Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Trinity College Dublin

Ph.D. Thesis
School of Physics

The Jacobi-Legendre framework for Machine Learning in Materials Investigation and Discovery

Candidate:
Michelangelo **Domina**

Supervisor:
Prof. Stefano **Sanvito**

Year 2024

Declaration for Plagiarism and Deposit

I, Michelangelo Domina, hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signature

Abstract

Machine-learning models have rapidly become fundamental tools in the study of materials properties. In the past few years there has been a surge of interest in the construction of new models and descriptors to accelerate the investigation of known materials and the discovery of new ones, mostly in the framework provided by electronic structure calculations.

This thesis revolves around the study of descriptors for machine-learning models, capable of describing the configuration of the system, while encoding the symmetries that are required to efficiently target properties of interest. Several cases will be treated, ranging from systems with magnetic degrees of freedom, to models that could predict potential energy surfaces, from models dedicated to the prediction of the electronic density, to ones for tensor and tensor fields. All of this will be done in the spirit of accelerating *ab-initio* calculations within a unified framework, denoted with the name "Jacobi-Legendre", here defined and meticulously investigated.

In the thesis, we will explore a model for magnetic systems, in which the spin degrees of freedom will be placed on the same footing as the description of the position of the atoms. We will then dedicate the core of this work to the definition and construction of the Jacobi-Legendre framework, starting from a model devoted to the prediction of the potential energy surface of a system. With the formalism in place, we will generalize the descriptors to the prediction of the electron density, proving how the reached accuracy enables to accelerate electronic-structure calculations. We will then define the formalism in full by presenting and exploring methods for the prediction of tensors and tensorial fields. The thesis will be concluded with a thorough study on how the use of multipolar-spherical harmonics can be beneficial in simplifying the definition of descriptors and in exposing limits of current approaches, while proposing new strategies based on the Jacobi-Legendre potentials.

Acknowledgements

First of all, I would like to thank Stefano for his availability, expertise, insight and kindness. Many of the results achieved in this thesis would not have been possible without his intuition and expertise. In the face of extraordinary situations, he managed to maintain a good atmosphere in this beautiful group and always provided all the necessary support.

I also thank Simone both professionally and personally, she has been there to help me in the most challenging moments.

Many thanks to Alessandro, who has always been available for clarifications and discussions, finding time even when this resource became scarce and precious.

I thank from the bottom of my heart all the colleagues in the group for creating and sustaining such a beautiful atmosphere. A special thanks to Urvesh, to whom I am grateful for both friendship and support. This thesis would have been much different without you.

Thanks to Matteo, Laura, Bruno, Hugo, Luke, and Mike; it has been a pleasure and an honor to go through this journey with you, and I am grateful for the mutual support we have given each other over the years and for your friendship. You have made every day enjoyable, even during the toughest times, and I can say that it is mainly thanks to you that this has been such a positive experience.

Thanks to Mario, who has always supported me, explaining the tricks of the trade and showing me the various stages of this journey. I am grateful to all the members of Office 3.18, Michael, Paddy, Rutchapon, Abisheik, Suman, Rajarshi, Eoghan, Willy: it is also thanks to you that the working atmosphere has always been so pleasant. A special thanks to Annie, Declan, and Eoin; it is always fun to spend time with you, and I am grateful for your friendship.

A special thank goes to Stefania N., who, with her availability, patience, and compassion, has allowed Dublin to become a place where I can feel at home.

A heartfelt thank goes to Prof. Messina, who gave me the tools and the rigour to face challenges in an inquisitive and formally critical way, and who believed in me and in my abilities during my time at the University of Palermo. I wish to thank also Prof. S. Tanaka who, with his care and support during my first experience abroad, prepared me to deal with the unavoidable hardship of studying and living somewhere far from home.

I wish to thank Prof. Guccione, for helping me in more than one critical phases before and after the adventure of the PhD studies, alongside her undisputed scientific skills she is able to support students which such care that I never found in anyone else, an for that I am extremely grateful. Finally, I wish to thank Prof. Mantegna, for helping me and supporting me during the difficult time of the choice of the PhD.

In general, I thank all the friends of the group and otherwise, Umit, Valerio, Anita, Akash, James N., Emanuele, Anna, Meric, Reena, Nina, Antik, Zahra, Cecilia, Sindre, Maria, Rui, Andrea, Anais, Jorge, Stella, and everyone else that has shared with me this time; thank you from the bottom of my heart for all these years together.

Among the friends who have supported me during these years, I thank Mikey very much: we faced an impossible situation together, and thanks to you and the kindness of your family, Mike, Maureen and Eric, everything turned out for the best. I will always be grateful to you all. A special thanks to Dan: you are an amazing person, and we have shared so much, ups and downs, that I wouldn't even know where to start. I want you to know that I will always be grateful for our friendship. A heartfelt thank you to James; your friendship is precious to me, as is the time we spend together. I hope the new adventure on the horizon will give you everything you deserve and more. Thanks to Annael; it was thanks to you and Mikey that Dublin seemed beautiful even though it was so far from home. Thanks to Jiv; you are an amazing person, and I greatly admire you; I am sure you will achieve your dreams. Thanks to Livia; you helped me during some of the toughest moments. Thanks to Camilla and Chiara; I will not forget how altruistically you dedicated yourselves to helping me in times of need. Thanks to the "Italian group" with whom I shared wonderful moments.

I will always thank my family, my aunts and uncle which supported me moving to Ireland. I wish to thank my mom and dad, who provided me with the tools, both material and mental, to reach this milestone. I carry you in my heart at all times, and a part of this thesis is dedicated to you.

Thanks to Roberto, I am proud to consider you part of my family; your friendship is one of the most important things in my life.

Finally, thanks to Stefania, who makes my life beautiful every day and knows how to make me happy with simple things as well as breathtaking ones. I know that when I am with you, I am at home; this work is largely dedicated to you.

Publications

- M. Domina, M. Cobelli, and S. Sanvito, *Spectral neighbor representation for vector fields: Machine learning potentials including spin*, Phys. Rev. B **105**, 214439 (2022).
- 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. Comput. Mater. **9**, 87 (2023).
- M. Domina, U. Patil, M. Cobelli, and S. Sanvito, *Cluster expansion constructed over Jacobi-Legendre polynomials for accurate force fields*, Phys. Rev. B **108**, 094102 (2023).

Conferences

Psi-k Conference, Lausanne, Switzerland (2022)

– M. Domina, U. Patil, M. Cobelli, and S. Sanvito. "Jacobi-Legendre potentials" (Poster).

– M. Domina, M. Cobelli, and S. Sanvito. "A spectral-neighbour representation for vector fields: machine-learning potentials including spin" (Poster).

Contents

Chapter 1

Introduction

In recent years, there has been a paradigm shift in the computational study of materials, with the crowning of Machine-Learning (ML) techniques as newly found essential tools in the study of condensed matter physics, molecular dynamics and thermodynamic properties. As for any other aspect of today daily life, ML is having a severe impact on science, enabling to go beyond the limits of traditional coding and so to greatly reduce the computational cost inherent the study of materials. The traditional approach involves quantum-level accurate *ab-initio* theories, that accurately predict the electronic structure, starting from the underlying quantum mechanical representation of atoms and electrons: an extremely successful example is given by density functional theory (DFT) [1, 2], which focuses on the study of the electronic density to infer several properties of interests, ranging from energies and forces, to localized magnetic moments or phonon-dispersion curves [3, 4]. While it is undeniable that these theories have sparked nothing short of a revolution in the computational study of electronic structures, accurate calculations when performed are usually characterized by an heavy computational overhead even on the most advanced high-performance computing clusters. This becomes particularly significant for the full characterization of known materials or for the discovery of new ones, where both the combinatorially complex problem of the available stoichiometries, and the characterization of the geometry of the system, must be tackled. Indeed, the ambitious goal behind the introduction of ML techniques in the field is to reach the same state-of-the-art-accuracy of computationally expensive *ab-initio* methods, but with a fraction of the cost. Starting from the seminal work of Blank and co-workers of Ref. [5], where a Neural Network (NN) was used to describe a potential energy surface (PES), the field have seen an exponential expansion, both in the development of progressively more sophisticated methods and tools, and in the application to the prediction of several properties of interest. Examples of applications are the search of stable ternary alloys [6], the prediction of the critical temperature [7, 8], the study of the atomization energies of molecules [9], covalent bonds

[10], the PES of catalysts [11], structure predictions [12], large-scale diffusions [13], study of the bulk and the surface phase (for example for copper [14]), Heusler compounds [15, 16], binary alloys [17], or phase changes (e.g., of GeTe [18]), just to mention a few.
ML also provides extremely useful tools for high-throughput investigations that, with the extensive use of shared datasets, such as the AFLOW [19] or the Materials project [20] repositories, significantly boost the search for new materials (see, for example, the short review of Ref. [21]): one example is given by the prediction of the thermodynamical stability (distance from the convex hull) for compounds (in a cubic perovskite geometry) constructed from 64 elements of the periodic table [22].

ML in a nutshell Let us now introduce the the most important ingredients for the definition of a ML model applied to the study of materials, before proceeding in presenting the goal of this thesis and its organization. Broadly speaking, we can say that a ML model is composed of three main parts: the first one, the descriptors (or fingerprints), are the way in which we characterize the system at hand, i.e., the way we encode and represent the degrees of freedom. At the other side of the spectrum is the quantity we want to target (or output): this can be the total energy of the system, or the forces, as well as other properties such as the polarizability or the atomization energies. At the core of a ML model is how the descriptors and the target are interpolated: this goes from the simple case of a linear regression on the descriptors, to similarity kernels, to even more complicated architectures such the ones provided by NNs. Efforts have been made in all these cases, with the definition of several linear-model based approaches [23–25] or kernel ones [26–29], up to deep NNs capable of tackling large and diverse datasets [30–33]. In particular, ML model must be trained, i.e., optimized with respect to the parameters of the model/architecture, on a dataset of choice, usually built by means of *ab-initio* methods. Indeed, all these architectures have some advantages and disadvantages: on the one hand, NNs are usually considered universal approximators [34, 35] and, while sacrificing the simplicity and immediate interpretability of linear models, given the large amount of parameters (weights) to optimize, they allow for a large flexibility in both the degrees of freedom of the Networks, and in the virtually unlimited choices for the construction of the architectures. This gives a significant advantage when, in particular, the available dataset are vast and diverse, containing a large number of systems, defined in different configurations. On the other hand, however, it is not uncommon to study very specific instances of a system where the *ab-initio* methods are expensive and data are not available. In such cases, not only does the dataset contain a (much) smaller number of datapoints, but it can also be difficult to populate: this is the scenario when a linear model, with highly tuned descriptors could be beneficial. Indeed, if constructing linear models necessitates of a careful design of the descriptors (in the case of NNs, this point is more shifted in the

design of the architecture), as implied in the very definition of linearity, i.e., a descriptor without the relevant features would be catastrophic for the performance of the model, linear models are well-suited for smaller dataset, for which the degrees of freedom are not easily saturated. Finally, a linear model can also benefit from specific active-learning schemes [36, 37], where the training of the model is performed in an optimal way with constant evaluations and selections regarding which new configurations should be progressively included in the training process. Kernel methods, in contrast, are somehow the middle ground between linear models and NNs: if, one one side, they retain an aspect of linearity, given that usually the target is defined as a linear interpolation of kernels, they can also incorporate high-degrees of non linearity, by means of the definition of the kernels themselves. The trade-off is usually on the cost of evaluating the kernels, since they usually require to compute similarity measures between the new configuration and all the ones in the dataset (this problem has been tackled, for example, by means of the introduction of recursive scheme for "higher-body" order fingerprints in Ref. [38]).

Training and loss functions While we will not dive into details about the training process or the best way to assess the performance of a ML model, it is important to define the most used strategies and tools. In particular, to train a ML model usually involves the minimization of a *loss function*, i.e., a function that describes the (dis)similarity between the predictions of the model and the actual target. Since the loss function contains the quantities predicted by the model itself, minimizing it is equivalent to a search of the optimal set of free parameters (usually called weights in the context of NNs, or expansion coefficients for linear models), that allow the descriptors to correctly reproduce the targeted quantities. Clearly, there is a large selection of loss functions and optimizers in literature, and an overview of the methods is beyond the scope of this thesis: to fix the ideas we will mostly use the squared L^2 norm.

Performance assessment Another important aspect for ML models is the assessment of the performance/accuracy. This is usually done by some metric: the most used ones are the mean absolute error (MAE) and the root mean squared error (RMSE). They are usually provided together, since the MAE gives more information about the average trend of the predictions, while the RMSE is more sensitive to outliers. Another important tool for visual inspection is the so-called parity plot, which is a scatter plot of all the predictions against the actual data. Since a good accuracy implies a good approximation of the target, visually this means that the more data are aligned with the 45° line, the better the model is performing. While we will make an extensive use of parity plots and

MAEs and RMSEs, it should be noted that, in the field of physical applications, these estimators do not provide a complete picture: indeed, most of the interest lies in applying the model to configurations that are outside the scope of the *ab-initio*-constructed dataset, namely in performing molecular dynamics or in allowing for a relaxation of the system. In this sense, only evaluating on specific tasks or properties of interest can give true insight on the model performances.

Descriptors and the role of symmetries A fundamental property in the construction and design of descriptors of ML models lies in the correct encoding of the symmetries of the targets. Indeed, when describing some property of a system, the descriptors should possess the same, and only, symmetries of the properties themselves. This is well clarified by the example of translations/rotations and energies. When dealing with a system in absence of any external fields, we can freely perform translation and rotations of the frame of reference, without varying the energy of the system. Thus, if descriptors that are not translationally or rotationally invariant are used, then the model has to "learn" to map all the different orientations of the system into the same energy. This produces suboptimal results. In fact, not only cannot the invariance be enforced exactly by a training procedure, but also, it requires inefficient usage of computational resources and an artificial enhancement of the dataset. Indeed, we also have to include translated or rotated configurations. Therefore, the encoding of the correct symmetries on the descriptors is of paramount importance in the design of ML models, and will be a central topic of this thesis. In particular, we will consider scalars, tensors and tensorial fields. For example, among the scalars, which are isometrically (under translation, rotation and inversion) invariant, energies have an ubiquitous role, and as such will be aimed by specific ML models, called ML potentials (MLPs)[39]. These models aim to predict the PES of the system, while also providing forces and components of the stress tensor. Another kind of target are the tensors, which also includes vectors (forces, dipoles), and which are "covariant" under a rotation, i.e., the components mix in a specific way. Finally we have scalar and tensorial *fields*, such as the electronic density or the magnetization vector. All these cases will be covered in depth in the thesis, and so we will postpone their analysis to subsequent chapters.

In this thesis work, we will specifically consider descriptors for linear models (or, at most, for kernel-based models), and we will not explicitly investigate NNs. However, already in the next chapter, we will explore the most important ideas that were born with the advent of NNs in materials science, such as the introduction of local representation, or the implementation of multi-body expansions. We refer to the two reviews of Refs. [40, 41], for a detailed presentation and evolution of NNs, which presents a time-ordered list of the key ideas and tools used in the field, and the challenges faced by new generations of ar-

chitectures and descriptors. Among such challenges, shared by virtually all the ML models available, we mention the inclusion of long-range interactions, which goes beyond the capabilities of local representations. This has been pursued, for example, in Ref. [42], which presented the long distance equivariant formalism (LODE) that introduced a power-law adapted atomic density, or in Ref. [43] which explicitly included electrostatic contributions in the representation. Another challenge is to describe magnetic environments, i.e., systems in which we have to associate vectorial degrees of freedom to the atoms. Contrary to the case of long range interactions, this has been analyzed explicitly in this thesis and will be explicitly tackled in Chapter 3.2.

1.1 The aims of this work

Before presenting the main objectives of this thesis work, we remark that the investigation will be focused on descriptors for *linear* models. The reason behind this choice is twofold. Firstly, as already mentioned, a linear model requires much stricter conditions on the descriptors: if the correct symmetries are not implemented, or if the descriptors lack descriptive capabilities, the linear model will catastrophically fail. This is, clearly, the trade-off to pay for a model, which represents the simplest scenario in the ML landscape. However, as second reason, a linear model provides an approach, which is closer in spirit to the rationale of a physics-based one. Indeed, linear models can be interpreted and can provide more insight by themselves, if compared to a much deeper approaches such as the one of NNs. Moreover, the study of descriptors for linear models is not restrictive, since descriptors that are suitable for linear regressions can be easily exported and be used in deeper NNs architecture. Also, maybe unsurprisingly, descriptors that naturally emerge from linear-cluster expansions are easily found also in kernel-based approaches: historically, indeed, the discovery of these descriptors proceeded the other-way round, from kernel approaches to applications to linear models. Thus, the scope of descriptors for linear models goes beyond their application in linear cases.

This thesis will address three main aspects of the construction of descriptors. As will be clarified presently, these areas are interconnected as will be shown across the thesis by means of a shared and uniform formalism. In this sense, we could say that a "zero-th" aspect of the thesis is indeed the definition of a unified mathematical framework, based on the spherical harmonics in general, and the multipolar-spherical harmonics in particular.

The first analysis will be done on the study of descriptors for systems with spin-degrees of freedom. While this topic has been addressed in previous works, we will focus here on the formulation of descriptors, which are in the same spirit of the ones written for MLP and used in both linear and kernel-based methods. It will also be a first example in which the use of the multipolar-spherical harmonics will not only prove to be beneficial, but also

natural.

The second topic that will be at the core of the thesis, is the formulation of a unified framework, here denoted as Jacobi-Legendre framework. While other frameworks are available, such as the one provided by kernel and density-based representations [26, 44] and the atomic cluster expansion [25], their shared mathematical background is the coupling of angular momentum channels [45], with respect to the expansion of an "atomic-density" field. This can be appreciated in the review of Ref. [46], where the evolution of the most common descriptors, and their relations, is shown in a phylogenetic tree. Indeed, a representation in terms of the atomic-density, has a few advantages: firstly it allows to obtain a *linear* scaling in terms of the number of atoms, contrary to the scaling of a classical multi-body expansion, where the number of bodies determines the computational scaling by means of a power-law. Also, it allows to reach any body-order in the representation, by performing the appropriate couplings of an arbitrary large number of angular channels. However, it also presents a few problems. Firstly, it does not maintain interpretability of the terms involved, since the resulting symmetries are imposed only after all the the channels have been coupled. In particular, it accounts for the construction of all possible tensors from the coordinate representation, with the final contraction accounting for the projection onto the space of interest, e.g., a rotationally invariant space. While this is not necessarily a problem, since everything can be always, in principle, written in terms of the internal coordinates representation, it introduces the non-trivial step of re-casting the expressions as, at least, contractions of vectors. Another point that requires care is that the coupling of angular momenta is not unique, but depends on the coupling scheme.
Indeed, for the last point of discussion of this thesis, we will show that having different coupling schemes causes problems in the expansion of higher-body orders terms: specifically since it is required to have at least a five-body representation to have a complete representation (as investigate by the works in Refs. [47, 48] and proved in Ref. [49]), it is important to investigate how different coupling schemes affect five-body terms. In doing so, we will provide a complete representation for five-body terms, which is independent of any coupling-scheme choice.

An internal-coordinate based framework In this thesis, we will focus on a cluster-expansion based framework expanded in terms of *internal coordinates*: the internal coordinates representation is the traditional way to approach the degrees of freedom of a physical system since, among other reasons, it allows to preserve a strong and almost visual intuition of all the terms of the expansion. We will also show that, on the one hand, this choice does not sacrifice the linear scaling achieved by the atomic-density field, and on the other, it introduces a new coupling scheme, which allows us to retain all the advantages of

having a direct representation in terms of internal coordinates. This aspect will be pivotal in the last chapter of the thesis, where the construction of a coupling-scheme-independent approach will be presented.

Not only this framework allows us to bridge the gap between the atomic-density-based formalism and the internal-coordinate representation, but we will also show how it can be naturally extended to represent scalar fields, tensor, and also tensor fields. In order to do so, we will define a simple strategy to construct representations of increasing complexity, while preserving the cluster-expansion and internal-coordinate based formalism. In particular, we will first introduce a MLP for energies, where the key ideas on the constraining of the expansion and on a linear-scaling representation will be presented and examined in great detail. A core point will be the formulation of a constrained and general basis for the radial expansion of the atomic potentials (Jacobi polynomials), and an accurate selection of the basis for the angular degrees of freedom (Legendre polynomials), so to allow a seamless generalization of the model. We will then move to the representation for scalar fields, which will focus on targeting the DFT electronic density. We will show how this can impact *ab-initio* calculations, by accelerating them, without compromising the accuracy of the density. Crucially, this representation will be derived completely by the MLP, in an almost straightforward way. Finally, we will show how to target tensor and tensor-fields. In particular, the model for tensors will be obtained with a new strategy, made available by the construction of the model for the scalar field. After the construction of the appropriate scalar field, the tensorial components will be obtained by a simple integration against spherical harmonics. This is a novel approach, which does not require an explicit introduction of the Wigner-D matrices, that will be used only for the proof of the rotational covariance of the tensorial components. The consequent construction of a model for tensor fields will be almost trivial, obtained by following the same strategy adopted to define the scalar-field model from the potential one. This is one of the first models for tensor fields defined in the literature, and it shares the same properties of the other models of the framework, i.e., it will be hierarchically improvable (from the underlying cluster-expansion strategy) and interpretable (from the internal-coordinates representation).

On the multipolar-spherical harmonics The last section of the thesis will connect all the topics touched by the thesis, from the multipolar-spherical harmonic formalism, used for the model on the atomic-magnetic moments, to the atomic cluster expansion (and all the methods based on the powerspectrum and bispectrum), to the coupling introduced for the Jacobi-Legendre potential, and based on the internal coordinate representation. Indeed, we will prove that the multipolar-spherical harmonics are the natural formalism at the foundation of the most used descriptors available. Not only they will be pivotal

in deriving the atomic cluster expansion formalism in a straightforward way, but the will also provide insight on the role of intermediate coupling channels in the expansion. Indeed, they will allow us to demonstrate the necessity of retaining all the intermediate channels and, in so doing, we will prove that the scheme proposed for the atomic cluster expansion is incomplete. Finally, we will show that, instead, the coupling provided by the Jacobi-Legendre potentials, despite being derived directly from an internal coordinate representation, not only is linear in the number of atoms, but also provides a complete representation for a five-body potentials.

1.2 Thesis organization

This thesis is organized as follows, with a graphical representation shown in Fig. 1.1.

Chapter 2 is devoted to a presentation of some of the most-widely used descriptors in the literature. We will introduce descriptors devoted to representation of scalars (energies), of tensors (dipole, polarizability, etc.) and we will conclude with an overview of models predicting scalar fields, e.g., the electron density.

In Chapter 3 we will introduce powerspectrum-based descriptors for magnetic systems, where the degrees of freedom associated with the spin will be defined on the same footing of the position related ones. We will show the performance of the model when applied to an iron-cluster toy-system, defined by transversal (Heisenberg) and longitudinal (Landau) excitations.

Chapter 4 will introduce the Jacobi-Legendre potential (JLPs), which constitutes the first step in the definition of the Jacobi-Legendre (JL) formalism. We will show, in depth, how a construction over the internal coordinates of the system automatically leads to a representation that satisfies the invariance of the energy under any isometry transformation. A major part of the section will be devoted to the definition of the radial basis, and on how to apply constraints (e.g., the locality of the representation) directly on the basis set. Also, the role of the symmetry under permutation of identical atoms will be investigated in detail. We will apply the model to a challenging carbon dataset, used in the training of the GAP17 [50] potential.

In Chapter 5 we will expand the JL framework by defining a model for the electronic density (a scalar field). The chapter will be divided in two parts: in the first we will introduce a general method to obtain a scalar field within the same formalism of the JLP, and we will perform a throughough investigation of the relevant properties. The second part will be devoted to applications, in which we will show the performance of the model when applied to examples of a molecule, metallic solids and a 2D material. By using the predicted densities, we will show how *ab-initio* calculations can be accelerated, without compromising the final accuracies.

In Chapter 6 we will complete the study of the JL framework, by defining the models for tensors and tensorial fields. This chapter will be fully devoted at the presentation of the methods, with great emphasis on how to obtain expressions for covariant quantities by a JL formalism devised to address scalar fields. The relevance of this step will be twofold. On the one hand, we will show a natural way to obtain the required covariance (by means of simple integrals against spherical harmonics). On the other hand, it will fully justify the choice of the Legendre polynomials made in Chapter 4, by showing how they are the natural choice of basis when dealing with scalar products. The chapter will be concluded with the presentation of one of the first models for tensor fields, which will inherit the same properties (multi-body decomposition, locality and linearity) of all the methods of the JL framework.

Chapter 7 will finally make manifest how the multipolar spherical harmonics are the natural basis-of-choice when dealing with cluster-expanded potentials. Indeed, we will show how they allow to seamlessly derive a complete atomic cluster expansion, while also exposing the limits (incompleteness and redundancies caused by the choice of the coupling scheme) of the original formalism, in particular for body-order terms greater than four. The chapter will be concluded with the presentation of a JLP-based expansion for the five-body order terms. Not only we will preserve the equivalence with a representation in internal coordinates (despite being linear in the number of atoms), but we will also obtain a complete expansion, while avoiding any degeneracy related to the choice of the coupling scheme.

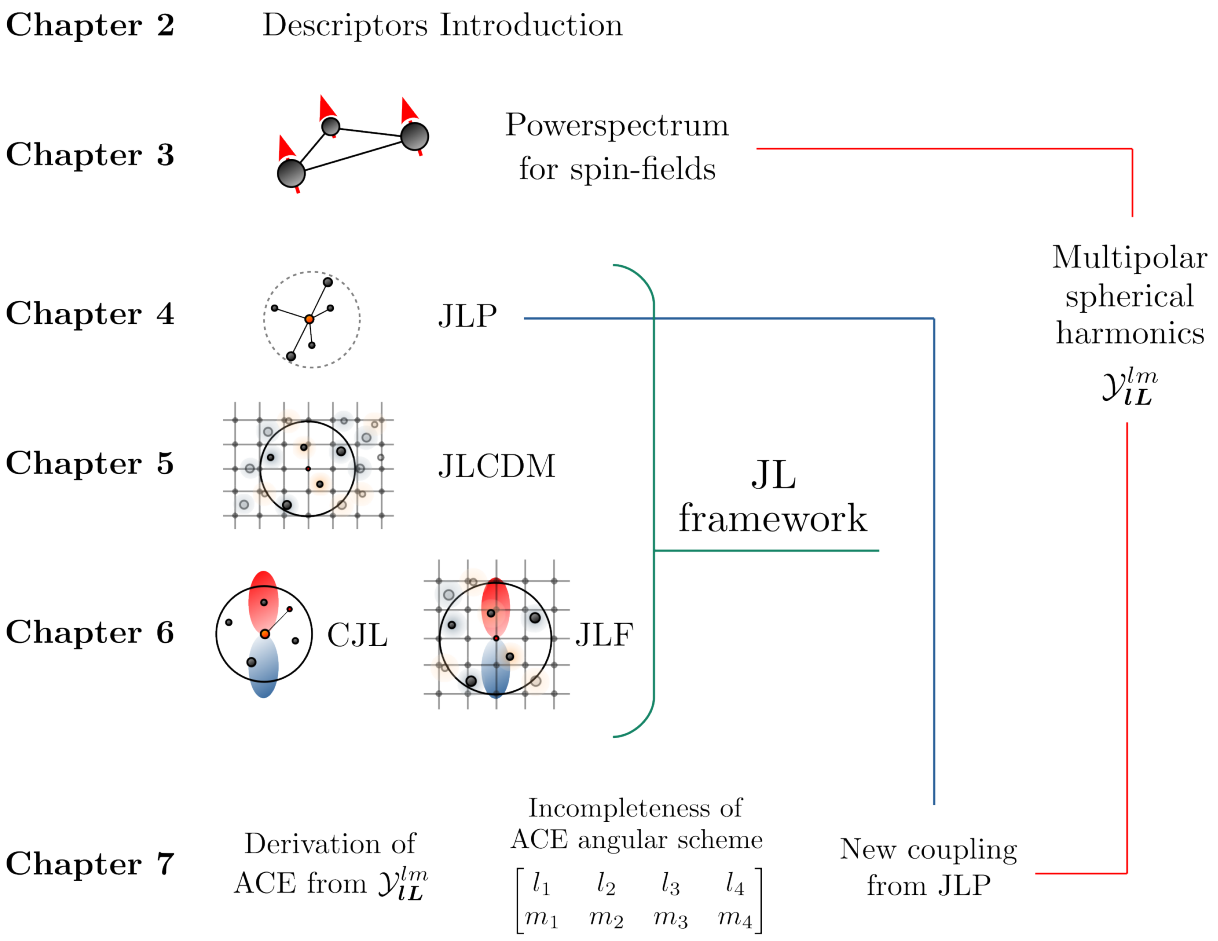

Figure 1.1: A graphical table of content for the thesis. In Ch. 2, we will present a few well-known descriptors, so to define the mathematical background of the thesis. Ch. 3 will be devoted to a method to treat magnetic materials, which will put the spin degrees of freedom on the same footing of the atomic positions. The formalism will heavily rely on the multipolar spherical harmonics. Chs. 4, 5 and 6 will be devoted to the construction of the JL framework. In particular, Ch. 4 will define the JLP, a potential to predict the PES of a system. Ch. 5 will introduce a model for the electronic charge density, so to target a scalar field. Ch. 6 will define a covariant model, which will be used to target tensors and tensor fields. Ch. 7 will revolve around the multipolar-spherical harmonics, and it will be shown that they constitute the natural basis for multi-body methods. Specifically, we will show the incompleteness of the original ACE-angular basis and will tie back to the JLP by the presentation of a new coupling-scheme independent approach to five-body order terms.

Chapter 2

Descriptors in literature and the formalism

In this chapter we are going to review several descriptors for scalars, tensors and scalar fields. The main goal of this chapter is twofold. Firstly we will introduce a coherent formalism and mathematical background for all the remainder of the thesis. The second reason is to give an overview of the descriptors, to introduce the key ideas, motivations and achievements. This will be instrumental to the following chapters, where we will introduce the general Jacobi-Legendre framework, which will push the investigation up to tensorial fields.

This chapter is essentially divided in three parts, mirroring the nature of the targets. The first one will be devoted to presenting MLP, and the second one to the main idea on how to introducing covariance in the descriptors, to target tensors. The last chapter will address the construction of descriptors for the electronic density (scalar fields). Importantly, a progressive construction has been followed, so that most of the relevant idea presented for a descriptor will be re-used and adapted in the subsequent sections.

2.1 Descriptors for Potential Energy Surfaces

The work of this thesis will focus on constructing a set of machine learning descriptors that can be easily interpreted and generalized to a vast landscape of different scenarios, ranging from scalars, e.g., *ab-initio* energies, to tensors, e.g., dipoles, polarizability, ending to tensor fields, e.g., non-spin-polarized and spin-polarized electronic densities. This chapter will be devoted to introduce, and rapidly review, the most used descriptors in literature, with the aim of building stable foundations for the subsequent main parts of this work. Following a constructive approach, we will first present descriptors for potential energy surfaces (PES), that target the energy of a system, to introduce the main ideas on

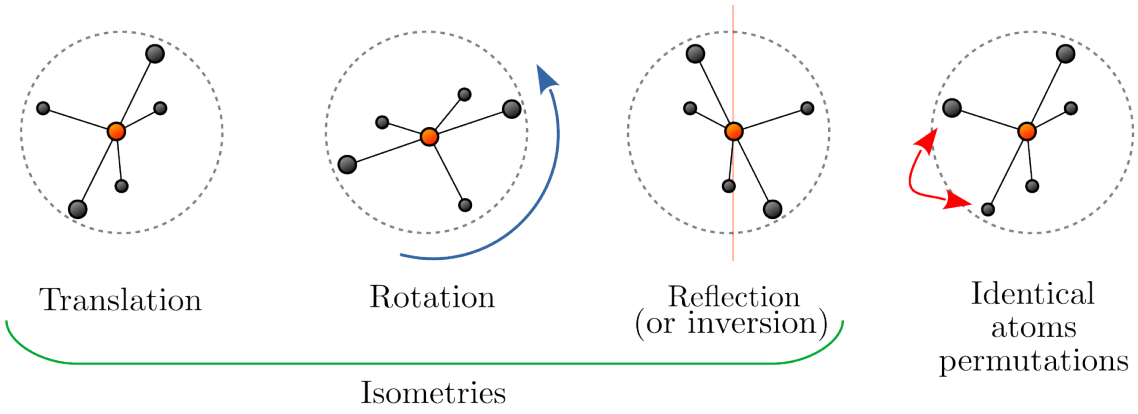

Figure 2.1: The figure shows the symmetries of scalars like the energy with respect to the atomic positions. From left to right: the symmetry under translation is automatically guaranteed by choosing an atom centered representation. The invariance under rotation and the one under inversion must be imposed with a careful choice of the features. Please note that, since in 3d space an inversion can always be written in terms of the composition between a rotation and a reflection, we will use inversion and reflection almost interchangeably in the thesis. Finally, the energy is symmetric under the swap of two identical atoms. All possible translation, rotations and reflections (or inversion), are generally addressed as isometries (distances- and angles-preserving operations).

how to construct quantities that possess desired symmetries (in this case scalars). This will be followed by a presentation on methods that aim to target tensorial quantities and scalar *fields*.

We will now proceed in discussing the Behler-Parinello symmetry functions [51], and the kernel method at the core of the Gaussian approximation potential (GAP) [26] [which is then expanded in the smooth overlap of atomic positions (SOAP) [27]]. We will also introduce the spectral neighbor analysis potential (SNAP)[24], and we will end this section with the atomic cluster expansion (ACE) [25].

2.1.1 Behler-Parrinello Symmetry Functions

The work from J. Behler and M. Parrinello of 2007 [51], introduced several key concepts that can be found in most of the subsequent works on descriptors. While the work focused mostly on Neural Networks (NN), the concepts of atomic contributions to the energies, cut-off function, and multi-body descriptors are the operative assumptions of the vast majority of the most successful descriptors and, as such, will be explored here in details.

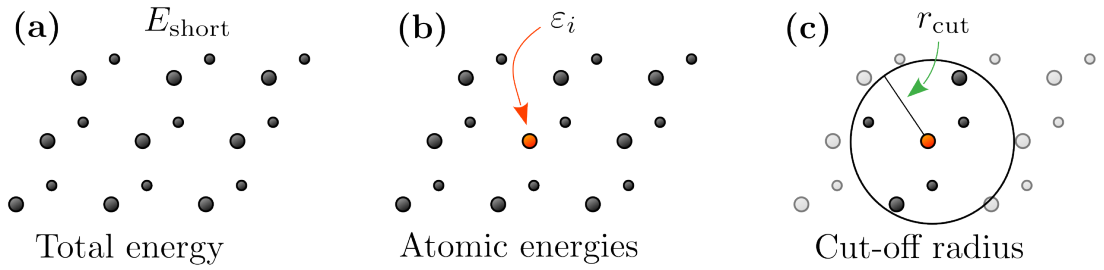

Figure 2.2: The figure shows the hypothesis underlying the definition of the vast majority of current descriptors for potential energy surfaces. (a) In this work we will explore only the short-ranged component of the energy of a system, E_{short}. (b) The short-ranged energy contribution is further separated in a sum of *atomic contributions*, in accordance with Eq. (2.1). (c) The atomic contributions are assumed to be local, accounting only for the interaction between atoms inside an optimized cut-off sphere of radius r_{cut}. This is implemented by introducing a cut-off function, f_c, as in Eq. (2.2), to make interactions progressively less relevant as the atomic distances approach the cut-off radius, all while preserving the continuity of the description.

The first assumption is that the energy can be partitioned in *atomic* contributions. Thus, if E is the energy of a system, we assume that the partition

$$E = \sum_{i}^{\text{atoms}} \varepsilon_i, \tag{2.1}$$

holds, where the sum runs over all the atoms in the system, and the *atomic energy*, ε_i is the contribution due to the i-th atom. As will be showed in the remainder of this thesis, the atomic energies are, *de-facto*, the main focus of the construction of machine learning descriptors. Focusing on this atomic terms, a crucial observation is that the only functional dependence on the system's properties can be written in terms of distances and angles. This is justified by the fact that, to efficiently describe a quantity of interest, the descriptors should possess only the symmetries of the target itself. Therefore, in the example of a scalar like the energy, the descriptors should mirror the same invariance rules of the energy. To identify, and to correctly encode the symmetries of the targets in MLPs, will constitute a central component of the thesis. In Fig. 2.1, the relevant symmetries for the energy are shown: they are translations, rotations, inversions (which, together, constitute all the possible "isometries") and invariance under permutations of identical atoms.

A problem that immediately arises when considering all possible distances and angles, is the applicability of this methods to very large or periodic systems. To address this, Behler and Parrinello introduced an optimizable cut-off distance (or radius), r_{cut}, for the atomic interactions: in this way, they enforced a locality principle, namely that atoms further

apart than the cut-off radius were not considered as interacting. To enforce smoothness and continuity of their description at the cut-off radius, they defined the following cut-off function

$$f_c(r_{ji}) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi r_{ji}}{r_{\text{cut}}}\right) + 1\right] & \text{for } r_{ji} \leq r_{\text{cut}}, \\ 0 & \text{for } r_{ji} > r_{\text{cut}}, \end{cases} \tag{2.2}$$

where $r_{ji} = |\mathbf{r}_i - \mathbf{r}_j|$, is the distance between the atoms i-th and j-th, located at $\mathbf{r}_i$ and $\mathbf{r}_j$ respectively. A concise overview of the key ideas discussed here is shown in Fig. 2.2.

The work of Behler and Parrinello then focused on the following descriptors

$$G_i^1 = \sum_{\substack{j \\ j\neq i}}^{\text{atoms}} e^{-\eta(r_{ji}-r_s)^2} f_c(r_{ji}), \tag{2.3}$$

$$G_i^2 = 2^{1-\zeta} \sum_{\substack{j,k \\ j\neq i, k\neq i}} (1 + \lambda\cos\theta_{ijk})^\zeta \, e^{-\eta(r_{ji}^2 + r_{ki}^2 + r_{jk}^2)} f_c(r_{ji}) f_c(r_{ki}) f_c(r_{jk}), \tag{2.4}$$

which are called *symmetry functions*. Here $\theta_{ijk} = \hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}$ is the scalar product between the direction connecting the j-th and the i-th atoms, $\hat{\mathbf{r}}_{ji}$, and the one connecting the k-th and i-th atoms, $\hat{\mathbf{r}}_{ki}$, namely, it contains the angular information of the three-body object centered on the atom i (the directions are defined as $\hat{\mathbf{r}}_{ji} = \mathbf{r}_{ji}/r_{ji}$). The parameter λ takes values $\lambda = \pm 1$. The sum for the first symmetry function, G_i^1, runs over all the atoms in the system that are not the central atom. It can be appreciated how G_i^1 is a sum of Gaussians centered in r_s, with width given by η, and with an embedding given by the cut-off function, to ensure the aforementioned smoothness. The total set of symmetry functions is then obtained by varying the center and the width, to increase the descriptive power of the total description. These functions are also a first example of two-body (2B) terms, since each of the addend depends only on the distance between two atoms. The other functions, G_i^2, are, instead, three-body (3B) terms, since they depend on three distances and one angle, as also mirrored by the presence of the double sum therein. The exponent ζ is another parameter that can be varied to obtain, again, more symmetry functions.

Finally, the atomic energy is assumed of the form

$$\varepsilon_i = f(\{G_i^1(r_s, \eta)\}, \{G_i^2(\eta, \zeta, \lambda)\}), \tag{2.5}$$

where the function f is approximated by a Neural-Network (NN), of optimized architecture, with the employed symmetry functions obtained by varying the inner parameters η, ζ, λ and r_s. Even if the symmetry functions display a significant descriptive power, mostly in combination with the use of NNs, they lack a few properties, such as higher-body

correlations or a systematically improvable scheme, which have been explicitly addressed in subsequent works, as we will show in the following.

2.1.2 GAP framework and SOAP

The Gaussian approximation potential (GAP) framework [26] and the smooth overlap of atomic positions (SOAP) [27], introduced new key ideas in the field, i.e., the *density trick* with its basis expansion, and the link between the investigation of higher-order correlations and the choice of appropriate coupling of angular momenta. We will now review these ideas. Please note that the same assumptions discussed in the previous section, namely, a short-ranged description, separation into atomic contributions and locality of the atomic environment, will be always implied in the following.

Explicitly, the starting point in constructing properly invariant descriptors is the atomic density

$$\rho_i(\mathbf{r}) = \sum_j^{\text{atoms}} w_{Z_j} \delta(\mathbf{r} - \mathbf{r}_{ji}) f_c(r_{ji}), \tag{2.6}$$

where $\delta(\mathbf{r})$ is a Dirac-delta function, w_{Z_i} are optimized weights that depend on the atomic species of the j-th atom, and $f_c(r_{ji})$ is the cut-off function, which ensures that the contribution from the neighbor atoms will smoothly vanish when approaching the cut-off boundary.

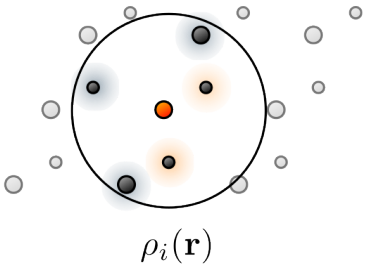

Figure 2.3: A pictorial representation of the atomic density defined in (2.6).

The density gives an atom-centered representation of the local environment, as can be appreciated by the the fact that all the atomic positions in Eq. (2.6) are always taken with respect to the position of the i-th atom. It is not necessary for the density to be made of delta functions only: indeed other choices of function, such as Gaussians centered in the atomic positions can be used[1]. In Fig. 2.3, we show a pictorial representation of the density, where the "localizing" function is placed on top of the atoms in the environment (in the case of a Dirac-δ, it would be an infinite spike), while the colors represent different species, determined by the weights.

The density is a function defined on the 3-dimensional space [albeit with compact support] and, as such, depends on the position vector $\mathbf{r}$. Thus, it can be expanded in a complete basis. Let us first separate the radial and angular components of the vector $\mathbf{r}$ as in

$$\rho_i(\mathbf{r}) = \rho_i(r, \hat{\mathbf{r}}), \tag{2.7}$$

[1]This is actually done in the GAP formalism. Here, however, we will treat only the case of delta functions, because they allow for a simpler analytical investigation.

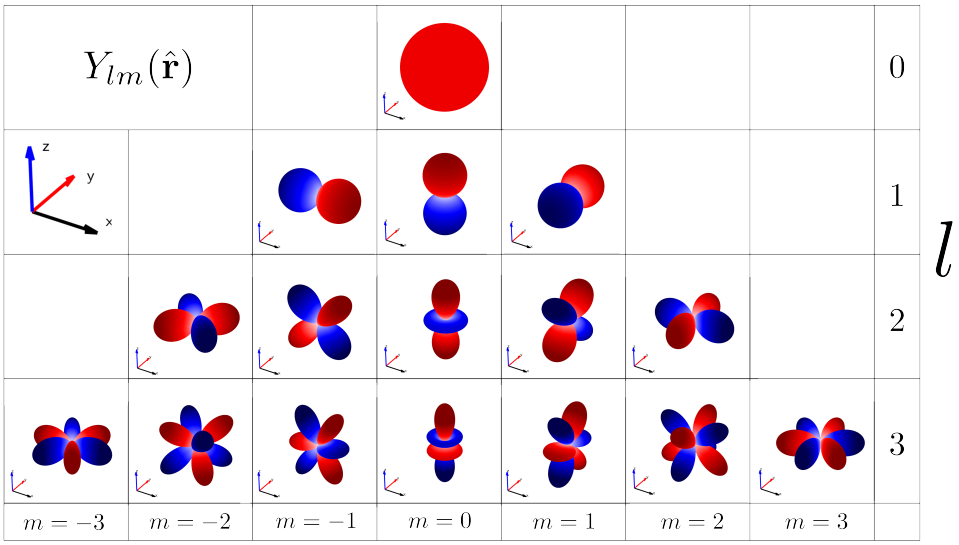

Figure 2.4: The spherical harmonics, Y_l^m, are at the core of the descriptors presented along this thesis. Here, the first 16 *real*-spherical harmonics are shown (they are obtained by the standard spherical harmonics by means of a unitary transformation, as shown in Ref. [52]). The plots are polar, with blue and red indicating positive and negative values, respectively.

so that we can expand the density in terms of radial basis of choice $R_{nl}(r)$ and in spherical harmonics $Y_l^m(\hat{\mathbf{r}})$ [45, 53], respectively. The choice of spherical harmonics will be of paramount importance in constructing rotationally invariant quantities, as will be shown shortly (see Fig. 2.4 for a visual representation). Note that this expansion can always be performed, since the spherical harmonics form an orthonormal basis for the squared-integrable functions on the sphere, namely, they form a basis of $L^2(S^2)$. This allows us to write

$$\rho_i(\mathbf{r}) = \sum_{nlm} c_{inlm} R_{nl}(r) Y_l^m(\hat{\mathbf{r}}), \tag{2.8}$$

where the expansion coefficients are given by

$$c_{inlm} = \int \mathrm{d}\mathbf{r}\, \rho_i(\mathbf{r}) R_{nl}(r) Y_l^{m*}(\hat{\mathbf{r}}) = \sum_j^{\text{atoms}} w_{Z_j} R_{nl}(r_{ji}) Y_l^{m*}(\hat{\mathbf{r}}_{ji}) f_c(r_{ji}). \tag{2.9}$$

Please note that we are assuming that the radial functions are real, orthonormal and complete. This last equation justifies the choice of the Dirac-delta functions, since they allow to perform the integration in a straightforward matter, by directly inserting Eq. (2.6) in the integral above. We have obtained that the description of the local environment is effectively encoded in the coefficients c_{inlm}. Crucially, the evaluation of the expression c_{inlm} depends linearly on the number of atoms in the neighborhood of the i-th atom. This property is the reason behind the construction of the density in the first place, and will be

shared by all the quantities that will be derived in this section. However, we still have to address the two questions on how to construct isometrically-invariant quantities, and the function on which to constitute the model to be used to approximate the energy in the fitting procedure. Note that, practically, we will always truncate the expansion at some optimized $n_{\max}$ (and also $l_{\max}$ if the angular expansion is decoupled from the radial one).

The powerspectrum The simplest possible rotationally invariant quantity that can be built is the *powerspectrum* defined, in the same spirit of the Fourier's spectra powerspectrum, as

$$p_{inn'l} := \sum_m c_{inlm} c^*_{in'lm} = \sum_m (-1)^m c_{inlm} c_{in'l-m}, \tag{2.10}$$

where $c^*_{n'lm} = (-1)^m c_{in'l-m}$ is the complex conjugate of c_{inlm}. We will now go into detail on the proof of the rotational invariance of the powerspectrum, to solidify ideas that will be crucial in the remainder of the thesis. The proof relies on the mixing rule of the spherical harmonics under a rotation, that explicitly reads [45]

$$Y_l^m(\hat{R}\hat{\mathbf{r}}) = \sum_{m'=-l}^{l} D^{l*}_{mm'}(\mathcal{R}) Y_l^{m'}(\hat{\mathbf{r}}), \tag{2.11}$$

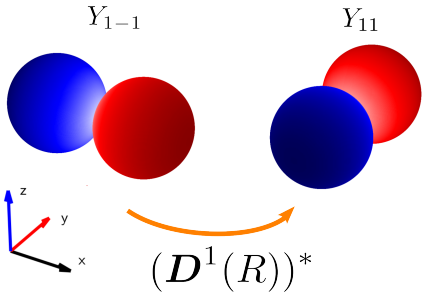

Figure 2.5: A counterclockwise rotation around the z-axis of $\pi/2$ is shown. The real spherical harmonic Y_{11} is mapped into Y_{1-1}.

where the matrix $\hat{R}$ is a representation of the rotation induced by the Euler angles $\mathcal{R}$, so that $\hat{R}\hat{\mathbf{r}}$ represents the versor obtained by applying the rotation to the original versor $\hat{\mathbf{r}}$. Here, the matrix $D^l_{mm'}$ is the so-called Wigner D-matrix, parameterized by $\mathcal{R}$. The mixing of spherical harmonics with different magnetic quantum number m can be easily understood in terms of Fig. 2.5. Importantly, however, the l is unaffected by the rotation, namely, spaces of different angular quantum number do not mix. The D-matrices are unitary, so they satisfy the relation

$$\sum_m D^l_{mm_1}(\mathcal{R}) D^{l*}_{mm_2}(\mathcal{R}) = \delta_{m_1 m_2}. \tag{2.12}$$

By looking at the explicit expression for the expansion coefficients c_{inlm} of Eq. (2.9), we notice that they incorporate a sum of complex-conjugated spherical harmonics. By applying a rotation to the whole system, i.e., applying the same rotation to each one of

the spherical harmonics, we deduce the following transformation rule

$$\hat{R} \,:\, c_{inlm} \rightarrow \sum_{m'} D^{l}_{mm'}(\mathcal{R}) c_{inlm'}. \tag{2.13}$$

Thus, from the definition of the powerspectrum, Eq. (2.10), we can directly prove its rotational invariance by noting that

$$p_{inn'lm} = \sum_{m} c_{inlm} c^{*}_{in'lm} \xrightarrow{R} \sum_{m_1 m_2} c_{inlm_1} c^{*}_{in'lm_2} \overbrace{\sum_{m} D^{l}_{mm_1} D^{l*}_{mm_2}}^{\delta_{m_1 m_2}} = \sum_{m} c_{inlm} c^{*}_{in'lm} = p_{inn'lm}, \tag{2.14}$$

where, in the last step, we re-labelled the dummy index m_1. Another important property of the powerspectrum is its invariance under inversion of the atomic positions, also called parity invariance. This can be shown by means of the addition theorem of the spherical harmonics [45, 53], a crucial property that will be extensively used in the remainder of this work. This theorem determines the connection between the spherical harmonics and the *Legendre polynomials*[2], establishing the following identity [54–56]

$$P_l(\cos\theta_{ijk}) = P_l(\hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}) = \frac{4\pi}{2l+1} \sum_{m} Y_l^{m*}(\hat{\mathbf{r}}_{ji}) Y_l^{m}(\hat{\mathbf{r}}_{ki}), \tag{2.15}$$

where the argument of the Legendre polynomials is the dot product between the versors $\hat{\mathbf{r}}_{ji}$ and $\hat{\mathbf{r}}_{ki}$. This, when inserted in Eq. (2.10), gives the alternative formulation of the powerspectrum

$$p_{inn'l} = \frac{2l+1}{4\pi} \sum_{jk}^{\text{atoms}} w_{Z_j} w_{Z_k} f_c(r_{ji}) f_c(r_{ki}) R_{nl}(r_{ji}) R_{n'l}(r_{ki}) P_l(\hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}). \tag{2.16}$$

This expression has the advantage that it manifestly shows the invariance of the powerspectrum under both rotations and inversions: it depends only on distances and on the scalar product $\hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}$, which are all isometrically invariant. Also, it allows to make a direct comparison with the Behler-Parrinello symmetry functions reported in the previous section, since $\cos\theta_{ijk} = \hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}$. In particular, we can appreciate how the powerspectrum is analogous to a three-body (3B) term, since it relays on two distances and one angle. The evaluation of this expression, however, does not scale linearly with the number of atoms inside the cut-off sphere, since it requires the calculation of the Legendre polynomials for each *pair* of atoms in the neighborhood.

[2]Please note that, among all the properties of the Legendre polynomials, we will make large use of the fact that they form a complete and orthonormal basis on the interval $[-1, 1]$.

The bispectrum The powerspectrum, depending only on the positions of three atoms (the central one, and two neighbours), is not complete with respect to the atomic environments defined by $\rho(\mathbf{r})$, in the sense that two different atomic environments, non connected by rotations or reflections, can have the same powerspectrum. Also, more generally, two different functions defined on the sphere can lead to the same powerspectrum, as explicitly shown by taking the two functions (from Ref. [27])

$$\begin{cases} f_1 = Y_2^2 + Y_2^{-2} + Y_3^3 + Y_3^{-3}, \\ f_2 = Y_2^1 + Y_2^{-1} + Y_3^2 + Y_3^{-2}. \end{cases} \tag{2.17}$$

For this reason, we will now define higher-body order invariant quantities, by appropriately connecting expansion coefficients c_{inlm} with different angular momentum channels. If the connection is carried using the Clebsh-Gordan (CG) coefficients, defined as [45]

$$C^{lm}_{l_1 m_1 l_2 m_2} = \langle lm | l_1 m_1 l_2 m_2 \rangle \,, \tag{2.18}$$

then we can construct the so-called *bispectrum* by means of the following coupling

$$b_{i\,ll_1l_2}^{nn_1n_2} := \sum_{mm_1m_2} c^*_{inlm} C^{lm}_{l_1 m_1 l_2 m_2} c_{in_1 l_1 m_1} c_{in_2 l_2 m_2}. \tag{2.19}$$

The rotational invariance of these objects can be easily verified[3] with the same strategy used above, and by employing the orthogonality of the CG coefficients and their relations with the Wigner-D matrices [45]. Explicitly, using

$$\sum_{m_1 m_2} C^{L_1 M_1}_{l_1 m_1 l_2 m_2} C^{L_2 M_2}_{l_1 m_1 l_2 m_2} = \delta_{L_1 L_2} \delta_{M_1 M_2}, \tag{2.20}$$

and

$$\sum_{m_1 m_2} C^{lm}_{l_1 m_1 l_2 m_2} D^{l_1}_{m_1 m_1'}(\mathcal{R}) D^{l_2}_{m_2 m_2'}(\mathcal{R}) = \sum_{m'} D^{l}_{mm'}(\mathcal{R}) C^{lm'}_{l_1 m_1' l_2 m_2'}, \tag{2.21}$$

we can write

$$\begin{aligned} \mathcal{Y}^{lm}_{l_1 l_2}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2) &:= \sum_{m_1 m_2} C^{lm}_{l_1 m_1 l_2 m_2} Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1) Y^{m_2}_{l_2}(\hat{\mathbf{r}}_2) \\ &\xrightarrow{R} \sum_{m_1' m_2'} \left[\sum_{m_1 m_2} C^{lm}_{l_1 m_1 l_2 m_2} D^{l_1 *}_{m_1 m_1'} D^{l_2 *}_{m_2 m_2'} \right] Y^{m_1'}_{l_1}(\hat{\mathbf{r}}_1) Y^{m_2'}_{l_2}(\hat{\mathbf{r}}_2) = \\ &= \sum_{m'} D^{l*}_{mm'} \sum_{m_1' m_2'} C^{lm'}_{l_1 m_1 l_2 m_2} Y^{m_1'}_{l_1}(\hat{\mathbf{r}}_1) Y^{m_2'}_{l_2}(\hat{\mathbf{r}}_2) = \sum_{m'} D^{l*}_{mm'} \mathcal{Y}^{lm'}_{l_1 l_2}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2), \end{aligned} \tag{2.22}$$

[3] Please, note that, here, we prove the rotation invariance by explicit calculation, not by exploiting representation of the SO(3) group, as has been done, for example, in Ref. [27].

where we have implied the dependence on the Euler angles, and we made use of the fact that the CG coefficients are real. In going from the first to the second line we applied a rotation $\hat{R}$ to both $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$, so that the above result can be written as

$$\mathcal{Y}^{lm}_{l_1 l_2}(\hat{R}\hat{\mathbf{r}}_1, \hat{R}\hat{\mathbf{r}}_2) = \sum_{m'} D^{l*}_{mm'}(\mathcal{R}) \mathcal{Y}^{lm'}_{l_1 l_2}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2). \tag{2.23}$$

The functions $\mathcal{Y}^{lm}_{l_1 l_2}$ are called *bipolar-spherical harmonics* [45], and form a complete and orthonormal set for the two points functions on the sphere. Comparing the result from (2.22) with the transformation rules for the spherical harmonics in (2.11), we see that the contraction of two spherical harmonics, $Y^{m_1}_{l_1}$ and $Y^{m_2}_{l_2}$, with a CG coefficient, $C^{lm}_{l_1 m_1 l_2 m_2}$, into the bipolar spherical harmonics, $\mathcal{Y}^{lm}_{l_1 l_2}$, obeys the same transformation rules of the Y^m_l spherical harmonic[4]. Here, this relation allows to prove that the bispectrum is invariant under rotation: indeed, the coupling

$$p^{lm}_{in_1 n_2 l_1 l_2} = \sum_{m_1 m_2} C^{lm}_{l_1 m_1 l_2 m_2} c_{in_1 l_1 m_1} c_{in_2 l_2 m_2}, \tag{2.24}$$

projects the product of the expansion coefficients in the (l, m) space of angular momentum and so, again, behaves like a spherical harmonics of indexes (l, m) under rotation. Therefore, if we perform the further contraction

$$b_{i\,\,ll_1 l_2}^{\,\,nn_1 n_2} = \sum_m c^*_{inlm} p^{lm}_{in_1 n_2 l_1 l_2} \tag{2.25}$$

we are constructing an object that follows the same transformation rules that were used to prove that the powerspectrum is invariant [please, compare with Eqs. (2.10) and (2.14)]. This proves that the bispectrum is indeed invariant under rotation.

Lastly, we remark that the bispectrum components acquire a phase factor of $(-1)^{l+l_1+l_2}$ under inversion. This can be easily proven by means of the transformation rule

$$Y^m_l(-\hat{\mathbf{r}}) = (-1)^l Y^m_l(\hat{\mathbf{r}}), \tag{2.26}$$

and by noticing that each term in the bispectrum contains three spherical harmonics, one for each angular momentum channel. Ultimately, this means that if we restrict our analysis to components for which the sum $l + l_1 + l_2$ is even, then the resulting quantity is invariant under the action of any element of the orthogonal group O(3), namely, it is isometrically invariant.

[4]This observation is of paramount importance and will lead the discussion for the rest of this chapter.

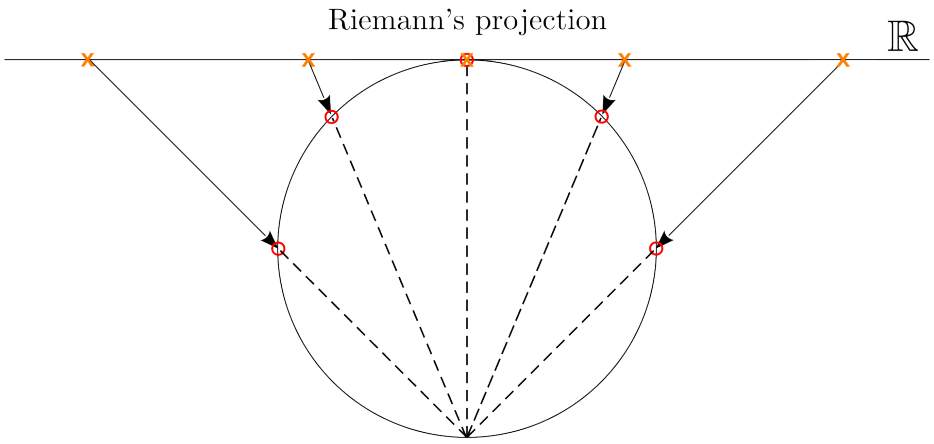

Figure 2.6: One of the possible 1-dimensional Riemann mappings. In this example, the real line, $\mathbb{R}$, is mapped on the circumference shown. The projection is done by following the dashed lines, with the north-pole being the fixed point of the projection.

Four-dimensional bispectrum We have already defined two invariant quantities that can be used as descriptors, the powerspectrum and the bispectrum. However, our attention has focused, so far, only on angular expansions of the atomic density. A way to incorporate also the radial information, and thus remove the need to select a specific radial basis, is to use a Riemann projection, as explicitly shown in both Refs. [26][27]. An example of one dimensional Riemann mapping is reported in Fig. 2.6. Explicitly, by considering a four-dimensional sphere S^3, with radius r_0, we can perform the following mapping from the 3d polar representation to a 4d one

$$\mathbf{r} = \begin{pmatrix} r\sin\theta\cos\phi \\ r\sin\theta\sin\phi \\ r\cos\theta \end{pmatrix} \longrightarrow \mathbf{u} = \begin{pmatrix} r_0\sin\theta_0\sin\theta\cos\phi \\ r_0\sin\theta_0\sin\theta\sin\phi \\ r_0\sin\theta_0\cos\theta \\ r_0\cos\theta_0 \end{pmatrix}, \tag{2.27}$$

where the new polar angle θ_0, being defined as $\theta_0 = \pi r/r_0$, incorporates the radial information. Here r_0 is a parameter, usually set to $r_0 = \frac{4}{3}r_{\text{cut}}$. The local density $\rho_i(\mathbf{r})$, is analogously promoted from a function in 3-dimensional (3d) space, to a function defined on the surface of a 4-dimensional sphere, and therefore can be expanded in hyperspherical harmonics, $U^l_{mm'}(\hat{\mathbf{u}})$ [the higher-dimensional analogous of the spherical harmonics [53]]. Explicitly,

$$\rho_i(\mathbf{r}) \xrightarrow{\mathbf{r}\to\hat{\mathbf{u}}} \rho_i(\hat{\mathbf{u}}) = \sum_j^{\text{atoms}} w_{Z_j}\delta(\hat{\mathbf{u}} - \hat{\mathbf{u}}_{ji}) f_c(r_{ji}) = \sum_l \sum_{m,m'=-l}^{l} c^l_{imm'} U^l_{mm'}(\hat{\mathbf{u}}), \tag{2.28}$$

with the expansion coefficients $c^j_{mm'}$ again obtained by an appropriate integration as

$$c^l_{imm'} = \int \mathrm{d}\hat{\mathbf{u}}\, \rho_i(\hat{\mathbf{u}}) U^{l*}_{mm'}(\hat{\mathbf{u}}) = \sum_j^{\text{atoms}} w_{Z_j} U^{l*}_{mm'}(\hat{\mathbf{u}}_{ji}) f_c(r_{ji}), \tag{2.29}$$

where $\mathrm{d}\hat{\mathbf{u}} = \sin^2\theta_0 \sin\theta \sin\phi \mathrm{d}\theta_0 \mathrm{d}\theta \mathrm{d}\phi$, to span the whole 4-dimensional (4d) surface. It can be shown that we can write similar contractions to that done for the 3d-powerspectrum and bispectrum, and again obtain rotationally invariant quantities. In this way, we can define the 4d-powerspectrum as

$$p_{il} := \sum_{mm'} c^{l*}_{imm'} c^l_{imm'} = \frac{l+1}{2\pi^2} \sum_{jk}^{\text{atoms}} w_{Z_j} w_{Z_k} f_c(r_{ji}) f_c(r_{ki}) C^1_l(\hat{\mathbf{u}}_{ji} \cdot \hat{\mathbf{u}}_{ki}), \tag{2.30}$$

where we used the generalized-addition theorem for spherical 4d-hyperspherical harmonics

$$C^1_l(\hat{\mathbf{u}} \cdot \hat{\mathbf{u}}') = \frac{2\pi^2}{l+1} \sum_{mm'} U^{l*}_{mm'}(\hat{\mathbf{u}}) U^l_{mm'}(\hat{\mathbf{u}}'), \tag{2.31}$$

with C^1_l being Gegenbauer polynomials (four-dimensional analogues of the Legendre polynomials) [53], evaluated on the scalar product between the two versors $\hat{\mathbf{u}}$ and $\hat{\mathbf{u}}'$. We can build an intuition around this scalar product, $\hat{\mathbf{u}}_{ji} \cdot \hat{\mathbf{u}}_{ki}$, by computing it explicitly: considering that the powerspectrum is invariant under rotation, we can always take the versors to be in the form

$$\hat{\mathbf{u}}_{ji} = \begin{pmatrix} 0 \\ 0 \\ \sin(\pi r_{ji}/r_0) \\ \cos(\pi r_{ji}/r_0) \end{pmatrix}, \quad \text{and} \quad \hat{\mathbf{u}}_{ki} = \begin{pmatrix} \sin(\pi r_{ki}/r_0)\sqrt{1-(\hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{ki})^2} \\ 0 \\ \sin(\pi r_{ki}/r_0)\hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{ki} \\ \cos(\pi r_{ki}/r_0) \end{pmatrix}, \tag{2.32}$$

and so the powerspectrum is invariant for any transformation that does not modify the scalar product

$$\hat{\mathbf{u}}_{ji} \cdot \hat{\mathbf{u}}_{ki} = \cos(\pi r_{ji}/r_0))\cos(\pi r_{ki}/r_0)) + \sin(\pi r_{ji}/r_0)\sin(\pi r_{ki}/r_0)\hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{ki}. \tag{2.33}$$

On the one hand, this expression shows that if the distances are left untouched, then the scalar product is invariant under rotations and reflections of the regular 3d-space, inheriting this property from the scalar product $\hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}$ (this can be also deduced by noticing that rotation of the 3d space form a subgroup of the rotations of the 4d one). However, this expression is invariant under *any* possible rotation of the 4d space, meaning that now we can perform rotations also around the new axis (the one defining the polar

angle θ_0) that, however, encodes the distances in the 3d-space. This implies that we have introduced new symmetries that involve, simultaneously, rotations and distortions of real-space distances. This can be further appreciated by the fact that the powerspectrum is invariant under any transformation that leaves the scalar product in Eq. (2.33) unchanged. These unphysical symmetries are usually addressed by forcing the neighborhood to be anchored to the north pole, explicitly modifying the density as

$$\rho_i(\hat{\mathbf{u}}) = \delta(\hat{\mathbf{u}}) + \sum_{j}^{\text{atoms}} w_{Z_j}\delta(\hat{\mathbf{u}} - \hat{\mathbf{u}}_{ji}) f_c(r_{ji}). \tag{2.34}$$

While this solves the problem from a practical point of view (the calculations are always done with the central atoms at the north pole), moreover it does not break the symmetries of the functions at play. We can go one step further, and construct also the 4-dimensional bispectrum, as

$$B_{ill_1l_2} = \sum_{\substack{mm'\\ m_1m_1'\\ m_2m_2'}} c^{l*}_{im'm} H^{lmm'}_{l_1m_1m_1',l_2m_2m_2'} c^{l_1}_{im_1'm_1} c^{l_2}_{im_2'm_2}, \tag{2.35}$$

which can, again, be proven to be invariant under any 4d rotation. The constants $H^{lmm'}_{l_1m_1m_1',l_2m_2m_2'}$, 4d analogues of the CG coefficients, are defined as

$$H^{lmm'}_{l_1m_1m_1',l_2m_2m_2'} = C^{lm}_{l_1m_1l_2m_2} C^{lm'}_{l_1m_1'l_2m_2'}. \tag{2.36}$$

As for the powerspectrum, this definition of bispectrum components is invariant also under unphysical rotations around the new polar axis.

The powerspectrum and the bispectrum are at the core of the modern formulation for descriptors in MLPs. They are ubiquitous, and constantly emerge from the most disparate scenarios. This is not surprising, since they are invariants that naturally emerge from the coupling of spherical harmonics. Indeed, they create a strong link between the search for new rotational invariants and the study of the coupling of angular momentum channels. The link is, at its core, established by the role of the CG coefficients in projecting the spherical harmonics onto the correct, invariant space. Having presented the most important descriptors for this thesis work, we will now proceed in discussing another pillar of MLPs, namely, the choice of the interpolating function connecting the descriptors to the targets.

The kernel method Our aim is to choose a way to reproduce the atomic energy ε_i, and so is, essentially, to make a choice of functional forms. A possible approach is the

"kernel" one, represented by an expression like

$$\varepsilon_i(\mathbf{q}) = \sum_{k=1}^{\text{config.}} \alpha_k K(\mathbf{q}, \mathbf{q}^{(k)}). \tag{2.37}$$

Here the vector $\mathbf{q}$ contains the descriptors of the local environment of which we want to calculate the energy, namely, the powerspectrum and/or the bispectrum. The vectors $\mathbf{q}^{(k)}$ contain the descriptors for the k-th atomic environment (or atomic *configuration*) in the training set. The coefficients α_k are determined by the fitting procedure, and $K(\mathbf{q}, \mathbf{q}^{(k)})$ is a similarity kernel, giving a measure on how similar the environments described by $\mathbf{q}$ and $\mathbf{q}^{(k)}$ are. One of the most useful interpretation of the kernel K is the one of a covariance matrix in the framework of Gaussian processes [see Ref. [57]]: this shows immediately that the kernel must be symmetric and positive definite. With the lens of the same general framework, an analytical formula for the coefficients α_k can also be obtained [57]

$$\alpha_k = \sum_{k'=1}^{\text{config.}} (\boldsymbol{K} + \gamma^2 \mathbb{1})^{-1}_{kk'} \varepsilon_{k'}, \tag{2.38}$$

where the matrix $\boldsymbol{K}$ contains the similarities between all the configurations in the dataset, namely, $K_{kk'} := K(\mathbf{q}^{(k)}, \mathbf{q}^{(k')})$, while ε_k is the atomic energy of the k-th configuration. The parameter γ is the regularisation constant. We also remark that we usually have access to the total energy, and not to the atomic ones. Therefore, assuming the same coefficients for all the local environments in the system, the expression to fit is usually

$$E = \sum_{i}^{\text{atoms}} \varepsilon_i(\mathbf{q}_i) = \sum_{k=1}^{\text{config.}} \alpha_k \sum_{i}^{\text{atoms}} K(\mathbf{q}_i, \mathbf{q}^{(k)}). \tag{2.39}$$

Here, we made the approximation that we use the same kernels for each pair of environments, implying that all the environment are formally equivalent. In general, this does not hold, and we could differentiate, for example, between different species.

The Gaussian Approximation Potential (GAP) framework

We are now in position to define the the Gaussian approximation potential (GAP) [26] framework as the one that encompasses the MLPs that are based on the 4d bispectrum components (calculated on Gaussian local densities), and on a kernel approximation of the energy. A first example of a kernel is given by the Gaussian one

$$K(\mathbf{B}_i, \mathbf{B}^{(k)}) = \exp\left\{ -\frac{1}{2} \sum_{ll_1l_2} \left[\left(B_{ill_1l_2} - B_{kll_1l_2} \right) / \theta_{ll_1l_2} \right]^2 \right\}, \tag{2.40}$$

where $\mathbf{B}_i$ is a vector containing all the 4d bispectrum components, $B_{ill_1l_2}$, defined in Eq. (2.35), and the superscript (k) indicates the k-th reference configuration in the training set. The widths $\theta_{ll_1l_2}$ are tunable parameters that must be optimized.

The Smooth Overlap of Atomic Positions (SOAP) The smooth overlap of atomic positions (SOAP) [27], while defined inside the GAP framewok, follows a specific kernel construction. It is based on the overlap integral S, defined as

$$S(\rho, \rho') = \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r})\rho'(\mathbf{r}). \tag{2.41}$$

where, for readability, we dropped the atomic index, i, from the local densities. Then, the rotationally invariant kernel

$$k(\rho, \rho') = \int \left|S(\rho, \hat{R}\rho')\right|^n \mathrm{d}\hat{R} = \int \mathrm{d}\hat{R} \left|\int \mathrm{d}\mathbf{r}\rho(\mathbf{r})\rho'(\hat{R}\mathbf{r})\right|^n, \tag{2.42}$$

is constructed, with n being a positive integer, and where the (Haar) integral[5] is performed over all the possible rotations. This kernel is then obtained by integrating the similarity between any two relative rotations of the environments described by ρ and ρ'. The exponent n magnifies the effect of the overlap[6], while the absolute value guarantees that the resulting object is always positive. Let us explicitly consider the two cases $n = 2$ and $n = 3$. Firstly, it holds that

$$S(\rho, \hat{R}\rho') = \sum_{\substack{nlm \\ n'l'm'm''}} c_{nlm} c'^{*}_{n'l'm''} D^l_{m'm''}(\mathcal{R}), \tag{2.43}$$

where we conveniently used the fact that ρ is real, and where c_{nlm} and c'_{nlm} are the expansion coefficients of ρ and ρ', respectively. Note that the above expression holds only if the radial basis is orthonormal, namely

$$\int \mathrm{d}r\, r^2 R_{nl}(r) R_{n'l'}(r) = \delta_{nn'}\delta_{ll'}. \tag{2.44}$$

From this, and by using the orthogonality of the Wigner-D matrices [45], that reads

$$\int \mathrm{d}R\, D^l_{mm'}(\mathcal{R}) D^{l_1 *}_{m_1 m'_1}(\mathcal{R}) = \frac{8\pi^2}{2l+1}\delta_{ll_1}\delta_{mm_1}\delta_{m'm'_1}, \tag{2.45}$$

we have that the $n = 2$ case reduces to the inner product between the powerspectra of ρ

[5] This integral can be performed by a parameterization in terms of the Euler angles.
[6] And, as will be shortly shown, increases the body order of the description.

and ρ'. Explicitly

$$k(\rho, \rho') = \sum_{nn'l} p_{nn'l} p'_{nn'l}, \tag{2.46}$$

where $p_{nn'l}$ is equivalent to the one defined in Eq. (2.10), up to a factor $\sqrt{8\pi^2/(2l+1)}$. Please note that, while the relations above hold, in the original SOAP formalism the density is written in terms of atom centered Gaussians and not Dirac-delta functions. The results are formally equivalent to the ones obtained here, but the actual analytical form of the expansion coefficients, c_{nlm}, changes. Analogously, it can be proven that the $n = 3$ cases can be written as the inner product between two bispectrum components. Explicitly

$$k(\rho, \rho') = \sum_{\substack{nn_1n_2 \\ ll_1l_2}} b^{nn_1n_2}_{ll_1l_2} b'^{nn_1n_2}_{ll_1l_2}, \tag{2.47}$$

where b and b' are proportional to the bispectrum components defined in Eq. (2.19), for the densities ρ and ρ', respectively. Finally, the formalism used the scaled kernel defined as

$$K(\rho, \rho') := \left(\frac{k(\rho, \rho')}{\sqrt{k(\rho, \rho) k(\rho', \rho')}} \right)^{\zeta}. \tag{2.48}$$

The right-hand side is elevated to an integer power ζ, to magnify the variations of the kernel with respect to changes in the atomic positions.

Potentials based on the GAP-framework have proven to be an important tool in investigating large portions of the energetic landscape of many systems of interest, with a significant reduction of the computational costs, all while preserving *ab-initio*-accuracy levels (see, for example, Ref. [58]). Not only were they capable of reaching high accuracy for specific systems (such as graphene in Ref. [59]), and interpolate between different phases, but they have been successfully applied also to large systems, for example in the case of silicon [60] and carbon [50, 58]. However, as shown in Ref. [61], there is a quite severe trade-off between the model performance and the computational overhead. This problem can be solved either by adopting a strategy to make the kernel calculations more efficient (such as in the TurboGAP of Ref. [62]), or by adopting new models based on the same descriptors, as shown in the next section.

2.1.3 SNAP

The spectral neighbours analysis potential (SNAP) [24] utilizes the 4d bispectrum $B_{jj_1j_2}$ in a *linear* fit of the energy (note that an inner-product kernel is equivalent to a linear model). In this way, the atomic energy is approximated by

$$\varepsilon_i \simeq \varepsilon_i^{\mathrm{SNAP}}(\mathbf{B}^i) = \beta_0^{Z_i} + \boldsymbol{\beta}^{Z_i} \cdot \mathbf{B}^i. \tag{2.49}$$

Here, $(\beta_0^{Z_i}, \boldsymbol{\beta}^{Z_i})$, are the coefficients to be fitted, while the vector $\mathbf{B}^i$ contains all the bispectrum components evaluated on the environment of the i-th atom. Despite is apparent simplicity and compactness (usually a SNAP model possesses 56 components for each species involved), it proved to be a very successful model in predicting the energetic landscape, as shown, for example, in the case of alloys and metals [17][63]. Also, by the exploiting the linearity of the model, the forces and the components of the stress tensor assume the simple form [note that $E = \sum_i \varepsilon_i$]

$$\begin{cases} \mathbf{F}^j = -\nabla_j E = -\boldsymbol{\beta} \cdot \sum_i^{\text{atoms}} \dfrac{\partial \mathbf{B}^i}{\partial \mathbf{r}_j}, \\ \mathbf{W} = -\sum_{j=1}^{N} \mathbf{r}_j \otimes \nabla_j E = -\boldsymbol{\beta} \cdot \sum_{j=1}^{N} \mathbf{r}_j \otimes \sum_{i=1}^{N} \dfrac{\partial \mathbf{B}^j}{\partial \mathbf{r}_j}. \end{cases} \tag{2.50}$$

We remark that, despite using the 4d bispectrum descriptors, analogous formulas can be obtained for any descriptors that describe the atomic energies by means of a linear regression.

Among the most important extension for SNAP, we mention here the quadratic SNAP (qSNAP, Ref. [64]), which expanded the linear regression in order to contain also quadratic products of bispectrum components, and the "explicit multi-element extension" of the SNAP potential (EME-SNAP, Ref. [65]) that tackled the problem of constructing a representation for multi-element systems. While, in general, these methods enhance the accuracy of the potential, they also largely increase the computational cost.

In the following section we will consider another model which combines the linearity of the SNAP with the descriptive power of the kernel methods. As will be made clear, this model will encompass both the powerspectrum and bispectrum, introducing also a possible new class of higher-body descriptors.

2.1.4 ACE

The Atomic Cluster Expansion (ACE) [25], as the name suggests, is a MLP which is based on the cluster expansion of the atomic energies ε_i with respect to an increasingly higher number of bonds considered. Here, we will use a slightly different, albeit equivalent, approach to the ACE formalism than the one of the original work. We will do to introduce strategies and expressions that will be useful for the remainder of this thesis.

The starting point for the construction of the ACE formalism is, again, the atomic energy ε_i, which is expanded in cluster contributions as

$$\varepsilon_i = \varepsilon_i^{(1)} + \sum_{(j)_i}^{\text{atoms}} v^{(2)}(\mathbf{r}_{ji}) + \sum_{(jk)_i}^{\text{atoms}} v^{(3)}(\mathbf{r}_{ji}, \mathbf{r}_{ki}) + \sum_{(jkp)_i}^{\text{atoms}} v^{(4)}(\mathbf{r}_{ji}, \mathbf{r}_{ki}, \mathbf{r}_{pi}) + \dots, \tag{2.51}$$

where $v^{(n)}$ is a n-cluster, or n-Body (nB) potential, and $\varepsilon_i^{(1)}$ is a constant shift. The first sum runs over all the atoms, up to a cut-off, in the environment of the i-th atom, the second sum runs over all the unique pairs of atoms in the same neighborhood, the third sum over all the unique triplets, and so on. We now introduce an orthonormal basis, $\{\phi_{nlm}\}$, to expand the density, in the same spirit as the one already used in Eq. (2.8). With basis elements defined as

$$\phi_{nlm}(\mathbf{r}) = R_{nl}(r) Y_l^m(\hat{\mathbf{r}}), \tag{2.52}$$

we obtain the expression[7]

$$\begin{aligned} \varepsilon_i = \varepsilon_i^{(1)} + \sum_j \sum_\nu a_\nu^{(2)} \phi_\nu(\mathbf{r}_{ji}) + \sum_{jk} \sum_{\nu_1 \nu_2}^{\nu_1 \geq \nu_2} a_{\nu_1\nu_2}^{(3)} \phi_{\nu_1}(\mathbf{r}_{ji}) \phi_{\nu_2}(\mathbf{r}_{ki}) + \\ + \sum_{jkp} \sum_{\nu_1\nu_2\nu_3}^{\nu_1 \geq \nu_2 \geq \nu_3} a_{\nu_1\nu_2\nu_3}^{(4)} \phi_{\nu_1}(\mathbf{r}_{ji}) \phi_{\nu_2}(\mathbf{r}_{ki}) \phi_{\nu_3}(\mathbf{r}_{ji}) + \dots, \end{aligned} \tag{2.53}$$

where the subscript ν is a short-hand for the collection of indexes $\nu = (n, l, m)$. We also assume that the nB-expansion coefficients $a^{(n)}$ are symmetric under index permutations, and so we can consider only ordered summations[8]. Crucially, in going from Eq. (2.51) to Eq. (2.53), we can notice how the summations over all the atoms are now unrestricted. This can be achieved by means of the following trick, which is shown for the particular case of the 3B terms: since the basis $\{\phi_\nu\}$ is complete, then the cases in the sum above for which $k = j$ can be written as

$$\sum_{j,k=j} \sum_{\nu_1\nu_2}^{\nu_1 \geq \nu_2} a_{\nu_1\nu_2}^{(3)} \phi_{\nu_1}(\mathbf{r}_{ji}) \phi_{\nu_2}(\mathbf{r}_{ji}) = \sum_{j,j=k} \sum_\nu \left[\sum_{\nu_1\nu_2}^{\nu_1 \geq \nu_2} a_{\nu_1\nu_2}^{(3)} c_{\nu_1\nu_2}^\nu \right] \phi_\nu(\mathbf{r}_{ji}), \tag{2.54}$$

where the coefficients $c_{\nu_1\nu_2}^\nu$ are the expansion coefficients of the product of two basis function in terms of a single one. Explicitly, if the basis functions are normalized, we have the

[7] Please, note that, contrary to the original ACE paper, the superscripts here will indicate the body order and not the correlation order.

[8] Any ordering of choice, e.g., lexicographic ordering, can be applied here.

expressions

$$\phi_{\nu_1}(\mathbf{r})\phi_{\nu_2}(\mathbf{r}) = \sum_{\nu} c^{\nu}_{\nu_1\nu_2}\phi_{\nu}(\mathbf{r}), \quad \text{with} \quad c^{\nu}_{\nu_1\nu_2} = \int \mathrm{d}\mathbf{r}\phi^*_{\nu}(\mathbf{r})\phi_{\nu_1}(\mathbf{r})\phi_{\nu_2}(\mathbf{r}). \tag{2.55}$$

This shows that, in going from the restricted to the unrestricted sum, we effectively introduced an additional 2B terms: crucially, we can be then re-absorb them by a re-definition of the expansion coefficients of lower-body order terms. Analogous considerations can be carried out for any of the terms in the expansion above [Eq. (2.53)].

If we now take the density of Eq. (2.6), and its expansion in the basis $\{\phi^*_{\nu}\}$ (note that the density is real)

$$\rho_i(\mathbf{r}) = \sum_j \delta(\mathbf{r} - \mathbf{r}_{ji}) = \sum_{\nu} A_{i\nu}\phi^*_{\nu}(\mathbf{r}), \tag{2.56}$$

then the expansion coefficients, that constitute the so-called *atomic basis*, read

$$A_{i\nu} = \int \mathrm{d}r\rho_i(\mathbf{r})\phi_{\nu}(\mathbf{r}) = \sum_j \phi_{\nu}(\mathbf{r}_{ji}). \tag{2.57}$$

Crucially, this step introduces the density trick in the cluster expansion, by allowing to re-write Eq. (2.53) as

$$\varepsilon_i = \varepsilon^{(1)}_i + \sum_{\nu} a^{(2)}_{\nu} A_{i\nu} + \sum_{\nu_1\nu_2}^{\nu_1\geq\nu_2} a^{(3)}_{\nu_1\nu_2} A_{i\nu_1}A_{i\nu_2} + \sum_{\nu_1\nu_2\nu_3}^{\nu_1\geq\nu_2\geq\nu_3} a^{(4)}_{\nu_1\nu_2\nu_3} A_{i\nu_1}A_{i\nu_2}A_{i\nu_3} + \dots. \tag{2.58}$$

Since the computational cost required for the evaluation of the elements of the atomic basis scales linearly with the numbers of atoms in the neighborhood, then one of the most important feats of the ACE expansion is the implementation of this scaling within a hierarchically-improvable, cluster-expansion-based scheme.

Symmetries Let us analyse the symmetry of the expansion in Eq. (2.58). The symmetry with respect to atomic permutation is naturally enforced by the summation over all the atoms in the atomic basis, $A_{i\nu}$, and by considering only ordered indexes ν (so that the expansion is symmetric under the exchange of atomic basis). The locality of the potentials is enforced by choosing a radial basis, $R_{nl}(r)$, that goes smoothly to zero at the cut-off radius. This is usually done by defining a complete set of orthonormal functions, with an envelope given by the cut-off function, $f_c(r_{ji})$.

Instead, care must be taken when considering the symmetry under rotation. Indeed, we can enforce it by selecting, from the summations in Eq. (2.58), only the combinations of indexes which result in a rotationally invariant quantity. The terms of the atomic basis, A_{inlm}, have the same formal role of the expansion coefficients c_{inlm} defined in the GAP

framework. This means that we can employ the coupling scheme of the powerspectrum and the bispectrum, for the 2B and 3B terms, respectively. This leads to[9]

$$\begin{cases} B^{(2)}_{in} = A_{in00}, \\ B^{(3)}_{in_1n_2l} = \sum_{m=-l}^{l} (-1)^m A_{in_1lm} A_{in_2l-m} = \sum_{m=-l}^{l} A_{in_1lm} A^*_{in_2lm}, \\ B^{(4)}_{i\,\substack{n_1n_2n_3\\ l_1l_2l_3}} = \sum_{m_1m_2m_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} A_{in_1l_1m_1} A_{in_2l_2m_2} A_{in_3l_3m_3}, \\ B^{(5)}_{i\,\substack{n_1n_2n_3n_4\\ l_1l_2l_3l_4}} = \sum_{m_1m_2m_3m_4} \begin{bmatrix} l_1 & l_2 & l_3 & l_4 \\ m_1 & m_2 & m_3 & m_4 \end{bmatrix} A_{in_1l_1m_1} A_{in_2l_2m_2} A_{in_3l_3m_3} A_{in_4l_4m_4}, \\ \dots . \end{cases} \tag{2.59}$$

We note the formal equivalence with the powerspectrum, $p_{nn'lm}$, of Eq. (2.10), and the 3B-term, $B^{(3)}$. We also rapidly mention that the 4B term, $B^{(4)}$, is equivalent to the bispectrum coupling of Eq. (2.19). Indeed, we can use the well-known 3j-Wigner symbols, proportional to CG coefficients and defined by the identity

$$\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} := \frac{(-1)^{j_1-j_2+M}}{\sqrt{2l_3+1}} C^{l_3-m_3}_{l_1m_1l_2m_2}. \tag{2.60}$$

The 3j-symbols are symmetric under any even permutation of columns, while they acquire a phase factor of $(-1)^{l_1+l_2+l_3}$ (directly linked with the effect of the reflection of the axis mentioned in section 2.1.2), under any odd permutation of columns. Since, it holds that $Y^{m*}_l = (-1)^m Y^{-m}_l$, then, we can manifest the equivalence between $B^{(4)}$ and the bispectrum (up to a proportionality factor) by simply re-labelling m_3 in $-m_3$. One of the most important achievements of the ACE formalism is that it paves the way to a systematic definition of higher-body-order terms. In particular, the rotationally invariant coupling introduced for the five-body term, $B^{(5)}$, is performed by the coefficients

$$\begin{bmatrix} l_1 & l_2 & l_3 & l_4 \\ m_1 & m_2 & m_3 & m_4 \end{bmatrix} = \sum_{lm} (-1)^m C^{lm}_{l_1m_1l_2m_2} C^{l-m}_{l_3m_3l_4m_4}, \tag{2.61}$$

which, as already discussed, can be interpreted as a first projection of $(l_1m_1l_2m_2)$ on the space at (lm), a second projection of $(l_3m_3l_4m_4)$ on the same space, and a final contraction, in the same spirit of the powerspectrum coupling. While this way of proceeding is systematic, care is required for the degeneracy of the coupling scheme, since one could choose among different, albeit same-space spanning, couplings (see Ref. [66]).

[9] We remark here again that we use a slightly different super-script notation with respect to the original ACE paper.

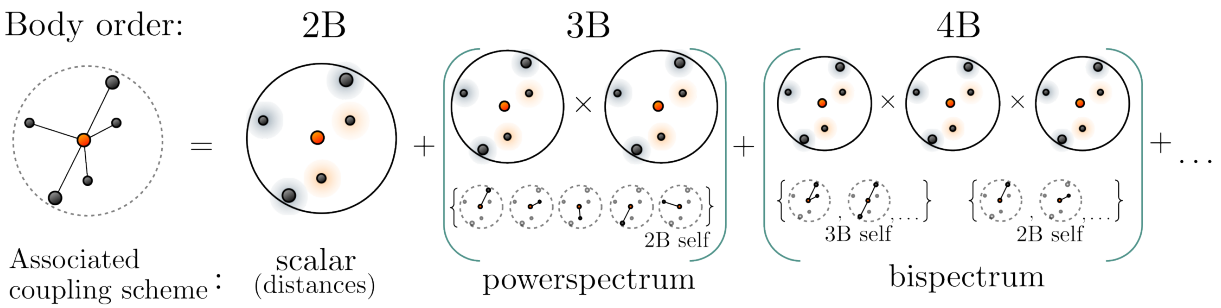

Figure 2.7: The main ingredients of the ACE framework are shown. The core of the formalism is the cluster expansion in a multi-body orders (the chemical potential is not explicitly shown). Then, the focus lies in recasting the expressions in term of coefficients of the atomic density. This is achieved by introducing self-energy terms which, depending only on lower body orders, can be always reabsorbed by means of coefficient re-definitions. Once the expression is written in terms of the density, everything is linear in the number of atoms in the neighborhood. Finally, only terms that produce scalar quantities are retained, by means of a coefficient selection based on the coupling of angular momenta: this allows to recover expressions which are analogous to the powerspectrum and the bispectrum.

Finally, by discarding all the terms for which the sum over the principal angular momentum number, $\sum_i l_i$, is odd, we can enforce also symmetry under reflection, and obtain the final expression for the atomic energy ε_i

$$\varepsilon_i = \sum_K \varepsilon_i^{(K)} = \sum_K \mathbf{c}^{(K)} \cdot \mathbf{B}_i^{(K)}. \tag{2.62}$$

Here, we grouped the terms by body order K, with the definition of the K-body atomic energies, $\varepsilon_i^{(K)} := \mathbf{c}^{(K)} \cdot \mathbf{B}_i^{(K)}$. Thus, the vector $\mathbf{c}^{(K)}$ contains all the K-body expansion coefficients. A graphical representation of the main point of this section are shown in Fig. 2.7. The ACE framework has achieved great success in combining the density trick with a systematically improvable way of reaching higher body order, which generalized all previous approaches. It achieves high performance on both organic molecules[67] and large-scale simulations[68]. Finally, an optimized and efficient implementation has been formulated and made available (see PACE, Ref. [69]). For all these reasons we will adopt the same cluster-expansion foundation in defining the Jacobi-Legendre descriptors in Chapter 4.

2.2 Descriptors for Tensors

In the previous section we have investigated a few of the most important MLP, designed to target scalar quantities, in particular the energy, of a given system. However, not all the quantities of physical interest behave are isometrically invariant. On the contrary, vectorial (velocities, forces, dipoles), and tensorial (stress tensor, polarizability) quantities

are ubiquitous. Indeed, to be able to efficiently predict tensor components is a crucial goal of any ML framework that aims to accelerating material discovery or investigation (for example in accelerated investigation of Raman [70, 71] and infrared [72] spectra). Therefore, lots of work has been focused on the construction of rotationally-*covariant* descriptors, namely descriptors that follow the same mixing of the components of a tensor, when a general rotation is applied.

In this section, we will first show how to cast the target in a form that clearly defines the effect of a rotation on the targets, so that a strategy to construct the descriptors could be easily outlined. We will then proceed in presenting a few of the most used covariant descriptors and models.

2.2.1 Spherical decomposition

Let us begin by showing how to decompose a tensor in its *spherical components*, to define the relations to enforce under rotations. A cartesian tensor of rank r (for example, a rank 2 stress tensor $\mathbf{W}$, in which we define the components in terms of the cartesian frame of reference, as W_{xx}, W_{xy}, ..., W_{zz}), can always be decomposed, by means of an appropriate change of basis, in terms of its spherical components. These are the components that, under a rotation of the frame of reference, follow the same transformation rules of the analogous spherical harmonics: indeed they are identified by three labels: the first indicating the angular quantum number l, ranging from 0 to the rank of the tensor. The second corresponding to the magnetic quantum number m, defined in the range $[-l, l]$. The third labelling which coupling scheme has been chosen: indeed, the spherical decomposition is usually not unique, and different combinations of cartesian components, despite being independent (in the sense that they do not mix under rotations), can share the same pairs (l, m). We will rapidly explore how this procedure is done, starting by the simplest case of a vector (tensors of rank 1) [please see Refs. [54, 73] for details].

Spherical decomposition of a vector We can define the spherical decomposition of a cartesian vector $\mathbf{V} = (V_x, V_y, V_z)$ by looking at the very definition of spherical harmonics of order 1, namely

$$Y_1^{\pm 1}(\hat{\mathbf{r}}) = \frac{1}{2}\sqrt{\frac{3}{2\pi}}(\mp r_x - i r_y), \qquad Y_1^0(\hat{\mathbf{r}}) = \sqrt{\frac{3}{4\pi}} r_z. \tag{2.63}$$

In complete analogy, we can define the spherical decomposition of the vector V as

$$V_{\pm 1} = \frac{1}{\sqrt{2}}(\mp V_x - i V_y) \qquad \text{and} \qquad V_0 = V_z, \tag{2.64}$$

that, by construction, will transform as the components of the spherical harmonic Y_1^q. This transformation can be written in matrix notation as

$$V_q = \sum_{i=x,y,z} u_{qi} V_i, \tag{2.65}$$

where $\mathbf{u} = (u_{qi})$ is the transformation matrix [defined by Eq. (2.64)]

$$\mathbf{u} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & -i & 0 \\ 1 & -i & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}. \tag{2.66}$$

Spherical decomposition of a tensor of rank 2 A cartesian tensor of rank 2 can be represented by a 3x3 square matrix, $\mathbf{M}$. We now use the trick of re-casting the matrix in terms of the outer product of two vectors $\mathbf{V}$ and $\mathbf{U}$, combined in a *dyad*, as

$$M_{ij} = V_i U_j \quad \text{where} \quad i, j = x, y, z. \tag{2.67}$$

If we now perform a spherical decomposition on each of the vectors of the dyad [by means of Eq. (2.64)] we are led to the expression

$$M_{q_1 q_2} = V_{q_1} U_{q_2} = \sum_{i,j=x,y,z} u_{q_1 i} u_{q_2 j} V_i V_j = \sum_{i,j=x,y,z} u_{q_1 i} u_{q_2 j} M_{ij}, \tag{2.68}$$

which shows how we can always write the components $M_{q_1 q_2}$ in terms of the cartesian ones, by means of the matrix $\mathbf{u}$. However, these components still do not follow the transformation rules of the spherical harmonics under a global rotation. Indeed, we recall that, as demonstrated in Eq. (2.22), by performing a coupling via the CG coefficients we can define quantities that are projected in the intended (l, m) space, i.e., that follow the same transformation rule of the spherical harmonic Y_l^m under rotation. Therefore, the spherical decomposition of a tensor of rank 2 is given by

$$M_l^m = \sum_{q_1=-1}^{1} \sum_{q_2=-1}^{1} C^{lm}_{1q_1 1q_2} M_{q_1 q_2}. \tag{2.69}$$

Given the selection rules of the CG, we can see that the only cases allowed are $l = 2, 1, 0$. As in the vectorial case, we do not have any degeneracy here, and so the spherical components of the matrix M are uniquely defined by the pairs (l, m).

Higher-rank tensors The procedure to decompose higher-rank tensors in spherical components follows the same recipe: introduce as many cartesian vectors as the tensor

rank, and cast them in terms of their spherical decomposition. However, when decomposing tensors of rank 3 or higher, we have to choose a coupling scheme. In the example of a rank 3 tensor, after constructing a *tryad* of vectors, we could decide to couple the first two by means of a CG coefficient, and then couple with the remaining vector. Or we could choose a completely different ordering. This implies that we have to keep track of the coupling scheme, since the same total (l, m) can be obtained by different routes. If we denote the rank 3 tensor by $\mathbf{T}$, the spherical components of order $l = 1$ (that transform as the spherical harmonic of order 1), can be obtained in the following three ways

$$\begin{cases} {}^{(0)}T_1^m = \sum_{q_3} C^{1m}_{001q_3} \left(\sum_{q_1 q_2} C^{00}_{1q_1 1q_2} T_{q_1 q_2 q_3} \right) = \dfrac{1}{\sqrt{3}} \sum_{q_1} (-1)^{1-q_1} T_{q_1 -q_1 m}, & \text{for } {}_{\mathbf{1+1}}\mathbf{0} + \mathbf{1}, \\ {}^{(1)}T_1^m = \sum_{q_{12} q_3} C^{1m}_{1q_{12}1q_3} \left(\sum_{q_1 q_2} C^{1q_{12}}_{1q_1 1q_2} T_{q_1 q_2 q_3} \right), & \text{for } {}_{\mathbf{1+1}}\mathbf{1} + \mathbf{1}, \\ {}^{(2)}T_1^m = \sum_{m_{12} q_3} C^{1m}_{2m_{12}1q_3} \left(\sum_{q_1 q_2} C^{2m_{12}}_{1q_1 1q_2} T_{q_1 q_2 q_3} \right), & \text{for } {}_{\mathbf{1+1}}\mathbf{2} + \mathbf{1}, \end{cases} \tag{2.70}$$

where we explicitly indicated the angular momentum coupling scheme as pre-indexes, so that, for example, ${}_{\mathbf{1+1}}\mathbf{0}$ indicates that we are considering the space $\mathbf{0}$ resulting from the coupling of $\mathbf{1} + \mathbf{1}$ angular-momentum spaces. If, on the one hand, these three cases do not mix (albeit they share the same (l, m)), the first case explicitly shows how a different choice of coupling scheme generally leads to different results. Indeed the last index of $T_{q_1 -q_1 m}$ is not on the same footing as the first two. We remark, however, that we can usually exploit the symmetries of the tensor $\mathbf{T}$ to reduce the degeneracy and obtain expressions that could even be independent of the choice of a particular coupling scheme [please, see Ref. [74]].

We have shown how to decompose a tensor in a manner that mimics the same transformation rules of the spherical harmonics. This is crucial for the construction of covariant descriptors, since we can now apply the same mathematical framework that led to the search for rotationally invariant quantities, again closely related with the transformation behaviour of the spherical harmonics.

2.2.2 SA-GPR framework: the λ-SOAP

We discuss here the symmetry-adapted Gaussian process regression framework (SA-GPR) introduced in Ref. [28]. The proposed kernel model is based on the same formalism of the SOAP model, and as such is called λ-SOAP. We will, indeed, follow the same recipe outlined in the SOAP section, by showing how to construct covariant descriptors, and by introducing the covariant λ-SOAP kernels.

A "covariant powerspectrum" The main underlying assumption for the definition of covariant descriptors is, again, the partitioning of a tensor in atomic contributions. So, if we have a tensor $\mathbf{T}$, we can write

$$\mathbf{T} = \sum_{i}^{\text{atoms}} \mathbf{T}_i, \tag{2.71}$$

which mirrors what has been already done for the energy. In a further analogy with MLPs, we will also assume that the atoms that contribute to $\mathbf{T}_i$ are the ones inside a cut-off sphere around the i-th atom. It is important, however, to remark that not all the atoms must be necessarily included in the sum above. If, for example, a few atoms are deemed to have a more relevant contribution to the total tensor than others, then we can neglect the non-relevant atoms and include only the significant ones in the sum above.

The last idea lies, again, in the introduction of the atomic density, $\rho(\mathbf{r})$, to represent our local environment, and its expansion in radial functions and spherical harmonics (as shown in Eq. (2.8)). Please note that the original λ-SOAP definition utilizes, again, atom-centered Gaussians in the definition of the density, and not Dirac-delta functions. This will not change the formal expression derived here, but it impacts the specific expression of the expansion coefficients, c_{inlm}. We have already shown the invariance of the powerspectrum and the bispectrum under rotation [see Eqs. (2.10) and (2.19)]. However, in proving the rotational invariance of the bispectrum components, we also proved that the quantity in Eq. (2.24), reported here for readability

$$p^{lm}_{in_1n_2l_1l_2} = \sum_{m_1m_2} C^{lm}_{l_1m_1l_2m_2} c_{in_1l_1m_1} c_{in_2l_2m_2}, \tag{2.72}$$

follows the same transformation rule, under rotation, of the spherical harmonic Y^m_l. We then already defined a covariant descriptors, which can be seen as a generalization of the standard powerspectrum[10]. Finally, it is clear that this new definition also encompasses the standard powerspectrum. Indeed, if $l = 0$, then $C^{00}_{l_1l_2m_1m_2} = \delta_{l_1l_2}\delta_{m_1m_2}$, and we obtain the powerspectrum [Eq. (2.10)] again.

Defining the λ-SOAP kernels Following the same derivation used for the SOAP kernel, one could be tempted to write the covariant kernel by a contraction of the covariant powerspectra of two atomic environments (here represented by their atomic density ρ and ρ'). Explicitly

$$k^l_{m_1m_2}(\rho, \rho') = \sum_{n_1n_2l_1l_2} p^{lm_1}_{n_1n_2l_1l_2}(\rho) p^{lm_2}_{n_1n_2l_1l_2}(\rho'), \tag{2.73}$$

[10]Please, note that the powerspectrum defined for the λ-SOAP is obtained by multiplying the definition in Eq. (2.72) by $(-1)^{l_1-l}/\sqrt{2l+1}$.

where we dropped the index i for readability. This is indeed the case, and the kernel above is a legitimate one when treating covariant targets, and a particular example of a λ-SOAP kernel. It is nevertheless interesting to outline how this kernel is constructed. We can consider again the overlap integral of Eq. (2.41) as our starting point, but now, instead of introducing the invariant kernel of Eq. (2.42), we can follow the same technique firstly introduced in Ref. [75] for forces (vectors), and then extended to any tensorial quantity in Ref. [28]. Essentially, this consists in defining the covariant kernel by means of the Wigner-D matrices as

$$k^l_{m_1m_2}(\rho,\rho') := \int \mathrm{d}\hat{R}\, D^l_{m_1m_2}(\mathcal{R})\Big|S(\rho,\hat{R}\rho')\Big|^n = \int \mathrm{d}\hat{R}\, D^l_{m_1m_2}(\mathcal{R})\left|\int \mathrm{d}\mathbf{r}\,\rho(\mathbf{r})\rho'(\hat{R}\mathbf{r})\right|^n. \tag{2.74}$$

Indeed, it can be easily proven that this kernel satisfies the property

$$\mathbf{k}^l(\hat{R}_1\rho,\hat{R}_2\rho') = \mathbf{D}^l(\mathcal{R}_1)k^l(\rho,\rho')\mathbf{D}^l(\mathcal{R}_2^{-1}), \tag{2.75}$$

which is the required transformation property for a covariant kernel. In the case of a density of the form of Eq. (2.6), or a density obtained by a sum of atom-centered Gaussians, the integrals can be computed analytically. For the case $n = 2$, we obtain exactly the kernel of Eq. (2.73), with the contraction of "covariant powerspectra". Please note that, as done for the invariant case, the kernels are again normalized by means of

$$K^l_{m_1m_2}(\rho,\rho') = \frac{k^l_{m_1m_2}(\rho,\rho')}{\sqrt{\|\mathbf{k}^l(\rho,\rho)\|_F\|\mathbf{k}^l(\rho',\rho')\|_F}}, \tag{2.76}$$

where $\|\cdot\|_F$ is the Frobenius norm.

The model With the definition of the covariant descriptors and the covariant kernel, we can now write the kernel model for the spherical components of a tensor $\mathbf{T}$ as

$$T^m_l = \frac{1}{N}\sum_i^N T^m_{i,l} = \frac{1}{N}\sum_{i=1}^N\left(\sum_J^{\text{data}}\sum_{m'=-l}^{l}\frac{\alpha^l_{Jm'}}{N_J}\sum_{j\in J}^{N_J}K^l_{mm'}(\rho_i,\rho^J_j)\right), \tag{2.77}$$

where N is the number of atoms in the system, $T^m_{i,l}$ is the contribution of the i-th environment to the tensor, the sum over J runs over all the systems in the training set, and where the sum over j runs over all the N_J atoms in the J-th system. Finally, $\alpha^l_{Jm'}$ are the coefficients to be evaluated/trained. The obtained model not only has the correct transformation properties, but can also benefit from the hierarchical nature of the kernel approach, so that higher-order kernels [in terms of the n exponent in Eq. (2.74)] can always be computed. However, the λ-SOAP kernels can get computationally heavy, in particular

when higher-order terms are evaluated. An application of the SA-GPR framework, where this problem was addressed, can be found in Ref. [76]. There, not only were kernels at different cut-off radii employed to increase the accuracy of the resulting model, but also, $l = 0$ kernels (standard SOAP) were multiplied by $l > 0$ ones, to increase the degree of non-linearity while not increasing the exponent n. Indeed, the transformation properties of a covariant object are unaffected if the object itself is decorated with a rotationally-invariant quantity. This same strategy is behind the following model, based on the SNAP formalism.

2.2.3 Extension of SNAP to tensorial quantities

Connecting a scalar descriptor with a covariant quantity is the approach proposed in the recent work from Ref. [77]. The descriptors defined for this model are bispectrum components decorated with spherical harmonics to ensure the covariant behaviour of the total object. This approach is an evolution of the one based on the 4d bispectrum, proposed in Ref. [78]. There, however, the systems were aligned with an appropriate reference configuration, to evaluate all the bispectrum components on the same footing. On the contrary, the new approach allows to have a build-in covariance directly in the descriptors. Indeed, the proposed expression for the tensor components is

$$T_l^m = \sum_i^N T_{i,l}^m = \sum_i^N \sum_{ll_1l_2} a_{ill_1l_2} B_{ill_1l_2} \overline{Y}_l^m(i), \tag{2.78}$$

where the first sum is over the N atoms in the system, and $B_{ill_1l_2}$ are the 4d-bispectrum components, used in the SNAP model, and given in Eq. (2.35)[11]. Here the term $\overline{Y}_l^m(i)$ is defined as

$$\overline{Y}_l^m(i) := \sum_j^{\text{atoms}} f_c(r_{ji}) Y_l^m(\hat{\mathbf{r}}_{ji}), \tag{2.79}$$

which is manifestly covariant, being a simple sum of spherical harmonics. The function f_c is a cut-off which, in the simplest case, is a discontinuous cut at the cut-off radius. The model is fully linear, and as such simple regression methods can be used to train the coefficients $a_{ill_1l_2}$ and, in this sense, this approach can be seen as a covariant extension of SNAP. Despite its simplicity and compactness the model is capable of reaching accuracies that can approximate the ones of the more complex λ-SOAP.

[11] Please note that, in the fitting procedure, the bispectrum components can be divided by the number of atoms to have normalized descriptors.

2.2.4 A general formula for covariant descriptors

There are a few other approaches that lean toward the prediction of tensorial quantities. We will not treat them explicitly here but, among the others, we mention the model devoted to the prediction of orbital-related block of an Hamiltonian that can be found in Ref. [79], or the N-body iterative contraction of equivariants (NICE) of Ref. [38]. These methods are closely related in spirit to the λ-SOAP kernel construction of Eq. (2.74). We will rapidly outline the core of these approaches, which lies in the following general construction for covariant quantities. Let us consider an atomic function F_i, centered on the i-th atom, and depending on the relative positions of all atoms in its surroundings $F_i = F_i(\{\mathbf{r}_{ji}\})$ [all the descriptors presented so far are sums of terms of this form]. The function defined as

$$G_l^m(\hat{\mathbf{s}};\{\mathbf{r}_{ji}\}) = \sum_{m'} Y_l^{m'}(\hat{\mathbf{s}}) \int \mathrm{d}\hat{R}\, D^l_{m'm}(\mathcal{R}) F_i(\{\hat{R}\mathbf{r}_{ji}\}), \tag{2.80}$$

transforms as the spherical harmonics of Y_m^l under a rotation of the atomic coordinates (here, $\hat{\mathbf{s}}$, is an unimportant auxiliary variable). This means that

$$G_l^m(\hat{\mathbf{s}};\{\hat{R}_1\mathbf{r}_{ji}\}) = \sum_{m'} D^{l*}_{mm'}(\mathcal{R}_1) G_l^{m'}(\hat{\mathbf{s}};\{\mathbf{r}_{ji}\}), \tag{2.81}$$

as proven by the following chain of identities

$$\begin{aligned}
G_l^m(\hat{\mathbf{s}};\{\hat{R}_1\mathbf{r}_{ji}\}) &= \sum_{m'} Y_l^{m'}(\hat{\mathbf{s}}) \int \mathrm{d}\hat{R}\, D^l_{m'm}(\mathcal{R}) F_i(\{\hat{R}_1\hat{R}\mathbf{r}_{ji}\}) \\
&= \sum_{m'} Y_l^{m'}(\hat{\mathbf{s}}) \int \underbrace{\mathrm{d}\hat{R}}_{=\mathrm{d}\left(\hat{R}\hat{R}_1\right)} \sum_{m_1} \underbrace{\left[\sum_{m_2} D^l_{m'm_2}(\mathcal{R}) D^l_{m_2m_1}(\mathcal{R}_1)\right]}_{=D^l_{m'm_1}(\mathcal{R}\mathcal{R}_1)} \underbrace{D^l_{m_1m}(\mathcal{R}_1^{-1})}_{=D^{l*}_{mm_1}(\mathcal{R}_1)} F_i(\{\hat{R}\mathbf{r}_{ji}\}) \\
&= \sum_{m_2} D^{l*}_{mm_1}(\mathcal{R}_1) \sum_{m'} Y_l^{m'}(\hat{\mathbf{s}}) \int \mathrm{d}\hat{R}\, D^l_{m'm}(\mathcal{R}) F_i(\{\hat{R}\mathbf{r}_{ji}\}) = \sum_{m_1} D^{l*}_{mm_1}(\mathcal{R}_1) G_l^{m_1}(\hat{\mathbf{s}};\{\mathbf{r}_{ji}\}).
\end{aligned} \tag{2.82}$$

Here, in going from the first to the second line, we used the inverse of the Wigner-D matrices

$$\sum_{m_1} D^l_{m_2m_1}(\mathcal{R}_1) D^l_{m_1m}(\mathcal{R}_1^{-1}) = \delta_{m_2m}.$$

In a formalism based on the Dirac-braket notation, similar to the one introduced in Refs. [44, 80], we defined the function

$$G_l^m(\hat{\mathbf{s}};\{\mathbf{r}_{ji}\}) = \langle\hat{\mathbf{s}}| \left[\int \mathrm{d}\hat{R}\hat{R}\, |lm\rangle\, F_i(\{\hat{R}\mathbf{r}_{ji}\})\right], \tag{2.83}$$

and we proved that the terms inside the square brackets transforms, under a rotation of the atomic coordinates, as the state $|lm\rangle$ [please, compare with Eq. (38) of Ref. [46]]. Finally we can understand the role of the Wigner-D matrices by comparing Eq. (2.74) and Eq. (2.80). Indeed, in both cases the rotation matrices act as projectors into the desired space. In particular, we can loosely read the function G_l^m as a contraction of the more general covariant kernel of Eq. (2.74) with a spherical harmonic.

2.3 Descriptors for Electron Density

Having explored descriptors for scalar and tensorial quantities, we close this chapter by outlined the construction of descriptors for scalar *fields*, defined on 3d space ($\mathbb{R}^3$), such as the DFT electron density.

Analogously to what was done for all the other descriptors above, we will consider only *atomic*-scalar fields, namely functions $n(\mathbf{r})$, that parametrically depend on the positions $\{\mathbf{r}_i\}$ of the atoms in a system, $n(\mathbf{r};\{\mathbf{r}_i\})$. As for the scalar and tensorial quantities, also the scalar fields have a well defined behaviour under symmetry operations. Indeed, let us consider a transformation $\hat{T}$, either a translation or a rotation. If we indicate by $n_{\hat{T}}$ the function obtained by a rotation of the system of atoms, then it must hold that

$$n_{\hat{T}}(\mathbf{r};\{\mathbf{r}_i\}) = n(\mathbf{r};\{\hat{T}\mathbf{r}_i\}) = n(\hat{T}^{-1}\mathbf{r};\{\mathbf{r}_i\}), \tag{2.84}$$

which is the property that we must enforce on the representation of a scalar field. A density function as defined in Eq. (2.6), satisfies this property, as can be explicitly seen by

$$\rho_{i,\hat{T}}(\mathbf{r}) = \sum_j^{\text{atoms}} w_{Z_j}\delta(\mathbf{r}-\hat{T}\mathbf{r}_{ji})f_c(r_{ji}) = \sum_j^{\text{atoms}} w_{Z_j}\delta(\hat{T}^{-1}\mathbf{r}-\mathbf{r}_{ji})f_c(r_{ji}) = \rho_i(\hat{T}^{-1}\mathbf{r}), \tag{2.85}$$

where we exploit the fact that, if $\hat{T}$ is either a translation or a rotation, then $\delta(\mathbf{r}-\hat{T}\mathbf{r}_{ji}) = \delta(\hat{T}^{-1}\mathbf{r}-\mathbf{r}_{ji})$. Indeed the use of an atomic density such as ρ_i will be the starting point for the descriptors introduced in the next section. For the remainder of the chapter, both $\rho(\mathbf{r})$ and $n(\mathbf{r})$ will refer to the electronic density.

2.3.1 SALTED

We present here the strategy behind the description of the electronic density adopted in the symmetry-adapted learning of three-dimensional electron density (SALTED) [29], which is based on the framework introduced in Refs. [81] and [82]. The SALTED descriptors, in analogy to all the other descriptors discussed so far, are constructed on a partition of the electronic density in atomic contributions. The atomic contributions are then expanded again in a radial and spherical harmonics basis, [in the same spirit of Eqs. (2.6)(2.8)(2.9)] as

$$\rho(\mathbf{r}) = \sum_{i}^{\text{atoms}} \rho_i(\mathbf{r}) = \sum_{i}^{\text{atoms}} \sum_{nlm} c_{inlm} R_{nl}(r) Y_l^m(\hat{\mathbf{r}}). \tag{2.86}$$

We already proved, in Eq.(2.13), how the coefficients c_{inlm} transform, under a rotation of the frame of reference, as the spherical harmonics Y_l^m [please, see Eq. (2.13)]. Thus, if $c_{inlm}^{\hat{R}}$ are the coefficients of the rotated density, $\rho_{\hat{R}}(\mathbf{r})$, then it holds that

$$c_{inlm}^{\hat{R}} = \sum_{m'} D_{mm'}^{l*}(\mathcal{R}) c_{inlm'}. \tag{2.87}$$

This means that the expansion coefficients are, indeed, covariant quantities, and, as such, can be addressed by a SA-GPR. Thus, we can predict the expansion coefficients by means of the covariant kernel defined in Eq. (2.74), as

$$c_{inlm} = \sum_{j}^{\text{data}} \sum_{m'=-l}^{l} \alpha_{nlm'}^{j} k_{mm'}^{l}(\rho_i, \rho_j) \delta_{Z_i Z_j}. \tag{2.88}$$

where the sum on j runs over the reference data, and where only same species' kernels are allowed. Obtaining the coefficients $\alpha_{nlm'}^{j}$ is the aim of the fitting procedure. While this method inherits all the versatility of the λ-SOAP approach, such as the possibility to reach higher-order kernels by means of larger n in Eq. (2.74), it also suffers from the high computational overhead of the descriptor calculations.

2.3.2 Adapted Symmetry Functions and SNAP

The next descriptors can be interpreted as an extension of the Behler-Parrinello symmetry functions, and have been introduced in the work from Ref. [83], where they were implemented within a NN architecture. They are based on the idea of covering the real 3d space with a uniform mesh of grid points, $\mathbf{r}_g$, and then evaluate the symmetry functions *centered on each* of the $\mathbf{r}_g$. The functions utilized are

$$S_k = C_k \sum_i^{\text{atoms}} \exp\left(\frac{-r_{gi}^2}{2\sigma_k^2}\right) f_c(r_{gi}), \tag{2.89}$$

$$V_k^\alpha = C_k \sum_i^{\text{atoms}} \frac{r_{gi}^\alpha}{2\sigma_k^2} \exp\left(\frac{-r_{gi}^2}{2\sigma_k^2}\right) f_c(r_{gi}), \tag{2.90}$$

$$T_k^{\alpha\beta} = C_k \sum_i^{\text{atoms}} \frac{r_{gi}^\alpha r_{gi}^\beta}{4\sigma_k^4} \exp\left(\frac{-r_{gi}^2}{2\sigma_k^2}\right) f_c(r_{gi}), \tag{2.91}$$

where $C_k = ((2\pi)^{3/2}\sigma_k)^{-1}$ is the normalization for the k-th Gaussian. Comparing the scalar fields, S_k, with Eq. (2.3), we notice the similarity with the two-body functions G_i^1, where the grid point $\mathbf{r}_g$ is formally treated on the same footing of an atom. The vectorial functions, V_k, and the tensorial ones, $T_k^{\alpha\beta}$, are obtained by differentiating the scalar fields. Indeed the operation of differentiation is arguably the simplest way to obtain covariant quantities from scalar ones (for example, the forces are obtained by differentiation of the energy terms). These features, designed specifically for a NN, are among the first examples of descriptors adapted to be evaluated on grid points. This allows for an introduction of the real-3D space, $\mathbb{R}^3$, on the same footing of the atoms. However, because the mesh is usually uniform in space, and because the descriptors must be evaluated for each grid point, these descriptors carry redundancies and a large computational cost. Also, since the functions above are derived from two-body symmetry functions, they are only one-body in the atoms [the other one being the grid point], and so the description is severely degenerate, in the sense that very different atomic configurations can result in the same fingerprints. This last problem was partially addressed by the more recent work work from Ref. [84], where the fingerprints role was taken by SNAP components, $B_{gll_1l_2}$, from Eq. (2.35), and again evaluated by sitting on a grid point, as graphically shown in Fig. 2.8. While this approach elevated the descriptors to be three-body in nature [three atoms and a grid point], it was still based on a NN architecture and in an uniform, evenly spaced, grid. Therefore most of the problematic aspects of the first method can be found also for the SNAP-based one.

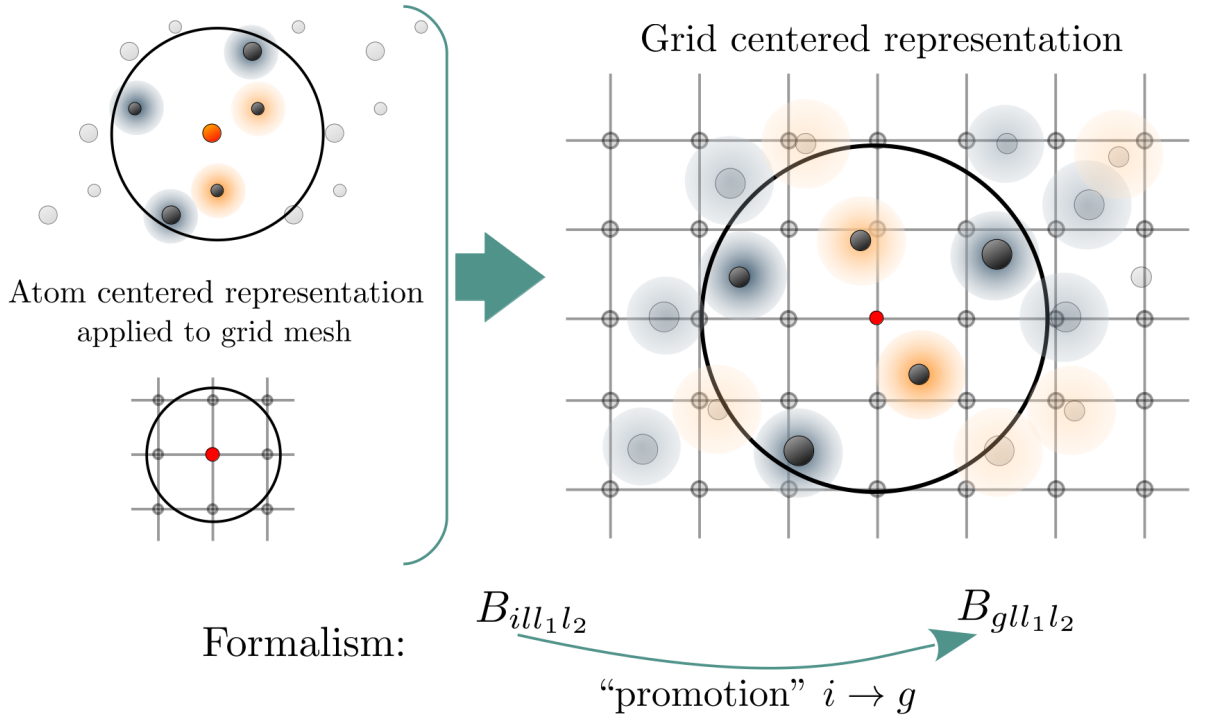

Figure 2.8: The main recipe to construct descriptors for a scalar field defined on the 3d-space is shown. The idea is to combine an atomic centered representation with a mesh of grid points covering the whole space. This carries fundamental assumptions, such as the idea of atomic decomposition of the density and of locality (enforced by the cut-off radius), into the new description. As such, the formalism is modified by simply applying a "promotion", from the central atom, i, to the grid point g (located at $\mathbf{r}_g$).

2.3.3 Conclusions

In this chapter, we have introduced the key ideas (atomic decomposition, locality, the density trick, the kernel methods and the grid trick) and mathematical tools (coupling of angular momenta, basis expansion, rotations encoded by means of the Wigner-D matrices) that constitute the foundation for the rest of this thesis. Starting from the next chapter, we will heavily employ the methods showed here, with the aim of defining descriptors that describe vector fields. From Chapter 4 to the end of the thesis, we will focus on the construction of a cluster-expansion-based machine-learning potential framework, constructed over internal coordinates, that will be able to encompass all the cases (from scalar to vector fields) described in this section.

Chapter 3

Powerspectrum for vector fields

This chapter is devoted to a review of the method presented in the published work 'A spectral-neighbour representation for vector fields: machine-learning potentials including spin" (in Ref. [85]), of which I am a co-author. My role in the article focused on the development of the mathematical framework and in part of its implementation. To maintain internal coherence with the rest of the thesis, we will follow a slightly different formalism, albeit equivalent, than the one used in the published manuscript.

3.1 Introduction

In the previous chapter, we have introduced several MLPs for accurate description of potential energy surfaces (PESs). However, all the descriptors shown are defined in terms of atomic positions only, and us such are not able to account for cases in which there is a non-position-related degeneracy in the state of the system, e.g., in the case ferromagnetic and antiferromagnetic states, as shown in Fig. 3.1. To investigate such PESs, one could always try to equip the MLP model with ad-hoc terms describing the magnetic interactions: this approach has been followed in Ref. [86], where a SNAP model was trained alongside a classical Heisenberg Hamiltonian. Or as done in Ref. [87], where a NN was constructed over a set of local spin interactions, in in terms of scalar products between the spins, as fingerprints. In this chapter, however, our aim will be on defining general descriptors that rely on as few assumption as possible regarding the functional form of the interaction. Efforts in this direction have been made in Ref. [88], where the Behler-Parrinello symmetry functions were generalized in order to carry the vectorial information of spin-collinear PESs (with fixed length of the spins). Similarly, in Ref. [89], NN-trained corrections were added to an Hamiltonian containing an Heisenberg and a Landau term (the corrections were formulated again in terms of modified symmetry functions). Another approach has been pursued in Ref. [90], where the magnetic vectors were treated on the

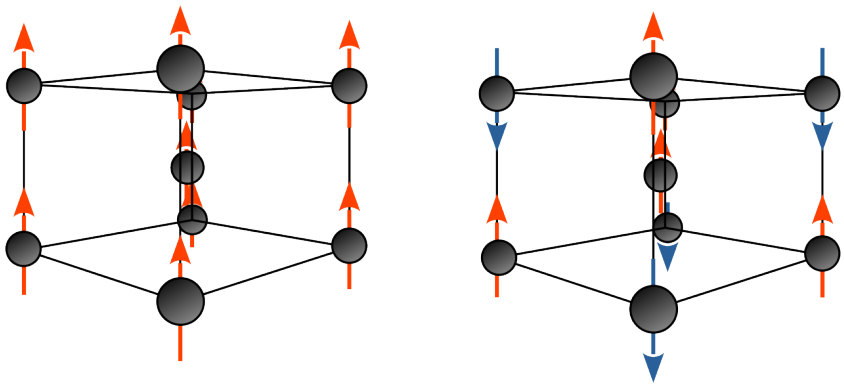

Figure 3.1: Two systems with the same atomic configurations but with different magnetic phases. Descriptors based on the atomic positions only are not able to distinguish between the two scenarios.

same footing of the atomic position vectors by constructing higher order tensors: this was done by means of appropriate external products in the same spirit of the Moment Tensor Potential framework [23]. It is also important to mention the generalization of the ACE model, aimed to encode general tensorial quantities [91], which was used very recently to train an ACE model on non-collinear iron systems [92]. Because the ACE's framework encompasses all possible coupling of angular momenta, it is always possible to map the method presented here in a low-body order expansion of a corresponding ACE potential. However, in this work, we will keep the vectorial nature of the quantities involved (local-semi-classical spins), and, inspired by the compactness of the SNAP potential, we will proceed only on defining a powerspectrum formalism. This is different to what done in the aforementioned ACE, where the vectors were described by delta functions. Our choice, instead, will result in a compact potential which, despite the small number of components involved, shows already a good descriptive power. Moreover, being grounded on the mathematical formalism of the powerspectrum, our method allows enough flexibility to be generalized to higher-order correlations by means of an appropriate kernel approach: this has been recently done by M.-T. Suzuki and co-workers [see Ref. [93]], where the construction introduced here is generalized to the GAP framework and to higher-body order expansions. As a final remark, the study of the powerspectrum in this new context allows to meticulously study properties that can be applied also to cases of atomic-position-based potentials.

This chapter will be structured as follows: we will first introduce the methods and show how to construct rotationally invariant quantities starting from a density representation of a vector field. This will employ the use of the spherical decomposition of the field and of the bipolar-spherical harmonics. We will then define the powerspectrum and will study, in detail, its most relevant properties. This will be followed by the construction

of a toy-model, simulating a cluster of iron atoms, to investigate the descriptive power of the powerspectrum in fully controlled scenarios. We will prove that a linear model that employs only the powerspectrum as descriptors is able to accurately describe a PES characterized by an Heisenberg interaction, with elements of spin-lattice coupling, when the training is performed on distorted ferromagnetic systems. We will then add Landau contributions to test the limits of the model when in presence of longitudinal interactions, and will compare the results with a GAP-based extension, capable of a natural introduction of non-linearity in the description.

3.2 Methods

The starting point of our discussion is the general atomic density (see Eq. (2.6))

$$\rho_i(\mathbf{r}) = \sum_{r_{ji} \leq r_{\text{cut}}} w_{Z_j} \delta(\mathbf{r} - \mathbf{r}_{ji}), \tag{3.1}$$

where the sum runs over the atoms in the cut-off sphere. The choice of the Dirac-delta functions to describe the local environment is not unique: other approaches are possible, such as the one used in Ref. [93], where Gaussians functions were employed (in the spirit of the GAP framework). We now assume that we can associate a vector, $\mathbf{v}_j$, to each atom in the system, i.e., we will define a vector field from the atomic density $\rho(\mathbf{r})$. To fix the ideas, we will explicitly consider the case of local-semi-classical spins, but this approach is independent on the nature of the field introduced. Thus, the newly defined vectorial field, $\boldsymbol{\rho}_i(\mathbf{r})$, is written as

$$\boldsymbol{\rho}_i(\mathbf{r}) = \sum_{r_{ji} \leq r_{\text{cut}}}^{\text{atoms}} w_{Z_j} \delta(\mathbf{r} - \mathbf{r}_{ji}) \mathbf{v}_j. \tag{3.2}$$

We now aim to choose a suitable basis for the expansion of the density. Normally, a vector is expanded in cartesian components as

$$\mathbf{v}_j = v_{j,x}\hat{\mathbf{e}}_x + v_{j,y}\hat{\mathbf{e}}_y + v_{j,z}\hat{\mathbf{e}}_z, \tag{3.3}$$

however, since our goal is to construct rotationally invariant quantities in terms of an harmonic representation, it is more appropriate to use the spherical versors (please, see Ref [54, 73], and section 2.2.1)

$$\hat{\mathbf{e}}_{\pm 1} = \mp \frac{1}{\sqrt{2}} (\hat{\mathbf{e}}_x \pm i\hat{\mathbf{e}}_y), \quad \text{and} \quad \hat{\mathbf{e}}_0 = \hat{\mathbf{e}}_z, \tag{3.4}$$

which transforms, under rotation, as the spherical harmonics Y_1^q, with $q = 0, \pm 1$. This

relation is the one already introduced in Eq. (2.64), where the decomposition of a vector in its spherical components was shown. Therefore, by using the matrix $\mathbf{u}$ of Eqs. (2.65) and (2.66), we can write[12]

$$\begin{aligned}\hat{\mathbf{v}}_j &= \sum_{p=x,y,z} v_{j,p}\hat{\mathbf{e}}_p = \sum_{pp'} v_{j,p} \underbrace{\sum_{q=0,\pm1} (u^{-1})_{pq}u_{qp'}}_{=\delta_{pp'}} \hat{\mathbf{e}}_{p'} \\ &= \sum_{q=0,\pm1} \underbrace{\left(\sum_p v_{j,p}(u^{-1})_{pq}\right)}_{:=v_{j,q}} \underbrace{\left(\sum_{p'} u_{qp'}\hat{\mathbf{e}}_{p'}\right)}_{=\hat{\mathbf{e}}_q} = \sum_{q=0,\pm1} v_{j,q}\hat{\mathbf{e}}_q,\end{aligned} \tag{3.5}$$

which shows how to evaluate the harmonic components of the vector $\hat{\mathbf{v}}_j$ by means of the inverse of the matrix $\mathbf{u}$. Explicitly, they read

$$v_{j,\pm1} = \frac{1}{\sqrt{2}}(\mp v_{j,x} + i v_{j,y}), \quad \text{and} \quad v_{j,0} = v_{j,z}. \tag{3.6}$$

Given the formal equivalence between the versors $\hat{\mathbf{e}}_q$ and the spherical harmonics Y_1^q (they also form orthonormal sets), we can go a step further and directly replace the basis versors with spherical harmonics as

$$\mathbf{v}_j = \sum_{q=0,\pm1} v_{j,q}\hat{\mathbf{e}}_q \xrightarrow{\hat{\mathbf{e}}_q \to Y_1^q(\hat{\mathbf{s}})} \mathbf{v}_{j,q}(\hat{\mathbf{s}}) = \sum_{q=0,\pm1} v_{j,q}Y_1^q(\hat{\mathbf{s}}), \tag{3.7}$$

where $\hat{\mathbf{s}}$ is an unimportant auxiliary variable that mirrors the orientation of the frame of reference (as such, it will undergo the same rotations applied to $\hat{\mathbf{r}}$). With this identification, the density reads

$$\boldsymbol{\rho}_i(\mathbf{r},\hat{\mathbf{s}}) = \sum_{r_{ji}\le r_{\rm cut}}^{\rm atoms} \sum_q w_{Z_j}\delta(\mathbf{r}-\mathbf{r}_{ji}) v_{j,q}Y_1^q(\hat{\mathbf{s}}). \tag{3.8}$$

We can now expand the density $\boldsymbol{\rho}_i$ in terms of a complete set of radial functions and spherical harmonics, in analogy to what was done in Eq. (2.8). Please note, however, that we have two spherical harmonics, in relation to the direction of the position vector, $\hat{\mathbf{r}}$, and the auxiliary variable, $\hat{\mathbf{s}}$, respectively. Explicitly

$$\boldsymbol{\rho}_i(\mathbf{r},\hat{\mathbf{s}}) = \sum_{n=0}^{n_{\rm max}} \sum_{l=0}^{n} \sum_{m=-l}^{l} \sum_{q=0,\pm1} c_{inlmq} R_{nl}(r) Y_l^m(\hat{\mathbf{r}}) Y_1^q(\hat{\mathbf{s}}), \tag{3.9}$$

[12]Please note that the matrix $\mathbf{u}$ is trivially invertible.

with the expansion coefficients given by[13]

$$c_{inlmq} = \int \mathrm{d}\mathbf{r}\mathrm{d}\hat{\mathbf{s}}\, \boldsymbol{\rho}_i(\mathbf{r},\hat{\mathbf{s}}) R_{nl}(r) Y_l^{m*}(\hat{\mathbf{r}}) Y_1^{q*}(\hat{\mathbf{s}}) = \sum_j^{\text{atoms}} w_{Z_j} v_{j,q} R_{nl}(r_{ji}) Y_l^{m*}(\hat{\mathbf{r}}_{ji}). \tag{3.10}$$

The radial functions chosen here are the same introduced for the Spherical-Bessel descriptors (firstly introduced in Ref. [94], and then expanded in Ref. [95]): we will then identify $R_{nl}(r)$ with the normalized functions $g_{n-l,l}(r)$ defined in Ref.[95]. We chose these functions for their regularity and for their smooth-vanishing behaviour at the cut-off radius. We can now use the bipolar spherical harmonics, $\mathcal{Y}^{JM}_{l_1 l_2}$, defined in Eq. (2.22), to cast the density $\boldsymbol{\rho}_i$ in the *coupled*-angular-momentum space. The procedure employs the completeness of the CG coefficients

$$\sum_{J=|l_1-l_2|}^{l_1+l_2} \sum_{M=-J}^{J} C^{JM}_{l_1 m_1 l_2 m_2} C^{JM}_{l_1 m_1' l_2 m_2'} = \delta_{m_1 m_1'} \delta_{m_2 m_2'}, \tag{3.11}$$

that, when plugged into Eq. (3.9), allow to write

$$\boldsymbol{\rho}_i(\mathbf{r},\hat{\mathbf{s}}) = \sum_{nl} \sum_{J=|l_1-1|}^{l_1+1} \sum_{M=-J}^{J} u_{inlJM} R_{nl}(r) \mathcal{Y}^{JM}_{l1}(\hat{\mathbf{r}},\hat{\mathbf{s}}), \tag{3.12}$$

with

$$u_{inlJM} = \sum_{mq} C^{JM}_{lm1q} c_{inlmq}, \quad \text{and} \quad \mathcal{Y}^{JM}_{l1}(\hat{\mathbf{r}},\hat{\mathbf{s}}) = \sum_{mq} C^{JM}_{lm1q} Y_l^m(\hat{\mathbf{r}}) Y_1^q(\hat{\mathbf{s}}). \tag{3.13}$$

As already proved in Eq. (2.22), the bipolar-spherical harmonics $\mathcal{Y}^{M}_{Jl_1 l_2}(\hat{\mathbf{r}},\hat{\mathbf{s}})$ transform, under a simultaneous rotation of the versors $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$, as the spherical harmonic Y_J^M does. Therefore, under such rotation, we have[14]

$$\boldsymbol{\rho}_i(\hat{\mathbf{r}},\hat{\mathbf{s}}) \xrightarrow{R^{-1}} \sum_{nlJM} \left[\sum_{M'} D^{J*}_{M'M}(\mathcal{R}^{-1}) u_{inlJM'} \right] R_{nl}(r) \mathcal{Y}^{JM}_{l1}(\hat{\mathbf{r}},\hat{\mathbf{s}}). \tag{3.14}$$

Thus, by exploiting the relation $D^{J*}_{M'M}(\mathcal{R}^{-1}) = D^{J}_{MM'}(\mathcal{R})$, the rotated expansion coefficients are given by

$$\hat{R} : u_{inlJM} \to \sum_{M'} D^{J}_{MM'}(\mathcal{R}) u_{inlJM'}, \tag{3.15}$$

[13] We are implying that the radial basis is orthonormal.

[14] Please note that, by rotating the versor $\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$, we are performing a passive rotation, namely a rotation of the frame of reference, here indicated with $\mathcal{R}^{-1}$. This notation, inverse of what adopted in the original paper [85], has been used here to maintain formal coherence with the rest of the thesis.

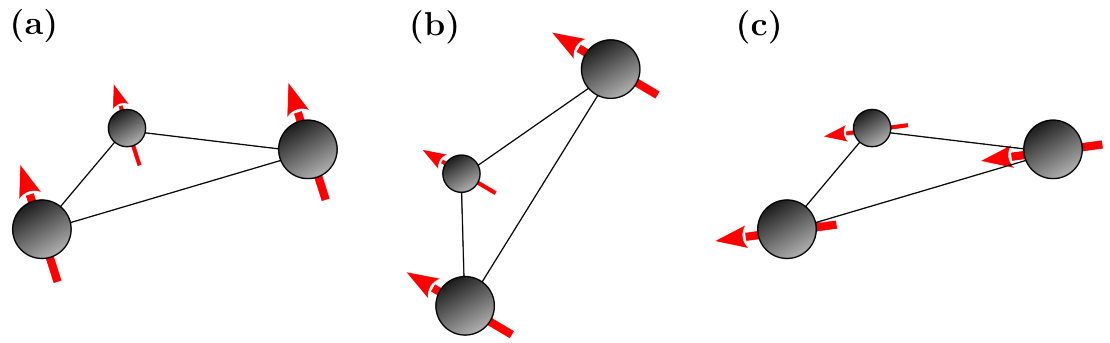

Figure 3.2: (a)-(b) Configurations that differ only by a simultaneous rotation of the atomic position and of the vector field: they will lead to the same powerspectrum, as defined in Eq. (3.16). (c) Configuration obtained from (a) by applying a rotation on the vector field. This will lead, in general, to a different powerspectrum from the one of (a)-(b).

which, formally, is the same transformation rule for the standard coefficients, c_{inlm} [please, see Eq. (2.13)]. Then, by the same argument used for the coefficients c_{inlm}, we can define a rotationally invariant powerspectrum for the expansion coefficients in the *coupled* scheme, u_{nlJM}, as

$$p_{inn'll'J} := \sum_M u_{inlJM} u^*_{in'l'JM}. \tag{3.16}$$

We remark that the most important feature of this powerspectrum is the invariance under *simultaneous* rotations of the atomic positions and the vector field, as graphically shown in Fig. 3.2: this is guaranteed by the coupling of the two angular-momentum channels, performed by casting the expansion in terms of the total angular momentum.

In the original manuscript of Ref. [85], we considered a simplified version of Eq. (3.16), one in which $n = n'$ and $l = l'$: this was done not only to reduce the number of components required, but also to deal with a real expression of the powerspectrum. It is worth investigating, however, when the powerspectrum is real[15]. If we take the complex conjugate of the expansion coefficients u_{nlJM}, and we use the symmetries $Y_l^{m*} = (-1)^m Y_l^{-m}$ and $v^*_{j,q} = (-1)^q v_{j,-q}$ (which can be easily verified from Eq. (3.6)), we get

$$\begin{aligned} u^*_{inlJM} &= \sum_{mq} c^*_{inlmq} C^{JM}_{lm1q} = \sum_{mq} (-1)^{m+q} c_{inl-m-q} C^{JM}_{lm1q} = \\ &= (-1)^M \sum_{mq} c_{inlmq} C^{JM}_{l-m1-q} = (-1)^{l+1-J+M} u_{inlJ-M}, \end{aligned} \tag{3.17}$$

obtained by the selection rule $M = m + q$ (imposed by the CG coefficients), and by re-labelling m and q in their opposite. Also, the last step exploits the symmetry rule $C^{lm}_{l_1 m_1 l_2 m_2} = (-1)^{l_1+l_2-l} C^{l-m}_{l_1-m_1 l_2-m_2}$ of the CG coefficients. We can now evaluate the com-

[15] What follows here, is an original analysis done for this thesis, non present in the published manuscript.

plex conjugate of the powerspectrum as

$$p^*_{inn'll'J} = (-1)^{l+l'} p_{inn'll'J}, \tag{3.18}$$

and deduce that the powerspectrum is real for even $l+l'$, and purely imaginary otherwise. One can then consider either different radial channels, n, or angular channels, l, being mindful to multiply by the imaginary unit all the cases for which $l+l'$ is odd. The role of this phase factor will be discussed shortly.

Since the powerspectrum $p_{inn'll'}$ is obtained by the contraction of total angular momentum channels, it is worth investigating if the two components of the density, the atomic positions and the vectors, have some independent symmetry on their own. This can be done by explicitly studying its behaviour under parity and time-reversal.

Parity In case of the parity symmetry, all the atomic positions change sign, i.e., $\mathbf{r}_{ji} \to -\mathbf{r}_{ji}$. We already know that the standard powerspectrum defined in Eq. (2.10) is isometrically invariant, and thus satisfies this property. However, in that case, we did not have any further coupling involved, such as the one with a vector field, treated here. To study how the powerspectrum components transform under parity, we first notice that the expansion coefficients c_{inlm}, as well as the coupled ones u_{inlm}, follow the same transformation rules of the spherical harmonics under an inversion of the arguments, i.e., $Y_l^{m*}(-\hat{\mathbf{r}}_{ji}) = (-1)^l Y_l^{m*}(\hat{\mathbf{r}}_{ji})$. Therefore, under parity, the powerspectrum transforms as

$$p_{inn'll'} \to (-1)^{l+l'} p_{inn'll'}. \tag{3.19}$$

If we compare this expression with Eq. (3.18), i.e., the one obtained by conjugating the powerspectrum, we notice how the two transformations lead to the very same multiplicative factor. Thus we deduce that selecting the real components of the powerspectrum, i.e., for $l+l'$ even, automatically satisfies the constrains that the powerspectrum is invariant under inversion of the atomic positions.

Time-reversal The time-reversal case can be seen as the analogous of the parity ones where the role of the atomic positions and the spins are swapped. Indeed, here we have that the atomic positions are left unchanged, while we invert the direction of all the vectors at play, i.e., $\mathbf{v}_j \to -\mathbf{v}_j$. Since the atomic positions and the vector components are treated on the same footing in the expression of the expansion coefficients c_{inlmq}, with the only exception being that the spherical-vector components behaves like the spherical harmonic of order $(1, q)$, then, by the same argument followed for the parity symmetry, the powerspectrum is always invariant under time reversal, since the emerging factor would

always be[16] $(-1)^{1+1} = 1$.

This analysis shows that the powerspectrum is always invariant under time reversal and can be constrained to be invariant under parity. For the remainder of this chapter we will consider only powerspectrum components such that $n = n'$ and $l = l'$, shortened to $p_{inlJ} := p_{innllJ}$ for readability.

The central atom This paragraph is devoted to show why care must be taken when the representation of the vector fields extends to the central atom of the atomic density. Indeed, since the origin of the frame of reference is, simultaneously, a fixed point for rotations around it, and a degenerate point for the representation in spherical harmonics, including a value of the field, there is the potential to break the rotational invariance. This can be easily proven in the special case of a system with just one, isolated, atom, i, in the center of the representation (the following discussion can be easily generalized to configurations with an actual neighborhood). In this case, the component u_{in100} reads

$$u_{in100} = w_{Z_i} R_{n1}(0) \sum_{mq} C^{00}_{1m1q} Y_1^{m*}(\hat{\mathbf{0}}) v_{i,q} = -\frac{1}{\sqrt{4\pi}} w_{Z_i} R_{n1}(0) v_{i,z}, \tag{3.20}$$

obtained by using $Y_l^{m*}(\hat{\mathbf{0}}) = \delta_{m0}\sqrt{\frac{2l+1}{4\pi}}$, $C^{00}_{l_1 m_1 l_2 m_2} = \delta_{l_1 l_2}\delta_{m_1 -m_2}(-1)^{l_1 - m_1}/\sqrt{2l_1+1}$, and $v_{i,0} = v_{i,z}$. Now, since the origin is a degenerate point for the representation in spherical harmonics, a rotation maps $Y_l^0(\hat{\mathbf{0}})$ in itself. This implies that u_{in100} is always proportional to the z-component of the vector field at the origin, and therefore, the powerspectrum is proportional to its magnitude, being $p_{in10} = |u_{in100}|^2$. Since the magnitude of the z component of a vector is manifestly not rotationally invariant, this proves that, when the value of the field in the origin is included in the description, at least one component of the powerspectrum is not invariant.

An intuitive way to look at this symmetry breaking is by considering that a vector in the origin of the frame of reference imposes a preferential direction: the representation acquires cylindrical symmetry in place of the spherical one. Thus, an operational way to solve the problem, would be to evaluate the powerspectrum after rotating the frame of reference: if the z-axis coincides with the direction of the vector in the origin, then the powerspectrum component p_{in10} coincides with the full magnitude of the vector (which is along z), and no directional ambiguity is present. However, the situation can be solved also if the radial basis vanishes at the origin for all $l > 0$. Indeed, as can be seen from the Eq. (3.20) above, when the radial function is coupled with the angular momentum l,

[16] Please note that the proof for the invariance of the powerspectrum under both parity and time-reversal is in stark contrast with the claim made in the preprint of Ref. [96], where it is said that the powerspectrum is not invariant under these two symmetry operations.

it can work as a "selector" of invariant components. Moreover, the case for $l = 0$ does not pose any problems, since the spherical harmonic $Y_0^0 = 1/\sqrt{4\pi}$ is a constant, and it can be easily proven that the powerspectrum components for $l = 0$ are proportional to the magnitude of the vector $\mathbf{v}_i$, which is rotationally invariant. The property of $R_{nl} = 0$, for $l > 0$ is indeed satisfied by the Spherical-Bessel descriptors [95], $g_{n-l,l}(r)$, the radial basis chosen for this work: in fact, these functions are constructed from the Spherical Bessel functions, $j_l(x)$, that vanish, for $l > 0$, at the origin of the frame of reference.

Heisenberg-like components In this small paragraph, we will evaluate the p_{in01} powerspectrum components, to develop a better insight into the powerspectrum properties. In this case, the expansion coefficients, u_{in01M}, are given by

$$u_{in01M} = \frac{1}{\sqrt{4\pi}} \sum_{j}^{\text{atoms}} w_{Z_j} v_{j,M} R_{n0}(r_{ji}), \tag{3.21}$$

so that the powerspectrum reads

$$p_{in01} = \frac{1}{4\pi} \sum_{jk}^{\text{atoms}} w_{Z_j} w_{Z_k} R_{n0}(r_{ji}) R_{n0}(r_{ki}) \mathbf{v}_j \cdot \mathbf{v}_k. \tag{3.22}$$

If, on one hand, the rotational invariance is manifest, we can also appreciate how these components resemble the analytical form of an Heisenberg model, with the coupling coefficients depending on the atomic coordinates. Please note that these components have expression because the $l = 0$ case washes out the information on the directionality of the atomic bonds (the spherical harmonics are constant): then, the only way to obtain a rotationally invariant quantities is by means of the scalar product $\mathbf{v}_j \cdot \mathbf{v}_k$. However, since the full powerspectrum includes also terms with $l > 0$, then its descriptive power goes beyond the simple Heisenberg form and, crucially, allows to couple the position degrees of freedom with the vectorial ones.

This concludes our investigation on the properties of the rotationally invariant powerspectrum for vectorial fields. We will now shortly discuss the model adopted in this work.

The fit of the energy We opted for an approach similar to the SNAP model one, i.e., we considered the linear fit of the energy with respect to the powerspectrum

$$E \simeq \boldsymbol{\theta} \cdot \sum_i \mathbf{p}_i = \sum_{n=0}^{n_{\text{nmax}}} \sum_{l=0}^{n} \sum_{J=|l-1|}^{J=l+1} \theta_{nlJ} \sum_i p_{inlJ}. \tag{3.23}$$

Here, we defined the two vectors, $\boldsymbol{\theta}$ and $\mathbf{p}_i$, containing, respectively, the coefficients of the fit and the powerspectrum components of the i-th configuration. The sum over n is, as usually done, truncated to an optimized $n_{\max}$. It is important to stress that, since the powerspectrum is at most quadratic in the magnitude of the vector (as can be seen noticing that the expansion coefficients u_{inlmq} are, at most, linear in the vectorial components), using a linear model could be detrimental when the expected interaction goes beyond quadratic terms. However, this can be addressed by considering, for example, a kernel-based model (as done in Ref. [93]). We will explicitly address this issue in the next section, and we will compare the performance of the linear model with a GAP-based one, for a toy-model of choice.

3.3 The Physical System

In order to test the descriptive power of our powerspectrum-based linear model, we chose to apply it to a toy model constructed over a spin-dependent Hamiltonian. The choice was driven by the necessity of perform a controlled benchmark of the methods in a fully parameterizable scenario, without the constraints imposed by spin-polarised *ab initio* calculations (or equivalent methods). Clearly, this is far from being an exhaustive investigation, albeit a necessary one.

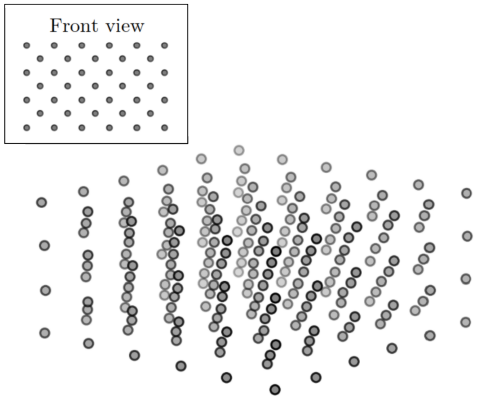

Figure 3.3: The system investigated: a rectangular block of 219 atoms in a *bcc* layout, constructed by stacking 6x6 and 5x5 squares.

The model was constructed over a rectangular clusters of 219 atoms, with a 6×6 atom base, each carrying a semi-classical spin, arranged in a *bcc* lattice. The system is shown in Fig. 3.3. We choose not to incorporate any periodic boundary conditions, to test the response of the model to different kinds of environment (bulk and surfaces) at once. The atoms were randomly displaced from the ideal *bcc* lattice to construct several datasets, of increasing degrees of distortion. The chosen Hamiltonian comprised an Heisenberg term and a Landau one, to explore both transverse and longitudinal degrees of freedom. Explicitly

$$H = H_{\mathrm{H}} + H_{\mathrm{L}}, \tag{3.24}$$

where

$$H_{\mathrm{H}} = -\frac{1}{2} \sum_{<i,j>} J_{ij}(r_{ij}) \mathbf{S}_i \cdot \mathbf{S}_j, \tag{3.25}$$

and

$$H_{\mathrm{L}} = \sum_i \left(A S_i^2 + B S_i^4 + C S_i^6 \right). \tag{3.26}$$

The Heisenberg term, H_{H}, is characterized by an exchange parameter that depends on the distance between the atoms, $J(r_{ij})$, so that the model incorporates a coupling between the spin degrees of freedom and the atomic positions (spin-phonons). Here the spins $\mathbf{S}_i$ are in units of $\hbar$, so that $S_i = |S_i| = M_i/g_e\mu_B$, with M_i being the corresponding magnetic moment and μ_B the Bohr magneton. The sum in H_H runs over pair of neighbors $< i, j >$. We choose the functional the following functional form for the exchange parameter J_{ij} [97]

$$J_{ij}(r_{ij}) = J_n(1 - \Delta r_{ij}/r_n)^3, \quad \text{with} \quad \Delta r_{ij} = r_{ij} - r_n, \tag{3.27}$$

where the distance r_n is the n-th neighbour distance in the ideal *bcc* lattice, and J_n is the coupling associated to the interaction between n-th neighbours in ideal conditions. The Landau term, H_{L}, depends on even powers of the magnitude of the local spins (longitudinal contributions), and is characterized by parameters A, B and C. We chose to specialize our description to *bcc* iron, with parameters taken from Ref. [98]: thus we considered an Heisenberg interaction extending to the second nearest neighbors, and with coupling constants $J_1 = 22.52$ meV and $J_2 = 17.99$ meV for the first and second neighbors respectively. Analogously, the Landau parameters have been set to $A = -440.987$ meV, $B = 150.546$ meV and $C = 50.769$ meV.

3.4 Numerical Results

In this section we will report the results obtained by training a powerspectrum model on the iron cluster. At first, we used a simplified version of the Hamiltonian, without the Landau contributions, i.e., containing only on Heisenberg term, $H = H_{\mathrm{H}}$, in which the lengths of the magnetic moments were fixed at 2.2 μ_{B}. We then incorporated also the Landau term H_L, and released the constraint on the fixed lengths.

3.4.1 The Heisenberg Model with lattice-dependent coupling constant

As first step, we directed our efforts in the construction of the training dataset. We built a total of six training sets, divided in two groups of three sets each.
The first group, called "ferromagnetic", has been constructed by randomly picking more than 200 atomic spins of the system (the actual number, between 200, and 219, was also draw at random), and orienting them along the z-axis. In this way, the system was in an al-

most z-aligned state. We then proceeded in displacing the atomic positions from the ideal *bcc* structure. The three training sets of the ferromagnetic group are then distinguished by the maximum random displacement allowed: explicitly the maximum distortion from the ideal lattice was of 5%, 10% and 20% of the lattice constant, for the first, second and third set respectively. This procedure was repeated till each dataset contained 100 configurations.

The construction of second group, called "random", followed the same strategy for the random displacement in the atomic positions, so that the dataset were characterized by 5%, 10% and 20% maximum displacement, respectively. However, in these cases we defined all the spins to be pointing in random directions. Also, we repeated the procedure to have again 100 different configurations in each of the three training sets.

In order to test the predictions of the model, we crafted three test sets for each of the six dataset used in the training. The first one was built by progressively aligning all the spins along the z-axis. Thus, it contains 219 configurations, in which the n-th configuration contains n randomly-selected spins aligned to the z-axis, while the other remained randomly oriented. Because this dataset was primarily devised to test the prediction across the magnetic landscape, the atoms were fixed in the ideal *bcc*-lattice positions. In contrast, the second and third test-sets were defined by the same atomic displacement of the relative dataset, to better gauge the response of the model to atomic displacements. In particular, the spins for the second test-set were selected to mirror the ferromagnetic configuration, with 200 randomly-selected spins aligned with the z-axis, and the remaining 19 randomly oriented. On the contrary, the third test set had all the spins randomly oriented in space. This was done to explicitly test those region of the PES which were either in the same energy range, or in a completely different one, of the energies explored during the training process.

The ferromagnetic dataset, with 10% displacement Here we discuss the results for the potential trained on the ferromagnetic dataset, in the case of 10% maximum displacement. The other two cases (5% and 20%) are reported in Table 3.1. They will not be discussed in detail, having similar performance of the 10% one.

We used a Ridge-regression fitting strategy, with an optimized regularization constant of $\alpha = 3.2 \times 10^3$, and a truncation parameter for the expansion of $n_{\max} = 4$: in this way, the resulting model was very compact, with only 35 features. The cut-off radius was optimized to be $r_c = 1.4$ (in units of the lattice constant). Given the limited size of the dataset, we opted for a 5-fold cross-validation procedure, with a 80/20 split ratio between training and validation sets. The obtained Mean Absolute Errors (MAEs) were $(4.83 \pm 0.15) \times 10^{-5}$ eV/atom and $(6.8 \pm 0.7) \times 10^{-5}$ eV/atom, for the training and the

"Ferromagnetic" training sets

	Training Parameters			MAE (10^{-5}eV/atom)			
Disp.	$n_{\max}$	r_{cut}	α	Train	Validation	Prediction	Upper
5%	5	1.4	1.7×10^2	1.3 ± 0.07	2.05 ± 0.14	7.3	6.9
10%	4	1.4	3.2×10^3	4.83 ± 0.15	6.8 ± 0.7	56	82
20%	6	1.35	3.9×10^4	13.9 ± 1.2	27 ± 3	61	110

"Random" training sets

	Training Parameters			MAE (10^{-5}eV/atom)			
Disp.	$n_{\max}$	r_{cut}	α	Train	Validation	Prediction	Upper
5%	5	1.45	1.9×10^2	1.74 ± 0.04	4.7 ± 0.3	13	4.9
10%	4	1.4	8.4×10^3	6.5 ± 0.3	12.3 ± 0.9	250	83
20%	4	1.4	2.1×10^4	23.0 ± 0.9	38 ± 3	10^3	320

Table 3.1: We report here the optimized hyperparameters and performances of the trained models. The upper table shows the cases trained on the "ferromagnetic" dataset, while the lower table shows the ones trained on the "random" dataset, with the indicated maximum value of the atomic-position displacement (5%, 10% and 20%). The cut-off radius, r_{cut} is given in units of the lattice constant. The other hyperparameters are the truncation of the powerspectrum expansion, $n_{\max}$, and the regularization constant of the Ridge regression, α. The training and validation values are obtained form a 5-fold cross-validation procedure, i.e., we report the mean and the standard deviation values. The "Prediction" column reports the MAEs evaluated on the first of the test set, i.e., the one with progressively aligned spins. The "Upper" column reports the MAEs on the same set, but calculated only for energies > -0.01 eV/atom, i.e., energies far from the ones in the training set for the potentials trained on the ferromagnetic dataset (upper table) and, instead, energies close to the ones of the random dataset (lower table).

validation set, respectively. Compared to the range of variability of the training set, of about 10^{-2} eV/atom, we see that this errors correspond to approximately 0.1% of the energies variability.

We then tested the prediction of the model on the first prediction set, i.e., the one obtained by progressively aligning all the spins of the system while keeping the atoms pinned in the ideal *bcc* lattice positions. The resulting MAE was of 5.6×10^{-4} eV/atom. The parity plot is shown in Fig. 3.4. Crucially, evaluating the MAE only for configurations with energy above -0.01 eV/atom, i.e., the ones with energies that are the furthest from the training set, gives a value of 8.2×10^{-4} eV/atom, which is comparable with the one on the full set. This suggests that the model is capable of achieving rather accurate results on the full energetic landscape, even in regions which as further from the ones explored during the training phase.

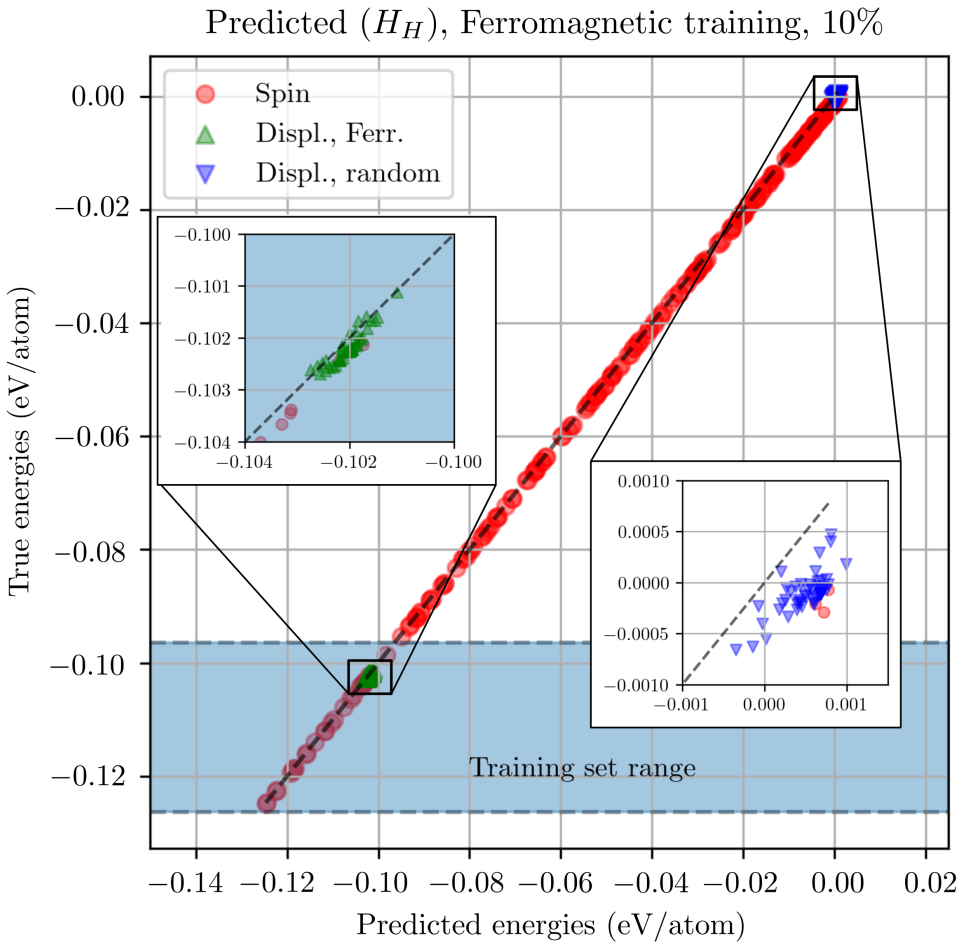

Figure 3.4: Predicted against actual energies for a Ridge-regression trained on the ferromagnetic dataset with 10% maximum atomic displacement. The actual energies are calculated analytically from the Heisenberg model, H_H, of Eq. (3.25), where the coupling constant exhibit also a spin-lattice coupling. The three different colours represent the different test set that we devised to probe the accuracy of the model. The red circles refer to the systems on an ideal *bcc* lattice, but with progressively z-aligned magnetic moments. The other dots represent different displacements of the atoms for near to ferromagnetic and paramagnetic configurations, respectively. The figure demonstrates the good agreement of the potential also for configurations which are energetically far from that used in the training (blue region). Zoom-in around different energy regions are displayed in the inserts.

An overview of the results for the other cases is given in Table 3.1: we can notice how increasing the degree of distortion reduces the quality of the model. However, the accuracies reached are still quite elevated across all cases, if compared to the range of energies involved. The parity plot for the worst case (for the case of 20% maximum distortion) is shown in Fig. 3.5(a), and confirms the good agreement between the model and the analytic energies.

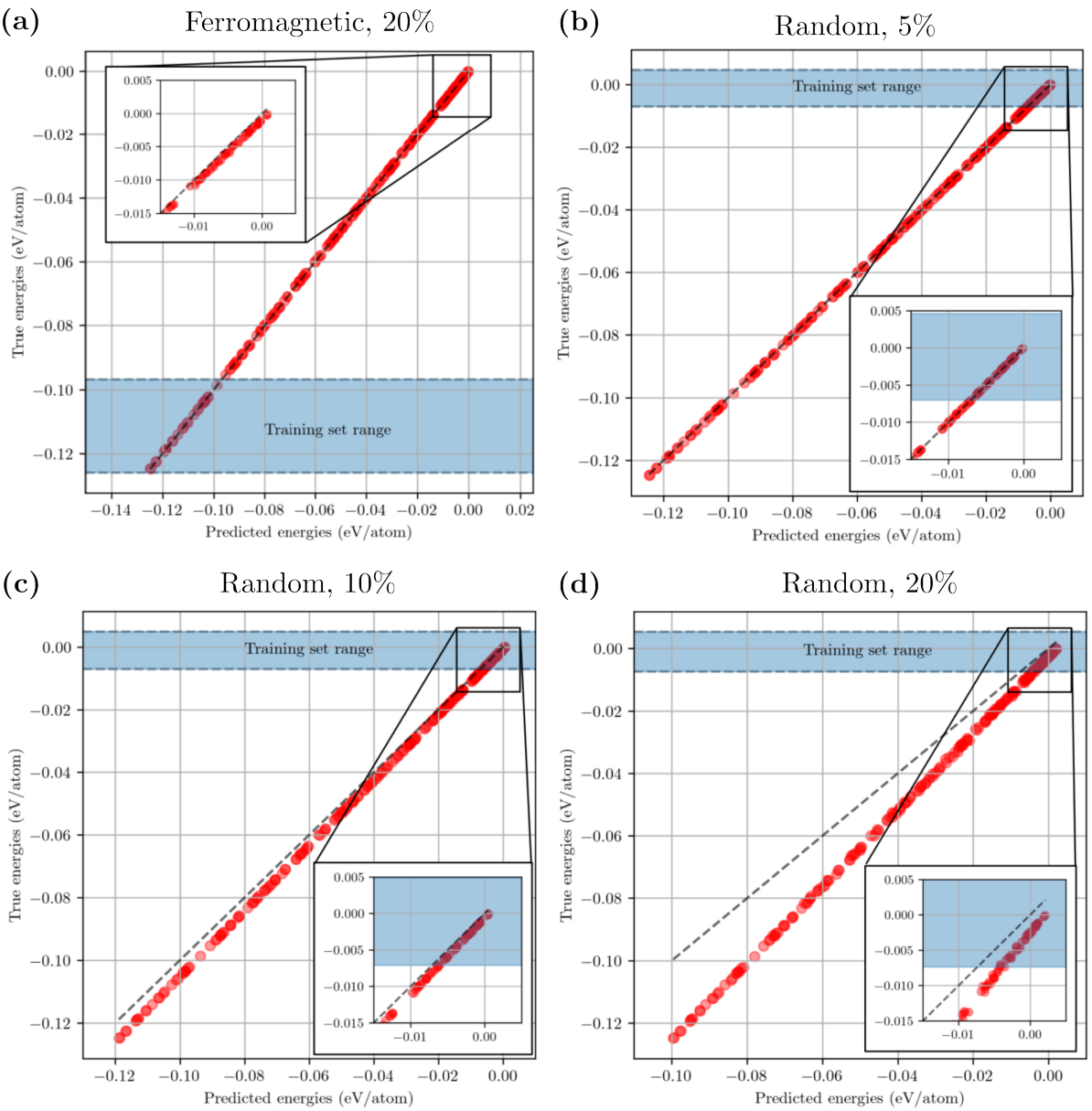

Figure 3.5: The figures show parity plots that are analogous to the one reported in Fig. 3.4 (please, note that the axis range and labels are the same). (a) We report the parity plot for the worst performing "ferromagnetic-trained" model, i.e., the one trained with a 20% maximum displacement. We can appreciate how the model retains good accuracies even when probed on energies which are far from the range explored during the training phase. The insert shows a magnification of the highest end of the energies range, the furthest one from the training set. (b)-(d) The parity plots for all the model training on randomly oriented spins are shown, with increasingly value of the maximum atomic displacement. We can appreciate how the models still performs quite accurately for the 5% and 10% cases, but that the accuracies rapidly deteriorate, down to the failure of the 20% case. We can also notice how the energy range explored by the training set is much more contained that the one spanned in (a). We remark that, contrary to (a), the inserts now show a magnification of the performance of the model in the region explored during the training.

Random datasets Unlike the previous case, the model trained on randomized spins proved to be less accurate in predicting the energies of the system. In fact, despite a roughly doubled MAE on the test sets, it failed in the prediction on ferromagnetic configurations, with an increasing loss in accuracy with higher displacements. This can be seen from the Table 3.1, where it is shown that the prediction on the aligning spin dataset is of the order of 10 meV/atom, despite the accuracy shown on the test set of 0.38 meV/atom. Also, it can be seen how the accuracy of the two model on the highest-energy structures is lower than the one obtained from the ferromagnetic case, despite the fact that the model trained on random configurations is indeed trained on the highest end of the energy spectrum. The parity plots in Fig.s 3.5(b)-(d) explicitly show how these models deviate from the parity line, in particular when extending to unseen region of the energy landscape. We attributed this failure to the fact that training on randomly oriented spin do not offer the correct strategy to train the model. Indeed the atomic environments are too homogeneous, and thus the descriptive capabilities of the models are severely impaired. This is indeed confirmed by the range of energies covered by the training set, also shown in Fig.s 3.5(b)-(d): when compared to the one of the model trained on the ferromagnetic case, we can see that, energy-wise, much fewer distinct cases are actually explored. This is in stark contrast with the model trained on the ferromagnetic cases, which was able to maintain high accuracy on all the predicted configurations.

3.4.2 Including longitudinal excitations via the Landau Term

We now include the Landau term H_L, to investigate the performance of the model against the complete Hamiltonian of Eq. (3.24). In this case we studied only the model corresponding to the ferromagnetic case, with a 10% maximum random displacement. Importantly, since the Landau term incorporates longitudinal excitations, we selected a randomly chosen length in the range from 1.8 μ_B to 2.3 μ_B, for the randomly oriented spins, while keeping the z-aligned ones constrained to 2.25 μ_B. We kept the same cross-validation procedure, with five 80/20 splits. The test set follows the same approach of the progressively aligning one introduced above. Here, however, we constrained the z-aligned magnetic moments to have a fixed length of 2.25 μ_B, while the length of the randomly oriented one varied in the range from[17] 1.9 μ_B to 2.3 μ_B. The optimized parameters are $n_{\max} = 4$ for the expansion truncation, $r_{\text{cut}} = 1.4$ (lattice constant) for the cut-off radius, and $\alpha = 2 \times 10^5$ for the regularisation constant. The MAEs obtained in this case are $(4.9 \pm 0.1) \times 10^{-4}$ eV/atom and $(6.0 \pm 0.5) \times 10^{-4}$ eV/atom, respectively, for the training and validation sets. Comparing to the model trained on the Heisenberg

[17] We use a slightly different range for the test set to ensure that the predictions are in regions where the energies are not too high with respect to the minimum energies explored.

terms only, we obtained a value of the MAE of roughly one order of magnitude larger. As already noted, this can be intuitively understood in terms of the dependence of the Landau term on the fourth and sixth power of the magnitude of the spins, in contrast with the quadratic behaviour of the powerspectrum. The MAE evaluated on the prediction set, is of 6.0×10^{-3} eV/atom, and the resulting parity plot is shown in Fig. 3.6. It is interesting to study the two extremes of the energy ranges, one compatible with the energies explored in the training phase and the other for the energies on randomly oriented spins. We then evaluated the MAEs for energies less than -0.32 eV/atom and above -0.26 eV/atom, and obtained the values 7.1×10^{-4} eV/atom and 1.0×10^{-2} eV/atom respectively. We can conclude that the model possesses a good accuracy in interpolating between the energies in the range of the training set, but that the performances deteriorate when the model is forced to adapt to unseen regions of the energy landscape, and in particular when is tested against configurations with randomly oriented spins.

We tested our hypothesis on the limitations imposed by the quadratic nature of the model, by combining the powerspectrum representation with a non-linear framework. In particular, we considered a GAP [26] model, as introduced in Eq. (2.40), but adapted to the vector-field powerspectrum. In this way, the atomic energy of the i-th atom can be written as

$$\varepsilon_i = \sum_{k=0}^{N_{\text{train}}} \theta_t \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{p}_i - \mathbf{p}_t)^2\right\} \tag{3.28}$$

where the sum is extended over all the atoms in the training set, and where $\mathbf{p}_i$ and $\mathbf{p}_t$ are the powerspectrum, respectively, of the i-th atom and of the t-th atomic configuration of in the training set. As extensively discussed in section 2.1.2, the GAP formalism is based on the definition of a similarity kernel, S, which is, generally, not linear. This allows us to describe energy contributions going beyond the quadratic order in the spin magnitude. We remark that the approach followed here is slightly different from the actual GAP one, since our density is still defined in terms of delta functions, contrary to the Gaussian formulation of GAP. The parity plot obtained from the GAP model is shown in Fig. 3.6. We can appreciate that we do not observe an increase in performance with respect to the linear model, in the range explored by the training set. Indeed, we see that the performance associated with the configurations having energy smaller than -0.32 eV/atom remains very close to that one of the linear model. However, the total MAE decreases to 4.5×10^{-4} eV/atom (compared to 7.1×10^{-4} eV/atom of the linear model) and, more importantly, the MAE for energies larger than -0.26 eV/atom is now reduced to 5.8×10^{-3} eV/atom, i.e., is halved. This is clearly showed in the figure, where the GAP model is able to stay closer to the parity line when the linear model deviates the most. Since, generally, one does not have the control on the orders of the spin contributions, we conclude that using a non-linear model could improve the performance in real scenarios,

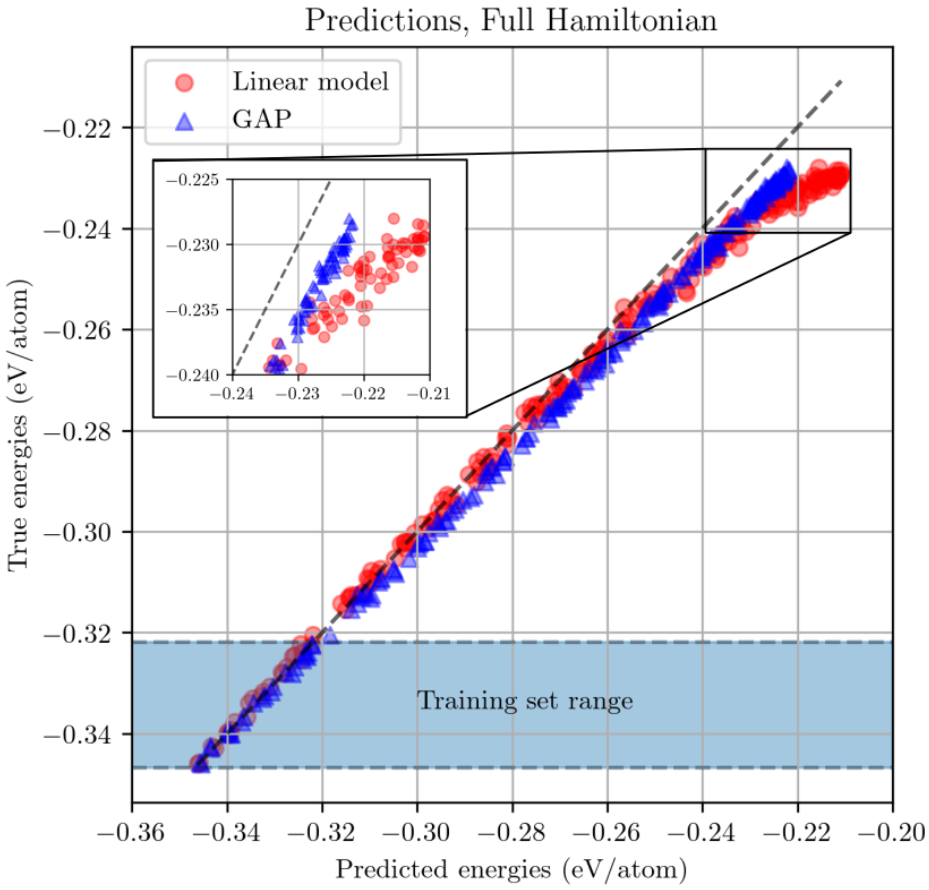

Figure 3.6: We show here the parity plot for the true vs predicted energy from the Landau model described by the full Hamiltonian of Eq. (3.24). The predictions are done on the dataset with progressively aligning spins. We can notice that, despite the high accuracy in the range explored by the training set, the linear model (represented by red circles) strongly deviates from the parity line for the highest-end of the energies. Instead, a GAP-adapted model, from Eq. (3.28) (blue triangles), while still showing a deviation from the parity line, manages to keep a good accuracy along all the full range of energies explored.

namely when the energies are obtained from an *ab-initio* method of choice.

3.5 Conclusions

In this chapter we have introduced a powerspectrum representation for vectorial fields. We developed the methods in depth, grounding our approach in the usual hypothesis of decomposition in atomic contributions of the energy, and of the short-ranged nature of the interaction. We then defined a vector field representation of the atomic environment, where the atomic positions are localized by Dirac-delta functions. Finally, relying on the harmonic decomposition of vectors, and on the rotational properties of the bipolar-spherical harmonics, we constructed the rotational invariant powerspectrum for a simultaneous rotation of the field and of the atomic positions. After a detailed investigation of

the analytical property of the powerspectrum, ranging from its behaviour under the action of fundamental symmetries, to the role of the vector field in the center of the frame of reference, we constructed a toy model to freely test the capabilities of the descriptors. We simulated a block of iron atoms placed in a *bcc* lattice, with an underlying Hamiltonian containing Heisenberg and Landau terms. The interaction between the spin degrees of freedom and the lattice was enforced by position-dependent coupling constants for the Heisenberg model. We then proceeded in constructing several datasets that could test the descriptive power of the powerspectrum for different degrees of displacement and orientation of the spins. In particular we followed a hierarchical approach, starting from the Heisenberg model only, and then adding Landau contributions. After choosing a linear model to build a better intuition on the advantages and limitations of the powerspectrum, we proceeded to the training and testing phases: we concluded that, for the Heisenberg model, the best strategy was to train the model on configurations with most of the spin aligned to the z axis, and only a small portion of the remaining ones randomly oriented. This allowed to obtain a model that was capable to achieve extremely high accuracies even for configurations that were energetically very different from the ones in the training set. We then included the Landau term. This proved to be challenging for a linear model, and we hypothesized that this was caused by the quadratic nature of the powerspectrum, with respect to the magnitude of the vector field, in contrast with the sixth power reached by the Landau term. Thus, we adopted the GAP framework to induce non-linearity in the powerspectrum. This proved to be beneficial and resulted in a more accurate model also for energy that were far from the range explored by the training set.

One of the most important underlying features of the model was the use of the bipolar-spherical harmonics to achieve the correct invariance under rotation. This confirms the relevance of the multi-polar spherical harmonics as the underlying natural language to define the powerspectrum and the bispectrum components, which was made manifest already in the previous chapter. We will heavily rely on the multi-polar spherical harmonics also in the last chapter of the thesis, when we will show another strategy to obtain the ACE formalism, and we will expose its limitation for the investigation of higher-body order terms. However, before that, we will now dive in the core of the thesis, by introducing the Jacobi-Legendre formalism. The next few chapters will be devoted to the definition of descriptors which are capable of predicting any quantity of choice, from scalar, to scalar fields, reaching also tensor and tensor fields, within a coherent and generalized framework. The construction of this framework will be tied up with the expansion in multi-polar spherical harmonics of the last chapter, when it will guide us to a new descriptor for the five-body terms in an ACE formalism.

Chapter 4

Jacobi-Legendre potentials

In this chapter, we will review the MLP introduced in the published work "Cluster expansion constructed over Jacobi-Legendre polynomials for accurate force fields", Ref. [99], of which I am a co-author. The main aim of the work was to introduced a competitive MLP, based on the internal coordinates representation for the atomic environments.

The idea behind this MLP was to generate descriptors that could be easily interpreted, and that would not be based on the formalism of the coupling schemes of angular momenta. Moreover, we wished to define a framework that could be naturally generalized and extended beyond the description of scalar quantities, making it suitable also for scalar fields, tensorial quantities and tensorial fields. In this spirit, this chapter introduces the foundation for all the remainder of the thesis, and thus most of the definitions given here will find full fruition when applied to the generalization proposed in subsequent chapters, when the full framework will be explored.

As will be shown below, the core of the Jacobi-Legendre formalism, introduced here, lies in the cluster expansion of the atomic contributions to the short-range energies of a system, in the same spirit as the one introduced for the ACE (please see section 2.1.4). These similarities, as well as the differences, will be investigated in detail in this chapter. We remark, however, that the discussion that will be initiated here will culminate in the last chapter of the thesis, where the benefits of using an internal-coordinate-based coupling scheme will be made more manifest, in particular in relation to the five-body order expansion.

This chapter will be structured as follows: firstly we will introduce the mathematical framework and the main assumptions for the descriptors. This will constitute also the main underlying assumptions for all the following chapters. Then we will proceed in a systematic analysis of different body-order potentials, $v^{(n)}$, starting from an in-depth overview of the two-body one, $v^{(2)}$, where we will explore many of the core strategies of this work. Our analysis will terminate at the five-body order, which will be only outlined.

A graphical abstract of this chapter is shown in Fig. 4.1, where the core ideas behind the Jacobi-Legendre potentials are shown. Importantly, after introducing the three-body order expansion, we will also discuss in detail how to achieve linear computational cost with the number of atoms in the cut-off, all without sacrificing the representation in internal coordinates. In doing so, we will manifest the connection between the formalism proposed here, and the approaches based on the density-trick and introduced in Chapter 2.

After a presentation of the features of the linear model, from how to compute and take into account forces, to the loss function used, we will show a first application of the potential, done on a well-known carbon dataset. We will conclude by showing that the Jacobi-Legendre formalism is able to reach state-of-the-art accuracies also when dealing with phonon-dispersion curves.

My role in the manuscript was devoted to the construction of the mathematical framework and to the investigating of the formal details of the potential, while the implementation part has been performed almost entirely by the other authors, under constant sharing of ideas on every aspect of the work.

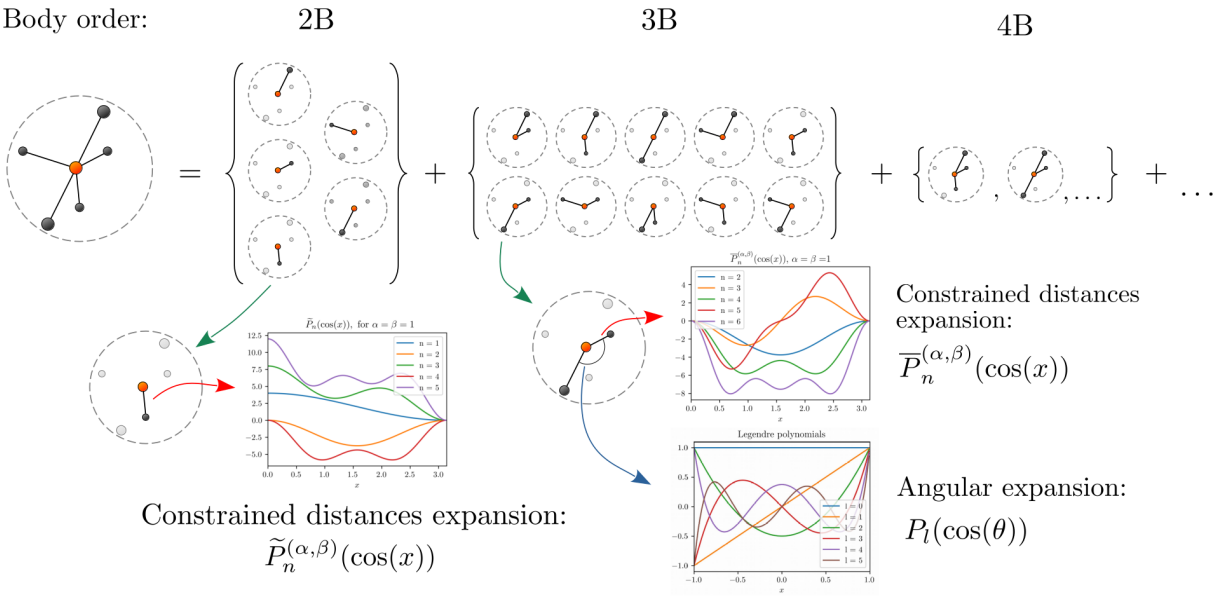

Figure 4.1: We show here a graphical abstract for the Jacobi-Legendre potentials. The main underlying assumptions is the factorization of the energy in atomic, short-ranged, contributions. In particular, this work heavily relies on the cluster expansion of the atomic energies in multi-body terms. Each energy contribution is then expanded in Jacobi polynomials (for the distances) and Legendre polynomials (for the angles). Crucially, we will enforce constraints on the expansion, so that only potentials encoding desired physical features will be taken into account.

4.1 A Cluster-Expansion based Machine-Learning Potential

The starting point for our discussion is the cluster expansion of the *atomic* energies in terms of internal coordinates. The main assumption adopted here is same of all the other descriptors presented so far, i.e., the factorization of the energy in atomic, short-ranged contributions. Explicitly, the atomic energies, ε_i, are defined by means of

$$E_{\text{short}} = \sum_{i}^{\text{atoms}} \varepsilon_i, \qquad \text{and} \qquad \varepsilon_i = \sum_{v}^{\substack{\text{body}\\ \text{order}}} \varepsilon_{Z_i}^{(v)}, \tag{4.1}$$

where we applied a cluster expansion for the atomic energies, with (please, compare with eq, (2.51))

$$\begin{cases} \varepsilon_{Z_i}^{(2)} = \sum_{(j)_i}^{\text{atoms}} v_{Z_j Z_i}^{(2)}(r_{ji}), \\ \varepsilon_{Z_i}^{(3)} = \sum_{(jk)_i}^{\text{atoms}} v_{Z_j Z_k Z_i}^{(3)}(r_{ji}, r_{ki}, s_{jki}), \\ \varepsilon_{Z_i}^{(4)} = \sum_{(jkp)_i}^{\text{atoms}} v_{Z_j Z_k Z_p Z_i}^{(4)}(r_{ji}, r_{ki}, r_{pi}, s_{jki}, s_{jpi}, s_{kpi}), \\ \varepsilon_{Z_i}^{(5)} = \sum_{(jkpq)_i}^{\text{atoms}} v_{Z_j Z_k Z_p Z_q Z_i}^{(5)}(r_{ji}, r_{ki}, r_{pi}, r_{qi}, s_{jki}, s_{jpi}, s_{jqi}, s_{kpi}, s_{kqi}, s_{pqi}), \\ \dots . \end{cases} \tag{4.2}$$

Here $\varepsilon_{Z_i}^{(1)}$ is a constant shift. By using the distances r_{ji} and the scalar products $s_{jki} := \hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}$, we already made manifest the choice of the internal coordinate as degrees of freedom of our representation. Indeed, given the use of internal coordinates, the potential is constrained to be rotationally invariant *by design*, in contrast with a spherical-harmonics-based construction, where the invariance must be reached by selecting the correct coupling schemes, i.e., by selecting only the components that belong to the rotationally invariant space. Moreover, this choice has two important properties: on one hand, the internal coordinates being a complete representation[18], it has the same descriptive power of other complete methods. On the other, we will show that we can actually recover an expansion in terms of spherical harmonics: this will have the double effect of re-introducing a linear scaling (in the same fashion of the density trick) and of showing a different, more hidden,

[18]The internal coordinate representation is complete up to the five-body terms, becoming over-complete for higher-body orders.

approach to the coupling of angular momenta. This new coupling will not only preserve, by design, the *isometrical* invariance of the expanded quantities, but it will also inherit all the properties of the internal coordinates representation, specifically it will make no reference to the order of the coupling of single angular-momentum channels. We will also make explicit reference to the atomic species Z_i, which will allow us to impose the correct symmetries on the expanded potentials, $v^{(n)}$, under permutations of identical atoms.

A crucial point in the construction of the potential will be the choice of the expansion basis for the radial and the angular variables of the potentials, which will be discussed at length in the following sections. This choice will follow three main ideas: the first will be to keep the formulation as general as possible, to reduce the implicit choices in the construction of the potential (this will lead to the choice of the Jacobi polynomials). The second will consist in maintaining consistency with the descriptions adopted by other MLP, specifically in constructing rotational invariant quantities (which will lead to an expansion in terms of Legendre polynomials). The last one, will be to be able to define a general framework able to tackle non-scalar targets, such as scalar fields, tensorial quantities or even tensorial fields. This will be clearly shown in the next chapters. We note that defining a potential in terms of the internal coordinates has also been pursued in a recently developed potential, designed as *proper orthogonal descriptors* (please, see Refs. [100, 101]).

We will now proceed in a detailed analysis of the potentials, $v^{(n)}$, defined in Eq. (4.2), in a hierarchical way, starting from the two-body ones and progressively increasing the body order.

4.1.1 Two-body (2B) potential

While the two-body potential, $v^{(2)}_{Z_j Z_i}(r_{ji})$, is the simplest case to investigate, it deserves a thorough analysis aimed to introduce core concepts that will be extensively exploited in the following sections. First of all, we remark that the class of potentials treated here are only short-ranged ones: indeed we will assume that the potentials will smoothly vanish for distances, r_{ji}, larger then an optimized cut-off radius, r_{cut}, i.e., $v(r_{ji}) \simeq 0$, for $r_{ji} \geq r_{\text{cut}}$. A consequence of this property is that we can always devise a continuous, one-to-one mapping, x, so that the significant distances are mapped in the interval $[-1, 1]$, i.e., $x : [r_{\text{min}}, r_{\text{cut}}] \to [-1, 1]$ (here we are taking into account also a minimum possible distance, r_{min}). For the remainder of the work we will always use the "cosine" map defined as

$$x \equiv x(r; r_{\text{min}}, r_{\text{cut}}) = \cos\left(\pi \frac{r - r_{\text{min}}}{r_{\text{cut}} - r_{\text{min}}}\right). \tag{4.3}$$

In the following, it will be shown that this map has some interesting properties when combined with the choice of the Jacobi polynomials. However, we remark that different

mappings are entirely possible. Now, since the mapping is a one-to-one, and so it is invertible, we can define the composition

$$v^{(2)}_{Z_jZ_i}(r_{ji}) = \underbrace{(v^{(2)}_{Z_jZ_i} \circ x^{-1})}_{=:f}(x_{ji}) = f^{(2)}_{Z_jZ_i}(x_{ji}), \tag{4.4}$$

where $x_{ji} := x(r_{ji})$ is used as practical shorthand. If we make the further operative assumption that the potential v is at least square integrable with respect to some weight $w(x)$, then it admits an expansion in an orthogonal-polynomials basis of choice [102]. Let us suppose that $\{P_n(x)\}$ is such basis. Thus we can write the expansion

$$v^{(2)}_{Z_jZ_i}(r_{ji}) = f^{(2)}_{Z_jZ_i}(x_{ji}) = \sum_n a_n^{Z_jZ_i} P_{nji}, \tag{4.5}$$

where we defined the shorthand $P_{nji} := P_n(x_{ji})$. We remark that the coefficients mirror the dependence of the potential on the atomic species Z_j and Z_i: indeed, in general, we expect a change in the functional form of the potential when considering atoms of different species.

Constrained functions We will now show one of the main ideas of our framework, which allows to encode desired constraints directly in the expansion of the potentials. Crucially, from Eq. (4.5), we can impose constraints on the function f by directly intervening on the expansion coefficients. For example, the short-ranged nature of the potentials $v^{(2)}_{Z_jZ_i}$ implies that the function f must vanish when $f(x(r_{\text{cut}})) = 0$. This can be imposed directly on the coefficients by evaluating

$$f(x(r_{\text{cut}})) = \sum_n a_n P_n(x(r_{\text{cut}})) = 0, \tag{4.6}$$

so that

$$a_0 = -(P_0(x(r_{\text{cut}})))^{-1} \sum_{n=1}^{\infty} a_n P_n(x(r_{\text{cut}})). \tag{4.7}$$

Here we assumed that $P_0(x(r_{\text{cut}}))$ is not zero: this is always possible because, the basis being complete, at least one polynomial is necessarily not vanishing at $x(r_{\text{cut}})$. By substituting a_0 from Eq. (4.7) back in Eq. (4.5), we get the new expansion

$$f(x_{ji}) = \sum_{n=1}^{\infty} a_n \widetilde{P}_{nji}, \tag{4.8}$$

where we now defined the "vanishing"-polynomials

$$\widetilde{P}_{nji} = P_{nji} - \frac{P_{0ji}}{P_0(x(r_{\text{cut}}))} P_n(x(r_{\text{cut}})). \tag{4.9}$$

This is just one example among all the possible constraints that can be imposed: similar ones can involve any other point, i.e., by substituting $x(r_{\text{cut}})$ with an arbitrary $\overline{x}$, or be imposed on derivatives, integrals, and any other linear operator acting on the polynomial basis. If more than one constraint is required, this procedure can be generalized by progressively imposing more conditions, until all of the requirements are met.

The Jacobi polynomials We introduce now the radial basis chosen for this work, i.e., the Jacobi polynomials, $\{P_n^{(\alpha,\beta)}\}$ [55]. The Jacobi polynomials constitute a general class of polynomials, that, for each choice of real numbers $\alpha, \beta > -1$, defines a complete and orthogonal basis for the square-integrable functions on the interval $[-1, 1]$. The pair (α, β) determines the inner on the interval $[-1, 1]$, by means of the weights $w^{(\alpha,\beta)}(x) = (1-x)^\alpha (1+x)^\beta$. The determination of the weights can also be interpreted as the choice of a metric on the interval $[-1, 1]$: in other words, portions of the interval are magnified (stretched) by this choice. This allows, in principle, to naturally magnify regions of the interval that are more relevant to the determination of the potential, neglecting other portions where, for example the data are scarce or absent. The Jacobi polynomials are also very general, encompassing basis functions such as the Legendre polynomials (for $\alpha = \beta = 0$), or the Gegenbauer polynomials (for $\alpha = \beta =$ half integers). We already saw an example of Gegenbauer polynomials in section 2.1.2, where the 4-dimesional bispectrum was introduced: this solidifies our metric-based intuition, since we can recover the same metric of the surface of a 4-dimensional hypersphere. Thus, optimizing the pair (α, β), allows to automatically select an adequate radial basis, with no need to make any implicit assumption on the choice of the basis set or of the metric. An example of Jacobi polynomial, evaluated on the cosine map defined in Eq. (4.3), is reported on the left-side of Fig. 4.2.

Specializing the constraining procedure of Eq. (4.9) to the expansion in Jacobi polynomials, we can define the vanishing-Jacobi polynomials as

$$\widetilde{P}_{nji}^{(\alpha,\beta)} = P_{nji}^{(\alpha,\beta)} - P_n^{(\alpha,\beta)}(-1), \tag{4.10}$$

where we used $P_0^{(\alpha,\beta)} = 1$, and the fact that the cosine map of Eq. (4.3) implies $x(r_{\text{cut}}) = -1$. In Fig. 4.2 we report examples of vanishing-Jacobi polynomials, as evaluated for the cosine mapping.

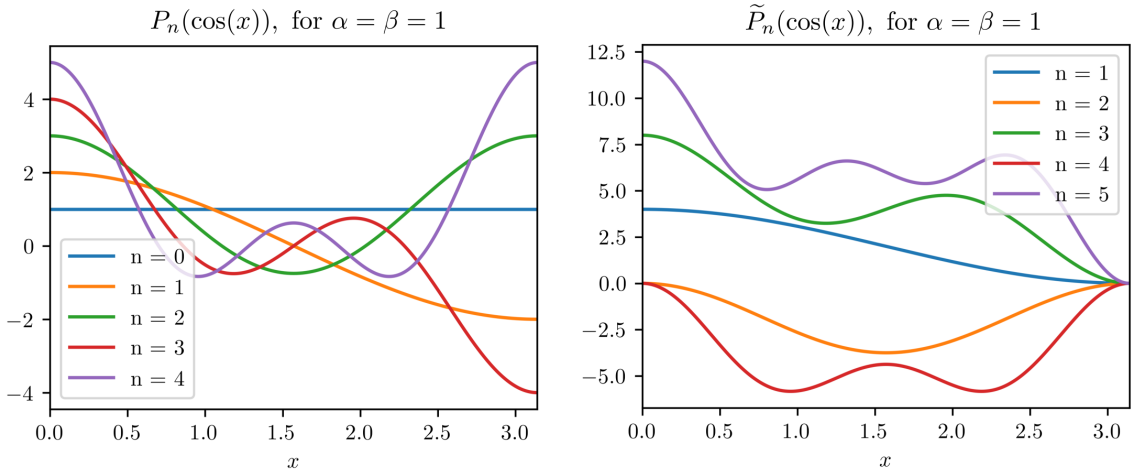

Figure 4.2: (Left) The figure shows the first five Jacobi polynomials for $\alpha = \beta = 1$. The cosine function encodes the effect of the cosine map of Eq. (4.3). Please, note that this map constrains the polynomials to have zero derivatives at the edges of the interval $[0, \pi]$. (Right) The figure show the first five vanishing-Jacobi polynomials, defined by means of Eq. (4.10), for the same α and β of the Left figure. It can be appreciated how the polynomials are constrained to vanish toward the right-end of the interval, which represents the cut-off radius.

Emergence of the cut-off function An interesting feature of the combination of the constraining procedure and the use of the cosine map, is the natural emergence of the cut-off function defined in Eq. (2.2). Indeed, we can use the series expansion for the Jacobi polynomials (see Ref. [55])

$$P_n^{(\alpha,\beta)}(x) = \frac{1}{2^n}\sum_{j=0}^{n}\binom{n+\alpha}{j}\binom{n+\beta}{n-j}(x-1)^{n-j}(x+1)^j, \tag{4.11}$$

and the cosine map $x \to \cos(x)$ so that we can write

$$P_n^{(\alpha,\beta)}(\cos(x)) = \sum_{j=0}^{n}\binom{n+\alpha}{j}\binom{n+\beta}{n-j}(-1)^{n-j}\sin^{2(n-j)}(x/2)\cos^{2j}(x/2), \tag{4.12}$$

where we used the identities

$$\cos(x) - 1 = -2\sin^2(x/2), \quad \text{and} \quad \cos(x) + 1 = 2\cos^2(x/2). \tag{4.13}$$

Now, the definition of the vanishing-Jacobi polynomial of Eq. (4.10), requires the evaluation of the above formula in $x = \pi$: due to the $\cos(x/2)$ terms, the only surviving

contribution is the one for $j = 0$. Thus we can write

$$\begin{aligned}\widetilde{P}_n^{(\alpha,\beta)}(\cos(x)) = \sum_{j=1}^{n}(-1)^{n-j}\binom{n+\alpha}{j}\binom{n+\beta}{n-j}\sin^{2(n-j)}(x/2)\cos^{2j}(x/2)+ \\ +(-1)^n\binom{n+\beta}{n}\left(\sin^{2n}(x/2)-1\right),\end{aligned} \tag{4.14}$$

where we separated the contribution for $j = 0$ from the rest. By using the identity

$$\sin^{2n}(x/2) - 1 = -\cos^2(x/2)\sum_{j=1}^{n}\sin^{2(n-j)}(x/2), \tag{4.15}$$

we can finally write

$$\widetilde{P}_n^{(\alpha,\beta)}(\cos(x)) = f_c(x)Q_n^{(\alpha,\beta)}(\cos(x)), \tag{4.16}$$

which shows the natural emergence of the cut-off function $f_c = (\cos(x) + 1)/2 = \cos^2(x)$. Here $Q_n^{(\alpha,\beta)}$ are functions defined by the sum

$$\begin{aligned}&Q_n^{(\alpha,\beta)}(\cos(x)) = \\ &= (-1)^n\sum_{j=1}^{n}\left[(-1)^j\binom{n+\alpha}{j}\binom{n+\beta}{n-j}\cos^{2(j-1)}(x/2) - \binom{n+\beta}{n}\right]\sin^{2(n-j)}(x/2).\end{aligned} \tag{4.17}$$

Since the Jacobi polynomials encompass a large class of orthogonal polynomials, we deduce that the crucial point of this derivation lies in the choice of the mapping. Another important point is the fact that, if we focus only on the vanishing-Jacobi polynomials, we do not need to take into account the cut-off function explicitly, contrary to what is usually done for other descriptors. If this allows for simpler analytic expressions, it has also the advantage that the derivatives are easier evaluated, and that no subsequent orthogonalization procedure has to take place, the expansion being based on the Jacobi polynomials, which are already orthogonal.

The 2B potential and the behaviour at the origin In conclusion, we have that the potential can be written in terms of vanishing-Jacobi polynomials, as

$$v_{Z_jZ_i}^{(2)}(r_{ji}) = \sum_{n=1}^{n_{\max}} a_n^{Z_jZ_i}\widetilde{P}_n^{(\alpha,\beta)}\left(\cos\left(\pi\frac{r_{ji}-r_{\min}}{r_{\text{cut}}-r_{\min}}\right)\right) =: \sum_{n=1}^{n_{\max}} a_n^{Z_jZ_i}\widetilde{P}_{nji}^{(\alpha,\beta)}, \tag{4.18}$$

with the vanishing-Jacobi polynomials, $\widetilde{P}_n^{(\alpha,\beta)}$, defined in Eq. (4.10). We will call this expansion the 2B-Jacobi-Legendre (2B-JL) expansion. Please note that the our definition of the potentials is symmetric under swap of the indices, i.e., $v_{Z_jZ_i}^{(2)} := v_{Z_iZ_j}^{(2)}$. This can be important when dealing with multi-species 2B clusters. The 2B-JL expansion is

characterized by five hyperparameters, which are allowed to be different among different-atomic-species clusters. The hyperparameters, along with their range of definition, are $\alpha, \beta \in (-1, \infty)$, $n_{\max} \in \mathbb{N}^+$, $r_{\text{cut}} \in (0, \infty)$ and $r_{\min} \in (-\infty, r_{\text{cut}})$. Given the large number of parameters to optimize, it could be necessary to perform some approximation: for example, in many cases we found that $r_{\min} = 0$ is an adequate choice. Also, since there is usually a large range for optimal (α, β) pairs, a simpler approach consists in optimize for $\alpha = \beta$. It is also worth to investigate the behaviour of the potential at the origin, i.e., when $r_{ji} = 0$. If we put $r_{\min} = 0$, we obtain

$$v^{(2)}_{Z_j Z_i}(0) = \sum_{n=1}^{n_{\max}} a_n^{Z_j Z_i} \widetilde{P}_n^{(\alpha,\beta)}(1), \tag{4.19}$$

and from the identity

$$\widetilde{P}_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n} - (-1)^n \binom{n+\beta}{n}, \tag{4.20}$$

we can conclude that the magnitude of the potential can become quite large at small distances. This behaviour is clearly determined by the coefficients $\{a_n^{Z_j Z_i}\}$: however, we can bias the choice of the hyperparameters so that we select only potentials that present a strong repulsive behaviour at small distances. This can be easily done by visualizing the function

$$v^{(2)}_{Z_j Z_i}(r) = \sum_{n=1}^{n_{\max}} a_n^{Z_j Z_i} \widetilde{P}_n^{(\alpha,\beta)} \left(\cos\left(\pi \frac{r - r_{\min}}{r_{\text{cut}} - r_{\min}} \right) \right), \tag{4.21}$$

with the learned coefficients. This procedure could be impaired if higher-body order terms are allowed to influence the behaviour at small distances. However, as will be shortly shown, we will impose further constraints to prevent this interference, and to encode the idea that the repulsive behaviour is essentially a 2B interaction.
This is in contrast with what is usually done in literature, when a repulsive function (usually exponential) is imposed on the small distances region of the potential, where the data are scarce or completely absent.

4.1.2 Three-body (3B) potential

This section is devoted to the definition of the expansion of the three-body (3B) potentials $v^{(3)}$. The construction follows closely the one for the two-body potential, with a first expansion in a product of two Jacobi polynomials, one for each distance. However, in this case, it is necessary to consider also the angular contributions given by the scalar product between the directions of the two bonds. These are expanded in terms of Legendre polynomials (hence the name Jacobi-Legendre expansion). Indeed, guided by their expansion

in a sum of products of spherical-harmonics (see (2.15)), they seemed to be the natural choice for the expansion of terms, depending on a scalar product between two versors. Not only will the use Legendre polynomials be pivotal in establishing a connection with other descriptors, but also, their decomposition in spherical harmonics will be crucial for the generalization of the JL expansion to scalar fields, tensors and tensor fields, as will be shown in following chapters. Explicitly, the unconstrained expression for the potential reads

$$v^{(3)}_{Z_jZ_kZ_i}(r_{ji}, r_{ki}, s_{jki}) = \sum_{n_1n_2l} a^{Z_jZ_kZ_i}_{n_1n_2l} P^{(\alpha,\beta)}_{n_1ji} P^{(\alpha,\beta)}_{n_2ki} P^{jki}_l, \tag{4.22}$$

where we introduced the shorthand $P^{jki}_l = P_l(\hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki})$, with P_l being a the l-th Legendre polynomial.

Constraining the potential Analogously to what has been done for the 2B case, we want to constrain the 3B potential to vanish whenever *at least one* of the distances approaches the cut-off radius. This can be done by following the very same procedure introduced in the previous sections, which, in this case, leads to the constraints (please, compare with the constraint (4.7))

$$\begin{cases} a^{Z_jZ_kZ_i}_{0n_2l} = -\sum_{n_1\geq 1}^{n_{\max}} a^{Z_jZ_kZ_i}_{n_1n_2l} P^{(\alpha,\beta)}_{n_1}(-1) & \text{for all } n_2, l, \\ a^{Z_jZ_kZ_i}_{n_10l} = -\sum_{n_2\geq 1}^{n_{\max}} a^{Z_jZ_kZ_i}_{n_1n_2l} P^{(\alpha,\beta)}_{n_2}(-1) & \text{for all } n_1, l. \end{cases} \tag{4.23}$$

The crucial point here is that the two distances are constrained independently, i.e., the potential goes to zero when one of the distances approaches the cut-off, independently from the value of the other. By plugging these constraints back in Eq. (4.22), it can be shown that the resulting expression can be obtained by simply substituting the Jacobi polynomials with their vanishing counterpart (the polynomials defined in Eq. (4.10)). As already mentioned in the previous section, we want to enforce the idea that the repulsive behaviour at small distances is characterized only by a 2B interaction. For this reason, we can further constrain the 3B potential to vanish also when at least one of the distances approaches zero. This brings the new set of constraints (please, compare again with the constraint (4.7), where, instead of $x(r_{\text{cut}})$ we have $x(0)$).

$$\begin{cases} a_{1n_2l}^{Z_jZ_kZ_i} = -(\widetilde{P}_1^{(\alpha,\beta)}(1))^{-1} \sum_{n_1\geq 2}^{n_{\max}} a_{n_1n_2l}^{Z_jZ_kZ_i} \widetilde{P}_{n_1}^{(\alpha,\beta)}(1) & \text{for all } n_2 \geq 1, \text{ and all } l, \\ a_{n_11l}^{Z_jZ_kZ_i} = -(\widetilde{P}_1^{(\alpha,\beta)}(1))^{-1} \sum_{n_2\geq 2}^{n_{\max}} a_{n_1n_2l}^{Z_jZ_kZ_i} \widetilde{P}_{n_2}^{(\alpha,\beta)}(1) & \text{for all } n_1 \geq 1, \text{ and all } l. \end{cases} \tag{4.24}$$

Plugging the constraints back in the potential expansion, we finally obtain

$$v_{Z_jZ_kZ_i}^{(3)}(r_{ji}, r_{ki}, s_{jki}) = \sum_{n_1,n_2=2}^{n_{\max}} \sum_{l=0}^{l_{\max}} a_{n_1n_2l}^{Z_jZ_kZ_i} \overline{P}_{n_1ji}^{(\alpha,\beta)} \overline{P}_{n_2ki}^{(\alpha,\beta)} P_l^{jki}, \tag{4.25}$$

in which we defined the "double"-vanishing-Jacobi polynomials, as

$$\overline{P}_n^{(\alpha,\beta)}(x) := \widetilde{P}_n^{(\alpha,\beta)}(x) - \frac{\widetilde{P}_n^{(\alpha,\beta)}(1)}{\widetilde{P}_1^{(\alpha,\beta)}(1)} \widetilde{P}_1^{(\alpha,\beta)}(x). \tag{4.26}$$

An example of double-vanishing-Jacobi polynomials is shown in Fig. 4.3. It is worth mentioning that, by means of the constraining procedure adopted, not only does the 3B potential vanish at both edges of the interval of definition (both at small distances and at the cut-off radius), but also there is a significant reduction in the number of independent expansion coefficients, from $(n_{\max}+1)^2(l_{\max}+1)$ in the unconstrained case, to $(n_{\max}-1)^2(l_{\max}+1)$ in the constrained one. As can be seen, the reduction is much more significant for relatively small $n_{\max}$. We also remark that alongside the same cluster-dependent hyperparameters of the 2B case, we now also have a truncation of the angular expansion up to the index $l_{\max} \in \mathbb{N}$ and thus each cluster carries six hyperparameters.

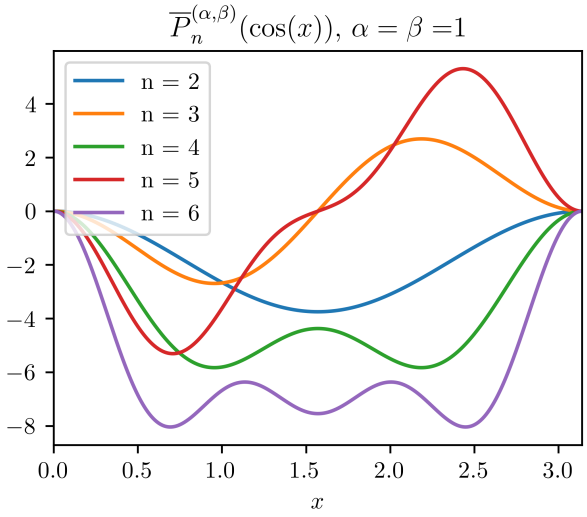

Figure 4.3: We show here the first five double-vanishing-Jacobi polynomials, as defined by Eq. (4.26), for the same $\alpha = \beta = 1$ chosen in Fig. (4.2). The figure shows how the polynomials are constrained to go to zero at both edges of the interval.

Symmetries of the potential Another important aspect of the 3B potentials lies in the ordering of the atomic numbers Z_j, Z_k and Z_i. Specifically, the first atomic species

refers to the one of j-th atom, which corresponds to the first distance, r_{ji}. Analogously, the second species refers to the one of the atom that is considered in the second distance, r_{ki}. Finally the last atomic species refers to the central atom. Thus, if we swap the role of the distances in the potential, we must also swap the functional dependence: we then deduce the symmetry rule

$$v^{(3)}_{Z_jZ_kZ_i}(r_{ji}, r_{ki}, s_{jki}) = v^{(3)}_{Z_kZ_jZ_i}(r_{ki}, r_{ji}, s_{jki})\,. \tag{4.27}$$

This is mirrored by the expansion coefficients $a^{Z_jZ_kZ_i}_{n_1n_2l}$, since the first index, n_1, refers to the first distance, while the second index, n_2, refers to the second distance. Therefore it must also hold that $a^{Z_jZ_kZ_i}_{n_1n_2l} = a^{Z_kZ_jZ_i}_{n_2n_1l}$. While this symmetry is important to define unique clusters and avoid redundancies, it becomes even more relevant when the atoms j and k belong to the same species, i.e., $Z_j = Z_k = Z$. In this case, the expansion coefficients become symmetric under the swap of Jacobi indexes $n_1 \leftrightarrow n_2$, i.e., they satisfy

$$a^{ZZZ_i}_{n_1n_2l} = a^{ZZZ_i}_{n_2n_1l}. \tag{4.28}$$

This is a fundamental symmetry, that must be imposed to enforce the invariance of the potentials with respect to permutations of identical atoms. This means that, in order to ensure the symmetry of the coefficients, the expansion becomes

$$\begin{aligned} v^{(3)}_{ZZZ_i}(r_{ji}, r_{ki}, s_{jki}) = &\sum_{n=2}^{n_{\max}} \sum_{l=0}^{l_{\max}} a^{ZZZ_i}_{nnl} \overline{P}^{(\alpha,\beta)}_{nji} \overline{P}^{(\alpha,\beta)}_{nki} P^{jki}_l + \\ &+ \sum_{\substack{n_1=2 \\ n_2=2 \\ n_1>n_2}}^{n_{\max}} \sum_{l=0}^{l_{\max}} a^{Z_jZ_kZ_i}_{n_1n_2l} \left[\overline{P}^{(\alpha,\beta)}_{n_1ji} \overline{P}^{(\alpha,\beta)}_{n_2ki} + \overline{P}^{(\alpha,\beta)}_{n_2ji} \overline{P}^{(\alpha,\beta)}_{n_1ki} \right] P^{jki}_l, \end{aligned} \tag{4.29}$$

where we separated the same indexes contributions, $n_1 = n_2 = n$, from the remaining cases, $n_1 \neq n_2$. We can simplify this expression by defining the 3B-Jacobi-Legendre (3B-JL) expansion

$$v^{(3)}_{Z_jZ_kZ_i}(r_{ji}, r_{ki}, s_{jki}) = \sum_{n_1n_2l}^{\text{unique}} a^{Z_jZ_kZ_i}_{n_1n_2l} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1ji} \overline{P}^{(\alpha,\beta)}_{n_2ki} P^{jki}_l \right), \tag{4.30}$$

where the first summation runs over all the unique coefficients with respect to a symmetry rule such as (4.28), e.g., we sum only over $n_1 \geq n_2$ in the example above. Instead, the second one runs over all the possible permutations of indexes that lead to equivalent coefficients (the swap $n_1 \leftrightarrow n_2$ in the example above). Thus, the definition given in Eq. (4.30) reduces to Eq. (4.29) when $Z_j = Z_k = Z$. This formalism can be easily

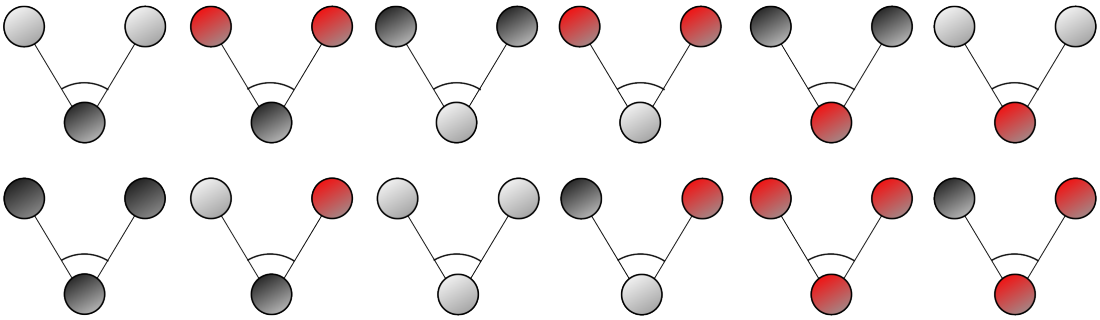

Figure 4.4: We graphically show the type of 3B clusters that should be considered when dealing with a 3-species system (here in black, red and white respectively). Please note that clusters obtained by a swap of the two atoms in the neighborhood are implicitly taken into account by means of the symmetry rule in Eq. (4.28).

generalized to higher-body order terms: indeed, it only requires us to define the list of symmetries that are satisfied by the expansion coefficients. Note, however, that higher-body orders will also introduce more than one angle, and, therefore, the symmetry will get more complicated than a simple swap of indexes.

On multi-species systems In this short paragraph, we address the problematic increase in the number of clusters with the number of atomic species in the system. Indeed, as shown in Fig. 4.4 for the simple case of just three species, the number of clusters can rapidly become intractable (also considering that all the hyperparameters are, in principle, independent for each cluster). A possible way to mitigate this problem is to implement a cluster selection, either based on physical knowledge, or obtained by introducing an initial "screening" phase, focused on the determination of which cluster can be neglected without affecting the performance of the model. Another strategy is to perform approximations, such as label two atomic species that are expected to behave in a similar way, as the same species. We stress that, however, there is not a good-for-all solution to this problem.

Linear scaling and connection with the powerspectrum Let us now discuss how the choice of Legendre polynomials naturally leads to a formulation of the 3B expansion in terms of the powerspectrum. Please note that what follows in this paragraph will be not used in the developed version of the potential. Indeed, the role of this paragraph is to show that a connection with other descriptors exists and that a linear scaling (with respect to the number of atom in the atomic neighborhood) can be achieved.

Here, we will exploit the connection between the powerspectrum and the expansion in Legendre polynomials shown in Eq. (2.16). The first step consists in using the addition

theorem for the spherical harmonics (from Eq. (2.15), and reported here for readability)

$$P_l(\hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}) = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} Y_l^{m*}(\hat{\mathbf{r}}_{ji}) Y_l^m(\hat{\mathbf{r}}_{ki}). \tag{4.31}$$

We also define the Jacobi-Legendre-(JL-)atomic basis (analogous to the atomic basis introduced in the ACE formalism, and strictly related to it)

$$(J_p L_q)^{i,Z}_{n_1 \dots n_p l_1 m_1 \dots l_q m_q} = \sum_{j \in Z} \left[\prod_{r=1}^{p} \overline{P}^{(\alpha,\beta)}_{n_r j i} \right] \left[\prod_{s=1}^{q} Y_{l_s}^{m_s}(\hat{\mathbf{r}}_{ji}) \right]. \tag{4.32}$$

Here, the indexes p and q shows how many double-vanishing Jacobi polynomials and Legendre polynomials are present in the expansion. Also, the summation runs over all the atoms of the same species in the surrounding of the central atom i: thus, crucially, the evaluation of the JL-atomic basis is *linear* with respect to the number of atoms in the neighborhood. We note that, since both Jacobi polynomials and Legendre polynomials are complete, this atomic basis could appear somehow redundant: indeed, there exist coefficients $c^n_{n_1 n_2}$, such that

$$\overline{P}^{(\alpha,\beta)}_{n_1 j i} \overline{P}^{(\alpha,\beta)}_{n_2 j i} = \sum_n c^n_{n_1 n_2} \overline{P}^{(\alpha,\beta)}_{n j i}. \tag{4.33}$$

This implies that the basis can be always reduced to linear combination of $(J_1 L_1)^{i,Z}_{nlm}$ terms (the same argument holds for the spherical harmonics). However, because the coupling coefficients $c^n_{n_1 n_2}$ are usually evaluated by performing numerical integrations and, more importantly, since they depend on the pair (α, β), we cannot build a look-up table. For this reasons we decided not to perform any basis reduction. We also note that the case $(J_1 L_1)^{i,Z}_{n_1 l_1 m_1}$ is equivalent to the ACE-atomic basis defined in Eq. (2.57), if the radial functions are identified with the double-vanishing-Jacobi and Legendre polynomials.
With this basis choice, the 3B-atomic energy term, $\varepsilon^{(3)}_{Z_i}$, becomes

$$\begin{aligned} \varepsilon^{(3)}_{Z_i} &= \sum_{(jk)_i}^{\text{atoms}} v^{(3)}_{Z_j Z_k Z_i}(r_{ji}, r_{ki}, s_{jki}) = \sum_{(jk)_i}^{\text{atoms}} \sum_{n_1 n_2 l}^{\text{unique}} a^{Z_j Z_k Z_i}_{n_1 n_2 l} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1 j i} \overline{P}^{(\alpha,\beta)}_{n_2 k i} P_l^{jki} \right) \\ &= \sum_{Z_1 \geq Z_2} \sum_{\substack{(j,k)_i \\ j \in Z_1 \\ k \in Z_2}}^{\text{atoms}} \sum_{n_1 n_2 l}^{\text{unique}} a^{Z_1 Z_2 Z_i}_{n_1 n_2 l} \frac{4\pi}{2l+1} \sum_{\text{symm.}} \left(\sum_{m=-l}^{l} \overline{P}^{(\alpha,\beta)}_{n_1 j i} \overline{P}^{(\alpha,\beta)}_{n_2 k i} Y_l^{m*}(\hat{\mathbf{r}}_{ji}) Y_l^m(\hat{\mathbf{r}}_{ki}) \right), \end{aligned} \tag{4.34}$$

where we firstly separated the contributions in terms of the atomic species, and then we applied the addition theorem. Now, the only thing that prevents us from using the JL-atomic basis of Eq. (4.32) is the constrained summation over all the *unique* pairs for

atoms. We first notice that the constraint holds only when the atomic species are the same. Then we can leverage on the fact that performing the inner sum over the indexes swap is equivalent to permuting j and k in the 3B-JL expansion. Therefore, we can release the constraint on the sum by simply adding and subtracting terms like

$$\frac{1}{2}\sum_{Z}\sum_{(j\in Z)_i}\sum_{n_1n_2l}^{\text{unique}} a^{ZZZ_i}_{n_1n_2l}\overline{P}^{(\alpha,\beta)}_{n_1ji}\overline{P}^{(\alpha,\beta)}_{n_2ji} = \frac{1}{2}\sum_{Z}\sum_{n_1n_2l}^{\text{unique}} a^{ZZZ_i}_{n_1n_2l}(J_2L_0)^{i,Z}_{n_1n_2}. \tag{4.35}$$

These can be interpreted as a "self-interaction" terms, since we can read them as if the central atom was to interact "twice" with the same atom in the environment. The self-interaction contributions are also discussed in the ACE formalism (see, for example Fig. 2.7 and Refs. [25, 67]) where, however, they are reduced to combination of 2B terms and then re-absorbed in the expansion. On the contrary, our aim here is to keep a clear separation between different body orders for two reasons: if, on one hand, different body orders are defined in terms of different hyperparameters (specifically, a different truncation $n_{\max}$, different cut-off radii r_{cut}, and different pairs (α,β)), on the other hand we aim in keeping a formal link between the internal coordinate representation and the atomic-basis one, so that it is possible to use one or the other, at need[19]. Finally, we can write

$$\varepsilon^{(3)}_{Z_i} = \sum_{Z_1\geq Z_2}\sum_{n_1n_2l}^{\text{unique}} b^{Z_1Z_2Z_i}_{n_1n_2l}\left(C^{(3),Z_1Z_2}_{in_1n_2l} - S^{(3),Z_1Z_2}_{in_1n_2}\right), \tag{4.36}$$

where we defined the coupling term, $C^{(3),Z_1Z_2}_{in_1n_2l}$, and the self-interaction, $S^{(3),Z_1Z_2}_{in_1n_2}$, as

$$\begin{cases} C^{(3),Z_1Z_2}_{in_1n_2l} := \dfrac{4\pi}{2l+1}\displaystyle\sum_{m=-l}^{l}(-1)^m(J_1L_1)^{i,Z_1}_{n_1lm}(J_1L_1)^{i,Z_2}_{n_2l-m}, \\ S^{(3),Z_1Z_2}_{in_1n_2} := \delta_{Z_1Z_2}(J_2L_0)^{i,Z_1}_{n_1n_2}. \end{cases} \tag{4.37}$$

The coefficients $b^{Z_1Z_2Z_i}_{n_1n_2l}$ are defined in terms of the $a^{Z_1Z_2Z_i}_{n_1n_2l}$ as

$$b^{Z_1Z_2Z_i}_{n_1n_2l} = \frac{a^{Z_1Z_2Z_i}_{n_1n_2l}}{1+\delta_{Z_1Z_2}\delta_{n_1n_2}}, \tag{4.38}$$

so that spurious multiplicative factors are correctly taken into account. We note how the coupling term is formally equivalent to the powerspectrum defined in Eq. (2.10) and, therefore, also to the term $B^{(3)}_{in_1n_2l}$ in Eq. (2.59) for the ACE model. We also note that our selection of unique coefficients, and subsequent sum over symmetries, is equivalent

[19] If there are not many atoms in the atomic environment, the computational overhead of evaluating the spherical harmonics and contracting them could make the linear-scaling scheme less efficient than the internal coordinates one.

to the lexicographic order $\nu_1 \geq \nu_2$ used in ACE (2.59) (please, see also Ref. [103]): we stress, however, that this equivalence will not hold when we will introduce an analogous expansion for the four-body term.

Crucially, in Eq. (4.36), we shifted from an expression which was quadratic in the number of neighbors inside the cut-off sphere (as can see from the constrained sum over all the unique pairs in Eq. (4.30)), to an equivalent expression which is, instead, linear.

Connection with "the density trick" Before proceeding with an in-depth analysis of the four-body terms, let us strengthen the connection between the "powerspectrum-like" term, given from $C^{(3),Z_1Z_2}_{in_1n_2l}$ and recovered from the enforcement of a linear scaling, and the powerspectrum obtained from the density trick and first introduced in Sec. 2.1.2. The main objective is to construct a neighbor density using the Jacobi polynomials as radial basis. The strategy is similar to the one showed in Sec. 2.1.2, where the definition of a map from the real space, $\mathbb{R}^3$, to the surface of a four-dimensional space, S^3, was used to define the new density $\rho_i(\mathbf{r}) \xrightarrow{\mathbf{r}\to\hat{\mathbf{u}}} \rho_i(\hat{\mathbf{u}})$, which was subsequently used in the SNAP model [24]. Analogously, we can directly use the cosine map, defined in Eq. 4.3, to construct a neighbor density, ρ_i, which is defined on the domain $[-1,1]\otimes S^2$, where S^2 is the surface of the unitary, 3-dimensional sphere. Explicitly, we have

$$\rho_i(\mathbf{r}) \xrightarrow{r\to x(r)} \rho_i(x,\hat{\mathbf{r}}), \tag{4.39}$$

where we recall that the cosine map, x, is defined from $[r_{\min}, r_{\text{cut}}]$ to $[-1,1]$, as

$$x(r) = \cos\left(\pi \frac{r - r_{\min}}{r_{\text{cut}} - r_{\min}}\right). \tag{4.40}$$

We can now proceed with a constructive approach, by *assuming* that the density $\rho_i^{JL}(x,\hat{\mathbf{r}})$ is given as

$$\rho_i^{JL}(x,\hat{\mathbf{r}}) := \sum_{nlm} c^{JL}_{inlm} P_n^{(\alpha,\beta)}(x) Y_l^m(\hat{\mathbf{r}}), \tag{4.41}$$

with the coefficients given by

$$c^{JL}_{inlm} := \sum_j^{\text{atoms}} P^{(\alpha,\beta)}_{nji} Y_l^{m*}(\hat{\mathbf{r}}_{ji}), \tag{4.42}$$

and where the sum runs over all the atoms inside the cut-off sphere. These coefficients can be easily obtained by means of the integral

$$c_{inlm} = \frac{1}{N_n^{(\alpha,\beta)}} \int \mathrm{d}\hat{\mathbf{r}} \int_{-1}^{1} \mathrm{d}x\, w^{(\alpha,\beta)}(x) \rho_i^{JL}(x,\hat{\mathbf{r}}) P_n^{(\alpha,\beta)}(x) Y_l^{m*}(\hat{\mathbf{r}}), \tag{4.43}$$

where we used the weight function[20] $w^{(\alpha,\beta)}(x) = (1-x)^\alpha(1+x)^\beta$, and $N_n^{(\alpha,\beta)}$ is a normalization constant given by the integral [55]

$$\int \mathrm{d}x\, w^{(\alpha,\beta)}(x) P_m^{(\alpha,\beta)}(x) P_n^{(\alpha,\beta)}(x) = \delta_{mn} N_n^{(\alpha,\beta)}, \tag{4.44}$$

which is explicitly computed as

$$N_n^{\alpha,\beta)} = \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{n!\Gamma(n+\alpha+\beta+1),} \tag{4.45}$$

with $\Gamma(x)$ being the Gamma function. Now, it is interesting to investigate if the defined density is indeed a localization function, with well-recognizable peaks in correspondence of the atomic positions. This can be done by considering the $l = m = 0$ case, since the angular dependence of the function is the same of the one of the Dirac-delta distributions of Sec. 2.1.2, which are, by definition, localized. Explcitly, we can investigate the sum

$$S_{in}^{(\alpha,\beta)} := \sum_n c_{in00}^{JL} = \frac{1}{\sqrt{4\pi}} \sum_n \sum_j^{\text{atoms}} P_{nji}^{(\alpha,\beta)}. \tag{4.46}$$

While a detailed analysis of this sum is outside the scope of our discussion, Fig. 4.5 (top) shows examples of truncated $S_{in}^{(\alpha,\beta)}$, evaluated over four randomly-selected atomic positions, and with $\alpha = \beta = 2$. On the one hand, the Figure reveals that the expression does indeed define a localization function. On the other hand, it also shows the presents of a divergence at the edges of the interval of definition. As mentioned before, this shows how the weight, $w^{(\alpha,\beta)}(x)$, can be seen as imposing a new metric on the interval, which is analogous to the different metric adopted by the choice of the SNAP formalism over the surface of the 4-dimensional sphere. This can be seen by plotting the same function multiplied by the weight $w^{(\alpha,\beta)}(x)$, as shown in Fig. 4.5 (bottom): not only this procedure regularizes the peaks' height, which shows how the weight function is responsible for a different metric on the interval, but also the divergences at the edges of the interval is now disappeared. Even more, we can appreciate how the derivatives of the resulting expression vanish at the edge of the interval, a fact that can be easily shown by calculating the derivative of $w^{(\alpha,\beta)}(x) S_{in}^{(\alpha,\beta)}$. In particular, it is easy to show that the derivative of order k, evaluated on the right or left edge, vanish if and only if $\alpha > k$ or $\beta > k$, respectively. We just mention that this property could be used as a further strategy to address the need for an explicit cut-off function, and could give an enhanced control over the behaviour of

[20] Technically, we are defining a function which belongs to the space $L^2_{w^{(\alpha,\beta)}}([-1,1] \otimes S^2)$, of square-integrable functions, with respect to the given weight. Given that the function is defined by means of its series expansion, and given that the series is always truncated to a suitable value, we can avoid any detail regarding the convergence of the proposed expression.

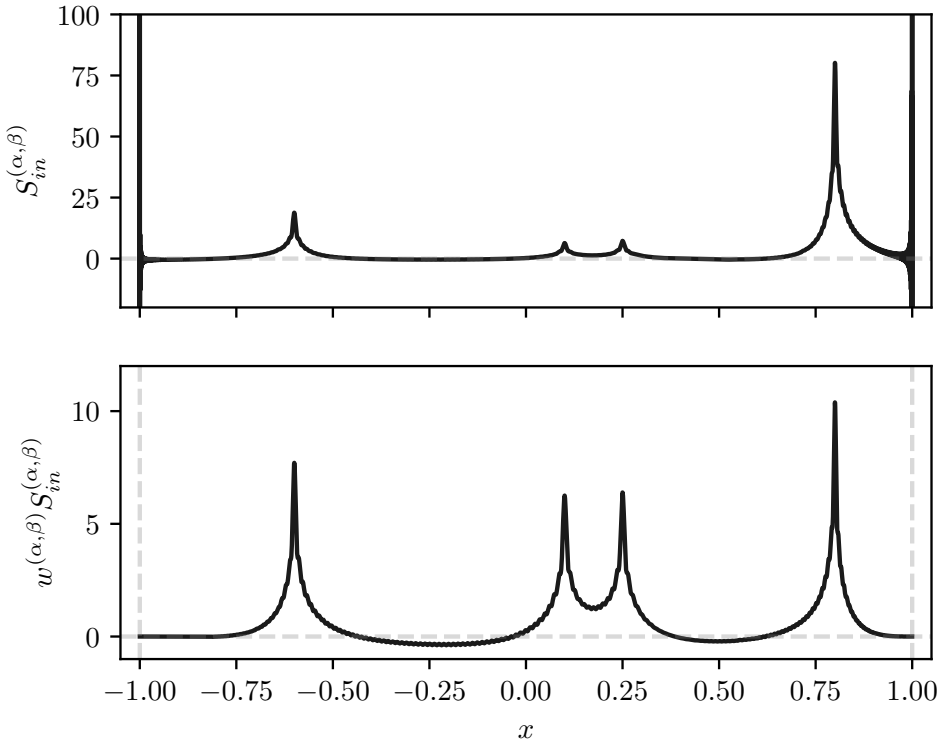

Figure 4.5: (Top) The Figure shows a possible example of the sum from Eq. 4.46. The atomic positions are at $x_1 = -0.6$, $x_2 = 0.1$, $x_3 = 0.25$ and $x_4 = 0.8$. This example is constructed with $\alpha = \beta = 2$, and only the case $n = 500$ is shown. It is also shown how the density possesses peaks on the atomic position and divergences at the edges of the domain, for $x = \pm 1$.
(Bottom) The Figure shows the same function scaled by the weight $w^{(\alpha,\beta)}(x)$. From the Figure, it can be appreciated how the terms $S_{in}^{(\alpha,\beta)}$ are based on an underlying metric over the interval $[-1, 1]$. Indeed, not only the divergences disappeared (with the function now smoothly vanishing at the edges), but also the peaks are on the same scale.

the derivatives at the edge of the interval.

With a well-defined neighbor density, we can now follow the same procedure discussed in Sec. 2.1.2, and construct the powerspectrum or the bispectrum, in complete analogy to what done for the GAP framework. Indeed, a powerspectrum can be easily defined by means of the contraction

$$p_{inn'l}^{JL} = \sum_m (-1)^m c_{inlm}^{JL} c_{in'l-m}^{JL}, \tag{4.47}$$

which can be used in either a linear model or a kernel-based model.

Finally, we can compare this expression with $C_{in_1n_2l}^{(3)}$ (note that, for the sake of com-

pactness, we are not considering details regarding the atomic species), explicitly

$$\begin{aligned} C^{(3)}_{in_1n_2l} &= \frac{4\pi}{2l+1}\sum_m (-1)^m (J_1L_1)^i_{n_1lm}(J_1L_1)^i_{n_2l-m} \\ &= \frac{4\pi}{2l+1}\sum_m (-1)^m \sum_{jk}^{\text{atoms}} \overline{P}^{(\alpha,\beta)}_{n_1ji} Y_l^m(\hat{\mathbf{r}}_{ji})\overline{P}^{(\alpha,\beta)}_{n_2ji} Y_l^{-m}(\hat{\mathbf{r}}_{ki}), \end{aligned} \tag{4.48}$$

and

$$p^{JL}_{in_1n_2l} = \sum_m (-1)^m c^{JL}_{inlm} c^{JL}_{in'l-m} = \sum_m (-1)^m \sum_{jk}^{\text{atoms}} P^{(\alpha,\beta)}_{n_1ji} Y_l^m(\hat{\mathbf{r}}_{ji}) P^{(\alpha,\beta)}_{n_2ki} Y_l^{-m}(\hat{\mathbf{r}}_{ki}). \tag{4.49}$$

The two expressions are almost identical, but for an unessential overall factor and for the appearance of double-vanishing Jacobi polynomials instead of standard ones. We will now show a method to address this difference. The underlying assumption is to assume that the energy is obtained by a *linear* combination of the powerspectrum components, explicitly

$$E_{\text{short}} = \sum_i^{\text{atoms}} \sum_{n_1n_2l} a_{n_1n_2l} p^{JL}_{in_1n_2l}, \tag{4.50}$$

for some coefficients $a_{n_1n_2l}$ and where the first sum runs over all the atoms in the system. This equation can be rewritten as

$$E_{\text{short}} = \sum_i \sum_{jk} v(r_{ji}, r_{ki}, \hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{ki}), \tag{4.51}$$

where, again, the potential v has been defined as

$$\begin{aligned} v(r_{ji}, r_{ki}, \hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{ki}) &:= \sum_{n_1n_2l} a_{n_1n_2l} \sum_m (-1)^m P^{(\alpha,\beta)}_{n_1ji} Y_l^m(\hat{\mathbf{r}}_{ji}) P^{(\alpha,\beta)}_{n_2ki} Y_l^{-m}(\hat{\mathbf{r}}_{ki}) \\ &= \sum_{n_1n_2l} \frac{2l+1}{4\pi} a_{n_1n_2l} P^{(\alpha,\beta)}_{n_1ji} P^{(\alpha,\beta)}_{n_2ki} P_l^{jki}, \end{aligned} \tag{4.52}$$

with the second line obtained by means of the addition theorem for the spherical harmonics. Now that we have defined a potential, we can apply again the constraints[21]

$$\lim_{r_1\to r_{\text{cut}}} v(r_1, r_2, \hat{\mathbf{r}}_1\cdot\hat{\mathbf{r}}_2) = 0 \quad \forall\, \mathbf{r}_2, \qquad \lim_{r_2\to r_{\text{cut}}} v(r_1, r_2, \hat{\mathbf{r}}_1\cdot\hat{\mathbf{r}}_2) = 0 \quad \forall\, \mathbf{r}_1, \tag{4.53}$$

and

$$\lim_{r_1\to 0} v(r_1, r_2, \hat{\mathbf{r}}_1\cdot\hat{\mathbf{r}}_2) = 0 \quad \forall\, \mathbf{r}_2, \qquad \lim_{r_2\to 0} v(r_1, r_2, \hat{\mathbf{r}}_1\cdot\hat{\mathbf{r}}_2) = 0 \quad \forall\, \mathbf{r}_1. \tag{4.54}$$

[21] Note that we are not considering $r_{\min}$ for practical reasons.

These expressions are equivalent to the ones used in deriving the double-vanishing-Jacobi polynomials. Indeed, it can be proven that applying these constraints results in an expansion for the energy in terms of a modified powerspectrum, namely

$$E_{\text{short}} = \sum_i \sum_{n_1 n_2 \geq 2} \sum_l a_{n_1 n_2 l} \overline{p}^{JL}_{i n_1 n_2 l}, \tag{4.55}$$

with

$$\overline{p}^{JL}_{i n_1 n_2 l} := \sum_m (-1)^m \overline{P}^{(\alpha,\beta)}_{n_1 j i} Y_l^m(\hat{\mathbf{r}}_{ji}) \overline{P}^{(\alpha,\beta)}_{n_2 k i} Y_l^{-m}(\hat{\mathbf{r}}_{ki}). \tag{4.56}$$

By comparing this expression with Eq. (4.48), we can see that

$$\overline{p}^{JL}_{i n_1 n_2 l} = \frac{2l+1}{4\pi} C^{(3)}_{i n_1 n_2 l}, \tag{4.57}$$

which establishes a connection between the JL formalism, in its linear-scaling form encoded in $C^{(3)}_{i n_1 n_2 l}$, and density trick. To summarize, this has been achieved by constructing a suitable neighbors density and by introducing a constraining procedure. In particular, the construction of the density ρ_i^{JL} allows to perform an explicit comparison with any other method that is based on the density trick formalism. It is important to mention, however, that the double-vanishing-Jacobi polynomials arise from the constraining procedure applied to a linear model. If another model is used then the constraining procedure could be different, and so the resulting expressions could deviate from the ones of the JL expansion. However, this is consistent with the fact that the double-vanishing-Jacobi polynomial were originally derived from a linear expansion.

We conclude by remarking that the formalism, which is based on the JL-atomic basis and the one in terms of the JL-neighbors density, while being crucial in establish a link with previous methods, will be relegated only to theoretical investigations. In the following, instead, we will explicitly use an internal-coordinate representation (the original JL-expansion of Eq. 4.30) keeping in mind that, if there are too many atoms in cut-off sphere (enough to justify the evaluation of the $2l+1$ spherical harmonics, and the $2l+1$ subsequent contractions over m), the linearised formalism could give a computational advantage.

4.1.3 Four-body (4B) potential

The four-body (4B) potential is expanded in complete analogy to what already done for the 3B case. Indeed the 4B-JL expansion reads

$$v^{(4)}_{Z_jZ_kZ_pZ_i}(r_{ji},r_{ki},r_{pi},s_{jki},s_{kpi},s_{jpi}) = \sum_{\substack{n_1n_2n_3\\l_1l_2l_3}}^{\text{unique}} a^{Z_jZ_kZ_pZ_i}_{\substack{n_1n_2n_3\\l_1l_2l_3}} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1ji} \overline{P}^{(\alpha,\beta)}_{n_2ki} \overline{P}^{(\alpha,\beta)}_{n_3pi} P^{jki}_{l_1} P^{jpi}_{l_2} P^{kpi}_{l_3} \right). \tag{4.58}$$

The range of all Jacobi indexes is always $[2, n_{\max}]$, and the one for the Legendre indexes is $[0, l_{\max}]$. It is worth noticing that the symmetry scheme introduced here, with the sum over the unique coefficients, is different from a lexicographic order. Indeed, permuting two atomic species in the definition of the potential, implies the swap of the corresponding distances and scalar products (angles). For example, the potential defined by the swap $Z_i \to Z_j$ satisfies the symmetry

$$v^{(4)}_{Z_kZ_jZ_pZ_i}(r_{ki},r_{ji},r_{pi},s_{jki},s_{jpi},s_{kpi}) := v^{(4)}_{Z_jZ_kZ_pZ_i}(r_{ji},r_{ki},r_{pi},s_{jki},s_{kpi},s_{jpi}). \tag{4.59}$$

On the one hand this means that we can always impose an ordering on the atomic number, as $Z_i \geq Z_j \geq Z_k$ and, on the other, this also implies the following symmetry relation for the expansion coefficients

$$a^{Z_kZ_jZ_pZ_i}_{\substack{n_2n_1n_3\\l_1l_3l_2}} = a^{Z_jZ_kZ_pZ_i}_{\substack{n_1n_2n_3\\l_1l_2l_3}}, \tag{4.60}$$

obtained by attributing the correct Jacobi and Legendre indexes to distances and angles, respectively. In particular, if the atoms in the atomic environment belong to the same species Z, we have the full chain of symmetries

$$a_{\substack{n_1n_2n_3\\l_1l_2l_3}} = a_{\substack{n_2n_1n_3\\l_1l_3l_2}} = a_{\substack{n_3n_2n_1\\l_3l_2l_1}} = a_{\substack{n_1n_3n_2\\l_2l_1l_3}} = a_{\substack{n_2n_3n_1\\l_3l_1l_2}} = a_{\substack{n_3n_1n_2\\l_2l_3l_1}} , \tag{4.61}$$

which must be taken into account in Eq. (4.58), to enforce the correct invariance under identical atoms permutations.

Linear scaling and differences with respect to the bispectrum coupling scheme By following the same procedure of the one shown for the 3B case, we can cast the 4B atomic energy, $\varepsilon^{(4)}$, in terms of the JL-atomic basis as

$$\varepsilon^{(4)}_i = \sum_{Z_1\geq Z_2\geq Z_3} \sum_{\substack{n_1n_2n_3\\l_1l_2l_3}}^{\text{unique}} b^{Z_1Z_2Z_3Z_i}_{\substack{n_1n_2n_3\\l_1l_2l_3}} \left[C^{(4),Z_1Z_2Z_3}_{i,\ \substack{n_1n_2n_3\\l_1l_2l_3}} - S^{(4),Z_1Z_2Z_3}_{i,\ \substack{n_1n_2n_3\\l_1l_2l_3}} \right], \tag{4.62}$$

where

$$C^{(4),Z_1Z_2Z_3}_{i,\ \substack{n_1n_2n_3\\ l_1l_2l_3}} = \Pi_{l_1l_2l_3} \sum_{m_1m_2m_3} (-1)^{\sum_m} (J_1L_2)^{i,Z_1}_{n_1l_1m_1l_2-m_2} (J_1L_2)^{i,Z_2}_{n_2l_3m_3l_1-m_1} (J_1L_2)^{i,Z_3}_{n_3l_2m_2l_3-m_3} \,. \tag{4.63}$$

Here, we used the shorthand $(-1)^{\sum_m} := (-1)^{m_1+m_2+m_3}$, and $\Pi_{l_1l_2l_3} := \prod_{\nu=1}^{3} 4\pi/(2l_\nu + 1)$. Again, the $b^{Z_1Z_2Z_3Z_i}_{\substack{n_1n_2n_3\\ l_1l_2l_3}}$ coefficients are proportional to the initial ones. The expression for the self-energy, $S^{(4)}_i$, is more involved and can be found in the SI of the original work from Ref. [99]. However, we mention that it is expressed only in terms of the atomic basis (J_1L_2), (J_2L_2) and (J_3L_0), so that the full expression is again linear with respect to the number of neighbors inside the cut-off sphere. Crucially, this expression is obtained by a simple re-arrangement of the terms of Eq. (4.58) and, as such, one can freely navigate between an expression in terms of internal coordinates and the one in terms of the JL-atomic basis. In particular, it is important to remark how the coupling scheme of angular momenta implied in the coupling term $C^{(4),Z_1Z_2Z_3}_{i,\ \substack{n_1n_2n_3\\ l_1l_2l_3}}$ is not equivalent to the one used in the definition of the bispectrum components[22],

$$B^{(4),Z_1Z_2Z_3}_{i,\ \substack{n_1n_2n_3\\ l_1l_2l_3}} = \sum_{m_1m_2m_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} (J_1L_1)^{i,Z_1}_{in_1l_1m_1} (J_1L_1)^{i,Z_2}_{in_2l_2m_2} (J_1L_1)^{i,Z_3}_{in_3l_3m_3}, \tag{4.64}$$

reported here for readability[23], and in the form proposed in the ACE model (see Eq. (2.59)). Indeed, the bispectrum coupling cannot be straightforwardly re-arranged in an equivalent expression in terms of internal coordinates. On the contrary, while the bispectrum coupling scheme is more compact, the one introduced here has the advantage that it fully mirrors a representation in terms of internal coordinates, which can be used at need (in the following we will use only the representation in terms of internal coordinates, and not the linearised one).

4.1.4 The Five-body (5B) potential

Proceeding to higher-body order terms, we investigate here how to construct the 5B-JL expansion. The main aim of this section is to show how the internal coordinate representation contains the minimal number of distances and angles that are needed to describe an isometrically-invariant-5B quantity. We remark that investigating a 5B potential is

[22]Please note that the internal coordinate representation is implicitly invariant under reflection. This implies that the two couplings span the same space only when the only bispectrum components retained are the ones for $l_1 + l_2 + l_3$ even.

[23]Note that this coupling can be obtained with the same procedure introduced in the previous section, starting from the JL-neighbor density of Eq. (4.41), ρ^{JL}_i. This approach indeed leads to the bispectrum coupling in terms of the JL-atomic basis.

necessary, since it is proven that a 4B representation is not complete, i.e., that two multi-atom configurations, albeit not related by an O(3) transformation, can lead to the same bispectrum components (please, see Ref. [49] for details). Before proceeding it is important to distinguish between the two notion of completeness that appear in this work. The first one, used throughout this thesis, refers to a complete representation of a function in terms of expansions over its degrees of freedom. This allows to define a basis and, being it the notion of completeness that refers to (invariant) functions evaluated on n bodies, is a definition concerned with the counting and characterization of degrees of freedom. The second notion of completeness is related to the unique representation of a neighborhood containing N atoms. For example, in the case of the the multi-body JL-expansion, our approach is to represent the contribution of the neighborhood in terms of progressively higher body-terms. Thus, the first concept of completeness relates to ways of representing the single n-body term, e.g., a 3B term can be uniquely represented by 2 distances and one angle. In contrast, the second form of completeness tries to answer the question on how to uniquely represent the neighborhood, and more generally, the entire system. If, on the one hand, the two notions are clearly related and indeed become equivalent for the same number of atoms, i.e., $n = N$, on the other hand they are very distinct: while a 3-body term can be described by 2 distances and one angle, a general neighborhood cannot be uniquely defined by a sum of 3-body contributions. Indeed the work of Ref. [49] proves that even a sum of 4-body contributions is not enough, and more recent works go in the direction of constructing a complete representation (see, for example, the recent work of Ref. [104]). While we mentioned the completeness with respect to the representation of the full neighborhood of atoms, from this point forward we will consider only the first kind of completeness and, in particular, this section will discuss an elementary hand-waving method to prove that an isometrically-invariant-5B term can be uniquely described by 4 distances and 6 scalar products.

We can now proceed with an expansion for the 5B terms and with a discussion over the completeness of the JL representation. The 5B-JL expansion is defined as

$$
\begin{aligned}
&v^{(5)}_{Z_jZ_kZ_pZ_qZ_i}(r_{ji}, r_{ki}, r_{pi}, r_{qi}, s_{jki}, s_{jpi}, s_{jqi}, s_{kpi}, s_{kqi}, s_{pqi}) = \\
&\quad = \sum_{\substack{n_1n_2n_3n_4\\ l_1l_2l_3l_4l_5l_6}}^{\text{unique}} a^{Z_jZ_kZ_pZ_qZ_i}_{\substack{n_1n_2n_3n_4\\ l_1l_2l_3l_4l_5l_6}} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1ji} \overline{P}^{(\alpha,\beta)}_{n_2ki} \overline{P}^{(\alpha,\beta)}_{n_3pi} \overline{P}^{(\alpha,\beta)}_{n_4qi} P^{jki}_{l_1} P^{jpi}_{l_2} P^{jqi}_{l_3} P^{kpi}_{l_4} P^{kqi}_{l_5} P^{pqi}_{l_6} \right),
\end{aligned}
\tag{4.65}
$$

and, as mentioned, is expanded over 4 distances, each carrying its own set of double-vanishing-Jacobi polynomials and 6 scalar products, addressed by the Legendre polynomials. An expression that scales linearly with the number of atoms is available also for the 5B case, but it will not be explicitly discussed here. Instead we will postpone a mention

to the linearised expression to the last chapter of the thesis, where we will compared the implied coupling scheme with the one introduced by the ACE model.

We can now focus on the angular part of the potential[24]. To simplify the notation, let us orderly relabel the versors as, $\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3$ and $\hat{\mathbf{r}}_4$. Since the potentials are isometrically invariant, we can always rotate the frame of reference to align the versor $\hat{\mathbf{r}}_1$ with z-axis. We can then perform another rotation around the z-axis (leaving the direction of the first versor unchanged) and bring the second versor on the xz plane. Finally, if the third versor has a negative y-coordinate, we can mirror the system with respect to the xz-plane, to impose a positive y-coordinate. Thus, the coordinates of the versors, in terms of polar and azimuthal angles, can be generally written as[25]

$$\hat{\mathbf{r}}_1 = (0,0), \quad \hat{\mathbf{r}}_2 = (\theta_2, 0), \quad \hat{\mathbf{r}}_3 = (\theta_3, \phi_3), \quad \text{and } \hat{\mathbf{r}}_4 = (\theta_4, \phi_4). \tag{4.66}$$

The three polar angles, as well as the angle ϕ_3, are all defined in the range $[0, \pi]$ (because the y coordinate of $\hat{\mathbf{r}}_3$ is positive), while the last azimuthal angle is defined in the range $[0, 2\pi]$. Crucially, this is the minimal set of angles needed to describe an isometrically invariant quantity defined on 4 versors.
Moving to a representation in terms of scalar products, we notice how the polar angles are unambiguously defined by

$$\hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_n = \cos\theta_n, \quad \text{for } n = 2, 3, 4. \tag{4.67}$$

Moreover, the azimuthal angle ϕ_3 is determined by inverting the expression

$$\hat{\mathbf{r}}_2 \cdot \hat{\mathbf{r}}_3 = \cos\theta_2 \cos\theta_3 + \sin\theta_2 \sin\theta_3 \cos\phi_3, \tag{4.68}$$

which allows us to obtain $\cos\phi_3$. This is enough to fully characterize ϕ_3 since it is defined on the range $[0, \pi]$. Please, note that all the angles considered so far are fully determined by the scalar products $\hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_n$.

A slightly different approach is required for the azimuthal angle ϕ_4, since its cosine does not carry enough information to unambiguously determine its value. Indeed we need to invert the system of equations

$$\begin{cases} \hat{\mathbf{r}}_2 \cdot \hat{\mathbf{r}}_4 = \cos\theta_2 \cos\theta_4 + \sin\theta_2 \sin\theta_4 \cos\phi_4, \\ \hat{\mathbf{r}}_3 \cdot \hat{\mathbf{r}}_4 = \cos\theta_3 \cos\theta_4 + \sin\theta_3 \sin\theta_4 (\cos\phi_3 \cos\phi_4 + \sin\phi_3 \sin\phi_4). \end{cases} \tag{4.69}$$

By inverting the first equation, we obtain the value of $\cos\phi_4$, which can then exploited

[24] Please note that, concerning the radial contributions, an expansion over four distances is already a complete representation.

[25] Again, since our targets are isometrically invariant, this is completely general.

in the second equation to retrieve the value of $\sin\phi_4$, and finally unambiguously obtain the angle ϕ_4. We showed two things here: firstly, all the five scalar products $\hat{\mathbf{r}}_i \cdot \hat{\mathbf{r}}_j$ are required to cover the minimal set of degrees of freedom defined in Eq. (4.66). Secondly, that each new versor, with an azimuthal angle defined on the full interval $[0, 2\pi]$, brings three more degrees of freedom in terms of scalar products: one for the polar angle, and two to characterize the azimuthal angle[26]. We then deduce that the representation in internal coordinates becomes over-complete in going to a six-body description: indeed, adding a new versor $\hat{\mathbf{r}}_5$, will bring four more scalar products, while only three are required to fully characterize the pair (θ_5, ϕ_5). This discussion also shows that, if the described quantity is not invariant under reflection, i.e., if ϕ_3 now take values in the range $[0, 2\pi]$, we need to define a new degree of freedom, which carries the information on the sign of $\sin\phi_3$. While we hypothesise that this problem could be solved by enhancing the description, for example introducing a new Legendre expansion in terms of the quantity $\widehat{\mathbf{r}_1 \times \mathbf{r}_2} \cdot \hat{\mathbf{r}}_3$, we will not consider this case here, postponing the discussion to a future analysis.

We conclude this section by noticing that the number of scalar products is strictly related to the number of indexes in the expansion: this brings some issues with the 5B terms presented in the ACE formalism, which carries only four indexes. We will explore this problem in the last chapter, where we will show, not only that at least 5 indexes are necessary to fully characterize an isometrically invariant 5B function, but also that, using 6 indexes allows us to disentangle the representation from the choice of a specific coupling scheme.

4.2 The Jacobi-Legendre Potential (JLP)

We can now define the complete Jacobi-Legendre potentials (JLPs), obtained by plugging together Eqs. (4.18), (4.30), (4.58) in (4.1) and (4.1). We adopt the same formalism introduced for the SNAP potential (see Eq. (2.49)), and symbolically write

$$\varepsilon^{\mathrm{JLP}}_{Z_i}(\mathbf{J}_i) = \varepsilon_{Z_i} + \sum_{v}^{\substack{\text{body}\\ \text{order}}} \mathbf{a}^{(v)}_{Z_i} \cdot \mathbf{J}^{(v)}_i, \tag{4.70}$$

[26]Note that, to gain information on the azimuthal angle, we actually need a scalar product (continuous quantity) and a boolean number (in this case, given by the sign of the sine of the angles). In general, this is enough to obtain a complete representation of an angle defined over the interval $[0, 2\pi]$. However, if on the one hand we are including more information than needed (the sine is a continuous quantity, against a boolean one), this allows to treat all the scalar products on the same footing, and thus it makes it easier to deal with the symmetries of the expression, in particular the invariance under permutation of identical atoms.

where we divided the expression in different body orders, with $\mathbf{a}_{Z_i}^{(v)}$ being the vector containing all the v-body-order-expansion coefficients (which will be learned in the training phase), and $\mathbf{J}_i^{(v)}$ being the vector containing the sums of all the v-body descriptors subtended to the same coefficient. The vector $\mathbf{J}_i$ is constructed by concatenating all of the $\mathbf{J}_i^{(v)}$ ones.

Forces and Stress tensor Given the linearity of the JLP model, we can calculate the forces and the components of the stress tensor[27] by (please, compare with Eq. (2.50))

$$E^{\mathrm{JLP}} = \sum_i^{\mathrm{atoms}} \varepsilon_{Z_i}^{\mathrm{JLP}}(\mathbf{J}_i) \Rightarrow \begin{cases} \mathbf{F}_j = -\nabla_j E = -\sum_v^{\substack{\mathrm{body}\\ \mathrm{order}}} \sum_i^{\mathrm{atoms}} \mathbf{a}_{Z_i}^{(v)} \cdot \dfrac{\partial \mathbf{J}_i^{(v)}}{\partial \mathbf{r}_j}, \\ \mathbf{W} = -\sum_j \mathbf{r}_j \otimes \nabla_j E = -\sum_{i=1} \sum_v^{\substack{\mathrm{body}\\ \mathrm{order}}} \mathbf{a}_{Z_i}^{(v)} \cdot \sum_{j=1} \mathbf{r}_j \otimes \dfrac{\partial \mathbf{J}_i^{(v)}}{\partial \mathbf{r}_j}. \end{cases} \tag{4.71}$$

In particular, from the first equation, we can also define the v-body contribution to the forces as

$$\mathbf{F}_j^{(v)} := -\sum_i^{\mathrm{atoms}} \mathbf{a}_{Z_i}^{(v)} \cdot \frac{\partial \mathbf{J}_i^{(v)}}{\partial \mathbf{r}_j}. \tag{4.72}$$

From the expressions above, it is clear that the only ingredient needed to evaluate the forces is the use of the chain rule with the derivative of the (double-)vanishing-Jacobi and Legendre polynomials, which are given by

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}x}\widetilde{P}_n^{(\alpha,\beta)}(\cos(x)) = \dfrac{\mathrm{d}}{\mathrm{d}x}P_n^{(\alpha,\beta)}(\cos(x)) = -\dfrac{\alpha+\beta+n+1}{2}\sin(x)P_{n-1}^{(\alpha+1,\beta+1)}(\cos(x)), \\ \dfrac{\mathrm{d}}{\mathrm{d}x}P_l(x) = \dfrac{\mathrm{d}}{\mathrm{d}x}P_l^{(0,0)}(x) = \dfrac{l+1}{2}P_{l-1}^{(1,1)}(x)\,, \\ \dfrac{\mathrm{d}}{\mathrm{d}x}\overline{P}_n^{(\alpha,\beta)}(\cos(x)) = \\ \qquad = -\dfrac{\sin(x)}{2}\left((\alpha+\beta+n+1)P_{n-1}^{(\alpha+1,\beta+1)}(\cos(x)) - (\alpha+\beta+2)\dfrac{\widetilde{P}_n^{(\alpha,\beta)}(-1)}{\widetilde{P}_1^{(\alpha,\beta)}(-1)}\right). \end{cases} \tag{4.73}$$

From the first expression we can appreciate how the vanishing-Jacobi polynomials have zero derivatives at the edges of the range of definition, i.e., for $r = 0$ and $r = r_{\mathrm{cut}}$, due to the presence of the sine functions (as can be seen from Fig.s 4.2 and 4.3). We notice that the second derivative does not share this property. However, should it be important to constrain also the second derivatives to go to zero, a possible solution could be to enforce an additional constraint on the expansion. In the second equation above we exploited

[27] Please note that the definition used here for the components of the stress tensor does not contain a normalization by the volume of the cell, and thus it appears in units of energy. If needed, dividing by the cell volume brings the expression in the standard pressure's units.

the fact that the Legendre polynomials are a particular case of Jacobi polynomials (for $\alpha = \beta = 0$), allowing us to evaluate all of the derivatives within the same framework.

This concludes the characterization of the descriptors defined for the JLP. In the next short paragraph we will outline the main aspects of the linear regression for the determination of the expansion coefficient.

Linear regression The learning process (evaluation of the expansion coefficients $\mathbf{a}_{Z_i}^{(v)}$), was driven by the minimisation of the widely used loss function

$$L = \|\mathbf{E} - \mathbf{J}_E\mathbf{a}\|_2^2 + c_F\|\mathbf{F} - \mathbf{J}_F\mathbf{a}\|_2^2 + c_W\|\mathbf{W} - \mathbf{J}_W\mathbf{a}\|_2^2, \tag{4.74}$$

written in terms of the is the 2-norm $\|\cdot\|_2$. Here the vector $\mathbf{E}$ represents all the energies in the training set (usually obtained by *ab-initio* calculations), $\mathbf{a}$ is the vector obtained by concatenating all the coefficients of the expansion, and $\mathbf{J}_E$ is the matrix whose rows contain the set of descriptors for one configuration of the training set. Similarly $\mathbf{F}$ is the vector of all the forces of the dataset, while $\mathbf{J}_F$ is again the matrix of the differentiated descriptors, in which each row refers to a specific configuration of the training set. We will train on each force component, so that, if the c-th configuration has N_c atoms, we will have $3N_c$ forces associated with that configuration. Analogously, we will train on each of the six components of the stress-tensor (6 for each configuration): the rows and columns of the matrix $\mathbf{J}_W$, refers to the configuration and to the descriptors for the stress-tensor components, respectively. Finally, c_F and c_W are coupling constants to be optimized. We can slightly modify the expression above and make the coupling constants, c_F and c_W, configuration dependent. Practically, this is equivalent to slightly changing the descriptors' definition. This could be useful, for example, when the configurations in the dataset have a different number of atoms and/or to weight the terms in the loss functions on the same footing. In the same spirit, we will consider per-atom energies (dividing the energies and the descriptors by the number of atoms) and we will divide all the forces by $\sqrt{3N_c}$, to bring energies and forces on the same footing with respect to the loss function.

Please note that other potentials make use of a non-linear-embedding function to target the energies. An example is given by the ACE potential which mirrors the functional form of a Finnis-Sinclair potential, and uses a two-targets approach (please, see Refs. [25, 105], and [106] for a discussion on the properties of the embedding functions). Here, however, we will follow the simplest case of a linear regression. We will investigate different embedding functions in future works.

	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
# of features	10	90	364

Table 4.1: Details of the JLP trained on the carbon dataset. In order to reduce the number of hyperparameters, we fixed α and β to be equal, and $r_{\min} = 0$. The model is relatively compact and comprises 465 (464 plus the intercept) features. This table has been reported from the main manuscript, Ref. [99].

4.3 Application: A JLP for Carbon

To test the performance of the JLPs, we decided to build a potential from the dataset presented in Ref. [50], and used to train the GAP17 potential. This choice was driven by the challenges that it presents. Firstly, the dataset contains a large variety of different carbon phases, from amorphous and surfaces, to crystalline structures (graphene, graphite diamond). Moreover, several phases show an high degree of non-locality (in the sense that they present a relatively large distance for the decaying of the forces between two atoms), which requires a careful tuning of the cut-off radius. Before fitting a JLP, we filtered the dataset, removing all the structures with an absolute maximum force components greater than 30 eV/Å. Moreover, we also removed configurations of carbon dimers that were present in the dataset, there included to improve the quality of two-body contributions for the original GAP17 potential [Ref. [50]], but unessential for our purposes. In total, out of the 4,080 configurations, we removed 37. The remaining 4,043 were split into a training (2,830 structures) and test (1,213 structures) sets.

The fitting procedure was done simultaneously on energies, forces and stress tensor, with the coupling coefficients optimized at $c_F = 0.5$ and $c_W = 0.075$, respectively. The actual fitting procedure was carried out by means of a singular value decomposition. We built a potential up to the 4B terms and, mirroring the locality analysis performed for the GAP17, we fixed the cut-off radius at 3.7 Å for every body-order. The optimised hyperparameters are reported in Table 4.1. We stress that the resulting model, with its 465 total features, is arguably very compact if compared to the variety of phases in the dataset. Also, we used the formula introduced and discussed in Ref. [107] (please see Eq. (25) therein) for the stress tensor, in place of the one in Eq. (4.71).

The accuracy of the fitted model on the training set, reported in terms of the Root Mean Squared Errors (RMSEs), was evaluated at 43.9 meV/atom for the energy, 0.781 eV/Å

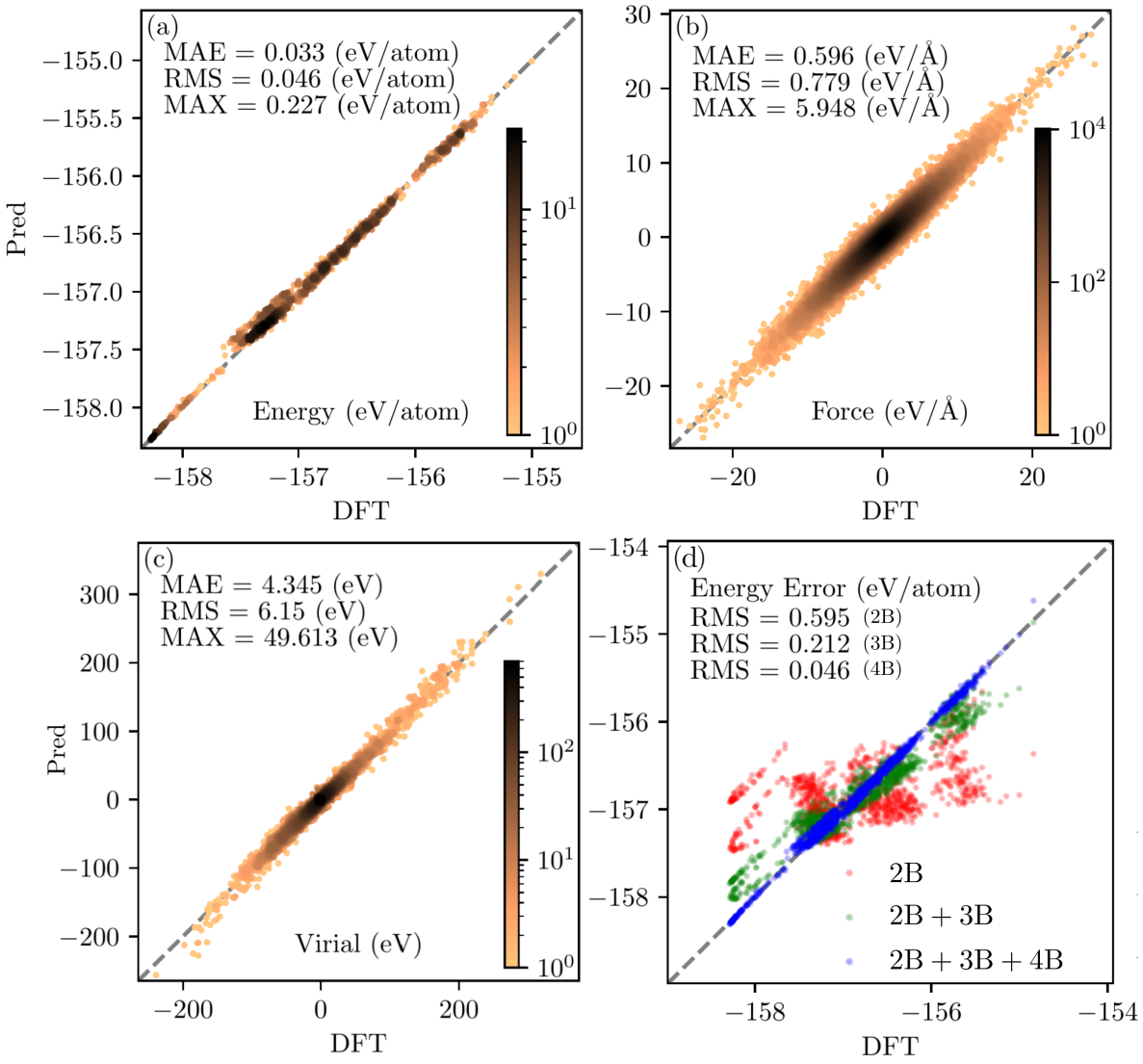

Figure 4.6: Parity plots obtained from the predictions on the test set for the (a) energies, (b) forces, (c) virial stress. The Mean Absolute Errors (MAEs) and Root Mean Square Error (RMSE) are reported for each plot, alongside the error on the worst prediction. The color code indicates the data density (number of datapoints). Figure (d) shows the prediction when a truncated potential is used, going from only 2B terms (in red), including also the 3B ones (in green), and up to the full 4B potential (in blue).

for the forces and 6.62 eV for the stress tensor. We found a similar accuracy for the predictions on the test set, evaluated at 46.6 meV/atom for the energy, 0.779 eV/Å for the forces and 6.15 eV for the stress tensor. The parity plots for the predictions on the test set is reported in Fig. 4.6(a)-(c). From the same figure we can appreciate that, despite the relatively-compact model used here, both forces and tensor components reach an high level of accuracy when compared with similar potentials [50, 58, 68].

It is also interesting to investigate how different body orders contribute to the total fit

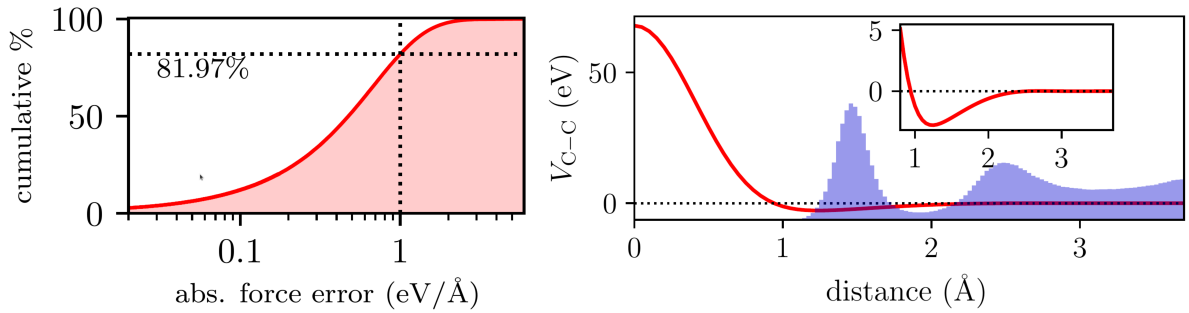

Figure 4.7: (Left) Cumulative distributions of the forces components. Following the the same discussion of Ref. [50], we selected the 1 eV/Å as a reference point. (Right) Pair-wise potential obtained from the JLP and Eq. (4.21). The insert shows a magnification of the curve around the minimum, while the background histogram is the distribution of the pair-wise distances of the full dataset.

of the JLP. This is explored in Fig. 4.6(d), where we show the parity plot (on the test set) obtained by truncating the fit to include only up to the 2B, 3B, or the full 4B potential, respectively[28]. From the figure we can see how it is necessary to include 4B contributions to interpolate all the different phases of the dataset, in accord with what was already concluded in Ref. [50] for the GAP17 potential. Interestingly, the 4B term allows for an interpolation among the lowest energy phases (graphene, graphite and diamond), which are otherwise separated in "branches". We also observe that the inclusion of the 4B term is probably necessary to discriminate among different atomic-local environments of the amorphous phases.

In order to further compare this JLP with the GAP17 potential, a plot of the cumulative distribution of the errors on the force components is shown in Fig. 4.7 (Left). A point on the curve represents the percentage of components which are predicted with an absolute error that is less than the indicated value. In particular, we observe around 81.97% of the forces with an error which is less than the reference value of 1 eV/Å: this can be compared with the similar evaluation performed in the GAP17, where the same reference value corresponded to 68.3% of force components.

Finally, we investigate the behaviour of the 2B potential, by plugging the fitted coefficients in Eq. (4.21) and plotting the resulting functional form, shown in Fig. 4.7 (Right). In the Figure, the 2B potential is compared with the histogram of pair-wise distances of the entire dataset. We can appreciate the natural emergence of a strong short-range repulsive behaviour, despite the lack of highly compressed structures. We remark that the 2B repulsion dictates the behaviour, at small distances, of the entire potential, because of the constraints imposed on higher-body terms. The potential also shows a minimum in close proximity to the first peak of the pair-distances distributions, as expected.

[28]This is only a truncation, the potential is not re-trained on lower order cases.

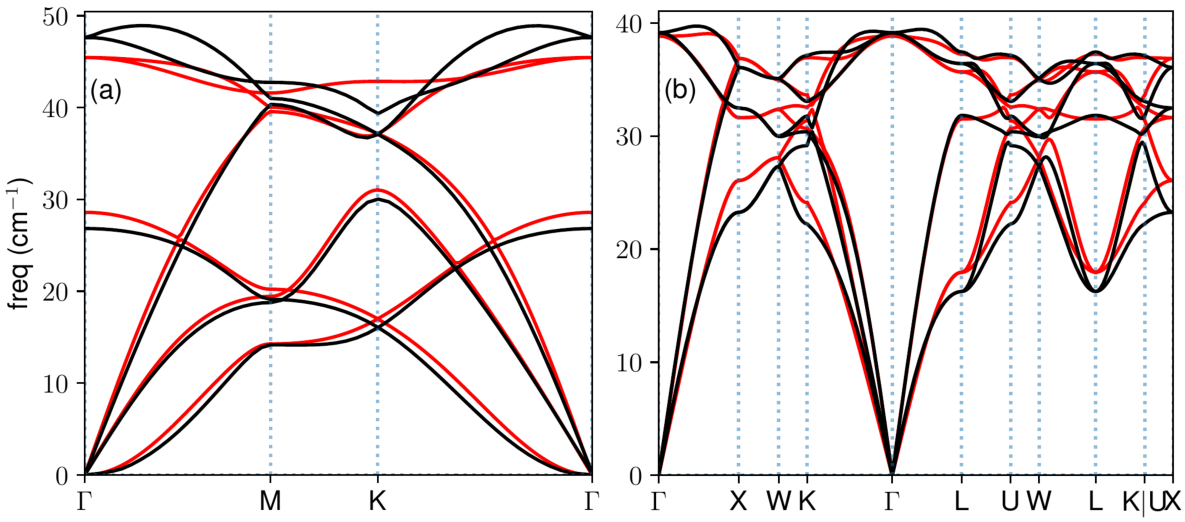

Figure 4.8: The phonon dispersion curves for graphene (left) and diamond (right) are shown. The red curve has been predicted with the JLP, while the black curve is from reference DFT calculations (details in the main text).

Phonon dispersion curves Given the relative high accuracy achieved for energies, forces and stress, we decided to challenge the JLP in the prediction of the phonons dispersion curves for graphene and diamond (using the phono3py package [108, 109]). As reference, we used the curves for crystalline diamond (mp-66) obtained from Materials Project [20], and the one for graphene from the phonon website [110], calculated by means of density functional perturbation theory and the ABINIT code [111]. The results are reported in Fig. 4.8 for both graphene (left) and diamond (right), with the JLP predicted curves are in red, and the reference calculations are in black. The figure shows how the JLP is able to reproduce the phonon dispersion curves, with the graphene being the more accurate between the two. We remark that perfect agreement is not expected, since the DFT dataset used in the training phase (from Ref. [50]) was constructed with the CASTEP code [112], and the phonons have been evaluated with finite differences. Instead, the reference was generated with the ABINIT code. Different DFT implementations could play a role in the difference between the two curves. Importantly, we do not find any negative frequencies at the Γ points. We can compare the phonon dispersion curves obtained here with the similar ones reproduced by the GAP17 potential, or the more recent GAP20 (introduced in Ref. [58], please see the SI therein). The GAP20 is a potential constructed with the same formalism of GAP17, but with a larger and more curated dataset, in particular including, among the others, structures from the work on graphene of Ref. [59]. Interestingly however, the potential obtained here, while being trained on the GAP17 dataset, has a much closer accuracy to the one of the GAP20 potential.

We also briefly discuss the newly developed ACE potential for carbon, discussed in

the recent work of Ref. [68]. This potential is based on a much larger dataset (17,293 structures in the training set) containing a multitude of different phases, ranging from the same ones used to train the JLP, to highly energetic or highly compressed cases. A comparison with our JLP is not trivial, and in particular a comparison of RMSEs could be meaningless: for example, if the whole dataset is considered, the RMSEs for the JLP are much lower than the one of the ACE, but the JLP is not tested on more challenging, highly energetic or highly compressed configurations. On the other hand if some of the more challenging structures are excluded, the RMSE for the energies gets very close, while the ACE has a halved RMSE compared to the one for the JLP for the forces, as can be seen by the more accurate phonon dispersion curves for graphene and diamond. Still, the JLP is trained on a much smaller dataset, and training on richer and broader scenarios would surely improve the accuracy of the potential. Indeed, in future works, we plan to investigate the performance of a JLP trained on a richer dataset to be able to do a more fair comparison and to test the limits of this potential.

4.4 Conclusions

In conclusion, we presented the Jacobi-Legendre formalism for the construction of an MLP. We showed how an approach based on the cluster expansion and on a representation in internal coordinates could be used in defining a competitive potentials, with no reference to any coupling scheme of angular momenta. We used Jacobi polynomials and Legendre polynomials to expand the atomic potential in terms of polynomials formulated in terms of distances and angles, respectively. In particular, the choice of the Jacobi polynomials allowed for a great flexibility in the optimization of the radial basis, with a minimal set of initial assumptions. A crucial strategy introduced in this work is the procedure used in constraining the potentials, which can be used to enforce physically-expected behaviours on the description. In this chapter, we only imposed the short-ranged nature of the potentials, forcing them to smoothly vanish for distances larger than an optimized cut-off radius. We also imposed a 2B interaction at short distances. However, the presented procedure, being completely general, can be adapted to virtually any non-contradictory set of local constraints. As an interesting by-product, we also showed the natural emergence of the widely used cosine-cut-off function and, by leveraging on the choice of Jacobi polynomials, we were able to bias the potential to exhibit a repulsive behaviour at short distances. Finally, we showed how to implement symmetries based on the atomic species of the atoms in the systems, and how to achieve linear scaling with the number of atoms within the cut-off volume, in particular, by exploiting the decomposition of Legendre polynomials in sums of products of spherical harmonics. The same strategy led to a formal comparison with other descriptors, such as the powerspectrum and the bispectrum.

We then tested the JLP against the challenges posed by the carbon dataset used to train the GAP17 potentials. The resulting model was capable of achieving highly accurate RMSEs for energies, forces and components of the stress tensor at once, despite its relative compactness (comprising 465 features), and the variety of phases described in the dataset. Finally we reproduced the zero-temperature phonon dispersion curves for graphene and diamond, obtaining satisfactory qualitative agreement.

This chapter represents the first step in the formulation of the complete JL framework, which will be undertaken in the rest of the thesis. Indeed, starting from the next chapter, we will show how the formalism introduced here can be naturally extended from the prediction of scalar quantities (such as the energy) to also scalar field. One of the main strengths of the JL expansion will be its versatility in naturally adapting to the most disparate targets, while keeping the core of the formalism essentially unchanged.

Chapter 5

The JLCDM for the Electron Density

This chapter will be devoted to the introduction of the Jacobi-Legendre charge density model (JLCDM) from our recent published work *Linear Jacobi-Legendre expansion of the charge density for machine learning-accelerated electronic structure calculations* [113]. This work aimed to build a model capable of predicting the converged real-space electronic-density, at a fraction of the DFT computational cost. This was achieved by a scalar-field-adapted-JL approach, obtained from a cluster-expansion of a grid-point representation of the electronic density.

In this spirit, the first aim of the chapter is to construct a model, which is able to accurately predict the electronic density, to reduce the computational overhead of DFT calculations. The second goal consists in taking the first step towards the generalization of the JL formalism, in the direction of defining the general framework that is at the core of this thesis. Indeed, we will show how the JL descriptors can be naturally generalized to go beyond the description of scalars and, together with the following chapter, we will show that the JL descriptors are, indeed, able to encompass all the quantities of interest in the acceleration of the investigation of materials.

The chapter is structured as follows: firstly, the methods will be presented, alongside the main strategies and ideas, which constitute the ground for the generalization of the JL formalism. In particular, we will show that the model possesses the correct transformation symmetries of a scalar field. Then, the full Jacobi-Legendre charge-density model (JLCDM) will be introduced, with particular care toward the preservation of the representation continuity at small distances. We will then proceed to discuss the adopted strategy for the grid-points sample strategy in the construction of the training set: indeed, we will show how a Gaussian sampling, with respect to the inverse of the magnitude of the density, allows us to remove unnecessary redundancies, while retaining the information needed to

high-quality reconstructions of the electronic density.

The accuracy of the model will then be tested on four systems of interest: benzene, aluminium, molybdenum and 2D MoS_2. In particular, we will show that the trained JLCDM is able to predict the density of phases that were not present in the training set. Finally, we will compare the energies and forces obtained from a non-self consistent approach based on the JLCDM-predicted electronic density. From a comparison against the fully converged ones, we will show that the JLCDM can approximately halve the number of steps required to obtain *ab-initio*-quality performances.

My role in the work has focused on how to adapt the Jacobi-Legendre formalism, introduced in the previous chapter to a scalar field and, specifically, to the description of the DFT electronic density.

5.1 Methods

The main aim for this chapter is the construction of a *linear* model for the electron density, $n(\mathbf{r})$, based on the JL formalism. The operative assumptions are the same of the ones already introduced in section 2.3: the first implicit assumption is that the density can be defined uniquely by the positions (and atomic species) of all the atoms in the system $\{\mathbf{r}_j\}$, i.e.,

$$n(\mathbf{r}) \equiv n(\mathbf{r}; \{\mathbf{r}_j\}). \tag{5.1}$$

In order to be able to practically deal with the position vector $\mathbf{r}$, we build a real-space-grid mesh $\{\mathbf{r}_g\}$, covering the real space $\mathbb{R}^3$. This is the same step taken in the works of Refs. [83] and [84]. The last assumption is that the value of the density at a grid point $\mathbf{r}_g$, depends only on the atoms inside a cut-off sphere of radius r_{cut}. This hypothesis, based on a *nearsightedness* principle [see Refs. [114, 115]], will be formally equivalent to the introduction of the cut-off radius for the scalar descriptors: in this way, we can identify a local environment that determines the value of the density at the point, $\mathbf{r}_g$. Finally, following the same recipe of the JLP method, we assume that we can write the electron density in the form of a cluster expansion as,

$$n(\mathbf{r}_g) = \sum_i n^{(1)}_{Z_i}(\mathbf{r}_g; \mathbf{r}_i) + \sum_{(j,i)} n^{(2)}_{Z_i Z_j}(\mathbf{r}_g; \mathbf{r}_i, \mathbf{r}_j) + \sum_{(i,j,k)} n^{(3)}_{Z_i Z_j Z_k}(\mathbf{r}_g; \mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \dots, \tag{5.2}$$

where the sums run, respectively, over single atoms, pairs and triplets inside a cut-off sphere centered on the grid point $\mathbf{r}_g$. As for the JLP, we divided the electron density in different body-order terms. In this sense, all the one-body (1B) contributions, $n^{(1)}(\mathbf{r}_g; \mathbf{r}_i)$, depend on one atom in the neighborhood (and on the grid point), the two-body (2B) order terms, $n^{(2)}(\mathbf{r}_g; \mathbf{r}_i, \mathbf{r}_j)$, depend on the position of two atoms, and so on for increasingly

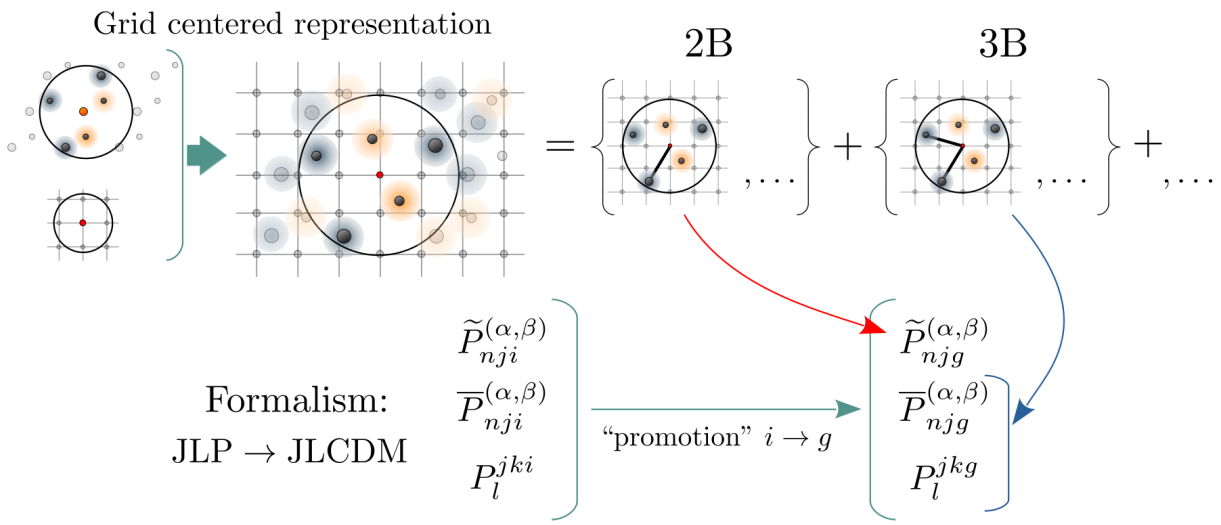

Figure 5.1: Graphical abstract of the core ideas for the construction of JLCDM models. We used the same grid formulation already introduced in Refs. [83, 84] [see Fig. 2.8], where, by centering the representation on a grid point $\mathbf{r}_g$, an atom-centered representation is adapted to a mesh of grid points that covers the whole space. The grid-centered representation is then expanded in a cluster expansion, similarly to what was done for the JLP. Finally, each distance contribution is expanded in (double-)vanishing-Jacobi polynomials, while the angular contributions are expanded in Legendre polynomials. The final formulation is formally identical to the one introduced for the JLP in 4.70, where we simply "promoted" the central atom to be a grid point.

higher-body orders. We also assumed that the functional form of the various terms depends only on the atomic species of the atoms involved, similarly to the analogous hypothesis for the potentials of the Jacobi-Legendre expansion.

Now, the strongest link with the JLP is that, if the frame of reference is translated on top of the grid point $\mathbf{r}_g$, then the value of the density is invariant under any rotation of the local environment around the origin. In this frame of reference, we have that all the terms in the sum above must be scalars, and so that they must depend only on distances and scalar products. Explicitly

$$n(\mathbf{r}_g) = \sum_i n^{(1)}_{Z_i}(r_{ig}) + \sum_{(j,i)} n^{(2)}_{Z_iZ_j}(r_{ig}, r_{jg}, s_{ijg}) + \sum_{(i,j,k)} n^{(3)}_{Z_iZ_jZ_k}(r_{ig}, r_{jg}, r_{kg}, s_{ijg}, s_{ikg}, s_{jkg}) + \dots, \tag{5.3}$$

which is formally equivalent to the cluster expansion introduced in (4.2), in which the grid point is used as a centre of the local environment, in analogy to the role of the central atom in the JLP's formalism. Note that r_{ig} is the distance between the grid point position and the i-th atom, and the relative scalar products are defined as $s_{ijg} = \hat{\mathbf{r}}_{ig} \cdot \hat{\mathbf{r}}_{jg}$. A graphical abstract for this construction is shown in Fig. 5.1. Crucially, we notice that this cluster expansion is consistent with the same constraints introduced in the Jacobi expansions of

the JLP, since each addend in the sum above is required to vanish when at least one of its distances approaches the cut-off radius. This will justify the use of the vanishing- and double-vanishing-Jacobi polynomials when we will further expand the body terms following the JL expansion recipe.

Transformation properties of the density Before introducing the actual expansion, it is important to prove that the above expansion is consistent with the definition of a scalar field. This can be done by demonstrating that the cluster expansion satisfies the transformation rules introduced in Eq. (2.84), and reported here for readability,

$$n_{\hat{T}}(\mathbf{r}_g; \{\mathbf{r}_i\}) = n(\mathbf{r}_g; \{\hat{T}\mathbf{r}_i\}) = n(\hat{T}^{-1}\mathbf{r}_g; \{\mathbf{r}_i\}), \tag{5.4}$$

where $n_{\hat{T}}(\mathbf{r}; \{\mathbf{r}_i\})$ is the density of the rotated system of atoms and where $\hat{T}$ is a general isometric transformation (any operation that leaves distances and angles unchanged, namely translation, rotation or inversion). In order to prove this transformation property, we can leverage the identity

$$\hat{T}\mathbf{r}_i - \mathbf{r}_g = \hat{T}\left(\mathbf{r}_i - \hat{T}^{-1}\mathbf{r}_g\right) , \tag{5.5}$$

which states that we can move the effects of the transformation from the grid point to the atoms, by factorizing out a global transformation. Since lengths are left unchanged by an isometric transformation, we have that

$$\left|\hat{T}\mathbf{r}_i - \mathbf{r}_g\right| = \left|\mathbf{r}_i - \hat{T}^{-1}\mathbf{r}_g\right|, \tag{5.6}$$

which implies that the distances undergo the required transformation. Similarly, because the scalar product between two vectors is not affected by a global isometry, we have

$$\left(\hat{T}\mathbf{r}_i - \mathbf{r}_g\right) \cdot \left(\hat{T}\mathbf{r}_j - \mathbf{r}_g\right) = \left(\mathbf{r}_i - \hat{T}^{-1}\mathbf{r}_g\right) \cdot \left(\mathbf{r}_j - \hat{T}^{-1}\mathbf{r}_g\right) . \tag{5.7}$$

This proves that the system of internal coordinates, written in terms of the atoms and a grid point, is indeed a good representation of the degrees of freedoms of scalar fields.

5.1.1 The Jacobi-Legendre Charge Density Model

We can now expand the body terms of the cluster expansion of Eq. (5.3) by means of the JL formalism (expand the dependence on each of the distances in terms of Jacobi polynomials and on each of the scalar products in terms of Legendre polynomials), and obtain

$$\begin{cases} n^{(1)}_{Z_i}(r_{ig}) = \sum_n a^{Z_i}_n \widetilde{P}^{(\alpha,\beta)}_{nig}, \\ n^{(2)}_{Z_iZ_j}(r_{ig}, r_{jg}, s_{ijg}) = \sum_{n_1n_2l}^{\text{unique}} a^{Z_iZ_j}_{n_1n_2l} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1ig} \overline{P}^{(\alpha,\beta)}_{n_2jg} P^{ijg}_l \right), \\ n^{(3)}_{Z_iZ_jZ_k}(r_{ig}, r_{jg}, r_{kg}, s_{ijg}, s_{ikg}, s_{jkg}) = \sum_{\substack{n_1n_2n_3 \\ l_1l_2l_3}}^{\text{unique}} a^{Z_iZ_jZ_k}_{\substack{n_1n_2n_3 \\ l_1l_2l_3}} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1ig} \overline{P}^{(\alpha,\beta)}_{n_2jg} \overline{P}^{(\alpha,\beta)}_{n_3kg} P^{ijg}_{l_1} P^{ikg}_{l_2} P^{jkg}_{l_3} \right), \\ \dots, \end{cases} \tag{5.8}$$

The sum on the unique indexes is the same as the one introduced in Eq. (4.30), and is necessary to enforce the correct invariance under permutation of identical atoms. We also used the same cosine map introduced for the JLP: indeed, the vanishing-Jacobi polynomials between the j-th atom and the grid point are given by

$$\widetilde{P}^{(\alpha,\beta)}_{nig} = \widetilde{P}^{(\alpha,\beta)}_n \left[\cos\left(\pi \frac{r - r_{\min}}{r_{\text{cut}} - r_{\min}} \right) \right], \tag{5.9}$$

and the Legendre polynomials evaluated on the scalar product between $\hat{\mathbf{r}}_{ig}$ and $\hat{\mathbf{r}}_{jg}$ are given by

$$P^{ijg}_l = P_l(\hat{\mathbf{r}}_{ig} \cdot \hat{\mathbf{r}}_{jg}).$$

Again, the range of indexes for the Jacobi polynomials is $[1, n_{\max}]$ for the 1B case, and $[2, n_{\max}]$ for all the others. Here, the role of the hyperparameter $r_{\min}$ is more important than in the case of the JLP. Indeed, it is not unusual for the grid points to get relatively close to the positions of the atoms. This means that, contrary to what usually happens for MLPs, the short-distances range is not only explored by the dataset, but it also has a significant impact on the performance of the model. Therefore, the constraint that the derivative of the potential vanishes at small distances (enforced by the cosine map, as can be seen in Fig. 4.2), can be detrimental to the quality of the fit. For this reason, using a negative $r_{\min}$ to push the point of zero derivative to the left, namely, to inaccessible-negative values of r, relieves the potential from this constraint, and allows the model to predict steep variations of the density in close proximity to the atomic positions.

Also, contrary to the JLP case, the choice of the double-vanishing-Jacobi polynomials is not uniquely guided by the condition that the value of the density in proximity to

an atomic position should be mainly characterized by a 1B interaction with that atom. Indeed, for reasons that will be shortly discussed, we have to remove $r_{\min}$ from the cosine map of the double-vanishing-Jacobi polynomials. Thus, for this JL expansion, we chose the polynomials

$$\overline{P}_{nig}^{(\alpha,\beta)} := \overline{P}_{n}^{(\alpha,\beta)}\left[\cos\left(\frac{\pi r_{ig}}{r_{\text{cut}}}\right)\right], \tag{5.10}$$

to expand the distance dependence of all the nB-order terms, but the 1B. The effect of the choice of the double-vanishing-Jacobi and of the removal of $r_{\min}$ is twofold. The first consequence can be appreciated by investigating the limiting case in which $i = g$, i.e., the grid point is on top of one of the atoms. By taking explicitly the 2B term, we have

$$n^{(2)}_{Z_iZ_j}(\underbrace{r_{gg}}_{=0}, r_{jg}, \underbrace{s_{gjg}}_{=1}) = \sum_{n_1n_2l}^{\text{unique}} a^{Z_iZ_j}_{n_1n_2l} \sum_{\text{symm.}} \left\{\widetilde{P}^{(\alpha,\beta)}_{n_1}\left[\cos\left(\pi\frac{r_{\min}}{r_{\text{cut}} - r_{\min}}\right)\right]\widetilde{P}^{(\alpha,\beta)}_{n_2jg}\right\}, \tag{5.11}$$

which is, effectively, a 1B term (since depends only on the distance r_{jg}) and, as such, have already been considered in the lower-order-body contributions. By setting $r_{\min} = 0$, and by using the double-vanishing-Jacobi polynomials, one removes these redundancies and the body orders are kept as formally separated as possible. The second consequence consists in preserving the continuity of the expansion. Indeed, if we take the limit process of the grid point approaching the atomic position, we have that, in general,

$$\lim_{r_{ig}\to 0} n^{(2)}_{Z_iZ_j}(r_{ig}, r_{jg}, s_{ijg}) \neq n^{(2)}_{Z_iZ_j}(0, r_{jg}, 1). \tag{5.12}$$

The cause for this inequality is that the direction of a zero-length vector is not well-defined, and so the scalar product representation jumps discontinuously to 1. Again, this discontinuity will be solved if both sides of the equation are constrained to be zero, which is enforced by the simultaneous use of the double-vanishing-Jacobi polynomials and of an $r_{\min} = 0$. Arguably, the question of a zero $r_{\min}$ can also arise for the simpler case of a JLP, where, however, the role of $r_{\min}$ is much less relevant, given that such small distances regions are rarely explored by the dataset.

The hyperparameters of this model are defined in the same way as the one introduced for the JLP, with each cluster, defined by its atomic species, carrying its own set. However, since we removed the $r_{\min}$ but from the 1B case, which does not have an $l_{\max}$, the number of hyperparameters is always five, regardless of the body order[29].

We can then define the Jacobi-Legendre Charge Density Model (JLCDM) by the sym-

[29]We remark that the hyperparameters for the 2B order term are $\alpha, \beta \in (-1, \infty)$, $n_{\max} \in [1, \infty)$, $r_{\text{cut}} \in (0, \infty)$, $r_{\min} \in (-\infty, r_{\text{cut}})$. For higher-body-order terms, $r_{\min}$ is removed, but we have $l_{\max} \in [0, \infty)$.

bolic linear expansion

$$n(\mathbf{r}_g) = \sum_{v}^{\substack{\text{body}\\ \text{order}}} \mathbf{a}^{(v)} \cdot \mathbf{J}^{(v)}, \tag{5.13}$$

where, similarly to what done in the definition of the JLP, Eq. (4.70), we have concatenated all the expansion coefficients of the same body-order in the vectors $\mathbf{a}^{(v)}$, and all the descriptors in the vectors $\mathbf{J}^{(v)}$. Crucially, the linearity of the model allows one to drastically reduce the number of expansion coefficients to fit (usually of the order of the thousands), with respect to Neural-Network (NN) based approaches, which requires the fit of a much larger number of weights (of the order of the millions). Finally, the expansion coefficients $\mathbf{a}^{(v)}$ can be determined by minimising a loss function of choice, typically analogous to the one showed in Eq. (4.74), but containing all the grid points that have been selected to evaluate the density. Indeed, the strategy used for sampling the grid points will be crucial for the performance of the model, as will be discussed in the next section.

5.2 Grid-point sampling strategy

Now that the model has been defined, we can address how to efficiently select training points out of the millions available in the mesh. The usual approach to the mesh is to choose a uniform drawing from a uniform, evenly-spaced, grid (we reviewed a few of the available methods in Sec. 2.3). We will now prove that this strategy is not optimal and carries a lot of redundancies. To make the discussion more concrete, we take an example of benzene molecule from Ref. [116], obtained from sampling molecular dynamics at 300K. A DFT calculation was performed on the sampled geometry, so a density in real space, defined over 5,832,000 grid points, was obtained. The calculations were carried out with the Vienna ab Initio Simulation Package (VASP) [117, 118], (please see the manuscript, Ref. [113], for the details of the settings used). An example of density is shown in Fig. 5.2(a). Crucially, a uniform grid-point mesh is sub-optimal as can be seen from Fig. 5.2(b): indeed, most of the points are drawn from regions in space where the density is negligible, far from any relevant contributions. Our approach, instead, aims to select relevant points with respect to the density value. Indeed, we follow the Gaussian probability density given by

$$P(\mathbf{r}_g) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(1/n(\mathbf{r}_g))^2}{2\sigma^2}\right), \tag{5.14}$$

so that points corresponding to an higher value of the density is selected with an higher probability, than the ones where the density was negligible. Here, the Gaussian broadness σ, is a parameter that can be optimized. An example of this selection is reported in

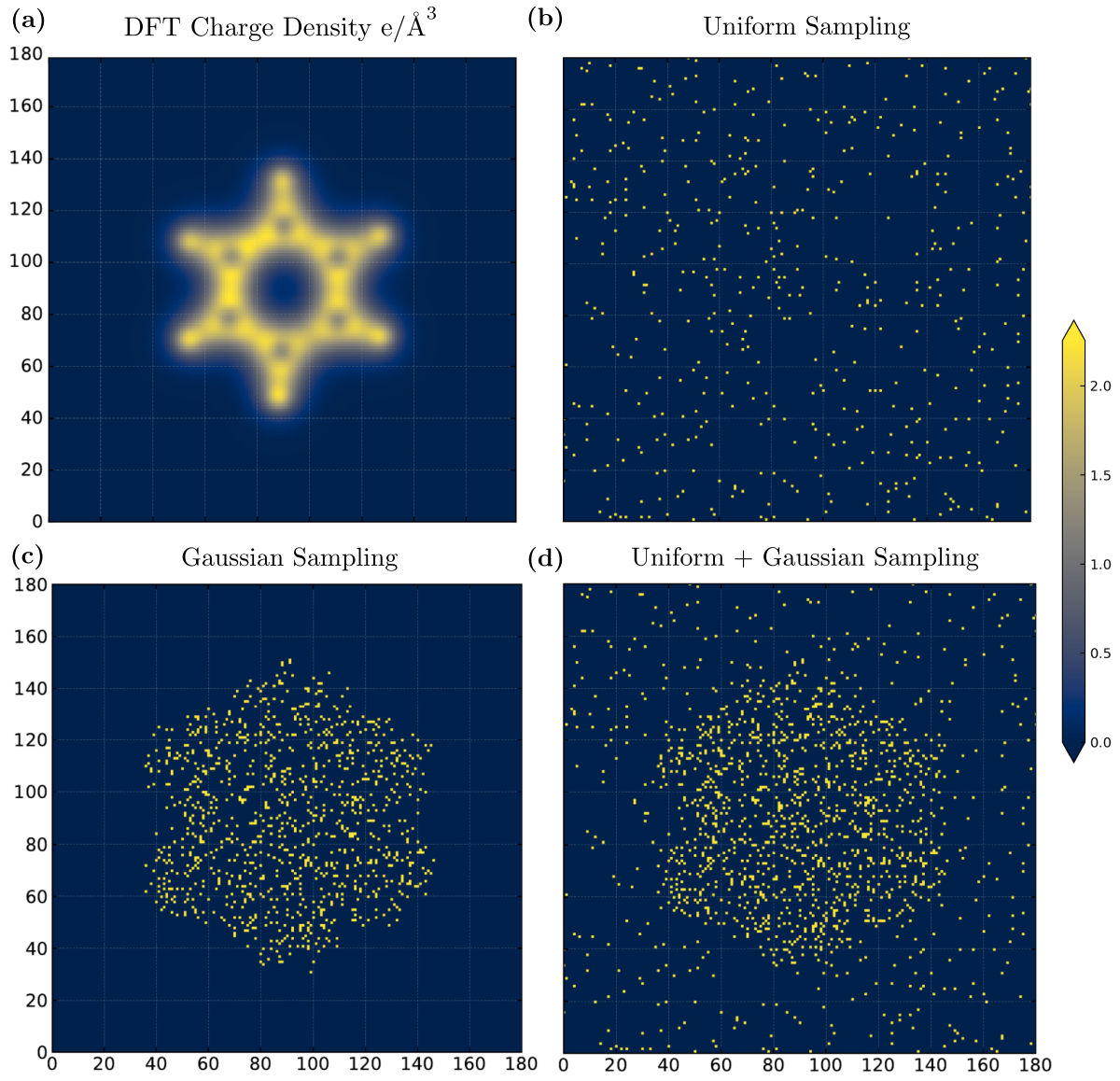

Figure 5.2: (a) Density obtained from a VASP calculation performed on one snapshot from the molecular-dynamics dataset of Ref [116]. (b) An example of uniform sampling performed on the grid. Most of the points are in positions were the value of the density is negligible and, thus, this approach carries a lot of redundancies. (c) Example of drawing from the density distribution of Eq. (5.14). The points are now more relevant, covering positions where the density has significant values. (d) The actual strategy used, an overlap of the uniform density from (b) and the Gaussian drawing from (c), devised to avoid underfitting of regions with small values of the density. Please note that the axis labels refers to the grid-point mesh.

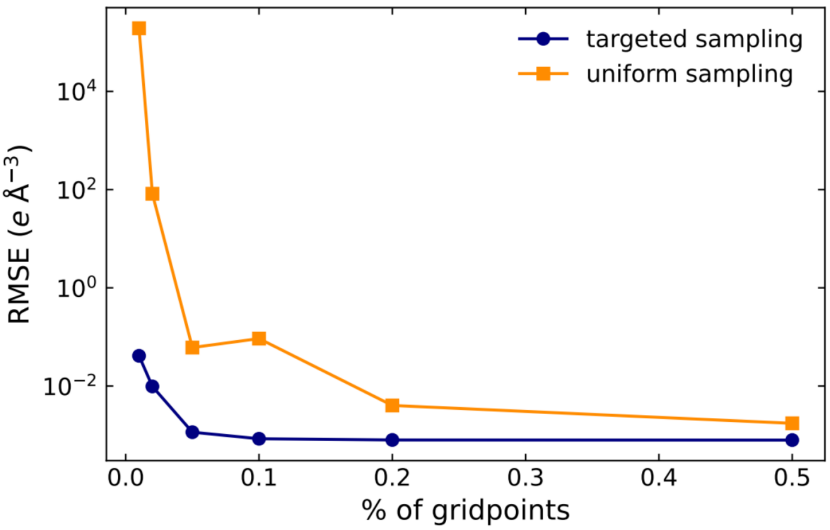

Figure 5.3: The Figure shows the error of the model against the percentage of sampled from the total number of grid points (in the main text). The label "targeted" refers to the "Gaussian + uniform" sampling strategy discussed in Fig. 5.2. As can be appreciated from the logarithmic scale on the y axis, the effect of the Gaussian sampling is a faster convergence of the error, and thus a smaller number of descriptors' calculations.

Fig. 5.2(c). Still, to avoid under-sampling the points with small density values, we finally considered a mixture of the Gaussian sampling with a uniform one, as shown in Fig. 5.2(d). This strategy allowed us to select *only a few thousand points* without compromising the accuracy of the models. To show this, in Fig. 5.3, it is reported the testing error with respect to the percentage of grid point employed in the training of a model for the benzene molecule: it can be seen the targeted-(mixed-)sampling strategy allows to reach convergence in the error already with 0.1% of the total number of grid points. Also, since the descriptors were evaluated for each grid point, we also obtained a significant reduction of the computational overhead.

5.3 Accuracy of the model

In order to probe the accuracy of the model, we performed several tests, on four different systems. Firstly, we investigated the performance on a benzene molecule, from the dataset reported above, of Ref. [116]. We then moved to aluminium and molybdenum, to probe the performance of the model on metallic solids.

The aluminium was chosen to closely compare our model against the ones based on neural network architecture discussed in section 2.3 (from Refs. [29, 83, 84]), Instead, the choice of molybdenum aimed to test the JLCDM on the prediction of less localized electronic densities. Finally, we focused on two dimensional MoS_2, where a model was

	Body	r_{cut}	n_{max}	l_{max}	r_{min}	α	β	σ	# features
Benzene	1B	2.80	27	–	-0.78	7.00	0.00	90 (50%)	1572
	2B	2.80	12	5	–	7.00	0.00		
Al	1B	4.08	15	–	-0.74	7.87	3.62	40 (40%)	120
	2B	4.08	6	6	–	5.87	1.75		
Mo	1B	4.04	20	–	-1.09	4.02	5.46	30 (60%)	812
	2B	4.04	12	11	–	-0.08	2.38		
2D MoS_2	1B	4.76	18	–	-0.93	6.72	6.97	40 (40%)	2346
	2B	4.76	11	10	–	5.07	5.07		

Table 5.1: We show here all the hyperparameters used in training the models. The cut-off radius, r_{cut}, and the distance, r_{min}, are in Å. The width σ determines the width of the sampling distribution of Eq. (5.14), while the value in parenthesis determine the percentage of points sampled from the Gaussian distribution relatively to the total number of selected grid points (the remaining were sampled from the uniform distribution). It can be appreciated how the number of features is always modest (if compared to similar model trained for neural networks), with the aluminium case being the most compact of all.

trained on 1H and 1T phases and then tested on the 1T' phase. This allowed to probe the transferability of the model to unseen phases. All of the optimized hyperparameters, for each one of the systems, are reported in Table 5.1, alongside the width of the Gaussian distribution used to sample the points. Please note that, to simplify the fitting procedure, we constrained all the clusters of the same body-order to have the same hyperparameters. All the models were trained up to the 2B order: while this could bring degeneracy in the representation (in the sense that non-isometric equivalent environments could be mapped in the same set of descriptors, and so in the same value of the density), we were able to reach satisfactory accuracy, while not compromising the compactness of the model. For the details on how the datasets have been constructed, and for the settings used for the *ab-initio* calculations, please see the main manuscript, Ref. [113].

Benzene The model for the benzene molecule used 30 molecular-dynamics snapshots for the training and other 30 for the test set, and was trained over 6,000 grid points. The optimized hyperparameters are reported in Table 5.1, and the resulting vanishing-Jacobi polynomials are shown in Fig. 5.4. As already mentioned, contrary to the case of the JLP where the short-distances range of the polynomials was not explored by the dataset, the grid points cover all the possible distances, from $r = 0$ to r_{cut}. On the one hand, this reduces the chances of over-fitting, and therefore higher n_{max} can be used. On the other hand, this also makes the role for the r_{min} more critical: in particular we allow for r_{min} to be negative, so that the derivative is not constrained to approach zero at $r = 0$. The MAE and the RMSE of the model are, respectively, 2.85×10^{-4} e/Å^{-3} and 1.033×10^{-3}

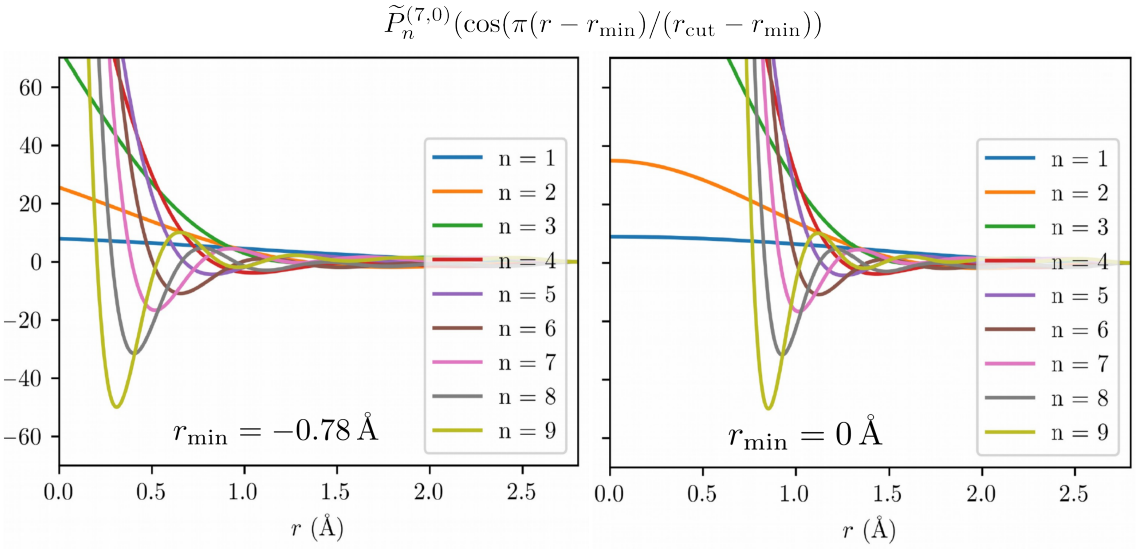

Figure 5.4: The figure reports the first nine vanishing-Jacobi polynomials used in the fit for the benzene models. The cut-off radius is $r_{\text{cut}} = 2.80$ Å. The two figures show the role of $r_{\min}$ in shifting the curve, so that the fitted model is not constrained to be flat in approaching the left edge ($r = 0$) of the interval.

e/Å^3 which, compared to the maximum value of the DFT density, ∼2.6 e/Å^3, show a performance which is quantitatively close to the converged DFT one. This is confirmed by the plots shown in Fig.s 5.5(a)-(b), where the difference between a fully converged DFT density and the predicted one is portrayed. There, we can appreciate how this difference does not present any relevant geometrical feature: we can interpret this fact as a confirmation that a 2B description of the local environment is indeed enough for this system.

Aluminium The training for aluminium involved only 10 training configurations and 10 test configurations. As can be seen from Fig.s 5.5(c)-(d), the de-localisation of the electronic density allows one to have an extremely compact model, with only 120 features. This result can be interpreted by noticing that, for aluminium, the electronic density is less localized around the atomic position, as can be seen from Fig. 5.5(d). Indeed, the magnitude of the electronic density does not vary much, regardless of the grid point explored. For this reason, the model does not have to interpolate between regions with a large value of the density and regions with a small one, and so a smaller number of features is required. The MAE and RMSE for the model are, respectively, 4.81×10^{-4} e/Å^3 and 6.1×10^{-4} e/Å^3, on par with similar results from much larger (in terms of number of weights) neural-networks-based models (please see Refs. [83] and [84]).

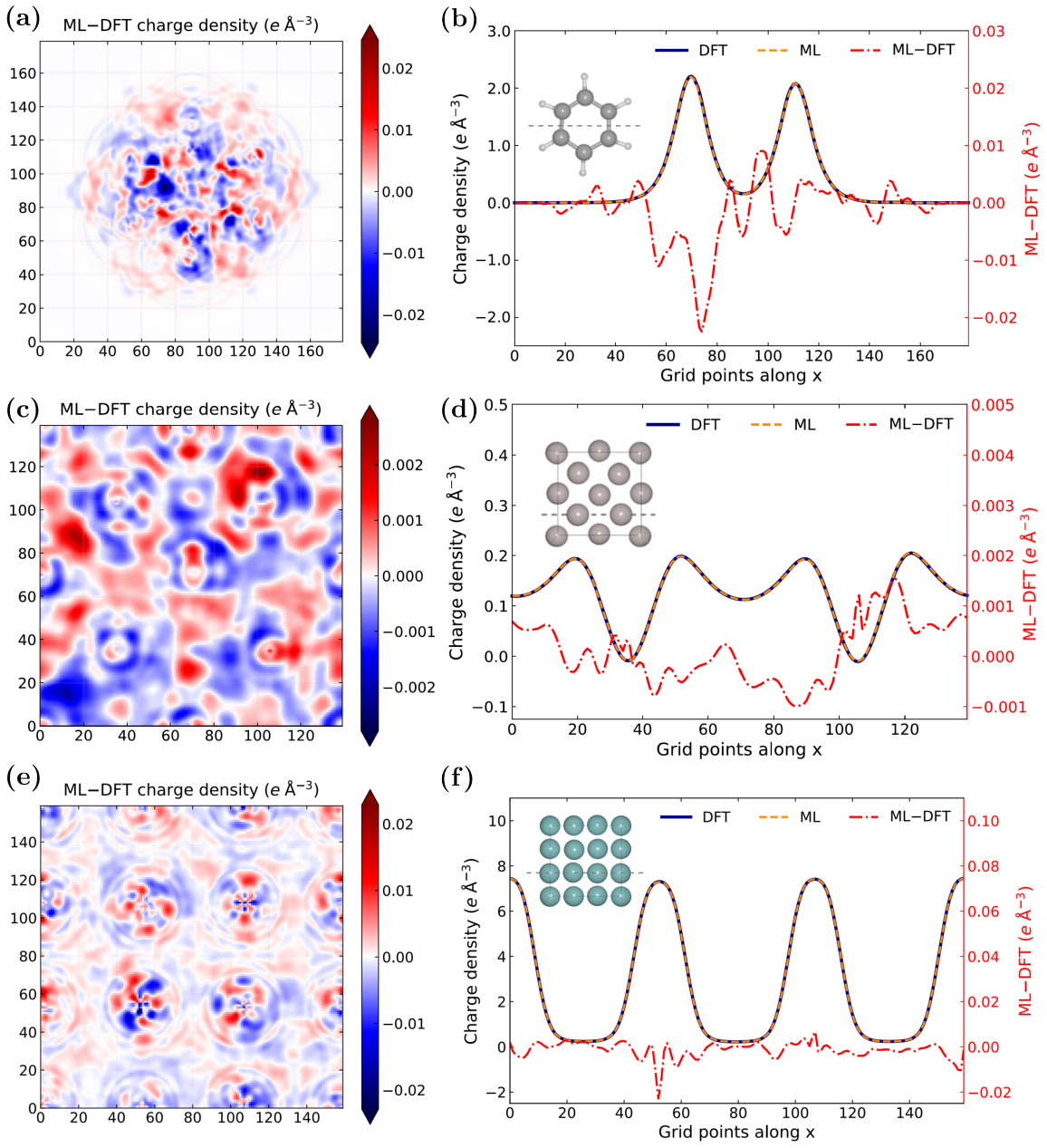

Figure 5.5: (a) The difference between the electronic density of one test-benzene configuration and the predicted one is shown (along the molecular plane). The lack of symmetries in the error distribution is an indication that the degeneracies of a 2B description do not play a significant role for this prediction. (b) The figure shows the true electronic density, the predicted one and their difference, along the cut in the insert. It can be appreciated how the scale of the difference is two orders of magnitude smaller than the actual value of the density. (c)-(d)Analogous plots for the aluminium case. It can be noticed how the density is delocalized, which can be interpreted as a justification for the compactness of the trained model (only 120 features, as shown in Table 5.1). (e)-(f) Analogous plot for molybdenum. Here the errors appear localized in a region around the atoms with geometrical features arising. Arguably, this features could be resolved by an higher body order expansion.

Molybdenum As for the aluminium case, the JLCDM for molybdenum was trained and tested on dataset of 10 configurations each. The accuracy of the model reached a MAE of 1.97×10^{-3} e/Å^3, and a RMSE of 2.82×10^{-3} e/Å^3. From Fig.s 5.5(e)-(f), it can be seen how the local density is much more localized around the atomic position. Indeed, contrary to what happened in the previous two cases, the difference between the converged-DFT density and the predicted one presents some geometrical features, such as a distinct radial distribution and recognizable patterns along the bond directions. Arguably, these features could be addressed and resolved by higher-order body expansions. However, given the high accuracy already reached, we postponed this investigation to future analysis.

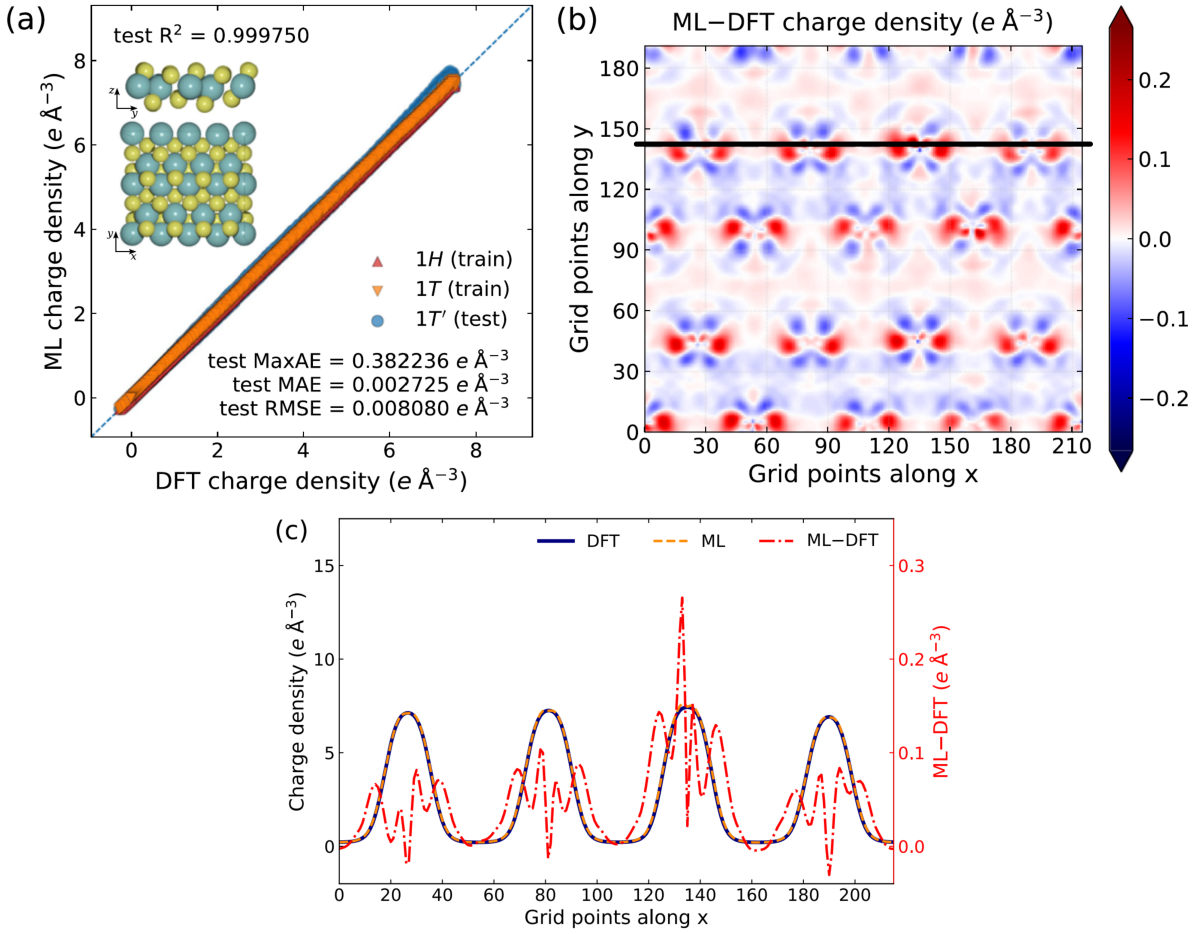

Figure 5.6: (a) Parity plot of the predicted values of the true vs predicted values for the 2D MoS_2 model. As the density gets larger, the prediction on the 1T' phase deviates more from the true value. However, the agreement is still good along all the full range considered. (b) A section of the difference between the predicted density and the true one, taken on the plane of the Mo atoms (xy-plane in this case). (c) Value of the true density against the difference with the predicted one. The section is taken along the solid line in (b). The difference in the two scales still shows a good accuracy in the prediction, despite being the less accurate prediction among the ones showed in this chapter.

2D MoS_2 By investigating a two-dimensional MoS_2 system, we also probed the performance of the model when applied to different phases of the same material. Indeed, we trained a JLCDM on the 1H and 1T phases, with 10 configurations as training set, while the test was performed on as many configurations disposed in the 1T' phase (which is a natural deformation of the 1T phase that occurs under relaxation [119]). We reported the grid-points parity plot in Fig. 5.6(a), alongside with the MAE and the RMSE of 2.725×10^{-3} e/Å^3 and 8.080×10^{-3} e/Å^3, respectively. The parity plot shows systematically more disagreement between the true density and the predicted one on the higher-end of the values range. Indeed, Fig. 5.6(b) shows a section for the difference DFT-predicted densities, where we can clearly recognize a distinguished geometrical pattern in the error. While this could surely be attributed to the different training-testing phases (primarily considering that the 1T' phase is a distortion of the 1T one), an higher-body order is still expected to improve the performance of the fit. Nevertheless, by the quality of the predicted density, we can deduce that the model was able to transfer to the unseen 1T' phase, despite its relative compactness (the model consists in only 2,346 features, as reported in Table 5.1).

Comparison between fully converged energies and forces and the JLCDM predicted ones As a final test, we used the predicted density to obtain energies and forces by means of non-self-consistent (NSC) cycles. Since VASP requires a few cycles for the diagonalization of the Hamiltonian to be efficiently carried out, we saw that we only needed 5 cycles (keeping the density fixed as the predicted one) to converge to accurate values of energies and forces. This is almost half of what required by fully self-consistent (SC) procedures, which converged after 9-12 cycles for all the systems investigated. We report the values of the difference between the NSC values of energies and forces, and the converged ones in Fig. 5.7. As expected, the worst prediction on both energies and forces was found for the 2D MoS_2 system, while the best ones was for the aluminium. In general, however, the accuracy of the energies is, at worst, of the order of meV/atom, which rivals with the accuracy of MLP models. It is important to notice that the accuracy for both aluminium and molybdenum is one order of magnitude higher than of the other two systems. Analogous conclusions can be drawn for the forces.

5.3.1 Comparison with other methods

In this section we outline a qualitative comparison with the methods that have been introduced in Sections 2.3.1 and 2.3.2. We remark that the following will not be a rigorous treating, but a discussion on the main conceptual differences and similarities.

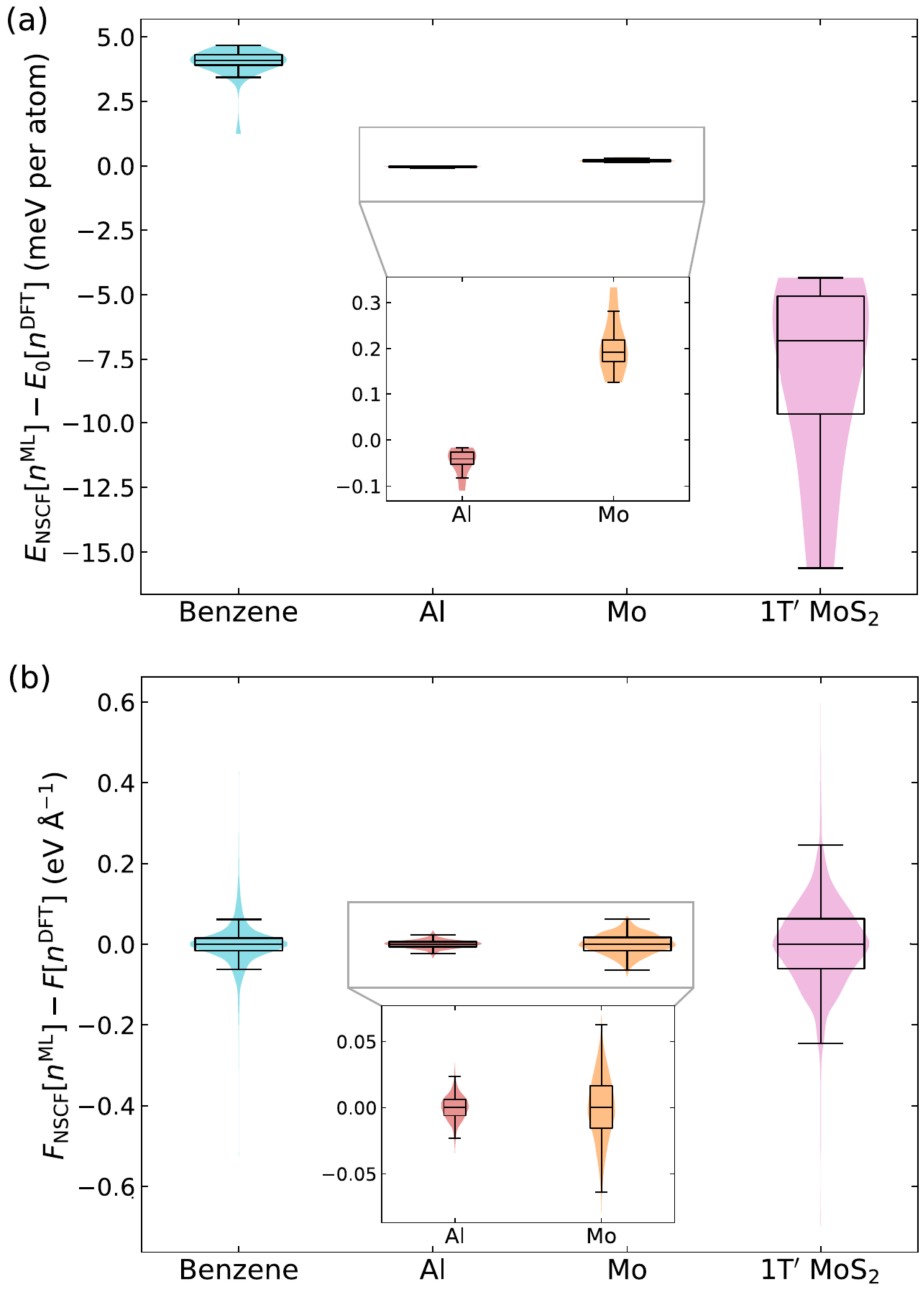

Figure 5.7: Violin plots for the difference between the NSC energies (a) and forces (b) against the fully converged ones, for all the systems studied in this chapter. The inserts show a magnification of the results for the aluminium and molybdenum systems, which are approximately one order of magnitude more accurate than the other two. In particular, the accuracies on the energies of, at worst, a few meV/atom, are on par with the accuracies reached by MLP models.

Adapted Symmetry Functions The main idea for the JLCDM is based, in part, on the same idea behind the extension of the Behler-Parrinello symmetry function to scalar fields, done in Ref. [83]. However, some important differences are present. The first one is that the JLCDM, being based on a multi-body expansion, is not limited on the number of

body that are represented by the descriptors. Indeed, since the Behler-Parrinello functions (and their derivatives) depend only on distances, they are intrinsically 2-body. Another important difference consist in the model itself. Whereas the work from Ref. [83] uses a Neural Network (NN) architecture, the JLDCM is linear. This has the advantage of making the model more interpretable and significantly more compact[30], while preserving similar accuracies[31].

SNAP The work from Ref. [84] uses the same core idea of the adapted symmetry functions one, but with the 4-dimensional bispectrum (as used in SNAP, see Sec. 2.1.2 and Ref. [24]) as features. Thus, while the JLCDM can reach any body order of choice, the 4d bispectrum components are 4B features. However, the 4d bispectrum is also more constrained, as it includes more symmetries than required (as shown in Sec. 2.1.2). Moreover the work is, again, based on a NN model.

SALTED A qualitative comparison between the JLCDM and the SALTED model of Ref. [29], based on the λ-SOAP is less straightforward. Specifically, the underlying main idea is different. Indeed, as shown in eq. 2.86, here reported for readability

$$\rho(\mathbf{r}) = \sum_{i}^{\text{atoms}} \rho_i(\mathbf{r}) = \sum_{i}^{\text{atoms}} \sum_{nlm} c_{inlm} R_{nl}(r) Y_l^m(\hat{\mathbf{r}}), \tag{5.15}$$

the decomposition of the density is done in terms of a Resolution of Identity (RI) ansatz. Explicitly, whereas the JLCDM is founded on an expansion of the density in terms of a multi-body terms, the SALTED model expand the density in terms of an atom-centered basis (RI basis) and then targets the coefficients with a λ-SOAP model. Usually the basis is already provided by the specific code, and so both the coefficients and the basis are readily available. Moreover, the coefficients can be used to initialize the DFT calculations. Thus, on the one hand, the SALTED approach requires the use of a covariant model (the λ-SOAP) to evaluate the coefficients for each atom of the system. On the other hand, the JLCDM uses a simpler (linear) model, but has to evaluate the descriptors for each grid point. While specific tasks could be specifically tuned for one model or the other (e.g., the JLCDM is more efficient in providing the value of the density in single points in space, while the SALTED is a good choice if an atom-centered decomposition is needed), it is not yet clear if there is a clear computational advantage between the two[32].

[30] Comparing the size of the models, we have $\sim 10^6$ weights for a NN-based approach against, at most, a few thousands coefficients for the JLCDM.

[31] It is important to mention that a NN model could be more suited for much larger and various datasets.

[32] A comparison could be made by evaluating the *total* number of coefficients to be evaluated with SALTED and the number of grid points required by the JLCDM.

5.4 Conclusions

In conclusion, we presented the first expansion of the JL formalism applied to the description of scalar fields and, in particular, specialized it to the prediction of the electronic density. By leveraging on the construction of previous works, we showed how the model is naturally obtained by "promoting" the central atom of the JLP to be the position of an arbitrary point in space. We justified this choice by showing that the resulting density satisfies the correct symmetry properties under rotation. We therefore applied the same constraints devised for the JLP and constructed a fully interpretable (given the internal coordinate representation) and systematically improvable model for the charge density, here denominated with JLCDM.

We devised an improved strategy for the selection of the relevant grid points to be used in the training phase with the aim of removing the redundancies of the grid-points mesh, while retaining as much information as possible. The JLCDM was then tested on four systems, to challenge it against a molecule (benzene), metallic solids (aluminium and molybdenum) and the 2D MoS_2. Not only we showed that the model was able to reach a high level of accuracy for all the cases explored, but we also the transferability on new phase not present in the training set, all while keeping a relatively small number of features. Finally we compared the fully-converged energies and forces with the ones obtained from the predicted densities. The results showed that, when compared to a full-SC scheme, the model allowed to a halving of the DFT cycles required to reach *ab-initio* accuracies.

If this chapter was devoted to extend the JL formalism from the prediction of scalar quantities to the treatment of scalar fields, in the next chapter we will complete the definition of the complete JL framework, including the description of tensors and tensor fields. In this spirit, this chapter and the previous one can be seen as the formal foundations for the reminder of the thesis.

Chapter 6

The Full JL framework

6.1 Introduction

This chapter is devoted to the presentation of the full JL framework, extending the expansion for scalar and scalar fields to cases of covariant quantities, namely tensors and tensor fields. We will first introduce the Covariant-JL (CJL) formalism, which targets the spherical components of a general tensor. The CJL will be founded on the same assumptions (atomic decomposition and locality) of all the previous sections of this thesis.

Specifically, the core idea will be of constructing a scalar function, followed by encoding the full information on the components of the tensor. We will then apply a similar expansion to the one introduced in the previous section: this phase will employ a simple "promotion" idea, in which the role of one neighbor atom is taken by a grid point, contrary to the grid-centered representation of the JLCDM. Finally, the tensor components will be retrieved by evaluating simple angular integrals on spherical harmonics (please note that, we will never mention Wigner-D matrices in this derivation). Not only will the resulting expression will be reduced to the JLP for the scalar case, but we will also prove that the expansion consists, essentially, of JLP-like-invariant terms, decorated with appropriate covariant contributions, obtained by projecting the spherical harmonics onto the desired space. The resulting descriptors will be fully hierarchical and expanded in a multi-body fashion. If the method has value in itself, since it combines the progressive expansion in multi-body terms with a simple derivation in terms of integrals on spherical harmonics only (contrary to the usual approaches in terms of Wigner-D matrices, shown in Sec. 2.2), it also complements the JLCDM for the acceleration of *ab-initio* calculations, since it allows us to target the PAW augmentation charges. This will be technically explored in the application section.

We will conclude the chapter by going beyond the description of tensors, with the introduction of descriptors for tensor fields: this will be one of the main results of the

thesis, obtained by organically combining the JLP, the JLCDM and the CJL formalisms. This achievement, which relies on everything developed so far across the thesis, will be the last missing piece for the definition of the full JL framework.

I remark that real applications of the methods of this chapter are still on-going. For this reason, I decided not to include any partial results that still necessitate final tests. However, I wish to mention that the CJL formalism, both stand-alone or applied to the PAW charges, is showing encouraging accuracies, at least at par with other models available in literature.

6.2 A recipe for cluster-expanded covariant models

This section will introduce the key ideas for the generalization of the JL formalism to tensorial quantities. In particular, we will show how the "promotion" strategy, introduced in the previous chapter, can be slightly modified to be applied also to tensorial cases.

Our main assumption, already exploited in previous works (please see Sec. 2.2), is that a general atomic tensor, $\boldsymbol{T}_l$, of rank l, can be separated into *atomic*-centered contributions. Explicitly, we write

$$\boldsymbol{T}_l = \sum_i^{\text{atoms}} \boldsymbol{T}_{i,l} \tag{6.1}$$

where the sum runs over the atoms of the system. We remark that not all the atoms must conform to the formula above. Indeed, if we expect only a few atomic environments to carry significant contributions to the full tensor $\boldsymbol{T}_l$, then we can neglect all the others and make the sum to run over the significant atoms only.

Real spherical decomposition In this preliminary paragraph we will show how to decompose a real tensor (with no imaginary components) in terms of a decomposition in *real*-spherical harmonics. Indeed, this is the case for the vast majority of tensors of interest in the computational study of materials.

As for the spherical decomposition showed in section 2.2.1, it is always possible to decompose a cartesian tensor into components that transform as real-spherical harmonics. In particular, we will follow the same approach as Ref. [52], and define the real spherical harmonics by means of the unitary transformation[33]

$$Y_{lm}(\mathbf{r}) := \begin{cases} (-1)^m \sqrt{2}\,\mathrm{Re}[Y_l^m(\mathbf{r})] & \text{for } m > 0, \\ Y_l^0(\mathbf{r}) & \text{for } m = 0, \\ (-1)^m \sqrt{2}\,\mathrm{Im}[Y_l^{-m}(\mathbf{r})] & \text{for } m < 0, \end{cases} \tag{6.2}$$

[33]Please note that the following definition has an additional $(-1)^m$ factor with respect to that of Ref. [52].

which can also be written in the matrix form

$$Y_{lm}(\hat{\mathbf{r}}) = \sum_{m'} U^l_{mm'} Y_l^{m'}(\hat{\mathbf{r}}). \tag{6.3}$$

Here the unitary matrix $\boldsymbol{U}$ is defined as

$$U^l_{mm'} = \delta_{m0} + \frac{1-\delta_{m0}}{\sqrt{2}}\left[H(m)\Big((-1)^m\delta_{m'm} + \delta_{m'-m}\Big) + iH(-m)\Big((-1)^{m+1}\delta_{m'-m} + \delta_{m'm}\Big)\right], \tag{6.4}$$

with $H(m)$ being the Heaviside function. The procedure used to obtain the real-spherical components of a tensor is the same as the one described in section 2.2.1, but with the additional unitary transformation introduced by $\boldsymbol{U}$. However, care must be taken after performing the required couplings. Indeed, for example, the $l = 1$ spherical components of a tensor of rank 2, T_1^q, are given by[34]

$$\begin{cases} T_1^{\pm 1} = \dfrac{1}{2}\left[\mp(T_{xz} - T_{zx}) - i(T_{yz} - T_{zy})\right], \\ T_1^0 = \dfrac{i}{\sqrt{2}}(T_{xy} - T_{yx}). \end{cases} \tag{6.5}$$

As it can be see from the T_1^0 component, which is left untouched by the unitary transformation $\boldsymbol{U}$, this term would be purely imaginary. Thus, we need to apply a further unitary transformation by multiplying all these terms by the inverse of the imaginary unit, $-i$, before proceeding with the transformation brought by $\boldsymbol{U}$. By doing that, the real components can be written as[35]

$$\begin{cases} T_{11} = \dfrac{1}{\sqrt{2}}(T_{yz} - T_{zy}) = \dfrac{1}{\sqrt{2}}(\mathbf{U}\times\mathbf{V})_x, \\ T_{10} = \dfrac{1}{\sqrt{2}}(T_{yz} - T_{zy}) = \dfrac{1}{\sqrt{2}}(\mathbf{U}\times\mathbf{V})_z, \\ T_{1-1} = \dfrac{1}{\sqrt{2}}(T_{zx} - T_{xz}) = \dfrac{1}{\sqrt{2}}(\mathbf{U}\times\mathbf{V})_y, \end{cases} \tag{6.6}$$

where the tensor $\boldsymbol{T}$ was represented in terms of a dyad, i.e., $\boldsymbol{T} = \mathbf{U}\otimes\mathbf{V}$. The expressions above show that the procedure produces indeed a real tensor (in this case it is a vector, since $l = 1$). In the following, we will always imply that the tensor $\boldsymbol{T}$ has already been decomposed in terms of its real harmonic components, i.e., we will investigate only terms in the form T_{lm}.

We conclude this paragraph by mentioning that the harmonic decomposition of fully

[34] Please note that there is no degeneracy in the partition of the angular momentum space for a tensor of rank 2.

[35] We follow the convention that terms with only subscripts are real, i.e., $T_{lm} = {}^R T_l^m$.

symmetric tensors undergoes a significant reduction of the components involved, as explicitly shown in Ref. [74]. This can be already appreciated for the components shown above: if the tensor $\boldsymbol{T}$ was fully symmetric under (cartesian) indexes swap, then all the $l = 1$ components would have vanished.

6.2.1 Constructing a scalar field

We are now introducing the core idea of this chapter: we want to create a *scalar field* from the atomic contributions $\boldsymbol{T}_{i,l}$. In doing so, we will establish a strong link with the construction of the JLCDM of the previous chapter. Indeed, we have already discussed how the JL formalism can be extended to encompass also scalar fields, by formally "promoting" the central atom to be a grid point, and then by performing a cluster expansion on the resulting terms. In this chapter we will do something similar, where the method will be formally equivalent to "promoting" an atom in the neighborhood of the central one.

The construction of the scalar field is done by introducing an *auxiliary* versor, $\hat{\mathbf{r}}_{gi}$, starting from the central atom and ending in a point in real space, here denoted with g and called grid point for convenience[36]. This implies that the object will now depend on the real space coordinates $\mathbf{r}_g$, namely we are indeed constructing a real-space field. The simplest way to define a scalar field is then

$$\mathcal{T}_{i,l}(\hat{\mathbf{r}}_{gi}) := \sum_{m=-l}^{l} T_{i,lm} Y_{lm}(\hat{\mathbf{r}}_{gi}), \tag{6.7}$$

where the components of the tensor can be easily extracted by integration over the solid angle

$$T_{i,lm} = \int \mathrm{d}\hat{\mathbf{r}}_{gi} \mathcal{T}_{i,l}(\hat{\mathbf{r}}_{gi}) Y_{lm}(\hat{\mathbf{r}}_{gi}), \tag{6.8}$$

considering the orthogonality of the spherical harmonics[37] Before proceeding, however, let us prove that the function $\mathcal{T}_{i,l}(\hat{\mathbf{r}}_{gi})$ is indeed a scalar field under rotation, i.e., that it satisfies the relation

$$\mathcal{T}_{i,l}\left(\widehat{\hat{R}^{-1}\mathbf{r}_g - \mathbf{r}_i}; \{\hat{\mathbf{r}}_j\}\right) = \mathcal{T}_{i,l}\left(\widehat{\mathbf{r}_g - \hat{R}\mathbf{r}_i}; \{\hat{R}\hat{\mathbf{r}}_j\}\right), \tag{6.9}$$

where $\hat{R}$ is a generic rotation, and where we explicitly showed the dependence of the field

[36] While the point must not necessarily belong to a grid mesh, keeping the same nomenclature of the previous chapter allows to preserve a coherent language for the full framework.

[37] A unitary transformation does not affect the orthogonality relations, so the real spherical harmonics are orthogonal themselves.

on the atomic positions, $\{\hat{\mathbf{r}}_j\}$. Please, note that we used the following compact notation

$$\widehat{\hat{R}^{-1}\mathbf{r}_g - \mathbf{r}_i} := \frac{\hat{R}^{-1}\mathbf{r}_g - \mathbf{r}_i}{\left|\hat{R}^{-1}\mathbf{r}_g - \mathbf{r}_i\right|},$$

to indicate the normalization of rather-lengthy expressions.

Proof that $\mathcal{T}_{i,l}(\hat{\mathbf{r}}_{gi})$ is a scalar field Let us prove that Eq. (6.9) holds. We can cast the definition of the scalar field $\mathcal{T}_{i,l}$ in the equivalent form[38]

$$\mathcal{T}_{i,l}(\hat{\mathbf{r}}_{gi}; \{\hat{\mathbf{r}}_j\}) = \sum_{m=-l}^{l} T^*_{i,lm}(\{\hat{\mathbf{r}}_j\}) Y_{lm}(\hat{\mathbf{r}}_{gi}) \tag{6.10}$$

which holds since the components $T^*_{i,lm}$ are real. Since we want to investigate the response of the field $\mathcal{T}^*_{i,lm}$ to an arbitrary rotation, it is useful to derive the transformation rules for the real spherical harmonics. They can be obtained by using the definition in terms of the unitary matrix $\boldsymbol{U}$ [see Eq. (6.3)] as

$$\begin{aligned} Y_{lm}(\hat{R}\hat{\mathbf{r}}) &= \sum_{m'} U^l_{mm'} Y_l^{m'}(\hat{R}\hat{\mathbf{r}}) = \sum_{m'm''} U^l_{mm'} D^{l*}_{m'm''}(\mathcal{R}) Y_l^{m''}(\hat{\mathbf{r}}) \\ &= \sum_{m'm''m'''} U^l_{mm'} D^{l*}_{m'm''}(\mathcal{R}) U^{l*}_{m'''m''} Y_{lm'''}(\hat{\mathbf{r}}), \end{aligned} \tag{6.11}$$

where the rotation was performed on the standard spherical harmonics by means of the Wigner-D matrices, followed by a re-casting in terms of the real spherical harmonics by means of the inverse of the matrix $\boldsymbol{U}$, i.e., $(\boldsymbol{U}^{-1})^l_{mm'} = U^{l*}_{m'm}$. Thus we can write

$$Y_{lm}(\hat{R}\mathbf{r}) = \sum_{m} {}^R D^{l*}_{mm'}(\mathcal{R}) Y_l^m(\mathbf{r}), \tag{6.12}$$

with the analogues of the Wigner-D matrices for the real spherical harmonics defined as

$${}^R D^l_{mm'}(\mathcal{R}) := \sum_{m''m'''} U^{l*}_{mm''} D^l_{m''m'''}(\mathcal{R}) U^l_{m'm'''}. \tag{6.13}$$

We can now proceed in evaluating the effect of applying a rotation[39] $\hat{R}^{-1}$ on the real space vectors $\mathbf{r}_g$. Explicitly, we write

[38] We use the form to exploit the orthogonality relations of the classical spherical harmonics.

[39] Please, note that we use here the inverse rotation, to express everything in terms of an active rotation on the system of atoms, instead of a passive rotation acting on the frame of reference, here represented by $\mathbf{r}_g$.

$$\begin{aligned}\mathcal{T}_{i,l}\left(\widehat{\hat{R}^{-1}\mathbf{r}_g-\mathbf{r}_i};\{\hat{\mathbf{r}}_j\}\right) &= \sum_{m=-l}^{l} T^*_{i,lm}(\{\hat{\mathbf{r}}_j\})Y_{lm}\left(\widehat{\hat{R}^{-1}\mathbf{r}_g-\mathbf{r}_i}\right)\\ &= \sum_{m} T^*_{i,lm}(\{\hat{\mathbf{r}}_j\}) \sum_{m'm''m'''} U^l_{mm'} \underbrace{D^{l*}_{m'm''}(\mathcal{R}^{-1})}_{=D^l_{m''m'}(\mathcal{R})} U^{l*}_{m'''m''} Y_{i,lm'''}\left(\widehat{\mathbf{r}_g-\hat{R}\mathbf{r}_i}\right) \qquad (6.14)\\ &= \sum_{m'''}\left[\sum_{m} {}^R D^{l*}_{m'''m}(\mathcal{R}) T_{i,lm}(\{\hat{\mathbf{r}}_j\})\right]^* Y_{i,lm'''}\left(\widehat{\mathbf{r}_g-\hat{R}\mathbf{r}_i}\right).\end{aligned}$$

In going from the first to the second equality, we have used the relation $(\widehat{\hat{R}^{-1}\mathbf{r}_g-\mathbf{r}_i}) = \hat{R}^{-1}(\widehat{\mathbf{r}_g-\hat{R}\mathbf{r}_i})$, to apply the rotation of real spherical harmonics. By comparing the terms inside the square brackets with Eq. (6.13), we can appreciate that this is indeed the expected formula for the transformation of the real-spherical components of the tensor $\boldsymbol{T}_{i,l}$. In other words, by construction, the real-spherical components of the tensor follows the same transformation rules of the real-spherical harmonics, namely,

$$T_{i,lm}(\{\hat{R}\hat{\mathbf{r}}_j\}) = \sum_{m'} {}^R D^{l*}_{mm'}(\mathcal{R}) T_{i,lm'}(\{\hat{\mathbf{r}}_j\}). \qquad (6.15)$$

Therefore, we finally have that

$$\mathcal{T}_{i,l}\left(\widehat{\hat{R}^{-1}\mathbf{r}_g-\mathbf{r}_i};\{\hat{\mathbf{r}}_j\}\right) = \sum_{m'''} T_{i,lm'''}(\{\hat{R}\hat{\mathbf{r}}_j\}) Y_{i,lm'''}\left(\widehat{\mathbf{r}_g-\hat{R}\mathbf{r}_i}\right) = \mathcal{T}_{i,l}\left(\widehat{\mathbf{r}_g-\hat{R}\mathbf{r}_i};\{\hat{R}\hat{\mathbf{r}}_j\}\right), \qquad (6.16)$$

which proves that $\mathcal{T}_{i,l}$ is indeed a scalar field.

6.2.2 Cluster expansion

Having established that Eq.(6.8) defines a scalar field, we can now propose a suitable expansion by follow closely the method outlined in the the previous chapter. The first step is to expand $\mathcal{T}_{i,l}(\hat{\mathbf{r}}_{gi})$ in a multi-body (cluster) expansion as,

$$\mathcal{T}_i(\hat{\mathbf{r}}_{gi}) = \mathcal{T}_i^{(1\mathrm{B})}(\hat{\mathbf{r}}_{gi}) + \mathcal{T}_i^{(2\mathrm{B})}(\hat{\mathbf{r}}_{gi}) + \mathcal{T}_i^{(3\mathrm{B})}(\hat{\mathbf{r}}_{gi}) + \ldots, \qquad (6.17)$$

where we have generalized the procedure by dropping the rank indexes l, which will be recovered by means of the integration described in Eq. (6.8). Here, the nB terms $\mathcal{T}_i^{(\mathrm{nB})}(\hat{\mathbf{r}}_{gi})$, consist of a sum of contributions, each depending on the position of n-1 atoms in the neighborhood of the i-th one. Since $\mathcal{T}_i(\hat{\mathbf{r}}_{gi})$ is a scalar field, all the terms on the right-hand side of the equation above must be scalar fields too. We then deduce immediately that

the 1B term must be a constant, which we assume to depend only on the atomic species of the i-th atom, Z_i, namely, $\mathcal{T}_i^{(1\mathrm{B})}(\hat{\mathbf{r}}_{gi}) = a_0^{Z_i}$.

Crucially, all the nB terms, $\mathcal{T}^{(\mathrm{nB})}$, being scalar fields, they can always be written in terms of degrees of freedom that mirror the transformation properties of a scalar field. This is done according to what shown in Sec. 5.1 of the last chapter, i.e., by introducing an internal system of coordinates that includes also the grid point. We then propose the following functional forms

$$\begin{cases} \mathcal{T}_i^{(1\mathrm{B})}(\hat{\mathbf{r}}_{gi}) = a_0^{Z_i}, \\ \mathcal{T}_i^{(2\mathrm{B})}(\hat{\mathbf{r}}_{gi}) = \sum\limits_{(j)_i} T_{ji}^{(2\mathrm{B})}(r_{ji}, s_{gji}), \\ \mathcal{T}_i^{(3\mathrm{B})}(\hat{\mathbf{r}}_{gi}) = \sum\limits_{(j,k)_i} T_{jki}^{(3\mathrm{B})}(r_{ji}, r_{ki}, s_{gji}, s_{gki}, s_{jki}), \\ \dots, \end{cases} \tag{6.18}$$

where we defined again the scalar products $s_{gji} := \hat{\mathbf{r}}_{gi} \cdot \hat{\mathbf{r}}_{ji}$. The first sum runs over all the atoms in the neighborhood of the i-th and, similarly, the second sum runs over all the possible pairs of atoms in the same neighborhood. In doing so, we introduce the last assumption for the proposed expansion, namely the locality of the representation, which is the final ingredient to link this formalism with the one developed for the JLCDMs. Therefore, we assume that two atoms will interact only up to an optimized cut-off radius, r_{cut}. Here we will explicitly consider only terms up to 3B, but, crucially, the procedure can be extended also to any higher-body order, e.g., by expanding $\mathcal{T}_i^{(4\mathrm{B})}(\hat{\mathbf{r}}_{gi})$ in sum of terms depending on three distances and six angles. This means that also the formalism proposed here, in line with the on for the JLP and the JLCDM, is hierarchical and systematically improvable.

We can now leverage the dependence on the internal coordinates of Eq. (6.36), and perform the same expansion in terms of (double)-vanishing Jacobi polynomials [see Eqs. (4.10) and (4.26)] for each distance, and in terms of Legendre polynomials for each angle (scalar product), respectively. This leads to the expansion

$$\begin{cases} \mathcal{T}_i^{(1\mathrm{B})}(\hat{\mathbf{r}}_{gi}) = a_0^{Z_i}, \\ \mathcal{T}_i^{(2\mathrm{B})}(\hat{\mathbf{r}}_{gi}) = \sum\limits_{(j)_i} \sum\limits_{nl} a_{nl}^{Z_j Z_i} \left(\delta_{l0} \widetilde{P}_{nji}^{(\alpha,\beta)} + (1-\delta_{l0}) \overline{P}_{nji}^{(\alpha,\beta)} \right) P_l^{gji}, \\ \mathcal{T}_i^{(3\mathrm{B})}(\hat{\mathbf{r}}_{gi}) = \sum\limits_{(j,k)_i} \sum\limits_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{\text{unique}} a_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{Z_j Z_k Z_i} \sum\limits_{\text{symm.}} \left(\overline{P}_{n_1 ji}^{(\alpha,\beta)} \overline{P}_{n_2 ki}^{(\alpha,\beta)} P_{l_1}^{gji} P_{l_2}^{gki} P_{l_3}^{jki} \right), \\ \dots. \end{cases} \tag{6.19}$$

We can immediately appreciate how the formulas above are very close to the ones derived in Eqs. (4.18) and (4.30) for the JLPs, and Eq. (5.8) for the JLCDMs. Indeed, by looking more closely at the contributions for $l = 0$ in the 2B term, and $l_1 = l_2 = 0$ in the 3B one, we have

$$\begin{cases} \left(\mathcal{T}_i^{(\mathrm{2B})}(\hat{\mathbf{r}}_{gi})\right)_{l=0} = \displaystyle\sum_{(j)_i}\sum_{n} a_{n0}^{Z_j Z_i}\widetilde{P}_{nji}^{(\alpha,\beta)}, \\ \left(\mathcal{T}_i^{(\mathrm{3B})}(\hat{\mathbf{r}}_{gi})\right)_{l_1=l_2=0} = \displaystyle\sum_{(j,k)_i}\sum_{\substack{n_1 n_2 \\ l_3}}^{\text{unique}} a_{\substack{n_1 n_2 \\ 00 l_3}}^{Z_j Z_k Z_i} \sum_{\text{symm.}} \left(\overline{P}_{n_1 ji}^{(\alpha,\beta)}\overline{P}_{n_2 ki}^{(\alpha,\beta)} P_{l_3}^{jki}\right), \end{cases} \tag{6.20}$$

which we recognize as the 2B and 3B contributions of the JLP expansion [please see again Eqs. (4.18) and (4.30)]. Thus, we can infer that all the properties of the JLP are inherited by this new expansion: we have the two hyperparameters, $\alpha, \beta > -1$, which define the optimal basis for the radial part, and the expansion over the Jacobi indexes is truncated to an optimized $n_{\max}$. The vanishing and double-vanishing Jacobi polynomials are defined by means of Eqs. (4.10) and (4.26), and by the shorthand notation

$$\widetilde{P}_{nji}^{(\alpha,\beta)} := \widetilde{P}_n^{(\alpha,\beta)}\left(\cos\left(\pi\frac{r_{ji}-r_{\min}}{r_{\mathrm{cut}}-r_{\min}}\right)\right), \quad \text{and} \quad \widetilde{P}_{nji}^{(\alpha,\beta)} := \overline{P}_n^{(\alpha,\beta)}\left(\cos\left(\pi\frac{r_{ji}}{r_{\mathrm{cut}}}\right)\right),$$

and

$$P_l^{jki} := P_l(\hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{ki}).$$

We remark that the double vanishing polynomials do not present any $r_{\min}$, to preserve the continuity of the expressions, as discussed in the study of Eq. (5.10). Also, given that our implied assumption is that the functions depend only on the atomic species, then the selection of the unique indexes, and the subsequent sum over the symmetric expressions, ensures the invariance under permutations of identical atoms. Finally, by comparing directly the expansion for the JLCDM in Eq. (5.8) and the one above, Eq. (6.20), we note that the most important difference is that the first is obtained by "promoting"[40], from the JLP framework, the central atom to be a grid point. Conversely, the one here can be thought as obtained by a promotion of one atom in the neighborhood. This observation makes it easier to construct the two models from the JLP, showing that we are navigating in a shared framework. As will be shown in the next section, promoting one atom in the neighborhood instead of the central one (as done for the JLCDM), allows to greatly simplify the evaluation of the integrals shown in Eq. (6.8).

Despite the many similarities between the expansion of Eq. (6.18), and the analogous ones for the JLP and the JLCDM, we also have, however, a few differences. First of all,

[40] As done in the previous chapter, the use of this promotion idea is just symbolic, and has more to do with an operative "substitute the i in the formula with g" than anything else.

we will never consider the distance between the central atom and the grid point, and thus we are left with one less Jacobi index [please, compare with Eq. (5.8)]. Also, the truncation for the Legendre indexes will be different between polynomials that depend on the position of the grid point, and the ones that contain only the atomic position, as will be clarified in the next section. Finally, the introduction of the Kronecker delta, δ_{l0}, to separate the zero-th component of the 2B term from the rest, is imposed to maintain the continuity of the representation. This removes the jump caused by the ill-definition of the scalar products when $j \to i$, in the same spirit of the removal of $r_{\min}$ from the double-vanishing-Jacobi polynomials (we remark here that the double-vanishing Jacobi polynomials smoothly vanish when the distance between the atoms tends to zero).

6.3 The components of the tensor from a scalar field

Equipped with the JL expansion, of Eq. (6.19), we can finally evaluate the integrals in Eq. (6.8) and obtain the desired covariant model. In particular, given the linearity of the integrals, we can integrate each of the nB terms separately. In Fig. 6.1, we graphically show the full construction of the model. Explicitly, we want to evaluate integrals of the form

$$T_{i,lm} = \sum_{v=1}^{\substack{\text{body}\\\text{order}}} T_{i,lm}^{(v\text{B})} = \sum_{v=1}^{\substack{\text{body}\\\text{order}}} \int \mathrm{d}\hat{\mathbf{r}}_{gi} \mathcal{T}_i^{(v\text{B})}(\hat{\mathbf{r}}_{gi}) Y_{lm}(\hat{\mathbf{r}}_{gi}), \tag{6.21}$$

where the vB contribution to the components, $T_{i,lm}^{(v\text{B})}$, are defined by the integrals on the right-hand side. We will consider each body order separately.

1B Terms. The 1B case is trivial, since $T_{i,lm}^{(1\text{B})}$ is a constant. Thus, the integral reads

$$T_{i,lm}^{(1\text{B})} = a_0^{Z_i} \int \mathrm{d}\hat{\mathbf{r}}_{gi} Y_{lm}(\hat{\mathbf{r}}_{gi}) = \sum_{m'} a_0^{Z_i} U_{mm'}^{l} \underbrace{\int \mathrm{d}\hat{\mathbf{r}}_{gi} Y_{l}^{m'}(\hat{\mathbf{r}}_{gi})}_{=\sqrt{4\pi}\delta_{l0}\delta_{m'0}} = \delta_{l0}\delta_{m0}\sqrt{4\pi} a_0^{Z_i}, \tag{6.22}$$

where we have used the definition of the real spherical harmonics in terms of the matrix $\boldsymbol{U}$ [Eq. 6.3], and used the fact that the 0-th component of the real spherical harmonics is identical to the one of the complex ones, i.e., $U_{m0}^{l} = \delta_{m0}$. Since this term is relevant only for the $l = 0$ case, it is a spherically-symmetric contribution. In the reminder, we will absorb the unessential coefficient $\sqrt{4\pi}$ in the expansion coefficients, by a re-definition of $a_0^{Z_i}$.

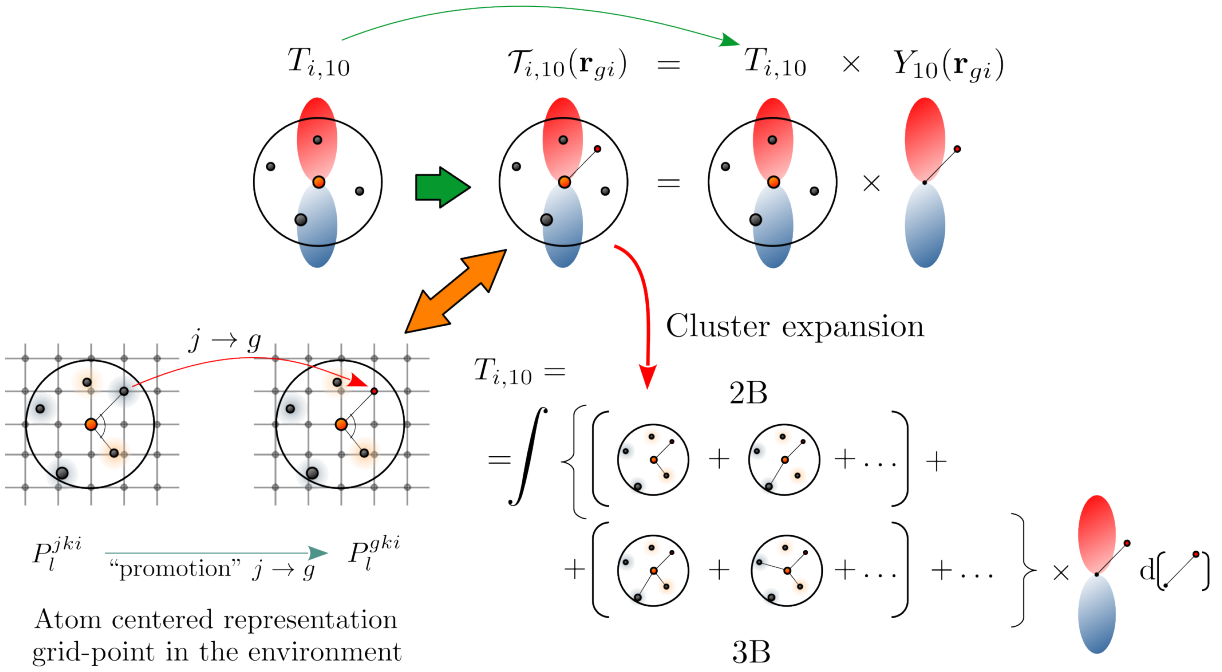

Figure 6.1: We show here a visual guide for the model presented in this chapter. Our aim is to target an atomic-related tensorial quantity, here graphically represented by the atomic environment and an associated harmonic component of the tensor (the polar graph emerging from the central atom). (Follow the green arrow) The first step is to construct an adequate scalar field that encodes the components of the tensor. This can be done naturally, by multiplying the spherical components of the tensor by the corresponding spherical harmonics. (Follow the orange arrow) The scalar field is then described by means of a recipe closely related to the one followed for the JLCDM models: indeed this is formally equivalent to "promote" one atom in the environment to be a grid point. This allows one to apply a cluster expansion to the scalar field, which implies an analogous expansion on the tensor components. The components are then retrieved by integrating each cluster contribution against the appropriate spherical harmonic.

2B Terms. The 2B contributions are given by the integrals

$$T^{(2\mathrm{B})}_{i,lm} = \int \mathrm{d}\hat{\mathbf{r}}_{gi} Y_{lm}(\hat{\mathbf{r}}_{gi}) \sum_{(j)_i} \sum_{nl'} a^{Z_j Z_i}_{nl'} \left(\delta_{l'0} \widetilde{P}^{(\alpha,\beta)}_{nji} + (1-\delta_{l'0}) \overline{P}^{(\alpha,\beta)}_{nji} \right) P^{gji}_{l'}. \tag{6.23}$$

The first term does not contain any dependence on the spherical harmonics and thus will lead again to a spherically symmetric term and to a factor $\sqrt{4\pi}$. Instead, to evaluate the second term, we want to expand the expression of the Legendre polynomials in terms of real spherical harmonics. This can be done by the use of the addition theorem for

spherical harmonics. Indeed it holds that

$$\sum_m \underbrace{Y^*_{lm}(\hat{\mathbf{r}}_1)}_{=Y_{lm}(\hat{\mathbf{r}}_1)} Y_{lm}(\hat{\mathbf{r}}_2) = \sum_{m'm''} \underbrace{\left[\sum_m U^{l*}_{mm'} U^l_{mm''}\right]}_{=\delta_{m'm''}} Y_l^{m'*}(\hat{\mathbf{r}}_1) Y_l^{m''}(\hat{\mathbf{r}}_2) = \sum_{m'} Y_l^{m'*}(\hat{\mathbf{r}}_1) Y_l^{m'}(\hat{\mathbf{r}}_2), \tag{6.24}$$

where we have used the fact that Y_{lm} is real and that the matrix $\boldsymbol{U}$ is unitary. Thus, from Eq. (2.15), we have that the addition theorem is preserved in the real spherical harmonics, and reads

$$P_l^{gji} = \frac{4\pi}{2l+1} \sum_m Y_{lm}(\hat{\mathbf{r}}_{gi}) Y_{lm}(\hat{\mathbf{r}}_{ji}). \tag{6.25}$$

This yields the integral

$$\int \mathrm{d}\hat{\mathbf{r}}_{gi} Y_{lm}(\hat{\mathbf{r}}_{gi}) P_{l'}^{gji} = \frac{4\pi}{2l'+1} \sum_{m'} Y_{l'm'}(\hat{\mathbf{r}}_{ji}) \underbrace{\int \mathrm{d}\hat{\mathbf{r}}_{gi} Y_{lm}(\hat{\mathbf{r}}_{gi}) Y_{l'm'}(\hat{\mathbf{r}}_{gi})}_{\delta_{ll'}\delta_{mm'}} = \frac{4\pi}{2l+1} Y_{lm}(\hat{\mathbf{r}}_{ji}), \tag{6.26}$$

evaluated by means of the orthogonality of the real-spherical harmonics.

Crucially, we can appreciate how this operation contains, at its core, the very justification on the use of the Legendre polynomials in the entire JL formalism. Indeed, we can always separate the Legendre polynomials in terms of spherical harmonics, not only allowing us to easily compare with other descriptors in literature, but also providing the natural tools to evaluate all the angular integrals with ease, all while not sacrificing the interpretability of the internal coordinates representation.

We can now write the full expression for the 2B contributions to the tensor components $T_{i,lm}$ as

$$T^{(2\mathrm{B})}_{i,lm} = \frac{4\pi}{2l+1} \sum_{j,j\neq i}^{\text{atoms}} \sum_n^{n_{\max}} \left[\delta_{l0} \frac{1}{\sqrt{4\pi}} a^{Z_j Z_i}_{n0} \widetilde{P}^{(\alpha,\beta)}_{nji} + (1-\delta_{l0}) a^{Z_j Z_i}_{nl} \overline{P}^{(\alpha,\beta)}_{nji} Y_{lm}(\hat{\mathbf{r}}_{ji})\right]. \tag{6.27}$$

We remark that, since the case $l = 0$ was left untouched by the integral projection, it is still proportional to the analogous 2B terms in JL-potentials [please compare with Eq. 4.18]. This is not surprising, since the $l = 0$ term is a scalar, exactly what the JLP targets; nevertheless it is a good check and demonstrates that we are exploring a coherent and general framework. Importantly, the expansion coefficients do not depend on the magnetic number m. Not only does this significantly reduce the number of coefficients that have to be determined, but also it ensures that the expressions are precisely covariant, since the fit does not affect the relation between coefficients of different magnetic numbers. Please note, again, that all the unessential multiplicative factors will be absorbed in a redefinition of the expansion coefficients.

3B Terms: the $l = 0$ case. For reasons that will be clear shortly, we divide the discussion on the 3B terms by firstly discussing the $l = 0$ case. The integral is explicitly written as

$$T^{(3\mathrm{B})}_{i,00}(\hat{\mathbf{r}}_{gi}) = \underbrace{\frac{1}{\sqrt{4\pi}}}_{=Y_{00}} \int \mathrm{d}\hat{\mathbf{r}}_{gi} \sum_{(j,k)_i} \sum_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{\text{unique}} a^{Z_j Z_k Z_i}_{\substack{n_1 n_2 \\ l_1 l_2 l_3}} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1 ji} \overline{P}^{(\alpha,\beta)}_{n_2 ki} P^{gji}_{l_1} P^{gki}_{l_2} P^{jki}_{l_3} \right). \tag{6.28}$$

If we now use the additional theorem twice, for each Legendre polynomial that depends on the grid points, we can evaluate the integral

$$\begin{aligned} \int \mathrm{d}\hat{\mathbf{r}}_{gi} P^{gji}_{l_1} P^{gki}_{l_2} &= \frac{(4\pi)^2}{(2l_1+1)(2l_2+1)} \sum_{m_1 m_2} Y_{l_1 m_1}(\hat{\mathbf{r}}_{ji}) Y_{l_2 m_2}(\hat{\mathbf{r}}_{ki}) \underbrace{\int \mathrm{d}\hat{\mathbf{r}}_{gi} Y_{l_1 m_1}(\hat{\mathbf{r}}_{gi}) Y_{l_2 m_2}(\hat{\mathbf{r}}_{gi})}_{\delta_{l_1 l_2}\delta_{m_1 m_2}} \\ &= \delta_{l_1 l_2} \frac{(4\pi)^2}{(2l_1+1)(2l_2+1)} \sum_{m_1} Y_{l_1 m_1}(\hat{\mathbf{r}}_{ji}) Y_{l_1 m_1}(\hat{\mathbf{r}}_{ki}) = \delta_{l_1 l_2} \frac{4\pi}{2l_1+1} P^{jki}_{l_1}, \end{aligned} \tag{6.29}$$

where, in the last step, we contracted the sum over m_1 into a single Legendre polynomial. By inserting this expression into Eq. (6.28), we see how we have a redundancy regarding the Legendre polynomials evaluated on $\hat{\mathbf{r}}_{ji} \cdot \hat{\mathbf{r}}_{ki}$, since they appear twice in the expression. However, we can reduce the expression by observing that the Legendre polynomials form a complete set for the functions defined on the interval $[-1, 1]$. Thus, we can reduce the product of the two polynomials by means of

$$P^{jki}_{l_1} P^{jki}_{l_2} = \sum_l c_{l_1 l_2 l} P^{jki}_l, \tag{6.30}$$

where the coefficients are given by[41]

$$c_{l_1 l_2 l} = \frac{2l+1}{2} \int_{-1}^{1} \mathrm{d}x P_{l_1}(x) P_{l_2}(x) P_l(x). \tag{6.31}$$

The actual evaluation of this coefficients is not relevant for our discussion. Instead, we can insert this expression back into the integral of Eq. (6.28) and, by defining new expansion coefficients $a^{Z_j Z_k Z_i}_{n_1 n_2 l}$ as

$$a^{Z_j Z_k Z_i}_{n_1 n_2 l} := \sqrt{4\pi} \sum_{l_1 l_3} \frac{1}{2l_1+1} a^{Z_j Z_k Z_i}_{\substack{n_1 n_2 \\ l_1 l_1 l_3}} c_{l_1 l_3 l}, \tag{6.32}$$

[41] Please note that the Legendre polynomials are not normalized, and so a factor $(2l+1)/2$ must be explicitly included.

following which we can write the $T^{(3B)}_{i,00}(\hat{\mathbf{r}}_{gi})$ components as[42]

$$T^{(3B)}_{i,00} = \sum_{(j,k)_i} \sum_{n_1 n_2 l_1}^{\text{unique}} a^{Z_j Z_k Z_i}_{n_1 n_2 l_1} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1 ji} \overline{P}^{(\alpha,\beta)}_{n_2 ki} P^{jki}_{l_1} \right). \tag{6.33}$$

Unsurprisingly, we just re-derived the 3B term of the JLP expansion, as it can be seen by comparing with Eq. (4.30). This is consistent with the fact that we are dealing with a scalar quantity.

3B Terms: $l > 0$ cases. We can finally consider all the $l > 0$ cases of the 3B expansion. The integral now reads

$$T^{(3B)}_{i,lm}(\hat{\mathbf{r}}_{gi}) = \sum_{(j,k)_i} \sum_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{\text{unique}} a^{Z_j Z_k Z_i}_{\substack{n_1 n_2 \\ l_1 l_2 l_3}} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1 ji} \overline{P}^{(\alpha,\beta)}_{n_2 ki} P^{jki}_{l_3} \int \mathrm{d}\hat{\mathbf{r}}_{gi} P^{gji}_{l_1} P^{gki}_{l_2} Y_{lm}(\hat{\mathbf{r}}_{gi}) \right). \tag{6.34}$$

Comparing this expression with the analogous ones from the JLP, we can see how each term is weighted by a covariant term. Indeed, this is the reason why we can consider a different expansion truncation for the pair (l_1, l_2) and for l_3: the pair completely encodes the covariant behaviour, while l_3 is devoted to the functional dependence on the atomic positions only.

The calculation of the integral is done by writing both Legendre polynomials as sums of products of spherical harmonics. This leads to integrals of the form

$${}^{R}G^{l_1 l_2 l}_{m_1 m_2 m} := \int Y_{l_1 m_1}(\hat{\mathbf{r}}_{gi}) Y_{l_2 m_2}(\hat{\mathbf{r}}_{gi}) Y_{lm}(\hat{\mathbf{r}}_{gi}) \mathrm{d}\hat{\mathbf{r}}_{gi} \,, \tag{6.35}$$

where ${}^{R}G^{l_1 l_2 l}_{m_1 m_2 m}$ is the analogous of the Gaunt coefficients [45, 52] for the real-spherical harmonics. An explicit calculation of this coefficients can be done by casting everything in terms of the complex spherical harmonics. Explicitly we write

$$\begin{aligned}
{}^{R}G^{l_1 l_2 l_3}_{m_1 m_2 m_3} &= \int Y_{l_1 m_1}(\hat{\mathbf{r}}) Y_{l_2 m_2}(\hat{\mathbf{r}}) \underbrace{Y^*_{l_3 m_3}(\hat{\mathbf{r}})}_{= Y_{l_3 m_3}(\hat{\mathbf{r}})} \mathrm{d}\hat{\mathbf{r}} \\
&= \sum_{m'_1 m'_2 m'_3} U^{l_1}_{m_1 m'_1} U^{l_2}_{m_2 m'_2} U^{l_3 *}_{m_3 m'_3} \int \mathrm{d}\hat{\mathbf{r}}\, Y^{m'_1}_{l_1}(\hat{\mathbf{r}}) Y^{m'_2}_{l_2}(\hat{\mathbf{r}}) Y^{m'_3 *}_{l_3}(\hat{\mathbf{r}}) \\
&= \sqrt{\frac{(2l_1+1)(2l_2+1)}{4\pi(2l_3+1)}} C^{l_3 0}_{l_1 0 l_2 0} \sum_{m'_1 m'_2 m'_3} U^{l_3 *}_{m_3 m'_3} C^{l_3 m'_3}_{l_1 m'_1 l_2 m'_2} U^{l_1}_{m_1 m'_1} U^{l_2}_{m_2 m'_2},
\end{aligned} \tag{6.36}$$

[42] We have also re-labelled l in l_1 to avoid confusion with the tensor component label.

where we have used the integral, in term of Clebsh-Gordan (CG) coefficients,

$$\int \mathrm{d}\hat{\mathbf{r}}\, Y_{l_1}^{m_1}(\hat{\mathbf{r}}) Y_{l_2}^{m_2}(\hat{\mathbf{r}}) Y_{l_3}^{m_3 *}(\hat{\mathbf{r}}) = \sqrt{\frac{(2l_1+1)(2l_2+1)}{4\pi(2l_3+1)}} C_{l_1 0 l_2 0}^{l_3 0} C_{l_1 m_1 l_2 m_2}^{l_3 m_3}, \tag{6.37}$$

for the product of three spherical harmonics. We remark that the coefficients ${}^{R}G_{m_1 m_2 m_3}^{l_1 l_2 l_3}$ can be pre-computed. Also, the terms $C_{l_1 0 l_2 0}^{l_3 0}$ impose that the coefficients are zero unless the sum $l_1 + l_2 + l_3$ is even.

The full 3B terms Finally, by absorbing all the unessential factors in the definition of the expansion coefficients, we obtain the full expansion for the 3B case

$$\begin{aligned} T_{i,lm}^{(3\mathrm{B})} = \delta_{l0} \sum_{(j,k)_i} \sum_{n_1 n_2 l_1}^{\text{unique}} a_{n_1 n_2 l_1}^{Z_j Z_k Z_i} \sum_{\text{symm.}} \left(\overline{P}_{n_1 j i}^{(\alpha,\beta)} \overline{P}_{n_2 k i}^{(\alpha,\beta)} P_{l_1}^{jki} \right) + \\ + (1-\delta_{l0}) \sum_{(j,k)_i} \sum_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{\text{unique}} a_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{Z_j Z_k Z_i} \sum_{\text{symm.}} \left(\overline{P}_{n_1 j i}^{(\alpha,\beta)} \overline{P}_{n_2 k i}^{(\alpha,\beta)} P_{l_3}^{jki} \sum_{m_1 m_2} {}^{R}G_{m_1 m_2 m}^{l_1 l_2 l} Y_{l_1 m_1}(\hat{\mathbf{r}}_{ji}) Y_{l_2 m_2}(\hat{\mathbf{r}}_{ki}) \right). \end{aligned} \tag{6.38}$$

Please note that the "real" Gaunt coefficients ${}^{R}G_{m_1 m_2 m}^{l_1 l_2 l}$ convey the appropriate coupling of real-spherical harmonics to reach the correct l space.

General recipe for higher-body orders In this short paragraph we outline the procedure to evaluate higher-body order terms. Firstly, the relevant internal coordinates of the scalar function must be identified and written in terms of an atom-centered formalism. For example, in the 4B case, we would have 3 distances (one for each body) and 6 angles (due to the presence of the grid point in the environment). Then a projection on the correct angular momentum space must be carried out, by means of an integration against the appropriate spherical harmonic. This can be always done by expanding each grid-dependent Legendre polynomial in sums of products of spherical harmonics, with the aid of the addition theorem. Finally, care must be taken to eliminate all the possible redundancies, such as that for the $l = 0$ case of the 3B-term above, which could be reduced to the form of a JLP-3B contribution.

6.3.1 The Covariant JL (CJL) model

Finally, we can define the Covariant-JL (CJL) model as

$$T_{lm} = \sum_{i}^{\text{atoms}} T_{i,lm} = \sum_{i} \left(T_{i,lm}^{(1\text{B})} + T_{i,lm}^{(2\text{B})} + T_{i,lm}^{(3\text{B})} + \dots \right), \tag{6.39}$$

where the body order terms are given by Eqs. (6.22), (6.27) and (6.38):

$$\begin{cases} T_{i,lm}^{(1\text{B})} = \delta_{l0}\delta_{m0} a_0^{Z_i}, \\ T_{i,lm}^{(2\text{B})} = \sum\limits_{j,j\neq i} \sum\limits_{n} \left(\delta_{l0} a_{n0}^{Z_j Z_i} \widetilde{P}_{nji}^{(\alpha,\beta)} + (1-\delta_{l0}) a_{nl}^{Z_j Z_i} \overline{P}_{nji}^{(\alpha,\beta)} Y_{lm}(\hat{\mathbf{r}}_{ji}) \right), \\ T_{i,lm}^{(3\text{B})} = \delta_{l0} \sum\limits_{(j,k)_i} \sum\limits_{n_1 n_2 l_1}^{\text{unique}} a_{n_1 n_2 l_1}^{Z_j Z_k Z_i} \sum\limits_{\text{symm.}} \left(\overline{P}_{n_1 ji}^{(\alpha,\beta)} \overline{P}_{n_2 ki}^{(\alpha,\beta)} P_{l_1}^{jki} \right) + \\ \quad + (1-\delta_{l0}) \sum\limits_{(j,k)_i} \sum\limits_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{\text{unique}} a_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{Z_j Z_k Z_i} \sum\limits_{\text{symm.}} \left(\overline{P}_{n_1 ji}^{(\alpha,\beta)} \overline{P}_{n_2 ki}^{(\alpha,\beta)} P_{l_3}^{jki} \sum\limits_{m_1 m_2} {}^{R}G_{m_1 m_2 m}^{l_1 l_2 l} Y_{l_1 m_1}(\hat{\mathbf{r}}_{ji}) Y_{l_2 m_2}(\hat{\mathbf{r}}_{ki}) \right), \\ \dots . \end{cases} \tag{6.40}$$

We remark once again that the coefficients of the expansion do not depend on the magnetic quantum number m, and that, crucially, this model is hierarchical, i.e., we can progressively reach higher-body orders to increase its accuracy. We also note that the model is written, whenever possible, in terms of internal coordinates, since it naturally separates the expansion into products of invariant and covariant contributions. Indeed, while it is true that the covariant terms can still absorb part of the functional dependence on the angles between the atoms, they are mostly devoted to discriminate between different angular momentum space. In other words, by avoiding the mixing of the invariant and covariant factors, one can independently optimise the terms, for example by choosing different hyperparameters between the factor that enforces the covariant behaviour of the expression, and the factor that is devoted to resolve the angular dependency of the expression with respect to the atomic positions. Also, since the target is written in terms of a decomposition in spherical harmonics, it is not possible to cast the expression solely in terms of internal coordinates: our method naturally separates all the terms that can be written in terms of internal coordinate (scalar) from the covariant ones. In this way, it produces a complete and coherent representation of a covariant nB term, even encoding the fact that the covariant properties of an object are not affected by the multiplication by a scalar.

We stress that, since the expansion coefficients depend only on the species of the atoms involved, atoms of the same species share the same coefficients. In this way, the amount of

coefficients to be fitted can be drastically reduced when in the presence of several atoms of the same species. Moreover, we recall that it is not necessary to take into account all the atoms in the sum of Eq. (6.39), or in the multi-body expansions, i.e., if channels connecting two or more atoms are found to have non-significant contributions, they can be neglected from any of the terms of Eq. (6.40). Given the analytical form of the CJL, each term can be manipulated at need, e.g., differentiated or integrated. Also, since the CJL is fully linear, it allows the use of any linear-solver method of choice.

6.4 Applications

At the moment, we have just started to test the full possibilities of the CJL. In particular we are considering the dataset from Ref. [28], where the dipole moments, $\boldsymbol{\mu}$, the polarizability tensor, $\boldsymbol{\alpha}$, and the hyperpolarizability tensor [120], $\boldsymbol{\beta}$, are predicted for water monomers, H_2O, water dimers, $(H_2O)_2$, and Zundel cations, $H_5O_2^+$. While the results are in their early stages, and thus are not reported here, they already show promising accuracy, at least at par with the ones reached by the models of Refs. [28, 77].

6.4.1 The PAW augmentation charges

We have also applied the CJL to the prediction of the projector-augmented-wave (PAW) [121] augmentation charges used in the VASP code [122]. The PAW formalism provides a way to regularize the rapid oscillations of the charge density near the atomic positions. The main idea is that it is beneficial to have a slowly varying behaviour of the charge density, to converge more rapidly to the ground state solution. However, the charge density usually presents steep variations in the vicinity of the atoms. For this reason, the density is first projected onto a well-behaved basis in the immediate atomic proximity, inside "PAW spheres" centered on the atoms. For this representation to be approximately equivalent to an all-electron one, an "augmentation" (or compensation) density must be introduced, which is then defined in terms of an atomic-centered expansion. This strategy is at the core of codes like VASP, and thus one must evaluate also the components of the compensation density at each step of the SC cycles. This was not done in the JLCDM formalism, where the PAW-compensation occupancies (the components of the compensation density) were not taken into account: in this sense, the JLCDM predicts only the "smooth" part of the density, and does not mirror the correct density near the atomic positions. Here, we prove that the CJL is complementary to the JLCDM, as it allows to predict the compensation charges. When both methods are applied together, we can predict the full density.

Our approach aims to target the compensation density around the i-th atom, $\widehat{n}_i(\mathbf{r})$,

defined as

$$\widehat{n}_i(\mathbf{r}) = \sum_{\boldsymbol{v}_1\boldsymbol{v}_2,lm} \rho_{\boldsymbol{v}_1\boldsymbol{v}_2} \widehat{Q}^{lm}_{\boldsymbol{v}_1\boldsymbol{v}_2}(\mathbf{r}), \tag{6.41}$$

where $\boldsymbol{v}_p = (k_p, l_p, m_p)$ is a convenient shorthand for the set of indexes that determine the radial (k_i) and angular (l_i, m_i) expansions, and where the terms $\widehat{Q}^{lm}_{\boldsymbol{v}_1\boldsymbol{v}_2}(\mathbf{r})$ are obtained by constraining the spherical-multipole moments of $\widehat{n}_i(\mathbf{r})$, as will be discussed shortly. Please note that we suppress the index i in the right-hand side for readability. However, even if not explicitly indicated, the origin of the frame of reference will be always placed on top of the i-th atom. The expansion coefficients, $\rho_{\boldsymbol{v}_p\boldsymbol{v}_q}$, are called occupancies. The only requirement for $\hat{n}_i(\mathbf{r})$ is that it must possess the same spherical-multipole moments of the function [43]

$$\overline{n}(\mathbf{r}) = \sum_{\boldsymbol{v}_1\boldsymbol{v}_2} \rho_{\boldsymbol{v}_1\boldsymbol{v}_2} Q_{\boldsymbol{v}_1\boldsymbol{v}_2}(\mathbf{r}), \tag{6.42}$$

with $Q_{\boldsymbol{v}_1\boldsymbol{v}_2}$ defined as

$$Q_{\boldsymbol{v}_1\boldsymbol{v}_2}(\mathbf{r}) = \sum_{\alpha=0}^{1} (-1)^\alpha \phi^*_{\alpha\boldsymbol{v}_1}(\mathbf{r}) \phi_{\alpha\boldsymbol{v}_2}(\mathbf{r}). \tag{6.43}$$

The actual form of the $\phi_{\alpha\boldsymbol{v}}(\mathbf{r})$ functions is unessential for our discussion (please see Refs. [121, 122] for details). However, they can be always written as

$$\phi_{\alpha\boldsymbol{v}}(\mathbf{r}) = c_{\alpha kl} R_{kl}(\mathbf{r}) Y_{lm}(\hat{\mathbf{r}}), \tag{6.44}$$

where we have absorbed all the m dependence in the $\rho_{\boldsymbol{v}_1\boldsymbol{v}_2}$ occupancies. This will be crucial for the development of the method here, since the coefficients $\rho_{\boldsymbol{v}_1\boldsymbol{v}_2}$ transform, under a rotation of the atomic positions, as the product of spherical harmonics $Y_{l_1m_1} Y_{l_2m_2}$. We can now obtain the moments of $Q_{\boldsymbol{v}_1\boldsymbol{v}_2}$, denoted with $q^{lm}_{\boldsymbol{v}_1\boldsymbol{v}_2}$, as[44]

$$\begin{aligned} q^{lm}_{\boldsymbol{v}_1\boldsymbol{v}_2} &:= \int_{\Omega_i} Q_{\boldsymbol{v}_1\boldsymbol{v}_2}(\mathbf{r}) r^l Y_{lm}(\hat{\mathbf{r}}) \mathrm{d}\mathbf{r} \\ &= \sum_\alpha (-1)^\alpha c^*_{\alpha k_1 l_1} c_{\alpha k_2 l_2} \underbrace{\int_0^{r_c} \mathrm{d}r\, r^{l+2} R_{k_1 l_1}(r) R_{k_2 l_2}(r)}_{=: c^l_{k_1 k_2}} \underbrace{\int \mathrm{d}\hat{\mathbf{r}}\, Y_{l_1 m_1}(\hat{\mathbf{r}}) Y_{l_2 m_2}(\hat{\mathbf{r}}) Y_{lm}(\hat{\mathbf{r}})}_{= {}^R G^{l_1 l_2 l}_{m_1 m_2 m}} \\ &= {}^R G^{l_1 l_2 l}_{m_1 m_2 m} c^l_{k_1 k_2} \sum_\alpha (-1)^\alpha c^*_{\alpha k_1 l_1} c_{\alpha k_2 l_2}, \end{aligned} \tag{6.45}$$

where we have used the definition of Eq. (6.35) for the coefficients ${}^R G^{l_1 l_2 l}_{m_1 m_2 m}$, and we have

[43] This function corresponds to the difference $(n^1 - \widetilde{n}^1)$ in the original reference, Ref. [122], Eqs. (24)-(27).

[44] Please, note that the integral are performed on the augmentation spheres around the i-th atom only, indicated by Ω_i.

defined the constants $c^l_{k_1k_2}$ as the moments of the product of the two radial basis. Finally, a function $\hat{n}_i(\mathbf{r})$ that possesses the same moments as $\overline{n}(\mathbf{r})$ can be defined as

$$\hat{n}_i(\mathbf{r}) = \sum_{lm,\boldsymbol{v}_1\boldsymbol{v}_2} \rho^{lm}_{\boldsymbol{v}_1\boldsymbol{v}_2} q^{lm}_{\boldsymbol{v}_1\boldsymbol{v}_2} g_l(r) Y_{lm}(\hat{\mathbf{r}}), \tag{6.46}$$

where $g_l(r)$ are arbitrary functions that must satisfy

$$\int_0^{r_c} g_l(r) r^l \mathrm{d}r = 1.$$

Comparing Eq. (6.46) and Eq. (6.41), we can make the identification

$$\widehat{Q}^{lm}_{\boldsymbol{v}_1\boldsymbol{v}_2}(\mathbf{r}) := q^{lm}_{\boldsymbol{v}_1\boldsymbol{v}_2} g_l(r) Y_{lm}(\hat{\mathbf{r}}). \tag{6.47}$$

Crucially, since $\hat{n}_i(\mathbf{r})$ is a scalar field, we can recognize the covariant nature of the expansion coefficients. Indeed, we can always write

$$\widehat{n}_i(\mathbf{r}) = \sum_{\substack{k_1k_2\\ l_1l_2lm}} \left(\sum_{m_1m_2} \rho^{k_1l_1m_1}_{k_2l_2m_2} q^{lm}_{k_1l_1m_1\ k_2l_2m_2} \right) g_l(r) Y_{lm}(\hat{\mathbf{r}}), \tag{6.48}$$

where the term in brackets transforms as the spherical harmonic Y_{lm} under a rotation of the system of atoms. For this reason, we can target the expansion coefficients

$$\hat{n}^{lm}_{i,k_1k_2l_1l_2} := \sum_{m_1m_2} \rho^{k_1l_1m_1}_{k_2l_2m_2} q^{lm}_{k_1l_1m_1\ k_2l_2m_2}, \tag{6.49}$$

with a CJL model. From Eqs. (6.48) and (6.45), it can be seen that the single terms $\hat{n}^{lm}_{ik_1k_2l_1l_2}$ are not necessarily real but, when performing the sums over l_1 and l_2, the only surviving components are the real ones, so that the final expression for $\hat{n}_i(\mathbf{r})$ is real. Thus, we can neglect any imaginary term: it would be washed out anyway. This also implies that we may restrict our investigation to expansions that are symmetric in the simultaneous exchange $(k_1l_1) \leftrightarrow (k_2l_2)$. With this in mind, and by leveraging on the covariant nature of $\hat{n}^{lm}_{i,k_1k_2l_1l_2}$, we can proceed with the construction of our CJL model.

The easiest approach is to divide our investigation in two subcases: the scalar case, for $l = 0$, and the general case $l > 0$. Importantly, we will re-define the expansion coefficients, making them absorb all the unessential factors, and write directly an expansion for $\hat{n}^{lm}_{i,k_1k_2l_1l_2}$ of Eq.(6.48). Please note that from now on we will use the shorthand notation $\boldsymbol{\mu} := (i, k_1k_2l_1l_2)$, for all the PAW-specific indexes, to write the more compact notation $\hat{n}^{lm}_{i,k_1k_2l_1l_2} := \hat{n}^{lm}_{\boldsymbol{\mu}}$.

Scalar case: $\boldsymbol{l = 0}$. We already showed that the scalar case can be always brought in the form of a JLP model [see Eq. (4.70)]. Thus, the proposed expansion reads[45]

$$\hat{n}^{00}_{\boldsymbol{\mu}} = a^{Z_i}_{\boldsymbol{\mu}} + \sum_{j,j\neq i}\sum_{n} a^{Z_jZ_i}_{\boldsymbol{\mu}n}\widetilde{P}^{(\alpha,\beta)}_{nji} + \sum_{(j,k)_i}\sum_{n_1n_2l'}^{\text{unique}} a^{Z_jZ_kZ_i}_{\boldsymbol{\mu}n_1n_2l'}\sum_{\text{symm.}}\left(\overline{P}^{(\alpha,\beta)}_{n_1ji}\overline{P}^{(\alpha,\beta)}_{n_2ki}P^{jki}_{l'}\right) + \dots . \quad (6.50)$$

Here, all the symmetries of the coefficients have been already discussed at length in Chapter 4. However, we remark that the enforcement of the symmetries on n_1 and n_2, imposed by means of the constrained summation on the "unique" coefficients, must be done for each pair (k_1, l_1). Also, since projecting onto the total space $l = 0$ requires $l_1 = l_2$ (as can be seen from the integrals in Eq. (6.45)), the symmetry under swap $k_1 \leftrightarrow k_2$, i.e., $\hat{n}^{00}_{i,k_1k_2l_1l_1} = \hat{n}^{00}_{i,k_2k_1l_1l_1}$, must be explicitly enforced, namely, by evaluating only the terms for $k_1 \geq k_2$.

General case: $\boldsymbol{l > 0}$ The general case can be written as

$$\begin{aligned}\hat{n}^{lm}_{\boldsymbol{\mu}} &= \sum_{j,j\neq i}\sum_{n} a^{Z_jZ_i}_{\boldsymbol{\mu}nl}\overline{P}^{(\alpha,\beta)}_{nji}Y_{lm}(\hat{\mathbf{r}}_{ji}) + \\ &+ \sum_{(j,k)_i}\sum_{\substack{n_1n_2\\ l'_1l'_2l'_3}}^{\text{unique}} a^{Z_jZ_kZ_i}_{\boldsymbol{\mu}n_1n_2l'_1l'_2l'_3l}\sum_{\text{symm.}}\left(\overline{P}^{(\alpha,\beta)}_{n_1ji}\overline{P}^{(\alpha,\beta)}_{n_2ki}P^{jki}_{l'_3}\sum_{m_1m_2}{}^{R}G^{l'_1l'_2l}_{m_1m_2m}Y_{l'_1m_1}(\hat{\mathbf{r}}_{ji})Y_{l'_2m_2}(\hat{\mathbf{r}}_{ki})\right) + \dots\end{aligned} \quad (6.51)$$

where, again, the symmetry rule for the swap $(k_1l_1) \leftrightarrow (k_2, l_2)$, can be enforced by hand by not evaluating redundant cases with respect to the symmetry $\hat{n}^{lm}_{i,k_2k_1l_2l_1} := \hat{n}^{lm}_{i,k_1k_2l_1l_2}$.

This full expansion allows to target the augmentation occupancies, i.e., the channels $\hat{n}^{lm}_{\boldsymbol{\mu}}$, and, when used in combination with a JLCDM, provides the last piece needed to predict the full electronic density. An application that employs a CJL model to describe the electronic density is under work. While the study is in its final stages, a few tests must still be carried out.

In particular, my role was in the formulation of the models, and so I will not report any partial result here. However, I can mention that we already observed encouraging-high accuracies on an application to the phase transition between the 1H and the 1T phases of a 2D MoS_2 system, despite an expansion up to the 3B terms only.

General approach for operators We conclude this section by mentioning that, guided by the fact that the expansion above does not have any memory of the rotation rule with respect to the uncoupled channels l_1 and l_2, another approach can be carried out. Indeed,

[45] We followed the convention that the last line of the subscripts of the expansion coefficients refers to the set of $\hat{n}$ terms targeted by the model, while every other index refers to the JL expansions.

if we want to preserve the nature of the uncoupled angular momenta, we can instead consider the full expansion [obtained by combining Eqs. (6.45) and (6.49)]

$$\hat{n}^{lm}_{i,k_1k_2l_1l_2} = \sum_{m_1m_2} {}^{R}G^{l_1l_2l}_{m_1m_2m}\rho_{\boldsymbol{v}_1\boldsymbol{v}_2}c^{l}_{k_1k_2}\sum_{\alpha}(-1)^{\alpha}c^{*}_{\alpha k_1l_1}c_{\alpha k_2l_2}. \tag{6.52}$$

It can be shown that, under rotation, the coefficients $\rho_{\boldsymbol{v}_1\boldsymbol{v}_2}$ transform as the product of two independent spherical harmonics: this is because they are obtained by multiplying the expansion coefficients of two orbital functions [as can be read from Eqs. (6.42) and (6.44)]. Thus, we can target them with the double CJL expansion

$$\rho_{\boldsymbol{v}_1\boldsymbol{v}_2} \simeq \rho^{\text{CJL}}_{\boldsymbol{v}_1\boldsymbol{v}_2} := \underbrace{T_{i,\boldsymbol{v}_1}T_{i,\boldsymbol{v}_2}}_{\text{CJL models}}. \tag{6.53}$$

This allows to expand $\rho^{\text{CJL}}_{\boldsymbol{v}_1\boldsymbol{v}_2}$ into body-orders [instead of targeting the full contraction in Eq. (6.52)] as

$$\rho^{\text{CJL}}_{\boldsymbol{v}_1\boldsymbol{v}_2} = \rho^{(1\text{B})}_{\boldsymbol{v}_1\boldsymbol{v}_2} + \rho^{(2\text{B})}_{\boldsymbol{v}_1\boldsymbol{v}_2} + \rho^{(3\text{B})}_{\boldsymbol{v}_1\boldsymbol{v}_2} + \dots, \tag{6.54}$$

where, using the expansion of Eqs. (6.39) and (6.40), we have[46]

$$\begin{cases} \rho^{(1\text{B})}_{\boldsymbol{v}_1\boldsymbol{v}_2} = T^{(1\text{B})}_{i,\boldsymbol{v}_1}T^{(1\text{B})}_{i,\boldsymbol{v}_2}, \\ \rho^{(2\text{B})}_{\boldsymbol{v}_1\boldsymbol{v}_2} = T^{(1\text{B})}_{i,\boldsymbol{v}_1}T^{(2\text{B})}_{i,\boldsymbol{v}_2} + T^{(2\text{B})}_{i,\boldsymbol{v}_1}T^{(1\text{B})}_{i,\boldsymbol{v}_2}, \\ \rho^{(3\text{B})}_{\boldsymbol{v}_1\boldsymbol{v}_2} = T^{(2\text{B})}_{i,\boldsymbol{v}_1}T^{(2\text{B})}_{i,\boldsymbol{v}_2} + T^{(1\text{B})}_{i,\boldsymbol{v}_1}T^{(3\text{B})}_{i,\boldsymbol{v}_2} + T^{(3\text{B})}_{i,\boldsymbol{v}_1}T^{(1\text{B})}_{i,\boldsymbol{v}_2}, \\ \dots. \end{cases} \tag{6.55}$$

This system of equations will not be investigated explicitly. Instead, it must be interpreted as a picture of the expected functional forms, with the Eqs. in (6.40) working as guidelines. Crucially, this method can be generalized to be applied to operators, e.g., in the prediction of Hamiltonians written in terms of atomic basis.

6.5 JL for Fields

We started our construction toward a general JL framework with the definition of a model for scalar quantities (JLP), which was then expanded, by means of a "promotion" to a grid-centered approach, to target also scalar fields (JLCDM). Then, in this chapter, we further generalized the model to predict also tensorial quantities. This was done by constructing an appropriate scalar field that was expanded on the idea of promoting an atom to be a grid point. At this point, we had full control of the harmonic components of

[46] Please, note that one of the body in the expansion is always fixed, being the central atom i.

the scalar fields, which allowed us to define the CJL. The last, perhaps natural, step is to combine the two type of promotions, and define a model that is able to target a general tensorial field. This can be of paramount importance in the study of materials. Like the JLCDM, which was able to accelerate DFT calculations by predicting the electronic density, being able to predict tensorial field allows us to target the magnetization vector, to potentially accelerate the study of non-collinear magnetic materials. Indeed, when describing materials with magnetic properties, the simple density is not enough, and instead we have to consider also the spin density. This is, when warranted, done by means of the introduction of a *vectorial* field, called magnetization, $\boldsymbol{m}(\mathbf{r})$. The magnetization is usually defined from a representation given in terms of the Pauli matrices (please see Ref. [3] for more details)

$$(n(\mathbf{r}), \boldsymbol{m}(\mathbf{r})) \xrightarrow{\text{Pauli mapping}} \begin{pmatrix} n + m_z & m_x - im_y \\ m_x + im_y & n - m_z \end{pmatrix} =: \begin{pmatrix} n^{\uparrow\uparrow} & n^{\uparrow\downarrow} \\ n^{\downarrow\uparrow} & n^{\downarrow\downarrow} \end{pmatrix}, \tag{6.56}$$

i.e., to describe the system we need both a scalar and a vectorial field.

As mentioned before, we have already all the ingredients to fully expand the JL formalism: a Jacobi-Legendre expansion for fields (JLF) can be formally obtained from the CJL model by following the same recipe discussed in the construction of the JLCDM from the JLP, namely, we can formally promote the position of the central atom in Eq. (6.40), $\mathbf{r}_i$, to be a grid point $\mathbf{r}_g$, from a mesh covering the entire space. Thus, the JLF expansion is defined as

$$\begin{cases} T^{(1\mathrm{B})}_{lm}(\mathbf{r}_g) = \sum\limits_{i}^{\text{atoms}} \sum\limits_{n} \left(\delta_{l0} a^{Z_i}_{n0} \widetilde{P}^{(\alpha,\beta)}_{nig} + (1-\delta_{l0}) a^{Z_i}_{nl} \overline{P}^{(\alpha,\beta)}_{nig} Y_{lm}(\hat{\mathbf{r}}_{ig}) \right), \\ T^{(2\mathrm{B})}_{lm}(\mathbf{r}_g) = \delta_{l0} \sum\limits_{(ij)} \sum\limits_{n_1 n_2 l_1}^{\text{unique}} a^{Z_i Z_j}_{n_1 n_2 l_1} \sum\limits_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1 ig} \overline{P}^{(\alpha,\beta)}_{n_2 jg} P^{ijg}_{l_1} \right) + \\ + (1-\delta_{l0}) \sum\limits_{(ij)} \sum\limits_{\substack{n_1 n_2 \\ l_1 l_2 l_3}}^{\text{unique}} a^{Z_i Z_j}_{\substack{n_1 n_2 \\ l_1 l_2 l_3}} \sum\limits_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1 ig} \overline{P}^{(\alpha,\beta)}_{n_2 jg} P^{ijg}_{l_3} \sum\limits_{m_1 m_2} {}^{R}G^{l_1 l_2 l}_{m_1 m_2 m} Y_{l_1 m_1}(\hat{\mathbf{r}}_{ig}) Y_{l_2 m_2}(\hat{\mathbf{r}}_{jg}) \right), \\ \dots . \end{cases} \tag{6.57}$$

This expansion correctly reproduces the JLCDM for the scalar field case, when $l = 0$ [please compare with Eq. (5.8)]. We will conclude this section by showing that the expansion satisfies the correct symmetries.

Given a general tensor field that depends on a point in space, $\mathbf{r}_g$, and, parametrically,

on the atomic positions $\mathbf{r}_i$, the following identities must hold

$$\begin{cases} \boldsymbol{F}(\mathbf{r}_g; \{\hat{t}\mathbf{r}_i\}) = \boldsymbol{F}(\hat{t}^{-1}\mathbf{r}_g; \{\mathbf{r}_i\}), & \text{Translations,} \\ \boldsymbol{F}(\mathbf{r}_g; \{\hat{R}\mathbf{r}_i\}) = \hat{R}\boldsymbol{F}(\hat{R}^{-1}\mathbf{r}_g; \{\mathbf{r}_i\}), & \text{Rotations} \\ \boldsymbol{F}(\mathbf{r}_g; \{\hat{P}\mathbf{r}_i\}) = \hat{P}\boldsymbol{F}(\hat{P}^{-1}\mathbf{r}_g; \{\mathbf{r}_i\}), & \text{Inversions/Parity.} \end{cases} \tag{6.58}$$

for translations, rotations, and inversion, respectively. This is analogous to the properties expressed in Eq. (5.4) for a scalar field, and in Eq. (6.9) for a tensor one. It is more useful, at this point, to write the transformation rules in some representation. Since we rely on a harmonic decomposition of the field, we can write

$$\begin{cases} F_{lm}(\mathbf{r}_g; \{\mathbf{r}_i + \mathbf{t}\}) = F_{lm}(\mathbf{r}_g - \mathbf{t}; \{\mathbf{r}_i\}), & \text{Translations,} \\ F_{lm}(\mathbf{r}_g; \{\hat{R}\mathbf{r}_i\}) = \sum_{m'} {}^R D^{l*}_{mm'}(\mathcal{R}) F_{lm'}(\hat{R}^{-1}\mathbf{r}_g; \{\mathbf{r}_i\}), & \text{Rotation} \\ F_{lm}(\mathbf{r}_g; \{-\mathbf{r}_i\}) = (-1)^l F_{lm}(-\mathbf{r}_g; \{\mathbf{r}_i\}), & \text{Inversion.} \end{cases} \tag{6.59}$$

where we have used the definition given in Eq. (6.13) for the matrices ${}^R D^{l*}_{mm'}(\mathcal{R})$. In particular, the last property is obtained from the analogous transformation for the spherical harmonics $Y_{lm}(-\mathbf{r}) = -Y_{lm}(\mathbf{r})$. To prove that the JLF expansion of Eq. (6.57) follows the correct transformation rules, let us start from the, arguably trivial, translation symmetry.

Translation symmetry All the terms in Eq. (6.57) depend on the position of the atoms with respect to the position of the grid point, i.e., there are only terms of the form $\mathbf{r}_i - \mathbf{r}_g$. Therefore the translation symmetry is proved by noticing that

$$\mathbf{r}_i - (\mathbf{r}_g + \mathbf{t}) = (\mathbf{r}_i - \mathbf{t}) - \mathbf{r}_g. \tag{6.60}$$

Rotation symmetry We can prove that the JLF expansion satisfies the correct transformation rules, by focusing separately on the invariant and on the covariant blocks of each term of the expansion. Indeed, the invariant block being in the same form of the JLCDM, it already satisfies the correct relations between the grid points and the atomic positions, namely, that a rotation of the atomic position corresponds to a counter-rotation of the grid points [please, see Eq.(5.4)]. Instead, the matrices ${}^R D^{l}_{mm'}(\mathcal{R})$ that are required in Eq. (6.59) are introduced by the covariant part[47]. Explicitly, by considering the 3B

[47] We already have proved that the covariant terms transform, globally, as the spherical harmonic Y_{lm}.

case, we can appreciate that the invariant and covariant terms as

$$\underbrace{\overline{P}^{(\alpha,\beta)}_{n_1 ig}\overline{P}^{(\alpha,\beta)}_{n_2 jg}P^{ijg}_{l_3}}_{\text{invariant}}\underbrace{\sum_{m_1 m_2}{}^{R}G^{l_1 l_2 l}_{m_1 m_2 m}Y_{l_1 m_1}(\hat{\mathbf{r}}_{ig})Y_{l_2 m_2}(\hat{\mathbf{r}}_{jg})}_{\text{covariant}}, \tag{6.61}$$

and, since we can always write

$$Y_{lm}(\widehat{\hat{R}\mathbf{r}_i - \mathbf{r}_g}) = Y_{lm}(\hat{R}(\mathbf{r}_i - \widehat{\hat{R}^{-1}\mathbf{r}_g})). \tag{6.62}$$

then the transformation rule of Eq. (6.59) is satisfied.

Inversion symmetry Also the inversion symmetry is proven by focusing only on the covariant block. Here, we will consider only the 3B case showed in Eq. (6.61), and we will outline the proof for all the other cases. Given the symmetry of the spherical harmonics under inversion, we notice that, under parity, a factor $(-1)^{l_1+l_2}$ arises from the covariant block. However, since the coupling terms, ${}^{R}G^{l_1 l_2 l}_{m_1 m_2 m}$, are zero unless the sum $l_1 + l_2 + l$ is even[48], then the sum of the two channels $l_1 + l_2$ must share the same parity with the projected angular momentum channel l, i.e., $(-1)^l = (-1)^{l_1+l_2}$. This is indeed the correct factor required by the last identity of Eq. (6.59). We only mention that the same proof can be applied to higher body order system, where, however, one has to take into account the parity of all the intermediate coupling steps.

6.6 Conclusions

In this chapter we concluded our discussion on the Jacobi-Legendre framework. We showed how to target tensors, operators, and general fields, within a unified and coherent framework. We leveraged the construction done in the previous chapter, where a scalar field was obtained by defining a "grid-point" centered JLP. Here, instead, we used this "promotion" on an atom in the neighborhood, and we proved that the result was, again, a scalar field. We then applied a cluster expansion on the field, with the aim of evaluating its projection on the correct angular momentum space, by means of simple angular integrations. As a check, we also proved that the scalar case, for $l = 0$, correctly reduced to a JLP.

Crucially, this chapter finally showed the benefits of the choice of the Legendre polynomials, introduced as soon as Chapter 4, for the formalism of JLP models: in particular we heavily exploited their expansion in products of spherical harmonics, which had a pivotal

[48] We recall that this can be seen from the CG coefficient $C^{l0}_{l_1 0 l_2}$ in Eq. (6.36), which vanishes unless $l_1 + l_2 + l$ is even.

role in evaluating the angular integrals. A cumbersome part of the derivation was to keep track of the relation between the complex spherical harmonics and the real ones: from this point of view, this chapter can be used as a bridge between the two representations, with all the transformations laid out. Indeed, it can be noted how the expressions are written so that the two formalism can be exchanged without really changing the relevant formulas.

Unfortunately, we were not able to report direct numerical applications, mostly due to the fact that studies are still ongoing. Nevertheless, we proceeded to present a possible application for the PAW formalism: this, together with the JLCDM, will allow us to predict the full electronic density, including also the rapidly oscillating behaviour around the atomic positions.

We closed this chapter by presenting the model for general fields, here called JLF. We proved that the model defined is, on the one hand, hierarchical and systematically improvable with respect to the cluster order, and on the other hand it satisfies the correct transformation rules under the action of any arbitrary rotation and/or inversion of the system investigated.

With both the CJL and the JLF, this chapter is crucial to fully define the JL framework, which otherwise would have been left incomplete. We showed how to navigate between all the different models, in going from a fully scalar one to a formalism for tensor fields. This is the main achievement of my thesis and is fully reported in Fig. 6.2: we graphically show there the complete JL formalism, alongside its core ideas and strategies.

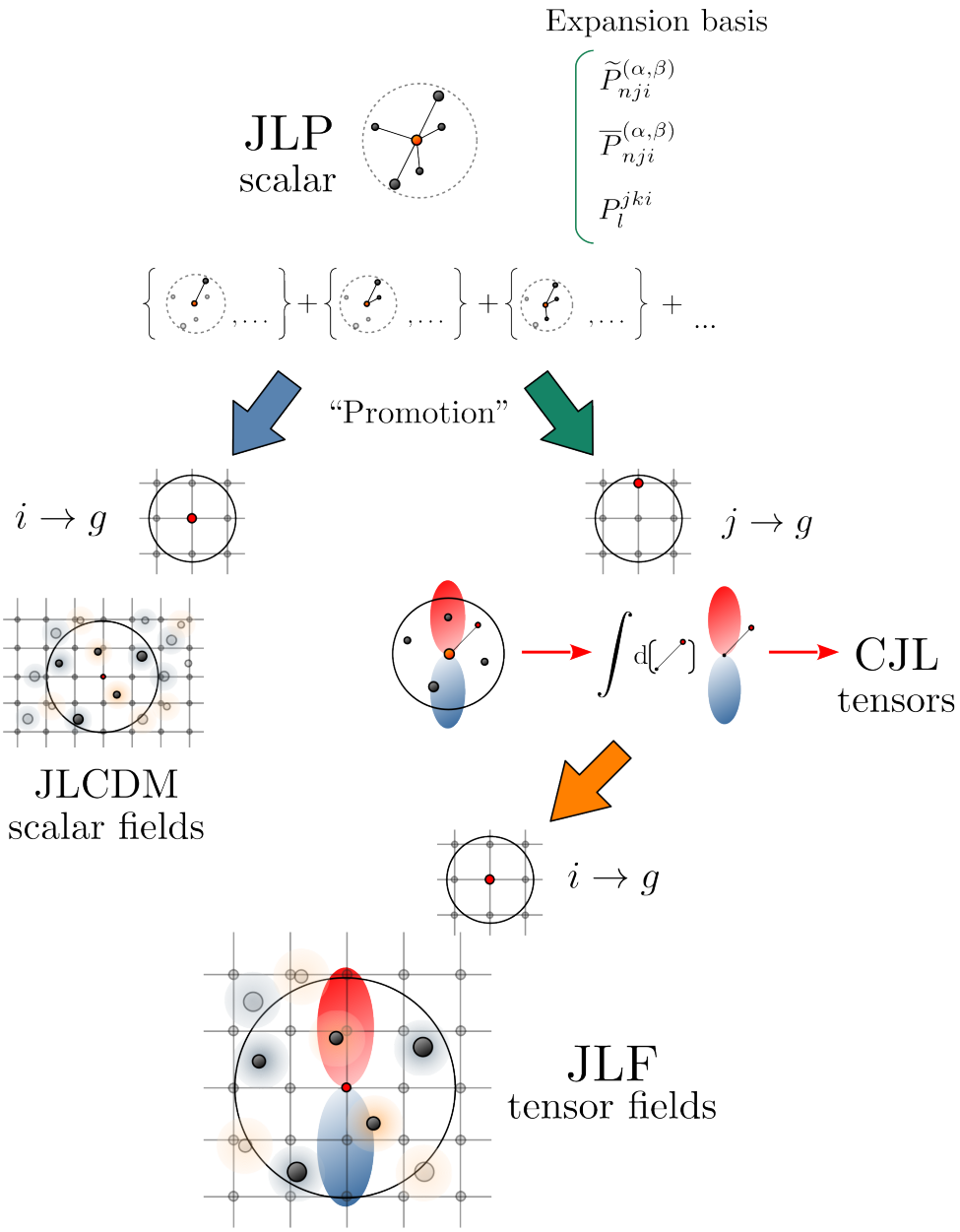

Figure 6.2: **The Jacobi-Legendre framework.** (From top to bottom) With the definition of the model for any general field we can now draw the graphical scheme for the full Jacobi-Legendre framework. It starts with the JLP and the cluster expansion in terms of vanishing- and double-vanishing-Jacobi polynomials. The expression obtained can be expanded to the case of a scalar field, by allowing one of the atoms to be, instead, a grid point. In the case in which the grid point takes the role of the central atom, we obtain the descriptors for scalar fields used in the JLCDM. Instead, if we allow one of the atoms in the neighbourhood to be the grid point, we obtain a scalar field that satisfies simple integration rules when its harmonic components are evaluated. This drives us to the definition of the CJL expansion for tensorial quantities introduced and explored in this chapter. Finally, we can again perform a "promotion" of the central atom to a grid point to obtain a model for general vector fields, decomposed in their harmonic representation, here called JLF.

Chapter 7

Multipolar expansion and the five body case

In this final chapter we will present a slightly different point of view for the definition of rotationally invariant quantities, based on the formalism of the multipolar-spherical harmonics. We have already exploited the property of the bipolar-spherical harmonics of Eq. 2.22 in defining the bispectrum, and in Section 3.2 for the construction of the powerspectrum from vectorial fields. In these cases, we hinted how the multipolar-spherical harmonics constitute a natural formalism for the treatment of multi-body MLPs. Here, we will show that using the multipolar-spherical harmonics allows us to obtain the desired symmetries in a seamless way, i.e., by means of a simple selection of the appropriate components of the basis. We will proceed in explicitly deriving the expansions up to the 5B term, here treated as the expansion of 4-point functions defined on the unitary sphere. In particular, this last case will introduce a discussion on the choice of coupling schemes and on the completeness of the ACE framework. By leveraging the multipolar-spherical harmonics we will expose a limitation in the basis used in the original ACE work, which will be used as a case study to show how the formalism can be used to derive and investigate the expansion of multi-body functions. Finally, this chapter will also be tied with the JL formalism, and we will show an alternative to the scheme induced by the coupling of angular momentum induced by the JLP expansion. Indeed, we will obtain expressions that not only are isometrically invariant and complete, but also, that do not depend on any coupling scheme. We will show how this allows for a simple treatment of the invariance under permutation of identical atoms.

The chapter will be structured as follows: we will first proceed in formally expanding multi-point functions, defined on the unitary sphere, in terms of multipolar-spherical harmonics, and we will show how to incorporate the rotational and parity invariance in the expression. We will then re-derive the powerspectrum and the bispectrum couplings,

as the natural coupling schemes induced by the rotationally invariant components of the multipolar spherical harmonics. Moreover, we will show how the 4-point functions require the explicit consideration of intermediate coupling channels, which are instead contracted in the original ACE. With the founding formalism in place, we will derive a general expression for an ACE model, which will is complete at any body order. This will be followed by a proof of the incompleteness of the original ACE formalism for any body-order larger than four. Again, the proof will heavily rely on the properties of the multipolar spherical harmonics. In particular, we will show how two different 5B potentials can be described by the same 5B terms of the ACE framework.

We will close this chapter by tying the discussion with the 5B expansion proposed for the JLP models. We will then present a new expansion that, in contrast to the general ACE or other approaches that require the coupling of angular momentum channels, will not only be complete and isometrically invariant, but also independent from any coupling scheme of angular momenta, despite being symmetric under permutation of identical atoms.

7.1 The Multipolar Spherical Harmonics

This section will introduce a formalism for multipolar spherical harmonics, which are implicitly used in all the descriptors based on the powerspectrum and on the bispectrum, and in particular in the multi-body expansions of the ACE potential. The approach will be the similar to the one adopted in Chapter 3.2, where the bipolar-spherical harmonics were first introduced. Also, a few of the properties that will be re-derived here are listed in Ref. [45].

7.1.1 Bipolar Spherical Harmonics

Let us consider a function F, depending on two versors $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$, such that it satisfies the expansion

$$F(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2) = \sum_{l_1 m_1 l_2 m_2} f_{l_1 m_1 l_2 m_2} Y_{l_1}^{m_1}(\hat{\mathbf{r}}_1) Y_{l_2}^{m_2}(\hat{\mathbf{r}}_2), \tag{7.1}$$

with the coefficients $f_{l_1 m_2 l_2 m_2}$ given by

$$f_{l_1 m_1 l_2 m_2} = \int \mathrm{d}\hat{\mathbf{r}}_1 \mathrm{d}\hat{\mathbf{r}}_2 \, F(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2) Y_{l_1}^{m_1 *}(\hat{\mathbf{r}}_1) Y_{l_2}^{m_2 *}(\hat{\mathbf{r}}_2). \tag{7.2}$$

We can now use the same procedure adopted to derive the spin-powerspectrum in

Chapter 3.2, based on the CG orthogonality (reported here for readability)

$$\sum_{l=|l_1-l_2|}^{l_1+l_2}\sum_{m=-l}^{l} C^{lm}_{l_1m_1l_2m_2}C^{lm}_{l_1m_1'l_2m_2'} = \delta_{m_1m_1'}\delta_{m_2m_2'}, \tag{7.3}$$

so that, the expansion of F can be re-written as

$$\begin{aligned} F(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2) &= \sum_{\substack{l_1m_1m_1'\\ l_2m_2m_2'}} f_{l_1m_1'l_2m_2'}\left(\sum_{lm} C^{lm}_{l_1m_1l_2m_2}C^{lm}_{l_1m_1'l_2m_2'}\right) Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1)Y^{m_2}_{l_2}(\hat{\mathbf{r}}_2)\\ &= \sum_{l_1l_2lm}\left(\sum_{m_1'm_2'} C^{lm}_{l_1m_1'l_2m_2'}f_{l_1m_1'l_2m_2'}\right)\left(\sum_{m_1m_2} C^{lm}_{l_1m_1l_2m_2}Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1)Y^{m_2}_{l_2}(\hat{\mathbf{r}}_2)\right)\\ &= \sum_{l_1l_2lm}\mathcal{F}^{lm}_{l_1l_2}\mathcal{Y}^{lm}_{l_1l_2}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2), \end{aligned} \tag{7.4}$$

where we have introduced the terms

$$\begin{cases} \mathcal{F}^{lm}_{l_1l_2} := \sum_{m_1m_2} C^{lm}_{l_1m_1l_2m_2}f_{l_1m_1l_2m_2},\\ \mathcal{Y}^{lm}_{l_1l_2}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2) := \sum_{m_1m_2} C^{lm}_{l_1m_1l_2m_2}Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1)Y^{m_2}_{l_2}(\hat{\mathbf{r}}_2). \end{cases} \tag{7.5}$$

We have then just performed a change of basis. This shows how the *bipolar-spherical harmonics*, $\mathcal{Y}^{lm}_{l_1l_2}$, form indeed a basis for the two-points functions on the unitary sphere. The bipolar-spherical harmonics are also orthonormal, as can be easily seen from

$$\begin{aligned} &\int \mathrm{d}\hat{\mathbf{r}}_1\mathrm{d}\hat{\mathbf{r}}_2\,\mathcal{Y}^{lm}_{l_1l_2}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2)\mathcal{Y}^{l'm'*}_{l_1'l_2'}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2) =\\ &= \sum_{\substack{m_1m_2\\ m_1'm_2'}} C^{lm}_{l_1m_1l_2m_2}C^{l'm'}_{l_1'm_1'l_2'm_2'}\underbrace{\int \mathrm{d}\hat{\mathbf{r}}_1\,Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1)Y^{m_1'*}_{l_1'}(\hat{\mathbf{r}}_1)}_{\delta_{l_1l_1'}\delta_{m_1m_1'}}\underbrace{\int \mathrm{d}\hat{\mathbf{r}}_2\,Y^{m_2}_{l_2}(\hat{\mathbf{r}}_2)Y^{m_2'*}_{l_2'}(\hat{\mathbf{r}}_2)}_{\delta_{l_2l_2'}\delta_{m_2m_2'}}\\ &= \delta_{l_1l_1'}\delta_{m_1m_1'}\delta_{l_2l_2'}\delta_{m_2m_2'}\underbrace{\sum_{m_1m_2} C^{lm}_{l_1m_1l_2m_2}C^{l'm'}_{l_1m_1l_2m_2}}_{\delta_{ll'}\delta_{mm'}} = \delta_{l_1l_1'}\delta_{m_1m_1'}\delta_{l_2l_2'}\delta_{m_2m_2'}\delta_{ll'}\delta_{mm'}. \end{aligned} \tag{7.6}$$

This is obtained by using the orthogonality of the spherical harmonics and of the Clebsh-Gordan (CG) coefficients. The bipolar spherical harmonics can be intuitively seen as the basis for the expansion of the function F in the space of coupled angular momentum (l,m). This holds since we have proved in Eq. (2.22) that a bipolar-spherical harmonic of order (l,m) follows the same rotation rules of the spherical harmonic Y^m_l under *simultaneous* rotation of the two versors $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$. This was also crucial in the

derivation of the spin-powerspectrum in Chapter 3.2. Now, if we assume that $\mathcal{F}^{lm}_{l_1l_2} = 0$ unless $l = m = 0$, then the function F behaves like a scalar for a simultaneous rotation of its arguments, which is exactly what we want for a MLP. In this cases since $C^{00}_{l_1m_1l_2m_2} = \delta_{l_1l_2}\delta_{m_1,-m_2}(-1)^{l_1-m_1}/\sqrt{2l_1+1}$, we have

$$\mathcal{Y}^{00}_{l_1l_2}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2) = \delta_{l_1l_2}\frac{(-1)^{l_1}}{\sqrt{2l_1+1}}\sum_{m_1}(-1)^{m_1}Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1)Y^{-m_1}_{l_1}(\hat{\mathbf{r}}_2), \tag{7.7}$$

and so the expansion of the function F reads

$$F_{\text{scalar}}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2) = \sum_l a_l \mathcal{Y}^{\text{scalar}}_l(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2), \tag{7.8}$$

where we have defined $a_l := (-1)^l\mathcal{F}^{00}_{ll}$, and $\mathcal{Y}^{\text{scalar}}_l := \mathcal{Y}^{00}_{l_1l_2}$ (please note that the factor $(-1)^l$ is irrelevant with respect to the orthogonality, and can be removed by a trivial unitary transformation). We stress here that F has only scalar (invariant under rotations) components. If we compare this expression with Eq. (2.10), we immediately recognise the powerspectrum coupling rule, here reported for readability, and adapted to the simplified case of no radial contributions [as done, for example in the introduction of the powerspectrum in Ref. [27]]

$$p_l := \sum_m (-1)^m c_{ilm}c_{il-m} = \sum_{jk}^{\text{atoms}}\sum_{m=-l}^{l}(-1)^m Y^m_l(\hat{\mathbf{r}}_{ji})Y^{-m}_l(\hat{\mathbf{r}}_{ki}). \tag{7.9}$$

We also note that a linear expansion of the function in powerspectrum components is exactly the formulation of the angular terms of the 3B ACE of Eqs. (2.59) and (2.62), reported here in detail

$$\begin{aligned}\varepsilon^{(3)}_i &:= \sum_{n_1n_2l}^{\text{ordered}} a_{n_1n_2l}B^{(3)}_{in_1n_2l} = \sum_{n_1n_2l}^{\text{ordered}} a_{n_1n_2l}\sum_{m=-l}^{l}(-1)^m A_{in_1lm}A_{in_2l-m} \\ &= \sum_{n_1n_2l}^{\text{ordered}} a_{n_1n_2l}\sum_{jk}^{\text{atoms}} R_{n_1l}(r_{ji})R_{n_2l}(r_{ki})\sum_{m=-l}^{l}(-1)^m Y^m_l(\hat{\mathbf{r}}_{ji})Y^m_l(\hat{\mathbf{r}}_{ki}).\end{aligned} \tag{7.10}$$

Indeed, we can already anticipate that this can be interpreted as the expansion of atomic potentials, $v^{(3)}$ [already introduced for the ACE model in Eq. (2.51), and for the JLP in Eq. (4.2)], in terms of the rotationally-invariant components of the bipolar spherical harmonics, as

$$v^{(3)}(\mathbf{r}_{ji}, \mathbf{r}_{ki}) = \sum_{n_1 n_2 l} a_{n_1 n_2 l} R_{n_1 l}(r_{ji}) R_{n_2 l}(r_{ki}) \mathcal{Y}_l^{\text{scalar}}(\hat{\mathbf{r}}_{ji}, \hat{\mathbf{r}}_{ki}), \tag{7.11}$$

which leads to the 3B-atomic energies by means of the relation[49].

$$\varepsilon_i^{(3)} = \sum_{jk}^{\text{atoms}} v^{(3)}(\mathbf{r}_{ji}, \mathbf{r}_{ki}). \tag{7.12}$$

This explicitly shows that the bipolar spherical harmonics are the natural basis for this expansion. In the following, we will focus only on functions that not only behave as scalars under reflection, but are also invariant under parity (so that they are invariant under any isometry). The case above is the only case in which this is always verified, since the powerspectrum coupling depends only on the scalar product $\hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_2$, [as shown in Eq. (2.15)]. In general, however, the invariance under inversion must be enforced, as it will be clear in the following sections.

7.1.2 Tripolar Spherical Harmonics

Following the same argument for the bipolar-spherical harmonics, we can now take into account a three point function F, with an expansion that reads

$$F(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3) = \sum_{\substack{l_1 m_1 \\ l_2 m_2 \\ l_3 m_3}} f^{l_1 m_1}_{l_2 m_2 \, l_3 m_3} Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1) Y^{m_2}_{l_2}(\hat{\mathbf{r}}_2) Y^{m_3}_{l_3}(\hat{\mathbf{r}}_3). \tag{7.13}$$

By exploiting again the orthogonality of the CG coefficients, this time twice, we can re-write the expansion as

$$F(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3) = \sum_{l_1 l_2 l_{12} l_3 l m} \mathcal{F}^{lm}_{(l_1 l_2) l_{12} l_3} \mathcal{Y}^{lm}_{(l_1 l_2) l_{12} l_3}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3), \tag{7.14}$$

with

$$\begin{cases} \mathcal{F}^{lm}_{(l_1 l_2) l_{12} l_3} := \displaystyle\sum_{\substack{m_1 m_2 m_3 \\ m_{12}}} C^{lm}_{l_3 m_3 l_{12} m_{12}} C^{l_{12} m_{12}}_{l_1 m_1 l_2 m_2} f^{l_1 m_1}_{l_2 m_2 \, l_3 m_3}, \\ \mathcal{Y}^{lm}_{(l_1 l_2) l_{12} l_3}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3) := \displaystyle\sum_{\substack{m_1 m_2 m_3 \\ m_{12}}} C^{lm}_{l_3 m_3 l_{12} m_{12}} C^{l_{12} m_{12}}_{l_1 m_1 l_2 m_2} Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1) Y^{m_2}_{l_2}(\hat{\mathbf{r}}_2) Y^{m_3}_{l_3}(\hat{\mathbf{r}}_3). \end{cases} \tag{7.15}$$

Here, the *tripolar-spherical harmonics*, $\mathcal{Y}^{lm}_{(l_1 l_2) l_{12} l_3}$ are defined by means of a chosen

[49] Please, note that we are not yet explicitly investigating how to obtain the ordered indexes rule necessary to enforce the invariance under atom permutations. This will be discussed in subsequent sections.

coupling scheme. They form an orthonormal basis set for three-point functions defined on the surface of the sphere, in strict analogy to the bipolar-spherical harmonics. For the scheme chosen above, we have first coupled the spherical harmonics depending on $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$. Then, the resulting object was coupled with the remaining spherical harmonic $Y_{l_3}^{m_3}(\hat{\mathbf{r}}_3)$. The choice of a coupling scheme implies a representation degeneracy, formally equivalent to the one obtained in Eq. (2.70) for the determination of the spherical components of a rank 3 tensor. We will now show how this degeneracy can be removed in the case of isometrically invariant functions. However, let us first project on the rotationally invariant space (constraining the expression to $l = m = 0$): the tripolar-spherical harmonics of interest are given by

$$\mathcal{Y}^{00}_{(l_1l_2)l_{12}l_3}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2,\hat{\mathbf{r}}_3) = (-1)^{l_1+l_2+l_3}\delta_{l_{12}l_3}\sum_{m_1m_2m_3}\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} Y_{l_1}^{m_1}(\hat{\mathbf{r}}_1)Y_{l_2}^{m_2}(\hat{\mathbf{r}}_2)Y_{l_3}^{m_3}(\hat{\mathbf{r}}_3), \tag{7.16}$$

where we have used the following relation between the CG coefficients and the well-known 3j-Wigner symbols [45]

$$C^{lm}_{l_1m_1l_2m_2} = (-1)^{-l_1+l_2-m}\sqrt{2l+1}\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & -m_3 \end{pmatrix}. \tag{7.17}$$

Again, we find that this coupling scheme is formally equivalent to the one introduced for the bispectrum, in Eq. (2.19) or, equivalently, the one for $B^{(4)}$ in the ACE coupling, Eq. (2.59), and reported here

$$B^{(4)}_{i\,\substack{n_1n_2n_3\\ l_1l_2l_3}} = \sum_{m_1m_2m_3}\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} A_{in_1l_1m_1}A_{in_2l_2m_2}A_{in_3l_3m_3}. \tag{7.18}$$

Moreover, the linear expansion of the function F now reads

$$F_{\text{scalar}}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2,\hat{\mathbf{r}}_3) = \sum_{l_1l_2l_3} a_{l_1l_2l_3}(-1)^{l_1+l_2+l_3}\sum_{m_1m_2m_3}\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} Y_{l_1}^{m_1}(\hat{\mathbf{r}}_1)Y_{l_2}^{m_2}(\hat{\mathbf{r}}_2)Y_{l_3}^{m_3}(\hat{\mathbf{r}}_3), \tag{7.19}$$

with the expansion coefficients defined as $a_{l_1l_2l_3} := \mathcal{F}^{00}_{(l_1l_2)l_3l_3}$. From this expansion, and from the symmetries of the 3j-Wigner symbols, we can see how different coupling schemes lead to the same expressions. Indeed, the 3j symbols are invariant under cyclic permutations of columns, and acquire the factor $(-1)^{l_1+l_2+l_2}$ under an odd permutation, e.g.,

$$\begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} = \begin{pmatrix} l_3 & l_1 & l_2 \\ m_3 & m_1 & m_2 \end{pmatrix} = (-1)^{\sum_l}\begin{pmatrix} l_2 & l_1 & l_3 \\ m_2 & m_1 & m_3 \end{pmatrix}. \tag{7.20}$$

Here we have introduced the useful shorthand notation $\sum_l = l_1 + l_2 + l_3$ to indicate the summation over all the angular momentum channels. In terms of the choice of the coupling scheme, this means that all the six sets of tripolar spherical harmonics (one for each coupling scheme) are always equivalent up to a sign when projected on the rotational invariant space. In other words, if we choose a coupling scheme that preserves the cyclic order of the indexes we will end up with an expression equivalent to $\{\mathcal{Y}^{00}_{(l_1l_2)l_{12}l_3}\}$, while if we change this ordering we will obtain $\{(-1)^{\sum_l}\mathcal{Y}^{00}_{(l_1l_2)l_{12}l_3}\}$. However, a change in the ordering/coupling scheme is not the only operation that introduces a factor $(-1)^{\sum_l}$. Indeed, because of the symmetry under parity of the spherical harmonics, here reported for readability

$$Y^m_l(-\hat{\mathbf{r}}) = (-1)^l Y^m_l(\hat{\mathbf{r}}), \tag{7.21}$$

we have that

$$\mathcal{Y}^{00}_{(l_1l_2)l_{12}l_3}(-\hat{\mathbf{r}}_1, -\hat{\mathbf{r}}_2, -\hat{\mathbf{r}}_3) = (-1)^{\sum_l}\mathcal{Y}^{00}_{(l_1l_2)l_{12}l_3}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3). \tag{7.22}$$

This means that an inversion causes the atomic basis to undergo the same factor of a non-cyclic permutation of the angular momenta in the coupling scheme. We immediately deduce that, if F is also invariant under parity, then its expansion will be constrained to cases such that the sum $l_1 + l_2 + l_3$ is even. Explicitly

$$F_{\text{iso}}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3) = \sum_{l_1l_2l_3}^{\sum_l=\text{even}} a_{l_1l_2l_3}\mathcal{Y}^{\text{scalar}}_{l_1l_2l_3}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3), \tag{7.23}$$

where we have defined

$$\mathcal{Y}^{\text{scalar}}_{l_1l_2l_3}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3) := \sum_{m_1m_2m_3} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} Y^{m_1}_{l_1}(\hat{\mathbf{r}}_1)Y^{m_2}_{l_2}(\hat{\mathbf{r}}_2)Y^{m_3}_{l_3}(\hat{\mathbf{r}}_3), \tag{7.24}$$

for $l_1+l_2+l_3$ even. Given the symmetry rules of the 3j symbols, an important by-product is that we restricted the expansion to cases in which the coupling coefficients are *independent* from any coupling scheme.

We have just proved that the expansion of a three-point-invariant function is provided by the tripolar-spherical harmonics, which inherit the completeness and orthogonality from the expansion in spherical harmonics. However, again, they yield a practical partition of the angular momentum space that allows us to seamlessly select the rotationally invariant space. In other words, an isometrically invariant function can be expanded as a linear combination of the orthonormal basis $\{\mathcal{Y}^{\text{scalar}}_{l_1l_2l_3}\}$, which is independent of any coupling scheme of angular momenta, when appropriately constrained. In doing so, we have also appreciated a connection between the parity invariance and the selection of couplings coefficients, which are symmetric under different choices of coupling schemes.

7.1.3 Quadrupolar Spherical Harmonics

In this section we will continue our progression and introduce an expansion for four-point functions. This case is relevant, since it is the first example where the choice of the coupling scheme matters, as will be shown by the explicit (and irreducible) presence of the intermediate channel in the final expressions.

The starting point is the expansion of a four-point function $F(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3, \hat{\mathbf{r}}_4)$ over spherical harmonics

$$F(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3, \hat{\mathbf{r}}_4) = \sum_{\substack{l_1 l_2 l_3 l_4 \\ m_1 m_2 m_3 m_4}} f^{\,l_1 m_1}_{\substack{l_2 m_2 \\ l_3 m_3 \\ l_4 m_4}} Y_{l_1}^{m_1}(\hat{\mathbf{r}}_1) Y_{l_2}^{m_2}(\hat{\mathbf{r}}_2) Y_{l_3}^{m_3}(\hat{\mathbf{r}}_3) Y_{l_4}^{m_4}(\hat{\mathbf{r}}_4). \tag{7.25}$$

As for the previous cases, by choosing a coupling scheme and appropriately introducing CG coefficients, we can re-cast the above expansion in terms of the *total* angular momentum channels (l, m) as

$$F(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3, \hat{\mathbf{r}}_4) = \sum_{lm} \sum_{\substack{l_1 l_2 l_3 l_4 \\ l_{12} l_{34}}} \mathcal{F}^{lm}_{(l_1 l_2) l_{12} (l_3 l_4) l_{34}} \mathcal{Y}^{lm}_{(l_1 l_2) l_{12} (l_3 l_4) l_{34}}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3, \hat{\mathbf{r}}_4), \tag{7.26}$$

where we have introduced the *quadrupolar-spherical harmonics*

$$\begin{aligned} &\mathcal{Y}^{lm}_{(l_1 l_2) l_{12} (l_3 l_4) l_{34}}(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3, \hat{\mathbf{r}}_4) \\ &:= \sum_{\substack{m_1 m_2 m_3 m_4 \\ m_{12} m_{34}}} C^{lm}_{l_{12} m_{12} l_{34} m_{34}} C^{l_{34} m_{34}}_{l_3 m_3 l_4 m_4} C^{l_{12} m_{12}}_{l_1 m_1 l_2 m_2} Y_{l_1}^{m_1}(\hat{\mathbf{r}}_1) Y_{l_2}^{m_2}(\hat{\mathbf{r}}_2) Y_{l_3}^{m_3}(\hat{\mathbf{r}}_3) Y_{l_4}^{m_4}(\hat{\mathbf{r}}_4). \end{aligned} \tag{7.27}$$

The adopted coupling scheme can be read by the order of the indexes of the CG coefficients, which shows which angular momentum channels are connected: firstly, we perform the coupling between the first two channels, l_1 and l_2, into l_{12}, as can be read from $C^{l_{12} m_{12}}_{l_1 m_1 l_2 m_2}$. Then, analogously, we couple the third and fourth channels, l_3 and l_4, in l_{34}, represented by $C^{l_{34} m_{34}}_{l_3 m_3 l_4 m_4}$. Finally, we couple l_{12} and l_{34}, so that the full expression is projected into the desired space, l, by means of $C^{lm}_{l_{12} m_{12} l_{34} m_{34}}$. Clearly, the coupling scheme here is much more relevant than in the previous case, since different coupling schemes, albeit spanning the same space, could lead to very different expansions. We remark that the quadripolar-spherical harmonics form an orthonormal basis for the four-point functions, as it can be easily verified by direct integration. We now project again into the rotationally invariant, $l = m = 0$, and consider only the scalar components of the

function F. The relevant quadripolar-spherical harmonics read

$$\begin{aligned}\mathcal{Y}^{\text{scalar}}_{(l_1l_2)l(l_3l_4)}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2,\hat{\mathbf{r}}_3,\hat{\mathbf{r}}_4) &:= \mathcal{Y}^{00}_{(l_1l_2)l_{12}(l_3l_4)l_{34}}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2,\hat{\mathbf{r}}_3,\hat{\mathbf{r}}_4)\\ &= \delta_{l_{12}l_{34}}\frac{(-1)^{l_{12}}}{\sqrt{2l_{12}+1}}\sum_{\substack{m_1m_2m_3m_4\\ m_{12}}}(-1)^{m_{12}}C^{l_{12}m_{12}}_{l_1m_1l_2m_2}C^{l_{12}-m_{12}}_{l_3m_3l_4m_4}Y^{m_1}_{l_1}Y^{m_2}_{l_2}Y^{m_3}_{l_3}Y^{m_4}_{l_4}.\end{aligned} \tag{7.28}$$

Please note that we have used the short hand $Y^{m_i}_{l_i} := Y^{m_i}_{l_i}(\hat{\mathbf{r}}_i)$, i.e., we will conveniently label with the same conventions of versors of the function. Also, to simplify the notation, we will re-label l_{12} and m_{12} with l and m, respectively, and we define $\mathcal{Y}^{\text{scalar}}_{(l_1l_2)l(l_3l_4)} := \mathcal{Y}^{00}_{(l_1l_2)l(l_3l_4)l}$. The expansion for the function F_{scalar}, then becomes

$$F_{\text{scalar}}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2,\hat{\mathbf{r}}_3,\hat{\mathbf{r}}_4) = \sum_{l_1l_2l_3l_4l}\mathcal{F}_{(l_1l_2)l(l_3l_4)}\mathcal{Y}^{\text{scalar}}_{(l_1l_2)l(l_3l_4)}(\hat{\mathbf{r}}_1,\hat{\mathbf{r}}_2,\hat{\mathbf{r}}_3,\hat{\mathbf{r}}_4), \tag{7.29}$$

with appropriate expansion coefficients $\mathcal{F}_{(l_1l_2)l(l_3l_4)}$. Finally, we can appreciate how, again, an inversion of the coordinates leads to the emergence of a factor $(-1)^{\sum_l}$. If we constrain the expansion to the case in which the sum $\sum_l$ is even, then we will again obtain a suitable basis for an isometrically invariant function.

Before proceeding with our investigation on the four-point functions, let us summarize the results obtained so far, with the aim of applying this formalism to the ACE framework. Firstly, we proved that if we have a multi-point function defined on a sphere, then we can expand it in multipolar spherical harmonics, which constitute a natural (and orthonormal) basis of choice. We also saw that, in order to constrain the expressions to be rotationally invariant, it suffices to consider only the $l = m = 0$ components of the total angular momentum of the multipolar spherical harmonics, justifying the use of this basis. Also, to impose parity invariance, we need to discard all the spherical harmonics with odd $\sum_l$, retaining only the even ones. We showed that, the more arguments that the function has, the more important becomes the choice of a coupling scheme for the definition of the multipolar spherical harmonics. Finally, we found that the rotationally invariant components of the bipolar-spherical harmonics are also the symmetric ones under inversion. In contrast, we had to impose the condition $\sum_l$ already for the tripolar case. As a by-product, this constraint removed the dependence of the expression from a specific coupling scheme. However, the coupling scheme becomes relevant for higher-point terms.

Let us apply this findings to the ACE case. We will re-derive the ACE formalism from the point of view of a multipolar decomposition of the atomic energies, and we will show how this approach provides insights on the incompleteness of the ACE-angular basis. Finally, we will tie back to the discussion with the JLP formalism, and we will show that it provides a way of defining a coupling-scheme independent expansion for the quadrupolar case.

7.2 An ACE Framework

In general, the rotationally invariant components of a v-polar spherical harmonic, can be written in a compact form as

$$\mathcal{Y}^{\text{scalar}}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}) := \mathcal{Y}^{00}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}) = \sum_{\boldsymbol{mM}} H^{\boldsymbol{lL}}_{\boldsymbol{mM}} \prod_{k=1}^{v} Y^{m_i}_{l_i}(\hat{\mathbf{r}}_i), \tag{7.30}$$

where we have introduced a few useful shorthand notations to make the expressions more readable. Here, the vector $\hat{\boldsymbol{R}} = (\hat{\mathbf{r}}_1, \dots \hat{\mathbf{r}}_v)$ contains all the versors of the the system, while the vectors $\boldsymbol{l} = (l_1, \dots l_v)$ and $\boldsymbol{m} = (m_1, \dots, m_v)$ contain all the indexes associated with the spherical harmonics. Finally, the vectors $\boldsymbol{L} = (L_1, \dots, L_{v-2})$ and $\boldsymbol{M} = (M_1, \dots, M_{v-2})$ represent all the intermediate angular momentum channels that arise from the coupling (if any are needed). Please note that we are not explicitly indicating any particular coupling scheme: we will not discuss the relation between different coupling schemes, we will only assume that a coupling scheme is fixed for each v. The coefficients $H^{\boldsymbol{lL}}_{\boldsymbol{mM}}$ are the core of the definition of the multipolar spherical harmonics, containing the correct product of CG coefficients need to project the expression into the space of zero angular momentum. The examples showed in the previous sections are, explicitly,

$$\begin{cases} H^{l_1 l_2}_{m_1 m_2} = \delta_{l_1 l_2}\delta_{m_1 -m_2}\dfrac{(-1)^{l_1+m_1}}{\sqrt{2l_1+1}} & \text{for } v=2, \\ H^{l_1 l_2 l_3}_{m_1 m_2 m_3} = \dfrac{(-1)^{l_3-m_3}}{\sqrt{2l_3+1}} C^{l-m_3}_{l_1 m_1 l_2 m_2} & \text{for } v=3, \\ H^{(l_1 l_2)L(l_3 l_4)}_{(m_1 m_2)M(m_3 m_4)} = \dfrac{(-1)^{L-M}}{\sqrt{2L+1}} C^{LM}_{l_1 m_1 l_2 m_2} C^{L-M}_{l_3 m_3 l_4 m_4} & \text{for } v=4. \end{cases} \tag{7.31}$$

With these choices, the multipolar spherical harmonics form a complete and orthonormal set for the rotationally-invariant v-points functions on the unitary sphere.

Let us now consider the cluster expansion of the atomic energies ε_i given in Eq. (2.51),

here reported for convenience

$$\varepsilon_i = \varepsilon_i^{(1)} + \sum_{j}^{\text{atoms}} v^{(2)}(\mathbf{r}_{ji}) + \sum_{jk}^{\text{atoms}} v^{(3)}(\mathbf{r}_{ji}, \mathbf{r}_{ki}) + \sum_{jkp}^{\text{atoms}} v^{(4)}(\mathbf{r}_{ji}, \mathbf{r}_{ki}, \mathbf{r}_{pi}) + \dots, \tag{7.32}$$

where we have already released the constraints on the sum over the atoms, by the same trick discussed in Section 2.1.4. We can see how the $(v+1)$-body term in the sum is in the form $v^{(v+1)}(\boldsymbol{R}_{\boldsymbol{j}i})$, with the definition

$$\boldsymbol{R}_{\boldsymbol{j}i} := (R_{\boldsymbol{j}i}, \hat{\boldsymbol{R}}_{\boldsymbol{j}i}) = (r_{j_1 i}, r_{j_2 i}, \dots r_{j_v i}, \hat{\mathbf{r}}_{j_1 i}, \hat{\mathbf{r}}_{j_2 i}, \dots \hat{\mathbf{r}}_{j_v i}),$$

and $\boldsymbol{j} := (j_1, \dots, j_v)$. This means that we can expand the potential in terms of a radial basis for the distances and v-polar-spherical harmonics for the angles. Explicitly

$$v^{(v+1)}(\boldsymbol{R}_{\boldsymbol{j}i}) = \sum_{\boldsymbol{nlL}}^{\sum_l = \text{even}} a_{\boldsymbol{nlL}} \left[\prod_{k=1}^{v} R_{n_k l_k}(r_{j_k i}) \right] \mathcal{Y}^{\text{scalar}}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}_{\boldsymbol{j}i}), \tag{7.33}$$

where we have already used the components of zero angular momentum, and we have restricted the sum to even $\sum_l$. In this way, the potential satisfies, almost by construction, the required transformation symmetries. Please note that, in order to have a complete expansion, the coefficients must also depend on the intermediate angular channels $\boldsymbol{L}$: this was explicitly shown for the expansion in terms of the quadrupolar-spherical harmonics [Eq. (7.29)]. We can now evaluate the $(v+1)$-body contribution to the atomic energy ε_i, here indicated with $\varepsilon_i^{(v)}$, as

$$\begin{aligned} \varepsilon_i^{(v+1)} &= \sum_{\boldsymbol{j}} v^{(v+1)}(\boldsymbol{R}_{\boldsymbol{j}i}) = \sum_{\boldsymbol{j}} \sum_{\boldsymbol{nlL}}^{\sum_l = \text{even}} a_{\boldsymbol{nlL}} \left[\prod_{k=i}^{v} R_{n_k l_k}(r_{j_k i}) \right] \mathcal{Y}^{\text{scalar}}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}_{\boldsymbol{j}i}) \\ &= \sum_{\boldsymbol{nlL}}^{\sum_l = \text{even}} a_{\boldsymbol{nlL}} \sum_{\boldsymbol{mM}} H^{\boldsymbol{lL}}_{\boldsymbol{mM}} \prod_{k=1}^{v} \sum_{j_k}^{\text{atoms}} R_{n_k l_k}(r_{j_k i}) Y^{m_i}_{l_i}(\hat{\mathbf{r}}_{j_k i}), \end{aligned} \tag{7.34}$$

where, in the second line, we have used the definition of Eq. (7.30) for $\mathcal{Y}^{\text{scalar}}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}_{\boldsymbol{j}i})$, and we have distributed the sum over $\boldsymbol{j}$ in the products. If we now use the atomic basis introduced in the ACE formalism, and reported here

$$A_{i n_k l_k m_k} := \sum_{j_k}^{\text{atoms}} R_{n_k l_k}(r_{j_k i}) Y^{m_i}_{l_i}(\hat{\mathbf{r}}_{j_k i}), \tag{7.35}$$

and define the invariant products

$$B^{(v+1)}_{\boldsymbol{inlL}} = \sum_{\boldsymbol{mM}} H^{\boldsymbol{lL}}_{\boldsymbol{mM}} \prod_{k=1}^{v} A_{in_k l_k m_k}, \tag{7.36}$$

we can finally derive an ACE framework, directly from the multipolar spherical harmonics, as

$$\varepsilon^{(v+1)}_i = \sum_{\boldsymbol{nlL}}^{\sum_l=\text{even}} a_{\boldsymbol{nlL}} B^{(v+1)}_{\boldsymbol{inlL}}. \tag{7.37}$$

Not only is this expression completely general, but crucially, it also shows that, once a specific radial basis is fixed, the invariant products $B^{(v+1)}_{\boldsymbol{inlL}}$ are fully determined by the coupling coefficients $H^{\boldsymbol{lL}}_{\boldsymbol{mM}}$ of the v-polar-spherical harmonics. However, the formalism proposed here is not the same as the original ACE, which will be discussed in the following section.

7.3 Incompleteness of the original ACE representation

We can now compare the expression obtained above with the original ACE model, of Ref [25]. We will consider only cases for $v \geq 4$, since they are the ones requiring the explicit definition of intermediate angular momenta $\boldsymbol{L}$, as it can be appreciated from the last Eq. in (7.31). As such, these are also the cases in which the choice of the coupling scheme matters. In particular, the work in Ref. [66] has been devoted to the removal of the redundancies and degeneracies arising from selecting a particular scheme in the case $v = 4$.

The framework proposed above, in Eq. (7.37), is different from the proposed one in the original ACE [25]. Indeed, by comparing the full scheme presented here with the one in the last line of Eq. (2.59) and with the coupling terms introduced in Eq. (2.61), we obtain that the original ACE is

$$\varepsilon^{(v+1),\text{ACE}}_i = \sum_{\boldsymbol{nl}}^{\sum_l=\text{even}} a_{\boldsymbol{nl}} B^{(v+1),\text{ACE}}_{\boldsymbol{inl}}, \tag{7.38}$$

namely, the expansion coefficients are assumed to not depend on the intermediate channels $\boldsymbol{L}$, and the invariant products are defined in terms of the sum

$$B^{(v+1),\text{ACE}}_{\boldsymbol{inl}} := \sum_{\boldsymbol{L}} B^{(v+1)}_{\boldsymbol{inlL}}. \tag{7.39}$$

In the original ACE formalism, this approach is shared by all the $v \geq 4$ cases, where all the intermediate channels are always contracted. We will now prove that this coupling scheme is incomplete[50]. The crucial point of the proof is that we can trace back all the steps from Eq. (7.30) to Eq. (7.37). Thus, it can be seen that, by assuming that the expansion coefficients do not depend on the intermediate angular momenta, we are implicitly preforming an expansion of the form

$$v^{(v+1)}(\boldsymbol{R}_{ji}) = \sum_{\boldsymbol{nl}}^{\sum_l=\text{even}} a_{\boldsymbol{nl}} \left[\prod_{k=1}^{v} R_{n_k l_k}(r_{j_k i})\right] \sum_{\boldsymbol{L}} \mathcal{Y}^{\text{scalar}}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}_{ji}). \tag{7.40}$$

Here, the sum over the intermediate angular momenta have been transferred over the v-polar-spherical harmonics, since they are the only terms that maintain their explicit dependence on the intermediate channels. Clearly, this is equivalent to using an "ACE-angular" basis defined as

$$\mathcal{Y}^{\text{ACE}}_{\boldsymbol{l}}(\hat{\boldsymbol{R}}_{ji}) := \sum_{\boldsymbol{L}} \mathcal{Y}^{\text{scalar}}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}_{ji}). \tag{7.41}$$

By construction, using this basis to expand the angular part of a potential $v^{(v+1)}(\hat{\boldsymbol{R}}_{ji})$ leads directly to the original ACE formalism of Eq. (7.38). In order to prove that this expansion cannot describe an arbitrary potential $v^{(v+1)}$, let us study the properties of the ACE-angular basis.

Properties of the ACE-angular basis Firstly, let us consider the inner product

$$\langle \mathcal{Y}^{\text{ACE}}_{\boldsymbol{l}} | \mathcal{Y}^{\text{ACE}}_{\boldsymbol{l'}} \rangle = \int \mathcal{Y}^{\text{ACE}}_{\boldsymbol{l}}(\hat{\boldsymbol{R}}) \mathcal{Y}^{\text{ACE}*}_{\boldsymbol{l'}}(\hat{\boldsymbol{R}})\, \mathrm{d}\hat{\boldsymbol{R}}, \tag{7.42}$$

where

$$\int \mathrm{d}\boldsymbol{R} := \int \mathrm{d}\hat{\mathbf{r}}_1 \ldots \mathrm{d}\hat{\mathbf{r}}_n, \tag{7.43}$$

i.e., we are integrating each variable over the respective solid angle. Since the v-polar spherical harmonics are orthonormal, the integral can be easily evaluated as[51]

$$\int \mathcal{Y}^{\text{ACE}}_{\boldsymbol{l}}(\hat{\boldsymbol{R}}) \mathcal{Y}^{\text{ACE}*}_{\boldsymbol{l'}}(\hat{\boldsymbol{R}})\, \mathrm{d}\hat{\boldsymbol{R}} = \sum_{\boldsymbol{LL'}} \underbrace{\int \mathcal{Y}^{\text{scalar}}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}) \mathcal{Y}^{\text{scalar}*}_{\boldsymbol{l'L'}}(\hat{\boldsymbol{R}})\, \mathrm{d}\hat{\boldsymbol{R}}}_{=\delta_{\boldsymbol{ll'}}\delta_{\boldsymbol{LL'}}} = \delta_{\boldsymbol{ll'}} \sum_{\boldsymbol{L}} \{\boldsymbol{lL}\}, \tag{7.44}$$

[50] In the sense that not all the v-points functions can be described by this choice of the coupling.
[51] We use the generalization of the Kronecker delta, so that $\delta_{\boldsymbol{LL'}} := \delta_{L_1 L'_1} \cdot \ldots \cdot \delta_{L_{v-2} L'_{v-2}}$

where the term $\{\boldsymbol{lL}\}$ is 1 if the intermediate angular momenta $\boldsymbol{L}$ are consistent with the initial channels $\boldsymbol{l}$ and the chosen coupling scheme, i.e., if they belong to the tree generated by $\boldsymbol{l}$, and 0 otherwise. This shows that the ACE-angular basis is orthogonal but not normalized. In practice, the square of the normalization constant, defined as

$$N_{\boldsymbol{l}}^2 := \delta_{\boldsymbol{ll'}} \sum_{\boldsymbol{L}} \{\boldsymbol{lL}\}, \tag{7.45}$$

counts how many intermediate channels lead to the rotationally invariant space, from the initial $\boldsymbol{l}$ and with the fixed coupling scheme.

Another crucial property is obtained by looking at a way to write the multipolar spherical harmonics in terms of the ACE-angular basis. In practice, we look for expansion coefficients[52] $U_{\boldsymbol{lL}}^{\boldsymbol{l'}}$ such that

$$\mathcal{Y}_{\boldsymbol{lL}}^{\text{scalar}}(\hat{\boldsymbol{R}}) = \sum_{\boldsymbol{l'}} U_{\boldsymbol{lL}}^{\boldsymbol{l'}} \mathcal{Y}_{\boldsymbol{l'}}^{\text{ACE}}(\hat{\boldsymbol{R}}). \tag{7.46}$$

The coefficients are obtained by means of the integral

$$\begin{aligned} U_{\boldsymbol{lL}}^{\boldsymbol{l'}} &= \frac{1}{N_{\boldsymbol{l'}}^2} \int \mathcal{Y}_{\boldsymbol{lL}}^{\text{scalar}}(\hat{\boldsymbol{R}}) \mathcal{Y}_{\boldsymbol{l'}}^{\text{ACE}*}(\hat{\boldsymbol{R}}) \, \mathrm{d}\hat{\boldsymbol{R}} \\ &= \frac{1}{N_{\boldsymbol{l'}}^2} \sum_{\boldsymbol{L'}} \underbrace{\int \mathcal{Y}_{\boldsymbol{lL}}^{\text{scalar}}(\hat{\boldsymbol{R}}) \mathcal{Y}_{\boldsymbol{l'L'}}^{\text{scalar}*}(\hat{\boldsymbol{R}}) \, \mathrm{d}\hat{\boldsymbol{R}}}_{=\delta_{\boldsymbol{ll'}}\delta_{\boldsymbol{LL'}}} = \frac{1}{N_{\boldsymbol{l}}^2} \delta_{\boldsymbol{ll'}}, \end{aligned} \tag{7.47}$$

which holds for every well-defined multipolar-spherical harmonic[53]. This means that, in general, we can write v-polar-spherical harmonics as

$$\mathcal{Y}_{\boldsymbol{lL}}^{\text{scalar}}(\hat{\boldsymbol{R}}) = \frac{1}{N_{\boldsymbol{l}}^2} \mathcal{Y}_{\boldsymbol{l}}^{\text{ACE}}(\hat{\boldsymbol{R}}). \tag{7.48}$$

We now have all the ingredients required for the aforementioned proof.

Incompleteness of the original ACE-angular basis and the importance of the intermediate coupling channels We will focus only on a v-point isometrically-invariant function, F, defined on the unitary sphere. Indeed, the radial part does not play any role here, and the following can be easily generalized to the case of general function in space by assuming that the coefficients depend on the relevant distances. We already know that the function F can always be expanded in terms of v-polar-spherical harmonics projected on the isometrically invariant space as,

[52] The expansion coefficients must be components of a unitary matrix.

[53] This is because, for the multipolar-spherical harmonic to be defined, $\boldsymbol{L}$ must belong to the tree generated by $\boldsymbol{l}$.

$$F(\hat{\boldsymbol{R}}) = \sum_{\boldsymbol{lL}}^{\sum_l=\text{even}} \mathcal{F}_{\boldsymbol{lL}} \mathcal{Y}^{\text{scalar}}_{\boldsymbol{lL}}(\hat{\boldsymbol{R}}), \tag{7.49}$$

where $\mathcal{F}_{\boldsymbol{lL}}$ are the expansion coefficients. It is crucial to notice that, to fully characterize the function F, we need all the expansion coefficients, namely we cannot ignore the dependence on the intermediate channels $\boldsymbol{L}$. We can now use the property of Eq. (7.47) and obtain

$$F(\hat{\boldsymbol{R}}) = \sum_{\boldsymbol{lL}}^{\sum_l=\text{even}} \mathcal{F}_{\boldsymbol{lL}} \frac{1}{N_l^2} \mathcal{Y}^{\text{ACE}}_{l}(\hat{\boldsymbol{R}}) = \sum_{\boldsymbol{l}}^{\sum_l=\text{even}} \frac{1}{N_l^2} \left[\sum_{\boldsymbol{L}} \mathcal{F}_{\boldsymbol{lL}} \right] \mathcal{Y}^{\text{ACE}}_{l}(\hat{\boldsymbol{R}}). \tag{7.50}$$

The expression above shows that the ACE-angular basis cannot discriminate between functions for which the sum of the expansion coefficients in the square brackets is the same. Equivalently, we can say that functions such that

$$\sum_{\boldsymbol{L}} \mathcal{F}_{\boldsymbol{lL}} = 0, \tag{7.51}$$

cannot be represented by the ACE-angular basis. Thus, if we sum an arbitrary function, F, and a function that cannot be represented by the ACE-angular basis, F^{null}, then, the new function $G := F + F^{\text{null}}$ will have the same ACE representation of F, while being generally different. Also, it is not difficult to construct functions for which the sum above vanishes. For instance, this is obtained by choosing random values for the coefficients, and then by imposing that one of them is the opposite of the sum of all the others. This is a simple strategy to generate such a non-representable function.

The immediate consequence of this proof is that a general potential cannot be completely characterized by the expansion proposed in Eq. (7.40), i.e., washing out the dependence of the expansion coefficients on the intermediate angular momentum eliminates non-reducible degrees of freedom.

In particular, this implies that the original ACE formalism is not able to properly describe the atomic energy terms

$$\varepsilon_i^{(v+1)} = \sum_{\boldsymbol{j}} v^{(v+1)}(\boldsymbol{R}_{\boldsymbol{j}i}), \tag{7.52}$$

for $v \geq 4$, since we could add any non-representable potential [such that the property of Eq. (7.51) holds] and obtain the same value of the atomic energy. This is similar to the discussion in Ref. [27], reported in Eq. (2.17), where it was observed that the powerspectrum alone was not able to describe atomic environment, since different functions may possess the same powerspectrum.

Before proceeding to the next and conclusive section of this Chapter, we remark again that this proof concerns the original ACE formalism, of Ref. [25]. Indeed, please note that more recent works (see, for example Refs. [38, 103]), use a complete representation, preserving the intermediate channels that appear in the coupling.

However, even the complete representation is not without problems, as we will see in the next section.

7.4 A 5B representation without angular-momentum couplings

We will now discuss the problem of enforcing the invariance of the atomic energy under permutation of identical atoms: this appears as a severe downside of the complete expansion in Eq. (7.37). Indeed, while the ACE-angular basis is not complete, it allows for a simple treatment of atomic permutations. This can be easily understood if we focus on the fact that each pair n_k, l_k refers to an atomic basis: imposing this invariance is equivalent to requiring that the order of the atomic basis in the products of Eq. (7.36) does not affect the atomic energy of Eq. (7.37). In other words, exchanging two (n_k, l_k) pairs must lead to the same result. This can be easily imposed if the coupling coefficients are not dependent on the intermediate angular momentum channels, as was the case for the ACE formalism. With this assumption, the symmetry would be imposed by the only requirement that the expansion coefficients $a_{\boldsymbol{nl}}$ are symmetric under the swap of any (n_k, l_k) pairs. This leads to an expansion in which only lexicographically ordered pairs are considered. However, in the presence of a complete representation, where also the intermediate channels $\boldsymbol{L}$ matter, imposing this constraint is more complicated, since exchanging two channels l_k and $l_{k'}$, in general, produces a different coupling scheme. This can be seen with a specific example: if the coupling scheme is such that it first imposes the coupling between the first two channels, $l_1 + l_2 \to L_{12}$, and then the third and fourth ones, $l_3 + l_4 \to L_{34}$, then swapping l_2 with l_3 will cause a change in the coupling scheme, i.e., we are now to couple the first channel with the third one and the second with the fourth. Unfortunately, in general, different coupling schemes produce different results, albeit spanning the same space. This problem exists already for the $\nu = 4$ case, has discussed in Ref. [66], where a selection of only non redundant contributions was performed algorithmically, by using the generalized Wigner symbols. Here, however, we will focus on a completely different approach, strictly tied, again, to the JL framework. This will bypass completely the need for the coupling

of angular momenta[54].

In section 4.1.4 we have already proved that a representation over internal coordinates, with 4 distances and 6 angles, is complete and irreducible. In particular, in Eq. (4.65), we saw that we can write the atomic potentials in terms of double-vanishing-Jacobi and Legendre polynomials as

$$v^{(5)}(\boldsymbol{R}_{ji}) = \sum_{\substack{n_1 n_2 n_3 n_4 \\ l_1 l_2 l_3 l_4 l_5 l_6}}^{\text{unique}} a^{\,n_1 n_2 n_3 n_4}_{l_1 l_2 l_3 l_4 l_5 l_6} \sum_{\text{symm.}} \left(\overline{P}^{(\alpha,\beta)}_{n_1 ji} \overline{P}^{(\alpha,\beta)}_{n_2 ki} \overline{P}^{(\alpha,\beta)}_{n_3 pi} \overline{P}^{(\alpha,\beta)}_{n_4 qi} P^{jki}_{l_1} P^{jpi}_{l_2} P^{jqi}_{l_3} P^{kpi}_{l_4} P^{kqi}_{l_5} P^{pqi}_{l_6} \right), \tag{7.53}$$

where the sum over unique coefficients and on the symmetries has been discussed in detail in Chapter 4. We can notice that the expansion over unique coefficients is exactly what we were looking for: it enforces the invariance of the expression, at the potentials' level, under the permutation of two identical atoms. We will not discuss explicitly the full chain of equivalences for the coefficients, which consists of 24 possible ways to swap the indexes. An example is the case in which we swap the first two atoms, the j-th and k-th ones. This implies that the two coefficients

$$a^{\,n_1 n_2 n_3 n_4}_{l_1 l_2 l_3 l_4 l_5 l_6} = a^{\,n_2 n_1 n_3 n_4}_{l_1 l_4 l_5 l_2 l_3 l_6} \tag{7.54}$$

must be equal. By exchanging all the possible atoms, it is possible to obtain the complete set of 24 equivalences.

We have already discussed, in Chapter 4, how to linearize the JLP with respect to the number of atoms in the cut-off sphere. This was accomplished by applying the addition theorem,

$$P_l^{jki} = \frac{4\pi}{2l+1} \sum_m Y_l^m(\hat{\mathbf{r}}_{ji}) Y_l^{m*}(\hat{\mathbf{r}}_{ki}), \tag{7.55}$$

to every Legendre polynomials of the expansion. We can do the same for the 5B expansion above. We do not report the full expression, which will not add anything to the discussion. Instead we just mention that it consists of a sum over $m_1, \ldots, m_6$, and of products of 12 spherical harmonics. If we now factorize together all the terms that refer to the same

[54]Here, we mean that the coupling of different channels will not be manifest, since the expressions will be naturally symmetric and irreducible. This is caused by the fact that the representation in internal coordinates is done in terms of angles and distances and, in this sense, the underlying coupling scheme is constructed in terms of the resulting angles and not on the resulting invariant properties. This makes the treatment much simpler, since the angles are naturally invariant, while a versor-based approach requires us to construct the invariance intentionally.

atom, and we evaluate the sum[55]

$$\varepsilon_i^{(5)} = \sum_{jkpq} v^{(5)}(\boldsymbol{R}_{ji}), \tag{7.56}$$

it can be shown that the resulting expression can be written in terms of the JL-atomic basis, already defined in Eq. (4.32) and reported here for readability,

$$(J_1 L_3)^i_{n\,l_1 l_2 l_3}_{m_1 m_2 m_3} := \sum_j^{\text{atoms}} \overline{P}^{(\alpha,\beta)}_{nji} Y^{m_1}_{l_1}(\hat{\mathbf{r}}_{ji}) Y^{m_2}_{l_2}(\hat{\mathbf{r}}_{ji}) Y^{m_3}_{l_4}(\hat{\mathbf{r}}_{ji}). \tag{7.57}$$

After some manipulation, the energy $\varepsilon_i^{(5)}$ can be written as a sum of contracted products of four JL-atomic basis, namely

$$\begin{aligned}\varepsilon_i^{(5)} = &\sum_{\substack{n_1 n_2 n_3 n_4 \\ l_1 l_2 l_3 l_4 l_5 l_6}}^{\text{unique}} b_{l_1 l_2 l_3 l_4 l_5 l_6}^{n_1 n_2 n_3 n_4} \times \\ &\times \sum_{\substack{m_1 m_2 m_3 \\ m_4 m_5 m_6}} (-1)^{\sum_m} (J_1 L_3)^i_{n_1\,l_1 l_2 l_3}_{m_1 m_2 m_3} (J_1 L_3)^i_{n_2\,l_1 l_4 l_5}_{-m_1 m_4 m_5} (J_1 L_3)^i_{n_3\,l_2 l_4 l_6}_{-m_2 -m_4 m_6} (J_1 L_3)^i_{n_4\,l_3 l_5 l_6}_{-m_3 -m_5 -m_6}.\end{aligned} \tag{7.58}$$

Here the coefficients b are defined in terms of the coefficients a by absorbing unessential factors obtained from the addition theorem of Eq. (7.55). We also defined the convenient shorthand $\sum_m = m_1 + \ldots + m_6$. We can notice that the sum over the possible symmetries has been lost from Eq. (7.53): this is caused by the summation over all possible atoms, which is equivalent to the sum over all possible swaps of indexes, and allows us to retain only the sum over the *unique* coefficients. This is in the same spirit of the lexicographic ordering of the ACE formalism, but with the main difference that, while finding equivalent coefficients is less straightforward, the expansion above is now complete. Crucially, not only is the linear scaling guaranteed by the fact that the evaluation of the JL-atomic basis of Eq. (7.57) is linear in the number of atoms inside the cut-off radius, but also all the other symmetries are enforced by construction, the representation in internal coordinates being isometrically invariant. Indeed, we have thus bypassed the need of a coupling scheme with all the redundancies and degeneracies. This was indirectly achieved by introducing a more complicated atomic basis, and by increasing the number of indexes to consider to 6: one more than that considered in the work on the generalized Wigner symbols of Ref. [66], and 2 more than that of the ACE-angular basis approach. We can make two observations: firstly, the evaluation of the JL-atomic basis does not require more than the evaluation

[55] Please note that we are following the ACE framework, in which the sum over the atoms is not restricted. The arising of self-energy term contributions have been already analysed in Chapter 4.

of all the spherical harmonics and the radial basis. Indeed, the remaining operations consist in simple products of the already computed values of the spherical harmonics. The bottleneck is then transferred on the coupling scheme, which can, nevertheless, be tackled in progressive contractions, with the aim of minimizing the number of calculations required. Secondly, we argue that the presence of more indexes is not necessarily detrimental for the expansion. Indeed, it is safe to assume that more indexes allow one to spread the information on more basis terms, and so, arguably, a smaller truncation parameter $l_{\max}$ could be used. Since the overhead is mostly caused by higher values of angular momenta, it also means that we should have a trade-off between the number of contractions and the highest value of the angular momentum reached. While this is still an hypothesis, we will focus our effort in investigating these speculations in future works.

In conclusion, the proposed expansion of Eq. (7.58), for the 5B-atomic energies, not only is complete, but it does not suffer from the irreducible redundancies caused by the choice of a coupling scheme, on one hand, and the necessity of enforcing the permutational invariance of identical atoms on the other. While a similar expansion for the 6B is not yet available (the representation in terms of the internal coordinates is over-complete when all the possible angles are taken into account), we hope that the strategy introduced here, which shifts the focus from the pure coupling of angular momenta to a more angle-based approach, combined with the versatility of an approach based on the multipolar-spherical harmonics, could lead to efficient approaches in the study of descriptors for isometrically invariant quantities.

Conclusions

Summary

This thesis focused on the study of descriptors for linear ML models applied to the study of materials. We explored the main ideas that drive the construction of the models, ranging from the encoding of the correct symmetries to strategies to simplify otherwise computationally expensive methods. We presented a model for the treatment of magnetic systems, able to bring on the same footing the atomic positions and the spin degrees of freedom. This was tested on a toy-system characterized by spin-lattice coupling with transversal and longitudinal spin excitations.

The central part of the thesis has been devoted for the definition of the JL framework, which encompasses all the descriptors based on the JL formalism. Here, we showed how the formalism is constructed, with the choice of the polynomial basis, the constraining procedure, the enforcement of symmetries and the generalization to the description of tensors and tensor fields. While all the descriptors are atomic, local and defined in terms of a cluster expansion, the main aspect lies in the choice of the internal coordinates. Indeed we proved that the choice of the internal coordinates is not restrictive. On the contrary, we proved that the JL formalism can be written in a form that scales linearly with the number of atoms, while preserving the natural interpretability given by a representation in terms of internal coordinates. Thus, the central part of the thesis has been devoted to the extension of JL models, by presenting simple strategies for the construction of covariant quantities. Crucially, we showed how JL models for potentials (JLP) and for the electronic density (JLCDM) can reach high accuracy levels and could be used in accelerating DFT calculations.

The last chapter has been devoted to showing how the multipolar spherical harmonics constitute a useful framework in the study of multi-body descriptors, allowing for an in-depth study of symmetries and completeness by virtue of their orthogonality relationships. Finally, by investigating five-body order terms, we showed how an internal-coordinate based formalism led to a complete representation, possessing all the required symmetry properties.

Future plans

In this final section we will discuss the next steps of our investigation, the new extensions of the JL formalism and some tests and applications that will be performed.

Spin-JLP An expansion of the JL formalism will aim to include spin degrees of freedom in a JLP. This is one natural direction for our work, combining the study on the spin-powerspectrum of Chapter 3 and the JLP formalism of Chapter 4, in terms of an internal coordinates representation.

Application of the JLP, pushing the limits on multi-component systems A significant part of our future effort will be devoted to test JLPs on diverse material systems. While this is necessary in itself, to gauge the applicability and limits, special effort will be dedicated to study multi-component systems, in the spirit of reducing the combinatorial scaling implied in a species-basis cluster expansion.

Introduction of long-range interaction While a discussion of long-range interactions was beyond the scope of the thesis, it is important to test the descriptors on systems in which the interactions are not only short-ranged. An extension of the JL formalism to include also these cases is not yet clear, but we aim to dedicate our efforts in pursuing such generalizations.

Application to quantum transport With the availability of a model to predict the electronic density, a possible subsequent step is to design an application for quantum transport. This requires the evaluation of non-equilibrium charge density of an open system. We are working on an expansion of the JLCDM to be applied to the SIESTA code [123, 124]. This should allow to accelerate quantum-transport calculations performed within the SMEAGOL [125] code framework (that relies on SIESTA), to avoid converging the density at each step of the SMEAGOL iterations.

Application of the CJL (In production) With the covariant formalism provided by the CJL, we are working to predict tensorial quantities of interest (such as the one provided by Ref. [75]). The result will be important for analyzing the performance and, eventually, the shortcomings of the formalism, in particular in comparison with other schemes available.

Prediction of the PAW-augmentation charges (In preparation) As mentioned in the main text, the JLCDM is not enough to fully predict the electronic density of DFT codes that rely on the PAW formalism. We are working on finalizing a project for the

prediction of the PAW-augmentation charges by means of a CJL model, and to use them to further accelerate DFT calculations.

Application of the JLF to spin-polarized electronic densities and magnetization vectors An important aspect of the JL framework lies in its generalization to tensor fields, which is still a largely unexplored territory of ML applied to electronic-structure calculations. In this sense, we will soon start explore the acceleration of DFT calculations in non-collinearly spin-polarized cases, expanding the JLCDM approach to some of the most challenging, and computationally heavy, tasks of DFT codes.

Coupling-scheme independent 5B framework (In preparation) We are working to finalize the study on the multipolar spherical harmonics applied to ML descriptors. We will further proceed in evaluating the advantages and disadvantages of using the proposed JLP-based linear expansion for 5B terms.

Bibliography

[1] P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas", Phys. Rev. **136**, B864 (1964).

[2] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects", Phys. Rev. **140**, A1133 (1965).

[3] R. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, 2004).

[4] R. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, Density-functional Theory of Atoms and Molecules (Oxford University Press, USA, 1994).

[5] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, "Neural network models of potential energy surfaces", J. Chem. Phys. **103**, 4129 (1995).

[6] M. Minotakis, H. Rossignol, M. Cobelli, and S. Sanvito, "Machine-learning surrogate model for accelerating the search of stable ternary alloys", Phys. Rev. Mater. **7**, 093802 (2023).

[7] J. Nelson and S. Sanvito, "Predicting the Curie temperature of ferromagnets using machine learning", Phys. Rev. Mater. **3**, 104405 (2019).

[8] J. Zhang, Z. Zhu, X.-D. Xiang, K. Zhang, S. Huang, C. Zhong, H.-J. Qiu, K. Hu, and X. Lin, "Machine Learning Prediction of Superconducting Critical Temperature through the Structural Descriptor", J. Chem. Phys. C **126**, 8922 (2022).

[9] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning", Phys. Rev. Lett. **108**, 058301 (2012).

[10] A. Lunghi and S. Sanvito, "A unified picture of the covalent bond within quantum-accurate force fields: From organic molecules to metallic complexes' reactivity", Sci. Adv. **5**, 5 (2019).

[11] H. Li, Z. Zhang, and Z. Liu, "Application of Artificial Neural Networks for Catalysis: A Review", Catal. **7**, 306 (2017).

[12] S. Wengert, G. Csányi, K. Reuter, and J. T. Margraf, "Data-efficient machine learning for molecular crystal structure prediction", Chem. Sci. **12**, 4536 (2021).

[13] W. Li, Y. Ando, E. Minamitani, and S. Watanabe, "Study of Li atom diffusion in amorphous Li3PO4 with neural network potential", J. Chem. Phys. **147**, 214106 (2017).

[14] N. Artrith and J. Behler, "High-dimensional neural network potentials for metal surfaces: A prototype study for copper", Phys. Rev. B **85**, 045439 (2012).

[15] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, and C. Wolverton, "Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds", Phys. Rev. Mater. **2**, 123801 (2018).

[16] S. Sanvito, C. Oses, J. Xue, A. Tiwari, M. Zic, T. Archer, P. Tozman, M. Venkatesan, M. Coey, and S. Curtarolo, "Accelerated discovery of new magnets in the Heusler alloy family", Sci. Adv. **3**, 4 (2017).

[17] X.-G. Li, C. Hu, C. Chen, Z. Deng, J. Luo, and S. P. Ong, "Quantum-accurate spectral neighbor analysis potential models for Ni-Mo binary alloys and fcc metals", Phys. Rev. B **98**, 094104 (2018).

[18] G. C. Sosso, G. Miceli, S. Caravati, J. Behler, and M. Bernasconi, "Neural network interatomic potential for the phase change material GeTe", Phys. Rev. B **85**, 174103 (2012).

[19] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations", Comput. Mater. Sci. **58**, 227 (2012).

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

[21] S. Curtarolo, G. Hart, M. Buongiorno Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design", Nat. Mater. **12**, 191 (2013).

[22] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the thermodynamic stability of solids combining density functional theory and machine learning", Chem. Mater. **29**, 5090 (2017).

[23] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials", MMS **14**, 1153 (2016).

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

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

[26] 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).

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

[28] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, "Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems", Phys. Rev. Lett. **120**, 036002 (2018).

[29] A. M. Lewis, A. Grisafi, M. Ceriotti, and M. Rossi, "Learning Electron Densities in the Condensed Phase", J. Chem. Theory and Comput. **17**, 7203 (2021).

[30] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost", Chem. Sci. **8**, 3192 (2017).

[31] H. Wang, L. Zhang, J. Han, and W. E, "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics", Comput. Phys. Commun. **228**, 178 (2018).

[32] C. Chen and S. P. Ong, "A universal graph deep learning interatomic potential for the periodic table", Nat. Comput. Sci **2**, 718 (2022).

[33] O. T. Unke and M. Meuwly, "PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges", J. Chem. Theory Comput. **15**, 3678 (2019).

[34] T. Mitchell, *Machine learning* (McGraw-Hill Education, 1997).

[35] A. Géron, *Hands-on machine learning with scikit-learn, keras, and tensorflow* (" O'Reilly Media, Inc.", 2022).

[36] E. V. Podryabinkin and A. V. Shapeev, "Active learning of linearly parametrized interatomic potentials", Computational Materials Science **140**, 171 (2017).

[37] V. Briganti and A. Lunghi, "Efficient generation of stable linear machine-learning force fields with uncertainty-aware active learning", Mach. Learn.: Sci. Technol. **4**, 035005 (2023).

[38] J. Nigam, S. Pozdnyakov, and M. Ceriotti, "Recursive evaluation and iterative contraction of N-body equivariant features", J. Chem. Phys. **153**, 121101 (2020).

[39] 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).

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

[41] E. Kocer, T. W. Ko, and J. Behler, "Neural Network Potentials: A Concise Overview of Methods", Annu. Rev. Phys. Chem. **73**, 163 (2022).

[42] A. Grisafi and M. Ceriotti, "Incorporating long-range physics in atomic-scale machine learning", J. Chem. Phys. **151**, 204105 (2019).

[43] Z. Deng, C. Chen, X.-G. Li, and S. P. Ong, "An electrostatic spectral neighbor analysis potential for lithium nitride", npj Comput. Mater. **5**, 75 (2019).

[44] M. J. Willatt, F. Musil, and M. Ceriotti, "Atom-density representations for machine learning", J. Chem. Phys. **150**, 154110 (2019).

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

[46] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Physics-inspired structural representations for molecules and materials", Chemical Reviews **121**, 9759 (2021).

[47] B. Parsaeifard, D. S. De, A. S. Christensen, 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).

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

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

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

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

[52] H. H. Homeier and E. Steinborn, "Some properties of the coupling coefficients of real spherical harmonics and their relation to Gaunt coefficients", J. Mol. Struct.: THEOCHEM **368**, 31–37 (1996).

[53] J. Avery and J. Avery, *Hyperspherical Harmonics and Their Physical Applications* (World Scientific Publishing Company Pte Limited, 2017).

[54] J. J. Sakurai and J. Napolitano, *Modern Quantum Mechanics*, 2nd ed. (Cambridge University Press, 2017).

[55] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, 1964).

[56] G. B. Arfken and H.-J. Weber, *Mathematical methods for physicists* (Academic Press Orlando, FL, 1972).

[57] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2005).

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

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

[60] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, "Machine Learning a General-Purpose Interatomic Potential for Silicon", Phys. Rev. X **8**, 041048 (2018).

[61] 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 Potentials", J. Phys. Chem. A **124**, 731 (2020).

[62] M. A. Caro, "Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials", Phys. Rev. B **100**, 024112 (2019).

[63] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, "Accurate force field for molybdenum by machine learning large materials data", Phys. Rev. Mater. **1**, 043603 (2017).

[64] Wood, Mitchell A. and Thompson, Aidan P., "Extending the accuracy of the SNAP interatomic potential form", J. Chem. Phys. **148**, 241721 (2018).

[65] M. A. Cusentino, M. A. Wood, and A. P. Thompson, "Explicit Multielement Extension of the Spectral Neighbor Analysis Potential for Chemically Complex Systems", J. Chem. Phys. A **124**, 5456 (2020).

[66] J. M. Goff, C. Sievers, M. A. Wood, and A. P. Thompson, "Permutation-adapted complete and independent basis for atomic cluster expansion descriptors", U.S. Department of Energy, Tech. Report, 2022.

[67] D. P. Kovács, C. v. d. 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).

[68] M. Qamar, M. Mrovec, Y. Lysogorskiy, A. Bochkarev, and R. Drautz, "Atomic Cluster Expansion for Quantum-Accurate Large-Scale Simulations of Carbon", J. Chem. Theory Comput. **19**, 5151 (2023).

[69] Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. P. 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. **7**, 1–12 (2021).

[70] S. Luber, M. Iannuzzi, and J. Hutter, "Raman spectra from ab initio molecular dynamics and its application to liquid S-methyloxirane", J. Chem. Phys. **141**, 094503 (2014).

[71] G. M. Sommers, M. F. C. Andrade, L. Zhang, H. Wang, and R. Car, "Raman spectrum and polarizability of liquid water from deep neural networks", Phys. Chem. Chem. Phys. **22**, 10592 (2020).

[72] M.-P. Gaigeot and M. Sprik, "Ab Initio Molecular Dynamics Computation of the Infrared Spectrum of Aqueous Uracil", J. Phys. Chem. B **107**, 10344 (2003).

[73] M. Weissbluth, *Atoms and Molecules* (Elsevier Science, 2012).

[74] A. Stone, "Transformation between cartesian and spherical tensors", Mol. Phys. **29**, 1461 (1975).

[75] A. Glielmo, P. Sollich, and A. De Vita, "Accurate interatomic force fields via machine learning with covariant kernels", Phys. Rev. B **95**, 214302 (2017).

[76] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, and M. Ceriotti, "Accurate molecular polarizabilities with coupled cluster theory and machine learning", PNAS **116**, 3401 (2019).

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

[78] A. Lunghi and S. Sanvito, "Surfing Multiple Conformation-Property Landscapes via Machine Learning: Designing Single-Ion Magnetic Anisotropy", J. Chem. Phys. C **124**, 5802 (2020).

[79] L. Zhang, B. Onat, G. Dusson, A. McSloy, G. Anand, R. Maurer, C. Ortner, and J. Kermode, "Equivariant analytical mapping of first principles Hamiltonians to accurate and transferable materials models", npj Comput. Mater. **8**, 158 (2022).

[80] M. J. Willatt, F. Musil, and M. Ceriotti, "Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements", Phys. Chem. Chem. Phys. **20**, 29661 (2018).

[81] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, "Transferable Machine-Learning Model of the Electron Density", ACS Cent. Sci. **5**, 57 (2019).

[82] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf, "Electron density learning of non-covalent systems", Chem. Sci. **10**, 9424 (2019).

[83] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad, "Solving the electronic structure problem with machine learning", npj Comput. Mater. **5** (2019).

[84] J. A. Ellis, L. Fiedler, G. A. Popoola, N. A. Modine, J. A. Stephens, A. P. Thompson, A. Cangi, and S. Rajamanickam, "Accelerating finite-temperature Kohn-Sham density functional theory with deep neural networks", Phys. Rev. B **104**, 035120 (2021).

[85] M. Domina, M. Cobelli, and S. Sanvito, "Spectral neighbor representation for vector fields: Machine learning potentials including spin", Phys. Rev. B **105**, 214439 (2022).

[86] S. V. Nikolov, M. A. Wood, A. Cangi, J.-B. Maillet, M.-C. Marinica, A. P. Thompson, M. P. Desjarlais, and J. Tranchida, "Data-driven magneto-elastic predictions with scalable classical spin-lattice dynamics", npj Comput. Mater. **7** (2021).

[87] H. Yu, C. Xu, X. Li, F. Lou, L. Bellaiche, Z. Hu, X. Gong, and H. Xiang, "Complex spin Hamiltonian represented by an artificial neural network", Phys. Rev. B **105**, 174422 (2022).

[88] M. Eckhoff and J. Behler, "High-dimensional neural network potentials for magnetic systems using spin-dependent atom-centered symmetry functions", npj Comput. Mater. **7**, 170 (2021).

[89] J. B. Chapman and P.-W. Ma, "A machine-learned spin-lattice potential for dynamic simulations of defective magnetic iron", Sci. Rep. **12**, 22451 (2022).

[90] I. Novikov, B. Grabowski, F. Körmann, and A. Shapeev, "Magnetic Moment Tensor Potentials for collinear spin-polarized materials reproduce different magnetic states of bcc Fe", npj Comput. Mater. **8**, 13 (2022).

[91] R. Drautz, "Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer", Phys. Rev. B **102**, 024104 (2020).

[92] M. Rinaldi, M. Mrovec, A. Bochkarev, Y. Lysogorskiy, and R. Drautz, *Non-collinear Magnetic Atomic Cluster Expansion for Iron*, arXiv:2305.15137v1, 2023.

[93] M.-T. Suzuki, T. Nomoto, E. V. Morooka, Y. Yanagi, and H. Kusunose, "High-performance descriptor for magnetic materials: Accurate discrimination of magnetic structure", Phys. Rev. B **108**, 014403 (2023).

[94] E. Kocer, J. K. Mason, and H. Erturk, "A novel approach to describe chemical environments in high-dimensional neural network potentials", J. Chem. Phys. **150**, 154102 (2019).

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

[96] H. Yu, Y. Zhong, J. Ji, X. Gong, and H. Xiang, *Time-reversal equivariant neural network potential and Hamiltonian for magnetic materials*, arXiv:2211.11403v2, 2022.

[97] P.-W. Ma, C. H. Woo, and S. L. Dudarev, "Large-scale simulation of the spin-lattice dynamics in ferromagnetic iron", Phys. Rev. B **78**, 024434 (2008).

[98] P.-W. Ma and S. L. Dudarev, "Longitudinal magnetic fluctuations in Langevin spin dynamics", Phys. Rev. B **86**, 054416 (2012).

[99] M. Domina, U. Patil, M. Cobelli, and S. Sanvito, "Cluster expansion constructed over Jacobi-Legendre polynomials for accurate force fields", Phys. Rev. B **108**, 094102 (2023).

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

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

[102] G. E. Andrews, R. Askey, and R. Roy, *Special Functions*, Encyclopedia of Mathematics and its Applications (Cambridge University Press, 1999).

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

[104] J. Nigam, S. N. Pozdnyakov, K. K. Huguenin-Dumittan, and M. Ceriotti, "Completeness of atomic structure representations", APL Machine Learning **2**, 016110 (2024).

[105] M. W. Finnis and J. E. Sinclair, "A simple empirical N-body potential for transition metals", Phil. Mag. A **50**, 45 (1984).

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

[107] 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).

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

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

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

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

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

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

[114] W. Kohn, "Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms", Phys. Rev. Lett. **76**, 3168 (1996).

[115] W. Kohn and A. Yaniv, "Locality principle in wave mechanics", PNAS **75**, 5270 (1978).

[116] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, B. Kieron, and K. R. Müller, "Bypassing the Kohn-Sham equations with machine learning", Nat. Commun. **8**, 872 (2017).

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

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

[119] G. Eda, T. Fujita, H. Yamaguchi, D. Voiry, M. Chen, and M. Chhowalla, "Coherent atomic and electronic heterostructures of single-layer mos2", ACS Nano **6**, PMID: 22799455, 7311–7317 (2012).

[120] C. Liang, G. Tocci, D. M. Wilkins, A. Grisafi, S. Roke, and M. Ceriotti, "Solvent fluctuations and nuclear quantum effects modulate the molecular hyperpolarizability of water", Phys. Rev. B **96**, 041407 (2017).

[121] P. E. Blöchl, "Projector augmented-wave method", Phys. Rev. B **50**, 17953 (1994).

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

[123] E. Artacho, D. Sánchez-Portal, P. Ordejón, A. García, and J. M. Soler, "Linear-Scaling ab-initio Calculations for Large and Complex Systems", Phys. Status Solidi B **215**, 809 (1999).

[124] E. Artacho, E. Anglada, O. Diéguez, J. D. Gale, A. García, J. Junquera, R. M. Martin, P. Ordejón, J. M. Pruneda, D. Sánchez-Portal, and J. M. Soler, "The SIESTA method; developments and applicability", J. Condens. Matter Phys. **20**, 064208 (2008).

[125] A. R. Rocha, V. M. García-Suárez, S. Bailey, C. Lambert, J. Ferrer, and S. Sanvito, "Spin and molecular electronics in atomically generated orbital landscapes", Phys. Rev. B **73**, 085414 (2006).