Extended-Ensemble Docking to Probe Evolution of Ligand Binding Sites in Proteins Undergoing Large-Scale Structural Changes for Their Function

Karan Kapoor, Sundar Thangapandian, and Emad Tajkhorshid*

Theoretical and Computational Biophysics Group, NIH Center for Macromolecular Modeling and Bioinformatics, Beckman Institute for Advanced Science and Technology, Department of Biochemistry, and Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

E-mail: emad@illinois.edu

Abstract

Proteins can sample a broad energetic landscape as they undergo conformational transition between different functional states. As key players in almost all cellular processes, proteins are important drug targets. Considering the different conformational states of proteins is therefore central for a successful drug-design strategy. Here we introduce a novel docking protocol, termed as extended-ensemble docking, pertaining to proteins that undergo large-scale (global) conformational changes during their function. In its application to multidrug ABC-transporter P-glycoprotein (Pgp), extensive non-equilibrium molecular dynamics simulations employing system-specific collective variables capturing the alternate access mechanism of Pgp, are first used to construct the transition cycle of the transporter. An extended set of conformational

states representing the full transition between the inward- and the outward-facing states of Pgp, is then used to seed high-throughput docking calculations of a set of known substrates, non-substrates, and modulators of the transporter. Large differences are observed in the predicted binding affinities in the conformational ensemble, with compounds showing stronger binding affinities in intermediate conformations compared to the starting crystal structure. Hierarchical clustering of the individual binding modes of the different compounds shows all ligands preferably bind to the large central cavity of the protein, formed at the apex of the transmembrane domain (TMD), whereas only small binding populations are observed in the previously described R and H sites present in the individual TMD leaflets. The central cavity is further clustered into two major subsites: first subsite preferably binds transported substrates and high-affinity inhibitors, whereas the second subsite shows preference for larger substrates and low-affinity modulators. These central sites along with the low-affinity interaction sites present in the individual TMD leaflets may respectively correspond to the proposed high- and low-affinity binding sites in Pgp. We propose further optimization strategy for developing more potent inhibitor of Pgp, based on increasing its specificity to the extended-ensemble of the protein instead of using a single protein structure, as well as its selectivity for the high-affinity binding site. In contrast to earlier in-silico studies using single static structures of Pgp, our results show good correlation with other experimental studies, pointing to the importance of incorporating the global conformational flexibility of proteins in future drug-discovery endeavors.

Introduction

Modern drug discovery is centered around the identification of suitable protein targets that play important roles in a specific disease, and the development of small molecules designed to either attenuate or promote the function of those target proteins.[1] Advances in structural determination techniques such as cryo-EM have led to a rapid increase in structurally known proteins that can help guide the drug discovery process.[2] These experimentally resolved

structures generally represent the starting or end states in the functional cycle of a protein, as the conformational intermediates (usually shorter-lived and higher-energy states) arising during the function are often inaccessible under the experimental conditions.[3] This reduces the possible structural diversity of the target that can be utilized in designing more efficient drug discovery strategies. Generating and targeting these conformational intermediates still remains a major challenge in successful drug discovery campaigns.

Computational methods are increasingly used to complement the experimental drug discovery setup.[4,5] Structure-based drug discovery methods like docking utilize the available three-dimensional structure of the target protein to make predictions for the best binding ligand candidates to the target.[6,7] This reduces the number of chemotypes to be experimentally tested, compared to high-throughput screening methods utilized alone without any prior knowledge of the possible interaction between the target and the small molecules.[8]

Traditionally, and given the structural determination challenges described above, only a single experimental (e.g., crystal or cryo-EM) structure or a single computational model (e.g., a homology model) of the target protein has been used for docking purposes, an approach we refer to as *s*ingle-point docking[9] (Fig. 1). As stated above, these structures generally represent a static snapshot of a major functional state of the target protein. More recently, the single-point docking approach has been improved to take into account thermal and conformational fluctuations of such major conformational states of the protein in an approach referred to as *e*nsemble docking; the approach has been shown to further enhance both the quality and the efficiency of lead predictions[10] and successfully applied to a number of protein targets[11,12] (Fig. 1). Ensemble docking makes use of molecular dynamics (MD) simulations for sampling the local conformational basin of the target generated in the vicinity of the starting protein structure.[13,14]

In the case of larger and more flexible protein targets though, which often display a complex conformational landscape with multiple local minima,[3,15] the limited timescales of classical equilibrium MD simulations may not cover the large spatial and temporal distances

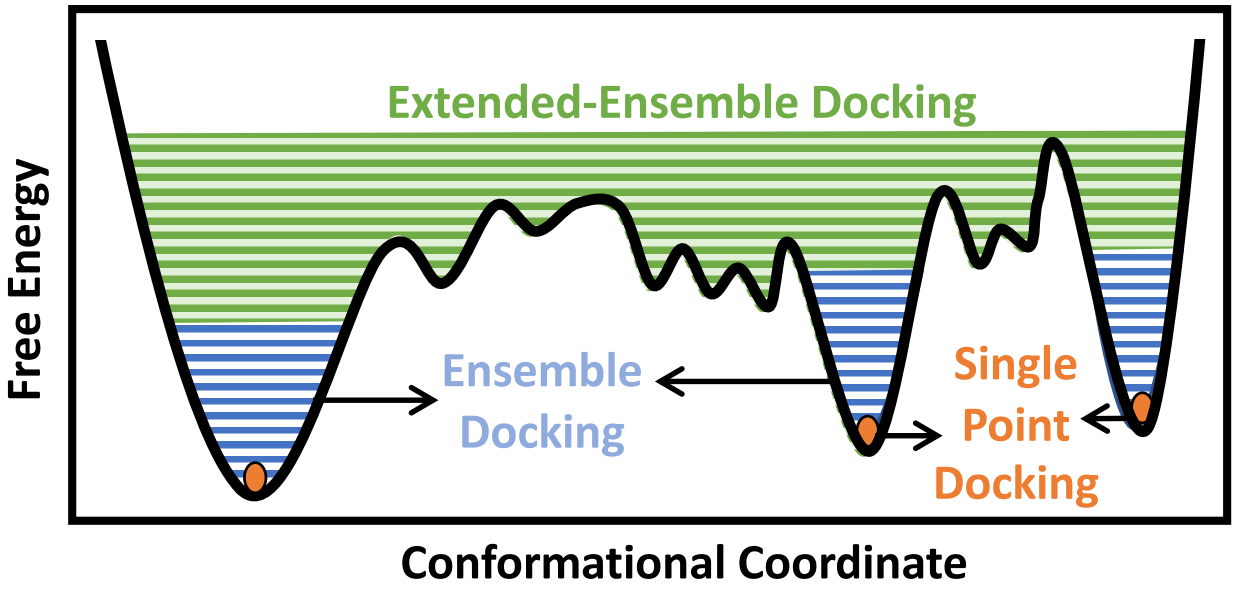

Figure 1: **Conformational spaces targeted by different docking approaches.** Classifying docking approaches in terms of the targeted conformational space of the protein. Single-point docking utilizes a single structure of the protein target, restricting the sampling to a single point (orange dots) in the conformational landscape of the protein. Ensemble docking utilizes an ensemble of protein structures, often generated using MD simulations, taking into account the local conformational space in the vicinity of the starting structure (blue lines). Extended-ensemble docking, the method introduced here, aims at taking into account the full functional cycle of the protein, generated, e.g., through the application of biased techniques to transition between the major functional states of the protein (green lines).

constituting the conformational heterogeneity of the target protein.[16] Non-equilibrium methods, on the other hand, allow us to expand our sampling of the phase space and to visit intermediate conformations formed during the transition between the major functional states of a protein.[17,18] This forms the basis of what we term as *extended-ensemble docking*, that uses the extended sampling offered by non-equilibrium MD for generating an extended-ensemble of the protein, sampling the full conformational transition pathway between its major states (Fig. 1). The extended-ensemble of the protein can then be used to seed high-throughput docking calculations for specific drug-discovery applications.

Here we report the first application of this approach to investigation of ligand binding in the multidomain ABC exporter, P-glycoprotein (Pgp),[19] a transporter that is over-expressed in the plasma membrane of cancer cells and a major cause for the development of multidrug resistance.[20,21] The transporter accomplishes its function as a cellular 'vacuum cleaner'

through a large central drug binding pocket (DBP), formed by its two opposing transmembrane domains (TMDs), which is known to bind different classes of molecules (Fig. S1).[22,23] The transporter follows the general 'alternating-access' model[24,25] for its function, where it alternates between the inward-facing (IF) and outward-facing (OF) states, alternatively exposing the DBP to the intracellular and extracellular environments. The large-scale structural transitions of the transporter are fueled by and coupled to ATP-driven dimerization and opening of the two nucleotide binding domains (NBDs) connected to the TMDs.[26–29]

A number of stduies employing molecular docking,[30–32] pharmacophore mapping,[33] machine learning,[34] and MD-based approaches[35] have tried to characterize the binding sites in Pgp. However, these have generated vastly conflicting results to each other and to earlier mutational studies;[36–38] it still remains debatable as to whether diverse molecules bind to distinct or overlapping binding sites in Pgp. A major limitation of the above in-silico studies is that they neglect the structural flexibility inherent to a flexible multidomain protein such as Pgp by targeting only a single crystal structure/homology model, providing only a snapshot of the myriad of possibilities of how different molecules can bind to the protein.

In the presented study, non-equilibrium, driven MD simulations, employing a set of system-specific reaction coordinates describing the alternate-access transitions of ABC transporters, is first used to sample the full transition cycle of Pgp between the IF and OF states. Docking of a set of substrates, non-substrate and modulators on the generated extended-ensemble of the protein reveals that compounds may bind with different binding affinities to different conformational states/intermediates of the protein. Preference of compounds for different "subsites" formed in the central binding cavity of Pgp, modulated by the conformational changes in the protein as it undergoes transition between the IF and OF states, may form the basis of broad substrate specificity of the transporter. Our results show close correlation with the previous experimental studies and additionally allow us to make predictions for developing more potent inhibitors of this multidrug resistance transporter.

Computational Approach

The main distinction of extended-ensemble docking when compared to other docking approaches is to first provide models for the intermediate conformational states of the protein that arise during the transition between structurally known functional states. Here, we generate the transition and the conformational intermediates involved through the application of non-equilibrium based methods. Then, instead of targeting a single experimental/modelled structure, the entire extended ensemble of the protein is targeted through high-throughput docking calculations (extended-ensemble docking). Furthermore, we also present a suitable protocol for clustering the predicted binding modes of compounds in the ensemble of structures in order to obtain physiologically relevant results related to the functional mechanism of the protein. Following are the details of the different steps involved in the protocol and its specific application to the multidrug transporter Pgp (Fig. 2).

I. Generating an extended conformational ensemble of the target protein

In the first step of the protocol, an extended ensemble of the target protein is generated. We start with the available (known/modeled) structures of major functional states (end states) of the protein. Experimentally resolved structures, e.g., through, crystallography or cryo-EM, might be affected by artifacts due to non-native environments used for fixing the biological sample.[39] In order to obtain relaxed structures, which are required for optimal sampling of the transition pathway computationally, one would need to first equilibrate them in a native environment, e.g., solution for globular proteins and a lipid bilayer for membrane proteins.

Subsequently, in order to describe the conformational transition between major functional states in a physiological context, we define a set of system-specific collective variables that capture the essential conformational changes in the protein during the transition. These collective variables need to be closely related to important/relevant structural features of

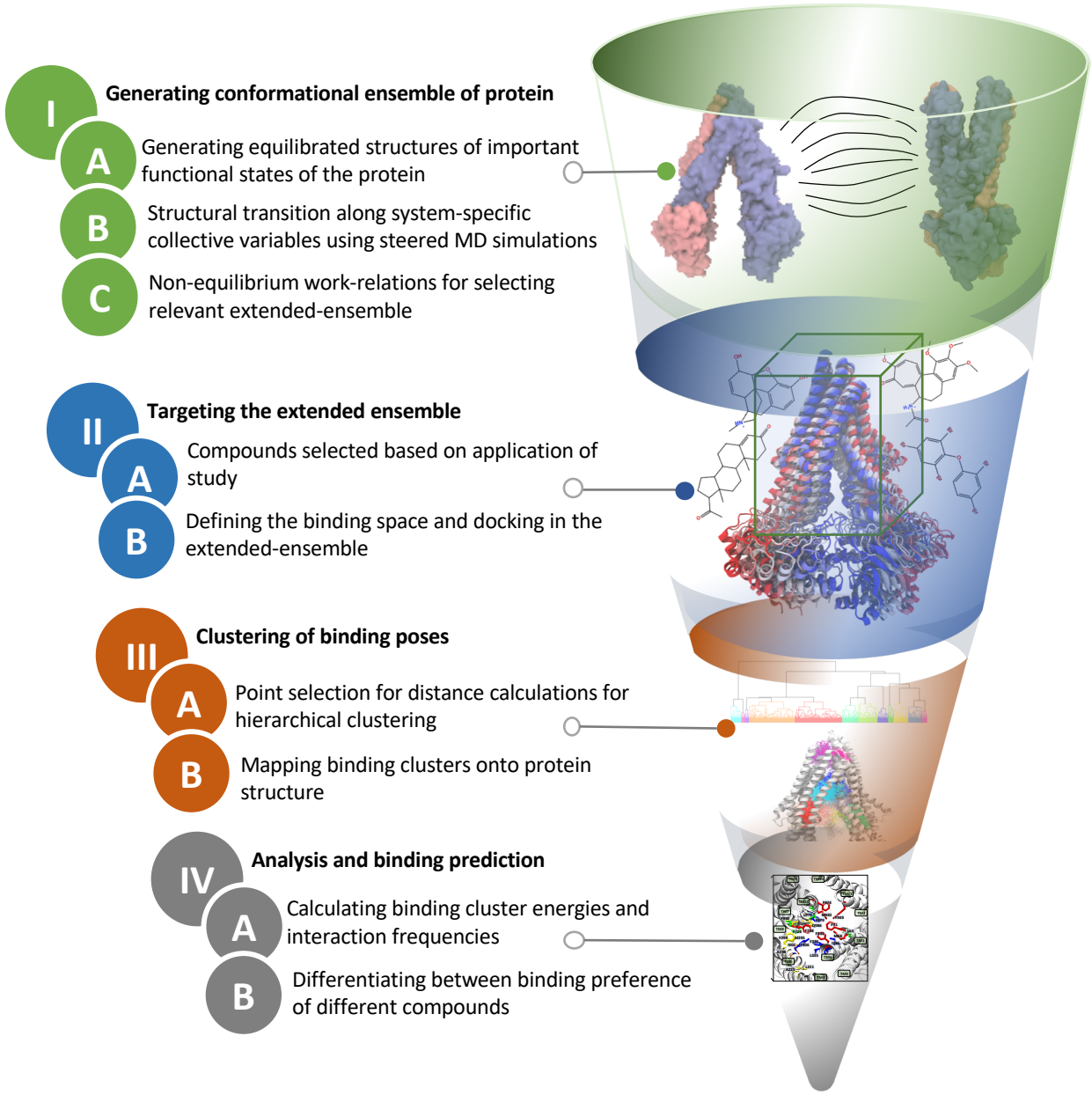

Figure 2: **Extended-ensemble docking protocol.** Flow-diagram showing different steps involved in the extended-ensemble docking protocol.

the different states of the protein; by applying time-dependent bias along these collective variable(s), for example by using non-equilibrium methods such as steered MD (SMD),[17,18] we can generate transition pathways connecting the end states.[40,41]

The amount of non-equilibrium work needed to induce the conformational change between the states during the driven transition provides an approximate quantitative measure for the relevance of the sampled transition pathway.[42,43] In addition, RMSD of the final structure obtained from the driven simulation with respect to the target structure provides a measure for the success of the prescribed pathway to complete the transition. Low RMSD to the target structure and low work values for a pathway point towards the efficiency and the quality of the transition pathway. The conformational ensemble of the protein can then be generated by selecting structures spanning the distance between the two end states, for example, by uniformly selecting representative conformations along the most optimal (lowest work) transition pathway.

In the following sections we describe each of these steps for the specific example of Pgp.

Equilibrated structures of functional states of Pgp

In the case of Pgp, crystallographic studies have been able to capture only the nucleotide-free IF states of the transporter thus far.[22,44–47] Additionally, these crystal structures show the NBDs of Pgp separated to different degrees, pointing to the conformational heterogeneity of the IF state in the transporter. MD simulations were thus employed to allow further sampling of the conformational space of the IF state. We had previously generated an equilibrated ensemble of PPgp IF structures,[48] using multiple independently run MD simulations of the crystal structure of mouse Pgp (PDB: 4M1M[22]) fully solvated and embedded in POPC/cholesterol lipid bilayers. As binding of ATP/Mg^{2+} to the NBDs of the transporter is a prerequisite for their dimerization and progression of the catalytic cycle of the transporter, ATP/Mg^{2+} were carefully placed into their respective binding sites in the NBDs following the protocol described by Wen et al.[49] before the subsequent equilibration of the protein.

Recently, a cryo-EM structure of Pgp in the OF state was published.[50] Due to the presence of a Q/E mutation at the catalytic site of the transporter, the structure has been suggested to be a 'dead-mutant', a non-catalytic form of Pgp incapable of substrate transport.[51] Other structural studies have instead suggested this structure to be a post-transport collapsed-occluded state of the transporter.[52] Additionally, studies using DEER spectroscopy have shown that Pgp samples a wider opening of the extracellular TMD gate in its OF state during the catalytic cycle,[29] not seen in the cryo-EM structure. Due to these reasons, we have generated our own equilibrated model of the OF state of Pgp using careful sequence alignment, homology modeling, and SMD simulations, and tested its stability in multiple MD simulations.[29] The distances between different regions of the transporter present in this state were found to match well with the spectroscopy studies described above.

System-specific collective variables for sampling conformational transition

Our previous work with the ABC transporters, including Pgp, has provided us with a set of collective variables shown to capture the alternate-access mechanism of ABC-transporters.[29,43] Conformational transition between the equilibrated IF and OF states of Pgp were carried out using these system-specific collective variables, capturing both the global conformational changes taking place in the protein during this transition, as well as local interactions crucial for the formation of the stable Pgp states (Fig. S2). These include two orientation-based collective variables, denoted as α and β, describing the closing of the TMD intracellular gate and the opening of the TMD extracellular gate during IF to OF transition, respectively. These orientation quaternions (α and β) were defined based on the bundling patterns of the Cα atoms of the TMD helices in IF and OF structures, respectively, and are associated with the relative rotation of the helices during transition between these two states. Additionally, one collective variable, denoted as SB, defines the formation of a salt bridge (K185-D993) interaction located in the middle of the TM helical region. For forming the correct dimerized state of the NBDs, 12 collective variables (collectively denoted as NBDi) defining the

distances between ATP and the interacting signature motif (LSGGQ) residues located on the opposite NBD, as well as X-loop interactions (D163-K1167 and E522-T806) that lock the NBDs with two of the intracellular helices located at the interface between the TMD and NBDs, were used (Table S1). All the target distances for NBDi collective variables were obtained from the high-resolution crystal structure of dimerized NBDs of HlyB (PDB: 1XEF[53]), an ABC exporter.

Structural transition using SMD simulations

We performed SMD simulations using different biasing-protocols to explore a wide-range of mechanistically distinct transition pathways in Pgp, each defined by a different sequential order of application of the four system-specific collective variables described above. These transition protocols are: P1 (α + NBDi + SB –> β), P2 (NBDi –> α + SB –> β), P3 (α + SB –> NBDi –> β), and P4 (α + NBDi + SB + β), where '+' denotes collective variables applied at the same time and '–>' denotes collective variables following after.

Multiple biased simulations were performed starting from 20 equilibrated ATP/Mg^{2+}-bound IF structures of Pgp embedded in explicit POPC:cholesterol lipid bilayers (described above) built using the CHARMGUI membrane builder module[54,55] and the modelled OF state as the target structure. The system was solvated and neutralized with Na^+ and Cl^- ions to achieve a net concentration of 150 mM. The final system contained $\sim$240,000 atoms with dimensions of $100\times100\times170$ Å^3. The simulations were carried out in an NPT ensemble, with the temperature maintained constant at 310 K using Langevin dynamics[56] with a damping coefficient of $\gamma = 0.5$ ps^{-1}, and the pressure maintained constant at 1 bar using the Nosé-Hoover Langevin piston method.[57,58] Long-range electrostatic forces were calculated using particle mesh Ewald (PME) method[59] with periodic boundary conditions. The non-bonded interactions were calculated with switching and cutoff distances of 10 Å and 12 Å, respectively. A 2 fs integration timestep was used for calculating the forces. All simulations were performed using NAMD2[60] with the CHARMM36 force-field representing pro-

tein, lipid, and nucleic acids, and TIP3P model for water molecules.[61–63] Each biased run was performed for 30 ns to bring the systems close to the target orientations and distances (Table S1). Harmonic force constants of 10^5 kcal/mol/rad^2 and 2 kcal/mol/Å^2 were used for the quaternion-based and distance-based collective variables, respectively. Subsequently, non-equilibrium work profiles and Cα RMSD with respect to the target OF structure for all runs were calculated using our in-house TCL scripts and VMD.[64]

Non-equilibrium work-relations for selecting relevant conformational ensemble of the protein

A heterogenous ensemble traversing the large conformational landscape of Pgp is desirable for the purpose of the current study, allowing delineation of binding preference of different molecules to a diverse set of protein conformations. For this purpose, the equilibrated IF structure showing largest conformational difference (measured in terms of RMSD) from the target OF structure was selected as the starting structure for running longer (100 ns) SMD simulations for the generation of the extended ensemble of the protein. The simulations were carried out using the most efficient biasing-protocol (in terms of work and RMSD values) obtained from the above shorter SMD runs. The relevant extended ensemble of the protein was then selected by taking 50 snapshots along the 100 ns SMD run, equally distributed in the collective variable phase-space. We also analyzed the conformational modulation of the possible binding pockets forming in Pgp over the course of the transition using Site Finder tool in Molecular Operating Environment (MOE),[65] that calculates both the size and the ligand binding propensity (based on the amino acid composition) of the predicted active binding pockets.

II. Targeting the extended ensemble

Structure-based drug-discovery generally utilize techniques like molecular docking for targeting either single structures of the protein target or a local ensemble of protein structures,

allowing targeting of the local conformational space in the vicinity of the starting structures. In the presented extended-ensemble docking protocol, we aim at utilizing the full functional cycle of the protein, transitioning between the major functional states of the protein, for docking purposes. As different docking methods are available for predicting the protein-ligand interactions, providing different levels of accuracy, the selection of a suitable docking method would depend on the specific goals of the study. Here we describe the specific steps involved in the docking study applied to Pgp.

Molecular docking

In the current study, we are interested in understanding the basis of broad substrate specificity in Pgp, and validate our predictions with experimental as well as other in-silico studies conducted on the transporter. For this reason, a previous single-point docking study[31] on the crystal structure of the IF state of Pgp (PDB: 4M1M[22]) was used as a point of comparison for the presented extended-ensemble study. From this single-point docking study, we used a reduced dataset of 14 compounds, containing substrates, non-substrates and high- and low-affinity modulators of Pgp. These selected molecules have been suggested to bind to overlapping sites in Pgp identified previously through mutagenesis studies.[36–38]

We further apply the same docking method used in the single-point docking study described above, allowing for direct comparison with the results from the current study. Docking of compounds was carried out using Autodock Vina[66] in the selected conformational ensemble of the transporter. With the protein centered at (0, 0, 0), a binding grid-box was defined around the TMD region of the protein. The box was centered at (0, 0, 40) with dimensions of 80, 54 and 80 Å in the x, y and z directions, respectively, covering the TMD region. The docking was carried out in an agnostic manner, i.e., the compounds were allowed to sample the entire binding space defined by the box without any constraints. A total of nine binding poses were generated for each docked compound in each protein conformation. A high value for the Autodock Vina exhaustiveness parameter (100) was used, allowing

ample sampling of the compounds in the defined binding space.

III. Clustering of binding poses

In the extended-ensemble docking, clustering of the predicted binding modes of the compounds into specific binding sites in the target protein can be a challenging problem due to the large-scale motions involved in the conformational transition of the protein. For example, clustering of the docked compounds based upon distance-based metrics, like RMSD calculated between the docked compounds, can incorrectly generate separate binding clusters for compounds binding to the same region of the protein. To overcome this problem, distances of each compound can instead be calculated with respect to an independent frame of reference defined in the three-dimensional space and then be used as input for clustering. Here, we present a clustering strategy and its example application to Pgp that can be generalized to other similarly flexible protein targets.

Point selection for distance calculations

In the case of Pgp, distances of each docked compound was calculated with respect to 6 independent reference points in the three-dimensional space: two variable points in the $+x$, -x directions and four fixed points in the $+y$, -y, $+z$, -z directions, respectively. As the closure of the TMD gate in the $\pm x$ direction drives the conformational changes leading to the IF-OF transition, selection of moving $\pm x$ points located on the protein helices involved in the conformational change (and fixed points in the $\pm y/\pm z$ directions) allow the generation of similar distance values for the compounds binding to the same region of the protein. With the protein centered at (0, 0, 0), a grid-box (same as used for docking above) was defined around the TMD region of the protein, with its center at (0, 0, 40) and dimensions of 80, 54 and 80 Å in the x, y and z directions, respectively (Fig. S3). Four fixed points in the $\pm y/\pm z$ directions were selected as the center of the four faces of the grid-box, lying at (0, 27, 40), (0, -27, 0), (0, 0, 0) and (0, 0, 80), respectively, whereas the two moving points in the $\pm x$

directions were selected as the Cα atoms present at the start of the TM1 and TM7 helices of Pgp, respectively. Subsequently, the distances of the center of mass of the predicted binding modes of the docked compounds were calculated from each of these reference points.

Clustering

The above six distances calculated for binding modes of each compound were used for hierarchical clustering in MATLAB.[67] In the first step, a dissimilarity matrix between each pair of objects (distances) present in the dataset was calculated. The dataset for all compounds, consisting of 6300 distances (50 protein conformations $\times$ 14 compounds $\times$ 9 binding poses for each compound), generated a dissimilarity matrix consisting of $(6300 \times (6300 - 1))/2$ pairs. In the second step, a hierarchical tree was constructed by linking together pairs of objects in close proximity using the distances from the dissimilarity matrix. The average linkage method was employed for linking the pair of objects.[68] In the third step, a distance cutoff was used for pruning the branches of the hierarchical tree, allowing partitioning of all points lying below a specific branch into separate clusters. Using trial and error, a distance cutoff of 15 was selected, providing a suitable partitioning of the hierarchical tree, with the different pruned branches representing clustering of compounds binding to different regions of the protein.

IV. Analysis and binding predictions

The binding behavior of different compounds in the extended-ensemble of the target protein can be characterized in terms of the range of predicted binding affinities of the respective binding modes of each compound present in each cluster, obtained from hierarchical clustering above, along with their relative population distribution. Mapping of these clusters onto the surface of the protein allows identification of the corresponding binding sites in the protein along with the residues displaying the highest overall contact frequencies with the predicted binding modes of each compound. In the specific case of Pgp, any heavy atom of

a protein residue lying within 4 Å of a heavy atom of the bound ligand was considered to form contacts with the ligand.

Results

Here we describe the results for the application of extended-ensemble docking to multi-drug transporter Pgp. The process of generating the extended ensemble of the transporter using SMD simulations is described in the supplementary section (Fig. S4, S5, S6, S7 and S8). We first evaluate the conformational modulation of the possible binding sites forming in Pgp as it transitions from the IF to OF state. Subsequent docking calculations carried out in the selected ensemble allow us to characterize the differential binding affinities displayed by compounds in the different conformations of the protein. Furthermore, these calculations allow us to differentiate between the different classes of binders based on their binding preference for specific sites in the DBP.

Evolution of binding pockets in Pgp during IF to OF transition

The calculation of possible active binding pockets in the extended-ensemble of Pgp allowed the evaluation of the changes in the shape/size as well as the propensity of ligand binding of these binding pockets due to the conformational changes in the protein. Three active binding pockets are seen to be present in the membrane encompassing region of the TMD forming the central DBP, with the pocket present at the apex of the TMD observed in all conformations to display high ligand binding propensities. An exception here is the snapshot-50 present in the OF state and showing a low propensity value, which is expected as the DBP in the OF structure becomes solvent exposed (Fig. S3 and Table S2). The largest size of this pocket, along with the highest ligand binding propensity, is observed in the snapshot-30, conformationally similar to the starting crystal structure, with a RMSD of 3.2 Å observed between the two structures. The active binding pockets present in the individual TMD

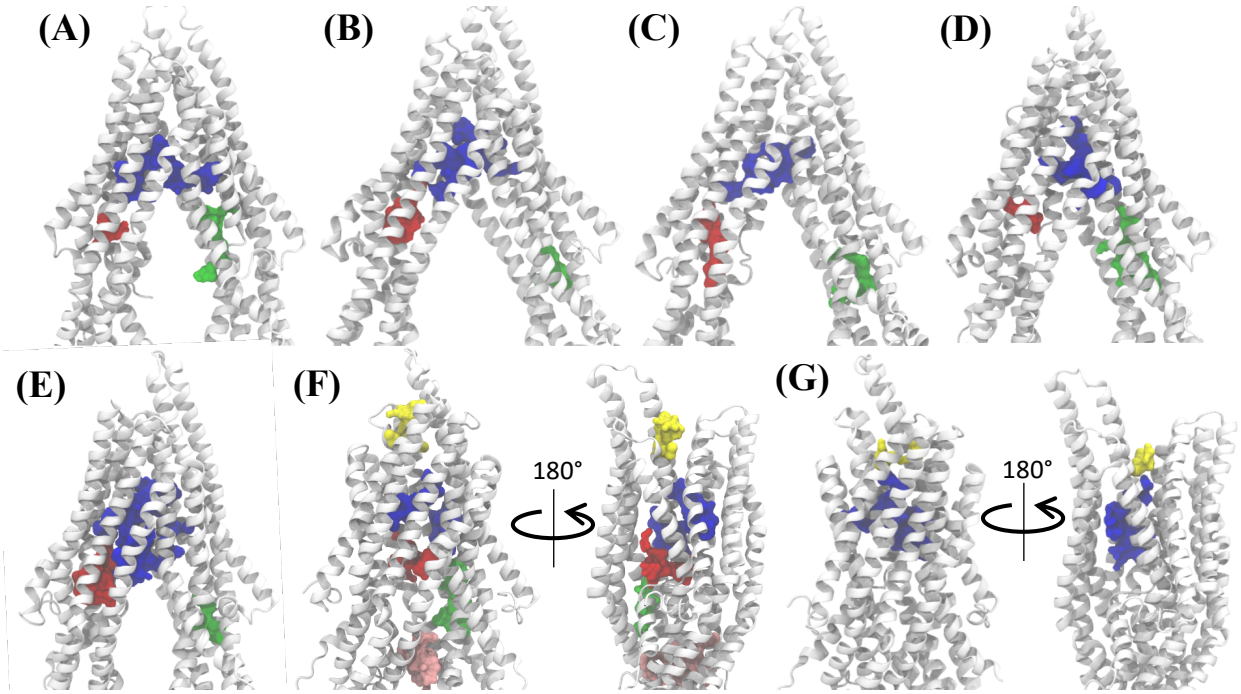

Figure 3: **Binding pocket predictions during IF to OF transition.** A) The predicted active binding pockets (calculated using MOE) present in the starting IF crystal structure are shown. The different binding pockets are represented by space filling representation in different colors. B), C), D), E), F) and G) Active binding pockets predicted in snapshots 1, 10, 20, 30, 40 and 50, of the extended-ensemble of the protein, respectively. The large central pocket (in blue) present in the top half of the DBP shows the highest ligand binding propensities and is present in all conformations of the protein.

leaflets show relatively smaller sizes in the different protein conformations and respectively display lower ligand binding propensities. Small additional binding pockets present on the extracellular and intracellular sides of TMD, showing low ligand binding propensities, are also seen to form in the later snapshots (snapshots 40 and 50) as the protein transitions to the OF state.

Differential binding affinities of compounds predicted in different protein conformations

Instead of targeting a single crystal/cryoEM/homology model, docking of different compounds was targeted on the selected conformational ensemble covering the transitional landscape of Pgp (Fig. 4A and B). This allowed us to evaluate the differential binding of these

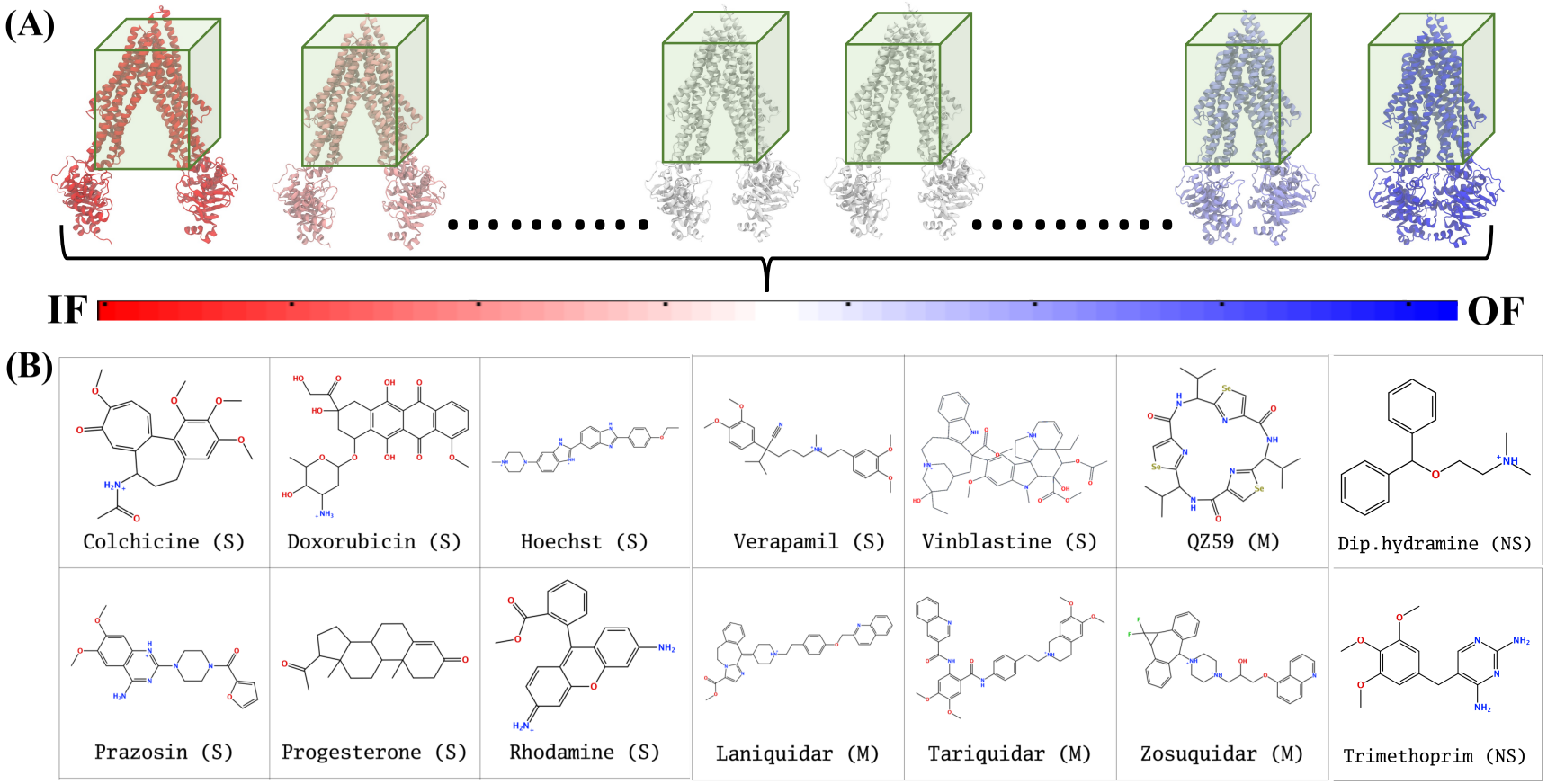

Figure 4: **Conformation selection and compounds used for docking.** (A) The extended-ensemble of Pgp for IF to OF transition (with color changing from red to white to blue between these states), generated by taking 50 images equally distributed along the conformational phase-space defined by the interacellular closing angle, extracellular opening angle and the distance between the two NBDs (Fig. S8). Molecular docking was carried out in a binding grid box (shown in green) defined around the TMD of the protein in all these conformations. (B) The chemical structures of the 14 compounds selected for docking purposes in the extended-ensemble of Pgp are shown. These compounds include known substrates (S), modulators (M) and non-substrates (NS) of Pgp.

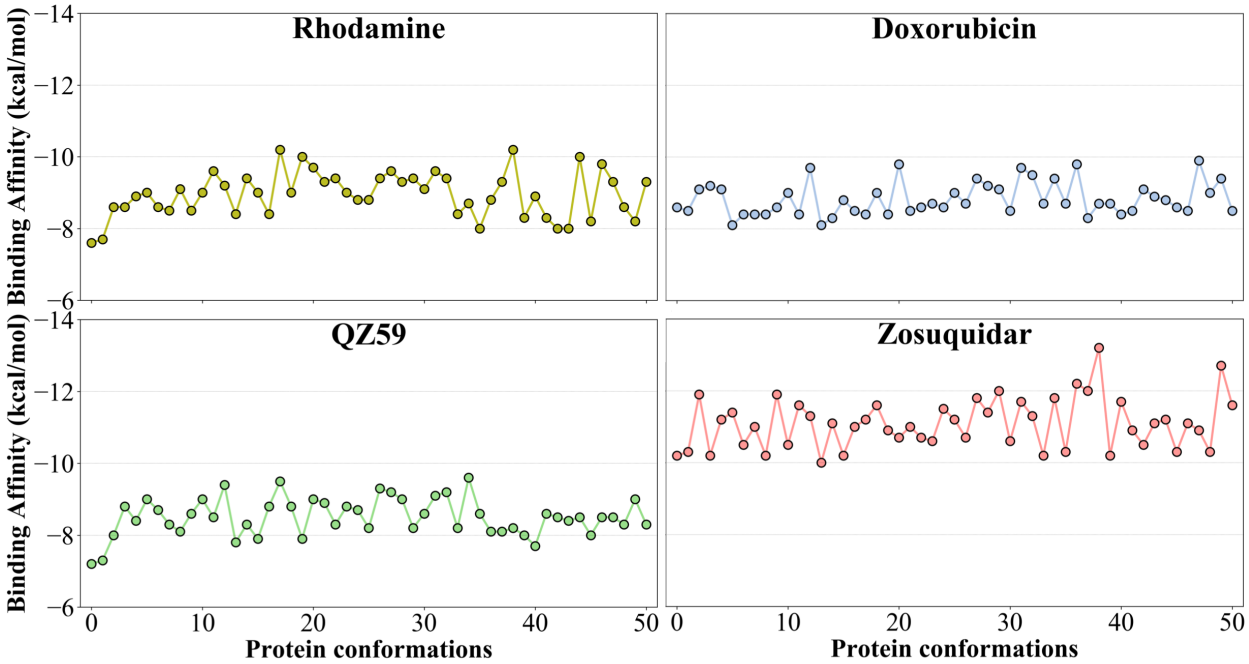

Figure 5: **Predicted binding affinities in the extended-ensemble of the protein.** The highest predicted binding affinities for four representative compounds (small substrate: rhodamine; larger substrate: doxorubicin; low-affinity modulator: QZ59; high-affinity modulator: zosuquidar) in each selected conformation of the protein are shown. Rest of the compounds are shown in Fig. S9. The 0th conformation represents the starting crystal structure of the IF state used in building the transition cycle, and the selected conformational ensemble along the IF to OF transition pathway is numbered 1 to 50. The predicted binding affinities are seen to fluctuate between different protein conformations, with the high-affinity modulators showing the highest binding affinities among all compounds.

compounds to the protein as it undergoes structural transition from IF to OF state. The compounds are seen to bind with different binding affinities in the different protein conformations (Fig. 5 and S9). All compounds show higher affinity binding modes (displaying higher negative docking scores) in conformations other than the starting crystal structure of the IF state, with the difference being as much as -4.1 kcal/mol as in the case of laniquidar. In general, high-affinity modulators of Pgp (laniquidar/tariquidar/zosuquidar) show higher binding affinities compared to low-affinity modulators (QZ59), and larger (doxorubicin/vinblastine) as well as smaller (hoechst/verapamil/progesterone/prazosin/rhodamine/colchicine) substrates. Non-substrates (diphenhydramine/trimethoprim) show the lowest binding affinities among all compounds.

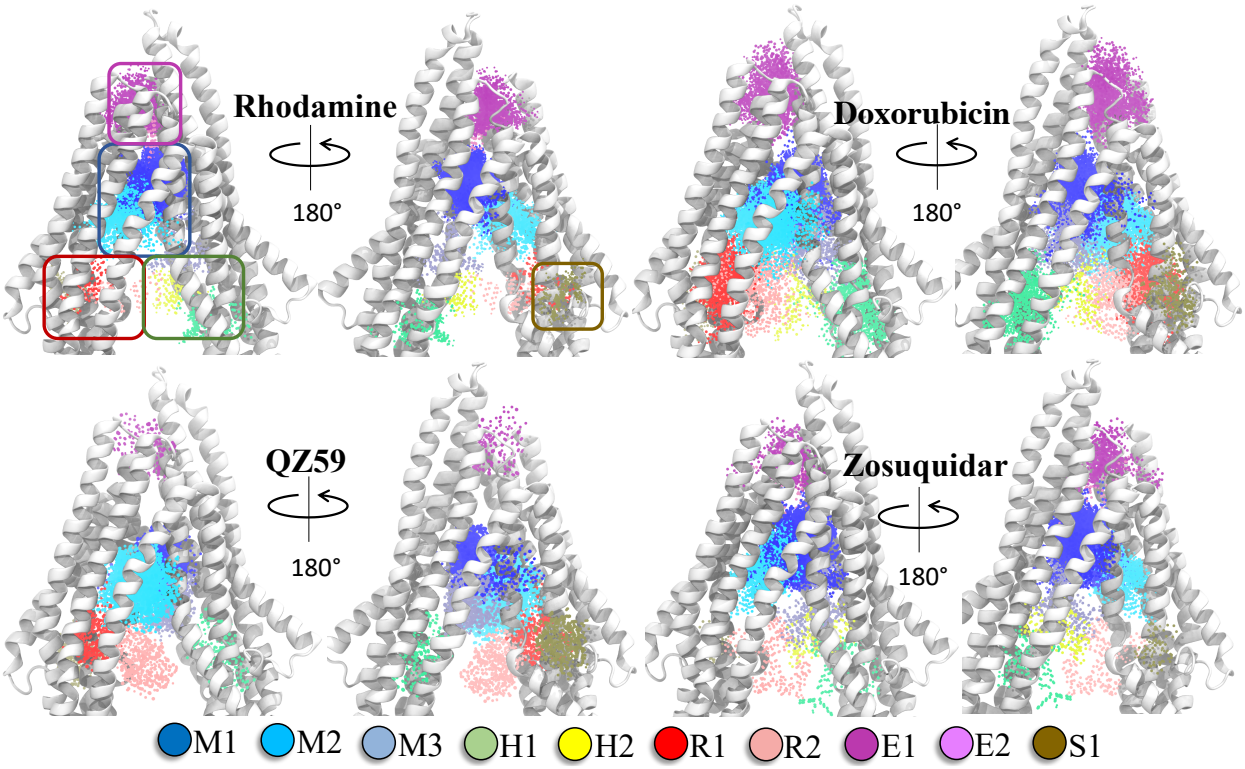

Figure 6: **Clustering of binding modes generated from docking of different compounds.** Clustering of the binding modes for four representative compounds in the extended-ensemble of Pgp was derived using distance-based hierarchical clustering (Fig. S10). Rest of the compounds are shown in Fig. S11. Only one representative IF conformation of the protein structure is shown for clarity. The different binding sites present in the TMD are enclosed by a rectangle (blue: M site; green: H site; red: R site; purple: E site; brown: S site) shown for first compound rhodamine. Binding clusters present in these binding sites are shown in different color, with heavy atoms of the clustered binding modes shown as points. The density of points in each cluster represents the relative cluster population. M1 and M2 subsites present at the apex of the TMD show the highest cluster populations for different compounds.

Clustering shows preferential binding of compounds to two major sites in DBP

The hierarchical clustering of binding modes of all compounds based on distances from 6 points selected in the three-dimensional space, as described in the methods section, allowed us to characterize their binding preference for different regions of the protein. The cutoff distance of 15 partitioned the resultant hierarchical tree into 12 separate clusters (Fig. S10). The relative populations of these clusters showed that clusters 4 and 5 together constituted 78% of all predicted binding modes, whereas only marginal number of binding modes

(< 1%) were observed in the clusters 1, 2, 9 and 12 (Table S3). Mapping the different clusters generated from heirarchical clustering onto the protein structure showed that clusters 1, 2 and 3 (named E1, E3 and E2 sites, respectively) bind to the extracellular side of the TMD, clusters 4, 5 and 11 (named M1, M2 and M3 sites, respectively) bind to the apex of the TMD, clusters 6 and 7 (named R2 and R1 sites, respectively) bind to TMD2, cluster 8 (named S1 site) binds to the outer membrane-facing side of TMD2, cluster 9 (named S2 site) binds to the bottom of the TMD and clusters 10 and 12 (named H1 and H2 sites, respectively) bind to TMD1. The above naming convention used for protein binding sites assigned to each of these clusters is further described in the discussion section.

High-affinity modulators and most substrates show high binding populations in the M1 site, located at the apex of the TMD where the two TMD leaflets (TMD1 and TMD2) meet, whereas low-affinity modulator and larger substrates show preferable binding in the M2 site, located at the apex and partially towards the side of the TMD2 (Fig. 6, S11 and Table 1). Overall, these two clusters present in the top half of the DBP show higher binding affinities compared to the other majorly populated clusters present in the protein (Fig. 7 and S12). The M3 site, present below M1 and towards the side of the TMD1, is lowly populated for all compounds except for larger substrates/low-affinity modulator (Fig. 6, S11 and Table 1). This cluster is seen to form only in the later snapshots of the protein (Fig. 7 and S12). R1 and R2 sites are present below the M2 site in TMD2, whereas H1 and H2 sites are present below the M3 site in TMD1 (Fig. 6 and S11). Low-affinity modulators and large substrates display moderate binding populations in R1/R2 sites, with doxorubicin also showing binding in H1 site (Table 1). The H2 site shows marginal binding populations for all compounds. In general, R1 and H1 sites are seen to form in snapshots 1-30, with the protein present in IF-open conformations, whereas R2 and H2 sites are sampled more in the later snapshots 30-50, with the protein present in IF-closed and OF conformations (Fig. 7 and S12). E1 and E2 sites are seen to be present at the top of the TMD for most compounds with the exceptions of low-affinity modulator QZ59 and larger substrates doxorubicin and vinblastine

(Fig. 6, S11 and Table 1). Both these site are found to be present in snapshots 30-50, with the protein transitioning to the OF state and the DBP opening towards the extracellular side (Fig. 7 and S12). Some of the compounds also show binding to an additional extracellular site E3 formed in the early snapshots of the protein and lying at the top and outside the TMD. This site may not represent a physiological site for the substrate transport, as the transporter is present in an IF state in these conformations with the DBP inaccessible from the extracellular side.

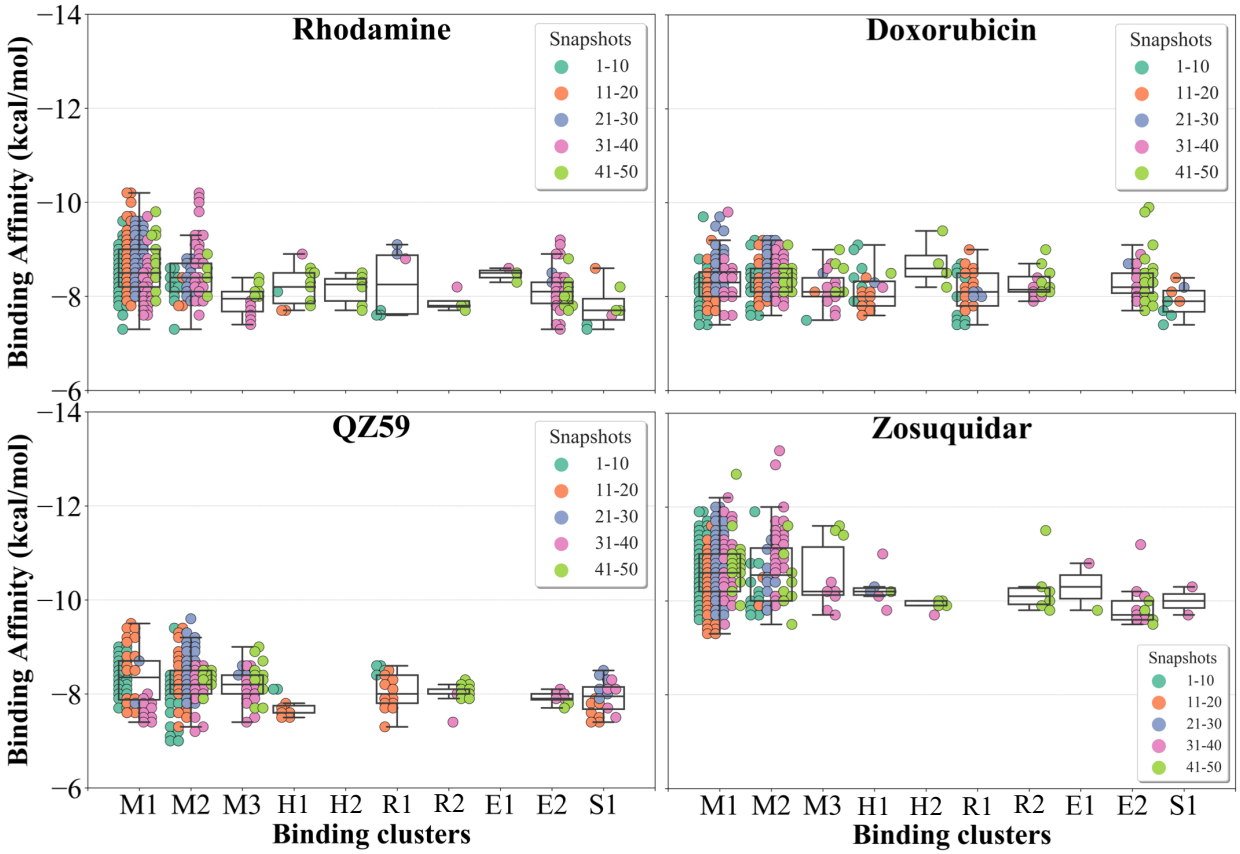

Figure 7: **Binding cluster energies.** The binding affinities of members (binding modes) of each cluster are shown as a swarm plot for four representative compounds. Rest of the compounds are shown in Fig. S12. The snapshots belonging to different set of conformations of the protein as it undergoes IF to OF transition are shown in different colors (defined in legend). Additionally, a boxplot, providing the median cluster values, Q1 and Q3 quartiles, as well as minimum (Q1 − 1.5 × interquartile range) and maximum (Q3 + 1.5 × interquartile range) binding affinity values, is overlaid on top of the swarm plot for each cluster. In general, M1 and M2 binding clusters, showing the highest binding populations, also display the strongest binding affinities for all compounds.

Other than the above ten binding sites, compounds show binding to two additional

subsidiary sites S1 and S2, with S1 present outside TMD2 and formed between the elbow helix and the broken (unstructured) region of TM12 helix, and S2 site formed between TMD4 and TMD6 at the bottom of the TMD (Fig. 6, S11 and Table 1). Out of these, the S2 site is seen to form only in the later protein conformations, and as in the case for the E3 site, may not represent a physiologically important site. This is because the transporter is present in OF-like state in these conformations, with the DBP inaccessible from the intracellular side, likely preventing substrate access to the DBP.

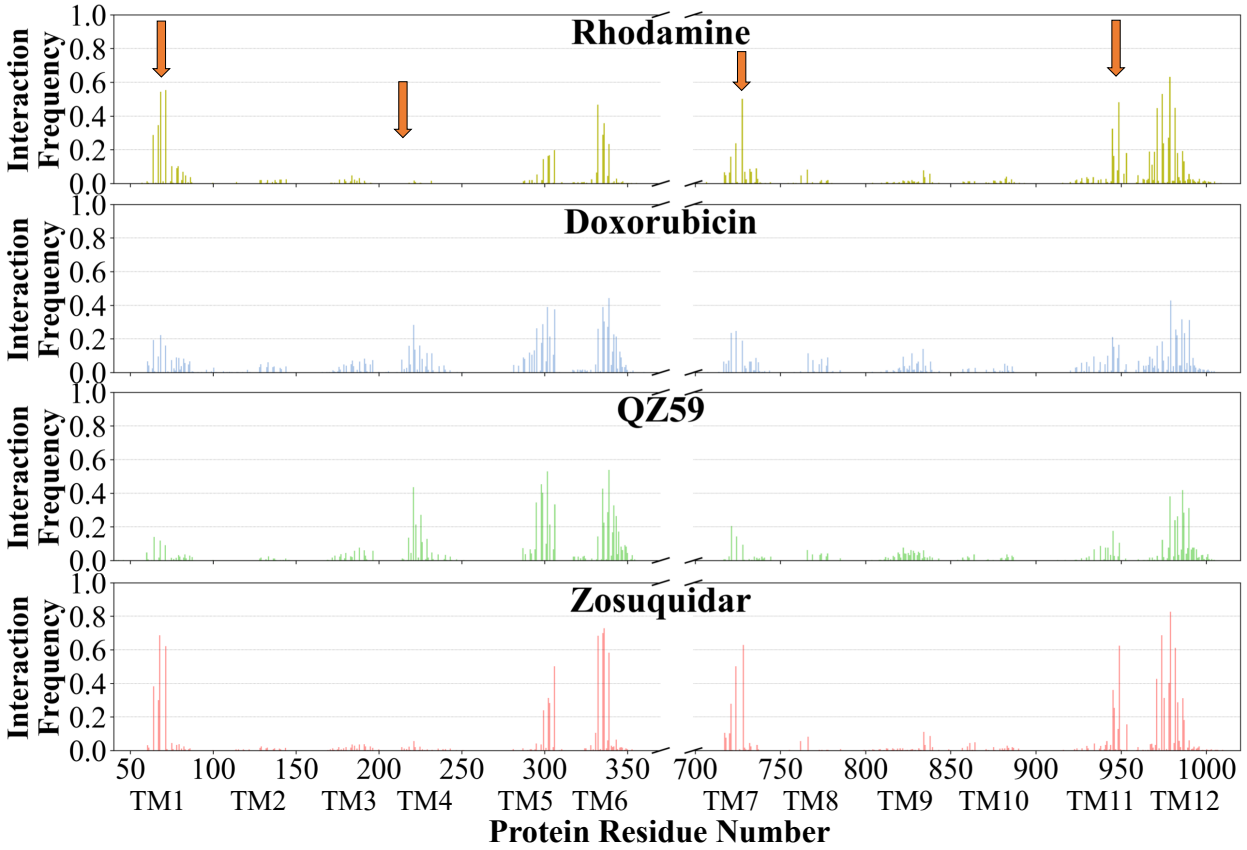

Figure 8: **Interaction frequency of binding site residues.** The normalized interaction frequencies of the binding site residues calculated for all binding modes for four representative compounds docked in the extended-ensemble of Pgp are shown. Rest of the compounds are shown in Fig. S13. The residues are assumed to interact with the docked compound if they are present within 4 Å of the bound ligand. The orange arrows point to the regions of the protein showing differences in interactions for different compounds. High-affinity modulators display the highest values for interaction frequencies with the binding site residues pointing to a specific mode of binding for these compounds.

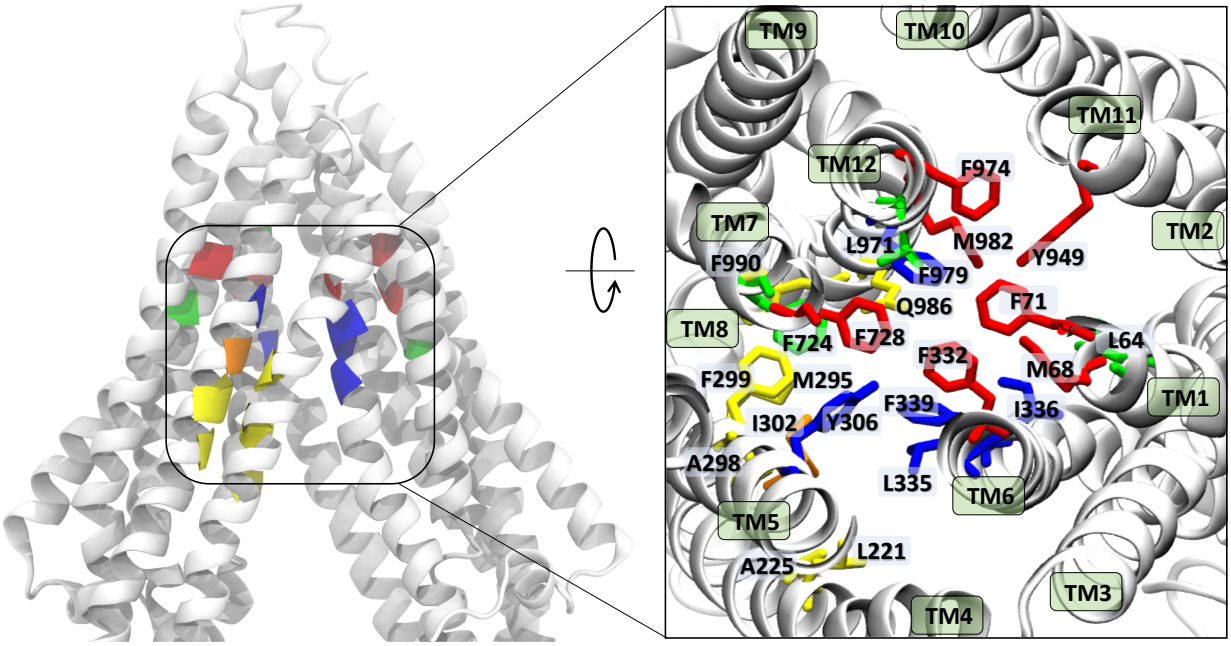

Figure 9: **Top binding site residues identified in Pgp.** Binding site residues in Pgp showing the highest (top-10) interaction frequencies (from Fig. 8 and S13) with the compound docked in the extended-ensemble of the protein are shown in cartoon representation of the protein backbone (left) and in stick representation (inset, right). The different TM helices of Pgp are labelled in the inset. The top binding site residues common to all compounds are shown in blue, residues common to smaller substrates/high-affinity modulators are shown in red, residues common to low-affinity modulators/larger substrates are shown in yellow, residues common to smaller substrates/low-affinity modulators are shown in orange, and residues showing preference for only smaller substrates are shown in green. Smaller substrates and high-affinity modulators share all binding site residues (forming the M1 subsite), whereas low-affinity modulators and larger substrates show binding preference for a different set of residues (forming the M2 subsite) present below the M1 subsite and partially overlapping with it.

Important binding residues present in major sites

The contacts formed by the residues of the protein with the predicted binding modes for each compound in the extended-ensemble of the protein were quantified in terms of interaction frequencies. Comparatively, high-affinity modulators (laniquidar, tariquidar and zosuquidar) and some substrates (progesterone, verapamil) show higher interaction frequencies with the same residues compared to low-affinity modulator (QZ59) and both larger (doxorubicin, vinblastine) and smaller substrates (colchicine, hoechst, prazosin, rhodamine) (Fig. 8 and Fig. 13). The non-substrate trimethoprim shows the lowest interaction frequencies among all compounds. Overall, most compounds form highly favorable interactions with residues from

TM1, TM5, TM6, TM7, TM11 and TM12 helices, that come together to form the DBP, with the exception of larger substrates and low-affinity modulator that show low interaction frequencies with residues from TM1, TM7 and TM11 but instead display higher interactions with TM4 residues.

Based on the highest (top-10) interaction frequencies calculated over all binding modes, all compounds display binding preference for a set of five core residues present in the DBP: Y306, L335, I336, F339, F979 (Fig. 9 and Table 2). High-affinity modulators and smaller substrates additionally show preference for seven common residues: M68, F71, F332, F728, Y949, F974, M982, whereas high-affinity modulators and low-affinity modulator/larger substrates do not share any residue other than the core residues. Low-affinity modulator/larger substrates on the other hand show preference for a different set of residues of the DBP: I221, I225, I295, I298, I299, M986 and M990.

Differential binding of substrates, modulators and non-substrates to major binding site

Ranking of the compounds based on their predicted binding affinities in the highly populated M1 site show that the high-affinity modulators bind with relatively the strongest binding affinities to the extended ensemble, whereas non-substrates of Pgp display the weakest binding affinities (Fig. 10A). The substrates and low-affinity modulator lie between these two extremes. High-affinity modulators and most substrates also display high binding populations in the M1 site, with the population lying between 40-80% for different compounds (Fig. 10B). In contrast, low-affinity modulator and larger substrates show low binding populations in the M1 site, and instead display higher binding populations in the second major binding site of the protein, the M2 site (Table 1).

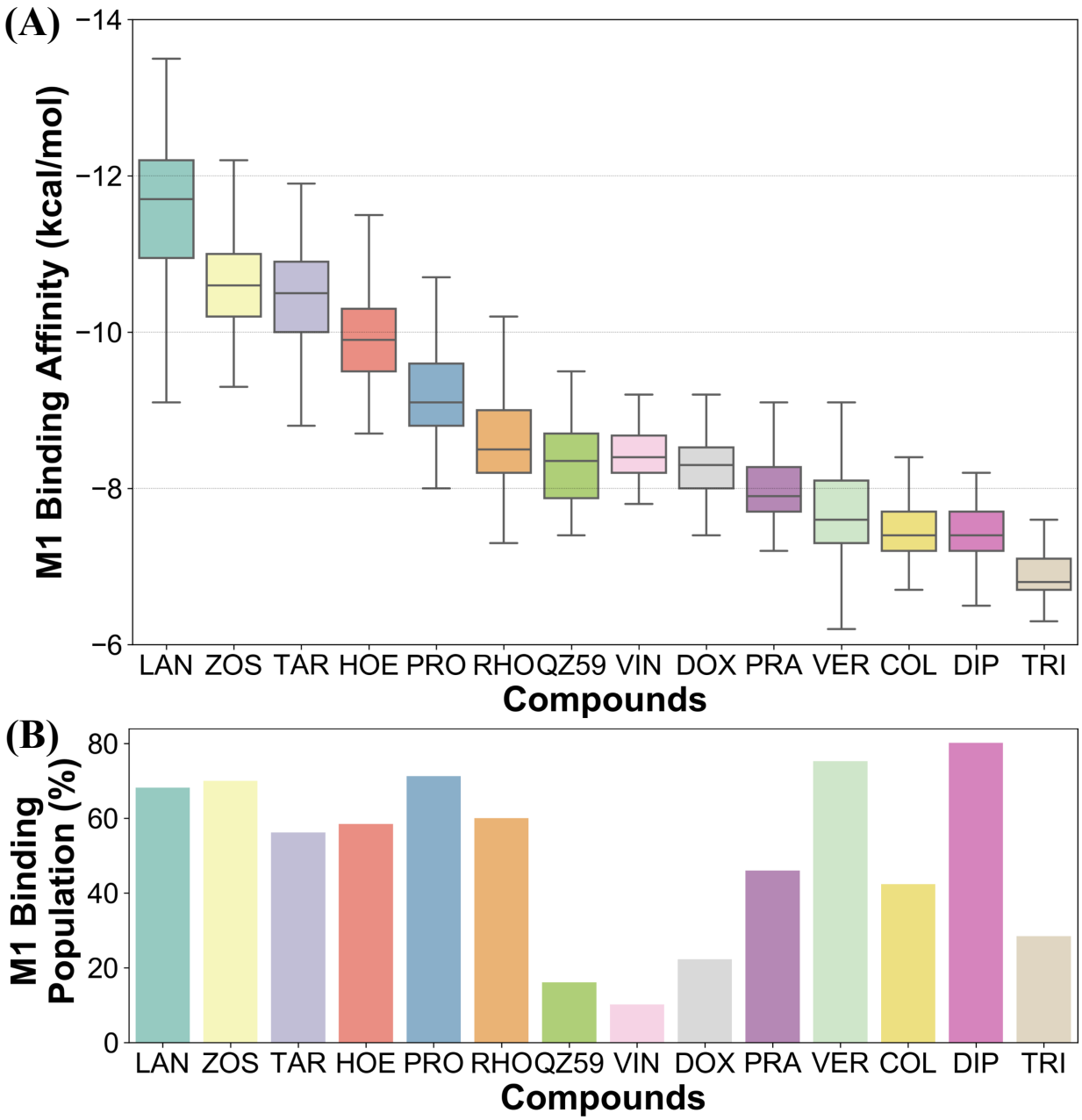

Figure 10: **Differentiating between different classes of compounds.** The binding site preference of different compounds is evaluated in terms of (A) the predicted binding-affinities calculated over the extended-ensemble of the protein in the M1 subsite and (B) the respective binding populations in the same site. Ranking the compounds based on their binding affinities separates the high-affinity modulators and the non-substrates to the two extremes, with low-affinity modulators and substrates lying between these two. Comparison of the binding site population in the M1 subsite further distinguishes the high-affinity modulators and smaller substrates (showing high populations in M1 subsite) from low-affinity modulators and larger substrates (showing higher populations in M2 subsite instead).

Discussion

We have put forth a protocol, termed as extended-ensemble docking, that can be applied in understanding the binding of small molecules as well as in drug discovery applications for flexible proteins undergoing large structural changes, both globally as well as locally in the binding sites (Fig. 2). In the presented study, we make use of a sophisticated SMD approach utilizing well-validated system-specific collective variables describing the alternate-access mechanism in the multidrug ABC transporter Pgp, allowing us to generate a physically relevant conformational ensemble of the protein. Multiple runs with different sequential combinations of the collective variables allow us to delineate the minimum action path, corresponding to the mechanistically most probable transition path connecting the two major functional states of Pgp. This is followed by a series of docking and clustering calculations, allowing us to further characterize the different binding sites in Pgp as they evolve during the transition cycle of the transporter and the differential binding of compounds to these.

The global conformational changes in Pgp also driving the structural changes in the binding sites of the protein are seen to modulate both the size/shape and the ligand binding propensities of these binding pockets. Radioligand binding assays characterizing the binding of Pgp substrates, as well as study measuring drug binding affinity and reactivity to an antibody sensitive to Pgp conformational state, have shown that binding sites in Pgp display differential binding affinities for substrates depending on the conformation of the transporter during its catalytic cycle.[69,70] In the presented study, these differences are reflected in the large differences observed in the predicted binding affinities of compounds in the extended ensemble of the protein (Fig. 5). All compounds are seen to display stronger binding affinities in an alternate Pgp conformations selected during the IF to OF transition compared to the starting crystal structure, showing that experimentally determined protein structures may only reflect one of the many possibilities for interactions between the transporter and its substrates.

Previous drug-drug interaction studies utilizing fluorescence spectroscopy,[71,72] kinetic

studies of substrate binding and transport[73,74] as well as competitive binding assays[75,76] have postulated the presence of at least three binding sites in the TMD of the transporter. These have been defined as H and R sites,[71,72] showing preference for substrates hoechst and rhodamine, respectively, and the modulatory M site, showing preference for vinblastine[37,75] and verapamil,[73,74] two compounds which have been described in literature to function as both modulators and substrates of Pgp.[77–80] Previous single-point docking study targeting the crystal structure of Pgp in the IF state structurally assigned the above sites to different regions in the TMD, with the docked compounds found to show binding preference for all three sites.[31] In the presented study, we redefine the TMD binding sites of Pgp into only one major substrate-modulator binding site, analogous to the M site and encompassing the top half of the DBP, whereas the sites present in the lower half of the DBP and in individual TMD leaflets, analogous to the H and R sites described above, show low binding populations and comparatively weaker binding affinities for all compounds and may in fact only function as poly-specific interacting sites. Here, as both hoechst and rhodamine display similarly low binding populations in the TMD1/TMD2 interacting sites, the respective H and R sites identified in our study have been structurally assigned following the same convention as the previous single-point study.[31] We also report the presence of an interaction site present on the outer surface of the TMD2 of Pgp, named as S site, as well as additional sites formed on the extracellular side of TMD as the transporter transitions to the OF state, named as E sites, that likely assist in the extrusion of molecules out of the transporter.

Based on mapping of the clustering results onto the protein structure, the M site is found to consist of 3 subsites, which we name here as M1, M2 and M3. The M1 subsite present in a central location at the apex of the TMD, allows interactions with residues from multiple helices present in both the TMD leaflets, and may represent the central transport site for substrate molecules (Fig. 6). High-affinity modulators, that can also function as third generation inhibitors of Pgp, are predicted to bind to the same M1 subsite with comparatively stronger binding affinities and higher interaction frequencies, pointing to a more selective

mode of binding for these compounds compared to the substrates (Fig. 7 and Fig. 8). These inhibitors, showing IC_{50} values in the nanomolar range,[81–83] may act as direct competitive inhibitors against substrate molecules binding to the same site. Recently published cryo-EM structures show binding to the same site for a substrate, taxol, as well as two inhibitor molecules of zosuquidar bound to Pgp, with the first molecule occupying the M1 subsite and the second occupying M2/M3 subsites.[52,84] Substrates verapamil and progesterone, that were previously discovered as first generation inhibitors of Pgp,[85] bind to the M1 subsite with high interaction frequencies similar to third generation inhibitors, although with comparatively weaker binding affinities. These molecules show IC_{50} values in the micromolar range and can function as Pgp inhibitors only at higher concentrations, at which they display excessive toxicity and thus not suitable for administration.[86] Larger low-affinity modulator QZ59 shows preference for the M2 subsite of the M site, with low binding populations also observed in the M1 and M3 subsites. Previously published crystal structure of Pgp with this modulator shows the compound bound to the same location present below and partially overlapping with the M1 site and towards the TMD2 leaflet of Pgp. Initial binding to the M2 subsite may also allow binding of additional molecules to the other subsites, further stabilizing the combined binding mode of two or more molecules. This is also evident from multiple crystal structures of Pgp in complex with different structural analogs of QZ59, where two molecules are seen to bind together to the M2-M1 or M3-M1 subsites of Pgp.[46] Larger substrates like doxorubicin and vinblastine also show a similar binding behavior to low-affinity modulators. Similar to QZ59, it is possible that at higher concentrations two or more of these substrates bind together to the different M subsites, allowing them to be eventually transported through the M1 transport site.

Mapping of the clustering results on the protein structure further showed that the two minor binding sites identified on separate TMD leaflets, analogous to H and R sites, each consisted of two subsites, H1 and H2, and R1 and R2, respectively (Fig. 6). The H1 and R1 subsites out of these, present in open-IF like conformations of the transporter, may be

involved in forming initial interactions with the molecules as they partition from the membrane into the two proposed drug-entry portals located on either sides of the transporter. Increase in the local concentration of molecules at these interaction sites may aid in their eventual movement up towards the central transport site. The S site present on the membrane facing side of the TMD2 may allow formation of encounter complexes of substrates with the transporter, and may function as an entry site for these molecules into Pgp. Similar site has also been identified in the crystal structure of Pgp in complex with QZ-Valine, a derivative of the QZ59 compound,[46] as well as in MsbA, a bacterial homolog of Pgp.[87]

The extracellular E site, consisting of E1 and E2 subsites, represent novel binding sites identified in this study formed as the TMD opens to the extracellular side in the OF state of the transporter. Out of these, E2 site, partially overlapping with the M1 site, represents the binding site in DBP opened towards the extracellular side. These sites may play a role in assisting the diffusion of the substrate out of the transporter as the DBP is exposed to the extracellular environment in the OF state by providing additional interaction sites for these molecules on the extracellular side of the protein.

Experiments using photoaffinity labelling of Pgp substrates as well as crystal structure studies have suggested the presence of high- and low-affinity binding sites in Pgp.[46,88,89] Potential of mean force calculations using umbrella sampling also identified multiple overlapping minimum free energy binding locations for different substrates inside the TMD of Pgp.[90,91] The work presented in this extended-ensemble docking study provides further evidence for this proposal, where the M1 and M2 subsites function as the high-affinity binding sites for the high-affinity modulators/smaller substrates and larger low-affinity modulators/substrates, respectively (Fig. 11A). These subsites form overlapping sites within the DBP of Pgp, with modulators and substrates showing preference for the same set of core residues (Fig. 9). The low-affinity binding sites proposed earlier comprise of the H, R and S sites identified in our study and present below the M site in the TMD of Pgp. Correspondingly, the predicted binding affinities display an increasing trend between these low- and high-affinity binding

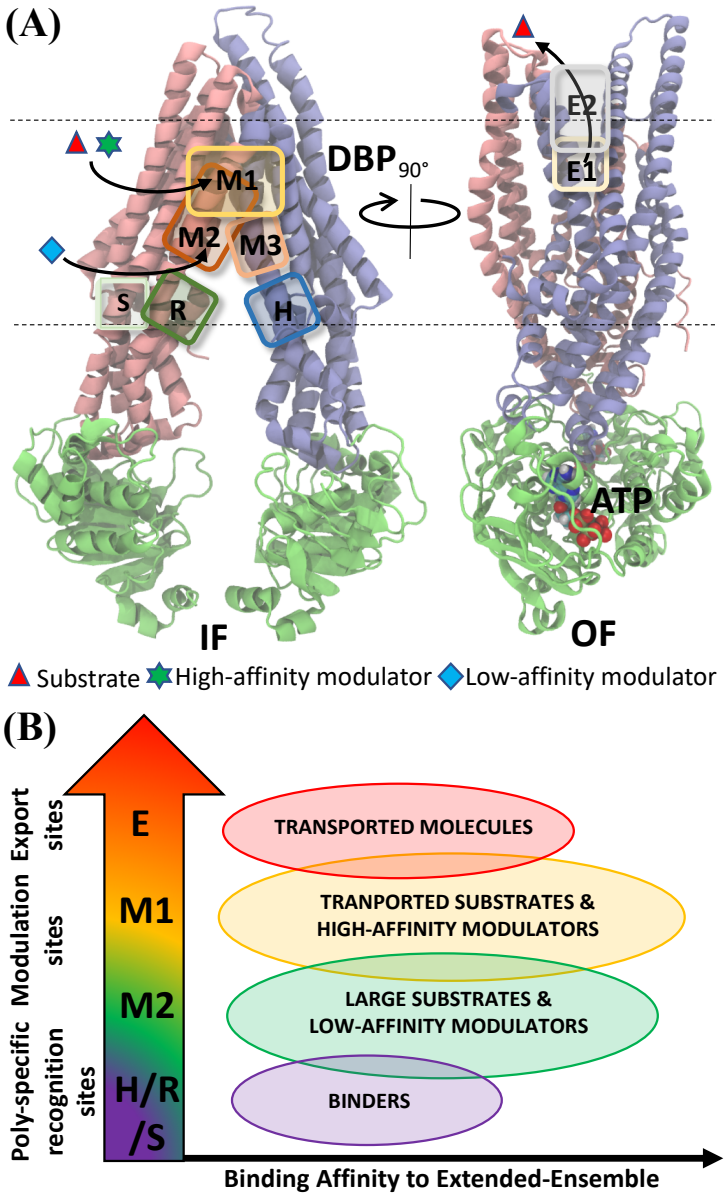

Figure 11: **Evolution of binding sites in extended-ensemble of Pgp.** (A) Different binding sites in DBP characterized in the extended-ensemble of Pgp are shown in representative IF (left) and OF (right) conformations. The two transmembrane leaflets are shown in blue (TMD1) and pink (TMD2) cartoon representation of the protein backbone, respectively, and the NBDs are shown in green. The major substrate binding site (M1), as well as the major low-affinity modulator/large substrate binding site (M2) are observed throughout the conformational transition of the protein, whereas extracellular sites are only observed as the transporter transitions to the OF state and may facilitate the export of substrates out of the transporter. (B) Combining together the predicted binding affinities of all compounds in the different binding sites obtained in the extended-ensemble, differences are observed between the poly-specific interaction sites (H/R/S), modulation sites (M1/M2) and the extracellular sites (E). These differences may facilitate in the unidirectional transport of the molecules from the inside to the outside of the cell.

sites, likely promoting the unidirectional diffusion of molecules towards the central transport site (M1) at the apex of the TMD and the eventual export through the extracellular sites (Fig. 11B).

The differential binding behavior of the compounds to the above subsites in the conformational ensemble further allowed us to successfully differentiate between the different classes of compounds binding to Pgp (Fig. 10). The availability of different binding sites, modulated by the conformational changes in the transporter as it undergoes transition between the IF and OF states, as well as modulation of binding-affinities of these sites during this transition, may form the basis of poly-specificity of Pgp towards different cellular substrates. Furthermore, molecules showing stronger binding affinity to the extended-ensemble of the protein instead of a single structure, as well as a strong selectivity for the M1 subsite, may provide a roadmap for the development of highly specific fourth-generation inhibitors of Pgp.

Conclusion

Large flexible protein like Pgp have proven to be difficult drug-targets, as a number of recent ventures in developing new drug molecules targeting these protein have failed. Further optimization strategies using crystal or cryo-EM structures can prove difficult, as these proteins constantly fluctuate between different conformational states, concurrently leading to structural changes in the available binding pockets. This also proves to be a continual roadblock for developing accurate structural-activity relationship models. Future endeavours in structure-based drug discovery targeting these proteins mandates new in-silico strategies, taking into account the flexibility and inter-talk between different domains of the protein as it undergoes transition between its major functional states. We have provided here a docking protocol pertaining to these proteins, by carrying out high-throughput virtual screening in an extended-ensemble of the multi-drug transporter Pgp instead of targeting a single

crystal/cryo-EM structure. We also provide novel insights into the differential binding preference of different classes of compounds to the transporter. Future applications to other flexible proteins systems may open a new frontier for in-silico drug design targeting important drug targets.

References

(1) Schenone, M.; Dančík, V.; Wagner, B. K.; Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nature chemical biology* **2013**, *9*, 232.

(2) Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nature methods* **2015**, *13*, 24.

(3) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein ensembles: how does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell. *Chemical reviews* **2016**, *116*, 6516–6551.

(4) Ou-Yang, S.-s.; Lu, J.-y.; Kong, X.-q.; Liang, Z.-j.; Luo, C.; Jiang, H. Computational drug discovery. *Acta Pharmacologica Sinica* **2012**, *33*, 1131–1140.

(5) Ramsay, R. R.; Popovic-Nikolic, M. R.; Nikolic, K.; Uliassi, E.; Bolognesi, M. L. A perspective on multi-target drug discovery and design for complex diseases. *Clinical and translational medicine* **2018**, *7*, 3.

(6) Anderson, A. C. The process of structure-based drug design. *Chemistry & biology* **2003**, *10*, 787–797.

(7) Batool, M.; Ahmad, B.; Choi, S. A structure-based drug discovery paradigm. *International journal of molecular sciences* **2019**, *20*, 2783.

(8) Antel, J. Integration of combinatorial chemistry and structure-based drug design. *Current opinion in drug discovery & development* **1999**, *2*, 224–233.

(9) Van Montfort, R. L.; Workman, P. Structure-based drug design: aiming for a perfect fit. *Essays in biochemistry* **2017**, *61*, 431–437.

(10) Ellingson, S. R.; Miao, Y.; Baudry, J.; Smith, J. C. Multi-Conformer Ensemble Docking to Difficult Protein Targets. *Journal of Physical Chemistry B* **2015**, *119*, 1026–1034, PMID: 25198248.

(11) Kapoor, K.; McGill, N.; Peterson, C. B.; Meyers, H. V.; Blackburn, M. N.; Baudry, J. Discovery of novel nonactive site inhibitors of the prothrombinase enzyme complex. *Journal of chemical information and modeling* **2016**, *56*, 535–547.

(12) Pi, M.; Kapoor, K.; Ye, R.; Hwang, D.-J.; Miller, D. D.; Smith, J. C.; Baudry, J.; Quarles, L. D. Computationally identified novel agonists for GPRC6A. *PloS one* **2018**, *13*.

(13) Evangelista, W.; Weir, R. L.; Ellingson, S. R.; Harris, J. B.; Kapoor, K.; Smith, J. C.; Baudry, J. Ensemble-based docking: From hit discovery to metabolism and toxicity predictions. *Bioorganic & medicinal chemistry* **2016**, *24*, 4928–4935.

(14) Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, O.; McCammon, J. A.; Miao, Y.; Smith, J. C. Ensemble Docking in Drug Discovery. *Biophysical Journal* **2018**, *114*, 2271–2278.

(15) Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature* **2007**, *450*, 964–972.

(16) Zuckerman, D. M. Equilibrium sampling in biomolecular simulations. *Annual review of biophysics* **2011**, *40*, 41–62.

(17) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochimica et Biophysica Acta* **2015**, *1850*, 872–877.

(18) Moradi, M.; Tajkhorshid, E. Computational recipe for efficient description of large-scale conformational changes in biomolecular systems. *Journal of Chemical Theory and Computation* **2014**, *10*, 2866–2880.

(19) Holland, I. B.; Blight, M. A. ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. *Journal of Molecular Biology* **1999**, *293*, 381–399.

(20) Fardel, O.; Lecureur, V.; Guillouzo, A. The P-glycoprotein multidrug transporter. *General Pharmacology: The Vascular System* **1996**, *27*, 1283–1291.

(21) Sharom, F. J. The P-glycoprotein multidrug transporter. *Essays in Biochemistry* **2011**, *50*, 161–178.

(22) Li, J.; Jaimes, K. F.; Aller, S. G. Refined structures of mouse P-glycoprotein. *Protein Science* **2014**, *23*, 34–36.

(23) Sharom, F. J. Complex Interplay between the P-Glycoprotein Multidrug Efflux Pump and the Membrane: Its Role in Modulating Protein Function. *Frontiers in Oncology* **2014**, *4*, 1–19.

(24) Mitchell, P. A general theory of membrane transport from studies of bacteria. *Nature* **1957**, *180*, 134–136.

(25) Jardetzky, O. Simple allosteric model for membrane pumps. *Nature* **1966**, *211*, 969–970.

(26) Oldham, M. L.; Davidson, A. L.; Chen, J. Structural insights into ABC transporter mechanism. *Current Opinion in Structural Biology* **2008**, *18*, 726–733.

(27) Locher, K. P. Structure and mechanism of ATP-binding cassette transporters. *Philosophical Transactions of the Royal Society of London B. (Biological Sciences)* **2009**, *364*, 239–245.

(28) Ward, A.; Reyes, C. L.; Yu, J.; Roth, C. B.; Chang, G. Flexibility in the ABC transporter MsbA: alternating access with a twist. *Proceedings of the National Academy of Sciences, USA* **2007**, *104*, 19005–19010.

(29) Verhalen, B.; Dastvan, R.; Thangapandian, S.; Peskova, Y.; Koteiche, H. A.; Nakamoto, R. K.; Tajkhorshid, E.; Mchaourab, H. S. Energy Transduction and Alternating Access of the Mammalian ABC Transporter P-glycoprotein. *Nature* **2017**, *543*, 738–741.

(30) Tarcsay, Á.; Keserű, G. M. Homology modeling and binding site assessment of the human P-glycoprotein. *Future medicinal chemistry* **2011**, *3*, 297–307.

(31) Ferreira, R. J.; Ferreira, M.-J. U.; dos Santos, D. J. Molecular docking characterizes substrate-binding sites and efflux modulation mechanisms within P-glycoprotein. *Journal of chemical information and modeling* **2013**, *53*, 1747–1760.

(32) Chufan, E. E.; Kapoor, K.; Sim, H.-M.; Singh, S.; Talele, T. T.; Durell, S. R.; Ambudkar, S. V. Multiple transport-active binding sites are available for a single substrate on human P-glycoprotein (ABCB1). *PloS one* **2013**, *8*.

(33) Silvente-Poirot, S.; Poirot, M. Cholesterol and cancer, in the balance. *Science* **2014**, *373*, 1445–1446.

(34) Klepsch, F.; Vasanthanathan, P.; Ecker, G. F. Ligand and structure-based classification models for prediction of P-glycoprotein inhibitors. *Journal of chemical information and modeling* **2014**, *54*, 218–229.

(35) Zhang, J.; Sun, T.; Liang, L.; Wu, T.; Wang, Q. Drug promiscuity of P-glycoprotein and its mechanism of interaction with paclitaxel and doxorubicin. *Soft Matter* **2014**, *10*, 438–445.

(36) Ambudkar, S. V.; Dey, S.; Hrycyna, C. A.; Ramachandra, M.; Pastan, I.; Gottesman, M. M. Biochemical, cellular, and pharmacological aspects of the multidrug transporter. *Annual review of pharmacology and toxicology* **1999**, *39*, 361–398.

(37) Safa, A. R. Identification and Characterization of the Binding Sites of P-Glycoprotein for Multidrug Resistance-Related Drugs and Modulators. *Current Medicinal Chemistry - Anti-Cancer Agents* **2004**, *4*.

(38) Loo, T. W.; Bartlett, M. C.; Clarke, D. M. Processing Mutations Disrupt Interactions between the Nucleotide Binding and Transmembrane Domains of P-glycoprotein and the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR). *Journal of Biological Chemistry* **2008**, *283*, 28190–28197.

(39) Palamini, M.; Canciani, A.; Forneris, F. Identifying and visualizing macromolecular flexibility in structural biology. *Frontiers in molecular biosciences* **2016**, *3*, 47.

(40) Moradi, M.; Tajkhorshid, E. Mechanistic picture for conformational transition of a membrane transporter at atomic resolution. *Proceedings of the National Academy of Sciences, USA* **2013**, *110*, 18916–18921.

(41) Moradi, M.; Enkavi, G.; Tajkhorshid, E. Atomic-level characterization of transport cycle thermodynamics in the glycerol-3-phosphate:phosphate transporter. *Nature Communications* **2015**, *6*, 8393.

(42) Moradi, M.; Sagui, C.; Roland, C. Calculating relative transition rates with driven nonequilibrium simulations. *Chemical Physics Letters* **2011**, *518*, 109–113.

(43) Moradi, M.; Tajkhorshid, E. Mechanistic picture for conformational transition of a membrane transporter at atomic resolution. *Proceedings of the National Academy of Sciences, USA* **2013**, *110*, 18916–18921.

(44) Aller, S. G.; Yu, J.; Ward, A.; Weng, Y.; Chittaboina, S.; Zhuo, R.; Harrell, P. M.; Trinh, Y. T.; Zhang, Q.; Urbatsch, I. L.; Chang, G. Structure of P-Glycoprotein reveals a molecular basis for poly-specific drug binding. *Science* **2009**, *323*, 1718–1722.

(45) Ward, A. B. et al. Structures of P-glycoprotein reveal its conformational flexibility and an epitope on the nucleotide-binding domain. *Proceedings of the National Academy of Sciences, USA* **2013**, *110*, 13386–13391.

(46) Szewczyk, P.; Tao, H.; McGrath, A. P.; Villaluz, M.; Rees, S. D.; Lee, S. C.; Doshi, R.; Urbatsch, I. L.; Zhang, Q.; Chang, G. Snapshots of ligand entry, malleable binding and induced helical movement in P-glycoprotein. *Acta Crystallographica D* **2015**, *71*, 732–741.

(47) Esser, L.; Zhou, F.; Pluchino, K. M.; Shiloach, J.; Ma, J.; Tang, W.-k.; Gutierrez, C.; Zhang, A.; Shukla, S.; Madigan, J. P.; Zhou, T.; Kwong, P. D.; Ambudkar, S. V.; Gottesman, M. M.; Xia, D. Structures of the multidrug transporter P-glycoprotein reveal asymmetric ATP binding and the mechanism of polyspecificity. *Journal of Biological Chemistry* **2017**, *292*, 446–461.

(48) Thangapandian, S.; Kapoor, K.; Tajkhorshid, E. Probing cholesterol binding and translocation in P-glycoprotein. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2020**, *1862*, 183090.

(49) Wen, P.-C.; Verhalen, B.; Wilkens, S.; Mchaourab, H. S.; Tajkhorshid, E. On the Origin of Large Flexibility of P-glycoprotein in the Inward-facing State. *Journal of Biological Chemistry* **2013**, *288*, 19211–19220.

(50) Kim, Y.; Chen, J. Molecular structure of human P-glycoprotein in the ATP-bound, outward-facing conformation. *Science* **2018**, *359*, 915–919.

(51) Dastvan, R.; Mishra, S.; Peskova, Y. B.; Nakamoto, R. K.; Mchaourab, H. S. Mechanism of allosteric modulation of P-glycoprotein by transport substrates and inhibitors. *Science* **2019**, *364*, 689–692.

(52) Alam, A.; Kowal, J.; Broude, E.; Roninson, I.; Locher, K. P. Structural insight into substrate and inhibitor discrimination by human P-glycoprotein. *Science* **2019**, *363*, 753–756.

(53) Zaitseva, J.; Jenewein, S.; Jumpertz, T.; Holland, I. B.; Schmitt, L. H662 is the linchpin of ATP hydrolysis in the nucleotide-binding domain of the ABC transporter HlyB. *EMBO Journal* **2005**, *24*, 1901–1910.

(54) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **2008**, *29*, 1859–1865.

(55) Jo, S.; Lim, J. B.; Klauda, J. B.; Im, W. CHARMM-GUI membrane builder for mixed bilayers and its application to yeast membranes. *Biophysical Journal* **2009**, *97*, 50–58.

(56) 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. *Journal of Computational Physics* **1977**, *23*, 327–341.

(57) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *Journal of Chemical Physics* **1994**, *101*, 4177–4189.

(58) Feller, S. E.; Zhang, Y.; Pastor, R. W. Constant pressure molecular dynamics simulation: The Langevin piston method. *Journal of Chemical Physics* **1995**, *103*, 4613–4621.

(59) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: an $N \cdot \log(N)$ method for Ewald sums in large systems. *Journal of Chemical Physics* **1993**, *98*, 10089–10092.

(60) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *Journal of Computational Chemistry* **2005**, *26*, 1781–1802.

(61) Huang, J.; MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry* **2013**, *34*, 2135–2145.

(62) Feller, S. E.; MacKerell, Jr., A. D. An Improved Empirical Potential Energy Function for Molecular Simulations of Phospholipids. *Journal of Physical Chemistry B* **2000**, *104*, 7510–7515.

(63) Foloppe, N.; MacKerell Jr., A. D. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry* **2000**, *21*, 86–104.

(64) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(65) Molecular Operating Environment (MOE), 2019.01. 2020.

(66) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2010**, *31*, 455–461.

(67) TheM̃athWorks, I. Matlab v.6. 2002.

(68) Maimon, O.; Rokach, L. *Data mining and knowledge discovery handbook*; Springer, 2005.

(69) Martel, V.; Racaud-Sultan, C.; Dupe, S.; Marie, C.; Paulhe, F.; Galmiche, A.; Block, M. R.; Albiges-Rizo, C. Conformation, Localization, and Integrin Binding of

Talin Depend on Its Interaction with Phosphoinositides. *Journal of Biological Chemistry* **2001**, *276*, 21217–21227.

(70) Bársony, O.; Szalóki, G.; Türk, D.; Tarapcsák, S.; Gutay-Tóth, Z.; Bacsó, Z.; Holb, I. J.; Székvölgyi, L.; Szabó, G.; Csanády, L.; Szakács, G.; Goda, K. A single active catalytic site is sufficient to promote transport in P-glycoprotein. *Scientific reports* **2016**, *6*, 24810.

(71) Shapiro, A. B.; Ling, V. Positively cooperative sites for drug transport by P-glycoprotein with distinct drug specificities. *European Journal of Biochemistry* **1997**, *250*, 130–137.

(72) Shapiro, A. B.; Ling, V. Transport of LDS-751 from the cytoplasmic leaflet of the plasma membrane by the rhodamine-123-selective site of P-glycoprotein. *European journal of biochemistry* **1998**, *254*, 181–188.

(73) Loo, T. W.; Clarke, D. M. Determining the dimensions of the drug-binding domain of human P-glycoprotein using thiol cross-linking compounds as molecular rulers. *Journal of Biological Chemistry* **2001**, *276*, 36877–36880.

(74) Loo, T. W.; Bartlett, M. C.; Clarke, D. M. Simultaneous Binding of Two Different Drugs in the Binding Pocket of the Human Multidrug Resistance P-glycoprotein. *Journal of Biological Chemistry* **2003**, *278*, 39706–39710.

(75) PASCAUD, C.; GARRIGOS, M.; ORLOWSKI, S. Multidrug resistance transporter P-glycoprotein has distinct but interacting binding sites for cytotoxic drugs and reversing agents. *Biochemical Journal* **1998**, *333*, 351–358.

(76) Callaghan, R.; Riordan, J. R. Synthetic and natural opiates interact with P-glycoprotein in multidrug-resistant cells. *Journal of Biological Chemistry* **1993**, *268*, 16059–16064.

(77) Tsuruo, T.; Iida, H.; Tsukagoshi, S.; Sakurai, Y. Overcoming of vincristine resistance in P388 leukemia in vivo and in vitro through enhanced cytotoxicity of vincristine and vinblastine by verapamil. *Cancer research* **1981**, *41*, 1967–1972.

(78) Wang, S.-Y.; Bonner, K.; Russell, C.; Wang, G. K. Tryptophan scanning of D1S6 and D4S6 C-termini in voltage-gated sodium channels. *Biophysical Journal* **2003**, *85*, 911–920.

(79) Wang, X. Q.; Li, L. N.; Chang, W. R.; Zhang, J. P.; Gui, L. L.; Guo, B. J.; Liang, D. C. Structure of C-phycocyanin from *Spirulina platensis* at 2.2 Å resolution: a novel monoclinic crystal form for phycobiliproteins in phycobilisomes. *Acta Crystallographica D* **2001**, *57*, 784–792.

(80) Rautio, J.; Humphreys, J. E.; Webster, L. O.; Balakrishnan, A.; Keogh, J. P.; Kunta, J. R.; Serabjit-Singh, C. J.; Polli, J. W. In vitro P-glycoprotein inhibition assays for assessment of clinical drug interaction potential of new drug candidates: a recommendation for probe substrates. *Drug metabolism and disposition* **2006**, *34*, 786–792.

(81) Mistry, P.; Stewart, A. J.; Dangerfield, W.; Okiji, S.; Liddle, C.; Bootle, D.; Plumb, J. A.; Templeton, D.; Charlton, P. In vitro and in vivo reversal of P-glycoprotein-mediated multidrug resistance by a novel potent modulator, XR9576. *Cancer research* **2001**, *61*, 749–758.

(82) Dantzig, A. H.; Shepard, R. L.; Cao, J.; Law, K. L.; Ehlhardt, W. J.; Baughman, T. M.; Bumol, T. F.; Starling, J. J. Reversal of P-glycoprotein-mediated multidrug resistance by a potent cyclopropyldibenzosuberane modulator, LY335979. *Cancer Research* **1996**, *56*, 4171–4179.

(83) van Zuylen, L.; Sparreboom, A.; van der Gaast, A.; Nooter, K.; Eskens, F.; Brouwer, E.; Bol, C.; De Vries, R.; Palmer, P.; Verweij, J. Disposition of docetaxel in the presence of

P-glycoprotein inhibition by intravenous administration of R101933. *European Journal of Cancer* **2002**, *38*, 1090–1099.

(84) Alam, A.; Kung, R.; Kowal, J.; McLeod, R. A.; Tremp, N.; Broude, E. V.; Roninson, I. B.; Stahlberg, H.; Locher, K. P. Structure of a zosuquidar and UIC2-bound human-mouse chimeric ABCB1. *Proc Natl Acad Sci USA* **2018**, *115*, 1973–1982.

(85) Balaeff, A.; Mahadevan, L.; Schulten, K. Modeling DNA loops using the theory of elasticity. *Physical Review E* **2006**, *73*, 031919.

(86) Kurosawa, M.; Okabe, M.; Hara, N.; Kawamura, K.; Suzuki, S.; Sakurada, K.; Asaka, M. Reversal effect of itraconazole on adriamycin and etoposide resistance in human leukemia cells. *Annals of hematology* **1996**, *72*, 17–21.

(87) Smriti, P. Z.; Mchaourab, H. S. Mapping daunorubicin-binding sites in the ATP-binding cassette transporter MsbA using site-specific quenching by spin labels. *The Journal of biological chemistry* **2009**, *284*, 13904.

(88) Dey, S.; Ramachandra, M.; Pastan, I.; Gottesman, M. M.; Ambudkar, S. V. Evidence for two nonidentical drug-interaction sites in the human P-glycoprotein. *Proceedings of the National Academy of Sciences* **1997**, *94*, 10594–10599.

(89) Al-Shawi, M. K.; Omote, H. The remarkable transport mechanism of P-glycoprotein: a multidrug transporter. *Journal of Bioenergetics and Biomembranes* **2005**, *37*, 489–496.

(90) Subramanian, N.; Condic-Jurkic, K.; Mark, A. E.; O'Mara, M. L. Identification of possible binding sites for morphine and nicardipine on the multidrug transporter P-glycoprotein using umbrella sampling techniques. *Journal of chemical information and modeling* **2015**, *55*, 1202–1217.

(91) Subramanian, N.; Schumann-Gillett, A.; Mark, A. E.; O'Mara, M. L. Probing the

pharmacological binding sites of P-glycoprotein using umbrella sampling simulations. *Journal of chemical information and modeling* **2018**, *59*, 2287–2298.

Table 1: Percentage distribution of binding modes of individual compounds in different binding sites in the extended ensemble of the protein.

	Compounds	**M1**	**M2**	**M3**	**H1**	**H2**	**R1**	**R2**	**E1**	**E2**	**E3**	**S1**	**S2**
Substrates	Colchicine	42.4	22	1.8	0.9	0.7	2	3.1	0	15.7	0.9	8.3	2.2
	Doxorubicin	22.3	40	5.1	5.3	0.9	8.3	3.6	0	9.8	1.8	1.8	1.1
	Hoechst	58.5	27.1	1.1	2.2	0	3.8	0.9	2.4	4	0	0	0
	Prazosin	46	28.5	0.4	7.4	0.2	5.3	0.5	0.5	7.6	2.5	0.9	0.2
	Progesterone	71.3	24	0	0.2	0.2	0.2	0	0.2	3.4	0	0.5	0
	Rhodamine	60	15.4	2.7	2.4	1.3	1.3	0.9	0.7	11.4	0.7	1.6	1.6
	Verapamil	75.3	16.4	0	1.1	0	1.6	0.5	0.9	2	0.2	1.8	0.2
	Vinblastine	10.2	51.1	8.5	1.3	1.8	5.3	14.7	0	4.4	0.9	1.3	0.5
Modulators	QZ59	16.1	54.3	7.4	2.5	0	4.5	4.7	0	1.8	2.2	4.5	2
	Laniquidar	68.2	19.6	0.7	0.7	0	0.2	2	1.3	3.8	0	3.1	0.4
	Tariquidar	56.2	27.2	0.5	1.3	2.9	4	1.8	1.8	3.6	0	0.5	0.2
	Zosuquidar	70	17.9	2.2	1.3	1.1	0	1.4	0.5	3.8	0	0.5	1.3
Non-substrates	Diphenhydramine	80.2	15.4	0	0.4	0	0.7	0	3.3	0	0	0	0
	Trimethoprim	28.5	31	1.1	10.2	0	9.1	0	1.1	14.6	1.4	3	0

Table 2: Top 10 binding residues for each compound showing highest interaction frequencies in the extended-ensemble of Pgp.

Compounds	**Top Binding Residues**
Colchicine	L64, M68, F71, F332, L335, I336, F728, F974, F979, M982
Doxorubicin	I221, I299, I302, Y306, L335, I336, F339, F979, M986, M990
Hoechst	F71, I302, Y306, F332, L335, F339, F728, Y949, F974, F979
Prazosin	M68, Y306, F332, L335, I336, F339, F724, Y949, F974, F979
Progesterone	M68, F71, Y306, F332, L335, I336, F339, F728, F974, F979
Rhodamine	M68, F71, F332, I336, F728, Y949, F971, F974, F979, M982
Verapamil	M68, F71, Y306, F332, L335, I336, F728, F974, F979, M982
Vinblastine	I221, I225, I295, I298, I299, I302, Y306, F339, M986, M990
QZ59	I221, I295, I298, I299, I302, Y306, L335, F339, F979, M986
Laniquidar	M68, F71, Y306, F332, L335, I336, F728, Y949, F974, F979
Tariquidar	M68, Y306, F332, L335, I336, F339, F728, F974, F979, M982
Zosuquidar	M68, F71, F332, L335, I336, F728, Y949, F974, F979, M982
Diphenhydramine	M68, F71, F332, F724, F728, Y949, F953, L971, F974, F979
TRI	F299, I302, Y303, Y306, F339, Q721, Q834, F979, A983, V987