The Jacobi-Legendre potential

M. Domina*, U. Patil*, M. Cobelli*, and S. Sanvito
School of Physics and CRANN Institute, Trinity College, Dublin 2, Ireland
(Dated: August 23, 2022)

Inspired by the cluster expansion method, we introduce a compact machine-learning potential constructed over Jacobi and Legendre polynomials. The constituent many-body contributions are separated, fully interpretable and adaptable to replicate the physical knowledge of the system, such as a repulsive behaviour at a small inter-atomic distance. Most importantly the potential requires a small number of features to achieve accuracy comparable to that of more numerically heavy and descriptor-rich alternatives. This is here tested for an organic molecule, a crystalline solid and an amorphous compound. Furthermore, we argue that the physical interpretability of the various terms is key to the selection and training of stable potentials.

Machine-learning potentials (MLPs) are rapidly becoming the gold standard for molecular dynamics and thermodynamical sampling in materials science [1]. The general idea is that of performing a high-dimension fit of the potential energy surface computed with an electronic ab-initio method, for instance with density functional theory, to obtain a numerical atomic energy functional for large-scale simulations. In particular, one aims at using a limited number of electronic structure data to interpolate the potential energy surface 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 [2–4] or non linear [5–7]. Furthermore, the target quantity may be different from the energy and may include electronic properties [8–10].

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 improve the model accuracy is that of designing descriptors invariant with respect to the symmetries of the target quantity. In the case of the energy, this results in descriptors which are invariant for translations, rotations and permutations of identical atoms. Typically, there is a subtle tradeoff between the model complexity, the descriptor type and the size and composition of the dataset needed to construct the MLP. However, despite the large variety of descriptors and models developed so far [11], they all appear to be susceptible to a wide-spread problem in machine learning, namely their interpretability. In fact, the descriptors themselves are often complex many-body functions, while the training process is driven by the optimization of a global metric such as the energy mean absolute error. As such, a MLP is often treated as a black box, with little control and understanding of the physics learned by the model itself. This can easily lead to the generation of MLPs unable to extrapolate to data that have not been seen during the training, and one must rely on numerical techniques to establish whether a particular atomic configuration is interpolated or extrapolated by the model [12].

In this letter we introduce a novel linear model built over a set of descriptors derived from the energy cluster expansion. Our MLP, that we name the Jacobi-Legendre Potential (JLP), while close in spirit to the recently introduced Atomic Cluster Expansion (ACE) [3, 13], offers a more compact alternative requiring a small number of features to achieve a high level of accuracy. This results in small, data-efficient, models more similar in size to the spectral neighbour atomic potentials (SNAPs) [2]. Each body-order contribution remains separated so to allow the reconstruction of the $N$-body interaction functional form in a systematic way. This aspect can be used to introduce an inductive-bias in the models by selecting, for example, only the ones with a repulsive short-range behaviour of the two-body interaction. As a consequence, one can construct physically sound and stable potentials in the extrapolation regime. The JLP is here validated against a benzene dataset and over datasets of both crystalline and amorphous periodic structures, namely copper and carbon, respectively. Finally, we will show how the reconstruction of the interaction at different $N$-body orders can guide us in the tuning of the model, providing more insights than those offered by the standard metrics commonly used to validate a fit.

Our proposed MLP accounts for short-ranged contributions to the total energy, $E_{\text{short}}$, and it is written as a multi-body expansion,

$$E_{\text{short}} = E_0 + E_2 + E_3 + E_4 \ldots ,$$

where the single-body contribution, $E_0$, is an energy offset depending on the number of atomic species present in the system, $E_2$ is the two-body (2B) interaction, depending only on atoms pairs, $E_3$ is the three-body (3B) energy and, in general, $E_n$ describes the $n$-body contribution ($n$B). Since the total energy is a scalar, it must be invariant under translations and rotations of the reference frame. Thus, we can write the 2B term as a sum of pair-wise potentials, each one depending only on the distance between atom pairs, explicitly

$$E_2 = \sum_{i \neq j} v_{2, i, j}^{(2)} (r_{i,j}).$$

We assume that the functional form of the 2B potentials
depends only on the atomic species of the atom involved, so that \( Z_i \) is the atomic number of the \( i \)-th atom located at \( r_i \), and \( r_{ji} = |r_j - r_i| \). By definition, it holds that \( v_{Z_i Z_j} = v_{Z_j Z_i} \). In the following we will use \( v_{ji} \) as a shorthand for \( v_{Z_i Z_j} \). It is useful to remark that the 2B term in Eq. (2) can be re-written, by an appropriately factorization, as a sum of local contributions, \( E_2 = \sum_i \varepsilon_i^{(2)} \), where \( \varepsilon_i^{(2)} \) is the energy associated to the \( i \)-th atom resulting from the pairwise interaction with its local atomic neighbourhood.

The short-ranged nature of the energy implies that the 2B potentials vanish at distances much larger than a characteristic interaction range. Thus, there exists a cut-off radius \( r_{\text{cut}} \), such that \( v_{ji}(r_{ji}) \approx 0 \) for \( r_{ji} > r_{\text{cut}} \). One can then expand the \( v_{ji} 's \) for \( r_{ji} < r_{\text{cut}} \) over Jacobi’s polynomials, \( P_{n}^{\alpha,\beta}(x) \), namely

\[
v^{(2)}_{ji}(r_{ji}) = \sum_{n=0}^{n_{\text{max}}} a_{n}^{ji} P_{n}^{\alpha,\beta}\left(\gamma \cos \left( \frac{r_{ji} - r_{\text{min}}}{r_{\text{cut}} - r_{\text{min}}} \right) \right), \tag{3}
\]

where the sum runs over the positive integers \( n > 0 \), and where we have truncated the expansion to a suitable order, \( n_{\text{max}} \). The shifted Jacobi polynomials employed in this expansion are defined as \( P_{n}^{\alpha,\beta}(x) = P_{n}^{\alpha,\beta}(x) - P_{n}^{\alpha,\beta}(\gamma) \) for \( -\gamma \leq x \leq 1 \) and 0 for \( x < -\gamma \). The constant \( 0 < \gamma \leq 1 \) has been introduced for numerical stability. Again, \( a_{n}^{ji} \) must be intended as a shorthand for \( a_{n}^{ji} Z_i Z_j \) and it holds that \( a_{n}^{ji} = a_{n}^{ji} \). This expansion has being constructed by introducing six hyper-parameters, namely \( \alpha, \beta, \gamma, r_{\text{cut}}, r_{\text{min}} \) and \( n_{\text{max}} \), with \( \alpha, \beta \) being real numbers greater than \(-1\). Each independent set of coefficients, namely sets of coefficients relative to different pairs of atomic species, can be fitted with different hyperparameters. This freedom allows one to perform fine tuning, but it also enlarges the hyper-parameter’s space when the system of interest contains several species. Importantly, the JLPs allow one to address this problem by mean of features selection on the potential used: for instance it is possible to remove potentials that are expected to be unessential or unify potentials, which share the same features, practically reducing the number of hyperparameters and features.

Following the same arguments introduced for the 2B term, and given that the relative position of three points in space can be uniquely determined by two distances and one angle, the 3B interaction can be written as

\[
E_3 = \sum_i \sum_{(j,k)} v^{(3)}_{Z_i Z_k Z_j}(r_{ji}, r_{kj}, \hat{r}_{ji} \cdot \hat{r}_{kj}), \tag{4}
\]

where the second sum runs over all the atoms pairs in the neighbourhood (within \( r_{\text{cut}} \)) of the \( i \)-th atom. Note that also \( E_3 \) can be re-casted in a local form, \( E_3 = \sum_i \varepsilon_i^{(3)} \).

In this case too we will use \( v^{(3)}_{jki} \) as a shorthand. By exploiting the short-ranged nature of the interaction, the \( v^{(3)}_{jki} \) potentials can be expanded as a linear combination of products of two Jacobi polynomials, one for each distance. Given the vectorial nature of the quantities involved, we chose to expand the angular dependence of the potentials over Legendre polynomials. Since these are a particular class of the Jacobi polynomials, such choice is not unique. Explicitly, the proposed expansion for the 3B potentials reads

\[
v^{(3)}_{jki}(r_{ji}, r_{kj}, \hat{r}_{ji} \cdot \hat{r}_{kj}) = \sum_{\substack{n_{1}, n_{2}, n_{3} \leq n_{\text{max}} \\sum_{l=0}^{n_{\text{max}}} a_{n_{1} n_{2} n_{3}}^{jki} \hat{P}_{n_{1}}^{(\alpha)} \hat{P}_{n_{2}}^{(\beta)} \hat{P}_{n_{3}}^{(\gamma)} \hat{r}_{ki}^{l}},
\]

where we have defined \( \hat{P}_{n}^{(\alpha)} = P_{n}^{(\alpha)}(\hat{r}_{ji} \cdot \hat{r}_{kj}) \), being \( P_{n}^{(\alpha)} \) a Legendre polynomial, and where we have introduced the compact notation \( \hat{P}_{n_{1}}^{(\alpha)} \) for a polynomial between the \( j \)-th and \( i \)-th atom [see Eq. (3)]. The expansion of the angular-dependent terms has been truncated at \( n_{\text{max}} \), thus introducing one additional hyper-parameter. Notably, the hyper-parameters chosen for the 3B potentials are generally different from their 2B counterparts. Analogously to the 2B case, \( a_{n_{1} n_{2} n_{3}}^{jki} \) is a shorthand for \( a_{n_{1} n_{2} n_{3}}^{jki} \).

In this case, the expansion coefficients are symmetric for simultaneous swap of the indexes \( j \leftrightarrow k \) and Jacobi indexes, \( n_{1} \leftrightarrow n_{2} \), namely \( a_{n_{1} n_{2} n_{3}}^{jki} = a_{n_{2} n_{1} n_{3}}^{jki} \). Because of this symmetry one must consider only indexes \( n_{1} \geq n_{2} \) when the atoms \( j \) and \( k \) belong to the same species, as shown in the supplementary material (SM). Clearly, the hyper-parameter space can become large when considering multi-species systems, making it necessary to perform system-based features selection. The 4B term can be expanded in a similar way. Considering that three distances and three angles can uniquely describe the relative position of four points in space in a local form, our expansion for the 4B term, \( E_4 \), consists in a linear combination of products of three shifted Jacobi polynomials, one for each distances, and three Legendre polynomials, one for each angle. The explicit expansion for the 4B term can be found in the SM, along with the required symmetry properties for the same-spectre case.

We can finally define a general expansion for the \( nB \) term, \( E_n \). In the local form, our potentials depend on \( n - 1 \) distances and \( (n - 1)(n - 2)/2 \) angles. The expansion is then obtained by associating a Jacobi polynomial to each distance and a Legendre polynomial to each angle. In general, each of the \( nB \) terms can be expanded as a linear combination of products of shifted Jacobi and Legendre polynomials, with the symmetries of the potentials mirrored in that of the indexes of the expansion coefficients and in the factorization of the terms, i.e. permutation of identical atoms, reflection, etc. Such linear expansion allows one to compute the expansion coefficients by simple linear regression. Furthermore, the forces can be calculated by appropriately differentiating the \( nB \) terms, given the simple differentiation rules of the Jacobi polynomials. The explicit expression of the forces derived from the 2B term is reported in the SM. Once the forces are evaluated, also the virial stress can be calculated as shown by Thompson et. al. [14]
FIG. 1. JLP for benzene constructed over the dataset of [15]. Panel a): JPL total energy against the DFT data. The model consist of 30 features (29+intercept) and has an RMS of 1.36 meV/atom on the test set. The energies have been translated by 133 eV/atom. Panel b): reconstructed 2B potentials $v_{CC}^{(2)}$, $v_{CH}^{(2)}$ and $v_{HH}^{(2)}$ [see Eq. (3)], as a function of the interatomic separation and parity distribution plot. In the insert it is shown the full behaviour for small distances, where the potentials are repulsive. Note that the potentials have a minimum close to the first nearest-neighbours distances and that the potentials smoothly vanish at the cut-off radius. Panel c): polar representation of the $v_{CCC}^{(3)}$ potential (the $v_{CCH}^{(3)}$ and $v_{CHC}^{(3)}$ are shown in the SM). The angle represents the angular dependence, while the radius encodes the distance between the central carbon and one carbon on the side. The distance with the second carbon is taken at 1.4 Å. The colour code describes the potential depth. The potential gets repulsive when $CCC < 90^\circ$, while the equilibrium angle, $120^\circ$, is in the attractive region.

FIG. 2. 2B potential reconstruction for a model trained on the copper dataset from Ref. [1] with the distribution of the pair distances from the dataset. The model consists in 14 features (13+intercept, of which 5 are relative to the 2B term and the remaining 8 relative to the 3B term), and with an RMS of 0.9 meV/atom. The insert show the behaviour at small distances, where the potential is strongly repulsive.

is that it allows us to reconstruct the form of the $n$-body potentials, $v^{(n)}$. For example, once the 2B coefficients $a_{ij}^{(2)}$ are learned, we can reconstruct the potentials $v^{(2)}_{ij}$ (x) by mean of Eq. (3). This gives us a fully interpretable potential, namely we can immediately establish the appropriateness of the hyper-parameters based on the behaviour of the potential obtained. We have observed that, while using a large maximum polynomial order $n_{\text{max}}$ often increases the accuracy (it reduces the root-mean-square error - RMS), the resulting potentials develop unphysical features, such as unjustified rapid oscillations or attractive behaviour at small distances. On the contrary, for smaller $n_{\text{max}}$, we obtain a repulsive behaviour in the same region, despite the training set never includes structures presenting close atoms. From this analysis, we conclude that the RMS is not a sufficient indicator of the performance of a potential and then, crucially, the hyperparameters will be tuned by simultaneously reducing the RMS and keeping physically-reasonable potentials shapes, for instance by ensuring the correct repulsive behaviour near $r_{\text{min}}$.

In this letter we report results obtained for benzene, copper and amorphous carbon. These systems have been chosen to test the JLPs over a broad range of chemical-bond distributions and structures, namely a molecule, a crystalline solid and an amorphous compound. The coefficients are always learned by linear regression.

Reference [15] provides the data for benzene, for which we have constructed a JLP based on 30 features (see SM for details on the dataset and the hyperparameters). The 2B energy is expanded over three potentials, $v_{CC}^{(2)}$, $v_{HH}^{(2)}$, $v_{CH}^{(2)}$, while the 3B one includes $v_{CCH}^{(3)}$, $v_{CCH}^{(3)}$ and $v_{CHC}^{(3)}$. Importantly, we found that including also the three remaining 3B potentials $v_{HHC}^{(3)}$, $v_{HCH}^{(3)}$ and $v_{HHC}^{(3)}$ does not improve the fit, allowing for 3B features selection. This gives us an RMSE of 1.76 meV/atom on the test set, as shown in Fig. 1. The reconstructed 2B potential (middle panel of Fig. 1) vanishes at $r_{\text{cut}}$, presents a repulsive behaviour for short atomic distances and a pronounced minimum around the nearest-neighbour bond length. Finally, the right-hand side panel of Fig. 1 shows the energy profile of the angular component of the 3B potential for CCC, where we can appreciate an attractive contribution for the angle formed among nearest-neighbour carbons ($120^\circ$) and repulsive for angles smaller than $90^\circ$. The other potentials are reported in the SM.

Next, fcc copper (data from [1]) is used to test the
FIG. 3. Summary of the results for the amorphous carbon dataset, [16]. The bottom panel shows the difference between the DFT and the predicted energy values. The distributions of the predicted and DFT points are in the right-hand side and top panel, respectively. The "2b" fit is obtained by using only the features belonging to the 2B part of the full model. Analogously, the "2b+3b" fit has been obtained by using the 2B and 3B features. The remaining fit represents the full 4B model. The total number of features used is 103 (with the intercept) and the RMS of the full model is 54.3 meV/atom.

Potential on a crystalline structure. Also in this case the model includes descriptors up to the 3B order. In Fig. 2 we report the 2B potential reconstruction, which presents a minima in correspondence of the position of the first neighbors and a repulsive short-range behaviour. The model has a RMS of 0.9 meV/atom over the test set, in line with the performance of the models tested in [1] (see SM for further details).

Our last example consists of amorphous carbon, with dataset taken from Ref. [16]. The results are reported in Fig. 3, where we present a JLP up to the 4B term, displaying an accuracy close to that of a GAP model, despite being constructed over only 103 features (102 + intercept). In the figure, we also show how the model performs when truncated to lower-body orders, namely 2B and 3B. We see that the 2B features of the potential, while reaching already a decent overall accuracy, are particularly inaccurate for low energies, where in the dataset we find the bulk crystalline structures. We then conclude that a body-order-wise analysis of the performance of the model and a simultaneous visual examination of the reconstructed potential can be used as a powerful tool to investigate a dataset, exposing different features such as different material phases. In the SM we report the 2B and 3B potentials reconstructions.

In conclusion, we are proposing a new linear, many-body, machine-learned potential, inspired by cluster expansion methods and constructed over the Jacobi and Legendre polynomials. Since terms corresponding to different interaction orders are fully separated, the potential is fully interpretable and displays physically meaningful features, such as a repulsive behaviour at small distances or minima mirroring the closest neighbours structure. We have provided examples for an organic molecule, a crystalline solid and an amorphous compounds, showing that the ability to interpret the various contributions is equally important to the energy accuracy to obtain stable and physically sound potentials.

Acknowledgments

This work has been supported by the Irish Research Council Advanced Laureate Award (IRCLA/2019/127), and by the Irish Research Council postgraduate program (MC). 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 Number - EuroCC-AF-3).

[1] 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).
[2] 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).
[3] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Phys. Rev. B 99, 014104 (2019).
[4] A.V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, Multiscale Modeling & Simulation 14, 1153 (2016).
[5] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. 98, 146401 (2007).
[6] 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).
[7] 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).
[8] A. Lunghi and S. Sanvito, A unified picture of the co-
valent bond within quantum-accurate force fields: from simple organic molecules to metallic complexes’ reactivity, Science Advances 5, eaaw2210 (2019).

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

[10] F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti and L. Emsley, Chem. Rev. 121, 9759 (2021).

[11] 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).

[12] 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).

[13] 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).

[14] V. L. Deringer and G. Csányi, Machine learning based interatomic potential for amorphous carbon, Phys. Rev. B 95, 094203 (2017).
The Jacobi-Legendre potential

Supplementary Material

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

I. USEFUL DEFINITIONS, EXPRESSIONS AND PROPERTIES

A. Definitions

The distances in the Jacobi-Legendre (JL) potentials are encoded in the shifted Jacobi polynomials, defined as

$$\tilde{P}^{(\alpha,\beta)}(x) = \begin{cases} P^{(\alpha,\beta)}(x) - P^{(\alpha,\beta)}(-\gamma) & \text{for } -\gamma \leq x \leq 1, \\ 0 & \text{for } x < -\gamma, \end{cases}$$

where $P^{(\alpha,\beta)}$ is a Jacobi polynomials and $\gamma$ is a constant $\in [0,1]$. A shorthand for the argument of the polynomials is

$$\tilde{P}^{(\alpha,\beta)}_{nji} = \tilde{P}^{(\alpha,\beta)}_{n}(\gamma \cos(\pi r_{ji} - r_{\min} - r_{\text{cut}})).$$

For the angular contribution we use the Legendre polynomial. In particular we use the shorthand notation

$$P_{l}^{jki} = P_{l}(\hat{r}_{ji} \cdot \hat{r}_{ki}),$$

where $P_{l}$ is a Legendre polynomial of order $l$, and $\hat{r}_{ji} = (r_{j} - r_{i})/r_{ji}$

B. Explicit formulas

1. Two-body term

In general, the 2B term can be written as

$$E_{2} = \sum_{n>0} \sum_{ij} a_{nji}^{ij} \tilde{P}^{(\alpha,\beta)}_{nji}.$$  \hspace{1cm} (4)

The expansion coefficients, however, depend on the atomic species of the atom considered. This, together with the symmetry property $a_{nji}^{ij} = a_{nji}^{ji}$, allows for simplifications based on the number of species. Here we report the explicit expressions of the potentials in terms of the number of species.

a. Mono-species system

For a system in which all the atoms belong to the same species $Z$, the expansion coefficients do not depend on the pair of atom considered. Thus, the expression becomes

$$E_{2} = \sum_{n>0} a_{nZ}^{Z} \sum_{ij} \tilde{P}^{(\alpha,\beta)}_{nji},$$

where

$$E_{2}^{Z_{p}Z_{q}} = \sum_{n>0} a_{nZ_{p}}^{Z_{q}} \sum_{i \in Z_{p}} \sum_{j \in Z_{q}} \tilde{P}^{(\alpha,\beta)}_{nji}, \hspace{1cm} \text{with } p, q = 1, 2.$$  \hspace{1cm} (5)

b. Two-species system

Having two species, $Z_1$ and $Z_2$, the 2B term can be written as

$$E_{2} = E_{2}^{Z_{1}Z_{1}} + E_{2}^{Z_{2}Z_{2}} + E_{2}^{Z_{1}Z_{2}} + E_{2}^{Z_{2}Z_{1}},$$

where

$$E_{2}^{Z_{p}Z_{q}} = \sum_{n>0} a_{nZ_{p}}^{Z_{q}} \sum_{i \in Z_{p}} \sum_{j \in Z_{q}} \tilde{P}^{(\alpha,\beta)}_{nji}, \hspace{1cm} \text{with } p, q = 1, 2.$$  \hspace{1cm} (6)

Note that, since $a_{nZ_{p}}^{Z_{q}} = a_{nZ_{q}}^{Z_{p}}$, then $E_{2}^{Z_{p}Z_{q}} = E_{2}^{Z_{q}Z_{p}}$. Also, in general, we may have different hyper-parameters for each non-equivalent term in Eq. (6).
c. Multi-species system  If the system if made of atoms belonging to the species \( Z_1, \cdots, Z_N \), then the 2B term can be written as

\[
E_2 = \sum_p^N E_2^p Z_p + \sum_{p \neq q}^N E_2^{p q} Z_p Z_q,
\]

(8)

where all the non-equivalent terms in the sum can have different hyper-parameters.

d. 2B forces  The expression for the 2B forces can be obtained by simple differentiation of the 2B term as

\[
F_j^{(2)} = -\frac{\partial E_2}{\partial r_j} = \sum_{n> i, \alpha \neq j}^{n_{\text{max}}} \sum_{\alpha} \sum_{\beta} a_{k}^{(\alpha,\beta)} Q_{nji}^{(\alpha,\beta)},
\]

(9)

with

\[
Q_{nji}^{(\alpha,\beta)} = \frac{(\alpha + \beta + n + 1) \gamma \pi}{r_{\text{cut}} - r_{\text{min}}} \sin \left( \pi \frac{r_{ji} - r_{\text{min}}}{r_{\text{cut}} - r_{\text{min}}} \right) P_{n-1}^{(\alpha+1,\beta+1)} \left( \gamma \cos \left( \pi \frac{r_{ji} - r_{\text{min}}}{r_{\text{cut}} - r_{\text{min}}} \right) \right) \hat{r}_{ji}
\]

(10)

where \( r_{ji} \leq r_{\text{cut}} \).

2. Three-body term

The three body term reads

\[
E_3 = \sum_i \sum_{(j,k)} v_{jki}^{(3)}(r_{ji}, r_{kj}, \hat{r}_{jki}),
\]

(11)

where we defined \( \hat{r}_{jki} = \hat{r}_{ji} \cdot \hat{r}_{ki} \) for simplicity. The second sum runs over all the atoms in the neighborhood of the atom \( i \)-th, such that each pair is taken once. As for the 2B potentials, the functional form is supposed to be dependent only on the species of the atoms \( j, k, i \). Thus \( v_{jki}^{(3)}(r_{ji}, r_{kj}, \hat{r}_{jki}) \) is to be intended as a shortcut for \( v_{Z_j Z_k Z_i}^{(3)}(r_{ji}, r_{kj}, \hat{r}_{jki}) \) which simply states that the ordering of the arguments in the function is arbitrary, but must be coherently followed once the choice is made: this removes any problems in the ordering when considering the possible pairs \((j, k)\) of the neighborhood. The expansion of the potentials then reads

\[
v_{jki}^{(3)}(r_{ji}, r_{kj}, \hat{r}_{jki}) = \sum_{n_{1}^{+}}^{n_{\text{max}}} \sum_{n_{2}^{+}}^{l_{\text{max}}} \sum_{l}^{l_{\text{max}}} \sum_{\alpha}^{a_{jki}^{(\alpha,\beta)}} \sum_{\beta}^{a_{kji}^{(\alpha,\beta)}} \tilde{P}_{n_{1}ji}^{(\alpha,\beta)} \tilde{P}_{n_{2}ki}^{(\alpha,\beta)} \tilde{P}_{l_{ji}}^{(\alpha,\beta)}
\]

(12)

where the expansion coefficients satisfy the symmetry property \( a_{jki}^{(\alpha,\beta)} = a_{kji}^{(\alpha,\beta)} \). Again, the expression \( a_{jki}^{\alpha,\beta} \) must be intended as a shortcut for \( a_{n_{1} n_{2} l}^{\alpha,\beta} \). Thus, if the \( j \)-th and \( k \)-th atoms belong to the same species, then we have that the coefficients are symmetric under the swap \( n_{1} \leftrightarrow n_{2} \) and therefore we use the expression

\[
v_{Z_j Z_k Z_i}^{(3)}(r_{ji}, r_{kj}, \hat{r}_{jki}) = \sum_{n_{1}^{+}}^{n_{\text{max}}} \sum_{n_{2}^{+}}^{l_{\text{max}}} \sum_{l}^{l_{\text{max}}} a_{n_{1} n_{2} l}^{\alpha,\beta} \tilde{P}_{n_{1}ji}^{(\alpha,\beta)} \tilde{P}_{n_{2}ki}^{(\alpha,\beta)} \tilde{P}_{l_{ji}}^{(\alpha,\beta)}
\]

(13)

Since the expression in square brackets are now symmetric under the swap \( n_{1} \leftrightarrow n_{2} \) we can take just indexes such that \( n_{1} \geq n_{2} \). This enforces the potential’s permutational invariance for atoms of the same species.

3. Four-body term

The 4B term, \( E_4 \), in the JL formalism explicitly reads

\[
E_4 = \sum_{i} \sum_{(j,k,m)} v_{jkm}^{(4)}(r_{ji}, r_{kj}, r_{mi}, \hat{r}_{jki}, \hat{r}_{jmi}, \hat{r}_{kmi}),
\]

(14)
where the second sum runs over all triplets of atoms in the neighborhood of the $i$-th atom and with the potential given by

$$v^{(4)}_{jkm}(r_{ji}, r_{ki}, r_{mi}, \hat{r}_{jmi}, \hat{r}_{kmi}) = \sum_{n_{1}^{j}}^{n_{max}} \sum_{n_{2}^{k}}^{n_{max}} a^{kmi}_{n_{1}n_{2}n_{3}l_{1}l_{2}l_{3}} \hat{p}(\alpha, \beta)^{\prime} \hat{p}(\alpha, \beta)^{\prime} p_{n_{1}ki}^{l_{1}} p_{l_{2}jmi}^{l_{2}} p_{kmi}^{l_{3}}. \tag{15}$$

Also here it is assumed that the functional form of the potential depends only on the atomic species involved. Thus $v^{(4)}_{jkm}$ and $a^{kmi}_{n_{1}n_{2}n_{3}l_{1}l_{2}l_{3}}$ are intended as a shorthand for $v^{(4)}_{j_{z}j_{z}Z_{m}Z_{i}}$ and $a^{Z_{i}Z_{k}Z_{m}Z_{i}}_{n_{1}n_{2}n_{3}l_{1}l_{2}l_{3}}$ respectively.

In order to avoid any ambiguity in the triplets ordering, the potentials are defined to satisfy the following chain of symmetries

$$v^{(4)}_{jkm}(r_{ji}, r_{ki}, r_{mi}, \hat{r}_{jmi}, \hat{r}_{kmi}) = v^{(4)}_{mjk}(r_{mi}, r_{ji}, r_{ki}, \hat{r}_{kmi}, \hat{r}_{jmi}, \hat{r}_{jmi}) = v^{(4)}_{kmj}(r_{ki}, r_{mi}, r_{ji}, \hat{r}_{jmi}, \hat{r}_{kmi}, \hat{r}_{kmi}) = v^{(4)}_{jmk}(r_{ji}, r_{mi}, r_{ki}, \hat{r}_{jmi}, \hat{r}_{kmi}, \hat{r}_{kmi}) = v^{(4)}_{mjk}(r_{mi}, r_{ji}, r_{ki}, \hat{r}_{jmi}, \hat{r}_{kmi}, \hat{r}_{jmi}), \tag{16}$$

which correspond to simultaneous-cyclic and non-cyclic permutations of the atomic ordering and functional dependence of the potentials. This is mirrored in the corresponding symmetries for the coefficients, analogously to the three-body case. Finally, the case in which two or more atoms belong to the same atomic species must be treated carefully, so that the right symmetry of the coefficients - which represents the invariance under permutations of identical atoms - is enforced.

4. Linear regression

The target quantity is the total energy of the system, which is then approximated by the JL energy written as

$$E_{tot} \simeq E_{JL} = E_{0} + E_{2} + E_{3} + E_{4} + \ldots. \tag{17}$$

Here the intercept is given by

$$E_{0} = \sum_{k}^{\text{species}} N_{k} \varepsilon_{k}^{(0)}, \tag{18}$$

where the sum runs over all the atomic species in the system, $\varepsilon_{k}^{(0)}$ are constant to be fitted and where $N_{k}$ is the number of atoms of the $k$-th species.

II. BENZENE MOLECULE

The data set for the benzene molecule has been taken from reference [1], where the geometries are sampled from long molecular dynamics trajectories at temperatures ranging from 300K to 400K. We use the provided test/train data, where the train set contains 6000 geometries from 300K, 300K+350K and 300K+400K trajectories and the test set contains 200 geometries at 300K.

The hyperparameters for the fitted model are reported in Table I, while the potentials $v^{(3)}_{HC}$ and $v^{(3)}_{CHC}$ are shown in Fig. 1.

III. CRYSTALLINE FCC COPPER

The data set for fcc copper has been taken from reference [2], where the configurations include the equilibrium geometry, various surfaces, strained geometries and geometries sampled from molecular dynamics trajectories, including liquids and vacancies. We use the provided test/train data, where the training-data set consists of 262 geometries and the test-data set consists of 31 geometries. The dataset was originally generated to benchmark several machine learning inter-atomic potentials. We report the hyper-parameters used for the fit in Table II, where it is possible to appreciate the model compactness. In Fig. 2 we show the parity plot for both the training and the test sets, showing that 14 features are enough to obtain a very good accuracy.
Two Body Three Body

|       | CC | HH | HC |       | CC | CCH | CCH |
|-------|----|----|----|-------|----|-----|-----|
| \( n_{\text{max}} \) | 5  | 6  | 4  | \( n_{\text{max}} \) | 2  | 1   | 1   |
| \( l_{\text{max}} \)   | –  | –  | –  | \( l_{\text{max}} \)   | 2  | 2   | 1   |
| \( r_{\text{cut}} \) (Å) | 2.4 | 2.5 | 3.3 | \( r_{\text{cut}} \) (Å) | 2.8 | 3.0 | 2.8 |
| \( \alpha = \beta \)     | 3.0 | 3.0 | 1.0 | \( \alpha = \beta \)     | 1.0 | 1.0 | 1.0 |
| \# of features | 5  | 6  | 4  | \# of features | 9  | 3   | 2   |

**TABLE I.** Hyperparameters for benzene. In order to reduce the complexity of the hyperparameter’s space we decided to set \( \alpha = \beta, \gamma = 1 \) and \( r_{\text{min}} = 0 \). Note that the number of features is small, at most 9, for each of the fitted potentials.

**FIG. 1.** On the left the potential \( v_{hCC}^{(3)}(1, r, \theta) \) and on the right the potential \( v_{cHCC}^{(3)}(1, r, \theta) \) where the HC bond is fixed at 1 Å.

|       | CuCu | CuCuCu |
|-------|------|--------|
| \( n_{\text{max}} \) | 7    | 2      |
| \( l_{\text{max}} \) | –    | 1      |
| \( r_{\text{cut}} \) (Å) | 4.8  | 4.8    |
| \( \alpha = \beta \) | 5.0  | 5.0    |
| \# of features | 7    | 6      |

**TABLE II.** The hyperparameters used to fit the copper dataset are reported. We set \( \alpha = \beta, \gamma = 1 \) and \( r_{\text{min}} = 0 \). The CuCu column shows the hyperparameters chosen for the 2B term, while the CuCuCu shows the ones for the 3B term. The total number of features is 14, including the intercept.

**IV. AMORPHOUS CARBON**

Here we used the data set from reference [3], where the geometries consists of crystalline (diamond) carbon, graphite, geometries sampled from high temperature molecular dynamics trajectories representing liquid or amorphous phases and surfaces. We use the bulk crystalline and amorphous structures from the provided test/train data. The final training set contains 2927 geometries and the test set contains 489 geometries. In Table III we report the chosen hyperparameters, while Fig. 3 shows the two body and three body potentials reconstruction.
FIG. 2. a) Parity plot for the JLP fitted on the copper dataset. The plot shows high accuracy despite the fact that the small number of features. Panel b) Reconstruction of the three body potential where one of the bond is fixed at 2.5 Å. It should be noted that features at small distances are washed out by the strong repulsive behaviour of the 2B potential shown in the main text.

|                | Two Body | Three Body | Four Body |
|----------------|----------|------------|-----------|
|                | CC       | CCCCC      | CCCC      |
| \( n_{\text{max}} \) | 7        | 5          | 1         |
| \( l_{\text{max}} \) | –        | 4          | 3         |
| \( r_{\text{cut}} \) (Å) | 2.8      | 3.0        | 3.1       |
| \( \alpha = \beta \) | 4        | 1          | 1         |
| # of features  | 7        | 75         | 20        |

TABLE III. Details of the JLP trained on the carbon dataset. To deduce the number of hyper parameters, we fix \( \alpha \) and \( \beta \) to be equal, \( \gamma \) to 1 and \( r_{\text{min}} \) to 0. The model is relatively compact with only 103 (102+intercept) features.
FIG. 3. Panel a): Two-body potential reconstruction. The insert shows the small-distances-repulsive behaviour of the potential. This behaviour is spontaneously arising in an extrapolation region (the dataset does not contain any bond smaller than 1 Å), with no underlying potentials or super-imposed constraints. A minima can be appreciated around the position of the closest neighbours. Panel b): Reconstruction of the three-body potential. This is a slice obtained by fixing one of the bond length at the equilibrium distance of 1.5 Å. While the potential appears large in the small-distances region, it should be noted that the 2B potentials is greatly larger in the same region. Most importantly, the potentials shows minima around 120° for the first neighbours length, and around 60° and in the range from ~145° to 180° for the second neighbours length (2.5 Å), coherently with crystals geometries.
[1] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K.-R. Müller, *Bypassing the Kohn-Sham equations with machine learning*, Nat Commun **8**, 872 (2017).

[2] 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).

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