2019-12-26

Characterizing protein-ligand binding using atomistic simulation and machine learning: Application to drug resistance in HIV-1 protease

Troy W. Whitfield
University of Massachusetts Medical School

Et al.

JCTC Journal of Chemical Theory and Computation

Biomolecular Systems

Characterizing protein-ligand binding using atomistic simulation and machine learning: Application to drug resistance in HIV-1 protease

Troy W. Whitfield, Debra A. Ragland, Konstantin B. Zeldovich, and Celia A. Schiffer

J. Chem. Theory Comput., **Just Accepted Manuscript** • DOI: 10.1021/acs.jctc.9b00781 • Publication Date (Web): 26 Dec 2019

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Characterizing protein-ligand binding using atomistic simulation and machine learning: Application to drug resistance in HIV-1 protease

Troy W. Whitfield,[*,†,⊥] Debra A. Ragland,[‡,¶] Konstantin B. Zeldovich,[§,‖] and Celia A. Schiffer[*,‡]

†*Department of Medicine, University of Massachusetts Medical School, Worcester, MA 01605*

‡*Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605*

¶*Current address: Clemson University, Clemson, SC 29634*

§*Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605*

‖*Current address: Sanofi Pasteur, Cambridge, MA 02139.*

⊥*Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605*

E-mail: Troy.Whitfield@umassmed.edu; Celia.Schiffer@umassmed.edu

Abstract

Over the past several decades, atomistic simulations of biomolecules, whether carried out using molecular dynamics or Monte Carlo techniques, have provided detailed insights into their function. Comparing the results of such simulations for a few closely related systems has guided our understanding of the mechanisms by which changes like ligand binding or mutation can alter function. The general problem of detecting and interpreting such mechanisms from simulations of many related systems, however, remains a challenge. This problem is addressed here by applying supervised and unsupervised machine learning techniques to a variety of thermodynamic observables extracted from molecular dynamics simulations of different systems. As an important test case, these methods are applied to understanding the evasion by HIV-1 protease of darunavir, a potent inhibitor to which resistance can develop via the simultaneous mutation of multiple amino acids. Complex mutational patterns have been observed among resistant strains, presenting a challenge to developing a mechanistic picture of resistance in the protease. In order to dissect these patterns and gain mechanistic insight on the role of specific mutations, molecular dynamics simulations were carried out on a collection of HIV-1 protease variants, chosen to include highly resistant strains and susceptible controls, in complex with darunavir. Using a machine learning approach that takes advantage of the hierarchical nature in the relationships among sequence, structure and function, an integrative analysis of these trajectories reveals key details of the resistance mechanism, including changes in protein structure, hydrogen bonding and protein-ligand contacts.

Introduction

Organisms use mutation to respond to changes in environment. In so doing, they can produce novel protein variants whose modified physical characteristics, such as structure or dynamics, may offer a functional advantage under the selective pressure imposed by the altered environment. When the environmental alteration is due to the presence of a therapeutic agent, the variants with an advantage are said to be "resistant". Some well known examples where the altered biophysical properties of protein variants are understood to confer resistance include the T790M mutation in EGFR kinase and mutations at or near the catalytic site of HIV-1 protease. In EGFR kinase, the T790M mutation induces changes in ATP binding, thereby evading inhibitors used in cancer therapy,[1,2] while in HIV-1 protease, resistance can occur through mutations that cause a loss of favorable interactions at the binding site with its inhibitors, accompanied by a relatively mild compromise in the binding and processing of natural substrates.[3–6] Despite important examples such as these, however, the general problem of relating physical changes induced by mutations with functional outcomes remains a significant challenge.

The inference problem outlined above is inherently hierarchical, with protein function being dictated by changes at the sequence level, but mediated by alterations in physical properties. A variety of machine learning techniques naturally lend themselves to deciphering such relationships. In this study, a combination of supervised and unsupervised machine learning strategies is assessed for this problem via an application to HIV-1 protease evasion of antiviral inhibition. Specifically, the singular value decomposition (unsupervised) and regularized regression (supervised) techniques are used to analyze thermodynamic observables, including mean protein-inhibitor van der Waals energies, mean hydrogen bond occupancies and protein $C_\alpha - C_\alpha$ distances, collected from atomistic simulations of several susceptible and resistant protein-ligand systems.

Principal component analysis[7] (PCA) and its formulation using the singular value decomposition[8,9] (SVD) represent venerable strategies for dimensionality reduction and pat-

tern detection. As such, these methods can be described as *unsupervised* machine learning[10] and while these techniques have been used for a long time in many fields, there is a history of application to atomistic simulations specifically. PCA and SVD analysis methods (and the closely related quasiharmonic analysis[11–13]) have been developed for characterizing protein molecular dynamics (or Monte Carlo[11,14]) trajectories.[15–17] These applications have been used to identify slow collective variables for chemical insight[14–17] and as an ingredient to model reduced dynamics.[18,19] More recently, autoencoders (neural networks) have been applied to trajectories from molecular simulations as an alternative dimensional reduction technique to PCA.[20–23]

Likewise, *supervised* machine learning techniques, where data labels are used to train predictions, have been applied to atomistic models and simulations. Some of these applications include using regularized kernel ridge regression on molecular geometry and charge descriptors to predict molecular atomization energies[24] or on basis functions to construct density functionals.[25,26] Neural networks have been used to define potential energy surfaces trained on quantum mechanical energies,[27,28] to define free energy surfaces,[29,30] or for charge assignment[31] for force field parameterization[32] and other applications.

For the most part, these applications of machine learning to atomistic simulations have been directed in an effort to construct efficient and accurate approximations for a many-body problem,[24–28,31–34] or to characterize the reaction coordinates for a given system.[14–23,35] In contrast, the approach presented here uses machine learning to detect an association between the occurrence of specific physical changes and functional outcomes among a set of systems that sample different phenotypes. Applied to HIV-1 protease, this problem corresponds to selecting a subset of chemically intuitive thermodynamic observables (e.g. mean hydrogen bond occupancy between donor A and acceptor B, etc.) that can be used as "features" to accurately distinguish between drug susceptible and resistant protease variants among a set of strains that includes examples of both classes. This problem is similar to scoring protein-ligand binding affinity from static (e.g. x-ray crystallographic) structures,[36–42] except that

the starting point here is a set of properties extracted from simulated systems at thermal equilibrium in aqueous solution, as opposed to fixed structures, with a clear emphasis placed on the interpretability of these features. Moreover, by modeling the dependence of altered physical properties on protein sequence, the current approach is designed to respect the hierarchy of mutations, physical properties and protein function described above.

The remainder of the paper is organized as follows. Background on the HIV-1 protease and its potent inhibitor darunavir[43,44] is presented in the following section, along with a description of the HIV-1 protease variants studied here. The Methods section includes a description of the simulations that were used and a brief outline of how SVD is applied to analyze multiple related systems, followed by strategies for combining SVD on simulation-based estimates of physical properties with regularized regression on protein amino acid sequence, and for a purely regression-based (supervised) approach. Results are presented for cross-variant analysis of HIV-1 protease binding to darunavir, characterizing the relationship between resistance to inhibition and changes in mean protease-inhibitor van der Waals energies, mean hydrogen bond occupancies and protein $C_\alpha - C_\alpha$ distances, respectively. Finally, a summary is given of the general utility of the techniques presented here, which can be applied to a variety of systems, along with the specific insights gained by applying them to study drug resistance in HIV-1 protease.

Background: HIV-1 protease inhibition and resistance

The human immunodeficiency virus type-1 (HIV-1) is an RNA retrovirus that leads to acquired immunodeficiency syndrome (AIDS) among infected individuals. The HIV-1 genome encodes the sequences of 19 proteins, including a protease enzyme that is required for the cleavage of polypeptide precursor molecules into mature proteins. HIV-1 protease is a homodimeric aspartic protease composed of 99 amino-acid monomers, with access to the active site controlled by a pair of flaps[45] (see Fig. S1). Because this protease is essential for the

production of infectious virions, it was an early and important target for drug therapies to inhibit its function.[46]

Darunavir[43] is an HIV-1 protease inhibitor with high binding affinity[44,47] that can be effective against strains where resistance to other inhibitors has developed.[43,44,47,48] Despite this effectiveness and the associated delay in the onset of protease resistance to darunavir inhibition, however, resistance has been observed in the presence of multiple simultaneous mutations.[49,50] As with some other inhibitors,[51] many of these mutations are distal from the catalytic site,[49] making their effects on inhibitor binding more difficult to interpret than those of proximal mutations, where the substrate envelope hypothesis, which predicts that HIV-1 protease inhibitors that fit within the overlapping consensus volume of the substrates are less likely to be susceptible to drug-resistant mutations,[4] is a useful guide. In order to gain mechanistic insight into the role that specific mutations play in this resistance, molecular dynamics simulations were previously carried out on 15 selected HIV-1 protease variants,[52] chosen to include drug susceptible wild-type controls, along with strains that are resistant to darunavir *in vivo*[50] and/or *in vitro*.[52] Here, we focus on using supervised machine learning to analyze thermodynamic observables, such as mean intra-protease hydrogen bond occupancies, collected from this set of simulations.

The sequence variants of HIV-1 protease that were selected for study with molecular dynamics simulations are listed in Table 1 and comprise enzymes that were determined, by various means, to be susceptible or resistant to darunavir inhibition. Shorthand names for each variant (e.g. "VSL", "VEG" etc. for variants with multiple mutations and descriptive names such as "V32I" for variants with a single mutation) are used throughout and are listed in Table 1. The amino acid sequence for each simulated protease variant, including two well-studied wild-type enzymes, the NL4-3 clone[54] and the Q7K autolysis resistant variant[55,56] of the ARV2/SF2 strain,[57] here simply called "SF-2", are listed in the accompanying multiple sequence alignment (Fig. S2). This panel of 15 HIV-1 protease variants includes several well known mutations, such as L10F, V11I, V32I, L33F, K43T, M46I, I47V, G48MV,

Table 1: Darunavir resistance in 15 selected HIV-1 protease variants. Variant names are colored according to their categorical label in the "susceptible" column. The amino acid sequence for each variant is given in a separate multiple sequence alignment (Fig. S2).

Name	K_i[a](pM)	$\frac{K_i}{K_i^{SF-2}}$	IC_{50}[b](nM)	$\frac{IC_{50}}{IC_{50}^{NL4-3}}$	EC_{50}[c](nM)	$\frac{EC_{50}}{EC_{50}^{NL4-3}}$	susceptible
SF-2	2	1					Yes
NL4-3			0.8	1	3.98	1	Yes
V32I	7	3.5					Yes
L33F							Yes[c]
L76V	3	1.5					Yes
I84V							Yes[c]
I93L							Yes[c]
DM	45	22					No
DRV^r8							No[c]
DRV^r10							No[c]
VSL			31.2	39	320	80	No
SLK			19.2	24	32.5	8	No
KY			89.6	112	1160	291	No
ATA				> 200			No
VEG				> 200	7800	1959	No

[a] Reported by Ragland *et al.*[53] [b] Reported by Varghese *et al.*[50] [c] Reported by Ragland *et al.*[52]

I50V, F53L, I54MSV, Q58E, G73ST, L67V, V82FA, I84V, L89V and L90M, that are associated with resistance to inhibition.[50,58,59] Individually, none of these mutations is known to significantly diminish the binding affinity of darunavir for the protease.[49] In different combinations, however, resistance has been observed[50] (see Table 1 and Fig. S2). As some of the mutations listed above are distal from the active site of the enzyme, understanding how they affect inhibitor binding is not straightforward. Indeed, while the complex mutational response that characterizes resistance (to darunavir inhibition) in HIV-1 protease makes uncovering mechanistic insight such a challenging problem, it also suggests machine learning as a particularly appropriate strategy to apply.

A variety of measurements to assess inhibitory activity are available, including inhibitory constants (K_i), the inhibitor concentration yielding 50% inhibition in the concentration-response curve (IC_{50}), or 50% effective concentrations (EC_{50}) from phenotypic dose-effect curves. In Table 1, some of the protease variants listed have one or more of these related[60]

measurements associated with them, while the susceptibility of other strains to darunavir inhibition is known only qualitatively from monitoring populations in cell culture experiments:[52] resistant stains are abundant under conditions where inhibitor concentration is high. When confronted with disparate and sometimes qualitative target data such as these, it is useful to cast the problem as one of classification. Accordingly, each HIV-1 protease variant listed in Table 1 is assigned a binary classification as either "susceptible" or "not susceptible" to darunavir inhibition.

Methods

Molecular dynamics simulations of inhibitor-bound HIV-1 protease variants

Each of the variants listed in Table 1 has previously[52] been simulated in complex with darunavir using molecular dynamics. Where darunavir-bound crystal structures were available[53] (the SF-2, V32I, L76V and DM variants), they were used as starting coordinates, including crystallographic water molecules, for the simulations prior to structural optimization, equilibration and data collection. Otherwise, homology models were used as initial coordinates.[52] The homology models were constructued from x-ray crystal structures to include darunavir and the important bridge water molecule between the inhibitor and the protease flaps. Tautomerization states were optimized using Epik[61,62] from the Schrödinger Suite and hydrogen-bond networks and protonation states were determined and optimized using PROPKA[63] at pH 7.0, with exhaustive sampling of water orientations and minimization of the hydrogen atom configurations of altered species. The protonation states for the catalytic aspartic acid residues were asymmetric.[53,64] Finally, interaction energies of hydrogen atoms were minimized under the OPLS2005[65] force field.

Simulations were carried out in an orthorhombic unit cell with periodic boundary conditions, explicit TIP3P aqueous solvent[66] at physiological (and electrostatically neutral) 150

nM NaCl concentration. The smooth particle mesh approximation[67] to the Ewald sum was used to evaluate Coulombic interactions. The isothermal-isobaric ensemble was simulated for a total of 300 ns (in three separate 100 ns simulations with randomly initialized velocities) at 300 K and 1 bar using the Desmond[68] implementation of the Martyna-Tobias-Klein extended system.[69] The OPLS2005[65] force field was used with multiple time steps[70] employed in the integrator for short-range (2 fs) and long-ranged (6 fs) interactions and a 9 Å cutoff applied for non-bonded interactions. Fast vibrational motions were constrained using the SHAKE[71] algorithm. In aggregate, 4.5 μs of simulations were collected.

Although simulations were carried out for both susceptible and resistant variants of HIV-1 protease in complex with darunavir, the inhibitor did not escape from the active site in any of the simulations. Examination of the root-mean-square displacement (RMSD) for the protease C_α atoms during these simulations reveals that the sampled conformations were within an RMSD of less than about 2 Å of the initial structures (Fig. S3). These conformations correspond to well-sampled bound states with no flap opening. Examination of the per-residue root-mean-square fluctuations (Fig. S4), however, indicates that increased fluctuations at the flap, flap elbow and cantilever are prevalent among the resistant variants.

In order to understand how various microscopic interactions observed in these simulations can be used to classify sequence variants as either susceptible or resistant, thereby gaining mechanistic insight into how the enzyme evades inhibition, thermodynamic averages for ligand-protease van der Waals interactions, intra-protease hydrogen bond occupancies and intra-protease $C_\alpha - C_\alpha$ distances collected from these simulations were analyzed here using unsupervised and supervised machine learning.

Unsupervised learning on trajectories from multiple systems

The singular value decomposition is used in a wide range of applications that encompasses pseudoinverse and optimization problems,[9] and signal processing, including dimensionality reduction. The familiar decomposition theorem is easily stated: for a $n \times p$ data matrix $\mathbf{D}$,

the singular value decomposition is

$$\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T,$$

where the non-zero elements of the diagonal matrix $\boldsymbol{\Sigma}$ are the singular values, $\mathbf{U}$ is a $n \times n$ unitary matrix in the column space of $\mathbf{D}$ and $\mathbf{V}$ is a $p \times p$ unitary matrix in the row space of $\mathbf{D}$. Since $\mathbf{U}$ and $\mathbf{V}$ are unitary it follows that

$$\mathbf{D}\mathbf{D}^T = \mathbf{U}\boldsymbol{\Sigma}^{\mathbf{2}}\mathbf{U}^T, \tag{1}$$

$$\mathbf{D}^T\mathbf{D} = \mathbf{V}\boldsymbol{\Sigma}^{\mathbf{2}}\mathbf{V}^T. \tag{2}$$

If the matrix $\mathbf{D}$ is *centered*, the singular value decomposition can be used to compute the *principal components* for a covariance matrix of the data. There are different choices for centering the data,[72–74] including subtracting the row or column means:

$$D_{ij} = D'_{ij} - \tfrac{1}{p}\textstyle\sum_{j=1}^{p} D'_{ij}, \tag{3}$$

$$D_{ij} = D'_{ij} - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} D'_{ij}, \tag{4}$$

where $\mathbf{D}'$ is the raw data matrix. When the row mean (eq. 3) is used, $\mathbf{D}\mathbf{D}^T$ becomes the covariance matrix of the rows, by substitution into eq. 1. The corresponding principal components are the left singular vectors, $\mathbf{u}_i$, which are conventionally ordered by the singular values s_i along the diagonal of $\boldsymbol{\Sigma}$. The variances of the principal components are proportional to s_i^2 (eq. 1). Alternatively, subtracting the column mean (eq. 4) makes $\mathbf{D}^T\mathbf{D}$ the correlation matrix of the columns.

Here, the interest is in applying the SVD to sets of p thermodynamic observables collected from atomistic simulations of n systems. The corresponding matrices of thermodynamic observables that are considered include mean per-residue ligand-protease van der Waals interactions, $\mathbf{D}^{\mathrm{LJ}}$ (computed using the Lennard-Jones potential from the OPLS2005[65] force

field), intra-protease hydrogen bond occupancies, $\mathbf{D}^{\mathrm{HB}}$, and intra-protease $\mathrm{C}_\alpha - \mathrm{C}_\alpha$ distances, $\mathbf{D}^{\mathrm{d}_{\mathrm{C}_\alpha}}$. In the case of the mean van der Waals energies, for example, there were $p = 64$ residues that had non-zero mean interaction energies with the ligand in at least one of the simulated systems. Following the machine learning nomenclature, these p observables can be called "features". Column centering (eq. 4) is used in conjunction with SVD to determine the "eigenfeatures", which are the right singular vectors, $\mathbf{v}_i$. The first eigenfeature, $\mathbf{v}_1$, defines the axis of highest variance in the space of the features, while the corresponding principal component score $s_1\mathbf{u}_1$ gives the coordinates for the n systems in the principal component space. Together, the right and left singular vectors can provide valuable insight into which features are most responsible for phenotypic changes. For example, if $s_1\mathbf{u}_1$ (or, equivalently, the unscaled left singular vector $\mathbf{u}_1$) effectively delineates the HIV-1 protease strains that are susceptible and resistant to its inhibitor, then projecting the eigenfeature $\mathbf{v}_1$ onto the original features can identify particularly important (i.e. large) components.

The ability to read mechanistic meaning from the easily interpreted components of a given $\mathbf{v}_k$ (e.g. van der Waals or hydrogen bonding interactions between specific pairs of atoms) may come at the expense of not finding the most compact low-dimensional representation of the data matrix (e.g $\mathbf{D}^{\mathrm{LJ}}$). By using SVD to carry out dimensional reduction, one is restricted to a linear transformation of the input data, yet it might be possible to find a lower-dimensional embedding that separates the susceptible and resistant strains via a non-linear transformation such as an autoencoder. Although it has recently been informative to relate the distributions of latent variables for an autoencoder representation of various systems to physically interpretable collective variables for phase transitions[75] and reaction coordinates,[76] choosing to employ SVD for the current application can nevertheless be motivated by the convenience of interpreting the individual components for a $\mathbf{v}_k$ of interest versus the weights of a neural network.[77]

Penalized regression for principal components

The unsupervised analysis outlined above, on p features extracted from simulations of n systems, addresses only half of the problem that was laid out in the Introduction: varying physical features can now be associated with a change in phenotype, but no information regarding how sequence (i.e. a set of specific mutations) underlies these changes is provided. This connection connection can be made by applying a supervised learning approach. For example, a linear model may be fitted to the component scores $s_i\mathbf{u}_i$. Linear models, as opposed to more general supervised approaches like support vector machine,[78] support vector regression,[79] random forest[80] etc., can have the advantage of easier interpretability.

The principal components are interpreted most readily when the spectrum of the singular values is sharply peaked—in other words, low in entropy[81]—and when phenotypic variation correlates with variation in one of the leading component scores. Without loss of generality, one can assume that at least one of the component scores, $s_1\mathbf{u}_1$ for example, can be used to delineate different phenotype classes. In this case, a relationship between this variance and sequence changes among the n simulated protein variants can be formulated using a linear model. After first defining, $\mathbf{u}_1 \equiv \mathbf{u}$, the (unscaled) score for each protein variant $1 \leq i \leq n$ can be written:

$$u_i = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i = \beta_0 + \sum_{j=1}^{m} x_{ij}\beta_j + \epsilon_i, \tag{5}$$

where $\beta_0, \boldsymbol{\beta}$ are the coefficients, ϵ_i is the random error associated with protein variant i and x_{ij} is the jth covariate of variant i, defined here as an indicator variable for amino acid substitutions at candidate site j:

$$x_{ij} = \begin{cases} 0 & \text{for} \quad \text{wild-type amino acid at candidate site } j \text{ of variant } i \\ 1 & \text{for} \quad \text{non-wild-type amino acid at candidate site } j \text{ of variant } i. \end{cases}$$

For the set of HIV-1 protease variants studied here (Table 1), there are $m = 50$ $(1 \leq j \leq m)$ sites that vary out of the 99 total residues present (see Fig. S2). The $n \times m$ covariate

matrix $\mathbf{X}$ for the panel of $n = 15$ protease variants is represented in eq. S1 (Supporting Information). Note that since the wild-type NL4-3 reference strain is included in $\mathbf{X}$, with $x^T_{\mathrm{NL4-3}}$ the null vector, then $u_{\mathrm{NL4-3}} = \beta_0$ is the intercept.

The coefficients in eq. 5 can be fitted by regression. In the protease panel (Table S1), however, the number of predictors (i.e. amino acid positions with sequence variation) is greater than the number of observations (i.e. sequence variants), so regularization[82,83] is important for avoiding overfitting. The coefficients were fitted, therefore, by solving[84]

$$\min_{\beta_0,\boldsymbol{\beta}} \left[\frac{1}{2n} \sum_{i=1}^{n} \left(u_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 + \gamma P_\alpha(\boldsymbol{\beta}) \right],$$

where

$$P_\alpha(\boldsymbol{\beta}) = \sum_{j=1}^{p} \frac{1-\alpha}{2} \beta_j^2 + \alpha |\beta_j| \tag{6}$$

is the elastic-net penalty.[83,85] In fitting the coefficients $\beta_0, \boldsymbol{\beta}$, the elastic net parameter ($\alpha \in [0,1]$) was $\alpha = 0.95$ and γ was chosen to minimize errors under five-fold cross-validation, with larger values of γ implying a larger penalty and a sparser (i.e. with a smaller number of non-zero coefficients), more interpretable, but perhaps less accurate solution. The relatively high value chosen for the elastic net parameter means that the l_1-norm penalty (i.e. the $\sum_j \alpha|\beta_j|$ term in eq. 6) is emphasized. Because the l_1-norm penalty leads to sparser solutions,[82] this emphasis serves to develop insight by selecting only the most important sites of amino acid variation for predicting $\mathbf{u}$. The elastic net penalty includes an l_2-norm penalty, the presence of which has been shown to aid in feature selection when correlation among features is present,[83,84] as is the case for the problems described here. Taken its own (i.e. when $\alpha = 0$), the l_2-norm penalty leads to ridge regression.[86]

In regression problems with correlated predictors, both the l_1-norm and l_2-norm penalty in eq. 6 can improve parameter estimation by shrinking the coefficients. Whereas correlated predictors will are retained under the l_2-norm penalty, the behavior under the l_1-norm penalty is different and the coefficients of some correlated predictors will be set to zero, thereby

selecting a subset of features that capture the strongest effects.[82,83] As a feature selection procedure for high dimensional regression problems, applying the elastic net penalty has the desirable characteristic of being a convex optimization problem, allowing for numerically efficient solutions. The feature selection associated with applying the l_1-norm penalty is desirable here precisely because it aids in the identification and interpretation of important mutations in eq. 5. One caveat, however, is that one might be interested in identifying highly correlated (or even collinear) features.

Finally, the estimates for non-zero coefficients derived from applying the l_1-norm penalty are known to be biased toward zero. As the principal motivation for using this penalty here is to help identify important features—in eq. 5, these features are mutations—any such bias is not generally of concern. When making predictions using the model, however, removing this bias can offer an improvement. A two-stage process, whereby features were selected using the elastic net penalty and another regression model was subsequently fitted by applying ordinary least squares to the selected features, called the "relaxed lasso",[87] or in the present case "relaxed elastic net", has been used here wherever coefficients have been reported.

Sequential penalized regressions

As an alternative to the partially unsupervised approach described above, supervised learning can be performed directly on p properties extracted from simulations of n systems. Such an approach can be an advantage in cases where no single principal component score, $s_i\mathbf{u}_i$, is able to adequately discriminate changes in phenotype among the different systems. Even in cases where one of the principal component scores effectively partitions different systems according to phenotype, however, interpreting the variances captured by different features may not be straightforward, as only a subset of the projections for the corresponding eigenfeature, $\mathbf{v}_i$, may vary according to phenotype. In such cases, the simplified feature selection permitted by beginning from a supervised learning approach, where the most phenotypically important features are identified without the need for an element-wise examination of eigenfeatures,

can be convenient.

As noted above, the partial and heterogeneous nature of the target data for these n systems (see Table 1) makes it appropriate to cast supervised learning on p features as a classification problem. A binary (e.g. susceptible versus resistant) classifier can be written as a logistic regression:

$$\log \frac{Pr(Y_i = 1|\mathbf{D}_i)}{Pr(Y_i = 0|\mathbf{D}_i)} = \beta_0 + \mathbf{D}_i^T \boldsymbol{\beta} + \epsilon_i = \beta_0 + \sum_{j=1}^{p} D_{ij}\beta_j + \epsilon_i, \tag{7}$$

where $Y_i \in \{0, 1\}$ is the binary response (i.e. taken from the "susceptible" column in Table 1) for system i and $Pr(Y_i = 1|\mathbf{D}_i)$ is the conditional probability of observing a positive response, given the covariate vector $\mathbf{D}_i$. The term on the left hand side of eq. 7 is the log-likelihood ratio, $\beta_0, \boldsymbol{\beta}$ are the coefficients, ϵ_i is the (logistic distributed) random error associated with system i ($1 \leq i \leq n$) and D_{ij} is the jth covariate of variant i. This model makes an obvious analogy with that in eq. 5, but the matrix of covariates in eq. 7 is now just the data matrix $\mathbf{D}$. In data-sets that are large enough to support multi-category responses, eq. 7 can be generalized to multinomial logistic regression[83] and in data-sets where the response variable is known quantitatively for each case (e.g. an inhibition constant has been measured for all variants in Table 1), the problem could be formulated as a multiple linear regression.

The coefficients in eq. 7 may be fitted by maximizing the likelihood (or equivalently, minimizing the negative log-likelihood) subject to a penalty:

$$\min_{\beta_0, \boldsymbol{\beta}} \left[-\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\beta_0, \boldsymbol{\beta}; \mathbf{Y}, \mathbf{D}) + \gamma P_\alpha(\boldsymbol{\beta}) \right], \tag{8}$$

where $P_\alpha(\boldsymbol{\beta})$ is the elastic net penalty (eq. 6), $\mathcal{L}$ is the log-likelihood for the logistic function and $\mathbf{Y} = \{Y_1, Y_2, ..., Y_n\}$. As above, regularization using an elastic net penalty was employed to limit overfitting and to facilitate feature selection. The elastic net parameter was $\alpha = 0.95$ and γ was chosen to minimize errors under three-fold cross-validation.

The procedure described above identifies a subset of the physical descriptors that, for a

given data matrix $\mathbf{D}$, can be used to predict phenotype classes. In the current application, for example, these descriptors can include van der Waals interactions between an inhibitor of HIV-1 protease and specific residues in the enzyme; the classes are strains that are "susceptible" versus "resistant" to darunavir inhibition. The subset of mutations that appears to control changes in these selected physical descriptors, or features, can be inferred by again appealing to regression on sequence descriptors, as in eq. 5. Unlike in eq. 5, however, the outcome variable is now the physical descriptor itself, rather than a principal component score. For a selected feature, k, one can take $D = D_k$ as the kth column of the data matrix and fit the following linear model:

$$D_k = \beta_0 + \mathbf{x}_k^T \boldsymbol{\beta} + \epsilon_k = \beta_0 + \sum_{j=1}^{m} x_{kj}\beta_j + \epsilon_k, \tag{9}$$

where there are $1 \leq k \leq n$ systems and the covariate matrix $\mathbf{X}$ is given as in eq. S1. In this fully supervised approach to analyzing thermodynamic observables collected from simulations of many related systems, phenotypically important features are identified using logistic regression (eq. 7) and changes in amino acid sequence are regressed on each selected feature (eq. 9) separately.

Results

Short-range protein-ligand interactions

As they are sensitive to changes in binding geometry, protein-ligand van der Waals interactions are important probes of affinity[38,39] and have been observed to change in response to resistance-associated mutations in HIV-1 protease.[3,52,53,88,89] Furthermore, although enthalpic terms other than van der Waals interactions also contribute to the free energy for ligand binding, changes in van der Waals interactions have proven to be among the most predictive physics-based features in machine learning estimates of binding affinity in HIV-1

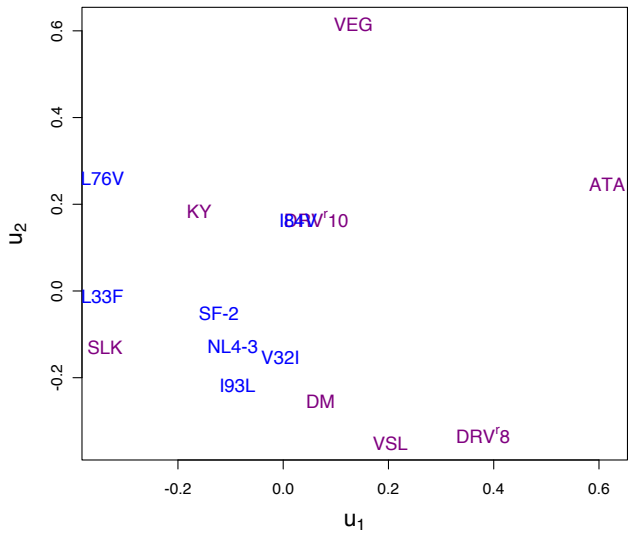

Figure 1: HIV-1 protease variants projected on top two principal component scores for the column centered mean per-residue protein-ligand van der Waals interactions.

protease.[38]

Within the simulations of 15 HIV-1 protease variants studied here, there were 64 residues that had non-zero mean van der Waals interactions with the ligand over the course of the simulations. Applying SVD to the column centered 15×64 matrix of these data, $\mathbf{D}^{\mathrm{LJ}}$, results in a sharply peaked spectrum of singular values (Fig. S5 (a)) with more than 41% of the variance accounted for by the first singular value and an additional 21% by the second singular value. Plotting the unscaled principal component scores, $\mathbf{u}_1$ and $\mathbf{u}_2$, against one another for these two singular values shows that the two wild type strains, NL4-3 and SF-2, are approximately co-localized, as expected (Fig. 1). There is also a significant segregation of susceptible and resistant strains along $\mathbf{u}_1$, with resistant protease variants placed at higher values ($p < 0.04$, Fig. S6 (a)). The correspondence between resistance to darunavir inhibition and $\mathbf{u}_1$ is not perfect, however, with two resistant strains, KY and SLK placed at lower scores among the susceptible strains (Fig. 1).

By examining the corresponding right singular vector, $\mathbf{v}_1$, one can assess the relative importance of different eigenfeature components (Fig. 2(a) and Fig. S7). Many of the mean van der Waals interactions between darunavir and specific protease residues are ubiquitously

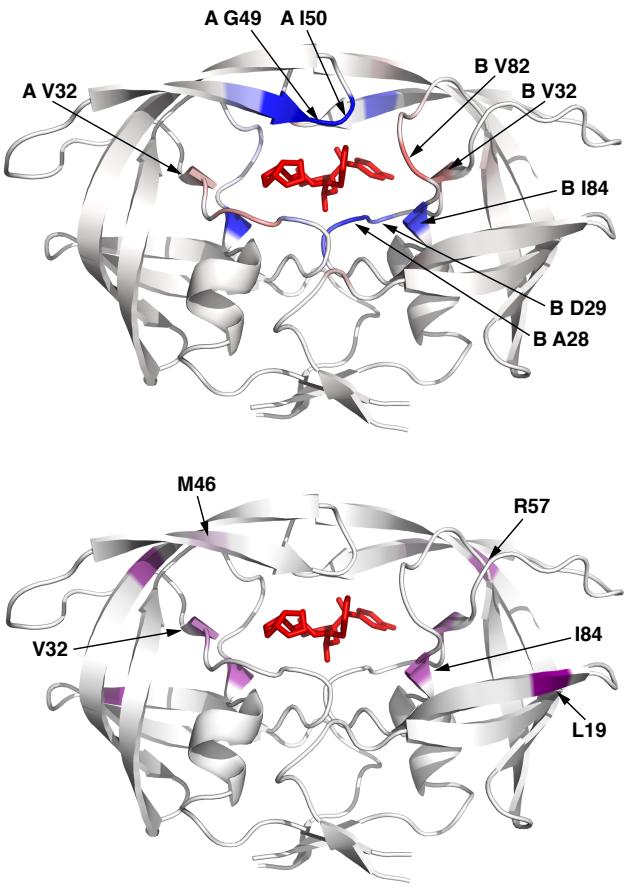

Figure 2: The resistance-associated signature of changes in ligand-protein van der Waals interactions is mapped onto the darunavir-bound structure[47] of HIV-1 protease. In (a), residues colored in darker blue indicate a greater loss in van der Waals interactions among resistant strains, as indexed by $\mathbf{v}_1$, while those colored in darker red indicate a greater gain. Residues colored in white showed no change. In (b), the fitted coefficients from penalized regression on $\mathbf{u}_1$ (eq. 5), defining a predictive subset of mutations, are colored darker violet for larger values of $\boldsymbol{\beta}$. All indicated coefficients are positive, meaning that amino acid substitutions at those residues imply an increase along $\mathbf{u}_1$. Here and elsewhere, the two monomeric subunits of the HIV-1 protease are labeled "chain A" and "chain B", with chain B located proximal to the sulfonamide moiety of darunavir.

weak among the simulated systems and therefore account for very little of the captured variance. Other interactions, by contrast, are important components of $\mathbf{v}_1$ that correspond to a loss in mean ligand-protein van der Waals energy among the resistant strains at that residue or to a gain. On balance, $\mathbf{v}_1$ catalogs a loss in van der Waals interaction energy corresponding to the development of resistance, yet there are some residues like V32 in both monomeric subunits and V82 located in the chain proximal to the sulfonamide moiety of

darunavir, where the opposite effect is observed (Fig. 2(a) and Fig. S7(b)) among the strains listed in Table 1. Some of the protease residues with the most altered mean ligand-protein van der Waals interactions across these variants are located in the flaps, above the catalytic site as shown in Fig. 2(a). These flap residues include I47, G48, G49 and I50. Isoleucine 84 is another active-site residue where van der Waals interactions with darunavir become less favorable in the resistant strains.

As noted above, $\mathbf{u}_1$ has a significant, but not perfect, correspondence with resistance among the HIV-1 protease variants listed in Table 1. Accordingly, one may find residues like G27, where the van der Waals interactions with the ligand are highly variable among our panel of protease variants and that have, therefore, large projections onto $\mathbf{v}_1$. The mean van der Waals interactions between darunavir and G27 vary significantly, yet they do so at least as much *within* the susceptible and resistant variants as *between* these two respective classes (Fig. S7(b)), suggesting that alterations in this interaction do not correlate with darunavir resistance. This example serves to emphasize that some of the features identified using SVD may appear important without really delineating the susceptible and resistant variants. For that task, supervised learning approaches can offer a direct solution.

Having made a connection between one of the principal components of the column centered matrix of ligand-protein van der Waals interactions and resistance to darunavir inhibition among a set of HIV-1 protease variants, it is natural to ask which specific mutations can best explain, or predict, changes in this eigenfeature. Carrying out penalized regression of $\mathbf{u}_1$ against the sequence predictors (eq. 5) identifies a subset of residues where mutations are predictive under cross-validation (Pearson's $r = 0.73$, mean square error=0.07, see Fig. S8(a)). The selected sites of mutation that predict $\mathbf{u}_1$ under this model, ordered by decreasing β_j are: L19, I84, R57, V32 and M46 (Fig. 2(b), Table S1). All of these coefficients are positive, meaning that mutations at any of these sites lead to increased $\mathbf{u}_1$. Some of these sites of mutation, like L19 and R57, are not generally associated with resistance to darunavir or to other protease inhibitors,[49,50,58,59] but their importance in the linear model for $\mathbf{u}_1$ is

nevertheless straightforward to understand: the L19Q mutation is present only in the ATA strain, which has a particularly large score, while R57K is present in the highly resistant VSL, ATA and VEG strains (see Fig. S2).

As a complementary approach, supervised machine learning on molecular properties collected from simulations of different systems (see Methods) can *directly* identify resistance-associated differences that may be masked by other sources of variance (e.g. tolerated variation among wild-type proteins). Carrying out penalized logistic regression on $\mathbf{D}^{\mathrm{LJ}}$ (eq. 8), results in the selection of several features that were also among the important components of $\mathbf{v}_1$ (see Fig. 2 and Fig. S7) from the SVD: the mean ligand-protease van der Waals interactions at residues G49 and I50 on chain A and at residues I47, A28 and D29 on chain B. Violin plots for these features are shown in Fig. 3(a), indicating that favorable van der Waals interactions are lost in the resistant variants (light blue violins) versus the susceptible variants (dark blue violins) for residues G49, I50 (chain A) and I47 (chain B), while these interactions can increase for residues A28 and D29 (chain B).

The logistic regression coefficients for the model are represented in Fig. 3(b) as directed edges that connect each physical feature (blue circles) according to its contribution to the classification of variants as susceptible or resistant (an outcome that is depicted with a red circle). The size and shade of these edges is determined by the coefficients of the model (eq. 8) in the following way: recalling that the encoding for variant i is $Y_i = 0$ for "susceptible" (see Table 1), the odds ratio for coefficient β_j is $e^{\beta_0+\beta_j}/e^{\beta_0} = e^{\beta_j}$ and sets the width of the edges. When the odds ratio is greater than 1 (black edges), resistance is *more* likely to occur as the predictor increases, while when the odds ratio is less than 1 (grey edges), resistance is *less* likely to occur as the predictor increases. For example, the increase that is observed among resistant protease variants for the mean van der Waals energy between the inhibitor and residue G49 (chain A) in Fig. 3(a) is rendered as a black edge in the graphical representation (Fig. 3(b)) of the fitted model. By contrast, the corresponding decrease that is observed in Fig. 3(a) for these interactions with residue A28 on chain B is rendered as

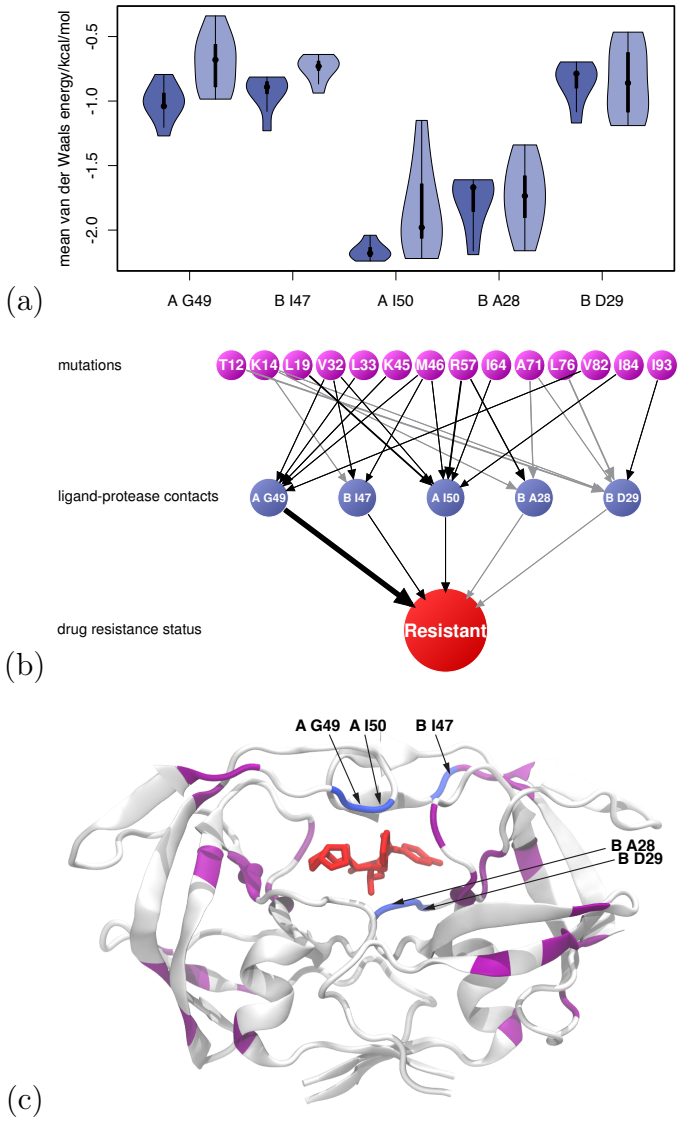

Figure 3: Hierarchy of regularized regression models to help decipher drug resistance mechanisms in HIV-1 protease. For each predictor (a) in a logistic regression model of resistance from mean per-residue ligand-protease van der Waals interactions, a linear model of sequence predictors has been fitted (b). The violin plots in (a) depict the kernel density of mean per-residue van der Waals energies for susceptible (left, darker blue) and resistant (right, lighter blue) protease variants. The directed edges shown in (b) correspond to fitted coefficients and are colored according to sign (black for positive, grey for negative) and sized according to importance. In (c) the physical and sequence features are projected onto the structure of the ligand-bound enzyme.

a grey edge in Fig. 3(b): the van der Waals interactions between the inhibitor and this residue tend to be more favorable among the resistant variants listed in Table 1, compared to the susceptible variants. The narrow width of the edge that connects residue A28 with

resistance status in Fig. 3(b) corresponds to a relatively small coefficient in the model.

As noted above, because a strong l_1-norm penalty is used in the fitting (eq. 8), the presence of correlated features will be attenuated in the final model. This behavior is by design, as the resulting model is sparse and therefore readily interpretable. Furthermore, one should not generally expect all of the most important $\mathbf{v}_1$ components from the SVD to appear as features in supervised learning: recall that not all of the variance captured using SVD is related to differences between susceptible and resistant strains. These considerations can help explain differences between the eigenfeature $\mathbf{v}_1$ (Fig. S7) and the relatively sparse set of features selected using penalized logistic regression (Fig. 3(a)).

For each of the important protein-inhibitor van der Waals interactions that were identified using supervised learning, a set of underlying sequence alterations can be inferred. As before, this inference is done using linear regression (eq. 9), but now separately on each of the features selected using logistic regression instead of on the relevant principal component scores (e.g. $\mathbf{u}_1$). The coefficients for linear regression of the selected van der Waals interactions against sequence features are visualized in Fig. 3(b) as directed edges that connect individual mutations (violet circles) with interactions (blue circles). These edges are shaded according to the sign of the coefficients, with black used for positive and grey used for negative coefficients. The widths of the edges are scaled by the magnitude of the coefficients, $|\beta_j|$.

As one example, consider the mutation at residue 46. In panel of HIV-1 protease variants studied here, this mutation is always from methionine to isoleucine (see Fig. S2) and is a well known resistance mutation[50,58,59] for protease inhibitors, including darunavir. In Fig. 3(b), this mutation makes a positive contribution to explaining the changes in mean darunavir-protease van der Waals interactions at three of the selected residues: G49 and I50 in chain A and I47 in chain B. The positive sign of these coefficients means that each of these interaction energies increases (i.e. becomes less favorable) when the M46I mutation is present (see Fig. S9).

To further illustrate the interpretation of the regression coefficients for sequence features in Fig. 3(b), consider mutations at residue 71. Mutations at residue 71, such as A71V, are generally described as "secondary" HIV-1 resistance mutations in the sense that they offer a reduction in inhibitor affinity only in the presence of additional mutations. In this context, the A71V mutation has been shown to alter resistance to different inhibitors, either in the form of enhancement[90] or diminution.[91] Among the protease variants studied here, the A71I and A71V mutations are associated with increased resistance (Fig. S2). This association can be seen via the grey edges (negative coefficients) in Fig. 3(b). When a mutation is present at residue 71, the van der Waals interactions between darunavir and residues 28 and 29 in chain B are reduced (i.e. become more favorable, see Fig. S10). These specific alterations in van der Waals interactions are, in turn, a distinguishing feature of the resistant and susceptible strains here.

Finally, consider the set of mutations that can be used to model changes in van der Waals interactions between darunavir and residue 50, chain A. This residue is located at the tip of one of the two flaps that control access to the catalytic site (Figs. S1, 2(a) and 3(c)). Although residue 50 is known to harbor so-called primary resistance mutations, such mutations are not prevalent in among the panel of HIV-1 protease variants studied here, which contains a single example: the VSL variant has an I50V substitution (Fig. S2). Instead, the observed loss in van der Waals affinity among the resistant strains appears to be accounted for by the mutations indicated in Fig. 3(b), each of which is represented as a black edge (positive coefficient). Several of these mutations, like V32I, M46I and I84V are primary resistance mutations whose effects include a loss in van der Waals affinity between darunavir and residue 50 (Fig. S11).

Mapping the selected mutations and alterations in van der Waals interactions onto the structure of the darunavir-bound HIV-1 protease, it is striking to note how many of the mutations are distal from the catalytic site. The resistance-associated alterations in the short-ranged van der Waals interactions are naturally localized near the inhibitor, while the

distal nature of the selected mutations implies that these changes are mediated by structural and/or dynamical alterations throughout the protease.

Intra-protein hydrogen bonding

Since the analysis of mean inhibitor-protease van der Waals interactions suggests that losses in affinity at key residues near the catalytic site are mediated by structural alterations elsewhere in the enzyme, it is reasonable to next interrogate alterations in intra-protease hydrogen bonding and protein geometry. Hydrogen bonds are essential determinants of protein secondary structure,[92] so resistance-associated changes in protein structure are likely to have signature alterations to hydrogen bonding.

Between the backbone intra-protease hydrogen bonds and those formed directly between darunavir and the protease, there were $p = 113$ non-zero interactions to consider, resulting in a 15×113 matrix, $\mathbf{D}^{\mathrm{HB}}$. Applying SVD to this column centered matrix of hydrogen bond occupancies results in a sharply peaked singular value spectrum (Fig. S5 (b)) with more than 34% of the variance accounted for by the first singular value and an additional 16% by the second singular value. Plotting the unscaled principal component scores, $\mathbf{u}_1$ and $\mathbf{u}_2$, against one another for these two singular values shows that the two wild type strains, NL4-3 and SF-2, are approximately co-localized (Fig. 4), as they were when applying the same type of analysis to the van der Waals interactions above. Examining the principal component scores, only $\mathbf{u}_2$ offers a significant ($p < 0.04$, Fig. S6 (b)) discrimination between the susceptible and resistant variants.

Note that the singular value decomposition offers no assurance that any singular value, let alone the largest singular value, will effectively account for *phenotypic* variance or classification. There is evidently some variability in intra-protease hydrogen bonding that does not affect resistance to inhibition. Nevertheless, inspection of $\mathbf{u}_2$ offers valuable insight regarding how intra-protease hydrogen bonds are perturbed among the darunavir-resistant strains.

By examining the right singular vector that is most relevant to resistance, $\mathbf{v}_2$, the impor-

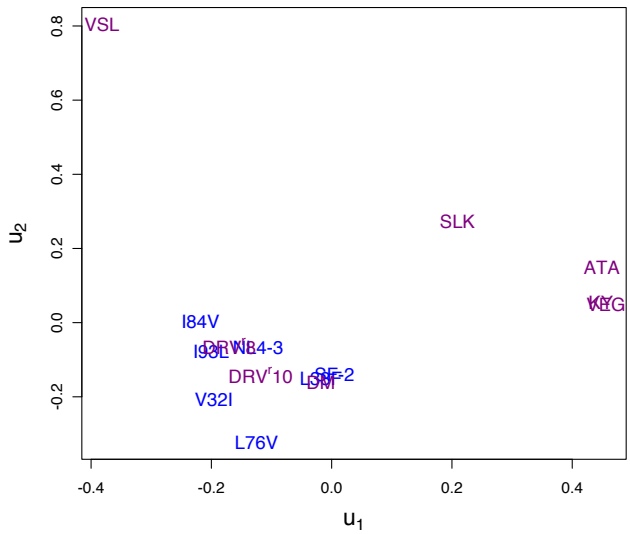

Figure 4: HIV-1 protease variants projected on top two principal component scores of the centered mean occupancies for intra-protease backbone hydrogen bonds.

tant components of this eigenfeature can be identified (Figs. 5, S12). Unlike with the van der Waals interactions between the inhibitor and protease, there is no notable overall loss in hydrogen bonding among the resistant variants, yet patterns of alterations are evident. Several of the important components of $\mathbf{v}_2$ are interactions between residues in "60s loop" and the β-sheet that begins at residue 70 (Fig. S12), sometimes called the "cantilever" (Fig. S1). The changes in hydrogen bonding occupancy for these residues, however, tend to be modest when comparing resistant and susceptible variants. Other large magnitude components of $\mathbf{v}_2$ include hydrogen bonding within each of the flap tips and between residues 67 and 12 in both chains. The former hydrogen bond occupancies are modestly increased, while the latter exhibit a dramatic decrease from the susceptible to the resistant variants. Perhaps not surprisingly, the hydrogen bond occupancy between the flap two tips (residues 50 and 51 in chains A and B, respectively) shows a decrease among the resistant variants (Figs. 5, S12).

To identify specific mutations that can explain $\mathbf{u}_2$, a regression model was fitted (eq. 5). The sequence features that were selected under cross-validation are highly predictive (Pearson's $r = 0.97$, mean square error=0.06, see Fig. S8(b)) and include mutations to L10 and I50 (Table S1), both of which are known to increase resistance to darunavir inhibition.[49]

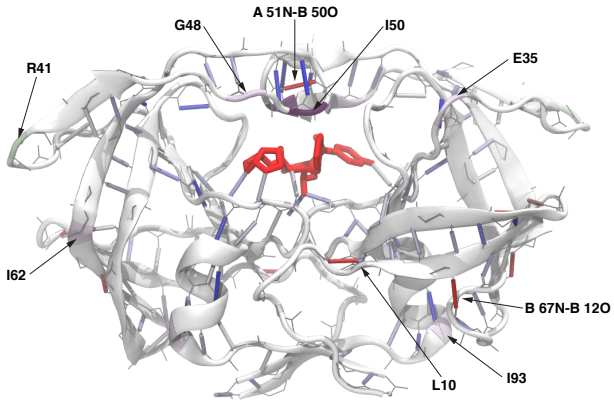

Figure 5: The resistance-associated eigenfeature of changes in intra-protease hydrogen bonding is mapped onto the darunavir-bound structure[47] of HIV-1 protease. Hydrogen bonds are depicted as cylinders colored in darker blue to indicate a greater gain in occupancy among resistant strains, as indexed by $\mathbf{v}_2$, while those colored in darker red indicate a greater loss. Hydrogen bonds colored in white showed no change in occupancy (see Fig. S12). The fitted coefficients from penalized regression on $\mathbf{u}_2$ (eq. 5), defining a predictive subset of mutations, are colored darker violet for larger positive values of $\boldsymbol{\beta}$ and green for negative values (i.e for β_{R41} only, see Table S1).

Among the HIV-1 protease variants listed in Table 1, the I50V mutation is present only in a single strain, VSL (Fig. S2). Including it in the regression model ensures that the high $\mathbf{u}_2$ score for VSL is accurately captured (Fig. 4). Likewise, the mutations at L10 (or collinear mutations at I54 that are not included in the model) delineate the highly resistant patient-derived variants in the panel. The remaining coefficients in the model are positive, with the exception of the negative coefficient for a mutation at residue 41. Mutations at residue 41, therefore, lead to a decrease in $\mathbf{u}_2$. The otherwise polymorphic R41K mutation is mostly present in the susceptible panel variants, making its contribution to the model clear, as $\mathbf{u}_2$ is indeed lower for these variants (Fig. 4).

Carrying out penalized logistic regression directly on the hydrogen bond occupancies in $\mathbf{D}^{\mathrm{HB}}$ identifies a sparse subset that can be used to classify HIV-1 protease variants as susceptible and resistant to darunavir inhibition (Fig. 6). These distinguishing hydrogen bonds include one formed between residue 67, located in a turn between a pair of β-sheets (sometimes called the "cantilever"[93] of the protease) and residue 12 in a β-sheet (sometimes

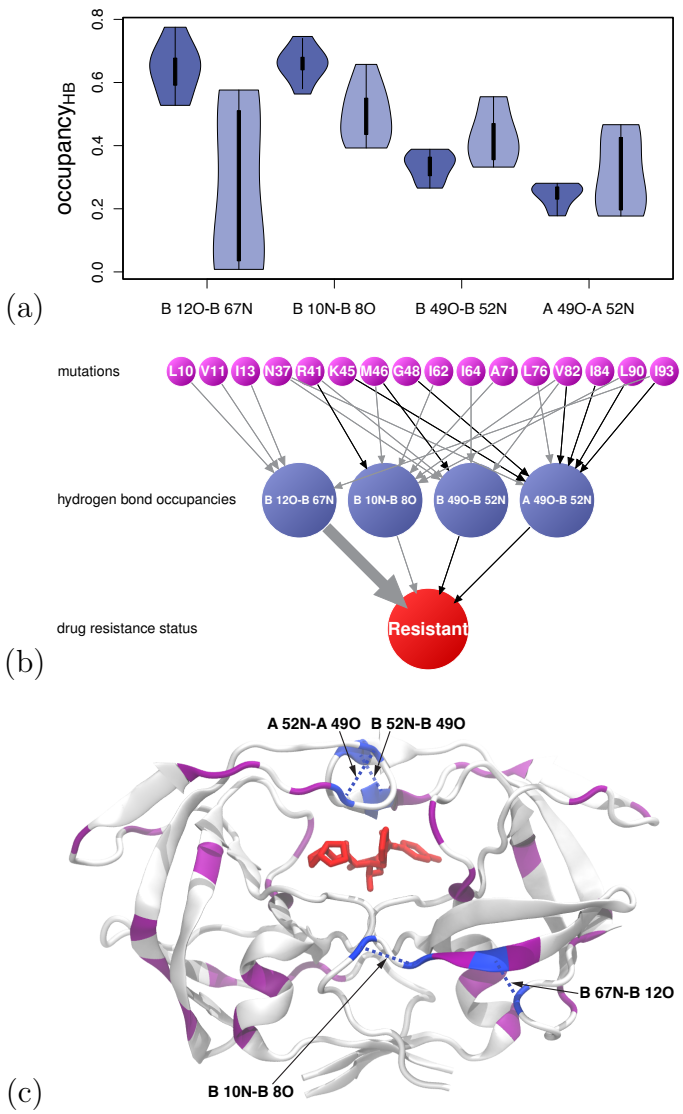

Figure 6: Visualization of the regularized regression model hierarchy used to to identify resistance-associated alterations in intra-protease hydrogen bonding. For each predictor (a) in a logistic regression model of resistance based upon the mean occupancies of intra-protease hydrogen bonds, a linear model of sequence predictors has been fitted (b). The violin plots in (a) depict the kernel density of mean occupancies for susceptible (left, darker blue) and resistant (right, lighter blue) protease variants. The directed edges shown in (b) correspond to fitted coefficients and are colored according to sign (black for positive, grey for negative) and sized according to importance. In (c) the physical and sequence features are projected onto the structure of the ligand-bound enzyme.

called part of the "fulcrum"[93]) in chain B. This hydrogen bond is severely disrupted among some of the resistant variants (Fig. 6(a)), suggesting a possible alteration in the coupling between these two domain elements of the enzyme. Another disrupted backbone hydrogen

bond was selected, connecting two residues, 8 and 10, that are located in a chain B turn near the core of the protease. This turn sits between the core β-sheet structures of the protein and those at the terminal dimer interface[94] (Fig. 6(c)). As both of these hydrogen bonds are disrupted in resistant variants, their corresponding coefficients in the model (eq. 8) are negative (grey edges in Fig. 6(b)). Note that, while these two hydrogen bonds in chain B were selected via penalized regression, their counterparts in the other monomer were similarly disrupted among resistant variants (Pearson's $r = 0.75$ and 0.74, respectively). Adding such correlated features to a linear model, however, generally offers little predictive benefit and a sparse set was selected under the l_1-norm bias of the elastic net penalty (eq. 9).

Two weaker features that were selected correspond to increased backbone hydrogen bond occupancies within each of the flaps (Fig. 6). These hydrogen bond occupancies have positive coefficients (black edges leading to the "resistant" status in Fig. 6(b)). These features can be omitted from the model, however, without significant impact on prediction: they are included in Fig. 6 only to identify the most informative of the remaining features.

The mutations that can predict each of the selected hydrogen bond occupancies were identified using penalized regression and are listed in Fig. 6(b). The substantial loss of hydrogen bonding observed among resistant strains between the chain B residues 67 and 12 appears to be controlled by mutations nearby residues: 10, 11, 12 and 93. These mutations are variously present among the highly resistant patient-derived strains that exhibit the greatest disruption of this hydrogen bond (Table S1). In each case, the model coefficient is negative (grey edges): the presence of amino acid substitutions here leads to a loss of hydrogen bonding. Mutations at residues 10 and 11 are known to confer resistance to inhibition,[50] including by darunavir,[49] while the polymorphic I93L mutation, located within the α-helix and physically close to residue 67, is considered an accessory mutation for resistance.

Disruptions of the hydrogen bond between chain B residues 8 and 10 can be modeled using mutations at residues 41, 46, 62 and 71. Recall that the polymorphic R41K mutation is present mainly in the susceptible variants in this study (Fig. S2), so that its coefficient

should be positive (black edge): this hydrogen bond occupancy increases among susceptible variants. Otherwise, however, the coefficients are negative (grey edges), including that for the A71V or A71I mutation.

Protein structure

Having identified resistance-related alterations among close-range inhibitor-enzyme contacts and more distal alterations in hydrogen bonding, one can also use the simulations to probe any corresponding structural changes that may mediate these effects. The distances between C_α atoms can be used to define protein structures: even a sparse set of C_α distances can be adequate for structural determination.[95]

Cataloging the distances between pairs of $N = 2 \times 99$ distinct C_α atoms in the HIV-1 protease dimer, there are $p = N(N-1)/2 = 19,503$ to consider, resulting in a $15 \times 19,503$ column centered matrix of mean distances, $\mathbf{D}^{\mathrm{d}_{\mathrm{C}_\alpha}}$. Applying SVD to this matrix yields a sharply peaked singular value spectrum (Fig. S5 (c)) with more than 42% of the variance accounted for by the first singular value and an additional 18% by the second singular value. Plotting the unscaled principal component scores, $\mathbf{u}_1$ and $\mathbf{u}_2$, against one another for these two singular values shows that the two wild type strains, NL4-3 and SF-2, are once more approximately co-localized (Fig. 7). Examining the principal component scores, only $\mathbf{u}_2$ offers a significant ($p < 0.006$, Fig. S6 (c)) discrimination between the susceptible and resistant variants. Considering the profound importance of hydrogen bonding for determining protein structure, it is unsurprising to observe that the same principal component score ($\mathbf{u}_2$) in the SVD for $\mathbf{D}^{\mathrm{HB}}$ and $\mathbf{D}^{\mathrm{d}_{\mathrm{C}_\alpha}}$, respectively, is most related to drug resistance.

Because of the extreme "large p, small n"[10] nature of $\mathbf{D}^{\mathrm{d}_{\mathrm{C}_\alpha}}$, a detailed accounting of each component of the $\mathbf{v}_2$ eigenfeature is not feasible. Nevertheless, examination of $\mathbf{v}_2$ reveals a relatively small subset of important components (Fig. 8, Fig. S13(a)). After mapping the components of $\mathbf{v}_2$ onto structural elements of the protease (Fig. 8), the most striking observation is that the β-sheets of the outer core (sometimes called the "cantilever",

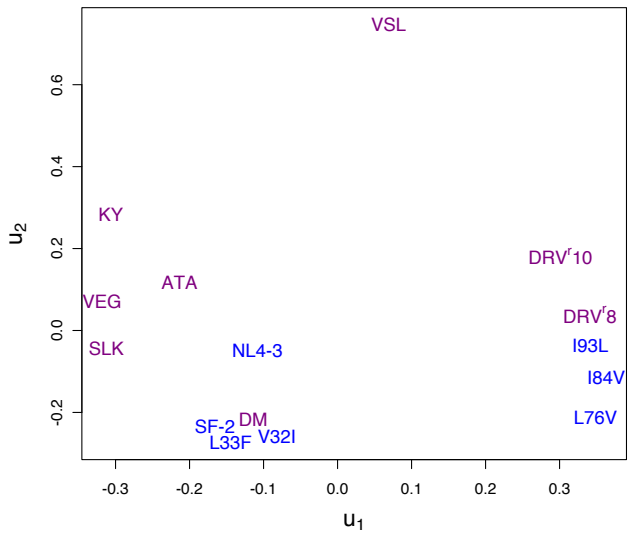

Figure 7: HIV-1 protease variants projected on top two principal component scores of the centered mean $C_\alpha - C_\alpha$ distances.

see Fig. S1) for each monomer are further apart from one another in the resistant versus the susceptible variants. This observation is also apparent from inspection just the $k = 75$ components with the largest absolute value, $|v_{2k}|$ (Fig. S13(b)). The inter-monomeric distances between these β-sheets, as measured using C_α atoms, are typically increased by about 1.5 among the resistant variants, where much greater variances in these distances are also observed. Exhibiting similar, but more modest increases among the resistant variants, are the intra-monomeric distances between this β-sheet and the terminal residues at the dimer interface.

Specific residues where mutations can predict $\mathbf{u}_2$ were identified by fitting a regression model (eq. 5). The sequence features that were selected under cross-validation are highly predictive (Pearson's $r = 0.98$, mean square error=0.04, see Fig. S8(c)) and include familiar resistance mutations to residues I50, K43, M46 and V82 along with known accessory mutations at A71 and I93 (Table S1). As was the case when the hydrogen bonding matrix, $\mathbf{D}^{\mathrm{HB}}$, was analyzed using SVD, the importance of the I50V mutation in modeling $\mathbf{u}_2$ here lies in accurately reproducing the high score of the VSL variant (Fig. 7). The I62V mutation that is included in the model (Table S1) is polymorphic, but is present only among resistant

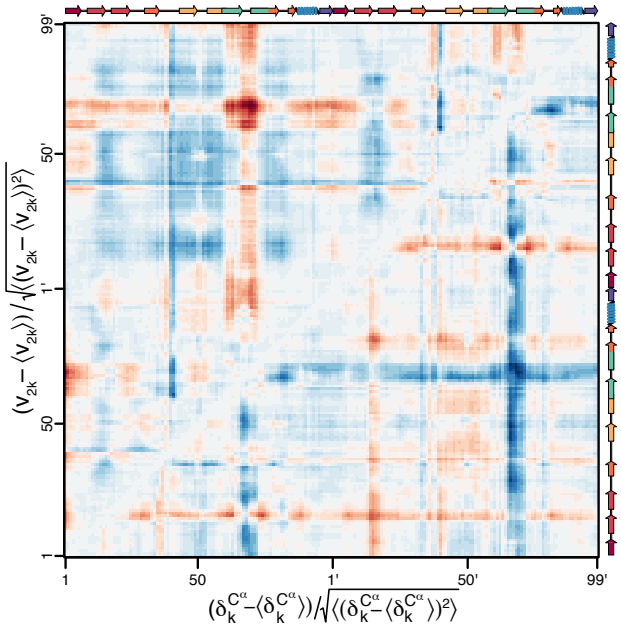

Figure 8: Resistance-associated variability in intra-protease $C_\alpha - C_\alpha$ distances reveals domain changes in geometry. For these data, the second singular value, s_2, best separates susceptible from resistant strains (see Figs. 7, S6(c)), so $\mathbf{v}_2$ is plotted in the upper triangle. For comparison, the difference in mean $C_\alpha - C_\alpha$ distance between susceptible and resistant strains, $\delta_k = \frac{1}{n_{\text{susc.}}} \sum^{n_{\text{susc.}}} {d_{C_\alpha}}^{(k)} - \frac{1}{n_{\text{res.}}} \sum^{n_{\text{res.}}} {d_{C_\alpha}}^{(k)}$, is shown in the lower triangle. For ease of visualization together in the same plot, both v_{2k} and δ_k have been standardized, with red indicating more positive and blue more negative values, respectively. Secondary structural elements and annotations are shown in the margins (see also Fig S1). There is a notable positive v_{2k} for the "cantilever" β-sheets between residues 59 and 75, corresponding to a more open protease structure (i.e. negative δ_k in the lower triangle) among the resistant variants.

protease variants studied here (Fig. S2). With one exception, the model coefficients are all positive, meaning that the predicted $\mathbf{u}_2$ increases under amino acid substitution. The single exception with a negative coefficient is the polymorphic K14R mutation that is prevalent among the susceptible variants.

As an alternative to the SVD-based approach above, directly selecting alterations of average distances between specific pairs of C_α atoms using penalized logistic regression identifies a sparse set of predictive distances. A compact set of four features is presented in Fig. 9(b), indicating a resistance-associated opening between the monomeric subunits and a corresponding compression of some structural elements within each monomer (Fig. 9(c)). Although the features selected under the elastic net penalty are sparse, it is clear from unsupervised learning that the structural alterations that they capture are concerted and involve each of the 70s β-sheets (i.e. cantilevers) as a whole (Figs. 8 and S13(b)). Apart from using SVD, an alternative way to detect the concerted nature of these alterations is to reduce the parameter, α, used in the elastic net penalty (eq. 6), thereby emphasizing the l_2-norm penalty (data not shown).

The mutations that were found to predict each of the selected distances between C_α atoms are indicated in Fig. 9(b). Among the list of these predictive mutations, most are familiar from analyzing the hydrogen bond occupancies, such as those at residues 10, 13, 46, 62, 71, 76, 82, 84 and 90. Residues 10 and 13 from this list flank residue 12 and were also implicated above in the resistance-associated disruption of the hydrogen bond formed between residues 12 and 67. Consistent with that observation, this same pair of mutations is found to predict an expansion between the two monomeric subunits of the enzyme (Figs. 9(b) and 9(c)).

With two exceptions, all of the mutations that were selected via separate regressions on each of the distance features in Fig. 9(a) correspond to concerted expansions and compressions of structural elements that were also characterized using SVD on $\mathbf{D}^{\mathrm{d}_{\mathrm{C}_\alpha}}$. That is, these resistance-associated structural changes occur when wild-type amino acids are substituted

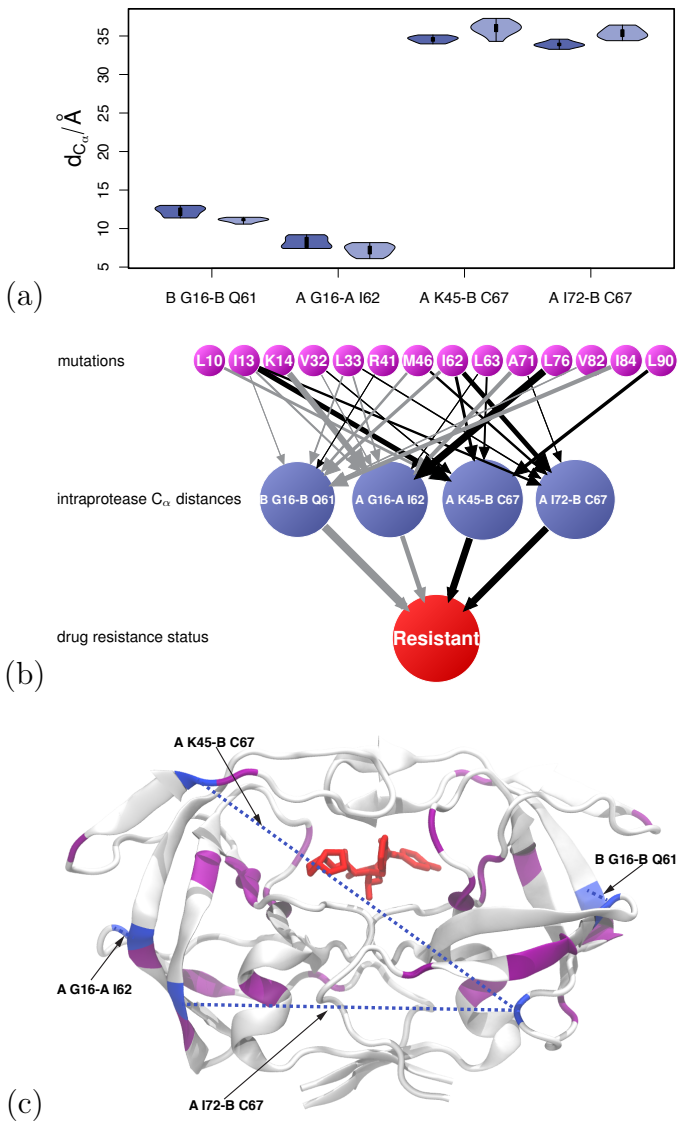

Figure 9: Visualization of the regularized regression model hierarchy used to to identify resistance-associated alterations in HIV-1 protease structure. For each predictor (a) in a logistic regression model of resistance based upon the mean distances between C_α atoms, a linear model of sequence predictors has been fitted (b). The violin plots in (a) depict the kernel density of mean occupancies for susceptible (left, darker blue) and resistant (right, lighter blue) protease variants. The directed edges shown in (b) correspond to fitted coefficients and are colored according to sign (black for positive, grey for negative) and sized according to importance. In (c) the physical and sequence features are projected onto the structure of the ligand-bound enzyme.

at the residues shown in Figs. 9(b) and 9(c). The exceptions to this trend occur at residues 41 and 76, where mutations are associated with the opposite structural alterations. The case of the polymorphic R41K mutation is straightforward to interpret and by now familiar

from the preceding analysis: it is mainly present among the susceptible strains. The case of the L76V mutation, however, is more interesting. In combination with other mutations, the L76V substitution confers resistance to darunavir[50,96–98] but hypersusceptibility to some other protease inhibitors.[96–98] Compensatory mutations are reported to include M46I, I54V, V82A, I84V and L90M.[97] Among the protease variants studied here (Fig. S2), L76V is present in only the DRV^{r}10 (resistant) and L76V (susceptible) variants. In each class of variants, susceptible and resistant, the strain with L76V present exhibits the largest distance between C_α atoms on residues 16 and 62 on chain A, perhaps because the smaller valine residue allows room for a less compressed contact.

The concerted structural alterations that delineate the susceptible and resistant variants in Table 1 are consistent with those anticipated from a previous examination of x-ray crystallographic structures of inhibitor- and product peptide-bound HIV-1 and SIV protease.[94] Based on this structural comparison, it was suggested that "domain orientation or movement may be a factor in the development of resistance" due to mutations that are distal from the active site of the HIV-1 protease but located at the interfaces of its structural domains.[94] Among the 15 protease variants studied here, such mutations include those at residues 10, 71, 89, 90 (Figs. S1, S2), physically located near the terminal dimer interface and at residues 20, 32, 33, 35, 45, 54, 63 and 77 (Figs. S1, S2). While some of these mutations are collinear with other mutations in the panel and therefore do not appear among the selected sequence features in Fig. 9(c), many of these "interfacial" mutations are present.

Discussion

Overall, applying SVD and regularized regression to data collected from atomistic simulations of many different HIV-1 protease sequence variants has provided useful insights into how physical alterations in the darunavir-protease complex control binding and lead to resistance. Moreover, relationships between resistance-related alterations in physical properties

and specific mutations were identified.

It has been a longstanding problem to characterize the mechanisms by which mutations throughout the HIV-1 protease lead to resistance, particularly for mutations that are located far from the active site of the enzyme. Using a combination of supervised and unsupervised machine learning techniques here has revealed that several such distal mutations, including the known resistance mutations at residues 10 and 11, appear to function by interfering with important hydrogen bonds within the enzyme, thereby causing broad structural changes that affect the short-range contacts with the inhibitor. Mutations elsewhere in the protease, like those at residue 46 in the flaps and residue 71 in the cantilever also contribute to these alterations, as do mutations that are more proximal to the active site, such as the primary resistance mutations at residues 32, 82 and 84.

These findings provide detailed support for the idea that alterations to the network of intra-protease hydrogen bonds are an important signature of drug resistance in HIV-1 protease.[52,53] Because much of the variability in this network among our panel appeared to be unrelated to drug resistance, however, the techniques presented here were particularly important for detecting such signatures. In other words, this analysis has emphasized the utility of filtering the noise in these data by applying either SVD or direct penalized regression.

Given a matrix of p thermodynamic observables extracted from n simulations, SVD offers a convenient route to dimensionality reduction, which can offer insight into how the observables (e.g. mean van der Waals interactions or hydrogen bonds etc.) relate to experimentally measured properties (e.g. resistance to inhibition) of the simulated systems. Examining the components of key singular vectors can characterize the relative importance of the p observables. Ideally, SVD can identify a coordinate that explains all differences among the measured properties. There is no assurance that this will be true, however, in which case supervised machine learning can provide additional insight.

Penalized regression was used here to sequentially apply supervised learning to simulation and sequence data, with the objective of extracting a compact set of readily interpretable

mutations and altered thermodynamic observables. To this end, the elastic net penalty was applied with the goal of finding the sparse solutions implied by emphasizing its l_1-norm penalty. Compared with SVD or other unsupervised learning techniques, this supervised approach has the advantage that the features selected by the procedure are, by design, able to distinguish the different phenotype classes. For example, the five inhibitor-protease van der Waals interactions that were selected can be used to accurately distinguish between HIV-1 protease variants that are susceptible and resistant to darunavir inhibition. If a broader set of features is desirable, to detect alterations throughout an element of protein secondary structure, for instance, the elastic net penalty can be parameterized to place a greater emphasis on its l_2-norm penalty.

While the results here demonstrate that changes among select thermodynamic observables collected in the bound ligand-inhibitor complex can predict resistance status and can be leveraged for useful insights into the resistance mechanism, the unbound states of the system(s) have not been explicitly examined. These unbound states are characterized by a separate protein and ligand, each in aqueous solution. For the present study, darunavir is common to all systems, so its properties in bulk solution cannot inform differences in resistance among the HIV-1 protease variants. It is possible, however, that unbound states of the protease variants affect resistance by, for example, altering the cross section for the ligand to encounter the active site.[99,100] A straightforward extension of the present methods would be to simulate such states and include their features among the predictors.

The methods presented here could also be readily extended to include finer details of the short-range inhibitor-protein interactions. This extension could be accomplished, for example, by separately considering m different molecular fragments, or moieties of the inhibitor. The data matrix $\mathbf{D}^{\mathrm{LJ}}$ would then include a longer list of $p' = m \cdot p$ predictors. Bearing in mind the limitations implied by the size of n relative to that of p', analyzing this $n \times p'$ matrix can offer more detailed insights into the development of resistance, thereby informing the design of improved inhibitors.

Conclusions

Overall, applying SVD and regularized regression to data collected from atomistic simulations of many different HIV-1 protease sequence variants has provided useful insights into how physical alterations in the darunavir-protease complex control binding and lead to resistance. Relationships between resistance-related alterations in physical properties and specific mutations were identified.

The methods presented here have been applied to study drug resistance in HIV-1 protease, yet these techniques are broadly applicable to data from a variety of atomistic simulations. For example, given structural information and measurements of binding affinity for a series of antibodies and a target protein (or possibly multiple targets), important interaction sites could be identified for refinement. Likewise, given structural and binding information, simulations for a series of protein-RNA/DNA complexes could be used to characterize the molecular details of specificity and carry out refinement.

Acknowledgement

T.W.W. is grateful to Oliver D. King and Francesca Massi for helpful discussionsat during the course of this study. This work was supported by grant P01GM109767 from the National Institutes of Health.

Supporting Information Available

The following files are available free of charge via the Internet at http://pubs.acs.org:

- supportingInformation.pdf: Supplementary equation, figures and table.

References

(1) Yun, C.-H.; Mengwasser, K. E.; Toms, A. V.; Woo, M. S.; Greulich, H.; Wong, K.-K.; Meyerson, M.; Eck, M. J. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 2070–2075.

(2) Park, J.; McDonald, J. J.; Petter, R. C.; Houk, K. N. Molecular Dynamics Analysis of Binding of Kinase Inhibitors to WT EGFR and the T790M Mutant. *J. Chem. Theory Comput.* **2016**, *12*, 2066–2078.

(3) Wang, W.; Kollman, P. A. Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14937–14942.

(4) Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C. A. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure* **2002**, *10*, 369–381.

(5) Altman, M. D.; Ali, A.; Reddy, G. S. K. K.; Nalam, M. N. L.; Anjum, S. G.; Cao, H.; Chellappan, S.; Kairys, V.; Fernandes, M. X.; Gilson, M. K.; Schiffer, C. A.; Rana, T. M.; Tidor, B. HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *J. Am. Chem. Soc.* **2008**, *130*, 6099–6113.

(6) Ali, A.; Bandaranayake, R. M.; Cai, Y.; King, N. M.; Kolli, M.; Mittal, S.; Murzycki, J. F.; Nalam, M. N. L.; Nalivaika, E. A.; Özen, A.; Prabu-Jeyabalan, M. M.; Thayer, K.; Schiffer, C. A. Molecular Basis for Drug Resistance in HIV-1 Protease. *Viruses* **2010**, *2*, 2509–2535.

(7) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* **1901**, *2*, 559–572.

(8) Golub, G. H.; Reinsch, C. Singular Value Decomposition and Least Squares Solutions. *Numer. Math.* **1970**, *14*, 403–420.

(9) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*; Cambridge University Press: New York, NY, USA, 2007.

(10) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer Series in Statistics; Springer New York Inc.: New York, NY, USA, 2016.

(11) Karplus, M.; Kushick, J. N. Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, *14*, 325–332.

(12) Levy, R. M.; Srinivasan, A. R.; Olson, W. K.; McCammon, J. A. Quasi-harmonic method for studying very low frequency modes in proteins. *Biopolymers* **1984**, *23*, 1099–1112.

(13) Teeter, M. M.; Case, D. A. Harmonic and quasiharmonic descriptions of crambin. *J. Phys. Chem.* **1990**, *94*, 8091–8097.

(14) Bahar, I.; Erman, B.; Haliloglu, T.; Jernigan, R. L. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry* **1997**, *36*, 13512–13523.

(15) García, A. E. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699.

(16) Romo, T. D.; Clarage, J. B.; Sorensen, D. C.; Phillips, G. N. Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. *Proteins* **1995**, *22*, 311–321.

(17) Andrews, B. K.; Romo, T.; Clarage, J. B.; Pettitt, B. M.; Phillips, G. N. Characterizing global substates of myoglobin. *Structure* **1998**, *6*, 587–594.

(18) Lange, O. F.; Grubmüller, H. Can Principal Components Yield a Dimension Reduced Description of Protein Dynamics on Long Time Scales? *J. Phys. Chem. B* **2006**, *110*, 22842–22852.

(19) Lange, O. F.; Grubmüller, H. Collective Langevin dynamics of conformational motions in proteins. *J. Chem. Phys.* **2006**, *124*, 214903.

(20) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **2018**, *148*, 241703.

(21) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5.

(22) Sultan, M. M.; Wayment-Steele, H. K.; Pande, V. S. Transferable Neural Networks for Enhanced Sampling of Protein Dynamics. *J. Chem. Theory Comput.* **2018**, *14*, 1887–1894.

(23) Hernández, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational encoding of complex dynamics. *Phys. Rev. E* **2018**, *97*, 062412.

(24) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(25) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.

(26) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **2017**, *8*, 872.

(27) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(28) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.

(29) Schneider, E.; Dai, L.; Topper, R. Q.; Drechsel-Grau, C.; Tuckerman, M. E. Stochastic Neural Network Approach for Learning High-Dimensional Free Energy Surfaces. *Phys. Rev. Lett.* **2017**, *119*, 150601.

(30) Sidky, H.; Whitmer, J. K. Learning free energy landscapes using artificial neural networks. *J. Chem. Phys.* **2018**, *148*, 104111.

(31) Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A. E.; Lokhov, A.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks. *J. Chem. Theory Comput.* **2018**, *14*, 4687–4698.

(32) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(33) Dolgirev, P. E.; Kruglov, I. A.; Oganov, A. R. Machine learning scheme for fast extraction of chemically interpretable interatomic potentials. *AIP Adv.* **2016**, *6*, 085318.

(34) Yao, K.; Herr, J. E.; Parkhill, J. The many-body expansion combined with neural networks. *J. Chem. Phys.* **2017**, *146*, 014106.

(35) Galvelis, R.; Sugita, Y. Neural Network and Nearest Neighbor Algorithms for Enhancing Sampling of Molecular Dynamics. *J. Chem. Theory Comput.* **2017**, *13*, 2489–2500.

(36) Ballester, P. J.; Mitchell, J. B. O. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169–1175.

(37) Durrant, J. D.; McCammon, J. A. NNScore: a neural-network-based scoring function for the characterization of protein-ligand complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1865–1871.

(38) Ding, B.; Wang, J.; Li, N.; Wang, W. Characterization of Small Molecule Binding. I. Accurate Identification of Strong Inhibitors in Virtual Screening. *J. Chem. Inf. Model.* **2012**, *53*, 114–122.

(39) Li, G.-B.; Yang, L.-L.; Wang, W.-J.; Li, L.-L.; Yang, S.-Y. ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions. *J. Chem. Inf. Model.* **2013**, *53*, 592–600.

(40) Wang, Y.; Guo, Y.; Kuang, Q.; Pu, X.; Ji, Y.; Zhang, Z.; Li, M. A comparative study of family-specific protein-ligand complex affinity prediction based on random forest approach. *J. Comput. Aided Mol. Des.* **2015**, *29*, 349–360.

(41) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2015**, *5*, 405–424.

(42) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296.

(43) Ghosh, A. K.; Kincaid, J. F.; Cho, W.; Walters, D. E.; Krishnan, K.; Hussain, K. A.; Koo, Y.; Cho, H.; Rudall, C.; Holland, L.; Buthod, J. Potent HIV protease inhibitors incorporating high-affinity P2-ligands and (R)-(hydroxyethylamino)sulfonamide isostere. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 687–690.

(44) De Meyer, S.; Azijn, H.; Surleraux, D.; Jochmans, D.; Tahri, A.; Pauwels, R.; Wigerinck, P.; de Béthune, M.-P. TMC114, a novel human immunodeficiency virus type 1 protease inhibitor active against protease inhibitor-resistant viruses, including a broad range of clinical isolates. *Antimicrob. Agents Chemother.* **2005**, *49*, 2314–2321.

(45) Navia, M. A.; Fitzgerald, P. M.; McKeever, B. M.; Leu, C. T.; Heimbach, J. C.; Herber, W. K.; Sigal, I. S.; Darke, P. L.; Springer, J. P. Three-dimensional structure

of aspartyl protease from human immunodeficiency virus HIV-1. *Nature* **1989**, *337*, 615–620.

(46) Craig, J. C.; Duncan, I. B.; Hockley, D.; Grief, C.; Roberts, N. A.; Mills, J. S. Antiviral properties of Ro 31-8959, an inhibitor of human immunodeficiency virus (HIV) proteinase. *Antivir. Res.* **1991**, *16*, 295–305.

(47) King, N. M.; Prabu-Jeyabalan, M.; Nalivaika, E. A.; Wigerinck, P.; de Béthune, M.-P.; Schiffer, C. A. Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *J. Virol.* **2004**, *78*, 12012–12021.

(48) McCoy, C. Darunavir: a nonpeptidic antiretroviral protease inhibitor. *Clin. Ther.* **2007**, *29*, 1559–1576.

(49) Rhee, S. Y.; Taylor, J.; Fessel, W. J.; Kaufman, D.; Towner, W.; Troia, P.; Ruane, P.; Hellinger, J.; Shirvani, V.; Zolopa, A.; Shafer, R. W. HIV-1 Protease Mutations and Protease Inhibitor Cross-Resistance. *Antimicrob. Agents Chemother.* **2010**, *54*, 4253–4261.

(50) Varghese, V.; Mitsuya, Y.; Fessel, W. J.; Liu, T. F.; Melikian, G. L.; Katzenstein, D. A.; Schiffer, C. A.; Holmes, S. P.; Shafer, R. W. Prototypical Recombinant Multi-Protease-Inhibitor-Resistant Infectious Molecular Clones of Human Immunodeficiency Virus Type 1. *Antimicrob. Agents Chemother.* **2013**, *57*, 4290–4299.

(51) Wu, T. D.; Schiffer, C. A.; Gonzales, M. J.; Taylor, J.; Kantor, R.; Chou, S.; Israelski, D.; Zolopa, A. R.; Fessel, W. J.; Shafer, R. W. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *J. Virol.* **2003**, *77*, 4836–4847.

(52) Ragland, D. A.; Whitfield, T. W.; Lee, S.-K.; Swanstrom, R.; Zeldovich, K. B.;

Kurt Yilmaz, N.; Schiffer, C. A. Elucidating the Interdependence of Drug Resistance from Combinations of Mutations. *J. Chem. Theory Comput.* **2017**, *13*, 5671–5682.

(53) Ragland, D. A.; Nalivaika, E. A.; Nalam, M. N. L.; Prachanronarong, K. L.; Cao, H.; Bandaranayake, R. M.; Cai, Y.; Kurt Yilmaz, N.; Schiffer, C. A. Drug resistance conferred by mutations outside the active site through alterations in the dynamic and structural ensemble of HIV-1 protease. *J. Am. Chem. Soc.* **2014**, *136*, 11956–11963.

(54) Adachi, A.; Gendelman, H. E.; Koenig, S.; Folks, T.; Willey, R.; Rabson, A.; Martin, M. A. Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *J. Virol.* **1986**, *59*, 284–291.

(55) Rosé, J. R.; Salto, R.; Craik, C. S. Regulation of autoproteolysis of the HIV-1 and HIV-2 proteases with engineered amino acid substitutions. *J. Biol. Chem.* **1993**, *268*, 11939–11945.

(56) Rutenber, E.; Fauman, E. B.; Keenan, R. J.; Fong, S.; Furth, P. S.; Ortiz de Montellano, P. R.; Meng, E.; Kuntz, I. D.; DeCamp, D. L.; Salto, R.; Rose, J. R.; Craik, C. S.; Stroud, R. M. Structure of a non-peptide inhibitor complexed with HIV-1 protease. Developing a cycle of structure-based drug design. *J. Biol. Chem.* **1993**, *268*, 15343–15346.

(57) Sanchez-Pescador, R.; Power, M. D.; Barr, P. J.; Steimer, K. S.; Stempien, M. M.; Brown-Shimer, S. L.; Gee, W. W.; Renard, A.; Randolph, A.; Levy, J. A. Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). *Science* **1985**, *227*, 484–492.

(58) Rhee, S.-Y.; Gonzales, M. J.; Kantor, R.; Betts, B. J.; Ravela, J.; Shafer, R. W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **2003**, *31*, 298–303.

(59) Shafer, R. W. Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.* **2006**, *194 Suppl 1*, S51–8.

(60) Cheng, Y.; Prusoff, W. H. Relationship between the inhibition constant (Ki) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099–3108.

(61) Shelley, J. C.; Cholleti, A.; Frye, L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pKa prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des.* **2007**, *21*, 681–691.

(62) Greenwood, J. R.; Sullivan, D. C. A. P.; Shelley, J. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput. Aided Mol. Des.* **2010**, *24*, 591–604.

(63) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and interpretation of protein pKa value. *Proteins* **2005**, *61*, 704–721.

(64) Adachi, M. et al. Structure of HIV-1 protease in complex with potent inhibitor KNI-272 determined by high-resolution X-ray and neutron crystallography. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 4641–4646.

(65) Banks, J. L. et al. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **2005**, *26*, 1752–1780.

(66) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(67) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577.

(68) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. Proceedings of the 2006 ACM/IEEE Conference on Supercomputing. New York, NY, USA, 2006.

(69) Martyna, G.; Tobias, D.; Klein, M. Constant-pressure molecular-dynamics algorithms. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

(70) Tuckerman, M.; Berne, B. J.; Martyna, G. J. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.* **1992**, *97*, 1990.

(71) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(72) Golyandina, N.; Nekrutkin, V.; Zhigljavsky, A. A. *Analysis of Time Series Structure: SSA and Related Techniques*; Monographs on Statistics and Applied Probability; Chapman & Hall/CRC Press: Boca Raton, FL, USA, 2001; Vol. 90.

(73) Wall, M. E.; Rechtsteiner, A.; Rocha, L. M. In *A Practical Approach to Microarray Data Analysis*; Berrar, D. P., Dubitzky, W., Granzow, M., Eds.; Kluwer: Norwell, MA, 2003; Chapter 5, pp 91–109.

(74) Zhang, L.; Marron, J. S.; Shen, H.; Zhu, Z. Singular Value Decomposition and Its Visualization. *J. Comput. Graph. Stat.* **2007**, *16*, 833–854.

(75) Wetzel, S. J. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Phys. Rev. E* **2017**, *96*, 022140.

(76) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J. Chem. Phys.* **2018**, *149*, 072301.

(77) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.

(78) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

(79) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A. J.; Vapnik, V. In *Advances in Neural Information Processing Systems 9*; Mozer, M. C., Jordan, M. I., Petsche, T., Eds.; MIT Press, 1997; pp 155–161.

(80) Ho, T. K. Random Decision Forests. Proceedings of the Third International Conference on Document Analysis and Recognition. Washington, DC, USA, 1995; pp 278–282.

(81) Alter, O.; Brown, P. O.; Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 10101–10106.

(82) Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal. Stat. Soc. B* **1996**, *58*, 267–288.

(83) Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; Monographs on Statistics and Applied Probability; Chapman & Hall/CRC Press: Boca Raton, FL, USA, 2015; Vol. 143.

(84) Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22.

(85) Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. Royal. Stat. Soc. B* **2005**, *67*, 301–320.

(86) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67.

(87) Meinshausen, N. Relaxed Lasso. *Comput. Stat. Data Anal.* **2007**, *52*, 374–393.

(88) Cai, Y.; Schiffer, C. A. Decomposing the Energetic Impact of Drug Resistant Mutations in HIV-1 Protease on Binding DRV. *J. Chem. Theory Comput.* **2010**, *6*, 1358–1368.

(89) Shen, Y.; Radhakrishnan, M. L.; Tidor, B. Molecular mechanisms and design principles for promiscuous inhibitors to avoid drug resistance: lessons learned from HIV-1 protease inhibition. *Proteins* **2015**, *83*, 351–372.

(90) Clemente, J. C.; Hemrajani, R.; Blum, L. E.; Goodenow, M. M.; Dunn, B. M. Secondary Mutations M36I and A71V in the Human Immunodeficiency Virus Type 1 Protease Can Provide an Advantage for the Emergence of the Primary Mutation D30N. *Biochemistry* **2003**, *42*, 15029–15035.

(91) Mittal, S.; Bandaranayake, R. M.; King, N. M.; Prabu-Jeyabalan, M.; Nalam, M. N. L.; Nalivaika, E. A.; Yilmaz, N. K.; Schiffer, C. A. Structural and thermodynamic basis of amprenavir/darunavir and atazanavir resistance in HIV-1 protease with mutations at residue 50. *J. Virol.* **2013**, *87*, 4176–4184.

(92) Creighton, T. E. *Proteins: structures and molecular properties*, 2nd ed.; W.H. Freeman and Company: New York, NY, USA, 1993.

(93) Harte, W. E.; Swaminathan, S.; Beveridge, D. L. Molecular dynamics of HIV-1 protease. *Proteins* **1992**, *13*, 175–194.

(94) Rose, R. B.; Craik, C. S.; Stroud, R. M. Domain flexibility in retroviral proteases: structural implications for drug resistant mutations. *Biochemistry* **1998**, *37*, 2607–2621.

(95) Crecca, C. R.; Roitberg, A. E. Using distances between α-carbons to predict protein structure. *Int. J. Quantum Chem.* **2008**, *108*, 2782–2792.

(96) Tartaglia, A.; Saracino, A.; Monno, L.; Tinelli, C.; Angarano, G. Both a Protective and a Deleterious Role for the L76V Mutation. *Antimicrob. Agents Chemother.* **2009**, *53*, 1724–1725.

(97) Young, T. P.; Parkin, N. T.; Stawiski, E.; Pilot-Matias, T.; Trinh, R.; Kempf, D. J.; Norton, M. Prevalence, Mutation Patterns, and Effects on Protease Inhibitor Susceptibility of the L76V Mutation in HIV-1 Protease. *Antimicrob. Agents Chemother.* **2010**, *54*, 4903–4906.

(98) Wiesmann, F.; Vachta, J.; Ehret, R.; Walter, H.; Kaiser, R.; Stürmer, M.; Tappe, A.; Däumer, M.; Berg, T.; Naeth, G.; Braun, P.; Knechten, H. The L76V mutation in HIV-1 protease is potentially associated with hypersusceptibility to protease inhibitors Atazanavir and Saquinavir: is there a clinical advantage? *AIDS Res. Ther.* **2011**, *8*, 7.

(99) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How does a drug molecule find its target binding site? *J. Am. Chem. Soc.* **2011**, *133*, 9181–9183.

(100) Limongelli, V.; Bonomi, M.; Parrinello, M. Funnel metadynamics as accurate binding free-energy method. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 6358–6363.

Graphical TOC Entry