A Surface-Accelerated String Method for Locating Minimum Free Energy Paths

Timothy J. Giese, Şölen Ekesan, Erika McCarthy, Yujun Tao, and Darrin M. York*

Laboratory for Biomolecular Simulation Research, Institute for Quantitative Biomedicine and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA

E-mail: Darrin.York@rutgers.edu

ABSTRACT: We present a surface-accelerated string method (SASM) to efficiently optimize low-dimensional reaction pathways from the sampling performed with expensive quantum mechanical/molecular mechanical (QM/MM) Hamiltonians. The SASM accelerates the convergence of the path by using the aggregate sampling obtained from the current and previous string iterations, whereas approaches like the string method in collective variables (SMCV) or the modified string method (MSM) update the path only from the sampling obtained from the current iteration. Furthermore, the SASM decouples the number of images used to perform sampling from the number of synthetic images used to represent the path. The path is optimized on the current best estimate of the free energy surface obtained from all available sampling, and the proposed set of new simulations are not restricted to be located along the optimized path. Instead, the umbrella potential placement is chosen to extend the range of the free energy surface and improve the quality of the free energy estimates near the path. In this manner, the SASM is shown to improve the exploration for a minimum free energy pathway in regions where the free energy surface is relatively flat. Furthermore, it improves the quality of the free energy profile when the string is discretized

*To whom correspondence should be addressed

with too few images. We compare the SASM, SMCV, and MSM using 3 QM/MM applications: a ribozyme methyltransferase reaction using 2 reaction coordinates, the $2'$-O-transphosphorylation reaction of Hammerhead ribozyme using 3 reaction coordinates, and a tautomeric reaction in B-DNA using 5 reaction coordinates. We show that SASM converges the paths using roughly 3 times less sampling than the SMCV and MSM methods. All three algorithms have been implemented in the FE-ToolKit package made freely available.

1 Introduction

The ability to model chemical reactions in the condensed phase[1] using molecular simulations has far-reaching implications to the study of catalysis in biological systems.[2,3] Advances in fast, accurate quantum mechanical force fields[4,5] and machine learning models[6–11] have greatly extended the scope of applications that can be routinely addressed. Nonetheless, simulations of complex reaction pathways remain computationally intensive, and ongoing development of new methods to improve the robustness and computational cost are important.

Reaction mechanisms can be characterized by calculating a free energy surface in a set of relevant reaction coordinates, the determination of the minimum free energy profile (MFEP) through the surface, and identification of key stationary points along the MFEP. Many methods for calculating free energy surfaces have been developed. These approaches can be categorized as:[12] methods which analyze equilibrium statistics obtained from umbrella sampling,[13–16] methods which analyze nonequilibrium statistics[17–19] based on the work of Jarzynski,[20] and methods that integrate auxiliary degrees of freedom, such as λ-dynamics[21–24] and metadynamics.[25,26] Similarly, there are two general approaches for locating a minimum free energy path.[27] The first approach is to sample the reaction over a wide range of reaction coordinate values to obtain a relatively complete picture of the free energy surface through which a path can be optimized. The second, and more cost-effective, approach is to use a chain-of-states method, such as nudged elastic band[28] or the string method,[29,30] to direct the sampling toward the MFEP, thereby reducing the amount of effort spent simulating irrelevant, high-energy regions of the free energy surface.

Many variations of the string method[31–38] have been developed that are capable of being applied to large-scale problems, like protein folding.[33,34] These applications often describe the path using a large number of reaction coordinates,[39] direct comparison of Cartesian coordinates,[29] path collective variables,[27,40] the use of the hills method,[25,41] or machine learning techniques.[42] Although string method development was originally motivated by the desire to use many reaction coordinates,[31,32,35] many examples can be found of their use in quantum mechanical/molecular mechanical (QM/MM) applications involving only a few reaction coordinates.[43–49] String methods, such as the one presented in Ref. 35, are particularly appealing because it is performed with standard umbrella sampling with harmonic biasing potentials, which are widely supported across simulation packages. Because QM/MM sampling is very costly, the present work seeks to optimize the string method described in Ref. 35 specifically for cases involving QM/MM simulations with a few reaction coordinates. The new method reduces the number of string iterations required to reach convergence because it uses the current estimate of the unbiased free energy to accelerate the exploration of flat regions of the surface. In this respect, the new method draws inspiration from ideas behind the metadynamics approach;[25,26] however, the new method only requires sampling obtained using standard harmonic biasing potentials.

We describe a new surface-accelerated string method (SASM) and compare it to two similar algorithms: the string method in collective variables[31,32] (SMCV), and the modified string method[35] (MSM). We have implemented all 3 of these methods in the ndfes software[46] freely distributed within the FE-ToolKit package.[50] The FE-ToolKit package has also been incorporated in the open source AmberTools simulation suite.[51] There are several key differences between the SASM and related string methods. First, the SASM is a hybrid of the two approaches for locating a MFEP (chain-of-states method versus calculation of a multidimensional free energy surface). Whereas the SMCV and MSM update the path from the sampling obtained in the most recent string iteration, the SASM optimizes the path on the current estimate of the multidimensional free energy surface calculated from the aggregate sampling of all string iterations. Second, the SASM decouples the number of images used to represent the path from the number of simulated images. The SMCV and MSM methods construct a new path by fitting a curve that interpolates a set of discrete control points obtained from a corresponding number of simulated images; therefore, if there was an insufficient number of images, the path may cut corners. By decoupling the representation of the path from the number of simulated images, the level of detail used to describe the path is not limited by the number of simulations. Third, unlike the SMCV and MSM, the SASM does not require the images to be simulated along the current estimate of the path. We take advantage of this by introducing alternating stages of "exploration" and "refinement" steps. The exploration steps propose new simulations offset from the path in the direction that the path is moving, and the refinement steps place simulations along the path in a manner that improves the phase space overlap.

We compare the progress of the string optimizations using the SMCV, MSM, and SASM with respect to the number of simulations per string, the sampling per simulation, and the spline representation of the path (either piecewise linear or Akima spline paths) in 3 applications. The first application uses 2 reaction coordinates to describe a ribozyme undergoing a methyl transfer reaction (MTR1)[52–54] (PDB ID 7V9E). The second application uses 3 reaction coordinates to model the 2′-*O*-transphosphorylation reaction of Hammerhead ribozyme (HHr)[55] (PDB ID 2OEU). The third application uses 5 reaction coordinates to optimize a tautomeric reaction pathway in B-DNA (PDB ID 113D).[56] Schematics of the 3 systems are shown in Figure 1. We demonstrate that the SASM converges the MFEP faster than the SMCV and MSM when we vary the amount of sampling. The SASM avoids artifacts that can occur in the path "reparametrization step" of the SMCV and MSM. Finally, we show that the SASM method will sample the path in an efficient manner that achieves good overlap between the biased simulations when the number of simulations is reduced.

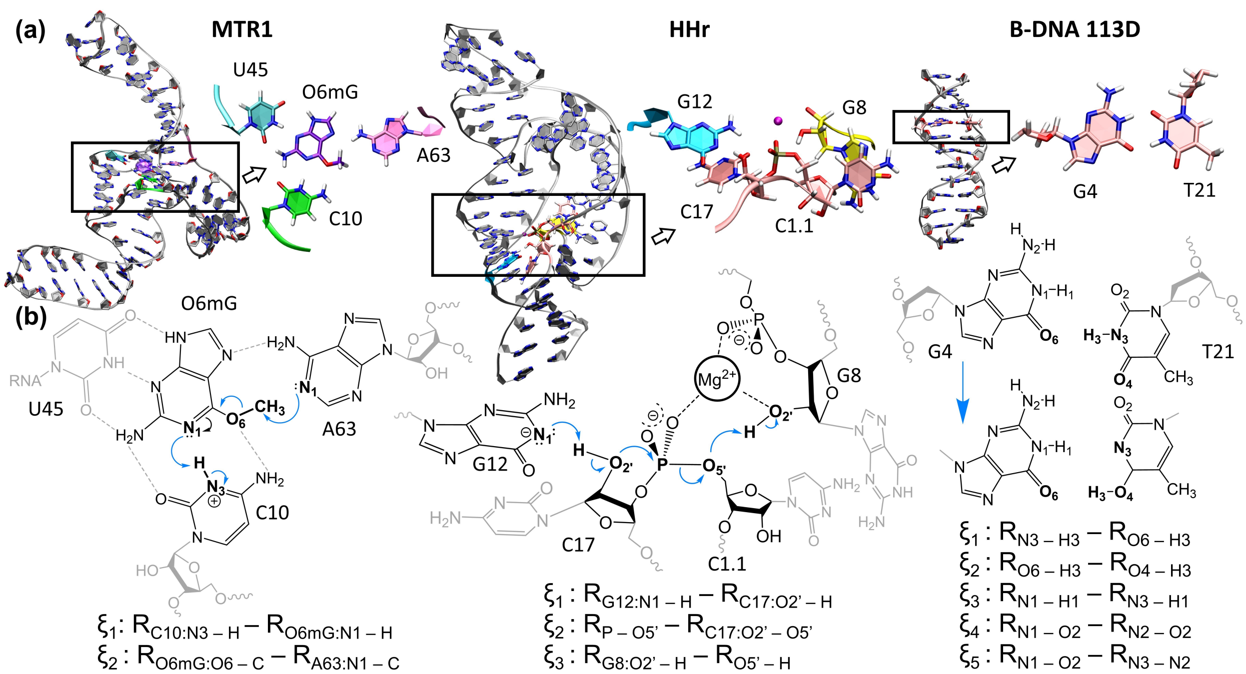

Figure 1: (a) The MTR1 ribozyme, HHr ribozyme, and a B-DNA with a GT wobble pair examined in this work. The rectangles highlight the active site region. (b) The reaction mechanisms and reaction coordinates. The B-DNA system is a tautomer reaction which transfers the T21 N3 proton to position O4 and reorganization of the G:T hydrogen bond network. The shown atomic configurations correspond to the reactant state. The black and gray atoms denote the QM region and nearby MM atoms, respectively.

2 Methods

2.1 String Method in Collective Variables

This section summarizes the SMCV method, which was originally described in Refs. 31 and 32. Let $\mathbf{x}$ and $\mathbf{q}(\mathbf{x})$ be the 3N array of atomic positions and N_{dim} reaction coordinate values, respectively. Umbrella sampling is performed at N_{img} images along the path using a biased potential energy function, U_n.

$$U_n(\mathbf{x}) = U(\mathbf{x}) + W(\mathbf{q}(\mathbf{x}); \mathbf{k}_n, \mathbf{q}_n) \tag{1}$$

Image n is biased by a potential W that is parametrized by N_{dim} harmonic force constants $\mathbf{k}_n$ and equilibrium positions $\mathbf{q}_n$. In other words, N_{dim} is the size of the reduced dimensional space of reaction coordinates.

$$W(\mathbf{q}(\mathbf{x}); \mathbf{k}_n, \mathbf{q}_n) = \frac{1}{2}\sum_{d=1}^{N_{\mathrm{dim}}} k_{nd}\left(q_d(\mathbf{x}) - q_{nd}\right)^2 \tag{2}$$

The algorithm for calculating the SMCV consists of the following steps.

1. Sample each of the N_{img} images along the path for some amount of time, Δt. The images differ by their biasing potentials, which center the harmonic potentials at discrete points along the current estimate of the path, $\mathbf{q}_n$.

2. Analyze the sampling to update (evolve) the reaction coordinate values, $\mathbf{q}_{\mathrm{c},n}$. The "control points", $\mathbf{q}_{\mathrm{c},n}$ are discrete estimates along the new path, but they do not necessarily uniformly discretize it. The calculation of the control points is sometimes called the "evolution step".

3. Construct a parametric curve that interpolates the control points. The parametric curve is the new estimate of the path.

4. Uniformly discretize the parametric curve to obtain the biasing potential centers for the next iteration. The construction of a new curve and its discretization is sometimes called the "reparametrization step".

The SMCV evolution step is given by eq. 3, where $q_{nd}^{(k)}$ is the value of the reaction coordinate d of image n at string iteration k, and $q_{\mathrm{c},nd}^{(k+1)}$ is a control point used to define the parametric curve in string iteration $k+1$, discussed in the next section. Each image is simulated for a length of time Δt, and $\langle \cdot \rangle_{\mathbf{k}_n,\mathbf{q}_n}$ denotes a time average obtained from image n.

$$q_{\mathrm{c},nd}^{(k+1)} = q_{nd}^{(k)} - \frac{\Delta t}{\gamma} \sum_{d'=1}^{N_{\mathrm{dim}}} M_{dd'}(\mathbf{k}_n^{(k)}, \mathbf{q}_n^{(k)}) \nabla G_{d'}(\mathbf{k}_n^{(k)}, \mathbf{q}_n^{(k)}) \tag{3}$$

$\nabla G_d(\mathbf{k}_n^{(k)}, \mathbf{q}_n^{(k)})$ approximates the free energy gradient about the point $\mathbf{q}_n^{(k)}$ in dimension d.

$$\nabla G_d(\mathbf{k}_n, \mathbf{q}_n) = -\left\langle \frac{\partial W(\mathbf{q}(\mathbf{x}); \mathbf{k}_n, \mathbf{q}_n)}{\partial q_{nd}} \right\rangle_{\mathbf{k}_n,\mathbf{q}_n} \tag{4}$$

$\mathbf{M}$ is closely related to a product of mass weighted Wilson B-matrices;[57] that is to say, $\nabla_a q$ is the gradient of the reaction coordinate value with respect to the atomic positions of atom a and m_a is an atomic mass.

$$M_{dd'}(\mathbf{k}_n, \mathbf{q}_n) = \left\langle \sum_{a=1}^{N} \frac{\nabla_a q_d(\mathbf{x}) \cdot \nabla_a q_{d'}(\mathbf{x})}{m_a} \right\rangle_{\mathbf{k}_n,\mathbf{q}_n} \tag{5}$$

γ is a friction coefficient, a parameter of the method. The numerical stability of the SMCV critically depends on the ratio $\Delta t \gamma^{-1}$. In Ref. 32, it was found that the method was stable when choosing $\gamma = 1500$ ps^{-1} when $\Delta t = 20$ fs. In the present work, we adjust γ to maintain this same ratio when Δt is varied. The construction of parametric curves and their uniform discretization are described in the next section.

2.2 Parametric Curves and the Reparametrization Step

We represent a continuous path as a parametric curve of reaction coordinates, $\mathbf{q}(p)$, where $p \in [0, 1]$ is a progress variable such that $p = 0$ and $p = 1$ denote two ends of the path. In other words, the path at string iteration k, $\mathbf{q}^{(k)}(p)$ is an array of N_{dim} one-dimensional splines that are chosen such that each spline interpolates the N_{img} control points, $q_{\mathrm{c},nd}^{(k)}$ located at a common set of progress control values, $p_{c,n}^{(k)}$.

$$q_d^{(k)}(p) \equiv q_d(p; \mathbf{q}_{c,d}^{(k)}, \mathbf{p}_c^{(k)}) \tag{6}$$

In the context of the SMCV (or similar string methods), the control points are the new estimates of the reaction coordinates after the evolution step (eq. 3). In some cases, one may choose to reduce the numerical noise in the path by first applying a smoothing procedure, in which case the control points are the reaction coordinate values after smoothing. The results presented in this work use a smoothing algorithm implemented in the ndfes software when the parametric curve is modeled with Akima spline functions,[58] but we do not apply smoothing to the control points when using piecewise linear paths. The details of the smoothing algorithm are described in the Supporting Information.

The parametric curve depends on the progress control values, which are interpreted as fractional arc lengths through the curve. If the path is a piecewise linear function connecting the control points, then the progress control values can be calculated from the Euclidean distance between adjacent points, as shown in eq. 7.

$$p_{c,n}^{(k)} \approx \begin{cases} 0, & \text{if } n = 1 \\ \frac{\sum_{m=2}^{n} \sqrt{\sum_{d=1}^{N_{\text{dim}}} \left(q_{md}^{(k)} - q_{m-1,d}^{(k)}\right)^2}}{\sum_{m=2}^{N_{\text{img}}} \sqrt{\sum_{d=1}^{N_{\text{dim}}} \left(q_{md}^{(k)} - q_{m-1,d}^{(k)}\right)^2}}, & \text{otherwise} \end{cases} \tag{7}$$

Alternatively, if the parametric curve is a set of Akima spline functions[58] (or any smooth interpolating function), then eq. 7 is only an approximation of the fractional arc lengths. Accurate values of the progress control values can be found by iteratively solving eq. 8, initiated from eq. 7.

$$p_{c,n}^{(k,i+1)} = \frac{\int_0^{p_{c,n}^{(k,i)}} \sqrt{\sum_{d=1}^{N_{\text{dim}}} \left(\frac{\partial q_d(p; \mathbf{q}_{c,d}^{(k)}, \mathbf{p}_c^{(k,i)})}{\partial p}\right)^2} dp}{\int_0^1 \sqrt{\sum_{d=1}^{N_{\text{dim}}} \left(\frac{\partial q_d(p; \mathbf{q}_{c,d}^{(k)}, \mathbf{p}_c^{(k,i)})}{\partial p}\right)^2} dp} \tag{8}$$

We terminate the iterative solution when $\sum_{m=1}^{N_{\text{img}}} \left(p_{c,m}^{(k,i+1)} - p_{c,m}^{(k,i)}\right)^2 < 10^{-16}$.

Given the the parametric spline representation of the path, the uniformly discretized images for string iteration $k+1$ is shown in eq. 9, where $p_n = (n-1)/(N_{\mathrm{img}}-1)$.

$$q_{nd}^{(k+1)} = q_d(p_n; \mathbf{q}_{\mathrm{c},d}^{(k+1)}, \mathbf{p}_{\mathrm{c}}^{(k+1)}) \tag{9}$$

2.3 Modified String Method

The modified string method (MSM) was originally presented in Ref. 35; it differs from the SMCV only by replacing the evolution step (eq. 3) with eq. 10.

$$q_{\mathrm{c},nd}^{(k+1)} = \langle q_d(\mathbf{x}) \rangle_{\mathbf{k}_n^{(k)},\mathbf{q}_n^{(k)}} \tag{10}$$

In other words, the control points for the new path are the mean observed positions of the reaction coordinates from the simulations performed along the current path. Upon finding the control points, a new parametric curve is fit. The curve is uniformly discretized to define the new positions of the biasing potentials.

2.4 Surface-Accelerated String Method

The surface-accelerated string method (SASM) constructs a N_{dim} dimensional free energy surface from the available sampling and optimizes a path on that surface. A decision is then made to place a new set of simulations, which may or may not be along the optimized path. When the new simulations are placed along the path, we refer to it as a “refinement step”. Alternatively, we allow for “exploration steps” that offset the simulations from the path in the direction that the path is moving.

The algorithm for calculating the SASM consists of the following steps.

1. Sample each of the N_{img} images for some amount of time.

2. Construct a N_{dim} dimensional unbiased free energy surface by analyzing the aggregate sampling produced from all simulations and string iterations. This is the best estimate of the free

energy surface from the available sampling. The N_{dim} dimensional space is discretized into bins, and the free energy value and the number of observed samples in each bin are tabulated.

3. Create a smooth representation of the free energy surface, such that the free energy value and gradient can be readily computed at any point in the space of reaction coordinates.

4. Use the free energy surface to optimize a MFEP in the space of reaction coordinates. This optimization procedure does not involve the generation of additional sampling. Instead, the optimization is performed on a fixed free energy surface using a series of "synthetic string iterations", described below.

5. If the current iteration is an even integer, then place the new simulations along the path. If the current iteration is an odd integer, then allow the new set of simulations to be displaced from the path by some amount in the direction that the path is moving.

The unbiased free energy can be calculated using established methods, such as the variational free energy profile method,[46,59,60] the multistate Bennett acceptance ratio (MBAR) method,[61] or the unbinned weighted histogram (UWHAM) method.[62,63] As discussed in Ref. 46, a smooth representation of the free energy surface can be made using one of many methods, including the use of Cardinal B-Splines,[64] radial basis functions,[65,66] or Gaussian process regression.[67] In the present work, we calculate the free energy surface by solving the MBAR/UWHAM equations to reweight the biased sampling. The samples are histogrammed, and the free energy of each bin is tabulated. We use fourth-order Cardinal B-splines to represent the surface as a smooth function. A mathematical description of the B-spline interpolation is provided in the Supporting Information for completeness. The free energy values are formally defined only in those regions whose histogram bins are occupied by at least one sample. In practice, we exclude all bins containing fewer than 10 samples because their free energy values are often unreliable.

To optimize a path on a fixed free energy surface, we adapt the MSM by replacing eq. 10 with eq. 11, where $F(\mathbf{q})$ is the value of the unbiased free energy at $\mathbf{q}$.

$$\mathbf{q}_{\mathrm{c},n}^{(k+1,s+1)} = \arg\min_{\mathbf{q}} \left\{ F(\mathbf{q}) + W(\mathbf{q}; \bar{\mathbf{k}}, \mathbf{q}_n^{(k+1,s)}) \right\} \tag{11}$$

$$\mathbf{q}_n^{(k+1,s+1)} = \mathbf{q}(p_n; \mathbf{q}_{\mathrm{c}}^{(k+1,s+1)}, \mathbf{p}_{\mathrm{c}}^{(k+1,s+1)}) \tag{12}$$

The $q_{\mathrm{c},nd}^{(k,s)}$ values are the control points of the synthetic images used to describe the path. Specifically, n indexes the synthetic image, d indexes the dimension, k is the string iteration, and s is the synthetic iteration. The number of synthetic images, N_{simg}, does not need to be the same as the number of images used to perform explicit simulations, N_{img}. In the present work, we use $N_{\mathrm{simg}} = 100$ to describe the path. The $\bar{\mathbf{k}}$ quantity appearing in eq. 11 is an $N_{\mathrm{dim}} \times 1$ array of force constants, chosen to be the average of the N_{img} simulations; that is, $\bar{k}_d = N_{\mathrm{img}}^{-1} \sum_{n=1}^{N_{\mathrm{img}}} k_{nd}^{(k)}$. Equation 11 is analogous to the MSM, but instead of performing a biased simulation of $3N$ atomic coordinates to obtain the reaction coordinate distribution means, one performs a minimization directly on a biased N_{dim} free energy surface. In other words, eq. 11 is a synthetic iteration that allows us to repeatedly propagate the string without producing additional sampling. The path is optimized with N_{siter} iterations (or until convergence is sufficiently met), such that $\mathbf{q}_{\mathrm{opt}}^{(k+1)}(p) \equiv \mathbf{q}(p; \mathbf{q}_{\mathrm{c}}^{(k+1,N_{\mathrm{siter}})}, \mathbf{p}_{\mathrm{c}}^{(k+1,N_{\mathrm{siter}})})$ is the best estimate of the MFEP from the available sampling. The optimized synthetic control points also serve as the initial guess for the path in the next string iteration: $\mathbf{q}_{\mathrm{c},n}^{(k+1,0)} = \mathbf{q}_{\mathrm{c},n}^{(k,N_{\mathrm{siter}})}$.

By optimizing the MFEP on the current estimate of the free energy surface, the N_{img} real images are no longer responsible for describing the path. Instead, their sole responsibility is to provide sampling to improve the the quality and range of the free energy surface. For this purpose, the SASM evolution step (eq. 13) includes two modifications relative to a simple uniform discretization.

$$\mathbf{q}_n^{(k+1)} = \mathbf{q}_{\mathrm{opt}}^{(k+1)}(p_n + \Delta p^{(k+1)}) + \Delta\mathbf{q}_n^{(k+1)} \tag{13}$$

The first modification is a shifting of the progress control points when discretizing the parametric curve,

$$p_n + \Delta p^{(k+1)} = \begin{cases} p_{n,0}, & \text{if } N(p_{n,0}) = 0 \\ p_{n,-1/3}, & \text{else if } N(p_{n,-1/3}) = 0 \\ p_{n,+1/3}, & \text{else if } N(p_{n,+1/3}) = 0 \\ p_{n,0}, & \text{else if } \mathrm{mod}(k,3) = 0 \\ p_{n,-1/3}, & \text{else if } \mathrm{mod}(k,3) = 1 \\ p_{n,+1/3}, & \text{else if } \mathrm{mod}(k,3) = 2 \end{cases} \tag{14}$$

where $p_{n,0}$ is a uniform discretization, and $p_{n,+1/3}$ and $p_{n,-1/3}$ shift the discretization by 1/3 of the distance to a neighboring image.

$$p_{n,x} = \max\left(0, \min\left(1, \frac{n-1+x}{N_{\mathrm{img}}-1}\right)\right) \tag{15}$$

$N(p)$ is the number of samples that have been observed at the point $\mathbf{q}_{\mathrm{opt}}^{(k+1)}(p)$. In other words, the first 3 cases in eq. 14 check whether there are gaps in the sampling along the path. If there is a gap, then sampling at that position is prioritized. The last 3 cases in eq. 14 are a schedule that is followed when no gaps in the sampling are detected. The schedule alternates between these displacements during the course of the string optimization to help ensure that one obtains sufficient sampling along the path in the event that one underestimated an appropriate value of N_{img}.

The second modification is the introduction of $\Delta\mathbf{q}_n^{(k+1)}$ which displaces the image in the direction of the path's movement.

$$\Delta\mathbf{q}_n^{(k+1)} = \begin{cases} 0, & \text{if } \mathrm{mod}(k,4) = 0 \\ [1 - \delta_{0,N(p_n+\Delta p^{(k+1)})}]h_1 \frac{\Delta\mathbf{q}_{\mathrm{min},n}^{(k+1)}}{|\mathbf{q}_{\mathrm{min},n}^{(k+1)}|}, & \text{if } \mathrm{mod}(k,4) = 1 \\ 0, & \text{if } \mathrm{mod}(k,4) = 2 \\ [1 - \delta_{0,N(p_n+\Delta p^{(k+1)})}]h_2 \frac{\Delta\mathbf{q}_{\mathrm{min},n}^{(k+1)}}{|\mathbf{q}_{\mathrm{min},n}^{(k+1)}|}, & \text{if } \mathrm{mod}(k,4) = 3 \end{cases} \tag{16}$$

We refer to $\Delta\mathbf{q}_n^{(k+1)} = 0$ as a refinement step that places the simulations along the path, and the other cases are exploration steps intended to better describe the free energy surface in the vicinity of path in the direction of its movement. The exploration steps accelerate the evolution of the string through flat areas of the free energy surface. The leading Kronecker delta function causes the exploration step to be skipped if a gap in the sampling was previously detected in eq. 14. The exploration direction is determined from the difference between the optimized paths of the current and previous iterations.

$$\Delta\mathbf{q}_{\min,n}^{(k+1)} = \mathbf{q}_{\text{opt}}^{(k+1)}(p_n + \Delta p^{(k+1)}) - \mathbf{q}_{\text{opt}}^{(k)}(p^*) \tag{17}$$

The value of p^* is the point on the previous path that is closest to the point $p_n + \Delta p^{(k+1)}$ on the current path.

$$p^* = \arg\min_p \left|\mathbf{q}_{\text{opt}}^{(k+1)}(p_n + \Delta p^{(k+1)}) - \mathbf{q}_{\text{opt}}^{(k)}(p)\right|^2 \tag{18}$$

The h_m values are the magnitude of the displacement, where w_d is a width assigned to each dimension. In the present work, we use $w_d = 0.15$ Å for all dimensions, which is also the width of the histogram bins used to construct the free energy surface.

$$h_m = \min\left(mw_1 \frac{|\Delta\mathbf{q}_{\min,n}^{(k+1)}|}{|\Delta q_{\min,n,1}^{(k+1)}|}, \cdots, mw_{N_{\text{dim}}} \frac{|\Delta\mathbf{q}_{\min,n}^{(k+1)}|}{|\Delta q_{\min,n,N_{\text{dim}}}^{(k+1)}|}\right) \tag{19}$$

If one imagines the point at $p_n + \Delta p^{(k+1)}$ as being located at the corner of a voxel, then eq. 19 can be interpreted as choosing the magnitude to be the maximum displacement that does not exceed the range of m voxels.

2.5 Computational Details

All QM/MM simulations in this work were performed with the sander molecular dynamics software[51] using a 1 fs integration time step. The SHAKE algorithm[68] was used to fix MM bonds involving hydrogen. The covalent bonds at the QM/MM boundary were capped with the hydrogen link-atom approach.[69,70] Electrostatics were calculated with the particle mesh Ewald method[71–73] adapted for use within semiempirical QM/MM simulations[74,75] using tinfoil boundary conditions[76,77] a 1 Å^3 reciprocal space grid, and 10 Å real space cutoffs. The Lennard-Jones interactions were similarly calculated to 10 Å and a long-range tail correction was included to account for the interactions beyond the cutoff.[78]

The MTR1 ribozyme (PDB ID 7V9E[54]) consists of 2,207 atoms with a net 66− charge. The ribozyme was solvated with 18,250 TIP4P/Ew waters, 113 sodium ions, and 47 chlorine ions in a truncated octahedron with real space lattice vectors of length 90.2 Å resulting in 75,367 particles and an ion concentration of 140 mM. The ff99OL3 RNA force field[79] and Joung and Cheatham[80] monovalent ion parameters have been used. Details regarding the preparation and equilibration of this system have already been reported elsewhere.[81] In brief, the pressure and temperature were equilibrated for 50 ns with the MM force field potential to maintain 1 atm and 298 K in the isothermal-isobaric ensemble using the Berendsen barostat[82] and Langevin thermostat[83] with a collision frequency of 5 ps^{-1}. At this point, the MM force field was replaced with the DFTB3 QM/MM potential using the "3ob" parameter set.[84] The QM region consists of 48 atoms with net 1+ charge, as illustrated in Figure 1. A QM/MM simulation of the reactant state was equilibrated for 12.5 ps in the canonical ensemble at 298 K. The DFTB3 QM/MM umbrella production sampling was similarly performed at constant temperature with 200 kcal mol^{-1} Å^{-2} force constants on the two reaction coordinates describing the transfer of a proton, $\xi_1 = R_{\text{C10:N3-H}} - R_{\text{O6mG:N1-H}}$, and methyl group, $\xi_2 = R_{\text{O6mG:O6-C}} - R_{\text{A63:N1-C}}$ as visualized in Figure 1.

The HHr ribozyme (PDB ID 2OEU[55]) consists of 2,020 atoms with a net 62− charge. The ribozyme was solvated with 13,319 TIP4P/Ew waters, 5 magnesium ions (replacing the crystal structure manganese ions), 86 sodium ions, and 34 chlorine ions in a truncated octahedron with real space lattice vectors of length 81.7 Å resulting in 55,421 particles and an ion concentration of 140 mM. The ff99OL3 RNA force field,[79] Joung and Cheatham[80] monovalent ion, and Li-Merz[85] 12-6-4 divalent ion parameters with Panteva[86,87] corrections, which ensure balanced interactions between metal ions and nucleic acids, have been used. Full details of the preparation and equilibration of this system has been reported elsewhere.[88] In brief, the pressure and temperature were equilibrated for 100 ns with the MM force field potential to maintain 1 atm and 298 K in the isothermal-isobaric ensemble. The MM force field was replaced with the AM1/d QM/MM potential.[89] The QM region consists of 85 atoms with net 1− charge. The QM region is illustrated in Figure 1; for clarity, the Mg^{2+} and the 4 waters directly coordinating the Mg^{2+} were including in the QM region. A QM/MM simulation of the reactant state was equilibrated for 50 ps in the canonical ensemble at 298 K. All AM1/d QM/MM umbrella production sampling was performed at constant temperature with 200 kcal mol^{-1} Å^{-2} force constants on the three reaction coordinates describing the proton transfer from the nucleophile to the general base, $\xi_1 = R_{\text{G12:N1-H}} - R_{\text{C17:O2'-H}}$, and phosphoryl transfer, $\xi_2 = R_{\text{P-O5'}} - R_{\text{C17:O2'-O5'}}$. and the proton transfer from the general acid to the leaving group, $\xi_3 = R_{\text{G8:O2'-H}} - R_{\text{O5'-H}}$.

The B-DNA sequence (PDB ID 113D)[56] consists of 762 atoms with a net 22− charge. The ribozyme was solvated with 5,151 TIP4P/Ew waters, 35 sodium ions, and 13 chlorine ions in a truncated octahedron with real space lattice vectors of length 59.3 Å resulting in 21,414 particles and an ion concentration of 140 mM. The system was modeled with the OL5 DNA force field[90] and Joung and Cheatham[80] monovalent ion parameter set. The system was prepared by minimizing the solvent environment and hydrogen positions while restraining the DNA heavy atoms, followed by a gradual heating of the system from 0 to 298 K over the course of 300 ps in the NVT ensemble, and the system density was equilibrated at 1 atm for 8 ns in the NPT ensemble. The MM force field was replaced with the AM1/d QM/MM potential,[89] where the QM region (the G4 and T21 nucleobases depicted in Figure 1) consists of 31 atoms with net neutral charge. A QM/MM simulation of the reactant state was equilibrated for 50 ps in the canonical ensemble at 298 K. All AM1/d QM/MM umbrella production sampling was performed at constant temperature with 200 kcal mol^{-1} Å^{-2} force constants for each of the reaction coordinates listed in Figure 1.

To start any string method, one must first construct a series of structures to be used as the initial guess. For the MTR1 reaction, we consider two initial guesses: a concerted guess that uniformly discretizes a line connecting the approximate position of the reactant state $\boldsymbol{\xi}_{\text{react.}} = (-1.4\,\text{Å}, -2.5\,\text{Å})$ to the product state $\boldsymbol{\xi}_{\text{prod.}} = (1.4\,\text{Å}, 2.5\,\text{Å})$, and a stepwise guess that uniformly discretizes a piecewise linear path connecting the reactant state, approximate intermediate state $\boldsymbol{\xi}_{\text{inter.}} = (1.4\,\text{Å},\ -2.5\,\text{Å})$, and the product state. For the HHR reaction, the initial guess discretizes a linear transformation between the approximate reactant state $\boldsymbol{\xi}_{\text{react.}} = (-1\,\text{Å},\ -2\,\text{Å},\ -1\,\text{Å})$ to the approximate product state $\boldsymbol{\xi}_{\text{prod.}} = (1\,\text{Å},\ 2\,\text{Å},\ 1\,\text{Å})$. Similarly, the initial guess for the B-DNA tautomer reaction discretizes a linear transformation between the reactant $\boldsymbol{\xi}_{\text{react.}}$=$(-0.86\,\text{Å}, -0.60\,\text{Å}, -2.0\,\text{Å}, -0.76\,\text{Å}, -2.6\,\text{Å})$ and product states $\boldsymbol{\xi}_{\text{react.}}$=$(0.67\,\text{Å}, 0.78\,\text{Å}, -0.82\,\text{Å}, 0.84\,\text{Å}, 0.22\,\text{Å})$. The atomic coordinates were generated from a sequence of short (200 fs) simulations that restart each image from the final structure of the previous image. After this scan was completed, each image was independently equilibrated for an additional 4 ps. The final coordinates from these equilibrations became the starting structures to initiate the string method.

The SMCV, MSM, and SASM were performed multiple times while varying the number of images and length of production sampling. The MTR1 simulations performed for 4 ps/image and 500 fs/image saved 400 samples/image and 250 samples/image, respectively. The HHr simulations performed for 625 fs/image and 312 fs/image saved 125 samples/image and 156 samples/image, respectively. The B-DNA simulations were performed for 1 ps/image and 200 samples/image were saved. In all cases, we analyze only the last 75% of saved samples when solving the MBAR/UWHAM equations.

3 Results and Discussion

Here we compare the SMCV, MSM, and SASM string methods using three reactive chemical systems having varying number of reaction coordinates. 1. A 2D example of an artificially engineered methyltransferase ribozyme (MTR1)[52] that catalyzes the methylation of a target adenine. 2. A 3D example of a naturally occurring hammerhead ribozyme (HHr)[55] that catalyzes site-specific RNA self-cleavage. 3. A 5D example of tautomerization in dG·dT wobble pairs that lead to misincorporation during replication.[91]

3.1 MTR1 catalytic mechanism

Evolutionary theories based on an RNA world[92,93] presumably would require RNA molecules to catalyze C-C and C-N bond formation essential for nucleic acid synthesis and early metabolic transformations. There are no known naturally occurring examples of RNA enzymes that have this ability. Recently, a methyltransferase ribozyme (MTR1) has been evolved *in vitro*[52] that binds O^6-methylguanine and catalyzes the methylation of a target adenine (A63) at the N1 position[53,94] (Figure 1a). Computational enzymology studies performed by our group,[81] in collaboration with Huang, Lilley and co-workers,[54] revealed a surprising sophisticated mechanism that involves a protonated cytosine residue that acts as an acid in order to facilitate site-specific C-N bond formation, broadening the range of known RNA-catalyzed chemistry and further demonstrating versatility of RNA catalysis.[95] In the computational study, we employed an early version of the string method and found it to be slowly convergent, making it extremely costly to perform *ab initio* QM/MM simulations. Hence, we use this as our first test system for developing improved string methods with accelerated convergence.

Figure 2 uses the MTR1 reaction to compare the progress of the SMCV, MSM, and SASM at string iterations 5, 15, 30, and 50. Each optimization was performed twice, starting from concerted and stepwise initial paths. Each string iteration samples 32 images for 4 ps/image (128 ps/iteration). The free energy surface is a best estimate made from the aggregate sampling of all iterations obtained from the 3 methods (38.4 ns of aggregate sampling). The black line is a reference MFEP, optimized on the aggregate free energy surface. The SMCV and MSM paths are Akima splines fit to the 32 evolved images, whereas the SASM paths are Akima splines fit to 100 synthetic images.

The three methods approach the MFEP at different rates. The SMCV and MSM make good progress during the first 15 iterations, but their progress stalls as they near the MFEP. This is due to the free energy surface becoming relatively flat near the MFEP. In contrast, the SASM gets closer to the MFEP at iteration 5 than the SMCV or MSM do at iteration 50. By placing the simulations around the path, the SASM is capable of exploring flat surfaces more efficiently.

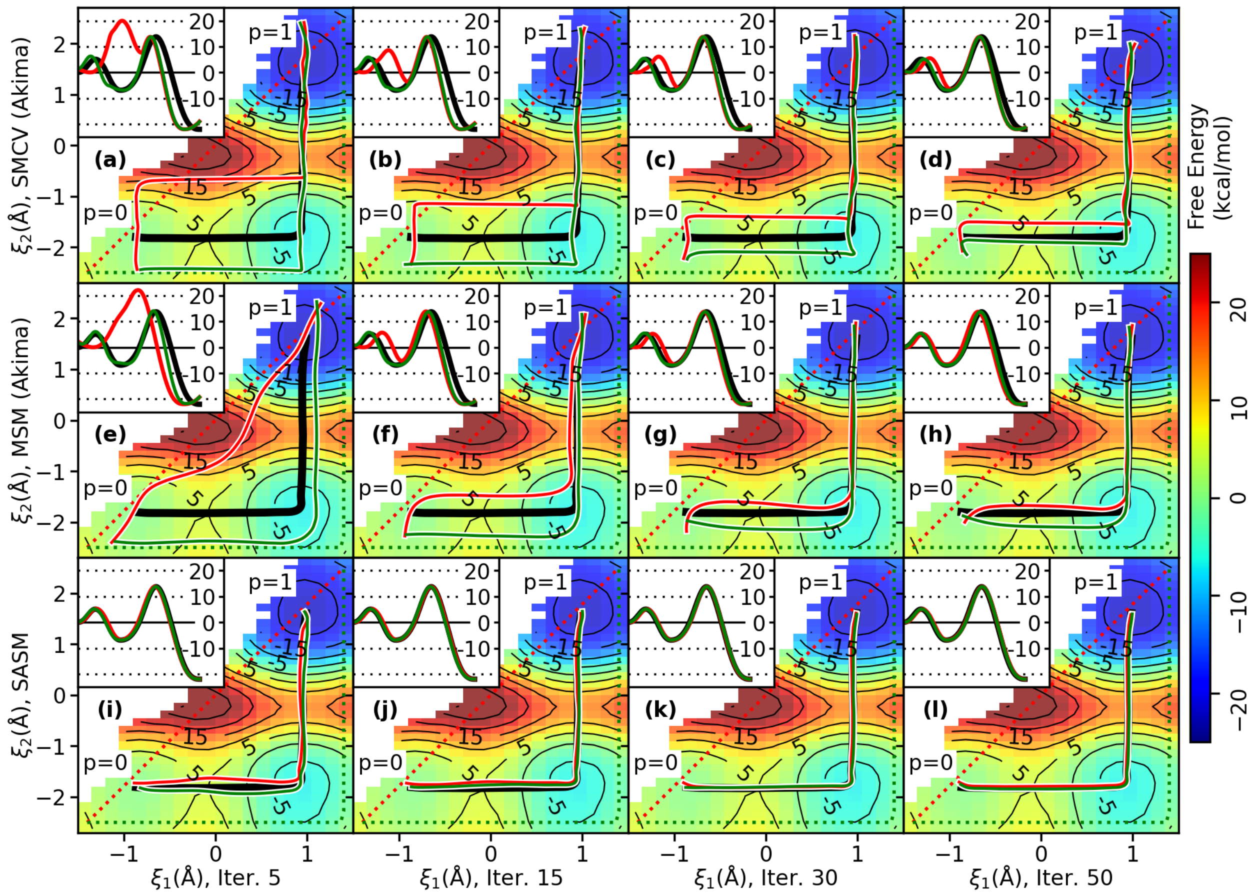

Figure 2: Progress of the string methods at several iterations of the MTR1 reaction starting from concerted (red lines) and stepwise (green lines) initial guess paths. The initial guesses are dashed lines. The colored areas are the best estimate of the free energy surface, calculated from the aggregate sampling produced by all string methods. The black line is the MFEP optimized on the best estimate of the surface. The insets are the reference free energy values along the paths (kcal/mol).

To test whether the conclusions drawn from Figure 2 are sensitive to the simulation time scale (time/image), we re-performed the string methods using only 500 fs/image of sampling. The resulting comparison (Supporting Information Figure S1) is nearly indistinguishable from Figure 2.

Figure 3 compares the string methods using fewer images and different spline representations of the path. The optimizations start from a concerted path, and each iteration samples 8 images for 4 ps/image. The colored areas are the current estimate of the free energy surface from the sampling produced by the current and previous iterations. The red line is the estimate of the path after the evolution step. The "x" marks are the proposed set of umbrella potential locations. The black line is the reference path shown in Figure 2. The insets display the free energy along the paths; the red line is the free energy of the current path from the available sampling, and the black line is the reference free energy along the reference path (made from 38.4 ns of aggregate sampling). The rows labeled "Akima" and "Linear" construct the parametric curve from Akima splines and piecewise linear functions, respectively. Whereas the SMCV and MSM paths are limited to 8 control points, the SASM path is constructed from Akima splines which interpolate 100 synthetic images optimized on the free energy surface.

Figure 3 illustrates that the SMCV and MSM methods are sensitive to the parametric form of the path when only a few (e.g., 8) images are simulated. Although the SMCV and MSM methods properly evolve the control points, both methods encounter artifacts within the reparametrization step when the path is modeled with Akima splines. The artifacts encountered by the SMCV are quite severe; reparametrization of the curve causes some images to be propagated in a direction away from the MFEP (Figure 3a-b), and the converged path differs significantly from the reference path (Figure 3d). The MSM similarly encounters artifacts between iterations 15 and 50, and it converges to an incorrect path (Figure 3l). The SMCV and MSM methods do approach the correct MFEP when using piecewise linear curves, however (see Figure 3h and p). The SASM does not exhibit artifacts using Akima splines because it is parametrized to 100 synthetic control points rather than 8 control points. By using more control points to define the path, the SASM also avoids corner cutting, which can be observed when using piecewise linear paths; for example, see the

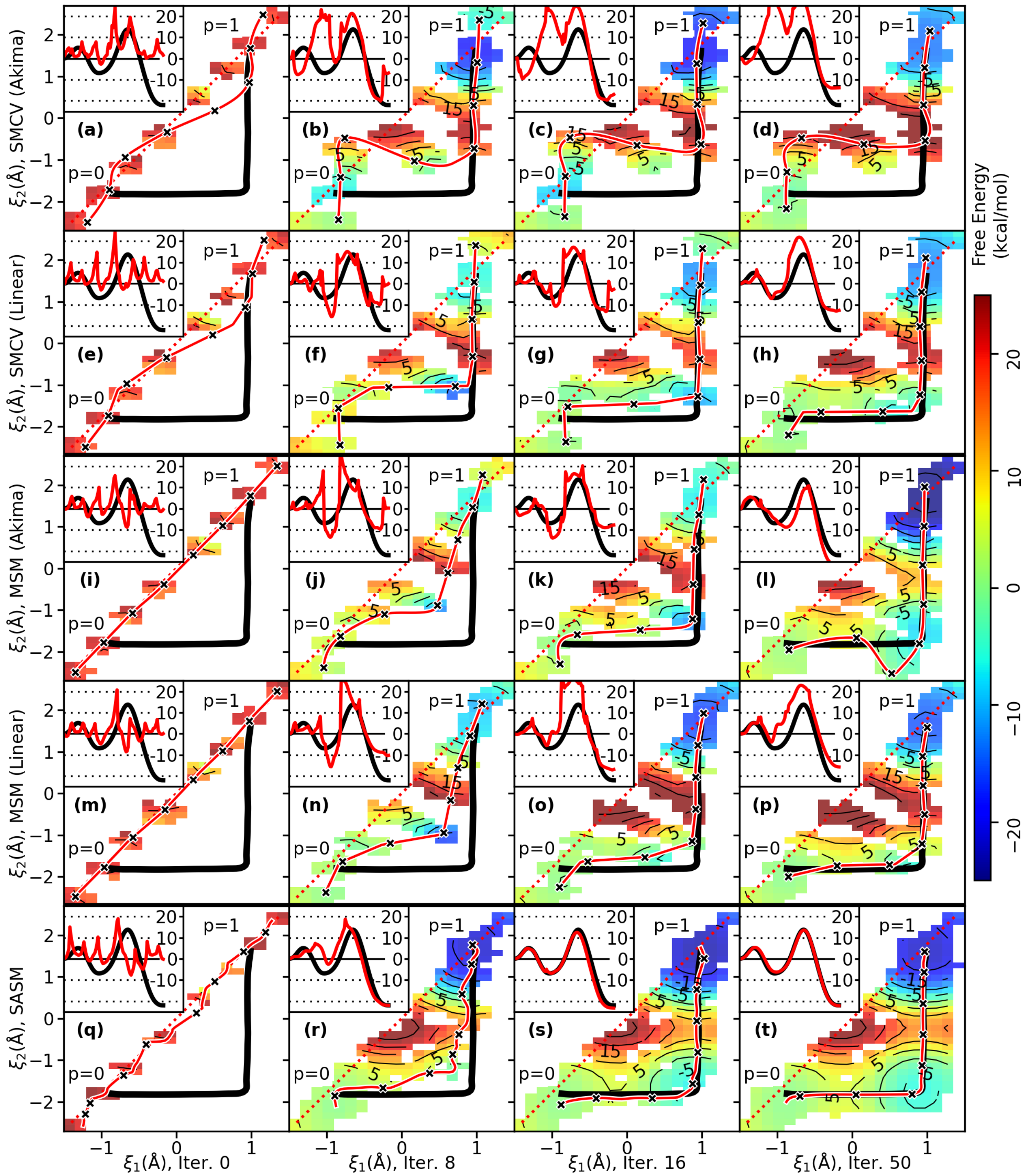

Figure 3: String iterations of MTR1 from a concerted (linear) initial guess (dashed red line). Each string is composed of 8 images, and each image is sampled for 4 ps. The solid red line is the current string, and the black "x" marks the next set of 8 simulations. The black line is a reference pathway, and the insets compare the current estimate of the free energy profile to the reference profile (kcal/mol). Parts a-d and e-h are the SMCV method using Akima and piecewise linear splines, respectively. Parts i-l and m-p similarly compare the MSM method. Parts q-t are the SASM method with 100 synthetic images.

intermediate state in Figure 3p.

When only 8 images are simulated, the progress of the SASM is modestly better than SMCV (Linear) and MSM (Linear); however, the SASM does a much better job at producing samples to analyze the free energy surface. As can be seen in the insets of Figures 3a-p, the limited number of images causes the SMCV and MSM to produce sampling that does not well overlap, resulting in noisy free energy profiles. In contrast, the SASM evolution step shifts the progress values to improve the sampling between the set of uniformly discretized points, and the exploration steps provide sampling around the path. Consequently, the SASM free energy profile after 15 iterations reproduces the reference profile very well (Figure 3s). In fact, the SASM profile after 15 iterations is better than the SMCV and MSM profiles after 50 iterations. The SASM placement algorithm attempts to fill the gaps in the sampling, which is easiest to observe in Figure 3q. After sampling the initial guess, the optimized path remains similar to the initial guess because all areas of the surface which have not been sampled are assumed to have a high free energy. The first 3 cases in eq. 14 propose new simulations in the unoccupied regions along the path.

3.2 HHr Mechanism

The hammerhead ribozyme (HHr)[55,96–98] is a metal-dependent small endonucleolytic self-cleaving RNA that has been extensively studied experimentally[99–101,101,102] and computationally,[103–109] and is an archetype model for RNA catalysis. The active site adopts an L-platform/L-scaffold architecture[110] with an L-pocket guanine residue that forms a divalent metal ion binding site enabling electrostatic interactions[88] to facilitate the reaction. The $2'O$-transphosphorylation mechanism can be described by three reaction coordinates, illustrated in Figure 1b.

Figure 4 extends the comparisons to the 3 dimensional HHr transphosphorylation reaction profiles. The dashed line is the concerted initial guess, and the remaining lines are the paths at a series of string iterations. The reference path shown in each image is provided as a visual aid. The reference path is the SASM MFEP after 300 iterations using 32 images and 625 fs/image of sampling. In other words, it is the MFEP optimized on the 3 dimensional surface produced from the analysis

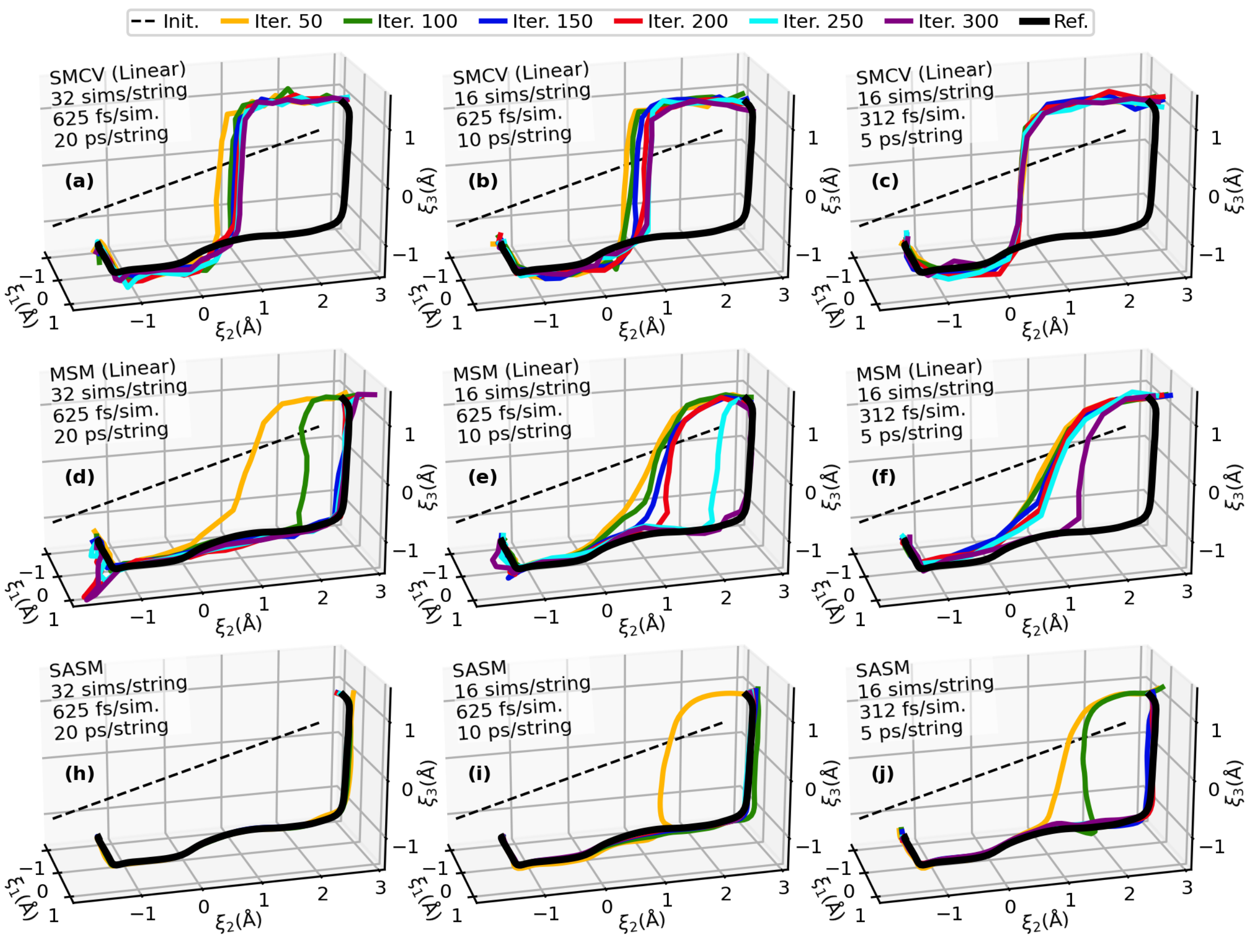

Figure 4: String iterations of HHr from a linear initial guess (dashed black line).

of 6 ns of aggregate sampling. The string methods were performed multiple times by varying the number of images and the amount of sampling. Each column of Figure 4 successively halves the amount of sampling per string.

All of the string methods predict that the first stage of the reaction transfers a proton (the ξ_1 coordinate) from the O2′ to the N1 position of the G12 general base (Figure 1b). The more interesting part of the comparison is the behavior of the paths in the ξ_2-ξ_3 plane, where ξ_2 is the phosphoryl transfer coordinate, and ξ_3 measures the proton transfer between the O5′ and the G8 general acid. The SMCV fails to locate the MFEP after 300 iterations, although it is possible that it may find the MFEP if iterated further.

The MSM locates the MFEP, but it requires many iterations. The MSM requires 150 iterations to locate the MFEP when performed with 32 images and sampled for 625 fs/image (Figure 4d). This corresponds to 3 ns of aggregate sampling. When the number of images is reduced to 16 (Figure 4e), the amount of sampling per iteration is reduced, but the MSM now requires 300 iterations (3 ns of aggregate sampling) to locate the MFEP. Further reduction in the amount of sampling requires more than 300 MSM iterations (Figure 4f). Notice that the progress of the MSM in Figures 4e-f does not significantly change from iterations 50 to 150, which would likely cause one to incorrectly believe that the path has converged. In fact, previous application of the MSM to the HHr reaction incorrectly concluded that the mechanism was concerted because of this behavior,[46] whereas the extended iterations presented in Figure 4 suggest that the MFEP is stepwise. The fundamental reason why MSM progress stalls is because the free energy gradient in the directions perpendicular to the path are quite small (Figure 5). The qualitative similarity between the MSM path at iteration 50 to the paths produced by SMCV is suggestive that the SMCV fails for a similar reason.

In this application, the SASM requires 3 times fewer iterations than MSM to reach convergence when using the same amount of sampling. Only 50 SASM iterations are required to converge the path using 32 images (Figure 4h) in comparison to 150 MSM iterations. When the number of images is reduced to 16 (Figure 4i), convergence is reached after 100 SASM iterations in comparison

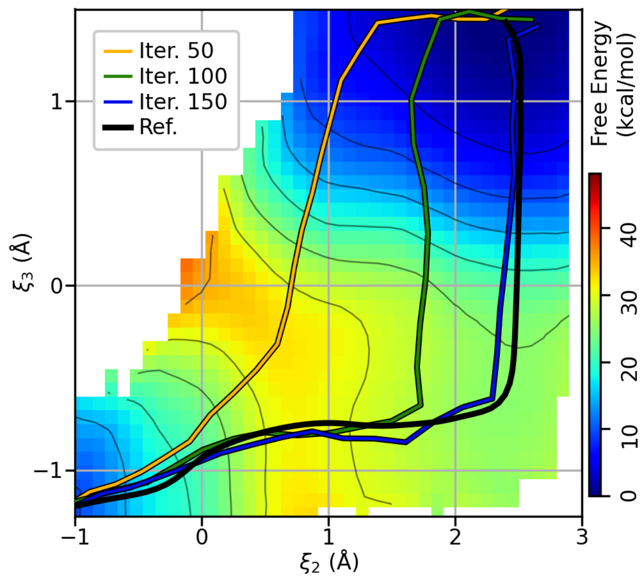

Figure 5: Two dimensional projection of the HHr free energy surface defined by the $\xi_1 = 0.95$ Å plane. The colored lines are the MSM paths at iterations 50, 100, and 150, and the black line is the SASM reference curve after 300 iterations, as shown in Figure 4d. The colored areas are the free energy values calculated from the aggregate sampling produced from 300 MSM and 300 SASM iterations (12 ns of sampling).

to 300 MSM iterations. The SASM requires fewer iterations because the synthetic string optimizations performed within the SASM can evolve the path to the fringes of the aggregate sampling, and the exploration steps increase the range of the free energy surface that can be used.

3.3 B-DNA G·T Wobble Tautomer Reaction

Rare tautomeric forms of nucleobases can cause Watson-Crick-like (WC-like) mispairs in DNA, and in turn lead to disease.[111] In the WC model, nucleobase pairs are in their "keto" form,[112] rather then "imino" or "enol" form. Recently, tautomerization has been reported for a G-T wobble pair ($G^{\text{enol}}T/U \leftrightarrow GT^{\text{enol}}/U^{\text{enol}}$) in B-DNA detected by NMR[91,113,114] and subsequently studied computationally.[115] This tautomerization reaction can be described by 5 reaction coordinates (Figure 1c).

A pathway in 5D cannot be visualized in the same way as the 2D and 3D systems, hence Figure 6a illustrates the convergence of the MFEP for the B-DNA tautomer reaction by calculating the root mean square deviation (RMSD) between the current estimate of the path and the initial guess using the 5 reaction coordinates. The SASM RMSD plateaus at 50 iterations, whereas the

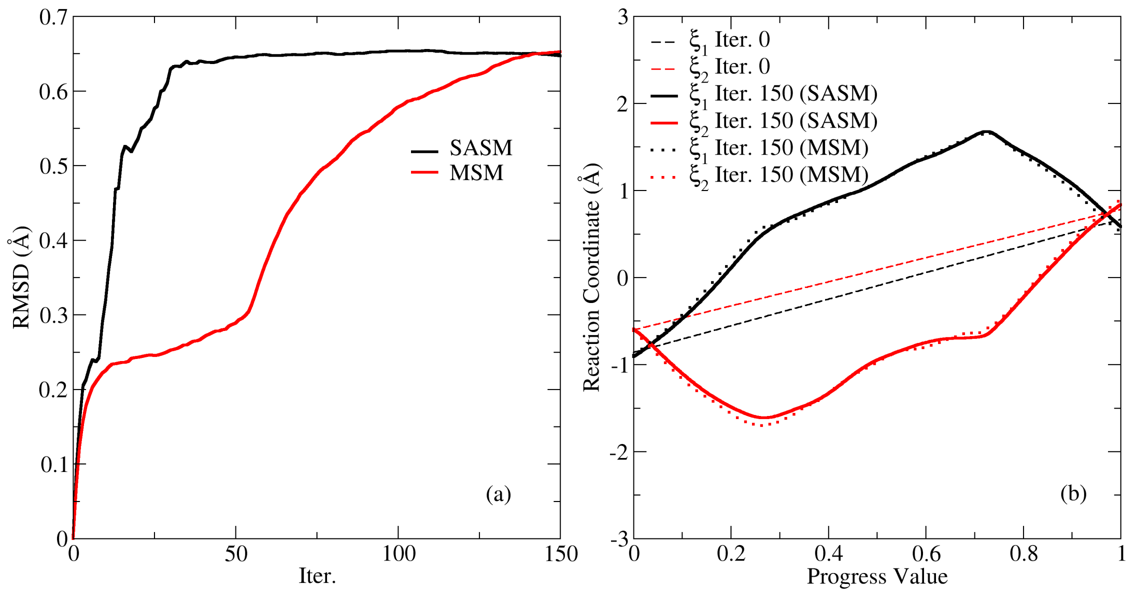

Figure 6: Convergence of the path describing the wGT→GT* tautomeric reaction in B-DNA. (a) The root mean square deviation of the 5 reaction coordinates relative to the concerted (linear) initial guess. (b) The ξ_1 and ξ_2 reaction coordinates along the initial and final pathways produced by the MSM and SASM. These two coordinates describe the proton transfer between N3-O6 and O6-O4, respectively.

MSM requires 150 iterations to reach the same RMSD. Figure 6b shows the initial and final profiles of the $\xi_1 = R_{\text{N3-H3}} - R_{\text{O6-H3}}$ and $\xi_2 = R_{\text{O6-H3}} - R_{\text{O4-H3}}$ reaction coordinates. The other 3 reaction coordinates are excluded from the figure to improve legibility. The initial path directly transfers the proton from N3 to the O4 position. The optimized paths instead transfer the proton from the N3 to O6 while shifting the hydrogen bond pattern of the G:T basepair. This is followed by the transfer of the proton from O6 to the O4 position. The SASM and MSM both produce the same path after 150 iterations. In summary, this application shows that the SASM can be extended to 5 dimensions and it can converge the path in fewer iterations than the MSM.

3.4 Computational Cost

Table 1 compares the CPU resources needed to perform the string methods on the HHr system with 32 images and 625 fs/image of sampling and the B-DNA system with 32 images and 1 ps/image of sampling. The measurements were performed on a single core of an Intel Xeon E5-2630 v3 processor, and the software was compiled with GCC 9.2.1. The timings can be decomposed into

Table 1: The number of CPU days required to perform MSM and SASM on the HHr and B-DNA systems for the specified number of iterations. Iteration 0 is the simulation and analysis of the initial path. Bold entries denote converged paths.

	HHr			B-DNA		
Iter.	T_{MSM}	T_{SASM}	$\frac{T_{\text{SASM}}}{T_{\text{MSM}}}$	T_{MSM}	T_{SASM}	$\frac{T_{\text{SASM}}}{T_{\text{MSM}}}$
0	0.51	0.51	1.00	0.19	0.19	1.00
10	5.56	5.56	1.00	2.07	2.08	1.00
25	13.14	13.16	1.00	4.89	4.96	1.01
50	25.77	**25.96**	1.01	9.60	**10.07**	1.05
100	51.03	52.48	1.03	19.01	22.57	1.19
150	**76.30**	81.12	1.06	**28.42**	40.20	1.41

two components: the resources used to perform the QM/MM simulations T_{sim}, and the resources used to perform the evolution step T_{evo}.

$$\begin{aligned} T(k) &= \sum_{k'=0}^{k} T_{\text{sim}}(k') + T_{\text{evo}}(k') \\ &=(k+1)T_{\text{sim}} + \sum_{k'=0}^{k} T_{\text{evo}}(k') \end{aligned} \tag{20}$$

The MSM times only include the resources used to perform the QM/MM simulations; the string evolution step (eq. 10) requires a negligible amount of effort, $T_{\text{evo}}(k) \approx 0$. The cost of performing MSM for the HHr and B-DNA systems are given by eqs. 21 and 22, respectively.

$$T_{\text{MSM}}^{\text{HHr}}(k) = (k+1)(32 \text{ images})\left(\frac{0.625 \text{ ps}}{\text{image}}\right)\left(\frac{1 \text{ CPU day}}{39.6 \text{ ps}}\right) \tag{21}$$

$$T_{\text{MSM}}^{\text{B-DNA}}(k) = (k+1)(32 \text{ images})\left(\frac{1 \text{ ps}}{\text{image}}\right)\left(\frac{1 \text{ CPU day}}{170 \text{ ps}}\right) \tag{22}$$

The SASM timings also include the cost of the evolution step, which is further decomposed into the resources used to solve the MBAR/UWHAM equations, T_{MBAR}, and the cost of performing an optimization on the resulting free energy surface, T_{opt}.

$$T_{\text{SASM}}(k) = T_{\text{MSM}}(k) + \sum_{k'=0}^{k} T_{\text{MBAR}}(k') + T_{\text{opt}}(k') \tag{23}$$

The solution of the MBAR/UWHAM equations formally scales $O(N_{\mathrm{dim}} N_{\mathrm{samples}} N_{\mathrm{states}})$, where N_{samples} is the number of samples to be reweighted and N_{states} is the number of states. The dimensionality does not vary with string iteration, and N_{samples} and N_{states} are both proportional to the number of iterations, leading to $T_{\mathrm{MBAR}}(k) \approx A(k+1)^2$, where A is coefficient fit to the observed times. This coefficient is 0.359 s and 0.876 s for the HHr and B-DNA systems, respectively. The quadratic dependence of $T_{\mathrm{MBAR}}(k)$ means that the aggregate cost for performing k string iterations scales cubically. The cost of performing the optimization formally scales $O(o^{N_{\mathrm{dim}}} N_{\mathrm{siter}} N_{\mathrm{simg}})$, where o is the order of the Cardinal B-spline, N_{siter} is the number of synthetic iterations, and N_{simg} is the number of synthetic images used to describe the path. These quantities are independent of string iteration, so $T_{\mathrm{opt}}(k) \approx B$, where B is 0.7 s and 19.5 s for the HHr and B-DNA systems, respectively.

The timings listed in Table 1 suggest that the SASM increases the computational cost by 1%-5% relative to the MSM for the first 50 iterations. This small increase is reflected in the high computational cost of performing QM/MM sampling. Although the SASM is more costly, it converges in fewer iterations. The SASM reduces the cost of converging the path by factors of 2.9 and 2.8 for the HHr and B-DNA systems, respectively. The SASM becomes increasingly expensive with respect to the number of iterations due to the cost of solving the MBAR/UWHAM equations from the aggregate sampling. To prevent the method from becoming too costly at high iterations, one could limit the analysis to the samples produced from the most recent 50 iterations, for example.

Figure 7 uses the first 50 SASM iterations of the B-DNA system to illustrate the cost of T_{MBAR} and T_{opt} as the number of reaction coordinates is varied. Although the sampling was performed with 5 reaction coordinates, we can measure T_{MBAR} and T_{opt} by ignoring 1-or-more of the reaction coordinates during the analysis. As previously discussed, the solution of the MBAR/UWHAM equations has a linear dependence on N_{dim}, and B-spline evaluations of the free energy surface has an exponential dependence on N_{dim}. The dashed lines are linear and exponential fits to the observed times. Figure 7 demonstrates that the SASM quickly becomes impractical when using more than 6 reaction coordinates due to high cost of evaluating the free energy of a high-dimensional surface. Another aspect to consider is that each added dimension further subdivides the samples into differ-

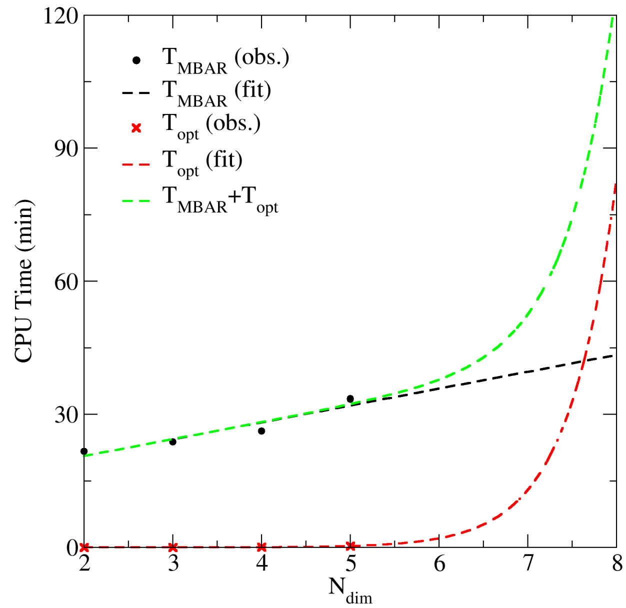

Figure 7: CPU time required to perform MBAR analysis (T_{MBAR}) and path optimization (T_{opt}) on the resulting free energy surface. The observed times were measured using the B-DNA sampling at iteration 50. The black and red dashed lines are linear and exponential fits to T_{MBAR} and T_{opt}, respectively.

ent histogram bins. For a fixed amount of sampling, each subdivision reduces the average number of samples per occupied bin and thus increases the uncertainty of the free energy in that region. For these reasons, we do not view the SASM as a replacement for the MSM when a large number of reaction coordinates is needed. Instead, the SASM is a complimentary tool specifically tailored to accelerate the convergence of low-dimensional pathways frequently encountered in QM/MM applications. In these situations, the added expense of generating free energy surfaces and optimizing paths from the available sampling is worthwhile to reduce the number of costly QM/MM evaluations.

4 Conclusions

We applied the SMCV, MSM, and SASM methods to QM/MM sampling of the MTR1, HHr, and B-DNA G·T mispair systems. These applications served to compare the behavior and performance of the string methods using 2, 3, and 5 reaction coordinates. The SASM is a new method developed in this work that is robust and has performance advantages for systems up to approximately 6-dimensions ($N_{\mathrm{dim}} \leq 6$). Rather than propagating the path from the sampling produced by the most recent set of images, the SASM uses the aggregate sampling from all string iterations. The sampling is used to construct the current best-estimate of a multidimensional free energy surface, and a MFEP is optimized on the surface. Consequently, the simulated images are no longer responsible for describing the parametric form of the path; their sole responsibility is to improve the quality and range of the sampling used to estimate the surface. The SASM exploits this freedom by alternating between "exploration" and "refinement" steps to rapidly traverse flat regions of the free energy surface.

Overall, the SMCV, MSM, and SASM methods are capable of converging to the correct MFEP if the right control parameters are found. In some cases, spline artifacts can be observed with the SMCV and MSM when only a few images (e.g., 8) are used. The SASM is found to be more robust, and it often requires approximately 1/3 of the string iterations to converge the MFEP. Analysis of computational timings indicate that the SASM increases the computational cost per string iteration by 5% or less relative to the MSM, but this is more than offset by requiring fewer iterations to reach convergence. The computational cost of representing a free energy surface with more than 6 reaction coordinates quickly becomes prohibitive; therefore, the SASM is not a blanket replacement for the MSM. Rather, it is a valuable tool that can be used to considerably accelerate convergence in QM/MM applications using a modest number of reaction coordinates.

Supporting Information Available

Descriptions of the procedures used to smooth the control points and Cardinal B-spline evaluation of the free energy, and a comparison of MTR1 profiles generated from reduced sampling. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Acknowledgments

The authors are grateful for financial support provided by the National Institutes of Health (No. GM62248) and the National Science Foundation (CSSI Frameworks Grant No. 2209718). Computational resources were provided by the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey; the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296 (supercomputer Expanse at SDSC through allocation CHE190067); and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin, URL: http://www.tacc.utexas.edu (supercomputer Frontera through allocation CHE20002). .

References

(1) Giese, T. J.; York, D. M. Quantum mechanical force fields for condensed phase molecular simulations. *J. Phys. Condens. Matter* **2017**, *29*, 383002.

(2) Gao, J.; Truhlar, D. G. Quantum Mechanical Methods for Enzyme Kinetics. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467–505.

(3) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. How enzymes work: Analysis by modern rate theory and computer simulations. *Science* **2004**, *303*, 186–195.

(4) Giese, T. J.; Huang, M.; Chen, H.; York, D. M. Recent Advances toward a General Purpose Linear-Scaling Quantum Force Field. *Acc. Chem. Res.* **2014**, *47*, 2812–20.

(5) Gao, J.; Truhlar, D. G.; Wang, Y.; Mazack, M. J.; Löffler, P.; Provorse, M. R.; Rehak, P. Explicit polarization: A quantum mechanical framework for developing next generation force fields. *Acc. Chem. Res.* **2014**, *47*, 2837–2845.

(6) Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chem. Engl.* **2017**, *56*, 12828–12840.

(7) Meuwly, M. Machine Learning for Chemical Reactions. *Chem. Rev.* **2021**, *121*, 10218–10239.

(8) Zeng, J.; Giese, T. J.; Ekesan, Ş.; York, D. M. Development of Range-Corrected Deep Learning Potentials for Fast, Accurate Quantum Mechanical/Molecular Mechanical Simulations of Chemical Reactions in Solution. *J. Chem. Theory Comput.* **2021**, *17*, 6993–7009.

(9) Giese, T. J.; Zeng, J.; Ekesan, Ş.; York, D. M. Combined QM/MM, Machine Learning Path Integral Approach to Compute Free Energy Profiles and Kinetic Isotope Effects in RNA Cleavage Reactions. *J. Chem. Theory Comput.* **2022**, *18*, 4304–4317.

(10) Pan, X.; Yang, J.; Van, R.; Epifanovsky, E.; Ho, J.; Huang, J.; Pu, J.; Mei, Y.; Nam, K.; Shao, Y. Machine-Learning-Assisted Free Energy Simulation of Solution-Phase and Enzyme Reactions. *J. Chem. Theory Comput.* **2021**, *17*, 5745–5758.

(11) Snyder, R.; Kim, B.; Pan, X.; Shao, Y.; Pu, J. Bridging semiempirical and *ab initio* QM/MM potentials by Gaussian process regression and its sparse variants for free energy simulation. *J. Chem. Phys.* **2023**, *159*, 054107.

(12) Kästner, J. Umbrella sampling. *WIREs Comput. Mol. Sci.* **2011**, *1*, 932–942.

(13) Torrie, G. M.; Valleau, J. P. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett.* **1974**, *28*, 578–581.

(14) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(15) McDonald, I. R.; Singer, K. Machine Calculation of Thermodynamic Properties of a Simple Fluid at Supercritical Temperatures. *J. Chem. Phys.* **1967**, *47*, 4766–4772.

(16) McDonald, I. R.; Singer, K. Examination of the Adequacy of the 12–6 Potential for Liquid Argon by Means of Monte Carlo Calculations. *J. Chem. Phys.* **1969**, *50*, 2308–2315.

(17) Adib, A. B. Free energy surfaces from nonequilibrium processes without work measurement. *J. Chem. Phys.* **2006**, *124*, 144111.

(18) Hummer, G. Fast-growth thermodynamic integration: Error and efficiency analysis. *J. Chem. Phys.* **2001**, *114*, 7330–7337.

(19) Zuckerman, D. M.; Woolf, T. B. Theory of a Systematic Computational Error in Free Energy Differences. *Phys. Rev. Lett.* **2002**, *89*, 180602.

(20) Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

(21) Knight, J. L.; Brooks, C. L., 3rd Lambda-dynamics free energy simulation methods. *J. Comput. Chem.* **2009**, *30*, 1692–1700.

(22) Kong, X.; Brooks, III, C. L. λ-dynamics: A new approach to free energy calculations. *J. Chem. Phys.* **1996**, *105*, 2414–2423.

(23) Liu, Z.; Berne, B. J. Method for accelerating chain folding and mixing. *J. Chem. Phys.* **1993**, *99*, 6071–6077.

(24) Tidor, B. Simulated annealing on free energy surfaces by a combined molecular dynamics and Monte Carlo approach. *J. Phys. Chem.* **1993**, *97*, 1069–1073.

(25) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.

(26) White, A. D.; Dama, J. F.; Voth, G. A. Designing Free Energy Surfaces That Match Experimental Data with Metadynamics. *J. Chem. Theory Comput.* **2015**, *11*, 2451–2460.

(27) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From to A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, 054103–054112.

(28) Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. 1998.

(29) E, W.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B* **2002**, *66*, 052301.

(30) E, W.; Ren, W.; Vanden-Eijnden, E. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.* **2007**, *126*, 164103.

(31) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **2006**, *125*, 024106.

(32) Ovchinnikov, V.; Karplus, M.; Vanden-Eijnden, E. Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI. *J. Chem. Phys.* **2011**, *134*, 085103–085128.

(33) Ovchinnikov, V.; Karplus, M. Investigations of α-helix $\leftrightarrow$ β-sheet transition pathways in a miniprotein using the finite-temperature string method. *J. Chem. Phys.* **2014**, *140*, 175103–175121.

(34) Vanden-Eijnden, E.; Venturoli, M. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* **2009**, *130*, 194103.

(35) Rosta, E.; Nowotny, M.; Yang, W.; Hummer, G. Catalytic Mechanism of RNA Backbone Cleavage by Ribonuclease H from Quantum Mechanics/Molecular mechanics simulations. *J. Am. Chem. Soc.* **2011**, *133*, 8934–8941.

(36) Khavrutskii, I. V.; Arora, K.; Brooks 3rd, C. L. Harmonic Fourier beads method for studying rare events on rugged energy surfaces. *J. Chem. Phys.* **2006**, *125*, 174108.

(37) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A growing string method for determining transition states: comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004**, *120*, 7877–7886.

(38) Pan, A. C.; Sezer, D.; Roux, B. Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B* **2008**, *112*, 3432–3440.

(39) Matsunaga, Y.; Komuro, Y.; Kobayashi, C.; Jung, J.; Mori, T.; Sugita, Y. Dimensionality of Collective Variables for Describing Conformational Changes of a Multi-Domain Protein. *J. Phys. Chem. Lett.* **2016**, *7*, 1446–1451.

(40) Kulshrestha, A.; Punnathanam, S. N.; Ayappa, K. G. Finite temperature string method with umbrella sampling using path collective variables: application to secondary structure change in a protein. *Soft Matter* **2022**, *18*, 7593–7603.

(41) Ensing, B.; Laio, A.; Parrinello, M.; Klein, M. L. A Recipe for the Computation of the Free Energy Barrier and the Lowest Free Energy Path of Concerted Reactions. *J. Phys. Chem. B* **2005**, *109*, 6676–6687.

(42) Badaoui, M.; Buigues, P. J.; Berta, D.; Mandana, G. M.; Gu, H.; Földes, T.; Dickson, C. J.; Hornak, V.; Kato, M.; Molteni, C.; Parsons, S.; Rosta, E. Combined Free-Energy Calculation and Machine Learning Methods for Understanding Ligand Unbinding Kinetics. *J. Chem. Theory Comput.* **2022**, *18*, 2543–2555.

(43) Ganguly, A.; Thaplyal, P.; Rosta, E.; Bevilacqua, P. C.; Hammes-Schiffer, S. Quantum Mechanical/Molecular Mechanical Free Energy Simulations of the Self-Cleavage Reaction in the Hepatitis Delta Virus Ribozyme. *J. Am. Chem. Soc.* **2014**, *136*, 1483–1496.

(44) Zhong, J.; Reinhardt, C. R.; Hammes-Schiffer, S. Role of Water in Proton-Coupled Electron Transfer between Tyrosine and Cysteine in Ribonucleotide Reductase. *J. Am. Chem. Soc.* **2022**, *144*, 7208–7214.

(45) Zhong, J.; Reinhardt, C. R.; Hammes-Schiffer, S. Direct Proton-Coupled Electron Transfer between Interfacial Tyrosines in Ribonucleotide Reductase. *J. Am. Chem. Soc.* **2023**, *145*, 4784–4790.

(46) Giese, T. J.; Ekesan, Ş.; York, D. M. Extension of the Variational Free Energy Profile and Multistate Bennett Acceptance Ratio Methods for High-Dimensional Potential of Mean Force Profile Analysis. *J. Phys. Chem. A* **2021**, *125*, 4216–4232.

(47) Reinhardt, C. R.; Sayfutyarova, E. R.; Zhong, J.; Hammes-Schiffer, S. Glutamate Mediates Proton-Coupled Electron Transfer Between Tyrosines 730 and 731 in *Escherichia coli* Ribonucleotide Reductase. *J. Am. Chem. Soc.* **2021**, *143*, 6054–6059.

(48) Li, P.; Soudackov, A. V.; Hammes-Schiffer, S. Fundamental Insights into Proton-Coupled Electron Transfer in Soybean Lipoxygenase from Quantum Mechanical/Molecular Mechanical Free Energy Simulations. *J. Am. Chem. Soc.* **2018**, *140*, 3068–3076.

(49) Ganguly, A.; Boulanger, E.; Thiel, W. Importance of MM Polarization in QM/MM Studies of Enzymatic Reactions: Assessment of the QM/MM Drude Oscillator Model. *J. Chem. Theory Comput.* **2017**, *13*, 2954–2961.

(50) Giese, T. J.; York, D. M. FE-ToolKit: The free energy analysis toolkit. https://gitlab.com/RutgersLBSR/fe-toolkit.

(51) Case, D. A.; Aktulga, H. M.; Belfon, K.; Cerutti, D. S.; Cisneros, G. A.; Cruzeiro, V. W. D.; Forouzesh, N.; Giese, T. J.; Götz, A. W.; Gohlke, H.; Izadi, S.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kurtzman, T.; Lee, T.-S.; Li, P.; Liu, J.; Luchko, T.; Luo, R.; Manathunga, M.; Machado, M. R.; Nguyen, H. M.; O'Hearn, K. A.; Onufriev, A. V.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Risheh, A.; Schott-Verdugo, S.; Shajan, A.; Swails, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Zhang, S.; Zhao, S.; Zhu, Q.; Cheatham 3rd, T. E.; Roe, D. R.; Roitberg, A.; Simmerling, C.; York, D. M.; Nagan, M. C.; Merz Jr, K. M. AmberTools. *J. Chem. Inf. Model.* **2023**, *63*, 6183–6191.

(52) Scheitl, C. P. M.; Ghaem Maghami, M.; Lenz, A.-K.; Höbartner, C. Site-specific RNA methylation by a methyltransferase ribozyme. *Nature* **2020**, *587*, 663–667.

(53) Scheitl, C. P. M.; Mieczkowski, M.; Schindelin, H.; Höbartner, C. Structure and mechanism of the methyltransferase ribozyme MTR1. *Nat. Chem. Biol.* **2022**, *18*, 547–555.

(54) Deng, J.; Wilson, T. J.; Wang, J.; Peng, X.; Li, M.; Lin, X.; Liao, W.; Lilley, D. M. J.; Huang, L. Structure and mechanism of a methyltransferase ribozyme. *Nat. Chem. Biol.* **2022**, *18*, 556–564.

(55) Martick, M.; Lee, T.-S.; York, D. M.; Scott, W. G. Solvent structure and hammerhead ribozyme catalysis. *Chem. Biol.* **2008**, *15*, 332–342.

(56) Hunter, W. N.; Brown, T.; Kneale, G.; Anand, N. N.; Rabinovich, D.; Kennard, O. The structure of guanosine-thymidine mismatches in B-DNA at 2.5-A resolution. *J. Biol. Chem.* **1987**, *262*, 9962–9970.

(57) Wilson, Jr., E. B.; Decius, J. C.; Cross, P. C. *Molecular Vibrations*; Dover Publications, Inc.: New York, 1980.

(58) Akima, H. A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures. *J. ACM* **1970**, *17*, 589–602.

(59) Lee, T.-S.; Radak, B. K.; Pabis, A.; York, D. M. A new maximum likelihood approach for free energy profile construction from molecular simulations. *J. Chem. Theory Comput.* **2013**, *9*, 153–164.

(60) Lee, T.-S.; Radak, B. K.; Huang, M.; Wong, K.-Y.; York, D. M. Roadmaps through free energy landscapes calculated using the multidimensional vFEP approach. *J. Chem. Theory Comput.* **2014**, *10*, 24–34.

(61) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(62) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J. Chem. Phys.* **2012**, *136*, 144102.

(63) Zhang, B. W.; Xia, J.; Tan, Z.; Levy, R. M. A Stochastic Solution to the Unbinned WHAM Equations. *J. Phys. Chem. Lett.* **2015**, *6*, 3834–3840.

(64) Milovanović, G. V.; Udovičić, Z. Calculation of coefficients of a cardinal B-spline. *Applied Mathematics Letters* **2010**, *23*, 1346–1350.

(65) Hardy, R. L. Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res.* **1971**, *76*, 1905–1915.

(66) Fornberg, B.; Wright, G. Stable computation of multiquadric interpolants for all values of the shape parameter. *Comput. Math. with Appl.* **2004**, *48*, 853–867.

(67) Li, P.; Jia, X.; Pan, X.; Shao, Y.; Mei, Y. Accelerated Computation of Free Energy Profile at ab Initio Quantum Mechanical/Molecular Mechanics Accuracy via a Semi-Empirical Reference Potential. I. Weighted Thermodynamics Perturbation. *J. Chem. Theory Comput.* **2018**, *14*, 5583–5596.

(68) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(69) Warshel, A.; Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249.

(70) Singh, U. C.; Kollman, P. A. A combined *ab initio* quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the CH_3Cl+Cl^- exchange reaction and gas phase protonation of polyethers. *J. Comput. Chem.* **1986**, *7*, 718–730.

(71) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(72) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(73) Petersen, H. G. Accuracy and efficiency of the particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 3668–3679.

(74) Nam, K.; Gao, J.; York, D. M. An efficient linear-scaling Ewald method for long-range electrostatic interactions in combined QM/MM calculations. *J. Chem. Theory Comput.* **2005**, *1*, 2–13.

(75) Walker, R. C.; Crowley, M. F.; Case, D. A. The implementation of a fast and accurate QM/MM potential method in Amber. *J. Comput. Chem.* **2008**, *29*, 1019–1031.

(76) Figueirido, F.; Del Buono, G. S.; Levy, R. M. On finite-size effects in computer simulations using the Ewald potential. *J. Chem. Phys.* **1995**, *103*, 6133–6142.

(77) de Leeuw, S. W.; Perram, J. W.; Smith, E. R. Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. *Proc. R. Soc. London, Ser. A* **1980**, *373*, 27–56.

(78) Giese, T. J.; York, D. M. A GPU-Accelerated Parameter Interpolation Thermodynamic Integration Free Energy Method. *J. Chem. Theory Comput.* **2018**, *14*, 1564–1582.

(79) Zgarbová, M.; Otyepka, M.; Šponer, J.; Mládek, A.; Banáš, P.; Cheatham III, T. E.; Jurečka, P. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886–2902.

(80) Joung, I. S.; Cheatham III, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(81) McCarthy, E.; Ekesan, Ş.; Giese, T. J.; Wilson, T. J.; Deng, J.; Huang, L.; Lilley, D. M. J.; York, D. M. Catalytic mechanism and pH dependence of a methyltransferase ribozyme (MTR1) from computational enzymology. *Nucleic Acids Res.* **2023**, *51*, 4508–4518.

(82) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(83) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **1992**, *32*, 523–535.

(84) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the seld-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

(85) Li, P.; Roberts, B. P.; Chakravorty, D. K.; Merz, Jr., K. M. Rational design of Particle Mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent. *J. Chem. Theory Comput.* **2013**, *9*, 2733–2748.

(86) Panteva, M. T.; Giambaşu, G. M.; York, D. M. Comparison of structural, thermodynamic, kinetic and mass transport properties of Mg^{2+} ion models commonly used in biomolecular simulations. *J. Comput. Chem.* **2015**, *36*, 970–982.

(87) Panteva, M. T.; Giambasu, G. M.; York, D. M. Force Field for Mg^{2+}, Mn^{2+}, Zn^{2+}, and Cd^{2+} Ions that have Balanced Interactions with Nucleic Acids. *J. Phys. Chem. B* **2015**, *119*, 15460–15470.

(88) Ekesan, Ş.; McCarthy, E.; Case, D. A.; York, D. M. RNA Electrostatics: How Ribozymes Engineer Active Sites to Enable Catalysis. *J. Phys. Chem. B* **2022**, *126*, 5982–5990.

(89) Lopez, X.; York, D. M. Parameterization of semiempirical methods to treat nucleophilic attacks to biological phosphates: AM1/d parameters for phosphorus. *Theor. Chem. Acc.* **2003**, *109*, 149–159.

(90) Zgarbova, M.; Sponer, J.; Otyepka, M.; Cheatham III, T. E.; Galindo-Murillo, R.; Jurecka, P. Refinement of the sugar-phosphate backbone torsion beta for AMBER Force Fields improves the description of Z- and B-DNA. *J. Chem. Theory Comput.* **2015**, *11*, 5723–5736.

(91) Kimsey, I. J.; Szymanski, E. S.; Zahurancik, W. J.; Shakya, A.; Xue, Y.; Chu, C.-C.; Sathyamoorthy, B. J.; Suo, Z.; Al-Hashimi, H. M. Dynamic basis for dG.dT misincorporation via tautomerization and ionization. *Nature* **2018**, *554*, 195–201.

(92) Kruger, K.; Grabowski, P.; Zaug, A.; Sands, J.; Gottschling, D.; Cech, T. Self-splicing RNA: autoexcision and autocyclisation of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* **1982**, *31*, 147–157.

(93) Gilbert, W. The RNA World. *Nature* **1986**, *319*, 618.

(94) Deng, J.; Shi, Y.; Peng, X.; He, Y.; Chen, X.; Li, M.; Lin, X.; Liao, W.; Huang, Y.; Jiang, T.; Lilley, D. M. J.; Miao, Z.; Huang, L. Ribocentre: a database of ribozymes. *Nucleic Acids Res.* **2023**, *51*, D262–D268.

(95) Wilson, T. J.; Lilley, D. M. J. The potential versatility of RNA catalysis. *Wiley Interdiscip. Rev. RNA* **2021**, *12*, 1651.

(96) Scott, W. G. Biophysical and biochemical investigations of RNA catalysis in the hammerhead ribozyme. *Q. Rev. Biophys.* **1999**, *32*, 241–294.

(97) Martick, M.; Scott, W. G. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* **2006**, *126*, 309–320.

(98) Leclerc, F. Hammerhead Ribozymes: True Metal or Nucleobase Catalysis? Where Is the Catalytic Power from? *Molecules* **2010**, *15*, 5389–5407.

(99) Blount, K. F.; Uhlenbeck, O. C. The Structure-Function Dilemma of the Hammerhead Ribozyme. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 415–440.

(100) Nelson, J. A.; Uhlenbeck, O. C. Hammerhead redux: does the new structure fit the old biochemical data? *RNA* **2008**, *14*, 605–615.

(101) Thomas, J. M.; Perrin, D. M. Probing general acid catalysis in the hammerhead ribozyme. *J Am Chem Soc* **2009**, *131*, 1135–1143.

(102) Ward, W. L.; Plakos, K.; DeRose, V. J. Nucleic acid catalysis: metals, nucleobases, and other cofactors. *Chem. Rev.* **2014**, *114*, 4318–4342.

(103) Lee, T.-S.; Silva-Lopez, C.; Martick, M.; Scott, W. G.; York, D. M. Insight into the role of Mg^{2+} in hammerhead ribozyme catalysis from x-ray crystallography and molecular dynamics simulation. *J. Chem. Theory Comput.* **2007**, *3*, 325–327.

(104) Lee, T.-S.; Silva Lopez, C.; Giambaşu, G. M.; Martick, M.; Scott, W. G.; York, D. M. Role of Mg^{2+} in hammerhead ribozyme catalysis from molecular simulation. *J. Am. Chem. Soc.* **2008**, *130*, 3053–3064.

(105) Lee, T.-S.; Giambaşu, G. M.; Sosa, C. P.; Martick, M.; Scott, W. G.; York, D. M. Threshold Occupancy and Specific Cation Binding Modes in the Hammerhead Ribozyme Active Site are Required for Active Conformation. *J. Mol. Biol.* **2009**, *388*, 195–206.

(106) Lee, T.-S.; York, D. M. Computational mutagenesis studies of hammerhead ribozyme catalysis. *J. Am. Chem. Soc.* **2010**, *132*, 13505–13518.

(107) Wong, K.-Y.; Lee, T.-S.; York, D. M. Active participation of the Mg^{2+} ion in the reaction coordinate of RNA self-cleavage catalyzed by the hammerhead ribozyme. *J. Chem. Theory Comput.* **2011**, *7*, 1–3.

(108) Chen, H.; Giese, T. J.; Golden, B. L.; York, D. M. Divalent Metal Ion Activation of a Guanine General Base in the Hammerhead Ribozyme: Insights from Molecular Simulations. *Biochemistry* **2017**, *56*, 2985–2994.

(109) Ganguly, A.; Weissman, B. P.; Piccirilli, J. A.; York, D. M. Evidence for a Catalytic Strategy to Promote Nucleophile Activation in Metal-Dependent RNA-Cleaving Ribozymes and 8-17 DNAzyme. *ACS Catal.* **2019**, *9*, 10612–10617.

(110) Gaines, C. S.; Piccirilli, J. A.; York, D. M. The L-platform/L-scaffold framework: a blueprint for RNA-cleaving nucleic acid enzyme design. *RNA* **2020**, *26*, 111–125.

(111) Mendiratta, G.; Ke, E.; Aziz, M.; Liarakos, D.; Tong, M.; Stites, E. C. Cancer gene mutation frequencies for the U.S. population. *Nat. Commun.* **2021**, *12*, 5961.

(112) Watson, J. D.; Crick, F. H. C. Molecular structure of nucleic acids. *Nature* **1953**, *171*, 737–738.

(113) Kimsey, I. J.; Petzold, K.; Sathyamoorthy, B.; Stein, Z. W.; Al-Hashimi, H. M. Visualizing transient Watson–Crick-like mispairs in DNA and RNA duplexes. *Nature* **2015**, *519*, 315–320.

(114) Szymanski, E. S.; Kimsey, I. J.; Al-Hashimi, H. M. Direct NMR Evidence that Transient Tautomeric and Anionic States in dG·dT Form Watson–Crick-like Base Pairs. *J. Am. Chem. Soc.* **2017**, *139*, 4326–4329.

(115) Li, P.; Rangadurai, A.; Al-Hashimi, H. M.; Hammes-Schiffer, S. Environmental Effects on Guanine-Thymine Mispair Tautomerization Explored with Quantum Mechanical/Molecular Mechanical Free Energy Simulations. *J. Am. Chem. Soc.* **2020**, *142*, 11183–11191.

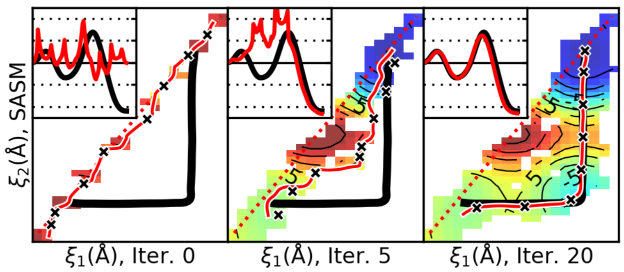

Figure 8: TOC Graphic