JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Pocket-Space Maps To Identify Novel Binding-Site Conformations in Proteins

Ian R. Craig,*[,||,†] Christopher Pfleger,[‡] Holger Gohlke,[‡] Jonathan W. Essex,[§] and Katrin Spiegel[†]

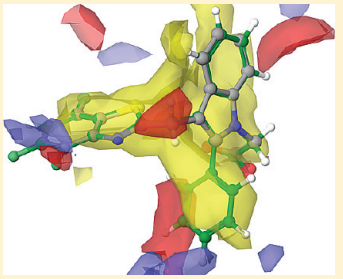
[†]Novartis Institutes for Biomedical Research, Wimblehurst Road, Horsham, West Sussex, RH12 5AB, U.K.

[‡]Mathematisch-Naturwissenschaftliche Fakultät, Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany

[§]School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K.

Supporting Information

ABSTRACT: The identification of novel binding-site conformations can greatly assist the progress of structure-based ligand design projects. Diverse pocket shapes drive medicinal chemistry to explore a broader chemical space and thus present additional opportunities to overcome key drug discovery issues such as potency, selectivity, toxicity, and pharmacokinetics. We report a new automated approach to diverse pocket selection, PocketAnalyzer^{PCA}, which applies principal component analysis and clustering to the output of a grid-based pocket detection algorithm. Since the approach works directly with pocket shape descriptors, it is free from some of the problems hampering methods that are based on proxy shape descriptors, e.g. a set of atomic positional coordinates. The approach is technically straightforward and allows simultaneous analysis of mutants, isoforms, and protein structures derived from multiple sources with different residue numbering schemes. The PocketAnalyzer^{PCA} approach is illustrated by the compilation of diverse sets of pocket shapes for aldose reductase and viral neuraminidase. In both cases this allows identification of novel computationally derived binding-site conformations that are yet to be observed crystallographically. Indeed, known inhibitors capable of exploiting these novel binding-site conformations are subsequently identified, thereby demonstrating the utility of PocketAnalyzer^{PCA} for rationalizing and improving the understanding of the molecular basis of protein−ligand interaction and bioactivity. A Python program implementing the PocketAnalyzer^{PCA} approach is available for download under an open-source license (http://sourceforge.net/projects/papca/ or http://cpclab.uni-duesseldorf.de/downloads).



INTRODUCTION

Advances in experimental and computational methodology, and improvements in hardware and instrumentation, have rapidly increased the rate at which molecular structures of proteins can be generated.[1] Multiple experimentally determined structures are often now available for a particular protein of interest, and it is also more frequently feasible to derive a large set of protein conformations using computational methods. This is driving structure-based drug design to move beyond a "one target, one structure" perspective to account for and embrace the flexibility of proteins.[2−4] Since medicinal chemists can use protein structures to guide their ligand design toward the formation of specific protein−ligand interactions, a diverse set of protein conformations presents an opportunity to explore chemical space more widely. This freedom increases the chances that key drug discovery issues such as potency, selectivity, ADME (absorption, distribution, metabolism, and excretion), and toxicity can be overcome. In the past, computational methods have confirmed the existence of different pocket shapes a posteriori, such as the additional subpocket in the HIV-integrase active site, exploited by the first marketed HIV-integrase inhibitor.[5,6] The goal of structure based computational chemistry is to identify druggable pocket shapes beforehand and guide chemistry to exploit alternative pocket shapes.

The analysis and interpretation of large volumes of protein structural information can be a lengthy process if visual inspection is required in order to detect and confirm diverse and potentially novel pocket shapes. Alternatively, computational approaches to diverse pocket selection have been devised that typically analyze Molecular Dynamics (MD) trajectories and often involve some form of clustering to bundle similar structures together, followed by selection of just one representative structure from each group.[7] Some form of dimensionality reduction is also common, Principal Component Analysis (PCA)[8] being a prominent example. These techniques provide a characterization of the dominant deformation modes of the binding pocket and map out the pocket conformational space.

Of critical importance is the set of coordinates to which techniques such as clustering and PCA are applied. The set of Cartesian coordinates for all active site C_α atoms can be a poor proxy for local binding pocket conformation because protein structures with similar C_α conformations may still have very different side-chain conformations and will thus incorporate very different binding pocket shapes. This can lead to dissimilar pocket conformations being clustered or mapped together, and to novel pocket shapes being missed. Thus, methods for select-

Received: April 13, 2011
Published: September 12, 2011

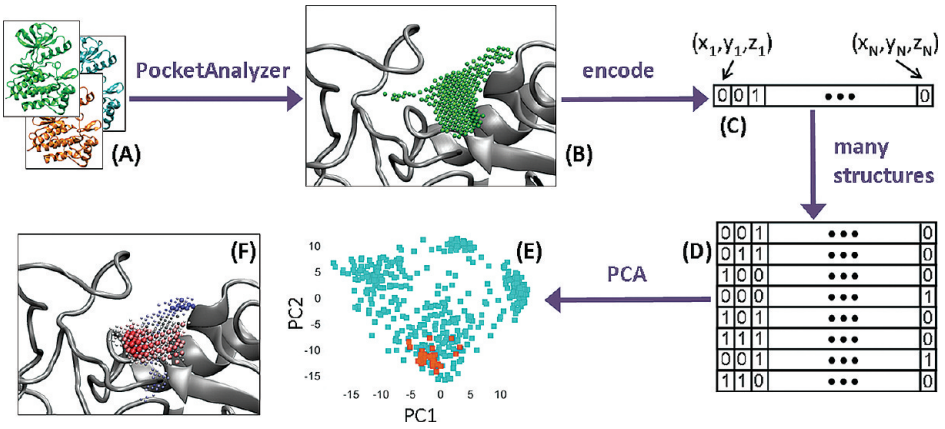


Figure 1. A graphical summary of the PocketAnalyzer^{PCA} approach. The PocketAnalyzer pocket detection algorithm is applied to each protein structure in the set (A). For each structure, the resulting pocket shape (B) is encoded as a row vector of integers (C). Merging the row vectors from each protein structure produces the pocket shape matrix (D). Projecting the row vectors onto principal components derived from the pocket shape matrix generates a map of the pocket conformational space (E). The principal components themselves describe the dominant changes in pocket shape within the set of protein conformations (F).

ing a diverse set of pocket conformations should instead use descriptors that account for all the atoms lining the pocket. An obvious choice is the set of Cartesian positions for all atoms bordering the binding site.[7,9] However, collating this set of descriptors can be technically awkward when comparing protein structures with different residue numberings and/or atom orderings. It also precludes the inclusion of protein structures with mutations of one or more binding pocket residues or structures of more than one isoform. Furthermore, there can be considerable ambiguity in defining the set of "binding site atoms", particularly for flexible proteins.

In this paper, we address the problem of diverse pocket selection from an alternative perspective by working directly with pocket shape descriptors rather than a set of proxy coordinates. In particular, we develop a procedure that reduces a large collection of structures of the same protein to a subset that retains a number of substantially distinct binding pocket conformations. The approach is based on applying PCA and clustering to the output of a grid-based pocket detection algorithm. The PCA yields (a) principal component (PC) eigenvectors, which reveal the dominant deformation modes of the pocket, and (b) PC projections ("scores"), which provide characterization and visualization of the pocket conformational distribution (PCD). Clustering of the PCD then results in an all-atom approach to diverse pocket selection that is free from the problems hampering methods based on atomic coordinates. As our method builds upon the PocketAnalyzer pocket detection code developed in the Gohlke group, we call it PocketAnalyzer^{PCA}. Although itself unpublished, the original PocketAnalyzer approach implements a variant of the LIGSITE algorithm.[10,11] Minor differences between PocketAnalyer and LIGSITE are described below. The approach is technically straightforward and allows simultaneous analysis of mutants, isoforms, and homologous proteins. Furthermore, although we focus on the identification of novel pocket conformations from MD simulations, our procedure is applicable to any source of atomistic protein structural information, and all combinations thereof.

Here, we apply PocketAnalyzer^{PCA} to two proteins that exhibit moderate but significant active site flexibility, namely aldose reductase and neuraminidase. Both proteins have been extensively studied, resulting in a well-characterized set of binding pocket conformations with which to compare the output of our approach to diverse pocket selection. Since the PocketAnalyzer^{PCA} technique is potentially applicable to problems beyond diverse pocket selection, the paper closes by highlighting some directions for future work.

MATERIALS AND METHODS

PocketAnalyzer^{PCA}. *Outline.* We first give a short overview of the PocketAnalyzer^{PCA} approach (Figure 1), before dealing with the methodological details in more depth. In the first step, a grid-based pocket detection algorithm is applied to an ensemble of protein structures. The identified pockets from each single protein structure are represented as a row vector of integers, each encoding the inclusion ("1") or exclusion ("0") of a particular grid point. The row vector describes the pocket shape corresponding to a specific conformation of the protein. Since the same set of grid-points is used to analyze each protein structure, the row vectors can be merged to produce a pocket shape matrix. Each column of this pocket shape matrix then represents the varying inclusion/exclusion of a particular grid-point in the pockets of an ensemble of protein structures.

The pocket shape matrix is subjected to Principal Component Analysis (PCA).[12] This results in (a) a set of PC eigenvectors, (b) a set of eigenvalues that correspond to the variance along each PC, and (c) the projections (or "scores") of each protein structure along each PC. The high-variance PCs describe the dominant changes in pocket conformation within the given set of protein structures. The scores characterize the distribution of pocket shapes along the PCs, providing a map of the pocket conformational space. Analysis and comparison of these pocket conformational distributions (PCDs) may provide a useful perspective on a variety of questions related to molecular recognition and protein dynamics. Of particular relevance to medicinal chemistry efforts is the rephrasing of the diverse pocket selection problem in terms of finding a small number of protein structures whose pockets nevertheless provide coverage of the significant regions of the PCD. In this work a clustering of the PCD is used to achieve this aim.

Pocket Detection. The pocket detection algorithm implemented within the PocketAnalyzer code is a variant of the LIGSITE

algorithm[10] very similar to that described by Stahl and co-workers.[11] As such, a grid map is defined across the protein where each grid point must meet a number of criteria in order to be included in the pocket. First, the grid point should not be within the van der Waals radius of any protein atom. Second, it must be sufficiently enclosed within the protein structure. This degree of buriedness is assessed by scanning away from the grid point along fourteen vectors (positive/negative in the *x*, *y*, and *z* axes and the four cube diagonals) and counting the number of directions in which protein atoms are encountered within a distance of 10 Å. A grid point is excluded if this count is less than the user-defined threshold *dob* (degree of buriedness). Third, the grid point must be surrounded by a certain number of other well-buried neighboring grid points, determined by the parameter *mnb* (minimal number of neighbors). Fourth, after clustering of the grid points meeting the preceding criteria, the point must belong to a cluster of size greater than the parameter *mcs* (minimal cluster size). The default values for the parameters are *dob* = 11, *mnb* = 15, and *mcs* = 50 at a grid-spacing of 0.8 Å.

The parameters can be used to optimize the pocket detection with respect to the protein system under investigation. Solvent-exposed pockets require a lower degree of buriedness threshold (*dob)* to adjust to the generally lower enclosure of grid-points in an open binding site. Smaller pockets might only be identified by setting a lower minimal cluster size threshold (*mcs*). The number of neighbors (*mnb*) affects the shape of the identified pockets. Increasing this value leads to pockets with a more globular shape, whereas lower values result in more disperse and filamentous pocket shapes, e.g. (sub)pockets that are connected by a tunnel. In this work the aldose reductase analysis used the default parameter settings, whereas the values have been modified for the neuraminidase analysis. In particular, the *dob* threshold parameter was reduced from the default value of 11 to 9 to account for the large and solvent-exposed binding pocket. As this then leads to disperse pocket shapes in the case of neuraminidase, the *mnb* value has been increased from the default of 15 to 18. Other parameters were kept at their default values. As for the grid-spacing, changing this value within the boundaries of 0.5 to 1.5 Å does not grossly affect the identification of pockets. We thus chose a grid spacing of 0.8 Å, equivalent to approximately one-half of a C−C bond distance, in order to detect pockets reliably with moderate computational effort.

Note that two differences between the LIGSITE/Stahl algorithm and the current one are that PocketAnalyzer: (a) does not ignore hydrogen atoms and (b) does not add a tolerance of 0.8 Å to the van der Waals radii. Regarding the former, in this work the Protein Preparation Wizard in *Maestro*[13] was used to add hydrogens to the crystal structures. The same utility was also used to optimize the resulting hydrogen-bonding networks.

Structural Alignment. A crucial aspect of the PocketAnalyzer^{PCA} approach is that all the protein structures are analyzed on the same set of grid points (currently initialized using the first structure in the list). For this to be meaningful, the protein structures must be aligned before submission to the pocket detection algorithm. Here, protein structures were aligned to minimize the rmsd between the Cartesian coordinates of the C_α atoms of a specific set of active site residues (defined below). As the structural alignment depends on the set of atoms chosen,[14] the question arises as to how sensitive the PocketAnalyzer^{PCA} approach is to a change in alignment. Notably, changing from an all C_α alignment to an active site C_α alignment caused little change in the resulting PCs and PCDs for large sets of protein structures (see Supporting Information Figure S1): although the PC scores of a few individual structures sometimes changed more significantly, the spectrum of pocket shapes returned by the clustering process generally remained the same. This demonstrates that the PocketAnalyzer^{PCA} approach is empirically rather insensitive to a switch between two sensible alignments.

For ALR, C_α atoms of the following manually selected active site residues were used for the alignment: 1ADS: W20, H46, V47, Y48, K77, W79, C80, H110, W111, T113, G114, F121, F122, L124, V130, N160, Q183, Y209, W219, V297, C298, L300, L301, S302, C303, H306, and Y309. For neuraminidase, the C_α atoms of all residues were used for the alignment.

Principal Component Analysis. PCA is used to reduce the dimensionality of the pocket shape matrix. Since all matrix elements are of the same type (i.e., binary integers representing grid-point inclusion/exclusion) the covariance matrix is directly diagonalized rather than first normalizing to the correlation matrix. Importantly, the PCA does not have to use all the grid-point inclusion/exclusion row vectors. Some can be withheld and then projected onto the resulting principal components. This provides a mechanism to investigate how well one set of structures (e.g., derived crystallographically) overlap or cover the PCD derived from another set (e.g., derived from a MD simulation).

To minimize the number of descriptors used by the PCA, only grid points that are included in the pocket of interest for at least one of the protein conformations are considered. For typical binding pockets only a few hundred grid-points meet this criterion, and the resulting PCA takes less than a minute for a few hundred protein structures. For both aldose reductase and neuraminidase the 'pocket of interest' in this work is the active (i.e., catalytic) site.

Clustering. Clustering is used to derive a small subset of protein structures whose pockets nevertheless provide coverage of the significant regions of the pocket conformational distribution. In particular, the CLARA algorithm[15] implemented in the R package *cluster*[16] is applied to group the protein structures according to their scores along the PCs. Retaining only the representative protein structure from each cluster leaves a subset of protein conformations corresponding to a diverse selection of pocket shapes. In CLARA the cluster representative is the medoid, i.e. the member of the cluster with minimal average Euclidean distance to the other members of the cluster. The PocketAnalyzer^{PCA} program performs this clustering via an external call to R.

The PocketAnalyzer^{PCA} program allows clustering of pocket shapes either on the original grid-point inclusion/exclusion descriptors or on the scores (i.e., projections) along a user-defined number of the PCs. Data sets with a steep scree plot, which shows the variance along the PC versus the PC index, could be meaningfully clustered using only a few PCs. In contrast, using all of the PCs is equivalent to clustering on the original grid-point variables. In between these extremes, based on their relatively flat scree plots, the examples presented here employ a CLARA clustering of the first 50 PC scores. Since this accounts for 83% of the variance in the aldose reductase data set, and 63% for neuraminidase, this corresponds to clustering on all pocket shape changes except small and/or infrequently observed ones.

Druggability Testing. SiteMap is used to assess the druggability of the cluster representatives.[17] The DScore druggability metric computed by SiteMap is a linear function of three pocket properties: size, enclosure (akin to an average degree of buriedness), and hydrophilicity. These descriptors are calculated for each pocket using a grid-based approach similar to the algorithm adopted by

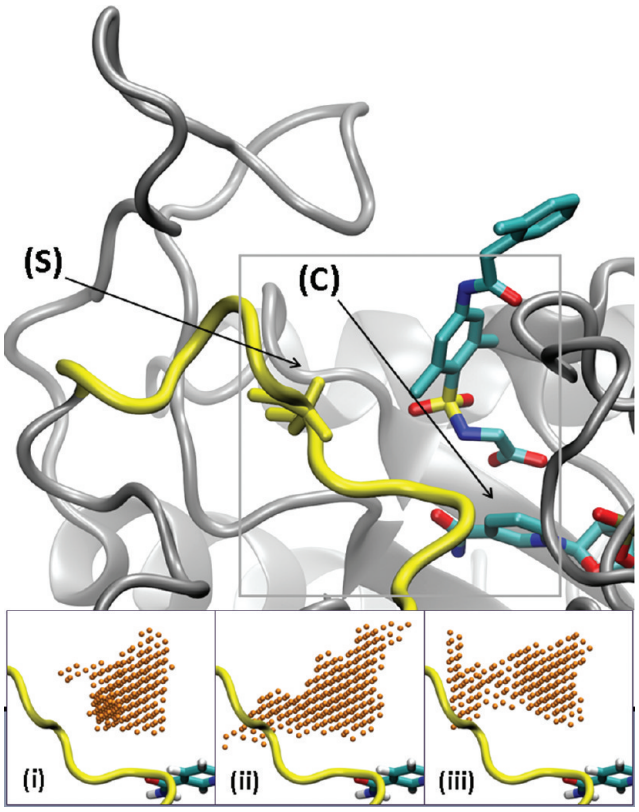


Figure 2. The large figure shows the active site of aldose reductase in an *apo*-like conformation (PDB 1EL3). The protein backbone is represented in gray ribbon with a more flexible section highlighted in yellow, as is the torsionally labile Leu300 residue. The selectivity pocket and the catalytic site are labeled (S) and (C) respectively. The ligand (IDD384) is shown right-of-center, and part of the $NADP^+$ cofactor is visible in the lower right-hand corner. For reference, the gray box indicates the area displayed in subsequent figures. The smaller figures below illustrate pocket shapes characterized by PocketAnalyzer for three representative ALR crystal structures: (i) 1ADS (*apo* conformation), (ii) 1PWL (*zenarestat*), and (iii) 2FZD (*tolrestat*).

PocketAnalyzer. More details on SiteMap and DScore can be found in the literature.[17] Many other methods for characterizing the physicochemical properties and druggability of protein pockets are available.[18−21]

Data Sets. *Aldose Reductase.* To provide a reference for active site conformations generated by MD simulations, a set of eight crystallographic ALR protein structures was assembled with the help of the CavBase module of Relibase+.[22,23] The ALR structures show two main active site conformations: an *apo* conformation in which the selectivity pocket illustrated in Figure 2 is closed (PDB codes 1ADS, 1EL3) and a group of conformations in which the selectivity pocket is open. Named after the ligands bound to them,[24] the latter are *tolrestat* (2FZD and 2FZB), *zenarestat* (1IEI and 1PWL), and *IDD594* (1US0 and 2R24). Note that (a) all ALR residue numberings in this work are those of the 1ADS structure, (b) the ligand in the 1PWL structure is actually *minalrestat* not *zenarestat*, and (c) the *apo* conformation was termed *'holo'* in some previous work.[24]

It is natural to speculate whether any known ALR inhibitors without a crystallographically confirmed binding mode might in fact occupy the novel pocket conformations identified by $PocketAnalyzer^{PCA}$ (see Results section). To address this issue, inhibitors with structural similarity to a ligand with an available crystallographic binding mode were aligned to that template using the substructure-based ligand alignment method available in ICM.[25] Five crystallographic binding modes were initially chosen as templates: *tolrestat* (from PDB code 2FZB), *zenarestat* (1IEI), *IDD384* (1EL3), *minalrestat* (1PWL), and *IDD594* (1US0). The first three bind the catalytic site of aldose reductase via an α-aminoacid substructure, *minalrestat* binds via a cyclic imide substructure, and *IDD594* via an α-hydroxyacid substructure. A set of 395 ALR inhibitors with $IC_{50} < 1\ \mu M$ was downloaded from the bindingDB,[26] of which 66 matched the substructure of at least one of the templates (see Supporting Information, List S9). Since at least 36 of the 66 substructure-matched bindingDB compounds were obvious derivatives of lidorestat,[27] its crystallographic binding mode from PDB structure 1Z3N was added as a sixth alignment template.

Neuraminidase. The set of crystallographic reference structures of neuraminidase N1 consists of an *apo* structure (2HTY), the oseltamivir bound structure in the closed 150-loop conformation (2HU4), the oseltamivir bound structure in the open 150-loop conformation (2HU0), the zanamivir bound *holo* structure (3B7E), a second *apo* structure (3BEQ), and the oseltamivir resistant H274Y mutant in complex with zanamivir and oseltamivir (3CKZ, 3CL0). To this set of neuraminidase N1 protein conformations, two structures of neuraminidase N8 are added in which the protein is bound to oseltamivir with the 150 loop in the open and closed positions (2HT7, 2HT8). All neuraminidase residue numberings in this work correspond to those of the 2HTY structure.

As for ALR, structural alignments to crystallographic binding-mode templates were used to predict whether known N1 inhibitors might bind to any of the novel pocket shapes identified by the $PocketAnalyzer^{PCA}$ protocol. A set of 64 N1 inhibitors with $IC_{50} < 1\ \mu M$ were downloaded from the bindingDB.[26] All but one of these matched the substructure of oseltamivir. In consequence, the only crystallographic binding-mode used was that of oseltamivir from PDB entry 2HU4.

MD Simulations. To generate computationally derived ensembles of ALR and neuraminidase structures, MD simulations were performed using the Amber 9 package.[28,29]

Protocol for Aldose Reductase. Three trajectories were initiated from the 1US0 crystal structure of ALR in complex with the carboxylic-acid type inhibitor *IDD594*. This initial structure was chosen due to its high resolution (0.66 Å) and to be consistent with previous simulations.[24] All crystallographic waters, citrate ions, and the ligand were deleted. The $NADP^+$ cofactor was retained. Side-chain tautomerization and protonation states were assigned with the help of the Protein Preparation Wizard in *Maestro*.[13] Hydrogens were deleted in *Maestro* before being added back in AMBER-compatible format using the *tleap* module. The Amber ff03 force-field[30] was used for the protein, and the Ryde parametrization was used for the cofactor.[31] The *tleap* module was also used to solvate the protein in a rectangular box of TIP3P water molecules, with a minimal distance between the solute and the boundary of the box of 11 Å.[32] Two potassium ions were then added to ensure overall charge neutrality. This resulted in a simulation cell with dimensions $81 \times 70 \times 81\ Å^3$ and a total of 37299 atoms.

One equilibration run was used to prepare all three trajectories, beginning with a two-step minimization approach. First, the protein and cofactor were held fixed while the solvent was minimized using 500 steps of the steepest descent method

followed by conjugate gradient minimization. In a second round of minimization, using the same approach, the restraints on the protein atoms were relaxed. The entire system was then heated from 5 to 300 K over 200 ps at constant volume. In this and all following simulations, a Langevin thermostat[33] with a collision frequency of 5 ps^{-1} was employed to regulate the temperature, and a time step of 1 fs was used. The SHAKE algorithm[34] was used throughout to constrain the lengths of bonds involving hydrogen, and the particle mesh Ewald method[35] was employed to treat long-range electrostatic interactions. The nonbonded cutoff was set at 12 Å.

The equilibration was completed with a second stage of dynamics, this time at constant pressure for 800 ps at a fixed temperature of 300 K. The pressure was regulated to a reference of 1 bar using AMBER's weak-coupling barostat[36] with a relaxation time of 4 ps.[29] Three production simulations were then launched from the equilibrated structure. These trajectories differed only in the sequence of random numbers used by the thermostat. Each of the simulations ran for 20 ns, and coordinates were saved at 500 ps intervals to give three sequences of 40 computationally derived protein conformations.

Protocol for Neuraminidase. To generate a computationally derived ensemble of neuraminidase conformations, MD simulations were initiated from the 2.5 Å resolution *apo* structure (PDB 2HTY) and the 2.4 Å resolution *holo* structure in complex with oseltamivir (PDB 2HU4). Topologies compatible with the Amber 9 program were prepared as outlined for aldose reductase. The calcium ion was retained as it is structurally important.[37] Force field parameters for oseltamivir have been assigned using the generalized amber force field (GAFF).[38] The atom types, charges, and prep-files for oseltamivir are included in the Supporting Information (Section S2). To achieve charge neutrality, three potassium ions were added. The systems were solvated with TIP3P water molecules as described above for ALR. This resulted in a final box size of 80 × 81 × 81 $Å^3$ and a total of 43201 atoms for the *apo* simulation and a box of size 79 × 82 × 82 $Å^3$ and a total of 43236 atoms for the two *holo* simulations.

The neuraminidase structure features eight disulfide bridges, which are expected to stabilize the protein framework. The minimization and equilibration protocol was shortened with respect to aldose reductase: in a first step, the entire system was minimized without restraints by 5000 steps of steepest descent. In a second step, the temperature was gradually increased during 60 ps to 300 K using the Andersen temperature coupling scheme, randomizing velocities every 1000 steps, and a time step of 2 fs. Harmonic restraints of 5 kcal mol^{-1} $Å^{-2}$ were applied to the protein backbone. The density of the system was allowed to adjust at the same time, again using AMBER's weak coupling barostat.[36] Finally, the harmonic restraints were released, and the system was allowed to relax during another 100 ps before starting the production run. Throughout all MD simulations, the SHAKE algorithm[34] and the particle mesh Ewald method[35] were used. The nonbonded cutoff was set at 12 Å. The two *holo* simulations differed only in the sequence of random numbers used by the thermostat. Each of the simulations ran for 20 ns, and coordinates were saved at 10 ps intervals to give three sequences of 200 computationally derived protein conformations. A greater number of conformations was extracted from the neuraminidase simulations than from the ALR trajectories to compensate for the higher mobility of the protein in the former.

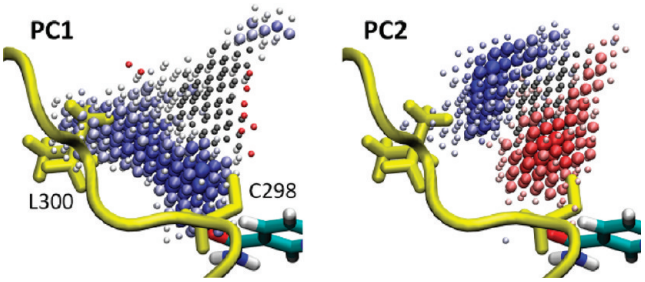


Figure 3. The highest-variance principal components of the aldose reductase data set: PC1 (left) and PC2 (right). Each is a linear combination of the original grid-point descriptors. The size of the sphere representing each grid-point reflects the absolute magnitude of its coefficient in the linear combination: the largest spheres indicate the largest coefficients, i.e. the most important grid-points for that particular PC. Any two grid-points with the same color are positively correlated: according to that particular PC, they tend to be in the pocket shape simultaneously. Any two grid points with different colors are negatively correlated: they do not tend to be in the pocket shape simultaneously. The dark gray spheres indicate grid-points which are included in more than 80% of the pocket shapes. Also shown for reference are the side-chains of residues C298 and L300.

RESULTS

PocketAnalyzer^{PCA} Applied to Aldose Reductase. PocketAnalyzer^{PCA} was first applied to aldose reductase (ALR). As a target for the prevention of several diabetic complications, ALR has been the subject of numerous drug design studies.[39] Crystallographic and computational investigations have already identified a set of distinct ALR active site conformations.[24,40−42] This makes ALR a useful benchmark system for diverse pocket selection, with the crystal structures revealing an active site of moderate flexibility driven more by changes in side-chain conformation rather than backbone rearrangement.

Some important components of the ALR active site are illustrated in Figure 2. Active site conformations identified by Klebe and co-workers[24,40−42] differ principally in the region of the "selectivity" pocket. In contrast, the region around the catalytic residues on the opposite side of the active site is markedly rigid.[24,43] The protein structure with PDB code 1ADS exemplifies the *apo* conformation in which the selectivity pocket is closed. The result is an active site of low volume mainly comprised of the residues surrounding the catalytic site and the $NADP^+$ cofactor. A side-chain rotation of residue L300, and an associated adjustment of certain backbone torsional angles, opens the selectivity pocket in the *tolrestat-*, *zenarestat-*, and *IDD594*-bound conformations.[24] However, differences in the positioning of other side-chains, particularly those of residues W111, F115, F122, and C303, result in a distinct selectivity pocket conformation in each case.

Diverse Pocket Selection. The PocketAnalyzer^{PCA} diverse pocket selection protocol was used to analyze and explore the active site conformations visited in the three MD simulations of aldose reductase (see Materials and Methods). The principal aim was to compare the MD-derived binding-site conformations with those observed in the set of available ALR crystal structures and to thus identify novel computationally generated pocket shapes. As a first step, the protein conformations contained in the crystallographic and MD data sets were structurally aligned and processed with PocketAnalyzer. A Principal Component Analysis was then applied to the grid-point inclusion/exclusion row

vectors corresponding to the 120 computationally derived structures (3 trajectories, 40 frames from each). In contrast, the grid-point row vectors corresponding to the 8 crystallographic pocket conformations were not submitted to the PCA but were subsequently projected onto the PCs resulting from the analysis of the MD-derived protein structures. However, including the crystallographic pocket conformations in the PCA makes little difference to the resulting PCs and the PCD (see Supporting Information Figure S3).

The two highest-variance PCs (out of 120 PCs in total) are shown in Figure 3. The first, labeled PC1, describes an expansion of the pocket in the direction of the flexible L300 loop, simultaneously opening both the selectivity pocket and another subpocket near residue C298. In contrast, the positive and negative lobes of PC2 show alternate, i.e. mutually exclusive, expansion into different areas. According to PC2, pocket conformations which include the grid-points toward the selectivity pocket and L300 (blue) tend not to include certain grid-points around the C298 subpocket (red). Since these high-variance principal components produced by PocketAnalyzer^{PCA} characterize the dominant changes in pocket shape within the structural data set, they provide a novel and easily visualized perspective on protein flexibility that is directly relevant to ligand binding and ligand design. It is worth noting that, although PC1 and PC2 are the highest variance principal components, in this case they are cumulatively responsible for only 21% of the data set's total variance. In general it may be necessary to consider more than two PCs to obtain a full picture of the possible pocket deformations.

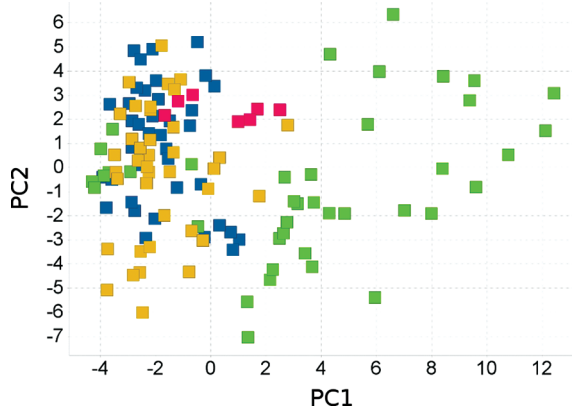


Figure 4. The pocket conformational distribution for the aldose reductase data set along PC1 and PC2. Pocket shapes derived from crystal structures are represented by red squares. Pocket shapes derived from the three trajectories MD1, MD2, and MD3 are colored blue, green, and yellow respectively.

PocketAnalyzer^{PCA} also generates a score for each pocket conformation along each PC by projecting the row vector of a pocket conformation onto the PC. These scores collectively characterize the pocket conformational distribution (PCD) and provide a map of the regions of pocket space explored in the current data set. Figure 4 exemplifies this with a plot of the scores along PC1 against those along PC2. Such plots enable comparisons to be made between the PCDs of different parts of the structural data set. For example, it is clear from Figure 4 that the MD simulations approach some of the crystallographic pocket conformations more closely than others.

A second observation from Figure 4 is that the three MD simulations explore different regions of pocket conformational space. In particular, the second trajectory visits an extensive region of pocket space that is neither covered by the other two trajectories nor the crystallographic structures. This comparison of their respective PCDs indicates that the conformational sampling is incomplete in these 20 ns MD simulations. However, combining multiple short trajectories increases the likelihood of achieving a better coverage of the pocket shape space.[24,44]

To address the diverse pocket selection problem, the CLARA clustering algorithm was then applied to cluster the ALR pocket shapes according to their principal component scores. Here, ten clusters were identified based on the pockets' scores along the first 50 PCs, which are cumulatively responsible for 83% of the total variance in the MD data set (see Supporting Information Figure S4). Figure 5 shows the cluster representatives and indicates the number of members for each resulting cluster.

The ten cluster representatives span a diversity of pocket shapes. There are small, *apo*-like conformations such as clusters 2, 3, 5, 6, and 7. In contrast, the selectivity pocket is open in clusters 4, 8, and 9 and in fact adopts a different conformation in each case.[24] A collapsed pocket shape is apparent for cluster 10, whereas clusters 1 and 9 extend into the novel C298 subpocket that is involved in PC1 and PC2 (see Figure 3). This subpocket is of some interest because, unlike the selectivity pocket, it is not exploited by any of the ligands in the publicly available ALR crystal structures.

Notably, while at least two cluster representatives are drawn from each of the three MD simulations (two from the first, three from the second, and two from the third), three cluster

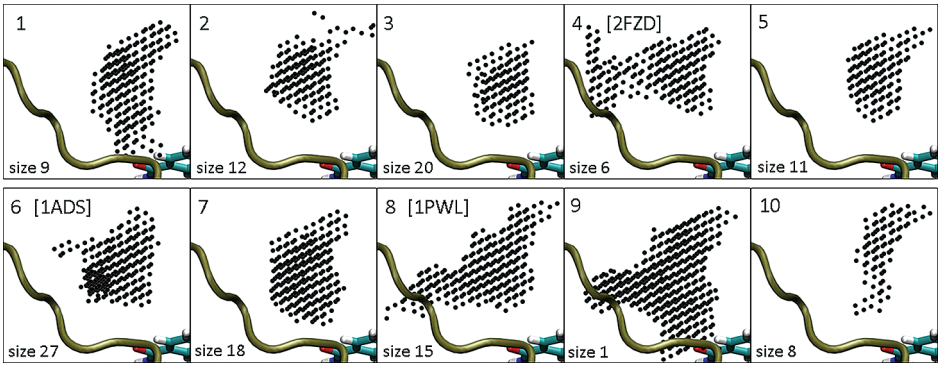


Figure 5. The diverse pocket selection for the active site of aldose reductase. The ten cluster representatives are shown along with the size of each cluster (number of members). Each representative is viewed from the same angle as Figure 2 but zoomed in to better discriminate between pocket shapes. For clarity, most of the protein is not shown — only the flexible L300 loop and the $NADP^+$ cofactor are included for reference.

Table 1. Druggability Analysis of the Ten ALR Cluster Representatives

cluster representative	source	DScore	size[a]	enclosure[b]	hydrophilicity[c]
1	MD 1	1.02	112	0.66	0.95
2	MD 1	0.94	88	0.65	1.02
3	MD 3	0.90	78	0.69	1.06
4	crystal	1.08	117	0.66	0.79
5	MD 3	0.97	124	0.66	1.12
6	crystal	0.98	95	0.68	1.06
7	MD 2	1.06	87	0.66	0.64
8	crystal	1.16	103	0.76	0.73
9	MD2	1.19	127	0.77	0.65
10	MD2	0.79	58	0.75	1.16

[a] Number of grid-points. [b] Fraction of radial rays which intersect the protein within a certain distance;[17] [c] In kcal mol^{-1} (but see SiteMap reference[17] for calibration).

representatives actually arise from crystallographically derived protein structures: number 4 is from 2FZD (*tolrestat*), number 6 is from 1ADS (*apo*), and number 8 is from 1PWL (*zenarestat*). The PocketAnalyzer^{PCA} diverse pocket selection has therefore automatically distinguished three of the four pocket conformations characterized by Sotriffer and co-workers.[24] The fourth crystallographic conformation (*IDD594*) is represented by structures 1US0 and 2R24 that are both members of cluster 8 along with 1PWL.

Within the context of the current clustering pattern, clusters which do not contain a single crystallographically derived pocket shape (which are all except clusters 4, 6, and 8) represent novel computationally derived ALR binding-site conformations. Undoubtedly, some of these "novel" conformations are in fact fairly similar to a pocket shape identified in one or other of the crystal structures. Others seem genuinely distinct, such as the collapsed pocket of cluster 10 or the pocket shapes expanding into the C298 subpocket as in cluster representatives 1 and 9.

Testing the Druggability of Novel Pocket Conformations. Although novel pocket conformations may indeed provide valuable opportunities for diversifying ligand design, the question arises as to whether any of these new computationally derived pocket shapes are realistically druggable. To address this issue, the druggability of all ten cluster representatives was analyzed with SiteMap[17] (see Materials and Methods section). The results are displayed in Table 1.

A higher DScore druggability metric is intended to indicate a more druggable pocket conformation, with a threshold of DScore = 0.98 taken to delineate "druggable" sites.[17] Pockets with DScore <0.83 are assumed to be "undruggable", and any in between are predicted to be "difficult". In Table 1 the MD-derived pocket conformations cover a wider range of druggabilities than the crystallographic pocket shapes. This is not entirely surprising, given that this observation compares conformations of protein−ligand cocrystal structures with those from ligand-free MD simulations. The more druggable computationally derived pockets are cluster representatives 1, 7, and 9, each of which has a DScore >1. Although cluster representative 7 is rather similar to *apo* conformations like 1ADS, the other two are more novel conformations that expand into the C298 subpocket. Cluster representative 9 is in fact much larger than is apparent from Figure 5; its full extent is shown in Figure 6. It is formed by a twist of the protein backbone at residues C298 and A299, with associated changes in side-chain conformation of C298 and Y209. Smaller adjustments of side-chain torsional angles occur for residues W112 and N160. The net effect is the opening of a channel past C298, underneath the L300 loop, connecting the ALR binding site with a second cavity as shown in Figure 6. The high DScore of this pocket derives in part from its size and in part from its low hydrophilic character.

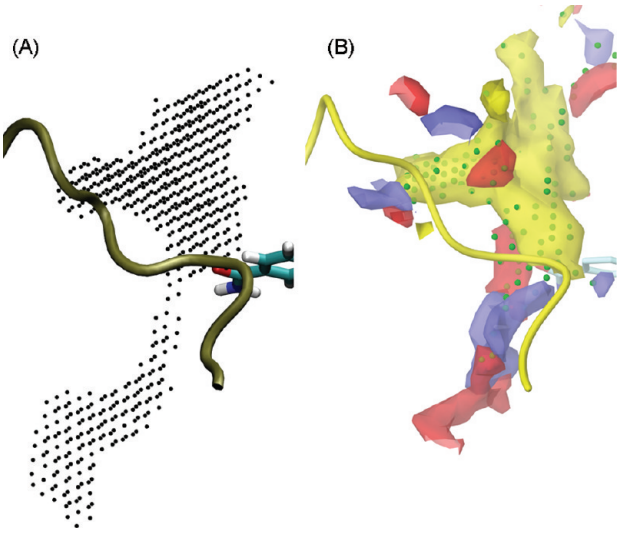


Figure 6. Full view of cluster representative 9 showing the open channel and the connection to a second cavity on the surface of the ALR protein structure. (A) The pocket shape identified by PocketAnalyzer. (B) The SiteMap analysis of the same protein structure, showing the hydrophobic (yellow), hydrogen-bond donor (blue), and hydrogen-bond acceptor (red) fields. SiteMap site points are represented as green spheres.

The predicted druggability of cluster representative 9 makes this computationally derived pocket conformation a potentially attractive target for ALR ligand design. Therefore, the pocket conformation was analyzed to determine if already known ALR inhibitors bind to this region. By comparing 52 publicly available ligand-bound wild-type ALR crystal structures, no ligand was found that entered the region of the novel C298 subpocket. This may suggest that in reality the cluster representative 9 pocket shape is rarely formed, or equivalently that it has an unfavorable free energy of formation. It seems plausible that the protein would have to adopt a relatively high energy conformation in order to open the C298 channel. Aspects such as protein strain and entropic changes are not explicitly incorporated into the current DScore druggability metric.

For the majority of known ALR inhibitors, however, there is no publicly available crystallographic information on the binding mode. Therefore, to address the question of whether these ligands might in fact occupy the C298 subpocket, those with structural similarity to a ligand with an available crystallographic binding mode were aligned to that template using the substructure-based ligand alignment method available in ICM.[25] The resulting ligand alignments provide predicted binding modes for a set of 66 substructure-matched ALR inhibitors (for details see Material and Methods).

Remarkably, the alignments of two lidorestat derivatives predict that they must bind to a pocket conformation similar to cluster

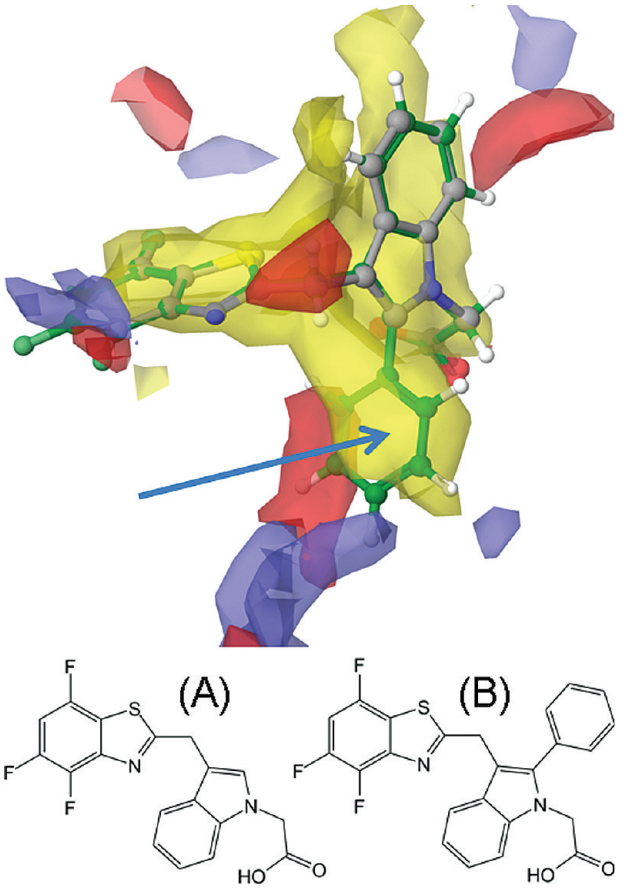


Figure 7. Ligand alignment of the 2-phenyl derivative of lidorestat (green; bindingDB monomerid = 16471) with the crystallographic binding mode of the parent compound from PDB structure 1Z3N (gray). The blue arrow indicates the occupation of the C298 subpocket by the 2-phenyl substituent. Lidorestat (A) and its 2-phenyl derivative (B) are sketched below the main figure. The SiteMap analysis of the cluster representative 9 is also shown in the same frame of reference following alignment of the associated protein structure with 1Z3N using the Align Binding Sites tool in *Maestro*.[13]

representative 9. The most striking example is shown in Figure 7, in which a phenyl substituent at the 2-position of the indole core of lidorestat is predicted to fill the hydrophobic part of the C298 subpocket. This phenyl-derivative of lidorestat has an IC_{50} of 100 nM in an *in vitro* ALR inhibition assay,[27] providing experimentally based evidence that the novel computationally derived pocket conformation identified by PocketAnalyzer^{PCA} is a plausible and druggable target for structure-based drug design and a potentially valuable opportunity to design novel ALR inhibitor chemotypes. Along these lines, since the phenyl-derivative of lidorestat also enters the more polar area toward the entrance to the channel, a hydrophilic meta or para substituent or the replacement of the phenyl ring with a nitrogen-containing heterocycle (e.g., pyridine, pyrrole, imidazole) might add further interactions with the protein.

PocketAnalyzer^{PCA} Applied to Neuraminidase. In a second example the PocketAnalyzer^{PCA} protocol was applied to viral neuraminidase subtype N1, which is a target for antiviral inhibitors useful in the treatment of influenza. Current drugs on the market include zanamivir and oseltamivir, both of which bind to the main catalytic (sialic acid-binding or SA) site of the enzyme. X-ray crystallography and extended MD simulations have revealed significant conformational mobility in this region, which has not been observed for other neuraminidase subtypes.[45] In particular, the 150-loop can adopt an open or closed conformation; in its open conformation, it gives access to a second cavity, adjacent to the SA site. A third cavity in the neuraminidase binding site has been identified by means of MD simulations close to the 430-loop.[7] In the following we therefore refer to (i) the SA-cavity where oseltamivir binds, (ii) the 150-cavity, and (iii) the 430-cavity. Even though no current drugs are known to bind to these latter two cavities, a technique called computational solvent mapping has identified low energy binding-sites for small molecular probes in both regions,[46] and a virtual screening campaign has identified several new chemotypes which could potentially target these sites.[7]

Recently, McCammon and co-workers have characterized the conformational mobility of the neuraminidase active site using extensive MD simulations.[47,48] In their work, all-atom root-mean-square deviation-based (rmsd) clustering on a subset of 62 residues lining the active site was used to identify and compare clusters of pocket conformations.[7] Here we show that similar results and insight about the binding site flexibility can be gained by applying the PocketAnalyzer^{PCA} diverse pocket selection protocol.

Diverse Pocket Selection. The PocketAnalyzer^{PCA} protocol was used to analyze and cluster snapshots extracted from three N1 MD simulations, including one *apo* and two *holo* simulations (see Materials and Methods). First, the aligned protein conformations are processed with the PocketAnalyzer algorithm and the grid-point inclusion/exclusion row vectors corresponding to the 600 computationally derived structures (3 trajectories, 200 frames from each) are fed into the PCA. The row vectors corresponding to nine crystallographic pocket conformations (see Material and Methods section) are withheld from the PCA but subsequently projected onto the resulting PCs. In Figure 8, the first two principal components, cumulatively responsible for 25% of the total variance, show that the most variable regions involve the 430-cavity and the 150-cavity, in agreement with the findings of Amaro et al.[47] The SA-cavity is more conserved, but in the *apo* simulation a side-chain movement of Arg152 causes some variation along a channel adjacent to the 150-loop. The PCD for the neuraminidase data set along PC1 and PC2 illustrates the different regions of pocket space explored by the three MD simulations (Supporting Information Figure S6).

To extract a representative set of pocket shapes we clustered the pockets calculated from the three MD simulations using their scores along the first 50 principal components, which are cumulatively responsible for 63% of the total variance in this data set (see Supporting Information Figure S5). This approach differs from the earlier rmsd-based clustering[7] in that rather than clustering each simulation separately, all 600 snapshots were analyzed simultaneously. We furthermore decided to define only 10 clusters to facilitate further processing and visual inspection of the resulting representative structures. It would however be straightforward to define a larger number of clusters or to extract additional structures from a given cluster of interest.

An overlay of the MD snapshots corresponding to the cluster representatives is shown in Figure 9 and compared to the corresponding overlay of the crystallographic protein structures. The most striking difference between the experimental and computational ensembles is the much larger conformational variability in the 430-loop in the MD simulations. In previous simulations,[47] the most dramatic conformational changes have

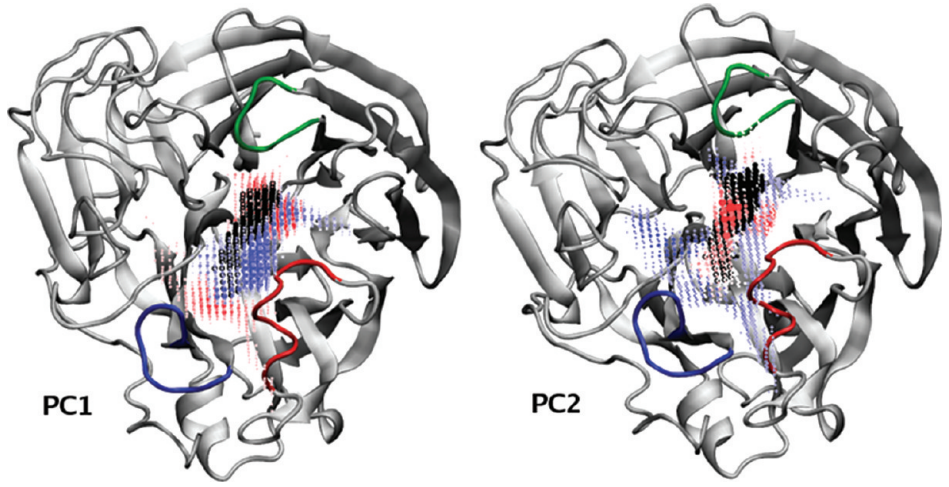


Figure 8. The highest variance principal components of the neuraminidase data set. The grid-points are sized and colored in the same way as in Figure 3. The 150-loop is shown in red, the 430-loop is shown in blue, and the SA-loop is shown in green. The latter comprises residues Pro245 to Ala250.

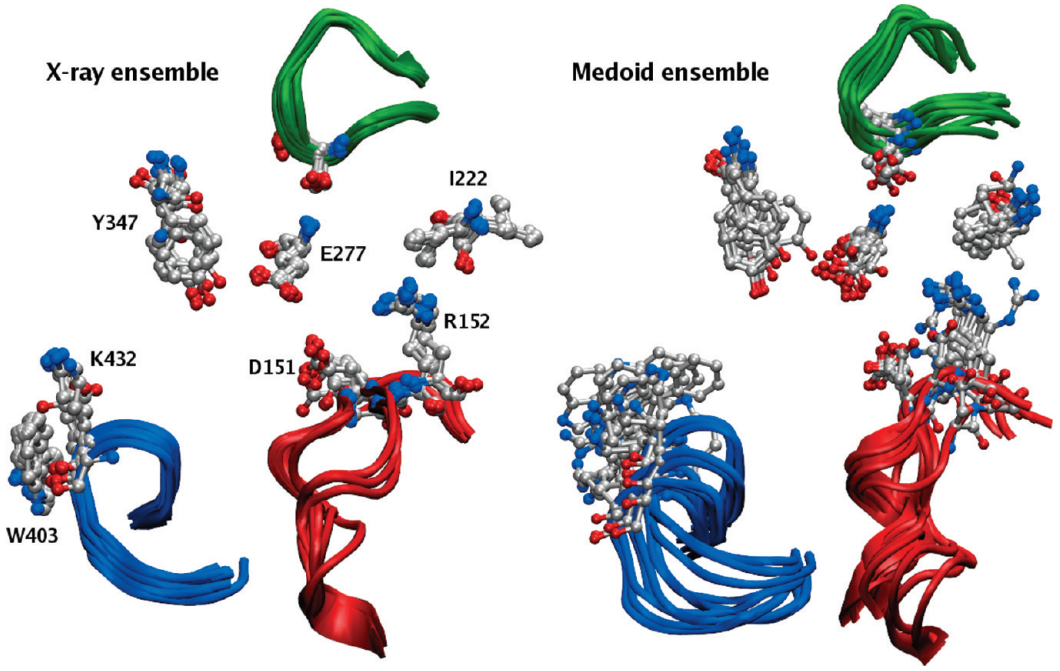


Figure 9. Left: Superposition of a representative set of *apo* and *holo* X-ray structures of N1 and N8 neuraminidases (see Material and Methods section). Right: Superposition of MD-derived cluster representatives selected by PocketAnalyzer[PCA]. The 150-, 430-, and SA-loops are color-coded as in Figure 8. Selected active site residues are shown in ball and stick representation.

been observed in the *apo* simulations, with the 430-loop reaching a wide-open conformation that results in a 430-loop C_α rmsd of 4 Å and a markedly increased solvent accessible surface area. In our case, it is the *holo* simulations that display the larger variations in the 430-loop, with an average 430-loop C_α rmsd of 4 Å (Supporting Information Figure S7, middle and lower panels). The opening of the 430-loop is accompanied by a positional change of Lys432 and Trp403, which swap positions (Figure 9). The 150-loop is more stable in the *apo* and first *holo* simulations, whereas it transitions into a wide-open state in the second *holo* simulation, leading to a final 150-loop C_α rmsd of 4 Å. This is again in agreement with previous MD simulations and leads to the opening of the 150-cavity.

More subtle differences in pocket-shapes are due to side-chain movements and are apparent in the pocket shapes shown in Figure 10. Clusters representatives 2, 3, and 4 originate from the *apo* simulation and reflect the absence of the extensive protein—ligand hydrogen-bond and salt-bridge network that exists in the oseltamivir-bound *holo* simulations. In the latter, the positively charged exocyclic amino group on the ligand interacts with the conserved Glu119 and Asp151 of the 150-loop. In addition, the ligand's carboxylic acid functionality interacts with Arg292, Arg371, and Tyr347. In the absence of the ligand, Asp151 is more mobile and moves outward, leading to a larger central pocket shape. Cluster representatives 8, 9, and 10 expand significantly into the 150-cavity; they all originate from the second *holo*

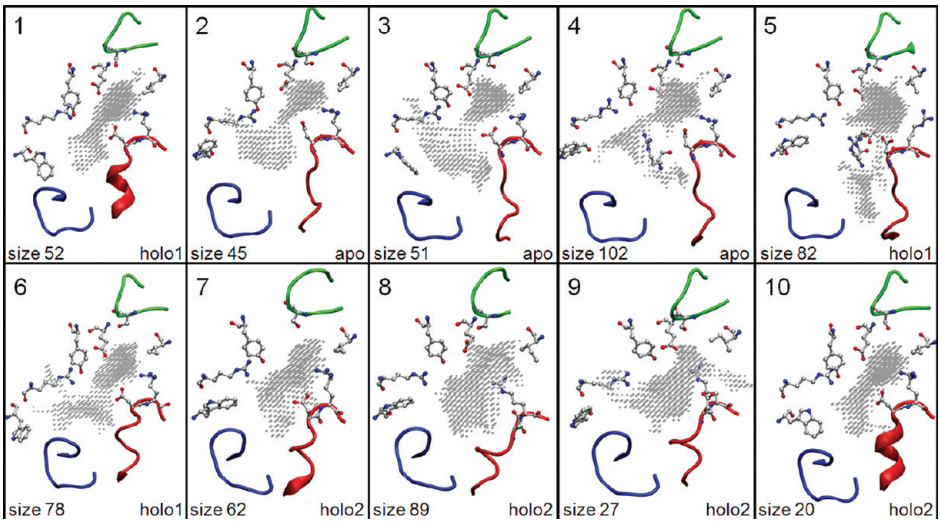


Figure 10. The diverse pocket selection for the neuraminidase data set. Only the 150 (red), 430 (blue), and SA-loop (green) are shown for clarity. A few residues lining the binding pocket are displayed in ball and stick representation in order to highlight side-chain movements that lead to changes in the pocket shape, such as Tyr347 and Asp151. The size of the cluster and source of its representative are indicated to the left and to the right of each pocket shape.

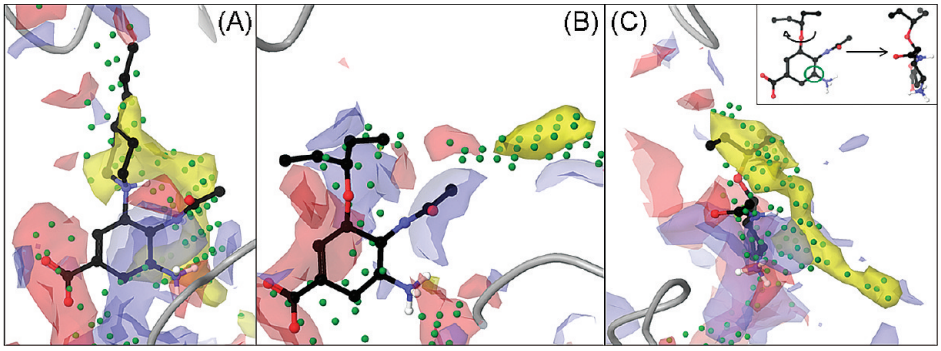


Figure 11. SiteMap analysis of three cluster representatives exhibiting small novel subpockets around the main SA-cavity. Fields are color-coded as for Figure 6. Green spheres represent the SiteMap site-points. For orientation, the SA-loop and part of the 150-loop are shown in gray ribbon at the top and bottom of each panel. (A) Cluster representative 6 showing the extension toward the SA-loop at the top of the figure. An oseltamivir derivative predicted to exploit this subpocket is superimposed (see text; bindingDB monomerid = 5261). (B) Cluster representative 4 showing the open channel and the hydrophobic patch (yellow) adjacent to the 150-loop. The channel runs horizontally across the middle of the figure. Oseltamivir is superimposed for reference. (C) Cluster representative 5 showing the opening of a hydrophobic tunnel in the floor of the catalytic site. This figure is rotated roughly 90° compared to the other panels, as illustrated for oseltamivir by the curly arrow in the inset. The green circle in the inset indicates the possible attachment point discussed in the text.

simulation which has the largest 150-loop rmsd. Pockets that extend into the 430-cavity were extracted from all three simulations, consider for example cluster representatives 1 (first *holo* simulation), 4 (*apo* simulation), and 7, 8, and 9 (second *holo* simulation).

To summarize the section up to this point, pocket shapes have been identified which correspond to: (a) the wide-open conformation of the 150-loop and (b) the wide-open conformation of the 430-loop. The PocketAnalyzer^{PCA} results thus broadly match the all-atom rmsd-based analysis of Cheng et al.[7]

Testing the Druggability of Novel Pocket Conformations. Of the ten pocket shapes in the diverse pocket selection in Figure 10, numbers 4, 5, and 6 were found to be of particular interest as novel (and potentially druggable) binding-site conformations of N1 neuraminidase. In making this selection, we have neglected pocket shapes that merely show expansion into the 150- and 430-cavities since similar binding-site conformations have been discovered and described in the earlier work discussed above. Instead, we focus on more subtle variations caused by the opening of small but distinct subpockets around the main SA-cavity.

The first of these is exemplified by cluster representative 6 and involves an extension of the pocket toward the SA-loop (see Figure 10). This coincides with the creation of a strongly hydrophobic patch in this region of the protein, which is at least partially responsible for cluster representative 6 having a relatively high predicted druggability; its SiteMap DScore is 1.05 as opposed to a median DScore of 0.97 (minimum 0.96, maximum 1.04) among the crystallographic binding-site conformations (see Supporting Information Table S8).

As for aldose reductase, an obvious question is whether any experimentally confirmed inhibitors of N1 neuraminidase might exploit this SA-loop subpocket. As above, we used structural alignment of known inhibitors to crystallographically derived binding-modes of structurally similar ligands to address this issue (for details see Materials and Methods). The result-

dx.doi.org/10.1021/ci200168b |*J. Chem. Inf. Model.* 2011, 51, 2666–2679

ing alignments provide predicted binding modes for 64 substructure-matched derivatives of oseltamivir. Within this set of submicromolar inhibitors, 95% involve a variation of the hydrophobic ether substituent that is directed toward the SA-loop in the crystallographic binding-mode of oseltamivir. The substructure alignments indeed predict that fire of the molecules occupy the novel SA-loop subpocket to some extent. One example is the 4-propylpiperidine derivative that is superimposed on the fields from the SiteMap analysis of cluster representative 6 in panel A of Figure 11. This molecule has an IC_{50} of 40 nM in an enzymatic N1 inhibition assay. Indeed, the report describing the synthesis of this compound also includes an illustration of a proprietary crystal structure of the related piperidine derivative.[49] This confirms the binding mode of these compounds and provides experimental support for plausibility of this novel binding-site conformation identified by PocketAnalyzer^{PCA}.

A second variation of the SA-cavity is exhibited by cluster representative 4 and is induced by a change in the conformation of the side-chain of Arg152. In particular, the guanidine group of Arg152 moves toward, and forms a salt bridge with, the side-chain of Glu277. This rearrangement opens a narrow channel that is adjacent to the 150-loop (Figure 11, panel B). Interestingly, this rotation of Arg152 side-chain would not be compatible with the binding modes of known SA-cavity ligands (e.g., oseltamivir and zanamivir) and, indeed, is only observed in the *apo* MD simulation. Furthermore, although the novel channel includes a moderately sized hydrophobic patch, the rearrangement leads to an overall decrease in depth and size of the catalytic binding side and hence a relatively low DScore of 0.92 (see Supporting Information Table S8). None of the substructure-matched oseltamivir derivatives are predicted to occupy this channel by the structural alignment approach discussed above. However, some support for the existence of this pocket conformation is provided by the earlier computational solvent mapping calculations.[46] In that work it was observed that a low energy cluster of chemical probe positions is found in the region of this channel.

Cluster representative 5 (Figure 11, panel C) exemplifies a third conformation of the SA cavity that, to the best of our knowledge, has not been previously observed in either experimental or computational studies. It involves the opening of a narrow but hydrophobic tunnel in the floor of the catalytic site. Again, none of the publicly available oseltamivir derivatives are predicted to enter this tunnel by the structural alignment approach. However, the tunnel is also present to some extent in cluster representatives 1, 2, 3, and 6, so it is a characteristic and recurrent feature of the simulations. The addition of several rather hydrophobic site-points results in high predicted druggability with a DScore of 1.08. Furthermore, the central cyclohexene ring of oseltamivir has an obvious attachment point for hydrophobic substituents that might access this tunnel (Figure 11). It will therefore be of interest to see if future medicinal chemistry studies report evidence of inhibitors binding to this novel pocket conformation.

■ CONCLUDING REMARKS

We have introduced the PocketAnalyzer^{PCA} methodology to address the problem of diverse pocket selection, i.e. how to reduce a large collection of structures of the same protein to a subset that retains a number of substantially distinct binding pocket conformations. Diverse pocket shapes drive medicinal chemistry to explore a broader chemical space and so present additional opportunities to overcome key drug discovery issues such as potency, selectivity, toxicity, and pharmacokinetics. The identification of diverse pocket shapes and novel binding-site conformations can therefore greatly assist the progress of structure-based ligand design projects.

The PocketAnalyzer^{PCA} approach combines a grid-based pocket detection algorithm with PCA and clustering. The resulting principal component (PC) eigenvectors reveal the dominant binding-site deformation modes within an ensemble of protein structures, and the corresponding PC scores provide characterization and visualization of the pocket conformational distributions. From a methodological point of view, the PocketAnalyzer^{PCA} approach provides a novel and complementary perspective on protein dynamics that may prove particularly relevant for ligand binding and drug design. PocketAnalyzer^{PCA} was primarily envisioned as a tool for analyzing trajectories of protein conformations produced by MD simulations. However, the procedure is applicable to any source of atomistic protein structural information and also to combinations of structures from several such sources, as in the examples presented above. Therefore, PocketAnalyzer^{PCA} may be useful for exploring the increasing volume of experimentally derived structural information resulting from high-throughput crystallography and advances in NMR-based techniques.

A technically related approach to a different problem combines pocket detection and clustering to track the opening and closing of transient binding pockets in protein−protein interaction surfaces along MD trajectories.[50] This ePOS method analyses each in a sequence of MD frames using the PASS pocket detection algorithm[51] and clusters the resulting pockets by the set of pocket lining atoms to define unique pockets and track their opening and closing as time progresses. The recently announced fpocket Web server uses a different pocket detection algorithm to perform a similar analysis.[52] A precedent for combining grid-based pocket characterization with PCA is the GRID/CPCA approach of Kastenholz et al.[53,54] However, rather than focusing on the analysis of many structures of the same protein (as here) the GRID/CPCA method is directed at comparing structures of different targets to derive insights that assist in improving compound selectivity.

When applied to aldose reductase, a protein with moderate binding-site flexibility and a well-characterized set of crystallographic binding-site conformations, PocketAnalyzer^{PCA} distinguishes three of the four crystallographically observed binding-site conformations previously reported by the Klebe group.[24] In addition, the approach identifies a number of distinct pocket shapes that have not been observed experimentally and which therefore represent novel computationally derived binding-site conformations. From a medicinal chemistry point of view, the most outstanding result is that one MD-derived pocket shape is particularly striking in its difference to the crystallographic conformations. A rotation of a short section of the protein backbone and accompanying adjustments in the positions of a few amino-acid side-chains open a channel connecting the active site with another pocket on the protein surface. Although the channel itself and the second pocket are rather polar, the 'entrance' to the channel forms a reasonably large hydrophobic subpocket, and SiteMap[17] analysis predicted good druggability for this novel conformation of the ALR active site. Indeed, subsequent alignment of known ALR inhibitors to the crystallographic binding modes of structurally similar ALR ligands identified a derivative of lidorestat that is predicted to fill the novel hydrophobic subpocket with a phenyl ring. This compound has an IC_{50} of

100 nM in an *in vitro* ALR inhibition assay,[27] providing experimental evidence that the novel computational derived pocket conformation identified by PocketAnalyzer^{PCA} is a plausible and druggable target for structure-based drug design against ALR.

In a second example, the PocketAnalyzer^{PCA} approach is used to derive a diverse set of binding-site conformations from viral neuraminidase. Similarly to ALR, the binding-site flexibility of neuraminidase is reasonably well-established, for example as a result of MD simulations,[48] and a number of distinct binding-site conformations have been characterized crystallographically. The PocketAnalyzer^{PCA} diverse pocket approach was found to identify a qualitatively similar range of binding-site conformations as a previously reported atom-based rmsd clustering method,[47,48] with the advantage of quickly highlighting conserved and variable regions in the pocket. The method also allows facile comparison of structures from different sources and direct visualization of differences in pocket shape rather than changes in proxy descriptors such as backbone and side-chain positions.

The N1 diverse pocket selection included three particularly interesting and novel subpockets adjacent to the main catalytic site of N1 neuraminidase. Alignment of known submicromolar N1 inhibitors to the crystallographic binding-mode of oseltamivir identified several molecules predicted to occupy the first of these subpockets, for example with a 4-propylpiperidine substituent. Indeed, the report describing the synthesis of this compound includes an illustration of a proprietary crystal structure of the unsubstituted piperidine derivative,[49] confirming the binding mode and the plausibility of this computationally derived binding-site conformation as a druggable target for N1 inhibition.

In addition to direct application to structure-based ligand design, the PocketAnalyzer^{PCA} protocol produces an ensemble of protein structures incorporating diverse and potentially novel pocket shapes that could be useful as input to numerous structure-based drug design methods.[55] For example, this would provide an effective way to account for protein flexibility in docking and virtual screening,[56−59] receptor-based pharmacophore generation,[3] and druggability analysis.[60] Applied in this way, PocketAnalyzer^{PCA} would be just one component in a larger drug discovery workflow, providing a rational approach to selecting an ensemble of protein conformations.[61,62]

Moving beyond the specific application of PocketAnalyzer^{PCA} to diverse pocket selection, in the future the approach may be more broadly applied to address questions regarding the effect of various perturbations on pocket conformational distributions. For example, when coupled with the appropriate MD simulations, PocketAnalyzer^{PCA} may provide a useful perspective on the change in binding-site conformation induced by factors such as mutation, allosteric modulation, solvent pH, and post-translational modification.

■ ASSOCIATED CONTENT

Supporting Information. Additional PCD plots and PCA scree plots for the ALR and NA data sets, rmsd plots for all MD simulations, and details of the oseltamivir force field parametrization. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ianrcraig@gmail.com.

Present Addresses

[‖]BASF SE, GVC/C - A030, 67056 Ludwigshafen, Germany.

■ ACKNOWLEDGMENT

We are grateful to Teresa Jimenez Vaquero for providing an initial version of the PocketAnalyzer code. Financial support from the Education Office of Novartis Institutes for Biomedical Research is gratefully acknowledged (I.C.).

■ ABBREVIATIONS:

ALR, aldose reductase; NA, neuraminidase; SA, sialic acid; PCA, principal component analysis; PC, principal component; PCD, pocket conformational distribution; MD, molecular dynamics; GAFF, generalized amber force field; rmsd, root mean square deviation

■ REFERENCES

(1) Dutta, S.; Burkhardt, K.; Young, J.; Swaminathan, G.; Matsuura, T.; Henrick, K.; Nakamura, H.; Berman, H. Data Deposition and Annotation at the Worldwide Protein Data Bank. *Mol. Biotechnol.* **2009**, *42*, 1–13.

(2) Ahmed, A.; Kazemi, S.; Gohlke, H. Protein flexibility and mobility in structure-based drug design. *Front. Drug Des. Discovery* **2007**, *3*, 455–476.

(3) Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. Developing a dynamic pharmacophore model for HIV-1 integrase. *J. Med. Chem.* **2000**, *43*, 2100–2114.

(4) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, *51*, 6237–6255.

(5) Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* **2004**, *47*, 1879–1881.

(6) Sotriffer, C. A.; Ni, H.; McCammon, J. A. Active site binding modes of HIV-1 integrase inhibitors. *J. Med. Chem.* **2000**, *43*, 4109–4117.

(7) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878–3894.

(8) Zhou, Z.; Madrid, M.; Evanseck, J. D.; Madura, J. D. Effect of a bound non-nucleoside RT inhibitor on the dynamics of wild-type and mutant HIV-1 reverse transcriptase. *J. Am. Chem. Soc.* **2005**, *127*, 17253–17260.

(9) Grant, B. J.; Rodrigues, A. P. C.; Elsawy, K. M.; McCammon, J. A.; Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **2006**, *22*, 2695–2696.

(10) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15*, 359–363.

(11) Stahl, M.; Taroni, C.; Schneider, G. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng.* **2000**, *13*, 83–88.

(12) Joliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer: New York, 2002.

(13) *Maestro*, version 9.1; Schrödinger, LLC: New York, 2010.

(14) Godzik, A. The structural alignment between two proteins: Is there a unique answer? *Protein Sci.* **1996**, *5*, 1325–1338.

(15) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, 1990.

(16) Available at http://cran.r-project.org/web/packages/cluster (accessed September 27, 2011).
(17) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.
(18) Carl, N.; Konc, J.; Janezic, D. Protein surface conservation in binding sites. *J. Chem. Inf. Model.* **2008**, *48*, 1279–1286.
(19) Carl, N.; Konc, J.; Vehar, B.; Janezic, D. Protein-Protein Binding Site Prediction by Local Structural Alignment. *J. Chem. Inf. Model.* **2010**, *50*, 1906–1913.
(20) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.
(21) Coleman, R. G.; Burr, M. A.; Souvaine, D. L.; Cheng, A. C. An intuitive approach to measuring protein surface curvature. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 1068–1074.
(22) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* **2003**, *326*, 607–620.
(23) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.
(24) Sotriffer, C. A.; Kramer, O.; Klebe, G. Probing flexibility and "induced-fit" phenomena in aldose reductase by comparative crystal structure analysis and molecular dynamics simulations. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 52–66.
(25) *ICM*, version 3.6; MolSoft, LCC; La Jolla, CA, 2010.
(26) Liu, T. Q.; Lin, Y. M.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.
(27) Van Zandt, M. C.; Jones, M. L.; Gunn, D. E.; Geraci, L. S.; Jones, J. H.; Sawicki, D. R.; Sredy, J.; Jacot, J. L.; DiCioccio, A. T.; Petrova, T.; Mitschler, A.; Podjarny, A. D. Discovery of 3-[(4,5,7-trifluorobenzothiazol-2-yl)methyl]indole-N-acetic acid (Lidorestat) and congeners as highly potent and selective inhibitors of aldose reductase for treatment of chronic diabetic complications. *J. Med. Chem.* **2005**, *48*, 3141–3152.
(28) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
(29) Case, D. A.; Darden, T. A.; Cheatham; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *Amber 9*; University of California: San Francisco, 2006.
(30) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. 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.
(31) Bryce, R. AMBER parameter database. http://www.pharmacy.manchester.ac.uk/bryce/amber (accessed March 15, 2010).
(32) 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.
(33) Izaguirre, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **2001**, *114*, 2090–2098.
(34) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of A System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
(35) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald - An N. Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
(36) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-Dynamics with Coupling to An External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
(37) Lawrenz, M.; Wereszczynski, J.; Amaro, R.; Walker, R.; Roitberg, A.; McCammon, J. A. Impact of calcium on N1 influenza neuraminidase dynamics and binding free energy. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 2523–2532.
(38) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
(39) Varkonyi, T.; Kempler, P. Diabetic neuropathy: new strategies for treatment. *Diabetes Obes. Metab.* **2008**, *10*, 99–108.
(40) Steuber, H.; Zentgraf, M.; Gerlach, C.; Sotriffer, C. A.; Heine, A.; Klebe, G. Expect the unexpected or caveat for drug designers: Multiple structure determinations using aldose reductase crystals treated under varying soaking and co-crystallisation conditions. *J. Mol. Biol.* **2006**, *363*, 174–187.
(41) Steuber, H.; Zentgraf, M.; Podjarny, A.; Heine, A.; Klebe, G. High-resolution crystal structure of aldose reductase complexed with the novel sulfonyl-pyridazinone inhibitor exhibiting an alternative active site anchoring group. *J. Mol. Biol.* **2006**, *356*, 45–56.
(42) Steuber, H.; Zentgraf, M.; La Motta, C.; Sartini, S.; Heine, A.; Klebe, G. Evidence for a novel binding site conformer of aldose reductase in ligand-bound state. *J. Mol. Biol.* **2007**, *369*, 186–197.
(43) Luque, I.; Freire, E. Structural stability of binding sites: Consequences for binding affinity and allosteric effects. *Proteins: Struct., Funct., Genet.* **2000**, 63–71.
(44) Caves, L. S. D.; Evanseck, J. D.; Karplus, M. Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Protein Sci.* **1998**, *7*, 649–666.
(45) Russell, R. J.; Haire, L. F.; Stevens, D. J.; Collins, P. J.; Lin, Y. P.; Blackburn, G. M.; Hay, A. J.; Gamblin, S. J.; Skehel, J. J. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **2006**, *443*, 45–49.
(46) Landon, M. R.; Amaro, R. E.; Baron, R.; Ngan, C. H.; Ozonoff, D.; McCammon, J. A.; Vajda, S. Novel druggable hot spots in avian influenza neuraminidase H5N1 revealed by computational solvent mapping of a reduced and representative receptor ensemble. *Chem. Biol. Drug Des.* **2008**, *71*, 106–116.
(47) Amaro, R. E.; Minh, D. D. L.; Cheng, L. S.; Lindstrom, W. M.; Olson, A. J.; Lin, J. H.; Li, W. W.; McCammon, J. A. Remarkable loop flexibility in avian influenza N1 and its implications for antiviral drug design. *J. Am. Chem. Soc.* **2007**, *129*, 7764.
(48) Amaro, R. E.; Xiaolin, C.; Ivaylo, I.; Dong, X.; McCammon, J. A. Characterizing Loop Dynamics and Ligand Recognition in Human and Avian Type Influenza Neuraminidases via Generalized Born Molecular Dynamics and End-Point Free Energy Calculations. *J. Am. Chem. Soc.* **2009**, *131*, 4702–4709.
(49) Lew, W.; Wu, H. W.; Chen, X. W.; Graves, B. J.; Escarpe, P. A.; MacArthur, H. L.; Mendel, D. B.; Kim, C. U. Carbocyclic influenza neuraminidase inhibitors possessing a C-3-cyclic amine side chain: Synthesis and inhibitory activity. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1257–1260.
(50) Eyrisch, S.; Helms, V. Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.* **2007**, *50*, 3457–3464.
(51) Brady, G. P.; Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383–401.
(52) Schmidtke, P.; Le Guilloux, V.; Maupetit, J.; Tuffery, P. fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **2010**, *38*, W582–W589.
(53) Afzelius, L.; Raubacher, F.; Karlen, A.; Jorgensen, F. S.; Andersson, T. B.; Masimirembwa, C. M.; Zamora, I. Structural analysis of CYP2C9 and CYP2C5 and an evaluation of commonly used molecular modeling techniques. *Drug Metab. Dispos.* **2004**, *32*, 1218–1229.
(54) Kastenholz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. E. J.; Fox, T. GRID/CPCA: A new computational tool to design selective ligands. *J. Med. Chem.* **2000**, *43*, 3033–3044.
(55) Perot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A. C.; Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discovery Today* **2010**, *15*, 656–667.

(56) Barril, X.; Morley, S. D. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* **2005**, *48*, 4432–4443.

(57) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511–524.

(58) Huang, S. Y.; Zou, X. Q. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 399–421.

(59) Virtanen, S. I.; Pentikainen, O. T. Efficient Virtual Screening Using Multiple Protein Conformations Described as Negative Images of the Ligand-Binding Site. *J. Chem. Inf. Model.* **2010**, *50*, 1005–1011.

(60) Egner, U.; Hillig, R. C. A structural biology view of target drugability. *Expert Opin. Drug Discovery* **2008**, *3*, 391–401.

(61) Bolstad, E. S. D.; Anderson, A. C. In pursuit of virtual lead optimization: Pruning ensembles of receptor structures for increased efficiency and accuracy during docking. *Proteins: Struct., Funct., Bioinf.* **2009**, *75*, 62–74.

(62) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *J. Chem. Inf. Model.* **2010**, *50*, 186–193.