Automatic Method for Identifying Reaction Coordinates in Complex Systems†

Ao Ma and Aaron R. Dinner*

Department of Chemistry, James Franck Institute and Institute for Biophysical Dynamics, The University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637

Received: September 30, 2004; In Final Form: December 2, 2004

To interpret simulations of a complex system to determine the physical mechanism of a dynamical process, it is necessary to identify the small number of coordinates that distinguish the stable states from the transition states. We develop an automatic method for identifying these degrees of freedom from a database of candidate physical variables. In the method neural networks are used to determine the functional dependence of the probability of committing to a stable state (committor) on a set of coordinates, and a genetic algorithm selects the combination of inputs that yields the best fit. The method enables us to obtain the first set of coordinates that is demonstrably sufficient to specify the transition state of the $C_{7eq} \rightarrow \alpha_R$ isomerization of the alanine dipeptide in the presence of explicit water molecules. It is revealed that the solute−solvent coupling can be described by a solvent-derived electrostatic torque around one of the main-chain bonds, and the collective, long-ranged nature of this interaction accounts for previous failures to characterize this reaction.

1. Introduction

Simulations of reactions are useful because they provide mechanistic insight that complements that readily obtainable from experiments. Processes of interest that are sufficiently fast can be accessed by straightforward integration of the equations of motion for the system from a suitable set of initial conditions.[1] Alternatively, if the reaction rate is dominated by the time required for fluctuations to reach particular dynamical bottlenecks, then transition path sampling (TPS)[2,3] and related methods[4−6] can be used. Both approaches yield a set of trajectories, each of which consists of a series of points in a many-dimensional phase space. To interpret these data and construct analytic theories that allow generalization and connection to experiments, it is necessary to identify a relatively small number of coordinates that govern the reaction dynamics of the system.[7−9]

In general, these key degrees of freedom are not the variables used for the integration of the equations of motion (for example, Cartesian coordinates). Rather, they are functions of the latter that distinguish the stable states of interest from the transition states separating them. In other words, dynamic trajectories projected onto the essential coordinates reflect the free energy as a function of those same variables. This correspondence implies that the coordinates used to parametrize the free energy are those that change most slowly during the reaction process and that the other degrees of freedom are sufficiently fast that they can be regarded as equilibrated at each point along the path. From a practical standpoint, identifying the physical variables that actually govern the rate-limiting dynamics accelerates the convergence of umbrella sampling simulations.[10,11] Transition-state theory can then be used to determine rates from differences in free energies between stable states and transition states.[12]

Unfortunately, expressions for reaction coordinates can be obtained by rigorous mathematical procedures[13] or by exhaustive exploration[14,15] for only the simplest of systems. Consequently, the coordinates that provide a good description of the dynamics are typically obtained by selecting combinations of order parameters based on physical intuition and then testing them for their ability to distinguish stable states from transition states in additional dynamical simulations.[7,16−18] Iteration is required in most cases because physical variables that play important roles are often nonobvious.[17] This "trial-and-error" approach is undesirable for a number of reasons: it requires considerable investment of a researcher's time, evaluation of the pertinence of each combination of order parameters to the reaction is computationally costly, and the lack of a well-defined metric of success prevents systematic refinement of guesses.

Thus, we seek to develop an automatic procedure for identifying reaction coordinates in complex systems. Ideally, such a method would construct the coordinates of interest from the simulation variables directly. Principal component analysis[19,20] and singular value decomposition[21,22] exist for this purpose, but we (unpublished results) and others[9] have found these methods inadequate. Consequently, we set a more modest goal for ourselves here. That is, we develop an efficient method to select and combine order parameters in a weighted fashion from a large pool of candidate physical variables. Specifically, we train artificial neural networks to reproduce committors[2,3,7,23,24] from up to three descriptors selected by a genetic algorithm. The genetic neural network (GNN) approach has been employed in other contexts;[25−29] the key insight in the present work is that the probability of committing to a given stable state can serve as the basis for a statistical procedure to relate physical variables to dynamics efficiently.

We use the method to identify coordinates that provide a good description of isomerization reactions of the alanine dipeptide in vacuum and in explicit solvent. As an aside, we show that the dynamics of the system in implicit solvent resembles much more closely that in vacuum than that in explicit solvent. The vacuum cases allow us to evaluate the method by comparing its output with results obtained manually.[17] We not only reproduce earlier findings but improve upon them systematically.

† Part of the special issue "David Chandler Festschrift".

* To whom correspondence should be addressed. Phone: (773) 702-2330. Fax: (773) 834-5250. E-mail: dinner@uchicago.edu.

10.1021/jp045546c CCC: $30.25
Published on Web 02/11/2005

The C_{7eq} to α_R transition in the presence of explicit water molecules provides a much more demanding test. Chandler and co-workers first studied the dynamics of this reaction and showed explicitly that it is sufficiently complex that the collective variables necessary to describe the solvent role in the mechanism cannot be determined by inspection,[9,17] in contrast to other reactions in aqueous solution.[16,18,30] The GNN method enables us to obtain, for the first time, a set of coordinates that is sufficient to specify the transition state of this complex reaction. Combination of a solute dihedral angle, a solute−solute pairwise distance, and a solvent-derived electrostatic torque around one of the rotating bonds results in prediction of committor probabilities with a root-mean-square (RMS) error of 0.133 (relative to a range of 1). The selection of the torque is significant because it reveals that the solute−solvent coupling is electrostatic rather than steric in nature; that the transition is governed by long-range interactions with the surrounding water molecules explains the inability to delineate the transition state with solvent distribution functions.[9] Namely, the water reorientation events that give rise to the fluctuations in the electrostatic field that lead to reaction can be as far away as the second solvation shell and in different locations relative to the solute in independent trajectories. By making possible the search of combinations of up to four physical variables from a database of 5812 candidates, the GNN method thus permits the identification of a specific, demonstrably important collective solvent coordinate not described elsewhere. Implications of the results for analytic models of dynamics in solvated systems are discussed.

2. Methods

In this section we first review key concepts. We then describe the procedure as well as tests to ensure that the results are physically meaningful.

2.1. Committor as a "Perfect" Reaction Coordinate. For a snapshot taken from a trajectory connecting stable states A and B, the committor p_B is defined as the likelihood that additional trajectories initiated from that configuration with momenta drawn isotropically from a Maxwell−Boltzmann distribution will commit to basin B prior to basin A.[23] In other words, p_B is essentially the transmission coefficient for reaching basin B from a given configuration.[24] Configurations in basin A (reactants) have $p_B = 0$, those in basin B (products) have $p_B = 1$, and transition states have $p_B = 0.5$ by construction. The definition of the transition state employed here differs slightly from the conventional one in which recrossings are prohibited for trajectories initiated from transition states with momenta in the forward direction. The alternate approach used in the present study (and others[7,9,16−18,23]) better lends itself to complex systems in which the reaction coordinate cannot be identified a priori. However, as defined above, p_B is limited to the characterization of configuration rather than phase space.

It is also important to distinguish between an "order parameter" and a "reaction coordinate". The former is any quantitative metric that identifies configurations in different stable states. In contrast, a reaction coordinate must be able to describe the complete dynamic process. That is, if one considers a reaction as a chain of states that connects two stable states, then a reaction coordinate must correlate closely with the index of every node in the chain, but order parameters need reflect only the structures at the two ends. Consequently, if a physical quantity is a good reaction coordinate, then it must be a good order parameter as well, but the inverse is not necessarily true. While determination of good order parameters can be difficult in its own right, identification of reaction coordinates is a much more challenging task; indeed, it is exactly the relative ease in finding order parameters that makes the transition path sampling method useful.[2,3]

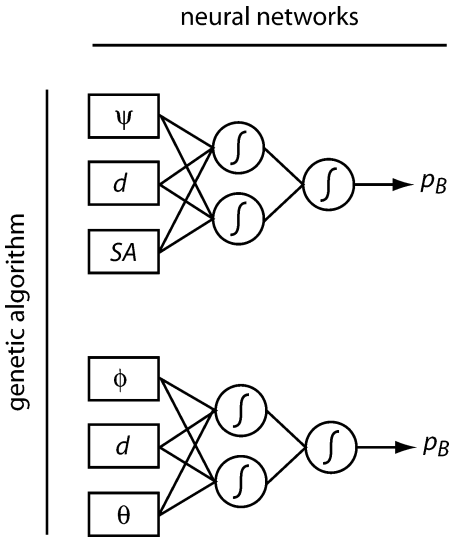

Figure 1. Genetic neural network. In the schematic neural networks with three inputs (symbolized by rectangles) and two nodes in the hidden layer (symbolized by vertical pairs of circles) are trained to reproduce p_B, and a genetic algorithm is used to select between different combinations of physical variables for the inputs.

We thus see that, by construction, p_B is a perfect reaction coordinate in the sense that it provides a quantitative description of the dynamic behavior of every state along a trajectory. At the same time p_B is not useful in itself for gaining mechanistic insight into a specific process because it does not relate in a transparent way to physical parameters of a system that can be varied computationally or experimentally in a controlled fashion. Due to its universality, p_B is a perfect but totally uninformative reaction coordinate. The key insight in the present study is that p_B can be used to guide an importance sampling procedure to search the space of physically meaningful variables efficiently.

2.2. Importance Sampling of Physical Variables. We use an artificial neural network to determine the functional dependence of p_B on combinations of coordinates. A genetic algorithm is then used to select different combinations of descriptors to identify the model that allows the best statistical fit (Figure 1). Thus, the method yields both the set of most meaningful physical variables and an analytic expression that combines them into a single approximate reaction coordinate. Details of the GNN procedure are given in the Supporting Information and elsewhere.[25,26,29] The essential feature introduced here is the scoring function. Because we seek to relate the committor probability to physical coordinates, we use the RMS error in predicted p_B. This metric is used to optimize the parameters in the neural networks and estimate the fitness of different combinations of physical variables in the genetic algorithm. Instead of p_B itself, we also tried basing the scoring function on a trinary variable that classified configurations as reactants, transition states, and products, but less accurate predictions were obtained (data not shown).

The method is statistical in nature and requires a database to train the neural networks. Each entry in the database consists of a list of values for the physical variables to be considered and p_B for a single configuration of the system. To avoid fitting certain regions along the reaction coordinate at the expense of others, it is important to obtain a set of structures that are distributed uniformly with respect to p_B. To this end, we used

transition path sampling[2,3] to harvest many reactive trajectories for each process of interest followed by an automated procedure described in the Supporting Information to select a subset of the saved configurations such that roughly equal numbers of database entries were distributed in bins of width 0.1 in p_B. However, we stress that TPS is not intrinsic to the method; any means of generating structures with a range of committor probabilities can be used.

It is worth noting that the most computationally costly part of the procedure is the evaluation of p_B for each entry in the database rather than the initial collection of reactive trajectories or the GNN search. For each p_B value it is necessary to shoot many (in most of the examples below, 100) additional trajectories, each of roughly the same length as a single reactive path. This fact accounts for the efficiency of the method. Here, p_B is calculated only once for a set of structures, while in the standard "trial-and-error" approach outlined in the Introduction this dynamical quantity is recalculated for each combination of coordinate values guessed to characterize the transition-state ensemble.

All the transition path sampling calculations were performed with CHARMM, version c31a1,[18,31] and the p_B values were obtained with a modified form of that program. The GNN search in the space of physical variables was carried out with HIPPO.[25,26] Additional simulation details are given with each example discussed below and in the Supporting Information.

2.3. Ensuring Significance. The RMS error in the predicted p_B will, in general, decrease as the number of inputs to the neural networks increases due to the greater number of adjustable parameters. To ensure that improved fits derive from increases in physically meaningful information, we divide the database into a training set and a test set. We use the former to optimize the weights of the neural networks and the latter to check the quality of the fit. As an additional check for physical significance, care was taken to ensure that the results were consistent across multiple independent trials with different random number generator seeds, which enter the procedure through both the initiation of the weights in the neural networks and mutations in the genetic algorithm.

To confirm the results of the method, each of the combinations of descriptors selected by the GNN procedure was subjected to the standard test.[2,3] Namely, we performed equilibrium simulations with a biasing potential to restrain the system to values of the physical variables believed to correspond to the transition-state ensemble and subsequently evaluated p_B for structures harvested from those simulations by shooting additional trajectories from those structures in the absence of any restraints. If this procedure yields a distribution of committor values centered closely around $p_B = 0.5$, then the selected coordinates are considered to provide a good description of the reaction.

Typically, the biasing potential used in the standard test restrains each of the selected physical variables with a separate harmonic term to its anticipated average value in the transition-state ensemble. Here, the trained neural network obtained in the GNN procedure provides an analytic expression that combines multiple coordinates into a single one (p_B^{GNN}). Consequently, we harvest putative transition-state structures with the umbrella potential term

$$V_{umb} = k(p_B^{GNN} - p_B^{targ})^2 \qquad (1)$$

where k is a force constant, p_B^{GNN} is the p_B value predicted by the GNN model for a given configuration, and p_B^{targ} is the target p_B value ($p_B^{targ} = 0.5$ to sample transition states preferentially).

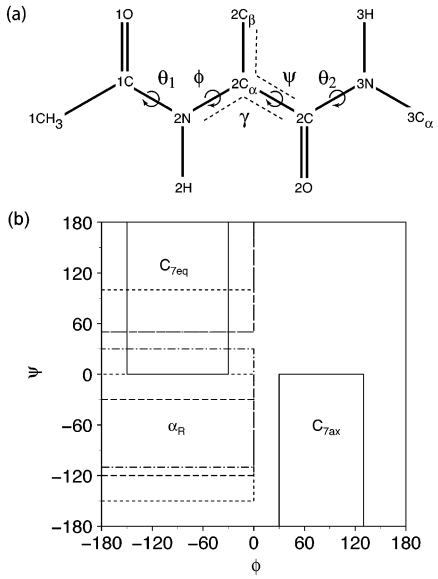

Figure 2. (a) Schematic of the alanine dipeptide. $1CH_3$ is a single extended atom. Dihedral angles that adequately describe the isomerization reactions in vacuum and implicit solvent are labeled: θ_1 (1O−1C−2N−2Cα), ϕ (1C−2N−2Cα−2C), ψ (2N−2Cα−2C−3N), and θ_2 (2O−2C−3N−3Cα). Dashed lines indicate the planes that define the improper dihedral angle $2C_\beta$−2C−2Cα−2N, denoted γ (see text and Table 1). (b) Stable states defining the reactions: $C_{7eq} \rightarrow C_{7ax}$ reaction in vacuum (solid black lines); $C_{7eq} \rightarrow \alpha_R$ in vacuum (long dash lines); explicit solvent (short dash lines); implicit solvent (dot-dash lines).

Because the derivatives of this term with respect to Cartesian coordinates are complicated, the Monte Carlo module in CHARMM[31,32] was used for these simulations. For $k = 100.0$ kcal/mol, 14% of structures fall in the range $0.49 < p_B^{GNN} < 0.51$ at 300 K, and these were the ones for which p_B was determined. The use of a single combined coordinate as the basis of the restraint is attractive because it allows individual coordinates to compensate each other as in the actual transition-state ensemble.

3. Results

To demonstrate the utility of the GNN approach, we applied it to isomerization reactions of the alanine dipeptide (Figure 2). Since our primary goal is to explore the GNN method, not to make direct connection with experiment, we chose to represent the alanine dipeptide with the CHARMM polar hydrogen topology and parameter sets[31,33] for convenience. We first compare the output of the method for the $C_{7eq} \rightarrow C_{7ax}$ transition in vacuum with results obtained for that reaction manually.[17] We then apply the method to the $C_{7eq} \rightarrow \alpha_R$ transition in vacuum, in implicit solvent, and in explicit solvent. The last of these represents a stringent test of the method since earlier studies[9,17] failed to yield coordinates that could be used to harvest additional transition states.

3.1. Reactions in Vacuum. *3.1.1. $C_{7eq} \rightarrow C_{7ax}$.* As mentioned above, we first examine the $C_{7eq} \rightarrow C_{7ax}$ transition, which corresponds to rotation around the 2N−2Cα bond. Basin A is defined as $-150° < \phi < -30°$ and $0° < \psi < 180°$, and basin B is defined as $30° < \phi < 130°$ and $-180° < \psi < 0°$ (Figure 2). As described in the Supporting Information we harvested 500 reactive trajectories, each of which was 1.2 ps in length. From these trajectories a database of 2100 configurations was prepared, and unless otherwise noted the database was partitioned into a training set of 2000 entries and a test set of 100 entries.

TABLE 1: Best Models for Isomerization Reactions of the Alanine Dipeptide[a]

			input RMS			model RMS	
reaction	n	inputs	train	test	frequency	train	test
$C_{7eq} \rightarrow C_{7ax}$ (100 trials for p_B)	1	1C−2N−2C$_\alpha$−2C (ϕ)	0.102	0.125	3/3	0.102	0.125
	2	1C−2N−2C$_\alpha$−2C (ϕ)	0.102	0.125	5/5	0.088	0.099
		1O−1C−2N−2C$_\alpha$ (θ_1)	0.292	0.289			
	3	1C−2N−2C$_\alpha$−2C (ϕ)	0.102	0.125	5/6	0.077	0.083
		1O−1C−2N−2C$_\alpha$ (θ_1)	0.292	0.289			
		2N−2C$_\alpha$−2C−3N (ψ)	0.156	0.187			
$C_{7eq} \rightarrow C_{7ax}$ (400 trials for p_B)	1	1C−2N−2C$_\alpha$−2C (ϕ)	0.111	0.103	3/3	0.111	0.103
	2	1C−2N−2C$_\alpha$−2C (ϕ)	0.111	0.103	3/3	0.061	0.056
		1O−1C−2N−2C$_\alpha$ (θ_1)	0.265	0.207			
	3	1C−2N−2C$_\alpha$−2C (ϕ)	0.111	0.103	10/10	0.050	0.042
		1O−1C−2N−2C$_\alpha$ (θ_1)	0.265	0.207			
		2C$_\beta$−2C−2C$_\alpha$−2N (γ)	0.174	0.175			
$C_{7eq} \rightarrow \alpha_R$ (vacuum)	1	2N−2C$_\alpha$−2C−3N (ψ)	0.111	0.117	3/3	0.110	0.117
	2	2N−2C$_\alpha$−2C−3N (ψ)	0.111	0.117	4/4	0.087	0.080
		2O−2C−3N−3C$_\alpha$ (θ_2)	0.289	0.283			
$C_{7eq} \rightarrow \alpha_R$ (explict solvent and instant coordinates)	1	2C$_\beta$−2C$_\alpha$−2C−2O (ψ')	0.176	0.154	3/3	0.176	0.154
	2	2C$_\beta$−2C$_\alpha$-2C−2O (ψ')	0.176	0.154	4/7	0.169	0.140
		$\mathcal{N}_{1C-2N}^{3H}$	0.222	0.207			
	3	2C$_\beta$−2C$_\alpha$−2C−2O (ψ')	0.176	0.154	3/5	0.167	0.133
		$r_{2H-2C\alpha}$	0.255	0.254			
		$\mathcal{N}_{1C-2N}^{3H}$	0.222	0.207			
$C_{7eq} \rightarrow \alpha_R$ (explict solvent and average coordinates)	1	2C$_\beta$−2C$_\alpha$−2C−2O (ψ')	0.176	0.154	3/3	0.176	0.154
	2	$\langle r_{2H-3H} \rangle_{solute}$	0.191	0.179	4/4	0.164	0.132
		$\langle \mathcal{N}_{1C-2N}^{3C\alpha} \rangle_{solute}$	0.177	0.159			
	3	$\langle r_{2H-3H} \rangle_{solute}$	0.191	0.179	3/3	0.161	0.126
		$\langle \mathcal{N}_{1C-2N}^{3C\alpha} \rangle_{solute}$	0.177	0.159			
		$\langle \mathcal{N}_{1C-2N}^{3N} \rangle_{solvent}$	0.197	0.178			
	4	$\langle r_{2H-3H} \rangle_{solute}$	0.191	0.179	3/5	0.160	0.120
		$\langle \mathcal{N}_{1C-2N}^{3C\alpha} \rangle_{solute}$	0.177	0.159			
		$\langle \mathcal{N}_{1C-2N}^{3N} \rangle_{solvent}$	0.197	0.178			
		$\langle r_{1O-2H} \rangle_{solute}$	0.256	0.251			
$C_{7eq} \rightarrow \alpha_R$ (implicit solvent)	1	2N−2C$_\alpha$−2C−2O (ψ'')	0.190	0.180	3/3	0.190	0.180
	2	2N−2C$_\alpha$−2C−3N (ψ)	0.190	0.184	5/5	0.123	0.113
		2O−2C−3N−3C$_\alpha$ (θ_2)	0.287	0.289			
	3	2N−2C$_\alpha$−2C−3N (ψ)	0.190	0.184	5/6	0.083	0.091
		2O−2C−3N−3C$_\alpha$ (θ_2)	0.287	0.289			
		2C$_\beta$−2C−2C$_\alpha$−2N(γ)	0.287	0.288			

[a] Number of inputs is n. $\mathcal{N}_{XY}^{Z}$ indicates the electrostatic torque around bond XY from solvent forces on atom Z. $\langle ... \rangle_X$ indicates an average over configurations with X fixed (where X is the solute or solvent).

The physical variables considered for each configuration included all pairwise distances (66 in total), all possible bond angles (15 in total), and all possible dihedral angles (21 in total, of which 5 were improper). It is important to stress that the sets of bond angles and dihedral angles each contain redundant information. For instance, the physical information contained in a dihedral angle might be the same as that contained in a combination of alternative internal coordinates in the database. We did not remove such redundant coordinates because the GNN search converges more quickly and requires a smaller number of inputs to the neural networks if coordinates that correlate more closely with p_B are present. We tried models with one, two, and three descriptors. In each case multiple runs were performed with different seeds for the random number generator, as indicated in Table 1.

The GNN procedure consistently selected from the 102 coordinates the three dihedral angles identified previously by hand.[17] As expected from the fact that it correlates strongly with the committor, the dihedral angle ϕ appeared in all the best models; by itself, it is capable of predicting p_B with an RMS error of 0.125. In contrast, the second and third dihedral angles (θ_1 and ψ) correlate only weakly with p_B (RMS errors of 0.288 and 0.187, respectively). Thus, the method is capable of selecting combinations of physical variables involving degrees of freedom that by themselves are not transparently relevant to the reaction.

Each of the models for the $C_{7eq} \rightarrow C_{7ax}$ transition in Table 1 was subjected to the standard test described in section 2.3. In other words, the umbrella potential term in eq 1 was used to harvest putative transition states (configurations with $p_B^{GNN} = 0.5$) and their actual committor values were subsequently calculated in the absence of any restraints (Figure 3). Consistent with ref 17, the dihedral angle ϕ by itself is incapable of defining the transition-state ensemble; most of the configurations harvested with the one-descriptor model fall in basin B (C_{7ax}). On the other hand, the two-descriptor model does yield adequate predictions as evidenced by the fact that the distribution of actual p_B values peaks around $p_B = 0.5$. Adding a third descriptor narrows the distribution only slightly. Comparable results were obtained by applying separate harmonic restraints to each of the coordinates in the models (inset to Figure 3).

Because the computational cost associated with the method is dominated by that required to evaluate p_B for the database, we investigated the dependence of the results on the number of entries in the training set. Specifically, we varied the training set size from 100 to 2000 configurations with the test set size fixed at 100. In all cases, the GNN procedure selected the two-descriptor model identified previously by hand.[17] The RMS error in p_B varied from 0.080 to 0.105, but there was no clear trend, which suggests that a training set size of 100 is adequate for this example. For more complex reactions, larger training set

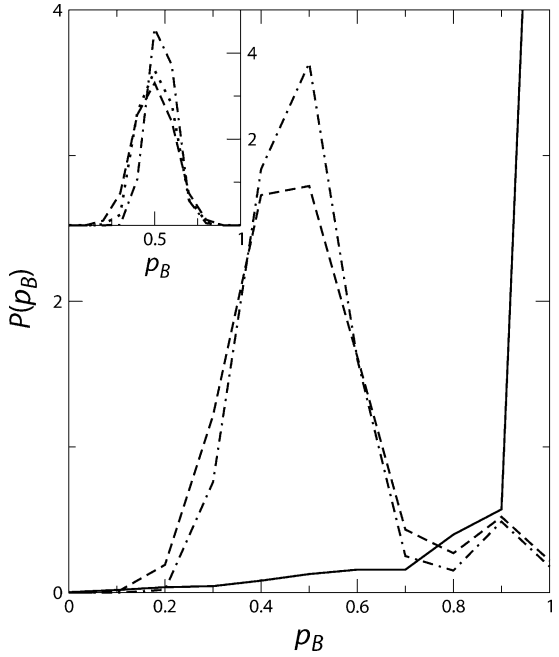

Figure 3. Committor distribution for the $C_{7eq} \rightarrow \alpha_R$ transition in vacuum. The ensemble of configurations was harvested with umbrella sampling with the restraint in eq 1 using the corresponding one-descriptor (solid line), two-descriptor (dashed line), and three-descriptor (dot-dash line) models in Table 1. (Inset) The ensemble of configurations was harvested with umbrella sampling with a separate harmonic restraint on each coordinate with the target values: $(\phi, \psi) = (-5.3°, -34.5°)$ (dashed line), $(\phi, \psi, \theta_1) = (-5.3°, -34.5°, 2.2°)$ (dotted line), and $(\phi, \psi, \gamma) = (-5.3°, -34.5°, -131.1°)$. The first three-descriptor model is based on p_B values calculated with 100 trajectories, and the second is based on p_B values calculated with 400 trajectories.

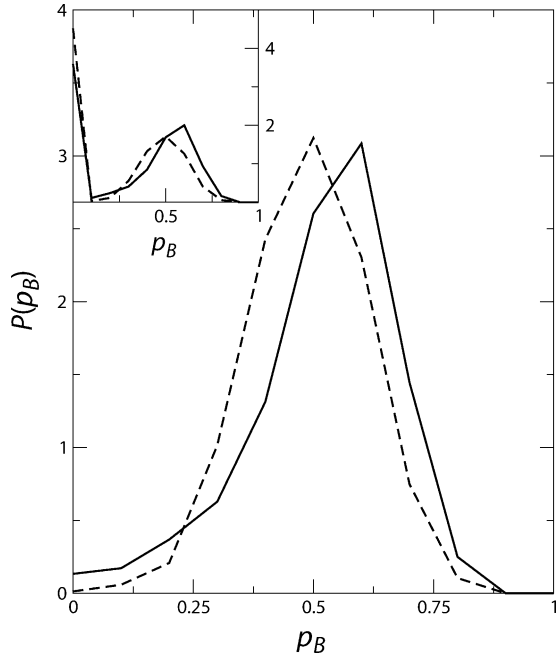

Figure 4. Committor distribution for the $C_{7eq} \rightarrow \alpha_R$ transition in vacuum. The ensemble of configurations was harvested with umbrella sampling with the restraint in eq 1 using the corresponding one- (solid line) and two-descriptor (dashed line) models in Table 1 and filtered to remove configurations with input values outside ranges in the training set. (Inset) The same without filtering.

sizes are likely to be required to adequately constrain models because larger numbers of descriptors must typically be combined to achieve the same degree of predictive accuracy, as described below.

Because p_B is computed by a stochastic procedure, there is an intrinsic statistical error in both the values input to the GNN procedure and those used to evaluate its performance. To determine if the method was yielding predictions at this limiting accuracy, we harvested another ensemble of 2700 configurations (partitioned into a training set of 2600 configurations and a test set of 100 configurations) with p_B values calculated using 400 instead of 100 trajectories. The RMS errors for the best two- and three-descriptor models decrease to 0.056 and 0.042, respectively (Table 1). Moreover, the p_B distribution obtained in the standard test noticeably narrows (inset to Figure 3). These improvements of roughly a factor of 2 are consistent with standard signal-to-noise scaling ($\sqrt{N}$, where N is the number of samples), which suggests that the method is indeed reaching limiting accuracy. The substitution of the improper dihedral angle γ for ψ (Figure 2a) as the third coordinate in the three-descriptor model reflects the fact that p_B predictions based on each of these physical variables by itself differ in RMS error by only 0.041, which is less than the intrinsic error in the p_B values in the original database. We thus see that the results of the method can be improved systematically to obtain a desired level of accuracy.

3.1.2. $C_{7eq} \rightarrow \alpha_R$. The second vacuum reaction that we studied is the $C_{7eq} \rightarrow \alpha_R$ transition, which corresponds to rotation around the $2C_\alpha-2C$ bond. We considered this isomerization for comparison with the corresponding transition in solution, which was also studied by Chandler and co-workers.[9,17] They did not study this transition in vacuum because the α_R basin is not stable in the empirical force field that they used (Amber 94[34]).

We harvested 600 reactive trajectories of 600 fs in length with TPS; basin A was defined as $-180° < \phi < 0°$ and $-120° < \psi < -30°$, and basin B was defined as $-180° < \phi < 0°$ and $50° < \psi < 180°$ (Figure 2). From these trajectories we used the automatic procedure described in the Supporting Information to prepare a database of 2500 configurations distributed uniformly in p_B. It was partitioned into a training set of 2400 configurations and a test set of 100 configurations. The physical variables included in the database were the same as those for the $C_{7eq} \rightarrow C_{7ax}$ transition.

The lowest RMS error in p_B was obtained with a two-descriptor model which consisted of dihedral angles ψ and θ_2 (Table 1 and Figure 2a). Addition of a third descriptor did not improve the GNN fit, and there was no consistent combination selected across multiple trials with different seeds. As was the case for the $C_{7eq} \rightarrow C_{7ax}$ transition, one descriptor (ψ) correlates strongly with p_B (RMS error of 0.12) by itself but the second ($2O-2C-3N-3C_\alpha$, denoted θ_2) does not (RMS error of 0.28).

To confirm that the models identified by the GNN method were physically meaningful, we subjected them to the standard test (Figure 4). Application of the test exactly as described in section 2.3 results in a bimodal distribution with peaks at $p_B \approx 0$ and 0.5 (inset to Figure 4). Inspection of the structures with $p_B \approx 0$ revealed that they have values of $\phi > 0$, while the structures in the training set were limited to $\phi < 0$. In other words, the equilibrium umbrella sampling procedure accesses regions of configuration space not sampled by the TPS calculations used to construct the database, and the neural network yields erroneous output when an input falls outside its original range. However, with no loss in predictive power, the structures harvested in the umbrella sampling procedure can be filtered or the Monte Carlo dynamics can be restricted a priori to limit the physical variables to their ranges in the training

set. As shown in Figure 4, the former eliminates the peak at $p_B \approx 0$ completely.

3.2. Reaction in Explicit Solvent. In this section we apply the GNN method to the $C_{7eq} \rightarrow \alpha_R$ transition in the presence of 198 TIP3 water molecules[33,35] subject to periodic boundary conditions in a cubic box of side length 18.4 Å. The stable state definitions were modified to reflect shifts in the locations of the minima on the free energy surface relative to the gas phase. Basin A was defined as $-180° < \phi < 0°$ and $-150° < \psi < 0°$, and basin B was defined as $-180° < \phi < 0°$ and $100° < \psi < 180°$ (Figure 2).[17] Addition of the solvent also slows the time to commit to a stable state. Six hundred thirty paths of length 3.3 ps were harvested with TPS, and a database of 2047 configurations (partitioned into a training set of 1947 configurations and a test set of 100 configurations) was prepared according to the procedure described in the Supporting Information.

As discussed above, no adequate characterization of this reaction existed previously, so the results obtained here are unbiased by prior knowledge of the forms of the coordinates to put in the database. Even though the nature of solvent participation in the reaction is unknown, there is evidence that water molecules are in some way directly involved.[9,17] Consequently, it is necessary to include coordinates that explicitly describe solute−solvent and solvent−solvent interactions in addition to the 102 internal coordinates introduced in section 3.1.1. These descriptors fall into four categories: (1) solute metrics pertinent to the solvation free energy, (2) quantities that characterize the two water molecules nearest the solute, (3) spherical-polar grid-based densities, and (4) decompositions of the energy function and its derivatives. The first category includes only the accessible surface area (1.4 Å radius probe), radius of gyration, and excluded volume of the solute (there were 53 such coordinates because different groups of atoms and cutoffs were tried). The other three categories are discussed in greater detail immediately below and in the Supporting Information.

For the second category we calculated the shortest distance between each solute atom and the three atoms in each water molecule. For each solute atom we then identified the two nearest water molecules and determined the three possible pairwise distances, the six possible pseudobond angles, and the three pseudodihedral angles characterizing the position of each water molecule relative to the solute atom. Since there are 12 solute atoms, this procedure results in $12 \times 2 \times (3 + 6 + 3) = 288$ pseudointernal coordinates.

Intrinsic to the second set of solvent descriptors is the problem that not only exchange with the bulk but small fluctuations can lead to changes in the identities of the two water molecules closest to each solute atom. This problem can be avoided by tabulating statistics for volumes of space around each solute atom. Specifically, the third category includes the numbers of oxygen atoms and hydrogen atoms and the net charge in solvation shells subdivided into angular regions; the specific radial and angular intervals considered are described in the Supporting Information. In total, 4096 such coordinates were added to the database. Similar coordinates were introduced in ref 9, but in contrast to that study, we did not examine the flux between grid cells because it is not well-defined for a single configuration of the system.

The last category of physical variables in the database involved decomposition of the energy function and its derivatives. In addition to the total solute intramolecular energy, we separately considered the van der Waals and Coulomb electrostatic interactions between selected groups of solute and solvent

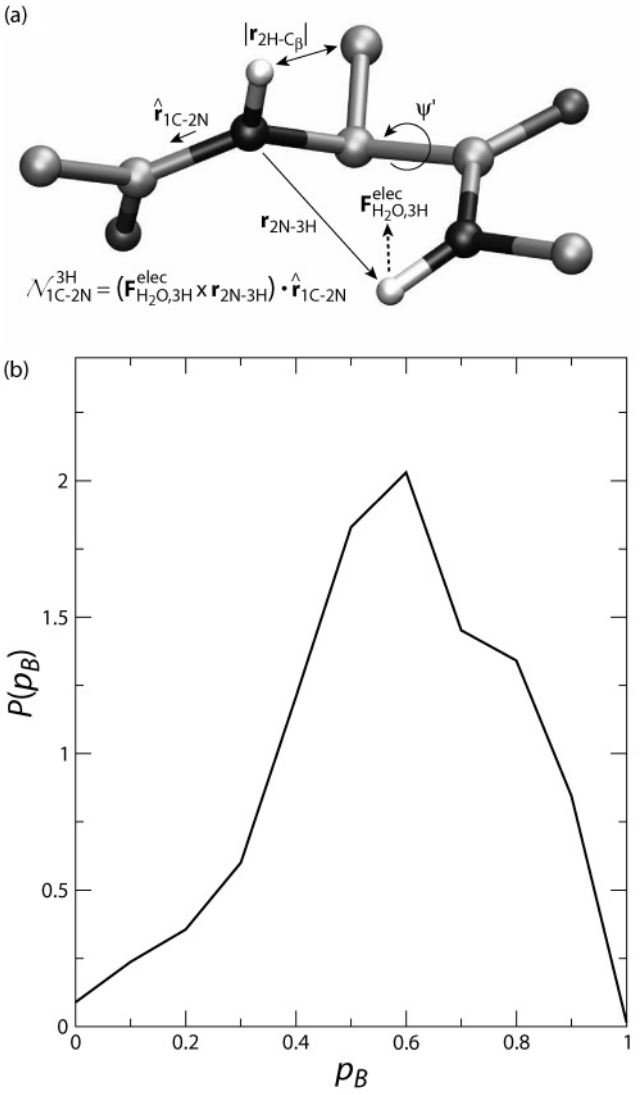

Figure 5. (a) Selected instantaneous descriptors for the $C_{7eq} \rightarrow \alpha_R$ transition in explicit solvent (Table 1). The structure shown is a transition state. $\mathbf{F}_{H2O,3H}^{elec} = -\nabla_{3H}\Sigma_i E_{i,3H}^{elec}$, where $E_{i,3H}^{elec}$ is the Coulomb electrostatic energy between atom i and solute atom 3H, the gradient is with respect to the position of atom 3H, and the sum runs over all solvent atoms. (b) Committor distribution for configurations harvested with a model corresponding to the descriptors indicated in part a.

atoms as detailed in the Supporting Information (1132 physical variables in total). Information about the forces associated with the solute−solvent interactions was incorporated into the database through the corresponding torques around main-chain bonds (141 coordinates). To the best of our knowledge, such torques have not been examined previously with regard to the reaction in question. However, like internal coordinates, they obviate the need to define arbitrary local coordinate systems for aligning configurations (as above and in ref 9).

The GNN procedure applied with the database described above yielded consistent one-, two-, and three-descriptor models (Table 1). Adding a fourth input did not reduce the RMS error in p_B. The best three-descriptor model has an RMS error of 0.133. It is composed of the dihedral angle $2C_\beta-2C_\alpha-2C-2O$ ($\psi' \approx \psi - 60°$ with a margin of error of approximately 10°), the distance $2H-2C_\beta$, and the torque around the 1C−2N bond associated with the electrostatic forces that the solvent exerts on the 3H atom ($\mathcal{N}^{3H}_{1C-2N}$) (Figure 5a). Because the intrinsic error in the calculation of the actual p_B derives from binomial

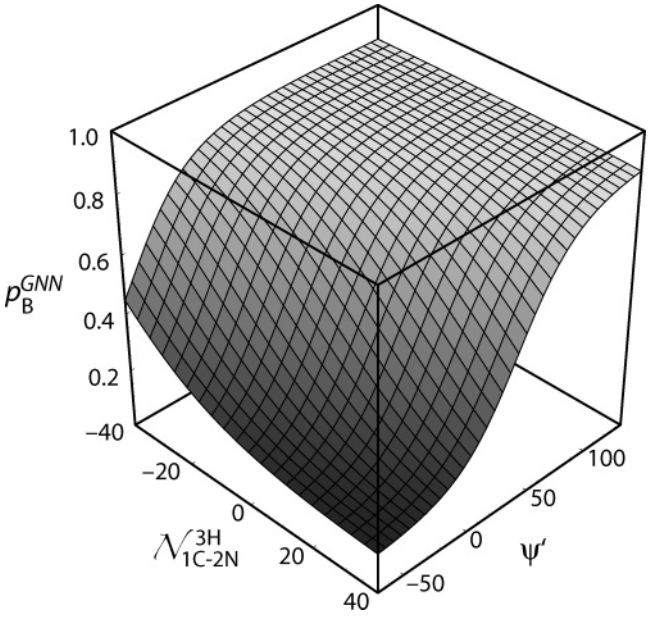

Figure 6. Predicted p_B for the $C_{7eq} \rightarrow \alpha_R$ transition in explicit solvent. The neural network used to harvest configurations in Figure 5b was fired for equally spaced values of ψ', $\mathcal{N}_{1C-2N}{}^{3H}$, and $r_{2H-2C\beta}$ over their ranges in the training set. The data shown are for $r_{2H-2C\beta} = 3.0$ Å. The units of torque are kcal/mol/radian, and those of ψ' are degrees.

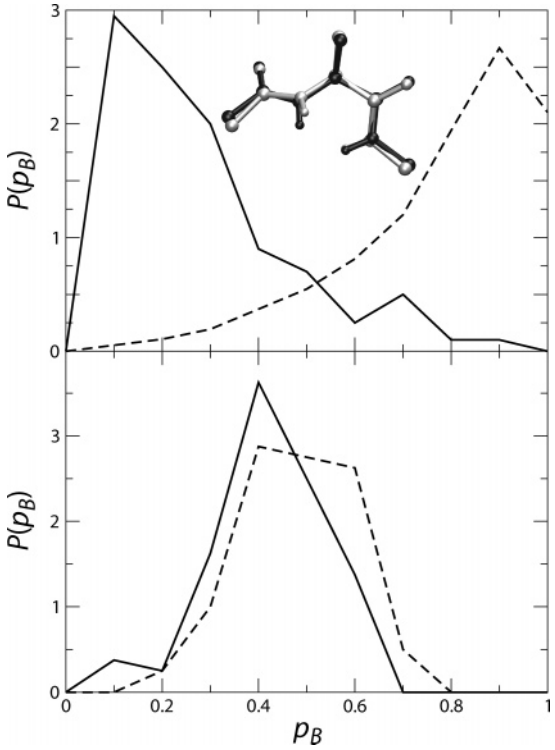

Figure 7. Committor distributions for ensembles of partially randomized structures. Two arbitrary transition states for the $C_{7eq} \rightarrow \alpha_R$ transition in explicit solvent were harvested, selected degrees of freedom were fixed, and the remaining degrees of freedom were randomized with molecular dynamics at 300 K (gray structure corresponds to solid line, and black structure corresponds to dashed line): Fixed solute, randomized solvent (top); fixed solvent, randomized solute (bottom).

statistics and 100 trajectories are used to calculate p_B, the limiting accuracy of the GNN procedure should be the same as in the vacuum cases (about 0.1). The fact that the RMS error is higher suggests that the characterization of the reaction is not complete; some secondary physical variable is not present in the database. Consistent with this idea, subjecting this model to the standard test results in a distribution that is broader than those obtained for the vacuum reactions (Figure 5b). Nevertheless, the shape of the distribution indicates that this combination of descriptors captures the essential dynamics. The GNN method thus represents a qualitative advance over previous approaches used to study the dynamics of this system[9,17] in that it is the first to yield a set of coordinates that is demonstrably sufficient to specify the transition state.

The GNN method identifies empirical relationships between dynamics and physical variables. As such, it is only the first step in analysis of a mechanism. In this regard, a very useful feature is that one can probe the functional dependence on the selected coordinates by firing the trained neural network. We carried out this procedure for the model in Figure 5 for input value ranges observed in the training set. As can be seen in Figure 6, the predicted transition states ($p_B^{GNN} = 0.5$) fall roughly along a diagonal line in the $\psi'-\mathcal{N}^{3H}_{1C-2N}$ plane. This dependence on both variables clarifies the behavior observed in ref 17; along $\psi' \approx 0°$, which corresponds to $\psi = 60°$ investigated in that study, the calculated committor values vary from 0.21 to 0.85. In the transition region p_B varies almost linearly with ψ' and $\mathcal{N}^{3H}_{1C-2N}$, consistent with the reasonably high accuracy obtained for this reaction with simple regression (Table 3 in the Supporting Information). However, in general, nonlinear fitting methods are needed (for example, see the corresponding results for the $C_{7eq} \rightarrow C_{7ax}$ transition in vacuum).

To better understand the role played by the selected coordinates, we performed the following numerical experiments. From the reactive trajectories we picked two arbitrary snapshots with $p_B = 0.5$, fixed the solute in its instantaneous conformation, and relaxed the solvent. For each selected transition state this procedure yields an ensemble of configurations for which we evaluated p_B in the absence of any constraints (Figure 7). The resulting distributions peak at the reactant and product basins. These data demonstrate that the solute configuration cannot by itself determine the committor, which is consistent with the identification of the torque by the GNN method. In contrast, the solvent configuration is sufficient to determine p_B. When an analogous experiment was performed in which we fixed the solvent and relaxed the solute, the values of p_B obtained clustered closely around 0.5 (Figure 7).

To confirm that the solvent influenced the solute dynamics primarily through electrostatic forces, we examined the distribution of dihedral angles sampled when the solvent was fixed and the solute was allowed to relax. These were compared with corresponding data obtained in the absence of the solute−solvent electrostatic interactions (but with the van der Waals terms) and in the absence of any interaction with the solvent starting with the same selected snapshots. As examples we show the distributions for the ψ torsional angle (Figure 8). Similar behavior was observed for other dihedral angles and for other starting configurations. The distributions obtained in the absence of electrostatic solute−solvent interactions differ remarkably from those with the full nonbond interactions and instead closely resemble those in vacuum. This correspondence suggests that the solute motion within the solvent cage is controlled mainly by the solute−solvent electrostatic interactions rather than by steric constraints.

It is important to stress that the present study is the first to show that the dynamics of the alanine dipeptide in aqueous solution depends on electrostatic, rather than steric, interactions with the surrounding water molecules. Monitoring the torque identified by the GNN method or the predicted p_B^{GNN} in individual trajectories revealed that stochastic water reorientation

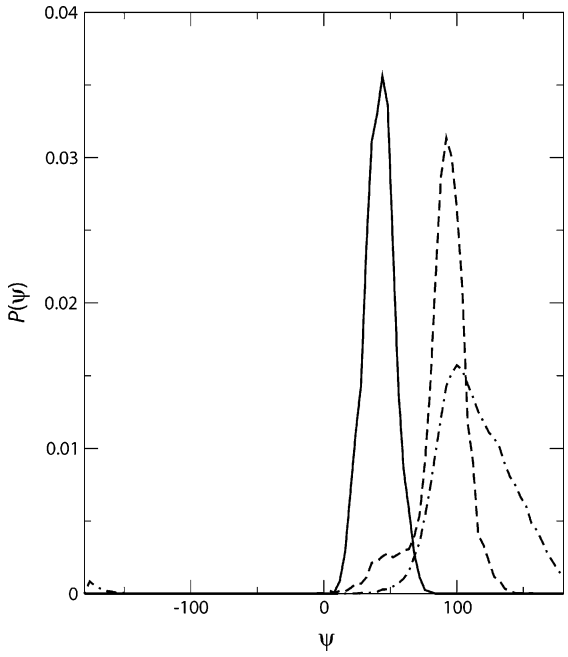

Figure 8. Distributions of the dihedral angle ψ in the presence of a fixed-solvent configuration from a representative transition state: Full potential (solid line); without solvent−solute electrostatic interactions (dashed line); vacuum (dot-dash line).

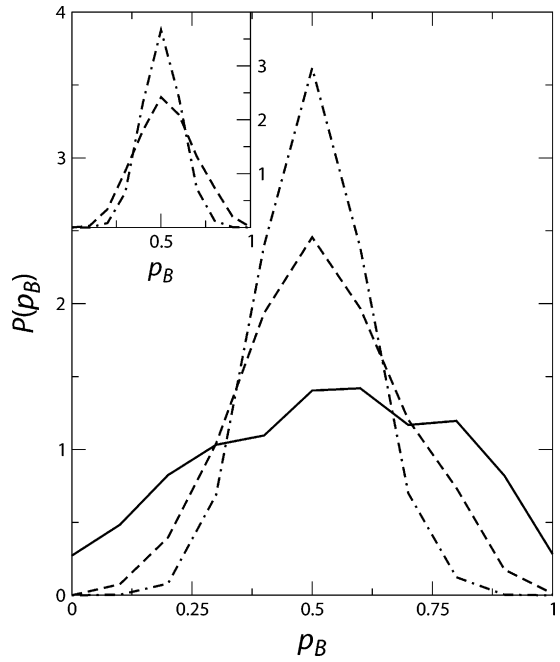

Figure 9. Committor distribution for the $C_{7eq} \rightarrow \alpha_R$ transition in implicit solvent. The ensemble of configurations was harvested with umbrella sampling with the restraint in eq 1 using the one- (solid line), two- (dashed line), and three-descriptor (dot-dash line) models in Table 1. (Inset) The ensemble of configurations was harvested with umbrella sampling with a separate harmonic restraint on each coordinate with the target values: $(\psi, \theta_2) = (38.26°, -3.29°)$ and $(\psi, \theta_2, \gamma) = (38.26°, -3.29°, -119.64°)$.

events as far away as the second solvation shell lead to fluctuations in these collective variables that permit the reaction to complete. However, water molecules in different places relative to the solute were responsible for these fluctuations in independent trajectories. The delocalized nature of the key solvent motions is consistent with the fact that the fluxes that McCormick and Chandler cite as important are observed in only 20−30% of the trials that they studied (see Figures 4 and 5 of ref 9) and explains their inability to use average populations of molecules in specific regions to predict transition states.

On the basis of the studies described immediately above, we added three sets of variables to the database. The first set consisted of the average solute internal coordinates obtained from equilibrium molecular dynamics simulations of the solute in the presence of the fixed solvent configuration for each structure in the database. The second set consisted of the corresponding averages of the torques considered previously, and the third set consisted of the averages of the torques obtained from simulations of the solvent in the presence of fixed solute configurations. With the expanded database, the GNN method yields consistent two-, three-, and four-descriptor models (Table 1). Not only is the RMS error in p_B improved but average coordinates are always selected over instantaneous ones. The shortcoming of such average coordinates is that they cannot be constrained readily in umbrella sampling simulations to apply the standard test. Consequently, they cannot be viewed as predictive, but they do shed light on the mechanism of isomerization in this system. Taken together with the numeric experiments, the preferential identification of coordinates that are averages over solute configurations in the presence of fixed solvent suggests that the solute dynamics are slaved to the relatively slow relaxation of the solvent cage. Unequivocal demonstration of this idea requires a time-correlation function analysis beyond the scope of the present study.

3.3. Reaction in Implicit Solvent. Because simulations in which the solvent is represented implicitly by terms in the energy function involve much less computational cost than ones with explicit solvent molecules, the former have become widespread. It is therefore of interest to compare the dynamics described above with those obtained with implicit solvent. Here, we study the alanine dipeptide with the analytical continuum solvent (ACS) energy term,[36−39] which is essentially a generalized Born method. To incorporate solvent friction effects, we used a Langevin integrator with friction coefficients based on the accessible surface area of each atom.[40]

TPS was used to harvest 1000 reactive trajectories. To reflect shifts in the locations of the stable states due to changes in the force field, basin A was defined as $-180° < \phi < 0°$ and $-110° < \psi < 30°$ and basin B defined as $-180° < \phi < 0°$ and $50° < \psi < 180°$ (Figure 2). The path length was set to 2.8 ps, which yielded an acceptance rate comparable to those in the TPS calculations above (about 15%). This commitment time is much closer to that of the reaction in explicit solvent than that of the reaction in vacuum, which suggests that the implicit solvent together with the selected representation of the friction captures certain features of aqueous dynamics.

From the reactive trajectories we constructed a database of 2500 configurations. The physical descriptors included were the redundant internal coordinates introduced in section 3.1.1. The best two-descriptor model identified by the GNN method combines the same two coordinates as were selected for this isomerization reaction in vacuum (Table 1), but the RMS error is slightly higher (0.113). The third variable, the improper dihedral angle γ (Figure 2a), lowers the RMS error to almost the level achieved in the vacuum case (0.090). Interestingly, in comparison with the gas-phase reaction, successive addition of the second and third coordinates results in more visible improvements in the prediction of p_B in both the test set (Table 1) and the standard test (Figure 9). Nevertheless, the mechanism in implicit solvent bears much more similarity to that in vacuum than that in explicit solvent, as expected from the fact that

variations in the positions of the water molecules in the latter results in qualitatively different dynamic behavior (Figure 7).

4. Discussion

In the present study we introduced a statistical method for identification of reaction coordinates in complex systems. Sample configurations distributed uniformly along a dynamic pathway are harvested, and their properties and committor values (p_B) are tabulated. Given this database neural networks are used to determine the functional dependence of p_B on sets of physical variables, and a genetic algorithm selects the combination of coordinates that yields the best fit. The GNN approach has been employed in other contexts;[25−29] the key insights in the present study are that p_B needs to be calculated only once for a set of representative structures and that a measure of the accuracy of a nonlinear statistical method for predicting p_B can serve as a scoring function for an importance sampling procedure that searches the space of physically meaningful variables relatively rapidly.

Application of the method involves several steps, and performance improvements in any of them will increase the overall efficiency. The protocol for the genetic algorithm and the minimization method for training the neural networks were not optimized for the specific problem at hand, so it is likely that they could be tuned. Transition path sampling is not intrinsic to the method but is probably the best available means for harvesting configurations with a uniform distribution of committor probabilities. Considerations for maximizing its efficiency are discussed elsewhere.[2,3] Because the calculation of p_B for the structures in the database is by far the most time-consuming step, its acceleration would result in the most significant gains in overall speed. A graph-based procedure introduced recently to estimate p_B directly from the initial set of reactive trajectories[41] could serve this purpose, but care would be necessary to avoid biasing the outcome of the GNN search through the choice of clustering method used to group configurations into nodes.

The method in its current form was capable of automatically combining nonobvious coordinates with more transparently relevant ones to reproduce results obtained previously by hand for the gas-phase $C_{7eq} \rightarrow C_{7ax}$ isomerization of the alanine dipeptide.[17] The secondary coordinates selected are dihedral angles that center on one of the relatively rigid peptide bonds, which suggests that the reaction dynamics are determined by a balance between the electrostatic interactions that stabilize the hydrogen bond and the cost to deform internal coordinates from their ideal geometries. Confirmation of this hypothesis requires more detailed mechanistic studies. In addition to identifying the most important physical variables, the method yielded an expression for p_B with predictive accuracy comparable to the intrinsic statistical error in that quantity. This functional relationship can be used to gain further insight into the role played by the selected coordinates.[27−29] Moreover, the fact that the method operates at the limiting accuracy allows models to be improved systematically by increasing the precision of the values in the database, in contrast to trial-and-error approaches.

The $C_{7eq} \rightarrow \alpha_R$ transition in aqueous solution was a challenging test. Because attempts to guess the variables that characterize the essential dynamics of the solvent had failed,[17] McCormick and Chandler introduced a systematic procedure based on average populations of water molecules in cells of a Cartesian grid superimposed on the system.[9] These authors argued that fluxes between particular cells correlated with reaction progress in 20−30% of the trajectories and that these motions were reflected in differences between stable states and transition states with respect to angularly restricted solvent radial distribution functions around certain solute atoms. However, these observations could not be used to harvest additional transition states to pass the standard test; instead, a bimodal distribution of committor probabilities peaked at $p_B = 0$ and 1 was obtained by picking configurations with a particular ψ dihedral angle value and low solvent occupancies in selected subvolumes.[9] Moreover, the nature of the solute−solvent coupling remained unclear because only local dipoles were examined and showed no predictive advantange.[9] Thus, prior to the present study the $C_{7eq} \rightarrow \alpha_R$ transition in aqueous solution remained poorly understood.

For the reaction described immediately above, the GNN method enabled us to test combinations of up to four coordinates drawn from a database of 5812 candidates to identify the first set of physical variables demonstrably sufficient to specify the transition state. A model based on the dihedral angle ψ', the pairwise distance between 2H and $2C_\beta$, and the torque around the 1C−2N bond arising from solvent electrostatic forces on the 3H atom gives an RMS error for the test set of 0.133 and a unimodal distribution centered on $p_B \approx 0.5$ when validated by the standard means. Subsequent expansion of the database to include coordinates that are averages over solute configurations in the presence of fixed solvent resulted in their preferential selection over instantaneous ones.

The electrostatic torque identified is a collective variable with contributions from water molecules beyond the first solvation shell. The importance of this coordinate was previously unappreciated and illuminates the shortcomings of the grid-based method introduced in ref 9. Because only a small number of water molecules must reorient or translate for the field to change, the key fluctuations are likely to be in different spatial regions in independent trials. Consistent with this idea, each of the fluxes identified by McCormick and Chandler appears in only 20−30% of the reactive trajectories studied.[9] The fact that the importance of any individual grid cell varies prevents the GNN from selecting these variables.

Full description of this complex reaction will require statistical characterization of the solvent motions that lead to changes in the electrostatic torque and a time-correlation analysis beyond the scope of the present study. However, numeric experiments presented in which either solvent or solute degrees of freedom were randomized suggest that the solute dynamics are slaved to relaxation of the solvent cage. This situation is reminiscent of electron transfer in polar solvents in the adiabatic limit. In that case, the reaction rate is dominated by the time required for the right solvent fluctuation to occur.[42] To the extent that this comparison is appropriate, it further accounts for the inability to predict transition states from a combination of the ψ dihedral angle and local solvent occupancies.[9] In the case of the dipeptide reaction, the solute configuration reports on the surrounding solvent structure. As a result, the solvent distribution functions are correlated with the internal coordinates, and only redundant information is obtained by examining both types of variables.

In an extensive analytical study van der Zwan and Hynes characterized several regimes of nonequilibrium solvation.[50,51] The model that they treated consisted of four linearly arranged dipoles that repelled each other with a dipole−dipole potential; the solvent was represented by the outer two dipoles, and the solute was represented by the inner two dipoles, which were also subject to an additional barrier to relative rotation. The physical variables that we identified for the explicit solvent isomerization indicate that it is in what van der Zwan and Hynes

call the "polarization cage" limit, in which the reacting dipoles are slaved by strong coupling to solvent dipoles with very large moments of inertia. Consistent with their observation that the reacting dipoles oscillate until the solvent dipoles relax and allow the reaction to proceed, the correlation times for ϕ and ψ in the presence of fixed solvent are about 0.5 ps, which is much shorter than the water dipole relaxation time.[43,44] This separation of time scales makes clear why average coordinates were selected over instantaneous ones in the present study.

The connection between the need to include solvent degrees of freedom in the analysis of a reaction and the time required for solvent reorganization relative to that for solute relaxation was explored by Truhlar and Garrett.[48] They studied a two-dimensional Grote−Hynes model[49] with solvent treated through an explicit degree of freedom. Parameters in the model were selected to correspond to the dissociation of an ion pair in solution,[16] with the coordinate other than the solvent degree of freedom taken to represent the distance between the ions. The two degrees of freedom were harmonically coupled to each other. At a high bath frequency (fast solvent) the solute coordinate was sufficient to determine p_B. At a low bath frequency (slow solvent) with the coupling constant increased to maintain the transmission coefficient at the same value the solute coordinate became inadequate for describing the dynamics; a bimodal committor distribution was obtained for phase space points at the solute coordinate value that corresponds to the maximum in the free energy profile projected onto that degree of freedom. This transition in distribution arises from the fact that as the coupling constant increases, phase space points with $p_B \approx 0$ and 1 increase in probability more quickly than those with $p_B \approx 0.5$ due to their relative Boltzmann factors.

Whether explicit consideration of solvent degrees of freedom is required will, in general, depend on the relative time scales of the key solute and solvent motions. We expect water to participate directly in the mechanisms of reactions that have solute relaxation times faster than the aqueous dipole fluctuation time, which is on the order of 10 ps.[43,44] These include the dipeptide isomerization studied here, proton[45] and electron transfers,[42] and ionic dissociation.[16] On the other hand, processes that involve large-scale motions of the solute but continuous local reorganization of the solvent will depend on the aqueous environment only in an average sense. For example, the transition state for the flipping of a terminal nucleotide in a three base pair oligomer was found to be described adequately by the interaction energies and distances between the unbinding base and its intra- and interstrand neighbors.[18] This is not to say that water is unimportant in this case; base stacking is driven by the hydrophobic effect,[46] and the commitment time is set by solvent friction.

We stress though that slow motions on the part of the solute do not in themselves exclude a structural role for water. For the collapse of a hydrophobic homopolymer in aqueous solution, ten Wolde and Chandler showed that the rate-limiting step is the nucleation of a large vapor bubble.[30] Thus, the time scale for certain solvent reorganization events can be very long. This case, in which both solute and solvent dynamics are slow, is likely to be representative of many molecular assembly and folding reactions. Care must be used in evaluating the importance of solvent in such processes. In a recent study of the folding of the 23-residue synthetic protein BBA5 it was argued that water does not play a structural role in the mechanism based on the observations that the collapse and drying of the hydrophobic core were concomitant and that committors were unchanged by solvent randomization.[47] However, these data are inconclusive for the following reasons. With regard to the former evidence, the radius of gyration actually decreases prior to formation of the vapor bubble in the homopolymer study discussed above.[30] As for the committor calculation, suppose that the region that separates the reactant and product basins is characterized by a free energy profile that is broad and flat (and possibly contains intermediates). Then even drastic changes to the solvent configuration will fail either to move the system out of the transition region or to prevent the system from returning to it relatively quickly in subsequent simulations used to evaluate the kinetic behavior. Unambiguous demonstration that collective solvent motions are not directly involved in this reaction will require at least observation of a unimodal committor distribution centered on $p_B = 0.5$ for a statistically significant number of configurations drawn from simulations in which the solvent is equilibrated for times longer than the overall commitment time in the presence of the fixed solute in structures corresponding to known transition states.

The identification of physical variables provides considerable mechanistic insight in itself. However, the method that we introduced could serve as the starting point for further quantitative analyses of reactions. For example, it might enable generalization of canonical transformation methods developed for the dynamics of simple systems[13] to more complex ones by facilitating integration over trivial coordinates to obtain an effective Hamiltonian. Alternatively, our method could be used to decompose a large system into component coarse-grained elements to access dynamics at long time scales.

Acknowledgment. We wish to thank Sung-Sau So for providing the HIPPO program and David Chandler, Stuart Rice, and Jiangbin Gong for critical readings of the manuscript. The research was supported in part by a Burroughs Wellcome Hitchings-Elion Fellowship to A.R.D. and the Louis Block Fund of The University of Chicago.

Supporting Information Available: Details of the GNN method, means of sampling p_B with uniform distribution, solvent descriptors, molecular dynamics simulations, umbrella sampling, and linear regression statistics for the models in Table 1. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids;* Oxford University Press: Oxford, 1987.
(2) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291.
(3) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. *Adv. Chem. Phys.* **2002**, *123*, 1.
(4) Elber, R.; Ghosh, A.; Cardenas, A.; Stern, H. *Adv. Chem. Phys.* **2003**, *126*, 93.
(5) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Appl. Phys.* **2003**, *93*, 2275.
(6) Zukerman, D. M.; Woolf, T. B. *Phys. Rev. E* **2000**, *63*, 016702.
(7) Dinner, A. R.; Karplus, M. *J. Phys. Chem. B* **1999**, *103*, 7976.
(8) Dinner, A. R.; Šali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. *Trends Biochem. Sci.* **2000**, *25*, 331.
(9) McCormick, T. A.; Chandler, D. *J. Phys. Chem. B* **2003**, *107*, 2796.
(10) Valleau, J. P.; Card, D. N. *J. Chem. Phys.* **1972**, *57*, 5457.
(11) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.
(12) Chandler, D. *J. Chem. Phys.* **1978**, *68*, 2959.
(13) Komatsuzaki, T.; Berry, R. S. *Adv. Chem. Phys.* **2002**, *123*, 79.
(14) Karplus, M.; Porter, R. N.; Sharma, R. D. *J. Chem. Phys.* **1965**, *43*, 3259.
(15) Karplus, M.; Sharma, R. D.; Porter, R. N. *J. Chem. Phys.* **1964**, *40*, 2033.
(16) Geissler, P. L.; Dellago, C.; Chandler, D. *J. Phys. Chem. B* **1999**, *103*, 3706.
(17) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877.

(18) Hagan, M. F.; Dinner, A. R.; Chandler, D.; Chakraborty, A. K. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13922.
(19) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495.
(20) Elmaci, N.; Berry, R. S. *J. Chem. Phys.* **1999**, *110*, 10606.
(21) Doruker, P.; Atilgan, A. R.; Bahar, I. *Proteins* **2000**, *40*, 512.
(22) Schmidt, M.; Rajagopal, S.; Ren, Z.; Moffat, K. *Biophys. J.* **2002**, *84*, 2112.
(23) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334.
(24) Northrup, S. H.; Pear, M. R.; Lee, C.-Y.; McCammon, J. A.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 4035.
(25) So, S.-S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
(26) So, S.-S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 5246.
(27) Dinner, A. R.; So, S.-S.; Karplus, M. *Proteins* **1998**, *33*, 177.
(28) Dinner, A. R.; Karplus, M. *Nat. Struct. Biol.* **2001**, *8*, 21.
(29) Dinner, A. R.; So, S.-S.; Karplus, M. *Adv. Chem. Phys.* **2002**, *120*, 1.
(30) ten Wolde, P. R.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6539.
(31) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
(32) Dinner, A. R. Monte Carlo Simulations of Protein Folding. Thesis, Harvard University, 1999.
(33) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902.
(34) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1.
(35) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(36) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578.
(37) Schaefer, M.; Bartels, C.; Leclerc, F.; Karplus, M. *J. Comput. Chem.* **2001**, *22*, 1857.
(38) Calimet, N.; Schaefer, M.; Simonson, T. *Proteins* **2001**, *45*, 144.
(39) Schaefer, M.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835.
(40) Smith, P.; Pettitt, B. M. *J. Phys. Chem.* **1993**, *97*, 6907.
(41) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415.
(42) Rossky, P. J.; Simon, J. D. *Science* **1994**, *370*, 263.
(43) Neumann, M. *J. Chem. Phys.* **1985**, *82*, 5663.
(44) Kurtović, Z.; Marchi, M.; Chandler, D. *Mol. Phys.* **1993**, *78*, 1155.
(45) Geissler, P. L.; Dellago, C.; Chandler, D.; Hutter, J.; Parrinello, M. *Science* **2001**, *291*, 2121.
(46) Chandler, D. *Nature* **2002**, *417*, 491.
(47) Rhee, Y. M.; Sorin, E. J.; Jayachandran, G.; Lindahl, E.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6456.
(48) Truhlar, D. G.; Garrett, B. C. *J. Phys. Chem. B* **2000**, *104*, 1069.
(49) Grote, R. F.; Hynes, J. T. *J. Chem. Phys.* **1980**, *73*, 2715.
(50) van der Zwan, G.; Hynes, J. T. *J. Chem. Phys.* **1983**, *78*, 4174.
(51) van der Zwan, G.; Hynes, J. T. *Chem. Phys.* **1984**, *90*, 21.