A generalizable, uncertainty-aware neural network potential for GeSbTe with Monte Carlo dropout

Sung-Ho Lee[a], Valerio Olevano[b,c], Benoit Sklénard[a,*]

[a] *Univ. Grenoble Alpes, CEA, Leti, F-38000, Grenoble, France*
[b] *Université Grenoble Alpes, F-38000, Grenoble, France*
[c] *CNRS, Institut Néel, F-38042, Grenoble, France*

Abstract

A Bayesian neural network potential (NNP) achieved with the Monte Carlo dropout approximation method is developed for GeSbTe alloys. The Bayesian NNP is shown to be more generalizable than its classical counterpart, yielding reasonable predictions on structures that are not directly in the training configurations, and is able to output uncertainty estimates for the predictions. Its application to a molecular dynamics (MD) simulation is also presented, and the validity of the obtained trajectory is evaluated by comparing it to Density Functional Theory (DFT).

1. Introduction

GeSbTe (GST) alloys are interesting materials for a variety of technological applications like phase change memories or photonic devices [1]. Molecular dynamics (MD) simulations are often used to survey their properties and to optimise their composition, but studying finite temperature properties such as phase transition or thermal conductivity often require long simulations of extended systems that are too expensive for *ab initio* methods. Neural network potentials (NNP), trained on Density Functional Theory (DFT) references, can be used to drive an MD to overcome this limitation, which were recently shown to enable simulations at scales that were previously impossible at a near-DFT level accuracy [2].

NNPs, however, can silently fail on structures that lie outside the learned configuration space, which poses some issues when relying upon them for atomistic simulations. One solution is to apply the Bayesian paradigm to the neural network to capture the model's uncertainty along with its predictions. This has previously been demonstrated for the different phases of carbon [3], but never to more complex materials containing more than one chemical species.

In this work, a Bayesian NNP using the Monte Carlo (MC) dropout technique will be used on GST materials to evaluate its ability to estimate the predictive uncertainty as well as its superior generalizability compared to classical NNPs.

2. Methodology

2.1. Classical and Bayesian HDNNP

A feedforward neural network (NN) is the most common form of a neural network. It is organised as a set of layers with each layer, in turn, containing a set of nodes. Each j^{th} node of an l^{th} layer $y_j^{[l]}$ in a feedforward NN is connected to all the nodes of the previous layer via the following relationship

$$y_j^{[l]} = \sigma^{[l]}\left(\sum_i w_{ij}^{[l]} \cdot y_i^{[l-1]} + b_j^{[l]}\right) \tag{1}$$

where σ is known as an activation function that introduces non-linearity to the network, and w and b are weight and bias parameters that are optimised during training. The index i runs over all nodes of the $l-1^{\text{th}}$ layer. This is also represented schematically in Fig. 1.

The High Dimensional NNP (HDNNP) [2] is a type of NNP that is composed of a set of Atomic Neural Networks (ANN) that each takes one atom of a given system as the input and outputs a so-called "atomic energy", whose sum provides the total energy of the system. The ANNs are simple feedforward NNs that share their weights and biases amongst the same chemical elements.

Classically, feedforward neural networks contain point weights and output point estimates. In contrast, Bayesian neural networks place a probability prior over the weights and the estimates are obtained as a distribution by sampling from the posterior [4].

Formally, this can be expressed as follows. Starting with the Bayes' theorem

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D,H)}{\int_H P(D,H')dH'} \tag{2}$$

substituting θ, representing the weights and biases of the neural network, for the *hypothesis* (H), and the input, target pair of the training dataset (D_x, D_y) for the *data* (D) and using the relationship $P(A|B)P(B) = P(A,B)$ the following equation is obtained

$$P(\theta|D) = \frac{P(D_y|D_x,\theta)P(\theta)}{\int_\theta P(D_y|D_x,\theta')P(\theta')d\theta'} \tag{3}$$

*Corresponding author
Email addresses: sung-ho.lee@cea.fr (Sung-Ho Lee), benoit.sklenard@cea.fr (Benoit Sklénard)

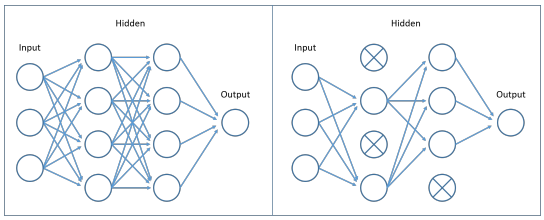

Figure 1: Schematic representations of a standard feedforward neural network (left) and a neural network with dropout, with the dropped nodes crossed out (right).

Lastly, with this Bayesian posterior of the parameters that take into account the training dataset, the predictive posterior can be used for Bayesian inference.

$$P(y|x, D) = \int_{\theta} P(y|x, \theta')P(\theta'|D)d\theta' \tag{4}$$

In practice, especially for complex functions like neural networks, computing and sampling from the posterior is intractable, mainly due to the difficulty associated with computing the evidence $\int_{\theta} P(D_y|D_x, \theta')P(\theta')d\theta'$ [5]. This is therefore performed indirectly by sampling $\theta_i \sim P(\theta|D)$ for $i = 0$ to N [4]. Monte Carlo dropout is one such approach to approximately sample from the posterior by using dropout at inference time [6].

Dropout was originally developed as a regularisation technique in which a set fraction of nodes of a neural network are randomly "dropped" (ie. set to zero) during training to prevent overfitting [7]. Similar to Eq. 1, dropout is expressed mathematically as follows

$$\begin{aligned} r_i^{[l]} &\sim \text{Bernoulli}(p) \\ \widetilde{y}_i^{[l-1]} &= r_i^{[l]} \cdot y_i^{[l-1]} \\ y_j^{[l]} &= \sigma\left(\sum_i w_{ij}^{[l]} \cdot \widetilde{y}_i^{[l-1]} + b_j^{[l]}\right) \end{aligned} \tag{5}$$

In standard dropout, the mask r is disabled at inference time and the values are rescaled according to the probability p to account for the increased number of connections. Therefore, for a given input, the output of the network is always constant. In contrast, the masking is maintained in Monte Carlo dropout and multiple forward passes through the network are performed for each input. The mask is drawn randomly from the Bernoulli distribution at each iteration which results in a distribution of outputs. The mean and standard deviation extracted from this distribution is then taken as the prediction and the corresponding uncertainty.

2.2. Computational Details

To generate the datasets, *ab initio* molecular dynamics (AIMD) calculations based on Density Functional Theory (DFT) were carried out using the VASP code [8, 9] at temperatures from 300 K to 1500 K at 300 K intervals for 20 ps for hexagonal GeTe, Sb_2Te_3 and $Ge_2Sb_2Te_5$ (containing 192, 240 and 216 atoms, respectively). 200 snapshots were then selected for each trajectory to compute accurate energies and forces. All DFT calculations were performed using the Perdew-Burke-Ernzerhof (PBE) [10] exchange-correlation functional and included the D3 dispersion correction to describe Van der Waals (VdW) interactions [11].

Our in-house library FFLearn was used to generate the NNPs, employing the symmetry functions proposed by Behler and Parrinello [12, 13] to encode the local environments around each atom for neural network input. After testing different configurations, all NNPs were defined to have 2 hidden layers with 15 nodes each. For the Bayesian NNPs, nodes were dropped with a probability of 10% save for the output layer that was kept at 0% to utilise all the available information, and sampled 100 times during inference.

3. Results and Discussion

For a neural network potential to be practicable for atomistic simulations, it must have the ability to generalize as much as possible to previously unseen configurations, and, perhaps more importantly, to recognise when it cannot.

Since a neural network potential can be thought of as a complex function modelling the potential energy surface (PES), regions on the PES that are outside the learned domain can lead to erroneous predictions. Being able to recognise when this occurs would make the NNP more reliable.

3.1. Comparison of Generalizability

In order to assess the generalizability, two NNPs (one Bayesian and one classical) were trained on the hexagonal GeTe 300 K, 600 K, and 900 K datasets (the training set) and made to predict the energies of the 1200 K and 1500 K snapshots (the test set). Root mean squared error (RMSE) between the predictions and the reference DFT energies was used as the accuracy metric, and the results are summarised in Table 1.

The higher temperature of the test sets was used to emulate structures drawn from outside the configuration space immediately spanned by the training set. The Bayesian NNP shows some promising results for the 1200 K dataset, with an RMSE of 53.1 meV/atom. The error becomes much greater for the 1500 K dataset which is expected to contain structures that are more amorphous-like, thus containing too many unknown configurations. Meanwhile, the classical NNP is unable to produce any valid results for either datasets.

Table 1: RMSE (meV/atom) of hexagonal GeTe between the DFT energy and the NNP predictions.

	Validation	1200 K	1500 K
Bayesian	3.18	53.1	1339.5
Classical	1.49	561.5	9185.0

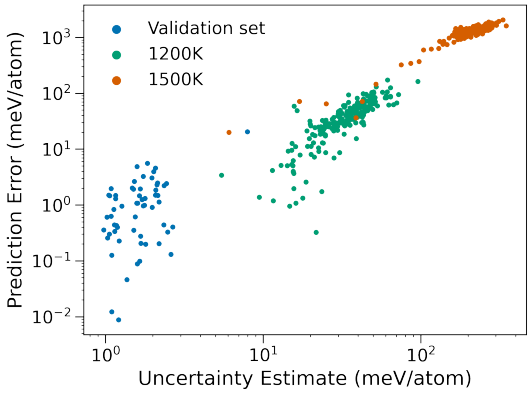

Figure 2: Uncertainty estimates for the hexagonal GeTe validation and test sets against the prediction errors given by the Bayesian NNP.

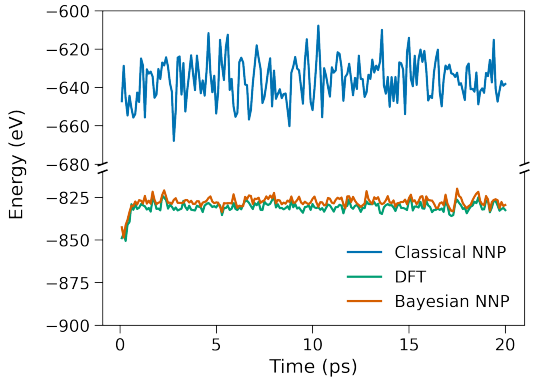

Figure 3: Energy predictions on $Ge_2Sb_2Te_5$ at 1200 K given by the classical and Bayesian NNPs trained on GeTe and Sb_2Te_3, compared to the DFT reference.

Table 2: RMSE (meV/atom) of $Ge_2Sb_2Te_5$ between the DFT energies and the NNP predictions at all temperatures.

Temperature (K)	300	600	900	1200	1500
Bayesian	70.4	50.0	32.0	14.6	25.6
Classical	992.7	964.2	933.6	902.2	455.5

3.2. Uncertainty estimation

The 1500 K example in the previous section highlights the importance of being able to estimate the predictive uncertainty. Fig. 2 shows that this is achievable with the Bayesian NNP, exhibiting a clear correlation between the uncertainty estimates and the prediction errors (computed as the absolute difference between the DFT reference energy and the NNP prediction). It is important to note, however, that this uncertainty is an estimate rather than a true reflection of the error, and care must be taken when interpreting it. Namely, they do not necessarily have a strictly linear relationship.

Making use of the uncertainty estimates typically involves setting a threshold, above which the structure is considered unknown to the model. Following this, an MD powered by an NNP (hereafter NNP-MD) can, for example, raise a warning or be stopped. The identified structure can also be used to improve the potential by means of active learning [14] or on-the-fly learning [15].

3.3. Predictions on $Ge_2Sb_2Te_5$

We now focus on $Ge_2Sb_2Te_5$ (GST225), a material that lies on the GeTe—Sb_2Te_3 tie-line on the ternary diagram. To compare the NNPs' generalizability to ternary alloys, a classical and a Bayesian NNP were each trained on the two extrema (GeTe and Sb_2Te_3) and tested on $Ge_2Sb_2Te_5$ at different temperatures. The results are summarised in Table 2, and the lowest error case (1200 K) is represented graphically in Fig. 3. Similar to the results in Sec. 3.1, the classical HDNNP fails to give meaningful predictions.

On the other hand, the relative success of the Bayesian NNP is surprising, given that it has not been supplied with any direct knowledge of the Ge–Sb or the Ge–Sb–Te relationships, or any permutations thereof. Nonetheless, these results hint at the possibility of modelling other GST compositions along the GeTe—Sb_2Te_3 tie-line with only a minimal dataset on the GST itself. As the dataset generation is the most expensive step to training NNPs, this could prove to be a very powerful development. However, more work is required to confirm these findings on different stoichiometries of GST, and to understand the mechanism behind this extrapolation capability before physically relevant properties can be calculated.

It should be noted that the particularly low RMSE for the 1200K dataset for the Bayesian NNP has no physical basis. Neural network training is a stochastic process that involves a few different random sampling operations. For example, the initial weights must be drawn from a probability distribution and the training dataset is randomly shuffled during training. In addition, as neural networks are very high dimensional, there exist many local minima. As such, different initial conditions practically always converge to different minima, which leads to slightly different results being produced for each new trained potential. Consequently, the lowest RMSE may lie at another temperature if the experiment was repeated, but the fluctuations in the values will remain minor. Though this effect is amplified here as GST225 lies far outside the trained configuration space (due to the lack of the Ge–Sb and the Ge–Sb–Te interactions in the training data, as mentioned above), it is still clear that the Bayesian NNP can give much more meaningful predictions than the classical.

3.4. Radial Distribution Function

To validate the Bayesian NNP on its ability to predict kinetic properties of periodic systems, an MD simulation was performed in the canonical (NVT) ensemble with the Nosé-Hoover thermostat at 1200 K for 20 ps followed by a simulation in the microcanonical (NVE) ensemble for 10 ps for the hexagonal GeTe using the LAMMPS [16] package. The potential was trained on all GeTe datasets mentioned in Sec. 2.2. The size of the simulation cell was intentionally kept relatively small (192 atoms; same as the training set structures) in order to compare against DFT.

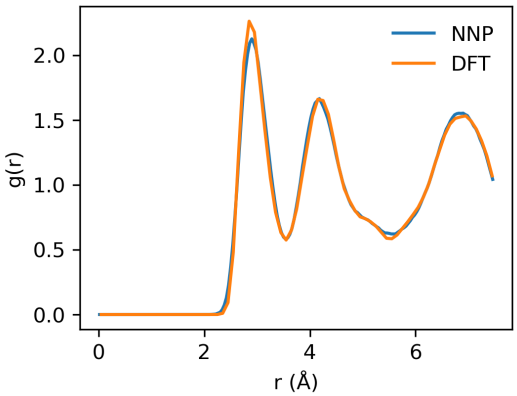

Figure 4: A comparison of the total radial distribution function for GeTe at 1200 K between the AIMD trajectory and the NNP-MD trajectory.

From the NVE trajectory, a time-averaged total radial distribution function (RDF; g(r)) was extracted, taking into account all interatomic distances (Ge – Ge, Ge – Te and Te – Te). As shown in Fig. 4, the RDF shows a good agreement between the NNP and DFT. In addition, the NNP-MD was completed ~35 times faster than the AIMD, running at a speed of 231.98 timesteps/s, compared to 6.65 timesteps/s for the DFT. Reducing the number of sampling steps for the Bayesian NNP will further increase the performance, and it is likely that even sampling only 10 times will result in only a minimal difference in the accuracy of the final prediction [6]. Moreover, as HDNNPs scale quasi-linearly with the number of atoms [2], while DFT follows cubic scaling with the number of electrons, this performance gain is only expected to increase as the system size grows.

In contrast, a classical NNP trained under the same conditions was unable to run an MD due to unphysically high forces being predicted, causing the simulation to crash. As a result, no RDF for the classical NNP could be computed for comparison.

4. Conclusion

In this work, a Bayesian neural network potential for GeSbTe was developed using Monte Carlo dropout [6]. It was shown to have an improved generalizability compared to its non-Bayesian counterpart on hexagonal GeTe, with insightful estimates on the uncertainties of the predictions. Molecular dynamics simulations performed with a Bayesian NNP exhibited good agreement with the *ab initio* MD, shown in the form of the radial distribution function. In addition, a significant improvement of the computational cost was also observed. These indicate that neural network potentials can reliably be used for studies of physical properties requiring atomistic simulations at large scales, such as the thermal conductivity. Furthermore, it surprisingly was able to reasonably calculate the energies of hexagonal $Ge_2Sb_2Te_5$, even in the absence of information on the two-body Ge – Sb and the three-body Ge – Sb – Te relationships in the training dataset. Future work will involve investigating this in more detail to better understand its underlying mechanism to try to take advantage of this for a cost-efficient training of a general GST potential.

Acknowledgment

This work was performed using HPC resources from GENCI–IDRIS (Grant 2021-A0110911995) and was partially funded by European commission through ECSEL-IA 101007321 project StorAIge and the French IPCEI program.

References

[1] S. R. Elliott, Chalcogenide Phase-Change Materials: Past and Future, International Journal of Applied Glass Science 6 (2015) 15–18. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijag.12107. doi:10.1111/ijag.12107, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijag.12107.

[2] J. Behler, Four Generations of High-Dimensional Neural Network Potentials, Chemical Reviews 121 (2021) 10037–10072. URL: https://pubs.acs.org/doi/10.1021/acs.chemrev.0c00868. doi:10.1021/acs.chemrev.0c00868.

[3] M. Wen, E. B. Tadmor, Uncertainty quantification in molecular simulations with dropout neural network potentials, npj Computational Materials 6 (2020) 124. URL: https://www.nature.com/articles/s41524-020-00390-8. doi:10.1038/s41524-020-00390-8.

[4] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, M. Bennamoun, Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users, IEEE Computational Intelligence Magazine 17 (2022) 29–48. URL: https://ieeexplore.ieee.org/document/9756596/. doi:10.1109/MCI.2022.3155327.

[5] A. Graves, Practical Variational Inference for Neural Networks, in: Advances in Neural Information Processing Systems, volume 24, Curran Associates, Inc., 2011, p. 9. URL: https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

[6] Y. Gal, Z. Ghahramani, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: Proceedings of The 33rd International Conference on Machine Learning, volume 48, 2016, pp. 1050–1059.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[8] G. Kresse, J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, Physical Review B 54 (1996) 11169–11186. URL: https://link.aps.org/doi/10.1103/PhysRevB.54.11169. doi:10.1103/PhysRevB.54.11169.

[9] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Physical Review B 59 (1999) 1758–1775. URL: https://link.aps.org/doi/10.1103/PhysRevB.59.1758. doi:10.1103/PhysRevB.59.1758.

[10] J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865–3868. URL: https://link.aps.org/doi/10.1103/PhysRevLett.77.3865. doi:10.1103/PhysRevLett.77.3865.

[11] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, The Journal of Chemical Physics 132 (2010) 154104. URL: https://doi.org/10.1063/1.3382344. doi:10.1063/1.3382344. arXiv:https://doi.org/10.1063/1.3382344.

[12] J. Behler, M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Physical Review Letters 98 (2007) 146401. URL: https://link.aps.org/doi/10.1103/PhysRevLett.98.146401. doi:10.1103/PhysRevLett.98.146401.

[13] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. (2011) 14.

[14] A. Krogh, J. Vedelsby, Neural Network Ensembles, Cross Validation, and Active Learning, in: Advances in Neural Information Processing Systems, volume 7, MIT Press, 1994. URL: https://proceedings.neurips.cc/paper/1994/hash/b8c37e33defde51cf91e1e03e51657da-Abstract.html.

[15] R. Jinnouchi, K. Miwa, F. Karsai, G. Kresse, R. Asahi, On-the-Fly Active Learning of Interatomic Potentials for Large-Scale Atomistic Simulations, The Journal of Physical Chemistry Letters 11 (2020) 6946–6955. URL: https://pubs.acs.org/doi/10.1021/acs.jpclett.0c01061. doi:10.1021/acs.jpclett.0c01061.

[16] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. i. t. Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, S. J. Plimpton, LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, Comp. Phys. Comm. 271 (2022) 108171. doi:10.1016/j.cpc.2021.108171.