Intrinsic map dynamics exploration for uncharted effective free-energy landscapes

Eliodoro Chiavazzo[a], Roberto Covino[b], Ronald R. Coifman[c], C. William Gear[d], Anastasia S. Georgiou[d], Gerhard Hummer[b,e], and Ioannis G. Kevrekidis[d,f,g,1]

[a]Energy Department, Politecnico di Torino, Turin 10129, Italy; [b]Department of Theoretical Biophysics, Max Planck Institute of Biophysics, 60438 Frankfurt am Main, Germany; [c]Department of Mathematics, Program in Applied Mathematics, Yale University, New Haven, CT 06510; [d]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544; [e]Institute of Biophysics, Goethe University, 60438 Frankfurt am Main, Germany; [f]The Program in Applied & Computational Mathematics, Princeton University, Princeton, NJ 08544; and [g]Institute for Advanced Study Technical University of Munich, 85748 Garching, Germany

Edited by Michael L. Klein, Temple University, Philadelphia, PA, and approved May 18, 2017 (received for review December 30, 2016)

We describe and implement a computer-assisted approach for accelerating the exploration of uncharted effective free-energy surfaces (FESs). More generally, the aim is the extraction of coarse-grained, macroscopic information from stochastic or atomistic simulations, such as molecular dynamics (MD). The approach functionally links the MD simulator with nonlinear manifold learning techniques. The added value comes from biasing the simulator toward unexplored phase-space regions by exploiting the smoothness of the gradually revealed intrinsic low-dimensional geometry of the FES.

free-energy surface | model reduction | machine learning | protein folding | enhanced sampling methods

A crucial bottleneck in extracting systems-level information from direct statistical mechanical simulations is that the simulations sample phase space "at their own pace" dictated by the shape and barriers of the effective free-energy surface (FES). In particular, this bottleneck is often a problem in molecular dynamics (MD). Long simulation times are "wasted" revisiting already explored locations in conformation space. Over the last 20 years, there has been a tremendous amount of effort invested, and many truly creative solutions have been proposed for biasing the simulations so as to circumvent this. Several techniques have now become a standard part of the simulator's toolkit, like umbrella sampling or SHAKE. Other biasing techniques, like importance sampling, milestoning, path sampling or metadynamics, and the nudged elastic band/string method have been also ingeniously formulated to help alleviate the above problem. It is worth mentioning also more recent methods based on machine learning, like reconnaissance metadynamics or diffusion map-directed MD. An incomplete list of works reporting about those methods can be found in refs. 1–10. Moreover, a recent review on dimensional reduction and enhanced sampling in atomistic simulations can be found in ref. 11. A crucial assumption that underpins many of these methods is that the dynamics are, effectively, low-dimensional: there exists a "good set of a few collective variables or coordinates" (also called reduction coordinates), in which one can write an effective Langevin or Fokker–Planck equation. It is the potential of this effective Langevin representation that we are trying to identify and exploit. One generally expects this effective Langevin representation to be a higher order, generalized one with memory terms (12). In effect, we will show here how we can construct "short memory" approximations with the help of collective variables (CVs) detected and updated "on the fly" using manifold learning.

If we knew the right CVs and had an "easy way" to create molecular conformations consistent with given values of these variables, then creating tabulated or interpolated effective FESs with a black box atomistic simulator and umbrella sampling would be "easy." By observing the dynamics of the MD in these few CVs, we can then straightforwardly estimate the local gradient of the effective potential and the local diffusivity in the effective Langevin description. Actually, "easily" does not do justice to the problem. In fact, estimating effective Langevin terms locally from simulations is a highly nontrivial estimation problem in the theory of stochastic differential equations (SDEs), and many careers in financial mathematics are made from studying it carefully. Here, we will conveniently assume that we have at our disposal "the current best" local stochastic estimation techniques available, so that we can go from observations of the unbiased dynamics (or the umbrella-sampled dynamics) to local effective SDE term coefficients.

Given an approximate effective FES in its few collective coordinates, we can then go ahead to perform tasks, like reaction rate estimation, with the explicit surrogate function or its tabulated form. Mathematical and computational tools for performing such tasks on explicit or tabulated functions of a few variables exist in the standard mathematical literature and will also be assumed known and "off the shelf" available from the optimization literature.

Although finding these invaluable good collective coordinates is difficult, it is at least reassuring to know that "the useful collective coordinate set" is not unique but rather, conveniently degenerate. For the sake of argument, let an FES be a 2D curved manifold. Any set of two basis vectors on any plane that is one to one with our FES would suffice to parametrize it and can be suitably used to navigate the manifold. Discovering good coordinates for

Significance

Direct simulations explore the dynamics of physical systems at their natural pace. Molecular dynamics (MD) simulations (e.g., of macromolecular folding) extensively revisit typical configurations until rare and interesting transition events occur. Biasing the simulator away from regions already explored can, therefore, drastically accelerate the discovery of features. We propose an enhanced sampling simulation framework, where MD and machine learning adaptively bootstrap each other. Machine learning guides the search for important configurations by processing information from previous explorations. This search proceeds iteratively in an algorithmically orchestrated fashion without advance knowledge of suitable collective variables. Applied to a molecular sensor of lipid saturation in membranes, a helix dissociation pathway not seen in millisecond simulations is discovered at the second iteration.

Author contributions: E.C. and I.G.K. designed research; E.C. performed research on benchmark 1; R.C. performed research on benchmark 2; R.R.C., C.W.G., and A.S.G. contributed to the data mining aspects of the problem; R.R.C., C.W.G., and A.S.G. assisted in writing the paper; G.H. contributed to the computations of benchmark 2; and E.C., R.C., G.H., and I.G.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: yannis@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1621481114/-/DCSupplemental.

describing a function based on data is at the crux of modern computer and data science research. This discovery is precisely our task here too: find and explore an effective FES and parametrize it by constructing a map in terms of useful collective coordinates.

We first discuss the simple, 1D case. As shown in *Inset* of Fig. 1, *Left*, when a physically meaningful reaction coordinate s is known a priori, then a procedure for extracting a good approximate FES from computational data is obvious: a possibly regular grid in the 1D phase space is constructed, and umbrella sampling is performed to estimate the potential of mean force. Alternatively, several parallel, appropriately initialized short runs can be used to estimate the local effective Langevin drift and diffusivity. Either way, an approximate FES with controlled approximation error can be interpolated.

However, good collective coordinates are not globally known in advance and must be generated as the computation progresses. Consider as an illustration the blue thick line in Fig. 1 representing a 1D manifold with the corresponding 1D effective FES in yellow. Let us assume that, after some initial simulation time, the solution trajectory becomes trapped in one of the energy wells as reported in Fig. 1, *Left*. Then, data mining can be applied to an ensemble of locally sampled configurations to (*i*) establish that the relevant manifold is 1D; (*ii*) learn its local parametrization; and thus, (*iii*) detect the boundary points of the manifold portion so far explored ("fathomed"). These boundary points can now be smoothly extended outward. Extension is not intended in time, but in the geometry of the manifold, parametrized locally by the first diffusion coordinate (13) or the first local principal component (LPC) close to each boundary. This extension takes us beyond the conformation space already explored and may well be against the local FES gradient, thus possibly leading to significant computational savings. The extension cannot but be an approximate one. In fact, as schematically indicated by black arrows in Fig. 1, *Left*, it acts as a "predictor," performing a Taylor series approximation of the manifold locally in the ambient space. A "corrector" step must follow: a short equilibration gives us unexplored conformations on the manifold beyond what we had already fathomed. The latter step is schematically illustrated by the red arrow following the black arrow in Fig. 1, and it can be possibly performed, e.g., by umbrella sampling using Plumed (14) or Colvars (15). New brief simulations are run, and new points on the manifold are collected (second blue point cloud in Fig. 1, *Left*), added to the database, and then fed to data mining to parametrize the augmented FES geometry. The extension procedure repeats again and again: new unexplored conformations keep being added, and the extended geometry of the effective FES is gradually revealed, leading to the discovery of new wells as schematically reported on the right side of Fig. 1, *Left*.

In higher dimensions, the basic approach remains the same, although its representation becomes more complicated. Let us consider a 2D "undulating" FES embedded in a high-dimensional ambient space as shown in Fig. 1, *Right*. The "color map" on this carpet denotes the effective FES contours; if, as we "walk" on the carpet, we estimate the local color gradient, we can use this information to help us direct our walking pattern. This dynamically adaptive exploration strategy should speed up the extraction of useful information, like a reaction rate, or the discovery of a saddle point.

In *Indiana Jones and the Last Crusade*, the hero walks on a glass mirror bridge that he cannot see. However, in the end, he takes some sand and throws it at his feet, so that the sand reveals the local shape of the bridge. This approach is precisely what we do in our intrinsic map dynamics (iMapD) with our "free-energy carpet." We start with simulations that have locally and partially sampled some location on the FES, thus representing the "sand we have poured around our own feet" or the first small cloud $C^{(1)}$ in Fig. 1, *Right*. However, now that a little of the low-dimensional geometry of the carpet is revealed, we can walk by taking a big step to a new location, pour some sand there, namely initialize MD conformations consistent with the new location in collective coordinate space, and start one or more unbiased simulations there. This new set of data is "the new sand" represented by the new little cloud $C^{(2)}$ in Fig. 1, *Right* to which we have stepped. The size of the step toward new locations depends on the smoothness of the carpet. One can easily intuit how the geometry is revealed from iterating this process: For a d-dimensional (here $d=2$) reduced description, the initial "seed" simulation will form an effectively d-dimensional cloud. We need to identify the $(d-1)$-dimensional boundary of this cloud (its "silver lining"), which in this case, is 1D. Starting from an ensemble of points on this 1D curve, we "take a step away" from the cloud, smoothly extrapolating the coordinates "as far as we trust their smoothness." Clearly, point extrapolation can be performed one by one or all in parallel computationally. As explained in ref. 16, we could possibly move along local geodesics. We can "trust" these geodesics only so far, because the carpet may "violently" curve, and our smooth extrapolation may not locally parametrize it any more.

By identifying the boundary, marching "outward" from a number M of points on it, creating M little d-dimensional clouds from simulations initialized at each extrapolation, and then, "integrating" the new M clouds in an atlas with the initial one again and again, we will "fathom" the carpet. Therefore, although the "local marching" from every boundary parametrization point may be done in local coordinates [e.g., local principal component analysis (PCA)] or more global coordinates [e.g., diffusion maps (DMAPs) geometric harmonics (17)], all new data points at every iteration can be integrated in our global geometry by either reprocessing them all together or creating an efficient database structure that allows transitioning from local coordinates of one cloud to local coordinates of the cloud next to it (one chart in an atlas to the next). This strategy is reminiscent of "simplicial continuation" (18) in following the parametric solutions of algebraic equations. We have a predictor step represented by our extrapolation in local reduction coordinates and then, a

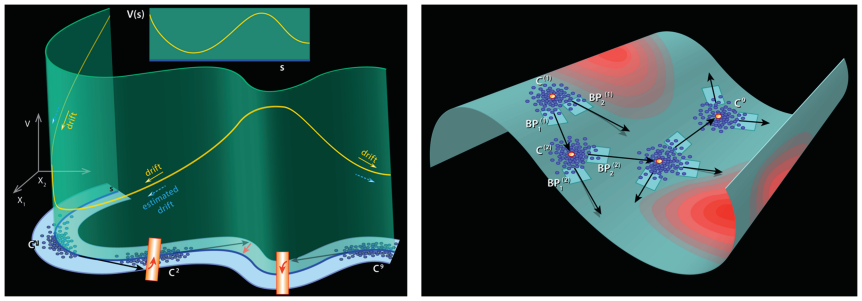

Fig. 1. Pictorial illustration of the iMapD exploration procedure with (*Left*) 1D and (*Right*) 2D effective FESs. In *Left Inset*, a good collective coordinate is already available—the collective coordinates in *Left* and *Right* are not a priori known. A full description is in the text.

corrector step that "brings us back down" to the FES. This corrector step might be discarding the fast initial transient of an unbiased simulation or running umbrella sampling constrained on the extrapolated coarse coordinates.

It is clear that we only need to move outward if we are to explore new areas; this goal can be achieved through good bookkeeping. Although extension is a nontrivial process to generalize, bookkeeping is easy in one dimension, easy (but nontrivial) in two dimensions, doable in three, and difficult in four. Practically, we expect the process to be easy to program for relatively low-dimensional (up to 4D) FESs, whereas it will require nontrivial computational geometry and programing in higher dimensions. This bookkeeping is precisely the same one necessary in multiparameter simplicial continuation for the tracking of solutions of algebraic equations (18, 19) [the available software package for continuation and bifurcation problems AUTO (20)].

Results

We discuss now the implementation of the proposed iMapD molecular simulation sampling approach.

Conforming to the literature and for the sake of clarity, our first benchmark illustration is the time-honored alanine dipeptide (10, 16, 21), here in implicit solvent with the Amber03 force field (22).

In our second application, we apply the iMapD algorithm to the transmembrane protein Mga2, which plays a key role in the regulation of lipid saturation levels in the yeast endoplasmic reticulum (ER). Recent simulations and experiments identified a unique rotation-based sensing mechanism to probe the membrane characteristics (23). In response to changes in lipid saturation, the 30-amino acid transmembrane helices (TMHs) anchoring Mga2 into the ER were found to rotate relative to each other in an Mga2 dimer, driven in part by packing effects acting on bulky protruding tryptophans. Just probing the rotational dynamics and charting the underlying free-energy landscape required millisecond-long MD simulations feasible only with a coarse-grained (CG) description (24, 25). However, even on this long timescale, only the TMH contact could be sampled, with TMH dissociation expected to occur on timescales orders of magnitude longer. Therefore, even in more than 3-ms simulations of a simplified CG description, the relevant configuration space of the dimer could not be sampled exhaustively.

Here, we show that, with iMapD, not only the competing Mga2-bound states but also the unbinding pathways can be discovered simply by strategic initialization of otherwise fully unbiased MD trajectories.

Benchmark 1. It has been long argued that alanine dipeptide admits a 2D reduced description in terms of two physically meaningful coordinates, namely the dihedral angles ϕ and ψ (21). While exploring the FES, our approach does not require such a priori knowledge of either the dimensionality or some physical meaning of the collective coordinates. Three successive stages of our exploration protocol are reported in Fig. 2. The protocol is initialized from a transient simulation segment, providing an ensemble of configurations visibly trapped within some initial potential well: this set of configurations is what we call the initial simulation data. Each stage of iMapD is composed of the following substeps.

Data mining. A manifold learning technique is used to discover a low-dimensional embedding for the data collected so far. Here, we use DMAPs (13). This discovery includes the selection of the appropriate dimension (d) of the manifold and its parametrization, here in terms of d leading diffusion coordinates (DC1, …, DCd).

Boundary detection. Using algorithms from the literature, such as alpha shapes (26, 27), or more generally, "wrapping" algorithms (28), we detect the $d-1$-dimensional boundary of the region explored by the available simulation data. In high dimensions, suitable algorithms are described in refs. 29 and 30.

Outward extension. At each boundary point, we take an outward step. In the current implementation, the latter step is approximately normal to the boundary in the tangent space of the low-dimensional manifold, and it is performed using LPCs in the ambient space. For each boundary point, (*i*) a fixed number of nearest neighbors is detected in ambient space; (*ii*) local PCA is performed on this set of neighbors with the local reduced dimension d_{loc} selected by a threshold for the maximum variance (details below and in *SI Text*); (*iii*) the center of mass of this local neighborhood is computed in the d_{loc}-dimensional PCA space; and then, (*iv*) "outward extension" of the manifold at the original boundary point is performed in the PCA low-dimensional space along the line segment passing through the local neighborhood center of mass and the boundary point itself. Geometric harmonics or Laplacian pyramids (17, 31) can alternatively be used for this purpose.

Lifting. Lifting is then performed from the extended LPCs to unexplored molecular configurations lying on/close to the extended manifold. Going from LPC directly to ambient space was satisfactory in this simple illustration. In general, however, equilibrated conformations consistent with the extended LPC values may be needed. As an example, those consistent and equilibrated conformations can be obtained through short, constrained, umbrella sampling runs (21), making use of Plumed (14) or Colvars (15).

New sampling/database updating on the extended manifold. Short simulation bursts are carried out from these new "extended" initial conditions. Possibly, several replicas can be generated from each condition when initialized with different Maxwell–Boltzmann velocities and/or different thermostat seeds. The new data are appended to the growing fathomed configuration database.

The procedure then repeats until new metastable configurations are detected. In this respect, we notice that the adopted machine learning algorithm comes with the intrinsic ability of detecting new states. When such an event occurs, the new sampled configuration set will appear in the low-dimensional diffusion coordinate representation as a new distinct cluster. Convergence can thus be inferred, and the search can be terminated when key features of the diffusion coordinate representation, like the number of clusters and connectivity, do not change anymore.

In Fig. 2, *Left*, an initial transient is visibly trapped in two nearby metastable configuration wells. DMAPs clearly suggest a 2D FES: Fig. 2, *Left Inset* shows this 2D manifold embedded in the space of the first three DMAP coordinates. The boundary points of the fathomed portion of the manifold are identified (Fig. 2, red circles) and extended outward (Fig. 2, green stars). Lifting via LPC is quite satisfactory here, and new sampling on the extended manifold is performed through simple unbiased short runs initialized at the lifted configurations. The resulting new configurations are appended to the growing simulation database, and a new round of data mining, boundary detection, and outward extension is shown in Fig. 2, *Center*, both in 2D projection and 3D embedding. This procedure is repeated one more time, leading to Fig. 2, *Right*, where two new folded metastable configurations have been discovered. What is important is that the manifold parametrization shown in Fig. 2, *Right* was not known at the beginning. Only the small portion of the manifold marked by the yellow ellipse in Fig. 2, *Right* was initially available. The geometry of the growing manifold beyond that initial ellipse and its adaptively growing parametrization have been gradually revealed as part of our exploration protocol.

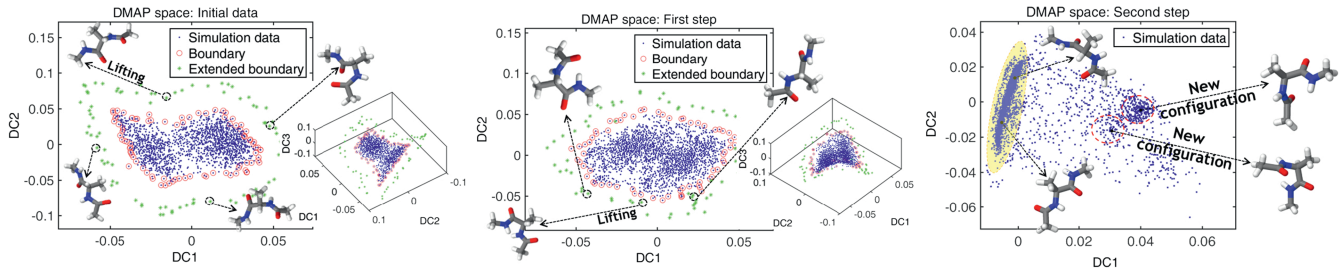

Fig. 2. (*Left*) A long initial trajectory of alanine dipeptide trapped in two nearby metastable wells is shown in the corresponding 2D DMAP projection (a 3D DMAP space embedding is also reported in *Inset*). Boundary points are identified (red circles) and extended outward (green stars). Local PCA suffices to lift to ambient space (see below). (*Center*) Short simulation runs are performed from previously extended boundary points. New configurations are generated and displayed (blue dots) in a 2D (and 3D) DMAP reduced space. (*Right*) After two steps, two new potential wells are reached by some of the simulation frames. The "starting" portion of the FES geometry accessed by the initial simulations is marked in yellow—the rest has been revealed through exploration.

For the sake of completeness, the results of the above exploration process are also reported in the popular Ramachandran plot in Fig. 3 in terms of the two physical coarse variables ϕ and ψ. The dipeptide, initially trapped in the basins in Fig. 3, *Upper Left*, is gradually forced toward new configurations that would not have naturally been visited in such a short simulation time period. In this relatively simple example, the low-dimensional FES "slow manifold" identified on the fly happens to also be the graph of a function above the two Ramachandran plot coordinates. In other words, the determinant of the Jacobian of the transformation from the "physical" coarse coordinates ϕ–ψ to the diffusion ones DC1–DC2 keeps the same sign and is neither too big nor too small on the data: it stays bounded away from zero and infinity, so that the transformation from physically meaningful to data-based collective coordinates is bi-Lipschitz (32). This consideration implies that the effectively 2D FES can be described equally well in terms of ϕ–ψ or our (evolving) DC1–DC2. If, however, this effectively 2D manifold "folded" over the Ramachandran plot variables, our data mining would still be able to correctly parametrize and extend the FES.

Before we elaborate on the steps, a few words about efficiency. In this example, the total computational time associated with all performed simulation bursts was estimated at ≈ 50 ps. It is known that, for this system, ≈ 150 ns of direct simulation are, on average, needed to observe the transition from the initial, lower free-energy configurations to the discovered, higher free-energy ones (the same system was simulated in ref. 10). This analysis yields an apparent computational speedup of three orders of magnitude for this rudimentary implementation, in line with what was observed in ref. 10, where the reinitialization did not involve extrapolation but rather, occurred at the "farthest reached" point in the leading diffusion coordinate. Because the computation there was 1D, boundary detection was straightforward. This technique, like reconnaissance metadynamics, also builds the exploration geometry and actually does it "seamlessly" without having to "jump and reinitialize" consistent molecular configurations. However, it is precisely this "jumping and reinitializing" that we feel is the most powerful element of our approach (see also the 1D illustrative example in Fig. S2): we do not have to wait to "fill in the wells" (as in metadynamics), and we do not need to sample the part of the geometry that we trust is smooth enough. We can take a step "as big as we trust" in the geometry and then, sample there, and in this way, we save a remarkable amount of computational time when exploration is the goal (*SI Text*). These two steps are determined by the length Δt of the unbiased sampling and the length c of the extension (*Materials and Methods* and *SI Text* have details). Some conceptual geometrical considerations are in order.

For a simple 1D SDE in terms of a known scalar variable x, reinitializations can be carried out with no effort, and the extension parameter c can be chosen as large as one likes. The main reason is that the support of the effective FES is not a curved 1D manifold in a high-dimensional space. In complex atomistic simulations, however, challenges to the practical implementation of the above procedure arise, because

the low-dimensional support (manifold) of the effective FES is typically curved and embedded in a high-dimensional phase space,

CG coordinates parametrizing this manifold are a priori unknown and need to be systematically discovered and "harmonized" with their incarnations at the previous step, and

reinitialization of the fine-scale simulator requires a lifting operator from the low-dimensional space up to ambient physical space as discussed in ref. 31.

Those are precisely some of the aspects addressed in this work. Our alanine example only provides a proof of concept illustration, because the code that we set up was far from optimal. We did not optimize the extension parameter c or the unbiased sampling time Δt, which was, instead, kept constant, and in this first

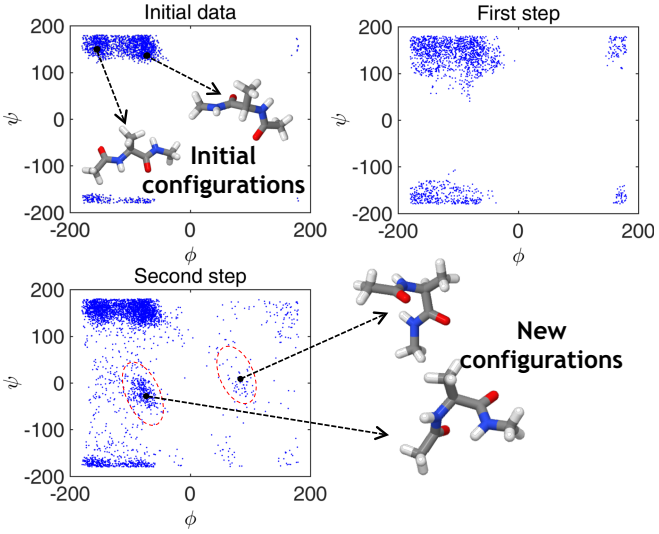

Fig. 3. The discovery process of Fig. 2 is redisplayed here as a Ramachandran plot of alanine dipeptide. Two steps are sufficient to reveal two initially unknown metastable configurations of the molecule. At each step, before outward extension of the boundary, we also performed global PCA filtering of the data noise, where 98% of the variance was retained (*Materials and Methods* and *SI Text*). Here (and below), physical coordinates are used only for representation purposes and are not used for computation.

attempt, at each step, we extended all detected boundary points. Although iMapD is, in principle, not limited in terms of the FES intrinsic dimension d, for simplicity, the current method implementation has been developed up to $d = 3$. This choice is also because of restrictions on the adopted alpha-shapes boundary detection (26, 27). Although the latter limitation can be overcome using alternative methods (29, 30), more care is needed when coping with the rapidly increasing number of boundary points with dimension d. To this end, future optimized implementations will include a smarter parametrization/selection of the boundary points to extend as well as a smarter selection of the unbiased sampling interval based on local estimates of the free-energy gradient. This selection is expected to follow the same principles discussed in detail and shown in ref. 16, where, however, the collective coordinates were already known. One might, for example, not extend points at which the effective FES rises steeper than a preset threshold. Finally, in estimating computational speedup, we should also include the cost of necessary intermediate steps, such as DMAPs, local PCA, and lifting.

Benchmark 2. Having shown the power of iMapD in applications to well-characterized model systems, we next use it to chart the configuration space of the biologically relevant Mga2 sensor of lipid saturation (23). For this challenging molecular system, even millisecond-long atomistic MD simulations proved insufficient to observe a dimer dissociation event. However, they provide us with an excellent reference for the dimeric-bound state (23). Fig. 4 shows the corresponding free-energy landscape as a function of the first two global DCs of the dimers, with the four highly populated clusters corresponding to local minima. Importantly, we do not use this surface to guide iMapD in any way, only to give the reader a global view of the progress in the search.

As the first step in iMapD, we run a burst of 10 short (100-ns) unbiased simulations initiated from the same starting configuration. The resulting trajectories sample their vicinity but do not escape the local free-energy minimum (black squares in Fig. 4*A*). We use the structures along these simulations to detect the boundary in the local DC representation, and from there, we project outward, building 16 new starting configurations. From each of these, we start another burst of 10 100-ns long unbiased simulations. Although most of these trajectories fall back to the starting cluster, many are able to escape from it, landing into new regions of the landscape and effectively discovering most of the highly populated clusters during the first phase of the expansion (blue circles in Fig. 4*A*). In the spirit of building a growing map of the landscape, we combine all configurations sampled so far and repeat the boundary detection and projection in newly calculated local DC to obtain new starting configurations. In a second iMapD round starting from them, we already visit all relevant regions of the landscape (blue circles in Fig. 4*B*).

For reference, we extend the initial unbiased simulations to estimate the timescales necessary to explore the landscape in a purely equilibrium approach. In Fig. 4*C*, we can see that, after running 10 simulations for 2 μs each, only one trajectory is able to leave the starting cluster. To discover all of the remaining clusters, each simulation must be run for 4 μs.

We now represent the exploration process by using two angles u and v that describe the relative orientation of the two TMHs in an Mga2 dimer (*Inset* in Fig. 5, *Lower*). We again took advantage of already available long-equilibrium simulations to calculate a reference FES as a function of u and v. Because of the identity of the two TMHs, the surface, shown in Fig. 5, is approximately mirror symmetric with respect to the bisector.

In iMapD, the first short unbiased simulations sample structures where the two reference tryptophans W10 face each other and consistent with what we saw in the global DC representation, are confined to the starting state. The first expansion leads to the discovery of two new states, which contain configurations where the W10s are far apart, pointing in one case to opposite directions and in the other, in the same direction. Taking into account the symmetry of the surface, the last relevant state is discovered during the second expansion. Importantly, in a few configurations at this stage, the two TMHs are actually separated, which represents a disassociation event of the dimer. This particular new configuration is then sampled for almost one-half of the time during a third expansion.

Discussion

In this work, we described, implemented, and tested iMapD, a geometry-based, machine learning-inspired approach to accelerate the extraction of information from atomistic and stochastic simulators. In particular, we focused on the exploration of effective FES. The unique enabling feature of iMapD is the performance of computations without prior assumptions on its reduced description in terms of CVs. In fact, according to a large body of literature, suitable CVs can be rather nonintuitive, even in simple systems (33, 34).

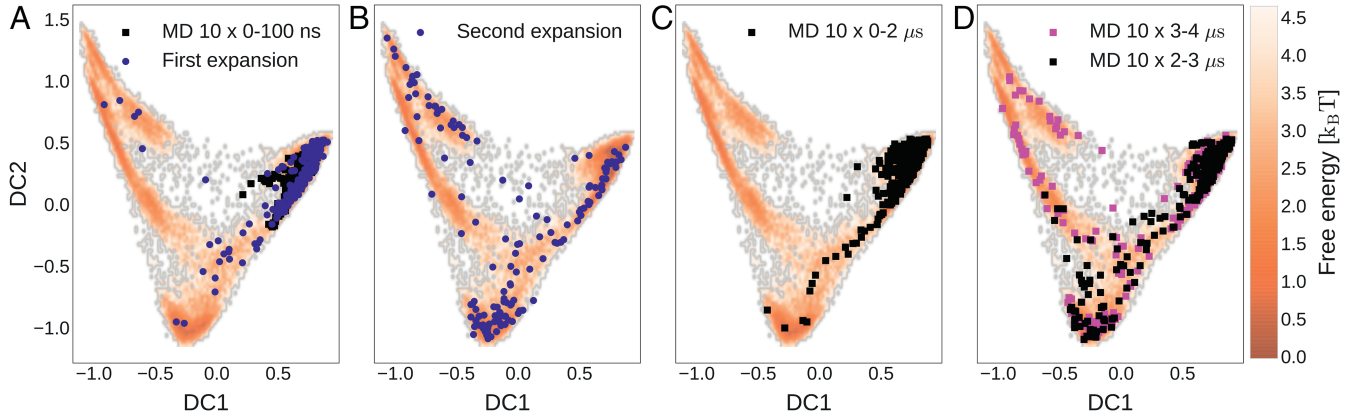

Fig. 4. Enhanced exploration of Mga2 dimer configurations represented on the FES as a function of the first two global DMAP coordinates. (*A*) Configurations sampled from 10 100-ns-long unbiased simulations initiated from a single configuration (black squares). Final configurations of 100-ns-long unbiased simulations initialized from the first set of 16 newly projected structures (blue circles). (*B*) Final configurations of 100-ns-long unbiased simulations started from the second set of 16 newly projected structures. Configurations from the initial 10 unbiased simulations that were extended and are here tracked (*C*) up to 2 μs and (*D*) from 2 to 3 μs (black squares) and from 3 to 4 μs (magenta squares). The FES was previously extracted from a 2.52-ms-long equilibrium simulation.

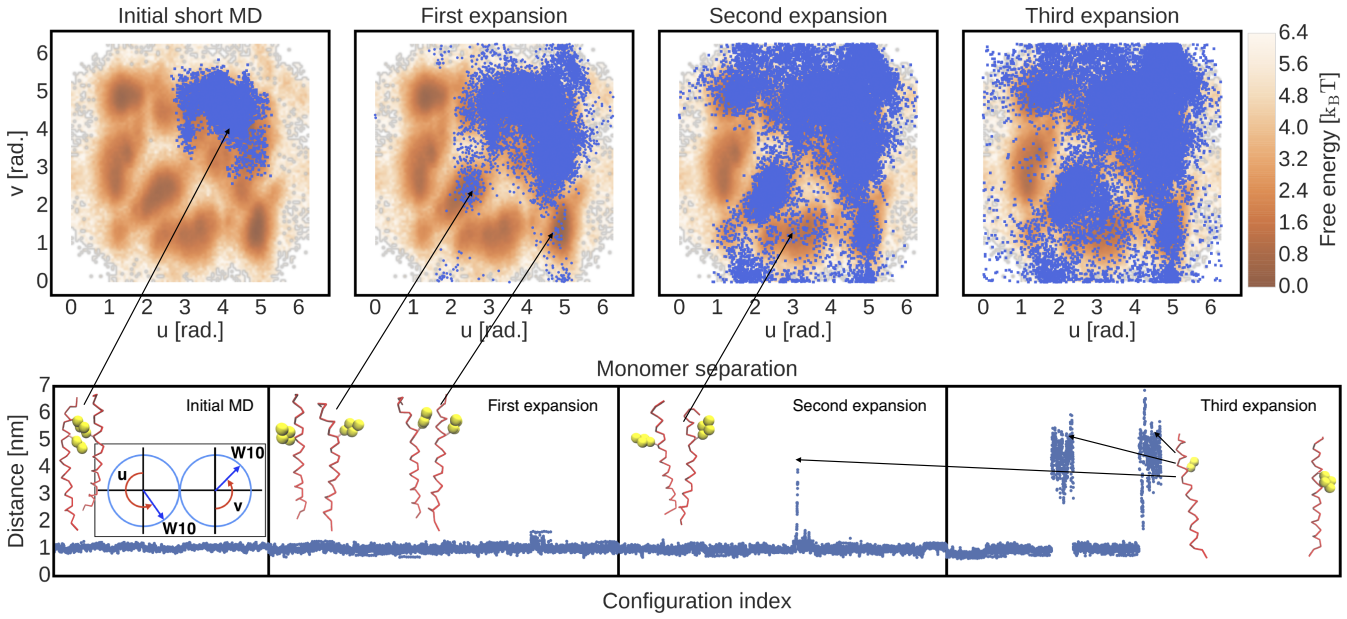

Fig. 5. (*Upper*) Enhanced exploration represented in u, v space. Angles were calculated on configurations sampled from cumulative trajectories simulated during the successive exploration phases and are represented as blue squares. The FES as a function of u, v was extracted from a 2.52-ms-long equilibrium simulation. (*Lower*) Distance separating the two TMHs monitored during all of the exploration phases. *Inset* shows different representative structures of the dimer, with the reference residue W10 shown in yellow and a schematic definition of u and v.

Importantly, our algorithm has been tested on CG simulations of Mga2 TMH dimers, a system of biological relevance with rich conformational dynamics in the microsecond to millisecond regime and beyond. Here, the two helices can make use of various contact interfaces, corresponding to the clusters in the FESs shown in Figs. 4 and 5.

Our setup mimics a situation in which only one structure is known and MD simulations are restricted to short timescales because of the size of the system. On the one hand, the number of computing cores that can be used to parallelize a single simulation might be limited by lack of resources or bounds in the scaling behavior; on the other hand, the dynamics of complex (bio)molecular systems are characterized by long correlation times. It stands to reason that, in such situations of practical interest, running many short independent simulations is often more effective than focusing on few long ones. However, a crucial element of this strategy is to select appropriate initial configurations to not get trapped in configuration space.

By using machine learning algorithms (DMAPs and PCA) we were able to infer new configurations from which to start bursts of short unbiased simulations, and we were able to efficiently discover new relevant structures of the Mga2 TMH dimers. Starting from a single initial structure, the iMapD algorithm was able, in only two iterations, to sample structures in the entire relevant configuration space of the dimer. We want to stress that all simulations that we have used are unbiased: after "intelligent" reinitialization, no unphysical force was added to steer the dynamics of the system. Efficient data reinitialization (detailed below and in *SI Text*) is a main enabling feature of this work.

To monitor the progress of the exploration, we used low-dimensional FESs calculated as a function of both machine learning coordinates (DMAP) and physical variables (angles u, v). These surfaces are representations of the configuration space of the dimer [i.e., when both TMHs are in close proximity ($\approx$ 1 nm)]. As we saw in Fig. 5 by monitoring the distance separating the two TMHs during the second and third expansions (and thus, after only a few 10s of microseconds of cumulative simulated time), the algorithm sampled a dissociation event.

In equilibrium simulations, these events occur on much longer timescales than the formation of the dimer itself. During more than 3-ms-long equilibrium simulations of the Mga2 dimer, we never observed a single dissociation event (23).

This problem has been recently addressed in the context of the same CG model by using metadynamics (35), where an unphysical history-dependent bias must be added on the distance separating the two TMHs, considered to be a priori a slow coordinate of the system. We applied a similar protocol to the Mga2 TMH dimer system as reported in *SI Text*. There, we show that the exploration efficiency of iMapD compares favorably with metadynamics (Fig. S1), and we illustrate the difficulty of capturing a single CV describing at the same time both the formation of a TMH dimer and the rotational dynamics in the dimer state. Moreover, we also quantify the artifacts caused by an unphysical biasing potential acting along a nonideal CV (Fig. S1).

In iMapD, instead, the algorithm "discovers" the slow coordinate corresponding to the separation between the two TMHs after having exhaustively explored the slow coordinates describing the conformational rearrangement in the dimer state shown in Figs. 4 and 5. One can say that the algorithm gradually and adaptively discovers a hierarchy (an "atlas") of slow coordinates. Whereas our exploration proceeds without explicit CVs, as a key difference to other methods, a postanalysis can certainly be used to identify the mechanisms and the associated CVs. For the Mga2 helix dimer, we first recovered the relative rotations of the two helices in the bound state, as seen previously in millisecond-long simulations. We then discovered a mechanism of dissociation, in which the amino acid contacts break as helices tilt against each other and lipids slide in between to trap the tilted state and trigger dissociation. This mechanism thus combines geometric and solvent coordinates and cannot easily be globally described by a combination of simple CVs (e.g., distances and angles).

Importantly, all of the short runs in iMapD are unbiased and appropriately, reweighted for the choice of initial conditions, can be exploited using standard methods as input in the construction of master equations or Markov state models. Moreover, because

we run many short simulations instead of few long ones, iMapD is inherently highly parallelizable.

The main attractive feature of the proposed approach is that it can explore low-dimensional effective FESs in high-dimensional configuration spaces without the need of relying on a priori knowledge of suitable collective coordinates. In fact, our coarse coordinates are progressively and adaptively revealed as computation progresses.

The main assumption on which our method is based is the same one that underpins most of the model reduction techniques in statistical mechanics: because of timescale separation, the system dynamics is mostly confined on low-dimensional (smooth) manifolds in phase space (31, 36–39). Our approach squarely aims at exploiting smoothness of the low-dimensional manifolds, which for the gradient systems of interest here act as the support of the FES governing molecular and other atomistic dynamics. Writing the expression jumping "as far as we trust the smoothness" above is then the pivot on which our approach "lives or dies." Two important issues, one relatively simple and one deeper, determine the tuning parameters of our algorithm. The first is the easier one: if the effective FES does not have hierarchical roughness, then there already exist two computational enabling technologies that support our algorithm. The first "enabler" has to do with the local scaling of the noise through a Mahalanobis-like distance (40), which combined with DMAPs-based data mining, conveniently factors out fast local oscillations. The curved fast local invariant measures, "half-moons" as we call them, are discussed in ref. 41.

The second enabler is straightforward. After factoring out these fast oscillations, we have a smooth surface, and now, we are faced with a numerical error control problem: the need for systematic adaptive step-size selection. We will not address this technical issue here; we simply note that the same computational machinery that, in traditional initial value problem solvers, allows one to make local error estimates can also be, in principle, used for our purposes. Performing the computation with one step and then performing it twice with one-half the step allows one to make a local "on line" error estimate and keep the computation below prescribed error bounds.

The second issue is deeper, and we will only pay lip service to resolving it, although we believe that what we suggest is "the right way" to go about it. This critical issue is caused by hierarchical roughness. In SDE language, this condition implies that our potentials are multiscale potentials and that our noises may not be just additive. Here, we revert to the discussion above about "what the best off-the-shelf estimation techniques" for multiscale diffusions and perhaps, not only diffusions but also, possibly Levy flight processes may be. In all of our discussion, we assumed that the effective equation is a Langevin or the associated Fokker–Planck. For simple "egg-carton"-like potentials, it is possible through ingenious but relatively straightforward tools (e.g., subsampling) to "go around" the roughness and estimate a smooth effective SDE (42–45).

However, what is more systematically missing is a round of data processing and if necessary, additional data collection for hypothesis testing. In a more general context, in 2007, we discussed this issue of "Deciding the nature of the coarse equation through microscopic simulations: The baby-bathwater scheme" (46). As the abstract of that paper states,

> "The effective coupling of microscopic simulators with macroscopic behavior requires certain decisions about the nature of the unavailable coarse equation … In the absence of an explicit formula … we propose, implement, and validate a simple scheme for deciding these and other similar questions about the coarse equation using only the microscopic simulator" (46).

We believe that the collection of data for hypothesis testing about the nature of unavailable effective SDEs is a nascent field, and we are cognizant of relatively few efforts in this direction. However, given the microscopic simulators, one can collect the data necessary for such algorithms. We believe that, although there will be technical difficulties and good mathematics in the process, this research area will advance significantly in the near future. Our approach will benefit from these advances.

While discussing estimation, there is another significant and less difficult item to consider: the exploitation of the estimated local potential gradient of the FES in informing its geometric exploration. To a large extent, this item has been discussed in ref. 16 when the collective coordinates were known. Here, the issues remain the same. When at the bottom of a well, we probably intend to move upward. We may want our steps to be along local geodesics; we may want our steps to maybe try to conform to level sets of the effective free energy. If we find a saddle, we may just kick a little "on the other side" and let the simulation find the new well bottom by itself. If we have disparate gradients in different directions (surfaces that look like the Grand Canyon), we might prefer not to go above some level set, because the simulation would never get there in a person's lifetime. If the code is to be useful, we appreciate that these decisions and several other common sense decisions have to be implemented in an automated fashion in the code.

Clearly, that option is a matter of effort and resources. Plumed (14) and Colvars (15) go quite a long way toward being a platform in which to incorporate what is done here and what we envision being done in an automated fashion. As shown in ref. 36, it is also important to recognize that one may have "exotic" manifolds that change dimensionality as the exploration proceeds. For instance, in these cases, we might have two dimensions narrowing to one dimension and then, maybe "opening back out," like a river delta, to two dimensions. We can, in principle, deal with that geometrically, but we do not discuss this in this paper referring to the work of Belkin et al. (47) and others (48). The important issue of adaptively determining the dimensionality of low-dimensional surfaces in data mining has been discussed in ref. 49.

It is fitting to close the discussion with a quotation from the 2016 review article by Peters (50):

> "However, the methods in this review share one overarching disadvantage. Human intuition remains the best source of trial coordinates and mechanistic hypotheses, and there is no procedure for having an epiphany. All current algorithms for optimizing reaction coordinates work within the space of chosen trial coordinates."

Our work here attempts such a "computer-assisted epiphany": by adaptively revealing the exploration geometry and exploiting its smoothness to guide additional exploration; it makes a step toward circumventing human intuition in the discovery phase. However, the rationalization of what has been discovered in terms of physically interpretable candidate coordinates is an important postprocessing step and significantly augments the overall value of the process (e.g., confirming that, in the last stages, the relative tilt between the two helices is "one-to-one" with the machine-discovered coordinates). A useful discussion of the "man-versus-machine" detected variables can be found in refs. 51 and 52.

Materials and Methods

DMAP: Mapping from Ambient Space to Reduced Space. The data mining substep of our procedure has been performed by the DMAP method, where density-invariant normalization of the affinity matrix has been performed. Full details on DMAP can be found in refs. 13 and 17, whereas the specific implementation followed in this work is provided in *SI Text*.

Lifting from Reduced to Ambient Space by Local PCA. At the core of our approach stands an effective procedure for extending and lifting from reduced to ambient space the boundary of the so-far explored FES portion. Let **B** be an arbitrary boundary point in the p-dimensional ambient space,

which we want to extend outward with respect to the previously simulated (available) point cloud. Let us identify the $(n-1)$ nearest neighbors of **B** in the ambient space. Let **X** be the $n \times p$ data matrix collecting the Cartesian coordinates of n points, namely the ones within the chosen neighborhood of **B**, including **B** itself.

Let d_{loc} be a PCA-based estimate of the FES local dimension as discussed in detail in *SI Text*. Let us consider the $n \times d_{loc}$ matrix **Y** collecting the reduced PCA coordinates of the n points of interest, namely the first d_{loc} columns of the matrix of principal component scores **S**. Let y_B and y_{center} be the PCA reduced coordinates of the above boundary point B and the center of mass of the considered neighborhood, respectively. Boundary points are extended in PCA space along the direction passing by y_B and y_{center}, so that a new point can be identified as $y_{new} = y_B + c\vec{\mathbf{v}}$, where c is a nonnegative scalar quantity stipulating how far we intend to extend the point **B** from the current location; however, $\vec{\mathbf{v}}$ is the (outward) unit vector along the chosen extension direction. Lifting the new point y_{new} into ambient space can be readily accomplished by a linear mapping:

$$\mathbf{Y}_{new} = y_{new}\tilde{\mathbf{C}} + \bar{\mathbf{X}}, \qquad [1]$$

where the $d_{loc} \times p$ matrix $\tilde{\mathbf{C}}$ is given by the (transposed) first d_{loc} columns of the matrix of loadings **C**, whereas **X** is the mean row vector, where each of the p ambient space coordinates is averaged over the n points in the chosen neighborhood. Full details on extension and lifting are provided in *SI Text*.

Computations with Ala Dipeptide. The Ala dipeptide is simulated with GROMACS 4.5.5 (53, 54) in a periodically replicated box with dimensions of $2 \times 2 \times 2$ nm^3. Solvent is treated implicitly using the Still generalized Born formalism with a cutoff of 0.8 nm. The temperature is maintained constant at 300 K by means of a velocity rescaling thermostat (55).

When searching for DMAP low-dimensional embedding, all configurations are first aligned to a reference configuration using the Kabsch algorithm (56, 57), and afterward, the standard Euclidean distance is used as pairwise dissimilarity function. The DMAP model parameter was set at $\epsilon = 0.35$ nm. When performing local PCA, at each boundary point, $n = 65$ nearest neighbors are considered, whereas the local dimension d_{loc} is automatically estimated by setting a threshold for maximum variance of 0.95.

The nonnegative scalar quantity c for local extension is chosen in the range $0.05 < c < 0.12$. Starting from each new extended configuration, two short bursts are performed, each time randomly reassigning Maxwell–Boltzmann velocities to atoms. Simulation bursts consist of 15,000 simulation steps with a time step of 0.02 fs. The latter unusually small time step is not essential for computations: it was chosen for convenience, because it ensured a sufficiently large number of samples along the burst trajectories.

Computations with Mga2.

Model and simulation details. The 30-aa-long transmembrane domain of Mga2 (sequence in single-letter code: RNDKMLIFFWIPLTLLLLTWFIMYKFGNQD) was modeled as an alpha helix in the MARTINI v2.2 force field (24, 25). We used the insane tool (58) to assemble for each simulation a $10 \times 10 \times 10$-nm box containing two Mga2 monomers, about 300 1-Palmitoyl-2-oleoylphosphatidylcholine (POPC) lipids, water, and ion beads corresponding to a 0.15 M NaCl concentration for a total amount of about 10,000 beads.

Each initial configuration was relaxed by using 15,000 steps of steepest descent and then equilibrated for 2 ns at a temperature of 303 K and pressure of 1 atm, restraining the positions of the protein beads compatibly with the pressure coupling. Temperature was kept constant with the velocity rescaling thermostat (55), and pressure was kept constant with the semiisotropic Berendsen barostat (59) during equilibration and the semiisotropic Parrinello–Rahman barostat (60) during the production runs.

All simulations were performed in GROMACS 4.6.7 (53, 54, 61, 62) using a time step of 20 fs.

Enhanced sampling details. We initially ran 10 independent simulations starting from the same structure, each 100-ns long. We saved configurations containing only the protein coordinates every 2 ns and aligned them in a self-consistent way to the average sampled configuration, removing translations and rotations with the Kabsch algorithm (56, 57). In particular, we first aligned the trajectory to an arbitrary configuration, calculated the average configuration, and used it to align the trajectory again, repeating the procedure until the rmsd between two consecutive average configurations was smaller than 0.01 nm. The alignment was done on the backbone beads of residues 3–28 of each monomer. Furthermore, we took into account the identity of the two monomers, which introduces an exchange symmetry in the system. We thus considered for every frame the structure with the smallest rmsd to the reference on swapping of the two monomers.

We calculated the first two DCs in the Cartesian space of the aligned configurations using the Euclidean metric and $\epsilon = 5$ nm and approximated the boundary of the obtained points with a convex hull. We projected each point on the boundary outward at a distance of $v = 5$ nm along the local PCA, which was calculated on its 100 nearest neighbors, with d_{loc} chosen to keep 95% of the original variance.

We added lipids around the projected dimer and solvated the resulting bilayer; then, we shortly equilibrated the system, obtaining 16 new configurations (Table 1). We then ran 10 independent 100-ns-long unbiased simulations for each new configuration by randomly initializing the velocities and merged the new trajectories with the initial ones.

After this first expansion, we repeated the entire procedure for a second and third time (expansions 2 and 3), obtaining 16 and 12, respectively, new configurations (Table 1). Newly discovered configurations where the two monomers are separated were excluded from the procedure, because they would dominate the representation in DC.

Global DMAP calculation. The FES as a function of the first two global diffusion coordinates shown in Fig. 4 was calculated using 24,815 configurations of the dimer sampled at equal times from a previously reported 2.52-ms-long equilibrium trajectory (23) that was self-consistently aligned to the average configuration as already explained.

To represent newly sampled configurations on the global DMAP landscape, we aligned them on the same average configuration described above and combined them with the configurations sampled from the long equilibrium trajectory, hence evaluating the DMAP on the combined set for each new trajectory. All DMAPs were calculated using the Euclidean metric and $\epsilon = 5$ nm.

Relative orientation calculation. The two angles u and v used in Fig. 5 define the relative rotation of the two alpha helices forming an Mga2 dimer; u is the angle defined counterclockwise between the orthogonal line to the direction connecting the centers of mass of the two helices and the vector pointing to residue W10 of the first monomer, and v is the angle defined counterclockwise between the orthogonal line to the direction connecting the centers of mass and the vector pointing to residue W10 of the second monomer (*Inset* in Fig. 5, *Lower*). Both angles have periodicity 2π. For chemically and structurally identical monomers, we would have mirror symmetry with respect to the line $u - v = 2\pi$.

The reference free-energy landscape of Fig. 5 was calculated by using 248,144 frames sampled at equal times from a 2.52-ms-long equilibrium trajectory (23). We calculated u and v both in the reference of the first monomer and the reference of the second monomer, in this way effectively

Table 1. Summary of performed MD simulations of the Mga2 dimer

Simulation	No. of starting structures	No. of simulations	Cumulative simulation time (μs)	*u*, *v* Discovered states	Cumulative count of dissociation events
1. Initial unbiased	1	10	1	1/4	0
2. Expansion 1	16	160	16	3/4	0
3. Expansion 2	16	150	15	4/4	1
4. Expansion 3	12	100	12	4/4	1
5. Reference unbiased	1	10	40	4/4	0
6. Equilibrium free energy	10	10	2,520	4/4	0

enforcing the monomer exchange symmetry of the system. Any residual deviation from the mirror symmetry in the surface is caused by some flexibility of the helices, which can be only approximately considered as rigid bodies.

Analysis and visualization of the data were performed with NumPy (63), SciPy (64), IPython (65), Matplotlib (66), and MDAnalysis (67). Molecular representations were made with VMD (68, 69).

ACKNOWLEDGMENTS. E.C., R.C., and I.G.K. acknowledge the hospitality and support of the Institute for Advanced Study Technical University of Munich. This work was partially supported by the US Air Force Office of Scientific Research (FA9950-17-1-00114), the US National Science Foundation (ECCS-1462241), and Defense Advanced Research Projects Agency (I.G.K.). E.C. acknowledges partial support of the Italian Ministry of Education through NANO-BRIDGE Project PRIN 2012 Grant 2012LHPSJC. R.C. and G.H. were supported by the Max Planck Society.

1. Abrams C, Bussi G (2013) Enhanced sampling in molecular dynamics using metadynamics, Replica-exchange, and temperature-acceleration. *Entropy (Basel)* 16:163–199.
2. Spiwok V, Sucur Z, Hosek P (2015) Enhanced sampling techniques in biomolecular simulations. *Biotechnol Adv* 33:1130–1140.
3. Weinan E, Vanden-Eijnden E (2010) Transition-path theory and path-finding algorithms for the study of rare events. *Annu Rev Phys Chem* 61:391–420.
4. Dellago C, Bolhuis PG (2009) Transition path sampling and other advanced simulation techniques for rare events. *Advanced Computer Simulation Approaches for Soft Matter Sciences III*, Advances in Polymer Science, eds Holm C, Kremer K (Springer, Berlin), Vol 221, pp 167–233.
5. Májek P, Elber R (2010) Milestoning without a reaction coordinate. *J Chem Theory Comput* 6:1805–1817.
6. A Beccara S, Škrbić T, Covino R, Faccioli P (2012) Dominant folding pathways of a WW domain. *Proc Natl Acad Sci USA* 109:2330–2335.
7. Abrams JB, Tuckerman ME (2008) Efficient and direct generation of multidimensional free energy surfaces via adiabatic dynamics without coordinate transformations. *J Phys Chem B* 112:15742–15757.
8. Chen M, Yu TQ, Tuckerman ME (2015) Locating landmarks on high-dimensional free energy surfaces. *Proc Natl Acad Sci USA* 112:3235–3240.
9. Maragliano L, Vanden-Eijnden E (2006) A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem Phys Lett* 426:168–175.
10. Zheng W, Rohrdanz M, Clementi C (2013) Rapid exploration of configuration space with diffusion-map-directed molecular dynamics. *J Phys Chem B* 117:12769–12776.
11. Rohrdanz MA, Zheng W, Clementi C (2013) Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annu Rev Phys Chem* 64:295–316.
12. Lei H, Baker N, Li X (2016) Data-driven parameterization of the generalized langevin equation. *Proc Natl Acad Sci USA* 113:14183–14188.
13. Coifman R, et al. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc Natl Acad Sci USA* 102:7426–7431.
14. Tribello G, Bonomi M, Branduardi D, Camilloni C, Bussi G (2014) Plumed2: New feathers for an old bird. *Comput Phys Commun* 185:604–613.
15. Fiorin G, Klein ML, Hénin J (2013) Using collective variables to drive molecular dynamics simulations. *Mol Phys* 111:3345–3362.
16. Frewen T, Hummer G, Kevrekidis IG (2009) Exploration of effective potential landscapes using coarse reverse integration. *J Chem Phys* 131:134104.
17. Lafon S (2004) Diffusion maps and geometric harmonics. PhD thesis (Yale University, New Haven, CT).
18. Allgower EL, Schmidt PH (1985) An algorithm for piecewise-linear approximation of an implicitly defined manifold. *SIAM J Numer Anal* 22:322–346.
19. Dankowicz H, Schilder F (2013) *Recipes for Continuation* (Society for Industrial and Applied Mathematics, Philadelphia, PA).
20. Doedel E, Oldeman BE (2009) *AUTO07p: Continuation and Bifurcation Software for Ordinary Differential Equations* (Concordia Univ, Montreal).
21. Hummer G, Kevrekidis IG (2003) Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J Chem Phys* 118:10762–10773.
22. Duan Y, et al. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24:1999–2012.
23. Covino R, et al. (2016) A Eukaryotic sensor for membrane lipid saturation. *Mol Cell* 63:1–11.
24. Marrink S, Risselada H, Yefimov S, Tieleman D, De Vries A (2007) The martini force field: Coarse grained model for biomolecular simulations. *J Phys Chem B* 111:7812–7824.
25. Monticelli L, et al. (2008) The martini coarse-grained force field: Extension to proteins. *J Chem Theory Comput* 4:819–834.
26. Edelsbrunner H, Kirkpatrick D, Seidel R (1983) On the shape of a set of points in the plane. *IEEE Trans Inf Theory* 29:551–559.
27. Edelsbrunner H, Mucke P (1994) Three-dimensional alpha shapes. *ACM Trans Graph* 13:43–72.
28. Edelsbrunner H (2003) *Surface Reconstruction by Wrapping Finite Sets in Space*, eds Aronov B, Basu S, Pach J, Sharir M (Springer, Berlin), pp 379–404.
29. Gear C, Chiavazzo E, Kevrekidis IG (2016) Manifolds defined by points: Parameterizing and boundary detection (extended abstract). *AIP Conf Proc* 1738:1–4.
30. Xia C, Hsu W, Lee L, Ooi B (2006) BORDER: Efficient computation of boundary points. *IEEE Trans Knowl Data Eng* 18:289–303.
31. Chiavazzo E, Gear CW, Dsilva CJ, Rabin N, Kevrekidis IG (2014) Reduced models in chemical kinetics via nonlinear data-mining. *Processes* 2:112–140.
32. Dsilva C, Talmon R, Rabin N, Coifman R, Kevrekidis IG (2013) Nonlinear intrinsic variables and state reconstruction in multiscale simulations. *J Chem Phys* 139:184109.
33. Bolhuis PG, Dellago C, Chandler D (2000) Reaction coordinates of biomolecular isomerization. *Proc Natl Acad Sci USA* 97:5877–5882.
34. Tribello GA, Ceriotti M, Parrinello M (2010) A self-learning algorithm for biased molecular dynamics. *Proc Natl Acad Sci USA* 107:17509–17514.
35. Lelimousin M, Limongelli V, Sansom MSP (2016) Conformational changes in the epidermal growth factor receptor: Role of the transmembrane domain investigated by coarse-grained metadynamics free energy calculations. *J Am Chem Soc* 138:10611–10622.
36. Chiavazzo E, Karlin I (2011) Adaptive simplification of complex multiscale systems. *Phys Rev E* 83:036706.
37. Chiavazzo E (2012) Approximation of slow and fast dynamics in multiscale dynamical systems by the linearized relaxation redistribution method. *J Comput Phys* 231:1751–1765.
38. Kevrekidis IG, Gear C, Hummer G (2004) Equation free: The computed-aided analysis of complex multiscale systems. *AIChe J* 50:1346–1355.
39. Kevrekidis IG, et al. (2003) Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis. *Commun Math Sci* 1:715–762.
40. Dsilva C (2015) Manifold learning for dynamical systems. PhD thesis (Princeton University, Princeton).
41. Singer A, Erban R, Kevrekidis IG, Coifman R (2009) Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proc Natl Acad Sci USA* 106:16090–16095.
42. Pavliotis G, Stuart A (2007) Parameter estimation for multiscale diffusions. *J Stat Phys* 127:741–781.
43. Kalliadasis S, Krumscheid S, Pavliotis G (2015) A new framework for extracting coarse-grained models from time series with multiscale structure. *J Comput Phys* 296:314–328.
44. Krumscheid S, Pradas M, Pavliotis G, Kalliadasis S (2015) Data-driven coarse graining in action: Modeling and prediction of complex systems. *Phys Rev E* 92:042139.
45. Calderon CP (2007) Fitting effective diffusion models to data associated with a "glassy" potential: Estimation, classical inference procedures, and some heuristics. *Multiscale Model Simul* 6:656–687.
46. Li J, Kevrekidis P, Gear C, Kevrekidis IG (2007) Deciding the nature of the coarse equation through microscopic simulations: The baby-bathwater scheme. *SIAM Rev* 49:469–487.
47. Belkin M, Que Q, Wang Y, Zhou X (2012) Graph laplacians on singular manifolds: Toward understanding complex spaces: Graph laplacians on manifolds with singularities and boundaries. arXiv:1211.6727.
48. Deutsch S, Medioni G (2017) Learning the geometric structure of manifolds with singularities using the tensor voting graph. *J Math Imaging Vis* 57:402–422.
49. Rohrdanz M, Zheng W, Maggioni M, Clementi C (2011) Determination of reaction coordinates via locally scaled diffusion maps. *J Chem Phys* 134:124116.
50. Peters B (2016) Reaction coordinates and mechanistic hypothesis tests. *Annu Rev Phys Chem* 67:669–690.
51. Frewen TA, et al. (2011) *Coarse Collective Dynamics of Animal Groups*, eds Gorban AN, Roose D (Springer, Berlin), pp 299–309.
52. Sonday BE, Haataja M, Kevrekidis IG (2009) Coarse-graining the dynamics of a driven interface in the presence of mobile impurities: Effective description via diffusion maps. *Phys Rev E* 80:031102.
53. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulations. *J Chem Theory Commun* 4:435–447.
54. van der Spoel D, et al. (2005) Gromacs: Fast, flexible, and free. *J Comput Chem* 26:1701–1718.
55. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126:014101.
56. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A* 32:922–933.
57. Kabsch W (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A* 34:827–828.
58. Wassenaar T, Ingólfsson H, Böckmann R, Tieleman D, Marrink S (2015) Computational lipidomics with insane: A versatile tool for generating custom membranes for molecular simulations. *J Chem Theory Comput* 11:2144–2155.
59. Berendsen H, Postma J, van Gunsteren W, DiNola A, Haak J (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684–3690.
60. Parrinello M, Rahman A (1980) Crystal structure and pair potentials: A molecular-dynamics study. *Phys Rev Lett* 45:1196–1199.
61. Abraham M, et al. (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2:19–25.

62. Pronk S, et al. (2013) Gromacs 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29:845–854.
63. van der Walt S, Colbert S, Varoquaux G (2011) The numpy array: A structure for efficient numerical computation. *Comput Sci Eng* 13:22–30.
64. Jones E, Oliphant T, Peterson P, et al. (2001) Scipy: Open source scientific tools for Python. Available at www.scipy.org/. Accessed November 12, 2015.
65. Perez F, Granger B (2007) Ipython: A system for interactive scientific computing. *Comput Sci Eng* 9:21–29.
66. Hunter J (2007) Matplotlib: A 2d graphics environment. *Comput Sci Eng* 9:90–95.
67. Michaud-Agrawal N, Denning E, Woolf T, Beckstein O (2011) Mdanalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* 32:2319–2327.
68. Humphrey W, Dalke A, Schulten K (1996) Vmd: Visual molecular dynamics. *J Mol Graph* 14:33–38.
69. Stone JE (1995) An efficient library for parallel ray tracing and animation. Master's thesis (University of Missouri, Rolla, MO).