Fast general two- and three-body interatomic potential

Sergey Pozdnyakov,[1] Artem R. Oganov,[1] Efim Mazhnik,[1] Arslan Mazitov,[2,3] and Ivan Kruglov[2,3]

[1]*Skolkovo Institute of Science and Technology Bolshoy Boulevard 30, bldg. 1, Moscow 121205, Russia*
[2]*Dukhov Research Institute of Automatics (VNIIA), Moscow 127055, Russia*
[3]*Moscow Institute of Physics and Technology, 9 Institutsky lane, Dolgoprudny 141700, Russia*

(Received 12 January 2023; accepted 15 March 2023; published 30 March 2023)

We introduce a new class of machine learning interatomic potentials—fast general two- and three-body potential (GTTP), which is as fast as conventional empirical potentials and require computational time that remains constant with increasing fitting flexibility. GTTP does not contain any assumptions about the functional form of two- and three-body interactions. These interactions can be modeled arbitrarily accurately, potentially by thousands of parameters not affecting resulting computational cost. Time complexity is O(1) per every considered pair or triple of atoms. The fitting procedure is reduced to simple linear regression on *ab initio* calculated energies and forces and leads to effective two- and three-body potential, reproducing quantum many-body interactions as accurately as possible. Our potential can be made continuously differentiable any number of times at the expense of increased computational time. We made a number of performance tests on one-, two- and three-component systems. The flexibility of the introduced approach makes the potential transferable in terms of size and type of atomic systems as long as they involve the same atomic species. We show that trained on randomly generated structures with just eight atoms in the unit cell, it significantly outperforms common empirical interatomic potentials in the study of large systems, such as grain boundaries in polycrystalline materials.

DOI: 10.1103/PhysRevB.107.125160

I. INTRODUCTION

In computational chemistry, the majority of calculations are performed within Born-Oppenheimer approximation [1], which states that the motion of atomic nuclei and electrons can be decoupled. Within this approximation, the potential energy of a system is completely defined by atomic positions, their types, and the total number of electrons in the system. Thus, the concept of potential energy surface (PES) is introduced as the functional dependence of the potential energy on the atomic positions. At each point, PES can be calculated by performing *ab initio* electronic structure calculations, where atomic positions are considered as the parameters of the electronic Hamiltonian. But such calculations are computationally very demanding, and simpler methods are typically used. One such method is density functional theory (DFT) [2,3], which significantly reduces the parameter space by introducing the charge density. Another example is the tight binding (TB) model [4], where the exact Hamiltonian is replaced by a parametrized matrix. Although these methods, especially DFT, remain quite accurate in many applications, they are still very computationally demanding, and thus it is hardly possible to use them for systems with more than several hundred atoms.

One possible way around this problem is to use conventional empirical interatomic potentials. In this approach, some fixed functional form with a few adjustable parameters is used for linking the potential energy and atomic positions. Such potentials are orders of magnitude faster, but their accuracy is limited, and for each type of compound, a different analytical form is needed. For example, different properties of metals are often modeled with the embedded atom method [5], modified embedded atom method [6], or angular-dependent potentials [7]. Organic compounds are usually simulated with AMBER, CHARMM, or other force fields (a good review can be found in Ref. [8]). Different chemical processes and reactions, polymerization, and isomerization can be studied with a reactive force field (ReaxFF) [9].

Another way is becoming increasingly popular nowadays—machine learning potentials. Regression problem is one of the standard problems of machine learning. Examples vary from the prediction of age by photo [10] to the prediction of the number of comments a blog post will receive based on its features [11]. The approximation of the PES can also be formulated as a regression problem, and the general scheme is the following: first, energies and forces are calculated by *ab initio* methods for some set of structures. Next, this dataset is used to fit some machine learning model, and after that, it can be used to efficiently and accurately predict energies and forces for new structures. A number of machine learning potentials were recently developed based on neural networks [12–20], Gaussian regression [21–23], linear regression [24–27], and other approaches [28–31].

Although conventional empirical potentials are the fastest, their accuracy is limited. Electronic structure calculations have the best accuracy, but they are computationally very demanding. Machine learning potentials represent a compromise between these two approaches.

In this paper, we report a general two- and three-body machine learning potential, which is as fast as conventional empirical potentials and, at the same time, is much more flexible.

The paper is structured as follows. In Sec. II we describe the methodology of the presented two- and three-body potential. Section III contains a theoretical comparison with the other interatomic potentials. The new class of atomic invariant descriptors is introduced in Sec. IV. Section V contains a generalization of the parametrization of the potential. In Sec. VI we report numerical experiments checking the effect of all hyperparameters, performance summary, computational cost, and extraction of chemically interpretable information from raw DFT calculations.

II. GENERAL TWO- AND THREE-BODY POTENTIAL

The real quantum interactions between the atoms in a chemical system can not be reduced to two- and three-body terms. But in most cases, the main contribution to the energy variance can be ascribed to two- and three-body interactions. So we decided to focus on them and construct a model, which is able to reproduce arbitrary two- and three-body interactions, at the same time being computationally efficient.

For the sake of simplicity, subsequent paragraphs contain a description of the potential for the case of a single atomic type. The generalization for multiple atomic species is described later.

In two- and three-body interactions approximation, the energy of the system (except additive constant) is given by

$$E = \sum_{i<j} E_2(\vec{r}_i, \vec{r}_j) + \sum_{i<j<k} E_3(\vec{r}_i, \vec{r}_j, \vec{r}_k), \tag{1}$$

where i, j, and k runs over all atoms in the system, $\vec{r}$ are the positions of corresponding atoms, E_2 and E_3 are the energies of pair and triple interactions, respectively.

A pair of atoms has one degree of freedom—the distance between them, while a triple of atoms has three degrees of freedom, which we decided to choose as three sides of the corresponding triangle. Thus, Eq. (1) can be rewritten as

$$\begin{aligned} E = &\sum_{i<j} \varphi_2(|\vec{r}_i - \vec{r}_j|) \\ &+ \sum_{i<j<k} \varphi_3(|\vec{r}_i - \vec{r}_j|, |\vec{r}_i - \vec{r}_k|, |\vec{r}_j - \vec{r}_k|), \end{aligned} \tag{2}$$

where φ_2 and φ_3 are one- and three-dimensional functions, which determine two- and three-body potentials. The summation in Eq. (2) scales as $O(N^3)$, where N is the number of atoms in the system, which is unacceptable. Thus, two cut-off radii R^2_{cut} and R^3_{cut} are introduced to discard long-range interactions. Now the summation in the first term is performed through only such pairs of atoms, where mutual distance is less than R^2_{cut}. Set of such pairs we will denote as $P(R^2_{\text{cut}})$. Summation in the second term we implemented in two variants—in the first one summation is performed over triples of atoms, where every side of the corresponding triangle does not exceed R^3_{cut}, and in the second over triples of atoms, where at least two sides do not exceed R^3_{cut}. Sets of proper triples we will denote as $T(R^3_{\text{cut}})$ for both variants. After such cutting, the complexity of the potential becomes the desired $O(N)$. The values of R^2_{cut} and R^3_{cut} represent the tradeoff between speed and accuracy. The higher R^2_{cut} and R^3_{cut}, the more accurate and slower the potential is. For different chemical systems, the best compromise between time and accuracy can be achieved with different variants of triples cutting. Thus, these two implemented ways to do it provide additional flexibility.

So, to determine the two- and three-body potential, one needs to determine functions φ_2 and φ_3 on finite domains. We decided to parametrize them in the form of piecewise polynomials on an equidistant grid. But the arbitrary coefficients for these polynomials are not suitable because the resulting PES approximation should obey certain continuity properties. For example, interatomic potentials are often used in molecular dynamics, where forces—derivatives of the energy with respect to atomic positions—are needed. Thus, PES approximation, and therefore functions φ_2 and φ_3 should be continuously differentiable. This means that one needs to impose additional stitching conditions on polynomial coefficients.

While the most prevalent demand for the potential is to be once continuously differentiable, sometimes a need for greater smoothness can arise. Our framework supports constructing arbitrarily many times continuously differentiable potentials.

Domain for the φ_2 is the interval from some $S_2 \geqslant 0$ to R^2_{cut}. It makes sense to choose $S_2 \neq 0$ because in all chemical systems there exists some minimal distance such that the probability of two atoms being closer is vanishingly small. In practice, after fitting the potential, we continue φ_2 from S_2 or even from some $C_2 > S_2$ to zero in accordance with the required smoothness in such a way that it tends to infinity at zero. This is needed to correctly handle such very rare situations as the ones in molecular dynamics when two atoms might come extremely close to each other. We use the equidistant grid containing $Q_2 + 1$ vertices, $Q_2 - 1$ inner vertices, and thus Q_2 intervals, which are enumerated from 0.

If constructed potential is required to be $k - 1$ times continuously differentiable, we use polynomials of order k and $\varphi_2(r)$ is given by

$$\varphi_2(r) = \sum_{l=0}^{k} a^l_p r^l, \tag{3}$$

where a^l_p is lth coefficient of the polynomial on the pth interval and $p = \lfloor Q_2 \frac{r - S_2}{R^2_{\text{cut}} - S_2} \rfloor$ is the index of the interval to which r belongs.

The values of polynomials and their $k - 1$ derivatives should match in all inner vertices. In addition, the value of the last polynomial and its $k - 1$ derivatives at R^2_{cut} should be equal to zero. Thus, arbitrary coefficients a^l_p are not suitable.

The way to ensure these stitching conditions is to use parametrization with cardinal B splines, which are a special case of B splines when the grid is equidistant. Cardinal B spline of kth order is the $k - 1$ times continuously differentiable (when $k > 1$) piecewise polynomial function of kth order on each interval, whose support consists of $k + 1$ equidistant intervals. Cardinal B splines of 0, 1, and 2nd order are shown in Fig. 1.

Cardinal B splines of arbitrary order can be calculated using the Cox-de Boor recursion formula [32,33].

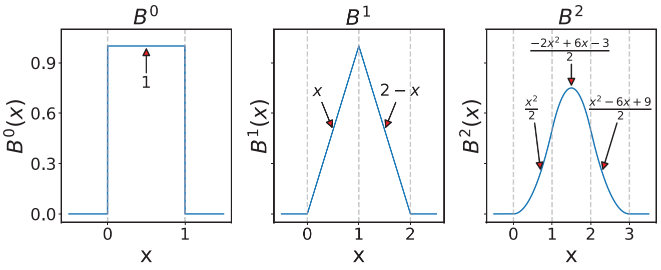

FIG. 1. Cardinal *B* splines of 0, 1, and 2nd order.

The new parametrization for $\varphi_2(r)$ is

$$\varphi_2(r) = \sum_{m=0}^{Q_2-1} c_m B_m^k(r), \tag{4}$$

where c_m are parametrization coefficients, $B_m^k(r)$ are cardinal B splines of order k and whose supports spread from $m-k$ to mth interval, see Fig. 2.

It is clear that any function in the form of Eq. (4) with arbitrary coefficients c_m is a piecewise polynomial and obeys necessary stitching conditions. Also, it can be shown [34] that any function in the form of Eq. (3), which obeys required stitching conditions, can be parametrized in the form of Eq. (4).

Hyperparameter k controls how many times φ_2 is continuously differentiable. But the greater this value, the higher the order of each polynomial and the higher are computational costs.

Now we will consider the three-dimensional φ_3 function, which determines three-body interactions. Its arguments are lengths of the sides of the triangle, which we denote as r_1, r_2, and r_3. The domain of φ_3 in the case of the first variant of triples cutting is the part of the cube $S_3 \leqslant r_1, r_2, r_3 \leqslant R_{\text{cut}}^3$, where r_1, r_2, and r_3 satisfy the triangle inequality.

Similarly to φ_2, we introduce an equidistant grid and put φ_3 to be polynomial on each elementary cube. Thus, φ_3 is given by

$$\varphi_3(r_1, r_2, r_3) = \sum_{l_1,l_2,l_3=0}^{k} b_{p_1,p_2,p_3}^{l_1,l_2,l_3} r_1^{l_1} r_2^{l_2} r_3^{l_3}, \tag{5}$$

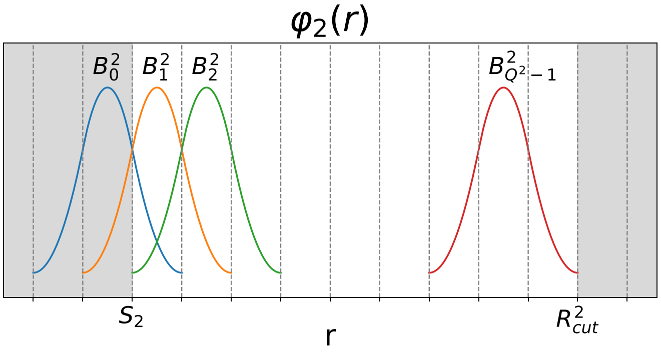

FIG. 2. Cardinal *B* splines parametrization.

where $b_{p_1,p_2,p_3}^{l_1,l_2,l_3}$ are coefficients of the three-dimensional polynomial placed in the elementary cube with indices p_1, p_2, p_3, $p_\alpha = \lfloor Q_3 \frac{r_\alpha - S_3}{R_{\text{cut}}^3 - S_3} \rfloor$, $\alpha = 1, 2, 3$.

As was stated earlier, arbitrary coefficients $b_{p_1,p_2,p_3}^{l_1,l_2,l_3}$ are not suitable, and thus three-dimensional cardinal B-splines parametrization is used. The three-dimensional cardinal B spline is given by

$$B_{m_1,m_2,m_3}^k(r_1, r_2, r_3) = B_{m_1}^k(r_1) B_{m_2}^k(r_2) B_{m_3}^k(r_3). \tag{6}$$

The φ_3 function should be symmetric with respect to permutations of the sides of the triangle. Thus symmetric combinations of three-dimensional cardinal B splines BS are used for the basis,

$$BS_{m_1,m_2,m_3}^k(r_1, r_2, r_3) = \sum_{\alpha_1,\alpha_2,\alpha_3} B_{\alpha_1,\alpha_2,\alpha_3}^k(r_1, r_2, r_3), \tag{7}$$

where the summation is taken through all permutations of m_1, m_2, m_3.

So possible parametrization for φ_3 can be given as

$$\varphi_3(r_1, r_2, r_3) = \sum_{0 \leqslant m_1 \leqslant m_2 \leqslant m_3 \leqslant Q_3 - 1} d_{m_1,m_2,m_3} BS_{m_1,m_2,m_3}^k(r_1, r_2, r_3). \tag{8}$$

This parametrization can be reduced because, due to triangle inequality, some terms in Eq. (8) will never affect the energy. Thus, the final parametrization is

$$\varphi_3(r_1, r_2, r_3) = \sum_{\{m_1,m_2,m_3\} \in Z} d_{m_1,m_2,m_3} BS_{m_1,m_2,m_3}^k(r_1, r_2, r_3), \tag{9}$$

where Z is defined as subset of $0 \leqslant m_1 \leqslant m_2 \leqslant m_3 \leqslant Q_3 - 1$, which contains only such $\{m_1, m_2, m_3\}$ that there exist such $\{r_1, r_2, r_3\}$ satisfying triangles inequality that $BS_{m_1,m_2,m_3}^k(r_1, r_2, r_3) \neq 0$.

In the case of the second variant of triples cutting, the domain for φ_3 is more complex, but still, the parametrization can be done in a similar manner.

So the fitting process of two- and three-body potential is reduced to determining the coefficients c_m and d_{m_1,m_2,m_3}. For this purpose, the functional dependence of the energy on these coefficients was investigated and turned out to be linear,

$$E = \sum_{i=0}^{Q_2-1} c_m D_m^2 + \sum_{\{m_1,m_2,m_3\} \in Z} d_{m_1,m_2,m_3} D_{m_1,m_2,m_3}^3, \tag{10}$$

where $D_m^2 = \sum_{\langle i,j \rangle \in P(R_{\text{cut}}^2)} B_m^k(|\vec{r}_i - \vec{r}_j|)$ and

$$D_{m_1,m_2,m_3}^3 = \sum_{\langle i,j,k \rangle \in T(R_{\text{cut}}^3)} BS_{m_1,m_2,m_3}^k(|\vec{r}_i - \vec{r}_j|, |\vec{r}_i - \vec{r}_k|, |\vec{r}_j - \vec{r}_k|)$$

Consequently, forces also depend linearly on the coefficients c_m and d_{m_1,m_2,m_3},

$$F_{q_\alpha} = -\frac{\partial E}{\partial r_{q_\alpha}} = \sum_{m=0}^{Q_2-1} c_m \left(-\frac{\partial D_m^2}{\partial r_{q_\alpha}} \right) + \sum_{\{m_1,m_2,m_3\} \in Z} d_{m_1,m_2,m_3} \left(-\frac{\partial D_{m_1,m_2,m_3}^3}{\partial r_{q_\alpha}} \right). \tag{11}$$

Thus, the fitting process is reduced to solving a linear regression problem, and the general scheme is the following:

For a given dataset, which contains structures and corresponding *ab initio* calculated energies and forces, we

(1) calculate values D^2_m, $D^3_{m_1,m_2,m_3}$, $\frac{\partial D^2_m}{\partial r_{q\alpha}}$, and $\frac{\partial D^3_{m_1,m_2,m_3}}{\partial r_{q\alpha}}$ for every structure,

(2) solve a joint linear regression problem, where input variables are values calculated at step 1, and target variables are energies and forces. The found coefficients of the linear model are c_m and d_{m_1,m_2,m_3}, and

(3) convert c_m and d_{m_1,m_2,m_3} to coefficients a^l_p and $b^{l_1,l_2,l_3}_{p_1,p_2,p_3}$.

After this the potential is ready since coefficients a^l_p and $b^{l_1,l_2,l_3}_{p_1,p_2,p_3}$ completely determine two- and three-body potential. In our implementation derivatives $\frac{\partial D^2_m}{\partial r_{q\alpha}}$ and $\frac{\partial D^3_{m_1,m_2,m_3}}{\partial r_{q\alpha}}$ are calculated analytically.

In the case of a multicomponent system, the energy is given by

$$E = \sum_{I\leqslant J} \sum_{\langle i,j\rangle \in P_{IJ}(R^2_{\rm cut})} \varphi_2^{I,J}(|\vec{r}_i - \vec{r}_j|) + \sum_{I\leqslant J\leqslant K} \sum_{\langle i,j,k\rangle \in T_{I,J,K}(R^3_{\rm cut})} \varphi_3^{I,J,K}(|\vec{r}_i - \vec{r}_j|, |\vec{r}_i - \vec{r}_k|, |\vec{r}_j - \vec{r}_k|), \tag{12}$$

where I, J, and K run through atomic species, $P_{IJ}(R^2_{\rm cut})$ are the sets of atomic pairs, where atoms have types I and J, $T_{I,J,K}(R^3_{\rm cut})$ are, analogously, sets of atomic triples, $\varphi_2^{I,J}$ and $\varphi_3^{I,J,K}$ are functions, which describe contributions to the energy from the pairs and triples with certain compositions.

If the total number of atomic species in the system is N_t, then the number of $\varphi_2^{I,J}$ and $\varphi_3^{I,J,K}$ functions is $\frac{N_t(N_t+1)}{2}$ and $\frac{N_t(N_t+1)(N_t+2)}{6}$, respectively. The parametrization for all these functions is the same as discussed earlier for the case of a one-component system with the only difference that the symmetry for the $\varphi_3^{I,J,K}$ is applied only through triangles sides, which are equivalent with taken into account atomic species. In other words, if all I, J, and K are the same, then the symmetry is applied through all three triangles' sides, and summation in Eq. (7) contains six terms, if two of I, J, and K are the same and the third is different, then the symmetry is applied only through two triangles sides and summation in Eq. (7) contains two terms, and if all I, J, and K are different, then symmetry is not applied, and summation in Eq. (7) contains one term, or, equivalently, initial three-dimensional cardinal B splines are used as basis functions. We denote the corresponding symmetric combinations as $BS^k_{IJK_{m_1,m_2,m_3}}$.

Also, for different symmetries, summation in Eq. (9) should be performed through different triples of indices, which we will denote as Z_{IJK}. Inequalities $m_\alpha \leqslant m_\beta$ should be satisfied only if r_α and r_β are equivalent in a triangle constructed from atoms with types I, J, and K; as earlier, triangles inequality cutting should be performed.

Eventually, the Eqs. (10) and (11) transform into

$$E = \sum_{I,J} \sum_{m=0}^{Q_2-1} c_{IJ_m} D^2_{IJ_m} + \sum_{I,J,K} \sum_{\{m_1,m_2,m_3\}\in Z_{IJK}} d_{IJK_{m_1,m_2,m_3}} D^3_{IJK_{m_1,m_2,m_3}} \tag{13}$$

and

$$F_{q_\alpha} = \sum_{I,J} \sum_{m=0}^{Q_2-1} c_{IJ_m} \left(-\frac{\partial D^2_{IJ_m}}{\partial r_{q_\alpha}} \right) + \sum_{I,J,K} \sum_{\{m_1,m_2,m_3\}\in Z_{IJK}} d_{IJK_{m_1,m_2,m_3}} \left(-\frac{\partial D^3_{IJK_{m_1,m_2,m_3}}}{\partial r_{q_\alpha}} \right), \tag{14}$$

where $D^2_{IJ_m} = \sum_{\langle i,j\rangle \in P_{IJ}(R^2_{\rm cut})} B^k_m(|\vec{r}_i - \vec{r}_j|)$ and $D^3_{IJK_{m_1,m_2,m_3}} =$

$$= \sum_{\langle i,j,k\rangle \in T_{IJK}(R^3_{\rm cut})} BS^k_{IJK_{m_1,m_2,m_3}}(|\vec{r}_i - \vec{r}_j|, |\vec{r}_i - \vec{r}_k|, |\vec{r}_j - \vec{r}_k|).$$

So single linear regression should be solved to simultaneously obtain all c_{IJ} and d_{IJK} coefficients and thus fit multicomponent two- and three-body potential.

It is a well-known fact that any continuous one-dimensional function can be approximated on the segment arbitrarily close in the form of Eq. (4) by reducing grid spacing or, which is the same, increasing Q_2 [35]. The same also applies to the three-body potential.

At the same time, complexity during the calculation of energies and forces does not depend on Q_2 and Q_3. Indeed, for every considered atomic pair or triple, the value of only one one- or three-dimensional polynomial of order k or its derivative should be calculated. Computational costs per single atomic pair or triple increase with hyperparameter k, but it only relates to desired smoothness of the potential and does not control the fitting flexibility. In practice, we use $k = 2$ for all potentials in this paper. In other words, the number of adjustable parameters does not affect computational time. It is especially beneficial in the case of multicomponent systems with a large number of atomic species where the number of these parameters can literally be thousands due to a large number of functions $\varphi_3^{I,J,K}$ and a large proportion of asymmetric or only partially symmetrical among them. In practice, the numbers of intervals of two- and three-body grids Q_2 and Q_3 are chosen long away in saturation area if the training dataset is big enough.

III. COMPARISON WITH OTHER INTERATOMIC POTENTIALS

The majority of existing conventional empirical potentials have a fixed functional form. Examples are Tersoff [36], Stillinger–Weber [37], and classical Lennard–Jones [38] potentials. These potentials have a fixed number of adjustable parameters, so their accuracy is limited. Sometimes, the resulting functional form is constructed from a set of one-dimensional functions parametrized by splines. Examples are Lenosky [39], and Zhang [40], where the three-body term in the modified embedded atom model (MEAM) is factorized as

$$\psi(r_1, r_2, \theta) = f_1(r_1)f_2(r_2)f_3(\theta), \tag{15}$$

where f_1, f_2, and f_3 are one-dimensional functions. This approach dramatically enriches the scope of functional forms it can parametrize, but it is clear that any three-dimensional function cannot be approximated arbitrarily close in the form of Eq. (15).

On the other hand, machine learning potentials are much more flexible, but their computational time increases with fitting flexibility. For neural networks, for instance, both expressivity and the number of multiplications in forward pass depend on the number of neurons, and thus, the larger capacity of the neural network comes at the cost of slower predictions. Some examples of the potentials based on the neural networks include deep potential molecular dynamics (DPMD) [41–43], Behler-Parrinello high-dimensional neural network potentials [12], recursively embedded atom neural networks (REANN) [44,45], higher order equivariant message passing neural networks (MACE) [46], strictly local equivariant deep learning interatomic potential (Allegro) [47], neural equivariant interatomic potentials (NequIP) [48], directional message passing neural network (DimeNet) [49], and geometric message passing neural network (GemNet) [50]. For kernel methods situation is the same. The functional form generated by such methods is given by

$$\text{prediction} = \sum_{q}^{N_{\text{samples}}} c_q K(\text{train sample}_q, \text{test sample}), \tag{16}$$

where the summation is over the whole training dataset or over the selected sparse points. Here again, the better fitting flexibility, which is determined by N_{samples}, comes at the cost of a more considerable computational cost. For the linear models, there is the same tradeoff. For instance ACE [51], MTP [24], and aPIP [52] express the energy as

$$\text{prediction} = \sum_{q}^{N_{\text{basis}}} c_q B_q(\{\vec{r}_i\}), \tag{17}$$

where $B_q(\{\vec{r}_i\})$ are the systematic basis functions of a collection of coordinates that describes the system. Here N_{basis} plays the same role as the N_{samples} for kernel methods.

The fundamental feature of our potential, which also can be cast to the form of Eq. (17) is that the domain of the functions $B_q(\{\vec{r}_i\})$, where they are not zero, is finite, and thus, it is not necessary to evaluate all of them given a single chemical configuration. Even more, the number of basis functions to be evaluated stays constant and does not depend on the total number of the N_{basis} basis functions used. The idea of finite support is also used in polynomial symmetry functions (PSF) [53] and in ultra Fast (UF) potentials [54]. In the case of PSF, it helps to significantly accelerate the computation of Behler-Parrinello symmetry functions [12], but later, on top of them, a neural network is applied, which shifts the overall computational cost from the conventional empirical potentials to the machine learning ones. Similar to GTTP, UF expresses two- and three-body potential in terms of the B-spline basis functions discussed above, but it lacks a number of important features. While we note that our potential can be cast to the form of Eq. (17) in order to highlight the ultimate source of computational efficiency, in practice, the computational scheme of GTTP is more efficient than that. Once potential is fitted, we never evaluate the B-spline basis functions. Instead, we explicitly convert the resulting functional form to the spline parametrization (the conversion of the coefficients c_m and d_{m_1,m_2,m_3} to a^l_p and $b^{l_1,l_2,l_3}_{p_1,p_2,p_3}$ mentioned in the previous section). Counting the required number of multiplications shows that this approach is way more efficient. For the two-body potential, in the case of the explicit evaluation of the basis functions, one needs to do $O(k^2)$ (where k is the order of B splines, practically we always use $k = 2$ for all the numerical experiments) multiplications per each pair of atoms. This number arises from the necessity to compute $k + 1$ basis functions, where each of them is given by a polynomial of order k. After the conversion to the spline parametrization, one needs to compute just one polynomial of order k, which costs only $O(k)$ multiplications. For the case of three-body potential, the difference is even more pronounced, $O(k^6)$ against $O(k^3)$ multiplications. On top of that, since we use symmetrization introduced in the Eq. (7) it is possible to compute only one three-dimensional polynomial for each "ordered" triplet of the atoms with the same specie in the system.

We should note that our functional form is limited to two- and three-body interactions, and thus, our potential is not a universal approximator, which is also called systematically improvable, in contrast to the methods [24,51,52] discussed above. Although, as it will be shown later, for many systems, the possibility to approximate just two- and three-body potential arbitrarily close is already enough to achieve good accuracy.

Thus, the presented potential is in the speed group of conventional empirical potentials and at the same time is flexible enough to approximate arbitrary two- and three-body interactions without any additional assumptions.

IV. NEW CLASS OF INVARIANT DESCRIPTORS

Usually, machine learning potentials are constructed in two steps. In the first step, a certain set of invariant descriptors is calculated, and in the second, it is fed to some machine learning algorithm. This is done because PES approximation should be invariant with respect to rotation, movement, reflection, and permutation of the identical atoms in the input structure. A good review of such descriptors is given in [23]. It is clear that descriptors $D^2_{IJ_m}$ and $D^3_{IJK_{m_1,m_2,m_3}}$ satisfy all the mentioned requirements along with smoothness with respect to atomic coordinates and, therefore, can be used along with arbitrary smooth machine learning algorithms (e.g., neural networks and kernel methods with smooth kernel generate smooth functions, whereas some machine learning methods—e.g., random forest—do not). Atomic versions of these descriptors are meant to describe local atomic neighborhoods and are defined as

$$\overset{\text{atomic}}{D^2_{I_m}} = \sum_{i \in \overset{\text{atomic}}{P_I}(R^2_{\text{cut}})} B^k_m(|\vec{r}_i - \vec{r}_{\text{central}}|) \tag{18}$$

TABLE I. Summary of aluminum datasets. N_s means the number of structures, N_a is the number of atoms in unit cell. In this particular case, it is identical for all structures within one dataset. F means scalar force components—projections on x, y, and z axes.

Notation	N_s	N_a	min E, $\frac{eV}{\text{Atom}}$	max E, $\frac{eV}{\text{Atom}}$	$\overline{E}$, $\frac{eV}{\text{Atom}}$	$\sqrt{\overline{(E-\overline{E})^2}}$, $\frac{eV}{\text{Atom}}$	$\sqrt{\overline{F^2}}$, $\frac{eV}{\text{Å}}$
Rand$_1$	20000	8	–3.75	54.84	4.06	7.15	22.87
Rand$_2$	7071	8	–3.75	–0.00	–2.15	1.15	5.00
Rand$_3$	2088	8	–3.75	–3.13	–3.48	0.19	0.69
MD	5000	108	–3.75	–3.69	–3.71	0.0034	0.35

and

$$\overset{\text{atomic}}{D^3_{IJ\ m_1,m_2,m_3}} = \sum_{\langle i,j\rangle\in \overset{\text{atomic}}{T_{IJ}}(R^3_{\text{cut}})} \overset{\text{atomic}}{BS^k_{IJ\ m_1,m_2,m_3}} \times (|\vec{r}_i-\vec{r}_j|, |\vec{r}_i-\vec{r}_{\text{central}}|, |\vec{r}_j-\vec{r}_{\text{central}}|), \tag{19}$$

where $\overset{\text{atomic}}{P_I}(R_{\text{cut}})$ is the set of neighbors with type I, $\overset{\text{atomic}}{T_{IJ}}(R^3_{\text{cut}})$ is, analogously, the set of pairs of neighbors with types I and J and $\overset{\text{atomic}}{BS^k_{IJ\ m_1,m_2,m_3}}$ are symmetric combinations of three-dimensional B splines where the central atom is considered to be inequivalent to any of its neighbors regardless of its type. We leave the analysis of these descriptors and the relationships between our descriptors and Behler–Parinello symmetry functions [13] to future work.

V. MORE GENERAL PARAMETRIZATION

In the case of $\varphi_2(r)$, when piecewise polynomial parametrization with polynomials of order k is used, there are $(k+1)Q_2$ initial degrees of freedom. If the potential should be $k-1$ times continuously differentiable, there are k stitching conditions in all inner vertices of the grid and in the right outer vertice, kQ_2 in total. So, there are $(k+1)Q_2 - kQ_2 = Q_2$ eventual degrees of freedom, which corresponds to the Q_2 coefficients in the cardinal B-splines parametrization [in the form of Eq. (4)]. But one can let the polynomials be of order k and require the potential to be only $k_d - 1$ times continuously differentiable, where $k_d < k$. In this case, there are $(k+1-k_d)Q_2$ eventual degrees of freedom. The corresponding cardinal B-splines parametrization is given by

$$\varphi_2(r) = \sum_{m=0}^{Q_2-1}\sum_{f=k_d}^{k} c_{f,m}B^f_m(r). \tag{20}$$

In the case of φ_3 three-dimensional cardinal B splines of not uniform order are defined as

$$B^{f_1,f_2,f_3}_{m_1,m_2,m_3}(r_1,r_2,r_3) = B^{f_1}_{m_1}(r_1)B^{f_2}_{m_2}(r_2)B^{f_3}_{m_3}(r_3). \tag{21}$$

The definition of the symmetric combinations $BS^{f_1,f_2,f_3}_{m_1,m_2,m_3}$ is analogous to the Eq. 7, where in summation f_1, f_2, and f_3 are also rearranged along with m_1, m_2, and m_3. All subsequent steps including the definition of atomic invariant descriptors $\overset{\text{atomic}}{D^2_{I_{f,m}}}$ and $\overset{\text{atomic}}{D^3_{IJ\ f_1,f_2,f_3,m_1,m_2,m_3}}$ are the same as before.

When the training dataset is large enough, there is no need to use $k > k_d$. Indeed, one can just put $k = k_d$, not affecting the smoothness of the potential, and increase Q_2 and Q_3 to ensure the same fitting flexibility. After this procedure, the smoothness and accuracy of the potential will be the same as before, and computational time will be lower since polynomials of lower order will have to be calculated.

But when the training dataset is not big enough, the use of $k > k_d$ may increase the accuracy of the potential since parametrization in the form of Eq. (20) along with lower Q_2 and Q_3 or bigger grid intervals may have better generalization capability.

VI. RESULTS

A. Aluminum

Aluminum is an example of a system where two- and three-body interaction approximation works well. To illustrate the performance of our potential, we applied it to four datasets. The first one contains 5000 steps of *ab initio* molecular dynamics simulation in the canonical (NVT) ensemble of aluminum with 108 atoms in the unit cell at 300 K and with volume $16.7\frac{\text{Å}^3}{\text{atom}}$. The second dataset consists of 20 000 random structures produced by a symmetric random structure generator from evolutionary algorithm USPEX [55–57], each with eight atoms, third is a subset of the second one and contains 7071 structures with negative energies and fourth is a subset of the third one and contains 2088 structures with energies less than $-3.13\ \frac{eV}{\text{atom}}$. The overview of these datasets is given in Table I. All *ab initio* calculations of energies and forces were performed using Vienna *Ab initio* Simulation Package (VASP) [58–60]. Projector-augmented wave (PAW) [61] method was used to describe core electrons and their interaction with valence electrons. The plane wave kinetic energy cutoff was set at 500 eV and Γ-centered k points with a resolution of $2\pi \times 0.05\text{Å}^{-1}$ were used.

The following several subsubsections contain an analysis of the hyperparameters of the developed potential. For the sake of brevity, thereinafter, we will understand forces as force components—projections on the x, y, and z axes. Error in energies per atom is a rather unphysical quantity since the total error per unit cell does not necessarily grow proportionally to the number of atoms in it. So, we decided to give all errors in energies per unit cell. Relative errors are calculated as the ratio of the absolute errors to the standard deviations of the corresponding values. All errors are given on the test samples and were obtained either by cross-validation or by explicit partitioning into train and test sets.

1. Relative importance

When solving the linear regression problem, the following minimization problem arises:

$$\min_{c,d} \frac{1}{\lambda}\left(\sum_i c_i^2 + \sum_i d_i^2\right) + \frac{W_E}{N_E}\sum_i (E_{\text{ab initio}_i} - E_{\text{predicted}_i}(c,d))^2 + \frac{W_F}{N_F}\sum_i (F_{\text{ab initio}_i} - F_{\text{predicted}_i}(c,d))^2, \tag{22}$$

where W_E and W_F are weights for the energies and forces, N_E and N_F are numbers of energies and forces in the dataset. λ is the usual L_2 regularization hyperparameter, which can be selected using standard techniques [62,63], while the influence of $Im = W_E/W_F$—relative importance of energies, should be investigated manually.

First of all, we investigated it on Rand$_2$ dataset. The other hyperparameters of the potential were put as $S_2 = S_3 = 1.0$ Å, $R^2_{\text{cut}} = 10.0$ Å, $R^3_{\text{cut}} = 5.0$ Å, $Q_2 = 27$, $Q_3 = 8$, $k = 2$, first variant of triples cutting. For each value of Im we measured RMS error in energies and forces. All errors were evaluated by 20-fold cross validation with random partitions. Results are shown in Fig. 3. It is very natural that the higher the value of Im, or, in other words, the higher priority the energies are given, the lower the error in energies and vice versa. But there is also another effect. The thing is that the number of energies in the dataset is much less than the number of forces. Indeed, structure, which contains N_a atoms, contributes one energy and $3N_a$ forces to the dataset. Thus, energies alone typically do not provide enough data to fit the potential, and training only on energies leads to overfitting. When the value of Im is very large, the potential is actually trained only on energies. So, one can expect that decreasing Im or taking into account the forces during the fitting can reduce the test error in energies. Figures 3(a), 3(d), 3(e), and 3(f) illustrate the dependence of test error in energies on the Im for different datasets and different potentials. In accordance with the reasons discussed earlier, all these dependencies consist of two plateaus and a well between them. The relative position of the plateaus and the size of the well depend on the interrelation between dataset size and the number of parameters in the potential.

Figure 3(b) illustrates the errors in forces. We observe qualitatively similar behavior in all studied cases.

Since we assume that the errors in energies and forces are equally important, we decided to choose the value of Im to minimize the product of these errors, which is plotted in Fig. 3(c).

2. Two-body hyperparameters

R^2_{cut} and R^3_{cut} represent the tradeoff between the accuracy and computational time. The higher R^2_{cut}, the more accurate the potential, but also slower. We measured the behavior of the error in energies for only two-body potential at various R^2_{cut} and different grid densities, namely 2, 4, 6, 8, and 10 intervals/Å, on the Rand$_2$ dataset. Results are shown in Fig. 4.

As can be seen from this plot, the RMS error converges to some nonzero limit, which is the limit of the accuracy of the two-body approximation.

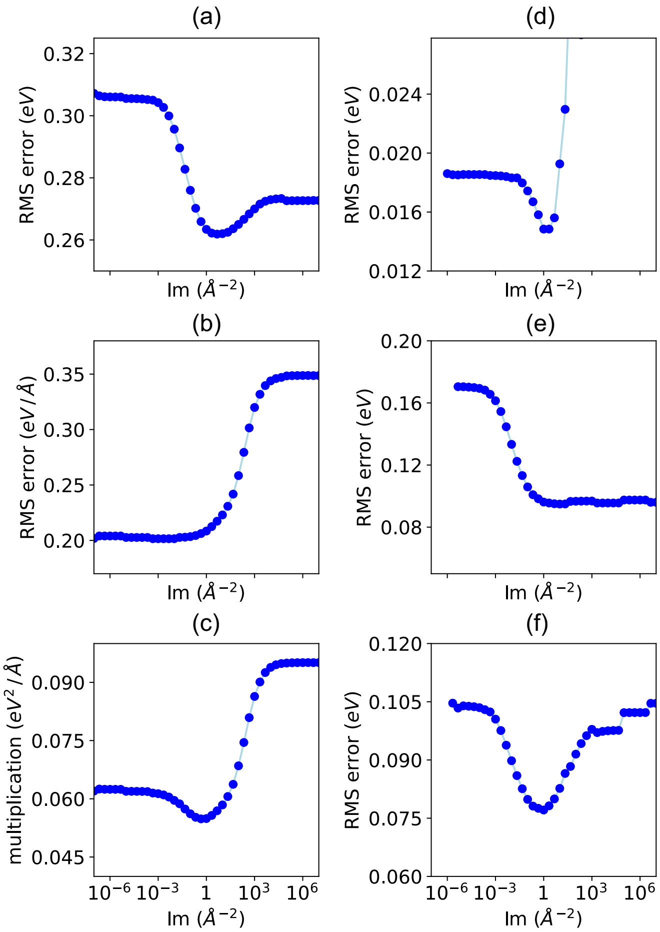

FIG. 3. influence of the relative importance of energies, Im hyperparameter. Panels [(a)–(c)] are related to Rand$_2$ dataset and illustrate cross-validation RMS errors in energies, forces, and their product respectively. Panels [(d)–(f)] illustrate errors in energies. (d) corresponds to the potential trained on one-tenth of the MD dataset, (e) and (f) to the potentials with a small and large number of parameters, respectively, trained on Rand$_3$ dataset.

For later calculations we have chosen $R^2_{\text{cut}} = 8$ Å and Q_2 corresponding to the grid density of 6 intervals/Å as hyperparameters at which the error almost completely converged.

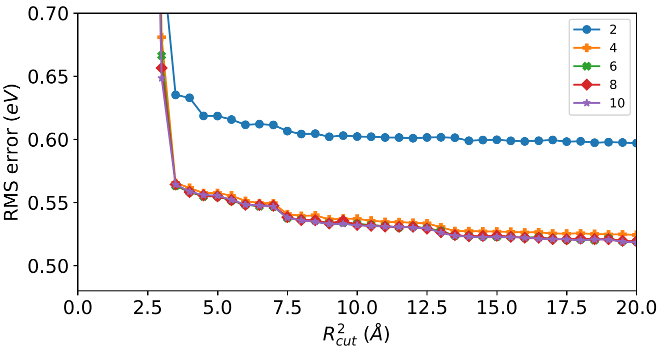

FIG. 4. Cross-validation RMS errors in energies for only two-body potential.

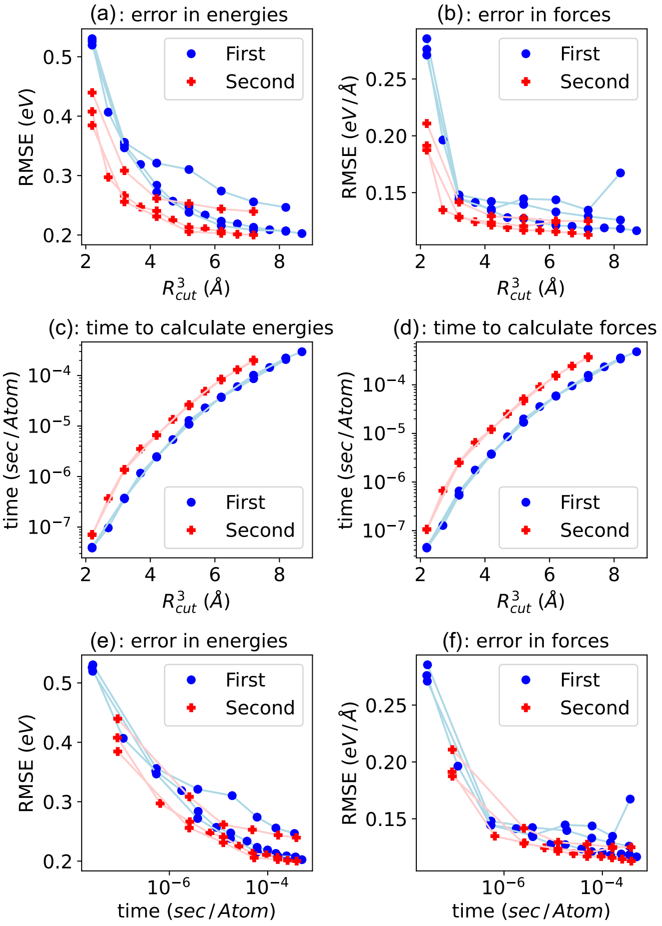

FIG. 5. Influence of three-body hyperparameters. All panels contain lines for several grid densities, namely 1, 2, and 3 intervals/Å, and first and second variants of triples cutting. Subplots (a) and (b) illustrate errors in energies and forces for different R^3_{cut}, (c) and (d) show computational time for energies and forces. Subplots (e) and (f) present a tradeoff between computational time and errors. Time on the horizontal axis corresponds to the simultaneous calculation of both energies and forces. All measurements were taken on one core of Intel Xeon CPU E5-2667 v4 for only three-body part, not including the construction of atomic neighborhoods. Times were averaged over a set of structures from the Rand$_2$ dataset. All standard errors of the mean do not exceed the size of the symbols.

3. *Three-body hyperparameters*

Now we fix hyperparameters of two-body potential found previously and measure the performance of two- and three-body potential with different three-body hyperparameters. As earlier, we performed calculations for various R^3_{cut} and several grid densities.

Figures 5(a) and 5(b) illustrate the behavior of errors in energies and forces with increasing R^3_{cut}. As expected at the same R^3_{cut}, the error is lower with the second variant of triples cutting because at the same R^3_{cut} the set of considered triples with the first variant of triples cutting is a subset of triples included with the second variant of triples cutting. But, for the same reason, the computational time with the second variant of triples cutting is higher at the same R^3_{cut}, as illustrated in Figs. 5(c) and 5(d). These figures also show that computational time indeed does not depend on Q_2 and Q_3 (lines for different densities almost coincide), and thus on fitting flexibility.

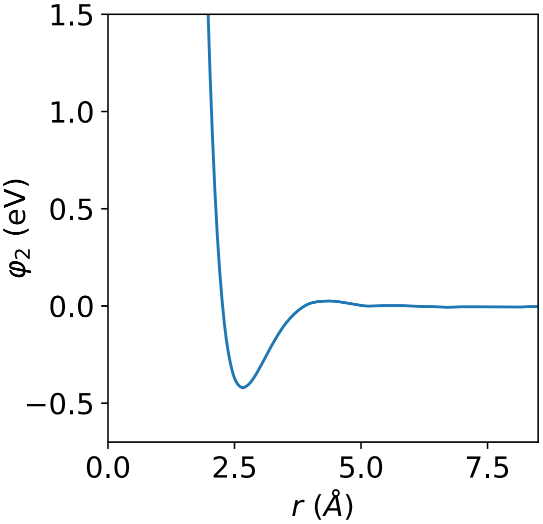

FIG. 6. Two-body potential trained on Rand$_2$ dataset for Al.

Figures 5(e) and 5(f) illustrate the tradeoff between accuracy and computational time. It appears that for this particular chemical system, the the second variant of triples cutting is slightly better.

We consider the $R^3_{\mathrm{cut}} = 5.2$ Å with second variant of triples cutting as sufficient. Q_3 was chosen to correspond grid density equal to 3 intervals/Å.

The resulting two-body potential is shown in Fig. 6. It has a very reasonable form resembling the one of Lennard-Jones potential even though our parametrization does not incorporate any physical assumptions and can approximate any function arbitrarily close. We show the restored three-body potential in the Supplemental Material [64].

We independently calculated two- and three-body contributions to the energy, and it appeared that the three-body part is an order of magnitude smaller. Namely, standard deviations of two- and three-body components on the Rand$_2$ dataset appeared to be 9.47 eV and 1.30 eV, respectively.

The presented analysis of the impact of R^2_{cut} and R^3_{cut} on accuracy and computational time of the potential provides guidance on how to choose the optimal values for these hyperparameters depending on the particular application of the potential. In general, for a new system, a similar analysis should be performed for the optimal selection of R^2_{cut} and R^3_{cut}. Alternatively, one can fit an extensive number of potentials for all possible selections of two- and three-body cutoff radiuses and do a Pareto front in terms of error and computational time, as we discuss in Sec. VI A 5.

4. *Performance summary*

Performance of the potential on the Rand$_2$ dataset is illustrated in Fig. 7.

For the other datasets, optimal hyperparameters of the potential were chosen in a similar manner, and they do not differ much.

TABLE II. Performance of GTTP for Al on energies. Absolute RMS errors are given in *meV* per unit cell(8 atoms/cell in case of Rand$_\alpha$ and 108 atoms/cell in case of MD). Relative errors are calculated as the ratio of the absolute error to the standard deviation. Rand$_\alpha$–MD cells illustrates errors after additive constant adjusting. See Fig. 8 and discussion in the text.

test on / train on	MD		Rand$_3$	Rand$_2$	Rand$_1$
MD	7.4, 2.01%	9.3, 2.55%	×	×	×
Rand$_3$	42.9, 11.74%		60.6, 3.89%	×	×
Rand$_2$	119.1, 32.61%		71.2, 4.56%	208.0, 2.27%	×
Rand$_1$	111.7, 30.57%		202.6, 12.99%	393.8, 4.3%	990.2, 1.73%

The numerical overview is given in Tables II and III. Dataset Rand$_3$ is a subset of Rand$_2$, which in turn is a subset of Rand$_1$. In Rand$_\alpha$–Rand$_\beta$ cells all energies and forces are predicted in a cross-validation cycle for Rand$_\alpha$ dataset with hyperparameters of the potential selected for Rand$_\alpha$, and later the error is measured only on values, which belong to Rand$_\beta$.

Rand$_\alpha$-MD cells illustrate the errors on MD of the potentials trained on Rand$_\alpha$. In the case of energies, these cells illustrate the errors after additive constant adjusting. Indeed, initially, there is a constant systematic error, see Fig. 8. It originates from both discrepancy between *ab initio* calculations and intrinsic error of the potential. In the case of different datasets, namely Rand and MD, *ab initio* calculations were performed with different parameters, which led to different ground state energies in both cases. Also, the potential itself predicts the ground-state energy is not absolutely correct. While contributing a relatively small part to the Rand$_\alpha$–Rand$_\beta$ errors, this makes a noticeable contribution in the case of Rand$_\alpha$–MD because the variability in the Rand$_\alpha$ datasets is much greater than in the MD, see Table I.

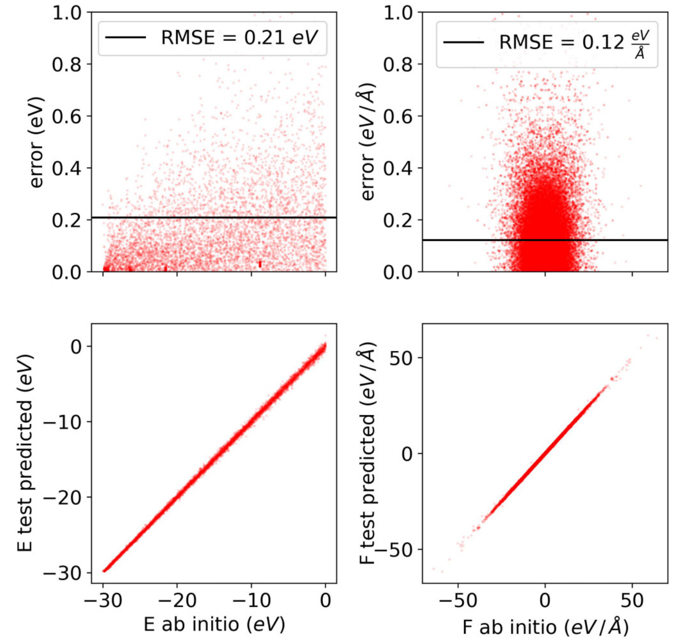

FIG. 7. Performance of the GTTP on the Rand$_2$ dataset. Energies and forces are predicted in the cross-validation cycle on the test samples.

The left subcell of MD–MD in Table II illustrates the "interpolation" error when the error is measured in a cross-validation cycle with random partitions, while the right subcell illustrates the "extrapolation" error when potential is trained on the first third of the timeline of molecular dynamics and tested on the last.

Thus, all errors presented in Tables II and III are measured on test samples.

Generally, the absolute error significantly depends on the variability in the dataset. The smaller part of phase volume is covered by the potential—the smaller is the absolute error and vice versa. Tables II and III also illustrate good transferability of the potential—being fitted to the beginning of the molecular dynamics trajectory, it can accurately describe system states from the last MD steps. In addition, it can, with satisfactory accuracy, predict energies and forces for structures with 108 atoms, being fitted to only structures with 8 atoms. Taking into account that the computational cost of acceptable accurate *ab initio* calculations scales cubically with system size, this property is especially useful. The performance on the MD dataset of the potential trained on Rand$_3$ is shown in Fig. 8.

5. *Computational time*

The hyperparameters of the potentials in previous sections were chosen far in saturation area, while it is possible to take smaller R^2_{cut} and R^3_{cut} to significantly reduce computational time and only slightly affect the accuracy. In order to investigate the tradeoff between time and accuracy, we fitted a number of potentials with different two- and three-body

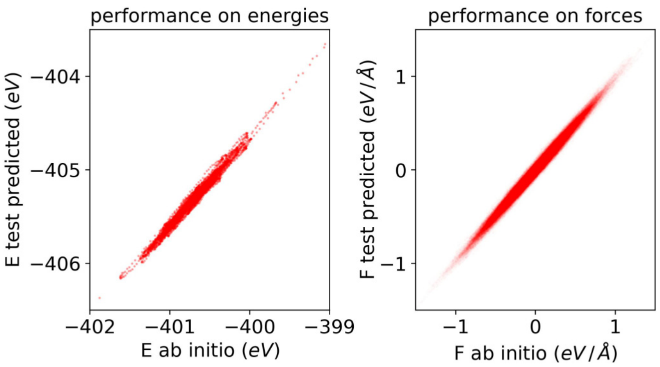

FIG. 8. Performance on the MD dataset of the potential trained on Rand$_3$ before additive constant adjusting. Note the discrepancy between vertical and horizontal axes in the energy graph, as discussed in the text.

TABLE III. Performance of GTTP for Al on forces. Absolute RMS errors are given in $\frac{\text{meV}}{\text{Å}}$. Relative errors are calculated as the ratio of the absolute error to the standard deviation.

test on train on	MD		Rand$_3$	Rand$_2$	Rand$_1$
MD	12.0, 3.47%	12.3, 3.55%	×	×	×
Rand$_3$	41.2, 11.85%		27.7, 4.0%	×	×
Rand$_2$	86.6, 24.94%		34.6, 5.01%	121.8, 2.44%	×
Rand$_1$	75.1, 21.63%		58.9, 8.52%	157.6, 3.15%	625.6, 2.74%

hyperparameters on the Rand$_2$ dataset. After that, we constructed the two-objective Pareto front, the first objective being computational time and the second one being the product of errors in energies and forces. To estimate errors, we used explicit partitioning into the train and test dataset with 80% of the structures in the training dataset. Times were measured within LAMMPS Molecular Dynamics Simulator [65] to simultaneously calculate energies, forces, and stress tensors, including constructing atomic neighborhoods on one core of Intel Xeon CPU E5-2667 v4. Also, we compared the Pareto front of our (GTTP) potential with the Pareto front of the moment tensor potential (MTP) [66]. The method of measuring time was the same in both cases. The result is shown in Fig. 9.

MTP is one of the fastest machine learning potentials. Namely, it was shown [24] that on the same dataset with the same accuracy, MTP is approximately 170 times faster than the Gaussian approximation potential (GAP) [21]. This was also confirmed in a recent study [67], where a comprehensive comparison of several machine learning potentials was performed. In spite of this, our potential convincingly outperforms MTP in the fast area. With increasing the computational time, the error of the GTTP converges to a nonzero limit, which is caused by the error of the two- and three-body interactions approximation itself. When this happens, the error of the systematically improvable MTP becomes lower. In case of forces the convergence is reached already at $10^{-6}\frac{\text{sec}}{\text{atom}}$, whereas in the case of energies it is reached at $10^{-6} - 10^{-5}\frac{\text{sec}}{\text{atom}}$. In the Supplemental Material [64], we provide a table with more detailed information about some potentials from the Pareto front.

When there are more than one atomic species, the potential energy surface is more complex, and, therefore, more parameters are required. Particularly, in GTTP, the number of parameters grows cubically with the number of atomic species. But at the same time, the computational cost of our potential does "not" increase with the number of parameters or with the number of atomic species. This is not the case for the majority of machine learning potentials and of MTP in particular, so one can expect that the relative performance of our potential will be even better on multicomponent systems.

B. Tungsten

The intrinsic flexibility of the potential makes it transferable in terms of the types and sizes of atomic systems. The example of tungsten demonstrates the good performance of GTTP for such huge systems as grain boundaries (GBs), which are among the most challenging subjects of computational chemistry [68]. For creating the potential, only the knowledge of randomly generated crystalline

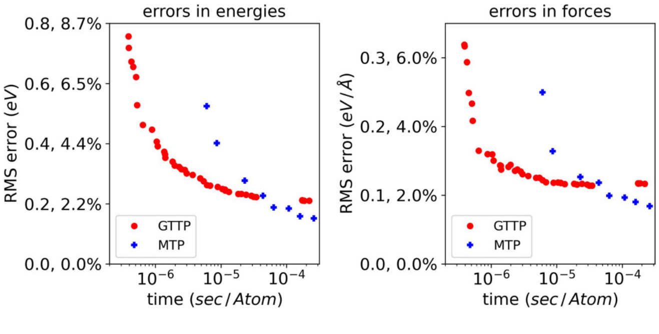

FIG. 9. Accuracy–computational time tradeoff. Times were averaged over a set of structures from the Rand$_2$ dataset. All standard errors of the mean do not exceed the size of the points on the plot. The computational cost of GTTP is an exact constant per each considered pair or triplet of atoms, not depending on the number the parameters used to approximate two- and three-body potentials. The computational cost grows because of the selection of bigger cutoff radiuses (independently for two- and three-body parts). The bigger cutoff radius simultaneously leads to better accuracy and a higher number of pairs/triplets of atoms to process.

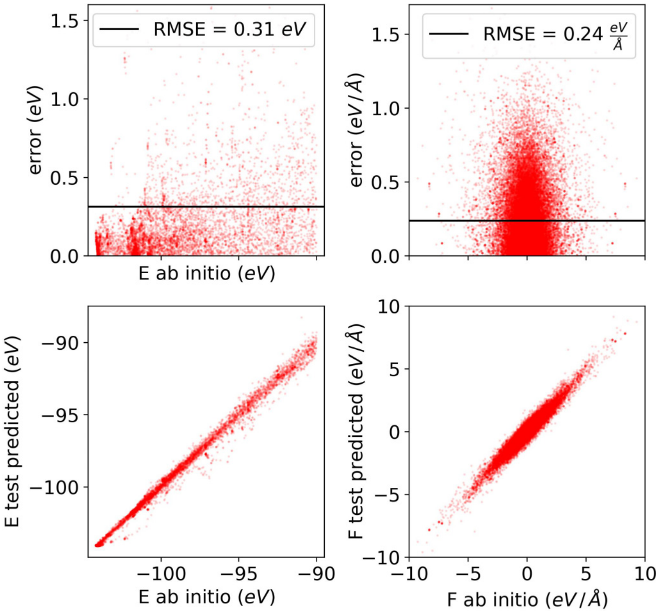

FIG. 10. Performance of GTTP for tungsten.

TABLE IV. Comparison of energies of $\Sigma 27(5\bar{5}2)[110]$ symmetric tilt GBs with EAM1, EAM2, GTTP potentials, and DFT. Within EAM potentials GB14 structure is unstable. Atomic density [n] is indicated in the second column. Root-mean-square error (RMSE) with respect to DFT was used as a quality metric of the algorithms. All data are given in J×m^{-2} units.

Label	**[n]**	**EAM1**	**EAM2**	**GTTP**	**DFT**
GB1	1/2	2,819		2,555	2,592
GB2	1/2	2,811		2,556	2,593
GB3	1/2	2,818		2,605	2,594
GB4	1/2	2,807		2,606	2,595
GB5	1/2	2,817		2,556	2,609
GB6	1/2	2,802		2,555	2,610
GB7	1/2	2,798		2,555	2,624
GB8	1/2	2,796		2,555	2,626
GB9	1/2	2,812		2,559	2,628
GB10	0	3,171		2,850	2,960
GB11	1/2		2,493	2.605	2,590
GB12	0		2,495	2,947	2,951
GB13	0		2,670	2,851	2,973
GB14	0			2,584	2,680
RMSE		0.203	0.321	0.065	

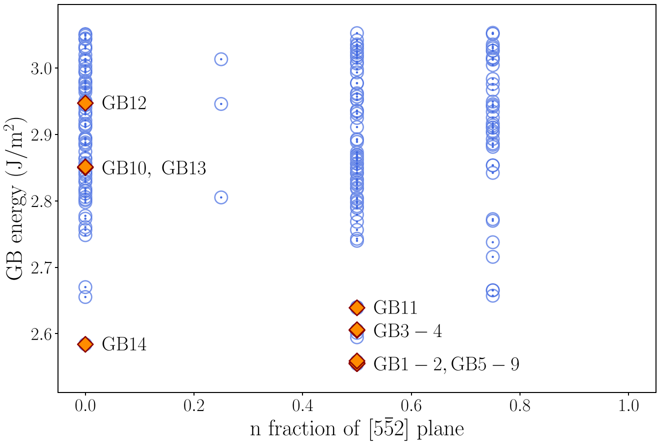

FIG. 12. Results of the evolutionary search with GTTP. The GB energy is plotted as a function of atomic density [n]. GB1–GB14 structures from Ref. [71] are marked with orange diamonds.

configurations of tungsten was used: the dataset consisted of 7286 structures with 8 atoms in the unit cell, and their energies varied from $-13.02\frac{eV}{\text{atom}}$ to $-11.25\frac{eV}{\text{atom}}$ with mean of $-12.35\frac{eV}{\text{atom}}$ and standard deviation of $0.48\frac{eV}{\text{atom}}$. The standard deviation of force components was $0.89\frac{eV}{\text{Å}}$. Values of R^2_{cut} and R^3_{cut} were set to 10.0 Å and 6.0 Å, respectively, and the first variant of triples pruning was chosen. The test errors of GTTP in energies and forces were $0.33eV$ (per unit cell) or 8.5% and $0.26\frac{eV}{\text{Å}}$ or 29%, which is illustrated in Fig. 10.

In order to test the performance of the constructed potential on GBs, we compared the results of grain boundaries structure prediction made using the USPEX code. In this paper, a family of $\Sigma 27(5\bar{5}2)[110]$ symmetric tilt GBs of tungsten with different atomic densities were predicted. The structures were subsequently relaxed using the LAMMPS code [65], employing EAM1 [69] and EAM2 [70] potentials. In order to verify their stability, *ab initio* calculations were performed. We used the same initial structures for the calculation with GTTP potential. The results of these three approaches are summarized in Table IV. The ground state of the $\Sigma 27(5\bar{5}2)[110]$ GB is demonstrated in Fig. 11.

Despite the good agreement between GTTP and DFT results, both these methods operated with the structures, which were previously generated by USPEX and relaxed by EAM potentials. Therefore, we performed the same evolutionary search but used our GTTP for structure relaxation. Figure 12 demonstrates the results of the search. Obtained GBs and their energies are marked by blue circles, while orange diamonds correspond to the most stable GBs predicted within EAM potentials [71]. The energy is plotted as a function of atomic density [n].

Thus, all the structures from Table IV were found by evolutionary search with GTTP. Comparison of the energy values shows that GTTP practically removes ambiguity in the ground state representation, which plagued EAM potentials, and provides 3–5 times better accuracy. It is worth noting that the metastable GB14 structure (Fig. 13) with [n] = 0, which was previously discovered in Ref. [72], was found by evolutionary search, while both EAM potentials treated it as an unstable one.

C. Performance on two- and three-component systems

To test our potential on multicomponent systems we applied it to titanium hydride and Li-intercalated anatase TiO_2. The titanium hydride dataset contained 17 335 steps of *ab initio* molecular dynamics trajectory with 108 titanium and 189 hydrogen atoms in the unit cell. This was taken from our recent study [73]. The force component standard

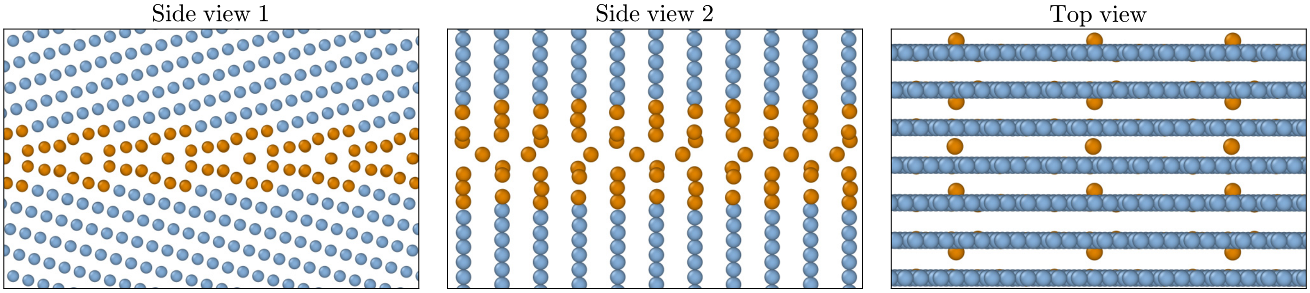

FIG. 11. Ground state of the $\Sigma 27[5\bar{5}2](110)$ GB from evolutionary search with GTTP, $\gamma_{\text{GB}} = 2.55$ J×m^{-2}.

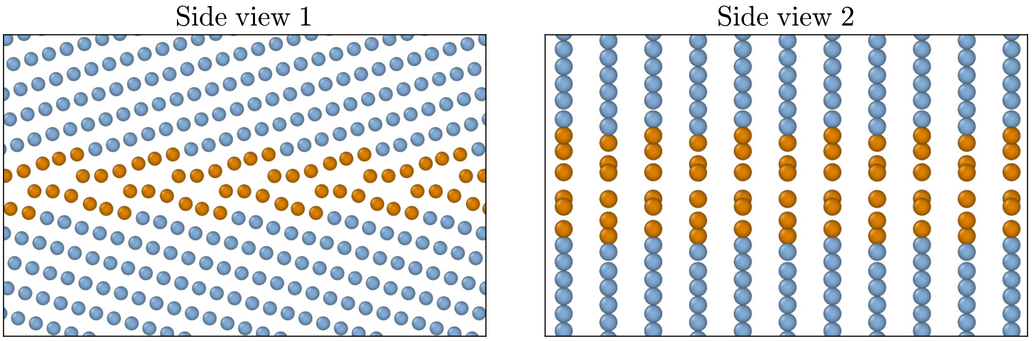

FIG. 13. Metastable GB14 structure from [72], and also predicted in this work with USPEX and GTTP, $\gamma_{\rm GB} = 2.58\ {\rm J}\times{\rm m}^{-2}$.

deviation is 0.92 $\frac{eV}{\text{Å}}$. We trained our potential on the first third of the molecular dynamics trajectory and tested on the last. We choose $R^2_{\rm cut} = 10.0$ Å, $R^3_{\rm cut} = 4.34$ Å and second variant of triples cutting. The error turned out to be 0.070 $\frac{eV}{\text{Å}}$ or 7.6%, which is illustrated in Fig. 14.

In case of Li-intercalated anatase $\rm TiO_2$ we used three datasets with random structures—$\rm Li_\mathit{x}TiO_2(1)$ [74], $\rm Li_\mathit{x}TiO_2(2)$, and $\rm Li_\mathit{x}TiO_2(3)$. Datasets $\rm Li_\mathit{x}TiO_2(2)$ and $\rm Li_\mathit{x}TiO_2(3)$ were generated by applying some mutations to the structures from the $\rm Li_\mathit{x}TiO_2(1)$ dataset. All datasets contain structures with 16 titanium and 32 oxygen atoms. The number of lithium atoms varied from 1 to 14 in $\rm Li_\mathit{x}TiO_2(1)$ and $\rm Li_\mathit{x}TiO_2(3)$, and was equal to 14 in $\rm Li_\mathit{x}TiO_2(2)$. Chosen hyperparameters of the potential are $R^2_{\rm cut} = 10.9$ Å, $R^3_{\rm cut} = 4.7$ Å. The numerical results of the performance of the potential are given in Table V and illustrated in Fig. 15.

The absolute error grows with the increase of standard deviations of force components or with the coverage of phase volume. But at the same time, the relative error decreases. We already faced this behavior for aluminum in Sec. VI A 4. The same situation was also observed in [75].

D. Chemical insights from raw data

Besides other advantages, our approach enables the extraction of interpretable information from large amounts of raw *ab initio* calculations, and further, we will consider carbon as an example. In order to fit the potential, we used a dataset containing 8353 random crystal structures, each with 8 atoms in the unit cell. The energy varied from $-8.9\ \frac{eV}{\rm atom}$ to $-5.0\ \frac{eV}{\rm atom}$. The resulting two- and three-body potentials are shown in Figs. 16 and 17. A more detailed picture of three-body potential is given in the Supplemental Material [64].

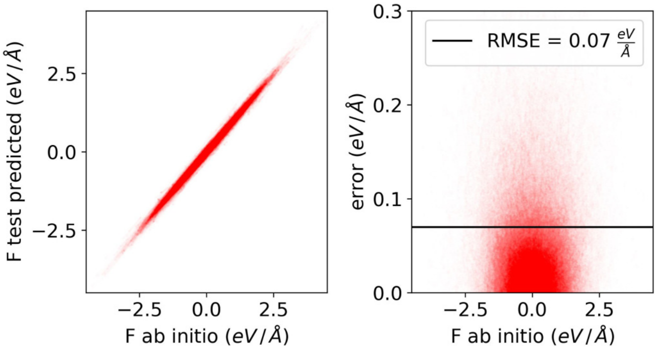

FIG. 14. Performance of GTTP on titanium hydride.

The position of the minimum of the two-body potential is 1.43 Å, which, as expected, corresponds to the C–C bond length (the C–C bond length is 1.40 Å in graphite, and 1.54 Å in diamond). The three-body potential has a very distinct minimum, which is also shown in the form of isosurface in Fig. 17, at the equilateral triangle with the side of 2.47 Å. This means that carbon should prefer crystal structures with such triangles. As Fig. 18 [76,77] shows, both graphite and diamond contain such equilateral triangles with the side of approximately 2.5 Å.

In addition, the importance of two- and three-body interactions in various systems can be studied. In order to do it, we gathered statistics for three archetypal cases—nearly-free-electron metal (aluminum), metal with a significant directional component of bonding (tungsten), and a covalent substance (carbon), which is shown in Table VI.

In the case of aluminum, the two-body description can reproduce most of the variability in energies and forces. The error of only two-body potential is relatively low, and, in the case of two- and three-body potential, the three-body part plays the role of small correction. The situation is the opposite for tungsten and carbon. In this case, the three-body interactions are very important, and moreover, correlations of higher order make a noticeable contribution to the energy variance.

VII. CONCLUSIONS

We have developed the framework for constructing two- and three-body potentials. Our methodology allows to model any two- and three-body interactions with arbitrary precision. At the same time, computational costs do not depend on the number of parameters or fitting flexibility and constitute a constant time per every considered pair or triple of atoms.

TABLE V. Performance of the two- and three-body potential on the Li-intercalated anatase

Dataset	Number of structures	$\sqrt{F^2}\ \frac{eV}{\text{Å}}$	RMSE F $\frac{eV}{\text{Å}}$
$\rm Li_\mathit{x}TiO_2(1)$	618	0.83	0.086, 10.3%
$\rm Li_\mathit{x}TiO_2(2)$	947	2.01	0.152, 7.6%
$\rm Li_\mathit{x}TiO_2(3)$	218	22.4	0.795, 3.5%

TABLE VI. Importance of two- and three-body interactions in aluminum, tungsten, and carbon. For every dataset, the following information is included: (1) standard deviation of *ab initio* energies in the dataset; (2) standard deviation of energies predicted by only two-body component of two- and three-body potential; (3) same for the three-body component; (4) RMSE error of only two-body potential; (5) RMSE error of two- and three-body potential; (6)–(10) the same for forces.

Dataset	Aluminum Rand$_2$	Tungsten	Carbon
STD E, *eV*	9.16	3.85	5.79
STD E 2-body, *eV*	9.47	5.70	7.50
STD E 3-body, *eV*	1.30	3.63	5.80
RMSE E only 2-body, *eV*	0.54, 5.9%	0.63, 16.5%	4.42, 76.3%
RMSE E 2- and 3-body, *eV*	0.21, 2.27%	0.33, 8.5%	2.02, 35.0%
STD F, $\frac{eV}{Å}$	5.00	0.89	5.81
STD F 2 body, $\frac{eV}{Å}$	4.92	0.77	5.64
STD F 3-body, $\frac{eV}{Å}$	0.48	1.08	2.83
RMSE F only 2-body, $\frac{eV}{Å}$	0.29, 5.8%	0.46, 51.9%	1.96, 33.8%
RMSE F 2- and 3-body, $\frac{eV}{Å}$	0.12, 2.4%	0.26, 29.3%	1.17, 20.2%

The assumption of two- and three-body interactions makes GTTP not to be a universal approximator even though two- and three-body potentials can be approximated arbitrarily accurately. The manifestation of this is a convergence of the accuracy to a nonzero limit in Fig. 9. One can not overcome this limit even with an infinite cutoff radius and infinitely flexible parametrization of two- and three-body potentials. Although, as we have shown in this paper, the resulting approximation of potential energy surface is very accurate for many systems, thus allowing to benefit from the computational efficiency of GTTP in atomistic simulations.

We applied our potential to aluminum, tungsten, titanium hydride, Li-intercalated anatase TiO$_2$, and carbon. In the case of aluminum, it showed great accuracy and good transferability properties—we found that the potential trained on only small random structures is able to describe with satisfactory accuracy large structures from a different distribution than in the training dataset. This is even more noticeable in the case of tungsten, where we used only random structures with just eight atoms in the unit cell as training dataset and then applied the potential to study large-scale grain boundaries in polycrystalline structures. We found that our potential significantly outperforms conventional EAM potentials specifically prepared for this purpose. In terms of RMSE of surface energy, our potential is 3–5 times better.

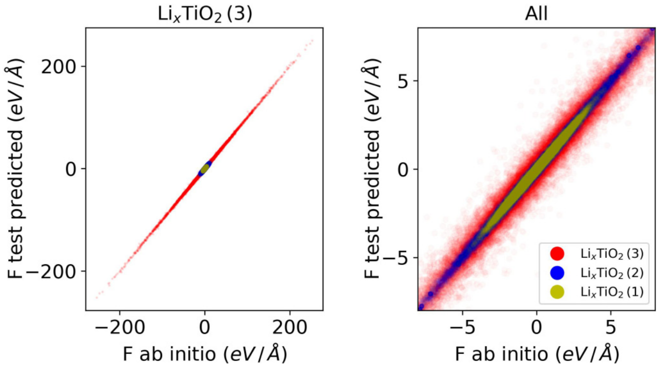

FIG. 15. Performance of GTTP on Li-intercalated anatase

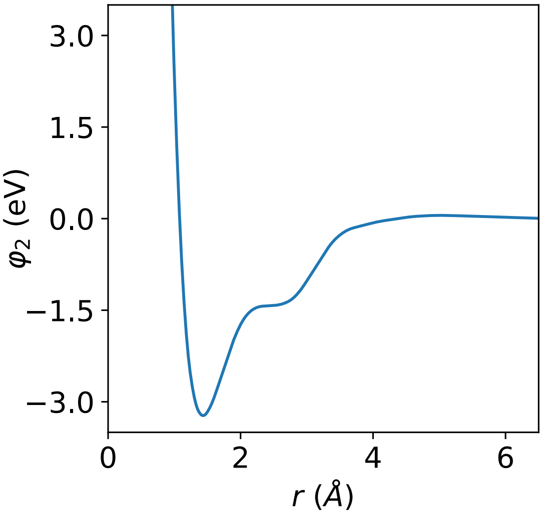

FIG. 16. Two-body potential for carbon.

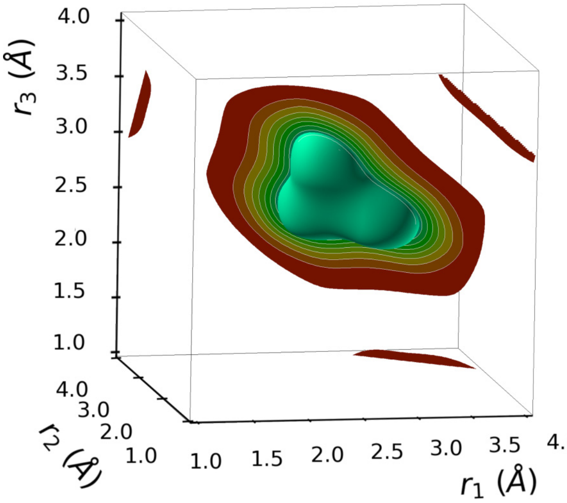

FIG. 17. –2.0 eV isosurface of the three-body potential for carbon (in the center).

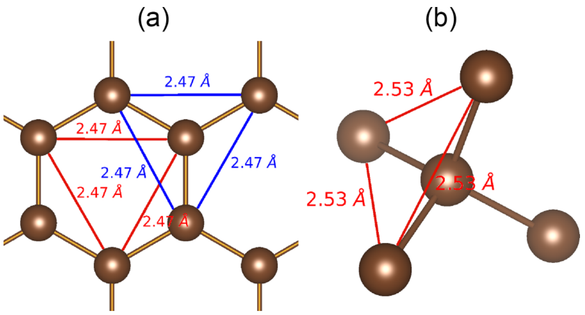

FIG. 18. Equilateral C–C–C triangles in (a) graphite and (b) diamond.

We studied the tradeoff between accuracy and computational time given by the developed potential on aluminum. We found that our potential has good accuracy already at the times of the order of $10^{-6} - 10^{-5}\,\frac{\text{sec}}{\text{atom}}$. Such computational efficiency can be beneficial for conducting very long molecular dynamics simulations. Another example is calculations involving huge systems.

The fitting procedure of our potential is very simple and reduces to linear regression. The number of hyperparameters is relatively small, and the influence of each of them was studied in detail. It is not necessary to search over hyperparameters for every new dataset from scratch. Q_2, Q_3, and Im can be transferred directly, while R^2_{cut} and R^3_{cut} can be chosen in such a way as to ensure the same number of considered pairs and triples of atoms. It approximately corresponds to the same average number of neighbors within the spheres of radii R^2_{cut} and R^3_{cut}.

In addition, the shape of the two- and three-body potential itself can provide useful chemical insights, as shown by the example of carbon. But such interpretations should be made with great care because the potential depends not only on the chemical properties of corresponding atoms but also on the distribution of structures in the training dataset, as well as on hyperparameters.

VIII. DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

We thank Alexandr Shapeev for sharing $Li_xTiO_2(1)$ dataset along with constructing the Pareto front of Moment Tensor Potential on aluminum. We are very grateful to Oleg Sergeev for his invaluable assistance in integrating our potential to LAMMPS and to Aleksey Yanilkin for useful discussions. Also, we thank Qiang Zhu for implementing structure prediction of GBs. A.R.O thanks Russian Science Foundation (Grant No. 19-72-30043). I.K. and A.M. thank the Russian Science Foundation (Grant No. 21-73-10261) for financial support.

S.P., A.R.O., and E.M. designed the research, S.P. developed the interatomic potential and tested it for different systems, A.M. and I.K. investigated GBs in tungsten. All authors wrote the manuscript.

[1] M. Born and R. Oppenheimer, Zur quantentheorie der molekeln, Ann. Phys. (Leipzig) **389**, 457 (1927).

[2] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys. Rev. **136**, B864 (1964).

[3] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. **140**, A1133 (1965).

[4] J. C. Slater and G. F. Koster, Simplified LCAO method for the periodic potential problem, Phys. Rev. **94**, 1498 (1954).

[5] M. S. Daw and M. I. Baskes, Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals, Phys. Rev. B **29**, 6443 (1984).

[6] M. I. Baskes, Modified embedded-atom potentials for cubic materials and impurities, Phys. Rev. B **46**, 2727 (1992).

[7] Y. Mishin and A. Lozovoi, Angular-dependent interatomic potential for tantalum, Acta Mater. **54**, 5013 (2006).

[8] J. W. Ponder and D. A. Case, Force fields for protein simulations, in *Advances in Protein Chemistry* (Elsevier, Amsterdam, 2003), Vol. 66, pp. 27–85.

[9] A. C. Van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, ReaxFF: A reactive force field for hydrocarbons, J. Phys. Chem. A **105**, 9396 (2001).

[10] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, Deep convolutional neural network for age estimation based on VGG-face model, arXiv:1709.01664.

[11] K. Buza, Feedback prediction for blogs, in *Data Analysis, Machine Learning and Knowledge Discovery* (Springer, New York, 2014), pp. 145–152.

[12] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. **98**, 146401 (2007).

[13] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. **134**, 074106 (2011).

[14] J. Behler, Perspective: Machine learning potentials for atomistic simulations, J. Chem. Phys. **145**, 170901 (2016).

[15] T. Morawietz and J. Behler, A density-functional theory-based neural network potential for water clusters including van der Waals corrections, J. Phys. Chem. A **117**, 7356 (2013).

[16] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, *Ab initio* quality neural-network potential for sodium, Phys. Rev. B **81**, 184107 (2010).

[17] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, Microscopic Origins of the Anomalous Melting Behavior of Sodium under High Pressure, Phys. Rev. Lett. **108**, 115701 (2012).

[18] J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, Metadynamics Simulations of the High-Pressure Phases of Silicon Employing a High-Dimensional Neural Network Potential, Phys. Rev. Lett. **100**, 185501 (2008).

[19] N. Artrith, T. Morawietz, and J. Behler, High-dimensional neural-network potentials for multicomponent systems: Application to zinc oxide, Phys. Rev. B **83**, 153101 (2011).

[20] P. E. Dolgirev, I. A. Kruglov, and A. R. Oganov, Machine learning scheme for fast extraction of chemically interpretable interatomic potentials, AIP Adv. **6**, 085318 (2016).

[21] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, Phys. Rev. Lett. **104**, 136403 (2010).

[22] A. P. Bartók, M. J. Gillan, F. R. Manby, and G. Csányi, Machine-learning approach for one-and two-body corrections to density functional theory: Applications to molecular and condensed water, Phys. Rev. B **88**, 054104 (2013).

[23] A. P. Bartók, R. Kondor, and G. Csányi, On representing chemical environments, Phys. Rev. B **87**, 184115 (2013).

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

[25] E. V. Podryabinkin and A. V. Shapeev, Active learning of linear interatomic potentials, Comp. Mat. Sci. **140**, 171 (2017).

[26] I. Kruglov, O. Sergeev, A. Yanilkin, and A. R. Oganov, Energy-free machine learning force field for aluminum, Sci. Rep. **7**, 8512 (2017).

[27] Z. Li, J. R. Kermode, and A. De Vita, Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces, Phys. Rev. Lett. **114**, 096405 (2015).

[28] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Big data meets quantum chemistry approximations: The δ-machine learning approach, J. Chem. Theory Comput. **11**, 2087 (2015).

[29] V. Botu and R. Ramprasad, Adaptive machine learning framework to accelerate *ab initio* molecular dynamics, Int. J. Quantum Chem. **115**, 1074 (2015).

[30] K. Yao, J. E. Herr, and J. Parkhill, The many-body expansion combined with neural networks, J. Chem. Phys. **146**, 014106 (2017).

[31] T. Mueller, A. G. Kusne, and R. Ramprasad, Machine learning in materials science: Recent progress and emerging applications, Rev. Comp. Chem. **29**, 186 (2016).

[32] C. de Boor, On calculating with *B*-splines, J. Approx. Theory **6**, 50 (1972).

[33] M. G. Cox, The numerical evaluation of b-splines, IMA J Appl Math **10**, 134 (1972).

[34] C. de Boor, in *A Practical Guide to Splines* 27 (Springer-Verlag, Berlin, 2001), pp. 94–101.

[35] C. de Boor, in *A Practical Guide to Splines* 27 (Springer-Verlag, Berlin, 2001), pp. 145–148.

[36] J. Tersoff, New empirical approach for the structure and energy of covalent systems, Phys. Rev. B **37**, 6991 (1988).

[37] F. H. Stillinger and T. A. Weber, Computer simulation of local order in condensed phases of silicon [Phys. Rev. B 31, 5262 (1985)], Phys. Rev. B **33**, 1451(E) (1986).

[38] J. E. Jones, On the Determination of Molecular Fields. II. From the Equation of State of a Gas, Proc. R. Soc. London A **106**, 463 (1924).

[39] T. J. Lenosky, B. Sadigh, E. Alonso, V. V. Bulatov, T. D. de la Rubia, J. Kim, A. F. Voter, and J. D. Kress, Highly optimized empirical potential model of silicon, Modell. Simul. Mater. Sci. Eng. **8**, 825 (2000).

[40] P. Zhang and D. R. Trinkle, A modified embedded atom method potential for interstitial oxygen in titanium, Comput. Mater. Sci. **124**, 204 (2016).

[41] L. Zhang, J. Han, H. Wang, R. Car, and W. E, Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics, Phys. Rev. Lett. **120**, 143001 (2018).

[42] L. Zhang, J. Han, H. Wang, R. Car, and W. E, End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems, *NeurIPS*, 2018.

[43] L. Zhang, D.-Y. Lin, H. Wang, R. Car, and W. E, Active learning of uniformly accurate interatomic potentials for materials simulation, Phys. Rev. Mater. **3**, 023804 (2019).

[44] Y. Zhang, C. Hu, and B. Jiang, Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation, J. Phys. Chem. Lett. **10**, 4962 (2019).

[45] Y. Zhang, J. Xia, and B. Jiang, Physically Motivated Recursively Embedded Atom Neural Networks: Incorporating Local Completeness and Nonlocality, Phys. Rev. Lett. **127**, 156002 (2021).

[46] I. Batatia, D. P. Kovács, G. Simm, C. Ortner, and G. Csányi, Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, Adv. Neural Info. Proc. Syst. **35**, 11423 (2022).

[47] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, and B. Kozinsky, Learning local equivariant representations for large-scale atomistic dynamics, Nat. Commun. **14**, 579 (2023).

[48] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nat. Commun. **13**, 2453 (2022).

[49] J. Gasteiger, J. Groß, and S. Günnemann, Directional message passing for molecular graphs, *ICLR*, 2020.

[50] J. Gasteiger, F. Becker, and S. Günnemann, Gemnet: Universal directional graph neural networks for molecules, *NeurIPS*, 2021.

[51] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Phys. Rev. B **99**, 014104 (2019).

[52] C. van der Oord, G. Dusson, G. Csányi, and C. Ortner, Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials, Mach. Learn.: Sci. Technol. **1**, 015004 (2020).

[53] M. P. Bircher, A. Singraber, and C. Dellago, Improved description of atomic environments using low-cost polynomial functions with compact support, Mach. Learn.: Sci. Technol. **2**, 035026 (2021).

[54] S. R. Xie, M. Rupp, and R. G. Hennig, Ultra-fast interpretable machine-learning potentials, arXiv:2110.00624.

[55] A. R. Oganov and C. W. Glass, Crystal structure prediction using ab initio evolutionary techniques: Principles and applications, J. Chem. Phys. **124**, 244704 (2006).

[56] A. R. Oganov, A. O. Lyakhov, and M. Valle, How evolutionary crystal structure prediction works and why, Acc. Chem. Res. **44**, 227 (2011).

[57] A. O. Lyakhov, A. R. Oganov, H. T. Stokes, and Q. Zhu, New developments in evolutionary structure prediction algorithm USPEX, Comput. Phys. Commun. **184**, 1172 (2013).

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

[59] 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).

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

[61] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B **59**, 1758 (1999).

[62] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer, New York, 2001), Chap. 7.

[63] D. J. MacKay, Bayesian interpolation, Neural Comput. **4**, 415 (1992).

[64] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevB.107.125160 for a detailed table for accuracy-computational time tradeoff and detailed pictures of extracted three-body potentials for aluminum and carbon.

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

[66] A. Shapeev (private communication).

[67] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood *et al.*, A performance and cost assessment of machine learning interatomic potentials, J. Phys. Chem. A **124**, 731 (2020).

[68] A. P. Sutton, in *Interfaces in Crystalline Materials* (Clarendon Press, New York, 1995), pp. 414–423.

[69] M.-C. Marinica, L. Ventelon, M. Gilbert, L. Proville, S. Dudarev, J. Marian, G. Bencteux, and F. Willaime, Interatomic potentials for modelling radiation defects and dislocations in tungsten, J. Phys.: Condens. Matter **25**, 395502 (2013).

[70] X. Zhou, H. Wadley, R. A. Johnson, D. Larson, N. Tabat, A. Cerezo, A. Petford-Long, G. Smith, P. Clifton, R. Martens *et al.*, Atomic scale structure of sputtered metal multilayers, Acta Mater. **49**, 4005 (2001).

[71] T. Frolov, W. Setyawan, R. Kurtz, J. Marian, A. R. Oganov, R. E. Rudd, and Q. Zhu, Grain boundary phases in bcc metals, Nanoscale **10**, 8253 (2018).

[72] W. Setyawan and R. J. Kurtz, *Ab initio* study of H, He, Li and be impurity effect in tungsten Σ3 {1 1 2} and Σ27 {5 5 2} grain boundaries, J. Phys.: Condens. Matter **26**, 135004 (2014).

[73] A. Mazitov, A. Oganov, and A. Yanilkin, Titanium-hydrogen interaction at high pressure, J. Appl. Phys. **123**, 235901 (2018).

[74] I. Novikov and A. Shapeev (private communication).

[75] V. L. Deringer and G. Csányi, Machine learning based interatomic potential for amorphous carbon, Phys. Rev. B **95**, 094203 (2017).

[76] K. Momma and F. Izumi, Vesta 3 for three-dimensional visualization of crystal, volumetric and morphology data, J. Appl. Cryst. **44**, 1272 (2011).

[77] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).