JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

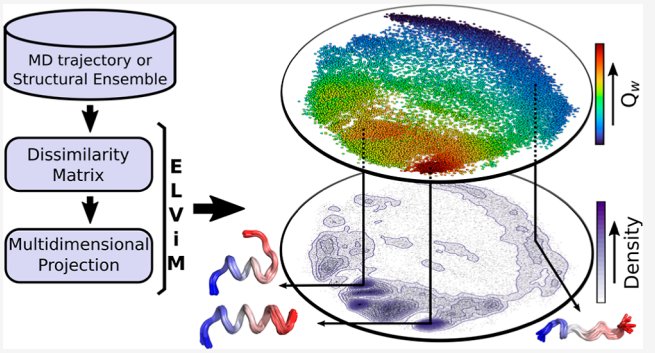
ELViM: Exploring Biomolecular Energy Landscapes through Multidimensional Visualization

Rafael Giordano Viegas, Ingrid B. S. Martins, Murilo Nogueira Sanches, Antonio B. Oliveira Junior, Juliana B. de Camargo, Fernando V. Paulovich, and Vitor B. P. Leite*

Cite This: J. Chem. Inf. Model. 2024, 64, 3443−3450 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Molecular dynamics (MD) simulations provide a powerful means of exploring the dynamic behavior of biomolecular systems at the atomic level. However, analyzing the vast data sets generated by MD simulations poses significant challenges. This article discusses the energy landscape visualization method (ELViM), a multidimensional reduction technique inspired by the energy landscape theory. ELViM transcends one-dimensional representations, offering a comprehensive analysis of the effective conformational phase space without the need for predefined reaction coordinates. We apply the ELViM to study the folding landscape of the antimicrobial peptide Polybia-MP1, showcasing its versatility in capturing complex biomolecular dynamics. Using dissimilarity matrices and a force-scheme approach, the ELViM provides intuitive visualizations, revealing structural correlations and local conformational signatures. The method is demonstrated to be adaptable, robust, and applicable to various biomolecular systems.



INTRODUCTION

In recent decades, molecular dynamics (MD) simulations have emerged as a potent tool for unraveling molecular systems' intricate kinetics and dynamics at the atomic level.[1] These simulations owe their robustness to significant strides in computational capabilities, modeling techniques, and sampling methods.[2−5] These advancements have extended simulation time scales and enabled the modeling of larger and more complex systems. However, analyzing the extensive data sets generated by MD simulations remains a formidable challenge. Researchers often rely on identifying meaningful reaction coordinates or collective variables to extract biophysically relevant information from molecular trajectories. Moreover, in recent years, machine learning techniques have also emerged as valuable tools for analyzing molecular trajectories.[6−8]

Analyzing molecular trajectories encounters several difficulties due to the large number of degrees of freedom, leading to a high-dimensional phase space. Each configuration in an MD trajectory can be conceptualized as a vector in a high-dimensional space. To illustrate, consider the description of the configurational space using only the alpha carbons of a protein with N residues; each conformation then exists as a vector within a multidimensional space encompassing $3N$ dimensions. Extracting meaningful features from this multidimensional space becomes arduous due to the so-called "curse of dimensionality". In such high-dimensional spaces, data points representing protein configurations are notably sparse, making fundamental analyses such as clustering exceptionally challenging.[6,9]

Various strategies have been employed to tackle these challenges, with dimensionality reduction (DR) techniques standing out. These methods aim to reduce the dimensionality of the configurational space, enabling a more manageable analysis and visualization. The difficulty relies in finding a good set of reaction coordinates. Another approach involves using DR techniques to map the multidimensional space onto a more manageable, lower-dimensional space. When the lower-dimensional space comprises only two or three dimensions, it facilitates the visualization of an MD trajectory using scatter maps, where each sampled conformation is represented by a data point. The underlying assumption is that the lower-dimensional representation can eliminate noise and redundant information while preserving only the relevant features of the multidimensional space.[6,8]

The first DR technique applied to MD trajectories was the principal component analysis (PCA),[10−12] a linear technique that works by diagonalizing a covariance matrix and projecting

Received: January 6, 2024
Revised: March 8, 2024
Accepted: March 8, 2024
Published: March 20, 2024

the data along the eigenvectors that retain the largest data variance, and it has been widely applied to several biological and chemical systems. Multidimensional scaling (MDS)[13−16] is a set of methods that aim to preserve pairwise distances or dissimilarities estimated in the high-dimensional space in the low-dimensional space. MDS methods differ in how distances are estimated and in how the optimization is carried out to represent these distances on the plane. However, both PCA and MDS are linear methodologies that rely on the hypothesis that most of the variance in the multidimensional data can be captured by a hyperplane.[6] Addressing limitations of these linear methods, recent advancements have introduced nonlinear methods such as kernel PCA,[17] Diffusion maps,[18] Isomap,[19] *t*-SNE,[20] UMAP,[21] Sketch-map,[22] and Encoder-Map.[23] While some methods seek a mapping between the high- and low-dimensional spaces, others are developed to visualize the data by optimizing the positions of points in the low-dimensional space to satisfy some cost function.

Another approach, inspired by statistical mechanics and spin glass systems principles, is the energy landscape theory (ELT),[24] which has been primarily used to describe the folding process. This theory depicts a funnel-shaped free-energy landscape biased toward the native state of a protein. In this representation, an ensemble of unfolded structures populates the high-energy portion of the landscape, which funnels toward the native structure. While this approach is not restricted to studying folding, it can be applied to a wide range of biomolecular systems and functions. Even though these processes are intrinsically multidimensional, it is often feasible to describe the kinetic and thermodynamic properties in terms of a few key quantities. However, one limitation of these techniques is that they require predefined reaction coordinates, potentially masking the richness and details of the problem.

This article elaborates on the energy landscape visualization method (ELViM), inspired by the ELT. Initially devised for visualizing the protein folding funnel,[25,26] ELViM represents a significant advancement by transcending the one-dimensional representation. ELViM is a multidimensional reduction method based on internal distances between pairs of structural conformations of the entire analyzed data set. Moreover, the method does not depend on reference structures or any other reaction coordinate. Through an iterative process, the method seeks to project an ensemble of conformations into two optimal dimensions, facilitating an intuitive visual analysis of the energy landscape. ELViM has been successfully applied in the study of various biomolecular systems, including an RNA tetraloop,[27] ordered proteins,[26,28−30] and intrinsically disordered peptides[31,32] and proteins.[33] We begin by discussing the general aspects of the method, subsequently delving into its intricacies, such as the dissimilarity metric, code details, iteration steps, and auxiliary tools, which are available at GitHub (https://github.com/VLeiteGroup/ELViM). We illustrate the method using the folding of the MP1 peptide and deliberate on ELViM's potential and caveats.

METHODS

Energy Landscape Visualization Method. Given an ensemble of conformations, the analysis using ELViM comprises two fundamental steps for generating lower-dimensional visualizations of the data set:

- Dissimilarity matrix calculation: In the first step, ELViM calculates a dissimilarity matrix. This matrix contains estimates of pairwise structural distances between the conformations in the data set.
- Multidimensional projection: In this process, each conformation from the molecular trajectory is represented as a data point in a two-dimensional space. This lower-dimensional representation is referred to as the "effective phase space" or simply the "ELViM projection".

In the upcoming sections, we provide detailed information regarding the main steps.

Dissimilarity Matrix. The initial implementation of ELViM was for lattice systems,[25] and a dissimilarity metric based on the ratio between the Jaccard index and the Jaccard distance was introduced. For out-of-lattice biomolecular systems, Oliveira et al.[26] proposed a novel dissimilarity metric based on the order parameter Q_w.[34,35] Considering two conformations, denoted as k and l, the similarity between them is defined as follows

$$q_w^{k,l} = \frac{1}{N_p} \sum_{i,j \in \text{pairs}} \exp\left[\frac{-(r_{i,j}^k - r_{i,j}^l)^2}{2\sigma_{i,j}^2}\right] \quad (1)$$

where $r_{i,j}^{\ k}$ and $r_{i,j}^{\ l}$ are the distances between the α-carbon of residues i and j from conformations k and l, respectively. N_p is a normalization constant equal to the number of α-carbon pairs, and $\sigma_{i,j}$ sets the similarity resolution of the metric and is defined by

$$\sigma_{i,j} = \sigma_0 |i - j|^{\epsilon} \quad (2)$$

typically, the values of σ_0 and ϵ are set to 1 Å and 0.15, respectively.[35] However, these parameters can be adjusted to fine-tune the dissimilarity metrics for different systems. The dissimilarity between the pair of conformations k and l is defined by

$$\delta_{k,l} = 1 - q_w^{k,l} \quad (3)$$

this dissimilarity measure is unitless and falls within the range $0 \leq \delta_{k,l} < 1$. A dissimilarity value equal to zero is only achieved when the two conformations are identical. Notably, this dissimilarity measure only relies on internal distances, not requiring structural alignment. While it is usual to consider only α-carbons for protein systems, this metric can be adapted to incorporate information from all atoms or other coarse-grained representations. After calculating dissimilarity values between every pair of conformations, these computed values are subsequently stored in a dissimilarity matrix used as input by the multidimensional projection technique. When there is a reference conformation (r), such as a native state, one can define a reaction coordinate with respect to this structure Q_w, such that $Q_w = q_w^{k,r}$. Q_w values can be used to color the data points and provide insights into the overall energy landscape.

Multidimensional Projection. The dissimilarity matrix provides information regarding the structural distances or dissimilarities between every pair of conformations. ELViM uses this dissimilarity matrix to perform a multidimensional projection procedure, which aims to project the data from the multidimensional conformational phase space onto a two-dimensional space while preserving as well as possible the relevant information. The outcome is a two-dimensional mapping, where each data point corresponds to a protein conformation, and the pairwise Euclidean distances in this reduced space are optimized to closely approximate the

https://doi.org/10.1021/acs.jcim.4c00034
J. Chem. Inf. Model. 2024, 64, 3443−3450

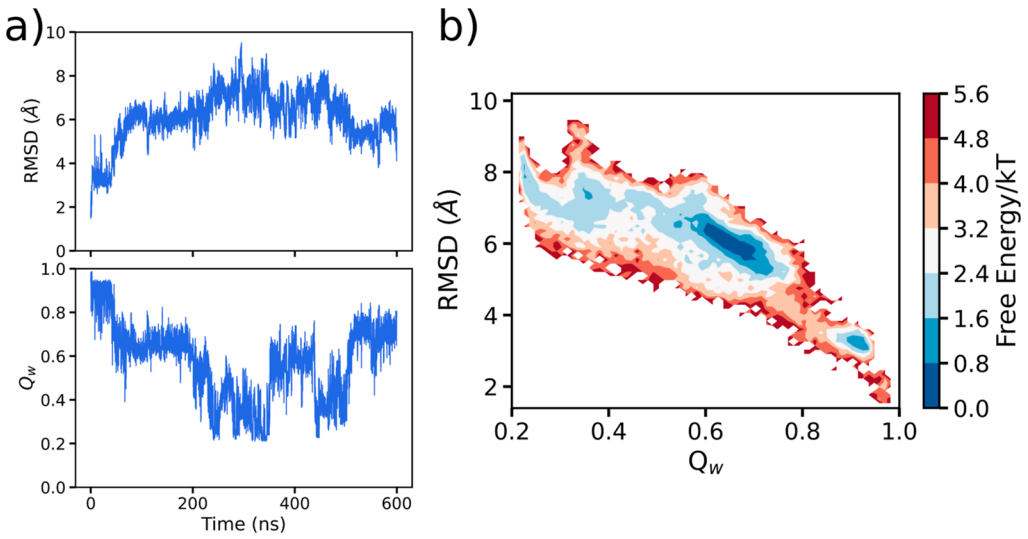


Figure 1. MD Trajectory. (a) Reaction coordinate, Q_w, and rmsd were both calculated using C_α coordinates and taking as reference the conformation that minimizes the energy of the α-helical structure. (b) 2D projection of the free-energy landscape as a function of the rmsd and Q_w.

dissimilarity values calculated within the multidimensional space. This mapping provides accessible and intuitive visualizations of structural correlations among the biological macromolecular conformations.

In ELViM, the multidimensional projection optimization is performed using the force scheme technique.[36] This algorithm, introduced by Tejada et al., is a standard method for multidimensional projection, offering a balance between precision and computational efficiency. It is referred to as a force approach or a force-directed method because it treats data points as masses interconnected by springs. This analogy captures how data points are "attracted to" or "repelled by" each other to optimize their pairwise distances to approximate the original dissimilarity. In this algorithm, each conformation k with coordinates $\vec{x}_k$ in the multidimensional phase space is projected to a data point x_k', represented in the Cartesian plane by the coordinates $\vec{x}_k'$. The projection begins with a random arrangement of data points to form the initial projection (X_0).

In each step, a conformation k acts as a reference, and the projected positions of all other conformations l ($l \neq k$) are slightly perturbed to adjust the distance in the plane $d_{k,l}(x_k', x_l')$ with the dissimilarity estimated in the multidimensional phase space $\delta_{k,l}(x_k, x_l)$. This adjustment is always carried out in the direction of the vector $\vec{v}_{k,l} = (\vec{x}_k' - \vec{x}_l')$ and is proportional to the difference between the dissimilarity and the Euclidean distance $(\delta_{k,l} - d_{k,l})$. Using a gradient-descent-like method, the cost function to be minimized is given by

$$E = \sum_{(k,l)} |\delta_{k,l} - d_{k,l}| \tag{4}$$

the degree of perturbation is controlled by a learning rate parameter, L_r, which may vary from an initial value to a predetermined minimum value $L_{r_{min}}$, both specified by the user. The learning rate for the i-th iteration is set to be

$$L_r[i] = \max\left\{ L_{r_0} * \left(1 + \frac{i}{\max_{it}}\right)^D, L_{r_{min}} \right\} \tag{5}$$

where L_{r0} is the initial learning rate value, D is the decay exponent, and $L_{r_{min}}$ is the minimum learning rate value. Note that combining the maximum function and the parameter $L_{r_{min}}$ allows the combination of an initial annealing of the learning rate, followed by constant learning rate steps. Typical value ranges are $1/8 < L_{r0} < 0.5$, $0.95 < D < 3$, and $0 < L_{r_{min}} < 1/8$.

The algorithm used in ELViM is detailed as follows:

Algorithm 1 ELViM-Force Scheme Algorithm

Initialize a projection X_0 with points randomly located.
while $i \leq max_{it}$ **do**
  for each randomly selected x_k' **do**
    for each $x_{l \neq k}'$ **do**
      $\delta_{k,l} \leftarrow \delta(x_k, x_l)$ ▷ ELViM dissimilarity
      $d_{k,l} \leftarrow \|\vec{x}_k' - \vec{x}_l'\|$ ▷ Euclidean distance
      $\vec{x}_l' \leftarrow \vec{x}_l' - L_r * (\delta_{k,l} - d_{k,l}) * (\vec{x}_k' - \vec{x}_l')/\|\vec{x}_k' - \vec{x}_l'\|$ ▷ vector from x_k' to x_l'
    end for
  end for
end while

ELViM Code Details. The ELViM main program is implemented in Python 3. It is based on the force scheme routine authored by F.V. Paulovich (https://github.com/fpaulovich/dimensionality-reduction). To handle molecular trajectories and extract the C_α coordinates for evaluating the dissimilarity matrix, the program utilizes MDtraj,[37] a Python library for molecular dynamics analysis. Additionally, the program benefits from parallel processing on CPU cores using Numba,[38] which optimizes Python code and executes it efficiently on CPU hardware.

ELViM accepts input molecular trajectory files in various formats recognized by MDTraj, including single trajectory files (e.g., PDB) or binary trajectories (e.g., XTC, DCD) along with a reference topology file (e.g., PDB). ELViM also offers the option to save the calculated dissimilarity matrix in a Python binary format for future use with different projection parameters. Alternatively, it allows the use of precomputed dissimilarity matrices. The program's output consists of a text file containing Cartesian coordinates representing the position of the biomolecular conformations within the effective phase space.

Simulation Details. To illustrate the utility of ELViM in analyzing the structural properties of a biomolecular system, we conducted the analysis with the antimicrobial peptide Polybia-MP1 (or MP1).[39] The MP1 peptide was constructed

https://doi.org/10.1021/acs.jcim.4c00034
J. Chem. Inf. Model. 2024, 64, 3443−3450

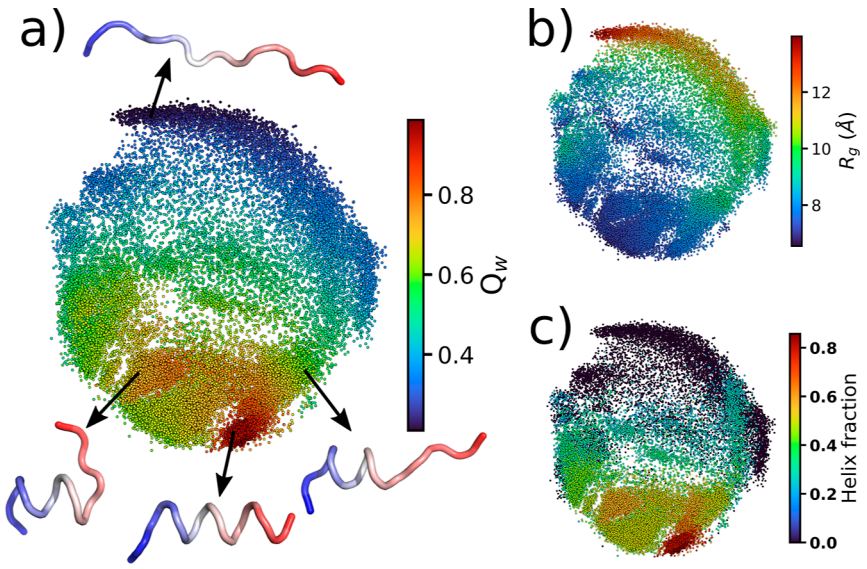


Figure 2. ELViM projection. Each conformation is represented by a dot. The *x*- and *y*-axes have been omitted since only pairwise distances provide meaningful information. A heatmap is employed to show the values of (a) reaction coordinate Q_w, (b) radius of gyration (R_g), and (c) helix content percentage for each conformation. Some conformations were selected from regions indicated by arrows and shown in carton with N-terminus in blue.

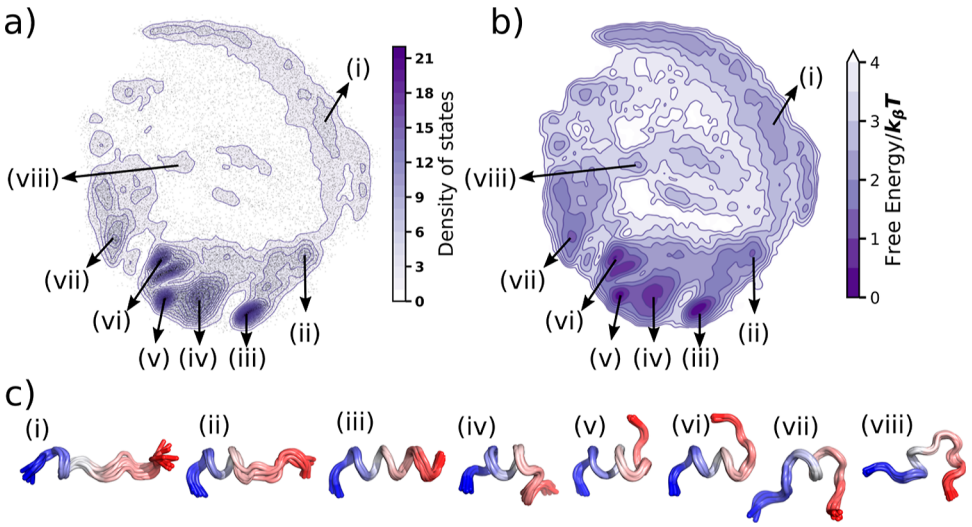


Figure 3. Density of states, free energy, and LCS. (a) The density of states is calculated using a Gaussian Kernel (KDE). The density is indicated by the colormap and contour plot superimposed to the ELViM projection dots in gray. (b) Free energy 2D profile estimated from the density of states is shown as a contour plot. (c) Data points from eight regions were manually selected, and 10 representative structures (LCS) of each region are depicted superimposed.

on an α helix configuration on VMD plugin Molefacture, solvated in a water box measuring 55 × 55 × 55 Å, containing NaCl ions in order to neutralize the system and also have 150 mM salt concentration. The system was equilibrated with 10,000 steps of conjugate gradient energy minimization and 10 ns of equilibrium MD with backbone restraint. The simulations were performed in the *NPT* ensemble, at 330 K, with a 2 fs time-step and periodic boundary conditions, using the software NAMD.[40] Temperature and pressure were modulated by Langevin thermostat[41] and Langevin piston,[42] respectively. The SHAKE algorithm[43] was used to constrain the lengths of covalent bonds, and the geometry of water molecules was preserved using the SETTLE algorithm.[44] The van der Waals interactions were calculated with a cutoff of 12 Å with a switching distance of 10 Å, and the long-range interaction was treated using the Particle Mesh Ewald method.[45]

RESULTS

In this example, we employed the ELViM to project an effective conformational phase space of the MP1 peptide sampled from an all-atom MD simulation. Along 600 ns of MD trajectory, we observed folding/unfolding events, when the peptide structure fluctuates between α-helix and random coil conformations. Figure 1 a displays the time evolution of the reaction coordinate, Q_w and the root-mean-square deviation (rmsd), both calculated having as a reference the conformation that minimizes the energy of the α-helical structure. In this case, Q_w equals 1 for an ideal α-helix and reached 0.2 for unfolded conformations.

https://doi.org/10.1021/acs.jcim.4c00034
J. Chem. Inf. Model. 2024, 64, 3443−3450

A common approach to comprehensively describe the sampled conformational phase space involves projecting the free-energy landscape onto a plane defined by order parameters or reaction coordinates of biological significance. In Figure 1 b, a free-energy estimation is presented through the projection of the rmsd and Q_w. Analysis of this surface reveals a predominant basin of partially folded conformations (with $Q \approx 0.65$), accompanied by two smaller basins: one representing the folded state, characterized by high values of Q_w and low rmsd values, and the other representing the unfolded state, characterized by low Q_w values and high rmsd values. Although this analysis is valuable, particularly in tasks such as estimating free energy barriers and discerning intermediate or metastable states, it does have limitations. Notably, this analysis does not offer detailed structural mapping, primarily due to the degeneracy associated with these order parameters, which would provide atomistic insights into molecular mechanisms.

The results of applying ELViM to analyze the MP1−MD trajectory are depicted in Figure 2. Dissimilarity was computed by using a σ_0 value of 1 Å. The learning rate varied from 0.3 to 0 with a decay of 0.95. In this representation of the landscape, each dot corresponds to a conformation. The axes have been omitted since only pairwise distances provide meaningful information. Data points that are projected near each other correspond to structurally similar conformations. Rotations and reflections about arbitrary axes can be performed without altering the global structure of the effective phase space.

In Figure 2a, the dots are color-coded based on their Q_w values, and we provide four representative structures. In this depiction, unfolded conformations are colored dark blue, while folded conformations are colored dark red. By coloring the dots according to various biophysically relevant variables, we can obtain different insights into the system. As additional examples, we have also colored the ELViM projection based on the radius of gyration R_g (Figure 2b) and the content of α-helix (Figure 2c).

To gain further insights into the folding landscape, it is essential to analyze how the density of data points varies throughout the effective conformational phase space. This measure estimates the density of states in the ELViM projection and allows for identifying basins formed by similar structures frequently visited during the simulation. Here, the probability density, ρ, is calculated using a Gaussian kernel density estimate (KDE) method, as implemented in scipy.[46] The results are depicted in Figure 3a, where we label some of the higher-density regions from (i) to (viii). Considering that we are analyzing a single-temperature classical MD simulation, the free energy two-dimensional (2D) profile of the projection can be calculated as

$$F = -k_\beta T \ln(\rho) \tag{6}$$

where the minimum value was set to zero. The resulting free energy surface (Figure 3b) shows that all local minima can be reached within the range of $2k_\beta T$.

Figure 3c presents representative structures from these selected regions, termed local conformational signatures (LCSs). To identify an LCS, we manually select an arbitrary region and calculate a matrix of distance−rmsd (drmsd)[47] values to find a reference conformation, which is defined as the conformation that minimizes the average drmsd. The LCS displays this conformation superimposed to its n closest neighbors according to drmsd values. It is noteworthy that this method is not a clustering analysis; rather, it is a visualization tool designed to provide a structural signature of conformations from arbitrarily selected regions. The representative structures from all high-density regions grant a broader perspective of the effective phase space. Furthermore, this approach allows us to compute contact maps and other biophysically relevant variables for conformations in an LCS. In cases where the data originate from unbiased simulations, high-density regions likely indicate local free energy minima within the overall free-energy landscape. Consequently, this analysis can provide essential insights, revealing atomistic details of these basins.

DISCUSSION

The ELViM algorithm starts with a random initialization of the projection and then optimizes distances to minimize the cost function (eq 4). To prevent dependencies on iteration order, points are randomly selected during optimization. Consequently, different runs produce different projections for the same set of parameters. If the complexity of the system is not too high, the projection stabilizes, and the different run outcomes are slightly different from each other, but the global structure of the effective phase space is maintained. As an example, we have generated three independent projections for the effective space of MP1 and compared them in Figure S1 of the Supporting Information. As previously discussed, only pairwise distances are meaningful in the ELViM projection, and rotations and reflections about arbitrary axes can be performed without changing the global structure of the effective phase space. As shown in Figure S1, we also arbitrarily selected seven local groups and showed how their position and composition are maintained throughout different replicas of the projection.

Based on our experience, a number of iterations equal to the square root of the number of conformations used in the projection typically suffices to achieve convergence. However, complex systems may require a larger number of iterations. To ensure the convergence of the projection with consistent global features, it is advisable to run and compare multiple projections. Consistency in the relative positioning of main basins is also a good indicator of convergence.

In the ELViM implementation, certain parameters can be adjusted to fine-tune the projection results. For instance, users can specify the parameters in $\sigma_{i,j}$ (eq 2), which determine the dissimilarity resolution. While the ELViM projections are usually robust when using the typical $\sigma_{i,j}$ parameter values indicated in the Methods section, extreme values may distort the projection. A small value tends to excessively increase the dissimilarity between data pairs, making it challenging to represent all the data points onto the plane. Conversely, a value that is too large may result in lower dissimilarity for very different structures, causing them to be placed within neighboring regions of the projection. To illustrate this dependency, we provide in the Supporting Information two ELViM projections for the MP1 system using different σ_0 parameters (Figure S2).

Initially, our implementation considered only C_α carbons as dissimilarity metrics. However, the code can be easily adapted to include other representations, such as using distances between all heavy atoms, as demonstrated in our RNA tetraloop study.[27] When atom indices differ from residue indices, adjustments to the $\sigma_{i,j}$ definition may be necessary. Modifications to the cost function (eq 4) can also be explored, such as squaring the residuals' differences between the

https://doi.org/10.1021/acs.jcim.4c00034
J. Chem. Inf. Model. 2024, 64, 3443−3450

dissimilarities and projected distances in eq 4. However, these modifications should be carefully evaluated for each system studied.

The learning rate parameter (eq 5) comprises three adjustable constants: the initial value L_{r0}, the final value L_{rmin}, and an exponent controlling the decay rate D, as discussed in the Methods section. This parameter acts as a scaling factor, which multiplies the residual difference between dissimilarities and distances, setting the size of the perturbation applied to every point in each iteration step. In this way, these parameters influence both convergence and the final projection outcome. Using excessively small learning rate values may confine data points within incorrect neighborhoods and significantly increase the required number of iterations to achieve convergence. Conversely, an excessively high learning rate may lead to an unstable projection. Large values of L_{rmin} may result in isolated structures resembling islands if they are considerably dissimilar from their nearest neighbors. An example of such behavior is provided in Supporting Information, Figure S3.

For projections with sufficient statistics, the probability density of each region in the projection phase space is associated with its free energy, as depicted in Figure 3. ELViM can offer an intuitive representation of the free-energy landscape as sampled during a simulation. The density of data points in the projection can be estimated by using a two-dimensional histogram or KDE, which can then be used to calculate free energy differences. In cases in which sampled conformations originate from enhanced sampling simulations or different conditions, data points may no longer carry the same weight. In such instances, ELViM can still provide a useful representation of the configurational space for visualization, but calculation of free energies may necessitate a reweighting procedure. For detailed examples, refer to our previous works.[31,32]

Finally, the occurrence of projection errors or distortions in dimensionality reduction and multidimensional visualization techniques should be acknowledged. These distortions are caused by data points misplaced in the projection.[9,48] These errors may be due to inherent mathematical limitations of projecting high-dimensional data onto a two-dimensional space.[9] In this context, some quality metrics have been proposed to quantify the extent to which distances or neighborhood relationships are preserved in the projection.[49,50] However, it is important to note that these metrics primarily assess the effectiveness of the optimization procedure but do not directly measure the method's capability to represent the unknown topology of the multidimensional phase space.[8]

When the ELViM projection was analyzed, misplaced points would behave as structural noise in a narrow neighborhood. In this sense, we suggest that the projection may be interpreted based on the prevalent structure signatures within a local neighborhood. We provide the LCSs tool to address this purpose, which identifies prevalent conformations within an arbitrarily selected region. By analyzing various LCSs comprising all of the high-density regions of the projection, it is possible to picture how representative conformational states are distributed throughout the projection. As previously stated, this tool provides a qualitative analysis, aiming at an intuitive interpretation of the projection. However, maintaining awareness of misplaced data points remains important to ensure the quality of the projection.

One intrinsic limitation of ELViM is the size and complexity of the investigated system. From a mathematical standpoint, multidimensional projection can be considered an ill-posed problem. The projection iteration procedure may converge to qualitatively different degenerate solutions, which may be viewed as alternate descriptions of the landscape. The limits of applicability of ELViM in terms of the size and complexity of a studied system have yet to be tackled. The complexity can be attributed to different factors, such as topology, types of secondary and tertiary structures, and folding mechanisms, which can lead to challenging representations of energy landscapes.

CONCLUSIONS

In conclusion, ELViM emerges as a valuable tool for unraveling the intricacies of biomolecular energy landscapes. Its ability to navigate high-dimensional configurational spaces and provide intuitive visualizations makes it a versatile choice for analyzing MD trajectories and structural ensembles. By eliminating the reliance on predefined reaction coordinates, ELViM allows for a more comprehensive exploration of the conformational phase space. Applying ELViM to the folding landscape of Polybia-MP1 demonstrates its efficacy in capturing dynamic transitions and revealing structural nuances. With its adaptable parameters and robust optimization scheme, ELViM stands as a promising method for researchers seeking a deeper understanding of biomolecular dynamics and functional mechanisms.

ASSOCIATED CONTENT

Data Availability Statement

Details for the MD simulation and the ELViM protocol are provided in the Materials and Methods section. A MD trajectory file and the Python script used to generate the ELViM projection are available on GitHub (https://github.com/VLeiteGroup/ELViM). PyMOL[51] was used for structural visualization.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00034.

Analysis of the ELViM reproducibility and study of the effects of the parameters σ_0 and L_r on the final outcome (PDF)

AUTHOR INFORMATION

Corresponding Author

Vitor B. P. Leite − *Department of Physics, São Paulo State University (UNESP), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, São Paulo 15054-000, Brazil;* orcid.org/0000-0003-0008-9079; Email: vitor.leite@unesp.br

Authors

Rafael Giordano Viegas − *Federal Institute of Education, Science and Technology of São Paulo (IFSP), Catanduva, São Paulo 15.808-305, Brazil; Department of Physics, São Paulo State University (UNESP), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, São Paulo 15054-000, Brazil;* orcid.org/0000-0002-6102-3375

Ingrid B. S. Martins − *Department of Physics, São Paulo State University (UNESP), Institute of Biosciences, Humanities*

https://doi.org/10.1021/acs.jcim.4c00034
J. Chem. Inf. Model. 2024, 64, 3443−3450

and Exact Sciences, São José do Rio Preto, São Paulo 15054-000, Brazil; orcid.org/0000-0001-9970-6035

Murilo Nogueira Sanches − *Department of Physics, São Paulo State University (UNESP), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, São Paulo 15054-000, Brazil;* orcid.org/0000-0001-9650-7989

Antonio B. Oliveira Junior − *Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

Juliana B. de Camargo − *Department of Physics, São Paulo State University (UNESP), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, São Paulo 15054-000, Brazil*

Fernando V. Paulovich − *Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00034

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the financial support from Prope/Unesp and the Brazilian agencies FAPESP (grants 2023/02219-1, 2022/08738-8, 2023/08101-2, and 2021/15028-4) and National Council for Scientific and Technological Development—CNPq (Grant 310017/2020-3). A.B.O.J. acknowledges the Robert A. Welch Postdoctoral Fellow program (grant C-1792). This research was supported by resources supplied by the National Laboratory for Scientific Computing (LNCC/MCTI, Brazil), the SDumont supercomputer, URL: http://sdumont.lncc.br, the Center for Scientific Computing (NCC/GridUNESP) of the São Paulo State University (UNESP) and the "Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP).

■ REFERENCES

(1) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429−452.

(2) Sidky, H.; Chen, W.; Ferguson, A. L. Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Mol. Phys.* **2020**, *118*, No. e1737742.

(3) Lazim, R.; Suh, D.; Choi, S. Advances in Molecular Dynamics Simulations and Enhanced Sampling Methods for the Study of Protein Systems. *Int. J. Mol. Sci.* **2020**, *21*, 6339.

(4) Schlick, T.; Portillo-Ledesma, S. Biomolecular modeling thrives in the age of technology. *Nat. Comput. Sci.* **2021**, *1*, 321−331.

(5) Sinha, S.; Tam, B.; Wang, S. M. Applications of Molecular Dynamics Simulation in Protein Study. *Membranes* **2022**, *12*, 844.

(6) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121*, 9722−9758.

(7) Noé, F.; De Fabritiis, G.; Clementi, C. Machine learning for protein folding and dynamics. *Curr. Opin. Struct. Biol.* **2020**, *60*, 77−84.

(8) Tribello, G. A.; Gasparotto, P. Using Dimensionality Reduction to Analyze Protein Trajectories. *Front. Mol. Biosci.* **2019**, *6*, 46.

(9) Nonato, L. G.; Aupetit, M. Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment. *IEEE Trans. Visualization Comput. Graphics* **2019**, *25*, 2650−2673.

(10) García, A. E. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696−2699.

(11) Jolliffe, I. *Principal Component Analysis*. In *Springer Series in Statistics*; Springer, 2002; .

(12) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Struct., Funct., Bioinf.* **1993**, *17*, 412−425.

(13) Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika* **1952**, *17*, 401−419.

(14) France, S. L.; Carroll, J. D. Two-Way Multidimensional Scaling: A Review. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **2011**, *41*, 644−661.

(15) Mead, A. Review of the Development of Multidimensional Scaling Methods. *J. R. Stat. Soc. - Ser. D Statistician* **1992**, *41*, 27.

(16) Pisani, P.; Caporuscio, F.; Carlino, L.; Rastelli, G. Molecular Dynamics Simulations and Classical Multidimensional Scaling Unveil New Metastable States in the Conformational Landscape of CDK2. *PLoS One* **2016**, *11*, No. e0154066.

(17) Schölkopf, B.; Smola, A.; Müller, K.-R. *Kernel Principal Component Analysis*; Artificial Neural Networks — ICANN'97: Berlin, Heidelberg, 1997, pp 583−588.

(18) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7426−7431.

(19) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319−2323.

(20) van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(21) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3*, 861.

(22) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13023−13028.

(23) Lemke, T.; Peter, C. EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J. Chem. Theory Comput.* **2019**, *15*, 1209−1215.

(24) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.

(25) Oliveira, A. B.; Fatore, F. M.; Paulovich, F. V.; Oliveira, O. N.; Leite, V. B. P. Visualization of Protein Folding Funnels in Lattice Models. *PLoS One* **2014**, *9*, No. e100861.

(26) Oliveira, A. B.; Yang, H.; Whitford, P. C.; Leite, V. B. P. Distinguishing Biomolecular Pathways and Metastable States. *J. Chem. Theory Comput.* **2019**, *15*, 6482−6490.

(27) Viegas, R. G.; Sanches, M. N.; Chen, A. A.; Paulovich, F. V.; Garcia, A. E.; Leite, V. B. P. Characterizing the Folding Transition-State Ensembles in the Energy Landscape of an RNA Tetraloop. *J. Chem. Inf. Model.* **2023**, *63*, 5641−5649.

(28) Sanches, M. N.; Parra, R. G.; Viegas, R. G.; Oliveira, A. B.; Wolynes, P. G.; Ferreiro, D. U.; Leite, V. B. Resolving the fine structure in the energy landscapes of repeat proteins. *QRB Discov.* **2022**, *3*, No. e7.

(29) Dias, R. V. R.; Pedro, R. P.; Sanches, M. N.; Moreira, G. C.; Leite, V. B. P.; Caruso, I. P.; de Melo, F. A.; de Oliveira, L. C. Unveiling Metastable Ensembles of GRB2 and the Relevance of Interdomain Communication during Folding. *J. Chem. Inf. Model.* **2023**, *63*, 6344−6353.

(30) da Silva, F. B.; Martins de Oliveira, V.; de Oliveira Junior, A. B.; Contessoto, V. d. G.; Leite, V. B. P. Probing the Energy Landscape of Spectrin R15 and R16 and the Effects of Non-native Interactions. *J. Phys. Chem. B* **2023**, *127*, 1291−1300.

(31) Sanches, M. N.; Knapp, K.; Oliveira, A. B., Jr.; Wolynes, P. G.; Onuchic, J. N.; Leite, V. B. Examining the Ensembles of Amyloid-β

Monomer Variants and Their Propensities to Form Fibers Using an Energy Landscape Visualization Method. *J. Phys. Chem. B* **2022**, *126*, 93−99.
(32) Martins, I. B. S.; Viegas, R. G.; Sanches, M. N.; de Araujo, A. S.; Leite, V. B. P. Probing Mastoparan-like Antimicrobial Peptides Interaction with Model Membrane Through Energy Landscape Analysis. *J. Phys. Chem. B* **2024**, *128*, 163−171.
(33) Oliveira Junior, A. B.; Lin, X.; Kulkarni, P.; Onuchic, J. N.; Roy, S.; Leite, V. B. P. Exploring energy landscapes of intrinsically disordered proteins: Insights into functional mechanisms. *J. Chem. Theory Comput.* **2021**, *17*, 3178−3187.
(34) Hardin, C.; Eastwood, M. P.; Prentiss, M. C.; Luthey-Schulten, Z.; Wolynes, P. G. Associative memory Hamiltonians for structure prediction without homology: α/β proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 1679−1684.
(35) Cho, S. S.; Levy, Y.; Wolynes, P. G. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 586−591.
(36) Tejada, E.; Minghim, R.; Nonato, L. G. On Improved Projection Techniques to Support Visual Exploration of Multi-Dimensional Data Sets. *Inf. Visual.* **2003**, *2*, 218−231.
(37) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.
(38) Lam, S. K.; Pitrou, A.; Seibert, S. *Numba: A LLVM-Based Python JIT Compiler*. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*: New York, NY, USA, 2015.
(39) Souza, B. M.; Mendes, M. A.; Santos, L. D.; Marques, M. R.; Cesar, L. M. M.; Almeida, R. N. A.; Pagnocca, F. C.; Konno, K.; Palma, M. S. Structural and functional characterization of two novel peptide toxins isolated from the venom of the social wasp Polybia paulista. *Peptides* **2005**, *26*, 2157−2164.
(40) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(41) Davidchack, R. L.; Handel, R.; Tretyakov, M. V. Langevin thermostat for rigid body dynamics. *J. Chem. Phys.* **2009**, *130*, 234101.
(42) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613−4621.
(43) Elber, R.; Ruymgaart, A. P.; Hess, B. SHAKE parallelization. *Eur. Phys. J. Spec. Top.* **2011**, *200*, 211−223.
(44) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−962.
(45) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(46) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261−272.
(47) Wallin, S.; Farwer, J.; Bastolla, U. Testing similarity measures with continuous and discrete protein models. *Proteins: Struct., Funct., Bioinf.* **2003**, *50*, 144−157.
(48) Ortigossa, E. S.; Dias, F. F.; Nascimento, D. C. d. Getting over High-Dimensionality: How Multidimensional Projection Methods Can Assist Data Science. *Appl. Sci.* **2022**, *12*, 6799.
(49) Pagliosa, P.; Paulovich, F. V.; Minghim, R.; Levkowitz, H.; Nonato, L. G. Projection inspector: Assessment and synthesis of multidimensional projections. *Neurocomputing* **2015**, *150*, 599−610.
(50) Espadoto, M.; Martins, R. M.; Kerren, A.; Hirata, N. S. T.; Telea, A. C. Toward a Quantitative Survey of Dimension Reduction Techniques. *IEEE Trans. Visualization Comput. Graphics* **2021**, *27*, 2153−2173.
(51) Schrödinger, LLC., *The PyMOL Molecular Graphics System*. 2010; https://github.com/schrodinger/pymol-open-source.

https://doi.org/10.1021/acs.jcim.4c00034
J. Chem. Inf. Model. 2024, 64, 3443−3450