JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

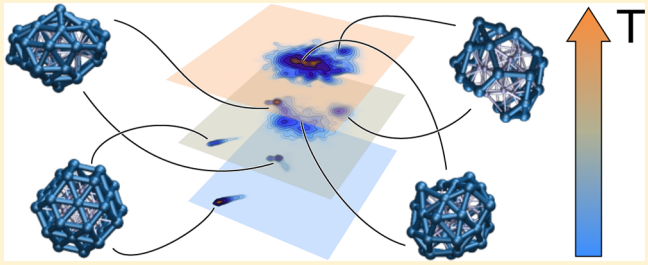
Demonstrating the Transferability and the Descriptive Power of Sketch-Map

Michele Ceriotti,*[,†] Gareth A. Tribello,[‡] and Michele Parrinello[‡]

[†]Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QZ, United Kingdom

[‡]Computational Science, Department of Chemistry and Applied Biosciences, ETH Zurich and Facoltà di Informatica, Instituto di Scienza Computationali, Università della Svizzera Italiana, Via Giuseppe Buffi 13, CH-6900, Lugano, Switzerland

ABSTRACT: Increasingly, it is recognized that new automated forms of analysis are required to understand the high-dimensional output obtained from atomistic simulations. Recently, we introduced a new dimensionality reduction algorithm, sketch-map, that was designed specifically to work with data from molecular dynamics trajectories. In what follows, we provide more details on how this algorithm works and on how to set its parameters. We also test it on two well-studied Lennard-Jones clusters and show that the coordinates we extract using this algorithm are extremely robust. In particular, we demonstrate that the coordinates constructed for one particular Lennard-Jones cluster can be used to describe the configurations adopted by a second, different cluster and even to tell apart different phases of bulk Lennard-Jonesium.



1. INTRODUCTION

Atomistic simulation methodologies are now frequently used to shed light on the atomic scale mechanisms that underlie experimentally observed phenomena. However, as the systems examined using simulations become progressively more and more complicated, the sheer resolution of the data that is obtainable from a simulation begins to present a problem. Atomistic simulations, by their very nature, provide high-dimensionality data, which oftentimes can only be interpreted by using physical/chemical intuition obtained from experiments. This is obviously problematic if we want to predict new chemical structures or novel reaction mechanisms based on simulations alone. Hence, there is a growing interest in using machine learning algorithms and smart visualization software to generate simplified representations of the data obtainable from atomistic simulations so that it can be more easily understood and interpreted by a human user.

Recently, we developed a new approach, sketch-map,[1,2] for visualizing the results from molecular dynamics (MD) and enhanced sampling simulations. In this approach, we use the high-dimensionality trajectory data obtained from an MD or enhanced sampling calculation to construct a two-dimensional representation of the free energy surface (FES). This representation is generated by first selecting a set of landmark frames from the trajectory and by then endeavoring to map out the spatial relationships between them in a lower-dimensionality space. The free energy, as a function of these bespoke collective variables (CVs), can then be calculated by projecting the remainder of the trajectory using an out-of-sample procedure.[2]

The free energy surfaces obtained with sketch-map coordinates provide a far richer view of the FES and the many basins that comprise it than those obtained when CVs based on physical intuition alone are used. Furthermore, biasing potentials can be constructed as a function of sketch-map coordinates in order to facilitate rapid exploration of configuration space.[2] A concern, particularly when sketch-map is used to generate bias potentials, is the extent to which these coordinates can discriminate between configurations that were not represented in the set of landmark frames from which the initial map was constructed. Sketch-map should give a detailed picture of the landscape in the immediate vicinity of the landmark points. However, if this is all it can do then it can only be used when a very thorough sampling of the free energy landscape is available. If this sort of detailed data is available, then further biased sampling on the landscape is probably not required. By contrast, if it is possible to use sketch-map coordinates to map configurations that are far from all of the landmark training points, then it is easy to conceive of enhanced sampling methods based on biasing sketch-map coordinates generated from an initial, cursory exploration of the energetically accessible parts of configuration space.[3−5] In addition, because sketch-map coordinates can identify unexpected stable configurations, they are particularly suitable for visualizing how the occupations of all the basins in the free energy landscape change when the underlying chemical system is perturbed. This sort of analysis could help when it comes to understanding how mutations or the presence of denaturant molecules affects protein structure or to understand how the free energy changes as systems cross phase boundaries. This analysis could even be used to understand the subtle differences

Received: December 2, 2012
Published: February 4, 2013

   dx.doi.org/10.1021/ct3010563 | *J. Chem. Theory Comput.* 2013, 9, 1521−1532

in the free energies obtained when the same system is simulated with two different force fields. However, when performing these sorts of comparisons it is important to remember that mappings from high to low dimensionality are generally not one-to-one. As such, the difference in free energy between two points on the surface is not a measure of the relative probability of the underlying configurations. The correct way to measure relative free energies is to define *regions*, $N(A)$, in CV space, that correspond to fluctuations around structures of interest. Integrating the probability distribution over these regions using $F_A = -k_B T \ln \int_{N(A)} \exp[-F(s)/k_B T]\, ds$ gives free energies, F_A, for the configurations that are well-defined and independent of the choice of CV.

To test the transferability of our sketch-map coordinates, we chose to re-examine some of the most extensively studied Lennard-Jones clusters.[6,7] For these simple systems, it is possible to calculate the positions of all the minima and transition states in the potential energy surfaces (PES) and to connect them all together to generate a disconnectivity graph that gives a sense of the global shape of the energy landscape and hence the properties of the model. In what follows, we will start by showing that analyzing the results of parallel tempering calculations using sketch-map gives results that are in agreement with what would be expected given the structure of the disconnectivity graph. We then show how the sketch-map coordinates constructed for LJ38 can be used to understand this system at a range of temperatures, to understand the physics of a second, completely different cluster, and even to tell apart different phases of bulk Lennard-Jonesium. These results provide a confirmation that coordinates generated by sketch-map are extremely robust. Clearly, they most definitely can classify structures from outside the initial, fitted set of landmarks and can thus be used in a wide variety of different contexts.

2. BACKGROUND

A large number of dimensionality reduction algorithms have been used to understand the high-dimensionality data output by atomistic simulations. These algorithms vary in sophistication, but all of them make assumptions about the way low-energy configurations are distributed across phase space, as illustrated in Figure 1.

The algorithm that is most commonly used to map trajectory data is principal component analysis (PCA).[8−10] This algorithm projects the high dimensionality data on the eigenvectors corresponding to the largest eigenvalues of the covariance matrix and assumes that the low-energy regions lie in a linear (d-dimensional) subspace of the full (D-dimensional) space as illustrated in Figure 1a. This assumption of global-linearity is not required when nonlinear manifold learning algorithms such as locally linear embedding,[11] Isomap,[12,13] and diffusion maps[14−17] are employed. However, as illustrated in Figure 1b, these algorithms still for the most part assume that around each point there is a neighborhood of fixed size where the accessible part of phase space resembles a d-dimensional Euclidean space. In our recent paper, we provided evidence that this assumption is invalid for data taken from a typical atomistic simulation. An analysis of the histogram of distances between trajectory frames showed that locally the distribution of points resembles that of a multivariate Gaussian in the high-dimensionality space. This distribution is compatible with our view of the energy landscape as being composed of energetic basins, in which the system fluctuates in the full dimensionality space about some mean

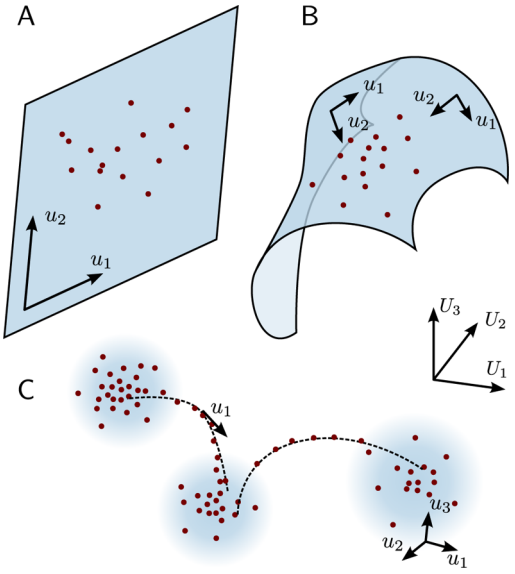


Figure 1. Configurations of data that can be visualized using dimensionality reduction algorithms. Panel A shows the sort of problem that can be tackled with PCA and linear methods. The points lie within a linear subspace (in this case a plane) in the full, three-dimensional space. Panel B shows the sort of problem that can be tackled with nonlinear manifold learning algorithms. The points lie on a curved surface, which at every point resembles a two-dimensional plane and which has been relatively uniformly sampled. Panel C shows the sort of data we obtain from a molecular dynamics trajectory. There are basins that are densely sampled and high-dimensional and transition pathways that are relatively poorly sampled and lower dimensional.

structure, that are then connected by a web of narrow transition pathways. Similar observations have also been reported by other researchers.[18] Two problems, which affect the performance of many manifold learning algorithms, will arise if the data has this structure. First, there will be a lot of noise in the vicinity of the energetic basins as a consequence of the thermal fluctuations. Second, and more importantly, there will be poor sampling at the transition states because the energy in these regions is far higher than the energy in the basins.

We recently introduced the sketch-map algorithm, which was designed with the problems discussed in the previous paragraph in mind. This algorithm is based on metric multidimensional scaling (MDS), which is at the heart of many other dimensionality reduction algorithms.[19] MDS generates a set of projections $\{x_i\}$ from a set of high-dimensionality landmark points $\{X_i\}$ by minimizing the following stress function:

$$\chi^2 = \sum_{i \neq j} [R_{ij} - r_{ij}]^2 \quad (1)$$

where R_{ij} is a measure of the dissimilarity between the high dimensional points X_i and X_j and r_{ij} is the Euclidean distance between their projections. Clearly, to minimize this stress, the projections, $\{x_i\}$, have to be arranged in the low-dimensional space so that the Euclidean distances between them match the dissimilarities between the high-dimensional points.

The Euclidean distance between points is the easiest and most natural quantity to use for the R_{ij} values in eq 1. However, reproducing these distances by arranging points in a lower dimensional space is difficult when we are dealing with trajectory data. On short length scales, there are high dimensionality features in the data because of thermal

dx.doi.org/10.1021/ct3010563 | J. Chem. Theory Comput. 2013, 9, 1521−1532

fluctuations. Worse still, if we rewrite the r_{ij} values in eq 1 in terms of R_{ij} and a relative error ε_{ij} (i.e., as $r_{ij} = R_{ij}(1 + \varepsilon_{ij})$), we find that each R_{ij} contributes $R_{ij}^2\varepsilon_{ij}^2$ to the final stress. That is to say, the greater the value of R_{ij}, the greater the penalty incurred in projections where $r_{ij} \neq R_{ij}$. This is far from ideal because longer Euclidean distances are unlikely to take nonlinear features into account as shown in Figure 1. In fact, many algorithms deliberately remove these longer Euclidean distances by either changing the way dissimilarity is measured[12,19] or by introducing weights so these distances contribute less to the stress.[20] Our way to resolve these problems is to introduce two sigmoid "filter" functions and to rewrite the stress as

$$\chi^2 = \sum_{i \neq j} [F(R_{ij}) - f(r_{ij})]^2 \tag{2}$$

where

$$f(r) = 1 - (1 + (2^{a/b} - 1)(r/\sigma)^a)^{-b/a} \tag{3}$$

and

$$F(R) = 1 - (1 + (2^{A/B} - 1)(R/\sigma)^A)^{-B/A} \tag{4}$$

These filters, $F(R)$ and $f(r)$, transform all the distances to values between zero and one. Distances less than σ are transformed to something similar to zero, while distances greater than σ are transformed to something close to one. As such, the difference $F(R_{ij}) - f(r_{ij})$ is small if both $R_{ij} < \sigma$ and $r_{ij} < \sigma$ or if both $R_{ij} > \sigma$ and $r_{ij} > \sigma$, even when $R_{ij} \neq r_{ij}$. Hence, rather than searching for projections in which all the distances between points are reproduced, the main aim in sketch-map is to ensure that points closer than σ are projected close together while those farther apart than σ are projected far apart. To see this point more clearly, it is useful to look again at the contribution that small relative errors, ε_{ij}, make to the stress function. Assuming $F \equiv f$, each contribution to the stress reads $[f(R_{ij}) - f(R_{ij} + \varepsilon_{ij}R_{ij})]^2$. Expanding the second term in this expression using a Taylor expansion in ε_{ij} gives $[f'(R_{ij})]^2R_{ij}^2\varepsilon_{ij}^2$ to leading order in ε_{ij}. This function is strongly peaked around σ, so small differences between the r_{ij} and R_{ij} values contribute more to the stress when $R_{ij} \approx \sigma$. This solves many of the problems discussed above as now the algorithm expends little effort on accurately reproducing the shortest and longest distances. In fact, these distances can be significantly distorted in the low-dimensional projection. This is good though because, as discussed above, strict constraints on these distances are often detrimental to the performance of the MDS algorithm.

Details on how to select parameters for sketch-map can be found in Appendix A and Appendix B. However, the parameter that most dramatically affects the results from a sketch-map calculation is σ. Any algorithm that works by minimizing an equation-2-like stress function will make the constraints on the shortest and longest distances less stringent. Therefore, when we set σ, we are essentially deciding what features will be displayed in the projection. In all probability it will not be possible to see the structure on length scales less than σ, as sketch-map will make little to no effort to accurately reproduce these distances. The value of σ should then be chosen by examining the data and deciding what short-range features can be safely ignored. For trajectory data, this is often the internal structure of the energetic basins—i.e., the thermal fluctuations.

The sums in eq 2 cannot run over the entire simulation trajectory as the cost of this analysis would then scale quadratically with simulation length. In fact, since we use the "point-wise global" optimization strategy described in ref 1, the cost would scale with the cube of the number of trajectory frames. To avoid this computational overhead, we start by projecting a subset of the points—the landmarks. A projection, x, for each of the remaining high dimensional points, X, is then generated by optimizing eq 2 for the landmarks and the additional point with fixed values for the projections of the landmarks. These constraints allow us to remove many terms from eq 2, which simplifies to

$$\delta^2(x) = \sum_{i=1}^{N} \{F[|X - X_i|] - f[|x - x_i|]\}^2 \tag{5}$$

The loop runs over the set of landmark frames, $|X-X_i|$ is the distance between the position of the ith landmark and the frame that is being projected, and $|x-x_i|$ is the distance between the projection of the frame and the projection of the ith landmark. When the new point X is in the vicinity of some of the landmarks, it is projected close to their projections. This is similar behavior to other, arguably simpler approaches[21,22] for performing the out-of-sample embedding that write the projection of a new high-dimensional configuration, X, as a weighted average of the projected landmarks, $x = \sum_i w_i x_i / \sum_i w_i$, where $w_i = \exp(-|X - X_i|/\lambda)$. Our new algorithm only comes into its own when the new point, X, is distant from all of the landmarks. With approaches based on weighted sums, the projection is by construction forced to lie within the smallest convex set containing all of the landmarks. As a result, points that are far away from all the landmarks end up being projected in the middle of the map (see Figure 2). This constraint is removed in our procedure because we do an explicit minimization of the stress function. Furthermore, when X is distant from all of the landmark points every distance $|X-X_i|$ in eq 5 is greater than σ and well into the tail region of $F(R)$, where $1 - F(R) \propto R^{-B}$ and $1 - f(r) \approx r^{-b}$, as discussed in Appendix B. As such, when we minimize eq 5 we are requiring each $|x-x_i|$ to be proportional to $|X-X_i|^{B/b}$. So when $b = B$, the algorithm insists that the new point should be projected so that the distances between the out-of-sample point and the landmarks, $|X-X_i|$, are the same as the distances between the

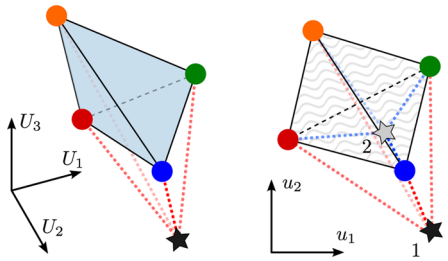


Figure 2. Figures showing how well sketch-map's out of sample procedure works. The left panel shows four points in three dimensions, while the right panel shows, using the same color code, projections of these points in two dimensions that were generated by minimizing eq 1. Projections for the star in the left panel were generated by minimizing a stress function and by using the weighted average described in the text. These projections are shown in black and gray, respectively. The projection generated by minimizing the stress function (black) is distant from all the other points in agreement with what is observed in the three-dimensional figure. In contrast, when the weighted sum is used the black star in the left panel is projected in between the four colored points.

dx.doi.org/10.1021/ct3010563 | J. Chem. Theory Comput. 2013, 9, 1521−1532

projection and the projections of the landmarks, $|x-x_i|$. In this special case, sketch-map's out-of-sample procedure behaves analogously to the algorithms used in GPS navigation systems to determine positions by measuring distances from a network of far-away satellites. This sensible treatment of points that are distant from the landmarks is important because, as discussed in what follows, it makes sketch-map coordinates extraordinarily resilient.

3. METHODS

We chose to examine two clusters of Lennard-Jones atoms in this work: the 38 and 55 atom clusters. Lennard-Jones 55 (LJ55) has a funnel shaped landscape with an easy to find global minimum.[23,24] This structure in the energy landscape ensures that at high temperatures there is a transition from an ordered, solid-like phase to a liquid-like phase because of the interplay between energy and entropy. By contrast, Lennard-Jones 38 (LJ38) has a doubly funneled energy landscape.[25] There is thus a solid−solid transition at moderate temperatures and a subsequent solid−liquid transition at higher temperatures. The solid−solid transition occurs because the energy landscape in the high-temperature phase is flatter, which ensures that the entropic contribution to the free energy of this structure is higher.

To sample the energy landscapes for the two clusters, we performed extensive parallel tempering calculations using Gromacs-4.5.5[26] patched with Plumed-1.3.[27] We set σ and ε in the Lennard-Jones potential and the mass, m, of the atoms equal to one and thus use time units, $t^* = (\varepsilon/m\sigma^2)^{1/2}$, and temperature units, $T^* = k_B T/\varepsilon$, throughout this paper. In our calculations, we used a time step of $0.001t^*$ and kept the temperatures fixed using the global thermostat of Bussi et al.[28] (relaxation time equal to 0.1 t^*). For both systems, a geometric distribution of temperatures was used, and swapping moves were attempted every 100 steps. For LJ38, 16 replicas were used, while for LJ55, 10 replicas were used. A restraining potential $(\kappa(r - r_0)^4$ with $\kappa = 0.4\varepsilon)$ was included to prevent sublimation of the clusters at higher temperatures.[29,30] This potential acts if the distance, r, between any atom and the center of mass of the cluster become greater than r_0. For LJ55, r_0 was set equal to 3.0σ, while for LJ38, r_0 was set equal to 2.25σ.

The first step in applying sketch map is to devise a high-dimensional description that takes the peculiarities of the system into account. Many of the dimensionality reduction algorithms that have been used to analyze trajectory data create their low-dimensionality maps based on the RMSD distances between a subset of the trajectory frames.[13,17,18] This is not a sensible approach for Lennard-Jones clusters as it does not incorporate the symmetry due to interchange of labels. Hence, in what follows we represent each configuration as a discretized probability distribution.[31,32] In our case, each of the values in this vector is calculated using

$$s_i = \frac{1}{N}\sum_{j=1}^{N}\int_{i-1/2}^{i+1/2} \mathrm{d}c \quad K(c - c_j) \tag{6}$$

where

$$K(x) = \begin{cases} 0 & |x| \geq 1/2 \\ (2 - 2|x|) & |x| < 1/2 \end{cases} \tag{7}$$

$$c_j = \sum_{k=1,k\neq j}^{N} c((r_{jk} - r_1)/(r_0 - r_1)) \tag{8}$$

and

$$c(y) = \begin{cases} 1 & y \leq 0 \\ 0 & y \geq 1 \\ [(y-1)^2(1+2y)] & 0 < y < 1 \end{cases} \tag{9}$$

In these expressions, the sums run over all the atoms in the system, r_{jk} is the distance between atom j and atom k, and the parameters r_0 and r_1 were set equal to 1.5σ and 1.3σ, respectively. Figure 3 shows that each s_i value measures the fraction of atoms with a coordination number between $i - 1/2$ and $i + 1/2$. Variables of this sort are clearly invariant to changes of labeling. Furthermore, they are able to detect changes in the coordination environment of the atoms caused by changes in the surface-to-bulk ratio, by phase transitions that alter the dominant crystal structure, or by the creation and annihilation of defects.

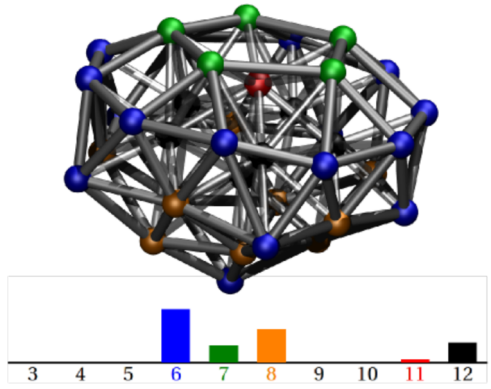


Figure 3. An explanation of the CVs we use to characterize the structures we find. Each CV measures the fraction of atoms with a particular coordination number. In the upper panel, the atoms are colored according to their coordination number. Atoms with low coordination numbers are found on the surface of the cluster, while those with higher coordination numbers are found in the bulk. The distribution of coordination numbers thus measures the surface to bulk ratio. The bar chart in the bottom panel shows the distribution of coordination numbers and serves as a key for the colors in the upper panel.

4. RESULTS

4.1. LJ38 at the Melting Point. There is a peak in the heat capacity curve for LJ38 at approximately $T_M = 0.18T^*$.[7] At this temperature, the cluster visits both solid-like and liquid-like configurations, which makes finding a low-dimensional representation particularly challenging. To construct the projection with sketch-map, we randomly selected 500 landmark points and set the sketch-map parameters equal to $\sigma = 0.125$, $A = 8$, $a = 1$, and $B = b = 2$ for the reasons discussed in Appendix B. The final result of the optimization is shown in Figure 4. [The parallel tempering calculations that were used to generate this data took about 112 h on 16 parallel processors. The sketch-map analysis then took about 3 h on a single node. Even for this computationally inexpensive system, the analysis of the data represents a tiny fraction of the total computational cost.] Sketch-map is able to clearly separate landmarks that

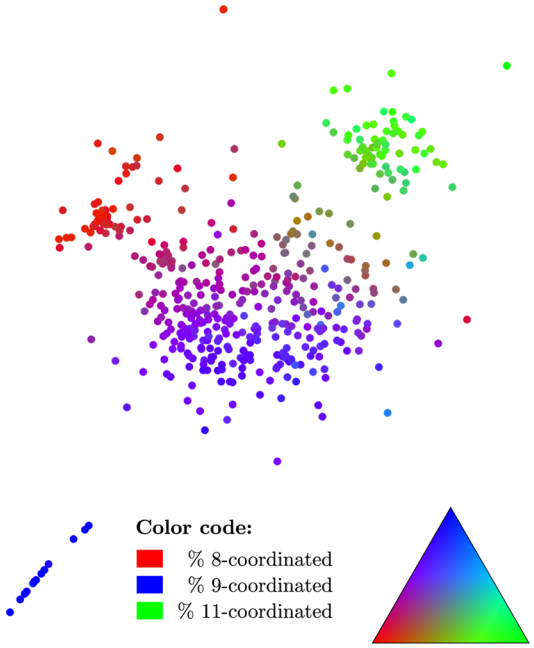


Figure 4. Projections for the set of landmark structures of LJ38 obtained from the melting temperature trajectory. Colors for the points are generated by taking the fractions of 8, 9, and 11 coordinated atoms and renormalizing so that these three components of the histogram sum to one. These renormalized values are then used to specify the degree to which red, blue, and green contribute to the final color, as illustrated in the key.

have different distributions of coordination environments, which is encouraging. We thus went ahead and projected the remainder of the points from the trajectory onto these coordinates and constructed the free energy surface shown in Figure 5. Many of the features one would expect given the structure of the disconnectivity tree are visible in this free energy surface. A minimum corresponding to the face-centered-cubic (fcc), truncated-octahedral global minima appears in the bottom left-hand corner of the surface. This feature is extended in the projection because at this temperature defective versions of this structure are energetically accessible. In fact, if you examine Figure 5, you can see a second minimum in the free energy near the fcc basin that corresponds to a particularly prevalent, defective version of this structure. The second lowest energy minimum in this energy landscape[25]—the incomplete Mackay icosahedron—appears in the top left corner of the FES. Once again, this structure is surrounded by minima corresponding to defective versions. Furthermore, at this temperature one of these defective minima is the most stable state. It is important to note that the sketch-map coordinates do not contain an explicit description of the pathway connecting the fcc and the icosahedral minima. This is not a failure of sketch-map however. The highest energy transition state on the lowest-energy pathway connecting these two states is more than 4ϵ higher in energy than the fcc minimum.[33] As such, the probability of adopting this configuration is vanishingly small. As a result, sketch-map fails to map out this pathway because the data it would require to do this is simply not there—these high energy configurations do not appear in the trajectory output by the replica at this temperature.

In Figure 5, the molten state appears in the center of the projection. It is composed of two broad featureless basins that are separated by a non-negligible barrier. The larger of these two basins is clearly a highly defective version of the incomplete Mackay icosahedron. However, the basin in the top right corner of the free energy surface does not resemble the structure in the fcc or in the icosahderal minima and is instead characterized by a large fraction of 11 coordinated atoms. Sketch map also identifies a shallow, metastable minimum with 5-fold symmetry. The fact that this high free-energy structure can be clearly distinguished further demonstrates the extent to which sketch-map variables capture the details in the free energy landscape.

4.2. Temperature Dependence for LJ38. The sketch-map coordinates in the previous section were constructed from landmarks selected at random from an extensive and well converged simulation. As such, the distribution of landmarks accurately reflects the underlying free energies of the configurations. In many cases, particularly when we want to use sketch-map to generate coordinates for enhanced sampling, we do not have access to data this rich. It is, therefore, important to understand how the quality of the landmarks affects the coordinates generated by sketch-map and to understand if sketch-map coordinates can also describe regions of configurational space where there are few or no landmarks.

When the temperature is raised, the trajectory will contain more liquid-like configurations and fewer solid-like configurations. The opposite will happen when the temperature is lowered. As such, a good way to test the robustness of the sketch-map coordinates is to construct a map at $T \ll T_M$ and to see how well it describes the free energy surface at T_M or higher. Among the landmarks used to construct the low temperature map, there will clearly be very few of the liquid-like configurations that will be populated at the higher temperature. Consequently, if the low-temperature coordinates can project the higher-temperature data sensibly, it suggests that they are very robust.

To quantitatively assess the quality of any projection, we use the following expression:

$$\chi^2 = \frac{1}{N(N-1)} \sum_{i,j=1}^{N} [F(|X_i - X_j|) - f(|x_i - x_j|)]^2 \quad (10)$$

where the X_is' are 10 000 of the high-dimensional configurations that are being projected, and the x_i's are their embeddings. This quantity is basically the sketch-map stress function (eq 2) computed for the points we are trying to embed so a small value implies that the out-of-sample points are being arranged relative to each other in a sensible manner. This is a particularly stringent test of the performance of the out-of-sample embedding procedure as the projection of each point is generated by optimizing eq 5—i.e. by trying to reproduce the position of each point relative to the landmarks. The distances between the out-of-sample points relative to each other are not used in the fitting procedure, so when the transformed distances are reproduced it suggests that the sketch-map coordinates capture the essential features in the free energy landscape. In addition, the fact that $[F(|X_i - X_j|) - f(|x_i - x_j|)]^2$ is a number between zero and one that is only equal to one when $|X_i - X_j| < \sigma$ and $|x_i - x_j| > \sigma$ or vice versa means that we can interpret eq 10 as the fraction of distances for which the out-of sample procedure has failed spectacularly.

Figure 6 shows free energy surfaces at the melting temperature, $T_M \approx 0.180T^*$, at a temperature well below T_M, $0.135T^*$, and at a temperature well above T_M, $0.225T^*$. To create this figure, three sets of sketch-map coordinates were

dx.doi.org/10.1021/ct3010563 | J. Chem. Theory Comput. 2013, 9, 1521−1532

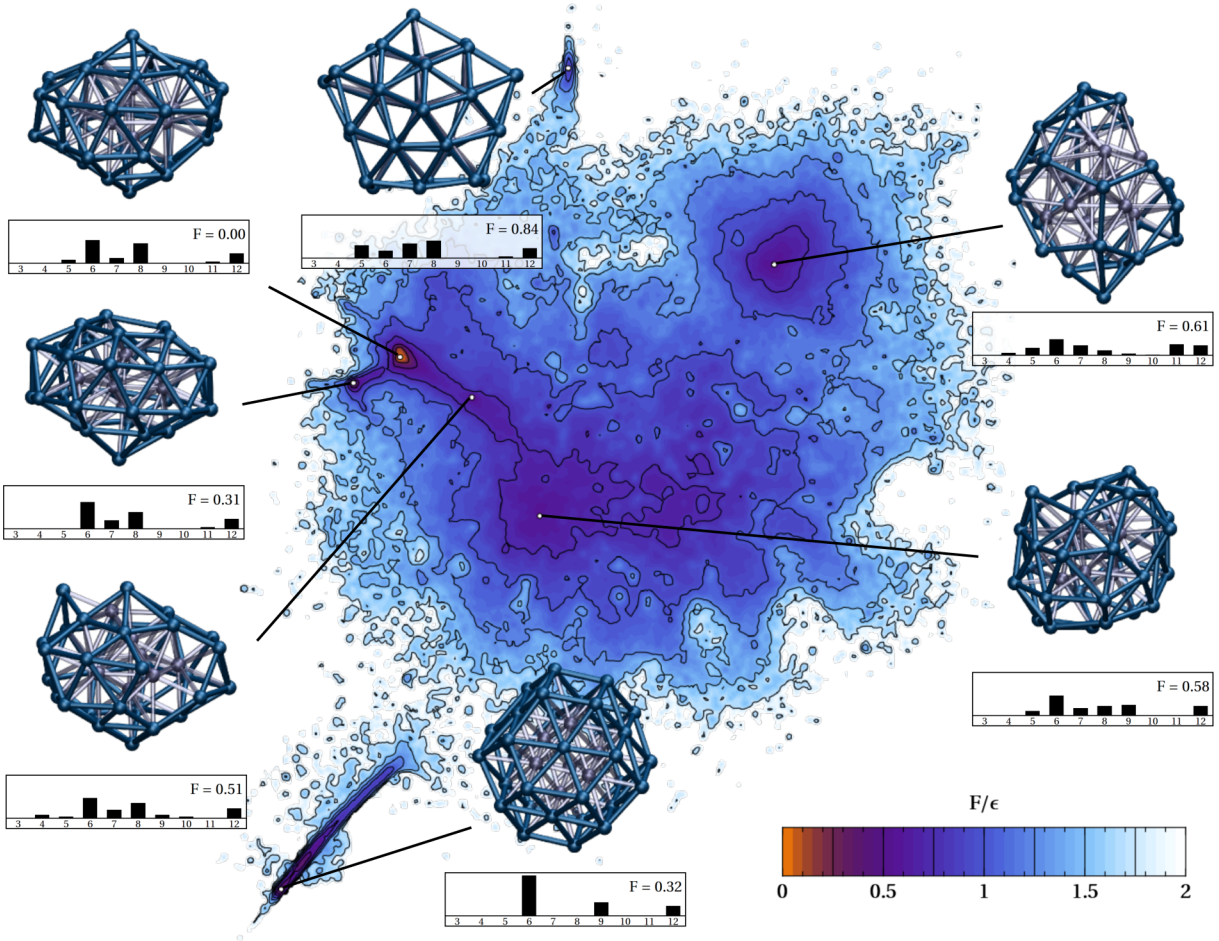


Figure 5. The free energy surface for LJ38 at $0.18T^*$ as a function of the sketch-map coordinates. This temperature is close to the peak in the heat capacity curve, so both the liquid and solid phases have substantial occupancies. This figure also shows where a number of representative configurations of LJ38 are projected together with their coordination number histograms.

generated by randomly selecting landmarks from the configurations visited by the replicas at these three temperatures. The nine free energy surfaces shown in Figure 6 are the result of projecting the data at each temperature on each of the three sets of sketch-map coordinates. It is clear that all three sets of sketch-map coordinates qualitatively represent the essential features of the free energy landscape across the phase transition. Even in the worst cases, the residual stress shows that less than 10% of the distances have $|X_i - X_j| > \sigma$ and $|x_i - x_j| < \sigma$ or vice versa.

The lowest-stress projections are those on the diagonal in Figure 6. These projections were constructed by randomly selecting landmarks from a trajectory at the temperature of interest. Clearly, having landmarks distributed in a way that reflects the underlying free energy does make a difference. The difference it makes is small, however, as is demonstrated by the free energy surfaces shown in the central column of Figure 6. These projections were generated using sketch-map coordinates constructed using randomly selected landmarks from the T_M trajectory. When this map is used, the stress is low for the projection of both the high and low temperature data because the set of landmarks contains representatives from both the solid and liquid parts of configuration space.

In Figure 6, the same qualitative picture of the physics emerges when the free energy surface is projected using sketch-map coordinates generated at any one of the three temperatures. All three maps show clearly that the solid-like configurations dominate at the low-temperatures and that the liquid-ones dominate at high-temperature. Furthermore, all three maps manage to separate the various defective versions of the incomplete Mackay icosahedron. This is remarkable, as it would be almost impossible to extract the free energy surface at $0.135T^*$ by reweighting[34,35] a simulation run at $T = 0.225T^*$. These two temperatures are on either side of a pseudo-phase transition, so there is almost no overlap between the configurations sampled during the two trajectories. A confirmation of this fact is provided by the top-central and bottom-central panels of Figure 6. At the lowest temperature, the system is almost exclusively confined to fcc-like structures that have negligible occupancies at the higher temperature. There are thus no liquid-like configurations among the landmarks used to construct the $T = 0.135T^*$ sketch-map and no FCC-like configurations among the landmarks used to construct the $T = 0.225T^*$ map. Nevertheless, the top-right and bottom-left panels of Figure 6 demonstrate that sketch-map's out-of-sample procedure works in spite of these deficiencies.

4.3. Testing the Sketch-Map Coordinates. Figure 6 proves that the out of sample procedure projects a given configuration in roughly the same location when there are similar configurations among the landmarks and when there are

dx.doi.org/10.1021/ct3010563 | J. Chem. Theory Comput. 2013, 9, 1521−1532

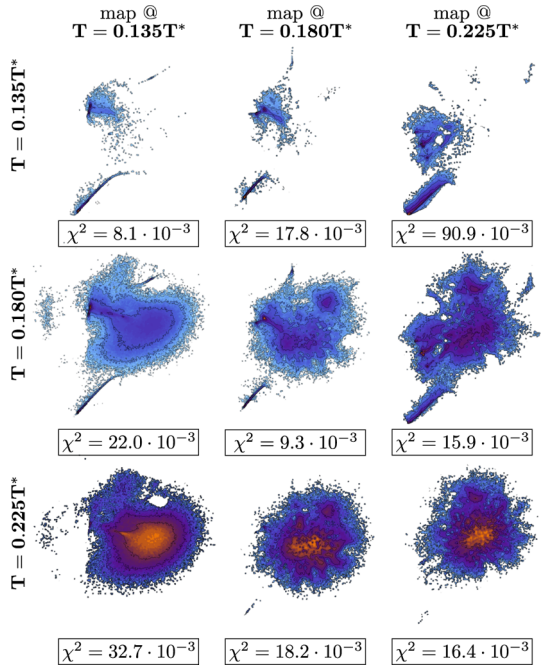


Figure 6. The free energy surface for LJ38 as a function of sketch-map coordinates at a range of temperatures. At the lowest temperature, the system spends most of its time trapped in the low-energy, solid-like minima in the landscape. As the temperature is increased and entropy starts to play a greater role, the system begins to spend a greater fraction of its time in the liquid-like basin. Meanwhile at the highest temperature, the entropic contribution to the free energy of the liquid is such that the solid-like basins no longer have substantial occupancies. To show how the free energy depends on temperature, we constructed sketch-map projections from three trajectories at different temperatures. We then constructed nine free energy surfaces—one for each of the three trajectories on each of the three sets of sketch-map coordinates.

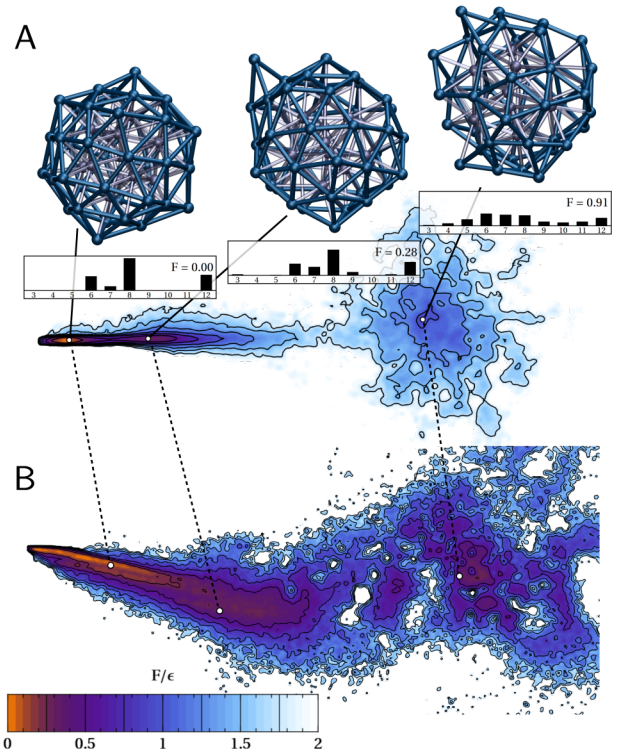


Figure 7. The free energy surface for LJ55 projected on a set of sketch-map coordinates generated from the LJ55 data (top) and a set of sketch-map coordinates generated from LJ38 data (bottom). The top free energy surface is more visually appealing, but the bottom one is still able to clearly separate the three main features in the energy landscape. Representative structures from these three basins are shown above the figure together with bar charts showing the coordination number histograms.

not. It is interesting to know how far we can push the bounds of this procedure. For instance, is it possible to project a second, completely different Lennard-Jones cluster using the sketch-map coordinates generated for LJ38? The histograms of coordination numbers that we use as high-dimensional descriptors for the clusters are normalized, so it is possible, in principle, to project an LJ55 trajectory on the LJ38 map. We thus took the data from the $T = 0.294T^*$ replica of our LJ55 parallel tempering calculation and projected it using the LJ38 map at $T = 0.180T^*$. At $T = 0.294T^*$, LJ55 is close to melting, so the trajectory contains both solid-like and liquid-like configurations, which makes it a challenging test. The results of this exercise are shown in the two free energy surfaces shown in Figure 7. These surfaces were constructed using sketch-map coordinates constructed from the LJ55 data and from the LJ38 data. The free energy surface constructed using the LJ55 data (top panel) is certainly more visually pleasing than the one constructed using the LJ38 data. The landscapes' features are seen in much clearer focus. However, in both representations the same three main features are identifiable: a basin corresponding to the minimum energy, icosahedral configuration, a basin corresponding to a defective version of the icosahedron that is implicated in surface diffusion,[36,37] and the liquid state. This suggests that sets of structures obtained from a cursory survey of the lay of the land can be used to construct a set of sketch-map coordinates that will project new structures in a reasonable location. These preliminary sketch-map coordinates can then be used, together with field-overlap metadynamics,[2] to accelerate the exploration of the free energy landscape and to extract relative free energies.

To test the sketch-map out-of-sample embedding procedure in an extreme scenario, we calculated the histogram of coordination numbers for a number of bulk LJ phases and projected them in two dimensions using the sketch-map coordinates generated from the melting-temperature trajectory for LJ38. As shown in Figure 8, we examined liquid, bcc, hcp, and fcc configurations for Lennard-Jones as well as the $I\bar{4}3d$ bulk phase discovered by Eshet et al.[38] All of these configurations are projected far from the LJ38 landmarks, which is good because these solid configurations clearly do not resemble any of the configurations of LJ38. Furthermore, with the exception of HCP and FCC, they are also projected far from each other. This second fact is remarkable, as it suggests that the sketch-map coordinates can discriminate between high-dimensional configurations that lie in a space that is disconnected from the space spanned by the LJ38 landmarks. [The dot product between the distribution of coordination numbers for the bcc and $I\bar{4}3d$ structures and all of the LJ38 landmarks is identically zero.] In fact, the only two configurations that the sketch-map coordinates constructed from the LJ38 data cannot distinguish are the hcp and fcc bulk structures. If you examine the histograms shown in Figure 8, however, it becomes clear that this is a failure of the high-dimensional description we have used. The histogram of coordination numbers is identical for the hcp and fcc bulk

dx.doi.org/10.1021/ct3010563 | J. Chem. Theory Comput. 2013, 9, 1521−1532

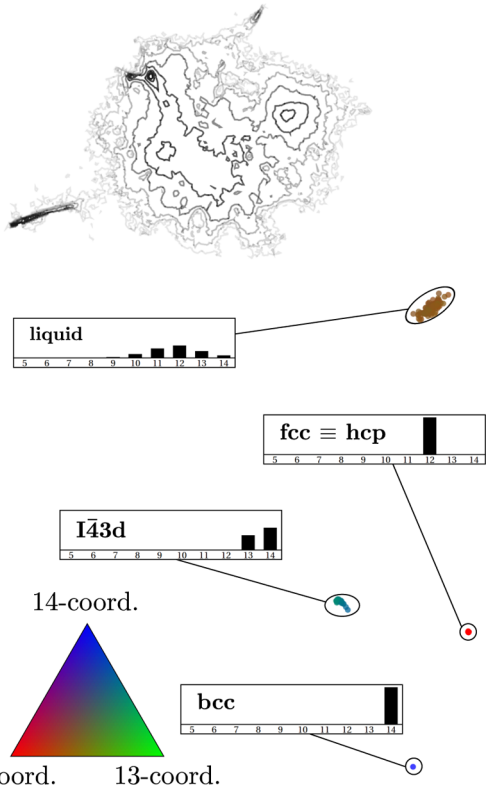


Figure 8. The projections obtained when a number of phases of bulk Lennard-Jones are projected using sketch-map coordinates generated from data on LJ38. The free energy surface for LJ38 is shown in the upper left-hand corner of the figure. All of the bulk structures are projected far from the structures of LJ38 and, with the exception of the hcp and fcc, are projected far from each other. Colors for the points are generated by taking the fractions of 12, 13, and 14 coordinated atoms and renormalizing so that these three components of the histogram sum to one. These renormalized values are then used to specify the degree to which red, blue, and green contribute to the final color, as illustrated in the key. The bar charts show the coordination number histogram for a representative structure of the specified bulk phase.

phases. This is well-known—to distinguish between the hcp and fcc structures, you have to examine the third coordination sphere as the first and second coordination spheres are identical.[39] As such, sketch-map coordinates constructed from the histogram of coordination numbers were never going to be able to distinguish fcc from hcp. This observation is important as it demonstrates that you still need to think about the physics when you use sketch-map. The high-dimensional coordinates must be able to discriminate between all possible structures. Sketch-map simply makes it easy to visualize differences between many degrees of freedom simultaneously and to thus incorporate many more degrees of freedom in the analysis. We could easily create sketch-map coordinates that can differentiate fcc and hcp by supplementing the histogram of coordination numbers with a few extra variables to describe the arrangement of atoms in the third coordination shell.

To be clear, we would not recommend using sketch-map coordinates constructed for LJ38 as universal coordinates for all phases of Lennard-Jones. To study LJ55, you should build sketch-map coordinates from an LJ55 trajectory. The exercises in this section are only there to demonstrate that sketch-map coordinates work even when the landmarks do not contain representatives of all the important structures.

5. CONCLUSION

Atomistic simulations produce data that is high-dimensional and thus impossible to comprehend without further analysis. Oftentimes, tools for performing this analysis are developed based on detailed knowledge of the chemistry/physics of the system. In many cases, however, one lacks such thorough understanding of the problem. It is then difficult to (a) know whether we have sampled all the low-energy parts of phase space and to (b) quantify the effect of small changes in the chemical environment.

It is increasingly recognized that the solutions to these problems lie in the MD trajectories themselves. Low-dimensional descriptions can be extracted by postprocessing trajectories, and these descriptions can then be used to accelerate the rate at which phase space is sampled or to understand how the energetics change when the chemical environment is perturbed. In this paper, we have extensively tested one such algorithm, sketch-map, by looking at a number of Lennard-Jones systems. This algorithm was designed to tackle issues specific to the problem of performing nonlinear dimensionality reduction on data coming from atomistic simulations, namely poor sampling at the transition states and the high-dimensional nature of thermal fluctuations. Our results show that sketch-map generates a useful map of phase space that can be used to construct revealing free energy surfaces. The sketch-map coordinates for LJ38 give a far richer picture than the average Steinhardt parameter[40]—a quantity that is often used for this system and which cannot distinguish the liquid state of LJ38 from the second lowest energy minimum in the potential energy surface.[25] More importantly, however, we have shown that the coordinates generated are extremely robust. The sketch-map coordinates constructed for LJ38 can be used to describe the free energy landscape of a second, completely different cluster or even to classify different bulk phases. This suggests that sketch-map will generate useful coordinates even if, because of the vagaries of sampling, the mapped trajectory does not visit all the energetically accessible parts of configuration space. Alternatively, if a chemical perturbation stabilizes configurations that were energetically inaccessible, sketch-map coordinates constructed for the unperturbed system will be able to recognize these new features in the energy landscape.

APPENDIX

A. Selecting an Appropriate σ Parameter

The most important parameter in sketch-map is σ. The discussions in the main text suggest that when we set σ, we are essentially deciding what features will be displayed in the projection. In all probability, it will not be possible to see the structure on length scales less than σ as sketch-map will make little to no effort to accurately reproduce these distances. The value of σ should be thus chosen by examining the data and deciding what short-range features can be safely ignored. For trajectory data, this is often the internal structure of the energetic basins—i.e., the thermal fluctuations.

Figure 9a shows what typical trajectory data looks like when it is projected on a pair of the high-dimensionality coordinates. There is clear clustering in the data, which, it seems reasonable to suppose, is because for much of the simulation time the

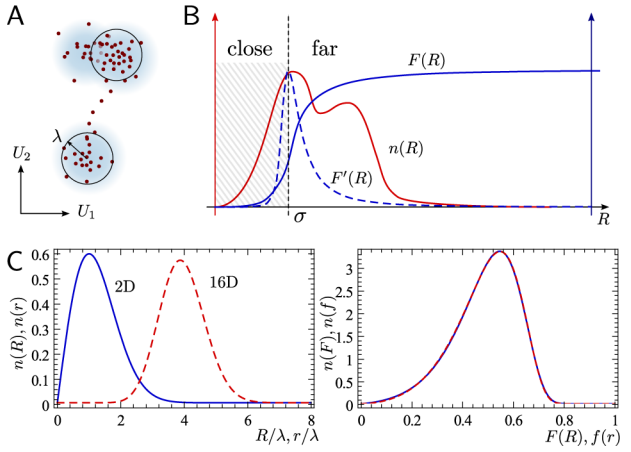


Figure 9. The figures describe how to select the parameters for sketch-map. Panel A is a cartoon showing how data points are distributed along two high-dimensional coordinates. There are clear clusters in the data, which we assume correspond to the basins in the energy landscape. As discussed in the text, we use the sizes of these clusters to choose σ. Panel B shows a cartoon of the histogram of pairwise distances between trajectory frames together with the sigmoid function we would use to sketch-map this set of points. The dashed line corresponds to the derivative of the sigmoid function. We tune σ so that the inflection point in the sigmoid function appears just before the first prominent peak in the histogram. Finally, panel C shows how setting $A \neq a$ allows one to resolve the problems that arise when you try to project the high-dimensionality thermal fluctuations into a low-dimensional space. On the left, the histogram of distances in a 16 dimensional Gaussian is shown together with the histogram of distances in a two-dimensional Gaussian. On the right, the distances from the 16-dimensional Gaussian and the two-dimensional Gaussian are transformed by two different sigmoid functions. The figure shows clearly that, although the original distributions do not resemble each other, the distributions of *differently transformed* distances match almost perfectly.

system is fluctuating about one of the minima in the energy landscape. This is precisely the sort of information we are not particularly interested in visualizing in our projections and that we don't want to focus on when we construct sketch-map projections. As such, the spatial extents of these clusters should inform any decision we make as to the value of σ. We should not be fooled, however, by looking at two-dimensional projections. The many directions orthogonal to those displayed will also contribute making distances far longer in N-dimensions than they appear to be in two. If we assume, however, that the fluctuations are isotropic, we can relate the radius we observe in two dimensions, λ, to the average distance between two D-dimensional points from the same basin using $\lambda(D-1)^{1/2}$. This sort of analysis of the two dimensional projections together with this equation is a good rule of thumb for setting σ.

When points are distributed in a high-dimensional space and you measure the full set of pairwise distances between them, you often find that the fraction of large distances is considerable. This is often just a consequence of the high dimensionality. In fact, we showed in our previous paper[1] that this part of the histogram of distances often resembles that obtained for a uniform distribution of points in the high-dimensionality space. It therefore seems reasonable to suppose that these long distances are not going to give us a great deal of information about any low-dimensionality features in the data. As such, a good way to check any putative value of σ is to examine the weight the algorithm would place on the reproduction of these distances. In practice, a good value for σ places the inflection point of the filter in the vicinity of the first prominent feature in the histogram of pairwise distances and thus gives little weight to the long distances.

B. Selecting the Other Sketch-Map Parameters

Changing the value of σ makes an enormous difference to the projection generated by sketch-map. Changing the other parameters has a much less drastic effect on the projection. These parameters are required as we need the filter functions to be smooth. If the filter functions are too sharp, the optimization is very poorly behaved.

The A and B parameters control the small-R and large-R tails of the filter, respectively. The histogram of distances between high-dimensional points (see Figure 9) can be used as a guide to set them. Generally, we want the filter function to go to zero quickly when $R < \sigma$, as we want a map in which points from the same energetic basin are projected almost on top of each other. A should thus be set to a large number—D is often a good first guess. We generally set $B \ll A$, as we want the tail of the function for large R to be considerably longer than the small R tail. As shown in Figure 9, we ideally would like this tail to not go to one before the histogram of distances goes to zero.

Our original intention when we developed sketch-map was, as we have described above, to make the constraints in the stress function on the shortest and longest distances less stringent. Clearly, an algorithm that works by minimizing an eq-2-like stress function with $f \equiv F$ fulfills this requirement. There are, however, good reasons for using different parameters in the two functions. $[F(x) - f(x)]$ equals zero for all x only if $f \equiv F$. Hence, if the two transfer functions are different, there is not necessarily a minimum in $[F(R) - f(r)]$ at $R = r$. This means that a sketch-map calculation run with two different filter functions will definitely not generate a mapping in which the distances between the projections are the same as the distances between the high-dimensional points. If the σ parameters are set differently, all distances will be uniformly scaled. If the a parameters are set differently, then distances less than σ will be distorted because for $R/\sigma \ll 1$, $[F(R)-f(r)] = 0$ when $r/\sigma \approx (R/\sigma)^{A/a}$. Similarly, when the b parameters are set differently, distances greater than σ will be distorted because for $R/\sigma \gg 1$, $[F(R) - f(r)] = 0$ when $r/\sigma \approx (R/\sigma)^{B/b}$. There is no reason to scale the map so σ can be set equal in the two functions. Similarly, we see no reason to deliberately distort the long distances so b and B can be set equal.

The reason for setting $A \neq a$ is again connected to the high-dimensionality features that are present in the way points are distributed in the high-dimensional space. If the system is inside one of the metastable basins in the energy landscape, it will fluctuate in all directions about the minimum energy structure in that basin. If the basin is harmonic, these fluctuations will give rise to a feature in the high-dimensional distribution of points that resembles a multivariate Gaussian. If we suppose, for the sake of simplicity, that this distribution is isotropic and that the variance along each direction is equal to λ, then two points selected at random from the basin will be separated by a distance on the order of $\lambda(D-1)^{1/2}$, as shown in the bottom panel of Figure 9. In other words, because of the high-dimensionality, two points from the same energetic basin can be far apart even if the spatial extent of the basin in each direction is small. In fact, this high-dimensionality effect can make it so that the separation between points in the same basin

dx.doi.org/10.1021/ct3010563 | J. Chem. Theory Comput. 2013, 9, 1521−1532

can become comparable to the separation between points in neighboring basins.

Obviously, we would like points in the same energetic basin to be projected close together and points in different basins to be projected far apart. In addition, it would be ideal if the projections of many points from a single basin in the free energy landscape together resembled a low-dimensional Gaussian as basins would then have an unambiguous signature in the map. Figure 9C shows that this will not happen if we attempt to match all the distances. The distribution of distances between points from the same basin will resemble that of a high-dimensional Gaussian and will thus be very different from the distribution of distances between points in a low dimensional Gaussian. This is important because a close match between these two histograms is a necessary (albeit not sufficient) condition for having the projections of points from the same basin arranged so that they resemble a low-dimensional Gaussian.

Thankfully, sketch-map does not try to match the distances between the high-dimensionality points and their projections. It instead transforms both distances by a filter function and endeavors to match these *transformed* distances. As such, the constraints on the distribution of distances between points from the same basin described above are no longer a prerequisite for having the projections of points from the same basin arranged so that they resemble a low-dimensional Gaussian. To satisfy this condition in sketch-map, we instead require that the histogram of $F(R_{ij})$ values and the histogram of $f(r_{ij})$ values match closely. This condition can be easily satisfied by tuning the parameters of the two filter functions separately. In particular, if we set a and A so that $a/A = d/D$, we can obtain the close match between the distributions of *transformed* distances between points taken from high and low dimensional Gaussian distributions shown in the right panel of Figure 9C.

C. Selecting Landmarks

In this work, landmarks were selected at random from trajectories, generated by extensive parallel tempering sampling. This ensured that the landmarks were distributed in a manner consistent with the underlying Boltzmann distribution for the temperature of the chosen replica. In our previous paper, we were analyzing data from a reconnaissance metadynamics simulation, so the relationship between the distribution of trajectory frames and the underlying free energy surface was unclear. We thus selected a set of landmarks, $\mathcal{X} = \{X_i\}$, using farthest point sampling (FPS), as this ensured that all the configurations visited during the trajectory were represented in the final map. In the FPS technique, you arbitrarily select a first point from the set of snapshots of the trajectory, $\mathcal{Z}$, that were collected during the simulation. Further points are then selected using the following criterion:

$$X_{j+1} = X: \min_{i \le j} |X_i - X| = \max_{Z \in \mathcal{Z}} \min_{i \le j} |X_i - Z| \tag{11}$$

where $|X_i - X|$ is the distance between configurations X and X_i. Selecting landmarks in this way ensures that all the sampled areas of phase space are represented in the set of landmark points, as at each stage the point selected is the one in $\mathcal{Z}$ that is farthest from all the points already selected. This perhaps solves the problems described above, but it does give us an algorithm that is rather sensitive to outliers. In addition, when this technique is used to select points from a parallel tempering or high-temperature MD simulation, we are not exploiting any of the valuable information about the relative probability of different configurations that is present in the trajectory. We therefore chose to develop a new two-stage procedure based on FPS but which does not totally ignore the probabilities of the configurations. There is a single parameter in this procedure, which controls the extent to which landmarks are chosen because they are in densely sampled parts of configuration space. When this parameter is set equal to one, the algorithm is equivalent to randomly selecting configurations. Setting it to zero corresponds roughly to selecting landmarks using FPS.

To select n landmarks from a total of N configurations using our new algorithm, we first select $(nN)^{1/2}$ points, $\mathcal{Y} = \{Y_i\} \subset \mathcal{Z}$, by farthest point sampling. We then proceed to count the number of points in $\mathcal{Z}$ that belong to the Voronoi polyhedron, $\mathcal{V}_i$, of each of the points in $\mathcal{Y}$ using

$$Z_j \in \mathcal{V}_i \Leftrightarrow |Z_j - Y_i| < |Z_j - Y_k| \ \forall \ k \neq i \tag{12}$$

The points selected using FPS are distributed uniformly across the space, so it is reasonable to assume that the associated Voronoi polyhedra all have approximately the same volume.[41] As such, the number of points within each polyhedron provides a rough estimate of the probability density in the neighborhood of the central point:

$$P_i = \frac{|\mathcal{V}_i|}{\sum_j |\mathcal{V}_j|} \tag{13}$$

It is straightforward to pick one of these polyhedra, $\mathcal{V}_k$, in accordance with these probabilities by exploiting an algorithm that is widely used in kinetic Monte Carlo.[42] If we then randomly pick one of the members of $\mathcal{V}_k$ and add it to the set of landmark frames, we have an algorithm that is equivalent to selecting points randomly from $\mathcal{Z}$. We selected landmarks for all of the sketch-map calculations in this paper by repeating this procedure n times and ensuring that we didn't pick the same point twice.

The advantage of this two stage procedure over simply selecting points randomly is that we can modify the probabilities (the P_i's) in the second step. That is to say, we can set the probability of selecting a given polyhedron to $P_i' = P_i^{\gamma}$. As illustrated in Figure 10, setting $\gamma > 1$ increases the differences in the probabilities and thus encourages the algorithm to only select landmarks from the most densely sampled regions. Setting $\gamma < 1$ has the opposite effect—encouraging the algorithm to ignore the underlying probabilities and to pick a set of landmarks that are more uniformly distributed over the space. In fact, we can interpret the γ parameter as an inverse temperature scaling. Setting $\gamma > 1$ is akin to selecting points according to a Boltzmann distribution at a temperature lower than the one at which the simulation was run, while setting $\gamma < 1$ is similar to selecting points according to a Boltzmann distribution at a higher temperature. This interpretation comes by considering the limiting case in which the high-dimensional description of the system is simply the Cartesian coordinates of the atoms. In this case the probabilities, P_i, are proportional to the Boltzmann distribution at the sampling temperature T.

To test the efficacy of this procedure, a further six sets of sketch-map coordinates were generated. To construct these coordinates, we selected landmarks from the 0.135, 0.180, and 0.225 T^* replicas of our parallel tempering simulation using the new landmark selection procedure with $\gamma = 1/4$ and $\gamma = 4$. We then projected the data from the replicas at each of these three temperatures on the nine sets of sketch-map coordinates and

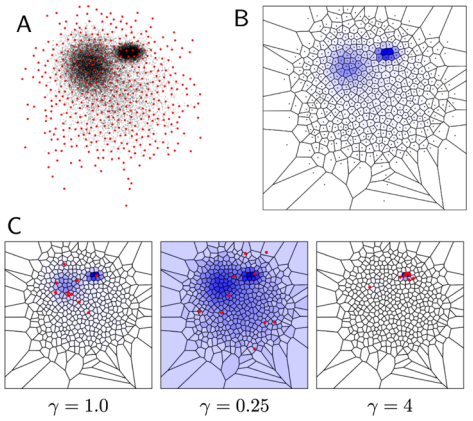


Figure 10. A schematic illustration showing how the two stage landmark selection algorithm works. The faint black points in panel A represent data points, while the red points are a set of landmarks selected by FPS. In panel B, the shapes of the Voronoi polyhedra for each of the red points in A are shown colored according to their Voronoi weight. Panel C shows the set of landmark configurations (red points) that would be selected from this data set for different values of the γ parameter. In these plots, the Voronoi polyhedra are colored in accordance with the probability of selecting a landmark from them.

evaluated the stress (eq 10). The results are shown in Table 1. The lowest-stress projections are still those constructed from landmarks selected at random from a trajectory at the projecting temperature. If we are trying to accumulate a free energy at a given temperature and we have a converged trajectory at the same temperature, it is best just to select landmark points at random. Setting $\gamma > 1$ or $\gamma < 1$ just disrupts the relation between the probability of selecting a point and the underlying free energy. Disrupting this relationship is only desirable when we are planning to use the sketch-map coordinates constructed by selecting landmarks from a trajectory at temperature T_1 to project a second, higher/lower-temperature trajectory. If we call the temperature in this second trajectory T_2 and we have $T_2 > T_1$, then, because at the higher temperature a greater volume of configuration space is energetically accessible, we need to select landmarks from sparsely sampled parts of configuration space where the potential energy is high. In contrast, if $T_2 < T_1$, we want to only select points from densely sampled regions of configuration space because, at the lower temperature, the system will be confined to regions where the potential energy is low. Table 1 shows that the results from the sketch-map projections are in accordance with this analysis. When sketch-map coordinates constructed from low-temperature trajectories are used to project higher-temperature data, the stress is lowered if $\gamma < 1$ and raised when $\gamma > 1$. By contrast, when we are using sketch-map coordinates constructed from high-temperature trajectories to project lower-temperature data, the stress is lowered when $\gamma > 1$ and raised when $\gamma < 1$.

Table 1. Comparison of Different Landmark Selection Protocols[a]

	map @ 0.135	map @ 0.180	map @ 0.225
random selection			
data @ T = 0.135	**8×10^{-3}**	18×10^{-3}	85×10^{-3}
data @ T = 0.180	22×10^{-3}	**9×10^{-3}**	16×10^{-3}
data @ T = 0.225	33×10^{-3}	18×10^{-3}	**16×10^{-3}**
two-stage selection, γ = 1/4			
data @ T = 0.135	*16×10^{-3}*	*34×10^{-3}*	80×10^{-3}
data @ T = 0.180	15×10^{-3}	*11×10^{-3}*	14×10^{-3}
data @ T = 0.225	27×10^{-3}	17×10^{-3}	*17×10^{-3}*
two-stage selection, γ = 4			
data @ T = 0.135	*20×10^{-3}*	9×10^{-3}	*97×10^{-3}*
data @ T = 0.180	*47×10^{-3}*	*18×10^{-3}*	15×10^{-3}
data @ T = 0.225	*76×10^{-3}*	*41×10^{-3}*	*18×10^{-3}*

[a]The table shows the residual stresses obtained when the data from simulations at three different temperatures is projected using sketch-map coordinates generated from landmarks selected using a variety of different protocols. The top panel in the table contains the stresses that were shown in Figure 6. In the lower two panels, we repeat this analysis using the new landmark selection protocol with the γ parameters described. The stresses obtained are underlined when they are lower than the corresponding stress for the map generated by selecting landmarks at random. The stresses obtained are in italics when the opposite is the case. As indicated in bold, the lowest-stress embedding is always the one constructed by randomly selecting landmarks from a trajectory at the same temperature. The new landmark procedure is only useful when you are trying to project high-temperature data using a map constructed from a low-temperature trajectory or vice versa.

Table 1 shows that adjusting γ has a significant effect when we are selecting landmarks from the 0.18 T^* replica and are using the resulting sketch-map coordinates to project the higher and lower temperature data. When we are selecting landmarks from the 0.135 and 0.225 T^* replicas, the effect changing γ has on the stress is much less marked. The reason for this is that, as discussed in the main text, the system undergoes a transition from a solid-like structure to a liquid-like structure between these two temperatures. Consequently, the higher-temperature trajectory samples a completely different part of configuration space to the lower temperature trajectory. As such, γ can no longer be interpreted as a temperature scaling.

AUTHOR INFORMATION

Corresponding Author

*E-mail: michele.ceriotti@chem.ox.ac.uk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Gábor Csányi, Ali Hassanali, Federico Giberti, and Davide Branduardi for useful discussions and also acknowledge funding from the European Union (Grant ERC-2009-AdG-247075) and the REA (Marie Curie IEF No. PIEFGA-2010-272402).

REFERENCES

(1) Ceriotti, M.; Tribello, G. A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023−13029.

(2) Tribello, G. A.; Ceriotti, M.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 5196−5201.

(3) Maragakis, P.; van der Vaart, A.; Karplus, M. *J. Phys. Chem. B* **2009**, *113*, 4664−4673.

(4) Tribello, G. A.; Ceriotti, M.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 17509−17514.

(5) Piana, S.; Laio, A. *J. Phys. Chem. B* **2007**, *111*, 4553−4559.

(6) Wales, D. J. *Energy Landscapes*; Cambridge University Press: Cambridge, U. K., 2003.

(7) Pártay, L. B.; Bartók, A. P.; Csányi, G. *J. Phys. Chem. B* **2010**, *114*, 10502−10512.
(8) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412.
(9) Garcia, A. E. *Phys. Rev. Lett.* **1992**, *68*, 2696−2699.
(10) Zhuravlev, P. I.; Materese, C. K.; Papoian, G. A. *J. Phys. Chem. B* **2009**, *113*, 8800−8812.
(11) Roweis, S. T.; Saul, L. K. *Science* **2000**, *290*, 2323−2326.
(12) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. *Science* **2000**, *290*, 2319−2323.
(13) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 9885−9890.
(14) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7432−7437.
(15) Coifman, R. R.; Lafon, S. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5−30.
(16) Belkin, M.; Niyogi, P. *Neural Comput.* **2003**, *15*, 1373−1396.
(17) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 13597−13602.
(18) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. *J. Chem. Phys.* **2011**, *134*, 124116.
(19) Borg, I.; Groenen, P. J. *Modern Multidimensional Scaling*, 2nd ed.; Springer: New York, 2005.
(20) Heiser, W. J. In *Classification and Related Methods*; North-Holland: Amsterdam, 1988; Chapter Multidimensional scaling with least absolute residuals, pp 455−462.
(21) Spiwok, V.; Kralova, B. *J. Chem. Phys.* **2011**, *135*, 224504.
(22) Branduardi, D.; Gervasio, F. L.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 054103.
(23) Doye, J. P. K.; Wales, D. J. *J. Chem. Phys.* **1995**, *102*, 9673−9688.
(24) Labastie, P.; Whetten, R. L. *Phys. Rev. Lett.* **1990**, *65*, 1567−1570.
(25) Doye, J. P. K.; Miller, M. A.; Wales, D. J. *J. Chem. Phys.* **1999**, *110*, 6896−6906.
(26) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(27) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. *Comput. Phys. Commun.* **2009**, *180*, 1961−1972.
(28) Bussi, G.; Donadio, D.; Parrinello, M. J. *J. Chem. Phys.* **2007**, *126*, 014101.
(29) Neirotti, J. P.; Calvo, F.; Freeman, D. L.; Doll, J. D. *J. Chem. Phys.* **2000**, *112*, 10340−10349.
(30) Calvo, F.; Neirotti, J. P.; Freeman, D. L.; Doll, J. D. *J. Chem. Phys.* **2000**, *112*, 10350−10357.
(31) Oganov, A. R.; Valle, M. *J. Chem. Phys.* **2009**, *130*, 104504.
(32) Tribello, G. A.; Cuny, J.; Eshet, H.; Parrinello, M. *J. Chem. Phys.* **2011**, *135*, 114109.
(33) Wales, D. J. *Mol. Phys.* **2002**, *100*, 3285−3306.
(34) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187−199.
(35) Bonomi, M.; Barducci, A.; Parrinello, M. *J. Comput. Chem.* **2009**, *30*, 1615−1621.
(36) Kunz, R. E.; Berry, R. S. *Phys. Rev. E* **1994**, *49*, 1895−1908.
(37) Kunz, R. E.; Berry, R. S. *Phys. Rev. Lett.* **1993**, *71*, 3987−3990.
(38) Eshet, H.; Bruneval, F.; Parrinello, M. *J. Chem. Phys.* **2008**, *129*, 026101.
(39) O'Keeffe, M.; Hyde, B. *Crystals Structures I. Patters and Symmetry*, 1st ed.; Mineralogical Society of America: Washington, DC, 1996.
(40) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. *Phys. Rev. Lett.* **1981**, *47*, 1297−1300.
(41) Shamos, M. I.; Hoey, D. Closest-point problems. *Proceedings of the 16th Annual Symposium on Foundations of Computer Science*, IEEE: Washington, DC, USA, 1975; pp 151−162.
(42) Voter, A. Introduction to the Kinetic Monte Carlo Method. In *Radiation Effects in Solids*; Sickafus, K., Kotomin, E., Uberuaga, B., Eds.; Springer: Netherlands, 2007; Vol. *235*, pp 1−23.