JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 4691−4705

pubs.acs.org/jcim

Probing Protein Allostery as a Residue-Specific Concept via Residue Response Maps

Hamed S. Hayatshahi,[†] Emilio Ahuactzin,[‡] Peng Tao,[§] Shouyi Wang,[∥] and Jin Liu*[,†]

[†]Department of Pharmaceutical Sciences, University of North Texas System College of Pharmacy, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, Texas 76107, United States

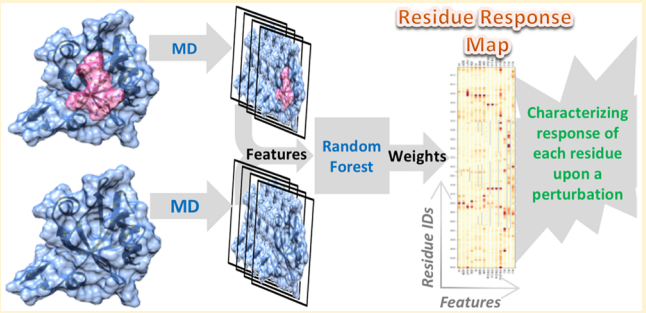
[‡]Harmony School of Innovation-Fort Worth, 8100 S. Hulen St., Fort Worth, Texas 76123, United States

[§]Department of Chemistry, Center for Drug Discovery, Design, and Delivery (CD4), Center for Scientific Computation, Southern Methodist University, Dallas, Texas 75275, United States

[∥]Department of Industrial, Manufacturing, & Systems Engineering, College of Engineering, University of Texas at Arlington, 701 S. Nedderman Dr., Arlington, Texas 76019, United States

Supporting Information

ABSTRACT: Allosteric regulation is a well-established phenomenon defined as a distal conformational or dynamical change of the protein upon allosteric effector binding. Here, we developed a novel approach to delineate allosteric effects in proteins. In this approach, we applied robust machine learning methods, including deep neural network and random forest, on extensive molecular dynamics (MD) simulations to distinguish otherwise similar allosteric states of proteins. Using the PDZ3 domain of PDS-95 as a model protein, we demonstrated that the allosteric effects could be represented as residue-specific properties through two-dimensional property-residue maps, which we refer to as "residue response maps". These maps were constructed through two machine learning methods and could accurately describe how different properties of various residues are affected upon allosteric perturbation on protein. Based on the "residue response maps", we propose allostery as a residue-specific concept, suggesting that all residues could be considered as allosteric residues because each residue "senses" the allosteric events through changing its single or multiple attributes in a quantitatively unique way. The "residue response maps" could be used to fingerprint a protein based on the unique patterns of residue responses upon binding events, providing a novel way to systematically describe the protein allosteric effects of each residue upon perturbation.



INTRODUCTION

Allostery has been an evolving concept.[1] The traditional definition considers the allosteric proteins as two-state switches that are concertedly[2] or sequentially[3] affected by an effector molecule. Some modern definitions consider allosteric proteins as conformational ensembles whose populations are shifted upon binding to an effector.[4−6] From a mechanistic viewpoint, binding to an effector propagates a signal through changing properties of a network of residues, including their conformational dynamics[7] and nonbonding interaction attributes,[8] which may not accompany observable conformational changes.[9] Therefore, it has been proposed that all proteins are intrinsically allosteric[10] because even for a classically considered nonallosteric protein, subtle conformational or dynamical changes can occur upon binding to a ligand. Recently, Berezovski and co-workers proposed that allosteric residues could be considered as a result of the allosteric communication induced via ligand binding and developed methods to detect allosteric sites based on per-residue energetic perturbation upon ligand binding, suggesting ligand binding may perturb the free energy of each residue.[11,12] Based on these insights, we hypothesize that the ligand-binding changes multiple attributes of a network of residues. Our hypothesis raises the question of whether the different attributes are perturbed to the same extent in all affected residues or each residue is affected differently. In other words, could different allosteric residues "sense" effector binding in different ways? With this broader view, we further hypothesize that all residues in a given protein are potentially allosteric with various responses and extents of response upon perturbation.

Here, we tested these hypotheses in the PDZ3 domain of PSD-95. PDZ3 is a well-known model for analyzing allosteric effects.[8,13−20] Two crystal structures of PDZ3 in bound and unbound states are available.[21] The bound structure includes a five-residue peptide ligand, which is bound in a groove walled with a helix (αB) and a sheet (βB). Petit et al. used isothermal titration calorimetry (ITC) and nuclear magnetic resonance

Received: May 31, 2019
Published: October 7, 2019

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

(NMR) to highlight dynamic allostery relative to an α-3 helix in PDZ3.[14] Gerek and Ozkan highlighted residues involved in allosteric pathways using perturbation response scanning (PRS).[16] McLaughlin et al. identified different mutational routes that could affect the binding specificity in PDZ3 via a high-throughput single mutation analysis.[17] In another work, Murciano-Calles et al. reported the allosteric effects of post-translational modifications of residues in PDZ3.[19] Recently, computational methods have been employed by multiple research groups to study the allosteric effects in PDZ3. Among these works is a Monte Carlo path generation simulation by Kaya et al., which revealed potential propagation routes of the allosteric signals in some proteins including PDZ3.[18] Kalescky et al. used molecular dynamics (MD) simulations to perform a rigid residue scan in the bound and unbound PDZ3 and identified allosteric residues that matched with previous experimental observations.[20] A more recent study by Kumawat and Chakrabarty revealed of the electrostatic interactions as the most significant hidden basis of dynamic allostery in PDZ3 using MD simulations.[8]

Besides PDZ3, MD simulations have been widely used to study allosteric effects. In one of the MD-based works by Van Wart et al., dynamical network analysis[22] was used to identify correlations among residue positions in MD trajectories to track down the allosteric networks between residues.[23] This approach implemented a graph theory on protein MD trajectories that assumes a time-dependent variable of protein residues as "nodes" and connects them with "edges". It was later employed and further developed by the same group to identify suboptimal paths that potentially convey allosteric messages among residues and was presented as an online program called weighted implementational of suboptimal paths (WISP).[24] Using this approach, they successfully identified allosteric signaling paths in an amidotransferase called HisH-HisF. The same theory was implemented in calculating the correlations in the MD trajectory and identifying the allosteric networks in thrombin.[25]

Kokh et al. investigated the local communication among residues to identify transiently formed binding pockets from MD trajectories and developed a program named TRAPP.[26] It analyzes the correlated pocket variations via principal component analysis (PCA), and calculates average deviations of the structure in the trajectory from a reference structure, followed by clustering the transient pockets into subpockets. As a result, the shape of the transient pockets can be recognized, and their similarity to other known pockets and their druggability can be analyzed. La Sala et al. have also presented another method for pocket detection and allosteric pocket−pocket communications via MD simulations.[27] They used the solvent-excluded surface to detect pockets in each MD snapshot and then track the exchange of atoms between adjacent pockets along the trajectory. A sequential tracking of pocket changes along the trajectory leads to the recognition of allosteric pathways between distantly located pockets.

Conformational dynamics (or entropic allostery) have also been studied via MD simulations as a mechanism for allosteric regulation.[7] In one such work, Guo, Pang, and Zhou have reported that binding of an effector to the WW domain of a protein called Pin1 changes the dynamics of the loops around the catalytic site, which is located in another domain of the protein.[28,29] This was found through performing MD simulations with artificial restraints on the WW domain, which led to similar effects on the catalytic domain as the effector binding. They also differentiated between fast time scale local dynamics and slow time scale global dynamics in allosteric signaling of this protein. Interestingly, they showed that binding the substrate to the catalytic site does not have the same dynamics effect on the WW domain where the effector binds. In contrast, Hertig et al. have described allostery as a bidirectional phenomenon.[30] In β_2AR, they also showed that changes in one direction propagate at a different speed from the changes in the opposite direction.[31] However, in a previous case, the researchers in the same group have introduced "artificial perturbations" as a technique to track down the allosteric signals with MD simulations.[32] They used steering MD as a means of artificial perturbation in fibronectin to allow force-based transitions to happen faster than a real biological system. This led to the discovery of allosteric signaling that had not previously seen in experimental studies.

Although many research works in the past tried to distinguish different mechanisms of allosteric response, it is still interesting to reveal whether different residues have the same or different contributions to the allosteric response. It is also not clear whether a certain property or a set of properties of each residue involved in an allosteric response changed upon a perturbation such as a ligand binding. Also, the potential capabilities of machine learning in identifying responses of each residue are not yet well explored. In this study, we use snapshots from MD simulations to characterize the response of each residue to a ligand-binding event. To determine the attribute/residue-specific responses upon ligand binding, we used hybrid models to calculate the extent of property responses in different residues upon ligand binding. These hybrid models would enjoy the sampling power of MD simulations in combination with predicting capabilities of machine learning methods. One set of models uses deep learning neural networks to compare the prediction accuracies of PDZ3 binding status using different residue attributes as descriptors. These models reveal whether residue attributes, such as position, nonbonding interaction, or dynamics, have different abilities to predict the binding status and hence are affected upon binding at different extent. The other set of models uses random forest to rank the residues based on their contribution to distinguish the PDZ3 binding state. These models assign significance indices to each residue in terms of its different attributes and show whether the attributes are perturbed to the same extent in all affected residues. We present the significance indices as two-dimensional diagrams called residue response maps. These maps fingerprint each protein in terms of its interaction with a specific binding ligand in a visually easy-to-follow fashion.

Machine learning models have been used widely to study the structures and interactions among biomolecules.[33,34] However, one of the challenges in using these methods, especially the deep learning neural networks is the requirement of large data set needed to train them. What empowers the machine learning models in our approach is that we integrate them with MD simulations, which provide a sufficient sampling pool of snapshots to train and validate the deep neural networks and the random forest models. The scheme of this approach is represented in Figure 1.

As shown in Figure 1, the modeling approach presented here starts with two sets of MD simulations. One set is initiated from the crystal structure of the unbound PDZ3 (ligand-free), whereas the other set is from the structures with PDZ3 complexed with the ligand (ligand-bound). Multiple residue-

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

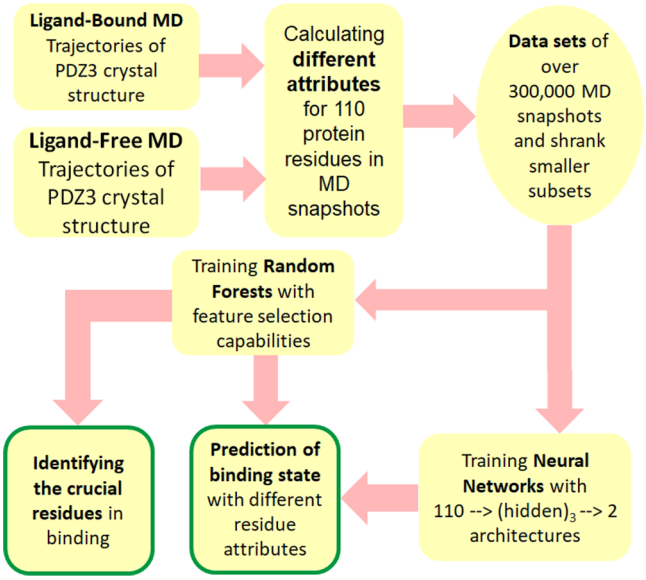


Figure 1. Scheme of the approach proposed in this work.

specific attributes are calculated for each MD snapshot. Each snapshot is also labeled based on whether it originated from the ligand-bound or ligand-free MD trajectories. With data of each residue-specific attribute, we generated a data set that contains records of MD snapshots from these two sets of MD simulations. We used each data set with a specific attribute to train deep neural networks and random forest models to predict whether each MD snapshot is from ligand-free or ligand-bound simulations. The accuracy of the prediction shows the prediction power of each attribute. We further ranked the residues based on their contribution to the prediction to identify residues with the strongest response using the feature weight values of the random forest models.

METHODS

MD Simulations and analysis. Molecular dynamics (MD) simulations were performed for the PDZ3 protein in its bound and unbound forms. To set up the MD simulations, residues 306−415 in the crystal structures of the bound and unbound PDZ3 (PDB codes: 1BE9 and 1BFE, respectively) were used as initial structures, respectively, and the rest of the amino acids were removed. The bound peptide was retained in the bound structure. tLEaP of Ambertools 16[35] was used to solvate the structures in octahedral boxes of TIP3P waters[36] with chloride ions to neutralize the system and 150 mM excess NaCl with Joung−Cheatham parameters[37] to resemble the physiological conditions. The Amber FF14SB force field was used to parametrize the protein atoms. Two copies of each simulation were provided. The starting structures were equilibrated with nine steps of minimization followed by the equilibration protocol described in a previous work.[38] Hydrogen mass repartitioning[39,40] was used to facilitate a 4 fs time step in the production phase, and SHAKE[41] was used to constrain bonds involving hydrogen atoms.

Each simulation copy was run for 1 μs as production using pmemd.cuda of Amber 16.[35] The production phase was done in 300 K with a Langevin thermostat[42,43] at constant pressure with a Monte Carlo barostat[44] with 5 ps pressure relaxation time. Particle mesh Ewald conditions[45,46] with a 9.0 Å cutoff were used for calculating the long-range interactions. The snapshots were saved to trajectories in every 1 ps. Snapshots with a minimum distance of more than 12.5 Å between the peptide and protein atoms were filtered out from the peptide bound trajectories. One in every 10 snapshots of the remaining frames (total of 338,538 bound and unbound frames) were used for further analysis, from which 59% were unbound structures and 41% were bound structures. The program cpptraj[47] of Ambertools 16[35] was used to analyze the trajectories.

Generating Descriptors. Different types of descriptors, defined as follows, were generated using a combination of cpptraj and python scripts for 110 residues (306−415) of each MD simulation snapshot:

(1) C_α descriptor was defined to describe the backbone position of the residues. It is the distance between the C_α atoms to the geometric center of the protein in each snapshot.

(2) GEOM descriptor was also defined to describe the position of the residues. It is the distance between the residues' geometric centers and the geometric center of the protein in each MD snapshot. This descriptor includes the side-chain positions in the calculations.

(3) FLUCT descriptor was defined to represent the projection of the residue fluctuation on each MD snapshot. To calculate the FLUCT values, we calculated the average protein structure by fitting all protein snapshots on the first MD snapshot, fitted all protein snapshots to this average structure, and calculated the FLUCT value of each residue at each snapshot as the root-mean-square distance (RMSD) between its structure in the snapshot and the protein average structure. The RMSD was calculated as follows:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \tag{E1}$$

where d_i is the distance between atom i in a given snapshot and the average structure, and N is the total number of atoms (hydrogen atoms are ignored here). FLUCT shows how far the structure of each residue at each snapshot is from its average structure.

(4) VDW descriptor represents the van der Waals interaction between each residue of the protein and all other residues at each MD snapshot. The VDW value for each residue (r) at each MD snapshot was calculated as follows:

$$\mathrm{VDW} = E^{\mathrm{t}}_{\mathrm{vdw}} - E^{\sim \mathrm{r}}_{\mathrm{vdw}} - E^{\mathrm{r}}_{\mathrm{vdw}} \tag{E2}$$

where $E^{\mathrm{t}}_{\mathrm{vdw}}$ is the total van der Waals interactions between all the atom pairs (ligand not included if exists), $E^{\sim \mathrm{r}}_{\mathrm{vdw}}$ is the total van der Walls interactions between all the atom pairs except atoms of residue (r), and $E^{\mathrm{r}}_{\mathrm{vdw}}$ is the total van der Waals interactions between the atom pairs in the residue (r). The cutoff for these calculations was set to 999 Å.

(5) ES descriptor represents the electrostatic interaction between each residue pair in the protein and all other residues in each MD snapshot. The ES value for each residue (r) was calculated as follows:

$$\mathrm{ES} = E^{\mathrm{t}}_{\mathrm{es}} - E^{\sim \mathrm{r}}_{\mathrm{es}} - E^{\mathrm{r}}_{\mathrm{es}} \tag{E3}$$

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

where E^{t}_{es} is the total electrostatic interactions between all atom pairs (ligand not included if exists), $E^{\sim r}_{es}$ is the total electrostatic interactions between all atom pairs except atoms in residue (r), and E^{r}_{es} is the total electrostatic interactions between atom pairs in residue (r). The cutoff for these calculations was set to 999 Å.

(6) NONB descriptor represents the total nonbonding interactions between each residue and all other residues in the protein. It is the sum of the VDW and ES values.

(7) SURF descriptor represents the contribution of atoms in each amino acid to the surface area of the free protein in Å^2 using the LCPO algorithm by Weiser et al.[48] (ligands were ignored in ligand-bound snapshots).

(8) DYN descriptors are representations of the dynamics of the residues in 10 ps, 1 ns, 25 ns, and 100 ns time scales. The DYN values are calculated as running RMSD in Å between heavy atoms of each residue at any MD snapshot and the same residue at a snapshot that occurs at the length of a given time scale earlier in the trajectory after the two snapshots were superimposed. For example, if the trajectories are saved every 10 ps, for 1 ns time scale, the first DYN value for each residue is the RMSD between its structure at the 100th snapshot and the initial snapshot, then the RMSD between the 101st snapshot and second snapshot and so on.

(9) PC descriptors are projections of the principal components on the trajectory snapshots. The principal components were calculated for each residue independently using a combined trajectory that included the ligand-bound and ligand-free trajectories. For each residue, the heavy atoms of the residue in all MD snapshots were fit to the first snapshot, and an average structure was generated. Then, the residue atoms in all the snapshots were refit to this average structure, and the principal components were calculated using the script presented in the Supporting Information. The first three components were considered to generate the PC descriptors (PC-1, PC-2, and PC-3) because these components were found sufficient to describe 90% of the residue motions.

Table 1 summarizes all descriptors and their abbreviations as used throughout the article. The data set containing the values of each descriptor was saved as a CSV file for all 338,538 snapshots. Sample scripts that were used for the calculation of descriptors are provided in the Supporting Information.

Deep Neural Networks. Deep neural networks are neural networks with more than a single hidden layer. We used these models to classify the protein conformations in each MD snapshot based on its ligand binding status. To train and test the neural networks, a combined set of the MD snapshots from the simulations with and without ligands was used. The snapshots were characterized by one of the above-mentioned descriptors of the 110 residues of the protein. The purpose of the classification in our work was not the prediction itself but to evaluate the importance of each feature in prediction accuracy. As a further measure of descriptor qualities, smaller subsets of the data sets were used to train the neural network, assuming that a given descriptor can be considered more efficient if it results in high classification accuracies when trained with a lower number of training samples. These subsets of the descriptor data sets were provided with systematic sieving of the data sets with different intervals of 10 ps (338,538 records), 100 ps (33,853 records), 1 ns (3385 records), and 10 ns (338 records), respectively.

Table 1. Descriptors of Per-Residue Properties (attributes) Calculated from MD Trajectories and Used to Train Machine Learning Models

Descriptor abbreviation	Description
C_α	Distance between C_α of the residue and the geometric center of the protein
GEOM	Distance between the geometric center of the residue and the geometric center of the protein
FLUCT	Root mean square deviation (RMSD) between the residue and its average structure in the trajectory
VDW	van der Waals interaction between the residue and the rest of the protein
ES	Electrostatic interaction between the residue and the rest of the protein
NONB	Total nonbonding (VDW+ES) interactions between the residue and the rest of the protein
SURF	Contribution of the residue to the protein surface area
DYN-10 ps	Dynamics of the residue in 10 ps time scale
DYN-1 ns	Dynamics of the residue in 1 ns time scale
DYN-25 ns	Dynamics of the residue in 25 ns time scale
DYN-100 ns	Dynamics of the residue in 100 ns time scale
PC-1	Projection of the first principal component
PC-2	Projection of the second principal component
PC-3	Projection of the third principal component

Feedforward with backpropagation neural networks were trained and tested using tensorflow 1.1[49] in 10 iterations of 5-fold cross-validations in which the records were shuffled in each iteration, and 80% were used for training, while 20% were used for testing in each cross-validation iteration. The models were trained using the Adam Optimizer algorithm[50] with a learning rate of 0.001 and a softmax classifier using cross entropy loss[51] as the cost function. The performance of the neural networks was measured with classification accuracy, which was defined as the rate of correctly classified snapshots in the test set. The accuracies of 5-fold cross-validation steps were averaged and used as the accuracy for each iteration, and then, the averaged accuracies of a total 10 iterations were reported and used to measure performance.

The architecture of the neural networks includes an input layer containing 110 cells, three hidden layers, and an output layer of two cells. The number of the cells in the hidden layers of the networks was optimized with several trial-and-error attempts targeting the highest consistent accuracies among many runs. The optimal architectures were found to be 110-$(500)_3$-2, 110-$(500)_3$-2, 110-$(200)_3$-2, and 110-$(50)_3$-2 for models fed with 338,538, 33,853, 3385, and 338 records, respectively. The same neural network models were trained and tested without considering the 12 pocket residues, which were defined as the ones that have less than 3 Å distance from the ligand in the bound crystal structure. These included residues 323, 324, 325, 326, 327, 328, 339, 372, 373, 376, 379, and 380. Also for each descriptor, two networks were trained with only the top 10 residues selected by the random forest models and a set of randomly chosen nonpocket residues. The optimal neural network architecture was found to be 10-$(40)_3$-2 for these models.

Random Forest Models. Random forest models can be used for classification and feature selection purposes. We used this method here for both purposes, i.e., (1) to validate the relative classification accuracies obtained with neural networks

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

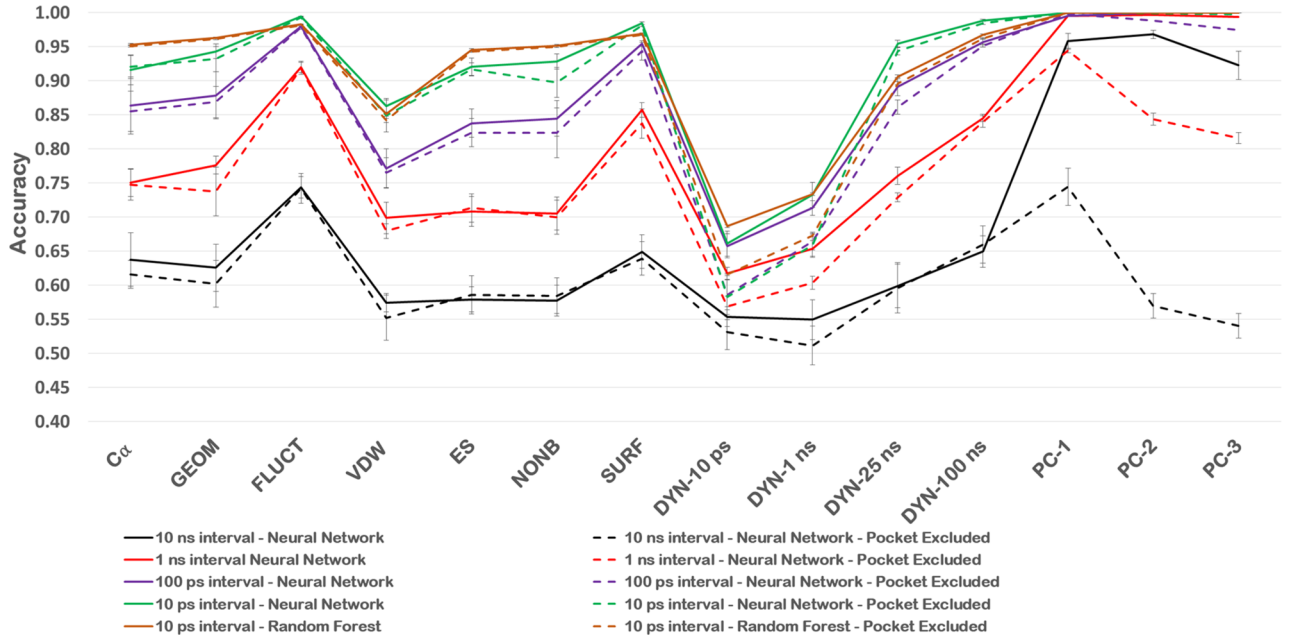


Figure 2. Accuracies of the neural network and random forest models for each descriptor. The models were designed to predict the classification of bound and unbound states using different residue descriptors. The accuracies of neural network models fed by data sets with snapshots of 10 ns, 1 ns, 100 ps, and 10 ps intervals are shown in black, red, purple, and green, respectively. The accuracies of the random forest models are shown in brown. The models trained with and without ligand-binding pocket residues are shown in solid and dashed lines, respectively. Error bars represent the standard deviation of 10 runs of neural networks and 100 runs of random forest models, respectively.

that were trained with 338,538 snapshots with and without the pocket residues and (2) to rank the residues according to their contribution to the classification ability.

To perform random forest modeling, we used the random forest classifier from the Scikit-Learn machine learning library.[52] For each data set, 75% of the 338,538 records with and without considering the 12 pocket residues were randomly sampled to train the model, and the remaining 25% records were used for test purposes. The randomized training and testing procedure was repeated 100 times, and averaged classification accuracies were reported. The accuracy was calculated as the number of correctly classified cases divided by the total number of test cases. The importance of each residue with regard to the prediction ability of each model was recorded, and the overall prediction ability of each residue was ranked based on the averaged importance level over all 100 models.

Elastic Network Model. Elastic network modeling (ENM) was performed using the DynOmics ENM 1.0 online tool.[53] The crystal structure of ligand-bound PDZ3 (PDB code: 1BE9) without the crystallographic water residues was used as the input structure. The bound peptide residues were used as the environment, and the protein residues were used as the system.

RESULTS

Multiple Attributes of Allosteric Residues Are Affected upon Ligand Binding. Multiple properties can be calculated for a protein from the molecular dynamics snapshots at the domain, residue, subresidue, and atomic levels. These properties describe the system in different ways and could be considered as different layers of information that potentially result in a comprehensive picture of the protein when being combined. From a protein allostery viewpoint, it is interesting to investigate the fluctuations of such properties at the residue level. We hypothesize that each one of such layers of information can differentiate the snapshots taken from the ligand-bound and ligand-free MD trajectories with different efficiencies. In other words, multiple residue attributes are affected upon ligand binding to a different extent. To test the hypothesis, we described the residues of the PDZ3 protein in 2 μs of ligand-bound and 2 μs of ligand-free MD simulations in terms of their positions, fluctuation, nonbonding interactions, surface area, and dynamics. Table 1 summarizes different residue descriptors that were calculated from MD trajectories and used to train and test predictive machine learning models. More details are provided in the Methods section.

We trained and optimized deep learning neural network models using the descriptors generated and calculated from MD trajectories based on a 5-fold cross-validation procedure that randomly separated the data set into 80% for training and 20% for validation in each run. The snapshots, taken every 10 ps in MD trajectories, were used to train models to distinguish the bound and unbound trajectories. To identify the residue attributes that have more efficient classification capabilities, we reduced the snapshots in the data sets by picking one snapshot in every 100 ps and 1 and 10 ns of MD trajectories as different snapshot offsets (Figure 2). We further trained random forest models using the data sets containing maximum data points (saved with 10 ps intervals). The random forest models resulted in similar prediction accuracies as the neural networks (Figure 2, brown solid line). The reported neural network accuracies were averaged over 10 average accuracies of cross-validation attempts. The random forest accuracies were averages of 100 runs of independent random forest models. Among all the residue descriptors, the projection of the principal components (PC) of the residues resulted in almost perfect prediction accuracy, followed by the residue fluctuations (FLUCT), surface area contributions (SURF), and position descriptors (GEOM and C_α). Comparing the

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

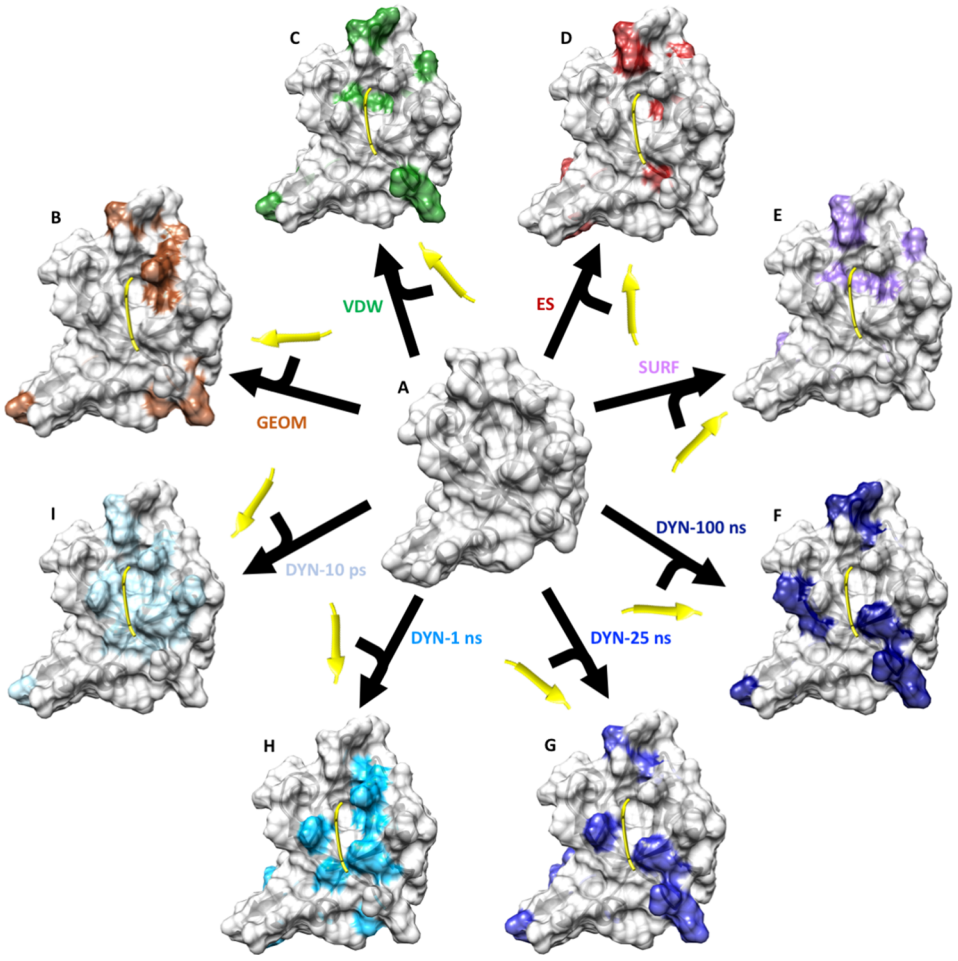


Figure 3. Schematic representations of how various residues are affected in different ways upon ligand binding of the PDZ3 domain. The crystal structures of the unbound state (A) and bound state (B–I) are shown. The binding ligand is shown in yellow. The top 15 residues selected by the random forest models, which were trained with different descriptors, are highlighted in different colors in each of the bound state (B–I). The descriptors used to identify these residues were marked on the arrow lines using the corresponding color.

efficiency of models trained with GEOM and C_α implies that incorporating the information about the side chains (GEOM) slightly improves the prediction accuracy. The nonbonding interactions generally resulted in poor predictions compared to other descriptors. However, the electrostatic interactions distinguished the binding states better than VDW interactions, implying that the electrostatic interactions are more affected upon ligand binding. The results showed that the efficiency of the dynamic properties depends heavily on the time scale of the dynamics being considered. Slow time scale dynamics resulted in much better prediction performance than fast time scale dynamics. It is worth noting that our purpose of using neural networks is not the prediction itself but to evaluate the importance of each feature in prediction. Therefore, even bad predictions have a valuable message. For example, if using a lower sample size the accuracy of a certain prediction drops drastically, it may suggest that the feature used for this prediction is not very informative to prediction binding. In other words, this feature may not be affected much upon binding.

To exclude the possibility that the classification capabilities mostly rely on information from the residues close to the binding site, rather than allosteric residues in farther distances, we trained the same neural networks and random forest models without the data from the pocket residues (Figure 2, dashed lines). The pocket residues were defined as residues that are located within a 3 Å distance from the bound peptide in the crystal structure (PDB code: 1BE9). Such definition excludes 12 residues (323, 324, 325, 326, 327, 328, 339, 372, 373, 376, 379, 380) from consideration for training the models. The accuracies of the predictive models of most descriptors did not drop dramatically when excluding the pocket residues (Figure 2), suggesting that the ability of the neural networks to distinguish the bound and unbound states does not mainly originate from the residues near the binding site. Rather, it relies on the allosteric residues. However, a higher pocket residue reliance on model accuracy can be observed when using the fast time scale dynamics and the projection of the principal components to train neural networks (Figure 2). The accuracies of all models that were trained with 10 ps and 1 ns dynamics data deteriorate significantly when the pocket residues are ignored. This is reasonable as the bound ligand is expected to change the fast fluctuations of binding pocket residues.

To investigate whether shorter simulations provide comparable results with the long simulations, we used all snapshots (saved every 10 ps) of the first 10 ns of bound and unbound trajectories to train the neural networks. Using these data sets resulted in prediction accuracies that are even better than using long trajectories (Figure S1), probably because the less diverse

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691–4705

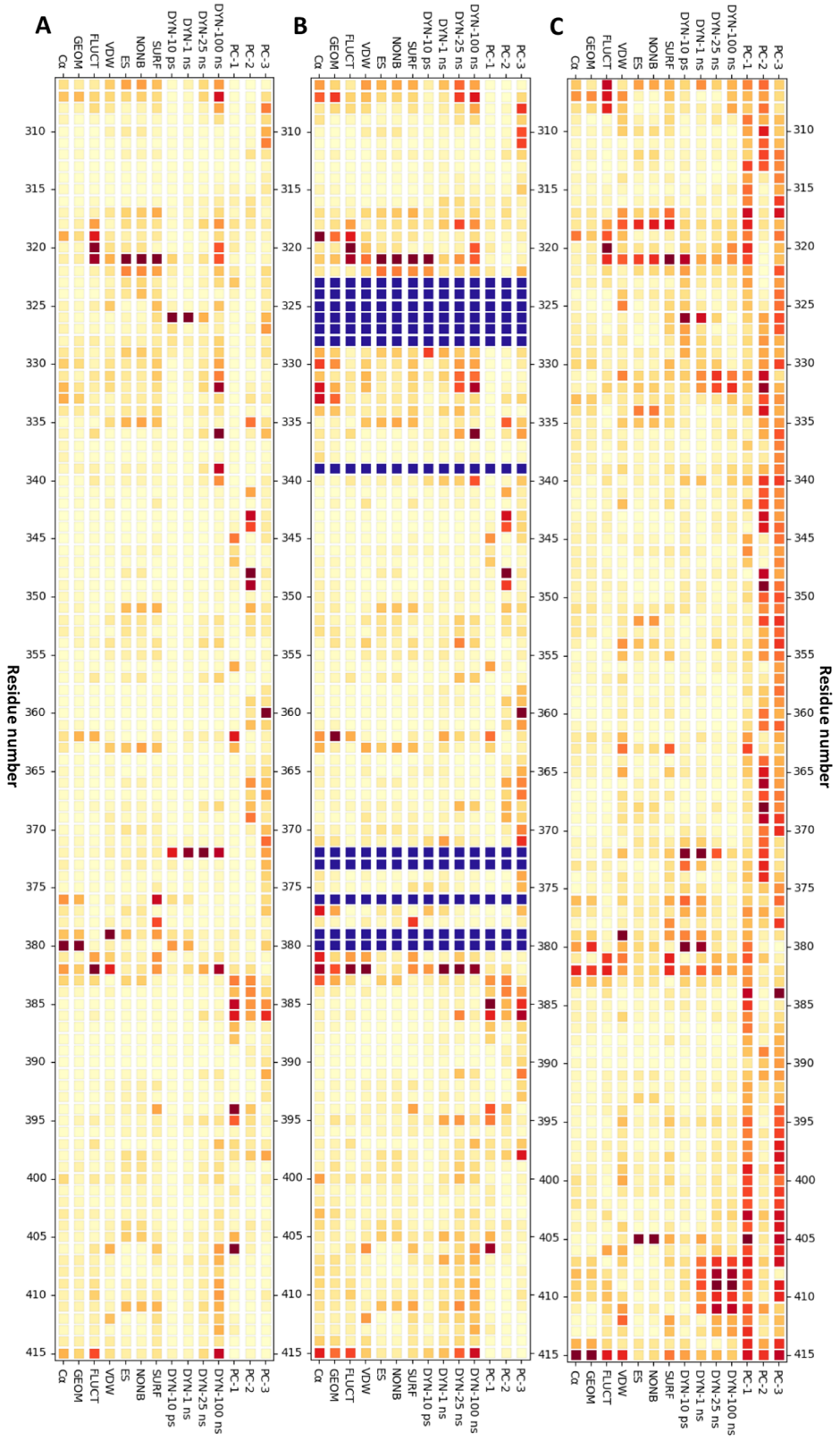


Figure 4. Residue response maps. This map illustrates the per-residue response to the ligand binding. The *x*-axis is the descriptors, while the *y*-axis is the residue number. The color intensity shows the extent of residue response upon ligand binding. The darker red color represents a stronger response upon binding. (A) Residue response maps calculated from random forest models with ligand-binding pocket residues. (B) Residue response maps calculated from random forest models without ligand-binding pocket residues. The excluded binding pocket residues are marked in solid blue. (C) Residue response maps calculated directly from MD trajectories. The color intensity in (C) depicts the absolute differences of the average property values between bound and unbound states.

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

Table 2. Average Uncertainties of Per-Residue Contribution Values to Random Forest Models[a]

Descriptor	Average uncertainty with pocket residues (%)	Average uncertainty without pocket residues (%)	Average uncertainty with pocket residues with short simulations (%)	Average uncertainty without pocket residues with short simulations (%)
C_α	21.06	19.36	93.11	81.5
GEOM	20.69	18.62	100.35	80.89
FLUCT	20.87	19.85	289.68	261.08
VDW	6.52	5.82	68.4	60.88
ES	10.96	9.85	110.98	97.22
NONB	11.75	10.68	118.8	106.47
SURF	15.58	14.78	158.32	148.09
DYN-10 ps	3.39	2.54	34.45	29.96
DYN-1 ns	3.96	3.03	54.22	48.69
DYN-25 ns	7.61	6.47	not applicable	not applicable
DYN-100 ns	13.54	12.44	not applicable	not applicable
PC-1	176.68	169.23	189.27	160.75
PC-2	210.06	202.74	418.32	403.92
PC-3	230.73	217.54	316.75	288.96

[a]The average uncertainties are averages of variation coefficients (standard deviation divided by the average over 100 runs) calculated for each random forest model over 110 residues. Short simulations include only the first 10 ns chunks of one bound and one unbound trajectory.

sample size of the shorter simulation increases the accuracies of the prediction. But the longer simulations provide more information about the residue responses, as revealed in later sections.

Allosteric Residues Are Affected in Different Ways upon Effector Binding. Considering the interdependence between properties of protein residues and ligand binding events, residues may change the protein affinity to a ligand and be affected by the binding ligand at the same time. The residues with a better ability to "sense" ligand binding have been considered as potential allosteric residues.[12] Various experimental and computational approaches have been used to identify such potential allosteric residues in model proteins such as PDZ3.[14−20] However, it is still unclear whether different allosteric residues sense the ligand binding in the same way (i.e., being affected through the changes in the same residue properties) or they respond to perturbations through changes of different residue properties. On the other hand, from a more practical viewpoint, it would be beneficial to know whether the changes in the same or different properties of these residues result in allosteric effects.

To answer these questions using the PDZ3 domain as a model protein, we used the random forest models, which is a well-established method with the inherent capability to quantitatively rank the contributions of features for prediction purpose. Here, we ranked the contribution of each residue to distinguish the bound and unbound snapshots of the MD trajectories. Figure 3 schematically represents the top 15 residues contributing most to the random forest models using different descriptors. The results show that the affected residues have different contributions to model accuracies when different residue properties are used to train the models, which essentially indicates that not all protein residues are affected in the same way.

Residue Response Maps. To quantitatively represent the per-residue contribution values to the random forest models, we presented the "residue response maps" showing the per-residue contribution values as color intensities (Figure 4). The residue response map quantitatively illustrates the changes of multiple properties of each residue upon the ligand binding, providing a novel and effective way to present the per-residue response to the perturbation. We hypothesize that the binding pocket residues undergo higher changes of their properties, hence contribute more to the random forest accuracies and mask the importance of other residues in allosteric positions in the response map. Thus, we prepared the maps from weights of random forest models (raw data presented in Tables S1 and S2) that were trained both with and without the data from the binding pocket residues (Figure 4A vs B, respectively). It is clear that excluding the pocket residues highlights the importance of some residues (darker red color) that were not important (lighter color) when pocket residues were considered.

To better interpret and evaluate the reliability of the maps, we calculated the uncertainty for each feature (Table 2). These uncertainties are the average of variation coefficients (standard deviations divided by averages) of residue importance values in 100 random forest runs. The lower average uncertainty indicates a more reliable residue descriptor because the selected set of residues with that descriptor is more consistent and reproducible during multiple runs. Among different descriptors, the lowest uncertainties belong to the short time scale dynamics, followed by the nonbonding interactions, making the residues suggested by these descriptors the most reliable ones. These residue properties are more converged in the simulation time scale than other residue properties, therefore suggesting more reproducible sets of important residues. In contrast, residues selected with the projections of the principal components suffer from dramatic uncertainty, making us hesitate to rely on the residue sets suggested by modes trained with these descriptors.

To investigate whether the same residues could be selected with shorter simulations, we prepared the residue response maps with the random forest models that are trained with 10 ns chunks of one bound and one unbound trajectory (Figure S2). Although these models result in almost perfect prediction accuracies, the uncertainties in the selection of the residues were much higher for these models comparing to models trained with longer simulations, and different residues were selected (Table 2). Therefore, we conclude that the residue sets suggested by the models trained with longer simulations are more reliable. The high binding prediction accuracies of the shorter simulations might be due to the overfitting of the models to limited sample size. Longer simulations sample the

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

conformational spaces more realistically and provide random forest models with a more diverse set of sample cases, which might make the predictions more complicated, but the selected residues more reliable.

Among different residue features, the projections of the principal components resulted in the selection of residues that have a poor agreement with both the list of experimentally known important residues and residues selected with other models. There are significant uncertainties for PC models in ranking the residues regarding their importance in ligand binding (Table 2), although these models result in almost perfect binding prediction. In other words, models that were trained with residue principal components data can predict the binding with any set of residues. This can also be inferred from the very busy PC columns on the maps that were calculated directly from the trajectories (Figure 4C).

To validate the random forest results, we prepared the residue response maps directly from the classical analyses of MD trajectories. To do this, the absolute difference of the property values between bound and unbound trajectories were mapped again with and without considering the pocket residues (Figure 4C). Comparing the maps from classical analyses of MD trajectories (Figure 4C) and from random forest model including the binding pocket residues (Figure 4A), it is clearly demonstrated that (1) most of the random forest models selected top residues (Figure 4A with darker red color) are also identified by MD trajectories (Figure 4C with darker red color), validating the random forest results and (2) random forest models identified much fewer top residues than MD analyses. Given that random forest models only consider top residues contributing to distinguish the bound and unbound state, the fewer identified top residues of the random forest model might result from removing noises from the MD analyses by excluding the property changes not directly related to binding. The random forest model by its feature selection nature selects the most important residue responses upon ligand binding, rather than highlighting all differences between the bound and unbound trajectories. Therefore, one of the advantages of the random forest approach is that it highlights the residue−feature pairs that are related to binding among many pairs that could be identified with the classical MD analyses approach. This advantage of random forest models stands out in PC descriptors (three columns on far right of the maps), where a small set of pairs are recognized as important by random forest models (Figure 4A) among many that were observed to be different by MD analyses (Figure 4C), demonstrating the capability of random forest models to exclude false-positive residues identified by MD analyses for certain properties. Another advantage of the random forest model is that it could exclude the binding pocket residues and highlight relevant allosteric residues. For example, residues Val386 and Glu395 were not picked by MD analyses (Figure 4C) or random forest models including the binding pocket residues (Figure 4A). However, these two residues stand out in random forest models excluding the binding pocket residues with the DYN-25 ns descriptor (Figure 4B). The mutation of Val386 or GLu395 has been reported to affect ligand binding experimentally,[54,55] demonstrating the capability of the random forest model to identify false-negative residues from MD analyses.

Experimentally Known Important Residues Glow in Residue Response Maps. Interestingly, most of the PDZ3 residues that were highlighted in the residue response maps have been reported previously as allosteric residues. Reading the maps (Figure 4) from the N-terminal residues, the initial three residues are located in a very flexible unstructured region, and the next region (residues 311−316) is a silence region, in which the residues are not highlighted in the maps. The following region includes a flexible loop (residues 317−322). The positions, nonbonding interactions, middle to long time scale dynamics, and some modes of motions of residues in this loop are highlighted in the maps.

The next region in the sequence is a flexible loop consisting of residues 329−335. Several residues in this loop are highlighted. These contributions are more significant when the pocket residues are ignored from random forest calculations (Figure 4B). Gly329 has high importance in models using total nonbonding interaction (NONB) and short time scale dynamics (DYN-10 ps and DYN-1 ns) as predictive descriptors. This residue has been reported to have the largest mutational effect[17] and relatively high allosteric response ratio as calculated by Grerek and Ozkan.[16] Our finding implies that the Gly329 probably experiences different dynamics in bound and unbound states. As this residue is relatively close to the N-terminal of the binding peptide, it is reasonable to expect that its short-range nonbonding interactions and low time scale vibrations are affected by ligand binding.

The negatively charged residues Glu331 and Asp332 are recognized by models utilizing van der Waals interactions (VDW) or the longer time scale dynamics (DYN-25 ns and DYN-100 ns) as the descriptor. It has been reported that succinimide cyclation of the Asp332 side chain alters peptide binding in PDZ3.[19] Although this residue does not interact with the binding ligand directly, it was proposed that this residue affects the local conformation of the loop; thus, the electrostatic interaction between the neighboring Glu331 and the ligand is disrupted.[19] Our results suggested that Asp332 senses the ligand binding as a change in its short-range nonbonding interactions and long time scale dynamics. It is reasonably expected that the change in such properties of this residue could affect binding as well. It has also been reported that mutation of Asp332 to proline, which has different flexibility and nonbonding interaction properties, affects ligand binding in PDZ3.[19]

Gly335 has also been shown to be sensitive to its deletion in the ^{15}N relaxation experiments.[14] Our results suggested that Gly335 is affected by the ligand through nonbonding interactions (according to VDW, ES, and NONB models). This fluctuation in nonbonding interactions might be related to the changes in the short time scale dynamics of Glu395 (a residue located in an α3 helix close to Gly335 and highlighted in the map), as it is expected that the fluctuations of the negatively charged side chain of Glu395 affect its surrounding environment through nonbonding interactions.

Residues 336−341 form the βG sheet. The long time scale motions of Ile336 show a significant contribution in distinguishing the bound state according to the map in Figure 4A and B. It has been reported that this residue is sensitive to mutation.[17] Here, our results showed that the dynamics of Phe340 at different time scales, especially at 100 ns, are affected by ligand binding. Phe340 is one of the residues that gave the highest fluctuation response upon perturbation by the perturbation scanning response (PRS) analysis.[16] Being highlighted in PRS experiments means that the random forces put on Phe340 cause a response in all protein residues. Our

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

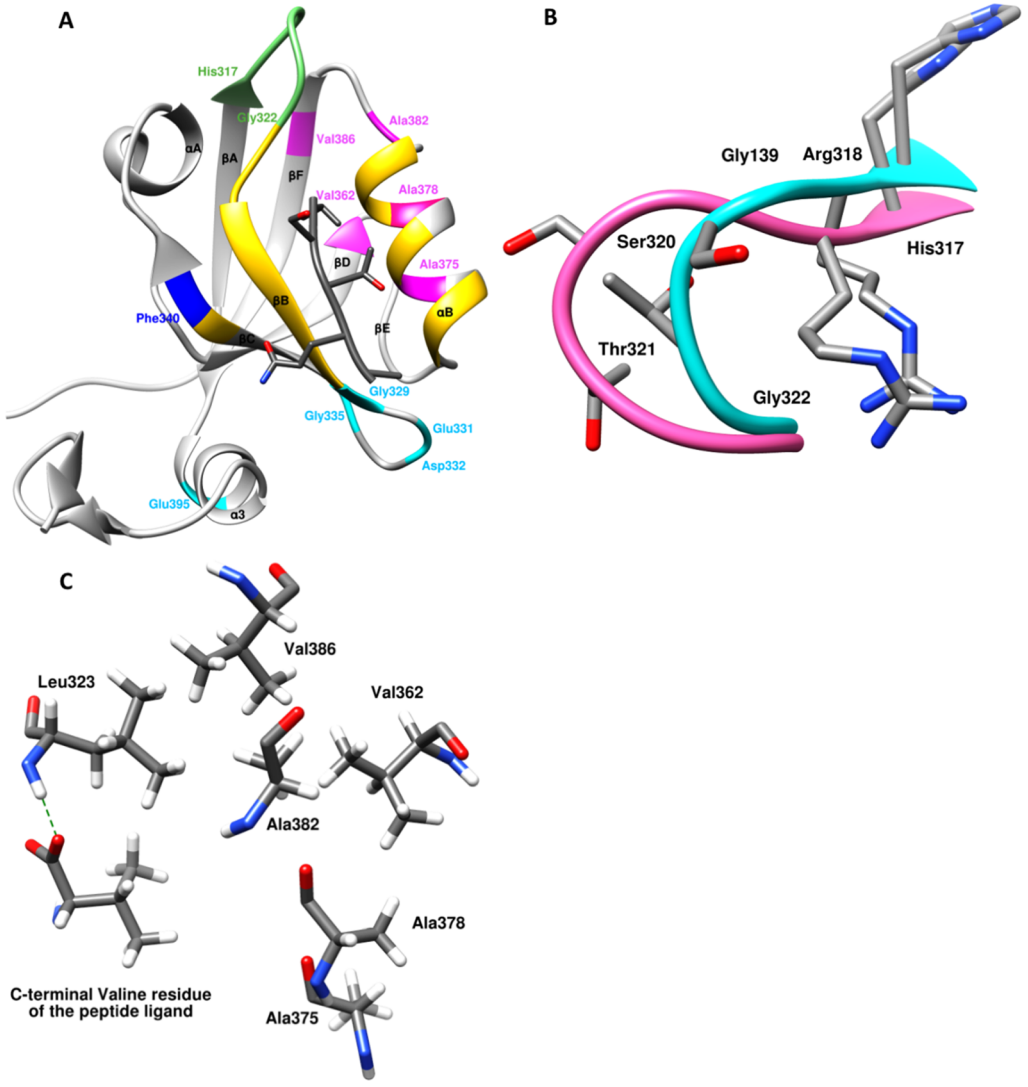


Figure 5. Structural information on the top selected residues by the random forest model. (A) Ribbon representation of the crystal structure of the PDZ3 protein (PDB code: 1BE9) highlighting some important residues recognized by random forest models. Residues in gold are considered as binding pocket residues and were not fed into the random forest models. (B) Residues 317−322 in the ligand-bound crystal structure (cyan) and unbound crystal structure (pink) of PDZ3, as the backbones of complete structures, are superimposed. Hydrogen atoms are not shown. (C) Localization of a network of hydrophobic residues of PDZ3 and the C-terminal of the peptide ligand. The green dashed line represents the key salt bridge between the peptide ligand and the backbone of Leu323.

results suggested that limiting the dynamics of this residue might also have an allosteric effect on ligand binding.

Moving down the sequence on the residue response map, Val362 is another highlighted residue that is part of the βD sheet. Val362 has been shown to have a relatively high response in PRS analysis as one of the highly weighted residues in the allosteric pathway of PDZ3.[16] Our results showed this residue has a high impact on distinguishing the bound and unbound protein conformations using fluctuations (FLUCT) or dynamics at 1 ns time scale (DYN-1 ns) as the descriptor. Also, its position is important in the prediction of the binding state, especially when its side chain is taken into account (GEOM), implying that its side chain is probably more affected upon ligand binding than its backbone. Another residue in this region is Val386 for which the dynamics at 25 ns time scale (Dyn-25 ns) are important according to our results. There are also other hydrophobic residues that directly or indirectly interact with valine residues (Figure 5C), which are highlighted in the residue response maps. These residues seem to make a network of hydrophobic interactions, and their different time scales are affected upon binding (Figure 5C).

The other region in the structure that has been reported as allosterically important in binding[14] is the α3 helix that includes residues 394−399. Our results showed that the dynamics of Glu395 at the 1 ns time scale is affected by binding. The residue close to Glu395 in space is Gly335. These residues interact with each other through nonbonding interactions. As noted before, Gly335 has been shown to be sensitive to its deletion in the ^{15}N relaxation experiments,[14] and its nonbonding descriptors are highlighted in the residue response map. This fluctuation in nonbonding interactions might be related to the changes in the short time scale dynamics of Glu395, as the fluctuations of the negatively charged side chain of Glu395 may affect its surrounding environment through nonbonding interactions. In the crystal structure,[21] Gly335 is close to Gly329 in space, which is in turn close to the ligand as mentioned above. This implies that the ligand affects short time scale vibrations and nonbonding

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

interactions of Gly329, which affects Gly335 and the α3 residues sequentially through nonbonding interactions. This nonbonding effect on the α3 helix has not been recognized in our models, although some residues in the helix show response to binding as perturbations in their longer time scale dynamics.

An interesting qualitative comparison of the predictions suggested by the residue response maps come from comparing these maps with the similarly presented high-throughput mutation sensitivity maps by McLaughlin et al.[17] Of course, a one-by-one agreement should not be expected between two maps at residue levels as they convey two messages that are different although related. McLaughlin et al. have mapped the cost of mutation of each residue on the protein to other amino acids (Figure 2b in McLaughlin et al.[17]). Two major sensitive regions that are highlighted in their work, are highlighted in our maps as well.

Validation of Residue Response Maps. To further validate our random forest results, we used the top 10 residues recognized by random forest models to retrain neural networks and compared them with the networks that used all nonpocket residues. As a control, we also trained the same networks with 10 randomly selected residues that did not contain any of the top 10 residues nor any of the residues in the binding site. The results show that the top 10 residues in each model predict the binding state significantly better than the 10 randomly selected residues, validating the significance of the top 10 residues in their contribution to the model accuracy. However, for many attributes, using the whole data set results in significantly better predictions than using only the top 10 residues (Figure S3), suggesting that more than 10 residues contribute significantly to the model accuracy. In other words, the top 20 or 30 residues may also be important for the model prediction capability.

Principal components, on the other hand, can result in the perfect prediction of binding status by using all residues as well as the top 10 residues and even randomly selected residues. Besides the high average uncertainties associated with these descriptors (Table 2) and very busy residue response maps calculated directly from the trajectories (Figure 4C), this insensitivity of the neural network accuracies to residues suggested that principal components may be a good descriptor to distinguish bound and unbound states, but the contribution differences among residues may not be significant.

We performed a quantitative and direct comparison between the random forest feature weights and the experimental results to check whether these weights add more insight into the MD simulations compared to the traditional MD analysis methods. To do this, we calculated the 10 ps and 1 ns dynamics values of methyl groups whose NMR order parameters have been reported by Lee.[56] We trained two random forest models that use these two data sets to determine the ligand binding states of MD snapshots. The random forest models could predict the binding states with accuracies of 58% and 61% with 10 ps and 1 ns dynamics data, respectively, suggesting that our machine learning model could reach reasonable accuracies based on the dynamics of key methyl groups. The relative low accuracies of these two models are expected, as the dynamics of methyl groups alone should not represent all the protein dynamics changes upon ligand binding. The average uncertainties for the suggested feature weights are 1.23% and 1.63% for prediction models using 10 ps and 1 ns dynamics data, respectively, suggesting the high fidelity of our machine learning model. These suggested weights from both models (Table S3) correlated with the experimental order parameters with a 0.40 coefficient. As a benchmark for the machine learning models, we calculated the difference between the root-mean-square fluctuations (RMSF) of the methyl groups in bound and unbound trajectories as a traditional method for representing the experimental order parameters (Table S3). The calculated RMSF values correlated with the experimental order parameters with a much lower coefficient of 0.12 compared to the random forest feature weights as 0.40. This benchmark demonstrates that the random forest feature weights, which are represented in this work as residue response maps, can improve our insight into the MD trajectories.

We further performed elastic network modeling (ENM) to verify whether the residues that are predicted to have slow dynamics based on the slow modes qualitatively agree with the residues that are highlighted in our DYN-100 ns residue response map. The average 10 slow modes from ENM are presented in the same way as random forest weights are presented in residue response maps (Figure S2C). A qualitative comparison between the two sets shows that residues 306−309, 320, 321, 331, and 332 and most residues in the 406−415 region are highlighted in both maps. However, some experimentally significant residues that are highlighted by random forest model are not identified by ENM. There are also some residues highlighted in ENM but missing in DYN-100 ns analysis. This is reasonable as the slow modes in ENM do not represent exact time scales and can point to time scales longer than 100 ns. For example, it is reasonable to consider the whole C-terminal region of the protein between residues 395 and 415 fluctuating at low frequency (probably in microsecond time scales) as the ENM modes suggest.

DISCUSSION AND CONCLUSION

It is well known and well described that protein residues located far from a binding site are affected upon ligand binding, and changes in the conformation of these residues upon binding to allosteric ligands can potentially regulate binding affinity of the protein with its major ligands. To gain more insight into these distal allosteric residues, it is important to know whether allosteric residues communicate with the binding site in similar or different ways. For any given allosteric protein, if all allosteric residues communicate with the binding site through the same main interactions (residue attributes in this study), such as the electrostatic interactions,[57] one just needs to focus on this main interaction and the location of allosteric residues when designing allosteric ligand. On the other hand, if all allosteric residues communicate with the binding site through different interactions, one should expect more complex behaviors in allosteric sites and will need to take more rational and systematic strategies in designing allosteric ligands because different properties of different allosteric sites might be needed to induce desired changes to regulate the protein binding to its ligand. Upon ligand binding, the internal energies of the protein redistribute[58] to respond to this binding perturbation, and this energetic redistribution may affect multiple properties for each residue. Therefore, we aimed to use hybrid models that benefit from the advantages of both MD and machine learning methods to reveal how different properties of each residue respond to ligand binding in the PDZ3 domain of PSD-95.

The neural networks trained with different properties of MD snapshots of ligand-bound and unbound states trajectories show that different properties have different potencies for

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

distinguishing the bound and unbound states snapshots. Among tested properties, position, fluctuations and long time scale dynamics of residues are more efficient in distinguishing the binding status in PDZ3. Having various potencies in predicting the binding state of PDZ3 shows that various residue properties are affected to different extents upon ligand binding. Therefore, we can describe different properties of residues as different layers of information about the protein when it is bound to a ligand. Each of these layers has a unique contribution to predicting the binding status of the protein. The information hidden in position, surface area, and nonbonding interactions layers might be inherited from the subtle differences in the initial structures because the simulations were initiated from the bound and unbound crystal structures independently. But the dynamics and fluctuation layers of information extracted from the molecular dynamics simulations are more representative of the differences induced by ligand binding during the simulations.

Neural networks and other modern deep learning models are very powerful predictive models that could be used for recognizing patterns and predicting states. However, the connections between their nodes and layers are mostly unclear. Specifically, the internal weights in these deep learning models cannot be easily used to select top features contributing to the prediction. Conversely the random forest model is a well-known feature selection model, using weighting methods for their predictions that can directly rank the features based on their importance. Therefore, we used this advantage of the random forest model to rank the contributing residues and used deep neural networks to validate our predictions. Using the random forest predictive models, we represented residue-specific allostery using residue response maps. Based on these maps, we demonstrate that all residues can be considered allosteric, and each of them "senses" the bound ligand in a unique way. Some residues are affected in many different ways, while others are affected in a certain way. Some residues change in their nonbonding interactions; some change in their dynamics properties. More residues are affected in a combination of these properties. Also, some residues are affected significantly and are likely the residues being identified as allosteric residues in experimental studies, while other residues are affected only slightly. This can be considered as analogous to social stress that results in different behavioral and emotional reactions in individuals living in society. As occurs in social stress, some reactions to the stress of ligand binding in a protein are more common and descriptive of the event, while some are less common. Among the common responses to the binding event in PDZ3 are the perturbation of residues' fluctuations and long time scale dynamics, and among the less common responses is the change in short time scale dynamics, which is more likely to be experienced by residues close to the binding site.

Although using different residue features gives us a more comprehensive perspective of the binding event, it should be noted that each of these different features has its pros and cons, thus should be considered with a weight. Some of these pros and cons can be inferred from the feature definitions. For example, positional features (C_α and GEOM) might have noises that are results of the translational degrees of freedom of the whole protein. In other words, the geometric center of the protein, which is the reference point for calculating these properties, may move upon binding, resulting in the noise in the values or washing out some important positional data. Also, the consistency of feature selection by the random forest models and certainty of feature weights can help us as an axillary tool to interpret the residue response maps. We reported these uncertainties as the mean variation coefficient of the importance values suggested by the random forest models in Table 2. The models with the least uncertainty are the most reliable models, and the residues selected by them should be considered the most meaningful. These include the non-bonding interactions and dynamics. On the other hand, the principal component models have the maximum uncertainties, and the residues suggested by them have the least agreement with the experiments. Initially, one would expect that the DYN and PC would convey similar meanings about the dynamics of the residues. However, our analyses here show that for training the machine learning models, PC descriptors result in high uncertainty. Considering such uncertainty issue, and also the easier physical interpretation of DYN descriptors, we believe that DYN descriptors are better descriptors for generating residue response maps.

The remaining question would be why are different responses observed among protein residues upon ligand binding to the protein? A hypothetical answer would interpret unique residue responses based on the unique position in the sequence, space, and chemical properties of the residue. Residues closer to the binding site are more likely to be affected by short-range nonbonding interactions directly from the ligand. The cascade of perturbation in such short-range interactions might fade with increasing distance from the binding site. But polar and charged residues at far distances from the binding site can be directly affected by the ligand through the electrostatic forces initiated by polar and charged moieties of the ligand. This would result in consequent perturbation in the position and dynamics of these ligands, which can start their own cascade of perturbations in their neighboring residues in time and space. If future experiments support this view about the allosteric effects in proteins, one might expect that in addition to their sequence and conformation proteins can potentially be described with "residue response maps" such as the one seen in Figure 4A, which represent the extent of specific changes in residue attributes. Supposing that every ligand is unique in its chemical properties, this description of proteins would also be ligand specific, and residue response maps can also be considered as fingerprints that uniquely describe a protein when it interacts with a specific ligand. We expect that residue response maps could give researchers useful guidance when they design drugs that target allosteric binding sites. These maps may provide useful information about the properties of allosteric binding sites. Furthermore, there can be some mechanistic speculations from the residue response maps. The top residues in C_α and GEOM models might be considered as residues involved in allosteric mechanisms based on the structural changes, while top residues in DYN models might be considered as residues involved in dynamics-driven allosteric mechanisms (Table S4). However, more evidence for these mechanistic speculations should be obtained from other experimental and computational approaches that focus on studying allosteric mechanisms.

In this work, we chose a well-studied allosteric ligand–protein system to test our hypothesis and evaluate the feasibility of our novel approach in analyzing protein residue characteristics related to an allosteric perturbation, which is the ligand binding in this case. There are some limitations in our

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

approach. We only tested one model protein in this study, which raises the question of whether this approach could be applied to other or larger proteins. We are in the process of testing larger proteins such as GPCR proteins and the CRISPR/Cas9 complex using this approach and will report our findings in separate publications.

Another limitation is the descriptor selection. In this work, we tried to choose a set of residue-specific parameters that could simply and efficiently describe the structure and dynamics of each residue per simulation frame. The more sophisticated or descriptive parameters can be used in future works to include residue−residue interactions and generate more comprehensive residue response maps. Here, we only tested very limited descriptors and demonstrated which descriptors could be more reliable to predict allosteric residues for one model protein. Whether these descriptors could be applied to other proteins remains unclear. Our exploration of descriptors is far from comprehensive, and it would be interesting for future studies to test more descriptors in other protein systems. In this work, we mainly focused on the response of individual residues. It is worthwhile for future studies to generate more comprehensive residue response maps to include the response of functionally important regions.

A third limitation lies in the machine learning algorithms that we used. The random forest model has been known to be a good algorithm for feature selections. However, testing other machine learning algorithms, such as information gain and decision tree, may provide more insights into the performance of machine learning algorithms on allostery studies. Another potential limitation of this method is that the reliability of machine learning outcomes depends on the convergence level of the MD simulations. This is similar to any other judgments done based on the MD simulations that the more converged the simulations are the more reliable the outcome will be. The protein system that we tested here is a relatively small one (110 residues), so we can expect relatively good sampling within 1 μs of simulation time (Figure S4). However, for larger systems, longer MD simulations should be performed.

Despite the above limitations, the results obtained in this study showed that combining machine learning models with atomistic descriptive models is a promising approach in computational structural biology and could give us new insight into the hidden structure−activity relationships in biological macromolecules. Such combinatorial approaches inherit the benefits of each developing component and are anticipated to lead to new important applications for these models. In this work, we demonstrated the usages of these approaches to (1) introduce a novel method for analysis of MD simulations to characterize the responses of each residue upon perturbation and (2) to introduce the residue response maps which can redefine allostery as a residue−property specific phenomenon. Such combinatorial approaches can be used in the future to analyze the effects of ligand binding, mutation and post-translational modifications in macromolecules.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00447.

Figure S1: Accuracies of neural network and random forest models trained with the first 10 ns chunk of bound and unbound trajectories. Figure S2: Residue response maps resulted from analysis of the first 10 ns chunk of bound and unbound trajectories with and without the binding pocket residues. Figure S3: Accuracies of neural network models utilizing the information from different residues of all MD snapshots. Figure S4: RMSD Histograms of two independent runs of the ligand-bound MD simulations. Table S1: Raw random forest weights calculated with the pocket residues. Table S2: Raw random forest weights calculated without the pocket residues. Table S3: Comparison between experimental order parameters of many methyl groups within the PDZ3 domain and the weights suggested by the random forest dynamics models and also the difference between the RMSF values of bound and unbound trajectories. Table S4: Top 10 residues identified by C_α and GEOM models as residues supposedly involved in allosteric mechanisms based on structural changes vs top 10 residues identified by DYN models as residues supposedly involved in allosteric mechanisms based on dynamics changes. Sample analysis scripts used to calculate residue properties from molecular dynamics trajectories. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jin.liu@unthsc.edu.

ORCID

Hamed S. Hayatshahi: 0000-0001-8639-7130

Peng Tao: 0000-0002-2488-0239

Jin Liu: 0000-0002-1067-4063

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the High-Performance Computing Center at the University of North Texas and the Texas Advanced Computing Center (TAAC) for providing computational resources for MD simulations and training of Deep Neural Networks related to this research work.

■ REFERENCES

(1) Liu, J.; Nussinov, R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Comput. Biol.* **2016**, *12*, No. e1004966.

(2) Monod, J.; Wyman, J.; Changeux, J. P. On the Nature of Allosteric Transitions: A Plausible Model. *J. Mol. Biol.* **1965**, *12*, 88−118.

(3) Koshland, D. E., Jr.; Nemethy, G.; Filmer, D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. *Biochemistry* **1966**, *5*, 365−385.

(4) Kumar, S.; Ma, B.; Tsai, C. J.; Wolfson, H.; Nussinov, R. Folding Funnels and Conformational Transitions via Hinge-Bending Motions. *Cell Biochem. Biophys.* **1999**, *31*, 141−164.

(5) Ma, B.; Kumar, S.; Tsai, C. J.; Nussinov, R. Folding Funnels and Binding Mechanisms. *Protein Eng., Des. Sel.* **1999**, *12*, 713−720.

(6) Tsai, C. J.; Kumar, S.; Ma, B.; Nussinov, R. Folding Funnels, Binding Funnels, and Protein Function. *Protein Sci.* **1999**, *8*, 1181−1190.

(7) Guo, J.; Zhou, H. X. Protein Allostery and Conformational Dynamics. *Chem. Rev.* **2016**, *116*, 6503−6515.

(8) Kumawat, A.; Chakrabarty, S. Hidden Electrostatic Basis of Dynamic Allostery in a PDZ Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E5825−E5834.

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

(9) Nussinov, R.; Tsai, C. J. Allostery Without a Conformational Change? Revisiting the Paradigm. *Curr. Opin. Struct. Biol.* **2015**, *30*, 17−24.
(10) Gunasekaran, K.; Ma, B.; Nussinov, R. Is Allostery an Intrinsic Property of All Dynamic Proteins? *Proteins: Struct., Funct., Genet.* **2004**, *57*, 433−433.
(11) Guarnera, E.; Berezovsky, I. N. Structure-Based Statistical Mechanical Model Accounts for the Causality and Energetics of Allosteric Communication. *PLoS Comput. Biol.* **2016**, *12*, No. e1004678.
(12) Tee, W. V.; Guarnera, E.; Berezovsky, I. N. Reversing Allosteric Communication: From Detecting Allosteric Sites to Inducing and Tuning Targeted Allosteric Response. *PLoS Comput. Biol.* **2018**, *14*, No. e1006228.
(13) Lockless, S. W.; Ranganathan, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999**, *286*, 295−299.
(14) Petit, C. M.; Zhang, J.; Sapienza, P. J.; Fuentes, E. J.; Lee, A. L. Hidden Dynamic Allostery in a PDZ Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 18249−18254.
(15) Lee, H. J.; Zheng, J. J. PDZ domains and their binding partners: structure, specificity, and modification. *Cell Commun. Signaling* **2010**, *8*, 8.
(16) Gerek, Z. N.; Ozkan, S. B. Change in Allosteric Network Affects Binding Affinities of PDZ Domains: Analysis Through Perturbation Response Scanning. *PLoS Comput. Biol.* **2011**, *7*, No. e1002154.
(17) McLaughlin, R. N., Jr.; Poelwijk, F. J.; Raman, A.; Gosal, W. S.; Ranganathan, R. The Spatial Architecture of Protein Function and Adaptation. *Nature* **2012**, *491*, 138−142.
(18) Kaya, C.; Armutlulu, A.; Ekesan, S.; Haliloglu, T. MCPath: Monte Carlo Path Generation Approach to Predict Likely Allosteric Pathways and Functional Residues. *Nucleic Acids Res.* **2013**, *41*, W249−W255.
(19) Murciano-Calles, J.; Corbi-Verge, C.; Candel, A. M.; Luque, I.; Martinez, J. C. Post-Translational Modifications Modulate Ligand Recognition by the Third PDZ Domain of the MAGUK Protein PSD-95. *PLoS One* **2014**, *9*, No. e90030.
(20) Kalescky, R.; Liu, J.; Tao, P. Identifying Key Residues for Protein Allostery Through Rigid Residue Scan. *J. Phys. Chem. A* **2015**, *119*, 1689−16700.
(21) Doyle, D. A.; Lee, A.; Lewis, J.; Kim, E.; Sheng, M.; MacKinnon, R. Crystal Structures of a Complexed and Peptide-Free Membrane Protein-Binding Domain: Molecular Basis of Peptide Recognition by PDZ. *Cell* **1996**, *85*, 1067−1076.
(22) Sethi, A.; Eargle, J.; Black, A. A.; Luthey-Schulten, Z. Dynamical Networks in tRNA: Protein Complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6620−6625.
(23) VanWart, A. T.; Eargle, J.; Luthey-Schulten, Z.; Amaro, R. E. Exploring Residue Component Contributions to Dynamical Network Models of Allostery. *J. Chem. Theory Comput.* **2012**, *8*, 2949−2961.
(24) Van Wart, A. T.; Durrant, J.; Votapka, L.; Amaro, R. E. Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *J. Chem. Theory Comput.* **2014**, *10*, 511−517.
(25) Bowerman, S.; Wereszczynski, J. Detecting Allosteric Networks Using Molecular Dynamics Simulation. *Methods Enzymol.* **2016**, *578*, 429−447.
(26) Kokh, D. B.; Richter, S.; Henrich, S.; Czodrowski, P.; Rippmann, F.; Wade, R. C. TRAPP: A Tool for Analysis of Transient Binding Pockets in Proteins. *J. Chem. Inf. Model.* **2013**, *53*, 1235−1252.
(27) La Sala, G.; Decherchi, S.; De Vivo, M.; Rocchia, W. Allosteric Communication Networks in Proteins Revealed through Pocket Crosstalk Analysis. *ACS Cent. Sci.* **2017**, *3*, 949−960.
(28) Guo, J.; Pang, X.; Zhou, H.-X. Two Pathways Mediate Interdomain Allosteric Regulation in Pin1. *Structure* **2015**, *23*, 237−247.
(29) Guo, J.; Zhou, H.-X. Dynamically Driven Protein Allostery Exhibits Disparate Responses for Fast and Slow Motions. *Biophys. J.* **2015**, *108*, 2771−2774.
(30) Hertig, S.; Latorraca, N. R.; Dror, R. O. Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations. *PLoS Comput. Biol.* **2016**, *12*, No. e1004746.
(31) Dror, R. O.; Arlow, D. H.; Maragakis, P.; Mildorf, T. J.; Pan, A. C.; Xu, H.; Borhani, D. W.; Shaw, D. E. Activation mechanism of the β2-adrenergic receptor. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (46), 18684−18689.
(32) Chabria, M.; Hertig, S.; Smith, M. L.; Vogel, V. Stretching Fibronectin Fibres Disrupts Binding of Bacterial Adhesins by Physically Destroying an Epitope. *Nat. Commun.* **2010**, *1*, 135.
(33) Tarca, A. L.; Carey, V. J.; Chen, X. W.; Romero, R.; Drăghici, S. Machine Learning and its Applications to Biology. *PLoS Comput. Biol.* **2007**, *3*, No. e116.
(34) Zhang, S. Application of Machine Learning in Drug Discovery and Development. *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques.* **2011**, 235.
(35) Case, D. A.; Betz, R. M.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Giese, T.J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R.C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; Kollman, P. A. *Amber 2016*; University of California: San Francisco, 2016.
(36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(37) Joung, I. S.; Cheatham, T. E., 3rd. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112*, 9020−9041.
(38) Hayatshahi, H. S.; Roe, D. R.; Galindo-Murillo, R.; Hall, K. B.; Cheatham, T. E. Computational Assessment of Potassium and Magnesium Ion Binding to a Buried Pocket in GTPase-Associating Center RNA. *J. Phys. Chem. B* **2017**, *121*, 451−462.
(39) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *J. Comput. Chem.* **1999**, *20*, 786−798.
(40) Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **2015**, *11*, 1864−1874.
(41) 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.
(42) Uberuaga, B. P.; Anghel, M.; Voter, A. F. Synchronization of Trajectories in Canonical Molecular Dynamics Simulations: Observation, Explanation, and Exploitation. *J. Chem. Phys.* **2004**, *120*, 6363−6374.
(43) Sindhikara, D. J.; Kim, S.; Voter, A. F.; Roitberg, A. E. Bad Seeds Sprout Perilous Dynamics: Stochastic Thermostat Induced Trajectory Synchronization in Biomolecules. *J. Chem. Theory Comput.* **2009**, *5*, 1624−1631.
(44) Åqvist, J.; Wennerström, P.; Nervall, M.; Bjelic, S.; Brandsdal, B. O. Molecular dynamics Simulations of Water and Biomolecules with a Monte Carlo Constant Pressure Algorithm. *Chem. Phys. Lett.* **2004**, *384*, 288−294.
(45) 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.
(46) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705

(47) Roe, D. R.; Cheatham, T. E., 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084−3095.
(48) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217−230.
(49) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467*, 2016.
(50) Kingma, D.; Ba, J. A. A Method for Stochastic Optimization. *arXiv:1412.6980*, 2014.
(51) Rubinstein, R. Y. Optimization of Computer Simulation Models with Rare Events. *European Journal of Operational Research* **1997**, *99*, 89−112.
(52) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Machine Learning Res.* **2011**, *12*, 2825−2830.
(53) Li, H.; Chang, Y.-Y.; Lee, J. Y.; Bahar, I.; Yang, L.-W. DynOmics: Dynamics of Structural Proteome and Beyond. *Nucleic Acids Res.* **2017**, *45*, W374−W380.
(54) Gianni, S.; Walma, T.; Arcovito, A.; Calosci, N.; Bellelli, A.; Engstrom, A.; Travaglini-Allocatelli, C.; Brunori, M.; Jemth, P.; Vuister, G. W. Demonstration of Long-Range Interactions in a PDZ Domain by NMR, Kinetics, and Protein Engineering. *Structure* **2006**, *14*, 1801−1809.
(55) Ye, F.; Liu, W.; Shang, Y.; Zhang, M. An Exquisitely Specific PDZ/Target Recognition Revealed by the Structure of INAD PDZ3 in Complex with TRP Channel Tail. *Structure* **2016**, *24*, 383−391.
(56) Lee, A. L. Contrasting Roles of Dynamics in Protein Allostery: NMR and Structural Studies of CheY and the Third PDZ Domain from PSD-95. *Biophys. Rev.* **2015**, *7*, 217−226.
(57) Kumawat, A.; Chakrabarty, S. Hidden Electrostatic Basis of Dynamic Allostery in a PDZ Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E5825−E5834.
(58) Liu, J.; Nussinov, R. Energetic Redistribution in Allostery to Execute Protein Function. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 7480−7482.

DOI: 10.1021/acs.jcim.9b00447
J. Chem. Inf. Model. 2019, 59, 4691−4705