Bayesian learning of thermodynamic integration and numerical convergence for accurate phase diagrams

V. Ladygin[1,2,3,*] I. Beniya[2] E. Makarov[2] and A. Shapeev[2,†]

[1]*Moscow Institute of Physics and Technology, Institutskiy Pereulok 9, Dolgoprudny, Moscow Region 141700, Russia*
[2]*Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Building 3, Moscow 143026, Russia*
[3]*California Institute of Technology, 1200 E. California Blvd, Pasadena, California 91125, USA*

Accurate phase diagram calculation from molecular dynamics requires systematic treatment and convergence of statistical averages. In this work we propose a Gaussian process regression based framework for reconstructing the free-energy functions using data of various origins. Our framework allows for propagating statistical uncertainty from finite molecular dynamics trajectories to the phase diagram and automatically performing convergence with respect to simulation parameters. Furthermore, our approach provides a way for automatic optimal sampling in the simulation parameter space based on a Bayesian optimization approach. We validate our methodology by constructing phase diagrams of two model systems, the Lennard-Jones and soft-core potential, and compare the results with the existing studies and our coexistence simulations. Finally, we construct the phase diagram of lithium at temperatures above 300 K and pressures below 30 GPa from a machine-learning potential trained on *ab initio* data. Our approach performs well when compared to coexistence simulations and experimental results.

I. INTRODUCTION

Computational materials science is a rapidly evolving field enabling the calculation of materials properties that have traditionally been accessible mostly by experiment. The phase diagram is one such aggregate property, answering the following question: Which phase of material will be stable under given conditions (temperature, pressure, composition)? Phase diagrams are thus indispensable as guidance for materials synthesis. Physically, a phase diagram can be thought of as a map of free energy of different phases of material; if we know the free-energy function of different phases, then we can tell which phase, or a mixture of phases, will be stable under a given condition.

CALPHAD (standing for CALCulation of PHAse Diagrams) [1,2] is by far the most prominent approach to constructing phase diagrams in practice. The core of CALPHAD is a classical fitting approach, using polynomial-like functions, to represent the Gibbs free-energy functions of different phases. The free energy is fitted mostly to the experimental data, and in this sense, the obtained phase diagram is an experimentally obtained materials property (at least when contrasted against modern *ab initio*–based materials modeling). With the rise of *ab initio* materials modeling, the experimental data can be supplemented by quantum-mechanical data when the former is not available (e.g., when a phase cannot be experimentally realized at given conditions).

There are efforts to further advance the CALPHAD methodology by applying modern data analysis algorithms. Early attempts are dedicated to the use of the Bayesian framework in combination with the classical CALPHAD approach [3–5]. In these works, the authors focus on a methodology for uncertainty prediction of model parameters of CALPHAD. More recent studies are aimed at overcoming the problem of database extension [6] with the use of first-principles data calculation. The further development of the approach [6] includes uncertainty estimation of predicted results based on uncertainty in model parameters via Monte Carlo Markov chains combined with the Bayesian inference [7,8]. Another problem related to a mixture of experimental and calculation results was approached recently in [9] with k-fold cross-validation of the input datasets.

From the side of computational materials science, state-of-the-art algorithms for the free-energy computation are grouped together under the umbrella of thermodynamic integration. The central idea of these algorithms is that one can easily compute derivatives of free energy from molecular dynamics (MD) simulations (or similar simulations for saving the Gibbs distribution), and hence the free energy can be obtained by integrating these data from a point where the free energy can be calculated exactly (e.g., at zero temperature) [10]. Thermodynamic integration, in its essence, is a way to obtain computational thermodynamic data and hence is not an algorithm that is competing with CALPHAD but rather complementing it.

Indeed, there are many works that use both ideas: Obtaining computational thermodynamic data and fitting it with simple functions of thermodynamic parameters for a number of systems. Examples of such works are studies of the

*vladimir.ladygin@phystech.edu
†a.shapeev@skoltech.ru

Lennard-Jones system [11], where the authors examine the dependence of melting temperature with respect to the size of the simulation cell. In [12], the reconstruction of the aluminum phase diagram via density functional theory (DFT) data is presented. In [13] authors reconstruct the P-T phase diagram of uranium with a combination of the USPEX algorithm [14] and standard thermodynamic integration technique for the free-energy calculation. In the work [13] authors focus on phase stability calculation under finite temperature conditions in specific points. In [15] a detailed phase diagram of water is explored.

In this work we propose a Gaussian process-based methodology to reconstruct the phase diagram based on thermodynamic data. Although we focus on nonexperimental data (computed from first principles or with empirical potentials) in the work, the methodology itself should be applicable to the experimental data as well. The essence of our methodology can simply be described as treating thermodynamic data coming from any source (zero-temperature limit or MD averages or coexistence simulations) as training data for a Gaussian process. Including numerical parameters (such as the number of atoms or the cutoff radius of the potentials) into the feature vector enables an automatic analysis of convergence and, moreover, taking the limit as numerical parameters go to infinity. Furthermore, the predictive variance of the Gaussian process naturally allows us to estimate the error of our prediction, including the statistical error originating from the finite MD trajectories, interpolation error arising from the finite number of conditions (temperature/pressure) at which simulations are run, and the error of extrapolation with respect to numerical parameters. Finally, Gaussian processes allow for an automatic assessment of the ratio by which the error in the quantity of interest (such as the melting temperature) can be reduced by running a simulation at given parameters, leading to autonomous algorithms of sampling the phase diagram points.

The presented methodology is applied to two model systems: Soft-core and Lennard-Jones potentials as a part of the validation procedure. We compare our calculations with the work [16] on the soft-core potential in which the authors estimated the melting curves in a wide temperature-pressure range. The Lennard-Jones system's properties, including the dependence on the simulation parameters, were investigated in various works. The points of interest on the Lennard-Jones phase diagram are the critical point where gas and liquid become indistinguishable and the triple point where all the three phases—gas, liquid, and solid—coexist. The location of the critical point strongly depends on the size of the simulation cell. In literature, there are two main approaches to study this dependence. The first one is associated with the projections of a simulation cell with periodic boundary conditions into the surface of a four-dimensional sphere [17]. Another approach relies on the use of cutoff radius bounded to the half-size of the simulation cell and corresponding long-range correction [18]. The critical point temperature estimations are not in agreement with each other with the given narrow confidence intervals. This fact shows that the approach for systematic phase transition calculation and uncertainty estimation is in demand in computational materials science. For validation of our results with respect to triple and melting points calculation

of the Lennard-Jones system, we have chosen the work of Ref. [11].

Finally, we apply our approach to a physical system, lithium, chosen because it undergoes various phase transitions under pressure. The CALPHAD approach for this material is based on experimental data from Refs. [19–21]. These works are attributed to fcc-liquid-bcc phase transitions. In [22] the phase diagram of lithium is examined in a wide temperature and pressure range from first-principles complementing experiment. Recently, in [23] authors compute the former transition lines with the existing classical potentials.

The paper is organized as follows. In Sec. II we present the theoretical aspect of the methodology and some implementation details. In particular, in Sec. II A we give the details of free-energy calculations, and in Sec. II B we introduce our Gaussian process regression approach. In Sec. III we show the results of the application of our approach to the phase diagram calculation of the model systems and lithium. Concluding remarks are given in Sec. IV.

II. THEORY

In the current section we will obtain the relations between free-energy derivatives and statistical averages as will be used by Gaussian processes. In what follows we distinguish the extensive and intensive quantities in the notation; the former will have a hat accent: $\hat{E}$, $\hat{V}$. The corresponding intensive, per-atom quantities are $E = \hat{E}/N$, $V = \hat{V}/N$, where N will denote the number of atoms.

A. Free energy

Let $\boldsymbol{x}$ be a configuration with N atoms enclosed in a volume $\hat{V}$ (we will interchangeably use $\hat{V}$ for the actual region in space and its measure) with periodic boundary conditions. Let $\hat{E}(\boldsymbol{x})$ be the potential energy of the interatomic interaction. We assume that the units for the temperature T are the same as for the energy $\hat{E}$; in other words, our Boltzmann constant is $k_B = 1$.

We define the free energy by

$$-T \log \int_{\hat{V}^N} \exp[-\hat{E}(\boldsymbol{x})/T] \mathrm{d}\boldsymbol{x} = \hat{F}_{\mathrm{ref}} - T\hat{S}, \qquad (1)$$

where $\hat{F}_{\mathrm{ref}}$ will be explicitly assigned later (differently for each phase), and we will call $\hat{S}$ the entropy. We will rely on comparing the absolute free energies of different phases (as opposed to the free-energy difference between phases); therefore it is important to choose $\hat{F}_{\mathrm{ref}}$ and $\hat{S}$ consistently across phases.

For solids we choose

$$F_{\mathrm{ref}}^{(\mathrm{s})} := E_0 + T\left(-\log(NV) + 1 - \tfrac{3}{2}\log(2\pi T)\right), \qquad (2)$$

where $E_0 = E_0(V)$ is the potential energy at zero temperature for the given volume. Here and in what follows the superscript (s) denotes the solid phase. Choosing $F_{\mathrm{ref}}^{(\mathrm{s})}$ in the form (2) is motivated by the fact that in this case the entropy admits a simple low-temperature expansion

$$S^{(\mathrm{s})} = -\frac{1}{2N} \log \det \hat{H} - \log(V) + O(T), \qquad (3)$$

where $\hat{H}$ is the Hessian of the energy $\hat{E}$ at the equilibrium configuration $\boldsymbol{x}_0$. For the derivation of (2) and (3), refer to Appendix A 1. A well-defined zero-temperature limit of $S^{(s)}$ such as (3) is important for reconstructing the free energy with a Gaussian process, as molecular dynamics can generate the data on derivatives of S and thus allow, without (3), one to reconstruct $S^{(s)} = S^{(s)}(T, V)$ only up to an additive constant.

Indeed, an *NVT*-thermostated molecular dynamics produces the averages of the form

$$\langle f \rangle := \frac{\int_{\hat{V}^N} f(\boldsymbol{x}) \exp[-\hat{E}(\boldsymbol{x})/T] d\boldsymbol{x}}{\int_{\hat{V}^N} \exp[-\hat{E}(\boldsymbol{x})/T] d\boldsymbol{x}}.$$

One can then find that

$$\frac{\partial S^{(s)}}{\partial T} = T^{-2} \langle E - E_0 \rangle - \frac{3}{2} T^{-1}, \quad \text{and} \quad (4)$$

$$\frac{\partial S^{(s)}}{\partial V} = T^{-1} \langle P - P_0 \rangle, \quad (5)$$

where $P_0 := -\frac{\partial E_0}{\partial V}$ is the pressure at zero temperature. Here and in what follows by P we denote the virial part of the pressure. The virial pressure is, in fact, easier to compute from molecular dynamics. The derivation of (4) and (5) is given in Appendix A 2.

Also, we consider the liquid and gas phases. A single free-energy curve can describe these phases because they are indistinguishable at temperatures above the critical one. Hence, we will denote the corresponding phase by the superscript (f) and refer to it as the fluid phase. For the fluid we simply choose ideal gas as a reference,

$$F_{\text{ref}}^{(f)} := -T \log(NV), \quad (6)$$

so that

$$\lim_{T \to \infty} S^{(f)} = \lim_{V \to \infty} S^{(f)} = 0. \quad (7)$$

This equality is a consequence of our definition of free energy in (1). In the limit of $T \to \infty$, the exponent in the integral is approaching 1. Hence, the integral itself is equal to NV. In the same manner, when $V \to \infty$, the interaction between particles is negligible [$\hat{E}(\boldsymbol{x}) \to 0$] and the integral also approaches NV.

The derivatives of $S^{(f)}$ are thus

$$\frac{\partial S^{(f)}}{\partial V} = T^{-1} \langle P \rangle, \quad \text{and} \quad (8)$$

$$\frac{\partial S^{(f)}}{\partial T} = T^{-2} \langle E \rangle. \quad (9)$$

The derivation of (8) and (9) is very similar to the corresponding formulas for the solid, and hence we omit such a derivation.

The harmonic (2) and ideal gas (6) limits are not always applicable—for instance, there are systems with solid phases being dynamically unstable at low temperature. For such systems, the limit (2) is irrelevant. In such cases we determine the additive constant of the free energy through fitting to the melting (or, more generally, coexistence) point of a phase. We find melting point at pressure P by solving the system of

equations

$$\begin{cases} \dfrac{\partial S^{(f)}}{\partial V^{(f)}} = \dfrac{P}{T} + \dfrac{1}{T} \dfrac{\partial F_{\text{ref}}^{(f)}}{\partial V^{(f)}} \\[2mm] \dfrac{\partial S^{(s)}}{\partial V^{(s)}} = \dfrac{P}{T} + \dfrac{1}{T} \dfrac{\partial F_{\text{ref}}^{(s)}}{\partial V^{(s)}} \\[2mm] S^{(f)} - S^{(s)} = \dfrac{F_{\text{ref}}^{(f)} - F_{\text{ref}}^{(s)}}{T} + \dfrac{P(V^{(f)} - V^{(s)})}{T} \end{cases} \quad (10)$$

with respect to the temperature T and specific volumes of solid and fluid, $V^{(f)}$ and $V^{(s)}$.

B. Gaussian process regression

The derivatives of the entropy from an *NVT* molecular dynamics (MD) cannot be obtained without some noise arising from averaging over a finite trajectory. Due to randomness in the initial state or in the thermostat, such a trajectory is random. Thus, the free energy that we reconstruct from the MD data is also random, but hopefully, it has a narrow distribution around the true free energy. The effect of a thermostat—let us consider a Langevin thermostat, for instance—consists of making a large number of small perturbations to the trajectory [24]. Thanks to the central limit theorem, it is hence reasonable to assume that averages over such a trajectory are distributed according to the Gaussian distribution. This assumption brings us to the Gaussian process framework.

In the Gaussian process framework, we assume that the data, and the reconstructed free energy, are distributed according to a multivariate Gaussian distribution. We assume zero mean—any prior information about a nonzero mean is already accounted for in F^{ref}. Further, we assume that the values of the free energy at different points (V_1, T_1) and (V_2, T_2) are correlated with covariance $\text{Cov}[S(V_1, T_1), S(V_2, T_2)] = k[(V_1, T_1), (V_2, T_2)]$. Such a distribution of functions $S(V, T)$ is called the *Gaussian process* (GP) and k is called *the kernel*. An simple example of the kernel is

$$k[(V_1, T_1), (V_1, T_1)] \sim \exp\left(-\frac{(T_1 - T_2)^2}{2\theta_T^2}\right)$$
$$\times \exp\left(-\frac{(V_1 - V_2)^2}{2\theta_V^2}\right). \quad (11)$$

A property of the Gaussian processes that will be very helpful in our application is that any linear functional of the Gaussian process is also Gaussian distributed. For example, the derivative with respect to volume [as, e.g., in (8)] at (V_1, T_1) is correlated with $S(V_2, T_2)$ as given by the following formula:

$$\text{Cov}\left(\frac{\partial S}{\partial V_1}(V_1, T_1), S(V_2, T_2)\right) = \frac{\partial}{\partial V_1} k[(V_1, T_1), (V_2, T_2)].$$

This allows us to make predictions based on data. In the most general case, each data point is a linear functional X on S, for example, $\langle S|X_1 \rangle = S(V, T)$, $\langle S|X_2 \rangle = \frac{\partial}{\partial V} S(V, T)$, etc. The data is usually given with noise, hence the input data to our Gaussian process is of the form $(X_1, Y_1, \Delta Y_1), (X_2, Y_2, \Delta Y_2), \ldots$, which means that $\langle S|X_i \rangle$ is measured (e.g., from molecular dynamics) as $Y_i \pm \Delta Y_i$, or to be precise, $\langle S|X_i \rangle$ is distributed according to the normal
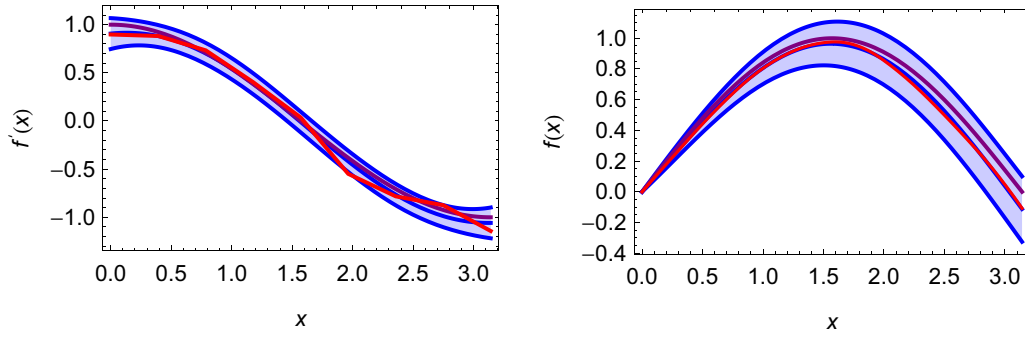
FIG. 1. Illustration of integration via the trapezoidal quadrature rule (red) and Gaussian process (blue) on noisy data. The left graph shows how the function $f(x) = \sin(x)$ is reconstructed, and the right one shows how the derivative is reconstructed. The Gaussian process has the same accuracy but produces a smooth result and also yields uncertainties of prediction within which the exact solution (purple) falls.

distribution

$$\langle S|X_i \rangle \sim \mathcal{N}(Y_i, (\Delta Y_i)^2).$$

We denote $\mathrm{Cov}(\langle S|X_1 \rangle, \langle S|X_2 \rangle) = k(X_1, X_2)$, extending the definition for the kernel. We assume that the uncertainties ΔY_i are all statistically independent from each other.

Suppose we want to make a prediction of $Y_* = \langle S|X_* \rangle$; for simplicity one can think of $Y_* = F(V_*, T_*)$. The Gaussian process framework is a particular case of the Bayesian framework in which the prediction problem is formulated as the following question: What is the most likely value of Y_* given data X_i, Y_i, and ΔY_i? To that end we form a joint distribution

$$\begin{bmatrix} Y \\ Y_* \end{bmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} K(X, X) + \mathrm{diag}(\Delta Y) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*). \end{pmatrix} \right),$$

where X, Y, and ΔY are the vectors composed of X_i, Y_i, and ΔY_i, which makes $K(X, X)$ a matrix composed of $k(X_i, X_j)$. From this, we find that Y_* is normally distributed with mean

$$\overline{Y_*} = K(X_*, X)[K(X, X) + \mathrm{diag}(\Delta Y)]^{-1} Y$$

and variance

$$\mathrm{Var}(Y_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) \\ + \mathrm{diag}(\Delta Y)]^{-1} K(X, X_*). \quad (12)$$

The variance of a nonlinear functional $\mathcal{F}(S)$, which is needed to predict the uncertainty of determining the melting point (10), is derived in Appendix A 4 in the limit of a large amount of data [in which can we can linearize $\mathcal{F}(S)$ around the mean prediction $\overline{S}$].

For a given application, one needs to find the right values of hyperparameters $\boldsymbol{\theta} = (\theta_T, \theta_V)$ in (11). This is done by maximizing the so-called marginal likelihood $p(Y|X, \boldsymbol{\theta})$, which is proportional to the probability that the underlying data is distributed according to the hyperparameters $\boldsymbol{\theta}$. The marginal likelihood is calculated according to the formula

$$\log p(Y|X, \boldsymbol{\theta}) = -\frac{1}{2} Y^T [K(X, X) + \mathrm{diag}(\Delta Y)]^{-1} Y \\ - \frac{1}{2} \log |K(X, X) + \mathrm{diag}(\Delta Y)| - \frac{n}{2} \log(2\pi), \quad (13)$$

where $|K|$ denotes the determinant of the matrix K and n is the dimension of the model (number of input points).

The Bayesian variance expressed in Eqs. (12) and (A3) does not directly depend on Y. This allows us to define the quantity

$$\mathcal{H}(Q, X^*) = -\log \frac{\mathbb{V}(Q|X \cup X^*)}{\mathbb{V}(Q|X)}, \quad (14)$$

which expresses an expected improvement of the variance of Q after adding a new point X^* to the dataset X. We will call (14) the information function. By maximizing $\mathcal{H}$ with respect to X^*, we find the point that is best in reducing the variance of the quantity Q. This gives rise to our active sampling algorithm, whose essence is to simply greedily add points with maximum information to the training set, one by one. For simplicity, when deciding which new point to add to the training set, we assume that we would add data with zero variance.

C. Illustrative example: Performing integration with a Gaussian process

Before applying Gaussian processes to reconstruct the free-energy function from its derivatives, let us consider an illustrative problem of simply reconstructing $f(x) = \sin(x)$ on the interval $[0, \pi]$ from its value $f(x) = 0$ and nine noisy values of $f'(x)$. To that end, we take nine evenly spaced points, $x_1 = 0$, $x_2 = \pi/8$, ..., $x_9 = \pi$, and consider $y_i = f(x_i) + \mathcal{N}(0, \sigma^2)$ (i.e., added to y_i some normally distributed noise) with variance $\sigma = 0.1$. The added noise simulates the statistical noise always present in averages taken over finite MD trajectories.

The results of the comparison are shown in Fig. 1. To mimic thermodynamic integration, the state-of-the-art method used for calculating the free energy, we use the second-order trapezoidal quadrature rule, shown with red in the figure. The "integration" with the Gaussian process was done as outlined in Sec. II B. We see that the accuracy of the two methods is comparable, but the Gaussian process yields smoother results and also gives an accurate confidence interval.
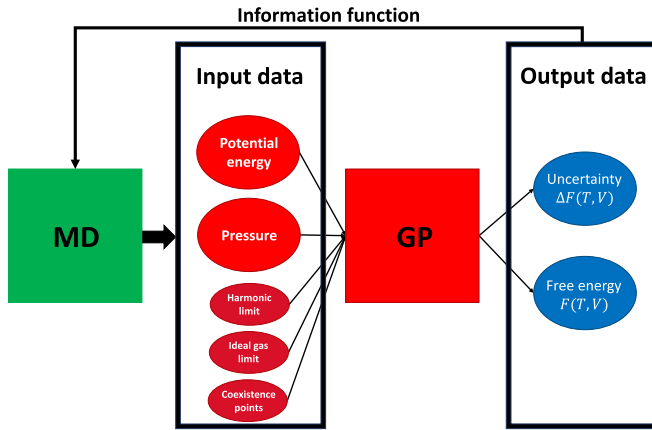
FIG. 2. General scheme of the Bayesian framework for calculating phase diagrams.

III. RESULTS AND DISCUSSION

A. Methods

The formulas from Sec. II A apply to any data regardless of their source. In this work, we use classical molecular dynamics implemented in the LAMMPS package [25] to generate the input data for the Gaussian process. Molecular dynamics simulations are performed in the canonical ensemble (NVT).

The core of classical MD simulations is the interatomic potential. An interatomic potential is a functional form that allows one to calculate the potential energy of the system. In our work we study the behavior of the systems described by the Lennard-Jones potential, the soft-core potential (being simply the repulsive term of the Lennard-Jones potential), and

the moment tensor potential [26,27]. For each of these functional forms, the total potential energy $E(\boldsymbol{x})$ can be partitioned into the sum of atomic contributions

$$E(\boldsymbol{x}) = \sum_i E(\boldsymbol{x}_i),$$

where $\boldsymbol{x}_i$ is the coordinate of the atom with the index i. The potential energy of a particle, $E(\boldsymbol{x}_i)$, is a sum of atomic contributions within finite sphere of radius r_{cut} called cutoff radius

$$E(\boldsymbol{x}_i) = \sum_{j:|\boldsymbol{x}_i-\boldsymbol{x}_j|<r_{\mathrm{cut}}} \varphi(|\boldsymbol{x}_i - \boldsymbol{x}_j|),$$

where φ is the function that describes potential energy of two-particle interaction with respect to their positions $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The potential energy of the soft-core and Lennard-Jones systems can be explicitly decomposed into a sum of pair contributions. Here and in what follows, the functional form of the potential energy of single pair interaction will be denoted by $\varphi(r)$.

The temperature T, specific volume V, cutoff radius of interatomic potential r_{cut}, and a number of atoms in the simulation cell N form the entries of the matrix $\boldsymbol{X}$ of the Gaussian process. The entries of the vector $\boldsymbol{Y}$ are formed from derivatives and values of the free energy. The derivatives are calculated via formulas (4), (5), (8), and (9) using the averaged potential energy $\langle E \rangle$ and the virial pressure $\langle P \rangle$ obtained from MD. An additive constant of the free energy is set via harmonic (2) and ideal gas (6) limits. Where these limits are not applicable, we determine the additive constant difference via (10). The standard deviation of the corresponding thermodynamic quantities over the MD trajectories forms entries of the vector $\boldsymbol{\Delta Y}$. An active sampling algorithm defined in Sec. II B is applied to reduce the uncertainty of the target quantities sys-
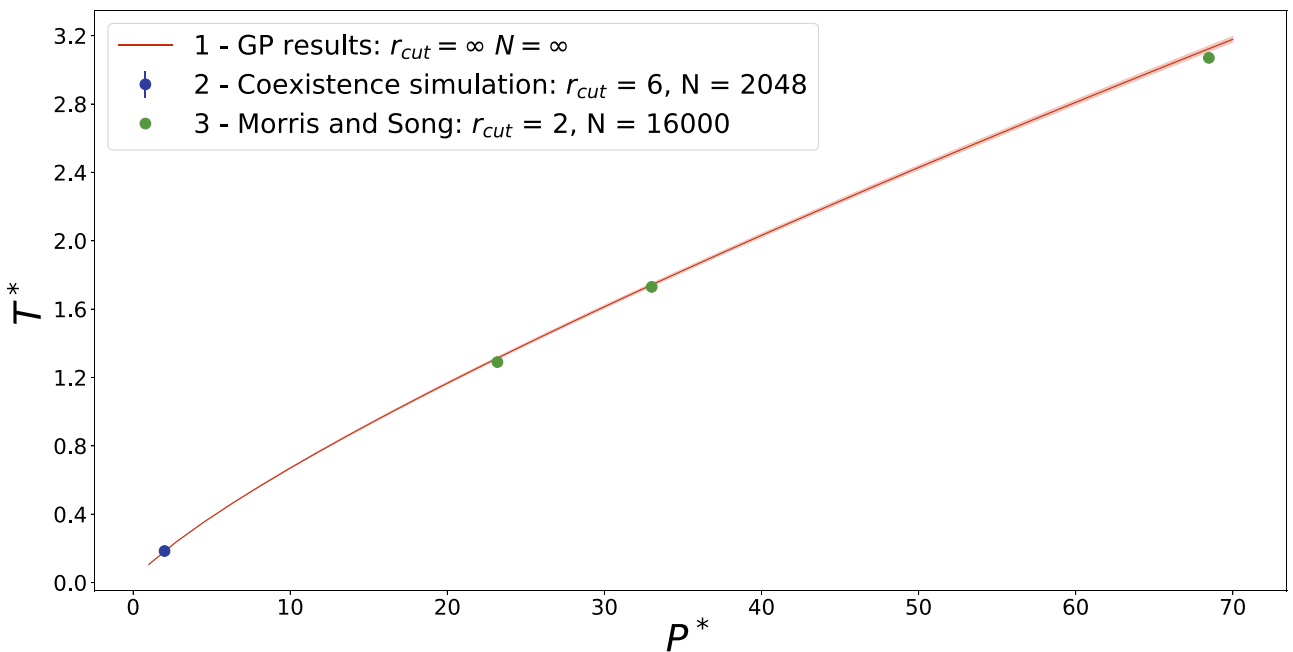

FIG. 3. Melting line of the soft-core potential plotted with the 68% confidence interval: The red curve represents the GP prediction of the melting curve at the limit of infinite cutoff and an infinite number of atoms; the blue point is our coexistence simulation result at the cutoff of 6 and number of atoms of 2048 for each phase. Green points are the results from [16], with the cutoff of 2 and the number of atoms of 16 000. Our data is seen to be in an excellent agreement with coexistence simulation results and previous studies.

tematically. The general scheme of our framework is shown in Fig. 2.

Analyzing and accelerating convergence with a Gaussian process

Gaussian processes will allow us to automate one more job that is traditionally done manually: The analysis and acceleration of the convergence with respect to numerical parameters, such as the number of atoms or the cutoff radius. The fundamental property of the Gaussian processes enabling this is what we have already used for the derivatives; taking the limit of a Gaussian process (e.g., with $N \to \infty$), being a linear operation, still yields a Gaussian process. Moreover, we choose the kernel functional form so that it reflects as much as possible the physical behavior of the system of interest, as we illustrate in the applications below. As a result, the trained Gaussian processes reproduce free energy with high accuracy.

B. Model system: Soft-core potential

We first validate our methodology on two model systems: Soft-core and Lennard-Jones potentials. In the soft-core potential, there is only one phase transition. This simplifies the comparison between our GP estimation and the existing data.

MD calculations in the case of soft-core and Lennard-Jones potentials are performed with a Metropolized Langevin thermostat [28] to avoid error related to a finite time step. The data presented below is given in the reduced Lennard-Jones units.

We consider the soft-core interatomic potential with pair interaction described by

$$\varphi(r) = \frac{4}{r^{12}},$$

where r is the distance between two atoms. Thanks to the simplicity of the soft-core potential, the free energy admits the following invariant transformation:

$$F(T, V) = -T \ln V + F(TV^4, 1).$$

Hence, the free-energy dependence for such a potential can be expressed as $F(T, V) \sim F(T, 1)$. This means that given a dependence of the free energy on the temperature at a certain volume, we know the dependence of the free energy in arbitrary volume and temperature range.

With this invariance in mind and also asymptotic behavior of the free energy for $V \to 0$, we define the GP kernel—a functional form that we use to estimate the correlation between the training set points X_1 and X_2:

$$k_{\text{scp}}(X_1, X_2) \sim \exp\left(-\frac{\left(\left(1 - \frac{T_1}{1+T_1}\right)^{\frac{1}{4}} - \left(1 - \frac{T_2}{1+T_2}\right)^{\frac{1}{4}} \right)^2}{2\theta_T^2} - \frac{\left(\frac{T_1}{1+T_1} - \frac{T_2}{1+T_2} \right)^2}{2\theta_T'^2} \right)$$

$$\times \left[1 + \frac{\theta_c^{10}}{c_1^5 c_2^5} \exp\left(-\left(\frac{1}{c_1^2} - \frac{1}{c_2^2} \right)^2 \frac{\theta_c^4}{2} \right) \right] \exp\left[-\left(\frac{1}{N_1} - \frac{1}{N_2} \right)^2 \frac{\theta_N^2}{2} \right],$$

where $\boldsymbol{\theta} = (\theta_T, \theta_T', \theta_c, \theta_N)$ are the hyperparameters of our model; T_1, T_2 are the temperatures of the first and second point; and c_1, c_2 and N_1, N_2 are the interatomic potential cutoff radius and the number of atoms in the simulation cell, correspondingly. The choice of the kernel reflects our knowledge of the system in the following way. We operate with the rescaled temperature $\frac{T}{1+T}$ instead of T to avoid divergence at an infinite temperature limit in the case of liquid. We modify the temperature dependence with a $\frac{1}{4}$ power term to account for the two-particle interaction in the liquid free energy at high temperature. The cutoff term depends on r_{cut} as c^{-5} (in fact, the dependency in the case of the soft-core potential is stronger, but we choose the form with c^{-5} to match the asymptotics for the Lennard-Jones system). Dependencies of lower order are taken into account via long-term correction. Finally, the free-energy dependence with respect to the number of atoms in the simulation cell is derived from Taylor series expansion and is proportional to $\frac{1}{N}$.

We compare our results of melting point calculation with coexistence simulation data at $P = 1$. The melting temperature predicted via GP is equal to 0.1849(8) [error in parenthesis corresponds to the 68% confidence interval, 0.1849(8) should be read as 0.1849 ± 0.0008]. The value obtained by our coexistence simulations is 0.1846(2). The obtained data are in agreement with respect to a given confidence interval. We note, however, that the GP provides a more reliable confidence interval being a natural statistical

estimation of uncertainty, unlike the coexistence simulations in which it is not easy to control all sources of error.

In the case of melting curve estimation, we compare our results with a direct calculation based on coexistence simulation and previous results from [16]. The obtained data is shown in Fig. 3. In this case, the GP estimation of the melting curve agrees well with our simulations. Our data is in good agreement with the previous results. The melting point at $P \approx 70$ can be seen to deviate by about 3σ (i.e., three standard deviations), which is statistically significant, but we attribute this to the low value of cutoff ($r_{\text{cut}} = 2$ for the green points taken from [16]).

C. Model system: Lennard-Jones potential

The Lennard-Jones potential was selected as a system with three phases and hence a more complex phase diagram. Also, the $\frac{1}{r^6}$ term of the Lennard-Jones potential results in the strong dependence of obtained data on the size of the simulation cell and cutoff radius of the interatomic potential. As in the soft-core case, the MD calculations are performed with a Metropolized Langevin thermostat [28].

The potential energy of the pair interaction of the Lennard-Jones system has a form

$$\varphi(r) = 4\left(\frac{1}{r^{12}} - \frac{1}{r^6} \right).$$

TABLE I. Critical point (index crit) estimation. r_{cut}^* denotes the cutoff radius of the interatomic potential with a long-range correction. Our results are in good agreement with previous studies.

Source	r_{cut}^*	System size	T_{crit}	ρ_{crit}
Ref. [17]	∞	∞	1.326(1)	0.316(1)
This work	∞	∞	1.327(1)	0.316(2)

Compared to the soft-core potential the Lennard-Jones potential has an additional term $\frac{1}{r^6}$. This leads to a phase diagram with three different phases: Solid, gas, and liquid. The last two can be described by a single free-energy curve because they are indistinguishable at temperatures above critical. The kernel for the Lennard-Jones free energy is similar to the kernel of the soft-core potential except that we treat the volume dependence of the free energy explicitly in the kernel,

$$k_{LJ}(X_1, X_2) \sim k_{scp}(X_1, X_2) \left(\exp\left(-\frac{(\rho_1 - \rho_2)^2}{2\theta_\rho^2} \right) \right.$$
$$\left. - \exp\left(-\frac{(\rho_1^2 + \rho_2^2)}{2\theta_\rho^2} \right) \right),$$

where ρ_i, simply defined as $\rho_i := V_i^{-1}$, are densities at first and second points. We subtract $\exp(-\frac{(\rho_1^2 + \rho_2^2)}{2\theta_\rho^2})$ from the Gaussian kernel to account for the zero-density (or infinite-volume) limit of the free energy (7).

We first validate our algorithm by computing the critical and triple points of the Lennard-Jones potential. The critical point is the point on the phase diagram where the liquid and gas phases become indistinguishable. Our results, along with the previous studies, are shown in Table I.

There are a lot of studies for comparison, but we have chosen [17] as the one with the most reliable values. In [17], the authors calculated the critical point using MC simulation and estimated effect of the system's finite size. Our results agree well with respect to confidence interval. Moreover, the estimated confidence interval is in agreement with the previous results.

We next estimate the triple point as a part of our validation procedure. An active sampling algorithm was used to improve the accuracy of the triple-point calculation significantly. An illustrative example of two steps of the active sampling algorithm is shown in Fig. 4.

Our results of the triple-point calculation, along with previous studies, are presented in Table II. Our results are in perfect agreement with the work of Ref. [11], where the authors performed the study of convergence with respect to the size of the system. From a comparison of the results with [29], one can deduce that the choice of r_{cut} strongly affects the triple-point calculation for the Lennard-Jones potential.

We next compare the melting line of the Lennard-Jones potential with our coexistence simulation results and the previous studies. The obtained data is shown in Fig. 5. Our results agree well with the previous studies. We have chosen [11] as reference data because the authors have examined the effect of the finite system size on the triple and melting points calculation. Compared to the study [11] we have treated cutoff and system size as explicit parameters of our model.

We next calculate the phase diagram of Lennard-Jones at the limit of infinite cutoff radius and system size. The phase diagram is presented in Fig. 6. As one can see, we have predicted the phase diagram and estimated the confidence interval of our calculation. Also, the phase diagram obtained via GP fitted with a limited number of points for both phases is in agreement with the previous studies.
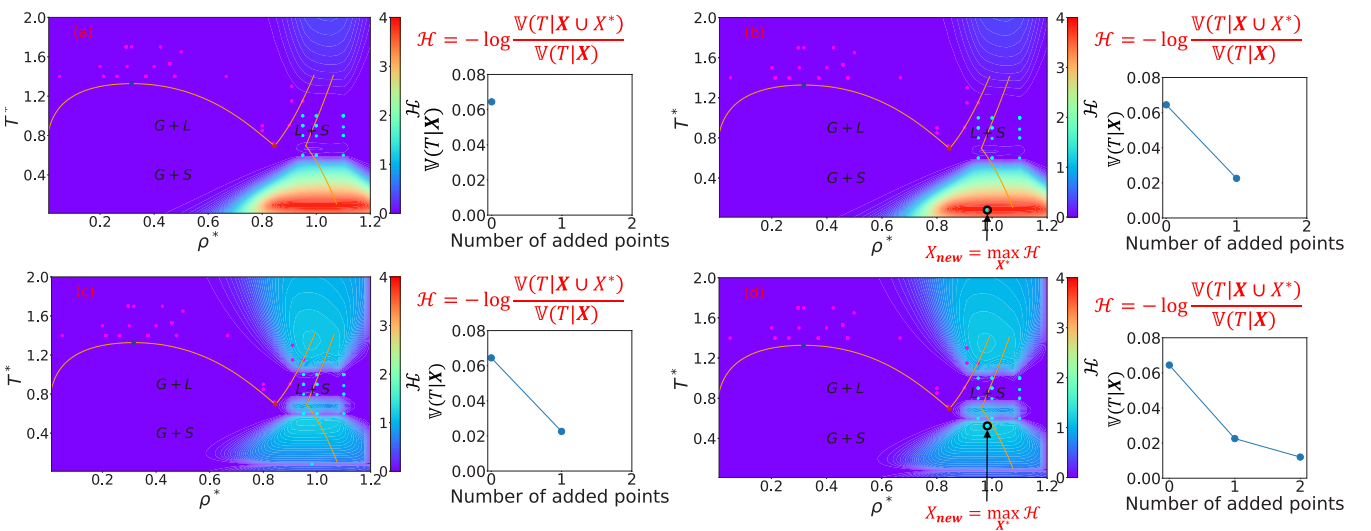

FIG. 4. Illustration of the active sampling algorithm applied to the calculation of the triple-point temperature estimation. Color plot represents the value of the information function $\mathcal{H}$ from (14) at a given point of the phase diagram. In step (a), the information function is calculated at each point. Then, in step (b), a new point X_{new} that maximizes the information function is added to the dataset. This finalizes the first iteration of the algorithm. Figures (c) and (d) correspond to the second iteration of the active sampling procedure. The proposed active sampling algorithm allows one to decrease the error of the target property systematically.

TABLE II. Triple-point (index tp) estimation. Indexes "gas," "liq," and "sol" denote gas, liquid, and solid phases, respectively. Our results agree well with existing studies.

Source	r_{cut}^*	System size	T_{tp}	ρ_{gas} $\times 10^{-3}$	ρ_{liq}	ρ_{sol}
Ladd and Woodcock [30]	2.5	1500	0.67(1)	–	0.818(4)	0.963(6)
Hansen [31]	–	864	0.68(2)	–	0.85(1)	–
Kofke [32]	–	236	0.698	–	0.854	0.963
Kofke [32]	–	932	0.687(4)	–	0.850	0.960
Ahmed and Sadus [29]	2.5	2048	0.661	–	0.864	0.978
Mastny and de Pablo [11]	6	∞	0.694(4)	–	–	–
This work	∞	∞	0.695(4)	1.9(1)	0.845(2)	0.961(1)

D. Physical system: Lithium

Finally, we apply our methodology to lithium, modeled with a machine-learning potential trained on DFT calculations. We compute the phase diagram of lithium at pressures below 30 GPa and temperatures above room temperature in this work. We hence examine the bcc, fcc, and liquid phases of lithium. In particular, the bcc phase is not stable at $T = 0$; therefore to accurately obtain the additive constant of the bcc-Li free energy, we rely on the harmonic limit for the fcc phase, fcc-liquid coexistence simulations to obtain the additive constant of the liquid phase, and bcc-liquid coexistence simulations to finally obtain the additive constant of the bcc phase.

Thermodynamic data is obtained with the use of a moment tensor potential (MTP) [26,27] as implemented in the MLIP software package [33]. We have trained a single MTP for all phases fitted on quantum-mechanical data in the examined temperature and pressure range. The MTP potential was actively trained on the fly by running MD simulations of fcc Li with 108 atoms and bcc Li with 128 atoms for a range of volumes covering the pressure range of interest and increasing temperature from normal conditions up to 900 K to observe melting. The default, level-16 MTP potential was used, with

the cutoff radius of 5 Å and the mindist value set to 1.4 Å. The DFT calculations were conducted with the plane-wave DFT, and a PAW pseudopotential with one electron treated as the valence electron, as implemented in the VASP package [34–36].

The kernel for the lithium fcc phase is chosen as

$$k_{Li}^{fcc}(x_1, x_2) \sim \exp\left(-\frac{(T_1 - T_2)^2}{2\theta_T^2}\right) \exp\left(-\frac{(\rho_1 - \rho_2)^2}{2\theta_\rho^2}\right)$$
$$\times \exp\left(-\left(\frac{1}{N_1} - \frac{1}{N_2}\right)^2 \theta_N^2\right).$$

For liquid and bcc phases we do not have reference at zero temperature. In order to account for divergence of entropy at this limit we modify temperature part of the kernel by adding the $\frac{1}{T}$ term:

$$k_{Li}^{bcc,liquid}(x_1, x_2)$$
$$\sim \left[\frac{\theta_T^2}{T_1 T_2} + \exp\left(-\frac{(T_1 - T_2)^2}{2\theta_T'^2}\right)\right]$$
$$\times \exp\left(-\frac{(\rho_1 - \rho_2)^2}{2\theta_\rho^2}\right) \exp\left[-\left(\frac{1}{N_1} - \frac{1}{N_2}\right)^2 \frac{\theta_N^2}{2}\right].$$
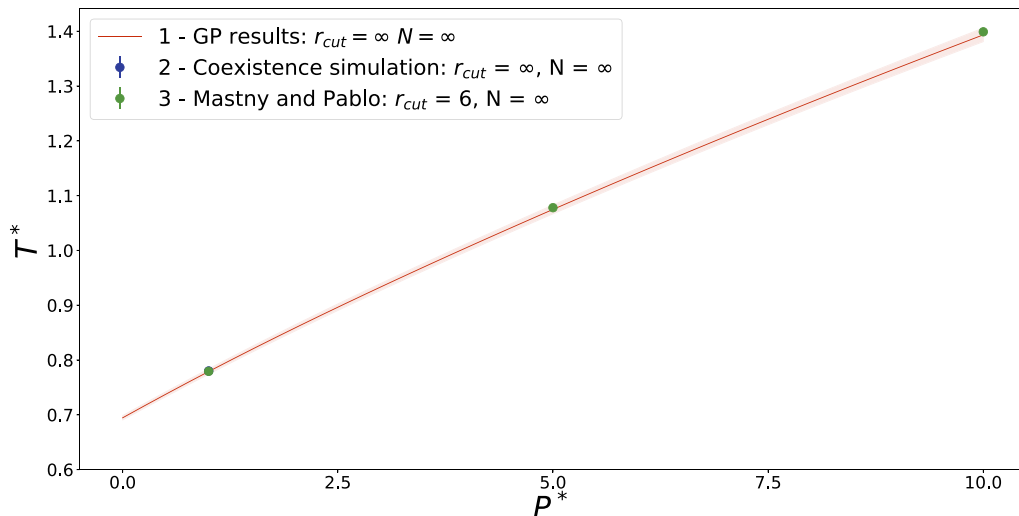

FIG. 5. Melting curve of the Lennard-Jones potential. The red curve corresponds to the calculation via GP. Rhombus is an extrapolation of coexistence simulation results to the limit of infinite r_{cut} and infinite system size. Green dots denote the existing results from [11]. Our results are in an excellent agreement with [11] and coexistence simulation data.
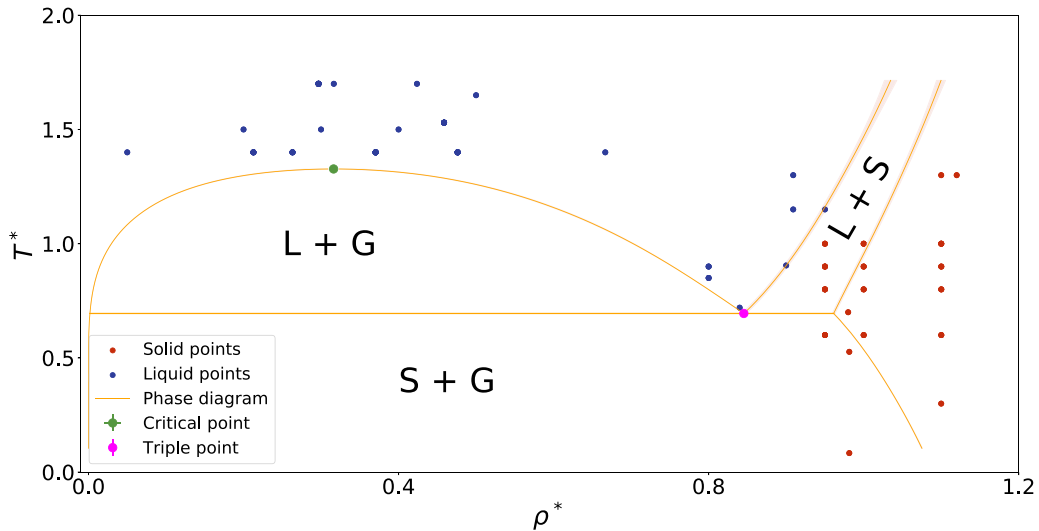
FIG. 6. The Lennard-Jones phase diagram at the limit of infinite r_{cut} and N with a confidence interval. Blue and red points are the training points for solid and liquid, respectively; the orange curve is a phase diagram estimated via GP. The green dot denotes the critical point. The pink dot represents the triple point. Only a few points of both solid and liquid phases are used to calculate all possible phase transitions of the Lennard-Jones system with high accuracy.

In the cases where we do not have reference data, we explicitly add melting points to the dataset. Using the MTP fitted on quantum-mechanical data, we cannot approach the limit of infinite temperature or volume (since the potential was not fitted at those conditions). Also, the bcc phase of lithium is dynamically unstable at zero temperature. Thus we use the bcc-liquid and fcc-liquid phase transition as reference data. For this reason, to validate our approach, we compare the prediction of GP with coexistence simulation results at various phase transition points. Results are presented in Table III. Predicted results lie well within the 95% confidence interval with respect to the coexistence simulation data. Finally, we compute the phase diagram of lithium in

a range of pressures below 30 GPa and a range of temperatures above 300 K extrapolated to the limit of infinite system size. To that end, we first train the MTP for these conditions by running short NVT-MD trajectories covering the chosen pressure and temperature conditions and actively learning the quantum-mechanical interaction on the fly [27,33,37]. In total, 290 108-atom and 128-atom (for fcc and bcc, respectively) crystalline and liquid configurations were selected and computed on DFT. We then switch off active learning and used the final potential with the developed methodology to construct the phase diagram of Li. The corresponding phase diagram, coexistence simulation results, and experimental data are shown in Fig. 7.

We observe that the GP estimates the phase diagram with high accuracy in a wide temperature and pressure range. Moreover, GP predicts the bcc-liquid phase transition accurately when compared to the experimental results [19,20]. The difference between our calculation and previous studies is less than 10 K. However, the fcc-liquid phase transition shows a greater divergence from the experimental data [21]. The difference could be due to inaccuracies of the DFT calculations as compared to the experimental data. The fcc-bcc phase transition below 500 K and above 300 K is not explored to the authors' knowledge. The available data around 300 K from [22] is in excellent agreement with the transition estimated via GP: Our bcc-fcc transition line goes directly through the region of pressures where both fcc and bcc phases were realized experimentally.

Also, we have calculated the pressure P_{tp} and temperature T_{tp} at the triple bcc-liquid-fcc point. P_{tp} is equal to 8.8 GPa; $T_{tp} = 501(2)$ K. The obtained result is in agreement with experimental estimations. As in the case of the Lennard-Jones potential, we have applied an active learning strategy to improve the accuracy of the triple points. An example of two steps of the algorithm applied to the lithium system is given in Fig. 8.

TABLE III. Comparison between the GP-predicted phase transition temperatures and the coexistence simulation results. The indices "coex" and "GP" denote coexistence simulation and estimations of the GP, respectively. "*" refers to the phase transition points used as free-energy references. The points not marked with "*" are the validation data, not used in the fitting of the GP. σ refers to the total confidence interval (combined coexistence and GP confidence intervals). ΔT is the absolute difference (error) between melting temperature predicted by coexistence simulation and GP ($\Delta T = |T_{coex} - T_{GP}|$). $\Delta T/\sigma$ is the error measured in standard deviations, value 1 is the expected value of the error, a value above 2 would indicate a statistically significant error. Our GP-based algorithm is thus in an excellent agreement with coexistence simulation results.

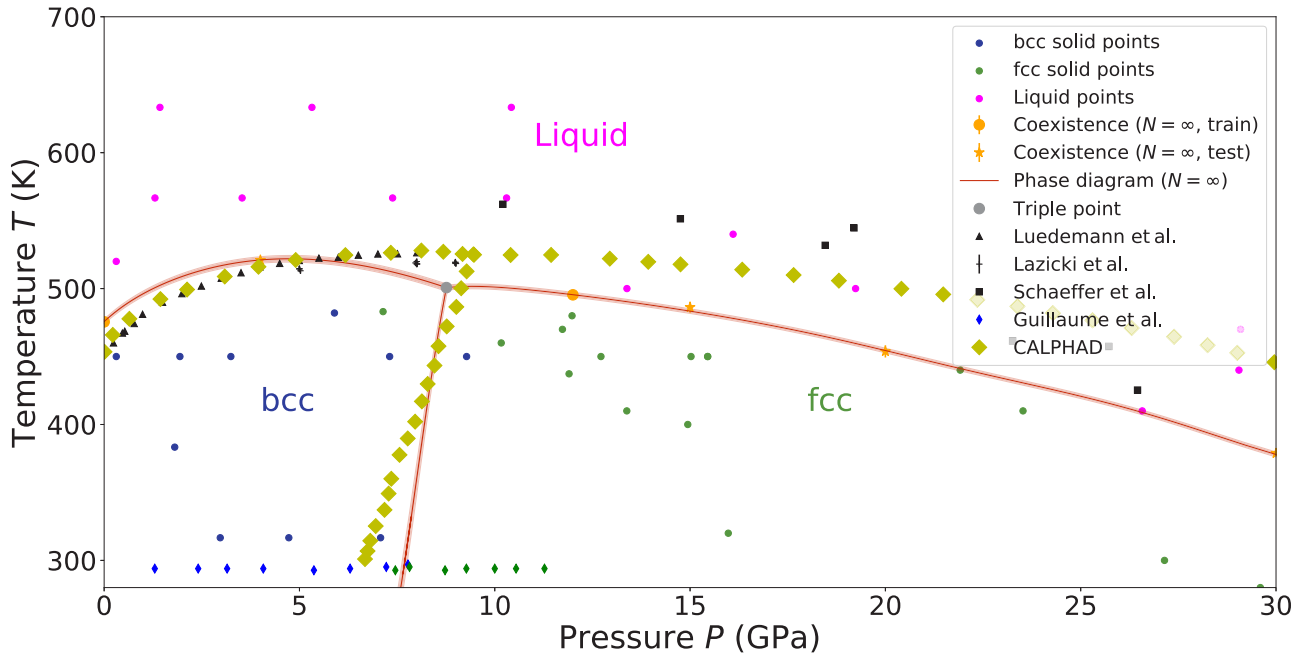| Phase transition	P, GPa	T_{coex}, K	T_{GP}, K	$\Delta T/\sigma$
fcc - liq*	12	495(2)	495(2)	0
fcc - liq	15	482(3)	483(2)	0.5
fcc - liq	20	456(2)	454(2)	0.7
fcc - liq	30	376.3(5)	378(2)	0.8
bcc - liq*	0	476(2)	476(2)	0
bcc - liq	4	521(3)	521(2)	0.03

FIG. 7. Lithium phase diagram at the infinite system size limit plotted with a confidence interval. Blue, green, and pink dots are training points for bcc, fcc, and liquid phases, respectively. The orange points are the coexistence simulation results. The red curve is the phase diagram obtained via GP. The gray dot is the location of the bcc-fcc-liquid triple point. The blue and green diamonds represent the fcc and bcc data from [22]. The experimental results [19–21] are denoted by black markers. The CALPHAD approximation, based on the data from [19–21], is shown by diamonds. Overall, we observe a very good agreement with the experimental data.

The total computational cost of constructing the phase diagram consists of three major parts: Obtaining the quantum-mechanical data for training the machine-learning potential, calculation of melting points, and running molecular dynamics to obtain the free-energy derivatives from statistical averages. The cost of the quantum-mechanical calculations
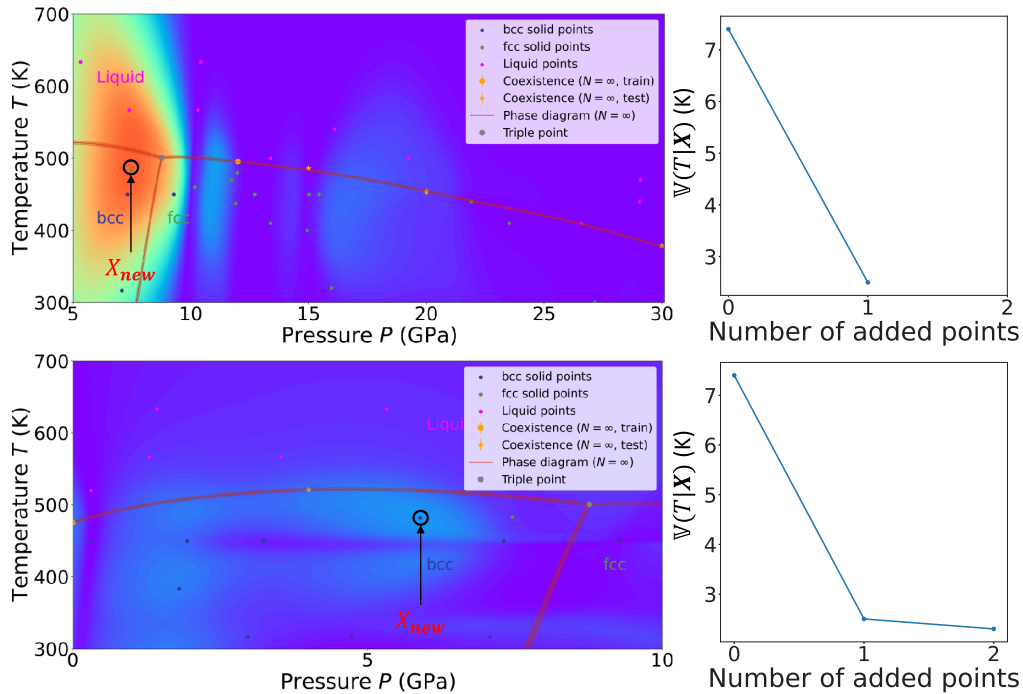

FIG. 8. The active sampling algorithm applied to the calculation of the triple-point temperature estimation of the lithium phase diagram. Color plot represents the value of the information function $\mathcal{H}$ from (14) at a given point of the phase diagram. The proposed active sampling algorithm allows one to optimize datasets corresponding to different phases.

was 28 000 CPU hours, about 100 hours per configuration (a typical calculation takes about 3 h when parallelized over 36 cores). The cost of the calculation of the two melting points used in the training set of the Gaussian process was about 10 000 h. Finally, the *NVT*-MD calculations took 16 000 CPU hours. The computational cost of the last part is comparable to the efficiency of a few melting-point calculations; however, our approach allows one to fit the entire free-energy surface with an error of less than 0.1 meV.

IV. CONCLUDING REMARKS

The construction of a phase diagram from atomistic simulation data is typically associated with a significant amount of manual work consisting of manual selection of numerical parameters, convergence tests, determining conditions at which to run simulations, analysis of the results, going back to the earlier stages if needed, etc. In the present work we have developed a Gaussian process-based methodology automating these stages of phase diagram calculation. In particular, the Gaussian process allows us to reconstruct the free-energy function based on various data sources (harmonic limit, zero-density limit, MD averages, and coexistence simulations). Furthermore, the dependence of the free energy on the numerical parameters can also be learned, which allows us to converge the results and estimate the exact value automatically together with the confidence interval that includes the error of extrapolation of the results with respect to the numerical parameters. On top of that, the Gaussian predictive variance allows for automatic sampling algorithms, automating the job of selecting the parameters for simulations. In view of these features of our approach, we believe that it will become increasingly useful with the rise of complex automatic protocols of calculating materials properties [38].

We have validated our algorithm on two model systems, soft-core potentials and the Lennard-Jones potentials, chosen as systems with plenty of available data in the literature. Our results agree very well with those from the literature, deviating only in those cases when the reference data we compare with is calculated with low values of convergence parameters (such as the Lennard-Jones cutoff radius). We then applied our methodology to lithium as an interesting example of a physical system. Comparison to our coexistence simulations as well as the existing experimental data shows a very good agreement, with discrepancies attributed to the error of DFT itself.

APPENDIX: DERIVATION OF FREE-ENERGY RELATIONS

1. Reference for solid

Let us expand the free energy (1) around $T = 0$. To that end we split x into two families of degrees of freedom: x_N

and $\tilde{x} = (x_1, \ldots, x_{N-1})$. Let $\beta := T^{-1}$. The determinant of the Jacobian of this transformation is 1 and hence we can write

$$\beta \hat{F}_{\text{ref}}(T) - \hat{S}(T) = -\log \int dx_N \int \exp\left(-\beta \hat{E}(\tilde{x}, x_N)\right) d\tilde{x}.$$

Because of translational symmetry, we can fix $x_N = 0$ in the inner (second) integral and hence

$$\beta \hat{F}_{\text{ref}}(T) - \hat{S}(T) = -\log \hat{V} - \log \int_{\hat{V}^{N-1}} \exp\left(-\beta E(\tilde{x}, 0)\right) d\tilde{x}.$$

Now note that for small T, only the energy near the ground state x_0 is relevant. Given a ground state x_0, it is repeated $(N - 1)!$ times in the integral because of permutation invariance in $E(\tilde{x}, 0)$. Hence

$$\beta \hat{F}_{\text{ref}}(T) - \hat{S}(T) = -\log \hat{V} - \log(N - 1)!$$
$$- \log \int_{\tilde{x} \sim x_0} \exp\left(-\beta E(\tilde{x}, 0)\right) d\tilde{x},$$

where the integration is taken specifically around the ground state x_0. $E(\tilde{x}, 0)$ has a nondegenerate Hessian around the ground state around which we will expand,

$$\beta \hat{F}_{\text{ref}}(T) - \hat{S}(T) = -\log \hat{V} - \log(N - 1)! + \beta \hat{E}_0$$
$$- \log \int \exp\left(-\langle x|\tilde{H}|x\rangle/(2T)\right) dx$$
$$+ O(T),$$

where $\tilde{H}$ is the Hessian of $E(\tilde{x}, 0)$ at the ground state. Carrying out the integration and calculating yields

$$\beta[\hat{F}_{\text{ref}}(T) - \hat{E}_0] - \hat{S}(T)$$
$$= -\log \hat{V} - \log(N - 1)!$$
$$\quad - \log[(2\pi T)/\det \tilde{H}]^{1/2} + O(T)$$
$$= -\log \hat{V} + \log(N) - \log N!$$
$$\quad - \frac{1}{2}(3N - 3)\log(2\pi T) + \frac{1}{2}\log \det \tilde{H} + O(T)$$
$$= -\log \hat{V} + \log(N) - N\log N + N - \log(2\pi N)$$
$$\quad + O(N^{-1})$$
$$\quad - \frac{1}{2}(3N - 3)\log(2\pi T) + \frac{1}{2}\log \det \tilde{H} + O(T)$$
$$= N[-\log N + 1 - \frac{3}{2}\log(2\pi T) + \frac{1}{2N}\log \det \tilde{H}]$$
$$\quad \times [-\log \hat{V} + \frac{1}{2}\log(2\pi) + \frac{3}{2}\log(T)] + O(T + N^{-1}).$$

We can now make the reverse change of variables from $\tilde{x}$ back to x, with $\det \tilde{H} = \det \hat{H}$, where $\hat{H}$ is the $(3N - 3) \times (3N - 3)$ Hessian of $\hat{E}$ computed at the ground state and projected onto the subspace orthogonal to the center of mass $x_c = \frac{1}{N}\sum_i x_i$. And finally, we will use the simplified version of this formula for the intensive quantities:

$$\beta[F_{\text{ref}}(T) - E_0] - S(T) = -\log N + 1 - \frac{3}{2}\log(2\pi T)$$
$$+ \frac{1}{2N}\log \det \hat{H} + O(T + N^{-1}).$$

Hence for solid, we choose

$$F_{\text{ref}}(T) := E_0 + T\left[-\log(NV) + 1 - \tfrac{3}{2}\log(2\pi T)\right],$$

which is the same as (2), and then (3) follows. We note that we could leave out V from F_{ref}, but then it would enter the expression for $\partial S/\partial V$ and hence create asymmetry with the liquid.

2. Derivative for solid

An *NVT*-thermostated molecular dynamics used in this work produces the averages

$$\langle f \rangle := \frac{\int f(\boldsymbol{x})\exp[-\beta\hat{E}(\boldsymbol{x})]d\boldsymbol{x}}{\int \exp[-\beta\hat{E}(\boldsymbol{x})]d\boldsymbol{x}}.$$

Here the integral is over $\hat{V}^N$, and we omit the region of integration when it is clear from the context.

Let us use this formula to find

$$\frac{\partial(\beta F_{\text{ref}} - S)}{\partial\beta} = -\frac{\partial}{\partial\beta}\log\int\exp[-\beta E(\boldsymbol{x})]d\boldsymbol{x}$$

$$= -\left(\int\exp[-\beta E(\boldsymbol{x})]d\boldsymbol{x}\right)^{-1}$$

$$\times \int[-E(\boldsymbol{x})]\exp[-\beta E(\boldsymbol{x})]d\boldsymbol{x}.$$

Then we can determine the relation between the mean potential energy and the free energy:

$$\langle E \rangle = \frac{\partial(\beta F_{\text{ref}} - S)}{\partial\beta} = -T^2\frac{\partial(\beta F_{\text{ref}} - S)}{\partial T}.$$

Hence

$$-\frac{\partial S}{\partial\beta} = -\frac{\partial(\beta F_{\text{ref}})}{\partial\beta} + \langle E \rangle$$

$$= -\frac{\partial}{\partial\beta}\left(\beta E_0 - \log N + 1 - \frac{3}{2}\log(2\pi T)\right) + \langle E \rangle$$

$$= -\frac{\partial}{\partial\beta}\left(\frac{3}{2}\log(\beta)\right) + \langle E - E_0 \rangle = \langle E - E_0 \rangle - \frac{3}{2}T,$$

or

$$\frac{\partial S(T, V)}{\partial T} = T^{-2}\langle E - E_0 \rangle - \frac{3}{2}T^{-1}. \tag{A1}$$

Let us find a similar expression for the partial derivative with respect to volume V:

$$\frac{\partial(\beta F_{\text{ref}} - S)}{\partial V} = -\frac{\partial}{\partial V}\log\int\exp[-\beta E(\boldsymbol{x})]d\boldsymbol{x}$$

$$= -\left(\int\exp[-\beta E(\boldsymbol{x})]d\boldsymbol{x}\right)^{-1}$$

$$\times \int\beta\left(-\frac{\partial E(\boldsymbol{x})}{\partial V}\right)\exp[-\beta E(\boldsymbol{x})]d\boldsymbol{x} - 1/V.$$

Assuming that $-\frac{\partial E(\boldsymbol{x})}{\partial V} = P(\boldsymbol{x})$, the formula can be rewritten as

$$\frac{\partial(\beta F_{\text{ref}} - S)}{\partial V} = -\left(\int\exp[-\beta e(\boldsymbol{x})]d\boldsymbol{x}\right)^{-1}$$

$$\times \int\beta P(\boldsymbol{x})\exp[-\beta E(\boldsymbol{x})]d\boldsymbol{x} - 1/V.$$

From this, we can derive the relation between the mean full pressure (sum of the ideal and virial parts) and free energy:

$$\langle P \rangle = -\frac{1}{\beta}\frac{\partial(\beta F_{\text{ref}} - S)}{\partial V} - \frac{1}{\beta V}.$$

Hence

$$-\frac{\partial S}{\partial V} = -\frac{\partial(\beta F_{\text{ref}})}{\partial V} - \beta\langle P \rangle$$

$$= -\frac{\partial}{\partial V}\left(\beta E_0 - \log[N(V)] + 1 - \frac{3}{2}\log(2\pi T)\right)$$

$$- \beta\langle P \rangle - \frac{1}{V}$$

$$= \beta(P_0 - \langle P \rangle),$$

or

$$\frac{\partial S}{\partial V} = \frac{\partial S(T, V)}{\partial V} = T^{-1}\langle P - P_0 \rangle. \tag{A2}$$

3. Calculation of the Hessian term

The term $N^{-1}\log\det\hat{H}$ can be calculated by integrating over the crystal Brillouin zone. Hessian is a matrix of second derivatives with respect to displacement. First, let us define Hessian for the interaction of two atoms. For this system, the Hessian is just a 3×3 matrix calculated as

$$H_{ij} = -\frac{\partial^2 E}{\partial r_i \partial r_j}, \qquad i, j \in \{1, 2, 3\}.$$

Here r_i, r_j are the components of the vector $\boldsymbol{r}$, and E is the potential energy of the system. The Hessian matrix for a larger system is constructed by adding such 3×3 blocks of two-atom interactions. For simplicity, we will only consider the fcc lattice case, which can be easily extended to an arbitrary lattice.

We start by denoting fcc the infinite fcc lattice with the point $(0, 0, 0)$ excluded. For a given vector in the k-space $\boldsymbol{k}$, cutoff r_{cut}, and the per-atom volume V, we define the Fourier transform of the Hessian matrix as

$$\tilde{H}_{ij} = \sum_{\substack{\boldsymbol{r} \in \textbf{fcc} \\ |\boldsymbol{r}| < r_{\text{cut}}}} H_{ij}(\mathbf{r})[1 - \cos(2\pi\boldsymbol{k}\cdot\boldsymbol{r})].$$

We also define the auxiliary function

$$\tilde{R}_{ij} = \sum_{\substack{\boldsymbol{r} \in \textbf{fcc} \\ |\boldsymbol{r}| < 1}} [1 - \cos(2\pi\boldsymbol{k}\cdot\boldsymbol{r})],$$

which will help us to integrate the singularity at $\boldsymbol{k} = (0, 0, 0)$.

The vectors $\boldsymbol{k}$ lie in the Brillouin zone of fcc lattice. The Brillouin zone Br of the fcc lattice is the set of points $\boldsymbol{k}$ defined as

$$\textbf{Br}\{\boldsymbol{k} : |\boldsymbol{k}|^2 \leqslant |\boldsymbol{k} - \boldsymbol{x}|^2, \boldsymbol{x} \in \textbf{bcc}\}.$$

Finally, we express the value of $\log\det\hat{H}$ as follows:

$$\log\det\hat{H} = 2\left(\int_{0 < k_1, k_2, k_3 < 1, \boldsymbol{k} \in \textbf{Br}} \log\det(\tilde{H}/\tilde{R})\right)$$

$$+ 6\left(\int_{0 < k_1, k_2, k_3 < 1, \boldsymbol{k} \in \textbf{Br}} \log\tilde{R}\right)$$

(note that $\tilde{H}$ is a 3×3 matrix, hence the factor of 6 in the second term). The material-independent part, $\int_{0 < k_1, k_2, k_3 < 1, \mathbf{k} \in \mathbf{Br}} \log \tilde{R}$, is integrated beforehand with an accuracy of 10^{-15}. The error of this calculation, for the purpose of feeding it to the Gaussian process regression framework, is as the mean square of accuracy and difference of $\log \det \hat{H}$ values for the nearest cutoffs.

4. Variance of a nonlinear functional

The melting point, defined as a system of equations (10), is a nonlinear functional of Gaussian processes. In this case we reduce the problem to a linear functional, obtained by expanding the original functional in the Taylor series around the mean of the Gaussian process. We denote the original nonlinear functional as $\mathcal{F} = \mathcal{F}(S)$. Then the functional is linearized near the solution and the quantity of interest is now approximated with a linear functional, $\mathcal{F}(S) \approx \mathcal{F}(\overline{S}) + \langle S - \overline{S}, J \rangle$, where $\overline{S}$ is the mean predicted entropy, and J is the gradient (Jacobian) of $\mathcal{F}$ evaluated at $\overline{S}$. We can then evaluate variance, similarly to (12):

$$\mathbb{V}[\mathcal{F}] \approx \mathbb{V}[J] = K(J, J) - K(J, X)^T K(X, X)^{-1} K(J, X).$$

$$(A3)$$

[1] N. Saunders and A. P. Miodownik, *CALPHAD (Calculation of Phase Diagrams): A Comprehensive Guide* (Elsevier, New York, 1998).

[2] P. J. Spencer, A brief history of Calphad, Calphad **32**, 1 (2008).

[3] E. Königsberger and G. Eriksson, A new optimization routine for ChemSage, Calphad **19**, 207 (1995).

[4] E. Königsberger, Improvement of excess parameters from thermodynamic and phase diagram data by a sequential Bayes algorithm, Calphad **15**, 69 (1991).

[5] M. Stan and B. Reardon, A Bayesian approach to evaluating the uncertainty of thermodynamic data and phase diagrams, Calphad **27**, 319 (2003).

[6] B. Bocklund, R. Otis, A. Egorov, A. Obaied, I. Roslyakova, and Z.-K. Liu, ESPEI for efficient thermodynamic database development, modification, and uncertainty quantification: Application to Cu-Mg, MRS Commun., **9**, 618 (2019).

[7] R. A. Otis and Z.-K. Liu, High-throughput thermodynamic modeling and uncertainty quantification for ICME, JOM **69**, 886 (2017).

[8] N. H. Paulson, B. J. Bocklund, R. A. Otis, Z.-K. Liu, and M. Stan, Quantified uncertainty in thermodynamic modeling for materials design, Acta Mater. **174**, 9 (2019).

[9] S. Zomorodpoosh, B. Bocklund, A. Obaied, R. Otis, Z.-K. Liu, and I. Roslyakova, Statistical approach for automated weighting of datasets: Application to heat capacity data, Calphad **71**, 101994 (2020).

[10] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Elsevier, New York, 2001), Vol. 1.

[11] E. A. Mastny and J. J. de Pablo, Melting line of the Lennard-Jones system, infinite size, and full potential, J. Chem. Phys. **127**, 104504 (2007).

[12] T. Sjostrom, S. Crockett, and S. Rudin, Multiphase aluminum equations of state via density functional theory, Phys. Rev. B **94**, 144101 (2016).

[13] I. A. Kruglov, A. Yanilkin, A. R. Oganov, and P. Korotaev, Phase diagram of uranium from ab initio calculations and machine learning, Phys. Rev. B **100**, 174104 (2019).

[14] C. W. Glass, A. R. Oganov, and N. Hansen, USPEX—Evolutionary crystal structure prediction, Comput. Phys. Commun. **175**, 713 (2006).

[15] A. Reinhardt and B. Cheng, Quantum-mechanical exploration of the phase diagram of water, Nat. Commun. **12**, 588 (2021).

[16] J. R. Morris and X. Song, The melting lines of model systems calculated from coexistence simulations, J. Chem. Phys. **116**, 9352 (2002).

[17] J. Caillol, Critical-point of the Lennard-Jones fluid: A finite-size scaling study, J. Chem. Phys. **109**, 4885 (1998).

[18] J. Pérez-Pellitero, P. Ungerer, G. Orkoulas, and A. D. Mackie, Critical point estimation of the Lennard-Jones pure fluid and binary mixtures, J. Chem. Phys. **125**, 054515 (2006).

[19] H. Luedemann and G. Kennedy, Melting curves of lithium, sodium, potassium, and rubidium to 80 kilobars, J. Geophys. Res. **73**, 2795 (1968).

[20] A. Lazicki, Y. Fei, and R. J. Hemley, High-pressure differential thermal analysis measurements of the melting curve of lithium, Solid State Commun. **150**, 625 (2010).

[21] A. M. J. Schaeffer, W. B. Talmadge, S. R. Temple, and S. Deemyad, High Pressure Melting of Lithium, Phys. Rev. Lett. **109**, 185702 (2012).

[22] C. L. Guillaume, E. Gregoryanz, O. Degtyareva, M. I. McMahon, M. Hanfland, S. Evans, M. Guthrie, S. V. Sinogeikin, and H. Mao, Cold melting and solid structures of dense lithium, Nat. Phys. **7**, 211 (2011).

[23] J. Dorrell and L. B. Pártay, Pressure-temperature phase diagram of lithium, predicted by embedded atom model potentials, J. Phys. Chem. B **124**, 6015 (2020).

[24] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, Oxford, England, 2017).

[25] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, J. Comput. Phys. **117**, 1 (1995).

[26] A. V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, Multiscale Model. Simul. **14**, 1153 (2016).

[27] K. Gubaev, E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev, Accelerating high-throughput searches for new alloys with active learning of interatomic potentials, Comput. Mater. Sci. **156**, 148 (2019).

[28] U. Grenander and M. I. Miller, Representations of knowledge in complex systems, J. Roy. Stat. Soc. Ser. B **56**, 549 (1994).

[29] A. Ahmed and R. J. Sadus, Solid-liquid equilibria and triple points of n-6 Lennard-Jones fluids, J. Chem. Phys. **131**, 174504 (2009).

[30] A. Ladd and L. Woodcock, Interfacial and co-existence properties of the Lennard-Jones system at the triple point, Mol. Phys. **36**, 611 (1978).

[31] J.-P. Hansen, Phase transition of the Lennard-Jones system. II. High-temperature limit, Phys. Rev. A **2**, 221 (1970).

[32] D. A. Kofke, Direct evaluation of phase coexistence by molecular simulation via integration along the saturation line, J. Chem. Phys. **98**, 4149 (1993).

[33] I. S. Novikov, K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, The MLIP package: Moment tensor potentials with MPI and active learning, Mach. Learn.: Sci. Technol. **2**, 025002 (2020).

[34] G. Kresse and J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B **47**, 558 (1993).

[35] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors us-ing a plane-wave basis set, Comput. Mater. Sci. **6**, 15 (1996).

[36] G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

[37] E. V. Podryabinkin and A. V. Shapeev, Active learning of linearly parametrized interatomic potentials, Comput. Mater. Sci. **140**, 171 (2017).

[38] J. Janssen, S. Surendralal, Y. Lysogorskiy, M. Todorova, T. Hickel, R. Drautz, and J. Neugebauer, pyiron: An integrated development environment for computational materials science, Comput. Mater. Sci. **163**, 24 (2019).