Neural Network Molecular Dynamics at Scale

Pankaj Rajak
Argonne Leadership Computing Facility
Argonne National Laboratory
Lemont, IL, U.S.
prajak@anl.gov

Kuang Liu
Collaboratory for Advanced Computing and Simulations
University of Southern California
Los Angeles, CA, U.S.
liukuang@usc.edu

Aravind Krishnamoorthy
Collaboratory for Advanced Computing and Simulations
University of Southern California
Los Angeles, CA, U.S.
kris658@usc.edu

Rajiv K. Kalia
Collaboratory for Advanced computing and Simulations
University of Southern California
Los Angeles, CA, U.S.
rkalia@usc.edu

Aiichiro Nakano
Collaboratory for Advanced Computing and Simulations
University of Southern California
Los Angeles, CA, U.S.
anakano@usc.edu

Ken-ichi Nomura
Collaboratory for Advanced Computing and Simulations
University of Southern California
Los Angeles, CA, U.S.
knomura@usc.edu

Subodh C. Tiwari
Collaboratory for Advanced Computing and Simulations
University of Southern California
Los Angeles, CA, U.S.
sctiwari@usc.edu

Priya Vashishta
Collaboratory for Advanced Computing and Simulations
University of Southern California
Los Angeles, CA, U.S.
priyav@usc.edu

***Abstract*—Neural network molecular dynamics (NNMD) simulations could revolutionize atomistic modeling of materials with quantum-mechanical accuracy at a fraction of computational cost. However, popular NNMD frameworks are generally implemented for a single computing node, and conventional energy-based NN models still suffer from large time-to-solution (T2S), prohibiting the application of NNMD to challenging materials simulations encompassing large spatiotemporal scales. Consequently, no leadership-scale NNMD simulation has thus far been reported. Here, we present a scalable parallel NNMD software (RXMD-NN) based on our scalable reactive molecular dynamics simulation engine named RXMD. RXMD-NN has achieved high scalability up to 786,432 IBM BlueGene/Q cores involving 1.7 billion atoms. Furthermore, we have achieved 4.6-fold reduction of T2S by using a novel network that directly predicts atomic forces from feature vectors. Reduced T2S has for the first time allowed the study of large-scale off-stoichiometry effect in a widely used phase change material, $Ge_2Se_2Te_5$, thereby resolving its "first-sharp diffraction peak mystery".**

Keywords—neural network, molecular dynamics, parallel computing, phase change materials, fist sharp diffraction peak, off-stoichiometry

I. Introduction

Molecular dynamics (MD) simulations follow the trajectories of all atoms to study material properties. For accurate description of materials, quantum MD (QMD) simulations compute interatomic forces quantum mechanically from first principles. The last decade has witnessed a surge of neural network MD (NMD) simulations, in which an artificial neural network (ANN) model is trained to reproduce the potential energy of QMD simulations [1-4]. Interatomic forces are then obtained by differentiating the energy with respect to atomic positions.

While NNMD substantially reduces the time-to-solution (T2S) while retaining the accuracy of QMD, it is still not fast enough to encompass necessary spatiotemporal scales to study many important material properties. This is primarily because most NNMD simulations use popular NNMD software packages such as the Atomic Energy Network (aenet) [5], which performs NNMD only on a single computing node using an external Python framework. A scalable parallel implementation of NNMD is urgently needed to perform large NNMD simulations on leadership-scale computers. The most widely used NNMD is based on an energy model where local atomic environment is encoded into a feature vector to train a neural network that predicts the potential energy of the system. With the energy model, atomic forces are obtained from analytical derivative of the neural network, which incurs an expensive force-evaluation step in every MD step. Another popular approach to machine-learning MD is Gaussian Approximation Potential (GAP) [6], where the potential energy surface is fitted against QMD dataset using Gaussian Process (GP). To predict atomic forces, however, the GP framework involves matrix inversion that scales O(N^3) with respect to the number of training examples, therefore it is not suitable for large-scale MD simulations.

II. Method Innovation

T2S-reducing force model: We have developed a direct force model employing a multi-layer perceptron model for NNMD simulations. An efficient feature vector design is essential to achieve accurate atomic-force prediction as well as fast evaluation of the feature vector during MD simulation. Based on the symmetric function proposed by Behler and Parrinello [7], we have designed a new feature vector incorporating (i) directional information of an atomic pair in the radial feature $G_{\sigma_i}^{2}(r_{ij})$ and (ii) bond vibrational modes (stretching $G_{\sigma_i}^{3a}(r_{ijk})$ and bending $G_{\sigma_i}^{3b}(r_{ijk})$) for an atomic

978-1-7281-7445-7/20/$31.00 ©2020 IEEE
DOI 10.1109/IPDPSW50202.2020.00167

triplet in the angular feature $G^3_{\sigma_i}(r_{ijk})$. These physically-based features are defined as

$$G^2_{\sigma_i}(r_{ij}) = \sum_j \frac{\sigma_{ij}}{|\vec{r}_{ij}|} e^{-\eta(r_{ij}-\mu)^2}$$

$$G^3_{\sigma_i}(r_{ijk}) = G^{3a}_{\sigma_i}(r_{ijk}) + G^{3b}_{\sigma_i}(r_{ijk})$$

$$G^{3a}_{\sigma_i}(r_{ijk}) = \sum_{jk} \frac{(1+\lambda\cos\theta_{jik})}{r_{ij}r_{ik}} \left\{ \frac{\sigma_{ij}}{r_{ij}}(r_{ij} - r_{ik}\cos\theta_{jik}) + \frac{\sigma_{ik}}{r_{ik}}(r_{ik} - r_{ij}\cos\theta_{jik}) \right\} S_t$$

$$G^{3b}_{\sigma_i}(r_{ijk}) = -\sum_{jk} \frac{(1+\lambda\cos\theta_{jik})^2}{r_{ij}r_{ik}} \left\{ \frac{\sigma_{ij}}{r_{ij}} + \frac{\sigma_{ik}}{r_{ik}} \right\} S_t$$

$$S_t = e^{-\eta(r_{ij}-\mu)^2} e^{-\eta(r_{ik}-\mu)^2} f_c(r_{ij}) f_c(r_{ik})$$

$$f_c(r_{ij}) = \begin{cases} 0.5\left(1+\cos\left(\frac{\pi r_{ij}}{r_c}\right)\right) & r_{ij} < r_c \\ 0 & r_{ij} \geq r_c \end{cases}$$

Here, r_{ij} is the interatomic distance between i-th and j-th atoms, η, μ and λ, are parameters of the modified symmetric function, r_c is the cutoff distance, and σ is the atomic coordinates in the x, y or z direction. We use one neural network to predict each of the atomic force components, F_x, F_y and F_z. This direct force prediction model eliminates the expensive evaluation of energy derivatives at every MD step as well as provides further parallelization opportunity by using one network for each force element. Fig. 1 schematically shows the force model, in which the physically-based feature functions are computed from local atomic environment, then subsequently are fed to the neural network to directly predict atomic force values. The ground truth and the predicted force values show an excellent agreement for Ge, Sb, and Te atoms in the x, y, and z directions respectively (Fig. 2).

Scalable parallel implementation: We have implemented the new NNMD force model in a linear-scaling parallel reactive MD software named RXMD [8]. RXMD was

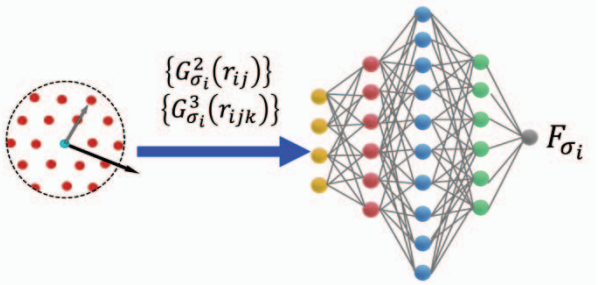

Fig. 1. Schematic of the neural-network force model. The physically-based symmetric functions, $G^2_{\sigma_i}(r_{ij})$ and $G^3_{\sigma_i}(r_{ijk})$, are computed from neighboring atomic coordinates. Neural network directly predicts atomic force F_σ.

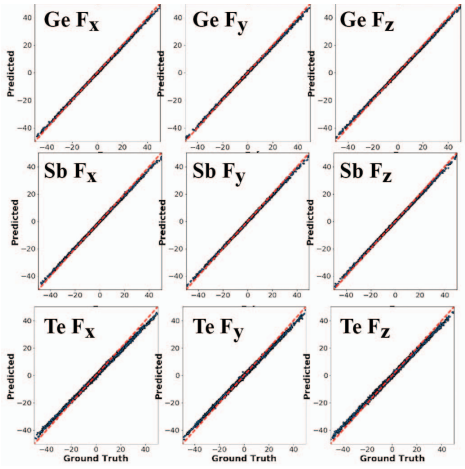

Fig. 2: Ground truth versus predicted force values of Ge, Sb, and Te atoms. Force components in the x, y, and z directions are labeled as F_x, F_y, and F_z, respectively.

originally developed for reactive MD simulations that scale from desktop computers to massively-parallel supercomputing platforms. To improve its parallel efficiency and T2S, a number of optimization techniques are used in RXMD, including a hybrid cell/neighbor list-based domain decomposition scheme to realize linear scaling for general n-tuple computations [9], an efficient six-way interprocess communication using Message Passing Interface (MPI) library, single-node performance improvement by Open Multi-Processing (OpenMP) multithreading and Shift-Collapse (SC) algorithm for provably minimum computation [10]. Though the number of combinations of parameters in the symmetry function is huge, its computational pattern is almost identical to the so called two-body and three-body interactions in MD, and thus these optimization techniques are readily applicable to the NNMD algorithm considered here. We also employ a mixed-precision approach that uses single precision for the symmetric function and force evaluation to reduce the amount of data transfer.

III. Performance Measurements

The performance of RXMD-NN has been tested on a $Ge_2Sb_2Te_5$ (GST) dataset using up to full-machine Mira at Argonne Leadership Computing Facility (ALCF). The Mira system consists of 48 computing racks, each of which has 1,024 nodes, providing 786,432 BlueGene/Q cores in total.

We first perform a weak-scaling test, in which the number of atoms per core N/P is kept constant. We measure the wall-clock time per simulation time step with scaled workloads — 25,920-atom GST system on each core (Fig. 3). By increasing the number of atoms linearly with the number of cores, the wall-clock time remains almost constant, indicating excellent scalability. To quantify the parallel efficiency, we first define the speed of the RXMD-NN code as a product of the total number of atoms and the number of NNMD time steps executed per second. The isogranular (or

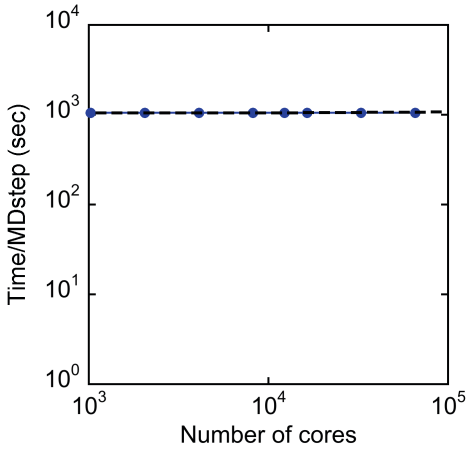

Fig. 3: Weak-scaling performance of the NNMD algorithm. Wall-clock time per MD step as a function of the number of BlueGene/Q cores up to 65,536. The number of GST atoms per MPI rank is kept constant as 25,920 that amounts to 1,698,693,120 GST atoms in total. Blue circle shows wall-clock time in second per MD step while black dashed line shows the theoretical peak performance.

weak-scaling) speedup is given by the ratio between the speed of P core and that of 1,024 cores as a reference system. With the granularity of 25,920 atoms per core, the parallel efficiency is 0.99 on 65,536 cores for a 1,698,693,120-atom system, shown in Fig. 3. This demonstrates a very high scalability of the RXMD-NN code.

We next perform a strong-scaling test by simulating GST with a total of 318,504,960 atoms. In this test, the number of cores ranges from P = 12,288 to 786,432, while keeping the total problem size constant. We measure the wall-clock time per NNMD time step as a function of P cores. The runtime is reduced by a factor of 55 on 786,432 cores compared with

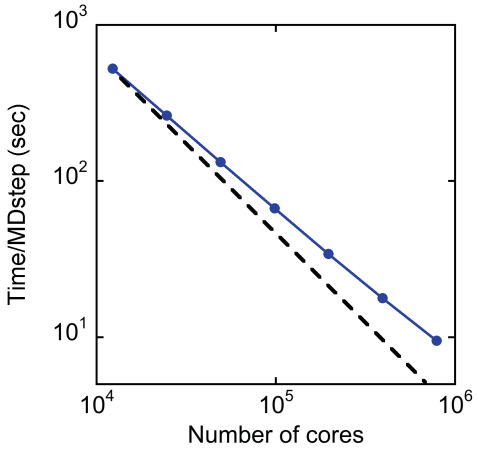

Fig. 4: Strong-scaling performance of the NNMD algorithm. The total problem size is kept constant as 318,504,960 GST atoms. Blue-line shows the obtained wall-clock time per MD step in seconds up to 786,432 BlueGene/Q cores, while black-dotted line shows the theoretical peak speed up using the smallest size benchmark on 12,288 BlueGene/Q cores as the reference.

the 12,288 cores run (*i.e.*, using 64-times larger number of cores); see Fig. 4. This signifies a strong-scaling speedup of 55, with the corresponding strong-scaling parallel efficiency of 0.861. It is more difficult to achieve high strong-scaling parallel efficiency compared with weak-scaling parallel efficiency, as the comparison of Figs. 3 and 4 suggests. This is due to the surge of communication/computation ratio as the workload per rank shrinks proportionally. With 64 times smaller system size of the weak-scaling test, the observed strong-scaling parallel efficiency is considered excellent.

IV. Applications

Thanks to the excellent scalability of the RXMD-NN code and reduced T2S by 4.6x compared to the standard energy model, we were able to address a challenging problem of medium-range structural order in off-stoichiometric GST for the first time. Phase change materials (PCM) are attracting great attention because of its novel engineering applications such as neuromorphic computing, photonics devices and scalable data storages [11-13]. GST is an archetypal PCM that shows fast switching behavior between high contrast amorphous and crystalline phases. Chalcogenide PCMs, such as GeTe and GST, exhibit reversible fast switching between crystal and amorphous phase by Joule heating. Around the phase transition temperature, GST shows orders-of-magnitude change in the electrical conductivity and the optical reflectivity, which has been used in optical rewritable media, and recently non-volatile memory devices. It is known that the performance of GST can be altered by their stoichiometry, however, investigating the non-stoichiometry effect poses a great challenge for QMD because their high computational cost prohibits from scaling the simulation system sufficiently large to describe a minute difference in the ratio of constituent atoms.

First sharp diffraction peak (FSDP) in neutron scattering experiment indicates the presence of medium-range structural order in amorphous materials with a network of randomly-distributed atoms. Although FSDP is an extremely useful indicator to characterize complex material phases, a large number of atoms, typically tens of thousands of atoms [14], is required to capture this salient feature using atomistic simulations even in stoichiometric systems. Investigating the non-stoichiometry effect on FSDP is expected to require even greater number of atomic configuration samples, therefore, this problem remains elusive because of the lack of a scalable simulation method for GST. While neutron scattering experiments indicate the existence of a clear FSDP peak, no MD simulation has thus far reproduced the feature.

We have applied the direct force model to investigate how the stoichiometry affects in GST to resolve this controversy. The initial configuration is created by QMD using NVT ensemble at 600 K and replicated by 8×8×8 times in the x, y, and z direction. The obtained system dimensions are 87.830×84.514×103.685 (Å^3). From the original $Ge_2Sb_2Te_5$ system with normal stoichiometry, we randomly remove

tellurium atoms to create two Te-deficient systems, namely $Ge_2Sb_2Te_{4.75}$ and $Ge_2Sb_2Te_{4.44}$. Table 1 summarizes the number of atoms per atom type in the three GST systems. We performed a series of melt-quench simulations using NNMD to create amorphous configuration. We have compared pair distribution function and bond-angle distribution to verify the obtained atomic configurations against QMD, and found the configurations generated by QMD and NNMD show excellent agreement.

TABLE 1. NUMBER OF ATOMS IN THE THREE GST SYSTEMS

System	Ge	Sb	Te
$Ge_2Sb_2Te_5$	5,760	5,760	14,400
$Ge_2Sb_2Te_{4.75}$	5,760	5,760	13,696
$Ge_2Sb_2Te_{4.44}$	5,760	5,760	12,800

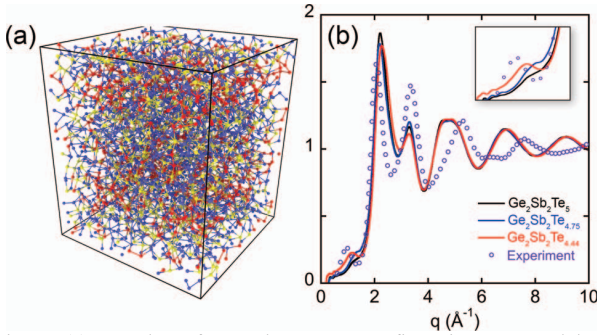

Fig. 5: (a) Snapshot of amorphous GST configuration generated by NNMD. Atoms are color-coded as red for Sb, yellow for Ge, and blue for Te, respectively (b) Neutron scattering structure factors $S_n(q)$ of GST systems with different stoichiometries, $Ge_2Sb_2Te_5$, $Ge_2Sb_2Te_{4.76}$, $Ge_2Sb_2Te_{4.44}$, and experiment. The inset shows a zoom-up view of $S_n(q)$ around small q value.

Fig. 5 shows the neutron scattering structure factors $S_n(q)$ of the three GST systems, $Ge_2Sb_2Te_5$, $Ge_2Sb_2Te_{4.76}$ and $Ge_2Sb_2Te_{4.44}$, along with an experimental result [15]. The inset of Fig.5(b) provides a close-up view of $S_n(q)$ around small q value signifying FSDP. The NNMD simulation shows that the reduction in the tellurium composition results in the increase of the peak height around q = 1.1(Å^{-1}), *i.e.* FSDP. Also, the peak position agrees well with the neutron scattering experiment, thereby resolving the FSDP mystery.

CONCLUSION

In this paper, we have presented a highly-efficient scalable NNMD algorithm to investigate phase change materials and its off-stoichiometry effect. Our neural network model directly predicts atomic forces based on physically-based features, which eliminates the expensive force evaluation step in the energy model. Taking advantage of a parallel MD engine that is highly-optimized for T2S, we have achieved nearly perfect parallel efficiency and 86% strong scaling efficiency using 786,432 BlueGene/Q cores of ALCF Mira. Our simulation shows a pronounced FSDP in the Te-deficient system, which indicates an increased medium-range structural ordering with deficient Te atoms.

ACKNOWLEDGEMENT

This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Science and Engineering Division, Grant DE-SC0018195. The simulations were performed at the Argonne Leadership Computing Facility under the DOE INCITE program, while scalable code development was supported by the Aurora ESP program.

REFERENCES

[1] J. Behler, "Constructing high-dimensional neural network potentials: a tutorial review," (in English), *International Journal of Quantum Chemistry,* vol. 115, no. 16, pp. 1032-1050, Aug 15 2015.

[2] K. Shimamura *et al.*, "Guidelines for creating artificial neural network empirical interatomic potential from first-principles molecular dynamics data under specific conditions and its application to alpha-Ag_2Se," (in English), *Journal of Chemical Physics,* vol. 151, no. 12, p. 124303, Sep 28 2019.

[3] V. L. Deringer, M. A. Caro, and G. Csányi, "Machine Learning Interatomic Potentials as Emerging Tools for Materials Science," *Advanced Materials,* vol. 1902765, p. 1902765, 2019.

[4] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," *npj Computational Materials,* vol. 3, p. 54, 2017.

[5] N. Artrith and A. Urban, "An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO_2," (in English), *Computational Materials Science,* vol. 114, pp. 135-150, Mar 2016.

[6] A. P. Bartok, M. C. Payne, R. Kondor, and G. Csanyi, "Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons," (in English), *Physical Review Letters,* vol. 104, no. 13, p. 136403, Apr 2 2010.

[7] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," *Journal of Chemical Physics,* vol. 134, 074106, 2011.

[8] K. Nomura, R. K. Kalia, A. Nakano, P. Rajak, and P. Vashishta, "RXMD: A scalable reactive molecular dynamics simulator for optimized time-to-solution," *SoftwareX,* vol. 11, p. 100389, 2020.

[9] M. Kunaseth, S. Hannongbua, and A. Nakano, "Shift/collapse on neighbor list (SC-NBL): fast evaluation of dynamic many-body potentials in molecular dynamics simulations," *Computer Physics Communications,* vol. 235, pp. 88-94, Feb 2019.

[10] M. Kunaseth, R. K. Kalia, A. Nakano, K. Nomura, and P. Vashishta, "A scalable parallel algorithm for dynamic range-limited n-tuple computation in many-body molecular dynamics simulation," *Proceedings of Supercomputing, SC13,* ACM/IEEE, 2013.

[11] K. Ding *et al.*, "Phase-change heterostructure enables ultralow noise and drift for memory operation," *Science,* vol. 23, p. eaay0291, 2019.

[12] W. Zhang, R. Mazzarello, and E. Ma, "Phase-change materials in electronics and photonics," *MRS Bulletin,* vol. 44, pp. 686-690, 2019.

[13] W. Zhang, R. Mazzarello, M. Wuttig, and E. Ma, "Designing crystallization in phase-change materials for universal memory and neuro-inspired computing," *Nature Reviews Materials,* vol. 4, pp. 150-168, 2019.

[14] A. Nakano, R. K. Kalia, and P. Vashishta, "First sharp diffraction peak and intermediate-range order in amorphous silica - finite-size effects in molecular-dynamics simulations," *Journal of Non-Crystalline Solids,* vol. 171, no. 2, pp. 157-163, Aug 1994.

[15] C. Qiao *et al.*, "The local structural differences in amorphous Ge-Sb-Te alloys," *Journal of Alloys and Compounds,* vol. 774, pp. 748-757, 2019.