Recognition of Protein Allosteric States and Residues: Machine Learning Approaches

Hongyu Zhou, Zheng Dong, and Peng Tao *

Allostery is a process by which proteins transmit the effect of perturbation at one site to a distal functional site upon certain perturbation. As an intrinsically global effect of protein dynamics, it is difficult to associate protein allostery with individual residues, hindering effective selection of key residues for mutagenesis studies. The machine learning models including decision tree (DT) and artificial neural network (ANN) models were applied to develop classification model for a cell signaling allosteric protein with two states showing extremely similar tertiary structures in both crystallographic structures and molecular dynamics simulations. Both DT and ANN models were developed with 75% and 80% of predicting accuracy, respectively. Good agreement between machine learning models and previous experimental as well as computational studies of the same protein validates this approach as an alternative way to analyze protein dynamics simulations and allostery. In addition, the difference of distributions of key features in two allosteric states also underlies the population shift hypothesis of dynamics-driven allostery model. © 2018 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.25218

Introduction

Allostery, which is referred to as a process by which proteins transmit the effect of perturbation at one site to a distal functional site, is fundamental to many biological regulations. Numerous studies have been conducted in the past half centuries. In the early 60s, two theoretical models, Monod–Wyman–Changeux (MWC)[1] and Koshland–Némethy–Filmer (KNF) models,[2] were proposed to explain significant conformational change observed in protein hemoglobin upon binding with oxygen molecules as concerted or sequential processes, respectively. Since then, protein allostery was commonly considered as the significant conformational change observed in protein structure upon local perturbation. However, there are many allosteric proteins being identified without significant conformational change upon perturbation. In contrast to the conformation-driven allostery observed in hemoglobin, new theoretical models were proposed as dynamics-driven allostery[3–5] or population shift among different states[6–10] to explain protein allostery without significant conformational changes. In these models, it was proposed that the external perturbations cause significant changes in the distribution of protein in different states, and lead to the change of free energy landscape related to protein allosteric functions. Various studies were carried out to distinguish different states through simulations[11–14] using principal component analysis based on the cross correlation matrix of protein simulations. Despite the progress made in these studies, further development is still necessary for better recognition of the different states of dynamics-driven allosteric proteins.

Identifying allostery-related residues and the pathways responsible for allosteric transformation is another challenge for the protein allostery studies. The theory for allosteric information transduction within the proteins has evolved from single pathway formed by residues into allosteric information transduction network model.[15] Numerous methods for identifying key allosteric residues from simulations have been developed recently.[11,16–19] These computational methods focus on correlation analysis related to protein dynamics. Potential contribution from simple geometric parameters, such as distances between residues or dihedral angles to allostery, has not been explored extensively.

In computer science, machine learning (ML) methods were developed for many purpose including pattern classification.[20] Due to their various advantages, ML methods have also been applied in computational biology.[21–24] Many ML methods are specialized in classification with high accuracy, and can also provide insights into the intrinsic differences in classification model. Therefore, ML methods are applied in this study to develop classification model with regard to protein allostery. Specifically, two widely applied ML methods, neural networks and decision tree models, are used to analyze geometric parameters including distances among residues and backbone dihedral angles, and develop prediction models to differentiate states of dynamics-driven allosteric proteins.

Neural network, also named as artificial neural network, was first proposed in the 1960s[25,26] to mimic the biological neural networks in animal brains. Recently, being developed as deep learning methods, the artificial neural network model has been widely used in many applications, including artificial intelligence and image recognition.[27,28] Since its initial application

H. Zhou, Z. Dong, P. Tao
Department of Chemistry, Center for Drug Discovery, Design, and Delivery (CD4), Center for Scientific Computation, Southern Methodist University, Dallas, Texas 75275
E-mail: ptao@smu.edu

Contract grant sponsor: Southern Methodist University Dean's Research Council research fund, and American Chemical Society Petroleum Research Fund; Contract grant number: 57521-DNI6

© 2018 Wiley Periodicals, Inc.

in computational chemistry in the 1990s,[29,30] artificial neural network model has been applied in rational drug design.[31–33] Being a nonlinear activation function, artificial neural network method is particularly suitable for modeling nonlinear relationships.[34]

Decision tree model, as another ML method, is widely used to identify key factors that contribute the most to the target states. In general, decision tree model is easy to apply on large amount of data with high dimensions. The resulted classification model based on the decision tree method is also easy to interpret related to the nature of the systems being studied.[35] Due to these advantages, decision tree model is often used to preprocess raw data in combination with other ML methods. Therefore, both artificial neural network and decision tree methods were applied in this study to develop prediction models for protein allostery.

The second PDZ domain (PDZ2) in the human PTP1E protein is a typical dynamics-driven allosteric protein upon binding with its allosteric effectors, and has been subjected to both experimental and computational investigations. Therefore, it is used as model system in this study and subjected to above two ML methods to develop classification models associated with its allosteric states. There are two goals to achieve in this study: developing theoretical prediction models to recognize two allosteric states of PDZ2 (unbound and bound) and identifying key geometric features that potentially drive allostery of this protein. It is expected that the selected ML methods could facilitate to reveal key features to influent the overall protein allosteric processes.

Methods

Molecular dynamics simulations

The initial structures of PDZ2 protein were obtained from Protein DataBank (PDB)[36] with codes as 3LNX and 3LNY, for the unbound and bound states, respectively (Fig. 1). These PDB structures were processed with hydrogen atoms added and solvated in a cubic water box as TIP3P model[37] with charge balancing ions as sodium and chlorine added. The systems were then subjected to energy minimization. Consequently, the systems were subjected to 12 picoseconds (ps) molecular dynamics (MD) simulations to gradually raise the temperature to 300K before being equilibrated via 10 nanoseconds (ns) isothermal–isobaric ensemble (NPT) MD simulations at 300 K and 1atm. Afterwards, canonical ensemble (NVT) Langevin MD simulations were carried out as the production runs. For all above simulations, 2 femtoseconds (fs) step size was used. The chemical bonds associated with hydrogen were fixed using SHAKE method.[38] Cubic periodic boundary condition (PBC) was applied in these simulations. The long-range electrostatic interactions were modeled using the particle mesh Ewald algorithm.[39] All simulations were carried out using CHARMM simulation package[40] version 40b1 and the CHARMM22 force field.[41] For both unbound and bound states of PDZ2, total of 13 simulations of 34 ns in length were carried out. For all trajectories, the initial 4 ns were discarded as equilibrium phase. Frames were saved every 10 ps. Therefore, 3000 frames were extracted from each 30 ns trajectory and subjected to the ML model analysis. Among 13 simulations of each state, 10 simulations were randomly selected as training set, and remaining three simulations were used as testing set. Cross-validation on training set was used to optimize the classification models and tested on test sets.

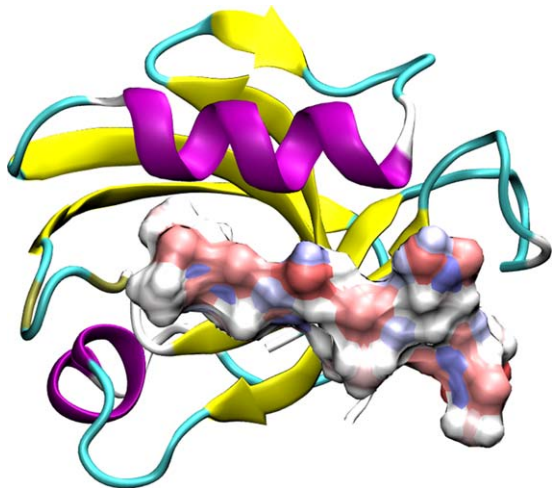

Figure 1. PDZ2 bound state with peptide. [Color figure can be viewed at wileyonlinelibrary.com]

Machine learning methods

The machine learning methods applied in this study include the artificial neural network (ANN) model, and the decision tree (DT) model. A typical ANN model consists of input layer, hidden layers, and output layer. Each layer consists a set of "nodes" interconnected with other nodes in the adjacent layer(s). These nodes contain activation functions. The connections among nodes are weighted by additional factors. During the training process of an ANN model, the original data from the training set were entered to the input layer and went through the hidden layer(s) before reaching the output layer. A feedback process called "back propagation" was employed to minimize the error at the output layer. The purpose of the back propagation is optimizing the activation functions and weights on internode connections to achieve the minimum prediction error at the output layer upon convergence.[42] When there is more than one hidden layer, ANN is also referred to as deep neural network model, which usually requires much higher computational cost in training process.[43] Therefore, only one hidden layer was used in the initial ANN model setup, and was shown to be sufficient. An additional regularization including L2 penalty term was used to avoid over-fitting problem in the training process. L2 penalty term was added to the ANN model when updating the weight of each node. This penalty term limits the changes of weights during each iteration to avoid over-fitting. Overall, the number of nodes in hidden layer and L2 penalty term was refined to achieve the highest accuracy.[44–46]

DT method has been widely used in strategy determination and identification of important factors. Combined with

chemical descriptors, DT method was also applied to predict chemical activities.[47] The DT method was also applied in this study to develop a classification model. It provides an efficient algorithm to identify how the results can be predicted from individual features based on the information entropy gain. The DT model implemented in scikit-learn package[48] was employed and refined to achieve the best predicative model in this study. Comparing to the ANN model, the classification or prediction model resulted from DT model is easier to interpret and understand.

Both pairwise distances for alpha carbons (Cα) and backbone dihedral angles (ψ and φ) were used as features to train the ANN and DT models. MSMbuilder[49] package was employed to extract full Cα pairwise distances and dihedral angles from simulation trajectories. For better performance of these two ML models, prescreening all features is necessary. Tree-based feature selection methods implemented in scikit-learn package[48] were applied to prescreen important features for the ML analyses presented in this study.

To assess the performance of each classification model, we calculated four summary metrics including accuracy, recall, precision, and F1 score, which are defined as

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{all}};\ \text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \quad (1)$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}};$$

$$F_1=2\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

where true positive (TP) and true negative (TN) are defined as the number of structures that are classified correctly into unbound and bound state. False positive (FP) and false negative (FN) are defined as the number of structures that are misclassified into the other states.

Analysis of MD trajectories

Root-Mean-Square Deviation (RMSD) and Root-Mean-Square Fluctuation (RMSF). The RMSD is used to measure the overall conformational change during the MD simulations with regard to a reference structure. For a molecular structure represented by Cartesian coordinate vector r_i (i=1 to N) of N atoms, the RMSD is calculated as the following:

$$\text{RMSD}=\sqrt{\frac{\sum_{i=1}^{N}\left(r_i^0-Ur_i\right)^2}{N}} \quad (2)$$

The Cartesian coordinate vector r_i^0 is the ith atom in the reference structure. The transformation matrix U is defined as the best-fit alignment between the PDZ2 structures along trajectories with respect to the reference structure.

RMSF is used to measure the fluctuation of atoms during MD simulations with respect to the averaged structure. RMSF_i of atom i for a given MD trajectory is defined as

$$\text{RMSF}_i=\sqrt{\frac{1}{T}\sum_{j=1}^{T}\left(v_i^j-\overline{v_i}\right)^2}, \quad (3)$$

where T is the total number of frames in the given MD trajectory, v_i^j is the coordinate atom i in the frame j, and $\overline{v_i}$ is the averaged coordinate of atom i in the given trajectory. This analysis is based on the simulation frames superimposed to the averaged structure of the given trajectory.

Principal Component Analysis (PCA). By applying quasi-harmonic analysis implemented in the CHARMM program, PCA was performed on the unbound and bound state simulations to obtain dominant modes in each state. Translational and rotational components were projected out for each frame. All analyses were carried out using CHARMM simulation package version 40b1.

Cross-correlation matrix is a measurement of the correlated movement of a set of atoms. Each matrix element is defined as

$$C_{ij}=\frac{c_{ij}}{c_{ii}^{1/2}c_{jj}^{1/2}}=\frac{\langle r_i r_j\rangle-\langle r_i\rangle\langle r_j\rangle}{\left[\left(\langle r_i^2\rangle-\langle r_i\rangle^2\right)\left(\langle r_j^2\rangle-\langle r_j\rangle^2\right)\right]^{1/2}}, \quad (4)$$

where C_{ij} is the measurement of the correlated movement between atoms i and j, c_{ij}, c_{ii}, and c_{jj} are the covariance matrix elements, and r_i and r_j are Cartesian coordinate vectors from the least-square fitted structures, hence with translation and rotation projected out. Matrix elements C_{ij} are between -1 and 1 with negative values indicating negative correlation and positive values indicating positive correlation between the motions of atoms i and j. It should be noted that the correlation is defined as related movement along the line between two points. Correlated movement along orthogonal paths yields a cross-correlation matrix element of zero.[50]

Dynamical Network Analysis. Potential allosteric pathways consisting residues identified by machine learning models were examined through dynamical network analysis using the *NetworkView* plugin implemented in VMD program.[51,52] In the dynamical network analysis, if the backbone alpha carbons of any residue pairs are within 4.5 Å for more than 75% of simulation time, these two residues are considered as being connected. The connection strength for each connected residue pair is weighted by correlation value of these two residues in the cross-correlation matrix. For any two residues not connected, optimal pathways may be identified through other connected residues and the connections among them.

Results

Prescreening features for further analysis

The pairwise distances for Cα and backbone dihedral angles were subjected to a prescreening process using DT model to select features for efficient machine learning analysis. All 26 trajectories for both unbound and bound PDZ2 states were used for the prescreening purpose. Total of 4371 Cα pair

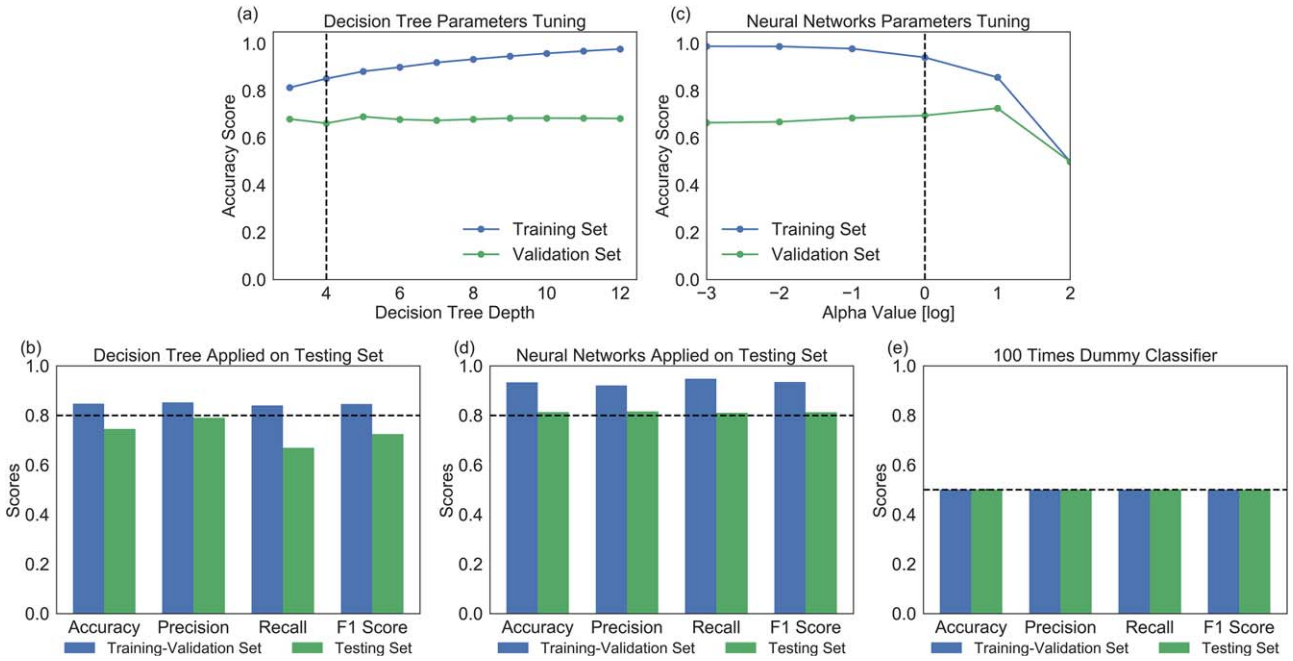

Figure 2. Machine learning models for PDZ2. a) Decision tree (DT) model parameters refinement, b) DT model testing results, c) artificial neural network (ANN) model parameters refinement, d) ANN model testing results, e) benchmark dummy classifier. [Color figure can be viewed at wileyonlinelibrary.com]

distances and backbone dihedral angles were subjected to the prescreening process. The number of important features that could be selected depends on the depth of DT model. With the depth n, the maximum number of features that can be covered in the model is 2^n–1. For feature prescreening purpose, to ensure that the DT model covers all the possible features in the affordable computational costs, the depth of DT model was set as 20. After training this DT model, total of 289 features each with importance greater than 0.1% were selected for the following analysis. Combined together, these 289 features contribute 90.0% as total importance to the model.

PDZ2 state classification by DT and ANN models

Using the preselected 289 features, the DT model was further refined through the following training procedure. Ten trajectories were randomly selected among 13 independent simulation trajectories as training set for the unbound and bound states of PDZ2, respectively. For each state, 10 selected trajectories were randomly divided into five groups each with two trajectories. For each 30 ns trajectory, 3000 frames evenly distributed along the trajectory were selected for the training and testing purpose. The five groups of trajectories of both unbound and bound states were subjected to five rounds of cross-validation process described as the following. In each round of the validation process, one group of both unbound and bound states trajectories was selected as the test set for validation purpose with the remaining four groups as the training set.

For the DT model, depths of the tree ranging from 3 to 12 were tested in the cross-validation process. With depths as 4 and 5, the best performance is achieved to avoid potential over-fitting problem (Fig. 2a). The DT model with depth 4 showed higher prediction power for the additional six simulations of unbound and bound states than the one with depth 5. Therefore, the DT model with depth 4 was selected as the final model. For ANN model, six different values of a parameter alpha, also referred to as learning rate, were tested for the best performance, with alpha as 1 (log(alpha)=0) leading to the best prediction model (Fig. 2b). For the best DT model with depth 4 and ANN model with alpha as 1, the prediction accuracy for the six testing trajectories is 75% and 80%, respectively (Figs. 2c and 2d). In addition, one dummy classifier was built to generate random predictions as a baseline comparison for the ANN and DT classifiers. Random dummy predictions were repeated 100 times, and the metrics calculated by averaging these 100 dummy classifications is 0.5 with standard deviation as 0.0034 (Fig. 2e). The differences between the baseline dummy classifier and the ANN or DT classifier suggest that, although the unbound and bound states have similar structure with less than 2 Å RMSD differences, these two states are clearly differentiable using machine learning methods.

One of the advantages about the two prediction models using machine learning methods is that they could calculate the probability of any given structure that belongs to either unbound or bound state. The distribution of this probability was calculated for all the testing trajectories using both DT and ANN models, and is plotted in Figure 3. In the distributions calculated using DT model, there are five peaks in each state. Each peak from one state overlaps with a corresponding peak from the other state. The major difference between each peak from two states is the height (Fig. 3a). For example, the unbound state simulations have the highest peak close to the unbound state end of x-axis. For the bound state, the highest peak is the closest to the bound state end of x-axis. However,

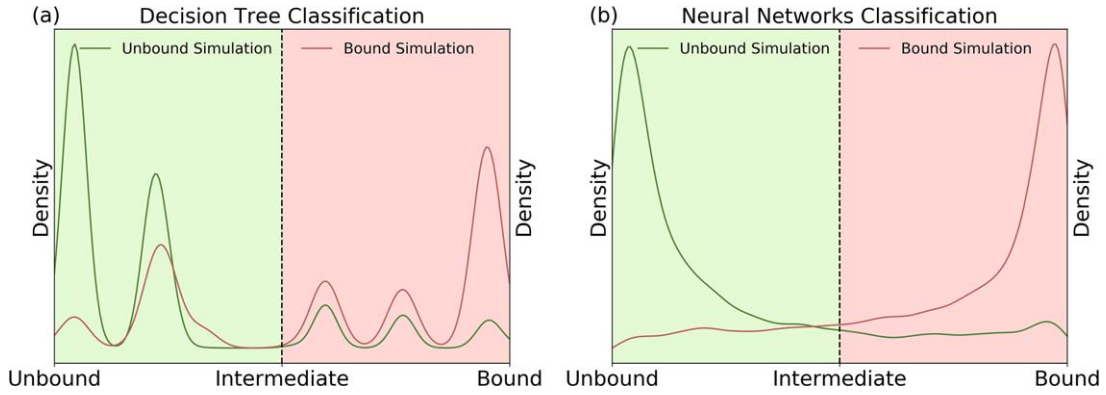

Figure 3. Probability distribution for unbound and bound states simulations: a) decision tree model and b) artificial neural network model. Unbound/intermediate/bound states are defined based on probabilities. [Color figure can be viewed at wileyonlinelibrary.com]

the second highest peak of the bound state is close to the unbound state end. In the ANN prediction model, the probability distribution of each state has only one major peak very close to each end of the *x*-axis, reflecting the high prediction accuracy of this model. In addition to the differentiation between two states, the calculated probabilities could also be utilized to select representative structures for various states, especially those different from both unbound and bound states, which are referred to as intermediate states. Using the probabilities calculated by the ANN model, the representative structures were selected for the unbound, bound, and intermediate states (Fig. 4). The colored arrows in unbound and bound states provide structural information differentiating these states from the intermediate state.

Identifying key residues

Another important implication of machine learning models is identifying the important features strongly correlated with allosteric states. In both DT and ANN models, the contribution from each feature to differentiate two states is calculated and can be used to rank the features. In the two models of this study, both Cα distances and backbone dihedral angles are used and ranked together based on their contributions. The top 10 features with the highest contributions are listed in Table 1 for the DT and ANN models, respectively. In the DT model, eight top features are Cα distances, while five of top ten features are Cα distances in the ANN model. Among the top 10 features, two models share three features (Cα distance between residues 38 and 71, backbone dihedral angle ψ connecting residues 1 and 2, backbone dihedral angle ϕ connecting residues 22 and 23). Among the top 10 features reported from the DT and ANN models, there are 19 different residues involved. Total of 16 among these 19 residues have been identified as related to PDZ2 allostery upon binding with the same peptide in several studies[53–56] The top three features listed in Table 1 from the DT and ANN models are subjected to further analysis described as the following.

Further analysis of the key residues

To illustrate the difference between the distributions of the unbound and bound states of PDZ2, a 2D-RMSD plot with reference to the crystal unbound and bound structures is shown in Figure 5a. The distribution plot shows that the bound state simulations sampled a region similar to the unbound state simulation, but covered larger conformational space. To further compare the simulations of the two states, distributions of three key features identified in the DT and ANN models (Cα distances between residues Lys38 and His71 and between residues Asn16 and Arg31, backbone dihedral angle ψ connecting residues Pro1 and Lys2) are plotted in Figures 5b–5d. Cα distance between Lys38 and Thr70 was not plotted because residue Thr70 is adjacent to residue His71. Interestingly, although the unbound and bound states have similar structures with low RMSD difference, the distributions of these three key features are significantly different between the two states. For the dihedral angle between residues Pro1 and Lys2, which

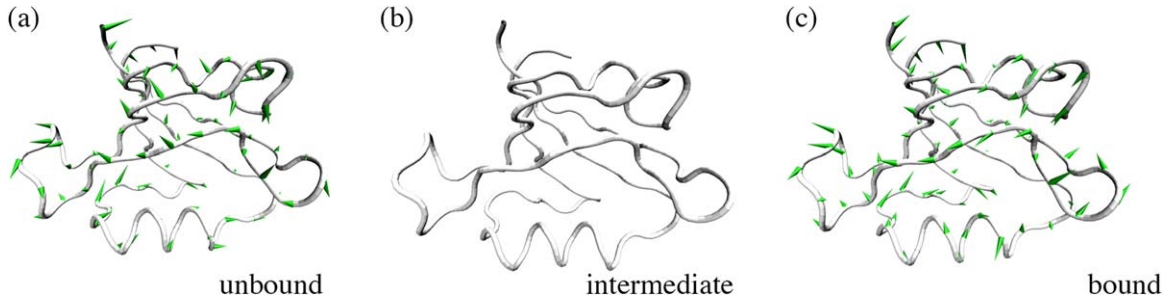

Figure 4. Representative structures for: a) unbound state, b) a representative intermediate state, and c) bound state. The colored arrows in unbound and bound states indicate the direction and magnitude of difference with reference to the intermediate state. [Color figure can be viewed at wileyonlinelibrary.com]

Table 1. Top 10 important features identified by decision tree and artificial neural networks models.

Decision tree		Neural networks	
Type	Residues	Type	Residues
Cα distance	38[b], 71[a,b]	ψ angle	1[b], 2[b]
ψ angle	1[b], 2[b]	Cα distance	38[b], 70[b]
Cα distance	16[a,b], 31[a,b]	Cα distance	38[b], 71[a,b]
Cα distance	31[a,b], 69[a,b]	φ angle	22[a,b], 23[b]
Cα distance	18[a,b], 28[b]	Cα distance	38[b], 73[b]
Cα distance	23[b], 31[a,b]	ψ angle	92, 93
Cα distance	31[a,b], 71[a,b]	ψ angle	22[a,b], 23[b]
Cα distance	7[b], 30[b]	Cα distance	24[b], 54
ψ angle	22[a,b], 23[b]	ψ angle	22[a,b], 23[b]
Cα distance	31[a,b], 52[b]	Cα distance	22[a,b], 73[b]

[a] Residue has already been identified by NMR studies.[53] [b] Residue has already been identified by other computational studies.[55,56]

appeared as the top feature in ANN model and the second most important feature in the DT model, the relative heights of two peaks are switched in the bound state compared with the unbound state. This observation is consistent with the population shift hypothesis,[8] that the free energy landscapes of two allosteric states are different upon perturbations despite the similarity of their structures. The distribution of the Cα distance between Lys38 and His71 is also significantly different between the two states. The most probable value of this distance in the bound state is larger than the one in the unbound state (Fig. 5b). The distribution of the Cα distance between residues Asn16 and Arg31 is peaked around 29 Å in both states. But the probability at the peak is much higher in the unbound state than in the bound state (Fig. 5d). Interestingly, the pairing residues for key Cα distances, Lys38:His71 and Asn16:Arg31 are far from each other and across the protein structure, as they are located either on or close to distal loop structures (Fig. 6). These results suggest that the correlated fluctuation of Lys38:His71 and Asn16:Arg31 or their associated secondary structures play a critical role to differentiate the unbound and bound states, and hence, serve as key factors related to the PDZ2 allostery.

In addition to the distribution analysis, the fluctuations of the key residues are another comparison between the different simulations. RMSF analysis could be used to measure the averaged structural fluctuations of each residue in dynamics simulations. PCA is a widely applied method to analyze the global motion of protein structures based on dynamics simulations. Therefore, we applied RMSF and PCA on the simulations of both unbound and bound states of PDZ2. In the RMSF plot (Fig. 7a), the four key residues Asn16, Arg31, Lys38, and His71, display rather high fluctuations. The cumulative contributions from PCA modes are plotted for both unbound and bound states in Figure 7b. For both states, the 20 modes with lowest frequencies account for more than 50% of the total variances. Therefore, the average of these modes was used to measure the fluctuation of each residue in principal component (PC)

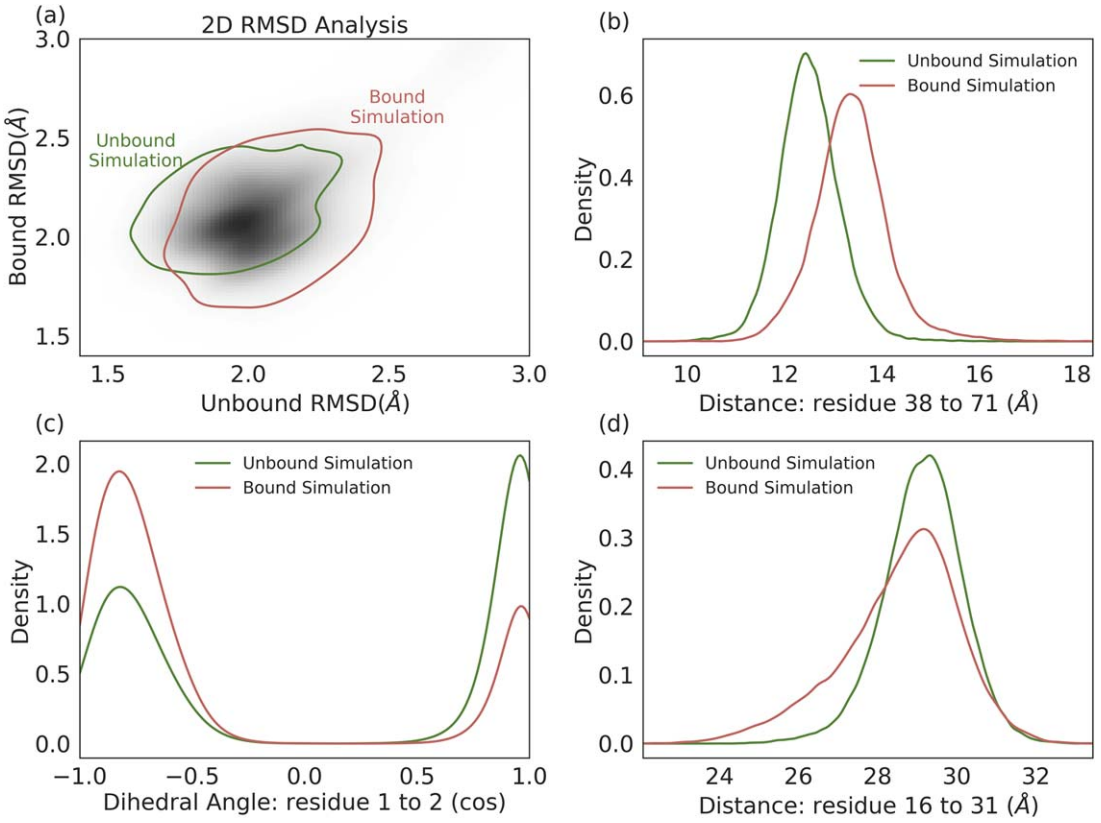

Figure 5. Distribution differences between the unbound and bound states for different features. a) 2D RMSD distribution, b) Cα distance between residue Lys38 and His71, c) dihedral angle between residue Pro1 and Lys2 (normalized by cosine value), d) Cα distance between residue Asn16 and Arg31. [Color figure can be viewed at wileyonlinelibrary.com]

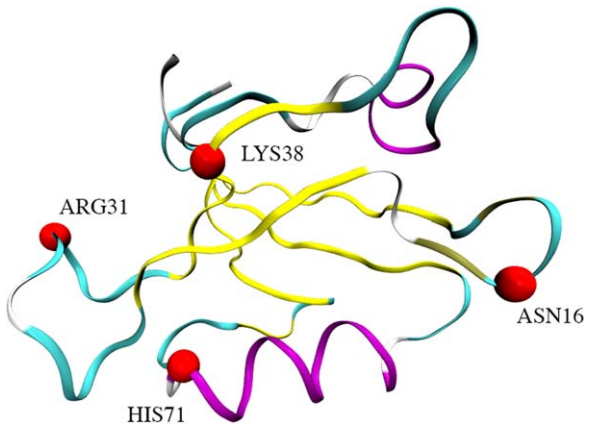

Figure 6. Four residues in the PDZ2 structure associated with the key Cα distances identified in the machine learning models. [Color figure can be viewed at wileyonlinelibrary.com]

vector space (Fig. 7c). Three residues, Asn16, Arg31, and His71 also display high fluctuations in the PC vector space.

PC1, the most dominant PC modes of unbound and bound states simulations, are illustrated in Figure 8. The loop between residues Val26 to Gly33 in both states displays higher fluctuation comparing with other part of the protein. Also, the bound state has a higher fluctuation than the unbound state. As shown in Figure 8, the fluctuation in that loop shows a trend to change the shape of the protein, which could be one of the reasons for the fluctuation difference between Asn16 and Arg31 as shown in Figure 5. Those differences between the PC1 modes in the two states could account for the allosteric effects.

Thus far, we focus on the distance distributions and residues fluctuations of key features identified by the machine learning models. The mechanisms how the key residue pairs are correlated with each other still remain unclear. Therefore, dynamical networks analysis,[51,52] a correlation-matrix-based method, was applied to identify potential allosteric pathways for Lys38:His71 and Asn16:Arg31 as the key residue pairs (Fig. 9). The analysis reveals that Val22 serves as one key residue involving correlation between Lys38 and His71. Although Val22 is not close to either Lys38 or His71 in sequence, it is located at the middle of these two residues in space and closer to Lys38 than to His71 (blue pathway in Fig. 9). In addition, four residues (Val22, Ile20, Leu18, Ser17) from the loop containing Asn16 and four residues (His32, Gly33, Gly34, and Tyr36) from the loop containing Arg31 form a communication pathway involved with the correlation between Asn16 and Arg31 (red pathway in Fig. 9). It is interesting that both pathways share the same residue Val22, which is also associated with multiple key features selected from the two machine learning models (Table 1) and other experimental and computational studies.[53–56]

Discussion

In this study, the decision tree and artificial neural networks models were applied to develop classification models of two allosterically related states of PDZ2 domain from PSD-95 protein. Principal component analysis of protein dynamics and RMS fluctuation analysis of individual residues were carried out to further evaluate the machine learning models. Dynamical network analysis was used to identify potential pathways accounting for the correlations among the key residues.

Classification of two states

In addition to the conformation-driven allostery, dynamics-driven allostery model plays increasingly important role in protein allostery from dynamical ensemble point of view.[8,57–59] In dynamics-driven allostery model, it is likely entropy instead of enthalpy that drives protein allostery because of the absence of significant conformational changes. There are some studies utilizing parameters associated with whole proteins instead of individual residues, such as RMSD, principal component analysis, and correlation matrix, to investigate protein allostery.[6,11,51,52] But there is still need for methods that differentiate allosteric states of proteins, build connections with individual residues, and provide guidance for mutagenesis studies to control protein allostery. Machine learning methods have been widely used in information technology classification applications,[60–62] and are regaining popularity in computational chemistry and biology.[63–65] One of the goals of this study is exploring new ways to differentiate protein allosteric states and build connections between protein allostery and individual residues. Therefore, in this study, the DT and ANN models were built to achieve more than 75% and 80% prediction accuracy to differentiate the unbound and bound allosteric states of PDZ2, respectively. More importantly, both models provide quantitative evaluation of features, which are associated with specific residues. The good agreement between the residues identified in both models and the previous experimental as well as

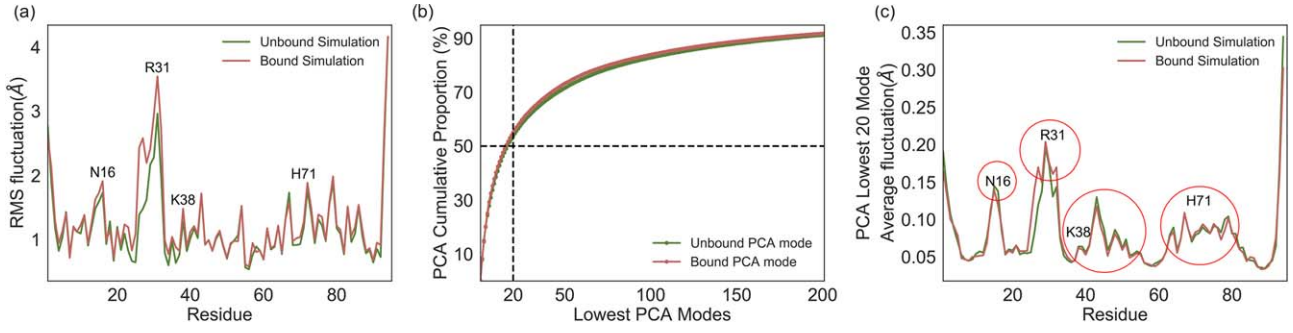

Figure 7. Residue fluctuations in RMSF method and PCA: a) RMSF, b) PCA cumulative variances, and c) fluctuation based on 20 PCA modes with lowest frequency for unbound and bound state simulations. [Color figure can be viewed at wileyonlinelibrary.com]

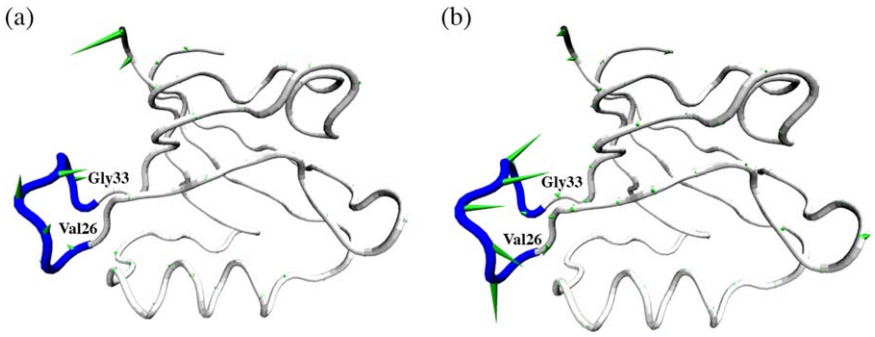

Figure 8. PC1 modes illustrated as porcupine plot: a) unbound state, b) bound state. Colored arrows indicate the direction and magnitude of movement. Val26-Gly33 loop is highlighted in blue color. [Color figure can be viewed at wileyonlinelibrary.com]

computational studies strongly suggests that the machine learning models could provide insight into protein allostery as compliment to other widely used analyses of protein simulations. The distinct distributions of key features in the unbound and bound states plotted in Figure 3 provide an alternative picture of population shift hypothesis underlying dynamics-driven protein allostery.[9]

One difficulty in protein allostery study is finding appropriate transition state with reference to distinct allosteric states.[9,18,57,58] Allosteric processes, especially dynamics-driven allostery, usually occur in short time scales, and are difficult to be characterized experimentally.[15,57] Using the quantitative machine learning models developed in this study, the distributions of simulations with regard to two allosteric states could be plotted (Fig. 4). The sampling located at the middle of two states could be considered as intermediate states and subjected to further analyses.

Given the effectiveness of the machine learning models presented in this study, one would logically expect that many other machine learning models could also be useful for analyzing simulations of protein allosteric states. Therefore comparison among different machine learning models for protein simulations will be the focus of future studies. In this study, the ANN model has better training and predication accuracy than the DT model. However to develop the accurate ANN model, the DT model is necessary to prescreen the potentially important features. This is mainly due to the different characteristics of these two methods. The DT model focuses on individual parameters as features for classification.[35] As contrast, the ANN model uses the combination of all features with different weights for classification purpose. In the future applications on different systems, caution should be used with regard to the choices and usage of machine learning models.

Identifying important features

An important strength displayed by machine learning models in this study is identifying key features for protein allostery. Both Cα distances and backbone dihedral angles can be easily used simultaneously for the development of accurate classification models. The fact that both distances and dihedral angles are among the top features suggests that many other order parameters of molecular simulation systems could be utilized for machine learning models for either allostery or other purposes such as computer-aided molecular design. The distributions of the selected individual features demonstrate significant difference between structurally similar allosteric states, and provide an alternative way of analyzing population shift of simulations upon allosteric or other perturbations on proteins. In addition, the key features specifically associated with individual residues provide unambiguous candidates for mutagenesis studies of proteins comparing to other studies using global descriptors of protein dynamics.[6,11]

The top 10 features identified in the DT and ANN models comprise 19 residues. It is significant that 16 of these 19 residues have been identified in one experimental NMR study and several computational studies.[53–56] For the top three features identified in this study, residues Pro1 and Lys2 serve as part of the allostery communication network identified in a protein network model.[56] Lys38 is regarded as one of the "hot residues" in another simulation study of PDZ2[54] as well as part of the communication network.[56] Asn16, Arg31, and His71 were identified as key allosteric residues in both NMR study[53] and other computational studies.[54–56] From biological point of view, Asn16 and His71 are located in the binding pocket, and could stabilize the binding peptide.[54,55] Residue Arg31 displayed a significant relaxation contribution value in a conformation exchange (R_{ex}) study.[47] Some experimental studies

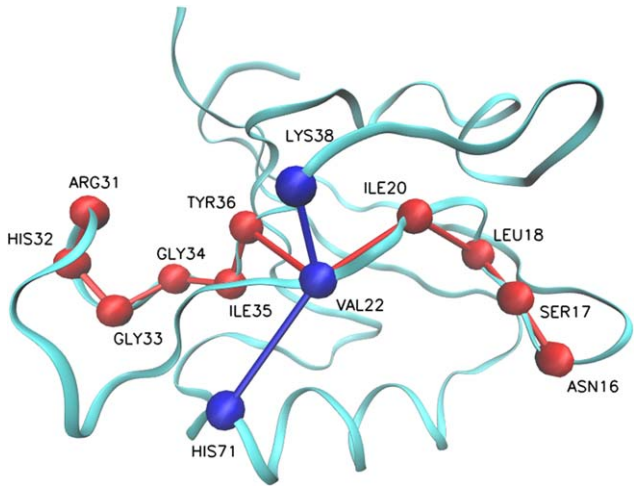

Figure 9. Network pathways for Lys38-His71 (blue) and Arg31-Asn16 (red). [Color figure can be viewed at wileyonlinelibrary.com]

also pointed out the residues located in $\beta1/\beta2$ loop including Asn16, and residues located in $\beta2/\beta3$ loop including Arg31 were important for peptide binding.[66] Considering that a large number of features associated with all the residues in the protein were treated equally to develop the classification models in this study, the overwhelming agreement with other studies strongly support the effectiveness of machine learning models for protein dynamics analysis.

The differences in the distributions of these key features between two states (Fig. 5) provide not only mechanistic insight into the machine learning models, but also a more quantitative view of population shift hypothesis of protein allostery. The different distribution of Cα distance between Lys38 and His71 may suggest the importance of secondary structures (a loop structure containing Lys38 and a helix structure containing His71, see Fig. 6) for the protein allostery. The key residue selections also agree with the residue fluctuation in the PCA and RMSF analysis (Fig. 7) in this study and another computational study of PDZ2.[56] The dynamical network analysis[51] identified two pathways containing additional residues which may play an important role for the communication between two key residue pairs (Fig. 9). The fact that residue Val22 being part of both pathways and also associated with several top key features identified in this study further support the notion that the machine learning models could be complimentary to the existing analysis tools of protein simulations by providing more insights related to individual residues. In general, these machine learning models could be applied to investigate the distribution differences between different states of dynamics-driven allosteric proteins, for which the conformational changes are not significant and the differences are difficult to be described by other analysis methods.

There might be concern that the important residues identified using ML methods are the outcome instead of the cause of allostery. According to the population shift hypothesis, distribution differences are essential for investigating the mechanism of allostery. Although not determined as either the cause or the outcome of allostery, it is an important step to identify the residues displaying distribution differences between two allosteric states. According to most experimental and computational studies about protein allostery, important residues behave differently through allosteric processes in the most cases. It may be possible that residues do not have any changes through allosteric processes but are fundamental to allostery effect. However probing these unlikely events is beyond the scope of this study.

Because the ML methods in this study do not require any *a priori* knowledge to identify most important residue pairs to differentiate two allosteric states, these models could serve as a complimentary method for the dynamical network analysis, which requires *a priori* knowledge about the source and target residues for the investigation of the potential allosteric pathways in the proteins of interest.

Conclusions

In this study, both decision tree and artificial neural network as machine learning models were applied to systematically investigate allosteric mechanism of PDZ2 protein upon binding with a peptide. Although there is no significant conformational change displayed between the unbound and bound states of PDZ2, two classification models developed in this study provide more than 75% of accuracy to differentiate these two states. Both models also provide a quantitative evaluation of the contributions from individual features to overall difference between the two states. Most residues associated with the important features including Cα distances and backbone dihedral angles have also been reported as key allosteric residues in both experimental and computational studies. Furthermore, the distributions of key features in different states provide alternative ways to analyze the population shift of protein ensemble upon allosteric perturbations. Additional analyses were also carried out for PDZ2 simulations using widely applied approaches including principal component analysis, RMS fluctuation analysis and dynamical network analysis, and showed good agreement with the machine learning results. Overall, the adopted machine learning methods on molecular dynamics simulations of protein in this study showed promise as a systematic and unbiased means to gain insight into protein allostery, especially the specific contribution from individual residues.

Acknowledgment

Computational time was provided by Southern Methodist University's Center for Scientific Computation and Texas Advanced Computing Center (TACC) at the University of Texas at Austin. The authors thank Drs. Gennady Verkhivker and Shouyi Wang for critical reading of the manuscript and fruitful discussions.

Keywords: allostery · machine learning · molecular dynamics · classification · protein

How to cite this article: H. Zhou, Z. Dong, P. Tao. *J. Comput. Chem*. **2018**, *39*, 1481–1490. DOI: 10.1002/jcc.25218

[1] J. Monod, J. Wyman, J.-P. Changeux, *J. Mol. Biol.* **1965**, *12*, 88.
[2] D. E. Koshland, G. Némethy, D. Filmer, *Biochemistry* **1966**, *5*, 365.
[3] A. Cooper, D. T. F. Dryden, *Eur. Biophys. J.* **1984**, *11*, 103.
[4] N. Popovych, S. Sun, R. H. Ebright, C. G. Kalodimos, *Nat. Struct. Mol. Biol.* **2006**, *13*, 831.
[5] R. G. Smock, L. M. Gierasch, *Science* **2009**, *324*, 198.
[6] Q. Cui, M. Karplus, *Protein Sci.* **2008**, *17*, 1295.
[7] J. Guo, H.-X. Zhou, *Chem. Rev.* **2016**, *116*, 6503.
[8] R. Nussinov, C. J. Tsai, *Curr. Opin. Struct. Biol.* **2015**, *30*, 17.
[9] C. J. Tsai, R. Nussinov, *PLoS Comput. Biol.* **2014**, *10*, e1003394.
[10] C. J. Tsai, A. del Sol, R. Nussinov, *J. Mol. Biol.* **2008**, *378*, 1.
[11] R. Kalescky, H. Zhou, J. Liu, P. Tao, *PLoS Comput. Biol.* **2016**, *12*, e1004893.
[12] R. D. Malmstrom, A. P. Kornev, S. S. Taylor, R. E. Amaro, *Nat. Commun.* **2015**, *6*, 7588.
[13] A. M. Ruschak, L. E. Kay, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E3454.
[14] P. Weinkam, J. Pons, A. Sali, *Proc. Natl. Acad. Sci. U S A* **2012**, *109*, 4875.
[15] C.-J. Tsai, A. del Sol, R. Nussinov, *Mol. Biosyst.* **2009**, *5*, 207.
[16] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, I. Bahar, *Biophys. J.* **2001**, *80*, 505.

[17] I. Bahar, T. R. Lezon, L.-W. Yang, E. Eyal, *Annu. Rev. Biophys.* **2010**, *39*, 23.
[18] W. Zheng, *Proteins: Struct. Funct. Bioinf.* **2010**, *78*, 638.
[19] H. Zhou, B. D. Zoltowski, P. Tao, *Sci. Rep.* **2017**, *7*, 46626.
[20] M. Donald, D. J. Spiegelhalter, C. C. Taylor, C. John, Machine Learning, Neural and Statistical Classification; Ellis Horwood: Horwood **1994**.
[21] R. Burbidge, M. Trotter, B. Buxton, S. Holden, *Comput. Chem.* **2001**, *26*, 5.
[22] D. E. Goldberg, J. H. Holland, *Mach. Learn.* **1988**, *3*, 95.
[23] H. Nielsen, S. Brunak, G. Von Heijne, *Protein Eng.* **1999**, *12*, 3.
[24] A. C. Tan, D. Gilbert, *Appl. Bioinf.* **2003**, *2*, 1.
[25] F. Rosenblatt, *Psychol. Rev.* **1958**, *65*, 386.
[26] M. Minsky, S. Papert, *An Introduction to Computational Geometry*; MIT Press: Cambridge, MA, **1969**.
[27] G. Giacinto, F. Roli, *Image Vis. Comput.* **2001**, *19*, 699.
[28] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, *Nature* **2016**, *529*, 484.
[29] J. Gasteiger, J. Zupan, *Angew. Chem. Int. Ed.* **1993**, *32*, 503.
[30] J. Gasteiger, J. Zupan, *VCH Verlagsgesellschaft*; Weinheim/VCH Publishers: New York, **1993**, *106*, 1367.
[31] M. Puri, Y. Pathak, V. K. Sutariya, S. Tipparaju, W. Moreno, Artificial Neural Network for Drug Design, Delivery and Disposition, Academic Press: San Diego, CA, **2015**.
[32] G. E. Dahl, N. Jaitly, R. Salakhutdinov, *Multi-task Neural Networks for QSAR Predictions*, ArXiv e-prints, **2014**, 1406, arXiv:1406.1231.
[33] J. C. Dearden, P. H. Rowe, *Artif. Neural Netw.* **2015**, *65*, 1260.
[34] S. Grossberg, *Neural Netw.* **1988**, *1*, 17.
[35] S. R. Safavian, D. Landgrebe, *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660.
[36] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.
[37] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926.
[38] J. Ryckaert, G. Ciccotti, H. J. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327.
[39] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1993**, *98*, 10089.
[40] B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comput. Chem.* **2009**, *30*, 1545.
[41] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, *J. Phys. Chem. B* **1998**, *102*, 3586.
[42] H. B. Demuth, M. H. Beale, O. De Jess, M. T. Hagan, *Neural Network Design*; Martin Hagan: Oklahoma State University, **2014**.
[43] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[44] M. Y. Park, T. Hastie, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2007**, *69*, 659.
[45] T. Evgeniou, M. Pontil, T. Poggio, *Adv. Comput. Math.* **2000**, *13*, 1.
[46] P. Abbeel, A. Y. Ng, Proceedings of the Twenty-First International Conference on Machine Learning, Alberta, Canada, July 4-8, 2004; ACM: New York, **2004**; p. 78.
[47] W. Tong, H. Hong, H. Fang, Q. Xie, R. Perkins, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525.
[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[49] M. P. Harrigan, M. M. Sultan, C. X. Hernández, B. E. Husic, P. Eastman, C. R. Schwantes, K. A. Beauchamp, R. T. McGibbon, V. S. Pande, *Biophys. J.* **2017**, *112*, 10.
[50] P. H. Hünenberger, A. E. Mark, W. F. van Gunsteren, *J. Mol. Biol.* **1995**, *252*, 492.
[51] J. Eargle, Z. Luthey-Schulten, *Bioinformatics* **2012**, *28*, 3000.
[52] A. Sethi, J. Eargle, A. A. Black, Z. Luthey-Schulten, *Proc. Natl. Acad. Sci. U S A* **2009**, *106*, 6620.
[53] E. J. Fuentes, C. J. Der, A. L. Lee, *J. Mol. Biol.* **2004**, *335*, 1105.
[54] Z. N. Gerek, S. B. Ozkan, *PLoS Comput. Biol.* **2011**, *7*, e1002154.
[55] Y. Kong, M. Karplus, *Proteins: Struct. Funct. Bioinf.* **2009**, *74*, 145.
[56] F. Raimondi, A. Felline, M. Seeber, S. Mariani, F. Fanelli, *J. Chem. Theory Comput.* **2013**, *9*, 2504.
[57] S. Hertig, N. R. Latorraca, R. O. Dror, *PLoS Comput. Biol.* **2016**, *12*, e1004746.
[58] B. A. Kidd, D. Baker, W. E. Thomas, *PLoS Comput. Biol.* **2009**, *5*, e1000484.
[59] H. N. Motlagh, J. O. Wrabl, J. Li, V. J. Hilser, *Nature* **2014**, *508*, 331.
[60] I. G. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*; IOS Press: Amsterdam, Netherlands, **2007**.
[61] J. W. Shavlik, T. G. Dietterich, *Readings in Machine Learning*; Morgan Kaufmann Publishers: Burlington, Massachusetts, **1990**.
[62] C. Taylor, D. Michie, D. Spiegalhalter, Eds. *Machine Learning, Neural and Statistical Classification*; Englewood Cliffs: New York, **1994**.
[63] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[64] M. K. Warmuth, J. Liao, G. Rätsch, M. Mathieson, S. Putta, C. Lemmen, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667.
[65] C. Helma, T. Cramer, S. Kramer, L. De Raedt, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402.
[66] G. Kozlov, D. Banville, K. Gehring, I. Ekiel, *J. Mol. Biol.* **2002**, *320*, 813.

Received: 27 December 2017
Revised: 2 March 2018
Accepted: 11 March 2018
Published online on 31 March 2018