Data Distillation for Neural Network Potentials toward Foundational Dataset

Gang Seob Jung *
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831
jungg@ornl.gov

Sangkeun Lee
Computational Sciences and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831
lees4@ornl.gov

Jong Youl Choi
Computational Sciences and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831
choij@ornl.gov

Abstract

Machine learning (ML) techniques and atomistic modeling have rapidly transformed materials design and discovery. Specifically, generative models can swiftly propose promising materials for targeted applications. However, the predicted properties of materials through the generative models often do not match with calculated properties through *ab initio* calculations. This discrepancy can arise because the generated coordinates are not fully relaxed, whereas the many properties are derived from relaxed structures. Neural network-based potentials (NNPs) can expedite the process by providing relaxed structures from the initially generated ones. Nevertheless, acquiring data to train NNPs for this purpose can be extremely challenging as it needs to encompass previously unknown structures. This study utilized extended ensemble molecular dynamics (MD) to secure a broad range of liquid- and solid-phase configurations in one of the metallic systems, nickel. Then, we could significantly reduce them through active learning without losing much accuracy. We found that the NNP trained from the distilled data could predict different energy-minimized closed-pack crystal structures even though those structures were not explicitly part of the initial data. Furthermore, the data can be translated to other metallic systems (aluminum and niobium), without repeating the sampling and distillation processes. Our approach to data acquisition and distillation has demonstrated the potential to expedite NNP development and enhance materials design and discovery by integrating generative models.

*Corresponding Author

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1 Introduction

Traditional approaches to materials design involve probing the physical and/or chemical properties of possible candidates with time-consuming and expensive trial-and-error studies. Therefore, deriving the principles for rational design from them is obstructed by the lack of sufficient and systematic data. Recent advances in physics-based computational atomistic modeling and simulations and machine learning (ML) techniques provide a novel avenue for exploring hypothetical materials to narrow the design space for materials with desired target properties [1, 2].

ML techniques have drastically advanced various sciences and engineering [3, 4, 5, 6, 7, 8]. Screening materials candidates through computational models with ML can accelerate new materials development and design, such as alloy [9], drugs [10], and polymer/protein engineering [11]. Although there are many challenges, from computational discovery to actual synthesis and deployment of the designed materials, the discovery is a critical starting point [12, 13].

To advance this computation-driven discovery as a starting point, two critical challenges must be addressed: the accuracy of ML models for predicting material properties and the effective generation of hypothetical materials. Although previous ML models work well for predicting the properties that are less sensitive to atomic coordinates, e.g., drug-likeness, it often becomes challenging to predict coordinate-sensitive properties, such as homo-lumo gap and vibration spectrum. This issue can be more severe with inorganic materials represented in the unit cell because their properties, such as band gap and elasticity, are more sensitive to the specific coordinates of atoms [14, 15, 16]. Therefore, obtaining the relaxed structure and the properties through *ab initio* calculations to confirm the prediction is often necessary. This redundant process can hinder the materials discovery. In many cases, the best properties predicted by the generated structures are less likely to be the same values after structural relaxation.

Recently developed ML-based forcefields (MLFFs) [17, 18, 19] have demonstrated a capability to predict potential energy surfaces (PESs) for atomic configurations with an accuracy comparable to *ab initio* electronic structure methods, but at a speed that is several orders of magnitude faster [20]. Neural network potentials (NNPs), a kind of MLFFs, utilize neural network to fit the interatomic interaction energies [21]. These have rapidly emerged due to their flexibility, accuracy, and efficiency. They are more suitable for large and complex systems than other MLFFs, as they can handle large data sets with many training data points.

Once NNPs can relax structure with similar accuracy as *ab initio* calculations, they can advance the materials discovery and screening more efficiently. However, NNPs, like other MLFFs, usually perform poorly outside their training domain and typically fail to predict unseen structures without appropriate data. Active learning (AL) [22] can help in improving the accuracy and exploring new structural data not included in the first stage when combining NNPs with enhanced sampling methods [23]. However, continuously acquiring more data and re-training still incurs significant computational costs. Therefore, it is desirable to have data in the beginning to handle the hypothetical structures as much as possible. Then, a smaller number of AL iterations are required for sufficient performance.

This study explores the feasibility of obtaining such "foundational" datasets for metallic systems using extended ensemble molecular dynamics (MD) techniques and data distillation. In MD simulations, conventional data sampling, e.g., isobaric-isothermal ensemble, has the local energy-minimum problem where sampling is more likely trapped in metastable or initial states. Therefore, it is challenging to sample amorphous structures or transient structures sufficiently. So many previous studies rely on changing temperatures from multiple known-initial structures to obtain various configurations for metallic/alloy systems [24, 25, 26]. In this study, we utilized the extended version of the multicanonical ensemble method [27, 28, 29], called the multiorder-multithermal ensemble (MOMT) [29] method. This method can sample possible transition states between solid and liquid phases with an accurate estimation of the density of states [30]. Also, we utilized the empirical embedded atom model (EAM) potentials [31, 32] instead of density functional theory calculations to evaluate the quality of sampled configuration.

Many similar configurations are sampled during either conventional or MOMT ensemble MD simulations. This can hinder the generalization of rarely sampled configurations. Therefore, based on

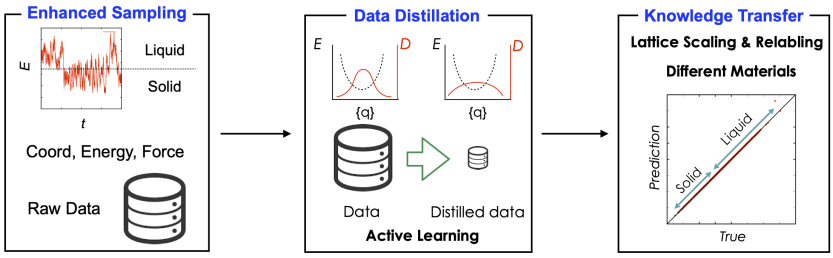

Figure 1: Schematic of workflow in the current study. First, we obtain the data of nickel (Ni) for two phases through the MOMT ensemble MD. The sampled data is biased to the equilibrium states. We reduce the amount of data without losing significant accuracy. The obtained knowledge on the configurations from Ni is translated to aluminum (Al) by scaling the lattice and relabeling.

the predicted atomic energy, we developed a data distillation process for the metallic system. Only 5% of the local configurational information can accurately describe the sampled structures ($\sim$1.0 meV/atom). Also, the derived NNP can describe relaxed FCC, BCC, and HCP structures with the correct cohesive energy rank, although those are not explicitly included in the training. Furthermore, we successfully demonstrated the distilled configurations can be translated to another metallic system without repeating the sampling and distillation processes.

2 Result and Discussion

A schematic of the workflow in the current study is shown in Figure 1. In the initial step, we sampled the configurations through the MOMT ensemble MD for liquid and solid phases. We utilized the previously developed scheme [30] to update the non-Boltzman weight factor for a broad sampling in both enthalpy and order parameter spaces. (See Supplementary Note 1). The method builds a bias for the visited states and allows overcoming the energy barrier to transit to other states. This approach is similar to well-known meta-dynamics or adaptive biasing force methods [33, 34]. One of the key differences of the multicanonical ensemble is that it utilizes the energy or enthalpy as a collective variable. Consequently, it can sample configurations with multiple reference temperatures. The difference in energy between the conventional MD and MOMT MD sampling is shown in Figure S1. Figure S1a shows the results of liquid and solid phases (108 nickel atoms) at the 2000K and 1 bar through the conventional isobaric-isothermal ensemble MD. Conventional sampling cannot overcome the energy barrier between the solid and liquid phases. However, the MOMT ensemble MD can sample both solid and liquid phases under the same conditions.

In this study, we utilized ANI-type NNP (See Supplementary Note 2). The results of the performance of NNPs through differently sampled configurations are summarized in Figure S2. The results confirm the previous findings that the training error (e.g., mean absolute error, MAE) values of training/validation data are insufficient to evaluate the NNPs because the values change because of configurational coverages [23, 35]. Although the data set from solid looks the best based on the training/validation MAE values. The MOMT sampled data demonstrates the best performance considering all sampled data. One notable fact is that even liquid phase-based NNP still decently predicts the solid phase while solid phase-based NNP poorly performs on the liquid data set.

Then, we evaluated the reliability of the trained NNPs by testing whether they can obtain relaxed FCC, BCC, and HCP structures through energy minimizations (See Supplementary Note 3). Table 1 shows the results. Although the FCC structure is the most stable state, only the NNP trained by the MOMT data could predict it correctly. However, even the NNP could not perfectly predict the other structures (BCC and HCP). A wider range of sampling could improve the reliability. Therefore, we sampled more to obtain 20,000 data points (Figure S3a). However, the method is inevitably sampling configurations of more probable states, not to visit next time by putting more bias and the sampled configurations can share similarity.

Therefore, we utilized active learning for data distillation to alleviate the data imbalance and generalize training NNPs. We performed 10 iterations for data distillation by the previously developed code [36]. We utilized the 5 models with randomly selected training (20%)/validation (80%) to estimate

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
FCC Data (errors)	2.56/-4.872 (+0.07/+0.004)	2.41/-4.905 (0.00/-0.072)	2.52/-4.831 (+0.08/+0.016)
LIQ Data (errors)	2.43/-4.889 (-0.06/-0.013)	2.49/-4.793 (+0.08/+0.040)	2.45/-4.905 (+0.01/-0.058)
MOMT Data (errors)	2.52/-4.895 (+0.03/-0.019)	2.46/-4.891 (+0.05/-0.058)	2.49/-4.889 (+0.05/-0.042)

Table 1: Results of energy minimization from the different initial structures, FCC, BCC, and HCP nickel through NNPs with 2,000 data points (blue: lowest energy, red: highest energy, green: middle, lb: bond length). The self-energy values utilized for each model are listed in Table S4.

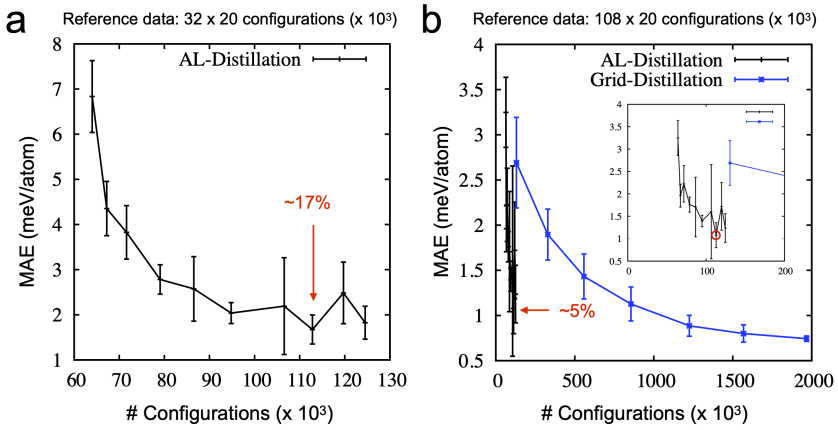

Figure 2: (a) The performance of data distillation through active learning for 32 nickel atoms system. The 17% data point can achieve MAE below 2.0 me/atom. The reference data for the MAE is MOMT sampled 20,000 frames (∼32×20,000 atomic configurations) (b) Comparison between grid-based distillation from 108 nickel atoms system and active learning-based distillation from 32 nickel atoms system. Only 5% of configurations can achieve MAE close to 1.0 meV/atom.

the atomic UQ from the standard deviation of the atomic energy predictions (See Supplementary Note 4). The purpose of the process is to reduce the amount of data without losing performance. An appropriate data reduction can help with fast training and performance, as shown in the previous study [35]. We chose 108 atom system as a reference because this is the minimum number to have a defective solid phase during the sampling, such as a stacking fault or HCP phase.

However, we found that the MOMT sampled configuration with a smaller unit cell of 32 atoms performs better as distilled data. Therefore, we used distilled data from the 32 atoms system. Figure 2a shows a good performance of data distillation on 32 nickel atoms system. 17% of configurations can achieve MAE lower than 2.0meV/atom for the total configurations. The effect is more drastic when we compare a physical more reasonable system (108 nickel atoms system) with manually distilled configurations in Figure 2b. The NNP trained from the distilled data shows a great performance (∼1.0meV/atom) with only 5% of information. As we hypothesized, widely sampled configurations can improve the energy rank of relaxed structures as shown in Table 2. The relaxed FCC, BCC, and HCP structures are well-matched with those referenced from EAM potentials (See Supplementary Note 5 for the performance between distilled onr and non-distilled one).

Finally, we investigated the generalization of the distilled information with the aluminum. Nickel and aluminum are FCC structures with different physical features regarding their energy, lattice parameters, and melting points. Aluminum has a larger lattice parameter (∼ 4.05Å) than nickel (∼3.52Å) and melt at a lower temperature (∼660°C) than nickel (1455°C). Although their lattice parameters are different, we can utilize the information by scaling the system size from nickel to aluminum (4.05/3.52∼1.15) and recalculating energy and forces based on aluminum potential. We

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 (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 2: Results of energy minimization from the different initial structures, FCC, BCC, and HCP nickel through NNPs from distilled data. (blue: lowest energy, red: highest energy, green: middle, l_b: bond length)

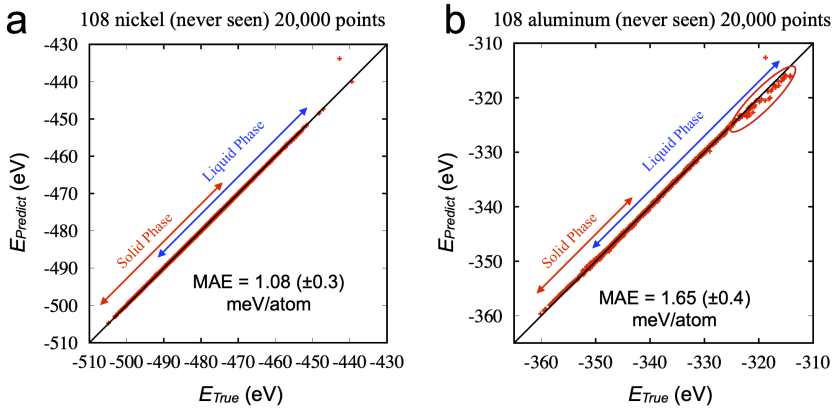

Figure 3: Scattering plot of the true and predicted energy values for 20,000 data points of 108 nickel atoms (1,000 points sampled for visualization). (b) Scattering plot of the true and predicted energy values for 20,000 data points of 108 aluminum atoms (1,000 points sampled for visualization). A red circle indicates the lower accuracy region, it accounts for rarely sampled energy region in the nickel system.

also performed MOMT ensemble MD simulations with aluminum with 108 atoms to obtain 20,000 configurational data points (Figure S3b). The data was not utilized to train NNP but to evaluate the NNP trained by the translated data from the nickel system.

Figure 3 shows the scatter plot of both nickel and aluminum. The NNP for aluminum shows a high accuracy except for a high-energy region. The NNP can also describe the rank of three closed-pack crystals, as shown in Table 3. The selected data through the nickel system are our essential knowledge for the reliable NNPs to describe solid and liquid phases. Then, this knowledge could be well-translated to the other metallic systems, aluminum (See Supplementary Note 6 for BCC-Niobium case).

3 Conclusions

In this short report, we demonstrate the effectiveness of the MOMT ensemble sampling for generating training data for metallic systems. The sampled data from MOMT includes both solid and liquid phases and covers a wider range of energy than conventional sampling, such as the isothermal-

	Al-FCC (32) l_b(Å); E_{tot}/atom (eV)	Al-BCC (54) l_b(Å); E_{tot}/atom (eV)	Al-HCP (48) l_b(Å); E_{tot}/atom (eV)
Reference (EAM)	2.86/-3.411	2.80/-3.309	2.82/-3.380
NNP (Translated Data) (errors)	2.87/-3.407 (+0.01/+0.04)	2.82/-3.333 (+0.02/-0.024)	2.81/-3.385 (-0.01/-0.005)

Table 3: Results of energy minimization from the different initial structures, FCC, BCC, and HCP aluminum through NNPs from translated data. (blue: lowest energy, red: highest energy, green: middle, l_b: bond length)

isobaric ensemble. Data imbalance issues exist because the MD sampling inevitably collects more configurations near the equilibrium state. Utilizing the active learning process based on atomic UQ, we successfully distilled the data to 5% without losing much accuracy. Furthermore, we investigated the obtained configurations that can be utilized for a very different system, aluminum by simple scaling and relabeling. Our results promise that there can be foundational data for these systems' potential energy surface (PES). In the future, we will explore the data set with DFT calculations. Once we establish the foundational data, it can revolutionize the developing process for NNPs and contribute more standard measurements for the model's performance.

References

[1] Kirstin Alberi, Marco Buongiorno Nardelli, Andriy Zakutayev, Lubos Mitas, Stefano Curtarolo, Anubhav Jain, Marco Fornari, Nicola Marzari, Ichiro Takeuchi, Martin L Green, et al. The 2019 materials by design roadmap. *Journal of Physics D: Applied Physics*, 52(1):013001, 2018.

[2] An Chen, Xu Zhang, and Zhen Zhou. Machine learning: accelerating materials development for energy storage and conversion. *InfoMat*, 2(3):553–576, 2020.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[4] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):83, 2019.

[5] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[6] David Pfau, James S Spencer, Alexander GDG Matthews, and W Matthew C Foulkes. Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Physical Review Research*, 2(3):033429, 2020.

[7] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[8] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[9] Mingwei Hu, Qiyang Tan, Ruth Knibbe, Miao Xu, Bin Jiang, Sen Wang, Xue Li, and Ming-Xing Zhang. Recent applications of machine learning in alloy design: A review. *Materials Science and Engineering: R: Reports*, 155:100746, 2023.

[10] Joshua Meyers, Benedek Fabian, and Nathan Brown. De novo molecular design and generative models. *Drug Discovery Today*, 26(11):2707–2715, 2021.

[11] Cheng Yan and Guoqiang Li. The rise of machine learning in polymer discovery. *Advanced Intelligent Systems*, 5(4):2200243, 2023.

[12] Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. Machine learning in drug discovery: a review. *Artificial Intelligence Review*, 55(3):1947–1999, 2022.

[13] Daniel P Tabor, Loïc M Roch, Semion K Saikin, Christoph Kreisbeck, Dennis Sheberla, Joseph H Montoya, Shyam Dwaraknath, Muratahan Aykol, Carlos Ortiz, Hermann Tribukait, et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature reviews materials*, 3(5):5–20, 2018.

[14] Yimo Han, Ming-Yang Li, Gang-Seob Jung, Mark A Marsalis, Zhao Qin, Markus J Buehler, Lain-Jong Li, and David A Muller. Sub-nanometre channels embedded in two-dimensional materials. *Nature materials*, 17(2):129–133, 2018.

[15] Hyoju Park, Gang Seob Jung, Khaled M Ibrahim, Yang Lu, Kuo-Lun Tai, Matthew Coupin, and Jamie H Warner. Atomic-scale insights into the lateral and vertical epitaxial growth in two-dimensional pd2se3–mos2 heterostructures. *ACS nano*, 16(7):10260–10272, 2022.

[16] Massimiliano Lupo Pasini, Gang Seob Jung, and Stephan Irle. Graph neural networks predict energetic and mechanical properties for models of solid solution metal alloy phases. *Computational Materials Science*, 224:112141, 2023.

[17] Jörg Behler and Gábor Csányi. Machine learning potentials for extended systems: a perspective. *The European Physical Journal B*, 94:1–11, 2021.

[18] Emir Kocer, Tsz Wai Ko, and Jörg Behler. Neural network potentials: A concise overview of methods. *Annual review of physical chemistry*, 73:163–186, 2022.

[19] Max Pinheiro, Fuchun Ge, Nicolas Ferré, Pavlo O Dral, and Mario Barbatti. Choosing the right molecular machine learning potential. *Chemical Science*, 12(43):14396–14413, 2021.

[20] Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021.

[21] Sönke Lorenz, Axel Groß, and Matthias Scheffler. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chemical Physics Letters*, 395(4-6):210–215, 2004.

[22] Burr Settles. Active learning literature survey. 2009.

[23] Gang Seob Jung, Jong Youl Choi, and Sangkeun Lee. Active learning of neural network potentials for rare events. 2023.

[24] Justin S Smith, Benjamin Nebgen, Nithin Mathew, Jie Chen, Nicholas Lubbers, Leonid Burakovsky, Sergei Tretiak, Hai Ah Nam, Timothy Germann, Saryu Fensin, et al. Automated discovery of a robust interatomic potential for aluminum. *Nature communications*, 12(1):1257, 2021.

[25] Fangjia Fu, Xiaoxu Wang, Linfeng Zhang, Yifang Yang, Jianhui Chen, Bo Xu, Chuying Ouyang, Shenzhen Xu, Fu-Zhi Dai, and Weinan E. Unraveling the atomic-scale mechanism of phase transformations and structural evolutions during (de) lithiation in si anodes. *Advanced Functional Materials*, page 2303936, 2023.

[26] Yuzhi Zhang, Haidi Wang, Weijie Chen, Jinzhe Zeng, Linfeng Zhang, Han Wang, and E Weinan. Dp-gen: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Computer Physics Communications*, 253:107206, 2020.

[27] Nobuyuki Nakajima, Haruki Nakamura, and Akinori Kidera. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *The Journal of Physical Chemistry B*, 101(5):817–824, 1997.

[28] Hisashi Okumura and Yuko Okamoto. Multibaric–multithermal ensemble molecular dynamics simulations. *Journal of computational chemistry*, 27(3):379–395, 2006.

[29] Yoshihide Yoshimoto. Extended multicanonical method combined with thermodynamically optimized potential: Application to the liquid-crystal transition of silicon. *The Journal of chemical physics*, 125(18), 2006.

[30] Gang Seob Jung, Yoshihide Yoshimoto, Kwang Jin Oh, and Shinji Tsuneyuki. Extended ensemble molecular dynamics for thermodynamics of phases. *arXiv preprint arXiv:2308.08098*, 2023.

[31] Y Zhang, R Ashcraft, MI Mendelev, CZ Wang, and KF Kelton. Experimental and molecular dynamics simulation study of structure of liquid and amorphous ni62nb38 alloy. *The Journal of chemical physics*, 145(20), 2016.

[32] MI Mendelev, MJ Kramer, Chandler A Becker, and M Asta. Analysis of semi-empirical interatomic potentials appropriate for simulation of crystalline and liquid al and cu. *Philosophical Magazine*, 88(12):1723–1750, 2008.

[33] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the national academy of sciences*, 99(20):12562–12566, 2002.

[34] Eric Darve, David Rodríguez-Gómez, and Andrew Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *The Journal of chemical physics*, 128(14), 2008.

[35] Gang Seob Jung, Hun Joo Myung, and Stephan Irle. Artificial neural network potentials for mechanics and fracture dynamics of two-dimensional crystals. *Machine Learning: Science and Technology*, 2023.

[36] Gang Seob Jung, Jong Youl Choi, Sangkeun Matthew Lee, and USDOE. Al-asmr: Active learning of atomistic surrogate models for rare events, 8 2023.