JCTC Journal of Chemical Theory and Computation

Article

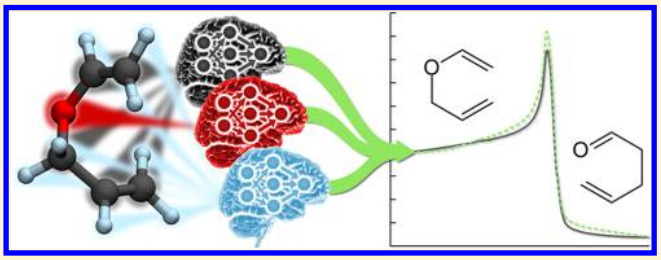
pubs.acs.org/JCTC

High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm

Michael Gastegger and Philipp Marquetand*

Institute of Theoretical Chemistry, University of Vienna, Währinger Str. 17, 1090 Vienna, Austria

ABSTRACT: Artificial neural networks (NNs) represent a relatively recent approach for the prediction of molecular potential energies, suitable for simulations of large molecules and long time scales. By using NNs to fit electronic structure data, it is possible to obtain empirical potentials of high accuracy combined with the computational efficiency of conventional force fields. However, as opposed to the latter, changing bonding patterns and unusual coordination geometries can be described due to the underlying flexible functional form of the NNs. One of the most promising approaches in this field is the high-dimensional neural network (HDNN) method, which is especially adapted to the prediction of molecular properties. While HDNNs have been mostly used to model solid state systems and surface interactions, we present here the first application of the HDNN approach to an organic reaction, the Claisen rearrangement of allyl vinyl ether to 4-pentenal. To construct the corresponding HDNN potential, a new training algorithm is introduced. This algorithm is termed "element-decoupled" global extended Kalman filter (ED-GEKF) and is based on the decoupled Kalman filter. Using a metadynamics trajectory computed with density functional theory as reference data, we show that the ED-GEKF exhibits superior performance — both in terms of accuracy and training speed — compared to other variants of the Kalman filter hitherto employed in HDNN training. In addition, the effect of including forces during ED-GEKF training on the resulting potentials was studied.



1. INTRODUCTION

Computational chemistry is a tightrope walk between accuracy and efficiency. Evidently, a sufficiently accurate potential energy surface (PES) is required in order to provide a reasonable description of a molecular system. However, when dealing with large systems or long simulation times, computational efficiency becomes an additional concern, and a compromise has to be found.

The most accurate PESs are provided by electronic structure methods, based on first-principles quantum mechanics, albeit at considerable computational expense, due to the explicit treatment of the many-electron problem. As a result, only systems of limited size and relatively short time-scales are accessible on a routine basis, when using these methods.[1,2]

Empirical force fields are several orders of magnitude faster to evaluate, since the PES is described as a sum of physically motivated analytic functions fitted to experimental or computational reference data. However, as a consequence of the fitting procedure involved, empirical potentials are only accurate for limited regions of the PES. Another drawback is the inability to describe bond breaking and bond formation events, as well as unusual coordination geometries and bonding situations, due to predefined atom types and the form of the analytic functions employed.[3,4] While so-called reactive force fields are capable of addressing this kind of problems, the accuracy of these approaches is ultimately still limited by the underlying physical approximations.[5]

A promising alternative is the use of machine learning (ML) techniques to construct PESs from electronic structure data. The highly flexible functional form of ML potentials allows for very accurate interpolation even of complicated PESs, as well as the description of complex and changing bonding patterns. Moreover, paired with a computational efficiency on par with empirical potentials, ML potentials can offer an accuracy comparable to high level electronic structure methods at the computational cost of force fields.[6]

Up to date, various ML approaches have been successfully applied to either the construction of PESs or the description of contributing terms, with Polynomial Fitting procedures,[7] Gaussian Processes,[8,9] Modified Shepard Interpolation,[10−13] Interpolating Moving Least Squares,[14,15] and Support Vector Machines[16] only being some of the most prominent examples. Particularly promising in this field are Neural Networks (NNs), a machine-learning technique inspired by the central nervous system.

The number of neural network potential energy surfaces (NN-PESs) has steadily increased over the course of the past decade, with applications ranging from surface science[17−25] to molecular systems.[26−45] For an exhaustive list of NN-PESs, see reviews.[46−48] These applications were accompanied by several advances to NN training algorithms and architectures, such as the inclusion of gradients in the training procedure.[42,49] Another example is the use of symmetrized NNs[43] or symmetry functions[18,50,51] to account for permutational invariance in the

Received: March 4, 2015
Published: April 16, 2015

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

input data and high dimensional representation schemes,[52,53] which employ multiple NNs for the construction of the PES.

An important step toward the routine use of NN-PESs in computational chemistry is the introduction of the high-dimensional neural network (HDNN) scheme by Behler and Parinello.[52] In the HDNN scheme, the PES is constructed from a sum of individual atomic energy contributions. These contributions depend on the atomic environment and are computed by a set of NNs, where a single NN is used for atoms of the same element. In order to describe the environment of the different atoms, a special set of atom centered symmetry functions (ACSFs) is employed.[51] The combination of these features allows HDNN potentials to overcome several limitations of standard NNs, the most important ones being limited system size, transferability of parameters, fixed amount of atoms and elements, and dependence of the NN-PES on translations and rotations in input space. The utility of HDNN potentials is demonstrated by a wide range of applications to solid state systems,[54−59] surface interactions,[21,22,60] and water clusters.[61−63]

Our work is based on these foundations and goes beyond in two ways: First, a novel training algorithm for HDNNs is presented. This "element-decoupled" Kalman filter is based on the decoupled formulation of the standard global extended Kalman filter (GEKF)[64] and exhibits superior performance in terms of training speed and accuracy of the final PES fits compared to other variants of the Kalman filter commonly used for HDNN training. Second, this algorithm is applied to the construction of HDNN-PESs for the Claisen rearrangement of allyl vinyl ether to 4-pentenal, which presents, to the best knowledge of the authors, the first application of the HDNN scheme to an organic reaction involving bond breaking and formation. Based on the resulting HDNN-PESs, the overall performance of the element-decoupled filter and its interpolation capabilities as well as the effect of the inclusion of forces during the training process are studied.

2. THEORETICAL BACKGROUND

This section is divided into two parts: The first part provides an outline of the NN architecture employed, with special focus on Behler−Parinello HDNNs. In the second part, the standard Kalman filter training algorithm will be reviewed, and adaptations thereof with regard to HDNNs will be discussed.

2.1. Neural Network Architecture. Similar to biological nervous systems, NNs consist of an arrangement of simple subunits, so-called neurons, often also referred to as nodes. These artificial neurons collect, process, and transmit incoming signals according to a predefined connection pattern, making NNs highly parallel, flexible, and robust nonlinear universal approximators capable of accurately fitting any continuous real function.[65−67] Due to these properties, NNs are typically employed in classification tasks, pattern recognition, and for the approximation of complicated functional relations, e.g. PESs.

One of the most frequently used NN architectures is the feed-forward multilayer perceptron.[68] An example for a NN of this type with one hidden layer and a single neuron in the output layer is depicted in Figure 1. In a feed-forward NN, signals are only transmitted into one direction. Starting with the vector of inputs $\mathbf{y}^0$ in the input layer, the output of a node j in layer l can be computed according to the regression

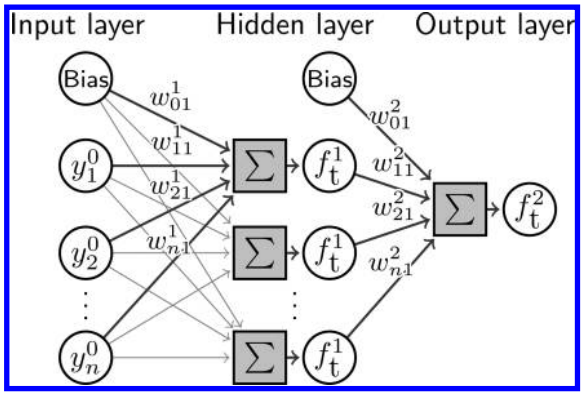


Figure 1. Feed-forward neural network with a single output node and one hidden layer. Bias nodes provide a constant offset to the transfer functions f^l_t. Nodes in adjacent layers are connected by the weights $\{w^l_{ij}\}$, the fitting parameters of the neural network.

$$y^l_j = f^l_t\left(w^l_{0j} + \sum_i^{n^{l-1}} w^l_{ij} y^{l-1}_i\right) \tag{1}$$

until the output layer is reached. The signals of the previous layer $\{y^{l-1}_i\}$ are scaled by a set of weight parameters $\{w^l_{ij}\}$, where w^l_{ij} is the weight connecting node i of the previous layer to node j in the current layer and w^l_{0j} is a bias weight, which provides an adjustable offset to the transfer function, hence adding additional flexibility. These weights are the adjustable parameters of the NN and are collected in the vector of weights $\mathbf{w}$, which has to be determined in the training process. After weighting and summation, a transfer function f^l_t is applied to the nodes in layer l. Typically, sigmoidal functions (e.g., hyperbolic tangents) are employed in the hidden layers, as they provide the NN with the ability to model any continuous, real function to arbitrary accuracy with a finite number of nodes in the hidden layer.[65−67] In the output layer, usually linear transfer functions are used. The output of the NN can be described as a function $f_{NN}(\mathbf{y}^0,\mathbf{w})$ depending on the input parameters and the vector of all weights.

Due to their robustness, computational efficiency, simplicity of training, and the availability of analytic gradients, feed-forward NNs are the type of NNs most commonly used in the construction of NN-PESs. However, several problems limit their applicability when employed in the interpolation of multidimensional PESs. Due to the predetermined structure of the NN after training, the output of the NN depends on the ordering of the inputs, in the case of a NN-PES the molecular coordinates. In a similar manner, the number of atoms that can be treated with such simple NNs must be proportional to the number of input nodes, making it impossible to treat molecules of different size with the same NN. In addition, the NN shows an unwanted variance with respect to translation and rotation of the molecular geometry if Cartesian coordinates are used.

One strategy to overcome these limitations is the HDNN scheme introduced by Behler and Parinello in 2007.[52] In this approach, the total energy E of a molecular system with N atoms is constructed from atomic energy contributions E_i according to

$$E = \sum_i^N E_i \tag{2}$$

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

The atomic energies depend on the chemical environment of atom i and are computed by individual NNs. Figure 2 shows an example for a Behler–Parinello HDNN.

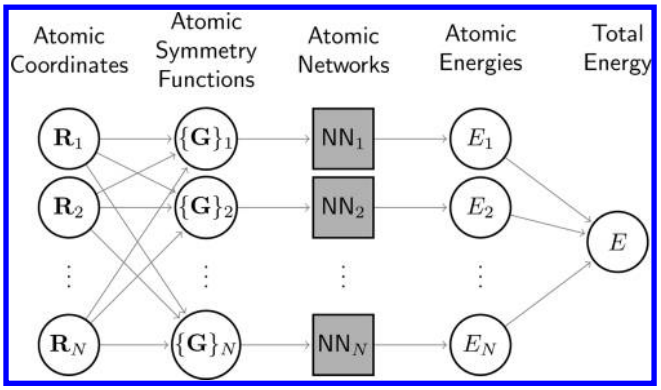


Figure 2. Scheme of a high-dimensional neural network of the Behler–Parinello type. The Cartesian coordinates $\{\mathbf{R}_i\}$ of a molecule are transformed into a set of atomic symmetry functions $\{\mathbf{G}_i\}$ describing an atom subject to its chemical environment. With these symmetry functions as inputs, the atomic contributions E_i to the total energy E are computed with atomic neural networks.

Since the total energy is now expressed as a sum of NN contributions E_i, any dependence on the ordering of inputs is eliminated. Moreover, by using one NN for nuclei of the same element, the HDNN can model molecular systems of arbitrary size. For every new atom, the sum in eq 2 is simply augmented by the energy contribution of the corresponding elemental NN. To ensure invariance with respect to translation and rotation of the molecule, the Cartesian coordinates $\{\mathbf{R}_i\}$ are transformed to a set of many-body symmetry functions $\{\mathbf{G}_i\}$, depending on the internuclear distances and angles between all atoms. These ACSFs describe the local chemical environment of atom i via radial and angular distributions of the surrounding nuclei.[51] An example for a radial ACSF is

$$G_i^{\mathrm{rad}} = \sum_{j\neq i}^{N_{\mathrm{atoms}}} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}) \tag{3}$$

while an angular distribution can be characterized by

$$G_i^{\mathrm{ang}} = 2^{1-\zeta} \sum_{j,k\neq i}^{N_{\mathrm{atoms}}} \left(1 + \lambda \frac{\mathbf{R}_{ij}\mathbf{R}_{ik}}{R_{ij}R_{ik}}\right)^{\zeta} e^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)} \times f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{4}$$

Here, R_{ij} is the distance between atoms i and j, $\mathbf{R}_{ij}$ is the vector $\mathbf{R}_j - \mathbf{R}_i$, and R_s, η, ζ, and λ are parameters which determine the shape of the ACSFs. Typically, a combination of radial and angular ACSFs $\{\mathbf{G}_i\}$, differing in the parameters R_s, η, ζ, and λ, is employed in the description of every atom. In addition, a radial cutoff R_c is introduced around the central atom in the form of a cutoff function

$$f_c(R_{ij}) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi R_{ij}}{R_c}\right) + 1\right], & R_{ij} \leqslant R_c \\ 0, & R_{ij} > R_c \end{cases} \tag{5}$$

This reduces the local chemical environment to the energetically relevant regions, leading to a linear scaling of the HDNN computation cost with system size.

Due to the well-defined functional form of the individual elemental NNs, an analytic expression for the HDNN forces can be derived. The force with respect to the Cartesian coordinates $\mathbf{R}_j$ of atom j is given by the relation

$$\mathbf{F}^{(j)} = -\sum_{i}^{N} \sum_{\alpha}^{M_i} \frac{\partial E_i}{\partial G_{i,\alpha}} \frac{\partial G_{i,\alpha}}{\partial \mathbf{R}_j} \tag{6}$$

where M_i is the number of ACSFs used to describe nucleus i.[51] The first term in eq 6 is the derivative of the elemental NNs with respect to the ACSFs, and the second term is the derivative of the ACSFs with respect to the Cartesian coordinates of atom j.

2.2. HDNN Training. A NN is defined by its structure and its weights $\mathbf{w}$. Once the structure – the number of hidden layers and nodes in the hidden layers – has been determined empirically, the initially random NN weights need to be optimized in order to reproduce the desired function, a process also called "training" of the NN.

Several different NN training algorithms exist, from gradient-based approaches, such as the so-called backpropagation algorithm,[69] to second-order methods, e.g. the Levenberg–Marquardt optimization.[70] One second-order training method that is particularly promising for the construction of NN-PESs is the GEKF.[71,72] In the GEKF, every sample (e.g., molecular geometry) is presented to the algorithm sequentially, immediately followed by an update of the weight vector $\mathbf{w}$. The correction of $\mathbf{w}$ is based on the current error and a weighted history of previous update steps. In the case of PES fitting, the error ν_k of the current update step k, also referred to as filter innovation, is

$$\nu_k = E_k - \tilde{E}_k \tag{7}$$

with E_k being the reference energy (from a quantum-chemical calculation), and $\tilde{E}_k$ being the corresponding NN-PES value. The direction in which $\mathbf{w}$ is changed in each update step is based on the current Jacobian matrix $\mathbf{J}_k$ – the NN gradient with respect to the weights – while the magnitude of the update is given by the filter covariance matrix $\mathbf{P}_k$. In short, $\mathbf{P}_k$ is the weighted history of Gauss–Newton approximations to the inverse Hessian of the modeling error surface and the main reason for the excellent training properties of the Kalman filter. For a detailed explanation of the GEKF for standard NNs as well as the associated recursion scheme and derivation thereof, we refer to refs 49, 71, and 72.

For a HDNN of the Behler-Parinello type, $\tilde{E}_k$ in eq 7 becomes the HDNN total energy given in eq 2. Since an individual NN is now used for every element with atomic number Z, the different elemental weight vectors $\mathbf{w}_k^{(Z)}$ need to be optimized simultaneously with the help of their respective error covariances $\mathbf{P}_k^{(Z)}$. One possible approach to HDNN training is to use an "atomic" Kalman filter (A-GEKF), where an update of the corresponding $\mathbf{w}_k^{(Z)}$ is performed for every atom i in the molecular sample of step k according to the regression scheme:

$$\mathbf{w}_k^{(Z)} = \mathbf{w}_{k-1}^{(Z)} + \mathbf{K}_k^{(i)} N_{\mathrm{atom}}^{-1} \nu_k \tag{8}$$

$$\mathbf{K}_k^{(i)} = \mathbf{P}_{k-1}^{(Z)} \mathbf{J}_k^{(i)} \left[\lambda_k^{(Z)} \mathbf{I} + \left(\mathbf{J}_k^{(i)}\right)^T \mathbf{P}_{k-1}^{(Z)} \mathbf{J}_k^{(i)}\right]^{-1} \tag{9}$$

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187–2198

$$\mathbf{P}_k^{(Z)} = (\lambda_k^{(Z)})^{-1}\left[\mathbf{I} - \mathbf{K}_k^{(i)}\left(\mathbf{J}_k^{(i)}\right)^T\right]\mathbf{P}_{k-1}^{(Z)} \tag{10}$$

Here, $\mathbf{K}_k^{(i)}$ is the Kalman gain matrix computed for atom i, $\mathbf{I}$ is the identity matrix, and the expression in brackets on the right-hand side of eq 9 is a scaling matrix. The λ_k are time-varying forgetting factors computed according to $\lambda_k = \lambda_{k-1}\lambda_0 + 1 - \lambda_0$, which are introduced in order to reduce the risk of convergence to local minima early in the training process.[71] Both λ_k and λ_0 are typically initialized close to unity. For the update of the weights (eq 8), the total error is averaged over the number of atoms N_{atom}, because the individual atomic contribution to the error is unknown. This assumption is problematic, especially in combination with atomic weight updates, since the weights of some elements are updated more frequently, leading to a smaller relative error in the weights when compared to the ones of less abundant elements. This uneven distribution of the errors imposes a severe limitation to the quality of the PES fit obtained.

An alternative is to model the HDNN scheme as one large composite NN. Using the sum rule, it is possible to show that the Jacobian $\mathbf{J}_k^{(Z)}$ associated with every element can be computed as

$$\mathbf{J}_k^{(Z)} = \sum_i^{N_{\text{atom}}} \mathbf{J}_k^{(i)}\delta_{Z_i,Z} \tag{11}$$

where $\delta_{Z_i,Z}$ is one if the element of the current atom Z_i corresponds to the elemental index Z of the Jacobian and zero otherwise. Instead of individual updates for all atoms, one update per element is now performed for every molecule, according to

$$\mathbf{w}_k^{(Z)} = \mathbf{w}_{k-1}^{(Z)} + \mathbf{K}_k^{(Z)} N_{\text{elem}}^{-1}\nu_k \tag{12}$$

$$\mathbf{K}_k^{(Z)} = \mathbf{P}_{k-1}^{(Z)}\mathbf{J}_k^{(Z)}\left[\lambda_k\mathbf{I} + \left(\mathbf{J}_k^{(Z)}\right)^T\mathbf{P}_{k-1}^{(Z)}\mathbf{J}_k^{(Z)}\right]^{-1} \tag{13}$$

$$\mathbf{P}_k^{(Z)} = \lambda_k^{-1}\left[\mathbf{I} - \mathbf{K}_k^{(Z)}\left(\mathbf{J}_k^{(Z)}\right)^T\right]\mathbf{P}_{k-1}^{(Z)} \tag{14}$$

In this "elemental" GEKF (E-GEKF), the bias due to the frequency of elements is eliminated, as every element is only updated once. However, it is still assumed that the elemental NNs contribute to the total error in a similar fashion. Since no clear guidelines on how to treat the total error in eq 12 exist, ad hoc corrections have to be introduced e.g. by weighting ν_k with the elemental fraction or, as is the case here, by dividing by the number of elements present in the molecule.

Compared to gradient-based NN training methods such as backpropagation, second-order algorithms based on the GEKF offer superior accuracy and convergence behavior, albeit at an increased computational cost. For every sample in the training data, several matrix operations including inversion have to be performed, and several passes over the training data, so-called "epochs", are required in order to obtain a suitably trained NN. However, the computational effort associated with HDNN training is still negligible compared to the generation of the electronic structure reference data, which currently represents the bottleneck in NN-PES construction. Furthermore, the efficiency of the GEKF can be increased by employing an adaptive threshold to the filter updates.[71] In this approach, the error of the current sample is compared to a threshold defined as a fixed fraction of the total root-mean-square error (RMSE) of the previous epoch. An update of the weights is only performed, if the error exceeds the threshold. In this way, the number of unproductive updates can be kept to a minimum, leading to an efficiency almost on par with standard backpropagation, while retaining all advantages of the GEKF.

3. A NEW TRAINING ALGORITHM

In the following, we present a new GEKF variant for HDNN training, which requires neither partitioning of the total error, nor any assumptions regarding its distribution among the subnets. This approach exhibits superior fitting performance and filter convergence behavior compared to the atomic and elemental GEKF algorithms. This improved filter algorithm is based on a variant of the GEKF initially proposed to reduce its computational effort, the so-called "decoupled" Kalman filter.[64,73]

3.1. Element-Decoupled Kalman Filter.

In the decoupled Kalman filter, sets of weights are treated as independent from each other in order to reduce the dimensionality of the matrix operations. Since it can be safely assumed that the weight vectors $\mathbf{w}_k^{(Z)}$ of the different elemental NNs in a HDNN fulfill this criterion, a decoupled Kalman scheme can be applied to HDNNs in the form of an "element-decoupled" GEKF (ED-GEKF). In this scheme, the computation of the Kalman gain in eq 13 is modified to

$$\mathbf{K}_k^{(Z)} = \mathbf{P}_{k-1}^{(Z)}\mathbf{J}_k^{(Z)}\left[\lambda_k\mathbf{I} + \sum_z^{N_{\text{elem}}}\left(\mathbf{J}_k^{(z)}\right)^T\mathbf{P}_{k-1}^{(z)}\mathbf{J}_k^{(z)}\right]^{-1} \tag{15}$$

where the expression in the brackets now depends on the subnets of all elements (indicated by z). In this way, the total error can be used in the weight update, without introducing additional assumptions regarding the distribution of individual atomic or elemental errors, making the "element-decoupled" GEKF an excellent training algorithm for the HDNN network architecture. Hence, eq 12 is modified to

$$\mathbf{w}_k^{(Z)} = \mathbf{w}_{k-1}^{(Z)} + \mathbf{K}_k^{(Z)}\nu_k \tag{16}$$

Moreover, it is also possible to extend the element-decoupled GEKF to include the forces in the training process. The force innovation associated with atom i is given by

$$\boldsymbol{\xi}_k^{(i)} = \mathbf{F}_k^{(i)} - \tilde{\mathbf{F}}_k^{(i)} \tag{17}$$

where $\mathbf{F}_k^{(i)}$ is the atomic reference force as computed by e.g. electronic structure methods, and $\tilde{\mathbf{F}}_k^{(i)}$ is the force obtained for the HDNN via eq 6. To incorporate forces in the element-decoupled scheme, only the weight update (eq 16) has to be changed to

$$\mathbf{w}_k^{(Z)} = \mathbf{w}_{k-1}^{(Z)} + \mathbf{K}_k^{(Z)}\nu_k + \mathbf{P}_{k-1}^{(Z)}\sum_i^{N_{\text{atom}}}\frac{\partial\tilde{\mathbf{F}}_k^{(i)}}{\partial\mathbf{w}_k^{(Z)}}\mathbf{B}_k^{(i)}\boldsymbol{\xi}_k^{(i)} \tag{18}$$

where $(\partial\tilde{\mathbf{F}}_k^{(i)})/(\partial\mathbf{w}_k^{(Z)})$ is the derivative of the HDNN forces with respect to the weights. $\mathbf{B}_k^{(i)}$ is a scaling matrix computed according to

$$\mathbf{B}_k^{(i)} = \left[\lambda_k\mathbf{I} + \sum_z^{N_{\text{elem}}}\left(\frac{\partial\tilde{\mathbf{F}}_k^{(i)}}{\partial\mathbf{w}_k^{(z)}}\right)^T\mathbf{P}_{k-1}^{(z)}\frac{\partial\tilde{\mathbf{F}}_k^{(i)}}{\partial\mathbf{w}_k^{(z)}}\right]^{-1} \tag{19}$$

This scaling matrix is required, since the derivation of the forces with respect to the weights produces a component vector for

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

every elemental NN and the contribution of each subnet to the overall atomic force error is unknown.

The ED-GEKF algorithm is straightforward to implement, and the associated computational performance is similar to or even better than the other filter variants. Compared to both the A-GEKF and E-GEKF, the presented algorithm exhibits a significant increase in the quality of the obtained PES fits, as well as improved convergence behavior (see Section 5). These properties open up new possibilities in the construction of HDNN potentials for a multitude of compound classes and reaction types. The ED-GEKF is expected to perform especially well for systems with several different elements (e.g., organic molecules, proteins), since the distribution of errors between the elemental subnets is inherently modeled by the algorithm. In addition, forces can be included in the training process in a consistent manner, albeit at an increased computational cost, due to the more expensive computation of the derivatives of the HDNN forces with respect to the weights compared to the standard Jacobian.

4. COMPUTATIONAL DETAILS

In order to apply the new training algorithm in the simulation of an organic reaction, reference electronic structure calculations were carried out with ORCA[74] at the BP86/def2-SVP[75−80] level of theory employing the resolution of identity approximation.[81,82] The obtained energies and gradients were augmented by the empirical D3 dispersion correction of Grimme[83] with the Becke−Johnson damping scheme.[84] A minimum energy reaction path of 500 intermediate geometries describing the transition between substrate and product was generated and optimized at the same level of theory with the WOELFLING chain-of-states method[85] implemented in TURBOMOLE.[86,87] Energies and gradients of these geometries were then recomputed with ORCA in order to ensure consistency of the reference data.

Born−Oppenheimer dynamics[88] and metadynamics[89] were carried out using electronic-structure gradients. Newton's equations of motion for the nuclei were integrated using the Velocity-Verlet algorithm[90] with timesteps of 0.5 fs. A Berendsen thermostat was employed in all simulations.[91] Initial velocities were sampled from a Maxwell−Boltzmann distribution at the corresponding bath temperature.[92] Metadynamics simulations were performed according to the scheme of Laio and Parinello,[89] utilizing the interatomic distances r_{O-C_5} and $r_{C_2-C_3}$ (see Figure 3) as collective variables to describe the Claisen rearrangement reaction.

Behler-type ACSFs were used to represent the chemical environment of the individual atoms in the HDNN-scheme.[51] A set of 10 radial and 32 angular distribution functions was employed for hydrogen and carbon atoms, while 7 radial and 20 angular functions were used for the single oxygen atom. The different number of ACSFs for oxygen and the other elements is due to combinatorial reasons. Since there is only one oxygen atom present in the molecule, it can only have C and H atoms as neighbors, while a carbon or hydrogen atom can have the neighbors C, H, and O. For all ACSFs a radial cutoff of 10.0 Å was used.

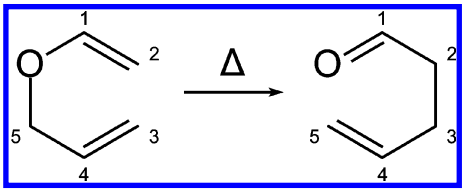


Figure 3. Thermal, aliphatic Claisen rearrangement reaction of allyl vinyl ether to 4-pentenal. Over the course of the reaction, the $O-C_5$ bond is broken, and a new bond is formed between C_2 and C_3.

Elemental subnets of different architectures were employed in the construction of the HDNN-PESs. The number of nodes in the input layer was constrained to the number of symmetry functions used in the description of the atomic environment (42 for hydrogen and carbon and 27 for oxygen), and a single node was used in the output layer. Hyperbolic tangent activation functions were used for the hidden layers, while a linear transformation was applied to the output layer of all networks. Since the elemental NN architectures can differ only in their respective hidden layers, a shorthand notation will be introduced, where e.g. C-40-40 refers to a subnet for the element carbon with two hidden layers of 40 nodes each. For the system at hand, elemental subnets of different size were tested, and it was found that two hidden layers of 40 nodes for every element offer the best compromise between accuracy and computational efficiency.

The weights of all NNs were initialized at random according to the scheme of Nguyen and Widrow.[93] In order to facilitate HDNN training, all symmetry functions derived from the reference data set and the corresponding reference energies were normalized to obtain an average of zero and a standard deviation of $\sigma = 1.0$ over the whole set. Overfitting of the reference data was detected by means of an early stopping procedure based on cross-validation.[94] In this approach, the reference data set is divided into a training and validation set. Only points in the training set are used in the construction of the HDNN potential, while the RMSEs of both data sets are monitored during training. At the onset of overfitting, the RMSE of the training set continues to decrease, while the RMSE of the validation set begins to increase. At this point, the fitting procedure is stopped, and the set of weights associated with the lowest validation RMSE is returned. In this work, the reference data was split into a training and validation set with a ratio of 9:1 in a random manner, and the training procedure was terminated after three successive training epochs with an increase in the validation RMSE. A ratio of 9:1 was chosen as a compromise between a sufficiently dense spacing of data points in the training set and a large enough validation set to reliably detect overfitting. The elemental covariance matrices required by the different Kalman filter variants were initialized as diagonal matrices $\mathbf{P}^{(Z)} = \delta^{-1}\mathbf{I}$, with a scaling factor $\delta = 0.01$. Values of $l_k = 0.99$ and $l_0 = 0.996$ were used for the time-varying forgetting schedule. The adaptive Kalman filter threshold was set to 0.9 times the RMSE of the training set in the previous epoch.[71]

Training of the HDNNs, as well as dynamics and metadynamics simulations, were carried out with a suite of programs developed in PYTHON[95] using the NUMPY package.[96]

5. RESULTS AND DISCUSSION

5.1. Model Reaction and Reference Data Generation.

To study the performance of the different Kalman filter variants and the effect of including forces into the training procedure, the aliphatic Claisen rearrangement of allyl vinyl ether was chosen as a model reaction (Figure 3). This reaction is a [3,3]-sigmatropic rearrangement, where the substrate allyl vinyl ether is converted thermally to the aldehyde 4-pentenal via bond

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

breaking and bond formation events. The transformation is irreversible and occurs at elevated temperatures. For an extensive review of this kind of rearrangement reaction, see ref 97 and references therein.

Several traits make this particular reaction an excellent subject of study for the construction of HDNN-PESs: The encountered molecules are comparatively small, rendering the generation of electronic structure reference data inexpensive. More than two different elements are present, leading to a multitude of chemical environments and allowing to test the applicability of the HDNN scheme to organic systems. Most importantly, different and changing bonding patterns have to be described in order to accurately model the rearrangement, which is one of the major advantages of ML potentials compared to standard empirical potentials.

The Claisen rearrangement reference data set for HDNN training was generated in a metadynamics run at 400 K using electronic structure energies and gradients. After an initial period of equilibration, Gaussians with a height of 6.28 kcal mol^{-1} and widths of 0.27 Å were deposited along the dimensions of the collective variables every 100 steps. The system was propagated for a total of 8.55 ps simulation time. Geometries, energies, and forces were collected every step, resulting in a total of 17100 molecules in the final reference data set.

It should be stressed at this point that the main focus of the present study is the accurate reproduction of a quantum chemically derived PES and not the accuracy of the underlying reference PES. Concerns regarding e.g. the character of the encountered species, quality of reaction barriers, and viability of the employed electronic structure method are therefore only of secondary importance. The BP86 functional was chosen in favor over more sophisticated methods primarily due to its overall computational cheapness and robustness, which further facilitated the metadynamics sampling procedure. All HDNN training methods presented in this work can be applied to any electronic structure method without additional adaptations.

5.2. ED-GEKF Performance. The performance of the different Kalman filter variants was studied based on HDNN-PESs constructed from HDNNs trained with the A-GEKF, E-GEKF, and the newly developed ED-GEKF algorithm, respectively. The employed HDNNs consisted of C-40-40, H-40-40, and O-40-40 subnets and were trained for 100 epochs. In order to account for the dependence of the final PES fit on the initial random partitioning of the data into training and validation sets, as well as the randomly initialized network weights, five HDNNs were trained with each filter, and the averaged results are reported here.

The RMSEs per molecule of the obtained HDNN-PES fits relative to the reference data set are shown in Figure 4. A clear improvement in the quality of the PESs is observed when going from the A-GEKF (0.99 kcal mol^{-1} RMSE for the training set and 1.16 kcal mol^{-1} for the validation set) over the E-GEKF (0.19 kcal mol^{-1} and 0.27 kcal mol^{-1}) to the ED-GEKF (0.08 kcal mol^{-1} and 0.13 kcal mol^{-1}). Especially the gain in accuracy obtained with the elemental and element-decoupled filter variants compared to the A-GEKF is pronounced. The reason for this behavior is most likely the elemental weight update in the A-GEKF algorithm, which is performed for each atom individually. This procedure introduces a bias in favor of more abundant elements, which — in combination with the assumption of an even distribution of the total error between

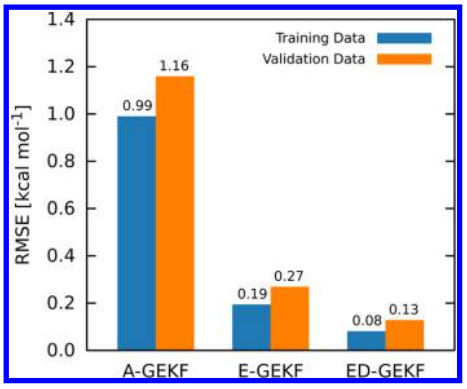


Figure 4. RMSEs (kcal mol^{-1}) for the potential energies predicted by HDNNs trained with different Kalman filter variants.

the individual subnets — ultimately limits the quality of PESs obtainable with this particular filter variant (see Section 2.2).

While the E-GEKF already shows superior performance compared to the A-GEKF, the HDNNs trained with the ED-GEKF reproduce the reference data even more accurately. Using the latter training algorithm, the RMSEs over the training and validation set are reduced to less than half of their corresponding E-GEKF values. Since the update frequency bias present in the A-GEKF is eliminated in both E-GEKF and ED-GEKF, this additional increase in fit quality can be solely attributed to the improved treatment of the error distribution between the subnets in the ED-GEKF. In contrast, no clear approach exists on how to distribute the total error between the respective subnets in the E-GEKF weight update (see eq 12), and one of the various imaginable empirical weighting schemes needs to be introduced. Several alternative schemes were investigated prior to this work, and the best results for the system at hand were achieved by averaging the total error evenly over the number of elements present in the current molecule N_{elem}. All data for the E-GEKF reported here was obtained with this weighting scheme. However, an unfortunate partitioning of the total error can easily lead to inferior fit accuracies on par with or even worse than those yielded by the A-GEKF algorithm, a problem not encountered with the ED-GEKF variant.

The aforementioned trends are even more pronounced, when the maximum deviations of the HDNN-PESs from the reference energies are compared (Figure 5). As above, changing from the A-GEKF to the other filters leads to a substantial improvement in the quality of the fit. The best accuracy is once

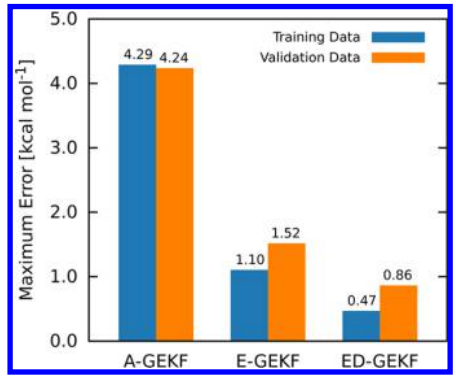


Figure 5. Maximum errors (kcal mol^{-1}) for the potential energies predicted by HDNNs trained with different Kalman filter variants.

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

again obtained with the ED-GEKF, exhibiting maximum errors smaller than 1.0 kcal mol^{-1} over the training as well as the validation data set. These results are an excellent demonstration for the high accuracy of HDNN-PESs produced by the ED-GEKF training algorithm.

With regards to computational efficiency, the A-GEKF algorithm performs worst, since the computation of the Kalman gain matrix, the weight update and the update of the covariance matrix are carried out for every individual atom present in the molecule. In contrast, these computations are only performed once per element in the E-GEKF and ED-GEKF variants, leading to a much better scaling behavior, especially for larger molecules. The major difference between the E-GEKF and ED-GEKF is that the scaling matrix (expression in the square brackets in eqs 13 and 15, respectively) has to be computed only once per molecular system in the ED-GEKF algorithm, while in the E-GEKF it is evaluated for every element. In order to calculate this expression, $n_{\mathrm{out}}^{(Z)} \times n_{\mathrm{out}}^{(Z)}$ matrices have to be inverted, where $n_{\mathrm{out}}^{(Z)}$ is the number of output nodes of the elemental subnets. Since a matrix inversion usually is an expensive operation in terms of computation time, the ED-GEKF is expected to perform even better than the E-GEKF in theory. In the special case of HDNNs, the elemental subnets possess only one output node (the atomic energy), and the matrix to be inverted is therefore a 1 × 1 matrix. Hence, the inversion operation reduces to a standard division, and virtually no difference in the computational efficiency between the E-GEKF and ED-GEKF variants is observed in praxis. In the implementation used in this work, both the ED-GEKF and E-GEKF algorithms are approximately 2.4 times faster than the A-GEKF for the system at hand from a purely computational perspective. It should be noted, however, that compared to the E-GEKF and A-GEKF variants, fewer iterations are needed for ED-GEKF to obtain accurate HDNN-PESs, making it the faster algorithm overall, as shown in the next section.

5.3. Filter Convergence. In addition to the quality of the fit, the evolution of the RMSEs over the course of the training procedure was investigated. The corresponding results for the training set are depicted in Figure 6. As before, the averages over five HDNNs for each filter variant are reported.

The atomic Kalman filter shows signs of convergence only at the end of the 100 training epochs, and even longer training periods would be required to fully converge the results. The E-GEKF can be considered converged after approximately 60 epochs, as only minor changes in the training RMSEs (smaller than 0.01 kcal mol^{-1}) are observed afterward. In case of the ED-GEKF algorithm, convergence is already reached after 40 epochs of training, and only small corrections to the network weights are performed after this point, once again demonstrating its excellent viability for HDNN training. Moreover, if training speed is of primary concern, sufficiently accurate HDNN-PESs with an average training RMSE of 0.12 kcal mol^{-1} can already be obtained after 20 training epochs. Even in this case, the accuracy of the ED-GEKF fit still exceeds the levels achieved by the A-GEKF and the E-GEKF variants after the full 100 epochs of training (0.99 kcal mol^{-1} and 0.19 kcal mol^{-1} training RMSE, respectively). Curves for the validation RMSEs, as well as the maximum deviations, are not shown here as they paint an identical picture.

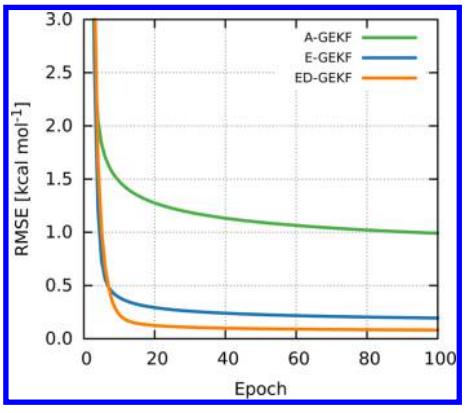


Figure 6. Evolution of the training set RMSEs (kcal mol^{-1}) during the training process for different GEKF variants.

5.4. Training including Forces. All HDNN-PESs presented in this work, up to this point, were created employing only energy data in the fitting procedure. It is, however, also possible to extend the ED-GEKF algorithm to perform a simultaneous fit of energies and forces (see Section 3.1). In order to study the effects of the inclusion of forces in the training procedure, a HDNN of the same dimensions as before (C-40-40, H-40-40, O-40-40) was trained with the force variant of the ED-GEKF (ED-GEKF+F). Due to the increased computational expense compared to the standard algorithm, only a single HDNN fit was performed. The required training time was further reduced by exploiting the fast convergence of the ED-GEKF and limiting the training to 40 epochs. The RMSEs over energies and forces obtained with the ED-GEKF+F and the other filter variants are given in Table 1. Since the trends observed for the training and validation data set are similar and the primary focus lies on the general performance of the different training algorithms, the RMSEs computed over the whole reference data set are reported here. In addition, due to the availability of only one HDNN for the ED-GEKF+F case, the values given for the A-GEKF, E-GEKF, and ED-GEKF are those obtained with the respective HDNNs exhibiting the smallest RMSE on the validation data set after training.

Table 1. RMSEs of Energies (kcal mol^{-1}) and Forces (kcal mol^{-1} Å^{-1}) over the Whole Reference Data Set

	RMSE	
filter type	energies	forces
A-GEKF	0.86	18.66
E-GEKF	0.20	12.40
ED-GEKF	0.08	11.84
ED-GEKF+F	0.17	6.79

For the HDNNs based on the energy data only, a similar pattern emerges for the forces as previously for the energies. While the A-GEKF is associated with the greatest deviation in the forces (18.66 kcal mol^{-1} Å^{-1}), the difference between the elemental and element-decoupled filters is less pronounced (12.40 kcal mol^{-1} Å^{-1} and 11.84 kcal mol^{-1} Å^{-1}, respectively) but still clearly in favor of the ED-GEKF, which produces the more accurate fit even in this case.

Compared to the standard element-decoupled algorithm, the force variant leads to a dramatic increase in the accuracy of the HDNN-PES forces. The RMSE of the forces over the whole reference data set is reduced to 6.79 kcal mol^{-1} Å^{-1}, almost half of the corresponding ED-GEKF value. However, while forces

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

are reproduced more accurately, a deterioration in the quality of the HDNN-PES energies is observed, with the RMSE increasing from 0.08 kcal mol^{-1} for the ED-GEKF HDNN to 0.17 kcal mol^{-1} for its ED-GEKF+F counterpart. This phenomenon originates from the second fit criterion introduced in the ED-GEKF+F algorithm in the form of the molecular forces. Since every molecule contributes only one energy but $3N_{atom}$ force components to the fitting process, a bias toward the fit of the forces over the energies is introduced. A similar strategy as in the combined function derivative approximation (CFDA) method by Pukrittayakamee and co-workers[42] can be employed to counteract this effect. By introducing a scaling factor for the force component of the weight update in eq 18, the relative importance of energies and forces during HDNN training can be regulated. However, an adaptation of this kind will be the subject of future research. Nevertheless, the deterioration in energies observed in this work is only minor (0.09 kcal mol^{-1}) compared to the greatly improved forces. This feature makes the force variant of the ED-GEKF algorithm especially attractive for cases where accurate forces are required, e.g. molecular dynamics simulations. In addition, since the ED-GEKF+F incorporates not only information on the energies but also information on the gradients (forces) in a similar manner to the CFDA method, no overfitting of the underlying PES should occur, thus eliminating the need for a cross-validation procedure. This implication will be tested in further studies.

5.5. HDNN Interpolation. One major feature of NNs is their capability to reliably interpolate data. To obtain insight into the interpolation performance of the HDNNs trained with the different Kalman filter variants, their ability to reproduce the electronic structure reaction profile of the Claisen rearrangement was studied. HDNN energies were computed for 500 intermediate geometries encountered along the reaction path (starting with allyl vinyl ether and ending in 4-pentenal) and compared to the corresponding reference values. The resulting reaction profiles are shown in Figure 7. RMSEs over

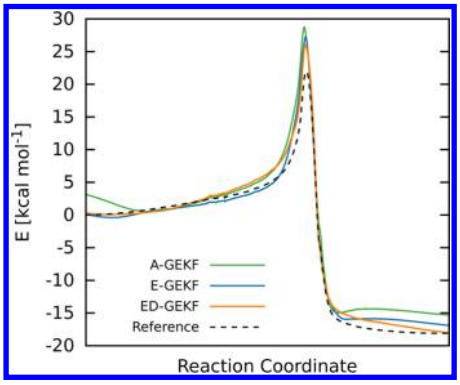


Figure 7. Reaction path of the Claisen rearrangement leading from allyl vinyl ether (left) to 4-pentenal (right) as obtained with the WOELFLING method implemented in TURBOMOLE. The curves for the different HDNNs correspond to the HDNN potential energies predicted for the geometries encountered along the profile.

the whole reaction path, as well as deviations of the HDNNs from the reference energies for the ether, the transition state (TS), and the aldehyde, are given in Table 2. As before, the HDNNs exhibiting the smallest validation RMSE for the respective filter algorithm were used.

Table 2. RMSEs of Energies (kcal mol^{-1}) along the Reaction Path and Deviations from the Reference Energy ΔE (kcal mol^{-1}) for Reactant (Ether), Transition State (TS), and Product (Aldehyde)

		ΔE		
filter type	RMSE	ether	TS	aldehyde
A-GEKF	2.48	3.12	6.81	2.88
E-GEKF	1.43	0.15	5.32	1.20
ED-GEKF	1.24	0.30	4.06	0.10

All investigated HDNN-PESs capture the general shape of the reaction profile, with a tendency to overestimate the barrier height. The HDNN trained with the A-GEKF algorithm yields the least accurate profile with a RMSE of 2.48 kcal mol^{-1}. In the region corresponding to allyl vinyl ether, the energy of the initial reactant is too high (ΔE = 3.12 kcal mol^{-1}), leading to an artificial minimum far from the original equilibrium structure of the ether. The onset of the A-GEKF's barrier, as well as the barrier itself, shows the greatest deviation from the reference energies out of all filter variants (ΔE = 6.81 kcal mol^{-1} for TS). The 4-pentenal region lies too high in energy (ΔE = 2.88 kcal mol^{-1}). In addition, a second artificial minimum is present immediately after the barrier. In case of the E-GEKF, better agreement with the reference curve is obtained (RMSE of 1.43 kcal mol^{-1}). The region preceding the barrier and the onset of the barrier are modeled accurately. Nevertheless, an artificial minimum can still be found close to the reactant. The barrier height is smaller than for the atomic filter variant (ΔE = 5.32 kcal mol^{-1} for TS). Although the region leading to the product is predicted closer to the reference (ΔE = 1.20 kcal mol^{-1}) by the E-GEKF, this part of the potential curve exhibits a similar shape as the one produced by the A-GEKF HDNN, showing an artificial minimum close to the barrier.

The most reliable reproduction of the original reaction profile is obtained by the ED-GEKF HDNN (RMSE of 1.24 kcal mol^{-1}). The region close to the ether is in good agreement with the reference. A small minimum is still present but less pronounced than in the other variants. The slightly larger error of the ED-GEKF (ΔE = 0.30 kcal mol^{-1}) compared to the E-GEKF (ΔE = 0.15 kcal mol^{-1}) observed directly at the equilibrium geometry of the reactant is due to the random nature of the initial HDNN weights and the partitioning of the reference data. Such small deviations from the overall trend that the ED-GEKF yields the highest quality fits are expected at individual points. The energies at the onset of the barrier are overestimated in the ED-GEKF, but the barrier height shows the smallest error compared to the reference energy (ΔE = 4.06 kcal mol^{-1} for TS). In the aldehyde region, energies are slightly overestimated, but near the product geometry, the HDNN energies closely resemble the reference values (ΔE = 0.10 kcal mol^{-1}). Unlike for the other filter variants, no additional minimum is present in the case of the ED-GEKF potential.

While the curve for the ED-GEKF is sufficiently accurate to demonstrate the interpolation capability of HDNNs trained with this algorithm, it still shows deviations from the electronic structure reaction profile, especially in the region of the barrier. The reason for this effect is that a single metadynamics trajectory was used in the generation of the reference data set. Due to this rather naïve sampling approach, no point in the reference data lies exactly on the reaction coordinate. Improved reference data sets could e.g. be generated with guided or self-consistent sampling procedures.[10,35,63]

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

A less sophisticated but yet extremely effective alternative is to include some critical points of the PESs into the training data. Here, three geometries corresponding to the equilibrium structures of allyl vinyl ether and 4-pentenal, as well as the TS structure of the rearrangement, were included in the reference data set to demonstrate the influence of additional training points on the interpolation performance. These geometries correspond to stationary points on the PES and are easily obtainable through routine electronic structure computations. HDNNs (C-40-40, H-40-40, O-40-40) were trained on this minimally expanded set for 100 epochs using the ED-GEKF. The reaction curve obtained for the HDNN with the smallest validation RMSE is depicted in Figure 8, along with the

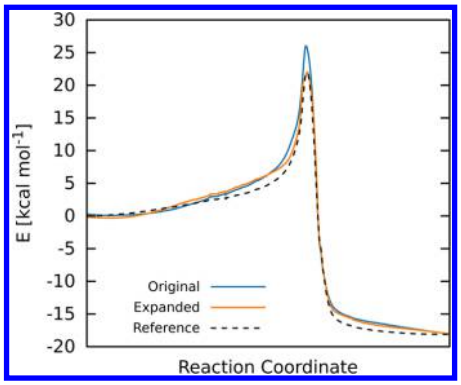


Figure 8. Comparison of the ED-GEKF HDNN reaction profiles obtained with the original reference data set and the data set expanded by the geometries of reactant, TS, and product.

reaction profile previously obtained for the ED-GEKF HDNN. Inclusion of only a few critical points greatly improves the accuracy of the HDNN reaction profile, reducing the RMSE to 0.91 kcal mol^{-1}. Especially the effect of incorporating the TS geometry is noticeable, as the barrier height is now reproduced accurately. The use of this extended reference data set also improves the performance of the other filter variants (not shown), although not to the same extent as in the ED-GEKF case.

Next, the distribution of the reference points in coordinate space is discussed. Although the 17100 points in the original reference data set are not situated directly on the reaction coordinate, the majority of them is nevertheless located in parts of the PES relevant for the transition. HDNN potentials trained on this data can therefore be expected to perform reasonably well in the description of the Claisen reaction, as demonstrated above.

In order to assess the performance of the HDNNs in regions of the PES not sampled during the initial metadynamics simulation, the bond dissociation of one of the hydrogen atoms bound to the C_S carbon atom of allyl vinyl ether was studied exemplarily. The C−H bond length was varied from 0.8 to 4.5 Å in 500 equidistant steps. The resulting potential energy profiles obtained for the electronic structure reference and the different HDNNs trained on the unexpanded reference data set are shown in Figure 9 for bond lengths between 0.9 and 1.5 Å.

Similar to before, the A-GEKF HDNN produces the worst fit, in this case only remotely reproducing the shape of the reference curve. The E-GEKF and ED-GEKF HDNNs perform well between bond lengths from 1.02 to 1.24 Å (up to energies of approximately 4 kcal mol^{-1}), with the ED-GEKF exhibiting a slightly smaller RMSE (0.36 kcal mol^{-1}) compared to the E-

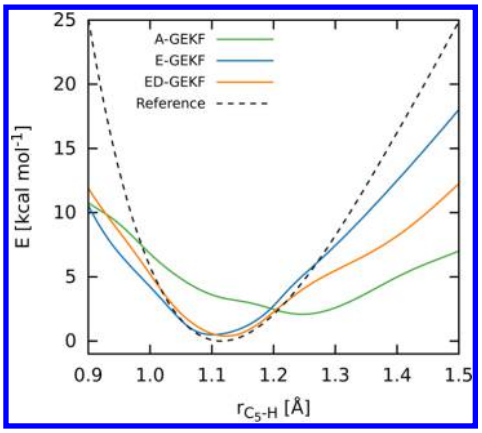


Figure 9. Potential energy profile for the dissociation of one of the hydrogen atoms bound to carbon atom C_S. The curves for the different HDNNs correspond to the HDNN potential energies predicted for the geometries encountered along the profile.

GEKF (0.53 kcal mol^{-1}) for this section of the PES. Such a behavior is expected, since this region of the reaction curve corresponds to the fluctuations of the C−H bond around the equilibrium bond length encountered during the metadynamics simulation and the E-GEKF and ED-GEKF HDNNs interpolate the reference data. Beyond this region, the HDNNs begin to extrapolate, and the quality of the HDNN-PESs deteriorates quickly, resulting in RMSEs close to 60 kcal mol^{-1} over the entire computed curve for all three HDNNs.

These observations are in accordance with the fact that NNs in general excel at interpolating data but perform poorly at extrapolation tasks. Because of this behavior, care should be taken to identify and avoid regions of the PES where extrapolation occurs. Approaches to address this problem exist, see e.g. ref 46. However, in those regions of the PES represented in the reference data set, an excellent fit can be obtained, provided a suitable training algorithm is chosen. Hence, while room for improvement still exists (e.g., addition of TS structure to the training set), the HDNN potential obtained with the ED-GEKF filter can be considered to give a reasonably accurate description of the degrees of freedom encountered in the Claisen reaction.

6. SUMMARY AND CONCLUSION

We report on a new training algorithm for high-dimensional neural networks (HDNNs) of the Behler−Parinello type and its application to the Claisen reaction of allyl vinyl ether. To the best of our knowledge, it is the first study of an organic reaction using HDNNs.

The training algorithm developed to generate the corresponding HDNNs has a substantially improved performance compared to other variants of the Kalman filter employed in HDNN training. In contrast to the latter, the new algorithm − termed element-decoupled global extended Kalman filter (ED-GEKF) − goes without the need for empirical weighting schemes or ad hoc assumptions. By employing the ED-GEKF during training, both root-mean-square errors and maximum fitting errors are reduced significantly. Moreover, fewer training periods and hence less time are required to obtain accurate HDNNs with the ED-GEKF in comparison to the other algorithms. The ED-GEKF can be extended to allow energies and forces to be fit simultaneously during training in a consistent manner, thus improving the description of the

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

HDNN forces tremendously. Especially applications which require highly accurate forces (e.g., molecular dynamics simulations) are expected to profit from this extension. The benefits of the ED-GEKF are expected to become even more pronounced for larger molecular systems or molecules involving many different elements. The ED-GEKF allows for the creation of HDNNs of unprecedented accuracy, putting the treatment of biological problems and complex organic reactions within reach.

In order to arrive at the aforementioned results, we used metadynamics and chain-of-states simulations to obtain the reaction path for the Claisen rearrangement of allyl vinyl ether to 4-pentenal. The potential energies were predicted by HDNNs trained on a set of reference points obtained with BP86/def2-SVP. Due to a naïve sampling scheme, none of the reference points lay on the reaction coordinate. We show that the accuracy of the predicted reaction profile can be drastically improved by including only a few critical points (reactant, transition state, and product geometries) in the reference data. This fact calls for the development of more efficient sampling procedures in the future in order to facilitate the automated generation of highly accurate potential energy surfaces for complicated systems and reactions with HDNNs.

AUTHOR INFORMATION

Corresponding Author

*Phone 43 1 4277 52764. Fax 43 1 4277 9527. E-mail: philipp.marquetand@univie.ac.at.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Leticia González, Christoph Flamm, and Jörg Behler for inspiring discussions. Allocation of computer time at the Vienna Scientific Cluster (VSC) is gratefully acknowledged.

REFERENCES

(1) Levine, I. N. *Quantum Chemistry*, 7th ed.; Prentice Hall: Boston, 2013.
(2) Friesner, R. A. *Ab initio* quantum chemistry: Methodology and applications. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6648−6653.
(3) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(4) Mackerell, A. D. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584−1604.
(5) Liang, T.; Shin, Y. K.; Cheng, Y.-T.; Yilmaz, D. E.; Vishnu, K. G.; Verners, O.; Zou, C.; Phillpot, S. R.; Sinnott, S. B.; van Duin, A. C. T. Reactive Potentials for Advanced Atomistic Simulations. *Annu. Rev. Mater. Res.* **2013**, *43*, 109−129.
(6) Handley, C. M.; Behler, J. Next generation interatomic potentials for condensed systems. *Eur. Phys. J. B* **2014**, *87*, 1−16.
(7) Brown, A.; Braams, B. J.; Christoffel, K.; Jin, Z.; Bowman, J. M. Classical and quasiclassical spectral analysis of CH_5^+ using an *ab initio* potential energy surface. *J. Chem. Phys.* **2003**, *119*, 8790−8793.
(8) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
(9) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(10) Ischtwan, J.; Collins, M. A. Molecular potential energy surfaces by interpolation. *J. Chem. Phys.* **1994**, *100*, 8080−8088.
(11) Wu, T.; Werner, H.-J.; Manthe, U. First-Principles Theory for the H + CH_4 → H_2 + CH_3 Reaction. *Science* **2004**, *306*, 2227−2229.
(12) Takata, T.; Taketsugu, T.; Hirao, K.; Gordon, M. S. Ab initio potential energy surface by modified Shepard interpolation: Application to the CH_3 + H_2 → CH_4 + H reaction. *J. Chem. Phys.* **1998**, *109*, 4281−4289.
(13) Crespos, C.; Collins, M. A.; Pijper, E.; Kroes, G. J. Multi-dimensional potential energy surface determination by modified Shepard interpolation for a molecule-surface reaction: H_2 + Pt(111). *Chem. Phys. Lett.* **2003**, *376*, 566−575.
(14) Dawes, R.; Thompson, D. L.; Guo, Y.; Wagner, A. F.; Minkoff, M. Interpolating moving least-squares methods for fitting potential energy surfaces: Computing high-density potential energy surface data from low-density *ab initio* data points. *J. Chem. Phys.* **2007**, *126*, 184108.
(15) Dawes, R.; Wang, X.-G.; Carrington, T., Jr. CO Dimer: New Potential Energy Surface and Rovibrational Calculations. *J. Phys. Chem. A* **2013**, *117*, 7612−7630.
(16) Vitek, A.; Stachon, M.; Kromer, P.; Snael, V. Towards the Modeling of Atomic and Molecular Clusters Energy by Support Vector Regression. International Conference on Intelligent Networking and Collaborative Systems (INCoS). 2013; pp 121−126.
(17) Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural network models of potential energy surfaces. *J. Chem. Phys.* **1995**, *103*, 4129−4137.
(18) Lorenz, S.; Scheffler, M.; Gross, A. Descriptions of surface chemical reactions using a neural network representation of the potential-energy surface. *Phys. Rev. B* **2006**, *73*, 115431.
(19) Ludwig, J.; Vlachos, D. G. *Ab initio* molecular dynamics of hydrogen dissociation on metal surfaces using neural networks and novelty sampling. *J. Chem. Phys.* **2007**, *127*, 154716.
(20) Manzhos, S.; Yamashita, K.; Carrington, T., Jr. Fitting sparse multidimensional data with low-dimensional terms. *Comput. Phys. Commun.* **2009**, *180*, 2002−2012.
(21) Carbogno, C.; Behler, J.; Groß, A.; Reuter, K. Fingerprints for Spin-Selection Rules in the Interaction Dynamics of O_2 at Al(111). *Phys. Rev. Lett.* **2008**, *101*, 096104.
(22) Behler, J.; Reuter, K.; Scheffler, M. Nonadiabatic effects in the dissociation of oxygen molecules at the Al(111) surface. *Phys. Rev. B* **2008**, *77*, 115421.
(23) Latino, D. A. R. S.; Fartaria, R. P. S.; Freitas, F. F. M.; Aires-de Sousa, J.; Silva Fernandes, F. M. S. Mapping Potential Energy Surfaces by Neural Networks: The ethanol/Au(111) interface. *J. Electroanal. Chem.* **2008**, *624*, 109−120.
(24) Latino, D. A. R. S.; Fartaria, R. P. S.; Freitas, F. F. M.; Aires-De-Sousa, J.; Silva Fer-nandes, F. M. S. Approach to potential energy surfaces by neural networks. A review of recent work. *Int. J. Quantum Chem.* **2010**, *110*, 432−445.
(25) Liu, T.; Fu, B.; Zhang, D. H. Six-dimensional potential energy surface of the dissociative chemisorption of HCl on Au(111) using neural networks. *Sci. China Chem.* **2013**, *57*, 147−155.
(26) Tafeit, E.; Estelberger, W.; Horejsi, R.; Moeller, R.; Oettl, K.; Vrecko, K.; Reibnegger, G. Neural networks as a tool for compact representation of *ab initio* molecular potential energy surfaces. *J. Mol. Graphics Modell.* **1996**, *14*, 12−18.
(27) Brown, D. F. R.; Gibbs, M. N.; Clary, D. C. Combining *ab initio* computations, neural networks, and diffusion Monte Carlo: An efficient method to treat weakly bound molecules. *J. Chem. Phys.* **1996**, *105*, 7597−7604.
(28) Houlding, S.; Liem, S. Y.; Popelier, P. L. A. A polarizable high-rank quantum topological electrostatic potential developed using neural networks: Molecular dynamics simulations on the hydrogen fluoride dimer. *Int. J. Quantum Chem.* **2007**, *107*, 2817−2827.
(29) No, K. T.; Chang, B. H.; Kim, S. Y.; Jhon, M. S.; Scheraga, H. A. Description of the potential energy surface of the water dimer with an artificial neural network. *Chem. Phys. Lett.* **1997**, *271*, 152−156.
(30) Cho, K.-H.; No, K. T.; Scheraga, H. A. A polarizable force field for water using an artificial neural network. *J. Mol. Struct.* **2002**, *641*, 77−91.

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

(31) Gassner, H.; Probst, M.; Lauenstein, A.; Hermansson, K. Representation of Intermolecular Potential Functions by Neural Networks. *J. Phys. Chem. A* **1998**, *102*, 4596−4605.
(32) Prudente, F. V.; Acioli, P. H.; Soares Neto, J. J. The fitting of potential energy surfaces using neural networks: Application to the study of vibrational levels of H_3^+. *J. Chem. Phys.* **1998**, *109*, 8801−8808.
(33) Rocha Filho, T. M.; Oliveira, Z. T.; Malbouisson, L. A. C.; Gargano, R.; Soares Neto, J. J. The use of neural networks for fitting potential energy surfaces: A comparative case study for the H_3^+ molecule. *Int. J. Quantum Chem.* **2003**, *95*, 281−288.
(34) Malshe, M.; Raff, L. M.; Rockley, M. G.; Hagan, M.; Agrawal, P. M.; Komanduri, R. Theoretical investigation of the dissociation dynamics of vibrationally excited vinyl bromide on an *ab initio* potential-energy surface obtained using modified novelty sampling and feedforward neural networks. II. Numerical application of the method. *J. Chem. Phys.* **2007**, *127*, 134105.
(35) Raff, L. M.; Malshe, M.; Hagan, M.; Doughan, D. I.; Rockley, M. G.; Komanduri, R. *Ab initio* potential-energy surfaces for complex, multichannel systems using modified novelty sampling and feedforward neural networks. *J. Chem. Phys.* **2005**, *122*, 084104.
(36) Agrawal, P. M.; Raff, L. M.; Hagan, M. T.; Komanduri, R. Molecular dynamics investigations of the dissociation of SiO_2 on an *ab initio* potential energy surface obtained using neural network methods. *J. Chem. Phys.* **2006**, *124*, 134306.
(37) Le, H. M.; Huynh, S.; Raff, L. M. Molecular dissociation of hydrogen peroxide (HOOH) on a neural network *ab initio* potential surface with a new configuration sampling method involving gradient fitting. *J. Chem. Phys.* **2009**, *131*, 014107.
(38) Manzhos, S.; Carrington, T., Jr. Using neural networks to represent potential surfaces as sums of products. *J. Chem. Phys.* **2006**, *125*, 194105.
(39) Le, H. M.; Raff, L. M. Cis → trans, trans → cis isomerizations and N-O bond dissociation of nitrous acid (HONO) on an *ab initio* potential surface obtained by novelty sampling and feed-forward neural network fitting. *J. Chem. Phys.* **2008**, *128*, 194310.
(40) Darley, M. G.; Handley, C. M.; Popelier, P. L. A. Beyond Point Charges: Dynamic Polarization from Neural Net Predicted Multipole Moments. *J. Chem. Theory Comput.* **2008**, *4*, 1435−1448.
(41) Le, H. M.; Dinh, T. S.; Le, H. V. Molecular Dynamics Investigations of Ozone on an Ab Initio Potential Energy Surface with the Utilization of Pattern-Recognition Neural Network for Accurate Determination of Product Formation. *J. Phys. Chem. A* **2011**, *115*, 10862−10870.
(42) Pukrittayakamee, A.; Malshe, M.; Hagan, M.; Raff, L. M.; Narulkar, R.; Bukkapatnum, S.; Komanduri, R. Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks. *J. Chem. Phys.* **2009**, *130*, 134101.
(43) Nguyen, H. T. T.; Le, H. M. Modified Feed-Forward Neural Network Structures and Combined-Function-Derivative Approximations Incorporating Exchange Symmetry for Potential Energy Surface Fitting. *J. Phys. Chem. A* **2012**, *116*, 4629−4638.
(44) Chen, J.; Xu, X.; Xu, X.; Zhang, D. H. Communication: An accurate global potential energy surface for the OH + CO → H + CO_2 reaction using neural networks. *J. Chem. Phys.* **2013**, *138*, 221104.
(45) Li, J.; Jiang, B.; Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. II. Four-atom systems. *J. Chem. Phys.* **2013**, *139*, 204103.
(46) Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930−17955.
(47) Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys.: Condens. Matter* **2014**, *26*, 183001.
(48) Handley, C. M.; Popelier, P. L. A. Potential Energy Surfaces Fitted by Artificial Neural Networks. *J. Phys. Chem. A* **2010**, *114*, 3371−3383.
(49) Witkoskie, J. B.; Doren, D. J. Neural Network Models of Potential Energy Surfaces:Prototypical Examples. *J. Chem. Theory Comput.* **2005**, *1*, 14−23.
(50) Behler, J.; Lorenz, S.; Reuter, K. Representing molecule-surface interactions with symmetry-adapted neural networks. *J. Chem. Phys.* **2007**, *127*, 014705.
(51) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
(52) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(53) Manzhos, S.; Carrington, T., Jr. A random-sampling high dimensional model representation neural network for building potential energy surfaces. *J. Chem. Phys.* **2006**, *125*, 084109.
(54) Artrith, N.; Hiller, B.; Behler, J. Neural network potentials for metals and oxides - First applications to copper clusters at zinc oxide. *Phys. Status Solidi B* **2013**, *250*, 1191−1203.
(55) Artrith, N.; Morawietz, T.; Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **2011**, *83*, 153101.
(56) Behler, J.; Martoňák, R.; Donadio, D.; Parrinello, M. Metadynamics Simulations of the High-Pressure Phases of Silicon Employing a High-Dimensional Neural Network Potential. *Phys. Rev. Lett.* **2008**, *100*, 185501.
(57) Behler, J.; Martoňák, R.; Donadio, D.; Parrinello, M. Pressure-induced phase transitions in silicon studied by neural network-based metadynamics simulations. *Phys. Status Solidi B* **2008**, *245*, 2618−2629.
(58) Eshet, H.; Khaliullin, R. Z.; Kühne, T. D.; Behler, J.; Parrinello, M. *Ab initio* quality neural-network potential for sodium. *Phys. Rev. B* **2010**, *81*, 184107.
(59) Khaliullin, R. Z.; Eshet, H.; Kühne, T. D.; Behler, J.; Parrinello, M. Graphite-diamond phase coexistence study employing a neural-network mapping of the *ab initio* potential energy surface. *Phys. Rev. B* **2010**, *81*, 100103.
(60) Seema, P.; Behler, J.; Marx, D. Adsorption of Methanethiolate and Atomic Sulfur at the Cu(111) Surface: A Computational Study. *J. Phys. Chem. C* **2013**, *117*, 337−348.
(61) Morawietz, T.; Behler, J. A Density-Functional Theory-Based Neural Network Potential for Water Clusters Including van der Waals Corrections. *J. Phys. Chem. A* **2013**, *117*, 7356−7366.
(62) Morawietz, T.; Behler, J. A Full-Dimensional Neural Network Potential-Energy Surface for Water Clusters up to the Hexamer. *Z. Phys. Chem.* **2013**, *227*, 1559−1581.
(63) Morawietz, T.; Sharma, V.; Behler, J. A neural network potential-energy surface for the water dimer based on environment-dependent atomic energies and charges. *J. Chem. Phys.* **2012**, *136*, 064103.
(64) Puskorius, G. V.; Feldkamp, L. A. Decoupled extended Kalman filter training of feedforward layered networks. IJCNN-91-Seattle International Joint Conference on Neural Networks. 1991; pp 771−777.
(65) Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Syst.* **1989**, *2*, 303−314.
(66) Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **1989**, *2*, 359−366.
(67) Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **1991**, *4*, 251−257.
(68) Bishop, C. M. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, 2006.
(69) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.
(70) Hagan, M.; Menhaj, M. Training feedforward networks with the Marquardt algorithm. *IEEE Trans. Neural Networks* **1994**, *5*, 989−993.
(71) Blank, T. B.; Brown, S. D. Adaptive, global, extended Kalman filters for training feedforward neural networks. *J. Chemom.* **2005**, *8*, 391−407.

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198

(72) Shah, S.; Palmieri, F.; Datum, M. Optimal filtering algorithms for fast learning in feedforward neural networks. *Neural Networks* **1992**, *5*, 779−787.
(73) Murtuza, S.; Chorian, S. F. Node decoupled extended Kalman filter based learning algorithm for neural networks. Proceedings of the 1994 IEEE International Symposium on Intelligent Control. 1994; pp 364−369.
(74) Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73−78.
(75) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098−3100.
(76) Dirac, P. A. M. Quantum Mechanics of Many-Electron Systems. *Proc. R. Soc., Ser. A* **1929**, *123*, 714−733.
(77) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822−8824.
(78) Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81*, 385−390.
(79) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.
(80) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.
(81) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. Auxiliary basis sets to approximate Coulomb potentials. *Chem. Phys. Lett.* **1995**, *240*, 283−290.
(82) Vahtras, O.; Almlöf, J.; Feyereisen, M. W. Integral approximations for LCAO-SCF calculations. *Chem. Phys. Lett.* **1993**, *213*, 514−518.
(83) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(84) Johnson, E. R.; Becke, A. D. A post-Hartree-Fock model of intermolecular interactions. *J. Chem. Phys.* **2005**, *123*, 024101−024101−7.
(85) Plessow, P. Reaction Path Optimization without NEB Springs or Interpolation Algorithms. *J. Chem. Theory Comput.* **2013**, *9*, 1305−1310.
(86) Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. Turbomole. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 91−100.
(87) *TURBOMOLE* V6.6 2014, a development of the University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989−2007, TURBOMOLE GmbH: since 2007. Available from http://www.turbomole.com (accessed 03.04.2015).
(88) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*, Reprint ed.; Cambridge University Press: Cambridge, 2012.
(89) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(90) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, *76*, 637−649.
(91) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(92) Tolman, R. C. *The Principles of Statistical Mechanics*, New ed.; Dover Publications: New York, 2010.
(93) Nguyen, D. H.; Widrow, B. Neural networks for self-learning control systems. *IEEE Control Systems Magazine* **1990**, *10*, 18−23.
(94) Plaut, D. C.; Nowlan, S. J.; Hinton, G. E. *Experiments on Learning by Back Propagation*; Technical Report; 1986.
(95) van Rossum, G.; Drale, F. L. *Python Reference Manual*; PythonLabs: Virginia, USA, 2001. http://www.python.org (accessed 03.04.2015).
(96) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22−30.
(97) Ziegler, F. E. The thermal, aliphatic Claisen rearrangement. *Chem. Rev.* **1988**, *88*, 1423−1452.

DOI: 10.1021/acs.jctc.5b00211
J. Chem. Theory Comput. 2015, 11, 2187−2198