Characterizing the function of domain linkers in regulating the dynamics of multi-specific biologics by microsecond molecular dynamics simulations and artificial intelligence

Bo Wang, Zhaoqian Su, and Yinghao Wu*

[1]Department of Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY, 10461

*Corresponding author:

Yinghao Wu

Phone: (718) 678-1232, Fax: (718) 678-1018, Email:yinghao.wu@einstein.yu.edu

ABSTRACT

Multi-domain proteins are not only formed through natural evolution but can also be generated by recombinant DNA technology. Because many fusion proteins can enhance the selectivity of cell targeting, these artificially produced molecules, called multi-specific biologics, are promising drug candidates, especially for immunotherapy. Moreover, the rational design of domain linkers in fusion proteins is becoming an essential step toward a quantitative understanding of the dynamics in these biopharmaceutics. We developed a computational framework to characterize the impacts of peptide linkers on the dynamics of multi-specific biologics. We constructed a benchmark containing six types of linkers that represent various lengths and degrees of flexibility and used them to connect two natural proteins as a test system. The microsecond dynamics of these proteins generated from Anton were projected onto a coarse-grained conformational space. The similarity of dynamics among different proteins in this low-dimensional space was further analyzed by a neural network model. Finally, hierarchical clustering was applied to place linkers into different subgroups based on the neural network classification results. The clustering results suggest that the length of linkers used to spatially separate different functional modules plays the most important role in regulating the dynamics of this fusion protein. Given the same number of amino acids, linker flexibility functions as a regulator of protein dynamics. In summary, we illustrated that a new computational strategy can be used to study the dynamics of multi-domain fusion proteins by a combination of long timescale molecular dynamics simulation, coarse-grained modeling, and artificial intelligence.

Introduction

Through evolutionary pathways, the majority of proteins are encoded with at least two structurally conserved domains in genomes of many organisms . Nowadays, domains from different species can also be genetically fused by recombinant DNA technology . Domains in either native or artificially generated fusion proteins are connected by peptide regions called domain linkers . While the tertiary structure of the individual domain remains unchanged, linkers are closely related to the relative orientation between different domains . For native proteins, these diversities of inter-domain positions and resulting dynamics are essential in many biological processes, such as signal transduction or transcriptional regulation . Thus, characterizing the impacts of linkers on the dynamics of multi-domain proteins is particularly important to understand their cellular functions. Moreover, multi-domain fusion proteins are becoming a promising category of bio-therapeutics, known as multi-specific biologics . The interplay among various pharmaceutical modules in these biologics is predominantly determined by linker properties. Direct fusion of functional modules without a linker often leads to impaired bioactivities . As a result, quantifying the dynamics of multi-specific biologics through the rational design of specific linkers is thought to be a crucial yet underexplored strategy in the development of next-generation biopharmaceutics .

Computational modeling possesses unique advantages over labor-intensive and time-consuming experimental approaches to test conditions that are difficult to attain in the laboratory on a systematic level. Among a large variety of different computational techniques, molecular dynamics (MD) simulation has already turned into a mature method that allows us to study the dynamics of biomolecules on atomic details . It has been used to analyze the positional fluctuations and correlated motions in multi-domain proteins . However, due to the intense consumption of computational resources, current applications of MD simulations are limited by the timescale they can reach. Fortunately, Anton—a supercomputer specializing in MD simulations—has recently become publicly available. The simulation performance on Anton and its upgraded version, Anton 2 , is nearly two orders of magnitude faster than other traditional supercomputers. In contrast, the new advancements in artificial intelligence (AI) have gained increasing attention from

the field of bioinformatics and demonstrated huge successes in its application to protein structure prediction . Therefore, the combination of AI and MD simulations provides a way to capture the complicated features in biomolecular systems. For instance, machine learning–based algorithms have been employed to train force fields or explore new conformational states by the given data generated from MD simulations.

In this work, we develop a computational framework to characterize the dynamics of multi-domain proteins. Two different ligands of immune receptors, major histocompatibility complex (MHC) and programmed death-ligand 1 (PD-L1) , are artificially connected by a peptide linker as a test model for multi-specific biologics. In the traditional two-signal hypothesis of T cell activation , an initial signal is provided by adhesion between T cell receptors (TCRs) on the surfaces of T cells and specific MHC-epitope molecules on the surface of antigen-presenting cells (APCs). This initial signal is followed by a secondary pathway through the engagement between co-regulatory receptors on T cell surfaces and their ligands on APCs. Secondary modulation can lead to either stimulatory signal (e.g., through the binding between tumor necrosis factor [TNF] and TNF receptor) or inhibitory signal (e.g., through the binding between PD-L1 and PD-1). Analogous to the natural response, we assume that, in our test system, the module of MHC can target the fusion protein to the specific T cell clones, while the other module, PD-L1, can direct the co-regulatory signal to inhibit the targeted T cells . As a result, this bi-specific biologics can hypothetically allow T cell targeting and modulation without the encounter of specific APCs.

For the test model, we construct a benchmark containing six types of linkers that connect MHC and PD-L1 with various lengths and degrees of flexibility. We utilize the supercomputer Anton 2 to carry out microsecond-level MD simulations on these linker-specific fusion proteins. Their dynamics are projected onto a low-dimensional space by coarse-graining the protein structure with a vector-based representation. We use a neural network–based classifier to compare the low-resolution motions between proteins of different linkers and further build a phylogenetic tree to summarize the classification results. We find that the six linkers hierarchically cluster into groups based on their length and dynamic features, suggesting that the global dynamics of this protein can be

effectively identified by its linker properties. Therefore, our results highlight the importance of linkers in orchestrating the motions between MHC and PD-L1 modules in this test system of bi-specific biologics. The computational strategy adopted in this study will potentially be useful to design new fusion proteins for future biopharmaceutics.

Methods

Linker benchmark and system preparation

The structural model of our test system was computationally constructed by fusing two functionally independent protein modules with a peptide linker. The structure of the MHC module was adopted from the protein databank (PDB) ID 3NWM, while the structure of the other PD-L1 module was adopted from the PDB ID 4Z18. The MHC module consists of 375 amino acids, including a heavy chain H-2K^{d} with 275 amino acids and a light chain β2m with 100 amino acids . To avoid instability during simulations, the target peptide in the groove of the original MHC was not modeled in the system. The PD-L1 module, in contrast, consists of two immunoglobulin structural domains, each of which contains about 100 amino acids . The linker region connects the C-terminus of the MHC light chain with the N-terminal domain of PD-L1. There are many linkers in the literature with different sequences and lengths. In order to provide a comprehensive and systematic study, we constructed a benchmark that contains six types of typical linkers. The first linker is called GS15, with a total length of 15 amino acids. The linker contains three repeats of a small fragment with four glycine followed by a serine. The second linker is called GS30, with six copies of GGGGS fragments. These two linkers are assumed to be intrinsically disordered due to the flexible feature of glycine . As a result, the initial structures of these two linkers were built by ModLoop . The third linker is called PLP15, which contains 15 consecutive prolines. The peptide of poly-proline can form the structure of α-helix with either right-handed or left-handed symmetry . Therefore, the initial structure of this linker was built by following the standard configuration of a right-handed α-helix. The fourth and fifth linkers are called PLPII15 and PLPII30, in which there are 15 and 30 consecutive proline, respectively. Different from PLP15, the initial structures of these two linkers were built by following the configuration of more extended left-handed α-helix . Finally, the sixth linker is called

PLrigid. It contains 20 amino acids with four repeats of fragments in total. Each fragment starts with a glutamic acid, followed by three consecutive amino acids of alanine, and finally, an arginine . The initial structure of this linker was built by following the standard configuration of a right-handed α-helix. The overall information of the linker benchmark is summarized in **Table 1**. The initial structures of six constructed fusion proteins using these linkers are shown in **Figure 1a** to **Figure 1f**. Their biological context will be described in the discussion.

The protocol of the Anton MD simulation

All equilibrium simulations of constructed fusion proteins with six different linkers were run on the Anton 2 supercomputer at the Pittsburgh Supercomputing Center . Proteins were solvated with water molecules and neutralized by adding Na+ and Cl– ions. As a result, the systems contain an average number of 252,000 atoms. For all Anton production runs, the isothermal–isobaric (NPT) ensemble was used with constant pressure (1 atm) and physiological temperature (310 K) using a Nosé–Hoover thermostat. We adopted the orthorhombic cells with approximate dimensions of 125 Å × 120 Å × 160 Å as simulation boxes for all systems, and periodic boundary conditions were imposed. A 2 fs time step was used for all simulations. We chose the CHARMM36m force field for proteins and the TIP4P-D water model. The water dispersion interactions were increased in the TIP4P-D model, enabling more accurate simulation of dynamics for proteins with intrinsic flexible linkers . The system-optimized simulation parameters were chosen by the Anton software. The Gaussian-split Ewald algorithm was used to compute the long-range electrostatic interactions with a 64 × 64 × 64 Å mesh. The cutoff for short-range non-bonded interactions was chosen to be at least 11 Å for all boxes. Consequently, a 2.5 μs trajectory was collected from the Anton 2 supercomputer for each bi-specific linker. This gives a total of six trajectories and an aggregate simulation time of 15 μs.

Vector-based coarse-graining of protein dynamics

The global conformation of a fusion protein can be described by a simplified vector-based model , if one only focuses on the relative orientation between its two

functional modules. The procedure to construct this model consisted of the following steps: First, a set of representative points was selected from the protein. These points were the center of the binding interface on the MHC, the center of mass of the MHC, the starting residue of the linker, the ending residue of the linker, the center of mass of PD-L1, and the center of the binding interface on PD-L1. After the selection of these points, a series of vectors were built by connecting these points with each other in the above order. Given these coarse-grained vectors as a virtual skeleton of the protein, the degrees of freedom that define its conformational changes were embodied through the internal coordinates along these vectors. These internal coordinates include the length of each vector between two consecutive points, the angle between every two consecutive vectors, and the dihedral formed among three consecutive vectors. Finally, if we neglected the local conformational fluctuations within each functional module, the conformations between two functional modules could be described by a minimal number of only six degrees of freedom along the vectors. They are the length of the linker (r), the two packing angles between functional modules and linkers (θ_1 and θ_2), and three packing dihedrals describing the relative rotations of functional modules (φ_1, φ_2, and φ_3). The vector-based virtual skeleton of the fusion protein and the definition of the six internal coordinates are illustrated in **Figure 1g**. As a result, the large-scale conformational changes of the fusion proteins within a certain amount of time Δt can be reflected by the difference in the values of these six internal coordinates between time t and $t+\Delta t$.

Identify dynamic similarity between linkers via a machine learning algorithm

A machine learning algorithm was used to identify the dynamics of fusion protein with one type of linker from another type in the benchmark. In detail, a feedforward back-propagation network was implemented. For a specific pair of linkers, the input neurons of the network are in six dimensions, the same as the internal coordinates along the virtual skeleton that are used to represent the global dynamics of fusion proteins. As described in the last section, each dimension gives the variation in the values of the corresponding internal coordinates from time t to time $t+\Delta t$, in which Δt is the time step providing information about the temporal correlation of conformational dynamics. The output is in one dimension, corresponding to the type of linker in comparison. The

network further contains a single hidden layer with four neurons. In this study, a sigmoid activation function was adopted. The weight of each neuron was modified using the back-propagation learning algorithm with a sum of square error function . The magnitude of the error sum in the learning process was monitored in each cycle, and the learning was terminated when the network converged.

The neural network algorithm was applied to all pairs of six linkers. For each pair, we used leave-one-out cross-validation to calibrate the test results after learning. Because the dynamics of fusion proteins for each linker can be represented by 2,000 sets of six-dimensional vectors, there are 4,000 vectors for a corresponding pair of linkers. As a result, the total cross-validation procedure for this pair of linkers contains 4,000 steps. Within each step, one vector was selected for testing, while the remaining vectors were used for training. During the learning process, vectors in the training set were fed into the neural network in a random order. After the network was trained, it was used to predict the type of given linker in the test set as input. When all 4,000 cross-validations were completed, the accuracy of prediction could be calculated. The accuracy is defined as the total percentage of correctly recognized linkers, belonging to either type of linkers in the pair, among all 4,000 predictions. High accuracy indicates that one of the linkers is highly distinguishable from the other one in the pair, suggesting that the dynamics between two fusion proteins are different. In contrast, low accuracy means that two linkers cannot be distinguished from each other based on their dynamic properties. After the calculations for accuracy, we can define the similarity between all the linkers in the benchmark.

The source codes and analysis results from our neural network–based classification program are available for download (https://github.com/wulab-github/AntonCGNN). This package also contains the structural ensembles generated from the Anton 2 simulations for all six types of linkers. They have been converted into the vector-based representation and are used as inputs of the neural network model. The source codes of the classification program are in FORTRAN77 format. The executable file is also provided. Detailed instructions about the program and

its output can be found in the repository. The program is free for academic users and works on a Linux platform.

Results

All linkers show large conformational fluctuations based on the MD simulation results

The MD simulations of the fusion proteins were carried out on Anton 2 supercomputers based on their all-atom structural models. The detailed procedure of system preparation and simulation setups are described in the methods. A 2.5 microsecond-long trajectory was generated to sample the conformational space for each of these six systems. For each system, the first 500 nanoseconds of the trajectory are used for equilibrium, while the next phase of 2 μs trajectory is used for recording. During the second phase, the conformation of each protein was recorded every 1 ns. In order to identify global conformational fluctuations of the proteins from the local conformational changes of the linker regions, we analyzed both the global root-mean-square difference (RMSD) of the entire protein and the local RMSD of the residues only in the linkers. The newly updated conformation was first superimposed onto the initial conformation of the recording phase by rigid-body superposition. We then calculated the backbone RMSD between the coordinates of Cα atoms in the new and initial conformations. The algorithm of rigid-body superposition uses the least-square minimization to generate the best rotation that fits the two sets of coordinates . This rotation is then applied to the new conformation so that it can be spatially aligned to the initial conformation before the calculation of RMSD. All the Cα atoms were used in the superposition in order to calculate the global RMSD, while only the Cα atoms in the linker region were used in the superposition and the calculation of the local RMSD.

We first plotted the local RMSD as a function of simulation time for all six types of linkers, as shown in **Figure 2a**. The color index of the curves in the figure is given on the top. The figure shows that the linker GS30 has the highest RMSD due to its length and high flexibility (red curve in **Figure 2a**). On the other hand, the linker PLP15 has much lower RMSD than all the other linkers (blue curve in **Figure 2a**). The average value of RMSD was less than 2 Å, indicating that the secondary structure of the right-

handed α-helix in the linker of PLP15 was maintained throughout the simulations. It is worth mentioning that the values of RMSD are dependent on the size of molecular systems. As a result, it is more meaningful to focus on the comparison of linkers with the same length. For instance, **Figure 2a** shows that, although the secondary structure of left-handed α-helix in the linker of PLPII15 was also maintained throughout the simulations, its RMSD (purple curve in **Figure 2a**) is much larger than that of the right-handed α-helix linker PLP15. This suggests that the more extended configuration of left-handed α-helix is much more flexible than right-handed α-helix is. In contrast, GS15 has the highest RMSD value (black curve in **Figure 2a**) among all three linkers consisting of 15 residues. Similarly, the RMSD of the linker PLPII30 (green curve in **Figure 2a**) is only slightly lower than GS30 (red curve in **Figure 2a**), confirming the flexibility of the left-handed α-helix. The observation of high flexibility in the left-handed α-helix is consistent with previous studies [42], which could be due to the lack of a stabilizing effect of intramolecular hydrogen bonds. Another possible reason is that the left-handed α-helices have a rise per residue that is almost twice as great as that of the right-handed α-helices, which makes them more exposed to solvent.

More interestingly, the linker of PLrigid shows a transition from low RMSD to high RMSD around 0.15 μs after equilibrium (orange curve in **Figure 2a**). Some snapshots of the linker configuration were taken from the trajectory around this time window, as shown in **Figure 3**. The figure suggests that the linker still had the helical structure at the time of 0.125 μs after equilibrium (**Figure 3a**). However, soon after that, the secondary structure at both ends of the linker started to melt (**Figure 3b** and **3c**). At 0.2 μs after equilibrium, all hydrogen bonds in the helical structure of the linker were broken (**Figure 3d**). Therefore, the low-RMSD to high-RMSD transition for PLrigid is because of its loss of secondary structure. The breaking of hydrogen bonds in the α-helix is further due to the reason that the Anton simulation was performed under a relatively high temperature of 310 K. While this temperature is biologically more relevant, our results demonstrated that some rigid linkers could be more flexible than was originally considered under room temperature.

The global RMSD for all residues in the fusion proteins is plotted as a function of simulation time for all six types of linkers in **Figure 2b**. The same color index is used. Different from the local RMSD, which highly depended on the type of linkers, we observed high values of global RMSDs for all six systems. For example, although the local RMSD of linker PLP15 is lower than 2 Å through all the 2 μs simulations, the global RMSD of the entire protein with PLP15 linker on average is around 20 Å (blue curves in **Figure 2a** and **2b**). **Figure 4** shows a few snapshots of the global configuration for this protein from the trajectory. The MHC module is highlighted in red, while the PD-L1 module is highlighted in green, and the linker in the middle is highlighted in gray. These snapshots clearly show that the local configuration of the linker region was maintained as α-helix throughout the entire simulation, while the changes in tertiary structures within each module can also be neglected. However, there are remarkable fluctuations in the relative orientations between the two modules, which results in the high value of the global RMSD. Therefore, our results indicate that small structural variations in rigid linkers can still lead to large inter-domain conformational fluctuations.

In summary, our microsecond MD simulations revealed that the levels of local structural dynamics are highly dependent on the sequence composition of different linkers. We also observed the loss of the presumably formed secondary structure during the simulation. However, the orientations between two modules show large fluctuations for all six fusion proteins, even if the linker itself is rigid. It is difficult to distinguish their differences purely based on the calculations of the global RMSD. Therefore, in the next section, we will use a set of simplified coordinate systems to quantitatively analyze the orientation between two modules.

The coarse-grained model captures the conformational preference for different linkers.

It is not sufficient to interpret the conformational fluctuations of proteins solely based on their RMSD profiles. In contrast, the impacts of different linkers on the global dynamics of multi-domain proteins can only be derived by focusing on a limited number of large-scale degrees of freedom. Therefore, we coarse-grained the protein structures by a vector-based model, as described in the **Methods** section. Using this model, the variations of protein structure between two functional modules were projected from MD

simulation trajectories onto a conformational space with six dimensions, $\{r, \theta_1, \theta_2, \varphi_1, \varphi_2, \varphi_3\}$, as shown in **Figure 1g**. We calculated the distribution of each of these six internal coordinates for each fusion protein in the benchmark. The means and standard deviations of these distributions, as well as their correlation coefficients for all six fusion proteins, can be found in the **Supporting Information** as **Table S1**. In order to show the similarity and difference between specific distributions, a few examples were further selected from the table and plotted in **Figure 5** for a more detailed comparison.

The statistical distribution of linker length r in the protein with linker PLP15 was calculated from Anton simulation and plotted as the red histogram in **Figure 5a**. In the same plot, the distribution of linker length in the protein with another linker GS30 forms the black histogram. The figure shows that the black histogram has a much wider distribution than the red histogram does, indicating that the variations of linker length in GS30 are much larger than they are in PLP15. This is because multiple hydrogen bonds formed in the poly-proline helix, making the linker PLP15 difficult to stretch. In contrast, the linker GS30 not only contains more amino acids but is also more flexible. Therefore, the two modules connected by this linker can be either much closer or much farther apart than the modules that are connected by PLP15. Similarly, the comparisons of two packing angles and dihedral between these two linkers are plotted in **Figure 5b** and **Figure 5c**, respectively. The two packing angles of both linkers show normal distributions. The peaks of both packing angles for PLP15 are around 135 degrees (blue and purple histograms), while the peaks of both packing angles for GS30 are around 45 degrees (red and black histograms). Moreover, the distributions of PLP15 show much smaller standard deviations. In terms of the packing dihedral φ_2, **Figure 5c** shows PLP15 for a normal distribution with a peak at 180 degrees, while a uniform distribution is observed with the range from −180 degrees to +180 degrees for GS180. These statistical results suggest that the relative orientations between modules in these two linkers have different preferences, and the packing in linker PLP15 is restricted to a smaller area of the conformational space.

In addition to the overall conformational preference, kinetic information about protein conformational changes within a given time window was obtained by analyzing

the variations of six internal coordinates from time t to $t+\Delta t$, which is defined as $\{\Delta r, \Delta\theta_1, \Delta\theta_2, \Delta\varphi_1, \Delta\varphi_2, \Delta\varphi_3\}$. For example, the variation of linker length Δr within the time interval Δt was derived by calculating the difference between the linker length at time t and the linker length at time $t+\Delta t$, while the time t was moved from the beginning to the end of the simulation trajectory and the window Δt was fixed to reflect the temporal correlation of conformational dynamics. We calculated these variations of all six internal coordinates for all six protein systems so that their distributions can be compared across systems with different types of linkers for a specific internal coordinate. The means and standard deviations for distributions of conformational variations can be found in the **Supporting Information** as **Table S2** for all internal coordinates of six fusion proteins, as well as their correlation coefficients. Moreover, the entire distributions of conformational variations along all internal coordinates were plotted in supplemental **Figure S3** as histograms with different color bars for all the linkers. In order to show the similarity and difference between specific distributions, a few examples were further selected from the table and plotted in **Figure 6** for a more detailed comparison.

In **Figure 6a**, the variations of linker length are compared between GS15 and GS30. The distribution of Δr obtained in linker GS15 was plotted by black histograms, while the same distribution obtained in GS30 was plotted by red histograms in the figure. The time interval of 1 ns was used to calculate the variations. In **Figure 6b**, we show the comparison of variations in packing angle θ_1 between the same linkers. The black histograms give the distribution of $\Delta\theta_1$ obtained in linker GS15, and the red histograms give the distributions of linker GS30 within the same time interval of 1 ns. Both **Figure 6a** and **Figure 6b** show that distributions in linker GS30 are much wider, indicating larger variations along the internal coordinates in GS30 than in GS15. Given the same short amount of time, this statistical result suggests that the protein containing linker GS30 undergoes much more diverse conformational changes than the protein containing linker GS15 does. Different from the comparison between GS15 and GS30, the comparison between PLP15 and PLPII15 shows similar distributions of variations formed in these two linkers. The variations of linker length in linkers PLP15 and PLPII15 are plotted as black and red histograms in **Figure 6c**, while the variations of packing angle θ_1 in linkers PLP15 and PLPII15 are plotted as black and red histograms in **Figure 6d**. The

time interval used to calculate the variation is also 1 ns. The overall distributions in these two linkers are very close, except that the variation of packing angle in PLPII15 is a little larger than that in PLP15, as shown in **Figure 6d**. This is due to the fact that the left-handed helix formed in PLPII15 is more flexible than the right-handed helix formed in PLP15.

It is worth mentioning that, although the conformational spaces of our tested proteins were sampled by the atomic scale simulations, a large amount of derived information was discarded after the vector-based coarse-graining. However, the application of supercomputer Anton 2 is still a necessary strategy for generating more realistic structural ensembles for these multi-domain proteins. The simulations purely based on the lower resolution models are not sufficient to capture their global dynamics. For instance, the coarse-grained force fields, such as MARTINI, suffer from not being able to appropriately maintain the secondary structures that are present in native proteins . Especially for the systems that contain disordered regions, even the simulation based on traditional all-atom force fields and water models could lead to over-compactness compared with the estimation from experiments . Therefore, coarse-grained structural modeling based on more advanced atomic simulations is an optimal combination to explore the low-dimensional conformational dynamics of different domain linkers with both accuracy and efficiency.

In summary, based on the statistical analysis of simulations, we revealed that conformational variations are highly distinctive between certain pairs of linkers but highly similar between others. In the next part, this kinetic similarity will be systematically evaluated via a neural network algorithm among all pairs of the six linker types.

Neural network classification results in a hierarchical structure of the linker benchmark

Using the vector-based coarse-grained model, the relative positions between two modules of a fusion protein can be represented by a low-dimensional space. Based on the definition of six internal coordinates in this space, we further discretized the dynamics of the protein’s large-scale conformational changes by calculating the variation along all the

internal coordinates between two specific time steps of the Anton MD simulations. Assuming that the basic features of different linkers are incorporated in these variations, we deduce that they can be used to characterize the dynamics of inter-module correlation of fusion proteins. Practically, for a group of proteins, the characterization of their dynamics can be implemented by performing all pairwise comparisons in the group. Here, these pairwise comparisons were carried out using the neural network classification method. Given a specific time interval, the discretized conformational variations along the internal coordinates were used as inputs to the algorithm. The underlying hypothesis is that the neural network method will be more likely to identify a pair of proteins from one another if the calculated variations in these two proteins are less similar. Conversely, the neural network method will not be able to identify a pair of proteins from each other if the variations in these two proteins are highly similar.

Based on this hypothesis, and following the procedure of cross-validation described in the **Methods**, we applied a feedforward back-propagation algorithm to estimate the dynamic similarity between all pairs of fusion proteins with six different types of linkers in the benchmark. For a specific pair, we calculated the accuracy of classification from the cross-validation, defined as the total percentage of correct prediction. Our calculations for all pairwise combinations are plotted as a two-dimensional matrix in **Figure 7a**. The indexes of the linker type are listed along the x and y axes, and the color scales of accuracy are shown next to the matrix. The figure indicates that the accuracy of identifying the protein dynamics with linker GS15 from linker GS30 is higher than 95%, suggesting that the conformational variations caused by these two linkers are highly distinguishable from each other. This is what we observed in **Figure 6a** and **Figure 6b**. From the contour plot, we also find that the accuracy to identify the protein dynamics with linker PLP15 from linker PLPII15 is lower than 50%, indicating that the neural network failed to identify the differences between these two linkers. This suggests that their conformational variations are highly similar, corresponding to our observations in **Figure 6c** and **Figure 6d**. In order to test the robustness of our classification algorithm, we changed the architecture of the neural network. Different numbers of hidden layers were added. For each layer, we further tried different numbers of neurons. Detailed analysis results for different architectures of neural networks are

summarized in the supplemental **Figure S2**. The figure shows that similar patterns of accuracy for all linker pairs were obtained from different architectures of network models, indicating the robustness of our classification results.

The time interval used to construct the conformational variations for the neural network was fixed at 1 ns. In order to test the protein dynamics under a longer timescale, we generated the conformational variations of proteins with different values of time intervals and fed them into the neural network. **Figure 7b** shows the classification results between linker GS15 and GS30 when the time interval increased from 1 ns to 100 ns. For each data point, the accuracy was calculated based on the cross-validation results. The accuracy and time intervals are plotted as x and y coordinates of the curve. The figure shows that the accuracy drops rapidly when the time interval increases from 1 ns to 10 ns. After 10 ns, however, the accuracy is gradually stabilized and oscillates around a lower value. This result suggests that the differences in conformational variations between linker GS15 and GS30 are larger within a relatively shorter time scale than 10 ns.

Based on the pairwise comparison of dynamic similarity between every two linkers in the benchmark, hierarchical clustering was further applied to organize these linkers into a higher-level structure. More specifically, the top-down divisive clustering algorithm was adopted. All six linkers were initially placed in one cluster. It was then split into two by a flat clustering method and the similarity among all linkers, corresponding to the accuracy calculated by neural network classification. The splitting process was iterated until each linker was in its own singleton cluster. As a result, the linkers with the lowest similarity were divided first, while the most similar linkers were placed in different branches on the lowest level. The clustering result was plotted as a phylogenetic tree in **Figure 7c**. The figure suggests that the six linkers in the benchmark can be split into two large groups. The first one contains GS15, PLP15, and PLPII15, in which PLP15 and PLPII15 are more similar in terms of their effect on the protein's global conformational dynamics. Thus, these two linkers form a further subgroup. The other group contains the remaining three linkers—GS30, PLPII30, and PLrigid. Within this group, the conformational dynamics mediated by GS30 and PLPII30 are more similar, so they are put into a subgroup. We found that the first group contains all small

linkers with 15 amino acids, but they are further separated based on flexibility. As a result, all longer linkers are in the second group, in which they are divided again based on the number of amino acids in each linker. Therefore, our results suggest that the length of linkers (number of amino acids in the linkers) used to spatially separate different functional modules play the most important role in regulating the dynamics of proteins. Given the same number of amino acids, linker flexibility—caused by the types of amino acids in the linkers—can function as a regulator of protein dynamics at the next level.

Concluding Discussion

Polypeptide linkers are used to spatially tether two contiguous protein domains. They allow sufficient flexibility in multi-domain proteins, facilitating versatile functions through regulating their inter-domain motions . Domain linkers can vary in composition, length, and structure. As a result, we constructed a benchmark that contains six different types of peptide segments. They are formed either by repeats of flexible and hydrophilic residues or by repeats of more rigid and hydrophobic ones. To test the impacts of these linkers on the dynamics of multi-domain proteins, we applied the benchmark to fusion proteins in which two functional modules are artificially connected with different linkers. The microsecond dynamics of all fusion proteins were simulated by a supercomputer at the atomic level. We showed that all six systems in the benchmark undergo large global conformational fluctuations, even if the local structures of some linkers are relatively well-preserved. The large-scale motions between the two functional modules were further described by a limited number of coarse-grained degrees of freedom. The variations along these degrees of freedom are used as input for a multi-layer neural network to identify proteins with different linkers. While the AI algorithm could successfully recognize the difference between some linkers, it failed in some other cases, indicating that the inter-domain dynamics between these linkers are highly similar. After the similarities of dynamics between all pairs of six linkers were calculated, they could be hierarchically classified into a tree-like topology. We found a correlation between the hierarchical structure and the linker properties, suggesting that a linker is an important determinant in mediating the dynamics of this multi-domain protein. Altogether, we

demonstrated the feasibility of a new computational strategy to study protein dynamics by a combination of long timescale MD simulation, coarse-grained modeling, and AI.

This paper demonstrated the possibility of analyzing protein dynamics by combining MD simulation with AI. We admit that the algorithms applied in our study or their parameters might not be the optimal choices. For instance, CHARMM36m is the force field used in our all-atom MD simulation. It is the recently updated version of the original CHARMM36 force field . CHARMM36m has been shown to greatly improve the structural properties of conformational ensembles generated for small, disordered peptides. However, we noticed the availability of other force fields that can also be used to simulate multi-domain proteins with disordered linkers. In particular, a newly developed force field called Amber99SB-disp has been benchmarked to attain high accuracy in simulations of disordered proteins . This force field used the combination of the TIP4P-D water model and the Amber99SB-ILDN force field as a starting point. The parameters in torsional angles, as well as in the Van der Waals interactions between protein and water, were then optimized iteratively until the observed discrepancies between simulations and experimental measurements were minimized on a benchmark dataset. It would be interesting to compare the dynamic properties of domain linkers from MD simulations generated from these different force fields.

The neural network model was selected in this study simply as a tool to distinguish the dynamics between different linkers. This task can be carried out by many other classification algorithms, including principal component analysis (PCA) and the classifier based on Kullback–Leibler (KL) divergence or Mahalanobis distance. Given the distributions of conformational variations in all six linkers, as illustrated in Figure S3, we admit that the similar classification results can be archived by these traditional methods. Therefore, for the datasets compared in current study, the neural network model is not computationally superior. However, as the future extension to more complicated systems with a higher level of divergence in the dynamics of linker regions, we expect that the neural-network-based model would provide more insights. PCA is more commonly used for high-dimensional datasets. It is not very efficient for application to our vector space, which only contains six degrees of freedom. The classifiers based on

Mahalanobis distance or KL divergence assume that the underlying probabilities of comparing datasets follow Gaussian distributions . In contrast, there are no assumptions regarding the underlying distributions of datasets for neural network–based classifiers. As a result, they are more flexible in identifying datasets with unknown or complicated distribution functions. There are other more advanced machine learning–based algorithms for data classification, such as supporting vector machine (SVM) and random forest . However, a systematic comparison of performance among these approaches is beyond the scope of this paper. Finally, it is worth mentioning that ENCORE is a software package that was recently developed to compare conformational ensembles generated from computational simulations . It would be interesting to apply ENCORE to the structural ensemble derived by this study and compare the outputs with our neural network model.

In order to verify whether the 2 μs MD simulations reached convergence, we calculated the root-mean-squared average correlation (RAC) for all six systems. The RAC function as recently developed to quantitatively analyze the convergence of time-series data under different time scales of a single trajectory . A more detailed definition of RAC can be found in the **Supporting Information** and supplemental **Figure S1**. **Figure S1a** shows that all calculated RAC curves decay as time interval increases. In order to further assess the convergence within a given simulation more carefully, we plotted the amplitude of slope in the RAC curves of all six systems in **Figure S1b**. The figure shows that the slopes of all RAC curves approach to zero when the length of time intervals increases to its maximal value. A closer look to the slope of RAC curves at the longest time scales is plotted by the inserted panel. The figure shows that although the amplitudes of slopes in all six systems are very small, they are still above 0, as indicated by the orange dashed line in the panel. This result suggests that the MD simulations in these six systems have not completely converged yet. This observation, however, would not affect the major conclusion drawn from our classification results. Instead of the entire conformational distributions, we are mainly focusing on the global conformational changes of a fusion protein within a short time scale of nanosecond, and further comparing these changes among proteins with different linkers. The uniform or Gaussian-like distributions of these conformational changes (**Figure S3**) indicate that the statistical convergence about the information of nanosecond-scale conformational

changes has been captured. The differences of conformational changes between linkers can further be appropriately characterized by the variations in these distributions. Therefore, we believe that our classification results would not be significantly changed by using MD simulations with longer timescale.

We have shown that the dynamics of bi-specific biologics studied in this work are closely related to the properties of the linker region. The dynamic properties of different linkers can be validated by various experimental approaches. For instance, the linker dynamics embedded in the structure ensembles generated from MD simulations can be captured by the profiles derived from small angle scattering (SAXS) experiments . The conformational dynamics of linkers are also reflected by variations in the distance between MHC and PD-L1, which can be validated by experiments, such as fluorescent resonance energy transfer (FRET) experiments. By further assuming that biologics with linkers of similar dynamic features have similar binding properties with their targeted receptors, our results offer the possibility to design linkers of multi-specific biologics so that their binding behaviors can be modulated. For instance, intuitively, the linkers in the subgroups of GS30 and PLPII30 can capture more cell surface receptors than other groups can due to their ability to search the local conformational space more thoroughly. As a result, we expect that these types of linkers are more sensitive to mediate the T cell co-regulatory pathways; this prediction can potentially be validated by T cell stimulation assays. The detailed analysis of how different linkers can regulate the binding between biologics and their corresponding receptors on T cell surfaces is beyond the scope of this work. Such research will be conducted in a follow-up study that integrates the MD simulations into our previously developed multiscale modeling framework .

Finally, our method serves as a foundation to evaluate the dynamics of other multi-domain biologics in which linkers are used to fuse different protein modulators. For instance, bi-specific T cell engagers (BiTEs) are a class of immunotherapeutic molecules that stimulate cytokine production by physically linking a T cell to a tumor cell . These molecules are also constructed of two protein fragments connected by a linker. While one fragment binds to a T cell-specific molecule, such as CD3, the other binds to an antigen on tumor cells. By future extension of our method, we will be able to estimate the

impacts of linkers on the dynamics of these systems and provide further insights into their effectiveness in T cell activation and tumor killing.

Acknowledgment

This work was supported by the National Institutes of Health under grant number R01GM120238. The work was also partially supported by a start-up grant from Albert Einstein College of Medicine. Computational support was provided by Albert Einstein College of Medicine High Performance Computing Center. Anton 2 computer time was provided by the Pittsburgh Supercomputing Center (PSC) through grant R01GM116961 from the National Institutes of Health. The Anton 2 machine at PSC was generously made available by D.E. Shaw Research.

Author Contributions

B.W. and Y.W. designed the research; B.W. and Y.W. performed the research; B.W., Z. S., and Y.W. analyzed the data; B.W. and Y.W. wrote the paper.

Additional Information

Competing Financial Interests: The authors declare no competing financial interests.

References

1. Bagowski, C.P., W. Bruins, and A.J. Te Velthuis, *The nature of protein domain evolution: shaping the interaction network*. Curr Genomics, 2010. **11**(5): p. 368-76.
2. Vogel, C., et al., *Structure, function and evolution of multidomain proteins*. Curr Opin Struct Biol, 2004. **14**(2): p. 208-16.
3. Yu, K., et al., *Synthetic fusion protein design and applications*. Biotechnol Adv, 2015. **33**(1): p. 155-164.
4. Wriggers, W., S. Chakravarty, and P.A. Jennings, *Control of protein functional dynamics by peptide linkers*. Biopolymers, 2005. **80**(6): p. 736-46.
5. Chen, X., J.L. Zaro, and W.C. Shen, *Fusion protein linkers: property, design and functionality*. Adv Drug Deliv Rev, 2013. **65**(10): p. 1357-69.
6. Bhaskara, R.M., A.G. de Brevern, and N. Srinivasan, *Understanding the role of domain-domain linkers in the spatial orientation of domains in multi-domain proteins*. J Biomol Struct Dyn, 2013. **31**(12): p. 1467-80.
7. Baldo, B.A., *Chimeric fusion proteins used for therapy: indications, mechanisms, and safety*. Drug Saf, 2015. **38**(5): p. 455-79.
8. Bai, Y., D.K. Ann, and W.C. Shen, *Recombinant granulocyte colony-stimulating factor-transferrin fusion protein as an oral myelopoietic agent*. Proc Natl Acad Sci U S A, 2005. **102**(20): p. 7292-6.
9. Bai, Y., and W.C. Shen, *Improving the oral efficacy of recombinant granulocyte colony-stimulating factor and transferrin fusion protein by spacer optimization*. Pharm Res, 2006. **23**(9): p. 2116-21.
10. Berger, S., P. Lowe, and M. Tesar, *Fusion protein technologies for biopharmaceuticals: applications and challenges: Editor Stefan R Schmidt*. mAbs, 2015. **7**(3): p. 456-60.
11. Perilla, J.R., et al., *Molecular dynamics simulations of large macromolecular complexes*. Curr Opin Struct Biol, 2015. **31**: p. 64-74.
12. Hollingsworth, S.A., and R.O. Dror, *Molecular dynamics simulation for all*. Neuron, 2018. **99**(6): p. 1129-43.
13. Wieczorek, G., and P. Zielenkiewicz, *Influence of macromolecular crowding on protein-protein association rates--a Brownian dynamics study*. Biophys J, 2008. **95**(11): p. 5030-6.
14. Ermakova, E., *Lysozyme dimerization: Brownian dynamics simulation*. J Mol Model, 2005. **12**(1): p. 34-41.
15. Haddadian, E.J., and E.L. Gross, *A Brownian dynamics study of the interactions of the luminal domains of the cytochrome b6f complex with plastocyanin and cytochrome c6: the effects of the Rieske FeS protein on the interactions*. Biophys J, 2006. **91**(7): p. 2589-600.
16. Roy, A., D.P. Hua, and C.B. Post, *Analysis of multidomain protein dynamics*. J Chem Theory Comput, 2016. **12**(1): p. 274-80.
17. Shaw, D.E., et al., *Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer*, in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 2014, IEEE Press: New Orleans, Louisiana. p. 41-53.
18. Angermueller, C., et al., *Deep learning for computational biology*. Mol Syst Biol, 2016. **12**(7): p. 878.
19. Hassanien, A.E., E.T. Al-Shammari, and N.I. Ghali, *Computational intelligence techniques in bioinformatics*. Comput Biol Chem, 2013. **47**: p. 37-47.

20. Wang, J., et al., *Machine learning of coarse-grained molecular dynamics force fields.* ACS Cent Sci, 2019. **5**(5): p. 755-67.
21. Degiacomi, M.T., *Coupling molecular dynamics and deep learning to mine protein conformational space.* Structure, 2019. **27**(6): p. 1034-40 e3.
22. Sun, C., R. Mezzadra, and T.N. Schumacher, *Regulation and function of the PD-L1 checkpoint.* Immunity, 2018. **48**(3): p. 434-52.
23. Chen, L., and D.B. Flies, *Molecular mechanisms of T cell co-stimulation and co-inhibition.* Nat Rev Immunol, 2013. **13**(4): p. 227-42.
24. Locksley, R.M., N. Killeen, and M.J. Lenardo, *The TNF and TNF receptor superfamilies: integrating mammalian biology.* Cell, 2001. **104**(4): p. 487-501.
25. MacEwan, D.J., *TNF ligands and receptors--a matter of life and death.* Br J Pharmacol, 2002. **135**(4): p. 855-75.
26. Sedger, L.M., and M.F. McDermott, *TNF and TNF-receptors: from mediators of cell death and inflammation to therapeutic giants - past, present and future.* Cytokine Growth Factor Rev, 2014. **25**(4): p. 453-72.
27. Wei, S.C., C.R. Duffy, and J.P. Allison, *Fundamental mechanisms of immune checkpoint blockade therapy.* Cancer Discov, 2018. **8**(9): p. 1069-1086.
28. Samanta, D., et al., *Structural and functional characterization of a single-chain peptide-MHC molecule that modulates both naive and activated CD8+ T cells.* Proc Natl Acad Sci U S A, 2011. **108**(33): p. 13682-7.
29. Pascolutti, R., et al., *Structure and dynamics of PD-L1 and an ultra-high-affinity PD-1 receptor mutant.* Structure, 2016. **24**(10): p. 1719-28.
30. van Rosmalen, M., M. Krom, and M. Merkx, *Tuning the flexibility of glycine-serine linkers to allow rational design of multidomain proteins.* Biochemistry, 2017. **56**(50): p. 6565-74.
31. Fiser, A., and A. Sali, *ModLoop: automated modeling of loops in protein structures.* Bioinformatics, 2003. **19**(18): p. 2500-1.
32. George, R.A., and J. Heringa, *An analysis of protein domain linkers: their classification and role in protein folding.* Protein Eng, 2002. **15**(11): p. 871-9.
33. Kumar, P., and M. Bansal, *Structural and functional analyses of PolyProline-II helices in globular proteins.* J Struct Biol, 2016. **196**(3): p. 414-25.
34. Sommese, R.F., et al., *Helicity of short E-R/K peptides.* Protein Sci, 2010. **19**(10): p. 2001-5.
35. Shaw, D.E., et al., *Millisecond-scale molecular dynamics simulations on Anton*, in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. 2009, ACM: Portland, Oregon. p. 1-11.
36. Piana, S., et al., *Water dispersion interactions strongly influence simulated structural properties of disordered protein states.* J Phys Chem B, 2015. **119**(16): p. 5113-23.
37. Henriques, J., C. Cragnell, and M. Skepo, *Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment.* J Chem Theory Comput, 2015. **11**(7): p. 3420-31.
38. Shan, Y., et al., *Gaussian split Ewald: A fast Ewald mesh method for molecular simulation.* J Chem Phys, 2005. **122**(5): p. 54101.
39. Wu, Y., et al., *OPUS-Dom: applying the folding-based method VECFOLD to determine protein domain boundaries.* J Mol Biol, 2009. **385**(4): p. 1314-29.
40. Li, J., et al., *Brief Introduction of back propagation (BP) neural network algorithm and its improvement*, in *Advances in Computer Science and Information Engineering: Volume 2*, D. Jin and S. Lin, Editors. 2012, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 553-8.

41. Kabsch, W., *A solution for the best rotation to relate two sets of vectors*. Acta Crystallographica Section A, 1976. **32**(5): p. 922-3.
42. Adzhubei, A.A., and M.J.E. Sternberg, *Left-handed polyproline II helices commonly occur in globular proteins*. J Mol Biol, 1993. **229**(2): p. 472-93.
43. Galindo-Murillo, R., D.R. Roe, and T.E. Cheatham, 3rd, *Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC)*. Biochim Biophys Acta, 2015. **1850**(5): p. 1041-58.
44. Kar, P., and M. Feig, *Recent advances in transferable coarse-grained modeling of proteins*. Adv Protein Chem Struct Biol, 2014. **96**: p. 143-80.
45. Robustelli, P., S. Piana, and D.E. Shaw, *Developing a molecular dynamics force field for both folded and disordered protein states*. Proc Natl Acad Sci U S A, 2018. **115**(21): p. E4758-66.
46. Endo, T., S. Ogishima, and H. Tanaka, *Standardized phylogenetic tree: a reference to discover functional evolution*. J Mol Evol, 2003. **57 Suppl 1**: p. S174-81.
47. Reddy Chichili, V.P., V. Kumar, and J. Sivaraman, *Linkers in the structural biology of protein-protein interactions*. Protein Sci, 2013. **22**(2): p. 153-67.
48. Huang, J., et al., *CHARMM36m: an improved force field for folded and intrinsically disordered proteins*. Nat Methods, 2017. **14**(1): p. 71-73.
49. Hughen, J., K. Hollon, and D. Lai, *Comparison of Mahalanobis distance, polynomial, and neural net classifiers*. 1990 Technical Symposium on Optics, Electro-Optics, and Sensors. Vol. 1294. 1990: SPIE.
50. Chen, J., B. Wang, and Y. Wu, *Structural characterization and function prediction of immunoglobulin-like fold in cell adhesion and cell signaling*. J Chem Inf Model, 2018. **58**(2): p. 532-42.
51. Tiberti, M., et al., *ENCORE: software for quantitative ensemble comparison*. PLoS Comput Biol, 2015. **11**(10): p. e1004415.
52. Wu, Y., et al., *Folding of small helical proteins assisted by small-angle X-ray scattering profiles*. Structure, 2005. **13**(11): p. 1587-97.
53. Chen, J. and Y. Wu, *A multiscale computational model for simulating the kinetics of protein complex assembly*. Methods Mol Biol, 2018. **1764**: p. 401-11.
54. Wang, B., et al., *Integrating structural information to study the dynamics of protein-protein interactions in cells*. Structure, 2018.
55. Huehls, A.M., T.A. Coupet, and C.L. Sentman, *Bispecific T-cell engagers for cancer immunotherapy*. Immunol Cell Biol, 2015. **93**(3): p. 290-6.

Figure Legends

Figure 1: In our test system, we computationally fused two functionally independent protein modules with six types of peptide linker. The structural models of these fusion proteins are shown in the figure. One module of the fusion proteins is MHC, which is shown in red, while the other is PD-L1, shown in green. The linkers connecting these two modules are shown in gray. The names of these linkers are as follows: **(a)** GS15, **(b)** GS30, **(c)** PLP15, **(d)** PLPII15, **(e)** PLPII30, and **(f)** PLrigid. We further coarse-grained these protein structures by a vector-based model so that the variations of protein structure between two functional modules could be represented by six internal coordinates **(g)**.

Figure 2: We analyzed the RMSD of six proteins from their MD simulation results. The local RMSD of the residues only in the linkers are plotted in **(a)** as a function of simulation time. The global RMSD for all residues in the fusion proteins is plotted in **(b)**. The colors of the curves correspond to the type of linkers, which are given on top of the figure. The figure shows that the levels of local structural dynamics are highly dependent on the sequence composition of different linkers. In contrast, the orientations between two modules show large fluctuations for all six fusion proteins, even if the linker is rigid.

Figure 3: The linker of PLrigid shows a transition from low RMSD to high RMSD around 0.15 μs after equilibrium. Some snapshots of the linker configuration were taken from the trajectory around this time window. These snapshots indicate that the linker still had the helical structure at the time of 0.125 μs after equilibrium **(a)**. However, soon after that, the secondary structure at both ends of the linker started to melt, as shown in **(b)** and **(c)**. At 0.2 μs after equilibrium, all hydrogen bonds in the helical structure of the linker were broken **(d)**.

Figure 4: We notice that although the local RMSD of linker PLP15 is lower than 2 Å through all the 2 μs simulations, the global RMSD of the entire protein with PLP15 linker on average is around 20 Å. Therefore, a few snapshots were taken from the trajectory of the protein with linker PLP15 to show its dynamics of global configuration. The snapshot in **(a)** is at the time of 0.25 μs after equilibrium. The snapshot in **(b)** is at the time of 0.5 μs after equilibrium. The snapshot in **(c)** is at the time of 1.0 μs after equilibrium. The

snapshot in **(d)** is at the time of 2.0 μs after equilibrium. The MHC module is highlighted in red, while the PD-L1 module is highlighted in green, and the linker in the middle is highlighted in gray.

Figure 5: We calculated the conformational distribution along the six internal degrees of freedom from the MD simulations and compared them between different linkers. Specifically, the distribution of linker length r in the protein with linker PLP15 is compared with linker GS30 in **(a)**. Similarly, the comparisons of two packing angles and dihedral between these two linkers are plotted in **(b)** and **(c)**, respectively.

Figure 6: Information about protein conformational changes within a given time window was obtained by analyzing the variations of six internal coordinates within the time window Δt. We calculated these variations along six internal coordinates and compared them across systems with different types of linkers. Specifically, the variations of linker length are compared between GS15 and GS30 in **(a)**. In **(b)**, we show the comparison of these two linkers for their variations in packing angle θ_1. In **(c)** and **(d)**, the variations of linker length and packing angle θ_1 are compared between PLP15 and PLPII15, respectively.

Figure 7: We applied the neural network algorithm to estimate the dynamic similarity between all six different types of linkers in the benchmark. For a specific pair, we calculated the accuracy of classification from the cross-validation. All pairwise combinations of calculated accuracy are plotted as a two-dimensional matrix in **(a)**. We further increased the time interval from 1 ns to 100 ns. For each time interval, we calculated the accuracy based on the cross-validation results. The accuracy to identify GS15 from GS30 is plotted in **(b)** as a function of time interval. Finally, based on the pairwise comparison of similarity among all linkers, we organized them into a hierarchic structure, which is plotted as a phylogenetic tree in **(c)**.

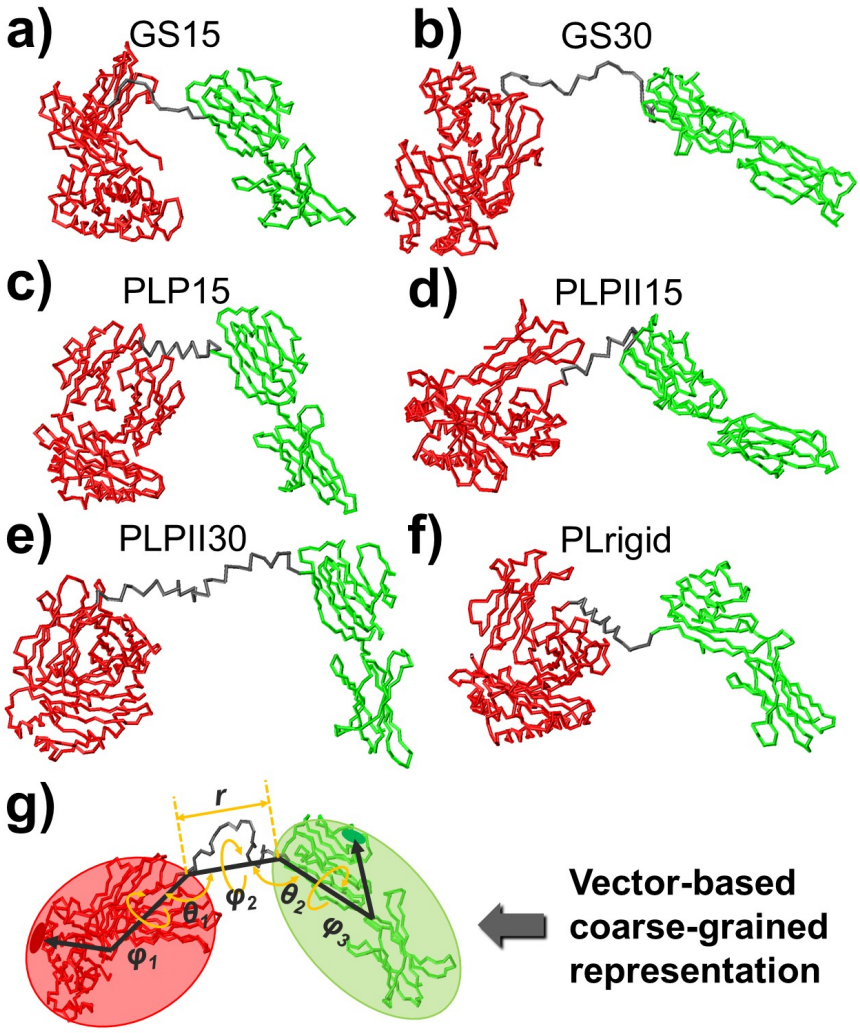
a)
GS15
b)
GS30
c)
PLP15
d)
PLPII15
e)
PLPII30
f)
PLrigid
g)
r
θ1
φ2
θ2
φ3
φ1
Vector-based
coarse-grained
representation

Figure 1

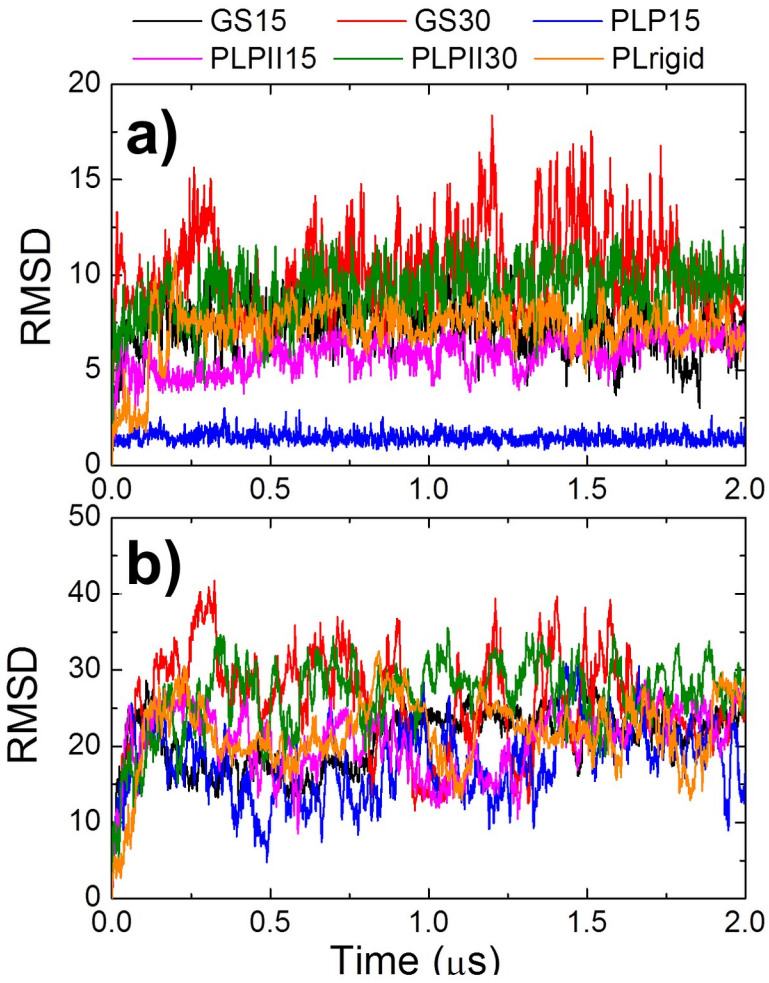
GS15
GS30
PLP15
PLPII15
PLPII30
PLrigid
a)
b)
RMSD
RMSD
Time (μs)
0
5
10
15
20
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50

Figure 2

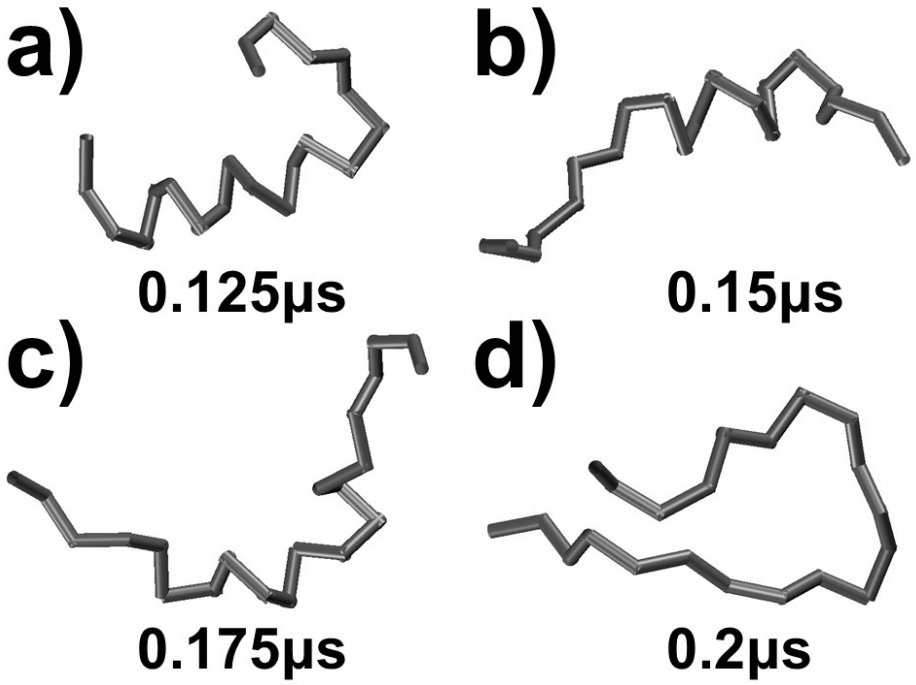

Figure 3

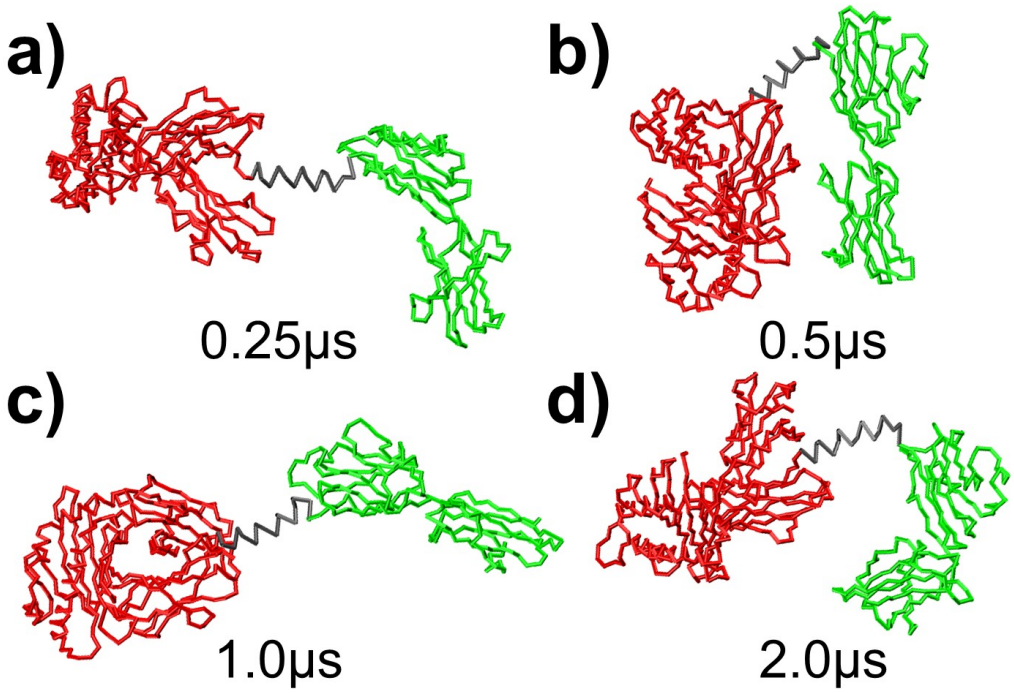

Figure 4

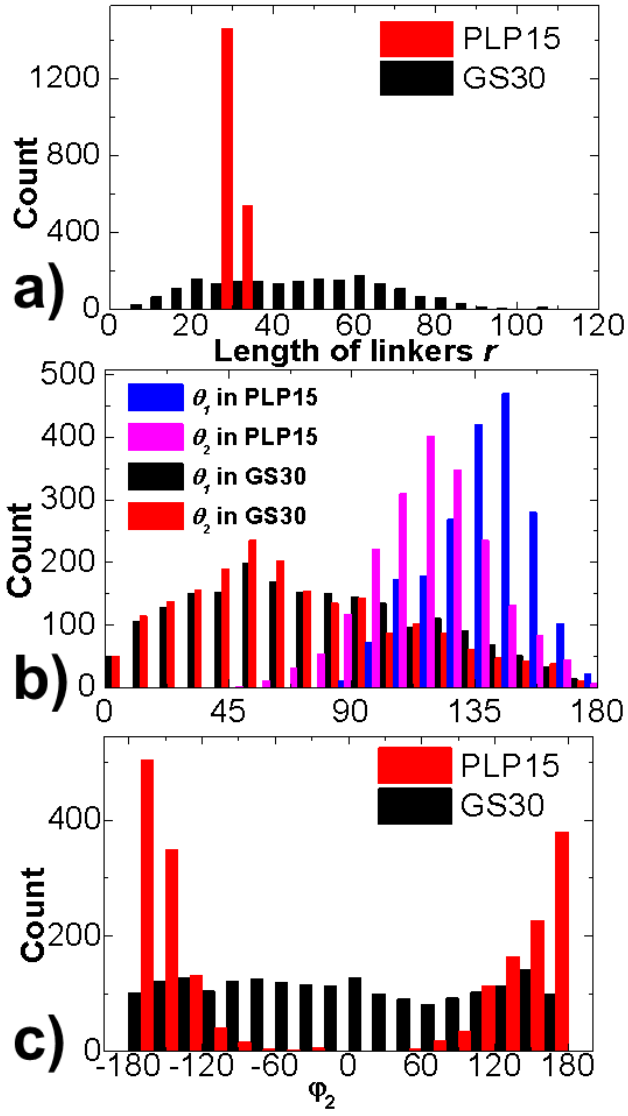
PLP15
GS30
Count
a)
Length of linkers r
θ1 in PLP15
θ2 in PLP15
θ1 in GS30
θ2 in GS30
b)
c)
φ2

Figure 5

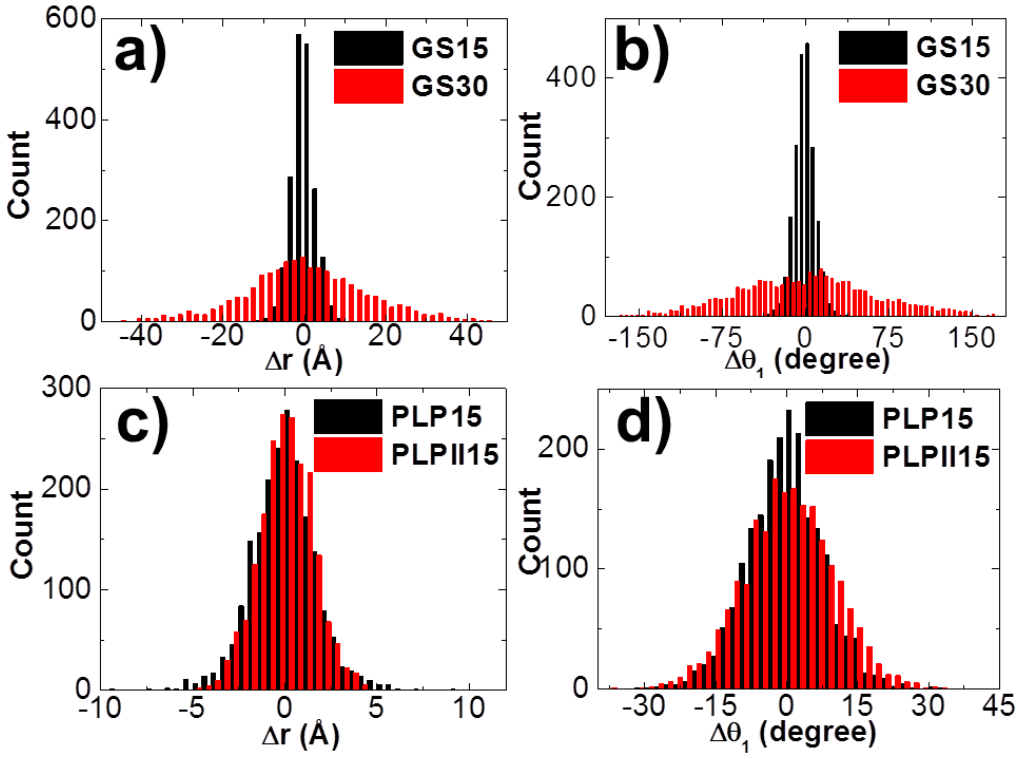
a)
GS15
GS30
Count
Δr (Å)
b)
GS15
GS30
Count
Δθ1 (degree)
c)
PLP15
PLPII15
Count
Δr (Å)
d)
PLP15
PLPII15
Count
Δθ1 (degree)

Figure 6

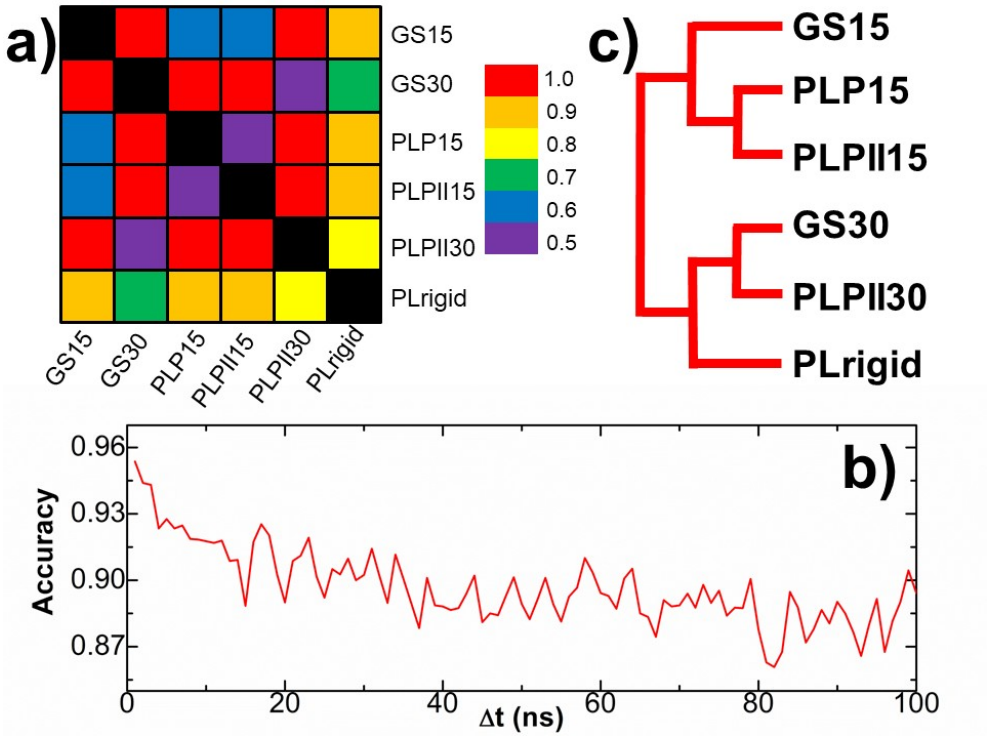
a)
GS15
GS30
PLP15
PLPII15
PLPII30
PLrigid
1.0
0.9
0.8
0.7
0.6
0.5
c)
GS15
PLP15
PLPII15
GS30
PLPII30
PLrigid
b)
Accuracy
0.96
0.93
0.90
0.87
0
20
40
60
80
100
Δt (ns)

Figure 7

Linker Name	Dynamic Property	Initial Structure	Sequence
GS15	Flexible	Loop	(GGGGS)×3
GS30	Flexible	Loop	(GGGGS)×6
PLP15	Rigid	Right-handed helix	15 prolines
PLPII15	Medium	Left-handed helix	15 prolines
PLPII30	Medium	Left-handed helix	30 prolines
PLrigid	Rigid	Right-handed helix	(EAAAR)×4

Table 1: Detailed information about the six linkers used in this study