# High Efficiency Configuration Space Sampling – probing the distribution of available states

Paweł T. Jochym* and Jan Łażewski

Institute of Nuclear Physics, Polish Academy of Sciences, Cracow, Poland
* pawel.jochym@ifj.edu.pl

January 20, 2021

## Abstract

**Substantial acceleration of research and more efficient utilization of resources can be achieved in modeling investigated phenomena by identifying the limits of system's accessible states instead of tracing the trajectory of its evolution. The proposed strategy uses the Metropolis-Hastings Monte-Carlo sampling of the configuration space probability distribution coupled with physically-motivated prior probability distribution. We demonstrate this general idea by presenting a high performance method of generating configurations for lattice dynamics and other computational solid state physics calculations corresponding to non-zero temperatures. In contrast to the methods based on molecular dynamics, where only a small fraction of obtained data is consumed, the proposed scheme is distinguished by a considerably higher, reaching even 80%, acceptance ratio.**

## Contents

**1 Introduction** **1**

**2 General idea of HECSS** **2**

**3 Sampling of probability distribution** **4**

**4 Convergence of derived quantities** **7**

**5 Conclusions** **8**

**References** **8**

## 1 Introduction

Every system can be successfully studied by methodical observation of its behaviour for a long enough time. However, especially for slowly changing characteristics, this could take proverbial eons. On the other hand, some elementary knowledge of possible features and existing constrains allows one to limit available states of the studied system and determine the probability distribution of these states in the configuration space. As a result, the system can be modeled based on its probable configurations. To illustrate this idea, we present its application to studies of vibrational properties of solids.

A number of problems in solid state physics connected with lattice dynamics can be effectively addressed with inter-atomic potential models constructed using data obtained from quantum mechanical calculations (e.g. Density Functional Theory – DFT). Probably the simplest of such models is harmonic approximation developed by Born and von Kármán at the beginning of the 20th century [1–3]. Over the years multiple increasingly more sophisticated models have been developed: Quasi-Harmonic approximation (QHA) [4], Temperature-Dependent Effective Potential [5–7], Self-Consistent Phonons (SCPH) [8] or Parlinski's approach [9], to name just a few. All the above mentioned schemes share common feature – they need an appropriate set of data to build a model of inter-atomic potential which is essential for this type of methods. The data set should correspond to the system at thermal equilibrium or other physical state. It is usually comprised of atomic positions as well as resulting energies and forces calculated with some quantum mechanical (e.g. DFT) or even effective potential method.

Presently, molecular dynamics is often used to investigate systems at non-zero temperature in thermal equilibrium. This is done either directly – by analysis of the MD trajectory – or as a source of configurations for building the mentioned effective models of the inter-atomic potential to be used in further analysis (e.g. with programs like ALAMODE [10, 11] or TDEP [7]). Both cases involve a very costly stage of running long MD calculations [12]. Since uncorrelated configurations from different parts of the phase space are required, they are generated by appropriate spacing of the sampling points over the computed trajectory or even by performing multiple independent MD runs. At the end only a small fraction of calculated configurations is used (typically 1-10%). Therefore, using MD in this context is exceedingly wasteful. This makes it not only very expensive but also useless for larger and more complicated systems (of hundreds or more atoms), where even static, single-point DFT calculations are challenging. In such cases running a 30000 steps MD becomes prohibitively expensive and impractical.

In this work we propose a new, High Efficiency Configuration Space Sampling (HECSS) method for modelling systems in non-zero temperature, including non-harmonic effects, without using MD trajectory. We also indicate its possible application to some additional cases like disordered systems or large, complicated systems.

## 2 General idea of HECSS

To reproduce the thermal equilibrium in the system, independent configurations of displacements consistent with a desired non-zero temperature should be selected. Having any initial approximations for the lattice dynamics of the system (e.g. standard harmonic approach [2, 4, 13]) one can estimate temperature-dependent atomic mean-square-displacements (MSD) from a small set of force-displacement relations. Using these MSD data as a first approximation, the atomic displacements with normal distribution around equilibrium positions can be easily generated. There is, however, a subtle issue around displacements generated this way – they are *uncorrelated* between atoms, while in reality atomic displacements are correlated at least for their close neighbours. For example, it is easy to see that a simultaneous out-of-phase movement of neighboring atoms towards or away from each other will generate larger changes in energy than a synchronous in-phase movement of the same atoms. The former configuration should be represented with lower probability than the later, instead of equal probability present in the above simplistic scheme. Thus, while the static configurations generation may be a correct direction in general, such a naive approach is not sufficient. One can see that some additional mechanism is required to adjust probability distribution of generated samples in order to accurately

reproduce configurations drawn from thermodynamic equilibrium ensemble. Classical statistical mechanics points to such a scheme for selection of configurations representing a system in thermal equilibrium.

The general form of the equipartition theorem says that a generalized virial for any phase space coordinate (i.e. generalized coordinate or momentum) is proportional to temperature when it is averaged over the whole ensemble:

$$\left\langle x_m \frac{\partial H}{\partial x_n} \right\rangle = \delta_{mn} k_B T. \tag{1}$$

If we assume ergodicity of the system, the ensemble average may be replaced with time average. For momenta this leads to the average kinetic energy per degree of freedom being equal to $k_B T/2$ and provides the kinetic definition of temperature. However, the relation holds also for derivatives of Hamiltonian with respect to positions. Considering relation (1) for some atomic displacement $q$ from the equilibrium configuration, and assuming the potential energy depends only on position, we can write position-dependent part of the Hamiltonian (i.e the potential energy $E_p(q)$) as a Taylor's expansion with respect to the atomic displacement $q$ from the equilibrium configuration:

$$E_p(q) = \sum_{n=2}^{\infty} C_n q^n, \tag{2}$$

where the expansion coefficients $C_n$ are, in general, functions of all remaining coordinates (displacements). The equipartition theorem (1) now takes the form:

$$k_B T = \left\langle q \sum_{n=2}^{\infty} n C_n q^{n-1} \right\rangle = \sum_{n=2}^{\infty} n C_n \left\langle q^n \right\rangle \tag{3}$$

and if we write $n$ as $(n-2)+2$ and divide both sides by 2 we get:

$$\langle E_p(q) \rangle = \frac{k_B T}{2} - \sum_{n=3}^{\infty} \frac{n-2}{2} C_n \left\langle q^n \right\rangle, \tag{4}$$

which is similar to the kinetic energy counterpart except for an additional term generated by the anharmonic part of the potential and defined by the third and higher central moments of the probability distribution of the displacements. If we can assume that the second term of the Eq. 4 is small in comparison with $k_B T$, we get a formula for the average potential energy of the system. Note that for harmonic systems the second part vanishes. For anharmonic systems omission of higher terms in Eq. 4 will provide first-order approximation of the mean potential energy. Only experience can tell us how good this approximation is and how wide its applicability range is. However, one should note that substantial higher-order terms are present only in parts of the formula connected with strongly anharmonic modes. Furthermore, for every atom in centro-symmetric position all odd-power moments vanish and the first non-zero moment is the fourth one. Finally, the formula for the potential energy of the whole system contains similar terms for all modes. Judging by extremely high efficiency of harmonic approximation for crystal lattice dynamics, we can expect that this averaging will make proposed approximation effective for a wide range of systems.

To sum up, MD provides a representation of the system with the properly distributed kinetic energy. For a single particle it is a Maxwell-Boltzmann distribution. By virtue of the central limit theorem (CLT) [14, 15], if we increase the number of particles we will approach at infinity (i.e. in the thermodynamical limit) a Gaussian distribution with the same average (the same mean) and the variance which is scaled as inverse number of

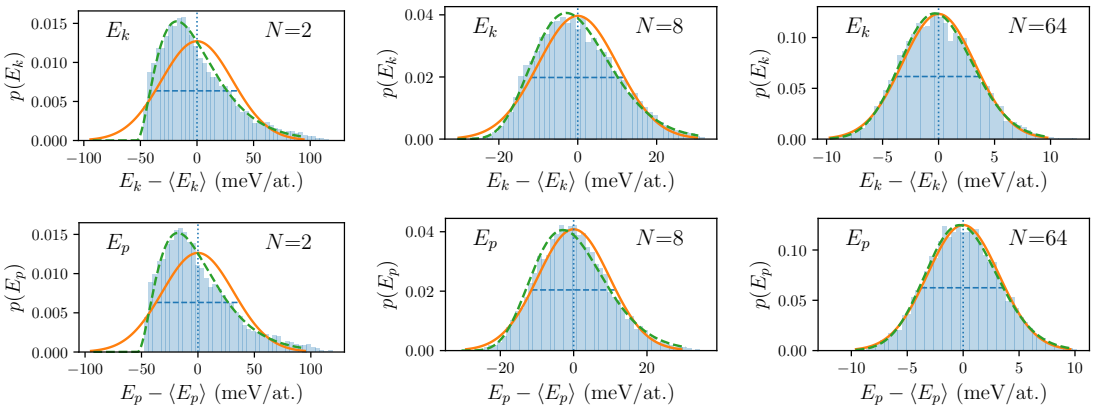


Figure 1: Probability distribution for single atom kinetic (upper) and potential (lower) energies averaged over $N = 2, 8$ and 64 randomly selected atoms. Solid orange lines show fitted normal distributions while dashed green lines show $\chi^2$ distribution for $3N$ degrees of freedom fitted to kinetic energy histograms. Data derived from the MD trajectory.

particles. As we can see for kinetic energy the relation is very simple whereas for the potential energy we have a quantity approximately close to temperature if the system is not too far from a harmonic one. Nevertheless, we do not know, in general, the form of the distribution of the potential energy. That constitutes substantial difficulty, which fortunately can be overcome by application of the CLT to calculate distribution of potential energy.

The CLT states that for any reasonable probability distribution, the distribution of the mean of the sample of the independent random variable drawn from it tends to the normal distribution with the same mean and variance scaled by the square root of the number of samples. The *reasonable* class is very broad here, certainly containing all physically interesting cases by virtue of requiring a finite variance and a well-defined mean. Thus, for potential energy per degree of freedom we can expect the probability distribution converging to the normal distribution:

$$\lim_{N\to\infty} p(E_p) = \mathcal{N}(\langle E_p \rangle, \sigma/\sqrt{3N}). \tag{5}$$

As shown above, one can approximate the $\langle E_p \rangle$ with the first term of Eq. 4 and the only unknown parameter in this formula is the variance of the distribution. Note that above expression is *independent* from the particular shape of the potential energy probability distribution for the single degree of freedom except of its mean $\langle E_p \rangle$ and variance $\sigma$.

However, we need to consider that the Eq. 5 is true *asymptotically*. At this point we need to decide if this relation has any practical use for *finite*, and preferably not too large, $N$. The common wisdom in statistical community states that for $N$ above $\approx 50$ the distribution of the average is practically indistinguishable from the true normal distribution, and even for smaller $N$, if the starting distribution is not too wild, the convergence is usually very quick.

## 3 Sampling of probability distribution

To verify if this "folk wisdom" holds true for the typical kinetic and potential distributions we have checked this hypothesis against actual MD data of a typical system. This test does not require high-accuracy forces and energies but demands ability to efficiently calculate

moderately sized systems (e.g. 1000 atoms). Thus, instead of using DFT as a source of energies/forces we have used effective potential model of the cubic 3C-SiC crystal. We have used LAMMPS [16] implementation of the potential with parameters derived in [17,18] and the NVT-MD implemented in ASAP3 module of the Atomistic Simulation Environment (ASE) [19]. High performance of this implementation allowed for $5 \cdot 10^4$ time steps (of 1 fs length) runs of the $5 \times 5 \times 5$ supercell (1000 atoms) to be executed on a single server in just a few hours.

The kinetic and potential energy probability distributions extracted from MD runs of systems of 2, 8 and 64 atoms (i.e. 6, 24, 192 degrees of freedom) are presented in Fig. 1. At this stage we are interested in the speed of convergence of the probability distribution, and this experiment shows that for typical distributions present in crystals the convergence is quite quick. Already at the $N_{DOF} = 24$ (i.e. 8 atoms) the deviation from the normal distribution is barely noticeable and at $N_{DOF} = 192$ (i.e. 64 atoms) it is indeed hardly visible. And that seems to hold true equally well for distributions of the kinetic and potential energy.

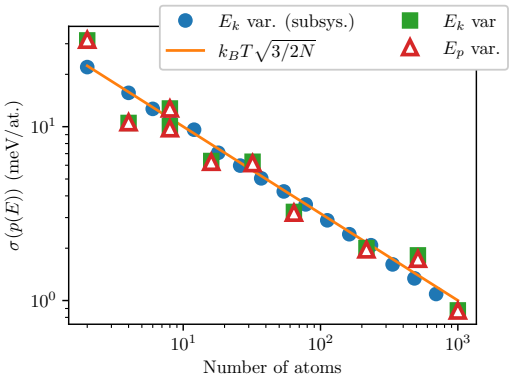


Figure 2: Variance of the energy distribution as a function of system size compared with prediction of the central limit theorem (orange line). Results for different numbers of randomly chosen coordinates of $5 \times 5 \times 5$ system (blue circles) were put together with variance of both the kinetic (green squares) and potential (red triangles) energies of smaller systems (defined in the text).

This simple example demonstrates that for our practical purposes we can expect the central limit theorem in application to energy distribution in crystals to hold above $\approx 30$ degrees of freedom, for both the kinetic and potential energies. This means that we can apply this approach even for very moderately sized systems of 10-20 or more atoms.

The energy distributions in Fig. 1, derived from the MD runs mentioned above, show clearly Gaussian distributions for both the kinetic and potential energies even for $N_{DOF} \approx 25$ degrees of freedom. Furthermore, the variance of these distributions plotted against the system's size and shown in Fig. 2 follows closely CLT prediction in Eq. 5. And this holds true for parts of a larger system (blue circles in Fig. 2) as well as for the whole smaller crystals (squares and triangles in Fig. 2). The dispersion of small systems' data in Fig. 2 is due to large temperature fluctuations in small sets of particles.

Thus, we have checked that, at least in our test case, the convergence to thermodynamic and CLT limits required by the Eqs 4 and 5 is quick enough to be useful in practical calculations for systems of just tens of atoms. The main problem now is that there is no direct access to potential energy and there is no way to invert relation from positions to potential energy – even in principle since the relation is many-to-one. Our goal here is to reproduce the potential energy distribution described by Eq. 5 and present in MD data by intelligently sampling the configuration space of the system – since this is the only input we can directly specify. Fortunately, computational statistics provides multiple algorithms dedicated to the task of sampling of indirectly specified probability distributions. In particular, the Metropolis-Hastings Monte Carlo [20] seems well suited to our purposes. To use it effectively we need to generate a prior distribution which covers the domain and, preferably, is fairly close to the target distribution. Obviously, we are unable to generate configurations corresponding to the distribution from Eq. 5 but we can use physically motivated approximation. We propose to approximate displacements of atoms in the

system by Gaussian probability distribution with variance tuned to the temperature and to the resulting energy. Our HECSS software package provides the Metropolis-Hastings implementation together with a tuned prior probability distribution generator. The tuning algorithm adjusts the variance of the atomic displacement distribution in each step: $\sigma_{n+1} = (1 + s(E_p(x_n)))\sigma_n$, according to the modified logistic sigmoid function:

$$s(E_p) = \delta \left( \frac{2}{1 + e^{-(E_p - E_0)/(w \cdot \sigma_{E_p})}} - 1 \right), \tag{6}$$

where $\sigma_{E_p} = k_B T \sqrt{3/2N}$ is the variance of the target potential energy distribution (5) and $\delta \approx 0.005 - 0.02$ is a small tuning parameter controlling the speed of the variance adjustment, while $w \approx 3$ controls the width of the prior distribution. Both parameters have substantial practical importance – they influence the effectiveness of the procedure – but play no fundamental role in the algorithm. Changing these parameters to the unsuitable values leads only to slower convergence of the procedure, since the Metropolis-Hastings algorithm is guaranteed to asymptotically produce the target distribution for any non-vanishing prior distribution. A good selection of the prior distribution means getting higher than 50% (in practice even above 80%) acceptance ratio instead of a few percent or even less if the prior distribution is very far from the target. The prior distribution we are proposing here is already of similar shape to the target one, thus it results in several additional samples at the start of the procedure to properly tune the width parameter – if it was not set correctly.

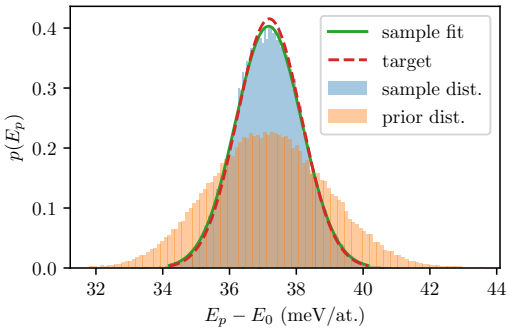


Figure 3: Prior energy probability distribution (orange filling) versus target distribution (blue filling). The lines indicate target distribution (red dashed line) and Gaussian distribution fitted to generated sample (green solid line).

The typical good relationship between prior and target distribution as well as the sampling produced by the proposed algorithm is illustrated in Fig. 3. Furthermore, the near-independent drawing of each step in the algorithm means that each sample from the produced set is potentially usable. Therefore, the burn-in period may be reduced to just a few samples required for tuning of the prior distribution parameters. The only source of possible correlations between samples in consecutive steps is the change in variance of the prior distribution, which is tuned after each step according to the sigmoid function (defined by Eq. 6). This is a very weak correlation since the variance is not supposed to change by more than $\delta \approx 0.5 - 2\%$. What is more, these parameters are independent from the size of the system and their values are not critical. The variance of the prior distribution, which is self-tuning, should be estimated with 20% accuracy in order for the burn-in period to be reduced to just one or two samples. Thus, the initial tuning may be performed using a small

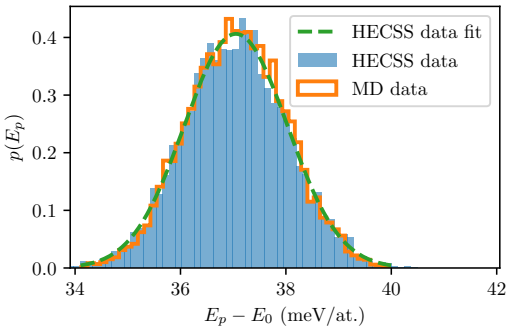


Figure 4: Probability distribution of potential energy per atom generated with HECSS scheme (blue shape) versus distribution extracted from the MD trajectory (orange contour). The dashed line indicates normal distribution fitted to HECSS sample.

supercell or even a primitive unit cell – depending on the system – by just recording the self-tuning trajectory of the algorithm and adapting initial parameters of the production run accordingly.

The end result of these algorithms is a series of samples (i.e. configurations) which reproduce expected probability distribution (5) of potential energy for the system in thermal equilibrium at the target temperature. The comparison between the potential energy probability distribution in the samples generated by HECSS and extracted from the MD run is depicted in Fig. 4.

## 4 Convergence of derived quantities

The results presented above demonstrate that it is possible to effectively generate samples with potential energy distributions consistent with the data from the MD trajectories. The remaining, much more difficult, question is whether these samplings indeed provide an appropriate representation of the system in thermal equilibrium at a given temperature. This issue may be tested in various ways. In this work we propose to check if the potential model built basing on the HECSS-generated displacement-force data provides phonon frequencies and lifetimes consistent with those derived from the MD trajectory data.

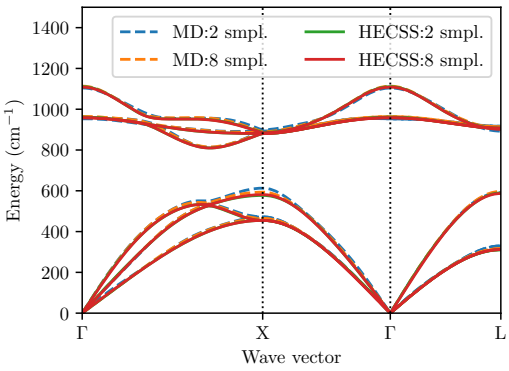


Figure 5: Consistency and convergence of phonon frequencies in 3C-SiC crystal determined with harmonic model derived from MD (dashed lines) and HECSS (solid lines) data.

Therefore, we have compared the results obtained from the calculations of 3C-SiC crystal with LAMPPS potential used in the previous section. The samples obtained by both methods (i.e. MD and HECSS) have been used to build force constants matrices for the material with ALAMODE program. Both second- and third-order force constants have been determined based on different number of samples. The resulting phonon frequencies and lifetimes are presented in Fig. 5 and Fig. 6, respectively. These findings demonstrate not only high-level of consistency between both data sets and models, but also similar convergence characteristics between both methods.

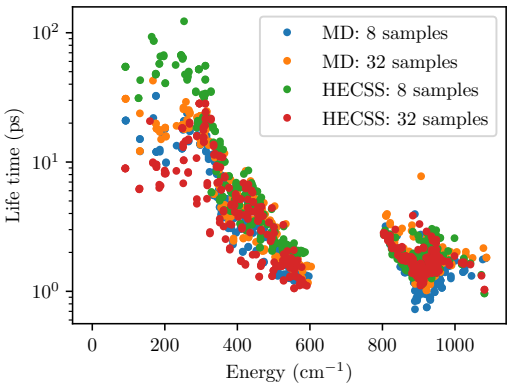


Figure 6: Consistency and convergence of 3C-SiC phonon lifetimes derived with third order model based on MD (blue and orange) and HECSS (green and red) generated data.

It is important to note that HECSS-generated data sets consist of first $N$ generated samples (after initial burn-in period of 3 samples), not the $N$ samples selected from the larger set, as it is done with MD trajectory. Obviously, if one were forced to run as many steps of HECSS algorithm as time steps, of the MD trajectory the whole effort would be pointless. It is evident that the results of both approaches are very similar, despite a large difference in necessary computational effort – which provides a clear justification for future application of the presented method to the

much more expensive DFT-based variant of the potential energy calculation.

## 5 Conclusions

We have introduced a new high efficiency configuration space sampling (HECSS) scheme as an alternative for application of Molecular Dynamics as a source of configurations representing systems at non-zero temperatures. The results presented above demonstrate potential of the proposed HECSS method to generate faithful configuration samplings for systems in thermal equilibrium, which can be used to investigate anharmonic effects present in crystalline solids. It is worth noting that this method is not limited to crystals or to only geometric degrees of freedom. In principle, it is possible to extend its applicability to magnetic degrees of freedom or disordered systems. Furthermore, due to its inherent ability to provide 3×number-of-atoms force-displacement data points per configuration, it reduces number of DFT calculations required for simple harmonic model determination. This reduction is much more pronounced in higher-order models, where number of independent variables is usually large. It should also be emphasized that the generated samples are drawn from the physically meaningful distribution and not from the non-physical, single axis displacements. This difference may become important if there is any substantial anharmonicity in the system, which couples degrees of freedom.

## Acknowledgments

The authors would like to express their gratitude to Krzysztof Parlinski, Przemysław Piekarz, Andrzej M. Oleś, and Małgorzata Sternik for very inspiring and fruitful discussions. This work was partially supported by National Science Centre (NCN, Poland) under grant UMO-2017/25/B/ST3/02586.

## References

[1] M. Born and T. V. Kármán, *Über Schwingungen in Raumgittern. (German) [On vibrations in space lattices]*, Physikalische Zeitschrift **13**(8), 297 (1912).

[2] A. A. Maradudin, *Theory of Lattice Dynamics in the Harmonic Approximation*, New York (1963).

[3] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, Oxford University Press (1988).

[4] B. Fultz, *Vibrational thermodynamics of materials*, Progress in Materials Science **55**(4), 247 (2010), doi:10.1016/j.pmatsci.2009.05.002.

[5] O. Hellman, I. A. Abrikosov and S. I. Simak, *Lattice dynamics of anharmonic solids from first principles*, Physical Review B **84**(18), 180301 (2011), doi:10.1103/PhysRevB.84.180301.

[6] O. Hellman, P. Steneteg, I. A. Abrikosov and S. I. Simak, *Temperature dependent effective potential method for accurate free energy calculations of solids*, Physical Review B **87**(10), 104111 (2013), doi:10.1103/PhysRevB.87.104111.

[7] O. Hellman and I. A. Abrikosov, *Temperature-dependent effective third-order interatomic force constants from first principles*, Physical Review B **88**(14), 144301 (2013), doi:10.1103/PhysRevB.88.144301.

[8] T. Tadano and S. Tsuneyuki, *Self-consistent phonon calculations of lattice dynamical properties in cubic $SrTiO_3$ with first-principles anharmonic force constants*, Physical Review B **92**(5), 054301 (2015), doi:10.1103/PhysRevB.92.054301.

[9] K. Parlinski, *Ab initio determination of anharmonic phonon peaks* **98**(5), 054305 (2018), doi:10.1103/PhysRevB.98.054305.

[10] Y. Oba, T. Tadano, R. Akashi and S. Tsuneyuki, *First-principles study of phonon anharmonicity and negative thermal expansion in $ScF_3$*, Physical Review Materials **3**(3), 033601 (2019), doi:10.1103/PhysRevMaterials.3.033601.

[11] T. Tadano, Y. Gohda and S. Tsuneyuki, *Anharmonic force constants extracted from first-principles molecular dynamics: Applications to heat transfer simulations*, Journal of Physics: Condensed Matter **26**(22), 225402 (2014), doi:10.1088/0953-8984/26/22/225402.

[12] N. Shulumba, O. Hellman and A. J. Minnich, *Intrinsic localized mode and low thermal conductivity of PbSe*, Physical Review B **95**(1), 014302 (2017), doi:10.1103/PhysRevB.95.014302.

[13] K. Parlinski, Z. Q. Li and Y. Kawazoe, *First-Principles Determination of the Soft Mode in Cubic $ZrO_2$*, Physical Review Letters **78**(21), 4063 (1997), doi:10.1103/PhysRevLett.78.4063.

[14] G. Pólya, *Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem*, Mathematische Zeitschrift **8**(3), 171 (1920), doi:10.1007/BF01206525.

[15] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, Wiley, New York (2003).

[16] S. Plimpton and B. Hendrickson, *A new parallel method for molecular dynamics simulation of macromolecular systems*, Journal of Computational Chemistry **17**(3), 326 (1996), doi:10.1002/(SICI)1096-987X(199602)17:3<326::AID-JCC7>3.0.CO;2-X.

[17] P. Erhart and K. Albe, *Analytical potential for atomistic simulations of silicon, carbon, and silicon carbide*, Physical Review B **71**(3), 035211 (2005), doi:10.1103/PhysRevB.71.035211.

[18] E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller and C. A. Becker, *The potential of atomistic simulations and the knowledgebase of interatomic models*, JOM **63**(7), 17 (2011), doi:10.1007/s11837-011-0102-6.

[19] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings *et al.*, *The atomic simulation environment—a Python library for working with atoms*, Journal of Physics: Condensed Matter **29**(27), 273002 (2017), doi:10.1088/1361-648X/aa680e.

[20] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika **57**(1), 97 (1970), doi:10.1093/biomet/57.1.97.