Check for updates

Received: 8 July 2022 | Revised: 21 September 2022 | Accepted: 28 September 2022

DOI: 10.1002/jcc.27035

RESEARCH ARTICLE

Journal of COMPUTATIONAL CHEMISTRY WILEY

Adaptive regularized Gaussian process regression for application in the context of hydrogen adsorption on graphene sheets

Gunnar Schmitz | Bastian Schnieder

Theoretische Chemie, Ruhr-Universität Bochum, Bochum, Germany

Correspondence
Gunnar Schmitz, Theoretische Chemie, Ruhr-Universität Bochum, Bochum D-44780, Germany.
Email: gunnar.schmitz@rub.de,

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: Ha 2588/9-1; German Research Foundation, Grant/Award Number: 215035359

Abstract

We present a Gaussian process regression (GPR) scheme with an adaptive regularization scheme applied to the QM7 and QM9 test set, several protonated water clusters and specifically to the problem of atomic hydrogen adsorption on graphene sheets. For the last system our goal is to achieve good predictive accuracy with only a few training points. Therefore, we assess for these systems a self-correcting multilayer GPR model, in which the prediction is corrected by a chain of additional GPR models. In our adaptive regularization scheme, we impose no noise on the training data, but use an approach based on the data itself to account for its impurity. The strength of this strategy is that the data points are treated differently based on their importance and that the regularization can still be controlled by a single parameter. We assess how the accuracy of the prediction depends on this parameter. We can show that the new regularization scheme as well as the multilayer approach results in more robust predictors. Furthermore, we demonstrate that the predictor can be in good agreement with the density-functional theory results.

KEYWORDS
atomistic potentials, Gaussian process regression, machine learning, regularization

1 | INTRODUCTION

Machine learning (ML) algorithms have gained a lot of interest in the recent years and also found their way into quantum chemistry as shown by recent work from different groups[1–8] and recent special issue on this topic[9] as well as reviews,[10–12] which indicate the impact of ML on theoretical chemistry. Algorithms like neural networks (NNs) as well as Kernel based methods like support vector regression (SVR), Kernel ridge regression (KRR) and Gaussian process regression (GPR)[1]have been adapted for chemical usage. For example, NN potentials (NNPs) were successfully used in molecular dynamic (MD) simulations for describing surface reactions.[13–15] Other applications show the usage for excited state dynamics.[16–18] Additionally, methods like NNs and GPR/KRR have been successfully applied for the simulation of vibrational spectra,[19–23] adsorption spectra[24] and other types of spectra. There is work on training GPR based ML potentials on the fly.[25–27] ML algorithms can be used to accelerate geometry optimizations,[28–30] to accelerate the convergence in self-consistent-field methods[31] or to estimate the Coupled Cluster correlation energy based on approximate MP2 like doubles amplitudes.[32] There are definitely many more applications, but we wanted to illustrate that ML techniques are NNs are very popular, but more and more work on GPR/KRR continues as viable alternative. Often the research objective gives an indication which algorithm to use. For example, some guidelines can be found in Ref. 33. NNs are a parametric form of fitting, while GPR/KRR perform non-parametric fitting. Having only limited training data GPR/KRR often yields a better accuracy than

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Journal of Computational Chemistry* published by Wiley Periodicals LLC.

NNs, but on the other hand NNs scale better with large training data sets. However, also linear scaling GPR approaches have been proposed[34] and are also partially used in Theoretical Chemistry.[35,36] An important concept in GPR is noise which one can attribute to the data. This noise can originate from experimental uncertainty or numerical precision, but is ultimately a tool to add flexibility to the model since the regressor has not to pass to every training point exactly anymore. In most cases, a constant noise is assigned to all data points. This is an homoscedastic approach, for example, it is assumed that all data points have the same finite variance. However, it would be beneficial if the magnitude of the noise added is based on the importance of the data point.[37,38] However, this is a rather complicated task, but when going a step back and take a general look at regularization, one can derive a simple approach to adapt the regularization based on the underlying data. This can be achieved by a further analysis of the co-variance matrix, which encodes the correlation of data points in the framework of GPR/KRR. An eigenvalue decomposition (EVD) of this matrix can be used to determine the number of significant training points, which are required to represent the model. By taking a look at the Eigen spectrum one usually observes that at some point the eigenvalues rapidly fall to zero. These small eigenvalues represent points, which are less important for the model, but also cause numerical problems. For example, the matrix might then be (numerically) not positive definite and the condition number gets large. Usually these problems are accounted for by adding a constant to the diagonal of the co-variance matrix, which is interpreted as noise imposed on the data and shifts the eigen spectrum by this value. Thus a small value should be applied in order to only introduce minimal changes to the original co-variance matrix. In this work we test a scheme in which the importance of a point is taken into account and the regularization differs based on the eigenvalue decomposition. The co-variance matrix is reconstructed using the EVD and a damped noise added to the diagonal, which mostly affects small eigenvalues and results in a new co-variance matrix which is closer to the original one.

Additionally, to this new regularization scheme we combine it with a multilayer GPR approach, in which different GPR models are chained together, where each layer tries to correct the remaining error of the previous layer(s) to the reference output signal. A similar idea was already presented in Ref. 39, but we combine it with an automatic deactivation of layers and the new regularization to achieve a stable predictor with high accuracy.

We apply the new regularization scheme to various well-known test sets for validation of the approach, but our main goal is to train potentials for hydrogen adsorption on graphene sheets. This will pave the way to answer questions in catalysis related to hydrogen spill over, which is interesting for hydrogen storage and hydration. Here, we want to emphasize that in the past decade hydrogen spillover from for example metal nanoparticles onto carbon material has gained interest since the spillover effect significantly enhances the carbon material's capability of storing hydrogen.[40–43] The first time hydrogen spillover was reported in 1964 by Khoobiar in Ref. 44, where he observed a transport of hydrogen atoms from Pt to WO_3. A potential application of this effect is in future fuel cell technology as hydrogen storage. Therefore, several experimental groups have tried to get a deeper understanding of the details of the hydrogen spillover mechanism, which however turned out quite difficult due to the complicated characterization of the carbon support and the generated hydrogen spillover species. Recently, Li et al.[45] and Wang et al.[46] showed that some functional groups on the surface as well as residual moisture can enhance the hydrogen spillover on platinum/nitrogen-doped mesoporous carbon composites. The processes caught also the interest of theoretical groups. For example Psofogiannakis et al.[47] shed light on the hydrogen spillover mechanism on Pt-Doped graphite using Density-functional Theory (DFT) using the coronene molecule and a Pt_4 cluster as model system. Furthermore, Wu et al.[48] studied the hydrogen storage capabilities on B-doped graphene with DFT and could show that the B-doping enhances the adsorption strength. There is for example a recent computational study on this matter,[49] where model systems of carbon-supported palladium nanoparticles are used to investigate the hydrogen spillover and the influence of nitrogen doping on this process by means of DFT. It was found that graphitic nitrogen atoms and also water molecules facilitate the hydrogen spillover and serve as anchoring sites for the spillover hydrogen atoms. However, further insight is needed and here ML techniques pave the way for long-term MD simulation or to tackle larger systems.

The paper is organized as follows. First, we will give a basic overview about the Theory behind GPR and the adaptive regularization we are introducing. This is followed by a recap of the self-correcting GPR framework and available Kernel functions, which can be used in different layers. Afterwards, it will be explained how GPR can be used to learn properties. In the current context of atomistic potentials, by a combination of using atom centered symmetry functions (ACSFs) and Kernel contraction. After this theory part, the computational details are outlined; and in the Result section, we first focus our attention on known quantum mechanical (QM) test sets as the QM7, QM9 test set as well as previously reported protonated water clusters. Based on these databases we assess the parameters in our regularization scheme and then apply it for the study of H adsorption on graphene sheets.

2 | THEORY

2.1 | Gaussian process regression

Gaussian process regression belongs to a class of ML algorithms based on Bayesian inference[34] and by definition a Gaussian process (GP) is a statistical distribution of functions, for which every finite linear combination of samples has a joint Gaussian distribution. Thus, we assume that our observed function is distributed according to a normal distribution of functions.

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma_n^2 \mathbf{I}) \tag{1}$$

In order to simplify things without effecting the generality of the approach a zero mean function $\mathbf{0}$ is assumed. The term $\mathbf{K}$ is the co-variance matrix, which elements i, j are obtained from a Kernel function $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, which encodes the correlation/closeness of the data points. The only crucial requirement of a Kernel function is that it is

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

positive definite. The term $\sigma_n^2\mathbf{I}$ is often added as regularization term and imposes a noise on the observations and can increase numerical stability. This trick helps especially if the co-variance matrix is (numerically) not strictly positive definite. This approach usually works, but in this work we will propose a scheme where this term is set to zero and an adaptive regularization scheme is used, which we describe in more detail in Section 2.2 and has the advantage that the changes to the original co-variance matrix is smaller.

In order to make predictions in the framework of GPR, we condition distribution to observations $\boldsymbol{f}_* \mid \boldsymbol{f} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which results in a new normal distribution with new mean function,

$$\mu(\boldsymbol{x}_i) = \sum_{j=1}^{n_{train}} \omega_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{2}$$

and co-variance matrix,

$$\boldsymbol{\Sigma} = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X})\mathbf{K}(\mathbf{X}, \mathbf{X})^{-1}\mathbf{K}(\mathbf{X}, \mathbf{X}_*) \ , \tag{3}$$

where the weights are obtained as $\omega = \boldsymbol{K}^{-1}\boldsymbol{f}$ and can be obtained by solving the linear systems of equations $\boldsymbol{K}\omega = \boldsymbol{f}$ for which we use the Bunch-Kaufmann decomposition which is numerically more stable than Cholesky decomposition, which is otherwise often applied for GPR.

The co-variance matrix of the posterior distribution can be used to judge the accuracy, since its diagonal elements represent the variance.

$$\mathbb{V}\left[\mathbf{f}_i^*\right] = \boldsymbol{\Sigma}_{ii} \ . \tag{4}$$

This is not an error bar in an absolute sense, but it is an indication if predictions might not be trustworthy. We found that a good conservative choice is to disregard predictions, when the variance is larger than $1 \times 10^{-1}\ \sigma^2{}_{\mathrm{n}}$.

2.2 | Adaptive regularization

The term in Equation (1) is a regularization term, which is interpreted as a noise for the data. It gives GPR more flexibility since the predictor has not to pass through each point exactly anymore. Mathematically this term just shifts the eigen spectrum of the co-variance matrix by a (small) constant and cures numerically instabilities if the co-variance matrix is numerically not fully positive definite. A drawback of this approach is that the same noise is added to each point. It would be more beneficial if it would only be added to those having a small eigenvalue. With this in mind, we apply a regularization scheme known in the multi-configuration time-dependent Hartree (MCTDH)[50–52] community, which is based on the eigen decomposition of the matrix and reconstructs the co-variance matrix as:

$$\widetilde{\mathbf{K}} = \mathbf{U}[\mathbf{D} + \varepsilon \cdot \exp(-\mathbf{D}/\varepsilon)]\mathbf{U}^T \ , \tag{5}$$

where ε is the term which controls the strength of regularization. $\mathbf{U}$ are the eigenvectors of $\mathbf{K}$, and $\mathbf{D}$ its eigenvalues. The new regularization scheme is extremely attractive since the re-constructed co-variance matrix is closer to the original one in the sense that the variance of the data, which can be extracted from the trace is less altered. If one measures the error as the difference in the trace of original matrix $\mathbf{K}$ and the re-constructed matrix $\widetilde{\mathbf{K}}$, which is either obtained by adaptive regularization (Equation 5) or by adding the noise $\widetilde{\mathbf{K}} = \mathbf{K} + \sigma_n^2\mathbf{I}$,

$$\delta = Tr\mathbf{K} - Tr\widetilde{\mathbf{K}} \tag{6}$$

the error in the conventional case is then,

$$\delta_{\sigma_n^2} = Tr\mathbf{K} - Tr\left(\mathbf{K} + \sigma_n^2\mathbf{I}\right) = n \times \sigma_n^2 \tag{7}$$

and for the adaptive case,

$$\delta_\varepsilon = Tr\mathbf{K} - Tr\left(\mathbf{U}[\mathbf{D} + \varepsilon \cdot \exp(-\mathbf{D}/\varepsilon)]\mathbf{U}^T\right) = \sum_n \varepsilon \exp - D_{ii}/\varepsilon \ , \tag{8}$$

which likely to be better since it does not directly scale with the number of training points n, but with the number of insignificant training points, which will also scale at some point with n, but the prefactor is largely reduced. To illustrate this for general matrices, we constructed random positive definite 5000×5000 matrices and calculated the errors in the traces with $\sigma^2{}_{\mathrm{n}} = \varepsilon = 1 \times 10^{-8}$ and gathered the results in Table 1. The average error using the adaptive regularization scheme is three orders of magnitude smaller. Additionally, we plotted the error in dependence of the size of the matrices in Figure 1. From this data, it is evident that first the error introduced with the adaptive regularization scheme is smaller and that the error scales better with the number of training points.

Note that in this study, the parameter ε will remain a user-defined parameter, but it is possible to treat this parameter as hyperparameter and optimize it. This can be an attractive option in the future.

2.3 | Self-correcting GPR

The adaptive regularization should give a more robust predictor since it gives a bias to the well-represented points in the training set. An additional ingredient for higher accuracy is to chain GPRs in layers to correct for the error in a previous layer on input each GP has the difference of the previous layer to the output signal.

$$\mathbf{v}_*^{(0)} \mid \mathbf{v} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}\right) \tag{9}$$

$$\mathbf{v}_*^{(1)} \mid \mathbf{v} - \mathbf{v}^{(0)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}\right) \tag{10}$$

$$\mathbf{v}_*^{(n)} \mid \mathbf{v} - \sum_i^{n-1} \mathbf{v}^{(i)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(n)}, \boldsymbol{\Sigma}^{(n)}\right) \tag{11}$$

The same idea was presented in the context of KRR by Dral and coworkers.[39] A strength of the multilayer approach is that different

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universitat Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Error in the trace of random positive definite 5000 × 5000 reconstructed matrices using the conventional approach and the adaptive regularization scheme δ_ε.

	1	2	3	4	5	6	7	8	9	10
$\delta_{\sigma_n^2}$	-5.0×10^{-5}	-5.0×10^{-5}	-5.0×10^{-5}	-5.0×10^{-5}	-5.0×10^{-5}	-5.0×10^{-5}	-5.0×10^{-5}	-5.0×10^{-5}	-5.0×10^{-5}	-5.0×10^{-5}
δ_ε	-1.0×10^{-8}	-1.0×10^{-16}	5.6×10^{-9}	-4.7×10^{-9}	-7.6×10^{-9}	-5.6×10^{-9}	-1.1×10^{-8}	1.0×10^{-8}	6.5×10^{-9}	-9.3×10^{-9}
Averages										
$\overline{\delta}_{\sigma_n^2}$	-5.0×10^{-5}									
$\overline{\delta}_\varepsilon$	-2.61×10^{-9}									

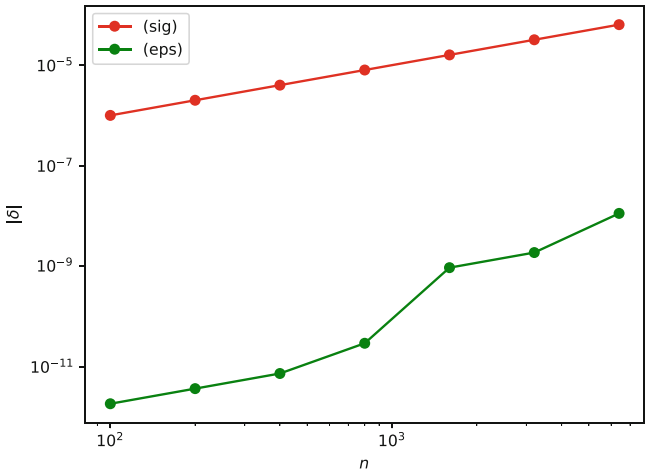


FIGURE 1 Absolute error in the trace of random positive definite matrices plotted against the size of the matrix.

Kernel functions can be applied to the different layers accounting for a different type of signal. The number and types of layer (Kernel) are chosen by the user. Since the self-correction scheme becomes unstable if the correction is small we deactivate a layer when the learned signal falls below a certain threshold T_δ, where the default is 1×10^{-7}.

2.4 | Kernel functions

The Kernel function encode a measure for the closeness of data points. We implemented the following Kernel functions, which can be used in the different layers. The length scale specific squared exponential Kernel (SCEXP):

$$k_{\mathbf{xy}} = \sigma_f^2 \exp\left(-0.5 \sum_i^d \frac{(x_i - y_i)^2}{l_i^2}\right) \tag{12}$$

The length scale specific Matérn 5/2 Kernel (SC52):

$$k_{\mathbf{xy}} = \sigma_f^2 \left(1 + \sqrt{5} \sum_i^d \frac{|x_i - y_i|}{l_i} + 5 \sum_i^d \frac{(x_i - y_i)^2}{3 l_i^2}\right) \times \exp\left(-\sqrt{5} \sum_i^d \frac{|x_i - y_i|}{l_i}\right) \tag{13}$$

The length scale specific Matérn 3/2 Kernel (SC32):

$$k_{\mathbf{xy}} = \sigma_f^2 \left(1 + \sqrt{3} \sum_i^d \frac{|x_i - y_i|}{l_i}\right) \exp\left(-\sqrt{3} \sum_i^d \frac{|x_i - y_i|}{l_i}\right) \tag{14}$$

The squared exponential Kernel (SQEXP):

$$k_{\mathbf{xy}} = \sigma_f^2 \exp\left(-0.5 \frac{\|\mathbf{x} - \mathbf{y}\|^2}{l^2}\right) \tag{15}$$

The Matérn 5/2 Kernel (SC52):

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$k_{xy} = \sigma_f^2 \left(1 + \frac{\sqrt{5}\|\mathbf{x}-\mathbf{y}\|}{l} + \frac{5\|\mathbf{x}-\mathbf{y}\|^2}{3l^2}\right) \exp\left(-\sqrt{5}\frac{\|\mathbf{x}-\mathbf{y}\|}{l}\right) \quad (16)$$

And the Ornstein Uhlenbeck Kernel (OUEXP):

$$k_{xy} = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|}{l}\right) \quad (17)$$

2.5 | Atomistic GPR potential

Our aim is to construct atomistic potentials using GPR. Employing atomistic potentials, we assume that the energy of a system E can be represented as[53]:

$$E = \sum_{\alpha} \sum_{i \in \alpha} \varepsilon_i^{\alpha} \quad , \quad (18)$$

where the sum runs over the atoms α, i are descriptors/features of atom α and are local atom specific energy functionals. As descriptor, we will employ atomic centered symmetry functions,[54] which will be described in more detail in Section 2.6. In this work, we apply this descriptor to the several test sets, but are also interested in the H adsorption on graphene sheets. For this specific case, it is enough to restrict the descriptor to a single atom and therefore the sum in Equation (18) only runs over the corresponding atom index. For this particular case the energies for training are calculated as:

$$\frac{E_{H_2}^{ads.}}{2} = E_{Support+H} - \left[E_{Support} + \frac{1}{2}E_{H_2}\right] \quad , \quad (19)$$

where $E_{Support\ +\ H}$ the energy of the support including the adsorbed H atom, $E_{Support}$ only the energy of the support and the energy of molecular hydrogen.

2.6 | Atomic centered symmetry functions

We implemented a slightly modified version of the atomic centered symmetry function (ACSF) G_i^1,G_i^2, and G_i^4 of Behler and co-workers as outlined in Ref. 54. The function G_i^1 is essentially a cut-off function and describes the range of interactions,

$$G_i^1 = \sum_{j \neq i} f_c(R_{ij}) \quad (20)$$

while the G_i^2 robes the angular environment around an atom,

$$G_i^2 = \sum_{j \neq i} \exp\left(-\exp(\eta)^2 R_{ij}^2 / R_c^2\right) f_c(R_{ij}) \quad (21)$$

and the G_i^4 function the angular part.

$$G_i^4 = 2^{1-\xi} \sum_{j,k \neq i (j \neq k)} (1 + \lambda \cos\theta_{ijk})^{\xi} \times \exp\left(-\exp(\eta)^2 / R_c^2 \left(R_{ij}^2 + R_{ik}^2 + R_{jk}^2\right)\right) \times f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \quad . \quad (22)$$

In the equation, above f_c is a cutoff function defined as:

$$f_c(R_{ij}) = \begin{cases} 0.5\left(1 + \cos\frac{\pi R_{ij}}{R_c}\right), \text{if} R_{ij} \leq R_c \\ 0, \text{if} R_{ij} > R_c \end{cases} \quad (23)$$

and θ_{ijk} is the angle between the triplet of atoms i, j and k. The parameter λ is either 1 or $-$ 1 and one should include both in a descriptor. The parameter R_c the cut-off radius and η and ξ control the sensitivity with respect to radial and angular information around atom i. Additionally, we propose the following modifications and introduce the functions,

$$G_i^{2,m} = \sum_{j \neq i} \frac{Z_i Z_j}{R_{ij}} \exp\left(-\exp(\eta)^2 (R_{ij} - R_s)^2 / R_c^2\right) f_c(R_{ij}) \quad (24)$$

And,

$$G_i^{4,m} = 2^{1-\xi} \sum_{j,k \neq i (j \neq k)} \frac{Z_i Z_j Z_k}{(R_{ij} R_{ik} R_{jk})^d} (1 + \lambda \cos\theta_{ijk})^{\xi} \times \exp\left(-\exp(\eta)^2 / R_c^2 \left((R_{ij} - R_s)^2 + (R_{ik} - R_s)^2 + (R_{jk} - R_s)^2\right)\right) \times f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \quad , \quad (25)$$

where R_s is a shift, d a hyperparameter and Z_i the nuclear charge (weighted case) or set to 1 (non-weighted case). Having functions with the same parameters, but varying the shift helps to efficiently probe the environment around an atom.

2.7 | Coulomb matrices

A different way to represent molecular systems for use in ML applications are Coulomb Matrices.[55] A coulomb matrix is calculated as follows:

$$C_{ij} = \begin{cases} 0.5 \cdot Z_i^{2.4}, \text{if} i = j \\ \frac{Z_i Z_j}{R_{ij}} \end{cases} \quad , \quad (26)$$

where Z_i and Z_j are the nuclear charges of atoms i and j. The Coulomb matrix gives an electrostatic descriptor for the system and the diagonal can be interpreted as the self-interaction of an atom. Coulomb matrices depend on the size of the molecular system and are thus not applicable to systems with varying sizes, but one can arrive at a more transferable descriptor by sorting the entries of the Coulomb Matrix column wise and add zeroes for systems which have less atoms.[56] We follow a similar path, but truncate the number of atoms per column to a value n_{cut} and refer to this in the subsequent parts as truncated Coulomb matrices.

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

This choice reflects a cutoff criterion to only consider interactions in a given vincinity. We want to note in passing that there are many choices on how to represent the chemical structure for use in ML. We restricted ourselves to ACSFs and Coulomb matrices, but there are many possibilities. For a recent overview, we refer to Ref. 57.

3 | KERNEL CONTRACTION

We want to learn a functional of the energy w.r.t. to atomic features. This type of decomposition is not accessible via quantum chemical calculations. Still, it is possible to learn the relation of different features and one observation at once following the ideas presented by Csányi and co-workers[35,53,58] and adapted by others.[59] To solve this issue let us assume a vector $\mathbf{P}$ of atomic features $\mathbf{p}_\alpha$ so that,

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}_\alpha \\ \mathbf{p}_\beta \\ \vdots \\ \mathbf{p}_\omega \end{pmatrix} = \begin{pmatrix} p_{\alpha,1} \\ \vdots \\ p_{\alpha,n} \\ p_{\beta,1} \\ \vdots \\ p_{\beta,n} \\ \vdots \\ p_{\omega,1} \\ \vdots \\ p_{\omega,n} \end{pmatrix}, \tag{27}$$

where n is the number of features per atom (e.g., number of atom centered symmetry functions) and the atom indices run from α to ω. We choose here ACSFs or truncated Coulomb matrices, but other choices are possible and found in the literature. It is only important to relate the features on the different atoms simultaneously to one observation, for example, the total energy we can modify the expression for the co-variance matrix as:

$$K_{ij} = \sum_{\alpha \in i} \sum_{\beta \in j} k\left(\mathbf{p}_\alpha^{(i)}, \mathbf{p}_\beta^{(j)}\right), \tag{28}$$

where $\mathbf{p}_\alpha^{(i)}$ are the features on atom α in the data point i. In the following, this should be denoted as Kernel contraction. The Kernel contraction allows to learn the atomic energy contribution of a single atom by learning the full energy. The only requirement is that the number of features and the used functional form (e.g., parameters of the ACSFs) are the same for all atoms.

The Kernel contraction allows to learn atom-wise properties and makes it possible to apply GPR to any chemical system. The basic steps for using GPR for molecular systems are summarized in Figure 2. Given the molecular structure as Cartesian coordinates, they have to be transformed in a set of descriptors, which in our case is done via ACSFs. A Kernel functions is used to correlate data points and Kernel contraction paves the ways for decompose a learned output to atomic contributions. The regression step is then plain application of GPR and the application falls again into the realm of quantum chemistry, where properties like energies and forces can be predicted.

4 | COMPUTATIONAL METHODS

All quantum chemical calculations were performed with the Turbomole program package[60] version 7.5.[61] For the DFT calculations, the *ridft* module was used. Basis sets applied in the DFT calculation include the def2-SVP[62] basis set. Furthermore, Grimme's D3 dispersion correction[63] was added. As functional the TPSS[64] (Tao, Perdew, Staroverov, and Scuseria exchange and correlation functional) was chosen.

The GPR generation was done using a local development version of the MidasCPP[65] program package. The GPR hyperparameters are optimized by maximizing the logarithm of the marginal likelihood[34] using the iRprop algorithm[66] as described in Ref. 28.

For the assessment of the regularization, we use the QM7 test set[55,67] and QM9 test[68,69] for atomization energies and the training sets of Schran et. al[70] which contains several conformations of the protonated water clusters H_2O, H_3O^+, H_5O_2+, H_7O_3+, H_9O. The data sets contain 6593, 7200, 9926, 10,329 and 20,664 structures, respectively. These test sets were partitioned in training and validation sets by applying a random shuffle and extracting 5% as validation set.

To generate the training set for the H adsorption on graphene sheets we created different graphene sheets and optimized a reaction path for the H atom approaching the graphene sheet using the *woelfling* program in the Turbomole package. For each point on the reaction path, we performed a short molecular dynamic simulation with 256 steps to get a configurational sampling. In the created data sets, high energy structures are removed to arrive at a more balanced training set.

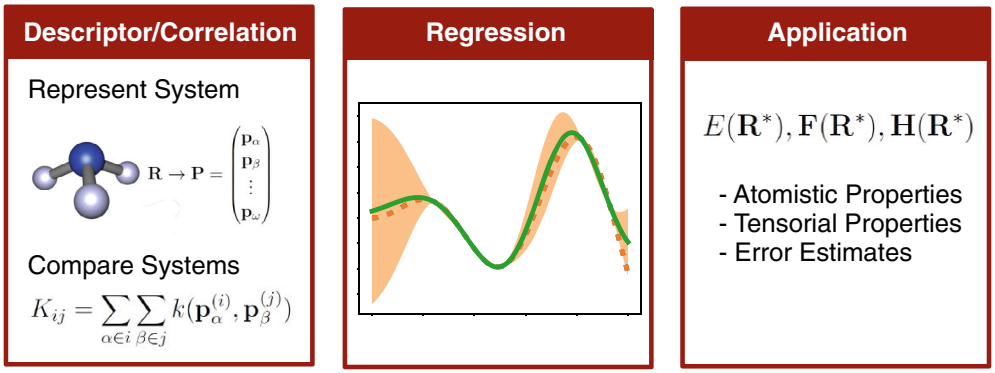


FIGURE 2 Schematic illustration about the steps involved in applying Gaussian process regression to chemical problems. First the molecular structure has to be represented and data points need to correlated via a Kernel, second the regression is performed and third the prediction is used for applications.

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5 | RESULTS AND DISCUSSION

5.1 | QM7 and QM9 test systems

We varied the input noise σ_n^2 (sig) as well as the regularization parameter ε (eps) for the QM7 test set of atomization energies. The hyperparameters remained fixed for this study for a better comparison. The dependence of the mean absolute error (MAE) to these parameters is shown in Figure 3. We see that with both approaches it is possible to reach chemical accuracy, but we also see that smaller errors are obtained for the adaptive regularization scheme compared to the regularization via noise term in case of larger values for σ_n^2 and ε. There is not much improvement when tightening the parameters below 1×10^{-7} and although the results look good in this case too tight values can cause over-fitting. Judging from the data a value of 1×10^{-6} to 1×10^{-7} seems to be a reasonable choice.

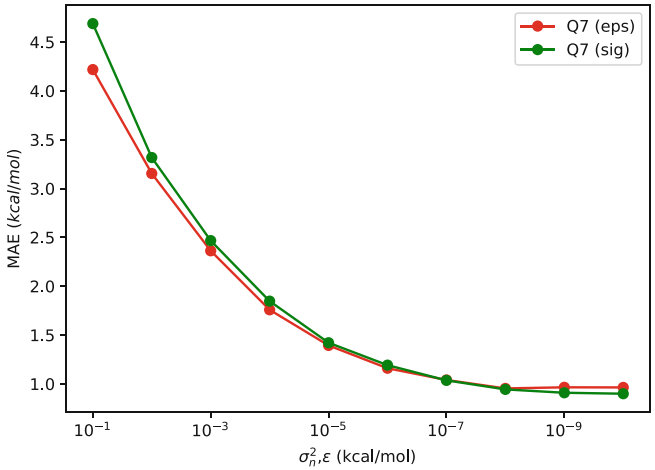


FIGURE 3 Mean absolute error (kcal/mol) of the GPR prediction of atomization energies based on the regularization parameters σ_n^2 and ε for the QM7 test set. As Kernel the SCEXP Kernel was used. In this assessment only predictions were included for which the variance is lower than 1×10^{-1} σ_n^2 or 1×10^{-1} ε.

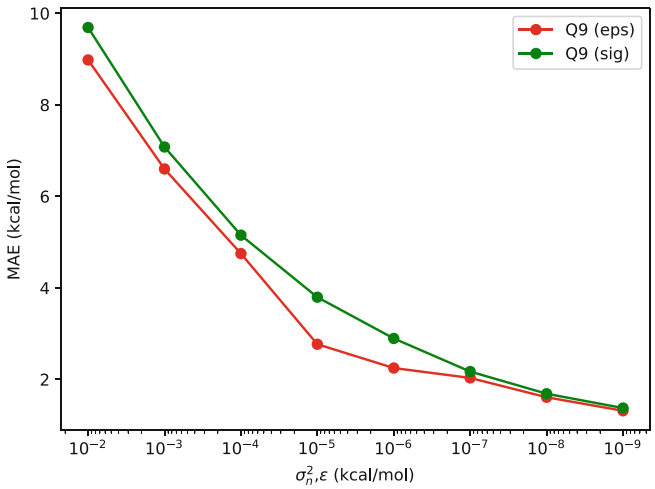


FIGURE 4 Mean absolute error (kcal/mol) of the GPR prediction of atomization energies based on the regularization parameters σ_n^2 and ε for the QM9 test set. As Kernel the SCEXP Kernel was used. In this assessment only predictions were included for which the variance is lower than 1×10^{-1} σ_n^2 or 1×10^{-1} ε. As descriptor we use a combination of different ACSFs including G_i^2,G_i^4, and weighted $G_i^{2,m}$ and $G_i^{4,m}$ functions.

The QM9 test set contains 133,730 structures. This is exceedingly large. For our purposes, we randomly extracted 26,746 points for training. For our assessment, we will again use the MAE w.r.t to the validation set. Figure 4 shows the MAE for the varied regularization parameters. Deviating from the previous test we did not go tighter than 1×10^{-9} since then over fitting occurred. We reach an MAE around 1.5 kcal/mol. It is to note that more accurate predictions were reported in the literature,[71] but we did not extensively optimize the hyperparameters in the GPR model nor the ACSFs. Thus, we can definitely still improve our results, but the main objective here is the comparison of both regularization schemes.

Setting the regularization to 1×10^{-6} seems to be a reasonable choice. For this choice, we studied how the amount of data influences

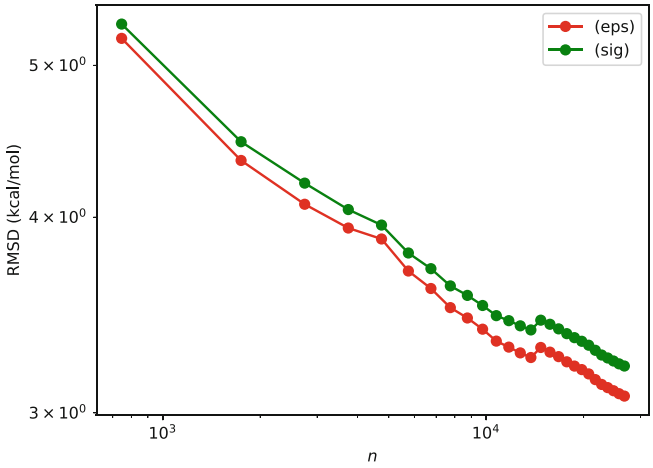


FIGURE 5 Dependence of the RMSD with respect to the amount of data for the QM9 data set using the adaptive regularization (eps) and the conventional (sig) approach. As Kernel the SCEXP Kernel was used. In this assessment only predictions were included for which the variance is lower than or 1×10^{-1} ε.

TABLE 2 Number of neglected predictions for both regularization schemes.

	N_{screen}	
ε,σ_n^2 (kcal/mol)	(eps)	(sig)
1×10^{-3}	0	0
1×10^{-4}	0	0
1×10^{-5}	0	1
1×10^{-6}	0	2
1×10^{-7}	1	7
1×10^{-8}	2	14
1×10^{-9}	3	17

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

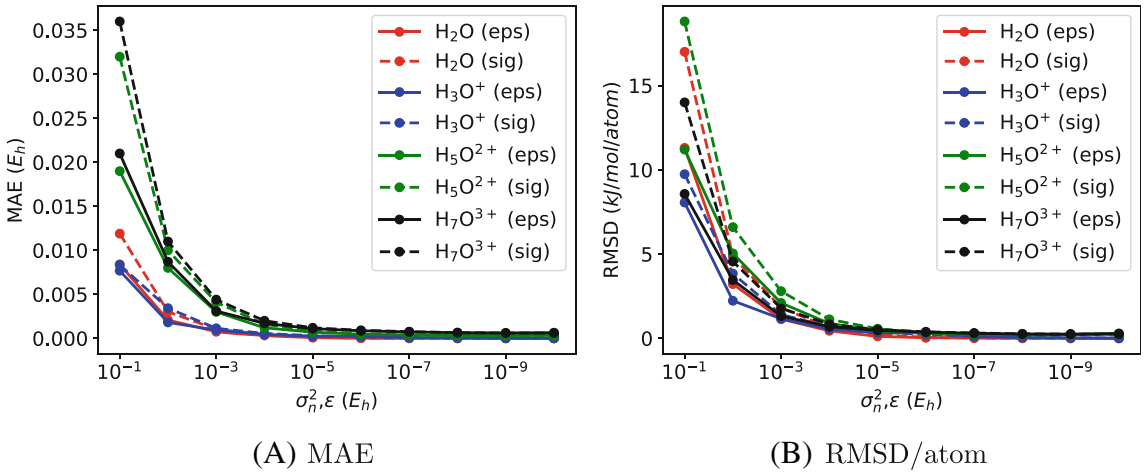


FIGURE 6 Mean absolute error MAE (E_h) of the GPR prediction of total energies based on the regularization parameters and ε as well as the RMSD per atom (kJ/mol) for systems of protonated water clusters. As Kernel the SCEXP Kernel was used. In this assessment only predictions were included for which the variance is lower than $1 \times 10^{-1}\,\sigma_n^2$ or $1 \times 10^{-1}\,\varepsilon$.

FIGURE 7 Dependence of the RMSD with respect to the amount of data using the adaptive regularization (eps) and the conventional (sig) approach. As Kernel the SCEXP Kernel was used. In this assessment only predictions were included for which the variance is lower than $1 \times 10^{-1}\,\sigma_n^2$ or $1 \times 10^{-1}\,\varepsilon$.

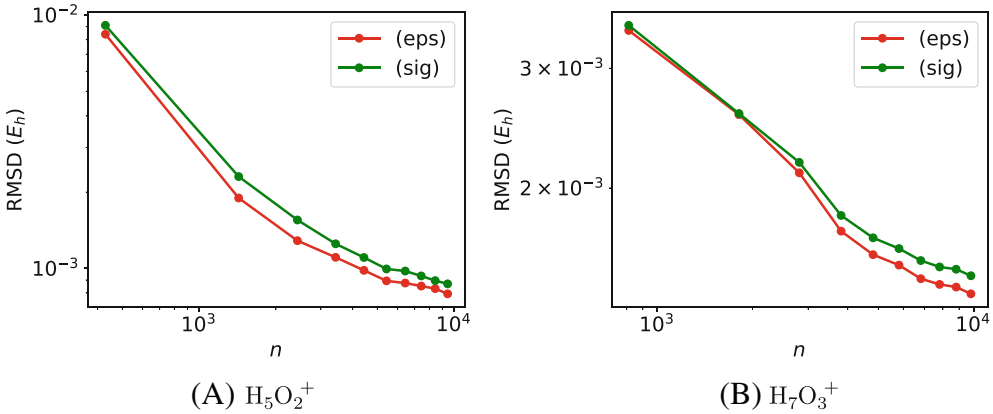
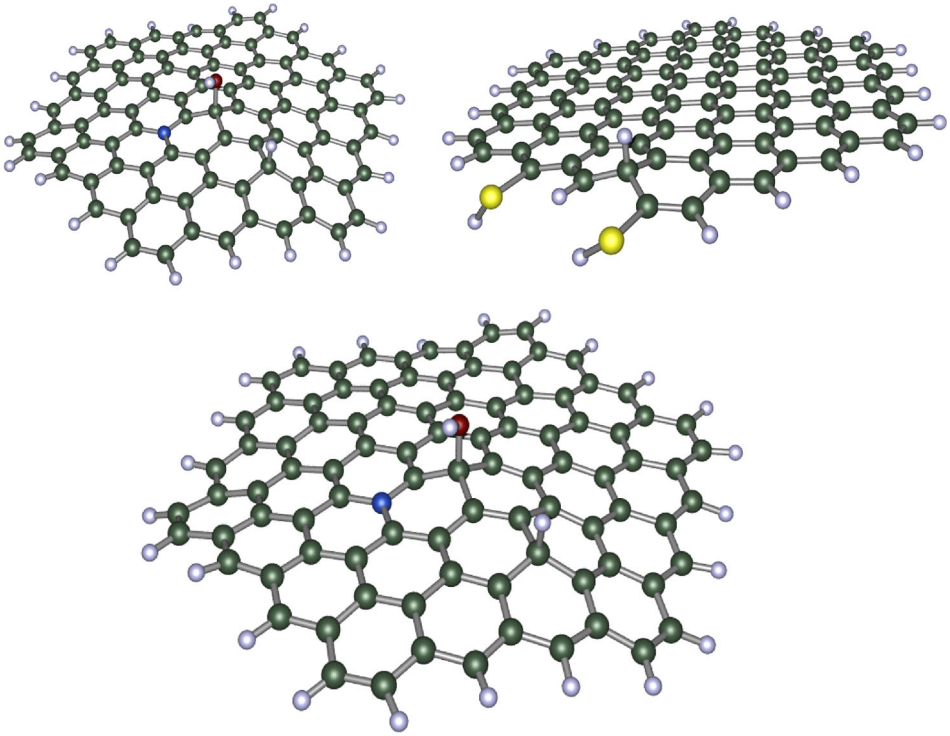


FIGURE 8 Examples of considered system for the H adsorption on graphene. Different functionalizations are accounted for



1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universitat Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the accuracy of the predictor in form of the root-mean-square deviation (RMSD) for both regularization schemes. For this purpose, we randomly selected smaller subsets of the training data and re-trained the predictor. The trend is depicted in Figure 5. Both regularization scheme perform equally well. For small amounts of the data, the conventional one is a bit more accurate and for large amount of data the adaptive one works slightly better.

Overall, the accuracy is quite similar, but the adaptive scheme has the more robust predictor. This can also be shown by our exclusion criterion, where we do not include predictions for which the variance is lower than or $1 \times 10^{-1}\ \sigma_n^2\ 1 \times 10^{-1}\ \varepsilon$. In production calculations, one would treat these predictions with caution and likely retrain the model including the exact data for these points. Table 2 lists the number of excluded predictions for both schemes. We observe that only very few points are removed using the adaptive regularization. Especially for low values for the regularizer the difference becomes apparent. This is an important feature for applications in Bayesian optimization and will likely result in fewer points being re-calculated.

5.2 | Protonated water clusters

In an next assessment step we considered the protonated water clusters mentioned in the computational details section. Again we varied the regularization parameters and calculated the MAE in a.u. and the RMSD per atom in kJ/mol. The data is depicted in Figure 6. The error statistics for the smaller systems are a bit better, but further tuning of the hyperparameters in GPR or in the ACSFs will most likely fix this issue. Furthermore, the main objective is to compare the two regularization scheme. Again, we observe a similar behavior. As expected the MAE and RMSD decreases when tightening the parameters and for larger parameters the adaptive scheme results in improved error statistics. The differences between both schemes are still somewhat small, but the adaptive regularization systematically works better. The curves start to flatten out around values of 1×10^{-6} for the regularization parameters. Like for the QM9 test set we studied how the amount of data influences the accuracy of the predictor. The trend is shown in Figure 7 and a bit more clear than in the case of the QM9 test case. We see that the adaptive regularization results in a more accurate predictor. It is to note that the differences are still small, but it is also to note that the differences are more pronounced for larger training sets. This makes sense since the adaptive scheme should put more emphasize on points, which are consistent with the model.

5.3 | H adsorption on graphene

5.3.1 | Initial test set

For the QM7 and QM9 test sets, we observed a good predictive accuracy and could asses computational settings. For the H adsorption on graphene we first want to explore how one can increase the accuracy of the ML model when only minimal amount of data is available.

TABLE 3 Error statistics for the different multilayer models m1 (SCEXP), m2 (SCEXP, SQEXP), and m3 (SCEXP, SQEXP, SC32). Energies are given in kJ/mol or for REL the relative error is reported in %.

	m1				m2				m3			
	ε (kJ/mol)		σ_n^2 (kJ/mol)		ε (kJ/mol)		σ_n^2 (kJ/mol)		ε (kJ/mol)		σ_n^2 (kJ/mol)	
	1×10^{-5}	1×10^{-6}	1×10^{-5}	1×10^{-6}	1×10^{-5}	1×10^{-6}	1×10^{-5}	1×10^{-6}	1×10^{-5}	1×10^{-6}	1×10^{-5}	1×10^{-6}
MAD	10.35	6.35	6.48	6.53	6.81	5.27	6.61	5.71	4.27	3.79	4.38	3.96
SD	11.79	7.50	7.02	7.59	3.48	3.68	3.24	3.28	3.24	2.77	3.94	2.72
REL	12.24	12.47	12.11	12.82	10.46	8.77	10.38	9.19	7.21	6.44	7.11	6.79
RMSD	15.48	7.81	8.01	7.92	7.65	5.90	7.37	6.35	5.06	4.33	5.48	4.46

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Error statistics for the single layer models m1 (SCEXP). Energies are given in kJ/mol or for REL the relative error is reported in %.

	ACSFs				Truncated coulomb matrix			
	ε (kJ/mol)		σ_n^2 (kJ/mol)		ε (kJ/mol)		σ_n^2 (kJ/mol)	
	1×10^{-5}	1×10^{-6}	1×10^{-5}	1×10^{-6}	1×10^{-5}	1×10^{-6}	1×10^{-5}	1×10^{-6}
N_{screen}	1	1	7	14	129	158	269	330
MAD	5.81	5.39	5.83	5.02	5.21	1.85	3.15	0.44
STD	9.57	8.54	9.06	7.27	5.20	3.06	5.37	0.52
REL	−2.58	−3.43	−0.62	0.15	−0.23	−0.29	−1.10	−0.01
RMSD	10.66	9.57	9.07	7.27	5.21	3.07	5.38	0.58

Note: In this assessment only predictions were included for which the variance is lower than $1 \times 10^{-1}\ \sigma_n^2$ or $1 \times 10^{-1}\ \varepsilon$. the number of neglected validation points is given as N_{screen}. The validation set contained 412 points. Both ACSFs and truncated coulomb matrices are compared as descriptor. In the latter case, we included up to 60 entries per column in the matrix for hydrogen.

Examples for the used systems are depicted in Figure 8. Therefore, we will in a first step combine the adaptive regularization with the self-correcting multilayer GPR approach. To indicate the benefits of the multilayer scheme, we discuss results from very initial explorations on data assembled from reaction path optimizations and single snapshots to calculate the H adsorption energy. We did calculations involving one, two, and three layers and employing the adaptive regularization scheme as well as the conventional noise scheme. In model 1 (m1) we employed a SCEXP Kernel, in model 2 (m2) a SCEXP Kernel and a SQEXP Kernel and in model 3 (m3) a SCEXP Kernel followed by a SQEXP and SC32 Kernel. The error statistics are gathered in Table 3. The error statistics contain the mean absolute deviation (MAD), the standard deviation (*SD*), the relative error (REL) and the root mean square deviation (RMSD).

We limit the assessment to values for the regularization of 1×10^{-5} and 1×10^{-6}. Using only one layer the accuracy is rather okay considering the limited amount of data used to train the model. However, the accuracy increases when adding more layers to the model. For the three layer system, we essentially reach chemical accuracy for the predictor with an RMSD around 4 kJ/mol. The relative error to the DFT reference is below 10%, which is good enough for most applications, especially considering that the evaluation of the predictor (once trained) virtually takes no time compared to a DFT calculation.

In this reported case, the adaptive regularization scheme and the conventional noise scheme behave equally well. In the next section, we want to target a larger training and validation set to examine the performance.

5.3.2 | Larger test sets

For the general test set and assessment, we followed the same recipe as in Section 5.3.1, but optimized the reaction path for more systems and performed small MD simulations for the points in the path to get a better configurational sampling. On the gathered data, we applied farthest point sampling to select the most diverse points and choose to remove close lying points This resulted in total in 6989 points. In our database, we added different functionalization and defects in the sheet. For this part, we rely on a single layer. We also experimented using in the self-correction multiple layers, but could only marginally improve the accuracy.

As first stage, we report the results using a single SCEXP layer and the values 1×10^{-5} and 1×10^{-6} for the regularization. For this study we employed ACSFs including the newly introduced $G^{2,m}$ and $G^{4,m}$ functions without weighting since we found them to perform better than the conventional ACSFs. Additionally, we employed truncated Coulomb Matrices for comparison. The error statistics can be found in Table 4. In this case, the adaptive regularization scheme and the conventional noise scheme perform quite well, and we can reach a relative error around 0.62%–3.4% compared to the DFT reference. We observe that we can achieve a better accuracy with truncated Coulomb matrices and that in this case the conventional regularization scheme gives a better accuracy, but this is only due to the fact that in our assessment we exclude as safeguard predictions, where the variance of the prediction is larger than 10^{-1} the input regularization parameter. It becomes evident that in the cases of the truncated Coulomb Matrices, much more predictions are screened out by this criterion. This also holds for the comparison of adaptive and conventional regularization scheme. Using the adaptive scheme the number of screened out predictions is at least reduced by a factor of 7. Reconstructing the co-variance matrix by altering the correlation to specific points increases the stability of the predictor. The RMSD is a bit larger than for the smaller test and training set, but this is due to the added diversity, which puts more emphasize on transferability.

The better performance of the adaptive regularization scheme can be explained by the increased amount of data. The increased size and diversity makes it mandatory to balance the data in order not to spoil the accuracy by a few badly represented points. This is achieved by our new regularization scheme by incorporating knowledge of the significance of a single data point. It is to be expected that the regularization scheme will in general perform better for larger data sets. The downside of this approach is the necessity of performing an eigen decomposition, which is more expensive than the usually used Cholesky decomposition. However, the arithmetic complexity is in both cases $\mathcal{O}(n^3)$, where n is the number of training points and both approaches can be combined with linear scaling variants[34] of GPR,

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

which removes this downside. In our current implementation the regularization scheme is used to reconstruct the co-variance matrix and afterwards it is decomposed again using the Bunch-Kaufmann decomposition. However, this is not necessary. It is possible to maintain the EVD decomposed form also for solving for the weights, which will boost the efficiency. Furthermore, it is to note that our model was trained on the "full "energy. It can be expected to increase the accuracy by using the delta-learning trick as shown in various Refs. 72, 73. In this context one can also expect to train a correction to the current model based on a fewer number of single points since the variations of the difference surface are smaller than those of the potential energy surface. As further validation for our approach we performed for one trained predictor a one-dimensional cut through the potential energy surface (PES) and calculated the adsorption energy of hydrogen using the predictive model and on the DFT level. Similar cuts are included in the training set for less dense intervals. A schematic illustration of the chosen distances are given in Figure 9. The *x* coordinate was fixed and the *y* coordinate was incremented from point to point by 0.1 a.u. Two cuts are depicted in Figure 10. For most parts of the cut the GPR results nicely follow the DFT reference and barriers are well reproduced. There is however a small double well minimum in the second cut not present in the DFT data and in general the agreement in the second cut is a bit worse. However, the error is still relatively small and a further sampling in this region and retraining will most likely solve the issue. Most barriers are nicely reproduced and overall the qualitative agreement is very good, although the shape of adsorption energy varies strongly with the position. It is noteworthy

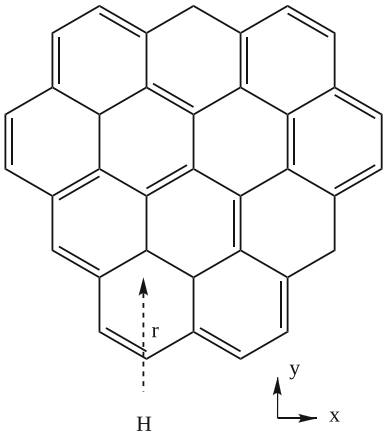


FIGURE 9 Schematic illustration of the scan along the distance of an H atom parallel to the graphene sheet. The *x* coordinate was fixed, and the scan went along the *y*-axis. Relaxation along the *z*-axis (not depicted) was possible.

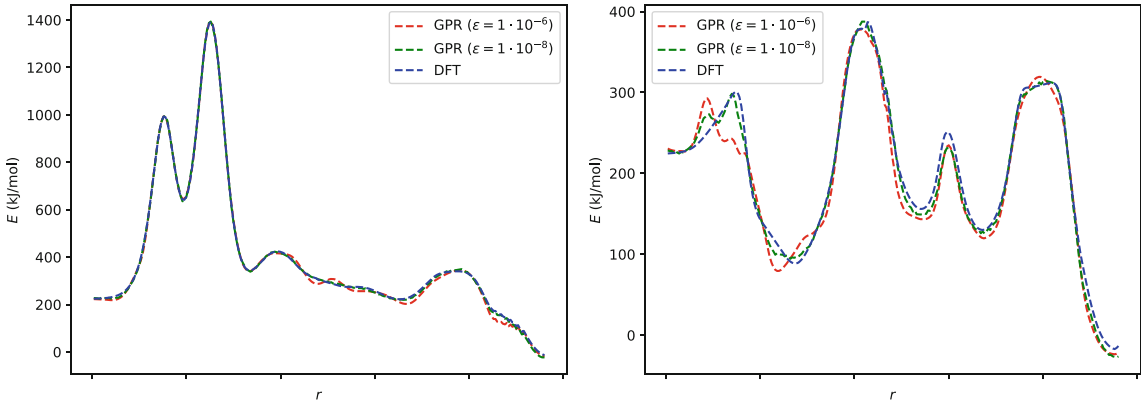
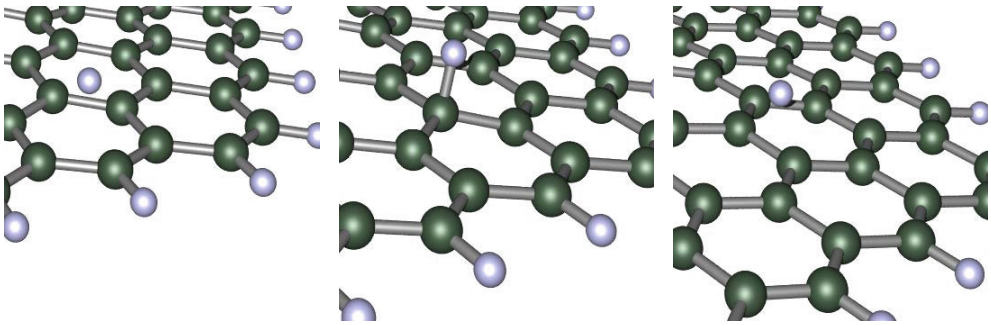


FIGURE 10 Different one-dimensional scans along the distance of H atom parallel to the graphene sheet. Depicted are the adsorption energies for the H atom. Shown are results based on GPR and DFT.



(A) Example Maximum I
$\Delta E_{\mathrm{DFT}}^{\mathrm{GPR}} = 5.1$ kJ/mol

(B) Example Minimum I
$\Delta E_{\mathrm{DFT}}^{\mathrm{GPR}} = 1.6$ kJ/mol

(C) Example Maximum II
$\Delta E_{\mathrm{DFT}}^{\mathrm{GPR}} = 3.4$ kJ/mol

FIGURE 11 Different positions of the H atom identified as minima and maxima on the PES. Additionally the errors compared to DFT (TPSS/def2-SVP) are given. These points refer to the first minima and maxima of Figure 10B.

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

that barriers are mainly large at the edge regions. Once adsorbed the energy landscape is on a small level and later the H atom can essentially even move freely on the surface. Based on this data structures, which can be associated to minima and maxima on the PES were extracted and analyzed. Examples are depicted in Figure 11 alongside the prediction errors compared to the DFT calculation using TPSS/def2-SVP. The errors are only a few kJ/mol. Looking at the identified structures, there are no surprises, but the introduced machinery is a first stepping stone to generate atomistic potentials, which can be used in MD simulations to investigate the H spillover effect. This will be part of a subsequent study where we train the potentials on a higher level of theory and include further defects on the graphene sheets. One current weakness of our trained potentials is that it is sort of two-dimensional in nature since it was trained mainly for the adsorbed H atom. To also get further information a next step would be to scan along the normal mode of the adsorbed H atom and generate additional data. However, this is work for a subsequent study and will at least double the amount of training data. We want to note in passing that this will be quite challenging since it turns out that the vibrational frequencies of the adsorbed H atoms vary largely based on the adsorbed position. On a different note, the building blocks of our work can be used in the reaction path optimization. Currently, we used optimized reaction paths a part of our training set, but only kept the optimized one. However, it is possible to turn this around and train a GPR based predictor on-the-fly during a reaction path optimization with the obtained energies and gradients to build a surrogate potential which is used for the relaxations and updated when the variance of the prediction gets too large.

6 | CONCLUSION

We presented a new regularization scheme for GPR, which adapts the regularization depending on the importance of a data point. This is achieved by an eigenvalue decomposition of the co-variance matrix and re-construction using a damped noise term. We assess the accuracy compared to the conventional regularization of adding a constant to the diagonal of the co-variance matrix, which can be interpreted as noise imposed on the data. In general cases the adaptive regularization scheme introduces smaller errors in the reconstructed co-variance as measured by the deviation in the trace. Furthermore, the error has a smaller dependence of the size of the matrix. This will especially be important for large- scale training sets. For the assessment, we use standard test sets like the QM7 and QM9 test sets as well as our own database for the H adsorption on graphene sheets. We observe that our new adaptive regularization scheme is numerically more stable and usually gives a better accuracy than the noise based scheme. Especially for large data sets the adaptive scheme is superior, which is already evident from the smaller dependence of the error on the size of the co-variance matrix. For our training data, we used a rather cheap level of theory, but this study was meant as proof-of-concept. In the future, we want to apply our approach to problems in catalysis and will create training data on a higher level of theory like Coupled Cluster. In this field, ML techniques combined with quantum chemical methods are game changer since the Machine-Learning model can be trained on smaller systems and then applied to larger systems to study the influence of structural parameters in long-term MD simulations. For this purpose, we want to combine our adaptive regularization scheme with linear scaling GPR variants. Work on this issue is already in progress in our laboratories. An additional ingredient will be a recent workflow we presented, which extracts reaction events in molecular dynamic simulation and post-processes them with DFT calculations. Combined with configurational sampling this will automatize the generation of training sets and the adaptive regularization will weight the data points according to their importance. Additionally an interesting perspective is to regard the regularization parameter as hyperparameter and optimize it alongside the other hyperparameters.

ACKNOWLEDGMENTS

Gunnar Schmitz thanks Ove Christiansen for helpful discussions on the adaptive regularization scheme. This work has been supported by the Deutsche Forschungsgemeinschaft within the framework of the Sino-German initiative "Novel Functional Materials for Sustainable Chemistry" through Grant Ha 2588/9-1. We also thank to the German Research Foundation (DFG) for funding this research within the framework of the SFB/Transregio 129 "Oxyflame" (project number 215035359). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The generated structures are available in the supporting information and further data is available on request from the authors.

ORCID

Gunnar Schmitz https://orcid.org/0000-0003-3677-4748

ENDNOTE

[1] GPR and KRR result in same working equations and will be used interchangeable in this paper

REFERENCES

[1] M. Rupp, *Int. J. Quantum Chem.* **2015**, *115*, 1058.
[2] J. Cui, R. V. Krems, *Phys. Rev. Lett.* **2015**, *115*, 073202.
[3] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.
[4] J. Behler, *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930.
[5] L. Mones, N. Bernstein, G. Csányi, *J. Chem. Theory Comput.* **2016**, *12*, 5100.
[6] B. Hartke, *J. Phys. Chem.* **1993**, *97*, 9973.
[7] A. R. Oganov, C. W. Glass, *J. Chem. Phys.* **2006**, *124*, 244704.
[8] T. C. Le, D. A. Winkler, *Chem. Rev.* **2016**, *116*, 6107.
[9] M. Rupp, O. A. von Lilienfeld, K. Burke, *J. Chem. Phys.* **2018**, *148*, 241401.
[10] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[11] B. Huang, N. O. Symonds, O. A. von Lilienfeld, *Quantum Machine Learning in Chemistry and Materials*, Springer International Publishing, Cham **2020**, p. 1883.
[12] P. O. Dral, in *Chemical Physics and Quantum Chemistry. Advances in Quantum Chemistry*, Vol. 81 (Ed: K. R. E. J. Brändas), Academic Press, Cambridge Massachusetts **2020**, p. 291.

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[13] H. Ghorbanfekr, J. Behler, F. M. Peeters, *J. Phys. Chem. Lett.* **2020**, *11*, 7363.
[14] V. Quaranta, J. Behler, M. Hellström, *J. Phys. Chem. C* **2019**, *123*, 1293.
[15] M. Hellström, V. Quaranta, J. Behler, *Chem. Sci.* **2019**, *10*, 1232.
[16] J. Westermayr, P. Marquetand, *Mach. Learn.: Sci. Technol.* **2020**, *1*, 043001.
[17] W.-K. Chen, X.-Y. Liu, W.-H. Fang, P. O. Dral, G. Cui, *J. Phys. Chem. Lett.* **2018**, *9*, 6702.
[18] J. Westermayr, F. A. Faber, A. S. Christensen, O. A. von Lilienfeld, P. Marquetand, *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025009.
[19] S. Manzhos, T. Carrington, *J. Chem. Phys.* **2006**, *125*, 194105.
[20] S. Manzhos, T. Carrington, *J. Chem. Phys.* **2008**, *129*, 224104.
[21] A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington, S. Manzhos, *J. Chem. Phys.* **2018**, *148*, 241702.
[22] G. Schmitz, D. G. Artiukhin, O. Christiansen, *J. Chem. Phys.* **2019**, *150*, 131102.
[23] G. Schmitz, E. L. Klinting, O. Christiansen, *J. Chem. Phys.* **2020**, *153*, 064105.
[24] B.-X. Xue, M. Barbatti, P. O. Dral, *J. Phys. Chem. A* **2020**, *124*, 7199.
[25] R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse, M. Bokdam, *Phys. Rev. Lett.* **2019**, *122*, 225701.
[26] R. Jinnouchi, F. Karsai, G. Kresse, *Phys. Rev. B* **2019**, *100*, 014105.
[27] R. Jinnouchi, K. Miwa, F. Karsai, G. Kresse, R. Asahi, *J. Phys. Chem. Lett.* **2020**, *11*, 6946.
[28] G. Schmitz, O. Christiansen, *J. Chem. Phys.* **2018**, *148*, 241704.
[29] A. Denzel, J. Kästner, *J. Chem. Phys.* **2018**, *148*, 094114.
[30] A. Denzel, J. Kästner, *J. Chem. Theory Comput.* **2018**, *14*, 5777.
[31] M. Gastegger, A. McSloy, M. Luya, K. T. Schütt, R. J. Maurer, *J. Chem. Phys.* **2020**, *153*, 044123.
[32] J. T. Margraf, K. Reuter, *J. Phys. Chem. A* **2018**, *122*, 6343.
[33] M. Pinheiro, F. Ge, N. Ferré, P. O. Dral, M. Barbatti, *Chem. Sci.* **2021**, *12*, 14396.
[34] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, Cambridge Massachusetts **2005**.
[35] A. P. Bartók, G. Csányi, *Int. J. Quantum Chem.* **2015**, *115*, 1051.
[36] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **2021**, *121*, 10073.
[37] Z. Li, X. Hong, K. Hao, L. Chen, B. Huang, *Chem. Eng. Res. Des.* **2020**, *157*, 162.
[38] I. A. Almosallam, M. J. Jarvis, S. J. Roberts, *Mon. Not. R. Astron. Soc.* **2016**, *462*, 726.
[39] P. O. Dral, A. Owens, S. N. Yurchenko, W. Thiel, *J. Chem. Phys.* **2017**, *146*, 244108.
[40] A. J. Lachawiec, G. Qi, R. T. Yang, *Langmuir* **2005**, *21*, 11418.
[41] C. I. Contescu, C. M. Brown, Y. Liu, V. V. Bhat, N. C. Gallego, *J. Phys. Chem. C* **2009**, *113*, 5886.
[42] J. L. Blackburn, C. Engtrakul, J. B. Bult, K. Hurst, Y. Zhao, Q. Xu, P. A. Parilla, L. J. Simpson, J.-D. R. Rocha, M. R. Hudson, C. M. Brown, T. Gennett, *J. Phys. Chem. C* **2012**, *116*, 26744.
[43] X. M. Liu, Y. Tang, E. S. Xu, T. C. Fitzgibbons, G. S. Larsen, H. R. Gutierrez, H.-H. Tseng, M.-S. Yu, C.-S. Tsao, J. V. Badding, V. H. Crespi, A. D. Lueking, *Nano Lett.* **2013**, *13*, 137.
[44] S. Khoobiar, *J. Phys. Chem.* **1964**, *68*, 411.
[45] Q. Li, A. D. Lueking, *J. Phys. Chem. C* **2011**, *115*, 4273.
[46] L. Wang, F. H. Yang, R. T. Yang, M. A. Miller, *Ind. Eng. Chem. Res.* **2009**, *48*, 2920.
[47] G. M. Psofogiannakis, G. E. Froudakis, *J. Phys. Chem. C* **2009**, *113*, 14908.
[48] H.-Y. Wu, X. Fan, J.-L. Kuo, W.-Q. Deng, *J. Phys. Chem. C* **2011**, *115*, 9241.
[49] L. Warczinski, C. Hättig, *J. Phys. Chem. C* **2021**, *125*, 9020.
[50] H.-D. Meyer, U. Manthe, L. Cederbaum, *Chem. Phys. Lett.* **1990**, *165*, 73.
[51] U. Manthe, H. Meyer, L. S. Cederbaum, *J. Chem. Phys.* **1992**, *97*, 3199.
[52] M. Beck, A. Jäckle, G. Worth, H.-D. Meyer, *Phys. Rep.* **2000**, *324*, 1.
[53] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.
[54] J. Behler, *J. Chem. Phys.* **2011**, *134*, 074106.
[55] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[56] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2013**, *9*, 3404.
[57] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chem. Rev.* **2021**, *121*, 9759.
[58] A. Grisafi, D. M. Wilkins, G. Csányi, M. Ceriotti, *Phys. Rev. Lett.* **2018**, *120*, 036002.
[59] M. Ceriotti, M. J. Willatt, G. Csányi, *Machine Learning of Atomic-Scale Properties Based on Physical Principles*, Springer International Publishing, Cham **2018**, p. 1.
[60] S. G. Balasubramani, G. P. Chen, S. Coriani, M. Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, A. Hellweg, B. Helmich-Paris, C. Holzer, U. Huniar, M. Kaupp, A. Marefat Khah, S. Karbalaei Khani, T. Müller, F. Mack, B. D. Nguyen, S. M. Parker, E. Perlt, D. Rappoport, K. Reiter, S. Roy, M. Rückert, G. Schmitz, M. Sierka, E. Tapavicza, D. P. Tew, C. van Wüllen, V. K. Voora, F. Weigend, A. Wodyński, J. M. Yu, *J. Chem. Phys.* **2020**, *152*, 184107.
[61] Development version of TURBOMOLE V7.5 2020, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007. https://www.turbomole.org.
[62] A. Schäffer, H. Horn, R. Ahlrichs, *J. Chem. Phys.* **1992**, *97*, 2571.
[63] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.
[64] J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, *Phys. Rev. Lett.* **2003**, *91*, 146401.
[65] O. Christiansen, D. G. Artiukhin, I. H. Godtliebsen, E. M. Gras, W. Győrffy, M. B. Hansen, M. B. Hansen, E. L. Klinting, J. Kongsted, C. König, MIDASCPP: Molecular interactions, dynamics and simulations chemistry program package in C++, https://midascpp.gitlab.io/.
[66] C. Igel, M. Hüsken, *Proc. Second Int. Symp. Neural Comput.* **2000**, 115.
[67] L. C. Blum, J.-L. Reymond, *J. Am. Chem. Soc.* **2009**, *131*, 8732.
[68] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864.
[69] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Scientif. Data* **2014**, *1*, 140022.
[70] C. Schran, J. Behler, D. Marx, *J. Chem. Theory Comput.* **2020**, *16*, 88.
[71] F. A. Faber, A. S. Christensen, B. Huang, O. A. von Lilienfeld, *J. Chem. Phys.* **2018**, *148*, 241717.
[72] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*, 2087.
[73] J. Kubečka, A. S. Christensen, F. R. Rasmussen, J. Elm, *Environ. Sci. Technol. Lett.* **2022**, *9*, 239.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: G. Schmitz, B. Schnieder, *J. Comput. Chem.* **2023**, *44*(6), 732. https://doi.org/10.1002/jcc.27035

1096987x, 2023, 6, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/jcc.27035 by Ruhr-Universität Bochum, Wiley Online Library on [29/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License