Supporting Information For:

KIF – Key Interactions Finder: A Program to Identify the Key Molecular Interactions that Regulate Protein Conformational Changes

Rory M. Crean,[1] Joanna S.G. Slusky,[2,3] Peter M. Kasson,[4,5] Shina Caroline Lynn Kamerlin[1,6]

1. Department of Chemistry – BMC, Uppsala University, BMC Box 576, S-751 23 Uppsala, Sweden
2. Center for Computational Biology, University of Kansas, Lawrence, KS 66047, USA
3. Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045, USA
4. Departments of Molecular Physiology and Biomedical Engineering, University of Virginia, Charlottesville, Virginia 22908, USA
5. Department of Cell and Molecular Biology, Uppsala University, BMC Box 596, Uppsala 751 23, Sweden
6. School of Chemistry and Biochemistry, Georgia Institute of Technology, 901 Atlantic Drive NW, Atlanta, GA 30332-0400, USA

Corresponding author email address: skamerlin3@gatech.edu

TABLE OF CONTENTS

S1. SUPPLEMENTARY METHODOLOGY S3

System Preparation for MD Simulations S3

Equilibration and Production MD Simulations S3

Additional MD Simulation Analysis S5

Classification of Simulation Frames S6

Classification and Regression Metrics Used for Machine Learning S7

Multi-collinearity and Model Selection S7

Measuring Calculation Timings S8

S2. SUPPLEMENTARY RESULTS S9

Limitations with the Machine Learning Generated Per Feature/Interaction Scores S9

Results for the PDZ3 Domain S12

S3. KIF VALIDATION S13

Generating a Synthetic Dataset to Validate KIF S13

S4. SUPPLEMENTARY TABLES S21

S5. SUPPLEMENTARY FIGURES S30

S6. SUPPLEMENTARY REFERENCES S47

S1. SUPPLEMENTARY METHODOLOGY

System Preparation for MD Simulations

We have previously prepared the required input and parameters files for simulations of protein tyrosine phosphatase 1B (PTP1B) starting from the closed and open WPD-loop conformational states in prior work, and used these previously prepared starting structures.[1-3] The end result of this preparation procedure was an octahedral water box of size such that no protein atom was within 10 Å of the box boundary and with counter ions added as necessary to neutralize the system. The protein atoms were described with the AMBER ff14SB force field with parameters for the phosphocysteine residue derived using restrained electrostatic potential calculations. The crystal structure used and histidine tautomerisation states used are provided in **Table S1**.

The setup for the PDZ3 domain simulations are described in **Methodology** section ("*Preparation and Simulation of the Model Systems*"). For the two Kemp eliminase (KE07) variants[4-6] studied here (R1 and R4), we again used starting structures obtained from our prior work,[7] with the crystal structures and histidine tautomerisation states used provided in **Table S1**.

Equilibration and Production MD Simulations

Simulations of PTP1B and the PDZ3 domain (with and without peptide) used the same equilibration and production simulation protocol, with these systems simulated using the AMBER18[8] package. Following system setup (described above), the equilibration protocol was performed as follows for both systems and all replicas: (1) All hydrogen atoms, water molecules and counterions were energy minimised with 500 steps of steepest descent and 500 steps of conjugate gradient minimization. This was done by applying 10 kcal mol^{-1} $Å^{-2}$ positional restraints on all other atoms in the system. (2) The system was heated from 50 K to

300 K in the NVT ensemble in 200 ps, with the above described restraints retained. (3) The size of the positional restraints was then reduced to 5 kcal mol^{-1} $Å^{-2}$ and only applied to the Cα atoms, with energy minimization again performed (with 500 steps of steepest descent followed by 500 steps of conjugate gradient minimization). (4) With the same restraints as step 3, the system was heated again from 25 K to 300 K over the course of 500 ps in the NVT ensemble. (5) Simulations were then performed in the NPT ensemble (300 K, 1 atm), with the time step changed from 1 fs to 2 fs. The 5 kcal mol^{-1} $Å^{-2}$ positional restraints were progressively released by reducing the restraint value in 1 kcal mol^{-1} $Å^{-2}$ blocks every 10 ps until the restraint was completed removed. Following this, a 10 ns equilibration simulation was performed for PTP1B and 100 ns equilibration simulation for PDZ3 before production MD simulations were began (with the production simulations using the same parameters as these equilibration simulations). The longer equilibration time for PDZ3 simulations was to allow the structure to relax after the removal of the peptide from the structure in order to simulate the unliganded state. Both NVT and NPT ensemble simulations used Langevin temperature control[9] with the collision frequency set to 1 ps^{-1}. Pressure control for the NPT ensemble simulations was maintained by a Berendsen barostat,[10] with the pressure relaxation time set to 1 ps. NVT simulations were ran with a timestep of 1 ps, whilst NPT simulations used a timestep of 2 ps, and the SHAKE algorithm[11] was used to constrain all bonds containing hydrogen atoms.

The Hamiltonian replica exchange molecular dynamics (HREX-MD)[12] simulations of the two KE07 variants were performed using the Amber ff99SB-ILDN[13] force field and TIP3P[14] water model using GROMACS 2018.4,[15] interfaced with PLUMED v2.5[16] (the ff99SB-ILDN was chosen over ff14SB,[17] as only force-fields embedded into the GROMCAS program can be used with this method). To prepare for production simulations, each system was subjected to 5000 steps of steepest descent and then conjugate gradient energy minimization. Systems

were then heated from 25 to 300 K in 500 ps in an NVT ensemble using Langevin temperature control,[9] with the collision frequency set to 1 ps^{-1}, and time step 1 fs was used. After this, each system was equilibrated in the NPT ensemble (300 K, 1 atm), with Langevin temperature control[9] and a Parrinello−Rahman barostat[18, 19] (pressure relaxation time of 1 ps) for temperature and pressure regulation respectively. All bonds to hydrogen were restrained using the P-LINCS algorithm[20] and a 10 Å non-bonded interaction cut-off was used, with long range electrostatic interactions evaluated using the Particle Mesh Ewald (PME) algorithm.[21] Consistent with the original study,[7] the production HREX-MD simulations were of 200 ns length each and six replicas were performed per system with lambda values exponentially scaled between 1.0 to 0.667. Exchanges between neighboring replicas were attempted every 1 ps, with analysis performed solely on the neutral (λ = 1) replica.

Additional MD Simulation Analysis

Dihedral angle measurements and C_α-atom root mean square deviation (RMSD) calculations for target variable assignment, alongside C_α-atom root mean square fluctuation (RMSF) calculations were all performed using CPPTRAJ.[22] The C_α-atom RMSD calculations (on PTP1B's flexible WPD-loop) and the C_α-atom RMSF calculations (on the PDZ3 domain and the KE07 variants) were performed in a similar manner. First, an average structure was generated by aligning only to residues with reasonable structural stability (highly flexible residues can reduce the quality of the fit). Then, all frames were aligned to this average structure, again using only those residues with defined secondary structure to perform the alignment. This aligned trajectory was then used to calculate either the C_α-atom RMSD or RMSF as appropriate. For PTP1B, the residues used for the alignment were: 15-26, 33-41, 69-84, 92-102, 106-109, 133-150, 153-162, 166-176, 188-201, 212-214, 220-237, 246-256, 264-281 (*i.e.* all residues with defined secondary structure). For the PDZ3 domain, the residues used for the alignment were: 311-414 (skipping the highly flexible N-terminal

residues). For the KE07 variants, the residues used for alignment were: 5-12, 32-42, 46-50, 58-72, 77-80, 86-95, 98-102, 111-119, 125-134, 137-142, 154-163, 185-194, 198-201, 208-216, 220-223, 235-243. The reference structures used for the C_α-atom RMSD calculations of PTP1B's WPD-loop were obtained from PDB ID: 6B90.[23] This structure contains both a closed and open WPD-loop structure, with the WPD-loop residues defined as residues 176-190.

The free energy landscape plots of the KE07 variants were generated by taking the per frame dihedral observations from our HREX-MD simulations to generate a 2D histogram. The bin widths were 4° for both W50's χ_1 and χ_2 dihedral angles. The raw bin populations were converted to a free energy values using the following formula:

$$G_i = -k_B T \, ln\left(\frac{N_i}{N_{max}}\right) \quad \textbf{(1)}$$

Where G_i is the free energy for bin i, k_B is the Boltzmann's constant, T is the temperature (300 K), and N_i is the number of observations in the bin i and N_{max} is the largest number of observations in any bin.

Classification of Simulation Frames

For PTP1B, each frame used in the PyContact[24] calculation was classified according to the WPD-loop RMSD with a rule based approach. The frame was defined as "Closed" (if the WPD-loop RMSD to the closed loop crystal structure was ≤ 1.5 Å) or "Open" (if the WPD-loop RMSD to the open loop crystal structure was ≤ 1.5 Å) and "Neither" (if the WPD-loop RMSD to both the closed and open crystal structure conformations was ≤ 1.5 Å or if the RMSD was >1.5 Å to both crystal structures). For the simulations of the PDZ3 domain, frames were classified according to if they were from the peptide bound or unliganded simulations.

Classification and Regression Metrics Used for Machine Learning

Three classification algorithms were used to build the ML classification models, (1) Random Forest (RF),[25] (2) categorical boosting (CatBoost)[26] and (3) eXtreme Gradient Boosting (XGBoost).[27] The criterion used for the training of each model was the default (Random Forest: the Gini impurity, CatBoost and XGBoost: logistic loss). The accuracy of each model was evaluated by calculating the F1-score. The ML regression models were built using the default loss functions provided by the three models (Random Forest and XGBoost: MSE, and CatBoost: RMSE), where MSE is the mean squared error and RMSE is the root mean squared error. Errors in the regression models were evaluated (on the validation datasets, see the **Methodology**) using both the RMSE and mean absolute error (MAE).

Multi-collinearity and Model Selection

Feature sets which contain multi-collinearity (features that are highly correlated to one another) can substantially impact the training of various ML models alongside providing potentially misleading per feature scores (*i.e.,* per feature importances). A typical approach to deal with multi-collinearity in ML pipelines is to cluster/group features that have high collinearity and select a single feature from each cluster/group for model building. Another alternative would be to perform some kind of dimensionality reduction, such as principal component analysis (PCA) on the feature set to generate a set of features with no-collinearity. Whilst these solutions would be perfectly acceptable if our goal was to use ML for prediction only, these solutions are non-ideal given our desire to determine all the important features (*i.e.,* interactions) for regulating the chosen target variable.

Therefore, we instead chose to select several ensemble ML models that are insensitive to multi-collinearity (from the perspective of building the actual models). Whilst the ML algorithms chosen can still give per feature scores that only showcase a few (of the many) important interactions (as we demonstrated with the results obtained in this manuscript for

both PTP1B and PDZ3), we were able to apply the techniques from our statistical analysis module instead in order to generate a better overview of the interaction networks.

Measuring Calculation Timings

To calculate the time taken to build the various models generated, we determined the value of a monotonic clock both before and after the process to be measured was performed (using the time.monotonic() method). Following this the difference in time was computed using the “datetime” libraries “timedelta()” function and converted into minutes. Timing were performed using a Dell Precision 3530 with an Intel Core i7-8850H with a 2.6 GHz clock speed.

S2. SUPPLEMENTARY RESULTS

Note that all **Supplementary Results** figures and tables are suffixed by "SRX", with X referring to the table or figure number in this section.

Limitations with the Machine Learning Generated Per Feature/Interaction Scores

Our package KIF provides the option to generate per feature/interaction and per residue scores from either a machine learning (ML) or statistical analysis module. In the following section we will demonstrate with the PDZ3 and PTP1B model systems how only some of the features/interactions that are notably associated with the target variable are required to accurately predict the target variables value. Whilst this would be perfectly fine (if not overkill) if one was applying these ML models towards prediction, this is undesirable if the goal is to identify many important features that describe the conformational differences. This (innate) limitation is not present in the methods available within the statistical analysis module, so we recommend a user apply methods from that module instead. This limitation is better highlighted with the PTP1B model system (as a very small number of features are required for an accurate classification model), so the results for PTP1B will presented first and in more detail.

Supplementary Results for PTP1B

As described in the main text, three machine learning (ML) models were built to distinguish between the WPD-loop open and closed states of PTP1B. All three classification ML models achieved an a F_1 score of 1.00 on the validation set with no misclassifications occurring at all (**Table SR1**). By analysing the per feature scores (*i.e.,* the feature importances) generated from each of these ML models one can readily identify why each model was so accurate (**Figure SR1**). That is, the conformational differences between the closed and open conformational states are enormous (the loop changes conformation by ~10 Å, **Figure 2B**), meaning only a few interactions in the local environment are more than

enough to accurately predict the conformational state of PTP1B. Whilst this would be perfectly fine (if not overkill) if one were applying these ML models towards prediction, this is undesirable if the goal is to identify many important features that describe the conformational differences.

Table SR1. The F_1 scores obtained during training (from the *k*-fold cross validation procedure) and the validation (*i.e.,* holdout) dataset for the machine learning performed on simulations of the PTP1B, alongside the time taken to build each model. [a]

	F_1 Score (cross validation)	**F_1 Score (holdout)**	**Time Taken (mins)**
All Residues			
Random Forest	1.00	1.00	1.1
CatBoost	1.00	1.00	2.3
XGBoost	1.00	1.00	0.6
No WPD-Loop Residues			
Random Forest	0.97	0.98	2.4
CatBoost	0.98	0.99	2.0
XGBoost	0.99	0.98	2.7

[a] The F_1 score is a metric that provides a balanced view of model accuracy, meaning it can be applied to both balanced and unbalanced datasets. The F_1 score was determined using the implementation in the sci-kit learn library.[28] The machine learning algorithms were trained to classify trajectory frames of PTP1B as being in the WPD-loop closed or open conformations (see the **Methodology** section).

One such way to try to "get around" the aforementioned issue would be to remove the "obvious" features which differ in the open and closed states and rebuild the models. We tested this by removing all the interactions/features which include a residue on the WPD-loop

(residues 176-190). With this approach, the F_1 scores on the validation dataset are still very high (between 0.98-0.99 for all three models, **Table SR1**), and there is clearly a greater spread in the per feature scores (especially in the case of the Random Forest model), **Figure SR1**. That said, this approach of removing certain local interactions is disadvantageous in that it is hard to imagine an ideal cut-off point, alongside the fact that local interactions that may still be of interest are lost when implementing a cut-off.

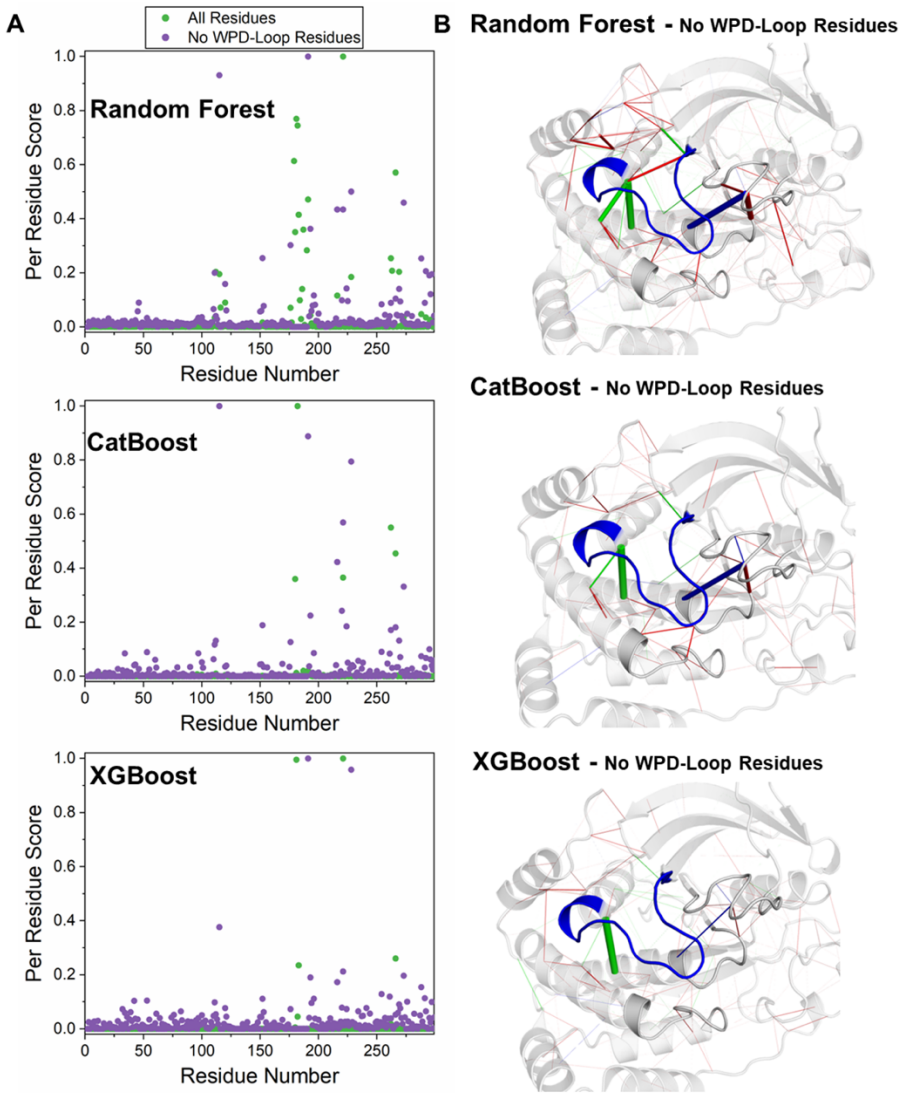

Figure SR1. (**A**) Per residue scores for the three machine learning (ML) models built for PTP1B to distinguish between the WPD-loop open and closed state. Each scatter plot contains scores for a model built using all residues (green dots) and all residues except for the

WPD-loop (purple dots). (**B**) Visualization of the per feature scores onto the structure of PTP1B for the three ML models built without using the WPD-loop residues interactions as features. The larger the cylinder the greater the per feature score and cylinders are colored according to their interaction type (red for hydrogen bonds; green for hydrophobic interactions and blue for salt bridges).

Results for the PDZ3 Domain

As described in the main text, three machine learning (ML) models were built to distinguish between the peptide bound and unliganded states of the PDZ3 domain. The per residue scores generated for the three ML models and two models from our statistical analysis package are provided in **Figure S3**. When comparing to the PTP1B model system described in the section above, the spread of scores is on average better/more aligned with the statistical models, which is likely the result of this classification being more challenging. That is, more interactions are needed to accurately classify the difference between the peptide bound and unliganded states of the PDZ3 domain as compared to the closed and open states of the PTP1B's WPD-loop, giving rise to a "larger spread" of scores across the residues.

S3. KIF VALIDATION

Note that all **Supplementary** figures and tables in this section are suffixed by "SVX", with X referring to the table or figure number in this section.

We note that the code used to generate the data for this section is available for viewing from the KIF GitHub repository at the following address:

https://github.com/kamerlinlab/KIF/tree/main/tutorials/synthetic_data_test

Generating a Synthetic Dataset to Validate KIF

In order to validate our package KIF (short for Key Interactions finder) is working correctly and is therefore able to identify important features from a dataset of molecular interactions, we constructed a model/synthetic dataset designed to replicate a typical MD simulation generated dataset that KIF would be applied to. This model dataset was constructed such that the features would have a range of pre-defined linear correlation values (*i.e.,* Pearson correlation coefficients). This would in effect mean that we would *a priori* know which features are "more important" than others. The basic idea behind this approach therefore is that KIF should be able to identify these known important features, if the methods within have been implemented correctly.

To construct this dataset we first produced observations for a continuous target variable by generating 10,000 samples from a normal distribution with a mean of 10 and standard deviation of 3. We then generated a second column with 10,000 observations from a normal distribution with the same parameters. This second column was used to generate 16 features with correlations to the target dataset between approximately -0.85 and +0.9. This was achieved by calculating the residuals from the fit of the two above described columns and subtracting the required amount from each observation in the 2nd column to achieve the desired linear correlation (LC) to the target variable. This process was repeated to generate

another 16 features only this time a uniform distribution with a minimum and maximum of 0 and 10 respectively was used as the second column. Finally, any features that had an observation below 0 was altered such that their minimum value was 0 (by simply adding a constant value to each observation). This reflects the reality that the features used within KIF would not be less than 0.

Comparison of the targeted linear correlations (LCs) to the measured LCs of each feature to the target variable (**Figure SV1A**) demonstrated the method worked well in generating a dataset with a diverse range of correlations to the target variable. Further, by determining the correlation between all the synthetic features to themselves (**Figure SV1B**), we observed the dataset to containing a large degree of multi-collinearity. This will therefore allow us to test whether having features/interactions with multi-collinearity impacts the methods within KIF. In summary, the resulting dataset consists of 36 continuous features, each with 10,000 observations and a range of correlations to the target variable between approximately -0.85 and +0.9.

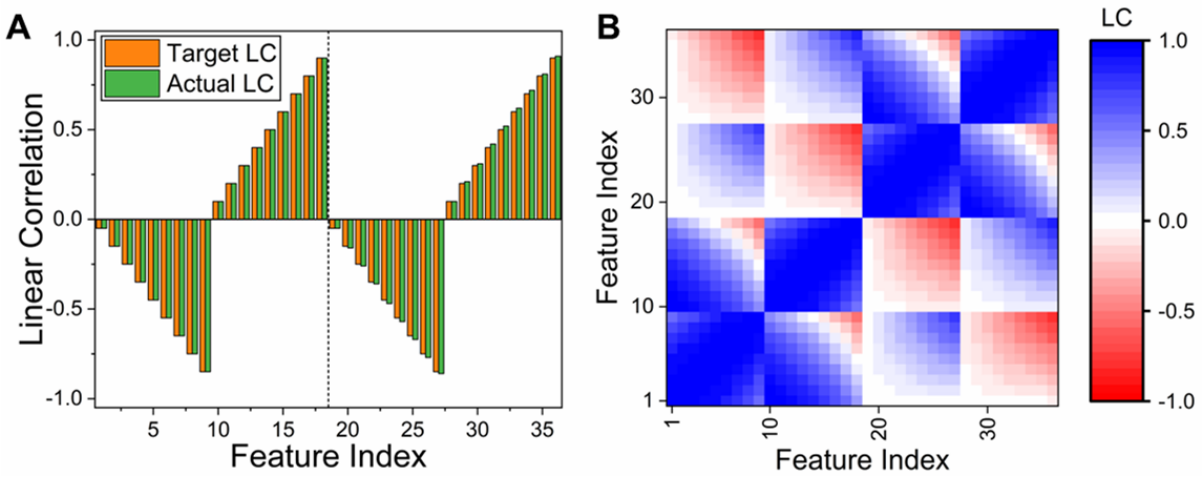

Figure SV1. (**A**) Targeted linear correlations (LCs) against the actual correlations for the 36 synthetically generated features. The first 18 features were obtained by sampling from a normal distribution, whilst the reaming 18 features were obtained by sampling from a uniform distribution, with the separation indicated by a black dotted line. (**B**) Feature

correlation matrix depicting the linear correlation between every feature to itself in the synthetic dataset.

The above dataset was generated with the target variable column being a continuous range of values, therefore creating a regression problem. As KIF can also perform classification we converted this continuous target into a categorical target, by diving the target up into two classes (for binary classification). Class assignments were made simply by whether each observation was greater than or lower than the mean target variable value. A visualisation of the class distributions for several features (**Figure SV2**) demonstrates this approach was effective in generating a categorical target variable with features having a range of associations to the target classes. Further and as depicted in **Figure SV2**, the more "important" features in the classification model are those with a higher LC to the continuous target variable. This means we can compare each KIF calculated per feature scores with this ordering to validate the classification models within KIF.

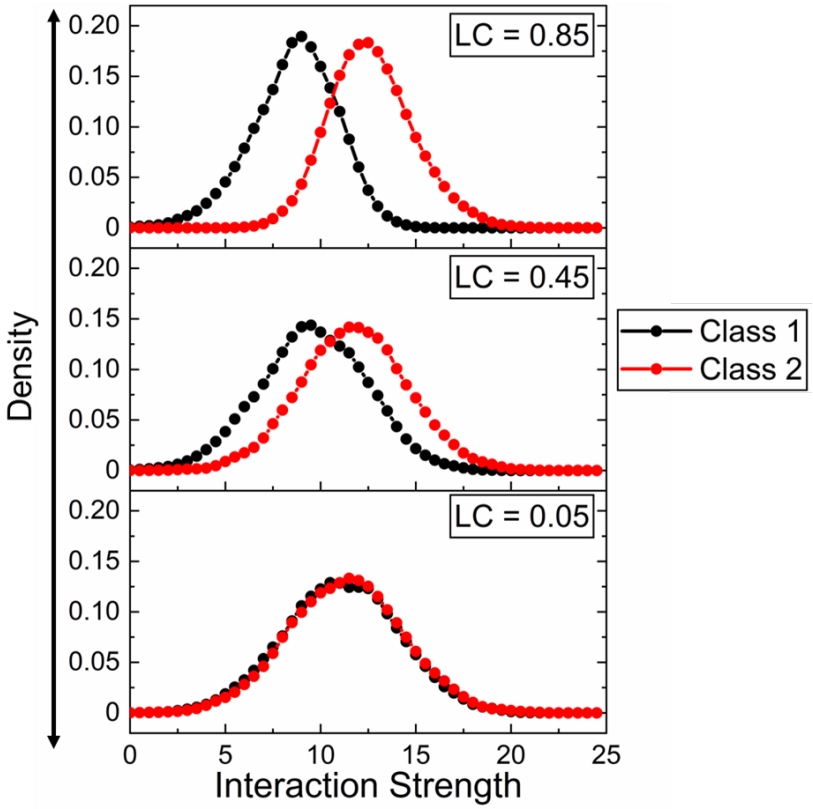

Figure SV2. Illustration of the class separation technique used to convert the synthetic dataset from a regression problem to a classification problem. Three exemplar features are shown with defined linear correlations (LCs) to the continuous target variable. The continuous target variable was divided into two classes based on if the observation was above or below the target variable's mean. The resulting probability densities of the three exemplar features for the two classes are shown. Probability densities were generated through gaussian kernel density estimation with the bandwidth set to 0.5 for all plots.

With our synthetic regression and classification datasets constructed we compared the per feature importance scores generated by all of the models available within KIF against their known absolute linear correlations (**Figure SV3**). The absolute value of the linear correlation was chosen as most of the metrics within KIF do not have a measure of directionality (*i.e.,* they don't have both negative and positive values, they only go from 0 to positive values). Further, comparisons between the per feature scores and their defined absolute linear

correlation was done using the Spearman's rank metric, which measures the ability to rank order the features. That is, as different types of metrics are being used, one would expect to be able to rank order the important from the less important features. With the above in mind, we can observe both the classification and regression statistical models perform well at reproducing the correct order with Spearman's rank values between 0.96-1.00 (**Figure SV3**). The machine learning (ML) methods on the other hand perform quite poorly on this synthetic dataset, in most cases identifying only a few features as important/high scoring (**Figure SV3**). These observations were expected given the nature of the synthetic dataset for the same reasons as discussed in detail in **Section SR2**. That is, the ML models only needed enough information to accurately predict the class or value and due to the dataset containing many features with high correlations to the target, lots of the features are effectively redundant. Consistent with the above, all ML models generated were highly predictive, with $R^2 \geq 0.99$ for regression and F1-scores ≥ 0.99 for classification for the 20% of the data excluded from the training (*i.e.,* the test set). For the above reason, and as discussed in the main text and in detail in **Section S2** of the **Supplementary Information**, we would recommend using the statistical methods to analyse most molecular interaction datasets, in order to capture all the relevant features/interactions.

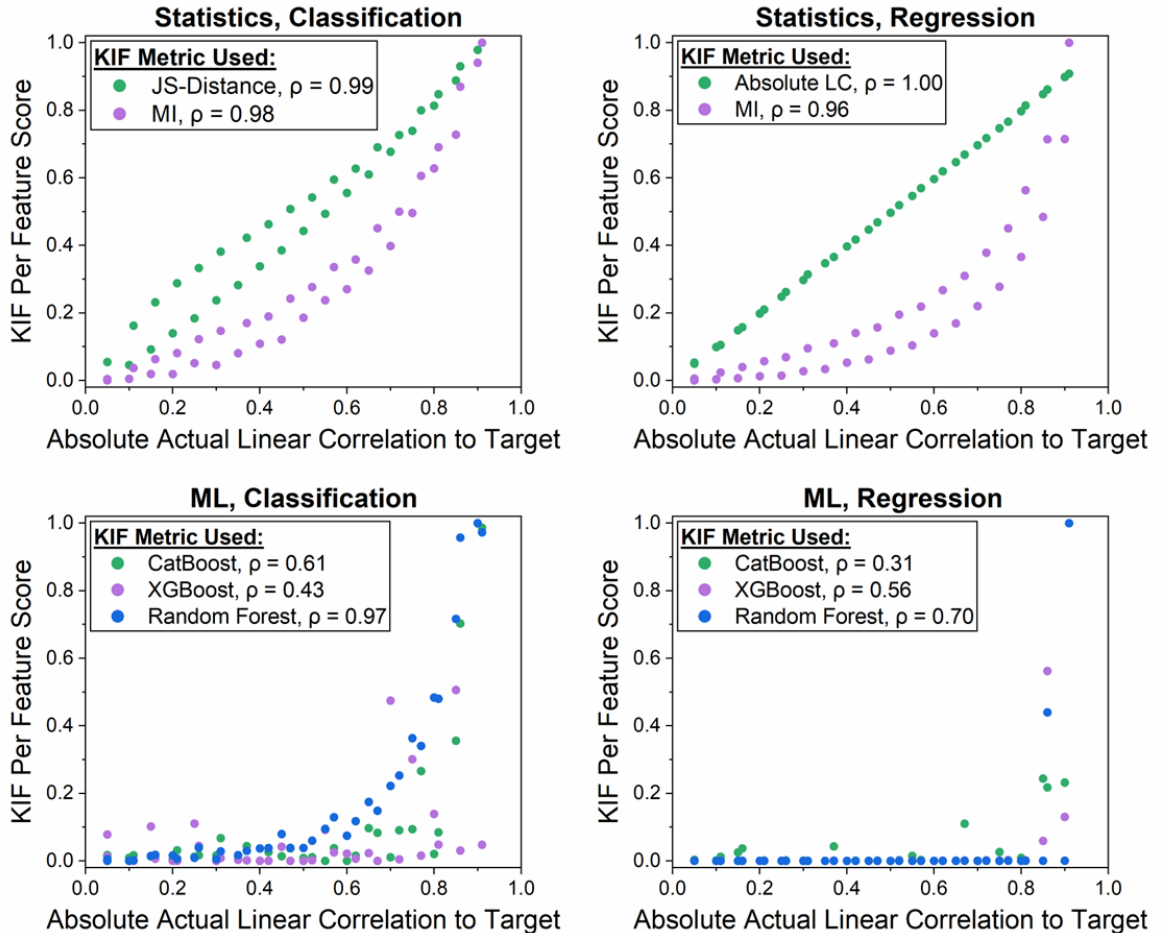

Figure SV3. Comparisons of the relationship between each feature in the synthetic dataset with a defined linear correlation to the target variable (the absolute value of this is the x-axis throughout) to the per feature score obtained from all of the metrics available with KIF. The figure is broken up into four panels, with the top panels showing results obtained from methods available within the statistical analysis module and bottom graphs from the machine learning (ML) module. The top and bottom panels are also divided up into whether the problem was a classification (left panels) or regression (right panels) problem. Provided with each scatter plot is the spearman's rank (ρ), which asses the ability to correctly rank order the features. Acronyms for the methods used are as follows: JS-Distance (Jensen Shannon distance); MI (mutual information), LC (Linear correlation).

Finally, we investigated the impact of duplicating features within the synthetic dataset. This was done in order to assess whether the duplicated features would obtain the same KIF score, which they theoretically (or at least ideally) should. To do this, we selected five columns with a range of LCs to the target variable (-0.65, -0.75, +0.1, +0.3 ,+0.9) and duplicated them. This dataset with the duplications was then run through the same KIF workflow as described above, and the KIF per feature scores of the duplicated features were evaluated (**Figure SV4**).

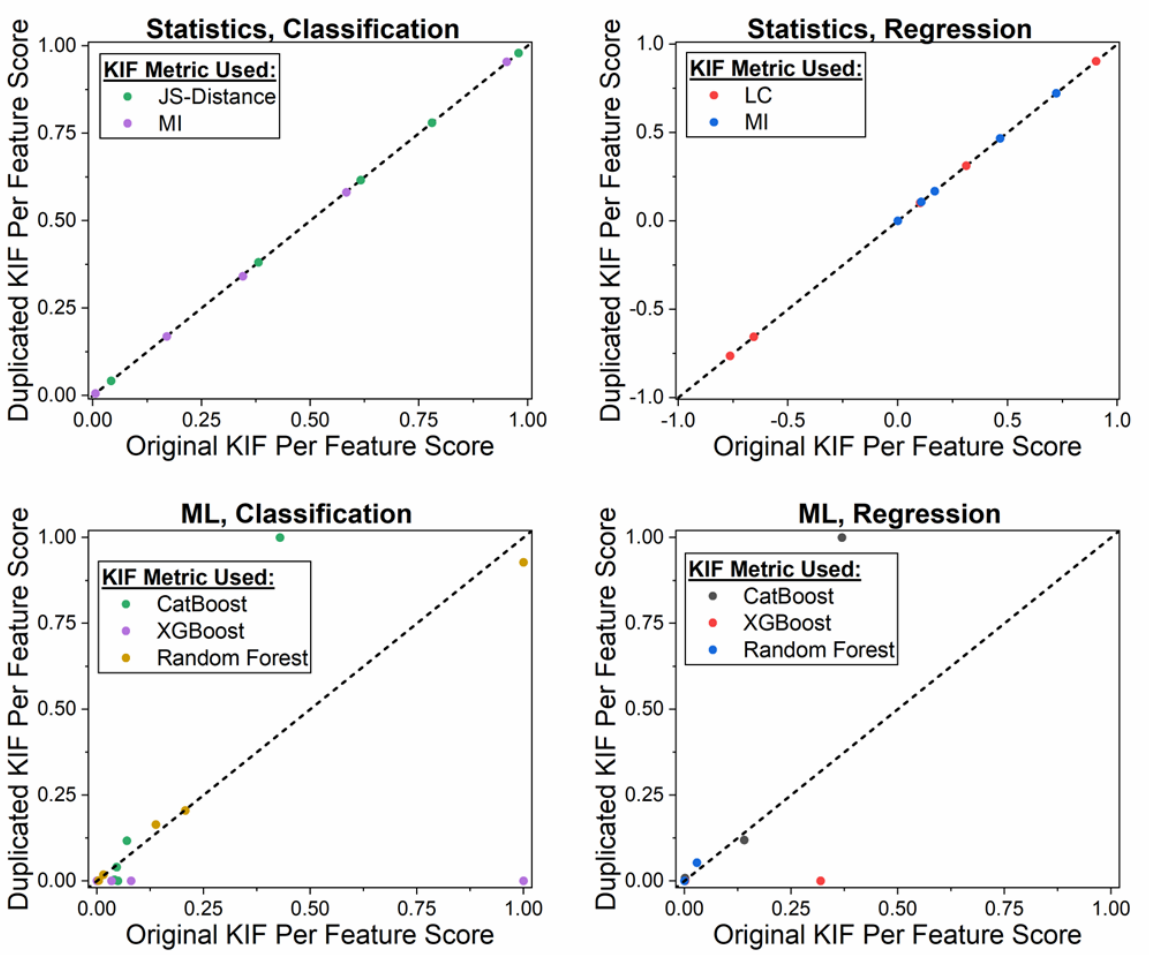

Figure SV4. The impact of the duplicating features on the KIF per features scores for all methods available within KIF. In each case, KIF was run with five duplicated features, meaning five sets of two columns have identical values. The KIF per feature score was then extracted for each duplicated feature pair and is plotted as x, y pairs on the graph above. A diagonal line of the form y=x is depicted to show if a point exists on the diagonal (*i.e.,* their x

and y values are identical). The figure is broken up into four panels, with the top panels showing results obtained from methods available within the statistical analysis module and bottom graphs from the machine learning (ML) module. The top and bottom panels are also divided up into whether the problem was a classification (left panels) or regression (right panels) problem. Acronyms for the methods used are as follows: JS-Distance (Jensen Shannon distance); MI (mutual information), LC (Linear correlation).

The results presented in **Figure SV4** demonstrate that all statistical methods contained within KIF were insensitive to the duplication of features, and correctly ranked them with the same score. This would be expected due to the way these methods were implemented, whereby each feature is evaluated one at a time, independent of all other features. In contrast, the ML methods did not reproduce the same scores for the duplicated feature pairs. This was also expected as there is no information gain to the ML model from adding the duplicated feature.

S4. SUPPLEMENTARY TABLES

Table S1. The crystal structures and histidine tautomerisation states used for all systems simulated.[a]

System	X-ray Structure	Tautomerisation State		
		HID	**HIE**	**HIP**
PTP1B				
Closed WPD-Loop	6B90[b]	214	94, 173, 175	25, 54, 60, 208, 296
Open WPD-Loop	6B90[b]	214	94, 173, 175	25, 54, 60, 208, 296
PDZ3 Domain				
Peptide Bound	1BE9	372	317	-
Unliganded	1BE9[c]	372	317	-
KE07 Variants				
R1	5D2W	244	84, 201, 209, 228	-
R4	3IIO	244	84, 201, 209, 228	-

[a] HID, HIE and HIP correspond to histidine side chains protonated on either the $N_{\delta 1}$, $N_{\epsilon 2}$, or both nitrogen atoms, the latter being positively charged. [b] Both the closed and open WPD-loop conformations are observed in this crystal structure. [c] The peptide was removed from the structure in order to run the simulations in the unliganded state.

Table S2. Description of the feature filtering performed on PTP1B simulations prior to performing the various analyses described in this work.[a]

Machine Learning and Statistical Analysis.	
Type of Filtering	**Description**
Residues Included	Tested retaining all non-WPD-loop residues and removing WPD-loop residue interactions (175-189)
Minimum Occupancy	Interaction kept only if present in ≥25% of frames
Average Interaction Strength	Interaction kept only when the average PyContact interaction strength score was ≥0.5
Interaction Type	Hydrogen bonds, salt bridges and hydrophobic interactions retained for analysis
Side- or Main-Chain Interactions	No filtering performed
Network Analysis	
Type of Filtering	**Description**
Residues Included	All included
Minimum Occupancy	Interaction kept only if present in ≥50% of frames.
Average Interaction Strength	No filtering performed
Interaction Type	No filtering performed
Side- or Main-Chain Interactions	No filtering performed

[a] Filtering was performed on the non-covalent interactions generated using PyContact.[24]

Table S3. Description of the feature filtering performed on simulations of the PDZ3 domain prior to performing the various analyses described in this work.[a]

Machine Learning and Statistical Analysis.	
Type of Filtering	**Description**
Residues Included	All PDZ3 domain residues, no peptide residues
Minimum Occupancy	Interaction kept only if present in ≥25% of frames
Average Interaction Strength	Interaction kept only average PyContact interaction strength score of ≥0.5
Interaction Type	No filtering performed
Side- or Main-Chain Interactions	No filtering performed
Network Analysis	
Type of Filtering	**Description**
Residues Included	All PDZ3 domain residues, no peptide residues
Minimum Occupancy	Interaction kept only if present in ≥ 50% of frames
Average Interaction Strength	No filtering performed
Interaction Type	No filtering performed
Side- or Main-Chain Interactions	No filtering performed

[a] Filtering was performed on the non-covalent interactions generated using PyContact.[24]

Table S4. Description of the feature filtering performed on simulations of selected KE07 variants prior to performing machine learning and statistical analysis.[a]

Filtering	Description
Residues Included	All residues included
Minimum Occupancy	Interaction kept only if present in ≥25% of frames
Average Interaction Strength	Interaction kept only average PyContact interaction strength score was ≥0.5
Interaction Type	No filtering performed
Side- or Main-Chain Interactions	No filtering performed

[a] Filtering was performed on the non-covalent interactions generated using PyContact.[24]

Table S5. The F1-scores obtained during training (from the *k*-fold cross validation procedure) and the validation (*i.e.,* holdout) dataset for the machine learning performed on simulations of the of the PDZ3 domain, alongside the time taken to build each model. [a]

	F_1 Score (cross validation)	F_1 Score (holdout)	Time Taken (mins)
Random Forest	0.97	0.98	4.3
CatBoost	0.99	0.99	1.4
XGBoost	1.00	0.99	3.3

[a] The F_1 score is a metric that provides a balanced view of model accuracy, meaning it can be applied to both balanced and unbalanced datasets. The F_1 score was determined using the implementation in the sci-kit learn library.[28] The machine learning methods were trained to classify trajectory frames of the PDZ3 domain as being either peptide bound or unliganded.

Table S6. Per feature scores and ranks determined for the PDZ3 domain, which contains either the residue H372 or F400.[a]

Residue 1	Residue 2	Interaction type	Per Feature Score	Feature Rank
G329 Main Chain	H372 Side Chain	Hydrogen bond	0.96	2
G330 Main Chain	H372 Main Chain	Hydrogen bond	0.93	5
I336 Side Chain	H372 Main Chain	van der Waals	0.72	10
I328 Main Chain	H372 Side Chain	Hydrogen bond	0.66	13
E331 Side Chain	H372 Side Chain	Hydrogen bond	0.65	15
G330 Main Chain	F400 Side Chain	van der Waals	0.94	4
G329 Main Chain	F400 Side Chain	van der Waals	0.74	8
E334 Side Chain	F400 Side Chain	van der Waals	0.45	31

[a] Per feature scores were determined from the statistical analysis part of our package using the Jensen-Shannon distance[29] as the metric (see the **Methodology** section). Per feature scores are normalized such that the max value is 1. Only features that are within the top 50 features are shown below and their "feature rank" indicates their position within the top 50.

Table S7. The relative populations of each active site conformational state obtained from our HREX-MD simulations of the R1 and R4 KE07 variants. The number of frames and the percentage population (in brackets) are shown.

Enzyme	Conformation[a]			
	A	**B**	**C**	**Other/None**
R1	6927 (69.3%)	532 (5.3%)	2059 (20.6%)	483 (4.83%)
R4	3623 (36.2%)	68 (0.68%)	5046 (50.5%)	1264 (12.6%)

[a] The conformation of each state was defined based on the W50's $\chi1$ and $\chi2$ dihedral angles, with the following definitions: Conformation A: $\chi1$ $\geq160°$ and $\leq210°$ and $\chi2$ $\geq45°$ and $\leq140°$; Conformation B: $\chi1$ $\geq250°$ and $\leq300°$ and $\chi2$ $\geq145°$ and $\leq220°$; Conformation C: $\chi1$ $\geq180°$ and $\leq230°$ and $\chi2$ $\geq235°$ and $\leq290°$; with all angle combinations classified therefore as Other/None.

Table S8. Impact of the N224D mutation on the conformation of the W50 side chain in the R1 and R4 KE07[4-7] variants.[a]

Interaction/Feature	R1		R4	
	LC	**MI**	**LC**	**MI**
Favor the "A" conformational state in the R1 variant				
N224D H228 Hydrogen bond	-0.47	0.23	-0.03	0.03
A9 H228 Hydrogen bond	-0.44	0.16	-0.14	0.02
A9 N224D Hydrogen bond	-0.34	0.09	-0.15	0.06
Y39 H228 Hydrogen bond	-0.37	0.19	0.08	0.10
Favour the "C" conformational state in the R1 variant				
F23 H228 Hydrogen bond (non-native)	0.41	0.15	0.08	0.10
N22 H228 Hydrogen bond	0.32	0.15	0.09	0.04
S225 H228 Hydrogen bond	0.37	0.08	0.32	0.12
S21 H228 Hydrogen bond (non-native)	0.22	0.07	-0.12	0.03

[a] LC stands for linear correlation and MI stands for mutual information.[30, 31] These are the two metrics used to determine per feature scores towards the target variable, see the **Methodology** section. The per feature scores for the interaction network surrounding N224 identified in the R1 variant are shown graphically in **Figure 8D**. The interactions are separated into those that favor either the "A" and "C" conformational states, determined by their linear correlation to the target variable, see the discussion in the main text. The two interactions which are non-native (not found in the respective crystal structure) are indicated.

Table S9. The values of the two regression error metrics used to evaluate the quality of the machine learning (ML) models generated for the R1 and R4 KE07 variants,[4-6] with the regression models trained to predict the value of the W50 χ_2 dihedral angle.[a]

	MAE[a] (°)	RMSE[b] (°)	Time Taken (minutes)
R1			
Random Forest	17.1	30.8	83.9
CatBoost	19.4	28.8	1.9
XGBoost	16.9	28.5	4.0
R4			
Random Forest	12.2	19.2	66.1
CatBoost	14.9	21.6	1.8
XGBoost	13.0	20.7	4.0

[a] MAE and RMSE denote the mean absolute error and root mean squared errors, respectively. The error in each model was determined using the validation set, which makes up 15% of the dataset (see the **Methodology** section). For both variants, three regression ML models were made, using the following algorithms: Random Forest,[25] Categorical boosting (CatBoost),[26] and eXtreme Gradient Boosting (XGBoost).[27]

Table S10. Impact of the K146E mutation between the R1 and R4 KE07 variants.[4-6,a]

Interaction/Feature	R1		R4	
	LC[b]	**MI**[b]	**LC**[b]	**MI**[b]
Loop to Helix Interactions				
E108 Y143 Hydrogen bond	-0.45	0.14	n.o.[b]	n.o.[b]
T104 Y143 Hydrogen bond	-0.21	0.08	-0.57	0.26
N103 S144 Hydrogen bond	n.o.[c]	n.o.[c]	-0.59	0.30
T104 T142 Hydrogen bond	-0.42	0.18	-0.20	0.06
Inter-Helix Interactions				
N103 A106 Hydrogen bond	-0.25	0.16	-0.54	0.31
T104 V107 Hydrogen bond	-0.52	0.29	-0.56	0.23
T104 E108 Hydrogen bond	-0.47	0.23	-0.60	0.51

[a] Per feature scores using our two different metrics are provided for interactions between the loop K146E is located on and to a nearby helix (*Loop to Helix Interactions*) alongside interactions within the neighboring helix itself (*Inter-Helix Interactions*). Refer to **Figure 8D** and the surrounding text for more context. [b] LC stands for linear correlation and MI stands for mutual information. These are the two metrics used to determine per feature scores towards the target variable, see the **Methodology** section. [c] n.o. stands for not observed, meaning the interaction did not occur in the simulation or was removed by the feature filtering steps (meaning it was infrequently observed or a very weak interaction).

S5. SUPPLEMENTARY FIGURES

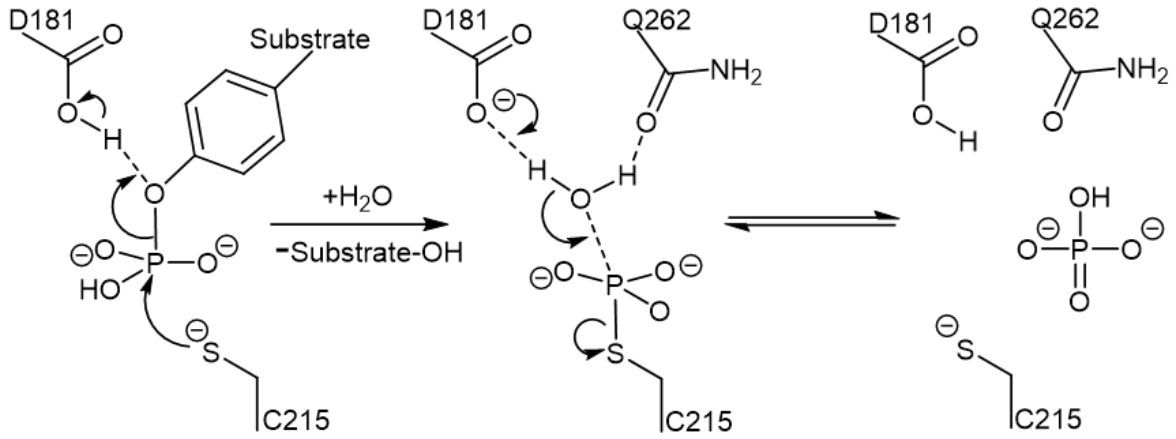

Figure S1. The catalytic mechanism of PTP1B.[32] In the first step, C215 (on the P-loop) acts as a nucleophile towards the phosphorylated substrate, whilst D181 (on the WPD-loop) acts as acid to help promote the reaction. In the second step, the enzyme is regenerated by a water molecule acting as a nucleophile to cleave the phosphoryl group from C215. This is promoted by D181 acting as a base to cleave a proton from the water molecule, and Q262 (on the Q-loop) hydrogen bonding with the other hydrogen atom on the nucleophilic water molecule. In this two-step mechanism, the second step (hydrolysis of the phosphoenzyme intermediate) is the rate limiting step,[33] and our simulations were thus of that intermediate in accordance with this, as outlined in the **Methodology** section.

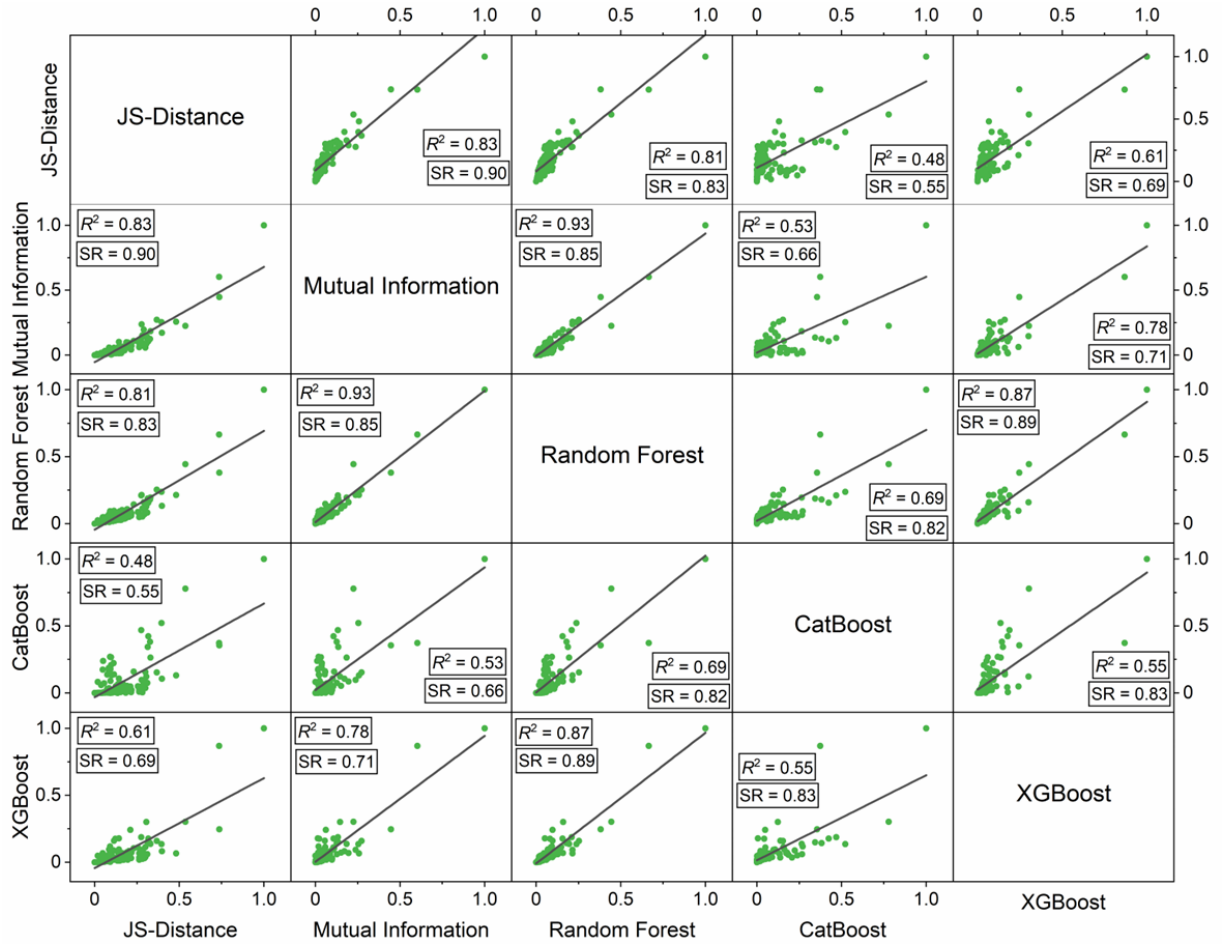

Figure S2. Scatter matrix for all 5 per residue score calculation methods used to study the PDZ3 domain (to classify simulation frames as either peptide bound or unliganded simulation frames). A linear fit for each scatter plot is shown as a black line and the R^2 value for the linear fit alongside the Spearman's rank (SR) is provided. The SR measures how similar the rank ordering of the per residue scores are. The mutual information[30, 31] and JS (Jensen-Shannon) distance[29] based approaches to calculate the per residue score are provided in the statistical analysis module, whilst the Random Forest,[25] CatBoost (Categorical Boosting),[26] and XGBoost (eXtreme Gradient Boosting)[27] methods are all available in the machine learning module.

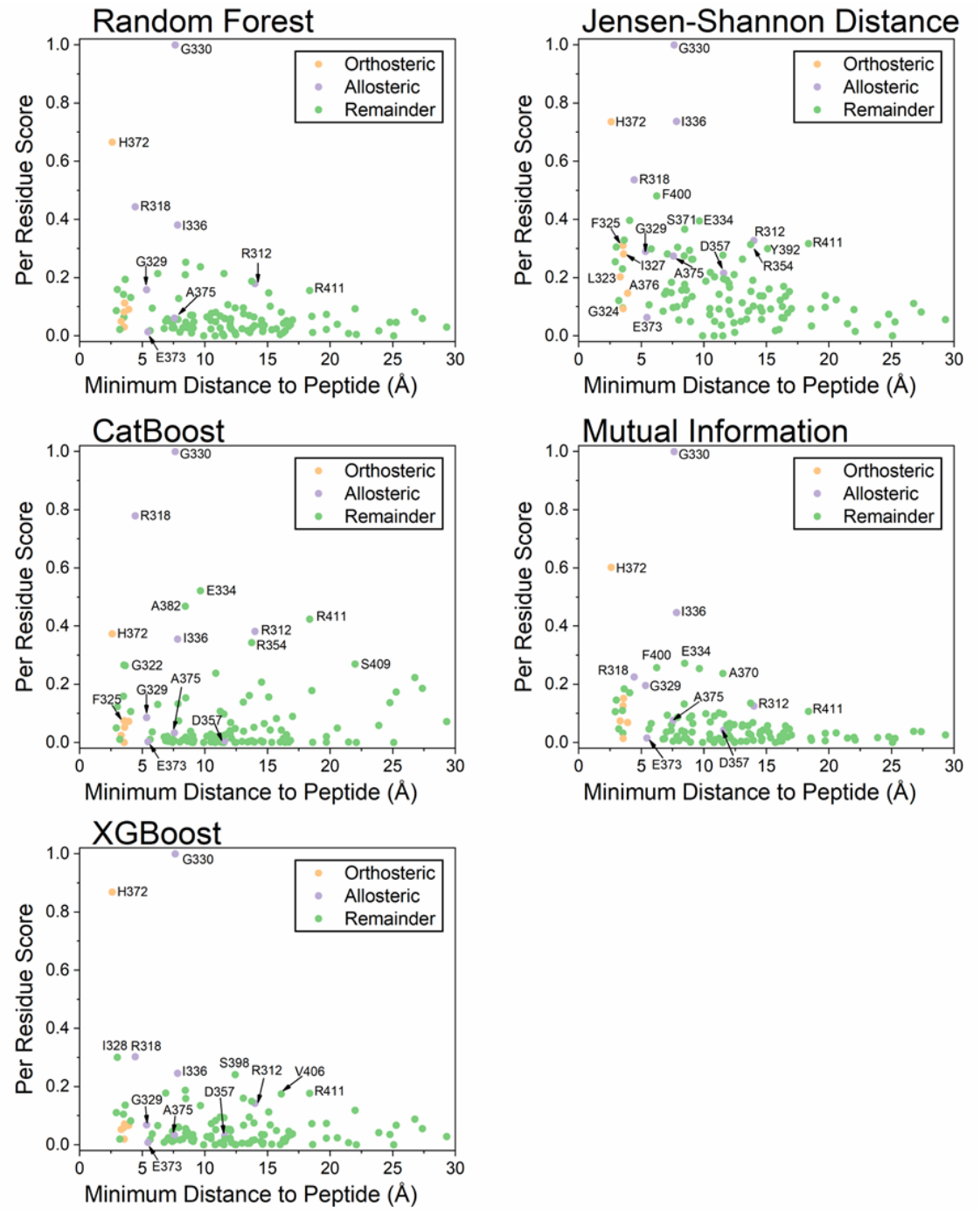

Figure S3. Calculated per residue scores for the PDZ3 domain, using the 5 methods available within our package. All models are based on distinguishing between the MD simulations of the PDZ3 domain both with and without peptide bound. The three panels on the left are from our machine learning module, and the two panels on the right are from our statistical analysis module. Residues are labelled and color coded according to their definitions provided by Faure *et al.*, obtained using exhaustive point mutation screening.[34]

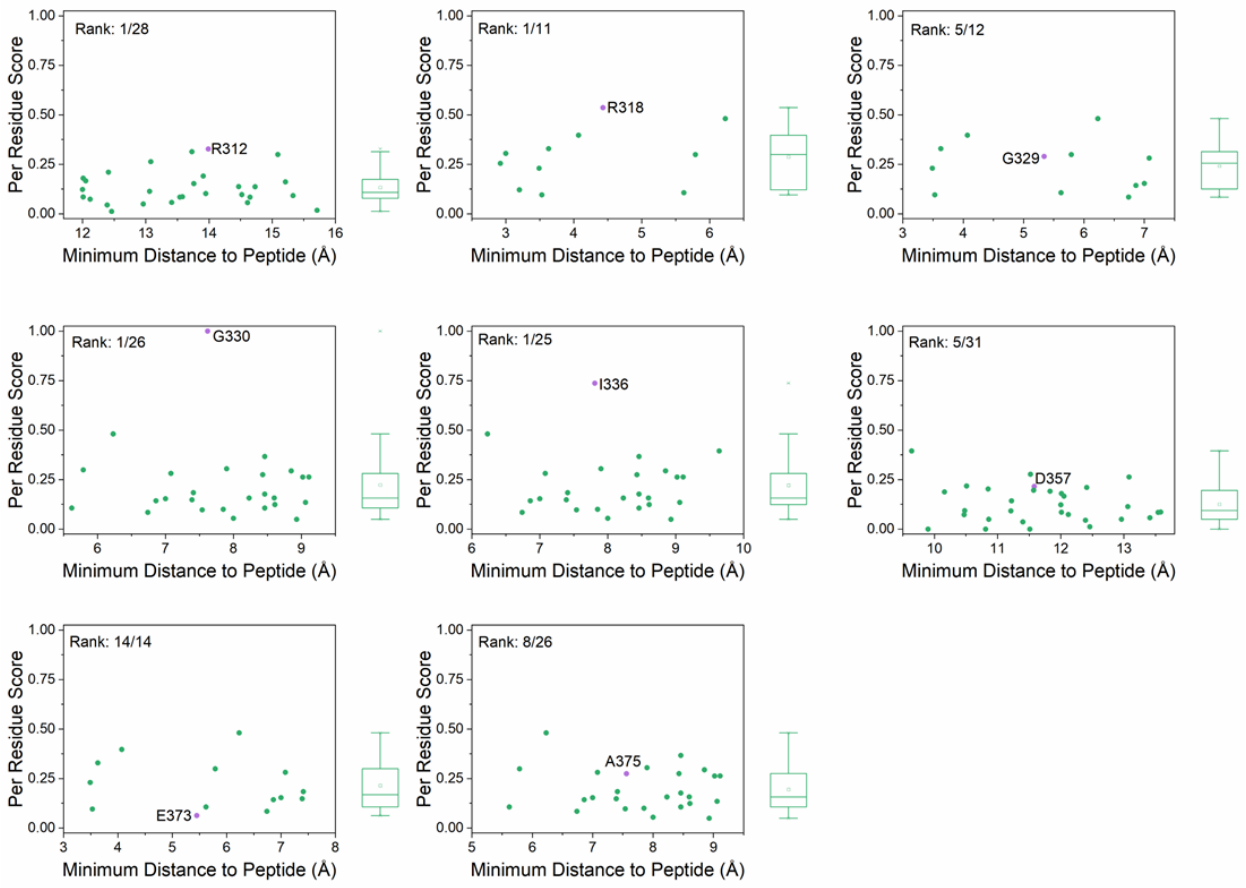

Figure S4. Comparison of the calculated per residue scores of the residues labelled by Faure *et al.,*[34] as "allosteric" (using exhaustive point mutation screening) to those labelled as "remainder" with a similar minimum heavy atom distance to the peptide. For each panel, an allosteric residue is compared to those remainder residues within 2 Å of the residue. The allosteric residue is colored purple and labelled, and the rank of the allosteric residue relative to the nearby remainder residues is provided. A box plot showing the distribution of the calculated scores is provided for each graph as well, with the box plot using the same y-axis scale as the scatter plots. The per residue scores were obtained from the JS-distance metric (see the **Methodology** section for further details).

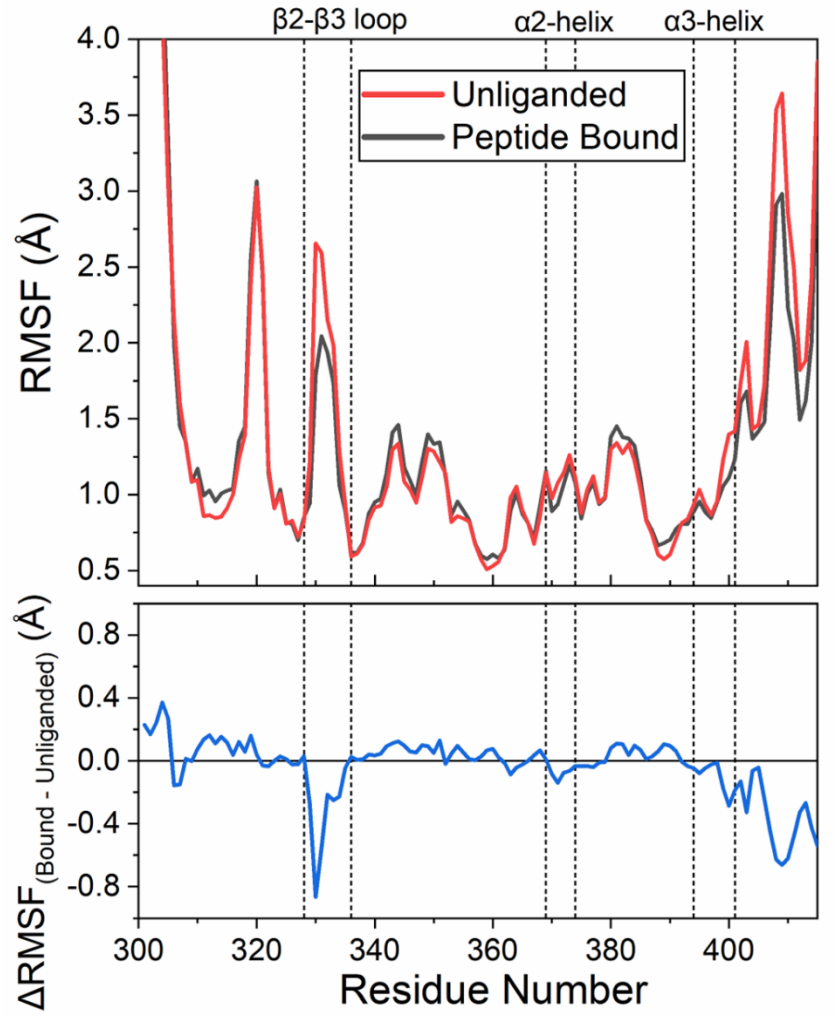

Figure S5. Calculated C_α-atom root mean squared fluctuations (RMSFs) for the simulations of the PDZ3 domain, in either the peptide bound or unliganded states. The ΔRMSF (peptide bound RMSF – unliganded RMSF) is plotted in the lower panel.

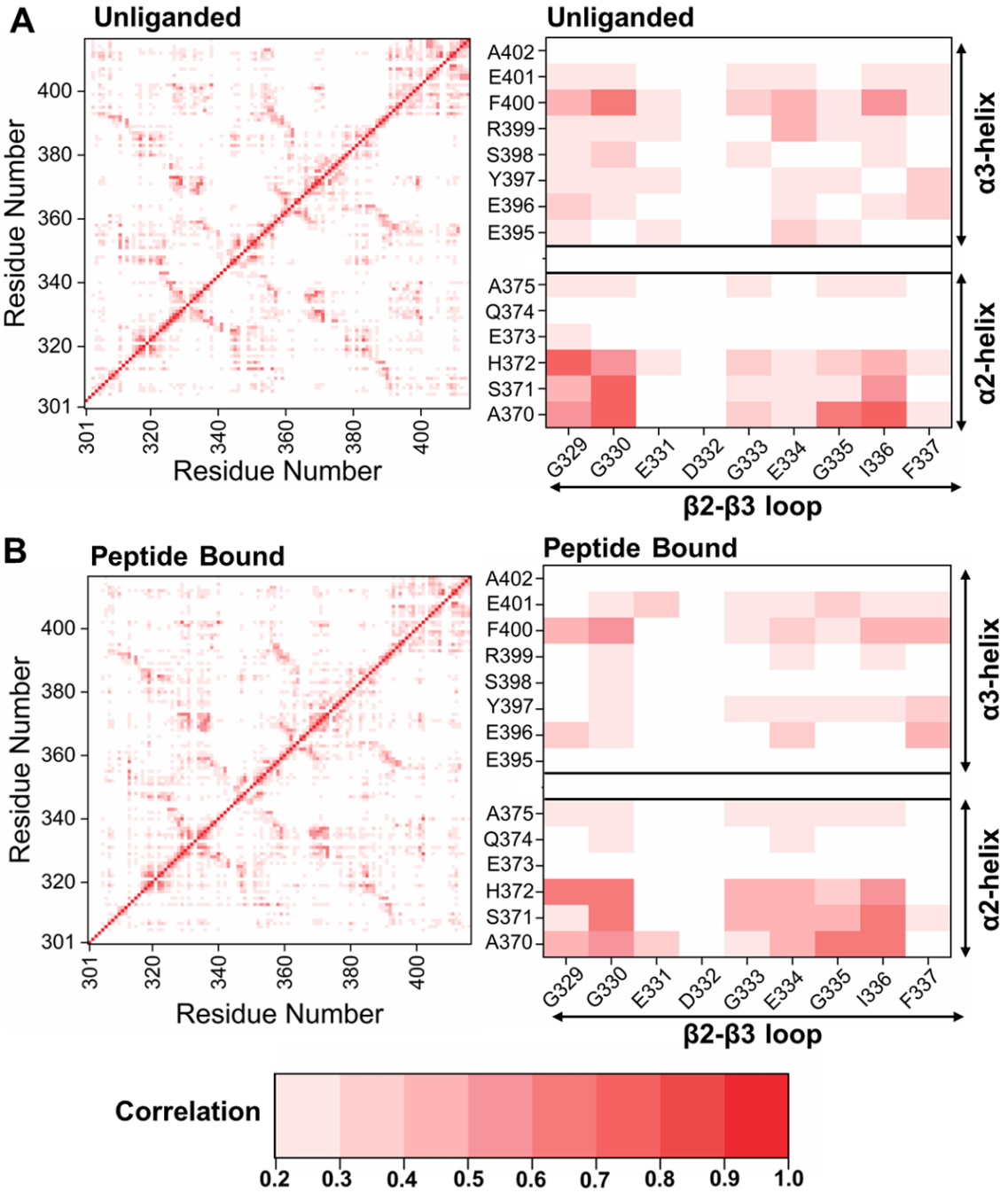

Figure S6. (**A**, **B**) The calculated linear correlation between each residue to every other residue during (**A**) unliganded and (**B**) peptide bound simulations of the PDZ3 domain. The panels on the right-hand side highlight the correlations between regions of interest, as described in the main text and shown in **Figure 4**. Residues correlations were determined directly from the non-covalent interactions (see the **Methodology** section), with the absolute correlation value colored according to the color bar. Any correlation less than |0.2| is shown in white

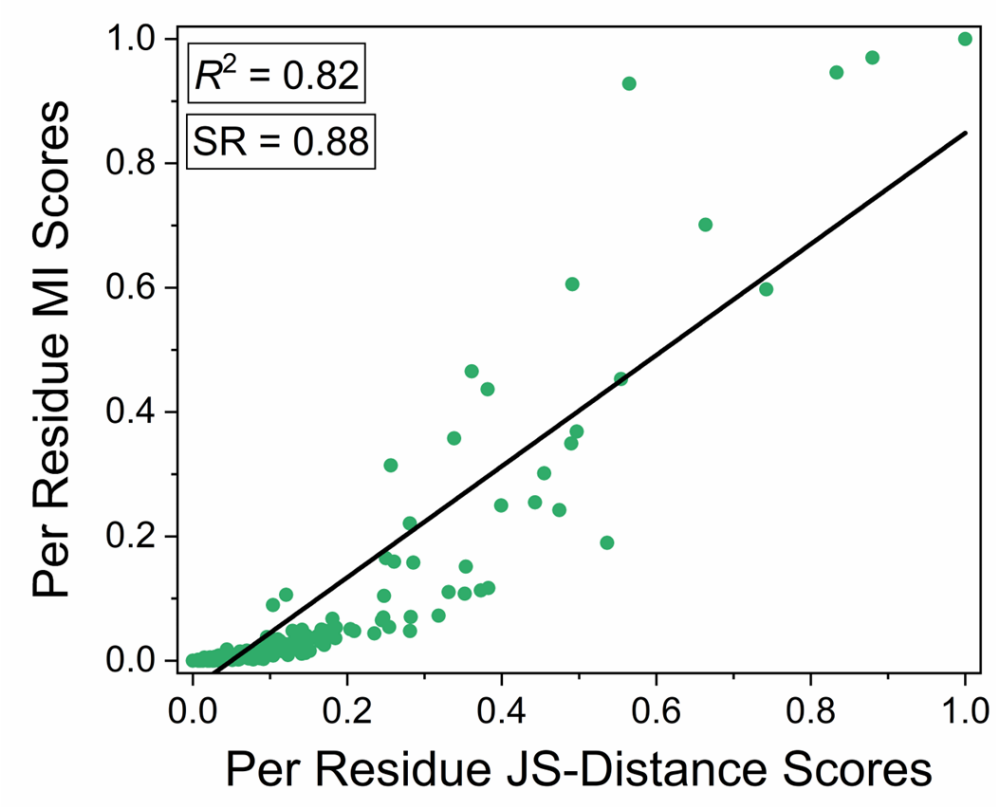

Figure S7. A comparison of the results obtained using the statistical analysis module of our program on our simulations of PTP1B, with the objective to distinguish between the two conformational states of the WPD-loop (see **Methodology** for further details). The calculated per residue scores for the two different metrics available (the Jensen Shannon (JS)-Distance,[29] and the mutual information (MI)[30, 31]), within our program are plotted against one another. A linear fit to the data is shown as a black line with the resulting R^2 value for the fit and the Spearman's rank (SR) is also provided. The SR measures how similar the rank ordering of the per residue scores are.

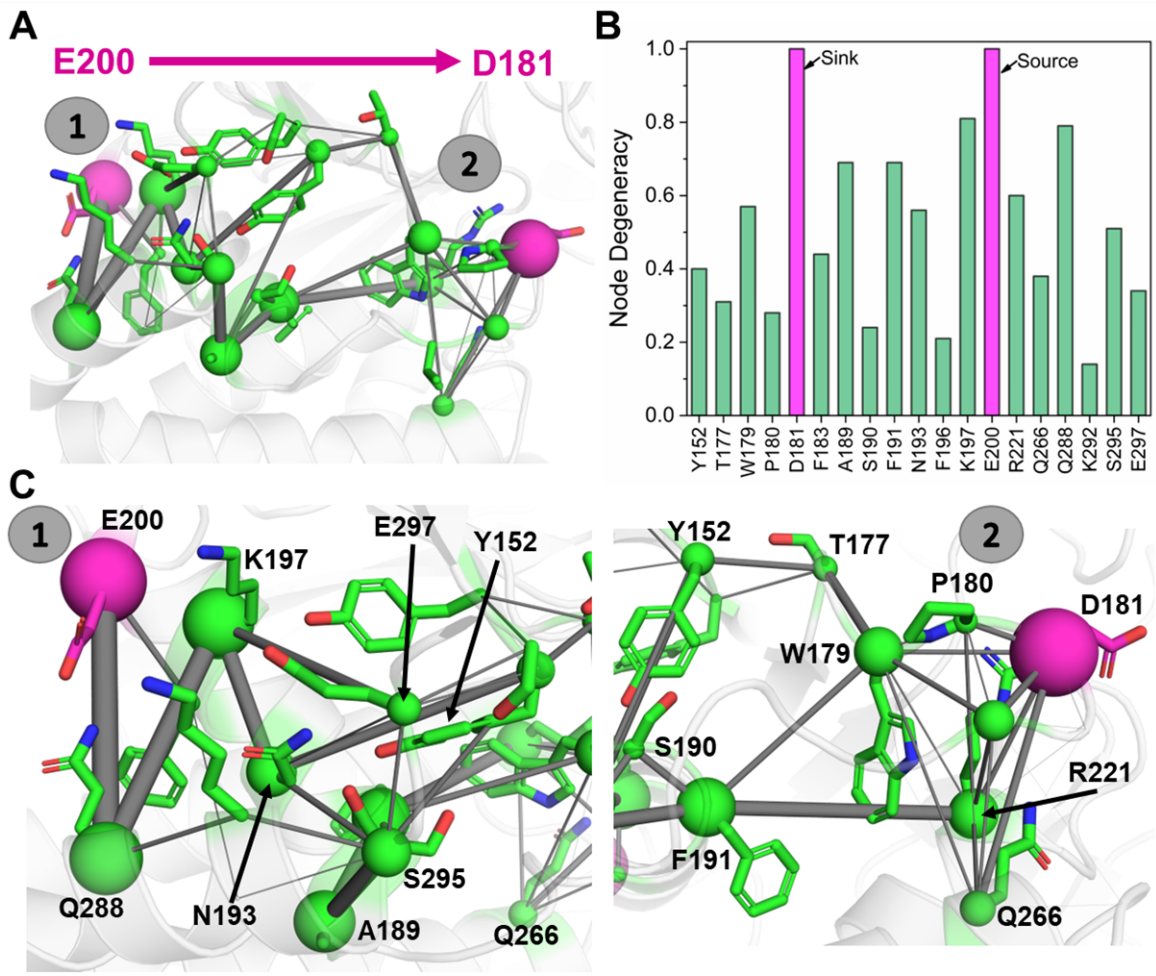

Figure S8. Overview of the WISP[35] calculations performed on PTP1B using E200 as the source and D181 as the sink. (**A**) Visualisation of the key residues and interactions along the path, with the source and sink residue colored purple. The nodes (spheres) represent the how frequently a given residue is present in the 500 paths generated, and edges (sticks between residues) indicate the frequency with which a certain residue-residue interaction is present in the 500 paths generated. In both cases, a larger size means an increased fraction and therefore increase importance in allosteric communication. (**B**) Node degeneracies (the fraction of pathways in which a given residue is present) for the 500 paths generated from E200 to D181. Only residues with node degeneracies ≥ 0.1 are shown. (**C**) A close up of two sections of the path, depicting the major residues and interactions present along the path. Equivalent results for the 500 pathways generated between G283 to D181 are shown in **Figure 6**.

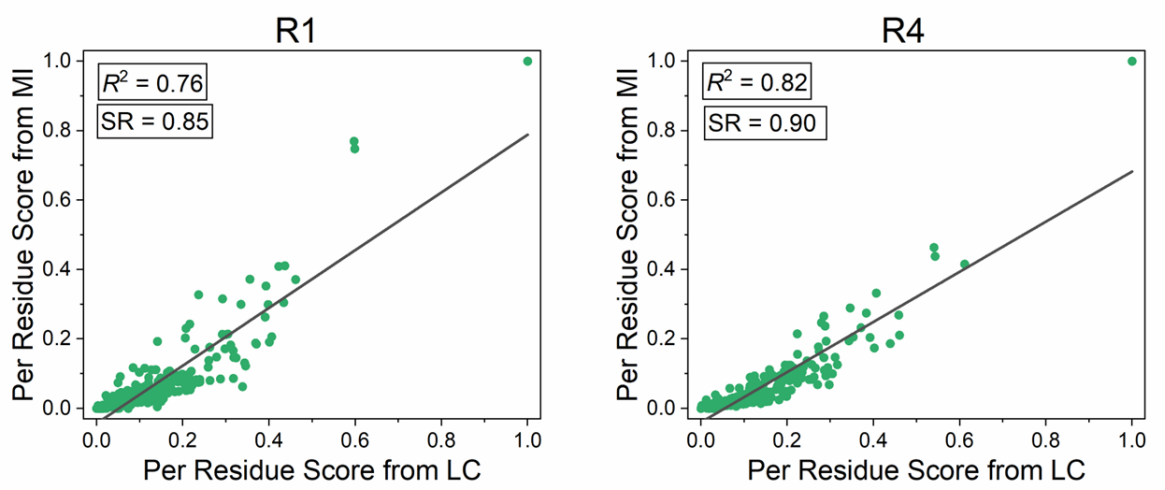

Figure S9. Comparison of the two different per residue scores calculation methods obtained from the regression analysis using the statistical modelling module for the R1 and R4 KE07 variants.[4-6] The target variable was the W50 χ_2 dihedral angle, and the metrics used were the linear correlation (LC) and the mutual information (MI)[30, 31] (see the **Methodology** section for further details). A linear fit for each scatter plot is shown as a black line and the R^2 value for the linear fit alongside the Spearman's rank (SR) is provided. The SR measures how similar the rank ordering of the per residue scores are.

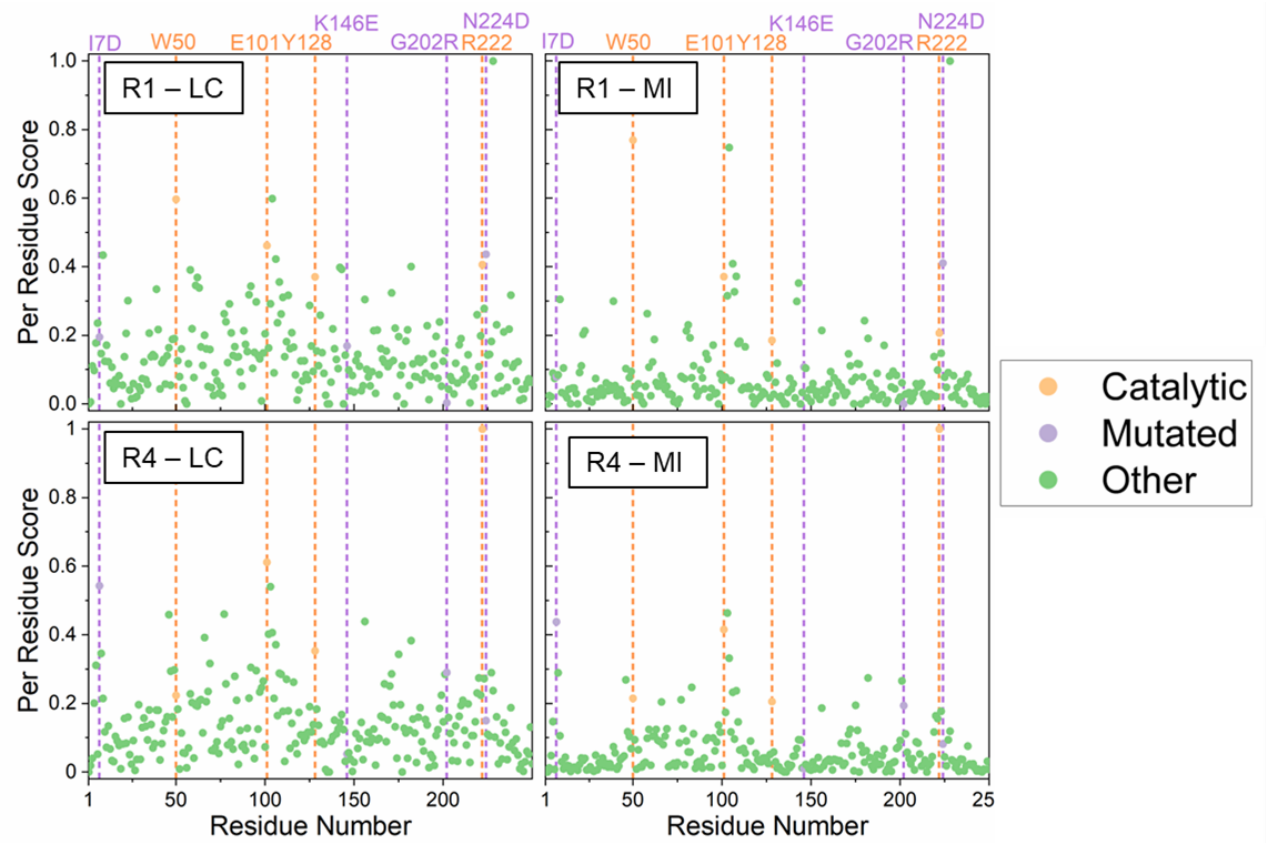

Figure S10. Per residue scores for the R1 and R4 KE07 variants,[4-6] obtained using the regression section of our statistical analysis package. In all cases, the target variable was the W50 χ_2 angle (continuous variable). The linear correlation (LC, left panels) and mutual information (MI, right panels) were used as the metrics to determine the per feature/interaction score and therefore per residue scores presented (see the **Methodology** section for further details). Residues that are considered key for catalysis (see **Figure 7C**) are shown in gold, and residues that are mutated between the R1 and R4 variants are shown in purple. All other residues are shown in green.

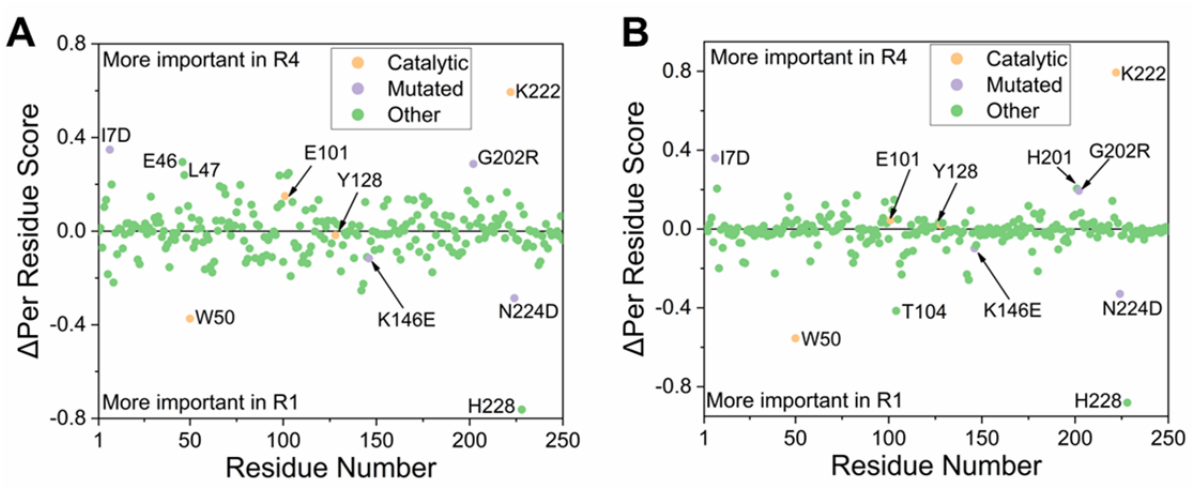

Figure S11. Difference (Δ) in per residue scores for the R1 and R4 KE07 variants,[4-6] obtained using the regression section of our statistical analysis package. The initial per residue scores were calculated for describing the difference in the W50 $\chi 2$ angle (and are shown in **Figure S10**). Shown here are (**A**) the difference in the per residue score values obtained using the linear correlation metric, and (**B**) the corresponding data obtained using the mutual information. Positive values indicate increased importance in the R4 variant, whilst negative scores indicate increased importance in the R1 variant.

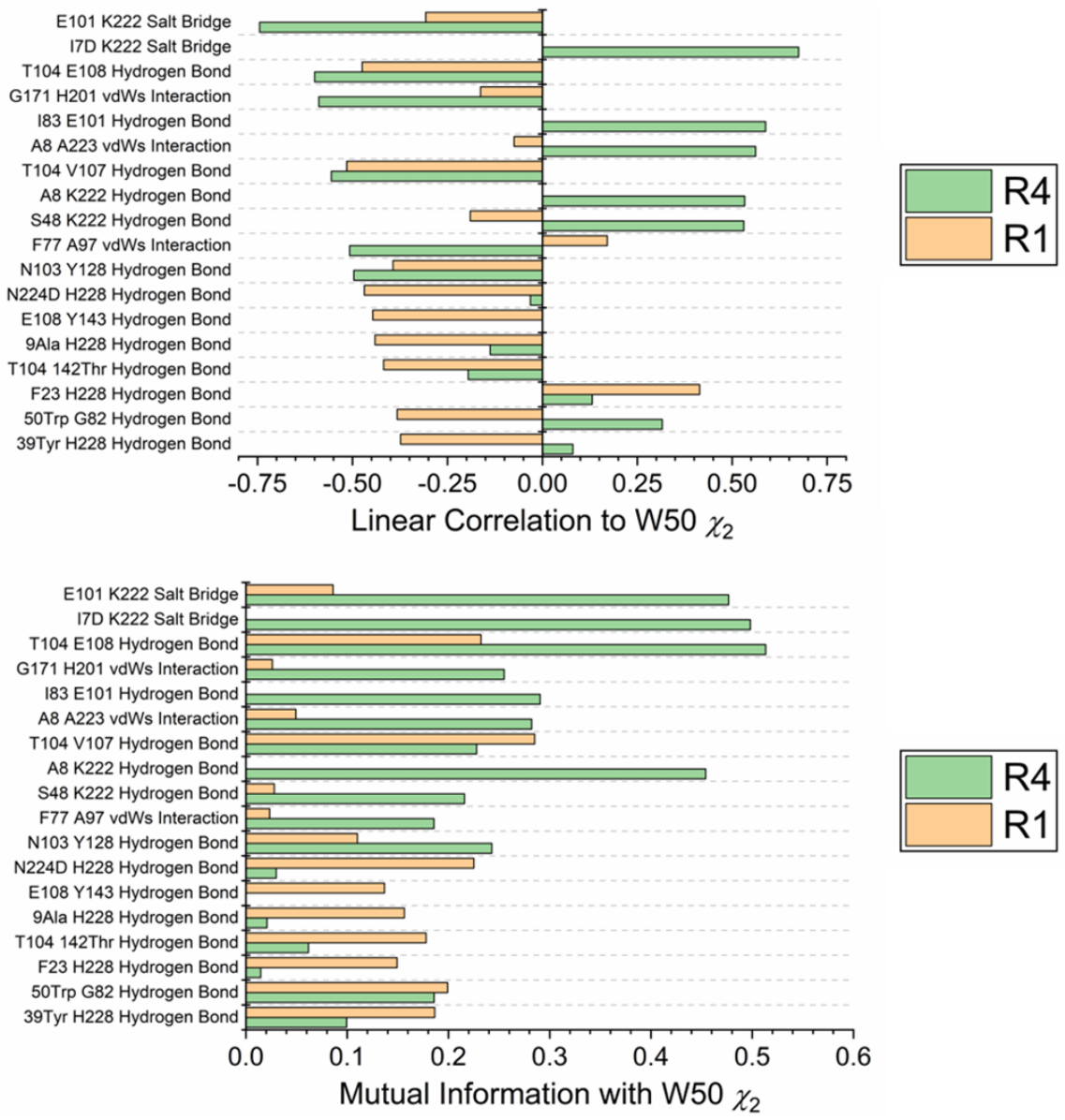

Figure S12. The top per feature scores obtained for both the R1 and R4 KE07 variants,[4-6] obtained from the regression analysis performed on the W50 χ_2 angle (see the **Methodology** section for further details). The top 10 features for both enzymes (as determined by their absolute linear correlation values) were taken forward to plot the graphs above. If a feature was in the top 10 for one variant and not the other, the corresponding per feature score was extracted for the other variant (provided that the interaction also exists in the other variant) and plotted as well.

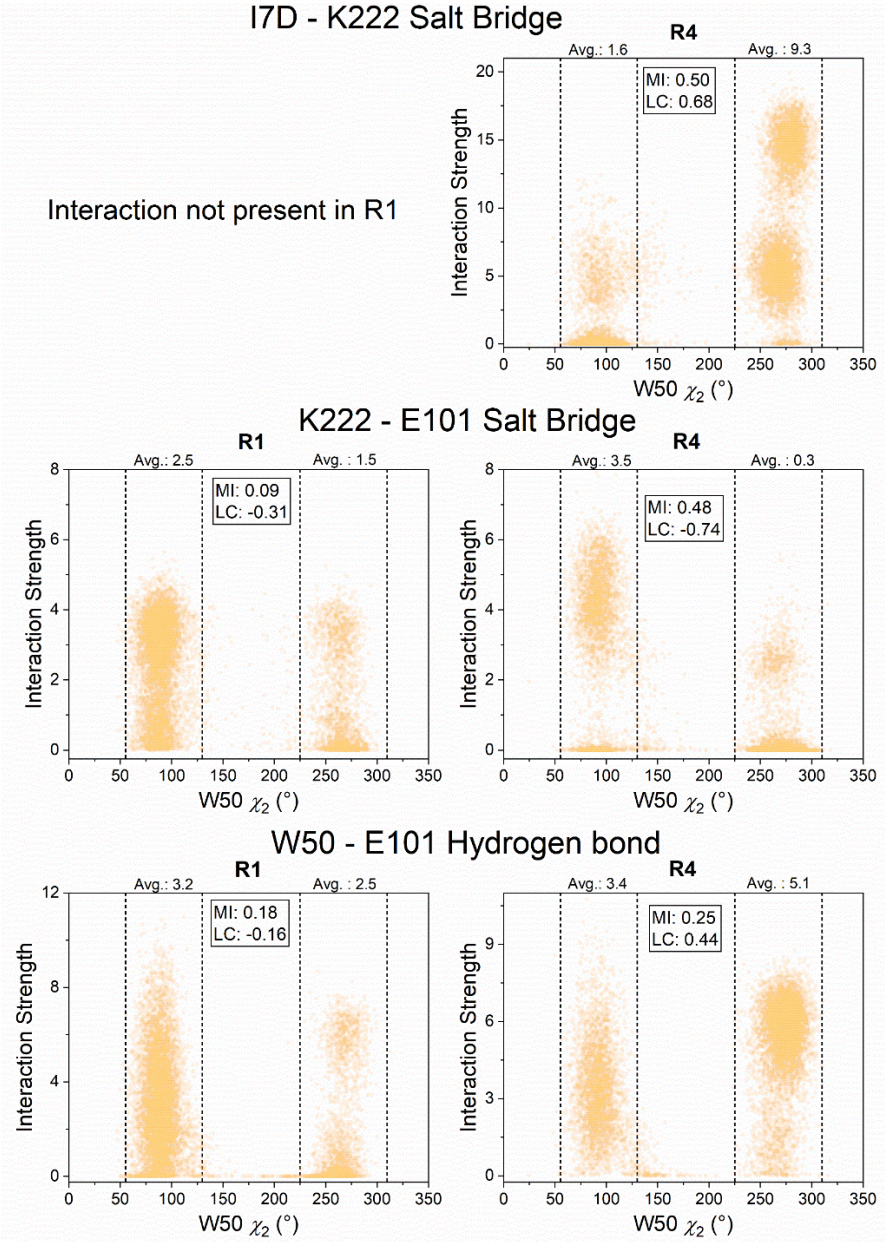

Figure S13. Scatter plots describing the role of the I7D mutation in the R1 and R4 KE07 variants[4-6] (I17 in R1, D17 in R4). The relationship between several interactions against the W50 χ_2 dihedral angle are plotted for both the R1 (top panels) and R4 (bottom panels) KE07 variants. The mutual information (MI)[30, 31] and linear correlation (LC) are provided for each scatter plot, with a negative LC indicating a stronger interaction in the "A" conformation. The average (Avg.) values of the interaction strength within the two ranges (depicted by dotted lines) are also provided, and describe the "A" (55-130°) and "C" (225-310°) conformational states.

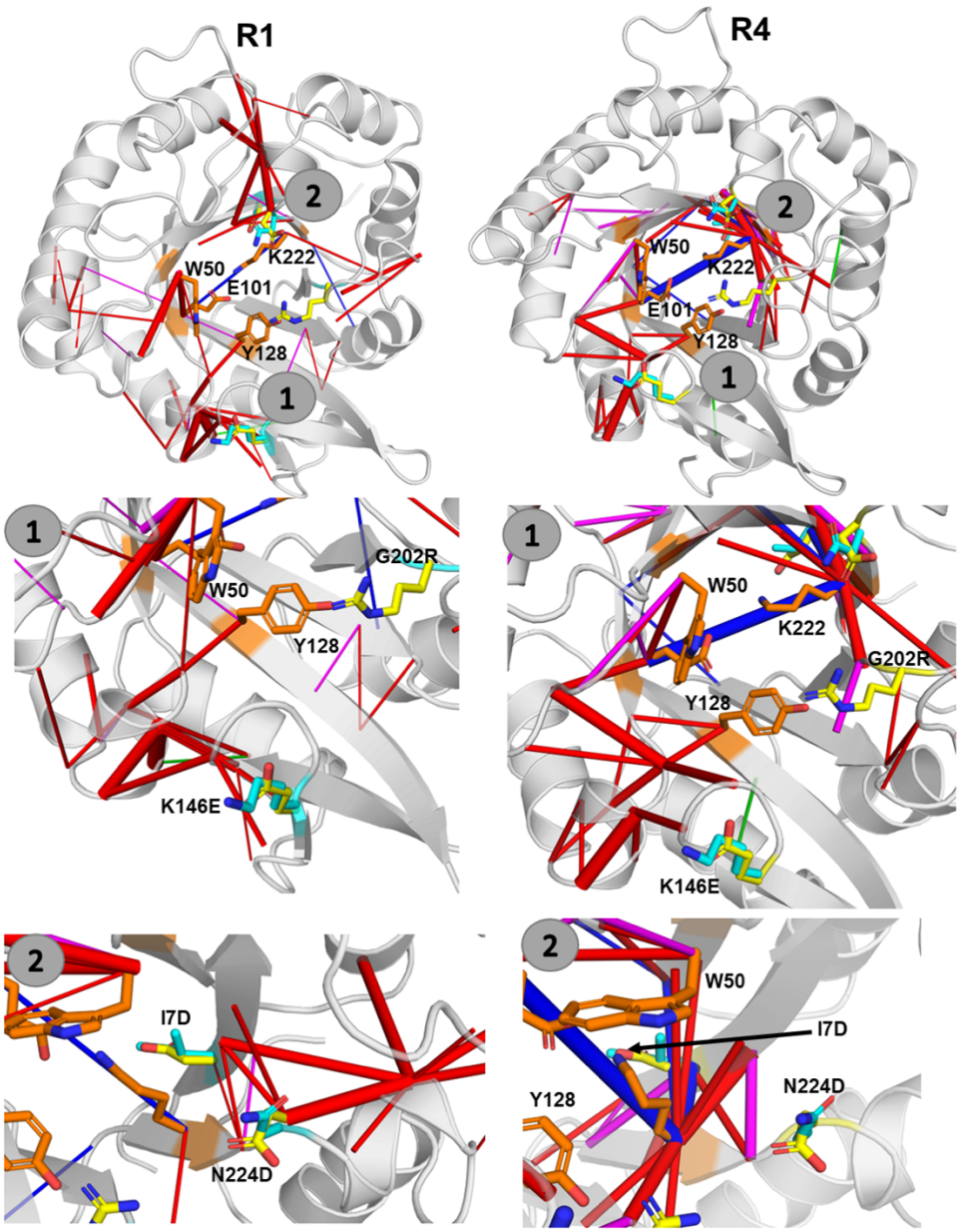

Figure S14. Visualization of the per feature scores determined for the R1 (left panels) and R4 (right panels) KE07 variants,[4-6] using our statistical analysis module with the mutual information metric[30, 31] used to calculate the per feature scores. The models were built to distinguish between the W50 χ_2 dihedral angle (see the **Methodology** section). The two lower panels focus on the mutation sites, with the R1 side chain (of the mutated residue) colored blue and the R4 side chain colored yellow. The larger the cylinder, the greater the score. Cylinders are colored according to their interaction type (red for hydrogen bonds, green for hydrophobic interactions, blue for salt bridges, and magenta for van der Waals interactions).

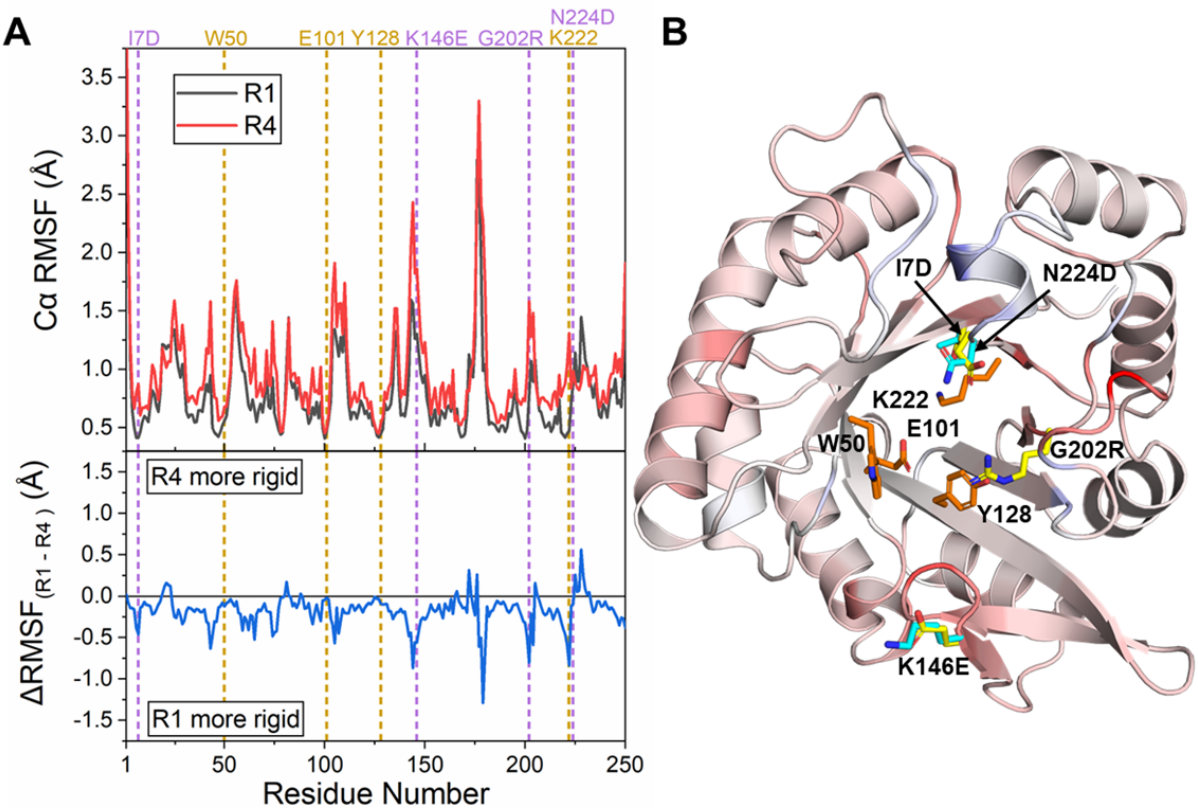

Figure S15. (**A**) Calculated C_α-atom root mean squared fluctuations (RMSFs) for the simulations of the R1 and R4 KE07 variants.[4-6] The ΔRMSF (R1 RMSF – R4 RMSF) is plotted in the lower panel. The location of the four mutations that separate the two KE07 variants are indicated in both plots in purple alongside the location of the four major residues that make up the active site (see **Figure 7**), which are colored in yellow. (**B**) Projection of the calculated ΔRMSF onto the structure of R1. Residues are color mapped from red (more rigid in the R1), through white (equally rigid in both KE07 variants) to blue (more rigid in R4). The key catalytic residues and mutations that separate R1 and R4 are indicated.

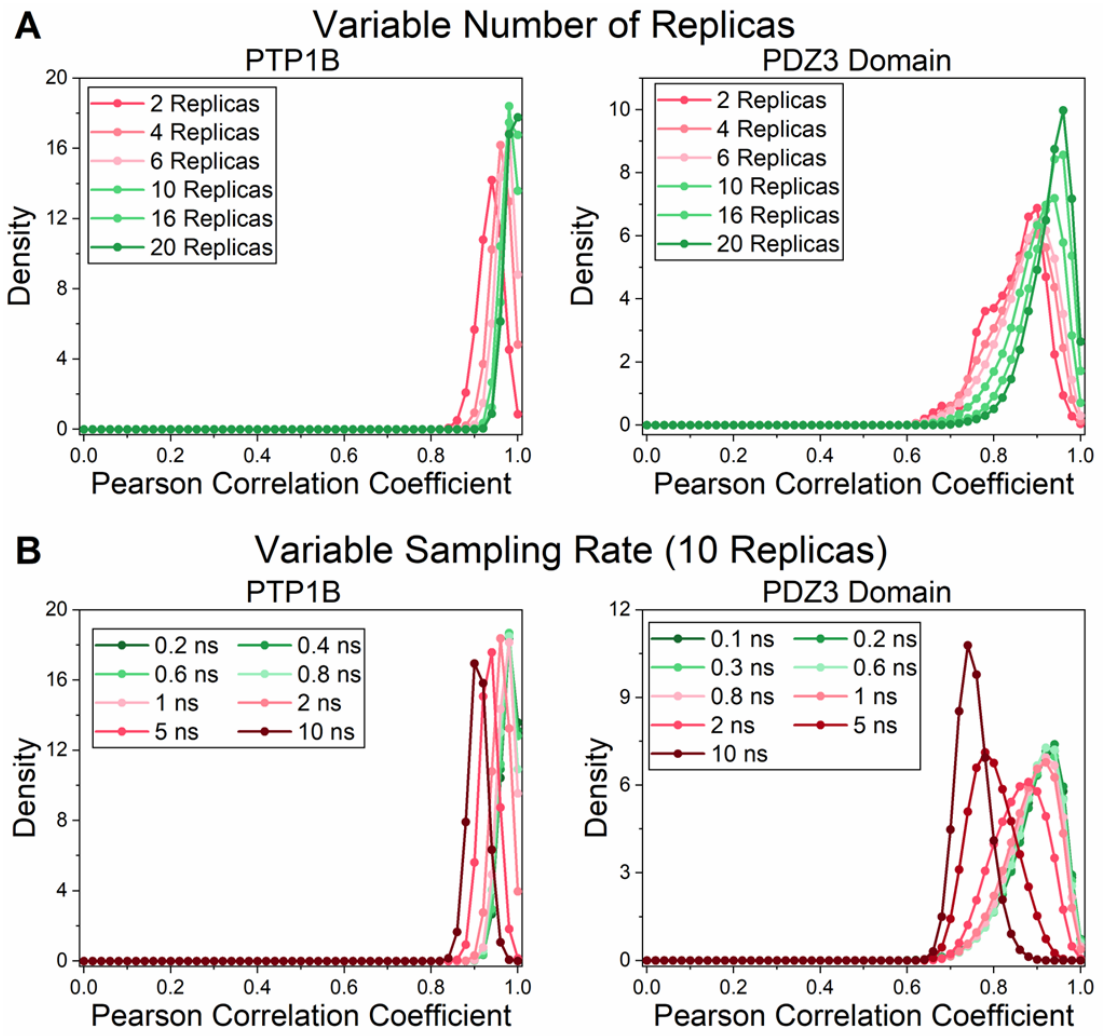

Figure S16. Gaussian kernel density estimates (KDEs) obtained from performing bootstrapping with replacement and random subsampling (5000 samples were generated for each scenario) on the per residue score calculations of PTP1B and the PDZ3 domain. (**A**) Bootstrapping with replacement and random subsampling performed using a differing number of simulation replicas. The Pearson correlation coefficient value for per residue scores obtained from each bootstrap sample against the result obtained from the 20 unique replicas, was used to build each KDE. (**B**) Bootstrapping with replacement and random subsampling using a random selection of 10 replicas for each sample with the time step between each frame taken altered (meaning those with larger timesteps also have correspondingly less frames available). For both PTP1B and the PDZ3 domain, an equal

number of simulations from both conformational states ("closed" and "open" WPD-loop or "peptide bound" and "unliganded" respectively) were used in the bootstrapping/resampling calculations, see the **Methodology** section for further details. Equivalent results using the Spearman's rank correlation are provided in **Figure 9**.

S6. SUPPLEMENTARY REFERENCES

[1] R. M. Crean *et al.*, J. Am. Chem. Soc. **143** (2021) 3830.

[2] R. Shen *et al.*, JACS Au **5** (2021) 646.

[3] R. Shen *et al.*, Chem. Sci. **13** (2022) 13524.

[4] D. Röthlisberger *et al.*, Nature **453** (2008) 190.

[5] O. Khersonsky *et al.*, J. Mol. Biol. **396** (2010) 1025.

[6] O. Khersonsky *et al.*, J. Mol. Biol. **407** (2011) 391.

[7] N.-S. Hong *et al.*, Nat. Commun. **9** (2018) 3900.

[8] D. A. Case *et al.*, AMBER 2018, University of California, San Francisco (2018).

[9] R. W. Pastor, B. R. Brooks, and A. Szabo, Mol. Phys. **65** (1988) 1409.

[10] H. J. C. Berendsen *et al.*, J. Chem. Phys. **81** (1984) 3684.

[11] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23** (1977) 327.

[12] G. Bussi, Mol. Phys. **112** (2014) 379.

[13] K. Lindorff-Larsen *et al.*, Proteins Struct. Func. Bioinformat. **78** (2010) 1950.

[14] W. L. Jorgensen *et al.*, J. Chem. Phys. **79** (1983) 926.

[15] D. van der Spoel *et al.*, J. Comp. Chem. **26** (2005) 1701.

[16] G. A. Tribello *et al.*, Comput. Phys. Commun. **185** (2014) 604.

[17] J. A. Maier *et al.*, J. Chem. Theory. Comput. **11** (2015) 3696.

[18] M. Parrinello, and A. Rahman, Phys. Rev. Lett. **45** (1980) 1196.

[19] M. Parrinello, and A. Rahman, J. Appl. Phys. **52** (1981) 7182.

[20] B. Hess, J. Chem. Theory Comput. **4** (2008) 116.

[21] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98** (1993) 10089.

[22] D. R. Roe, and T. E. Cheatham, J. Chem. Theory. Comput. **9** (2013) 3084.

[23] D. A. Keedy *et al.*, eLife **7** (2018) 1.

[24] M. Scheurer *et al.*, Biophys. J. **114** (2018) 577.

[25] L. Breiman, Mach. Learn. **45** (2001) 5.

[26] A. V. Dorogush, V. Ershov, and A. Gulin, arXiv (2018) 1810.11363.

[27] T. Chen, and C. Guestrin, arXiv (2016) 1603.02754.

[28] F. Pedregosa *et al.*, J. Mach. Learn. Res. **12** (2011) 2825.

[29] J. Lin, IEEE Trans. Inform. Theory **37** (1991) 145.

[30] C. E. Shannon, Bell Syst. Tech. J. **27** (1948) 379.

[31] J. G. Kreer, IEEE Trans. Inf. Theory **3** (1957) 208.

[32] Z.-Y. Zhang, Acc. Chem. Res. **36** (2003) 385.

[33] D. S. Cui *et al.*, J. Am Chem. Soc. **141** (2019) 12634.

[34] A. J. Faure *et al.*, Nature **604** (2022) 175.

[35] A. T. van Wart *et al.*, J. Chem. Theory Comput. **10** (2014) 511.