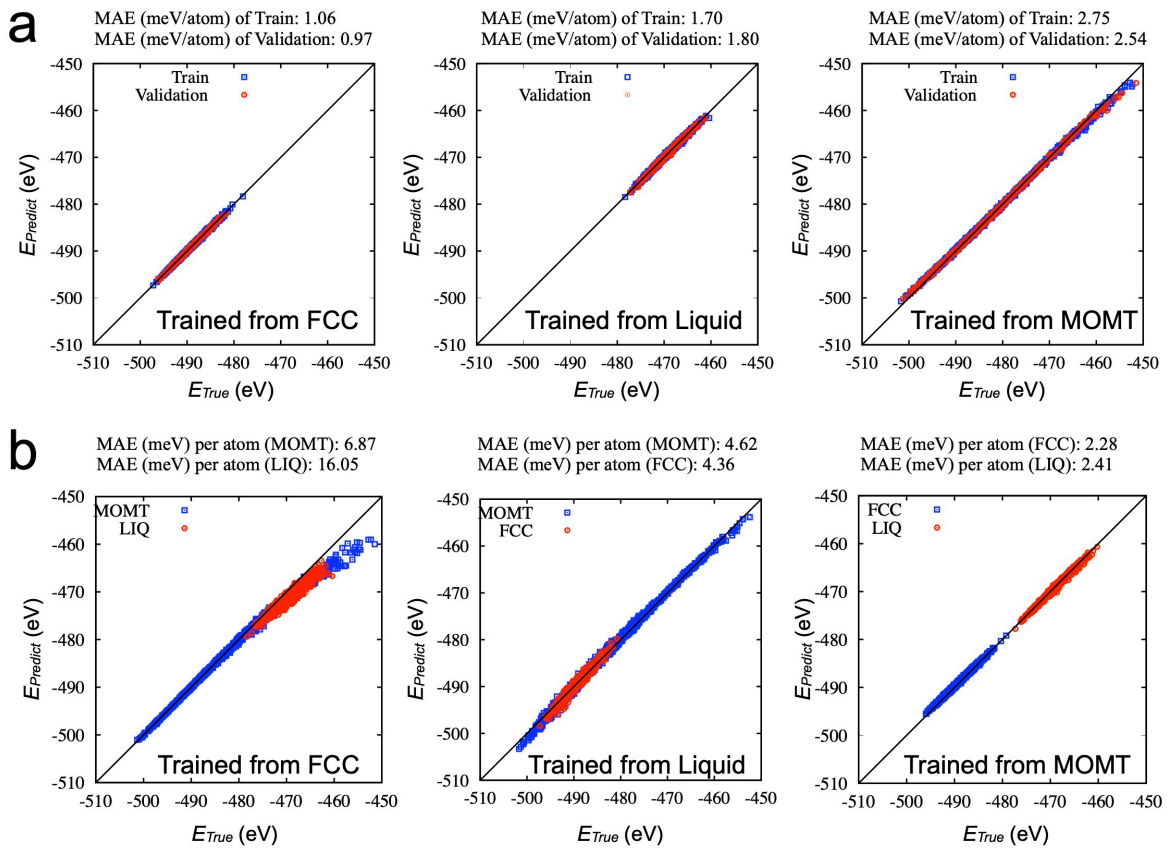

Figure S2 Mean absolute error (MAE) of NNPs trained from three different sampled data. (**a**) MAE of training/validation sets (**b**) MAE of other data (unseen) sets for each NNP model.

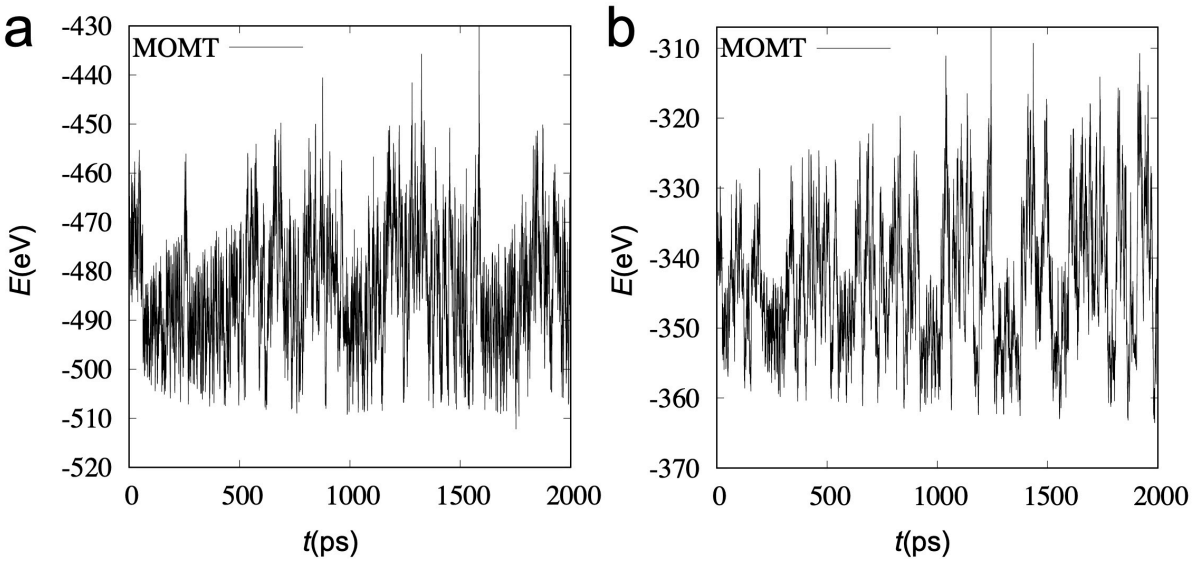

Figure S3 Times series of the MOMT ensemble MD simulations: (**a**) 108 nickel atoms at 2,000 K and 1 bar (**b**) 108 aluminum atoms at 1,000 K and 1 bar. The sampling region can change due to the reference temperature and maximum partial enthalpy. The range of the *y* label is set by scaling the *y* range of nickel. In the aluminum case, it can sample higher energy regions than nickel.

SI Reference

1 Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* **117**, 1-19, doi:https://doi.org/10.1006/jcph.1995.1039 (1995).

2 Zhang, Y., Ashcraft, R., Mendelev, M., Wang, C. & Kelton, K. Experimental and molecular dynamics simulation study of structure of liquid and amorphous Ni62Nb38 alloy. *The Journal of chemical physics* **145**, 204505 (2016).

3 Mendelev, M. I., Kramer, M. J., Becker, C. A. & Asta, M. Analysis of semi-empirical interatomic potentials appropriate for simulation of crystalline and liquid Al and Cu. *Philosophical Magazine* **88**, 1723-1750, doi:10.1080/14786430802206482 (2008).

4 Wang, F. & Landau, D. P. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Physical Review Letters* **86**, 2050-2053, doi:10.1103/PhysRevLett.86.2050 (2001).

5 Yoshimoto, Y. Extended multicanonical method combined with thermodynamically optimized potential: Application to the liquid-crystal transition of silicon. *The Journal of Chemical Physics* **125**, 184103, doi:10.1063/1.2363987 (2006).

6 Yoshimoto, Y. Melting of MgO studied using a multicanonical ensemble method combined with a first-principles calculation. *Journal of the Physical Society of Japan* **79**, 034602 (2010).

7 Jung, G. S., Yoshimoto, Y., Oh, K. J. & Tsuneyuki, S. Extended Ensemble Molecular Dynamics for Thermodynamics of Phases. *arXiv preprint arXiv:2308.08098* (2023).

8 Gao, X., Ramezanghorbani, F., Isayev, O., Smith, J. S. & Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *Journal of Chemical Information and Modeling* **60**, 3408-3415, doi:10.1021/acs.jcim.0c00451 (2020).

9 Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv [cs.LG]* (2020).

10 Behler, J. & Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters* **98**, 146401, doi:10.1103/PhysRevLett.98.146401 (2007).

11 Devereux, C. *et al.* Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *Journal of Chemical Theory and Computation* **16**, 4192-4202, doi:10.1021/acs.jctc.0c00121 (2020).

12 Jung, G. S., Myung, H. & Irle, S. Artificial neural network potentials for mechanics and fracture dynamics of two-dimensional crystals**. *Machine Learning: Science and Technology* **4**, 035001, doi:10.1088/2632-2153/accd45 (2023).

13 Jung, G. S., Choi, J. Y. & Lee, S. Active Learning of Neural Network Potentials for Rare Events. *Chemxriv* (2023).