frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 09 July 2019
doi: 10.3389/fmolb.2019.00047

Machine Learning Classification Model for Functional Binding Modes of TEM-1 β-Lactamase

Feng Wang[1], *Li Shen*[1], *Hongyu Zhou*[1], *Shouyi Wang*[2], *Xinlei Wang*[3] *and Peng Tao*[1]*

[1] *Department of Chemistry, Center for Scientific Computation, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, TX, United States,* [2] *Department of Industrial, Manufacturing, and Systems Engineering, University of Texas at Arlington, Arlington, TX, United States,* [3] *Department of Statistical Science, Southern Methodist University, Dallas, TX, United States*

TEM family of enzymes is one of the most commonly encountered β-lactamases groups with different catalytic capabilities against various antibiotics. Despite the studies investigating the catalytic mechanism of TEM β-lactamases, the binding modes of these enzymes against ligands in different functional catalytic states have been largely overlooked. But the binding modes may play a critical role in the function and even the evolution of these proteins. In this work, a newly developed machine learning analysis approach to the recognition of protein dynamics states was applied to compare the binding modes of TEM-1 β-lactamase with regard to penicillin in different catalytic states. While conventional analysis methods, including principal components analysis (PCA), could not differentiate TEM-1 in different binding modes, the application of a machine learning method led to excellent classification models differentiating these states. It was also revealed that both reactant/product states and apo/product states are more differentiable than the apo/reactant states. The feature importance generated by the training procedure of the machine learning model was utilized to evaluate the contribution from residues at active sites and in different secondary structures. Key active site residues, Ser70 and Ser130, play a critical role in differentiating reactant/product states, while other active site residues are more important for differentiating apo/product states. Overall, this study provides new insights into the different dynamical function states of TEM-1 and may open a new venue for β-lactamases functional and evolutional studies in general.

Keywords: TEM-1 β-lactamase, functional binding modes, structural analysis, random forest classification, machine learning, molecular dynamics

OPEN ACCESS

Edited by:
Gennady Verkhivker, Chapman University, United States

Reviewed by:
Elif Ozkirimli, Bogaziçi University, Turkey
Pavel Srb, Academy of Sciences of the Czech Republic (ASCR), Czechia

***Correspondence:**
Peng Tao
ptao@smu.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *11 February 2019*
Accepted: *11 June 2019*
Published: *09 July 2019*

Citation:
Wang F, Shen L, Zhou H, Wang S, Wang X and Tao P (2019) Machine Learning Classification Model for Functional Binding Modes of TEM-1 β-Lactamase. Front. Mol. Biosci. 6:47. doi: 10.3389/fmolb.2019.00047

INTRODUCTION

Antibiotic resistance against almost all the existing antibiotics presents a major risk to global health. Among many other factors, β-lactamases as a group of proteins that hydrolyze antibiotics play a key role in antibiotic resistance. The serine β-lactamases, which utilize a serine residue to hydrolyze the β-lactam ring-based antibiotics, and zinc based β-lactamases, are the two main groups of β-lactamases in general. Class A β-lactamases are one dominant subgroup in serine β-lactamases and are highly diversified. TEM-1, the most commonly encountered β-lactamase in Gram-negative bacteria, belongs to the Class A β-lactamases (Bradford, 2001). The structure and potential catalytic

mechanisms of TEM-1 have been studied extensively as a model system of Class A β-lactamases (Lamotte-Brasseur et al., 1991, 1999; Jelsch et al., 1992; Fonzé et al., 1995; Maveyraud et al., 1998; Petrosino et al., 1998; Minasov et al., 2002; Díaz et al., 2003; Hermann et al., 2003; Golemi-Kotra et al., 2004; Roccatano et al., 2005; Savard and Gagné, 2006; Doucet et al., 2007). The catalytic mechanism of TEM-1 can be divided into acylation and deacylation steps using penicillin as an example. The acylation step leads to an acylenzyme Michaelis-complex intermediate with a covalent bond formed between the Ser70 residue and ring opening product of penicillin β-lactam ring. This covalent bond in the acylenzyme intermediate is further hydrolyzed during the deacylation step, leading to an ineffective β-lactam ring-opening product detached from the enzyme. Catalytic functions of key residues at and surrounding an active site have been investigated extensively with some ongoing controversy (Oefner et al., 1990; Herzberg and Moult, 1991; Lamotte-Brasseur et al., 1991, 1992, 1994; Strynadka et al., 1992, 1996; Matagne et al., 1998). The active site of TEM-1 contains several conserved residues that are important for catalysis: Ser70, Lys73, Lys234, Glu166, and Ser130 (Fisette et al., 2010). Here and in the rest of the article, the sequence numbering of Ambler et al. (1991) is used to be consistent with the general literature about TEM-1 (Savard and Gagné, 2006; Doucet et al., 2007; Fisette et al., 2010). It is also believed that some residues, including Asn170, Ala237, Ser235, and Arg244, help to stabilize the acylenzyme intermediate. Although not fully determined, the contribution of these residues to TEM-1 catalytic mechanisms have been investigated extensively (Zafaralla et al., 1992; Stec et al., 2005; Marciano et al., 2009; Stojanoski et al., 2015; Palzkill, 2018). In addition, an allosteric site consisted of helixes 11 (residue 219–226) and 12 (residues 271–289) of TEM-1 were proposed (Horn and Shoichet, 2004). Two novel inhibitors were reported to destabilize the TEM-1 at high temperature. The two inhibitors can bind to the allosteric site in TEM-1, which locates in between helices 11 and 12. The allosteric site is 16 Å away from the active site. It was proposed that TEM-1 conformational changes were transmitted by a key catalytic residue, Arg244 (Horn and Shoichet, 2004). In another study, the allosteric site of TEM-1 was further detected through binding with a β-lactamase inhibitor protein (BLIP). It was suggested that the connections between active site and allosteric site may be modulated by the helix 10 region (residues 218–230) and Tpr229 in TEM-1 (Meneksedag et al., 2013). The allosteric site helixes 11 and 12 were also proposed as a cryptic pocket formation of TEM-1 (Oleinikovas et al., 2016). In addition, the residues P226-W229-P252 were identified as a PWP triad to stabilize the helix 10 region (Avci et al., 2016, 2018).

One important aspect of TEM-1 for its function is dynamics. Therefore, the molecular dynamics (MD) simulations were carried out to characterize dynamical properties of TEM-1 binding with benzyl penicillin molecule. A so-called Ω loop spans residues 163 through to 180 (including the key Glu166 residue for catalysis), and forms one edge of the active site (Dideberg et al., 1987; Herzberg and Moult, 1987; Moews et al., 1990; Jelsch et al., 1993; Vanwetswinkel et al., 2000). Some earlier MD simulations showed that the Ω loop was rather stable even with the absence of the ligand (Díaz et al., 2003). The whole TEM-1 has also been shown to be unusually rigid with limited motions on the picosecond-to-nanosecond time scale through a nuclear magnetic resonance (NMR) spectroscopy study (Savard and Gagné, 2006). Through more extended simulations and NMR studies, a variety of motions displayed by Ω loop are revealed to be potentially important for catalysis (Fisette et al., 2010). Another simulation study of TEM-1 binding with benzylpenicillin suggested that a substrate binding led to increased flexibility of Ω loop while making TEM-1 globally more rigid (Fisette et al., 2012). In addition to benzylpenicillin as a substrate, simulations were also carried out for TEM-1 bound with another two antibiotics, amoxicillin and ampicillin, to reveal that even the subtle differences in chemical structures of ligands could also regulate the substrate recognition (Pimenta et al., 2013).

One overlooked aspect of TEM-1's function is the binding with antibiotics and their hydrolysis product. Penicillin, for example, could bind with TEM-1 as favorable substrate, while the hydrolysis product of penicillin needs to leave the binding pocket for the turnover of this enzyme. Given the rigidity and sensitivity of the TEM-1 structure to the ligand, the response of protein dynamics to the ligand, in different chemical states through catalysis, could be significant and important for its function, however, this remains under-appreciated. One of the reasons for this is probably due to the fast turnover rate, which does not allow for a reliable experimental probe of the protein binding with ligands during its quick catalytic cycles. MD simulations provide an alternative way to scrutinize the difference between the binding modes of protein with similar ligands. However, due to the rigidity of TEM-1 and the similarity between two ligands of interest, some special analysis tools would be necessary for the purpose of comparison.

Machine learning methods are computational tools that construct data-driven prediction models based on training data. In recent years, machine learning methods have been successfully applied in computational chemistry (Husic and Pande, 2018), including pharmaceutical data analysis (Burbidge et al., 2001), protein–ligand binding affinity prediction (Ballester and Mitchell, 2010; Decherchi et al., 2015) and MD simulations based on machine learning analysis of quantum-mechanical forces (Li et al., 2015; Cortina and Kasson, 2018; Shcherbinin and Veselovsky, 2019). Recently, we have introduced two widely applied machine learning algorithms, a decision tree and an artificial neural network, to build classification models to differentiate two allosteric states of the second PDZ domain (PDZ2) in the human PTP1E protein as a dynamics-driven allosteric protein (Zhou et al., 2018). Despite the lack of a significant conformational change between two states of PDZ2, it was demonstrated that both algorithms could build effective prediction models and provide reliable quantitative evaluation of the contributions from individual residues to overall difference between the two states.

In this study, we applied another machine learning algorithm, random forest, to build models. Random Forest (Breiman, 2001) is a supervised learning algorithm that relies on an ensemble method to create an entire forest of random uncorrelated

decision trees, in order to achieve a more accurate and stable prediction. It has been found to be very useful in a wide scope of applications, due to its superior performance in classification and regression problems, as well as its ease of use and flexibility. The recognition of TEM-1 against ligands in different states is interrogated through simulations studies. The random forest method as an effective machine learning technique has been applied to analyze the simulations of TEM-1 in different binding states and evaluate the contribution from every residue and related secondary structures to the recognition of ligands in different states of TEM-1. Potential key residues could be identified based on their feature importance generated from the machine learning model of the simulation data of TEM-1 in different states. The TEM-1 hydrolysis mechanism is of great interest and has been subjected to extensive computational studies focusing on the TEM-1 active site or nearby residues (Díaz et al., 2001; Meroueh et al., 2005; Roccatano et al., 2005; Sgrignani et al., 2014). However, the potential contribution from protein dynamics in different states to catalysis has been largely overlooked. We hypothesize that TEM-1 in different catalytic states, including binding states with reactant and product, are differentiable and could provide further mechanistic details if subjected to appropriate analyses.

Therefore, the current study focuses on the development of classification models to differentiate dynamics of TEM-1 in different functional states and on obtaining information to correlate protein dynamics with individual residues regardless their positions relate to the active site. The dynamics of different states are compared with each other in the training process, governed by the random forest method. In the random forest method, the contribution from each residue to the overall classification model was measured as importance of features (Zhou et al., 2019). A higher importance value of a feature represented a higher contribution in classifying different functional states. Using the feature importance, important structures and residues identified by this computational study are also in agreement with previous studies of this enzyme. The analysis about active and allosteric sites of TEM-1 also sheds new light on the allosteric component of TEM-1 functions. The remainder of the paper is organized in four parts: computational methods, results, discussion, and conclusion.

COMPUTATIONAL METHODS

Molecular Dynamics (MD) Simulations

Three states of TEM-1 were subject to molecular dynamics (MD) simulations. TEM-1 bound with benzyl penicillin (**Figure 1A**) is referred to as the reactant state; TEM-1 bound with product of hydrolyzing benzyl penicillin (**Figure 1B**) is referred to as the product state, and TEM-1 alone without a ligand is referred to as the apo state. No crystal structure is available for TEM-1 binding with penicillin either as a reactant or product. The complex structure related to TEM-1 catalysis against penicillin with the best quality is an intermediate structure (PDB ID: 1fqg), which has been used for various computational studies. Therefore, this crystal structure was used to generate all three states of TEM-1, based on a hypothesis that equilibrium simulations could lead to sufficient sampling in these functional states. CHARMM molecular simulation program suite, version 40b1, was used to prepare and set up the systems (Halgren, 1992). Hydrogen atoms were added to the crystal structure of TEM-1 bound with benzyl penicillin using the hydrogen position construction facility (HBUILD) of the CHARMM. The benzyl penicillin ligand was removed to create the apo state of TEM-1. The benzyl penicillin structure was also modified using CHARMM internal coordinate editing functions to produce the benzyl penicillin hydrolysis product. CHARMM36 force field was used for TEM-1(Best et al., 2012). The CHARMM General Force Field (CGenFF) was generated for the benzyl penicillin and the benzyl penicillin hydrolysis product using online server ParamChem (https://cgenff.paramchem.org/). All systems are solvated in a water box using a TIP3P model with the addition of sodium and chloride ions to balance the charge and reproduce typical physiological ion concentrations.

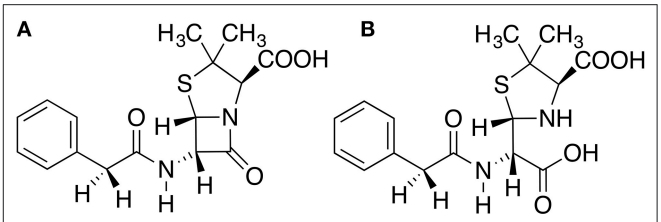

FIGURE 1 | Chemical structures of **(A)** Benzyl penicillin, **(B)** the hydrolysis product of benzyl penicillin.

The simulation boxes were subjected to 5,000 steps of the steepest descent energy minimization and further energy minimization using the adopted basis Newton-Raphson (ABNR) method until the total gradient of the system was lower than 0.02 kcal/mol•Å. Subsequently, the minimized simulation systems were subjected to 24 picoseconds (ps) isothermal-isobaric (NPT) ensemble equilibrium, gradually raising the temperature from 100 to 300 K. The system was then equilibrated via NVT ensemble MD simulations at 300 K. The time step for MD simulations is 2 fs, with all the bonds associated with hydrogen being fixed during the simulation using SHAKE method (Ryckaert et al., 1977). Periodic boundary condition was used in all simulations, and electrostatic interactions were calculated using the particle mesh Ewald method (Darden et al., 1993). For each state, five independent 100 ns NVT ensemble MD simulations were carried out as the production runs after 10 ns of equilibration. OpenMM simulation package was used to carry out the production MD simulations (Friedrichs et al., 2009; Eastman and Pande, 2015; Eastman et al., 2017).

Analysis of MD Simulations

Root-Mean-Square Deviation (RMSD)

RMSD is used to measure the difference in conformation for each snapshot of the MD simulations from 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:

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

Where r_i^0 is the Cartesian coordinate vector of the i^{th} atom in the reference structure. The transformation matrix U is defined as the best-fit alignment between the TEM-1 structure along trajectories with respect to the reference structure.

Root-Mean-Square Fluctuation (RMSF)

RMSF is used to measure the fluctuation of conformation for each frame of the trajectories from the averaged structure.

$$RMSF_i = \left[\frac{1}{T}\sum_{t=1}^{T}\left|r_i(t) - r_i\right|^2\right]^{\frac{1}{2}}, \quad (2)$$

Where T is the time period and $\overline{r_i}$ is the averaged position of atom i over the whole time period.

Principal Component Analysis (PCA)

For each state, PCA was performed by projecting each of the extracted 25,000 frames from five independent trajectories on the principal normal modes. The analysis was carried out using mdtraj package (McGibbon et al., 2015) and scikit-learn library in python (Pedregosa et al., 2011). PCA is a method to reduce the dimensionality of the motion of molecules. It can extract the dominant modes of the motion from a trajectory of molecular dynamic simulation. The normal modes for PCA (Jolliffe, 2011) were obtained through diagonalizing the correlation matrix of the atomic position in one trajectory. The correlation matrix element is calculated by

$$C_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}} = \frac{\langle r_i r_j\rangle - \langle r_i\rangle\langle r_j\rangle}{\sqrt{\left[\left(\langle r_i^2\rangle - \langle r_i\rangle^2\right)\left(\langle r_j^2\rangle - \langle r_j\rangle^2\right)\right]}}, \quad (3)$$

Where C_{ij} is the Pearson correlation coefficient between atoms i and j.

The distributions of three TEM-1 states simulations in the PCA projection space are normalized and plotted as a density contour graph. The distribution density function was estimated by the Gaussian kernels (Scipy 1.2.1) (Turlach, 1993; Bashtannyk and Hyndman, 2001; Scott, 2015; Silverman, 2018).

Random Forest Model

The random forest classification was used in this study to develop classification models for the three states of TEM-1. The python package scikit-learn v0.20.3 was used to carry out the training and testing using this model. For each independent 100 ns simulation of all states, 5,000 frames were evenly extracted as the training and testing data. For each state, four simulations among five production runs were randomly selected as the training set with the remaining simulation used as the testing set. For each selected frame from the simulation, all the pairwise distances among the α carbons (Cα) of TEM-1 backbone are extracted as the features for training purpose. A total of 263 TEM-1 amino acid residues result in 34,453 pairwise distances as the training features. As a pre-step before the classification, the feature selection is carried out using the random forest classification model. Following a previous study to build feature selection using machine learning methods (Zhou et al., 2018), all features are pre-screened to select features accounting for 98.0% as total importance. The apo/product model has 901 features out of the total of 34,453 features. Similarly, after the feature selection, the reactant/product model has 1,170 features, the non-product/product model has 964 features and the apo/reactant model has 1,923 features for their classification models. The final classification models were developed using these preselected features. The number of preselected features for four training models with all preselected features are provided in the **Supplementary Material**.

A random forest algorithm was built on the decision tree models. First, training data was randomly divided into numerous sets and decision tree models were built based on each set. Then all the decision tree models were combined to generate final random forest classification model (Breiman, 2001; Geurts et al., 2006; Louppe, 2014). The random forest algorithm implemented in scikit-learn v0.20.3 (ensemble.RandomForestClassifier) was employed in this study. The number of decision trees generated in the random forest model (referred to as n_estimator) was varied for the best performance with the highest training and validation accuracy (**Supplementary Figure 1**). For each model, the number of decision trees to obtain the highest accuracy of validation was selected for the final classification model.

The random forest method was employed for two purposes in this study, including feature prescreening and classification model developing. In feature prescreening, the feature importance generated from preliminary random forest training process is assigned to each feature. All features are sorted based on their feature importance. The features with the sum of their importance accounting for 98% are selected for the final classification model. These pre-screened features of each classification model present in this study are listed in the **Supplementary Material**. The final classification models were trained using the pre-screened features and with new set of feature importance generated from the training process. The new set of feature importance is used for further analyses presented in this study.

Scores

In this study, the scores including accuracy, precision, recall, and F1 score were used to evaluate the performance of each classification model. The python package v0.20.3 (Pedregosa et al., 2011; Buitinck et al., 2013) was employed to generate these four scores. The accuracy score is defined as

$$accuracy = \frac{1}{N}\sum_{i=0}^{N-1} 1(\hat{y}_i = y_i), \quad (4)$$

where N is the number of samples, $\hat{y}_i$ is the predicted label and y_i is the true label for the i_{th} sample.

In a binary classification task, such as the classification models in this study with two labels, the predictions of the model are evaluated as the following. Positive/negative labels are used to reflect the prediction made by the model. True/false are used to represent whether the predicted labels correspond to the observed labels (real labels). Accordingly, precision, recall and F1 scores are defined as the following.

$$precision = \frac{tp}{tp + fp}, \tag{5}$$

$$recall = \frac{tp}{tp + fn}, \tag{6}$$

$$F1 = 2\frac{precision^{*}recall}{precision + recall}, \tag{7}$$

Term *tp* (true positive) represents the situation that the model gives positive prediction and the observed label is indeed positive. Term *fp* (false positive) represents that the model gives positive prediction, but the observed label is negative. Term *fn* (false negative) represents that the model gives negative prediction, but the observed label is actually positive. F1 score is a weighted mean of the precision and recall.

Feature Importance

The importance of each feature is generated by random forest algorithm based on Gini impurity (Equation 8). A higher importance represents a more important feature in distinguishing different states. The Gini importance implemented in python package scikit-learn v0.20.3 was used in this study and briefly introduced in the Equations (8–12) as the following.

The feature importance was calculated as Gini impurity:

$$Gini\ impurity = \sum_{i=1}^{C} -f_i(1 - f_i), \tag{8}$$

where f_i is the frequency of a label at a node, and C is the number of labels.

In the random forest models, many decision trees are constructed for training purpose. All the predictions from these individual trees are collected to make the final random forest classification model. The importance (n_j) of a node *j* in each decision tree was represented by Gini impurity:

$$n_j = w_j C_j - \sum_{1}^{m} w_{m(j)} C_{m(j)}, \tag{9}$$

where w_j is the weighted number of samples reaching node *j*, C_j is the impurity value of node *j*, and m is the number of child nodes of the tree.

The feature importance of feature *i* on decision tree is calculated as:

$$f_i = \frac{\sum_{1}^{s} n_j}{\sum_{k \in all\ nodes} n_k}, \tag{10}$$

where *s* is the times of node *j* split on feature *i*.

The normalized feature importance in a decision tree is calculated through:

$$norm\, f_i = \frac{f_i}{\sum_{j\ \in all\ features\ in\ a\ tree} f_j}, \tag{11}$$

The final feature importance in random forest classification is calculated as:

$$F_i = \frac{\sum_{j \in all\ trees} norm\, f_i}{N}, \tag{12}$$

where $norm\ f_i$ is the normalized feature importance values of a decision tree, *N* is the total number of trees (Breiman, 2001; Geurts et al., 2006; Pedregosa et al., 2011; Louppe, 2014).

In our classification models, the features are pairwise Cα distances. To evaluate the importance of each amino acid residue, all the feature importance of the pairwise distances relating to each residue are summed up and divided by two to generate the importance of a residue. Then the total importance of 263 residues were accumulated and the importance percentage of each residue could be calculated based on the total importance. The value of importance percentage represents the ability of a residue to differentiate three states. In other words, the importance could help to evaluate the contribution from a residue to differentiate three states in dynamic motions.

RESULTS

TEM-1 Three States Simulations Analysis

The time evolution of the RMSD of TEM-1 in five independent simulation sets in apo, reactant, and product states are plotted in **Figure 2**. All RMSD values were calculated with reference to the TEM-1 crystal structure. The averaged RMSD values are 1.5, 1.3, and 1.1 Å for the apo, reactant, and product states, respectively. The plots suggest that the TEM-1 is rather stable with low RMSD fluctuations in all three states. Among three states, the apo state displays the highest TEM-1 fluctuation, and the product state displays the lowest TEM-1 fluctuation. To address the concern of the simulation convergence, we also calculated the accumulative entropy of TEM-1 in each state along each independent simulation (**Supplementary Figure 2**). All three states display clear convergence tendency in each simulation.

RMSF of individual residues was calculated for each state using all five simulations and plotted in **Figure 3**. In agreement with the RMSD results, TEM-1 in the apo state has the highest fluctuation for most part of the protein (blue dashed line in **Figure 3**). However, TEM-1 in both the reactant and product states also displays higher fluctuation than the apo state in certain part, revealing that the binding with ligands and the type of ligand do exert a subtle impact on protein dynamics.

Then, we carried out PCA using all 15 simulations from three states as an attempt to develop a model differentiating three states of TEM-1. The simulations of each state are projected onto the surface as contour plots with normalization using the first principal component (PC1) and second principal component (PC2) (**Figure 4**). Overall, all three states largely overlap with

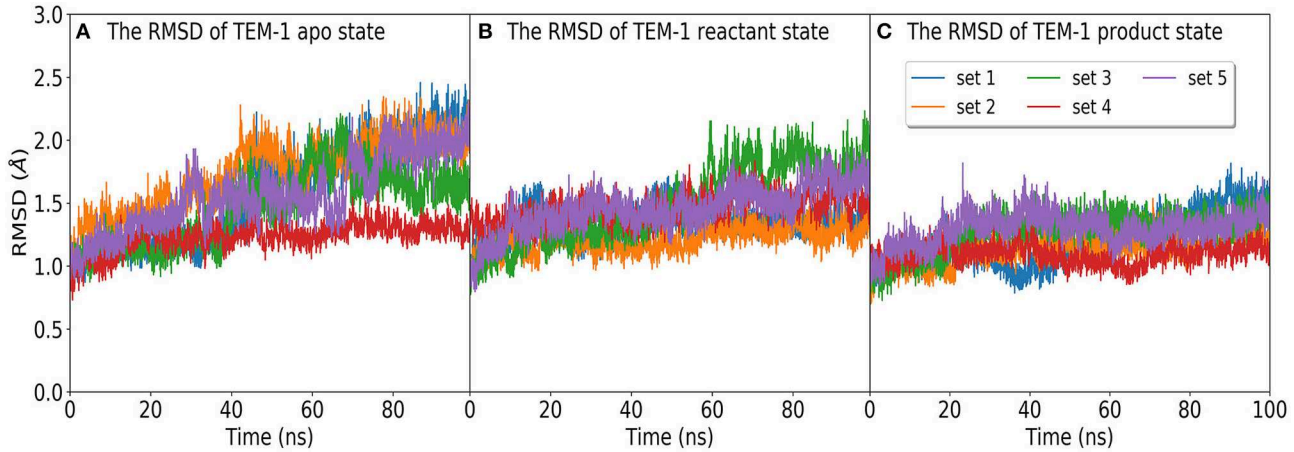

FIGURE 2 | The RMSD distribution of molecular dynamics simulations of TEM-1 in **(A)** apo state, and binding with benzyl penicillin in **(B)** reactant and **(C)** product states. In each state, the RMSD are calculated in five independent 100 ns simulations labeled as set 1 to set 5.

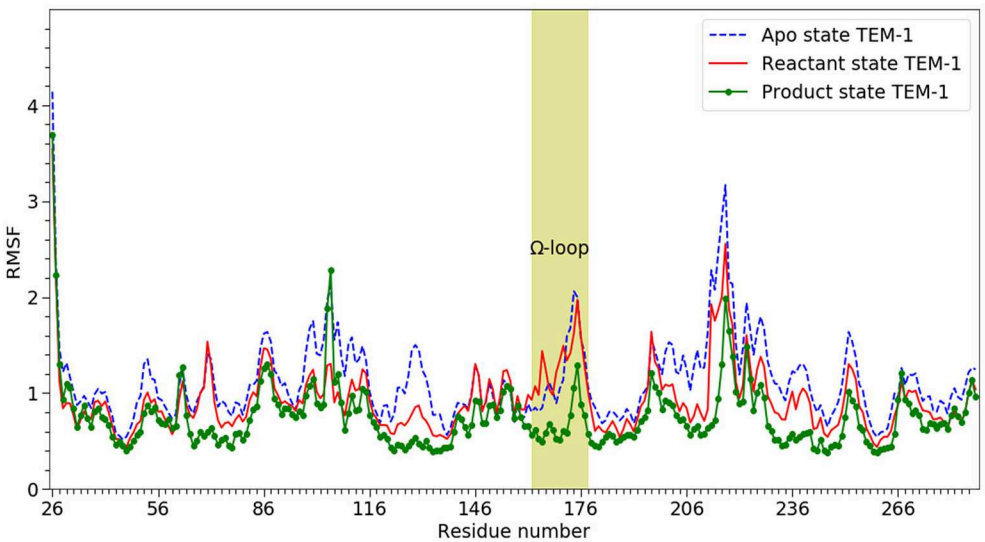

FIGURE 3 | The RMSF of α-carbons (Cα) from 26 to 288 on TEM-1 β-lactamase in apo (blue dash line), reactant (red line) and product states (dot line), Ω loop (residue 163–180) highlighted. All three states have overall similar distribution but with significant difference. The product shows the lowest overall RMSF. The apo state show the highest overall RMSF.

each other on the PC1/PC2 surface, and each state has two or three minima, which are referred to as attraction basins. The reactant and product states cover similar areas and largely overlap with each other, with their attraction basins close to each other. The apo state has different attraction basins and has much narrower distribution than the other two states. The PCA results reflect that the TEM-1 structure is generally rigid without significant global conformational change. However, the subtle differences among the distributions of TEM-1 in different states in the PCA space do indicate the shift in population of TEM-1 in different binding states. The following analysis using the random forest model provides more insight into these subtle differences.

Random Forest Model

The training and testing results of the random forest model for all three states, including accuracy, precision, recall, and F1 scores, are plotted in **Figure 5**. Classification models were developed to differentiate between apo and product states, reactant and product states, non-product (combining the apo and reactant states) and product states, as well as between apo and reactant states. For the classification model to differentiate the reactant and product states, the training with cross-validation provides high performance, and testing provides better than 87% accuracy in all categories (**Figure 5A**), suggesting that the TEM-1 reactant and product states are highly differentiable using the Cα pairwise distances as protein structural information. Slightly better scores

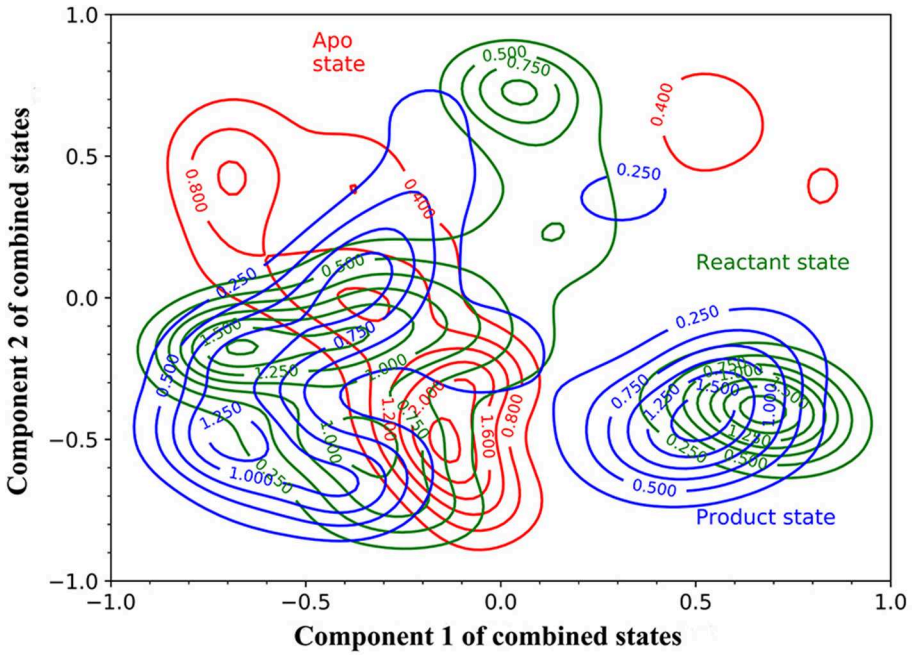

FIGURE 4 | The projection of the simulations of TEM-1 in apo (red), reactant (green) and, product (blue) states onto Component 1 and Component 2 of combined states. Components 1 and 2 are the first and second components from the principal component analysis (PCA) based on the simulations of all three states. The projection on to components 1 and 2 are normalized.

are obtained for the classification model to differentiate the apo and product states (**Figure 5B**). These results show that the TEM-1 in the product state is clearly distinguishable from TEM-1 in the apo and reactant states. However, distinguishability between the apo and reactant states of TEM-1 is significantly lower than the first two pairs (**Figure 5C**), suggesting that these two states share significant similarity in terms of protein backbone structural distributions represented as Cα pairwise distances. To further test this, both apo and reactant states are combined together to be considered as non-product state vs. product state. A classification model differentiating the non-product and the product states is built with cross-validation performance measures close to 100% and testing performance measures ranging between 82 and 99% (**Figure 5D**), similar to the models for apo/product and reactant/product pairs.

As part of preliminary study, two other widely applied machine learning methods, artificial neural network and support vector machine methods, were also applied to develop classification models for TEM-1. Both methods produced models with performance worse than random forest model (**Supplementary Figures 3, 4**). In addition, the random forest method provides importance numerical value for each feature, which could be used to search for key residues and functional groups in protein structure. Therefore, the remainder of the study focuses on random forest model result.

Secondary Structures Contribution

In random forest classification models, each Cα pair is given an importance value reflecting its contribution for the classification model. These values could be used to evaluate, to some extent, the importance of individual amino acid residues. We first used these values to evaluate the contribution of secondary structures in TEM-1, with regard to the differences among different states. For each secondary structure, all the importance values associated with residues in that structure are summed together and divided by two as the overall importance. Three well performing classification models, apo/product, reactant/product, and non-product/product, are used for this comparison purpose. The TEM-1 structure is divided into β-sheets, α-helices, coils and turns as secondary structures and the residues inclusive in these structures. The β-sheet and α-helices of TEM-1 are defined in a previous study (Savard and Gagné, 2006), and are commonly used in general literatures of TEM-1 (Simm et al., 2007; Fisette et al., 2010, 2012). The definition of coils and turns in the database of secondary structure assignments (DSSP) are used in this study (Kabsch and Sander, 1983). There are some coils and turns with just one or two residues. Some of them have small importance values. For simplification, when such a short coil or turn is adjacent to another coil or turn, they are combined as a new coil or turn structure for analysis. However, if a short coil or turn is between β-sheets or α-helices, it was kept by itself.

We further calculate the importance of individual secondary structures and plot it in **Figure 6**. All five β-sheets in TEM-1 have importance values lower than 5% (**Figure 6A**), indicating that the β sheets may not play an important role, with regard to ligand binding. There are 11 helices with varying lengths in TEM-1. Most helices have low importance (**Figure 6B**). The only exception is helix (69–85), which has overall importance close

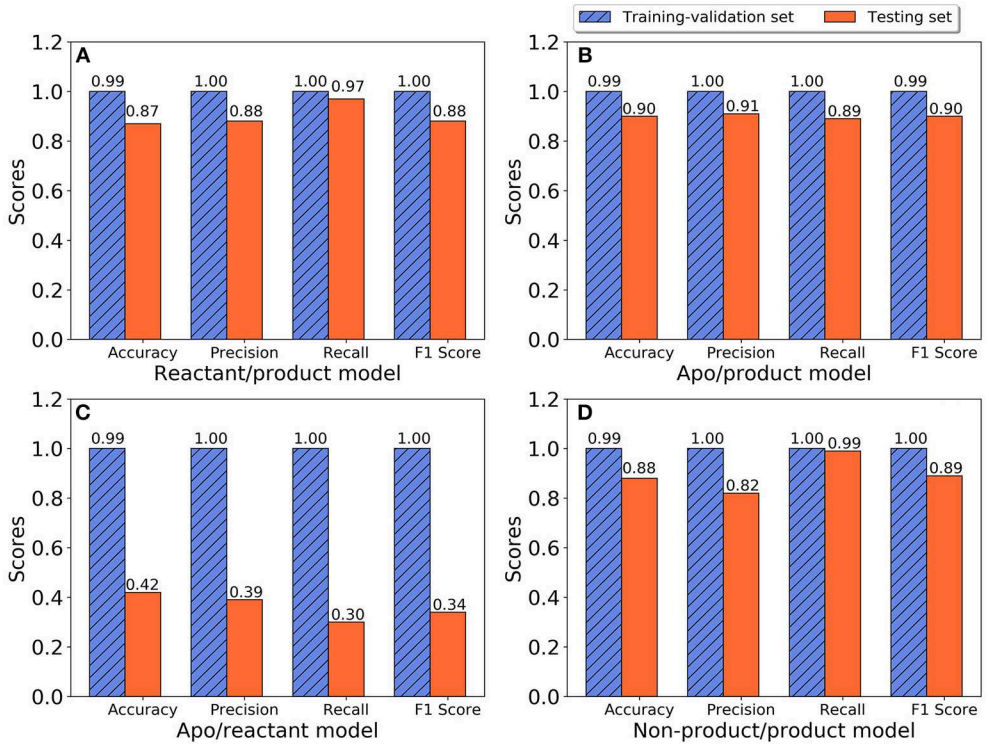

FIGURE 5 | The performance of random forest classification models in accuracy, precision, recall, and F1 scores for training-validation set (blue shadow) and testing set (red). **(A)** Reactant and product states model; **(B)** Apo and product states model; **(C)** Apo and reactant states model; **(D)** Non-product and product states model.

to 16% in the reactant/product model (**Figure 6B**), and also one of the helices around the active site of TEM-1 (**Figure 7** green transparent surface).

There are 10 short fragments being considered as random coils in TEM-1. Among this, residues 213–215 coil shows the highest importance in all three models (**Figure 6C**), which is illustrated and highlighted as cyan structure in **Figure 7**. The second important coil is residues 129–131, with three residues accounting for more than 8% importance in the non-product/product model and around 5% in the other two models. Both 213–215 and 129–131 (highlighted as red structure in **Figure 7**) coils are adjacent to the active site.

There is a total of 15 turn structures in TEM-1, some with significant difference among three classification models. The importance of the residue 216–220 turn (highlighted as yellow structure in **Figure 7**) is the highest on average among all turn structures, followed by residues 102–108 turn (highlighted as green structure in **Figure 7**). Both turns are positioned as gate to cap the TEM-1 active site.

For a better understanding of each residue, the mapping of importance percentage of each residue in TEM-1 obtained from the machine learning training process is plotted in **Figure 8** (divided into three parts A, B, and C). The serial numbers of residues from the PDB file that start from 26 to 111 are used in **Figure 8A**, from 112 to 198 are used in **Figure 8B** and from 199 to the end 288 are used in **Figure 8C**. The overall distributions of TEM-1 individual residue importance based on different classification models resemble each other. Residue 213 has the highest percentage (9.3%) in the apo/product model (**Figure 8C**), which is also the highest percentage for a single residue among all three models. In reactant/product model, residue 70 has the highest percentage as 8.4% (**Figure 8A**). In all three models, residues 67–73, 103–107, 127–135, 162–171, 176–182, and 210–220 have relative high importance percentages in all three models. Interestingly, these residue regions were proposed to undergo conformational changes in a previous NMR study (Savard and Gagné, 2006).

For each model, the top 10 residues with the highest percentages are listed and illustrated with the TEM-1 structure in **Figures 9A–F**. Most of the key residues identified through the classification model are not on either helices or strands secondary structures. However, few active site residues are among the top 10 residues (illustrated in green in **Figures 9D–F**). The percentages of active site residues are significantly different, which is plotted for all three models (**Supplementary Figure 5**). Ser70 from the TEM-1 active site has significantly high importance in the reactant/product model. Ser70 in the other two models, and all other active site residues, only display importance lower than 3%. These are in the agreement that the TEM-1 active site is generally rigid for the purpose of catalysis.

We further investigate the distribution of residues importance with reference to the active site. The importance of residues lying within a certain distance range (i.e., between 4 and 5 Å) from the active site residues are accumulated and

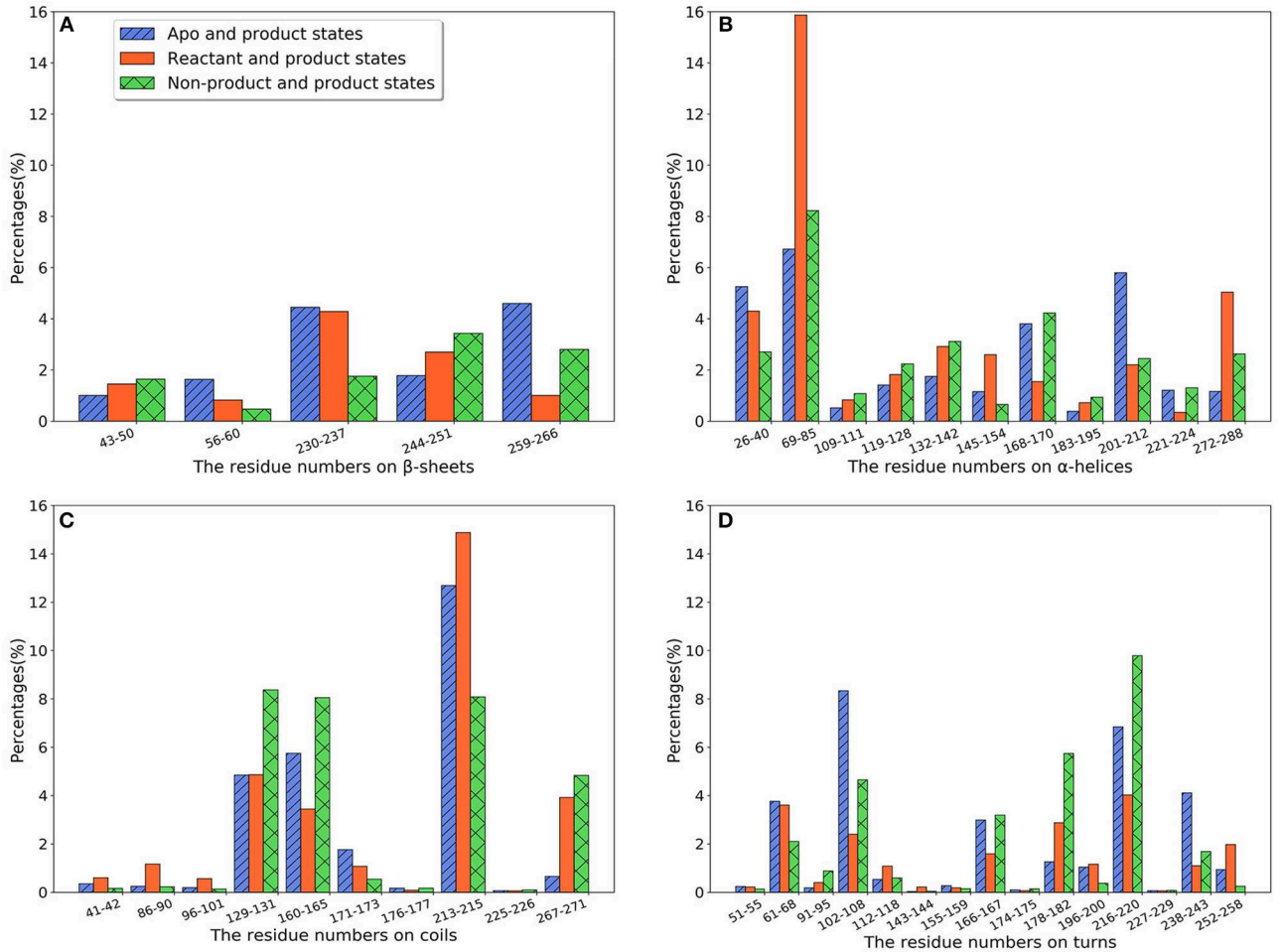

FIGURE 6 | The total feature importance of individual secondary structure in TEM-1: **(A)** β-sheets; **(B)** α-helix; **(C)** Random coils; **(D)** Turns. Each secondary structure is labeled by residue number. There are five β-sheets and 11 α-helices with varying sizes in TEM-1 structure.

normalized by the number of residues within a distance range, which is shown in **Figure 10A**. There are clearly three peaks of importance for the shells around 4, 7, and 10 Å away from the active site. The sums of importance of residues away from the active site region in the three models are plotted in **Figure 10B**. The accumulative importance of residues surrounding the active site is smoothly increasing along the distance.

The Conformational Analysis

In three states classifications, the key residues are identified based on the feature importance obtained from the classification models. However, the conformational changes in three states are very important for detecting the catalytic mechanism of TEM-1 bound with penicillin G complex. Therefore, further conformational analysis is carried out based on the selected key residues with top feature importance. Among the top 10 residues based on their accumulative feature importance, Tyr105 as a gatekeeper of the active site could stabilize the ligand binding (Doucet and Pelletier, 2007; Doucet et al., 2007). However, the interaction between Asn132 and Tyr105 may perturb the stabilizations (Wang et al., 2002). And a mutant of Asn105 has been proposed to create disruptive steric clashes with Asn132 and destabilize the ligand binding (Doucet and Pelletier, 2007). Asn132 is also a special residue, which was proposed to provide additional space for active site (Swarén et al., 1998). Therefore, the distance between Cα atoms of Tyr105 and Asn132 was selected for further analysis to reveal detailed conformational change relevant to functional states. In addition, the interaction between Lys73 and Asn132 was reported as important residues for TEM-1's catalytic function (Swarén et al., 1998). Accordingly, the Cα atoms distance between Lys73 and Asn132 is subjected to further analysis in this study. Two residues Gln39 and Thr269 among the top 10 residues are distal from the active site. Thr269 is really close to the allosteric site Helices 12 (Residue 272–288) identified in previous study (Horn and Shoichet, 2004). To reveal potential correlation between the active site and Gln39 as well as Thr269, the Cα atoms distance from Ser70 as the center of active site to these two residues are also subjected to further analysis.

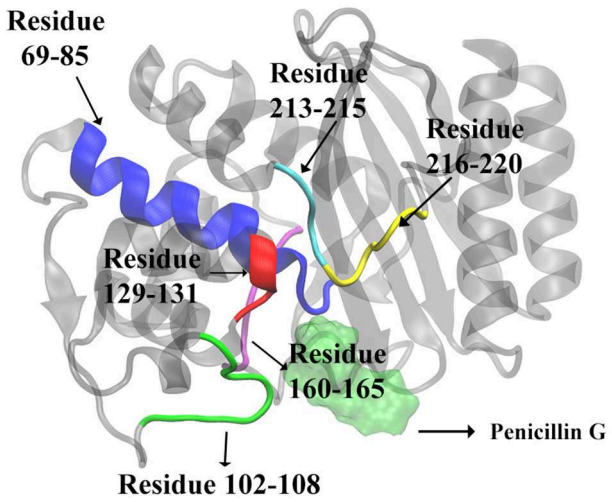

FIGURE 7 | The secondary structures of TEM-1 with significant total feature importance. Residues 69 through 85 as α-helix (blue), residues 213 through 215 as random coil (cyan); residues 216 through 220 as turn (yellow); Residues 129 through 131 as random coil (red), residues 102 through 108 as turn (green), residues 160 through 165 as random coil (purple). The ligand penicillin-G molecule is also illustrated as green transparent surface. The residue 160–165 is behind of the residue 69–85 in this view.

The density distributions of Cα atom distances of Tyr105-Asn132, Lys73-Asn132, Ser70-Gln39, Ser70-Thr269, and residue pairs for all three TEM-1 states are plotted in **Figure 11**. The Cα atom distance distribution of Tyr105-Asn132 has only one main peak close to 6 Å for reactant state (**Figure 11A**). However, the conversion from reactant to product leads to a second peak between 8 and 9 Å. Interestingly, the apo state without a ligand shows a similar distance distribution to the product state of this pair with two peaks between 6–7 Å and 8–9 Å. The density distribution of Lys73-Asn132 Cα atom distance has two peaks in the reactant state, one close to 9 Å and one between 10 and 11 Å (**Figure 11B**). The conversion to the product leads to only one peak around 9.2 Å of this distribution. In apo state, this distribution has a peak around 9.3 Å and a small shoulder about 10.3 Å.

For Ser70-Gln39 pair, the distributions of their Cα atom distance in all three states have only one peak (**Figure 11C**), which are located at 23.8, 24, and 24.5 Å for the apo, reactant and product states, respectively. Similarly, the density distributions of Ser70-Thr269 Cα atom distance also have only one peak for all three states, all between 19 and 20 Å (**Figure 11D**). These analyses demonstrated that the key residues with high feature importance do behave significantly in different functional states of protein. The residues Lys73, Asn132, Gln39, Ser70, and Thr269 are illustrated in the TEM-1 apo, reactant and product aligned structures with green transparent surface representing the ligand penicillin G binding pocket (**Figure 12**).

We further investigated four groups including Ω loop (residues 163–180), residues 213–220 including a turn and random coil structure and residues 102–108 as a turn structure, which are related to structures with high importance percentages illustrated in **Figure 7**. The helix 12 (residues 272–288) with high importance (>5%) in reactant/product model is also included. To reveal a potentially significant conformational change of these groups, the RMSD of these groups with the TEM-1 (1fqg) crystal structure as a reference are calculated and plotted in **Figure 13**. In TEM-1 bound with inhibitors, helix 11 (residues 219–226) and helix 12 (residues 272–288) were identified as an allosteric site (Horn and Shoichet, 2004). In the classification models generated in this study, helix 11 has a low feature importance and residues 213–220 have high importance. The RMSD distributions of residues 213–220 and helix 12 as potential allosteric sites are plotted in **Figures 13B,C**. The RMSD of residues 102–108 as a turn structure containing key residue Tyr105 is plotted in **Figure 13D**. The positions of the four residues group in TEM-1 are also illustrated in **Figure 12**. Interestingly, although Ω loop has high importance percentage, the RMSD distributions of Ω loop in three states are similar with each other displaying one main peak around 0.7 Å (**Figure 13A**). It indicated that Ω loop is not very flexible, agreeing with some NMR studies (Roccatano et al., 2005; Bös and Pleiss, 2009; Fisette et al., 2010). On the contrary, the RMSD distributions of 213–220 turn are significantly different among three states. In the reactant state, there are two main peaks around 1.2 and 2 Å and one small peak around 2.5 Å. In the product state, the RMSD distribution shift toward lower values with three peaks around 0.8, 1.3, and 2.5 Å. In the apo state, there is a dominant peak around 1.3 Å with a smaller peak around 2.6 Å. This clearly revealed significant conformational changes of this turn structure. The RMSD densities of helix 12 (residues 272–288) are similar in all three states with only one peak around 0.4 Å (**Figure 13C**), suggesting little conformational change of this secondary structure. The RMSD densities of residues 102–108 turn have one dominant distribution in three states (**Figure 13D**). The reactant and product states have the peak smaller than 0.4 Å. The apo state has the peak larger than 0.4 Å. These analyses demonstrate that the conformational change may play important role only in a limited local structure to differentiate functional states.

DISCUSSIONS

The role of protein dynamics in catalysis is becoming essential in understanding enzyme's catalytic mechanisms. TEM-1 is one of the proteins that has been interrogated for the correlation between dynamics and functions both experimentally and computationally (Farmer et al., 2017; Modi and Ozkan, 2018). In a detailed study of TEM-1 using NMR, the backbone motion of several TEM-1 mutants at Tyr105 was characterized and linked to its enzymatic function, because the residue in TEM-1 plays a key role in substrate differentiation and stabilization (Doucet et al., 2007). Coincidently, Tyr105 is identified as the second most important residue to differentiate the apo and product states in the current study (**Figure 9A**). The NMR study of TEM-1 also revealed that the mutations at residue 105 led to the change

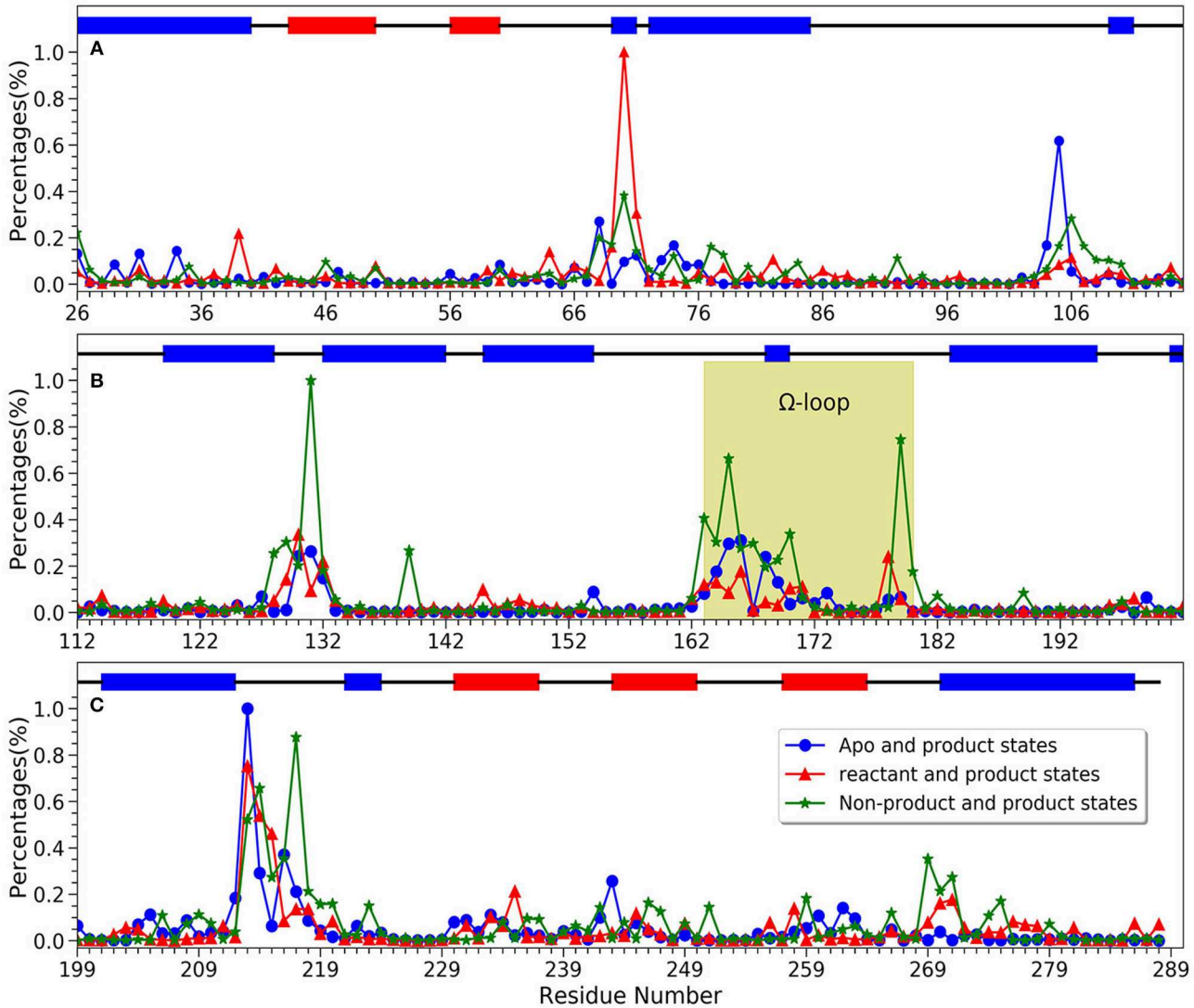

FIGURE 8 | The accumulative feature importance of each residue in TEM-1. The blue circles represent the apo and product states classification model, the red triangles represent the reactant and product states classification model, and the green stars represent the non-product and product states classification model. On the top of each sub-figure, the β-sheets and α-helices are labeled as red and blue rectangles, respectively. Ω-loop is highlighted in yellow. **(A)** Residues 26 through 115; **(B)** Residues 112 through 202; **(C)** Residues 199 through 288.

of backbone motion exceeding the TEM-1 active site and with a wide range of motion time scales. Interestingly, many key residues discovered in this study to be important for TEM-1 dynamical functional states are in a good agreement with the comprehensive NMR study.

The comparison among NMR spectroscopy of TEM-1 mutants showed that the most significant effect on backbone amide motion, marked as chemical shift differences, occur in the residues in 68–80, 100–115, 120–140, 163–170, 213–218, and 235–246 regions (Doucet et al., 2007). All these regions have significant feature importance from all classification modes developed in the current study (**Figure 8**). In general, the chemical shift differences observed in NMR spectroscopy have no direct connections with protein dynamics. But the backbone amide chemical shift is sensitive to the hydrogen bonding interactions in protein (Paramasivam et al., 2018). In another study, it was proposed that TEM-1 with mutant Tyr105 displayed effects on the backbone amide chemical shift of wild-type TEM-1 and can reduce the catalytic efficiency of TEM-1 binding with benzyl penicillin complex (Doucet et al., 2004). Although the backbone amide chemical shift difference is caused by the Tyr105 mutation of TEM-1 in the reference, there is indeed a relationship between the chemical shift difference and the catalytic efficiency for TEM-1 with benzyl penicillin complex. Therefore, the correlation between feature importance of key residues with the backbone amide chemical shift differences may help us to further understand the meaning of the machine learning based classification model. It is possible that the

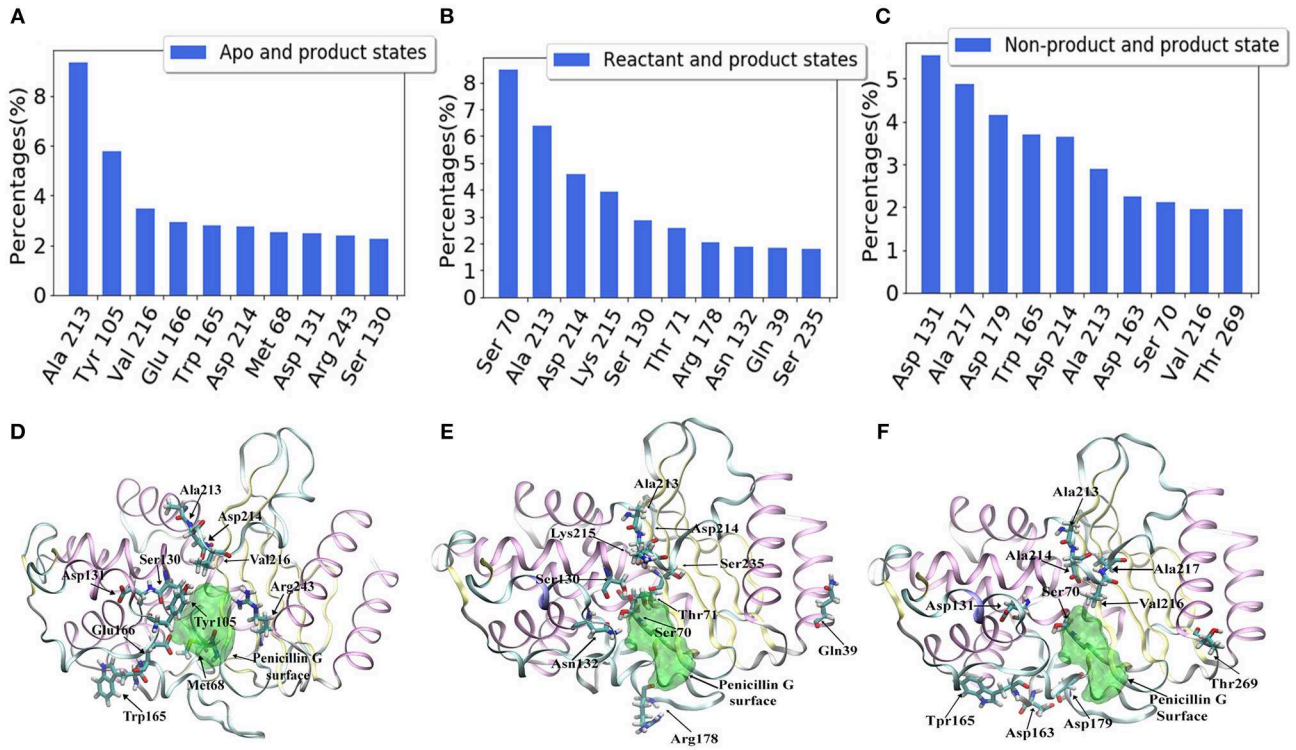

FIGURE 9 | Top 10 residues with the highest feature importance in TEM-1 β-lactamase based on classification models: **(A)** Apo/product model, **(B)** Reactant/product model, **(C)** Non-product/product model; The structures of top 10 residues with the highest feature importance and their positions comparing with the active site of TEM-1 in each model: **(D)** apo/product model, **(E)** reactant/product model, and **(F)** non-product/product model. The ligand penicillin-G molecule is also illustrated as green transparent surface. The active site of TEM-1 is the pocket holding the penicillin-G molecule.

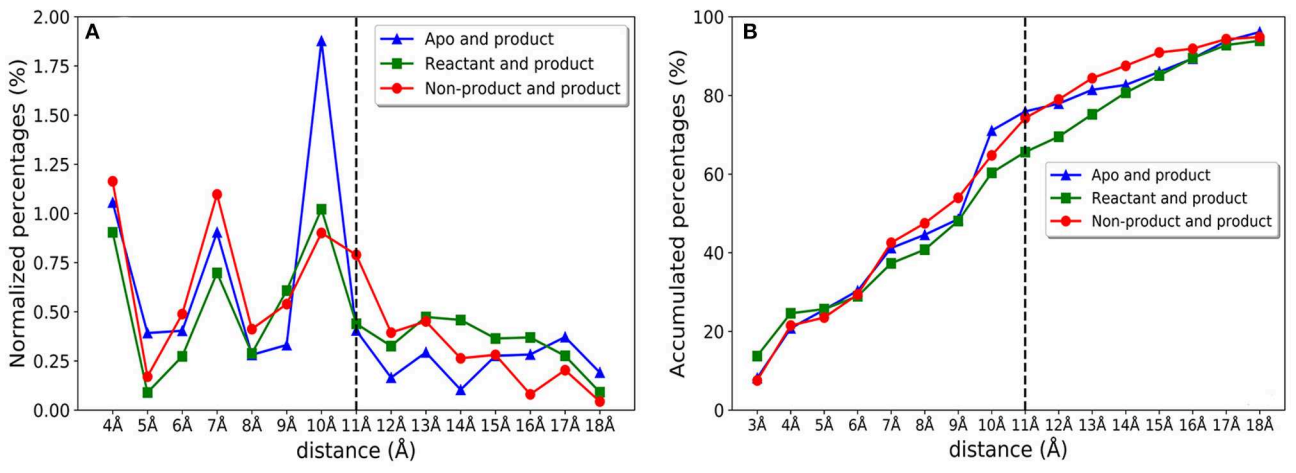

FIGURE 10 | The feature importance of residues with reference to the distance from active site for apo/product, reactant/product, and non-product/product classification models. **(A)** The normalized feature importance of residues within certain distance from active site (using 1 Å window). For example, the importance percentage 0.25% in 5 Å represents the importance percentage of all residues located in distance range 4–5 Å away from the active site; **(B)** Accumulative feature importance of residues with a certain distance from the active site. For example, the importance percentage of 5 Å represents the importance percentage including all the residues within 5 Å away from the active site.

backbone amide motion indicated by the NMR spectroscopy is well-coupled with the backbone Cα motion, which is used to construct features for the machine learning training models in this study. Further comparison also shows remarkable agreement at the individual residue level. Some conserved residues and residues at the so-called active site wall showed significant NMR

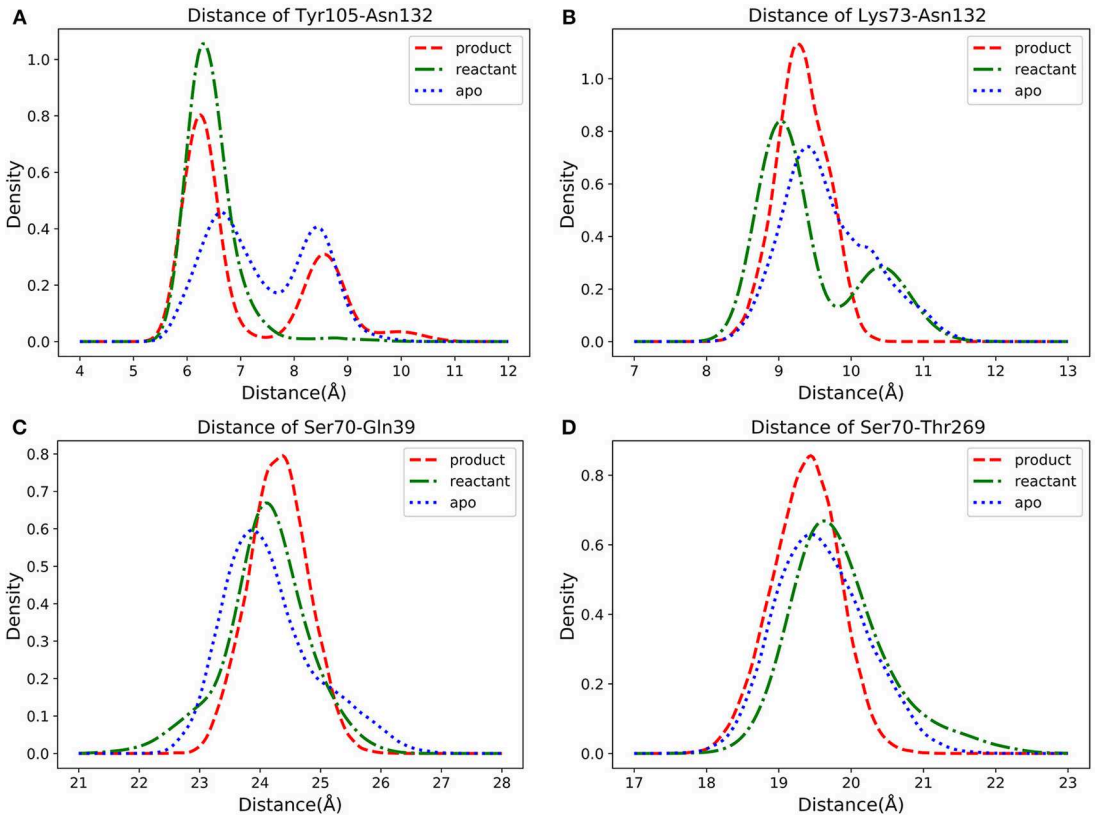

FIGURE 11 | The density distributions of pairwise α carbon atoms distance in apo (blue dot line), reactant (green dot dashed line) and product (red dashed line) states: **(A)** Tyr105 and Asn132, **(B)** Lys73 and Asn132, **(C)** Ser70 and Gln39, **(D)** Ser70 and Thr269.

TABLE 1 | The key residues from current study and a NMR study.

Residues with high feature importance[a]	Adjacent key NMR residues[b]
Met68, Ser70	Thr71
Ser130, Asp131	Met129 Asn132
Asp163, Trp165, Glu166	Arg164, Glu168
Arg178, Asp179	Thr181
Ala213, Asp214, Ala217	Lys215, Val216, Gly218
Ser235	Lys234
Thr269	Met270

[a] *Current study*
[b] *Table 4 in a NMR study (Doucet et al., 2007).*

relaxation parameter changes between the wild type and the most significantly different Y105D mutant (Doucet et al., 2007). Six residues (Asn132, Tyr105, Lys215, Val216, Thr71, and Arg243) among the 21 residues with the highest important features from the classification modes in this study (**Figure 9**) are among the key residues for the local dynamic effects identified in the NMR study. Many more residues (a total of 14) selected by the feature importance are also in the adjacent region within the key residues selected in the NMR study (**Table 1**).

Comparison between the NMR spectroscopy between wild type and Y105D mutant also revealed that significant local differences in the regions of residues 70–80, 124–135, and most importantly in 211–221. Our analysis shows that these regions display high accumulative feature importance as various secondary structures, such as residues 70–80 belonging to α-helix, residues 124–135 spreading across random coil and α-helix, and residues 211–221 containing both random coil and turn structures (**Figure 6**).

Ω loop (163–180) is a key secondary structure close to the ligand binding site in TEM-1 and important for its catalytic function. A previous NMR and MD simulation work showed that Ω loop displayed limited flexibility with the key translational component (Bös and Pleiss, 2009) It was proposed that the Ω loop is a key structural feature for substrate binding and recognition (Fisette et al., 2012). It was observed in the same study that the inter-Ω loop salt bridge between Arg164 and Asp179 is prone to be affected by the substrate binding, while the Arg164-Thr71 interaction is stabilized by the ligand binding. Accordingly, the Ω loop shows significant and various importance in our three classification models, with the most significance in the non-product/product model. Residues Asp163, Arg164, Trp165, and Asp179 are very important residues (>3% in **Figure 8B** Ω loop green highlighted part) for the non-product/product differentiation model. Residues Trp165, Glu166, and Glu168 are important residues (>2% in **Figure 8B** Ω loop green highlighted

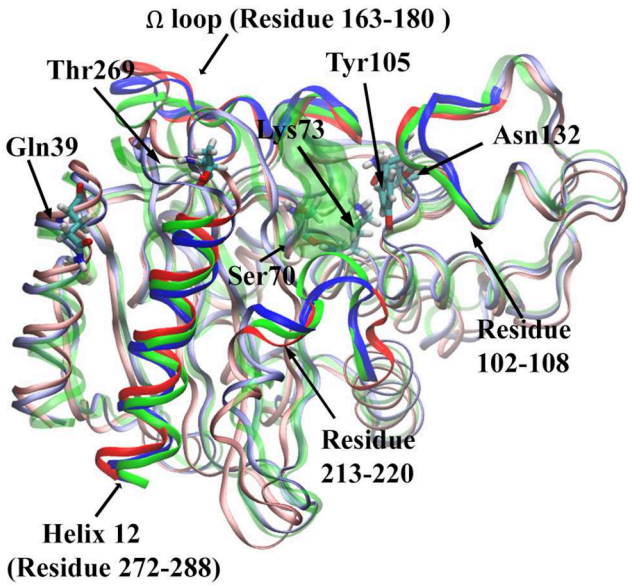

FIGURE 12 | The structure of TEM-1 in apo (blue), reactant (green) and product (red) states. Ω loop (residues 163–180), helix 12 (residues 272–288), residues 102–108, and residues 213–220 are highlighted in each state with same colors. Also key residues Gln39, Thr269, Ser70, Lys73, Tyr105, and Asn132 are labeled. The ligand penicillin-G molecule is represented as green transparent surface.

part) for the apo/product differentiation model. In comparison, the Ω-loop is somewhat less important in the reactant/product model than in the other two models, indicating the importance of differentiating the product from other states. In the non-product/product model, both Arg164 (close to 0.3% percentages of importance) and Asp179 (close to 0.8% percentages of importance) are emphasized as important residues. The Asp179 and Arg164 locate at the entrance of the active site and form the inter-Ω loop salt-bridge to stabilize the loop. In reactant/product and apo/product models, the importance of Arg164 and Asp179 are not obvious, the combination of apo and reactant magnify their importance in non-product/product model. We hypothesized that the interaction between Arg164 and Asp179 exist in all three states to stabilize the loop. Both hydrolyzed benzyl penicillin and benzyl penicillin molecules as substrates can strengthen the interaction. That may be the reason why the overall Ω loop does not carry high importance percentage in reactant/product model. The overall Ω loop is more stable in reactant and product states than in the apo state. In addition, Trp165 is highlighted in both non-product/product and apo/product models, which indicates that Tpr165 is a key residue to classify the apo/reactant and product states. Therefore, it is likely that Tpr165 plays an important role in de-acylation step of the catalytic mechanism, which is also mentioned in experimental study (Petrosino et al., 1998). Another key residue for acylation, Glu166, has a relative high importance in apo/product model. We proposed that Glu166 is not only as a general base in acylation (Minasov et al., 2002) but also very important in the de-acylation step. These detailed comparisons with experimental study provided further insight into the functions of the Ω loop of ligand binding in addition to enzyme catalysis.

The NMR study suggested the key Ω loop motion was in the microsecond (μs)-millisecond (ms) time scale, which was beyond the current simulation study. However, it was also pointed out that the Ω loop dynamics is more focused and less random than other secondary structures even at a large time scale. The good agreement and complimentary comparison between the classification models developed in this study and previous NMR studies of TEM-1 suggests the effectiveness of the machine learning method in the application of protein dynamics and functional analysis. The usage of Cα distance as training features from extensive MD simulations for training practically bridges among protein dynamics with inter-residue correlation, regardless the distance region within the framework of different functional states.

Asn132 was identified as a residue controlling the size of the TEM-1 active site cavity. Distance distribution analysis of Lys73 and Asn132 reveals that the binding with reactant effectively compresses the active site into a closed active site and creates a minor open state representing by two peaks of Cα distance distribution in reactant state (**Figure 11B**). However, the product binding state only has one main peak as a closed active site without a minor open state. This could be a key dynamical difference between reactant and product binding states. The interaction between Tyr105 and Asn132 also related to the active site. Opposite to the Lys73 and Asn132 Cα distance distribution, the Cα distance distribution of Tyr105 and Asn132 changes from single dominant peak in reactant state to double peaks in the product state (**Figure 11A**). The difference of the apo state distribution from both reactant and product states also sheds light on these TEM-1 functional states. Helix 11 (residues 219–226) and 12 (residues 272–288) were proposed as an allosteric site with 3–7 Å shift in helix 11 and 1–3 Å shift in helix 12 comparing to the apo structure (Horn and Shoichet, 2004; Avci et al., 2018). The significant conformational change of residues 213–220 as a turn and random coil structure adjacent to helix 11 could be coupled with the allosteric function residing in this region. The similarity of the helix 12 RMSD distributions shared by all three states warrants further study to elucidate the allosteric mechanism associated with this local structure (**Figures 13B,C**).

It could be a concern that the initial structures for apo, reactant and product state, generated from the same crystal structure (1FQG) in catalytic intermediate state, may not present three target states well. To address this concern, we collected a total of eight crystal structures of wild type TEM-1 in apo states and five crystal structures of wild type TEM-1 binding with various ligands from PDB, including the one with penicillin used as starting structure in this study, as reference structures for the simulations. The averaged RMSDs of each functional state simulations with reference to these crystal structures were calculated and plotted in **Supplementary Figure 6**. It is interesting that the product state simulations consistently have lower RMSDs with reference to all 13 crystal structures, including both apo and holo states of TEM-1 and the structure used in this

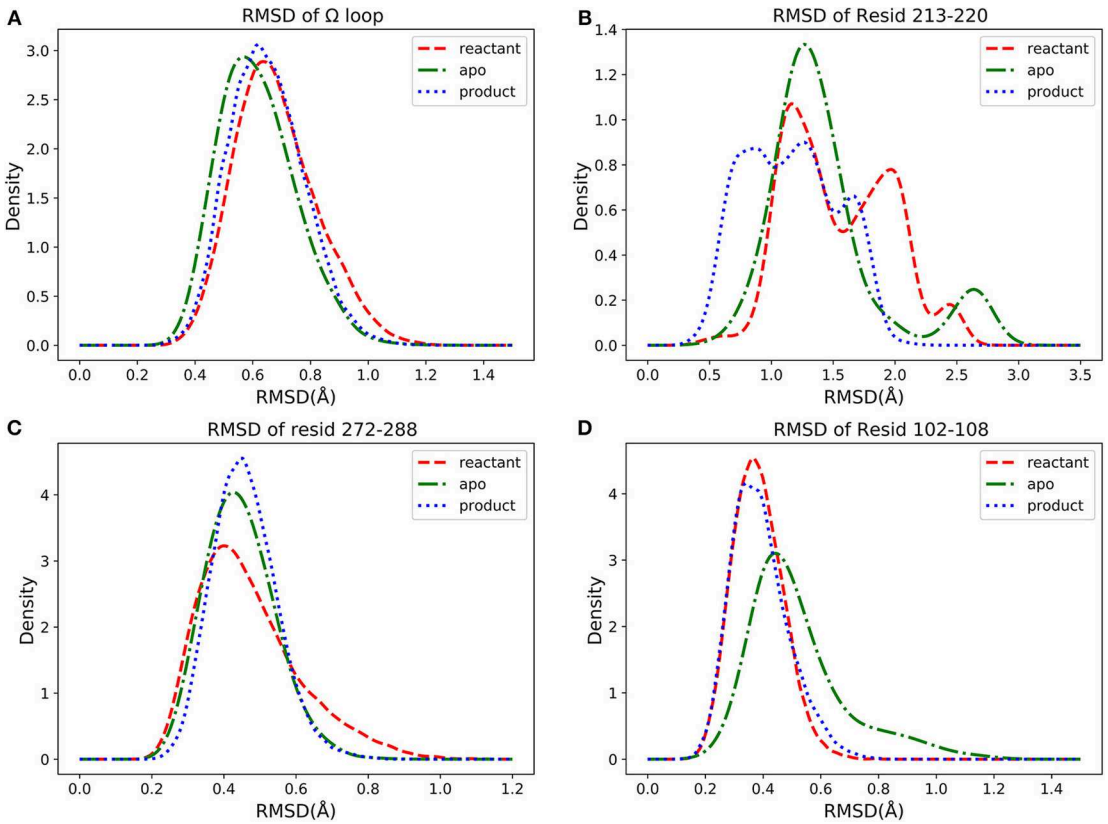

FIGURE 13 | The density distributions of residues groups' RMSDs in apo (green dot dashed line), reactant (red dashed line) and product (blue dot line) states: **(A)** RMSD of Ω loop (residues 163–180), **(B)** RMSD of residues 213–220, **(C)** RMSD of helix 12 (residues 272–288), **(D)** RMSD of residues 102–108.

study, than both apo and reactant state simulations. In addition, both apo and reactant states simulations consistently have similar RMSDs with reference to these TEM-1 crystal structures. Although these results could prove either the simulations are sufficient for the sampling of each state or not, these results are consistent with our results that the apo and reactant states are similar to each, and both are different from the product state. It may suggest that binding with the catalysis product is a dynamically stable state for TEM-1 and contributes to the catalytic activities of TEM-1 against antibiotics. This could lead to some intrinsic dynamical properties of TEM-1 in different functional states, which warrant further in-depth studies.

CONCLUSION

In this study, we developed classification models for TEM-1 β-lactamases in different binding modes against penicillin using a machine learning method called random forest. Using the backbone Cα distances of all residue pairs as the features for the model training purpose, the developed classification models effectively correlate the global protein dynamics with the individual residue correlation, with regard to the different binding modes. The feature importance generated from the classification model training process was used to evaluate the contribution from individual residues, as well as secondary structures in TEM-1, to each model. It is shown that the random coil structures carry the highest feature importance among secondary structures, including α-helix, β-strands, and turns. It may indicate that the motions of coils contribute significantly to the differences among three states, and lead to more flexibility of random coils than in other secondary structures. Accordingly, the protein flexibility is proposed to be a key factor in ligand recognition of TEM-1. Detailed comparison also revealed that the individual key residues identified from the machine learning models not only have a good agreement with the NMR study, but also provide unprecedented insight into the function of individual residues with regard to differentiating protein in different binding modes. Specifically, it is suggested that some catalytically important residues at the active site are also critical for recognizing the hydrolyzed product of antibiotics. Overall, this study demonstrates that machine learning methods provides effective tools to analyze protein dynamics in different binding modes and produce intriguing insight into the correlation between protein functional states and various structural levels.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

FW wrote the manuscript and carried out the four independent MD simulations for three states (1,200 ns) and performed all the analysis. LS carried out 1 MD simulation for three states (300 ns). HZ provided some scripts of machine learning. SW and XW authors contributed to the final version of the manuscript. PT contributed to the final version of the manuscript and supervised the project.

FUNDING

The work was supported by National Science Foundation under a CAREER Grant [1753167] and SMU Dean's Research Council research grant. Computational time was provided by Southern Methodist University's Center for Scientific Computation.

ACKNOWLEDGMENTS

Computational time was provided by Southern Methodist University's Center for Scientific Computation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2019.00047/full#supplementary-material

REFERENCES

Ambler, R. P., Coulson, A. F., Frère, J. M., Ghuysen, J. M., Joris, B., Forsman, M., et al. (1991). A standard numbering scheme for the class A beta-lactamases. *Biochem. J.* 276(Pt 1), 269–270. doi: 10.1042/bj2760269

Avci, F. G., Altinisik, F. E., Karacan, I., Senturk Karagoz, D., Ersahin, S., Eren, A., et al. (2018). Targeting a hidden site on class A beta-lactamases. *J. Mol. Graph. Model.* 84, 125–133. doi: 10.1016/j.jmgm.2018.06.007

Avci, F. G., Altinisik, F. E., Vardar Ulu, D., Ozkirimli Olmez, E., and Sariyar Akbulut, B. (2016). An evolutionarily conserved allosteric site modulates beta-lactamase activity. *J. Enzyme Inhib. Med. Chem.* 31, 33–40. doi: 10.1080/14756366.2016.1201813

Ballester, P. J., and Mitchell, J. B. O. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* 26, 1169–1175. doi: 10.1093/bioinformatics/btq112

Bashtannyk, D. M., and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Stat. Data Anal.* 36, 279–298. doi: 10.1016/S0167-9473(00)00046-3

Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., et al. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 dihedral angles. *J. Chem. Theory Comput.* 8, 3257–3273. doi: 10.1021/ct300400x

Bös, F., and Pleiss, J. (2009). Multiple molecular dynamics simulations of TEM beta-lactamase: dynamics and water binding of the omega-loop. *Biophys. J.* 97, 2550–2558. doi: 10.1016/j.bpj.2009.08.031

Bradford, P. A. (2001). Extended-spectrum beta-lactamases in the 21st century: characterization, epidemiology, and detection of this important resistance threat. *Clin. Microbiol. Rev.* 14, 933–951. doi: 10.1128/CMR.14.4.933-951.2001

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., et al. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint*. arXiv:1309.0238.

Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* 26, 5–14. doi: 10.1016/S0097-8485(01)00094-8

Cortina, G. A., and Kasson, P. M. (2018). Predicting allostery and microbial drug resistance with molecular simulations. *Curr. Opin. Struct. Biol.* 52, 80–86. doi: 10.1016/j.sbi.2018.09.001

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98, 10089–10092. doi: 10.1063/1.464397

Decherchi, S., Berteotti, A., Bottegoni, G., Rocchia, W., and Cavalli, A. (2015). The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nat. Commun.* 6:6155. doi: 10.1038/ncomms7155

Díaz, N., Sordo, T. L., Merz, K. M., and Suárez, D. (2003). Insights into the acylation mechanism of class A β-lactamases from molecular dynamics simulations of the TEM-1 enzyme complexed with benzylpenicillin. *J. Am. Chem. Soc.* 125, 672–684. doi: 10.1021/ja027704o

Díaz, N., Suárez, D., Sordo, T. L., and Merz, K. M. (2001). Acylation of class A β-lactamases by penicillins: a theoretical examination of the role of serine 130 and the β-lactam carboxylate group. *J. Phys. Chem. B* 105, 11302–11313. doi: 10.1021/jp012881h

Dideberg, O., Charlier, P., Wéry, J. P., Dehottay, P., Dusart, J., Erpicum, T., et al. (1987). The crystal structure of the beta-lactamase of *Streptomyces albus* G at 0.3 nm resolution. *Biochem. J.* 245, 911–913. doi: 10.1042/bj2450911

Doucet, N., De Wals, P.-Y., and Pelletier, J. N. (2004). Site-saturation mutagenesis of Tyr-105 reveals its importance in substrate stabilization and discrimination in TEM-1 β-lactamase. *J. Biol. Chem.* 279, 46295–46303. doi: 10.1074/jbc.M407606200

Doucet, N., and Pelletier, J. N. (2007). Simulated annealing exploration of an active-site tyrosine in TEM-1 beta-lactamase suggests the existence of alternate conformations. *Proteins* 69, 340–348. doi: 10.1002/prot.21485

Doucet, N., Savard, P.-Y., Pelletier, J. N., and Gagné, S. M. (2007). NMR investigation of Tyr105 mutants in TEM-1 β-lactamase: dynamics are coorrelated with function. *J. Biol. Chem.* 282, 21448–21459. doi: 10.1074/jbc.M609777200

Eastman, P., and Pande, V. S. (2015). OpenMM: a hardware independent framework for molecular simulations. *Comput. Sci. Eng.* 12, 34–39. doi: 10.1109/MCSE.2010.27

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., et al. (2017). OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comp. Biol.* 13:e1005659. doi: 10.1371/journal.pcbi.1005659

Farmer, J., Kanwal, F., Nikulsin, N., Tsilimigras, M. C. B., and Jacobs, D. J. (2017). Statistical measures to quantify similarity between molecular dynamics simulation trajectories. *Entropy* 19:646. doi: 10.3390/e19120646

Fisette, O., Gagné, S., and Lagüe, P. (2012). Molecular dynamics of class A β-lactamases—effects of substrate binding. *Biophys. J.* 103, 1790–1801. doi: 10.1016/j.bpj.2012.09.009

Fisette, O., Morin, S., Savard, P.-Y., Lagüe, P., and Gagné, S. M. (2010). TEM-1 backbone dynamics—insights from combined molecular dynamics and nuclear magnetic resonance. *Biophys. J.* 98, 637–645. doi: 10.1016/j.bpj.2009.08.061

Fonzé, E., Charlier, P., To'Th, Y., Vermeire, M., Raquet, X., Dubus, A., et al. (1995). TEM1 β-lactamase structure solved by molecular replacement and refined structure of the S235A mutant. *Acta Crystallogr. D Biol. Crystallogr.* 51, 682–694. doi: 10.1107/S0907444994014496

Friedrichs, M. S., Eastman, P., Vaidyanathan, V., Houston, M., Legrand, S., Beberg, A. L., et al. (2009). Accelerating molecular dynamic simulation on graphics processing units. *J. Comput. Chem.* 30, 864–872. doi: 10.1002/jcc.21209

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi: 10.1007/s10994-006-6226-1

Golemi-Kotra, D., Meroueh, S. O., Kim, C., Vakulenko, S. B., Bulychev, A., Stemmler, A. J., et al. (2004). The importance of a critical protonation state

and the fate of the catalytic steps in class A β-lactamases and penicillin-binding proteins. *J. Biol. Chem.* 279, 34665–34673. doi: 10.1074/jbc.M313143200

Halgren, T. A. (1992). The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.* 114, 7827–7843. doi: 10.1021/ja00046a032

Hermann, J. C., Ridder, L., Mulholland, A. J., and Höltje, H.-D. (2003). Identification of Glu166 as the general base in the acylation reaction of class A β-lactamases through QM/MM modeling. *J. Am. Chem. Soc.* 125, 9590–9591. doi: 10.1021/ja034434g

Herzberg, O., and Moult, J. (1987). Bacterial resistance to beta-lactam antibiotics: crystal structure of beta-lactamase from Staphylococcus aureus PC1 at 2.5 A resolution. *Science* 236, 694–701. doi: 10.1126/science.3107125

Herzberg, O., and Moult, J. (1991). Penicillin-binding and degrading enzymes. *Curr. Opin. Struct. Biol.* 1, 946–953. doi: 10.1016/0959-440X(91)90090-G

Horn, J. R., and Shoichet, B. K. (2004). Allosteric inhibition through core disruption. *J. Mol. Biol.* 336, 1283–1291. doi: 10.1016/j.jmb.2003.12.068

Husic, B. E., and Pande, V. S. (2018). Markov state models: from an art to a science. *J. Am. Chem. Soc.* 140, 2386–2396. doi: 10.1021/jacs.7b12191

Jelsch, C., Lenfant, F., Masson, J. M., and Samama, J. P. (1992). β-lactamase TEM1 of *E. coli* crystal structure determination at 2.5 Å resolution. *FEBS Lett.* 299, 135–142. doi: 10.1016/0014-5793(92)80232-6

Jelsch, C., Mourey, L., Masson, J.-M., and Samama, J.-P. (1993). Crystal structure of *Escherichia coli* TEM1 β-lactamase at 1.8 Å resolution. *Proteins* 16, 364–383. doi: 10.1002/prot.340160406

Jolliffe, I. (2011). "Principal component analysis," in *International Encyclopedia of Statistical Science,* ed M. Lovric (Berlin: Springer), 1094–1096.

Kabsch, W., and Sander, C. (1983). DSSP: definition of secondary structure of proteins given a set of 3D coordinates. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211

Lamotte-Brasseur, J., Dive, G., Dideberg, O., Charlier, P., Frère, J. M., and Ghuysen, J. M. (1991). Mechanism of acyl transfer by the class A serine β-lactamase of *Streptomyces albus* G. *Biochem. J.* 279, 213–221. doi: 10.1042/bj2790213

Lamotte-Brasseur, J., Jacob-Dubuisson, F., Dive, G., Frère, J. M., and Ghuysen, J. M. (1992). *Streptomyces albus* G serine beta-lactamase. Probing of the catalytic mechanism via molecular modelling of mutant enzymes. *Biochem. J.* 282(Pt 1), 189–195. doi: 10.1042/bj2820189

Lamotte-Brasseur, J., Knox, J., Kelly, J. A., Charlier, P., Fonzé, E., Dideberg, O., et al. (1994). The structures and catalytic mechanisms of active-site serine β-lactamases. *Biotechnol. Genet. Eng. Rev.* 12, 189–230. doi: 10.1080/02648725.1994.10647912

Lamotte-Brasseur, J., Lounnas, V., Raquet, X., and Wade, R. C. (1999). pKa calculations for class A β-lactamases: influence of substrate binding. *Protein Sci.* 8, 404–409. doi: 10.1110/ps.8.2.404

Li, Z., Kermode, J. R., and De Vita, A. (2015). Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* 114:096405. doi: 10.1103/PhysRevLett.114.096405

Louppe, G. (2014). Understanding random forests: from theory to practice. *arXiv preprint*. arXiv:1407.7502.

Marciano, D. C., Brown, N. G., and Palzkill, T. (2009). Analysis of the plasticity of location of the Arg244 positive charge within the active site of the TEM-1 β-lactamase. *Protein Sci.* 18, 2080–2089. doi: 10.1002/pro.220

Matagne, A., Lamotte-Brasseur, J., and FrÈRe, J.-M. (1998). Catalytic properties of class A β-lactamases: efficiency and diversity. *Biochem. J.* 330, 581–598. doi: 10.1042/bj3300581

Maveyraud, L., Pratt, R. F., and Samama, J.-P. (1998). Crystal structure of an acylation transition-state analog of the TEM-1 β-lactamase. Mechanistic implications for class A β-lactamases. *Biochemistry* 37, 2622–2628. doi: 10.1021/bi972501b

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., et al. (2015). MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* 109, 1528–1532. doi: 10.1016/j.bpj.2015.08.015

Meneksedag, D., Dogan, A., Kanlikilicer, P., and Ozkirimli, E. (2013). Communication between the active site and the allosteric site in class A beta-lactamases. *Comput. Biol. Chem.* 43, 1–10. doi: 10.1016/j.compbiolchem.2012.12.002

Meroueh, S. O., Fisher, J. F., Schlegel, H. B., and Mobashery, S. (2005). Ab Initio QM/MM study of class A β-lactamase acylation: dual participation of Glu166 and Lys73 in a concerted base promotion of Ser70. *J. Am. Chem. Soc.* 127, 15397–15407. doi: 10.1021/ja051592u

Minasov, G., Wang, X., and Shoichet, B. K. (2002). An ultrahigh resolution structure of TEM-1 β-lactamase suggests a role for Glu166 as the general base in acylation. *J. Am. Chem. Soc.* 124, 5333–5340. doi: 10.1021/ja0259640

Modi, T., and Ozkan, B. S. (2018). Mutations utilize dynamic allostery to confer resistance in TEM-1 β-lactamase. *Int. J. Mol. Sci.* 19:E3808. doi: 10.3390/ijms19123808

Moews, P. C., Knox, J. R., Dideberg, O., Charlier, P., and Frère, J.-M. (1990). β-lactamase of *Bacillus licheniformis* 749/C at 2 Å resolution. *Proteins* 7, 156–171. doi: 10.1002/prot.340070205

Oefner, C., D'Arcy, A., Daly, J. J., Gubernator, K., Charnas, R. L., Heinze, I., et al. (1990). Refined crystal structure of β-lactamase from *Citrobacter freundiiindicates* a mechanism for β-lactam hydrolysis. *Nature* 343, 284–288. doi: 10.1038/343284a0

Oleinikovas, V., Saladino, G., Cossins, B. P., and Gervasio, F. L. (2016). Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *J. Am. Chem. Soc.* 138, 14257–14263. doi: 10.1021/jacs.6b05425

Palzkill, T. (2018). Structural and mechanistic basis for extended-spectrum drug-resistance mutations in altering the specificity of TEM, CTX-M, and KPC β-lactamases. *Front. Mol. Biosci.* 5:16. doi: 10.3389/fmolb.2018.00016

Paramasivam, S., Gronenborn, A. M., and Polenova, T. (2018). Backbone amide 15N chemical shift tensors report on hydrogen bonding interactions in proteins: a magic angle spinning NMR study. *Solid State Nucl. Magn. Reson.* 92, 1–6. doi: 10.1016/j.ssnmr.2018.03.002

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: http://jmlr.org/papers/v12/pedregosa11a.html

Petrosino, J., Cantu, C., and Palzkill, T. (1998). β-lactamases: protein evolution in real time. *Trends Microbiol.* 6, 323–327. doi: 10.1016/S0966-842X(98)01317-1

Pimenta, A. C., Martins, J. M., Fernandes, R., and Moreira, I. S. (2013). Ligand-induced structural changes in TEM-1 probed by molecular dynamics and relative binding free energy calculations. *J. Chem. Inf. Model.* 53, 2648–2658. doi: 10.1021/ci400269d

Roccatano, D., Sbardella, G., Aschi, M., Amicosante, G., Bossa, C., Nola, A. D., et al. (2005). Dynamical aspects of TEM-1 β-lactamase probed by molecular dynamics. *J. Comput. Aided Mol. Des.* 19, 329–340. doi: 10.1007/s10822-005-7003-0

Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327–341. doi: 10.1016/0021-9991(77)90098-5

Savard, P.-Y., and Gagné, S. M. (2006). Backbone dynamics of TEM-1 determined by NMR: evidence for a highly ordered protein. *Biochemistry* 45, 11414–11424. doi: 10.1021/bi060414q

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York, NY: John Wiley & Sons.

Sgrignani, J., Grazioso, G., De Amici, M., and Colombo, G. (2014). Inactivation of TEM-1 by Avibactam (NXL-104): insights from quantum mechanics/molecular mechanics metadynamics simulations. *Biochemistry* 53, 5174–5185. doi: 10.1021/bi500589x

Shcherbinin, D., and Veselovsky, A. (2019). "Analysis of protein structures using residue interaction networks," in *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process,* ed C. G. Mohan (Cham: Springer International Publishing), 55–69.

Silverman, B. W. (2018). *Density Estimation for Statistics and Data Analysis*. New York, NY: Routledge.

Simm, A. M., Baldwin, A. J., Busse, K., and Jones, D. D. (2007). Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 β-lactamase. *FEBS Lett.* 581, 3904–3908. doi: 10.1016/j.febslet.2007.07.018

Stec, B., Holtz, K. M., Wojciechowski, C. L., and Kantrowitz, E. R. (2005). Structure of the wild-type TEM-1 β-lactamase at 1.55 Å and the mutant enzyme Ser70Ala at 2.1 Å suggest the mode of noncovalent catalysis for the mutant enzyme. *Acta Crystallogr. D Biol. Crystallogr.* 61, 1072–1079. doi: 10.1107/S0907444905014356

Stojanoski, V., Chow, D.-C., Hu, L., Sankaran, B., Gilbert, H. F., Prasad, B. V. V., et al. (2015). A triple mutant in the Ω-loop of TEM-1 β-lactamase changes the substrate profile via a large conformational change and an altered general base for catalysis. *J. Biol. Chem.* 290, 10382–10394. doi: 10.1074/jbc.M114.633438

Strynadka, N. C. J., Adachi, H., Jensen, S. E., Johns, K., Sielecki, A., Betzel, C., et al. (1992). Molecular structure of the acyl-enzyme intermediate in β-lactam hydrolysis at 1.7 Å resolution. *Nature* 359, 700–705. doi: 10.1038/359700a0

Strynadka, N. C. J., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R., et al. (1996). Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. *Nat. Struct. Biol.* 3, 233–239. doi: 10.1038/nsb0396-233

Swarén, P., Maveyraud, L., Raquet, X., Cabantous, S., Duez, C., Pédelacq, J.-D., et al. (1998). X-ray analysis of the NMC-A β-lactamase at 1.64-Å resolution, a class A carbapenemase with broad substrate specificity. *J. Biol. Chem.* 273, 26714–26721. doi: 10.1074/jbc.273.41.26714

Turlach, B. A. (1993). *Bandwidth Selection in Kernel Density Estimation: A Review*. Berlin: Citeseer.

Vanwetswinkel, S., Avalle, B., and Fastrez, J. (2000). Selection of β-lactamases and penicillin binding mutants from a library of phage displayed TEM-1 β-lactamase randomly mutated in the active site Ω-loop11Edited by A. *R. Fersht. J. Mol. Biol.* 295, 527–540. doi: 10.1006/jmbi.1999.3376

Wang, X., Minasov, G., and Shoichet, B. K. (2002). Noncovalent interaction energies in covalent complexes: TEM-1 β-lactamase and β-lactams. *Proteins* 47, 86–96. doi: 10.1002/prot.10058

Zafaralla, G., Manavathu, E. K., Lerner, S. A., and Mobashery, S. (1992). Elucidation of the role of arginine-224 in the turnover processes of class A beta-lactamases. *Biochemistry* 31, 3847–3852. doi: 10.1021/bi00130a016

Zhou, H., Dong, Z., and Tao, P. (2018). Recognition of protein allosteric states and residues: machine learning approaches. *J. Comput. Chem.* 39, 1481–1490. doi: 10.1002/jcc.25218

Zhou, H., Dong, Z., Verkhivker, G., Zoltowski, B. D., and Tao, P. (2019). Allosteric mechanism of the circadian protein vivid resolved through markov state model and machine learning analysis. *PLoS Comp. Biol.* 15:e1006801. doi: 10.1371/journal.pcbi.1006801

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Wang, Shen, Zhou, Wang, Wang and Tao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.