Efficient and insightful descriptors for representing molecular and material space

Thèse n. TODO 2023
Présentée le December 10, 2023
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 Name Surname, président du jury
Prof Name Surname, directeur de thèse
Prof Name Surname, rapporteur
Prof Name Surname, rapporteur
Prof Name Surname, rapporteur

Lausanne, EPFL, 2023

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, 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 `scikit-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, December 10, 2023 A. G.

Abstract

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 the synthesis conditions. A viable alternative is to calculate material properties using computer simulations, that make it possible to exploit advances in parallel computing to construct databases with millions of entries, and to obtain results that are internally consistent. The quantitative accuracy of these predictions, however, is dependent on the quality of the reference electronic structure calculations, increasing the computational effort and reducing the breadth of the searches. Data-driven approaches have been applied to reduce the cost of accurate computational studies, by using only a small number of reference calculations for a representative subset of materials space, and using them to train surrogate models that predict inexpensively the outcome of such calculation on new materials. The way materials structures are processed into a numerical description as input of machine learning algorithms is crucial to obtain efficient and computationally inexpensive models. Recent advancements in the design of information-efficient representations based on atomic densities have embedded novel types of information, such as neighborhood environments or pair descriptions.

Despite the rapid development in offloading calculations to more dedicated hardware, these enhancements nevertheless substantially increase the cost of the representation that 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 they encapsulate. Insights from such analyses 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 into understanding the inherent nature of the information capacity of these representations remains mostly uncharted. This thesis introduces a set of measures that facilitate quantitative analysis concerning the relationship between features and datasets, thereby assisting in such decision-making processes and providing valuable insights to the academic community. We demonstrate how these set of measures can be applied to analyse representations that are built in terms of body-order correlations of the 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.

Abstract

Additionally, a considerable amount of effort has been dedicated to optimize the basis set involved featurization of the representation, typically driven by heuristic considerations on the behavior of the regression target. This thesis showcases a scheme that utilizes splines to approximate the basis expansion coefficients, paving the way for expansive optimization methods to create more effective basis functions at no additional cost during simulation time. This is pivotal in simulations targeting materials encompassing a high variety of chemical species or relying on qualitative collective variables.
Lastly, complementing these efforts is the integration of the developed methods into well-maintained and thoroughly documented packages. This integration facilitates an enhancement of the existing methods and the incorporation of the methods into new workflows. Specifically, this thesis introduces an implemented framework for metadynamics simulation with machine learing interatomic potentials in composition with message-passing interface of `LAMMPS` It also explores the results obtained using this framework, particularly focusing on the finite-size effecs of the Curie point in barium titanate. Additionally, the thesis outlines potential future developments in creating a modular machine learning ecosystem for atomistic simulations.

Contents

Introduction

The discovery of new materials is one of the core pillars of technology, as every technology relies on a material and, needless to say, would not exist without it[1]. The search for new materials is bound by thermodynamic laws which tell what configurations are stable and can therefore be considered as potential material. methods provide approximate stability criteria which are in good agreement with experiments[2] making them a viable tool for the screening of new materials[3, 4, 5, 6]. 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[7], leading to the exploration of previously computationally intractable problems, such as the thermal conductivity of amorphous germanium telluride[8]. These methods are based on transforming geometrical, physical and chemical information into a vector representation, referred as descriptor, to then use it as features in a machine learning model. The development of expressive and computational inexpensive descriptors[9, 10] has lead 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 application.

The efficient computation of expressive descriptors is a challenging problem which 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 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.

The first part of this thesis presents a collection of measures that serve as toolkit to guide the choice of the descriptor and model. The second part discusses the implementation of machine learning model models as interatomic potentials, covering on one hand the efficient implementation of descriptors and data-driven models, and on the other hand their deployment into molecular dynamics software.

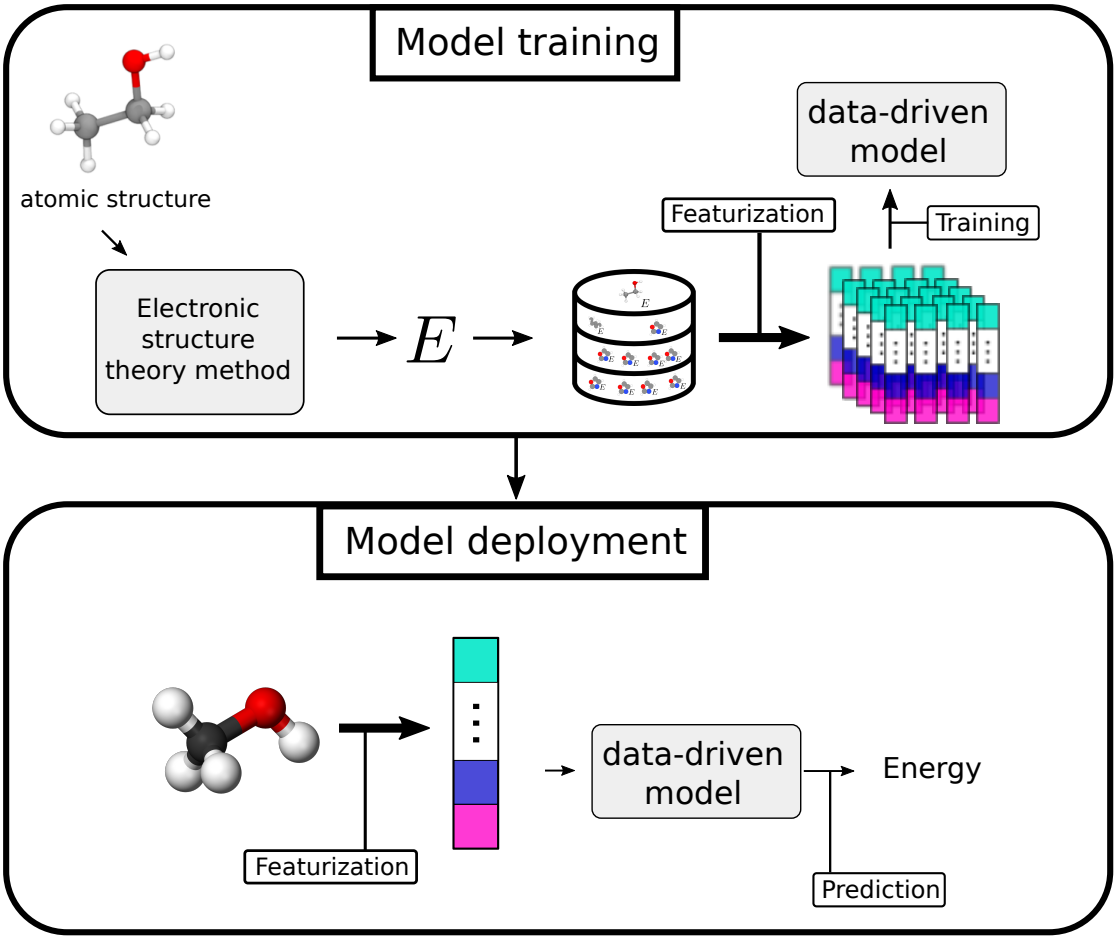

Figure 1: A schematic showing the idea of high-throughput calculations with data-driven model that serves as surrogate model to bypass the expensive electronic structure theory calculations after a training the model.

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 coeffients. 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.

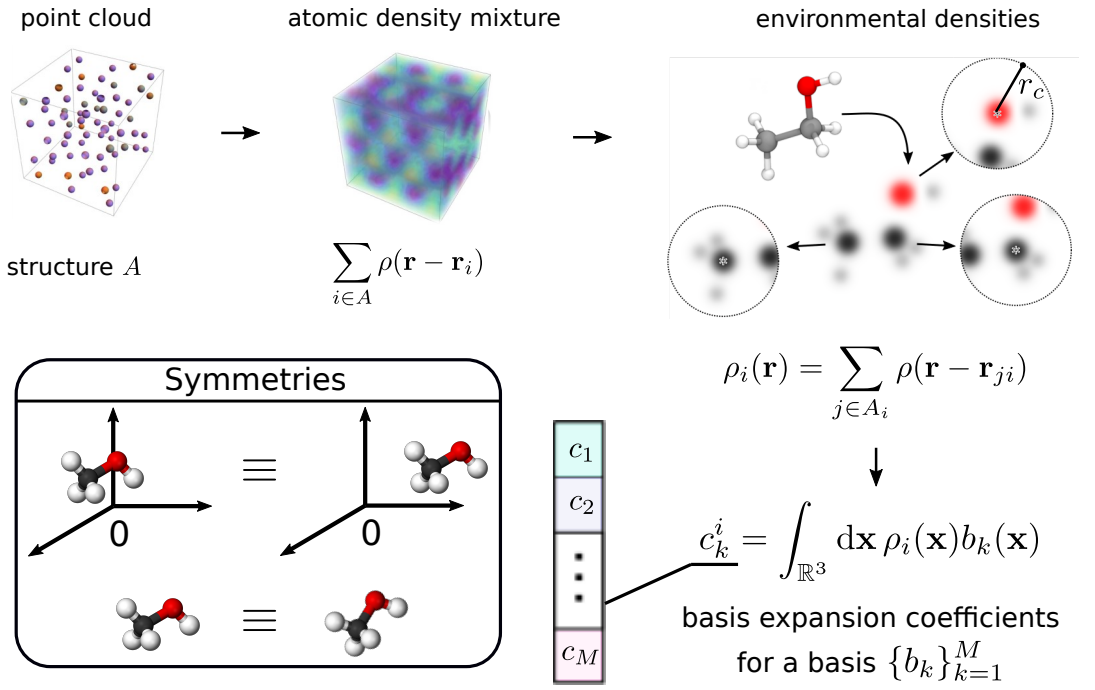

Figure 1.1: A schematic showing the featurization of an atomic structure A based on the atom-centered density correlation functions. The figure of the atoms in the box are retrieved from Ref. [18]. The figure of the atomic environments is retrieved from Ref. [17]. The methanol molecule is retrieved from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

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 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

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

This class of representations has been named *atom-centered density correlation* (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. This fact can be extended to the L^p norm

distance and the Wasserstein W_p metric, a generalization of the Earth mover distance similar as L^p is a generalization of L^1

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

Furthermore, 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 a description [28]. These can be seen one dimensional projections of the higher order ACDC functions and do not consider the whole correlation space. It has been shown that descriptors in this category can reach comparable accuracies to other methods [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) [29]. 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 [30]

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

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

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

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 [31] and in angular momentum theory [32] 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

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

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

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

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

$$\propto \sum_l P_l(\hat{\mathbf{r}} \cdot \hat{\mathbf{r}}') \quad \text{(addition theorem, where } P_l \text{ Legendre polynomial [33])} \tag{1.15e}$$

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

$$D_{m_1 m_1'}^{l_1}(\hat{R}) D_{m_2 m_2'}^{l_2}(\hat{R}) = \sum_{lmm'} D_{mm'}^l(\hat{R}) C_{mm_1 m_2}^{ll_1 l_2} C_{m'm_1'm_2'}^{ll_1 l_2} \tag{1.17}$$

where $C_{\mu m_1 m_2}^{ll_1 l_2}$ are the real Clebsch-Gordan coefficients [32, 34]. 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

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

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

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), formally derived in Ref. [34]. 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 [34, 35, 36].

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].

1.3.1 Density trick

It becomes apparent from Eqs. (1.19) 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 fact has been referret to as the *density trick*. It shifts the computation from the evaluation of the basis expansions across all ν-tuples to the computation of tensor products between the expansion coefficients of order ν as indicated the derivation in Eq. (1.19b). For example, the cost of the order 2 coefficients in Eq. (1.19b) without the usage of the density trick requires the evaluation of $O(\binom{N}{2}) = O(N^2)$ triplets for each of the $O(M^2)$ basis functions resulting in a total time complexity of $O(M^2N^2)$. In principle, the number of basis functions can be chosen arbitrary, we compare however a number of basis functions that is equal to the number of basis function acquired by applying the density trick. With the density trick one has to compute the expansion coefficients for the density $O(MN)$ to then to increase the order by a contracted tensor product scaling as $O(M^2)$ resulting in a total time complexity of $O(MN + M^2)$. Even though the scaling favors the use of the density trick, when embedding splines into the computation of the features, the slower iteration over all $(\nu+1)$-body parts can be balanced out by omitting the computational cost of tensor product and the model as shown in Ref . [37].

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 as a means to apply the density trick to produce higher-order coefficients. An extensively employed representative of using the density trick with Dirac δ densities is named *atomic cluster expansion* (ACE) [25].

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 mitigating the convergence in L^2 of the sphere. 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. [35]. Due to this limitation, recent advancements have been more focused on improving the computation of the spherical harmonics itself by exploiting recursive relationships [45] or switching to a Cartesian tensor basis [38, 46, 47]

$$\mathbf{T}^{(\nu)} = \hat{\mathbf{r}}_{1i} \otimes \cdots \otimes \hat{\mathbf{r}}_{\nu i} \quad \text{(Cartesion moment tensor of order } \nu\text{)}. \tag{1.21}$$

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 [37, 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

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

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

$$c^{ij}_{nlm} = c^{ij}_{nl} c^{ij}_{lm}. \quad (1.23)$$

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]

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

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 **??**. 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

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

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]

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

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 2. The other approach learns a embedding c_{ak}^{ij} 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

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

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 Symmetry-adapted data-driven basis optimization

Several algorithmic recipes for the construction of basis have been proposed [38, 39, 42, 52] 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) [53], and that it can be practically, and inexpensively, evaluated as a numerical basis with striking similarities to ideas in electronic-structure methods [54]. 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.

2.1 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 [55].

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

$$C^{al}_{nn'} = \frac{1}{N}\sum_i \sum_m c^i_{nlm} c^i_{n'lm} \tag{2.1}$$

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 Eq. 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

$$\sum_n U^{al}_{qn} c^i_{anlm} = c^i_{aqlm}. \tag{2.2}$$

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^{al}_q. 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. 55. When combining Eq. (2.2) with Eq. (1.23), however, one sees that the contracted coefficients can be evaluated directly

$$c^i_{aqlm} = \sum_{j \in A_i} \delta_{aa_j} c^{ij}_{ql} c^{ij}_{lm}. \tag{2.3}$$

using the contracted radial integrals

$$c^{ij}_{ql} = \sum_n U^{a_j l}_{qn} c^{ij}_{nl} \tag{2.4}$$

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 **??**. 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 (2.1) and diagonalize it, finding the contraction coefficients U^{al}_{qn}
3. Evaluate the contracted radial integrals using Eq. (2.4), over a dense radial grid
4. Use a spline approximation to evaluate directly the radial integrals (2.3) 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

$$R_{aql}(r) \equiv \sum_n U^{al}_{qn} c_{anl} R_{anl}(r). \tag{2.5}$$

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 Sec. 2.2.

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 Sec. 2.2.

2.1.1 Mixed-species basis

Even though Eq. (2.1) is defined separately for different species a, it is also possible to compute cross-correlations between different elemental channels, defining

$$C^{l}_{nn'} = \frac{1}{N}\sum_{i}\sum_{m} c^{i}_{anlm} c^{i}_{a'n'lm} \tag{2.6}$$

as done in the NICE framework[55] following ideas proposed in Ref. 17, resulting in coefficients that combine information on multiple species

$$c^{i}_{qlm} = \sum_{n} U^{l}_{q;an} c^{i}_{anlm}, \tag{2.7}$$

similar in spirit to the alchemical contraction discussed in Ref. 24. It is worth noting that although the NICE code[56] contains the infrastructure to compute these contractions as a post-processing of the primitive basis, the implementation we propose in `librascal`[57] 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:

$$\begin{aligned} c^{i}_{qlm} &= \sum_{an} U^{l}_{q;an} c^{i}_{anlm} && (2.8)\\ &= \sum_{an} U^{l}_{q;an} \sum_{j\in A_i} \delta_{aa_j} c^{ij}_{nlm} && (2.9)\\ &= \sum_{j\in A_i}\sum_{n} U^{l}_{q;a_j n} c^{ij}_{nlm}. && (2.10) \end{aligned}$$

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.

2.1.2 Supervised basis set optimization

For a given number of radial functions, and a target data set, the data-driven contracted basis (2.2) 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, 58]. 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) [59, 60]. 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

$$c_{q_\gamma 00} = \sum_n U^{a0}_{q_\gamma n} c_{n00}. \tag{2.11}$$

which is a special case of Eq. (2.5) for $l = 0$, where $U^{a0}_{q_\gamma n}$ is obtained as the orthogonalized PCovR projector, as discussed in Refs. 59, 60, 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.

2.1.3 Multispectrum

We discuss the general case of "multispectra" in the frame of the N-body iterative construction of equivariant (NICE) features[55], but analogous considerations apply to similar many-body descriptors such as the atomic cluster expansion (ACE)[42, 52] or the moment tensor potential (MTP)[38], and is likely to be relevant also for covariant neural networks[61, 62]. The NICE iteration increases the body order of features that describe correlations between ν neighbors by combining lower order features as described in Eq. (1.18).

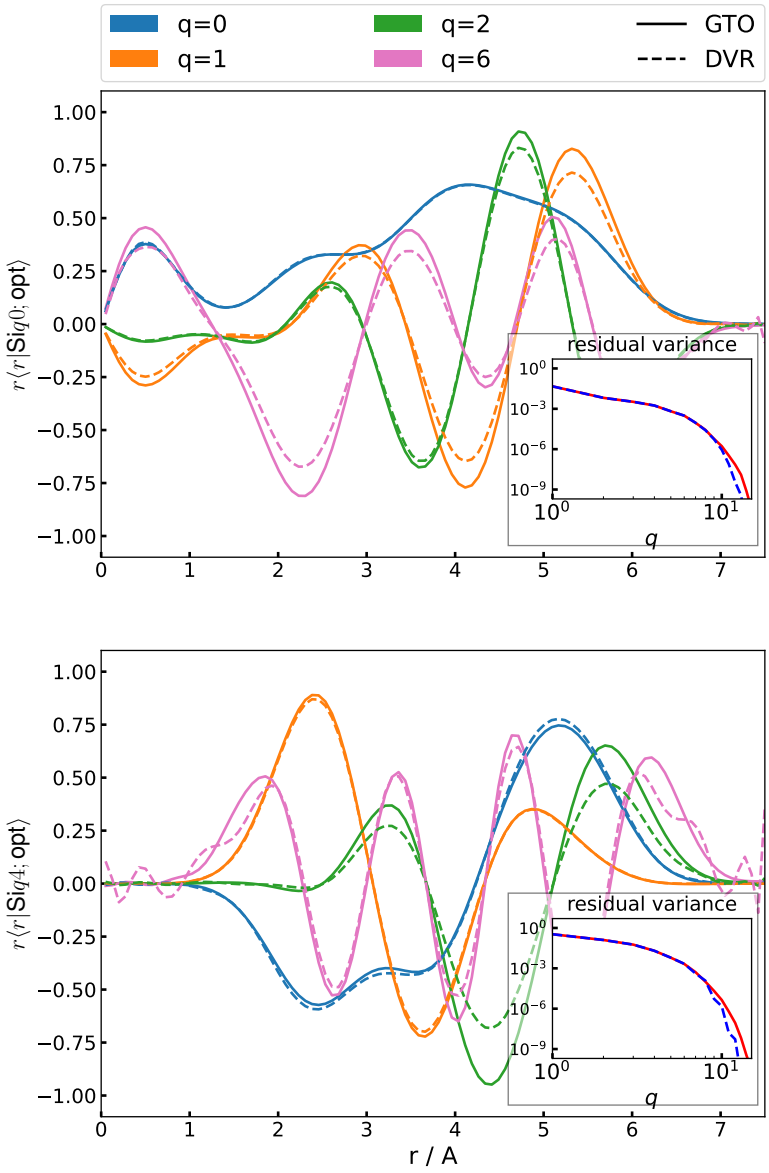

Figure 2.1: Several examples of the optimized radial basis functions on the silicon dataset for $l = 0$ and $l = 4$ using DVR and GTO as primitive basis contracted from $n_{\max} = 20$, with $r_{\text{cut}} = 6$.

2.2 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. 7, and the prediction of atomization energies for the organic molecules from the QM9 dataset [63]. 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[64].

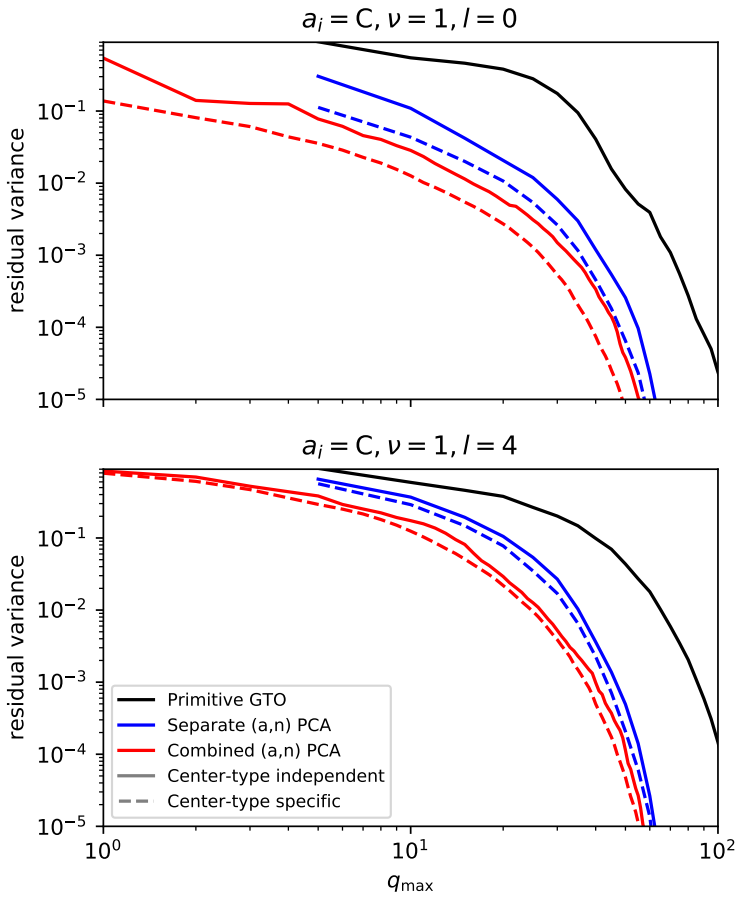

Figure 2.2: Convergence of the residual variance for the expansion coefficients of the density as a function of the number radial basis functions q_{max}, computed for the QM9 dataset and for environments centered on a C atom. The different series correspond to a GTO basis of increasing size (black), 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.

2.2.1 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

$$RV = 1 - \frac{\sum_i \sum_{q=1}^{q_{\max}} \left| c^i_{qlm} \right|^2}{\sum_i \sum_{n=1}^{n_{\max}} \left| c^i_{nlm} \right|^2} \tag{2.12}$$

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 Fig. 2.1. 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 neighbour type along the radial index, or compute a covariance matrix that combines the (a,n) indices. Figure 2.2 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.

2.2.2 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 correlations, and compute the global feature space reconstruction error[65] (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[66] and farthest point sampling (FPS)[67, 68], 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 a GTO basis of the same size (Figure 2.3). 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 2.1.3, 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 (Fig. 2.4, 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 correlation (Fig. 2.4, 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.

2.2.3 Regression models

The accuracy of a Gaussian approximation potential based on SOAP features, trained using both energy and forces , seen in Fig. 2.5 shows an improvement of the cross-validation error for the most aggressive truncation of the feature space (up to $n_{\mathrm{max}} \approx 6$ for forces, and $n_{\mathrm{max}} \approx 4$ 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

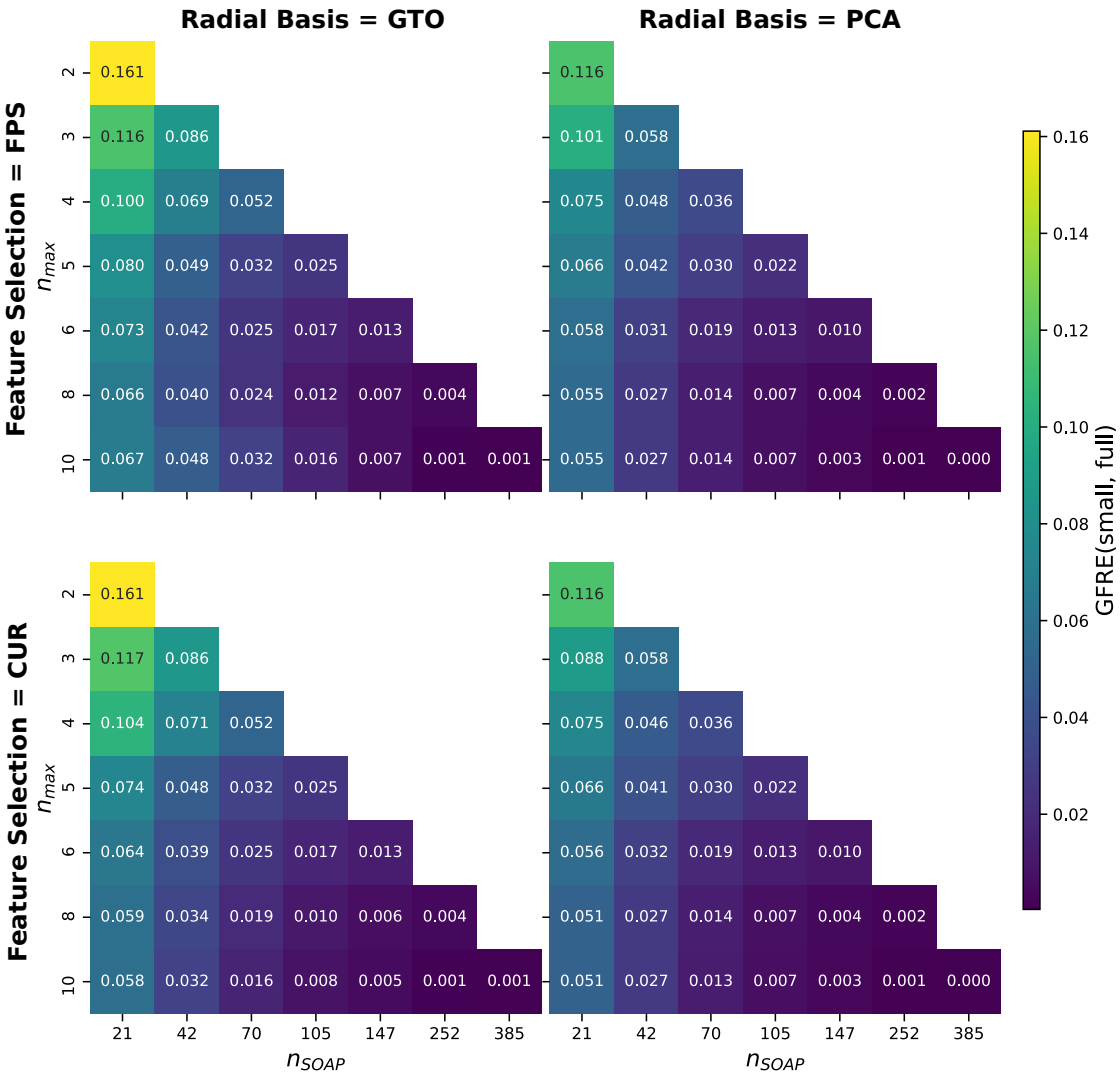

Figure 2.3: Feature space reconstruction errors for the power spectrum, resulting from the truncation of the radial basis and from the selection of a subset of the power spectrum entries using a deterministic CUR scheme and FPS. The "full" feature space is approximated with the power spectrum features, 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.

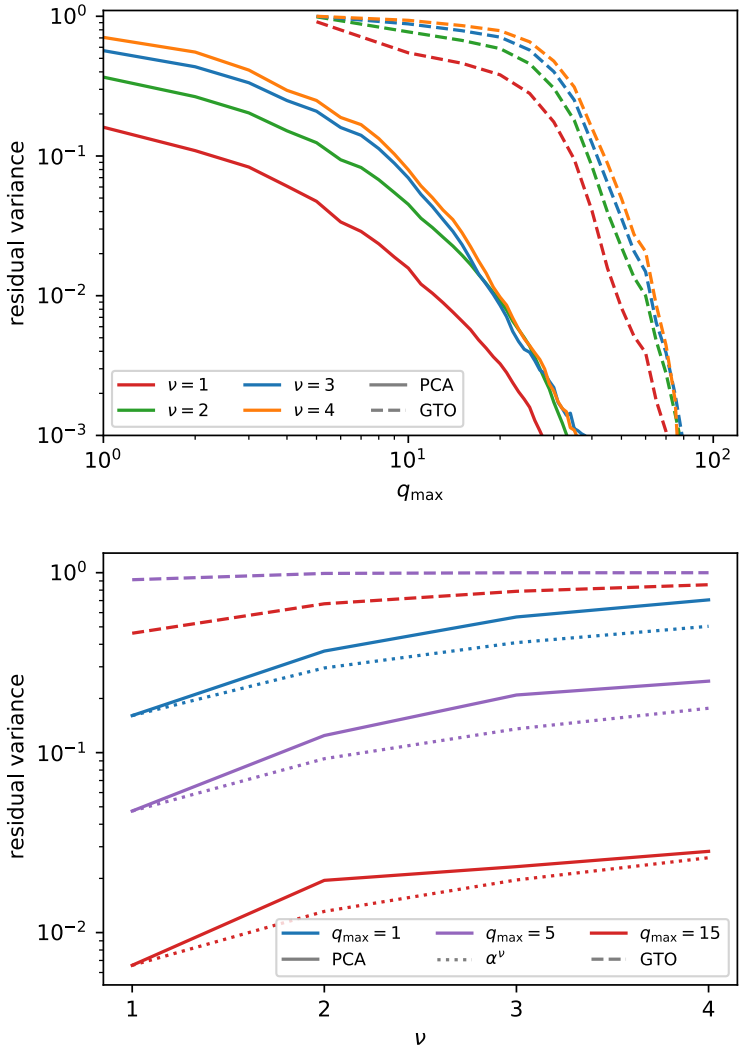

Figure 2.4: Residual variance for the multispectra computed for the QM9 dataset. For each body order, the baseline variance is taken to be that associated with the NICE features built starting from a "full" vector of density coefficients ($n_{\max} = 20, l_{\max} = 5$) – summing over the contributions 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.

comparison with Fig. 2.3, 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^i_n ($l = 0$). 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 Fig. 2.6) 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 Fig. 2.6 (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

$$c_{0_\gamma} \underset{\gamma \to 0}{=} \sum_n w_n c^i_n \approx y_i. \tag{2.13}$$

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 Fig. 2.6 (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 Fig. 2.7, 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 $q_{\text{max}} = 5$ (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^3_{max}, 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 $n_{\text{max}} = 20$ 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 Fig. 2.7. The plot compares the NICE models computed with $q_{\text{max}} = 50$ and an aggressive truncation of the body-order iteration, with the more balanced settings from Ref. 55 ($n_{\text{max}} = 12$, $l_{\text{max}} = 7$, $\nu_{\text{max}} = 5$, 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 $l_{\text{max}} = 5$ and $n_{\text{max}} = 5$) 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.

2.3 Future work

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[69] 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 correlation 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.

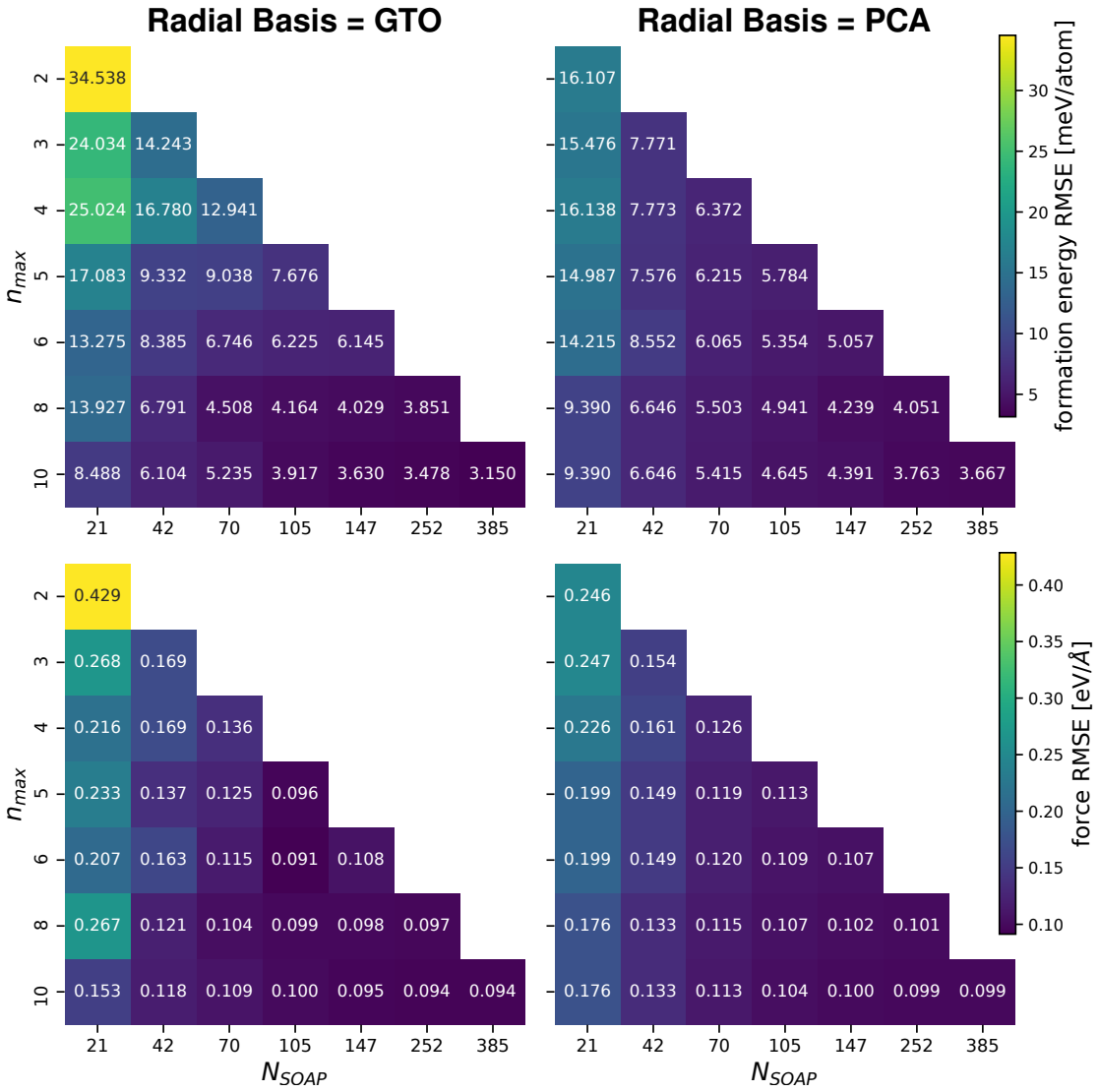

Figure 2.5: Energy and force RMSE for a Gaussian approximation potential based on the power spectrum, fitted to the Si dataset, plotted as a function of the number of radial functions $n_{\max}(q_{\max})$ and sparsification of the SOAP features, n_{SOAP} (using CUR selection).

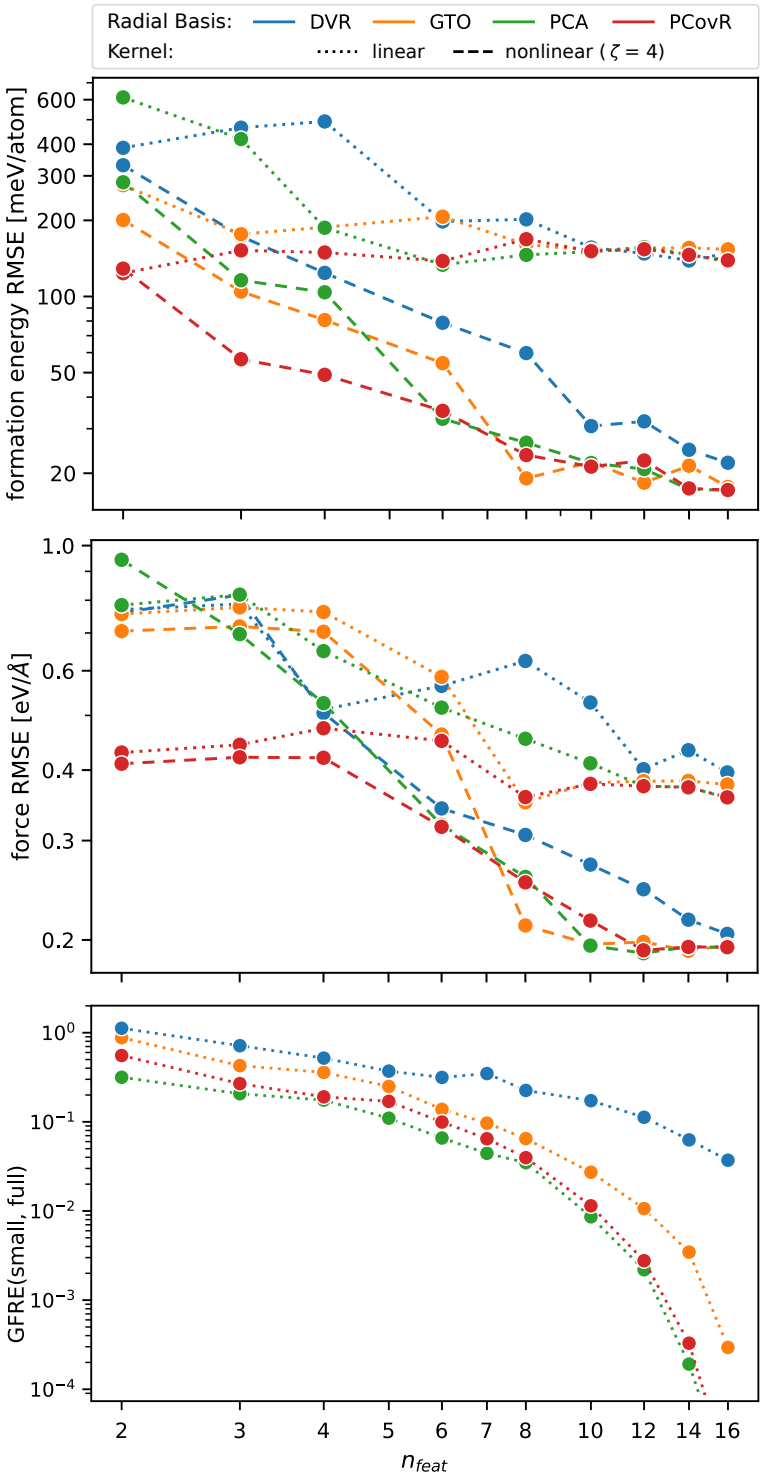

Figure 2.6: Energy (top) and force (center) 5-fold cross-validation RMSE and GFRE (bottom), computed on the silicon dataset for models based on the radial spectrum $|\overline{\rho_i^{\otimes 1}}\rangle$, as a function of the number of radial functions. Different curves correspond to a primitive DVR 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.

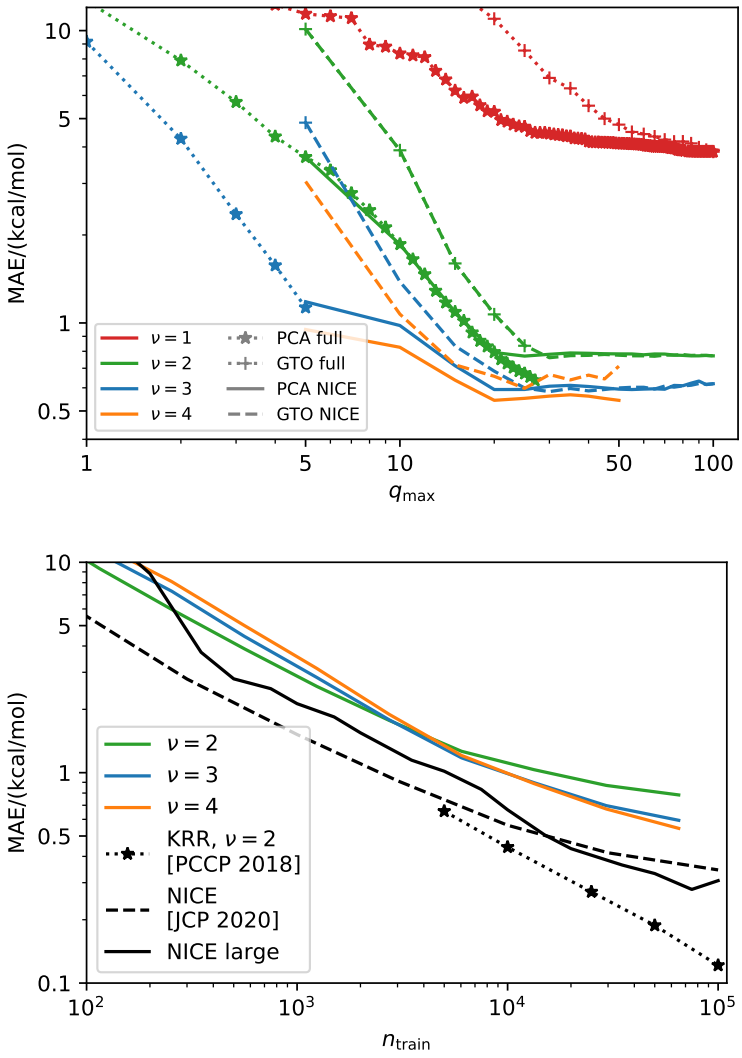

Figure 2.7: Convergence of ML models of the atomization energy of molecules from the QM9 dataset. (top) Convergence as a function of the (a, n) radial basis size, comparing a primitive GTO basis and an optimal PCA contraction, for different body orders of the features. For large q_{max} it is necessary 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. 55 (dashed line); a kernel model based on the power spectrum, using parameters analogous to those in Ref. 24 (dotted line).

Bibliography

[1] Renzo Tomellini, Johan Veiga Benesch, and Aud Alming. Commentary: Fostering innovation in materials sciences and engineering. *APL Materials*, 1(1):011001, 2013.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7):074106, 2011.

[10] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 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] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134(7), 2011.

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

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

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

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

[34] Jigyasa Nigam, Sergey Pozdnyakov, and Michele Ceriotti. Recursive evaluation and iterative contraction of n-body equivariant features. *The Journal of Chemical Physics*, 153(12), 2020.

[35] Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

[36] 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.

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

[53] 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.

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

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

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

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

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

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.