pharmaceuticals

MDPI

Review

Data-Driven Molecular Dynamics: A Multifaceted Challenge

Mattia Bernetti [1], Martina Bertazzo [2,†] and Matteo Masetti [3,*]

1 Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Bonomea 265, I-34136 Trieste, Italy; mbernett@sissa.it
2 Computational Sciences, Istituto Italiano di Tecnologia, via Morego 30, I-16163 Genova, Italy; martina.bertazzo@iit.it
3 Department of Pharmacy and Biotechnology, Alma Mater Studiorum—Università di Bologna, via Belmeloro 6, I-40126 Bologna, Italy
* Correspondence: matteo.masetti4@unibo.it
† Current affiliation: Global Research Informatics/Computational Chemistry, Evotec (France) SAS, 31100 Toulouse, France.

Received: 25 August 2020; Accepted: 16 September 2020; Published: 18 September 2020

Abstract: The big data concept is currently revolutionizing several fields of science including drug discovery and development. While opening up new perspectives for better drug design and related strategies, big data analysis strongly challenges our current ability to manage and exploit an extraordinarily large and possibly diverse amount of information. The recent renewal of machine learning (ML)-based algorithms is key in providing the proper framework for addressing this issue. In this respect, the impact on the exploitation of molecular dynamics (MD) simulations, which have recently reached mainstream status in computational drug discovery, can be remarkable. Here, we review the recent progress in the use of ML methods coupled to biomolecular simulations with potentially relevant implications for drug design. Specifically, we show how different ML-based strategies can be applied to the outcome of MD simulations for gaining knowledge and enhancing sampling. Finally, we discuss how intrinsic limitations of MD in accurately modeling biomolecular systems can be alleviated by including information coming from experimental data.

Keywords: machine learning; dimensionality reduction; reaction coordinates; collective variables; Markov state models; maximum entropy principle; experimental data

1. Introduction

The idea of exploiting computers and information technology for assisting the drug discovery process goes back to the early 1960s when the pioneering works by Corwin Hansch and Toshio Fushita laid the ground for quantitative structure-activity/property relationship (QSAR/QSPR) models [1,2]. Since then, several computational approaches have flourished and, as soon as adequate computational resources have reached widespread availability, computer-aided drug discovery (CADD) has become a valuable asset for both academia and industry [3]. Nowadays, different kinds of computational methods are virtually employed in all realms of science and, in the field of medicinal chemistry, they are widely recognized as an integral part of any modern drug discovery endeavor [4]. Many of these approaches do rely on the availability of experimental data to draw hypotheses, identify patterns, and make inferences (inductive learning) [5]. Thanks to recent advances in experimental techniques, like high-throughput screening and array measurements, among others, we are today in the position of disclosing the full potential of these computational tools [6]. Nevertheless, mining and integrating large-scale datasets coming from different sources is far from trivial. This issue is a declination of

the well-known problem of big data, which refers to an overwhelming amount of information ("big" in terms of volume and diversity) that challenges the possibility of taking advantage of it [7]. It is a common perception that the current resurgence of artificial intelligence (AI) methods, including brand new techniques like deep learning (DL) algorithms, will ultimately provide a suitable framework to address this task [8,9].

Over the past decade, another class of computational methods has gained increasing popularity in the field of drug discovery and development. These are based on the application of fundamental theories to predict the properties of materials and can be ascribed to deductive learning [5]. Molecular dynamics (MD) is one of such approaches that stem from the possibility of simulating the temporal evolution of systems (in jargon, the trajectory) based on a microscopic description thereof [10]. Given that a suitable representation of intra- and inter-molecular forces is provided (the so-called force field) and thus interactions between biomolecules are properly modeled, MD allows investigating molecular-level mechanisms and extracting relevant observables related to the process under investigation. Protein–ligand (un)binding mechanisms (i.e., "dynamic docking"), binding free energies, and even kinetic constants are examples of useful outcomes for drug design when MD simulations are applied to pharmaceutically relevant systems [11–13]. It is therefore not surprising that MD methods are nowadays routinely employed in drug discovery for complementing experiments and orthogonal computational techniques like molecular docking and virtual screening [14,15]. Notwithstanding their descriptive and predictive power, MD historically suffered from two main drawbacks: the limited accuracy of force fields and the short length of simulations compared to the real physical time required to observe the investigated biological events. Thanks to the advancements in high-performance (HPC) and/or distributed computing, the second issue is alleviating, as the gap between experimental and computational timescales is constantly decreasing. The straight drawback, however, is that trajectories and output data, in general, are steadily growing in size, calling for adequate strategies for extracting relevant information when the big data regime approaches. Even from this standpoint, AI-based methods are optimally equipped to cope with such increased complexity. It is worth highlighting that machine learning (ML) has long been adopted by the MD community to analyze simulations of biomolecular systems [16]. Nevertheless, not only today are these approaches becoming routine for processing MD-derived data, but the growing awareness of their potential is also boosting their development for gathering insight in an automated fashion, for informing subsequent simulations, and even for analyzing and guiding the dynamics in a seamless way (see Figure 1, paths "1" and "2"). Prominent parallel processing strategies are also starting to be explored for dealing with large-scale MD data analysis [17]. These include efficient tools such as Hadoop, an open-source implementation of MapReduce [18], which, differently from classical HPC frameworks with dedicated storage nodes, are instead based on localized storage on the compute nodes. By leveraging on such architectures that facilitate access to data, the performance of analysis algorithms can be remarkably improved. While such applications to MD-generated data are still in their infancy, they have the potential to become remarkable instruments as hardware resources lead towards the big data domain. A detailed discussion of this topic is out of the scope of the present review and will thus not be covered.

The big data problem refers not only to the quantity and diversity of data but also to their rate of production and trustworthiness of their sources (also known as "veracity"). In this context, it is worth highlighting that a class of emerging analysis methods is explicitly concerned with the combination of MD-derived data and experimental information to deal with the above-mentioned force field issue. Specifically, such strategies aim at reducing systematic inaccuracies due to the limitations of the force field used to model the biomolecular systems [19]. Albeit not strictly inside the ML domain, these approaches represent an innovative way of making sense from MD trajectories or guiding new simulations by enforcing the agreement with available experimental knowledge (see Figure 1, path "3").

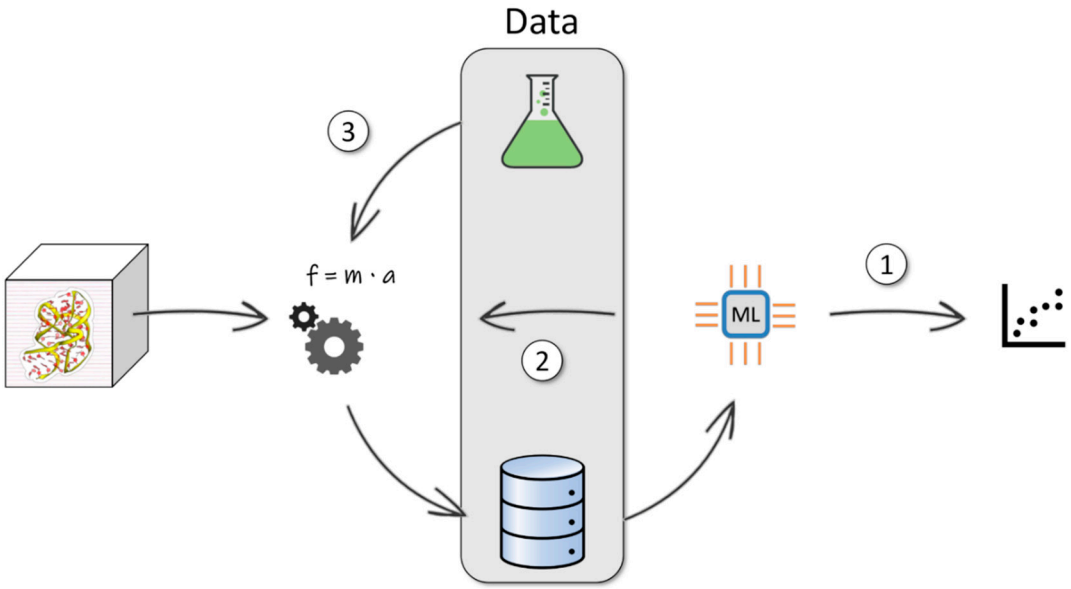

Figure 1. Pictorial representation of data exploitation in molecular dynamics (MD) simulations. Note that the source of data can be either computational (the very output of MD simulations, paths "1" and "2") or experimental (path "3"). Path "1" refers to the use of machine learning (ML) methods for the conventional analysis step performed a posteriori once the MD data have been generated. Path "2" depicts a loop where ML methods enter during the simulations to inform subsequent MD runs (specifically consisting of simulation runs, data generation, and ML-based data analysis). This loop can be either discontinuous (MD/ML resampling) or seamless (on-the-fly MD/ML).

In this review, we summarize the different classes of ML methods applied to the analysis of MD trajectories (hereafter referred to as MD/ML approaches) holding great potential in the field of biomolecular simulations. Here, we stress that some of these methods have been widely employed for decades, like clustering and principal component analysis (PCA), while others have been introduced only very recently, and their application to real-life drug discovery cases is yet to come. Finally, we provide an overview of the emerging methods to include experimental information in MD simulations for better modeling biomolecular systems.

2. Learning from Molecular Dynamics Trajectories

Molecular dynamics is concerned with the time evolution of systems under the classical laws of motion. The integrator is at the very heart of any MD engine, as it takes care of solving Newton's equations iteratively for discrete steps in time (i.e., the time-step). To satisfy a stable and accurate integration, one needs to choose the time-step to be significantly smaller than the characteristic oscillations of the system under consideration, typically 1–2 fs. This has two important implications. First, reaching experimentally relevant timescales for the events one wishes to investigate (micro-, milli-seconds, and possibly more) becomes a formidable computational task [15]. Second, for MD to be informative, the trajectory must be saved to a disk at a sufficiently high pace, thus generating a huge amount of data that requires further analysis. To deal with the former issue, apart from mere technological advances, several and highly diversified methods have been developed over the years. A large class of such methods relies on the notion of reaction coordinates (or collective variables, CVs), which are functions of the atomic positions whereby the investigated event is accelerated through external biases in the form of additional forces or potentials. The computation of most common CVs can be already found implemented either in MD codes, such as NAMD [20,21] and AMBER [22], or in dedicated software such as PLUMED [23]. Among the most popular approaches exploiting CVs, we mention umbrella sampling [24], steered dynamics [25,26], adaptive biasing force [27], and metadynamics [28,29]. Here, for simplicity, we refer collectively to these methods as "biased

sampling" and we redirect the interested reader to more specific reviews for further details [30,31]. Biased sampling methods can be highly informative because not only do they speed up the observation of "rare events" but, under proper simulative conditions, they also allow to retrieve the underlying free energy landscape. Unfortunately, choosing the proper CVs is not always straightforward, and much chemical intuition and/or trial-and-error procedures are often required [32].

Concerning the second issue, namely coping with a big data scenario, the application of ML methods can be of great benefit to the analysis of MD trajectories (Figure 1, path "1"). Furthermore, ML can also be integrated into MD protocols to optimize the production of MD trajectories. In particular, concerning the latter point, ML methods can help in identifying the CVs required for biased sampling simulations (Figure 1, path "2"). Notably, this can either be done subsequently, that is running biased sampling of CVs after they are identified from the analysis of one or more explorative MD simulations (MD/ML resampling), or via on-the-fly protocols (on-the-fly MD/ML). On-the-fly learning and sampling approaches probably represent the most elegant way of combining MD with ML, as they relieve the user of rather subjective choices and often tedious rounds of simulation and analysis. We note that on-the-fly MD/ML can be also declined in what we call here "guided sampling". This class of methods bears some similarities with adaptive sampling procedures like weighted ensemble [33,34], among others. In particular, they share the feature of taking care of launching and controlling repeated sequences of multiple MD simulations in an automated fashion. In addition to assisting the identification of CVs, in this case, ML methods help in optimally identifying the starting states for each sequence of MD runs. In practice, the states are drawn so as to "guide" the system towards undersampled or even previously unexplored regions of configurational space, thus achieving a wider exploration without the need of introducing external biases.

ML methods are generally classified into two broad categories: supervised and unsupervised learning (Table 1) [35]. While unsupervised learning deals with the identification of patterns among data, the ultimate goal of supervised learning is to disclose the relationships (if any) between dependent and independent variables. In this case, the learning procedure is carried out by partitioning the available dataset into two chunks: the training and the test set. The former is employed to train the model, while the latter is exploited to validate its performance in predicting the dependent variables for the subset of data that was not considered during training. It is important to recognize that, in the context of the analysis of MD trajectories, ML methods most often take as input convenient representations of the configuration of the system over time, rather than bare atomic coordinates [36]. These representations must satisfy symmetry invariances (rigid body translations and rotations as well as permutations of identical atoms) and can be described as a vector of features, like dihedral angles, or CVs in general, like contact maps, and so on. Notably, in the ML jargon, this process of mapping the Cartesian coordinates into the space of selected features of interest is typically referred to as "featurization" [8].

Table 1. Classification of the most popular ML algorithms.

Class	Learning Task	Method
Supervised Learning	Regression	Linear regression [1] Non-linear regression Support vector regression (SVR) Artificial neural network (ANN)
	Classification	Logistic regression (LR) [1] Linear discriminant analysis (LDA) [1] Support vector machines (SVR) k-nearest neighbor (kNN) Decision trees/random forests Artificial neural network (ANN)
Unsupervised Learning	Clustering	Hierarchical agglomerative/divisive k-means/-medoid Gaussian mixture models (GMM) [1] Density-based (DBSCAN) Self-organizing maps (SOM)
	Dimensionality Reduction	Principal component analysis (PCA) [1] Kernel-PCA (kPCA) [1] Independent component analysis (ICA) [1] Multidimensional scaling (MDS) [1] Isometric feature mapping (IsoMap) [1] Locally linear embedding (LLE) Diffusion maps (dMaps) [1] Artificial neural network (ANN) [1]

[1] Examples of this ML method have been described in the context of MD analysis and are reported in the text.

2.1. Unsupervised Learning Methods

2.1.1. Clustering and PCA: The Grand Old Tools of Trajectory Analysis

Unsupervised learning can be distinguished into clustering and dimensionality reduction (Figure 2). A more technical discussion of the use of these methods in the context of MD simulations can be found in an excellent review by M. Ceriotti [36]. Briefly, clustering methods attempt to partition input data into classes where members of the same group can be considered more similar between them than members belonging to different groups (Figure 2A). To implement the concept of similarity, one must define a way to measure distances in the feature space. Typically, the pairwise atomic root-mean-squared-deviation (RMSD) is used to this aim:

$$RMSD_{ij} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\mathbf{x}_k^i - \mathbf{x}_k^j\right)^2}, \quad (1)$$

where N is the number of atoms in the summation, and $\mathbf{x}_k^i$ and $\mathbf{x}_k^j$ are features of the system in two different configurations, i and j, sampled in an MD trajectory. When dealing with MD simulations, a common practice is to use atomic coordinates as input features, but other metrics can be envisioned. In this context, cluster analysis is very helpful for identifying the metastable states (i.e., local free energy basins) explored during the dynamics and it represents a popular way to analyze MD trajectories [37]. It is out of the scope of the present work to review such kind of approaches, as the variety of methods and the extent of applications would require an entire article to fully describe them [38]. We limit ourselves by mentioning an on-the-fly MD/ML approach introduced by Tribello et al., where the sampling is biased making use of a sophisticated clustering procedure (Gaussian mixtures) as a central ingredient [39]. The method, dubbed reconnaissance metadynamics, accelerates the dynamics through a repertoire of CVs that are expected to provide the best local approximation of each free energy well that is visited during the simulation. Cluster analysis is performed at regular intervals on the collected trajectory, and the identified clusters are then exploited for tuning a one-dimensional CV that will be used for biasing the dynamics until the next round of analysis [39].

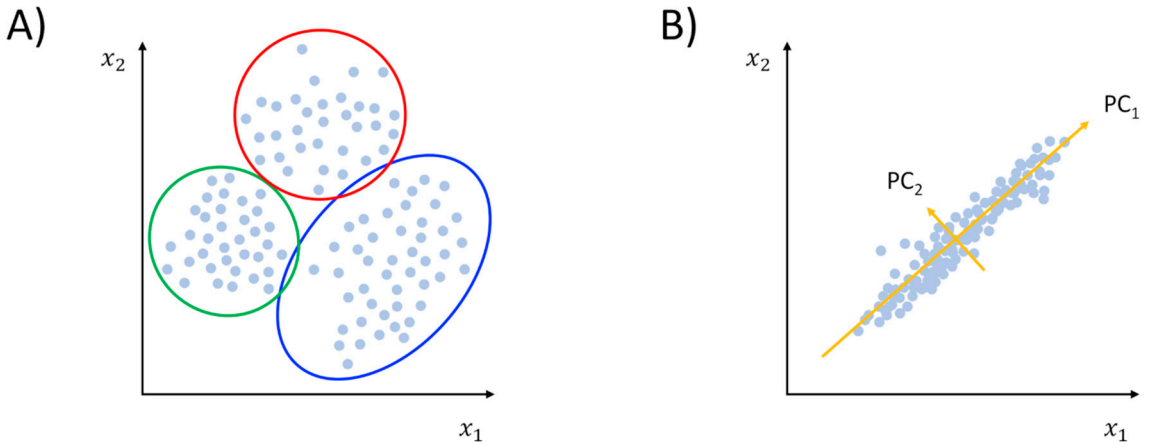

Figure 2. Pictorial representation of the unsupervised learning class of methods: cluster analysis (panel (**A**)) and dimensionality reduction (panel (**B**), principal component analysis (PCA) is displayed as a representative example).

While clustering helps in identifying groups of configurations basing on feature similarity, dimensionality reduction (or manifold learning) represents a variety of methods whose aim, as the name suggests, is to reduce the dimension of the feature vectors [36,40]. In other words, dimensionality reduction methods seek to find a low-dimensional (low-d) manifold embedded in the high-dimensional (high-d) space represented by the input data structure [32]. The procedure is rooted in the possibility of spotting redundancies and correlations that are often found in large data samples [35]. Once the low-d space is obtained, the interpretability of data is in general improved. One should keep in mind that such simplification always comes with a certain degree of information loss. Thus, finding an optimal tradeoff between the two can be difficult in some cases. The forerunner of all dimensionality reduction methods is certainly principal component analysis (PCA), which has become a popular MD analysis tool under the moniker of "essential dynamics" [41]. Technically speaking, PCA provides a linear transformation of the feature vectors in a way that best captures the variance of data. The outcome of the dimensionality reduction is therefore a set of eigenvectors (or principal components, PCs) ranked by the decreasing fraction of the total variance explained (eigenvalues, see Figure 2B). While in principle this can be applied to all sorts of features, it is common practice to apply PCA to atomic coordinates. Specifically, eigendecomposition can be performed by diagonalizing the covariance matrix of atomic fluctuations, whose elements are

$$C_{ij} = \left\langle \mathbf{x}_i \cdot \mathbf{x}_j \right\rangle, \tag{2}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the positional fluctuation vectors of atoms i and j, while the angle brackets denote the average evaluated over the entire simulation (ensemble average). We note, in passing, that in the one-dimensional case, the elements of the normalized covariance matrix ($C_{ij} = \langle x_i \cdot x_j \rangle / \sqrt{\langle x_i \rangle^2 \langle x_j \rangle^2}$) correspond to the Pearson correlation coefficient (r) between the two variables.

From a practical standpoint, this powerful ML analysis tool is used to identify the correlated motions of proteins or biomolecules in general. This is especially relevant as the projection of the trajectory on the first few principal components allows a rather straightforward identification of metastable states and transitions among them without the need of resorting to a cumbersome visual inspection of configurations. Thus, PCA can be in principle employed to study mechanisms underlying conformational transitions, ranging from minor local rearrangements up to entire folding processes (but see below for caveats). As in the case of cluster analysis, even providing a partial list of the most recent applications of PCA in the context of MD analysis would be unfeasible. We rather highlight here that the principal components extracted from MD simulations can be thought of as a set of CVs, and as such, they can be used for biased MD/ML resampling. This idea has been pursued in a form of restrained dynamics ("essential dynamics sampling" [42]) and metadynamics as well [43]. We note,

however, that while PCs can be considered as good order parameters whenever they allow to clearly distinguish among the most relevant states, for a number of reasons, they are not also necessarily good CVs for biased sampling. Indeed, despite its conceptual simplicity and ease of use, which pushed the implementation in several MD analysis tools over the years [22,23,44–48], PCA is not free from limitations. A technical drawback is related to the fact that PCA is typically fed with the Cartesian coordinates of a given subset of atoms (usually the backbone or Cα atoms in the case of proteins). In order to remove irrelevant motions like rigid body translations and rotation, it is customary to align the frames of the trajectory on a reference structure, which is often times the starting configuration or an average conformation. Thus, the results of the dimensionality reduction are somewhat dependent upon the choice of both the reference structure and the atoms used for finding the optimal alignment. A possibility to bypass this problem is choosing a different feature space, like internal coordinates [32]. For example, PCA on dihedral angles can be performed [49,50], and this approach has been recently used by Ferraro et al. to rationalize the change in the efficacy of a series of congeneric modulators of the dopamine D3 receptor [51].

A more elegant choice over PCA, however, is taking advantage of multidimensional scaling (MDS), a distinct, but somehow related ML tool. Differently from PCA, MDS (sometimes also referred to as principal coordinate analysis [52]) operates directly on pairwise distances between conformations (like the RMSD), thus avoiding the optimal alignment problem. Hence, in MDS, the problem can be reformulated as finding the embedding that best preserves the distances evaluated in the high-d space. MDS comes in two flavors: the original algebraic formulation, also known as "classical" MDS, and an optimization procedure through iterative algorithms (distance scaling, or "metric" MDS). The idea behind classical MDS is to transform the distance matrix into an inner product matrix that can be further diagonalized as in PCA. The ground for this reasoning is that, in Euclidean space, distances (D_{ij}) are related to inner products as follows:

$$D_{ij}^2 = \left|\mathbf{x}_i - \mathbf{x}_j\right|^2 = |\mathbf{x}_i|^2 + \left|\mathbf{x}_j\right|^2 - 2\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle. \tag{3}$$

Thus, after a procedure called double centering which takes care of the fact that inner products depend on the origin while distances do not, by inverting this relationship, one obtains the desired inner product matrix [52]. The set of obtained eigenvectors has much the same significance as in PCA, even though MDS is considered more general as it can also be applied to non-Euclidean high-d spaces [32]. Conversely, still under the assumption of a linear projection, in the simplest form of metric MDS, one rather optimizes the loss function [36]:

$$l_{ij}^2 = \sum\nolimits_{ij} \left(D_{ij} - d_{ij}\right)^2, \tag{4}$$

where D_{ij} and d_{ij} are the distances in the high- and low-d spaces, respectively. An important requirement of dimensionality reduction methods is the ability to map new high-d data points into a previously obtained embedding. This is the so-called "out-of-sample" problem that affects MDS and related methods. We note that PCA is devoided from this limitation, as new data points can be easily projected on to the PCs using the same linear transformation employed to carry out the dimensionality reduction. To the best of our knowledge, the first application of linear MDS in the analysis of MD trajectories was reported by Troyer and Cohen as early as 1995 [53]. More recently, Pisani et al. reported on an interesting application of MDS for mapping the conformations explored by the CDK2 protein kinase during MD simulations [54]. Notably, the embedding was constructed using a pool of experimentally derived structures, and an appropriate out-of-sample extension was devised to map the trajectories points on the previously derived low-d space [54].

Another quite serious problem with PCA is that, by construction, principal components only provide a linear mapping of the high-d space of input data. Thus, meaningful results can only be obtained if input data are linearly correlated. While at the bottom of free energy wells the motion

of biomolecules might satisfy the quasi-harmonic approximation, in general, this assumption is no longer valid in the proximity of transition state regions. This means that studying complex and highly non-linear rearrangements like protein folding, while technically feasible, can lead to arguable results. For the same reason, while PCA can be used to gain mechanistic insight at a qualitative level, it should not in general be used for extracting rates related to the process under investigation. The same reasoning applies to linear MDS.

2.1.2. Beyond Linear Dimensionality Reduction

The concept of non-linearity in dimensionality reduction methods can be implemented in several ways. Perhaps, the simplest way to achieve this is through kernel PCA (kPCA) that can be thought of as a generalization of PCA. The idea behind kernel methods is to exploit a non-linear transformation of the input data into some feature space $\mathbf{\Phi}(\mathbf{x})$ of higher dimensionality, with the hope of finding linear correlations in this new space. In particular, the kernel is a function that represents the inner product in the feature space:

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x}_i). \tag{5}$$

By diagonalizing the matrix whose elements correspond to this product, one obtains principal components like in PCA, but in this case, one attempts to capture non-linearity in the high-dimensional space through the definition of the kernel itself. The advantage is that one does not need to explicitly compute the mapping function $\mathbf{\Phi}(\mathbf{x})$, as the kernel matrix can be readily obtained by the input data using a polynomial, exponential, or sigmoid function. Indeed, the covariance matrix that is diagonalized in PCA can be considered as the simplest possible kernel function (the inner product). As an example, Antoniou and Schwartz successfully employed a polynomial kernel ($k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\mathbf{x}_i \cdot \mathbf{x}_j\right)^n$) to extract the CV describing the enzymatic reaction of the lactate dehydrogenase enzyme [55].

From a different standpoint, non-linearity can also be addressed within the MDS framework. Non-metric MDS (nMDS) can be considered as a form of non-linear MDS. Instead of attempting to preserve pairwise distances, it focuses on preserving their ranking in the high-d space. This is a useful approach to be considered when, rather than the exact value of the distance, the relationship among input data is thought to be more relevant. nMDS has been adopted to map the configurational space of the villin headpiece during folding trajectories that were previously generated through exceptionally long MD simulations for that time [56], and it was found to be superior to PCA and conventional cluster analysis [57]. Differently, Sketch Map is a non-linear metric MDS method introduced by Ceriotti et al. that seeks to preserve middle-ranged proximities in a way to collapse or amplify distances for points that are found below or above some characteristic length that is specific for the considered data structure (and that must be priorly assessed) [58]. The non-linear mapping is obtained through a modification of the loss function usually employed in metric MDS (Equation (4)):

$$l_{ij}^2 = \sum\nolimits_{ij} \left[F\left(D_{ij}\right) - f\left(d_{ij}\right)\right]^2, \tag{6}$$

where F and f are sigmoidal functions that are dependent on the choice of the aforementioned characteristic length. The rationale behind this approach is that, for complex molecular transitions like those observed in typical MD simulations, the noise in data due to thermal fluctuations will prevail in the proximity of minima, while undersampling will characterize the transitions between them [59]. Thus, by properly tuning the characteristic length, only the essential features of the high-d space will be preserved in the low-d embedding providing, as the name implies, a sketch of the entire energy landscape is visited by the system. Sketch Map comes with an efficient optimization strategy based on the selection of "landmark" (i.e., representative) points in the high-d space as well as a procedure to cope with the out-of-sample problem [58,59]. In the original implementation, Sketch Map was tested to reduce the dimensionality of the polyalanine-12 peptide using the 24-dimensional space of the backbone dihedral angles as input features [58]. Later, the same authors extended

the methodology to carry out biased MD/ML resampling of a previously generated embedding in analogy with metadynamics (therefore, the method was called field-overlap metadynamics) [60]. Recently, among the several applications, Bellucci et al. applied Sketch Map to describe conformational changes of the 16–22 segment of the β-amyloid peptide upon binding to a gold surface [61].

Isometric feature mapping (Isomap) is another non-linear dimensionality reduction method that builds on MDS [62]. Rather than evaluating the Euclidean distance in the high-d space, Isomap estimates the geodesic distance, which is the distance along a straight line in a curved manifold. In particular, the geodesic distance is computed finding the shortest path through a network analysis performed on the high-d space (Figure 3) [62]. This approximation holds only in the limit of very dense sampling, and, when this requirement is fulfilled, the computation becomes highly inefficient. A variant of the original algorithm specifically designed for big datasets such as the output of MD simulations is the so-called scalable Isomap (ScIMAP) proposed by Clementi and co-workers [63]. ScIMAP alleviates the computational burden by choosing random landmark points and approximating the distances only between these points and the remaining ones, instead of calculating all the pairwise shortest paths [63]. This method has been used to map the conformational space and compute the conformational free energy of coarse-grain models of the Src homology domain 3 (SH3) protein and a 22-residues β-hairpin [63,64]. Isomap has also been extended for its use in the context of biased MD/ML resampling simulations. Notably, the out-of-sample problem and the requirement of a smooth mapping of the configurational space for computing biasing forces (i.e., the differentiability) have been elegantly bypassed by Spiwok and Králová [65] through a generalization of the path CVs previously introduced by Branduardi et al. [66]. These variables are nowadays referred to as "Property Maps" within the PLUMED community [23,67], but we highlight here that an Isomap embedding can also be used as a CV space through the more general "Smooth and Nonlinear Data-Driven CV" (SandCV) formalism developed by Hashemian et al. [68]. Finally, we mention that Isomap was recently used by Schuetz et al. to map the unbinding pathways of drug-like molecules from their target as obtained by high-effective temperature MD simulations [69]. The projection of these pathways onto the low-d space was then clustered using the Fréchet distance as a metric with the aim to gain insight on the unbinding mechanism of the considered molecules [69].

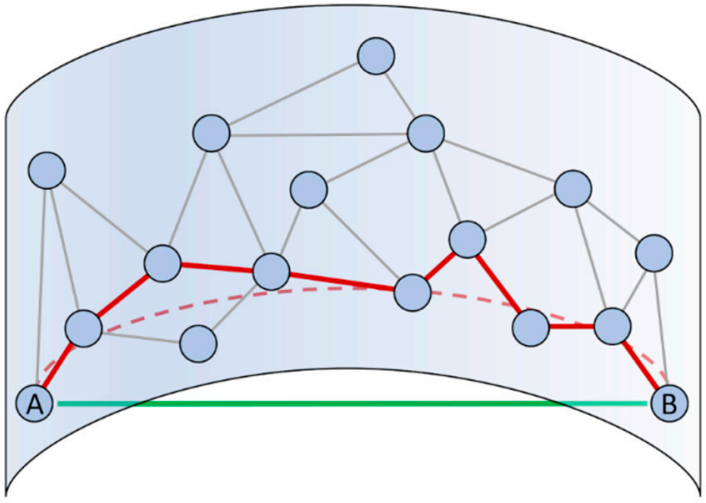

Figure 3. Schematic representation of the difference between the Euclidean and geodesic distance (green solid and red dashed lines, respectively) evaluated in a curved manifold. The network-based nearest neighbor approximation of the geodesic distance provided by Isomap is also shown (red solid lines).

Aside from the above-discussed methods, the issue of non-linearity in dimensionality reduction can be addressed from another perspective. Starting from the limitations of conventional PCA, namely the linear approximation and the often overlooked problem that by construction only collinear

motions can be detected as correlated (see Equation (2)), Lange and Grubmüller devised a generalized measure of correlation which rests on statistical mechanics arguments and information theory [70]. This generalized correlation coefficient (r_{MI}) builds on Shannon's mutual information (MI) between random variables, and, in analogy with the Pearson correlation coefficient r, it is conceived to return a value of 1 for fully correlated motions and 0 when no correlation is found [70]. By minimizing this MI measure in a procedure known as full correlation analysis (FCA), one obtains maximally uncoupled collective coordinates [71]. This is a form of independent component analysis (ICA, see below). Differently from PCA, where eigenvectors are ranked according to the amplitude of motion, the authors proposed a ranking based on the anharmonicity of the modes as assessed through the estimate of their negentropy [71]. For the investigated systems, FCA modes turned out to better describe conformational states and provided a better description of the transition pathway among basins than PCA, suggesting an improved ability to capture functional motions over linear methods [71]. As an example, FCA has been successfully employed to detect functional motions of the HIF-2α PAS-B domain that are possibly involved in assisting ligand (un)binding [72].

2.1.3. Including Dynamical Information into Geometric Dimensionality Reduction

All the methods described in previous sections are based on (linear or non-linear) static properties of the high-d space. A step forward towards a complete mechanistic interpretation of the simulated events can be taken by including some dynamical information on the derivation of the reduced dimensionality space. Diffusion maps are one such example that attempt to preserve the dynamic proximity between configurations visited in the high-d space [73]. To do so, diffusion maps employ a Gaussian kernel (hence, it corresponds to a form of kPCA):

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathrm{e}^{-\frac{RMSD_{ij}^2}{2\varepsilon^2}}, \tag{7}$$

where ε is a characteristic timescale below which the metric can be considered a meaningful representation of the transition between the two configurations $\mathbf{x}_i$ and $\mathbf{x}_j$ [32,73]. With this definition of the kernel, the principal components correspond to the eigenvectors of the Fokker–Planck equation, and therefore they should provide a faithful interpretation of the dynamics of the system. Specifically, apart from the trivial zeroeth mode, the lower ranked diffusion coordinates (DC) correspond to the slowest collective motions of the system, and they can be used as CVs for further sampling [32,73]. Several variants of diffusion maps have been proposed over the years, including the locally scaled diffusion map (LSDMap) by Clementi and coworkers which is an extension of the original method specifically designed to cope with noisy data like that of MD simulations [74]. In particular, the authors introduced an algorithm for detecting the intrinsic dimensionality and the local timescale for each configuration, thus avoiding artifacts in the embedding arising from a uniform choice of the ε parameter [74]. LSDMap allowed the authors to extract well-behaved CVs and to estimate rates through Kramers' rate theory [13,74]. Notably, the DCs captured by LSDMap are global coordinates representing the slowest modes of the entire molecule, while the definition of "local" information would be required for efficiently guiding the dynamics through on-the-fly MD/ML sampling. This idea is exploited in the so-called diffusion map-directed MD (DM-d-MD), where local DCs are estimated by periodically computing DCs, and restarting the simulation in the slowest mode [75]. In an extended version (extended DM-d-MD), the method was combined with a reweighting scheme ensuring the possibility to recover the Boltzmann distribution despite the artificial dynamics [76]. Another method that couples MD and on-the-fly non-linear manifold learning based on diffusion maps is intrinsic map dynamics (iMapD) developed by Chiavazzo et al. [77].

As already stressed, conventional PCA has several drawbacks, including the fact that only a linear correlation can be detected, and not entirely. Moreover, including dynamic information into the PCA framework is problematic, as PCs are not necessarily independent (even though they are orthogonal, see Figure 4). In order to overcome such limitations, Naritomi and Fuchigami introduced

the time structure-based independent correlation analysis (tICA) method [78]. Differently from PCA, as already mentioned, ICA is an ML approach that attempts to extract components that are as statistically independent as possible. The tICA method differs from conventional ICA in that it also includes information on time dependency among the extracted eigenvectors. Accordingly, the usual time-independent covariance matrix of Equation (2) is replaced by a time-lagged covariance matrix:

$$C_{ij}(\tau) = \left\langle \mathbf{x}_i(t) \cdot \mathbf{x}_j(t+\tau) \right\rangle, \tag{8}$$

where τ is a given simulation lag time that must be properly chosen. By diagonalizing the time-lagged covariance matrix, one obtains eigenvectors (independent components, IC) that are no longer orthogonal to each other. Among the interesting properties of this formalism, we mention that the eigenvalues λ_i provide information of the timescales of the associated IC, and in the special case of an autocorrelation function with a single exponential decay, the corresponding timescale t_i can be expressed as

$$t_i = -\frac{\tau}{\ln \lambda_i}. \tag{9}$$

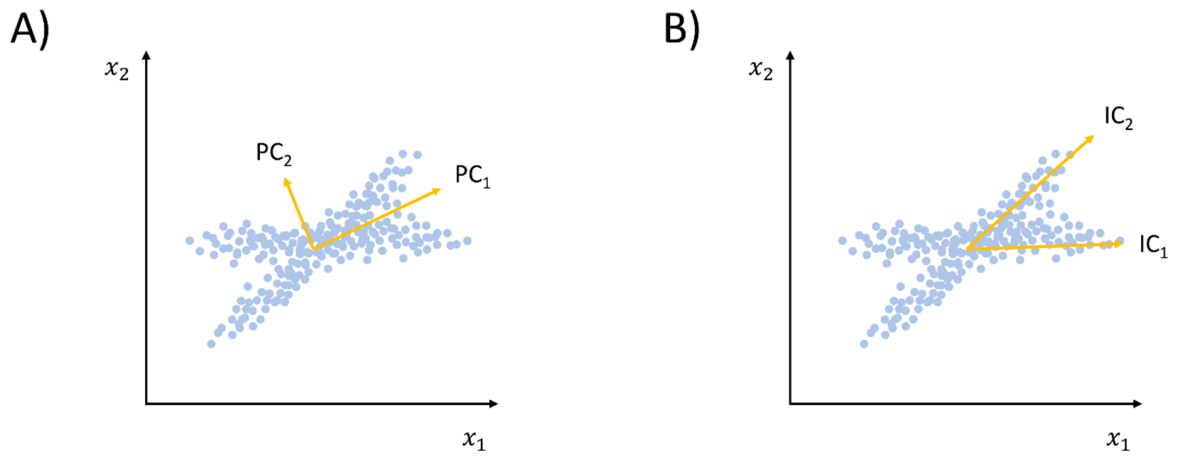

Figure 4. Difference between the components extracted through PCA (panel (**A**)) and a generic independent coordinate analysis (ICA) method (panel (**B**)). In specific cases, ICA provides a better description of the high-d data structure, as the eigenvectors identified are not necessarily restrained to the orthogonality relationship.

Thus, not only can tICA be used to gain mechanistic insight regarding the motion, but it is also useful to get a rough estimate of the associated characteristic timescales [78]. For the analysis to be robust, however, it is critical to choose the lag time appropriately, as any IC with smaller timescales than τ might represent an artifact due to thermal fluctuations rather than representing a true mode of motion [78]. Over the years, tICA has become a popular tool for extracting kinetically relevant CVs in the community of Markov state models (MSM) analysis [79–82]. Specifically, it has been proven that tICA can provide an optimal approximation of the true eigenvectors of the Markov transition matrix [83,84]. In practice, tICA is employed at the very beginning of MSM construction to map the usually high-dimensional input features into a lower-dimensional space that captures the relevant dynamics of the system, on which the subsequent clustering to identify the microstates is then performed. Moreover, it has been recently shown that tICA can be successfully used as CVs for MD/ML resampling (tICA-Metadynamics) even in a low-data regime [85]. A non-linear extension obtained through a Gaussian kernel (landmark kernel tICA-metadynamics) has also been proposed [85].

2.1.4. Neural Networks and Deep Learning

During the last couple of years, active research in the field of dimensionality reduction for analyzing simulation data has mostly been focused on the investigation of the potential offered by deep learning methods like neural networks (NN). Among these, auto-associative neural networks (ANNs, or autoencoders) represent a class of unsupervised ML methods based on the sequential use of two NNs. The first network is used for encoding the low-d embedding (often called the "latent space" in this context) from the input features, while the second takes care of decoding the compressed information of the latent space for reconstructing the original high-dimensionality (see Figure 5) [86]. Each network is composed of a layer of "neurons", whose activation is defined as

$$h_i = f(\boldsymbol{w}_i \mathbf{x} + b_i) = f(\sum_{j=1}^{N} w_{ij} x_j + b_i), \tag{10}$$

where f is a non-linear activating function (a sigmoid function), $\mathbf{x}$ is the input vector, w_{ij} are the elements of the weight matrix of the layer, and b_i is the biases of the layer. A crucial advantage of autoencoders over the methods discussed in the previous paragraph is that the difference among the original high-d space of data and the reconstructed version of it can be used as a direct measure of the performance of the ML method. In this way, autoencoders can be trained to obtain an optimal non-linear low-d embedding. From a practical standpoint, training the network corresponds to optimizing weights and biases to minimize the reconstruction error through iterative procedures where each minimization step is referred to as an "epoch". To control the magnitude of weights and biases during training, regularization terms are usually considered [86].

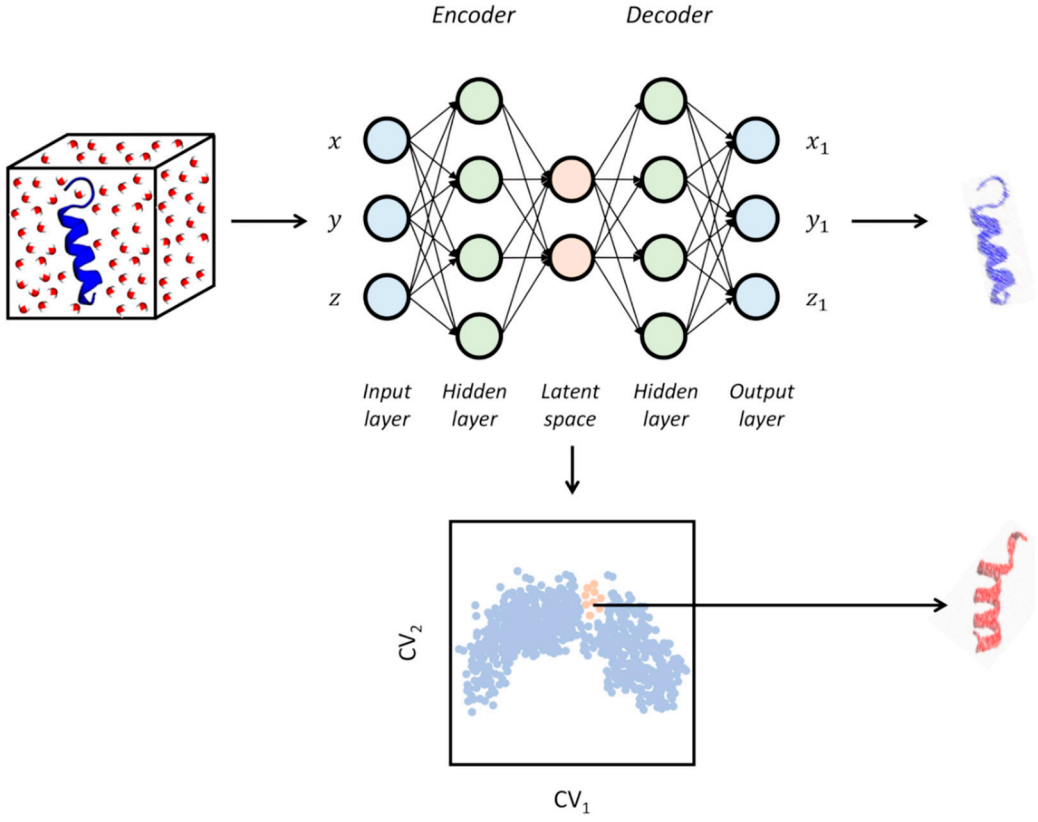

Figure 5. Schematic representation of an autoencoder. Basing on the conformations sampled through the MD simulations (protein in blue ribbons with surrounding water molecules), a latent space can be learned and trained (blue dots in the bottom plot) in a way to reproduce at best the original input data structure (blurred blue protein on the right). The latent space information can also be used to generate previously unexplored conformations (red dots in the bottom plot and red protein on the right).

As previously noted, this is a field that is rapidly evolving. However, without claiming exhaustiveness, we can identify two major classes of applications of DL in the context of MD simulations, even though there are no conceptual boundaries between them, and overlaps can be envisioned. The first focuses on the unsupervised extraction of statistically relevant information like the equilibrium population of states with a particular emphasis on the estimation of rate constants. As we discussed in the previous chapter, a "kinetically relevant" low-d embedding is instrumental to this aim. The second group of methods is more oriented to the automatic extraction of relevant CVs for on-the-fly MD/ML sampling or later use. In both cases, a mechanistic interpretation of the events that occurred during the simulation is also guaranteed. Time-lagged autoencoders like TAE, which extends the domain applicability of autoencoders to the modeling of time-series data, fall in the first class of methods [87]. Similarly, the variational dynamics encoder (VDE) is able to capture the relevant dynamics of complex processes through a non-linear embedding by adding Gaussian noise regularization (the so-called variational autoencoder, VAE) [88]. Closely related to these methods, the main goal of VAMPnets is rather to replace the well-established pipeline of tICA extraction, clustering, and MSM kinetic model building through a fully automated deep neural network, relieving the user from subjective choices and error-prone steps [89].

Concerning the methods focused on CV discovery, we note that the low-d space obtained through autoencoders is by construction a differentiable function of the input coordinates and it is devoided from the out-of-sample problem, making this class of ML methods intrinsically superior over conventional dimensionality reduction methods for CV extraction, MD/ML resampling, and even for on-the-fly MD/ML sampling. Molecular enhanced sampling with autoencoders (MESA) developed by Chen and Ferguson is based on successive rounds of non-linear CV discovery and biased sampling of these CVs [90,91]. Specifically, MESA is an on-the-fly MD/ML-guided sampling protocol which can be summarized as follows: generation of initial training data through previous unbiased or biased MD simulations, autoencoder-based CV discovery, boundary detection for identifying unexplored regions of the CV space, enhanced sampling in the low-d embedding, convergence assessment, and, finally, free energy estimation. A similar procedure is exploited in the reweighted autoencoded variational Bayes for enhanced sampling (RAVE) method proposed by Tiwary and coworkers [92]. Conversely, EncoderMap is an NN method developed by Lemke and Peter which combines the advantages of autoencoders with the loss function employed by Sketch Map (Equation (6)) to get a better defined and interpretable low-d embedding [93]. Remarkably, EncoderMap takes full advantage of the potentialities offered by autoencoders, as it not only allows to obtain a differentiable function mapping from the high- to the low-d space, but it can also be used for backward mapping. From this standpoint, the mapping function linking the low-d embedding to the original high-d space can be used to generate previously uncharted molecular configurations. As the authors stated, this unique feature provided by this class of methods can be used as a new type of molecular modeling. This possibility was further investigated in the improved variant EncoderMap(II) implementing the ability to reproduce both short-ranged and long-ranged features, which is essential for preserving chemical accuracy in the generation of conformations for large and even multi-domain proteins [94]. From a different standpoint, it is worth mentioning the release of the python package named Anncolvar from Spiwok and coworkers [95]. This package allows the training of a neural network for CV extraction and resampling within the PLUMED program [23]. To the best of our knowledge, this is the first example of the implementation of an autoencoder specifically designed for its use in the field of biomolecular simulations and with an eye to the community of researchers using enhanced sampling. This further underscores the importance and popularity gained by these ML methods in the context of MD simulations.

We wish to conclude this section on DL methods with a word of caution. Like clearly stated by Sultan et al., care must be taken when using autoencoders for analyzing MD trajectories, as their superior performance compared to more conventional dimensionality reduction methods is not free from potential pitfalls [96]. Most importantly, the black-box nature of NNs makes it hard to understand what the autoencoder actually learns, potentially leading to identifying a low-d embedding

that cannot necessarily be considered a good CV for biased MD/ML resampling [96]. It is also true, however, that some of the above-reported methods are specifically designed to include the information of the dynamics, and therefore this risk should be contained. In their work, Sultan et al. introduced a tICA-VDE extension that is optimally suited to extract relevant and transferable CVs [96].

2.2. Supervised Learning Methods

Compared to the plethora of unsupervised learning methods described in Section 2.1, the use of supervised ML methods for analyzing and biasing MD simulations is much more limited in the literature. Typically, the tasks that are considered by supervised learning can be distinguished in regression and classification (see Figure 6). Regression problems deal with the construction of quantitative predictive models relating some continuous dependent variables to the independent variables. In this case, ML methods are used to quantify such a relationship provided that a linear or non-linear model function is supplied by the user. Conversely, classification problems deal with the construction of qualitative predictive models able to predict the categorical class labels for a given observation. Some supervised ML methods, such as decision trees and ANN, can be used for both classification and regression problems with opportune measures. Other algorithms, such as linear regression for regression problems or logistic regression for classification problems, cannot easily be exploited for both types of tasks.

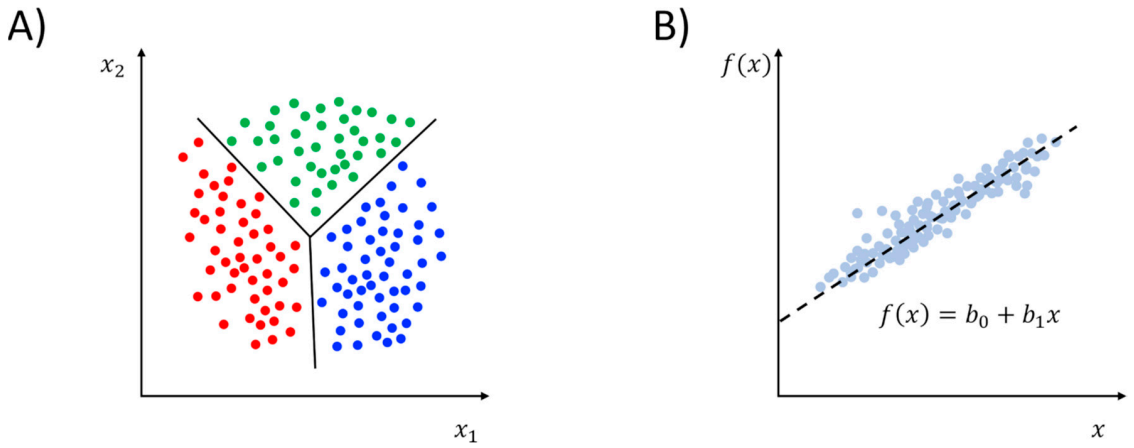

Figure 6. Pictorial representation of supervised learning class of methods: classification (panel (**A**), linear discriminant analysis (LDA) is displayed as an example) and regression (panel (**B**), linear regression displayed as an example,).

The use of regression in the context of the analysis of MD trajectories has been pioneered by Hub and de Groot [97]. The authors observed that, in the context of biomolecular simulations, dimensionality reduction is often carried out for getting mechanistic insight into the system under investigation. This is especially relevant in the case of proteins that are typically known to achieve their biological function, like catalysis, gating, and signal transduction, among others, through collective atomic motions. When dimensionality reduction is performed with well-established methods like PCA, however, the collective motions that one gets are by construction the widest ones, but they are not necessarily directly involved in the biological function [97]. This relation between function and motion is achieved by introducing a functional quantity f so that for each frame of the trajectory it can return a single value. This functional quantity can be any observable that might be relevant to describe the function one wishes to characterize, like atomic distances, binding sites' volume, solvent-accessible surfaces, and so on [97]. Then, assuming that f is a linear function of PCs, through a least-squares optimization procedure, a quantitative model of the observable as a function of the PCs can be obtained much like in PC regression (PCR). This corresponds to maximizing the Pearson's correlation coefficient, but the MI can also be maximized to include the non-linear dependency of f

as a function of atomic coordinates. The vectors that the method finds out are a linear combination of PCs, and the one displaying the largest correlation with the given observable is referred to as maximally correlated motion (MCM). However, as the authors pointed out, this vector is unaware of the underlying free energy landscape, therefore a sort of correction is also devised in order to obtain a physically meaningful coordinate that represents the most probable collective motion that determines the MCM (ensemble-weighted MCM, ewMCM) [97]. In order to avoid overfitting, a cross-validation procedure is implemented which envisions the partition of the trajectory into a training set and a test set. The method is called functional mode analysis (FMA), and the MCM/ewMCM can be also used for MD/ML resampling through biased simulations [97]. The initial assumption that deviations in f are mostly determined by PCs was later weakened in a generalization of the method (partial least-squares FMA, PLS-FMA) that simultaneously optimizes both the model and the basis vectors, yielding to more robust models with a substantially smaller number of components [98]. PLS-FMA was tested on the T4 lysozyme and Trp cage, and then applied to the yeast and human aquaporin channels (Aqy1 and AQP1, respectively), and the CLC-ec1 chloride antiporter using the active site geometry, hydrophobic solvent-accessible surface, channel gating, water permeability, and dihedral angles as functional observables [98]. Very recently, PLS-FMA has been used to identify allosteric communication pathways between the activation gate and the selectivity filter of potassium channels [99].

Whenever one wishes to identify motions involved in the discrimination between states, rather than determining the variation on a continuous observable, classification methods are more suitable than regression. One such method, linear discriminant analysis (LDA), seeks to identify the optimal hyperplane separating the two or more groups of data (that, unlike in cluster analysis, must be known a priori). Accordingly, partial least-squares LDA (PLSA-DA) was used by Peters and de Groot to analyze a series of bound and unbound ubiquitin complexes [100]. By labeling the trajectories according to the binding state (−1 for unbound, +1 for bound), they trained a model returning a vector which maximized the difference of the projection of structures from different classes while minimizing the difference from the same class. By doing so, they observed that the conformations accessible to the bound ubiquitin were partially overlapping those of the unbound ubiquitin, suggesting that conformational selection was the preferred recognition mechanism over induced-fit [100]. The linear discriminant analysis with ITERative procedure (LDA-ITER) is another method specifically designed to overcome a potential bias affecting PLS-DA and related to the fact that the projection vector is obtained through the averaged structures belonging to the two classes, making the results strongly dependent on the anisotropy of the investigated proteins [101].

LDA has also been recently used to approach the problem of identifying optimal CVs for biased MD/ML resampling. In particular, this class of supervised ML can be used to train CVs in the special case of previous knowledge of the end states. From this standpoint, the following methods resemble in spirit the already mentioned path CV framework [66]. A variant of LDA called harmonic LDA (HLDA) has been indeed recently introduced by Parrinello and coworkers as a CV suited to distinguish between two metastable states. This can be derived from short unbiased MD simulations initiated in the end states and through a series of features for mapping these states in a high-d space [102]. The method was further generalized to a multiclass problem (MC-HLDA) in order to treat more than two states simultaneously [103]. The obtained CV space was proven to be effective for reconstructing the free energy surface of chemical reactions through metadynamics, but was unable to lead to a converged free energy surface when a more complex problem like the folding of chignolin was considered [104]. Finally, we mention that other supervised ML methods other than LDA (like support vector machines, and logistic regression) have also been employed for the automatic detection of CVs for MD/ML resampling [105].

3. Learning from Molecular Dynamics Trajectories and Experimental Data

As hardware capacity advances and methods to improve sampling are optimized, the observation of biomolecular events on biologically relevant timescales is gradually becoming more accessible [10,106].

Related to this, possible limitations of the empirical models underlying MD, i.e., the empirical force fields, also become more evident as a result. In fact, deficiencies can appear when a meaningful comparison with reference experimental data is conducted. However, such comparison may display disagreement. While possible limitations in the model become apparent, we can however make optimal use of the available experimental knowledge to improve the quality of the simulations and account for deficiencies thereof. Indeed, the combination of experimental and theoretical sources of information is emerging as an effective strategy to get insights into the structural and functional features of biomolecules [107–109]. As a final note, we highlight that failures are typically ascribed to the MD simulations since they are the result of an empirical model. While this is undoubtedly reasonable, it is nevertheless advisable to include experimental information with criticism, as, in general, any source of data can be affected by errors of systematic, statistical, and procedural nature.

3.1. Validating MD Simulations through Comparison with Experiments

A typical pipeline when exploiting MD is to first set up and carry out the simulations of the biomolecular system, then analyze the resulting trajectories to compute relevant quantities of interest, and finally perform a comparison of the computed outcomes with a measure obtained experimentally. In such a way, what we do is to validate the results of the simulations against reference experimental data. To ensure a fair and meaningful comparison, it is essential that the MD runs are performed under conditions that are as close as possible to those in which the experimental measures were conducted. In this respect, for instance, it was widely reported over the years how ionic conditions of the bulk, both in terms of ionic nature and strength, can play a critical role when biomolecular processes are investigated [110–113]. With such considerations taken into account, the validation procedure reports on to what extent the empirical model possibly succeeds, or fails, in generating realistic and reliable results. Figure 7A depicts the fairly simple principle behind this routine procedure. For each snapshot of the MD trajectory generated by the simulation of a biomolecular system, we compute an observable of interest that we wish to compare with the experiments. A natural choice would be, among all the predicted values from the MD simulation, to pick the ones that provide the best match. The corresponding MD snapshots would thus represent the system configurations that best reproduce the reference experimental data. However, the experimental measure is most times conducted on heterogeneous systems, as the equilibrium population typically comprises a variety of different states of the biomolecule. As a result, the measurement does not reflect a single configuration but is instead an average over the whole ensemble of states. This is true for popular methods that are routinely employed to validate simulation data such as nuclear magnetic resonance (NMR) [114–116], small-angle X-ray scattering (SAXS) [117], double electron–electron resonance (DEER) [118], and Förster resonance energy transfer (FRET) [119]. Therefore, in order to pursue a meaningful validation, one should first average the observable value over all the MD trajectory frames and then compare this result with the reference experimental one.

3.2. Improving the Agreement through Ensemble Reweighting

If the validation procedure reports a disagreement between predicted (from the MD runs) and measured (in the experiments), one can still make optimal use of the available data through a reweighting strategy. In particular, the MD trajectory frames can be re-assigned suitable weights that allow improving the agreement with the experiments. The procedure is illustrated in Figure 7B. In this scheme, the configurations which are assigned higher weights are those that will contribute more to the final average, that is to the predicted value for the observable. In other words, such structures will be more representative of the experimental ensemble. As a result, the inaccuracies deriving from the employed empirical force field are corrected by including the guidance of the experimental data, thus improving the agreement. The use of the maximum entropy principle [120,121] is gaining popularity to implement such a strategy [19,122]. Specifically, given an initial (prior) distribution, by exploiting the maximum entropy principle, it is possible to identify a new distribution (posterior)

that is as close as possible to the original one and matches the experimental reference data. In other words, the initial distribution is subject to the least possible perturbation that allows improving the agreement with experimental observation. Under the maximum entropy framework, the new distribution can be represented as

$$P_{ME}(\mathbf{x}) \propto e^{-\lambda(s(\mathbf{x}))} P_0(\mathbf{x}). \tag{11}$$

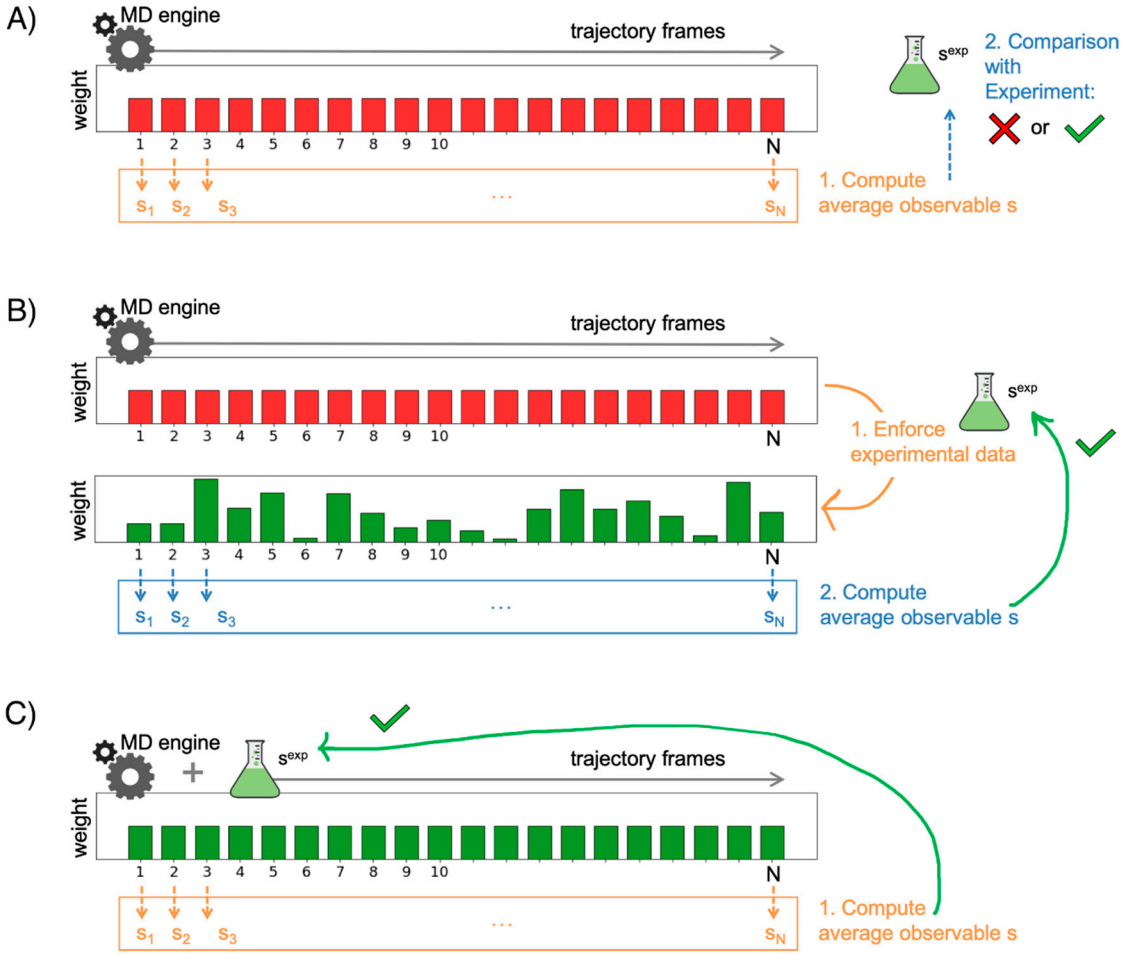

Figure 7. Using MD simulations in combination with experimental information. (**A**) Through a validation procedure, it is possible to estimate the agreement between computed quantities (average observable s, in the figure) and reference experimental data (s^{exp}). (**B**) Correction of the sampled data through a reweighting procedure can improve the agreement between predicted (MD trajectory) and measured (experiments). (**C**) Enforcing the experimental information in an on-the-fly fashion, the sampled ensemble is restrained to best match the experimental one.

This form gives all the possible posterior distributions P_{ME} which are the closest to the prior distribution P_0. Here, $s(\mathbf{x})$ is the observable of interest computed for each configuration $\mathbf{x}$ sampled in the MD simulations. Among the posterior distributions P_{ME}, the one that gives the best match with the experiment is sought. Such a latter requirement can be fulfilled by solving a minimization problem aimed at identifying the suitable value of the parameter λ [19]. This reweighting framework based on the maximum entropy principle has shown to be effective when applied to biomolecular systems of a different kind [123–128]. The procedure has been exploited by Bottaro et al. to reconstruct the conformational ensemble of four model tetranucleotides using extensive atomistic MD simulations and NMR experimental data [123]. Employing the simulation data alone, a significant disagreement with respect to the experiments was apparent. Thus, the simulated ensemble was refined through

reweighting using the NMR experimental data, including nuclear Overhauser effect (NOE) intensities and scalar couplings, in order to improve the agreement. A similar pipeline using NOEs experimental data to refine atomistic MD simulations was also pursued on a longer RNA construct, of 29-nucleotide length, belonging to the SINEUP family [126]. Other applications using NMR data were conducted on a nonapeptide, where simulated ensembles via MD were used in conjunction with ^{3}J coupling and RDC data [127], and on an intrinsically disordered protein, where coarse-grained simulations were reweighted with RDC measurements [125]. Reweighting using hydrogen–deuterium exchange combined with mass spectrometry (HDX-MS) data was also considered by Bradshaw et al. [128]. The scheme was first explored using artificial HDX-MS representing a conformational ensemble of the periplasmic binding protein TeaA that rapidly interconverted between its open and closed states. Such data were used to reweight bias-exchange metadynamics simulations. The procedure was then applied to the amino acid transporter LeuT membrane protein using experimental HDX-MS data. Różycki et al. included information from SAXS measurements performed at high-salt and low-salt concentrations to reweight coarse-grained simulations of CHMP3, a key protein of the ESCRT protein assembly [124]. As a result of the procedure, further insights into the conformations adopted by CHMP3 in its activated and autoinhibited states were obtained [124].

Noteworthy, the reweighting procedure is carried out as an analysis on outcome trajectories of MD simulations, i.e., after the MD runs are performed. Thus, a striking advantage of this procedure is that it can be repeated using additional experimental data, or different ones, with no need of performing new simulations. A further aspect is the possibility to include a regularization term, that can be introduced in the expression to model the experimental error and other sources of possible errors [19]. Indeed, in this respect, we note that the forward model, which is the form through which the observable is computed from the MD snapshots, can also incorporate errors. The effect of including such regularization, which the ML community is familiar with, is to soften the restraint towards the experimental reference. Thus, while the average computed after applying the reweighting procedure is going to match the experimental reference by construction, the application of a reweighting procedure with a regularization term is going to result in a computed average that sits between the one predicted from the prior with no reweighting and the experimental one. Finally, we note that the approach is not devoid of limitations. In particular, for a meaningful reweighting to be applicable, a certain degree of overlap is required between the sampled conformations and those comprised in the experimental ensemble, which is not the case when the domain sampled by the MD simulations contains a scarce number of configurations consistent with the experiment, if at all [129,130]. In such cases, the inapplicability of the reweighting procedure becomes apparent as either a large fraction of the total weight is assigned to one or few frames, or the minimization is not able to converge to a suitable λ. An illustrative example of this issue in a one-dimensional model is provided in the insightful review by Cesari et al. [19].

3.3. Enforcing Experimental Information during the Simulations

Another option that exploits the maximum entropy strategy and that in principle avoids the just-mentioned condition consists of including the experimental knowledge during the MD simulations [19,131]. In practice, the ensemble average computed from the simulations is enforced to match the experimental one in an on-the-fly fashion. This is achieved by modifying the system potential energy through the inclusion of an additional term, which has the effect of constraining the system towards configurations in better agreement with the reference experimental data [132,133]. A schematic depiction is given in Figure 7C. While, in principle, such a procedure has the advantage of producing a sampling domain which is, by construction, more consistent with the reference data, this nevertheless implies having chosen the target average value for the observable a priori, before starting the simulation. In other words, in the case where additional or different experimental data become available, within this framework, a new simulation needs to be performed from scratch. The approach was successfully applied to RNA nucleosides and dinucleotides to enforce NMR

experimental data during replica-exchange MD simulations [132]. In particular, the information from ^{3}J scalar couplings was exploited to guide the enhanced sampling simulations. Force field corrections to match the experimental reference were thus identified and were then validated over independent NMR solution experiments.

A different approach with the same purpose of instructing MD simulations by taking advantage of experimental knowledge is based on a multi-replica strategy [134–136]. Specifically, multiple replicas of the system are simulated with MD at the same time. Then, the average of the interesting observable over such replicas is enforced to match the experimental value. As a result, a restrained ensemble is obtained. Notably, in the limit of a large number of replicas employed, the method has been shown to produce the same ensemble of configurations like the one generated through the maximum entropy scheme used on-the-fly [131,137,138]. The multi-replica strategy was used by Best and Vendruscolo to generate an ensemble of structures of the third fibronectin type III domain from human tenascin that was consistent with available NMR data [135]. Similarly, the conformational variability of the ubiquitin protein in solution was probed using MD simulations and enforcing experimental information from NMR relaxation experiments using the multi-replica approach [136]. A more recent study relying on the use of multiple replicas was conducted by Hermann and Hub, where the strategy was used to enforce SAXS experimental data in an on-the-fly fashion during MD simulations of intrinsically disordered proteins [139].

Finally, inspired by this replica approach, the metainference method was further devised [140]. Metainference combines the mentioned multi-replica scheme with Bayesian inference. The inclusion of the statistical basis of the latter allows one to tune the strength of the restraints towards the reference experimental data. In such a way, all possible sources of errors, including errors in the experimental data or in the forward model, are taken into account. The method and its declination where metainference is combined with metadynamics (metadynamic metainference) [141] have been applied to diverse biological systems and take advantage of different sources of experimental information. Heller et al. studied the binding of the small molecule ligand 10058-F4 to the disordered protein c-Myc using metadynamic metainference simulations with experimental restraining consisting of NMR data, specifically backbone chemical shifts [142]. In a similar biological context, metadynamic metainference and NMR chemical shifts were exploited by Hultqvist et al. to get insights into the interaction of the two disordered proteins CID and NCBD [143]. Backbone chemical shifts were also used by Buckle et al. to investigate the interaction of the SNa15 peptide with non-native mineral surfaces [144]. Finally, concerning NMR, RDC data were employed in the metainference framework by Weber and coworkers to study the conformational space accessible to the LC protein [145]. Interestingly, cryo-EM experimental data were demonstrated to suitable to be integrated in a metainference scheme. In particular, Bonomi and coworkers took advantage of cryo-EM information in the effort of characterizing the structure and dynamics of the integral membrane receptor STRA6 [146]. Another example was reported by Vahidi et al. in their investigation on the structural dynamics of the ClpP proteolytic complex, where cryo-EM data were used to perform metainference [147]. Finally, SAXS intensities were also explored as a source of experimental information to be used with metainference. Paissoni and coworkers exploited this strategy to investigate the conformational ensemble of K63-linked diubiquitin [148] and to refine models of nucleic acid-protein complexes [149]. Similarly, Kooshapur et al. used SAXS experimental data in a metainference framework to derive a structural model of a complex between an RNA-binding protein and a microRNA [150].

4. Conclusions and Perspectives

In this review, we have summarized the currently available strategies that can be exploited to make optimal use of a constantly increasing volume of data in the field of molecular dynamics simulations applied to pharmaceutically relevant biological systems. Specifically, we have shown that this wealth of data can either be the output of MD simulations or come from experimental sources. The available information can be then exploited to inform subsequent calculations or to improve the prediction of

relevant observables. We have also shown that some of the most popular analysis tools that have historically been employed in the field of MD simulations pertain to the domain of machine learning methods, and we have provided an overview of the most influential approaches belonging to the classes of supervised and unsupervised ML methods. For each considered approach, relevant examples of applications in the field of biomolecular simulations, and more specifically to drug design, have been briefly discussed when available. Finally, we have summarized the simulative and analysis approaches that exploit experimental knowledge to improve the quality of computational predictions.

In summary, we have shown how the increasing richness of data (up to the regime of big data) is prompting a shift in the methodologies employed in the field of molecular dynamics in favor of more automatized and less human-dependent procedures. This has started to change not only the way we are analyzing but also the way we are conceiving MD simulations as a whole. In fact, the boundary between data production (i.e., the trajectory above all) and data analysis, that have traditionally been considered as separated processes (see Figure 1, path "1"), is getting less sharp in the newest MD/ML implementations (path "2" in the same figure). Similarly, but from a different perspective, experimental data, which are usually only considered during force field development, are gaining a key role in post-processing or even in guiding MD simulations (Figure 1, path "3"). These advances will ultimately lead to more efficient and/or predictive data-driven MD simulations with important implications for the entire community of biological simulations, including applications in the field of computational drug discovery.

Author Contributions: Conceptualization, M.M. and M.B. (Mattia Bernetti); methodology, M.M., M.B. (Mattia Bernetti), and M.B. (Martina Bertazzo); writing—original draft preparation, M.M. and M.B. (Mattia Bernetti); writing—review and editing, M.B. (Mattia Bernetti), M.B. (Martina Bertazzo), and M.M.; visualization, M.B. (Mattia Bernetti), M.B. (Martina Bertazzo), and M.M.; supervision, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Giovanni Bussi and Sergio Decherchi are acknowledged for providing useful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178–180. [CrossRef]
2. Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626. [CrossRef]
3. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E.W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334. [CrossRef] [PubMed]
4. Schaduangrat, N.; Lampa, S.; Simeon, S.; Gleeson, M.P.; Spjuth, O.; Nantasenamat, C. Towards reproducible computational drug discovery. *J. Cheminform.* **2020**, *12*, 9. [CrossRef]
5. Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21*, 151. [CrossRef] [PubMed]
6. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250. [CrossRef]
7. Hu, Y.; Bajorath, J. Entering the 'big data' era in medicinal chemistry: Molecular promiscuity analysis revisited. *Future Sci. OA* **2017**, *3*, FSO179. [CrossRef]
8. Lavecchia, A. Deep learning in drug discovery: Opportunities, challenges and future prospects. *Drug Discov. Today* **2019**, *24*, 2017–2032. [CrossRef]
9. Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119*, 10520–10594. [CrossRef]
10. Hollingsworth, S.A.; Dror, R.O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143. [CrossRef]
11. Gioia, D.; Bertazzo, M.; Recanatini, M.; Masetti, M.; Cavalli, A. Dynamic Docking: A Paradigm Shift in Computational Drug Discovery. *Molecules* **2017**, *22*, 2029. [CrossRef] [PubMed]

12. Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937. [CrossRef] [PubMed]
13. Bernetti, M.; Masetti, M.; Rocchia, W.; Cavalli, A. Kinetics of Drug Binding and Residence Time. *Annu. Rev. Phys. Chem.* **2019**, *70*, 143–171. [CrossRef] [PubMed]
14. De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59*, 4035–4061. [CrossRef] [PubMed]
15. Borhani, D.W.; Shaw, D.E. The future of molecular dynamics simulations in drug discovery. *J. Comput. Aided Mol. Des.* **2012**, *26*, 15–26. [CrossRef] [PubMed]
16. Daidone, I.; Amadei, A. Essential dynamics: Foundation and applications. *Wires Comput. Mol. Sci.* **2012**, *2*, 762–770. [CrossRef]
17. Klein, M.; Sharma, R.; Bohrer, C.H.; Avelis, C.M.; Roberts, E. Biospark: Scalable analysis of large numerical datasets from biological simulations and experiments using Hadoop and Spark. *Bioinformatics* **2017**, *33*, 303–305. [CrossRef]
18. Dean, J.; Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Commun. ACM* **2008**, *51*, 107–113. [CrossRef]
19. Cesari, A.; Reißer, S.; Bussi, G. Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments. *Computation* **2018**, *6*, 15. [CrossRef]
20. Phillips, J.C.; Hardy, D.J.; Maia, J.D.C.; Stone, J.E.; Ribeiro, J.V.; Bernardi, R.C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, 044130. [CrossRef]
21. Fiorin, G.; Klein, M.L.; Hénin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **2013**, *111*, 3345–3362. [CrossRef]
22. Case, D.A.; Cheatham Iii, T.E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688. [CrossRef] [PubMed]
23. Tribello, G.A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604–613. [CrossRef]
24. Torrie, G.M.; Valleau, J.P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199. [CrossRef]
25. Grubmüller, H.; Heymann, B.; Tavan, P. Ligand Binding: Molecular Mechanics Calculation of the Streptavidin-Biotin Rupture Force. *Science* **1996**, *271*, 997. [CrossRef] [PubMed]
26. Isralewitz, B.; Gao, M.; Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* **2001**, *11*, 224–230. [CrossRef]
27. Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.* **2008**, *128*, 144120. [CrossRef] [PubMed]
28. Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562. [CrossRef]
29. Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100*, 020603. [CrossRef]
30. Abrams, C.; Bussi, G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2014**, *16*, 163–199. [CrossRef]
31. Decherchi, S.; Masetti, M.; Vyalov, I.; Rocchia, W. Implicit solvent methods for free energy estimation. *Eur. J. Med. Chem.* **2015**, *91*, 27–42. [CrossRef] [PubMed]
32. Rohrdanz, M.A.; Zheng, W.; Clementi, C. Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions. *Annu. Rev. Phys. Chem.* **2013**, *64*, 295–316. [CrossRef] [PubMed]
33. Zuckerman, D.M.; Chong, L.T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu. Rev. Biophys.* **2017**, *46*, 43–57. [CrossRef] [PubMed]
34. Betz, R.M.; Dror, R.O. How Effectively Can Adaptive Sampling Methods Capture Spontaneous Ligand Binding? *J. Chem. Theory Comput.* **2019**, *15*, 2053–2063. [CrossRef] [PubMed]
35. Ferguson, A.L. Machine learning and data science in soft materials engineering. *J. Phys. Condens. Matter* **2017**, *30*, 043002. [CrossRef] [PubMed]
36. Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys.* **2019**, *150*, 150901. [CrossRef]

37. Shao, J.; Tanner, S.W.; Thompson, N.; Cheatham, T.E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3*, 2312–2334. [CrossRef]
38. Bottegoni, G.; Rocchia, W.; Cavalli, A. Application of Conformational Clustering in Protein–Ligand Docking. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Springer: New York, NY, USA, 2012; pp. 169–186.
39. Tribello, G.A.; Ceriotti, M.; Parrinello, M. A self-learning algorithm for biased molecular dynamics. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 17509. [CrossRef]
40. Tribello, G.A.; Gasparotto, P. Using Dimensionality Reduction to Analyze Protein Trajectories. *Front. Mol. Biosci.* **2019**, *6*, 46. [CrossRef]
41. Amadei, A.; Linssen, A.B.M.; Berendsen, H.J.C. Essential dynamics of proteins. *Proteins Struct. Funct. Bioinform.* **1993**, *17*, 412–425. [CrossRef]
42. Amadei, A.; Linssen, A.B.M.; de Groot, B.L.; van Aalten, D.M.F.; Berendsen, H.J.C. An Efficient Method for Sampling the Essential Subspace of Proteins. *J. Biomol. Struct. Dyn.* **1996**, *13*, 615–625. [CrossRef]
43. Spiwok, V.; Lipovová, P.; Králová, B. Metadynamics in Essential Coordinates: Free Energy Simulation of Conformational Changes. *J. Phys. Chem. B* **2007**, *111*, 3073–3076. [CrossRef] [PubMed]
44. Kutzner, C.; Páll, S.; Fechner, M.; Esztermann, A.; de Groot, B.L.; Grubmüller, H. More bang for your buck: Improved use of GPU nodes for GROMACS 2018. *J. Comput. Chem.* **2019**, *40*, 2418–2431. [CrossRef] [PubMed]
45. Glykos, N.M. Software news and updates carma: A molecular dynamics analysis program. *J. Comput. Chem.* **2006**, *27*, 1765–1768. [CrossRef] [PubMed]
46. Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S.; Domański, J.; Dotson, D.; Buchoux, S.; Kenney, I.; et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In Proceedings of the Python in Science Conference 2016, Austin, TX, USA, 11–17 July 2016; pp. 98–105. [CrossRef]
47. McGibbon, R.T.; Beauchamp, K.A.; Harrigan, M.P.; Klein, C.; Swails, J.M.; Hernández, C.X.; Schwantes, C.R.; Wang, L.-P.; Lane, T.J.; Pande, V.S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532. [CrossRef]
48. Grant, B.J.; Rodrigues, A.P.C.; ElSawy, K.M.; McCammon, J.A.; Caves, L.S.D. Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics* **2006**, *22*, 2695–2696. [CrossRef]
49. 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]
50. 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]
51. Ferraro, M.; Decherchi, S.; De Simone, A.; Recanatini, M.; Cavalli, A.; Bottegoni, G. Multi-target dopamine D3 receptor modulators: Actionable knowledge for drug design from molecular dynamics and machine learning. *Eur. J. Med. Chem.* **2020**, *188*, 111975. [CrossRef] [PubMed]
52. Becker, O.M. Geometric versus topological clustering: An insight into conformation mapping. *Proteins Struct. Funct. Bioinform.* **1997**, *27*, 213–226. [CrossRef]
53. Troyer, J.M.; Cohen, F.E. Protein conformational landscapes: Energy minimization and clustering of a long molecular dynamics trajectory. *Proteins Struct. Funct. Bioinform.* **1995**, *23*, 97–110. [CrossRef] [PubMed]
54. Pisani, P.; Caporuscio, F.; Carlino, L.; Rastelli, G. Molecular Dynamics Simulations and Classical Multidimensional Scaling Unveil New Metastable States in the Conformational Landscape of CDK2. *PLoS ONE* **2016**, *11*, e0154066. [CrossRef] [PubMed]
55. Antoniou, D.; Schwartz, S.D. Toward Identification of the Reaction Coordinate Directly from the Transition State Ensemble Using the Kernel PCA Method. *J. Phys. Chem. B* **2011**, *115*, 2465–2469. [CrossRef] [PubMed]
56. Freddolino, P.L.; Schulten, K. Common Structural Transitions in Explicit-Solvent Simulations of Villin Headpiece Folding. *Biophys. J.* **2009**, *97*, 2338–2347. [CrossRef] [PubMed]
57. Rajan, A.; Freddolino, P.L.; Schulten, K. Going beyond Clustering in MD Trajectory Analysis: An Application to Villin Headpiece Folding. *PLoS ONE* **2010**, *5*, e9890. [CrossRef] [PubMed]
58. Ceriotti, M.; Tribello, G.A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13023. [CrossRef]
59. Ceriotti, M.; Tribello, G.A.; Parrinello, M. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *J. Chem. Theory Comput.* **2013**, *9*, 1521–1532. [CrossRef]

60. Tribello, G.A.; Ceriotti, M.; Parrinello, M. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 5196. [CrossRef]
61. Bellucci, L.; Ardèvol, A.; Parrinello, M.; Lutz, H.; Lu, H.; Weidner, T.; Corni, S. The interaction with gold suppresses fiber-like conformations of the amyloid β (16–22) peptide. *Nanoscale* **2016**, *8*, 8737–8748. [CrossRef]
62. Tenenbaum, J.B.; Silva, V.d.; Langford, J.C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319. [CrossRef]
63. Das, P.; Moll, M.; Stamati, H.; Kavraki, L.E.; Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9885. [CrossRef] [PubMed]
64. Stamati, H.; Clementi, C.; Kavraki, L.E. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 223–235. [CrossRef] [PubMed]
65. Spiwok, V.; Králová, B. Metadynamics in the conformational space nonlinearly dimensionally reduced by Isomap. *J. Chem. Phys.* **2011**, *135*, 224504. [CrossRef] [PubMed]
66. Branduardi, D.; Gervasio, F.L.; Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, 054103. [CrossRef]
67. Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G.A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P.G.; Bottaro, S.; Branduardi, D.; et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16*, 670–673. [CrossRef]
68. Hashemian, B.; Millán, D.; Arroyo, M. Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables. *J. Chem. Phys.* **2013**, *139*, 214101. [CrossRef]
69. Schuetz, D.A.; Bernetti, M.; Bertazzo, M.; Musil, D.; Eggenweiler, H.-M.; Recanatini, M.; Masetti, M.; Ecker, G.F.; Cavalli, A. Predicting Residence Time and Drug Unbinding Pathway through Scaled Molecular Dynamics. *J. Chem. Inf. Model.* **2019**, *59*, 535–549. [CrossRef]
70. Lange, O.F.; Grubmüller, H. Generalized correlation for biomolecular dynamics. *Proteins Struct. Funct. Bioinform.* **2006**, *62*, 1053–1061. [CrossRef]
71. Lange, O.F.; Grubmüller, H. Full correlation analysis of conformational protein dynamics. *Proteins Struct. Funct. Bioinform.* **2008**, *70*, 1294–1312. [CrossRef]
72. Masetti, M.; Falchi, F.; Recanatini, M. Protein Dynamics of the HIF-2 alpha PAS-B Domain upon Heterodimerization and Ligand Binding. *PLoS ONE* **2014**, *9*. [CrossRef]
73. Ferguson, A.L.; Panagiotopoulos, A.Z.; Debenedetti, P.G.; Kevrekidis, I.G. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 13597. [CrossRef] [PubMed]
74. Rohrdanz, M.A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.* **2011**, *134*, 124116. [CrossRef] [PubMed]
75. Zheng, W.; Rohrdanz, M.A.; Clementi, C. Rapid Exploration of Configuration Space with Diffusion-Map-Directed Molecular Dynamics. *J. Phys. Chem. B* **2013**, *117*, 12769–12776. [CrossRef] [PubMed]
76. Preto, J.; Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19181–19191. [CrossRef]
77. Chiavazzo, E.; Covino, R.; Coifman, R.R.; Gear, C.W.; Georgiou, A.S.; Hummer, G.; Kevrekidis, I.G. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E5494. [CrossRef]
78. Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions. *J. Chem. Phys.* **2011**, *134*, 065101. [CrossRef]
79. Husic, B.E.; Pande, V.S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396. [CrossRef]
80. Chodera, J.D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135–144. [CrossRef]
81. Schwantes, C.R.; McGibbon, R.T.; Pande, V.S. Perspective: Markov models for long-timescale biomolecular dynamics. *J. Chem. Phys.* **2014**, *141*, 090901. [CrossRef]
82. Wang, W.; Cao, S.; Zhu, L.; Huang, X. Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules. *Wires Comput. Mol. Sci.* **2018**, *8*, e1343. [CrossRef]

83. Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102. [CrossRef] [PubMed]
84. Schwantes, C.R.; Pande, V.S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009. [CrossRef] [PubMed]
85. M. Sultan, M.; Pande, V.S. tICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *J. Chem. Theory Comput.* **2017**, *13*, 2440–2447. [CrossRef]
86. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504. [CrossRef] [PubMed]
87. Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **2018**, *148*, 241703. [CrossRef]
88. Hernández, C.X.; Wayment-Steele, H.K.; Sultan, M.M.; Husic, B.E.; Pande, V.S. Variational encoding of complex dynamics. *Phys. Rev. E* **2018**, *97*, 062412. [CrossRef]
89. Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5. [CrossRef]
90. Chen, W.; Ferguson, A.L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* **2018**, *39*, 2079–2102. [CrossRef]
91. Chen, W.; Tan, A.R.; Ferguson, A.L. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *J. Chem. Phys.* **2018**, *149*, 072312. [CrossRef]
92. Ribeiro, J.M.L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J. Chem. Phys.* **2018**, *149*, 072301. [CrossRef]
93. Lemke, T.; Peter, C. EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J. Chem. Theory Comput.* **2019**, *15*, 1209–1215. [CrossRef] [PubMed]
94. Lemke, T.; Berg, A.; Jain, A.; Peter, C. EncoderMap(II): Visualizing Important Molecular Motions with Improved Generation of Protein Conformations. *J. Chem. Inf. Model.* **2019**, *59*, 4550–4560. [CrossRef] [PubMed]
95. Trapl, D.; Horvacanin, I.; Mareska, V.; Ozcelik, F.; Unal, G.; Spiwok, V. Anncolvar: Approximation of Complex Collective Variables by Artificial Neural Networks for Analysis and Biasing of Molecular Simulations. *Front. Mol. Biosci.* **2019**, *6*, 25. [CrossRef] [PubMed]
96. Sultan, M.M.; Wayment-Steele, H.K.; Pande, V.S. Transferable Neural Networks for Enhanced Sampling of Protein Dynamics. *J. Chem. Theory Comput.* **2018**, *14*, 1887–1894. [CrossRef] [PubMed]
97. Hub, J.S.; de Groot, B.L. Detection of Functional Modes in Protein Dynamics. *PLoS Comput. Biol.* **2009**, *5*, e1000480. [CrossRef]
98. Krivobokova, T.; Briones, R.; Hub, J.S.; Munk, A.; de Groot, B.L. Partial Least-Squares Functional Mode Analysis: Application to the Membrane Proteins AQP1, Aqy1, and CLC-ec1. *Biophys. J.* **2012**, *103*, 786–796. [CrossRef]
99. Kopec, W.; Rothberg, B.S.; de Groot, B.L. Molecular mechanism of a potassium channel gating through activation gate-selectivity filter coupling. *Nat. Commun.* **2019**, *10*, 5366. [CrossRef]
100. Peters, J.H.; de Groot, B.L. Ubiquitin Dynamics in Complexes Reveal Molecular Recognition Mechanisms Beyond Induced Fit and Conformational Selection. *PLoS Comput. Biol.* **2012**, *8*, e1002704. [CrossRef]
101. Sakuraba, S.; Kono, H. Spotting the difference in molecular dynamics simulations of biomolecules. *J. Chem. Phys.* **2016**, *145*, 074116. [CrossRef]
102. Mendels, D.; Piccini, G.; Parrinello, M. Collective Variables from Local Fluctuations. *J. Phys. Chem. Lett.* **2018**, *9*, 2776–2781. [CrossRef]
103. Piccini, G.; Mendels, D.; Parrinello, M. Metadynamics with Discriminants: A Tool for Understanding Chemistry. *J. Chem. Theory Comput.* **2018**, *14*, 5040–5044. [CrossRef] [PubMed]
104. Mendels, D.; Piccini, G.; Brotzakis, Z.F.; Yang, Y.I.; Parrinello, M. Folding a small protein using harmonic linear discriminant analysis. *J. Chem. Phys.* **2018**, *149*, 194113. [CrossRef] [PubMed]
105. Sultan, M.M.; Pande, V.S. Automated design of collective variables using supervised machine learning. *J. Chem. Phys.* **2018**, *149*, 094106. [CrossRef] [PubMed]
106. Dror, R.O.; Dirks, R.M.; Grossman, J.P.; Xu, H.; Shaw, D.E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429–452. [CrossRef] [PubMed]
107. Ward, A.B.; Sali, A.; Wilson, I.A. Integrative Structural Biology. *Science* **2013**, *339*, 913. [CrossRef]

108. Bottaro, S.; Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **2018**, *361*, 355. [CrossRef]
109. Bonomi, M.; Heller, G.T.; Camilloni, C.; Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106–116. [CrossRef]
110. Joung, I.S.; Cheatham, T.E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041. [CrossRef]
111. Allnér, O.; Nilsson, L.; Villa, A. Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.* **2012**, *8*, 1493–1502. [CrossRef]
112. Ibragimova, G.T.; Wade, R.C. Importance of Explicit Salt Ions for Protein Stability in Molecular Dynamics Simulation. *Biophys. J.* **1998**, *74*, 2906–2911. [CrossRef]
113. Ross, G.A.; Rustenburg, A.S.; Grinaway, P.B.; Fass, J.; Chodera, J.D. Biomolecular Simulations under Realistic Macroscopic Salt Conditions. *J. Phys. Chem. B* **2018**, *122*, 5466–5486. [CrossRef] [PubMed]
114. Case, D.A. Chemical shifts in biomolecules. *Curr. Opin. Struct. Biol.* **2013**, *23*, 172–176. [CrossRef] [PubMed]
115. Tolman, J.R.; Ruan, K. NMR Residual Dipolar Couplings as Probes of Biomolecular Dynamics. *Chem. Rev.* **2006**, *106*, 1720–1736. [CrossRef] [PubMed]
116. Karplus, M. Vicinal Proton Coupling in Nuclear Magnetic Resonance. *J. Am. Chem. Soc.* **1963**, *85*, 2870–2871. [CrossRef]
117. Bernadó, P.; Mylonas, E.; Petoukhov, M.V.; Blackledge, M.; Svergun, D.I. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **2007**, *129*, 5656–5664. [CrossRef] [PubMed]
118. Jeschke, G. DEER Distance Measurements on Proteins. *Annu. Rev. Phys. Chem.* **2012**, *63*, 419–446. [CrossRef] [PubMed]
119. Piston, D.W.; Kremers, G.-J. Fluorescent protein FRET: The good, the bad and the ugly. *Trends Biochem. Sci.* **2007**, *32*, 407–414. [CrossRef]
120. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
121. Caticha, A. Relative Entropy and Inductive Inference. *Aip Conf. Proc.* **2004**, *707*, 75–96. [CrossRef]
122. Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Combining Experiments and Simulations Using the Maximum Entropy Principle. *PLoS Comput. Biol.* **2014**, *10*, e1003406. [CrossRef]
123. Bottaro, S.; Bussi, G.; Kennedy, S.D.; Turner, D.H.; Lindorff-Larsen, K. Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Sci. Adv.* **2018**, *4*, eaar8521. [CrossRef] [PubMed]
124. Różycki, B.; Kim, Y.C.; Hummer, G. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* **2011**, *19*, 109–116. [CrossRef] [PubMed]
125. Sanchez-Martinez, M.; Crehuet, R. Application of the maximum entropy principle to determine ensembles of intrinsically disordered proteins from residual dipolar couplings. *Phys. Chem. Chem. Phys.* **2014**, *16*, 26030–26039. [CrossRef] [PubMed]
126. Podbevšek, P.; Fasolo, F.; Bon, C.; Cimatti, L.; Reißer, S.; Carninci, P.; Bussi, G.; Zucchelli, S.; Plavec, J.; Gustincich, S. Structural determinants of the SINE B2 element embedded in the long non-coding RNA activator of translation AS Uchl1. *Sci. Rep.* **2018**, *8*, 3189. [CrossRef]
127. Leung, H.T.A.; Bignucolo, O.; Aregger, R.; Dames, S.A.; Mazur, A.; Bernèche, S.; Grzesiek, S. A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *J. Chem. Theory Comput.* **2016**, *12*, 383–394. [CrossRef]
128. Bradshaw, R.T.; Marinelli, F.; Faraldo-Gómez, J.D.; Forrest, L.R. Interpretation of HDX Data by Maximum-Entropy Reweighting of Simulated Structural Ensembles. *Biophys. J.* **2020**, *118*, 1649–1664. [CrossRef]
129. Shen, T.; Hamelberg, D. A statistical analysis of the precision of reweighting-based simulations. *J. Chem. Phys.* **2008**, *129*, 034103. [CrossRef]
130. Rangan, R.; Bonomi, M.; Heller, G.T.; Cesari, A.; Bussi, G.; Vendruscolo, M. Determination of Structural Ensembles of Proteins: Restraining vs Reweighting. *J. Chem. Theory Comput.* **2018**, *14*, 6632–6641. [CrossRef]
131. Pitera, J.W.; Chodera, J.D. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445–3451. [CrossRef]
132. Cesari, A.; Gil-Ley, A.; Bussi, G. Combining Simulations and Solution Experiments as a Paradigm for RNA Force Field Refinement. *J. Chem. Theory Comput.* **2016**, *12*, 6192–6200. [CrossRef]

133. Reißer, S.; Zucchelli, S.; Gustincich, S.; Bussi, G. Conformational ensembles of an RNA hairpin using molecular dynamics and sparse NMR data. *Nucleic Acids Res.* **2019**, *48*, 1164–1174. [CrossRef] [PubMed]
134. Fennen, J.; Torda, A.E.; van Gunsteren, W.F. Structure refinement with molecular dynamics and a Boltzmann-weighted ensemble. *J. Biomol. NMR* **1995**, *6*, 163–170. [CrossRef] [PubMed]
135. Best, R.B.; Vendruscolo, M. Determination of Protein Structures Consistent with NMR Order Parameters. *J. Am. Chem. Soc.* **2004**, *126*, 8090–8091. [CrossRef] [PubMed]
136. Lindorff-Larsen, K.; Best, R.B.; DePristo, M.A.; Dobson, C.M.; Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **2005**, *433*, 128–132. [CrossRef] [PubMed]
137. Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **2013**, *138*, 094112. [CrossRef]
138. Roux, B.; Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **2013**, *138*, 084107. [CrossRef]
139. Hermann, M.R.; Hub, J.S. SAXS-Restrained Ensemble Simulations of Intrinsically Disordered Proteins with Commitment to the Principle of Maximum Entropy. *J. Chem. Theory Comput.* **2019**, *15*, 5103–5115. [CrossRef]
140. Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2016**, *2*, e1501177. [CrossRef]
141. Bonomi, M.; Camilloni, C.; Vendruscolo, M. Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics. *Sci. Rep.* **2016**, *6*, 31232. [CrossRef]
142. Heller, G.T.; Aprile, F.A.; Bonomi, M.; Camilloni, C.; De Simone, A.; Vendruscolo, M. Sequence Specificity in the Entropy-Driven Binding of a Small Molecule and a Disordered Peptide. *J. Mol. Biol.* **2017**, *429*, 2772–2779. [CrossRef]
143. Hultqvist, G.; Åberg, E.; Camilloni, C.; Sundell, G.N.; Andersson, E.; Dogan, J.; Chi, C.N.; Vendruscolo, M.; Jemth, P. Emergence and evolution of an interaction between intrinsically disordered proteins. *eLife* **2017**, *6*, e16059. [CrossRef] [PubMed]
144. Buckle, E.L.; Prakash, A.; Bonomi, M.; Sampath, J.; Pfaendtner, J.; Drobny, G.P. Solid-State NMR and MD Study of the Structure of the Statherin Mutant SNa15 on Mineral Surfaces. *J. Am. Chem. Soc.* **2019**, *141*, 1998–2011. [CrossRef] [PubMed]
145. Weber, B.; Hora, M.; Kazman, P.; Göbl, C.; Camilloni, C.; Reif, B.; Buchner, J. The Antibody Light-Chain Linker Regulates Domain Orientation and Amyloidogenicity. *J. Mol. Biol.* **2018**, *430*, 4925–4940. [CrossRef] [PubMed]
146. Bonomi, M.; Pellarin, R.; Vendruscolo, M. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. *Biophys. J.* **2018**, *114*, 1604–1613. [CrossRef]
147. Vahidi, S.; Ripstein, Z.A.; Bonomi, M.; Yuwen, T.; Mabanglo, M.F.; Juravsky, J.B.; Rizzolo, K.; Velyvis, A.; Houry, W.A.; Vendruscolo, M.; et al. Reversible inhibition of the ClpP protease via an N-terminal conformational switch. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E6447. [CrossRef]
148. Paissoni, C.; Jussupow, A.; Camilloni, C. Determination of Protein Structural Ensembles by Hybrid-Resolution SAXS Restrained Molecular Dynamics. *J. Chem. Theory Comput.* **2020**, *16*, 2825–2834. [CrossRef]
149. Paissoni, C.; Jussupow, A.; Camilloni, C. Martini bead form factors for nucleic acids and their application in the refinement of protein-nucleic acid complexes against SAXS data. *J. Appl. Crystallogr.* **2019**, *52*, 394–402. [CrossRef]
150. Kooshapur, H.; Choudhury, N.R.; Simon, B.; Mühlbauer, M.; Jussupow, A.; Fernandez, N.; Jones, A.N.; Dallmann, A.; Gabel, F.; Camilloni, C.; et al. Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1. *Nat. Commun.* **2018**, *9*, 2479. [CrossRef]