Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels

Pavlo O. Dral, Alec Owens, Sergei N. Yurchenko, and Walter Thiel

Citation: The Journal of Chemical Physics **146**, 244108 (2017); doi: 10.1063/1.4989536
View online: http://dx.doi.org/10.1063/1.4989536
View Table of Contents: http://aip.scitation.org/toc/jcp/146/24
Published by the American Institute of Physics

THE JOURNAL OF CHEMICAL PHYSICS **146**, 244108 (2017)

Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels

Pavlo O. Dral,[1,a)] Alec Owens,[1,2] Sergei N. Yurchenko,[2] and Walter Thiel[1]
[1]*Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany*
[2]*Department of Physics and Astronomy, University College London, Gower Street, WC1E 6BT London, United Kingdom*

(Received 2 February 2017; accepted 9 June 2017; published online 27 June 2017)

We present an efficient approach for generating highly accurate molecular potential energy surfaces (PESs) using self-correcting, kernel ridge regression (KRR) based machine learning (ML). We introduce structure-based sampling to automatically assign nuclear configurations from a pre-defined grid to the training and prediction sets, respectively. Accurate high-level *ab initio* energies are required only for the points in the training set, while the energies for the remaining points are provided by the ML model with negligible computational cost. The proposed sampling procedure is shown to be superior to random sampling and also eliminates the need for training several ML models. Self-correcting machine learning has been implemented such that each additional layer corrects errors from the previous layer. The performance of our approach is demonstrated in a case study on a published high-level *ab initio* PES of methyl chloride with 44 819 points. The ML model is trained on sets of different sizes and then used to predict the energies for tens of thousands of nuclear configurations within seconds. The resulting datasets are utilized in variational calculations of the vibrational energy levels of CH_3Cl. By using both structure-based sampling and self-correction, the size of the training set can be kept small (e.g., 10% of the points) without any significant loss of accuracy. In *ab initio* rovibrational spectroscopy, it is thus possible to reduce the number of computationally costly electronic structure calculations through structure-based sampling and self-correcting KRR-based machine learning by up to 90%. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4989536]

I. INTRODUCTION

Molecular potential energy surfaces (PESs) are often computed using high-level *ab initio* theory, notably in theoretical rovibrational spectroscopy where sub-wavenumber accuracy is targeted for calculated transition frequencies. Such sophisticated quantum mechanical (QM) calculations can be extremely costly, and they have to be repeated for tens of thousands of nuclear geometries. Particularly challenging are larger systems with many degrees of freedom that suffer from the so-called "curse of dimensionality." Machine learning (ML) techniques offer a means to significantly reduce the necessary computational time.[1]

It has been shown that neural networks (NNs) can be used to interpolate between points along the coordinate of simple reactions,[2,3] to improve the accuracy of semiempirical quantum mechanics/molecular mechanics calculations,[4] and to calculate accurate energies of conformations and clusters of organic and inorganic species.[5–9] Kernel ridge regression (KRR) can reduce errors of low-level QM methods such that atomization enthalpies of organic isomers can be reproduced with an accuracy close to higher-level QM methods.[10,11] Gaussian process models are able to sample low-energy regions of the PES,[12] correct density functional theory energies of water,[13] predict energies and forces in Si_n clusters,[14] and construct global PESs for N_4.[15] ML can also be employed for accurate yet computationally fast PES representation; for example, NNs have been used to compute molecular vibrational energy levels,[16–20] reaction probabilities,[20] and reaction rates,[21] whilst potentials based on KRR,[22] NNs,[23–34] or Gaussian process models[35] have been utilized in molecular dynamics (MD) simulations.

ML performs best in the interpolation regime and can fail spectacularly when used for extrapolation. For relatively large training sets and high error tolerance, as in many chemical applications, training on a random sample of data points can give useful results.[10,11,15,36,37] Random sampling is straightforward and has minimal computational cost involved but it does not ensure *per se* that ML avoids extrapolation. Thus, it is preferable to ensure that ML is applied only for interpolation by using better sampling techniques. Even a simple manual choice of the critical points on the PES can improve the accuracy of ML significantly.[2]

Automated sampling is necessary for more complex PESs and several approaches have been reported in the literature. When training ML potentials, structures have been screened on their input vector elements to see if they are within the minimum and maximum values of the training set, which can then be expanded to include "outside-the-range" structures.[31,38] NN PESs of simple reactions have been re-fitted with additional points when the initial NN potential has

[a)]Electronic mail: dral@kofo.mpg.de

encountered problems.[3] Similarly, the error of the ML potential can be monitored during dynamics and ML can be retrained on-the-fly by including additional points when errors become too large.[39] In the case of very simple systems like H_3, the training set can be generated on a coarse grid.[3]

Because of their inherent flexibility, different NNs can give very different energies for the same structure, and hence, training different NNs on the same data can also highlight problem structures that should be incorporated into the training set.[7,33] Another approach exploits the fact that the most important regions of the PES are likely to be covered by standard MD simulations. Snapshots from the MD trajectories can thus be used in the training set to produce a more robust ML PES.[8,22,25,28,40,41] For better sampling of vibrational phase space, previous studies have employed stretched bond distances in MD simulations[28] or a random displacement to sample along the vibrational normal modes.[9] In other work, structures have been chosen from MD trajectories based on various clustering approaches, which reduce the redundancies in the NN training set.[42] For calculations of rotation-vibration spectra where low-energy regions of the PES are more important, probability distribution functions that favor structures near equilibrium have been used for pseudorandom sampling.[16–18]

Here, we report a new and efficient method for generating highly accurate molecular PESs which uses KRR-based machine learning. In our approach, we initially define a set of grid points (nuclear configurations) that cover the relevant low-energy part of the PES, which can be done using relatively inexpensive energy estimates (see Sec. II C). Thereafter, our sampling relies on interpolation between structures rather than energies, i.e., the sampling procedure itself does not require energy calculations or the training of several ML models. We investigate the effect of structure-based sampling on the accuracy of the ML model in comparison with random sampling. We also explore what error reduction can be achieved through ML self-correction.

To illustrate our KRR-based ML model, we utilize a high-level *ab initio* PES for methyl chloride ($CH_3{}^{35}Cl$).[43] This PES, denoted CBS-35^{HL}, is based on extensive explicitly correlated coupled cluster calculations with extrapolation to the complete basis set (CBS) limit and incorporates a range of higher-level additive energy corrections. These include core-valence electron correlation, higher-order coupled cluster terms, scalar relativistic effects, and diagonal Born-Oppenheimer corrections. A considerable amount of computational time was spent generating this PES. For example, a single point of the CBS-35^{HL} PES at the equilibrium geometry required 9 separate calculations and a combined total of 26.7 h on a single core of an Intel Xeon E5-2690 v2 3.0 GHz processor. Building a reliable PES requires tens of thousands of nuclear geometries for polyatomic species, and the computational time increases for distorted configurations due to slower energy convergence. Efficient ML techniques that can reduce the necessary computational effort are therefore highly desirable.

The paper is structured as follows: In Sec. II, we present the KRR-based self-correcting ML model, the molecular descriptor, and the sampling procedure. Details on the variational calculation of vibrational energy levels are also given. In Sec. III, the ML model is evaluated for different sizes of the training set, and vibrational energies are computed using several ML model PESs. Concluding remarks are offered in Sec. IV and an outlook is provided in Sec. V.

II. METHODS

A. Machine learning

The chosen ML technique is based on KRR[44] and is similar to the approach used for ML across chemical compound space[11,36,37,45,46] and for calculating relative stabilities of organic isomers.[10,11] Below, we give details of the KRR approach relevant to this work. All ML calculations were performed using our own program package `MLatom`.[47]

Some property value $Y^{\mathrm{ML}}(\mathbf{M}_i)$ of a nuclear configuration i represented by the molecular descriptor $\mathbf{M}_i$ (discussed in Sec. II B) is calculated using the expression,[44]

$$Y^{\mathrm{ML}}(\mathbf{M}_i) = \sum_{j=1}^{N_{\mathrm{train}}} \alpha_j K\left(\mathbf{M}_i, \mathbf{M}_j\right), \tag{1}$$

where the sum runs over N_{train} configurations represented by the molecular descriptors $\{\mathbf{M}_j\}$ of the training set. The regression coefficient α_j refers to the configuration $\mathbf{M}_j$ and K is the kernel. Here we use the Gaussian kernel which is equivalent to the form often employed in the literature,[36,37]

$$K\left(\mathbf{M}_i, \mathbf{M}_j\right) = \exp\left(-\frac{D\left(\mathbf{M}_i, \mathbf{M}_j\right)^2}{2\sigma^2}\right), \tag{2}$$

where σ is the kernel width, and $D\left(\mathbf{M}_i, \mathbf{M}_j\right)$ is the Euclidean distance (the L_2 norm) between descriptors $\mathbf{M}_i$ and $\mathbf{M}_j$ (vectors of size N_M) defined as

$$D\left(\mathbf{M}_i, \mathbf{M}_j\right) = \sqrt{\sum_{a}^{N_M} \left(M_{i,a} - M_{j,a}\right)^2}. \tag{3}$$

This definition of distances between molecular structures[36,37,45] and a similar definition based on the L_1 norm (instead of the L_2 norm used here)[10,11,37,46,48–51] were used previously to compare molecular geometries[48] and learn various molecular properties.[10,11,36,37,45,46,49–51] We have also tested the Laplacian kernel employing the L_1 norm, but the PESs obtained with this kernel showed much larger errors than those obtained with the Gaussian kernel.

Training the ML model involves finding the regression coefficients $\boldsymbol{\alpha}$, i.e., solving the minimization problem,

$$\min_{\alpha} \sum_{j}^{N_{\mathrm{train}}} \left[Y^{\mathrm{ML}}\left(\mathbf{M}_j\right) - Y^{\mathrm{ref}}\left(\mathbf{M}_j\right)\right] + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}, \tag{4}$$

which has a known analytical solution[44]

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y}^{\mathrm{ref}}, \tag{5}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{K}$ is the kernel matrix with the elements calculated using Eq. (2) for all pairs of the training set points, $\mathbf{Y}^{\mathrm{ref}}$ is a vector with reference property values, and λ is the so-called regularization parameter that ensures the transferability of the model to new nuclear configurations.[36,45]

Two additional improvements were made to the approach outlined above which were not considered in earlier work:[10,11,36,37,45] (i) we sample data to ensure that ML interpolates between the training points and does not extrapolate beyond them (see Sec. II C), and (ii) we use a nested, self-correcting ML approach. For the latter, ML is first trained on the reference *ab initio* energies $\mathbf{E}^{\text{ref}}$ [$\mathbf{Y}^{\text{ref}}$ in Eq. (5)] and then used to make a first-layer estimate of the deformation energy $E^{\text{layer 1}}$ [Y^{ML} in Eq. (1)]. The next layer corrects errors of the energies estimated in the previous layer, and so on. For example, the second layer ML model is trained on $\mathbf{\Delta E}^{\text{ref, layer 1}} = \mathbf{E}^{\text{ref}} - \mathbf{E}^{\text{layer 1}}$ and is then used to calculate $\mathbf{\Delta E}^{\text{ML, layer 1}}$ corrections for the prediction set which are summed up with layer 1 predicted energies to obtain layer 2 energies, i.e., $\mathbf{E}^{\text{layer 2}} = \mathbf{\Delta E}^{\text{ML, layer 1}} + \mathbf{E}^{\text{layer 1}}$. The dependence of the performance of the ML model on the number of layers is discussed in Sec. III A.

In order to determine the optimal values of the hyperparameters σ and λ for each layer, we sample 80% of the training set points into a sub-training set using the same sampling procedure that was employed for the training set (see Sec. II C). The points with deformation energies less than 10 000 cm^{-1} are taken from the remaining 20% into the validation set. Using the ML model trained on the sub-training set, we search for values of the hyperparameters which give the lowest root-mean-square error (RMSE) for the validation set. This is performed using a simple logarithmic grid search.[37,52] By optimizing the hyperparameters such as to obtain a better description below 10 000 cm^{-1}, we ensure an adequate treatment of the spectroscopically most relevant part of the PES.[53]

The above self-correcting scheme is similar in spirit to a two-stage NN model that has been introduced to remedy peculiarities of NNs,[17,18] such as the presence of ill-defined regions ("holes") in NN PESs as a result of overfitting. In this previous work, the NN potential was first fit with as few nodes as possible to eliminate holes, which resulted in a large RMSE; to obtain a PES of reasonable quality, the NN potential was then re-fit by incorporating more nodes to eliminate residual errors.

In our study, overfitting is primarily prevented by using the regularization parameter λ, and the problem of holes on the PES is addressed by our sampling procedure (discussed in Sec. II C). The KRR approach makes fitting straightforward as there is no need to manually adjust any of the parameters. As a result, the RMSE of our one-layer ML model is already quite low, while multi-layer ML models possess still slightly smaller RMSEs (see Sec. III A). This is in contrast to two-stage NN models where the RMSE of the first fit is one order of magnitude larger than that of the second fit.[17]

B. Molecular descriptor

The success of ML is largely determined by the choice of an appropriate molecular descriptor. Many of the molecular descriptors proposed in the literature are functions of the internuclear distances, for example, the Coulomb matrix,[36,37] the Bag of Bonds,[54] the atom-centered symmetry functions,[1,26] the bispectrum of the neighbor density,[35] the smooth overlap of atomic positions,[14] and others.[19] We designed and tested many descriptors for CH_3Cl, but overall the most accurate was a vector with ten elements corresponding to the ten interatomic pairs. Each element is defined as the corresponding internuclear distance in the near-equilibrium reference geometry of CH_3Cl (r^{eq}) divided by the current distance (r), e.g., $r^{\text{eq}}_{\text{C–Cl}}/r_{\text{C–Cl}}$ for the C–Cl atomic pair. This form of descriptor ensures that the ML model is rotationally and translationally invariant.

Since the molecular descriptor also has to be atom index invariant, we sort the hydrogen nuclei by the sum of their internuclear repulsions with the four other nuclei for structure-based sampling (see Sec. II C). Simple sorting of hydrogen nuclei in the molecular descriptor may however lead to instabilities in regions where the order of hydrogen nuclei changes. To avoid this problem, we employ a normalized permutational invariant kernel in our ML calculations (Sec. II A) as suggested in the literature,[13,55]

$$\overline{K}\left(\mathbf{M}_i, \mathbf{M}_j\right) = \frac{\sum_{\hat{P}}^{N_{perm}} K\left(\mathbf{M}_i, \hat{P}\mathbf{M}_j\right)}{\sqrt{\sum_{\hat{P}}^{N_{perm}} K\left(\mathbf{M}_i, \hat{P}\mathbf{M}_i\right)}\sqrt{\sum_{\hat{P}}^{N_{perm}} K\left(\mathbf{M}_j, \hat{P}\mathbf{M}_j\right)}}, \tag{6}$$

where $\hat{P}$ permutes the order of hydrogen nuclei. There are $N_{perm} = 3! = 6$ permutations of three hydrogen nuclei. We found that results obtained using the normalized permutational invariant kernel are superior to those obtained using a sorted molecular descriptor (see Sec. III A).

C. Grid points and sampling

The target PES needs to be evaluated on a large number of pre-defined grid points (nuclear configurations). These grid points can be determined through rather inexpensive initial energy computations. In our previous work on CH_3Cl, this was done by creating a sparse grid along each one-dimensional cut of the PES, calculating single-point energies using a reliable but relatively cheap *ab initio* method [e.g., CCSD(T)-F12b/VTZ-F12], fitting a polynomial function to get rough energy estimates, and randomly selecting all the remaining points using an energy-weighted Monte Carlo type sampling algorithm to cover the desired low-energy PES region.[43] This procedure, which was inexpensive computationally, produced a grid of 44 819 geometries with energies up to $hc \cdot 50\,000$ cm^{-1} (h is the Planck constant and c is the speed of light). In our present proof-of-principle study, we use this grid, which was employed for the final CBS-35^{HL} PES of CH_3Cl,[43] and partition it into training and test sets using the structure-based sampling procedure described below; the energies for the training set are taken from the available *ab initio* results,[43] while those for the remaining grid points are predicted essentially for free using ML.

In our sampling procedure, we select nuclear configurations for the training set based on the relative distance of the molecular descriptors. This is done such that all the remaining points used for prediction are within the boundaries of the training set or very close to it. The first point of the training set is taken closest to equilibrium. The second point of the training set is the one among all grid points that has the largest distance from the first point. For each remaining point on the grid, we calculate its distance to all points in the training set using Eq. (3) and determine the shortest distance to any point

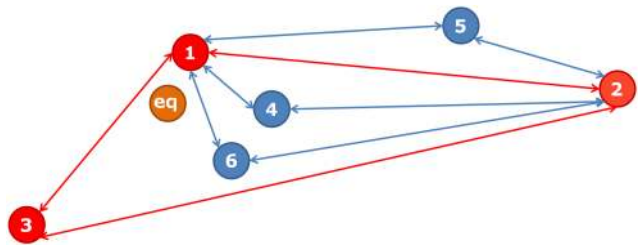

FIG. 1. Illustration of sampling 3 points into the training set out of 6 points in two-dimensional space. Point 1 is closest to equilibrium, while point 2 is farthest apart from point 1, so they are both included into the training set. Point 3 has the shortest distance to point 1 of the current training set, which is longer than any shortest distance of points 4, 5, and 6 to the points in the current training set (1 and 2). Thus, point 3 is included into the training set (red), while the remaining points 4–6 are left for the prediction set (blue).

in the training set. We then include the grid point that has the "longest" shortest distance into the training set, as illustrated in Fig. 1. This procedure is repeated until the required number of points is selected for the training set. The other remaining points are used for prediction, and by construction they lie within the training set or very close to it; at least one of their distances to the points in the training set should be shorter than the shortest distance between the points in the training set.

This sampling procedure is closely related to the farthest-point traversal iterative procedure used to select points such that they are as distant as possible from the previously selected points. In this respect, structure-based sampling can also be viewed as a way to obtain a training set as diverse as possible.

The sampling procedure outlined above selects a training set of predetermined size from a larger set of predefined structures. The same sampling principles can be applied to test whether additional structures (beyond the initially chosen set) should be included into the training set. If this is the case, the ML model needs to be re-trained (similar to an approach described in Ref. 31).

D. Variational calculations

In this work, we use the nuclear motion program TROVE[56,57] for computing vibrational energy levels. Since rovibrational calculations have previously been reported for CH_3Cl,[43,58] we summarize only the key aspects relevant for the present study.

In variational calculations, the PES must be represented analytically. To do this, we introduce the coordinates,

$$\xi_1 = 1 - \exp\left(-a(r_0 - r_0^{\mathrm{eq}})\right), \tag{7}$$

$$\xi_j = 1 - \exp\left(-b(r_i - r_1^{\mathrm{eq}})\right), \quad j = 2,3,4, \quad i = j - 1, \tag{8}$$

where $a = 1.65\,\text{Å}^{-1}$ for the C–Cl bond length r_0, and $b = 1.75\,\text{Å}^{-1}$ for the three C–H bond lengths r_1, r_2, and r_3. For the angular terms,

$$\xi_k = (\beta_i - \beta^{\mathrm{eq}}), \quad k = 5,6,7, \quad i = k - 4, \tag{9}$$

$$\xi_8 = \frac{1}{\sqrt{6}}(2\tau_{23} - \tau_{13} - \tau_{12}), \tag{10}$$

$$\xi_9 = \frac{1}{\sqrt{2}}(\tau_{13} - \tau_{12}), \tag{11}$$

where β_1, β_2, and β_3 are the $\angle(\mathrm{H}_i\mathrm{CCl})$ interbond angles, and τ_{12}, τ_{13}, and τ_{23} are the dihedral angles between adjacent planes containing $\mathrm{H}_i\mathrm{CCl}$ and $\mathrm{H}_j\mathrm{CCl}$. Here r_0^{eq}, r_1^{eq}, and β^{eq} are the reference equilibrium structural parameters of CH_3Cl.

The potential function (maximum expansion order $i + j + k + l + m + n + p + q + r = 6$) is given by the expression,

$$V(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6, \xi_7, \xi_8, \xi_9) = \sum_{ijk\ldots} \mathrm{f}_{ijk\ldots} V_{ijk\ldots} \tag{12}$$

and contains the terms,

$$V_{ijk\ldots} = \{\xi_1^i \xi_2^j \xi_3^k \xi_4^l \xi_5^m \xi_6^n \xi_7^p \xi_8^q \xi_9^r\}^{C_{3\mathrm{v}}(\mathrm{M})}, \tag{13}$$

which are symmetrized combinations of different permutations of the vibrational coordinates ξ_i and transform according to the A_1 representation of the $C_{3\mathrm{v}}(\mathrm{M})$ molecular symmetry group.[59]

The $\mathrm{f}_{ijk\ldots}$ expansion coefficients are determined through a least-squares fitting to the *ab initio* or ML data. Weight factors of the form[60]

$$w(E_i) = \left(\frac{\tanh\left[-0.000\,6 \times (E_i - 15\,000)\right] + 1.002\,002\,002}{2.002\,002\,002}\right) \times \frac{1}{NE_i^{(w)}} \tag{14}$$

are used in the fittings. Here $E_i^{(w)} = \max(E_i, 10\,000)$ and the normalization constant $N = 0.0001$ (all values in cm^{-1}). Larger weights (w) are assigned to lower deformation energies (E_i), which correspond to more spectroscopically important regions of the PES. As shown in Fig. 2, this form ensures that structures with energies up to $10\,000\ \mathrm{cm}^{-1}$ above equilibrium have weights near unity, whilst other configurations are significantly downweighted with increasing energy.

To ensure a reliable comparison, the CBS-35^{HL} PES expansion parameter set was used to fit the *ab initio* and ML-generated datasets. This contained 414 parameters and included linear expansion terms. For this reason, we fixed the values of $r_0^{\mathrm{eq}} = 1.7775$ Å, $r_1^{\mathrm{eq}} = 1.0837$ Å, and $\beta^{\mathrm{eq}} = 108.445°$ to be the same as those used for the CBS-35^{HL} PES. Each fit employed Watson's robust fitting scheme,[61] which reduces the

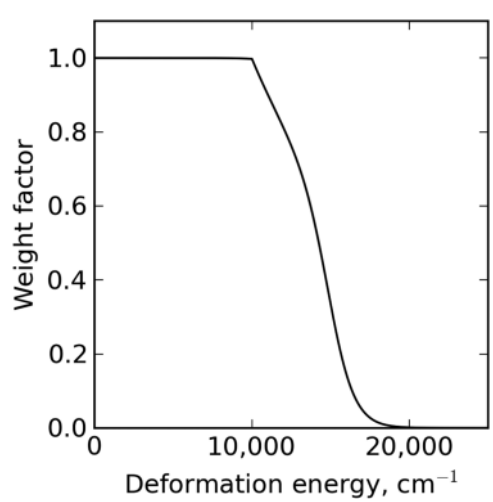

FIG. 2. Decay of unitless weight factors calculated using Eq. (14) with increasing deformation energy in cm^{-1}.

weights of outliers and lessens their influence in determining the final set of parameters.

III. RESULTS AND DISCUSSION

In this section, we investigate the accuracy of the ML models trained on 50%, 25%, and 10% of the available 44 819 deformed structures and their *ab initio* energies used to construct the CBS-35^{HL} PES.[43]

To evaluate the accuracy of predicted energies (E^{est}), we employ standard error measures as well as the weighted root-mean-square error (wRMSE),

$$\mathrm{wRMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(E_i^{\mathrm{est}} - E_i^{\mathrm{ref}}\right)^2 w\left(E_i^{\mathrm{ref}}\right)}. \qquad (15)$$

Weights $w\left(E_i^{\mathrm{ref}}\right)$ are calculated using Eq. (14).

A. Optimal machine learning model

We first examine how the accuracy of ML calculated deformation energies for the prediction set depends on the number of ML layers and the sampling procedure. The results in Table I indicate that the errors are significantly lowered by adding the second layer for all models except for the one trained on 10% of points selected with structure-based sampling. Including a third layer significantly reduces the errors in a few cases, while a fourth layer does not yield any noticeable improvements. We therefore expect that adding further layers will make no difference.

For practical purposes, the three-layer model appears to be sufficient. However, in the following, we have applied the four-layer model because the computational cost of including additional layers is rather low and we have not observed any significant accumulation of numerical noise. It takes around 4 h to optimize the ML hyperparameters using a fairly inefficient optimization algorithm, around a minute to train and only a couple of seconds to make predictions with the 50%-ML model on 20 cores of an Intel(R) Xeon(R) CPU E5-2687W v3 @ 3.10 GHz processor.

As shown in Table I, random sampling is clearly inferior to structure-based sampling; all four-layer ML models trained on randomly drawn points have wRMSEs significantly higher than those of the four-layer, structure-based sampling ML models. Interestingly, self-correction works well even for random sampling: it reduces the wRMSEs by 32%–44% and the standard deviations by 70%–90%. Despite this the remaining error of 4.13 ± 0.87 cm^{-1} for the randomly sampled, four-layer ML models trained on 50% of grid points are still much higher than the wRMSE of 0.37 cm^{-1} for the four-layer, structure-based sampling ML model trained on the same number of grid points. Standard deviations for random sampling are also relatively high and increase from 0.87 to 1.67 cm^{-1} when going from 50% to 10% training sets. There is no such problem with structure-based sampling which provides a unique training set—an important practical advantage for high-accuracy applications. As for the computational cost of sampling, it takes ca. 9 h to sample 50% from 44 819 data points on 12 Intel(R) Xeon(R) CPU X5680 @ 3.33 GHz processors.

As the training set becomes relatively small, structure-based sampling may under-represent training points in low-energy regions, e.g., 10% training points drawn from the entire grid contain only 195 structures with deformation energies below 10 000 cm^{-1} (compare the number of configurations with deformation energies below 1000 and 10 000 cm^{-1},

TABLE I. wRMSEs in deformation energies predicted by ML models trained on 50%, 25%, and 10% of the reference data for the remaining 50%, 75%, and 90% of the grid points. wRMSEs are calculated using Eq. (15) and are given in cm^{-1}. The number of training set structures with an *ab initio* deformation energy below 1000 cm^{-1} ($N_{<1000}$) and below 10 000 cm^{-1} ($N_{<10000}$) is also given for comparison.

			Number of layers			
N_{train}	$N_{<1000}$[a]	$N_{<10000}$[b]	1	2	3	4
Structure-based sampling from unsliced data						
22 409 (50%)	22	3985	6.57	0.37	0.37	0.37
11 204 (25%)	1	1033	3.21	1.87	1.60	1.60
4 481 (10%)	1	195	5.00	4.99	4.83	4.83
Structure-based sampling from data sliced into three regions						
22 408 (50%)	131	7215	2.31	0.62	0.62	0.62
11 203 (25%)	19	3348	2.90	2.59	2.58	2.58
4 480 (10%)	1	1191	4.42	3.63	3.63	3.63
Random sampling[c]						
22 409 (50%)	573 ± 12	7971 ± 53	6.07 ± 2.96	4.13 ± 0.87	4.13 ± 0.87	4.13 ± 0.87
11 204 (25%)	288 ± 13	4002 ± 36	8.41 ± 8.23	4.76 ± 0.90	4.75 ± 0.88	4.75 ± 0.88
4 481 (10%)	115 ± 12	1599 ± 35	14.73 ± 12.63	8.76 ± 1.67	8.73 ± 1.67	8.73 ± 1.67

[a] $N_{<1000} = 1145$ for the entire grid of 44 819 points.
[b] $N_{<10000} = 15\,935$ for the entire grid of 44 819 points.
[c] Standard deviations were calculated for 16, 20, and 30 various randomly drawn training sets for ML trained on 50%, 25%, and 10% of grid points, respectively.

columns $N_{<1000}$ and $N_{<10000}$ in Table I). Sorting the geometries by their distance to the equilibrium structure (which correlates strongly with the deformation energies, Fig. 3), followed by slicing the data into several regions and sampling points from each of these regions, can produce a more energy-balanced training set. Looking at Fig. 3, one can argue that splitting the set into three regions with an equal number of structures should be close to optimal. The first region includes the most important structures with deformation energies below 10 000 cm^{-1} and a significant portion of structures with energies between 10 000 and 20 000 cm^{-1}. The second region mainly includes configurations with energies between 10 000 and 20 000 cm^{-1} but also a considerable number of geometries above and below this region. The third slice includes all the remaining high-energy structures.

Structure-based sampling from each of the above regions gives a value of $N_{<10000}$ close to what is expected from random sampling, e.g., 10% of the training points drawn from the sliced grid contain 1191 structures with deformation energies below 10 000 cm^{-1} (Table I). As a result, the wRMSE of the ML model trained on the latter training set (3.63 cm^{-1}) is lower than the wRMSE of the ML model trained on 10% of the points drawn from unsliced data (4.83 cm^{-1}). However, such slicing does not generate training sets that are as diverse as possible, and therefore, the errors of the ML models become higher for the sliced grids as the training set increases in size (the training sets with 25% and 50% grid points, Table I). Thus, we recommend slicing only for very small training sets where low-energy structures are under-represented.

In the following, we will discuss four-layer ML models trained on 50%, 25%, and 10% of the points drawn using structure-based sampling from the available unsliced 44 819 grid points. We refer to these ML models as 50%-ML, 25%-ML, and 10%-ML, respectively. We also compare with one of the randomly sampled, four-layer ML models referred to as r50%-ML, which has been chosen at random from the ML models trained on 50% of grid points. In addition, we compare with the four-layer ML model trained on 10% of the points drawn using structure-based sampling from the dataset sliced into three regions. This model is referred to as s10%-ML in the following.

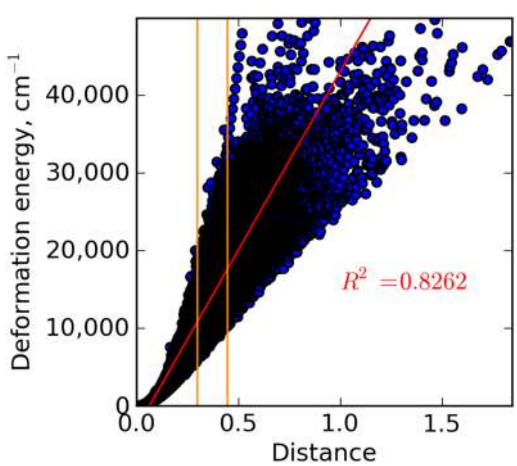

FIG. 3. Correlation between *ab initio* deformation energies in cm^{-1} and unitless distances to the near-equilibrium structure calculated using Eq. (3). A linear trend line is shown in red with its R^2 value (0.83). Orange vertical lines slice the data into three regions with equal numbers of data points in the training set. Each data point is represented by a blue dot with a black edge; hence, the most densely populated areas are black.

A more detailed analysis of the ML model errors, listed in Table II, reveals that r50%-ML has the largest outliers with a wRMSE of 4.14 cm^{-1}, which is more than twice as large as that of 25%-ML (wRMSE of 1.60 cm^{-1}) for their respective prediction sets. The non-weighted RMSE of r50%-ML (167.19 cm^{-1}) is more than four times higher than the non-weighted RMSE of 10%-ML (39.63 cm^{-1}). Moreover, the RMSE of r50%-ML for energies below 1000 cm^{-1} is practically the same as the respective RMSE of 50%-ML, but for energies below 10 000 cm^{-1}, the RMSE is higher than that of 50%-ML, despite the fact that many more points from this region are included into the training set (compare $N_{<1000}$ and $N_{<10000}$ for these two models in Table I). These observations provide strong evidence for the superiority of structure-based sampling. s10%-ML has RMSEs for energies below 1000 and

TABLE II. Number of grid points in the training (N_{train}) and prediction (N_{predict}) sets, largest positive outlier (LPO) and largest negative outlier (LNO), mean signed errors (MSEs), mean absolute errors (MAEs), RMSEs for entire sets (all), for structures with reference deformation energies below 1000 and 10 000 cm^{-1}, and wRMSEs in cm^{-1} for the training and prediction sets of the 50%-ML, r50%-ML, 25%-ML, 10%-ML, and s10%-ML models and the entire grid of 44 819 points.

	50%-ML	r50%-ML	25%-ML	10%-ML	s10%-ML
			Training set		
N_{train}	22 409	22 409	11 204	4 481	4 480
LPO	0.00	4.18	0.00	0.00	0.00
LNO	0.00	−4.14	0.00	0.00	0.00
MSE	0.00	0.00	0.00	0.00	0.00
MAE	0.00	0.03	0.00	0.00	0.00
RMSE (all)	0.00	0.11	0.00	0.00	0.00
RMSE (<10 000)	0.00	0.03	0.00	0.00	0.00
RMSE (<1 000)	0.00	0.07	0.00	0.00	0.00
wRMSE	0.00	0.03	0.00	0.00	0.00
			Prediction set		
N_{predict}	22 410	22 410	33 615	40 338	40 339
LPO	319.75	2 015.44	1 035.63	1 617.61	1 481.69
LNO	−476.20	−6 919.28	−1 060.38	−2 859.37	−2 190.33
MSE	0.02	−11.62	0.18	0.19	1.85
MAE	0.82	25.12	3.47	11.27	19.54
RMSE (all)	6.23	167.19	16.12	39.63	61.96
RMSE (<10 000)	0.20	1.38	1.12	4.26	1.19
RMSE (<1 000)	0.08	0.07	0.16	1.40	0.25
wRMSE	0.37	4.14	1.60	4.83	3.63
			Entire grid		
N		44 819			
LPO	319.75	2 015.44	1 035.63	1 617.61	1 481.69
LNO	−476.20	−6 919.28	−1 060.38	−2 859.37	−2 190.33
MSE	0.01	−5.81	0.14	0.17	1.66
MAE	0.41	12.57	2.60	10.14	17.58
RMSE (all)	4.41	118.22	13.96	37.60	58.79
RMSE (<10 000)	0.17	0.98	1.09	4.24	1.14
RMSE (<1 000)	0.08	0.07	0.16	1.40	0.25
wRMSE	0.26	2.93	1.39	4.59	3.44

10 000 cm^{-1} close to those of 25%-ML, while the respective RMSEs of 10%-ML are much higher. On the other hand, 10%-ML has a much lower non-weighted RMSE, mean absolute error (MAE), and MSE than s10%-ML. Thus, slicing clearly improves the description of the low-energy region at the cost of other regions. Apparently, for sparse training data, the benefits of a better description of the low-energy region achieved by slicing outweigh the disadvantage of an overall worse description of the PES, which is exemplified by the lower wRMSE of s10%-ML compared to the wRMSE of 10%-ML.

From Fig. 4, we see that deformation energies predicted by the 50%-ML, r50%-ML, 25%-ML, and s10%-ML models correlate nicely with the reference *ab initio* energies; the R^2 value is always larger than 0.999. Deformation energies predicted by the 10%-ML model (not shown in Fig. 4) correlate better ($R^2 = 0.999\,977$) with the reference energies than the energies predicted by the s10%-ML model (consistent with the above conclusions). Clearly, 50%-ML is superior to all other models and has the best correlation with far fewer outliers and smaller residual errors. This is particularly relevant for high-accuracy work as we will see in Sec. III B.

As for the effect of the training set size, illustrated in Fig. 5, wRMSEs in the prediction set drop from 144 cm^{-1} for the 1%-ML model to 0.05–0.06 cm^{-1} for the 85–99%-ML models. Interestingly, the 1%-ML model trained on only 448 grid points may still be regarded as a chemically meaningful representation of the PES since its non-weighted RMSE is only 0.77 kcal/mol (271 cm^{-1}). The error drops very quickly to 12.29 cm^{-1} for the 5%-ML model and is below 1.00 cm^{-1} for 35% and above, finally becoming smaller than 0.5 cm^{-1} for training sets with 50% or more of all configurations.

Regarding the four-layer, structure-based sampling ML models that were tested extensively, wRMSEs grow significantly from 50%-ML (0.37 cm^{-1}) to 25%-ML (1.60 cm^{-1}), to s10%-ML (3.63 cm^{-1}), and to 10%-ML (4.83 cm^{-1}). We further investigate the effect of the training set size for high-resolution spectroscopy applications in Sec. III B, where we report vibrational energies using PESs based on the 50%-ML, r50%-ML, 25%-ML, 10%-ML, and s10%-ML models. We also note that the wRMSEs for all four models remain rather small. One important factor that contributes to such high accuracy is the use of a permutational invariant kernel—the alternative approach of using the sorted molecular descriptor (Sec. II B) causes an increase in the corresponding wRMSEs by a factor of 2–3.

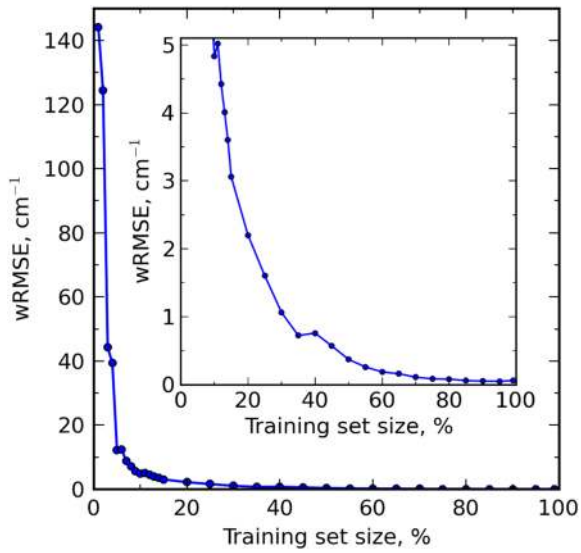

FIG. 5. Dependence of wRMSE in the prediction set (in cm^{-1}) of the four-layer, structure-based sampling ML models as a function of the training set size (in %). In all cases, sampling was done from unsliced data. The plot starts with a 1% training set size and ends at 99%. The plot in the inset starts with a 10% training set size.

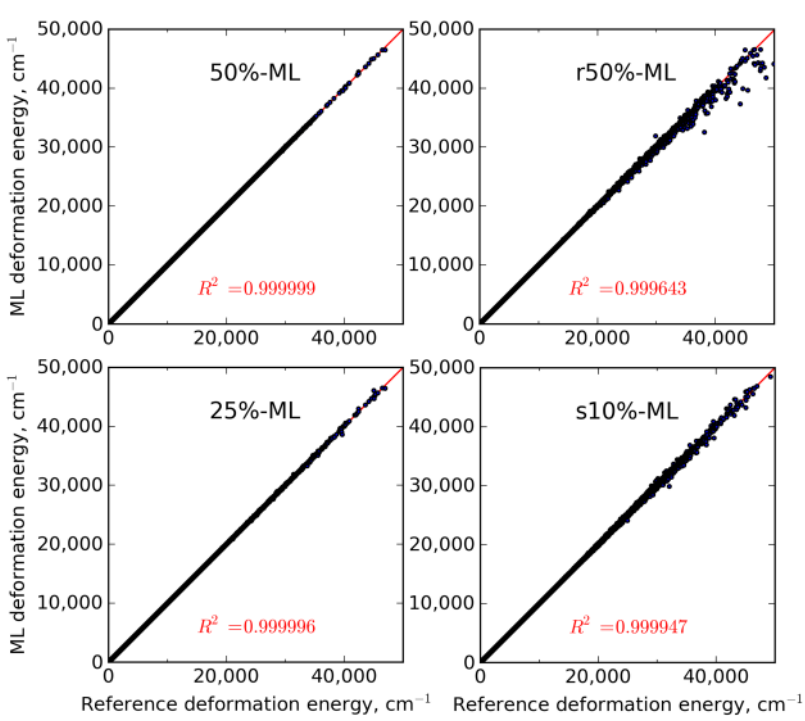

FIG. 4. Correlation between reference *ab initio* deformation energies and deformation energies predicted by 50%-ML, r50%-ML, 25%-ML, and s10%-ML for their respective prediction sets. Linear trend lines are shown in red with their R^2 value. Each data point is represented by a blue dot with a black edge; hence, the most densely populated areas are black.

The analytic representation employed for the CBS-35^{HL} (discussed in Secs. II D and III B) was fitted with a wRMSE of 3.00 cm^{-1} for the entire grid (see Sec. III B). For 50%-ML, the training set, prediction set, and entire grid of 44 819 points were reproduced with wRMSEs of 0.00, 0.37, and 0.26 cm^{-1}, respectively (Table II). Our ML approach, which has a defined analytic form, could therefore provide a more accurate description of the PES based on fewer training points and could possibly be employed directly in variational calculations. However, this is beyond the scope of this work, and the application of our ML technique in variational calculations will be the focus of future research.

Further fine-tuning of the ML models is possible by using the anisotropic Gaussian kernel with multiple kernel widths instead of a single σ parameter (as, for instance, in Ref. 13). In test calculations, we have found that such ML models may reduce errors somewhat (e.g., by 10% for 10%-ML), but there is a substantial increase in the complexity of the ML model and the computational cost for large training sets (due to parameter optimization), and hence, we decided to use one single kernel width parameter in this study.

B. Vibrational energy levels

For the 50%-ML, r50%-ML, 25%-ML, 10%-ML, and s10%-ML PESs, the $f_{ijk\ldots}$ expansion coefficients of the potential function given in Eq. (12) were determined in a least-squares fitting to the 44 819 grid points. The results of the fittings are listed in Table III. In addition, PESs were determined for the five associated training sets and the results are also included in Table II. We see that the fits of the CBS-35^{HL} PES, the ML model PESs, and the training set PESs are of a similar accuracy, with the exception of the 10%-ML fit which exhibits significantly larger errors. For the other fits, the RMSEs below 10 000 cm^{-1} range between 1.17 and 2.08 cm^{-1}, with wRMSE values between 3.01 and 3.59 cm^{-1} for the ML-based PESs and up to 3.60 cm^{-1} for the training set PESs. The mean errors are particularly low for the 50%-ML PES, its training set PES, and the 25%-ML PES (0.16, 0.24, and 0.02 cm^{-1}, respectively).

In TROVE calculations, the Hamiltonian was represented as a power series expansion around the equilibrium geometry in terms of nine vibrational coordinates and was constructed numerically using an automatic differentiation method.[57] The coordinates used were identical to those given in Eqs. (7)–(11), except for the kinetic energy operator where linear expansion terms, e.g., $(r - r^{eq})$, replace the Morse oscillator functions for the stretching modes. The kinetic and potential energy operators were truncated at the 6th and 8th orders, respectively, and atomic mass values were employed throughout. Calculations were carried out using a medium-sized vibrational basis set with a polyad truncation number of $P_{max} = 10$ (see Refs. 43 and 58 for further details). The basis set was

TABLE III. Number of expansion parameters $f_{ijk\ldots}$ and fitting wRMSEs[a] (in cm^{-1}) for the fits of the CBS-35^{HL} PES, ML model PESs, and training set PESs. Largest positive outlier (LPO) and largest negative outlier (LNO), mean signed errors (MSEs), mean absolute errors (MAEs), RMSEs for entire sets (all), for structures with reference deformation energies below 1000 and 10 000 cm^{-1}, and wRMSEs in cm^{-1} of the fitted functions with respect to the entire grid of the CBS-35^{HL} energies.

	CBS-35^{HL}	50%-ML	r50%-ML	25%-ML	10%-ML	s10%-ML
			ML model PESs			
No. of parameters	414	414	414	412	402	409
Fitting wRMSE	0.82	0.83	0.89	0.99	1.39	1.13
LPO	2717.33	2718.84	2546.10	2700.54	2506.24	2868.59
LNO	−6039.88	−6039.02	−6033.13	−6023.04	−5826.77	−5716.54
MSE	0.20	0.16	0.56	0.02	0.81	7.38
MAE	20.82	20.83	21.01	21.83	31.73	24.63
RMSE (all)	102.24	102.28	102.13	102.69	113.78	101.39
RMSE (<10 000)	1.18	1.18	1.22	1.22	3.66	1.37
RMSE (<1 000)	0.33	0.33	0.33	0.34	1.08	0.40
wRMSE	3.00	3.01	3.09	3.19	5.40	3.59
			Training set PESs			
No. of parameters		414	414	414	410	411
Fitting wRMSE		0.98	0.82	0.94	0.50	0.99
LPO		2548.77	2813.94	2368.88	1963.23	2431.71
LNO		−6128.55	−6033.18	−6185.91	−6267.97	−6214.42
MSE		0.24	0.75	1.50	−4.04	1.35
MAE		21.03	21.00	21.95	24.11	23.94
RMSE (all)		103.82	102.55	105.37	105.88	109.00
RMSE (<10 000)		1.17	1.18	1.30	2.08	1.18
RMSE (<1 000)		0.35	0.33	0.33	0.62	0.32
wRMSE		3.01	3.02	3.15	3.60	3.58

[a]Fitting wRMSEs are relative to the PES being fitted and not to the CBS-35^{HL} data. Note also that the weights differ slightly from Eq. (14) because Watson's robust fitting scheme[61] was employed (see Sec. II D).

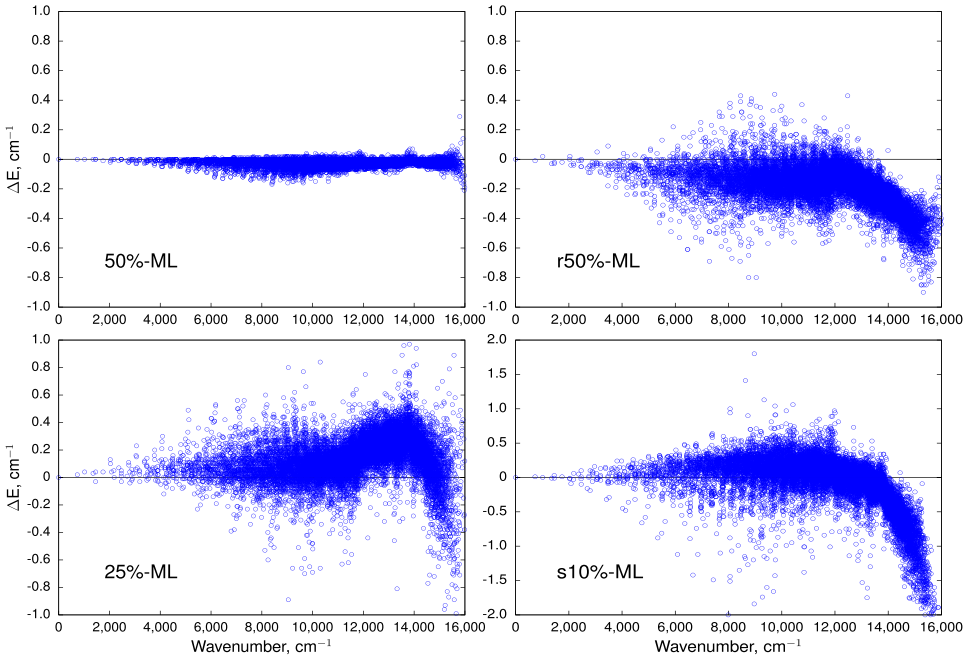

FIG. 6. Residual errors ($\Delta E = E_{\mathrm{CBS-35^{HL}}} - E_{\mathrm{ML}}$) of computed vibrational energy levels using the 50%-ML, r50%-ML, 25%-ML, and s10%-ML PESs, with respect to the CBS-35^{HL} PES values. Note the different scales for s10%-ML.

constructed using a multi-step contraction scheme and contained 16 829 vibrational basis functions.

In Fig. 6, we plot residual errors, $\Delta E = E_{\mathrm{CBS-35^{HL}}} - E_{\mathrm{ML}}$, of computed vibrational energy levels using the 50%-ML, r50%-ML, 25%-ML, and s10%-ML model PESs with respect to the CBS-35^{HL} PES values. The RMSE and mean absolute deviation (MAD) for energies up to 5000 and 10 000 cm^{-1} are also listed for each model in Table IV. Comparing 50%-ML with r50%-ML, it is clear that structure-based sampling produces results that are far more reliable than random sampling. The residual errors are consistently smaller and more uniform for the energy range considered. The 25%-ML and s10%-ML models still perform reasonably well but errors steadily increase with energy. The 10%-ML model (not shown in Fig. 6) has deteriorated and no longer gives accurate predictions (errors are much higher than 1 cm^{-1}, Table IV).

TABLE IV. Root-mean-square error (RMSE) and mean absolute deviation (MAD) of computed vibrational energy levels for the ML model PESs and training set PESs up to 5000 and 10 000 cm^{-1} (166 and 3606 levels, respectively), with respect to the original CBS-35^{HL} PES values.

	50%-ML	r50%-ML	25%-ML	10%-ML	s10%-ML
ML model PESs					
RMSE (<5 000 cm^{-1})	0.02	0.10	0.09	1.61	0.14
MAD (<5 000 cm^{-1})	0.01	0.08	0.07	1.29	0.10
RMSE (<10 000 cm^{-1})	0.04	0.18	0.16	1.75	0.28
MAD (<10 000 cm^{-1})	0.03	0.15	0.12	1.44	0.21
Training set PESs					
RMSE (<5 000 cm^{-1})	0.06	0.08	0.12	0.30	0.12
MAD (<5 000 cm^{-1})	0.05	0.06	0.10	0.25	0.10
RMSE (<10 000 cm^{-1})	0.12	0.14	0.19	0.74	0.32
MAD (<10 000 cm^{-1})	0.11	0.09	0.14	0.61	0.24

TROVE assigns quantum numbers to the eigenvalues by analyzing the contribution from the basis functions. This is how we match energy levels computed with different ML model PESs. However, given the approximate nature of the labeling scheme, these assignments can occasionally differ between the surfaces. This tends to happen mostly at higher energies (above 10 000 cm^{-1}) but does not necessarily mean that the energy levels are mismatched—they have simply been labeled differently. This occurs for 2% of the computed values in the case of 50%-ML, 9% for r50%-ML, 12% for 25%-ML, and 18% for s10%-ML. For 10%-ML, this percentage rises dramatically to 46% providing further evidence that this ML model is no longer reliable for high-accuracy applications.

We also computed vibrational energies using the training set PESs (also listed in Table IV). The errors are reasonably small and structure-based sampling again performs better than random sampling. For 50%-ML, 25%-ML, and s10%-ML, the predicted spectra are more reliable than the results of the respective training set PESs (constructed from 50%, 25%, and 10% of the entire CBS-35^{HL} PES dataset, respectively), but this does not hold for the r50%-ML model, and there is even a marked deterioration in the case of 10%-ML.

IV. CONCLUSIONS

We propose a procedure for building highly accurate PESs using KRR-based machine learning. Our approach employs structure-based sampling to ensure that machine learning is applied in the interpolation regime where it performs best. Data slicing in terms of the energy distribution is recommended for a balanced representation of the PES for very small training sets. Self-correction capabilities are introduced into the ML model by including additional ML layers.

In a pilot study, we explored the merits of our ML model using a recently published high-level *ab initio* PES of CH_3Cl as an example. Several ML models were built and trained using training sets of different sizes and a different number of data

slices, and their performance was assessed by comparisons with the original *ab initio* energies at the grid points in the training set and the prediction set. An excellent agreement was found for the 50%-ML model, which reproduces the PES with sub-wavenumber accuracy.

For five selected ML-based PESs, the vibrational energy levels of CH_3Cl were computed using variational TROVE calculations, in complete analogy to the published *ab initio* work.[43] The results clearly show that structure-based sampling produces more accurate ML models than random sampling. The structure-based sampling 50%-ML model gives an excellent agreement with the *ab initio* reference data for the vibrational energy levels (a RMSE of 0.04 cm^{-1} in the range up to 10 000 cm^{-1}). The accuracy deteriorates slightly for the 25%-ML model (a RMSE of 0.16 cm^{-1}) and for the s10%-ML model (a RMSE of 0.28 cm^{-1}, training set sampled from the dataset sliced into three regions) and quite strongly for the 10%-ML model (a RMSE of 1.75 cm^{-1}, training set sampled from the unsliced dataset).

The evidence from the present pilot study suggests that the number (and computational cost) of electronic structure calculations in high-level *ab initio* studies of rovibrational spectroscopy may be reduced by up to 90% (depending on the user needs and the initial grid size) by using structure-based sampling and self-correcting ML models, with a minimal loss of accuracy. We expect that this will also hold for small molecules other than CH_3Cl (in the absence of obvious reasons to suspect anything else). Of course, this should be examined in future work, which should also aim at establishing standard protocols for such ML studies. Finally, given the fact that our ML model is available in analytic form, it seems worthwhile to explore whether it can be used directly in TROVE-type variational calculations (without an intermediate fit to a standard polynomial form).

V. OUTLOOK

The objective of this study was not to improve upon the existing *ab initio* PES of CH_3Cl but to demonstrate how the computational cost of building such a PES can be substantially reduced by performing fewer *ab initio* calculations and by interpolating efficiently with KRR-based ML. We have shown that this can be done by the following procedure:

1. Generate a large and dense grid of deformed structures using established techniques (as outlined in Sec. II C and Ref. 43).
2. Select points from this grid into the training set by using structure-based sampling.
3. Calculate the energy for each point in the training set as accurately as possible using a high-level *ab initio* method.
4. Train the self-correcting ML model on the training set geometries and the high-level *ab initio* energies.
5. Predict the energies for the remaining grid points using self-correcting machine learning.
6. Calculate rovibrational levels variationally using TROVE.
7. Repeat steps 2 to 6 by including more points from the grid into the training set using structure-based sampling until the calculated rovibrational levels converge.

We plan to apply this procedure to generate accurate PESs for other small molecules. During these studies, we will also investigate additional ways to reduce even further the computational cost of generating accurate PESs by combining our procedure with the so-called Δ-ML approach.[11] In this approach, ML is trained on differences between high-level and less demanding but still reliable low-level *ab initio* results; the Δ-ML model is then applied to correct low-level energies for the remaining grid points (similar to previous work on the water PES[13]). We anticipate that this approach will allow us to investigate larger systems such as organic molecules with several carbon atoms.

ACKNOWLEDGMENTS

P.D. thanks Georgios Gerogiokas for software performance related advice and Alexey Dral for discussions. The authors thank the reviewers for their valuable comments.

[1] J. Behler, J. Chem. Phys. **145**, 170901 (2016), see also references therein.
[2] M. Gastegger and P. Marquetand, J. Chem. Theory Comput. **11**, 2187 (2015).
[3] B. Kolb, B. Zhao, J. Li, B. Jiang, and H. Guo, J. Chem. Phys. **144**, 224103 (2016).
[4] L. Shen, J. Wu, and W. Yang, J. Chem. Theory Comput. **12**, 4934 (2016).
[5] A. Bholoa, S. D. Kenny, and R. Smith, Nucl. Instrum. Methods Phys. Res., Sect. B **255**, 1 (2007).
[6] M. Malshe, R. Narulkar, L. M. Raff, M. Hagan, S. Bukkapatnam, P. M. Agrawal, and R. Komanduri, J. Chem. Phys. **130**, 184102 (2009).
[7] N. Artrith and J. Behler, Phys. Rev. B **85**, 045439 (2012).
[8] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).
[9] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).
[10] P. O. Dral, O. A. von Lilienfeld, and W. Thiel, J. Chem. Theory Comput. **11**, 2120 (2015).
[11] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 2087 (2015).
[12] K. Toyoura, D. Hirano, A. Seko, M. Shiga, A. Kuwabara, M. Karasuyama, K. Shitara, and I. Takeuchi, Phys. Rev. B **93**, 054112 (2016).
[13] A. P. Bartók, M. J. Gillan, F. R. Manby, and G. Csányi, Phys. Rev. B **88**, 054104 (2013).
[14] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[15] J. Cui and R. V. Krems, J. Phys. B: At., Mol. Opt. Phys. **49**, 224001 (2016).
[16] S. Manzhos and T. Carrington, J. Chem. Phys. **125**, 084109 (2006).
[17] S. Manzhos, X. G. Wang, R. Dawes, and T. Carrington, J. Phys. Chem. A **110**, 5295 (2006).
[18] S. Manzhos, R. Dawes, and T. Carrington, Int. J. Quantum Chem. **115**, 1012 (2014).
[19] M. Majumder, S. E. Hegger, R. Dawes, S. Manzhos, X.-G. Wang, C. Tucker, Jr., J. Li, and H. Guo, Mol. Phys. **113**, 1823 (2015), see also references therein.
[20] K. Shao, J. Chen, Z. Zhao, and D. H. Zhang, J. Chem. Phys. **145**, 071101 (2016).
[21] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, J. Chem. Phys. **103**, 4129 (1995).
[22] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017).
[23] S. Hobday, R. Smith, and J. Belbruno, Modell. Simul. Mater. Sci. Eng. **7**, 397 (1999).
[24] S. Hobday, R. Smith, and J. BelBruno, Nucl. Instrum. Methods Phys. Res., Sect. B **153**, 247 (1999).
[25] L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley, and R. Komanduri, J. Chem. Phys. **122**, 084104 (2005).
[26] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[27] J. Behler, R. Martonak, D. Donadio, and M. Parrinello, Phys. Rev. Lett. **100**, 185501 (2008).
[28] A. Pukrittayakamee, M. Malshe, M. Hagan, L. M. Raff, R. Narulkar, S. Bukkapatnum, and R. Komanduri, J. Chem. Phys. **130**, 134101 (2009).
[29] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, Phys. Rev. B **81**, 184107 (2010).

[30]R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, Phys. Rev. B **81**, 100103(R) (2010).
[31]J. Behler, Phys. Chem. Chem. Phys. **13**, 17930 (2011), see also references therein.
[32]R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, Nat. Mater. **10**, 693 (2011).
[33]T. Morawietz, V. Sharma, and J. Behler, J. Chem. Phys. **136**, 064103 (2012).
[34]J. Behler, Int. J. Quantum Chem. **115**, 1032 (2015), see also references therein.
[35]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[36]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[37]K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. **9**, 3404 (2013).
[38]V. Botu and R. Ramprasad, Int. J. Quantum Chem. **115**, 1074 (2015).
[39]Z. Li, J. R. Kermode, and A. D. Vita, Phys. Rev. Lett. **114**, 096405 (2015).
[40]P. Geiger and C. Dellago, J. Chem. Phys. **139**, 164105 (2013).
[41]T. Suzuki, R. Tamura, and T. Miyazaki, Int. J. Quantum Chem. **117**, 33 (2016).
[42]F. Häse, S. Valleau, E. Pyzer-Knapp, and A. Aspuru-Guzik, Chem. Sci. **7**, 5139 (2016).
[43]A. Owens, S. N. Yurchenko, A. Yachmenev, J. Tennyson, and W. Thiel, J. Chem. Phys. **142**, 244306 (2015).
[44]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer-Verlag, 2009), p. 763.
[45]O. A. von Lilienfeld, Int. J. Quantum Chem. **113**, 1676 (2013).
[46]R. Ramakrishnan, M. Hartmann, E. Tapavicza, and O. A. von Lilienfeld, J. Chem. Phys. **143**, 084111 (2015).
[47]P. O. Dral, MLatom: A Package for Atomistic Simulations with Machine Learning, version 0.9, Max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr, Germany, 2013–2017.
[48]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Sci. Data **1**, 140022 (2014).
[49]T. Bereau, D. Andrienko, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 3225 (2015).
[50]M. Rupp, R. Ramakrishnan, and O. A. von Lilienfeld, J. Phys. Chem. Lett. **6**, 3309 (2015).
[51]F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Phys. Rev. Lett. **117**, 135502 (2016).
[52]M. Rupp, Int. J. Quantum Chem. **115**, 1058 (2015).
[53]On one occasion (of the ML models trained on 50% of the randomly drawn grid points), the RMSE of the training set in the first layer became unreasonably high due to a very small value of the parameter λ. In this case, we increased λ to 10^{-6} for the first layer and re-trained the ML model.
[54]K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. **6**, 2326 (2015).
[55]A. P. Bartók and G. Csányi, Int. J. Quantum Chem. **115**, 1051 (2015).
[56]S. N. Yurchenko, W. Thiel, and P. Jensen, J. Mol. Spectrosc. **245**, 126 (2007).
[57]A. Yachmenev and S. N. Yurchenko, J. Chem. Phys. **143**, 014105 (2015).
[58]A. Owens, S. N. Yurchenko, A. Yachmenev, J. Tennyson, and W. Thiel, J. Quant. Spectrosc. Radiat. Transfer **184**, 100 (2016).
[59]P. R. Bunker and P. Jensen, *Molecular Symmetry and Spectroscopy*, 2nd ed. (NRC Research Press, Ottawa, 1998).
[60]H. Partridge and D. W. Schwenke, J. Chem. Phys. **106**, 4618 (1997).
[61]J. K. G. Watson, J. Mol. Spectrosc. **219**, 326 (2003).