AIP | The Journal of Chemical Physics

A comparative analysis of clustering algorithms: O2 migration in truncated hemoglobin I from transition networks

Pierre-André Cazade, Wenwei Zheng, Diego Prada-Gracia, Ganna Berezovska, Francesco Rao, Cecilia Clementi, and Markus Meuwly

Citation: The Journal of Chemical Physics **142**, 025103 (2015); doi: 10.1063/1.4904431

View online: http://dx.doi.org/10.1063/1.4904431

View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/142/2?ver=pdfcov

Published by the AIP Publishing

A comparative analysis of clustering algorithms: O_2 migration in truncated hemoglobin I from transition networks

Pierre-André Cazade,[1] Wenwei Zheng,[2,a)] Diego Prada-Gracia,[3] Ganna Berezovska,[1] Francesco Rao,[3] Cecilia Clementi,[2] and Markus Meuwly[1,b)]

[1]*Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*
[2]*Department of Chemistry, Rice University, 6100 Main St., Houston, Texas 77005, USA*
[3]*School of Soft Matter Research, Freiburg Institute for Advanced Studies, Albertstrasse 19, 79104 Freiburg im Breisgau, Germany*

(Received 18 June 2014; accepted 5 December 2014; published online 12 January 2015)

The ligand migration network for O_2–diffusion in truncated Hemoglobin N is analyzed based on three different clustering schemes. For coordinate-based clustering, the conventional k–means and the kinetics-based Markov Clustering (MCL) methods are employed, whereas the locally scaled diffusion map (LSDMap) method is a collective-variable-based approach. It is found that all three methods agree well in their geometrical definition of the most important docking site, and all experimentally known docking sites are recovered by all three methods. Also, for most of the states, their population coincides quite favourably, whereas the kinetics of and between the states differs. One of the major differences between k–means and MCL clustering on the one hand and LSDMap on the other is that the latter finds one large primary cluster containing the Xe1a, IS1, and ENT states. This is related to the fact that the motion within the state occurs on similar time scales, whereas structurally the state is found to be quite diverse. In agreement with previous explicit atomistic simulations, the Xe3 pocket is found to be a highly dynamical site which points to its potential role as a hub in the network. This is also highlighted in the fact that LSDMap cannot identify this state. First passage time distributions from MCL clusterings using a one- (ligand-position) and two-dimensional (ligand-position and protein-structure) descriptor suggest that ligand- and protein-motions are coupled. The benefits and drawbacks of the three methods are discussed in a comparative fashion and highlight that depending on the questions at hand the best-performing method for a particular data set may differ. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4904431]

I. INTRODUCTION

The difficulty of characterizing long-time scale dynamics in complex systems that exhibit several states is a fundamental problem in chemistry and biology. With the increasing availability of computational facilities, it is now routine to generate extended trajectories for large macromolecules in explicit solvent and to explore the fundamental dynamics of biomolecular processes, including protein folding, enzyme catalysis, signal transduction, and ligand binding. It is therefore paramount to formulate and validate methods to analyze the spatial and temporal evolutions of the system dynamics.

One method which has attracted considerable interest in analyzing the dynamics by which a system moves from one state to another is based on a transition network analysis (TNA), also known as Markov state modeling.[1–4] TNs are discrete representations of states or clusters and contain information about the possible pathways between the states. In graph theoretical terms, the conformational states are represented by nodes or vertices, whereas the transitions between them correspond to the edges. The kinetics between the nodes can be recovered by a master equation[1] or by kinetic Monte Carlo (KMC) methods.[5–9] KMC is particularly suitable for situations in which the time scale separation between different motions of interest is large, such as in protein folding. TNs have found several applications in protein folding,[10–18] enzyme catalysis,[19,20] ligand migration,[21] and studies of electron spin resonance.[22]

For a transition network analysis, the original data set needs to be clustered. In the present work, three different clustering methods are applied to the same data set and the ensuing coarse grained dynamics is followed based on this clustering. The methods include k–means, kinetics-based, and collective variable (CV)-based methods. The k–means method is based on geometrically clustering the data around a set of predefined (here, based on previous experimental and computational studies[23–25]) or continuously updated centers (Sec. II B 1). The kinetics-based approach used here relies on the Markov Clustering (MCL) method which partitions space on a fine mesh. This provides a large number of microstates which are then kinetically lumped together into clusters (macrostates) (Sec. II B 2). Alternatively, CV-based approaches (Sec. II B 3) can be used to define the clusters. For more details on CVs, the reader is referred to a recent review.[26] LSDMap employed here preserves the so-called “diffusion distances” which can be understood as the “ease” to transit from one configuration to a neighboring one by considering the number of possible pathways between them, and therefore is

[a)] Current address: Laboratory of Chemical Physics, NIDDK, NIH, Bethesda, Maryland 20892, USA.
[b)] Electronic mail: m.meuwly@unibas.ch

suitable for quantitative estimation of the transition barrier and rate.[27]

Ligand migration in globular proteins offers the possibility to compare experimentally and computationally characterized pockets with the states found by the clustering methods. Some clustering algorithms (MCL, LSDMap) do not make any assumption about the nature of the states (cluster centers), whereas others do (k–means). Hence, if different independent clustering methods find the same or largely similar states, and if they furthermore agree with experimentally characterized states, it is likely that they constitute a meaningful set of states of the system. It is of great general interest to benchmark different clustering methods on the same system for which experimental data are available for validation. One such protein for which the internal pockets have been analyzed is the truncated hemoglobin trHbN of *M. tuberculosis*.[23–25] The bacterial species is responsible for human tuberculosis and evades macrophage killing by neutralizing toxic agents, such as nitric oxide (NO) by oxidizing NO to nitrate (NO_3^-).[28,29] For this, O_2–access to the protein active site is required. The structure of trHbN is shown in Figure 1. Efficient NO detoxification in trHbN has been attributed to the presence of an almost continuous tunnel through the protein that ensures rapid ligand transfer to the heme where the chemical reaction takes place.[23,30,31] Hence, trHbN is an important example of the direct involvement of ligand migration in a physiologically relevant process.

The tunnels in trHbN consist of two orthogonal branches connecting the heme distal pocket to the protein surface at two distinct sites.[31] The crystal structure of trHbN with Xe atoms under pressure revealed five distinct docking sites along the two branches of the tunnel.[23] Furthermore, atomistic Molecular Dynamics (MD) simulations have provided a structural and spectroscopic characterization of NO and O_2 localization as physiologically relevant probes.[24,25] The resulting connectivity network provided insight into the ligand migration pathways and exit channels (Figure 2). While NO docks in the Xe2 pocket, O_2 preferably localizes in the DS2 pocket. Therefore, protein-ligand interactions appear to be *specific*.

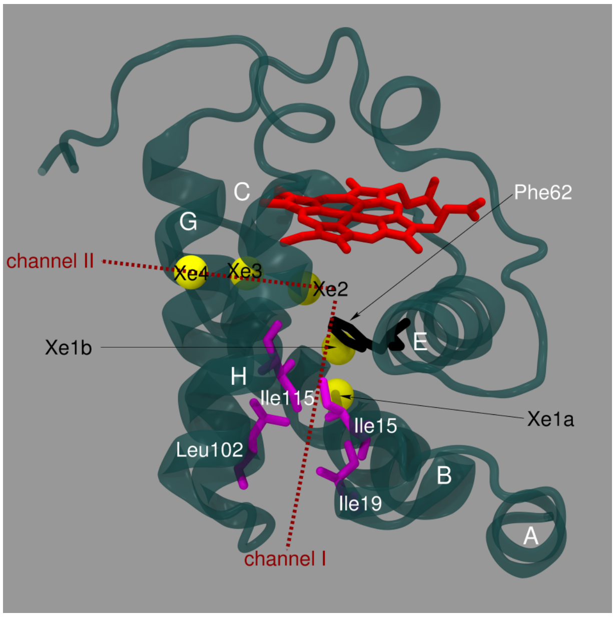

FIG. 1. X-ray structure of trHbN (PDB code 1s56 Ref. 23). The backbone is represented in petrol and the helices are labeled. The heme and Phe62 residues are shown in red and black sticks, respectively. The five xenon pockets found in the X-ray data are represented by yellow spheres. The residues around the channel I entrance are represented in magenta sticks. The red dashed lines indicate the two channels of the protein network.

O_2 is a particularly relevant physiological ligand and early studies of its migration were interpreted in terms of unspecific random diffusion.[32–34] More recent work found that O_2 follows specific tunnels (see above),[35–37] may involve multiple pathways, and is likely to be coupled (slaved) to the protein motion.[38–42] The possibility for a ligand to follow multiple pathways was, for example, demonstrated experimentally and through MD simulation for CO in Mb and O_2-migration in flavoenzymes.[38,43,44] The studies also suggest that small ligands can move through the bulky regions of the protein governed by thermal fluctuations of the protein and that ligand migration follows defined routes through the protein matrix.[42,45–60] Multiple pathways and active migration make a kinetic analysis of the ligand dynamics particularly relevant in view of the different time scales involved.

The purpose of the present work is to compare three different clustering algorithms applied to the same underlying MD data set characterizing the diffusion and extensive transition dynamics involving around 90 000 transitions of O_2 in NO-bound trHbN. The main questions asked in this work are (a) do different clustering algorithms applied to a close-to-complete microscopic sampling find the same “states” as characterized by their geometrical characteristics?, (b) how do the populations of the states compare?, and (c) does the kinetics of the states found from the three clustering algorithms differ and how closely do they trace the original reference trajectories? To answer these questions, the states involved are determined from different methods and then used in a transition network analysis. Finally, particularly important states in the network and their connectivity to the network are investigated in more detail.

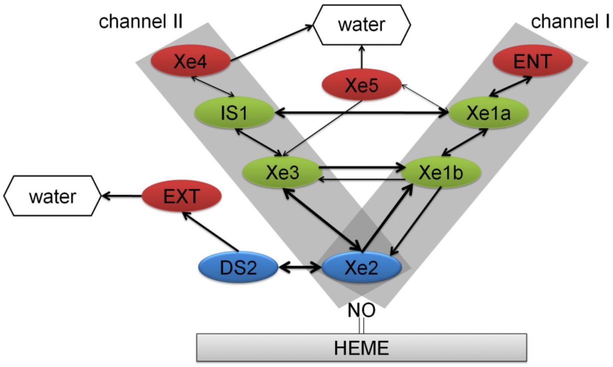

FIG. 2. Connectivity network of the O_2-ligand docking sites in trHbN. The observed ligand transitions among various protein pockets are indicated by arrows and line widths illustrate relative fluxes (thick for large flux, thin for small flux). The loss of ligand to the bulk solvent is indicated by arrows towards water. Docking sites which open toward the bulk (red), the “inner” network (green), and pockets around the reactive site (blue) are color-coded. The role of Xe3 as a hub is evident from this representation. The isolated PDS2 pocket is not represented for clarity.

II. METHODS

A. Molecular dynamics simulations

Details on the simulation conditions can be found in the previously published work.[25] Overall, 32 independent trajectories of 2 ns each are available for analysis from which the connectivity network is characterized. A simulation time of 2 ns for a trajectory was found to be sufficient for O_2 to extensively sample the entire network because free energy barriers between neighboring sites are typically around 1 kcal/mol or lower.[25] Moreover, beyond 2 ns of simulation, it becomes highly probable to find O_2 in the solvent as had also been found for NO in a previous study.[21,24] The runs are initially started in one of the 5 pockets identified experimentally[23,31] or from previous atomistic simulations.[24] From pockets Xe1, Xe2, and Xe3, 10 independent trajectories were started because these pockets were extensively sampled in exploratory runs. Sampling from pockets Xe4 and Xe5 was only carried out for one trajectory each because they are solvent exposed and ligands leave these sites on time scales of a few hundred picoseconds. Snapshots are written every 100 fs which leads to close to $6 \cdot 10^5$ snapshots analyzed involving more than 90 000 transitions.

B. Clustering methods

As a first step for a more coarse grained investigation of the O_2–dynamics within the protein, partitioning of the space visited by the unbound ligand into representative states is required. Various clustering methods and algorithms have been devised to address this issue.[61,62] In the following, we briefly summarize the different approaches employed in the present work, including k–means clustering,[61–64] MCL,[65] and the locally scaled diffusion map (LSDMap) methods.[66] All of them are applied to the same data set consisting of the Cartesian coordinates of O_2 from reoriented trajectories with the Fe-atom at the origin and the least-squares plane containing the four nitrogen atoms of the heme group of the protein in the xy–plane. The total number of configurations available is $6.4 \cdot 10^5$ of which $5.8 \cdot 10^5$ were analyzed, except for LSDMap, as explained below. The known xenon pockets are those identified in the X-ray structure of trHbN and complemented by MD studies[23,25] totaling 13 clusters: ENT, Xe1a, Xe1b, Xe2, DS2, EXT, Xe3, IS1, IS3, Xe4, Xe5, PDS2, WAT, and DUM.

1. *k–means clustering*

k–means clustering[61–64] aims at partitioning data space into Voronoi cells. In this approach, only the number of desired clusters is specified as input. The local optimal partitioning is found iteratively from a set of k clusters whose centers $\mathbf{m} = (m_{1x}, m_{1y}, m_{1z}, \ldots, m_{kx}, m_{ky}, m_{kz})$ need to be initially guessed. The algorithm can be summarized as follows:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \mathbf{m}_i\|^2, \tag{1}$$

with $\mathbf{x}$ the coordinates of the data points and $\mathbf{m}$ the coordinates of the k cluster centers. The explicit algorithm involves assignment of a new data point x_p to one of the sets $S_i (\forall\, 1 \le i \le k)$ according to

$$S_i^{(\nu)} = \left\{x_p : \left\|x_p - m_i^{(\nu)}\right\| \le \left\|x_p - m_j^{(\nu)}\right\| \;\forall\, 1 \le j \le k\right\} \tag{2}$$

and then updating the cluster centers $\mathbf{m}$

$$\mathbf{m}_i^{(\nu+1)} = \frac{1}{|S_i^{(\nu)}|} \sum_{\mathbf{x}_j \in S_i^{(\nu)}} \mathbf{x}_j, \tag{3}$$

where the variable ν loops over the iterations. The distance between the old $\mathbf{m}_\mathbf{i}^{(\nu)}$ and the new $\mathbf{m}_\mathbf{i}^{(\nu+1)}$ centers is used as the convergence criterion, and the clustering is considered to be converged when the sum of distances fulfills $\mathbf{m}_\mathbf{i}^{(\nu+1)} - \mathbf{m}_\mathbf{i}^{(\nu)} = 0$.

The procedure used in the present work is an extended version of the k–means algorithm in that two cutoffs, r_c^l and r_c^s, are used around each center. The larger cutoff, r_c^l, rejects configurations which are too far from any center to be included in the clustering (Eq. (4)), whereas the smaller cutoff, r_c^s, is used to determine which configurations will be used to optimize each center (Eqs. (5) and (6)). This is necessary to facilitate convergence and to obtain a more robust clustering. In what follows, the cutoffs are $r_c^s = 1.7$ Å and $r_c^l = 6.12$ Å. These values were chosen such that (a) the cluster centers remain close to the initial guesses from the X-ray structures and MD results and (b) the number of unassigned data (size of the "DUM" cluster) is lowest.

$$S_{\text{DUM}}^{(t)} = \left\{x_p : \|x_p - m_i^{(t)}\| > r_c^l \;\forall\, 1 \le i \le k\right\}, \tag{4}$$

$$S_i'^{(t)} = \left\{x_p : \left\|x_p - m_i^{(t)}\right\| \le r_c^s \;\forall x_p \in S_i^{(t)}\right\}, \tag{5}$$

$$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i'^{(t)}|} \sum_{\mathbf{x}_j \in S_i'^{(t)}} \mathbf{x}_j. \tag{6}$$

In general, the result of a k–means clustering depends on the initial guess and the number of cluster centers. For the present case, an obvious initial assignment uses the structurally observed pockets[25] described above. However, in order to scrutinize the final clustering performed along these lines, a clustering without initial guess (KMWG) was carried out. In this case, the first data entry was used as the first center $\mathbf{m}$.

2. *Markov clustering algorithm*

In contrast to k–means, the Markov CLustering Algorithm (MCL)[65] does not require an initial guess concerning the size and structure of the final clusters. The method is based on the analysis of a transition network which is obtained by (*i*) mapping the MD trajectory onto a discrete set of microstates; and (*ii*) building a transition network in which the nodes are the microstates and a link is placed between them if two microstates are visited one after the other along the trajectory.[67]

Within MCL, several methods have been proposed so far for the definition of these microstates, ranging from secondary structure strings,[67] RMSD clustering,[68] and order parameter fluctuations.[69] Before clustering, a transition network was built with the microstates corresponding to the cells of the rectangular grid (1 Å side) of the Cartesian region [−65.0, 65.0],[−65.0, 65.0],[−65.0, 65.0], each cell representing a microstate of the O_2 molecule. The discretization process represents a fine grained description of the molecular dynamics, resulting in a total of 24 664 microstates visited by the O_2 ligand along the 5.8×10^5 snapshots analyzed.

The discrete microstates time series built from the continuous O_2 trajectories were then mapped onto a transition network[67,70] with the properties described further below. In this network, nodes (i.e., microstates) are weighted with the total number of times they appear along the discrete trajectories. Links are added to the graph between any pair of nodes appearing successively, accounting this way for the total number of transitions observed. Finally, for each link detailed balance was imposed by averaging the number of transitions in both directions. Given the exhaustive sampling (98 163 transitions in total) that was possible in the present case, this was only partially necessary because the original trajectories mostly satisfied detailed balance already.

The application of the MCL algorithm on this transition network results in a set of kinetically homogeneous states. As is shown elsewhere, the resulting partitioning reflects the properties of the underlying free-energy surface where the clusters represent free-energy basins.[71,72] The kinetic model provided by these clusters satisfies the Markov property[73] which is not generally true for any type of clustering.[2]

MCL is based on the behavior of a random walker on the network guided by the transition probabilities. The algorithm can be summarized as follows: (*i*) build a transition matrix (TM) where each element T_{ji} represents the transition probability from node j to node i (i.e., sum over columns equal to one); (*ii*) compute $T^2 = T \times T$; (*iii*) take the pth power ($p > 1$) of every element of T^2 and normalize each column to one; and (*iv*) go back to step (*ii*). After several iterations, MCL converges to a matrix $T_{\mathrm{MCL}}(p)$ invariant under transformations (*ii*) and (*iii*). Only a few lines of $T_{\mathrm{MCL}}(p)$ have several nonzero entries that give the clusters as separated basins (usually, there is exactly one nonzero entry per column). Step (*iii*) reinforces high-probability walks at short time at the expense of low-probability ones. The parameter p determines the granularity of the clustering, see Figure 10. For large p, the random walks are likely to end up in small "basins of attraction" of the network, resulting in several small clusters. For $p = 1$, the clustering is not sensitive to any barrier and only one cluster covering the whole network is obtained. Thus, in free-energy language, the value of p determines to what barrier heights the algorithm is sensitive to. Small values of p split the network along the highest barriers.[71] As p increases, successively lower barriers are detected and the number of states increases, see Figure 10. For practical applications, p was found to typically vary between 1.1 and 1.8.[71,72] For dynamical processes with a clear separation of time scales, results are robust for different values of p.

3. Locally scaled diffusion map

The locally scaled diffusion map (LSDMap) is based on the kernel

$$K_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\varepsilon_i \varepsilon_j}\right), \qquad (7)$$

where $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the root mean square deviation between two configurations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and local scale ε_i represents the scale in configuration space around $\boldsymbol{x}_i$ within which the underlying manifold is almost linear and can be approximated by a hyperplane tangent to the manifold. The procedure to estimate this configuration-dependent local scale is detailed in previous work.[27] The exponential kernel K_{ij} can be considered as the "ease" with which $\boldsymbol{x}_i$ can diffuse into $\boldsymbol{x}_j$. A normalized version of this kernel approximates the Markov TM between pairs of configurations in the data set, and the eigenvectors of this matrix serve as the diffusion coordinates (DCs), representing the essential slowest motions of the system. LSDMap is useful to define Markovian clusters because configurations with barriers between them will be more distant in the low-dimensional CVs from LSDMap (diffusion coordinates) than in the original configuration space, and vice versa.

In order to render the calculation computationally manageable, a subset of the entire MD trajectories is chosen for LSDMap analysis by taking every fifth frame of the original trajectories. One of the original 32 trajectories is not included in the final analysis with LSDMap because most of the frames in that trajectory were far away from the frames of all the other trajectories. Similarly, 2 shorter trajectories were discarded giving a total of 29 trajectories. More detailed analysis on that trajectory suggests that most of the data points correspond to the less interesting situations in which O_2 is in the solvent instead of the protein. Therefore, 29 trajectories with each 4000 frames ($1.16 \cdot 10^5$ frames in total) are taken into account in LSDMap analysis.

The configurations represented by the first three DCs were clustered via the topological mode analysis tool ToMATo.[74] The method clusters the high-dimensional data set according to the spatial density function and merges the inconsequential clusters into noise. In the spatial density functions, the height of the peak is the density of the basin, and the position of the peak is the candidate cluster center. The cluster is filtered out as noise by checking the density difference between the height of the peak and the valley in the middle of the peak of interest and its neighboring peaks, that is, by checking the prominent part of the density peak. The free parameters used for ToMATo are the number n_N of neighbors to estimate the density, which was 200; the cutoff distance $r_c = 0.09$ Å in DC space to build the nearest neighbor network; and the threshold $\epsilon = 0.08\%$ of the number of points in the data set on the prominence of the density peak to consider a cluster as noise.

For each clustering, the method-specific parameters were varied to explore the sensitivity of the clusterings and to optimize discrimination between neighboring states. For KM and KMWG (see further below), initial tests without cutoffs lead to non-convergence. However, using an inner and outer cutoff for defining the clusters leads to ready convergence of the clusterings and changes of the present cutoffs away from the values used in the final clusterings modified the populations only in the percent range without changing the identity and number of clusters.

For MCL, an explicit example of the influence of the granularity parameter p on the clustering is given further below. For $p = 1$, the MCL algorithm detects 1 cluster. With larger values of p, the number of clusters increases but not all of them are statistically significant. The choice of p, which is usually between 1 and 2, is usually based on minimizing the number of clusters and the noise.

For LSDMap, the free parameter is a cutoff in the analysis of the noise spectra when determining the local scales.[75] It has been shown that the definition of states with LSDMap is robust to changes of this parameter (see Appendix A of Ref. 75). The results are also robust against changes of n_N since the data are densely sampled in the regions of interest. The value of r_c used to build the nearest neighbor network is directly related to the number of final clusters and the percentage of data clustered in the final states. When increasing r_c from 0.03 to 0.09 Å, the number of clusters decreases until it levels off at 0.08 and 0.09 Å. We choose $r_c = 0.09$ Å because over 99% of the data points are clustered, facilitating the comparison with the other two methods. With a suitable choice of r_c to build the nearest neighbor network, the results are also robust against ϵ to filter the noise.

C. O_2 Migration kinetic model

Once the clusters have been determined, they can be used as the "states" of a kinetic model by means of a transition network approach.

The transition network: The nodes of a transition network (TN) correspond to the states obtained from the clustering (see Sec. II B). A weight C_i is assigned to each node i accounting for the total number of times the system has visited that particular state. The total number of transitions $\tilde{C}_{ji}$ observed from state i to state j within two consecutive snapshots corresponds to the weight of the directed links of the network. Ideally, for equilibrium sampling, the transition count should be symmetric to satisfy detailed balance, i.e., $\tilde{C}_{ij} = \tilde{C}_{ji}$. However, due to finite sampling, these values are only nearly symmetric.[2,15,76] For this reason, symmetricity is imposed with the averaged weight for links $C_{ji} = (\tilde{C}_{ij} + \tilde{C}_{ji})/2$.

The network can be described by a single transition probability matrix $T(\Delta t)$ of the Markov chain with matrix elements

$$T_{ji}(\Delta t) = \frac{C_{ji}}{\sum_k C_{ki}}, \tag{8}$$

which represent the probability of the ligand to be found in state j at time $t + \Delta t$ given that it was in state i at time t. Hence, T is a row-stochastic matrix, given that

$$\sum_j T_{ji}(\Delta t) = 1, \tag{9}$$

and T describes the time evolution of a first order kinetic model according to

$$\vec{p}(t + \Delta t) = T\vec{p}(t), \tag{10}$$

where $\vec{p}(t)$ is the probability distribution along the states at time t. By construction, the stationary solution of Eq. (10) is given by the normalized occupation

$$T_i = \frac{C_i}{\sum_j C_j}. \tag{11}$$

The TN described above with its compact matrix representation captures the temporal evolution of the entire system. As such, a Markov model can be built to describe ligand migration in trHbN with the states provided by the three clustering methods: k–means, MCL, and LSDmap.

Validation through first passage time distributions: The First Passage Time (FPT) distribution corresponds to the distribution of times to reach a given target state from any other snapshot of the trajectory. In other words, the distribution of the time the system needs to be in state j at $t + \tau$, given that it is in any state i ($i \neq j$) at time t.

In order to compare the FPTs collected from the original MD trajectories, with those from the transition network, a random walk was run on the TN derived from each of the three clusterings. This random walk generates a discrete stochastic trajectory depending only on the transition probabilities between clusters. Arrival times from the random walk depend on the definition of the target state only and not on the detailed clustering of the data. For the original MD trajectory, the target state is defined as snapshots which correspond to the network state of interest obtained with a certain clustering method while for the random walk trajectories the target is represented by a network state as given in Table I derived from a particular clustering method.

III. RESULTS

In the present work, the dynamics of unbound O_2 in trHbN sampled from MD simulations has been analyzed following the different methods outlined in Sec. II.

A. Structural characterization of the states

Clustering from k–means: For k–means, two strategies were pursued. In one, the clustering was initialized from centers found either from experiment (Xe-pressurized X-ray scattering[23,31]) or from atomistic simulations.[24,25] They correspond to cavities naturally present in the protein. In the other, KMWG, initialization of the clustering only depends on the first data entry. In both cases, the quantity of interest is the population of each state (see Table I) and its geometrical center. The most populated states are DS2 and EXT which correspond to the main docking site for O_2 in the reactive site and to an exit route around the reactive site, respectively. The next most populated clusters are Xe1b, Xe1a, and ENT all along channel I which was found to be the uptake route for O_2.[25] From the main clusters (those found in X-ray experiments), Xe3 is the least populated one which is in accord with previous studies[24,25] and identifies it as a metastable state.

Clustering from MCL: As a result after analyzing the 29 trajectories (20 000 frames each), the network consists of 24 664 nodes and 98 163 links. Two weakly connected (depending on the details of the clustering) components are found: 62 nodes with 3.4% of the total weight and 24 602 nodes with the remaining population. The small component (PDS2) appears because there was a trajectory visiting a region of the Cartesian space not visited by any of the other 28 trajectories.

TABLE I. Comparison between the clusters found by KM, KMWG, MCL, LSDMap clustering, and experiment.[23]

	KM		KMWG		MCL (p = 1.6)		LSDMap	Expt.[23]
Label	r (Å)	Pop. (%)	r (Å)	Pop. (%)	r (Å)	Pop. (%)	Pop. (%)	r (Å)
ENT	16.6	7.8	16.3	5.6	16.1	5.9		
Xe1a	14.1	8.6	15.7	6.9	14.1	15.3[a]	40.3[b]	13.5
Xe1b	11.9	10.8	10.2	8.7	11.2	11.6	19.6	
Xe2	6.6	9.7	...	...	6.0	8.9	7.5[a]	6.4
DS2	5.4	14.2	6.0	21.3[b]	6.0	15.6	9.6	
EXT	8.1	12.8	5.6	6.1	5.9	7.2	11.2[a]	
Xe3	9.3	6.3	9.8	6.4	10.1	9.1[b]	0.2	10.9
IS1	12.9	8.5	13.9	11.8[a]	...	0.8		
IS3	11.1	2.7	...	...				
Xe4	15.8	7.2	16.0	4.7	14.8	3.1[a]	10.2	15.8
Xe5	11.6	4.4	8.9	6.5	8.7	1.7		
PDS2	5.1	4.6	6.7	3.4	5.7	3.4	4.6	
WAT	...	5.5	...	4.2			3.4[c]	

[a] Two clusters were merged.
[b] Three clusters were merged.
[c] Five clusters were merged.

MCL applied to the network described above with granularity $p = 1.6$ yields 2358 clusters (i.e., states) with different populations. A large number of clusters with low population are found by the algorithm due to finite size sampling. This can be observed cumulatively representing the population of the states (Figure 3). As such, only the 16 most populated clusters representing ≈ 84% of the total number of snapshots were retained. The 17th cluster in the ranked list, which is the first discarded, has a negligible population of 0.7%.

Clustering from LSDMap: Clustering of the data with LSDMap leads to ≈ 99% of the analyzed microstates assigned to 16 clusters. We did not cluster 100% of microstates because some of the events are in low populated minima of the LSDMap free energy landscape and increase noise (Figure 4). The 1st and 2nd DC characterize the motion of O_2 inside the protein while the higher-order DCs correspond to O_2 diffusing out of the protein towards water. The free energy profile as a function of the 1st and 2nd DC is shown in Figure 5. The barriers indicated qualitatively agree with previous free energy simulations.[25] However, it should be pointed out that connectivity along configurational coordinates—as exhibited in umbrella sampling simulations—may lead to very different free energy landscapes compared to connectivity along diffusional coordinates. In other words, depending on the collective variable used for projecting low-dimensional representations of the full free energy surface, the shapes and barriers may considerably vary as was previously found for protein folding.[77]

Figure 4 reports a two-dimensional representation of the data. Direct inspection of the clustered data suggests that the minimum at negative 1st DC corresponds to DS2/EXT

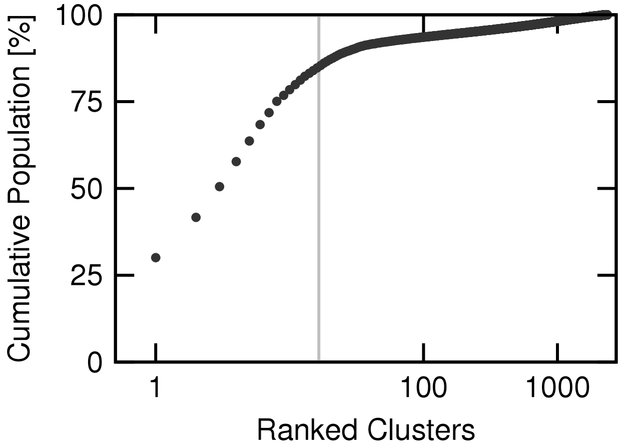

FIG. 3. Cumulative population of the MCL clusters. The 16 most populated clusters—before the gray vertical line—represent 84% of the total number of snapshots contained in the O_2 set of trajectories.

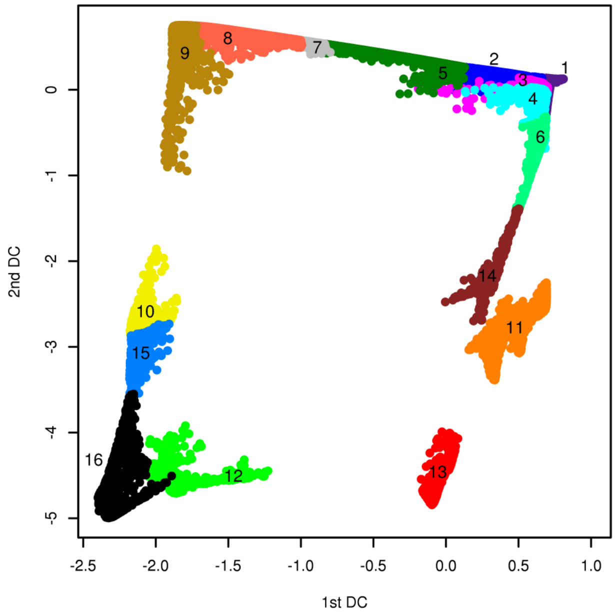

FIG. 4. LSDMap free energy profile and clustering plot. The color dots show the configurations in different clusters with respect to 1st and 2nd DC.

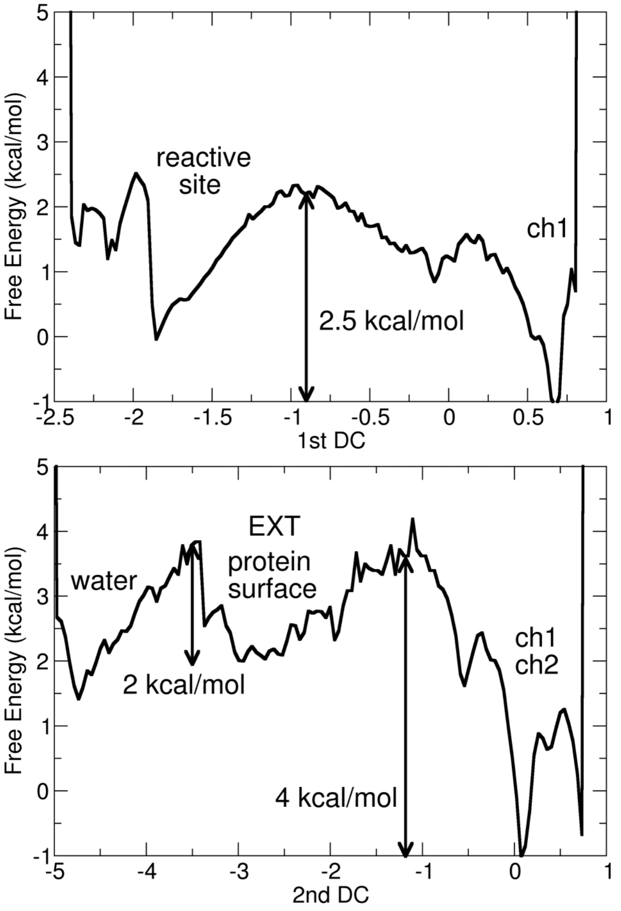

FIG. 5. LSDMap free energy as a function of the 1st DC (top) and the 2nd DC (bottom).

states (reactive site), whereas the minimum at positive values corresponds to a large basin involving ENT, Xe1a, and IS1. Hence, the 1st DC corresponds to the transition between the entrance of channel I and the reactive site. The free energy barrier is about 2 kcal/mol and is consistent with experiments on O_2 in flavoenzymes, which suggests that the barriers are sufficiently low to overcome at room temperature.[38] This is also in agreement with the computational results from previous free energy simulations (umbrella sampling) for this migration pathway.[25] States ENT, Xe1a, and IS1 are also found to overlap nicely with facile exchanges between these pockets separated by barriers of ≈ 1 kcal/mol. States 11, 12, and 13 at the negative values of 2nd DC are near the other exit route of O_2 from the protein. Therefore, the 2nd DC corresponds to the transition from states including Xe1, Xe5, Xe4, and EXT (minimum at positive 2nd DCs) towards water (minimum at negative 2nd DCs). These are consistent with the four exit routes observed in a previous work of the protein network,[25] in which the ligand leaves the protein in a multiple step process with barrier between 2 and 4 kcal/mol.

Comparison: Table I summarizes the results of the three clustering methods. The 13 clusters found by k–means (including the solvent state WAT) are reported with the distance to the Fe atom of the heme group and the population of each cluster. k–means clustering starts from initial guesses for the cluster centers which are then continuously updated as clustering proceeds. The initial guesses were the experimentally and computationally characterized pockets from previous work. Hence, the k–means centers and clusters are taken as the reference to be compared with the MCL and LSDMap methods. It is now of interest whether less biased methods, such as MCL and LSDMap, find states at comparable protein locations. In order to investigate the dependence of the k–means method on the initial guess, k–means clusterings without pre-assigned cluster centers ("k–means without guess" KMWG) were carried out using the same cutoffs as for k–means with a guess for the centers: the outer cutoff ($r_c^{\text{out}} = 6.12$ Å) decides whether or not a microstate belongs to the cluster and the inner cutoff ($r_c^{\text{in}} = 1.7$ Å) decides whether or not a microstate contributes to updating of the cluster center. Thus, a new cluster with its center at the microstate coordinate is created when a microstate is separated from any other center by more than r_c^{out}. As for the original algorithm, cluster centers are updated by averaging the coordinates of the microstates within r_c^{in} around the current center. At the end of every cycle if the clustering is not converged, empty clusters are removed, while clusters with centers at a distance less than r_c^{out} are lumped together, with a new center defined as the average of the other two. The procedure is repeated until convergence is reached. This results in a total of 267 clusters among which only 17 have a population larger than 1%.

These 17 clusters were analyzed and compared to the k–means clustering with initial guess in the same fashion as was done for MCL and LSDMap. The results are summarized in Table I. The two k–means methods yield similar results. KMWG allows to more easily assign individual events to a specific cluster for wide states such as ENT or Xe4 but lacks resolution for states close to each other as for Xe2, DS2, and EXT which are lumped together in one state and leads to the large population of DS2 and the non-existence of Xe2 which is not sufficiently separated from DS2. This could probably be improved by modifying the two cutoffs. However, the same parameters were employed for a one-to-one comparison between KM and KMWG, respectively. A comparison between the two clusterings suggests that using as a guess the experimentally and computationally known docking sites of the protein is a meaningful and beneficial procedure for this system.

In the following, the k–means, MCL, and LSDMap clusterings are compared. This is done by comparing the assignment of each frame to a specific cluster within a given clustering algorithm. Taking the k–means clustering as a reference, it is important to note that the frames of a k–means cluster can correspond to two or three different clusters in MCL and LSDMap, respectively, see Figure 6. Considering the Fe–O_2 distance, the locations and the populations of the Xe1b, Xe2, DS2, EXT, IS3, Xe4, and PDS2 states agree to within a few percent between k–means and MCL and k–means and LSDMap, respectively, as can be seen in Table I. This is also evident from the black squares at the intersections between these states in Figure 6. Three states—ENT, Xe1a, and IS1—are merged in one minimum with LSDMap. This shows that they are diffusionally similar but less so topologically. Within LSDMap, these states can be further separated by considering

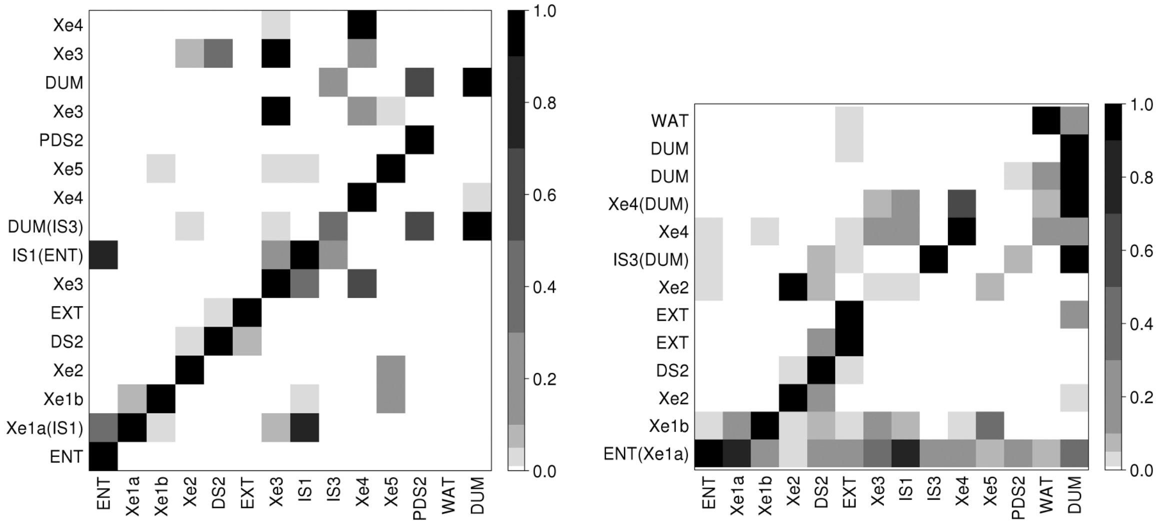

FIG. 6. Projection of the MCL (left) and LSDMap (right) clusters on the k–means clusters. For each MCL/LSDMap cluster (row), the projection is normalized on the maximum component. The scale is provided on the side of each plot.

higher-order DCs which correspond to more rapid motions. Similarly, MCL clustering finds that Xe1a and IS1, though structurally different, are in the same macrostate (Figure 6). This means that these states interconvert rapidly so that they appear as one cluster form a kinetic point of view. This is consistent with the small ≈ 1 kcal/mol energy barrier between these states.[25] Five distinct states defined by LSDMap correspond to WAT because the MD data with O_2 in the solvent are less sampled and disconnected. The main difference between k–means and LSDMap is that Xe5 disappears in the LSDMap analysis. This state is primarily sampled in the trajectory excluded from LSDMap analysis because most of the frames in that trajectory correspond to O_2 in water (see Sec. II). The other contributions of Xe5 originate from Xe1b and Xe2 in LSDMap, which suggests that these states are kinetically close. This is consistent with the results from MCL in which Xe5 from k–means contributes to a significant amount of Xe2 and Xe1b (Figure 6).

In Figure 6, the overlapping population between MCL clusters (left hand side) or LSDMap clusters (right hand side) and k–means clusters is shown. For MCL (see above) the clusters match nicely with those found by the k–means method. They are all defined with a single major component. This confirms that the computationally and experimentally determined pockets are a meaningful starting point for cluster centers. The only exceptions are ENT, Xe1a, and IS1 which appear lumped in the MCL approach, whereas they are well defined pockets in Cartesian space.

The population distributions $p(r_{\mathrm{Fe-O_2}})$ of the most important macrostates are explicitly reported in Figure 7. For this, the distance between the Fe and the free O_2 ligand was determined for every snapshot belonging to a particular cluster when using a specific clustering scheme. The data are arranged according to decreasing population as reported in Table I from left to right and top to bottom. The first observation is that all states found from k–means (red) and MCL (blue) overlap favourably (see also below). Therefore, the geometrical definition of the states found from k–means and MCL largely coincide. More detailed analysis of the most populated cluster—Xe1a and IS1—reveals that a small population at large Fe–O_2 separation (≈ 18 Å) in k–means is present. This part of the distribution structurally belongs to ENT. This is not the case for MCL since the similarity is based on kinetics rather than geometrical criteria. Conversely, LSDMap does not find separate clusters for the Xe1a and IS1 states. Instead, the most populated state contains almost all of Xe1a, IS1, and ENT, and small portions of several other states (see Table I and Figure 6). This is an entropic minimum containing quite heterogeneous structures, but the time scales of the motions inside this minimum are shorter than those corresponding to the first three DCs used in clustering. Structures with low Cartesian proximity are clustered for this state (a similar example is the unfolded state in protein folding). Therefore, a much broader structure distribution of this minimum as a function of the Fe–O_2 separation is found than with Xe1a/IS1 from k–means and MCL in Figure 7. On the other hand, k–means, MCL, and LSDMap largely overlap for clusters DS2, EXT, Xe1b, and Xe2. This could be further quantified by determining the overlap integral between the distributions. The fact that LSDMap does not detect Xe3 as a separate state is related to the very diffusive and dynamical nature of this site and its implied role as a hub.

A more quantitative analysis is carried out for the combined Xe1a and IS1 state (see left upper panel in Figure 7). For this, all events found by k–means and MCL, respectively, are considered. The total number of events is 113 920 out of which 77 268 are found by both methods. Hence, the two methods overlap for 68% of the events. The remaining 20 386 and 16 266 events are found by k–means and MCL, respectively. This is graphically illustrated in Figure 8 where events for the overlapping (red), k–means-only (blue), and MCL-only (green) distributions are reported. This representation also highlights a limitation in Figure 7 which suggests a somewhat larger overlap between the k–means and MCL-clustered data due to working with a one-dimensional descriptor (Fe–O_2 distance).

Contrary to k–means and KMWG, MCL, and LSDMap do not make assumptions concerning the spatial location of

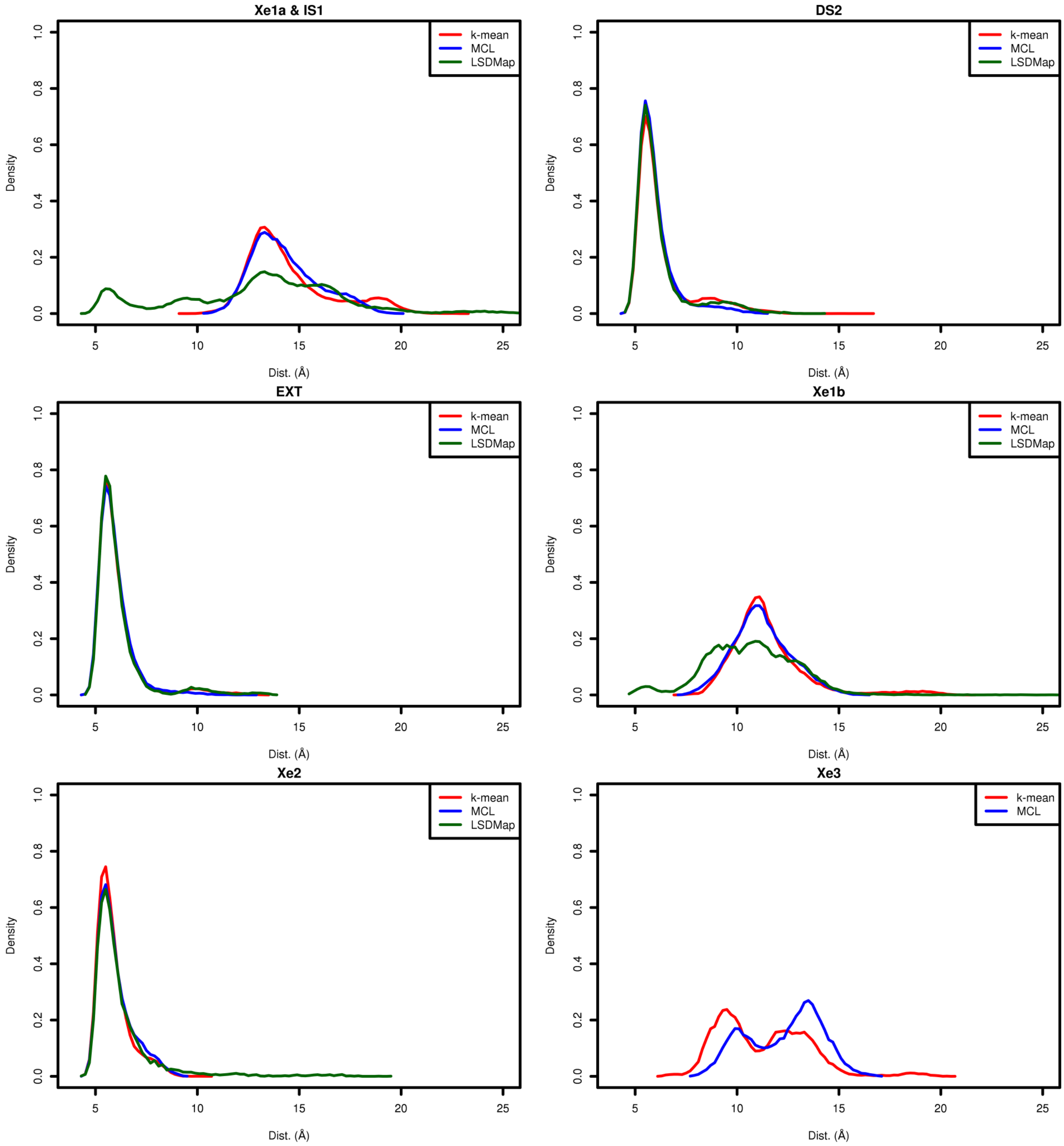

FIG. 7. Population density as a function of the Fe–O_2 distance. For each cluster, the distributions for k–means, for MCL, and for LSDMap are plotted as an histogram. Only the most populated macrostates are shown.

the cluster centers relative to the structurally known pockets. As one of the major conclusions of the present work—and to answer the first question formulated in the Introduction—it is found that the majority of cluster centers ("states") found from k–means, MCL, and LSDMap coincides with the known Xe pockets from the X-ray structure and from previous simulations. These methods verify the ability of the current MD simulation and various analyses to capture the major binding sites of the system from a kinetic point of view in addition to the structures. Also, the global minimum found in LSDMap corresponds to three structural pockets: ENT, Xe1a, and IS1 (see Figure 7). A similar correspondence is found for the MCL clustering for which Xe1a and IS1 are lumped together. These three states form a single basin from a free energy point of view which is consistent with the previous observation[25] of rapid exchange between those sites. Finally, all three methods find structurally related clusters but the difference is in the population, i.e., "size," of the clusters. This answers the second question put forward in the Introduction.

It is important to highlight that differences in the clustering primarily reflect the fact that different clustering schemes are sensitive to different underlying physical processes in the original trajectories. Each clustering (k–means, KMWG, MCL, and LSDMap) was stringently tested for convergence in its own right. Also, the underlying MD data are nearly converged, as a recent analysis of the coupled ligand-protein dynamics has revealed.[78]

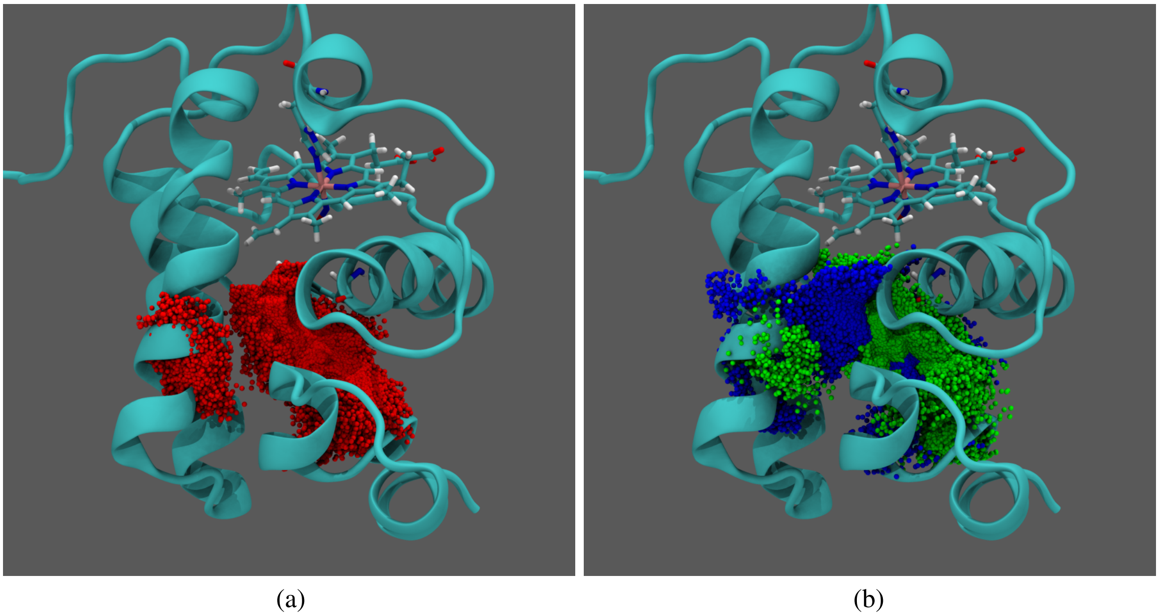

FIG. 8. Three-dimensional comparison of the Xe1a/IS1 states between k-means and MCL methods. (a) The overlapped structures of the two methods are shown in red. 77 268 of 113 920 structures (68%) are found in both methods. (b) The remaining non-overlapped structures of the two methods are shown in blue (18%, k-means only) and green (14%, MCL only).

B. Coarse grained dynamics

As a measure of the similarity between the original MD trajectory and the kinetic models based on k–means, MCL, and LSDMap clustering, the FPT distributions to the target state are studied. For this, the states DS2, EXT, and Xe1a are considered. The FPT for the MD trajectory was calculated as a distribution of times to reach a given target state from any time frame of the trajectory. For the networks, the FPT was obtained by running a random walk with 10^6 steps on the TM.

Figure 9 reports the FPT-distributions for the MD trajectory (red curve) and the random walk on the network (grey and black curves). The panels corresponding to clustering methods based on k–means, MCL, and LSDMap are denoted by KM, MCL, and LSD, respectively. Typically, only the shortest component of the population decay is accounted for by the random walk, whereas the slower components are not correctly reproduced with a one-dimensional descriptor. Overall, FPT-distributions from the MCL-clustering better trace the MD data. The standard deviation (grey shading) of the FPT from MCL clustering is significantly larger than that for the other two. This might be due to the significantly larger number of nodes in the MCL network, which is 2358, compared to 14 and 16 for k–means and LSDMap, respectively.

To answer the third question formulated in the Introduction, the clustering and FPT derived from sampling the transition matrices do differ quite significantly from the FPT distributions based on the MD simulations, depending on the state considered. None of the clusterings is able to faithfully capture all aspects of the dynamics from the atomistic simulations when using a one-dimensional descriptor. MCL correctly describes the short-time dynamics while missing certain features on the several hundred picosecond time scale. K-means and LSDMap are less suited to follow the kinetics. One of the reasons for this is the fact that only a one-dimensional descriptor was used.

In order to further explore this point and to quantify the sensitivity of the ensuing coarse grained dynamics on the parameter p employed in MCL clustering, a two-dimensional clustering including the protein-RMSD has been carried out. The clustering using two descriptors (ligand position and protein-RMSD) has been described in detail before.[78]

Two clusterings using a two-dimensional descriptor (the O_2 coordinates and the protein-RMSD relative to the X-ray crystal structure (Protein Data Bank entry 1IDR)) with MCL were carried out, one for $p = 1.12$ and the other for $p = 1.13$. The effect of a different granularity parameter p is to modify the barrier separating two basins from each other, as illustrated in Figure 10(a). This leads necessarily to a different partitioning of the time frames between the clusters and thus to different clusters and their populations, as is schematically shown in Figure 10(b). With increasing p, state 10 (left part of panel (b)) decomposes into states 16, 32, and 46 (the numbering is arbitrary, right part of panel (b)) because the clustering becomes more sensitive to lower barriers. The difference between clustering with $p = 1.12$ (squares) and $p = 1.13$ is the fact that certain frames are not clustered at all (squares with crosses) with the larger value for p, or that new frames are included in the clustering (squares with circles). The corresponding distribution functions $P(\rho)$, where ρ is the protein RMSD, are reported in Figure 10(c). There is almost a one-to-one correspondence for the distribution functions from clusterings with different p, as is evident in the left panel of Figure 10(c): The sum of the probability distributions $P(\rho)$ for states 16, 32, and 46 with $p = 1.13$ (grey curve) is almost identical to $P(\rho)$ for state 10 with $p = 1.12$ (red curve). The coarse grained dynamics from the two clusterings is compared in Figure 10(d) (grey and red lines). The data

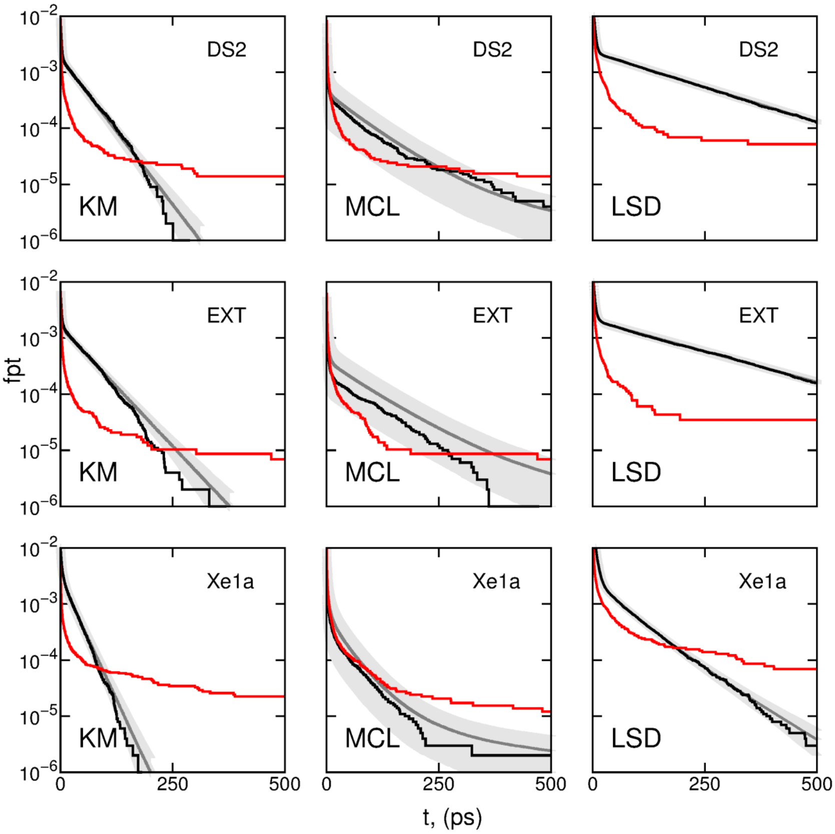

FIG. 9. FPT distributions for the states DS2, EXT, and Xe1a for each of the clustering methods: k–means on the left panels, MCL in the middle, and LSDMap on the right. Each panel contains the average FPT and standard deviation over 10^3 random walks each of length 10^6 on a given network (dark gray curve and light gray region, respectively). The black curves depict an exemplary random walk on the network, while the red curves correspond to the FPT distributions based on MD trajectories clustered according to k–means, MCL, and LSDMap clustering, respectively. Hence, even the FPT distributions from MD differ because they are based on different clusterings. For the random walks, FPT distributions are normalized by the length of the random walk while for the MD trajectories FPTs are normalized by the length of the trajectory.

reported are the FPT distribution averaged over 1000 RW trajectories and the corresponding standard deviation for $p = 1.12$ (dark and light red) and $p = 1.13$ (dark and light grey). The black line is the FPT distribution from the MD trajectory. Both coarse grained dynamics faithfully reproduce the short-time dynamics and favourably trace the rate of population change at later times. Comparing the results from Figures 10(d) and 9 (panel DS2, MCL) suggests that including the protein motion as a second descriptor has a non-negligible effect on how accurately the coarse-grained model is able to capture the real dynamics.[78]

IV. DISCUSSION

Ligand migration in TrHbN is an ideal test to compare and benchmark different clustering schemes because the cluster centers can be compared with experimental data and results from independent simulations and because the ligand migration network can be exhaustively sampled. This is in contrast to protein folding problems where typically only the structure of the native state is known and intermediates are difficult to characterize both, by experiment and computation. Furthermore, time scales for protein folding are sufficiently long to prevent one for carrying out a statistically significant number of folding trajectories. Also, ligand migration in trHbN involves about one dozen states which makes it different from other common test systems such as alanine dipeptide for which there are only a few states available.[17,27,71]

In the present case, the network can be rigorously sampled from atomistic MD simulations as the ligand migration barriers are known to be low[25] and several potential metastable states have been directly characterized in both experiments and previous simulations. A comparative assessment of the methods and particular practical aspects are discussed in the following. For this it is important to recall that the three clustering schemes are based on quite different assumptions. While k–means requires initial guesses for the cluster centers, MCL and LSDMap do not. However, with the structural data at hand, such a guess is much better defined compared to, e.g., the situation in protein folding for which a guess for meaningful intermediate (on-pathway intermediates) is much more difficult and potentially ill-defined. Hence, we do not expect k–means to fail primarily because of poor choices for the cluster centers in the present application. On the other hand, k–means and LSDMap start from a relatively small number of states (tens to hundreds), whereas MCL starts from tens of thousands of states and progressively reduces this number.

First, it is noted that all three methods find similar states. Specifically, the ENT, Xe1, Xe2, IS1, Xe4, DS2, EXT, and PDS2 states are found by all methods. The Xe3, Xe5, and IS3

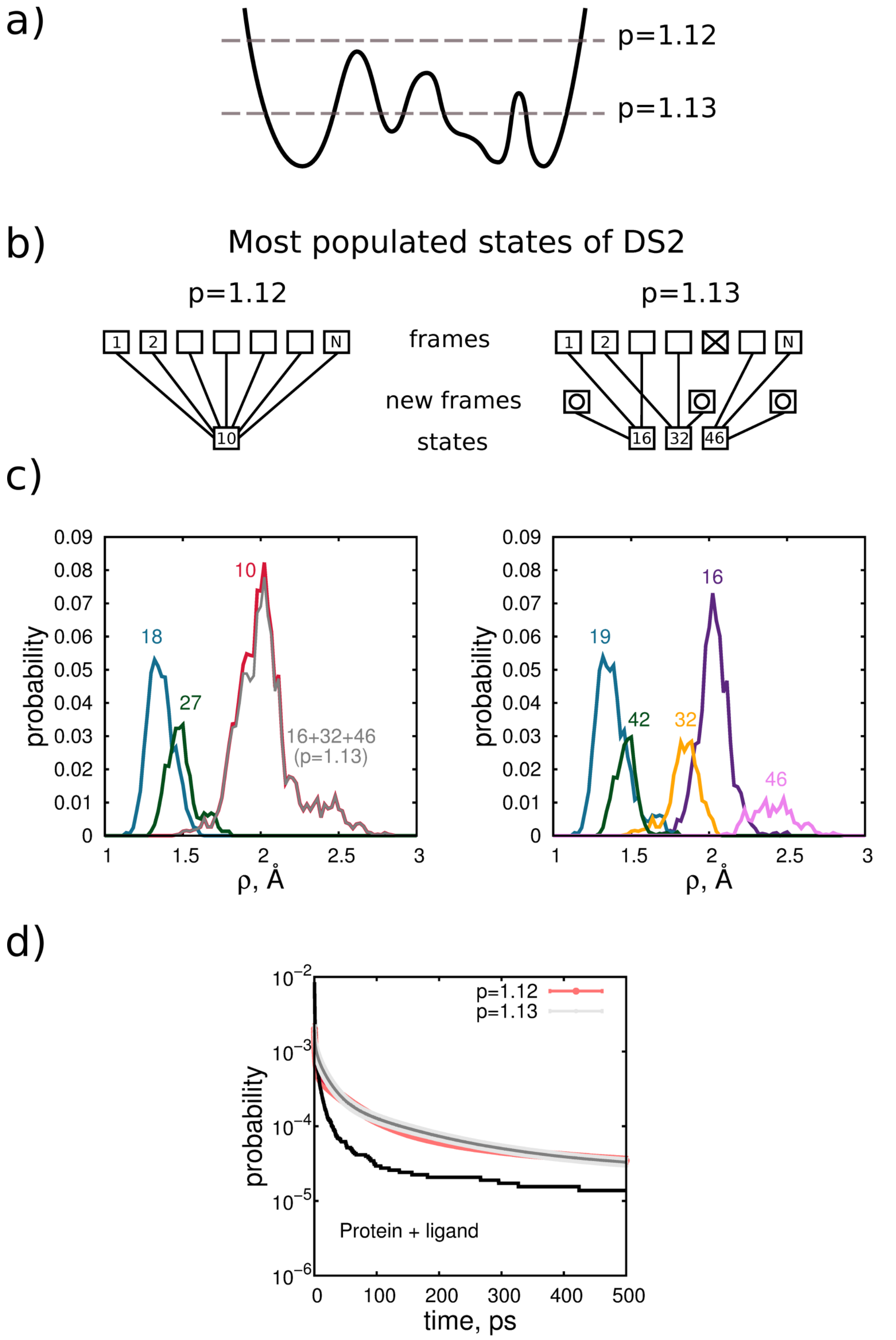

FIG. 10. Sensitivity of the first passage time distribution on the granularity parameter p in MCL for a two-dimensional extension of network analysis accounting for coupling of ligand and protein degrees of freedom. A detailed description is given in the text.

states are not present for LSDMap. As discussed above, Xe3 is too dynamic to be detected as a separate state with LSDMap, whereas Xe5 is only populated in one of the 32 trajectories which happens to be the one excluded from analysis with LSDMap because it mainly samples O_2 in water. This confirms its role as a major hub which needs to acquire and release the ligand(s) as they travel the network. It is also confirmed by the small free energy barrier by which this state is separated from the surrounding states, as was found from umbrella sampling simulations.[25] Comparing k–means and MCL—the two Cartesian-coordinate based clustering schemes—it is found that structurally, Xe1a, Xe2, Xe3, and Xe4 compare within 1 Å or better to the experimentally determined positions. The occupation of the most prominent states agrees favourably for all three methods, as evidenced in Figure 7 and Table I. Specifically, the comparison between k–means and MCL is noteworthy, whereas certain populations found by LSDMap are considerably larger (Xe1a) or smaller (Xe3) compared to the other two methods. It is found that clustering in diffusion space leads to a large primary cluster (population of 40%) which consists of three individual states found with k–means or MCL. Adding the populations of ENT, Xe1a, and IS1 from k–means or MCL yields a population of $\approx 25\%$ which is somewhat smaller than from LSDMap.

The dynamics of the states can also be compared from each clustering scheme. Using the states found from each clustering algorithm, the dynamics of the DS2, EXT, and Xe1a state from k–means and MCL largely overlap on the 500 ps time scale, whereas LSDMap gives somewhat slower dynamics. The main differences primarily occur for the long-time behaviour, see Figure 9. A network model that correctly describes the original MD kinetics presents a good match between the first-passage time distribution obtained directly from the MD trajectory and the one calculated from the network dynamics. This indicates that the network model minimizes the re-crossings on top of the barriers and that the original MD dynamics can be mapped into a Markov process. Our results suggest that the first-passage-time distributions determined from the transition matrices best cover the explicit MD results for MCL clustering. Conversely, k–means and LSDMap fail to capture the long-time dynamics.

In an effort to elucidate the findings from the present work for the non-expert of these types of data analysis, a brief list is compiled in the following.

- *K-means and KMWG:* This approach is highly intuitive. Once given the position of the cluster centers, all the snapshots of a trajectory are grouped to the closest centers. While in most cases clusters centers are randomly chosen (hence requiring thousands of centers to work with), the pre-knowledge of the binding pockets in the present case allowed us to select those as initial centers for the clustering. As it was illustrated in the paper, this approach gave a very nice characterization of the pockets. One caveat is that this method is purely geometrical, no information on the dynamics is used to group the snapshots. This leads to artifacts when considering the dynamics of the found clusters. In fact, small mismatches at the top of the barrier easily lead to re-crossings and a faster network dynamics. Hence, k–means is certainly an efficient method to group the snapshots of a MD trajectory into a set of pre-determined states (like in our case) but its use in building accurate/meaningful kinetics models is limited.
- *MCL:* The main advantage of this method is that it uses dynamics information to identify the states. As illustrated in the main text, MCL correctly identified the binding pockets without any *a priori* knowledge of their positions. Moreover, being based on the dynamics, it provided the best match to the original MD trajectory when it came to the comparison of the long time behavior of the first-passage-times distributions, see Figure 9. In our opinion, MCL has two main drawbacks. First, it requires a discretization of the original MD trajectory into a large set (i.e., thousands) of microstates or clusters of small sizes. For this step, many approaches can be used, like RMSD clustering[73] and k–means (with thousands of randomly placed centers) to a simple space discretization as in the present case. This step can be highly automatized, but in some cases results can be sensitive to the initial discretization step[73] or it is difficult to be performed.[79,80] Second, MCL requires a parameter p to be tuned which represents the level of granularity of the clustering. The smaller the parameter the higher the barriers separating the states. An advantage of MCL is that it assures that the found clusters are split along the barriers separating them, i.e., molecular states are never artificially split into two,[71] making the choice of this parameter only a question of the level of detail needed for the analysis.
- *LSDMap.* LSDMap uses short geometric distance (RMSD) to approximate local kinetic information (the “ease” with which one frame of structure can diffuse into another) and then merge them to extract a set of global, but geometrically unspecified reaction coordinates. LSDMap is largely independent of a geometrical reaction coordinate. The main limitation of using LSDMap as a clustering method is that it does not provide the definition of clusters directly, but instead a set of reaction coordinates characterizing different time scales of motions of the system. Here, on a first trial of using LSDMap to build TN, we used the clustering method ToMATo[74] on the reduced dimensionality space defined by LSDMap. The dynamics of the global minimum (Xe1a) is well characterized by the first two DCs, but the dynamics of some other states may require higher-order coordinates. The number of DCs used for multiple-minimum system and the choice of a topological clustering method on the DC space introduce another level of complexity. LSDMap involves the eigenvalue decomposition of an $N \times N$ matrix, with N is the number of frames for the analysis, and therefore is limited by the capacity of the computational resources. Here, only one-fifth of the entire data set was used for the analysis. However, it is found that clusters defined in this manner are largely consistent with the other two methods. The main advantage of LSDMap is that it provides a set of reaction coordinates decoupling the motions on different time scales, with the lower-order DCs corresponding to the slower motions. Therefore, for a system with a separation of time scales, the method serves as a meaningful way to obtain a low-dimensional free energy projection by using the first few DCs. Free energy barriers (e.g., barrier height of the migration pathway in this study) are usually well preserved as the rapid degrees of freedom are decoupled from the slow ones. The clusters on the free energy landscape also present a diffusion point of view on the stability of the states which sometimes differs from the structural view (ENT, Xe1a, and IS1). In addition, a simple visualization of relative position of the (meta)stable states in two dimensions help achieve a better understanding of the migration pathways and mechanisms of the system.

V. CONCLUSIONS

Starting from an extensive sampling of the O_2 migration in TrHbN, three different clustering schemes found largely identical docking sites (Xe1a, Xe1b, Xe4, and EXT) within the protein, which are also known from experiment and previous simulations.[23,25] The clusterings differ regarding states which are more diffusive or kinetic in nature and can be considered metastable from the diffusion point of view (LSDMap). All three clustering schemes find similar geometrical characterization of the states. However, the population of the states together with the kinetics between the states can differ considerably.

For a more complete assessment of the ligand dynamics, including the protein conformational degrees of freedom is likely to be relevant but outside the scope of the present work. The fact that protein and ligand dynamics is coupled to some degree is also suggested by previous work which found that the two cannot be easily separated for carbon monoxide in myoglobin.[42]

ACKNOWLEDGMENTS

Financial support from the Swiss National Science Foundation through funding of Project No. 200021-117810 and the NCCR MUST is gratefully acknowledged (to M.M.). C.C. and W.Z. thank the support from NSF (CDI-type I Grant Nos. 0835824, CHE-1152344, and CHE-1265929), and the Welch Foundation (C-1570). We would like to thank Sabyashachi Mishra for helpful discussions.

[1] N. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, Netherlands, 1981).

[2] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571–6581 (2004).

[3] F. Noé and S. Fischer, Curr. Opin. Struct. Biol. **18**, 154–162 (2008).

[4] F. Noé, J. Chem. Phys. **128**, 244103 (2008).

[5] D. A. Evans and D. J. Wales, J. Chem. Phys. **121**, 1080–1090 (2004).

[6] A. Bortz, M. Kalos, and J. Lebowitz, J. Comput. Phys. **17**, 10–18 (1975).

[7] A. Voter, Phys. Rev. B **34**, 6819–6829 (1986).

[8] K. A. Fichthorn and W. H. Weinberg, J. Chem. Phys. **95**, 1090–1096 (1991).

[9] W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy, J. Phys. Chem. B **113**, 11702–11709 (2009).

[10] S. Muff and A. Caflisch, Proteins **70**, 1185–1195 (2008).

[11] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U. S. A. **101**, 14766–14770 (2004).

[12] D. J. Wales, Mol. Phys. **100**, 3285–3305 (2002).

[13] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, J. Chem. Theory Comput. **1**, 515–526 (2005).

[14] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415–425 (2004).

[15] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).

[16] F. Noé, D. Krachtus, J. C. Smith, and S. Fischer, J. Chem. Theory Comput. **2**, 840–857 (2006).

[17] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).

[18] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou, J. Phys. Chem. B **108**, 6582–6594 (2004).

[19] S. Yang and B. Roux, PLoS Comput. Biol. **4**, e1000047 (2008).

[20] S. Yang, N. K. Banavali, and B. Roux, Proc. Natl. Acad. Sci. U. S. A. **106**, 3776–3781 (2009).

[21] S. Mishra and M. Meuwly, Biophys. J. **99**, 3969–3978 (2010).

[22] D. Sezer, J. H. Freed, and B. Roux, J. Phys. Chem. B **112**, 11014–11027 (2008).

[23] M. Milani, A. Pesce, Y. Ouellet, S. Dewilde, J. Friedman, P. Ascenzi, M. Guertin, and M. Bolognesi, J. Biol. Chem. **279**, 21520–21525 (2004).

[24] S. Mishra and M. Meuwly, Biophys. J. **96**, 2105–2118 (2009).

[25] P. -A. Cazade and M. Meuwly, Chem. Phys. Chem. **3**, 4276–4286 (2012).

[26] M. A. Rohrdanz, W. Zheng, and C. Clementi, Annu. Rev. Phys. Chem. **64**, 295–316 (2013).

[27] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).

[28] A. Couture, S.-R. Yeh, B. A. Wittenberg, J. B. Wittenberg, Y. Ouellet, D. L. Rousseau, and M. Guertin, Proc. Natl. Acad. Sci. U. S. A. **96**, 11223–11228 (1999).

[29] H. Ouellet, Y. Ouellet, C. Richard, M. Labarre, B. Wittenberg, J. Wittenberg, and M. Guertin, Proc. Natl. Acad. Sci. U. S. A. **99**, 5902–5907 (2002).

[30] A. Pesce, M. Milani, M. Nardini, and M. Bolognesi, Methods Enzymol. **436**, 303–315 (2008).

[31] M. Milani, A. Pesce, Y. Ouellet, P. Ascenzi, M. Guertin, and M. Bolognesi, EMBO J. **20**, 3902–3909 (2001).

[32] D. B. Calhoun, J. M. Vanderkooi, G. V. Woodrow, and S. W. Englander, Biochemistry **22**, 1526–1532 (1983).

[33] D. B. Calhoun, J. M. Vanderkooi, and S. W. Englander, Biochemistry **22**, 1533–1539 (1983).

[34] W. M. Vaugham and G. Weber, Biochemistry **9**, 464–473 (1970).

[35] L. Chen, A. Y. Lyubimov, L. Brammer, A. Vrielink, and N. S. Sampson, Biochemistry **47**, 5368–5377 (2008).

[36] L. Piubelli, M. Pedotti, G. Molla, S. Feindler-Boeckh, S. Ghisla, M. S. Pilone, and L. Pollegioni, J. Biol. Chem. **283**, 24738–24747 (2008).

[37] T. Hiromoto, S. Fujiwara, K. Hosokawa, and H. Yamaguchi, J. Mol. Biol. **364**, 878–896 (2006).

[38] R. Baron, C. Riley, P. Chenprakhon, K. Thotsaporn, R. T. Winter, A. Alfieri, F. Forneris, W. J. H. van Berkel, P. Chaiyen, M. W. Fraaije, A. Mattevi, and J. A. McCammon, Proc. Natl. Acad. Sci. U. S. A. **106**, 10603–10608 (2009).

[39] B. J. Johnson, J. Cohen, R. W. Welford, A. R. Pearson, K. Schulten, J. P. Klinman, and C. M. Wilmot, J. Biol. Chem. **282**, 17767–17776 (2007).

[40] K. Furse, D. Pratt, C. Schneider, A. Brash, N. Porter, and T. Lybrand, Biochemistry **45**, 3206–3218 (2006).

[41] R. Daigle, M. Guertin, and P. Laguee, Proteins: Struct., Funct., Bioinf. **75**, 735–747 (2009).

[42] N. Plattner and M. Meuwly, Biophys. J. **102**, 333–341 (2012).

[43] X. Huang and S. G. Boxer, Nat. Struct. Biol. **1**, 226–229 (1994).

[44] J. Cohen, A. Arkhipov, R. Braun, and K. Schulten, Biophys. J. **91**, 1844–1857 (2006).

[45] R. Elber and M. Karplus, J. Am. Chem. Soc. **112**, 9161–9175 (1990).

[46] D. A. Case and M. Karplus, J. Mol. Biol. **132**, 343–368 (1979).

[47] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou, Nature **280**, 558–563 (1979).

[48] V. Šrajer, T. Y. Teng, T. Ursby, C. Pradervand, Z. Ren, S. Adachi, W. Schildkamp, D. Bourgeois, M. Wulff, and K. Moffat, Science **274**, 1726–1729 (1996).

[49] V. Šrajer, Z. Ren, T. -Y. Teng, M. Schmidt, T. Ursby, D. Bourgeois, C. Pradervand, W. Schildkamp, M. Wulff, and K. Moffat, Biochemistry **40**, 13802–13815 (2001).

[50] F. Schotte, M. Lim, T. A. Jackson, A. V. Smirnov, J. Soman, J. S. Olson, G. N. Phillips, Jr., M. Wulff, and P. A. Anfinrud, Science **300**, 1944–1947 (2003).

[51] M. Schmidt, K. Nienhaus, R. Pahl, A. Krasselt, S. Anderson, F. Parak, G. U. Nienhaus, and V. Šrajer, Proc. Natl. Acad. Sci. U. S. A. **102**, 11704–11709 (2005).

[52] A. Sato, Y. Gao, T. Kitagawa, and Y. Mizutani, Proc. Natl. Acad. Sci. U. S. A. **104**, 9627–9632 (2007).

[53] C. Bossa, M. Anselmi, D. Roccatano, A. Amadei, B. Vallone, M. Brunori, and A. Di Nola, Biophys. J. **86**, 3855–3862 (2004).

[54] A. Ostermann, R. Waschipky, F. Parak, and G. Nienhaus, Nature **404**, 205–208 (2000).

[55] F. Schotte, P. Anfinrud, G. Hummer, and M. Wulff, Biophys. J. **86**, 525A (2004).

[56] P. Banushkina and M. Meuwly, J. Phys. Chem. B **109**, 16911–16917 (2005).

[57] P. Banushkina and M. Meuwly, J. Chem. Phys. **127**, 135101 (2007).

[58] M. Ceccarelli, R. Anedda, M. Casu, and P. Ruggerone, Proteins: Struct., Funct., Bioinf. **71**, 1231–1236 (2008).

[59] Y. Nishihara, S. Hayashi, and S. Kato, Chem. Phys. Lett. **464**, 220–225 (2008).

[60] L. Maragliano, G. Cottone, G. Ciccotti, and E. Vanden-Eijnden, J. Am. Chem. Soc. **132**, 1010–1017 (2010).

[61] H. Steinhaus, Bull. Acad. Polon. Sci **1**, 801–804 (1956).

[62] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, 1967), p. 14.
[63] J. Hartigan, *Clustering Algorithms* (John Wiley & Sons, Inc., 1975).
[64] J. Hartigan and M. Wong, Appl. Stat. **28**, 100–108 (1979).
[65] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, Nucleic Acids Res. **30**, 1575–1584 (2002).
[66] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U. S. A. **103**, 9885 (2006).
[67] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299–306 (2004).
[68] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch, J. Chem. Phys. **122**, 184901 (2005).
[69] G. Berezovska, D. Prada-Gracia, S. Mostarda, and F. Rao, J. Chem. Phys. **137**, 194101 (2012).
[70] D. Gfeller, D. de Lachapelle, P. D. L. Rios, G. Caldarelli, and F. Rao, Phys. Rev. E **76**, 026113 (2007).
[71] D. Gfeller, P. D. L. Rios, A. Caflisch, and F. Rao, Proc. Natl. Acad. Sci. U. S. A. **104**, 1817–1822 (2007).
[72] D. Prada-Gracia, J. Gómez-Gardeñes, P. Echenique, and F. Falo, PLoS Comput. Biol. **5**, e1000415 (2009).
[73] F. Rao and M. Karplus, Proc. Natl. Acad. Sci. U. S. A. **107**, 9152–9157 (2010).
[74] F. Chazal, S. Oudot, P. Skraba, and L. J. Guibas, "Persistence-based clustering in Riemannian manifolds," in *Computational Geometry (SCG 11), 27th Annual ACM Symposium on Computational Geometry, Paris, France, 13–15 June* (ACM, 2011), pp. 97–106.
[75] W. Zheng, M. A. Rohrdanz, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 144109 (2011).
[76] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[77] P. G. Bolhuis, Biophys. J. **88**, 50–61 (2005).
[78] P.-A. Cazade, G. Berezovska, and M. Meuwly, "Coupled protein–ligand dynamics in truncated hemoglobin N from atomistic simulations and transition networks," Biochim. Biophys. Acta (2014) (in press).
[79] F. Rao, S. Garrett-Roe, and P. Hamm, J. Phys. Chem. B **114**, 15598–15604 (2010).
[80] D. Prada-Gracia, R. Shevchuk, P. Hamm, and F. Rao, J. Chem. Phys. **137**, 144504 (2012).