JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

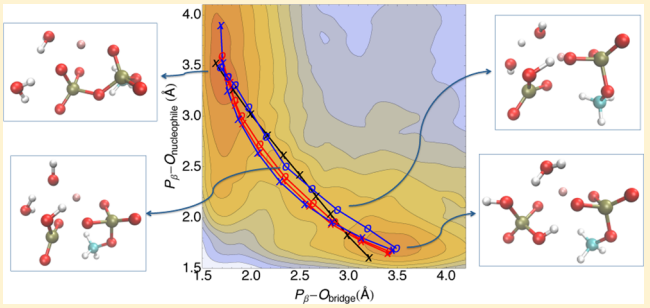
Finding Chemical Reaction Paths with a Multilevel Preconditioning Protocol

Seyit Kale,[†,‡] Olaseni Sode,[†,‡,§,∥,#] Jonathan Weare,[‡,∥,⊥] and Aaron R. Dinner*[,†,‡,§,∥]

[†]Department of Chemistry, [‡]James Franck Institute, [§]Institute for Biophysical Dynamics, [∥]Computation Institute, [⊥]Department of Statistics, University of Chicago, Chicago, Illinois 60637, United States

[#]Computing, Environment, and Life Sciences, Argonne National Laboratory, Argonne, Illinois 60439, United States

ABSTRACT: Finding transition paths for chemical reactions can be computationally costly owing to the level of quantum-chemical theory needed for accuracy. Here, we show that a multilevel preconditioning scheme that was recently introduced (Tempkin et al. *J. Chem. Phys.* **2014**, *140*, 184114) can be used to accelerate quantum-chemical string calculations. We demonstrate the method by finding minimum-energy paths for two well-characterized reactions: tautomerization of malonaldehyde and Claissen rearrangement of chorismate to prephanate. For these reactions, we show that preconditioning density functional theory (DFT) with a semiempirical method reduces the computational cost for reaching a converged path that is an optimum under DFT by several fold. The approach also shows promise for free energy calculations when thermal noise can be controlled.



INTRODUCTION

A central goal of quantum chemistry is to elucidate mechanisms and energetics of reactions. Both can be obtained from minimum (free) energy paths connecting reactants to products.[1] Methods such as conjugate peak refinement,[2] the nudged elastic band,[3,4] the string method,[5−7] and transition path sampling[8] refine a guess for the reaction path iteratively. These methods have yielded important insights in quantum chemical contexts[9−11] but are very computationally demanding because the energies and forces must be evaluated many times. While less expensive electronic-structure methods can often be used to generate reasonable reactant and product structures for single-point energy calculations, they are less reliable for intermediates and transition states and can lead to qualitatively incorrect results when searching for reaction paths.[12] Thus, methods for accelerating reaction path calculations while maintaining a desired level of theory are needed.

Recently, we proposed a multilevel (ML) preconditioning approach for accelerating the convergence of iterative molecular calculations and demonstrated its use for coupling fine-grained (FG) and coarse-grained (CG) models.[13] Preconditioning is a standard technique in numerical optimization; it can be viewed as a variable transformation that enables a root finding algorithm to converge in fewer steps.[14] In practice, the variable transformation is implicit. In our ML scheme, we perform a nested iteration that enables us to evaluate a CG model repeatedly to search for an optimum while always enforcing that it be a stable solution of a FG model. This approach is qualitatively different than one that combines the models sequentially, which is not guaranteed to lead to convergence. Previous studies[15−18] have used preconditioning to couple models with different resolutions but have employed schemes that require linearizing the approximate model.

In this article, we explore an analogous procedure for quantum chemical calculations. Here, a less computationally expensive, presumably less accurate method plays the role of the CG model and a more computationally expensive, presumably more accurate model plays the role of the FG model. An advantage in this context is that both models have the same number of degrees of freedom, simplifying their coupling. Hence, throughout the paper we refer to them as the reference (R) and preconditioning (P) levels of theory. We first demonstrate that our multilevel scheme accelerates refinement of the minimum energy paths of two chemical reactions: tautomerization of malonaldehyde (Figure 1) and Claissen rearrangement of chorismate to prephanate (Figure 2). We choose these reactions because they have served as prototypes for intrinsic reaction coordinate search and the reaction progress can be readily traced via the chemical bonds formed and broken.[19,20] We then illustrate how the approach can be extended to room-temperature free energy

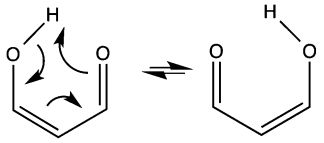


Figure 1. Reaction diagram for malonaldehyde tautomerization. Arrows indicate the flow of electron pairs.

Received: September 22, 2014
Published: November 7, 2014

   dx.doi.org/10.1021/ct500852y | *J. Chem. Theory Comput.* 2014, 10, 5467−5475

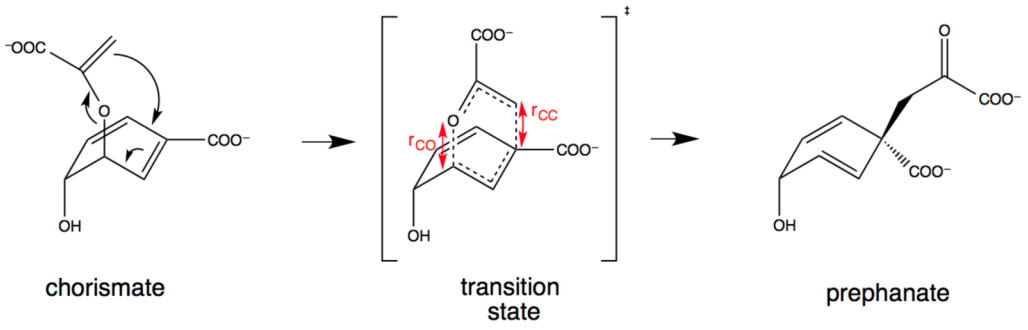


Figure 2. Claissen rearrangement of chorismate to prephanate. Red arrows indicate the distances that are used as CVs. Curved black arrows indicate the flow of electron pairs.

simulations with a phosphate hydrolysis reaction. While our approach is general, we specifically employ the string method for finding reaction paths; our preconditioning couples density functional theory (DFT) with semiempirical (SE) force fields. We provide evidence that better correspondence between the models leads to greater reduction in the number of iterations.

THEORY

Motivation. The theory behind nonlinear preconditioning can be understood by considering the simpler, linear case, as employed in Newton's method. We begin by observing that path refinement is mathematically equivalent to a multidimensional root finding problem, such that at the stationary solution, x_*, the component of the force in the direction orthogonal to the path vanishes:

$$f(x_*) = 0 \tag{1}$$

Fixed-point iteration solves eq 1 using

$$x_{m+1} = x_m - f(x_m) \tag{2}$$

For an initial state x_0 chosen sufficiently close to x_*, this iteration converges quickly when $f'(x_*)$ is nearly one (i.e., when $f(x) \approx x + c$) and slowly otherwise. Newton's method accelerates convergence to x_* by using the solution to a linear subproblem to define the next iterate x_{m+1} given the current iterate x_m. The linear subproblem can be written

$$g(x_{m+1}) = 0 \tag{3}$$

where

$$g(x) = f(x_m) + f'(x_m)(x - x_m) \tag{4}$$

which can be solved exactly at much lower cost than the original root finding problem.

Defining $y = g(x)$, we can write an alternative fixed-point iteration:

$$y_{m+1} = y_m - f(g^{-1}(y_m)) \tag{5}$$

or, in terms of the original variable,

$$g(x_{m+1}) - g(x_m) + f(x_m) = 0 \tag{6}$$

For the g defined above in eq 4, (observing that $g(x_m) = f(x_m)$), eq 6 reduces to Newton's method. Newton's method improves convergence because the function $f(g^{-1}(y))$ is closer to y (ignoring addition by a constant) than $f(x)$ was to x (ignoring addition by a constant).

This same principle applies for more general choices of g. Indeed, at convergence $g(x_{m+1}) = g(x_m)$ and eq 6 becomes $f(x_m) = 0$. As long as g is a reasonable approximation of f, we can expect the iteration in eq 6 to converge faster than fixed-point iteration. Of course it is also essential that the equation $g(x) = y$ can be solved relatively cheaply. These observations are the basis of the algorithms derived below.[13]

Formulation in Terms of the String Method. The string method[5−7] is a chain-of-states method in which the full system configuration is projected onto a smaller subspace of collective variables (CVs) that are expected to include the slowest varying degrees of freedom. The path is represented by discrete instances of the system (images) that interpolate an arc between two stable states. The goal is to gradually relax the arc to the most likely minimum energy path based on the local gradient. In the formulation that we employ,[6,7] the gradient is estimated by launching unbiased molecular dynamic simulations from image points and following their trajectories. The component of image displacement in the direction parallel to the string is removed by periodically modifying the image positions so that they are nearly equidistant along the string.

Switching our notation to string operators, we obtain successively better reaction pathways φ_i via

$$\varphi_{m+1} = \mathcal{S}_R(\varphi_m) \tag{7}$$

where $\mathcal{S}_R$ denotes the string operator for the reference model. Here, $\mathcal{S}_R$ represents the collective impact of releasing free dynamics trajectories from image points, averaging over the CVs, smoothing, and reparametrizating the resulting arc (see Computational Details). We seek a fixed-point solution φ_* that remains unaltered when $\mathcal{S}_R$ is applied to it, that is,

$$\varphi_* = \mathcal{S}_R(\varphi_*) \tag{8}$$

This equality is typically satisfied within thermal error when φ_* is in the vicinity of the minimum free energy path.

Notice that eq 7 is of the form of eq 2 with $f(\varphi) = \varphi - \mathcal{S}_R(\varphi)$. As in our earlier development, we define a change of variables $\xi = g(\varphi)$:

$$g(\varphi) = \varphi - \Delta\mathcal{S}_P$$

The operator $\mathcal{S}_P$ is analogous to $\mathcal{S}_R$ but with the string operations applied to a computationally inexpensive model (the preconditioning model, denoted P). Following the same steps as we did to write down eqs 5 and 6, we now obtain for the string case

$$\varphi_{m+1} = \Delta\mathcal{S}_P(\varphi_{m+1}) + \mathcal{S}_R(\varphi_m) - \Delta\mathcal{S}_P(\varphi_m) \tag{9}$$

dx.doi.org/10.1021/ct500852y | J. Chem. Theory Comput. 2014, 10, 5467−5475

The parameter $0 < \Delta < 1$ is defined by the user and controls the stability of the multilevel iteration—when the value of Δ is larger, the P model influences the iteration more strongly. For the examples in this study, we use $\Delta = 1.0$. Again, φ_* remains the solution of both eqs 8 and 9 because near convergence the additional terms on the right-hand side of eq 9 cancel.

From the practical point of view, eq 9 is nontrivial to solve because φ_{m+1} appears on both sides of the equality. One can approximate φ_{m+1} by solving the recursive relation

$$\varphi_{m+1}^{k+1} = \Delta\mathcal{S}_P(\varphi_{m+1}^{k}) + [\mathcal{S}_R(\varphi_m) - \Delta\mathcal{S}_P(\varphi_m)] \quad (10)$$

where k is a counter that runs up to a preset number K (typically 5 or 10). Thus, there is a nested iteration. The bracketed "correction" term on the right-hand side remains constant throughout the inner loop iterations.

The final path from the inner loop is fed back to the outer loop as the next guess for the solution:

$$\varphi_{m+1} = \varphi_{m+1}^{K} \quad (11)$$

The reformulation in eq 10 has the convenient feature that it only requires evaluations of the computationally inexpensive P model. The computationally expensive model enters only in the constant correction term, which is evaluated relatively few times.

COMPUTATIONAL DETAILS

Molecular Dynamics (MD). Constant temperature MD trajectories are performed using CP2K version 2.4, employing the QUICKSTEP module.[21] Becke, Lee, Yang, and Parr (BLYP)[22,23] and Perdew, Burke, and Ernzerhof (PBE)[24] exchange-correlation functionals are used in conjunction with the DZVP basis set.[25] Core electrons are represented via GTH pseudopotentials.[26] We use a plane wave cutoff of 300 Ry for the finest grid. Semiempirical force fields PM3,[27] AM1,[28] PM6,[29] and SCC-DFTB,[30] are as implemented in CP2K. Equal integration timesteps are used for DFT and SE dynamics (0.2 fs for malonaldehyde; 0.5 fs otherwise). Constant temperature is maintained via stochastic velocity rescaling.[31] Stable states and saddle points are calculated at the same levels of theory to verify string paths. In the cases of the near-zero-temperature reactions, we further verify the results by minimizing the energies of the end point structures of the final string, and the structures with the highest energies along the final strings are refined to obtain saddle points; the saddle point of malonaldehyde is found via an intrinsic reaction coordinate search, and that of chorismate is found via the dimer method.[32]

Parallelization. Our parallelization is based on the Swift[33] programming language, which abstracts various aspects of job submission. One string iteration of each image (steps 3 through 6 below) represents a separate Swift task that is submitted once its dependencies are ready. Swift initializes a Java script on the login node that interacts with the job submission interface of the computer. To reduce queue waiting times, Swift uses a customized resource provisioning service ("coaster")[34] such that compute node agents are not released back to the resource pool as long as there are more tasks to execute and queue upper limits permit doing so. The coaster script provides this feature by recording port indices and submitting new tasks through those channels. The current framework can be readily distributed over a grid of heterogeneous resources, although we did not pursue this option for the present work.

NEAR-ZERO-TEMPERATURE STRING SIMULATIONS

ML String Protocol. We use the following sequence of steps to execute the ML string equations in practice:

1. Generate an initial guess for the reaction path in the CV space by linearly interpolating between the end points. These end points are not required to be well-defined basins, as the end points of the string naturally relax to nearby minima.
2. Generate initial molecular configurations for the N images (for the examples presented, $N = 16$) by dragging the system (represented by the computationally inexpensive P model) from the reactant to the product basin with steered molecular dynamics. During the dynamics, we apply a harmonic restraint with a force constant of k_{drag} = 10 000 kcal/mol·Å^{-2}) to each CV and shift its minimum progressively from one image to the next over the course of 500 MD steps at $T = 1$ K for a total of 7500 steps.
3. Perform one iteration of the string method for both the R and P models separately, as follows.
   (a) Quench the images to their current positions by performing 50 MD steps at 0.01 K in the presence of harmonic restraints k_{freeze} = 10 000 kcal/mol·Å^{-2}.
   (b) Randomize the swarm by performing 50 MD steps at 1 K with a harmonic restraint with force constant k_{loose} = 100 kcal/mol·Å^{-2}.
   (c) Determine the displacement of each image by propagating each image for 50 MD steps at 1 K. The final resulting CV point is the average CV taken over all MD trajectory steps.
   (d) Smooth and reparametrize the resulting arcs such that image points are equidistant from each other along the CV hypersurface. Smoothing is performed as in refs 35 and 36 with a smoothing coefficient of $\kappa = 0.1$. The first and last image points are held fixed during this step.
4. Calculate the difference between R and P string updates from step 3c. The result is a vector in the CV space that has the same units and dimension as the string vector. This is the correction term that is added to each P update in the inner loop.
5. Iterate the inner loop, evaluating only the computationally inexpensive P model:
   (a) Generate an initial point by dragging each image for 250 steps at 1 K to the new CV points obtained from the R part of 3c.
   (b) Perform an iteration of the string method (following the protocol in steps 3a–d) for only the P model. Add the correction term each iteration and drag the P system to the resulting point. Repeat for $K = 5$ times.
6. For both models, drag the systems to the final point from the inner loop (step 5b) via 250 steps at 1 K.
7. Return to step 3a until convergence.

EXAMPLES

We first test our multilevel preconditioning scheme by applying it to two well-characterized reactions: the tautomerization of malonaldehyde and the Claissen rearrangement of chorismate to prephanate. For each system, we use a DFT method as the reference model and compare different preconditioning models

dx.doi.org/10.1021/ct500852y | J. Chem. Theory Comput. 2014, 10, 5467−5475

to determine the properties that give the best speedups. We show that in all cases the multilevel scheme converges to a minimum energy path of the reference model, even when the preconditioning model favors a qualitatively different path.

Malonaldehyde Tautomerization. The enol of malonaldehyde can undergo intramolecular proton transfers (Figure 1). Estimates for the hopping barrier vary with the level of theory, from 0.9 to 3.7 kcal/mol with DFT functionals[37] to 3.9−4.4 kcal/mol with coupled cluster approaches.[19,37,38] Hartree−Fock and early SE methods overestimate this and similar conversion barriers, probably as a result of underestimating electron correlation, and, consequently, also, the stabilization from conjugation in the transition state.[39,40] We tested four protocols, an R-only one with the BLYP functional (black, Figure 3), and three ML runs with SE preconditioners PM3 (red), PM6 (green), and SCC-DFTB (blue).

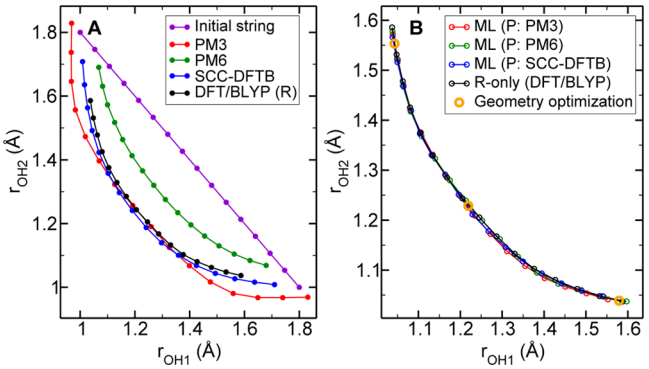


Figure 3. Reaction pathways for malonaldehyde tautomerization after 100 R- or P-only (A) or 20 ML (B) iterations of refinement.

We use as CVs the oxygen−hydrogen distances of the breaking and forming bonds, $r_{O_1H^*}$ and $r_{O_2H^*}$ where H* denotes the shared proton. All four force fields in this study predict approximately circular, symmetric arcs for the reaction path projected onto these coordinates (Figure 3A). The ML runs relax to the same circular arc as does the BLYP simulation (Figure 3B), passing through the saddle at $r_{O_1H^*} = r_{O_2H^*} = 1.22$ Å in agreement with earlier BLYP findings (1.223 Å via DZP[38] and 1.227 Å via cc-pVTZ[37]). The barrier heights for the converged paths are fairly close to each other at ≈1.32 kcal/mol even though P-only predictions are different by up to an order of magnitude (Figure 4). BLYP is known to underestimate this barrier, particularly with smaller basis sets;[37] our DZVP value falls in between the published values of 1.0 kcal/mol via DZP[38] and 2.0 kcal/mol via cc-pVTZ.[37]

To visualize the speed of convergence, we plot the norm of the displacement vector at each step of the iteration in Figure 5. This quantity should approach zero near convergence. We find that the ML scheme reaches the BLYP path 3−5 times faster than the R simulation. A similar speedup can be observed by following the barrier in the energy profile step-by-step (Figure 6).

Claissen Rearrangement. Conversion of diaxial chorismate to prephanate is a prototypical Claissen rearrangement, one of the best known pericyclic reactions (Figure 2). This reaction is a demanding testcase for our ML preconditioning scheme because the transition state has a cyclic, highly delocalized structure with significant electron correlation that is not

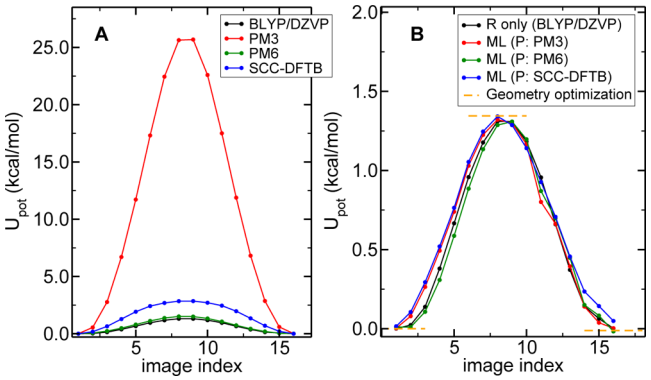


Figure 4. Potential energy profiles along the string for the malonaldehyde tautomerization after 100 R- or P-only (A) or 20 ML (B) iterations of refinement. Color convention is as in Figure 5. The predicted barrier height is ≈1.32 kcal/mol. Energies from geometry optimization are indicated with orange dashed lines.

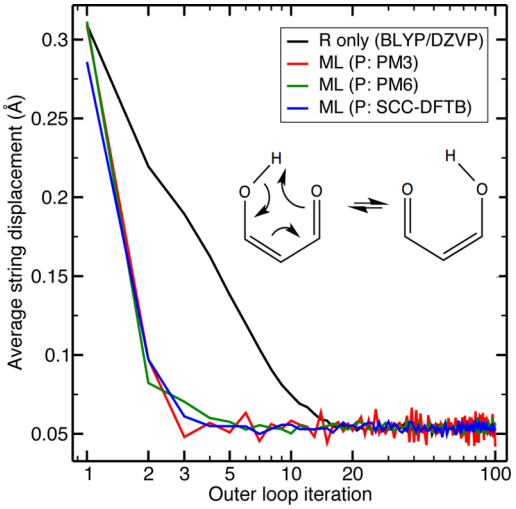


Figure 5. Convergence of the malonaldehyde tautomerization for the string method with only DFT (black) and in the ML preconditioning scheme with various inexpensive P models (red, PM3; green, PM6; and blue, SCC-DFTB). Progress is measured by the norm of the net string displacement as projected on the CV subspace and averaged over all images. Inset shows reaction diagram.

captured well by SE force fields.[12,41] Low levels of theory typically favor two-stage reaction scenarios (as opposed to concerted bond formations and ruptures), which are often hard to reconcile with experimental evidence.[12]

For the string method, we use the CO and CC distances indicated in Figure 2 as the CVs (r_{CO} and r_{CC}). We use DFT with the PBE functional for the reference model. In addition to PM3, we also precondition with DFT with the BLYP functional. Although the latter is of comparable computational cost to the reference, we consider it to explore how the properties of the preconditioning model affects the convergence (measured in the number of outer loop iterations).

The initial and final paths predicted by the two DFT methods indeed resemble each other more closely than that of PM3 (Figure 7A). Starting from an initial linear interpolation (black), both PBE (red) and BLYP (blue) converge to an extended transition state characterized by simultaneously long CO and CC distances around central images. PM3 (green), on the other hand, predicts a compact transition state, and the

dx.doi.org/10.1021/ct500852y | J. Chem. Theory Comput. 2014, 10, 5467−5475

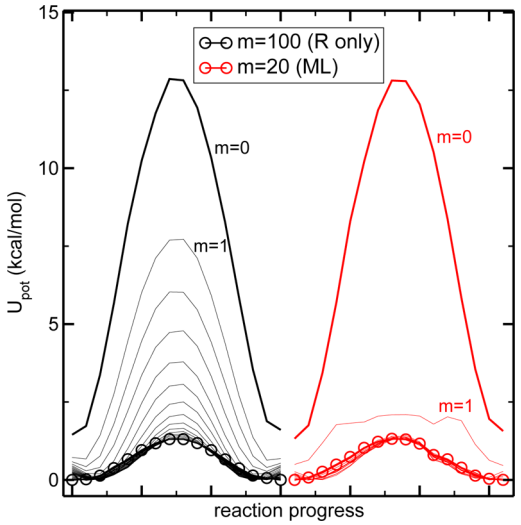


Figure 6. Monotonic relaxation of the potential energy profiles in R-only and ML (P: PM3) refinements for malonaldehyde tautomerization. Initial profiles and first iterations are indicated as $m = 0$ and $m = 1$, respectively. Last energies (100th R-only iteration and 20th ML iteration) are indicated with circles for each image. All iterations are shown.

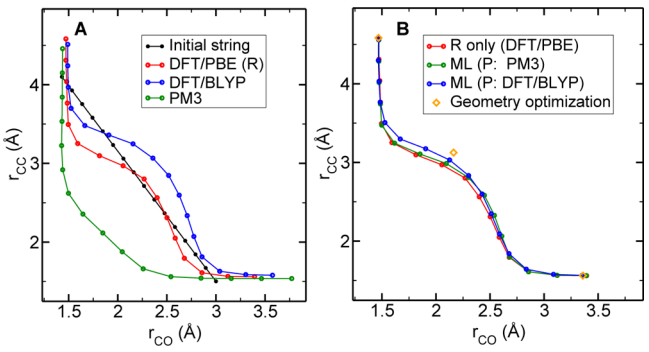


Figure 7. Minimum energy pathways for the Claissen rearrangement. (A) Predictions from each method by itself: PBE (R, red), BLYP (blue), and PM3 (green) alone. (B) R-only path compared against the two ML strings. Points from geometry optimization are in orange.

bonds that are ruptured maintain their covalent character longer as indicated by their near-equilibrium distances (~1.5 Å) in the basins. In comparison, all final ML paths agree fairly well with the PBE prediction (Figure 7B) as well as the basin and saddle points from geometry optimization (orange). As for the energies, PBE predicts the activation barrier and reaction enthalpy to be 22.0 and −19.7 kcal/mol (Figure 8A, red), which is close to earlier estimates of 17.5 and −20.8 kcal/mol, respectively, using the same basis set but a lower plane wave cutoff and different software.[20] The BLYP and PBE barriers are similar (19.0 and −17.2 kcal/mol, respectively), while the PM3 barrier is far higher (53.1 kcal/mol; the reaction enthalpy is −23.9 kcal/mol). ML energies (Figure 8B) are comparable with PBE.

The rate of convergence, as measured by the norm of the displacement vector averaged over all images, is shown in Figure 9. In terms of the number of reference model evaluations (outer loop iterations), we see that preconditioning with BLYP leads to a greater acceleration than preconditioning with PM3. For BLYP, an inner loop iteration costs about as much as the reference model's outer loop iteration. Because there are 5 inner loop iterations plus one energy evaluation for

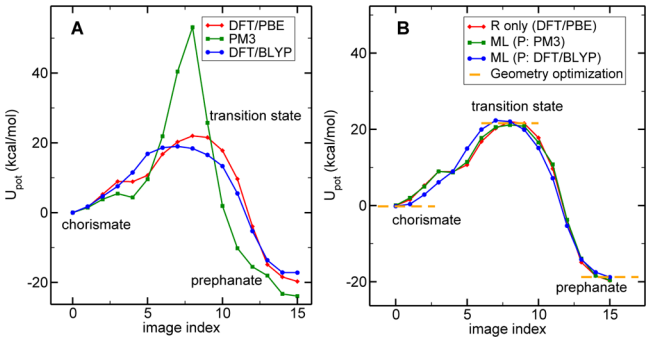


Figure 8. Potential energy profiles for the Claissen rearrangement. (A) PBE, PM3, and BLYP alone. (B) PBE compared with the results from the ML scheme with PM3 (green) and BLYP (blue) preconditioning. Energies from geometry optimization are indicated with orange dashed lines.

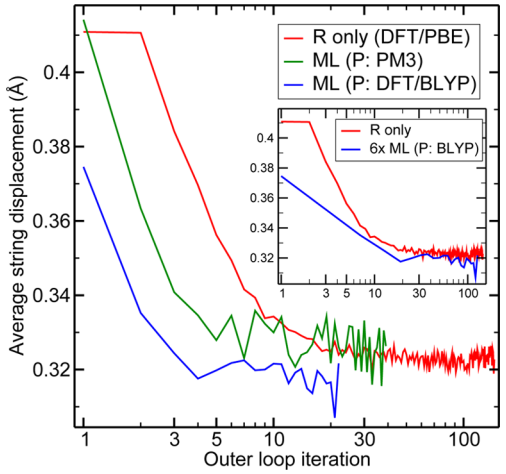


Figure 9. Convergence of ML path refinement of the Claissen reaction. Convergence is expressed in terms of the norm of the displacement vector as projected over the CVs and averaged over all images. Abscissa is number of R iterations. The inset shows a version of the BLYP ML curve in which the number of iterations is scaled by the number of energy evaluations per outer loop iteration (6), which provides a comparison in terms of rough computational cost.

the outer loop, we multiply the ML iterations by 6 in the inset to Figure 9 to obtain a rough estimate of relative computational costs. This shows that the speedup comes from shifting the work into the inner loop.

FINITE-TEMPERATURE STRING SIMULATIONS

We now consider extension of the ML scheme to a reaction in which entropy plays a significant role at room temperature. The finite-temperature case is more challenging as the preconditioning amplifies noise in the reaction path optimization, and we show how we suppress this effect. The specific example that we consider is hydrolysis of monomethylpyrophosphate (MPP) in the presence of two water molecules. We show results obtained with PM6 as the reference force field and PM3 as the preconditioning force field because this choice allows convergence of both R- and ML-string calculations in a feasible amount of time. However, we have also performed multilevel string calculations for this reaction with DFT with the PBE functional as the R model and the SE method PM6 as the P model, and we obtained a comparable initial reduction in the number of outer loop iterations.

dx.doi.org/10.1021/ct500852y | J. Chem. Theory Comput. 2014, 10, 5467−5475

Table 1. Initial MPP End Points[a]

CV	image 1 (reactant)	image 10 (product)
$P_\alpha - P_\beta$	2.96	4.50
$P_\beta - O_{br}$	1.61	3.20
$P_\beta - O_n$	3.50	1.60
$O_n - H_{n,1}$	0.96	2.00
$O_c - H_{n,1}$	2.40	0.96
$O_c - H_{c,1}$	0.96	2.00
$O_\beta - H_{c,1}$	1.36	2.00
$O_n - H_{n,2}$	0.96	0.96 (fixed)
$O_c - H_{c,2}$	0.96	0.96 (fixed)

[a]Abbreviations are as follows: (br)idge, (n)ucleophile, and (c)atalytic. Distances are in Ångstrom.

ML String Protocol.

1. Generate an initial guess for the reaction path in the CV space by linearly interpolating between the end points in Table 1.
2. Generate initial molecular configurations for the N images (here, N = 10) by dragging the system from the reactant to the product basin with steered molecular dynamics at the reference level of theory (PM6). During the dynamics, we apply a harmonic restraint with a force constant of k_{drag} = 5000 kcal/mol·Å^{-2} to each CV and shift its minimum progressively from one image to the next over the course of 10 000 MD steps at T = 300 K for a total of 90 000 steps. These structures are also used to start the PM3 calculations.
3. Perform one iteration of the string method for both the R and P models separately, as follows.
   (a) Equilibrate the images at their current positions to their desired temperatures in the presence of harmonic restraints k_{equil} = 5 000 kcal/mol·Å^{-2}. We equilibrate the reference model at 300 K for 2500 MD steps; images for the preconditioning model are equilibrated at 10 K for 500 MD steps.
   (b) Determine the displacement of each image. For the reference model, propagate a swarm of 25 copies of the system at the reference level of theory, each for 10 MD steps at 300 K under a harmonic restraint with force constant $k_{loose,R}$ = 5 kcal/mol·Å^{-2}. Each member of the swarm begins with different velocities selected at random from a Maxwell–Boltzmann distribution. For the preconditioning model, propagate a single copy for 10 MD steps at 10 K under a harmonic restraint with force constant $k_{loose,P}$ = 25 kcal/mol·Å^{-2}. In each case, determine the image displacements by averaging over all copies and trajectory steps associated with each image. We exclude points that are outside a cutoff to ensure that displacements are limited in extent. The cutoff is 0.25 Å in the CV space defined by the first three CVs in Table 1, which involve only heavy atoms, and 0.50 Å in the CV space defined by the remaining CVs, which involve hydrogen atoms.
   (c) Apply each displacement to its image if and only if the energy averaged over the swarm points within the cutoff is lower than the energy averaged over the equilibration steps in step 3a.
   (d) Smooth and reparametrize the resulting arcs five consecutive times such that image points are nearly equidistant in the CV space. We use a smoothing coefficient of κ = 0.1. The first and last image points are held fixed during this step.
4. Calculate the difference between the net displacements of the reference and preconditioning models at this iteration, d_m. Let the correction term (eq 10) be the average

$$d_m^{\mathrm{avg}} = \frac{\sum_{i=1}^{m} w_0^{m-i} d_i}{\sum_{i=1}^{m} w_0^{m-i}} \tag{12}$$

   where m is the present interation and w_0 = 0.75 is a constant that controls the relative weights of newer and older iterations. Set the contributions from protonic CVs to zero. These degrees of freedom are very noisy due to the fact that the two force fields can favor different protonation states for a given heavy atom configuration.
5. Iterate the inner loop, evaluating only the preconditioning model:
   (a) Generate an initial point by dragging each image for 500 MD steps at 10 K to the new CV points obtained from the reference model part of step 3d.
   (b) Perform an iteration of the string method for only the preconditioning model. Add the average correction term (eq 12) each iteration and drag the system to the resulting point. Smooth and reparametrize as in step 3d. Average over any prior preconditioning iterations (within this inner loop) such that

$$\varphi_{m+1}^{k,\mathrm{avg}} = \frac{\sum_{i=1}^{k} w_0^{k-i} \varphi_{m+1}^{i}}{\sum_{i=1}^{k} w_0^{k-i}} \tag{13}$$

   where m denotes the outer loop iteration, and k the inner loop iteration. The weighting is geometric, as in eq 12, with the same biasing constant w_0. Repeat for K = 10 times.
6. Geometrically average the difference between the final point from step 5b and the starting point from step 3b over any prior outer loop displacements using forms analogous to the inner-loop equations (eqs 12 and 13). Scale the resulting displacement by 0.9, and drag both systems to the resulting target image position. We drag the reference model at 300 K for 2500 MD steps; we drag the preconditioning model at 10 K for 500 MD steps. We use the same force constants as in step 3a.
7. Return to step 3a until convergence.

Example. Our choice of phosphate ester hydrolysis is motivated by the discrepancies reported between potential surface mapping[42−44] and targeted molecular dynamics[45−47] studies in characterizing the most likely pathway. There has been debate about the extent of associativity.[48] In a fully associative scenario (the S_N2 limit), the nucleophilic water inserts itself head-on, followed by the rupture of the bridging bond. This corresponds to a trigonal bipyramidal transition state, that is, a pentacoordinated β-phosphorus. In the alternative (S_N1) limit, a dissociative scenario,[49] a trigonal planar species is sandwiched in between the nucleophile and the anhydride bridging oxygen at distances as large as the sum of van der Waals radii (~3.3 Å).[48] Furthermore, this picture implies that the phosphodiester bond is ruptured before the nucleophile can be incorporated. Naturally, the preferred path depends on the electrostatic environment and the pK_a of the leaving group.[42]

dx.doi.org/10.1021/ct500852y | J. Chem. Theory Comput. 2014, 10, 5467−5475

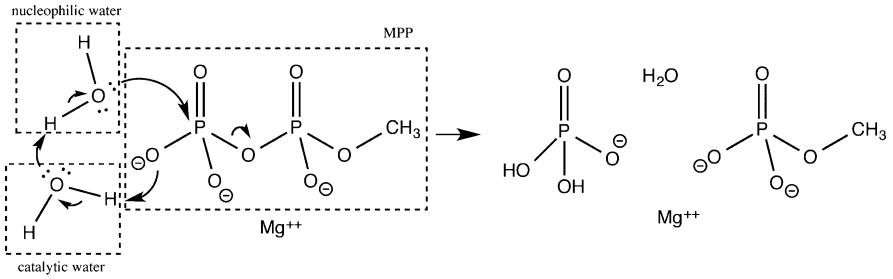


Figure 10. Diagram of hydrolysis of MPP catalyzed by two water molecules. Arrows indicate directions of electron pair movements. Nucleophilic and catalytic water molecules are indicated.

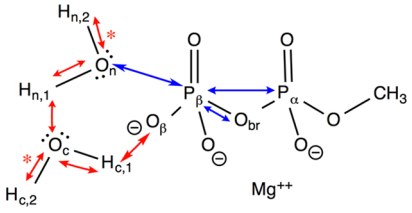


Figure 11. CVs used in the MPP example. Blue and red arrows indicate nonprotonic and protonic CVs, respectively. Bonds marked with an asterisk are fixed. Notations follow Table 1.

We consider hydrolysis in the presence of two water molecules, one nucleophilic and the other catalytic (Figure 10). For visualization purposes only, the reaction path is projected onto the breaking and forming oxygen−phosphate distances, those of the β-phosphorus with the anhydride bridging or the nucleophilic water oxygens (Figure 11). This representation is known as a More O'Ferrall−Jencks (MOFJ) plot. A perfectly associative reaction corresponds to a path that goes through the lower left corner, while a perfectly dissociative path would go through the upper right corner.

The space orthogonal to the MOFJ plot involves many protonic degrees of freedom. The diversity of possible protonation states demands consideration of additional coordinates when simulating the reaction. We use the nine CVs that are indicated in Table 1 and Figure 11. They include the breaking and forming bonds, as well as two inert O−H bonds that are fixed to reduce competing protonation pathways. Additionally, the CVs include four oxygen-proton CVs (protonic CVs); these are free to vary except at the string end points. For the terminal images, we fix the O−H distances at 0.96 Å to select for specific reaction and product protonation states. As noted in step 4 of the ML string protocol, only nonprotonic CVs (i.e., P_α−P_β, P_β−O_n, and P_β−O_{br}) contribute to the correction term.

As mentioned above, we use the PM6 force field as the reference and precondition it with low-temperature (T = 10 K) PM3. As can be seen in Figure 12, the PM6 reaction path consists of protonation of a phosphate by the catalytic water, followed by attack by the nucleophilic water and neutralization by transferring a proton between the waters (Figure 12). The ML simulations reprise this path (Figure 12), reaching it in about 3-fold fewer iterations (Figure 13).

Energies of the transition states found by reference and ML strings are comparable (Figure 14); however, basins exhibit differences because of hydrogen bonding rearrangements of water. This is more prominent in the product basin where the ML scheme finds more favorable interactions between the catalytic water and magnesium ion. To verify that these are indeed PM6 solutions, we switch off ML and continue PM6

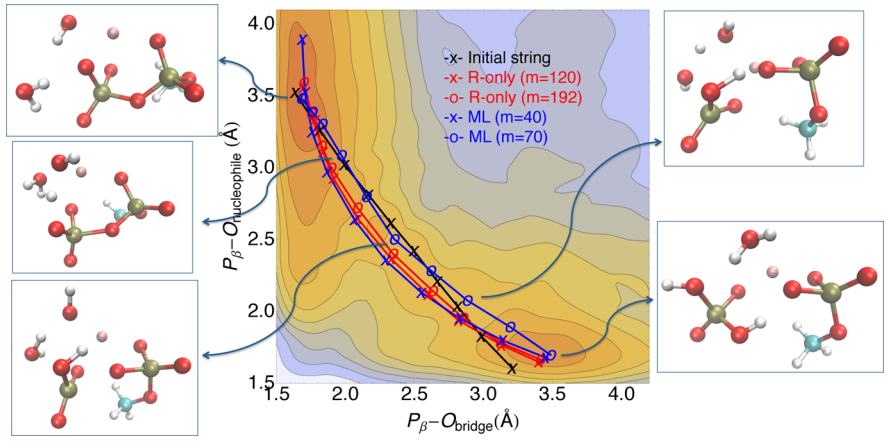


Figure 12. MOFJ plot for MPP hydrolysis. Abscissa is P_β-O_{br} distance, ordinate is P_β-O_n distance. Red and blue string pairs indicate first and last R and ML paths included in averaging potential energies. Insets are representative snapshots from the final ML path. Initial path is in black. Contoured landscape is from well-tempered metadynamics[50] (WTM) to provide information about the energy landscape local to the final string pathway. WTM setup uses 10 walkers[51] that are restarted every 500 MD steps from the final points of R-only images. Gaussian hills of height 1 kcal/mol are deposited every 50 steps on the space spanned by nonprotonic CVs. For WTM, ΔT is 5000 K. To reduce water evaporation in WTM, reflective boundaries are applied on all CVs at 5.0 Å. Walkers are integrated using the same MD time step and thermostat as in string trajectories. Over ∼51 000 hills are collected to reproduce the free energy surface. These simulations show that the transition state region is fairly flat, so that the variations in the paths are likely to reflect thermal fluctuations. Free energy contours are spaced by 5 kcal/mol.

dx.doi.org/10.1021/ct500852y | J. Chem. Theory Comput. 2014, 10, 5467−5475

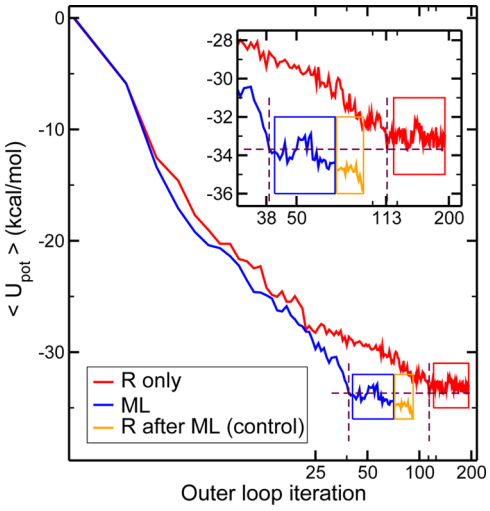


Figure 13. Convergence of MPP path optimization: R-only (red) and ML (blue). We measure convergence by the decrease of the R-model potential energy averaged over all images. Orange indicates control R-only sequence starting from the last ML point. Boxes show the plateau regions for the most likely minimum energy paths. Inset shows an expanded view near convergence.

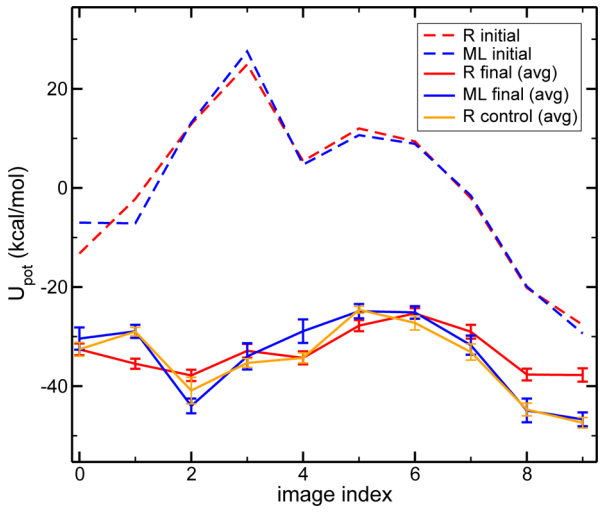


Figure 14. Initial and final energy profiles for MPP hydrolysis. Final profiles are averaged over the regions indicated inside the boxes from Figure 13. Bars show standard deviation.

string simulations for another 20 iterations. The path remains stable. We estimate activation barrier and the reaction enthalpy as 8.0 and −14.8 kcal/mol. While the level of theory and limited solvent representation employed here preclude drawing conclusions about the reaction, the simulations show that the ML preconditioning scheme can accelerate convergence of complex systems at room temperature. A caveat is that, achieving this speedup requires averaging the displacements and the correction term to control the noise, which the preconditioning amplifies.

CONCLUSIONS

Here, we show that a multilevel preconditioning scheme can accelerate quantum-chemical path optimization. We illustrate the scheme with the string method, but the approach as described could be immediately applied to other path finding methods[3,4] (likewise, geometry optimization, as discussed in ref 13). The scheme differs from traditional strategies for combining different levels of theory in that it uses information from both models at all times, in a way that guarantees that convergence is to a path that is stable under the reference model. As is true for almost any multidimensional optimization problem, this solution is not guaranteed to be the global minimum. The examples that we considered suggest that preconditioning with a model that better corresponds to the reference model in energy leads to a greater decrease in the number of reference model evaluations—essentially, more work can be shifted to the preconditioning model (the inner loop of the nested iteration). The multilevel approach suggested here is closely related to parallel-in-time integration, and it is possible that techniques developed to accelerate parallel-in-time integration could also be adapted to the present context (see ref 52).

Of course, the speedup in CPU time depends on the relative costs of the two models. The SE methods that we primarily use for preconditioning here are negligible in computational cost compared with the reference DFT. Given that a wide range of methods are computationally inexpensive in comparison to ab initio methods that account for electron correlation, especially with large basis sets, we expect significant speedups to be possible in applications that demand chemical accuracy. It will be interesting to explore such applications, as well as further developments to the method such as selectively choosing between preconditioning models to capture their different strengths as an optimization progresses.

AUTHOR INFORMATION

Corresponding Author

*E-mail: dinner@uchicago.edu. Phone: +1 (773) 702-2330. Fax: +1 (773) 702-4180.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research is supported by the National Institute of Health (NIH) Grant Number 5 R01 GM109455-02, and the National Science Foundation (NSF) through the Center for Multiscale Theory and Simulation (CHE-1136709). Computational resources were provided by the University of Chicago Research Computing Center (RCC). Additional resources were provided by NIH through resources of the Computation Institute and the Biological Sciences Division of the University of Chicago and Argonne National Laboratory, under grant S10 RR029030-01. We thank John H. Weare for helpful discussions and Michael Wilde for assistance with the parallel Swift framework.

REFERENCES

(1) Fukui, K. *Acc. Chem. Res.* **1981**, *14*, 363−368.
(2) Fischer, S.; Karplus, M. *Chem. Phys. Lett.* **1992**, *194*, 252−261.
(3) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9901−9904.
(4) Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9978−9985.
(5) E, W.; Ren, W. Q.; Vanden-Eijnden, E. *Phys. Rev. B.* **2002**, *66*, 052301.
(6) Maragliano, L.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2007**, *446*, 182−190.
(7) Pan, A. C.; Sezer, D.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3432−3440.
(8) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. *Adv. Chem. Phys.* **2002**, *123*, 1−78.

(9) Valiev, M.; Yang, J.; Adams, J. A.; Taylor, S. S.; Weare, J. H. *J. Phys. Chem. B* **2007**, *111*, 13455−13464.
(10) Saen-Oon, S.; Schramm, V. L.; Schwartz, S. D. *Z. Phys. Chem.* **2008**, *222*, 1359−1374.
(11) Sheppard, D.; Terrell, R.; Henkelman, G. *J. Chem. Phys.* **2008**, *128*, 134106.
(12) Houk, K. N.; Gonzalez, J.; Li, Y. *Acc. Chem. Res.* **1995**, *28*, 81−90.
(13) Tempkin, J. O. B.; Qi, B.; Saunders, M. G.; Roux, B.; Dinner, A. R.; Weare, J. *J. Chem. Phys.* **2014**, *140*, 184114.
(14) Nocedal, J.; Wright, S. J. *Numerical Optimization*; Springer Series in Operation Research and Financial Engineering; Springer: New York, 2006.
(15) Qiao, L.; Erban, R.; Kelley, C. T.; Kevrekidis, I. G. *J. Chem. Phys.* **2006**, *125*, 204108.
(16) Schlegel, H. B. *Theor. Chim. Acta* **1984**, *66*, 333−340.
(17) Fischer, T. H.; Almlof, J. *J. Phys. Chem.* **1992**, *96*, 9768−9774.
(18) Zhao, Z. J.; Wang, L. W.; Meza, J. *Phys. Rev. B* **2006**, *73*, 193309.
(19) Tautermann, C. S.; Voegele, A. F.; Loerting, T.; Liedl, K. R. *J. Chem. Phys.* **2002**, *117*, 1962−1966.
(20) Crespo, A.; Scherlis, D. A.; Marti, M. A.; Ordejon, P.; Roitberg, A. E.; Estrin, D. A. *J. Phys. Chem. B* **2003**, *107*, 13728−13736.
(21) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103−128.
(22) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098−3100.
(23) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789.
(24) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(25) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. *Can. J. Chem.* **1992**, *70*, 560−571.
(26) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703−1710.
(27) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209−220.
(28) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(29) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173−1213.
(30) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260−7268.
(31) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.
(32) Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **1999**, *111*, 7010−7022.
(33) Wilde, M.; Hategan, M.; Wozniak, J. M.; Clifford, B.; Katz, D. S.; Foster, I. *Parallel Comput.* **2011**, *37*, 633−652.
(34) Hategan, M.; Wozniak, J. M.; Maheshwari, K. *Proceedings of the Fourth IEEE International Conference on Utility and Cloud Computing*, 2011; pp 114−121.
(35) Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194103.
(36) Dickson, A.; Warmflash, A.; Dinner, A. R. *J. Chem. Phys.* **2009**, *130*, 074104.
(37) Sadhukhan, S.; Munoz, D.; Adamo, C.; Scuseria, G. E. *Chem. Phys. Lett.* **1999**, *306*, 83−87.
(38) Barone, V.; Adamo, C. *J. Chem. Phys.* **1996**, *105*, 11007−11019.
(39) Morpurgo, S.; Bossa, M.; Morpurgo, G. O. *J. Mol. Struct.: THEOCHEM* **1998**, *429*, 71−80.
(40) Benderskii, V. A.; Vetoshkin, E. V.; Irgibaeva, I. S.; Trommsdorff, H. P. *Chem. Phys.* **2000**, *262*, 393−422.
(41) Lyne, P. D.; Mulholland, A. J.; Richards, W. G. *J. Am. Chem. Soc.* **1995**, *117*, 11345−11350.
(42) Kamerlin, S. C. L.; Sharma, P. K.; Prasad, R. B.; Warshel, A. *Q. Rev. Biophys.* **2013**, *46*, 1−132.
(43) Plotnikov, N. V.; Prasad, B. R.; Chakrabarty, S.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **2013**, *117*, 12807−19.
(44) Prasad, B. R.; Plotnikov, N. V.; Warshel, A. *J. Phys. Chem. B* **2013**, *117*, 153−163.
(45) Glaves, R.; Mathias, G.; Marx, D. *J. Am. Chem. Soc.* **2012**, *134*, 6995−7000.
(46) Akola, J.; Jones, R. O. *J. Phys. Chem. B* **2003**, *107*, 11774−11783.
(47) Harrison, C. B.; Schulten, K. *J. Chem. Theory Comput.* **2012**, *8*, 2328−2335.
(48) Mildvan, A. S. *Proteins: Struct.,Funct., Bioinf.* **1997**, *29*, 401−416.
(49) Admiraal, S. J.; Herschlag, D. *Chem. Biol.* **1995**, *2*, 729−739.
(50) Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.
(51) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. *J. Phys. Chem. B* **2006**, *110*, 3533−3539.
(52) Bylaska, E.; Weare, J. Q.; Weare, J. H. *J. Chem. Phys.* **2013**, *139*, 074114.