JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Describing the Conformational Landscape of Small Organic Molecules through Gaussian Mixtures in Dihedral Space

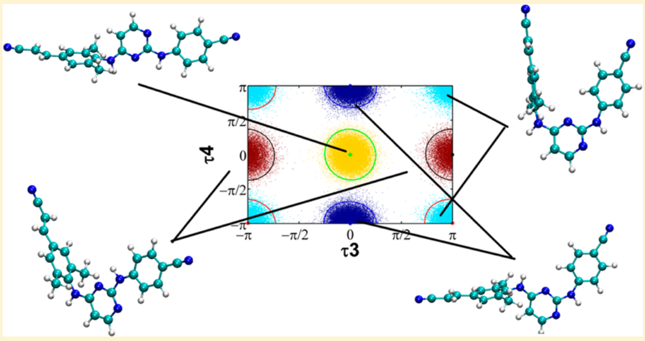
Pasquale Pisani,[†] Paolo Piro,[‡] Sergio Decherchi,[†] Giovanni Bottegoni,*,[†] Diego Sona,[‡] Vittorio Murino,[‡] Walter Rocchia,*,[†] and Andrea Cavalli[†,§]

[†]Department of Drug Discovery and Development, Istituto Italiano di Tecnologia, via Morego, 30, 16163 Genova, Italy

[‡]Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, via Morego, 30, 16163 Genova, Italy

[§]Dept. of Pharmacy and Biotechnology, University of Bologna, Via Belmeloro 6, 40126 Bologna, Italy

S *Supporting Information*



ABSTRACT: Due to the well-known structure−function paradigm, conformational equilibrium plays a major role in molecular recognition. Therefore, a deep understanding of the conformational profile of small organic molecules is an essential prerequisite to modern computer-assisted drug design. However, a thorough analysis and a meaningful representation of the conformational landscape of drug-like molecules remains a challenge. The thermodynamic equilibrium of conformational states can be described in terms of probability density function (PDF) defined in the space of the relevant degrees of freedom of the system. In principle, this PDF could be estimated by traditional histogram methods, which are, however, hampered by several limitations when the variables forming the space are more than two or three. Here, we present an unsupervised parametric fitting procedure based on cluster analysis, aimed at estimating the PDF in the conformational space of small drug-like molecules with low sensitivity to data dimensionality. Indeed, data are represented in the dihedral space of the molecule and clustered using a simple adaptation of the standard k-means algorithm for periodic data. In the final step of the analysis, the PDF is derived as a linear combination of multivariate circular Gaussian distributions. We show that exploiting the analytic properties of Gaussian distributions, the proposed approach makes it possible to analyze the conformational ensemble in higher dimensional spaces with several advantages over the histogram-based methods. The posterior analysis of the PDF also helps identify a minimal subset of variables able to provide a meaningful representation of the conformational space. We tested our approach on alanine dipeptide, alanine tetrapeptide, and rilpivirine with satisfactory results compared to standard histogram-based methods and to those based on chemical intuition.

INTRODUCTION

Drugs exert their action binding to a macromolecular counterpart. The recognition process is deeply affected by the ability of each molecular partner to adopt a conformation compatible with the binding event. For this reason, an accurate representation of the conformational ensemble of a drug in solution could be extremely valuable to design better drugs.[1,2] In solvent, organic molecules exist as ensembles of rapidly interconverting conformations. The macroscopic properties of these systems are the result of a dynamic equilibrium induced by their energetic landscapes: drugs can explore multiple conformational basins that vary in size and depth and that are separated by energetic barriers of different heights.[3] While the number of rotatable bonds clearly affects the complexity of molecular conformational populations, a general relationship between the flexibility of a molecule and its binding affinity with respect to a given target could not be clearly established.[4,5] The understanding of molecular recognition process requires investigations at the molecular level and accurate models. According to the conformational selection paradigm, the two partners of a binding process must pre-exist in free energy conformational minima favorable to the binding event.[1] However, the entire binding process is the results of a delicate balance between electrostatic force, nonpolar interactions, solvent contribution, and, put in a different perspective, between enthalpic and entropic effects.[3] All these factors can largely distort the free energy landscape of a single molecule in the solvent with respect to that of the same molecule in a complex. This reorganization is recognized as a key factor to understand molecular recognition, as well as drug potency, specificity, and resistance.[4,5] This is the main reason why the study of thermally accessible conformational substates of binding partners, even if they are different from the lowest

Received: October 29, 2013

energy conformation, can be a useful and computationally less demanding alternative approach to the simulation of the full binding event, which is often out of reach of the present computational methods.

Interestingly, physics-based computational studies can provide several direct insights into many aspects of these mechanisms. In the past decade, the development of better models and optimized algorithms, coupled with an ever increasing computing power, led to robust ways to effectively generate ensembles of conformations, closely approximating the ergodic behavior of small molecules in a realistic environment.[6] Within the framework of the statistical mechanics, the conformational equilibrium can be properly described by a probability density function (PDF) in an n-dimensional space, provided that an exhaustive sampling procedure has been carried out and that the coordinates defining the space represent relevant degrees of freedom of the system.[7] Each point in this space corresponds to a number of conformations sharing the same reduced coordinates, often referred to as collective variables (CVs). In this way, mapping the conformational populations of drug-like molecules in water, it is possible to estimate the free energy needed to reach the so-called bioactive conformation, that is, the conformation that a molecule assumes upon binding. In general, it represents a free energy penalty for reorganizing the distribution of conformations assumed in the unbound form to that/those compatible with the binding event.[8] This phenomenon has been described as "conformer focusing".[9] If we discretize the conformational space in k different conformational states and simplify the analysis by considering the case where only one conformer k is left for the bound ligand, than we can define this binding reorganization free energy of the k-th state as

$$\Delta F_k = -k_B T \log p_k$$

where k_B is the Boltzmann constant and p_k is the relative population of the k-th state.

Once a sufficient sampling has been achieved, there are still two main issues undermining a fruitful exploitation of the available computational machinery in order to design better drug molecules using conformational information: (i) how to get a handy expression for the PDF starting from the acquired sampling; (ii) how to infer from the PDF a subset of relevant degrees of freedom able to concisely describe the main conformational states of a molecule. As a matter of fact, data acquisition and representation become problematic when dealing with high dimensional spaces.

The simplest approach to PDF estimation is based on the use of histograms with predefined bin size. However, while being simple and rather intuitive, this method has some limitations. The discretization of continuous variables in finite size intervals leads to systematic errors. Moreover, when data is sparse, the analysis process can cause numerical instabilities.[10,11] Finally, since the complexity of the PDF representation grows exponentially with the number of degrees of freedom, the cost in terms of computer memory might become another important issue to address.

As per the last point, while it is always possible to project the conformational features of a system on a low-dimensional space, it is far from obvious how this can be achieved minimizing the loss of relevant information. This is equivalent to identify the set of the CVs that best account for the phenomenon of interest. In the context of molecular recognition, we require CVs to be able to identify and separate the most relevant conformational basins for a molecule in water. The rationale for this is based on the assumption that the energy of the bioactive conformation is in line with (and, in any case, never exceedingly higher than) the range of energies explored by the molecule in the solvent. In other words, the interaction with the target is mainly stabilizing an already existing low energy basin.[12] Our initial choice of putative CVs starts from internal coordinates, routinely used for separating the roto-traslational degrees of freedom from internal movements.[13−15] In particular, in this work, we focus on dihedral angles calculated only considering heavy atoms. Therefore, we ignore other internal coordinates that are much less influential on the definition of a conformational state: (i) bond lengths and planar angles, which do not deviate significantly from their average values; (ii) the rotation of symmetrical terminal groups (e.g., a methyl group); (iii) dihedral angles which involve hydrogen atoms; (iv) bonds between members of the same symmetrical aromatic ring. In the end, we focused on a subset of the internal coordinates composed by proper dihedral angles that represent rotatable bonds. We aimed at further characterizing this reduced set by pinpointing the dihedral angles that best represent the overall conformational information.

In the context of adaptive umbrella sampling, Maragakis and colleagues[16] described a method to represent the PDF over the conformational space using a parametric fitting procedure that scales efficiently with the dimensionality of the system. In their work, the PDF in the dihedral space was estimated using a Gaussian Mixture Model (GMM).[17] A mixture model is a probability model for representing a possibly convoluted random variable as a combination of a set of simpler random variables. GMM were also used by Tribello and co-workers to realize a multiple CV bias potential in Reconaissance Metadynamics.[18] In our work, GMM describes the distribution of all possible configurations of a conformational collective state C belonging to the dihedral space. This distribution, having an unknown number of modes, is approximated by a linear combination of Gaussian distributions. However, as pointed out in ref 16, some questions remain open: (i) how to choose the optimal number of components in order to avoid overfitting? (ii) how to exploit the parametric form of the PDF to pinpoint conformational basins and to avoid brute force search of PDF maxima (i.e., free energy minima)? (iii) how to use this representation to partition the conformational populations?

Here, we attempt to address these issues using an approach which relies on k-means clustering and on the expectation-maximization algorithm (EM) to estimate the parameters of the GMM.[19,20] These techniques are widely used in the machine learning community where issues such as those mentioned above have already been faced. K-means is a method for cluster analysis, which partitions N samples into k clusters so that each observation belongs to the cluster with the closest mean. EM is an iterative method for finding maximum likelihood estimates of parameters in statistical models. Although both problems are computationally challenging, there are efficient heuristics that converge quickly to a locally optimal solution.[21] Our overall approach works particularly well when the number of variables is relatively small, so that the number of samples counterbalances the dimensionality of the problem. Hence, it appears particularly suitable for drug-like compounds, which have on average 5 or 6 rotatable bonds, seldom exceeding 12.[22]

We tested our combined approach on several test bed systems. A first simple system was the alanine dipeptide (Figure 1), a widely used test bed for the calculation of nontrivial free

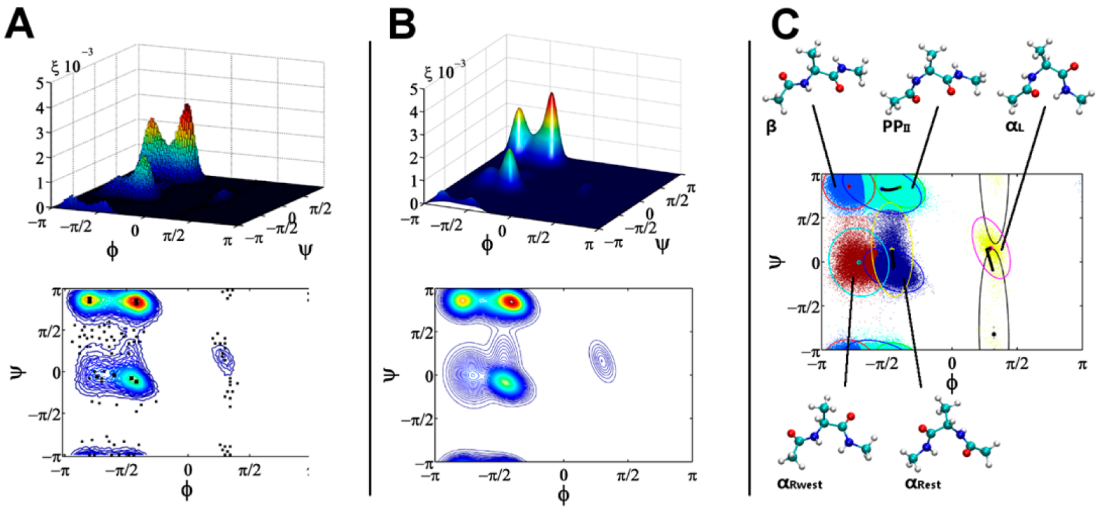


Figure 1. Conformational landscape of alanine dipeptide simulated in a TIP3P water box. (A) Histogram and contour plot. Black dots pinpoint maxima found with exhaustive search and a probability density greater than 10^{-5}. (B) GMM calculated using k-means++/EM and contour plot. (C) Conformational states recovered by soft clustering and representative structures.

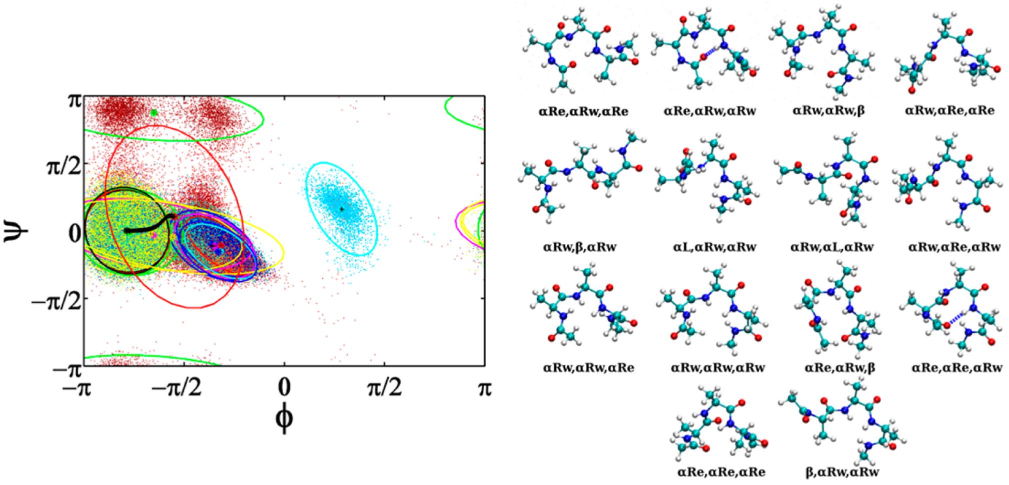


Figure 2. Analysis of the trajectory of the alanine tetrapeptide in a water box return 15 clusters. On the left, a subset of 50K points are plotted using φ and ψ dihedrals of the central residue. Representatives of each cluster are reported on the right.

energy surfaces and low energy pathways.[23−25] Here, we studied alanine dipeptide simulated in explicit solvent. More complex and realistic tests were performed on alanine tetrapeptide (Figure 2) and {(4-[[4-[[4-[(1E)-2-cyanoethenyl]-2,6-dimethylphenyl]amino]-2-pyrimidinyl]amino]-benzonitrile} also known as rilpivirine (Figure 3). The former is composed by three alanine residues terminally capped, which has been used as a prototype for the study of helix formation.[26−29] The latter is an inhibitor of the reverse transcriptase of HIV-1 virus (HIV1-RT) and it is a drug currently marketed for the treatment of HIV infections. This molecule was selected because experimental and in silico data were already available.[8,30] Moreover, because of its average number of rotatable bonds and because of the presence of symmetric groups, rilpivirine represents an ideal test case to illustrate the applicability of our method to drug-like compounds.

As mentioned above, we are also interested in the automatic selection of the most significant CVs in order to both reduce the dimensionality and make the representation intuitive and easy to handle. We devised a simple yet effective method based on the analysis of the covariance matrix of GMM.

The approach outlined here is, to the best of our knowledge, the first attempt to set up a fully automated protocol for the characterization of the conformational space of drug-like molecules as well as for the choice of an effective CV subset that helps identify the bioactive conformations.

■ METHODS

Molecular Dynamics Simulations. Alanine dipeptide and alanine tetrapeptide were built using Maestro graphical user interface as implemented in the Schrödinger software suite.[31] The coordinates of rilpivirine were extracted from the crystal structure of the complex formed by the inhibitor with HIV-1 reverse transcriptase deposited in the Protein Data Bank (2ZD1).[30] Point charges were derived from the electrostatic potential calculated after geometry optimization at the B3LYP/6-31G* level of theory with Gaussian 09,[32] following the RESP procedure. Amber FF99SB-ILDN[33] force field was used for alanine dipeptide and alanine tetrapeptide, while GAFF[34] was used to parametrize rilpivirine. All the systems were minimized

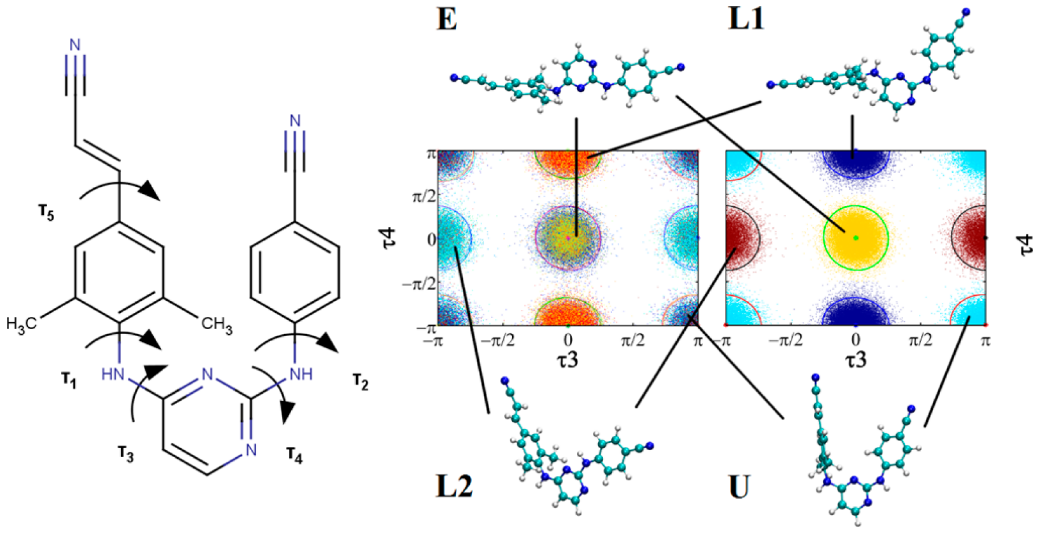


Figure 3. Rilpivirine exists in aqueous solvent as an ensemble of four different conformational states. If cluster analysis is performed in the space of the five dihedral angles of the molecule, 16 clusters are returned (scatter plot on the left). If the right symmetry is imposed on τ_1 and τ_2, symmetrically equivalent clusters are merged (plot on the right).

using steepest descent method for 500 steps and then conjugate gradient for 500 steps. Replica Exchange Molecular Dynamics (REMD)[35] simulations were performed using a Langevin thermostat with a collision frequency of 2 ps^{-1}. The time step was 2 fs. Bonds involving hydrogen atoms were restrained to their equilibrium length with the SHAKE algorithm.[36] The systems were heated using three steps of 5 ps each in the NVT ensemble with gradually reduced harmonic constraint on the heavy atoms at the temperature of 100, 200, and 300 K, respectively. A run of 10 ps in the NPT ensemble using a macroscopic pressure of 1 atm maintained with a weak-coupling scheme (Berendsen) and a relaxation time of 2 ps was used to equilibrate the water box. Each step was followed by minimization. All the simulations were performed using a short-range nonbonded cutoff of 8 Å, whereas long-range electrostatics was treated with the Particle Mesh Ewald method using a grid spacing of 1.0 Å and a cubic spline interpolation.[37] All the systems were solvated using TIP3P water molecules. The alanine dipeptide water box was composed of 396 molecules, using a cell margin distance from the solute of 8 Å in each dimension. Alanine tetrapetide box was composed of 539 water molecules, while rilpivirine was solvated using 736 molecules. REMD simulations were performed with 26 replicas across a geometrical distribution of temperatures between 300 and 600 K. The simulation time was 200 ns for each replica. Samples were collected every 2 ps in order to minimize correlations for a total of 100 000 samples per replica. Temperature exchanges between the i-th and the j-th replica were attempted with a frequency of 1 ps^{-1} and were accepted with probability:

$$\rho = \min\{1, \exp[-(\beta_i - \beta_j)(E_i - E_j)]\}$$

were β is defined as $(k_B T)^{-1}$, E is the energy of the system and i,j are the indexes of adjacent replicas. The simulations were performed using the Amber Molecular Dynamics Package version 10.[38]

The complex HIV1-RT/rilpivirine was prepared starting from the crystal structure deposited in the Protein Data Bank (2ZD1).[30] Amber FF99SB-ILDN[33] force field was used for protein parametrization, while rilpivirine was parametrized using the same procedure previously described for the simulation in water. The complex was solvated using 72.028 water molecules and neutralized using 8 chlorine negative ions for a total of 232.097 atoms. The system was minimized, thermalized and equilibrated using the same protocol and parameters of the previous simulations. A 20 ns production run of plain molecular dynamics was performed. The simulation of the complex was performed using the GPU version of the PMEMD program included in Amber Molecular Dynamics Package version 11.[39]

Data Reweighting. Suppose we have K different thermodynamic conditions, indexed by $i \in \{1,...,K\}$, characterized by trajectory probability densities:

$$p_i(\boldsymbol{x}) = Z_i^{-1} q_i(\boldsymbol{x}); \quad q_i(\boldsymbol{x}) \equiv e^{-\beta_i u_i(\boldsymbol{x})}; \quad Z_i \equiv \int d\boldsymbol{x}\, q_i(\boldsymbol{x})$$

where $q_i(\boldsymbol{x}) > 0$ is the unnormalized density, $u_i(\boldsymbol{x})$ is a potential that can be characterized by a modification of the Hamiltonian and Z_i an unknown normalization constant. It must be noted that, in our notation, vector (lowercase) and matrix (uppercase) arrays are written in bold. It is possible to recover the statistics of interest $p_0(\boldsymbol{x})$ using the multistate acceptance ratio method[10] to calculate the weights of all the sampled collected during the K simulations:

$$w_{in} = \frac{Z_i}{Z_0} e^{\beta_i u_i(\boldsymbol{x}_n)}$$

where w is the weight of a sample, i is the index of the replica and n is the index of the sample in each replica. The reference implementation of the multistate acceptance ratio, pyMBAR 2.0beta, was used to perform the calculations.[10]

Weighted K-Means Clustering. Given a set X = $\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N\}$ of observations, where each observation $x_i \in \mathbb{R}^d$ is a d-dimensional real vector, k-means clustering aims at partitioning the N observations into k sets ($k \leq N$) $S = \{S_1,S_2,...,S_k\}$ so as to minimize the within-cluster weighted sum of squared Euclidean distances:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in S_i} w_j \| \boldsymbol{x}_j - \boldsymbol{\mu}_i \|^2$$

where $w_j \in \mathbb{R}$ is the weight assigned to the j-th sample and $\boldsymbol{\mu}_i$ is the average of the observations belonging to cluster S_i. In particular, circular distances were calculated considering the periodicity of the dihedrals space. To determine an optimal clustering, the algorithm starts with an initial set of random k means. Then, it proceeds iteratively by alternating two steps: (i) each observation is assigned to the cluster with the closest mean; (ii) the new means are calculated as the centroids of the observations in the new cluster set. In our case, X was a set of vectors of dihedral angles and the initial set of means was chosen using the k-means++ initialization,[40] which guarantees better convergence. In k-means++, the first center is randomly selected with a uniform probability among the data points; then, the other centers are selected according to a weighted probability distribution, in which the probability of a given sample to be selected grows proportionally with the square of the distance from the nearest center. This approach rewards the separation among the clusters. For brevity, in the rest of the paper we will refer to the overall hard clustering procedure as *k-means++*. The partitioning obtained with the clustering is used as input for the estimation of a parametric representation of the PDF.

PDF Estimation. Periodic variables can be described using the multivariate Von Mises distribution.[41] However, if the sample are sufficiently concentrated around their mean, this distribution can be approximated by a multivariate Gaussian distribution where the difference between a sample and the mean is translated back to the reference periodic interval.[18] Therefore, a Gaussian distribution can still be used together with its useful properties as long as one takes into account the periodicity when calculating differences and averages.[41] The procedure can manipulate circular variables with different periodicity. The circular mean $\bar{\mu}$ is calculated by projecting the angular data from the original periodic window, ranging from *low* to *high* to the space between 0 and 2π:

$$\alpha_i = \frac{\boldsymbol{x}_i - \text{low}}{\text{high} - \text{low}} 2\pi$$

and then converting all dihedral angles (in radians) to complex numbers.

$$\bar{\boldsymbol{\mu}} = \arg \frac{1}{N} \sum_{i=1}^{N} \exp(i\alpha_i)$$

The mean obtained in this way can then be projected back to the original space.

For the covariance matrix, all the circular differences between the data and the mean are calculated. The matrix is then calculated in a linear space centered on the mean.

For each run of the k-means++, a PDF was estimated as a linear combination of multivariate Gaussian distributions (Gaussian Mixture Model or GMM):

$$\text{GMM} = \sum_{i=1}^{k} \pi_i G_i(\boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i)$$

where $G(\boldsymbol{\mu};\sum)$ is a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\sum$, while each $\pi_i \geq 0$ is a mixing coefficient. The mixing coefficients account for the normalization condition:

$$\sum_{i=1}^{k} \pi_i = 1$$

The parameters of the GMM were calculated using the Expectation Maximization (EM) algorithm.[20,42] Given a set of observations of a random variable distributed according to a parametric distribution, a Maximum Likelihood Estimation (MLE) consists in finding the maximum of the joint distribution of all N observations in the parameter space of the model.

$$\text{MLE} = \arg\max p(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N|\boldsymbol{\theta}) = \arg\max \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta})$$

where $\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N$ are the N observations, $\boldsymbol{\theta}$ is the vector of all the parameters of the model and the joint probability is calculated as a product under the assumption of independent and identically distributed samples.

EM is a two-step procedure for finding Maximum Likelihood solutions that, for multimodal distributions, converges to a local maximum. For a GMM, we can define the ≪responsibility≫ taken by the i-th component for ≪explaining≫ the observation x (this is the *a posteriori* probability that the sample is generated from the i-th component given a set of parameters for the mixture) as the value of the i-th component evaluated at $\boldsymbol{x}_n$ suitably normalized:

$$\gamma_{in} = w_n \frac{\pi_i G(\boldsymbol{x}_n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_k^{j=1} \pi_j G(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = w_n \frac{\omega_i(\boldsymbol{x}_n)}{\sum_k^{j=1} \omega_j(\boldsymbol{x}_n)}$$

where each sample $\boldsymbol{x}_n$ has a weight w_n calculated with the method described previously. In the E-step the current set of parameters $\{\boldsymbol{\mu}_j,\sum_j\}$ are used to evaluate the responsibilities. In the M-step the distribution parameters are re-estimated (using the current responsibilities) to maximize the likelihood of the data. For Gaussian distributions, this MLE can be done in closed form and for GMM it is

$$\pi_i = \frac{\sum_N^{n=1} \gamma_{in}}{\sum_N^{n=1} w_n}$$

$$\boldsymbol{\mu}_i = \frac{\sum_N^{n=1} \gamma_{in}\boldsymbol{x}_n}{\sum_N^{n=1} \gamma_{in}}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_N^{n=1} \gamma_{ni}(\boldsymbol{x}_n - \boldsymbol{\mu}_i)(\boldsymbol{x}_n - \boldsymbol{\mu}_i)^T}{\sum_N^{n=1} \gamma_{in}}$$

The clusters calculated using k-means++ were used to initialize the estimation of the parameters of the GMM. This procedure was iterated increasing the value of k by one, until the marginal gain in the log-likelihood function reached a plateau. This elbow was identified using a threshold of 1×10^3 as a stop criterion: when the gain was below that threshold the algorithm stopped. In the last step, each sample was assigned a probability of belonging to a given component. Each sample was assigned to the component with the highest probability. Since the algorithm converges to a local minimum, 100 independent restarts were performed to extend the search space and the one presenting the best final log-likelihood was selected. This approach based on a combination of k-means++ and EM turned out to be more accurate than a simple parametrization based only on k-means++. At the same time, our strategy provided GMMs with similar accuracy with respect to EM preceded by random initialization while depending on fewer components. Details on the comparison among these three initialization methods are reported in the Supporting Information.

Clustering and Gradient-Quadratic Search of Modes. The representation of the PDF in terms of a GMM can be easily exploited to perform a clustering of the conformational ensemble. For a GMM composed by k components, it is possible to assign to each sample k probability densities. These numbers can be interpreted as the probability of "belonging" to each specific component. The representative sample for each state is calculated as the sample with the highest probability of belonging to that component.

This clustering can be further refined using information about the modes of the GMM (i.e., local maxima). In fact, the modes of a PDF representing the conformational equilibrium of a molecule correspond to free energy minima, which can, in turn, identify metastable conformational states. If a conformational basin distribution has non-Gaussian tails, it can happen that more than one component of the mixture are needed to represent it. In the present context, we are estimating a multivariate PDF with an undetermined number of modes. We can take advantage of the GMM representation to automatically identify the different conformational states by enumerating their modes. The modes can be found using a gradient-quadratic search.[43] This approach benefits from the analytical tractability of Gaussian distributions. In fact, the gradient $\boldsymbol{g}$ and Hessian $\boldsymbol{H}$ of a Gaussian mixture with respect to the independent variables $\boldsymbol{x}$ exist $\forall\ x \in \mathbb{R}^d$:

$$\boldsymbol{g} = \sum_{i=1}^{k} \pi_i \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{x})$$

$$\boldsymbol{H} = \sum_{i=1}^{k} \pi_i \boldsymbol{\Sigma}_i^{-1}[(\boldsymbol{\mu}_i - \boldsymbol{x})(\boldsymbol{\mu}_i - \boldsymbol{x})^T - \boldsymbol{\Sigma}_i]\boldsymbol{\Sigma}_i^{-1}$$

The procedure consists of a hill-climbing algorithm that starts from every centroid of the mixture. It is straightforward to use quadratic maximization combined with gradient ascent. In particular, the quadratic maximization moves from $\boldsymbol{x}$ toward a zero-gradient point that can be a maximum, a minimum, or a saddle point. For maximization, the Hessian can only be used if it is negative definite (i.e., all its eigenvalues are negative). If the Hessian is not negative definite, the gradient ascent is used to move in the direction of the gradient. Once a point for which $\boldsymbol{g} = 0$ is found, the Hessian can confirm that the point is a maximum by checking that $\boldsymbol{H} < 0$. The algorithm is controlled by five parameters: (i) the minimum length of a gradient ascent step; (ii) the minimum gradient norm, below which the gradient is considered numerically zero; (iii) the minimum absolute difference between two modes to be assimilated as the same (this parameter controls whether two or more clusters identified in the first part of the clustering must be merged); (iv) the maximum value of the algebraically largest eigenvalue for the Hessian to be considered negative definite; (v) maximum number of iterations to limit the computation time. The values of parameters suggested in ref 41 and references therein were here adopted.

Identification of the Most Informative Variables. The approach we use to select the most informative variables to represent the conformational properties of a molecule is derived from Principal Component Analysis (PCA), and it is based on the covariance matrix of the class means, which was calculated using the formula:

$$\boldsymbol{\Sigma}_{\text{clust}} = \sum_{i=1}^{k} w_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{\text{TOT}})\,(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{\text{TOT}})^T$$

were $\boldsymbol{\mu}_{\text{TOT}}$ is the total average, while $\boldsymbol{\mu}_i$ and w_i are the mode and the weight of each cluster, respectively. Eigenvalues and eigenvectors of this matrix were calculated and the eigenvector with the highest eigenvalue was selected. The absolute value of each component of this eigenvector represents the weight of each variable on the linear projection over the axis represented by the eigenvector itself. The variable associated with the greatest weight is selected as the most informative. In order to find the second most informative variable, a reduced PDF is obtained by integrating over the previously selected variable. The integration of the GMM can be performed analytically by excluding the rows and columns corresponding to the selected variable from the covariance matrix and from the mean. This procedure can easily be repeated on the reduced PDFs to obtain an ordered set of the most informative variables of the distribution.

■ RESULTS

Alanine Dipeptide. Alanine dipeptide, an alanine residue terminally capped, is a widely adopted benchmark for free energy estimation methods. As reported in the literature, the conformational space of this system can be well described using only two dihedral angles (φ and ψ).[23,24] The system we simulated was composed by one alanine dipeptide molecule and 396 TIP3P water molecules. Twenty-six replicas of the system were sampled by Replica Exchange Molecular Dynamics (REMD). Data were collected for a total of 2 600 000 snapshots and reweighted using the MBAR method.[10] The free energy surface of alanine dipeptide in water is composed by four different conformational states.[44,45] As reported in Figure 1, all the conformational basins can be explored using the REMD protocol. Histogram analysis (Table 1 and Figure 1) reveals a landscape in agreement with what was previously reported. The histogram was calculated dividing each dimension in 100 bins. Maxima were found using exhaustive search and were ranked according to their probability density

Table 1. Conformational Populations of Alanine Dipeptide in TIP3P Water

	alanine dipeptide (TIP3P water)			
	histogram		GMM	
conformations	population	maxima	population	maxima
PP_{II}	34.6%	−73.8;149.4 −73.8;156.6	42.0%	−73.5;152.1
β	25.9%	−142.2;153.0 −142.2;160.2	20.9%	−141.6;154.7
α_{Rest}	24.7%	−81.0;−12.6 −73.8;−19.8	22.4%	−81.3;−13.3
α_{Rwest}	9.9%	−131.4;−9.0 −124.2;−12.6 −120.6;1.8	12.6%	−127.8;−6.9
α_{L}	1.6%	55.8;27 52.2;−34.2 52.2;−41.4	2.1%	53.8, 28.1

calculated in each bin. In order to avoid spurious maxima, bins in poorly populated regions ($P < 10^{-5}$) were discarded. Conformational basins identified are α_R, α_L, β, and PP_{II}, with the α_R state, which can be further separated into two subdomain: α_{Rest} and α_{Rwest}. According to an exhaustive search of local minima, there is more than one minimum near the center of each basin. In this case, after visual inspection of the contour plot (Figure 1A), the minimum of each basin was expressed as the arithmetic mean among the energy values of the local minima. In Table 1, all the minima found near the center of each basin are explicitly reported. α_{Rest} has a minimum at $(\varphi;\psi) = (-77.4;-16.2)$, $P = 24.7\%$, and $F = 0.84$ kcal/mol; α_{Rwest} has a minimum at $(\varphi;\psi) = (-125.4;-6.6)$, $P = 9.9\%$, and $F = 1.4$ kcal/mol; α_L has minimum at $(\varphi;\psi) = (-53.4;34.2)$, $P = 1.6\%$, and $F = 2.47$ kcal/mol; β has a minimum at $(\varphi;\psi) = (-142.2;\ 156.6)$, $P = 25.9\%$ and $F = 0.81$ kcal/mol; PP_{II} has a minimum at $(\varphi;\psi) = (-73.8;153.0)$ with $P = 34.6\%$ and $F = 0.64$ kcal/mol. Other minima, located in transition regions and tail of the basins, were discarded from further analysis. However, they are graphically reported in the contour plot of Figure 1A as black dots. The GMM method converges toward a representation composed by eight components. Results are in line with those obtained with the histogram analysis. The gradient-quadratic merged some components and converged toward five modes (Figure 1B), which accurately correspond to the five basins identified with the histogram method (Table 1): α_{Rest} with minimum at $(\varphi;\psi) = (-81.3;-13.3)$, $P = 22.4\%$, and $F = 0.84$ kcal/mol; α_{Rwest} with minimum at $(\varphi;\psi) = (-127.8;-6.9)$, $P = 12.6\%$, and $F = 1.24$ kcal/mol; α_L with minimum at $(\varphi;\psi) = (53.8;28.1)$, $P = 2.1\%$, and $F = 2.33$ kcal/mol; β with minimum at $(\varphi;\psi) = (-141.6;154.7)$, $P = 20.9\%$, and $F = 0.94$ kcal/mol; PP_{II} with minimum at $(\varphi;\psi) = (-73.5;\ 152.1)$ with $P = 42.0\%$ and $F = 0.52$ kcal/mol. When components are merged, their weighted average (using the weights of the respective components) was taken as reference value for the mean. The comparison between the populations calculated using the two methods (Figure 1 and Table 1) shows that the GMM representation, while being much more compact, retains the same information supplied by the histogram method and avoids spurious local maxima. The results of soft clustering are shown as a scatter plot in Figure 1C. We also explored this model system by means of Accelerated MD.[46] For all intents and purposes the proposed protocol of analysis is independent of the sampling procedure, provided that the latter is exhaustive. Details are provided in the Supporting Information (SI).

Alanine Tetrapeptide. Alanine tetrapeptide contains three alanine residues terminally capped, so its conformational space can be described using six dihedral angles (two for each alanine residue). Alanine-based polypeptides, both in silico and in experiments, show a shift in the equilibrium toward right-handed α-helix conformations.[47] GMM method returns a model composed by 17 components. As shown in Figure 2 and Table 2, results of our simulation are in agreement with those previously reported. For each alanine residue of the tetrapeptide, a label was assigned according to the value of the relative φ/ψ dihedrals couple. The nomenclature of the conformational states of each alanine residue follows the nomenclature used for alanine dipeptide in water. The probability of finding a full α_R configuration, calculated aggregating all the clusters representing $\alpha_{Rest}/\alpha_{Rwest}$-only classes, is about 94%. In two cases, more than one component converged to the same mode: clusters 2 and 8 toward a mode representing an "α_{Rest},α_{Rwest},α_{Rwest}" configuration, clusters 9 and 16 toward a mode representing an "α_{Rwest},α_{Rest},α_{Rwest}" configuration. Visual inspection of representative conformations for each component confirmed that in both cases the component could be merged in the same conformational basin.

Table 2. Conformational Populations of Alanine Tetrapeptide in TIP3P Water

	alanine tetrapeptide (TIP3P water)	
conformations	populations	maxima
α_{Rest}, α_{Rwest}, α_{Rest}	14.45%	−61.0; −19.0; −132.6; 7.0; −73.1; −11.8
α_{Rest}, α_{Rwest}, α_{Rwest}	14.3%	−56.8; −19.0; −130.0; 18.5; −138.7; 3.4
	11.02%	−64.6; −21.0; −136.7; 4.5; −139.8; 8.3
α_{Rwest}, α_{Rwest}, β	0.73%	−138.7; −1.8; −116.2; −2.0; −133.9; 158.8
α_{Rwest}, α_{Rest}, α_{Rest}	5.79%	−140.9; −0.3; −63.1; −23.9; −74.7; −9.2
α_{Rwest}, β, α_{Rwest}	1.13%	−117.5; −5.6; −130.0; 155.0; −119.5; −2.1
α_L, α_{Rwest}, α_{Rwest}	0.64%	52.9; 28.2; −205.5; −0.9; −109.3; −1.2
α_{Rwest}, α_L, α_{Rwest},	1.28%	−120; −3.5; 46.3; 42.7; −124.5; 1.3
α_{Rwest}, α_{Rest}, α_{Rwest}	8.57%	−140.9; −0.1; −64.5; −18; −133.5; 7.4
	2.27%	−139.9; 0.1; −68.9; −20.9; −132.4; 7.7
α_{Rwest}, α_{Rwest}, α_{Rest}	6.96%	−140.8; −0.5; −139.5; 8.2; −74.3; −11.8
α_{Rwest}, α_{Rwest}, α_{Rwest}	11.89%	−141.4; 0.5; −140.3; 6.8; −137.7; 9.2
α_{Rest}, α_{Rwest}, β	0.73%	−62.4; −20.4; −124.7; 3.4; −120.4; 154.1
α_{Rest}, α_{Rest}, α_{Rwest}	12.2%	−57.3; −27.4; −67.3; −11.4; −132.6; 3.2
α_{Rwest}, α_{Rwest}, α_{Rest}	0.93%	−100.2; −8.0; −115.5; 4.1; 55.0; 22.5
α_{Rest}, α_{Rest}, α_{Rest},	6.53%	−59.4; −27.6; −66.0; −19.3; −76.2; −10.2
α_{Rwest}, α_{Rwest}, α_{Rwest}	0.58%	−117.3; 157.3; −140.7; 2.4; −116.3; 4.4

Rilpivirine. Rilpivirine belongs to the class of dyaril-pyrimidines. This class of compounds is characterized by a pyrimidine ring substituted bye two aromatic moieties. In this specific case, the molecule displays five rotatable bonds denoted $\tau_1-\tau_5$, respectively (Figure 3). According to data reported in the literature, it is sufficient to take into account only τ_3 and τ_4 to separate the most relevant conformational states.[8] In particular, there are four nonsymmetrically equivalent states, labeled E, L_1, L_2, and U since the shapes adopted by the molecule resemble those of the respective roman letters. τ_1 and τ_2 represent rotations of symmetric aromatic rings, while τ_5 does not appreciably deviates from the equilibrium value throughout the simulation. For these reasons, these dihedral angles do not seem informative in describing the overall shape of the molecule. The results obtained by means of the GMM method without taking into account the symmetry around τ_1 and τ_2 are reported in Table 3. Sixteen clusters were returned: four symmetrical clusters for each of the expected conformations. Symmetries were suitably imposed in the k-means++/EM procedures by changing the periodicity of the variables. For τ_1 and τ_2 a periodicity of π was imposed. The algorithm correctly merges the symmetrically equivalent clusters into four: E with minimum at $(\tau_3;\tau_4) = (0.1;-0.2)$, $P = 36.98\%$, and $F = 0.6$ kcal/mol; L_1 with minimum at $(\tau_3;\tau_4) = (-0.2;180.0)$, $P = 49.71\%$, and $F = 0.42$ kcal/mol; L_2 with minimum at $(\tau_3;\tau_4) = (179.6;0.3)$, $P = 6.3\%$, and $F = 1.66$ kcal/mol; U with minimum at $(\tau_3;\tau_4) = (179.9;-179.6)$, $P = 7.01\%$, and $F = 1.59$ kcal/mol. The U state corresponds to the conformation adopted by rilpivirine when bound to HIV1-RT.[30] All the results are in agreement with the results reported by Okumura and coworkers in ref 8, within the limits of uncertainties and different setup of the simulations.

Dimensionality Reduction. In the second part of our analysis, we focused on the problem of dimensionality

Table 3. Conformational Populations of Rilpivirine in TIP3P Water

	rilpivirine (TIP3P water)					
	no symmetry			symmetry		
	clusters	maxima	population	clusters	maxima	population
E	2	89.8; 180.0; 0.6; −0.7	8.34%	1	0.1	36.98%
	3	−90.8; 0.2; 0; 0.6	10.29%		−0.2	
	7	90.8; 0.2; 0; 0.1	7.38%			
	8	−89.7; 179.7; −0.1; −0.8	10.97%			
L_1	1	−90.1; 179.8; 0.1; −179.7	14.06%	2	−0.2	49.71%
	6	−90; 0.9; −0.3; −179.6	13.53%		180.0	
	9	91; 178.8; −0.1; 179.5;	10.50%			
	10	90.2; 0.6; −0.3; 179.4	11.62%			
L_2	4	92.0; −179.6; 178.7; 0.2	1.86%	3	179.6	6.3%
	11	−92.1; −176.8; −179.9; 2.3	0.97%		0.3	
	12	−91.2; −1; 179.9; −1.8	1.4%			
	13	90.5; −0.9; 179.9; 0.9	2.07%			
U	5	90.1; −1.0; 179.6; 177.6	2.04%	4	179.9	7.01%
	14	−88.8; 2.1; 179.9; −178.4	1.44%		−179.6	
	15	90.5; −179.6; −179.4; −179.7	1.83%			
	16	−87.9; −177.9; 179.7; −177.0	1.7%			

reduction, with the aim to select the most informative variables with an unbiased and automatic method. Usually, CVs selection is performed manually according to users' experience, prior knowledge, and "chemical intuition". This problem is independent of the estimation of the PDF and the partitioning of the conformational ensemble in the higher dimensional space, so we focused only on rilpivirine simulation data. We decided to exploit the simplicity of the parametric representation to simplify the problem of dimensionality reduction. In particular, we focused on linear methods. A standard approach is the Principal Component Analysis (PCA), the basis of quasi-harmonic analysis and essential dynamics methods[48−50] the application of which has been also extended to the internal variables of the dihedral space.[51,52] Here, as explained in the methods section, we used the covariance matrix of the *means* class to identify the variables that better preserved the separation among the clusters. This dimensionality reduction approach allowed the unbiased selection of the variables that best describe the separation among different conformational states, namely τ_3 and τ_4, in good agreement with previous reports and chemical intuition. The PDF of the joint probability density function $P(\tau_3,\tau_4)$ was obtained from the original GMM analytically by integration.

HIV1-RT/Rilpivirine Complex. In the final part of the work, the complex formed by rilpivirine and wild-type HIV1 Reverse Transcriptase was simulated using plain molecular dynamics. The simulation was analyzed in order to provide an estimate of the entropic change upon binding.

In the first step, we performed an analysis on the ensemble of rilpivirine simulated in water using a GMM composed by 100 components. For this analysis, the components were not merged according to the gradient-quadratic search described in the Methods, resulting in the ensemble being partitioned in exactly 100 clusters. The use of a high number of clusters was chosen to allow a finer description of the conformational variability of the molecule in water (Figure 4 blue color). The population of each conformational state was correctly recovered, in particular: E was composed by 31 clusters (36.7% of samples), L_1 by 38 clusters (49.7%), L_2 by 16 clusters (5.6%), and U by 15 clusters (7.0%). In the second step, samples of the HIV1-RT/rilpivirine complex MD simulation were collected and classified according to the clustering done in the previous step. As shown in Figure 4 (red color), while in the bound state, rilpivirine is able to explore only one of the four main conformational states, namely that labeled as U. This state corresponds to a local minimum in water. More precisely, only a subset (12 out of 15) of the clusters corresponding to the U conformation is explored by the bound ligand, suggesting a further conformational restraining induced by the protein.

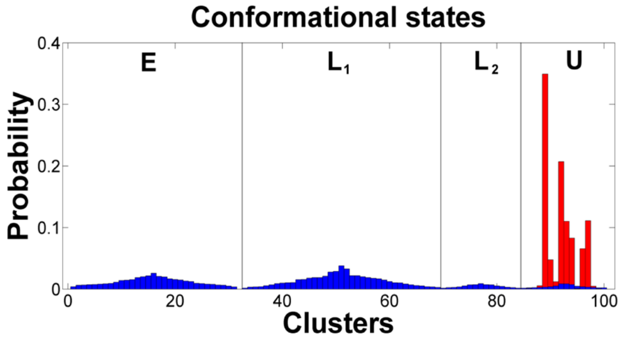


Figure 4. Cluster analysis of rilpivirine conformational ensemble in water using 100 clusters. All of the four conformational states (E, L_1, L_2, and U) are populated (blue bars). In the HIV1-RT/rilpivirine complex, only some clusters belonging to the U conformational state are populated (red bars).

DISCUSSION AND CONCLUSIONS

We have presented here a method that relies on the use of Gaussian Mixtures Models to approximate the probability density function describing the conformational equilibrium of small molecules in aqueous solution. The PDF fitting does not require any prior knowledge on the number or on the location of conformational states. The GMM method displays an improved scalability with respect to the size of the system relative to histogram-based approaches. A histogram requires a number of parameters that grows asymptotically with the order of $O(n^d)$ where n is the number of bins and d is the number of variables. This means that, using 100 bin for each dimension, as in the case of alanine dipeptide both in vacuum and water, 100^2 bins are needed. This figure rises up to 100^5 for rilpivirine and 100^6 for alanine tetrapeptide. The representation of a GMM with k components requires k parameters for the weights, kd parameters for the means and $kd(d+1)$ parameters for the covariance matrices. Thus, the computational complexity grows asymptotically with the order of $O(kd^2)$. In spite of this huge difference in the complexity of the representation, GMM efficiently preserve all the information contained in the histograms. As an example, in Figure 5, the histogram of the

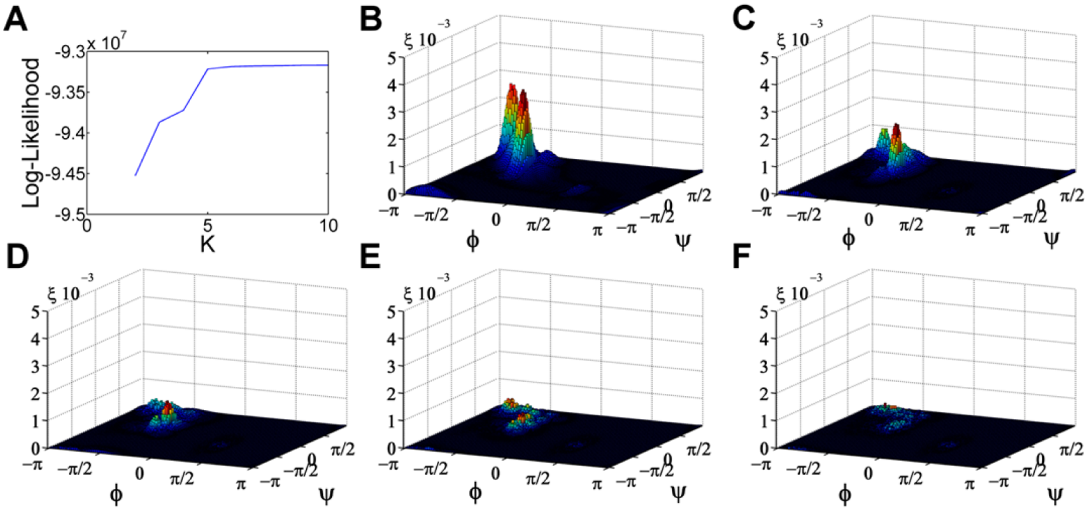


Figure 5. Log-likelihood is the probability to observe the samples of a data set given a set of parameters. For a GMM approximating an arbitrary smooth PDF, with respect to the number of components, it is a (theoretically) monotonically increasing function that reaches a plateau. (A) The log-likelihood of the GMM estimated for alanine dipeptide in vacuum (see SI for details). (B) The absolute difference between the value of the PDF calculated as a histogram and the value of the PDF calculated as a GMM, in each bin, for alanine dipeptide in vacuum is plotted for $K = 2$. (C) $K = 3$. (D) $K = 4$. (E) $K = 5$. (F) $K = 10$. The difference decreases as the number of components increases until it reaches values below the threshold of uncertainty in the estimation. Beyond this point, adding further components becomes useless.

alanine dipeptide in water was taken as a reference solution and the absolute difference with respect to the PDF modeled as a GMM was plotted. Differences were calculated evaluating and normalizing the probability densities represented by the GMM on a 100 × 100 grid, with each point of the grid centered in the corresponding bin of the histogram. The calculation was done for increasing values of k. The differences between the two methods decrease when the number of components increases. Moreover, using GMM, expressions particularly useful in molecular dynamics and statistical mechanics calculations such as those for gradient, Hessian matrix, related quantities (and their logarithmic forms) are available in closed form.

The strategy proposed to locate maxima for the GMM by an approximation of the PDF with arbitrary covariance matrices, in general, does not guarantee that all modes are recovered. However, it is to be noted that only the modes corresponding to more densely populated basins are expected to have a physical significance. In general, GMM modes can be located also outside the convex hull of component centroids and can even exceed the number of components itself.[53] For instance, a new "spurious" mode appears whenever the tails of elongated components intersect. A clear example is reported in ref 52. However, tails represent regions poorly populated; hence, in the present context, it is safer to ignore them. In our implementation, if all the most populated basins are correctly sampled, this does not constitute a problem, since the search for maxima starts from the centroids of the component, which should be already close to a meaningful mode.

On the same line, the main drawback of this approach is that the PDF estimate is less accurate in those regions of the domain that are poorly sampled during the simulation. While REMD is globally able to efficiently visit the entire conformational space, in the target replica (usually, the one at 300 K) the sampling of the high-energy regions, corresponding to the highest barriers, can be insufficient. This happens because barriers are lowered and sampled only for the replica simulated at a very high temperature. However, the weights of these replicas are exponentially lowered during the reweighting, so that the effects on the estimated PDF are minor. For this reason, we focused on the discretization of the conformational space in different states and we did not attempt to quantify energetic barriers or to identify transition states. However, it is worth pointing out that GMM and histograms are equally affected by this problem. Sampling in high-energy regions could be improved using bias approaches that push the system toward high-energy configurations. This issue is usually tackled using umbrella sampling or adaptively biased simulations. In their work, Maragakis and colleagues[16] described the Gaussian Mixture Umbrella Sampling (GAMUS) method to bias simulations using GMM. GAMUS is an adaptive biasing technique, which can be used to escape free energy minima and improve sampling in poorly sampled regions. Here, we propose a slightly different way to estimate the parameters and a method to choose the number of components for the GMM. In GAMUS, the number of components of the mixture is initially set and then iteratively increased, while parameters are estimated using multiple trials of the EM algorithm. In our approach, we perform multiple trials of k-means++ to feed the EM algorithm and the number of components is chosen using the log-likelihood of the PDF. In this way, it is possible to automatically choose the best number of components for the representation and avoid overfitting. It is important to highlight that, while resting on a heuristic method for the initialization, the overall final protocol is robust. In fact, the output does not vary across multiple independent instances of the procedure (data shown in SI).

We also showed how it is possible to exploit the properties of GMM to get a soft clustering of the conformational ensemble and to calculate thermodynamic properties of the conformational ensemble. As an example, in our tests, we exploited the GMM representation to perform a soft clustering of the conformational ensembles and to calculate the free energy penalties needed to constrain a small molecule to each free energy minimum (which correspond to the modes of the PDF). According to the so-called *conformational selection* model for binding, we can consider each minimum as a different binder to a given target and calculate the free energy penalty needed to limit the accessible conformational space to that specific state. When fast algorithms are needed, for instance in virtual screening, this issue is addressed using crude approximations such as energetic penalties proportional to number of rotatable bonds.[4,5] On the other hand, current sampling techniques can provide an exhaustive and accurate sampling of small-molecules conformational populations. This opens the possibility for a deeper analysis of this aspect of protein−ligand binding. The case of rilpivirine is particularly interesting. Rilpivirine binds HIV1-RT only in the U conformation,[30] so the reorganization free energy is predicted to be $F = 1.59$ kcal/mol as calculated from a population of $P = 7.01\%$. This value is significant if compared to the measured binding free energy (∼ −10 kcal/mol) and comparable to the value calculated in ref 8. In order to further characterize the differences between the conformational ensemble of the ligand in water and that in the bound state, we also simulated the complex. The comparison between the two ensembles (see Figure 4) highlights two quite interesting aspects: (i) the presence of a conformational selection mechanism upon binding, where only conformations belonging to an otherwise scarcely populated U basin were observed; (ii) a further restriction on the conformational freedom within state U, due to the interaction with the protein. The complex formation shows a substantial loss of freedom of the τ_3 and τ_4 angles.

Here, we also approached the problem of identifying the minimum subset of degrees of freedom that is required to describe the conformational ensemble. Low dimensional representations are more intuitive and easier to analyze, but they are useful only as long as they convey the relevant information. If this goal is met, the information gathered in comparatively simpler models, such as those that consider the small molecule in aqueous environment, is instrumental to the efficient simulation of more complex and more interesting cases such as the binding to its macromolecular target. In fact, accelerating the sampling of the conformational space just along the previously characterized most relevant degrees of freedom can save a huge amount of computational effort. We showed that standard linear methods of dimensionality reduction, even in their simplest form such as PCA, can take advantage of the GMM representation.

In principle, our method is not limited to the analysis of conformations in terms of dihedral angles. The very same approach can be used to independently analyze a high number of CVs. Ideally, it should be possible to (i) analyze correlations between CVs and (ii) select in an unsupervised and unbiased way the CVs which best preserve the clustering in the original higher-dimensional space.

In conclusion, we have reported on a novel theoretical and computational approach to the identification of conformations of organic molecules bearing different levels of chemical complexity. This approach can be utilized in several chemical endeavors ranging from drug discovery, to organic and biological chemistry. We have shown that our approach can perform, in terms of accuracy, at the same level of other previously reported methodologies in the field. In addition, our approach allows reducing conformational dimensionality and therefore computational effort. More importantly, reduced dimensionality can also allow reaching faster and more accurate convergence in free energy and kinetic estimations, an issue that often arises when running enhanced sampling calculations in chemical and biophysical settings. In one case of pharmaceutical relevance, that of rilpivirine and HIV1-RT, the approach suggested in the present work was able to provide a limited set of relevant degrees of freedom and a consequent partitioning of the conformational space of the solvated ligand apt to differentiate the "bioactive" state from the others by making no use of any a priori knowledge on either the target system or the binding process.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Details about (i) alanine dipeptide simulated in vacuum, (ii) accelerated MD simulations, (iii) the robustness of the generated GMM notwithstanding a k-means++ initialization, (iv) a comparison between k-means++, random-EM, and k-means++ plus EM initialization methods, and (v) the selection of a subset of the most relevant CVs. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Email: giovanni.bottegoni@iit.it.
*Email: walter.rocchia@iit.it.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

PDF, probability density function; GMM, Gaussian mixture model; MLE, Maximum Likelihood Estimation; HIV-1 RT, HIV-1 Reverse Transcriptase

■ REFERENCES

(1) Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* **2009**, *5*, 789−96.

(2) Weikl, T. R.; von Deuster, C. Selected-Fit Versus Induced-Fit Protein Binding: Kinetic Differences and Mutational Analysis. *Proteins: Struct., Funct., Bioinf.* **2009**, *75*, 104−10.

(3) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure* **2009**, *17*, 489−98.

(4) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization Upon Binding. *J. Med. Chem.* **2004**, *47*, 2499−510.

(5) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−41.

(6) Gallicchio, E.; Levy, R. M. Advances in All Atom Sampling Methods for Modeling Protein-Ligand Binding Affinities. *Curr. Opin. Struct. Biol.* **2011**, *21*, 161−6.

(7) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72*, 1047−69.

(8) Okumura, H.; Gallicchio, E.; Levy, R. M. Conformational Populations of Ligand-Sized Molecules by Replica Exchange Molecular Dynamics and Temperature Reweighting. *J. Comput. Chem.* **2010**, *31*, 1357−67.

(9) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein−Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880−4.

(10) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105−10.

(11) Kobrak, M. N. Systematic and Statistical Error in Histogram-Based Free Energy Calculations. *J. Comput. Chem.* **2003**, *24*, 1437−46.

(12) Wang, Q.; Pang, Y. P. Preference of Small Molecules for Local Minimum Conformations When Binding to Proteins. *PLoS One* **2007**, *2*, e820.

(13) Eckart, C. Some Studies Concerning Rotating Axes and Polyatomic Molecules. *Phys. Rev.* **1935**, *47*, 552−558.

(14) Kudin, K. N.; Dymarsky, A. Y. Eckart Axis Conditions and the Minimization of the Root-Mean-Square Deviation: Two Closely Related Problems. *J. Chem. Phys.* **2005**, *122*, 224105−2.

(15) Gō, N.; Scheraga, H. A. On the Use of Classical Statistical Mechanics in the Treatment of Polymer Chain Conformation. *Macromolecules* **1976**, *9*, 535−542.

(16) Maragakis, P.; van der Vaart, A.; Karplus, M. Gaussian-Mixture Umbrella Sampling. *J. Phys. Chem. B* **2009**, *113*, 4664−73.

(17) Piro, P.; Pisani, P.; Bottegoni, G.; Sona, D.; Rocchia, W.; Cavalli, A.; Murino, V. Fitting and Simplification of Mixture for Clustering Conformational Populations of Small Organic Molecules. In *33th International Conference on Chemical and Biological Engineering (ICCBE)*, Stockholm, Sweden, 2012.

(18) Tribello, G. A.; Ceriotti, M.; Parrinello, M. A Self-Learning Algorithm for Biased Molecular Dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 17509−14.

(19) MacQueen, J., Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. And Probability (Berkeley, Calif., 1965/66)*; Univ. California Press: Berkeley, Calif., 1967; Vol. Vol. *I*, Statistics, pp 281−297.

(20) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **1977**, *39*, 1−38.

(21) Arthur, D.; Manthey, B.; Roglin, H. K-Means Has Polynomial Smoothed Complexity. *50th Annual IEEE Symposium on Foundations of Computer Science*, Atlanta, GA, Oct. 25−27, 2009; IEEE: Atlanta, GA, 2009; pp 405−414.

(22) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein−Ligand Interaction. *Proteins: Struct., Funct., Bioinf.* **2002**, *49*, 457−71.

(23) Hermans, J. The Amino Acid Dipeptide: Small but Still Influential after 50 Years. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 3095−6.

(24) Jalkanen, K. J.; Elstner, M.; Suhai, S. Amino Acids and Small Peptides as Building Blocks for Proteins: Comparative Theoretical and Spectroscopic Studies. *J. Mol. Struct.: THEOCHEM* **2004**, *675*, 61−77.

(25) Kim, Y. S.; Wang, J.; Hochstrasser, R. M. Two-Dimensional Infrared Spectroscopy of the Alanine Dipeptide in Aqueous Solution. *J. Phys. Chem. B* **2005**, *109*, 7511−21.

(26) Tazaki, K.; Shimizu, K. Molecular Dynamics Simulations in Aqueous Solution: Application to Free Energy Calculation of Oligopeptides. *J. Phys. Chem. B* **1998**, *102*, 6419−6424.

(27) Tobias, D. J.; Brooks, C. L., 3rd Thermodynamics and Mechanism of α-Helix Initiation in Alanine and Valine Peptides. *Biochemistry* **1991**, *30*, 6059−70.

(28) Tsai, M. I.; Xu, Y.; Dannenberg, J. J. Ramachandran Revisited. Dft Energy Surfaces of Diastereomeric Trialanine Peptides in the Gas Phase and Aqueous Solution. *J. Phys. Chem. B* **2009**, *113*, 309−18.

(29) Miick, S. M.; Martinez, G. V.; Fiori, W. R.; Todd, A. P.; Millhauser, G. L. Short Alanine-Based Peptides May Form 3(10)-Helices and Not α-Helices in Aqueous Solution. *Nature* **1992**, *359*, 653−5.

(30) Das, K.; Bauman, J. D.; Clark, A. D., Jr.; Frenkel, Y. V.; Lewi, P. J.; Shatkin, A. J.; Hughes, S. H.; Arnold, E. High-Resolution Structures of Hiv-1 Reverse Transcriptase/Tmc278 Complexes: Strategic Flexibility Explains Potency against Resistance Mutations. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1466−71.

(31) *Maestro*, version 9.3; Schrödinger, LLC: New York, 2012.

(32) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision B.01; Gaussian, Inc.: Wallingford, CT, 2009.

(33) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99sb Protein Force Field. *Proteins* **2010**, *78*, 1950−8.

(34) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−74.

(35) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(36) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(37) Pratt, L. R.; Hummer, G. *Simulation and Theory of Electrostatic Interactions in Solution: Computational Chemistry, Biophysics, and Aqueous Solutions*; American Institute of Physics Inc.: Santa Fe, NM, 1999; p 534.

(38) Case, D.A.; T. A. D, Cheatham, T.E.; , III, Simmerling, C.L.; Wang, J.; Duke, R.E.; Luo, R.; Crowley, M.; R. C.Walker,Zhang, W.; Merz, K.M.; B.Wang, Hayik, S.; Roitberg, A.; , Seabra, G.; , I. Kolossváry, K. F.Wong, Paesani, F.; Vanicek, J.; X.Wu, Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *Amber 10*, University of California: San Francisco, CA, 2008.

(39) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I. ; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *Amber 11*; University of California: San Francisco, CA, 2010.

(40) Arthur, D.; Vassilvitskii, S. K-Means++: The Advantages of Careful Seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*; Society for Industrial and Applied Mathematics: New Orleans, LA, 2007; pp 1027−1035.

(41) Mardia, K. V.; Hughes, G.; Taylor, C. C.; Singh, H. A Multivariate Von Mises Distribution with Applications to Bioinformatics. *Can. J. Stat.* **2008**, *36*, 99−109.

(42) Redner, R.; Walker, H. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Rev.* **1984**, *26*, 195−239.

(43) Carreira-Perpinan, M. A. Mode-Finding for Mixtures of Gaussian Distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1318−1323.

(44) Poon, C.-D.; Samulski, E. T.; Weise, C. F.; Weisshaar, J. C. Do Bridging Water Molecules Dictate the Structure of a Model Dipeptide in Aqueous Solution? *J. Am. Chem. Soc.* **2000**, *122*, 5642−5643.

(45) Hu, H.; Elstner, M.; Hermans, J. Comparison of a QM/MM Force Field and Molecular Mechanics Force Fields in Simulations of Alanine and Glycine "Dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in Water in Relation to the Problem of Modeling the Unfolded Peptide Backbone in Solution. *Proteins* **2003**, *50*, 451−63.

(46) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919−29.

(47) Vila, J.; Williams, R. L.; Grant, J. A.; Wojcik, J.; Scheraga, H. A. The Intrinsic Helix-Forming Tendency of L-Alanine. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 7821−5.

(48) Ichiye, T.; Karplus, M. Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins* **1991**, *11*, 205−17.

(49) Garcia, A. E. Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696−2699.

(50) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412−25.

(51) Mu, Y.; Nguyen, P. H.; Stock, G. Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 45−52.

(52) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations. *J. Chem. Phys.* **2007**, *126*, 244111.

(53) Carreira-Perpiñán, M. Á.; Williams, C. K. I. On the Number of Modes of a Gaussian Mixture. In *Proceedings of the 4th International Conference on Scale Space Methods in Computer Vision*; Springer-Verlag: Isle of Skye, U.K., 2003; pp 625−640.