International Journal of
Molecular Sciences

MDPI

Review

The Two-Step Clustering Approach for Metastable States Learning

Hangjin Jiang [1] **and Xiaodan Fan** [2,*]

[1] Center for Data Science, Zhejiang University, Hangzhou 310058, China; jianghj@zju.edu.cn
[2] Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China
* Correspondence: xfan@cuhk.edu.hk

Abstract: Understanding the energy landscape and the conformational dynamics is crucial for studying many biological or chemical processes, such as protein–protein interaction and RNA folding. Molecular Dynamics (MD) simulations have been a major source of dynamic structure. Although many methods were proposed for learning metastable states from MD data, some key problems are still in need of further investigation. Here, we give a brief review on recent progresses in this field, with an emphasis on some popular methods belonging to a two-step clustering framework, and hope to draw more researchers to contribute to this area.

Keywords: molecular dynamics simulation; metastable states; energy landscape

Citation: Jiang, H.; Fan, X. The Two-Step Clustering Approach for Metastable States Learning. *Int. J. Mol. Sci.* **2021**, *22*, 6576. https://doi.org/10.3390/ijms22126576

Academic Editor: Daniel Svozil

Received: 31 May 2021
Accepted: 14 June 2021
Published: 19 June 2021

1. Introduction

Proteins are basic building blocks of life, which carry out most essential functions in a cell such as catalysation, signal transduction, gene regulation, molecular modification, etc. These capabilities depend on their three-dimensional biomolecular structures, which also undergo reversible transitions between alternative structures (also called conformations). Different conformations have different Gibbs free energy. The free energy landscape of the conformational space is rugged with a number of high-energy barriers. These barriers partition the conformational space into a set of low-energy wells, which are called metastable states. See Figure 1 for an illustration. Conformations belonging to one metastable state do not easily change into conformations belonging to another metastable state. For more details on the free energy landscape of proteins, we refer to Finkelstein and Ptitsyn [1].

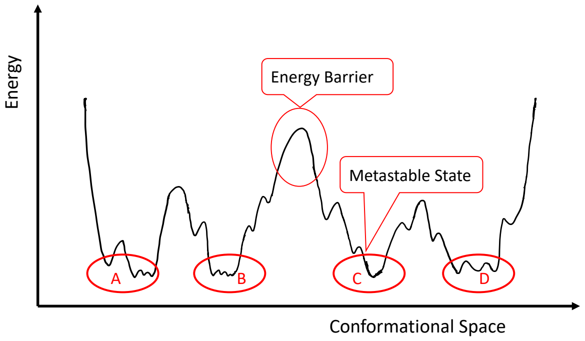

Figure 1. An illustration of the free energy landscape of a conformational space. There are four metastable states labeled as A, B, C and D. Conformations belonging to one metastable state, for example state B, do not easily change into conformations belonging to another metastable state, e.g., state C, due to the energy barrier between them.

The elucidation of the energy landscape and the conformational dynamics is crucial for understanding many biological processes, such as protein folding [2] and RNA folding [3], and for deciphering diseases related to improper conformational changes, such as Alzheimer's disease, Mad cow disease, Huntington's disease and Parkinson's disease [4]. Several experimental methods have been proposed to study stable structures or conformational changes, such as X-ray imaging [5], nuclear magnetic resonance [6], single-molecule fluorescence resonance energy transfer [7] and Cryo-electron [8]. Computational methods have also been proposed to predict protein structure from primary sequences, such as the generative probabilistic model by Boomsma et al. [9], the sequential Monte Carlo method by Wong et al. [10], critical assessment of protein structure prediction experiment [11–13] and deep neural network methods [14–17], including Google's AlphaFold [18]. However, these methods focus mainly on the protein-folding problem [19], which aims at the conformation of the lowest free energy. They cannot provide global dynamic information on conformational changes at the atomic level.

Molecular Dynamics (MD) simulations [20], which simulate conformational trajectories, have emerged and are the major source of global dynamic information at the atomic level. More specifically, MD simulations sample from a conformational space by evolving the structure based on Newton's equations of motion. Each evolution produces a trajectory formed by a sequence of conformations at times $t = 0, \tau, 2\tau, \cdots, n\tau$, where τ denotes the observation interval. To handle the rugged energy landscape as shown in Figure 1, generalized ensemble algorithms, such as multicanonical algorithm [21] and Replica Exchange [22], are used in MD simulations to generate a wider sampling by helping the simulation trajectories pass through energy barriers with a higher probability and avoid trapping in local modes [23].

Due to the high computational cost of MD simulations, the timescale of MD trajectories is usually shorter than the typical real conformational dynamics. To bridge the timescale gap, Markov state models (MSMs) [24–30] were commonly used to reproduce the long-time conformational dynamics of biomolecules using MD data, see for example, Chodera and Noé [31], Wang et al. [32], and Husic and Pande [30] for a review on the status of MSMs studies. Based on MSMs, current methods for identifying metastable states from MD data mostly take a two-step clustering approach. In this review of methods for learning metastable states from MD data, we provide a detailed discussion on this two-step clustering framework, check some popular methods within this framework as well as some initiatives beyond this framework. We hope this brief review would draw more researchers to break through this two-step clustering framework for better detection of metastable states.

2. Learning Metastable States from MD Data

Statistically, learning metastable states from MD data estimates the distribution of conformations over the structural space. Given the molecular data (trajectories of conformations as shown in Figure 2A) from MD simulations, we want to estimate the density function $f(x) = \sum_{i=1}^{k} q_i f_i(x|A_i)$, where $\{A_i : i = 1, 2, \cdots, k\}$ is a disjoint partition of the conformation space Ω, i.e., $A_i \cap A_j = \emptyset$ and $\cup_{i=1}^{k} A_i = \Omega$. A_i corresponds to the basins or metastable states of the energy landscape, q_i is the probability of conformation x belonging to basin A_i, and k is the unknown number of metastable states. Note that $f(x)$ is a multi-mode density function. Specifically, it has k modes, and each $f_i(x)$ has one mode in its region A_i.

Taking the free energy landscape in Figure 1 for example, we may write the structural density function as $f(x) = \sum_{i=1}^{4} q_i f_i(x|A_i)$, with (A_1, A_2, A_3, A_4) obtained by partition the conformation space according to the energy barriers between four basins (A, B, C, D). The aim of MD data analysis is to recover basins (A, B, C, D) from data.

Before discussing the difference between estimating the structural density function and traditional density function, we shall emphasize the biological property behind the partition $\{A_i : i = 1, 2, \cdots, k\}$ in the structural density function. Specifically, conformations belonging to the same basin (partition) shall not only have geometrical similarity at key parts but also have dynamical similarity. However, global geometrical similarity in structural space may not necessarily lead to dynamical similarity due to energy barriers. In other words, two conformations exist such that they are geometrically similar, i.e., the geometrical distance between them is smaller than some threshold; however, we may rarely observe dynamical transitions between them along the trajectories.

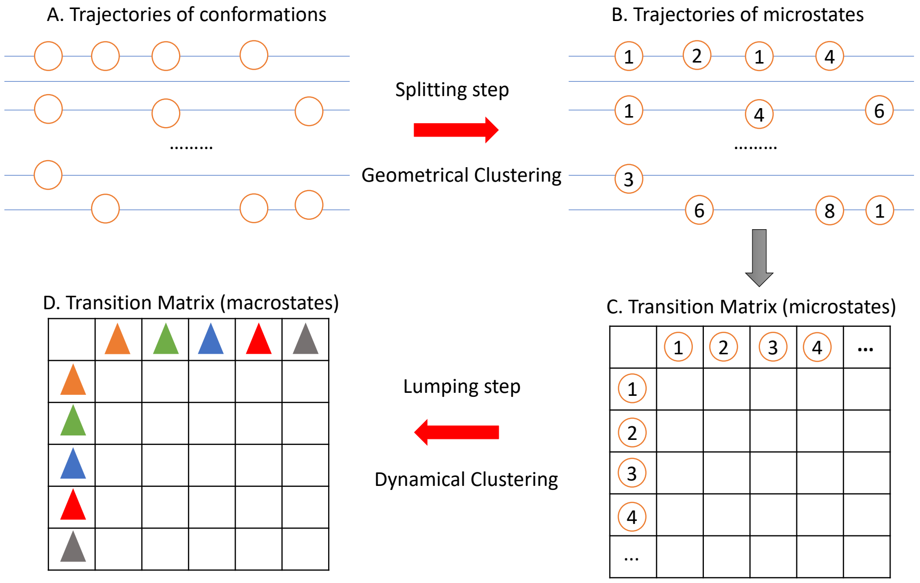
A. Trajectories of conformations
B. Trajectories of microstates
Splitting step
Geometrical Clustering
1 2 1 4
1 4 6
....
3
6 8 1
C. Transition Matrix (microstates)
1 2 3 4 ...
D. Transition Matrix (macrostates)
Lumping step
Dynamical Clustering

Figure 2. Workflow of the two-step clustering framework for learning metastable states. (**A**) Trajectories of conformations obtained from MD simulations. Each circle represents a different conformation. (**B**) Trajectories of microstates resulted from the splitting step. This step uses only the geometrical information to cluster conformations with high geometrical similarity into a microstate. Circles of the same number represent the conformations belonging to a same microstate. (**C**) The transition matrix between microstates, which counts the number of jumps between them along the trajectories. (**D**) Transition matrix between macrostates obtained from the lumping step by clustering microstates into macrostates. Each macrostate is a collection of microstates. Solid triangles with different colors represent different macrostates.

In the framework of traditional density estimation, the key is to get the best estimation of the density, i.e., q_i and $f_i(x)$. In other words, global geometrical similarity is the only concern in traditional density estimation, and the special biological property is ignored. Thus, Bayesian sequential partition [33], which extends the idea of classification tree [34] for estimating a high-dimensional density function, can not be applied here. In the framework of Markov state model for learning metastable states, one estimates directly the partition $\{A_i : i = 1, 2, \cdots, k\}$ and ignores q_i and $f_i(x)$ because they are unimportant to metastable states.

In summary, the difficulty underlying estimating the structural density function is how to recover the partition $\{A_i : i = 1, 2, \cdots, k\}$ and satisfy the biological property that conformations belonging to the same partition have both the geometrical and dynamical similarity. This difficulty increases with the complexity of the molecule under study. Note that there are two important parts for learning metastable states: the partition of the conformational space and the number of metastable states k. The metastable structure in each partition is defined as the conformation with lowest free energy.

3. The Two-Step Clustering Framework

A two-step clustering approach is widely used for identifying metastable states from MD data. This is due to the fact that the conformational space where MD simulations sample from is essentially a high-dimensional, continuous coordinate space. Therefore, even if the raw simulation trajectories may contain thousands of conformations, very few transitions between any specific pair of these conformations will be observed. To overcome this sparsity problem at the conformation level, a two-step clustering framework, a clever idea, is commonly adopted for analyzing the trajectories of conformations. See Figure 2 for an overview of this two-step clustering framework.

Firstly, a splitting step is used to group conformations into a number of microstates according to their structural (geometric) similarity. In this step, we introduce a new concept, called microstate, which is defined as a set of conformations with high geometrical similarity. Actually, conformations belonging to the same microstates are assumed to have both geometrical similarity and dynamical (kinetic) similarity, which ensures the fast converting among them. It is expected to observe more transitions between microstates than between conformations; thus, we can hopefully get a statistically stable transition matrix between microstates. Since this step uses only the geometric information of conformations, we refer it to as the geometric clustering step.

Secondly, a lumping step is used to cluster further microstates into macrostates (also called metastable states) based on the transition matrix between microstates. Thus, a macrostate (metastable state) is as a set of microstates with high dynamical similarity. This step depends on the dynamic information between microstates; thus, it is referred to as the dynamical clustering step.

To have relatively stable jumps between microstates, the number of microstates should be selected carefully. If it is too large, the transition frequency between microstates will be very low. If it is too small, a microstate may contain conformations that are separated by energy barriers. Both situations will prevent the detection of true metastable states. In addition, ignoring the geometrical information in the lumping step is problematic and gives undesired results. In the following, we dive into details of this two-step clustering framework.

3.1. The Splitting Step: Geometrical Clustering

The splitting step corresponds to the transition from Figure 2A to Figure 2B. The input of this step is the vector data in R^m of n samples, where n is the total number of conformations in all trajectories, and m is the dimension of the molecule, depending on the pre-processing of the MD data. A conformation can be represented by the coordinates of all atoms or its torsion angles. Thus, the dimension m of these two different representations may be different. K-means [35] and K-medoids [36] algorithms are widely used in this step due to their easy implementation.

To improve the efficiency of these two algorithms, dimension reduction methods such as principle component analysis are applied before geometrical clustering. The principle components (PCs) of coordinates and that of torsion angles are commonly used as the representation of a conformation, see for example Mu et al. [37], Altis et al. [38]. For more information about principle component analysis (PCA) of molecular dynamics, we refer to Sittel et al. [39], where a detailed comparison of PCA on the use of Cartesian and internal coordinates is given.

The main concerns about K-means/K-medoids algorithms are as follows. Firstly, both of them give a local optimum instead of a global optimum due to the large sample size. This means there may be some conformations belonging to the same microstates that are clustered into different microstates, and thus leads to bad basins. Researchers usually try to run the K-means or K-medoids algorithm multiple times to get better results. Second, the aim of K-means/K-medoids algorithms, essentially, is to obtain an η-cover of the vector space with centers $\{\mathbb{P}_1, \mathbb{P}_2, ...\mathbb{P}_k\}$ such that for each conformation $\mathbb{P}$, there is an $\mathbb{P}_i$, such that $d(\mathbb{P}, \mathbb{P}_i) \leq \eta$, where $d(\cdot, \cdot)$ is a distance function defined for any two vectors, and η can be understood as the similarity threshold for defining microstates. It is impossible for us to find a suitable η and k in real applications, which implies the inevitable difficulty of K-means/K-medoids algorithms in the splitting step to give satisfactory microstates.

In principle, any clustering algorithm (see Jain [40] for a review) taking vector data as input can be used for geometrical clustering. The computational burden and the quality of resultant microstates are the main concerns in this step. In addition, there are methods proposed to improve the quality of microstates from this step. We will discuss them in Section 3.3.

3.2. The Lumping Step: Dynamical Clustering

The lumping step corresponds to the transition from Figure 2C to Figure 2D. The input of this step is the transition matrix between microstates obtained from the splitting step. There are many different strategies for this dynamical clustering. We introduce below the representative ones and discuss their performance on an MD alanine dipeptide dataset, a well-understood molecule with six metastable states. The reason for choosing this historic dataset is that we know the ground truth of its metastable states, which is crucial for us to understand the performance of each method.

The MD trajectory data are taken from Chodera et al. [41], which consists of 974 20-ps NVE simulations with conformations stored every 0.1 ps, and there are 194,800 conformations in the dataset. The detailed simulation information can be found in Chodera et al. [41]. The conformation space of the alanine dipeptide can be represented by two torsion angles ϕ and ψ [41]; thus, it is a simple molecule often taken as a benchmark. Figure 3 shows the scatter plot of ϕ-ψ of these 194,800 conformations with transition matrixes between its six metastable states, shown in Table 1, where the partition of the conformational space into six clusters (metastable states) follows that given in Chodera et al. [41]. Note that these six metastable states are given by a manual partition according to the estimated landscape from the parallel tempering simulation [41]. To eliminate the impact of microstates from the splitting step on the lumping step, the microstates of the alanine dipeptide are obtained by a grid method that partitions the ϕ-ψ space into 80 × 80 grids and takes each non-empty grid as a microstate.

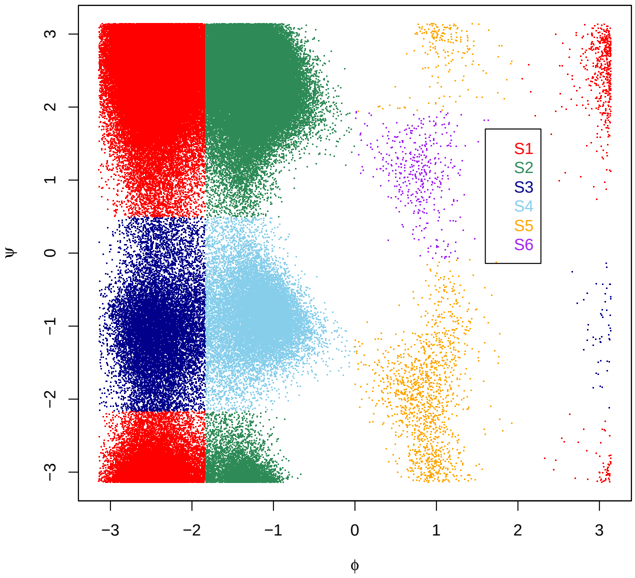

Figure 3. The scatter plot of ϕ-ψ of the alanine dipeptide with $\phi, \psi \in [-\pi, \pi]$. The partition of ϕ-ψ space into six clusters follows that given in Chodera et al. [41].

Table 1. Transition matrix of the benchmark clusters of the alanine dipeptide with S1–S6 shown in Figure 3.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**
S1	0.9457	0.0477	0.0062	0.0004	0.0000	0.0000
S2	0.0609	0.9365	0.0004	0.0021	0.0000	0.0002
S3	0.0403	0.0021	0.8939	0.0636	0.0000	0.0000
S4	0.0020	0.0090	0.0526	0.9356	0.0008	0.0000
S5	0.0013	0.0013	0.0000	0.0098	0.9718	0.0158
S6	0.0000	0.0401	0.0000	0.0000	0.0519	0.9080

Sum of diagonals: 5.591479
Mean of diagonals: 0.9319131
Minimal of diagonals: 0.8939

Perron Cluster Cluster Analysis (PCCA, Deuflhard et al. [42]) and Its Variants. PCCA is based on two important observations of transition matrix P between microstates: (P1) if P has an s block-diagonal structure, its eigenvalue $\lambda = 1$ is s-fold, which is used to identify the number of macrostates; (P2) the sign structure of the eigenvector corresponds to the assignment of macrostates. Thus, the idea of PCCA is mathematically solid and easy to implement. In practice, one should input the number of clusters (metastable states) generated from PCCA, which is difficult to estimate for real applications.

In PCCA, the true transition matrix between microstates is assumed to be block-diagonal, i.e., $D = \mathrm{diag}(D_{11}, D_{22}, \cdots, D_{kk})$, where k is the number of macrostates, and the observed transition matrix $P = D + E$, where E is the perturbation matrix representing error of the observations. This assumption may fail sometimes. Consider a special case where microstates are macrostates, we find that the true transition matrix can not be block-diagonal (see Table 1 for example). Secondly, PCCA can be understood as finding a macrostate assignment, based on property (P1) and (P2), by maximizing the sum of diagonals of the transition matrix between macrostates, i.e., the metastability of macrostates. However, this kind of object may not directly lead to macrostates with the biological property that conformations belonging to the same partition have both the geometrical and dynamical similarity. That is, although conformations within the same macrostates obtained by PCCA have high dynamical similarity, the geometric similarity between conformations belonging to the same macrostates are not guaranteed. This is partly due to its ignorance of geometric information when clustering microstates into macrostates. To make this point clear, we show in Figure 4A the clustering results of alanine dipeptide from PCCA by setting the number of clusters (metastable states) as its true number of clusters 6, which is very different from the reference clustering labels shown in Figure 3. Typically, metastable states S3 and S4 in Figure 3 are recognized as one metastable state in Figure 4A. However, PCCA gives satisfied results according to the corresponding transition matrix given in Table 2, which has a sum of diagonals close to that of the true transition matrix in Table 1. These facts together imply that the mathematical optimal solution provided by PCCA may not be biologically meaningful.

Different versions of PCCA, such as PCCA+ [43] and Flux PCCA (FPCCA, Beauchamp et al. [44]), are proposed to improve its robustness to random perturbations. PCCA+ needs users to give the range of number of clusters (metastable states). Figure 4B,C shows the clustering results of the alanine dipeptide from PCCA+ with different ranges, and the results are undesired.

Gibbs Sampling Algorithm (GSA, Wang et al. [45]). GSA is based on a Poisson model assuming that the observed number of jumps between macrostates follows a Poisson distribution. Taking the transition matrix between macrostates as a parameter, given this Poisson model, we have the likelihood function for the macrostates assignment of microstates. The transition matrix and the macrostates assignment of microstates is learned by maximizing the likelihood function. To the best of our knowledge, this is the first attempt on statistical modeling of the transition matrix between macrostates. The major point of GSA is to take the unknown number of macrostates as input. Figure 4D–F shows the clustering results of alanine dipeptide from GSA with different numbers of clusters. When users specify the right number of clusters, as shown in Figure 4D, GSA gives quite good results. However, when users specify a bad one, the results will be bad, as shown in Figure 4E,F.

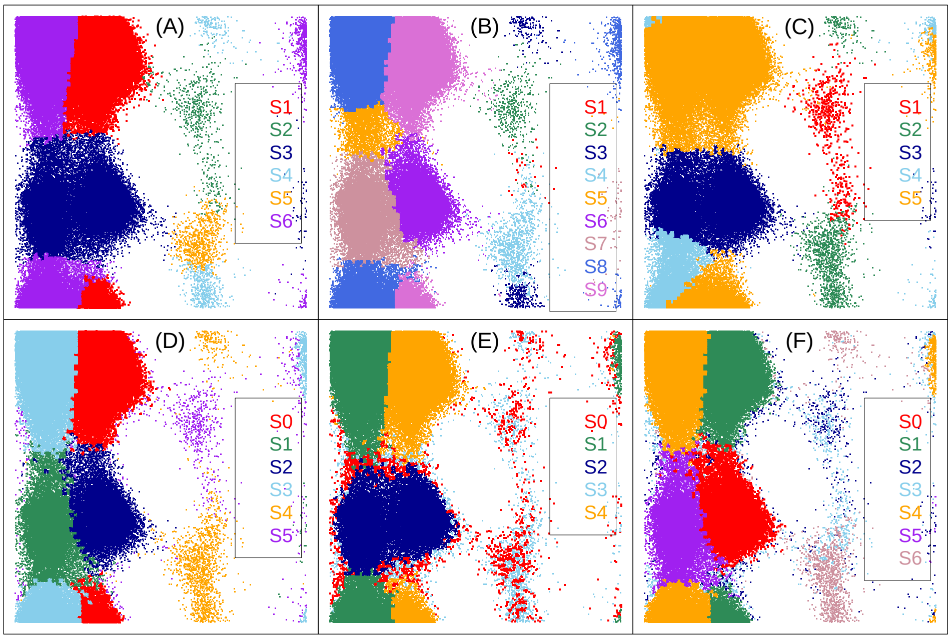

Figure 4. Clustering results of the alanine dipeptide from PCCA, PCCA+ and Gib algorithms. The axis is the same as that in Figure 3. (**A**) PCCA with 6 clusters; (**B**) PCCA+ with estimated number of cluster belonging to [3, 9]; (**C**) PCCA+ with estimated number of cluster belonging to [5, 7]; (**D**) GSA with 6 clusters; (**E**) GSA with 5 clusters; (**F**) GSA with 7 clusters.

Table 2. Transition matrix between clusters of the alanine dipeptide obtained by PCCA with S1–S6 shown in Figure 4A.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**
S1	0.9352	0.0003	0.0018	0.0000	0.0000	0.0626
S2	0.0477	0.9131	0.0000	0.0068	0.0324	0.0000
S3	0.0042	0.0000	0.9752	0.0000	0.0004	0.0202
S4	0.0000	0.0032	0.0000	0.9104	0.0816	0.0048
S5	0.0000	0.0269	0.0175	0.0672	0.8884	0.0000
S6	0.0508	0.0000	0.0068	0.0000	0.0000	0.9424

Sum of diagonals: 5.564797
Mean of diagonals: 0.9274662
Minimal of diagonals: 0.8884

Most Probable Pathway (MPP, Jain and Stock [46]). The idea underlying MMP is straightforward: it merges the microstate with its neighboring microstates on its most probable pathway that has the lowest free energy. The merit of MPP is that it does not require an estimated value or range on number of clusters (metastable states) as input. However, due to the discreteness of MD data, there exits undesired cases; for a microstate belonging to state A, we may observe its most probable pathway leads to a microstate with the lowest free energy belonging to state B. According to the principle of MPP, we should merge this microstate into state B, which is undesired. To make this point clear, we show in Figure 5A the clustering results of alanine dipeptide from MPP. As shown in the figure, some microstates belonging to state S4 are clustered wrongly into states S1 and S3.

Minimum Variance Clustering Approach (MVCA, [47]). Husic et al. [47] considered a different strategy in the second step by using symmetric Jensen–Shannon divergence to measure the similarity between microstates and Ward's minimum variance criterion to do the agglomerative clustering. Following this idea, one may use other distance metrics to measure the similarity between microstates, and use agglomerative clustering to cluster microstates into macrostates. We show in Figure 5B the clustering results of the alanine dipeptide from MVCA.

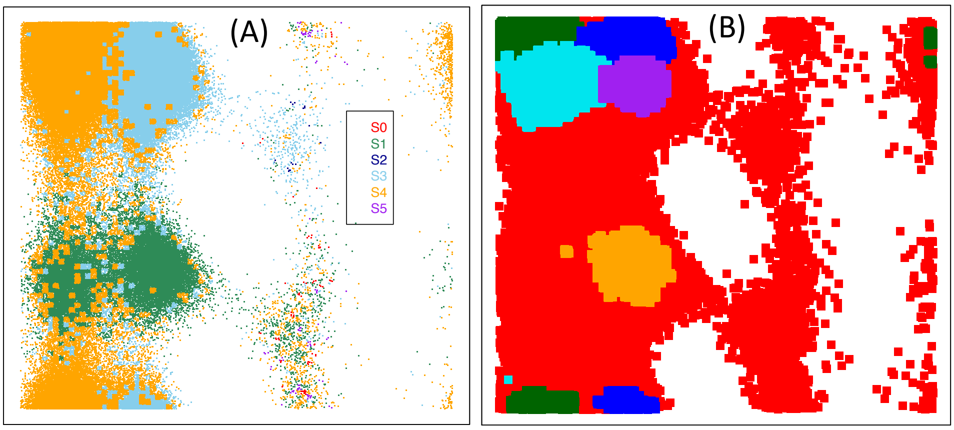

Figure 5. Clustering results of the alanine dipeptide from MPP (**A**) and MVCA (**B**). The axis is the same as that in Figure 3. Different colors in the figure present different clusters.

In summary, we discussed the performance of different strategies for the lumping step by providing the same (and almost ideal) microstates to different methods. That is, the performance of each method only depends on the strategy for the lumping step. According to results shown in Figures 4 and 5, the performance PCCA(+), MPP, MVCA and GSA depends strongly on the number of macrostates from the algorithm. Importantly, the scatterplots from PCCA(+), MPP and MVCA are different from the ground truth shown in Figure 3, although their underlying principle is well-understood. This may be partially caused by the fact that the lumping step uses only the dynamical information between microstates but ignores the geometrical information between them. GSA shows good performance when microstates are well defined and the number of macrostates is correctly specified, which are quite difficult as discussed before.

3.3. Refinements to The Framework

Researchers have noticed two shortcomings for the above two-step procedure [25,48,49] when applied to complex systems: (1) the quality of microstates is not guaranteed, which has a strong impact on the downstream analysis; (2) some poorly sampled states may dominate the coarse-grained model. In the following, we discuss a few strategies to refine the two-step framework.

Iterative framework to improve microstates. An iterative framework [48,49] was proposed to obtain better microstates. It goes as follows: (a) splitting macrostates by geometrical clustering using stepwise K-means algorithm by incorporating other information, such as escape probability; (b) lumping microstates into macrostates by dynamical clustering using PCCA, PCCA+ or any other method; (c) repeating (a–b) until it converges. Note that there is only one macrostate in the fist iteration. Essentially, this kind of iterative algorithm is to improve the quality of microstates. In other words, if we have another way to get better microstates, this iterative framework may not be helpful in basin estimation. This iterative method may improve the quality of microstates, but it still can not make sure the resultant microstates are good enough.

Other methods for the splitting step. Ignoring the dynamic information between conformations, geometric clustering is just a clustering problem with vectors as input. Based on this observation, all other methods for vector clustering [40] are applicable to the current problem. For example, Sittel and Stock [50] and Liu et al. [51] proposed a density-based clustering [52,53] method to cluster conformations into microsates and then used MPP or PCCA to further cluster microstates into macrostates. Taking a density-based clustering method in the splitting step avoids the difficulty of local convergence from K-means or K-medoids algorithms. However, selecting the bandwidth, the key parameter determining the number of clusters, for density-based clustering is still in need of further investigation. Furthermore, dimension reduction methods other than PCA are suggested to extract important coordinates before going to the splitting step, see Sittel and Stock [54] for a discussion.

Reducing the impact of poorly sampled microstates. Bayesian agglomerative clustering engine (BACE, Bowman [55] and Hierarchical Nyström expansion graph (HNEG, Yao et al. [56] are proposed to amend the shortcoming of PCCA and PCCA+, that they tend to identify poorly sampled states as being kinetically distinct from their neighbors [25]. Specifically, BACE identifies coarse-grained states by finding sets of states that have the same kinetics (i.e., transition probabilities to other states) within statistical uncertainty through Bayes factor. See Bowman [55] for details. HNEG attempts to solve this problem by placing more emphasis on well-sampled states than poorly sampled ones based on the idea that the whole transition matrix can be approximated by a sub-stable transition matrix between well-sampled states. The metastable macrostates are obtained by applying PCCA (PCCA+) to the sub-stable transition matrix. In other words, HNEG is an improved version of PCCA and PCCA+. We refer to Bowman et al. [57] for a review and a comparison on the performance of some of these methods, where the authors pointed out PCCA (PCCA+) has a similar performance with BACE and HNEG, but it is better than MPP.

For complex systems, the shortcomings of the lumping step discussed before still exist. The additional difficulty comes from the splitting step, which, as discussed in Section 3.1, is from the following two fundamental facts: (1) K-means/K-medoids algorithms are very difficult to converge, and (2) the number of microstates from them is also very difficult to determine. Although different methods are proposed to improve the quality of microstates, we still do not know whether they are good enough for downstream analysis on complex systems, as these methods work in an intuitive way. Thus, we expect new ideas to overcome these limitations. The optimal reaction coordinates [58] that treat the free energy as a function of reaction coordinates is a good example on this direction.

4. Some Extensions

In previous sections, we gave a brief review on methods under the two-step clustering framework for learning metastable states. Here, we discuss some extensions for modeling MD data beyond MSMs.

Deep neural networks (DNN) for learning molecular dynamics. DNN has rapidly developed in recent years due to its successful application to image processing. See LeCun et al. [59] for a review. Researchers are motivated to apply DNN to other areas including predicting protein folding and exploring the landscape of proteins. For example, Wu et al. [60] and Mardt et al. [17] proposed a deep generative Markov state model based on deep neural network to learn molecular dynamics and sample conformations from conformation space. The key elements in their DNN models are (1) a DNN encoding the coordinates information into latent space, (2) a Markov transition model between elements in latent space and (3) a generative model decoding from latent space to coordinates information. These DNN models are promising. However, to train the DNN, we should know the number of macrostates first, which is unknown to us. How to learn automatically the number of macrostates from MD data is still open.

DNN is a powerful tool for many problems, especially for image processing. However, a lack of explanation of results from deep learning is the key point that hinders its application on other areas, for example, biology. It is well known that obtaining MD data is time-consuming, and it is helpful to get it from deep learning, which is potential direction for future works. For more discussions on machine learning methods for MD data analysis, we refer to Noé [61].

Improvements on MSMs. Markov state models assume a Markov chain on a discretization of the state space. However, it is difficult to apply to high-dimensional biomolecular systems. The quality and reproducibility of MSMs are therefore limited. Differently, projected Markov Models (PMMs, Noé et al. [62]) only assume that the full phase space molecular dynamics is Markovian, and a projection of this full dynamics is observed on the discrete states. However, estimating PMMs is very difficult. In addition, Dynamic Graphical Models (DGMs, Olsson and Noé [63]) are proposed to deal with the case where the size of global metastable states grow exponentially with the system size. Similar to how spins interact in the Ising model, DGMs describe molecules as assemblies of coupled subsystems, and the change of each subsystem state is only governed by the states of itself and its neighbors. We refer to their original paper for more details about PMMs and DGMs.

5. Discussion and Outlook

In this paper, we reviewed some popular methods for learning metastable states from molecular dynamics data, and most of them belong to a two-step clustering framework including a splitting step and a lumping step. The performance of popular methods is illustrated based on MD data of the alanine dipeptide.

In the splitting step, one wants to obtain microstates by clustering conformations into microstates while ignoring the dynamical information underlying them. K-means and K-medoids are commonly used in this step. However, they suffer from drawbacks such as being stuck in a local mode and not easily obtaining microstates with biological properties that conformations belonging to the same partition have both the geometrical and dynamical similarity. Density-based clustering is used to avoid the drawback of these two methods, but it introduces a new difficulty on selecting the threshold to define the local density. Microstates from the splitting step have a strong impact on the downstream analysis. Bad microstates may lead to bad metastable states. For simple systems such as alanine dipeptide, a grid clustering method is available for geometrical clustering to get better microstates. However, for proteins such as HP35 NLE/NLE, methods such as MPP will inevitably give poor results due to bad microstates.

In the lumping step, there are many methods proposed for dynamical clustering. Each method has its own philosophy. PCCA has solid mathematical foundation, the idea behind MPP is straightforward, and GSA is based on a Poisson model. The key problem underlying them is their failures on estimating the number of metastable states; however, PCCA and GSA should take this unknown number as input. In addition, they strongly rely on the microstates from the splitting step. Bad microstates inevitably lead to bad macrostates no matter what method is used in the lumping step.

The key features of this two-step framework are (1) the quality of microstates has a strong impact on the quality of macrostates and (2) a separate consideration of geometric clustering and dynamical clustering. That is, the geometric information of conformations is only used in geometric clustering, and dynamical clustering uses only the dynamic information. Although an iterative framework is proposed to refine the microstates, we still can not make sure the conformations belong to the same microstates from the splitting step with the biological property. Thus, further investigation should focus on how to obtain high-quality microstates and how to combine the geometric and dynamic information to learn the metastable states.

Deep neural networks (DNN) are proposed to learn molecular dynamics from MD data, and thus learn the metastable states. However, it can not learn the number of metastable states. Instead, it should take this number as input. In other words, how to design a DNN for learning automatically the number of metastable states is still open. How to verify the DNN learned from MD data also needs further study.

Another interesting problem related to metastable states learning is to explore the relationship between local geometric similarity and dynamical similarity. We believe that the dynamic similarity between conformations comes from the local similarity in geometry. The advantage of DNN is to learn this relationship automatically from MD data, which is hidden from us and loses its interpretation. However, the two-step clustering framework ignores this point.

Finally, all of these methods do not take into consideration the statistical uncertainty of the MD data. There is great need of a full statistical model for MD data. GSA is a statistical model for the transition matrix between macrostates, and it has a good performance when we know the true numbers of macrostates and microsates are well defined. This fact sheds some light on statistical modeling on MD data. Based on the success of DNN, we should incorporate dimension reduction/variable selection into the statistical model for MD data, which would enable us to infer metastable states and the number of macrostates in a full statistical manner.

Author Contributions: Conceptualization, X.F.; methodology, H.J.; validation, H.J. and X.F.; formal analysis, H.J.; investigation, X.F.; resources, H.J.; data curation, H.J.; writing—original draft preparation, H.J.; writing—review and editing, H.J. and X.F.; visualization, H.J.; supervision, X.F.; project administration, X.F.; funding acquisition, H.J. and X.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by The National Natural Science Foundation of China (No. 11901517) and three grants from the Research Grants Council of the Hong Kong SAR, China (No. 400913, No. 14203915 and No. 14173817).

Acknowledgments: The authors would like to thank Guest Editor Daniel Svozil for the invitation, and thank the anonymous reviewers for their insightful suggestions to improve this paper.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Finkelstein, A.V.; Ptitsyn, O. *Protein Physics: A Course of Lectures*; Academic Press: Cambridge, MA, USA, 2002.
2. Schor, M.; Mey, A.S.; MacPhee, C.E. Analytical methods for structural ensembles and dynamics of intrinsically disordered proteins. *Biophys. Rev.* **2016**, *8*, 429–439. [CrossRef] [PubMed]
3. Sponer, J.; Bussi, G.; Krepl, M.; Banas, P.; Bottaro, S.; Cunha, R.A.; Gil-Ley, A.; Pinamonti, G.; Poblete, S.; Jurecka, P.; et al. RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chem. Rev.* **2018**, *118*, 4177–4338. [CrossRef] [PubMed]
4. Selkoe, D.J. Folding proteins in fatal ways. *Nature* **2003**, *426*, 900. [CrossRef] [PubMed]
5. Chapman, H.N.; Fromme, P.; Barty, A.; White, T.A.; Kirian, R.A.; Aquila, A.; Hunter, M.S.; Schulz, J.; DePonte, D.P.; Weierstall, U.; et al. Femtosecond X-ray protein nanocrystallography. *Nature* **2011**, *470*, 73. [CrossRef] [PubMed]
6. Kabsch, W.; Rösch, P. Nuclear magnetic resonance: Protein structure determination. *Nature* **1986**, *321*, 469. [CrossRef] [PubMed]
7. Ha, T. Single-molecule fluorescence resonance energy transfer. *Methods* **2001**, *25*, 78–86. [CrossRef]
8. Carroni, M.; Saibil, H.R. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods* **2016**, *95*, 78–85. [CrossRef]
9. Boomsma, W.; Mardia, K.V.; Taylor, C.C.; Ferkinghoff-Borg, J.; Krogh, A.; Hamelryck, T. A generative, probabilistic model of local protein structure. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 8932–8937. [CrossRef] [PubMed]
10. Wong, S.W.; Liu, J.S.; Kou, S. Exploring the conformational space for protein folding with sequential Monte Carlo. *Ann. Appl. Stat.* **2018**, *12*, 1628–1654. [CrossRef]
11. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Rost, B.; Hubbard, T.; Tramontano, A. Critical assessment of methods of protein structure prediction—Round VII. *Proteins Struct. Funct. Bioinform.* **2007**, *69*, 3–9. [CrossRef]
12. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Rost, B.; Tramontano, A. Critical assessment of methods of protein structure prediction—Round VIII. *Proteins Struct. Funct. Bioinform.* **2009**, *77*, 1–4. [CrossRef] [PubMed]
13. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1011–1020. [CrossRef]
14. Lena, P.D.; Nagata, K.; Baldi, P.F. Deep spatio-temporal architectures and learning for protein structure prediction. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2012; pp. 512–520.
15. Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **2017**, *13*, e1005324. [CrossRef] [PubMed]
16. Hou, J.; Adhikari, B.; Cheng, J. DeepSF: Deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics* **2017**, *34*, 1295–1303. [CrossRef]

17. Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5. [CrossRef] [PubMed]
18. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **2019**. [CrossRef] [PubMed]
19. Dill, K.A.; MacCallum, J.L. The protein-folding problem, 50 years on. *Science* **2012**, *338*, 1042–1046. [CrossRef] [PubMed]
20. Karplus, M.; McCammon, J.A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Mol. Biol.* **2002**, *9*, 646. [CrossRef] [PubMed]
21. Berg, B.A.; Neuhaus, T. Multicanonical algorithms for first order phase transitions. *Phys. Lett. B* **1991**, *267*, 249–253. [CrossRef]
22. Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151. [CrossRef]
23. Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Pept. Sci. Orig. Res. Biomol.* **2001**, *60*, 96–123. [CrossRef]
24. Bowman, G.R.; Huang, X.; Pande, V.S. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* **2009**, *49*, 197–201. [CrossRef] [PubMed]
25. Huang, X.; Yao, Y.; Bowman, G.R.; Sun, J.; Guibas, L.J.; Carlsson, G.; Pande, V.S. Constructing multi-resolution Markov state models (MSMs) to elucidate RNA hairpin folding mechanisms. In *Biocomputing 2010*; World Scientific: Singapore, 2010; pp. 228–239.
26. Lane, T.J.; Bowman, G.R.; Beauchamp, K.; Voelz, V.A.; Pande, V.S. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413–18419. [CrossRef]
27. McGibbon, R.T.; Pande, V.S. Learning kinetic distance metrics for Markov state models of protein conformational dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 2900–2906. [CrossRef]
28. Schwantes, C.R.; McGibbon, R.T.; Pande, V.S. Perspective: Markov models for long-timescale biomolecular dynamics. *J. Chem. Phys.* **2014**, *141*, 090901. [CrossRef] [PubMed]
29. Nüske, F.; Wu, H.; Prinz, J.H.; Wehmeyer, C.; Clementi, C.; Noé, F. Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias. *J. Chem. Phys.* **2017**, *146*, 094104. [CrossRef]
30. Husic, B.E.; Pande, V.S. Markov state models: From an art to a science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396. [CrossRef]
31. Chodera, J.D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135–144. [CrossRef]
32. Wang, W.; Cao, S.; Zhu, L.; Huang, X. Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1343. [CrossRef]
33. Lu, L.; Jiang, H.; Wong, W.H. Multivariate density estimation by Bayesian sequential partitioning. *J. Am. Stat. Assoc.* **2013**, *108*, 1402–1410. [CrossRef]
34. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; Routledge: London, UK, 1984.
35. Vassilvitskii, S.; Arthur, D. k-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7–9 January 2007; pp. 1027–1035.
36. Reynolds, A.P.; Richards, G.; Rayward-Smith, V.J. The application of k-medoids and pam to the clustering of rules. In Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning, Exeter, UK, 25–27 August 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 173–178.
37. Mu, Y.; Nguyen, P.H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins Struct. Funct. Bioinform.* **2005**, *58*, 45–52. [CrossRef]
38. Altis, A.; Nguyen, P.H.; Hegger, R.; Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **2007**, *126*, 244111. [CrossRef]
39. Sittel, F.; Jain, A.; Stock, G. Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *J. Chem. Phys.* **2014**, *141*, 07B605_1. [CrossRef]
40. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
41. Chodera, J.D.; Swope, W.C.; Pitera, J.W.; Dill, K.A. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.* **2006**, *5*, 1214–1226. [CrossRef]
42. Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of almost invant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Its Appl.* **2000**, *315*, 39–59. [CrossRef]
43. Deuflhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Its Appl.* **2005**, *398*, 161–184. [CrossRef]
44. Beauchamp, K.A.; McGibbon, R.; Lin, Y.S.; Pande, V.S. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 17807–17813. [CrossRef]
45. Wang, W.; Liang, T.; Sheong, F.K.; Fan, X.; Huang, X. An efficient Bayesian kinetic lumping algorithm to identify metastable conformational states via Gibbs sampling. *J. Chem. Phys.* **2018**, *149*, 072337. [CrossRef]
46. Jain, A.; Stock, G. Identifying metastable states of folding proteins. *J. Chem. Theory Comput.* **2012**, *8*, 3810–3819. [CrossRef] [PubMed]
47. Husic, B.E.; McKiernan, K.A.; Wayment-Steele, H.K.; Sultan, M.M.; Pande, V.S. A minimum variance clustering approach produces robust and interpretable coarse-grained models. *J. Chem. Theory Comput.* **2018**, *14*, 1071–1082. [CrossRef]

48. Chodera, J.D.; Singhal, N.; Pande, V.S.; Dill, K.A.; Swope, W.C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **2007**, *126*, 155101. [CrossRef]
49. Sheong, F.K.; Silva, D.A.; Meng, L.; Zhao, Y.; Huang, X. Automatic state partitioning for multibody systems (APM): An efficient algorithm for constructing Markov state models to elucidate conformational dynamics of multibody systems. *J. Chem. Theory Comput.* **2014**, *11*, 17–27. [CrossRef] [PubMed]
50. Sittel, F.; Stock, G. Robust density-based clustering to identify metastable conformational states of proteins. *J. Chem. Theory Comput.* **2016**, *12*, 2426–2435. [CrossRef] [PubMed]
51. Liu, S.; Zhu, L.; Sheong, F.K.; Wang, W.; Huang, X. Adaptive partitioning by local density-peaks: An efficient density-based clustering algorithm for analyzing molecular dynamics trajectories. *J. Comput. Chem.* **2017**, *38*, 152–160. [CrossRef]
52. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*; KDD: Portland, OR, USA, 1996; Volume 96, pp. 226–231.
53. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496. [CrossRef] [PubMed]
54. Sittel, F.; Stock, G. Perspective: Identification of collective variables and metastable states of protein dynamics. *J. Chem. Phys.* **2018**, *149*, 150901. [CrossRef]
55. Bowman, G.R. Improved coarse-graining of Markov state models via explicit consideration of statistical uncertainty. *J. Chem. Phys.* **2012**, *137*, 134111. [CrossRef] [PubMed]
56. Yao, Y.; Cui, R.Z.; Bowman, G.R.; Silva, D.A.; Sun, J.; Huang, X. Hierarchical Nyström methods for constructing Markov state models for conformational dynamics. *J. Chem. Phys.* **2013**, *138*, 174106. [CrossRef] [PubMed]
57. Bowman, G.R.; Meng, L.; Huang, X. Quantitative comparison of alternative methods for coarse-graining biological networks. *J. Chem. Phys.* **2013**, *139*, 121905. [CrossRef]
58. Krivov, S.V. Protein Folding Free Energy Landscape along the Committor-the Optimal Folding Coordinate. *J. Chem. Theory Comput.* **2018**, *14*, 3418–3427. [CrossRef]
59. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
60. Wu, H.; Mardt, A.; Pasquali, L.; Noe, F. Deep generative Markov state models. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 3975–3984.
61. Noé, F. Machine Learning for Molecular Dynamics on Long Timescales. *arXiv* **2018**, arXiv:1812.07669.
62. Noé, F.; Wu, H.; Prinz, J.H.; Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.* **2013**, *139*, 11B609_1. [CrossRef] [PubMed]
63. Olsson, S.; Noé, F. Dynamic graphical models of molecular kinetics. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 15001–15006. [CrossRef] [PubMed]