Combined free energy calculation and machine learning methods for understanding ligand unbinding kinetics

Magd Badaoui,[1,2] Pedro J Buigues,[2] Dénes Berta,[2] Gaurav M. Mandana,[1] Hankang Gu,[2] Callum J Dickson,[3] Viktor Hornak,[3] Mitsunori Kato,[4] Carla Molteni,[5] Simon Parsons,[6] Edina Rosta[1,2*]

[1] Department of Chemistry, King's College London, London SE1 1DB, United Kingdom

[2] Department of Physics and Astronomy, University College London, London WC1E 6BT, United Kingdom

[3] Computer-Aided Drug Discovery, Global Discovery Chemistry, Novartis Institutes for BioMedical Research, 181 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

[4] Schrödinger Inc., New York, New York, USA

[5] Department of Physics, King's College London, London WC2R 2LS, United Kingdom

[6] School of Computer Science, University of Lincoln, Lincoln LN6 7TS, United Kingdom

* e.rosta@ucl.ac.uk

ABSTRACT

The determination of drug residence times, which define the time an inhibitor is in complex with its target, is a fundamental part of the drug discovery process. Synthesis and experimental pharmacokinetics measurements are, however, expensive, and time-consuming. In this work, we aimed to obtain drug residence times computationally. Furthermore, we propose a novel algorithm to identify molecular design objectives based on ligand unbinding kinetics. We designed an enhanced sampling technique to accurately predict the free energy profiles of the ligand unbinding process, focusing on the free energy barrier for unbinding. Our method first identifies unbinding paths determining a corresponding set of internal coordinates (IC) that form contacts between the protein and the ligand, then iteratively updates these interactions during a series of biased molecular-dynamics (MD) simulations to reveal the ICs important for the whole of the unbinding process. Subsequently, we performed finite temperature string simulations to obtain the free energy barrier for unbinding using the set of ICs as a complex reaction coordinate. Importantly, we also aimed to enable further design of drugs focusing on improved residence times. To this end, we developed a supervised machine learning (ML) approach that uses as input unbiased "downhill" trajectories from the transition state (TS) ensemble of the string unbinding path. We demonstrate that our ML method can identify key ligand-protein interactions driving the system through the TS. Some of the most important drugs for cancer treatment are kinase inhibitors. One of these kinase targets is Cyclin Dependent Kinase 2 (CDK2), an appealing target for anticancer drug development. Here, we tested our method using three different CDK2 inhibitors for potential further development of these compounds. We compared the free energy barriers obtained from our calculations with those observed in available experimental data. We highlighted important interactions at the distal ends of the ligands that can be targeted for improved residence times. Our method provides a new tool to determine unbinding rates, and to identify key structural features of the inhibitors that can be used as starting points for novel design strategies in drug discovery.

KEYWORDS

Ligand-protein unbinding, molecular kinetics, CDK2, free energy calculations, machine Learning, Collective Variable selection, CV identification, feature selection methods, Transition State Analysis, MLTSA.

I. INTRODUCTION

Two essential factors describe the interaction between a drug and its target: binding affinity and residence time.[1] While the binding affinity describes the intermolecular interaction between the ligand and the protein, the residence time defines the timescale of the interaction.[2,3] Even if a drug interacts strongly with its target (high binding affinity), a short residence time can significantly reduce the efficacy of the drug.[4] The binding affinity arises from the thermodynamic relation between the stable bound and unbound states. The residence time, however, is determined by the path connecting those states, in particular, at the transition state of the unbinding pathway. Accordingly, promising hit candidates with high affinity have been discarded for the next step of the drug discovery process due to their low residence time.[5] Traditionally, drug discovery focused on finding compounds that interact with high binding affinity to a specific target. Recently, it is recognized that predicting pharmacokinetic properties is also a vital part of the drug design process.[6,7]

A major challenge in drug discovery is finding a fast and reliable method to predict kinetics of ligand-protein interactions.[8] Importantly, for experimental determination of ligand kinetics, ligands first need to be synthesized, which can be expensive and time-consuming even for a moderate number of compounds. Different experimental methods have been used to obtain kinetics of ligand-receptor unbinding, such as radioligand binding assays, fluorescence methods, chromatography, isothermal titration calorimetry (ITC), surface plasmon resonance (SPR) spectroscopy, and nuclear magnetic resonance (NMR) spectroscopy.[6,9]

Radioligand binding assays and fluorescence binding assays require binding with radiolabelled ligands, where they exploit the physical-chemical characteristics of the ligand between their free and complexed forms with the target. Several successful assays have been used to predict ligand-protein unbinding, for example fluorescence resonance energy transfer (FRET)[10] or fluorescence correlation spectroscopy (FCS).[11] These methods can suffer from interference (especially fluorescence), lack of accuracy for short residence times, and high cost/hazard in the case of radioligands.[12] SPR is the most widely used assay to measure rate constants associated with (k_{on} and k_{off}) of ligand-receptor unbinding. The receptors are immobilized to a sensor that can distinguish the protein from its ligand-free form to its bound forms. This method is label-free; however, the attachment of the protein to the probe may influence the activity of the protein, due to conformational changes.[12] To offer a screening approach that alleviates these difficulties, various computational techniques have been proposed as alternatives to estimating the kinetics of unbinding events.[13,14]

Molecular dynamics (MD) is a powerful computational tool to understand at an atomistic level the behaviour of biological processes such as protein-ligand interactions.[15] Unbiased MD simulations were successfully used in the initial stage of drug discovery process, using either multiple independent relatively short simulations,[16] or using specialized computer architecture, such as ANTON, where microsecond long simulations are readily accessible.[14] However, due to the limited timescales typically accessible via MD simulations, it is often challenging to obtain sufficient statistical sampling required to calculate kinetic and thermodynamic properties accurately. Drug-protein unbinding processes occur on long timescales, typically ranging from millisecond to hours, depending on the nature and the strength of the interaction between the ligand and target. Some drugs, for example, Aclidinium, Deoxyconformycin, or Tiotropium, have a half-life of hours,[17] requiring prohibitively long time scale simulations and highly demanding computer resources, therefore enhanced sampling methods are required.[18]

To accelerate the simulations and sample rare events, different enhanced sampling techniques have been proposed to predict free energy barriers and uncover the kinetics of biological events.[19,20] These methods include free-energy perturbation,[21,22] metadynamics (MetaD),[23,24] temperature-accelerated MD (TAMD),[25] steered MD (SMD),[26] milestoning,[27] umbrella sampling (US),[28] replica exchange,[29] scaled MD,[30] smoothed potential MD,[31] transition path sampling,[32] τ-Random Acceleration Molecular Dynamics Simulations (τ-RAMD)[33] and more recently a combination of enhanced MD with machine learning.[34,35] For most of these methods, a key factor is the identification of a collective variable (CV), representing a physical pathway, that allows the calculation of the free energy profile.[36] Hence, correct identification of appropriate CVs becomes a problem, with very few practical ways to build them properly.[37,38,39] These methods have already been used for ligand unbinding: for example, MetaD was used to predict the ligand-protein unbinding of p38 MAP kinase bound to type II inhibitors,[40] where depending on the set of CVs chosen, different values for k_{off} were obtained, and the closest k_{off} to the experimental data is still one order of magnitude lower. More recently, using the combination of MetaD with quantum mechanics/molecular mechanics (QM/MM) simulations, a more accurate prediction of the kinetics can be achieved.[41] The residence times of Sunitinib and Sorafenib in complex with the human endothelial growth factor receptor 2 have been calculated using SMD.[42] SMD was also used to calculate the unbinding free energy profile for TAK-632 and PLX4720 bound to B-RAF.[43] In both works, the ligands could be distinguished qualitatively to assess shorter, or longer residence times, however, the predicted free energy barriers for the unbinding were significantly lower than the experimental data.

To produce accurate free-energy profiles using biased simulations with many important degrees of freedom, we need to define an ideal set of CVs that map the full path of the reaction coordinate.[44,45] Usually, the vectors that describe this manifold are selected based on *a priori* chemical/physical intuition, typically based on the initial binding pose of the ligand. The same set of CVs are then kept constant and used for the full simulation. Considering only CVs from an initial structure implies possibly neglecting essential interactions that occur during the unbinding process, thus significantly affecting the free energy calculation. Additionally, structures resolved by X-ray crystallography or cryo-EM may capture the system in metastable states, which does not always reflect appropriate conformers for ligand binding.

In this work, we introduce a novel enhanced sampling method to obtain accurate free energy barriers for ligand-protein unbinding. Unlike existing methods, we also propose a method that subsequently can identify key molecular features determining the unbinding kinetics. We suggest an iterative way of assigning our CVs during the unbinding trajectory and then using these CVs as the driving force to pull the ligand out from the pocket and to perform the sampling for accurate free energy calculations. Similarly to *e.g.*, τ-RAMD (which, however does not provide a free energy profile), there is no need to *a priori* select CVs; these naturally arise from the unbinding trajectories that build a reliable path of unbinding taking the flexibility and dynamics of the system into consideration.

The CVs extracted from our trajectories sufficiently describe a full pathway for the unbinding process. Subsequently, we optimize this path in the space of the identified CVs to obtained a minimum free energy profile using the finite temperature string method.[46] While different unbinding trajectories might lead to slightly different variations due to multiple local minima along the paths, we typically expect that the main transition state ensembles would be captured by all of these paths similarly after the convergence to the minimum free energy pathway. This is the main underlying assumption behind the finite temperature string method, which was proven to work very well even for complex systems.[47,48] Our results accordingly show little variations in the unbinding free energy barriers using different starting pathways for free energy calculations.

In addition to determining unbinding rates, we also aim to identify key molecular descriptors that provide guidance for further design of drugs based on improved residence times. We propose a systematic approach to identify key low-dimensional sets of internal coordinates using machine learning (ML) approaches. Machine learning methods have been widely successful in multidimensional data driven problems, which are also applied to biomolecular simulations to determine key CVs.[49–51] Here, we develop a novel approach making use of our obtained string unbinding pathway and, within that, the knowledge of the transition state (TS) ensemble. We explored two different ML methods in this study: Neural Networks (NN),[52] which provide efficient training on complex high-dimensional data, and Gradient Boosting Decision Trees (GBDT),[53] which allow straightforward evaluation of feature importances (FI).[54] We generate unbiased "downhill" trajectories initiated at our TS, and use these to train a ML model which predicts the fate of binding or unbinding.

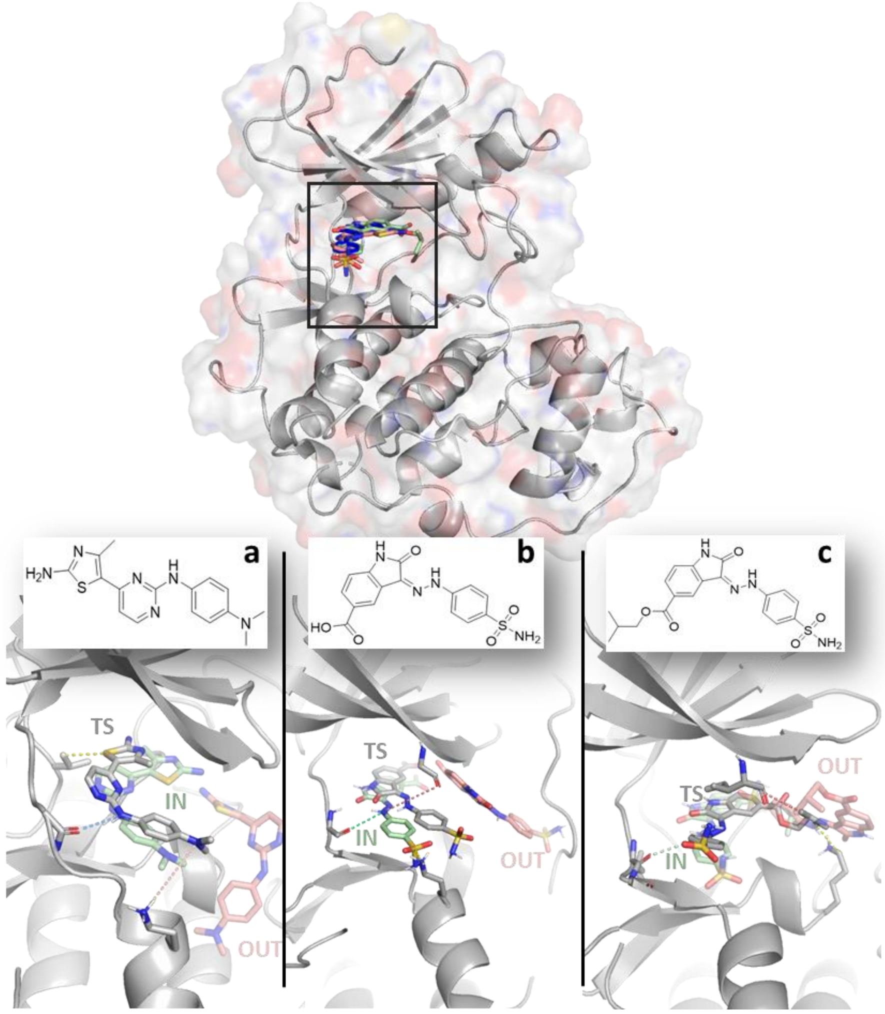

FIG. 1. CDK2 bound to three different ligands: **a** thiazolyl-pyrimidine derivative (18K), **b** oxindole carboxylic acid derivative (60K), and **c** carboxylate oxindole derivative (62K), originated from PDB structures 3sw4, 4fku, 4fkw, respectively. Structural details of the ATP pockets are shown for the three systems (bottom), with the ligands in the bound (green sticks), unbound (red sticks), and transition states (grey sticks). Dashed lines depict key interactions.

To test this approach on a simple analytical model system, we generated trajectory data using a collection of 1D model potentials, including one selected double-well potential. Our results demonstrate that our novel ML analysis can identify the key features correlated to this selected double-well potential to define the end states and thus can be used for key feature selection successfully. To demonstrate the applicability and accuracy of this approach on challenging complex biomolecular systems, we obtained free energy barriers for three ligands bound to CDK2 with PDB IDs of 3sw4 (*18K*), 4fku (*60K*), and 4fkw (*62K*) (FIG. 1).[55] Cyclin Dependent Kinase 2 (CDK2) is a crucial regulator in eukaryotic cell growth, deregulation of CDK2 has been associated with unscheduled cell proliferation resulting in cancer progression and aggressiveness.[56,57] Selective inhibition of this protein makes it an appealing target in treating multiple tumours of specific genotypes.[58] Several molecules are currently under clinical evaluation as CDK2 inhibitors for cancer treatment, such as AT759,[59] AG-024322,[60] Dinaciclib,[61] Roniciclib,[62] Milciclib.[63] Furthermore, CDK2 is an ideal benchmark system with its relatively small size and well-documented kinetic data for the binding of a range of different molecules.[55]

II. METHODS

All MD simulations are carried out in NAMD 2.12, [64] using AMBER ff14SB force field for the protein,[65] and using the general Amber force field (GAFF) for the ligands.[66] The MD simulation setup are detailed in Supplementary Note 1.

UNBINDING SIMULATIONS

Our unbinding method is illustrated algorithmically in FIG. 2. An explorational unbiased MD simulation of at least 20 ns was performed to identify the initial interactions between the protein and the ligand in the bound state. These initial simulations allow us to define the first set of CVs describing all distances between the heavy atoms of the ligand and the heavy atoms of the protein smaller than d_{in} = 3 Å, our interaction cut-off. The identified interactions will generate a single one-dimensional CV as the sum of these M distances, d_i, and will be used for iteratively biasing the simulations to observe an unbinding trajectory.

At every iteration, we will define our bias as a harmonic restraint: $V = \frac{1}{2}k\left(D - \sum_{i=1}^{M} d_i\right)^2$, where $D = D_0 + (Md_{tar})$. Here, we aim to reach the target value D for our 1D CV starting from the initial value at the beginning of the n^{th} iteration D_0. d_{tar} is the incremental factor, set to 1 Å, representing the average increase we aim to achieve per distance for the next iteration. The targeted D value will be reached progressively within the next 10 ns long MD simulation for every iteration. The force constant, k, was set to 20 kcal/mol/Å^2.

At the end of each iteration, the biased trajectory is analyzed, and novel interactions are identified, within d_{in} of the ligand, that are present for more than half of the total simulation time (*i.e.*, 5 ns). These novel interactions are then added to the list of interactions that define the main CV for the next iteration. Additionally, we also re-evaluate existing interactions. If a distance during the last 5 ns of the trajectory exceeds d_{out} = 6 Å or its variance exceeds d_{var} = 1 Å, then the distance is removed from the main CV in the next iteration. This exclusion factor will ensure that once a protein-ligand atom pair distance has exceeded d_{out}, and therefore there is no significant interaction between these atoms, we no longer bias this interaction. Similarly, loosely interacting atom pairs have higher distance fluctuations, and thus the corresponding weak interaction does not need to be included in the bias.

To reduce the number of interactions between the ligand and the protein and to remove redundancies, we combine atoms that are part of an equivalent group where a rotational degree of freedom can interconvert the atoms from one to the other (for example, benzene ring or carboxylic groups, see Fig. S1). Here, we consider the centre of mass of that functional group and not the individual atoms.

The iterative process will end when no more distances are present in the main CV from the last iteration *n*, thus there are no more stable interactions between the ligand and the protein, suggesting that the ligand is outside the binding pocket. Fig. S2.1-S2.9 represent the distances included in the unbinding trajectories.

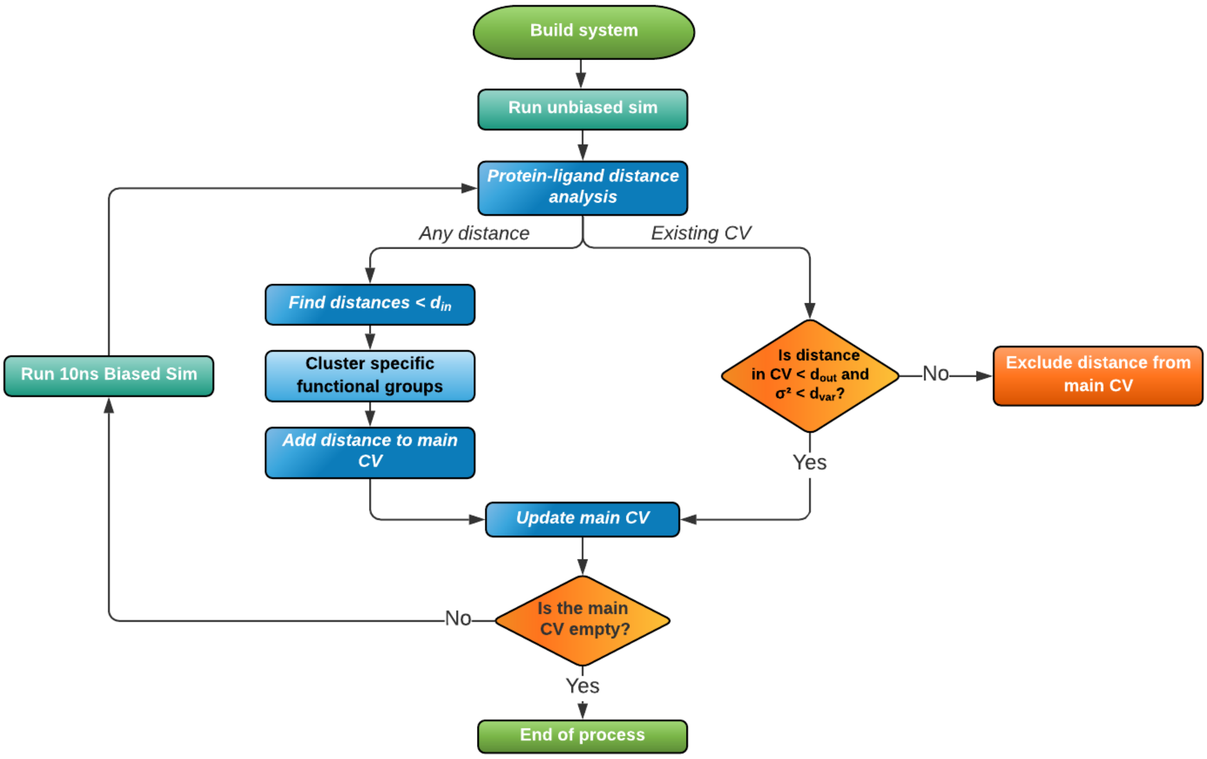
Build system
Run unbiased sim
Protein-ligand distance analysis
Any distance
Existing CV
Find distances < d_{in}
Cluster specific functional groups
Add distance to main CV
Is distance in CV < d_{out} and $\sigma^2 < d_{var}$?
No
Exclude distance from main CV
Yes
Run 10ns Biased Sim
Update main CV
No
Is the main CV empty?
Yes
End of process

FIG. 2. Flowchart illustrating the steps for the unbinding protocol.

B. FREE ENERGY CALCULATIONS

Once the ligand is outside of the binding pocket, to determine the minimum free energy path for the unbinding trajectory, we use the finite-temperature string method.[46] The initial path and the full set of distances (CVs) are taken from the obtained unbinding trajectory.[46,67,68] We extract the these CV values for each interatomic distance along the initial unbinding path to construct the minimum free energy unbinding pathway iteratively, building a string of 100 windows in the coordinate space. For each window and each CV, we apply a position restrain equidistantly along the initial fitted string, using a force constant of 20 kcal/mol/Å^2. We perform biased simulations using these restraints for a total time of 5 ns per window. From the obtained set of trajectories, a high-order (8) polynomial fitting is applied using the average values for each collective coordinate to build the subsequent set of refined CV positions. The procedure is carried out iteratively until the convergence of the free energy profiles and the pathway. This is verified by ensuring that the maximal change of each CV between subsequent iterations is below 7% (or 0.3 Å) from the previous iteration. By adding multiple overlapping biasing potentials along the dissociation pathways which are parametrized via the identified CVs, the string simulations can sufficiently sample the high dimensional path describing the full unbinding trajectory in detail. Finally, to obtain the corresponding Potential of Mean Force (PMF), we unbias the simulations using the binless implementation[46] of the weighted histogram analysis method (WHAM).[62]

C. MACHINE LEARNING TRANSITION STATE ANALYSIS (MLTSA)

We developed a Machine Learning Transition State Analysis (MLTSA) method to identify novel descriptors that determine the fate of a trajectory from the TS, which is applicable to unbinding simulations, but also suitable for other applications as a low-dimensional feature selection method for highly complex processes where a TS region is identified. In our case, the novel molecular interactions between the drug molecule and the protein for unbinding provide key signatures that determine the unbinding kinetics.

To test the validity of the MLTSA, we created an analytical model and compared the ability of two ML approaches to detect correlated features: a Multi-Layer Perceptron (MLP) architecture NN model and Gradient Boosting Decision Trees (GBDT), a common ML approach in feature selection.

The analytical model was based on using multidimensional trajectories generated via a set of one-dimensional (1D) free energy potentials (see details in SI Analytical Model System). Two types of potentials were used, both a set of single-well (SW) and double-well (DW) potentials. We used all but one of the DW potentials as "noise" and one of the DW potentials to define the outcome of the process, as the decisive coordinate to classify trajectories as "IN" or "OUT". We generated trajectories using Langevin dynamics along 25 1D potentials. We used these trajectories to define 180 input features analogously to our observable CV-s by computing linear combinations of the original coordinates (see details in SI Analytical Model System). In our example, 11 of these 180 contained the selected DW potential with some non-zero coefficient (see Table S1). We used these set of CVs to train the ML methods to predict the trajectory outcomes. Importantly, we aimed to identify the CVs that had the largest coefficients for our key selected DW potential.

We trained the MLP to analyze the model datasets of the downhill trajectories and predict their possible outcome from early on data, *i.e.*, at 30-60 steps of the downhill trajectory for the analytical model. The training was performed using the Scikit-learn library.[70] We trained a simple model with an MLP Classifier architecture, using three main layers (input, hidden, and output) with as many input nodes as input features depending on the system of study (for the analytical model 180 were used, for CDK2 see Table S1.II), fully connected to a hidden layer with 100 hidden neurons and ending in an output layer with one output node each for IN or OUT classifications. The model was optimized using the Adam solver[71] and using the ReLu[72] function as an activation function for the hidden layer. The training was done with a learning rate of 0.001, iterating over data until convergence or upon reaching the maximum number of iterations (500 epochs). Convergence is determined by the tolerance and the number of epochs with no change in loss. When there are 10 consecutive epochs with less than 0.0001 improvement on the given loss, the training stops, and convergence is reached. The same parameters were used for both the analytical model and CDK2 data.

We also tested the GBDT model using the Scikit-learn library as a comparison to the MLP approach. This method provides FI that enable the ranking and identification of relevant features. We trained 500 decision stumps as weak learners for GBDT minimizing a logistic loss function, with a learning rate of 0.1. The criterion for the quality of the splits was the Friedman Mean Squared Error (MSE), with a minimum of 2 samples to split an internal node, and a minimum of 1 sample to be at a leaf node. The maximum depth of the individual regression estimators was 3, without a limit on the maximum number of features to consider as the best split, without maximum on leaf nodes and using a validation fraction of 0.1. Same parameters were used for both the analytical model system and the CDK2 simulations.

The flowchart of the MLTSA method is illustrated in Fig. S3. For the analytical model, we run 180 trajectories for the ML training and a separate validation set with 50 additional unseen trajectories. Following the flowchart, after labelling them as “IN” or “OUT” using the decisive coordinate, we created a dataset for the ML algorithms containing 180 features per frame. We trained the ML models at different time frames (see Fig. S5) to observe the evolution of the accuracy throughout the simulations. This allows us to find a time range in the simulations where the classification problem is neither hard nor too trivial. Using this range, we trained the MLP model to analyze the importance of the features with our novel method. In a similar fashion to feature permutation[73,74], or other model inspection techniques[75–77], the MLTSA uses the Global Mean (GM) approach[76], which swaps the value of each feature, one at a time with the mean value of the feature across all data used for training. This altered dataset is used for prediction again expecting to get the same accuracy as the training on non-correlated features and an accuracy drop on the correlated features, which depends on the level of correlation. For the comparison with GBDT and its FI, we trained the model on the same time and fetched the FI from the model to compare it with the accuracy drop analysis.

For the application of the MLTSA on CDK2, first we identified the approximate TS location by selecting the last simulation frames from the highest energy five windows near the TS point of the obtained PMF. From each of these five starting coordinates, we then run 50 independent unbiased MD simulations 5 ns long each. We classify and label these short ‘downhill’ trajectories by considering a combination of two key distances (see Table S1.I), to identify which simulations finish either in a ligand bound position (IN) or in a ligand unbound position (OUT). We then select the starting structure (i.e., our TS) that provides the closest to a 1:1

ratio of IN and OUT events amongst these trajectories, and we run 200 additional 5 ns-long unbiased MD simulations with this starting point. We consider all interatomic distances (heavy atoms only) between the ligand and the protein that are within 6 Å at the TS starting position and determine the values of these distances along downhill trajectories. These constitute a dataset of distances for each simulation trajectory, and we aim to select the most important features from these with our MLTSA method.

The number of epochs and convergence of the loss function for each model can be found in the Tables S4.I – S4.III and Fig. S6. Thus, using the frames coming from the multiple short unbiased MD simulation trajectories starting from our TS, we provided a dataset of distances extracted along the trajectory, as well as the future outcome of the IN or OUT events as the desired answer/classification. We performed the training with trajectories of several different lengths as well as different time frames (Fig. S4), to observe the predicted accuracy at different time ranges along the simulations. From all the available trajectories for each system we reserve a part for further validation to avoid the overfitting of our model. The rest is used for training, with all frames from the trajectories concatenated and randomly mixed, then split in different fractions as training (0.7) and test (0.3) sets, which is used to assess if the model is learning appropriately. The trained model is additionally verified to have a similar prediction accuracy on the unseen trajectories.

Using our trained model, we assess which features are the most important for the model to predict whether the simulation is classified as bound (IN) or unbound (OUT). To do so, we apply our own feature reduction approach (FR), in which every single distance (i.e., feature) is excluded one-by-one from the analysis, and we calculate the drop in accuracy compared to the full set of distances present. Differently from the standard approach,[66] where the real value of each excluded feature is replaced with a zero, here we replace the value for each excluded feature with the global mean of that selected feature across the simulations, thus cancelling the variance of the aforementioned feature.

III. RESULTS AND DISCUSSIONS

MLTSA analytical test and validation

ML training on the model potential-derived trajectories was performed with both MLP and GBDT ML methods. We performed the MLP training at different time frames and trajectory lengths, from the 0th time step to the 500th step in intervals ranging from 10 to 150 frames at a time to assess the accuracy through time (Fig S5). Using a suitable time range consisting of the 30th-60th simulation steps from each trajectory, the trained ML methods found the classification problem accurately solvable, but not too trivial. We performed 100 replicas of the full process from generating 180 new independent simulations for each replica and performing the ML training. The MLP achieved an average test accuracy over 94% and an average validation accuracy over 93% whereas the GBDT achieved over 99% on the test set and 91% on the validation set. One notable difference between the training on both models is the computational time consumed, on MLP the 100 replicas took 5.85h whereas the GBDT took 8.33h which is 1.4x times longer than the MLP.

The highest correlated features (colored depending on the correlation level, color bar in Fig. 3 right panel) were correctly identified by both MLP and GBDT models. For GBDT, only the top three features show a high FI value (labels added to datapoints in Fig. 3), whereas the rest of the correlated features ranging from α~34% up to ~60% do not show a significant FI value. In addition, three features (#48, #89 and #136) despite having 40.34%, 34.80% and 35.48% mixing coefficients, respectively, GBDT did not capture their correlation, showing values very

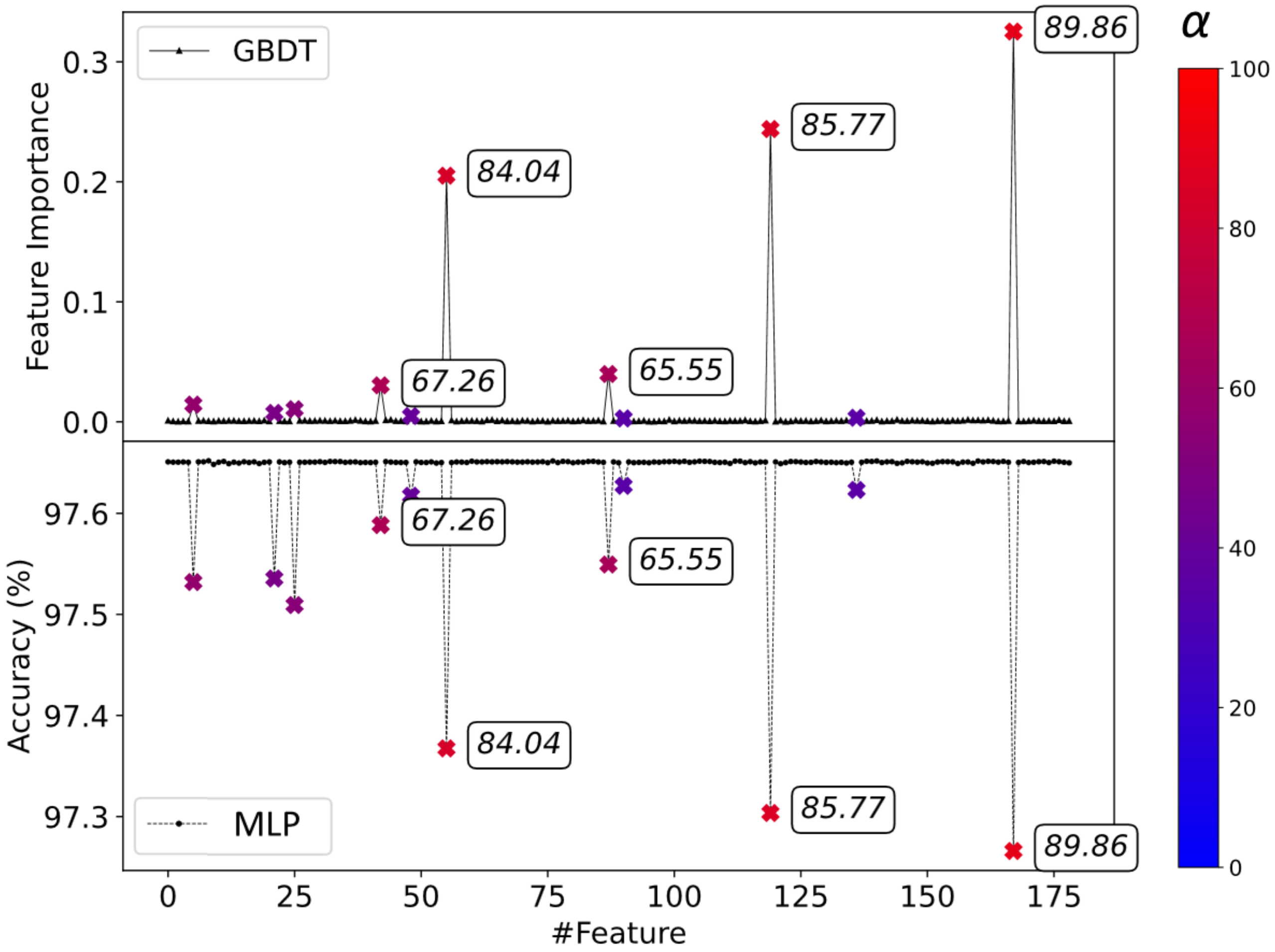

FIG. 3. Comparison between GBDT (top) and MLTSA with NN (bottom) feature analysis methods for the 1DW dataset. Correlated features are marked from blue (0%) to red (100%) depending on the mixing coefficient, α (x symbols, color scale on the right, five highest mixing coefficients also displayed for the datapoints). Uncorrelated features (small black symbols) are at 0 FI for GBDT and show no loss of accuracy for MLTSA with MLPs. Correlated features all show a significant accuracy drop for the MLP, while only the top correlated features have high FI using GBDT.

To identify the selected DW potential and its highest correlated features from the dataset, we calculated the accuracy drop (Fig. S3) using the trained MLP and compared this approach with the FI using GBDT. Results of both feature analysis are found in Fig. 3 for the 1DW dataset and in Fig. S10 for the 5DW potentials dataset.

close to 0. For the MLP, the top three distances are similarly captured as in the FI with the highest accuracy drops. Importantly, all correlated features have a non-zero accuracy drop, showing that they are correctly identified.

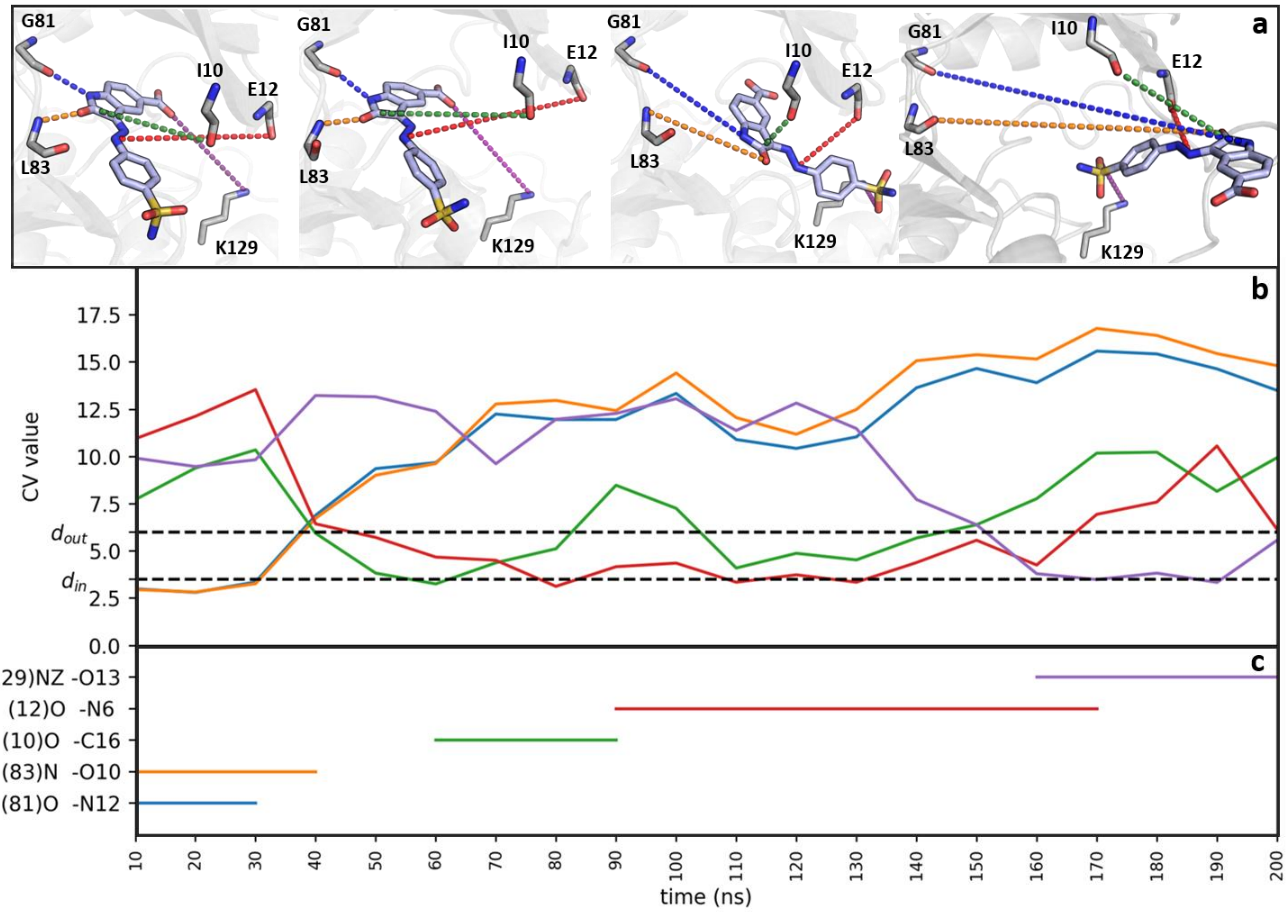

FIG. 3. a Unbinding trajectory of ligand 60K represented as selected snapshots along the trajectory at 0, 60, 90 and 160 ns from left to right. A representative set of distances used for the bias are shown as coloured dashed lines (for the full set of distances used, please refer to Fig. S2.1-S2.9), **b** Representative distance values during the trajectory. The lower dashed line is the cut-off below which an interaction is included in the main CV, the upper cut-off is the value above which the distance is excluded from the CV. **c** Representative distances included in the CV along the unbinding trajectory.

Using the dataset with increased complexity consisting of 5 DW potentials and 15 correlated features (Fig. S10), we observed a similar performance of the two ML methods. GBDT correctly captured and ranked the top three features (#8, #25 and #35). However, most other important features scored a FI value very close to 0. Out of 15 correlated features, GBDT hasn't identified 12 of them with high FI, whereas the MLP captured all of them. However, the MLP accuracy drop did not rank the top four features in the correct order, scoring the 3rd most correlated feature with the biggest accuracy drop.

Considering both analytical models, we found that which GBDT has a higher specificity to rank the top correlated features in the correct order, MLP has a higher sensitivity and captures all correlated features but cannot necessarily identify the highest ranked ones quantitatively using the accuracy drop as the measure. Therefore, a combination of the two ML methods can further help identify the most important features. In more complex systems, this performance might not be directly generalizable, however, due to the simple linear correlation of the CVs of this model.

CDK kinase unbinding free energy calculations

For each system, we performed three independent simulation replicas starting from the respective equilibrated system. For each replica, we performed the initial unbiased MD simulation, followed by our unbinding trajectory determination procedure and subsequently calculating the minimum free energy path and the corresponding free energy profile using the finite temperature string method.

Fig. 3 shows a representative result of the unbinding process for selected interactions. First distances (blue and orange) are identified from the initial unbinding trajectory. Later in the unbinding process at 60 ns a new interaction is found (green line) and at 90, and 160 ns more distances are included in the main CV (red and purple, respectively). Additionally, interactions are progressively being removed as they are breaking. Details of the selected CVs during the unbinding iterations are in the panels of Fig. S2.1-2.9 for every replica.

Overall, while the identified CVs in different replicas vary, a few common key CVs are present in all unbinding trajectories within different replicas. However, even if the actual unbinding pathways have differences for different replicas, as seen by looking at the different distances found along the path (Fig. S2), they are all expected to pass through the same TS ensemble. This can be confirmed from the free energy profiles as well, as there is only one key barrier corresponding to the breaking of the drugs with the His84 H-bonding contact (FIG. 5)[68], suggesting that the different replicas share this same TS ensemble indeed, despite the slightly varying pathways and identified CVs along the path.

Table I. Ligand binding kinetic and thermodynamic values of the three systems from Dunbar et al.[55] and calculated results obtained from our computational simulations. $\Delta G^{\ddagger}_{calc}$ was calculated using the Arrhenius equation: $k=k_BT/h \exp(\Delta G^{\ddagger}/k_BT)$ at 298 K.[79]

PDB	Ligand	K_D (M)	k_{on} $[M^{-1}s^{-1}]$	k_{off} $[s^{-1}]$	$\Delta G^{\ddagger}_{exp}$ (kcal/mol)	$\Delta G^{\ddagger}_{calc}$ (kcal/mol)
3sw4	18K	9.61E-07	1.00E+05	0.0823	18.93(±0.17)	**16.29(±0.21)**
4fku	60K	9.86E-08	1.32E+05	0.0133	20.01(±0.12)	**9.96(±1.5)**
4fkw	62K	4.73E-08	6.49E+04	0.00261	20.97(±0.05)	**20.27(±1.06)**

The energy barrier extracted from the PMF of our simulations qualitatively agrees with the experimental results and are very well reproducible within the same system (Table I and FIG. 4). The shape of the free energy profile is also consistent amongst the replicas, however the exact shape of the free energy present for that replica (Fig. S8 and Table S5).

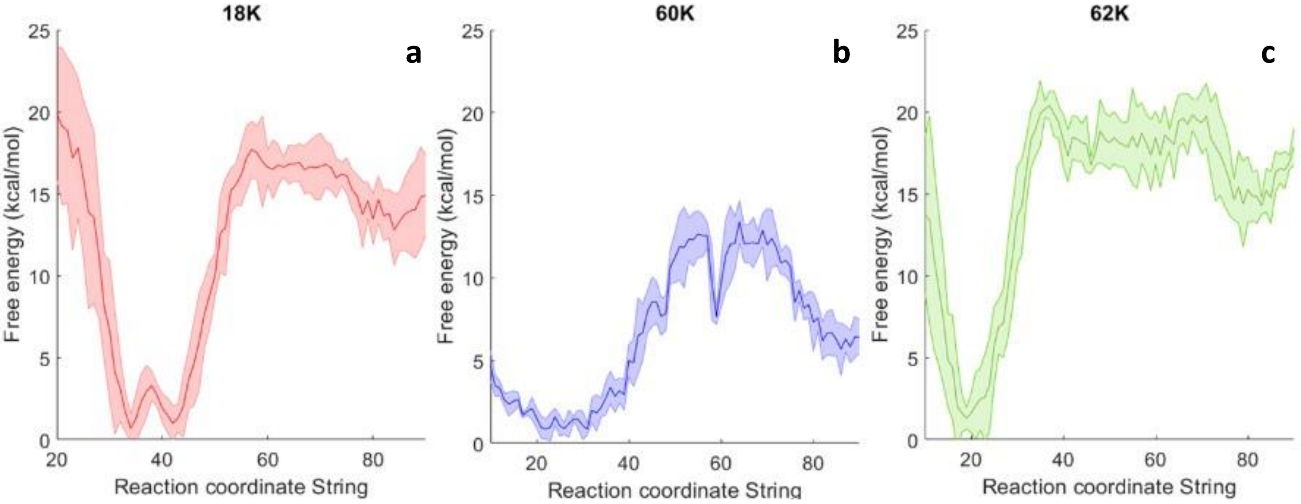

FIG. 4. PMF of the unbinding path for 18K (**a**) 60K (**b**), and 62K (**c**). The free energy profile is obtained from a representative replica, the standard error, shown as shaded area are obtained by dividing the full dataset into 4 subgroups.

Generally, higher number of CVs results in a broader TS peak (e.g., Fig. S8, ligand 62K).

For ligands 18K and 62K, we obtained very similar kinetic data to the experimental one, while ligand 60K shows a larger deviation of ~10 kcal/mol from the experimental value (FIG. 4). Importantly, comparing the same ligand within the three different replicas in all the three system provide very similar free energy barriers, expressed with a low standard error. Our energy barriers are able to reproduce the high energy barriers also seen experimentally thanks to the introduction of numerous key CVs that are not only taken from the initial ligand-bound conformation but instead introduced along the unbinding path (Fig. 3).

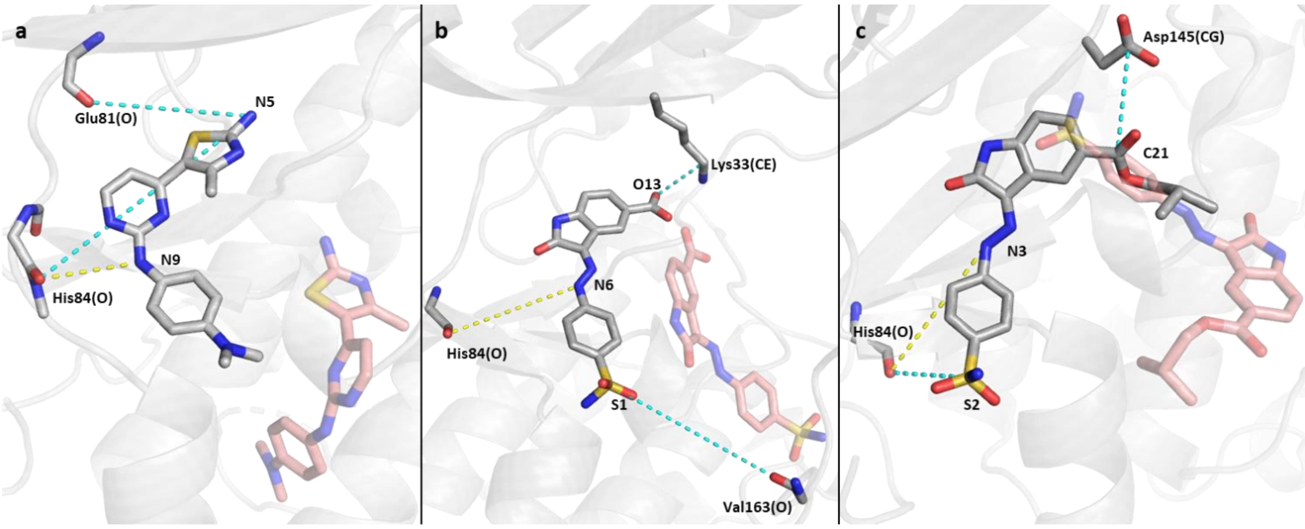

FIG. 5. CVs obtained from the unbinding of 18K (a), 60K (b), and 62K (c); representative distances shown in dashed lines (yellow: interaction from the initial structure, cyan: interaction found during the unbinding

trajectory), red sticks represent the coordinate of the ligand when it is outside the pocket. These distances appear in each of the three replicas for each system.

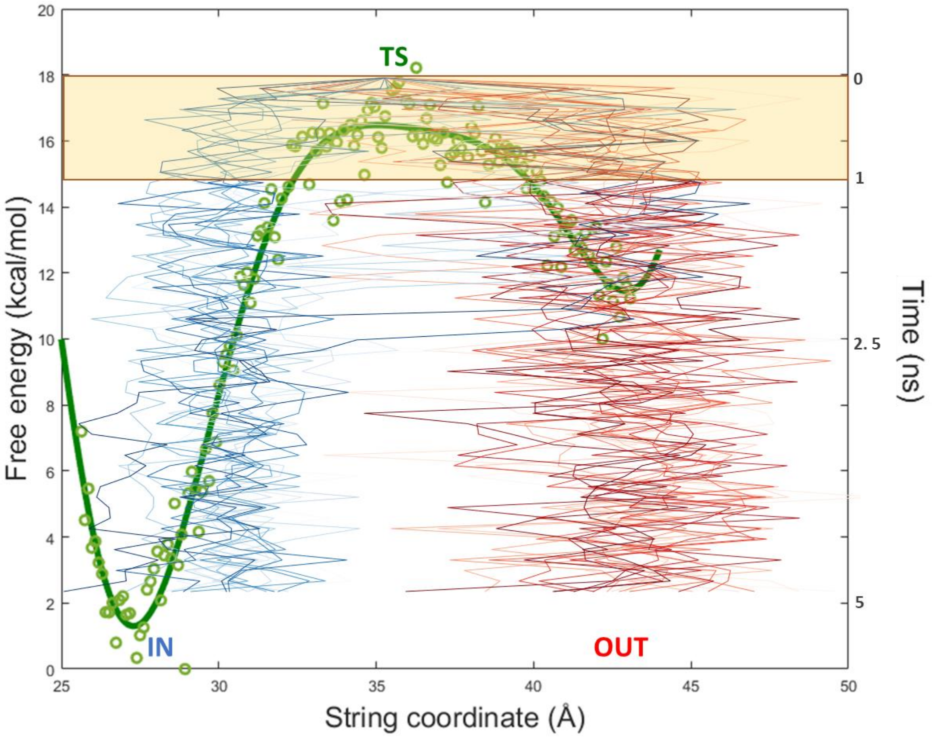

FIG. 6. Representation of the PMF of ligand 62K along the String coordinate and the path of multiple downhill trajectories started at the TS (in green) for further analysis. From the TS coordinate as a starting point, a set of simulations leading to both an IN position (blue) and an OUT position (red) are represented as lines. The green dots illustrate the free energy profile datapoints obtained from the WHAM calculation using the string window as string coordinate, and as a green line, the fitting obtained from the green dots. The yellow shade represents the simulation time portion used for analysis during our machine learning-based approach.

This H-bond was reported as a key interaction in many ligands in complex with CDK2/CDK5.[80,81] These distances were found and included from the initial unbiased simulation in each of the three systems before the unbinding procedure. However, during the unbinding trajectory, once this important H-bond between His84 and the ligand is broken, new interactions are formed, for varying time scales. For 18K, in all the three replicas, H-bonds are formed with the exocyclic amino group of the ligand (N5) and the backbone oxygen of Glu81 and subsequently with the backbone oxygen of His84. 60K and 62K molecules present a sulphonamide terminal group, which, during the trajectory, interacts with Val163 and His84 of CDK2.

To analyze which distances are the most important at the TS region, we implemented our MLTSA method. Starting with three datasets of 172 (60K), 139 (62K) and 148 (18K) independent downhill trajectories for each system, and initial set of CVs of over 170 (Table S1.II), we obtained a shortlist of distances for each system that are major determinants for the prediction of whether a molecule ends up in the bound or unbound states (FIG. 6). By training with trajectory data from up to 0.3 ns of each downhill simulation, the model can predict with high accuracy the IN or OUT outcome of the trajectories, more specifically: 80.11% for 18K, 90.44% for 60K and 93.83% for 62K. The effectiveness of the ML training is confirmed by comparing the

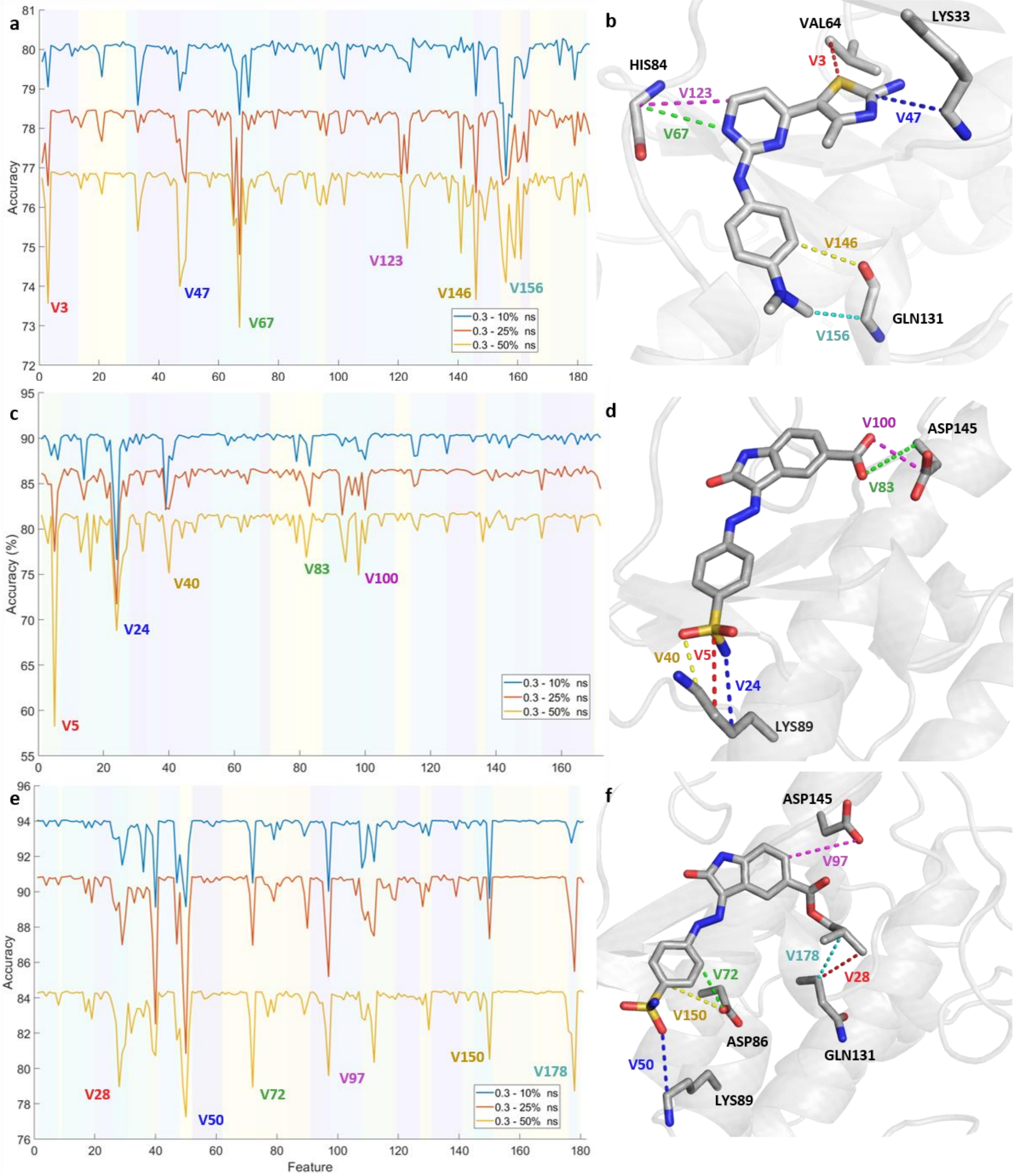

FIG. 7. Identification of the essential distances (Feature Reduction) from the largest accuracy drop using the last 50% (yellow), 25% (red), and 10% (blue) of the frames up to the first 0.3 ns of the simulations for a: 18K, c: 60K, e: 62K. The different shades in the background group the different features according to the atom of the ligand involved. Features presenting significant decrease in accuracy are labelled (see Table S1.II) and portrayed as a 3D representation on the right side of each plot: b:18K, d:60K, and f:62K.

accuracy of predicting the trajectory outcomes using our original final free energy reaction coordinate that was used to classify the trajectories at 5 ns (see Fig. S6, S7 and Tables S4.I-S4.III for additional results with analysis performed at different simulation lengths). Importantly, the ML model is able to predict the outcome more accurately at early times (before ~0.3 ns), than using the best possible prediction via the string reaction coordinate: with above ~80% accuracy versus ~60-70%, respectively for the ML and the standard reaction coordinate (Fig. S9.I-S9.III).

Using the trained model, we then performed a feature reduction analysis to identify which CV features affect the overall prediction ability of the ML model the most. For all three molecules we were able to select the most important structural features (FIG. 7 a, c, e), that lead to the significant reduction of the prediction accuracy, when such feature is eliminated (kept as a constant value fed to the ML, see Fig. S3 for details), while other features do not affect the overall accuracy of the predictions.

We also compared the validity of the feature reduction approach with GBDT to identify FIs. The results obtained show broad similarity with our main MLTSA approach (Fig. S11.I – S11.III) and outperform the baseline approach without ML. This suggests that alternative ML models may also be used successfully and further validate our results.

IV. CONCLUSIONS

Optimizing ligand unbinding kinetics is a very challenging problem for small molecule drug discovery and design, that can lead to the development of drugs with superior efficacy. To tackle this, we have developed a new method, which allows us to calculate the free energy barrier for the ligand unbinding process, therefore providing quantitative information about the residence time of a specific ligand. Our method involves an exploration step, where a ligand unbinding path is determined together with key collective variables that describe this path. Subsequently, we perform accurate free energy calculations using the complete set of identified interactions as CVs along the unbinding path via the finite temperature string method. This provides us with the free energy barriers and an ensemble of structures at the transition state of the ligand unbinding process. The novelty of the method lies in the combination of automated iterative addition and removal of the collective variables determining an unbinding trajectory, which allows us to discover novel interactions not available *a priori*, based on the interactions from the bound structure. We found that while the unbinding trajectories show different paths between different replicas for the same system, our method nevertheless identifies the key interactions important during the unbinding process and provides consistent free energy barriers. The combination of generating an initial path and identifying the important CVs for the unbinding process with the string method for accurate free energy calculations using high dimensional reaction coordinates provide an efficient way to obtain quantitative kinetics of ligand unbinding.

We tested this method using a well-studied cancer drug target, CDK2, using three drug molecules with measured kinetic profiles. We obtained high energy barriers corresponding to experiments using our method, which demonstrates the fundamental importance of determining a well-selected, high-dimensional set of CVs for the correct description of the process and kinetics results.

We explored analytical 180-dimensional systems using one or multiple DW potentials. We performed the ML analysis both with GBDT and MLP methods. Our results demonstrate for simple linear mixing models that they both can capture correctly the most important correlated features. The MLP is a faster approach, it was more sensitive to correlated features, however, sometimes could not rank the top features correspondingly. On the contrary, the GBDT feature importances could skip lowly correlated features in a dataset but can more accurately rank the FIs of the top selected features sacrificing higher computational time. Thus, we suggest that a joint approach of both models might complement each other to identify relevant CVs. Nonetheless, future studies with non-linear correlated time series can further help to explore the performances of these and other ML methods. Importantly, analogous analysis can be performed for various complex processes, including ones with multiple states as possible outcomes.

To aid the kinetics-based design of novel compounds, we also developed a novel method, MLTSA, that allows us to identify the most important features involved at the TS of the unbinding. We generated multiple trajectories initiated at the TS, which either terminated in the bound state or in the unbound state. We then trained a multilayer perceptron ML algorithm to predict the outcome of the trajectories by using a set of CVs and data drawn from the initial segment of the trajectories only. By doing so, we were able to demonstrate that the ML was able to predict the trajectory outcomes with much higher accuracy than using the original set of CVs used for the free energy calculations. A feature importance analysis was further employed to then identify the key CVs and the corresponding structural features that determined the fate of the trajectories, therefore are the most important descriptors of the TS.

In addition to binding rates, we also aimed to identify specific molecular features and interactions with the target protein that allows us to design kinetic properties of the ligand. Using our ML methods, we identified multiple interactions between the protein and specific parts of the ligands that were of major importance for the trajectories to pass the TS. Important protein-ligand interactions at the TS-bound poses for CDK2 correspond to functional groups of the distal ends of the ligands. Besides His84, a known key residue for interaction with multiple CDK2/4 inhibitors, here we also identified additional common interactions within CDK2 across the ligands, for example between Lys89 and the sulfonamide groups or between Asp145 and the carboxylic group and the ester group for 60K and 62K, respectively. Importantly, to perform this analysis, we require the knowledge of the TS structures as well as the MLTSA analysis using a set of trajectories from these initial points. Our algorithms enable us to uncover novel design objectives for a kinetics-based lead optimization process.

SUPPLEMENTARY MATERIALS

See Supplemental Material at [URL will be inserted by publisher] for simulation set-up for MD; atom clustering for the unbinding; unbinding distances; MLTSA flowchart and distances used; training results for MLTSA; all replica free energy profiles; validation for ML analysis and GBDT results.

ACKNOWLEDGEMENT

This project made use of time on ARCHER granted via the UK High-End Computing Consortium for Biomolecular Simulation, HECBioSim (http://hecbiosim.ac.uk), supported by EPSRC (grant no. EP/R013012/1) and ERC (project 757850 BioNet).

REFERENCES

(1) Copeland, R. A. *Evaluation of Enzyme Inhibitors in Drug Discovery: A Guide for Medicinal Chemists and Pharmacologists: Second Edition*; John Wiley and Sons, 2013. https://doi.org/10.1002/9781118540398.

(2) Copeland, R. A. The Drug-Target Residence Time Model: A 10-Year Retrospective. *Nature Reviews Drug Discovery*. Nature Publishing Group February 2016, pp 87–95. https://doi.org/10.1038/nrd.2015.18.

(3) Bernetti, M.; Masetti, M.; Rocchia, W.; Cavalli, A. Kinetics of Drug Binding and Residence Time. *Annu. Rev. Phys. Chem. Annu. Rev. Phys. Chem. 2019* **2019**, *70*, 143–171. https://doi.org/10.1146/annurev-physchem-042018.

(4) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug–Target Residence Time and Its Implications for Lead Optimization. *Nat. Rev. Drug Discov.* **2006**, *5* (9), 730–739. https://doi.org/10.1038/nrd2082.

(5) Lu, H.; Tonge, P. J. Drug-Target Residence Time: Critical Information for Lead Optimization. *Current Opinion in Chemical Biology*. NIH Public Access August 2010, pp 467–474. https://doi.org/10.1016/j.cbpa.2010.06.176.

(6) Bernetti, M.; Cavalli, A.; Mollica, L. Protein-Ligand (Un)Binding Kinetics as a New Paradigm for Drug Discovery at the Crossroad between Experiments and Modelling. *MedChemComm*. Royal Society of Chemistry March 2017, pp 534–550. https://doi.org/10.1039/c6md00581k.

(7) Ruiz-Garcia, A.; Bermejo, M.; Moss, A.; Casabo, V. G. Pharmacokinetics in Drug Discovery. *J. Pharm. Sci.* **2008**, *97* (2), 654–690. https://doi.org/10.1002/jps.21009.

(8) Schuetz, D. A.; de Witte, W. E. A.; Wong, Y. C.; Knasmueller, B.; Richter, L.; Kokh, D. B.; Sadiq, S. K.; Bosma, R.; Nederpelt, I.; Heitman, L. H.; Segala, E.; Amaral, M.; Guo, D.; Andres, D.; Georgi, V.; Stoddart, L. A.; Hill, S.; Cooke, R. M.; De Graaf, C.; Leurs, R.; Frech, M.; Wade, R. C.; de Lange, E. C. M.; IJzerman, A. P.; Müller-Fahrnow, A.; Ecker, G. F. Kinetics for Drug Discovery: An Industry-Driven Effort to Target Drug Residence Time. *Drug Discovery Today*. Elsevier Ltd June 2017, pp 896–911. https://doi.org/10.1016/j.drudis.2017.02.002.

(9) Darling, R. J.; Brault, P. A. Kinetic Exclusion Assay Technology: Characterization of Molecular Interactions. *Assay and Drug Development Technologies*. Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA December 2004, pp 647–657. https://doi.org/10.1089/adt.2004.2.647.

(10) Rose, R. H.; Briddon, S. J.; Hill, S. J. A Novel Fluorescent Histamine H 1 Receptor Antagonist Demonstrates the Advantage of Using Fluorescence Correlation Spectroscopy to Study the Binding of Lipophilic Ligands. *Br. J. Pharmacol.* **2012**, *165* (6), 1789–1800. https://doi.org/10.1111/j.1476-5381.2011.01640.x.

(11) Herrick-Davis, K.; Grinde, E.; Cowan, A.; Mazurkiewicz, J. E. Fluorescence Correlation Spectroscopy Analysis of Serotonin, Adrenergic, Muscarinic, and Dopamine Receptor Dimerization: The Oligomer Number Puzzle. *Mol. Pharmacol.* **2013**, *84* (4), 630–642. https://doi.org/10.1124/mol.113.087072.

(12) De Jong, L. A. A.; Uges, D. R. A.; Franke, J. P.; Bischoff, R. Receptor-Ligand Binding Assays: Technologies and Applications. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*. Elsevier December 2005, pp 1–25. https://doi.org/10.1016/j.jchromb.2005.10.002.

(13) Bruce, N. J.; Ganotra, G. K.; Kokh, D. B.; Sadiq, S. K.; Wade, R. C. New Approaches for Computing Ligand–Receptor Binding Kinetics. *Curr. Opin. Struct. Biol.* **2018**, *49*, 1–10. https://doi.org/10.1016/j.sbi.2017.10.001.

(14) Wolf, S.; Lickert, B.; Bray, S.; Stock, G. Multisecond Ligand Dissociation Dynamics from Atomistic Simulations. *Nat. Commun.* **2020**, *11* (1). https://doi.org/10.1038/s41467-020-16655-1.

(15) Huggins, D. J.; Biggin, P. C.; Dämgen, M. A.; Essex, J. W.; Harris, S. A.; Henchman, R. H.; Khalid, S.; Kuzmanic, A.; Laughton, C. A.; Michel, J.; Mulholland, A. J.; Rosta, E.; Sansom, M. S. P. P.; van der Kamp, M. W. Biomolecular Simulations: From Dynamics and Mechanisms to Computational Assays of Biological Activity. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2019**, *9* (3), e1393. https://doi.org/10.1002/wcms.1393.

(16) Huang, D.; Caflisch, A. The Free Energy Landscape of Small Molecule Unbinding. *PLoS Comput. Biol.* **2011**, *7* (2), e1002002. https://doi.org/10.1371/journal.pcbi.1002002.

(17) Dahl, G.; Akerud, T. Pharmacokinetics and the Drug-Target Residence Time Concept. *Drug Discovery Today*. Elsevier Current Trends August 2013, pp 697–707. https://doi.org/10.1016/j.drudis.2013.02.010.

(18) Lotz, S. D.; Dickson, A. Unbiased Molecular Dynamics of 11 Min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions. *J. Am. Chem. Soc.* **2018**, *140* (2), 618–628. https://doi.org/10.1021/jacs.7b08572.

(19) Nunes-Alves, A.; Kokh, D. B.; Wade, R. C. Recent Progress in Molecular Simulation Methods for Drug Binding Kinetics. *Current Opinion in Structural Biology*. Elsevier Ltd October 2020, pp 126–133. https://doi.org/10.1016/j.sbi.2020.06.022.

(20) Decherchi, S.; Cavalli, A. Thermodynamics and Kinetics of Drug-Target Binding by Molecular Simulation. *Chemical Reviews*. American Chemical Society December 2020, pp 12788–12833. https://doi.org/10.1021/acs.chemrev.0c00534.

(21) Jorgensen, W. L.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *Cit. J. Chem. Phys.* **1985**, *83*, 3050. https://doi.org/10.1063/1.449208.

(22) Jorgensen, W. L.; Thomas, L. L. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput.* **2008**, *4* (6), 869–876. https://doi.org/10.1021/ct800011m.

(23) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (20), 12562–12566. https://doi.org/10.1073/pnas.202427399.

(24) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. Kinetics of Protein-Ligand Unbinding: Predicting Pathways, Rates, and Rate-Limiting Steps. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (5), E386–E391. https://doi.org/10.1073/pnas.1424461112.

(25) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120* (24), 11919–11929. https://doi.org/10.1063/1.1755656.

(26) Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K. Steered Molecular Dynamics; Springer, Berlin, Heidelberg, 1999; pp 39–65. https://doi.org/10.1007/978-3-642-58360-5_2.

(27) Faradjian, A. K.; Elber, R. Computing Time Scales from Reaction Coordinates by Milestoning. *J. Chem. Phys.* **2004**, *120* (23), 10880–10889. https://doi.org/10.1063/1.1738640.

(28) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

(29) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151. https://doi.org/10.1016/S0009-2614(99)01123-9.

(30) Schuetz, D. A.; Bernetti, M.; Bertazzo, M.; Musil, D.; Eggenweiler, H. M.; Recanatini, M.; Masetti, M.; Ecker, G. F.; Cavalli, A. Predicting Residence Time and Drug Unbinding Pathway through Scaled Molecular Dynamics. *J. Chem. Inf. Model.* **2019**, *59* (1), 535–549. https://doi.org/10.1021/acs.jcim.8b00614.

(31) Mollica, L.; Decherchi, S.; Zia, S. R.; Gaspari, R.; Cavalli, A.; Rocchia, W. Kinetics of Protein-Ligand Unbinding via Smoothed Potential Molecular Dynamics Simulations. *Sci. Rep.* **2015**, *5* (1), 11539. https://doi.org/10.1038/srep11539.

(32) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. TRANSITION PATH SAMPLING : Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53* (1), 291–318. https://doi.org/10.1146/annurev.physchem.53.082301.113146.

(33) Kokh, D. B.; Amaral, M.; Bomke, J.; Grädler, U.; Musil, D.; Buchstaller, H. P.; Dreyer, M. K.; Frech, M.; Lowinski, M.; Vallee, F.; Bianciotto, M.; Rak, A.; Wade, R. C. Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2018**, *14* (7), 3859–3869. https://doi.org/10.1021/acs.jctc.8b00230.

(34) Evans, R.; Hovan, L.; Tribello, G. A.; Cossins, B. P.; Estarellas, C.; Gervasio, F. L. Combining Machine Learning and Enhanced Sampling Techniques for Efficient and Accurate Calculation of Absolute Binding Free Energies. *J. Chem. Theory Comput.* **2020**, *16* (7), 4641–4654.

https://doi.org/10.1021/acs.jctc.0c00075.

(35) Lamim Ribeiro, J. M.; Tiwary, P. Toward Achieving Efficient and Accurate Ligand-Protein Unbinding with Deep Learning and Molecular Dynamics through RAVE. *J. Chem. Theory Comput.* **2019**, *15* (1), 708–719. https://doi.org/10.1021/acs.jctc.8b00869.

(36) Hovan, L.; Comitani, F.; Gervasio, F. L. Defining an Optimal Metric for the Path Collective Variables. *J. Chem. Theory Comput.* **2019**, *15* (1), 25–32. https://doi.org/10.1021/acs.jctc.8b00563.

(37) Tribello, G. A.; Ceriotti, M.; Parrinello, M. Using Sketch-Map Coordinates to Analyze and Bias Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (14), 5196–5201. https://doi.org/10.1073/pnas.1201152109.

(38) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of Reaction Coordinates via Locally Scaled Diffusion Map. *J. Chem. Phys.* **2011**, *134* (12), 124116. https://doi.org/10.1063/1.3569857.

(39) Abrams, C.; Bussi, G. Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. *Entropy* **2013**, *16* (1), 163–199. https://doi.org/10.3390/e16010163.

(40) Casasnovas, R.; Limongelli, V.; Tiwary, P.; Carloni, P.; Parrinello, M. Unbinding Kinetics of a P38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. *J. Am. Chem. Soc.* **2017**, *139* (13), 4780–4788. https://doi.org/10.1021/jacs.6b12950.

(41) Haldar, S.; Comitani, F.; Saladino, G.; Woods, C.; Van Der Kamp, M. W.; Mulholland, A. J.; Gervasio, F. L. A Multiscale Simulation Approach to Modeling Drug-Protein Binding Kinetics. *J. Chem. Theory Comput.* **2018**, *14* (11), 6093–6101. https://doi.org/10.1021/acs.jctc.8b00687.

(42) Capelli, A. M.; Costantino, G. Unbinding Pathways of VEGFR2 Inhibitors Revealed by Steered Molecular Dynamics. *J. Chem. Inf. Model.* **2014**, *54* (11), 3124–3136. https://doi.org/10.1021/ci500527j.

(43) Niu, Y.; Li, S.; Pan, D.; Liu, H.; Yao, X. Computational Study on the Unbinding Pathways of B-RAF Inhibitors and Its Implication for the Difference of Residence Time: Insight from Random Acceleration and Steered Molecular Dynamics Simulations. *Phys. Chem. Chem. Phys.* **2016**, *18* (7), 5622–5629. https://doi.org/10.1039/c5cp06257h.

(44) Rydzewski, J.; Valsson, O. Finding Multiple Reaction Pathways of Ligand Unbinding. *J. Chem. Phys.* **2019**, *150* (22), 221101. https://doi.org/10.1063/1.5108638.

(45) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. Constrained Reaction Coordinate Dynamics for the Simulation of Rare Events. *Chem. Phys. Lett.* **1989**, *156* (5), 472–477. https://doi.org/10.1016/S0009-2614(89)87314-2.

(46) Rosta, E.; Nowotny, M.; Yang, W.; Hummer, G. Catalytic Mechanism of RNA Backbone Cleavage by Ribonuclease H from Quantum Mechanics/Molecular Mechanics Simulations. *J. Am. Chem. Soc.* **2011**, *133* (23), 8934–8941. https://doi.org/10.1021/ja200173a.

(47) Wang, H.; Huang, N.; Dangerfield, T.; Johnson, K. A.; Gao, J.; Elber, R. Exploring the Reaction Mechanism of HIV Reverse Transcriptase with a Nucleotide Substrate. *J. Phys. Chem. B* **2020**, *124* (21), 4270–4283. https://doi.org/10.1021/ACS.JPCB.0C02632.

(48) Ovchinnikov, V.; Karplus, M.; Vanden-Eijnden, E. Free Energy of Conformational Transition Paths in Biomolecules: The String Method and Its Application to Myosin VI. *J. Chem. Phys.* **2011**, *134* (8), 85103. https://doi.org/10.1063/1.3544209.

(49) Jung, H.; Covino, R.; Hummer, G. *Artificial Intelligence Assists Discovery of Reaction Coordinates and Mechanisms from Molecular Dynamics Simulations*; arXiv, 2019.

(50) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2019**, *71*, 361–390. https://doi.org/10.1146/annurev-physchem-042018-052331.

(51) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**. https://doi.org/10.1021/ACS.CHEMREV.0C01195.

(52) Burger, H. C.; Schuler, C. J.; Harmeling, S. Image Denoising: Can Plain Neural Networks Compete with BM3D? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; 2012; pp 2392–2399. https://doi.org/10.1109/CVPR.2012.6247952.

(53) Rao, H.; Shi, X.; Rodrigue, A. K.; Feng, J.; Xia, Y.; Elhoseny, M.; Yuan, X.; Gu, L. Feature Selection Based on Artificial Bee Colony and Gradient Boosting Decision Tree. *Appl. Soft Comput.* **2019**, *74*, 634–642. https://doi.org/10.1016/J.ASOC.2018.10.036.

(54) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science (80-.).* **2006**, *313* (5786), 504–507. https://doi.org/10.1126/SCIENCE.1127647.

(55) Dunbar, J. B.; Smith, R. D.; Damm-Ganamet, K. L.; Ahmed, A.; Esposito, E. X.; Delproposto, J.; Chinnaswamy, K.; Kang, Y. N.; Kubish, G.; Gestwicki, J. E.; Stuckey, J. A.; Carlson, H. A. CSAR Data Set Release 2012: Ligands, Affinities, Complexes, and Docking Decoys. *J. Chem. Inf. Model.* **2013**, *53* (8), 1842–1852. https://doi.org/10.1021/ci4000486.

(56) Malumbres, M.; Barbacid, M. Cell Cycle, CDKs and Cancer: A Changing Paradigm. *Nature Reviews Cancer*. Nat Rev Cancer March 2009, pp 153–166. https://doi.org/10.1038/nrc2602.

(57) Otto, T.; Sicinski, P. Cell Cycle Proteins as Promising Targets in Cancer Therapy. *Nature Reviews Cancer*. Nature Publishing Group January 2017, pp 93–115. https://doi.org/10.1038/nrc.2016.138.

(58) Tadesse, S.; Caldon, E. C.; Tilley, W.; Wang, S. Cyclin-Dependent Kinase 2 Inhibitors in Cancer Therapy: An Update. *Journal of Medicinal Chemistry*. American Chemical Society May 2019, pp 4233–4251. https://doi.org/10.1021/acs.jmedchem.8b01469.

(59) Wyatt, P. G.; Woodhead, A. J.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Davis, D. J.; Devine, L. A.; Early, T. R.; Feltell, R. E.; Lewis, E. J.; McMenamin, R. L.; Navarro, E. F.; O'Brien, M. A.; O'Reilly, M.; Reule, M.; Saxty, G.; Seavers, L. C. A.; Smith, D. M.; Squires, M. S.; Trewartha, G.; Walker, M. T.; Woolford, A. J. A. Identification of N-(4-Piperidinyl)-4-(2,6-Dichlorobenzoylamino)-1H-Pyrazole-3-Carboxamide (AT7519), a Novel Cyclin Dependent Kinase Inhibitor Using Fragment-Based X-Ray Crystallography and Structure Based Drug Design. *J. Med. Chem.* **2008**, *51* (16), 4986–4999.

https://doi.org/10.1021/jm800382h.

(60) Jessen, B. A.; Lee, L.; Koudriakova, T.; Haines, M.; Lundgren, K.; Price, S.; Nonomiya, J.; Lewis, C.; Stevens, G. J. Peripheral White Blood Cell Toxicity Induced by Broad Spectrum Cyclin-Dependent Kinase Inhibitors. *J. Appl. Toxicol.* **2007**, *27* (2), 133–142. https://doi.org/10.1002/jat.1177.

(61) Parry, D.; Guzi, T.; Shanahan, F.; Davis, N.; Prabhavalkar, D.; Wiswell, D.; Seghezzi, W.; Paruch, K.; Dwyer, M. P.; Doll, R.; Nomeir, A.; Windsor, W.; Fischmann, T.; Wang, Y.; Oft, M.; Chen, T.; Kirschmeier, P.; Lees, E. M. Dinaciclib (SCH 727965), a Novel and Potent Cyclin-Dependent Kinase Inhibitor. *Mol. Cancer Ther.* **2010**, *9* (8), 2344–2353. https://doi.org/10.1158/1535-7163.MCT-10-0324.

(62) Ayaz, P.; Andres, D.; Kwiatkowski, D. A.; Kolbe, C. C.; Lienau, P.; Siemeister, G.; Lucking, U.; Stegmann, C. M. Conformational Adaption May Explain the Slow Dissociation Kinetics of Roniciclib (BAY 1000394), a Type i CDK Inhibitor with Kinetic Selectivity for CDK2 and CDK9. *ACS Chem. Biol.* **2016**, *11* (6), 1710–1719. https://doi.org/10.1021/acschembio.6b00074.

(63) Caporali, S.; Alvino, E.; Starace, G.; Ciomei, M.; Brasca, M. G.; Levati, L.; Garbin, A.; Castiglia, D.; Covaciu, C.; Bonmassar, E.; D'Atri, S. The Cyclin-Dependent Kinase Inhibitor PHA-848125 Suppresses the in Vitro Growth of Human Melanomas Sensitive or Resistant to Temozolomide, and Shows Synergistic Effects in Combination with This Triazene Compound. *Pharmacol. Res.* **2010**, *61* (5), 437–448. https://doi.org/10.1016/j.phrs.2009.12.009.

(64) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *Journal of Computational Chemistry*. John Wiley and Sons Inc. December 1, 2005, pp 1781–1802. https://doi.org/10.1002/jcc.20289.

(65) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(66) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(67) Lans, I.; Medina, M.; Rosta, E.; Hummer, G.; Garcia-Viloca, M.; Lluch, J. M.; González-Lafont, À. Theoretical Study of the Mechanism of the Hydride Transfer between Ferredoxin-NADP+ Reductase and NADP+: The Role of Tyr303. *J. Am. Chem. Soc.* **2012**, *134* (50), 20544–20553. https://doi.org/10.1021/ja310331v.

(68) E, W.; Ren, W.; Vanden-Eijnden, E. Finite Temperature String Method for the Study of Rare Events†. *J. Phys. Chem. B* **2005**, *109* (14), 6688–6693. https://doi.org/10.1021/jp0455430.

(69) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The Weighted Histogram Analysis Method for Free-energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021. https://doi.org/10.1002/jcc.540130812.

(70) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.;

Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.

(71) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263–274. https://doi.org/10.1021/ci500747n.

(72) Nair, V.; Hinton, G. E. *Rectified Linear Units Improve Restricted Boltzmann Machines*.

(73) Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26* (10), 1340–1347. https://doi.org/10.1093/BIOINFORMATICS/BTQ134.

(74) Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional Variable Importance for Random Forests. *BMC Bioinforma. 2008 91* **2008**, *9* (1), 1–11. https://doi.org/10.1186/1471-2105-9-307.

(75) Krause, J.; Ng, K.; Perer, A. Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models Interpreting and Visualizing Machine Learning Models View Project The T1DI Study View Project Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. **2016**. https://doi.org/10.1145/2858036.2858529.

(76) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2016**, *13-17-August-2016*, 1135–1144.

(77) Hooker, G.; Mentch, L. Please Stop Permuting Features An Explanation and Alternatives.

(78) Verikas, A.; Bacauskiene, M. Feature Selection with Neural Networks. *Pattern Recognit. Lett.* **2002**, *23* (11), 1323–1335. https://doi.org/10.1016/S0167-8655(02)00081-8.

(79) Suardíaz, R.; Jambrina, P. G.; Masgrau, L.; González-Lafont, À.; Rosta, E.; Lluch, J. M. Understanding the Mechanism of the Hydrogen Abstraction from Arachidonic Acid Catalyzed by the Human Enzyme 15-Lipoxygenase-2. A Quantum Mechanics/Molecular Mechanics Free Energy Simulation. *J. Chem. Theory Comput.* **2016**, *12* (4), 2079–2090. https://doi.org/10.1021/acs.jctc.5b01236.

(80) Li, Y.; Zhang, J.; Gao, W.; Zhang, L.; Pan, Y.; Zhang, S.; Wang, Y. Insights on Structural Characteristics and Ligand Binding Mechanisms of CDK2. *International Journal of Molecular Sciences*. MDPI AG April 2015, pp 9314–9340. https://doi.org/10.3390/ijms16059314.

(81) Patel, J. S.; Berteotti, A.; Ronsisvalle, S.; Rocchia, W.; Cavalli, A. Steered Molecular Dynamics Simulations for Studying Protein-Ligand Interaction in Cyclin-Dependent Kinase 5. *J. Chem. Inf. Model.* **2014**, *54* (2), 470–480. https://doi.org/10.1021/ci4003574.