Toward high-throughput predictive modeling of protein binding/unbinding kinetics

See Hong Chiu[1], Lei Xie[1,2,*]

[1]The Graduate Center, The City University of New York, U. S. A.
[2]Department of Computer Science, Hunter College, The City University of New York, U. S. A.

[*]To whom correspondence should be addressed
Email: lei.xie@hunter.cuny.edu

Abstract

One of the unaddressed challenges in drug discovery is that drug potency determined *in vitro* is not a reliable indicator of drug efficacy and toxicity in humans. Accumulated evidences suggest that the *in vivo* activity is more strongly correlated with the binding/unbinding kinetics than the equilibrium thermodynamics of protein-ligand interactions (PLI) in many cases. However, existing experimental and computational techniques are both insufficient in studying the molecular details of kinetics process of PLI. Consequently, we not only have limited mechanistic understanding of the kinetic process but also lack a practical platform for the high-throughput screening and optimization of drug leads based on their kinetic properties. Here we address this unmet need by integrating energetic and conformational dynamic features derived from molecular modeling with multi-task learning. To test our method, HIV-1 protease drug complexes are used as a model system. Our integrated model provides us with new insights into the molecular determinants of the kinetics of PLI. We find that the coherent coupling of conformational dynamics between protein and ligand may play a critical role in determining the kinetic rate constants of PLI. Furthermore, we demonstrated that Normal Mode Analysis (NMA) is an efficient method to capture conformational dynamics of the binding/unbinding kinetics. Coupled with the multi-task learning, we can predict combined k_{on} and k_{off} accurately with an accuracy of 74.35%. Thus, it is possible to screen and optimize compounds based on their kinetic property. Further development of such computational tools will bridge one of the critical missing links between *in vitro* drug screening and *in vivo* drug efficacy and toxicity.

Significance

Drug efficacy and side effect are often dependent on the life-time rather than the binding affinity of drug-target complex. However, existing paradigm in drug discovery mainly focus on screening and optimizing drug leads based on their binding affinity to the receptor. The ignorance of kinetic process of drug binding and unbinding seriously hinders the development of efficient and safe therapeutics. For the first time, we integrate physically-based modeling with multi-task learning to investigate the molecular determinants of protein binding kinetics as well as efficiently and accurately predict the kinetic rate constants of drug-target complex. Such computational tools will allow us not only to elucidate novel mechanisms of protein binding/unbinding process but also to screen and optimize compounds based on their kinetic property. This will bridge one of the critical missing links between *in vitro* drug screening and *in vivo* drug efficacy and toxicity.

Introduction

Drug discovery is a costly and risky process. It often costs more than two billion dollars and takes more than ten years. Only about one third of drugs in phase III clinical trials reach the market. Target-based and cell-based screening are the two major approaches in the early stage of drug discovery. In both of these two technologies, one of the unaddressed fundamental challenges is that drug potency measured *in vitro* may not be a reliable indicator of drug efficacy and toxicity in the human body. In the compound screening and lead optimization, equilibrium thermodynamics constants such as half maximal inhibitory concentration (IC_{50}) or dissociation constant (K_d) have been used as the measures of drug potency for years. As molecules in the human body are in a non-equilibrium condition, the activity of a drug depends, not only on how

strong it interacts with the protein, but also how easy it hits the target and how long it resides in the target. Increasing body of evidence suggests that drug activity *in vivo* is not defined by equilibrium conditions measured *in vitro*, but rather depends on the residence time ($\tau = 1/ k_{off}$) of the receptor-ligand complex *in vivo* in a number of cases [1]. The longer residence time will increase the efficacy of the drug. For example, geldenamycin has low affinity for Heat shock protein (Hsp90) *in vitro* with $IC_{50} \sim 1$ μM, in comparison to its nanomolar effects *in vivo* [1,2]. Copeland et al. analyzed the results of the experiment of mutation-based resistance to inhibitors of HIV-1 protease studied by Maschera et al., and concluded that the essential factor for sustained drug efficacy *in vivo* is the residence time but not the affinity of the drug molecule on its target [3]. Pan et al. reported that residence time is highly correlated with functional efficacy of a series of agonists of the A_{2A} adenosine receptor ($r^2 = 0.95$), but there is little correlation with binding affinity ($r^2 = 0.15$) [4]. Furthermore, Dahl and Akerud disclosed that long residence time has predictability value only when k_{off} is slower than the pharmacokinetics elimination, which is defined as the elimination rate of the drug from the target vicinity [5]. On the other hand, the on-target side effect could be reduced by reducing the drug residence time. Thus, a drug with optimal efficacy and toxicity profile should have a balanced k_{on} and k_{off}. Since IC_{50} and K_d depend on the measurement of the combined effect of k_{on} and k_{off}, they are actually insufficient to explain the impact of binding/unbinding kinetic on drug action, as the same value of K_d can come from infinite number of combinations of k_{on} and k_{off}. Additionally, since K_d is dependent on the free energy difference between the bound and unbound states but is independent on the transition state of protein-ligand interaction (PLI), it is inadequate to elucidate the binding/unbinding kinetics of PLI [4,6].

Experimental techniques for the study of PLI kinetics such as surface plasmon resonance, fluorescence polarization, isothermal titration calorimetry, and mass spectrometry are not only expensive and time-consuming but also insufficient of providing detailed molecular characterization of the PLI kinetics process [7-9]. Computational modeling plays an increasing role in elucidating the binding/unbinding process of PLI. Molecular dynamics (MD) simulations have been reported to be capable to capture the binding process, from beginning to end, in full atomic detail. Unfortunately, the power of MD simulations is limited due to the fact that protein-ligand binding event takes place in a time scale ranging from microseconds up to hours and days. For the majority of the binding processes, they are infeasible for MD simulations. For this reason, metadynamics and other conformational sampling techniques have been developed not only to improve sampling in MD simulations of a system where ergodicity is hindered by the form of the system's energy landscape, but also adopted as a powerful technique for reconstructing the free-energy surface as a function of few selected degrees of freedom.

Buch et al. presented a kinetic model for the binding process of serine protease β-trypsin and inhibitor benzamidine obtained from MD simulations of free ligand binding. In addition to the kinetic pathway of the binding process, the binding free energy and the kinetic constants (k_{on} and k_{off}) of the process were also reported. This study reveals that benzamidine moves between two metastable intermediate states: S2 and S3; and reaches the bound state through S3 [10]. Gervasio et al. applied a metadynamics method successfully to the docking of ligands on flexible receptors in water solution. The method is able not only to find the docked geometry and to predict the binding affinity ($\Delta G_{binding}$) but also to explore the entire docking process from the solution to the docking cavity, including barriers and intermediate minima [11]. Even though these progresses are remarkable, metadynamics is not yet feasible to study the whole

binding/unbinding process of PLI on a large scale. In addition, the choice of collative variables in the metadynamics simulation is not a trivial task.

With the increasing availability of protein binding kinetics data [12,13], data-driven modeling provides an alternative and efficient solution to studying the PLI kinetics. Several predictive models for kinetic constants of protein-protein interaction (PPI) have been developed [14,15]. However, the molecular attributes in these models only covered static structural characteristics such as the percentage of residues in α-helix, the buried surface area of protein, the proportion of charged residues and the proportion of polar atoms at the interface, and the energetic features such as hydrogen bonding potential and the interfacial electrostatic interaction energy between interfacial residues. These features may not sufficiently capture conformational dynamics of the PLI kinetic processes. In addition, existing methods predict k_{on} and k_{off} independently. As a matter of fact, they are dependent in nature. To our knowledge, few methods are available for the large-scale modeling of the binding/unbinding kinetics of PLI with explicit dynamic features, as well as predicting k_{on} and k_{off} simultaneously.

To tackle the above problems, we integrate energetic and conformational dynamic features derived from efficient molecular modeling with state-of-the-art multi-task learning (MTL) approach. In this study, ligand-bound HIV-1 proteases are used as an example to build models. In addition to Electrostatic Energy (EE) and van der Waals Energy (VDWE), which are derived from all-atom Molecular Dynamics simulation [16,17] and environmental-dependent electrostatic potential energy [18], Relative Movement of Ligand-Residue (RMLR) and Relative Movement of Residue-Residue (RMRR) that represent the dynamics impact of ligand binding on the amino acid residues are derived from Normal Mode Analysis (NMA) analysis and used to train machine learning models.

Our multi-facet statistical analysis consistently shows that conformational dynamic features, such as RMLR, are as important as energetic features, particularly EE, in predicting k_{on}, k_{off}. Based on these findings, we propose that coherent conformational dynamic coupling between protein and ligand may play a critical role in determining the kinetic rate constants of PLI. Furthermore, we demonstrated that NMA is an efficient method to capture conformational dynamic features of the binding/unbinding kinetics of PLI. Coupled with the state-of-the-art multi-target classification as well as multi-target regression, it is possible for us to screen and optimize compounds based on the binding/unbinding kinetics of PLI in a high-throughput fashion. The further development of such computational tools will bridge one of the critical missing links between *in vitro* drug potency and *in vivo* drug efficacy and safety, thereby accelerating drug discovery process.

Results

1. Characteristics of data set

In this study, we focused on using HIV-1 protease complex structure to investigate the conformational dynamics and develop predictive model of ligand binding/unbinding kinetics. The HIV-1 protease is an excellent model system for our purpose. First, thirty-nine HIV-1 protease inhibitors have experimentally determined k_{on} and k_{off} under the same condition [19,20]. They provide reasonable number of high quality data points for the data-driven modeling. Second, abundant data of HIV-1 protease inhibitor resistance mutation (PIRM) are available. They can be used to validate the predictive model. Third, both unbound and complex structures of HIV-1 protease are released in Protein Data Bank [21]. The apo- and holo-conformations are the basis for our analysis.

When mapping the 39 HIV-1 protease inhibitors on the 2-dimensional space of k_{on} and k_{off}, as shown in Figure 1, all FDA-approved drugs were clustered in the upper-left corner with high k_{on} and low k_{off}. Based on the criteria of $\log_{10}k_{off} = -2$ and $\log_{10}k_{on} = 5.6$, which will put all FDA approved drugs in a single class and evenly distribute the inhibitors into four different classes, with the labels (0,0), (0,1), (1,0), and (1,1) (see Supplementary Table S1). It is noted that several inhibitors such as A037 have the similar value of K_d, which is equal to k_{off}/k_{on}, to that of the approved drugs, but fall into different classes from the FDA-approved drugs in the 2D map. It suggests that atomic interactive constant K_d alone is not sufficient to determine the drug effect.

Ten inhibitors have solved HIV-1 complex structures in PDB. For the remaining inhibitors whose complex structures have not been experimentally determined, protein-ligand docking software eHiTS [22] is applied to predict its binding pose. The receptor is chosen from one of the ligand-bound HIV-1 complexes with the co-crystallized ligand structure similar to the docked ligand structure. Whenever possible, the common fragment of the co-crystallized and the docked ligand is used as a constraint to select the final binding pose of the docked ligand, such that the RMSD of superimposed common fragments is minimal. An example is shown in Supplementary Figure S1.

Binding site amino acid residues that are involved in the HIV-1 protease inhibitor interactions are determined using the change of solvent assessable surface area (SASA) upon ligand binding. As depicted on Figure 2, there are total 44 amino acids on both chains of the HIV-1 dimmer.

2. Characterization of protein-ligand interaction using the directionality of normal modes

Normal Mode Analysis (NMA) is a powerful computational method to identify and characterize the slowest molecular deformational motions with large amplitude, which are widely involved in biological functions of macromolecules, but inaccessible by other methods. Ligand binding and unbinding events are often on a long-time scale ranging from milliseconds to days, far beyond the current capability of MD simulations. Coarse-grained NMA may allow us to extract important dynamic information on protein-ligand binding/unbinding processes. Since the presence of solvent damping dramatically slows down the large-amplitude motions of bio-molecules, the timescales of molecular motions in reality are much longer than what can be estimated from the eigenvalues of NMA that are calculated in vacuum. In other words, solvent damping causes a discrepancy on a timescale between NMA and real molecular motions. However, the study conducted by Ma revealed that the presence of solvent has a minor impact on eigenvectors, which are determined by the potential surface only [23]. Thus, the information provided by the eigenvectors for the directionality of conformational transitions could be used to study dynamic processes in the time-scale of real situations.

In this study, NMA was conducted using iMod [24]. The directionality of normal modes of the residues in the binding site is used to characterize the conformational dynamic features of binding and unbinding event. Specifically, two data sets including Relative Movement of Ligand-Residue (DS-RMLR), and Relative Movement of Residue-Residue (DS-RMRR) were derived from NMA analysis. Both DS-RMLR and DS-RMRR cover the 10 lowest frequency modes, where DS-RMLR illustrates the relative directionality of normal modes between ligand and residue, and DS-RMRR illustrates the change of directionality of normal modes of binding site residues upon the ligand binding. As an example, Figure 3A depicts the superposition of the 44 residue eigenvectors of the aligned apo structure and the DMP bound structure of 1st normal

mode. It illustrates the shift of the eigenvectors of the 44 residues of HIV-1 protease upon ligand binding. Figure 3B illustrates the relative displacements of the 44 ligand-residue pairs in the DMP bound HIV-1 complex.

3. Characterization of ligand-residue interaction energy

Residue decomposed Pairwise Interaction Energy (PIE) and its two constituting components including Electrostatic Energy (EE) and van der Waals Energy (VDWE), between the ligand and the binding site residue of HIV-1 protease, are calculated from all-atom Molecular Dynamics (MD) simulation and environmental-dependent electrostatic potential energy. The values of PIE, EE, and VDWE, which characterize various energetic aspects of ligand-residue interaction, are used to build three data sets: DS-PIE, DS-EE, and DS-VDWE.

4. Structural determinants of protein-ligand binding/unbinding

We use the energetic and conformational dynamic attributes derived from MD simulation and NMA to train a multi-target machine learning (MTML) model for the classification prediction of kinetic rate constants. In total, there are five principal training data sets including DS-PIE, DS-EE, DS-VDWE, DS-RMLR, and DS-RMRR. Each of them comprises thirty-nine cases with each case comprising 44 attributes.

MTML is defined as follows: Given a set of learning examples D of the form (X,Y), where $X = (x_1, x_2,\ldots, x_k)$ is a vector of k training attributes and $Y = (y_1, y_2,\ldots, y_t)$ is a vector of *t* target attributes, learn a model that, given a new unlabeled example X, can predict the values of all target attributes Y simultaneously. When y_i is categorical, the problem is known as classification. In this study, the yi is binarized value of k_{on} and k_{off} as shown in Figure 1.

Random Forest Predictive Clustering (RF-Clus) is applied for the task of MTML. RF-Clus outperforms other MTML algorithms in the benchmark studies [25]. In addition, it can handle high-dimensional features, e.g. in the situation where the number of attributes is much higher than the number of cases, and can select the importance of attributes (amino acid residues) that contribute to the accuracy of k_{on}/k_{off} prediction. The model was run on the iteration numbers of 100, 200, 250, and 500 in the leave-one-out cross-validation experiment.

Table 1 shows the selected features in the descending order of score of importance. Consequently, sixteen, fifteen, thirteen, and fourteen features were selected from DS-RMLR, DS-RMRR, and DS-EE, and DS-PIE. These identified key residues consist of three motifs: an N-terminal motif (R8, L10), a charged motif (L23, D25, G27, A28, D29, D30, and V32), and a motif corresponding to flap region (residue 43-58), as shown in Figure 4. Both the N-terminal motif and the charged motif are common to DS-PIE, DS-RMLR, DS-RMRR and DS-EE. The flap region is identified by DS-RMRL and DS-RMRR.

All-atom MD simulations have shown that the conformational dynamics of flap region (residue 43-58) plays a key role in the ligand binding process of HIV-1 protease [26,27,28]. Consistent with this observation, the residues in the flap region are identified as key kinetic features with signification displacement upon ligand binding. The recapitulation of the findings from the MD simulation provides a validation to the data-driven approach in the paper. The charged motif in the active site participates in substrate peptide recognition. Specifically, D25 and D29 form hydrogen bonds with substrate peptide. Additionally, R8 and D30 can interact with polar side chains or distal main chain groups in longer substrate peptides. Moreover, the mutation of L10, L23, and V32 lead to drug resistance of HIV protease inhibitors.

5. Combined electrostatic and conformational dynamic features can predict k_{on}/k_{off} accurately

We examine the impacts of the energetic and conformational dynamic features on the prediction of k_{on}/k_{off} accuracy by building different MTML models in three stages. First, we use energetic features (data sets: DS-EE, DS-PIE, DS-VDWE) to build the MTML model. Second, in order to evaluate if the normal mode directionality features can be used to predict the ligand binding and unbinding process, we apply DS-RMRR, DS-RMLR and DS-RMLR+DS-RMRR comprising 88 training attributes in the feature vectors to build the MTML model. Third, we integrate the properties of conformational dynamics and energetics by adding the RMLR features to DS-EE (data set DS-EE+DS-RMLR) to build the MTML model.

For the models trained by DS-EE, DS-PIE, DS-VDWE, DS-RMLR, DS-RMRR, DS-RMLR+DS-RMRR, and DS-EE+DS-RMLR, the highest prediction accuracy of $\log_{10}k_{on}$ are 71.79, 69.23, 43.59, 69.23, 51.28, 69.23, and 76.92% respectively (Figure 5A), the highest prediction accuracy of $\log_{10}k_{off}$ are 76.92, 66.67, 56.41, 71.79, 64.10, 71.79, and 71.79% respectively (Figure 5B), and the highest prediction accuracy of the combined four-class $\log_{10}k_{on}/\log_{10}k_{off}$ are 71.79, 66.66, 47,43, 69.23, 57.69, 70.51, and 74.35% respectively (Figure 5C).

Among the three models trained by the energetic features, the prediction accuracy of the combined four-class $\log_{10}k_{on}/\log_{10}k_{off}$ given by the DS-EE and DS-PIE models are significantly higher than a random guess (50%) by 21.79 and 16.66% respectively, but the accuracy given by the DS-VDWE model is lower than random by 2.57%. These results suggest that in the case of HIV-1 protease, electrostatic interaction plays a key role in the binding/unbinding process, and

the Electrostatic Energy features are more accurate in predicting k_{on} and k_{off} than the features of van der Waals Energy and Pairwise Interaction Energy.

For all the three models trained by the normal mode directionality features, the prediction accuracy of the combined four-class $\log_{10}k_{on}/\log_{10}k_{off}$ is higher than random. Although the accuracy given by the DS-RMRR model is only slightly higher than random by 7.69%, the accuracy given by the DS-RMLR and DS-RMLR+DS-RMRR models are significantly higher than random by 19.23, and 10.51% respectively. These results suggest that the normal mode directionality can capture the information on the ligand binding and unbinding process.

Comparing with the prediction accuracy of the combined four-class $\log_{10}k_{on}/\log_{10}k_{off}$ given by the DS-EE and DS-EE+DS-RMLR models shows that integrating the conformational dynamic features into the energetic features increases the accuracy from 71.79 to 74.35 by 2.56%. Consequently, it implies that the electrostatic interaction and conformational dynamics are jointly responsible for the binding kinetics of HIV protease.

Discussions

1. Coherent receptor-ligand movement is one of the structural determinants of protein binding/unbinding kinetics

Consistent with the all-atom MD simulation, the MTML model trained with the relative directionality of normal mode between residue and ligand recapitulates the role of flap region in the binding kinetics of HIV protease. It is known that electrostatic interaction between a charged drug and a charged receptor impacts the kinetic rate constants [29,30]. Specifically, k_{on} is sensitive to long-range electrostatic interaction, and k_{off} tend to be influenced more by short-range interactions such as hydrogen bonds, salt bridges and van der Waals contacts [31]. The

majority of the residues selected in this study are hydrophobic; the exceptions are the catalytic D25 and D29, which are able to form hydrogen bonds with the main chain groups of substrate peptides, and R8, D30 and K45 which can interact with polar side chains or distal main groups in longer substrate peptides. The MTML model ranks these charged residues more important than the flap region in their contribution to the prediction accuracy. In addition, the MTML model can achieve high prediction accuracy using the electrostatic energy alone. These indicate that the electrostatic interaction is one of the major factors in determining the binding/unbinding kinetics of HIV protease.

Interestingly, in addition to the electrostatic interactions, the directionality of ligand binding site residue movement also has strong correlations with the kinetic constants. Not only the similar residues are selected from DS-RMLR to those from DS-EE, the best performed MTML model is obtained from the combination of DS-RMLR and DS-EE. Based on this observation, we propose that the coherent movement between the ligand and the receptor may play a critical role in determining the ligand binding and unbinding kinetics. As shown in Figure 6, even two protein-ligand complexes have the same non-covalent interactions with the same intensity, they may have different kinetic constants due to the different relative movements between the ligand atom and the receptor atom. It is not surprising, as the non-covalent interactions, especially hydrogen-bonding, depends on the relative directionality of atomic pairs. The relative movement may change the directionality of the interaction, thus weaken (even break) or strengthen the interaction. Thus, the coherent conformational dynamics coupling could be one of key structural determinants of protein binding/unbinding kinetics. This has not been observed before.

2. High-throughput predictive modeling of ligand binding and unbinding kinetics

In spite of recognized importance of protein-ligand binding and unbinding kinetics in the drug discovery, few efficient computational tools are available to screen and optimize chemical compounds based on the binding and unbinding kinetics. With the increasing availability of experimentally determined k_{on}/k_{off} data [14,15], data-driven approach is an appropriate choice for the development of a high-throughput predictive model of ligand binding and unbinding kinetics [32]. However, two questions remain to be answered in developing an effective and efficient machine learning model. First, what are the molecular determinants of ligand binding and unbinding kinetics so that they can be used as features to train a high-quality machine learning model with the minimum impact of over-fitting, and false correlation? Second, what are the suitable machine learning algorithms that can handle high-dimensional data and predict k_{on}/k_{off} simultaneously? For the first time, we have shown that NMA could be an efficient tool to capture the conformationally dynamic information of the ligand binding and unbinding kinetics. The features derived from the NMA could be used to enhance the performance of the machine learning model. Moreover, recently developed multi-target classification algorithms such as RF-Clus could be adopted to train a machine learning model that can predict dependent k_{on}/k_{off} simultaneously.

Although this proof-of-concept study demonstrates the potential of integrating physically-based modeling with multi-target machine learning in understanding the molecular determinants, and developing high-throughput predictive model of ligand binding and unbinding kinetics, there is plenty of space to improve the methodology. Since solvation effect causes a discrepancy on a timescale between real molecular motion and NMA that are calculated in

vacuum, it is expected that NMA coupled with an implicit or explicit solvation model may provide more information on the conformational dynamics of ligand binding process. As water plays a critical role in the ligand binding, the explicit incorporation of the water molecule in the binding site may improve the accuracy of simulation. The global and local geometry of binding pocket could be another important feature [33,34]. In the current study, the ligand is treated as a single rigid body. As a matter of fact, the flexibility of the ligands may have impacts on the kinetic rate constants. As shown in supplemental information Figure S2, both of the values of $\log_{10}k_{on}$ and $\log_{10}k_{off}$ are weakly correlated with the ligand flexibility that is characterized by the number of rotatable bonds. The general trend is that the k_{on} and k_{off} decrease as the number of the ligand rotatable bonds increases. It suggests that the performance of MTML model could be further improved by incorporating the ligand properties. We group the k_{on}/k_{off} into four classes and use the classification model to predict the class and to select features. In practice, it could be more useful to predict the real value of k_{on} and k_{off} simultaneously. It requires a multi-target regression model, which is an active area of research in machine learning.

There are three different models of conformational ensemble of protein-ligand complex. They are the model of induced fit mechanism which is adopted by HIV-1 protein-ligand complex [1,35], the model of selected fit mechanism [36,37], and the model of three step mechanism [38,39]. The mechanism of model determines the on-rate and off-rate equations. For example, the induced fit on-rate is limited by the diffusional rate of encounter complex formation of the proteins in their unbound conformational ensemble, but the off-rate is dependent on the equilibrium between the ground state complex and the excited state complex [35]. Since all the training data sets in this study only cover the characteristics of the ground state HIV-1 complex, the ignorance of the characteristics of the excited state HIV-1 complex could induce deficiency

in the predictive model. In summary, the further development of predictive modeling tools of ligand binding and unbinding kinetics will bridge one of the critical missing links between in vitro drug potency and in vivo drug efficacy and safety on a large scale, thereby accelerating drug discovery process.

Materials and Methods

Figure 7 depicts the workflow of computational procedure in this study, which includes four phases: Phase 1 concerns the structure construction of 3D ligand-bound HIV-1 protease complex. Phase 2 addresses the identification of ligand binding site residues. Phase 3 targets the construction of the five principal data sets. Phase 4 is machine learning computation.

In brief, chemical structures of HIV protease inhibitors were converted into 3D conformation from their 2D structure. Then, the ligand was docked in the HIV protease if no co-crystallized structures exist. Normal Mode Analysis (NMA) was performed for both apo- and holo-structure for each inhibitor. Relative Movement of Ligand-Residue (RMLR) and Relative Movement of Residue-Residue (RMRR) that represent the conformational dynamics impact of ligand binding on the binding site residues were derived from NMA analysis. In addition, Pairwise Interaction Energy as well as its two components, van der Waals Energy and Electrostatic Energy between the ligand and amino acid residues, were derived from the 20 ns all-atom Molecular Dynamics simulation and environmental-dependent electrostatic potential energy. Finally, conformational dynamics and thermodynamics features, individually or combined, are used to train multi-target machine learning models.

Table 1. Key residues identified from four data sets: DS-RMLR, DS-RMRR, DS-EE, and DS-PIE. The feature selection criterion is the frequency of attribute occurrence ≥ 25%. Most of the residues are in chain A, and only three residues (with * superscript) are in chain B. Residues whose mutation lead to drug resistance are underlined.

Residue	DS-RMLR		DS-RMRR		DS-EE		DS-PIE	
	Frequency	Score	Frequency	Score	Frequency	Score	Frequency	Score
R8	35.86	0.78	56.46	0.73	64.79	0.74	52.32	0.73
L10	39.13	0.75	61.16	0.71	34.72	0.77	43.04	0.75
L23	45.60	0.71	54.18	0.71	40.69	0.72	35.52	0.76
D25	44.21	0.69	45.02	0.73	37.41	0.75	49.28	0.70
G27	42.85	0.68	39.96	0.73	30.62	0.75	35.40	0.72
A28	34.70	0.68	39.67	0.70	35.23	0.71	27.48	0.75
D29	28.18	0.71	41.37	0.67	28.36	0.73	48.60	0.67
D30	30.95	0.70	39.79	0.67	25.90	0.72	28.24	0.71
V32	29.35	0.66	37.83	0.66	25.15	0.72	29.84	0.70
K45	37.77	0.65	35.79	0.68			27.98	0.70
I47	30.44	0.68	35.19	0.66	25.54	0.68		
G48	26.88	0.67	26.10	0.66				
G49	34.44	0.63	31.57	0.63	33.23	0.66		
I50			36.64	0.61				
A52	26.08	0.66	27.07	0.62			26.73	0.64
F53	27.90	0.66						
L76							26.63	0.62
P81					25.54	0.63	25.78	0.61
R8*	30.55	0.61						
D25*							31.25	0.60
D29*					28.13	0.65		

Figure legends

Figure 1. Discretization of k_{on} and k_{off} of HIV protease inhibitors. Results of the discretization based on the criteria set at $\log_{10}k_{off}$ = -2 (x-axis) and $\log_{10}k_{on}$ = 5.6 (y-axis). Thirty-nine training records were discretized into four binary classes: (0,0), (0,1), (1,0), and (1,1).

Figure 2. Forty-four HIV-1 residues selected by SASA program and the 26 drug resistant mution residues. The chains A and B of HIV-1 are in transparent gray and green ribbons, respectively, with their flap regions (residue id: 43 – 58) in pink cartoon and active site (residue id: 25 – 29) in blue cartoon. The 22 SASA residues on the chain A are represented by 5 red beads for the charged residues and 17 green beads for the neutral residues. The 26 drug resistant mutation residues are depicted in lines on the chain B including the 12 PIRM residues near the binding site in red and the 16 PIRM residues outside the binding site in green.

Figure 3. Directionality of normal mode. (A) Superposition of the 44 residue eigenvectors of the aligned apo HIV-1 structure (red) (PDB code:3IXO) and the DMP bound HIV-1 structure (blue) (PDB code: 1QBS) of 1st normal mode. (B) Eigenvector displacements of the 44 DMP (red) – residue (green/blue) pairs in the DMP bound HIV-1 complex (1st normal mode). Green/blue arrows are the eigenvectors of the 22 residues of chain A/B respectively.

Figure 4. Twenty-one key residues. Chains A and B of HIV-1 protease are in transparent grey and transparent yellow ribbons, respectively. The three residues located on the chain B are labeled with * superscript. Charged residues including L23, D25, G27, A28, D29, D30 and V32

are in green. Residues located on the N-terminal including R8 and L10 are in red. Residues located in the flap region including K45, I47, G48, G49, I50, A52, and F53 are in blue; L76 and P81 located near the flap region are in pink.

Figure 5. Prediction accuracy of (A) $\log_{10}k_{on}$, (B) $\log_{10}k_{off}$, and (C) the combined four-class $\log_{10}k_{on}/\log_{10}k_{off}$. The number in parentheses is the iteration number used in the experiment.

Figure 6. (A) Coherent conformational coupling. The relative movement between ligand atom and receptor atom will not change the distance and directionality of the interaction, thus the intensity of interaction will not be changed. (B) Incoherent conformational coupling. The relative movement between ligand atom and receptor atom will alter the distance or directionality of the interaction. As a result, the interaction could be weaken or broken.

Figure 7. Schema of methodology.

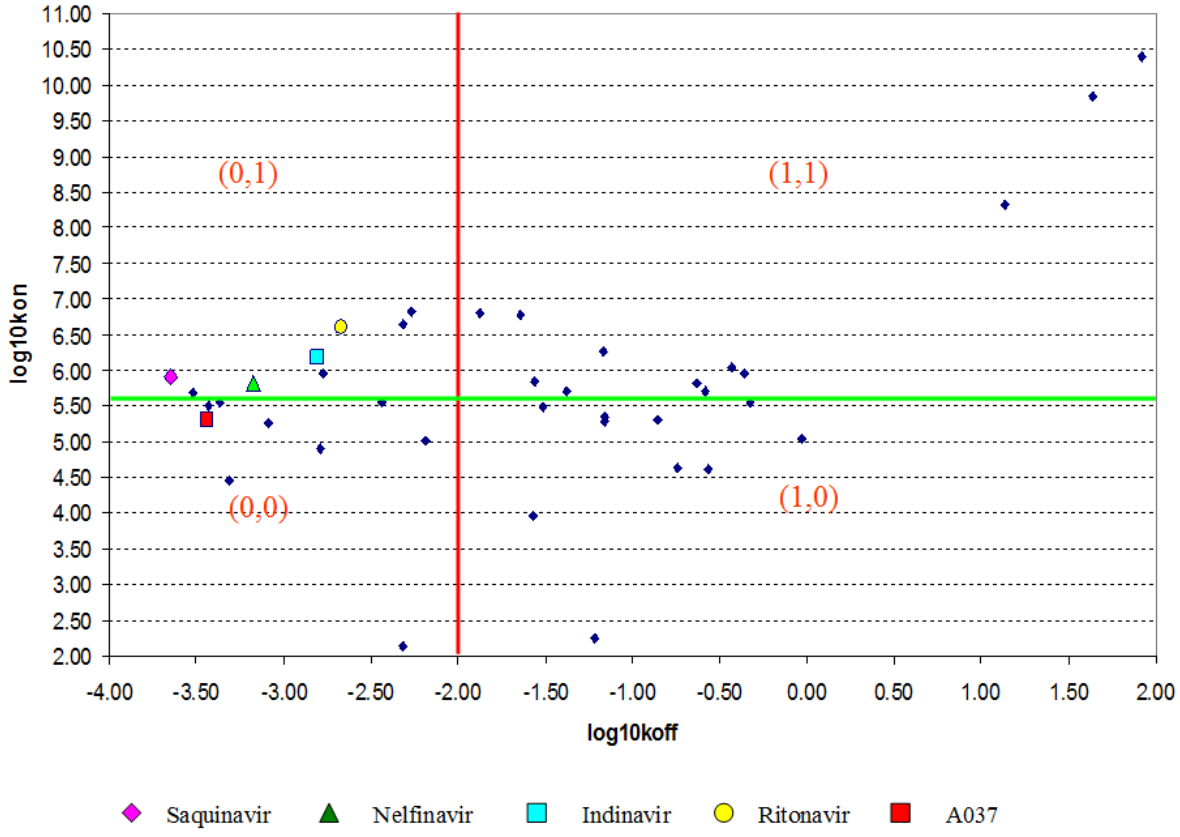

Figure 1

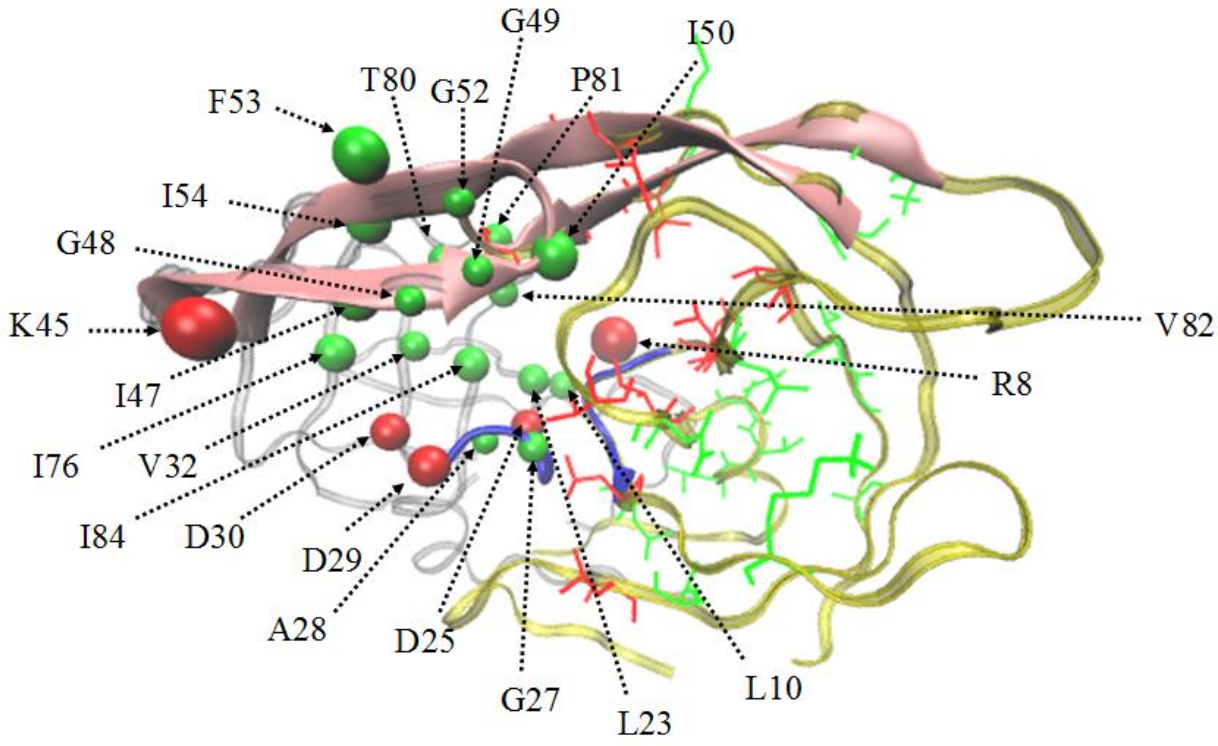

Figure 2

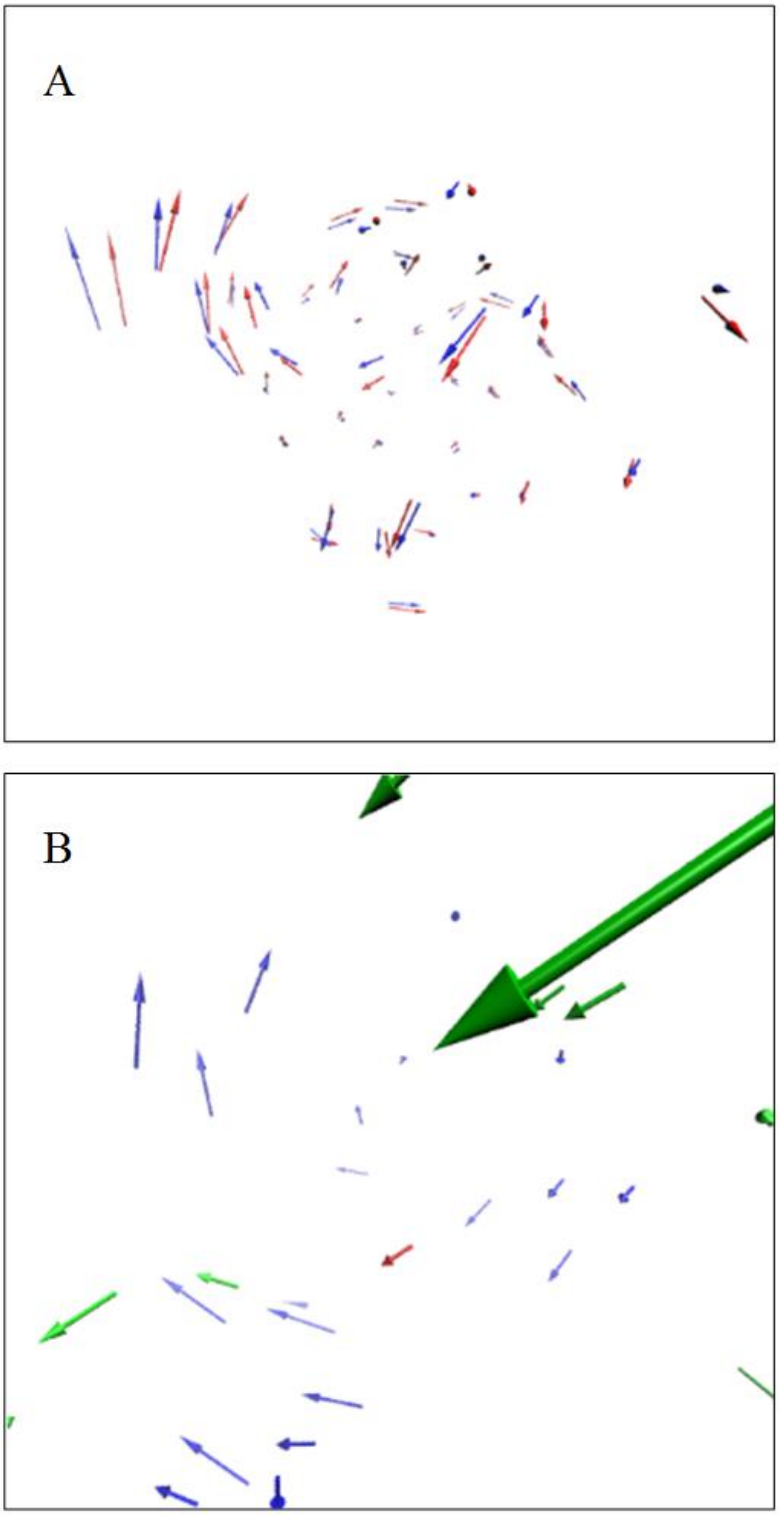

Figure 3

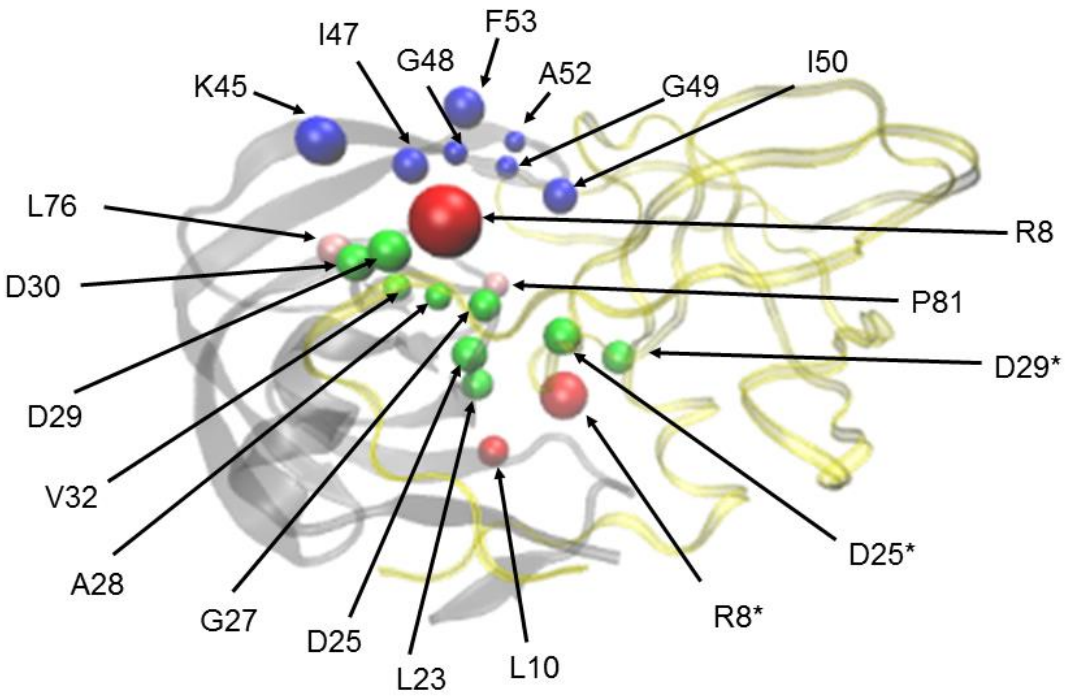

Figure 4

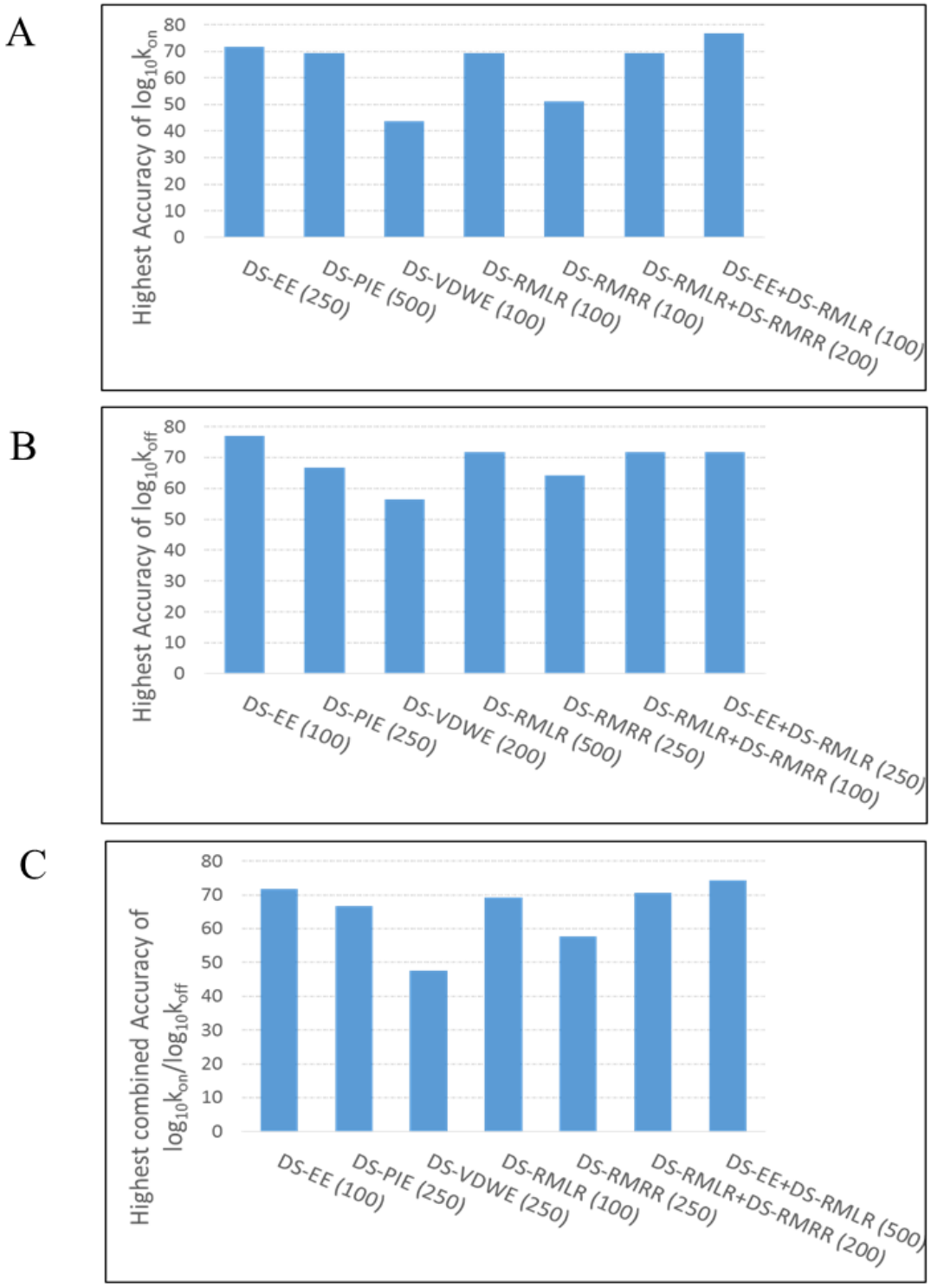

Figure 5

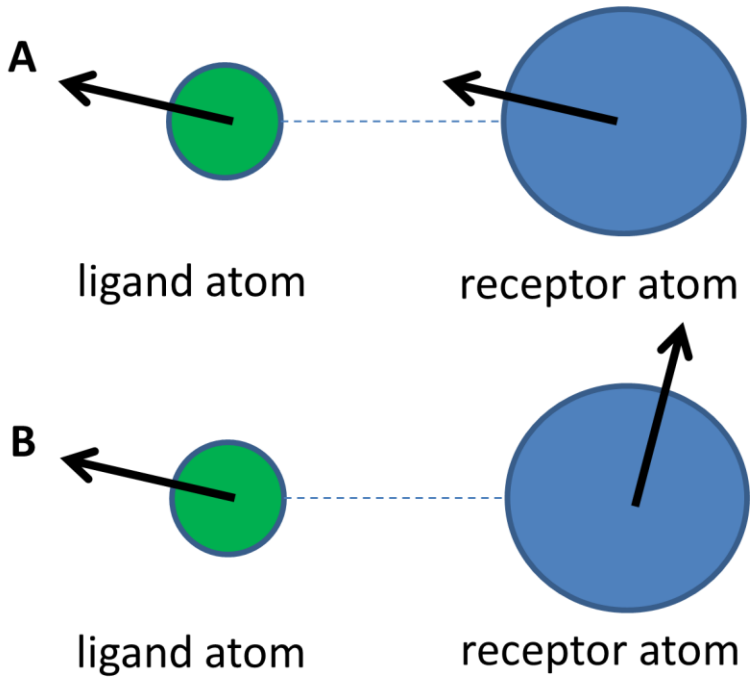

Figure 6

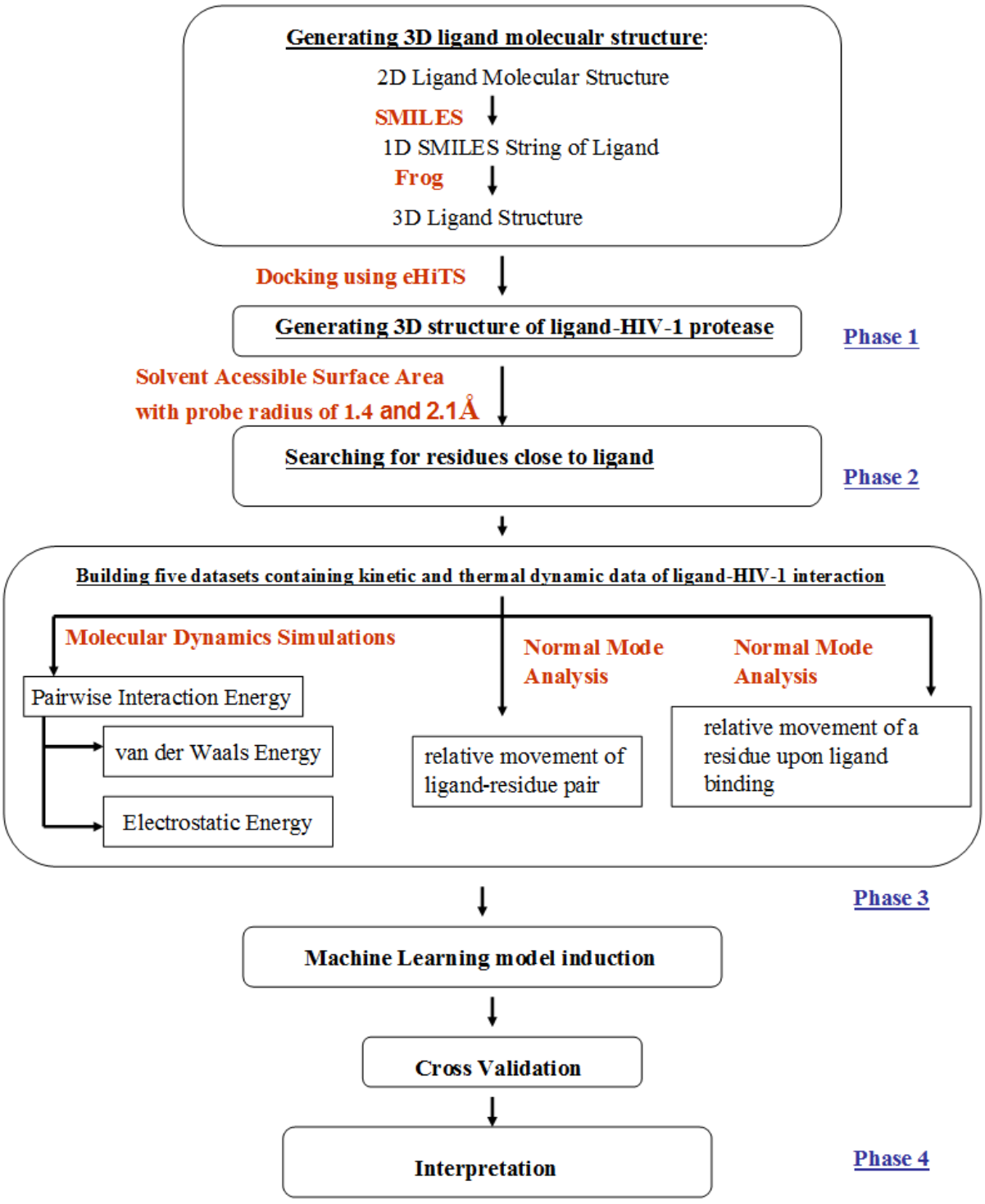

Figure 7

Supporting Information

Table S1. Results of the discretization. There are 9, 8, 10, and 12 training records in the binary classes of (0,0), (0,1), (1,0), and (1,1), respectively. Each record was identified by its corresponding ligand name.

Binary Class	(0,0)	(0,1)	(1,0)	(1,1)
Ligand I.D.	A037	Saquinavir	B347	B369
	B429	B440	B365	B388
	B409	Nelfinavir	A016	A021
	A038	Indinavir	A024	B355
	B412	B408	A047	A030
	B439	Ritonavir	A023	B322
	B268	Amp	A017	B425
	B277	U75875	B249	A045
	B435		A018	B295
			A015	B376
				A008
				DMP323
No. of records in the class	9	8	10	12

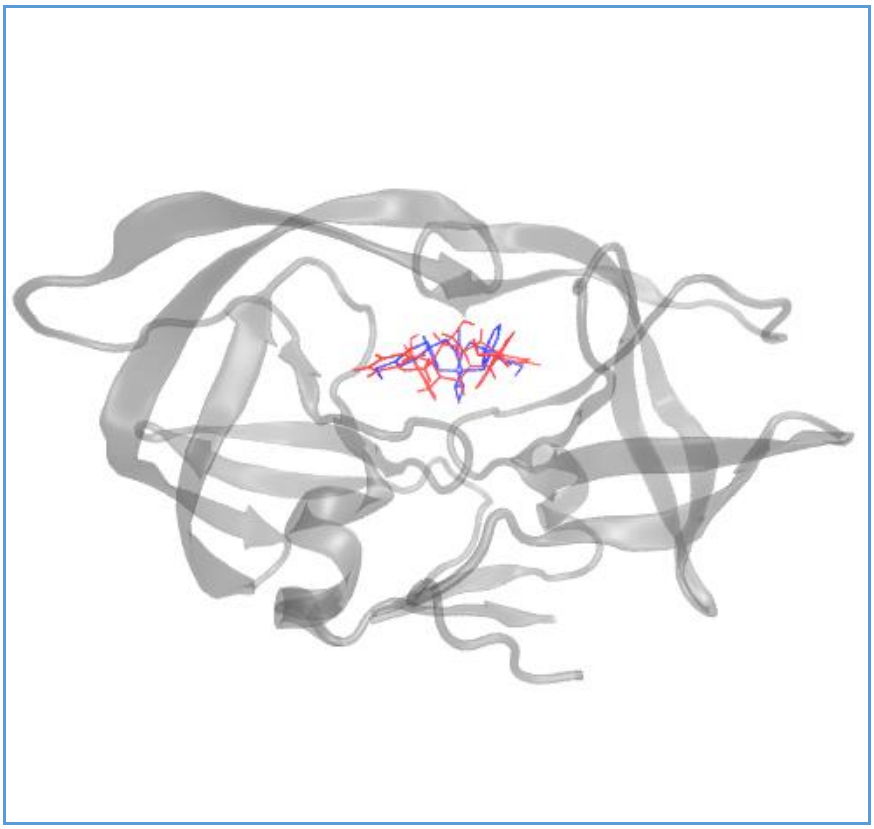

Figure S1. Docking ligand into HIV-1 protease. DMP (blue) is the co-crystalized ligand in HIV-1 protease (PDB code: 1QBS). Ligand A008 (red) is docked into the HIV-1 of 1QBS.

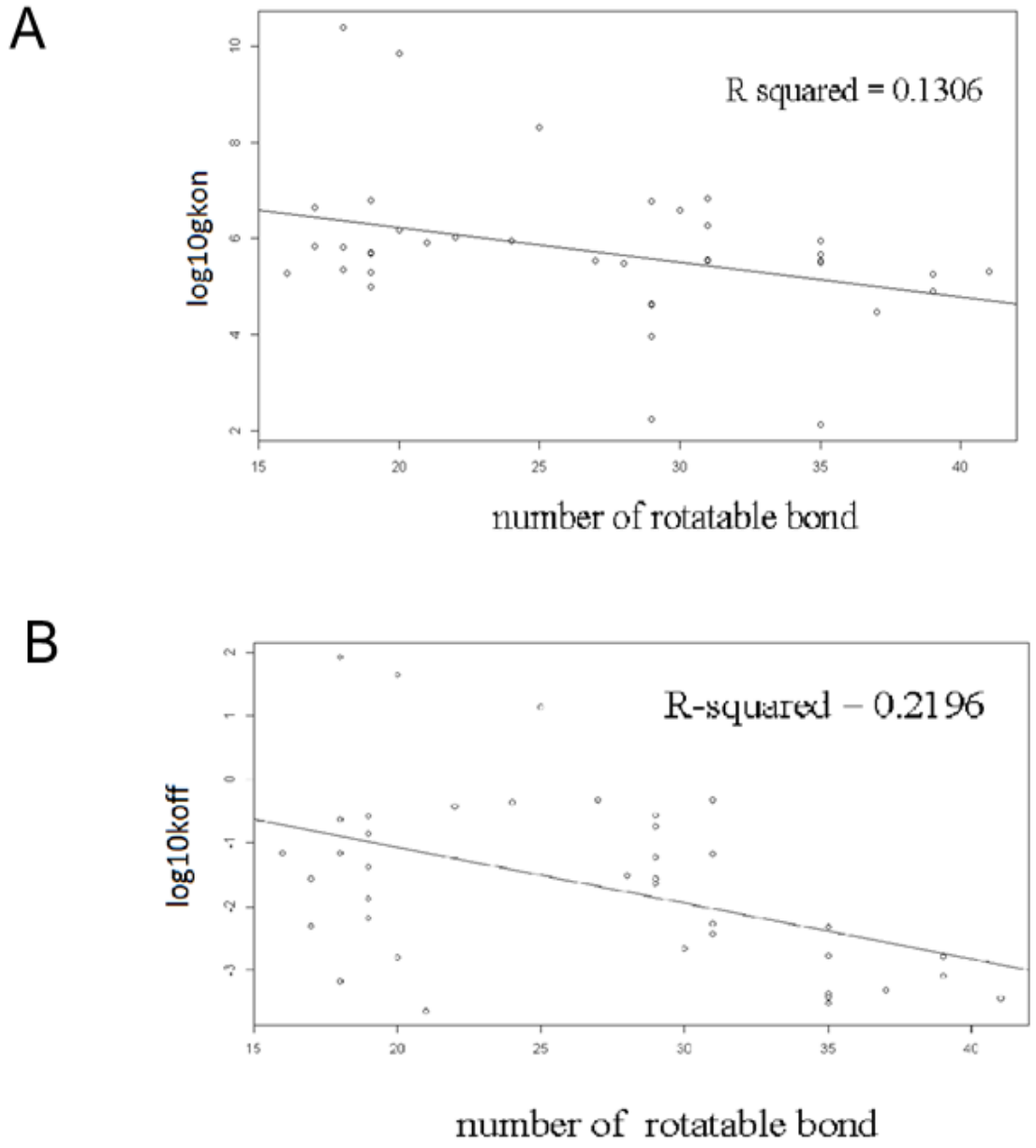

Figure S2. Number of ligand rotatable bond versus $\log_{10}k_{on}$ / $\log_{10}k_{off}$. Each data point represents one sample of ligand-HIV-1 complex.