Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

Unveiling the Structural Features that Regulate Carbapenems Deacylation of KPC-2 via 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] 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.)

Supporting Figures

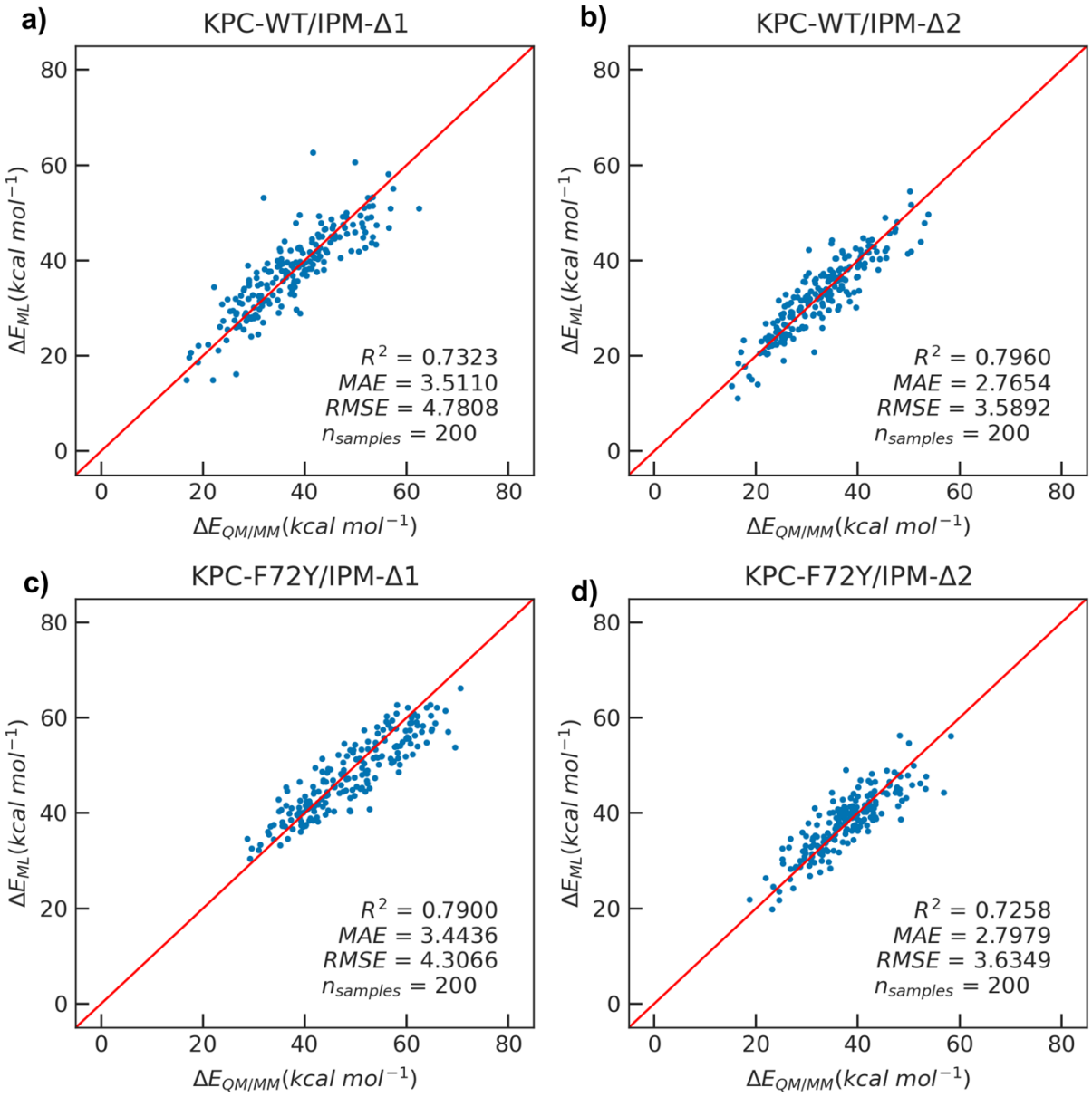

Fig. S1. The **linear regression** model performance on four systems. (a) Model performance on the KPC-WT/IPM-Δ1 system; (b) the KPC-WT/IPM-Δ2 system; (c) the KPC-F72Y/IPM-Δ1 system; (d) the KPC-F72Y/IPM-Δ2 system. R^2, MAE, RMSE, $n_{samples}$ refer to the coefficient of determination, the mean absolute error, the root-mean-squared error, and the number of samples, respectively.

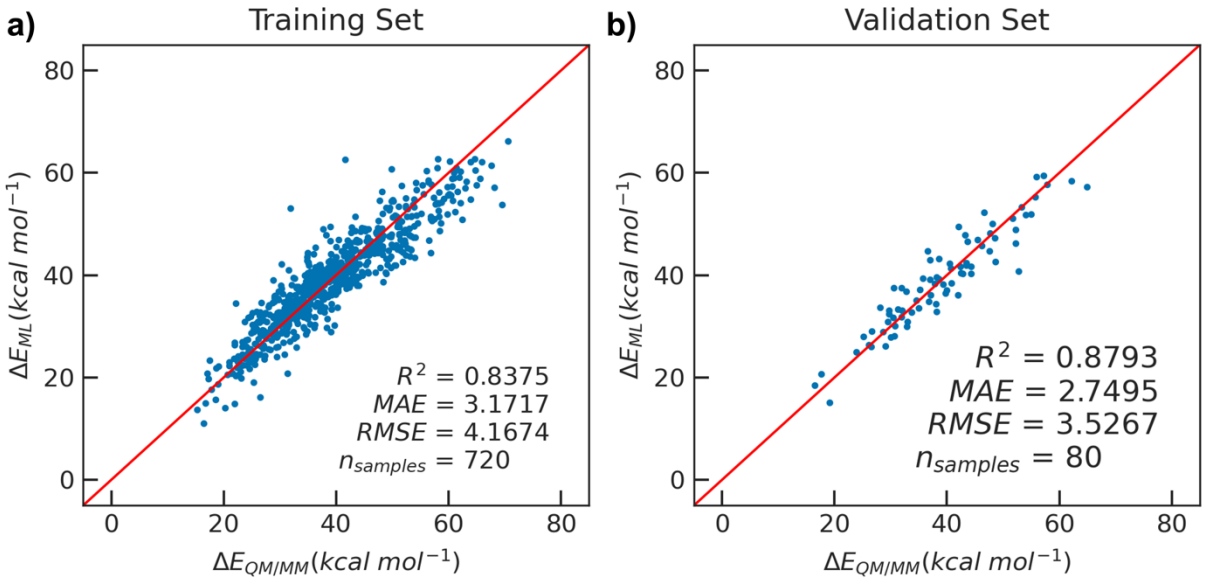

Fig. S2. The performance of the **Linear Regression** (LR) model on the training and validation set. a) The performance of LR model on the training set, where R2, MAE, RMSE, $n_{samples}$ refers to the coefficient of determination, mean absolute error, root mean square error and the number of samples, b) The performance of the LR model on the validation set respectively. $\Delta E_{QM/MM}$ refers to the barrier energy obtained from QM/MM calculation, while ΔE_{ML} stands for the barrier energy predicted by the LR model.

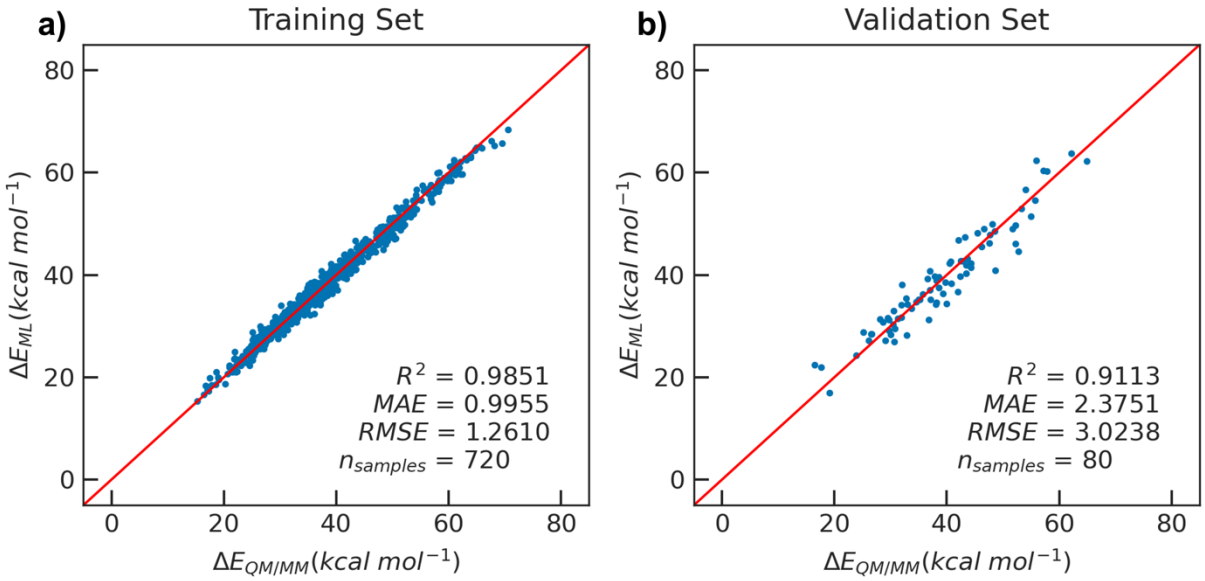

Fig. S3. The performance of the **XGBoost** model on the training and validation sets. a) The performance of the XGBoost model on the training set, where R2, MAE, RMSE, $n_{samples}$ refers to the coefficient of determination, mean absolute error, root mean square error and the number of samples, b) The performance of the XGBoost model on the validation set. $\Delta E_{QM/MM}$ refers to the barrier energy obtained from the QM/MM calculations, while ΔE_{ML} stands for the barrier energy predicted by the XGBoost model.

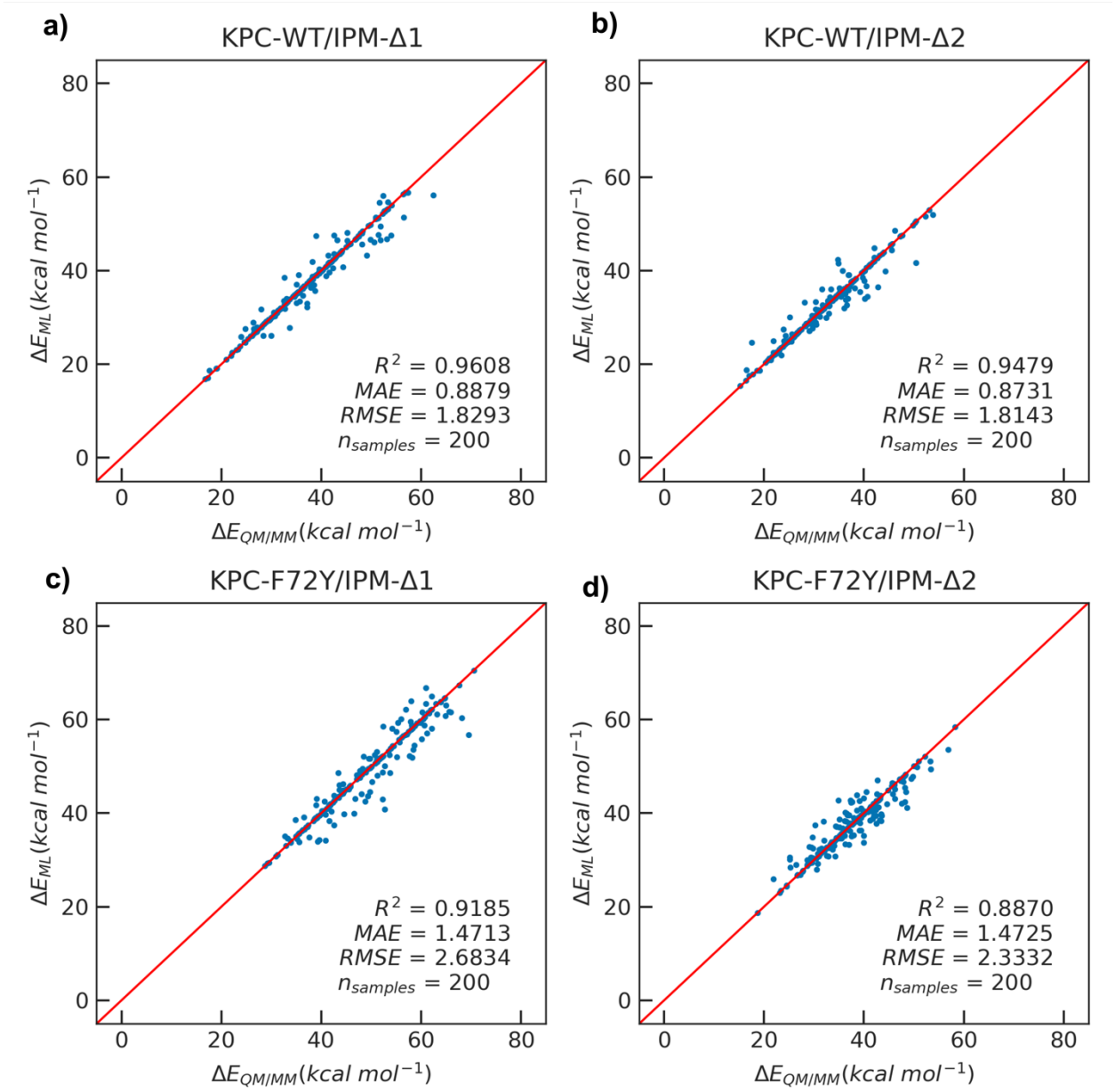

Fig. S4. The **Support Vector Machine (SVM)** model performance on the four systems. (a) Model performance on the KPC-WT/IPM-Δ1 system; (b) the KPC-WT/IPM-Δ2 system; (c) the KPC-F72Y/IPM-Δ1 system; (d) the KPC-F72Y/IPM-Δ2 system. R^2, MAE, RMSE, $n_{samples}$ refer to the coefficient of determination, the mean absolute error, the root-mean-squared error, and the number of samples, respectively.

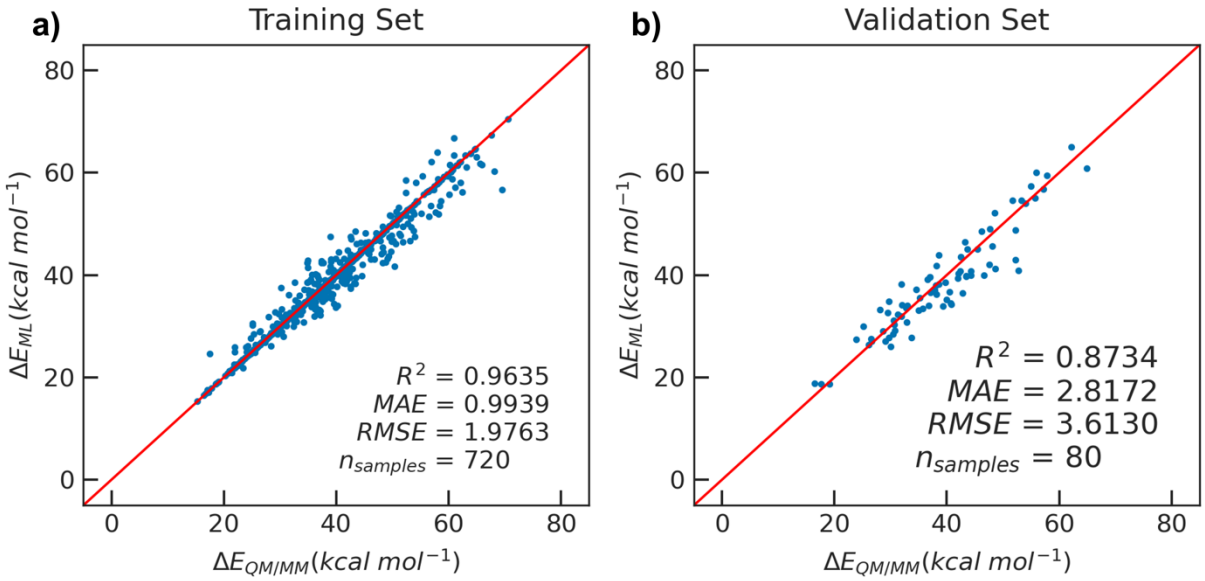

Fig. S5. The performance of **Support Vector Machine (SVM)** model on the training and validation sets. a) The performance of the SVM model on the training set, where R^2, MAE, RMSE, $n_{samples}$ refers to the coefficient of determination, mean absolute error, root mean square error and the number of samples, b) The performance of the SVM model on the validation set. $\Delta E_{QM/MM}$ refers to the barrier energy obtained from the QM/MM calculations, while ΔE_{ML} stands for the barrier energy predicted by the SVM model.

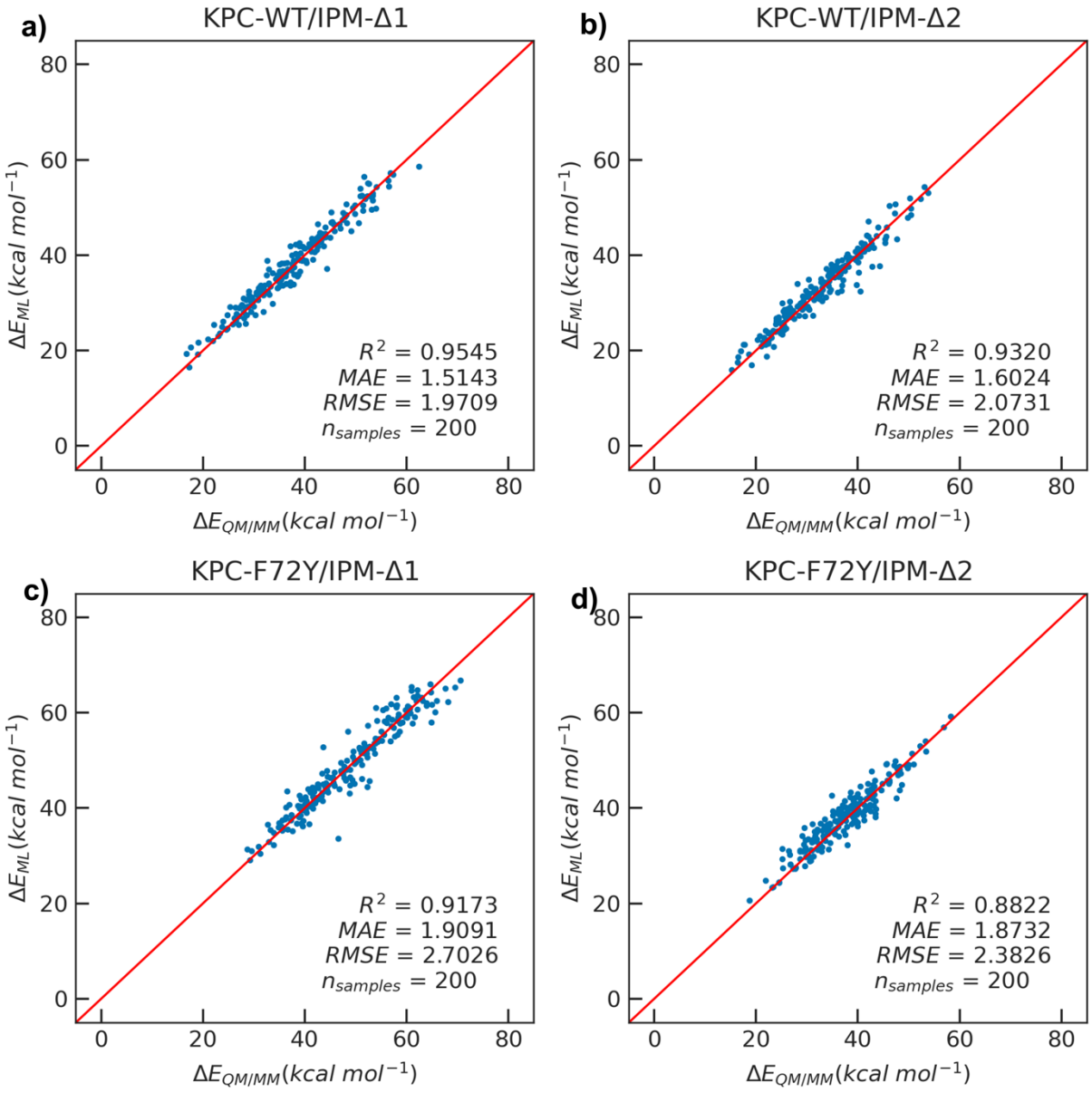

Fig. S6. The **Neural Network** model performance on the four systems. (a) Model performance on the KPC-WT/IPM-Δ1 system; (b) the KPC-WT/IPM-Δ2 system; (c) the KPC-F72Y/IPM-Δ1 system; (d) the KPC-F72Y/IPM-Δ2 system. R^2, MAE, RMSE, $n_{samples}$ refer to the coefficient of determination, the mean absolute error, the root-mean-squared error, and the number of samples, respectively.

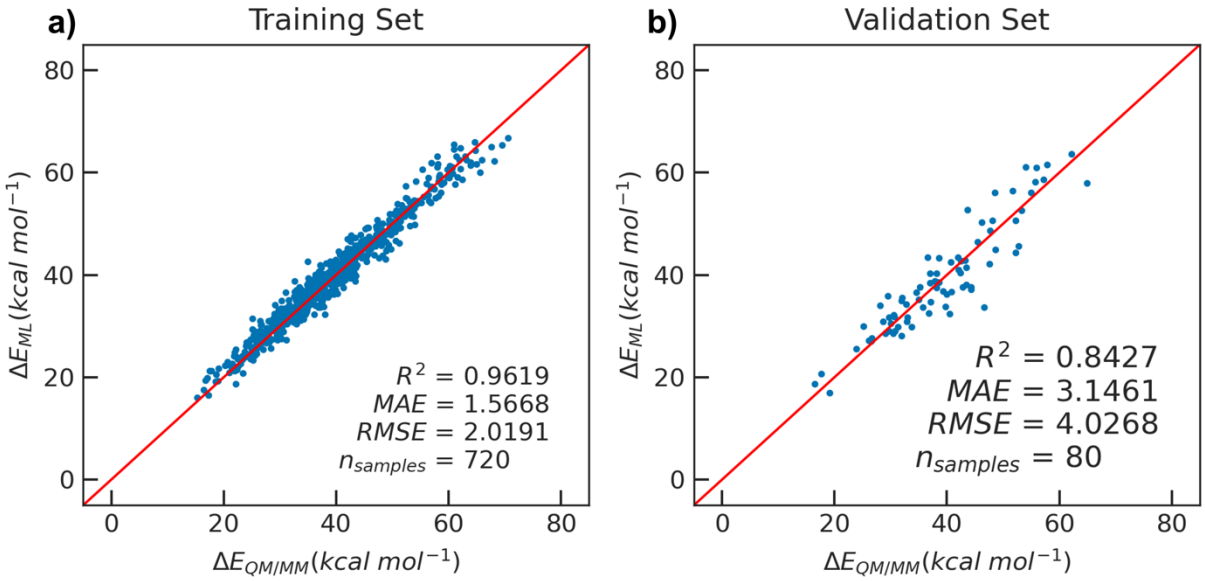

Fig. S7. The performance of the **Neural Network** (NN) model on the training and validation sets. a) The performance of the NN model on the training set, where R2, MAE, RMSE, $n_{samples}$ refer to coefficient of determination, mean absolute error, root mean square error and the number of samples, b) The performance of the NN model on the validation set. $\Delta E_{QM/MM}$ refers to the barrier energy obtained from the QM/MM calculations, while ΔE_{ML} stands for the barrier energy predicted by the neural network model.

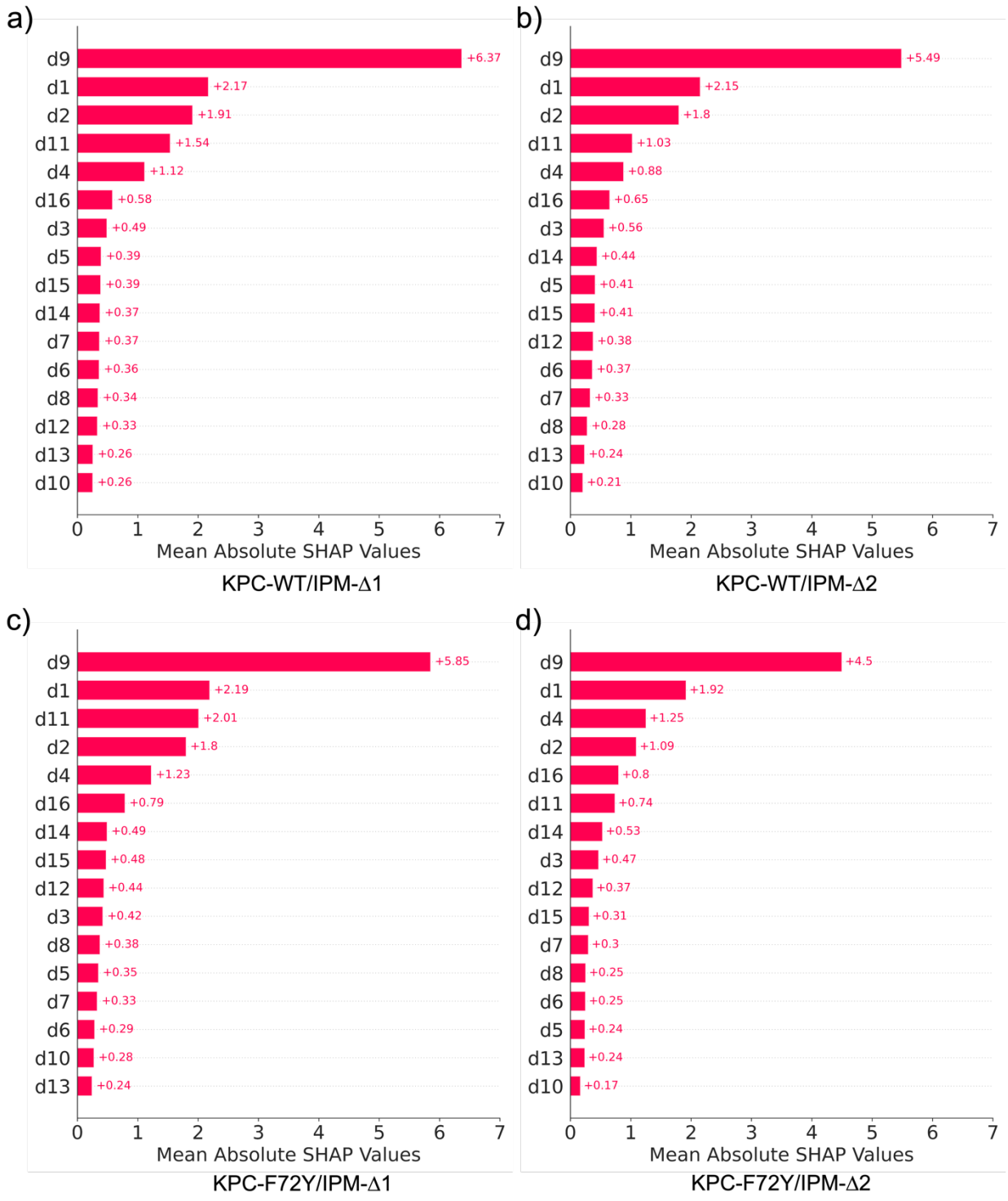

Fig. S8. Mean absolute SHAP values for the 16 conformational features. Mean absolute SHAP values for (a) the KPC-WT/IPM-Δ1 system; (b) the KPC-WT/IPM-Δ2 system; (c) the KPC-

F72Y/IPM-Δ1 system; and (d) the KPC-F72Y/IPM-Δ2 system. The rankings of the distances follow the attributed importance for each feature. See Fig. 3

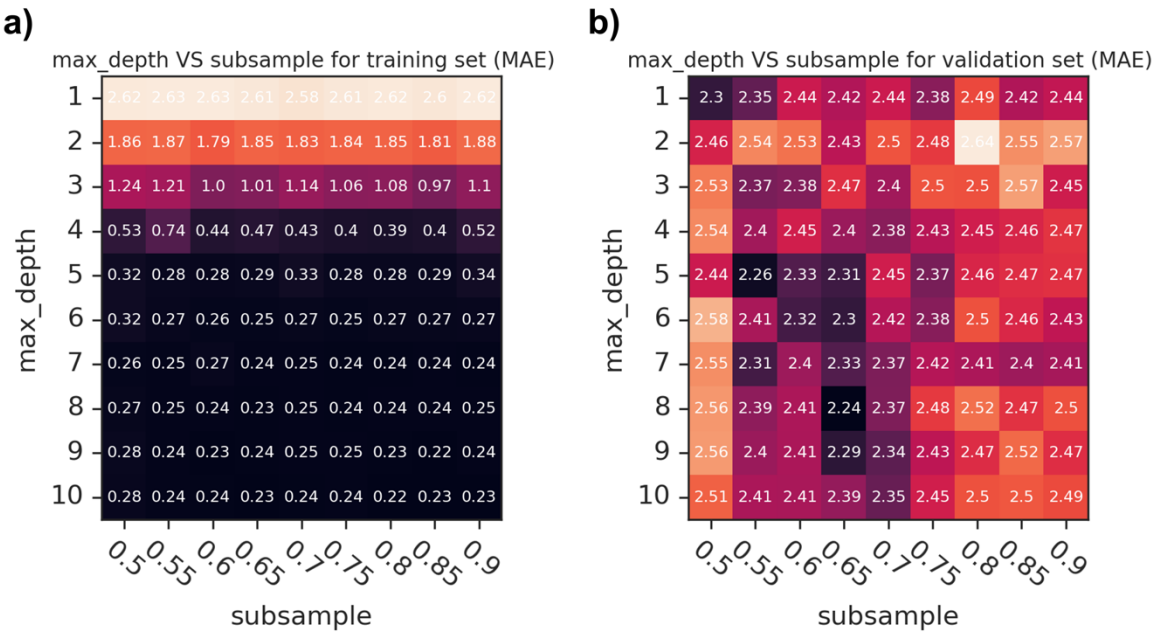

Fig. S9. The grid search for parameter max_depth and subsample. a) The grid search for the training set, values in every block are MAE between the energy calculated by QM/MM method and energy predicted by ML method. b) The grid searching process for the validation set.

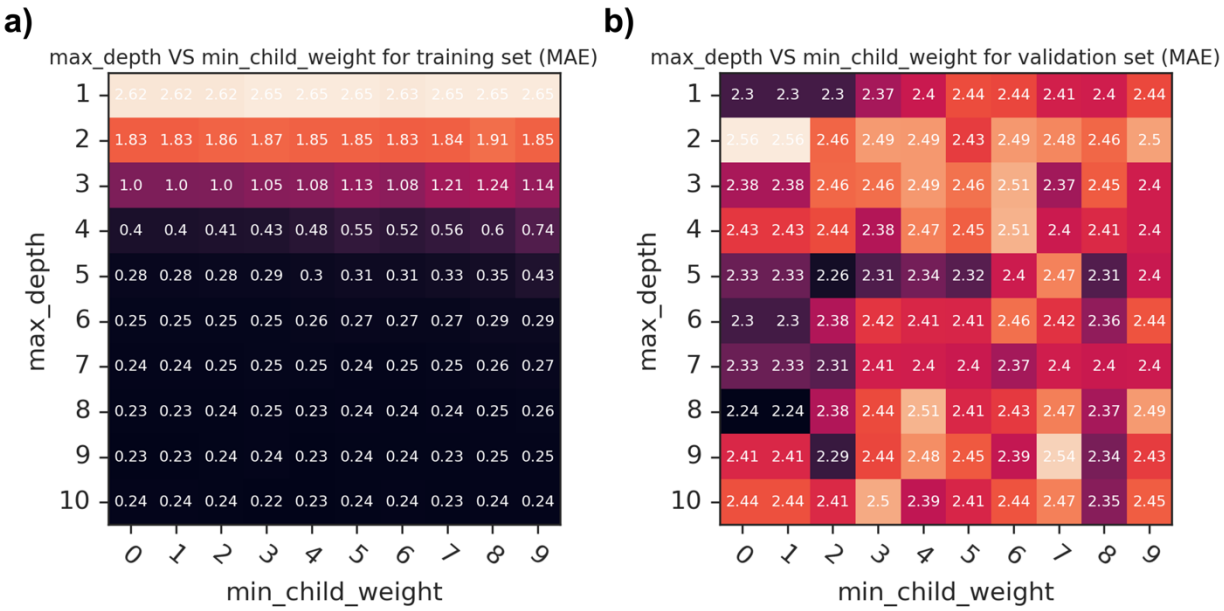

Fig. S10. The grid search for the parameter max_depth and min_child_weight. a) The grid search for the training set, values in every block are MAE between barrier energies calculated by the QM/MM method and the barrier energies predicted by the ML method. b) The grid searching process for the validation set.

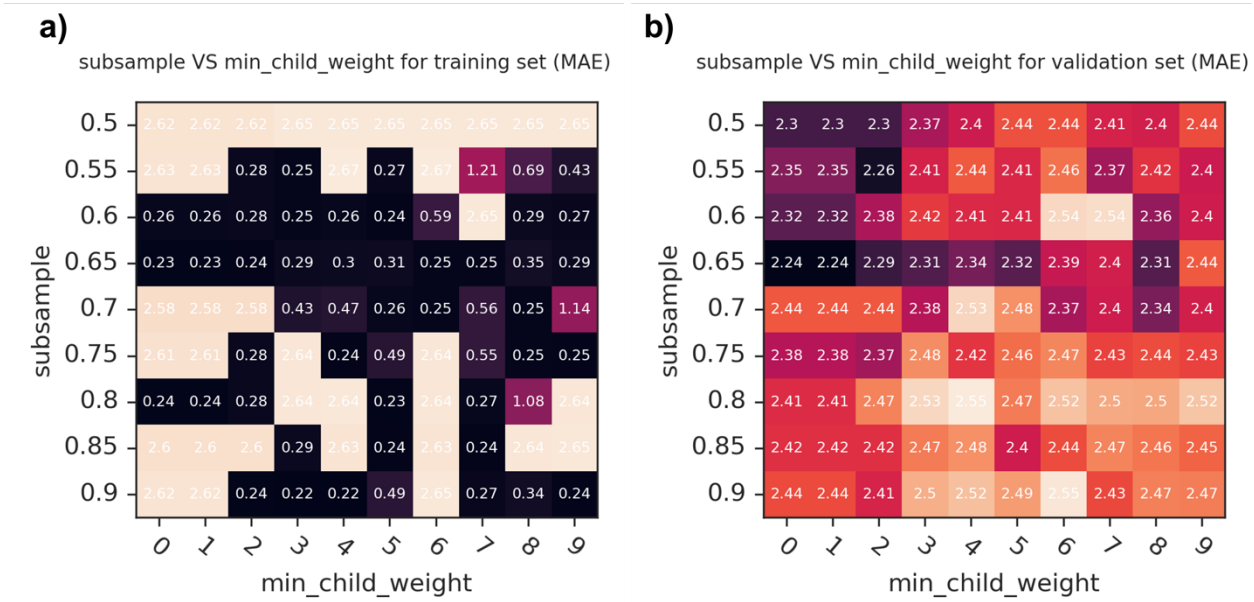

Fig. S11. The grid search for the parameter subsample and min_child_weight. a) The grid search for the training set, values in every block are the MAE between barrier energies calculated by the QM/MM method and barrier energies predicted by the ML method. b) The grid searching process for the validation set.

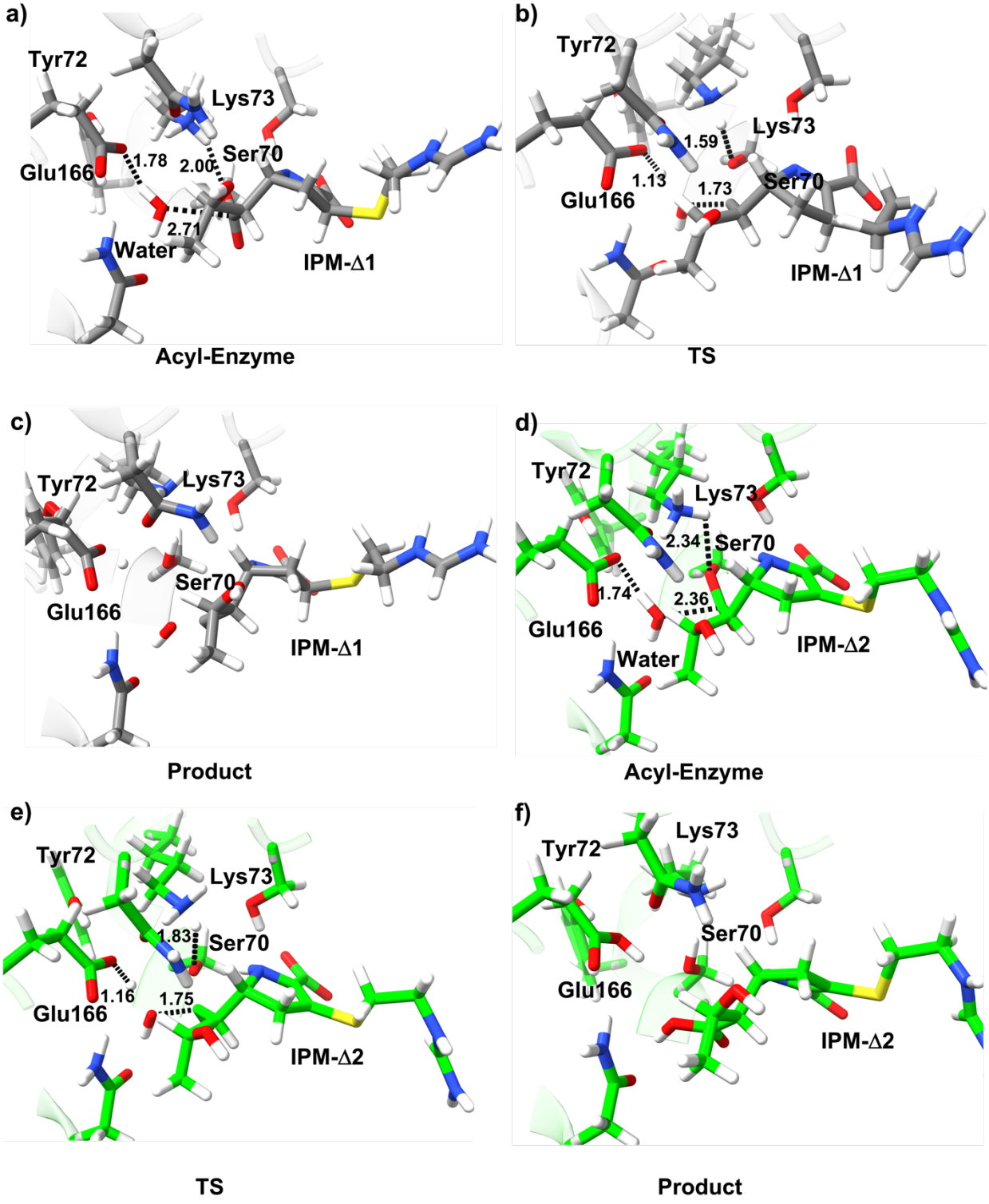

Fig. S12. Active site structure of MEPs with the lowest barrier energy for KPC-F72Y/IPM-Δ1(grey) and KPC- F72Y/IPM-Δ2 (green). TS refers to transition state (TS). Three important

distances (Glu166 Oε2 - Water H1, IPM C7 - Water O, Lys73 Hζ1- Ser70 Oγ), which involve proton transfer, and nucleophilic attack, are marked as the black dashed line with Å unit. The carbon atoms are colored as gray in KPC-F72Y/IPM-Δ1 and green in KPC-F72Y/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-F72Y/IPM-Δ2 and number 200 for KPC-F72Y/ IPM-Δ1.

Supporting Tables

Table S1. Average distance (Å) for all features.

Referring name	Distance name	KPC-WT/IPM-**Δ**1	KPC-WT/IPM-**Δ**2	KPC-F72Y/IPM-**Δ**1	KPC-F72Y/IPM-**Δ**2
d1	Phe72 Hζ (Tyr72 Hη) – Glu166 Oε2	3.34	3.11	1.66	1.65
d2	Lys73 Hζ2 – Glu166 Oε2	2.16	2.02	2.39	2.15
d3	Water H1 – Glu166 Oε2	1.65	1.65	1.65	1.70
d4	Water H1 – Glu166 Oε1	2.51	2.42	2.48	2.33
d5	Asn170 Hδ2 – Glu166 Oε1	1.92	1.88	1.95	1.87
d6	Water H2 – Asn170 Oδ	1.77	1.76	1.79	1.79
d7	Lys73 Hζ1 – Water O	2.82	3.11	2.92	3.16
d8	Lys73 Hζ2 – Asn132 Oδ	1.86	1.91	1.85	1.90
d9	Water O – IPM C7	2.87	2.80	2.95	2.83
d10	Lys73 Hζ1 – Ser70 Oγ	2.12	2.17	2.07	2.09
d11	IPM 6αOH – Water O	3.47	3.25	2.87	2.74
d12	IPM 6αOH – Asn132 Oδ	3.61	3.66	3.39	3.35
d13	IPM 6αOH – Glu166 Oε1	4.17	4.09	3.61	3.20
d14	IPM 6αOH – Glu166 Oε2	4.59	4.61	4.17	4.02
d15	Lys73 Hζ1 – Ser130 Oγ	3.08	3.20	2.85	3.06
d16	Ser130 Hγ – IPM N4	1.90	2.59	1.93	2.55

XGBoost method hyperparameter grid search

Max_depth, learning_rate, subsample ratio, and min_child_weight, four hyperparameter for XGBoost model training have been investigated to obtain the optimal model to accurately build the relationship between conformations' features and barrier energy.

Learning rate is the easiest parameter to select due to the fact that too small or too large values leads to an unconverged training process. Therefore, 0.1, the default value for the XGBoost model, was chosen in the training process.

Max_depth, the most important parameter, was searched from 1 to 10 to get the optimal value. We noticed that the performance of the XGBoost model on the training set gradually increases when max_depth gets larger, while the performance shows only modest changes on the validation set. Max_depth as 3 and subsample as 0.6 was chosen to avoid the overfitting and underfitting problem (Fig. S9). Parameter min_child_weight is not as important as other parameters for the XGBoost training (Fig. S11). min_child_weight parameter as 1 was chosen because it exhibits the best performance on the validation set.

Free energy difference between the imipenem tautomers

The free energy difference between the KPC-WT/IPM-Δ1 and KPC-WT/IPM-Δ1 tautomer states were investigated using the dual-topology DFTB/MM thermodynamic integration (TI) scheme. For computational efficiency, the QM region was reduced to include only the Ser70 side chain (partitioned between Cα and Cβ) and the entire covalently bonded IPM ligand. We note that the active atoms form a neural zwitterion region.

We replicated the QM region in the system topology (*via* the CHARMM REPLICATE command). The first subsystem (the "reactant") was set as the KPC-WT/IPM-Δ1 tautomeric state and the replicated subsystem (the "product") as the KPC-WT/IPM-Δ2. The dual topology TI simulation were performed using the build-in SCCDFTB module of CHARMM (*via* the SCCDFTB LAMD DTOP command). We note that the dual-topology scheme suffers from the end-point instability due to the scaling of the sub-system (and especially the QM) potentials, thus long timescale simulations are hardly feasible. We further note that using the default 3ob parameters caused the positively charged IPM tails to deform during the dynamics runs. Consequently, we used the DFTB3/3ob-f/C36m level of theory, which permits stable covalent bonds, for the TI simulations. The order parameters were sampled at $\lambda = 0.1, 0.3, 0.5, 0.7$, and 0.9. In each sampling window, the dynamics of the perturbed system were propagated at a 1 fs timestep for 2.5 ps at each TI step. The gradient of the perturbed system potential to the value of the order were collected every 5 fs. The final free energy difference was computed by trapezoidal numerical integration. The resulting log files and the dual topology coordinate files (psf and cor) were provided in the Zenodo repository (see Data Availability).

We report that the free energy of the KPC-WT/IPM-Δ1 states is higher by 3.97 kcal mol^{-1} than KPC-WT/IPM-Δ2. The free energy difference aligns with the experimental observation that

the two tautomeric states were interchangeable since that the tautomerization paths would differ only by the reported free energy difference. Also, we expect similar results for the F72Y systems since that the mutation points are far from the tautomer rings. Finally, due to limited simulation time, this result should be interpreted qualitatively instead of quantitatively.