Cluster expansion constructed over Jacobi-Legendre polynomials for accurate force fields

M. Domina*, U. Patil*, M. Cobelli*, and S. Sanvito
School of Physics and CRANN Institute, Trinity College, Dublin 2, Ireland

(Dated: March 22, 2024)

We introduce a compact cluster expansion method constructed over Jacobi and Legendre polynomials to generate highly accurate and flexible machine-learning force fields. The constituent many-body contributions are separated, interpretable, and adaptable to replicate the physical knowledge of the system. In fact, the flexibility introduced by the use of the Jacobi polynomials allows us to impose, in a natural way, constraints and symmetries to the cluster expansion. This has the effect of reducing the number of parameters needed for the fit and of enforcing desired behaviors of the potential. For instance, we show that our Jacobi-Legendre cluster expansion can be designed to generate potentials with a repulsive tail at short interatomic distances, without the need of imposing any external function. Our method is here continuously compared with available machine-learning potential schemes, such as the atomic cluster expansion and potentials built over the bispectrum. As an example, we construct a Jacobi-Legendre potential for carbon by training a slim and accurate model capable of describing crystalline graphite and diamond, as well as liquid and amorphous elemental carbon.

I. INTRODUCTION

Machine-learning potentials (MLPs) are rapidly becoming the gold standard for molecular dynamics and thermodynamical sampling in materials science [1–4]. The general idea is that of performing a high-dimension fit of the potential energy surface (PES) computed with an electronic *ab-initio* method, for instance, with density functional theory (DFT), to obtain a numerical atomic energy functional for large-scale simulations. In particular, one aims at using a conveniently limited number of electronic structure data to interpolate the PES at an accuracy comparable with that of the electronic method itself. MLPs can thus be defined as parametric functions that associate to a given chemical structure the system energy. The mathematical relation between the input features describing a structure, often called *descriptors*, and the output target can be either linear [5–7] or non linear [8–10]. Furthermore, the target quantity may be different from the energy and may include electronic properties [11–14], or even tensorial quantities [15–17].

The specific descriptors choice is crucial to the construction of a MLP. It is commonly agreed that a strategy to drastically reduce the size of the training set and to improve the model accuracy is that of designing descriptors invariant with respect to the symmetries of the target quantity. In the case of the total energy, this results in descriptors which are invariant for translations, rotations, and permutations of identical atoms. In principle, one can then combine any choice of descriptors with any desired machine-learning model, going from simple regressions to neural networks of various complexity to kernel-based schemes. Typically, there is a subtle tradeoff between the model complexity, the descriptor type, and the size and composition of the data set needed to construct the MLP. Complex many-body descriptors [18] are usually combined with linear models, while simpler structure representations are used as input to deep-learning algorithms. In both cases, there may be issues of interpretability, namely, it is not always transparent what the level of physics learned by the model itself is. As a consequence one often relies on numerical techniques to establish whether a particular atomic configuration is interpolated or extrapolated by the model [19].

In this paper we introduce a linear model built over a set of descriptors derived from the energy cluster expansion. Our MLP, that we name the Jacobi-Legendre potential (JLP), is close in spirit to the recently introduced Atomic Cluster Expansion (ACE) [5, 20]. In fact, given the completeness of the ACE [21], one can establish a one-to-one mapping between the two potentials. Importantly, our JLPs adopt internal coordinates, so they are, in essence, expansions of the N-body potentials in orthogonal polynomials evaluated on distances and angles between atoms. As such, the JLPs are not affected by issues concerning the invariant coupling of different angular momenta channels [9, 22]. Our use of the internal coordinates is closer to the recently developed proper orthogonal descriptors (PODs) [23, 24]. Here, however, we retain the spherical harmonics formalism by mean of the Legendre polynomials, so a comparison between the JLPs and other well-known potentials can be naturally drawn. Our scheme also makes extensive use of Jacobi polynomials, of which the Legendre ones are a particular case.

Given a central atom, one of the most important properties of MLPs is the achievement of linear scaling for the time required to compute the local descriptors with respect to the number of surrounding atoms, up to an optimized cutoff distance. Here, we will show that linear scaling can be achieved for the JLPs too and, in doing so, we will establish a link between well-known MLPs and the internal coordinate representation adopted in this paper.

The use of an explicit expansion over orthogonal and complete polynomials gives several advantages, such as the enforcement of symmetries and local constraints. Indeed, our potentials are constructed so key properties, such as the smooth vanishing contribution at the cut-off radius, arise naturally without the need of introducing *ad hoc* cutoff functions. In fact, these properties are enforced by applying constraints on the expansion coefficients. Crucially, the procedure introduced in our paper is completely general, so not only the number of coefficients to learn can be substantially reduced but also the physical knowledge of the PESs can be introduced in a natural way.

For instance, a desired feature arising from the choice of the Jacobi polynomials and of the constraining procedure, is that, by appropriately tuning the hyperparameters, a repulsive behavior naturally emerges for the potential at small distances. This is obtained without introducing any external repulsive function. Moreover, while it is not generally possible to completely separate the body-order contributions, we formally avoid any mixing between them. This allows us to reconstruct the N-body functional dependence in terms of the learned coefficients. As a consequence, by combining these two properties, one can introduce an inductive bias in the models by selecting, for example, only the hyperparameters that lead to a repulsive short-range behavior of the two-body (2B) interaction. Since small distances are usually absent from the training set, a direct consequence is that the potentials naturally possess a physically meaningful behavior in this extrapolated regime.

The paper is organized as follows. An extensive Methods section presents in detail each body order of the expansion, with a discussion on the relevant properties of each term. Then, the potential is fitted to the carbon data set used to train the GAP17 potential of Ref. [25]. The result of the fit on energies, forces, and stress are reported. Furthermore, we will close this case study by presenting the phonon dispersion curves for graphene and diamond, as predicted by the trained JLP model.

II. METHODS

In this section, we introduce the JLPs. This class of potentials is based on the total energy cluster (many-body) expansion. Therefore, after a discussion of the main idea behind such a strategy, we will proceed with the systematic introduction of each many-body term and their associated technical details. Note that a similar strategy can also be used to construct a JLP-like model for quantities different from the energy, such as the charge density at a particular point in space [26].

A. Introduction

An overview of the strategy behind the construction of the JLPs is provided in Fig. 1. In general, it is reasonable to assume that the total energy of a system, E, can be partitioned into a short- and a long-range contribution. Our proposed MLP accounts only for the short-ranged part, E_{short}, that can be further expanded over terms vanishing at distances larger than a characteristic interaction range. In particular, we follow the well-known strategy of a multi-body expansion for the energy and write

$$E_{\mathrm{short}} = E_1 + E_2 + E_3 + E_4 \dots \,. \tag{1}$$

Here the single-body contribution, E_1, is an energy offset depending on the number of atomic species present in the system, E_2 is the 2B energy, depending only on atoms pairs, E_3 is the three-body (3B) energy, depending on triplets of atoms and, in general, E_n describes the n-body (nB) energy term.

A second essential assumption is that we can decompose each of the nB energy terms in local quantities, such that each term can be written as a sum of atom-centered contributions. Explicitly, this writes

$$E_n = \sum_{i}^{\mathrm{atoms}} \varepsilon_i^{(n)} \,, \tag{2}$$

with $n \geq 2$, and where the sum runs over all possible atoms in the system. Each local contribution to the nB energy, $\varepsilon_i^{(n)}$, depends only on the local neighborhood of the i-th atom (the red atom in Fig. 1), up to a cut-off distance r_{cut}.

In essence, the JLPs consist of a linear expansion of the $\varepsilon_i^{(n)}$ contributions. As such, at the core, the JLP is closely related to linear MLPs such as the spectral neighbor analysis potential (SNAP)[6], the moment tensor potentials (MTPs) [7], and the ACE [5, 20]. Since the successful generalization of the coupling scheme of the power spectrum (a 3B representation) and the bispectrum [a four-body representation] [9, 22] to any higher-body order, first introduced in the ACE potentials, all new potentials build from the same set of assumptions (many-body expansion of the energy and locality), differ in the way of constructing the basis functions, or on the introduction of completely new basis sets [1, 2]. The JLPs are not different in these regards. Based on a particular choice of basis functions (radial and angular), they are also complete, so a one-to-one mapping between the terms of a JLP and the analogous ones of the ACE is possible. In particular, as the name suggests, we chose the Jacobi polynomials as the radial basis and the Legendre polynomials as the angular one.

The choice of Jacobi polynomials [27], $P_n^{(\alpha,\beta)}(x)$, is motivated by their dependence on the two real parameters, α and β, which can lead to a broad selection of different orthogonal polynomials. Two classical examples are the Legendre polynomials ($\alpha = \beta = 0$) and the Chebyshev polynomials of the second kind ($\alpha = \beta = 1/2$). Thus, treating α and β as hyper-parameters allows one to optimize the radial basis set, and removes the need for manually choosing the best basis. In contrast, we have

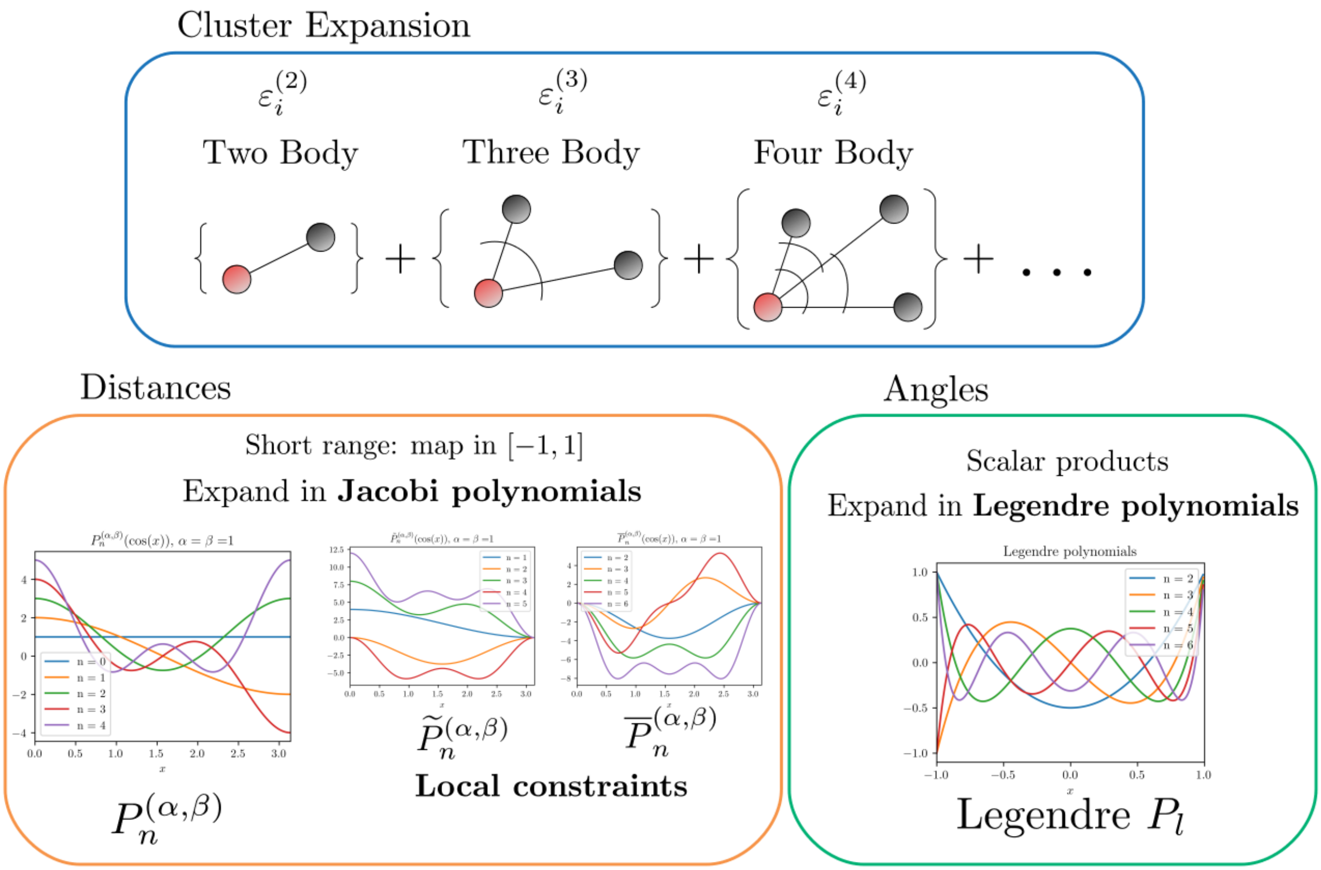
Cluster Expansion
$\varepsilon_i^{(2)}$
Two Body
$\varepsilon_i^{(3)}$
Three Body
$\varepsilon_i^{(4)}$
Four Body
+ + + . . .
Distances
Short range: map in $[-1,1]$
Expand in **Jacobi polynomials**
$P_n^{(\alpha,\beta)}(\cos(x))$, $\alpha = \beta = 1$
n = 0
n = 1
n = 2
n = 3
n = 4
x
$\widetilde{P}_n^{(\alpha,\beta)}(\cos(x))$, $\alpha = \beta = 1$
n = 1
n = 2
n = 3
n = 4
n = 5
$\overline{P}_n^{(\alpha,\beta)}(\cos(x))$, $\alpha = \beta = 1$
n = 2
n = 3
n = 4
n = 5
n = 6
$P_n^{(\alpha,\beta)}$
$\widetilde{P}_n^{(\alpha,\beta)}$
$\overline{P}_n^{(\alpha,\beta)}$
Local constraints
Angles
Scalar products
Expand in **Legendre polynomials**
Legendre polynomials
n = 2
n = 3
n = 4
n = 5
n = 6
x
Legendre P_l

FIG. 1. The workflow of the linear model presented in this paper. We first decompose the total energy over the terms of a local multibody expansion, as in Eq. (1). Each contribution is then further expanded over atomic contributions [Eq. (2)], $\varepsilon_i^{(n)}$, which depend on the distances and the angles between a central atom (in red) and atoms in its neighborhood (in black). For instance, the two-body term consists only of one distance, the three-body one of two distances and one angle, etc. By assuming short-range interaction, the distances are then mapped onto the interval $[-1, 1]$, so each distance-dependent term can be expanded as products of Jacobi polynomials. The angles are then mapped onto scalar products, so the functional dependence on the angles can be similarly expanded in terms of Legendre polynomials. Crucially, the expansions on the distances is locally constrained, so effective polynomials will be employed in place of the Jacobi polynomials.

chosen the Legendre polynomials not only since they lead to a certain homogeneity in the representation (being the Legendre polynomials a particular instance of the Jacobi ones), but also for their strong relation with the spherical harmonics. This means that a spherical harmonics decomposition can always be performed, a key feature for achieving computational-linear scaling with respect to the number of atoms (neighbors) inside the interaction cut-off sphere.

After performing the expansion over the chosen basis, we will present a general way for constraining the expansion coefficients, so known physical (and local) properties of the system can be encoded directly in the descriptors at any body order. As a byproduct of applying the constraint on Jacobi polynomials, we will show the natural emergence of the widely used cut off function, $f_c = (1 - \cos(x))/2$. In this case, a cut-off function, f_c, is not externally imposed on the basis set, but instead emerges naturally from the formalism.

Finally, since it has been proved that all 4B descriptors mentioned before are not complete, in the sense that one could find two distinct local environments with the same set of descriptors [28], or manifold with slow-varying fingerprints with respect to a similarity measure [29, 30], we will explicitly investigate the JLP up to the five-body (5B) order term, E_5. However, we anticipate that the choice of the Jacobi polynomial as a basis set, and the associated constraining procedure, can also be applied to other potentials. Indeed, they can be exported easily to other multibody expansion approaches, so, for example, one could use the constrained-Jacobi basis as a radial basis for ACE.

B. The Two-Body Term

In this section, we introduce the expansion of the 2B energy term, E_2. Since the total energy is a scalar, it

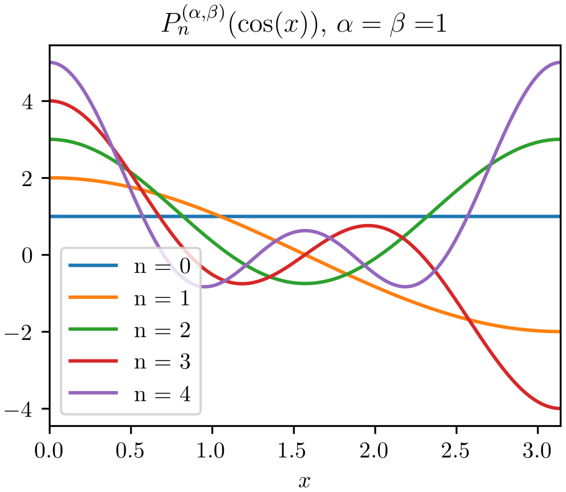

FIG. 2. First five Jacobi polynomials, $P_n^{(\alpha,\beta)}$, with $\alpha = \beta = 1$. In the plot, the functions are composed with a cosine, to provide a better idea of the representation used in the expansion of Eq. (8). It can be appreciated how the derivative is zero at both edges of the domain.

must be invariant under translations and rotations of the reference frame. A possible way to satisfy such invariance is to assume that the energy depends only on the distances between atom pairs, and that this dependency is realized by a continuous function (potential), $v^{(2)}$. Furthermore, we assume that the actual functional form depends only on the atomic species of the atoms involved, so, if Z_i is the atomic number of the atom located at the position $\boldsymbol{r}_i$, and $r_{ji} = |\boldsymbol{r}_j - \boldsymbol{r}_i|$, we have

$$v^{(2)} \equiv v^{(2)}(r_{ji}; Z_j, Z_i) \equiv v^{(2)}_{Z_j Z_i}(r_{ji}), \tag{3}$$

and

$$E_2 = \sum_{\substack{ij \\ j\neq i}} v^{(2)}_{Z_j Z_i}(r_{ji}) \, . \tag{4}$$

The 2B potential, $v^{(2)}_{Z_j Z_i}$, is thus defined symmetric under the exchange $Z_j \leftrightarrow Z_i$, namely, $v^{(2)}_{Z_j Z_i} = v^{(2)}_{Z_i Z_j}$. Note that, in principle, one can still explicitly distinguish non-equivalent atoms belonging to the same species, by introducing "virtual" species.

It is useful to remark that the 2B term in Eq. (4) can be recast in the form of Eq. (2), where $\varepsilon_i^{(2)} = \sum_{j\neq i} v^{(2)}_{Z_j Z_i}(r_{ji})$ is the energy associated to the i-th atom resulting from the pairwise interaction with its local atomic neighborhood. Note that these local contributions are well-defined because of the short-ranged nature of the interaction. Thus, there exists a natural cut-off radius r_{cut}, such that $v_{Z_j Z_i}(r_{ji}) \simeq 0$ for $r_{ji} > r_{\text{cut}}$.

We now provide the proposed expansion for the potentials $v^{(2)}_{Z_j Z_i}$, followed by its derivation. The expansion is

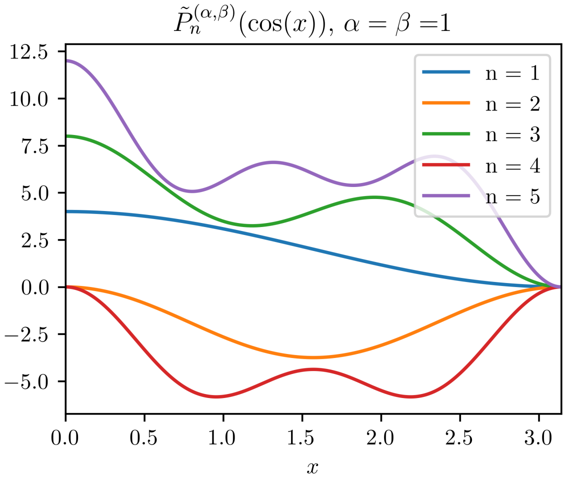

FIG. 3. First five vanishing-Jacobi polynomials, $\widetilde{P}_n^{(\alpha,\beta)}$ ($\alpha = \beta = 1$) as defined in Eq. (6), derived from the Jacobi polynomials of Fig. 2. These polynomials are constrained to vanish at the right-hand side edge of the domain.

$$v^{(2)}_{Z_j Z_i}(r_{ji}) = \sum_{n=1}^{n_{\max}} a_n^{Z_j Z_i} \widetilde{P}_n^{(\alpha,\beta)} \left(\cos\left(\pi \frac{r_{ji} - r_{\min}}{r_{\text{cut}} - r_{\min}} \right) \right) , \tag{5}$$

where the sum is truncated to a suitable polynomial order, $n_{\max}$, and where $a_n^{Z_j Z_i}$ are the expansion coefficients for the n-th order. The vanishing-Jacobi polynomials, $\widetilde{P}_n^{(\alpha,\beta)}$, employed here, are defined in terms of the Jacobi polynomials, $P_n^{(\alpha,\beta)}$, as

$$\widetilde{P}_n^{(\alpha,\beta)}(x) = P_n^{(\alpha,\beta)}(x) - P_n^{(\alpha,\beta)}(-1) \quad \text{for } -1 \leq x \leq 1 \, , \tag{6}$$

for $n \geq 1$. Thus, the $\widetilde{P}_n^{(\alpha,\beta)}$ have the property to vanish at the right-hand side extreme of their domain, namely, at r_{cut}. The Jacobi polynomials are shown in Fig. (2), while the corresponding vanishing-Jacobi polynomials are in Fig. 3. The expansion presents five hyper-parameters, α, β, r_{cut}, $r_{\min}$, and $n_{\max}$, with α and β being real numbers greater than -1. We will refer to Eq. (5) as the 2B-Jacobi-Legendre (2B-JL) expansion.

At this point, it should be noted that there can be a different set of hyperparameters for each different atomic species. As such, in the case of many-species compounds, the hyper-parameter space can potentially become rather large. Then, it may be desirable to take system-based approximations or to perform feature selection.

We will now present the arguments leading to Eq. (5). To make the formalism more readable, we define the compact notation

$$\widetilde{P}_{nji}^{(\alpha,\beta)} \equiv \widetilde{P}_n^{(\alpha,\beta)} \left(\cos\left(\pi \frac{r_{ji} - r_{\min}}{r_{\text{cut}} - r_{\min}} \right) \right) , \tag{7}$$

which will be widely used throughout this paper. As already remarked, the potentials should vanish for distances larger than the interaction cut-off radius. With only this constraint in mind, we can expand the potential in terms of Jacobi polynomials as

$$v^{(2)}(r) = \sum_{n=0}^{n_{\max}} a_n P_n^{(\alpha,\beta)}\left(\cos\left(\pi r/r_{\mathrm{cut}}\right)\right) , \tag{8}$$

where, for simplicity, we set $r_{\min} = 0$ and omit the explicit dependence on the atomic species.

We have chosen the Jacobi polynomials, $P_n^{(\alpha,\beta)}(x)$, since they are complete and orthogonal over the interval $x \in [-1, 1]$, with respect to the weight function $w^{\alpha\beta}(x) = (1 - x)^{\alpha}(1 + x)^{\beta}$. Furthermore, as already noted, their generality, parameterized through the real coefficients α and β, allows one to perform automatic searches of the most efficient basis set, without any additional hypothesis. We found that, in most cases, there is a large range of optimal α and β values, so we usually reduce the number of hyperparameters by constraining the search to $\alpha = \beta$.

Note that we are not introducing any cut-off function in the expansion to force a smooth-vanishing behavior at the cut-off radius. Also, the a_0 coefficient is not present in the sum of Eq. (5), while it still appears in Eq. (8). We will now impose the right behavior on the expansion coefficients, so the resulting potential vanishes by construction at the cut-off radius. The result of this approach is similar, in a sense, to models with naturally vanishing radial functions, such as the spherical-Bessel descriptors [31]. Explicitly, we constrain the expression in Eq. (8) to satisfy the condition $v^{(2)}(r_{\mathrm{cut}}) = 0$. Then, since $P_0^{(\alpha,\beta)} = 1$, we obtain that the first coefficient must satisfy

$$a_0 = -\sum_{n\geq 1}^{n_{\max}} a_n P_n^{(\alpha,\beta)}(-1) . \tag{9}$$

By inserting this expression back into Eq. (8), we finally obtain Eq. (5). It is worth mentioning that this procedure can be easily generalized to impose any constraint to the functional form of the potential, so local physical knowledge of the system can be enforced in the description itself. An example of a further constraint will be shown for higher-body terms.

We can then interpret the vanishing Jacobi polynomials, defined in Eq. (6), as the radial basis obtained when expanding functions vanishing at the left-hand side limit of the interval $[-1, 1]$ (in our mapping, the point $x = -1$ is mapped onto the cut-off distance). As a final remark, the expansion coefficients $a_n^{Z_j Z_i}$ inherit the same symmetry properties of the potential, namely, they are symmetric under the exchange of the atomic species, $a_n^{Z_j Z_i} = a_n^{Z_i Z_j}$.

In closing this section, it must be mentioned that the 2B-JL expansion suffers from the same scaling problem of most of the established MLPs when dealing with multiple species. In fact, the number of pair-wise potentials that one can define scales quadratically with the number of species, so system-based approximations are required for complex chemical compositions. This problem will become more severe for the higher-body terms. Despite the fact that the current implementation of the JLP can already deal with multi-species compounds, in this paper we will show an application for a single-species (carbon) system, postponing the explicit investigation of multi-species potentials to future works.

Emergence of the cut-off function from the constraints

A relevant property of the 2B-JL expansion is that, as rigorously proved in Appendix A, we can factorize the vanishing Jacobi polynomials as

$$\widetilde{P}_n^{(\alpha,\beta)}(\cos x) = f_c(x) Q_n^{(\alpha,\beta)}(\cos(x)) , \tag{10}$$

where $f_c(x)$ is the well-know cut-off function $f_c(x) = (1 + \cos(x))/2$, introduced in Ref. [8], and the $Q_n^{(\alpha,\beta)}(\cos(x))$ are functions explicitly defined in Appendix A. As far as we know, the functions $Q_n^{(\alpha,\beta)}$ are not equivalent to other functions already used in the MLPs literature. While the property described by Eq. (10) establishes a strong connection between our expansion and other potentials which use the cut-off function, it is important to stress that, within the JLPs, $f_c(x)$ arises naturally from the choice of the radial basis and the constraining method implemented. As such, it is not an embedding function, as one can clearly see in Fig. 3. Among the advantages of this approach there is that, since the Jacobi polynomials are already complete and orthogonal, no further orthogonalization procedure has to take place. Also, we do not have to explicitly evaluate the derivative of the cut-off function when calculating the forces, since we can simply use the derivative of the (vanishing-) Jacobi polynomials. Finally, by imposing the constraint of Eq. (9), we are reducing the number of coefficients to learn: this is particularly relevant for higher-body terms, as will be shown in the following sections.

C. The Three-Body Term

In this section, we will discuss the linear expansion of the 3B energy term, E_3. While the core strategy is the same as the one employed in the previous section, we will introduce here a Legendre expansion for the angular dependence of the cluster. We will also impose a further constraint on the Jacobi polynomials and we will discuss the role of symmetries when considering atoms of the same species.

Following the same approach introduced in the previous section, we assume that E_3 can be written as a sum

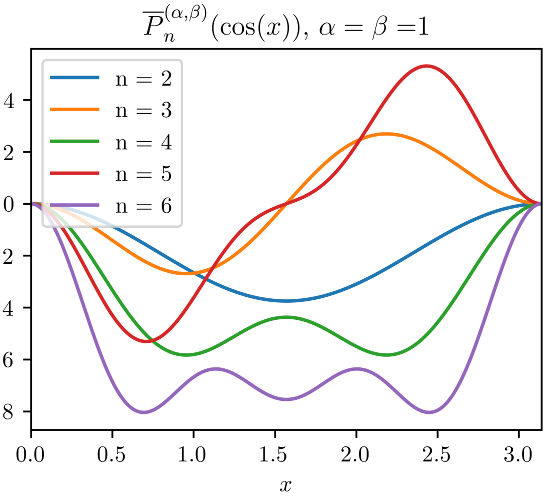

FIG. 4. The first five double-vanishing Jacobi polynomials, $\overline{P}_n^{(\alpha,\beta)}$ (here plotted for $\alpha = \beta = 1$), as defined in Eq. (13), and derived from the vanishing ones shown in Fig. 3. The polynomials are constrained to vanish at both edges of the domain.

of local 3B potentials, $v^{(3)}$, as

$$E_3 = \sum_i^{\text{atoms}} \sum_{(j,k)_i} v^{(3)}_{Z_j Z_k Z_i}(r_{ji}, r_{ki}, \hat{\boldsymbol{r}}_{ji} \cdot \hat{\boldsymbol{r}}_{ki}) \,, \tag{11}$$

where the first sum runs over all the atoms in the system and the second one runs over all the atoms pairs in the neighborhood (within r_{cut}) of the i-th atom (the red atom in Fig. 1). Here, to ensure the translational and rotational invariance of the descriptors, we consider only internal coordinates between the central atom i and the atoms j and k in the surroundings. Therefore, only the distances r_{ji} and r_{ki} and the scalar products $\hat{\boldsymbol{r}}_{ji} \cdot \hat{\boldsymbol{r}}_{ki}$ (essentially the angle defining a 3B cluster), are taken into consideration.

The functional form of the potential $v^{(3)}_{Z_j Z_k Z_i}$ depends on the ordering of the atomic species numbers Z_j, Z_k, and Z_i. Specifically, the first atomic species refers to the first distance, the second atomic species to the second distance, while the last one refers to the central atom. Thus, it holds that

$$v^{(3)}_{Z_k Z_j Z_i}(r_{ki}, r_{ji}, s_{jki}) = v^{(3)}_{Z_j Z_k Z_i}(r_{ji}, r_{ki}, s_{jki}) \,, \tag{12}$$

where s_{jki} is a short-hand notation for $\hat{\boldsymbol{r}}_{ji} \cdot \hat{\boldsymbol{r}}_{ki}$. Put into words, if we exchange the species of the atoms in the environment, we will also have to exchange their distances. From now on, we will use $v^{(3)}_{jki}$ as shorthand notation for $v^{(3)}_{Z_j Z_k Z_i}$.

By adopting the same workflow followed in constructing the 2B case, we now give an expression for the 3B JL expansion, and then we provide its derivation. The 3B-JL expansion reads

$$v^{(3)}_{jki}(r_{ji}, r_{ki}, s_{jki}) = \sum_{n_1, n_2 = 2}^{n_{\max}} \sum_{l=0}^{l_{\max}} a^{jki}_{n_1 n_2 l} \overline{P}^{(\alpha,\beta)}_{n_1 ji} \overline{P}^{(\alpha,\beta)}_{n_2 ki} P^{jki}_l \,, \tag{13}$$

where $P^{jki}_l = P_l(s_{jki})$ is the Legendre Polynomial, P_l, evaluated on the scalar product s_{jki}. The first sum runs on all the n_1 and n_2 in the interval [2,$n_{\max}$]. The 3B expansion introduces a new hyperparameter, $l_{\max}$, which sets the level of truncation of the angular expansion. The coefficients $a^{jki}_{n_1 n_2 l}$ have to be intended as a compact form for $a^{Z_j Z_k Z_i}_{n_1 n_2 l}$. Crucially, we use here the double-vanishing Jacobi polynomials, $\overline{P}^{(\alpha,\beta)}_n(x)$, which can be defined in terms of the vanishing ones as (see Fig. 4)

$$\overline{P}^{(\alpha,\beta)}_n(x) = \widetilde{P}^{(\alpha,\beta)}_n(x) - \frac{\widetilde{P}^{(\alpha,\beta)}_n(1)}{\widetilde{P}^{(\alpha,\beta)}_1(1)} \widetilde{P}^{(\alpha,\beta)}_1(x) \,, \tag{14}$$

for $n \geq 2$. From this definition, it can be seen that the double-vanishing polynomials not only vanish smoothly at the cut-off distance ($x = -1$) but also for small distances ($x = 1$). By employing these polynomials, the repulsive behavior at short distances is not influenced by the 3B-JL expansion and, as such, is completely determined by the 2B expansion. Note that these polynomials have been devised for the case in which $r_{\min}$ is small. If this hypothesis does not hold, we suggest a case-by-case investigation of the most appropriate polynomials or constraints to use.

The derivation of the expansion in Eq. (13) follows the same strategy presented in detail for the derivation of the 2B-JL expansion, Eq. (5). Since the distances and the scalar product are independent variables, we expand the functional dependence of the potential on the distances in terms of a product of two Jacobi polynomials, one for each distance. Then, the scalar product dependence is expanded in terms of Legendre polynomials. Analogously to the constraint adopted in the 2B case, we constrain the expansion to vanish at the cut-off radius. Here, however, we impose the potential to vanish when *at least one* of the distances approaches the cut-off, independently of the value of the other distance or of the angular part. Crucially, applying independent constraints on the variables at play, allows us to severely reduce the number of free coefficients, when compared to the 2B case. Indeed, the constraints explicitly read [please, compare with the constraint introduced in Eq. (9)]

$$\begin{cases} a_{0 n_2 l} = -\displaystyle\sum_{n_1 \geq 1}^{n_{\max}} a_{n_1 n_2 l} P^{(\alpha,\beta)}_{n_1}(-1) & \text{for all } n_2, l, \\ a_{n_1 0 l} = -\displaystyle\sum_{n_2 \geq 1}^{n_{\max}} a_{n_1 n_2 l} P^{(\alpha,\beta)}_{n_2}(-1) & \text{for all } n_1, l, \end{cases}$$

so the expression can be re-casted in terms of products of vanishing Jacobi polynomials and Legendre polynomials only. However, we can further constrain the number of free coefficients by imposing that the potentials also vanish when at least one of the distances approaches zero. In this way we impose a condition also on the $a^{jki}_{1n_2l}$ and $a^{jki}_{n_11l}$ coefficients. In doing so, we obtain the double-vanishing polynomials and the 3B-JL expansion of Eq. (13). Note that in the unconstrained case we have $(n_{\max}+1)^2(l_{\max}+1)$ free coefficients, while in the double-constrained one these are only $(n_{\max}-1)^2(l_{\max}+1)$. As such, we deduce that the reduction in the number of coefficients is quite severe for relatively low $n_{\max}$. Another relevant reduction in the number of free parameters is induced through the symmetries of the coefficients, when atoms of the same species are taken into account, as explained in detail in the following section.

Symmetries of the coefficients

We can explicitly read the role of the indexes in the expansion coefficients $a^{jki}_{n_1n_2l}$ of Eq. (13), by noting that the first index, n_1, refers to the expansion on the first argument of the potential $v^{(3)}_{jki}$ (the distance between the atoms j and i), while the second one expands the second distance. Thus, the symmetry property of the potentials described by Eq. (12) directly implies that $a^{jki}_{n_1n_2l} = a^{kji}_{n_2n_1l}$, namely, that the expansion coefficients are symmetric with respect to the simultaneous exchange of the species indexes $Z_j \leftrightarrow Z_k$ and of the Jacobi indexes, $n_1 \leftrightarrow n_2$. While this is effectively just a reordering of the arguments of the potential, with appropriate re-labelling, it becomes relevant in the case of identical atoms. Indeed, if the atoms j and k belong to the same atomic species Z, then they are indistinguishable, making the potential invariant under the exchange of the first and the second arguments (the two distances). Then, one needs to enforce the same symmetry on the coefficients, namely, they must be symmetric under Jacobi-index exchange alone, $a^{ZZZ_i}_{n_1n_2l} = a^{ZZZ_i}_{n_2n_1l}$.

We can then re-cast the 3B-JL expansion for the same atom species, $Z_j = Z_k = Z$, as

$$
\begin{aligned}
&v^{(3)}_{ZZZ_i}(r_{ji}, r_{ki}, s_{jki}) \\
&\quad= \sum_{n_1=2}^{n_{\max}} \sum_{l=0}^{l_{\max}} a^{ZZZ_i}_{n_1n_1l} \overline{P}^{(\alpha,\beta)}_{n_1ji} \overline{P}^{(\alpha,\beta)}_{n_1ki} P^{jki}_l \\
&\quad+ \sum_{\substack{n_1=2\\ n_2=2\\ n_1>n_2}}^{n_{\max}} \sum_{l=0}^{l_{\max}} a^{Z_jZ_kZ_i}_{n_1n_2l} \left[\overline{P}^{(\alpha,\beta)}_{n_1ji} \overline{P}^{(\alpha,\beta)}_{n_2ki} + \overline{P}^{(\alpha,\beta)}_{n_2ji} \overline{P}^{(\alpha,\beta)}_{n_1ki} \right] P^{jki}_l \, .
\end{aligned} \tag{15}
$$

Equation (15) explicitly shows the application of the symmetries for $n_1 \neq n_2$. Now, we can introduce the more practical expression

$$
\begin{aligned}
&v^{(3)}_{Z_jZ_kZ_i}(r_{ji}, r_{ki}, s_{jki}) \\
&\quad= \sum_{n_1n_2l}^{\text{unique}} a^{Z_jZ_kZ_i}_{n_1n_2l} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1ji} \overline{P}^{(\alpha,\beta)}_{n_2ki} P^{jki}_l \right) ,
\end{aligned} \tag{16}
$$

which also encompasses the cases for different species and is easily generalized to higher-body order expansion terms. Here, the first sum runs over indexes that lead to non-equivalent coefficients with respect to the symmetries of the potential (in this case, indexes such that $n_1 \geq n_2$), while the second sum runs over all the permutations of indexes that refers to equivalent coefficients (in this case the exchange $n_1 \leftrightarrow n_2$). If the atoms j and k belong to two different species, then the expression reduces to the simple form of Eq. (13). In contrast, if the j-th and k-th atoms are of the same species, we end up with the formula in Eq. (15). It must be noted that, not only is this expression crucial to enforce the role of identical atoms, but it also roughly halves the number of free coefficients in the expansion. Finally, we conclude by noting that, while in the case of the 3B expansion there is no difference between the symmetrization in Eq. (16) and the lexicographic order introduced for the ACE coefficients (see Ref. [21] for details), these are indeed different in the generalization to the 4B case, as will be shown in Sec. II E.

D. Linear Scaling and the JL Atomic Basis

Before presenting the 4B-JL expansion, we discuss here the scaling of the 3B-JL expansion, with respect to the number of neighbors inside the cut-off volume. Indeed, by inserting Eq. (16) into the expression of Eq. (11) for the 3B energy, E_3, it is clear that the computational time to evaluate the 3B-JL expansion scales quadratically with the number of neighbours surrounding a central atom. This is because one has to explicitly look around for all possible pairs of atoms. Such scaling makes the formalism unpractical, when the number of atoms inside the cut-off sphere becomes relatively large. Most of the MLPs used in literature have solved this problem by achieving linear scaling with respect to the number of neighbors. Importantly, the 3B-JL expansion, being strictly tied to the powerspectrum components [22], can also be rearranged to reach the same scaling. In this rather technical section, we will mainly discuss the results of such "linearization", laying the formalism for a similar discussion in the 4B case. The formal derivation is then presented in the Supplemental Material (SM), Ref. [32]. As such, what is presented here can be considered a short review of the results obtained for other MLPs, in particular, for the power-spectrum case. Crucially, we will maintain the equivalence with the "internal coordinates representation" of the 3B term, Eq. (16), so one could freely move between the linear scaling formalism and the internal coordinate one, the latter being more advantageous for a small number of neighbors inside the cut-off volume.

We start by remarking that the choice of Legendre polynomial as an expansion basis was primarily driven by its natural decomposition in terms of a sum of products of spherical harmonics, Y_l^m, namely,

$$P_l(\hat{\boldsymbol{r}}_1\cdot\hat{\boldsymbol{r}}_2)=\frac{4\pi}{2l+1}\sum_{m=-l}^{l}(-1)^m Y_l^m(\hat{\boldsymbol{r}}_1)Y_l^{-m}(\hat{\boldsymbol{r}}_2)\,. \tag{17}$$

By exploiting this property, and by combining Eq. (16) with Eq. (11), one can prove that the 3B local energy term, $\varepsilon_i^{(3)}$, (defined so $E_3=\sum_i\varepsilon_i^{(3)}$), can be written as

$$\varepsilon_i^{(3)}=\sum_{\substack{Z_1Z_2\\Z_1\geq Z_2}}\sum_{n_1n_2l}^{\text{unique}} b^{Z_1Z_2Z_i}_{n_1n_2l}\left[C^{(3),Z_1Z_2}_{in_1n_2l}-S^{(3),Z_1Z_2}_{in_1n_2}\right], \tag{18}$$

where the first sum runs over the atomic species present in the system. The coefficients $b^{Z_1Z_2Z_i}_{n_1n_2l}$ are simply proportional to $a^{Z_1Z_2Z_i}_{n_1n_2l}$, as shown in the SM [32], so the equivalence in going from Eqs. (11)-(13) to Eq. (18), is preserved. We refer to the coefficient $C^{(3),Z_1Z_2}_{in_1n_2l}$ as the coupling term, which is obtained by including, on top of proper pairs of neighbor atoms, also the degenerate terms in which the central atom is allowed to interact twice with the same atom in the environment, namely, we accept the cases $(j,j)_i$ in the sum of Eq. (11). These "self-interacting" terms, $S^{(3),Z_1Z_2}_{in_1n_2}$, must be then removed, and so they are subtracted in Eq. (18).

Explicitly, the coupling and the self-interacting term are written as

$$C^{(3),Z_1Z_2}_{in_1n_2l}=\frac{4\pi}{2l+1}\sum_{m=-l}^{l}(-1)^m A^{Z_1}_{in_1lm}A^{Z_2}_{in_2l-m}\,, \tag{19}$$

and

$$S^{(3),Z_1Z_2}_{in_1n_2}=\delta_{Z_1Z_2}\sum_{j\in Z_1}\overline{P}^{(\alpha,\beta)}_{n_1ji}\overline{P}^{(\alpha,\beta)}_{n_2ji}\,, \tag{20}$$

where $\delta_{Z_1Z_2}$ is the Kronecker-delta. Here, we have adopted a "species-wise" atomic basis from the one defined for the ACE potential (see Ref. [5]), namely,

$$A^{Z}_{inlm}=\sum_{j\in Z}\overline{P}^{(\alpha,\beta)}_{nji}Y_l^m(\hat{\boldsymbol{r}}_{ji})\,, \tag{21}$$

where the radial basis has been specialized to the double-vanishing Jacobi polynomials. Also, we note that Eq. (19) is proportional to the power-spectrum components [22] or to the analogous rotationally invariant product $B^{(2)}_{in_1n_2l}$ introduced for the ACE potential. The crucial point here is that the coupling term in Eq. (19) is written over the species-wise atomic basis of Eq. (21). Since the A^Z_{inlm} basis scales linearly with the number of neighbors of the i-th atom, then we can evaluate the coupling term with a linear cost. This, together with the fact that the self-energy also scales linearly with respect to the number of neighbors, makes the computational scaling of the entire local energy, Eq. (18), linear in the numbers of neighbor atoms.

Incidentally, we note that we can write the product of the double-vanishing Jacobi polynomials in Eq. (20) in terms of a linear combination of double-vanishing Jacobi polynomials, namely, there are coefficients $c_n^{n_1n_2}$, such that

$$\overline{P}^{(\alpha,\beta)}_{n_1ji}\overline{P}^{(\alpha,\beta)}_{n_2ji}=\sum_{n=2}^{n_1+n_2}c_n^{n_1n_2}\overline{P}^{(\alpha,\beta)}_{nji}\,. \tag{22}$$

The coefficients $c_n^{n_1n_2}$ are usually calculated by numerical integration. This shows that the self-energy term can be re-casted as a linear combination of A^Z_{in00} too, and that it is, as expected, an effective 2B contribution. However, given the possible different cut-off radii of the 2B and 3B potentials and the relative different truncation, $n_{\max}$, we will keep the body orders as formally separated as possible, and we will not absorb the self-interaction terms back into lower body orders [33].

Let us now define a practical extension of the atomic basis A^Z_{inlm}, so to simplify the discussion for higher-order terms. We define the JL-atomic basis as

$$(J_pL_q)^{i,Z}_{n_1\ldots n_pl_1m_1\ldots l_qm_q}=\sum_{j\in Z}\left[\prod_{r=1}^{p}\overline{P}^{(\alpha,\beta)}_{n_rji}\right]\left[\prod_{s=1}^{q}Y_{l_s}^{m_s}(\hat{\boldsymbol{r}}_{ji})\right]. \tag{23}$$

This also includes the atomic basis A^Z_{inlm}, since

$$(J_1L_1)^{i,Z}_{nlm}=A^Z_{inlm}\,. \tag{24}$$

However, the definition in Eq. (24) allows us to take more than one double-vanishing Jacobi and one Legendre polynomial at once.

By looking at Eq. (22) (the same property holds for the Legendre polynomials) one could appreciate how all the components of Eq. (23) can be reduced to linear combinations of A^Z_{inlm}. Therefore, the definition of the JL-basis could appear unnecessary. However, since the coefficients $c_n^{n_1n_2}$ must be evaluated by integration, it can be more convenient to directly use the JL-atomic basis rather than performing the necessary integrations and contractions. It is important to note that evaluating the elements of the JL-atomic basis is still linear with the number of neighbors of the i-th atom: the only scaling affected is in terms of the number of the components involved, namely, the number of polynomials in the product.

Finally, we can now write the coupling term and the self-energy over the JL-atomic basis as

$$\begin{cases} C^{(3),Z_1Z_2}_{in_1n_2l}=\dfrac{4\pi}{2l+1}\displaystyle\sum_{m=-l}^{l}(-1)^m(J_1L_1)^{i,Z_1}_{n_1lm}(J_1L_1)^{i,Z_2}_{n_2l-m}\,,\\ S^{(3),Z_1Z_2}_{in_1n_2}=\delta_{Z_1Z_2}(J_2L_0)^{i,Z_1}_{n_1n_2}\,. \end{cases} \tag{25}$$

The JL-atomic basis will be used as the general framework for the analogous analysis of the linear scaling in the 4B case.

E. The Four-Body Term

For the 4B case we will follow the very same steps presented for the 3B one. We start by expanding the 4B energy contribution, E_4, as

$$E_4 = \sum_{i}^{\text{atoms}} \sum_{(j,k,p)_i} v^{(4)}_{jkpi}(r_{ji}, r_{ki}, r_{pi}, s_{jki}, s_{kpi}, s_{jpi}) \;, \tag{26}$$

where, analogously to E_3 in Eq. (11), the second sum runs over all the triplets of atoms in the neighborhood of the i-th atom. As for the 3B case, $v^{(4)}_{jkpi}$ is a shorthand form for $v^{(4)}_{Z_j Z_k Z_p Z_i}$.

The 4B potential, $v^{(4)}$, depends on three distances and three angles, so any triplets of atoms in the neighborhood of the i-th one is uniquely determined up to a reflection. The JL-4B expansion is then simply obtained by generalizing Eq. (16) to the case in which we have three double-vanishing Jacobi polynomials and as many Legendre polynomials, namely,

$$\begin{aligned} & v^{(4)}_{jkpi}(r_{ji}, r_{ki}, r_{pi}, s_{jki}, s_{kpi}, s_{jpi}) \\ & = \sum_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}}^{\text{unique}} a^{jkpi}_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1 ji} \overline{P}^{(\alpha,\beta)}_{n_2 ki} \overline{P}^{(\alpha,\beta)}_{n_3 pi} P^{jki}_{l_1} P^{jpi}_{l_2} P^{kpi}_{l_3} \right) . \end{aligned} \tag{27}$$

The range of the Jacobi indexes is again $[2, n_{\max}]$, while that of the Legendre ones is $[0, l_{\max}]$, where both $n_{\max}$ and $l_{\max}$ require optimization. By adopting the same formalism of Eq. (16), the symmetries of the potential are implemented in the expansion by construction. As for the 3B case, we have that $a^{jkpi}_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}}$ is a shorthand for $a^{Z_j Z_k Z_p Z_i}_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}}$.

It is useful to explicitly investigate the symmetries for the case in which the atoms in the neighborhood belong to the same species. By associating the Jacobi indexes n_1, n_2 and n_3 to the first, second, and third distances, respectively, and analogously associating the Legendre indexes to the scalar products, we impose the following symmetries on the expansion coefficients:

$$\begin{aligned} a_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}} &= a_{\substack{n_2 n_1 n_3 \\ l_1 l_3 l_2}} = a_{\substack{n_3 n_2 n_1 \\ l_3 l_2 l_1}} \\ &= a_{\substack{n_1 n_3 n_2 \\ l_2 l_1 l_3}} = a_{\substack{n_2 n_3 n_1 \\ l_3 l_1 l_2}} = a_{\substack{n_3 n_1 n_2 \\ l_2 l_3 l_1}} \;. \end{aligned} \tag{28}$$

The first identity states that, when exchanging the first two atoms, we have to simultaneously exchange the relative distances from the central atom (swapping the Jacobi indexes n_1 and n_2) and the angles formed with the remaining atom (exchanging the Legendre indexes l_2 and l_3). All the other identities can be interpreted in a similar way. The equivalences in Eq. (28) give us the unique set of indexes to use in Eq. (27), so the number of parameters to learn is reduced by roughly a factor of 6. The second sum in Eq. (27) will then run over all the indexes permutations involved in Eq. (28), mostly resulting in a sum of six terms, similarly to what was explicitly shown in Eq. (15).

We conclude this section by remarking that the use of double-vanishing polynomials in the 4B-JL expansion allows us to implement an even more severe reduction in the number of free coefficients compared to the 3B case.

F. 4B Linear Scaling: connection with the Bispectrum

A linear scaling with the number of atoms in the neighborhood volume can also be achieved for the 4B case. Indeed, this is even more important than for lower-body orders, since otherwise the scaling would be cubic with the number of neighbors. The backbone of the demonstration is similar to the one adopted for the 3B case, so, by using the property of Eq. (16) and the JL-atomic basis defined in Eq. (23), we can write the local energy term, $\varepsilon^{(4)}_i$, as

$$\begin{aligned} \varepsilon^{(4)}_i = & \sum_{Z_1 \geq Z_2 \geq Z_3} \sum_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}}^{\text{unique}} b^{Z_1 Z_2 Z_3 Z_i}_{\substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}} \\ & \times \left[C^{(4), Z_1 Z_2 Z_3}_{i, \substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}} - S^{(4), Z_1 Z_2 Z_3}_{i, \substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}} \right] , \end{aligned} \tag{29}$$

where the coupling term for the 4B is given by

$$\begin{aligned} C^{(4), Z_1 Z_2 Z_3}_{i, \substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}} = & \frac{(4\pi)^3}{(2l_1+1)(2l_2+1)(2l_3+1)} \\ & \times \sum_{m_1 m_2 m_3} (-1)^{m_1+m_2+m_3} (J_1 L_2)^{i, Z_1}_{n_1 l_1 m_1 l_2 - m_2} \\ & \times (J_1 L_2)^{i, Z_2}_{n_2 l_3 m_3 l_1 - m_1} (J_1 L_2)^{i, Z_3}_{n_3 l_2 m_2 l_3 - m_3} \;. \end{aligned} \tag{30}$$

The corresponding expression for the self-energy, $S^{(4)}_i$, is more involved and, for the sake of brevity, is reported in the SM, Ref. [32]. Here, we just wish to mention that it is obtained from linear combinations of products of the basis terms $(J_1 L_2)$, $(J_2 L_2)$, and $(J_3 L_0)$.

The coupling scheme described in Eq. (30) differs from the bispectrum-component coupling scheme [6, 22], while being strictly related to it. Indeed, the bispectrum writes in the ACE flavour [5] as

$$\begin{aligned} B^{(3), Z_1 Z_2 Z_3}_{i \, \substack{n_1 n_2 n_3 \\ l_1 l_2 l_3}} = & \sum_{m_1 m_2 m_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \\ & \times A^{Z_1}_{i n_1 l_1 m_1} A^{Z_2}_{i n_2 l_2 m_2} A^{Z_3}_{i n_3 l_3 m_3} \;, \end{aligned} \tag{31}$$

where the $3j$-Wigner symbol [34] is introduced and A_{inlm} is the atomic basis of Eq. (21). Furthermore, the JL-atomic basis terms, $(J_1 L_2)$, can be written as a linear combination of the A^Z_{inlm}:

$$\begin{aligned}(J_1L_2)^{i,Z}_{nl_1m_1l_2m_2} &\\ = \sum_{lm}(-1)^m\sqrt{\frac{(2l+1)(2l_1+1)(2l_2+1)}{4\pi}} &\qquad (32)\\ \times\begin{pmatrix} l_1 & l_2 & l\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix} l_1 & l_2 & l\\ m_1 & m_2 & -m\end{pmatrix}A^Z_{inlm}\,. &\end{aligned}$$

This is directly derived from the product rule for two spherical harmonics. From this expression, one could write the coupling terms $C^{(4)}_i$ as a linear combination of bispectrum components $B^{(3)}_i$ (see SM, Ref. [32]). This linear combination represents a way of combining the bispectrum components so the final result is explicitly written in terms of internal coordinates only. Crucially, our argument shows that adopting the coupling scheme in $C^{(4)}_i$ could give an advantage over the bispectrum components, since it allows us to maintain the equivalence between the expression in Eq. (30) and the analogous one from Eqs. (26) and (27). Then, an intuitive and equivalent closed expression (in terms of internal coordinates) remains available for any case in which the number of atoms in the neighbours is relatively small, so one could opt between Eq. (29) and Eqs. (26) and (27) as needed.

G. The Five-Body Term

The 5B term can be obtained by direct generalization of the 4B case. Indeed, the procedure is analogous, namely, the energy contribution, E_5, is partitioned in local components, which consist of a sum of local 5B potentials, $v^{(5)}_{jkpqi}$. These depend on four distances and six angles, so they can be expanded as a linear combination of products of four double-vanishing Jacobi polynomials and six Legendre polynomials. The resulting expression is analogous to the one obtained in Eq. (18) for the 4B case. The symmetry properties of the potentials are also treated in the same way, resulting in a reduction of the number of coefficients up to roughly a factor of 24 when dealing with identical atoms.

H. behavior at the origin

A common practice in MLPs is to introduce an external function to impose a repulsive behavior when the interatomic distance becomes small. Here, since we are using the double-vanishing Jacobi polynomials for all body orders beyond two, the only term affecting the behavior at small distances is the 2B, given in Eq. (5). We can then obtain some insight into the behavior of the potential by evaluating Eq. (5) at the origin. Indeed, if $r_{\min}=0$, we obtain

$$v^{(2)}_{Z_jZ_i}(0)=\sum_{n=1}^{n_{\max}} a_n^{Z_jZ_i}\widetilde{P}^{(\alpha,\beta)}_n(1)\,. \qquad (33)$$

From the identity

$$\widetilde{P}^{(\alpha,\beta)}_n(1)=\binom{n+\alpha}{n}-(-1)^n\binom{n+\beta}{n}\,,$$

we can conclude that the magnitude of the potential at the origin can become very large for a high enough n. Therefore, by biasing the hyper-parameters so the potential is positive at the origin, we can produce a strongly repulsive behavior almost by construction, with no use of any external function.

This observation must be checked on a case-by-case base, an operation that can be performed visually by simply looking at the potential. Indeed, once the best expansion coefficients are available, it is possible to plot the function

$$v^{(2)}_{Z_jZ_i}(x)=\sum_{n=1}^{n_{\max}} a_n^{Z_jZ_i}\widetilde{P}^{(\alpha,\beta)}_n\left(\cos\left(\pi x/r_{\rm cut}\right)\right)\,, \qquad (34)$$

and analyze the behavior near the origin. Since small distances are usually in an extrapolation region of the potential, with little to no data corresponding to such distances present in the training set, a visual investigation of the 2B potential could also return us some intuition on the behavior of the model when dealing with extrapolation attempts to unseen atomic distributions.

I. Forces and Stress

In this section, we outline the general recipe to calculate the forces and the virial-stress tensor. Given the linearity of the expressions associated with the JL expansion, one only needs the derivative of the (double-)vanishing Jacobi and of the Legendre polynomials, from which all the relevant quantities can be evaluated.

Since the multi-body expansion of the energy, Eqs. (1), and the fact that E_1 is just an energy offset, the n-body contribution to the force of an atom at position $\boldsymbol{r}_a$, is given by

$$\boldsymbol{F}^{(n)}_a=-\frac{\partial E_n}{\partial \boldsymbol{r}_a}\,, \qquad (35)$$

whereas the total force is obtained by summing over all the n-body contributions, $\boldsymbol{F}_a=\sum_n\boldsymbol{F}^{(n)}_a$.

As it can be seen from Eq. (4) and (5), the evaluation of the 2B force contribution, $\boldsymbol{F}^{(2)}_a$, requires only the application of the chain rule and the derivative of the vanishing polynomials, namely,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\widetilde{P}^{(\alpha,\beta)}_n(\cos(x)) = \frac{\mathrm{d}}{\mathrm{d}x}P^{(\alpha,\beta)}_n(\cos(x)) = &\qquad (36)\\ = -\frac{\alpha+\beta+n+1}{2}\sin(x)P^{(\alpha+1,\beta+1)}_{n-1}(\cos(x))\,. &\end{aligned}$$

This expression shows that the derivative of the potential smoothly vanishes ($\boldsymbol{F}^{(2)}_a=\boldsymbol{0}$) at the origin and at the cutoff radius (when $x=0,\pi$). Furthermore, from Eq. (36)

we can appreciate that the force can be written solely in terms of Jacobi polynomials. This results in a linear expansion that can be easily implemented or analytically investigated.

Analogously, we can evaluate the 3B contribution to the forces by differentiating the E_3 term. This implies that we need to calculate [see Eqs. (13) and (16)]

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{r}_a}\sum_{\text{symm.}}\left(\overline{P}^{(\alpha,\beta)}_{n_1 ji}\overline{P}^{(\alpha,\beta)}_{n_2 ki}P^{jki}_{l}\right)= \\ &=\sum_{\text{symm.}}\frac{\partial}{\partial \boldsymbol{r}_a}\left(\overline{P}^{(\alpha,\beta)}_{n_1 ji}\overline{P}^{(\alpha,\beta)}_{n_2 ki}P^{jki}_{l}\right), \qquad (37)\end{aligned}$$

where we are able to exchange the sum and the derivative, since the former acts only on the Jacobi and Legendre indexes. Therefore, in evaluating the derivative of the product, we can use again the chain-rule and the differentiation formula for the Legendre polynomials, namely,

$$\frac{\mathrm{d}}{\mathrm{d}x}P_l(x)=\frac{\mathrm{d}}{\mathrm{d}x}P^{(0,0)}_l(x)=\frac{l+1}{2}P^{(1,1)}_{l-1}(x)\,, \qquad (38)$$

where we have used the fact that the Legendre polynomials are obtained from the Jacobi polynomials by setting $\alpha=\beta=0$. Finally, we also need the differentiation rule for double-vanishing Jacobi polynomials

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}x}\overline{P}^{(\alpha,\beta)}_{n}(\cos(x))= \\ &=-\frac{\sin(x)}{2}\Bigg((\alpha+\beta+n+1)P^{(\alpha+1,\beta+1)}_{n-1}(\cos(x)) \\ &\qquad -(\alpha+\beta+2)\frac{\widetilde{P}^{(\alpha,\beta)}_{n}(-1)}{\widetilde{P}^{(\alpha,\beta)}_{1}(-1)}\Bigg)\,. \qquad (39)\end{aligned}$$

The 4B and 5B contributions to the forces are evaluated in the same way, and these do not introduce any further ingredient to obtain an analytical form. The expression for the forces in the case of the JL-atomic basis will be explicitly discussed in future works. Finally, we can also obtain the virial-stress tensor by means of the formula discussed in Ref. [35] [see Eq. (25)].

J. Linear Regression

To select the optimal expansion coefficients for each body term, we minimize the widely used loss function

$$L=\|\boldsymbol{E}-\boldsymbol{J}_E\boldsymbol{a}\|_2^2+c_F\|\boldsymbol{F}-\boldsymbol{J}_F\boldsymbol{a}\|_2^2+c_W\|\boldsymbol{W}-\boldsymbol{J}_W\boldsymbol{a}\|_2^2\ ,$$

where the vector $\boldsymbol{E}$ represents all the energies in the training set (obtained by *ab-initio* calculations), $\boldsymbol{a}$ is the vector of all the coefficients of the expansion, $\boldsymbol{J}_E$ is the matrix, whose rows contain the set of descriptors for one configuration of the training set. Similarly $\boldsymbol{F}$ is the vector of all the forces of the dataset, while $\boldsymbol{J}_F$ are the appropriate differentiated descriptors. Note that, explicitly, we will train on each component of the forces for each atom in the system. This means that, if the i-th configuration has N_i atoms, we will have $3N_i$ forces associated with that configuration. The vector of the components of the stress tensor for each training point is $\boldsymbol{W}$. In this case, we will train independently on each of the six components, for any of the configurations in the training set. Finally, c_F and c_W are coupling constants to be optimized, and $\|\cdot\|_2^2$ is the square of the vector two-norm. While the use of a multi-target scheme, embedded in a non-linear function, can be used to increase the accuracy of the model [36], we remark that we follow here a simple linear approach.

	Two body	Three body	Four body
$n_{\max}$	10	6	4
$l_{\max}$	–	5	3
r_{cut} (Å)	3.7	3.7	3.7
$\alpha=\beta$	1	1	1
No. of features	10	90	364

TABLE I. Details of the JLP trained on the carbon dataset from Ref. [25]. In order to reduce the number of hyperparameters, we fix α and β to be equal, and $r_{\min}=0$. The model is relatively compact and comprises 465 (464 plus the intercept) features.

The minimization procedure that will be adopted for the remainder of this paper, where results on a mono-species system are shown, will be based on the singular value decomposition (SVD). We stress that we will not regularize the energy offset, E_1. Furthermore, instead of using the total energies, we will always consider the energy per atom in the training set.

Here we wish to remark that the coupling constants, c_F and c_W, can also depend on the specific configuration, a fact that can be seen as a configuration wise rescaling of the descriptors and targets. This is useful, in particular, when the configurations have a different number of atoms. As a direct example, the loss function used in the next section, will have all the forces and the relative descriptors divided by $\sqrt{3N_i}$, where N_i is the number of atoms in the configuration. This is performed in order to weigh the energies, forces, and stress, on a similar footing in the minimization procedure. Another advantage of such a configuration-wise weighting scheme is that the energy offset per atom, E_1, can be written analytically in terms of the per-atom average energy, average descriptors, and the linear fitting coefficients.

III. A JLP FOR CARBON

As an application of the method described here, we have fitted a JLP on the carbon data set used to fit the GAP17 potential of Ref. [25]. We have opted for this dataset, since it presents several challenges. Firstly,

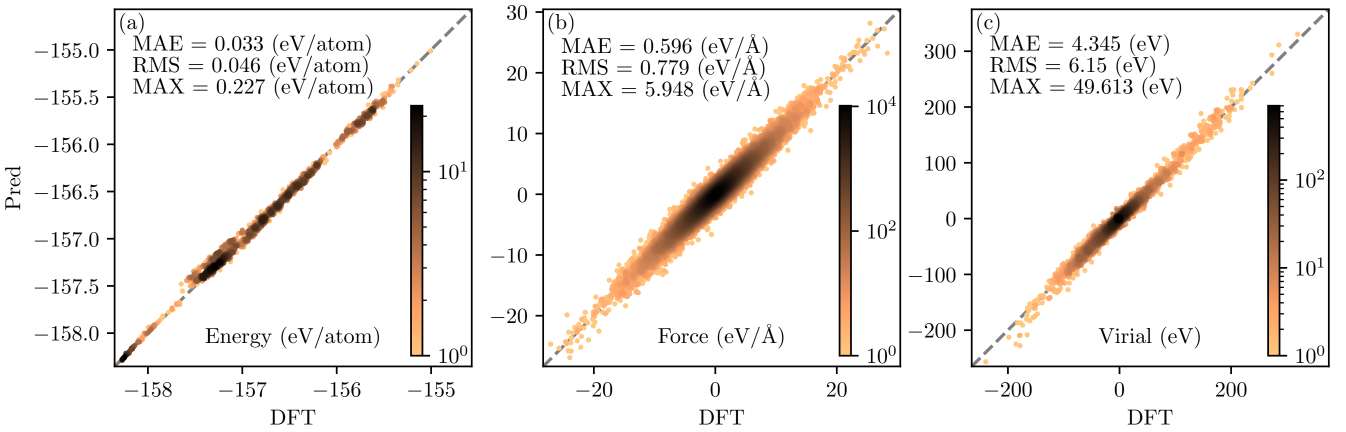

FIG. 5. Parity plots computed over the test set for the (a) energies, (b) forces, (c) virial stress. The mean absolute errors (MAEs) and root mean square error (RMSE) are reported for each plot, alongside the error on the worst prediction. The color code indicates the data density (number of points).

the dataset is made of different phases of carbon, ranging from crystalline structures (graphene, graphite, diamond), to surfaces and amorphous phases. In addition, some phases present a relative large distance for the decay of the forces between two atoms, as explicitly shown in the same Ref. [25]. This is mirrored in the choice of the appropriate cut-off radius. For the fit we have removed all the carbon dimers and any structures with absolute maximum force components greater than 30 eV/Å. In total we have thus removed 37 structures of which 30 are the carbon dimers used to fit the 2B GAP and seven other structures, which do not satisfy the maximum force criteria. The remaining 4,043 structures are split into a training set of 2,830 and a testing one of 1,213. The structure index of all the training and testing structures are given in the SM, Ref. [32].

We use energy, forces, and virial stress to fit the linear model. The hyper-parameters for the final potential are summarized in Table I. Following the analysis on the locality of Ref. [25], we have kept the same cut-off radius as for the GAP17 model, namely, 3.7 Å. The coupling constant c_F and c_W, of the loss function, are 0.5 and 0.075, respectively. Finally, the descriptors have been calculated in their internal coordinate form and the cluster expansion is truncated at the 4B order. This gives us a potential defined over 465 features.

For the fitted model, we find that the training-set root mean squared errors (RMSEs) are 43.9 meV/atom for the energy, 0.781 eV/Å for the forces and 6.62 eV for the stress. As shown in Fig. 5, reporting the parity plots for the test set, the corresponding RMSEs are 46.6 meV/atom for the energy, 0.779 eV/Å for the forces and 6.15 eV for the stress, namely, they are of the same quality as for the training set (the parity plots for the training set are reported in the SM, Ref. [32]). We observe that the structures, which deviate the most from the energy-parity plot in Fig. 5(a), correspond to all carbon in the amorphous phase. These appear to be slightly more difficult to be dealt with by the JLP. Furthermore, we wish to remark that, as can be appreciated in Fig. 5(c), the predicted components of the virial-stress appear to be in good agreement with the DFT ones.

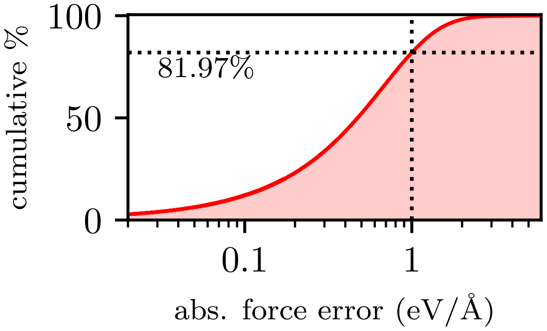

FIG. 6. The cumulative distribution of the test-set predicted forces. The model shows that approximately 81.97% of the structures have an error below 1 eV/Å.

In Fig. 6, we report the cumulative distribution of the error on the forces for the test set. The curve represents the percentage of structures, which have an error below the one indicated. As a reference, we explicitly consider the case of 1 eV/Å, which was taken as reference for the GAP17 potential (see Ref. [25]). The remarkably high value of 81.97%, compared to the 68.3% of the GAP17 potential, shows the capability of the JLP in correctly predicting the force components.

As remarked in Section II H, the JL potential naturally shows a repulsive behavior at a short distance, without the inclusion of any external fast-varying function. This is made clear in Fig. 7, where we show the C-C potential obtained by plotting Eq. (34) with the fitted 2B-

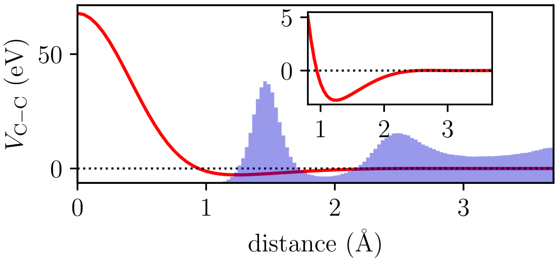

FIG. 7. Reconstruction of the 2B potential from Eq. (34) (red curve). The inset shows a magnification around the minimum, while the histogram reports the pair-distance distribution of the entire data set. Qualitatively, the potential shows a strong repulsive behavior for small distances and a minimum, which is consistent with the position of the first peak in the pair-distances distribution.

expansion coefficients. From the figure one can clearly identify the short-distance repulsive behavior of the 2B potential, which arises naturally from the 2B expansion coefficients. Furthermore, the potential shows a shallow minimum close to the first peak in the radial distribution function computed over the fitting data set (blue shadow). We stress here once again that the repulsive behavior is completely determined by the 2B coefficients, since all the other body-order terms are written in term of the double-vanishing Jacobi polynomials, which vanish at short distances. As a consequence, Fig. 7 gives us complete information about the repulsive behavior of the entire potential.

We then employed the optimized JLP, to predict the phonons dispersion curves for graphene and diamond, using the PHONO3PY package [37, 38]. The results are reported in Fig. 8, where the reference phonon dispersion for crystalline diamond (mp-66) was obtained from materials project [39] and for graphene was obtained from the phonon website [40]. These reference calculations have been performed using density functional perturbation theory and the ABINIT code [41].

As one can appreciate from the figure, the agreement between the JLP-computed phonon bands and the DFT reference ones is quite remarkable, for both the acoustic and optical branches, with no negative frequencies present at the Γ points. The largest disagreement is generally found for the optical branches and it is of the order of 2 cm^{-1} (see, for instance, the graphene bands at around 45 cm^{-1}). Note that this is a particularly challenging test, since the training data set has an energy spread of several eV/atom, while the energy differences computed in the finite-difference scheme used here are a few meV/atom from the equilibrium energy. This means that our JLP is able to describe, on the same footing, both the low-energy physics of crystalline carbon around equilibrium, and high-energy liquid and amorphous structures. Qualitatively, the performance of the JLP is much closer to the one of the more recent and accurate GAP20 potential, as it can be seen in the Supplemental Material of Ref. [42]. This result is achieved despite the fact that GAP20 was trained on an extended version of the GAP17 data set, which specifically includes, among many others, more graphene data points (from Ref. [43]). Here we stress again that we did not add any features to the 465 of Table I, and we did treat the low energy configurations on the same footing of the high energy ones, with the weights depending only on the number of atoms in the configuration, as clarified at the end of the previous section. Note also that perfect agreement is not even expected. In fact, the DFT data set used to train the JPL model was obtained with the CASTEP code [44] and the phonon via finite differences, while our DFT reference has been generated with ABINIT [41] and density functional perturbation theory. Additional differences can also be ascribed to the different pseudopotentials used and to details in the DFT implementation.

IV. IMPLEMENTATION DETAILS

The JLP is implemented in a Python module, which is currently undergoing optimization in sight of a future release. The computationally intensive part of calculating the descriptors and the polynomial expansions is implemented in CYTHON. The current implementation is serial, and, for example, takes 26.8 ms to compute the descriptors (energies, forces, and stress), for a randomly selected periodic structure from the training set used in the previous section. Such structure contains 64 atoms in the unit cell, while the calculations have been performed on an Intel i7-9600 processor system with 16GB RAM. However, we wish to remark that, since the JL formalism presented in this paper keeps the evaluation of the descriptors for each local environment independent from the others, future effort will point towards making these calculations running in parallel, as they are the bottleneck of our current implementation. We therefore want to stress that the calculation time provided here is not yet informative of the actual optimal performance of the JL descriptors, as it can still be significantly reduced.

V. CONCLUSIONS

In conclusion, we have introduced all the necessary formalism to develop a general cluster expansion for the total energy, where the different body-order terms are systematically separated. This is designed for the short-range chemical-bond-related part of the total energy, which is written as the sum of individual atomic contributions. The core idea is that of expanding the different body-order terms, representing the inter-atomic distances over Jacobi polynomials and the structural angles over Legendre polynomials, which are a special case of the Jacobi ones. This is an extremely general representation, giving us ample flexibility when constructing the

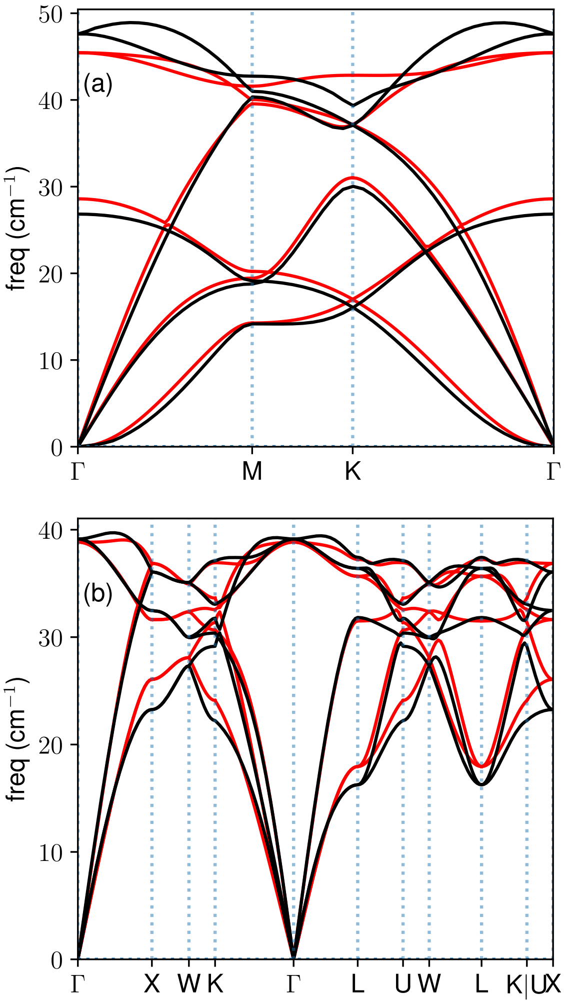

FIG. 8. Phonon spectra for (a) graphene and (b) diamond computed with the optimised JLP described in the text (red lines). The reference DFT calculations (black lines) have been obtained with density functional perturbation theory as implemented in the ABINIT code.

potential.

An important feature is that one can impose both constraints and symmetries on the coefficients of the expansion, a practice that allows us to implement desired behaviors of the potential in a natural way. For example, one can impose the potential to vanish at a desired cut-off distance, by simply imposing a set of conditions over the zero-order coefficients of the Jacobi polynomial expansion. In a similar way, one can constraint the expansion of all the body-order terms larger than two to vanish at the origin, so the short-distance behavior of the potential is solely determined by the 2B contribution. This, in turn, can be designed to display repulsive behavior at short distances.

Furthermore, the implementation of physical symmetries over the expansion allows us to drastically reduce the number of independent coefficients to determine. Also, the calculation of the descriptors that define the cluster expansion, is proved to scale linearly with the number of atoms in the cut-off volume. The demonstration of such scaling is rooted in the decomposition of the Legendre polynomials over spherical harmonics, a feature that allows us to map our representation on known many-body atomic bases such as the powerspectrum, the bispectrum and those introduced in the ACE method.

The formalism introduced here is put to the test for a quite complex dataset, namely, the carbon one used to construct the GAP17 potential. This comprises crystalline graphite and diamond, as well as a multitude of liquid and amorphous carbon structures. We then show that a 4B relatively compact model, containing 465 features, and trained over energies, forces and stress tensor, is capable of achieving extremely competitive RMSEs across all quantities. Furthermore, the same potential reproduces, quite accurately, the zero-temperature phonon band structure of both graphene and diamond, demonstrating accuracy at both low and high energies. We believe that the JLP introduced here adds to the burgeoning field of MLPs, bringing a versatile tool where symmetry and constraints can be implemented in a natural and efficient way. The ability to separate the different body orders and the possibility to construct relatively compact models, make the JLP a strong candidate for the calculation of PES both in data-rich and data-poor situations.

ACKNOWLEDGMENTS

This paper has been supported by the Irish Research Council Advanced Laureate Award (No. IRCLA/2019/127), and by the Irish Research Council postgraduate program (MC). U.P. acknowledges the Qatar National Research fund for additional financial support (Grant No. NPRP12S-0209-190063). We acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing (ICHEC) and Trinity Centre for High Performance Computing (TCHPC) for the provision of computational resources. We acknowledge support from ICHEC via the academic flagship program (Project No. - EuroCC-AF-3). We have extensively used NUMPY [45], CYTHON [46], SCIKIT-LEARN [47] and MATPLOTLIB [48].

Appendix A: Proof of property (10)

We derive here a series expansion for the vanishing Jacobi polynomials and we will prove the property Eq. (10). The series expansion for the Jacobi Polynomials is (Ref. [27])

$$P_n^{(\alpha,\beta)}(x) = \frac{1}{2^n}\sum_{j=0}^{n}\binom{n+\alpha}{j}\binom{n+\beta}{n-j}(x-1)^{n-j}(x+1)^j\,. \tag{A1}$$

Performing the substitution $x \to \cos x$, where $x = \pi(r - r_{\min})/(r_{\text{rcut}} - r_{\min})$, we get

$$P_n^{(\alpha,\beta)}(\cos x) = \sum_{j=0}^{n} (-1)^{n-j} \binom{n+\alpha}{j} \binom{n+\beta}{n-j} \sin^{2(n-j)}(x/2) \cos^{2j}(x/2) \,. \tag{A2}$$

Evaluating the expression at $x = \pi$, which means evaluating the polynomial at the cut-off, makes all the terms of the summation vanish except for the case $j = 0$. Therefore, by means of Eq. (6), we get a series expansion for the vanishing Jacobi polynomials:

$$\begin{aligned} \widetilde{P}_n^{(\alpha,\beta)}(\cos x) = & \\ = \sum_{j=1}^{n} (-1)^{n-j} \binom{n+\alpha}{j} \binom{n+\beta}{n-j} \sin^{2(n-j)}(x/2) \cos^{2j}(x/2) & \\ -(-1)^n \binom{n+\beta}{n} (\sin^{2n}(x/2) - 1) \,. & \end{aligned} \tag{A3}$$

Finally, by using the identity

$$\sin^{2n}(x/2) - 1 = -\cos^2(x/2) \sum_{j=1}^{n} \sin^{2(n-j)}(x/2), \tag{A4}$$

we prove the property Eq. (10),

$$\widetilde{P}_n^{(\alpha,\beta)}(\cos x) = f_c(x) Q_n^{(\alpha,\beta)}(\cos(x)) \,, \tag{A5}$$

where

$$\begin{aligned} & Q_n^{(\alpha,\beta)}(\cos(x)) \\ & = \sum_{j=1}^{n} \left[(-1)^{n-j} \binom{n+\alpha}{j} \binom{n+\beta}{n-j} \cos^{2(j-1)}(x/2) \right. \\ & \qquad \left. + (-1)^n \binom{n+\beta}{n} \right] \sin^{2(n-j)}(x/2) \,, \end{aligned} \tag{A6}$$

and $f_c(x) = \cos^2(x/2) = (1 + \cos(x))/2$.

[1] J. Behler, *Four Generations of High-Dimensional Neural Network Potentials*, Chem. Rev. **121**, 16, 10037-10072, (2021).

[2] E. Kocer, T. W. Ko, J. Behler, *Neural Network Potentials: A Concise Overview of Methods*, Ann. Rev. Phys. Chem. **73**, 163-186, (2022).

[3] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Machine Learning Force Fields*, Chem. Rev. **121**, 10142 (2021).

[4] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A.V. Shapeev, A.P. Thompson, M.A. Wood and S.P. Ong, *Performance and Cost Assessment of Machine Learning Interatomic Potential*, J. Phys. Chem. A **124**, 731 (2020).

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

[6] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles and G. J. Tucker, *Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials*, J. Comp. Phys. **285**, 316 (2015).

[7] A.V. Shapeev, *Moment tensor potentials: A class of systematically improvable interatomic potentials*, Multiscale Modeling & Simulation **14**, 1153 (2016).

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

[9] 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).

[10] S. De, A.P. Bartók, G. Csányi and M. Ceriotti, *Comparing molecules and solids across structural and alchemical space*, Phys. Chem. Chem. Phys. **18**, 13754 (2016).

[11] M. Domina, M. Cobelli and S. Sanvito, *A spectral-neighbour representation for vector fields: machine-learning potentials including spin*, Phys. Rev. B **105**, 214439, (2022).

[12] A. Lunghi and S. Sanvito, *A unified picture of the covalent bond within quantum-accurate force fields: from simple organic molecules to metallic complexes' reactivity*, Science Advances **5**, eaaw2210 (2019).

[13] A. Lunghi and S. Sanvito, *Surfing multiple conformation-property landscapes via machine learning: Designing magnetic anisotropy*, J. Chem. Phys. C **124**, 5802 (2019).

[14] F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti and L. Emsley, *Chemical shifts in molecular solids by machine learning. Chemical shifts in molecular solids by machine learning*, Nat Commun **9**, 4501 (2018).

[15] A. Glielmo, P. Sollich, and A. De Vita, Phys. Rev. B **95**, 214302 2017).

[16] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Phys. Rev. Lett. 120, 036002 (2018).

[17] V. H. A. Nguyen and A. Lunghi, *Predicting tensorial molecular properties with equivariant machine learning models*, Phys. Rev. B **105**, 165131, (2022).

[18] F. Musil, A. Grisafi, A.P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, *Physics-Inspired Structural Representations for Molecules and Materials*, Chem. Rev. **121**, 9759 (2021).

[19] I. S. Novikov, K. Gubaev, E. V. Podryabinkin and A. V. Shapeev, *The MLIP package: moment tensor potentials with MPI and active learning*, Mach. Learn.: Sci. Technol. **2**, 025002 (2021).

[20] Y. Lysogorskiy, C.v.d. Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt,M. Mrovec, A. Thompson, G. Csányi, C. Ortner and R. Drautz, *Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon*, npj Comput Mater **97**, 7 (2021).

[21] G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, C. Ortner, *Atomic cluster expansion: Completeness, efficiency and stability*, J. Comp. Phys. **454**, 110946 (2022).

[22] A. P. Bartók, R. Kondor, G. Csányi, *On representing*

chemical environments, Phys. Rev. B **87**, 184115 (2013).

[23] N. C. Nguyen, A. Rohskopf, *Proper orthogonal descriptors for efficient and accurate interatomic potentials*, J. Comp. Phys. **480**, 112030 (2023).

[24] N. C Nguyen, *Fast proper orthogonal descriptors for many-body interatomic potentials*, Phys. Rev. B, **107**, 144103 (2023).

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

[26] B. Focassio, M. Domina, U. Patil, A. Fazzio and S. Sanvito, *Linear Jacobi-Legendre expansion of the charge density for machine learning-accelerated electronic structure calculations*, npj Comp. Mater. **9**, 87 (2023).

[27] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover, Mineola, NY, (1972), ch. 22.

[28] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Incompleteness of Atomic Structure Representations*, Phys. Rev. Lett. **125**, 16, 166001 (2020).

[29] M. Parsaeifard, D. S. De, A. S. Chrinstensen, F. A Faber, E. Kocer, S. De, J. Behler, O A. von Lilienfeld, and S. Goedecker, *An assessment of the structural resolution of various fingerprints commonly used in machine learning*, Mach. Learn.: Sci. Technol. **2**, 015018 (2021).

[30] B. Parsaeifard, and S. Goedecker, *Manifold of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions*, J. Chem. Phys. **156**, 034302 (2022).

[31] E. Kocer, J.K. Mason and H. Erturk, *Continuous and optimally complete description of chemical environments using Spherical Bessel descriptors*, AIP Adv. **10**, 015021 (2020).

[32] See Supplemental Material at [URL to be inserted by publisher] for a formal derivation of the linear scaling formalism, and for the different body-order contributions to the trained Carbon JLP model.

[33] D. P. Kovács, C. van der Oord, J. Kucera, A. E. A. Allen, D. J. Cole, C. Ortner, and G. Csányi, *Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE*, J. Chem. Theory Comput. **17**, 7696 (2021).

[34] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum Theory of Angular Momentum* (World Scientific, Singapore, 1988).

[35] A. P. Thompson, S. J. Plimpton, and W. Mattson, *General formulation of pressure and stress tensor for arbitrary many-body interaction potentials under periodic boundary conditions*, J. Chem. Phys. **131**, 154107 (2009).

[36] A. Bochkarev, Y. Lysogorskiy, S. Menon, M. Qamar, M. Mrovec, and R. Drautz, *Efficient parametrization of the atomic cluster expansion*, Phys. Rev. Materals, **6**, 013804 (2022).

[37] A. Togo, L. Chaput, and I. Tanaka, *Distributions of phonon lifetimes in Brillouin zones*, Phys. Rev. B, **91**, 094306 (2015).

[38] A Togo, *First-principles Phonon Calculations with Phonopy and Phono3py*, J. Phys. Soc. Jpn., **92**, 012001-1-21 (2023).

[39] 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, *Commentary: The Materials Project: A materials genome approach to accelerating materials innovation.* APL materials. **1**, 1 (2013).

[40] `https://henriquemiranda.github.io/phononwebsite/phonon.html?json=http://henriquemiranda.github.io/phononwebsite/localdb/graphene/data.json`

[41] G. Petretto, S. Dwaraknath, H. P. C. Miranda, D. Winston, M. Giantomassi, M. J. Van Setten, X. Gonze, K. A. Persson, G. Hautier and G. M. Rignanese. *High-throughput density-functional perturbation theory phonons for inorganic materials.* Sci Data, **5**, 180065 (2018).

[42] P. Rowe, V. L. Deringer, P. Gasparotto, G. Csányi, A. Michaelides, *An accurate and transferable machine learning potential for carbon*, J. Chem. Phys. **153**, 034702 (2020).

[43] P. Rowe, G. Csányi, D. Alfè, A. Michaelides, *Development of a machine learning potential for graphene*, Phys. Rev. B, **97**, 054303 (2018).

[44] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. J. Probert, K. Refson and M. C. Payne, *First principles methods using CASTEP*, Z. Kristallogr. **220**, 567 (2005).

[45] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, and R. Kern, *Array programming with NumPy*, Nature **585**, 357 (2020).

[46] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn, and K. Smith, *Cython: The best of both worlds.*, Comput. Sci. Eng. **13**, 31 (2010).

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine Learning in Python.*, J. Mach. Learn. Res. **12**, 2825 (2011).

[48] J. D. Hunter, *Matplotlib: A 2D graphics environment*, Comput. Sci. Eng. **9**, 90 (2007).