molecules

MDPI

Review

Machine Learning Generation of Dynamic Protein Conformational Ensembles

Li-E Zheng [1,†], Shrishti Barethiya [2,†], Erik Nordquist [2] and Jianhan Chen [2,*]

[1] Department of Gynecology, The First Affiliated Hospital of Fujian Medical University, Fuzhou 350005, China; zelie@sina.com
[2] Department of Chemistry, University of Massachusetts Amherst, Amherst, MA 01003, USA; sbarethiya@umass.edu (S.B.); enordquist@umass.edu (E.N.)
* Correspondence: jianhanc@umass.edu
† These authors contributed equally to this work.

Abstract: Machine learning has achieved remarkable success across a broad range of scientific and engineering disciplines, particularly its use for predicting native protein structures from sequence information alone. However, biomolecules are inherently dynamic, and there is a pressing need for accurate predictions of dynamic structural ensembles across multiple functional levels. These problems range from the relatively well-defined task of predicting conformational dynamics around the native state of a protein, which traditional molecular dynamics (MD) simulations are particularly adept at handling, to generating large-scale conformational transitions connecting distinct functional states of structured proteins or numerous marginally stable states within the dynamic ensembles of intrinsically disordered proteins. Machine learning has been increasingly applied to learn low-dimensional representations of protein conformational spaces, which can then be used to drive additional MD sampling or directly generate novel conformations. These methods promise to greatly reduce the computational cost of generating dynamic protein ensembles, compared to traditional MD simulations. In this review, we examine recent progress in machine learning approaches towards generative modeling of dynamic protein ensembles and emphasize the crucial importance of integrating advances in machine learning, structural data, and physical principles to achieve these ambitious goals.

Keywords: autoencoder; Boltzmann generator; collective variable; dimension reduction; enhanced sampling; generative adversarial network; latent space; neural network; physics-informed machine learning; transfer learning

check for updates

Citation: Zheng, L.-E.; Barethiya, S.; Nordquist, E.; Chen, J. Machine Learning Generation of Dynamic Protein Conformational Ensembles. *Molecules* **2023**, *28*, 4047. https://doi.org/10.3390/molecules28104047

Academic Editor: Claudia Andreini

Received: 10 April 2023
Revised: 4 May 2023
Accepted: 9 May 2023
Published: 12 May 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Proteins are the major functional macromolecules in biology, which play critical and diverse roles in virtually all cellular processes and are involved in numerous human diseases, including cancers, neurodegenerative diseases, and diabetes [1–3]. A central property of proteins is that their amino acid sequence (and thus their chemical structure) encodes highly specific three-dimensional (3D) structural properties to support their function. Enormous efforts have been invested in experimental determination of the high-resolution structures of proteins, using a range of techniques, including nuclear magnetic resonance (NMR), X-ray crystallography, and more recently, cryogenic electron microscopy (Cryo-EM) [4,5]. These efforts have now provided an arguably complete coverage of all protein families and possible folds, with over 200,000 protein structures publicly available through the RCSB Protein Data Bank (PDB) database [6]. In parallel with these developments, dramatic advances have been made in leveraging available structures and multi-sequence alignments for the prediction of protein structure from sequence information alone [7,8]. These efforts culminated in recent development of AlphaFold [9] and RoseTTAFold [10],

which are end-to-end deep machine learning (ML) methods capable of generating high-quality structures for the entire proteomes [11]. Most recently, large language models have also emerged as powerful ML tools for discovering structural and functional properties of proteins from massive sequence databases [12]. For example, ESMfold from Meta trained with a masked language modeling objective can develop attention patterns that capture structure contacts and recover atomic protein structures that are comparable to AlphaFold2 predictions [13]. Together, these powerful tools have drastically expanded the structural coverage of proteins [6,14] and are having transformative impacts in biological and biomedical research [15,16].

Notwithstanding the remarkable successes of single protein structure prediction [17], the need for additional developments is well-recognized [18–21]. In particular, existing structure prediction tools largely aim to generate a single structure for a given sequence; yet, there is not a single "native" state for all proteins [22]. The structures of proteins can change, depending on the environment, such as changes in temperature, pH, or ligand binding, as well as post-translational modifications (PTMs) [23]. More fundamentally, proteins are dynamic in nature and their dynamic properties are essential to how proteins work in biology and how they can be targeted for therapeutic interventions [24]. NMR relaxation analysis is one of the most powerful approaches for deriving the magnitude and timescale of internal protein motions at residue level [25–27]. Multiple structures can be determined for various functional states of the same protein. Nonetheless, experimental characterization of dynamic properties and conformational transitions of proteins is challenging and severely limited in spatial and temporal resolutions [28]. Instead, physics-based molecular modeling and simulation have been the workhorses for generating ensembles of dynamic structures and conformational transition paths of proteins at atomistic resolutions [29–33]. These simulations have greatly benefited from efficient GPU-accelerated molecular dynamics (MD) algorithms [34–39], advanced sampling techniques [40–47], and steadily improved general-purpose protein force fields [48–50]. The reach of MD simulations has also been drastically expanded by the development of the special-purpose Anton supercomputers [51]. Despite these advances, a persisting bottleneck of atomistic MD simulations for generation of dynamic protein ensembles is the computational cost. In general, comprehensive sampling of the dynamic conformational ensemble is only feasible for small and simple systems. As such, there has been a long history and great need of leveraging data-driven ML methods to accelerate MD simulations and/or to directly generate dynamic protein ensembles [52–57].

In this short review, we summarize recent progresses in the development and application of deep generative models for biomolecular modeling, particularly those developed for generating dynamic structure ensembles of proteins in various contexts. We note that several outstanding reviews already provide in-depth discussions of how ML can be used to learn the energy landscape and collective variables (CVs), derive coarse-grained models, and generate protein structure ensembles [52–59]. This review will mainly focus on the most recent works published in the last three years, as identified from Web of Science searches using various combinations of keywords, including "machine learning", "protein conformation", "protein structure", and "ensemble" as well as references of related papers. We will also discuss the challenges of generative models for complex protein ensembles and how the incorporation of physical knowledge may be critical for overcoming these limitations.

2. A Rich Continuum of Protein Structures and Dynamics for Function

As illustrated in Figure 1, the range of protein conformational dynamics in nature can be roughly classified into four general categories of increasing complexity and thus difficulty for characterization and prediction. The simplest case is local conformational dynamics within a largely well-defined native fold. Such dynamics include atomic thermal fluctuations around the native structure, which measure the local rigidity. Such rigidity information can often be inferred from the crystal B-factors [60] or derived readily from short MD simulations. More importantly, certain local regions, such as loops of a protein,

can have nontrivial dynamic properties and sample a range of conformations relevant to the function. For example, the anti-apoptotic Bcl-xL protein [61] contains a BH3-only protein binding interface that adopts many different conformations within the ~50 experimental structures in PDB (Figure 1A). Atomistic simulations with enhanced sampling show that this interface is inherently dynamic and suggests many rapidly interconverting conformations [62,63]. Interestingly, all previous observed conformers are well-represented in the MD-generated ensemble, highlighting the importance of predicting and generating dynamic ensembles of local loops or regions for understanding protein function. Note that simulation of the dynamic ensemble for even a relatively modest local region is computationally intensive, requiring over 16 μs sampling time in the case of Bcl-xl, even with enhanced sampling [63].

The second major class of functional dynamics include proteins that undergo large-scale conformational transitions between two or more major states, which can be triggered by a wide range of cellular stimuli, including ligand binding, PTMs, and changes in the solution conditions (e.g., pH, temperature, and ionic strength) [64–66]. Figure 1B illustrates a drastic conformational transition of the COVID-19 spike protein trimer in the pre- and post-fusion states, as driven by interaction with the host membrane [67]. Understanding the molecular mechanisms and details of these large-scale conformational transitions is crucial for understanding protein function and for developing rational strategies of therapeutic interventions targeting these proteins. Experimentally, it may be possible to capture different conformations that correspond to various function states, but some states may require conditions difficult to replicate under structural determination conditions and these states may only be transiently accessible [65,68]. It is even more challenging to experimentally resolve the transition pathways [69,70] and molecular modeling, and simulations are generally required [71,72]. As will be discussed further, this has been one of the areas in which ML and generative models have made major impacts, especially when combined with MD simulations [52,53,58].

The third and fourth classes of functional protein dynamics include proteins that can remain partially or fully disordered under physiological conditions [73–75]. These proteins are referred to as intrinsically disordered proteins (IDPs) and are the most challenging to characterize, both experimentally and computationally. These proteins make up ~30% of all eukaryotic proteins and are key components of the regulatory networks that dictate virtually all aspects of cellular decision-making [76]. Deregulated IDPs are associated with many diseases including cancers, diabetes, and neurodegenerative and heart diseases [77–79]. Importantly, as illustrated in Figure 1C, IDPs must be described using dynamic structural ensembles. These ensembles are not random and often contain nontrivial transient local and long-range structures that are crucial to their function [80–82]. Examples are also emerging to show that IDPs can remain unstructured, even in specific complexes and functional assemblies [83–89]. Figure 1D illustrates how the N-terminal transactivation domain of tumor suppressor p53 remains highly dynamic in the specific complex with cyclophilin D, a key regulator of the mitochondrial permeability transition pore (PTP) [90]. Such a dynamic mode of specific protein interactions seems much more prevalent than previously thought [91–93]. Arguably, the key to a quantitative and predictive understanding of IDPs and their dynamic interactions is the ability to accurately describe their dynamic conformational equilibria under relevant biological contexts. Such a capability is also critical for developing effective strategies for targeting IDPs in therapeutics, where they are considered a promising but difficult new class of drug targets [94–96]. For example, the disordered C-terminal region of protein tyrosine phosphatase 1B (PT1B), a key protein in breast cancers, can be targeted by a small natural product, trodusquemine [97]. The drug's binding induces a shift in the dynamic conformational equilibrium of the C-terminal region of PT1B that allosterically disrupts HER2 signaling and inhibits tumorigenesis [98].

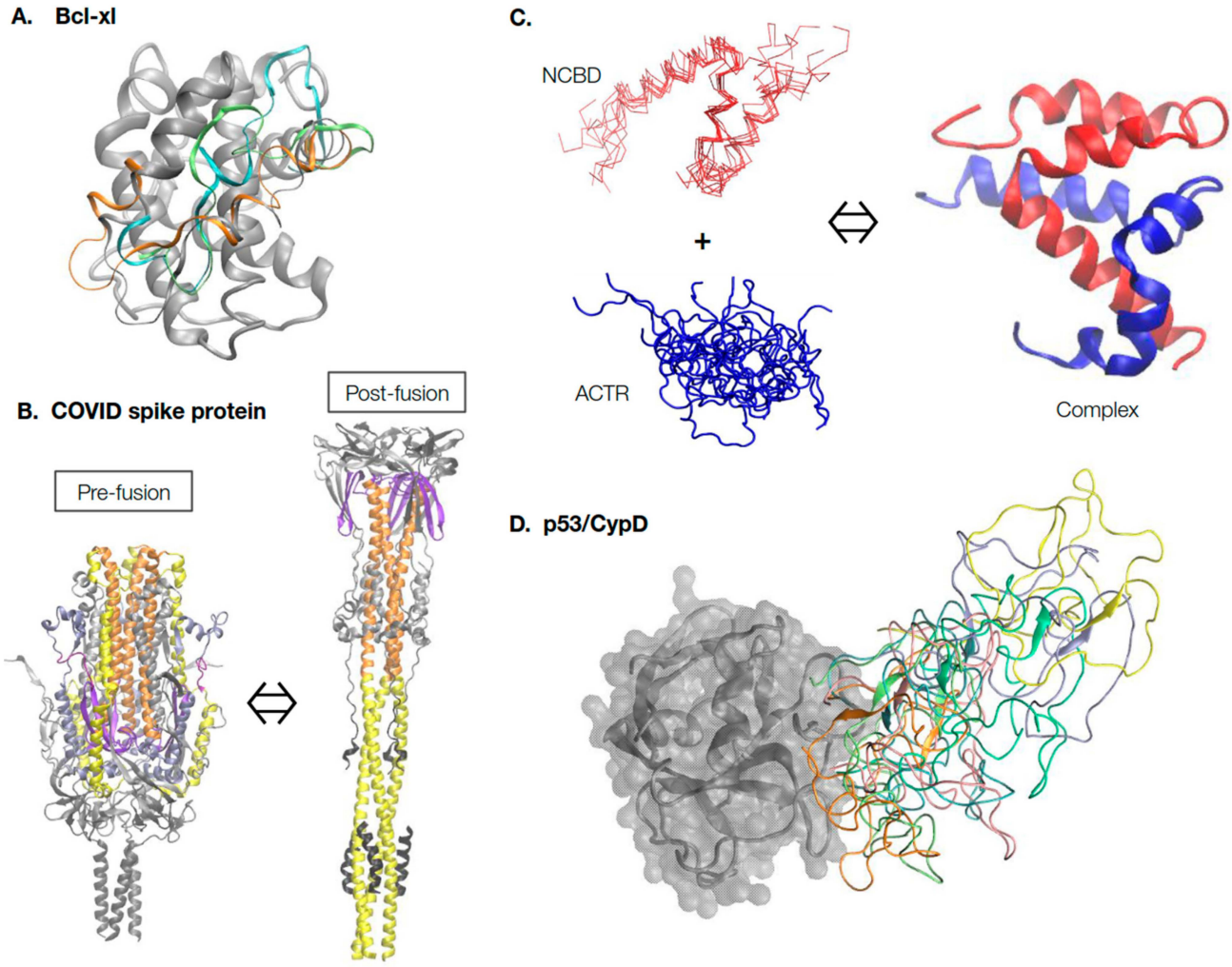

Figure 1. Continuum of protein structure and dynamics. (**A**) Inherent conformational dynamics of the BH3-only binding interface are crucial for the functioning of the Bcl-xl protein. Multiple representative conformations of the binding interface, shown in different colors, were generated using enhanced sampling simulations in explicit solvent [63]. (**B**) The COVID-19 spike protein undergoes dramatic large-scale conformational transitions in the pre-fusion and post-fusion states. The structures were extracted from Cryo-EM models (PDB: 6xr8 and 6xra [67]) and, for clarity, only common and resolved segments are shown. The central helices are shown in orange, heptad repeat 1 in yellow, and the fusion peptide proximal region in purple. Animations of the transition can be found on poteopedia.org. (**C**) Intrinsically disordered proteins ACTR and NCBD undergo binding-induced disorder-to-order transition to form the folded complex. The complex structure was taken from PDB: 1kbh [99], and the disordered ensembles of ACTR and NCBD were generated using coarse-grained (CG) MD simulations [100]. Note that while ACTR is fully disordered, free NCBD is a molten globule with essentially fully-formed helices. (**D**) Dynamic interactions of the N-terminal domain (NTD) of tumor suppressor p53 with the folded mitochondrial PTP regulator protein Cyclophilin D (CypD). CypD is shown in gray; multiple dynamic conformations of p53 NTD were extracted from previous CG MD simulations [90] and shown in different colors.

3. Generative Deep Learning for Biomolecular Modeling

Generative deep learning models are a class of neural networks that aim to learn the regularities of the training dataset and capture such regularities using appropriate probability distribution functions [54,101]. The learned probability distribution functions should be smooth, and can thus be used for generating new samples that are similar to the original dataset, with generally low computational costs. For biomolecular modeling, latent variable models such as variational autoencoders (VAEs) [102,103] and generative adversarial networks (GANs) [104] have been particularly suitable due to lower computational cost compared to sequential autoregressive models [105]. In both VAE and GAN models, the goal is to learn a lower-dimensional representation of the data in the so-called latent space, either explicitly (in VAEs) or implicitly (in GANs). VAEs learn a probabilistic encoder-decoder model that maps the input data to the latent space and back, while GANs train a generator that draws new samples from the latent space, and a discriminator that distinguishes between real and fake samples. The performance of both VAEs and GANs critically depends on careful training and tuning of the model hyperparameters, such that a good latent space representation exists and can be reliably learned.

Briefly, autoencoders (AE) are unsupervised ML algorithms that use neural network (NN) architecture to compress input data into knowledge (latent) representations. As illustrated in Figure 2A, an AE consists of two sub-models: an encoder and a decoder. The encoder is a generally a convolutional or feed-forward NN that reduces the high-dimensional input to lower-dimensional latent space, and the decoder NN transforms the latent space back to the original high-dimensional output. In other words, latent space is the input for the decoder. The number of encoder layers and decoder layers are generally the same. The AE is trained by optimizing the loss function to minimize the loss between the original and reconstructed outputs. For proteins, the input can be a set of Cartesian or internal coordinates. After training, the latent representation contains an information-dense and relatively lossless projection of the ensemble. Such latent spaces can be utilized in multiple ways. They could be used as CVs for calculating free energy surfaces, constructing Markov state models, or guiding additional MD simulations [106]. They could also be directly used for direct generation of new structures of the same biomolecule. For the latter, VAEs are often used where the distribution in the latent space is constrained to have a smooth normal distribution. GANs are also a type of unsupervised deep learning model; they have gained significant attention in recent years due to their ability to generate new data that is similar to a given training dataset. GANs also consist of two deep neural networks, namely a generator and a discriminator (Figure 2B). Generators are trained to generate new structures of the biomolecules by sampling from the prior distribution, whereas discriminators classify the generated structure as fake versus real. Both the sub-models are trained together in such a way that the generator produces structures that can fool the discriminator and discriminator is iteratively updated in order to accurately classify the structures.

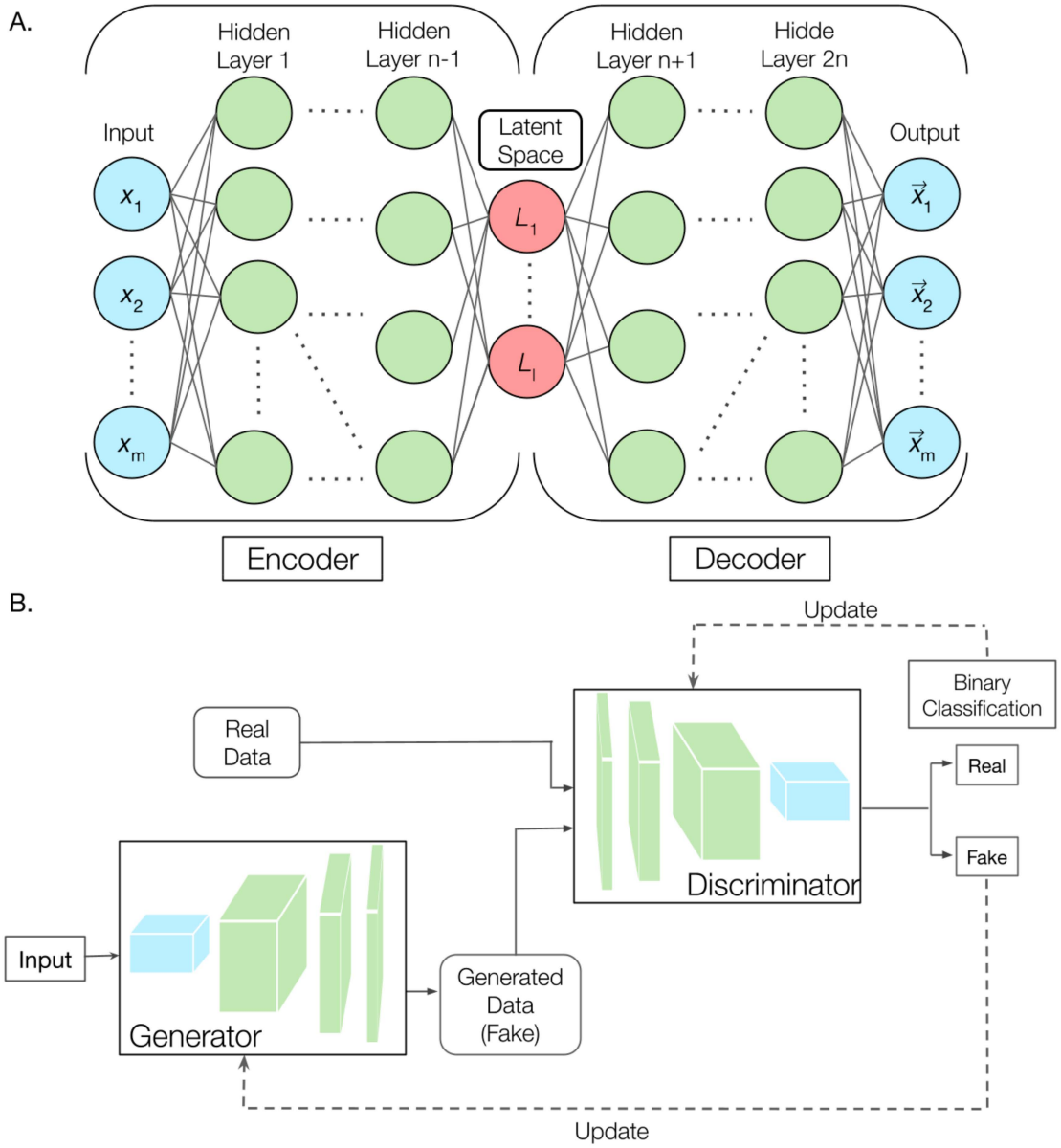

Figure 2. Generative ML approaches for protein dynamics. (**A**) An autoencoder consists of two fully-connected NN. The encoder compresses the m-dimensional input data vector (x_1, x_2, . . . , x_m) in several orders in dimensions into the l-dimensional latent space vector (L_1, L_2, . . . , L_l, $l << m$), while the encoder decompresses the latent space and reconstructs the original high-dimensional form ($\vec{x_1}, \vec{x_2}, \ldots, \vec{x_m}$). (**B**) GANs contain two networks. The generator takes input that is of a predefined latent space and generates a similar output to realistic data using a deconvolutional neural network. The discriminator is a convolutional neural network that classifies whether the data is real or fake. The generator is further updated by using this discriminator output to perform better in the next epoch in order to fool the discriminator, and the discriminator also backpropagates to update and reduce loss in order to more accurately classify the data.

4. ML Approaches to Identify CVs and Drive Enhanced Sampling in MD Simulations

A key application for ML in the generation of protein functional ensembles is to aid in the discovery of low-dimensional CVs that can distinguish important functional states, particularly those that could monotonically describe the transition between the states and are suitable for usage in enhanced sampling MD simulations such as umbrella sampling, adaptive sampling, and metadynamics [52,59]. Even though such approaches are not strictly generative models, a key advantage of integrating ML and MD is that the resulting

conformations can be unbiased in order to yield proper thermodynamic ensembles. To be useful for this purpose, CVs must be of low enough dimension and, more critically, capture the slowest fluctuating degree(s) of freedom of protein conformation transitions [107]. The later requirement is highly nontrivial, in order to account for complex diffusive protein dynamics. Slow dynamics in other degrees of freedom not captured in the CVs leads to hidden barriers and hinders MD sampling [108]. It remains debatable if high-dimensional conformational fluctuation of biological proteins could be effectively reduced to representation using a few "reaction coordinates" (RCs) [109]. As such, data-driven discovery of CVs for protein dynamics remains an active area of research in recent years [110–115]. In the following, we highlight a few representative deep learning approaches in this direction.

In the so-called deeply enhanced sampling of protein (DESP) method developed by Salawu [110], a neural network (NN) is trained alongside the MD simulations to learn the latent space for representing the conformational space sampled. Biasing potentials are then introduced in the latent space, using the Kullback-Leibler (KL) divergence, to discourage MD from revisiting the conformational space already sampled. This approach was evaluated using a model helical peptide A_{12} and a small protein GB98. The results suggest that DESP can sample much broader conformational space compared to both conventional and accelerated MD within the same amount of simulation time. Importantly, unbiased probability distributions also could be uncovered. DESP is in principle generalizable to other proteins. However, additional work is required to understand how the quality of the latent space learned, as well as the choice of biasing potentials, affect the sampling efficiency. Further analysis is also required to evaluate how accurate and/or physical the additional conformational space sampled by DESP is. The apparent lack of reversible transitions in the DESP trajectories is concerning, which may reflect a critical limitation of using KL divergence directly to drive the sampling of new conformational space. Tao and co-workers [114] have compared the efficacies of AE and VAE for driving MD exploration of protein conformational space. Here, random points were selected in the latent space to initiate new MD trajectories. Using the adenosine kinase as a model system, it was shown that the latent space learned in VAE is superior to the AE-derived one for generating unsampled conformational space, likely due to the normal and smoother latent space distribution in VAE.

Tiwary and co-workers recently adapted reweighted auto-encoded variational Bayes for enhanced sampling (RAVE) for efficient sampling of protein loop conformations [112]. RAVE is based on the principle of the predictive information bottleneck (PIB), a predictive model for describing the evolution of a given dynamical system that encodes high dimensional input into low dimensional representations. PIB can be learned in an iterative manner, similar to autoencoders, and interpreted as the RC for usage in metadynamics enhanced sampling. The input to RAVE was generated using the automatic mutual information noise omission (AMINO) method to reduce the redundancies among a large set of raw order parameters (OP) that reflect generic features of protein contacts. The RC in RAVE is constructed as a linear combination of selected bias functions of the OP output from AMINO. Metadynamics trajectories generated using the RC as a bias variable are fed back to RAVE to further optimize the RC. This iterative process between enhanced sampling and RAVE learning continues until multiple transitions between different metastable states are sampled. Applied to protein T4 lysozyme, it was found that the functional states and free energy surfaces generated using the above protocol successfully recapitulate the loop conformational stabilities of the wild-type enzyme and three mutants. Furthermore, it was observed that the number of OPs required for the RC decreases in the mutants, suggesting increased cooperativity of conformational fluctuation (and thus decreased global flexibility). This work represents an interesting development towards an automated procedure for atomistic simulations of protein loop dynamics. It is not clear, though, whether the RAVE/AMINO approach can be generalized in order to sample either large-scale conformational transitions or disordered protein conformational ensembles (classes 2–4 of Figure 1).

Reinforced learning (RL)-based algorithms [116,117] have also been adapted for promoting the exploration of slow degrees of freedom of complex protein conformational fluctuations. Recently, a multiagent RL-based adaptive sampling (REAP) approach has been designed for sampling of rare states along user-defined CVs [113]. REAP is initialized by running short MD simulations, followed by conformational clustering to discretize the action space. The smallest clusters are selected as candidates. The reward is calculated for each cluster and optimization is performed by summing the reward as a weighted sum of the candidate CVs. Conformations with the highest rewards initiate the new set of simulations, minimizing redundant exploration. The process is repeated until either convergence or the desired final state is reached. REAP is effective if the user-defined CVs capture well the range of conformational space to be sampled. Multiagent LEAP allows more effective sampling by learning from independent simulations initiated from distinct starting conformations. This is achieved by introduction of a *stakes* function to modulate how rewards are attributed to different agents for discovering new states. The benefit of this multiagent formulation comes from two features. Conformations are labeled and utilized by the agent who discovers them, and therefore, each agent computes the rewards from different data points. Additionally, the agents share information within an action space to tell other agents what conformations they have already discovered. It was demonstrated that the multiagent REAP algorithm is more efficient in sampling the loop motion of Src kinase and driving large-scale conformational transitions of the transporter OsSWEET2b. The same multiagent formalism was shown to work well with other adaptive sampling techniques, including "least counts" and AdaptiveBandit. A key limitation of both single-agent and multi-agent REAP algorithms, though, is their dependence on user-defined CVs for prioritizing sampling. As illustrated in the works above, identification of such CVs can be highly nontrivial for protein conformational fluctuations in general. It is also not clear how one can reconstruct unbiased ensembles from REAP trajectories, a capacity which will be important for functional studies. The pros and cons of REAP and related adaptive sampling techniques remain to be thoroughly examined, especially compared to weighted ensemble methods that can recover both kinetic rates and thermodynamic stabilities of long-timescale processes [118].

5. Directly Sampling Conformational Space using ML-Derived Latent Representation

Learning low-dimensional representations of protein conformational ensembles is of interest for more than driving MD sampling. In principle, if one can learn the latent space representation from a relatively limit set of conformations, (e.g., generated by short MD simulations), one could then sample in the latent space to directly decode and generate new high-dimensional structures of the same protein. The requirement here is that the latent space representation is smooth and continuous. Such a generative approach can be dramatically faster than MD simulations. Nonetheless, the complexity and sheer size of the conformational space should never be under-appreciated, even for small proteins. Functionally relevant conformational states are vanishingly small in comparison, and importantly, they can be well-separated in the Cartesian space. It is not entirely clear if the desired mapping of high-dimensional protein conformational space to a smooth low-dimensional latent space representation exists, or, if such a mapping can be reliably learned from an incomplete set of pre-generated conformations. Nonetheless, direct sampling of protein conformational space using ML-derived latent representation is a highly attractive strategy and has continued to attract intense interest in recent years [119–123].

The ability of using the latent space to generate physically plausible structures has been evaluated using a set of proteins of different sizes, topologies and dynamic properties [119]. The results suggest that over 98% of reconstructed structures generated by AE for rigid proteins were classified as valid by a random forest (RF) classifier. For flexible proteins with multiple functional states, it was observed that VAE trained using both open and closed conformations could provide a reliable interpolation between these states. However, if only one of the states is available, VAE would fail to provide accurate extrapolation to

distinct states not seen in the training set. This observation highlights the challenges of using latent presentation in direct sampling of novel conformational spaces of complex biomolecules. Hay and co-workers [121] tested whether an AE could be used to map MD generated conformations onto a pre-defined low-dimensional space (e.g., the first and second principal components) for subsequent prediction of new conformations, as a proof-of-principle. Using a flexible short peptide Ala_{13} and the protein calmodulin (CaM), the approach demonstrated modest success and could be most suitable for generating new initial structures for seeding additional MD simulations. Using pre-defined latent space is unusual and likely will be ineffective in general.

Degiacomi and co-workers [120] later described a 1D convolutional NN (CNN) that was directly trainable with protein structures to learn a latent representation of the conformational space. To address the challenge of extrapolating to novel conformational spaces, a key development here was the design of a new physics-based loss function that resembled the classical molecular mechanics force field. The loss function contained physics-motivated terms to enforce the covalent geometry and minimize steric clashes. Applied to a protein enzyme MurD with multiple open, close and intermediate states available, it was demonstrated that the CNN trained with physical constraints was capable of predicting the correct transition paths between the open and close states without any intermediate conformation provided. Intriguingly, the authors further showed that they could transfer features learned from one protein to others, which dramatically reduces the number of training samples for one protein and provides superior performance in generating novel low-energy conformations. It will be very interesting to see how this approach can be extended to generate dynamic protein ensembles in general, such as those of dynamic loops or IDPs.

Success has also been demonstrated recently using VAE to learn the dynamic conformational space of IDPs using short MD simulations and then generate full disordered ensembles [123]. Here, the objective was to use the smallest amount of MD sampling and generate the most accurate full ensembles, equivalent to those derived from much longer MD simulations. Three IDPs, namely, Q15, Aβ40, and ChiZ, were used as model systems where multi-μs MD trajectories had been previously generated. The dimension of the latent space was empirically chosen to be 0.75 N_{res} (number of residues). It was shown that only ~ the first 10% of the MD trajectories were necessary to regenerate the full ensemble of small IDP Q15, but the amount of training data required increased rapidly for larger IDPs such as ChiZ. The authors also evaluated the effects of latent space dimension and input features (e.g., dihedral angles instead of Cartesian coordinates). The results showed that larger latent vectors do not improve accuracy and the choice of input coordinates does not significantly impact model performance. Curiously, evaluation of VAE performance has focused on the accuracy of regenerating individual conformers, as measured by root-mean-square-distance (RMSD). It would be more desirable to examine ensemble distributions of key local and global properties, such as overall size, residual structures, long-range contacts, etc. One could also examine the details of conformational substates using clustering and principal component analysis, in comparison to the available ensembles generated by much longer MD trajectories. Furthermore, the three IDP ensembles appear largely random and devoid of nontrivial local structures and transient long-range organizations. Yet, many biologically relevant IDPs, especially those involved in cellular signaling and regulation, are not random coils, and contain important residual structures [124,125]. As such, it remains to be established whether such relatively standard VAE framework would be adequate for generative modeling of IDPs in general.

6. One-Shot Generation of Dynamic Protein Conformational Ensembles

An ultimate "one-shot" generative ML model would take the sequence of the protein and generate a full ensemble of the most relevant conformations. This is an extremely ambitious goal for proteins with high-dimensional and complex conformation space. Progress towards this goal has thus far been more limited [126–128]. Boltzmann generators are a

type of generative model directly trained on the energy function of the system [127]. They can learn to sample from equilibrium distributions without directly learning the probability density function (e.g., from short MD trajectories). Instead, it learns to sample from a dimensionless energy function u(**x**) using a generative network and reweighting procedures. The generative network maps latent space samples from a simple prior distribution P(**z**) (e.g., Gaussian) to high-probability samples from the target distribution $P(\mathbf{x}) \sim e^{-u(\mathbf{x})}$. The probability of generating a configuration can be computed using the change-of-variables equation if the generative network is an invertible transformation. Invertible networks are called flows, and they can be stacked to create deep invertible neural networks. Even though the loss function can be designed to balance the sampling of low-energy states and the diversity of the sampled conformational space, Boltzmann generators show a strong tendency to suffer from mode collapse and generate similar conformations from a single metastable state. Instead, it needs to be combined with training by example, from existing experimental structures or short MD simulations. Additional terms can also be added to drive Boltzmann generators to sample along a pre-defined RC. This framework has mainly been demonstrated on relatively simple toy systems. Application to complex molecules such as proteins will lead to unrealistic structures with distorted covalent geometries and severe atomic overlaps. Instead, the technique requires careful separation of various degrees of freedom into Cartesian (backbone) and internal (sidechains) coordinate sets. It was demonstrated that the eventual generator could sample a key X-O loop conformational transition of protein BPTI that occurs on a millisecond-timescale. Clearly, extension of Boltzmann generators for one-shot generation of dynamic protein conformational ensembles requires much additional work. It has been argued that training solely on the energy is unlikely to be adequate [128].

Feig and co-workers recently described a conditional generative model for generating disordered protein conformational ensembles using the sequence alone as input [126]. The model, referred to as IdpGAN, is trained with long MD trajectories of a large set of IDPs, using a standard GAN architecture with multilayer perceptrons. The feasibility of IdpGAN is first demonstrated using trajectories from coarse-grained (CG) MD simulations. The results show that IdpGAN is highly effective in generating realistic disordered ensembles for an arbitrary IDP sequence, as characterized by a number of metrics, including protein overall dimension, contact distributions, and correlation of multiple pairs. The approach was further demonstrated by retraining the model using all-atom implicit solvent simulations. Impressively, conformational ensembles generated by IdpGAN do not merely capture overall structural properties such as compaction, but also contain realistic sequence-specific local structures such as residual helices. This is a highly nontrivial result that highlights the great potential of IdpGAN. It has been noted that IdpGAN relies on training on exhaustive trajectories of a carefully curated set of IDPs. This is a requirement that may be extremely difficult to meet for larger IDPs with nontrivial local structures and transient long-range interactions (such as proteins p53 [129] and tau [130]). It will be interesting to further evaluate the resolution of IdpGAN, for example, to resolve the effects of mutations of PTMs at one or few sites, which may be important in studies of regulatory IDPs.

7. Conclusions and Future Directions

Recent breakthroughs in ML approaches have transformed the studies of protein structure and function. The problem of predicting a native structure of protein has essentially been solved. A key new frontier is expanding these ML approaches to help describe dynamic fluctuations and transitions of proteins that are crucial to its functions. For this, much effort has been dedicated to generative ML models such as VAEs and GANs (Figure 2). A common objective of these generative models is to learn a low-dimensional latent representation of the high-dimensional conformational space of the protein, which can then be used to guide additional MD sampling and construct free energy surfaces. If restrained to be smooth and continuous, the latent space could also be used for direct generation of novel conformations not seen in the training sample. A fundamental challenge,

however, is that the possible conformational space of proteins is vast, and functionally relevant states are vanishingly small and often well segregated. It remains to be established to what degree the dimensionality of protein conformational space could be reduced, whether the complex and discontinuous distributions of low-energy states could be mapped to smooth and continuous ones in the latent space, and how the mapping can be reliably learned. As such, only limited successes have been demonstrated at this point. Generative modeling of dynamic protein conformations has only been feasible in relatively simple cases such as protein loop motions, or when sufficient examples are available for training, such as for relatively simple transitions between to known states or for highly disordered proteins with limited nontrivial local and global structural features.

It is generally recognized that it is likely critical to incorporate physics towards generative ML models that may be more generally applicable to nontrivial proteins [106,131–134]. For example, Boltzmann generators are directly trained on the energy functions of the system to generate independent samples of low-energy states, even though it is also evident that training on energy alone is unlikely to be adequate for complex biomolecules [127,128]. It should be noted that the physical principles of molecular interactions arguably contain all the information needed for generating the complete conformational ensemble of any protein, though this method is generally impractical due to its computational cost. On the other hand, the large number of available experimental structures, as well as numerous high-quality MD trajectories on many proteins, may already contain enough of the information needed to train transferable ML models for generating dynamic protein conformational ensembles, given the sequence. Developing such models will require advances on multiple fronts.

1. Deeper and more systematic understanding of if and how the complexity of the high-dimensional conformational space of proteins can be mapped to a smooth and continuous space of sufficiently low dimensionality. This may be investigated using ultra-long trajectories of a large set of folded and unfolded proteins, such as those from Shaw and co-workers, using the latest balanced atomistic protein force field [49], as well as those generated using extensive enhanced sampling simulations [124,125,135]. A challenge here is to ensure the quality and convergence of the structural data, particularly for IDPs.
2. Consistent and rigorous evaluation of the performance of generative ML models that is widely accepted and adopted by the community. At present, most developments are benchmarked using different custom examples and special test-cases. There is only minimal cross-comparison needed to rigorously establish the strengths and weaknesses of various approaches, such as over-fitting, computational cost, and interpretability. Again, ultra-long MD trajectories or highly-converged conformational ensembles generated from enhanced sampling simulations of proteins of different sizes, topologies and dynamic properties could be used as a standard benchmark set. Similar practices have been instrumental in the development of methods for predicting protein structure [8], protein-protein interaction [136], and protein-ligand interaction [137].
3. New ML approaches for integrating information from diverse datasets, such as protein structures, sequence alignment, and MD trajectories, and incorporating physical principles of molecular interactions, such as various empirical protein energy functions. ML has been under accelerated development in recent years, with many exciting ideas emerging, such as large language models (LLMs) [12], deep transfer learning [138], and diffusion models [139]. Furthermore, the computational infrastructure now allows much larger and increasingly complex models to be trained with extremely large datasets [140].

At present, it is hard to clearly envision how generative ML models of protein conformational ensemble may look or how generally applicable these models will be across the spectrum of protein dynamics relevant to biological function (Figure 1). However, development of ML and artificial intelligence tools has rapidly accelerated in recent years,

and the trend should continue in the foreseeable future. Notwithstanding the challenges described in this review, generative modeling of protein conformational dynamics has immense potential to completely transform how we study protein structure, dynamics, and function in biology and medicine, and therefore offers a bright future for the field.

Author Contributions: Writing—drafting and editing, L.-E.Z., S.B. and J.C.; writing—review and editing, S.B. and E.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by National Institutes of Health grant R35 GM144045 (to J.C.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank Xiaorong Liu and Yumeng Zhang for providing protein structural ensembles used in Figure 1. The authors also thank anonymous reviewers for critical suggestions.

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

Sample Availability: Not applicable.

References

1. Bushweller, J.H. Targeting transcription factors in cancer-from undruggable to reality. *Nat. Rev. Cancer* **2019**, *19*, 611–624. [CrossRef] [PubMed]
2. Kastenhuber, E.R.; Lowe, S.W. Putting p53 in Context. *Cell* **2017**, *170*, 1062–1078. [CrossRef] [PubMed]
3. Berlow, R.B.; Dyson, H.J.; Wright, P.E. Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation. *J. Mol. Biol.* **2018**, *430*, 2309–2320. [CrossRef] [PubMed]
4. Yip, K.M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **2020**, *587*, 157–161. [CrossRef] [PubMed]
5. Rout, M.P.; Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **2019**, *177*, 1384–1403. [CrossRef]
6. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P.A.; Crichlow, G.V.; Dalenberg, K.; Duarte, J.M.; et al. RCSB Protein Data Bank (RCSB.org): Delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.* **2023**, *51*, D488–D508. [CrossRef]
7. Kuhlman, B.; Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 681–697. [CrossRef]
8. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* **2021**, *89*, 1607–1617. [CrossRef]
9. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
10. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. [CrossRef]
11. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Zidek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly accurate protein structure prediction for the human proteome. *Nature* **2021**, *596*, 590–596. [CrossRef]
12. Bepler, T.; Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* **2021**, *12*, 654–669.e3. [CrossRef] [PubMed]
13. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv* **2022**. [CrossRef]
14. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439–D444. [CrossRef] [PubMed]
15. Thornton, J.M.; Laskowski, R.A.; Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* **2021**, *27*, 1666–1669. [CrossRef] [PubMed]
16. Borkakoti, N.; Thornton, J.M. AlphaFold2 protein structure prediction: Implications for drug discovery. *Curr. Opin. Struct. Biol.* **2023**, *78*, 102526. [CrossRef]
17. Pearce, R.; Zhang, Y. Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* **2021**, *297*, 100870. [CrossRef]
18. Lane, T.J. Protein structure prediction has reached the single-structure frontier. *Nat. Methods* **2023**, *20*, 170–173. [CrossRef]
19. Moore, P.B.; Hendrickson, W.A.; Henderson, R.; Brunger, A.T. The protein-folding problem: Not yet solved. *Science* **2022**, *375*, 507. [CrossRef]

20. Ourmazd, A.; Moffat, K.; Lattman, E.E. Structural biology is solved—Now what? *Nat. Methods* **2022**, *19*, 24–26. [CrossRef]
21. Ruff, K.M.; Pappu, R.V. AlphaFold and Implications for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433*, 167208. [CrossRef] [PubMed]
22. Frauenfelder, H.; Sligar, S.G.; Wolynes, P.G. The energy landscapes and motions of proteins. *Science* **1991**, *254*, 1598–1603. [CrossRef] [PubMed]
23. Guo, J.; Zhou, H.X. Protein Allostery and Conformational Dynamics. *Chem. Rev.* **2016**, *116*, 6503–6515. [CrossRef] [PubMed]
24. Miller, M.D.; Phillips, G.N., Jr. Moving beyond static snapshots: Protein dynamics and the Protein Data Bank. *J. Biol. Chem.* **2021**, *296*, 100749. [CrossRef]
25. Sugase, K.; Konuma, T.; Lansing, J.C.; Wright, P.E. Fast and accurate fitting of relaxation dispersion data using the flexible software package GLOVE. *J. Biomol. NMR* **2013**, *56*, 275–283. [CrossRef]
26. Palmer, A.G., 3rd. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* **2004**, *104*, 3623–3640. [CrossRef]
27. 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]
28. Bonomi, M.; Heller, G.T.; Camilloni, C.; Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106–116. [CrossRef]
29. Lane, T.J.; Shukla, D.; Beauchamp, K.A.; Pande, V.S. To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58–65. [CrossRef]
30. 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]
31. Best, R.B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2017**, *42*, 147–154. [CrossRef] [PubMed]
32. Mackerell, A.D. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584–1604. [CrossRef]
33. Gong, X.; Zhang, Y.; Chen, J. Advanced Sampling Methods for Multiscale Simulation of Disordered Proteins and Dynamic Interactions. *Biomolecules* **2021**, *11*, 1416. [CrossRef]
34. Brooks, B.R.; Brooks, C.L.; Mackerell, A.D.; Nilsson, L.; Petrella, R.J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614. [CrossRef] [PubMed]
35. Case, D.A.; Cheatham, T.E., III; Darden, T.A.; Duke, R.E.; Giese, T.J.; Gohlke, H.; Goetz, A.W.; Greene, D.; Homeyer, N.; Izadi, S.; et al. *AMBER 2017*; University of California: San Francisco, CA, USA, 2017.
36. Eastman, P.; Friedrichs, M.S.; Chodera, J.D.; Radmer, R.J.; Bruns, C.M.; Ku, J.P.; Beauchamp, K.A.; Lane, T.J.; Wang, L.-P.; Shukla, D.; et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2012**, *9*, 461–469. [CrossRef] [PubMed]
37. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. [CrossRef]
38. Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kal, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802. [CrossRef]
39. Gotz, A.W.; Williamson, M.J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R.C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542–1555. [CrossRef]
40. Zhang, W.H.; Chen, J.H. Accelerate Sampling in Atomistic Energy Landscapes Using Topology-Based Coarse-Grained Models. *J. Chem. Theory Comput.* **2014**, *10*, 918–923. [CrossRef]
41. Moritsugu, K.; Terada, T.; Kidera, A. Scalable free energy calculation of proteins via multiscale essential sampling. *J. Chem. Phys.* **2010**, *133*, 224105. [CrossRef]
42. Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151. [CrossRef]
43. Liu, P.; Kim, B.; Friesner, R.A.; Berne, B.J. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13749–13754. [CrossRef]
44. Mittal, A.; Lyle, N.; Harmon, T.S.; Pappu, R.V. Hamiltonian Switch Metropolis Monte Carlo Simulations for Improved Conformational Sampling of Intrinsically Disordered Regions Tethered to Ordered Domains of Proteins. *J. Chem. Theory Comput.* **2014**, *10*, 3550–3562. [CrossRef]
45. Peter, E.K.; Shea, J.E. A hybrid MD-kMC algorithm for folding proteins in explicit solvent. *Phys. Chem. Chem. Phys.* **2014**, *16*, 6430–6440. [CrossRef]
46. Zhang, C.; Ma, J. Enhanced sampling and applications in protein folding in explicit solvent. *J. Chem. Phys.* **2010**, *132*, 244101. [CrossRef]
47. Zheng, L.Q.; Yang, W. Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering. *J. Chem. Theory Comput.* **2012**, *8*, 810–823. [CrossRef]
48. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.L.; Grubmuller, H.; MacKerell, A.D., Jr. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71–73. [CrossRef]
49. Robustelli, P.; Piana, S.; Shaw, D.E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E4758–E4766. [CrossRef]

50. Best, R.B.; Zheng, W.; Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **2014**, *10*, 5113–5124. [CrossRef]
51. Shaw, D.E.; Grossman, J.P.; Bank, J.A.; Batson, B.; Butts, J.A.; Chao, J.C.; Deneroff, M.M.; Dror, R.O.; Even, A.; Fenton, C.H.; et al. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In Proceedings of the SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, New Orleans, LA, USA, 16–21 November 2014; pp. 41–53.
52. Gkeka, P.; Stoltz, G.; Barati Farimani, A.; Belkacemi, Z.; Ceriotti, M.; Chodera, J.D.; Dinner, A.R.; Ferguson, A.L.; Maillet, J.B.; Minoux, H.; et al. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *J. Chem. Theory Comput.* **2020**, *16*, 4757–4775. [CrossRef]
53. Chen, M. Collective variable-based enhanced sampling and machine learning. *Eur. Phys. J. B* **2021**, *94*, 211. [CrossRef] [PubMed]
54. Hoseini, P.; Zhao, L.; Shehu, A. Generative deep learning for macromolecular structure and dynamics. *Curr. Opin. Struct. Biol.* **2021**, *67*, 170–177. [CrossRef] [PubMed]
55. Lindorff-Larsen, K.; Kragelund, B.B. On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins. *J. Mol. Biol.* **2021**, *433*, 167196. [CrossRef]
56. Ramanathan, A.; Ma, H.; Parvatikar, A.; Chennubhotla, S.C. Artificial intelligence techniques for integrative structural biology of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2021**, *66*, 216–224. [CrossRef]
57. Wang, Y.; Lamim Ribeiro, J.M.; Tiwary, P. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2020**, *61*, 139–145. [CrossRef]
58. Allison, J.R. Computational methods for exploring protein conformations. *Biochem. Soc. Trans.* **2020**, *48*, 1707–1724. [CrossRef]
59. Noe, F.; De Fabritiis, G.; Clementi, C. Machine learning for protein folding and dynamics. *Curr. Opin. Struct. Biol.* **2020**, *60*, 77–84. [CrossRef] [PubMed]
60. Sun, Z.; Liu, Q.; Qu, G.; Feng, Y.; Reetz, M.T. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem. Rev.* **2019**, *119*, 1626–1665. [CrossRef]
61. Delbridge, A.R.D.; Grabow, S.; Strasser, A.; Vaux, D.L. Thirty years of BCL-2: Translating cell death discoveries into novel cancer therapies. *Nat. Rev. Cancer* **2016**, *16*, 99–109. [CrossRef]
62. Liu, X.; Beugelsdijk, A.; Chen, J. Dynamics of the BH3-Only Protein Binding Interface of Bcl-xL. *Biophys. J.* **2015**, *109*, 1049–1057. [CrossRef]
63. Liu, X.R.; Jia, Z.G.; Chen, J.H. Enhanced Sampling of Intrinsic Structural Heterogeneity of the BH3-Only Protein Binding Interface of Bcl-xL. *J. Phys. Chem. B* **2017**, *121*, 9160–9168. [CrossRef] [PubMed]
64. Orellana, L. Large-Scale Conformational Changes and Protein Function: Breaking the in silico Barrier. *Front. Mol. Biosci.* **2019**, *6*, 117. [CrossRef] [PubMed]
65. Korzhnev, D.M.; Kay, L.E. Probing invisible, low-populated States of protein molecules by relaxation dispersion NMR spectroscopy: An application to protein folding. *Acc. Chem. Res.* **2008**, *41*, 442–451. [CrossRef] [PubMed]
66. Noe, F.; Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162. [CrossRef]
67. Cai, Y.; Zhang, J.; Xiao, T.; Peng, H.; Sterling, S.M.; Walsh, R.M., Jr.; Rawson, S.; Rits-Volloch, S.; Chen, B. Distinct conformational states of SARS-CoV-2 spike protein. *Science* **2020**, *369*, 1586–1592. [CrossRef]
68. de Lera Ruiz, M.; Kraus, R.L. Voltage-Gated Sodium Channels: Structure, Function, Pharmacology, and Clinical Indications. *J. Med. Chem.* **2015**, *58*, 7093–7118. [CrossRef]
69. Rangl, M.; Schmandt, N.; Perozo, E.; Scheuring, S. Real time dynamics of Gating-Related conformational changes in CorA. *Elife* **2019**, *8*, e47322. [CrossRef]
70. Chung, H.S.; Eaton, W.A. Protein folding transition path times from single molecule FRET. *Curr. Opin. Struct. Biol.* **2018**, *48*, 30–39. [CrossRef]
71. Sands, Z.; Grottesi, A.; Sansom, M.S. Voltage-gated ion channels. *Curr. Biol.* **2005**, *15*, R44–R47. [CrossRef]
72. Jensen, M.O.; Jogini, V.; Borhani, D.W.; Leffler, A.E.; Dror, R.O.; Shaw, D.E. Mechanism of voltage gating in potassium channels. *Science* **2012**, *336*, 229–233. [CrossRef] [PubMed]
73. Wright, P.E.; Dyson, H.J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29. [CrossRef] [PubMed]
74. Uversky, V.N. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. *Front. Phys.-Lausanne* **2019**, *7*, 10. [CrossRef]
75. Hatos, A.; Hajdu-Soltesz, B.; Monzon, A.M.; Palopoli, N.; Alvarez, L.; Aykac-Fas, B.; Bassot, C.; Benitez, G.I.; Bevilacqua, M.; Chasapi, A.; et al. DisProt: Intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **2020**, *48*, D269–D276. [CrossRef] [PubMed]
76. Iakoucheva, L.M.; Brown, C.J.; Lawson, J.D.; Obradovic, Z.; Dunker, A.K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **2002**, *323*, 573–584. [CrossRef]
77. Uversky, V.N.; Oldfield, C.J.; Dunker, A.K. Intrinsically disordered proteins in human diseases: Introducing the D-2 concept. *Annu. Rev. Biophys.* **2008**, *37*, 215–246. [CrossRef]
78. Tsafou, K.; Tiwari, P.B.; Forman-Kay, J.D.; Metallo, S.J.; Toretsky, J.A. Targeting Intrinsically Disordered Transcription Factors: Changing the Paradigm. *J. Mol. Biol.* **2018**, *430*, 2321–2341. [CrossRef]

79. Giri, R.; Kumar, D.; Sharma, N.; Uversky, V.N. Intrinsically Disordered Side of the Zika Virus Proteome. *Front. Cell. Infect. Microbiol.* **2016**, *6*, 144. [CrossRef]
80. Boehr, D.D.; Nussinov, R.; Wright, P.E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5*, 789–796. [CrossRef]
81. Smock, R.G.; Gierasch, L.M. Sending signals dynamically. *Science* **2009**, *324*, 198. [CrossRef]
82. White, J.T.; Li, J.; Grasso, E.; Wrabl, J.O.; Hilser, V.J. Ensemble allosteric model: Energetic frustration within the intrinsically disordered glucocorticoid receptor. *Philos. Trans. R. Soc. B-Biol. Sci.* **2018**, *373*, 20170175. [CrossRef] [PubMed]
83. Mittag, T.; Marsh, J.; Grishaev, A.; Orlicky, S.; Lin, H.; Sicheri, F.; Tyers, M.; Forman-Kay, J.D. Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure* **2010**, *18*, 494–506. [CrossRef] [PubMed]
84. McDowell, C.; Chen, J.; Chen, J. Potential Conformational Heterogeneity of p53 Bound to S100B(betabeta). *J. Mol. Biol.* **2013**, *425*, 999–1010. [CrossRef] [PubMed]
85. Wu, H.; Fuxreiter, M. The Structure and Dynamics of Higher-Order Assemblies: Amyloids, Signalosomes, and Granules. *Cell* **2016**, *165*, 1055–1066. [CrossRef] [PubMed]
86. Krois, A.S.; Ferreon, J.C.; Martinez-Yamout, M.A.; Dyson, H.J.; Wright, P.E. Recognition of the disordered p53 transactivation domain by the transcriptional adapter zinc finger domains of CREB-binding protein. *Proc. Natl. Acad. Sci. USA* **2016**, *18*, 494–506. [CrossRef]
87. Csizmok, V.; Orlicky, S.; Cheng, J.; Song, J.H.; Bah, A.; Delgoshaie, N.; Lin, H.; Mittag, T.; Sicheri, F.; Chan, H.S.; et al. An allosteric conduit facilitates dynamic multisite substrate recognition by the SCFCdc4 ubiquitin ligase. *Nat. Commun.* **2017**, *8*, 13943. [CrossRef] [PubMed]
88. Borgia, A.; Borgia, M.B.; Bugge, K.; Kissling, V.M.; Heidarsson, P.O.; Fernandes, C.B.; Sottini, A.; Soranno, A.; Buholzer, K.J.; Nettels, D.; et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **2018**, *555*, 61. [CrossRef] [PubMed]
89. Clark, S.; Myers, J.B.; King, A.; Fiala, R.; Novacek, J.; Pearce, G.; Heierhorst, J.; Reichow, S.L.; Barbar, E.J. Multivalency regulates activity in an intrinsically disordered transcription factor. *eLife* **2018**, *7*, e36258. [CrossRef]
90. Zhao, J.; Liu, X.; Blayney, A.; Zhang, Y.; Gandy, L.; Mirsky, P.O.; Smith, N.; Zhang, F.; Linhardt, R.J.; Chen, J.; et al. Intrinsically Disordered N-terminal Domain (NTD) of p53 Interacts with Mitochondrial PTP Regulator Cyclophilin D. *J. Mol. Biol.* **2022**, *434*, 167552. [CrossRef]
91. Fuxreiter, M. Fuzziness in Protein Interactions-A Historical Perspective. *J. Mol. Biol.* **2018**, *430*, 2278–2287. [CrossRef]
92. Weng, J.; Wang, W. Dynamic multivalent interactions of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2019**, *62*, 9–13. [CrossRef]
93. Miskei, M.; Antal, C.; Fuxreiter, M. FuzDB: Database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Res.* **2017**, *45*, D228–D235. [CrossRef] [PubMed]
94. Wojcik, S.; Birol, M.; Rhoades, E.; Miranker, A.D.; Levine, Z.A. Targeting the Intrinsically Disordered Proteome Using Small-Molecule Ligands. In *Intrinsically Disordered Proteins*; Rhoades, E., Ed.; Academic Press: Cambridge, MA, USA, 2018; Volume 611, p. 703.
95. Ruan, H.; Sun, Q.; Zhang, W.L.; Liu, Y.; Lai, L.H. Targeting intrinsically disordered proteins at the edge of chaos. *Drug Discov. Today* **2019**, *24*, 217–227. [CrossRef] [PubMed]
96. Chen, J.; Liu, X.; Chen, J. Targeting intrinsically disordered proteins through dynamic interactions. *Biomolecules* **2020**, *10*, 743. [CrossRef] [PubMed]
97. Santofimia-Castano, P.; Rizzuti, B.; Xia, Y.; Abian, O.; Peng, L.; Velazquez-Campoy, A.; Neira, J.L.; Iovanna, J. Targeting intrinsically disordered proteins involved in cancer. *Cell. Mol. Life Sci.* **2020**, *77*, 1695–1707. [CrossRef]
98. Krishnan, N.; Koveal, D.; Miller, D.H.; Xue, B.; Akshinthala, S.D.; Kragelj, J.; Jensen, M.R.; Gauss, C.M.; Page, R.; Blackledge, M.; et al. Targeting the disordered C terminus of PTP1B with an allosteric inhibitor. *Nat. Chem. Biol.* **2014**, *10*, 558–566. [CrossRef]
99. Demarest, S.J.; Martinez-Yamout, M.; Chung, J.; Chen, H.W.; Xu, W.; Dyson, H.J.; Evans, R.M.; Wright, P.E. Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **2002**, *415*, 549–553. [CrossRef]
100. Liu, X.; Chen, J.; Chen, J. Residual Structure Accelerates Binding of Intrinsically Disordered ACTR by Promoting Efficient Folding upon Encounter. *J. Mol. Biol.* **2019**, *431*, 422–432. [CrossRef]
101. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2016.
102. Pinheiro Cinelli, L.; Araújo Marins, M.; Barros da Silva, E.A.; Lima Netto, S. Variational Autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*; Cinelli, L.P., Marins, M.A., Barros da Silva, E.A., Netto, S.L., Eds.; Springer International Publishing: Cham, Switzerland, 2021; p. 111.
103. Welling, D.P.K.a.M. An Introduction to Variational Autoencoders. *Found. Trends Mach. Learn.* **2019**, *12*, 307–392.
104. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
105. Parshakova, T.; Andreoli, J.-M.; Dymetman, M. Global Autoregressive Models for Data-Efficient Sequence Learning. *arXiv* **2019**, arXiv:1909.07063.
106. Noe, F.; Tkatchenko, A.; Muller, K.R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390. [CrossRef] [PubMed]

107. Mardt, A.; Pasquali, L.; Wu, H.; Noe, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5. [CrossRef] [PubMed]
108. Yang, W. Advanced Sampling for Molecular Simulation is Coming of Age. *J. Comput. Chem.* **2016**, *37*, 549. [CrossRef]
109. Tribello, G.A.; Gasparotto, P. Using Dimensionality Reduction to Analyze Protein Trajectories. *Front. Mol. Biosci.* **2019**, *6*, 46. [CrossRef] [PubMed]
110. Salawu, E.O. DESP: Deep Enhanced Sampling of Proteins' Conformation Spaces Using AI-Inspired Biasing Forces. *Front. Mol. Biosci.* **2021**, *8*, 587151. [CrossRef]
111. Kukharenko, O.; Sawade, K.; Steuer, J.; Peter, C. Using Dimensionality Reduction to Systematically Expand Conformational Sampling of Intrinsically Disordered Peptides. *J. Chem. Theory Comput.* **2016**, *12*, 4726–4734. [CrossRef]
112. Smith, Z.; Ravindra, P.; Wang, Y.; Cooley, R.; Tiwary, P. Discovering Protein Conformational Flexibility through Artificial-Intelligence-Aided Molecular Dynamics. *J. Phys. Chem. B* **2020**, *124*, 8221–8229. [CrossRef]
113. Kleiman, D.E.; Shukla, D. Multiagent Reinforcement Learning-Based Adaptive Sampling for Conformational Dynamics of Proteins. *J. Chem. Theory Comput.* **2022**, *18*, 5422–5434. [CrossRef]
114. Tian, H.; Jiang, X.; Trozzi, F.; Xiao, S.; Larson, E.C.; Tao, P. Explore Protein Conformational Space With Variational Autoencoder. *Front. Mol. Biosci.* **2021**, *8*, 781635. [CrossRef]
115. Moritsugu, K. Multiscale Enhanced Sampling Using Machine Learning. *Life* **2021**, *11*, 1076. [CrossRef]
116. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement Learning: A Survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
117. Wang, X.; Wang, S.; Liang, X.; Zhao, D.; Huang, J.; Xu, X.; Dai, B.; Miao, Q. Deep Reinforcement Learning: A Survey. *IEEE Trans. Neural. Netw. Learn Syst.* **2022**. *early access*. [CrossRef] [PubMed]
118. Zuckerman, D.M.; Chong, L.T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu. Rev. Biophys.* **2017**, *46*, 43–57. [CrossRef]
119. Degiacomi, M.T. Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure* **2019**, *27*, 1034–1040.e1033. [CrossRef] [PubMed]
120. Ramaswamy, V.K.; Musson, S.C.; Willcocks, C.G.; Degiacomi, M.T. Deep Learning Protein Conformational Space with Convolutions and Latent Interpolations. *Phys. Rev. X* **2021**, *11*, 011052. [CrossRef]
121. Jin, Y.; Johannissen, L.O.; Hay, S. Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins* **2021**, *89*, 915–921. [CrossRef]
122. Tatro, N.J.; Das, P.; Chen, P.-Y.; Chenthamarakshan, V.; Lai, R. ProGAE: A Geometric Autoencoder-Based Generative Model for Disentangling Protein Conformational Space. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
123. Gupta, A.; Dey, S.; Hicks, A.; Zhou, H.-X. Artificial intelligence guided conformational mining of intrinsically disordered proteins. *Commun. Biol.* **2022**, *5*, 610. [CrossRef]
124. Schrag, L.G.; Liu, X.; Thevarajan, I.; Prakash, O.; Zolkiewski, M.; Chen, J. Cancer-Associated Mutations Perturb the Disordered Ensemble and Interactions of the Intrinsically Disordered p53 Transactivation Domain. *J. Mol. Biol.* **2021**, *433*, 167048. [CrossRef]
125. Zhao, J.; Blayney, A.; Liu, X.; Gandy, L.; Jin, W.; Yan, L.; Ha, J.H.; Canning, A.J.; Connelly, M.; Yang, C.; et al. EGCG binds intrinsically disordered N-terminal domain of p53 and disrupts p53-MDM2 interaction. *Nat. Commun.* **2021**, *12*, 986. [CrossRef]
126. Janson, G.; Valdes-Garcia, G.; Heo, L.; Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **2023**, *14*, 774. [CrossRef]
127. Noe, F.; Olsson, S.; Kohler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*, eaaw1147. [CrossRef] [PubMed]
128. Patel, Y.; Tewari, A. RL Boltzmann Generators for Conformer Generation in Data-Sparse Environments. *arXiv* **2022**, arXiv:2211.10771.
129. Liu, X.; Chen, J. Residual Structures and Transient Long-Range Interactions of p53 Transactivation Domain: Assessment of Explicit Solvent Protein Force Fields. *J. Chem. Theory Comput.* **2019**, *15*, 4708–4720. [CrossRef]
130. Mukrasch, M.D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. Structural polymorphism of 441-residue tau at single residue resolution. *PLoS Biol.* **2009**, *7*, e34. [CrossRef] [PubMed]
131. Yang, Y.; Perdikaris, P. Physics-informed deep generative models. *arXiv* **2018**, arXiv:1812.03511.
132. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [CrossRef]
133. Jagtap, A.D.; Kharazmi, E.; Karniadakis, G.E. Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems. *Comput. Methods Appl. Mech. Eng.* **2020**, *365*, 113028. [CrossRef]
134. Yang, L.; Zhang, D.; Karniadakis, G.E. Physics-Informed Generative Adversarial Networks for Stochastic Differential Equations. *SIAM J. Sci. Comput.* **2020**, *42*, A292–A317. [CrossRef]
135. Liang, C.; Savinov, S.N.; Fejzo, J.; Eyles, S.J.; Chen, J. Modulation of Amyloid-beta42 Conformation by Small Molecules Through Nonspecific Binding. *J. Chem. Theory Comput.* **2019**, *15*, 5169–5174. [CrossRef]
136. Wodak, S.J.; Janin, J. Modeling protein assemblies: Critical Assessment of Predicted Interactions (CAPRI) 15 years hence: 6TH CAPRI evaluation meeting April 17–19 Tel-Aviv, Israel. *Proteins* **2017**, *85*, 357–358. [CrossRef]
137. Mobley, D.L.; Gilson, M.K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531–558. [CrossRef] [PubMed]

138. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A Survey on Deep Transfer Learning. *arXiv* **2018**, arXiv:1808.01974.
139. Jing, B.; Erives, E.; Pao-Huang, P.; Corso, G.; Berger, B.; Jaakkola, T. EigenFold: Generative Protein Structure Prediction with Diffusion Models. *arXiv* **2023**, arXiv:2304.02198.
140. Leiter, C.; Zhang, R.; Chen, Y.; Belouadi, J.; Larionov, D.; Fresen, V.; Eger, S. ChatGPT: A Meta-Analysis after 2.5 Months. *arXiv* **2023**, arXiv:2302.13795.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.