A multiscale approach for computing gated ligand binding from molecular dynamics and Brownian dynamics simulations

S. Kashif Sadiq,[*,†,‡,¶] Abraham Muñiz Chicharro,[†,§] Patrick Friedrich,[†] and Rebecca C. Wade[†,‖,⊥]

†*Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany*

‡*Genome Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany*

¶*Infection Biology Unit, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C/Doctor Aiguader 88, 08003 Barcelona, Spain*

§*Faculty of Biosciences, Heidelberg University, Im Neuenheimer Feld 234, 69120 Heidelberg, Germany*

‖*Center for Molecular Biology (ZMBH), DKFZ-ZMBH Alliance, Heidelberg University, Im Neuenheimer Feld 282, 69120 Heidelberg, Germany*

⊥*Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany*

E-mail: kashif.sadiq@embl.de

Abstract

We develop an approach to characterise the effects of gating by a multi-conformation protein consisting of macrostate conformations that are either accessible or inaccessible

to ligand binding. We first construct a Markov state model of the apo-protein from atomistic molecular dynamics simulations from which we identify macrostates and their conformations, compute their relative macrostate populations and interchange kinetics, and structurally characterise them in terms of ligand accessibility. We insert the calculated first-order rate constants for conformational transitions into a multi-state gating theory from which we derive a gating factor γ that quantifies the degree of conformational gating. Applied to HIV-1 protease, our approach yields a kinetic network of three accessible (semi-open, open and wide-open) and two inaccessible (closed and a newly identified, 'parted') macrostate conformations. The 'parted' conformation sterically partitions the active site, suggesting a possible role in product release. We find that the binding kinetics of drugs and drug-like inhibitors to HIV-1 protease falls in the slow gating regime. However, because γ=0.75, conformational gating only modestly slows ligand binding. Brownian dynamics simulations of the diffusional association of eight inhibitors to the protease - that have a wide range of experimental association constants ($\sim 10^4$ - 10^{10} $M^{-1}s^{-1}$) - yields gated rate constants in the range $\sim$0.5-5.7 $\times$ 10^8 $M^{-1}s^{-1}$. This indicates that, whereas the association rate of some inhibitors could be described by the model, for many inhibitors either subsequent conformational transitions or alternate binding mechanisms may be rate-limiting. For systems known to be modulated by conformational gating, the approach could be scaled computationally efficiently to screen association kinetics for a large number of ligands.

1 Introduction

Molecular association is central to virtually all biomolecular processes from regulation of metabolic networks and signal transduction to macromolecular assembly and drug efficacy in pharmacology. Extensive effort over several decades has led to the development of many computational methods to characterize molecular binding affinities.[1] Increasingly, though, the role of binding kinetics, not just affinity, is becoming more widely appreciated in determining the behaviour of biomolecular interaction networks and especially *in vivo* drug

efficacy.[2,3]

The binding of two molecules - such as a ligand to a protein - first requires them to encounter each other through a diffusional process, imposing an upper, diffusion-controlled limit - on the association rate.[4–6] Factors such as steric specificity, hydrodynamic interactions and binding site accessibility can all cause reductions from this limit, whilst ligand-protein interaction energies can either enhance or further impede the association rate - for example, through electrostatic steering.[5,7–12]

Protein conformational changes can also modulate binding. A canonical view of these changes usually depicts two primary protein conformations in thermodynamic equilibrium, active and inactive, where the active conformation corresponds to the native ligand-bound state.[13] Both conformations are accessible to ligand binding. In a conformational selection (CS) mechanism, a ligand selects to bind only to the active conformation within the protein population ensemble. Induced fit (IF) is the opposite extreme, where the ligand first binds to the inactive conformation and then the protein is induced to transition into the active conformation by forming the required intermolecular contacts consistent with the native ligand-bound state. Both mechanisms have been shown to play a role in binding processes and their association kinetics.[14–16] In general, however, both mechanisms can contribute to a given binding process.[13,14] Indeed, the relative contribution of IF versus CS increases with both ligand concentration[17,18] and the rate of protein conformational transition.[18,19] The existence of more molecular conformations can be described by a dock and coalesce (DC) mechanism[20,21] in which simultaneous formation of all required native bound interactions upon or after initial protein-ligand binding is unlikely,[13] thus precluding either purely CS or IF binding alone. This mechanism has been applied to the binding of highly disordered systems such as intrinsically disordered proteins (IDPs).[22–24]

Dynamic conformational changes in the protein can also alter binding site accessibility, introducing a gating step in the initial ligand binding process.[25–27] Gating theory for the interchange between two conformations, open and closed,[27–31] suggests the timescale of in-

terchange (τ_c) can play a significant role that depends on whether it is faster ($\tau_c << \tau_d$) or slower ($\tau_c >> \tau_d$) than the characteristic timescale of protein-ligand diffusional encounter (τ_d), or in between. For fast gating, the ligand effectively always encounters the protein in an accessible (open) state - thus the gated (k^g) and ungated (k^{ug}) association rate constants are equal. For slow gating, $k^g = \gamma k^{ug}$, where γ, the gating factor, represents the probability of finding the protein in a ligand-accessible state.

Brownian dynamics (BD) simulation methods offer a computationally inexpensive route for calculating ungated diffusional association kinetics,[32] in part due to frequent use of rigid bodies, simplified forcefields and implicit solvent models. They are well-established for protein-protein association rate constants[33,34] and applied to protein-ligand systems.[35,36] However, use of rigid structures poses challenges, especially becuase many biomolecular systems exhibit considerable flexibility and where CS, IF or DC mechanisms can be significant contributors to the binding kinetics. Rigid-body BD alone may therefore be applicable to mainly diffusional-limited binders. Multiscale methods that combine BD with molecular dynamics (MD) simulation methods may offer a possible solution to such limitations.[35] A method, known as SEEKR, has been proposed where BD simulations are performed to reach the first milestone, followed by spawning of MD simulations at subsequent milestones.[37–40] Such methods have yet to be tested on protein-ligand systems with large conformational flexibility.

Furthermore, a plethora of methods have emerged that aim to compute ligand-receptor binding kinetics.[41,42] Approaches based on molecular dynamics (MD) simulations have become a focal point in this development process.[43] Unbiased high-throughput all-atom MD simulations coupled to Markov state models (MSMs) have proven a rigorous route to extract the kinetics and thermodynamics of biomolecular processes, as well as a detailed description of their respective pathways, including phenomena such as receptor-ligand,[15,44–47] and receptor-receptor binding,[48] protein-folding,[49–52] conformational changes[46,53–56] and macromolecular assembly.[57]

MSMs usually involve the projection of the high-dimensional conformational space of atomistic features into a lower-dimensional manifold of collective variables (CVs), followed by discretization of the resulting space into non-overlapping microstates. Assignment of the underlying MD simulation ensemble in terms of a sequence of these microstates allows inter-state transition probabilities to be calculated at time intervals chosen to ensure a memoryless jump process, the condition of 'Markovianity'.[58] The resulting transition matrix, together with methods for further coarse-graining the MSM can then be used to probabilistically group together microstates into a smaller set of macrostates, that represent kinetically distinct interpretable conformations of the system as well as the corresponding coarse-grained equilibrium distribution and transition kinetics.[59]

Advantageously, MSMs can account for detailed changes across complex degrees of freedom, such as ligand-receptor binding pathways and therefore can quantify the degree of CS, IF or DC involved.[48] Disadvantageously, they impose a huge computational demand in order to resolve such kinetics accurately. Estimation of association rate constants may require simulating up to $\sim 10^3$-fold longer than the timescale of the binding process[3] due to the dilute solute concentration in a typical MD simulation box. Furthermore, the choice of initial features, the process of dimensionality reduction to obtain CVs and clustering methods for discretisation can all affect the quality of the model.[60] Significant developments have been made in each of these aspects over the last decade,[61] including the derivation of a variational principle[62] that enables selection of CVs to be systematized based on maximizing the decorrelation time of linear combinations of atomistic metrics, an approach known as time-independent component analysis, tICA.[63–65] Furthermore, adaptive sampling techniques[66–68] that explore the conformational space more efficiently and neural network approaches that learn optimal CVs[69] can result in orders-of-magnitude reductions in required sampling.

Despite these major developments, application of MSMs remains prohibitive when scaled to multiple numbers of receptor-ligand systems, especially when typical experimentally measured associaton rate constants (k_{on}^{exp}) for many clinically relevant compounds are $k_{on}^{exp} \sim 10^6$

$M^{-1}s^{-1}$ or slower.[3] The high computational demand of unbiased MD has motivated a number of enhanced sampling methods that introduce biasing potentials and/or biased exploration, allowing faster exploration of rarely accessible regions of the conformational space. Recent developments in methods such as metadynamics,[70,71] bias-exchanged metadynamics,[72] Gaussian-accelerated molecular dynamics (GAMD),[73] weighted ensemble sampling,[74] scaled MD[75] and adaptive multilevel splitting[76] have been applied to rate constant calculations. Unfortunately, such methods either require predefined CVs - which are not always known - or the use of biasing potentials makes recovering the original kinetics challenging.[77]

For the class of proteins that exhibit multiple conformational states that may gate ligand binding, we seek here to combine the advantages of MSMs for rigorously computing conformational kinetics with the advantages of BD approaches for rapidly computing ungated rate constants. HIV-1 protease is a suitably flexible, multi-conformation protein on which to develop such an approach and where the relative population, kinetic interchange and functional relevance of different conformations still remain incompletely understood as does the role of such conformations in gating ligand binding. The protease consists of a C_2-symmetric homodimer with a pair of flexible β-hairpin structures, termed 'flaps' that gate access to an active site[78] that structurally recognises[79,80] and cleaves a number of sequence specific peptidic junctions in viral Gag and GagPol polyproteins.[81–83] Most crystal structures of HIV-1 protease are of ligand-bound complexes - such structures are invariably characterised by a 'closed' flap conformation[84,85] where the flaps extend beyond but curl towards each other (Figure 1). A handful of apo-protein conformations exist. These include the well-characterised semi-open conformation[86] differentiated by a reversal of flap handedness with respect to the symmetry axis of the dimer - with flaps extending beyond but curling away from each other (Figure 1). A spectrum of more open flap structures[87] exist, marked by an absence of interflap interactions as well as the characteristic closed conformation. Structurally asymmetric conformations and flap dynamics have also been reported.[88]

Natural substrate recognition, transition state stabilisation and peptide cleavage occur

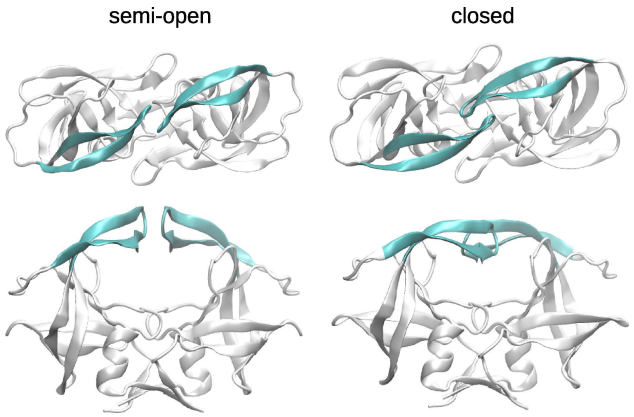


Figure 1: Characteristic conformations of the HIV-1 protease, illustrating a semi-open apo-state (PDB:1HHP)[86] and a closed state (PDB: 1KJ7),[80] required for both substrate cleavage and observed for inhibitor binding. For clarity, the substrate is omitted from the closed state structure depicted here. The homodimeric protein is shown in ribbon representation with the gating flaps in cyan.

in the closed conformation.[79] Solution NMR experiments on the apo protein have reported both a fast (<10 ns) and slow ($\sim$100μs) flap transition suggested to be flap tip curling and wide opening, respectively, as well as a dynamic equilibrium between three states of which the semi-open state is the dominant population.[89,90] More recent NMR studies are inconsistent with this view, reporting a mixture of states where the closed conformation is favoured.[91] Although lateral binding or unbinding of inhibitors, small peptide substrates or auto-associating peptides to either the closed or semi-open conformations may be possible,[53,72,92,93] steric considerations require an open state to admit peptidic cleavage sites within Gag or GagPol polyproteins or natural subtrate analogs. MD simulations have also reversibly captured a mixture of closed, semi-open, open and wide-open states in the apo protein - open, characterised by the flaps curling back from semi-open to expose the active site and wide-open, by a significantly larger flap-tip separation[27,94–96] - with a range of estimated timescales for flap opening ($\sim$10 ns), wide-opening ($\sim$500 ns) and closing ($\sim$50 ns) from a semi-open state.

BD simulations in a coarse-grained model of a two state open-closed system suggested a slow gating regime and showed good agreement with experimental rate constants for a

number of inhibitors in the 10^5-10^6 $M^{-1}s^{-1}$ range.[27] However, surface plasmon resonance (SPR) measurements association rate constants for a large array of inhibitors of HIV-1 protease, including the 9 FDA-approved drugs in clinical use,[97] showed k_{on}^{exp} values ranging widely from 10^2-10^{10} $M^{-1}s^{-1}$. This suggests that a significant degree of CS, IF or mixed DC meachnism may be at work during ligand binding and possibly in a ligand-dependent way. Furthermore, as more than two states exist and the open state is an intermediate state, then gated binding to an open state would still only be the first step of the process. A second step, requiring some degree of structural rearrangement is subsequently necessary to arrive at the final bound closed state. Simulations have suggested this second step may be rate limiting[98] and support the idea that different ligands may vary in their preferred binding mechanism.[99,100]

Here, using HIV-1 protease as a model system, we report a novel method that combines BD and MSMs built from all-atom MD simulations to calculate the extent of conformation gated ligand binding. This is done by combining the two approaches within a theoretical multi-state gating framework. We build on previous gating theory[25–27] to determine a gating factor in terms of the accessible and inaccessible ensemble of conformational macrostates of a MSM. Next we reconstruct the full apo-protein conformational kinetics of HIV-1 PR from a MSM based on all-atom MD simulations, identify accessible and inaccessible states - including a previously unidentified state - from which we obtain the resulting gating factor. Finally, we compute the ungated association kinetics for several ligands binding to each of the set of determined accessible states using BD simulations, applying the theoretical framework to arrive at final gated rate constants. This results in fast and scalable calculations of absolute values of protein-ligand biomolecular association rate constants that approach experimental values, where gating is the dominant factor in limiting binding and allows us, in some cases, to dissect various binding mechanisms.

Theory

We build on the theoretical formalism of MSMs[101] and gating theory,[25–27] in order to derive a term for the required gating factor to modulate ligand binding association kinetics in terms of the transition kinetics between conformational macrostates exhibited by the unbound target receptor.

Let us consider a Markov state model (MSM) of a stochastic process $\{X_n\}$ which is sufficiently described by a transition matrix $\mathbf{T}$. The elements of the matrix are the conditional probabilities

$$\mathrm{P}_{ij}(\tau) = \mathrm{P}(X_{t_0+\tau} = j | X_{t_0} = i), \tag{1}$$

describing the probability of a transition from a current state i to the next state j within the lag time τ. A MSM fulfilling the Chapman-Kolomogrov equation $\mathbf{T}(n \cdot \tau) = \mathbf{T}(\tau)^n$, in other words a vector holding the description of the state distribution at any time $t = t_0 + n\tau$, can be calculated from an initial distribution $\mathbf{p}(t_0)$ by

$$\mathbf{p}(t)^{\mathrm{T}} = \mathbf{p}(t_0)^{\mathrm{T}} \mathbf{T}(\tau)^n. \tag{2}$$

After sufficient time the probability distribution approaches the stationary distribution $\pi = \mathbf{p}(t \to \infty)$, which contains the probabilities of the single states in equilibrium. Further application of the transition matrix on π has no effect

$$\pi^{\mathrm{T}} = \pi^{\mathrm{T}} \mathbf{T}(\tau), \tag{3}$$

which forms an eigenvector equation. When such a stationary distribution exists a property, the detailed balance, can be derived from Eq.3 and 1.

$$\pi_i \cdot \mathrm{P}_{ij} = \pi_j \cdot \mathrm{P}_{ji}. \tag{4}$$

This property is physically important and describes the reversibility of the MSM.

The MSM transition matrix is estimated from the observed/counted transitions between the states in the single MD simulations. An important parameter for the accuracy of the estimation is the lag time τ. Ideally, the eigenvalues λ_i of the MSM should be independent of τ. Therefore τ should be chosen such that the implied relaxation time scales are approximately constant

$$t_i(\tau) = -\frac{\tau}{\ln|\lambda_i(\tau)|}. \tag{5}$$

To obtain those t_i for a MSM, the lag time might easily exceed the given time scales and therefore our interest lies only in the largest u eigenvalues $\lambda_i(\tau)$, which represent the slowest state transitions. For projected MSMs built on the reduced set of slow-transition eigenvalues, Eq.2 does not in general hold anymore. Rather, in order to coarse-grain an MSM into a meaningful and smaller set of macrostates, a hidden markov model (HMM) can be built to describe u such states at any time step $k\tau$.

The HMM can then be expressed as:

$$\mathbf{P}(k\tau) = \mathbf{\Pi}^{-1}\chi^{\mathrm{T}}\tilde{\mathbf{\Pi}}\tilde{\mathbf{P}}^k(\tau)\chi, \tag{6}$$

where $\tilde{\mathbf{P}} \in \mathbb{R}^{u\times u}$ is a transition matrix of u kinetic important coarse-grained states and $\tilde{\mathbf{\Pi}} = \mathrm{diag}(\tilde{\pi}_1, ..., \tilde{\pi}_u)$ contains the corresponding stationary distribution of the coarse-grained u-state HMM. χ is a matrix, whose elements $\chi_{Ij} \in \mathbb{R}^{u\times v}$ hold the probability to observe a macrostate I, while the system is in microstate j. $\mathbf{\Pi} = \mathrm{diag}(\pi_1, ..., \pi_v)$ is the stationary distribution of the v-state MSM.

To calculate a coarse-grained matrix, the microstates are grouped into a membership matrix $\mathbf{B} \in \mathbb{R}^{v\times u}$, the elements of which contain the probability that a microstate i participates in a macrostate J and are computed by linear transformation of the first u eigenvectors λ_i

(PCCA+ and PCCA++ methods). χ is then calculated as

$$\chi = \tilde{\mathbf{\Pi}}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{\Pi}. \tag{7}$$

Also the $u \times u$ transition matrix of Eq.6 can be calculated from the membership matrix $\mathbf{B}$ as

$$\tilde{\mathbf{P}} = (\mathbf{B}^{\mathrm{T}}\mathbf{B})\mathbf{B}^{\mathrm{T}}\mathbf{P}\mathbf{B}, \tag{8}$$

where $\mathbf{P}$ is the $v \times v$ MSM transition matrix. Together with Eq.7, Eq.8 is used to estimate the HMM.

A transition rate matrix $\mathbf{K}$ can then be derived, which describes the state transitions in terms of the actual kinetics. The relation between the rate and probability matrix is given by the equation

$$\tilde{\mathbf{P}}(\tau) = \exp(\tau\mathbf{K}), \tag{9}$$

where τ is here again the lag time at which the given HMM is constructed.

Let us now consider a multistate receptor that permits significant ligand access to only a single accessible state r, as described previously.[27] In a regime where ligand binding is limited by conformational gating to r and not by a subsequent induced fit association rate constant (k^{if}), the overall bimolecular association rate constant (k_{on}) is dominated by the gated association rate constant (k^{g}) and can be expressed as:

$$\mathrm{k}_{on} \sim \mathrm{k}^{g} = \gamma \mathrm{k}^{ug}, \tag{10}$$

where γ=k_{acc}/(k_{acc}+k_{nacc}), is the conformational gating factor and represents the probability of the ligand finding the apo-receptor in state r (Figure 2). The first-order rate constants

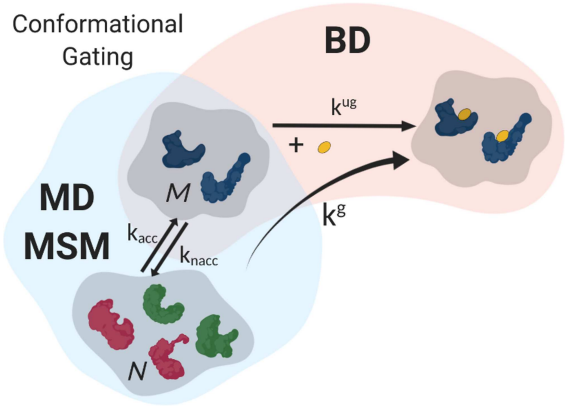


Figure 2: Schematic of coupling approach between BD and MD/MSM to compute gated ligand binding to accessible states of a multiconformation receptor. A conformational gating factor (γ), representing the probability of accessing the accessible states is derived from the previously described hidden Markov model (HMM) transition matrix of the apo-state conformational kinetics (see Theory), where γ=$\mathrm{k}_{acc}/(\mathrm{k}_{acc}+\mathrm{k}_{nacc})$ and where k_{acc} and k_{nacc} are the first-order rate constants for transitioning into the accessible and non-accessible states, respectively. BD simulations are then performed for binding to the accessible states yielding a second-order ungated association rate constant (k^{ug}). For ligand-receptor complexes in which other binding mechanisms are not rate-limiting, the overall association kinetics can be approximated by a gated rate constant, k^{g}=$\gamma\mathrm{k}^{ug}$.

k_{acc}=$\sum_{i\neq r}\mathrm{k}_{ir}$ and k_{nacc}=$\sum_{j\neq r}\mathrm{k}_{rj}$ describe the overall transition kinetics between accessible state r and all other non-accessible states in the apo-receptor. Here, k_{ij} represent the (off-diagonal) elements of the transition rate matrix $\mathbf{K}$, representing the net flux from a single metastable state i to another metastable state j, and which can be computed directly from the HMM of the apo-receptor, as described above.

We now expand this model more generally to the case of multiple accessible states. The apo-state space is divided into two disjoint subsets, containing all non-accessible N and all accessible M states. The first-order conformational rate constants between the two sets of states can then be expressed as the sum of transition rates from non-accessible $n \in N$ to accessible $m \in M$ states such that $\mathrm{k}_{acc} = \sum \mathrm{k}_{nm}$ and vice versa $\mathrm{k}_{nacc} = \sum \mathrm{k}_{mn}$. A gating factor ($\gamma$) can then be defined for the overall transition of the accessible states:

$$\gamma = \frac{\sum_{N,M} \mathrm{k}_{nm}}{\sum_{N,M}(\mathrm{k}_{mn} + \mathrm{k}_{nm})}. \tag{11}$$

However, γ alone does not contain the information about the respective contributions of the accessible states. To account for these we define the state specific gating factor $\Gamma_m = \gamma\epsilon_m$, where ϵ_m represents the relative population of state $m \in M$ and which can also be computed from the HMM. By combining with ungated rate constants (k_m^{ug}) computed from BD simulations for each state m, the overall gated association rate constant (k^g) for the set of states, M, is then given by:

$$\mathrm{k}^g = \sum_M \Gamma_m \cdot \mathrm{k}_m^{ug} \tag{12}$$

$$= \gamma \sum_M \epsilon_m \cdot \mathrm{k}_m^{ug}. \tag{13}$$

With this formalism, the combined ungated rate constant (k^{ug}) for the set of states, M, is then given by:

$$\mathrm{k}^{ug} = \sum_M \epsilon_m \cdot \mathrm{k}_m^{ug}, \tag{14}$$

and in the limiting case where there is only one accessible metastable state, $\epsilon_m = 1$, we recover the single-state gating factor. Extending the previous formalism[27] to multiple accessible states, 'slow' and 'fast' gating regimes can be differentiated in relation to the characteristic diffusion time $\tau_d = r_c^2/D$, where r_c is the collision distance between the molecules and D is the relative translational diffusion constant by:

$$\mathrm{k}^g = \gamma \sum_M \epsilon_m \cdot \mathrm{k}_m^{ug}, \tau_c >> \tau_d \tag{15}$$

$$\mathrm{k}^g = \sum_M \epsilon_m \cdot \mathrm{k}_m^{ug}, \tau_c << \tau_d, \tag{16}$$

where τ_c is the characteristic timescale of conformational gating given by:

$$\tau_c = \frac{1}{\sum_{N,M}(\mathrm{k}_{mn} + \mathrm{k}_{nm})}. \tag{17}$$

Materials and Methods

Initial preparation

Atomic coordinates for wildtype dimeric HIV-1 protease in the semi-open conformation were extracted from the 1HHP crystal structure[86] in the Protein Data Bank.[102] Those for a wide-open conformation were taken from the reported structure in our previous study.[95] The standard AMBER forcefield (ff03)[103] was used to describe all protease parameters. A dianionic state was assigned to the catalytic dyad as, at physiological pH, the unbound protease is unprotonated, whilst mono-protonation is thought to occur upon or after ligand binding.[104–107] The VMD software package[108] and the LEAP module of the AMBER 16 software package[109] were then used to build the initial systems. Each system was solvated using atomistic TIP3P water[110] in a cubic box extending at least 7 Å around the complex and then electrically neutralised with an ionic concentration of 0.15 M NaCl, resulting in fully atomistic systems for both semi-open and wide-open conformations of ∼40,000 atoms.

Equilibration and production MD simulations

All equilibration and production simulations were performed using ACEMD.[111] During equilibration the positions of the heavy protein atoms were restrained by a 10 kcal/mol/Å^2 spring constant and conjugate-gradient minimisation was performed for 2000 steps. The hydrogen atoms and water molecules were then unrestrained and allowed to evolve for a total of 500 ps at 300 K in the NVT ensemble to ensure thorough solvation of the complex and to prevent premature flap collapse.[112] The magnitude of the restraining spring constant was set to 1 kcal/mol/Å^2 for the following 500 ps, then to 0.05 kcal/mol/Å^2 for 500 ps, then finally to

zero for 500 ps. The temperature was maintained at 300 K using a Langevin thermostat with a low damping constant of 0.1/ps and the pressure was maintained at 1 atm using a Nose-Hoover barostat. The sytem was finally equilibrated for 6 ns of unrestrained simulation in the isothermal-isobaric ensemble (NPT). The long range Coulomb interaction was handled using a GPU implemention[113] of the particle mesh Ewald summation method (PME)[114] with grid size 84x84x84. A non-bonded cut-off distance of 9 Å was used with a switching distance of 7.5 Å. For the equilibration runs, the SHAKE algorithm[115] was employed on all atoms covalently bonded to a hydrogen atom with an integration timestep of 2 fs. Production simulations took place in the NVT ensemble. The hydrogen mass repartitioning scheme[116] was used, allowing a timestep of 4 fs. All other parameters were kept the same as in the equilibration phase. The production data set consisted of 465 and 499 simulations started from semi-open and wide-open conformations respectively, each for 130 ns with coordinate generation every 100 ps.

Our molecular simulation protocol for the HIV-1 protease has been previously validated using NMR S^2 order parameters.[95]

Determination of collective variable space

Before MSMs can be constructed from MD simulation data, the high-dimensional configuration space of atomistic coordinates needs to be projected onto a low-dimensional, but meaningful, collective-variable (CV) space. Finding the appropriate CV space is still a major challenge, despite considerable efforts at systemisation.[63–65] A minimum fluctuation alignment (MFA) method, introduced previously[53] and developed further here, was used to determine a set of information-rich collective variables that describe flap conformations in HIV-1 PR.

Minimum fluctuating (MF) residues for the protease were identified as residues 23, 24, 85 and 89-92 of each monomer. The homodimeric C_2 symmetry of the protease makes identical MF residue subsets on both monomers a sensible choice of vector termini. The x-axis was

thus chosen to be the vector between the center of mass (COM) of the backbone atoms of residues 89-92 of the second and first monomers respectively. A second vector was selected (between the COM of residues 23,24 and 85 of each monomer respectively). The resultant z-axis therefore approximated the C_2 symmetry axis. The selected residues are either within the α-helix of the protease or within the β-sheet region that supports each side of the active-site and thus exhibit minimal RMSFs compared to other residues.[95] Due to the choice of the x-axis, the resultant y-axis was similar to the vector describing the binding conduit of bound peptides in the active site.[79,80]

A 3D metric based on the re-aligned cylindrical polar components of the I50-I50′ C_α-atom flap tip separation vector, $\lambda(\lambda_r, \lambda_\Theta, \lambda_z)$, was then devised. A wrapping angle of 50° was added to all values of λ_Θ with a wrapping modulus of 360°. This extends the original 1D Cartesian λ_x metric used earlier[53] and was employed as a set of collective variables to describe the conformational transitions of the system.

Markov state model

The software package pyEMMA2[101] was used for Markov state model construction and analysis. The 3D λ-space was used as the set of collective variables in which to construct the models. The data were subsequently clustered into 100 microstates using k-means clustering in this 3D CV-space. Discretized state labels were assigned and Markov state models (MSMs) with Bayesian error estimation were built at various lag-times (τ) and the computed relaxation timescales analyzed as a function of τ. Constant relaxation timescales for the slowest transitions were exhibited at τ= 30 ns, significantly shorter than the length of all trajectories used, with time scale separation suggesting a set of five kinetically distinct macrostates (C1-C5). A subsequent MSM was then estimated using a lag time of 30 ns. The 100-state MSM at τ= 30 ns was coarse-grained into a five-state hidden Markov model (HMM) and validated using a generalized Chapman-Kolmogorov test. Fuzzy clustering of microstates into macrostates was performed and the resulting membership ($\mathbf{B}$) and observa-

tion (χ) distribution matrices computed. The transition matrix ($\tilde{\mathbf{P}}$) between the smaller set of five meaningful macrostates (C1-C5) was then computed using Equations 7 and 8. The stationary distribution of the five states and the transition kinetics between these states was then extracted from the HMM.

Structural characterization

Representative snapshots of the kinetically distinct macrostates (C1-C5) were extracted for structural characterization. Macrostate C1 was noted to be an antisymmetric doublet-state (see Results), thus representative structures were extracted separately for each of the corresponding substates C1a and C1b. For each macrostate, the microstate with the largest observation probability was selected. Within each corresponding microstate, conformers were ranked in order of proximity in the λ-space to the centroid and the closest 1000 conformers extracted for analysis. Microstates for C1a and C1b corresponded to those with the first and second highest observation probability, respectively, for macrostate C1. In addition, various structural features from the extracted ensembles were compared with respect to the first replica of a previous ensemble MD simulation of the HIV-1 protease in a closed conformation and bound to the SP1-NC natural octapeptide.[117,118] Analysis was implemented using Python scripts based on the MDAnalysis toolkit.[119] Distributions of the root-mean-squared-deviations (RMSDs) of the C_α backbone with respect to semi-open (1HHP) and closed (1KJ7) reference crystal structures were computed as were the distributions of the absolute flap-tip separation (λ). Instantaneous hydrogen bonds were computed based on a donor-acceptor distance threshold of $\leq$3.5 Å and a donor-hydrogen-acceptor angle of $\geq$150°. Only those hydrogen bonds with an exhibition frequency $f_{hb} \geq 0.5$ were considered stable - the corresponding mean distances and angles were computed from the exhibited snapshots. Hydrophobic contacts were determined from sidechain-sidechain distance analysis; the minimum non-hydrogen atom distance between pairs of sidechains was computed for each conformer. A hydrophobic contact was assigned to hydrophobic residue pairs with a

mean minimum distance $\leq$ 5 Å across the corresponding ensemble/trajectory. Volumetric maps of the averaged weighted atomic densities for each apo-conformer ensemble and the natural octapeptide from a former study[117] were computed with the Volmap tool in VMD[108] using a grid spacing of 0.5 Å.

Brownian dynamics simulations

Atomistic coordinates for the 100 conformers most proximal to the microstate centroid in each structural ensemble were extracted to use in Brownian dynamics (BD) simulations to compute diffusional association rate constants for a set of inhibitors with experimentally determined rate constants.[97] Protein partial atomic charges were assigned using the ff14SB forcefield.[120] Atomistic coordinates and partial charges for the inhibitors were extracted from a previous study.[121] PQR files were generated using the PDB2PQR module in AMBER18.[122] Electrostatic potential grids for the protein structures and for the inhibitors were computed using APBS[123] with a grid spacing of 1 Å, an ionic concentration of 150 mM and a temperature of 298K, corresponding to experimental conditions, as well as a solvent dielectric constant of 78, a solute dielectric constant of 2 and an ionic radius of 1.5 Å. For the calculation of the electrostatic potential, the linearized Poisson-Boltzmann equation was solved under the single Debye-Hückel dielectric boundary condition. All BD simulations were carried out using SDA7.2.3,[124] and association rates constants were computed using the Northrup-Allison-McCammon algorithm.[32] SDA employs an effective charge model (ECM),[125] where the charge on a reduced number of centers is fitted to reproduce, in a uniform dielectric medium, the potential computed with APBS for a heterogeneous dielectric medium in the region around the molecule. ECM charges for the protein were determined using the standard protocol in SDA, those for the ligands were determined from a newly developed protocol.[126] Electrostatic and hydrophobic desolvation grids were generated with a grid spacing of 1 Å. The ionic strength was set to zero for computing the electrostatic grid, with electrostatic and hydrophobic desolvation factors of 0.36 and -0.013 assigned, respectively. The solute

and probe radii were both set to 1.4 Å.

Translational and rotational diffusion coefficients were calculated for all solutes using Hydropro[127] with an atomic element radius (AER) of 2.9 Å and 1.2 Å for protein and ligands respectively, with a σ_{min} of 1.0 Å and a σ_{max} of 2.0 Å. For each macrostate-ligand combination, a diffusional association rate constant was computed from a set of 100000 trajectories (100 protein conformers $\times$ 1000 replicas). Each trajectory was started with an initial center-to-center distance of 150 Å between the solutes and stopped when the solutes reached a seperation of 250 Å. A variable time step was used, linearly ranging from 1 ps at a solute-solute separation of 50 Å to 20 ps at a separation of 90 Å. A distance threshold of 3 Å was set for pairs of contacts to be assigned as independent reaction criteria that described formation of a receptor-ligand encounter complex. Reaction criteria included polar interactions at 3.5 Å defined in terms of acceptor-donor pairs, as well as methyl, halogen and aromatic interactions at 4.5 Å, defined in terms of π-π pairs from the above-mentioned crystal structures of the corresponding complexes, filtering out any residues that did not match the sequence of the apo-protein (PDB: 1HHP) and including symmetrically equivalent contacts due to the C_2 symmetry of the protease dimer. Reaction criteria coordinates were taken for each BD simulation after superimposition of the C_α backbone of residues 1-43 and 58-99 in each monomer of a given apo-conformer with the corresponding original inhibitor-bound complex. An encounter complex was counted whenever a sufficient number of contacts from the subset of pre-defined reaction criteria was observed within a specified distance. The association rate constant was initially calculated using a range of definitions of encounter complex in terms of number of contacts, $\{1, ..., 5\}$ and their corresponding distances from a minimum of 3 Å and with increasing window intervals of 0.5 Å up to 20 Å. Ungated association rate constants were subsequently computed by requiring the formation of 3 independent contacts shorter than 4 Å.

BD docking simulations were performed for each of the ligands diffusing to each of the

protein macrostates under the same reaction criteria as for the association rate calculations to check if the reaction criteria corresponded to ligands binding within the active site of the protease. 10000 trajectories were run for protein conformers with the highest number of encounters obtained in the association rate constant calculation simulations, and complexes were recorded at a 4 Å reaction criterion window distance.

Results

Ensemble molecular dynamics simulations resulted in $\sim$1.25 $\times 10^6$ molecular conformers for subsequent analysis. In order to reveal and characterize meaningful conformational states from this ensemble, we constructed a converged Markov state model (MSM) based on a suitable projection of the atomistic coordinates into a three-dimensional collective variable space. This enabled us to identify a small set of discretised macrostates of the system, compute their equilibrium distribution and transition kinetics, and then analyse representative constituent conformers belonging to each macrostate to reveal distinct structural features. This in turn allowed the differentiation of the macrostates that were accessible and non-accessible to ligand binding and the subsequent computation of a gating factor. Brownian dynamics simulations enabled computation of the ungated association rate constant to accessible states from which the overall gated rate constant could be computed for a set of ligands.

Kinetic network of HIV-1 protease conformations

The 3D λ-state-space in cylindrical polar coordinates (λ_r, λ_Θ, λ_z) was chosen as a collective variable space in which to construct Markov state models (MSMs). The λ-state-space was discretised into 100 microstates. MSMs constructed using Bayesian error estimation across a range of lag-times (τ) show independence of the relaxation timescale (t_n) for the slowest-order transitions at τ = 30 ns (Figure 3(a)) and a significant timescale separation between

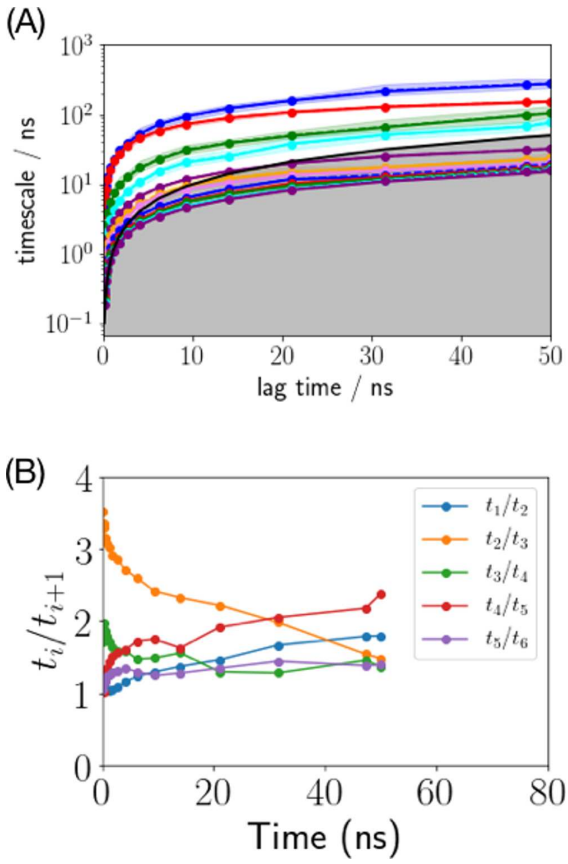


Figure 3: (a) Relaxation timescales (t_n) for the dominant non-stationary eignvectors of the MSMs constructed in the λ-space at increasing lag times (τ) with Bayesian error estimation. Independence of the relaxation timescale with lag time is observed at $\tau = 30$ ns. (b) Plot of the ratio of proximal relaxation timescales (t_i/t_{i+1}) shows a signifcant timescale separation of $t_4/t_5 > 2$ at $\tau = 30$ ns suggesting the existence of five kinetically meaningful metastable states.

the fourth and fifth non-stationary eigenvectors with $t_4/t_5 > 2$ (Figure 3(b)).

Consequently a MSM was constructed with Bayesian error estimation using 100 microstates and $\tau = 30$ ns and the microstate equilibrium distribution computed. A potential of mean force (PMF) landscape, computed for 2D projections of the λ-state-space and weighted by this equilibrium distribution (Figure 4(A)), samples a large region, ranging from $0 < \lambda_r < 40$ Å, $0 < \lambda_\Theta < 360°$ and $-15 < \lambda_z < 15$ Å. This PMF differentiates several significant energetic minima with $\lambda(\lambda_r, \lambda_\Theta, \lambda_z)$ centers at $\lambda \sim$(8 Å,40°,±5 Å), $\lambda \sim$(5 Å,100°,0 Å), $\lambda \sim$(20 Å,150°,0 Å) and $\lambda \sim$(8 Å,300°,0 Å), with minima values of -0.5, -4.5, -3.5 and -2.5 kcal/mol respectively, as well as one saddle point with a center at $\lambda \sim$(10 Å,120°,0 Å) and an energy of -2.0 kcal/mol.

Timescale separation analysis implies that five kinetically meaningful macrostates exist, with the two slowest transitions in the system being around or larger than the 100 ns-timescale. Based on this a corresponding five-state coarse-grained hidden Markov model (HMM) was computed. This further clustered the microstates into kinetically separated macrostates, C1-C5 (Figure 4(A)). This model was validated by performing a generalized Chapman-Kolmogorov (C-K) test (Supporting Figure S1(A)). The C-K test showed good agreement with the predicted probability decay from the model at $\tau = 30$ ns ($p_{IJ}^k(\tau)$) and the expected decay from models constructed at multiple (k) lag times ($p_{IJ}(k\tau)$). Moreover, the probabilities, χ, of observing each of the 100 microstates across the five macrostates (see Theory) were computed and projected onto the λ-space (Supporting Figure S1(B)). This analysis confirmed that the five macrostates also exhibited mutual exclusivity of microstate probabilities and therefore could be interpreted as distinct states. The microstates that constitute most macrostates correspond well with the energetic minima in the PMF. The only exception is C2, whose microstates seem to be primarily distributed in a transitional region between C4 and C5.

The HMM yielded a kinetic network of the stationary distributions and their respective first-order transition kinetics (Figure 4(B)). The five states, C1-C5, exhibit different

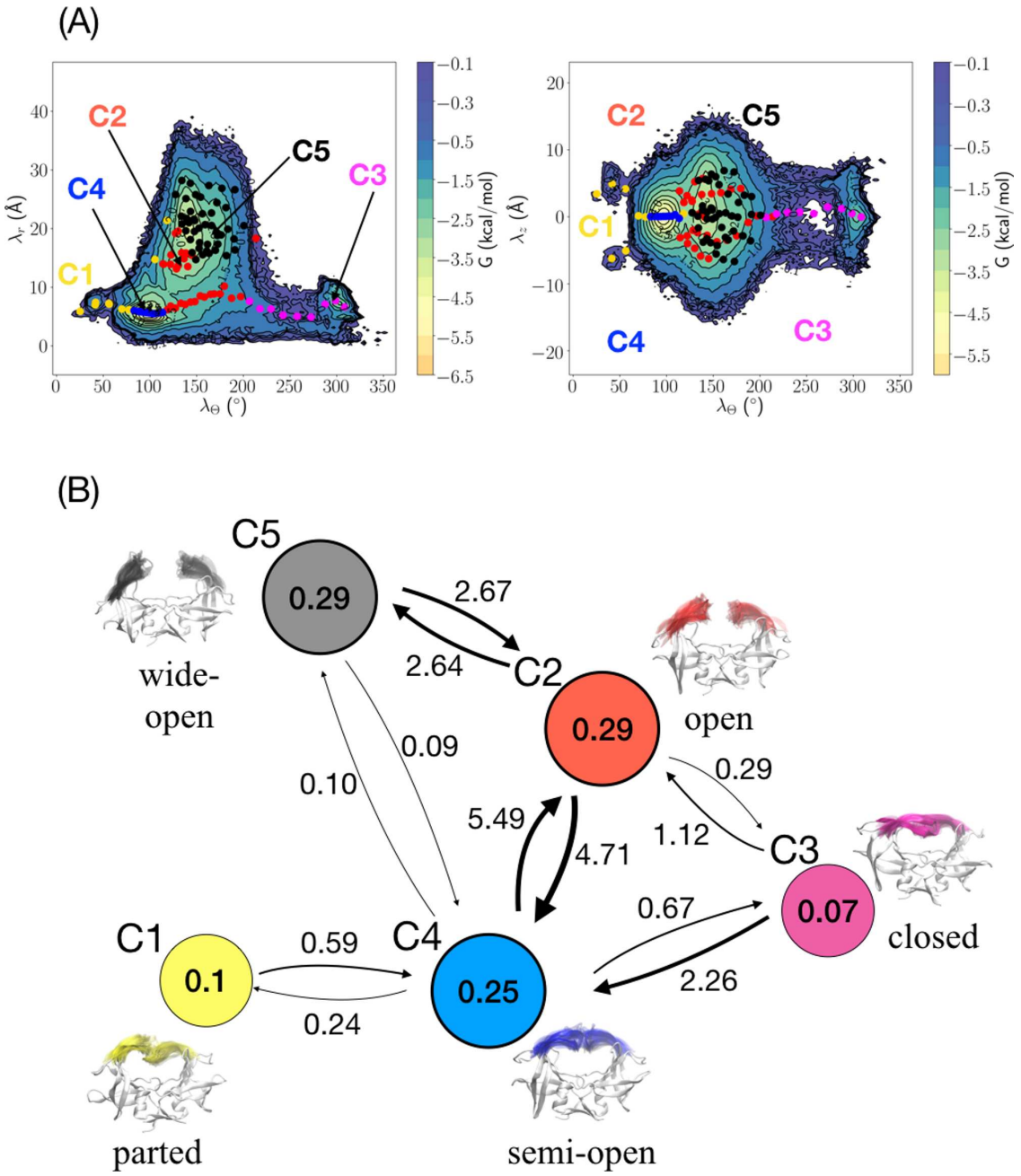


Figure 4: (A) Potential of mean force (PMF) of HIV-1 PR in the apo-state in the $\lambda(\lambda_r,\lambda_\Theta,\lambda_z)$-collective variable space, projected onto λ_r-λ_Θ and λ_Θ-λ_z. Values are in kcal/mol. Energetic minima correspond to frequently sampled regions of the λ-space weighted by the equilibrium distribution of discretised microstates of the MSM. A HMM based on five macrostates (C1-C5) has clusters of corresponding microstates whose λ-centroids are overlayed on the PMF. (B) Kinetic network between the five coarse-grained macrostates, corresponding to conformations C1-C5, determined by the hidden Markov model (HMM). Stationary probabilities are shown in the circles, forward and reverse first order transition kinetics are shown by arrows with values in μs^{-1}. Displayed structures for each macrostate are superpositions of conformers belonging to the most probable microstate given the corresponding macrostate.

stationary probabilities (ρ) with $\rho_1 = 0.1$, $\rho_2 = 0.29$, $\rho_3 = 0.07$, $\rho_4 = 0.25$ and $\rho_5 = 0.29$ corresponding to zero-centered Boltzmann-weighted free energies of ∼-0.2, -0.8, 0, -0.7 and -0.8 kcal/mol, respectively. Thus the free energy differences betwen all five states are thermodynamically barely indistinguishable from thermal noise (∼0.6 kcal/mol).

Nonetheless, there are significant differences in the transition kinetics between the various states. Only direct fluxes faster than 100 μs are shown ($\geq 3\times10^{-4}/\tau = 0.01\ \mu\text{s}^{-1}$). Forward and reverse transitions between C2 and C5, as well as between C2 and C4, are well within the sub-μs timescale ($> 1\ \mu\text{s}^{-1}$), although the rate for direct transition between C4 and C5 is ∼0.1 μs^{-1}. The C1 conformation only directly accesses the C4 conformation, and does so with kinetics slower than the μs-timescale, whilst the C3 conformation accesses both C2 and C4 conformations with forward and reverse kinetics faster and slower than 1 μs^{-1}, respectively. Only the C4 conformation directly accesses all other conformations on a timescale faster than 100 μs.

Structural characterization of principal conformations

The conformational features of representative structures corresponding to each of the macrostates were analyzed. Probabilistic clustering of microstates into macrostates via the observation probability matrix ($\chi_{I,j}$) enables selection of the most likely microstates that represent each macrostate (Supporting Table S1). For each macrostate, the most probable microstate largely coincides with the corresponding minimum or minima in the λ-space (Supporting Figure S2). Macrostate C1 is characterized by two distinct minima in the λ-space equidistantly separated from the origin along λ-z. This implies a symmetry-related doublet state. Moreover the two most probable microstates for C1 correspond to each of these minima - thus C1 was split into two sub-states, C1a and C1b, in subsequent analyses corresponding to the $+\lambda_z$ and -λ_z regions respectively. A set of 1000 closest conformers was thus extracted from the most probable microstate of each macrostate based on the ranked proximity of each snapshot to the corresponding microstate centroid in the 3D cylindrical polar λ-space.

Visual inspection of each set of conformers reveals that several macrostates correspond to known conformations of HIV-1 protease (Figure 4(B)), as well as a previously unidentified conformation. Hydrogen bonds and sidechain-sidechain contacts between each flap and the opposite (*trans*) monomer were analyzed (Supporting Tables S2 and S3). C2 and C5 correspond to open and wide-open conformations, respectively, where the flaps are pulled back enabling sterically unhindered access to the active site (Figure 4(B)). These two conformations exhibit no *trans* flap-monomer interactions. Macrostates C3 and C4 correspond to the characteristic closed and semi-open conformations, respectively, where the flaps extend over the active site but with opposite handedness (Figure 5 and Figure 6(A)-(B)). In C3, the I50 sidechain interacts with the characteristic array of hydrophobic residues (V32′, I47′, I54′, I80′ and I84′) on the opposite monomer and symmetrically vice versa. In C4, these contacts are missing, and are instead replaced predominantly a smaller set of hydrophobic contacts (I50-I50′,I50-F53′ and F53-I50′) and a symmetric pair of inter-glycine amide hydrogen bonds (G51–N-H–O-G49′ and G51′–N-H–O-G49). C1a and C1b are identified as anti-symmetric sub-states of a novel conformation, where one flap is extended as in the semi-open conformation whilst the other orients under it and traverses across the active site towards the other monomer (Figure 5). They share a subset of hydrophobic contacts with both C3 (C1a: I50-V32′ and I50-I47′, C1b: V32-I50′ and I47-I50′) and C4 (C1a: F53-I50′, C1b: I50-F53′) but also exhibit sidechain contacts (C1a: I50-D30′ and M46-I50′, C1b: D30-I50′ and I50-M46′) and a backbone hydrogen bond (C1a: G48′-N-H–O-I50, C1b: G48-N-H–O-I50′) that are unique to them and which serve to stabilise these conformations. In particular, the I50-D30′/D30-I50′ contacts bring the corresponding flap tip diagonally across to the opposite side of the active site.

Observations of C3 and C4 are further confirmed by unimodal distributions of flap-RMSD with respect to crystal-structures of closed and semi-open conformations for which C3 and C4 exhibit the smallest mean RMSDs, 1.8±0.5 Å and 2.0±0.4 Å, respectively (Figure 6(C)). All other macrostates exhibit significantly larger mean RMSDs with respect to both crystal

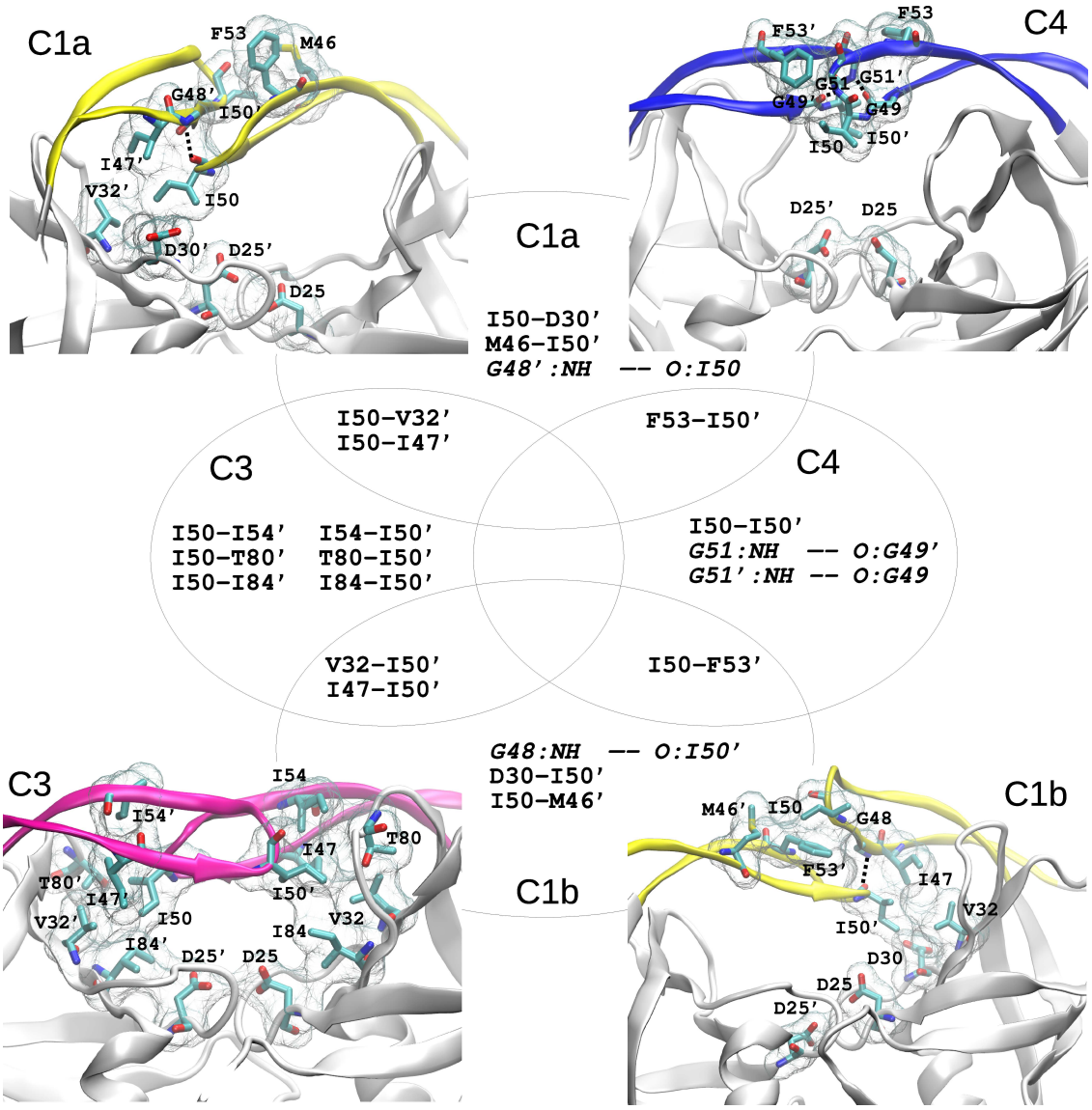


Figure 5: Mututally exclusive and shared sidechain-sidechain and hydrogen bond interactions between each flap region and the alternate monomer across different macrostates. C2 and C5 do not form any such interactions and are not displayed. Hydrogen bonds are marked with dashed black lines and noted in italics. The catalytic dyad D25/D25′ is shown for structural context.

structures. Analysis of the distribution of absolute flap tip separation distances (λ) shows that C3 and C4 all exhibit peak distances substantially lower than 10 Å, whilst C1a and C1b both peak at $\sim$10 Å(Figure 6(D)). By contrast, C2 and C5 exhibit a relatively broad distribution of distances peaking at 15 Å and 20 Å respectively, implying that only these two states provide sufficient access to peptidomimetic ligands binding in the active site. The λ-distributions of all conformers for each of the top microstates of every macrostate exhibit very similar profiles respectively to the macrostate conformer sets selected here (Supporting Figure S2).

Cross-sections of averaged weighted atomic density over the conformers of each set provide further insight into the steric active-site accessibility of the corresponding macrostates (Figure 6(E)-(H)). The averaged density of a closed bound reference complex of HIV-1 protease with the SP1-NC cleavage peptide was calculated based on previous simulations.[117,118] Superimposing densities calculated individually for the apo-protease (gray) and ligand (red) from this complex are consistent with both a snugness of fit and the inaccessibility of binding the enzymatically viable closed conformation (Supporting Figure S3). Furthermore, the comparable density profiles of the protein between the C3 conformation in the apo-protease simulations and the ligand-bound reference complex (Supporting Figure S3) further confirm that C3 corresponds to the closed conformation of the protease in the presence of bound ligand. The ligand exhibits significant density between -5 Å $< x <$ 5Å, -15 Å $< y <$ 15Å, 5 Å $< z <$10 Å. To provide corresponding steric context to the macrostates, the density of the peptide from this analysis was superimposed with those calculated from each of the representative conformer sets. Cross-sections in the x-y plane reveal that at z = 8 Å, states C1a, C1b and C3 exhibit partial occlusion at the edges of the active site, whilst states C2, C4 and C5 exhibit a zero-density lateral exit pathway in the y-direction. Moreover, states C2-C5 exhibit a zero-density groove corresponding to the active site that can accommodate the reference peptide density. However, by z = 12 Å, only macrostates C2 and C5 maintain this groove, whilst the flap extensions in the other conformations incur protein density that closes

off peptide escape in the z-direction. Lateral cross-sections in the x-z and y-z planes confirm that only the C2 and C5 macrostates maintain a zero-density pathway out of the active site along the z-direction. Thus, combining exit pathways from the y- and z-directions, only macrostates C2, C4 and C5, corresponding to open, semi-open and wide-open conformations, respectively are selected as ligand accessible states for rigid body binding simulations.

Gated ligand binding of peptidomimetic inhibitors

A set of peptidomimetic ligands (Figure 7), covering a range of experimentally determined association rate constants, k_{on} $\sim 10^4$ - 10^{10} $\mathrm{M}^{-1}\mathrm{s}^{-1}$, were chosen for Brownian dynamics simulations to each of the three chosen (C2, C4 and C5) accessible states of HIV-1 protease. The closest 100 conformers to the centroid of the most probable microstate were extracted for each of the accessible states. All three sets of conformers exhibit similar translational diffusion constants ($< D_T(C2) > = 9.90 \pm 0.08 \times 10^{-11}\ m^2/\mathrm{s}$, $< D_T(C4) > = 1.00 \pm 0.01 \times 10^{-10}\ m^2/\mathrm{s}$), $< D_T(C5) > = 9.80 \pm 0.08 \times 10^{-11}\ m^2/\mathrm{s}$) and similar radii of gyration ($r_g(C2) = 1.99 \pm 0.03 \times 10^{-9}\ m$, $r_g(C4) = 1.91 \pm 0.02 \times 10^{-9}\ m$, $r_g(C5) = 2.04 \pm 0.02 \times 10^{-9}\ m$). Translation diffusion constants and radii of gyration for the ligands range between 2.88 - 4.58 $\times\ 10^{-10}\ m^2/\mathrm{s}$ and 4.43 - 7.83 $\times\ 10^{-10}\ m$, respectively, yielding a characteristic diffusional timescale, τ_d, that ranges between $\sim$ 12 - 16 ns depending on the ligand. By contrast, based on the outcome of the MSM and using Equation 17, the characteristic protein conformational transition timescale, τ_c, is computed to be almost an order of magnitude larger ($\tau_c = 85$ ns), suggesting a 'slow' gating regime ($\tau_c >> \tau_d$). Thus, there is a kinetic gating factor to binding and via Equation 11, this is computed to be $\gamma = 0.75$.

Ungated rate constants (k_m^{ug}) were then computed from BD simulations to each of the macrostates m=C2,C4 and C5, where the value for each state was computed using ensemble averaging over the set of 100 conformers that represented it. For each state, a number of complexes were formed at the reaction criteria corresponding to the bound state (see Methods). BD docking simulations were performed for each of the ligands binding to each

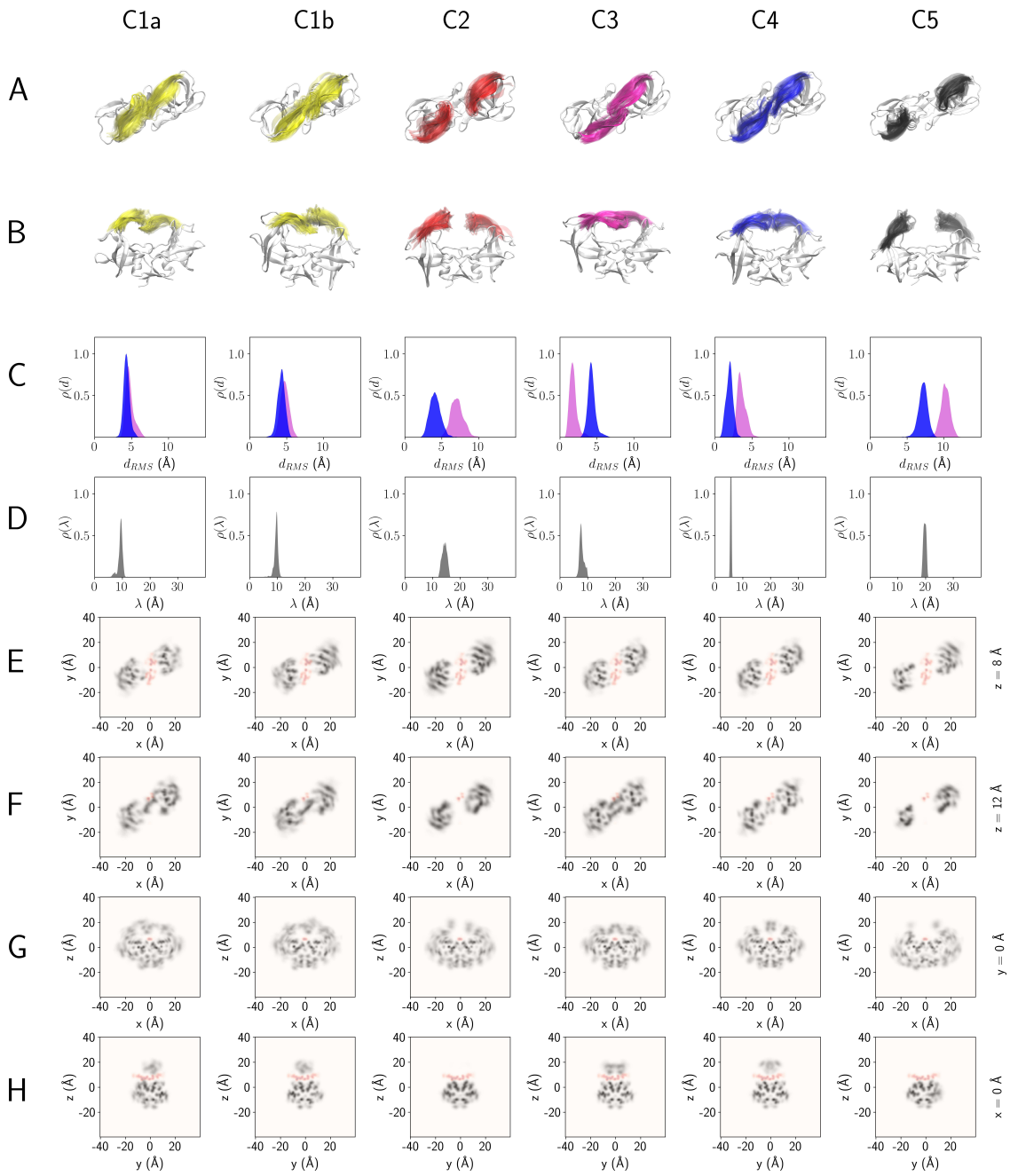


Figure 6: Structural characterization of representative conformers of each macrostate of HIV-1 apo-protease. (A) Top (x-y plane) and (B) side (x-z plane) views of ribbon representations of superimposed structures in each conformational ensemble, (C) Distribution ($\rho(d)$) of root-mean-square deviations (d_{rms}) of C_α atoms of protease flaps (residues 43-58 in each monomer) with respect to crystallographic reference closed (pink) and semi-open (blue) structures. (D) Distribution of ($\rho(\lambda)$), the absolute flap tip separation distance (λ). (E-H) Cross-sections of superimposed volumetric density between the SP1-NC natural peptide in a closed-bound reference structure (red) and representative conformational ensembles of each macrostate (grey) in the x-y plane at z =8 Å (E) and z = 12 Å (F), the x-z plane at y = 0 Å (G) and the y-z plane at x = 0 Å (H).

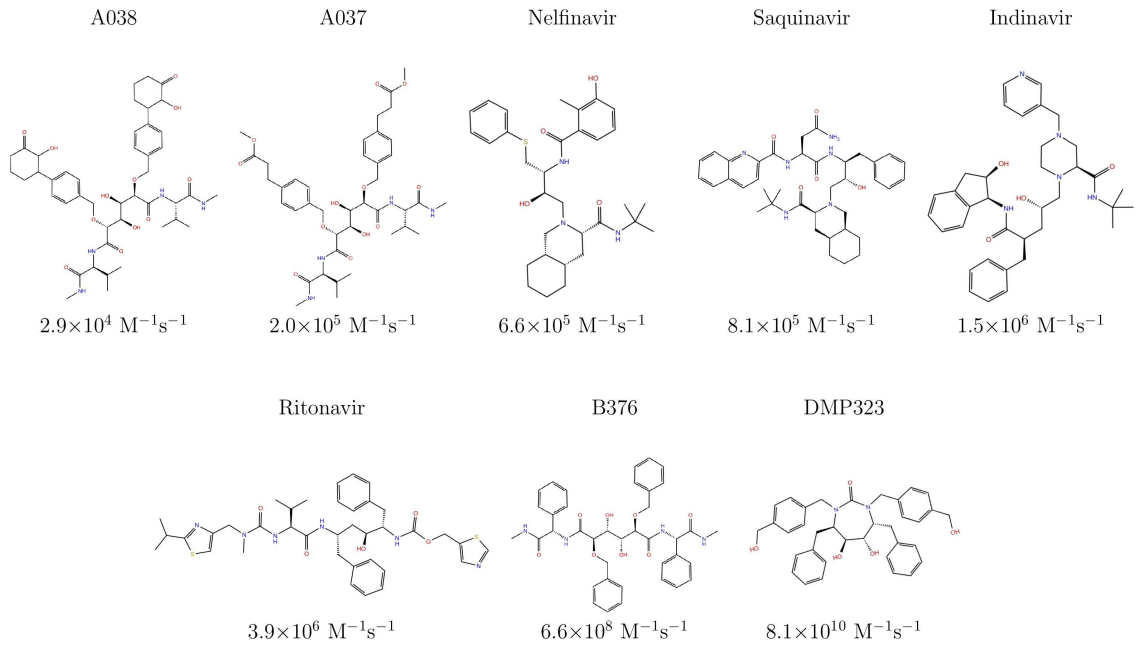


Figure 7: Chemical structures of the set of peptidomimetic ligands for which association rate constants were computed and which have a wide range of experimentally determined association rate constants, k_{on}^{exp} $\sim 10^4$ - 10^{10} $M^{-1}s^{-1}$.

of the accessible protein macrostates under these reaction criteria and the most popular clusters inspected. These confirmed that fulfillment of the reaction criteria corresponds to ligands binding within the active site of the protease (Supporting Figure S4). Similarly, BD simulations were carried out for ligands binding to the inaccessible states (C1 and C3). Lack of complexes formed within the active site of these two states thus further confirmed assignment of them as inaccessible (Supporting Figure S4).

Values for the ungated rate constants (k_m^{ug}) were combined to give the overall ungated rate constants for the set of accessible states (k^{ug}) and finally multiplied by the gating factor γ to obtain the gated rate constants (k^g). Our calculations show, as expected, that binding to the C4 (semi-open) state is slower than to the C2 (open) state and also slower than binding to the C5 (wide-open) state for all ligands except Nelfinavir and Saquinavir (Table 1). Furthermore, binding for most ligands is faster to C2 than C5, occasionally by several-fold, whilst for Ritonavir and B376 it is only marginally slower. The resulting combined ungated rate constants (k^{ug}) are almost all within one order of magnitude of each other,

ranging from k^{ug}∼60-750 μM^{-1}s^{-1} and, as the gating factor is 0.75, the gated rate constant range is only marginally reduced (k^{g}∼45-565 μM^{-1}s^{-1}). This results in k^{g} being up to several orders faster than the corresponding experimental association rate constants (k_{on}^{exp}) for most ligands, similar for B376, and several orders of magnitude slower for the fastest binder, DMP323 (Table 1).

Discussion and Conclusions

We have developed an approach to characterize the effects of gating by a multi-conformation protein consisting of subsets of accessible and inaccessible states on ligand association rate constants. Our approach first involves the construction of a kinetic network model of the apo-protein from atomistic MD simulations from which we identify meaningful conformations, compute their relative populations and interchange kinetics, structurally characterizing them in terms of ligand accessibility. We integrate the calculated first-order rate constants for conformational transitions into a multi-state gating theory from which we derive a characteristic timescale for conformational transitions (τ_c) and a gating factor γ. This factor modulates the association rate constants of ungated ligand binding to the subset of accessible conformations (k^{ug}) - which we compute here for a set of ligands from Brownian dynamics simulations - when τ_c is significantly slower than the characteristic timescale for diffusional encounter (τ_d) and enables us to quantify the degree of conformational gating by the protein.

Applied to HIV-1 protease, the kinetic network model of the apo-protein has macrostates corresponding to canonical crystallographic conformations - namely the well-characterized closed (C3) and semi-open (C4) states, as well as open (C2) and wide-open (C5) states previously reported from atomistic and coarse-grained simulations. It also reveals a previously unreported (C1) state (Figure 4). This latter conformation is an asymmetric doublet macrostate (C1a and C1b) in which the β-hairpin moiety, termed 'flap', of one monomer, reaches into and across the active site, making contacts with the opposite monomer and vice

Table 1: Comparison of experimental association rate constants (k_{on}^{exp}) with computed ungated constants for individual (k_{m}^{ug}) and multiple-combined (k^{ug}) accessible states and corresponding gated rate constants (k^{g}), computed with a gating factor of γ = 0.75. All rate constants are given in units of $\mu\mathrm{M}^{-1}\mathrm{s}^{-1}$.

Ligand	k_{C2}^{ug}	k_{C4}^{ug}	k_{C5}^{ug}	k^{ug}	k^{g}	k_{on}^{exp}
A038	417.00 ± 5.24	103.72 ± 1.50	330.67 ± 5.70	292.48 ± 4.27	219.36 ± 3.20	0.03 ± 0.00
A037	349.53 ± 2.81	89.69 ± 2.76	262.07 ± 5.66	240.71 ± 3.79	180.53 ± 2.84	0.20 ± 0.04
Nelfinavir	156.96 ± 3.70	84.54 ± 2.58	61.77 ± 0.80	101.89 ± 2.35	76.42 ± 1.76	0.66 ± 0.30
Saquinavir	103.56 ± 3.02	48.59 ± 3.21	32.32 ± 2.11	62.11 ± 2.76	46.58 ± 2.07	0.82 ± 0.16
Indinavir	1052.30 ± 8.50	269.74 ± 4.65	866.02 ± 10.31	751.50 ± 7.97	563.63 ± 5.98	1.53 ± 0.24
Ritonavir	242.21 ± 4.06	39.90 ± 1.28	250.47 ± 5.67	184.16 ± 3.79	138.12 ± 2.84	3.92 ± 0.11
B376	631.82 ± 6.36	112.97 ± 2.71	672.54 ± 4.10	489.77 ± 4.47	367.33 ± 3.35	205.00 ± 187.00
DMP323	533.32 ± 2.98	157.72 ± 2.26	480.90 ± 5.78	401.87 ± 3.74	301.40 ± 2.81	25200.00 ± 9990.00

versa. These contacts are a mixture of those unique to the conformation (especially the flap tip I50 with the distal D30′), some only otherwise exhibited by the semi-open conformation, and others only otherwise by the closed conformation (Figure 5). Structural characterization reveals that only this state exhibits significant overlap in the averaged atomic density when superimposed with a putative natural substrate bound in the closed conformation. The other characterized states fit within an existing putative model of conformational change across the catalytic cycle[128] - preferring a semi-open state in the apo protein, transitioning to an open/wide-open state during ligand encounter, then to the closed bound reactive state, and finally via an open state again for product release, back to the semi-open state. Our results suggest that C1 sterically partitions the active site, thus we term it 'parted' here. It may therefore assist in separating post-cleavage peptide fragments that are no longer covalently connected - and thus be favoured during product release, especially as it lies structurally on the pathway between closed and semi-open. CGMD simulations have suggested product release can occur even in the closed conformation.[129] The parted conformation is kinetically disjoint from all other states except the semi-open in our apo-protein model - future atomistic simulations may reveal whether transition to the parted conformation from closed becomes kinetically favourable during product release.

Our kinetic model suggests that, individually, all of the macrostate conformations are almost thermodynamically indistinguishable from each other. The stationary distribution favours the subset of accessible (semi-open, open and wide-open) states, but even combined they are only ∼1 kcal/mol more favourable than the sum of inaccessible states (parted and closed). Furthermore, there is comparatively rapid interchange between semi-open, open and wide-open states on the sub-μs timescale whilst being an order of magnitude slower from these states to either parted or closed conformations. Interestingly, this results in a slow gating regime for ligand binding ($\tau_c >> \tau_d$), but because the calculated gating factor is not much smaller than unity ($\gamma = 0.75$), this can still be interpreted as the protease almost always being accessible or open to binding. This view broadly agrees with our previous

study[95] but here, we can quantify the estimated timescale of transition from semi-open to wide-open to be ∼500 ns, in good agreement with a previous kinetic model.[96] However, our model also suggests an equistable population for the wide-open conformation compared to the other conformations, rather than a rarely accessed state. The role of the wide-open state is still ambiguous as ligand binding does not seem to require it. Indeed, we find binding of almost all ligands studied here is faster to the open than to the wide-open state (Table 1). The relative stability of the wide-open state may instead be compatible with alternate roles. It has been observed that the protease makes direct non-specific interactions with the viral RNA - dramatically enhancing catalytic activity.[130] Whilst the mechanism for this remains unclear, the wide-open state may further modulate RNA binding and thus indirectly regulate catalysis. The immature HIV-1 ribonucleoprotein forms a phase-separated condensate that allows influx of protease for its subsequent maturation.[131] The conformational flexibility of the protease, including its ability to form the wide-open state may aid its mobility in such a crowded environment.

Possible slower conformational transitions that are not observed in our simulations may also play a role. Solution NMR experiments attributed a ∼100μs transition to a wide-open state.[89] However, our model suggests that the transition to wide-opening occurs on the sub-μs timescale. It is possible that the slower transition observed by NMR may be to yet another state. If such a state was inaccessible to ligand binding then it would significantly reduce the gating factor, and in turn reduce the gated association rate constant for all ligands.

The approach reported here enables us to dissect the potential contribution of conformational gating to a set of accessible states during the process of ligand binding and can be used to suggest possible binding mechanisms. For bi-conformational systems where the active and inactive conformations correspond to accessible (open) and inaccessible (closed) states, respectively, gating can be a modulating factor to the conformational selection (CS) mechanism. However, the complexity of HIV-1 protease draws out an important distinction between accessibility and active state. Even though a small ligand could bind directly to the

active conformation (closed), this is largely inaccessible to peptidomimetic ligand binding, precluding a pure CS mechanism. Conformational gating therefore modulates accessible but intermediate binding states. For a pure induced fit (IF) mechanism, the ligand would need to initially bind a single conformation-gated accessible state, followed by a transition into the active (closed) conformation that may or may not be rate-limiting. However, several alternate complex binding mechanisms consistent with a dock and coalesce (DC) model may be possible, some of which are not addressed by our model.

In this study, we intrinsically decouple gating kinetics from ligand approach. However, proximity of a ligand to the protease may affect flap gating kinetics through transient intermolecular interactions. Therefore, even though the calculated gating factor suggests the protease active site is mostly accessible to binding, proximal ligand approach and/or loose binding away from the active site may alter this factor. Secondly, changes in protein and ligand conformations upon interaction are not accounted for in our rigid-body BD simulations. Transient interactions that involve flexible adaptation of a ligand to the protein along the binding pathway may result in either over- or underestimation of the ungated rate constant. Moreover, such interactions may allow a reactive flux through binding pathways different to first binding an accessible conformation. For example, a ligand binding loosely to the lateral exterior of the active site in a closed conformation, may kinetically trap the protease so that it cannot readily transition to an open state. However, this may be followed by partial opening or relaxation of the flaps to enable threading of the ligand into the active site to adopt a native bound state, with the rate constant depending on the context-specific flexibility of both the ligand and the protease and the ability to form transient interactions along the binding pathway. This deviation from pure CS, through a set of DC steps has been observed as a minor reactive flux in previous MD simulations of immature HIV-1 protease self-association.[53] Conversely, initial encounters distal from the binding site may rapidly direct the ligand towards the active site, possibly by lower-dimensional diffusion across the protease surface. Similarly, another rationale for the existence of states such as the parted

and wide-open may be to facilitate this rapid binding mechanism.

Here, we studied a set of inhibitors of HIV-1 protease whose association rate constants measured by surface plasmon resonance (SPR) vary across a wide range: k_{on}^{exp} $\sim 10^4$ - 10^{10} $M^{-1}s^{-1}$. However, the calculated gated rate constants (k^g) all lie within a narrow range from $\sim$0.5-5.7 $\times$ 10^8 $M^{-1}s^{-1}$. Only one inhibitor (B376), a small peptidomimetic ligand, exhibits a computed gated rate constant similar to the experimental association rate constant. The dominant binding mechanism for this inhibitor and those with a similar experimental rate constant is likely to one where binding is rate-limited by the initial conformation-gated diffusion into the active site and the subsequent transition to the native bound state is very fast.

Most inhibitors studied here are large peptidomimetic ligands (A038, A037, nelfinavir, saquinavir, indinavir and ritonavir) for which $k^g >> k_{on}^{exp}$ (Table 1) but for which k_{on}^{exp} varies by only $\sim$100-fold. These ligands are between $\sim$100-10000 fold slower in binding than B376. Larger peptidomimetic chains have a greater degree of conformational freedom than smaller ones which is likely to be a dominant factor in this observed decrease in association rate constant. Their association may be limited by the gating of the conformational changes of the ligand or by subsequent protein-ligand changes to to the native bound configuration, both of which are not considered here. This mechanism is compatible with earlier findings from CGMD simulations on ligand binding.[98] However, it is possible that these ligands also exhibit the above-mentioned transient induced interactions not captured by our approach that may impede binding either by affecting the gating factor, ungated rate constant and/or by an alternate mechanism. Differential combinations of these contributions may explain the rate constant variation within this set of compounds.

Finally, the inhibitor DMP323, a smaller and more rigid cyclic urea compound, exhibits $k^g << k_{on}^{exp}$ and is therefore not limited by gating of the diffusional association. Indeed, its rate constant exceeds the Smoluchowski diffusion limit. It is possible this fast rate constant could be due to an initial first encounter distal from the binding site followed by rapid surface

diffusion to a final bound state. Previous MD simulations of a similar rapid cyclic urea binder with a $k_{on}^{exp} \sim 10^9$ $M^{-1}s^{-1}$ have suggested a binding mechanism where the ligand associates to the semi-open conformation and then induces an open conformation.[99,100] However, it is also possible that the DMP323 measurement is an experimental artifact due to upper limits for reliable association rate constant measurements by SPR.[97]

Although, HIV-1 protease ligands likely exhibit several binding mechanisms, other proteins with gated binding sites may be largely modulated by the gating itself. In such cases, our approach may extend beyond suggested classification of different binding mechanisms and allow the computation of association rate constants that correlate with the overall experimental rate constants for different inhibitors. Finally, because our method only requires computing the apo-protein kinetics once from atomistic simulations and then using the extracted conformations for simulating the binding of each ligand using computationally inexpensive independent Brownian dynamics simulations that can be run in parallel, it could be effectively scaled for such protein systems for high-throughput ligand ranking.

Acknowledgement

The authors thank Dr. Neil Bruce for assistance with SDA and discussions on the computation of rate constants, Dr. Gaurav Ganotra and Dr. Ariane Nunes-Alves for assistance in parameterizing ligand partial charges and diffusion coeffiecients, respectively, Dr. Ricard Argelaguet and Dr. Stefan Richter for technical assistance, Prof. Gianni De Fabritiis for sharing previous simulation trajectories from the GPUGRID project and Dr. Toni Giorgino, Prof. Gilles Mirambeau, Prof. Andreas Meyerhans and Prof. Ulrich Schwarz for discussions. This work was supported by the EU/EFPIA Innovative Medicines Initiative (IMI) Joint Undertaking, K4DD (grant no. 115366 to R.C.W.), the European Union Horizon 2020 Framework Programme (https://ec.europa.eu/programmes/horizon2020/en) for Research and Innovation under the Specific Grant Agreement no. 720270 (Human Brain Project SGA1, to

R.C.W.), the Volkswagen Foundation "Experiment! Funding Initiative" (grant no. 93874 to S.K.S.), the amfAR Mathilde Krim Fellowship in Basic Biomedical Research (grant no. 108680 to S.K.S.) and the Klaus Tschira Foundation.

Supporting Information Available

Figure S1. (A) Generalized Chapman-Kolmogorov test and (B) Distribution of microstate observation probabilities (χ) projected onto the λ-space.

Table S1. Top five most likely microstates that represent each macrostate.

Figure S2. Centroids of most probable microstate of each macrostate in the λ-space.

Table S2. Hydrogen bonds between flap and alternate monomers.

Figure S3. Cross sections of superimposed volumetric density of natural peptide and protein from a complex of HIV-1 protease.

Table S3. Sidechain-sidechain contacts between flap and alternate monomers.

Figure S4. Representative docked structures of peptidomimetic ligands to macrostates of HIV-1 protease.

Data. Simulation trajectories and output data pertaining to the results of the manuscript together with scripts for result generation and further analysis can be found on Zenodo at DOI: 10.5281/zenodo.5006612. Additional analysis scripts related to this data set can be found on Github at: https://github.com/kashifsadiq/hiv1pr-msm/.

References

(1) de Ruiter, A.; Oostenbrink, C. Advances in the calculation of binding free energies. *Current opinion in structural biology* **2020**, *61*, 207–212.

(2) Swinney, D. C. The role of binding kinetics in therapeutically useful drug action. *Current opinion in drug discovery & development* **2009**, *12*, 31–39.

(3) others,, et al. Kinetics for Drug Discovery: an industry-driven effort to target drug residence time. *Drug Discovery Today* **2017**, *22*, 896–911.

(4) Collins, F. C.; Kimball, G. E. Diffusion-controlled reaction rates. *Journal of colloid science* **1949**, *4*, 425–437.

(5) Stroppolo, M.; Falconi, M.; Caccuri, A.; Desideri, A. Superefficient enzymes. *Cellular and Molecular Life Sciences CMLS* **2001**, *58*, 1451–1460.

(6) Barzykin, A.; Seki, K.; Tachiya, M. Kinetics of diffusion-assisted reactions in micro-heterogeneous systems. *Advances in colloid and interface science* **2001**, *89*, 47–140.

(7) Schlosshauer, M.; Baker, D. A general expression for bimolecular association rates with orientational constraints. *The Journal of Physical Chemistry B* **2002**, *106*, 12079–12083.

(8) Northrup, S. H.; Erickson, H. P. Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proceedings of the National Academy of Sciences* **1992**, *89*, 3338–3342.

(9) Song, Y.; Zhang, Y.; Shen, T.; Bajaj, C. L.; McCammon, J. A.; Baker, N. A. Finite element solution of the steady-state Smoluchowski equation for rate constant calculations. *Biophysical Journal* **2004**, *86*, 2017–2029.

(10) Shaul, Y.; Schreiber, G. Exploring the charge space of protein–protein association: a proteomic study. *Proteins: Structure, Function, and Bioinformatics* **2005**, *60*, 341–352.

(11) Wade, R. C.; Gabdoulline, R. R.; Lüdemann, S. K.; Lounnas, V. Electrostatic steering and ionic tethering in enzyme–ligand binding: Insights from simulations. *Proceedings of the National Academy of Sciences* **1998**, *95*, 5942–5949.

(12) Gabdoulline, R. R.; Kummer, U.; Olsen, L. F.; Wade, R. C. Concerted simulations reveal how peroxidase compound III formation results in cellular oscillations. *Biophysical journal* **2003**, *85*, 1421–1428.

(13) Pang, X.; Zhou, H.-X. Rate constants and mechanisms of protein–ligand binding. *Annual review of biophysics* **2017**, *46*, 105–130.

(14) Boehr, D. D.; Nussinov, R.; Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology* **2009**, *5*, 789–796.

(15) Silva, D.-A.; Bowman, G. R.; Sosa-Peinado, A.; Huang, X. A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS Comput Biol* **2011**, *7*, e1002054.

(16) Changeux, J.-P.; Edelstein, S. Conformational selection or induced fit? 50 years of debate resolved. *F1000 biology reports* **2011**, *3*.

(17) Hammes, G. G.; Chang, Y.-C.; Oas, T. G. Conformational selection or induced fit: a flux description of reaction mechanism. *Proceedings of the National Academy of Sciences* **2009**, *106*, 13737–13741.

(18) Greives, N.; Zhou, H.-X. Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. *Proceedings of the National Academy of Sciences* **2014**, *111*, 10197–10202.

(19) Zhou, H.-X. From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions. *Biophysical journal* **2010**, *98*, L15–L17.

(20) Ou, L.; Matthews, M.; Pang, X.; Zhou, H.-X. The dock-and-coalesce mechanism for the association of a WASP disordered region with the Cdc42 GTP ase. 2017.

(21) Qin, S.; Pang, X.; Zhou, H.-X. Automated prediction of protein association rate constants. *Structure* **2011**, *19*, 1744–1751.

(22) Zhou, H.-X. Intrinsic disorder: signaling via highly specific but short-lived association. *Trends in biochemical sciences* **2012**, *37*, 43–48.

(23) Zhou, H.-X.; Pang, X.; Lu, C. Rate constants and mechanisms of intrinsically disordered proteins binding to structured targets. *Physical Chemistry Chemical Physics* **2012**, *14*, 10466–10476.

(24) Rogers, J. M.; Oleinikovas, V.; Shammas, S. L.; Wong, C. T.; De Sancho, D.; Baker, C. M.; Clarke, J. Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proceedings of the National Academy of Sciences* **2014**, *111*, 15420–15425.

(25) McCammon, J. A.; Northrup, S. H. Gated binding of ligands to proteins. *Nature* **1981**, *293*, 316–317.

(26) Szabo, A.; Shoup, D.; Northrup, S. H.; McCammon, J. A. Stochastically gated diffusion-influenced reactions. *The Journal of Chemical Physics* **1982**, *77*, 4484–4493.

(27) Chang, C.-E.; Shen, T.; Trylska, J.; Tozzini, V.; McCammon, J. A. Gated binding of ligands to HIV-1 protease: Brownian dynamics simulations in a coarse-grained model. *Biophysical journal* **2006**, *90*, 3880–3885.

(28) Zhou, H.-X. Theory of the diffusion-influenced substrate binding rate to a buried and gated active site. *The Journal of chemical physics* **1998**, *108*, 8146–8154.

(29) Zhou, H.-X.; Wlodek, S. T.; McCammon, J. A. Conformation gating as a mechanism for enzyme specificity. *Proceedings of the National Academy of Sciences* **1998**, *95*, 9280–9283.

(30) Zhou, H.-X.; McCammon, J. A. The gates of ion channels and enzymes. *Trends in biochemical sciences* **2010**, *35*, 179–185.

(31) McCammon, J. A. Gated diffusion-controlled reactions. *BMC biophysics* **2011**, *4*, 1–5.

(32) Northrup, S. H.; Allison, S. A.; McCammon, J. A. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *The Journal of Chemical Physics* **1984**, *80*, 1517–1524.

(33) Gabdoulline, R. R.; Wade, R. C. Simulation of the diffusional association of barnase and barstar. *Biophysical journal* **1997**, *72*, 1917–1929.

(34) Gabdoulline, R. R.; Wade, R. C. Protein-protein association: investigation of factors influencing association rates by Brownian dynamics simulations. *Journal of molecular biology* **2001**, *306*, 1139–1155.

(35) Luty, B. A.; El Amrani, S.; McCammon, J. A. Simulation of the bimolecular reaction between superoxide and superoxide dismutase: synthesis of the encounter and reaction steps. *Journal of the American Chemical Society* **1993**, *115*, 11874–11877.

(36) Sung, J. C.; Wynsberghe, A. W. V.; Amaro, R. E.; Li, W. W.; McCammon, J. A. Role of secondary sialic acid binding sites in influenza N1 neuraminidase. *Journal of the American Chemical Society* **2010**, *132*, 2883–2885.

(37) Zeller, F.; Luitz, M. P.; Bomblies, R.; Zacharias, M. Multiscale simulation of receptor–drug association kinetics: Application to neuraminidase inhibitors. *Journal of chemical theory and computation* **2017**, *13*, 5097–5105.

(38) Votapka, L. W.; Jagger, B. R.; Heyneman, A. L.; Amaro, R. E. SEEKR: simulation enabled estimation of kinetic rates, a computational tool to estimate molecular kinetics and its application to trypsin–benzamidine binding. *The Journal of Physical Chemistry B* **2017**, *121*, 3597–3606.

(39) Jagger, B. R.; Lee, C. T.; Amaro, R. E. Quantitative ranking of ligand binding kinetics with a multiscale milestoning simulation approach. *The journal of physical chemistry letters* **2018**, *9*, 4941–4948.

(40) Jagger, B. R.; Ojha, A. A.; Amaro, R. E. Predicting Ligand Binding Kinetics Using a Markovian Milestoning with Voronoi Tessellations Multiscale Approach. *Journal of Chemical Theory and Computation* **2020**, *16*, 5348–5357.

(41) Bruce, N. J.; Ganotra, G. K.; Kokh, D. B.; Sadiq, S. K.; Wade, R. C. New approaches for computing ligand–receptor binding kinetics. *Current opinion in structural biology* **2018**, *49*, 1–10.

(42) Nunes-Alves, A.; Kokh, D. B.; Wade, R. C. Recent progress in molecular simulation methods for drug binding kinetics. *Current Opinion in Structural Biology* **2020**, *64*, 126–133.

(43) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry* **2016**, *59*, 4035–4061.

(44) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **2011**, *108*, 10184–10189.

(45) Gu, S.; Silva, D.-A.; Meng, L.; Yue, A.; Huang, X. Quantitatively characterizing the ligand binding mechanisms of choline binding protein using Markov state model analysis. *PLoS Comput Biol* **2014**, *10*, e1003767.

(46) Plattner, N.; Noé, F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nature communications* **2015**, *6*, 1–10.

(47) Paul, F.; Wehmeyer, C.; Abualrous, E. T.; Wu, H.; Crabtree, M. D.; Schöneberg, J.; Clarke, J.; Freund, C.; Weikl, T. R.; Noé, F. Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations. *Nature communications* **2017**, *8*, 1–10.

(48) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature chemistry* **2017**, *9*, 1005.

(49) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19011–19016.

(50) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9 (1- 39). *Journal of the American Chemical Society* **2010**, *132*, 1526–1528.

(51) Bowman, G. R.; Pande, V. S. Protein folded states are kinetic hubs. *Proceedings of the National Academy of Sciences* **2010**, *107*, 10890–10895.

(52) Sborgi, L.; Verma, A.; Piana, S.; Lindorff-Larsen, K.; Cerminara, M.; Santiveri, C. M.; Shaw, D. E.; de Alba, E.; MunÌČoz, V. Interaction networks in protein folding via atomic-resolution experiments and long-time-scale molecular dynamics simulations. *Journal of the American Chemical Society* **2015**, *137*, 6506–6516.

(53) Sadiq, S. K.; Noé, F.; De Fabritiis, G. Kinetic characterization of the critical step in HIV-1 protease maturation. *Proceedings of the National Academy of Sciences* **2012**, *109*, 20449–20454.

(54) Xia, J.; Deng, N.-j.; Levy, R. M. NMR relaxation in proteins with fast internal motions and slow conformational exchange: model-free framework and Markov state simulations. *The Journal of Physical Chemistry B* **2013**, *117*, 6625–6634.

(55) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nature communications* **2014**, *5*, 1–11.

(56) Banerjee, R.; Yan, H.; Cukier, R. I. Conformational transition in signal transduction: metastable states and transition pathways in the activation of a signaling protein. *The Journal of Physical Chemistry B* **2015**, *119*, 6591–6602.

(57) Perkett, M. R.; Hagan, M. F. Using Markov state models to study self-assembly. *The Journal of chemical physics* **2014**, *140*, 214101.

(58) Chodera, J. D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology* **2014**, *25*, 135–144.

(59) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *The Journal of chemical physics* **2013**, *139*, 184114.

(60) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics* **2011**, *134*, 174105.

(61) Husic, B. E.; Pande, V. S. Markov state models: From an art to a science. *Journal of the American Chemical Society* **2018**, *140*, 2386–2396.

(62) Noé, F.; Nuske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation* **2013**, *11*, 635–655.

(63) Naritomi, Y.; Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *The Journal of chemical physics* **2011**, *134*, 065101.

(64) Schwantes, C. R.; Pande, V. S. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *Journal of chemical theory and computation* **2013**, *9*, 2000–2009.

(65) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **2013**, *139*, 015102.

(66) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced modeling via network theory: Adaptive sampling of Markov state models. *Journal of chemical theory and computation* **2010**, *6*, 787–794.

(67) Doerr, S.; De Fabritiis, G. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *Journal of chemical theory and computation* **2014**, *10*, 2064–2069.

(68) Hruska, E.; Abella, J. R.; Nüske, F.; Kavraki, L. E.; Clementi, C. Quantitative comparison of adaptive sampling methods for protein dynamics. *The Journal of chemical physics* **2018**, *149*, 244119.

(69) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature communications* **2018**, *9*, 1–11.

(70) Cavalli, A.; Spitaleri, A.; Saladino, G.; Gervasio, F. L. Investigating drug–target asso-

ciation and dissociation mechanisms using metadynamics-based algorithms. *Accounts of chemical research* **2015**, *48*, 277–285.

(71) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the National Academy of Sciences* **2015**, *112*, E386–E391.

(72) Pietrucci, F.; Marinelli, F.; Carloni, P.; Laio, A. Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *Journal of the American Chemical Society* **2009**, *131*, 11811–11818.

(73) Wang, J.; Miao, Y. Peptide Gaussian accelerated molecular dynamics (Pep-GaMD): Enhanced sampling and free energy and kinetics calculations of peptide binding. *The Journal of Chemical Physics* **2020**, *153*, 154109.

(74) Dickson, A.; Lotz, S. D. Multiple ligand unbinding pathways and ligand-induced destabilization revealed by WExplore. *Biophysical journal* **2017**, *112*, 620–629.

(75) Mollica, L.; Decherchi, S.; Zia, S. R.; Gaspari, R.; Cavalli, A.; Rocchia, W. Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Scientific reports* **2015**, *5*, 1–12.

(76) Teo, I.; Mayne, C. G.; Schulten, K.; Lelièvre, T. Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time. *Journal of chemical theory and computation* **2016**, *12*, 2983–2989.

(77) Miao, Y. Acceleration of biomolecular kinetics in Gaussian accelerated molecular dynamics. *The Journal of chemical physics* **2018**, *149*, 072308.

(78) Wlodawer, A.; Miller, M.; Jaskólski, M.; Sathyanarayana, B. K.; Baldwin, E.; Weber, I. T.; Selk, L. M.; Clawson, L.; Schneider, J.; Kent, S. B. H. Conserved folding

in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* **1989**, *245*, 616–621.

(79) Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C. A. How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. *J. Mol. Biol.* **2000**, *301*, 1207–1220.

(80) 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.

(81) Pettit, S. C.; Sheng, N.; Tritch, R.; Erickson-Viitanen, S.; Swanstrom, R. The regulation of sequential processing of HIV-1 gag by the viral protease. *Advances in Experimental Medicine and Biology* **1998**, *436*, 15–25.

(82) Pettit, S. C.; Henderson, G. J.; Schiffer, C. A.; Swanstrom, R. Replacement of the P1 amino acid of human immunodeficiency virus type 1 gag processing sites can inhibit or enhance the rate of cleavage by the viral protease. *J. Virol.* **2002**, *76*, 10226–10233.

(83) Pettit, S. C.; Clemente, J. C.; Jeung, J. A.; Dunn, B. M.; Kaplan, A. H. Ordered processing of the human immunodeficiency virus type 1 gagpol precursor is influenced by the context of the embedded viral protease. *Journal of Virology* **2005**, *79*, 10601–10607.

(84) Palese, L. L. Conformations of the HIV-1 protease: a crystal structure data set analysis. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2017**, *1865*, 1416–1422.

(85) Hassan, S.; Srikakulam, S. K.; Chandramohan, Y.; Thangam, M.; Muthukumar, S.; Gayathri Devi, P.; Hanna, L. E. Exploring the conformational landscapes of HIV protease structural ensembles using principal component analysis. *Proteins: Structure, Function, and Bioinformatics* **2018**, *86*, 990–1000.

(86) Spinelli, S.; Liu, Q.; Alzari, P.; Hirel, P.; Poljak, R. The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie* **1991**, *73*, 1391–1396.

(87) Martin, P.; Vickrey, J. F.; Proteasa, G.; Jimenez, Y. L.; Wawrzak, Z.; Winters, M. A.; Merigan, T. C.; Kovari, L. C. "Wide-open" 1.3 A structure of a multidrug-resistant HIV-1 protease as a drug target. *Structure* **2005**, *13*, 1887–1895.

(88) Torbeev, V. Y.; Raghuraman, H.; Hamelberg, D.; Tonelli, M.; Westler, W. M.; Perozo, E.; Kent, S. B. Protein conformational dynamics in the mechanism of HIV-1 protease catalysis. *Proceedings of the National Academy of Sciences* **2011**, *108*, 20982–20987.

(89) Ishima, R.; Freedberg, D. I.; Wang, Y. X.; Louis, J. M.; Torchia, D. A. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure* **1999**, *7*, 1047–1055.

(90) Freedberg, D. I.; Ishima, R.; Jacob, J.; Y-X.Wang,; Kustanovich, I.; Louis, J. M.; Torchia, D. A. Rapid structural fluctuations of the free HIV protease flaps in solution: Relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Sci.* **2002**, *11*, 221–232.

(91) Roche, J.; Louis, J. M.; Bax, A. Conformation of inhibitor-free HIV-1 protease derived from nmr spectroscopy in a weakly oriented solution. *ChemBioChem* **2015**, *16*, 214–218.

(92) Sadiq, S. K.; Wan, S.; Coveney, P. V. Insights into a mutation-assisted lateral drug escape mechanism from the HIV-1 protease active site. *Biochemistry* **2007**, *46*, 14865–14877.

(93) Chang, C.-E. A.; Trylska, J.; Tozzini, V.; Andrew McCammon, J. Binding pathways of ligands to HIV-1 protease: coarse-grained and atomistic simulations. *Chemical biology & drug design* **2007**, *69*, 5–13.

(94) Hornak, V.; Okur, A.; Rizzo, R. C.; Simmerling, C. HIV-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. *J. Am. Chem. Soc.* **2006**, *128*, 2812–2813.

(95) Sadiq, S. K.; De Fabritiis, G. Explicit solvent dynamics and energetics of HIV-1 protease flap opening and closing. *Proteins* **2010**, *78*, 2873–2885.

(96) Deng, N.-j.; Zheng, W.; Gallicchio, E.; Levy, R. M. Insights into the dynamics of HIV-1 protease: a kinetic network model constructed from atomistic simulations. *Journal of the American Chemical Society* **2011**, *133*, 9387–9394.

(97) Markgren, P.-O.; Schaal, W.; Hämäläinen, M.; Karlén, A.; Hallberg, A.; Samuelsson, B.; Danielson, U. H. Relationships between structure and interaction kinetics for HIV-1 protease inhibitors. *Journal of medicinal chemistry* **2002**, *45*, 5430–5439.

(98) Kang, M.; Roberts, C.; Cheng, Y.; Chang, C.-e. A. Gating and intermolecular interactions in ligand-protein association: coarse-grained modeling of HIV-1 protease. *Journal of chemical theory and computation* **2011**, *7*, 3438–3446.

(99) Huang, Y.-m. M.; Raymundo, M. A. V.; Chen, W.; Chang, C.-e. A. Mechanism of the association pathways for a pair of fast and slow binding ligands of HIV-1 protease. *Biochemistry* **2017**, *56*, 1311–1323.

(100) Miao, Y.; Huang, Y.-m. M.; Walker, R. C.; McCammon, J. A.; Chang, C.-e. A. Ligand binding pathways and conformational transitions of the HIV protease. *Biochemistry* **2018**, *57*, 1533–1541.

(101) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A software package for estimation, validation, and analysis of Markov models. *Journal of chemical theory and computation* **2015**, *11*, 5525–5542.

(102) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; ; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(103) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(104) Hyland, L. J.; Tomaszek, T. A., Jr.; Meek, T. D. Human immunodeficiency virus-1 protease. 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanism. *Biochemistry* **1991**, *30*, 8454–8463.

(105) Wang, Y.-X.; Freedberg, D. I.; Yamazaki, T.; Wingfield, P. T.; Stahl, S. J.; Kaufman, J. D.; Kiso, Y.; Torchia, D. A. Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the Asymmetric drug KNI-272. *Biochemistry* **1996**, *35*, 9945–9950.

(106) Kovalskyy, D.; Dubyna, V.; Mark, A. E.; Korenelyuk, A. A molecular dynamics study of the structural stability of HIV-1 protease under physiological conditions: The role of Na+ ions in stabilizing the active site. *Proteins: Struct. Funct. Bioinf.* **2005**, *58*, 450–458.

(107) Wittayanarakul, K.; Hannongbua, S.; Feig, M. Accurate prediction of protonation state as a prerequisite for reliable MM-PB(GB)SA binding free energy calculations of HIV-1 protease inhibitors. *J. Comput. Chem.* **2008**, *29*, 673–685.

(108) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(109) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer

programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Comm.* **1995**, *91*, 1–41.

(110) 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.

(111) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theor. Comput.* **2009**, *5*, 1632–1639.

(112) Meagher, K. L.; Carlson, H. A. Solvation Influences Flap Collapse in HIV-1 Protease. *Proteins: Struct. Funct. Bioinf.* **2005**, *58*, 119–125.

(113) Harvey, M.; De Fabritiis, G. An implementation of the smooth particle mesh ewald method on gpu hardware. *Journal of Chemical Theory and Computation* **2009**, *5*, 2371–2377.

(114) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–9593.

(115) 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.

(116) Feenstra, K.; Hess, B.; Berendsen, H. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **1999**, *20*, 786–798.

(117) Sadiq, S. K.; Coveney, P. V. Computing the role of near attack conformations in an enzyme-catalyzed nucleophilic bimolecular reaction. *Journal of chemical theory and computation* **2015**, *11*, 316–324.

(118) Sadiq, S. K. Fine-Tuning of Sequence Specificity by Near Attack Conformations in Enzyme-Catalyzed Peptide Hydrolysis. *Catalysts* **2020**, *10*, 684.

(119) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry* **2011**, *32*, 2319–2327.

(120) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015**, *11*, 3696–3713.

(121) Ganotra, G. K.; Wade, R. C. Prediction of drug–target binding kinetics by comparative binding energy analysis. *ACS medicinal chemistry letters* **2018**, *9*, 1134–1139.

(122) Case, D.; Ben-Shalom, I.; Brozell, S.; Cerutti, D.; Cheatham III, T.; Cruzeiro, V.; Darden, T.; Duke, R.; Ghoreishi, D.; Gilson, M.; Gohlke, H.; Goetz, A.; Greene, D.; Harris, R.; Homeyer, N.; Huang, Y.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C.; Smith, J.; Salomon-Ferrer, R.; Swails, J.; Walker, R.; Wang, J.; Wei, H.; Wolf, R.; Wu, X.; Xiao, L.; DM, Y.; Kollman, P. AMBER 2018. *University of California, San Francisco* **2018**,

(123) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences* **2001**, *98*, 10037–10041.

(124) Martinez, M.; Bruce, N. J.; Romanowska, J.; Kokh, D. B.; Ozboyaci, M.; Yu, X.; Öztürk, M. A.; Richter, S.; Wade, R. C. SDA 7: A modular and parallel implemen-

tation of the simulation of diffusional association software. *Journal of computational chemistry* **2015**, *36*, 1631–1645.

(125) Gabdoulline, R. R.; Wade, R. C. Effective charges for macromolecules in solvent. *The Journal of Physical Chemistry* **1996**, *100*, 3868–3878.

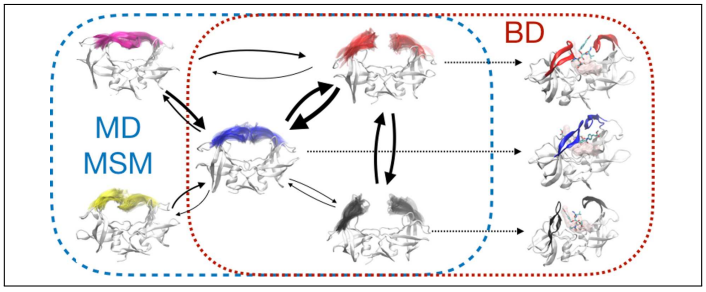
(126) Ganotra, G. K. Computational studies of drug-binding kinetics. Ph.D. thesis, Heidelberg University, 2020.

(127) de la Torre, J. G.; Huertas, M. L.; Carrasco, B. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophysical journal* **2000**, *78*, 719–730.

(128) Karthik, S.; Senapati, S. Dynamic flaps in HIV-1 protease adopt unique ordering at different stages in the catalytic cycle. *Proteins: Structure, Function, and Bioinformatics* **2011**, *79*, 1830–1840.

(129) Trylska, J.; Tozzini, V.; Chang, C.-e. A.; McCammon, J. A. HIV-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics. *Biophys J* **2007**, *92*, 4179–4187.

(130) Potempa, M.; Nalivaika, E.; Ragland, D.; Lee, S.-K.; Schiffer, C. A.; Swanstrom, R. A direct interaction with RNA dramatically enhances the catalytic activity of the HIV-1 protease in vitro. *Journal of molecular biology* **2015**, *427*, 2360–2378.

(131) Lyonnais, S.; Sadiq, S. K.; Lorca-Oró, C.; Dufau, L.; Nieto-Marquez, S.; Escriba, T.; Gabrielli, N.; Tan, X.; Ouizougun-Oubari, M.; Okoronkwo, J.; Reboud-Ravaux, M.; Gatell, J. M.; Marquet, R.; Paillart, J.-C.; Meyerhans, A.; TisnÃľ, C.; Gorelick, R. J.; Mirambeau, G. The HIV-1 ribonucleoprotein dynamically regulates its condensate behavior and drives acceleration of protease activity through membrane-less granular phase-separation. *bioRxiv* **2019**, 528638.

Graphical TOC Entry