Efficient and insightful descriptors for representing molecular and material space

Présentée le Février 22, 2024
École polytechnique fédérale de Lausanne
Faculté des sciences et techniques de l'ingénieur
Laboratoire de science computationnelle et modélisation
(en anglais *Computational Science and Modeling*)
Programme doctoral en science et génie des matériaux

pour l'obtention du grade de Docteur ès Sciences par

Alexander Jan Goscinski

acceptée sur proposition du jury:

Prof Véronique Michaud, président du jury
Prof Ceriotti Michele, directeur de thèse
Prof Herbst Michael, rapporteur
Dr. Caro Miguel, rapporteur
Dr. Rupp Matthias, rapporteur

Lausanne, EPFL, 2024

What I cannot create,
I do not understand.
— Richard Feynman

To my mother, sister and brother

Acknowledgements

I want to thank my supervisor Prof. Michele Ceriotti for the investment of his lifetime to train me and for his patience, Dr. Michael J. Willatt and Kevin K. Huguenin-Dumittan for the cultivating discussions about math, Dr. Félix Musil and Max Veit for the collaboration on the strenuous development of `librascal`, Dr. Guillaume Fraux for teaching me how to develop long-lasting software, Jigyasa Nigam, Sergey N. Pozdnyakov and Filippo Bigi for the discussions on atomistic representations and their efficiency, Dr. Bruno Loureiro and Giovanni Piccioli for the discussions on statistical learning theory, Dr. Federico Grasselli, Dr. Lorenzo Gigli, Dr. Chiheb B. Mahmoud and Dr. Davide Tisi for teaching me physics, Dr. Martin Uhrin for the discussions on the inverse design problem, Dr. Nataliya Lopanitsyna and Dr. Andrea Anelli for the support, Prof. Rose K. Cersonsky, Dr. Ben A. Helfrecht and Sergei Kliavinek for the collaboration on `scikit-matter`, João Prado, Taylor Baird and Divya Suman for the collaboration on `scicode-widgets`, Dr. Sanggyu Chong and Matthias Kellner for introducing me to the world of chemistry, Dr. Philip Loche and Joe Abbott for the collaboration on the laboratories software stack and all people that have been a member of the laboratory of computational science and modeling (COSMO) during my time for making it a pleasant working environment to discuss and share ideas.

Lausanne, January 18, 2024 A. G.

Abstract

Data-driven approaches have been applied to reduce the cost of accurate computational studies on materials, by using only a small number of expensive reference electronic structure calculations for a representative subset of the materials space, and using them to train surrogate models that predict inexpensively the outcome of such calculations on an extensive space of configurations spanned by the subset. The way materials structures are processed into a numerical description as input of machine learning algorithms is crucial to obtain efficient models, and has advanced significantly in the last decade, putting forth enhancements in the embedding of geometrical and chemical information. Despite the rapid development of offloading calculations to more dedicated hardware, these enhancements nevertheless substantially increase the cost of the numerical description, which remains a crucial factor in simulations. It is therefore vital to delve deeper into the design space of representations to understand the type of information the numerical descriptions encapsulate. Insights from such analyzes aid in making more informed decisions regarding the trade-off between accuracy and performance. While a substantial amount of work has been undertaken to compare representations concerning their structure-property relationship, a thorough exploration of the inherent nature and the information capacity of these representations remains mostly unexplored. This thesis introduces a set of measures that facilitate quantitative analysis concerning the relationship between features, thereby assisting in such decision-making processes and providing valuable insights to the academic community. We demonstrate how these measures can be applied to analyze representations that are built in terms of many-body correlations of atomic densities. For this form of featurization, we investigate the impact of different choices for the functional form, the basis functions, and the induced feature space determined by the similarity measure and metric space. We employ these measures subsequently on featurizations with basis functions optimized to the dataset to show the higher information capacity in comparison to an unoptimized. We show how these well-established optimization methods based on the covariance or correlation matrix, such as principal component analysis, can be applied in a manner that preserves symmetries. The scheme utilizes splines to bypass the optimization during prediction time, permitting the adoption of more expansive optimization methods in the future. Complementing these efforts is the integration of the developed methods into well-maintained and thoroughly documented packages, facilitating advancements and incorporation into new workflows. As a showcase of this development, we present a framework for running metadynamics simulations that incorporates a machine learning interatomic potential into the molecular dynamics engine `LAMMPS` to exploit its message-passing interface implementation of the domain decomposition. This enabled us to study finite-size effects in the paraelectric-ferroelectric phase transition in barium titanate. Born out of this software development, a way forward is presented for a more modular software ecosystem for the flexible construction of data-driven interatomic potentials with immediate deployment into simulations.

Zusammenfassung

Datengesteuerte Ansätze wurden eingesetzt, um die Kosten für genaue rechnerische Untersuchungen von Werkstoffen zu senken, indem nur eine kleine Anzahl von teuren Referenzberechnungen der elektronischen Struktur für eine repräsentative Teilmenge des Materialraums verwendet wird und mit ihnen Surrogatmodelle trainiert werden, die kostengünstig das Ergebnis solcher Berechnungen für einen umfangreichen Raum von Konfigurationen vorhersagen, der durch die Teilmenge abgedeckt wird. Die Art und Weise, wie Materialstrukturen in eine numerische Deskription als Input für Algorithmen des maschinellen Lernens verarbeitet werden, ist entscheidend um effiziente Modelle zu kreieren und hat sich in den letzten zehn Jahren erheblich weiterentwickelt, die durch Verbesserungen bei der Einbettung geometrischer und chemischer Informationen erzielt wurden. Trotz der rasanten Entwicklung bei der Auslagerung von Berechnungen auf speziellere Hardware erhöhen diese Verbesserungen jedoch die Kosten der numerischen Deskription erheblich, was nach wie vor ein entscheidender Faktor bei Simulationen ist. Daher ist es von entscheidender Bedeutung, den Design Space von Darstellungen tiefer zu durchdringen, um zu verstehen, welche Art von Informationen die numerischen Deskriptionen einschließen. Die Erkenntnisse aus solchen Analysen helfen dabei, fundiertere Entscheidungen über den Kompromiss zwischen Genauigkeit und Leistung zu treffen. Während eine beträchtliche Menge an Arbeit geleistet wurde, um Repräsentationen hinsichtlich ihrer Struktur-Eigenschafts-Beziehung zu vergleichen, bleibt eine gründliche Erforschung der inhärenten Natur und der Informationskapazität dieser Repräsentationen weitgehend unerforscht. In dieser Arbeit wird eine Reihe von Maßen vorgestellt, die eine quantitative Analyse der Beziehungen zwischen den Merkmalen erleichtern und so bei solchen Entscheidungsprozessen helfen und der akademischen Gemeinschaft wertvolle Erkenntnisse liefern. Wir demonstrieren, wie diese Maße zur Analyse von Repräsentationen eingesetzt werden können, die auf Vielkörper-Korrelationen der atomaren Dichte beruhen. Für diese Form der Featurisierung untersuchen wir die Auswirkungen verschiedener Entscheidungen für die funktionale Form, die Basisfunktionen und den induzierten Feature Space, der durch das Ähnlichkeitsmaß und den metrischen Raum bestimmt wird. Wir wenden diese Maße anschließend auf Featurisierungen mit auf den Datensatz optimierten Basisfunktionen an, um die höhere Informationskapazität im Vergleich zu einer nicht optimierten zu zeigen. Wir zeigen, wie diese gut etablierten Optimierungsmethoden, die auf der Kovarianz- oder Korrelationsmatrix basieren, wie z. B. die Hauptkomponentenanalyse, auf eine Weise angewendet werden können, die Symmetrien bewahrt. Das Schema nutzt Splines, um die Optimierung während der Vorhersagezeit zu umgehen, sodass in Zukunft auch umfangreichere Optimierungsmethoden eingesetzt werden können. Ergänzt werden diese Bemühungen durch die Integration der entwickelten Methoden in gut gewartete und sorgfältig dokumentierte Pakete, die die Weiterentwicklungen und die Einbindung in neue Arbeitsabläufe erleichtern. Als Beispiel für diese Entwicklung stellen wir ein Framework für die Durchführung von Metadynamiksimulationen vor, der ein interatomares Potential für maschinelles Lernen

in die Molekulardynamik-Engine LAMMPS integriert, um deren Message Passing Interface Implementierung der Domänenzerlegung zu nutzen. Dies ermöglichte uns die Untersuchung von Finite-Size-Effekten beim paraelektrischen-ferroelektrischen Phasenübergang in Bariumtitanat. Aus dieser Softwareentwicklung hervorgehend wird ein Weg zu einem modulareren Software-Ökosystem für die flexible Konstruktion datengesteuerter interatomarer Potentiale mit sofortigem Einsatz in Simulationen vorgestellt.

List of Publications directly related to this thesis

- Félix Musil, Max Veit, Alexander Goscinski, Guillaume Fraux, Michael J Willatt, Markus Stricker, Till Junge, and Michele Ceriotti. Efficient implementation of atom-density representations. *The Journal of Chemical Physics*, 154(11), 2021.
- Alexander Goscinski, Guillaume Fraux, Giulio Imbalzano, and Michele Ceriotti. The role of feature space in atomistic learning. *Machine Learning: Science and Technology, 2(2):025028*, 2021.
- Alexander Goscinski, Félix Musil, Sergey Pozdnyakov, Jigyasa Nigam, and Michele Ceriotti. Optimal radial basis for density-based atomic representations. *The Journal of Chemical Physics*, 155(10), 2021.
- Alexander Goscinski, Victor Paul Principe, Guillaume Fraux, Sergei Kliavinek, Benjamin Aaron Helfrecht, Philip Loche, Michele Ceriotti, and Rose Kathleen Cersonsky. scikit-matter: A suite of generalisable machine learning methods born out of chemistry and materials science. *Open Research Europe*, 3:81, 2023.
- Lorenzo Gigli, Alexander Goscinski, Michele Ceriotti, and Gareth A. Tribello. Modeling The ferroelectric phase transition in barium titanate with DFT accuracy and converged sampling. *arXiv preprint arXiv:2310.12579*, 2023.

Contents

Introduction

In high-throughput material design, large databases of materials are searched for candidates with desirable characteristics. So far, searches based on experimental data have been limited in scope due to the vast combinatorial space of materials, the heterogeneous quality of available data, and the difficulty in separating the intrinsic properties of a material from those that are contingent on the processing or synthesis conditions. In the last decades, it has been shown that ab initio quantum chemistry methods provide approximate stability criteria that are in good agreement with experiments [1] making them a viable tool for the screening of new materials [2, 3, 4, 5]. The quantitative accuracy of these predictions, however, is dependent on the quality of the reference electronic structure calculations. Higher levels of theory are often necessary to model the electronic structure properties correctly, but they come with an additional computational effort, thereby reducing the breadth of the searches. Due to the vast number of possible atomic structures to be considered, the efficiency of these methods is crucial. Data-driven methods have become an efficient extension, reducing expensive quantum chemistry calculations to a bare minimum while reaching close-to-ab initio accuracy over a wide configuration space [6], leading to the exploration of previously computationally intractable problems, such as the thermal conductivity of amorphous germanium telluride [7], and moving simulations more into the role as guidance for experiments [8]. These methods are based on transforming geometrical, physical, and chemical information into a vector representation, referred to as descriptor, to then use it as input of a machine learning model. The development of expressive and computationally inexpensive descriptors [9, 10] has led to applications in a wide range of areas [11, 12, 13]. Efficient descriptors are therefore essential for state-of-the-art high-throughput material design applications. The efficient computation of expressive descriptors is a challenging problem that has seen a wide range of proposals [9, 10, 14, 15]. When used to build an interatomic potential or to predict other atomic-scale properties, representations are used together with different supervised learning schemes, so it is difficult to disentangle the interplay of the descriptor, regression method, and target property that combine to determine the accuracy and computational cost of the different methods [16]. A deeper understanding of these descriptors is therefore essential, especially considering that the efficiency of accurate potentials is still a limiting factor for the research that can be conducted on materials.

In the first chapter of this thesis, atomistic descriptors based on geometrical information and developed in the last two decades are formalized in a mathematically grounded theory utilizing concepts of representation theory to connect the set of numerical descriptions to functional spaces. In the second chapter, we continue to present a set of measures that can serve for quantitative analysis to guide the choice of the descriptor and model and to provide insight into the inherent nature of the descriptors. Insights are presented on how

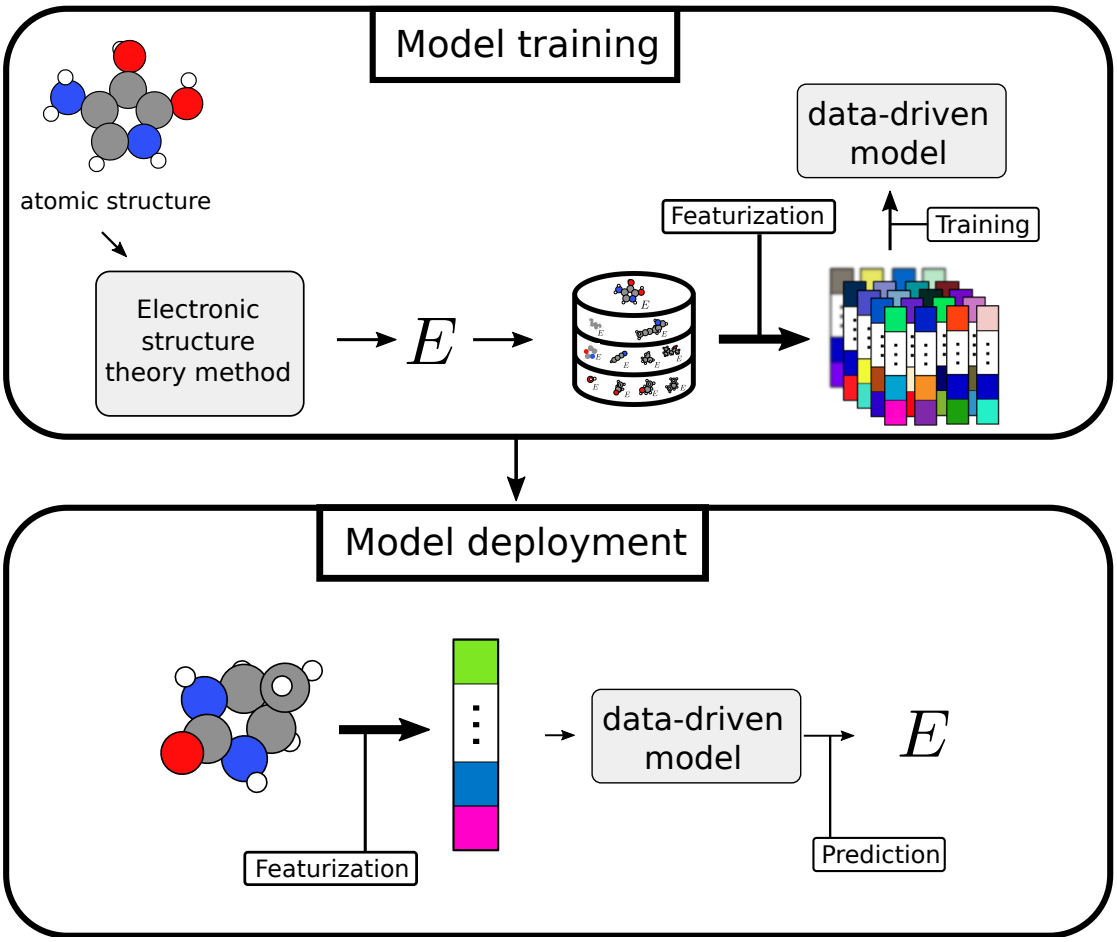

Figure 1: A schematic showing the idea of high-throughput calculations with a data-driven model that serves as surrogate model to bypass the expensive electronic structure theory calculations after training the model.

changes in the smearing, cutoff, radial scaling, body-order or switch of the metric space to the Wasserstein metric affects the information content of the feature space induced by the metric. In Chapter 3 we present how symmetries can be embedded into data-driven optimization methods based on the covariance or correlation matrix to improve the radial basis functions. The information capacity is evaluated with the metrics developed in Chapter 2 and benchmarks on the regression performance for a single species and a multi-species dataset are shown. The last chapter focuses on the implementation of a machine learning interatomic potential (MLIP) and showcases its application to study finite-size effects in paraelectric-ferroelectric phase transitions in barium titanate. The specifications of the spline used for the basis optimization presented in Chapter 4 are written out, and the effect of the grid on performance and accuracy is briefly discussed. Finally, drawbacks in extensibility and deployment of the developed MLIP software are outlined, and future directions for a modular machine learning software ecosystem for atomistic simulations are presented.

1 Theory of atomistic representations

Physical properties, such as energies, dipole moments and polarizabilities, all exhibit symmetries that can be exploited to facilitate the construction of a surrogate model that learns a relationship to such properties from geometric information. By embedding the symmetries into the numerical description of the atomic structure the hypothesis space is reduced that needs to be considered by the learning algorithm, thereby resulting in more effective models. This chapter covers the theory and computation of symmetrized features on the atomic-scale. A similar approach as in Ref. [17] is taken that introduces the topic by utilizing concepts from representation theory to give a more profound understanding of the approaches existing in the field. We therefore begin with introducing the representation $f_A : \Omega \to \mathbb{R}$ of an atomic structure A on a smooth manifold $\Omega \subseteq \mathbb{R}^z$, where we use Ω to consider different encodings of the atomic structure A. In its simplest form, an encoding can be the atomic positions $\mathbf{q} \in \mathbb{R}^{3N}$ of structure A. To construct a numerical description for a structure A that can be used as input for a data-driven model, we project on its representation with an orthonormal set of *basis functions* $\{b_k : \Omega \to \mathbb{R}\}_{k=1}^{M}$, to obtain a set of *expansion coefficients* $\{c_k \in \mathbb{R}\}_{k=1}^{M}$ from the basis expansion

$$c_k = \int_\Omega \mathrm{d}\mathbf{x}\, f_A(\mathbf{x})\, b_k(\mathbf{x}) \text{ for } k = 1, \dots M. \tag{1.1}$$

For a lot of cases the orthonormality constraint of the basis is relaxed, since the orthonormalization can be seen as part of the learning algorithm. The choice of the representation space as well as the basis is essential for an effective numerical description, i.e. a description that captures information with fewer number coefficients. We will refer to the whole process of transforming a structure A to a representation and then to a numerical description as *featurization* of structure A. A widely-used family of representation spaces is based on higher orders of atom-density-based functions. This family of representation spaces is introduced and it is shown how invariances can be efficiently embedded into the computation of the expansion coefficients. Additionally, general characteristics of basis functions deployed in atomic-scale models are presented as well as different practices to transform the expansion coefficients into inputs for data-driven models.

1.1 Atom-centered density

A majority of developed atomistic descriptors can be seen as different approaches to construct the expansion coefficients based on a family of functions, that originates from the structural

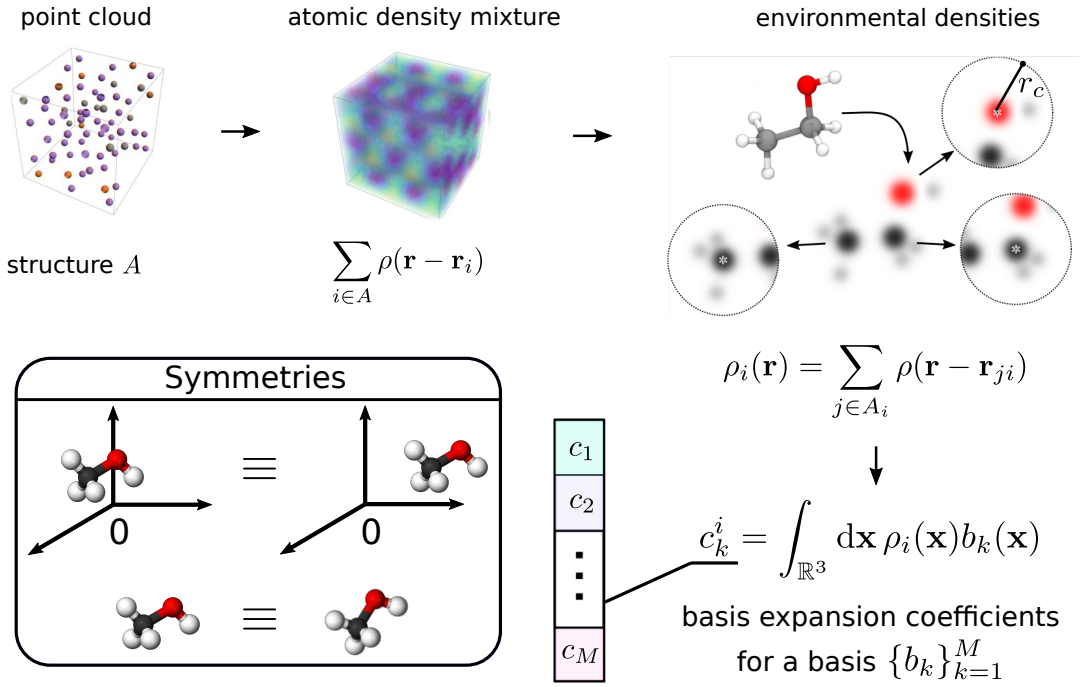

Figure 1.1: A schematic showing the featurization of an atomic structure A based on the atom-centered density correlations functions. The figure of the atoms in the box are retrieved from Ref. [18]. The figure of the atomic environments is retrieved from Ref. [17]. The methanol molecule is retrieved from Ref. [19].

density function [20]

$$\sum_{i \in A} \rho(\mathbf{r} - \mathbf{r}_i), \quad \rho : \mathbb{R}^3 \to \mathbb{R} \quad \text{(atomic density)}, \tag{1.2}$$

where $\mathbf{r}_i \in \mathbb{R}^3$ is the position of the ith atom in the atomic structure A and ρ is an arbitrary function decaying from its origin. Commonly, a Gaussian g or a Dirac δ function are chosen as atomic density ρ. A widely adapted approach to impose translational invariance, i.e. independence of the center of the structure, is to describe the atomic structure as a sum of atomic environment contributions

$$\rho_i(\mathbf{r}) = \sum_{j \in A_i} \rho(\mathbf{r} - \mathbf{r}_{ji}), \quad \rho_i : \mathbb{R}^3 \to \mathbb{R} \quad \text{(environmental density of atom } i\text{)}, \tag{1.3a}$$

where A_i is the set of all atoms that are in the *environment* of atom i, in most applications defined as the set of atoms within a certain distance, the *cutoff*, and $\mathbf{r}_{ji}$ is the direction vector $\mathbf{r}_j - \mathbf{r}_i$. While we refer to ρ_i as environmental density in this thesis, the term atomic density is however frequently more loosely used to refer also to ρ_i [20]. This approach further aligns with the partitioning of a structure property y_A into local atomic contributions

$$\sum_{i \in A} y_i = y_A \tag{1.4}$$

which is motivated by the heuristical observation that atomic properties decay with their distance to the center, a concept commonly referred to as *locality* or *nearsightedness* [21].

1.2 Hierarchy of invariant representations

As a large family of physical quantities are invariant under rotations of the atomic structure, it is thus required to account for rotational invariance in the process of relating atomic structures to such quantities. Rotational invariance can be embedded into the representation by simply introducing a Haar integral over the rotation group $SO(3)$

$$\overline{\rho_i^{\otimes 1}}(\mathbf{r}) = \int_{SO(3)} \mathrm{d}\hat{R}\, \rho_i(\hat{R}\mathbf{r}) \quad \text{(ACDC of order 1)} \tag{1.5}$$

which can be further extended to higher-order correlations of the density

$$\overline{\rho_i^{\otimes \nu}}(\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(\nu)}) = \int_{SO(3)} \mathrm{d}\hat{R}\, \rho_i(\hat{R}\mathbf{r}^{(1)}) \ldots \rho_i(\hat{R}\mathbf{r}^{(\nu)}) \quad \text{(ACDC of order } \nu). \tag{1.6}$$

This class of representations has been named *atom-centered density correlations* (ACDC) functions [22]. Although it is theoretically possible to evaluate this integral numerically, it does not offer an efficient means to determine the expansion coefficients. Hence, it is essential to choose suitable candidates for the density ρ and the basis set $\{b_k\}_{k=1}^{M}$ that yield an efficient solution for the integral.

1.2.1 Solution for Dirac δ densities

An explicit solution of the integral over $SO(3)$ can be expressed for the Dirac δ densities. Here we present the solutions for order 1 and 2

$$\sum_j \int_{SO(3)} \mathrm{d}\hat{R}\, \delta(\hat{R}\mathbf{r} - \mathbf{r}_{ji}) \propto_r \sum_j \delta(r - r_{ji}) \quad \text{(order 1)}, \tag{1.7a}$$

$$\sum_{jk} \int_{SO(3)} \mathrm{d}\hat{R}\, \delta(\hat{R}\mathbf{r} - \mathbf{r}_{ji}) \delta(\hat{R}\mathbf{r}' - \mathbf{r}_{ki}) \propto_r \sum_{jk} \delta(r - r_{ji}) \delta(r' - r_{ki}) \delta(\theta - \theta_{jki}) \quad \text{(order 2)}, \tag{1.7b}$$

where we use $\mathbf{r}$ and $\mathbf{r}'$ as shorter notation to refer to $\mathbf{r}^{(1)}$ and $\mathbf{r}^{(2)}$. We use $\propto_r$ to omit constant factors and radial terms r that appear due to the integration over the rotation group. These factors are not essential, since typically a radial scaling term is added to the density to control the general scaling [9, 23, 24, 25]. The correlation function of order 1 naturally results in a description of the distance to atom i and the order 2 function in the two distances and an angle with respect to atom i. This can be generalized to higher orders retrieving a decomposition into different body-order contributions as it is done for interatomic potentials. This expansion makes the close relationship clear between ACDC-based descriptors to interatomic potentials that decompose the energy into different body-order contributions.

1.2.2 Ordered support of representation

One early-developed approach to obtain a numerical input has been to directly use the discrete many-body information (e.g. 2-body distances in the environment) in form of a concatenated vector. The vector is then sorted to achieve permutational invariance [26, 27, 28]. The sorting of the many-body information approach can be connected to the ACDC descriptors by a change of the metric space from the L^1 norm distance to the Earth mover's distance (EMD) [17]. The relationship can be clearly seen by using fact that the EMD between two distributions p and p′

can be connected to the L^1 distance between the inverses of the cumulative density functions P and P' of those distributions

$$\mathrm{EMD}(\mathrm{p},\mathrm{p}') = \int_0^1 \mathrm{d}s \left| P^{-1}(s) - P'^{-1}(s) \right|, \text{ with } P(x) = \int_{-\infty}^{s} \mathrm{p}(x). \tag{1.8}$$

Then the EMD between two order 1 ACDC functions using the Dirac δ function as the atomic densities is equal up to a normalization factor dependent on the number of atoms to the L^1 difference between their sorted distances vectors [17]. This fact can be extended to the L^p norm distance denoted by the term W_p referring to the more common naming *Wasserstein distance* for the EMD.

$$W_p(\mathrm{p},\mathrm{p}')^p = \int_0^1 \mathrm{d}s \left| P^{-1}(s) - P'^{-1}(s) \right|^p. \tag{1.9}$$

It is not clear how the Wasserstein metric applied to higher orders of the ACDC functions changes the nature of the representation, since for higher dimensions a form as in Eq. (1.9), reproducing the Wasserstein distance by the L^p distance for a representation, is not known. An approach that extends this idea to higher orders utilizes sorted angles or sorted torsions as description [28]. These can be seen as one-dimensional projections of the higher-order ACDC functions and do not consider the whole coherent space of correlations. It has been shown that descriptors in this category can reach comparable accuracies to the common methods that can be induced from the Euclidean metric [27, 28] with energy accuracies close to 1 kcal/mol on the QM9 and QM7b dataset. These descriptions nevertheless have undesired characteristics with regard to differentiability and extensibility. The sorted quantities have discontinuities with respect to changes in the atomic positions that emerge when two distances are swapped due to the sorting, which is problematic for predictions of derivatives. These discontinuities can however be mitigated by smoothening the descriptor within a similarity or distance measure. Another problem is that their size depends on the number of atoms in the local environment being represented, thus they are often padded with zero values impeding their application to neighborhoods with diverse number of atoms.

1.2.3 Fixed basis set

Considering the order 1 solution in Eq. (1.7a) with an additional radial factor r

$$f(r) = r \sum_j \delta(r - r_{ji}), \tag{1.10}$$

we can retrieve the 2-body distances used for sorted distance by using a basis set of the form $\{\delta(r - r_{ji}) : \mathbb{R} \to \mathbb{R}\}_{j \in A_i}$ with a subsequent sorting. From this point of view the cause of the discontinuities exihibited by the descriptors in the last section can be attributed to the variation of the basis function across atomic environments. A natural solution is therefore the usage of the same basis set across all environments in all structures [9, 10, 25]. By expanding on the ACDC function using the Dirac δ function for the densities with the basis functions $\{b_k^\nu\}_k$ of different orders one obtains coefficients of the form

$$\int_{\mathbb{R}^3} \mathrm{d}\mathbf{r}\, b_k^1(r)\delta(r - r_{ji}) \propto_r b_k^1(r_{ji}), \tag{1.11a}$$

$$\int_{\mathbb{R}^3\times\mathbb{R}^3} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}'\, b_k^2(r, r\prime, \theta(\mathbf{r}\cdot\mathbf{r}'))\delta(r' - r_{ji})\delta(r - r_{ki})\delta(\theta(\mathbf{r}\cdot\mathbf{r}') - \theta_{jki}) \propto_r b_k^2(r_{ji}, r_{ki}, \theta_{ijk}). \tag{1.11b}$$

One widely-used representative of this approach is the Behler-Parrinello symmetry function (BPSF) [9]. In the next section we discuss how this approach is extended for Gaussian densities.

1.2.4 Radial and angular decomposition

For Gaussian densities the order 1 expression in Eq. (1.5) can be analytically solved by exploiting properties of the Gaussian function [29]

$$\int_{SO(3)} \mathrm{d}\hat{R}\, g(\hat{R}\mathbf{r}-\mathbf{r}_{ji}) = \int_{SO(3)} \mathrm{d}\hat{R}\, \exp\big(\|\hat{R}\mathbf{r}-\mathbf{r}_{ji}\|^2/(2\sigma^2)\big) \tag{1.12a}$$

$$= 8\pi^2 \sinh\Big(r\, r_{ji}^2/2\sigma^2\Big)(r\, r_{ji}/2\sigma^2)\exp\Big(-(r^2+r_{ji}^2)/4\sigma^2\Big) \tag{1.12b}$$

$$\approx \frac{1}{r_{ij}}\exp\big(-((r-r_{ji})^2)/4\sigma^2\big) \tag{1.12c}$$

which gives approximatively a Gaussian density in radial space. A solution of the integral for higher orders requires a more complex derivation utilizing mathematical properties of spherical harmonics $Y_m^l(\hat{\mathbf{r}}):\mathbb{R}^2\to\mathbb{R}$. Spherical harmonics have been studied extensively in invariant theory [30] and in angular momentum theory [31] which make them a suitable candidate to solve the integral in Eq. (1.6). Consequently, to exploit the mathematical properties of spherical harmonics the atomic density must be reexpressed in form of spherical harmonics. We extend the spherical harmonics by a complete orthonormal radial basis $\{R_n(r):\mathbb{R}\to\mathbb{R}\}_{n=1}^{\infty}$ to cover the radial part of the density. Then the atomic density can be reformulated as

$$c_{nlm}^i = \int_{\mathbb{R}^3} \mathrm{d}\mathbf{r}\, R_n(r)\, Y_m^l(\hat{\mathbf{r}})\rho_i(\mathbf{r}), \tag{1.13a}$$

$$\rho_i(\mathbf{r}) = \sum_{nlm} c_{nlm}^i R_n(r)\, Y_m^l(\hat{\mathbf{r}}). \tag{1.13b}$$

This reformulation allows us to solve Eq. (1.6) for order 2. The radial basis can be extracted out of the integral as it is not affected

$$\int_{SO(3)} \mathrm{d}\hat{R}\,\rho_i(\hat{R}\mathbf{r})\rho_i(\hat{R}\mathbf{r}') = \sum_{nn'} R_n(r)R_{n'}(r') \sum_{ll'mm'} c_{nlm}c_{n'l'm'} \int_{SO(3)} \mathrm{d}\hat{R}\, Y_m^l(\hat{R}\hat{\mathbf{r}})\, Y_{m'}^{l'}(\hat{R}\hat{\mathbf{r}}'). \tag{1.14}$$

For solving the integral we can omit the radial part and coefficients outside of the integral for simplicity

$$\sum_{ll'mm'} \int_{SO(3)} \mathrm{d}\hat{R}\, Y_m^l(\hat{R}\hat{\mathbf{r}})\, Y_{m'}^{l'}(\hat{R}\hat{\mathbf{r}}') \tag{1.15a}$$

$$= \sum_{ll'mm'} \int_{SO(3)} \mathrm{d}\hat{R} \sum_u D_{mu}^l(\hat{R})\, Y_u^l(\hat{\mathbf{r}}) \sum_{u'} D_{m'u'}^{l'}(\hat{R})\, Y_{u'}^{l'}(\hat{\mathbf{r}}') \quad (\mathbf{D}(\hat{R}) \text{ is the Wigner D-matrix}) \tag{1.15b}$$

$$= \sum_{ll'uu'} Y_u^l(\hat{\mathbf{r}})\, Y_{u'}^{l'}(\hat{\mathbf{r}}') \sum_{mm'} \int_{SO(3)} \mathrm{d}\hat{R}\, D_{mu}^l(\hat{R}) D_{m'u'}^{l'}(\hat{R}) \tag{1.15c}$$

$$\propto \sum_{lu} Y_u^l(\hat{\mathbf{r}})\, Y_u^l(\hat{\mathbf{r}}') \quad \text{(orthogonality Wigner D-matrix)} \tag{1.15d}$$

$$\propto \sum_{l} P_l(\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}') \quad \text{(addition theorem, where } P_l \text{ Legendre polynomial [32])} \tag{1.15e}$$

Incorperating the radial part and the coefficients back into the above solution we obtain

$$\sum_{nn'l} c_{nn'l} R_n(r) R_{n'}(r') P_l(\hat{\mathbf{r}} \cdot \hat{\mathbf{r}}'), \text{ with } c_{nn'l} = \sum_m c_{nlm} c_{n'lm}. \tag{1.16}$$

The integral for higher orders can be further solved by exploiting the fact that the product of Wigner D-matrices can be decomposed into a linear combination of Wigner D-matrices

$$D^{l_1}_{m_1 m_1'}(\hat{R}) D^{l_2}_{m_2 m_2'}(\hat{R}) = \sum_{lmm'} D^{l}_{mm'}(\hat{R}) C^{ll_1l_2}_{mm_1m_2} C^{ll_1l_2}_{m'm_1'm_2'} \tag{1.17}$$

where $C^{ll_1l_2}_{\mu m_1 m_2}$ are the real Clebsch-Gordan coefficients [31, 33]. This relationship was initially utilized in Ref. [10] to generate order 3 functions, commonly referred to as *bispectrum*. Subsequently, it has been formulated into a recursive expression to derive higher-order functions of the form

$$\overline{\rho_i^{\otimes \nu+1}}(\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(\nu+1)}) = \sum_{k_{\nu+1}} c_{k_{\nu+1}} f^{\nu+1}_{k_{\nu+1}}(\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(\nu+1)}), \tag{1.18a}$$

$$c_{k_{\nu+1}} = \sum_{k_\nu, k_1} c_{k_\nu k_1} c_{k_1} c_{k_\nu}, \tag{1.18b}$$

where we can separate between coefficients of the form c_{k_ν} that depend solely on the order ν function, and further coefficients of the form $c_{k_\nu k_1}$ that couple the order ν and 1 functions. The coefficients $c_{k_\nu k_1}$ are connected to the Clebsch-Gordan coefficients in Eq. (1.17), the relationship is more formally derived in Ref. [33]. Due to the polynomial increase of the feature size with body-order, there exist various strategies for compressing the basis coefficients in high-dimensional space [33, 34, 35].

1.3 Basis expansion

Solving the integral by expanding the atomic density onto a certain basis as shown in Eq. (1.15) naturally enforces the same choice for the basis set to solve for the expansion coefficients

$$c^i_{nlm} = \int_{\mathbb{R}^3} \mathrm{d}\mathbf{r} R_n(r) Y^l_m(\hat{\mathbf{r}}) \rho_i(\mathbf{r}) \quad \text{(spherical expansion coefficients)}, \tag{1.19a}$$

$$c^i_{nn'l} = \sum_m c^i_{nlm} c^i_{n'lm} = \int_{\mathbb{R}^3 \times \mathbb{R}^3} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' R_n(r) R_{n'}(r') P_l(\hat{\mathbf{r}} \cdot \hat{\mathbf{r}}') \int_{SO(3)} \mathrm{d}\hat{R} \rho_i(\hat{R}\mathbf{r}) \rho_i(\hat{R}\mathbf{r}') \quad \text{(order 2)}. \tag{1.19b}$$

The order 2 expansion coefficients are frequently referred to as *smooth overlap of atomic positions* (SOAP) [10]. Note that while we motivated the decomposition of the representation into an angular and radial part as a means to solve the Haar integral for higher orders, one can also motivate this decomposition for the Dirac δ density. An extensively employed representative of Dirac δ densities is named *atomic cluster expansion* (ACE) [25].

1.3.1 Density trick

The Eqs. (1.19) show that the order 2 coefficients can be computed from the spherical expansion coefficients. Taking into account the recursion formula presented in Eq. (1.18) it becomes

evident that all higher-order correlations can be constructed from the spherical expansion coefficients. This way of propagating to higher orders has been referred to as the *density trick*. It shifts the computation from the evaluation of the basis expansions across all ν-tuples to the computation of the contracted tensor products between the expansion coefficients of order ν with the spherical expansion coefficients. For example, using the density trick to compute the order 2 coefficients in Eq. (1.19b) for an environment one has to compute the M expansion coefficients for a density with N neighbors resulting in a complexity of $O(MN)$ to then increase the order by a contracted tensor product scaling as $O(M^2)$ resulting in a total time complexity of $O(MN + M^2)$ to obtain M^2 expansion coefficients. Without the usage of the density trick it requires the evaluation of $O(\binom{N}{2}) = O(N^2)$ neighbor pairs for each of the $O(M^2)$ basis functions resulting in a total time complexity of $O(M^2N^2)$ for M^2 expansion coefficients. Even though the scaling favors the use of the density trick, when replacing the combined feature computation and model prediction by a spline, one can directly evaluate the $(\nu+1)$-body potential, thereby eliminate altogether the evaluation of features, effectively setting $M = 1$ [36, 37]. Despite the remaining $O(N^2)$ cost, it is drastically faster than comparable methods utilizing the density trick as shown in the benchmarks in Ref. [37].

1.3.2 Radial basis

The radial basis consists of one-dimensional functions defined on a compact domain $R_n : [0, r_c] \to \mathbb{R}$, where the cutoff r_c forms one of the hyperparameters of the radial basis. A variety of radial basis functions, such as shifted-Gaussians [10], Chebyshev polynomials [25, 38] or Gaussian type orbitals [39], have been proposed in the literature. These functions all share certain characteristics that have been shown to positively imptact the learning performance. One key characteristic is the decay of the density coupled with an increasing spread with respect to the radial distance as it deemphasizes the importance of information far from the center. It is motivated by the principle of nearsightedness [21] that underpins, as discussed in Section 1.1 the decomposition of the structural representation into local atom-centered contributions. To reduce redundancy within the basis set, and considering that the dot product serves as a natural measure of similarity, orthogonality is enforced between the basis functions, thereby avoiding redundancy

$$\int_{\mathbb{R}} \mathrm{d}\mathbf{r}\, R_n(\mathbf{r}) R_{n'}(\mathbf{r}) = 0 \quad \text{for } n \neq n' \quad \text{(orthogonality)}. \tag{1.20}$$

If the chosen basis does not inherently provide orthogonality, it is typically enforced a posteriori with a Löwden orthogonalization [40]. Another shared characteristic is the uniform distribution of the basis functions across the interval $[0, r_c]$ providing an initial guess for representing the radial space [39, 41, 42]. This can be proceeded by a subsequent optimization step of the basis according to the radial distribution of the dataset.

1.3.3 Angular basis

The choice of the spherical harmonics as angular basis is essentially fixed, since they form an irreducible representation of $SO(3)$ and thus cannot be further compressed without diminishing the representation space. One direction of angular dependent optimization has been therefore to bias the construction of the radial basis for each angular channel separately by

the criteria of maximal variance [43] or maximal smoothness [44]. While in principle similar optimizations across the angular channels, mixing them, are possible, a strict preservation of the angular channels in the spherical harmonics is needed to propagate to higher orders as shown in Eq. (1.18) or Ref. [34]. Due to this limitation, recent advancements have been more focused on improving the computation of the spherical harmonics itself by exploiting recursive relationships in its gradients [45] or switching to a Cartesian tensor basis [38, 46, 47]

$$\mathbf{T}^{(\nu)} = \hat{\mathbf{r}}_{ji} \otimes \cdots \otimes \hat{\mathbf{r}}_{ji} \in \mathbb{R}^{3^\nu} \quad \text{(Cartesion moment tensor of order } \nu\text{)}. \tag{1.21}$$

While the Cartesian moment tensor forms a reducible angular basis, in return it allows a more efficient computation of the angular components at a minimal loss of accuracy [48].

1.3.4 Decomposition of the basis expansion

When deriving an expresion for the spherical expansion coefficients in Eq. (1.19a), the coefficients naturally decompose into neighbor contributions

$$c^i_{nlm} = \sum_{j \in A_i} \int_{\mathbb{R}^3} \mathrm{d}\mathbf{r}\, R_n(r)\, Y^l_m(\hat{\mathbf{r}}) \rho(\mathbf{r} - \mathbf{r}_{ji}) = \sum_{j \in A_i} c^{ij}_{nlm} \text{ (neighbor expansion coefficients)}. \tag{1.22}$$

Each neighbor coefficient further decomposes into a *radial expansion coefficient* c^{ij}_{nl} and *angular expansion coefficient* c^{ij}_{lm}

$$c^{ij}_{nlm} = c^{ij}_{nl} c^{ij}_{lm}. \tag{1.23}$$

The dependency of the radial coefficients on the angular component appears due to the coupling of the radial and angular contributions in the Gaussian density. This coupling limits the choice of the type of radial basis function that lead to an analytical expression of the integral. If no analytical solution can be derived a numerical integration is required that is typically more costly [39]. One approach has been therefore to express the atomic density into a form that disentangles the radial and angular contribution [49]

$$\rho_i(\mathbf{r}) = \rho_{i,r}(r)\, \rho_{i,\perp\mathbf{r}}(\hat{\mathbf{r}}). \tag{1.24}$$

Nevertheless, this approach removes the information that is encoded in this coupling. Instead, the information can be preserved at minimum computational cost by splining the radial expansion coefficients. For each coefficient nl, the one-dimensional function $f^{nl} : [0, r_c] \to \mathbb{R}$, that returns the radial expansion coefficients $f(r_{ji}) = c^{ij}_{nl}$, is splined. More technical details about the splining of the radial expansion coefficients can be found in Section 4.1. As this approach avoids the cost of the integration it opens the door to a wider choice of the basis while preserving the information contained in the coupling [43, 44, 50].

So far we only expressed the coefficients for the case of a single chemical species. To extend the representation so it can encapsulate different information for each species, an additional channel for each neighbor species a_j of atom j is included into the coefficients separating the neighbor contributions into different dimensions

$$c^i_{anlm} = \sum_{j \in A_i} \delta_{aa_j} c^{ij}_{nlm}. \tag{1.25}$$

Including the species information increases the dimensionality of the numerical description by a multiplicative factor dependent on the number of species. This growth in feature size becomes even more severe when combining it with an increase in the body-order. Therefore, two major approaches for the compression of the species channels have been proposed. One approach linearly combines all n_{species} species channels to a reduced number of n_{pseudo} channels [50, 51]

$$c^i_{bnlm} = \sum_a U_{ba} c^i_{anlm}, \quad \mathbf{U} \in \mathbb{R}^{n_{\text{pseudo}} \times n_{\text{species}}} \tag{1.26}$$

with b often referred as *pseudo species*. This approach and its extension to an optimization of the radial basis is in more detail explored in Chapter 3. The other approach learns a embedding c^{ij}_{ak} for each species a that is seperate from the basis expansion coefficients. An embedding can be expressed as an *one-hot encoding* of the species a with a subsequent linear transformation. An one-hot encoding of species a is a vector with one nonzero entry at the dimension corresponding to species a.

$$\mathbf{z}_a = [0, \dots, 1, 0, \dots, 0] \in \mathbb{R}^{n_{\text{species}}}. \tag{1.27}$$

Then for each species a linear weight can be learned that when multiplied with $\mathbf{z}_a$ returns the embedding

$$c^{ij}_{ak} = [\mathbf{U}\mathbf{z}_a]_k, \quad \mathbf{U} \in \mathbb{R}^{n_{\text{embedding}} \times n_{\text{species}}}, \quad \text{(embedding)} \tag{1.28}$$

This embedding is subsequently combined with the basis expansion coefficients by an addition or a nonlinear transformation [41]. The former approach retains a clear formulation of the chemical information entering the model that is performing the prediction, while the latter one looses interepretability due to the combination of the embedding and basis coefficients. Both approaches can be extended to include species information from the central atom in their numerical description.

2 Measures of information capacity

In the last chapter we learned about the core ideas that underlie most of the existing atom-centred description schemes that are particularly well-suited to model additive, extensive properties, and the incorporation of geometric and atom permutation symmetries. While incorporation of symmetries makes representations much more data efficient, it raises subtle issues of whether the mapping from structure to descriptor is injective or not [10, 52, 53]. Many of the structural representations that fulfill these symmetry requirements are closely related to one another, corresponding to projections of n-body correlations of the atom density [17, 24]. Yet, comparing them is not straightforward. When used to build an interatomic potential, or to predict another atomic-scale property, representations are used together with different supervised learning schemes, so it is difficult to disentangle the interplay of descriptor, regression method, and target property that combine to determine the accuracy and computational cost of the different methods. [16] Juxtaposing alternative choices of representations is complicated by the fact that non-linear transformations are often applied as a part of the data processing algorithm, and so it would be equally important to be able to analyze the effect of these transformations. Efforts to compare different choices of descriptors have been mostly focused this far on a comparison of compressibility [54, 55], their ability to represent atomic structures uniquely [10, 53, 56, 57], their role in constructing a metric [58, 59] and their sensitivity to perturbations of the atomic structure [55, 60].

Here we propose a strategy to compare feature spaces both in terms of their mutual information content – which we define transparently as the ability to linearly or non-linearly reconstruct each other – and in terms of the amount of deformation that has to be applied to match the common information between the two. Note that the definition we use here differs from that used in information-theoretical treatments, based on Shannon entropy – which is difficult to compute in high dimensions [61], and does not reflect as naturally the behavior of different features when used in the context of atomistic machine learning.

This strategy is demonstrated by applying it to elucidate several issues related to the behavior of density-based representations. First, we investigate the role of the basis and of the density smearing in the practical implementation of 3-body density correlation features; we then estimate the loss of information that one incurs by truncating the description to low body-order of correlations; finally, we discuss the role of the metric used to compare two structures, by testing the commonly used Euclidean distance against kernel-induced and Wasserstein-type metrics. An open-source implementation of functions to compute these quantities is provided in the package `scikit-matter`[62].

2.1 Comparing feature spaces

Consider a dataset $\mathscr{D} = \{x_i\}$ containing n items. For a given choice of features $\mathscr{F}$, each item is described by an $m_{\mathscr{F}}$-dimensional feature vector $\mathbf{x}_i$. As a whole, the dataset is described by a feature matrix $\mathbf{X}_{\mathscr{F}}^{\mathscr{D}} \in \mathbb{R}^{n \times m_{\mathscr{F}}}$. We consider all of the feature matrices in this work to be standardized, i.e. centred and scaled so as to have zero mean and unit variance for the selected data set. Consider a second featurization $\mathscr{F}'$. We want to be able to compare the behavior of different choices of feature spaces when representing the dataset $\mathscr{D}$, e.g. which of two sets of features have more expressive power, and how much distorted is one representation relative to the other.

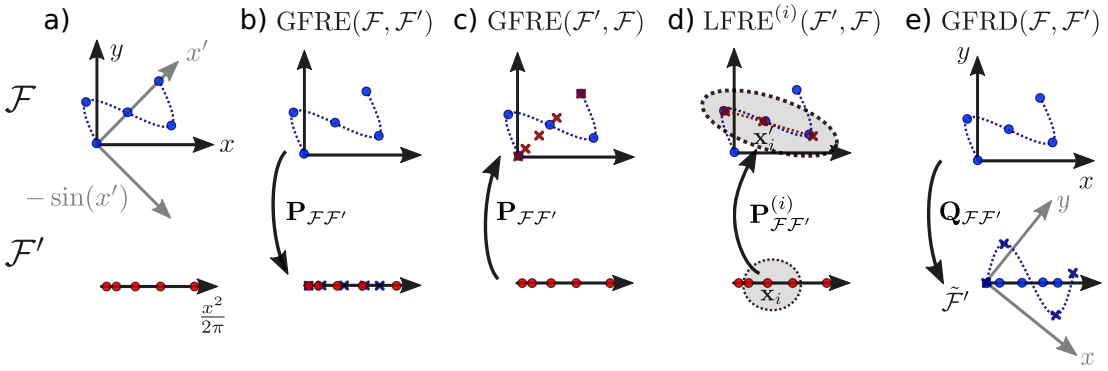

Figure 2.1: A schematic representation of the different measures of feature-space dissimilarity we introduce in this work, discussed from left to right. The figure considers a dataset containing five samples, embedded in a two-dimensional feature space $\mathscr{F}$ and a one-dimensional feature space $\mathscr{F}'$. As shown, the relationship between the two embeddings can involve arbitrary linear and non-linear transformations (panel a). The global feature space reconstruction error (GFRE) defined in Equation (2.2) amounts to finding the best linear mapping between the two feature spaces. This measure is not symmetric: in this example the GFRE($\mathscr{F}, \mathscr{F}'$) (panel b) is smaller than the GFRE($\mathscr{F}', \mathscr{F}$) (panel c), since $\mathscr{F}$ contains an additional nonzero dimension. The local version of the reconstruction error (LFRE) defined in Equation (2.7) makes it possible to probe whether a non-linear map exists between the two spaces: in this case, the sinus function can be approximated in neighborhood of each sample $\mathbf{x}_i$ by a linear map $\mathbf{P}^{(i)}_{\mathscr{F}\mathscr{F}'}$ defined in Equation (2.6); by treating each neighborhood separately it is possible to achieve a low LFRE($\mathscr{F}', \mathscr{F}$) (panel d). Finally, the global feature space reconstruction distortion (GFRD) defined in Equation (2.5) determines whether the two featurizations are connected by an orthogonal transformation $\mathbf{Q}_{\mathscr{F}\mathscr{F}'}$ defined in Equation (2.4), by finding the best alignment between $\mathscr{F}$ and the approximation of $\mathscr{F}'$ that can be obtained as a linear projection of $\mathscr{F}$. Even though GFRE($\mathscr{F}, \mathscr{F}'$) is small, one of the components of $\mathscr{F}$ is scaled down to zero, resulting in a large value of GFRD($\mathscr{F}, \mathscr{F}'$) (panel e).

2.1.1 Global feature space reconstruction error

As a simple, easily-interpretable measure of the relative expressive power of $\mathscr{F}$ and $\mathscr{F}'$, we introduce the global feature space reconstruction error GFRE$^{\mathscr{D}}(\mathscr{F}, \mathscr{F}')$, defined as the mean-square error that one incurs when using the feature matrix $\mathbf{X}_{\mathscr{F}}$ to linearly regress $\mathbf{X}_{\mathscr{F}'}$. In this work we compute the GFRE by a 2-fold split of the dataset, i.e. compute the regression weights

$\mathbf{P}_{\mathscr{F}\mathscr{F}'}$ over a train set $\mathscr{D}_{\text{train}}$ composed of half the entries in $\mathscr{D}$,

$$\begin{aligned}\mathbf{P}_{\mathscr{F}\mathscr{F}'} &= \underset{\mathbf{P}\in\mathbb{R}^{m_{\mathscr{F}}\times m_{\mathscr{F}'}}}{\operatorname{argmin}} \left\| \mathbf{X}_{\mathscr{F}'}^{\mathscr{D}_{\text{train}}} - \mathbf{X}_{\mathscr{F}}^{\mathscr{D}_{\text{train}}}\mathbf{P} \right\| \\ &= \left({\mathbf{X}_{\mathscr{F}}^{\mathscr{D}_{\text{train}}}}^T \mathbf{X}_{\mathscr{F}}^{\mathscr{D}_{\text{train}}} \right)^{-1} (\mathbf{X}_{\mathscr{F}}^{\mathscr{D}_{\text{train}}})^T \mathbf{X}_{\mathscr{F}'}^{\mathscr{D}_{\text{train}}}\end{aligned} \tag{2.1}$$

and then compute the error over the remaining test set $\mathscr{D}_{\text{test}}$

$$\text{GFRE}^{\mathscr{D}}(\mathscr{F},\mathscr{F}') = \sqrt{\left\| \mathbf{X}_{\mathscr{F}'}^{\mathscr{D}_{\text{test}}} - \mathbf{X}_{\mathscr{F}}^{\mathscr{D}_{\text{test}}}\mathbf{P}_{\mathscr{F}\mathscr{F}'} \right\|^2 / n_{\text{test}}}, \tag{2.2}$$

averaging, if needed, over multiple random splits. The GFRE is a positive quantity, which is equal to zero when there is no error in the reconstruction, and that is usually bound by one[1]. For numbers of features larger than n_{train}, the covariance matrix is not full rank, and one needs to compute a pseudoinverse. Without loss of generality, one can regularize the regression to stabilize the calculation. In this paper, we computed the pseudoinverse by means of a singular value decomposition, and we determined the optimal regularization in terms of the truncation of the singular value spectrum, using 2-fold cross-validation over the training set to determine the optimal truncation threshold. Often, it is also useful to observe the behavior of the GFRE in the absence of any regularization: overfitting is in itself a signal of the instability of the mapping between feature spaces. In general, $\text{GFRE}^{\mathscr{D}}(\mathscr{F},\mathscr{F}')$ is not symmetric. If $\text{GFRE}^{\mathscr{D}}(\mathscr{F},\mathscr{F}') \approx \text{GFRE}^{\mathscr{D}}(\mathscr{F}',\mathscr{F}) \approx 0$, $\mathscr{F}$ and $\mathscr{F}'$ contain similar types of information; if $\text{GFRE}^{\mathscr{D}}(\mathscr{F},\mathscr{F}') \approx 0$, while $\text{GFRE}^{\mathscr{D}}(\mathscr{F}',\mathscr{F}) > 0$, one can say that $\mathscr{F}$ is more descriptive than $\mathscr{F}'$: this is the case, for instance, one would observe if $\mathscr{F}'$ consists of a sparse version of $\mathscr{F}$, with some important and linearly-independent features removed; finally, if $\text{GFRE}^{\mathscr{D}}(\mathscr{F},\mathscr{F}') \approx \text{GFRE}^{\mathscr{D}}(\mathscr{F}',\mathscr{F}) > 0$, the two feature spaces contain different, and complementary, kinds of information and it may be beneficial to combine them to achieve a more thorough description of the problem.

2.1.2 Global feature space reconstruction distortion

The feature space reconstruction error gives insights into whether a feature space can be inferred by knowledge of a second one. However, having both a small $\text{GFRE}^{\mathscr{D}}(\mathscr{F},\mathscr{F}')$ and $\text{GFRE}^{\mathscr{D}}(\mathscr{F}',\mathscr{F})$ does not imply two feature spaces are identical. Even though they contain similar amounts of information, one feature space could give more emphasis to some features compared to the other, which can eventually result in different performance when building a model. To assess the amount of distortion of $\mathscr{F}'$ relative to $\mathscr{F}$, we introduce the global feature space reconstruction distortion $\text{GFRD}^{\mathscr{D}}(\mathscr{F},\mathscr{F}')$. To evaluate it, we first compute the singular value decomposition of the projector Equation (2.1), $\mathbf{P}_{\mathscr{F}\mathscr{F}'} \approx \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, and then use it to reduce the two feature spaces to a common basis, in which the reconstruction error is zero, because the residual has been discarded

$$\tilde{\mathbf{X}}_{\mathscr{F}} = \mathbf{X}_{\mathscr{F}}\mathbf{U} \quad \tilde{\mathbf{X}}_{\mathscr{F}'} = \tilde{\mathbf{X}}_{\mathscr{F}}\mathbf{\Sigma}. \tag{2.3}$$

When the second feature space $\mathscr{F}'$ has a lower dimensionality than $\mathscr{F}$, some combinations of the starting features are not used to compute $\tilde{\mathscr{F}}'$. In this case, we pad $\mathbf{\Sigma}$ with zeros, so that $\tilde{\mathscr{F}}'$ has the same dimensionality $m_{\mathscr{F}}$ as the starting space. This choice ensures that the GFRD

[1] This is due to the fact that feature matrices are standardized, and so $\left\| \mathbf{X}_{\mathscr{F}'}^{\mathscr{D}_{\text{test}}} \right\| / n_{\text{test}}$ is of the order of one

takes the same value it would have in the case $\mathscr{F}'$ had the same dimensionality as $\mathscr{F}$, but lower rank. In the opposite case, with $m_{\mathscr{F}} < m'_{\mathscr{F}}$, padding $\mathbf{\Sigma}$ and $\mathbf{U}$ with zeros, or truncating $\mathbf{V}$, yields the same GFRD.

We can then address the question of whether $\tilde{\mathbf{X}}_{\mathscr{F}}$ and $\tilde{\mathbf{X}}_{\mathscr{F}'}$ are linked by a unitary transformation (in which case the GFRD should be zero), or there is a distortion involved. A possible answer involves solving the orthogonal Procrustes problem [63] – i.e. finding the orthogonal transformation that "aligns" as well as possible $\tilde{\mathbf{X}}_{\mathscr{F}}$ to $\tilde{\mathbf{X}}_{\mathscr{F}'}$:

$$\begin{aligned}\mathbf{Q}_{\mathscr{F}\mathscr{F}'} =& \underset{\mathbf{Q}\in\mathbb{U}^{m\times m}}{\operatorname{argmin}} \left\| \tilde{\mathbf{X}}^{\mathscr{D}_{\text{train}}}_{\mathscr{F}'} - \tilde{\mathbf{X}}^{\mathscr{D}_{\text{train}}}_{\mathscr{F}} \mathbf{Q} \right\| \\ =& \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T,\end{aligned} \tag{2.4}$$

where $\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T = (\tilde{\mathbf{X}}^{\mathscr{D}_{\text{train}}}_{\mathscr{F}})^T \tilde{\mathbf{X}}^{\mathscr{D}_{\text{train}}}_{\mathscr{F}'}$. The amount of distortion can then be computed by assessing the residual on the test set,

$$\text{GFRD}^{\mathscr{D}}(\mathscr{F},\mathscr{F}') = \sqrt{\left\| \tilde{\mathbf{X}}^{\mathscr{D}_{\text{test}}}_{\mathscr{F}'} - \tilde{\mathbf{X}}^{\mathscr{D}_{\text{test}}}_{\mathscr{F}} \mathbf{Q}_{\mathscr{F}\mathscr{F}'} \right\|^2 / n_{\text{test}}}. \tag{2.5}$$

If desired, the error can be averaged over multiple random splits of the reference data set $\mathscr{D}$.

2.1.3 Local feature space reconstruction error

A downside of the global feature comparison schemes introduced above is that the linear nature of the regression means that they cannot detect if $\mathscr{F}$ and $\mathscr{F}'$ contain analogous information, but differ by a non-linear transformation. In the next Section we discuss how one can generalize the schemes to use kernel features, that can also be used to detect non-linear relationships between the original feature spaces. An alternative approach is to compute a local version of the feature space reconstruction error, $\text{LFRE}^{\mathscr{D}}(\mathscr{F},\mathscr{F}')$, loosely inspired by locally-linear embedding [64]. To compute the LFRE, a local regression is set up, computed in the k-neighborhood $\mathscr{D}^{(i)}_{k-\text{neigh}}$ around sample i – the set of k nearest neighbors of sample i, based on the Euclidean distance between the samples in $\mathscr{F}$ – to reproduce the $\mathscr{F}'$ features using $\mathscr{F}$ as input features, centred around their mean values $\bar{\mathbf{x}}_{\mathscr{F}'}$ and $\bar{\mathbf{x}}_{\mathscr{F}}$.

A local embedding of $\mathbf{x}_i$ is determined as

$$\tilde{\mathbf{x}}'_i = \bar{\mathbf{x}}_{\mathscr{F}'} + (\mathbf{x}_i - \bar{\mathbf{x}}_{\mathscr{F}})\mathbf{P}^{(i)}_{\mathscr{F}\mathscr{F}'}, \tag{2.6}$$

where $\mathbf{P}^{(i)}_{\mathscr{F}\mathscr{F}'}$ contains the regression weights computed from $\mathscr{D}^{(i)}_{k-\text{neigh}}$. The local feature space reconstruction error is given by the residual discrepancy between the $\mathscr{F}'$ counterpart of the i-th point and its local embedding (2.6):

$$\text{LFRE}^{\mathscr{D}}(\mathscr{F},\mathscr{F}') = \sqrt{\sum_i \left\| \mathbf{x}'_i - \tilde{\mathbf{x}}'_i \right\|^2 / n_{\text{test}}}. \tag{2.7}$$

Inspecting the error associated with the reconstruction of individual points can reveal regions of feature space for which the mapping between $\mathscr{F}$ and $\mathscr{F}'$ is particularly problematic. Similarly, one can compute a local version of GFRD, that could be useful to detect strong local distortions that might indicate the presence of a singularity in the mapping between two feature spaces.

2.1.4 Bending space: comparing induced feature spaces

It is often possible to substantially improve the performance of regression or dimensionality reduction algorithms, without explicitly changing the feature vectors. This can be achieved by introducing a (non-linear) similarity measure to compare $\mathbf{x}_i$, which takes the form of a kernel function $k(\mathbf{x},\mathbf{x}')$, or a dissimilarity measure which takes the form of a distance $d(\mathbf{x},\mathbf{x}')$.

Let us recall that a positive-definite kernel induces a kernel distance by the relation[65]

$$d_k(\mathbf{x},\mathbf{x}')^2 = k(\mathbf{x},\mathbf{x}) + k(\mathbf{x}',\mathbf{x}') - 2k(\mathbf{x},\mathbf{x}'), \tag{2.8}$$

and that any negative-definite distance can be used to build positive-definite kernels such as the substitution kernel[66]

$$k_d^{\mathbf{x}_0}(\mathbf{x},\mathbf{x}') = -\frac{1}{2}(d(\mathbf{x},\mathbf{x}')^2 - d(\mathbf{x},\mathbf{x}_0)^2 - d(\mathbf{x}_0,\mathbf{x}')^2), \mathbf{x}_0 \in \mathscr{F} \tag{2.9}$$

or the radial basis function (RBF) kernel

$$k_d^{\mathrm{RBF}}(\mathbf{x},\mathbf{x}') = \exp\left(-\gamma d(\mathbf{x},\mathbf{x}')^2\right), \quad \gamma \in \mathbb{R}_+ \tag{2.10}$$

A positive definite kernel induces a feature space $\mathscr{H}$, commonly known as reproducing kernel Hilbert space (RKHS), in which the similarity measure can be expressed as a dot product:

$$k(\mathbf{x},\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{x}'), \quad \mathbf{x},\mathbf{x}' \in \mathscr{F},\ \boldsymbol{\phi} : \mathscr{F} \to \mathscr{H}. \tag{2.11}$$

While in general $\boldsymbol{\phi}(\mathbf{x})$ is not known, for a given dataset $\mathscr{D}$ it is possible to approximate the RKHS features by using a kernel principal component analysis [67]. Since linear regression in RKHS features is equivalent to kernel ridge regression, we simply use kernel features computed on the training dataset $\mathscr{D}_{\text{train}}$ to reduce the problem of comparing kernel (or distance) induced features to that of comparing explicit features, and use GFRE and GFRD as defined in Eqs. (2.2) and (2.5). It is possible to re-formulate these measures in an explicit kernelized form, as well as to compute low-rank approximations of the kernel to reduce the computational cost for very large datasets (see e.g. Ref. 54 for a pedagogic discussion). In this paper we simply use the explicit RKHS features, that can be obtained by diagonalizing the kernel matrix $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, with $K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$, and defining

$$\mathbf{X}_{\mathscr{H}} = \mathbf{U}\boldsymbol{\Lambda}^{-1/2}, \tag{2.12}$$

which is then standardized as we do for any other set of features. To define a feature space associated with a metric, rather than a kernel, we first center the squared distance matrix (which is equivalent to computing a substitution kernel analogous to Equation (2.9)) and then proceed similarly by diagonalizing the resulting matrix.

2.1.5 Dataset selection

We use four different datasets, chosen to emphasize different aspects of the problem of representing atomic structures: A *random methane* dataset consisting of different random displacements of the four hydrogen atoms around the central carbon atom to cover the complete configurational space of CH_4 structures; A *carbon* dataset of approximately 10′000 minimum

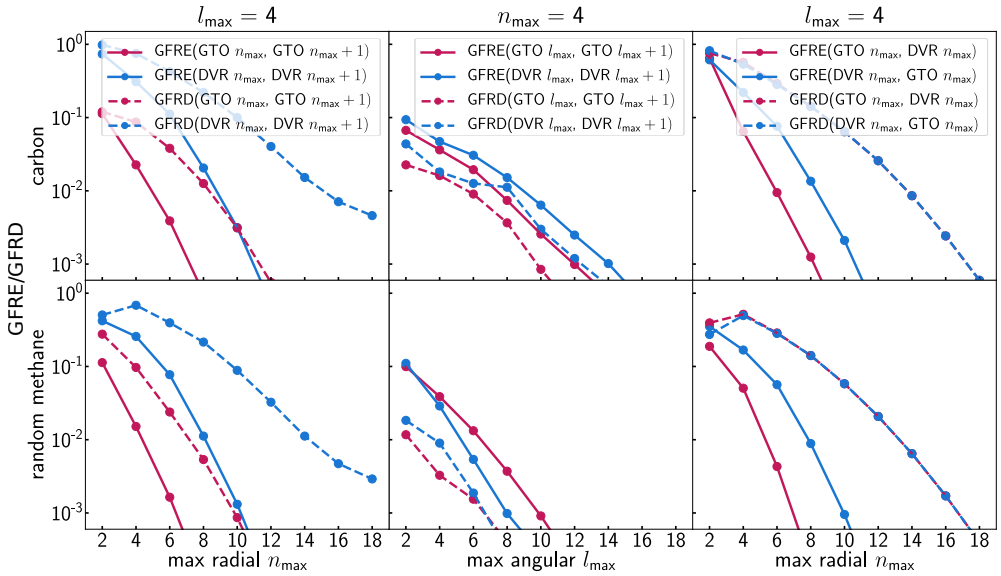

Figure 2.2: Comparison of the GFRE and GFRD for increasing numbers of radial (left, with fixed $l_{max} = 4$) and angular (middle, with fixed $n_{max} = 4$) basis functions. On the right, an explicit comparison of the two basis sets in terms of GFRE(GTO, DVR), GFRE(DVR, GTO) and the corresponding measures of distortion.

energy carbon structures, obtained as the result of ab initio random structure search [68, 69], as an example for a realistic dataset of condensed phase structures; A *degenerate methane* dataset composed of two groups of methane structures (which we refer to as $\mathscr{X}^+$ and $\mathscr{X}^-$), each associated with a 2D manifold parameterised by two parameters (u, v): structures with $v = 0$ in the two manifolds have exactly the same C-centred 3-body correlations, despite being different (as discussed in Ref. 53); A *displaced methane* dataset, which consists in an ideal, tetrahedral CH_4 geometry with one hydrogen atom pulled away from the central carbon atom, as an example of a set of structures that are distinguished by a clearly identifiable structural feature, here the C−H distance.

2.2 Comparing ACDC representations

While different discretizations of the abstract vectors on a basis are a matter of computational convenience and affect the computational cost of different approaches [16], but their descriptive power becomes equivalent in the limit of a complete basis set. We demonstrate the use of the GFRE, LFRE and GFRD to assess with quantifiable measures the effect of some of the different choices one can make when designing a representation.

2.2.1 SOAP and BPFS

We begin by considering two practical realizations of atom-centred symmetrized features of order $\nu = 2$ in Equation (1.16) as implemented in `librascal`[70], and the BPSF[9] as implemented in the n2p2 package[71]. We consider two different basis sets here, Gaussian-

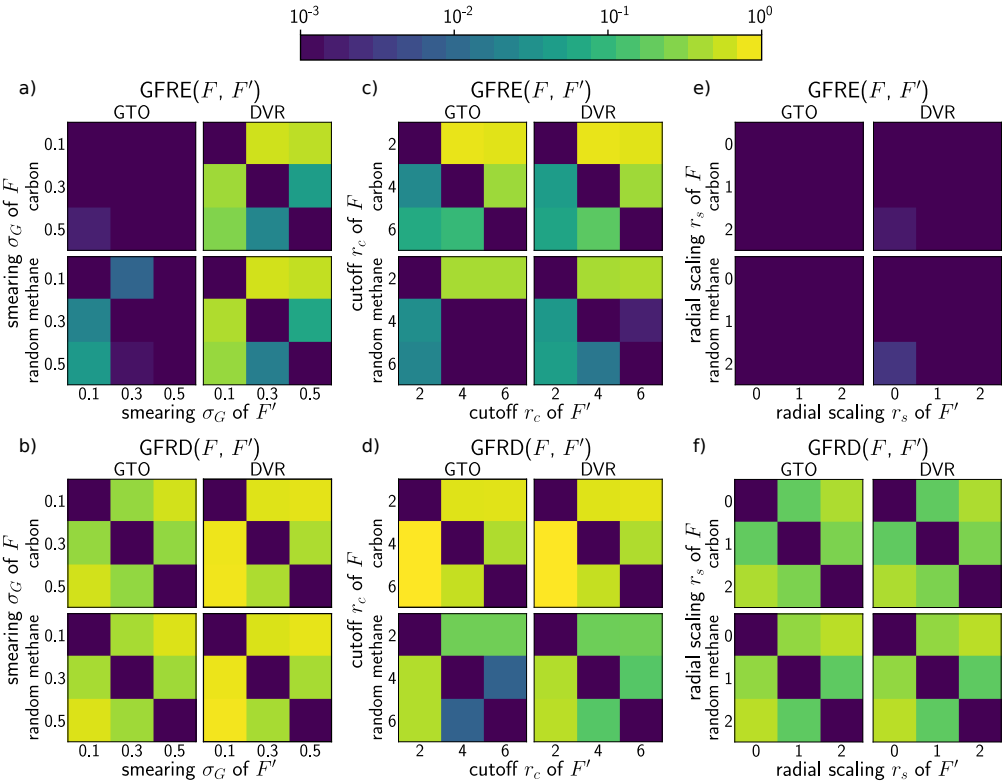

Figure 2.3: Comparison of the GFRE (top) and GFRD (bottom) a),b) for different smearing σ_G ($r_c = 4$Å) c),d) for different cutoff values ($\sigma_G = 0.5$Å), and e),f) for different radial scaling exponents ($r_c = 4$Å, $\sigma_G = 0.5$Å). For all comparisons $(n_{max}, l_{max}) = (10, 6)$ were used. The feature specified by the row is used to reconstruct the feature specified by the column.

type orbitals (GTO)

$$R_n^{\mathrm{GTO}} = N_n r^n \exp\left(-b_n r^2\right), \tag{2.13a}$$

$$\text{with } N_n = \frac{2}{\sigma_n^{2n+3}\Gamma((n+3)/2)}, \quad b_n = 1/(2\sigma_n), \quad \sigma_n = r_c \max(\sqrt{n}, 1)/n_{\max}, \tag{2.13b}$$

that are orthogonalized with respect to each other, and a discrete variable representation (DVR) basis

$$R_n^{\mathrm{DVR}} = \sqrt{w_n}\delta(r - r_n) \tag{2.14}$$

where r_n are Gaussian quadrature points and w_n their corresponding weights. For both bases, the integral (1.23) can be evaluated analytically, and the density coefficient computed as a sum over the neighbors of the i-th atom.

Even though they can be seen as a projection on an appropriate basis of the symmetrized atom density that underlies SOAP [17], BPSF are usually computed as a sum over tuples of neighboring atoms of functions of interatomic angles and distances. Among the many functional forms that have been proposed [72] we consider the two-body functions

$$G_i^{(2)} = \sum_j e^{-\eta(r_{ji} - R_s)^2} \cdot f_c(r_{ji}), \tag{2.15}$$

and the three-body functions

$$G_i^{(3)} = 2^{1-\zeta} \sum_j \sum_{k \neq j} (1 + \lambda \cdot \cos \hat{\mathbf{r}}_{ij} \cdot \hat{\mathbf{r}}_{ik})^{\zeta} \cdot e^{-\eta(r_{ji}^2 + r_{ik}^2 + r_{jk}^2)} \cdot f_c(r_{ji}) f_c(r_{ik}) f_c(r_{jk}), \tag{2.16}$$

where f_c is a cutoff function, and $\eta, \zeta, \lambda, R_s$ are parameters that define the shape of each BPSF. We generate systematically groups of symmetry functions of different size by varying the values of these parameters following the prescriptions discussed in Ref. 73. The list of values for the BPSF parameters we used are supplied in supplementary information.

GTO and DVR radial basis. We start by considering the convergence of the SOAP representation with different choices of radial basis. Figure 2.2 demonstrates the convergence with the number of radial functions $n_{\max}$ and angular momentum channels $l_{\max}$ (in a Cauchy sense, i.e. comparing results for successive increments of these parameters). Overall, the GTO basis converges faster than DVR for most cases, both in terms of GFRE and GFRD. The slower radial convergence of the GFRD indicates that even as the discretization approaches convergence, the changing position of peaks and nodes of the basis functions gives different emphasis to interatomic correlations over different ranges. This is consistent with the observation that, particularly for small $(n_{\max}, l_{\max})$, regression accuracy depends on the number of basis functions in a way that is not necessarily monotonic. When considering the convergence of the angular component $l_{\max}$, GTO and DVR show nearly identical error decay, indicating that the convergence of the radial and angular basis are largely independent of each other.

The faster convergence of the GTO basis suggests that, for a given $n_{\max}$, a representation expanded on this basis should contain a greater amount of information on the structure. This is reflected in the direct comparison of the two bases, GFRE(GTO $n_{\max}$, DVR $n_{\max}$) < GFRE(DVR $n_{\max}$, GTO $n_{\max}$) for small $n_{\max}$. When both basis set have converged, they become essentially equivalent. Since the two representations are related to each other by a unitary transformation, GFRD(GTO $n_{\max}$, DVR $n_{\max}$) $\to 0$ as $n_{\max} \to \infty$.

Gaussian smearing. The Gaussian smearing used in SOAP features works as a parameter controlling the balance between local resolution and the smoothness of the mapping between Cartesian coordinates and symmetrized density features. A small σ_{G} value can identify minute changes more accurately, but a too small value for σ_{G} can lead to ill-conditioned regression, as the features associated with different structures show little overlap with each other. In fact, there is a tight interplay between the density smearing, the choice of the basis set, and the regularization of a regression model. As seen in Figure 2.3(a,b), in the case of the smooth GTO basis set there is relativley little reconstruction error, and in general smaller σ_{G} values give a better reconstruction of large-σ_{G} features than vice versa. The opposite is true for the δ-like DVR basis: the GFRE for DVR is larger than in the case of GTO, and it is harder to reconstruct large-σ_{G} features from their sharp-Gaussian counterparts than vice versa. It should be also added that, without an automatic choice of regularization, results depend greatly on the way the feature mapping is executed. In particular, sharp-to-smooth mapping can lead to major overfitting problems, with GFRE becoming much larger than one for the test set. Even in cases where the GFRE is small, the feature space distortion is large, which highlights the fact that the Gaussian smearing changes significantly the emphasis given to different structural correlations, and can therefore affect the accuracy of regression models.

Radial cutoff and scaling. One of the most important hyperparameters when defining an atom-centred representation is the cutoff distance, which restricts the contributions to the density to the atoms with $r_{ji} < r_c$. Figure 2.3(c,d) shows that the GFRE captures the loss of information associated with an aggressive truncation of the environment, with very similar behavior between GTO and DVR bases. The figure also reflects specific features of the different data sets: for instance, GFRE($r_c = 4$ Å, $r_c = 6$ Å) is close to zero for the random methane data set, because there are no structures where atoms are farther than 4 Å from the centre of the environment. GFRE > 0 also when mapping long-cutoff features to short-range features, although the reconstruction error is much smaller than in the opposite direction. This indicates the need for an increase in $n_{\max}$ to fully describe the structure of an environment when using a large value of r_c, which is consistent with the greater amount of information encoded within a larger environment. The GFRD plot also underscores the strong impact of the choice of r_c on the emphasis that is given to different parts of the atom-density correlations. This effect explains the strong dependency of regression performance on r_c, and the success of multi-scale models that combine features built on different lengthscales [74]. A similar modulation of the contributions from different radial distances can be achieved by scaling the neighbor contribution to the atom-centred density by a decaying function, e.g. $1/(1+(r_{ji}/r_0)^s)$. This approach has proven to be very effective in fine-tuning the performance of regression models using density-based features [24, 75, 76]. As shown in Figure 2.3(e,f), this is an example of a transformation of the feature space that entails essentially no information loss – resulting in a very small GFRE between different values of the scaling exponent s. However, it does result in substantial GFRD, providing additional evidence of how the emphasis given by a set of features to different inter-atomic correlations can affect regression performance even if it does not remove altogether pieces of structural information.

Behler-Parinello symmetry functions. BPSF can be seen as projections of the same, abstract symmetrized density features that underlies the construction of SOAP features. While the latter representation is usually implemented using an orthogonal set of basis functions, BPSFs are non-orthogonal, and are usually chosen based on a careful analysis of the inter-atomic correlations that are relevant for a given system [9, 77, 78], or selected automatically out of a large pool of candidates [73]. Figure 2.4 shows clearly that an orthogonal basis set provides a more effective strategy to converge a representation than the grid-based enumeration of the non-linear hyperparameters of non-orthogonal basis functions. GFRE(SOAP, BPSF) < GFRE(BPSF, SOAP) for all feature set sizes and both data sets. As usual, we remark that zero reconstruction error does not imply equivalence for regression purpose: the GFRD remains very high even for the largest feature set sizes.

Given that, in real scenarios, one would usually combine systematic enumeration of BPSF features with an automatic selection method[73], we also use the feature reconstruction framework to investigate the convergence of the automatic screening procedure, i.e. the error in reconstructing the full vector based on the first m features chosen with a CUR decomposition-based procedure[73, 79]. Figure 2.5 shows that a few dozens CUR-selected features allow to almost-perfectly reconstruct the full feature vector. The convergence is particularly fast for BPSF, where $m = 50$ leads to a minuscule GFRE, indicating that the non-orthogonal features are highly redundant, and explaining the saturation in model performance that was observed in Ref. 73.

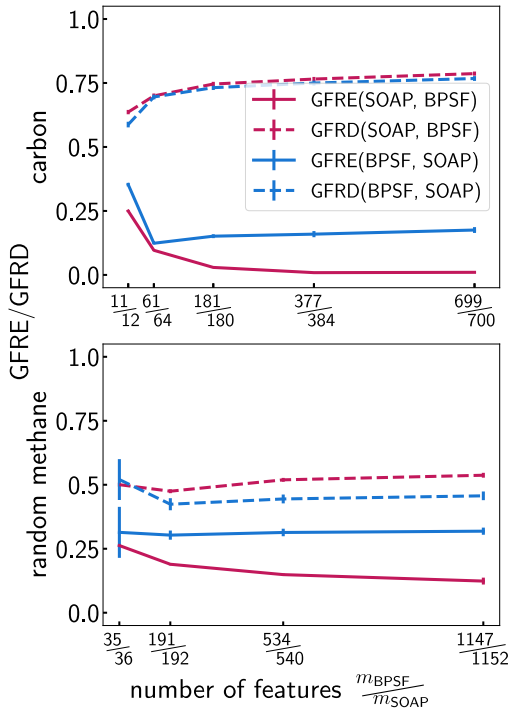

Figure 2.4: Comparison of the GFRE and the GFRD between SOAP(GTO) and BPSF features with systematically-increasing sizes of the feature vectors. BPSF features are generated by varying over a grid the hyperparameters entering the definitions of $G^{(2)}$ and $G^{(3)}$, following Ref. 73. SOAP expansion truncation parameters $(n_{\text{max}}, l_{\text{max}})$ are adjusted to approximately match the number of BPSF features.

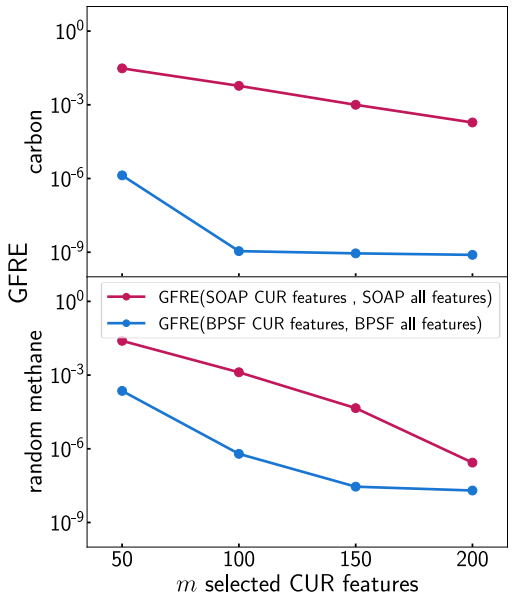

Figure 2.5: Convergence of a CUR approximation of the full SOAP/BPSF feature vectors (the largest size considered in Figure 2.4) with number of retained features.

$(n_{\max}, l_{\max})$	m_{SOAP}	$t_{\text{SOAP-GTO}}$ / s	$t_{\text{SOAP-DVR}}$ / s	m_{BPSF}	t_{BPSF} / s
(2,2)	12	8.08 ± 0.30	7.95 ± 0.03	11	3.90 ± 0.14
(4,3)	64	10.58 ± 0.03	9.57 ± 0.02	61	10.43 ± 0.22
(6,4)	180	13.87 ± 0.03	11.87 ± 0.14	181	30.20 ± 0.41
(8,5)	384	17.58 ± 0.05	14.50 ± 0.02	377	66.55 ± 0.55
(10,6)	700	23.02 ± 0.04	18.57 ± 0.03	699	124.60 ± 0.78

Table 2.1: Timings in seconds for the evaluation of SOAP features using GTO ($t_{\text{SOAP-GTO}}$) and DVR ($t_{\text{SOAP-DVR}}$) as radial basis, and of BPSF (t_{BPSF}), using $r_c = 4$ Å, on the entire carbon dataset. The SOAP discretization parameters are chosen to approximately match the number of features m_{SOAP} and m_{BPSF}. The measurements have been conducted on a single Intel(R) Xeon(R) CPU E3-1245 v6 @ 3.70GHz core, using librascal [70] for SOAP features and n2p2 [80] for BPSF.

Computational cost In this work we focus on the comparison of different kinds of features in terms of their information content, without commenting on the computational overhead associated with their evaluation, or the application of non-linear transformations. Computational cost depends on implementation choices, and can be optimized for usage patterns that differ from those that we apply here. However, the effort associated with the evaluation of a model plays an important role in determining its ultimate usability. To provide some context for our experiments, we report in Table 2.1 the timings for the evaluation of SOAP and BPSF features with the same parameters used in this Section. These timings show clearly that the computational savings afforded by the use of a DVR basis are not sufficient to offset the reduced information content with respect to the GTO basis. When comparing BPSF and SOAP, the clearest difference is that the cost of evaluating the former scales linearly with the number of features, while the cost of evaluating SOAP features is sublinear since it is dominated by the calculation of the density expansion coefficients c^i_{nlm} and is therefore sublinear with respect to the invariants. This difference in scaling is due to the different mechanism for evaluating 3-body terms, that scales quadratically with the number of neighbors for BPSF, and only linearly for SOAP, underscoring how a careful selection of the most relevant features is very important in the context of a BPSF framework.

2.2.2 Body order feature truncation.

The examples in Section 2.2.1 demonstrate the impact of implementation details and hyperparameters choices on the information content of features that are all equivalent to a three-body correlation of the atom density. A more substantial issue is connected to the use of representations based on different ν-body correlations of a decorated atom density, which is equivalent to the pair correlation function (2-body, $\nu = 1$), to the SOAP power spectrum (3-body, $\nu = 2$) or to the bispectrum (4-body, $\nu = 3$). Different orders incorporate conceptually distinct kinds of information: when used in linear regression, different density correlation orders correspond to a body-order expansion of the target property [17, 38, 81, 82, 83], and the link between the convergence of the body-order expansion and the injectivity of the structure-feature map is an open problem, with known counter-examples showing that low values of ν are insufficient to achieve a complete representation of an atomic environment [53].

Figure 2.6 shows that high-order features cannot be recovered as linear functions of lower-order features, while an approximate (if not complete) reconstruction of lower-ν components

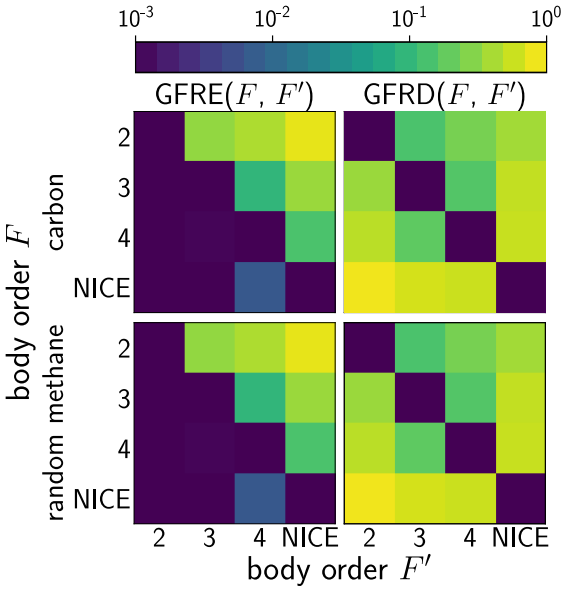

Figure 2.6: GFRE and GFRD body order comparison using GTO as radial basis function, $r_c = 4$Å, $\sigma_G = 0.5$Å and $(n_{\max}, l_{\max}) = (6,4)$. NICE features were computed keeping the top 400 equivariant components at each level of the body-order iteration, and keeping invariant components up to $\nu = 4$.

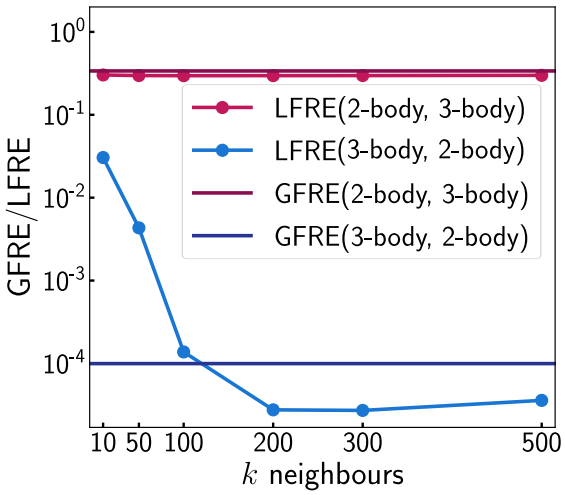

Figure 2.7: Convergence of the LFRE between 2 and 3-body density correlation features (using GTOs as radial basis, $r_c = 4$Å, $\sigma_G = 0.5$Å and $(n_{\max}, l_{\max}) = (6,4)$) with increasing number of neighbors.

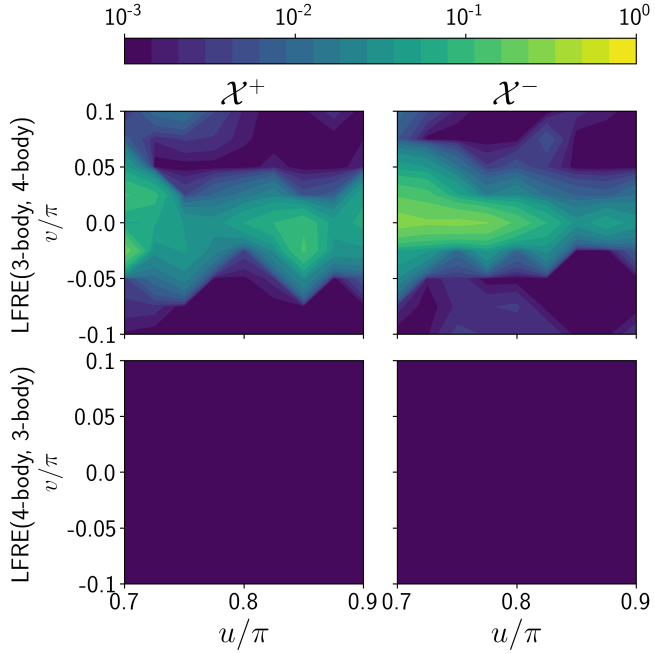

Figure 2.8: Pointwise LFRE for the structures from the degenerate methane dataset as a function of the structural coordinates (u, v) for $(n_{\max}, l_{\max}) = (6, 4)$ and $k = 15$ neighbors.

based on high-ν components is possible. Reconstructing features of different order entails a large amount of distortion, with the GFRD approaching one in most cases. We also include in the comparison features obtained with the recently-developed N-body iterative contraction of equivariants (NICE) framework, that identifies the most important features for each ν value, and uses them to compute $(\nu+1)$-order features [33]. Keeping 400 features for each body order is sufficient to achieve perfect reconstruction of 2 and 3-body features, but not for the 4-body (bispectrum) term, which cannot be reconstructed fully with 400 NICE features. Considering however that GFRE(NICE, $\nu = 3$) $\ll$ GFRE($\nu = 3$, NICE), one can infer that the of information loss associated with truncating the body order expansion is more severe than when restricting the number of 4-body features.

The comparison of features of different order can also be used to elucidate the role of the (non-)linearity of the mapping between feature spaces. Figure 2.7 compares global and local feature reconstruction errors between 2 and 3-body density correlation features, for the random CH_4 data set. In the case of the low-to-high body order reconstruction, the LFRE is only marginally lower than its global counterpart, indicating that the large GFRE($\nu = 1, \nu = 2$) is a consequence of lower information content and not only of the linear nature of the map. The reverse case is also revealing: for small k-neighborhood sizes, LFRE($\nu = 2, \nu = 1$) > GFRE($\nu = 2, \nu = 1$), because the small number of neighbors included in the model reduce the accuracy of the feature reconstruction map. When the number of neighbors approaches the intrinsic dimensionality of the $\nu = 2$ features, instead, LFRE < GFRE – because the reconstruction is based on a locally-linear map that can approximate a non-linear relationship between features. As k approaches the full train set size, the LFRE approaches the GFRE, as the locality of the mapping is lost.

The LFRE also makes it possible to identify regions of phase space for which the construction of a mapping between feature spaces is difficult or impossible. Consider the case of the

degenerate manifold discussed in Ref. 53. The dataset includes two sets of CH_4 environments, and those parameterised by $v = 0$ cannot be distinguished from each other using 3-body ($\nu = 2$) features. Figure 2.8 shows the LFRE for each point along the two manifolds. When trying to reconstruct 3-body features using as inputs 4-body features (that take different values for the two manifolds) the LFRE is essentially zero. When using the 3-body features as inputs, instead, one observes a very large error for points along the degenerate line, while points that are farther along the manifold can be reconstructed well. This example demonstrates the use of the LFRE to identify regions of feature space for which a simple, low-body-order representation is insufficient to fully characterize the structure of an environment, and can be used as a more stringent, explicit test of the presence of degeneracies than the comparison of pointwise distances discussed in Ref. 53.

2.2.3 Kernel-induced feature spaces

With the exception of the trivial, scalar-product form, a kernel introduces a non-linear transformation of the feature space, potentially allowing to obtain more accurate regression models. A crucial aspect of kernel methods is the fact that this non-linear transformation gives rise to a linear feature space that is defined by the combination of the kernel and the training samples – or the active samples in the case of sparse kernel methods. We can then use our feature-space reconstruction framework to compare quantitatively the linear feature space with the kernel-induced features. We do so using a radial basis function kernel, varying the γ parameter. In the $\gamma \to 0$ limit the RBF kernel becomes roughly linear, and the non-linearity increases with growing γ. The use of standardized input features means that γ is effectively unitless, and also standardize the kernel-induced features. To reduce noise, given that the kernel matrices are often very ill-conditioned, we only retain the RKHS features that are associated with the largest eigenvalues, preserving those that together contribute to approximately 99% of the variance. 1 Figure 2.10 plots the GFRE and GFRD for the mapping of linear and RBF features computed for 2 and 3-body density correlations. The non-linear nature of the transformation is apparent in the increase in the GFRE(linear,RBF) for larger values of γ, for both $\nu = 1$ and $\nu = 2$. The transformation is not entirely lossless, as evidenced by the fact that the reverse GFRE is also non-zero. The GFRE(RBF,linear) becomes particularly large for very large values of the γ parameter. This can be understood from the fact that the decay of the kernel becomes very sharp, and it only provides information about the nearest neighbors of each point – effectively leading to an ill-conditoned regression problem as we show in more detail below.

Having assessed the impact of non-linear kernel features on a single body order representation, we can then investigate whether a non-linear transformation helps inferring high-body order correlations from low-body-order features. This is relevant because the use of non-linear kernels has been proposed [81] (and used in practice for a long time [10, 84]) as a strategy to describe many-body effects on atomistic properties. We compute the GFRE for promoting $\nu = 1$ (2-body) to $\nu = 2$ (3-body) and $\nu = 2$ to $\nu = 3$ features for different values of the RBF kernel γ. In Figure 2.10 we show these curves for both the usual GFRE definition (that involves a separate test set) and for a prediction carried out on the train set. These results show that while a non-linear kernel does allow a low-body-order model to discern higher body-order features, it does so in a poorly transferable way: high-γ models show much reduced GFRE for train-set predictions, but lead to a degradation in the feature reconstruction for the test set.

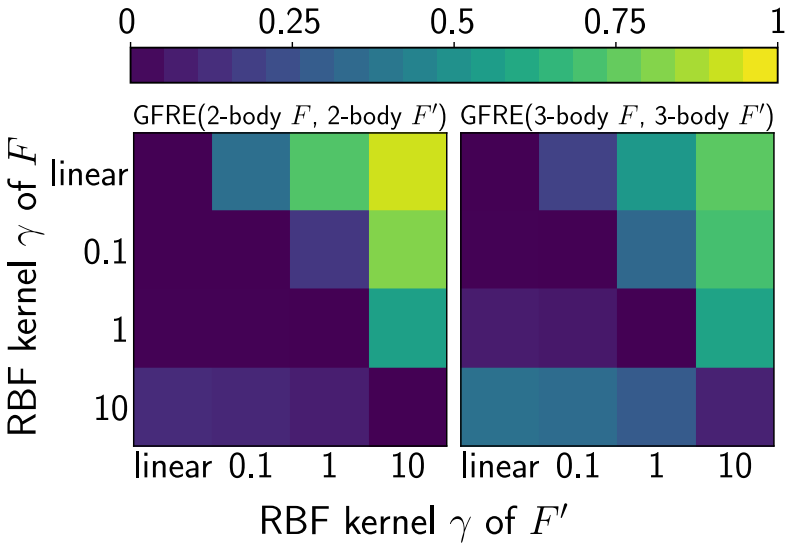

Figure 2.9: GFRE on the random methane dataset for interconverting the linear 2-body (left) and 3-body (right) feature spaces with those induced by a RBF kernel with different inverse kernel width γ.

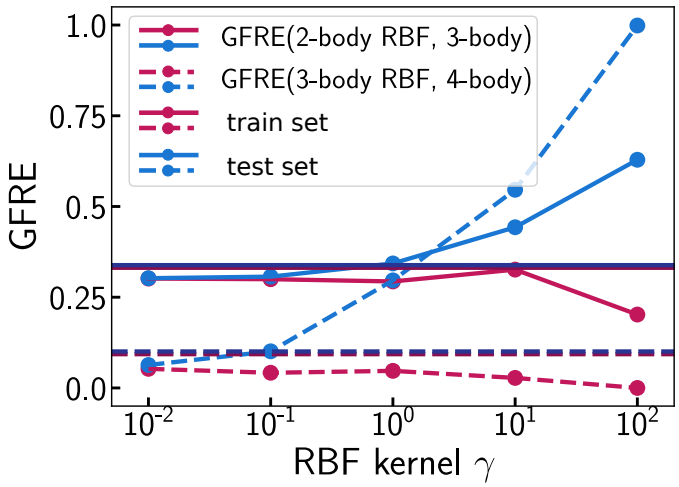

Figure 2.10: GFRE on the random methane dataset curves as a function of the γ hyperparameter of k_E^{RBF}. Values for train and test sets are plotted separately. The horizontal lines correspond to the GFRE of the linear features.

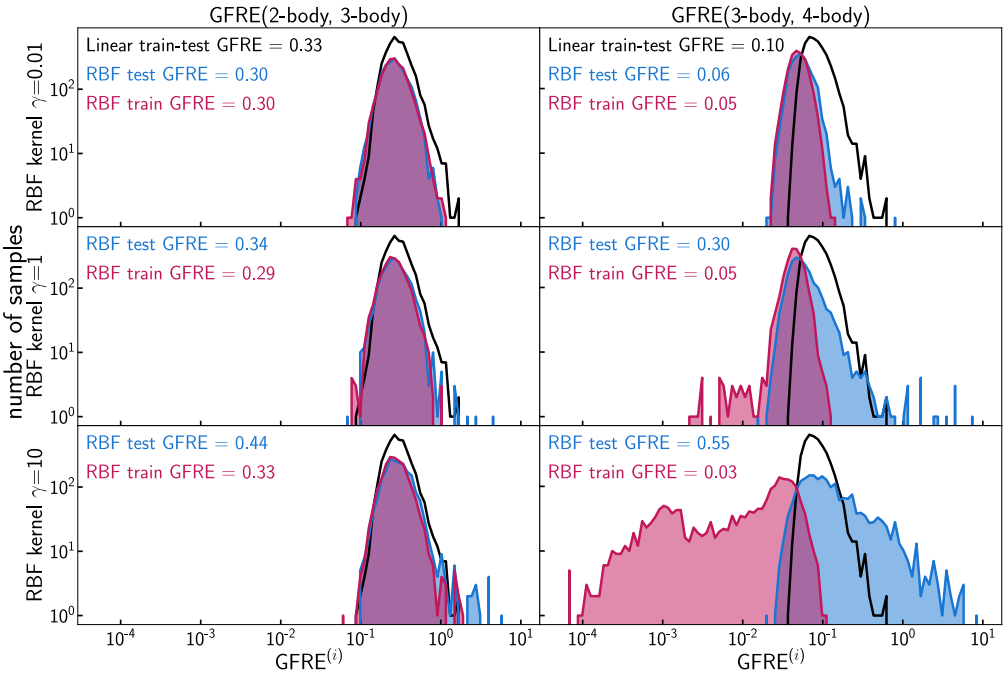

Figure 2.11: Histograms of the pointwise reconstruction error for $2 \to 3$ (left) and $3 \to 4$ (right) body order features, using a RBF kernel with different values of γ (top to bottom, $\gamma = 0.01$, 1.0, 10) to reconstruct the higher body order features. Red curves refer to the train set points, blue curves to the test set, and the black line correspond to the linear train-test set GFRE, that serves as a reference.

Only low-γ models show a small improvement in the test-set GFRE compared to an entirely linear mapping. In this regime, the RBF kernel is dominated by the low-exponent components of the Gaussian expansion, vindicating the choice of low-order polynomial kernels, that are used in most of the published SOAP-based potentials. A better understanding of the effect of a non-linear feature space transformation can be obtained by analyzing the distribution of reconstruction errors for individual samples

$$\mathrm{GFRE}^{(i)}(\mathscr{F},\mathscr{F}') = \left\| \mathbf{x}'_i - \mathbf{x}_i \mathbf{P}_{\mathscr{F}\mathscr{F}'} \right\|. \tag{2.17}$$

The histograms for this "pointwise GFRE" (Figure 2.11) show that increasing the non-linearity of the kernel does indeed allow to reconstruct more accurately a fraction of both the test and the train set. When extrapolating the mapping to points that have not been seen before, however, there is an increasingly large fraction of outliers for which the reconstruction is catastrophically poor.

The pointwise errors are also revealing of the different nature of the $\nu = 1 \to \nu = 2$ and $\nu = 2 \to \nu = 3$ cases. In the former case, the clear lack of information in the 2-body descriptor makes it impossible, even for a highly non-linear kernel, to obtain an accurate reconstruction of higher body-order features. In the latter case, instead, the train set reconstruction become nearly perfect with large γ – indicating that despite the existence of degenerate manifolds of configurations [53] it is possible to reconstruct 4-body features using only 3-body inputs, for structures that are not exactly on the degenerate manifold. However, the increasingly large tail of very high test-set GFRE samples suggests that this mapping is not smooth, and rather unstable. When building a regression model for a property that depends strongly on 4-body

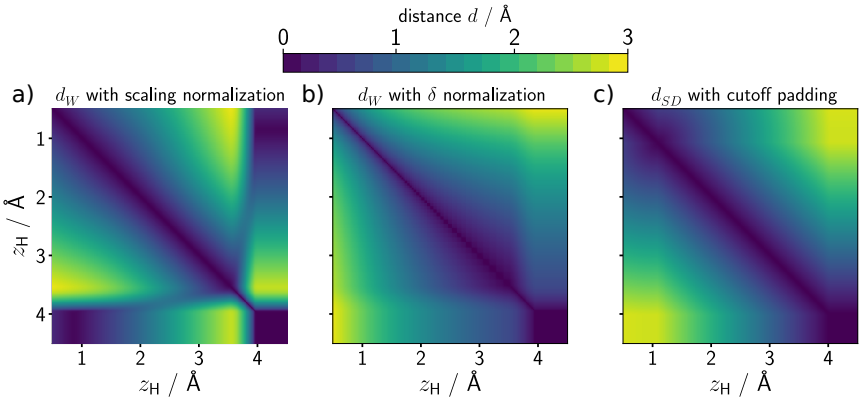

Figure 2.12: Distance between two *displaced methane* configurations with different values of z_H, computed using a Wasserstein distance using (a) scaling normalization; (b) cutoff δ normalization; (c) Euclidean distance between sorted interatomic distance vectors.

terms, this instability may translate into poor extrapolative power for a non-linear model based on 3-body features.

2.2.4 Wasserstein metric

As an example of the transformation induced by a non-Euclidean metric we consider the effect of using a Wasserstein distance to compare $\nu = 1$ density correlation features. The Wasserstein distance is defined as the minimum "work" that is needed to transform one probability distribution into another – with the work defined as the amount of probability density multiplied by the extent of the displacement [85, 86, 87]. The EMD has been used to define a "regularized entropy match" kernel to combine local features into a comparison between structures [59], to obtain permutation-invariant kernels based on Coulomb matrices [88]. Here we use the Wasserstein distance to compare two-body ($\nu = 1$) features, that can be expressed on a real-space basis and take the form of one-dimensional probability distributions.

The formal definition of the Wasserstein distance of order 2 between two probability distributions $p(r)$ and $p'(r)$ defined on a domain M reads

$$W(p,p')^2 = \inf_{\gamma\in\Gamma(p,p')} \int_{M\times M} d(r,r')^2 \,\mathrm{d}\gamma(r,r'), \tag{2.18}$$

where $\Gamma(p,p')$ is the set of all joint distributions with marginals p and p'. For 1-dimensional distributions, $W(p,p')$ can be expressed as the 2-norm of the difference between the associated inverse cumulative distribution function (ICDF) P^{-1} of two environments, $W(p,p')^2 = \int_0^1 \left|P^{-1}(s) - P'^{-1}(s)\right|^2 \mathrm{d}s$, with $P(r) = \int_0^r p(r)\,\mathrm{d}r$

In order to express the symmetrized 2-body correlation function as a probability density, we first write it on a real-space basis r, and evaluate it on 200 Gaussian quadrature points, that we also use to evaluate the CDF and its inverse. We then proceed to normalize it, so that it can be interpreted as a probability density. We estimate the integral of the distribution (that effectively counts the number of atoms within the cutoff distance)

$$Z_i = \int_0^{r_c} \overline{\rho_i^{\otimes 1}}(r)\,\mathrm{d}r, \tag{2.19}$$

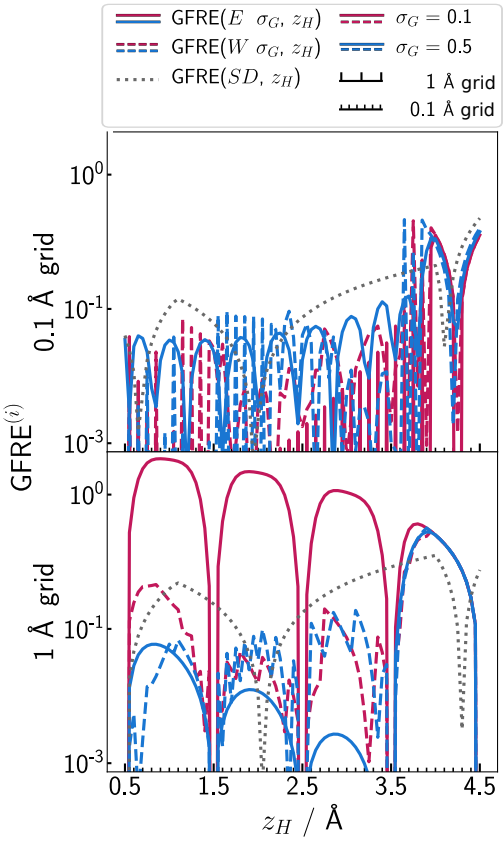

Figure 2.13: Errors when reproducing the atomic displacement z_H for a fine (top) and coarse (bottom) grid of training points, and different Gaussian σ_G and metrics. A constant regularization that discards singular values smaller than 10^{-3} has been applied to all pointwise GFRE calculations.

and the maximum value of the integral over the entire dataset $Z_{\mathscr{D}}$. A simple scaling of the correlation function

$$p_i^{\mathrm{s}}(r) = \frac{1}{Z_i}\overline{\rho_i^{\otimes 1}}(r) \tag{2.20}$$

distorts the comparison between environments with different numbers of atoms. To see how, we use the *displaced methane* dataset, in which three atoms in a CH_4 molecule are held fixed in the ideal tetrahedral geometry, at a distance of 1Å from the carbon centre. The fourth atom, aligned along the z axis, is displaced along it, so that each configuration is parameterised by a single coordinate z_{H}. Figure 2.12(a) shows the distance computed between pairs of configurations with different z_{H}, demonstrating the problem with the renormalized probability (2.20): p^{s} loses information on the total number of atoms within the cutoff, and so once the tagged atom moves beyond r_{c} the remaining CH_3 environment becomes indistinguishable from an ideal CH_4 geometry.

One can obtain a more physical behavior when atoms enter and leave the cutoff by introducing a δ-like "sink" at the cutoff distance, defining

$$p_i^{\delta}(r) = \frac{1}{Z_{\mathscr{D}}}\left[\overline{\rho_i^{\otimes 1}}(r) + (Z_{\mathscr{D}} - Z_i)\delta(r - r_c)\right]. \tag{2.21}$$

Figure 2.12b shows that with this choice the Wasserstein metric between $p_i^{\delta}(r)$ reflects the distance between the moving atoms. With this normalization, in fact, the Wasserstein metric corresponds to a smooth version of the Euclidean metric computed between vectors of sorted interatomic distances [17], shown in Figure 2.12c. The distortions that can be seen in the comparison between Figure 2.12b,c are a consequence of the Gaussian smearing, the smooth cutoff function, and the $SO(3)$ integration that modulates the contribution to $\overline{\rho_i^{\otimes 1}}(r)$ coming from atoms at different distances.

Having defined a meaningful normalization and a probabilistic interpretation of the radial density correlation features, we can investigate how the feature space induced by a Wasserstein metric relates to that induced by an Euclidean distance. Figure 2.13 shows the error in the reconstruction of z_{H} for the *displaced methane* dataset when restricting the training set to 0.05Å and 1.0Å spaced grids. Using a Euclidean distance with a sharp σ_{G} leads to a highly non-linear mapping between the displacement coordinate and feature space, and a linear model cannot interpolate accurately between the points of a sparse grid. A Wasserstein metric, on the other hand, measures the minimal distortion needed to transform one structure into another, and so provides a much more natural interpolation along z_{H}, which is robust even with a sharp density and large spacing between training samples. It is worth stressing that the sorted distance metric – which effectively corresponds to the δ density limit of the Wasserstein metric – performs rather poorly, and cannot even reproduce the training points. This is because the mapping between feature space and z_{H} is not exactly linear, changing slope when z_{H} crosses 1Å (because the sorting of the vector changes) and 4Å (because one atom exits the cutoff). The sorted-distances feature space does not have sufficient flexibility to regress this piecewise linear map, as opposed to its smooth Wasserstein counterpart.

Having rationalized the behavior of the Wasserstein metric for a toy model, we can test how it compares to the conventional Euclidean metric on a more realistic data set. We consider in particular the AIRSS *carbon* data set, and compare different levels of density smearing as well as Euclidean and Wasserstein metrics. Figure 2.14 paints a rather nuanced

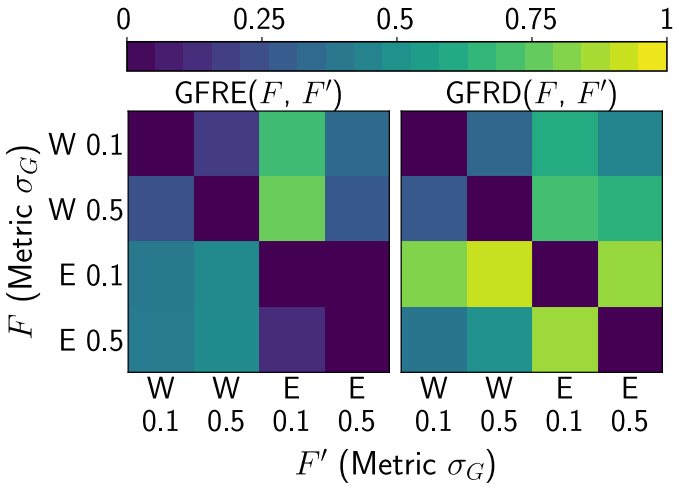

Figure 2.14: Comparison of GFRE and GFRD for the *carbon* dataset, using sharp ($\sigma_G = 0.1$Å) and smooth ($\sigma_G = 0.5$Å) radial SOAP features, as well as Euclidean (E) and Wasserstein (W) metrics.

picture of the relationship between the linear and the Wasserstein-induced feature spaces. The GFRE is non-zero in both directions, meaning that (in a linear sense) Wassertein and Euclidean features provide complementary types of information. Smearing of the density has a small effect on the Wasserstein metric, so that both GFRE($W(\sigma_G = 0.1\text{Å}), W(\sigma_G = 0.5\text{Å})$) and GFRD($W(\sigma_G = 0.1\text{Å}), W(\sigma_G = 0.5\text{Å})$) are small, whereas for Euclidean features – as observed in Section 2.2.1 – changing σ_G induces small information loss, but a large distortion of feature space. Overall, there is no sign of the pathological behavior seen in Figure 2.13, which is an indication that (at least for 2-body features) the *carbon* dataset is sufficiently dense, and that the better interpolative behavior of the EMD does not lead to a more informative feature space.

2.3 Conclusion

Applications of machine learning to atomistic modelling suggest that the featurization that is chosen to represent a molecule or material can be equally or more important than the choice of regression scheme [75]. This has led to the proliferation of approaches to build descriptors, that often differ from each other only in implementation details. The framework we introduce in this work enables a comparison of alternative choices of representations that does not depend on the target property, and makes it possible to determine objectively which of two features contains more information – based on a feature-space reconstruction error – and how much distortion is present in the way they describe the information that is common between the pair – based on a measure of feature-space distortion. Even though the framework is linear in nature, it can be generalized to account for non-linear relationships between feature spaces, either by using kernel-induced features, or by decomposing the feature comparison problem into a collection of local mappings.

Using this framework we demonstrate that the choice of basis set can affect substantially the convergence of SOAP features, and that for instance Gaussian type orbitals are more stable in the limit of small density smearing than the discrete variable representation basis. In practice the convergence of the representation with the number of basis functions should

be considered together with the computational cost of the basis. The analytical expression for GTOs involve special functions that are usually harder to evaluate than those that appear for a DVR basis. This overhead, however, is not sufficient to compensate for the reduced information content, and can be avoided altogether by using a spline approximation for the special functions. In general, computational cost depends substantially on the details of the implementation, and should be assessed in an end-to-end manner as a function of the specific use case. We also show quantitatively that a systematic orthogonal basis is much more effective in describing the atom density than the heuristic symmetry functions of the Behler-Parrinello kind – notwithstanding the considerable success that the latter approach has had in the construction of neural-network-based interatomic potentials [89].

A more systematic difference between atomistic machine-learning frameworks arises from the choice of the order of inter-atomic correlations that underlies the representation. We show that atom density correlation features of high body order make it possible to approximately reconstruct low-body order features, while the opposite is not true. Even when using a non-linear (or locally-linear) mapping, reconstructing 3-body features from 2-body information is virtually impossible. The 3-to-4-body mapping is more subtle: an overall reconstruction based on a linear model is not possible, but a local mapping works well, provided that the structures are far from the manifold of structures for which the 3-body description is not injective. The associated transformation, however, is highly non-linear, and a kernel model that can reconstruct 4-body features shows poor transferability outside of the training set, which hints at similar shortcomings whenever one wanted to use it to learn a property that depends strongly on 4-body correlations. Even though an overall linear reconstruction is not possible, the ν-to-$(\nu+1)$-body mapping error decreases with increasing ν, indicating that less information is added with higher body-orders. This is consistent with the satisfactory results that have been obtained in the regression of atomistic properties using only 3-body information [76, 81], even though the fundamental incompleteness of 3-body features has been shown to have implications for the asymptotic learning performance [53]. An analysis based on the GFRE might help determine the high-order correlations that provide the highest amount of information, and combined with an iterative scheme to evaluate the corresponding features [33] provide a strategy to increase model accuracy with an affordable computational effort.

We also investigate the effect of changing the metric used to compare features, by juxtaposing the Euclidean distance (that is induced by a linear description of the feature space) with a Wasserstein metric, that can be applied to the comparison of n-body correlation features when expressed as real-space distributions. We find that – with an appropriate normalization – the Wasserstein distance can be seen as a proxy of the minimal amount of distortion needed to transform an environment into another, and that this behavior induces smooth interpolation between sparse reference points, contrary to what is observed for the Euclidean distance. However, both an aggressive smearing of the atom density, and the use of a more realistic data set cure the pathological behavior of the linear featurization, so that the Wasserstein metric should not be regarded as superior to the Euclidean one, but complementary. Generalizing the Wasserstein metric to higher body-order correlations, which induce a higher-dimensional feature space that is more likely to be sparsely populated, would be an interesting further research direction.

It further is not clear how the change to the Wasserstein metric generalizes to ACDC functions

of higher order. Existing approaches only employ sorted, one-dimensional projections of higher orders [28]. These approaches can be linked to one-dimensional projections of higher-order ACDC functions in the same way as the sorted distances descriptor is linked to the ACDC function of order 1. For the coherent higher-order space, however, it is not clear what form the induced feature map has. In this case the metric is not negative definite, thus the same substitution kernel used to retrieve the sorted distances descriptor is not positive definite and therefore does not induce a RKHS. While the positive definiteness can be relaxed by extending the concept of RKHS to reproducing kernel Krein space (RKKS), there is no bijection between the indefinite kernel and the RKKS [90] which means there is no unique feature map that can be deduced as a representation from this approach.

The analysis in this work can be extended to compare atom-density representation with a broader class of descriptors based on topological [91, 92, 93] or physicochemical information [94, 95, 96] as well as property-dependent representations induced by neural network frameworks [97, 98]. Even more broadly, an objective measure of the relative effectiveness of features can guide the development of machine learning schemes for any problem that depends strongly on the strategy used to obtain a mathematical description of its inputs. The feature space reconstruction error and distortion can be incorporated into any machine learning frameworks to drive feature selection algorithms [73, 96, 99] or to ensure that implementation choices that improve computational efficiency do not cause a degradation in the resolving power of the resulting features.

3 Symmetry-adapted data-driven basis optimization

Several algorithmic recipes for the construction of basis have been proposed [38, 39, 42, 82] that aim at achieving computational efficiency, and/or at being best adapted to the specific requirement of a given fitting problem, typically the construction of a machine learning model of the potential energy. We bring these considerations to their logical conclusion, by showing that a data-driven basis to expand the atom density, that is optimal in terms of the information content for a given number of functions, can be built as a contraction of a larger primitive basis set, similarly to what is routinely done in quantum chemistry for Gaussian type orbitals (GTOs) [100], and that it can be practically, and inexpensively, evaluated as a numerical basis with striking similarities to ideas in electronic-structure methods [101]. Using an effective basis reduces the number of features that are needed to encode the same information, and thereby reduces the training and prediction time of the resulting machine learning (ML) models. We demonstrate the accuracy, and the computational efficiency, of this approach for both the construction of machine learning potentials for materials, and for the prediction of molecular properties.

3.1 Unsupervised optimization

Principal component analysis has been proposed to compute the data-driven contractions of equivariant features that represent in the most informative way the variability of a dataset as part of the N-body iterative contraction of equivariant (NICE) frameworks [33].

We propose to apply this procedure on the density coefficients as a mean to determine a data-driven radial basis. Keeping different chemical species separate, this amounts to computing the rotationally invariant covariance matrix

$$C^{al}_{nn'} = \frac{1}{N}\sum_i \sum_m c^i_{nlm} c^i_{n'lm} \tag{3.1}$$

where the summation over m results from the Haar integral over the rotation group and can be derived the same way as the order 2 ACDC function in Equation 1.15. For each (a,l) channel, one diagonalizes $\mathbf{C}^{al} = \mathbf{U}^{al}\mathbf{\Lambda}^{al}(\mathbf{U}^{al})^T$, and computes the optimal coefficients

$$\sum_n U^{al}_{qn} c^i_{anlm} = c^i_{aqlm}. \tag{3.2}$$

Note that we compute $\mathbf{C}^{al}$ without centering the density coefficients. For $l > 0$, the mean ought to be zero by symmetry (although it might not be for a finite dataset), and even for the totally symmetric, $l = 0$ terms, density correlation features are usually computed in a way that is more consistent with the use of non-centered features.

The number of contracted numerical coefficients $q_{\max}$ can be chosen inspecting the eigenvalues Λ_q^{al}. At first, it might appear that in order to evaluate the contracted basis one has to compute the full set of $n_{\max}$ coefficients, and this is how the idea was applied in Ref. 33. When combining Equation (3.2) with Equation (1.23), however, one sees that the contracted coefficients can be evaluated directly

$$c^i_{aqlm} = \sum_{j \in A_i} \delta_{aa_j} c^{ij}_{ql} c^{ij}_{lm}. \tag{3.3}$$

using the contracted radial integrals

$$c^{ij}_{ql} = \sum_n U^{a_j l}_{qn} c^{ij}_{nl} \tag{3.4}$$

that can be computed over r, approximated with cubic splines in the range $[0, r_c]$, and then evaluated at exactly the same cost as for a spline approximation of the radial integrals of a primitive basis of size $q_{\max}$. The exact mathematical form and implementation details of the splines can be found in Section 4.1. Splining does not affect the invariant behavior of the atom-density features, and introduces minute discrepancies relative to the analytical basis that do not affect the quality of the resulting models. Thus, the procedure we propose entails the following steps:

1. Compute the density coefficients (1.23) for a representative dataset, using *any* primitive basis, and a large $n_{\max}$
2. Compute the covariance (3.1) and diagonalize it, finding the contraction coefficients U^{al}_{qn}
3. Evaluate the contracted radial integrals using Equation (3.4), over a dense radial grid
4. Use a spline approximation to evaluate directly the radial integrals (3.3) for the first $q_{\max}$ optimal features, and use the coefficients in subsequent ML steps

Even though this framework only needs the contracted radial expansion coefficients (1.23), one can also compute and inspect the "optimal radial basis" that corresponds to the optimized coefficients

$$R_{aql}(r) \equiv \sum_n U^{al}_{qn} c_{anl} R_{anl}(r). \tag{3.5}$$

For a given dataset, these functions are optimal in the sense that when truncated to $q_{\max} < n_{\max}$, they describe the greatest fraction of the variance for the local atom-density coefficients, and unique in the sense that they are independent on the choice of the primitive basis, in the limit in which the latter is complete, as demonstrated in Section 3.2.

For a given dataset, these functions are optimal in the sense that when truncated to $q_{\max} < n_{\max}$, they describe the greatest fraction of the variance for the local atom-density coefficients,

and unique in the sense that they are independent on the choice of the primitive basis, in the limit in which the latter is complete, as demonstrated in Section 3.2.

3.1.1 Mixed-species basis

Even though Equation (3.1) is defined separately for different species a, it is also possible to compute cross-correlations between different elemental channels, defining

$$C^{l}_{nn'} = \frac{1}{N}\sum_{i}\sum_{m} c^{i}_{anlm} c^{i}_{a'n'lm} \tag{3.6}$$

as done in the NICE framework[33] following ideas proposed in Ref. 17, resulting in coefficients that combine information on multiple species

$$c^{i}_{qlm} = \sum_{n} U^{l}_{q;an} c^{i}_{anlm}, \tag{3.7}$$

similar in spirit to the alchemical contraction discussed in Ref. 24. It is worth noting that although the NICE code[102] contains the infrastructure to compute these contractions as a post-processing of the primitive basis, the implementation we propose in `librascal`[70] computes the contracted coefficients directly. However, it only implements the less information-efficient separate (a,n)-PCA strategy. An implementation that evaluates directly the combined contraction would incur an overhead because every neighbor would contribute to every q channel irrespective of their species:

$$\begin{aligned} c^{i}_{qlm} &= \sum_{an} U^{l}_{q;an} c^{i}_{anlm} && (3.8)\\ &= \sum_{an} U^{l}_{q;an} \sum_{j\in A_i} \delta_{aa_j} c^{ij}_{nlm} && (3.9)\\ &= \sum_{j\in A_i}\sum_{n} U^{l}_{q;a_j n} c^{ij}_{nlm}. && (3.10) \end{aligned}$$

Given however that the cost of evaluating the density coefficients is usually a small part of the calculation of density-correlation features[39, 49], we expect that this approach should be in general preferable compared to the calculation of a large primitive basis, and to a two-step procedure in which element-wise optimal functions are further contracted into mixed-element coefficients.

3.1.2 Supervised basis set optimization

For a given number of radial functions, and a target data set, the data-driven contracted basis (3.2) provides the most efficient description of the atom-centred density in terms of the fraction of the retained variance. The most effective variance-preserving compression however does not guarantee that the features are the most effective to predict a given target property. In fact, it has already been shown that SOAP features tend to emphasize correlations between atoms that are far from the atomic center, which can lead to a counter-intuitive degradation of the model accuracy with increasing cutoff radius[24, 74]. This effect can be contrasted by introducing a radial scaling[23, 24] that de-emphasizes the magnitude of the atom density in the region far from the central atom. By applying this scaling – or other analogous tweaks[49] – to the atom density before it is expanded in the primitive basis, one

ensures that the optimal basis is also built with a similar focus on the structural features that contribute more strongly to the target property. In other terms, the information-optimal basis set we introduce here can be combined with a heuristic or data-driven optimization of the underlying density representation, to reflect the scale and resolution of the target property.

Another possibility is to extend the scheme to incorporate a supervised target y_i in the selection of the optimal basis using principal covariates regression (PCovR) [54, 103]. PCovR is a simple linear scheme that can be tuned to provide a projection of features to a low-dimensional latent space that combines an optimal variance compression target with that of providing an accurate linear approximation of the desired target property. Since $l > 0$ contributions of the features have zero mean, the optimization problem can be combined with a supervised component only for $l = 0$, and yields an optimal basis

$$c_{q_\gamma 00} = \sum_n U^{a0}_{q_\gamma n} c_{n00}. \tag{3.11}$$

which is a special case of Equation (3.5) for $l = 0$, where $U^{a0}_{q_\gamma n}$ is obtained as the orthogonalized PCovR projector, as discussed in Refs. 54, 103, using a mixing parameter γ, that determines how strong the emphasis of the optimization should be on minimizing the residual variance or the error in regressing the target.

3.1.3 Multispectrum

We discuss the general case of "multispectra" in the frame of the N-body iterative construction of equivariant (NICE) features[33], but analogous considerations apply to similar many-body descriptors such as the atomic cluster expansion (ACE)[42, 82] or the moment tensor potential (MTP)[38], and is likely to be relevant also for covariant neural networks[104, 105]. The NICE iteration increases the body order of features that describe correlations between ν neighbors by combining lower order features as described in Equation (1.18).

3.2 Results on silicon and QM9

To illustrate the construction and use of an optimal radial basis we present examples for two very different problems: the construction of a general-purpose potential for silicon, based on the training dataset from Ref. 6, and the prediction of atomization energies for the organic molecules from the QM9 dataset [106]. These two examples are complementary: the silicon potential involves a single chemical species, uses forces for training and aims to predict the properties of arbitrary distorted configurations. The QM9 energy model involves multiple elements, but only minimum-energy structures, and, despite its limitations, has been widely used as a benchmark of new representations for molecular machine learning[75].

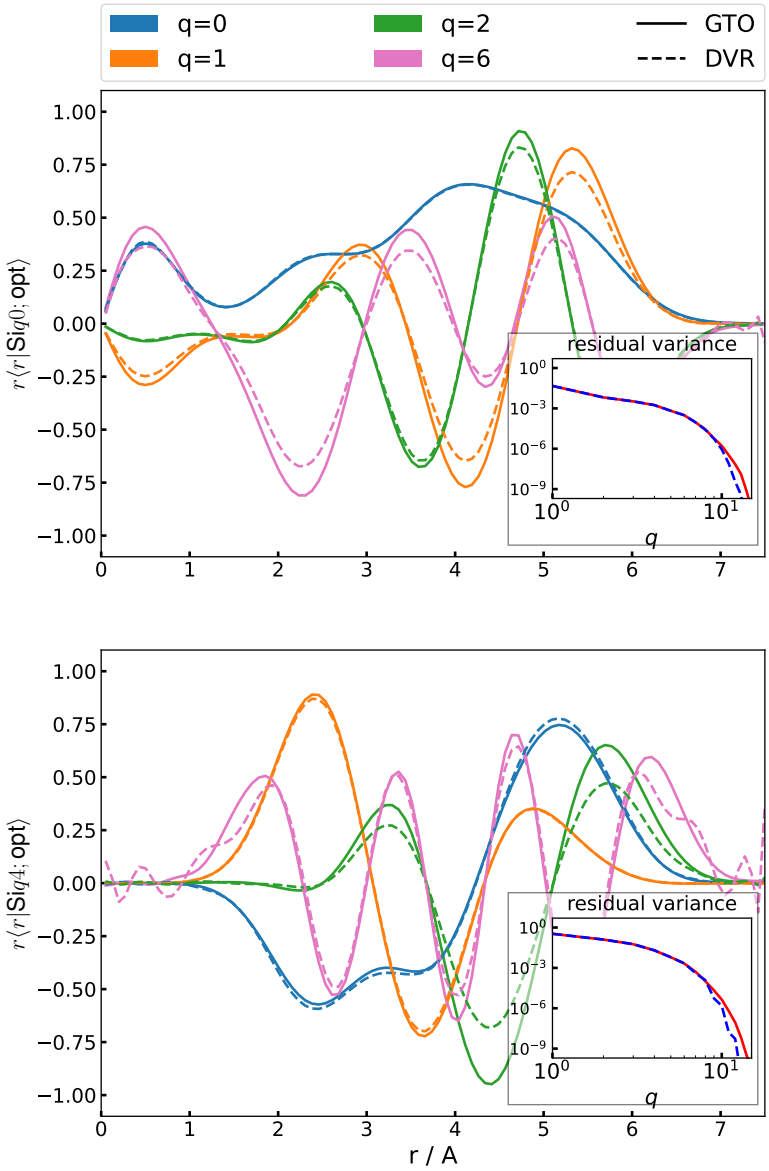

Figure 3.1: Several examples of the optimized radial basis functions on the silicon dataset for $l = 0$ and $l = 4$ using DVR and GTO as primitive basis contracted from $n_{\max} = 20$, with $r_{\text{cut}} = 6$.

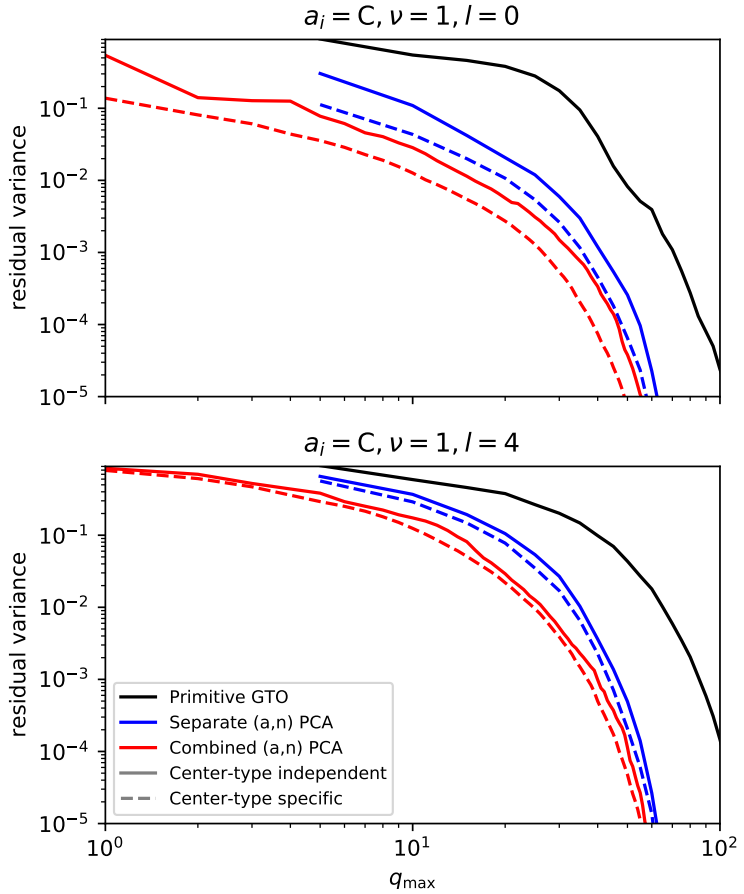

Figure 3.2: Convergence of the residual variance for the expansion coefficients of the density as a function of the number radial basis functions q_{max}, computed for the QM9 dataset and for environments centered on a C atom. The different series correspond to a GTO basis of increasing size (black), to an optimal basis computed for each neighbor density by separating (blue) or by mixing chemical and radial channels (a, n) (red). Full lines use the same basis irrespective of the species of the central atom, dashed lines correspond to a basis optimized specifically for C-centered environments.

3.2.1 Convergence of the density expansion

We begin by considering the convergence of the density expansion, by considering a large primitive basis and then increasing $q_{\max}$ monitoring the residual variance

$$RV = 1 - \frac{\sum_i \sum_{q=1}^{q_{\max}} \left| c^i_{qlm} \right|^2}{\sum_i \sum_{n=1}^{n_{\max}} \left| c^i_{nlm} \right|^2} \tag{3.12}$$

that measures the amount of information lost relative to that contained in the large-$n_{\max}$ primitive basis description. For the Si dataset, the residual variance decays rapidly with increasing number of optimal basis functions, as shown in Figure 3.1. The figure also shows the shape of the optimal radial functions, and demonstrate that the same radial functions can be obtained starting from either of the DVR or GTO bases implemented in `librascal`: the discrepancy increases for higher indices q, but can be reduced by increasing the size of the primitive basis, at no cost during the evaluation of the optimal splined basis. Furthermore, the optimal functions reflect some "sensible" expectations – highly oscillating functions are associated with low covariance eigenvalues, the functions decay at the cutoff distance , and higher angular momentum functions are peaked at larger distances, consistent with the greater variability in the angular distribution at large r.

In the multi-species case, exemplified by the QM9 dataset, there are several possible choices for the contraction strategy. First, one can compute a different contraction depending on the species of the central atom (center-type specfic), or use the same basis functions independent of a_i (center-type independent). Second, one can contract separately the density contribution from each neighbor type along the radial index, or compute a covariance matrix that combines the (a,n) indices. Figure 3.2 shows the convergence of the explained variance for the four possible cases, compared to the baseline of a primitive GTO basis of increasing size - which shows by far the slowest convergence of the explained variance, requiring almost 100 radial channels ($n_{\max} = 20$, for the 5 species present) to reduce the importance of features below 10^{-4}. The same level can be achieved with $q_{\max} \sim 50$ when performing separate PCAs for each neighbor species, and $q_{\max} \sim 30$ when computing jointly the correlations between radial and elemental channels. Performing a separate PCA depending on the species of the central atom accelerates slightly the convergence of the explained variance.

3.2.2 Convergence of density correlations features

We now turn to considering how the truncation of the density expansion basis affects the evaluation of higher-order features, focusing in particular on the invariant components. We begin analyzing the convergence of the power spectrum computed for the Si dataset. We take the SOAP features computed with a large $n_{\max} = 20$ as the "full" description of three-body correlation, and compute the global feature space reconstruction error[107] (GFRE) that measures how accurately the full feature space can be reconstructed using SOAP features that are built from a truncated density expansion. Given that SOAP features are usually subselected using a low-rank matrix approximation (CUR) approach[73] and farthest point sampling (FPS)[108, 109], we also investigate the interplay between the density expansion optimization, and this further feature reduction step.

Using an optimal density expansion basis systematically improves the GFRE compared to

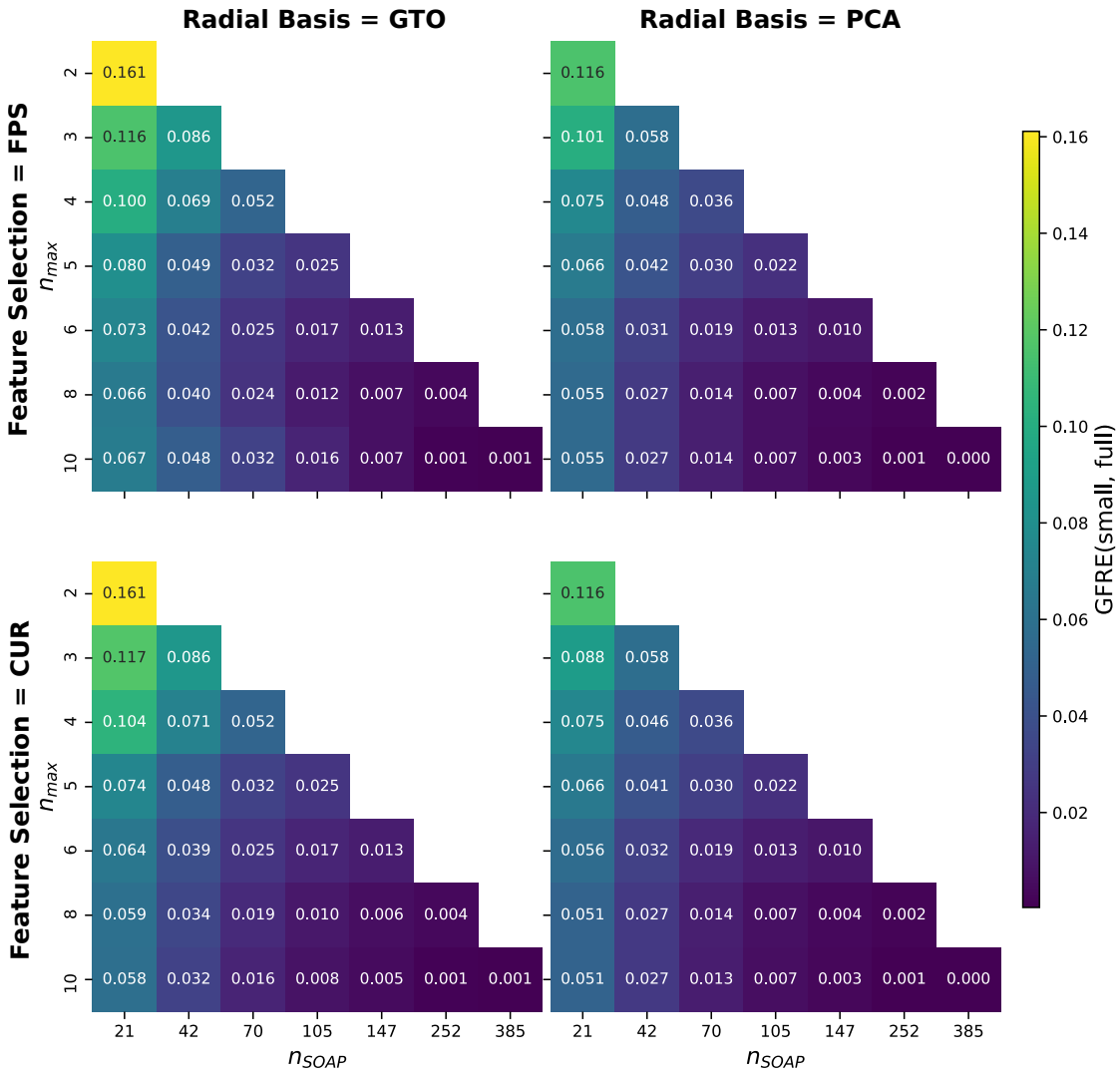

Figure 3.3: Feature space reconstruction errors for the power spectrum, resulting from the truncation of the radial basis and from the selection of a subset of the power spectrum entries using a deterministic CUR scheme and FPS. The "full" feature space is approximated with the power spectrum features, computed using a GTO basis with ($n_{\mathrm{max}} = 20, l_{\mathrm{max}} = 6$), and we compare the convergence obtained by using a smaller GTO basis against a truncated optimal basis of the same size.

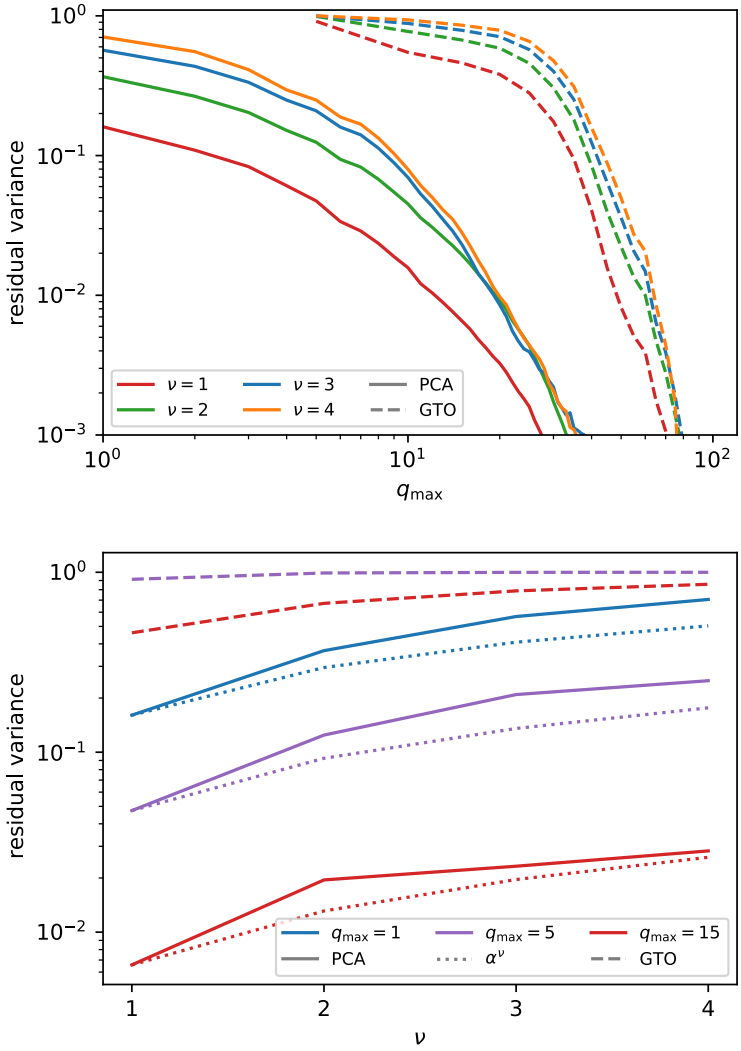

Figure 3.4: Residual variance for the multispectra computed for the QM9 dataset. For each body order, the baseline variance is taken to be that associated with the NICE features built starting from a "full" vector of density coefficients ($n_{\text{max}} = 20, l_{\text{max}} = 5$) – summing over the contributions from all atoms in a representative sample of the QM9 dataset. We compare results for a small GTO basis (dashed lines) against those for an optimal basis (full lines) determined using a separate PCA procedure depending on the chemical nature of the central atom, and using a combined (a, n) covariance. (top) Different colors correspond to order-ν multispectra. $\nu = 1$ and $\nu = 2$ terms are computed in full; for the $\nu > 2$ terms the NICE contraction has been converged so that the discarded variance at each iteration is smaller than that due to the truncation of the density coefficients. (bottom) Comparison of the residual variance for fixed radial/chemical basis size and different orders of multispectrum. Dotted lines indicate the behavior one would expect if the retained variance followed exactly a multiplicative behavior.

a GTO basis of the same size (Figure 3.3). This is true both for the full-sized SOAP vector, and for a subselection of the invariant power spectrum entries based on a deterministic CUR algorithm, as well as on FPS. This suggests that using an optimal radial basis as the building block of higher-order spectra yields feature vectors that can be easily compressed further, which is important to reduce the cost of evaluating SOAP based models. The cost of different parts of the feature evaluation (density expansion, invariant calculation, kernel evaluation, gradients ...) depends subtly on the composition of the system and the various convergence parameters [39]. When evaluating a Gaussian process regression model, the calculation of the invariant features and of the kernel values is often dominant, and so the possibility of aggressively subselecting SOAP features with little performance loss is as important as the reduction in the number of radial basis size.

The same efficient compression is observed for the QM9 dataset, when extending the construction to higher-order features and to a multi-component system. Despite the fact that, as discussed in Section 3.1.3, there is no formal guarantee that the optimal density coefficients are also optimal to build high-ν equivariants, we find in practice that the PCA basis leads to much faster convergence of the bispectrum and the trispectrum compared to the primitive basis (Figure 3.4, top panel). The truncation of the density coefficients affects the multispectra in a way that is qualitatively similar to a multiplicative behavior : the impact of an incomplete description of the density gets amplified by taking successive orders of correlations (Figure 3.4, bottom panel). Given that the raw number of multispectrum components grows exponentially as q_{max}^{ν}, the density basis truncation has a dramatic effect in reducing the size of the multispectrum vector. This observation may be extremely important in the construction of systematic high-body order expansions such as NICE or ACE, and in particular in the extension of these approaches to multiple chemical species. The very efficient feature reduction that can be achieved by combining (a, n) channels at the density level shall make it much easier to avoid the exponential increase of complexity of high-body order models with growing chemical diversity.

3.2.3 Regression models

The accuracy of a Gaussian approximation potential based on SOAP features, trained using both energy and forces , seen in Figure 3.5 shows an improvement of the cross-validation error for the most aggressive truncation of the feature space (up to $n_{\text{max}} \approx 6$ for forces, and $n_{\text{max}} \approx 4$ for energy), but no improvements for large n_{max}. For the largest feature set the primitive GTO basis can be up to 10% more accurate than the corresponding optimal-basis model. A comparison with Figure 3.3, that shows that the PCA basis is objectively more informative than the primitive basis, suggests that an effect similar to the degradation of performance with increasing environment cutoff radius might be at play here: for this dataset size, the GTO basis, which becomes smoother for large distances, is better suited to build a potential with limited amounts of training data. The fact that the GTO basis may be fortuitously better adapted to this specific regression problem is also suggested by the non-monotonic convergence of the error. Depending on the value of n_{max}, the GTO functions are distributed so as to span the $[0, r_c]$ range . Particularly for small n_{max}, the varying positions of maxima and nodes of the orthogonalized GTOs emphasize different portions of the atomic environment, and can produce such a non-monotonic trend, particularly in the limit of a relatively small train set size. The PCA basis, on the other hand, is constructed to provide a progressively more complete

description of the atom density for the specific training set, resulting in a more regular, mostly monotonic convergence.

These effects can be investigated more easily by considering a 2-body model, that uses only the radial coefficients c_n^i ($l = 0$). The comparison between the GTO and the DVR basis (the former being vastly superior in terms of linearly decodable mutual information content, as seen from the GFRE in the bottom panel of Figure 3.6) is far from clear-cut, with GTOs giving the worst results for forces with $n_{\max} = 4, 6$. The optimal PCA basis is usually comparable with - but not substantially better than - the best result between GTO and DVRs, for each size of the basis. The relative performance of different basis sets is similar when using a linear model and a polynomial kernel, although the nonlinear model reaches an accuracy that is approximately 6 times better for energies and two times better for forces. We extend the optimal basis to a PCovR optimization ($\gamma = 0.1$) with the energies as supervised component to determine the contraction coefficients of the basis: as shown in Figure 3.6 (top, center), this PCovR optimal basis yields much better accuracies in the small $q_{\max}$ range. In fact, by taking the "pure regression", $\gamma \to 0$ limit of PCovR, one would obtain a basis that, for a linear model, yields an accuracy comparable to a fully-converged 2-body potential even with $q_{\max} = 1$. This is because the coefficients are built so that a linear regression performed for the $q_{\max}$-dimensional features would match as well as possible the predictions of a linear model based on the full primitive basis

$$c_{0_\gamma} \underset{\gamma \to 0}{=} \sum_n w_n c_n^i \approx y_i. \tag{3.13}$$

Thanks to the spline approximation of the optimal basis, c_{0_γ} with $\gamma \to 0$ can be computed at the cost of a single radial function evaluation, much as it would be the case for a pair potential. The use of a nonlinear model based on the same radial spectrum features provides the simplest test of transferability for the PCovR-optimized basis beyond ridge regression. Even though for very small $q_{\max}$ there is a noticeable improvement (up to a factor of 2 for the force RMSE and $q_{\max} = 2$) against primitive and PCA-optimized bases, the advantage is quickly lost for larger bases, where the variance reduction plays the leading role in driving the selection of radial basis even for small α. As shown in Figure 3.6 (bottom), the improved regression accuracy of PCovR-optimized basis functions comes at a necessary cost in terms of reconstruction error - even though with an intermediate value of the mixing parameter they achieve higher information content than either of the primitive bases, as measured by the GFRE.

The advantages of using an optimized radial basis become much clearer for the QM9 dataset. As shown in Figure 3.7, there is a dramatic improvement of performance at all body orders when using a PCA-contracted (a, n) basis, with the improvement becoming more and more substantial for higher ν. For the bispectrum features with $q_{\max} = 5$ (effectively only one channel per species), the use of a combined basis leads to a 5-fold reduction of the test error compared to the primitive GTO basis, and makes it possible to reach the symbolic threshold of 1 kcal/mol MAE. In other terms, an optimal PCA contraction achieves an accuracy comparable to a primitive GTO basis which is roughly 2 times larger. Given that the number of bispectrum ($\nu = 3$) features scales as $q_{\max}^3$, this translates into an order of magnitude improvement in computational efficiency for the QM9 predictions. For larger basis sets, and for $\nu > 3$, it becomes necessary to truncate the construction of the multispectra, which within the current implementation of the NICE framework is achieved with further PCA contractions applied at each iteration. In order to be able to use a consistent PCA threshold up to the full primitive GTO basis (which contains $n_{\max} = 20$ radial terms per chemical species) we need to use a

rather aggressive truncation, which results in clear performance loss, as evidenced by the saturation of the model accuracy with increasing $q_{\max}$.

The interplay of the truncation of the density coefficients, the thresholding heuristic, and the use of the features in a linear or a nonlinear model, is evident in the lower panel of Figure 3.7. The plot compares the NICE models computed with $q_{\max} = 50$ and an aggressive truncation of the body-order iteration, with the more balanced settings from Ref. 33 ($n_{\max} = 12$, $l_{\max} = 7$, $\nu_{\max} = 5$, 1000 invariant features per body order), with a "large NICE" model which includes 53880 features (up to $\nu = 4$, built upon a relatively small spherical expansion with $l_{\max} = 5$ and $n_{\max} = 5$) and with a kernel ridge regression (KRR) model that uses the same parameters as in Ref. 24 (i.e. using only the power spectrum and a nonlinear kernel). The details of the NICE construction affect substantially the stability and the accuracy of the model in the high-n_{train} limit, that vary by a factor of two. Furthermore, a nonlinear model based on low-body order features is the most accurate, and reaching a MAE of 0.12kcal/mol with $n_{\text{train}} = 10^5$. Even though a thorough investigation of these aspects is beyond the scope of the present work, the understanding of the interplay between the truncation of the density basis and the information loss at higher body order that we discuss here shall support more systematic studies in the future.

3.3 Conclusion

The realisation that most of the widely adopted representations for machine learning of atomistic properties can be seen as a discretization of interatomic correlations naturally points to the importance of determining the most expressive and concise basis to expand the atom density. For a given dataset it is possible to uniquely define a basis that is optimal in terms of its ability to linearly compress the information encoded in the variance of the density coefficients, which can be determined as a contraction of any complete primitive basis, and evaluated efficiently by approximating it with splines.

We have explored with numerical experiments the implications of this choice to evaluate higher-order correlations of the density, and to build linear and nonlinear regression models of the energy for both condensed-phase silicon and small organic molecules. Our study indicates that the optimization of the density basis has a dramatic impact on the information content of higher-order features, but that achieving the ultimate accuracy also requires tuning the basis to reflect the sensitivity of the target property to changes in the atomic configurations. A more intuitive approach may be to perform this tuning at the level of the atomic density, e.g. modulating the amplitude and resolution of atomic contributions depending on the distance from the central atom. An "unsupervised" optimal basis would then provide the most concise, and systematically-convergent, discretization of this tuned atomic density.

Another possible strategy involves the use of supervised criteria in the construction of the basis, as we have demonstrated applying PCovR to the construction of an optimal $\nu = 1$ basis. A systematic investigation of the effect of varying the parameters of PCovR, as well as the use of PCov-style feature selection[110] in the construction of the multi-spectra, is a promising direction for further research. One of the challenges is that it is only meaningful to apply the linear reasoning that underlie PCovR to optimize features with the same equivariant properties as the targets, and so the $l > 0$ channels of the density coefficients cannot be optimized with a

straightforward application of this scheme. One approach to addressing this challenge has been to transition from a closed-form optimization to gradient descent one. The optimization of the decomposition matrix is achieved by propagating the gradients from the prediction loss associated with higher-order invariant features back to the radial expansion coefficients. While this approach has proven effective for optimizing the chemical decomposition [50], simultaneously optimizing the chemical and radial channels introduces numerical instabilities in the decomposition matrix optimization, which is an issue that remains to be addressed.

The performance gains associated with the use of an optimal basis are much clearer in the presence of multiple chemical elements, in particular when using a combined basis in which radial channels associated with different species are considered together in the construction of the symmetry-adapted feature covariance matrix. This combined basis can capture the same amount of information of a primitive basis that is 3 to 5 times larger, and is essential to the efficient construction of high-order density correlations features, given that we show analytically how the loss of information that is due to a truncated basis becomes worse with increasing ν. It shall help accelerate the convergence of the schemes, such as NICE, ACE, MTP, that rely on very high body order terms. We show that linear NICE models built on high-order combinations of the optimal basis yield much lower error than those constructed on a GTO basis of similar size, even though the truncation of the body order iteration, or introducing nonlinearities, can also affect, positively or negatively, convergence.

The determination of the optimal basis is much less demanding than the fitting of even the simplest models. After fitting, the evaluation of the contracted basis involves no overhead over a primitive basis of equal size, thanks to the use of a spline approximation. Given that it provides consistently higher information content, and that it results in models that have comparable (for silicon) or much better (for QM9) accuracy than standard choices of orthogonal bases, we recommend adopting this scheme in any machine learning approach that requires representing an atomic density – particularly for systems that involve many chemical species, or for frameworks that rely on the evaluation of high-order density correlations.

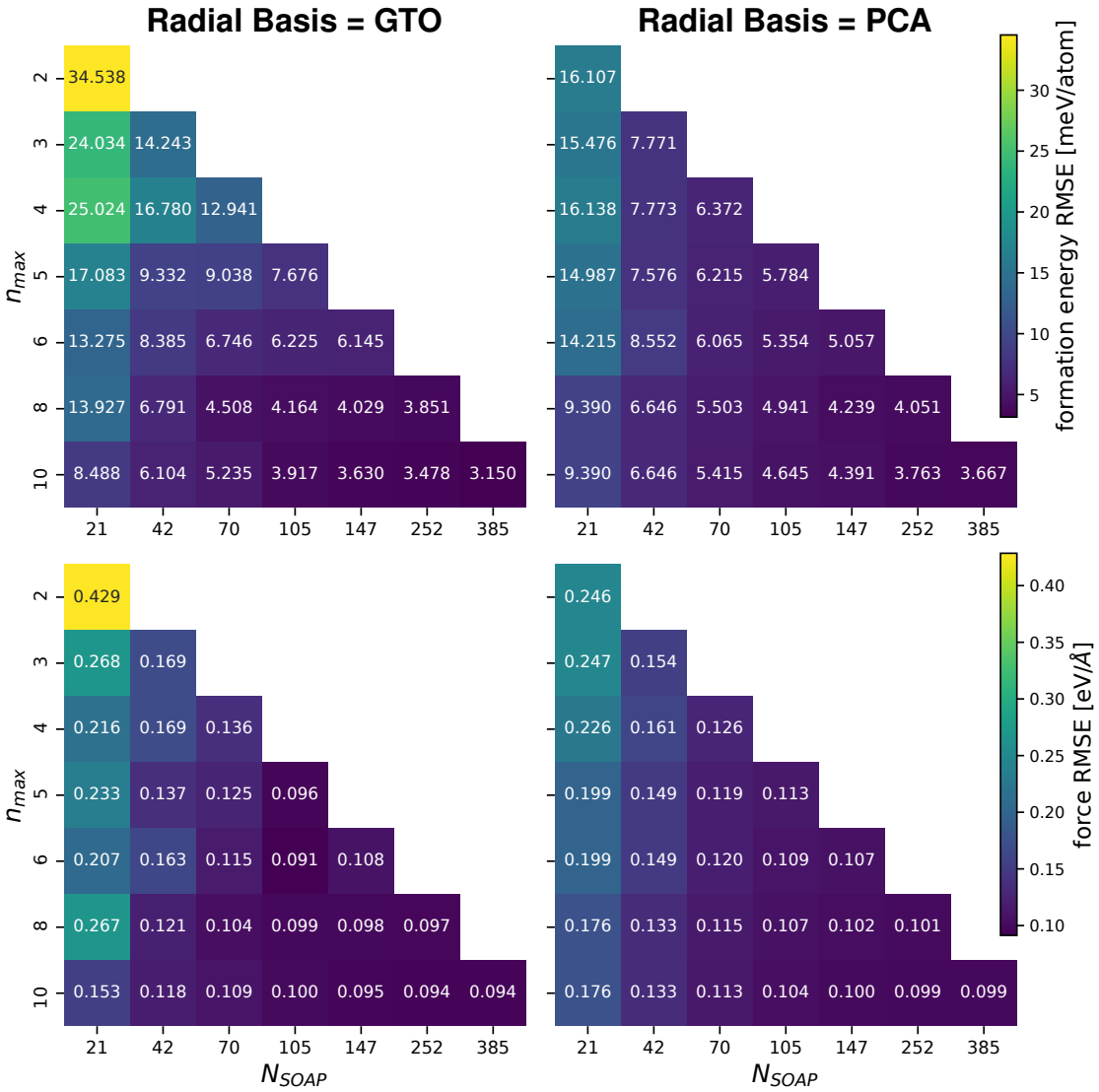

Figure 3.5: Energy and force RMSE for a Gaussian approximation potential based on the power spectrum, fitted to the Si dataset, plotted as a function of the number of radial functions $n_{\max}(q_{\max})$ and sparsification of the SOAP features, n_{SOAP} (using CUR selection).

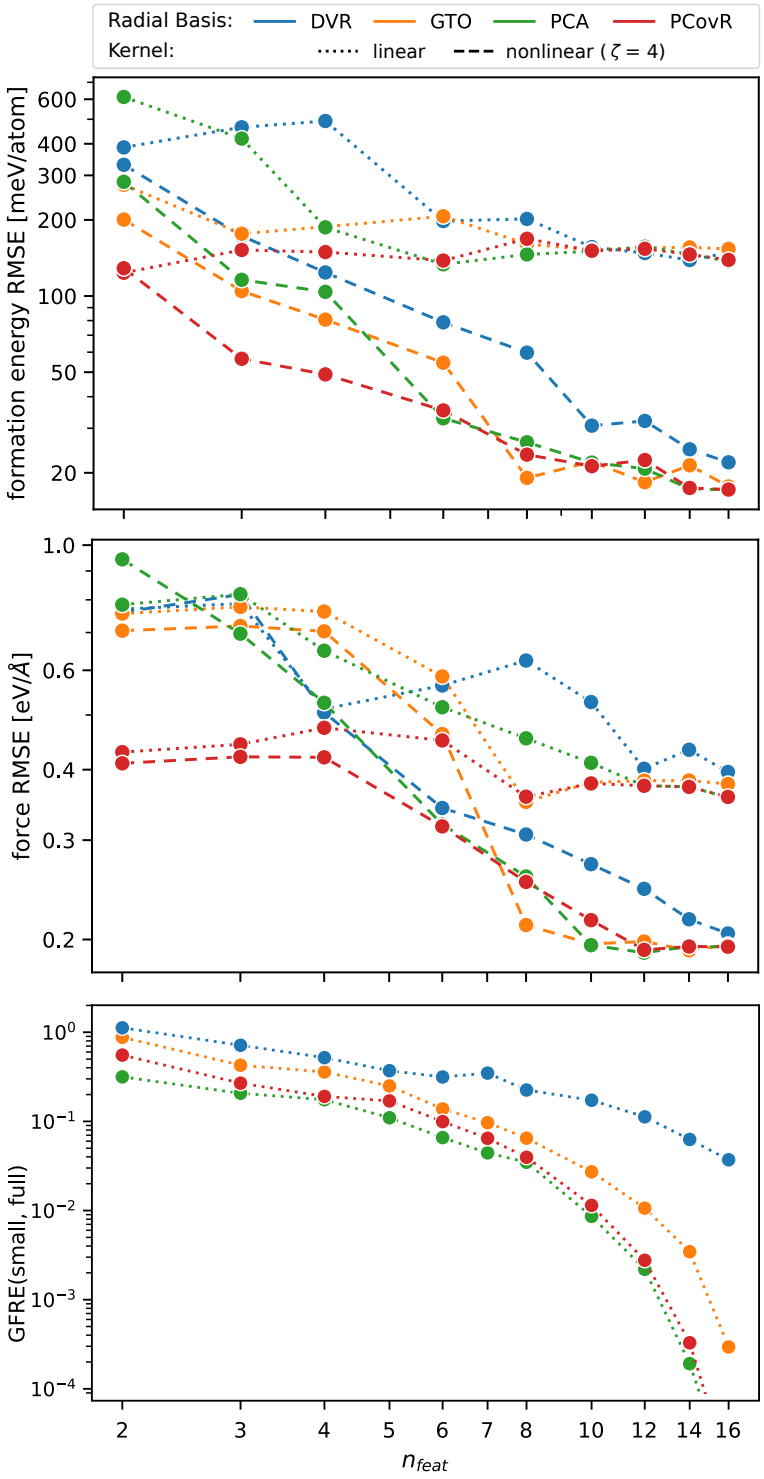

Figure 3.6: Energy (top) and force (center) 5-fold cross-validation RMSE and GFRE (bottom), computed on the silicon dataset for models based on the radial spectrum $|\overline{\rho_i^{\otimes 1}}\rangle$, as a function of the number of radial functions. Different curves correspond to a primitive DVR and GTO basis, and to the optimal (PCA and PCovR) contracted bases. The PCovR contraction is performed with $\gamma = 0.1$. Full lines correspond to a linear model, and dashed lines to a polynomial kernel with exponent $\zeta = 4$. The GFRE is computed relative to a $n_{\max} = 20$ GTO basis.

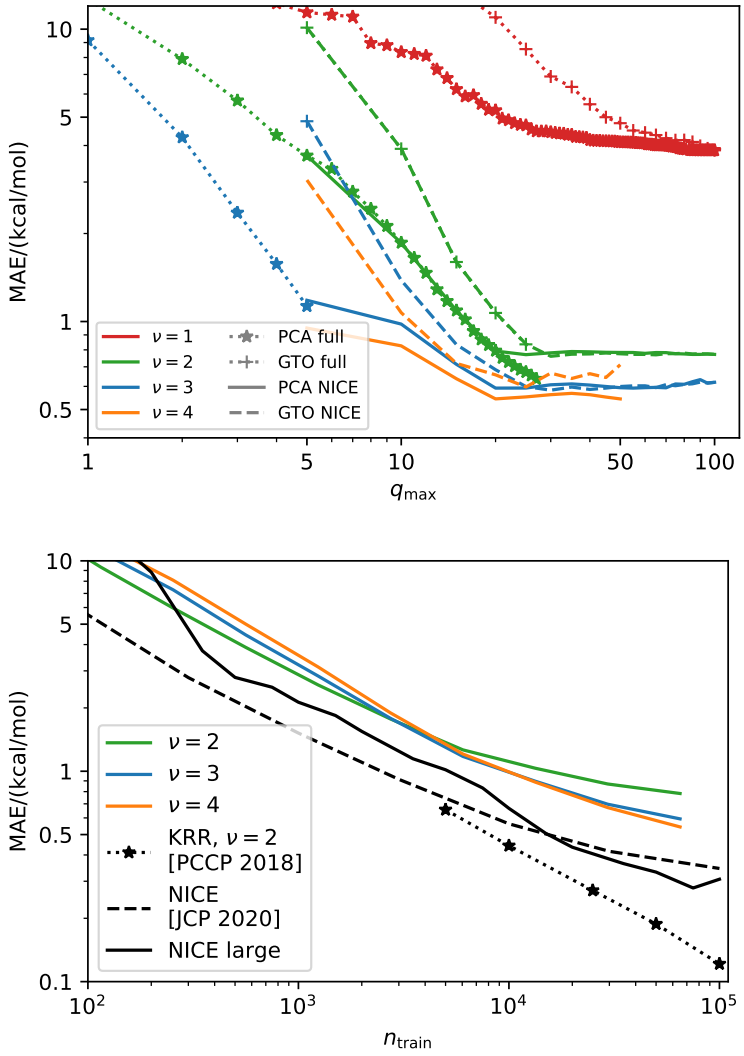

Figure 3.7: Convergence of ML models of the atomization energy of molecules from the QM9 dataset. (top) Convergence as a function of the (a, n) radial basis size, comparing a primitive GTO basis and an optimal PCA contraction, for different body orders of the features. For large q_{max} it is necessary to truncate aggressively the NICE iteration, which results in a plateau of the accuracy with large q_{max}. All curves are trained and tested on a set of 65'000 structures, up to the largest q_{max} which could fit into 1TB of memory. (bottom) Learning curves are obtained with linear models built on the PCA optimal features of increasing body order. All coloured curves are computed with $q_{max} = 50$, and the same truncation parameters as in the top panel. For comparison, we show a selection of bespoke models, with black lines: a large NICE model (full line) using 53390 features; the NICE model from Ref. 33 (dashed line); a kernel model based on the power spectrum, using parameters analogous to those in Ref. 24 (dotted line).

4 Implementation of short-range machine learning interatomic potentials

Interatomic potentials have been used for a long time to approximate the potential energy surface in classical molecular dynamics (MD) simulation by decomposing the systems' total energy into body-order contributions. Classical molecular dynamics can be performed by solving the classical equations of motion

$$H(\mathbf{p},\mathbf{q}) = \frac{\mathbf{p}^2}{2m} + V(\mathbf{q}), \quad \frac{\partial \mathbf{p}}{\partial t} = \frac{\partial V(\mathbf{q})}{\partial \mathbf{q}}, \quad \frac{\partial \mathbf{q}}{\partial t} = \frac{\mathbf{p}}{m} \quad \text{(classical Hamiltonian mechanics).} \tag{4.1a}$$

For an interatomic potential $V(\mathbf{q})$ takes the form

$$V(\mathbf{q}) = \sum_{i=1} V_1(\mathbf{r}_i) + \sum_{i,j=1} V_2(\mathbf{r}_i,\mathbf{r}_j) + \sum_{i,j,k=1} V_3(\mathbf{r}_i,\mathbf{r}_j,\mathbf{r}_k) + \dots \quad \text{(interatomic potential).} \tag{4.2a}$$

Due to limitation in computational power, the development of accurate interatomic potentials historically relied on a meticulous hand-tuning of parametric models called *empirical interatomic potential*. They were constrained by the chemical and phase spaces they could predict accurately [111, 112]. However, with the increased computational power available today, the shift towards data-driven models emerged, allowing a more automated construction for fitting models on data generated from empirical potentials [113] or ab initio calculations [114, 115, 116]. Remarkably, the current generation of machine learning interatomic potentials (MLIPs) has reached the capability to extensively cover chemical [50] and phase space [6] accurately.

In this chapter, we discuss my contributions to the software ecosystem that facilitate the deployment of MLIPs into a MD software. These contributions include my work on the `librascal` package [70], instrumental for constructing MLIPs based on the SOAP featurization of atomic structures, and its interface to `LAMMPS` [117], an advanced classical MD code, enabling studies on ferroelectric phase transitions in barium titanate [118], and on the transport properties of lithium ortho-thiophosphate [119]. To ensure that a code base remains manageable, strategic integration with existing software solutions is essential, wherever appropriate. This requires not only a solid understanding of relevant mathematical methods and software packages but also proficiency in software design patterns. This combination of requirements poses significant challenges on the development of enduring and robust MLIP packages. In the final section we discuss how we can encounter these challenges by a standardized abstract data type allowing a more flexible construction of MLIPs and by that

addressing the need for more efficient workflows [8].

4.1 Implementation of cubic splines for featurization

One of main contributions to `librascal` has been the implementation of a cubic spline to interpolate the radial expansion coefficients as expressed in Equation (1.23). For each coefficient nl, the one-dimensional function $f^{nl} : [0, r_c] \to \mathbb{R}$ is splined. The function f^{nl} maps the distance $r \in [0, r_c]$ to the neighbor contribution of the radial expansion coefficient as in Equation (1.23). For the construction of the cubic spline the targeted interval $[0, r_c]$ is further partitioned into a set of subintervals $[r_1, r_2], \ldots, [r_K, r_{K+1}]$ where $0 = r_1 < r_2 < \ldots < r_K < r_{K+1} = r_c$. Then cubic splines are polynomials of degree 3 $p_k(r) = A_k + B_k r + C_k r^2 + D_k r^3$ on the interval $[0, 1]$ with the boundary conditions

$$p_k^{nl}(0) = f^{nl}(r_k) \text{ and } p_k^{nl}(1) = f^{nl}(r_{k+1}) \text{ for } k = 1, \ldots, K, \tag{4.3a}$$

$$\left(\frac{\partial p_k^{nl}}{\partial r}\right)_{r=1} = \left(\frac{\partial p_{k+1}^{nl}}{\partial r}\right)_{r=0} \text{ for } k = 1, \ldots, K-1, \tag{4.3b}$$

$$\left(\frac{\partial^2 p_k^{nl}}{\partial r^2}\right)_{r=1} = \left(\frac{\partial^2 p_{k+1}^{nl}}{\partial^2 r}\right)_{r=0} \text{ for } k = 1, \ldots, K-1, \tag{4.3c}$$

$$\left(\frac{\partial^2 p_1^{nl}}{\partial r^2}\right)_{r=0} = 0 \text{ and } \left(\frac{\partial^2 p_K^{nl}}{\partial r^2}\right)_{r=1} = 0 \quad \text{(natural boundary conditions).} \tag{4.3d}$$

These $4K$ boundary conditions can be rearranged to a tridiagonal linear system solving for the $4K$ unknowns [120]

$$\begin{pmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 4 & 1 & \\ & & & 1 & 2 & \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_{K-1} \\ B_K \end{pmatrix} = \begin{pmatrix} 3(f^{nl}(r_2) - f^{nl}(r_1)) \\ 3(f^{nl}(r_3) - f^{nl}(r_1)) \\ \vdots \\ 3(f^{nl}(r_K) - f^{nl}(r_{K-2})) \\ 3(f^{nl}(r_K) - f^{nl}(r_{K-1})) \end{pmatrix}. \tag{4.4a}$$

This system can be solved in linear time with time complexity $O(2K)$ by iterating two times through the matrix following a Gaussian elimination scheme. Note that the exact form of the polynomial p_k can be chosen arbitrary and the same procedure can be used to solve for the $4K$ coefficients. For the implementation in `librascal`, we followed closely the implementation of Ref. [121] that uses Lagrange polynomials of degree 3 in their derivation. The grid points $\{r_k\}_{k=0}^{K+1}$ are incrementally adjusted until the function approximation error falls beneath a user-specified error threshold. The approximation error can be estimated by sampling points in the intervals (r_k, r_{k+1}) and computing the difference between the function f^{nl} and the spline p^{nl}. During the evaluation of the implementation, we found that conditioning on both the relative and absolute error yields the most robust results in terms of interpolation accuracy and convergence of the grid size. While algorithms for adaptive grids can be advantageous to control the grid size and thus reduce memory requirements, for atomistic applications a uniform grid does not reach a memory-intensive regime while having a minimal impact on the prediction accuracy. We analyzed the effect of the splining accuracy on the prediction accuracy more rigorously on NMR chemical shieldings of hydrogen environments (the dataset can be found in Ref. [122]). We trained the model for different splining accuracies and compared

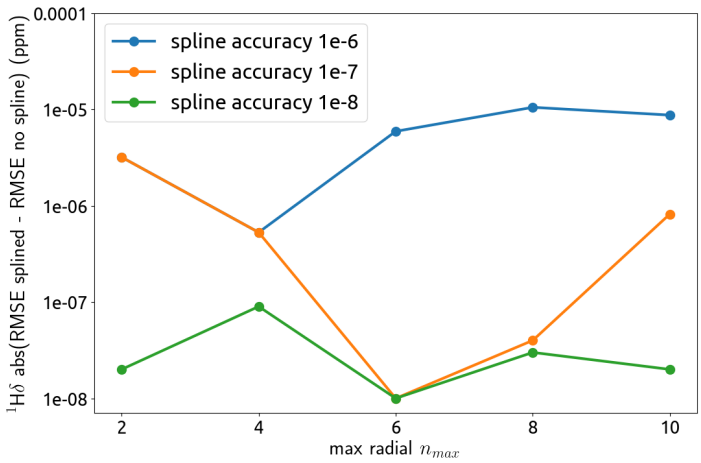

Figure 4.1: The relationship of the spline accuracy to the difference of the prediction error compared to using no spline for a linear model using SOAP descriptors with $l_{\max} = 9$. NMR chemical shieldings of hydrogen environments were used as target property using the datasets and same train-test setup as in Ref. [122]. We trained the model for different splining accuracies and compared the difference to a model without splining. The effect of the spline on the difference is several orders of magnitude below the DFT accuracy in the order of 10^{-1}, while for all accuracies the grid consists of less than 2048 points.

the difference to a model without splining. The results presented in Figure 4.1 show that the change in the prediction error is several order of magnitude below the DFT accuracy, which is estimated to be in the order of 10^{-1}, while for all accuracies the grid consists of fewer than 2048 points. Compared to an adaptive grid, the uniform grid reduces the asymptotic complexity of determining the subinterval for evaluation from a logarithmic $O(\log K)$ binary tree search to a constant $O(1)$ lookup. Moreover, as the grid is constructed for each nl channel, the setup and the evaluation of the spline opens the possibility of parallelizing the spline evaluation over all channels. Viable forms of parallelization encompass multithreading, the use of single instruction multiple data (SIMD) instructions, or leveraging graphics processing unit (GPU) acceleration.

4.2 Interfacing with molecular dynamics packages

Molecular dynamics (MD) packages, such as LAMMPS, GROMACS, CP2K, and i-PI [117, 123, 124, 125], commonly separate the time evolution into separate modules: The time integrator to propagate equation of motions, the thermo- and barostats and the calculation and the potential energy and its forces. As the computation of the potential energy only depends the atomic positions and species, and the dynamics require only the energy and forces, is therefore a well-suited point for separating it from the rest of the code base. One significant benefit of this approach is to avoid reimplementing established time integrators, thermo- and barostats. Although their implementation may appear straightforward, developing a robust code base that transparently handles edge cases for the non-expert user is nevertheless time-consuming task, demanding extensive documentation. Furthermore, well-established MD software packages, such as `GROMACS` [126] and `LAMMPS` [117], offer a variety of parallelization strategies that significantly improve the speed of interatomic potentials. These include a message-passing

interface (MPI) for the domain decomposition, CUDA- and OpenMP-based multithreading as well as SIMD abstraction modules for hardware-adaptive compute kernels. The embedding of hardware dependent parallelization strategies, such as customized CUDA and compute kernels, necessitates adapting the interatomic pontential code to these specific kernel routines. In contrast, MPI-based domain decomposition only requires dividing the potential into contributions of atomic-environments within a cutoff, as outlined in Equation (1.4). To provide the MPI parallelization to external developped short-range potentials MD codes further calculates the neighbor list on which the short-range potential must be computed. The neighbor list in MD engines is a list of atom centers and neighbors within a cutoff, similar as the atomic environment has been defined in Equation (1.3a). For short-range potentials the force acting on atom k can thus be evaluated by

$$\mathbf{F}_k = -\frac{\partial E_A}{\partial \mathbf{r}_k} = -\sum_{i\in A}\frac{\partial E_i}{\partial \mathbf{r}_k} = -\sum_{i\in A}\sum_{j\in A_i}\frac{\partial E_i}{\partial \mathbf{r}_{ji}}\frac{\partial \mathbf{r}_{ji}}{\partial \mathbf{r}_k}, \quad \text{where } \frac{\partial \mathbf{r}_{ji}}{\partial \mathbf{r}_k} = \begin{cases} 1, & k=i \\ -1, & k=j \\ 0, & \text{else,} \end{cases} \tag{4.5}$$

where A represents a structure and A_i the atomic environment around atom i. It is evident from Equation (4.5) that the forces acting on atom k solely depend on the derivatives of its and its neighboring atomic environments. Consequently, these forces can be parallelized by dividing the cells into separate domains that only include a subset of all atoms and its local environments. Each domain can then be assigned to an independent hardware component allowing an embarrassingly parallel evaluation of the forces. Communication between the domains becomes only necessary if one wants to avoid a recomputations of the partial forces $\partial E_i/\partial \mathbf{r}_{ij} = -\partial E_i/\partial \mathbf{r}_{ji}$ where atom i and atom j belong to two different domains. As this presents a tradeoff between the number of evaluations of the potential and the number of MPI-communications, MD engines such as `LAMMPS` prodived a tunable parameter that can choose between the two options, see the parameter *newton* in the `LAMMPS` manual [127]. Global communication is only required when aggregating the local energy contributions into the total energy, as explained in the `GROMACS` manual [128] by the *nstcalcenergy* parameter. These communications can be performed by the MD engine without any additional information from the short-range potential, since the MD engine controls the division of the neighbor list into domains. Thus, the implementation of these communications can be abstracted out of the short-range potential code, thereby enabling a modular design that does not need to account for any MPI-communication. In the case of the study on barium titanate [118], the MPI parallelization of `LAMMPS` played a crucial role in investigating the impact of long-range structural correlations on the Curie temperature, due to the necessity of conducting simulations on large cell sizes.

4.3 Implementation of kernel models with forces

The abundance of ML packages available today [129], such as `scikit-learn` [130], `PyTorch` [131], prompts the question of whether these developments are needed. Most of these packages, however, are focused on applications in generic data analysis and computer vision and are therefore not suited to the specific needs of MLIPs, which include domain-specific featurization of 3D structures and gradient inference, as the calculation of forces requires. In this section we discuss the implementation of positions gradients for kernel based MLIPs, the

ones used in our recent study on barium titanate [118]. For a comprehensive introduction to kernel models please refer to Ref. [132]. We focus in this section on the extension to kernel models with position gradients in the context of atomistic learning. For a kernel k fitted on the samples $\{\mathbf{c}_t \in \mathbb{R}^M\}_{t=1}^N$ and targets $\{y_t \in \mathbb{R}\}_{t=1}^N$ the optimal weights $\boldsymbol{\alpha} \in \mathbb{R}^N$ are retrieved as solution of the minimization problem

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}' \in \mathbb{R}^N}{\operatorname{argmin}}\, \ell(\boldsymbol{\alpha}') \tag{4.6a}$$

$$\ell(\boldsymbol{\alpha}) = \sum_n^N \| \sum_{t=1}^N \alpha_t k(\mathbf{c}_t, \mathbf{c}_n) - y_n \|^2 \tag{4.6b}$$

where the $\ell(\boldsymbol{\alpha})$ is the loss function. It can be expressed in vectorial form as

$$\ell(\boldsymbol{\alpha}) = \|\mathbf{K}\boldsymbol{\alpha} - \mathbf{y}\|^2. \tag{4.7a}$$

This problem can be easily solved by setting the derivative of $\ell(\boldsymbol{\alpha})$ to zero

$$0 = \frac{\partial \ell(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \tag{4.8a}$$

$$0 = 2\mathbf{K}^T\mathbf{K}\boldsymbol{\alpha} - 2\mathbf{K}\mathbf{y} \tag{4.8b}$$

$$\mathbf{K}\mathbf{y} = \mathbf{K}^T\mathbf{K}\boldsymbol{\alpha} \tag{4.8c}$$

$$\mathbf{K}\mathbf{y} = \mathbf{K}\mathbf{K}\boldsymbol{\alpha}, \text{ since } \mathbf{K} = \mathbf{K}^T \tag{4.8d}$$

which gives $\mathbf{K}^{-1}\mathbf{y}$ as solution for $\boldsymbol{\alpha}$. The solution can subsequently be used to evaluate the target property at an arbitrary point i by the relationship

$$\sum_{t=1}^N \alpha_t k(\mathbf{c}_t, \mathbf{c}_i) = y_i. \tag{4.9}$$

Since the energy is a property of a structure, given a kernel k acting on the featurization of two environments, the structural kernel becomes

$$\sum_{t=1}^N \alpha_t \sum_{i \in A} \sum_{t_{i'} \in A^{(t)}} k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i) = \sum_{i \in A} E_i = E, \tag{4.10}$$

where we use the notation $A^{(t)}$ to index structure t.

4.3.1 Training with forces

We include the gradients wrt. the atomic position $\mathbf{r}_k$ of atom k into the picture. Note that the training points used to construct the kernel are independent from the points for which the gradients are evaluated. This is important so the gradient operator only acts on the target structure and not on the training structures. We get an expression for the negative forces by taking the derivative of the energy in Equation (4.10) wrt. the position of atom k in structure A

$$\frac{\partial E}{\partial \mathbf{r}_k} = \sum_{t=1}^N \alpha_t \sum_{t_{i'} \in A^{(t)}} \sum_{i \in A} \frac{\partial k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i)}{\partial \mathbf{r}_k}, \quad \mathbf{r}_k \in A \tag{4.11}$$

Similar as to the expression of the forces in Equation (4.5) we can evaluate the derivative of the kernel

$$\frac{\partial k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i)}{\partial \mathbf{r}_k} = \sum_{j \in A^{(t)}_{t_{i'}}} \frac{\partial k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i)}{\partial \mathbf{c}_{t_{i'}}} \frac{\partial \mathbf{c}_{t_{i'}}}{\partial \mathbf{r}_{j t_{i'}}} \frac{\partial \mathbf{r}_{j t_{i'}}}{\partial \mathbf{r}_k}, \quad \mathbf{r}_k \in A. \tag{4.12}$$

To account for the use of atomic forces as additional property we can extend the loss function by adding one more term

$$\ell(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \| \sum_{t=1}^{N} \alpha_t \sum_{t_{i'} \in A^{(t)}} \sum_{i \in A^{(n)}} k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i) - E^{(n)} \|^2 + \sum_{n=1}^{N} \sum_{k \in A^{(n)}} \| \sum_{t=1}^{N} \alpha_t \sum_{t_{i'} \in A^{(t)}} \sum_{i \in A^{(n)}} \frac{\partial k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i)}{\partial \mathbf{r}_k} - \frac{\partial E^{(n)}}{\partial \mathbf{r}_k} \|^2. \tag{4.13}$$

To reformulate this loss into a vectorial form we define the following quantities

$$[\mathbf{F}]_{(n,k,p)} = \frac{\partial E^{(n)}}{\partial r_k^{(p)}} \tag{4.14a}$$

$$[\mathbf{K}_{NN}]_{n,t} = \sum_{t_{i'} \in A^{(t)}} \sum_{i \in A^{(n)}} k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i), \tag{4.14b}$$

$$[\mathbf{K}_{N_\partial N}]_{(n,k,p),t} = \sum_{t_{i'} \in A^{(t)}} \sum_{i \in A^{(n)}} \frac{\partial k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i)}{\partial r_k^{(p)}}, \tag{4.14c}$$

where (n, k, p) is the flattened multi-index specifying the structure, the atom and the Cartesian dimension and N_∂ denotes the number of these quantities over all structures. More formally defined as

$$N_\partial = 3 \sum_{n=1}^{N} |A^{(n)}|, \tag{4.15}$$

where we use $|A^{(n)}|$ to describe the number of atoms in the structure $A^{(n)}$. Then we can reformulate the loss again into vectorial form

$$\ell(\boldsymbol{\alpha}) = \|\mathbf{K}_{N+N_\partial,N}\boldsymbol{\alpha} - \mathbf{y}\|^2, \tag{4.16a}$$

$$\text{with} \quad \mathbf{K}_{N+N_\partial,N} = \begin{bmatrix} \mathbf{K}_{N,N} \\ \mathbf{K}_{N_\partial,N} \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix}, \tag{4.16b}$$

$$\mathbf{K}_{N,N} \in \mathbb{R}^{N,N}, \qquad \mathbf{K}_{N_\partial,N} \in \mathbb{R}^{N_\partial,N}, \mathbf{E} \in \mathbb{R}^N, \qquad \mathbf{F} \in \mathbb{R}^{N_\partial} \tag{4.16c}$$

where $\mathbf{F}$ are the stacked gradients and $\mathbf{K}_{N_\partial N}$ is the matrix from the stacked sum-product of the feature and kernel gradients as defined in the loss 4.9, explicitely written as follows

By including the gradients in the loss function, the memory requirements for the kernel matrix increase dramatically, as N_∂ grows with the total number of atoms over all structures while N is the number of structures. A fruitful approach is therefore to perform a low-rank approximation of the kernel matrix by projecting the N_∂ points onto a fixed number T of representative points, defined as *pseudo points*. For the well-established sparse *Gaussian approximation potential* (GAP) [84] the *subset of regressor* (SoR) [133, 134, 135] method is used. The SoR method derives the kernel weights from the low-rank approximation of a positive-definite kernel matrix. We can extend $\mathbf{K}_{N_\partial+N,N}$ to a positive-definite matrix by taking the kernel gradients as additional

training basis functions into account, thereby deriving the following expression for the energy

$$\sum_{t=1}^{N} \alpha_t \sum_{i \in A} \sum_{t_{i'} \in A'^{(t)}} k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i) + \sum_{t=1}^{N} \sum_{k \in A'^{(t)}} \beta_{tk} \sum_{t_{i'} \in A'^{(t)}} \sum_{i \in A} \frac{\partial k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i)}{\partial \mathbf{r}_k} = E. \tag{4.17}$$

Redoing the same derivation with this energy prediction brings a matrix of the form $\mathbf{K}_{N_\partial+N, N_\partial+N}$, explicitely specified in Ref. [136]. The low-rank approximation of this full kernel has the form

$$\tilde{\mathbf{K}} = \mathbf{K}_{N_\partial+N,T} \mathbf{K}_{TT} \mathbf{K}_{T,N_\partial+N} \approx \mathbf{K} \tag{4.18a}$$

where the relationship $\tilde{\mathbf{K}} \approx \mathbf{K}$ is exact in the case where $\mathbf{K}$ has rank T and if T linearly-independent points of $\mathbf{K}$ are chosen as pseudo points [137]. The SoR method allows an efficient solution for the loss corresponding to this approximated kernel

$$\ell(\boldsymbol{\alpha}) = \|\tilde{\mathbf{K}}\boldsymbol{\alpha} - \mathbf{y}\|. \tag{4.19}$$

Instead of solving the kernels weights as $\tilde{\boldsymbol{\alpha}} = \tilde{\mathbf{K}}^{-1}\mathbf{y}$, the SoR allows to reformulate the evaluation for a new structure A

$$\boldsymbol{\alpha}_T = (\mathbf{K}^T_{N_\partial+N,T} \mathbf{K}_{N_\partial+N,T} + \mathbf{K}^T_{T,T})^{-1} \mathbf{K}^T_{N_\partial+N,T} \mathbf{y}, \quad E^A = \mathbf{k}_T \boldsymbol{\alpha}_T, \tag{4.20a}$$

where $\mathbf{k}_T$ is a vector with the kernel entries between the the pseudo points and a structure A evaluated as

$$[\mathbf{k}_T]_t = \sum_{t_{i'} \in A^{(t)}} \sum_{i \in A} k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i), \tag{4.21a}$$

$$[\mathbf{k}_T]_{t_k} = \sum_{t_{i'} \in A^{(t)}} \sum_{i \in A} \frac{\partial k(\mathbf{c}_{t_{i'}}, \mathbf{c}_i)}{\partial r_k^{(p)}}, \tag{4.21b}$$

depending if pseudo point t corresponds to a structures or a gradient in one of the Cartesian directions.

Omitting the computation of the kernel in the complexity analysis and assuming $T < N_\partial + N$, the complexity of the computation of the weights $\boldsymbol{\alpha}_T$ decreases from $O((N_\partial + N)^3)$ to $O(T^3 + T^2(N_\partial + N)) = O(T^2(N_\partial + N))$. Furthermore, the complexity of the evaluation for a new structure A decreases from $O(N_\partial + N)$ to $O(T)$.

In sparse GAP only individual environmental features are chosen as pseudo points neglecting any gradients. It has not been studied yet how much the absence of gradients in the pseudo points affects the learning performance [138]. The evaluation costs are however dramatically reduced, since the kernel gradients are more expensive to evaluate. Additionally, the environmental features among the structures exhibit high similarity, thus choosing a small representative set of environments is a suitable strategy to reduce cost. From the fact that we only use environments as pseudo points we obtain a simplified expression for the inference of new forces from Eqs. (4.11) and (4.12)

$$-\sum_{t=1}^{T} \alpha_t \sum_{i \in A} \sum_{j \in A_i} \frac{\partial k(\mathbf{c}_t, \mathbf{c}_i)}{\partial \mathbf{c}_i} \frac{\partial \mathbf{c}_i}{\partial \mathbf{r}_{ji}} \frac{\partial \mathbf{r}_{ji}}{\partial \mathbf{r}_k} = -\frac{\partial E}{\partial \mathbf{r}}. \tag{4.22}$$

4.3.2 Efficient inference of forces

Starting from the expression in Equation (4.22) we can rearrange the order of the sums to make the calculation more efficient

$$\frac{\partial E}{\partial \mathbf{r}_k} = \sum_{m=1}^{M}\sum_{t=1}^{T}\alpha_t \sum_{i\in A}\sum_{j\in A_i}\frac{\partial k(\mathbf{c}_t, c_{im})}{\partial c_{im}}\frac{\partial c_{im}}{\partial \mathbf{r}_{ji}}\frac{\partial \mathbf{r}_{ji}}{\partial \mathbf{r}_k} \tag{4.23a}$$

$$= \sum_{i\in A}\sum_{j\in A_i}\underline{\sum_{m=1}^{M}\sum_{t=1}^{T}\alpha_t \frac{\partial k(\mathbf{c}_t, c_{im})}{\partial c_{im}}\frac{\partial c_{im}}{\partial \mathbf{r}_{ji}}\frac{\partial \mathbf{r}_{ji}}{\partial \mathbf{r}_k}} \tag{4.23b}$$

$$= \sum_{i\in A}\sum_{j\in A_i}\underline{\sum_{m=1}^{M}\frac{\partial c_{im}}{\partial \mathbf{r}_{ji}}\frac{\partial \mathbf{r}_{ji}}{\partial \mathbf{r}_k}\sum_{t=1}^{T}\alpha_t \frac{\partial k(\mathbf{c}_t, c_{im})}{\partial c_{im}}}. \tag{4.23c}$$

In fact moving the sum over t forward and extracting the feature gradient out of the sum reduces the complexity from $O(MT)$ to $O(M+T)$ of the term underlined in Equation (4.23).

4.4 Metadynamic framework embedding MLIP

To study the paraelectric-ferroelectric phase transitions in barium titanate [118] we used metadynamics [139] to accelerate the sampling of the transitions. In metadynamics the sampling of a transition is accelerated by introducing a bias potential $B(\mathbf{s})$ that acts on a collective variable (CV) $\mathbf{s} \in \mathbb{R}^d$ constructed from the atomic coordinates $S(\mathbf{q}) = \mathbf{s}$ where d is usually much smaller $3|A|$, the number of coordinates of the system. The purpose of building a CV is to use a low-dimensional variable that models the phase transition well such that the action of an external bias potential can enable transitions that would not take place on the time scale of standard MD. The extended potential of metadynamics is thus a function of

$$\tilde{V}(\mathbf{q}) = V(\mathbf{q}) + B(S(\mathbf{q})), S : \mathbb{R}^{3|A|} \to \mathbb{R}^d. \tag{4.24}$$

The free energy surface of a given phase transition can then be retrieved from the dynamics associated with the extended potential up to a constant by the relation

$$F(s) = -B(s) - T\log\left(\tilde{P}(s)\right) + \text{const.}, \tag{4.25a}$$

$$\text{where } \tilde{P}(s) \propto \int_{\mathbb{R}^{3|A|}} \mathrm{d}\mathbf{q}\, \exp\left(-\frac{\tilde{V}(\mathbf{q})}{T}\right)\delta(s - S(\mathbf{q})). \tag{4.25b}$$

In practice the exact derivation of the free energy surface requires an iterative reweighting from the trajectories as the bias term is increased during the dynamics to further drive the phase transition. To take advantage of the MPI parallelization implemented in `LAMMPS` we conceptualized a framework that distributes the computations needed for a metadynamic simulation as schematically shown in Figure 4.2. The potential energy $V(\mathbf{q})$ is computed with `librascal` utilizing `LAMMPS` for the domain decomposition of the neighborlist, the bias potential is computed with the software package `PLUMED` and the dynamics associated with the extended potential $\tilde{V}(\mathbf{q})$ is run with the package `i-PI`. For the study on barium titanate the collective variable is based on the spherical expansion coefficients, as motivated in Ref. [140]. Thus further required an interface between `librascal` and `PLUMED`.

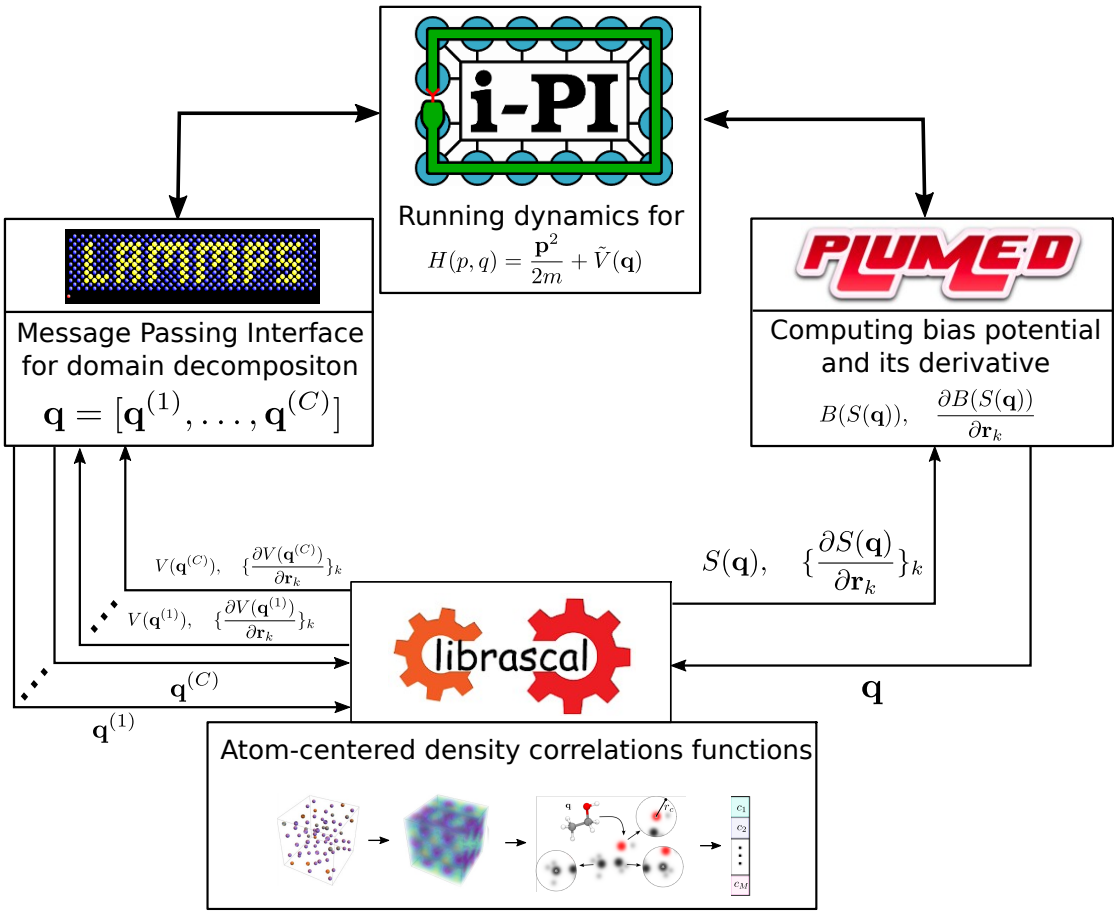

Figure 4.2: A schematic showing the interwork of the software packages to run metadynamic simulations to study paraelectric-ferroelectric phase transitions in barium titanate. The domain decomposition of LAMMPS is noted as $\mathbf{q}$ for the full atomic position vectors and $[\mathbf{q}^{(1)},\ldots,\mathbf{q}^{(C)}]$ for the positions corresponding to the C domains.

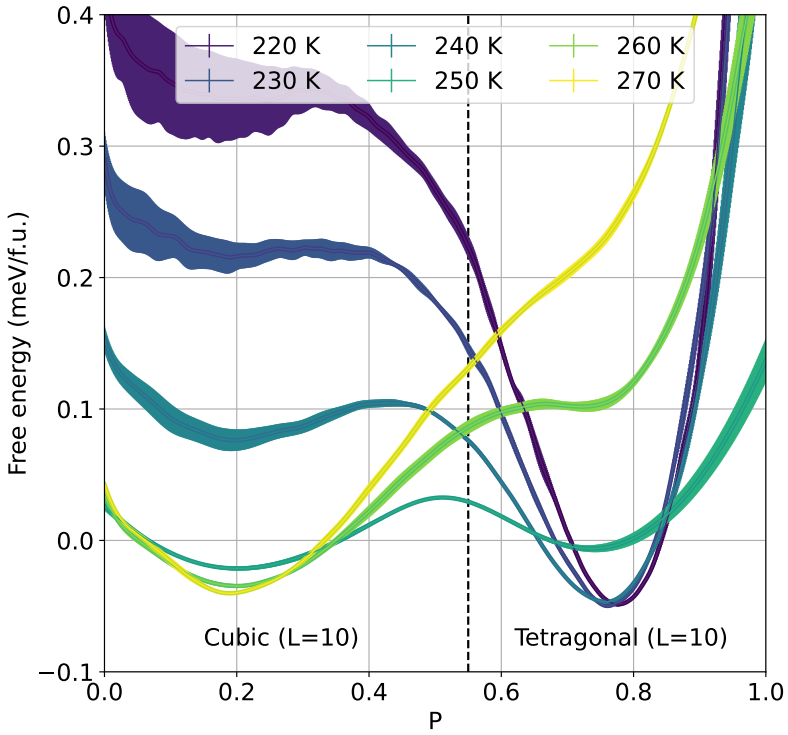

Figure 4.3: Free energy surfaces as a function of the collective variable for temperatures above and below the transition temperature (249 K) for a $10 \times 10 \times 10$ cell. The shaded areas correspond to errors on the free energy, computed as the standard deviation on the mean of the free energy estimates on 4 independent blocks for each metadynamics simulation.

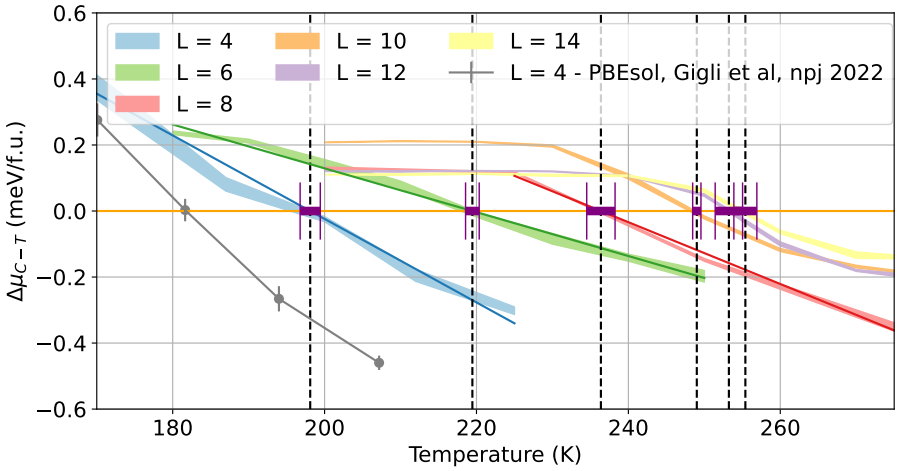

Figure 4.4: Difference between the chemical potential of the cubic and tetragonal phases as a function of the temperature for different simulation box sizes L.

4.4.1 Finite-size convergence of the Curie point in barium titanate

In previous work [140] a transition between the cubic and tetragonal phases was observed in unbiased MD simulations of $4 \times 4 \times 4$ supercells. These transitions occur in unbiased MD because the cell is relatively small. When a larger supercell is employed spontaneous transitions become exceedingly rare and the system remains stuck in the energetic minimum that corresponds to the cubic or tetragonal phase for the duration of the simulation, even outside of the range of stability of that phase. For these larger systems a simulation bias is thus required to drive transitions between the two phases. Figure 4.3 shows that metadynamics simulations using the order parameter based on the Ti-centered spherical expansion coeffients in the notation as in Equation (1.25)

$$P = \sqrt{(\sum_{i \in A} c^i_{O11-1})^2 + (\sum_{i \in A} c^i_{O110})^2 + (\sum_{i \in A} c^i_{O11+1})^2}, \tag{4.26}$$

can be used to drive transitions for $10 \times 10 \times 10$ supercells. This figure shows the free energy surfaces (FES) that emerge from these metadynamics simulations. These free energy surfaces were obtained by reweighting using the iterative trajectory reweighting (ITRE) method [141]. Block averaging was used to estimate the errors on the estimates of the free energy shown in Figure 4.3.

Figure 4.3 clearly shows that there is a minimum for high CV values when the temperature is low and the system is in the tetragonal phase and ferroelectric. This minimum is replaced by a minimum at a low value of the CV when the temperature is high and the system is in the cubic phase and paraelectric. At intermediate temperatures two minima are observed as one phase is metastable.

To extract the difference in chemical potential between the tetragonal and cubic phases we performed clustering using the probabilistic analysis of molecular motifs algorithm (PAMM) [142]. This clustering technique assigns two probabilities $\theta_1(s_i)$ and $\theta_2(s_i)$ to each CV value s_i. $\theta_1(s_i)$ is the likelihood that the corresponding frame is from the cubic basin, while $\theta_2(s_i)$ measures the likelihood that it is from is tetragonal basin. The chemical potential difference per formula unit between the two phases can thus be estimated as:

$$\Delta\mu_{\text{C-T}} = -\frac{k_B T}{L^3} \log\left(\frac{\sum_i \theta_1(s_i) w_i}{\sum_i \theta_2(s_i) w_i}\right)$$

where the sum runs over all the trajectory frames, L^3 is the number of unit cells, T is the temperature and w_i is the weight for each trajectory frame obtained from ITRE.

Figure 4.4 shows how $\Delta\mu_{\text{C-T}}$ changes with temperature for a range of differently-sized supercells. This quantity is initially positive for all cell sizes indicating that the tetragonal phase is more stable than the cubic one at low temperatures. It becomes negative at high temperatures when the relative stabilities of the two phases reverses. The Curie point can be determined from Figure 4.4 by finding the temperature at which $\Delta\mu_{\text{C-T}}$ is zero. In Figure 4.4 these temperatures are indicated by the vertical dashed lines. Errors on these estimates of the Curie temperature are also indicated. To determine these errors we divided each trajectory into four blocks and obtain four separate estimates for each $\Delta\mu_{\text{C-T}}$ value. Variances were computed from these four estimates so the shaded areas in Figure 4.4 indicate the (1-σ) confidence limits. The Curie temperature for each system size was extracted by drawing a line of best fit through

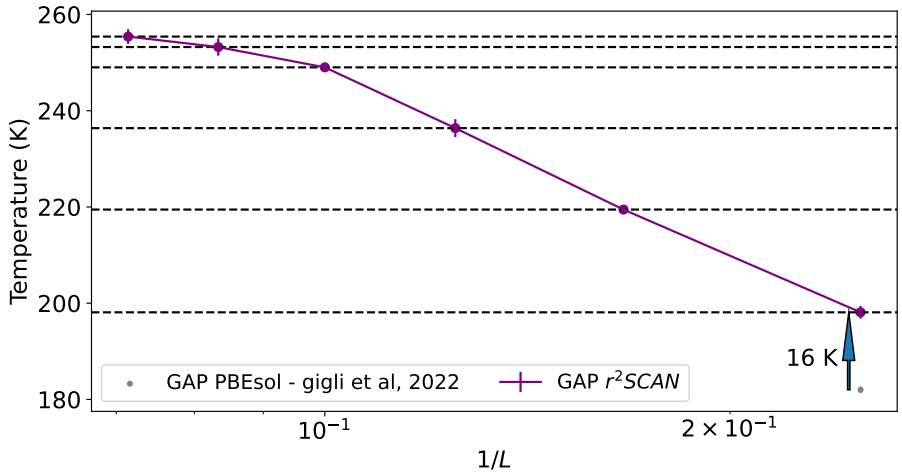

Figure 4.5: Predicted Curie point at a function of the inverse of the box size $1/L$. The grey dot at the bottom right shows the result from [140]. The more accurate functional used in this work shifts the transition temperature upwards by 16 K as indicated by the vertical arrow. However, this upward shift is smaller than the increases in transition temperature that are seen for the larger systems simulated in this work.

the estimates of $\Delta\mu_{\text{C-T}}$. Propagated errors from this fitting then yield an estimate of the error on the transition temperature.

Figure 4.4 clearly shows that the Curie temperature increases with system size. Furthermore, these differences in transition temperature are for the most part statistically significant. Figure 4.5 indicates the size dependence for the transition temperature more clearly. In this figure the transition temperature is shown as a function of the inverse box size. It is only the $14 \times 14 \times 14$ system that has a transition temperature that is compatible with the smaller $12 \times 12 \times 12$ system. For all other system sizes the transition temperature is underestimated. The observation of significant finite-size effects here contradicts the analysis provided in Ref. [140] that relied on extrapolating the dielectric constant in the high-temperature regime with a Curie-Weiss law. Interestingly, this indirect approach underestimates the finite-size effects. If the aim is to extract accurate thermodynamics simulating large system sizes is thus essential.

4.5 Future directions

The development of `librascal` has revealed several shortcomings in the design that required to rethink the way MLIP are developed. In the following we discuss these problems and give a perspective how these will be addressed in future software development.

4.5.1 Modularity of featurization and model construction

A crucial disadvantage in the design of `librascal` was the entanglement of the featurization and the model construction inside one library. This entanglement is caused by the need to efficiently store and compute the forces as indicated in Equation (4.23c). A strict separation of the computation of the kernel and the dot product between the weights and the kernel, as implemented in `scikit-learn` [143], results in a less efficient implementation. An implementation that performs the summations in Equation (4.23c) in the same order requires

the information of the sample dimension (structure, atom, neighbor). In `librascal` for that reason an abstract data type was implemented that makes the features and gradients accesible by these quantities. A domain-agnostic approach as in packages such as `scikit-learn` that cannot retrieve this information from their passed data type, in this case `numpy` arrays [144], is therefore not possible. Furthermore, handling operations that consistently act on features and gradients requires to keep track of the features that correspond to gradients, since multiple gradients can correspond to the same features but derived with respect to different quantities (e.g. Cartesian coordinates and/or neighbor). By incorperating metadata into the data type, these operations can be implemented more generically and thus be reused for multiple model implementations. These problems are addressed by the development of a new abstract data type implemented in the software package `metatensor`. This abstract data type acts as a container of the feature and its gradients with individual labels for each entry in the dimensions. With this kind of abstract data type, `numpy` like data operations that take the structural characteristics of the data into account can be implemented. The model and featurization can be separated into two different modules since the relevant structural information of the features and gradients is kept in the data type. That effort is currently in progress with the development of `rascaline` and `torch_spex`, packages for the featurization of atomic structures, returning them as `metatensor` data containers, and `metatensor-learn` a subpackage for the model construction from `metatensor` data containers.

4.5.2 Serialization of MLIPs

ML models exhibit various trade-offs between accuracy, evaluation time, training cost, and hyperparameter optimization. Consequently, the choice of a suitable model depends on the system of interest, size of the dataset, and available computational resources. In the last decade, the rapid development of machine learning models has produced numerous interfaces for MD engines such as `LAMMPS` [145, 146, 147, 148, 149]. The current software infrastructure, however, causes considerable friction between the ML package and the MD code when switching the model. In fact, not only does the MD software require a recompilation for a different interface and an installation of the new package but also a whole new workflow for training and deployment needs to be typically created. In addition, model development frequently takes place in higher-level languages like `Python` or `Julia`, owing to their flexibility. The choice of the programming language for model development narrows down its deployment to MD packages that provide compatible interfaces for that exact language; otherwise, they require an implementation of the model in one of the supported languages that the MD engine provides. This requires an additional implementation of a process that saves the model to a file that can be loaded back from the low-level language interface. The process of exporting an abstract data type to file is named *serialization*. This additional development work hinders the application of numerous developed ML models in MD packages written low-level programming languages, which are essential for conducting research on large systems or running long-term simulations.

Similar challenges are present in the industry, where models trained with diverse ML packages must be adapted to various hardware architectures and software stacks. This diversity complicates the consistent deployment of the same package version across different devices. To address this, an open intermediate representation for ML models, known as *open neural network exchange* (ONNX), has been developed [150]. However, ONNX is currently not a

suitable candidate to serialize MLIPs, since it does not support the inference of gradients.

The *open knowledgebase of interatomic models* (OpenKIM) [151] attempts to solve this issue by developing abstract representations of data and processing directives necessary for molecular simulations. It unifies several MD packages pair potential interfaces into one common interatomic potential interface, named the KIM API [152]. Although this approach reduces the number of required interfaces, it doesn't comprehensively cover all relevant MD packages. Notably, packages like `GROMACS` and `CP2K` [124] are absent. Most MD packages supported by the KIM API are implemented in higher-level languages, which in the first place do not present a hurdle for MLIP developers to interface with. However, this solution does not allow a full development of MLIPs in a high-level language, as it still necessitates the creation of an interface with the KIM API using a low-level language such as C, C++, or Fortran.

It is due to the shortcomings of the existing solutions that we pursue a software ecosystem that leverages TorchScript as a serialization format. The utilization of TorchScript not only supports the inference of gradients but also allows model utilization in C++, overcoming the high-level to low-level language barrier. Additionally, it offers advanced model optimization utilities, such as kernel fusioning, to enhance complex models' efficiency. Considering these capabilities, TorchScript appears to be a promising candidate to standardize the landscape of MLIPs.

Conclusion

We presented a set of measures that can be used to analyze different design choices of the representation, including the effect of smearing, radial scaling, body order, type of density, or induced metric space. The analysis in Chapter 2 showed that with this set of measures, one can retrieve insights about the nature of the features. In Chapter 3, however, when applied in the analysis of the optimized radial basis, the quantitive results retrieved with these measures did not translate to an improved supervised error. This discrepancy between unsupervised and supervised information is a well-known fact in the design of features, as, for example, distant neighbors are the ones that contribute the most to the variance of the features but contribute the least to the local energies; therefore a radial scaling is typically applied to suppress the distant contributions. The proposed approach of directly analyzing the features thus has limited applications for understanding the learning performance of these features. A promising direction is their utilization on features from black box models as NN to understand what kind of information these black box models exhibit by reconstructing the black box model features from features with a mathematically interpretable functional form (e.g., body-order decomposition).

We studied data-driven optimization methods based on the covariance and correlation matrix for the radial basis that preserve symmetries, and we compared them to unoptimized bases. We observed that, in contrast to the unoptimized DVR basis, the GTO basis performed similarly or even better than the unsupervised data-driven one for the single-species dataset. This indicates that the radial-dependent smoothening present in the GTO basis is an important factor for an efficient radial basis. In the analysis of Chapter 3, we assessed that a change in smearing is not always linearly retrievable from features with lower or higher smearings, unlike a change in radial scaling that is completely retrievable. This supports the fact that the choice of smearing can have a significant effect on the features in terms of information content. We therefore suggest future work on a supervised optimization of the basis with a radial-dependent smearing. The strength of the radial distance can be controlled by a hyperparameter that can be optimized on a two-dimensional spline to circumvent the cost of the basis function evaluation.

In the last chapter, we presented parts of the implementation of the package `librascal` that were relevant for the interface implementation to `LAMMPS`. The interface was essential for the construction of an efficient metadynamics framework with `i-PI` and `PLUMED` that we employed for the study of finite-size effects in paraelectric-ferroelectric phase transitions in barium titanates. We find that the transition from the cubic to tetragonal phase occurs at a temperature of 254 K in our simulations. This value compares much more favorably with the experimental value of 393 K than the value of 182 K that was obtained in previous,

similar calculations [140]. Part of the discrepancy can be attributed to the less accurate DFT functional that was used in the this previous work. However, the main source of error comes from very large finite-size effects that only converge when the simulation box is larger than $12 \times 12 \times 12$ unit cells which emphasizes the need for simulations with large systems.

Finally, we pointed out design flaws in `librascal` which we addressed by presenting a future direction for the development of MLIPs based on an abstract data type that allows for the reuse of operations in the implementation of ML models that can act simultaneously and consistently on features and gradients by keeping the required metadata within the data type. Since the efficient manipulation of atomistic data and the implementation of interfaces to MD engines are essential factors in the development of MLIPs, we hope that this effort will enable more flexible and efficient workflows to promote research on more challenging problems.

Bibliography

[1] Martin Jansen. Conceptual inorganic materials discovery–a road map. *Advanced Materials*, 27(21):3229–3242, 2015.

[2] Gerbrand Ceder, Y-M Chiang, DR Sadoway, MK Aydinol, Y-I Jang, and Biying Huang. Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature*, 392(6677):694–696, 1998.

[3] Martin P Andersson, Thomas Bligaard, Arkady Kustov, Kasper E Larsen, Jeffrey Greeley, Tue Johannessen, Claus H Christensen, and Jens K Nørskov. Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts. *Journal of Catalysis*, 239(2):501–506, 2006.

[4] Kesong Yang, Wahyu Setyawan, Shidong Wang, Marco Buongiorno Nardelli, and Stefano Curtarolo. A search model for topological insulators with high-throughput robustness descriptors. *Nature materials*, 11(7):614–619, 2012.

[5] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials*, 15(10):1120–1127, 2016.

[6] Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Phys. Rev. X*, 8(4):041048, December 2018. https://link.aps.org/doi/10.1103/PhysRevX.8.041048.

[7] Gabriele C Sosso, Davide Donadio, Sebastiano Caravati, Jörg Behler, and Marco Bernasconi. Thermal transport in phase-change materials from atomistic simulations. *Physical Review B*, 86(10):104301, 2012.

[8] Choongseok Chang, Volker L. Deringer, Kalpana S. Katti, Veronique Van Speybroeck, and Christopher M. Wolverton. Simulations in the era of exascale computing. *Nature Reviews Materials*, 8(5):309–313, March 2023.

[9] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134(7), 2011.

[10] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):184115, May 2013.

[11] Aria Mansouri Tehrani, Anton O Oliynyk, Marcus Parry, Zeshan Rizvi, Samantha Couper, Feng Lin, Lowell Miyagi, Taylor D Sparks, and Jakoah Brgoch. Machine learning directed search for ultraincompressible, superhard materials. *Journal of the American Chemical Society*, 140(31):9844–9853, 2018.

[12] Gabriele C Sosso, Volker L Deringer, Stephen R Elliott, and Gábor Csányi. Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials. *Molecular Simulation*, 44(11):866–880, 2018.

[13] Yasemin Basdogan, Mitchell C Groenenboom, Ethan Henderson, Sandip De, Susan B Rempe, and John A Keith. Machine learning guided approach for studying solvation environments. *Journal of Chemical Theory and Computation*, 2019.

[14] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[15] Haoyan Huo and Matthias Rupp. Unified representation for machine learning of molecules and crystals. *arXiv preprint arXiv:1704.06439*, 13754, 2017.

[16] Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V. Shapeev, Aidan P. Thompson, Mitchell A. Wood, and Shyue Ping Ong. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *J. Phys. Chem. A*, page acs.jpca.9b08723, January 2020.

[17] Michael J. Willatt, Félix Musil, and Michele Ceriotti. Atom-density representations for machine learning. *J. Chem. Phys.*, 150(15):154110, April 2019.

[18] Juhwan Noh, Jaehoon Kim, Helge S Stein, Benjamin Sanchez-Lengeling, John M Gregoire, Alan Aspuru-Guzik, and Yousung Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5):1370–1384, 2019.

[19] Wikipedia. Alcohol (chemistry) — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Alcohol%20(chemistry)&oldid=1177472846, 2023. [Online; accessed 15-October-2023].

[20] Felix Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Physics-inspired structural representations for molecules and materials. *Chemical Reviews*, 121(16):9759–9815, 2021.

[21] Emil Prodan and Walter Kohn. Nearsightedness of electronic matter. *Proceedings of the National Academy of Sciences*, 102(33):11635–11638, 2005.

[22] Jigyasa Nigam, Sergey Pozdnyakov, Guillaume Fraux, and Michele Ceriotti. Unified theory of atom-centered representations and message-passing machine-learning schemes. *The Journal of Chemical Physics*, 156(20), 2022.

[23] Bing Huang and O. Anatole Von Lilienfeld. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.*, 145(16), 2016.

[24] Michael J. Willatt, Félix Musil, and Michele Ceriotti. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Phys. Chem. Chem. Phys.*, 20(47):29661–29668, 2018.

[25] Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019.

[26] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O Anatole Von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters*, 6(12):2326–2331, 2015.

[27] James Barker, Johannes Bulin, Jan Hamaekers, and Sonja Mathias. Localized coulomb descriptors for the gaussian approximation potential. *arXiv preprint arXiv:1611.05126*, 2016.

[28] Bing Huang and O. Anatole von Lilienfeld. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *The Journal of Chemical Physics*, 145.

[29] Félix Musil and Michele Ceriotti. Machine learning at the atomic scale. *CHIMIA International Journal for Chemistry*, 73(12):972–982, 2019.

[30] JS Dowker. Spherical harmonics, invariant theory and maxwell's poles. *arXiv preprint arXiv:0805.1904*, 2008.

[31] AP Yutsis and AA Bandzaitis. Theory of angular momentum in quantum mechanics. *Vil'nyus*, 1965.

[32] Alan Robert Edmonds. *Angular momentum in quantum mechanics*. Princeton university press, 1996.

[33] Jigyasa Nigam, Sergey Pozdnyakov, and Michele Ceriotti. Recursive evaluation and iterative contraction of N -body equivariant features. *J. Chem. Phys.*, 153(12):121101, September 2020.

[34] Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch– Gordan Nets: a Fully Fourier Space Spherical Convolutional Neural Network. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[35] Tao Yan, Jiamin Wu, Tiankuang Zhou, Hao Xie, Feng Xu, Jingtao Fan, Lu Fang, Xing Lin, and Qionghai Dai. Fourier-space diffractive deep neural network. *Physical review letters*, 123(2):023901, 2019.

[36] Sergey Pozdnyakov, Artem R Oganov, Efim Mazhnik, Arslan Mazitov, and Ivan Kruglov. Fast general two-and three-body interatomic potential. *arXiv preprint arXiv:1910.07513*, 2019.

[37] Stephen R Xie, Matthias Rupp, and Richard G Hennig. Ultra-fast interpretable machine-learning potentials. *npj Computational Materials*, 9(1):162, 2023.

[38] Alexander V. Shapeev. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.*, 14(3):1153–1173, January 2016.

[39] Félix Musil, Max Veit, Alexander Goscinski, Guillaume Fraux, Michael J Willatt, Markus Stricker, Till Junge, and Michele Ceriotti. Efficient implementation of atom-density representations. *The Journal of Chemical Physics*, 154(11), 2021.

[40] Lucjan Piela. Appendix j - orthogonalization. In Lucjan Piela, editor, *Ideas of Quantum Chemistry (Second Edition)*, pages e99–e103. Elsevier, Oxford, second edition edition, 2014.

[41] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

[42] Genevieve Dusson, Markus Bachmayr, Gábor Csányi, Ralf Drautz, Simon Etter, Cas van der Oord, and Christoph Ortner. Atomic cluster expansion: Completeness, efficiency and stability. *Journal of Computational Physics*, 454:110946, 2022.

[43] Alexander Goscinski, Félix Musil, Sergey Pozdnyakov, Jigyasa Nigam, and Michele Ceriotti. Optimal radial basis for density-based atomic representations. *The Journal of Chemical Physics*, 155(10), 2021.

[44] Filippo Bigi, Kevin K Huguenin-Dumittan, Michele Ceriotti, and David E Manolopoulos. A smooth basis for atomistic machine learning. *The Journal of Chemical Physics*, 157(23), 2022.

[45] Filippo Bigi, Guillaume Fraux, Nicholas J. Browning, and Michele Ceriotti. Fast evaluation of spherical harmonics with sphericart. *The Journal of Chemical Physics*, 159(6):064802, 08 2023.

[46] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[47] Guillem Simeon and Gianni De Fabritiis. Tensornet: Cartesian tensor representations for efficient learning of molecular potentials. *arXiv preprint arXiv:2306.06482*, 2023.

[48] Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V Shapeev, Aidan P Thompson, Mitchell A Wood, et al. Performance and cost assessment of machine learning interatomic potentials. *The Journal of Physical Chemistry A*, 124(4):731–745, 2020.

[49] Miguel A Caro. Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials. *Physical Review B*, 100(2):024112, 2019.

[50] Nataliya Lopanitsyna, Guillaume Fraux, Maximilian A. Springer, Sandip De, and Michele Ceriotti. Modeling high-entropy transition metal alloys with alchemical compression. *Phys. Rev. Mater.*, 7:045802, Apr 2023.

[51] Michael J Willatt, Félix Musil, and Michele Ceriotti. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Physical Chemistry Chemical Physics*, 20(47):29661–29668, 2018.

[52] O. Anatole von Lilienfeld, Raghunathan Ramakrishnan, Matthias Rupp, and Aaron Knoll. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quantum Chem.*, 115(16):1084–1093, August 2015.

[53] Sergey N Pozdnyakov, Michael J Willatt, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Incompleteness of Atomic Structure Representations. *Phys. Rev. Lett.*, 125:166001, 2020.

[54] Benjamin Helfrecht, Rose K Cersonsky, Guillaume Fraux, and Michele Ceriotti. Structure-property maps with Kernel Principal Covariates Regression. *Mach. Learn.: Sci. Technol.*, July 2020.

[55] Berk Onat, Christoph Ortner, and James R. Kermode. Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials. *J. Chem. Phys.*, 153(14):144106, October 2020.

[56] Jonathan E Moussa. Comment on "fast and accurate modeling of molecular atomization energies with machine learning". *Physical review letters*, 109(5):059801, 2012.

[57] Ali Sadeghi, S Alireza Ghasemi, Bastian Schaefer, Stephan Mohr, Markus A Lill, and Stefan Goedecker. Metrics for measuring distances in configuration spaces. *The Journal of chemical physics*, 139(18):184118, 2013.

[58] Li Zhu, Maximilian Amsler, Tobias Fuhrer, Bastian Schaefer, Somayeh Faraji, Samare Rostami, S Alireza Ghasemi, Ali Sadeghi, Migle Grauzinyte, Chris Wolverton, and Stefan Goedecker. A fingerprint based metric for measuring similarities of crystalline structures. *J. Chem. Phys.*, 144(3):034203, January 2016.

[59] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18(20):13754–13769, 2016.

[60] Behnam Parsaeifard, Deb Sankar De, Anders Steen Christensen, Felix Andreas Faber, Emir Kocer, Sandip De, Jörg Behler, Anatole von Lilienfeld, and Stefan Goedecker. An assessment of the structural resolution of various fingerprints commonly used in machine learning. *Mach. Learn.: Sci. Technol.*, August 2020.

[61] Kari Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of machine learning research*, 3(Mar):1415–1438, 2003.

[62] Alexander Goscinski, Victor Paul Principe, Guillaume Fraux, Sergei Kliavinek, Benjamin Aaron Helfrecht, Philip Loche, Michele Ceriotti, and Rose Kathleen Cersonsky. scikit-matter: A suite of generalisable machine learning methods born out of chemistry and materials science. *Open Research Europe*, 3:81, 2023.

[63] Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, March 1966.

[64] S T Roweis and L K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–6, December 2000.

[65] Bernhard Schölkopf. The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307, 2001.

[66] Bernard Haasdonk and Claus Bahlmann. Learning with distance substitution kernels. In *Joint pattern recognition symposium*, pages 220–227. Springer, 2004.

[67] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[68] Chris J Pickard and R J Needs. Ab initio random structure searching. *J. Phys. Condens. Matter*, 23(5):053201, February 2011.

[69] Chris J. Pickard. AIRSS data for carbon at 10gpa and the C+N+H+O system at 1gpa, 2020.

[70] Félix Musil, Max Veit, Till Junge, Markus Stricker, Alexander Goscinski, Guillaume Fraux, Rose Cersonsky, Michael J Willatt, Andrea Grisafi, and Michele Ceriotti. librascal – A scalable and versatile library to generate representations for atomic-scale learning. https://github.com/cosmo-epfl/librascal.

[71] Andreas Singraber, Tobias Morawietz, Jörg Behler, and Christoph Dellago. Parallel multistream training of high-dimensional neural network potentials. *Journal of chemical theory and computation*, 15(5):3075–3092, 2019.

[72] Jörg Behler. Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations. *Phys. Chem. Chem. Phys. PCCP*, 13(40):17930–55, October 2011.

[73] Giulio Imbalzano, Andrea Anelli, Daniele Giofré, Sinja Klees, Jörg Behler, and Michele Ceriotti. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.*, 148(24):241730, June 2018.

[74] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.*, 3(12):e1701816, December 2017.

[75] Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.*, 13(11):5255–5264, November 2017.

[76] Federico M. Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. Chemical shifts in molecular solids by machine learning. *Nat. Commun.*, 9(1):4501, December 2018.

[77] K. V.Jovan Jose, Nongnuch Artrith, and Jörg Behler. Construction of high-dimensional neural network potentials using environment-dependent atom pairs. *J. Chem. Phys.*, 136(19):194111, 2012.

[78] Jörg Behler. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.*, 115(16):1032–1050, August 2015.

[79] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[80] Andreas Singraber. n2p2 - a neural network potential package.

[81] Aldo Glielmo, Claudio Zeni, and Alessandro De Vita. Efficient nonparametric n -body force fields from machine learning. *Phys. Rev. B*, 97(18):184307, May 2018.

[82] Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B*, 99:014104, Jan 2019.

[83] Ryosuke Jinnouchi, Ferenc Karsai, Carla Verdi, Ryoji Asahi, and Georg Kresse. Descriptors representing two- and three-body atomic distributions and their effects on the accuracy of machine-learned inter-atomic potentials. *J. Chem. Phys.*, 152(23):234102, June 2020.

[84] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters*, 104(13):136403, 2010.

[85] SS Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Its Appl.*, 18(4):784–786, 1974.

[86] Scott D. Cohen and Leonidas Guibas. The earth mover's distance: Lower bounds and invariance under translation. pages 1–44, 1997.

[87] Marco Cuturi. Permanents, transportation polytopes and positive definite kernels on histograms. *Int. Jt. Conf. Artif. Intell. IJCAI*, pages 732–737, 2007.

[88] Onur Çaylak, Anatole von Lilienfeld, and Björn Baumeier. Wasserstein metric for improved quantum machine learning with adjacency matrix representations. *Mach. Learn.: Sci. Technol.*, June 2020.

[89] Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.*, 145(17):170901, November 2016.

[90] Dino Oglic and Thomas Gaertner. Learning in reproducing kernel Krein spaces. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3859–3867. PMLR, 10–15 Jul 2018.

[91] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1):1–12, 2017.

[92] Christopher Sutton, Luca M Ghiringhelli, Takenori Yamamoto, Yury Lysogorskiy, Lars Blumenthal, Thomas Hammerschmidt, Jacek R Golebiowski, Xiangyue Liu, Angelo Ziletti, and Matthias Scheffler. Crowd-sourcing materials-science challenges with the nomad 2018 kaggle competition. *npj Computational Materials*, 5(1):1–11, 2019.

[93] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. In *Advances in Neural Information Processing Systems*, pages 8466–8478, 2019.

[94] Ghanshyam Pilania, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran, and Ramamurthy Ramprasad. Accelerating materials property predictions using machine learning. *Scientific reports*, 3(1):1–6, 2013.

[95] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

[96] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.*, 2(8):083802, August 2018.

[97] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148(24):241722, June 2018.

[98] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

[99] Martín Leandro Paleico and Jörg Behler. A bin and hash method for analyzing reference data and descriptors in machine learning potentials. *Machine Learning: Science and Technology*, 2(3):037001, 2021.

[100] Ansgar Schäfer, Hans Horn, and Reinhart Ahlrichs. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *The Journal of Chemical Physics*, 97(4):2571–2577, August 1992.

[101] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.*, 180(11):2175–2196, November 2009.

[102] Sergey Pozdnyakov. NICE libraries. https://github.com/cosmo-epfl/nice.

[103] Sijmen de Jong and Henk A.L. Kiers. Principal covariates regression. *Chemometrics and Intelligent Laboratory Systems*, 14(1-3):155–164, April 1992.

[104] Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant Molecular Neural Networks. In *NeurIPS*, page 10, 2019.

[105] Benjamin Kurt Miller, Mario Geiger, Tess E. Smidt, and Frank Noé. Relevance of rotationally equivariant convolutions for predicting molecular properties. *ArXiv Prepr. ArXiv200808461*, 2020.

[106] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data*, 1:1–7, August 2014.

[107] Alexander Goscinski, Guillaume Fraux, Giulio Imbalzano, and Michele Ceriotti. The role of feature space in atomistic learning. *Machine Learning: Science and Technology*, 2(2):025028, 2021.

[108] Y Eldar, M Lindenbaum, M Porat, and Y Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.*, 6(9):1305–15, January 1997.

[109] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. Demonstrating the transferability and the descriptive power of sketch-map. *J. Chem. Theory Comput.*, 9(3):1521–1532, March 2013.

[110] Rose K Cersonsky, Benjamin Helfrecht, Edgar Albert Engel, Sergei Kliavinek, and Michele Ceriotti. Improving Sample and Feature Selection with Principal Covariates Regression. *Mach. Learn.: Sci. Technol.*, May 2021.

[111] Frank H Stillinger and Thomas A Weber. Computer simulation of local order in condensed phases of silicon. *Physical review B*, 31(8):5262, 1985.

[112] Jerry Tersoff. Empirical interatomic potential for silicon with improved elastic properties. *Physical Review B*, 38(14):9902, 1988.

[113] Thomas B Blank, Steven D Brown, August W Calhoun, and Douglas J Doren. Neural network models of potential energy surfaces. *The Journal of chemical physics*, 103(10):4129–4137, 1995.

[114] Alex Brown, Bastiaan J Braams, Kurt Christoffel, Zhong Jin, and Joel M Bowman. Classical and quasiclassical spectral analysis of ch 5+ using an ab initio potential energy surface. *The Journal of chemical physics*, 119(17):8790–8793, 2003.

[115] Sönke Lorenz, Matthias Scheffler, and Axel Gross. Descriptions of surface chemical reactions using a neural network representation of the potential-energy surface. *Physical Review B*, 73(11):115431, 2006.

[116] Jörg Behler and Michele Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98(14):146401, April 2007.

[117] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.*, 271:108171, 2022.

[118] Lorenzo Gigli, Alexander Goscinski, Michele Ceriotti, and Gareth A. Tribello. Modeling the ferroelectric phase transition in barium titanate with DFT accuracy and converged sampling. *arXiv preprint arXiv:2310.12579*, 2023.

[119] Lorenzo Gigli, Davide Tisi, Federico Grasselli, and Michele Ceriotti. Mechanism of charge transport in lithium thiophosphate. *arXiv preprint arXiv:2310.15679*, 2023.

[120] Richard H Bartels, John C Beatty, and Brian A Barsky. Hermite and cubic spline interpolation. *An Introduction to Splines for Use in Computer Graphics and Geometric Modelling*, pages 9–17, 1998.

[121] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[122] Federico M Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. Chemical shifts in molecular solids by machine learning. *Nature communications*, 9(1):1–10, 2018.

[123] B Hess, C Kutzner, D van der Spoel, and E Lindahl. {GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation}. *J Chem Theory Comput*, 4(3):435–447, 2008.

[124] Thomas D Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z Khaliullin, Ole Schütt, Florian Schiffmann, et al. Cp2k: An electronic structure and molecular dynamics software package-quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics*, 152(19), 2020.

[125] Venkat Kapil, Mariana Rossi, Ondrej Marsalek, Riccardo Petraglia, Yair Litman, Thomas Spura, Bingqing Cheng, Alice Cuzzocrea, Robert H Meißner, David M Wilkins, et al. i-pi 2.0: A universal force engine for advanced molecular simulations. *Computer Physics Communications*, 236:214–223, 2019.

[126] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[127] LAMMPS developers. LAMMPS Users Manual - Commands - newton command, 2023.

[128] GROMMACS developers. GROMACS Users Manual - nstcalcenergy, 2023.

[129] Ibrahim Haddad. Artificial intelligence and data in open source. https://www.linuxfoundation.org/research/artificial-intelligence-and-data-in-open-source, mar 2022.

[130] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[131] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[132] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[133] Grace Wahba, Xiwu Lin, Fangyu Gao, Dong Xiang, Ronald Klein, and Barbara Klein. The bias-variance tradeoff and the randomized gacv. *Advances in Neural Information Processing Systems*, 11, 1998.

[134] Alex Smola and Peter Bartlett. Sparse greedy gaussian process regression. *Advances in neural information processing systems*, 13, 2000.

[135] Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[136] Albert P. Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry*, 115(16):1051–1057, 2015.

[137] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

[138] Michele Ceriotti, Michael J Willatt, and Gábor Csányi. Machine learning of atomic-scale properties based on physical principles. *Handbook of Materials Modeling: Methods: Theory and Modeling*, pages 1911–1937, 2020.

[139] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212, 2020.

[140] Lorenzo Gigli, Max Veit, Michele Kotiuga, Giovanni Pizzi, Nicola Marzari, and Michele Ceriotti. Thermodynamics and dielectric response of BaTiO3 by data-driven modeling. *npj Computational Materials*, 8(1):1–17, September 2022. Number: 1 Publisher: Nature Publishing Group.

[141] F. Giberti, B. Cheng, G. A. Tribello, and M. Ceriotti. Iterative Unbiasing of Quasi-Equilibrium Sampling. *Journal of Chemical Theory and Computation*, 16(1):100–107, January 2020.

[142] Piero Gasparotto, Robert Horst Meißner, and Michele Ceriotti. Recognizing Local and Global Structural Motifs at the Atomic Scale. *J. Chem. Theory Comput.*, 14(2):486–498, February 2018.

[143] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[144] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[145] LAMMPS developers. LAMMPS Users Manual - Build LAMMPS - Packages with extra build options - ML-IAP, 2023.

[146] LAMMPS developers. LAMMPS Users Manual - Build LAMMPS - Packages with extra build options - ML-PACE, 2023.

[147] LAMMPS developers. LAMMPS Users Manual - Build LAMMPS - Packages with extra build options - ML-POD, 2023.

[148] LAMMPS developers. LAMMPS Users Manual - Build LAMMPS - Packages with extra build options - ML-QUIP, 2023.

[149] LAMMPS developers. LAMMPS Users Manual - Build LAMMPS - Packages with extra build options - ML-HDNNP, 2023.

[150] Junjie Bai, Fang Lu, Ke Zhang, et al. Onnx: Open neural network exchange. https://github.com/onnx/onnx, 2019.

[151] Daniel S Karls, Matthew Bierbaum, Alexander A Alemi, Ryan S Elliott, James P Sethna, and Ellad B Tadmor. The openkim processing pipeline: A cloud-based automatic material property computation engine. *The Journal of Chemical Physics*, 153(6), 2020.

[152] Ryan S. Elliott and Ellad B. Tadmor. Knowledgebase of interatomic models (kim) application programming interface (api), 2011.

List of Figures