PCCP

Unveiling the Structural Features that Regulate Carbapenems Deacylation in KPC-2 Through QM/MM and Interpretable Machine Learning

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-08-2022-003724.R2
Article Type:	Paper
Date Submitted by the Author:	18-Nov-2022
Complete List of Authors:	Yin, Chao; Southern Methodist University, Deparment of Chemistry Song, Zilin; Southern Methodist University, Department of Chemistry Tian, Hao; Southern Methodist University Dedman College of Humanities and Sciences, Chemistry Palzkill, Timothy; Baylor College of Medicine, Pharmacology Tao, Peng; Southern Methodist University, Department of Chemistry; Southern Methodist University

SCHOLARONE™
Manuscripts

Unveiling the Structural Features that Regulate Carbapenems Deacylation in KPC-2 Through QM/MM and Interpretable Machine Learning

Authors:

Chao Yin,[a] Zilin Song,[a] Hao Tian,[a] Timothy Palzkill,[b] Peng Tao[a,*]

Affiliations:

[a] Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, Texas 75205, United States;

[b] From the Department of Pharmacology and Chemical Biology, Baylor College of Medicine, Houston, Texas 77030, United States; orcid.org/0000-0002-5267-0001;

***Author to whom any correspondence should be addressed:**

ptao@smu.edu (P.T.)

Abstract

Resistance to carbapenem β-lactams presents major clinical and economical challenges for the treatment of pathogen infections. The fast hydrolysis of carbapenems by carbapenemase-producing bacterial strains enables the effective deactivation of carbapenem antibiotics. In this study, we aim to unravel the structural features that distinguish the notable deacylation activity of carbapenemases. The deacylation reactions between imipenem (IPM) and the KPC-2 class A serine-based β-lactamases (ASβLs) are modeled with combined Quantum Mechanical/Molecular Mechanical (QM/MM) minimum energy pathway (MEP) calculations and interpretable machine-learning (ML) methods. We firstly applied a dual-level computational protocol to achieve fast sampling of QM/MM MEPs. A tree-based ensemble ML model was employed to learn the MEP activation barriers from the conformational features of the KPC-2/IPM active site. The barrier-predicting model was then unboxed using the Shapley Additive Explanations (SHAP) importance attribution methods to derive mechanistic insights, which were also verified by additional QM/MM wavefunction analysis. Essentially, we show that potential hydrogen bond interactions to the general base and the tautomerization states of the carbapenem pyrroline ring could concertedly regulate the activation barrier of KPC-2/IPM deacylation. Nonetheless, we demonstrate the efficacy of interpretable ML to assist the analysis of QM/MM simulation data for robust extraction of human-interpretable mechanistic insights.

Introduction

β-Lactam-resistant bacterial strains challenge public health and sustainable economic development from various aspects. β-Lactamases have long been identified as the immediate cause of β-lactam antibiotic resistance encountered in most resistant strains. In particular, the resistance to carbapenems, a series of β-lactam drugs that are of great clinical importance, has also emerged due to their effective hydrolysis mediated by carbapenemases[1–4].

Carbapenemases belonging to the class A Serine-based β-lactamases (ASβLs) family that hydrolyze the β-lactams substrates through a generally conserved acylation – deacylation mechanism[5–7]. The acylation half of the reaction is triggered by the nucleophilic attack of Ser70 hydroxyl to the β-lactam carbonyl. Notably, while being highly conserved in ASβLs, the acylation pathways of the reaction have shown mechanistic flexibility on the residues acting as the general bases[3]. In the subsequent deacylation step, the deacylation water attacks the acyl-enzyme ester carbon while synergistically delivering one of its protons to the Glu166 carboxyl, which is the only viable general base for deacylation (Fig. 1)[5]. Relatedly, the mechanism of β-lactams hydrolysis differs in the class B zinc-based β-lactamases and the class D β-lactamases. In the case of class B β-lactamases, the zinc ions in the active site mediates the nucleophilic water attacks and the rapid ligand dissociation,[8–12], while class D β-lactamases use a carboxylated lysine (Lys73-CO_2) as the general base for both acylation and deacylation.[13,14]

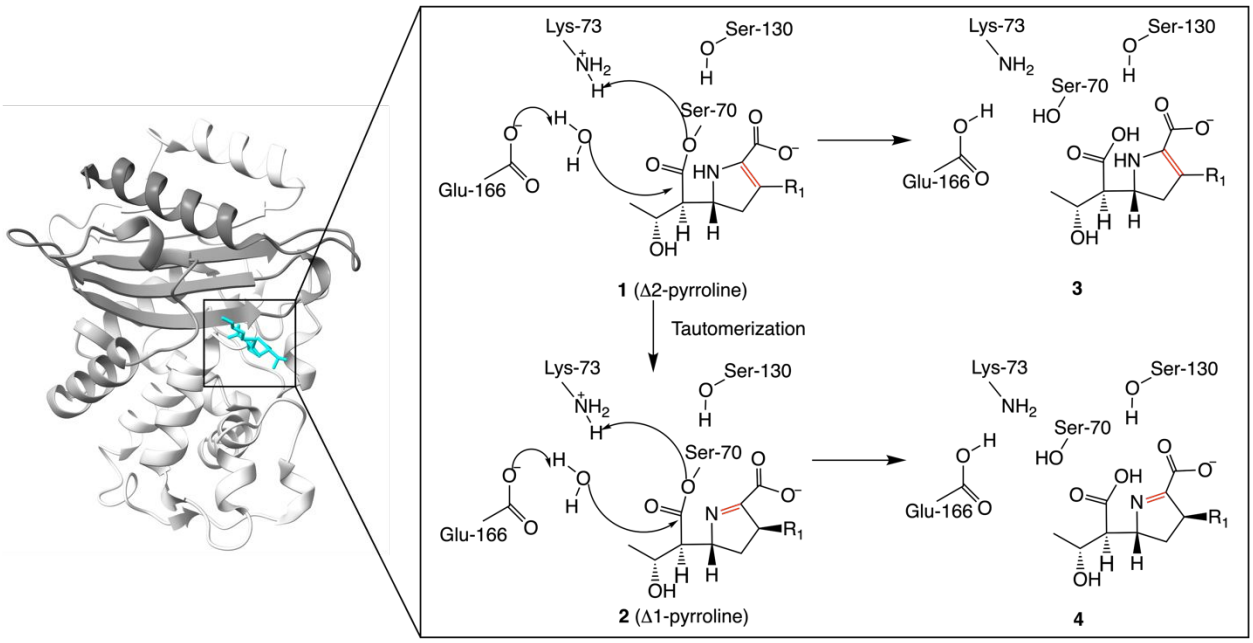

Fig. 1. The crystal structure of KPC-2/IPM (PDB: 6XJ8)[15] and the mechanisms of the deacylation step in KPC-2/IPM hydrolysis. The red bonds highlight the Δ1 and Δ2 tautomerization states on IPM pyrroline.

The *Klebsiella pneumoniae* carbapenemases (KPC) family of ASβLs has been identified as a frequent cause of antibiotic drug resistance[16]. The KPC-2 variant of the KPC sub family has been investigated by pioneering experimental efforts. In particular, the kinetic study of Mehta *et al*.[15] reported that the carbapenem resistance driven by KPC-2 stems from the effective deacylation of the acyl-enzyme intermediate. In addition, changes of the local environment surrounding the general base (Glu166) were reported to impact the catalytic activity of KPC-2 and related carbapenemases[17–22]. Specifically, it was reported that the hydrogen bond between the general base Glu166 and Tyr72 in the KPC-2 Phe72Tyr (KPC-F72Y) variant reduces the basicity of Glu166 and impede the deacylation reaction of carbapenemases.[17] Extensive evidence have further demonstrated that the deacylation activity of carbapenemases is correlated with the tautomerization states of the conserved five-member pyrroline ring in the carbapenem scaffold.[23–

[25] Computational efforts utilizing the combined Quantum Mechanical/Molecular Mechanical methods (QM/MM) have also been employed to derive mechanistic insights into various β-lactam hydrolysis by β-lactamases[5,26–32]. Recently, Chudyk *et al*.[28] reported that the deacylation of meropenem catalyzed by ASβL-carbapenemases is also related to the orientation of the 6α-hydroxyethyl groups on the substrate. While various structural and kinetic features have been proposed to impact deacylation activity in carbapenemases in general, the correlation between the local environment of Glu166 as the general base, the pyrroline tautomerization state, and the orientation of the carbapenem 6α-hydroxyethyl is yet to be clarified.

In light of the on-going emergence of machine-learning (ML) techniques to approximate complex biophysical and chemical observables[33–37], the explainability and interpretability of ML models has been a focus to understand the underlying mechanism basis of the studied problem. Machine-learning is data-driven approaches that can learn patterns from existing data without the *a prior* knowledge on the variable correlations.[38,39] Among ML approaches, supervised learning methods[40,41] such as linear regression[42], decision trees[43], random forest (RF)[44], support vector machines (SVM)[45], and deep learning (DL)[46] have been widely used and have been applied in many aspects of chemistry[47–49]. Compared to other ML methods, the tree-based Extreme Gradient Boosting model (XGBoost)[50] model is superior in both prediction performance and explainability. The XGBoost method has been widely used in Quantitative Structure–Activity Relationships analysis[51,52], prediction of reaction barriers [53], reaction yield[54], and drug discovery[55].

Different explainable ML (XML) models, including anchor explanations[56], counterfactual explanations[57], integrated gradients[58] and the Shapley Additive Explanations (SHAP) method for tree-based models[59,60], have been proposed to facilitate ML explainability for various regression methods. The XGBoost model inherits the linear explainability from the tree-based models,

making the SHAP method for tree models as the optimal XML model to unveil the underlying mechanistic basis of KPC-2/IPM deacylation. In addition, the combined scheme of the XGBoost and SHAP methods has been commonly applied in different fields.[61–63]

In this computational study, we applied QM/MM minimum energy pathway (MEP) calculations and XML methods to unveil the structural features that control the deacylation activity of KPC-2 carbapenemases with the antibiotic imipenem (IPM). We focus on the deacylation reaction in four model systems: the wild-type KPC-2 and IPM-Δ2 tautomer (KPC-WT/IPM-Δ2), the wild-type KPC-2 and IPM-Δ1 tautomer (KPC-WT/IPM-Δ1), the Phe72Tyr mutant KPC-2 and IPM-Δ2 tautomer (KPC-F72Y/IPM-Δ2), and the Phe72Tyr mutant KPC-2 and IPM-Δ1 tautomer (KPC-F72Y/IPM-Δ1). We first present the computational QM/MM workflow that enables fast sampling of QM/MM MEPs. The XGBoost model was employed to learn the deacylation energy barriers from the conformational features selected from the acyl-enzyme reactant conformations. The impacts from essential structural factors to the deacylation barrier heights were quantified by the native feature importance of the XGBoost model and the SHAP methods[59,60]. Most importantly, we reveal the interplay between the major structural factors that regulate the KPC-2/IPM deacylation reactivity using our integrated computational schemes.

Computational Methods

System setup

The KPC-2 crystal complex with a hydrolyzed IPM molecule (PDB: 6XJ8)[15] was used as the starting structure and the mutant residue Ala170 was modified to Asn170 as in the wild type enzyme. The CHARMM General Force Field (CGenFF)[64] parameters of the IPM ligand in its unbound form were generated using the CGenFF portal (https://cgenff.umaryland.edu). The protonation states on titratable amino acid residues were set as the default protonation states from

the CHARMM36 (C36) topologies[65]. Specifically, all Arg and Lys residues were protonated, all His residues were modelled as singly protonated on Nδ1 position, while Asp and Glu residues were deprotonated. In addition, the Cys69 and Cys238 residues were connected as the disulfide bridge conserved in most ASβL-carbapenemases.[66,67] The KPC-2/IPM complex was immersed in an 80 Å × 80 Å × 80 Å cubic box of TIP3P solvent molecules to ensure a minimum distance of 10 Å between the enzyme complex and the boundary of the simulation box. Sodium and chloride ions were added to neutralize the total charge of the system. 300 steps of steepest descent minimizations using the classical potentials were firstly performed on the solvent molecules with the enzyme complex fixed in place. Then, 3,000 steps of adopted-basis Newton Raphson (ABNR) minimizations were performed on the simulation system with the following residues fixed in place: Ser70, Phe72, Lys73, Ser130, Asn132, Glu166, Asn170, IPM, and deacylation water (DW). Due to the inability of the CGenFF parameters to treat the covalent bond between Ser70 and the β-lactam carbonyl, we switched to semi-empirical QM/MM method to further relax the model system.

The single link atom scheme was used to partition the covalent bonds between the QM and MM region, which were defined as the Cα – Cβ bonds on the amino acid residues. The third-order Density Functional Tight Binding theory (DFTB3)[36] was used as the QM Hamiltonian while the rest of the system was treated with the classical potentials. The acyl-enzyme complex (the reactant conformation for the KPC-WT/IPM-Δ2 deacylation pathway) was created by minimizing the QM/MM system with necessary distance-based quadratic bias potentials. We note that the IPM-Δ2 tautomer was created by pulling the excess hydrogen on Ser70 hydroxyl (Ser70 Hγ) onto the IPM β-lactam nitrogen. The biased minimization creates the KPC-WT/IPM-Δ2 acyl-enzyme conformation, which was further relaxed for 5,000 ABNR steps with no bias potential or positional

constraint at the DFTB3/C36 level. From the optimized KPC-WT/IPM-Δ2 conformation, we created the KPC-WT/IPM-Δ1 states by imposing distance-based restraints to pull the proton on IPM N4 to IPM C2 on its *S* stereoisomer side with further minimizations. We note that the IPM-(*S*)-Δ1 tautomer state has been reported in crystal structures in KPC-2/IPM acyl-enzyme complex[17]. The KPC-F72Y systems (KPC-F72Y/IPM-Δ1 and KPC-F72Y/IPM-Δ2) were then created by mutating Phe72 in the wild-type systems to Tyr72 (Fig. 2).

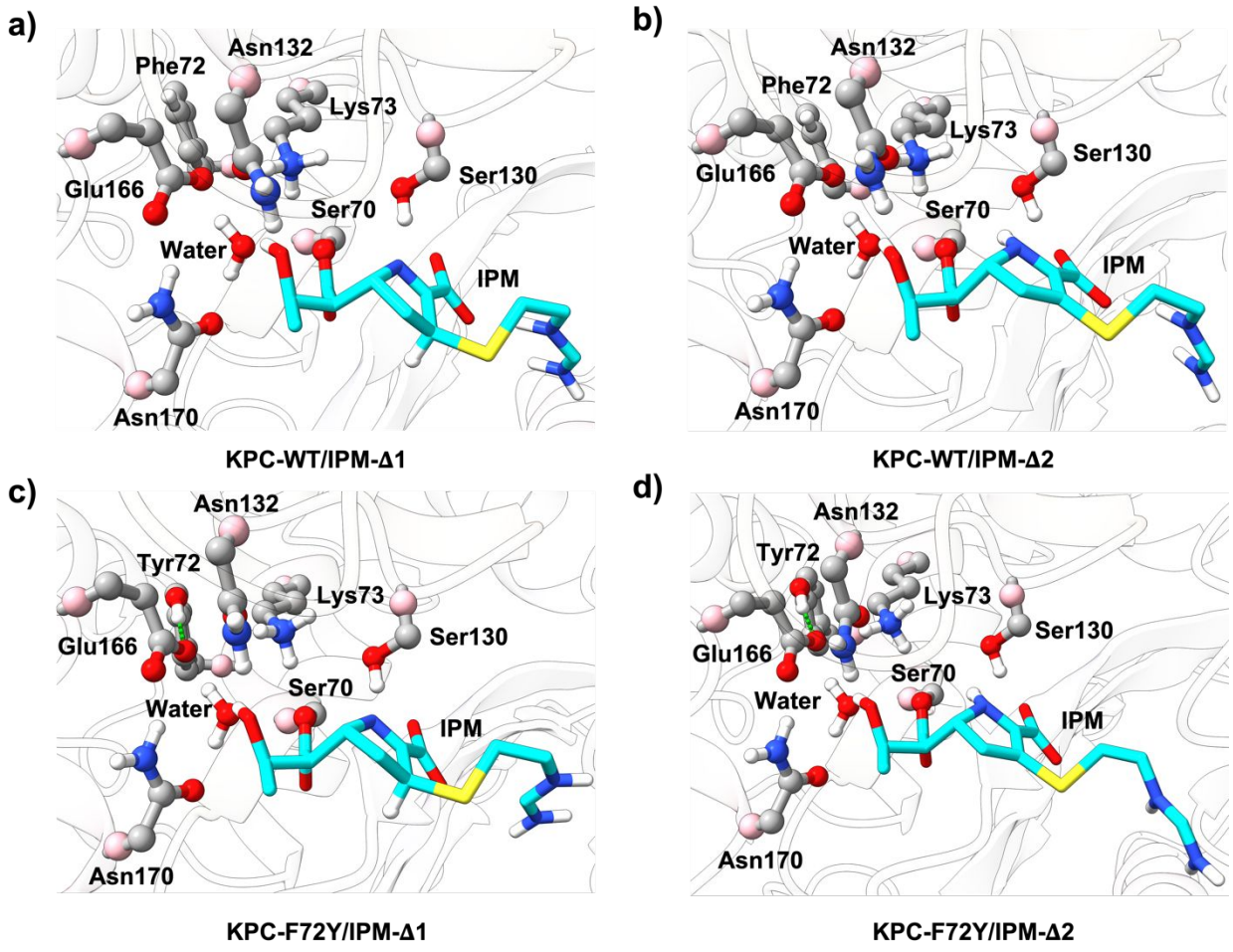

Fig. 2. Active site configurations for the four modeling systems. (a) KPC-WT/IPM-Δ1. (b) KPC-WT/IPM-2; (c) KPC-F72Y/IPM-Δ1; (d) KPC-F72Y/IPM-Δ2. The carbon atoms of residues for QM region are colored as grey, and the carbon atoms of IPM ligand are colored with cyan. All hydrogen link-atoms, nitrogen, oxygen, sulfur, hydrogen atoms are colored in pink, blue, red,

yellow, and white, respectively. The extra hydrogen bond between Tyr72 and Glu166 in KPC-F72Y systems is presented as a dashed green line.

To effectively sample the external configurations, we returned to the pure MM treatment of the four systems. The link atoms were temporarily removed from the simulation box and the MM host Cα atoms were rescaled to its atomic unit mass. The key reacting atom groups (the Ser70 hydroxyl, the Lys73 amino, the Ser130 hydroxyl, the Glu166 Oε2, the IPM bicyclic rings, and the deacylation water, DW) were fixed in place to retain their QM optimized orientations. The four system was gradually heated from 110 K to 310 K in 50 ps with explicit velocity scaling. Isothermal and isobaric (NPT) equilibration dynamics was then performed for 350 ps with the system temperatures maintained at 310 K using the Hoover thermostat and pressure at 1 atm with the Langevin piston method[69]. Each simulation system was subjected to 100 ns NVT dynamics sampling with positional constraints on the aforenoted key reacting groups. Conformational snapshots were collected at a 500 ps interval, leading to a total number of 800 sampled configurations from the four systems (200 snapshots per system).

In this study, all molecular dynamics (MD) simulations were integrated at 1 fs time steps. The SHAKE algorithm[70] was applied to constrain the solvent molecules as rigid bodies. The nonbonding part of the classical interactions were treated explicitly within 12 Å. The Van der Waals interactions were smoothed to zero at 16 Å. The long-range electrostatic interactions were treated with the particle mesh Ewald (PME)[71] summation under periodical boundary conditions. All MD simulations were performed with CHARMM[41] and OpenMM[74].

QM/MM MEPs

For each of the sampled 800 configurations, we first rebuilt the QM/MM partitioning scheme. Each configuration was first minimized with the DFTB3/C36 level of theory with the MM residues beyond 4 Å of the QM region fixed in place, which produces the reactant acyl-enzyme states. The initial product states were obtained from minimizations with 500 kcal mol^{-1} $Å^{-1}$ restraining forces on the atoms involved in the deacylation reaction (catalytic water, Lys73 Hζ1, Ser70 Oγ, Ser70Hγ, and IPM C7). The final product states were created by further minimizing the initial product configurations without restraints.

The chain-of-states Reaction Path with the Holonomic Constraints (RPwHC) method of Brokaw *et al.*[75] was applied for the calculation of the DFTB3/C36 MEPs. The initial guess of the MEPs was obtained from linearly intercepting the Cartesian space between each pair of the acyl-enzyme reactant and the deacylated product configurations with 36 replicated structures (replicas). A kinetic energy potential force of 0.05 kcal mol^{-1} $Å^{-1}$ was adopted for all MEP calculations. While the RPwHC method enforces equal mass-weighted root-mean-square distances between adjacent replica images, the masses of reacting hydrogen atoms (Water H1, H2 and Lys73 Hζ1) were scaled by a weighting factor of 50 to capture their continuous displacement along the MEPs. The MEPs were considered to converge when the energy change of the whole chain between each minimization step was lower than 0.01 kcal mol^{-1}. The DFTB3/C36 optimized replicas along each MEP were subjected to single point calculations at Density Functional Theory level to obtain accurate energetic profiles. The B3LYP hybrid functional[76,77] with the 6-31++G** basis set[78] plus the Becke-Johnson damped version of the D3 dispersion correction[79] was used as the high-level counterpart for the single point energy refinement(B3LYP-D3(BJ)/6-31+G**/C36). All QM/MM

calculations were performed with the DFTB3 module of CHARMM[80,81] and the CHARMM/Q-Chem interface[82].

Machine Learning

While the potential energy barriers on the sampled MEPs can be regarded as dependent on the acyl-enzyme conformations, the high dimensionality of the conformational space of the QM active site prevents the effective identification of key structural factors that regulate the height of the activation barriers. We selected the key reaction coordinates and potential hydrogen bonds in all reactants (acyl-enzyme) conformations as the input features (Fig. 3). The key reaction coordinates were selected as the bond formation distances during the deacylation. The potential hydrogen interactions were identified with the donor – acceptor conformations that satisfied the Baker-Hubbard criteria in at least one of the acyl-enzyme structures. The features for the potential hydrogen interactions were extracted as the hydrogen – acceptor distances. The postprocessing of the molecular conformations used the MDAnalysis package[83].

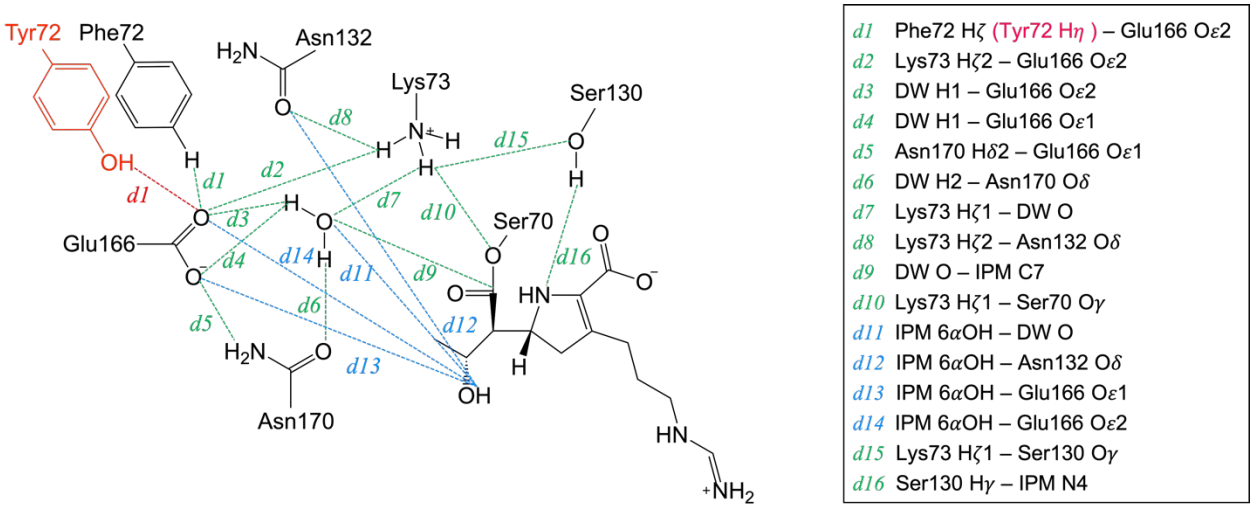

Fig. 3. Features selection scheme, all green dashed lines are potential hydrogen bond in the active site (except d1 in the KPC-WT systems and d9). Residue Tyr72, only existing in the KPC-F72Y systems, is colored as red. Feature d1 (green) is the distance between atom Phe72 Hζ and atom

Glu166 Oε2 in the KPC-WT systems. Feature d1 (red) is the distance between atom Tyr72 Hη and Glu166 Oε2 in the KPC-F72Y systems. All hydrogen bonds involving with IPM 6α hydroxyl are colored as blue. Feature d9 is the distance between deacylation water O atom and IPM C7 atom, which is assumed to be critical for the deacylation step of carbapenem hydrolysis.

All reactants leading to reaction pathways with various reaction barriers were used as the training data due to the following considerations. The large number of reactant conformations were used in the training dataset to ensure the generality of the conclusions about the importance of the features used for model development. Reaction pathways with higher barriers also provide additional information on the correlation between geometric features and the reaction barriers, which is helpful for identifying the features that are strongly correlated with the high reaction barriers.

Four ML methods, linear regression, the XGBoost method, SVM, and neural networks, were applied to learn the correlation between the acyl-enzyme (reactant) conformational feature vectors and deacylation barriers. 720 conformations were included in the training set and the remaining 80 conformations as the validation set by using the stratified splitting which prevents sample unbalance. High performance was observed not only on training set but also on the validation set based on the XGBoost, SVM, and neural network models (Fig. 6 and Fig. S1-S7). The kernel function in the SVM model makes it hard to interpret, and the nonlinear activations of the neural network also complicates its explanation. Therefore, XGBoost was chosen as the machine learning model in this study. The hyperparameters of the XGBoost model were selected *via* a grid search strategy which minimizes the square-error between the QM/MM barrier energy and the predicting barrier energy by XGBoost model (Table 1, and Fig. S9-S11). The best learning rate for all cases is 0.1. The grid search shows that the max_depth parameter plays a significant role in the

performance of the training model. The Mean Absolute Error (MAE) between the barrier energy calculated by the QM/MM method and the predicted barrier energy by the XGBoost method ranges from 2.70 to 0.24 kcal/mol on the training set with max_depth varying from 1 to 9. We chose max_depth as 3 as a balance between overfitting and underfitting. We selected 0.6 for subsample and 1 for min_child_weight since the model has a good MAE (2.37 kcal/mol) on the validation set. The linear regression, XGBoost and SVM models are implemented in scikit-learn package[84], and neural network model is carried out in Tensorflow[85] and Keras[86], while the SHAP method are performed with SHAP package[59].

Table 1. The grid search of optimal hyperparameters for the XGBoost model

Hyperparameters	Range	Optimal value
Learning rate(η)	0.01,0.1,1	0.1
Max depth	1-9	3
Subsample ratio	0.5-0.9	0.6
Min_child_weight	0-9	1

The SHAP method[60] was used to interpret the ML model and explore the mechanism of KPC-2/IPM hydrolysis. The SHAP method attributes feature importance for each sample as the feature's contribution to the deviation between the sample output and the expectation of the model outputs. In brief, the resulting feature importance from the SHAP method on one input sample accounts for the contribution of the feature to the difference between the predicted output and the expectation of the overall model outputs. The assigned SHAP values are also additive for each data sample: the SHAP values from each feature sum up to the total deviation of the corresponding output and the expectation of the model output. By determining the SHAP values of all features on all the training examples, one could quantify the impact of the input feature on the model predictions.

Results and discussion

Deacylation Barriers

The deacylation barriers of the four simulated systems are listed in Fig. 4. The exponential averaged barriers (ΔE_{EA}) were calculated as:

$$\Delta E_{EA} = -RT\ln\left(\frac{1}{N}\sum_{i=1}^{N}\exp\left(-\frac{\Delta E_i}{RT}\right)\right) \#(1)$$

with R being the ideal gas constant; T being the temperature, ΔE_i being the potential energy barrier on the i-th MEP, and N being the total number of MEPs. Accordingly, we rank the deacylating activity of the four systems as: KPC-WT/IPM-Δ2 (18.32 kcal mol^{-1}) > KPC-WT/IPM-Δ1 (19.65 kcal mol^{-1}) > KPC-F72Y/IPM-Δ2 (21.93 kcal mol^{-1}) > KPC-F72Y/IPM-Δ1 (31.60 kcal mol^{-1}). Although the general hypothesis suggests that a tetrahedral intermediate may exist during the deacylation process leading to two transition states, only one transition state was observed along MEPs in all four systems without the tetrahedral intermediates as energy minima. Attempts to characterize tetrahedral intermediates did not lead to stable structures as local minimum using the level of theory employed in this study, suggesting that the tetrahedral intermediates may not be sufficiently stable to carry mechanistic significance. In addition, due to the similarity between the IPM-Δ1 and IPM-Δ2 systems, the TSs observed in these two systems are structurally similar. However, the chemical distinction between these two systems governs that these the TSs from these two systems are chemically distinct from each other despite the structural similarity. Machine learning methods are superior in capturing subtle yet meaningful differences and producing classification models with these meaningful differences encoded and are employed in the subsequent study using these MEPs as training data.

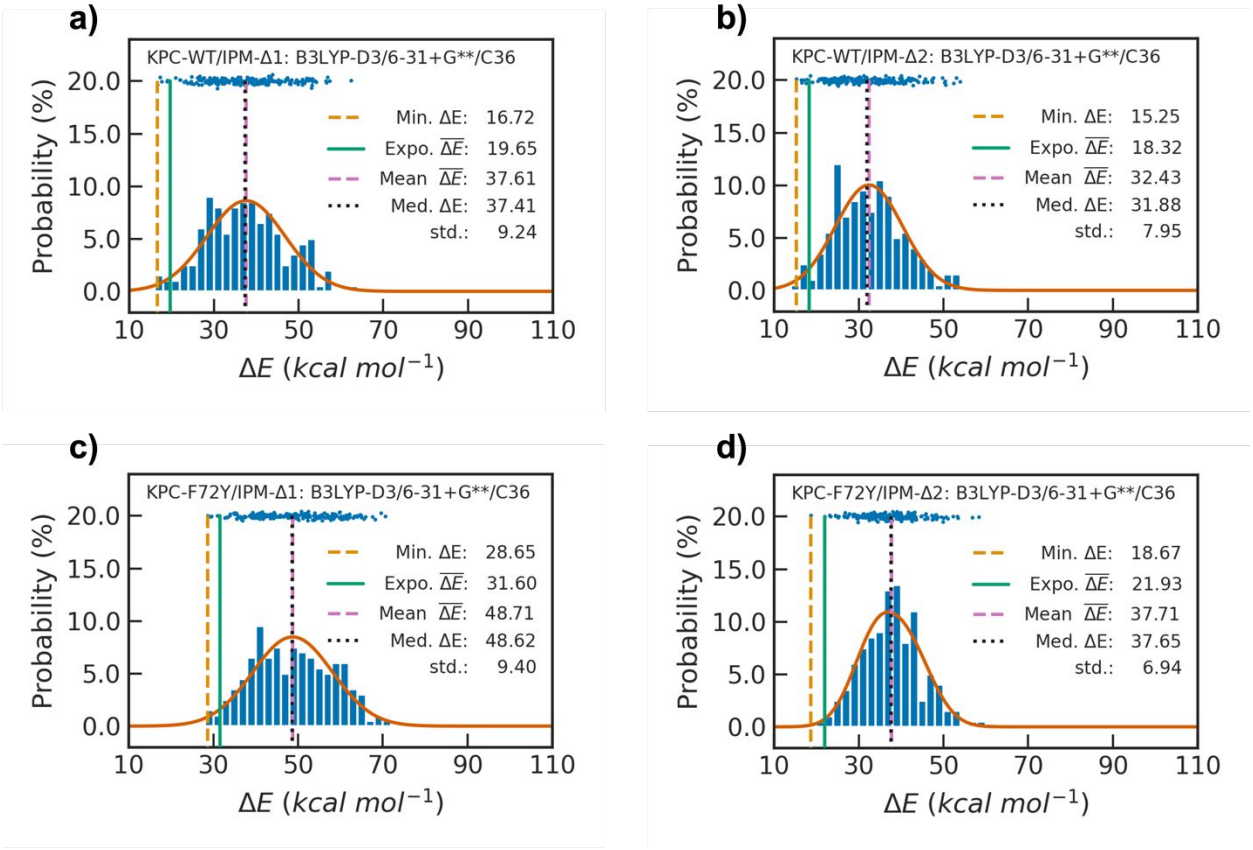

Fig. 4. The distribution of the deacylation barriers of the four modeled systems at B3LYP-D3(BJ)/6-31+G**/C36 level of theory. (a) KPC-WT/IPM-Δ1; (b) KPC-WT/IPM-Δ2; (c) KPC-F72Y/IPM-Δ1; and (d) KPC-F72Y/IPM-Δ2. Min, Expo, Mean, Med, and std refers to the minimum, the exponential average, the mean average, the median, and the standard deviation of the barrier energies, respectively.

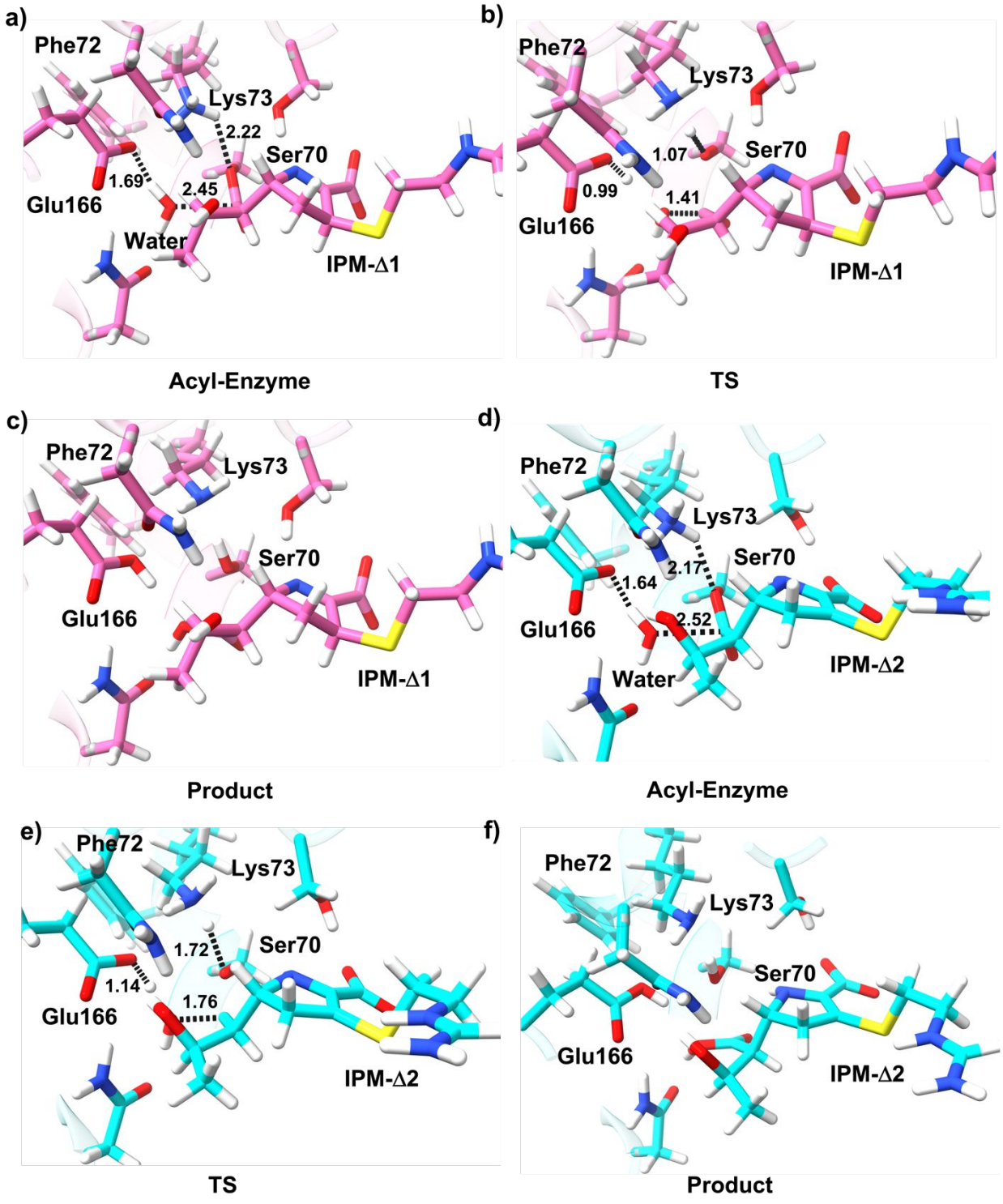

Fig. 5. Active site structure of MEPs with lowest barrier energy for KPC-WT/IPM-Δ1(pink) and KPC-WT/IPM-Δ2 (cyan). TS refers to transition state (TS). Three important distances (Glu166

Oε2 - Water H1, IPM C7 - Water O, Lys73 Hζ1- Ser70 Oγ), which involves the proton transfer, and nucleophilic attack, are marked as the black dashed line with Å unit. The carbon atoms are colored as pink in KPC-WT/IPM-Δ1 and cyan in KPC-WT/IPM-Δ2. The hydrogen, nitrogen, oxygen, and sulfur atoms are colored as white, blue, red, and yellow respectively. The minimum MEPs pathway is number 100 for KPC-WT/ IPM-Δ2 and number 37 for KPC-WT/ IPM-Δ1.

Experimental enzyme kinetic studies[15,17] have shown that the deacylation rates (k_3) for the wild type and Phe72Tyr mutant KPC-2 for IPM are 56 s^{-1} and 0.02 s^{-1}, which approximates deacylation free energy barriers of 15.05 kcal mol^{-1} and 19.77 kcal mol^{-1}, respectively. In our calculations, the minimum deacylation barriers on the IPM-Δ2 states are 15.25 kcal mol^{-1} for the KPC-WT systems and 18.67 kcal mol^{-1} for KPC-F72Y systems. Moreover, the exponential averaged barriers of the KPC-WT MEPs are lower than those of the KPC-F72Y systems, which is in agreement with the experimental observations.

Additionally, the minimum barrier energy heights of the IPM-Δ2 states are lower than those of the IPM-Δ1 states (16.72 kcal mol^{-1} and 28.65 kcal mol^{-1}). The IPM-Δ2 systems have a smaller exponential averaged barrier compared to the IPM-Δ1 systems MEPs. Notably, the significant reaction barrier difference between the IPM-Δ1 and IPM-Δ2 pathways in the KPC-F72Y system demonstrates that the Δ2 form is preferentially hydrolyzed, agreeing with observations for other Class A β-lactamases.[23–25] In contrast, the barrier energy difference between the IPM-Δ1 and IPM-Δ2 pathways in the KPC-WT system seems trivial. We also carried out free energy calculations using a thermodynamic integration (TI) method to estimate the free energy difference between the complexes with IPM-Δ1 and IPM-Δ2, for the wild type and mutant, respectively. Our calculations show that the complexes with IPM-Δ1 are slightly lower than the complexes with IPM-Δ2 for both

wild type and mutant, in agreement with an experimental finding that the free energy difference between these two tautomeric states is small[87] (See Supporting Information for details).

Machine Learning and Model Interpretation

The XGBoost model predicts the deacylation barrier heights with the coefficient of determination (R^2) above 0.9 and MAE lower than 1.3 kcal mol^{-1} (Fig. 6). The SHAP method is employed to analyze the impact of every element in the feature vector on the deacylation reaction (Fig. 7). For the hydrogen bonds represented by the features d2 (Lys73 Hζ2 – Glu166 Oε2), d3 (DW H1 – Glu166 Oε2), d4 (DW H1 – Glu166 Oε1), d5 (Asn170 Hδ2 – Glu166 Oε1), d6 (DW H2 – Asn170 Oδ), d8 (Lys73 Hζ2 – Asn132 Oδ), d12 (IPM 6αOH – Asn132 Oδ), d13 (IPM 6αOH – Glu166 Oε1), and d14 (IPM 6αOH – Glu166 Oε2), their SHAP values distributions show that the formations of these hydrogen bonds, which are indicated by the blue points (the shorter hydrogen – acceptor distances and therefore stronger hydrogen bonding interactions), contributes negatively to the deacylation barrier (Fig. 7). On the other hand, the hydrogen bonding features d7 (Lys73 Hζ1 – DW O), d11 (IPM 6αOH – DW O), and d15 (Lys73 Hζ1 – Ser130 Oγ) have SHAP values distributions opposite to those of the above-mentioned features, implying that the hydrogen bonds related to features d7, d11, and d15, lead to increase of the barrier energy. As for the key reacting coordinates, the positive SHAP values of the nucleophilic attack of DW on IPM β-lactam carbonyl (d9, DW O – IPM C7) on the samples with longer distance (red points) demonstrate the increase of the reaction barrier.

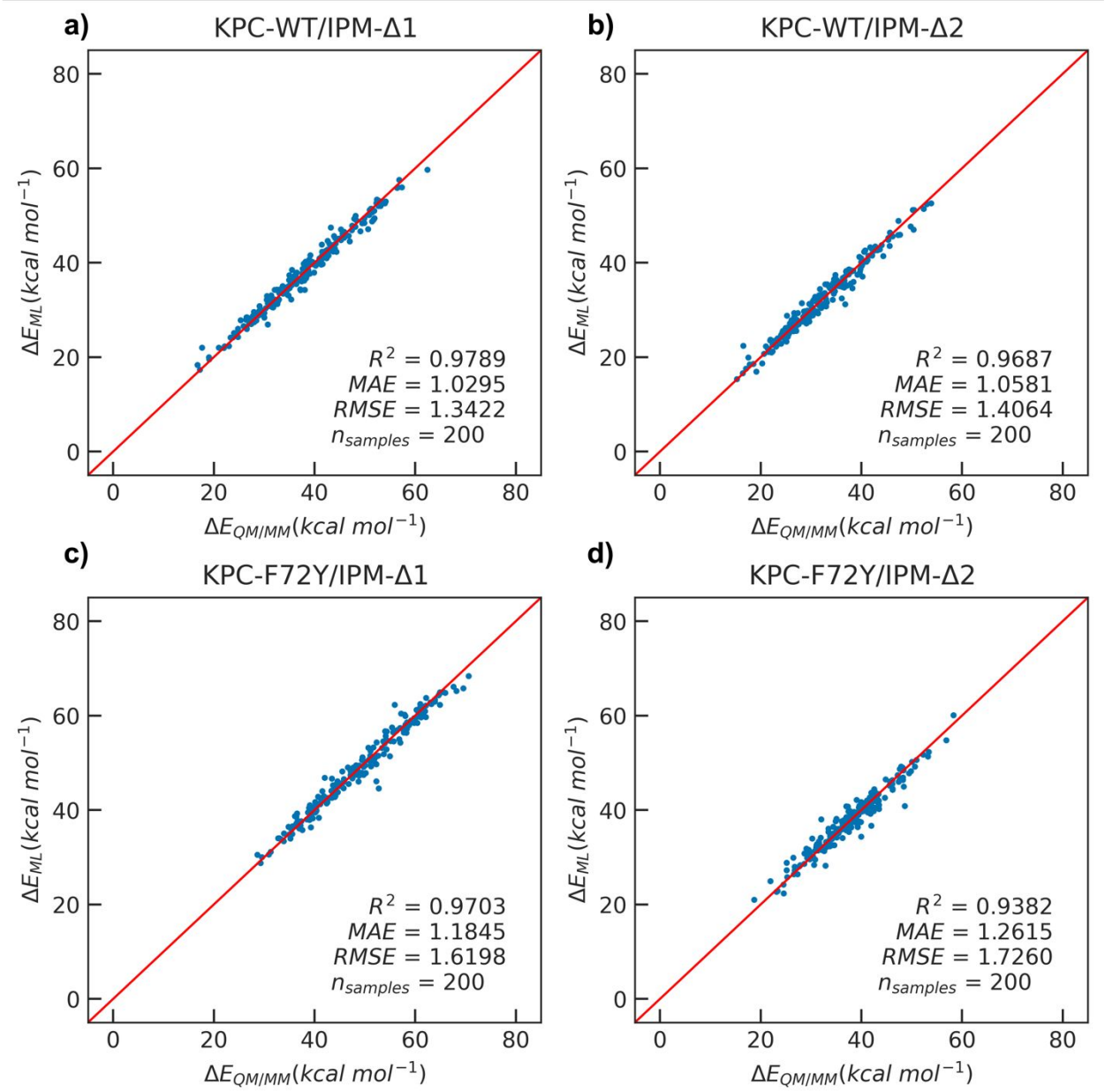

Fig. 6. The XGBoost model performances on four systems. (a) Model performances on the KPC-WT/IPM-Δ1 system; (b) the KPC-WT/IPM-Δ2 system; (c) the KPC-F72Y/IPM-Δ1 system; (d) the KPC-F72Y/IPM-Δ2 system. R^2, MAE, RMSE, $n_{samples}$ refer to the coefficient of determination, the mean absolute error, the root-mean-squared error, and the number of samples, respectively.

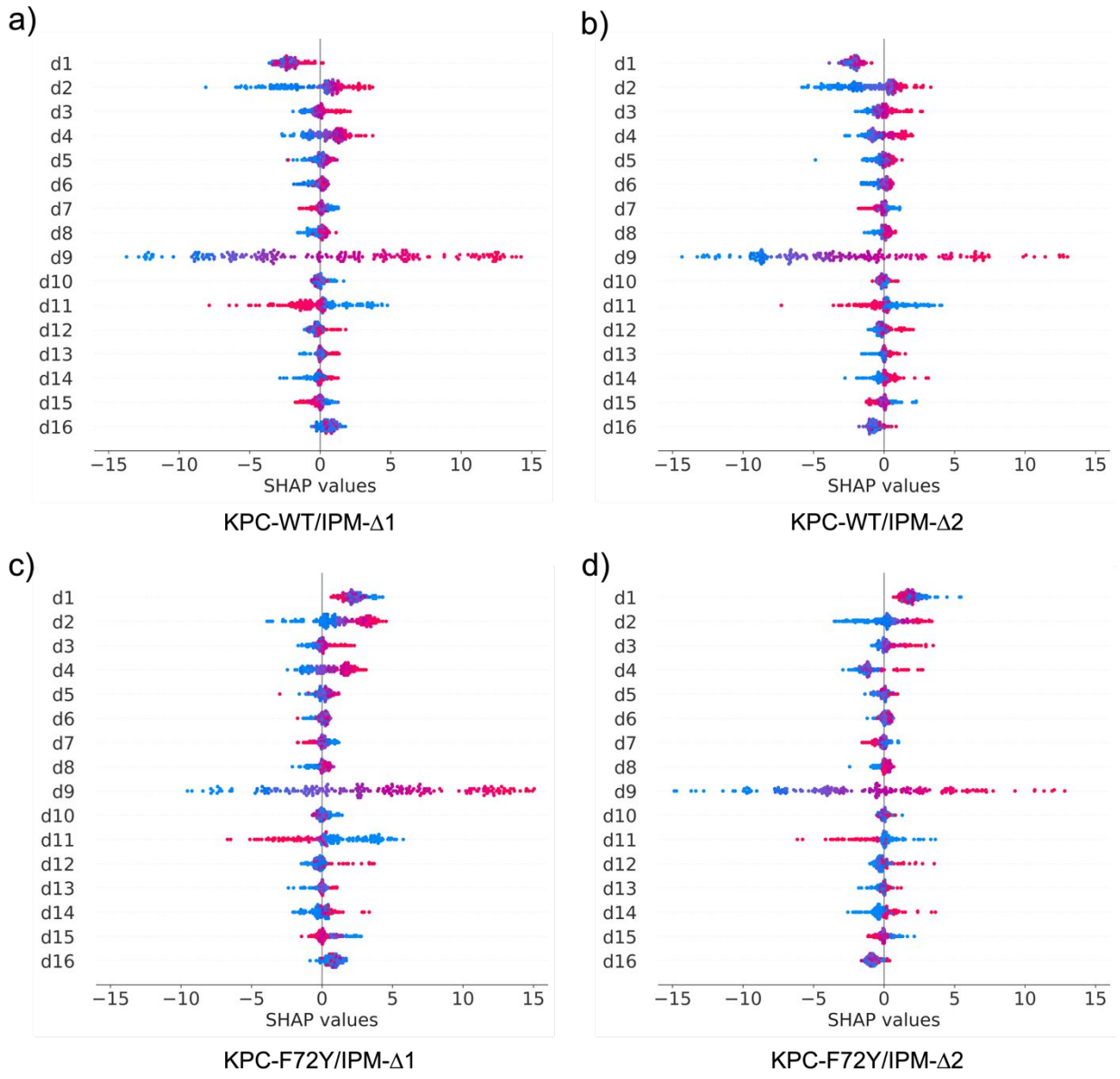

Fig. 7. The SHAP values obtained from the XGBoost model. The SHAP values on each structural feature of the conformation data from (a) the KPC-WT/IPM-Δ1 system, (b) the KPC-WT/IPM-Δ2 system; (c) the KPC- F72Y /IPM-Δ1 system; and (d) the KPC-F72Y/IPM-Δ2 system. The *x*-axis is the SHAP values attributed to each sample. The values of the atomic distances (the structural features) are noted by the color scheme of the scatter plots: from the shorter (blue) to longer (red) values of the distances.

Noteworthily, the SHAP values for feature d1 (Phe72 Hζ– Glu166 Oε2) in the KPC-WT systems are all negative for all sampled distances (Fig. 7.a and Fig. 7.b). In contrast, the SHAP values of d1 are all positive for the KPC-F72Y systems (Fig. 7.c and Fig. 7.d). This observation suggests that the extra hydrogen bond represented by d1 in the KPC-F72Y systems leads to an increase of the barrier energy and slows the deacylation rate. On the other hand, the feature d16 (Ser130 Hγ – IPM N4) has a similar SHAP values distribution regardless of their distance. Specifically, for the KPC-WT/IPM-Δ1 and KPC-F72Y/IPM-Δ1 systems, the major distribution of d16 has negative SHAP values (Fig. 7.a and Fig. 7.c). For the KPC-WT/IPM-Δ2 and KPC-F72Y/IPM-Δ2 systems, the major distribution of d16 has positive SHAP values (Fig. 7.b and Fig. 7.d). This suggests that the hydrogen bond between IPM N4 and Ser130 Oγ represented by feature d16 in the IPM-Δ1 systems leads to an increase of barrier energies. On the other hand, the weaker forms of this hydrogen bond in the IPM-Δ2 systems, indicated by longer distances of d16 (Fig. 8.a), leads to decrease of barrier energies.

Mean absolute SHAP values were calculated to quantify the overall feature contributions and to determine the dominant structural factors for the deacylation reaction (Fig. S8). The features d9 (DW O – IPM C7), d1 (Phe72 Hζ (Tyr72 Hη) – Glu166 Oε2), d11 (IPM 6αOH – DW O), d2 (Lys73 Hζ2 – Glu166 Oε2), d4 (IPM 6αOH – Glu166 Oε2), and d16 (Ser130 Hγ - IPM N4) are the top 6 most important features to the barrier energy in all four systems. Interestingly, these six features contain the three most important structural factors regulating the deacylation reaction of KPC-2/IPM hydrolysis: the local environment of Glu166, IPM pyrroline tautomerization, and the IPM 6α hydroxyethyl orientation.

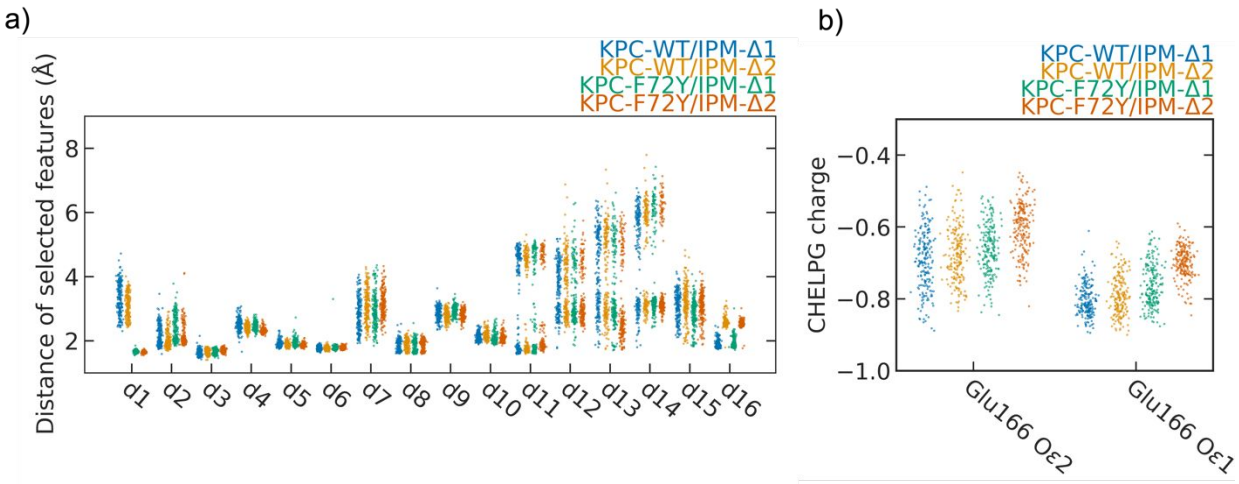

Fig. 8. Distance distribution of the selected features and the atomic charge of Glu166 carboxyl oxygen atoms. (a) The distance distribution for the selected features for the four systems. Labels on the x-axis represent the selected features. (b) The CHELPG charges for two oxygen atoms in the Glu166 carboxyl.

The impact of local environment around Glu166

The general base residue for the deacylation, Glu166, could accept hydrogen bonds from multiple neighbor residues, including Lys73 and Asn170, as well as Tyr72 in the KPC-F72Y system. These hydrogen bonds can be divided into two classes based on their effect on the deacylation step of KPC-2/IPM hydrolysis: favorable or unfavorable to the deacylation reaction.

The existence of a potential hydrogen bond donated from Tyr72 to Glu166 in the KPC-F72Y system was proved by the less than 2 Å distance distribution of feature d1 (Tyr72 Hη – Glu166 Oε2) (Fig. 8.a). The effect of deacylation inhibition of the additional hydrogen was first confirmed by the QM/MM MEPs barrier energy profile (Fig. 4). The negative SHAP values for feature d1 in the KPC-WT system and all positive SHAP values of this feature in the KPC-F72Y system show that this hydrogen bond hinders the deacylation reaction in the KPC-F72Y system (Fig. 7). Further

analysis utilizing the CHELPG charge populations[88] for the Glu166 carboxyl was performed to investigate how the additional hydrogen bond affects the Glu166 residue. The atomic charge analysis shows that the Glu166 carboxyl oxygen atoms have less negative charge in the KPC-F72Y systems (Fig. 8.b). This indicates that the hydrogen bond between Glu166 and Tyr72 reduces the basicity of Glu166 as the general base, which not only aligns with the work of Furey *et al*.[17], but also is consistent with the Hirvonen *et al.* finding that the hydration of the general base in the active site reduces the deacylation efficiency[20,21].

On the other hand, the features d2 (Lys73 Hζ2 – Glu166 Oε2), d3 (DW H1 – Glu166 Oε2), and d4 (DW H1 – Glu166 Oε1) are shown to favor the deacylation. Hata *et al.*[89] proposed that the Glu166 – Lys73 – Ser70 hydrogen bond network plays an important role in the proton migration in the deacylation step, where the hydrogen bond between Glu166 and Lys73 helps the deacylation reaction. In this study, the effect of deacylation assistance for the hydrogen bond donating from Lys73 to Glu166 is indicated by the negative SHAP values (the decrease of barrier energies) for the blue points (stronger hydrogen bonding interaction) of the feature d2. The features d3 (DW H1 – Glu166 Oε2) and d4 (DW H1 – Glu166 Oε1) also favor the deacylation reaction as they reduce the proton migration distance between the catalytic water and Glu166.

Table 2. Average distance (Å) between atoms involving in the deacylation step.

System	KPC-WT/IPM-Δ1	KPC-WT/IPM-Δ2	KPC-F72Y/IPM-Δ1	KPC-F72Y/IPM-Δ2
IPM 6αOH – DW O (d11)	3.47	3.25	2.87	2.74
DW O – Lys73 Hζ1 (d7)	2.82	3.11	2.92	3.16
DW O – IPM C7 (d9)	2.87	2.80	2.95	2.83

The impact from the tautomerization states of the IPM pyrroline

The IPM pyrroline ring could undergo tautomerization during the formation of acyl-enzyme and generates two potential tautomers IPM-Δ1 and IPM-Δ2 for the deacylation. The barrier heights based on the QM/MM MEPs calculations show that the IPM-Δ2 system is more active than the IPM-Δ1 system in KPC-F72Y mutant, while the deacylated products in both tautomer states can be produced in the KPC-WT system.

Three structural features d7 (Lys73 Hζ1 – DW O), d9 (DW O – IPM C7), and d16 (Ser130 Hγ – IPM N4) are shown to be correlated with the IPM pyrroline tautomerization states. Among them, the feature d9 (DW O – IPM C7) is the most significant factor in the deacylation reaction of KPC-2/IPM hydrolysis due to its largest mean absolute SHAP values. The feature d9 in the IPM-Δ2 systems leads to more decrease of barrier energy than in the IPM-Δ1 systems, which is demonstrated by that the d9 in the IPM-Δ2 systems have more samples with negative SHAP values than those in the IPM-Δ1 systems (Fig. 7). The comparison of feature d9 also reveals the effect of pyrroline ring tautomerization on the nucleophilic attack distance. The mean nucleophilic attack distance (d9) is 2.80 Å and 2.83 Å for the KPC-WT/IPM-Δ2 and KPC-F72Y/IPM-Δ2 systems, respectively, which are smaller than those in the KPC-WT/IPM-Δ1 (2.87 Å) and KPC-F72Y/IPM-Δ1 (2.95 Å) systems (

Table **2**).

Besides the nucleophilic attack distance, feature d7 (Lys73 Hζ1 – DW O) is another important factor which behaves differently in the IPM-Δ1 and IPM-Δ2 systems. More positive SHAP values appear in the IPM-Δ1 states than in the IMP-Δ2 states (Fig. 7). The values of feature d7 in the IPM-Δ1 systems (2.82 Å and 2.92 Å for the KPC-WT/IPM-Δ1 and KPC-F72Y/IPM-Δ1 systems,

respectively) are also smaller than those in the IPM-Δ2 systems (3.11 Å and 3.16 Å for the KPC-WT/IPM-Δ2 and KPC-F72Y/IPM-Δ2 systems respectively), indicating a stronger interaction between Lys73 and the catalytic water in the IPM-Δ1 systems (

Table **2**). This shorter interacting distances consequently lead to an increase of barrier energy for the deacylation reaction. These results reveal the different contributions from feature d7 to activities of KPC wild type and mutant against imipenem in different tautomerization states.

Feature d16, the distance between IPM N4 and Ser130, is yet another feature influenced by the pyrroline ring tautomerization. Observation of d16 distances shows that the hydrogen bond between Ser130 Oγ and IPM N4 is stronger in the IPM-Δ1 systems. This hydrogen bond hinders the deacylation reaction as most samples of the feature d16 have positive SHAP values in the IPM-Δ1 systems (Fig. 7.a and Fig. 7.c).

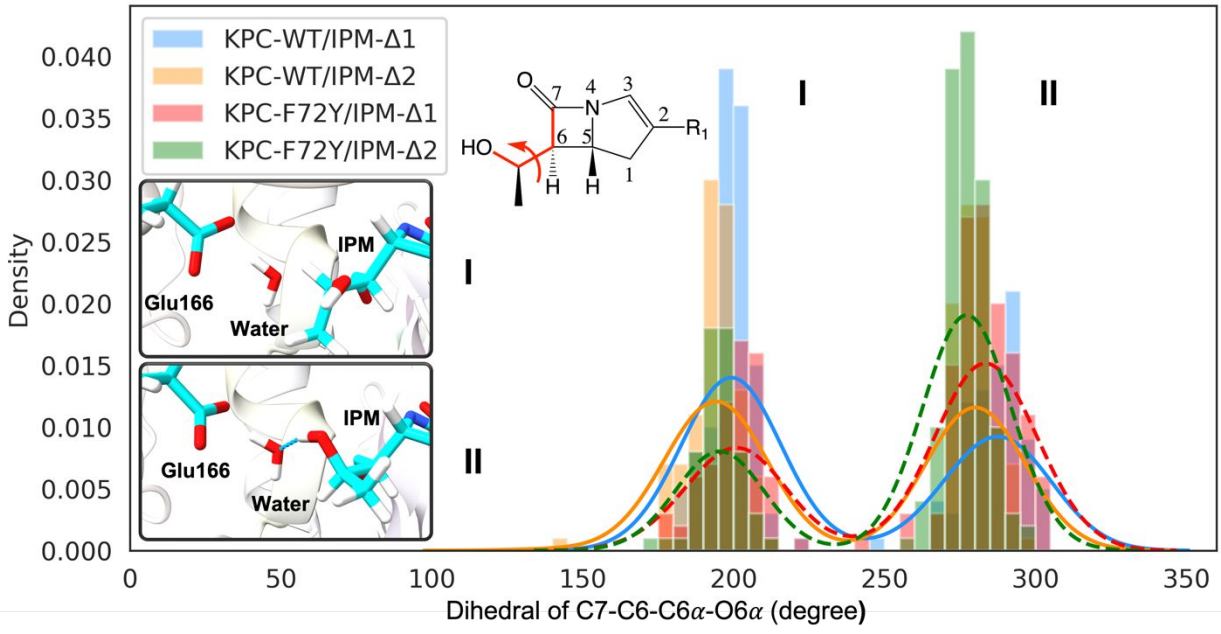

Fig. 9. Orientations of the 6α-hydroxyethyl group in KPC-2/IPM acyl-enzyme conformations. 1). The blue, yellow, red, and green histograms represent the density distribution of the C7-C6-C6α-

O6α dihedral angle in KPC-WT/IPM-Δ1, KPC-WT/IPM-Δ2, KPC-F72Y/IPM-Δ1 and KPC-F72Y/IPM-Δ2, respectively. The dashed blue, dashed yellow, solid red, and solid green lines are the density distribution of this dihedral angle in KPC-WT/IPM-Δ1, KPC-WT/IPM-Δ2, KPC-F72Y/IPM-Δ1, and KPC-F72Y/IPM-Δ2 systems, respectively. 2). The panel of conformation I represents the C7-C6-C6α-O6α dihedral angle of 200°. In this state, the IPM 6α hydroxyethyl group forms hydrogen bonds with solvent molecules. The carbon, nitrogen, hydrogen, and oxygen atoms are colored as cyan, blue, white, and red, respectively. The panel of conformation II represents the C7-C6-C6α-O6α dihedral angle of 280°. In this state, the IPM 6α hydroxyethyl group serves as a hydrogen bond donor to the deacylation water. The color scheme of panel conformation II is the same as the panel conformation I. 3). The IPM structure is shown with the dihedral of C7-C6-C6α-O6α highlighted in red.

The impact from the orientation of IPM 6α hydroxyethyl

The IPM 6α hydroxyethyl has been proposed to play an important role in regulating the deacylation step of carbapenem hydrolysis[21,28]. Recently, Chudyk *et al.*[28] reported that the acylated carbapenem in ASβL-carbapenemases adopts two main 6α hydroxyethyl orientations in the acyl-enzyme complex. Accordingly, both orientations of IPM 6α hydroxyethyl group were observed for all four systems. These two orientations were represented by the dihedral angle values of C7-C6-C6α-O6α around 200° and 280°, respectively, referred to as conformations I and II (Fig. 9). In conformation I, the IPM 6α hydroxyethyl group adapts an orientation, donating a hydrogen bond to solvent molecules. In conformation II, the IPM 6α hydroxyethyl group mainly hydrogen bonds with the catalytic water. It also forms hydrogen bonds with Glu166 and Asn132 in some

snapshots in the MD simulation as conformation II. Note that those two conformations were also observed in our previous work[32], the IPM hydrolysis by GES-5, in which Conformation I with dihedral of C7-C6-C6α-O6α resting at 210° was found to be the preferred state. Interestingly, more conformation II states with the hydrogen bond to catalytic water mentioned above are observed in the acyl-enzyme reactant of the KPC-F72Y systems than in the KPC-WT systems. The observation reveals that the mutation of Phe72Tyr in the KPC-F72Y systems leads a tight interaction between the IPM 6α hydroxyethyl group and the catalytic water through a hydrogen bond.

Four features (d11, d12, d13, and d14) were selected to represent the interactions between the IPM 6α hydroxyethyl group and the residues in the active site. The feature d11 (IPM O6*α* – DW O) was selected as the representative interaction due to its larger mean absolute SHAP values than the other three features in all four systems. Positive SHAP values for blue points and negative SHAP values for red points of the feature d11 indicate that the existence of this hydrogen bond between the IPM 6α hydroxyethyl group and the catalytic water deactivates the catalytic water and slows the deacylation rate with the increased barrier energy (Fig. 7). This finding is consistent with the work of Hirvonen *et al.* [21], that the formation of the hydrogen bond between the catalytic water and the 6α-hydroxyethyl group of carbapenem is unfavorable for carbapenem hydrolysis by the OXA-48 β-lactamase. Additionally, there are more negative points of the d11 feature in the KPC-WT systems than those in the KPC-F72Y systems, suggesting the mutation of Phe72Tyr in the KPC-F72Y systems increase the barrier energy contributed by the feature d11 compared with those in the KPC-WT systems.

The impact of local environments around the catalytic water

Two groups of features representing the interaction between catalytic water and residues in the active site were found with different impact. Features d3 (DW H1 – Glu166 O*ε*2), d4 (DW H1

– Glu166 Oε1), and d6 (DW H2 – Asn170 Oδ) favor the deacylation reaction given their SHAP values distribution. The interactions between catalytic water and Glu166, Asn170 are suggested to reduce the barrier energy with their negative SHAP values for short distances and positive values for long distances (Fig. 7). On the other hand, features d7 (Lys73 Hζ1 – DW O), d9 (DW O – IPM C7), and d11(IPM 6αOH – DW O) could inhibit the deacylation due to the opposite SHAP values distribution. Features d7 and d11 represent hydrogen bonds with catalytic water as acceptor and hinder the deacylation reaction. On the contrary, features d3, d4, and d6 represent hydrogen bonds with catalytic water as donor and are beneficial to the deacylation reaction. Therefore, it is suggested that the hydrogen bond interactions with catalytic water as acceptor impair the nucleophilic attack to the tetrahedral intermediate and slow the deacylation reaction rate. Conversely, the hydrogen bonds with catalytic water as the donor help the proton migration and favor the deacylation reaction.

Conclusions

In this study, we investigate the deacylation reaction of KPC-2/IPM hydrolysis using QM/MM calculations. 800 QM/MM MEPs of deacylation reactions for four systems (KPC-WT/IPM-Δ1, KPC-WT/IPM-Δ2, KPC-F72Y/IPM-Δ1 and KPC-F72Y/IPM-Δ2), were calculated. Our QM/MM calculations show that not only the Phe72Tyr mutation but also the IPM tautomerization leads to a higher barrier energy for the KPC-2/IPM deacylation (though the IPM-Δ1 hydrolysis is still energetically favorable in the KPC-WT).

We further applied the XGBoost model assisted by the SHAP method to analyze the barrier energies using conformational features of the acyl-enzyme reactant states in order to provide insight into the mechanism of the deacylation reaction of KPC-2/IPM. The effect of specific features and the dominant factors of the deacylation reaction could be determined by the mean

absolute SHAP values as well as their distributions. We identified three factors highly impacts the deacylation reaction of KPC-2/IPM hydrolysis based on the ML model. First, Tyr72 forms an additional hydrogen bond with Glu166 in the mutant KPC-F72Y system. This hydrogen bond is shown to inhibit the deacylation step by reducing the basicity of the general base. Second, the tautomerization states on the ligand pyrroline rings is correlated with the hydrogen bonding interactions between Lys73 and the DW. The IPM-Δ2 tautomer has been previously proposed to stabilize the tetrahedral intermediate during the deacylation[32]. Meanwhile, the hydrogen bonding donated by Lys73 (as in IPM-Δ1 states) would decrease the nucleophilicity of the DW. Both effects would synergistically contribute to the deacylation inefficiency observed in the IPM-Δ1 systems, especially in the KPC-F72Y mutant. Third, the IPM 6α-hydroxyethyl group adapts two orientations. In one orientation, the hydrogen bond to catalytic water hampers the deacylation step by causing a longer nucleophilic attack distance. In additional, this orientation is more often observed in the KPC-F72Y systems showing that the local environment changes of Glu166 also have significant impacts on the orientation of IPM 6α hydroxyethyl group. Hydrogen bonds formed between the catalytic water and the IPM 6α-hydroxyethyl group as well as Lys73 collectively regulate the catalytic water behaviors.

Lastly, in this study, we showed that the combination of the XGBoost model and the SHAP method could effectively assist the analysis of KPC-2/IPM hydrolysis QM/MM MEPs and provide the mechanistic insights into different interactions in the active site. Finally, our study demonstrates the potential of explainable Machine Learning for understanding the mechanism of enzyme catalysis.

Data availability

All data and codes reported in the current study are publicly available at DOI: 10.5281/zenodo.7114981.

Author contributions

C.Y.: Conceptualization, methodology, formal analysis, visualization, data curation, writing original draft, writing – review and editing; Z.S.: Conceptualization, methodology, writing – review and editing; T.P.: Writing – review and editing; P.T.: Supervision, conceptualization, writing – review and editing, project administration, and funding acquisition.

Acknowledgement

This material is based upon work supported by the National Science Foundation under a CAREER Grant No. 1753167. T.P. is supported by NIH grant AI32956. Computational time was provided by the Southern Methodist University's Centre for Research Computing.

Conflicts of Interests

The authors declare no competing interests.

Reference

1 K. M. Papp-Wallace, A. Endimiani, M. A. Taracila and R. A. Bonomo, *Antimicrob. Agents Chemother.*, 2011, **55**, 4943–4960.
2 J. Walther-Rasmussen and N. Høiby, *J. Antimicrob. Chemother.*, 2007, **60**, 470–482.
3 T. Palzkill, *Front. Mol. Biosci.* 2018, 5.
4 A. M. Queenan and K. Bush, *Clin. Microbiol. Rev.*, 2007, **20**, 440–458.
5 F. Fonseca, E. I. Chudyk, M. W. van der Kamp, A. Correia, A. J. Mulholland and J. Spencer, *J. Am. Chem. Soc.*, 2012, **134**, 18275–18285.
6 O. A. Pemberton, X. Zhang and Y. Chen, *J. Med. Chem.*, 2017, **60**, 3525–3530.
7 N. P. Krishnan, N. Q. Nguyen, K. M. Papp-Wallace, R. A. Bonomo and F. van den Akker, *PLOS ONE*, 2015, **10**, e0136813.
8 M.-N. Lisa, A. R. Palacios, M. Aitha, M. M. González, D. M. Moreno, M. W. Crowder, R. A. Bonomo, J. Spencer, D. L. Tierney, L. I. Llarrull and A. J. Vila, *Nat. Commun.*, 2017, **8**, 538.
9 S. BOUNAGA, A. P. LAWS, M. GALLENI and M. I. PAGE, *Biochem. J.*, 1998, **331**, 703–711.
10 G. Bahr, L. J. González and A. J. Vila, *Chem. Rev.*, 2021, **121**, 7957–8094.

11 C. K. Das and N. N. Nair, *Phys. Chem. Chem. Phys.*, 2017, **19**, 13111–13121.
12 R. Tripathi and N. N. Nair, *ACS Catal.*, 2015, **5**, 2577–2586.
13 L. Maveyraud, D. Golemi-Kotra, A. Ishiwata, O. Meroueh, S. Mobashery and J.-P. Samama, *J. Am. Chem. Soc.*, 2002, **124**, 2461–2465.
14 V. Thakkur, C. K. Das and N. N. Nair, *ACS Catal.*, 2022, **12**, 10338–10352.
15 S. C. Mehta, I. M. Furey, O. A. Pemberton, D. M. Boragine, Y. Chen and T. Palzkill, *J. Biol. Chem.*, 2021, **296**, 100155.
16 H. Yigit, A. M. Queenan, G. J. Anderson, A. Domenech-Sanchez, J. W. Biddle, C. D. Steward, S. Alberti, K. Bush and F. C. Tenover, *Antimicrob. Agents Chemother.*, 2001, **45**, 1151–1161.
17 I. M. Furey, S. C. Mehta, B. Sankaran, L. Hu, B. V. V. Prasad and T. Palzkill, *J. Biol. Chem.*, 2021, **296**, 100799.
18 H. Frase, M. Toth, M. M. Champion, N. T. Antunes and S. B. Vakulenko, *Antimicrob. Agents Chemother.*, 2011, **55**, 1556–1562.
19 S. D. Kotsakis, V. Miriagou, E. Tzelepi and L. S. Tzouvelekis, *Antimicrob. Agents Chemother.*, 2010, **54**, 4864–4871.
20 V. H. A. Hirvonen, A. J. Mulholland, J. Spencer and M. W. van der Kamp, *ACS Catal.*, 2020, **10**, 6188–6196.
21 V. H. A. Hirvonen, T. M. Weizmann, A. J. Mulholland, J. Spencer and M. W. van der Kamp, *ACS Catal.*, 2022, **12**, 4534–4544.
22 P. S. Levitt, K. M. Papp-Wallace, M. A. Taracila, A. M. Hujer, M. L. Winkler, K. M. Smith, Y. Xu, M. E. Harris and R. A. Bonomo, *J. Biol. Chem.*, 2012, **287**, 31783–31793.
23 M. Kalp and P. R. Carey, *Biochemistry*, 2008, **47**, 11830–11837.
24 R. L. Charnas and J. R. Knowles, *Biochemistry*, 1981, **20**, 2732–2737.
25 L. W. Tremblay, F. Fan and J. S. Blanchard, *Biochemistry*, 2010, **49**, 3766–3773.
26 J. C. Hermann, C. Hensen, L. Ridder, A. J. Mulholland and H.-D. Höltje, *J. Am. Chem. Soc.*, 2005, **127**, 4454–4465.
27 E. I. Chudyk, M. A. L. Limb, C. Jones, J. Spencer, M. W. van der Kamp and A. J. Mulholland, *Chem Commun*, 2014, **50**, 14736–14739.
28 E. I. Chudyk, M. Beer, M. A. L. Limb, C. A. Jones, J. Spencer, M. W. van der Kamp and A. J. Mulholland, *ACS Infect. Dis.*, 2022, **8**, 1521–1532.
29 Z. Song, H. Zhou, H. Tian, X. Wang and P. Tao, *Commun. Chem.*, 2020, **3**, 134.
30 Z. Song, F. Trozzi, T. Palzkill and P. Tao, *Org. Biomol. Chem.*, 2021, **19**, 9182–9189.
31 Z. Song, F. Trozzi, H. Tian, C. Yin and P. Tao, *ACS Phys. Chem. Au*, 2022, **2**, 316-330.
32 Z. Song and P. Tao, *Electron. Struct.*, 2022, **4**, 034001.
33 J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev and A. E. Roitberg, *Nat. Commun.*, 2019, **10**, 2903.
34 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 13890.
35 P. Bleiziffer, K. Schaller and S. Riniker, *J. Chem. Inf. Model.*, 2018, **58**, 579–590.
36 M. J. Latallo, G. A. Cortina, S. Faham, R. K. Nakamoto and P. M. Kasson, *Chem. Sci.*, 2017, **8**, 6484–6492.
37 G. A. Cortina, J. M. Hays and P. M. Kasson, *ACS Catal.*, 2018, **8**, 2741–2747.
38 S. B. Kotsiantis, I. D. Zaharakis and P. E. Pintelas, *Artif. Intell. Rev.*, 2006, **26**, 159–190.
39 M. I. Jordan and T. M. Mitchell, *Science*, 2015, **349**, 255–260.
40 S. B. Kotsiantis, I. Zaharakis and P. Pintelas, *Emerg. Artif. Intell. Appl. Comput. Eng.*, 2007, **160**, 3–24.

41 A. Singh, N. Thakur and A. Sharma, in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 1310–1315.
42 D. W. Hosmer Jr, S. Lemeshow and R. X. Sturdivant, *Applied logistic regression*, John Wiley & Sons, 2013, vol. 398.
43 S. B. Kotsiantis, *Artif. Intell. Rev.*, 2013, **39**, 261–283.
44 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
45 J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua and A. Lopez, *Neurocomputing*, 2020, **408**, 189–215.
46 Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436–444.
47 A. C. Mater and M. L. Coote, *J. Chem. Inf. Model.*, 2019, **59**, 2545–2559.
48 P. O. Dral, *J. Phys. Chem. Lett.*, 2020, **11**, 2336–2347.
49 G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa and A. Fazzio, *J. Phys. Mater.*, 2019, **2**, 032001.
50 T. Chen and C. Guestrin, *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, 785–794.
51 R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma and E. M. Gifford, *J. Chem. Inf. Model.*, 2016, **56**, 2353–2360.
52 V. Svetnik, T. Wang, C. Tong, A. Liaw, R. P. Sheridan and Q. Song, *J. Chem. Inf. Model.*, 2005, **45**, 786–799.
53 S. Vargas, M. R. Hennefarth, Z. Liu and A. N. Alexandrova, *J. Chem. Theory Comput.*, 2021, **17**, 6203–6213.
54 J. Dong, L. Peng, X. Yang, Z. Zhang and P. Zhang, *J. Comput. Chem.*, 2022, **43**, 289–302.
55 Z. Wu, T. Lei, C. Shen, Z. Wang, D. Cao and T. Hou, *J. Chem. Inf. Model.*, 2019, **59**, 4587–4601.
56 M. T. Ribeiro, S. Singh and C. Guestrin, *Proc. AAAI Conf. Artif. Intell.*, 2018, **32**.
57 S. Wachter, B. Mittelstadt and C. Russell, 2018. DOI: 10.48550/arXiv.2010.10596
58 M. Sundararajan, A. Taly and Q. Yan, 2017. DOI: 10.48550/arXiv.1703.01365
59 S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, *Nat. Mach. Intell.*, 2020, **2**, 56–67.
60 S. M. Lundberg and S.-I. Lee, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017, vol. 30.
61 Y. Meng, N. Yang, Z. Qian and G. Zhang, *J. Theor. Appl. Electron. Commer. Res.*, 2021, **16**, 466–490.
62 S. Zhang, T. Lu, P. Xu, Q. Tao, M. Li and W. Lu, *J. Phys. Chem. Lett.*, 2021, **12**, 7423–7430.
63 W.-L. Ye, C. Shen, G.-L. Xiong, J.-J. Ding, A.-P. Lu, T.-J. Hou and D.-S. Cao, *J. Chem. Inf. Model.*, 2020, **60**, 4216–4230.
64 K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. Mackerell, *J. Comput. Chem.*, 2010, **31**, 671–690.
65 R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig and A. D. MacKerell, *J. Chem. Theory Comput.*, 2012, **8**, 3257–3273.
66 C. A. Smith, Z. Nossoni, M. Toth, N. K. Stewart, H. Frase and S. B. Vakulenko, *J. Biol. Chem.*, 2016, **291**, 22196–22206.
67 P. Swarén, L. Maveyraud, X. Raquet, S. Cabantous, C. Duez, J.-D. Pédelacq, S. Mariotte-Boyer, L. Mourey, R. Labia, M.-H. Nicolas-Chanoine, P. Nordmann, J.-M. Frère and J.-P. Samama, *J. Biol. Chem.*, 1998, **273**, 26714–26721.
68 M. Gaus, X. Lu, M. Elstner and Q. Cui, *J. Chem. Theory Comput.*, 2014, **10**, 1518–1537.

69 S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, *J. Chem. Phys.*, 1995, **103**, 4613–4621.
70 V. Kräutler, W. F. van Gunsteren and P. H. Hünenberger, *J. Comput. Chem.*, 2001, **22**, 501–508.
71 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J. Chem. Phys.*, 1995, **103**, 8577–8593.
72 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187–217.
73 B. R. Brooks, C. L. Brooks III, 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 and M. Karplus, *J. Comput. Chem.*, 2009, **30**, 1545–1614.
74 P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLoS Comput. Biol.*, 2017, **13**, e1005659.
75 J. B. Brokaw, K. R. Haas and J.-W. Chu, *J. Chem. Theory Comput.*, 2009, **5**, 2050–2061.
76 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
77 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
78 W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257–2261.
79 H. Schröder, A. Creon and T. Schwabe, *J. Chem. Theory Comput.*, 2015, **11**, 3163–3170.
80 Q. Cui, M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus, *J. Phys. Chem. B*, 2001, **105**, 569–585.
81 H. L. Woodcock, M. Hodošček and B. R. Brooks, *J. Phys. Chem. A*, 2007, **111**, 5720–5728.
82 E. Epifanovsky, A. T. B. Gilbert, X. Feng, J. Lee, Y. Mao, N. Mardirossian, P. Pokhilko, A. F. White, M. P. Coons, A. L. Dempwolff, Z. Gan, D. Hait, P. R. Horn, L. D. Jacobson, I. Kaliman, J. Kussmann, A. W. Lange, K. U. Lao, D. S. Levine, J. Liu, S. C. McKenzie, A. F. Morrison, K. D. Nanda, F. Plasser, D. R. Rehn, M. L. Vidal, Z.-Q. You, Y. Zhu, B. Alam, B. J. Albrecht, A. Aldossary, E. Alguire, J. H. Andersen, V. Athavale, D. Barton, K. Begam, A. Behn, N. Bellonzi, Y. A. Bernard, E. J. Berquist, H. G. A. Burton, A. Carreras, K. Carter-Fenk, R. Chakraborty, A. D. Chien, K. D. Closser, V. Cofer-Shabica, S. Dasgupta, M. de Wergifosse, J. Deng, M. Diedenhofen, H. Do, S. Ehlert, P.-T. Fang, S. Fatehi, Q. Feng, T. Friedhoff, J. Gayvert, Q. Ge, G. Gidofalvi, M. Goldey, J. Gomes, C. E. González-Espinoza, S. Gulania, A. O. Gunina, M. W. D. Hanson-Heine, P. H. P. Harbach, A. Hauser, M. F. Herbst, M. Hernández Vera, M. Hodecker, Z. C. Holden, S. Houck, X. Huang, K. Hui, B. C. Huynh, M. Ivanov, Á. Jász, H. Ji, H. Jiang, B. Kaduk, S. Kähler, K. Khistyaev, J. Kim, G. Kis, P. Klunzinger, Z. Koczor-Benda, J. H. Koh, D. Kosenkov, L. Koulias, T. Kowalczyk, C. M. Krauter, K. Kue, A. Kunitsa, T. Kus, I. Ladjánszki, A. Landau, K. V. Lawler, D. Lefrancois, S. Lehtola, R. R. Li, Y.-P. Li, J. Liang, M. Liebenthal, H.-H. Lin, Y.-S. Lin, F. Liu, K.-Y. Liu, M. Loipersberger, A. Luenser, A. Manjanath, P. Manohar, E. Mansoor, S. F. Manzer, S.-P. Mao, A. V. Marenich, T. Markovich, S. Mason, S. A. Maurer, P. F. McLaughlin, M. F. S. J. Menger, J.-M. Mewes, S. A. Mewes, P. Morgante, J. W. Mullinax, K. J. Oosterbaan, G. Paran, A. C. Paul, S. K. Paul, F. Pavošević, Z. Pei, S. Prager, E. I. Proynov, Á. Rák, E. Ramos-Cordoba, B. Rana, A. E. Rask, A. Rettig, R. M. Richard, F. Rob, E. Rossomme, T. Scheele, M. Scheurer, M. Schneider, N. Sergueev, S. M. Sharada, W. Skomorowski, D. W. Small, C. J. Stein, Y.-C. Su, E. J. Sundstrom, Z. Tao, J. Thirman, G. J. Tornai, T. Tsuchimochi, N. M. Tubman, S. P. Veccham, O. Vydrov, J. Wenzel, J. Witte, A. Yamada, K. Yao, S. Yeganeh, S. R. Yost, A. Zech, I. Y. Zhang, X. Zhang, Y. Zhang,

D. Zuev, A. Aspuru-Guzik, A. T. Bell, N. A. Besley, K. B. Bravaya, B. R. Brooks, D. Casanova, J.-D. Chai, S. Coriani, C. J. Cramer, G. Cserey, A. E. DePrince, R. A. DiStasio, A. Dreuw, B. D. Dunietz, T. R. Furlani, W. A. Goddard, S. Hammes-Schiffer, T. Head-Gordon, W. J. Hehre, C.-P. Hsu, T.-C. Jagau, Y. Jung, A. Klamt, J. Kong, D. S. Lambrecht, W. Liang, N. J. Mayhall, C. W. McCurdy, J. B. Neaton, C. Ochsenfeld, J. A. Parkhill, R. Peverati, V. A. Rassolov, Y. Shao, L. V. Slipchenko, T. Stauch, R. P. Steele, J. E. Subotnik, A. J. W. Thom, A. Tkatchenko, D. G. Truhlar, T. Van Voorhis, T. A. Wesolowski, K. B. Whaley, H. L. Woodcock, P. M. Zimmerman, S. Faraji, P. M. W. Gill, M. Head-Gordon, J. M. Herbert and A. I. Krylov, *J. Chem. Phys.*, 2021, **155**, 084801.

83 N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, *J. Comput. Chem.*, 2011, **32**, 2319–2327.

84 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

85 Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, 2015.

86 F. Chollet and others, 2015. https://keras.io. Access date: 06/01/2022.

87 C. T. Lohans, E. I. Freeman, E. van Groesen, C. L. Tooke, P. Hinchliffe, J. Spencer, J. Brem and C. J. Schofield, *Sci. Rep.*, 2019, **9**, 13608.

88 C. M. Breneman and K. B. Wiberg, *J. Comput. Chem.*, 1990, **11**, 361–373.

89 M. Hata, Y. Fujii, M. Ishii, T. Hoshino and M. Tsuda, *Chem. Pharm. Bull. (Tokyo)*, 2000, **48**, 447–453.