Permutationally Invariant, Reproducing Kernel-Based Potential Energy Surfaces for Polyatomic Molecules: From Formaldehyde to Acetone

Debasish Koner and Markus Meuwly*

Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland

E-mail: m.meuwly@unibas.ch

Abstract

Constructing accurate, high dimensional molecular potential energy surfaces (PESs) for polyatomic molecules is challenging. Reproducing Kernel Hilbert space (RKHS) interpolation is an efficient way to construct such PESs. However, the scheme is most effective when the input energies are available on a regular grid. Thus the number of reference energies required can become very large even for penta-atomic systems making such an approach computationally prohibitive when using high-level electronic structure calculations. Here an efficient and robust scheme is presented to overcome these limitations and is applied to constructing high dimensional PESs for systems with up to 10 atoms. Using energies as well as gradients reduces the number of input data required and thus keeps the number of coefficients at a manageable size. Correct implementation of permutational symmetry in the kernel products is tested and explicitly demonstrated for the highly symmetric CH_4 molecule.

Introduction

The dynamics of molecular system is entirely governed by the underlying potential energy surface (PES) which describes the inter- and intramolecular interactions. Often, such PESs are computed from reference data based on electronic structure calculations using both, regular or more random coordinate grids. As the study of the dynamics of molecular systems requires energies and gradients, determining them '*on the fly*' (i.e. *ab initio* molecular dynamics) can be computationally prohibitive, in particular when high-level methods such as second order Møller-Plesset (MP2), multi reference configuration interaction (MRCI), or coupled cluster with singles, doubles, and perturbative triples (CCSD(T)) are used together with large basis sets. Therefore, constructing an analytical representation of the ab initio PES is a meaningful and advantageous alternative to accurately and efficiently describe intramolecular interactions.

Developing accurate and computationally and data-efficient representations of potential energies for multidimensional systems is a challenging task. There are several approaches to describe the energetics of a molecular PES: (i) fitting functional forms based on a single or double many body expansion[1] such as the London-Eyring-Polanyi-Sato (LEPS)[2] or Aguado-Paniagua (AP) surfaces,[3] (ii) permutationally invariant polynomials (PIPs),[4] (iii) interpolation by cubic splines,[5] or modified Shepard interpolation,[6,7] (iv) kernel based methods including reproducing kernel Hilbert space (RKHS),[8,9] Gaussian progress (GP) regression,[10] or (v) Neural network (NN) based representations.[11,12] The popular functional terms (e.g. LEPS, AP) based on many body expansions can provide accurate and computationally efficient representations for tri- and tetra-atomic systems.[13–15] However, using them for polyatomic systems is quite challenging as the many body expansion becomes more complicated. Interpolation methods are computationally expensive for multidimensional PESs whereas PIP, GP, and NN approaches can be applied efficiently to construct high-dimensional PESs.[12,16,17]

RKHS interpolation has been shown to provide highly accurate PESs for spectroscopic applications[18] and reaction dynamics[19] as well as for molecular dynamics (MD) simulations. For small molecules (diatomic and triatomic)[19–24] this method is advantageous over other methods as it reproduces the precalculated on-grid energies 'exactly', captures the long range interactions correctly if appropriate kernel polynomials are chosen and results in smooth PESs with continuous gradients.[25,26] For a single energy evaluation for an unknown molecular structure the RKHS method needs to sum over all training samples.[8] However, if the *ab initio* energies for training structures are provided on a regular grid, the kernel functions can be decomposed into only two to five terms which is much smaller than the training set size.[27] The sum then runs over these few terms which can be precomputed and stored in a look up table. Hence, with this *fast* RKHS approach the computational cost scales almost linearly with the number of data points[9,27] and very accurate PESs can be constructed for systems using a dense grid. The fast-evaluation method was later modified to use partially filled grids with similar efficiency.[28]

It has been shown that within a high dimensional model representation (HDMR), RKHS can be used to construct PESs. RKHS-HDMR works beyond conventional tensor-product constructs and with successive multilevel decomposition procedures which reduces multidimensional interpolation to independent low dimensional interpolation.[29] This approach can also be used for non-rectangular grids. An application of the RKHS-HDMR approach to a low-dimensional (3d) system has been reported for CH_2 as an example.[29] In a more recent study, the RKHS-HDMR approach has been tested for the ten dimensional Friedman target function but not for a PES.[30] However, the use of RKHS for all degrees of freedoms in constructing PESs for larger (i.e. four or more atoms) molecular systems is scarce in the literature. Rather, a RKHS representation is used for selected degrees of freedom, e.g. the van der Waals separation (R) whereas analytical expressions are employed for the remaining degrees of freedom as was done for tetra- and penta-atomic van der Waals complexes/molecules e.g., OH–HCl,[31] HCN-HCl[32] and NH_3–He.[33]

One of the main difficulties in using grid-based interpolation methods is their unfavourable scaling with increasing dimensionality of the problem. Although the fast RKHS approach[27] allows for near-independent data set size construction and evaluation of a RKHS, the requirement of a rectangular grid-based reference data set structure makes this approach highly computationally expensive in terms of storage memory and number of operations. Even with a partially filled grid the fast RKHS implementation scales as 2^M where M is the number of dimensions/degrees of freedom, which makes it unmanageable for more than four atom species. Sampling the configuration space more densely near the stationary structures, e.g. around minima and saddle points, can significantly reduce the number of input energies.[28] But in practice using only a small number of structures and energies leads to uneven RKHS PESs with discontinuous gradients. On the other hand, including gradient information for a configuration provides information about the likely behavior of the PES in surrounding regions which is encoded in the coefficients or parameters of an analytical PES. Hence, the analytical PES provides a smooth behavior in the neighbourhood of a training grid point even if only fewer numbers of configurations are used for training.

It has been shown for permutationally invariant polynomials (PIPs) applied to CH_4 that by using gradients along with energies in the input data set, smooth and accurate PESs can be obtained using fewer input data.[16] From energy and gradient information for only 100 configurations, randomly selected from an *ab initio* molecular dynamics (AIMD) simulation, a PIP-based PES was constructed with root mean square errors of 8.8 cm^{-1} and 39.8 cm^{-1}/a_0 for energy and gradients, respectively. The harmonic frequencies from the normal mode analysis using those PIP PESs were within 1 cm^{-1} compared with the *ab initio* frequencies. Subsequently, this approach was applied to N-methyl acetamide (NMA) to construct PESs for *trans*-NMA[34] and a full dimensional PES for NMA[35] with a root mean squared fitting error ranging from 26.8 cm^{-1} for full PIP and 148.9 cm^{-1} when a fragment-based approach was used whereby the energies used in the fitting covered a range up to ~ 3.5 eV.

Here, we introduce an efficient and robust approach to represent highly accurate PESs for molecules with four to ten atoms using RKHS interpolation with reciprocal power decay kernels. Gradients are used along with the energies to determine the coefficients for the tensor product form of the kernels. The formulation is applied to systems ranging from formaldehyde (CH_2O, 4 atoms) to acetone (CH_3COCH_3, 10 atoms). Molecular symmetry is included explicitly in the tensor product expansion of the kernel polynomials and is demonstrated to yield accurate RKHS-based results for the highly symmetric CH_4 molecule. First, the methodological developments are discussed. Next, RKHS-based PESs are determined for illustrative examples and the harmonic frequencies are determined as a validation of the methods. Finally, conclusions are drawn.

Methods

RKHS with Energies and Gradients

Within the RKHS formalism[36] potential energies for a system can be expressed as a linear combination of reproducing kernel functions using a set of known energies $V(\mathbf{x})$ at different configurations $\mathbf{x}$. The representer theorem[37] for a general functional relationship $y = f(\mathbf{x})$ states that $f(\mathbf{x})$ can always be approximated as a linear combination of suitable functions

$$f(\mathbf{x}) \approx \widetilde{f}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x}_i) \tag{1}$$

where α_i are coefficients and $K(\mathbf{x}, \mathbf{x}')$ is a kernel function. The reproducing property asserts that $f(x') = \langle f(x), K(x, x') \rangle$ where $\langle \cdot \rangle$ is the scalar product and $K(x, x')$ is the kernel.[36]

Popular choices for $K(\mathbf{x}, \mathbf{x}')$ for representing PESs are polynomial kernels

$$K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle^d \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product and d is the degree of the polynomial. It is also possible to include knowledge about the long range behaviour of the physical interactions into the kernel function itself.[25,38]

The coefficients α_i (Eq. 1) can be determined such that $\widetilde{f}(\mathbf{x}_i) = y_i$ for all input $\mathbf{x}_i$ in the dataset, i.e.

$$\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y} \tag{3}$$

where $\boldsymbol{\alpha} = [\alpha_i \cdots \alpha_N]^{\mathrm{T}}$ is the vector of coefficients, $\mathbf{K}$ is an $N \times N$ matrix with entries $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ called kernel matrix[39,40] and $\mathbf{y} = [y_1 \cdots y_N]^{\mathrm{T}}$ is a vector containing the N observations y_i in the data set. Since the kernel matrix is symmetric and positive-definite by construction, Cholesky decomposition[41] can be used to efficiently solve Eq. 3. Once the coefficients α_i have been determined, unknown values y_* at arbitrary positions $\mathbf{x}_*$ can be estimated as $y_* = \widetilde{f}(\mathbf{x}_*)$ using Eq. 1.

In practice the solution of Eq. 3 is only possible if the kernel matrix $\mathbf{K}$ is not ill-conditioned. Fortunately, even if $\mathbf{K}$ is ill-conditioned, an approximate (regularized) solution can be obtained for example by Tikhonov regularization.[42] This amounts to adding a small positive constant λ to the diagonal of $\mathbf{K}$, such that

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\,\mathbf{y} \tag{4}$$

is solved instead of Eq. 3 when determining the coefficients α_i (here, $\mathbf{I}$ is the identity matrix). Adding $\lambda > 0$ to the diagonal of $\mathbf{K}$ damps the magnitude of the coefficients $\boldsymbol{\alpha}$ and increases the smoothness of $\widetilde{f}$. While this has the effect that the known values in the data set are only *approximately* reproduced by Eq. 1, i.e. strictly $\widetilde{f}(\mathbf{x}_i) \neq y_i$, perhaps counterintuitively, it can *increase* the overall quality of predictions for unknown $\mathbf{x}_*$: In cases where the values y_i are noisy, reproducing them exactly also reproduces the noise, which is unlikely to generalise well to unknown data. Therefore, this method of determining the coefficients can also be used to prevent over-fitting and is known as kernel ridge regression (KRR).

When applied to represent discrete data for energies, the PES can be written as

$$V(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x'_i}) \tag{5}$$

where α_i are coefficients and $K(\mathbf{x}, \mathbf{x}')$ is the reproducing kernel and $\mathbf{x'_i}$ represents the training set which are the geometries for which energies have been determined from electronic structure calculations. The coefficients are then determined from the known ab initio energies for N configurations by solving the linear equations

$$\begin{pmatrix} K(\mathbf{x}_1, \mathbf{x'}_1) & K(\mathbf{x}_1, \mathbf{x'}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x'}_N) \\ K(\mathbf{x}_2, \mathbf{x'}_1) & K(\mathbf{x}_2, \mathbf{x'}_2) & \cdots & K(\mathbf{x}_2, \mathbf{x'}_N) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x}_N, \mathbf{x'}_1) & K(\mathbf{x}_N, \mathbf{x'}_2) & \cdots & K(\mathbf{x}_N, \mathbf{x'}_N) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{pmatrix} \tag{6}$$

This procedure gives an exact solution on the grid points $\mathbf{x'_i}$. The explicit matrix form (Eq. 6) for Eq. 1 is given to clarify how the structure of $K(\mathbf{x}, \mathbf{x_i})$ changes once gradients are included in constructing the RKHS (see below).

For an M-dimensional problem, the multi-dimensional kernel can be written as a direct product

$$K(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{M} k_j(x, x') \tag{7}$$

where $k_j(x,x')$ are 1D kernels. Multidimensional reproducing kernels can therefore be used to represent the p-body interaction energies of a system.

Within a many body expansion, the total potential energy of a system can be decomposed into a sum of p-body interactions $V^{(p)}$. For a molecule with n atoms, each p-body term consists of ${}^nC_p$ p-body interactions, where ${}^nC_p$ is the binomial coefficient. The total potential for an n-atomic species is therefore

$$V = \sum_{p=1}^{n} \sum_{i=1}^{{}^nC_p} V_i^{(p)} \tag{8}$$

In practice Eq. 8 is truncated at $p = 3$ or 4, i.e. contributions up to three- and 4-body terms are included which is what is also done in the present work.

One dimensional, reciprocal power reproducing kernels have been shown to describe diatomic potentials with high accuracy on the interval $[0, \infty]$.[8,25] The general expression for a $k^{[n,m]}$ reproducing polynomial kernel is

$$k^{[n,m]} = n^2 x_>^{-(m+1)} B(m+1, n)\,{}_2F_1\left(-n+1, m+1; n+m+1; \frac{x_<}{x_>}\right) \tag{9}$$

where, n and m are the smoothness and asymptotic reciprocal power parameters, whereas $x_<$ and $x_>$ are the smaller and larger value of x, respectively. $B(a,b)$ in Eq. 9 is the beta function $B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$ and ${}_2F_1(a,b;c;z)$ is Gauss' hypergeometric function.[8] These kernel polynomials can also be used to construct an M-dimensional reproducing kernels as a function of radial dimensions by direct product relations. In the present study, each term of p-body interaction energy is represented as an M-dimensional ($M = {}^pC_2$) reproducing kernel constructed from M reciprocal power kernels for M interatomic distances r_j. The full kernel is then

$$K(\mathbf{r},\mathbf{r}') = \sum_{p=1}^{n} \sum_{l=1}^{{}^nC_p} \prod_{j=1}^{{}^pC_2} k_j(r_j, r'_j) \tag{10}$$

and

$$V(\mathbf{r}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{r},\mathbf{r}') \tag{11}$$

Here, $\mathbf{r}$ is a vector containing all pairwise interatomic distances of an n-atomic system, $\mathbf{r} = \{r_h | h = 1, 2, 3 \cdots, ^n C_2\}$. In this study different reciprocal power kernels were tested, and it is found that $k^{[3,5]}$, $k^{[3,1]}$ and $k^{[3,0]}$ kernels perform best to construct mono/multidimensional kernels for 2-, 3-, and 4-body interaction energies, respectively.

Derivatives of the potential with respect to the distance coordinates can be calculated by simply replacing the reproducing kernels $K(\mathbf{r},\mathbf{r}')$ by their derivatives $K'(\mathbf{r},\mathbf{r}')$. Then the gradients of the total potential with respect to a Cartesian coordinates x_i are

$$\frac{dV}{dx_i} = \sum_{h=1}^{{}^nC_2} \frac{dV}{dr_h} \frac{dr_h}{dx_i} \tag{12}$$

and

$$\frac{dV}{dr_h} = \sum_{i=1}^{N} C_i K'(\mathbf{r},\mathbf{r}') \tag{13}$$

If the PES is faithfully represented by the RKHS, its derivative is also a good approximation of the gradients.

In a next step, the gradients - which are also available from the electronic structure calculations - are included in Eq. 6 which yields

$$\begin{pmatrix}
K(\mathbf{x}_1,\mathbf{x'}_1) & K(\mathbf{x}_1,\mathbf{x'}_2) & \cdots & K(\mathbf{x}_1,\mathbf{x'}_N) \\
K'_{x1}(\mathbf{x}_1,\mathbf{x'}_1) & K'_{x1}(\mathbf{x}_1,\mathbf{x'}_2) & \cdots & K'_{x1}(\mathbf{x}_1,\mathbf{x'}_N) \\
K'_{y1}(\mathbf{x}_1,\mathbf{x'}_1) & K'_{y1}(\mathbf{x}_1,\mathbf{x'}_2) & \cdots & K'_{y1}(\mathbf{x}_1,\mathbf{x'}_N) \\
K'_{z1}(\mathbf{x}_1,\mathbf{x'}_1) & K'_{z1}(\mathbf{x}_1,\mathbf{x'}_2) & \cdots & K'_{z1}(\mathbf{x}_1,\mathbf{x'}_N) \\
\vdots & \vdots & \ddots & \vdots \\
K'_{xn}(\mathbf{x}_1,\mathbf{x'}_1) & K'_{xn}(\mathbf{x}_1,\mathbf{x'}_2) & \cdots & K'_{xn}(\mathbf{x}_1,\mathbf{x'}_N) \\
K'_{yn}(\mathbf{x}_1,\mathbf{x'}_1) & K'_{yn}(\mathbf{x}_1,\mathbf{x'}_2) & \cdots & K'_{yn}(\mathbf{x}_1,\mathbf{x'}_N) \\
K'_{zn}(\mathbf{x}_1,\mathbf{x'}_1) & K'_{zn}(\mathbf{x}_1,\mathbf{x'}_2) & \cdots & K'_{zn}(\mathbf{x}_1,\mathbf{x'}_N) \\
\vdots & \vdots & \ddots & \vdots \\
K(\mathbf{x}_N,\mathbf{x'}_1) & K(\mathbf{x}_N,\mathbf{x'}_2) & \cdots & K(\mathbf{x}_N,\mathbf{x'}_N) \\
K'_{x}(\mathbf{x}_N,\mathbf{x'}_1) & K'_{x}(\mathbf{x}_N,\mathbf{x'}_2) & \cdots & K'_{x}(\mathbf{x}_N,\mathbf{x'}_N) \\
K'_{y}(\mathbf{x}_N,\mathbf{x'}_1) & K'_{y}(\mathbf{x}_N,\mathbf{x'}_2) & \cdots & K'_{y}(\mathbf{x}_N,\mathbf{x'}_N) \\
K'_{z}(\mathbf{x}_N,\mathbf{x'}_1) & K'_{z}(\mathbf{x}_N,\mathbf{x'}_2) & \cdots & K'_{z}(\mathbf{x}_N,\mathbf{x'}_N) \\
\vdots & \vdots & \ddots & \vdots \\
K'_{xn}(\mathbf{x}_N,\mathbf{x'}_1) & K'_{xn}(\mathbf{x}_N,\mathbf{x'}_2) & \cdots & K'_{xn}(\mathbf{x}_N,\mathbf{x'}_N) \\
K'_{yn}(\mathbf{x}_N,\mathbf{x'}_1) & K'_{yn}(\mathbf{x}_N,\mathbf{x'}_2) & \cdots & K'_{yn}(\mathbf{x}_N,\mathbf{x'}_N) \\
K'_{zn}(\mathbf{x}_N,\mathbf{x'}_1) & K'_{zn}(\mathbf{x}_N,\mathbf{x'}_2) & \cdots & K'_{zn}(\mathbf{x}_N,\mathbf{x'}_N)
\end{pmatrix}
\begin{pmatrix}
\alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N
\end{pmatrix}
=
\begin{pmatrix}
V_1 \\ dV_1/dx1 \\ dV_1/dy1 \\ dV_1/dz1 \\ \vdots \\ dV_1/dxn \\ dV_1/dyn \\ dV_1/dzn \\ \vdots \\ V_N \\ dV_N/dx1 \\ dV_N/dy1 \\ dV_N/dz1 \\ \vdots \\ dV_N/dxn \\ dV_N/dyn \\ dV_N/dzn
\end{pmatrix}
\tag{14}$$

For a species with n atoms and N configurations $\mathbf{x}$ for which energies have been computed, the left-hand side matrix in Eq. 14 has dimension $(3n+1)N \times N$. Eq. 14 can be solved using a least square fitting algorithm. Here, the 'dgelss' subroutine from the LAPACK library[43] is used to solve the set of linear equations.

To better represent important (i.e. low-energy) regions of the PES, a weighted fit is performed. The weights w_i for each point have been chosen as

$$w_i = \frac{\Delta V}{\Delta V + (V_i - V_{\mathrm{min}})} \tag{15}$$

where ΔV is either a constant (here 4 eV) or the maximum energy of the training set relative to the minimum ($\Delta V = V_{\mathrm{max}} - V_{\mathrm{min}}$), and V_i is the relative energy of a configuration i with respect to the minimum energy of the system V_{min}. In this way, a larger weight is assigned to structures close to the equilibrium. A similar weight function is also used for the gradients

$$w_i = \frac{\Delta g}{\Delta g + |g_i|}. \tag{16}$$

The maximum value of Δg is 10 eV/a_0.

Symmetrized RKHS

One of the main challenges when constructing a multidimensional PES is to maintain the symmetry of the PES with respect to interchanging equivalent atoms. Configurations for all permutations of equivalent atoms are to be included. The most straightforward way is to include all permutationally equivalent configurations with the same energies in the training data set. However, this increases the size of the training data set, which also increases the evaluation cost in RKHS for an energy evaluations the sum runs for all the training structures. Also to obtain the coefficients the set of linear equations are solved numerically which may lead to a mismatch between energies of two equivalent structures due to numerical inaccuracies. Hence, it is advantageous to rather explicitly symmetrize the total kernel polynomial $K(\mathbf{r}, \mathbf{r}')$ (see Eq. 10) by expanding it as a linear combination of all equivalent structures of a molecule.

$$K_{\rm sym}(\mathbf{r},\mathbf{r}') = \sum_{i=1}^{S} K_i(\mathbf{r},\mathbf{r}'), \tag{17}$$

where S is the number of equivalent configurations. A similar strategy was followed in constructing PESs from PIPs for which symmetrized basis functions were generated by adding products of all 'monomials' for a molecule considering permutations of equivalent atoms.[44]

An example is given here for the CH_4 molecule. All permutations with respect to four equivalent H atoms are shown in Figure 1. Atom positions are assigned by 'a' through 'e', while different atoms can be distinguished by different colors. The order of the interatomic distances with respect to positions are given in Table 1 for all permutations. For CH_4 there are four and six equivalent CH and HH distances, respectively. To define a 1D kernel two bond distances are required: either $k(x,x')$ or $k(y,z')$ where x and x' are the same pairwise distance (here the C-H or H-H distances) and y and z' are two distances that need to be explicitly symmetrized (here two H-H or two C-H distances for symmetry-related hydrogen atoms). In the absence of symmetry, ten 1D kernels $((1^2 \times 4) + (1^2 \times 6))$ for interatomic distances define the basis set for RKHS $k(r_{\rm ab}, r'_{\rm ab}), k(r_{\rm ac}, r'_{\rm ac}), \cdots, k(r_{\rm de}, r'_{\rm de})$ (only one configuration is possible e.g. configuration 1 in Figure 1). However, using symmetry each configuration has 24 permutations which leads to 52 1D kernels ($4^2 + 6^2 = 52$ for the four CH and six HH bonds) for interatomic distances to complete the basis set for RKHS. All 52 1D basis kernel functions are reported in Table 1 i.e. $[k(r_{\rm yr}, r'_{\rm yr}), \cdots, k(r_{\rm mg}, r'_{\rm mg})]$, $[k(r_{\rm yr}, r'_{\rm yr}), \cdots, k(r_{\rm mg}, r'_{\rm bg})]$, $\cdots$, $[k(r_{\rm yr}, r'_{\rm yg}), \cdots, k(r_{\rm mg}, r'_{\rm rb})]$. It is to be noted that Table 1 contains 240 kernel functions in total whereas many of them $((6 \times (4 \times 4) + 4 \times (6 \times 6))$ are equivalent. The 2-body interaction energy is then the sum of all these 240 1D kernel functions.

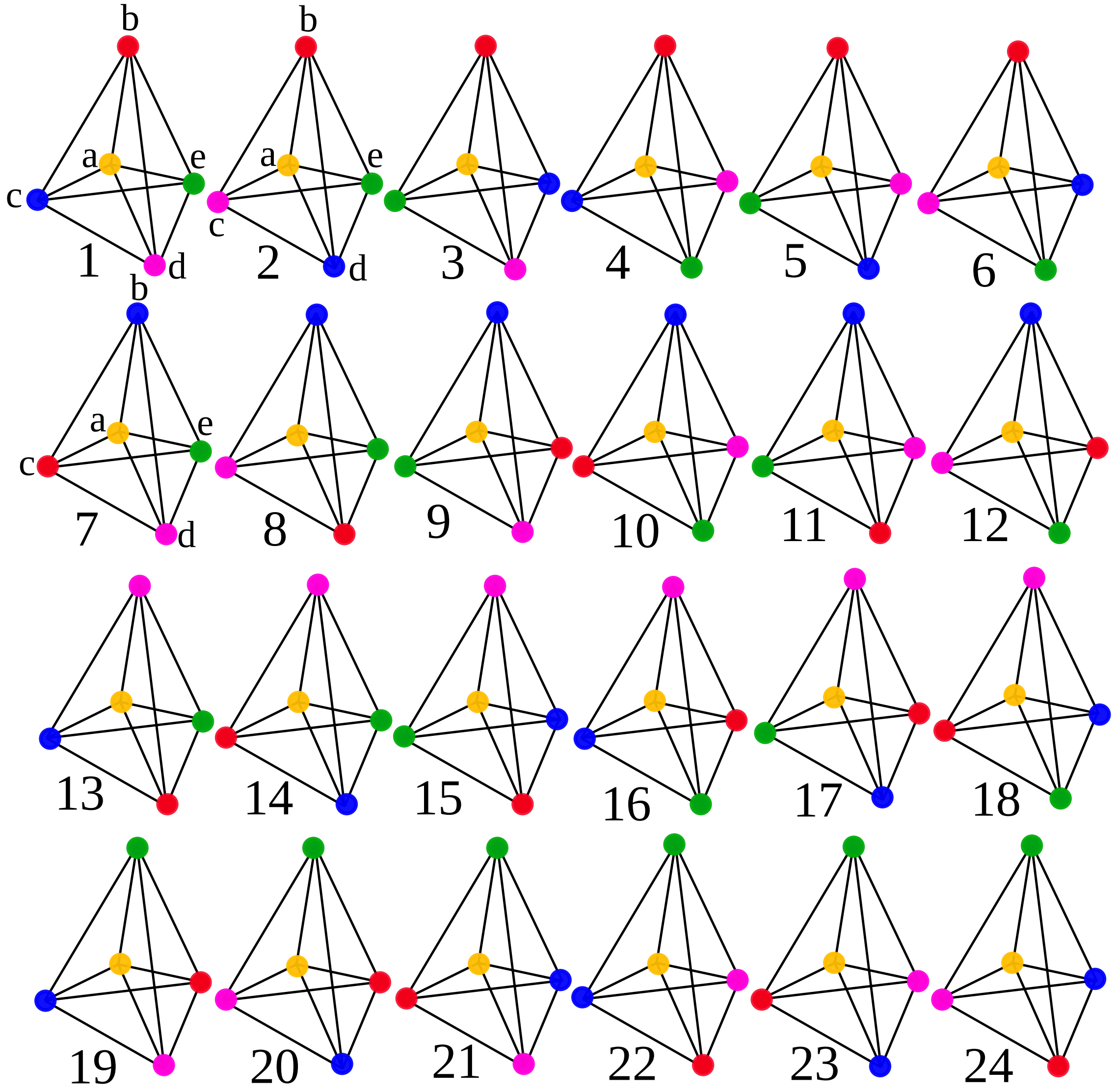

Figure 1: All 24 permutations of H atoms in CH_4 molecule. Atoms are represented by color, yellow (y) is for the carbon atom and red (r), blue (b), magenta (m) and green (g) for the hydrogen atoms. Positions of the atoms are denoted by 'a', 'b', 'c', 'd' and 'e'.

Table 1: Symmetrization order of interatomic distances for equivalent CH_4 structures. Interatomic distances between two different atoms/positions are $r_{ij} = r_{ji}$. Atom positions and color indices are defined in Figure 1.

Configurations	r_{ab}	r_{ac}	r_{ad}	r_{ae}	r_{bc}	r_{bd}	r_{be}	r_{cd}	r_{ce}	r_{de}
1	r_{yr}	r_{yb}	r_{ym}	r_{yg}	r_{rb}	r_{rm}	r_{rg}	r_{bm}	r_{bg}	r_{mg}
2	r_{yr}	r_{ys}	r_{yb}	r_{yg}	r_{rm}	r_{rb}	r_{rg}	r_{mb}	r_{mg}	r_{bg}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
23	r_{yg}	r_{yr}	r_{yb}	r_{ym}	r_{gr}	r_{gb}	r_{gm}	r_{rb}	r_{rm}	r_{bm}
24	r_{yg}	r_{ym}	r_{yr}	r_{yb}	r_{gm}	r_{gr}	r_{gb}	r_{mr}	r_{mb}	r_{rb}

Similarly, multidimensional product kernels for 3 or 4-body interaction energies can also be constructed from such 1D kernels. Note that p-body interactions must be considered for all permutations. For example, in the absence of symmetry, the CH_4 molecule has ${}^5C_4 = 5$ four body terms while including symmetry there are ${}^5C_4 \times 4! = 120$ four body terms. An explicit example for all 2-, 3-, and 4-body terms for the case of CH_2O is given in the supporting information.

To determine all combinations of the 2-, 3-, and 4-body terms an automated procedure is required that handles all possible symmetry terms and also to eliminate redundancies. For this, an in-house python[45] code was written using the 'itertools' module. The software generates both, the required symmetrized form of the RKHS and efficient fortran source code. A related strategy was followed recently when constructing the fitting coefficients for PESs represented as PIPs.[4]

Generation of the Reference Data Sets

Although much higher levels of theory could in principle be used, in particular for the smaller systems, the reference calculations in the present work were carried out at the density functional theory (DFT) level for convenience and illustration. All electronic structure calculations were performed using the Orca 4.0[46] software using the B3LYP functional[47,48] and cc-pVDZ[49] basis set, similar to previous work on the PIP-based PES for NMA.[34] 'Very tight' SCF convergence (10^{-9} hartree) criteria along with the largest grid ('grid7') for the Lebedev integration were used in all calculations. The structures of all molecules were optimized and harmonic frequencies were determined. Then, reference structures were sampled using an in-house written code as described in Ref. 50 at different temperatures (20 to 2500 K) by distorting the equilibrium structures and randomly displacing the atoms along the normal modes. For each of the systems, energies and gradients were calculated for 4000 to

10000 reference structures. From this reference data, $N_{\text{train}} = 1600$ to 2500 structures were used for constructing the RKHS (see Table 2) and $N_{\text{test}} = 800$ to 1000 structures, randomly drawn from the remaining data, were used for testing. Here, it is worth to be mentioned that all structures with energies larger than 4 eV with respect to the global minimum were excluded from the reference and the test set.

Results

Quality and Extrapolation of the PESs

First the quality of the resulting potential energy surfaces is discussed. Unless otherwise stated, all RKHS-PESs were constructed from using energies and gradients. As an example the data set generated and used in constructing the multidimensional PESs for formaldehyde is reported in Figure 2. It shows the total data set (brown), the reference set (blue), and the extrapolation set (red) which extends to considerably higher energies. This last data set is used to assess the extrapolation capabilities of the RKHS-based PESs for structures (sampled at 5000 K), potentially far outside the configurations used for generating the RKHS representation. One of the potential shortcomings of certain machine learning approaches for inter- and intramolecular PESs is their limitation as valid *interpolators* but not to extrapolate well beyond the structures used to generate the model.

The performance of the RKHS-based PES for the test set is illustrated in Figure 3. Both, energies and forces are very accurately described as the RMSE and MAE of 0.0003 kcal/mol and 0.0002 kcal/mol for energies and 0.004 kcal/mol/Å and 0.002 kcal/mol/Å for forces (gradients) demonstrate. For the coefficient of determination, R^2, one finds $1 - R^2 = 4 \times 10^{-9}$ and $1 - R^2 = 2 \times 10^{-8}$ for energies and forces, respectively, see Table 2.

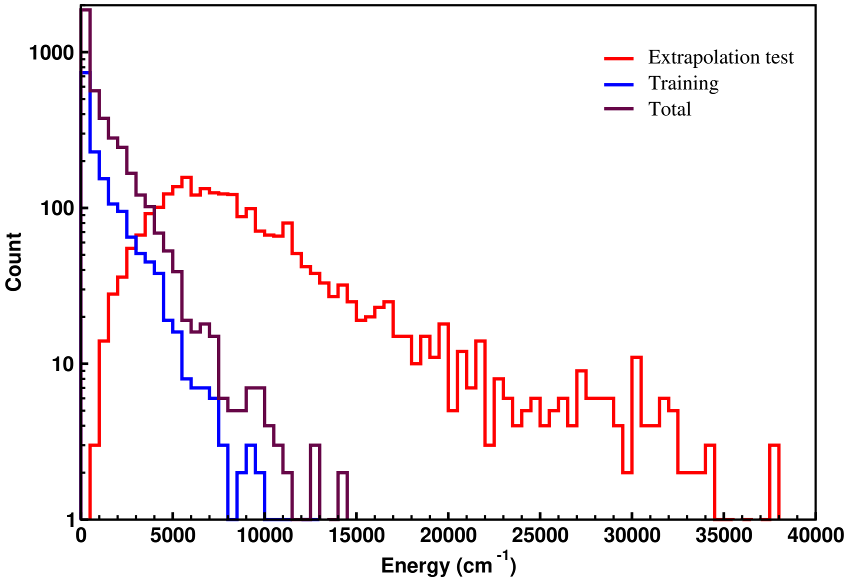

Figure 2: Distribution of the reference and extrapolation data set for CH_2O. The distribution of the total data set (4001 points, brown) along with 1600 reference energies (blue lines) and 2500 extrapolation energies (red lines). The counts are given on a logarithmic scale.

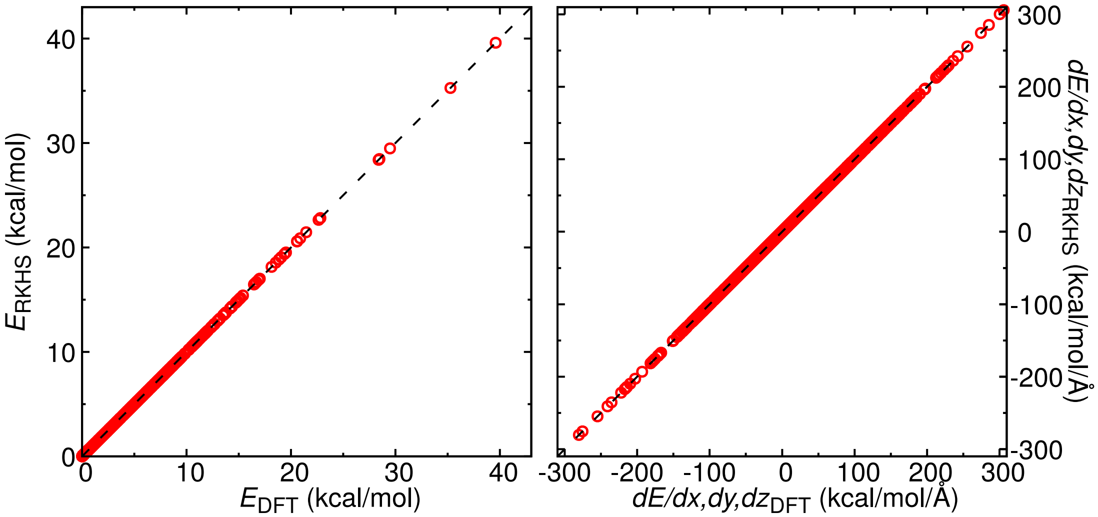

Figure 3: Correlation between the energies (right) and gradients (left) for CH_2O molecule obtained from DFT calculations and predicted by the RKHS PES for 800 test data set. The RMSE and MAE for the PESs of all molecules are reported in Table 2.

Although the performance on the test data is very favourable, an even more important aspect of molecular PESs in particular when used in atomistic simulation is their validity and quality for structures far away from those they were trained on. This is required for stable and meaningful MD simulations. The extrapolation capability is demonstrated in Figure 4

Table 2: Molecules, their sizes ($N_{\text{atom}} \equiv n$), and the number of training $\mathrm{N}_{\text{train}}$ and test N_{test} structures used. For each molecule the root mean squared error (RMSE), mean absolute error (MAE) for energies (kcal/mol) and forces (kcal/mol/Å) and Pearson correlation coefficient calculated for N_{test} test data is given.

Molecule	N_{atom}	N_{train}	N_{test}	RMSE	MAE	$1-R^2$	RMSE	MAE	$1-R^2$
					Energy			Force	
CH_2O	4	1600	800	0.0003	0.0002	4×10^{-9}	0.0044	0.0021	2×10^{-8}
CH_4	5	2400	1000	0.0018	0.0013	9×10^{-8}	0.0098	0.0048	5×10^{-7}
HCOOH	5	2400	1000	0.0015	0.0007	2×10^{-7}	0.0161	0.0069	2×10^{-6}
CH_3OH	6	2400	1000	0.0205	0.0102	5×10^{-6}	0.1064	0.0550	6×10^{-5}
CH_3CHO	7	2400	1000	0.0246	0.0124	4×10^{-6}	0.1067	0.0580	8×10^{-5}
CH_3NO_2	7	2500	1000	0.0181	0.0092	1×10^{-5}	0.0974	0.0525	9×10^{-5}
CH_3COOH	8	2500	1000	0.0188	0.0093	6×10^{-7}	0.0919	0.0483	5×10^{-5}
CH_3CONH_2	9	2500	1000	0.0431	0.0132	2×10^{-6}	0.1190	0.0571	5×10^{-5}
CH_3COCH_3	10	2500	1000	0.1019	0.0659	2×10^{-5}	0.3067	0.2002	3×10^{-4}

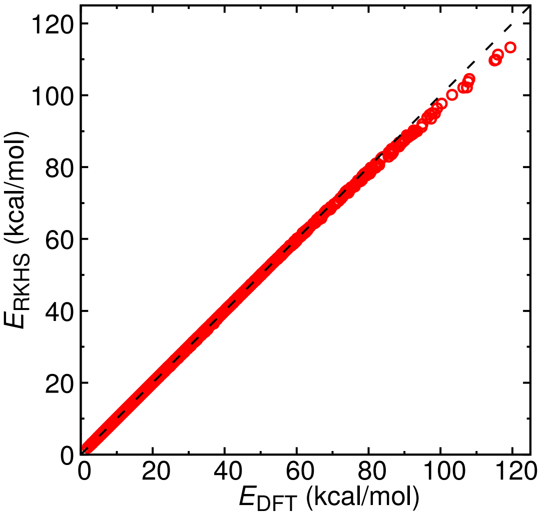

Figure 4: Performance on the 2500 structures for CH_2O from the extrapolation data set (red line in Figure 2), sampled at 5000 K. Correlation between the energies obtained from DFT calculations and predicted by the RKHS PES trained on energies and gradients for 1600 structures. The RKHS prediction has an RMSE of 0.532 kcal/mol, MAE of 0.114 kcal/mol with $R^2 = 0.99913$).

which demonstrates that the RKHS PES for CH_2O remains accurate for energies three times higher than for the energies in the reference and test set. Up to energies $\sim$ 100 kcal/mol above the global minimum the RMSE is better than 0.5 kcal/mol which allows reliable MD simulations even at high temperatures.

The supporting information provides similar information for the CH_4 molecule, i.e. the energy distribution for all energies, those used for constructing the RKHS-PES and those used for testing (see Figure S2 and the validation of the RKHS-PES as the correlation of energies and gradients between the reference calculations and the evaluation of the RKHS-PES (Figure S3). Very accurate predictions can also been achieved in this case.

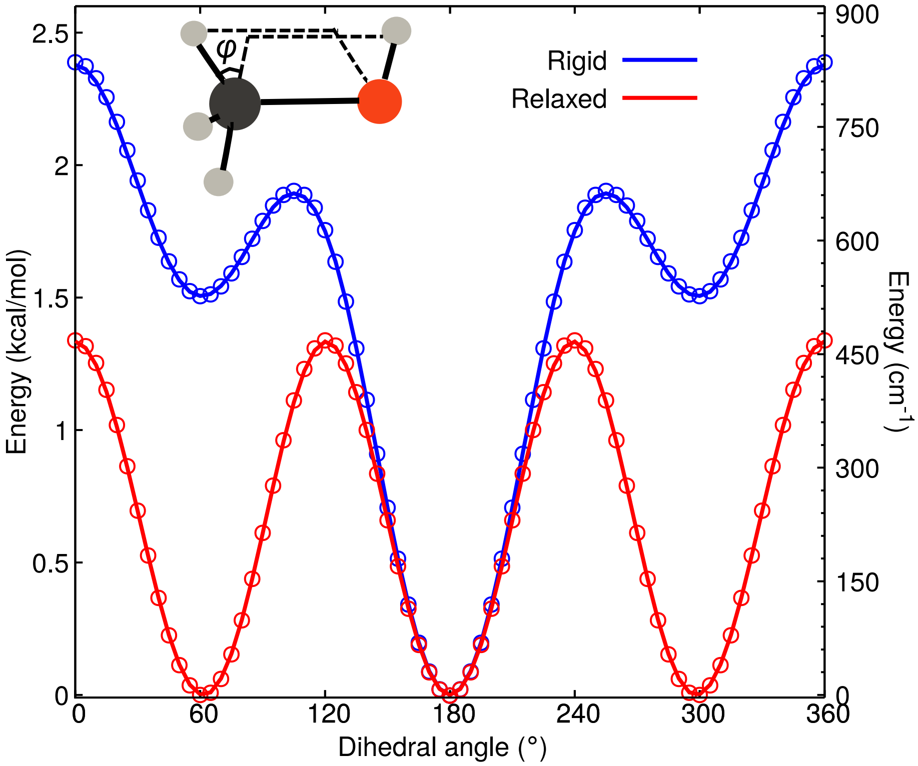

Figure 5: Potential energies obtained from DFT calculations (open circles) and RKHS PES (solid lines) as a function of the H-C-O-H dihedral angle in CH_3OH. Blue line shows energies for a rigid scan changing only one H-C-O-H dihedral angle and the red line shows energies for a relaxed scan where the molecule is optimized for each value of the H-C-O-H angle. The definition of the dihedral angle is shown at top left; filled circles represent different atoms, gray black and red color represent the H, C and O atoms, respectively.

A typical cut through the global potential energy surface is afforded by considering 1-dimensional energy functions along particular internal degrees of freedom. One degree of freedom that is particularly challenging in empirical energy function ("force field") development are dihedral torsions. Figure 5 reports the potential energy profiles along the H-C-O-H torsion in CH_3OH for a rigid and a relaxed scan. In a rigid scan potential energies are calculated for different values of the H-C-O-H dihedral angle while keeping all other degrees of freedom frozen at the equilibrium geometry. Conversely, in a relaxed scan the structure of the molecule is optimized for a given value of the H-C-O-H dihedral angle. Both scans from the RKHS PES accurately reproduce the reference B3LYP data. The symmetry of the molecule (i.e. permutations among the methyl hydrogens) is also preserved in the RKHS PES.

To quantify the advantage of the "energy+gradient" based RKHS method over the "energy-only" data set (where only energies are used as an input to obtain the coefficients, see Eq. 6) energy and force learning curves on the test data sets are calculated for CH_4. The "learning curves" for the RMSE (red lines) and MAE (blue lines) for both energies and forces, using "energy only" (dashed) and "energy+gradient" (solid), are shown in Figure S4. When using "energy only" (dashed curves), both, energies (left panel) and forces (right panel) continuously improve as the size of the training set increases and further improvements appear to be possible beyond 6×10^{-4} kcal/mol for energies and 6×10^{-3} kcal/mol/Å for the largest training set ($N_{\mathrm{train}} = 9600$). However, for the forces the "energy+gradient" approach reaches similar accuracy as the "energy only" RKHS using 1/6 of the data (i.e. $N_{\mathrm{train}}^{\mathrm{energy+gradient}} = 1600$ vs. $N_{\mathrm{train}}^{\mathrm{energy}} = 9600$). Hence, including gradient information explicitly in the RKHS, see Eq. 14, reduces the number of coefficients which also speeds up the RKHS evaluation. The energy learning curves from using "energy+gradient" in constructing the RKHS-PESs appear to saturate with ($N_{\mathrm{train}} = 3200$) at similar values for RMSD and MAE. This is because the weights of the gradients are $3n$ times larger than those for the energies, where n is the total number of atoms of the molecule.

Quality of Normal Mode Frequencies from RKHS-PESs

Normal mode frequencies are useful computational observables to compare the performance of fitted PESs with the reference calculations they are based on.[34] Harmonic frequencies were calculated for the molecules using the ASE package[51] by linking the RKHS PESs as an external energy calculator. Table 3 compares the normal mode frequencies from the B3LYP/cc-pVDZ calculations with those from the RKHS-represented PESs for CH_2O, HCOOH, and CH_4. Besides the remarkable accuracy (difference < 1 cm^{-1} for every mode) with which the kernel-represented PESs are capable of describing the reference calculation for all examples considered, maintaining the correct symmetry and degeneracy in the case of CH_4 is most notable. In particular, the RKHS PES exactly (for the HCH bend) or very closely (for the CH stretch) maintains the two triply degenerate modes at 1309 cm^{-1} and 3146 cm^{-1}, respectively, as it should be. This also underlines the correct implementation of permutational invariance in the formulation.

Table 3: Harmonic frequencies (in cm^{-1} and rounded to full wavenumbers) and zero point energies (in eV) for CH_2O, HCOOH and CH_4 computed using the reference B3LYP/cc-pVDZ calculations (Ref.) and calculated from their RKHS-PES (RKHS). The RKHS-PESs were trained on energies and gradients. The RMSD between reference values and those from the RKHS PESs is well below 1 cm^{-1}.

	CH_2O		HCOOH		CH_4	
mode	Ref.	RKHS	Ref.	RKHS	Ref.	RKHS
1	1186	1186	627	627	1309	1309
2	1252	1252	700	701	1309	1309
3	1514	1514	1046	1046	1309	1309
4	1831	1831	1138	1137	1530	1529
5	2862	2862	1311	1311	1530	1530
6	2914	2914	1394	1393	3025	3025
7			1843	1843	3146	3145
8			3031	3031	3146	3146
9			3676	3677	3146	3146
ZPE	0.717	0.717	0.916	0.917	1.206	1.206

A broader overview of all harmonic frequencies for all compounds in Table 2 is shown in

Figure 6. These normal mode frequencies are from the RKHS-PESs trained on energies and gradients. For the 124 normal mode frequencies the overall MAE between reference calculations and frequencies determined on the RKHS-PESs is 4.1 cm^{-1} with $R^2 = 0.99995$. This is consistent with the high accuracy of the energies and forces reported in Table 2. Here it is worth to be mentioned that for larger molecules low frequency (< 200 cm^{-1}) modes contribute most to the error. This is consistent with recent work using PIPs for a full-dimensional PES for N-methyl acetamide for which some of the low-frequency modes differ up to ~ 30 cm^{-1}.[34] It should be emphasised that such accuracy is independent of the quality of the electronic structure method used for the reference calculations. In other words, if energies and forces are available at a considerably higher level of theory (e.g. CCSD(T) with a large basis set) the same performance in reproducing such reference data as that reported here is expected which provides a very high accuracy but computationally efficient energy function with analytical gradients.

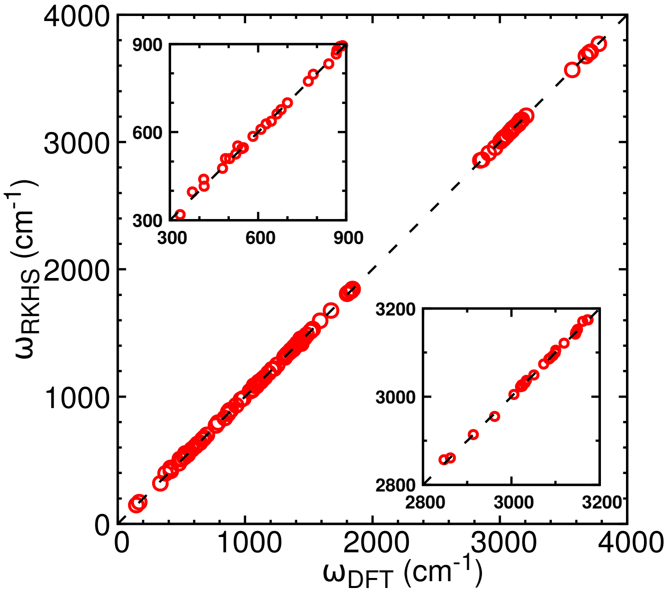

Figure 6: Correlation between the harmonic frequencies for all the systems considered obtained from DFT calculations and RKHS PESs with an RMSE of 6.7 cm^{-1} and MAE of 4.1 cm^{-1}, and $R^2 = 0.99995$. The insets show magnifications of the low- and high-frequency vibrations. All RKHS PESs are based on energy+gradients.

Another property of interest concerns the change (ideally "improvement") of an observable (here normal modes) as the number of training data N_{train} increases. This is reported in Figure 7 for RKHS-PESs trained on "energies only" and "energies + gradients". When energies only are used for training the RKHS PES for CH_4 an average error better than 1 cm^{-1} requires $N_{\mathrm{train}} \sim 3200$ training data whereas including energies and gradients in generating the RKHS-PES already achieves this with $N_{\mathrm{train}} \sim 400$. This should be compared with the findings for the learning curves in Figure S3 that report a similar performance for "energy only" and "energy+gradients" for $N_{\mathrm{ref}} = 9600$ and $N_{\mathrm{ref}} = 1600$, respectively. This is attributed to the additional information the gradients provide about the local curvature around every structure for which an energy is available. Furthermore, the curves in Figure 7 behave very differently for "energy only" and "energy+gradients" used in constructing the RKHS-PES. Whereas the PES trained on "energies only" appears to have two slopes (up to $N_{\mathrm{train}} \sim 400$ and beyond $N_{\mathrm{train}} > 800$ with a local maximum deviation at $N_{\mathrm{train}} \sim 800$), normal modes determined on the "energy+gradients" trained PESs continuously improve until $N_{\mathrm{train}} \sim 1600$ to an average error of 0.2 cm^{-1} after which they level off within the fluctuation bars. Probably this is the maximum accuracy that can be achieved for harmonic frequencies. Again it is to be mentioned that the Hessian is calculated numerically in ASE.

Discussion and Conclusions

The present work introduces an extension of RKHS-based PESs[8] to polyatomic molecules. Combining energy and force information to construct tensor-product based kernels up to 4-body interactions is shown to yield highly accurate PESs for molecules ranging from formaldehyde to acetone. Using "energy + gradients" for constructing the RKHS-PES requires between a factor of 6 to 10 less reference data than working with "energy only". The RKHS-PESs are very accurate and extrapolate well to structures with considerably higher

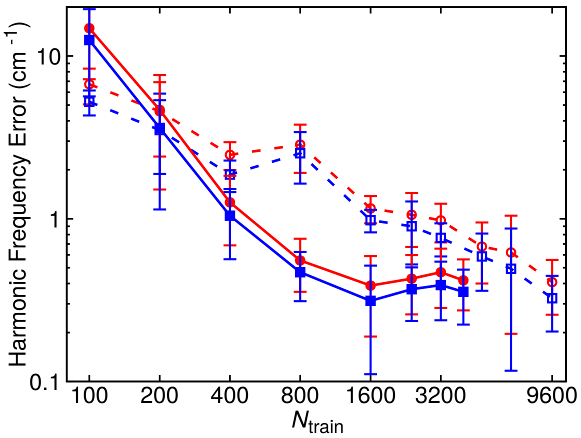

Figure 7: Root mean squared difference for harmonic frequencies for CH_4 from using "energy only" (dashed lines and open symbols) and "energy+gradient" (solid lines and filled symbols) training. For a given number of training data each model is trained for five times for random data set. Average values and standard deviations (error bars) of the RMSE and MAE are shown as red and blue, respectively.

energies, see Figure 4. This is not guaranteed for NN-learned PESs as recent work on acetaldehyde[52] with the PhysNet[12] NN-architecture has shown. Unless structures at the highest energies are included, many of the MD trajectories become invalid as the energies and forces generated from the NN are inconsistent with the true energies and forces compared with the reference electronic structure calculations.

The harmonic modes computed from the RKHS PES and from the reference electronic structure calculations (here B3LYP/cc-pVDZ) are within 1 cm^{-1} for small molecules and within 5 cm^{-1} for larger molecules except for low frequency modes ($<$ 200 cm^{-1}). Similar observation were also made for *cis*- and *trans*-NMA using PIP-based PESs.[34,35] Such performance naturally extends to reference data computed at a much higher level of theory. Hence, for systems with up to 10 atoms considered here the only limitation will be the computing time required for generating the training and test data set.

To achieve an agreement between reference data and that from the representation (here RKHS) for arbitrary configurations or even low-dimensional projections (e.g. a torsional potential) for bonded terms is extremely challenging for empirical force fields. As an example, earlier versions of the CHARMM force field[53] had to be empirically corrected by introducing the CMAP correction[54] to account for deficiencies in the dihedral potentials. Because the number of dihedral terms is large and primarily responsible for secondary and tertiary structural changes in peptides and proteins, specifically improving these contributions to empirical force fields appears to be a useful possibility. It is also worth to point out that the RKHS PES is permutationally invariant for the equivalent methyl H atoms which is also seen in Figure 5. These findings also extend to larger molecules as demonstrated for dihedral scans for acetone as reported in Figure 8. The relaxed scan from the reference B3LYP/cc-pVDZ calculations and the RKHS PES agree very well except around the top of the barrier where they differ by ~ 25 cm^{-1}. Both the methyl group and also the methyl hydrogens in each group preserved their symmetry in the RKHS PES.

Another future application of the methods discussed here are molecular dynamics simulations of small molecules on global, anharmonic and fully coupled RKHS PESs. As an example, the infrared spectrum for CH_4 in the gas phase is reported in Figure 9. This simulation was carried out with a suitably modified version of the CHARMM molecular simulation program[55] to use energies and forces from the RKHS-PES. The PES trained on 2400 structures using both energies and gradients was used. The time step in this simulation was 0.1 fs and the simulation temperature was 300 K. First, the system is heated to the simulation temperature, equilibrated for 7 ps and then an equilibrium NVE simulation was carried out for 250 ps. Total energy is conserved to within 0.015 kcal/mol, see inset of Figure 9, which underlines that the forces in the RKHS are correctly implemented.

Finally, the possibility to extend the methodology introduced here to intermolecular interactions is mentioned. The present work was concerned with the "bonded interactions" when comparing with empirical force field technology.[56–58] However, for condensed phase simula-

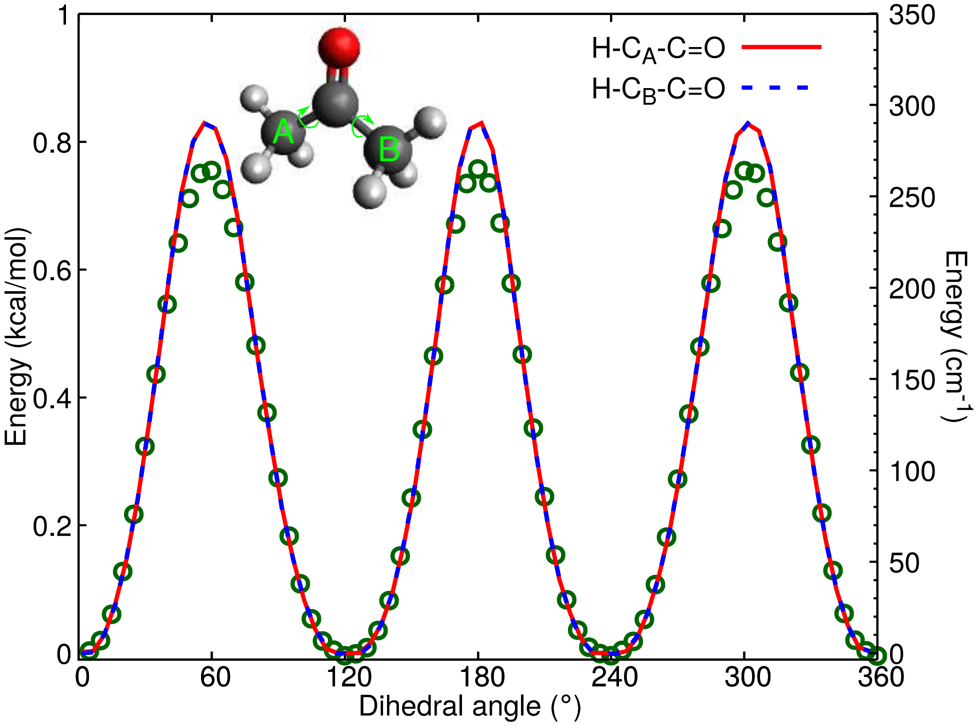

Figure 8: Potential energies obtained from B3LYP/cc-pVDZ calculations (green open circles) and RKHS PES (solid red and dashed blue lines) as a function of H-C-C-O dihedral angles in CH_3COCH_3 (acetone). A relaxed scan is performed for both the dihedral angles where the molecule is optimized for each points.

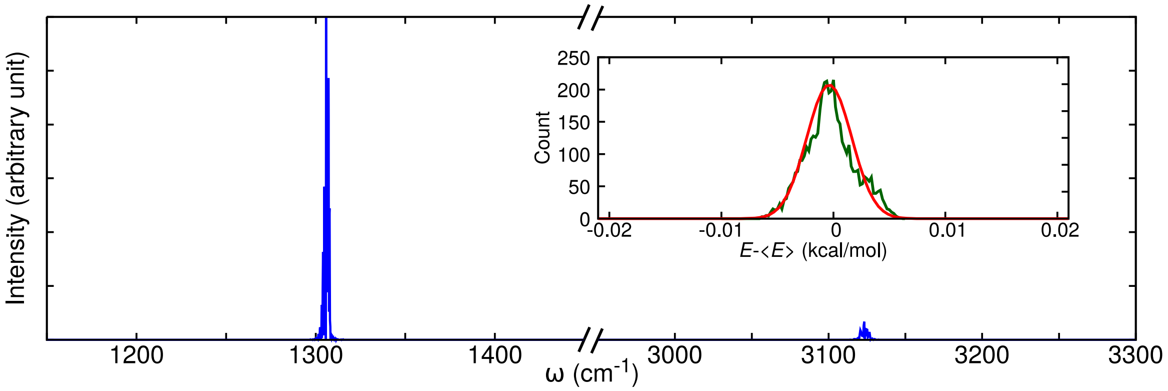

Figure 9: IR spectrum for CH_4 obtained from the dipole moment autocorrelation function and subsequent fast Fourier transformation. The molecular dipole moment was computed by using Mulliken point charges from DFT calculations for the equilibrium structure. The infrared active modes (triply degenerate HCH bend and triply degenerate CH stret modes) are at 1306 cm^{-1} and 3123 cm^{-1}, respectively. As required, the totally symmetric, infrared inactive CH stretch mode at 3025 cm^{-1}, see Table 3, does not appear in the infrared spectrum. The inset shows the distribution of the total energy fluctuation around its $\langle E \rangle$ (green line) in the MD simulations with a superimposed Gaussian function (red line).

tions, nonbonded interactions between, e.g., a solute and the surrounding solvent need to be determined and available as well. One future possibility is to combine the accurate RKHS-PESs discussed here with accurate multipolar electrostatic models (possibly augmented by polarization).[59,60] Alternatively, developing an RKHS-based fragment approach can be envisaged to treat molecular dimers and trimers.

In conclusion, the RKHS technique which has already been found to be highly beneficial for the study of reactive processes[19,22,24,61] and spectroscopic studies[18,21,62] has been considerably extended to treat the intramolecular degrees of freedom for molecules with up to 10 atoms. Together with further developments this approach is expected to provide a way towards quantitative gas- and condensed-phase simulations.

Acknowledgment

The authors acknowledge financial support from the Swiss National Science Foundation (NCCR-MUST and Grant No. 200021-7117810), the AFOSR, and the University of Basel.

References

(1) Varandas, A. J. C. *Advances in Chemical Physics*; John Wiley & Sons, Inc., 2007; pp 255–338.

(2) Porter, R. N.; Karplus, M. Potential Energy Surface for H_3. *J. Chem. Phys.* **1964**, *40*, 1105–1115.

(3) Aguado, A.; Paniagua, M. A New Functional form to Obtain Analytical Potentials of Triatomic Molecules. *J. Chem. Phys.* **1992**, *96*, 1265–1275.

(4) Qu, C.; Yu, Q.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces. *Annu. Rev. Phys. Chem.* **2018**, *69*, 151–175.

(5) Xu, C.; Xie, D.; Zhang, D. H.; Lin, S. Y.; Guo, H. A new ab initio potential-energy surface of $HO_2(X^2A'')$ and quantum studies of HO_2 vibrational spectrum and rate constants for the $H+O_2 \leftrightarrow O+OH$ reactions. *J. Chem. Phys.* **2005**, *122*, 244305.

(6) Shepard, D. A Two-Dimensional Interpolation Function for Irregularly-Spaced Data. Proceedings of the 1968 23rd ACM National Conference. New York, NY, USA, 1968; pp 517–524.

(7) Crespos, C.; Collins, M. A.; Pijper, E.; Kroes, G. J. Application of the modified Shepard interpolation method to the determination of the potential energy surface for a molecule-surface reaction: H_2+Pt(111). *J. Chem. Phys.* **2004**, *120*, 2392–2404.

(8) Ho, T.-S.; Rabitz, H. A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations. *J. Chem. Phys.* **1996**, *104*, 2584.

(9) Unke, O. T.; Meuwly, M. Toolkit for the Construction of Reproducing Kernel-Based Representations of Data: Application to Multidimensional Potential Energy Surfaces. *J. Chem. Inf. Model* **2017**, *57*, 1923–1931.

(10) Rasmussen, C. E. *Gaussian Processes in Machine Learning*; Springer, Berlin, Heidelberg, 2004.

(11) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(12) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theor. Comput.* **2019**, *15*, 3678–3693.

(13) Koner, D.; Panda, A. N. Quantum Dynamical Study of the He + NeH^+ Reaction on a New Analytical Potential Energy Surface. *J. Phys. Chem. A* **2013**, *117*, 13070–13078.

(14) Paukku, Y.; Yang, K. R.; Varga, Z.; Truhlar, D. G. Global Ab Initio Ground-State Potential Energy Surface of N_4. *J. Chem. Phys.* **2013**, *139*, 044309.

(15) Koner, D.; Barrios, L.; González-Lezana, T.; Panda, A. N. Scattering study of the Ne + $NeH^+(v_0 = 0, j_0 = 0) \rightarrow NeH^+$ + Ne reaction on an ab initio based analytical potential energy surface. *J. Chem. Phys.* **2016**, *144*, 034303.

(16) Nandi, A.; Qu, C.; Bowman, J. M. Using Gradients in Permutationally Invariant Polynomial Potential Fitting: A Demonstration for CH_4 Using as Few as 100 Configurations. *J. Chem. Theor. Comput.* **2019**, *15*, 2826–2835.

(17) Unke, O. T.; Koner, D.; Patra, S.; Käser, S.; Meuwly, M. High-dimensional potential energy surfaces for molecular simulations: from empiricism to machine learning. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 013001.

(18) Salehi, S. M.; Koner, D.; Meuwly, M. Vibrational Spectroscopy of N_3^- in the Gas and Condensed Phase. *J. Phys. Chem. B* **2019**, *123*, 3282–3290.

(19) Koner, D.; Bemish, R. J.; Meuwly, M. The $C(^3P) + NO(X^2\Pi) \rightarrow O(^3P) + CN(X^2\Sigma^+)$, $N(^2D)/N(^4S) + CO(X^1\Sigma^+)$ reaction: Rates, branching ratios, and final states from 15 K to 20 000 K. *J. Chem. Phys.* **2018**, *149*, 094305.

(20) Hollebeek, T.; Ho, T.-S.; Rabitz, H.; Harding, L. B. Construction of reproducing kernel Hilbert space potential energy surfaces for the $^1A''$ and $^1A'$ states of the reaction $N(^2D)+H_2$. *J. Chem. Phys.* **2001**, *114*, 3945–3948.

(21) Koner, D.; San Vicente Veliz, J. C.; van der Avoird, A.; Meuwly, M. Near dissociation states for H_2^+–He on MRCI and FCI potential energy surfaces. *Phys. Chem. Chem. Phys.* **2019**, *21*, 24976–24983.

(22) San Vicente Veliz, J. C.; Koner, D.; Schwilk, M.; Bemish, R. J.; Meuwly, M. The $N(^4S)+O_2(X^3\Sigma_g^-) \leftrightarrow O(^3P)+NO(X^2\Pi)$ Reaction: Thermal and Vibrational Relaxation Rates for the $^2A'$, $^4A'$ and $^2A''$ States. *Phys. Chem. Chem. Phys.* **2020**, *22*, 3927–3939.

(23) Pezzella, M.; Koner, D.; Meuwly, M. Formation and Stabilization of Ground and Excited-State Singlet O_2 upon Recombination of ^3P Oxygen on Amorphous Solid Water. *J. Phys. Chem. Lett.* **2020**, *11*, 2171–2176.

(24) Koner, D.; San Vicente Veliz, J. C.; ; Bemish, R. J.; Meuwly, M. Accurate Reproducing Kernel-Based Potential Energy Surfaces for the Triplet Ground States of N_2O and Dynamics for the $N+NO \leftrightarrow O+N_2$ Reaction. *arXiv preprint arXiv:2002.02310* **2020**,

(25) Soldán, P.; Hutson, J. M. On the long-range and short-range behavior of potentials from reproducing kernel Hilbert space interpolation. *J. Chem. Phys.* **2000**, *112*, 4415–4416.

(26) Ho, T.-S.; Rabitz, H. Proper construction of ab initio global potential surfaces with accurate long-range interactions. *J. Chem. Phys.* **2000**, *113*, 3960–3968.

(27) Hollebeek, T.; Ho, T.-S.; Rabitz, H. A fast algorithm for evaluating multidimensional potential energy surfaces. *J. Chem. Phys.* **1997**, *106*, 7223–7227.

(28) Hollebeek, T.; Ho, T.-S.; Rabitz, H. Efficient potential energy surfaces from partially filled ab initio data over arbitrarily shaped regions. *J. Chem. Phys.* **2001**, *114*, 3940–3944.

(29) Ho, T.-S.; Rabitz, H. Reproducing kernel Hilbert space interpolation methods as a paradigm of high dimensional model representations: Application to multidimensional potential energy surface construction. *J. Chem. Phys.* **2003**, *119*, 6433–6442.

(30) Luo, X.; Lu, Z.; Xu, X. Reproducing kernel technique for high dimensional model representations (HDMR). *Comput. Phys. Commun.* **2014**, *185*, 3099 – 3108.

(31) Wormer, P. E. S.; Kłos, J. A.; Groenenboom, G. C.; van der Avoird, A. Ab initio computed diabatic potential energy surfaces of OH–HCl. *J. Chem. Phys.* **2005**, *122*, 244325.

(32) van der Avoird, A.; Bondo Pedersen, T.; Dhont, G. S. F.; Fernández, B.; Koch, H. Ab initio potential-energy surface and rovibrational states of the HCN-HCl complex. *J. Chem. Phys.* **2006**, *124*, 204315.

(33) Gubbels, K. B.; Meerakker, S. Y. T. v. d.; Groenenboom, G. C.; Meijer, G.; van der Avoird, A. Scattering resonances in slow NH_3–He collisions. *J. Chem. Phys.* **2012**, *136*, 074301.

(34) Qu, C.; Bowman, J. M. A fragmented, permutationally invariant polynomial approach for potential energy surfaces of large molecules: Application to N-methyl acetamide. *J. Chem. Phys.* **2019**, *150*, 141101.

(35) Nandi, A.; Qu, C.; Bowman, J. M. Full and fragmented permutationally invariant polynomial potential energy surfaces for trans and cis N-methyl acetamide and isomerization saddle points. *J. Chem. Phys.* **2019**, *151*, 084306.

(36) Aronszajn, N. Theory of Reproducing Kernels. *Trans. Amer. Math. Soc.* **1950**, *68*, 337–404.

(37) Schölkopf, B.; Herbrich, R.; Smola, A. J. A Generalized Representer Theorem. International Conference on Computational Learning Theory. 2001; pp 416–426.

(38) Hollebeek, T.; Ho, T.-S.; Rabitz, H. Constructing multidimensional molecular potential energy surfaces from ab initio data. *Annu. Rev. Phys. Chem.* **1999**, *50*, 537–570.

(39) Müller, K.-R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* **2001**, *12*.

(40) Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel methods in machine learning. *Ann. Stat.* **2008**, 1171–1220.

(41) Golub, G. H.; Van Loan, C. F. *Matrix Computations*; JHU Press Baltimore, 2012; Vol. 3.

(42) Tikhonov, A. N.; Arsenin, V. I.; John, F. *Solutions of Ill-Posed Problems*; Winston Washington, DC, 1977; Vol. 14.

(43) Anderson, E.; Bai, Z.; Dongarra, J.; Greenbaum, A.; McKenney, A.; Du Croz, J.; Hammarling, S.; Demmel, J.; Bischof, C.; Sorensen, D. LAPACK: A Portable Linear Algebra Library for High-Performance Computers. Proceedings of the 1990 ACM/IEEE Conference on Supercomputing. Washington, DC, USA, 1990; p 2âĂŞ11.

(44) Braams, B. J.; Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577–606.

(45) Python Software Foundation, Python 3.0, https://www.python.org/. `https://www.python.org/`.

(46) Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1327.

(47) Becke, A. Density-Functional Thermochemistry. III. The Role of Exact Exchange . *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(48) Lee, C.; Yang, W.; Parr, R. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(49) Dunning, T. H. J. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen . *J. Chem. Phys.* **1989**, *90*, 1007.

(50) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(51) Larsen, A. H.; Mortensen, J. J. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(52) Käser, S.; Unke, O. T.; Meuwly, M. Isomerization and Decomposition Reactions of Acetaldehyde Relevant to A tmospheric Processes from Dynamics Simulations on Neural Network-Based Potential Energy Surfaces. *arXiv preprint arXiv:2003:08171, accepted in J. Chem. Phys.* **2020**,

(53) MacKerell, A.; Bashford, D.; Bellott, M.; Dunbrack, R.; Evanseck, J.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*.

(54) MacKerell, A.; Feig, M.; Brooks, C. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(55) Brooks, B. R.; Brooks III, C. L.; MacKerell, Jr., A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. et al. CHARMM: The Biomolecular Simulation Program. *J. Comp. Chem.* **2009**, *30*, 1545–1614.

(56) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theor. Comput.* **2012**, *8*, 3257–3273.

(57) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(58) Jorgensen, W. L.; Tirado-Rives, J. The OPLS potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(59) Kramer, C.; Gedeck, P.; Meuwly, M. Atomic Multipoles: Electrostatic Potential Fit, Local Reference Axis Systems and Conformational Dependence. *J. Comp. Chem.* **2012**, *33*, 1673–1688.

(60) Bereau, T.; Kramer, C.; Meuwly, M. Leveraging Symmetries of Static Atomic Multipole Electrostatics in Molecular Dynamics Simulations. *J. Chem. Theor. Comput.* **2013**, *9*, 5450–5459.

(61) Dörfler, A. D.; Eberle, P.; Koner, D.; Tomza, M.; Meuwly, M.; Willitsch, S. Long-range versus short-range effects in cold molecular ion-neutral collisions. *Nat. Commun.* **2019**, *10*, 5429.

(62) Koner, D.; Schwilk, M.; Patra, S.; Bieske, E. J.; Meuwly, M. N_3^+: Full-Dimensional Potential Energy Surface, Vibrational Energy Levels and Ground State Dynamics. *arXiv preprint arXiv:22004.12404* **2020**,

Supporting Information: Permutationally Invariant, Reproducing Kernel-Based Potential Energy Surfaces for Polyatomic Molecules: From Formaldehyde to Acetone

Debasish Koner and Markus Meuwly*

Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland

E-mail: m.meuwly@unibas.ch

Symmetrized RKHS for CH_2O

Here the explicit expressions for non-symmetrized and symmetrized preserving permutational invariance for 2-, 3-, and 4-body terms are given explicitly for formaldehyde.

2-, 3- and 4-body Terms Without Symmetry

Formaldehyde has two permutationally invariant hydrogen atoms which leads to two equivalent configurations for any structure, see Figure S1. Table S1 shows the permutationally invariant interatomic distances for the two equivalent configurations, configuration 1 and configuration 2. Without symmetry, there are six different interatomic distances, i.e. $r_{\rm ab}$,

r_{ac}, r_{ad}, r_{bc}, r_{bd} and r_{cd}. They form six 1D kernel basis functions (6×1^2) for the total kernel polynomial, i.e., $k(r_{\mathrm{ab}}, r'_{\mathrm{ab}}), k(r_{\mathrm{ac}}, r'_{\mathrm{ac}}), \cdots, k(r_{\mathrm{cd}}, r'_{\mathrm{cd}})$.

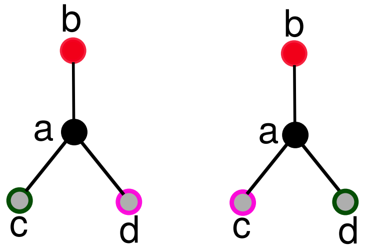

Figure S1: The two permutationally invariant configurations of CH_2O molecule, configuration 1 (left) and configuration 2 (right). Atoms are represented by color, black (b) is for the carbon atom, red (r) is for the oxygen atom, and gray with green border (g) and gray with magenta border (m) for the hydrogen atoms. Positions of the atoms are denoted by 'a', 'b', 'c', and 'd'.

Table S1: Symmetrization order of interatomic distances for two equivalent CH_2O configurations. Interatomic distances between two different atoms/positions are $r_{ij} = r_{ji}$. Atom positions and color indices are defined in Figure S1.

Configuration	r_{ab}	r_{ac}	r_{ad}	r_{bc}	r_{bd}	r_{cd}
1	r_{br}	r_{bg}	r_{bm}	r_{rg}	r_{rm}	r_{gm}
2	r_{br}	r_{bm}	r_{bg}	r_{rm}	r_{rg}	r_{mg}

In the following, the 1D kernels used for the 2-, 3-, and 4-body terms are $k^{[3,5]}$, $k^{[3,1]}$, and $k^{[3,0]}$, see main text. Then the unsymmetrized 2-body interaction energy of CH_2O can be expressed as a sum of these 1D kernel functions,

$$K_{\mathrm{2b}}(\mathbf{r}, \mathbf{r}') = k^{[3,5]}(r_{\mathrm{ab}}, r'_{\mathrm{ab}}) + k^{[3,5]}(r_{\mathrm{ac}}, r'_{\mathrm{ac}}) + k^{[3,5]}(r_{\mathrm{ad}}, r'_{\mathrm{ad}}) + k^{[3,5]}(r_{\mathrm{bc}}, r'_{\mathrm{bc}}) + \\ k^{[3,5]}(r_{\mathrm{bd}}, r'_{\mathrm{bd}}) + k^{[3,5]}(r_{\mathrm{cd}}, r'_{\mathrm{cd}}).$$

There are four (${}^4C_3$) 3-body interactions in the CH_2O molecule. Each 3-body interaction energy is a by product of three 1D kernel functions. Without considering symmetry the

3-body interaction energies are

$$
\begin{aligned}
K_{\mathrm{3b}}(\mathbf{r},\mathbf{r}') = k^{[3,1]}(r_{\mathrm{ab}},r'_{\mathrm{ab}}) \times k^{[3,1]}(r_{\mathrm{ac}},r'_{\mathrm{ac}}) \times k^{[3,1]}(r_{\mathrm{bc}},r'_{\mathrm{bc}}) \\
+ k^{[3,1]}(r_{\mathrm{ab}},r'_{\mathrm{ab}}) \times k^{[3,1]}(r_{\mathrm{ad}},r'_{\mathrm{ad}}) \times k^{[3,1]}(r_{\mathrm{bd}},r'_{\mathrm{bd}}) \\
+ k^{[3,1]}(r_{\mathrm{ac}},r'_{\mathrm{ac}}) \times k^{[3,1]}(r_{\mathrm{ad}},r'_{\mathrm{ad}}) \times k^{[3,1]}(r_{\mathrm{cd}},r'_{\mathrm{cd}}) \\
+ k^{[3,1]}(r_{\mathrm{bc}},r'_{\mathrm{bc}}) \times k^{[3,1]}(r_{\mathrm{bd}},r'_{\mathrm{bd}}) \times k^{[3,1]}(r_{\mathrm{cd}},r'_{\mathrm{cd}}).
\end{aligned}
$$

There is only one (${}^4C_4$) 4-body interaction in the $\mathrm{CH_2O}$ molecule. Each 4-body interaction energy can be written as a product of six 1D kernel functions. Without considering symmetry the 4-body interaction energies can be written as

$$
\begin{aligned}
K_{\mathrm{4b}}(\mathbf{r},\mathbf{r}') = k^{[3,0]}(r_{\mathrm{ab}},r'_{\mathrm{ab}}) \times k^{[3,0]}(r_{\mathrm{ac}},r'_{\mathrm{ac}}) \times k^{[3,0]}(r_{\mathrm{ad}},r'_{\mathrm{ad}}) \times k^{[3,0]}(r_{\mathrm{bc}},r'_{\mathrm{bc}}) \times \\
k^{[3,0]}(r_{\mathrm{bd}},r'_{\mathrm{bd}}) \times k^{[3,0]}(r_{\mathrm{cd}},r'_{\mathrm{cd}}).
\end{aligned}
$$

2-, 3- and 4-body Terms With Symmetry

If two fold symmetry is included, there are four types of interatomic distances in the $\mathrm{CH_2O}$ molecule i.e. one CO, two CH, two OH and one HH. They form 10 (1^2+2^2+2^2+1^2) 1D kernel basis functions for the total kernel polynomial, and include [$k(r_{\mathrm{br}},r'_{\mathrm{br}})$, $k(r_{\mathrm{bg}},r'_{\mathrm{bg}})$, $k(r_{\mathrm{bm}},r'_{\mathrm{bm}})$, $k(r_{\mathrm{rg}},r'_{\mathrm{rg}})$, $k(r_{\mathrm{rm}},r'_{\mathrm{rm}})$, $k(r_{\mathrm{gm}},r'_{\mathrm{gm}})$] for the CO, CH, OC, and HH distances of configuration 1, and [$k(r_{\mathrm{br}},r'_{\mathrm{br}})$, $k(r_{\mathrm{bg}},r'_{\mathrm{bm}})$, $k(r_{\mathrm{bm}},r'_{\mathrm{bg}})$, $k(r_{\mathrm{rg}},r'_{\mathrm{rm}})$, $k(r_{\mathrm{rm}},r'_{\mathrm{rg}})$, $k(r_{\mathrm{gm}},r'_{\mathrm{mg}})$] for configuration 2. It should be noted that there are 12 kernel functions (see Table S1) of which $k(r_{\mathrm{br}},r'_{\mathrm{br}})$ and $k(r_{\mathrm{gm}},r'_{\mathrm{gm}})$ appear twice.

The symmetrized 2-body interaction energy of $\mathrm{CH_2O}$ is then a sum of these 12 1D kernel functions

$$K_{2\mathrm{b}}(\mathbf{r},\mathbf{r}') = 2\times k^{[3,5]}(r_{\mathrm{br}},r'_{\mathrm{br}}) + k^{[3,5]}(r_{\mathrm{bg}},r'_{\mathrm{bg}}) + k^{[3,5]}(r_{\mathrm{bm}},r'_{\mathrm{bm}}) + k^{[3,5]}(r_{\mathrm{rg}},r'_{\mathrm{rg}}) + k^{[3,5]}(r_{\mathrm{rm}},r'_{\mathrm{rm}}) + k^{[3,5]}(r_{\mathrm{bg}},r'_{\mathrm{bm}}) + k^{[3,5]}(r_{\mathrm{bm}},r'_{\mathrm{bg}}) + k^{[3,5]}(r_{\mathrm{rg}},r'_{\mathrm{rm}}) + k^{[3,5]}(r_{\mathrm{rm}},r'_{\mathrm{rg}}) + 2\times k^{[3,5]}(r_{\mathrm{gm}},r'_{\mathrm{gm}})$$

For each equivalent configuration there are four (${}^4C_3$) 3-body interactions in the CH_2O molecule. Considering symmetry the 3-body interaction energy is

$$\begin{aligned}
K_{3\mathrm{b}}(\mathbf{r},\mathbf{r}') = {} & k^{[3,1]}(r_{\mathrm{br}},r'_{\mathrm{br}})\times k^{[3,1]}(r_{\mathrm{bg}},r'_{\mathrm{bg}})\times k^{[3,1]}(r_{\mathrm{rg}},r'_{\mathrm{rg}}) \\
& + k^{[3,1]}(r_{\mathrm{br}},r'_{\mathrm{br}})\times k^{[3,1]}(r_{\mathrm{bg}},r'_{\mathrm{bm}})\times k^{[3,1]}(r_{\mathrm{rg}},r'_{\mathrm{rm}}) \\
& + k^{[3,1]}(r_{\mathrm{br}},r'_{\mathrm{br}})\times k^{[3,1]}(r_{\mathrm{bm}},r'_{\mathrm{bg}})\times k^{[3,1]}(r_{\mathrm{rm}},r'_{\mathrm{rg}}) \\
& + k^{[3,1]}(r_{\mathrm{br}},r'_{\mathrm{br}})\times k^{[3,1]}(r_{\mathrm{bm}},r'_{\mathrm{bm}})\times k^{[3,1]}(r_{\mathrm{rm}},r'_{\mathrm{rm}}) \\
& + k^{[3,1]}(r_{\mathrm{bg}},r'_{\mathrm{bg}})\times k^{[3,1]}(r_{\mathrm{bm}},r'_{\mathrm{bm}})\times k^{[3,1]}(r_{\mathrm{gm}},r'_{\mathrm{gm}}) \\
& + k^{[3,1]}(r_{\mathrm{bg}},r'_{\mathrm{bm}})\times k^{[3,1]}(r_{\mathrm{bm}},r'_{\mathrm{bg}})\times k^{[3,1]}(r_{\mathrm{gm}},r'_{\mathrm{gm}}) \\
& + k^{[3,1]}(r_{\mathrm{rg}},r'_{\mathrm{rg}})\times k^{[3,1]}(r_{\mathrm{rm}},r'_{\mathrm{rm}})\times k^{[3,1]}(r_{\mathrm{gm}},r'_{\mathrm{gm}}) \\
& + k^{[3,1]}(r_{\mathrm{rg}},r'_{\mathrm{rm}})\times k^{[3,1]}(r_{\mathrm{rm}},r'_{\mathrm{rg}})\times k^{[3,1]}(r_{\mathrm{gm}},r'_{\mathrm{gm}})
\end{aligned}$$

Finally, the two fold symmetry, 4-body interaction energies can be written as

$$\begin{aligned}
K_{4\mathrm{b}}(\mathbf{r},\mathbf{r}') = {} & k^{[3,0]}(r_{\mathrm{br}},r'_{\mathrm{br}})\times k^{[3,0]}(r_{\mathrm{bg}},r'_{\mathrm{bg}})\times k^{[3,0]}(r_{\mathrm{bm}},r'_{\mathrm{bm}})\times k^{[3,0]}(r_{\mathrm{rg}},r'_{\mathrm{rg}})\times \\
& k^{[3,0]}(r_{\mathrm{rm}},r'_{\mathrm{rm}})\times k^{[3,0]}(r_{\mathrm{gm}},r'_{\mathrm{gm}}) \\
& + k^{[3,0]}(r_{\mathrm{br}},r'_{\mathrm{br}})\times k^{[3,0]}(r_{\mathrm{bg}},r'_{\mathrm{bm}})\times k^{[3,0]}(r_{\mathrm{bm}},r'_{\mathrm{bg}})\times k^{[3,0]}(r_{\mathrm{rg}},r'_{\mathrm{rm}})\times \\
& k^{[3,0]}(r_{\mathrm{rm}},r'_{\mathrm{rg}})\times k^{[3,0]}(r_{\mathrm{gm}},r'_{\mathrm{gm}})
\end{aligned}$$

CH_4 RKHS PES

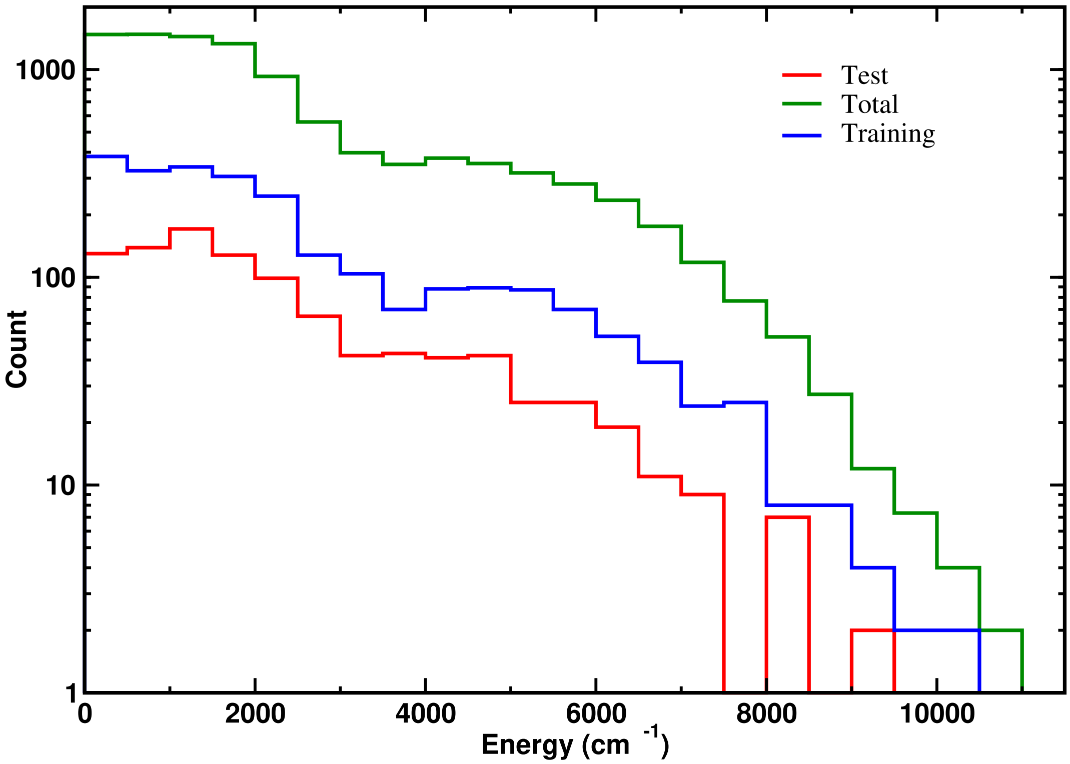

Figure S2: Distribution of the reference data set for CH_4. Distribution of all 10000 reference energies (green) along with 2400 training energies (blue) and 1000 test energies (red).

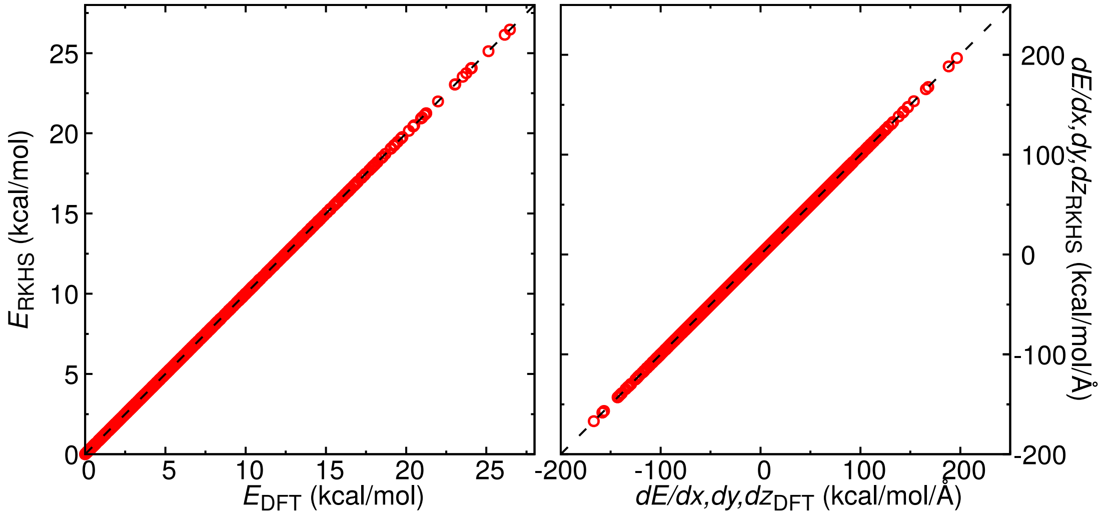

Figure S3: Correlation between reference (DFT) and prediction by the RKHS-PES for energies (left) and gradients (right) for CH_4 for $N_{\mathrm{test}} = 1000$. The RMSE, MAE and $(1 - R^2)$ for the energies are 0.0018, 0.0013 kcal/mol and 9×10^{-8}, respectively. For the gradients the RMSE, MAE and $(1 - R^2)$ are 0.0098, 0.0048 kcal/mol/Å and 5×10^{-7}, respectively.

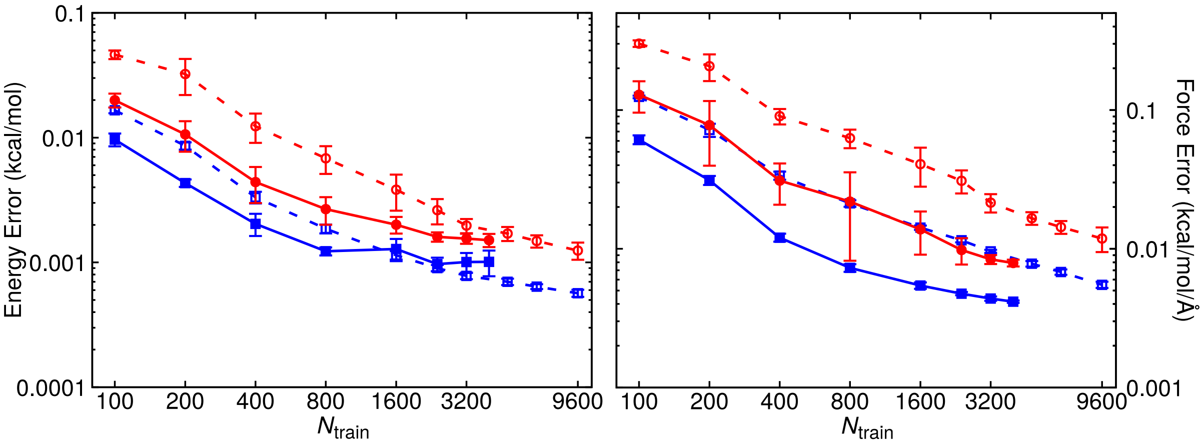

Figure S4: Energy (left) and force (right) learning curve for CH_4 for "energy only" (dashed lines and open symbols) and "energy+force" (solid lines and filled symbols). For a given number N_{ref} each model is generated five times using randomly drawn reference data. Average values and standard deviations (error bars) of the RMSE (red) and MAE (blue) are shown, respectively.