JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Identification of Mutational Hot Spots for Substrate Diffusion: Application to Myoglobin

David De Sancho,[†,∥,⊥] Adam Kubas,[‡,#] Po-Hung Wang,[‡,▽] Jochen Blumberger,[‡] and Robert B. Best*[,§]

[†]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

[‡]Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, United Kingdom

[§]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, United States

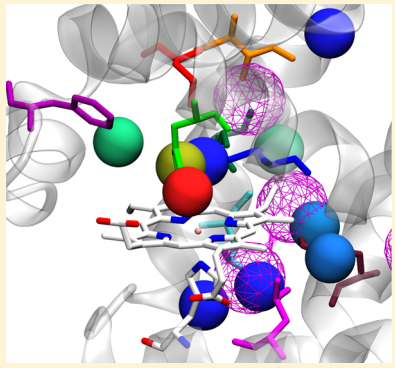
[∥]CIC nanoGUNE, Tolosa Hiribidea 76, 20018 Donostia-San Sebastian, Spain

[⊥]IKERBASQUE, Basque Foundation for Science, Maria Diaz de Haro 3, 48013 Bilbao, Spain

[▽]Theoretical Molecular Science Laboratory, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

Supporting Information

ABSTRACT: The pathways by which small molecules (substrates or inhibitors) access active sites are a key aspect of the function of enzymes and other proteins. A key problem in designing or altering such proteins is to identify sites for mutation that will have the desired effect on the substrate transport properties. While specific access channels have been invoked in the past, molecular simulations suggest that multiple routes are possible, complicating the analysis. This complexity, however, can be captured by a Markov State Model (MSM) of the ligand diffusion process. We have developed a sensitivity analysis of the resulting rate matrix, which identifies the locations where mutations should have the largest effect on the diffusive on rate. We apply this method to myoglobin, which is the best characterized example both from experiment and simulation. We validate the approach by translating the sensitivity parameter obtained from this method into the CO binding rates in myoglobin upon mutation, resulting in a semi-quantitative correlation with experiments. The model is further validated against an explicit simulation for one of the experimental mutants.



INTRODUCTION

The diffusion of small molecules inside proteins is often essential to their function. Perhaps the best known example is the binding of oxygen and carbon monoxide by hemoglobin and myoglobin. Early crystal structures of myoglobin did not reveal a clear access path to the heme pocket, implying a role for protein dynamics.[1] The same theme occurs in the diffusion of enzyme substrates from the solvent to buried active sites, such as in P450 cytochromes[2] or flavoenzymes,[3,4] or from one active site to another in multi-enzyme complexes such as tryptophan synthase[5] and carbon monoxide dehydrogenase/acetyl-CoA synthase.[6] The study of the ligand migration pathways is hence key for a fundamental understanding of protein function and critical also for our ability to engineer these systems, for example, to enhance substrate diffusion or reduce active site access for inhibitors.

The dynamics of the diffusion process and the conformational states involved can be partially resolved via a combination of ultrafast spectroscopy and time-resolved crystallography. Molecular dynamics (MD) simulations can provide a complementary, and more detailed, picture of the mechanisms by which small molecules are able to reach the protein active sites or binding sites. By far the best characterized model system is myoglobin, due to its suitability for both ultrafast spectroscopy and crystallography.[7,8] Currently experiments[9−18] and simulations[19−35] have reached a general consensus on the protein cavities occupied by the gas molecules and the access tunnels to the heme group. An interesting avenue of research is the possibility of engineering the ligand diffusion process for proteins of biomedical or industrial interest using site-directed mutagenesis. However, the engineering of enzymes for modulating the access of ligands is still primarily guided by visual inspection of the structures and human intuition. Computational approaches have the potential to direct such experimental efforts by providing candidates for mutation sites using a more quantitative method.

Here, we present a general approach to calculate the effect of mutations on the binding kinetics of ligand molecules to proteins. We approximate the dynamics of gas molecules within the protein and exchange with the solvent via a Markov state model (MSM), which we show to provide a good description at relevant time scales.[36] We then present a method to analyze the sensitivity to mutation of diffusive rates obtained from the MSM to identify local "hot spots" where mutations would be expected to have the largest effect. We have applied our method

Received: December 17, 2014

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

to CO diffusion in myoglobin, for which the abundant experimental data available can serve to validate the results from our approach. We find good correspondence between the effects of mutations predicted by our calculations and experiment. Lastly, we show how the effect of mutations identified by this perturbative approach can be more accurately quantified by resampling relevant transitions in the rate matrix with new simulations.

THEORY

Master Equation. We start from the chemical master equation that describes the evolution of probability in a network of stochastic transitions (see Figure 1a) between a set of metastable states (or microstates)

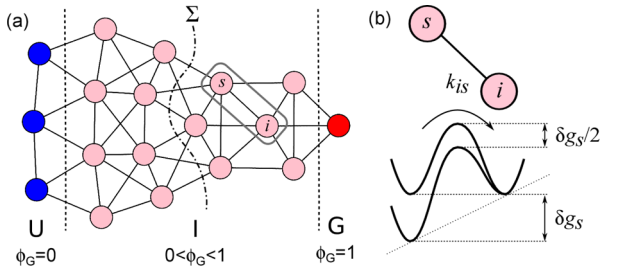


Figure 1. Theoretical framework. (a) Scheme of a network of metastable states (nodes). Edges are shown between pairs of nodes that are connected by direct transitions. The network is divided into three regions, separated by dashed lines, corresponding to the unbound state (U, blue, with $\phi_G = 0$), the geminate or bound state (G, red, with $\phi_G = 1$), and an intermediate region (I, pink, with $0 < \phi_G < 1$). The dot-dashed curve marks the dividing surface $\sum$ across which the reactive flux is calculated. A perturbation is introduced in microstate s that affects all the connected microstates i (only one such pair is shown with a gray frame for clarity). (b) Schematic free energy diagram for the perturbed microstate s and a connected microstate i. Upon a change of δg_s in the free energy of microstate s there is a change of $\delta g_s/2$ in the barrier to reach microstate i.

$$\frac{d\mathbf{p}(t)}{dt} = \mathbf{K}\mathbf{p}(t) \tag{1}$$

Here, $\mathbf{K}$ is the rate matrix, with elements k_{ji} corresponding to the rate coefficients for the transition between microstates $i \rightarrow j$, and $\mathbf{p}$ is the vector with the probabilities of all microstates in the network.

Sensitivity Analysis. We are interested in the sensitivity of an overall reaction rate (here the binding rate, k_{+1}) with respect to mutations in the network. The sensitivity upon mutation of a microstate s, α_s, is defined as the partial derivative of the binding rate with respect to changes in the free energy of the microstate

$$\alpha_s = \frac{\partial}{\partial g_s}\ln(k_{+1}) \tag{2}$$

The binding rate is calculated from the rate matrix $\mathbf{K}$ as the steady state rate for reaching the active site, using the Berezhkovskii–Hummer–Szabo (BHS) method,[37] which allows us to estimate the flux from the rate coefficients of the metastable network. Using this, the derivative in eq 2 can be expressed in terms of the elements in the rate matrix $\mathbf{K}$. We use linear free energy relationships to approximate the variation of the microscopic rate coefficients with respect to changes in the stability of a microstate.

Reactive Flux and Commitment Probabilities. We obtain the binding rate, k_{+1}, using the expression of flux over population $k_{+1} = J_{U\rightarrow G}/f_U$, where $J_{U\rightarrow G}$ is the total flux through an interface separating the unbound and bound states, and f_U is the fractional population of unbound states in equilibrium. We calculate the flux using the BHS expression[37]

$$J_{U\rightarrow G} = \sum_{j\in G^*, i\in U^*} k_{ji}p_{eq}(i)[\phi_G(j) - \phi_G(i)] \tag{3}$$

The sum in eq 3 runs over pairs of states i and j at the unbound (U^*) and bound (G^*) sides of a dividing surface $\sum$ (such as that illustrated in Figure 1a), which are connected (i.e., $k_{ji} > 0$); $p_{eq}(i)$ is the equilibrium probability of state i, and $\phi_G(i)$ is the commitment probability of state i. The values of ϕ_G (also termed the committor or p_{fold} in the context of protein folding) are defined to be 0 and 1 for microstates in the unbound and geminate states, respectively, i.e., $\phi_G(i \in U) = 0$ and $\phi_G(i \in G) = 1$. For the intermediate region, the values of ϕ_G are determined by solving[37]

$$\sum_j \phi_G(j)k_{ji} = \sum_{j\in I}\phi_G(j)k_{ji} + \sum_{j\in G}k_{ji} = 0,\ i \in I \tag{4}$$

We note that the equations for the committors and the reactive flux have an exact correspondence with those derived in transition path theory.[38,39]

Effects of Mutations on the Microscopic Rates. To reasonably approximate the effects of the mutation near a microstate on the rate coefficients of $\mathbf{K}$, we use linear free energy relationships (Figure 1b). For a mutation causing a change δg_s to the free energy of microstate s, we assume that the "transition state" for the $s \rightleftharpoons i$ transition is halfway between microstates s and i and the perturbed forward ($s \rightarrow i$) and backward ($s \leftarrow i$) rate constants are thus

$$k_{is}(\delta g_s) = k_{is}e^{\beta\delta g_s/2} \text{ and } k_{si}(\delta g_s) = k_{si}e^{-\beta\delta g_s/2} \tag{5}$$

respectively. Hence, for negative δg_s (i.e., stabilizing mutation), we get $k_{is}(\delta g_s) < k_{is}$ (i.e., slowdown of the $s \rightarrow i$ rate) and for positive δg_s (i.e., destabilizing mutation), we get $k_{is}(\delta g_s) > k_{is}$ (i.e., speedup of the $s \rightarrow i$ rate), according to intuition.

Calculation of Sensitivity Parameter. We calculate the sensitivity to point mutations in microstate s (eq 2) as

$$\alpha_s = \frac{\partial}{\partial g_s}\ln k_{+1} = \frac{\frac{\partial J}{\partial g_s}f_U - J\frac{\partial f_U}{\partial g_s}}{k_{+1}f_U^2} \tag{6}$$

This involves estimating the derivative of the flux (eq 3) with respect to the free energy of the mutated state, g_s,

$$\begin{aligned}\frac{\partial J_{U\rightarrow G}}{\partial g_s} &= \sum_{j\in G^\star, i\in U^\star}\frac{\partial}{\partial g_s}(k_{ji}p_{eq}(i)[\phi_G(j) - \phi_G(i)]) \\ &= \sum_{j\in G^\star, i\in U^\star}\left(\frac{\partial k_{ji}}{\partial g_s}p_{eq}(i)[\phi_G(j) - \phi_G(i)] + k_{ji}\frac{\partial p_{eq}(i)}{\partial g_s}\right. \\ &\quad \left. \times [\phi_G(j) - \phi_G(i)] + k_{ji}p_{eq}(i)\frac{\partial[\phi_G(j) - \phi_G(i)]}{\partial g_s}\right)\end{aligned} \tag{7}$$

The derivatives (i) $\partial k_{ji}/\partial g_s$, (ii) $\partial p_{eq}(j)/\partial g_s$, and (iii) $\partial[\phi_G(j) - \phi_G(i)]/\partial g_s$ are determined as

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX–XXX

i. The derivative of the rate coefficients, when one of the microstates involved is the mutated state s, is obtained from the linear free energy relation for the rates:

$$\frac{\partial k_{is}}{\partial g_s} = \frac{\beta}{2} k_{is} \text{ and } \frac{\partial k_{si}}{\partial g_s} = -\frac{\beta}{2} k_{si} \tag{8}$$

ii. The derivative of the equilibrium population is readily obtained from the conservation of population:

$$\frac{\partial p_{\text{eq}}(j)}{\partial g_s} = \begin{cases} -\beta p_j (1 - p_j), \text{ if } j = s \\ \beta p_j p_s, \text{ if } j \neq s \end{cases} \tag{9}$$

iii. We solve for the derivative of $\phi_G(i)$ after differentiating eq 4.

$$\sum_{j \in I} \left(\frac{\partial k_{ji}}{\partial g_s} \phi_G(j) + k_{ji} \frac{\partial \phi_G(j)}{\partial g_s} \right) + \sum_{j \in G} \frac{\partial k_{ji}}{\partial g_s} = 0, \; i \in I \tag{10}$$

In practice, we can rewrite this expression as a vector equation of dimension $|I|$.

$$\frac{\partial \mathbf{A}}{\partial g_s}\mathbf{b} + \mathbf{A}\frac{\partial \mathbf{b}}{\partial g_s} + \frac{\partial \mathbf{c}}{\partial g_s} = 0 \tag{11}$$

where for $i \in I$ we define

$$A_{ji} = k_{ij} \text{ for } j \in I; \; b_i = \phi_G(i); \text{ and } c_i = \sum_{j \in G} k_{ji} \tag{12}$$

Then we can calculate the derivative of the ϕ_G using

$$\frac{\partial \mathbf{b}}{\partial g_s} = \left(-\frac{\partial \mathbf{c}}{\partial g_s} - \frac{\partial \mathbf{A}}{\partial g_s}\mathbf{b} \right) \mathbf{A}^{-1} \tag{13}$$

METHODS

Molecular Dynamics Simulations. We have run atomistic molecular dynamics simulations of sperm whale myoglobin (PDB id: 1mbn) using the Amber ff03 protein force field[40] in explicit TIP3P water[41] and in the presence of 20 CO molecules (corresponding to a total of 12,436 atoms with a gas concentration of 266 mM, see Figure 2a). Dynamics was propagated using a stochastic leapfrog algorithm implemented in the Gromacs package (version 4.6.5)[42] with a time step of 2 fs for a total of 500 ns. A temperature of 300 K was controlled by coupling the simulation to a Langevin thermostat with a friction of of 1 ps^{-1}, and a pressure of 1 bar was maintained using a Parrinello−Rahman barostat.[43] We use the protonation state of the protein described by Onufriev and co-workers based on available empirical evidence.[33] This is important as the protonation state has been shown to determine the conformation of His64, which can act as a gate for the entry of ligand molecules to the binding pocket.[44] We note that in simulations with an alternative protonation state the binding mechanism of the CO molecules changes considerably (Figure S3, Supporting Information). For the CO, we use a well established three-site model[45] from our previous work.[46]

The high concentration of gas molecules (a CO pressure of 1 bar corresponds to 0.93 mM concentration) was used to enhance sampling of transitions in the MSM. We have checked

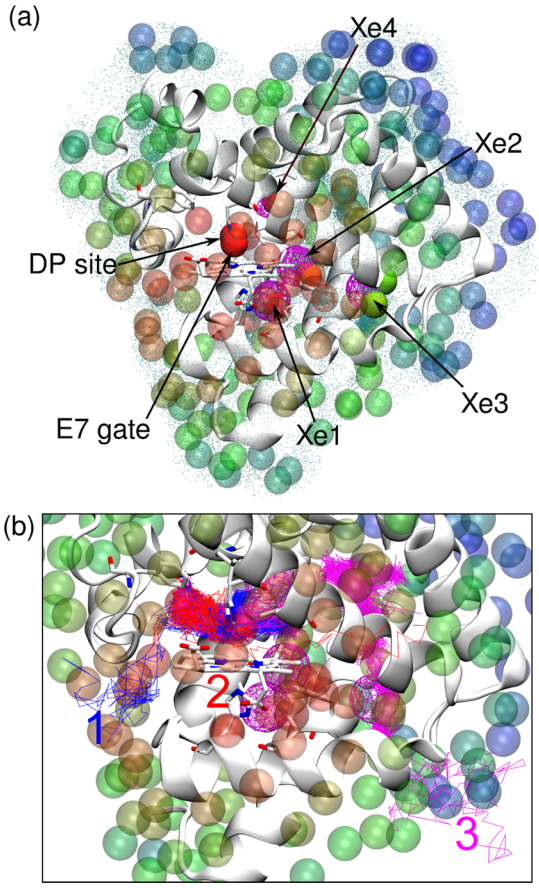


Figure 2. MSM microstates and access pathways. (a) Structure of myoglobin (white ribbon) with the positions of CO molecules from the MD trajectory (cyan dots) used in the clustering. Spheres represent the clusters corresponding to (nonsolvent) microstates used in the construction of the Markov state model, colored from red to blue based on their proximity to heme group. The clusters matching relevant xenon sites within myoglobin (purple mesh) as shown as solid spheres. (b) Access pathways to the geminate site. Lines represent trajectory paths corresponding to entry events through the three most important access ports.

that the protein dynamics is not substantially altered in the presence of the gas molecules relative to a simulation of the protein in water. Both in the absence and presence of CO molecules, the global C^{α}-RMSD of the myoglobin remains within 0.1−0.2 nm of the native structure; the profiles of residue-wise RMSD from native for the two simulations are also very similar to each other, differing by less than 0.05 nm for most residues (Figures S1 and S2, Supporting Information).The initial positions for the CO were generated by replacing at random water molecules from the simulation box. In previous work, we have shown how the gas molecules are able to rapidly permeate the protein, reaching a stable partition coefficient between water and protein.[46] In the case of myoglobin, we have verified that there is little dependence of the kinetic parameters obtained from the MSM (see below) on the number of gas molecules used (Figures S4 and S5, Supporting Information).

Markov State Model. We derive a master equation/Markov state model (MSM) from the simulation data. The procedure requires the following steps: (i) discretization of the simulation in order to define the microstates of the model, (ii) assignment of transitions among this set of microstates, (iii)

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

estimation of the transition and rate matrices based on the state trajectory, (iv) analysis, and (v) validation. We briefly describe these steps below.

Discretization. To define the microstates of the model, we first cluster the Cartesian coordinates of the virtual atoms (VCO) at the center of mass of the CO molecules. In order to do this, we overlay all snapshots of the simulation on the initial structure, using the α carbons of the myoglobin molecule as reference, and recenter all molecules to be within this primary cell. Only gas molecule positions that are closer than 6 Å of the protein α carbons or heme atoms are used in the clustering, as every other possible position of the gas molecule is lumped into the "solvent" microstate. Clustering is carried out using the Daura algorithm,[47] as implemented in the Gromacs package,[42] with a cutoff of 3 Å. From this procedure, we obtain 434 clusters of which we keep only the most populated set, accounting for 95% of the total population. This procedure results in 194 clusters, which we use together with the solvent cluster to construct the MSM.

Assignment. To identify transitions between pairs of microstates i and j, we use transition-based assignment (TBA),[48] also termed in the literature "core set MSMs".[49] For each of the individually defined clusters, we calculate the distribution of distances from the cluster center to its members. We define the transition-based assignment distance for that cluster, $r_{\mathrm{TBA}}(i)$, as the radius that contains 80% of the molecules corresponding to that cluster. This procedure results in a set of nonoverlapping spheres corresponding to the different clusters. Transitions from one state i to another state j are identified when the virtual atom of the gas molecule is closer to the center of cluster j than $r_{\mathrm{TBA}}(j)$. For transitions from a state i to the solvent, we require that the gas molecule has really left the vicinity of the protein, imposing a cutoff distance for escape of 6 Å from any protein α carbon or heavy atom from the heme group.

We use the simulation data of each individual CO molecule as an independent trajectory, assuming that the results are uncorrelated with the other gas molecules in the system. Each trajectory is assigned following the rules described above. We calculate the number of transitions between every pair of microstates, n_{ji}, after a given lag time (Δt). This results in the transition count matrix $\mathbf{N}(\Delta t)$. For the calculation of the transition probability matrix and the rate count matrix, we use only the largest strongly connected subset of the network selected using Tarjan's algorithm.[50]

Estimation. As in previous work,[51] we first determine the transition probability matrix $\mathbf{T}(\Delta t)$ using the maximum likelihood estimator[52]

$$t_{ji}(\Delta t) = n_{ji}(\Delta t)/\sum_k n_{ki}(\Delta t) \tag{14}$$

where $t_{ji}(\Delta t)$ are the elements of the transition probability matrix, i.e., the probability that initially being in state i a gas molecule is found in state j after a lag time Δt.

As before, we calculate the elements of the rate matrix $\mathbf{K}$ using the following approximation[51]

$$k_{ji} \approx \begin{cases} t_{ji}(\Delta t)/\Delta t & \text{for } i \neq j \\ -\sum_{j\neq i} k_{ji} & \text{for } i = j \end{cases} \tag{15}$$

The approximation in eq 15 becomes exact in the limit $\Delta t \to 0$.

The rates obtained from the individual CO trajectories correspond to a concentration of $\sim 1/V^{\mathrm{sim}}_{\mathrm{H_2O}}$. In order to obtain the rates at a reference 1 mM concentration, we scale the rate coefficients connecting the solvent microstate with those inside the protein by $V^{\mathrm{sim}}_{\mathrm{H_2O}}/V^{\mathrm{o}}_{\mathrm{H_2O}}$, where $V^{\mathrm{o}}_{\mathrm{H_2O}}$ is the volume per molecule of gas at a 1 mM gas concentration as before.[6,46,53,54]

Analysis. We compute equilibrium populations from the right eigenvector of the stationary mode of the rate matrix (ψ^{R}_0) and relaxation times τ_i from its eigenvalues λ_i as $\tau_I = -1/\lambda_i$. Errors for these quantities are obtained by a bootstrap method.[55] Each bootstrap sample was generated by randomly drawing trajectory segments from the pool of simulations with repetition until the same amount of data as in the original data set is obtained.

Validation. In order to validate the Markov state model, we carry out a series of tests. First, we check that the equilibrium distribution is equivalent to the population of state visits, as should be the case for a long simulation where the molecules have had enough time to equilibrate (Figure S6, Supporting Information). Then we carry out a number of tests derived from the Chapman−Kolmogorov equation. The first of these involves checking for the convergence of the relaxation times of the rate matrix at different values of the lag time (Δt). The value of the lag time must be chosen so that the relaxation times of the system are converged[56] (Figure S7, Supporting Information). Additionally, we run a test to compare the result from the propagation of the MSM and the simulation data[52] (Figure S8, Supporting Information). According to both criteria at lag times longer than 10 ps, the model produces results that are indistinguishable within error.

Estimation of Free Energy Changes for Experimental Mutations. In order to make a direct comparison with experiment, we use a database of kinetic data for mutations aiming to affect the E7 gate (H64 and F46), DP site (L29), Xe4 cavity (I28, V68 and I107), and Xe1 cavity (L89, L104, F138)[57] (Table 1). The rate we can make a comparison with is the biomolecular rate constant k'_{entry}, which corresponds to the access from the solvent to the initial docking site, from which the gas molecule would be able to bind to the heme group.[57] Although this experimenteal data set corresponds to O_2 binding rates, we assume that the relative changes will be the same for CO, as the mutations have primarily a steric effect and the molecules are similar in size.

In order to compare our results with this data set of mutants, we require an estimate of the free energy change that the mutation induces in the microstates of the MSM. Here, we use a very simple description of these effects. First, we assume that a given mutation in the protein will only affect the closest microstate s, and hence, the overall effect of the mutation can be obtained as

$$\Delta \ln k_{+1} = \Delta g_{\mathrm{mut}} \times \alpha_s \tag{16}$$

Our estimate of Δg_{mut} assumes that the changes in the amino acid volume upon mutation will translate in a decrease or increase of the cavity volume v_s and its population. Therefore, the change in population will be proportional to the relative changes in the cavity volume for the WT protein

$$\Delta g_s = -k_{\mathrm{B}}T \ln\left(\frac{p_s + p_s \Delta v_s/v_s}{p_s}\right) = -k_{\mathrm{B}}T \ln(1 + \Delta v_s/v_s) \tag{17}$$

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 1. Dataset of Entry Rates for O_2 Binding to Myoglobin from Olson et al.[57]

protein	k'_{entry} μM^{-1} s^{-1}
WT	34 ± 7
E7 gate mutations	
H64A	410
H64W	8.6
F46A	110
F46W	35
docking site mutations	
L29A	78
L29W	≥6
Xe4 cavity mutations	
I28A	49
I28W	14
V68A	44
V68W	0.2
I107A	59
I107W	21
Xe1 cavity mutations	
L89G	31
L89W	57
L104A	50
L104W	28
F138A	48
F138W	27

We assume that the change in the cavity volume (Δv_s) is the same as the change in amino acid volume from WT to mutant but with opposite sign. We use the amino acid volumes reported by Zamyatnin.[58] The reference cavity volume for the WT (v_s) is approximated as that of a sphere with radius equal to the RMSD of the particle positions for that cluster scaled by a single adjustable parameter, which we set to be $\lambda = 1/5$ to improve the correlation with experiment. The justification for doing this is that each cavity fluctuates, and so the RMSD of the particles assigned to a given cluster will generally be larger than the cavity size. In those cases where the estimated volume change from amino acid replacement is strongly negative and larger in magnitude than the actual reference volume, then a very small volume (10^{-6} Å^3) is assigned. In this case, the site should clearly be completely blocked, and so we assign it a very high free energy so that it is essentially never visited.

RESULTS

MSM Captures Essential Features of Binding of CO to Myoglobin. In order to characterize the diffusion of CO molecules into myoglobin, we have analyzed equilibrium simulations of the protein in the presence of the gas with a Markov state model. As a first test for our computational model, we check whether it correctly captures pockets identified in the protein via structural studies. In particular, we can compare our clusters with the sites occupied by xenon in equilibrium crystal structures and which have been shown to be populated by CO in time-resolved crystallographic experiments (Figure 2a). The xenon sites, Xe1, Xe2, Xe3, and Xe4, which are transiently visited by CO molecules after photodissociation,[17] are all found by the clustering algorithm without any preconditioning. There is a one-to-one mapping between the Xe1, Xe2, and Xe4 sites and single gas molecule clusters, while two of our clusters are equally close to the Xe3 site. In addition, the distal pocket (DP) (here referred to as the geminate site (G)) is identified by the clustering, and a cluster also appears immediately adjacent to His64, the key residue in the E7 gate that is widely accepted to be the main entry gate to myoglobin. These results indicate that all of these protein cavities are populated preferentially in our simulations, in accord with experiment.

In the 500 ns of equilibrium trajectories, we find seven instances in which gas molecules reach the geminate site (Figure 2b). These occur through diverse pathways that are summarized in Figure 2b. Three of the entries occur via a side-tunnel reaching Phe46 (pathway 1), two occur via His64 (pathway 2), and another two via a back gate (pathway 3). The first two pathways can be mapped onto the main access portal identified in a previous simulation study by Onufriev et al.[33] with the entry clusters corresponding to a broad region in the vicinity of the E7 gate. The back gate into the active site of the protein is widely regarded as being of minor relevance. We note, however, that in simulations carried out with a different protonation state, access to the geminate site was found to occur predominantly via this back gate, resulting in a very similar net binding rate (Figure S3, Supporting Information).

Estimate of Binding Rates and Committors. The advantage of the MSM methodology is that by stitching together the transitions from the 20 different CO molecules analyzed as independent trajectories we are able to integrate the information corresponding to the different pathways into a unified kinetic model. In Figure 3, we show the spectrum of relaxation times corresponding to the different modes of the MSM. The slowest relaxation has a characteristic time scale of 5.7 ns, while there are two more modes with similarly slow time

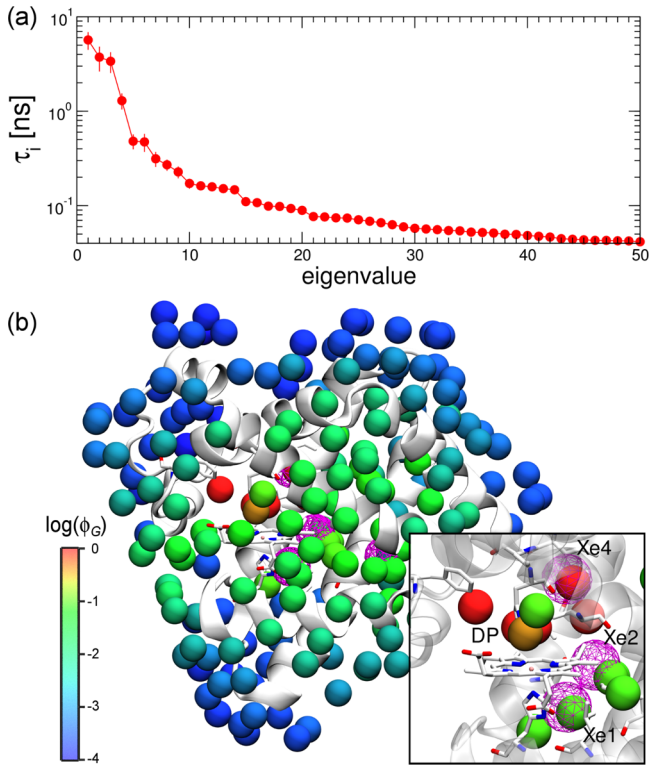


Figure 3. Markov state model for CO diffusion in myoglobin. (a) Relaxation times for the first 50 eigenvalues (λ_i) from the rate matrix $\mathbf{K}(\Delta t)$ obtained with a lag time $\Delta t = 10$ ps. (b) Commitment probabilities or ϕ_G values. Inset emphasizes the gas sites with $\phi_G > 0.1$ relative to the positions of the Xe clusters and DP site.

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

scales. Inspecting the eigenvectors of the rate matrix, we can identify the states that are exchanging with each relaxation time. The slowest mode (λ_1) corresponds to the exchange between the solvent cluster with a series of microstates in the proximity of the heme group, including those at the geminate site and Xe1 and Xe2 clusters (i.e., the associated process is the binding of the CO from the solvent to the protein interior). The next two eigenmodes (λ_2 and λ_3) correspond to the exchange between these internal clusters and those in the back entry to myoglobin (one of which is close to the Xe3 site). This description is coincident with that provided by Onuvrief and co-workers, with two main discrete pathways connected by a bottleneck (in this case corresponding to the separation emerging from modes λ_2 and λ_3).

While there are many accessible sites within the protein, it is not immediately evident which are most important and at what point the gas molecule is more likely to bind. This can be quantified by computing the committors for binding to the geminate (or DP) binding site, ϕ_G, defined here as the probability that gas molecules will reach the geminate site before reaching the solvent, i.e., the critical value of $\phi_G = 0.5$ signifies an equal probability of binding or dissociation. We use the BHS method[37] (see Methods section) to determine the values of the committors directly from the rate matrix. In this case, we define the end states corresponding to unbound, $\phi_G(U) = 0$, and bound, $\phi_G(G) = 1$, as the solvent microstate and the DP or geminate site, respectively (Figure 2b). For every other microstate between these, the committors can be calculated analytically from the rate matrix using eq 4.[37] By definition, the highest value of the committor $\phi_G = 1$ corresponds to the geminate site, and sites in its vicinity also have very high values ($0.89 < \phi_G < 1$, shown in red in Figure 3b). The cluster in front of His64, a key residue in the entry through the E7 gate, has a value of $\phi_G \simeq 0.3$, i.e., being the closest to the definition of a transition state based on this criterion (orange in Figure 3b). For the rest of the clusters, the values of the committor rapidly decrease. For microstates either on the surface of myoglobin but in the proximity of His64 or in the internal pockets of the protein (including Xe1 and Xe3), we obtain $\phi_G \simeq 0.1$ (green in Figure 3b), while for the microstates at the surface of the protein, we get $\phi_G \simeq 0$ (blue in Figure 3b)

The BHS method also allows for us to calculate fluxes for each of the microscopic transitions, so we can calculate the reactive flux through a dividing line $\sum$ that we place just before reaching the geminate state. Using eq 3, we recover the binding rate, $k_{+1} = 646\ \mu M^{-1}\ s^{-1}$ and reversing the definition of the end states also the dissociation rate $k_{-1} = 15\ \mu s^{-1}$. These estimates are somewhat faster than the experimental values of 12 μM^{-1} s^{-1} and 5.3 μs^{-1},[59] resulting in a slightly large binding constant ($K = k_{+1}/k_{-1} = 43\ M^{-1}$) relative to experiments (2.2 M^{-1}). The fast dynamics may partly be explained by the low viscosity of the TIP3P water model, resulting in a diffusion coefficient of CO in water about 3 times higher than in experiment.[60]

Sensitivity of Binding Rate to Mutations. Our main focus is to capture the relative effects of changes in one microstate of the network on the global dynamics. In particular, we would like to predict the effect of mutations on the rate of binding of the gas molecules to the heme. While one approach to doing this would be to directly simulate all mutants of interest, it would be computationally demanding and subject to considerable statistical error. To overcome these problems, we instead apply a perturbative approach to the rate matrix for the wild type protein. To do this, we consider mutations as either increasing or decreasing the volume accessible to gas molecules in a specific site within the protein, thus changing the relative free energy of that site. From the change in free energy, we estimate changes in microscopic rate coefficients k_{ji} via a linear free energy relation, allowing us to determine changes in overall rates. The sensitivity of the binding rate k_{+1} to mutations at a specific site s in the diffusion network can be characterized by the sensitivity parameter $\alpha_s = \partial \ln k_{+1}/\partial g_s$, where g_s is the free energy of site s (see Methods section).

In Figure 4a, we present the calculated sensitivity parameter, α_s, determined from the rate matrix using eqs 5 −13. The

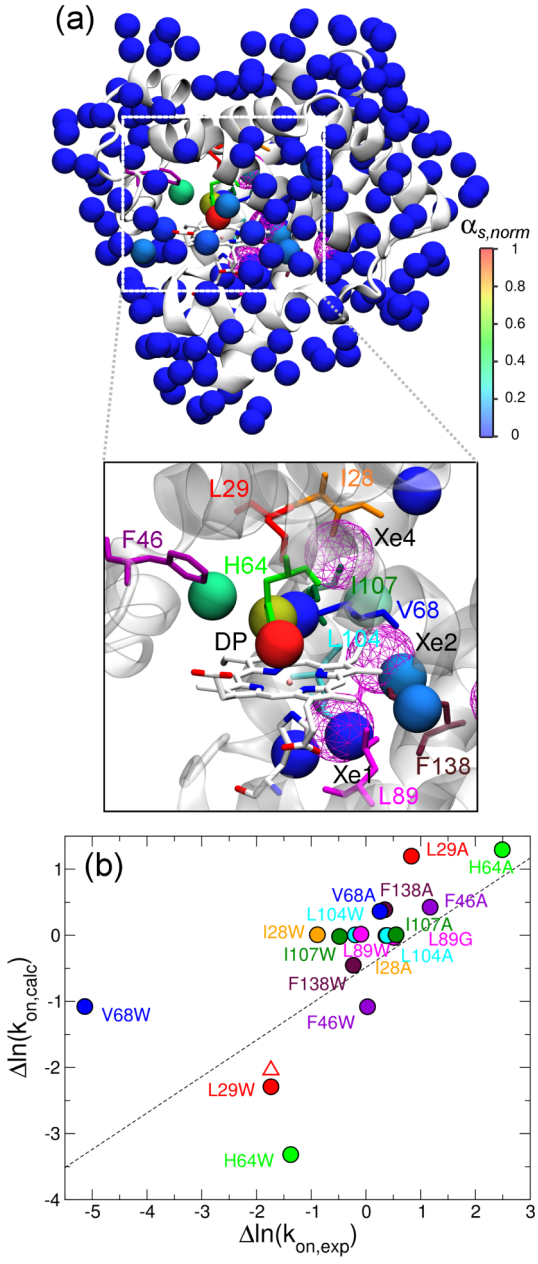


Figure 4. Sensitivity analysis of the MSM. (a) Normalized sensitivity parameters for each of the microstates, colored onto the structure. Inset shows a close-up view emphasizing the positions of the experimental mutations. (b) Correlation of experimental and calculated changes in the logarithm of the on rate. Color code is the same in both panels. Triangle corresponds to the result from the explicit simulation of the L29W mutant.

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

parameter α_s can be interpreted as the effect that an infinitesimal microstate blocking mutation would have on the binding rate, with a negative sign indicating that a decrease in the population (increase in free energy g_s) would translate into a decrease in the overall binding rate. As expected, we find that changes in the stability of microstates near the protein surface have a very little effect on the rates (blue in Figure 4a). However, in the vicinity of the DP site, the effects are much stronger, with $\partial k_{+1}/\partial g_s$ becoming large and negative (green to red in Figure 4). If we zoom in to the region that our analysis proposes is most sensitive to mutations, we find that the microstates with larger sensitivity are close to some of the sites that were experimentally mutated to block the Xe2 and Xe4 cavities (Table 1 and Figure 4a, inset). A particularly strong effect is observed around the H64 residue, the key amino acid for entry via the E7 gate. The mutation of this residue to a tryptophan resulted in a considerable decrease in the binding rate of O_2 to myoglobin.[57] Another residue that exhibited considerable sensitivity in experiments was L29, which upon mutation to a tryptophan resulted in a decrease in the entry rate by a factor of about 6. The largest effect of a blocking mutation in experiments resulted from V68W, which intended to affect the Xe2 site. In our model, the sensitivity of the Xe2 is larger than for most other clusters. However, the simulation model seems to emphasize the effect of the E7 gate pathway.

Validation of Sensitivity Analysis. In order to make a more direct comparison with the experiments, we make a rough guess of the free energy change of the clusters (δg_s) based on the original cluster volume and the volume change between swapped amino acid residues (see Methods section). Then we can calculate the changes in the rates due to a change in free energy Δg_s of the site s, as $\Delta \ln k_{+1,\mathrm{calc}} = \Delta g_s \alpha_s$. Using this coarse-grained description of the effects of the mutation, we obtain the values of $\Delta \ln k_{+1,\mathrm{calc}}$ corresponding to each of the mutants in the data set (Figure 4b). We find a semi-quantitative agreement with the experimental results, with a correlation coefficient of $R = 0.62$. In fact, if we remove the most notable outlier (V68W) from the correlation, the correlation coefficient increases to $R = 0.79$. The model seems to overestimate the effect of the E7 gate while it underestimates the effects in alternative pathways. The model correctly captures the relative insensitivity of the rates to most of the alanine and glycine mutations, the lack of response to the Xe1 cavity mutations, and the larger effects resulting from the mutation in the vicinity of the Xe4 and docking sites. The residues L29, V68, and H64 are all identified as hot spots for mutation, although their rank order is not correctly captured, due to a failure of our coarse-grained description of the effects of the mutations on the free energy of the site. The discrepancy could be due to uncertainties in the effective volume estimate for the tryptophan, whose orientation in the mutant may determine the strength of the perturbation. Effects other than volume (e.g., electrostatics) or structural relaxation in the actual mutant, which are not included in our simple model, may also be important.

As a second attempt to validate the results of our sensitivity calculation, we have explicitly simulated one of the most sensitive mutants in the experiments (L29W). Instead of constructing an MSM from scratch for the mutant, we assume that the mutation principally affects the closest microstate, s, to the mutated residue and all its connected clusters, i for $\forall\, i$: $k_{is} \neq 0$ or $k_{si} \neq 0$. We have run simulations from those clusters as initial positions for a single CO molecule. This allowed us to recalculate the column elements k_{is} for all transitions out of the mutated site s to connected sites i and k_{ji} for all transitions out of each connected site i to all other sites j connected to it and repeat the calculation of the binding rate, k_{+1}. Via this approach, we are able to recover a binding rate of 64.1 μM^{-1} s^{-1}. The change in the binding rate from the explicit simulation $\Delta \ln k_{+1} = -2.04$ is then very close to that predicted by the model, $\Delta \ln k_{+1} = -2.29$ and also very close to the experimental change $\Delta \ln k_{+1} = -1.74$ (Figure 4).

CONCLUSIONS

We have presented a new approach for quickly determining the sensitivity of diffusion on rates for protein substrates to mutations in the protein interior. The methodology can also be coupled with calculations that explicitly consider the chemical attachment of the ligand to the binding site.[61,62] This approach can serve as an initial screening for carrying out explicit simulations of mutants or for experimentalists to direct their protein engineering studies. The method is based on inferring a kinetic Markovian model from atomistic simulation data, from which we calculate a sensitivity metric for the effects of mutants on the rates.

We have applied the method to the study of ligand binding in myoglobin. The MSM that we have constructed is in good agreement with the current models for gas diffusion within this protein obtained from simulation and experiment and sheds new light on the role of the E7 gate as an entry port to myoglobin and on the time scales for communication between the primary and secondary pathways, which are comparable to the slowest relaxation time in the system. Our approach is validated in two different ways. First, using a set of simple assumptions to estimate the free energy change of a model microstate upon mutation in myoglobin, we find semi-quantitative agreement with experiments. Second, we test our results against the explicit simulation of one of the experimentally relevant mutants (L29W), again finding remarkable agreement. In future work, we will address a more refined estimation of the free energy changes, although the results of the very coarse approach used here are encouraging.

The method proposed will make it possible to identify "hot spots" for mutations in proteins that should maximally affect ligand diffusion. A similar approach would also be more generally applicable to analyze the sensitivity of global rate coefficients obtained from any MSM to the stability of its constituent microstates.

ASSOCIATED CONTENT

Supporting Information

Additional molecular visualizations, details of the MSM methodology and results, and analytical expressions for the sensitivity parameter. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: robertbe@helix.nih.gov.

Present Address

[#]Max Planck Institute for Chemical Energy Conversion, Stiftstr. 34-36, 45470 Mülheim an der Ruhr, Germany.

Notes

The authors declare no competing financial interest.

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

■ ACKNOWLEDGMENTS

D.D.S. was supported by EPSRC Grant EP/J016764/1 and A.K. by EPSRC Grant EP/J015571/1. R.B.B. was supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health. J.B. thanks the Royal Society for a University Research Fellowship. This work was carried out on the HECToR and Archer computing facilities (Edinburgh), access to which was granted through the Materials Chemistry Consortium (EPSRC Grants EP/F067496 and EP/L000202). This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). The authors acknowledge the use of the UCL Legion High Performance Computing Facility (Legion@UCL) and associated support services in the completion of this work. D.D. acknowledges Jacob Stevenson for useful discussions.

■ REFERENCES

(1) Kendrew, J. C.; Dickerson, R. E.; Strandberg, B. E.; Hart, R. G.; Davies, D. R.; Phillips, D. C.; Shore, V. C. *J. Am. Chem. Soc.* **2005**, *127*, 16961−16968.
(2) Cojocaru, V.; Winn, P. J.; Wade, R. C. *Biochim. Biophys. Acta* **2007**, *1770*, 390−401.
(3) Baron, R.; Riley, C.; Chenprakhon, P.; Thotsaporn, K.; Winter, R. T.; Alfieri, A.; Forneris, F.; van Berkel, W. J. H.; Chaiyen, P.; Fraaije, M. W.; Mattevi, A.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 10603−10608.
(4) Baron, R.; McCammon, J. A.; Mattevi, A. *Curr. Opin. Struct. Biol.* **2009**, *19*, 672−679.
(5) Hyde, C. C.; Ahmed, S. A.; Padlan, E. A.; Miles, E. W.; Davies, D. R. *J. Biol. Chem.* **1988**, *263*, 17857−17871.
(6) Wang, P.-h.; Bruschi, M.; de Gioia, L.; Blumberger, J. *J. Am. Chem. Soc.* **2013**, *135*, 9493−9502.
(7) Brunori, M.; Bourgeois, D.; Vallone, B. *J. Struct. Biol.* **2004**, *147*, 223−234.
(8) Elber, R. *Curr. Opin. Struc. Biol.* **2010**, *20*, 162−167.
(9) Austin, R. H.; Beeson, K. W.; Eisenstein, L.; Frauenfelder, H.; Gunsalus, I. C. *Biochemistry* **1975**, *14*, 5355−5373.
(10) Gibson, Q. H.; Regan, R.; Elber, R.; Olson, J. S.; Carver, T. E. *J. Biol. Chem.* **1992**, *267*, 22022−34.
(11) Huang, X.; Boxer, S. G. *Nat. Struct. Biol.* **1994**, *1*, 226−229.
(12) Scott, E. E.; Gibson, Q. H. *Biochemistry* **1997**, *36*, 11909−11917.
(13) Lim, M.; Jackson, T. A.; Anfinrud, P. A. *Nat. Struct. Mol. Biol.* **1997**, *4*, 209−214.
(14) Ostermann, A.; Waschipky, R.; Parak, F. G.; Nienhaus, G. U. *Nature* **2000**, *404*, 205−208.
(15) Merchant, K. A.; Noid, W. G.; Akiyama, R.; Finkelstein, I. J.; Goun, A.; McClain, B. L.; Loring, R. F.; Fayer, M. D. *J. Am. Chem. Soc.* **2003**, *125*, 13804−13818.
(16) Nienhaus, K.; Deng, P.; Kriegl, J. M.; Nienhaus, G. U. *Biochemistry* **2003**, *42*, 9647−9658.
(17) Schotte, F.; Lim, M.; Jackson, T. A.; Smirnov, A. V.; Soman, J.; Olson, J. S.; Phillips, G. N.; Wulff, M.; Anfinrud, P. A. *Science* **2003**, *300*, 1944−1947.
(18) Bourgeois, D.; Vallone, B.; Arcovito, A.; Sciara, G.; Schotte, F.; Anfinrud, P. A.; Brunori, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 4924−4929.
(19) Case, D. A.; McCammon, J. A. *Ann. N.Y. Acad. Sci.* **1986**, *482*, 222−233.
(20) Elber, R.; Karplus, M. *Science* **1987**, *235*, 318−321.
(21) Quillin, M. L.; Li, T.; Olson, J. S., Jr,; Duo, G. N. P.; Ikeda-Saito, Y.; Regan, M.; Carlson, R.; Gibson, M.; Li, Q. H.; Elber, R. *J. Mol. Biol.* **1995**, *245*, 416−436.
(22) Carlson, M. L.; Regan, R. M.; Gibson, Q. H. *Biochemistry* **1996**, *35*, 1125−1136.
(23) Vitkup, D.; Petsko, G. A.; Karplus, M. *Nat. Struct. Mol. Biol.* **1997**, *4*, 202−208.
(24) Meller, J.; Elber, R. *Biophys. J.* **1998**, *74*, 789−802.
(25) Nutt, D. R.; Meuwly, M. *Biophys. J.* **2003**, *85*, 3612−3623.
(26) Nutt, D. R.; Meuwly, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 5998−6002.
(27) Bossa, C.; Anselmi, M.; Roccatano, D.; Amadei, A.; Vallone, B.; Brunori, M.; Nola, A. D. *Biophys. J.* **2004**, *86*, 3855−3862.
(28) Bossa, C.; Amadei, A.; Daidone, I.; Anselmi, M.; Vallone, B.; Brunori, M.; Nola, A. D. *Biophys. J.* **2005**, *89*, 465−474.
(29) Goldbeck, R. A.; Bhaskaran, S.; Ortega, C.; Mendoza, J. L.; Olson, J. S.; Soman, J.; Kliger, D. S.; Esquerra, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 1254−1259.
(30) Cohen, J.; Schulten, K. *Biophys. J.* **2007**, *93*, 3591−3600.
(31) Ceccarelli, M.; Anedda, R.; Casu, M.; Ruggerone, P. *Proteins* **2008**, *71*, 1231−1236.
(32) Anselmi, M.; Nola, A. D.; Amadei, A. *Biophys. J.* **2008**, *94*, 4277−4281.
(33) Ruscio, J. Z.; Kumar, D.; Shukla, M.; Prisant, M. G.; Murali, T. M.; Onufriev, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9204−9209.
(34) Mishra, S.; Meuwly, M. *Biophys. J.* **2010**, *99*, 3969−3978.
(35) Maragliano, L.; Cottone, G.; Ciccotti, G.; Vanden-Eijnden, E. *J. Am. Chem. Soc.* **2010**, *132*, 1010−1017.
(36) Chodera, J. D.; Noé, F. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.
(37) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(38) E, W.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123*, 503−523.
(39) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.
(40) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(42) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845−854.
(43) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182−7190.
(44) Boechi, L.; Arrar, M.; Martí, M. A.; Olson, J. S.; Roitberg, A. E.; Estrin, D. A. *J. Biol. Chem.* **2013**, *288*, 6754−6762.
(45) Straub, J. E.; Karplus, M. *Chem. Phys.* **1991**, *158*, 221−248.
(46) Wang, P.-h.; Best, R. B.; Blumberger, J. *J. Am. Chem. Soc.* **2011**, *133*, 3548−3556.
(47) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.
(48) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.
(49) Schutte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105.
(50) Tarjan, R. *SIAM J.Comput.* **1972**, *1*, 146−160.
(51) De Sancho, D.; Mittal, J.; Best, R. B. *J. Chem. Theory Comput.* **2013**, *9*, 1743−1753.
(52) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(53) Wang, P.-h.; Best, R. B.; Blumberger, J. *Phys. Chem. Chem. Phys.* **2011**, *13*, 7708−7719.
(54) Wang, P.-h.; Blumberger, J. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 6399−6404.
(55) Efron, B. *Ann. Stat.* **1979**, *7*, 1−26.
(56) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.
(57) Olson, J. S.; Soman, J.; Phillips, G. N. *IUBMB Life* **2007**, *59*, 552−562.
(58) Zamyatnin, A. *Annu. Rev. Biophys. Bioeng.* **1984**, *13*, 145−165.
(59) Carver, T. E.; Rohlfs, R. J.; Olson, J. S.; Gibson, Q. H.; Blackmore, R. S.; Springer, B. A.; Sligar, S. G. *J. Biol. Chem.* **1990**, *265*, 20007−20.

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(60) Yeh, I.-C.; Hummer, G. *J. Am. Chem. Soc.* **2002**, *124*, 6563−6568.
(61) Harvey, J. N. *Faraday Discuss.* **2004**, *127*, 165−177.
(62) Kubas, A.; De Sancho, D.; Best, R. B.; Blumberger, J. *Angew. Chem., Int. Ed.* **2014**, *126*, 4165−4168.

DOI: 10.1021/ct5011455
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX