Minimum mode saddle point searches using Gaussian process regression with inverse-distance covariance function

Olli-Pekka Koistinen [†,‡] Vilhjálmur Ásgeirsson [‡] Aki Vehtari [†]
Hannes Jónsson [‡,§]

[†] Department of Computer Science, Aalto University, Espoo, Finland

[‡] Science Institute and Faculty of Physical Sciences, University of Iceland, Reykjavík, Iceland

[§] Department of Applied Physics, Aalto University, Espoo, Finland

hj@hi.is

Abstract

The minimum mode following method can be used to find saddle points on an energy surface by following a direction guided by the lowest curvature mode. Such calculations are often started close to a minimum on the energy surface to find out which transitions can occur from an initial state of the system, but it is also common to start from the vicinity of a first order saddle point making use of an initial guess based on intuition or more approximate calculations. In systems where accurate evaluations of the energy and its gradient are computationally intensive, it is important to exploit the information of the previous evaluations to enhance the performance. Here, we show that the number of evaluations required for convergence to the saddle point can be significantly reduced by making use of an approximate energy surface obtained by a Gaussian process model based on inverse inter-atomic distances, evaluating accurate energy and gradient at the saddle point of the approximate surface and then correcting the model based on the new information. The performance of the method is tested with start points chosen randomly in the vicinity of saddle points for dissociative adsorption of an H_2 molecule on the Cu(110) surface and three gas phase chemical reactions.

1 Introduction

In systems characterized by a smooth potential energy surface, the transition state between the initial and final state of an atomic rearrangement event is, within the harmonic approximation to transition state theory, placed using information about a first-order saddle point on the energy surface, i.e., a location with zero gradient and exactly one negative eigenvalue of the Hessian matrix. Finding first order saddle points is then the essential task when identifying the mechanisms and estimating the rates of transitions. The transition state is taken to be a hyperplane going through the saddle point with normal parallel to the eigenvector corresponding to the negative eigenvalue.

In chain-of-states methods, such as the nudged elastic band method,[1,2] saddle points are found by calculating a minimum energy path between the initial and final state and identifying the energy maximum along the path. There, both the initial and final state of the transition are specified. In another type of algorithms, only the initial state is specified and the saddle point found by climbing up the energy surface without specifying the final state of the transition. Such calculations are often started from the vicinity of the initial state minimum to find out which transitions can occur, but it is also common to start from somewhere close to the saddle point with an initial guess obtained from intuition or from approximate minimum energy path calculations.[3,4] Early algorithms of this sort required the evaluation of the full Hessian, the matrix of the second derivatives of the energy with respect to the coordinates, and calculation of all the eigenvalues and eigenvectors (see ref 5 for a review). In a more efficient formulation, the minimum mode following method, only the eigenvector corresponding to the lowest eigenvalue is found and used to guide the search for the saddle point(s) without a need to evaluate the Hessian matrix.[6–8]

In this article, we choose to find the minimum mode using the dimer method.[6,9–11] A dimer is here a pair of points in a configuration space, separated by a small fixed distance. The dimer is first rotated around its midpoint to find the orientation that gives the lowest total energy of the two configurations. This gives the direction of the lowest curvature mode of the Hessian, the minimum mode.[12] The dimer is then translated towards the saddle point by reversing the force (negative energy gradient) component in this direction. The movements are based only on the energy and the gradient of the energy and thus do not require calculation of the Hessian matrix.

In systems where accurate evaluations of energy and its gradient are computationally expensive, it is important to exploit the information in previous evaluations to enhance the performance. Here, we show that Gaussian process (GP) regression[13–16] can be used to significantly reduce the number of evaluations required for convergence to saddle points. The basic scheme is similar to the one used for nudged elastic band calculations in the GP-NEB method:[17–19] a regular minimum mode following calculation is performed to find a saddle point on an approximate surface obtained by a Gaussian process model, accurate energy and gradient are then evaluated at that point, and the model is subsequently refined based on the new evaluations. If no information about the energy surface is available in the beginning, it is useful to perform initial rotations with accurate evaluations before starting to translate the dimer. We show that GP regression can be used also to reduce the number of evaluations required for finding the lowest curvature mode in this initial rotation phase.

A similar general scheme for saddle point searches starting from a configuration close to a saddle point has been presented by Denzel and Kästner who use a stationary Matérn covariance function to build the GP model.[20] Here, we use a more expressive covariance function based on inverted inter-atomic distances, coupled with a robust stopping criterion, as suggested in ref 19 and compare the performance to stationary covariance functions. The inverse-distance covariance function makes the method more robust especially when the calculation is started far from the saddle point where the atomic forces are large.

The performance of the GP-dimer method is tested with start points up to 3 Å away from saddle points for dissociative adsorption of an H_2 molecule on the Cu(110) surface and three gas phase chemical reactions. With the largest start distances, the number of energy and gradient evaluations is found to be reduced by an order of magnitude compared to the regular dimer method.

2 Dimer method

In this section, we review the principles of the dimer method for finding the minimum mode[6,9–11] and present details for two variants of the algorithm, here referred to as CG-dimer[10] and LBFGS-dimer.[11] They are used as references for comparing the performance with our GP-dimer method. The LBFGS-dimer algorithm is used also as a part of the GP-dimer method as described in the following section.

A dimer is defined as a pair of points in a configuration space, referred to as image 1, $\mathbf{R}_1$, and image 2, $\mathbf{R}_2$. The small distance between $\mathbf{R}_1$ and $\mathbf{R}_2$ is kept constant, and half of this distance is referred to as the dimer separation, $\Delta_{\mathbf{R}}$ (here 10^{-2} Å as recommended in ref 9). The middle point of the dimer is denoted by $\mathbf{R}_0$, and the orientation vector $\hat{\mathbf{N}}$ is a unit vector that points from $\mathbf{R}_0$ towards $\mathbf{R}_1$. The dimer energy is defined as the sum $E_1 + E_2$, where E_1 and E_2 denote the energy of the system at $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively. The direction of lowest curvature of energy at $\mathbf{R}_0$ corresponds to the orientation of minimum dimer energy, which is obtained by rotating the dimer around $\mathbf{R}_0$ so that the rotational force is zeroed. Denoting the force (negative energy gradient) acting on $\mathbf{R}_i$ by $\mathbf{F}_i$ and the component of $\mathbf{F}_i$ perpendicular to the dimer by $\mathbf{F}_i^{\perp} = \mathbf{F}_i - (\mathbf{F}_i \cdot \hat{\mathbf{N}})\hat{\mathbf{N}}$, the scaled rotational force acting on $\mathbf{R}_1$ is defined by

$$\mathbf{F}_{\mathrm{rot}} = (\mathbf{F}_1^{\perp} - \mathbf{F}_2^{\perp})/\Delta_{\mathbf{R}}. \tag{1}$$

As suggested by Olsen et al.,[9] it is more efficient to evaluate the force at the middle point $\mathbf{R}_0$ instead of $\mathbf{R}_2$ and extrapolate the force at $\mathbf{R}_2$ as $\mathbf{F}_2 = 2\mathbf{F}_0 - \mathbf{F}_1$.

Each rotation iteration is performed within a plane spanned by unit vectors $\hat{\mathbf{N}}$ and $\hat{\mathbf{\Omega}}$, where the steepest descent direction of rotation for image 1 is $\hat{\mathbf{\Omega}} = \mathbf{F}_{\mathrm{rot}}/||\mathbf{F}_{\mathrm{rot}}||$. In CG-dimer and LBFGS-dimer, $\hat{\mathbf{\Omega}}$ is modified based on previous rotation iterations according to nonlinear conjugate gradient[21,22] or limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)[23,24] algorithms, respectively. According to the approach of Heyden et al.,[10] a rough estimate for the optimal rotational angle is first calculated based on $\mathbf{F}_0$ and $\mathbf{F}_1$ as

$$\omega^* = \frac{1}{2}\arctan\frac{(\mathbf{F}_1 - \mathbf{F}_0)\cdot\hat{\mathbf{\Omega}}}{\Delta_{\mathbf{R}}\,|C|}, \tag{2}$$

where $C = (\mathbf{F}_0 - \mathbf{F}_1)\cdot\hat{\mathbf{N}}/\Delta_{\mathbf{R}}$ is the curvature of the energy along the dimer. After the preliminary rotation of ω^*, the orientation vector of the dimer is given by

$$\hat{\mathbf{N}}^* = \hat{\mathbf{N}}\cos\omega^* + \hat{\mathbf{\Omega}}\sin\omega^* \tag{3}$$

and the rotation direction by

$$\hat{\mathbf{\Omega}}^* = -\hat{\mathbf{N}}\sin\omega^* + \hat{\mathbf{\Omega}}\cos\omega^*. \tag{4}$$

After evaluating the force $\mathbf{F}_1^*$ at $\mathbf{R}_1^* = \mathbf{R}_0 + \Delta_{\mathbf{R}}\hat{\mathbf{N}}^*$, the optimal rotational angle based on a local quadratic approximation to the energy surface is given by

$$\omega = \begin{cases} \dfrac{1}{2}\arctan\dfrac{b_1}{a_1}, & \text{if } \dfrac{b_1}{a_1} \geq 0 \\ \dfrac{1}{2}\arctan\dfrac{b_1}{a_1} + \dfrac{\pi}{2}, & \text{if } \dfrac{b_1}{a_1} < 0, \end{cases} \tag{5}$$

where

$$b_1 = (\mathbf{F}_0 - \mathbf{F}_1)\cdot\hat{\mathbf{\Omega}}/\Delta_{\mathbf{R}} \tag{6}$$

and

$$a_1 = \frac{b_1\cos(2\omega^*) - (\mathbf{F}_0 - \mathbf{F}_1^*)\cdot\hat{\mathbf{\Omega}}^*/\Delta_{\mathbf{R}}}{\sin(2\omega^*)}. \tag{7}$$

The orientation vector of the dimer after the rotation is then given by

$$\hat{\mathbf{N}}^{\text{new}} = \hat{\mathbf{N}} \cos\omega + \hat{\mathbf{\Omega}} \sin\omega \tag{8}$$

and the new location of image 1 by

$$\mathbf{R}_1^{\text{new}} = \mathbf{R}_0 + \Delta_{\mathbf{R}} \hat{\mathbf{N}}^{\text{new}}. \tag{9}$$

The rotation direction in the end of the rotation, needed for the following rotation iteration in CG-dimer, is given by

$$\hat{\mathbf{\Omega}}_{\text{end}} = -\hat{\mathbf{N}} \sin\omega + \hat{\mathbf{\Omega}} \cos\omega. \tag{10}$$

Here, the rotation iterations are stopped if the preliminary rotational angle ω^* is estimated to be below five degrees,[11] if the actual rotational angle ω is below this threshold, or if a prescribed maximum number of consecutive rotation iterations is reached. In the two latter cases, the curvature of energy along the new orientation vector $\hat{\mathbf{N}}^{\text{new}}$ can be estimated as

$$C^{\text{new}} \approx C + a_1(\cos(2\omega) - 1) + b_1 \sin(2\omega). \tag{11}$$

After the rotation phase, the middle point of the dimer is translated in order to advance towards the saddle point. The translational force is obtained by inverting the component of $\mathbf{F}_0$ parallel to the dimer:

$$\mathbf{F}_{\text{trans}} = \mathbf{F}_0 - 2\mathbf{F}_0^{\parallel}, \tag{12}$$

where $\mathbf{F}_0^{\parallel} = (\mathbf{F}_0 \cdot \hat{\mathbf{N}})\hat{\mathbf{N}}$. This allows the dimer to climb upwards on the energy surface in the direction of the minimum mode while moving towards lower energy in directions perpendicular to the minimum mode. In CG-dimer and LBFGS-dimer, also the translations are modified according to conjugate gradient and L-BFGS algorithms, respectively. If the curvature along the dimer is positive, the dimer is assumed to be in a convex region where all eigenvalues of the Hessian matrix are positive. In this case, a step of a predefined length is taken in the opposite direction of $\mathbf{F}_0^{\parallel}$ to make the dimer climb up from the energy basin as quickly as possible. Here, this step length is set to 0.1 Å, which is also the maximum step length for the translation iterations.[9] The calculation is considered to have converged when the maximum component of force $\mathbf{F}_0$ at the middle point of the dimer is below a threshold T_0, which is here set to $T_0 = 0.01$ eV/Å.

2.1 CG-dimer

The first of the two reference algorithms, referred to here as CG-dimer, follows mainly the details presented by Heyden et al.[10] In this algorithm, only one rotation iteration, if any, is performed between translations. Thus, each rotation phase includes two or three energy and force evaluations. Since the initial orientation of the dimer is chosen randomly in our test cases, a larger maximum number of rotations, equal to the number of degrees of freedom in the system, is used in the first rotation phase to stabilize the algorithm.

In CG-dimer, we apply separate nonlinear conjugate gradient algorithms[21,22] to choose the rotational plane and translational search direction, as suggested previously.[6] Given a rotation direction $\hat{\mathbf{\Omega}}$, the rotation proceeds as presented above. For translations, a preliminary step is first taken and the middle point of the dimer moved to

$$\mathbf{R}_0^* = \mathbf{R}_0 + \frac{\hat{\mathbf{\Gamma}} \cdot \mathbf{F}_{\text{trans}}}{2\,|C|} \hat{\mathbf{\Gamma}}, \tag{13}$$

where $\hat{\mathbf{\Gamma}}$ is a unit vector parallel to the search direction.[10] After evaluating the force $\mathbf{F}_0^*$ at $\mathbf{R}_0^*$, the middle point of the dimer is then moved to the estimated zero point of the translational force component parallel to the search direction:

$$\mathbf{R}_0^{\text{new}} = \mathbf{R}_0 - \frac{\hat{\mathbf{\Gamma}} \cdot \mathbf{F}_{\text{trans}}}{\hat{\mathbf{\Gamma}} \cdot (\mathbf{F}_{\text{trans}}^* - \mathbf{F}_{\text{trans}})}(\mathbf{R}_0^* - \mathbf{R}_0), \tag{14}$$

where $\mathbf{F}_{\text{trans}}^* = \mathbf{F}_0^* - 2(\mathbf{F}_0^* \cdot \hat{\mathbf{N}})\hat{\mathbf{N}}$.

In the conjugate gradient algorithm for the translations, the search direction $\hat{\mathbf{\Gamma}}$ is parallel to a conjugated force vector $\mathbf{\Gamma}$, which is a linear combination of the current and previous translational force vectors. $\mathbf{\Gamma}$ can be expressed recursively as

$$\mathbf{\Gamma} = \mathbf{F}_{\text{trans}} + \beta \mathbf{\Gamma}^{\text{old}}, \tag{15}$$

where $\mathbf{\Gamma}^{\text{old}}$ is the conjugated force vector in the previous translation iteration and the coefficient β_{trans} is here given by the Polak-Ribière formula:[22]

$$\beta_{\text{trans}} = \max\left\{0, \frac{(\mathbf{F}_{\text{trans}} - \mathbf{F}_{\text{trans}}^{\text{old}}) \cdot \mathbf{F}_{\text{trans}}}{\mathbf{F}_{\text{trans}}^{\text{old}} \cdot \mathbf{F}_{\text{trans}}^{\text{old}}}\right\}, \tag{16}$$

where $\mathbf{F}_{\text{trans}}^{\text{old}}$ is the previous translational force vector. In the first iteration, $\mathbf{\Gamma}$ is set equal to $\mathbf{F}_{\text{trans}}$. As noted previously,[6] increasing translational force in the search direction would lead to a step backwards against the search direction, which indicates that the dimer may still be in a convex area in spite of negative estimated curvature C in the direction of the dimer. In this case, a step of a predefined length (here 0.1 Å) is taken in the search direction. Without further restrictions, however, a negative step against the search direction may occur also if the search direction itself is opposite to the current translational force vector. This would as well trigger the predefined step and might lead to a trap where the dimer bounces between two locations. To prevent this kind of situations, we set β_{trans} to zero when the correction vector $\beta_{\text{trans}}\mathbf{\Gamma}^{\text{old}}$ in eq 15 becomes longer than the current translational force vector $\mathbf{F}_{\text{trans}}$. In addition, we reset the memory of conjugate directions when the number of conjugated iterations reaches the number of degrees of freedom in the system or if a predefined step length is used due to positive C, negative step against the search direction or excessive step length.

Analogously to the conjugate gradient algorithm described above for the translations, the rotation direction $\hat{\mathbf{\Omega}}$ is parallel to a conjugated force vector $\mathbf{\Omega}$ defined recursively based on the current rotational force vector $\mathbf{F}_{\text{rot}}$, the previous rotational force vector $\mathbf{F}_{\text{rot}}^{\text{old}}$ and the previous conjugated force vector $\mathbf{\Omega}^{\text{old}}$. The only difference is that $\mathbf{\Omega}^{\text{old}}$ needs to be rotated on the previous rotational plane to be aligned with $\hat{\mathbf{\Omega}}_{\text{end}}^{\text{old}}$, which is the rotation direction in the end of the previous iteration (eq 10).[6] Thus, the recursive expression for $\mathbf{\Omega}$ is given by

$$\mathbf{\Omega} = \mathbf{F}_{\text{rot}} + \beta_{\text{rot}} ||\mathbf{\Omega}^{\text{old}}|| \hat{\mathbf{\Omega}}_{\text{end}}^{\text{old}}, \tag{17}$$

where

$$\beta_{\text{rot}} = \max\left\{0, \frac{(\mathbf{F}_{\text{rot}} - \mathbf{F}_{\text{rot}}^{\text{old}}) \cdot \mathbf{F}_{\text{rot}}}{\mathbf{F}_{\text{rot}}^{\text{old}} \cdot \mathbf{F}_{\text{rot}}^{\text{old}}}\right\}. \tag{18}$$

The coefficient β_{rot} is set to zero when the correction vector $\beta_{\text{rot}}||\mathbf{\Omega}^{\text{old}}||\hat{\mathbf{\Omega}}_{\text{end}}^{\text{old}}$ in eq 17 becomes longer than the current rotational force vector $\mathbf{F}_{\text{rot}}$, and the memory of rotational conjugate directions is reset when the number of conjugated rotation iterations reaches the number of degrees of freedom in the system or if rotational convergence is reached.

2.2 LBFGS-dimer

The second reference algorithm, referred to here as LBFGS-dimer, follows mainly the details presented by Kästner and Sherwood.[11] In this algorithm, the rotations are continued until convergence unless the maximum of ten consecutive rotation iterations is reached. If the number of degrees of freedom is less than ten, we use this number as the maximum. To reduce the number of evaluations between the consecutive rotation iterations to one, the force $\mathbf{F}_1^{\text{new}}$ at the new location of image 1 is estimated as

$$\mathbf{F}_1^{\text{new}} \approx \frac{\sin(\omega^* - \omega)}{\sin\omega^*}\mathbf{F}_1 + \frac{\sin\omega}{\sin\omega^*}\mathbf{F}_1^* + \left(1 - \cos\omega - \sin\omega\tan\frac{\omega^*}{2}\right)\mathbf{F}_0. \tag{19}$$

In LBFGS-dimer, the rotational plane and translational search direction are chosen using separate limited-memory BFGS algorithms.[23,24] Given a rotation direction $\hat{\mathbf{\Omega}}$, the rotation proceeds similarly as in CG-dimer. For translations, the L-BFGS algorithm gives also a step length in addition to the search direction, and thus no preliminary step is needed.

The L-BFGS algorithm approximates an inverse Hessian matrix implicitly based on information stored from previous iterations. The memory of L-BFGS includes displacement vectors $\boldsymbol{\delta}_{\mathbf{x}}^i, i = 1, 2, \ldots, M$, between the locations and $\boldsymbol{\delta}_{\mathbf{F}}^i, i = 1, 2, \ldots, M$, between the effective forces in the i^{th} and $(i-1)^{\text{th}}$ last iteration counting backwards from the current iteration. The size of the memory, M, is here limited to the number of degrees of freedom in the system, and the inverse Hessian is initialized to an identity matrix scaled by $\lambda = (\boldsymbol{\delta}_{\mathbf{F}}^1 \cdot \boldsymbol{\delta}_{\mathbf{x}}^1)/||\boldsymbol{\delta}_{\mathbf{F}}^1||^2$.[24] If the memory is empty, the scaling factor is set to $\lambda = 0.01$ Å^2/eV. The optimal displacement vector $\boldsymbol{\delta}_{\mathbf{x}}$ for the current iteration is obtained by the following recursive procedure,[23] where $\boldsymbol{\Psi}$ is initialized to the effective force vector:

For $i = 1, 2, \ldots, M$:

Set $\alpha^i \leftarrow \dfrac{\boldsymbol{\Psi} \cdot \boldsymbol{\delta}_{\mathbf{x}}^i}{\boldsymbol{\delta}_{\mathbf{F}}^i \cdot \boldsymbol{\delta}_{\mathbf{x}}^i}$.

Set $\boldsymbol{\Psi} \leftarrow \boldsymbol{\Psi} - \alpha^i \boldsymbol{\delta}_{\mathbf{F}}^i$.

Set $\boldsymbol{\delta}_{\mathbf{x}} \leftarrow \lambda\boldsymbol{\Psi}$.

For $i = M, M-1, \ldots, 1$:

Set $\boldsymbol{\delta}_{\mathbf{x}} \leftarrow \boldsymbol{\delta}_{\mathbf{x}} + \left(\alpha^i - \dfrac{\boldsymbol{\delta}_{\mathbf{F}}^i \cdot \boldsymbol{\delta}_{\mathbf{x}}}{\boldsymbol{\delta}_{\mathbf{F}}^i \cdot \boldsymbol{\delta}_{\mathbf{x}}^i}\right)\boldsymbol{\delta}_{\mathbf{x}}^i$.

In the L-BFGS algorithm for the translations, $\boldsymbol{\delta}_{\mathbf{x}}^i$ are displacements of the middle point $\mathbf{R}_0$ and the effective force is the translational force $\mathbf{F}_{\text{trans}}$. The new location of $\mathbf{R}_0$ is simply given by $\mathbf{R}_0^{\text{new}} + \boldsymbol{\delta}_{\mathbf{x}}$. The memory of L-BFGS is reset, if a predefined step length is used due to positive C or excessive step length.

For the rotations, $\boldsymbol{\delta}_{\mathbf{x}}^i$ are given by changes of the orientation vector $\hat{\mathbf{N}}$ during previous rotation iterations and the effective force is the rotational force $\mathbf{F}_{\text{rot}}$. After estimating $\boldsymbol{\delta}_{\mathbf{x}}$ according to the recursive procedure described above, the rotation direction $\hat{\mathbf{\Omega}}$ is given by a unit vector parallel to

$$\boldsymbol{\delta}_{\mathbf{x}}^{\perp} = \boldsymbol{\delta}_{\mathbf{x}} - (\boldsymbol{\delta}_{\mathbf{x}} \cdot \hat{\mathbf{N}})\hat{\mathbf{N}}, \tag{20}$$

which is the component of $\boldsymbol{\delta}_{\mathbf{x}}$ perpendicular to the dimer. The memory of the L-BFGS algorithm for rotations is reset after each translation step.

3 GP-dimer method

In this section, the GP-dimer method is described. There, each iteration involves the energy surface being modeled using Gaussian process regression, dimer calculations performed on the approximate surface, and the GP model then refined after evaluating the accurate energy and force at the saddle point determined on the approximate surface. The method can be seen as a surface walking version of the GP-NEB method.[17–19] The dimer calculations on the approximated surface are performed using LBFGS-dimer[11] with some modifications.

3.1 Gaussian process regression

A Gaussian process[13–16] model defines the joint probability distribution of the function values $\mathbf{f} = [f(\mathbf{x}^{(1)}), f(\mathbf{x}^{(2)}), \ldots, f(\mathbf{x}^{(N)})]^\mathsf{T}$ at any finite set of input locations $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}]^\mathsf{T}$ as a multivariate Gaussian $p(\mathbf{f}) = \mathcal{N}(\mathbf{m}, K(\mathbf{X}, \mathbf{X}))$, where $\mathbf{m} = [m(\mathbf{x}^{(1)}), m(\mathbf{x}^{(2)}), \ldots, m(\mathbf{x}^{(N)})]^\mathsf{T}$ is defined by mean function $m(\mathbf{x})$ and the notation $K(\mathbf{X}, \mathbf{X}')$ stands for a covariance matrix with elements $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}'^{(j)})$ defined by covariance function $k(\mathbf{x}, \mathbf{x}')$. In the applications of the GP-dimer method presented here, the energy surface is modeled as a function of a $3N_\mathrm{m}$-dimensional coordinate vector

$$\mathbf{x} = [x_{1,1}, x_{1,2}, x_{1,3}, x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{N_\mathrm{m},1}, x_{N_\mathrm{m},2}, x_{N_\mathrm{m},3}]^\mathsf{T}$$

including the coordinates for moving atoms $1, 2, \ldots, N_\mathrm{m} \in A_\mathrm{m}$. The system may also involve a set atoms with fixed coordinates, denoted by A_f, but here those atoms are taken into account in the GP model only if some of the moving atoms has been within the radius of 5 Å from the frozen atom during the GP-dimer algorithm. As suggested in ref 19, the prior probability model of the energy surface is defined here as a GP with mean function $m(\mathbf{x}) = 0$ and covariance function

$$k_{1/r}(\mathbf{x}, \mathbf{x}') = \sigma_\mathrm{c}^2 + \sigma_\mathrm{m}^2 \exp\left(-\frac{1}{2} \sum_{i \in A_\mathrm{m}} \sum_{\substack{j \in A_\mathrm{m}, j>i \\ \vee \\ j \in A_\mathrm{f}}} \frac{\left(\frac{1}{r_{i,j}(\mathbf{x})} - \frac{1}{r_{i,j}(\mathbf{x}')} \right)^2}{l^2_{\phi(i,j)}} \right), \tag{21}$$

where

$$r_{i,j}(\mathbf{x}) = \sqrt{\sum_{d=1}^{3} (x_{i,d} - x_{j,d})^2}$$

is the distance between atoms i and j, $\phi(i,j)$ is the atom pair type for pair (i,j), and $l_{\phi(i,j)}$ is the length scale for that pair type. Weakly informative prior distributions $p(\sigma_\mathrm{m}) = \mathcal{N}(0, \max\{1\ \mathrm{eV}^2, (\Delta_\mathbf{y}/3)^2\})$ and $p(l_\phi) = \mathcal{N}(0, \max\{1\ \text{Å}^{-2}, (\Delta_\mathbf{X}/3)^2\})$ are set for the magnitude σ_m and length scales $l_\phi, \phi = 1, 2, \ldots, N_\phi$, with $\Delta_\mathbf{y}$ representing the range of energy values in the training data set and $\Delta_\mathbf{X}$ representing the maximum difference between the data points based on difference measure

$$\mathcal{D}_{1/r}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i \in A_\mathrm{m}} \sum_{\substack{j \in A_\mathrm{m}, j>i \\ \vee \\ j \in A_\mathrm{f}}} \left(\frac{1}{r_{i,j}(\mathbf{x})} - \frac{1}{r_{i,j}(\mathbf{x}')} \right)^2}. \tag{22}$$

The constant term σ_c^2 is set to the square of the mean of the observed energy values but no lower than 1 eV^2. The only differences to the model suggested for the GP-NEB method in ref 19 are the lower limits for the constant term and for the variances of the prior distributions of the hyperparameters. Since the initial training data set in the GP-dimer method is focused around one start point, small $\Delta_\mathbf{X}$ and $\Delta_\mathbf{y}$ would lead to unnecessarily restrictive priors for the

magnitude σ_m and length scales l_ϕ in the beginning if no lower limits for the variances were used.

For comparisons, alternative GP models with stationary covariance functions are also implemented for the GP-dimer method. Following the notation of ref 19, we define the squared exponential covariance function as

$$k_x(\mathbf{x},\mathbf{x}') = \sigma_\mathrm{c}^2 + \sigma_\mathrm{m}^2 \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2l^2}\right) \tag{23}$$

and the Matérn-5/2 covariance function as

$$k_x^{\mathrm{M}-5/2}(\mathbf{x},\mathbf{x}') = \sigma_\mathrm{c}^2 + \sigma_\mathrm{m}^2\left(1+\frac{\sqrt{5}||\mathbf{x}-\mathbf{x}'||}{l}+\frac{5||\mathbf{x}-\mathbf{x}'||^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}||\mathbf{x}-\mathbf{x}'||}{l}\right). \tag{24}$$

Since the dimer method relies on the curvature properties of the energy surface, Matérn covariance functions with a lower smoothness parameter ($\nu < 2$) are not good choices for this application. The priors of the hyperparameters are defined similarly as for $k_{1/r}$, but the prior variance of the length scale is based on the regular distance in the $3N_\mathrm{m}$-dimensional coordinate space.

The evaluations of energy and force (negative energy gradient) are regarded as accurate up to floating point presentation accuracy, and thus Gaussian noise with a small variance is assumed to be included in both energy (noise variance $\sigma^2 = 10^{-8}$ eV^2) and force evaluations (noise variance $\sigma_\mathrm{d}^2 = 10^{-8}$ eV^2/Å^2) to avoid numerical problems. Given a training data set $\{\mathbf{X},\mathbf{y}\}$, where $\mathbf{y} = [y^{(1)}, y^{(2)}, \ldots, y^{(N)}]^\mathsf{T}$ includes evaluated energy values from N locations $\mathbf{X}$, and a noise covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{I}_N$ with $\mathbf{I}_N$ denoting an identity matrix, the hyperparameters $\boldsymbol{\theta} = \{\sigma_\mathrm{m}, l_1, l_2, \ldots, l_{N_\phi}\}$ can be optimized by maximizing the marginal posterior probability density $p(\boldsymbol{\theta}\,|\,\mathbf{y},\mathbf{X}) \propto p(\boldsymbol{\theta})p(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\theta})$, where $p(\boldsymbol{\theta}) = p(\sigma_\mathrm{m})\prod_{\phi=1}^{N_\phi} p(l_\phi)$ and

$$p(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}\,|\,\mathbf{0}, K(\mathbf{X},\mathbf{X}) + \mathbf{\Sigma}) \tag{25}$$

is the marginal likelihood of $\boldsymbol{\theta}$. The GP approximation for the energy $f(\mathbf{x}^*)$ at any location $\mathbf{x}^*$ is then obtained by GP regression as the mean of the posterior predictive distribution of $f(\mathbf{x}^*)$ conditional on the optimized hyperparameters $\boldsymbol{\theta}$,

$$\mathrm{E}[f(\mathbf{x}^*)\,|\,\mathbf{y},\mathbf{X},\boldsymbol{\theta}] = K(\mathbf{x}^*,\mathbf{X})(K(\mathbf{X},\mathbf{X})+\mathbf{\Sigma})^{-1}\mathbf{y}, \tag{26}$$

and the approximation for the partial derivative of the energy with respect to coordinate $x_{i,d}^*$ is given by

$$\mathrm{E}\left[\frac{\partial f(\mathbf{x}^*)}{\partial x_{i,d}^*}\,\middle|\,\mathbf{y},\mathbf{X},\boldsymbol{\theta}\right] = \frac{\partial K(\mathbf{x}^*,\mathbf{X})}{\partial x_{i,d}^*}(K(\mathbf{X},\mathbf{X})+\mathbf{\Sigma})^{-1}\mathbf{y}, \tag{27}$$

where the elements of $\partial K(\mathbf{x}^*,\mathbf{X})/\partial x_{i,d}^*$ are obtained by differentiating the covariance function.

The derivatives of the covariance function are needed also when including the force evaluations in the training data set.[25–28] When $\mathbf{y}$ is extended to include partial derivatives of f (components of negative force), the training covariance matrix $K(\mathbf{X},\mathbf{X})$ and covariance vector $K(\mathbf{x}^*,\mathbf{X})$ are extended correspondingly to include prior covariances between the energy and derivative values

$$\mathrm{Cov}\left[\frac{\partial f(\mathbf{x})}{\partial x_{i,d}}, f(\mathbf{x}')\right] = \frac{\partial}{\partial x_{i,d}}\mathrm{Cov}\big[f(\mathbf{x}), f(\mathbf{x}')\big] = \frac{\partial k(\mathbf{x},\mathbf{x}')}{\partial x_{i,d}} \tag{28}$$

and the covariances between the derivatives

$$\mathrm{Cov}\left[\frac{\partial f(\mathbf{x})}{\partial x_{i_1,d_1}}, \frac{\partial f(\mathbf{x}')}{\partial x'_{i_2,d_2}}\right] = \frac{\partial^2}{\partial x_{i_1,d_1}\partial x'_{i_2,d_2}}\mathrm{Cov}\big[f(\mathbf{x}), f(\mathbf{x}')\big] = \frac{\partial^2 k(\mathbf{x},\mathbf{x}')}{\partial x_{i_1,d_1}\partial x'_{i_2,d_2}}, \tag{29}$$

and the noise covariance matrix $\mathbf{\Sigma}$ is extended to include the noise variances for the force evaluations (σ_{d}^2) on the diagonal. Expressions for the derivatives of covariance functions $k_{1/r}$, k_x and $k_x^{\mathrm{M}-5/2}$ with respect to the atom coordinates $x_{i,d}$ and the hyperparameters $\boldsymbol{\theta}$ can be found in ref 19. The latter are useful in the hyperparameter optimization, which is here performed with the scaled conjugate gradient algorithm[29] implemented in the GPstuff toolbox.[30]

3.2 Algorithm description

If no information about the energy surface or the minimum energy orientation of the dimer is available in the beginning, it is useful to perform initial rotations with accurate evaluations to find the lowest curvature mode before starting to translate the dimer. These rotations can be done through a regular rotation scheme using either the conjugate gradient or the L-BFGS approach, but we choose to utilize GP regression also in the initial rotation phase. A similar initial phase where the lowest curvature mode is found by GP regression iterations is applied also in ref 20. With only a middle point and a randomized orientation given for the initial dimer, the GP-dimer algorithm proceeds as follows:

1. Evaluate accurate energy E_0 and force $\mathbf{F}_0$ at the middle point $\mathbf{R}_0$.
2. Check final convergence using accurate force $\mathbf{F}_0$.
3. Evaluate accurate energy E_1 and force $\mathbf{F}_1$ at $\mathbf{R}_1$.
4. Check rotational convergence using accurate forces $\mathbf{F}_0$ and $\mathbf{F}_1$.
5. Repeat initial rotations until rotational convergence:
   (a) Update the GP model based on the energy and force evaluations.
   (b) Rotate the dimer until rotational convergence using the GP approximation of the energy gradient.
   (c) Evaluate accurate energy E_1 and force $\mathbf{F}_1$ at $\mathbf{R}_1$.
   (d) Check rotational convergence using accurate forces $\mathbf{F}_0$ and $\mathbf{F}_1$.
6. Repeat GPR iterations until final convergence:
   (a) Update the GP model based on the energy and force evaluations.
   (b) Rotate and translate the dimer until early stopping or convergence using the GP approximation of the energy gradient.
   (c) Evaluate accurate energy E_0 and force $\mathbf{F}_0$ at $\mathbf{R}_0$.
   (d) Check final convergence using accurate force $\mathbf{F}_0$.

Figure 1 shows a two-dimensional illustration of the progression of the GP-dimer algorithm when finding a saddle point for dissociative adsorption of an H_2 molecule on a Cu(110) surface with fixed positions for the copper atoms. This example system is the same as the one used for testing the GP-NEB algorithm in ref 19. Each of the four graphs present a cut of the GP approximation to the energy surface based on energy and force data evaluated at the points marked by + signs. The pink and red bars represent the dimer in the beginning and end of the initial rotation round or GPR iteration, respectively. Starting from an initial dimer coinciding with the two-dimensional cut of the coordinate space, the lowest curvature mode of the accurate energy surface is found after two initial rotation rounds, and the GP model extrapolates a saddle point close to the correct location already based on the four data points evaluated around the start point. After one more evaluation at the saddle point found on the approximate energy

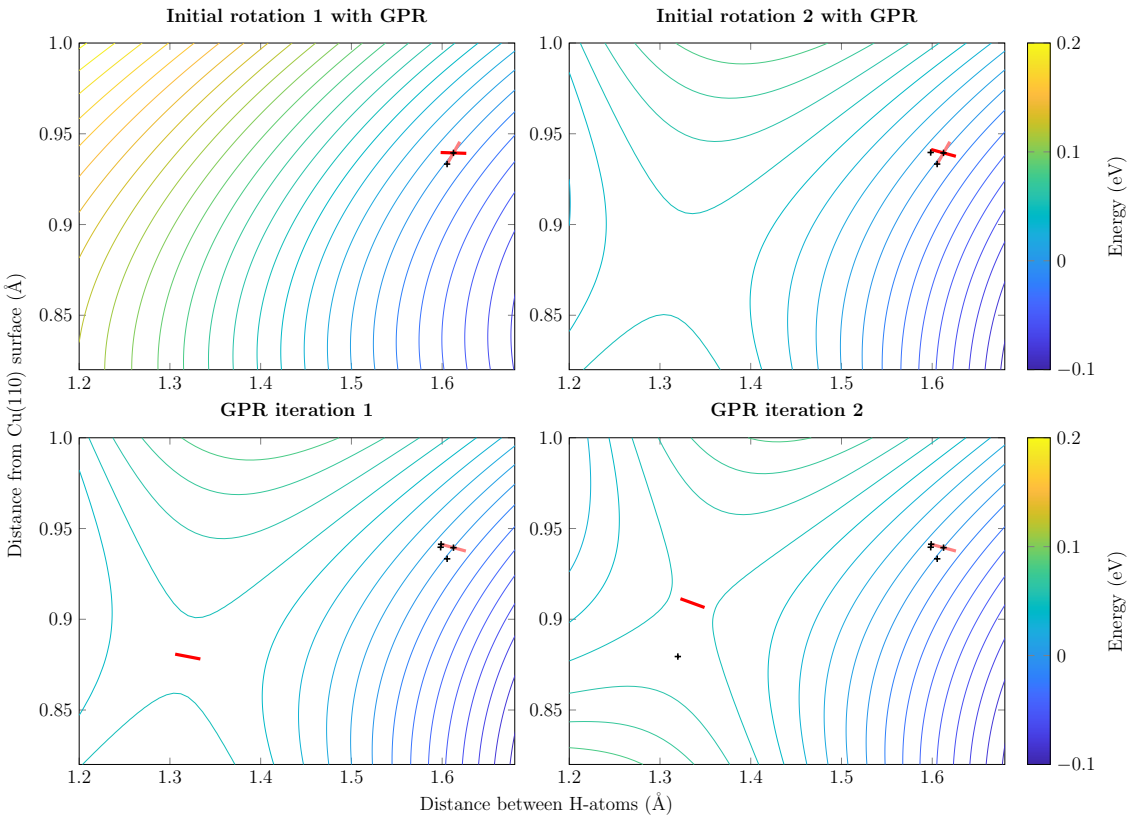

Figure 1: A two-dimensional cut through the energy surface of an H_2 molecule interacting with a Cu(110) surface. The H-H molecular axis lies in a plane parallel to the surface and perpendicular to the close packed rows of Cu-atoms. The upper graphs present GP approximations to the energy surface for the two initial rotation rounds and the lower graphs for the two GPR iterations of the GP-dimer algorithm. The training data points, marked with + signs, include both energy and force evaluations. The pink and red bars represent the dimer in the beginning and end of the initial rotation round or GPR iteration, respectively.

surface, the GP approximation is corrected and the middle point of the dimer converges to the correct saddle point.

Convergence of the GP-dimer algorithm has been reached when the maximum component of the accurate force $\mathbf{F}_0$ is below the final convergence threshold T_0 (here 0.01 eV/Å). Rotational convergence in the initial rotation phase is checked by calculating the preliminary rotational angle ω^*, given by eq 2, using the accurate forces $\mathbf{F}_0$ and $\mathbf{F}_1$. If more than one initial rotation rounds have been performed, also the angle between the converged orientations in the current and previous round is taken into account as an alternative criterion. The initial rotation phase is stopped when either of these angles is below T_ω (here 5 degrees). The maximum number of initial rotation rounds is set to the number of degrees of freedom in the system.

During the initial rotation rounds, we use the rotation scheme of LBFGS-dimer with forces approximated by the GP model to find the lowest curvature mode on the approximate energy surface. As an exception to LBFGS-dimer, the new force $\mathbf{F}_1^{\text{new}}$ is not estimated with eq 19 after the rotation iterations but the GP approximation is used also there. A tighter convergence threshold $T_\omega^{\text{GP}} = \min\{0.01 \text{ rad}, T_\omega/10\}$ is used for both the preliminary rotational angle ω^* and the realized rotational angle ω, given by eq 5. Each initial rotation round is here started from the same initial orientation.

The LBFGS-dimer algorithm is used also for dimer relaxation in the actual GPR iterations where the dimer is both rotated and translated on the approximate energy surface. Again, the GP approximation of the new force $\mathbf{F}_1^{\mathrm{new}}$ is used instead of estimating $\mathbf{F}_1^{\mathrm{new}}$ with eq 19 or estimating the new curvature C^{new} with eq 11 after any rotation iteration. The rotational convergence threshold T_ω^{GP} for ω^* or ω is now set to 0.01 rad and the convergence threshold T_0^{GP} for the maximum component of $\mathbf{F}_0$ on the approximate energy surface is set to 1/10 of the lowest accurate maximum component $\mathbf{F}_0$ evaluated so far. Here, dimer relaxation is always started from the same initial location with orientation obtained from the initial rotation phase.

In ref 19, the GP model based on inverse inter-atomic distances is coupled with an early stopping criterion constraining relative changes in the inter-atomic distances during the GP-NEB algorithm. The same early stopping criterion is used here for dimer relaxation inside the GPR iterations: After each translation iteration, there needs to exist an evaluated configuration $\mathbf{x}_{\mathrm{eval}}$ so that

$$\forall i \in \mathrm{A_m}\, \forall j \in \mathrm{A_m} \cup \mathrm{A_f} : \frac{2}{3} r_{i,j}(\mathbf{x}_{\mathrm{eval}}) < r_{i,j}(\mathbf{R}_0) < \frac{3}{2} r_{i,j}(\mathbf{x}_{\mathrm{eval}}). \tag{30}$$

If this condition does not hold, the last translation iteration is rejected and dimer relaxation stopped. Another early stopping criterion is based on the regular difference measure $\mathcal{D}_x$: After each translation iteration, there needs to exist an evaluated configuration $\mathbf{x}_{\mathrm{eval}}$ so that

$$\mathcal{D}_x(\mathbf{R}_0, \mathbf{x}_{\mathrm{eval}}) < L_x^{\mathrm{es}}. \tag{31}$$

This criterion with $L_x^{\mathrm{es}} = 0.5$ Å is applied also when using the stationary covariance functions k_x or $k_x^{\mathrm{M}-5/2}$.

To guarantee that the early stopping criterion in eq 30 cannot be triggered by a single translation step taken from an evaluated data point, a following limitation rule is set for the step length of translation iterations inside the GPR iterations:[19] An individual atom $i \in \mathrm{A_m}$ cannot move more than 99% of

$$\min_{j \in \mathrm{A_f} \cup \mathrm{A_m} \setminus \{i\}} r_{i,j}(\mathbf{R}_0)/6,$$

where the minimum is taken over all inter-atomic distances from that atom to any other atom in $\mathbf{R}_0$. If this limit is exceeded, the whole displacement vector is shortened so that the displacement of atom i is at the limit. A corresponding limitation rule to accompany the early stopping criterion in eq 31 is obtained by limiting the displacement vector to 99% of L_x^{es}. If this limit is exceeded, the displacement vector is simply shortened to the limit.

As in ref 19, an activation distance of 5 Å is applied here for the frozen atoms. This means that a frozen atom is taken into account in the covariance function of the GP model only if some of the moving atoms has been within the radius of 5 Å from the frozen atom in the configuration of the middle point of the dimer during the algorithm. The distances from the moving atoms to inactive frozen atoms are checked on each translation iteration inside the GPR iterations, and if new frozen atoms are activated, the GP model is updated.

4 Results

In this section, we present results for tests of the GP-dimer method with start points chosen randomly within the vicinity of saddle points for a dissociative adsorption of a hydrogen molecule on a Cu(110) surface and three different gas phase chemical reactions. The performance of the GP-dimer method with the inverse-distance covariance function $k_{1/r}$ and two stationary covariance functions, k_x and $k_x^{\mathrm{M}-5/2}$, is reported in terms of the required number of energy and force evaluations and compared to two variants of the regular dimer method, described above as CG-dimer and LBFGS-dimer. In addition, we compare the number of evaluations required

for finding the lowest curvature mode of the energy surface when performing initial rotations using the Gaussian process regression, conjugate gradient, or L-BFGS approach.

4.1 Application to H_2 dissociation on Cu(110)

Our first example transition is a dissociative adsorption of an H_2 molecule on the Cu(110) surface. The same transition has been used in ref 19 for testing the GP-NEB algorithm for finding the minimum energy path between given initial and final states. In the test system here, the two H-atoms are allowed to move, whereas the Cu-atoms are frozen. The energy surface is described in ref 1. The start point of the algorithm is chosen by a random displacement of 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, 1, 2, or 3 Å from the saddle point of the example transition in the six-dimensional coordinate space. Ten different start points and randomly chosen initial orientations are used for each distance. Figure 1 illustrates the progression of the GP-dimer algorithm in an easy example where the start point and the initial orientation coincide with the same two-dimensional cut of the coordinate space as the saddle point.

Figure 2 shows the number of energy and force evaluations required for convergence to a saddle point with the GP-dimer method and the two variants of the regular dimer method, CG-dimer (orange) and LBFGS-dimer (blue). In addition to the inverse-distance covariance function $k_{1/r}$ (green), GP-dimer results are shown also for the squared exponential covariance function k_x (red) and Matérn-5/2 covariance function $k_x^{\mathrm{M}-5/2}$ (violet). In almost all cases, GP-dimer requires less evaluations than the regular dimer methods, and the difference increases when moving the start point farther away from the saddle point of the example transition. With start points closer than 0.5 Å to the saddle point, there are only small differences in the performance between the three covariance functions in the GP-dimer calculations, but the benefits of the inverse-distance covariance function become visible with larger distances. Figure 3 shows an example of the behaviour of the GP-dimer method with different covariance

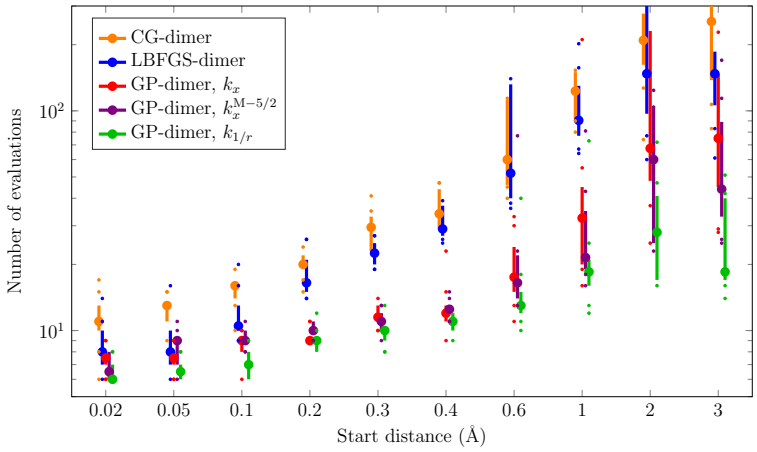

Figure 2: Number of energy and force evaluations required for convergence to a saddle point in the H_2/Cu(110) example using the regular CG-dimer (red) and LBFGS-dimer (blue) methods and the GP-dimer method with the squared exponential (red), Matérn-5/2 (violet), and inverse-distance (green) covariance functions. The distance of the start point of the calculation from the example saddle point is shown on the horizontal axis, and the vertical axis represents the number of evaluations in logarithmic scale. The large dots present the median number of evaluations among ten randomly chosen start positions. The bars present the interval between the third and eighth largest numbers, and the two smallest and largest numbers are presented by small dots if not included in the interval.

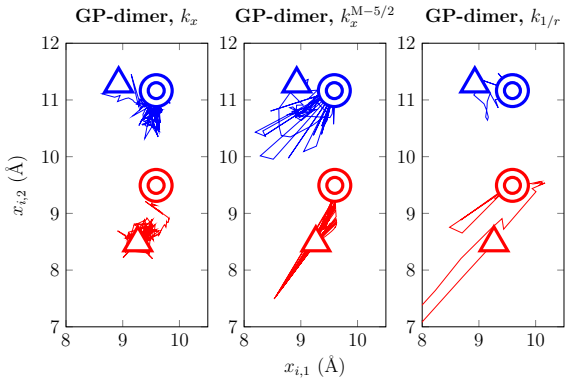

Figure 3: Example of the movement of the two H-atoms (blue and red) during the GP-dimer algorithm in the H_2/Cu(110) example when using the squared exponential (left), Matérn-5/2 (middle), and inverse-distance (right) covariance functions. The locations of the atoms are shown as projections on a plane parallel to the Cu(110) surface. The triangles represent the start configuration, and the double circles represent the saddle point where the algorithm converges.

functions with a start distance of 3 Å. The blue and red lines present the movement of the two H-atoms projected on the plane of the Cu(110) surface during the algorithm. When the inverse-distance covariance function is used (right), convergence is reached after 40 evaluations. With Matérn-5/2 (middle) and squared exponential (left) covariance functions, convergence to the same saddle point requires 170 and 228 evaluations, respectively.

4.2 Application to chemical reactions

Another set of test examples studied here involves three chemical reactions ($N_{\mathrm{m}} \leq 14$). The electronic structure computations for the energy and atomic forces are performed using the PM3 semi-empirical approach[31] as implemented within the ORCA suite of programs.[32] Reaction 1 is a simple addition of N_2O and ethylene to form oxadiazole, reaction 2 is the rearrangement of allyl vinyl ether to form 1-pentene-5-one, and reaction 3 is the removal of sulfurdioxide from butadiene sulfone.[33] The activation energies for the example reactions are relatively high: 1.86 eV, 3.36 eV, and 2.41 eV, respectively. For each of the reactions, the saddle point used as the center of start points is confirmed to have a single negative eigenvalue of the Hessian. The reactant, saddle point, and product state configurations of the example reactions are illustrated in Figure 4 alongside the corresponding result graphs.

Start points for the dimer calculations are chosen by a random displacement of 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.6, or 1 Å from the example saddle point in the $3N_{\mathrm{m}}$-dimensional coordinate space. Ten different start points and randomly chosen initial orientations are again used for each value of the distance. As shown in Figure 4, the pattern of the results is quite similar for each of the three test examples. Unlike in the H_2/Cu(110) example, the LBFGS-dimer method performs here clearly better than CG-dimer. The three variants of the GP-dimer method require again significantly less evaluations than the regular dimer methods, but the difference between the stationary and inverse-distance covariance functions starts to become appreciable already at 0.1 Å. From some start positions, the algorithms may end up in configurations where the energy and force evaluations fail. In such cases, the number of required evaluations is considered to be above 300.

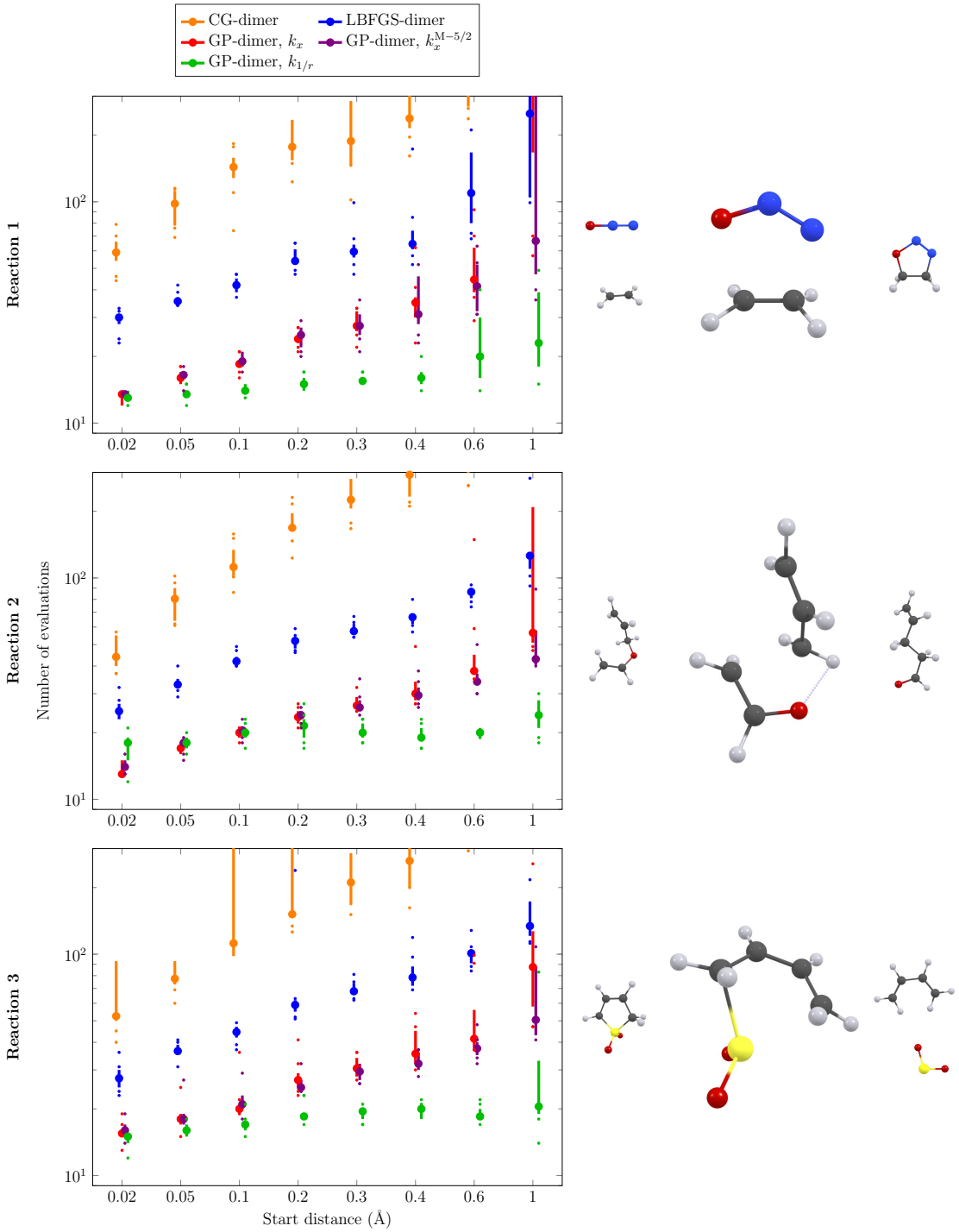

Figure 4: Number of energy and force evaluations required for convergence to a saddle point in three test examples using the regular CG-dimer (red) and LBFGS-dimer (blue) methods and the GP-dimer method with the squared exponential (red), Matérn-5/2 (violet), and inverse-distance (green) covariance functions. The distance of the start point of the calculation from the saddle point of the example reaction is shown on the horizontal axis, and the vertical axis represents the number of evaluations in logarithmic scale. The large dots represent the median number of evaluations among ten randomly chosen start positions. The bars present the interval between the third and eighth largest numbers, and the two smallest and largest numbers are represented by small dots if not included in the interval. The reactant, saddle point and product state configurations for each example reaction are visualized with the following atom colors: C: dark gray, H: light gray, O: red, N: blue, S: yellow.

4.3 Initial rotations

In the GP-dimer method, the initial rotation phase is performed using a Gaussian process regression approach where the lowest curvature mode at the start point is found by rotating the dimer on the approximate energy surface and refining the GP model based on accurate evaluations. To demonstrate the savings as compared to the conjugate gradient or L-BFGS approaches, we present results from separate tests where the rotations at the start point are continued until the preliminary rotation angle ω^*, given by eq 2, is below five degrees. If ω^* is based on estimated forces, the rotational convergence is confirmed by evaluating the accurate forces and the rotations are continued if necessary.

Figure 5 presents the number of evaluations required for rotational convergence in the H_2/Cu(110) example with the conjugate gradient (red), L-BFGS (blue), and GP regression approach (red, violet, and green for squared exponential, Matérn-5/2, and inverse-distance covariance functions, respectively). For the GP regression approach, the median of the number of evaluations remains between four and six with any of the three covariance functions. The median for the L-BFGS approach is 1–3 evaluations and the median for the conjugate gradient approach 3–11 evaluations larger than for the GP regression approach, but those results include more outliers. Due to the local nature of the training data set in the initial rotation phase, there are only small differences between the stationary and inverse-distance covariance functions.

Figure 6 presents corresponding results for the three chemical reaction examples. Again, the GP regression approach consistently converges with less evaluations than the comparison methods, although the difference to the L-BFGS approach is small, especially when using the stationary covariance functions. The performance of the conjugate gradient approach is significantly worse.

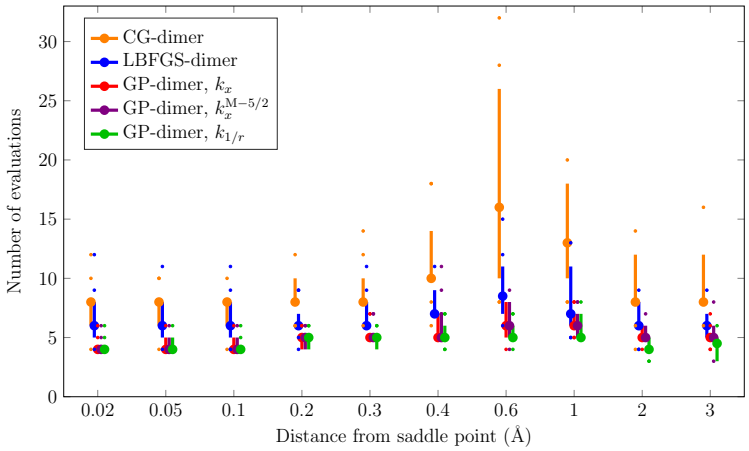

Figure 5: Number of energy and force evaluations required for rotations to the lowest curvature mode of energy in the H_2/Cu(110) example with the conjugate gradient (red) and L-BFGS (blue) approaches and with the Gaussian process regression approach using the squared exponential (red), Matérn-5/2 (violet), and inverse-distance (green) covariance functions. The distance between the location of the middle point of the dimer and the saddle point of the example transition is shown on the horizontal axis. The large dots present the median number of evaluations among ten randomly chosen start positions. The bars present the interval between the third and eighth largest numbers, and the two smallest and largest numbers are presented by small dots if not included in the interval.

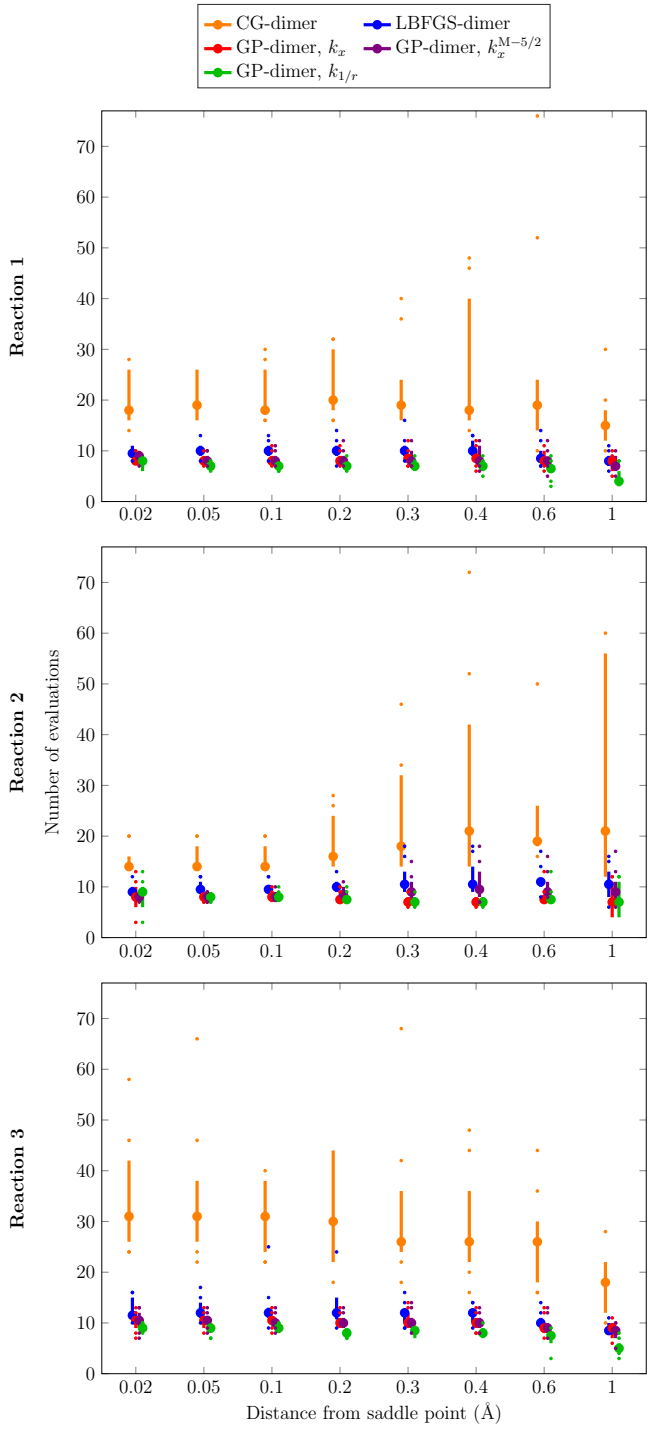

Figure 6: Number of energy and force evaluations required for rotations to the lowest curvature mode of energy in the three chemical reaction examples with the conjugate gradient (red) and L-BFGS (blue) approaches and with the Gaussian process regression approach using the squared exponential (red), Matérn-5/2 (violet), and inverse-distance (green) covariance functions. The distance between the location of the middle point of the dimer and the saddle point of the example reaction is shown on the horizontal axis. The large dots present the median number of evaluations among ten randomly chosen start positions. The bars present the interval between the third and eighth largest numbers, and the two smallest and largest numbers are presented by small dots if not included in the interval.

5 Discussion

The results presented here show that the inverse-distance covariance function suggested for GP-NEB calculations in ref 19 is beneficial also when applying the Gaussian process regression approach to minimum mode following calculations. The improved covariance function and the accompanied early stopping criterion are especially important when the start point for the calculation is not close to a saddle point. Already when the start point is displaced 0.1 Å from a saddle point, the GP-dimer method with the inverse-distance covariance function may require significantly smaller number of energy and force evaluations to reach convergence than when using a stationary covariance function. Our results show also generally that the GP regression approach (no matter which covariance function is used) gives advantage also when starting really close to the saddle point compared to the usual implementations of the dimer method based on conjugate gradient or L-BFGS algorithms. This applies to the initial rotations as well as to the translations of the dimer in the climb up the energy surface. Even though stationary covariance functions give convergence in the examples presented here, similar problems can arise as seen in GP-NEB calculations where the inclusion of large atomic forces can lead to failure in calculations based on stationary covariance functions.[19] Ultimately, the GP-dimer method can be used in repeated saddle point searches starting from a given local minimum to map out relevant low-lying saddle points in long time scale simulations.[3,4] We expect the robustness of the inverse-distance covariance approach will then be even more important.

We have assumed here that no information about the energy surface is available in the beginning of the minimum mode following calculation, only the coordinates and a random orientation of the dimer. If other information is available, the initial rotation phase may be unnecessary. This is the case, for example, if the start point has been obtained from an NEB calculation on some approximate energy surface, based for example on a lower level of electronic structure theory, or if an NEB calculation has converged only to a large tolerance. In these cases, information from the NEB calculation can be utilized when training the GP model, and that can make the GP regression approach even more useful.

After finding a saddle point, all eigenvalues of the Hessian at the saddle point are typically required to estimate the transition rate using the harmonic approximation to transition state theory. This can involve substantial computational effort for large systems. The GP model learned during the minimum mode following calculation gives a probability distribution for the energy surface, which can be used to estimate the Hessian and its eigenvalues as well as the uncertainty of these estimates and the calculated transition rate. If the rate cannot be estimated reliably enough, new energy and force calculations can be performed in a systematic way to update the GP model until the required confidence levels have been reached. Uncertainties of the second derivatives could also be utilized for deciding when to evaluate also image 1 (once or even repeatedly until rotational convergence) during the algorithm, which may become beneficial when starting far away from a saddle point.

Acknowledgements

This work was financially supported by the Academy of Finland (Grant 278260; Flagship programme: Finnish Center for Artificial Intelligence, FCAI, Grants 320181, 320182, 320183) and the Icelandic Research Fund. O.-P.K. was supported by the Finnish Cultural Foundation (Grant 180536) and V.Á. by a doctoral fellowship from the University of Iceland Research Fund. Computational resources were provided by the Aalto Science-IT project and the Icelandic High Performance Computing services located at the University of Iceland.

References

[1] Mills, G.; Jónsson, H.; Schenter, G. K. Reversible work based transition state theory: application to H_2 dissociative adsorption. *Surf. Sci.* **1995**, 324, 305.

[2] Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998; pp 385–404.

[3] Henkelman, G.; Jónsson, H. Long time scale kinetic Monte Carlo simulations without lattice approximation and predefined event table. *J. Chem. Phys.* **115**, 9657 (2001).

[4] Gutiérrez, M. P.; Argáez, C.; Jónsson, H. Improved minimum mode following method for finding first order saddle points. *J. Chem. Theory Comput.* **13**, 125 (2017).

[5] Ásgeirsson, V.; Jónsson, H. Exploring potential energy surfaces with saddle point searches. In *Handbook of Materials Modeling: Methods: Theory and Modeling*; Andreoni, W., Yip, S., Eds.; Springer: Cham, 2018.

[6] Henkelman, G.; Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **1999**, 111, 7010.

[7] Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **59**, 3969 (1999).

[8] Malek, R.; Mousseau, N. Dynamics of Lennard-Jones clusters: a characterization of the activation-relaxation technique. *Phys. Rev. E* **62**, 7723 (2000).

[9] Olsen, R. A.; Kroes, G. J.; Henkelman, G.; Arnaldsson, A.; Jónsson, H. Comparison of methods for finding saddle points without knowledge of the final states. *J. Chem. Phys.* **2004**, 121, 9776.

[10] Heyden, A.; Bell, A. T.; Keil, F. J. Efficient methods for finding transition states in chemical reactions: comparison of improved dimer method and partitioned ration function optimization method. *J. Chem. Phys.* **2005**, 123, 224101.

[11] Kästner, J.; Sherwood, P. Superlinearly converging dimer method for transition state search. *J. Chem. Phys.* **2008**, 128, 14106.

[12] Voter, A. F. Hyperdynamics: accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **1997**, 78, 3908.

[13] O'Hagan, A. Curve fitting and optimal design for prediction. *J. Royal Stat. Soc. B* **1978**, 40, 1.

[14] MacKay, D. J. C. Introduction to Gaussian processes. In *Neural Networks and Machine Learning*; Bishop, C. M., Ed.; Springer-Verlag: Berlin, 1998; pp 133–166.

[15] Neal, R. M. Regression and classification using Gaussian process priors. In *Bayesian Statistics 6*; Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M., Eds.; Clarendon Press: Oxford, 1999; pp 475–501.

[16] Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, 2006.

[17] Koistinen, O.-P.; Maras, E.; Vehtari, A.; Jónsson, H. Minimum energy path calculations with Gaussian process regression. *Nanosyst.: Phys. Chem. Math.* **2016**, 7, 925. A slightly corrected version available as e-print arXiv:1703.10423.

[18] Koistinen, O.-P.; Dagbjartsdóttir, F. B.; Ásgeirsson, V.; Vehtari, A.; Jónsson, H. Nudged elastic band calculations accelerated with Gaussian process regression. *J. Chem. Phys.* **2017**, 147, 152720.

[19] Koistinen, O.-P.; Ásgeirsson, V.; Vehtari, A.; Jónsson, H. Nudged elastic band calculations accelerated with Gaussian process regression based on inverse inter-atomic distances. Preprint ChemRxiv:8850440, 2019.

[20] Denzel, A.; Kästner, J. Gaussian process regression for transition state search. *J. Chem. Theory Comput.* **2018**, 14, 5777.

[21] Fletcher, R.; Reeves, C. M. Function minimization by conjugate gradients. *Comput. J.* **1964**, 7, 149.

[22] Polak, E.; Ribière, G. Note sur la convergence de méthodes de directions conjuguées. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique* **1969**, 3, 35.

[23] Nocedal, J. Updating quasi-Newton matrices with limited storage. *Math. Comput.* **1980**, 35, 773.

[24] Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, 45, 503.

[25] O'Hagan, A. Some Bayesian numerical analysis. In *Bayesian Statistics 4*; Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M., Eds.; Clarendon Press: Oxford, 1992; pp 345–363.

[26] Rasmussen, C. E. Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals. In *Bayesian Statistics 7*; Bernardo, J. M., Dawid, A. P., Berger, J. O., West, M., Heckerman, D., Bayarri, M. J., Smith, A. F. M., Eds.; Clarendon Press: Oxford, 2003; pp 651–659.

[27] Solak, E.; Murray-Smith, R.; Leithead, W. E.; Leith, D. J.; Rasmussen, C. E. Derivative observations in Gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems 15*; Becker, S., Thrun, S., Obermayer, K., Eds.; MIT Press: Cambridge, 2003; pp 1057–1064.

[28] Riihimäki, J.; Vehtari, A. Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Converence on Artificial Intelligence and Statistics*; Teh, Y. W., Titterington, M., Eds.; Proceedings of Machine Learning Research, 2010; pp 645–652.

[29] Bishop, C. M. *Neural Networks for Pattern Recognition*; Clarendon Press: Oxford, 1995; pp 282–285.

[30] Vanhatalo, J.; Riihimäki, J.; Hartikainen, J.; Jylänki, P.; Tolvanen, V.; Vehtari, A. GPstuff: Bayesian modeling with Gaussian processes. *J. Mach. Learn. Res.* **2013**, 14, 1175.

[31] Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989**, 10, 209.

[32] Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, 8, e1327.

[33] Birkholz, A. B.; Schlegel, H. B. Using bonding to guide transition state optimization. *J. Comput. Chem.* **2015**, 36, 1157.