Database optimization for empirical interatomic potential models

Pinchao Zhang and Dallas R Trinkle

Department of Materials Science and Engineering, University of Illinois, Urbana-Champaign, Champaign, IL, USA

E-mail: zhang119@illinois.edu and dtrinkle@illinois.edu

Received 13 February 2015, revised 25 May 2015
Accepted for publication 2 July 2015
Published 11 August 2015

Abstract

Weighted least squares fitting to a database of quantum mechanical calculations can determine the optimal parameters of empirical potential models. While algorithms exist to provide optimal potential parameters for a given fitting database of structures with corresponding energy-related predictions and to estimate prediction errors using Bayesian sampling, defining an optimal fitting database based on potential predictions remains elusive. A testing set of structures and energy-related predictions provides an empirical measure of potential transferability. Here, we propose an objective function for fitting databases based on testing set errors. The objective function allows the optimization of the weights in a fitting database, the assessment of the adding or removing of structures in the fitting database, or the comparison of two different fitting databases. To showcase this technique, we consider an example Lennard-Jones potential for Ti, where modeling multiple complicated crystal structures is difficult for a radial pair potential. The algorithm finds different optimal fitting databases, depending on the objective function of potential prediction error for a testing set.

Keywords: empirical potential, database optimization algorithm, Bayesian sampling, potential transferability, weight optimization

(Some figures may appear in colour only in the online journal)

1. Introduction

Atomic-scale simulations have the capability to predict the properties of defect structures that are often inaccessible by experimental techniques [1–8]. These predictions require accurate and efficient calculations of energies and forces on atomic configurations that sample a variety of atomic environments, and would even represent different bonding configurations. Accurate quantum mechanical methods are difficult to scale to large simulation systems and
long simulation times, while empirical interatomic potentials offer increased computational efficiency at a lower level of accuracy. Maximizing the efficiency of computational material science studies requires the development of potentials that are transferable, i.e. capable of predicting properties outside their fitting range, and accurate for static and dynamic calculations.

Without direct transferable derivations of interatomic potentials from quantum mechanical methods, empirical interatomic potential modeling approximates quantum mechanical methods with different potential functional forms. The potential fitting community has proposed a variety of functional forms for empirical potentials, including embedded-atom method (EAM) [4, 9], modified embedded-atom method (MEAM) [2, 10, 11] and charged-optimized many-body potential (COMB) [12, 13]. There have been multiple implementations of different potential functional forms for various materials [7, 9, 14–19]. Even for the same type of materials, such as Cu [7, 20, 21] and Si [10, 14, 22–24], different empirical interatomic potential models are proposed for different applications with different transferabilities. There are advanced techniques to optimize the potential parameters based on a weighted least-squares regression to a fitting database of experimental or quantum mechanical calculation data [4, 20], including the force-matching method [25] for empirical interatomic potential parameter optimization. In force-matching, a fitting database includes quantum mechanical force calculations for diverse atomic environments to obtain realistic empirical potential models. To study the transferability of the empirical potential model, Frederiksen et al applied Bayesian statistics to empirical interatomic potential models: instead of using the best fit, an ensemble of neighboring parameter sets reveal the flexibility of the model [26]. They showed that the standard deviation of the potential predictions is a good estimate of the true error. However, even with these advances, the determination of empirical interatomic potentials relies on the selection and weighting of a fitting database without a clear, quantitative guide for the impact on predictions.

To address the issue of fitting database selection, we present an automated, quantitative fitting-database optimization algorithm based on prediction errors in a testing set using Bayesian statistics. We construct an objective function of the prediction errors in the testing set to optimize the relative weights of a fitting database. The weight change represents three conventional operations on a fitting database. We demonstrate the optimization algorithm with a simple interatomic potential model: Lennard-Jones potential fitting of Ti crystal structures. We choose this example because a radial potential has difficulty describing the stability of different crystal structures of a transition metal. The new algorithm also helps to understand the transferability of the empirical potential model for the structures in the testing set.

We start with a brief introduction of the fitting procedure of empirical potential models and the mathematical description of the potential fitting process. Next, we discuss Bayesian error estimation as it applies to our problem. Then we define an objective function with a testing set, and use this quantitative measure to devise an algorithm to optimize a fitting database. Lastly, we demonstrate this new approach on an example system with clear limitations: Lennard-Jones potential for Ti.

2. Introduction and difficulties in empirical potentials fitting

Empirical potential models predict energies and forces more efficiently than density-functional theory (DFT) [27, 28]. DFT calculations can provide accurate energies and forces for at most a few thousand atoms, while interatomic potentials can easily simulate millions of atoms. However, empirical potential fitting requires great care to construct accurate potentials. In atomic-scale simulations, a structure \( \alpha \) consists of atoms with chemical identities \( \lambda_m \) at positions \( \mathbf{R}_m : \alpha = \{ (\mathbf{R}_m, \lambda_m) \} \), and the total energy is a function of those identities and positions.
Empirical interatomic potentials are nonlinear functions of atomic positions, described by a potential parameters $\theta$, where

$$E_{\theta}(\theta) \equiv E(\{\{\mathbf{r}_i, \chi_i\}\}; \theta) = \frac{1}{2} \sum_{mn} V_{mn}^{E,\chi}(\mathbf{r}_m - \mathbf{r}_n; \theta) + \frac{1}{3} \sum_{mnl} V_{mnl}^{E,\chi}(\mathbf{r}_m - \mathbf{r}_n, \mathbf{r}_n - \mathbf{r}_l; \theta) + \cdots,$$

(1)

for interatomic potential functions $V_{m}^{E,\chi}$ between $M$ atoms of chemical identity $\{\chi_1, \ldots, \chi_k\}$. A general empirical interatomic potential that reproduces all DFT energy calculations accurately is computational intractable, since it would require a large number of many-body terms. Rather, we are interested in simpler potentials that provide accurate results for a smaller domain of atomic configurations including perfect crystals and defect structures under various thermodynamic conditions; this includes potentials that may not be easily written in the form of equation (1), such as EAM and MEAM potentials. Besides energies and forces, empirical potentials can compute quantities related to energies and their derivatives such as lattice constants, defect formation energies, and elastic constants. We define a structure property $A_\alpha$ as a measurable quantity that depends on the energy or energy derivatives with respect to position of a structure $\alpha$. One can evaluate the value for a structure property using DFT, experiments, or empirical potentials. The structure property function $A_\alpha(\theta)$ is the potential prediction for the structure property as a function of the potential parameters, and is highly nonlinear in the parameters $\theta$.

The goal of the potential optimization is to find a set of parameters that provide accurate structure property predictions. Potentials are optimized against a fitting database: a weighted set of structure property values for the potential to reproduce through structure property functions. This requires nonlinear optimization to find the ‘best parameters’. These best parameters are tested against a different set of structure properties (the testing set) to determine the transferability of the empirical potential. If the potential predictions are unsatisfactory—which is quite common at the start of potential optimization—manual modifications are made to the fitting database to improve potential transferability: adding and removing structure property values, or tuning the weights. With the modified fitting database, one would optimize the parameters and examine the structure property predictions in the testing set. This circuit repeats until the differences of the structure property predictions and the structure property values are acceptable. If the potential remains untransferable for the testing set structure properties, one may either reduce the transferability demands by removing structure property values from the testing set, or change the potential functional form to increase flexibility.

The manual modifications rely on non-quantitative connections between the fitting database and testing set predictions, and preclude automation. Normally, structure properties are chosen to sample different atomic environments, e.g. force-matching [25], and the weights tuned to achieve a ‘balance’ of errors. Here, we construct an objective function that captures this balance of errors in the testing set. We can quantify how modifications to the fitting database affect predictions in the testing set, and propose an algorithm to automatically optimize the fitting database according to the testing set. A mathematical description of the potential fitting process is essential for quantitative analysis and automation of empirical potential fitting.

3. Mathematical description of empirical potentials fitting process

The fitting database $F$ is a set $\{(A_\alpha, w_\alpha)\}$ of structure property values $A_\alpha$ with an associated database entry index $\alpha$ and non-negative relative weight, $w_\alpha \geq 0$. We impose a trivial constraint
that the sum of all the weights are equal to one, as only relative weight values are important. The finite set of structure property values with positive weights contributes to the fitting, while structure property values with zero weights do not. In this way, we can consider a fitting database that includes all possible structure property values that could be taken into account for potential fitting, while assigning a weight of 0 to any value that is not used. Then all modifications to a fitting database entail changes in weights: (1) tuning: modify a non-zero weight to a new non-zero weight, (2) adding: change a zero weight to a non-zero weight, (3) removing: change a non-zero weight to zero. The prediction error $\epsilon_\theta(\theta)$ is the 2-norm of the difference of a structure property function and structure property value of a structure $\theta$,

$$
\epsilon_\theta(\theta) = \|A^\theta(\theta) - A^\theta_{ref}\|_2,
$$

where $A^\theta(\theta)$ is the potential prediction of a structure property function given parameters $\theta$, $A^\theta_{ref}$ is the structure property value from DFT or experiment and $\|\cdot\|_2$ denotes the 2-norm of a $d$-dimensional vector $x$

$$
\|x\|_2 = \left(\sum_{n=1}^{d} |x_n|^2\right)^{\frac{1}{2}}.
$$

If we want to fit the energy of a structure $\theta$, we use the difference in energy between a structure $\theta$ and a reference structure $0$,

$$
A^\theta = E^\theta - E^0.
$$

The energy prediction error is

$$
\epsilon_\theta(\theta) = |(E^\theta(\theta) - E^0(\theta)) - (E^\theta_{ref} - E^0_{ref})|.
$$

If we want to fit the forces on atoms of a structure $\theta$, we use a $3N$ vector of forces on the $N$ atoms. The force prediction error is

$$
\epsilon_\theta(\theta) = \|F^\theta(\theta) - F^\theta_{ref}\|_2.
$$

The weighted summed squared error function for a fitting database $F$ is

$$
S(\theta, F) = \sum_{\theta \in F} w_\theta \epsilon_\theta^2(\theta).
$$

One minimizes $S(\theta, F)$ in the fitting database to find the optimal parameter set of the potential model. The testing set $T$ is a set of structure properties $A^\beta_T$ with an associated database entry index $\beta$. There are no relative weights for structures in a testing set; rather, it represents the domain of atomic environments that one expects the empirical potential to predict accurately. There is no requirement that the fitting database and the testing set overlap; however, we expect that the fitting database structure property functions will often be included in the testing set.

To compare different databases against a single testing set $T$, we define the ‘objective function’. The objective function of the fitting database $F$ is, given the testing set $T$,

$$
O(F|T) = \sum_{\beta \in T} \ln t(\langle \epsilon_\theta^2(\theta) \rangle_\beta),
$$

where $t$ is the threshold function, and $\langle \epsilon_\theta^2(\theta) \rangle_\beta$ is the mean squared error of a structure property function in the testing set evaluated using the Bayesian error estimation method, which will be discussed in section 4. The threshold function $t(x)$ prevents the logarithm of the error from reaching $-\infty$ as the error approaches zero,
\[
\ell(x) = \begin{cases} 
  x^2/4\epsilon_0^2 + \epsilon_0^2 & : x < 2\epsilon_0^2 \\
  x & : x \geq 2\epsilon_0^2 
\end{cases}.
\] (9)

We can choose different error tolerances \( \epsilon_0 \) for different structure properties. Equation (8) is monotonic and approximates the relative difference in errors in the testing set. The relationship between the objective function and relative errors is described in appendix A. We obtain the optimal weights in the fitting database by minimizing the objective function. The analytic derivative with respect to the relative weights is available with the knowledge of \( \partial (\ell^2(\theta))_\theta / \partial \omega_a \) in section 4. Standard gradient descent algorithms can be used to minimize the objective function, such as conjugate gradient and Broyden–Fletcher–Goldfarb–Shanno [29]. Next, we develop an automated algorithm to optimize the objective function with respect to the weights in the fitting database to minimize the relative errors in the testing set.

4. Database optimization algorithm using Bayesian statistics

Bayesian inference provides a systematic framework to infer the flexibility of the models from data using conditional probability [26, 30]. Bayesian statistics treats model parameters as random variables with a probability distribution given by the posterior distribution. Bayesian inference measures the flexibility of the model by generating an ensemble of good fits in the neighborhood of the best fit. We draw samples from the posterior probability distribution of the parameters, \( P(\theta|F) \), to estimate the errors of the predictions. According to the Bayes’ theorem, the posterior distribution of the parameters is a product of the prior distribution \( \pi(\theta) \) and the likelihood function \( L(F|\theta) \),

\[
P(\theta|F) \propto \pi(\theta) \times L(F|\theta),
\] (10)

where the prior distribution \( \pi(\theta) \) includes the information about the potential model before we take the fitting data into account and the likelihood function is the likelihood of obtaining the fitting database provided with a parameter set \( \theta \). Here we use the maximally unbiased prior distribution of a uniform distribution over a measurable set \( \mathcal{H} \) of allowed parameters sets,

\[
\pi(\theta) = \left[ \int_{\mathcal{H}} d\theta \right]^{-1},
\] (11)

though other choices are possible. Assuming the errors are independent and identically normally distributed, the likelihood function takes the form of the multivariate normal distribution [26, 31]

\[
L(F|\theta) \propto \exp \left( -\frac{1}{W} \sum_{a \in \hat{F}} w_a \ell_a^2(\theta) \right),
\] (12)

where

\[
W = S(\theta^{MLE}, F).
\] (13)

\( \theta^{MLE} \) is the maximum likelihood estimation (MLE) of the likelihood function. The likelihood function is independent of the sum of the weights, and a zero-weighted database entry would appear not to be present in the optimization.

The Bayesian prediction of a function \( A(\theta) \) is the conditional expectations or mean
The Bayesian error is the mean squared error of the Bayesian prediction:

$$\langle \epsilon_\theta^2(\theta) \rangle_F = \langle A(\theta) \rangle_F - A^2 + \text{var}_F(A(\theta)),$$

where \(\text{var}_F(A(\theta)) = \langle A^2(\theta) \rangle_F - \langle A(\theta) \rangle_F^2\) is the variance of the Bayesian prediction. The covariance of two functions \(A_a(\theta)\) and \(A_p(\theta)\) represents the correlation between two functions:

$$\text{cov}_F(A_a(\theta), A_p(\theta)) = \langle A_a(\theta) A_p(\theta) \rangle_F - \langle A_a(\theta) \rangle_F \langle A_p(\theta) \rangle_F.$$

The quantities of interest above such as mean, variance and covariance share the same form of expectation value of different functions where the probability distribution of random variable \(\theta\) is the posterior distribution. For complicated high-dimensional parameter space such as empirical potentials, the integral in equation (14) cannot be evaluated in closed form, and the high-dimensionality makes direct numerical quadrature converge slowly. We instead use Markov Chain Monte Carlo (MCMC) \([32]\) to sample the posterior distribution of the potential parameter \(\theta\) and calculate the integral in equation (14) numerically. The chain of \(N\) samples will contain a set of \(N\) independent samples \(\{\theta_n\}\) (where \(N/\text{samples} \text{ is the autocorrelation length}), and the numerical estimate of the mean is

$$\langle A_a(\theta) \rangle_F \approx \frac{1}{N} \sum_{n=1}^{N} A_a(\theta_n),$$

with a sampling error of \(\sqrt{\text{var}_F(A_a(\theta))/N}\). The best set of parameters \(\theta^{\text{MLE}}\) defines the empirical potential for predictions, and the ensemble of parameters \(\{\theta_n\}\) allows the estimation of errors on those predictions.

All the averages are implicit functions of the relative weights in the fitting database. The derivative of a Bayesian prediction with respect to weight is

$$\frac{\partial \langle A(\theta) \rangle_F}{\partial w_i} = \frac{\partial}{\partial w_i} \left( \frac{1}{N} \sum_{n=1}^{N} A_a(\theta_n) \right),$$

and

$$\frac{\partial \log L(F|\theta)}{\partial w_i} = \frac{\partial^2 \langle \epsilon_\theta^2(\theta) \rangle_F}{\partial w_i^2} + \frac{\partial w}{\partial w_i} \log L(F|\theta),$$

where

$$\frac{\partial \epsilon_\theta^2(\theta)}{\partial w_i} = \frac{\partial}{\partial w_i} \left( \frac{1}{N} \sum_{n=1}^{N} A_a(\theta_n) \right).$$

We can calculate any Bayesian mean derivative with respect to weight in terms of covariance from the MCMC sampling chain. Note that

$$\frac{\partial \log L(F|\theta)}{\partial w_i} = -\frac{\partial^2 \langle \epsilon_\theta^2(\theta) \rangle_F}{\partial w_i^2} + \frac{\partial w}{\partial w_i} \log L(F|\theta).$$

The derivative of \(W\) with respect to weight is found using the chain rule,

$$\frac{\partial W}{\partial w_i} = \frac{\partial S(\theta, F)}{\partial w_i} \bigg|_{\theta^{\text{MLE}}}.$$
\[ + \sum_n \frac{\partial S(\theta, F)}{\partial \theta_n} \bigg|_{\theta_{n, MLE}} \frac{\partial \theta_{n, MLE}}{\partial w_n} \]

(22)

\[ = \epsilon_n^2(\theta_{MLE}), \]

(23)

as \( \theta_{MLE} \) is an extremum of \( S(\theta, F) \). Applying equations (19)–(23) to equation (15) yields

\[ \frac{\partial (\epsilon_2^2(\theta))_F}{\partial w_n} = \text{cov}_{\theta} \left( \epsilon_n^2(\theta), -\frac{\epsilon_n^2(\theta) + \epsilon_n^2(\theta_{MLE}) \log L(F|\theta)}{W} \right) \]

(24)

\[ = -\frac{1}{W} \left( C_{\theta}^F = \frac{\epsilon_n^2(\theta_{MLE})}{W} \sum_f w_f C_{f} \right), \]

(25)

where

\[ C_{\theta}^F = C_{\theta}^F = \text{cov}_{\theta}(\epsilon_n^2(\theta), \epsilon_n^2(\theta)). \]

(26)

Therefore, we can calculate the analytic derivative of \( O(F;T) \) with respect to weights with Bayesian error estimation method and further optimize the objective function. Finally, note that as our likelihood function is independent of \( \sum_n w_n \),

\[ \sum_n w_n \frac{\partial O(F;T)}{\partial w_n} = 0, \]

(27)

The optimal weights found by minimizing \( O(F;T) \) determine the adding and removing of structures in the fitting database. According to the definition of the fitting database, we fit the structures with positive weight, and all the other structures that do not contribute to the fitting will have a weight of zero. The optimal weight value can be determined even for structures not presently in the fitting database. A structure can be added to the fitting database when its optimal weight value is positive, since the adding of that structure decreases the relative error in the testing set. A structure can be removed from the fitting database when its optimal weight value is zero.

Figure 1 outlines the new automated empirical interatomic potential fitting algorithm using Bayesian statistics. We build a testing set, which describe the expected transferability domain of the potential in the first place. Then we construct a finite fitting database consisting of a set of structure property values from DFT calculations, and assign each structure property with a non-negative relative weight. We use non-linear weighted least squares optimizers to find the MLE, and use MCMC [32] sampling of the posterior distribution in equation (10) to generate the ensemble of parameters. We calculate the mean-squared errors of structure property function predictions in the testing set, the objective function, and its gradients using the parameter ensemble. Next, we obtain the optimal weight set for the fitting database by minimizing the objective function using a conjugate-gradient method, and assess the adding or removing of structures to the fitting database according to the weights during optimization. This step can take advantage of structure searching methods [33], for example, to identify candidate structures, though we do not do so here. We repeat the circuit with the modified relative weight set of the fitting database until the objective function value converged.

The testing set is the key component of this approach not only because the objective function consists of the mean squared errors of the testing set structure properties, but also one wants the empirical potential predictions for structure properties in the testing set to have
small errors—whether that is known from comparison with DFT calculations or estimated from Bayesian sampling without DFT. With the relative errors in the testing set minimized, any weight deviation from the optimal will result in an increase in relative errors. This means that while we could choose weights to reduce the error of one or several testing set structure property function predictions, it will worsen the predictions of other structures and the trade-off is not worthwhile. An optimal fitting database however does not guarantee a reliable empirical potential model. The optimization algorithm provides the best possible empirical potential for a given a fitting database and a given testing set, but it has no judgment on
whether the optimal Bayesian errors are below the expected error thresholds; they can, in fact, be quite large. This can occur if the empirical potential model does not contain the relevant physics to describe the atomic environments in the testing set, which reduces transferability. Then, we must—for predictive empirical potential methods—decide to improve the potential model itself or remove structures from the testing set to optimize for reduced transferability.

5. Implementation on Lennard-Jones Potential fitting of Ti

5.1. Potential form and calculation details

We apply the database optimization algorithm to a simple empirical interatomic potential model, the Lennard-Jones potential. The Lennard-Jones potential is a two-parameter pair potential:

\[
V_2(r; r_0, E_b) = \begin{cases} 
4E_b \left( \frac{r_0}{r} \right)^{12} - \left( \frac{r_0}{r} \right)^6 & : r \leq r_{\text{cutoff}} \\
0 & : r > r_{\text{cutoff}}
\end{cases}
\]

(28)

where \(E_b\) is the binding energy of a dimer with a separation of \(\sqrt{2}r_0\). We choose the cutoff radius \(r_{\text{cutoff}} = 3r_0\) and the allowable parameters are \(r_0 > 0, E_b > 0\).

The DFT calculations for six different crystal structures of Ti are performed with vasp [34, 35], a plane-wave density functional code. We apply a Ti ultrasoft Vanderbilt type pseudopotential [36], with a plane-wave cutoff energy of 400 eV for energy convergence of 0.3 meV/atom [15]. The \(k\)-point meshes for different structures are, \(16 \times 16 \times 12\) for hcp, \(32 \times 32 \times 32\) for bcc, \(24 \times 24 \times 24\) for fcc, \(16 \times 16 \times 16\) for hexagonal, \(8 \times 8 \times 8\) for A15 and \(12 \times 12 \times 20\) for \(\omega\), with Methfessel–Paxton [37] smearing parameter of 0.2 eV to obtain an energy accuracy of 1 meV/atom [15, 38]. The energy versus volume data includes four different structures with volume of the unit cell as \(0.95V_0, 0.975V_0, 1.025V_0\) and \(1.05V_0\), where \(V_0\) is the unit cell volume of the equilibrium structure. The fitting databases contain various energy differences and energy versus volume data combinations among the six crystal structures. We generate the Markov chain of the potential parameters using the Metropolis–Hasting algorithm [32]. The potential parameter ensemble contains \(10^4\) independent parameter sets from the MCMC simulation with \(10^6\) attempted steps, with an auto-correlation length of approximately 100. We use a reweighting scheme discussed in appendix B to approximate the objective function values for all possible sets of weights with only one sampling run. Since a radial potential model does not describe the physics of metallic bonding, we expect that the Lennard-Jones potential will not be transferable for all six structures. We expect our algorithm to identify the lack of transferability of LJ potential in the optimization. We systematically consider different types...
of fitting databases and testing sets with this in mind. For simplicity, we define $E_{AB}$ as the energy difference between structure A and B, $E_{\text{vac}}$ as the vacancy formation energy, $-w_{AB}$ as the weight of the energy difference between structure A and B, and $-w_{AE \text{ vol}}$ as the weight of the energy versus volume of structure A. We start with a simple fitting database that contains two energy differences, and a testing set with the same entries.

5.2. Two-structured fitting database

Figure 2 shows the objective function as a function of the weight in two different fitting database and testing set combinations. Figure 2(a) shows the objective function of a fitting database with $-E_{\text{bcc}}$ and $-E_{\text{A15-fcc}}$. The objective function has a unique minimum with an optimal relative weight ratio of the two structures. Moreover, if we calculate the derivative of the objective function with respect to weight at endpoints (where one weight is zero), we can see that each derivative of the objective function with respect to the weights indicates that the other structure should be included in the fitting database. Therefore the optimal weight value for both fitting database structures are positive, and we refer this as a ‘mixed’ fitting database.

On the other hand, figure 2(b) shows the objective function of a fitting database with $-E_{\text{fcc}}$ and $-E_{\text{hcp}}$. The objective function reaches minima at one of the endpoints, which means that a fitting database containing both $E_{\text{hcp}}$ and $E_{\text{fcc}}$ has higher relative errors for the testing set than a ‘pathological’ fitting databases with only one structure. This is due to the non-transferability between hcp and fcc structures. We refer to these pathological cases as ‘unmixed’ fitting databases.

Figure 3 shows the result of optimizing all possible combinations of two-structured fitting databases with two energy differences sharing a common reference structure. Databases with physical MLEs with positive $E_b$ and $r_0$ are either mixed or unmixed two-structured fitting databases. Most mixed fitting databases include fcc, bcc, hex and A15 structures and most unmixed fitting databases includes hcp or $\omega$ energy differences. By exploring a wide phase space (six crystal structures of Ti) of Lennard-Jones potential fitting, we have shown that the database optimization algorithm offers an automated, systematic and quantitative way of analyzing empirical potential model fitting with different fitting databases and testing sets.
5.3. Three-structured fitting database

Figure 4 is a Gibbs triangle (so that $\sum w_i = 1$) contour plot of the objective function for a three-structured fitting database including $E_{\text{bcc-fcc}}$, $E_{\text{hex-fcc}}$, and $E_{\text{A15-fcc}}$. The fitting database includes three energy differences with a unique reference structure, where the testing set remains the same entries as the fitting database. While all three of the two-structured fitting databases are mixed fitting databases, the minimum occurs between $E_{\text{bcc-fcc}}$ and $E_{\text{hex-fcc}}$. We start with an equal initial weight (center in Gibbs triangle contour) and the optimal weight values are $w_{E_{\text{bcc-fcc}}} = 0.46$, $w_{E_{\text{hex-fcc}}} = 0.54$ and $w_{E_{\text{A15-fcc}}} = 0$. The weight of $E_{\text{A15-fcc}}$ that changes from positive to zero means that the algorithm automatically removes $E_{\text{A15-fcc}}$ from the fitting database. The gradient of $w_{E_{\text{A15-fcc}}}$ from the two-structured fitting database $E_{\text{bcc-fcc}}$ and $E_{\text{hex-fcc}}$ is positive meaning adding $E_{\text{A15-fcc}}$ increases the total relative errors in the testing set. Although adding $E_{\text{A15-fcc}}$ can reduce the prediction error of $E_{\text{A15-fcc}}$, it will increase the prediction errors for the other two structures and the trade-off is not worthwhile. Figure 5 shows comparison of the prediction distributions evaluated at the initial weights and the optimal weights. The prediction distribution shows the prediction error trade-off of removing $E_{\text{A15-fcc}}$ from the fitting database. We are able to reduce the prediction error of $E_{\text{bcc-fcc}}$ and $E_{\text{hex-fcc}}$ more than the increase in the prediction error of $E_{\text{A15-fcc}}$. Moreover, any weight deviation from the optimal weight will lead to an increase of the total relative error of the three energy differences.
differences. It shows that the new database optimization algorithm is able to automatically provide quantitative information about the transferability of the potential model.

5.4. Energy differences and volume changes

We now apply the algorithm to larger fitting databases and testing sets. The testing set includes five energy differences using hcp as the reference structure and six energy versus volume data. For each energy versus volume data, we use four structures with unit-cell volumes of \(0.95V_0, 0.975V_0, 1.025V_0\) and \(1.05V_0\), while \(V_0\) is the volume for equilibrium structure. The weights for each energy in an energy/volume curve are held to a single value. The fitting database consists of all hcp energy differences and hcp energy versus volume data—but not the other energy versus volume data. We start with an equal initial weight and the only positive optimal weights are \(w_{\text{hcp-hex}} = 0.46\), \(w_{\text{hcp-hex}} = 0.54\) and \(w_{\text{hcp-bcc}} = 0.0\).

In figure 6, the predictions for hcp energy versus volume data improve significantly compared to the initial equal weight guess. Figure 7 shows that the optimal fitting database offers a close prediction of the shape of fcc energy versus volume curve, which is expected since the fcc and hcp structures have the same first nearest neighbor atoms. Bayesian errors of the four bcc energy differences are too large to have good predictions for either lattice constant or bulk modulus for bcc. Similarly, in figure 8, sloppy predictions for hex and \(\omega\) energy versus volume data are obtained from the optimal fitting database. The optimal weight set provides the predictions with
the least amount of relative error in the testing set, the best possible prediction for the testing set is still unacceptable indicating that the optimal Lennard-Jones potential with the given optimal fitting database is not transferable for testing set including energy versus volume data for bcc, hex, and $\alpha$, and cannot be made so without a different potential functional form.

5.5. Structures without DFT calculations in the testing set

Figure 9 presents the prediction distributions of the vacancy formation energy before and after optimization, where no related DFT calculation is in the testing set. We add the hcp vacancy configuration to the testing set, where the structure property is the vacancy formation energy. The testing set consists of all hcp energy differences, all six energy versus volume data and the single hcp vacancy configuration. The fitting database includes five hcp energy differences and hcp energy versus volume data. The non-zero optimal weight values are $w_{\text{hex-fcc}} = 0.46, w_{\text{hex-fcc}} = 0.54$ and $w_{\text{A15-fcc}} = 0.0$. We use $10^5$ independent samples to generate the distribution.

Figure 5. Three-structured fitting database prediction distribution for bcc + hex + A15 with fcc as reference structure. The first row of distributions are calculated with equal weights and the second row are calculated with the optimal weight set, $w_{\text{bcc-fcc}} = 0.46, w_{\text{hex-fcc}} = 0.54$ and $w_{\text{A15-fcc}} = 0$. We use $10^5$ independent samples to generate the distribution.

Figure 5. Three-structured fitting database prediction distribution for bcc + hex + A15 with fcc as reference structure. The first row of distributions are calculated with equal weights and the second row are calculated with the optimal weight set, $w_{\text{bcc-fcc}} = 0.46, w_{\text{hex-fcc}} = 0.54$ and $w_{\text{A15-fcc}} = 0$. We use $10^5$ independent samples to generate the distribution.
energy prediction. It suggests that the prediction will be accurate if the DFT calculation locates within the high likelihood parameter neighborhood of the empirical potential prediction.

6. Discussion and conclusion

We combine conventional potential parameter optimization methods and the Bayesian sampling technique to propose a new database optimization algorithm to build transferable empirical potential models. We choose an objective function as a function of prediction errors in the testing set and minimize the objective function to obtain optimal weights in the fitting database. We can mathematically determine adding and removing of structures to the fitting database, and automate the potential fitting process. Moreover, we are able to include structures
without DFT calculation in the testing set and optimize the Bayesian errors of these structures. The algorithm is demonstrated by a simple empirical potential model, Lennard-Jones potential fitting for Ti. We go through all possible combinations of two-structured and three-structured fitting databases and analyze the behavior of the objective function with respect to weight change. More complicated fitting databases and testing sets are taken into account. The new algorithm leads to the best possible empirical interatomic potential model based on the related fitting database and testing set.

Figure 8. Prediction for the hex and ω energy versus volume curve with fitting database including five hcp energy differences and hcp energy versus volume data, and testing set including five hcp energy differences and all six energy versus volume data. (a) $E_{\text{hex}}(V/V_0)$ and (b) $E_{\omega}(V/V_0)$ with the optimal weight set from figure 6.

Figure 9. Prediction for the hcp vacancy formation energy with a fitting database that includes five hcp energy differences and hcp energy versus volume data, and a testing set that includes all hcp energy differences, all six energy versus volume data and a hcp vacancy configuration. The top figure shows the prediction distribution calculated at equal weights in the fitting database. The bottom figure shows the prediction distribution calculated at optimal weights in the fitting database.
The database optimization algorithm is applicable for more complicated potential models. The mathematical definition is general for any potential functional form. One can easily calculate the Bayesian errors for more complicated potentials such as EAM or MEAM. The computational effort of the algorithm scales with the Monte Carlo steps required, which has the advantages dealing with high dimensional parameter space. We suggest that our algorithm could be used to build new EAM/MEAM potentials with designed transferability.

Acknowledgments

The authors thank Y Chen, H Wu, and M Fellinger for useful comments. This work was supported by the Office of Naval Research through ONR Award No. N000141210752.

Appendix A. The objective function and relative errors

The difference of the logarithm of the Bayesian errors for one structure property function prediction from two different fitting databases, $F_1$ and $F_2$ is

$$\ln(\epsilon_{\beta, F_1}^2) - \ln(\epsilon_{\beta, F_2}^2) = \ln \left( \frac{\langle \epsilon_{\beta, F_1}^2 \rangle_{F_1}}{\langle \epsilon_{\beta, F_2}^2 \rangle_{F_2}} \right) = \ln \left( 1 + \frac{\langle \epsilon_{\beta, F_1}^2 \rangle_{F_1} - \langle \epsilon_{\beta, F_2}^2 \rangle_{F_2}}{\langle \epsilon_{\beta, F_2}^2 \rangle_{F_2}} \right)$$

(A.1)

If $\langle \epsilon_{\beta, F_1}^2 \rangle_{F_1} - \langle \epsilon_{\beta, F_2}^2 \rangle_{F_2}$ is small, then $\langle \epsilon_{\beta, F_1}^2 \rangle_{F_1} \approx \frac{1}{2}(\langle \epsilon_{\beta, F_1}^2 \rangle_{F_1} + \langle \epsilon_{\beta, F_2}^2 \rangle_{F_2})$, and

$$\ln(\epsilon_{\beta, F_1}^2) - \ln(\epsilon_{\beta, F_2}^2) \approx \frac{\langle \epsilon_{\beta, F_1}^2 \rangle_{F_1} - \langle \epsilon_{\beta, F_2}^2 \rangle_{F_2}}{\langle \epsilon_{\beta, F_2}^2 \rangle_{F_2}} \approx \frac{2(\epsilon_{\beta, F_1}^2)_{F_1}}{\langle \epsilon_{\beta, F_1}^2 \rangle_{F_1} + \langle \epsilon_{\beta, F_2}^2 \rangle_{F_2}} = \frac{2(\epsilon_{\beta, F_2}^2)_{F_2}}{\langle \epsilon_{\beta, F_1}^2 \rangle_{F_1} + \langle \epsilon_{\beta, F_2}^2 \rangle_{F_2}}.$$  

(A.2)

Therefore the difference of the logarithm of the Bayesian errors is equivalent to the relative error difference. Minimizing the objective function with respect to weight in equation (8) is to minimize the relative errors in the testing set. We can also compare potential models using different fitting databases for the same testing set.

Appendix B. Reweighting of the sampling chain

The reweighting technique has been used in Monte Carlo simulations to avoid unnecessary resampling [39, 40]. When weights are changed in the database, we need to reevaluate the mean in equation (14) using equation (17). A change to the weights in the fitting database changes the likelihood function to $L(F^* | \theta)$. We rewrite $\langle A(\theta) \rangle_{F^*}$ as

$$\langle A(\theta) \rangle_{F^*} = \frac{\int P(\theta | F^*) A(\theta) P(\theta | F) d\theta}{\int P(\theta | F^*) P(\theta | F) d\theta} = \frac{\int L(F^* | \theta) \frac{A(\theta) P(\theta | F)}{L(F | \theta)} d\theta}{\int L(F^* | \theta) \frac{P(\theta | F)}{L(F | \theta)} d\theta}.$$ 

(B.1)

$$\approx \frac{\sum_{i=1}^{N} \frac{A(\theta_i) L(F^* | \theta_i)}{L(F | \theta_i)}}{\sum_{i=1}^{N} \frac{L(F^* | \theta_i)}{L(F | \theta_i)}}.$$ 

(B.2)

Thus a reweighting term is assigned to the original data, and provides new predictions without requiring a new sampling chain.
References

[1] Daw M S and Baskes M I 1984 Phys. Rev. B 29 6443–53
[2] Baskes M I 1992 Phys. Rev. B 46 2727–42
[3] Xu W and Moriarity J A 1996 Phys. Rev. B 54 6941–51
[4] Daw M S, Foiles S M and Baskes M I 1993 Mater. Sci. Rep. 9 251–310
[5] Wirth B, Odette G, Maroudas D and Lucas G 2000 J. Nucl. Mater. 276 33–40
[6] Lilleodden E, Zimmerman J, Foiles S and Nix W 2003 J. Mech. Phys. Solids 51 901–20
[7] Mishin Y, Mehl M J, Papaconstantopoulos D A, Voter A F and Kress J D 2001 Phys. Rev. B 63 224106
[8] Woodward C, Trinkle D R, Hector L G and Olmsted D L 2008 Phys. Rev. Lett. 100 045507
[9] Li Y, Siegel D J, Adams J B and Liu X Y 2003 Phys. Rev. B 67 125101
[10] Baskes M I, Nelson J S and Wright A F 1989 Phys. Rev. B 40 6085–100
[11] Baskes M I and Johnson R A 1994 Modelling Simul. Mater. Sci. Eng. 2 147
[12] Yu J, Simnott S B and Philpot S R 1994 Phys. Rev. B 81 125328
[13] Lenosky T J, Sabidgh B, Alonso E, Bulatov V V, de la Rubia T D, Kim J, Voter A F and Kress J D 2000 Modelling Simul. Mater. Sci. Eng. 8 825
[14] Hennig R G, Lenosky T J, Trinkle D R, Rudin S P and Wilkins J W 2008 Phys. Rev. B 78 054121
[15] Park H, Fellinger M R, Lenosky T J, Tipton W W, Trinkle D R, Rudin S P, Woodward C, Wilkins J W and Hennig R G 2012 Phys. Rev. B 85 214121
[16] Liu X Y, Adams J B, Ercolessi F and Moriarty J A 1996 Modelling Simul. Mater. Sci. Eng. 4 293
[17] Fellinger M R, Park H and Wilkins J W 2010 Phys. Rev. B 81 144119
[18] Sheng H W, Kramer M J, Cadien A, Fujita T and Chen M W 2011 Phys. Rev. B 83 134118
[19] Foiles S M, Baskes M I and Daw M S 1986 Phys. Rev. B 33 7983–91
[20] Zhou X et al. 2001 Acta Mater. 49 4005–15
[21] Baskes M I 1987 Phys. Rev. Lett. 59 2666–9
[22] Tersoff J 1986 Phys. Rev. Lett. 56 632–5
[23] Balamane H, Halicioglu T and Tiller W A 1992 Phys. Rev. B 46 2250–79
[24] Ercolessi F and Adams J B 1994 Europhys. Lett. 26 583
[25] Frederiksen S L, Jacobsen K W, Brown K S and Sethna J P 2004 Phys. Rev. Lett. 93 165501
[26] Hohenberg P and Kohn W 1964 Phys. Rev. B 33 864–71
[27] Kohn W and Sham L J 1965 Phys. Rev. A 140 1133–8
[28] Nocedal J and Wright S J 2006 Numerical Optimization (Springer Series in Operations Research and Financial Engineering) 2nd edn (New York: Springer)
[29] Bolstad W M 2007 Introduction to Bayesian Statistics 2nd edn (New York: Wiley)
[30] Hogg R V, McKean W J and Craig A T 2004 Introduction to Mathematical Statistics 5th edn (New York: Wiley)
[31] Givens G H and Hoeting J A 2005 Computationsal Statistics (New York: Wiley)
[32] Feng J, Hennig R G, Ashcroft N W and Hoffmann R 2008 Nature 451 445
[33] Kresse G and Hafner J 1993 Phys. Rev. B 47 558–61
[34] Kresse G and Furthmüller J 1996 Phys. Rev. B 54 11169–86
[35] Vanderbilt D 1990 Phys. Rev. B 41 7892–5
[36] Methfessel M and Paxton A T 1989 Phys. Rev. B 40 3616–21
[37] Hennig R G, Trinkle D R, Bouchet J, Srivivasan S G, Albers R C and Wilkins J W 2005 Nat. Mater. 4 129–33
[38] Ferrenberg A M and Swendsen R H 1988 Phys. Rev. Lett. 61 2635–8
[39] Ferrenberg A M and Swendsen R H 1989 Phys. Rev. Lett. 63 1195–8