life

MDPI

Article

Multiscale Enhanced Sampling Using Machine Learning

Kei Moritsugu

Graduate School of Medical Life Science, Yokohama City University, Yokohama 230-0045, Japan; moritugu@yokohama-cu.ac.jp; Tel.: +81-45-508-7233

Abstract: Multiscale enhanced sampling (MSES) allows for an enhanced sampling of all-atom protein structures by coupling with the accelerated dynamics of the associated coarse-grained (CG) model. In this paper, we propose an MSES extension to replace the CG model with the dynamics on the reduced subspace generated by a machine learning approach, the variational autoencoder (VAE). The molecular dynamic (MD) trajectories of the ribose-binding protein (RBP) in both the closed and open forms were used as the input by extracting the inter-residue distances as the structural features in order to train the VAE model, allowing the encoded latent layer to characterize the difference in the structural dynamics of the closed and open forms. The interpolated data characterizing the RBP structural change in between the closed and open forms were thus efficiently generated in the low-dimensional latent space of the VAE, which was then decoded into the time-series data of the inter-residue distances and was useful for driving the structural sampling at an atomistic resolution via the MSES scheme. The free energy surfaces on the latent space demonstrated the refinement of the generated data that had a single basin into the simulated data containing two closed and open basins, thus illustrating the usefulness of the MD simulation together with the molecular mechanics force field in recovering the correct structural ensemble.

Keywords: enhanced sampling; multiscale enhanced sampling (MSES); machine learning; variational autoencoder (VAE); ribose-binding protein

check for updates

Citation: Moritsugu, K. Multiscale Enhanced Sampling Using Machine Learning. *Life* **2021**, *11*, 1076. https://doi.org/10.3390/life11101076

Academic Editor: Yuichi Togashi

Received: 24 September 2021
Accepted: 8 October 2021
Published: 12 October 2021

1. Introduction

The construction of the free energy surface (FES) from a well-converged structural ensemble is the ultimate goal of computational biophysics in understanding biomolecular functions. To this aim, various kinds of enhanced sampling methods have been developed [1–8] since it would be difficult for a straightforward application of a brute-force molecular dynamic (MD) simulation to cover the whole configurational space. Despite recent advances in supercomputing and GPU hardware, there remains a limitation in the application of the methods to large biomolecules, especially when a number of comparative systems need to be simulated such as the binding of different chemical compounds.

Previous works have proposed the multiscale enhanced sampling (MSES) in which the sampling of the target protein model at an atomistic resolution is enhanced by coupling with the accelerated dynamics of the associated coarse-grained (CG) model [9–18]. The multiscale simulation system comprises an all-atom model of proteins with its surrounding solvents together with the molecular mechanics (MM) force field ($\mathbf{r}_{MM}$; N degrees of freedom) and a CG model of the proteins ($\mathbf{r}_{CG}$; M degrees of freedom). The potential energy is given by the following equation:

$$V = V_{MM}(\mathbf{r}_{MM}) + V_{CG}(\mathbf{r}_{CG}) + k_{MMCG}[\chi_{MM}(\mathbf{r}_{MM}) - \chi_{CG}(\mathbf{r}_{CG})]^2 (\equiv V_{MMCG}), \qquad (1)$$

where V_{MM} and V_{CG} are the MM and CG potential energy functions, respectively. V_{MMCG} is the coupling term between MM and CG with a coupling constant, k_{MMCG}, which is useful for accelerating the MM dynamics via coupling with CG. $\chi_{CG}(\mathbf{r}_{CG})$ is defined by the K collective variables of the CG coordinates, and $\chi_{MM}(\mathbf{r}_{MM})$ is a K-dimensional vector that is a projection of $\mathbf{r}_{\text{MM}}$ onto the associated K-dimensional space. The intrinsic free energy surface solely from V_{MM} can be obtained by eliminating bias from the coupling V_{MMCG} via Hamiltonian replica exchange in which many replicated systems are assigned various values of k_{MMCG} that range from a large value to zero [19]. The condition $K < M << N$ ensures a high exchange probability irrespective of the size of MM, N, which achieves excellent scalability that is applicable to large protein systems in solution [9]. Moreover, the approximation of an adiabatic separation [20] allows for further efficient sampling by using a single copy of CG that does not experience counterforce from MM, while replicated MM systems are simulated as the restraint MDs between $\chi_{\text{MM}}(\mathbf{r}_{\text{MM}})$ and $\chi_{\text{CG}}(\mathbf{r}_{\text{CG}})$ with various k_{MMCG} values [13].

In MSES, V_{CG} can be arbitrarily chosen based on prior knowledge or experimental data, depending on which subspace is selected for enhanced sampling. However, constructing a CG model that requires the proper selection of a set of parameters is a difficult task, especially for those who are not familiar with computation. In this article, an extension that utilizes a machine learning approach instead of adopting the CG model is proposed—the variational autoencoder (VAE) [21]. VAE is a (deep) neural network that learns encodings for the input data by trying to reconstruct high-dimensional data through low-dimensional encoded data in latent space. This is also a complex generative model of data through the latent space as a representation of compressed data, which is then useful for not only image generation and pattern recognition, but also for the systematic discovery of data-driven collective variables (CVs) from biomolecular simulation data [22–26]. The proper selections of nonlinear CVs is expected in an automatic manner, which is expected to be more efficient than the methods that use linear CVs such as principal component (PC) analysis.

The present application of VAE to MSES, termed as "VAE-driven MSES", is summarized as follows (Figure 1): suppose that there are two different structures, such as closed and open conformations, and that the enhanced sampling is aimed at interpolating the two structures. The MD simulations starting at the two structures are performed and the trajectories of both the closed MD and open MD are used as inputs of the network. Inter-residue distances $\boldsymbol{d}_{\text{input}}$ are used as the structural features, motivated by the recent successes of high-quality protein structure prediction such as AlphaFold [27]. The input $\boldsymbol{x}$ needs to be normalized via f as the values from 0 to 1. Since the latent layer encodes information regarding the difference in the structural dynamics of the closed and open forms, the interpolated data can be efficiently generated in the latent space $\boldsymbol{z}$. The generated data is then decoded to output $\boldsymbol{y}$, allowing for $f^{-1}(\boldsymbol{y}) = \boldsymbol{d}_{\text{output}}$ to be calculated. $\boldsymbol{d}_{\text{output}}$ is then used as the distance restraints on the MM simulations, thus driving the MM conformational sampling via the scheme of MSES.

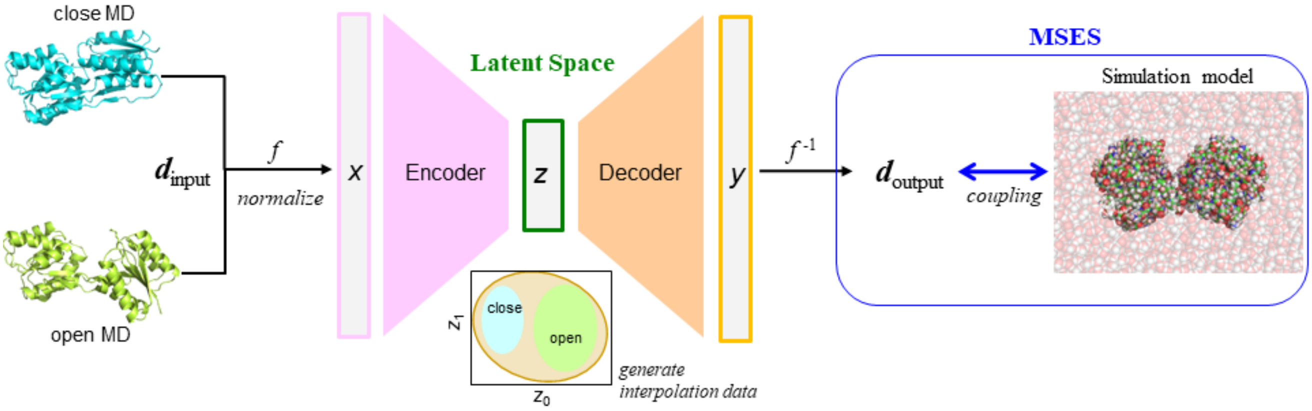

Figure 1. Scheme of VAE-driven MSES. Two trajectories of the closed and open MDs were used as input $\boldsymbol{x}$ for encoding, after normalizing the associated inter-residue distances $\boldsymbol{d}_{\text{input}}$ via f. In the latent space $\boldsymbol{z}$, the interpolated data between the closed and open forms were generated and then decoded to $\boldsymbol{y}$, allowing for $f^{-1}(\boldsymbol{y}) = \boldsymbol{d}_{\text{output}}$ to be calculated. $\boldsymbol{d}_{\text{output}}$ was thus used to drive the sampling of the simulation model such as the all-atom protein structure, including explicit solvent.

The proposed method was applied to the structural change of the ribose-binding protein (RBP). RBP is a periplasmic protein that binds ribose and undergoes an open-close motion of the two ligand-binding domains between the ligand-bound/closed form and the ligand-unbound/open form. The validity of VAE-driven MSES was demonstrated by comparing the derived RBP structural ensemble with that from the original MSES adopting the CG model. The dynamics in the latent space of VAE was also examined to show how the generated data in the reduced space could work for the enhanced sampling. The extrapolation or prediction of the open structure using only the MD trajectories in the closed form with and without the ligand (i.e., when the open structure was blind) was also attempted, which illustrates the usefulness of the MSES simulation together with the molecular mechanics force field in refining the generated data via VAE so as to recover the correct structural ensemble.

2. Materials and Methods

2.1. Model Construction and MD Simulation

MD simulations were carried out for RBP with and without a ribose molecule as the ligand. The present study aimed at an enhanced sampling in the ligand-unbound form, i.e., for large-scale structural change seen in two crystal structures—the ligand-bound/closed form and the ligand-unbound/open form, since the dynamics in the ligand-bound form is considered to be within the fluctuation around the closed structure. To prepare for the subsequent enhanced sampling, three MD simulations were performed, in as short as 100 ns each, for (1) the ligand-bound/closed form (termed herein as "holo"), (2) the ligand-unbound/closed form (termed herein simply as "close"), and (3) the ligand-unbound/open form (termed herein as "open"). The holo and open simulation models were taken from the crystal structures of the Protein data Bank (PDB) entries 2dri [28] and 1ba2 [29], respectively. The R67D point mutation in the holo form was converted into a wild type. The model of the closed simulation was constructed by removing the ligand from the holo model. For the three models, rectangular simulation boxes were constructed, with a margin of 12 Å from the boundary of the simulation box, fully solvated by 20,000–25,000 TIP3P water molecules [30] and potassium and sodium ions at a concentration of ~150 mM, resulting in 70,000–80,000 atoms in total. AMBER ff14SB [31] was used for the potential energy of the protein, and GLYCAM06 [32] for β-D-ribopyranose. The MD simulations were performed by AMBER 16 [33] under constant temperature and pressure (NPT) conditions at P = 1 atm and T = 300 K, using a Berendsen's barostat and Langevin dynamics to control the temperature settings and 1.0 ps^{-1} as the collision frequency. The particle mesh Ewald method [34] was employed for the electrostatic interactions. The time step was 2 fs, using constraining bonds that involve hydrogen atoms via the SHAKE algorithm [35]. The three MD simulation trajectories were used for analyses that were taken every 10 ps.

2.2. Motion Tree

Protein dynamics can be simplified by a set of inter-domain motions once the number of domains is determined, which are then taken as rigid structural units. The regions of such domains are defined by the hierarchical clustering of inter-residue distance fluctuation to construct a tree diagram named "Motion Tree" [36]. The Motion Tree illustrates a pair of domains and the magnitude of the associated domain motion at each node in a hierarchical manner. In previous studies, the Motion Trees were calculated and utilized to analyze complicated MD simulation trajectories [37–39] as well as crystal structure ensembles [14].

The distance fluctuation as a metric of the hierarchical clustering is calculated as:

$$D_{mn} = \langle \Delta d_{mn}^2 \rangle^{1/2}, \tag{2}$$

where Δd_{mn} is the distance between the Cα atoms of residues m and n, and < . . . > is the average over the structural ensemble. Here, both the closed and open MD simulation trajectories were used as the structural ensemble for constructing the Motion Tree. The inter-residue distances between the two defined domains (domain 1 and domain 2) were then used as the features or the input variables of the subsequent VAE (see below), instead of using all the inter-residue distances in an inefficient manner.

2.3. Variational Autoendcoder

In this study, the VAE architecture used consisted of seven fully connected layers containing N_{inp}, 1000, 1000, L, 1000, 1000, and N_{inp} nodes (three layers in both encoder and decoder parts, and a coding layer comprising L nodes). Test calculations using different numbers of layers and nodes were found to yield similar results. As the input data, the inter-residue distances (between Cα atoms; $\boldsymbol{d}_{\text{input}}$) for domain 1 (114 residues) and 2 (131 residues), excluding 1–2 (virtual bond) and 1–3 (virtual angle) interactions, were chosen (see Results below), i.e., the number of elements for $\boldsymbol{d}_{\text{input}}$ was N_{inp} = 14,448. The normalization for the input data was calculated as follows:

$$x = f(\boldsymbol{d}_{\text{input}}) = 0.4(1/\boldsymbol{d}_{\text{input}} - 0.14), \tag{3}$$

where $\boldsymbol{d}$ is in the unit of [nm]. Here, the inverse of $\boldsymbol{d}$ was taken, since closer inter-residue distances will have more information on the structure description such as the atom contacts. Other choices for the input data and f that can make the neural network more efficient and interpretable are also possible, and will be studied in future work. The model was trained by optimizing the sum of reconstruction loss and the Kullback–Leibler divergence for 300 epochs using the Adam optimizer, which was implemented using PyTorch v1.9.1 (https://pytorch.org/, accessed on 12 October 2021). The training data of 20,000 structures were taken from two 100 ns trajectories of the closed and open MDs. The coding of the python script was described by [21].

2.4. Multiscale Enhanced Sampling

The VAE-driven MSES simulation was performed for RBP in the ligand-unbound form. The V_{MM} was the same as in the MD simulations. The coupling potential V_{MMCG} in Equation (1) was replaced by the harmonic distance constraints of $\boldsymbol{d}_{\text{output}}(t)$ that were imposed on the N_{inp} = 14,448 Cα atom pairs between the two dynamic domains. The Hamiltonian replica exchange was carried out every 40 ps using 10 replicas with k_{MMCG} = 0, 0.000008, 0.000014, 0.000025, 0.000043, 0.000072, 0.000119, 0.000194, 0.000312, and 0.00049 kcal/mol/Å^2. The number of replica exchanges was 5000, corresponding to a 200 ns simulation. The trajectories of the unbiased (k_{MMCG} = 0) replica were taken as the target structural ensemble that were used for analyses.

For comparison, the original MSES using the CG model was also carried out. The 271 Cα atoms in RBP were chosen as the CG coordinates. V_{CG} was set as the double-well model that embedded the two closed and open crystal structures so as to drive the structural changes between the two structures [40]. The CG simulation was performed using Langevin dynamics, with a friction constant of 1 ps^{-1} under constant temperature conditions (NVT) of T = 1000 K in order to satisfy the condition of adiabatic separation, with a heavy mass of 10,000 amu [13]. The other parameters on the MSES simulation were the same as described above for the VAE-driven MSES. The MSES simulations were performed with my own script.

3. Results and Discussion

3.1. MD Trajectory Analysis for Processing VAE Input Data

The short-time (100-ns) MD simulations of RBP in both the closed and open forms were performed in order to generate the training data for VAE to optimize the latent space, which characterized the difference between the two structural dynamics. Since the success of the neural network will rely strongly on the processing of the input data, the MD simulation trajectories were firstly analyzed in detail to find out the structural features that were useful in describing the RBP inter-domain motion.

To do this, the Motion Tree was calculated from the closed and open MD trajectories (see Methods). The Tree illustrated clear definitions of the two dynamic domains at Node 1 that move like rigid bodies—domain 1 (114 residues; residues 10–30, 39–102, 234–262) and domain 2 (131 residues; residues 103–233)—after removing flexible regions such as the N-terminal and C-terminal segments as well as the β-sheet (residues 31–38), which were defined at the descendant Nodes 2 and 3 (Figure 2). This information was then used to define the inter-residue distances between the two domains as the input data, excluding the other inter-residue distances that would make the difference between the closed and open forms rather ambiguous.

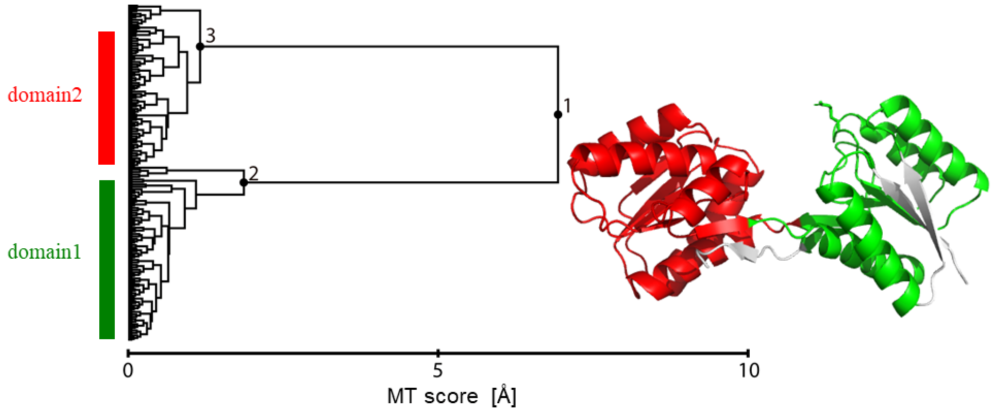

Figure 2. Motion Tree constructed from two trajectories of the closed MD and open MD. This defines two dynamic domains (shown on the right panel in green: residues 10–30, 39–102, 234–262, and in red: residues 103–233) that move like rigid bodies, after removing flexible regions defined at Nodes 2 and 3.

PC analysis was also performed from the closed and open MD trajectories to determine a set of uncorrelated harmonic distribution functions as linear combinations. The motion of the small domain (domain 1) relative to the large domain (domain 2) was visualized by calculating and diagonalizing the variance–covariance matrix of the interatomic fluctuations in the small domain after superimposing the large domain on a reference coordinate. The first and second PCs, with contributions of 0.93 and 0.04, respectively, were then used to draw the FESs. Figure 3a,b show the distant distributions of the closed and open MDs, mainly along PC1. The difference in the distribution width also indicates much more fluctuation during the open MD than the closed MD.

3.2. *Generation of Interpolation Data between Closed and Open Forms via Variational Autoendcoder*

The VAE network was trained so that the high-dimensional input $\boldsymbol{x}$ was maximally reconstructed as the output $\boldsymbol{y}$ through the low-dimensional encoding $\boldsymbol{z}$ for both the closed and open MD structural ensembles. The number of nodes in the latent space L seems to be the most important factor for the success of the dimensional reduction or the choice of proper CVs. To determine the best choice of L, the coincidence between $\boldsymbol{x}$ and $\boldsymbol{y}$ was then examined as a function of L.

Here, the two quantities—the averaged relative difference, $< |\boldsymbol{x} - \boldsymbol{y}| / \boldsymbol{x} >$, and the correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$—were used to compare $\boldsymbol{x}$ and $\boldsymbol{y}$, where these were calculated for each node of the input $\boldsymbol{x}$ and the output $\boldsymbol{y}$ that varied according to the 10,000 closed MD and 10,000 open MD structures, i.e., the number of the calculated data was N_{inp} = 14,448. The associated 2D maps were drawn for L = 2, 5, and 10 (Figure 4), showing sufficient coincident, i.e., a marginally small difference and a high correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$ for the three cases. Note that the data with the correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$ < 0.7 are mostly a consequence of intrinsically very small differences in the averaged inter-residue distances in between the closed and open MDs. A slight improvement or a shift of the peak to the upper left was observed in the 2D map at L = 2 relative to the maps at L = 5 and 10 (see Figure 4). Furthermore, a smaller L is more useful for the visualization and interpretation of the latent space. Thus, L = 2 was chosen in this study.

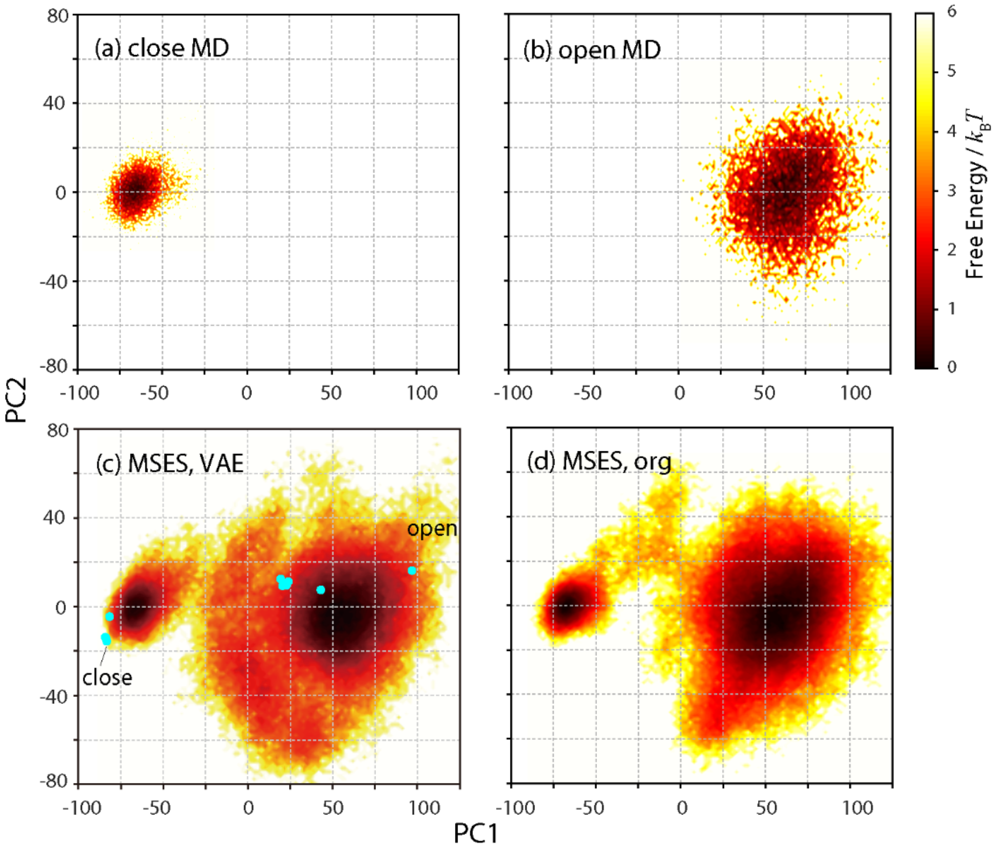

Figure 3. Free energy surfaces in units of k_BT along PC1 and PC2 axes that were determined by both the closed and open MD trajectories. (**a**) closed MD, (**b**) open MD, (**c**) VAE-driven MSES, and (**d**) original MSES. In (**c**), all the 11 chains in 7 crystal structures deposited in PDB are also plotted in cyan, including the closed and open structures of the present simulation models.

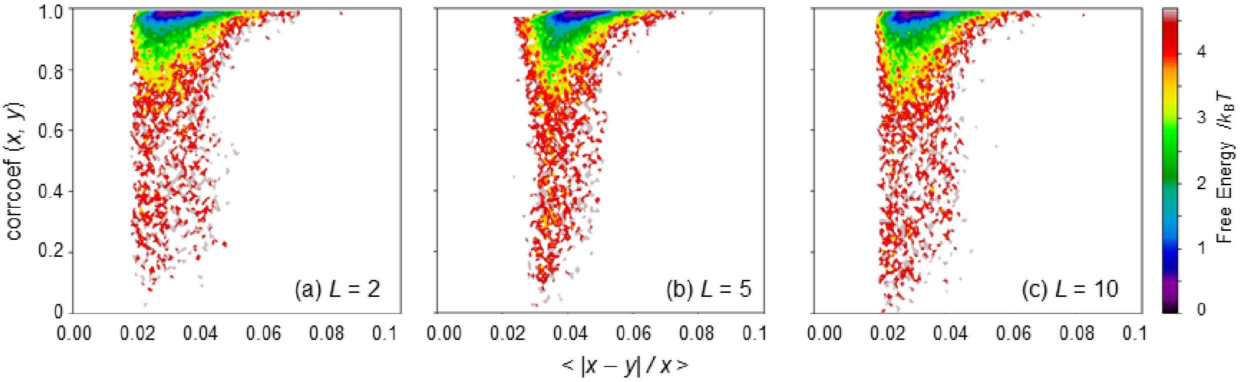

Figure 4. Two-dimensional maps along the averaged relative difference, < $|\,\boldsymbol{x} - \boldsymbol{y}\,|$ / $\boldsymbol{x}$ >, and the correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input and output of VAE (see Figure 1). Free energies in units of k_BT are shown. (**a**) L = 2, (**b**) L = 5, and (**c**) L = 10, where L is the number of nodes used in the latent layer.

The dynamics seen in the closed and open MD simulations were then examined in the latent space with L = 2, i.e., the 2D maps along z_0 and z_1 (Figure 5a,b). The two landscapes show quite distant distributions along z_0, indicating the direction as the CV that most characterizes the difference between the two structural dynamics. In contrast, the distributions along z_1 seem almost the same. This implies that z_1 is unrelated to the structural change between the closed and open forms. In fact, the correlation coefficient of z_0 and z_1 in the training data was very small (0.01).

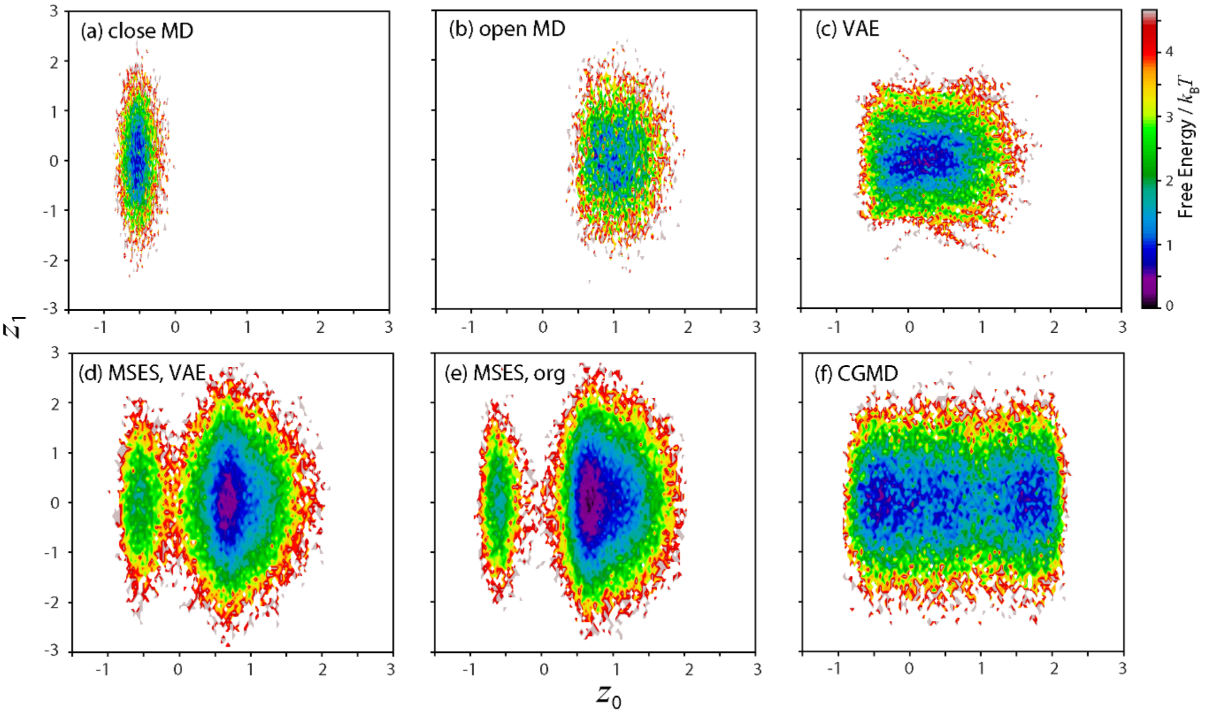

Figure 5. Two-dimensional maps along latent space variables, z_0 and z_1. Free energies in units of k_BT are shown. (**a**) closed MD, (**b**) open MD, (**c**) generated time-series data by VAE, (**d**) VAE-driven MSES, (**e**) original MSES, and (**f**) coarse-grained MD, which was applied to the original MSES.

The data that interpolate the closed and open forms were then generated in the latent space. The resultant dimensional reduction to L = 2 made this possible in various ways, compared with the situation when the data interpolation was carried out in the high-dimensional space $\boldsymbol{x}$ (N_{inp} = 14,448). Another advantage of VAE is that the network also allows for the decoding of the generated data to the inter-residue distances in order to be adopted into MSES in a straightforward manner. In order to correspond with the MSES scheme, the generated data must be continuous or a kind of time-series data, $\boldsymbol{z}$ (t). Here, the simplest strategy was taken to fulfill this requirement: (I) a discrete sequence of $\boldsymbol{z}$, $\boldsymbol{z}$ (n) {n = 1, 2, . . . } was built by randomly taking $\boldsymbol{z}$ from the pools of both the closed and open MDs one by one, such as $\boldsymbol{z}$ (1) = $\boldsymbol{z}_{\mathrm{close}}$ (1), $\boldsymbol{z}$ (2) = $\boldsymbol{z}_{\mathrm{open}}$ (1), $\boldsymbol{z}$ (3) = $\boldsymbol{z}_{\mathrm{close}}$ (2), $\boldsymbol{z}$ (4) = $\boldsymbol{z}_{\mathrm{open}}$ (2), and so on; and (II) a linear interpolation with equally spaced intervals was performed between $\boldsymbol{z}$ (n) and $\boldsymbol{z}$ (n+1), where the number of intervals was set to 25, corresponding to 25 × 40 ps (i.e., the timestep of the replica exchange in MSES) = 1 ns, the timescale for simulating the open-close structural change. The resultant FES of $\boldsymbol{z}$ (t) covered that of both the closed and open MD simulations (Figure 5c).

3.3. VAE-Driven MSES

The time series of the inter-residue distances, $\boldsymbol{d}_{\mathrm{output}}$ (t), was calculated by decoding $\boldsymbol{z}$ (t) to the output $\boldsymbol{y}$ via the VAE network and the calculation of f^{-1} ($\boldsymbol{y}$) = $\boldsymbol{d}_{\mathrm{output}}$, which was then adopted in the VAE-driven MSES. The Hamiltonian replica exchange using 10 copies resulted in a large average acceptance ratio of the replica exchange (0.32), indicating a high sampling efficiency. The derived FES on PC1 and PC2 covered the whole configuration, including the FESs from both the closed and open MDs as well as the 11 RBP crystal structures (all chains deposited in 7 PDB entries, 1ba2, 1dbp, 1dri, 1drk, 1urp, 2dri, and

2gx6) [41], indicating the success of the present enhanced sampling (Figure 3c). To validate the proposed method, the original MSES adopting the CG model was also performed and compared. The resultant FES on PC1 and PC2 was in reasonable agreement with that of the VAE-driven MSES (Figure 3d). Agreement with other enhanced sampling methods was also seen through umbrella sampling [42] and CGMD [41].

The dynamics in the latent space was also examined. To do this, the structural ensembles from the calculated VAE-driven MSES, the original MSES, and the associated CGMD were encoded to z via the VAE network. The resultant 2D maps in the latent space are shown in Figure 5d–f. The FES of VAE-driven MSES contained two basins related to the basins from the closed and open MDs (Figure 5d). Since the ligand-unbound form of RBP was simulated, the open basin was intrinsically more stable than the closed basin, which was also seen in the FES along PC1 and PC2 (see Figure 3c). More importantly, this is quite different from the FES from the generated data by simple interpolation between the closed and open MDs, z (t) (see Figure 5c). This result illuminates the usefulness of MD simulations that employ the molecular mechanics force field in refining the generated data in the latent space of VAE. The FES of the original MSES (Figure 5e) was comparable with that of the VAE-driven MSES, although the FES of the CGMD (Figure 5f), which was used to drive the MM conformation in the original MSES, was much broader than the FES from z (t) (see Figure 5c). This result also indicates that the data of the inter-residue distances can be roughly estimated, e.g., based on the minimum requirement for the information on both the closed and open MDs to be included, since the subsequent MSES simulations are used to recover the correct structural ensemble in terms of statistical mechanics.

3.4. Prediction of Open Structure via Closed and Holo MDs with VAE-Driven MSES

Finally, an attempt to obtain the RBP structural ensemble in the ligand-unbound form using only the closed structure was made (while keeping the open structure blind), which would lead to the prediction of the open structure. To do this, it was assumed that the dynamics response in the closed form on the ligand unbinding would extrapolate the structural change from the closed to the open forms, which is in accord with the linear response theory stating that the protein structural change is reflected by its intrinsic dynamics [43]. In short, the holo and closed MD trajectories were used to generate the data of the target structural ensemble extending to the open structure, which was accomplished in the latent space of VAE. It must be noted that such data extrapolation can be performed as a rough estimation owing to the subsequent refinement of the structural ensemble by MSES.

For this purpose, both the holo and the closed MD trajectories were used as the training data to construct the new VAE network. By taking $L = 10$, it was possible to identify two latent space variables (z'_0 and z'_9, where "[*dash*]" indicates the variables in different networks from those previously defined) that characterized the differences between the holo and the closed dynamics (although small differences were seen in the average of the remaining eight z' variables in between the holo and the closed MDs); Figure 6a,b clarifies the shift of the z'_0–z'_9 distribution to the bottom right via ligand unbinding. The need for an increased L is probably due to a smaller magnitude and more complication in the difference in dynamics seen in between the holo and the closed MDs, rather than in between the closed and the open MDs.

The time-series data, z' (t), was then generated as follows: (I) the average of z' was calculated for the holo and closed MDs separately as $< z' >_{\text{holo}}$ and $< z' >_{\text{close}}$, and (II) the z' distribution extending to the open form, z'_{ext} was built by expanding z'_{close} to the direction of

$$< z' >_{\text{close}} - < z' >_{\text{holo}}, \text{ i.e., } z'_{\text{ext}} = z'_{\text{close}} + \alpha\, [< z' >_{\text{close}} - < z' >_{\text{holo}}], \tag{4}$$

where α is the scaling factor. In this study, $\alpha = 5$ was determined as the maximum value before the Cα atom distance reached 30 Å between residue 231 and 238, comprising the hinge between the two domains (at which point the hinge is fully extended). z' (t) was then generated by connecting the distributions of z'_{close} and z'_{ext}, in the same way as the previous interpolation between z_{close} and z_{open} (see Section 3.2). The 2D map of the derived z' (t) certainly expanded toward the bottom right from the distribution of the closed MD (Figure 6c).

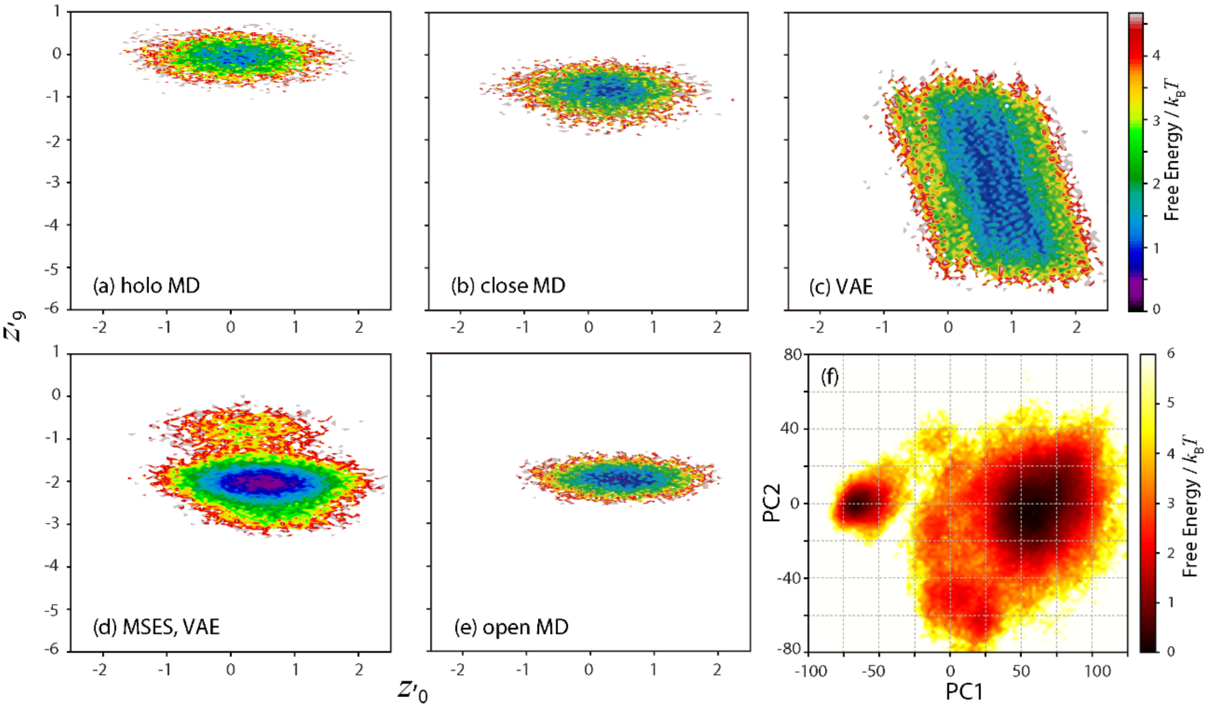

Figure 6. Extrapolation or prediction of the open structure using holo and closed MDs. (**a**–**e**) Two-dimensional maps along two latent space variables, z'_0 and z'_9. Free energies in units of k_BT are shown. (**a**) holo MD, (**b**) closed MD, (**c**) generated time-series data by VAE, (**d**) VAE-driven MSES, and (**e**) open MD. In (**f**), the free energy surface calculated from the structural ensemble of VAE-driven MSES along PC1 and PC2 are also shown in units of k_BT.

The VAE-driven MSES simulation was then carried out by adopting z' (t). The 2D map in the latent space calculated from the derived structural ensemble (Figure 6d) contained two basins related not only to the closed MD but also to the open MD (Figure 6e), demonstrating the structural sampling in the open form without using any information in the open structure. Note that the distribution of z' (t) (see Figure 6c) was much broader than that of the MSES structural ensemble (see Figure 6d). This actually led to a smaller average acceptance ratio of the Hamiltonian replica exchange (0.26), using the same number of copies and k_{MMCG} values, and indicating a slightly decreased sampling efficiency. Nevertheless, the derived FES on PC1 and PC2 (Figure 6f) was almost the same as that from the VAE-driven MSES using the data of the closed and open MDs (see Figure 3c). It can then be concluded that the ability of the VAE-driven MSES to refine the extrapolated data in the latent space as a rough estimation, i.e., using a structural criterion that the inter-domain hinge does not distort, allowed for the prediction of the RBP structural dynamics in the ligand-unbound form, or of the open structure as the most stable state, using only the MD simulations in the closed form.

4. Conclusions

In this study, an extension of the MSES simulation using VAE, a machine learning approach, was proposed and applied to the domain motion of RBP. This method allowed the dynamics in the reduced subspace generated by VAE to perform an enhanced sampling of an all-atom protein structure. Here, the closed and open MD trajectories of RBP were used as the input of the training data after normalizing the residual-residue distances as the structural features. The Motion Tree constructed from the two trajectories was utilized to define the dynamic domains and the inter-domain residue pairs. The trained

VAE model could characterize the difference in the structural dynamics of the closed and open forms in the encoded latent space. The interpolated data were then generated in the low-dimensional latent space by simply connecting the distributions of the closed and open MDs, which were then decoded to the time-series data of the inter-residue distances and used to drive the structural sampling at an atomistic resolution via the MSES scheme. The derived structural ensemble was found to be in reasonable agreement with that from the original MSES adopting the CG model, thus validating the proposed VAE-driven MSES.

The dynamics in the latent space was then examined, since the sampling efficiency of the MSES simulation would rely on the dimensional reduction or the proper choices of CVs via VAE. The distribution calculated from the MSES structural ensemble after encoding this in the latent space contained two basins related to the closed and open forms, reflecting the 2D FES on PC1 and PC2. In contrast, the generated data in the latent space yielded a single basin combining the two closed and open basins. The difference in the latent space distributions indicates that the subsequent MSES can accurately refine the consequent structural ensemble on the basis of statistical mechanics together with the molecular mechanics force field. In this sense, allowing for a rough estimation of the generated data via VAE is the advantage of the present method, although it is rather time-consuming to perform additional MD simulations after the VAE network optimization. This is because it is still difficult for solely a deep neural network to complete the prediction of the whole structural dynamics, especially for comparative systems such as the binding of different ligands and in various kinds of surrounding solvents. Essentially, the prediction of the unknow open structure was demonstrated to be possible by using only the closed MDs with and without the bound ligand and by extrapolating the data related to the open-close motion.

Funding: This research received no external funding.

Data Availability Statement: Data supporting the reported results and the source code on the VAE implementation are available from the author upon request.

Acknowledgments: This research was supported by the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP21am0101109, and by MEXT Grants-in-Aid for Scientific Research (C), 20K03883. The computations were performed at the Graduate School of Medical Life Science at Yokohama City University. This research also used the computational resources of the HPCI system, provided by Kyushu University through the HPCI System Research Project (Project ID: hp200005, hp210009). I am grateful to Hiroshi Fujisaki and Akinori Kidera for their helpful comments.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Schlitter, J.; Engels, M.; Kruger, P.; Jacoby, E.; Wollmer, A. Targeted Molecular-Dynamics Simulation of Conformational Change —Application to the T[−]R Transition in Insulin. *Mol. Simul.* **1993**, *10*, 291–308. [CrossRef]
2. Grubmuller, H. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E* **1995**, *52*, 2893–2906. [CrossRef] [PubMed]
3. Sugita, Y.; Okamoto, Y. Replica—Exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151. [CrossRef]
4. Wang, F.G.; Landau, D.P. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053. [CrossRef]
5. Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566. [CrossRef]
6. Maragliano, L.; Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **2006**, *426*, 168–175. [CrossRef]
7. Morishita, T.; Itoh, S.G.; Okumura, H.; Mikami, M. Free-energy calculation via mean-force dynamics using a logarithmic energy landscape. *Phys. Rev. E* **2012**, *85*, 066702. [CrossRef]
8. Kamiya, M.; Sugita, Y. Flexible selection of the solute region in replica exchange with solute tempering: Application to protein-folding simulations. *J. Chem. Phys.* **2018**, *149*, 072304. [CrossRef]

9. Moritsugu, K.; Terada, T.; Kidera, A. Scalable free energy calculation of proteins via multiscale essential sampling. *J. Chem. Phys.* **2010**, *133*, 224105. [CrossRef]
10. Moritsugu, K.; Terada, T.; Kidera, A. Disorder-to-order transition of an intrinsically disordered region of sortase revealed by multiscale enhanced sampling. *J. Am. Chem. Soc.* **2012**, *134*, 7094–7101. [CrossRef]
11. Moritsugu, K.; Terada, T.; Kidera, A. Multiscale enhanced sampling driven by multiple coarse-grained models. *Chem. Phys. Lett.* **2014**, *616*, 20–24. [CrossRef]
12. Moritsugu, K.; Terada, T.; Kidera, A. Energy Landscape of All-Atom Protein-Protein Interactions Revealed by Multiscale Enhanced Sampling. *PLoS Comput. Biol.* **2014**, *10*, e1003901. [CrossRef]
13. Moritsugu, K.; Terada, T.; Kidera, A. Multiscale enhanced sampling for protein systems: An extension via adiabatic separation. *Chem. Phys. Lett.* **2016**, *661*, 279–283. [CrossRef]
14. Moritsugu, K.; Terada, T.; Kidera, A. Free-Energy Landscape of Protein-Ligand Interactions Coupled with Protein Structural Changes. *J. Phys. Chem. B* **2017**, *121*, 731–740. [CrossRef]
15. Moritsugu, K.; Terada, T.; Kokubo, H.; Endo, S.; Tanaka, T.; Kidera, A. Multiscale enhanced sampling of glucokinase: Regulation of the enzymatic reaction via a large scale domain motion. *J. Chem. Phys.* **2018**, *149*, 072314. [CrossRef] [PubMed]
16. Moritsugu, K.; Nishi, H.; Inariyama, K.; Kobayashi, M.; Kidera, A. Dynamic recognition and linkage specificity in K63 di-ubiquitin and TAB2 NZF domain complex. *Sci. Rep.* **2018**, *8*, 16478. [CrossRef] [PubMed]
17. Moritsugu, K.; Nishino, Y.; Kidera, A. Inter-lobe Motions Allosterically Regulate the Structure and Function of EGFR Kinase. *J. Mol. Biol.* **2020**, *432*, 4561–4575. [CrossRef] [PubMed]
18. Yasar, F.; Bernhardt, N.A.; Hansmann, U.H. Replica-exchange-with-tunneling for fast exploration of protein landscapes. *J. Chem. Phys.* **2015**, *143*, 224102. [CrossRef]
19. Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058–9067. [CrossRef]
20. Rosso, L.; Tuckerman, M.E. An adiabatic molecular dynamics method for the calculation of free energy profiles. *Mol. Simul.* **2002**, *28*, 91–112. [CrossRef]
21. Kingma, D.P.; Welling, M. Auto-encoding variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
22. Hernandez, 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] [PubMed]
23. 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]
24. Noe, F.; De Fabritiis, G.; Clementi, C. Machine learning for protein folding and dynamics. *Curr. Opin. Struct. Biol.* **2020**, *60*, 77–84. [CrossRef] [PubMed]
25. Noé, F.; Tkatchenko, A.; Muller, K.R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys.Chem.* **2020**, *71*, 361–390. [CrossRef] [PubMed]
26. 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] [PubMed]
27. 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]
28. Bjorkman, A.J.; Binnie, R.A.; Zhang, H.; Cole, L.B.; Hermodson, M.A.; Mowbray, S.L. Probing Protein-Protein Interactions - the Ribose-Binding Protein in Bacterial Transport and Chemotaxis. *J. Biol. Chem.* **1994**, *269*, 30206–30211. [CrossRef]
29. Bjorkman, A.J.; Mowbray, S.L. Multiple open forms of ribose-binding protein trace the path of its conformational change. *J. Mol. Biol.* **1998**, *279*, 651–664. [CrossRef]
30. Jorgensen, W.D. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
31. Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713. [CrossRef]
32. Kirschner, K.N.; Yongye, A.B.; Tschampel, S.M.; Gonzalez-Outeirino, J.; Daniels, C.R.; Foley, B.L.; Woods, R.J. GLYCAM06: A generalizable Biomolecular force field. Carbohydrates. *J. Comput. Chem.* **2008**, *29*, 622–655. [CrossRef] [PubMed]
33. Case, D.A.; Cheatham, T.E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688. [CrossRef] [PubMed]
34. Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald—An N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092. [CrossRef]
35. Ryckaert, J.P.; Ciccotti, G.; Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341. [CrossRef]
36. Koike, R.; Ota, M.; Kidera, A. Hierarchical Description and Extensive Classification of Protein Structural Changes by Motion Tree. *J. Mol. Biol.* **2014**, *426*, 752–762. [CrossRef]
37. Moritsugu, K.; Koike, R.; Yamada, K.; Kato, H.; Kidera, A. Motion Tree Delineates Hierarchical Structure of Protein Dynamics Observed in Molecular Dynamics Simulation. *PLoS ONE* **2015**, *10*, e0131583. [CrossRef]
38. Moritsugu, K.; Ito, T.; Kidera, A. Allosteric response to ligand binding: Molecular dynamics study of the N-terminal domains in IP3 receptor. *Biophys. Physicobiol.* **2019**, *16*, 232–239. [CrossRef] [PubMed]

39. Koike, R.; Takeda, S.; Maeda, Y.; Ota, M. Comprehensive analysis of motions in molecular dynamics trajectories of the actin capping protein and its inhibitor complexes. *Proteins-Struct. Funct. Bioinform.* **2016**, *84*, 948–956. [CrossRef] [PubMed]
40. Maragakis, P.; Karplus, M. Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. *J. Mol. Biol.* **2005**, *352*, 807–822. [CrossRef] [PubMed]
41. Orellana, L.; Yoluk, O.; Carrillo, O.; Orozco, M.; Lindahl, E. Prediction and validation of protein intermediate states from structurally rich ensembles and coarse-grained simulations. *Nat. Commun.* **2016**, *7*, 12575. [CrossRef] [PubMed]
42. Ravindranathan, K.P.; Gallicchio, E.; Levy, R.M. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *J. Mol. Biol.* **2005**, *353*, 196–210. [CrossRef] [PubMed]
43. Ikeguchi, M.; Ueno, J.; Sato, M.; Kidera, A. Protein structural change upon ligand binding: Linear response theory. *Phys. Rev. Lett.* **2005**, *94*, 078102. [CrossRef] [PubMed]