The Journal of Chemical Physics

Optimal radial basis for density-based atomic representations

Cite as: J. Chem. Phys. **155**, 104106 (2021); https://doi.org/10.1063/5.0057229
Submitted: 18 May 2021 • Accepted: 19 August 2021 • Published Online: 09 September 2021

Alexander Goscinski, Félix Musil, Sergey Pozdnyakov, et al.

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Optimal radial basis for density-based atomic representations

Cite as: J. Chem. Phys. 155, 104106 (2021); doi: 10.1063/5.0057229
Submitted: 18 May 2021 • Accepted: 19 August 2021 •
Published Online: 9 September 2021

Alexander Goscinski, Félix Musil, Sergey Pozdnyakov, Jigyasa Nigam, and Michele Ceriotti[a]

AFFILIATIONS

Laboratory of Computational Science and Modeling, Institute of Materials, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

[a]Author to whom correspondence should be addressed: michele.ceriotti@epfl.ch

ABSTRACT

The input of almost every machine learning algorithm targeting the properties of matter at the atomic scale involves a transformation of the list of Cartesian atomic coordinates into a more symmetric representation. Many of the most popular representations can be seen as an expansion of the symmetrized correlations of the atom density and differ mainly by the choice of basis. Considerable effort has been dedicated to the optimization of the basis set, typically driven by heuristic considerations on the behavior of the regression target. Here, we take a different, unsupervised viewpoint, aiming to determine the basis that encodes in the most compact way possible the structural information that is relevant for the dataset at hand. For each training dataset and number of basis functions, one can build a unique basis that is optimal in this sense and can be computed at no additional cost with respect to the primitive basis by approximating it with splines. We demonstrate that this construction yields representations that are accurate and computationally efficient, particularly when working with representations that correspond to high-body order correlations. We present examples that involve both molecular and condensed-phase machine-learning models.

I. INTRODUCTION

Machine-learning algorithms for atomistic simulations rely heavily on the transformation of structural information and chemical composition into descriptors or features.[1–3] An effective molecular representation should be invariant (or more generally, equivariant) with respect to symmetry operations,[3–9] capable of differentiating between inequivalent configurations,[10] and sensitive to atomic deformations.[11,12] In broad terms, it should encode in the most efficient way the relationships between a structure and the properties one is interested in predicting.[13] Even though many alternative approaches have been proposed to construct a representation that fulfills (at least partly) these requirements,[14] it has become clear that most of the existing schemes are strongly connected to each other and can be seen as projections on different choices of basis functions of the correlations of the atom density,[15,16] or equivalently of a cluster expansion of interactions.[17,18] Besides the importance of these considerations to determine the formal relation between different frameworks, the choice of basis function affects the prediction quality[19] and the efficiency of a basis in terms of linearly decodable mutual information.[20] Consequently, several algorithmic recipes for the construction of basis have been proposed[6,18,21,22] 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, which 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),[23] and that it can be practically, and inexpensively, evaluated as a numerical basis with striking similarities to ideas in electronic-structure methods.[24] Using an effective basis reduces the number of features that are needed to encode the same information, thereby reducing 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.

II. THEORY

We use the bra-ket notation originally introduced in Refs. 15 and 16 and discussed in detail in Ref. 14. An atomic structure A is

represented in terms of its atom density,

$$\langle a\mathbf{x}|A;\rho\rangle = \sum_i \delta_{aa_i}\langle \mathbf{x}|\mathbf{r}_i;g\rangle, \tag{1}$$

where $\langle \mathbf{x}|\mathbf{r}_i;g\rangle \equiv g(\mathbf{x}-\mathbf{r}_i)$ is a Gaussian of width σ_a centered on the position $\mathbf{r}_i$ of the ith atom and a_i is an index that indicates the chemical species of that atom. Translational symmetrization breaks this global atom density into a sum of atom-centered neighbor densities $\left|A;\left\langle\rho^{\otimes 2}\right\rangle_{\mathbb{R}^3}\right\rangle = \sum_i |A;\rho_i\rangle$,

$$\langle a\mathbf{x}|A;\rho_i\rangle = \sum_{j\in A}\delta_{aa_j}\langle \mathbf{x}|\mathbf{r}_{ji};g\rangle f_{\rm cut}(r_{ji}), \tag{2}$$

where $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$, and we introduce a smooth cutoff function $f_{\rm cut}$ to restrict the range of the environment.

It is convenient to express $\langle a\mathbf{x}|A;\rho_i\rangle$ on a basis of spherical harmonics $Y_l^m(\hat{\mathbf{x}}) \equiv \langle \hat{\mathbf{x}}|lm\rangle$ and radial functions $R_{nl}(x) \equiv \langle x|nl\rangle$,

$$\langle anlm|A;\rho_i\rangle = \int \mathrm{d}\mathbf{x}\,\langle nl|x\rangle\,\langle lm|\hat{\mathbf{x}}\rangle\,\langle a\mathbf{x}|A;\rho_i\rangle. \tag{3}$$

Regardless of the choice of $\langle x|nl\rangle$, one can evaluate the density coefficients as a sum over neighbors,

$$\begin{aligned}\langle anlm|\rho_i\rangle &= \sum_j \delta_{aa_j}\langle nlm|\mathbf{r}_{ji};g\rangle \\ &= \sum_j \delta_{aa_j}\langle nl|r_{ji};g\rangle\langle lm|\hat{\mathbf{r}}_{ji}\rangle,\end{aligned} \tag{4}$$

where $\langle nl|r;g\rangle$ is a radial integral,

$$\langle nl|r;g\rangle = 4\pi e^{-\frac{r^2}{2\sigma_a^2}}\int_0^\infty \mathrm{d}x\,x^2\langle nl|x\rangle e^{-\frac{x^2}{2\sigma_a^2}}\,\mathrm{i}_l\left(\frac{xr}{\sigma_a^2}\right), \tag{5}$$

that can be computed analytically for some choices of basis or approximated numerically and computed as a spline for each radial and angular channel pair.[22] The $\sigma_a \to 0$ limit corresponds to a δ-like density, which is used in alternative implementations of the density correlation features,[6,18,21] and can be evaluated as easily on any discrete basis. We discuss in the supplementary material some considerations on the practical evaluation of density coefficients.

A. Optimal density basis

Principal component analysis (PCA) has been used 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.[25] We propose to apply this procedure to the first-order equivariants—that correspond to 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 (see the supplementary material),

$$C^{al}_{nn'} = \frac{1}{N}\sum_i\sum_m \langle anlm|\rho_i\rangle\langle\rho_i|an'lm\rangle, \tag{6}$$

where the summation over m ensures that the covariance is independent of the orientation of structures in the dataset. 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,

$$\langle aqlm;\mathrm{opt}|\rho_i\rangle = \sum_n U^{al}_{qn}\langle anlm|\rho_i\rangle. \tag{7}$$

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 $q_{\rm max}$ 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 $n_{\rm max}$ coefficients, and this is how the idea was applied in Ref. 25. When combining Eq. (7) with Eq. (4), however, one sees that the contracted coefficients can be evaluated directly,

$$\langle aqlm;\mathrm{opt}|\rho_i\rangle = \sum_j \delta_{aa_j}\langle aql;\mathrm{opt}|r_{ji};g\rangle\langle lm|\hat{\mathbf{r}}_{ji}\rangle, \tag{8}$$

using the contracted radial integrals,

$$\langle aql;\mathrm{opt}|r;g\rangle = \sum_n U^{al}_{qn}\langle nl|r;g\rangle, \tag{9}$$

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 $q_{\rm max}$. The exact mathematical form and implementation details of the splines can be found in Ref. 22. Splining does not affect the equivariant behavior of the atom-density features and introduces minute discrepancies relative to the analytical basis, which do not affect the quality of the resulting models. Thus, the procedure we propose entails the following steps:

1. compute the density coefficients (4) for a representative dataset, using *any* primitive basis, and a large $n_{\rm max}$,
2. compute the covariance (6) and diagonalize it, finding the contraction coefficients U^{al}_{qn},
3. evaluate the contracted radial integrals using Eq. (9), over a dense radial grid,
4. use a spline approximation to evaluate directly the radial integrals (8) for the first $q_{\rm max}$ optimal features, and use the coefficients in subsequent ML steps.

Even though this framework only needs the contracted radial *integrals* (5), one can also compute and inspect the "optimal radial basis" that corresponds to the optimized coefficients,

$$\langle x|aql;\mathrm{opt}\rangle \equiv \sum_n U^{al}_{qn}\langle x|nl\rangle. \tag{10}$$

For a given dataset, these functions are optimal in the sense that when truncated to $q_{\rm max} < n_{\rm max}$, 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. III.

1. Mixed-species basis

Even though Eq. (6) is defined separately for different species a, it is also possible to compute cross correlations between different elemental channels, defining

$$C^{l}_{an;a'n'} = \frac{1}{N}\sum_i\sum_m \langle anlm|\rho_i\rangle\langle\rho_i|a'n'lm\rangle, \tag{11}$$

as performed in the NICE framework[25] following ideas proposed in Ref. 16, resulting in coefficients that combine information on multiple species,

$$\langle qlm;\mathrm{opt}|\rho_i\rangle = \sum_n U^{l}_{q;an}\langle anlm|\rho_i\rangle, \tag{12}$$

similar in spirit to the alchemical contraction discussed in Ref. 15. It is worth noting that although the NICE code[26] contains the infrastructure to compute these contractions *as a post-processing of the primitive basis*, the implementation we propose in librascal[27] 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,

$$\begin{aligned}\langle qlm;\mathrm{opt}|\rho_i\rangle &= \sum_j\sum_{an} U^{l}_{q;an}\delta_{aa_j}\langle nlm|\mathbf{r}_{ji};g\rangle\\ &= \sum_j\sum_{n} U^{l}_{q;a_jn}\langle nl|r_{ji};g\rangle\langle lm|\hat{\mathbf{r}}_{ji}\rangle\\ &= \sum_j\langle a_jql;\mathrm{opt}|r_{ji};g\rangle\langle lm|\hat{\mathbf{r}}_{ji}\rangle.\end{aligned} \tag{13}$$

Given, however, that the cost of evaluating the density coefficients is usually a small part of the calculation of density correlation features,[22,28] 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. Supervised basis set optimization

For a given number of radial functions, and a target dataset, the data-driven contracted basis (7) provides the most efficient description of the atom-centered 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 smooth overlap of atomic positions (SOAP) features tend to emphasize correlations between atoms that are far from the atomic center, which can lead to a counterintuitive degradation of the model accuracy with increasing cutoff radius.[15,29] This effect can be contrasted by introducing a radial scaling[13,15] 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[28]—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).[30,31] 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,

$$\langle r|aq0;\mathrm{opt};\gamma\rangle = \sum_n U^{a0;\gamma}_{qn}\langle r|n0\rangle, \tag{14}$$

which is a special case of Eq. (10) for $l=0$, where $U^{a0;\gamma}_{qn}$ is obtained as the orthogonalized PCovR projector, as discussed in Refs. 30 and 31, using a mixing parameter γ, which determines how strong the emphasis of the optimization should be on minimizing the residual variance or the error in regressing the target.

B. Density correlation features

In the vast majority of applications, the density coefficients are not used directly in applications but are combined to build higher order invariant or equivariant features.[3,8,16,18] For example, the power spectrum (i.e., SOAP invariant features[3]) can be computed as

$$\left\langle a_1n_1;a_2n_2;l|\overline{\rho_i^{\otimes 2}}\right\rangle \propto \frac{1}{\sqrt{2l+1}}\sum_m \langle a_1n_1lm|\rho_i\rangle\langle a_2n_2lm|\rho_i\rangle^{*}, \tag{15}$$

where the density coefficients can be either those obtained from primitive basis functions truncated at increasing $n_{\max}$ or those from an optimal basis containing $q_{\max}$ terms. For this work, we use primarily the orthogonalized GTO basis introduced in Ref. 22, which compares favorably in terms of information content[14,20] with a discrete variable representation (DVR) basis (a family of orthogonal polynomials), as well as with the alternative GTO basis used in DScribe[32] and the shifted-Gaussian basis of QUIP.[33]

We discuss the general case of "multispectra" in the frame of the N-body iterative construction of equivariant (NICE) features,[25] but analogous considerations apply to similar many-body descriptors, such as the atomic cluster expansion (ACE)[18,25] or the moment tensor potential (MTP),[6] and are likely to be relevant also for covariant neural networks.[9,34] We consider the case of a single chemical species to keep a notation that is by necessity quite cumbersome as simple as possible, but the generalization is trivial. The NICE iteration increases the body order of features that describe correlations between ν neighbors $\left\langle Q|\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu}\right\rangle$ (Q is a generic index that labels the features, and λ, μ, and σ are the indices that describe their behavior with respect to rotations and inversion) by combining lower order features,

$$\begin{aligned}\left\langle Q;nlk|\overline{\rho_i^{\otimes(\nu+1)};\sigma;\lambda\mu}\right\rangle \propto &\sum_m\left\langle n|\overline{\rho_i^{\otimes 1};lm}\right\rangle\\ &\times\left\langle Q|\overline{\rho_i^{\otimes\nu};(\sigma(-1)^{l+k+\lambda});k(\mu-m)}\right\rangle\\ &\times\langle lm;k(\mu-m)|\lambda\mu\rangle,\end{aligned} \tag{16}$$

using Clebsch–Gordan coefficients $\langle lm; l'm'|l''m''\rangle$ in an expression analogous to the sum of angular momenta. The $\nu = 1$ equivariants are nothing but the density coefficients,

$$\left\langle n|\overline{\rho_i^{\otimes 1}; \sigma; l\,m}\right\rangle = \delta_{\sigma 1}\langle nlm|\rho_i\rangle^\star, \tag{17}$$

and one can compute invariant descriptors by retaining only the $\left\langle Q|\overline{\rho_i^{\otimes \nu}; 1; 00}\right\rangle$ terms using the other components only as computational intermediates.

1. Change of basis for the multispectrum

First, we investigate the relation between the multispectra computed in an arbitrary radial basis and in the optimal basis obtained from the principal components of the density coefficients,

$$\begin{aligned}
&\left\langle Q; qlk; \text{opt}|\overline{\rho_i^{\otimes(\nu+1)}; \sigma; \lambda\mu}\right\rangle \\
&\quad \propto \sum_m \langle qlm; \text{opt}|\rho_i\rangle^\star \left\langle Q|\overline{\rho_i^{\otimes\nu}; \sigma((-1)^{l+k+\lambda}); k(\mu-m)}\right\rangle \\
&\quad\quad \times \langle lm; k(\mu-m)|\lambda\mu\rangle \\
&\quad = \sum_m \langle lm; k(\mu-m)|\lambda\mu\rangle \sum_n U^l_{qn}\langle nlm|\rho_i\rangle^\star \\
&\quad\quad \times \left\langle Q|\overline{\rho_i^{\otimes\nu}; \sigma((-1)^{l+k+\lambda}); k(\mu-m)}\right\rangle \\
&\quad = \sum_n U^l_{qn}\left\langle Q; nlk|\overline{\rho_i^{\otimes(\nu+1)}; \sigma; \lambda\mu}\right\rangle.
\end{aligned} \tag{18}$$

In other terms, the change of basis can be achieved by constructing the multispectrum using the density coefficients in the optimal radial basis or by applying the transformation to each (n_ν, l_ν) term in the multispectrum computed in the original basis. The transformation of the multispectrum is given by a block-diagonal matrix composed of $\mathbf{U}^l$.

2. Truncation of the multispectrum

Among the consequences of Eq. (18) is the fact that—if the optimal basis is not truncated so that $\mathbf{U}^l$ enacts an orthogonal transformation—the change to the optimal basis preserves the magnitude of the multispectrum,

$$\sum_{q=1}^{n_{\max}}\left|\left\langle Q; qlk; \text{opt}\left|\overline{\rho_i^{\otimes(\nu+1)}; \sigma; \lambda\mu}\right\rangle\right|^2 = \sum_{n=1}^{n_{\max}}\left|\left\langle Q; nlk\left|\overline{\rho_i^{\otimes(\nu+1)}; \sigma; \lambda\mu}\right\rangle\right|^2. \tag{19}$$

More generally, truncating the basis to include $q_{\max}$ optimized basis functions reduces the norm of the multispectrum by the same multiplicative factor at each iteration,

$$\begin{aligned}
&\sum_{q=1}^{q_{\max}}\sum_{lkQ}\sum_{\sigma\lambda\mu}\left|\left\langle Q; qlk; \text{opt}\left|\overline{\rho_i^{\otimes(\nu+1)}; \sigma; \lambda\mu}\right\rangle\right|^2 \\
&\quad = \sum_{q=1}^{q_{\max}}\sum_{lm}\left|\left\langle q\left|\overline{\rho_i^{\otimes 1}; l\,m}\right\rangle\right|^2 \times \sum_{Q\sigma kp}\left|\left\langle Q\left|\overline{\rho_i^{\otimes\nu}; \sigma; kp}\right\rangle\right|^2,
\end{aligned} \tag{20}$$

which can be derived exploiting the orthogonality of CG coefficients (see the supplementary material). One sees how (if the compound index Q was expanded to indicate the $q_\nu l_\nu k_\nu$ terms at each order ν) the norm of the multispectrum can be expanded into a product of terms coming from each order, and the errors introduced by truncation accumulate as a product. As a sidenote, the combination of Eqs. (19) and (20) implies that, for each environment, the norm of the ν-spectrum should equal the norm of the corresponding one-spectrum raised to the power ν when summing over all the equivariant components. This provides a stringent test to estimate the amount of information that is lost when contracting, subselecting, or truncating the angular momentum of the equivariant components during the iterative construction of high-body order features.

3. Principal component basis for multi-spectra

The derivation of (20) applies to each environment A_i separately and does not translate exactly into an expression for the retained variance (which involves an average over the training set). A similar issue arises when addressing the question of what is the best radial basis (again in terms of variance retained for a given level of truncation) that one can use to apply the NICE iteration for a specific feature Q and intermediate angular momentum state k. In building the covariance, we sum over (σ, λ, μ)—i.e., we look for a single transformation that applies to all terms that derive from combinations of $\left\langle Q|\overline{\rho_i^{\otimes\nu}; s; kp}\right\rangle$ with the density coefficients,

$$\begin{aligned}
NC^l_{nn'}(\nu; Q; k) &= \sum_{i\sigma\lambda\mu}\left\langle Q; nlk\left|\overline{\rho_i^{\otimes(\nu+1)}; \sigma; \lambda\mu}\right\rangle\right. \\
&\quad \times \left\langle\overline{\rho_i^{\otimes(\nu+1)}; \sigma; \lambda\mu}\right|\left. Q; n'lk\right\rangle \\
&= \sum_i\sum_m\left\langle n\left|\overline{\rho_i^{\otimes 1}; l\,m}\right\rangle\right\langle\overline{\rho_i^{\otimes 1}; l\,m}\left|n'\right\rangle\right. \\
&\quad \times \sum_{\sigma p}\left|\left\langle Q\left|\overline{\rho_i^{\otimes\nu}; \sigma; kp}\right\rangle\right|^2.
\end{aligned} \tag{21}$$

This expression corresponds to a covariance matrix of the density coefficients, which is built by weighting the contribution from each environment by the magnitude of $\left\langle Q|\overline{\rho_i^{\otimes\nu}; \sigma; kp}\right\rangle$. Thus, the optimal combinations that are determined for $\nu = 1$ are not necessarily equal to those needed in further iterations. Computing a different radial basis for each NICE iteration would be extremely cumbersome; in what follows, we provide evidence that the basis optimized for the density coefficients provides an effective compression even for the higher-order terms in the multispectrum.

III. RESULTS

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. 35, and the prediction of atomization energies for the organic molecules from the QM9 dataset.[36] These two examples are complementary: The silicon potential involves a single chemical species, uses forces for training, and aims to predict the properties of arbitrarily 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.[37]

A. 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,

$$RV = 1 - \frac{\sum_i \sum_{q=1}^{q_{\max}} |\langle qlm; \text{opt}|\rho_i\rangle|^2}{\sum_i \sum_{n=1}^{n_{\max}} |\langle nlm|\rho_i\rangle|^2}, \tag{22}$$

which 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 an increasing number of optimal basis functions, as shown in Fig. 1. The figure also shows the shape of the optimal radial functions and demonstrates that the same radial functions can be obtained starting from either 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 [even if the raw basis exhibits much larger spillover (see the supplementary material)], 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 specific) 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 2 shows the convergence of the residual 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, requiring almost 100 radial channels ($n_{\max} = 20$ for the five species present) to reduce the residual below 10^{-4}. The same level can be achieved with $q_{\max} \sim 50$ when

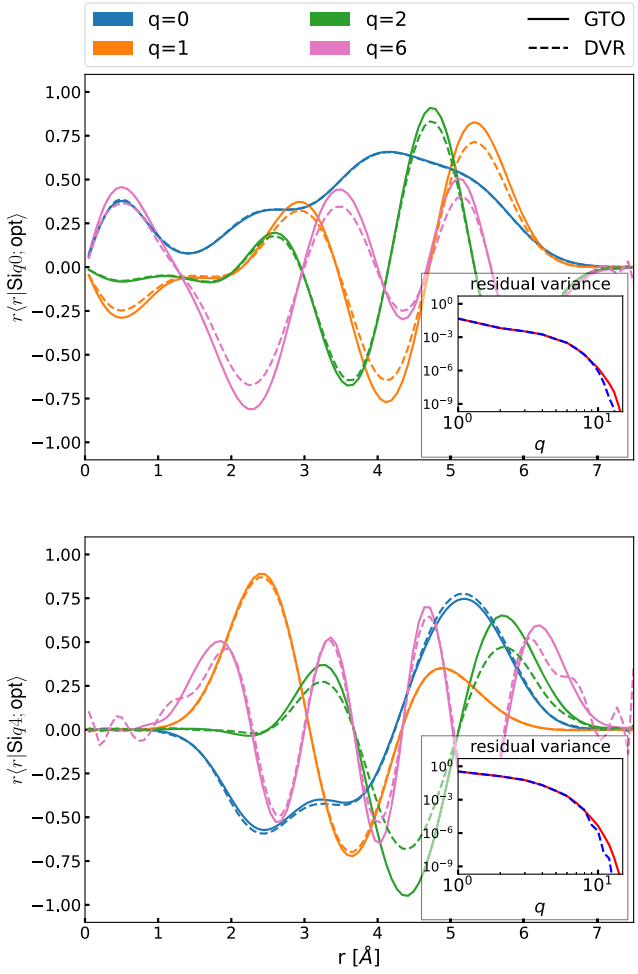

FIG. 1. Several examples of the optimized radial basis functions on the silicon dataset for $l = 0$ and $l = 4$ using DVR and GTO as the primitive basis contracted from $n_{\max} = 20$, with $r_{\text{cut}} = 6$.

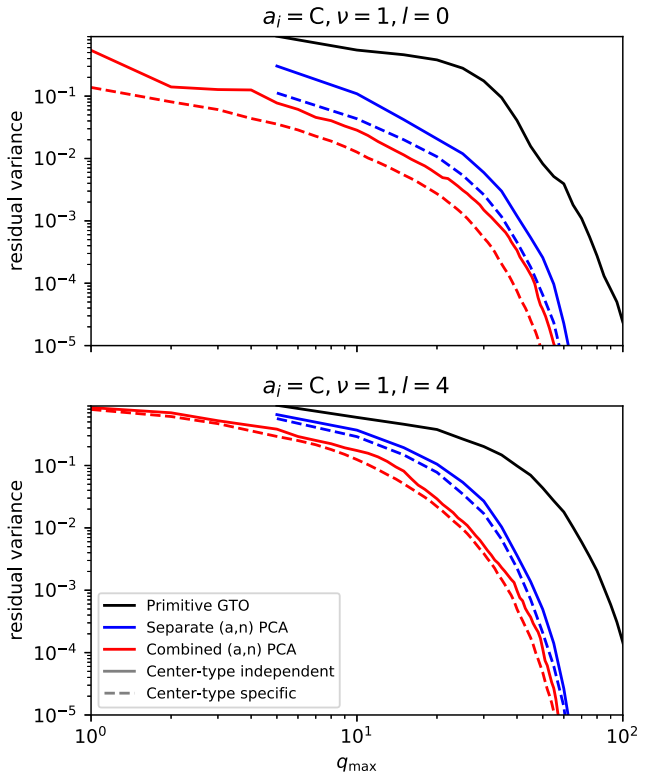

FIG. 2. Convergence of the residual variance for the expansion coefficients of the density as a function of the number of radial basis functions $q_{\max}$, computed for the QM9 dataset and for environments centered on a C atom. The different series correspond to a GTO basis of increasing size (black) and to an optimal basis computed for each neighbor density by separating (blue) or by mixing chemical and radial channels (a, n) (red). The full lines use the same basis irrespective of the species of the central atom, and the dashed lines correspond to a basis optimized specifically for C-centered environments.

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 residual variance.

B. Convergence of density correlation 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[20] (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[38] or farthest point sampling (FPS),[39,40] 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 (Fig. 3). This is true for both the full-sized SOAP vector and 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

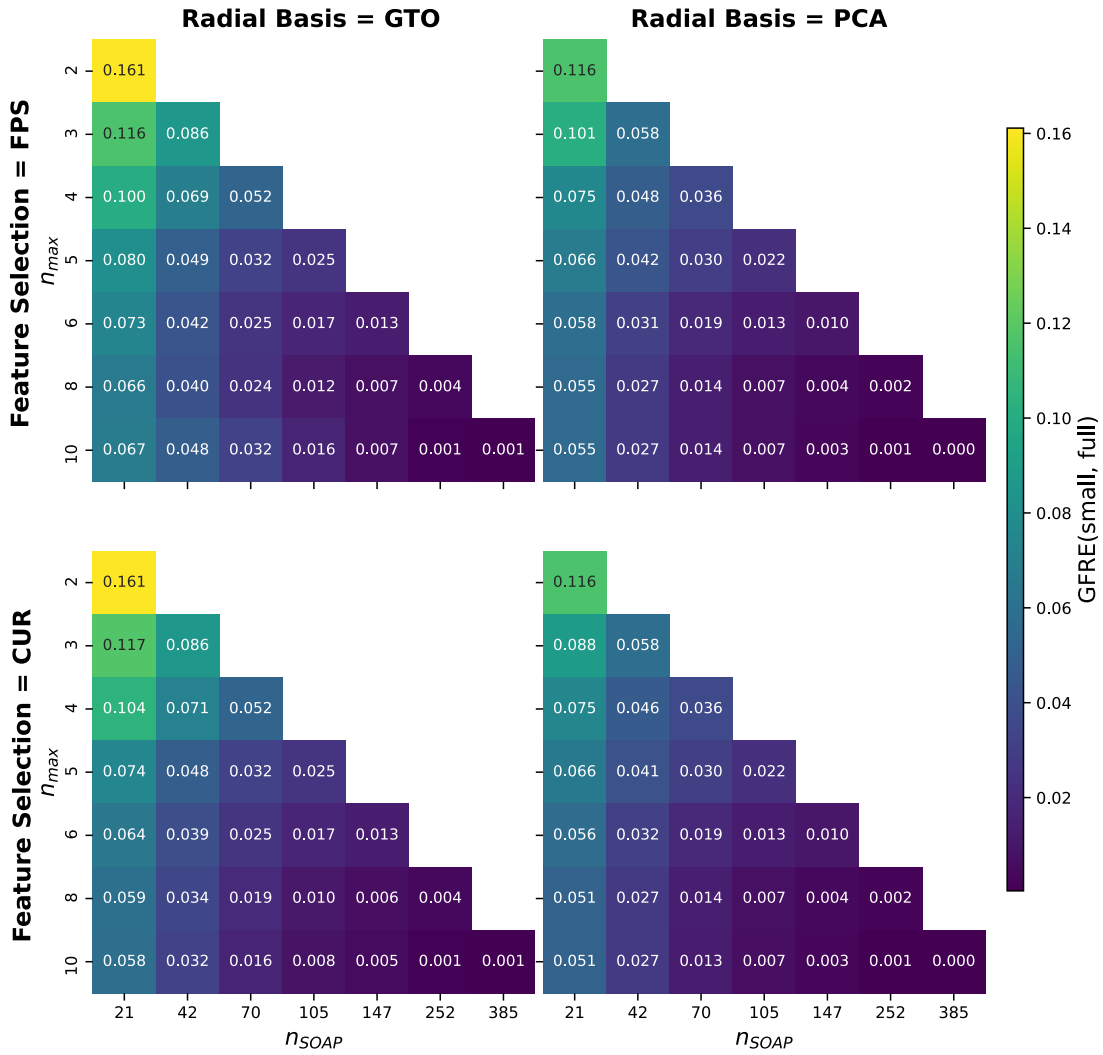

FIG. 3. Feature space reconstruction errors for the power spectrum, resulting from the truncation of the radial basis and from the selection of a subset of the power spectrum entries using a deterministic CUR scheme and FPS. The "full" feature space is approximated with the power spectrum features, computed using a GTO basis with $n_{max} = 20$, $l_{max} = 6$, and we compare the convergence obtained by using a smaller GTO basis against a truncated optimal basis of the same size.

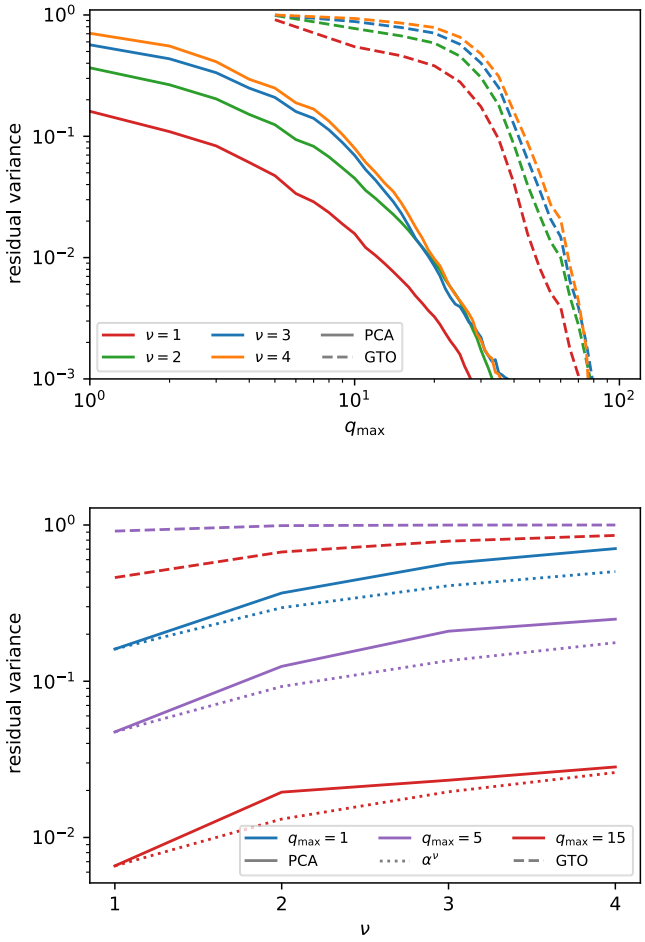

FIG. 4. Residual variance for the multispectra computed for the QM9 dataset. For each body order, the baseline variance is taken to be that associated with the NICE features built starting from a "full" vector of density coefficients ($n_{max} = 20, l_{max} = 5$)—summing over the contributions from all 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) The 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 the multispectrum. The dotted lines indicate the behavior one would expect if the retained variance followed exactly the multiplicative behavior given in Eq. (20).

models. The cost of different parts of the feature evaluation (density expansion, invariant calculation, kernel evaluation, gradients, and so on) depends subtly on the composition of the system and the various convergence parameters.[22] When evaluating a Gaussian process regression model, the calculation of the invariant features and 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 Sec. II B, 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 a much faster convergence of the bispectrum and the trispectrum compared to the primitive basis (Fig. 4, top panel). The truncation of the density coefficients affects the multispectra in a way that is qualitatively similar to what predicted by Eq. (20): The impact of an incomplete description of the density gets amplified by taking successive orders of correlation (Fig. 4, bottom panel). Given that the raw number of multispectrum components grows exponentially as q^ν_{max}, 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 in the complexity of high-body order models with growing chemical diversity.

C. Regression models

The accuracy of a Gaussian approximation potential based on SOAP features, trained using both energy and forces (details in the supplementary material), seen in Fig. 5 shows an improvement of the cross-validation error for the most aggressive truncation of the feature space (up to $n_{max} \approx 6$ for forces and $n_{max} \approx 4$ 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 Fig. 3, which 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 (see the supplementary material). Particularly for small n_{max} and for a relatively small train set size, 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. 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 two-body model that uses only the $\left\langle n|\overline{\rho_i^{\otimes 1}}\right\rangle \propto \langle n00|\rho_i\rangle$ features. The comparison between the GTO and the DVR basis (the former being vastly superior in terms of linearly decodable mutual information content,[20] as seen from the GFRE in the bottom panel of Fig. 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

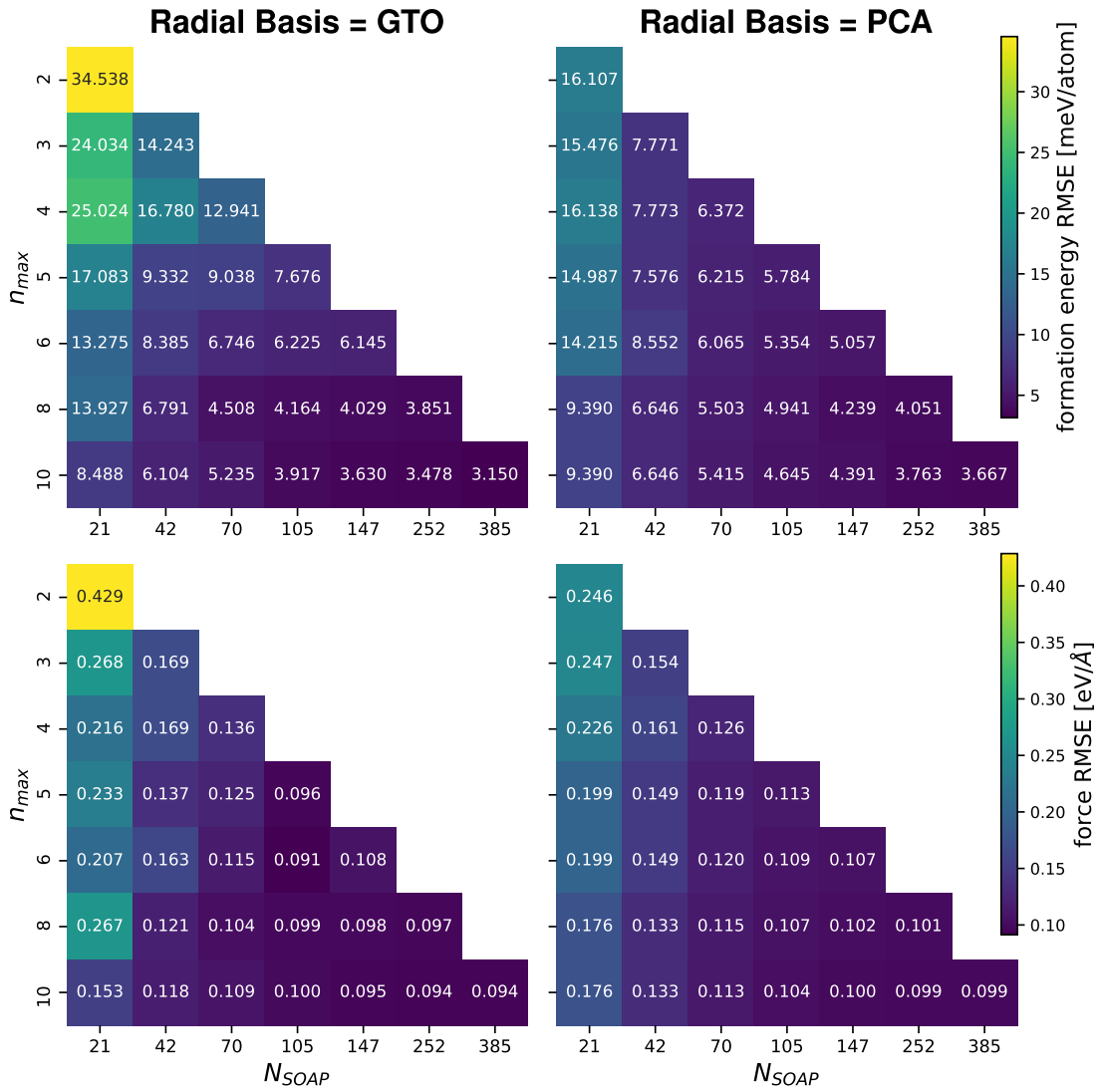

FIG. 5. Energy and force RMSE for a Gaussian approximation potential based on the power spectrum, fitted to the Si dataset, plotted as a function of the number of radial functions $n_{\max}(q_{\max})$ and sparsification of the SOAP features, n_{SOAP} (using CUR selection).

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 six 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 a supervised component to determine the contraction coefficients of the basis: As shown in Fig. 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 two-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,

$$w_0^{\mathrm{opt}}\left\langle q=0;\mathrm{opt};\gamma\to 0\middle|\overline{\rho_i^{\otimes 1}}\right\rangle \approx \sum_n w_n \left\langle n\middle|\overline{\rho_i^{\otimes 1}}\right\rangle = \tilde{y}(A_i). \qquad (23)$$

Thanks to the spline approximation of the optimal basis, $\left\langle 0;\mathrm{opt};\gamma\to 0\middle|\overline{\rho_i^{\otimes 1}}\right\rangle$ 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 root-mean-square error (RMSE) and $q_{\max} = 2$] against primitive and PCA-optimized bases, the advantage is quickly

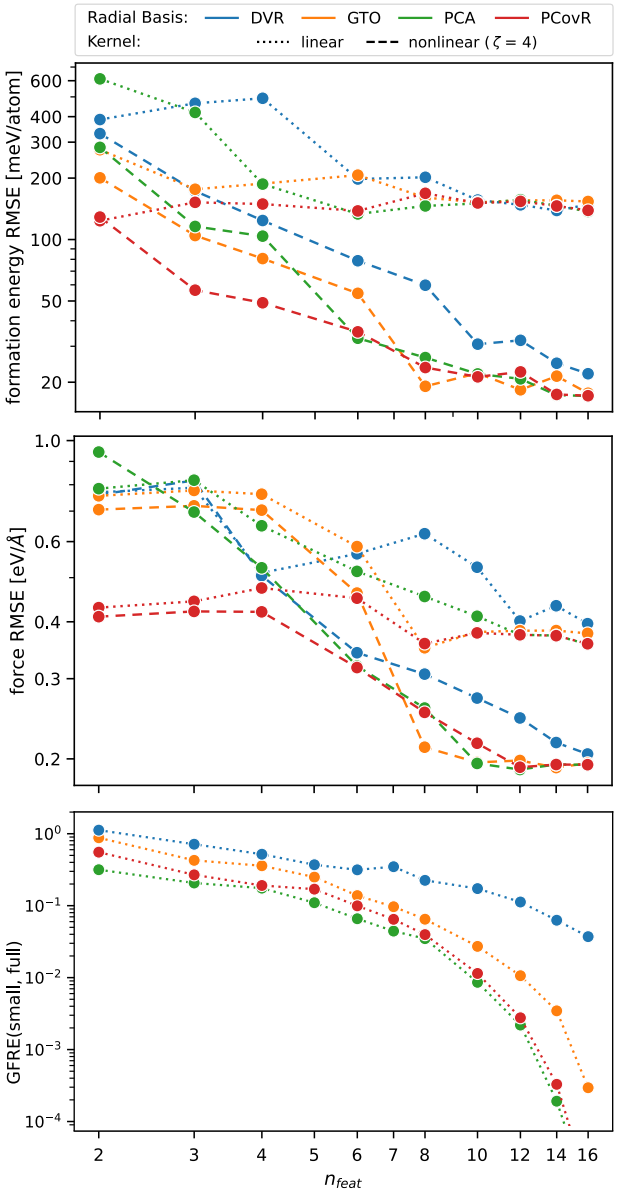

FIG. 6. Energy (top) and force (center) fivefold cross-validation RMSE and GFRE (bottom), computed on the silicon dataset for models based on the radial spectrum $\langle \rho_i^{\otimes 1} |$, as a function of the number of radial functions. Different curves correspond to a primitive DVR and GTO basis and to the optimal (PCA and PCovR) contracted bases. The PCovR contraction is performed with $\gamma = 0.1$. The full lines correspond to a linear model, and the dashed lines correspond to a polynomial kernel with exponent $\zeta = 4$. The GFRE is computed relative to a $n_{max} = 20$ GTO basis.

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. 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. 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 fivefold reduction in the test error compared to the primitive GTO basis and makes it possible to reach the symbolic

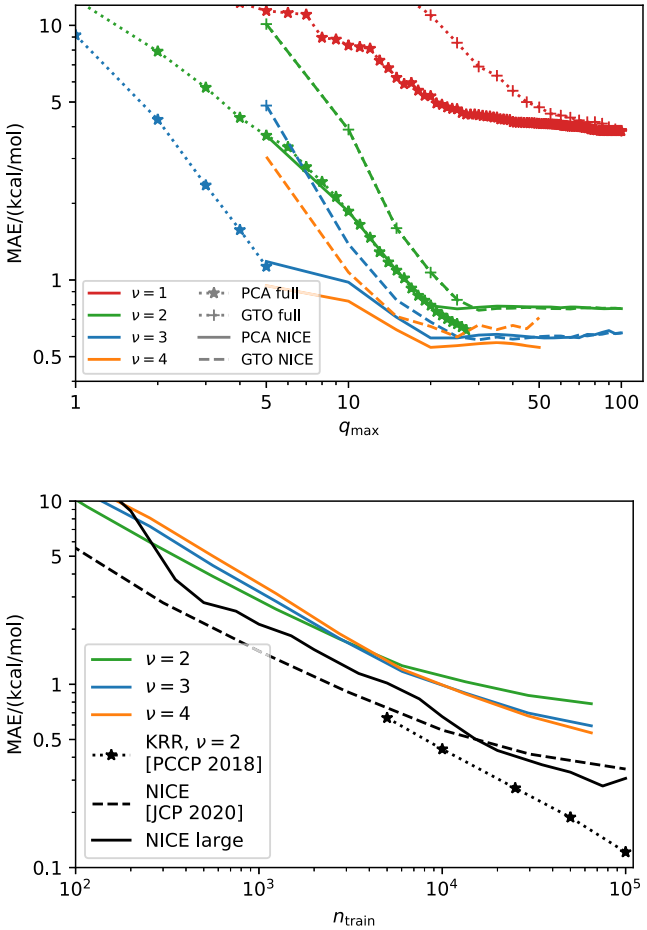

FIG. 7. Convergence of ML models of the atomization energy of molecules from the QM9 dataset. (Top) Convergence as a function of the (a, n) radial basis size, comparing a primitive GTO basis and an optimal PCA contraction, for different body orders of the features. For large q_{max}, it is necessary 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 1 TB of memory. (Bottom) The learning curves are obtained with linear models built on the PCA optimal features of increasing body order. All colored 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 53 390 features; the NICE model from Ref. 25 (dashed line); and a kernel model based on the power spectrum, using parameters analogous to those in Ref. 15 (dotted line).

threshold of 1 kcal/mol mean absolute error (MAE). In other terms, an optimal PCA contraction achieves an accuracy comparable to a primitive GTO basis, which is roughly two times larger. Given that the number of bispectrum ($\nu = 3$) feature 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 $n_{\max} = 20$ 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. 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. 25 ($n_{\max} = 12$, $l_{\max} = 7$, $\nu_{\max} = 5$, 1000 invariant features per body order), with a "large NICE" model that includes 53 880 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. 15 (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, which vary by a factor of 2. Furthermore, a nonlinear model based on low-body order features is the most accurate, reaching a state-of-the-art MAE of 0.12 kcal/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 high-body order that we discuss here shall support more systematic studies in the future.

IV. CONCLUSIONS

The realization 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, both analytically and 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[41] in the construction of the multispectra, 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 behavior as the targets, and so the $l > 0$ channels of the density coefficients cannot be optimized with a straightforward application of this scheme to the fitting of (scalar) potential.

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 3–5 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, and 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.

SUPPLEMENTARY MATERIAL

The supplementary material contains additional derivations and more detailed benchmarks of the methods discussed in the main text.

ACKNOWLEDGMENTS

F.M., J.N., and M.C. acknowledge the support by the NCCR MARVEL, funded by the Swiss National Science Foundation (SNSF). A.G., S.P., and M.C. acknowledge the support from the Swiss National Science Foundation (Project No. 200021-182057).

DATA AVAILABILITY

The data that support the findings of this study are openly available in the NICE libraries[26] and LIBRASCAL.[27]

REFERENCES

[1] J. Behler, "Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations," Phys. Chem. Chem. Phys. **13**, 17930–17955 (2011).

[2] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).

[3] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[4] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).

[5] B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," Int. Rev. Phys. Chem. **28**, 577–606 (2009).

[6] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," Multiscale Model. Simul. **14**, 1153–1173 (2016).

[7] A. Glielmo, P. Sollich, and A. De Vita, "Accurate interatomic force fields via machine learning with covariant kernels," Phys. Rev. B **95**, 214302 (2017).

[8] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, "Symmetry-adapted machine learning for tensorial properties of atomistic systems," Phys. Rev. Lett. **120**, 036002 (2018).

[9] B. Anderson, T. S. Hy, and R. Kondor, "Cormorant: Covariant molecular neural networks," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, 2019), p. 10.

[10] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Incompleteness of atomic structure representations," Phys. Rev. Lett. **125**, 166001 (2020).

[11] B. Onat, C. Ortner, and J. R. Kermode, "Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials," J. Chem. Phys. **153**, 144106 (2020).

[12] B. Parsaeifard, D. S. De, A. S. Christensen, F. A. Faber, E. Kocer, S. De, J. Behler, A. von Lilienfeld, and S. Goedecker, "An assessment of the structural resolution of various fingerprints commonly used in machine learning," Mach. Learn.: Sci. Technol. **2**, 015018 (2020).

[13] B. Huang and O. A. von Lilienfeld, "Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity," J. Chem. Phys. **145**, 161102 (2016).

[14] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Physics-inspired structural representations for molecules and materials," Chem. Rev. **121**, 9759–9815 (2021).

[15] M. J. Willatt, F. Musil, and M. Ceriotti, "Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements," Phys. Chem. Chem. Phys. **20**, 29661–29668 (2018).

[16] M. J. Willatt, F. Musil, and M. Ceriotti, "Atom-density representations for machine learning," J. Chem. Phys. **150**, 154110 (2019).

[17] J. M. Sanchez, F. Ducastelle, and D. Gratias, "Generalized cluster description of multicomponent systems," Physica A **128**, 334–350 (1984).

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

[19] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, and S. P. Ong, "Performance and cost assessment of machine learning interatomic potentials," J. Phys. Chem. A **124**, 731 (2020).

[20] A. Goscinski, G. Fraux, G. Imbalzano, and M. Ceriotti, "The role of feature space in atomistic learning," Mach. Learn.: Sci. Technol. **2**, 025028 (2021).

[21] M. Bachmayr, G. Csanyi, R. Drautz, G. Dusson, S. Etter, C. van der Oord, and C. Ortner, "Atomic cluster expansion: Completeness, efficiency and stability," arXiv:1911.03550 (2019).

[22] F. Musil, M. Veit, A. Goscinski, G. Fraux, M. J. Willatt, M. Stricker, and M. Ceriotti, "Efficient implementation of atom-density representations," J. Chem. Phys. **154**, 114109 (2021).

[23] A. Schäfer, H. Horn, and R. Ahlrichs, "Fully optimized contracted Gaussian basis sets for atoms Li to Kr," J. Chem. Phys. **97**, 2571–2577 (1992).

[24] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, "*Ab initio* molecular simulations with numeric atom-centered orbitals," Comput. Phys. Commun. **180**, 2175–2196 (2009).

[25] J. Nigam, S. Pozdnyakov, and M. Ceriotti, "Recursive evaluation and iterative contraction of *N*-body equivariant features," J. Chem. Phys. **153**, 121101 (2020).

[26] S. Pozdnyakov, NICE libraries, https://github.com/cosmo-epfl/nice, 2020.

[27] F. Musil, M. Veit, T. Junge, M. Stricker, A. Goscinki, G. Fraux, and M. Ceriotti, LIBRASCAL, https://github.com/cosmo-epfl/librascal.

[28] M. A. Caro, "Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials," Phys. Rev. B **100**, 024112 (2019).

[29] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules," Sci. Adv. **3**, e1701816 (2017).

[30] S. de Jong and H. A. L. Kiers, "Principal covariates regression," Chemom. Intell. Lab. Syst. **14**, 155–164 (1992).

[31] B. A. Helfrecht, R. K. Cersonsky, G. Fraux, and M. Ceriotti, "Structure-property maps with Kernel principal covariates regression," Mach. Learn.: Sci. Technol. **1**, 045021 (2020).

[32] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "DScribe: Library of descriptors for machine learning in materials science," Comput. Phys. Commun. **247**, 106949 (2020).

[33] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).

[34] B. K. Miller, M. Geiger, T. E. Smidt, and F. Noé, "Relevance of rotationally equivariant convolutions for predicting molecular properties," arXiv:2008.08461 (2020).

[35] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, "Machine learning a general-purpose interatomic potential for silicon," Phys. Rev. X **8**, 041048 (2018).

[36] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 140022 (2014).

[37] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, "Prediction errors of molecular machine learning models lower than hybrid DFT error," J. Chem. Theory Comput. **13**, 5255–5264 (2017).

[38] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials," J. Chem. Phys. **148**, 241730 (2018).

[39] Y. Eldar, M. Lindenbaum, M. Porat, and Y. Y. Zeevi, "The farthest point strategy for progressive image sampling," IEEE Trans. Image Process. **6**, 1305–1315 (1997).

[40] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Demonstrating the transferability and the descriptive power of sketch-map," J. Chem. Theory Comput. **9**, 1521–1532 (2013).

[41] R. K. Cersonsky, B. A. Helfrecht, E. A. Engel, S. Kliavinek, and M. Ceriotti, "Improving sample and feature selection with principal covariates regression," Mach. Learn.: Sci. Technol. **2**, 035038 (2021).