2D Depiction of Protein−Ligand Complexes

Alex M. Clark* and Paul Labute

Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montréal, Québec, Canada H3A 2R7

Received April 24, 2007

A method is presented for the automated preparation of schematic diagrams for protein−ligand complexes, in which the ligand is displayed in conventional 2D form, and the interactions to and between the residues in its vicinity are summarized in a concise and information-rich manner. The structural entities are arranged to maximize aesthetic ideals and to properly convey important distance relationships. The diagram is annotated with calculated hydrogen bonds, a substitution contour, solvent exposure, chelated metals, covalently bound linkages, $\pi-\pi$ and $\pi-$cation interactions, and, for series of complexes, conserved residues and interactions. Residues, cofactors, ions, and solvent components are drawn in cartoon form as adjuncts to the ligand. The method can be applied to aligned sets which contain multiple ligands, or multiple members of a protein family, in which case the ligand orientations and protein residue placement will show consistent trends throughout the series.

INTRODUCTION

The ready availability of structural data for protein−ligand complexes, both from experimental determination, such as by X-ray crystallography, and by postulated models, such as computational docking, has made it necessary to find ways to easily interpret the results. The spatial orientation of the ligand inside the binding pocket and the pertinent interactions occurring between the constituent species provide important cues in efforts to design new and better drugs.

While a large number of software packages are available for visualizing the 3D structures, the options for producing a schematic 2D summary of a protein−ligand complex are comparatively few. In practice, the 2D and 3D visualization methods are complementary. A rotatable stereographic 3D view provides a faithful representation of the available data but requires considerable operator time and practice to thoroughly perceive and is often an inconvenient way to communicate information among peers. A flattened schematic representation can be designed to be very quickly understood and can be readily shared as straightforward picture data or printed hardcopy. However, the relevance of the features which are highlighted in the schematic form are only as good as the discretion of the artist who produced the diagram, and three-dimensional conformations are typically not easy to faithfully reproduce in two dimensions.

The popular program LIGPLOT[1] has been commonly used to automate the process of drawing such diagrams, as an alternative to studying the system using 3D tools and manually producing a 2D pictorial equivalent. In this work, we revisit the objective of producing an aesthetically pleasing 2D diagram of a protein−ligand complex, with the goal of conveying representative geometric positioning and summarizing pertinent interaction features. More recent attempts to achieve these objectives favor a ligand-centric algorithm,[2] which we also use in order to provide a strong emphasis on stylistic clarity. This in turn allows a high volume of information to be condensed into a diagram which remains easy for the beholder to interpret.

Considering also that protein−ligand interaction data are frequently encountered in series, such as crystal structures of the same protein with different ligands, homologous protein series, or multiple poses produced by docking results, we have anticipated the need to produce a corresponding series of diagrams in which the relative orientations of the ligands are shown, and the positions of the protein residues remain fixed. In this way, we can also annotate differing residues and conserved/nonconserved interactions across the series, without requiring the observer to mentally realign the images.

In the remainder of this work, we describe how the algorithm operates, beginning with a molecular representation of protein and ligand structures, and culminating with a planar layout, giving rise to a fully annotated diagram. We will then discuss how these features can be used to clearly understand and communicate structural data relevant to drug design efforts.

METHODS

An example of the result of the algorithm which we present in this work is shown in Figure 1, which represents a bound PDE 5 inhibitor (PDB code 1UDT).[3,4] As can be seen, the central feature of the diagram is the ligand, which is drawn using conventional aesthetics, in a way that balances reproduction of the actual 3D orientation against visual clarity. The protein residues and other active site constituents such as solvent molecules are arranged around the ligand in cartoon form, with hydrogen-bonding interactions shown as dotted lines. Other information about the spatial environment around the ligand and protein atoms is encoded in the form of a substitution proximity contour and hues denoting solvent accessible surface area.

* To whom correspondence should be addressed. Phone: (514) 393-1055. Fax: (514) 874-9538. E-mail: aclark@chemcomp.com.

10.1021/ci7001473 CCC: $37.00
Published on Web 08/23/2007

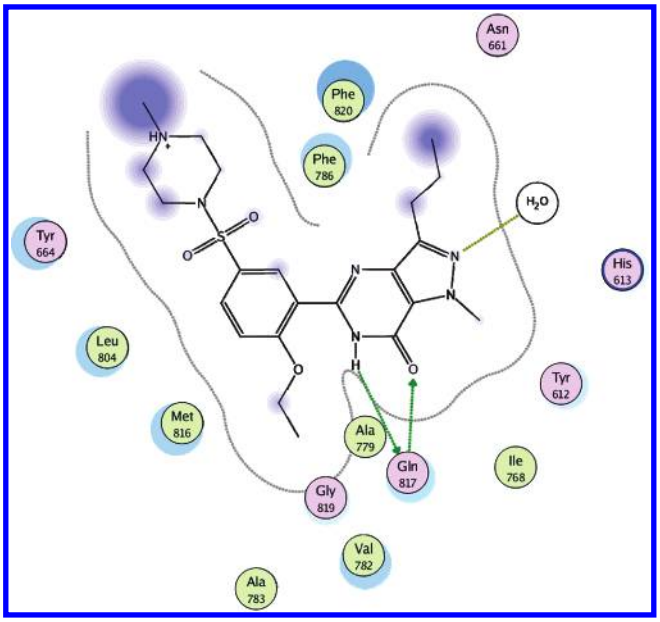


Figure 1. PDE5 inhibitor (PDB code 1UDT), rendered using default settings. A total of 14 protein residues and one solvent molecule are shown in proximity around the ligand, with hydrogen-bonding interactions shown where detected. The ligand is further annotated with a proximity contour and magnitude of solvent exposure, while the protein residues are annotated with a reduction of solvent exposure.

The remainder of this section will discuss the manner in which these properties are calculated and presented.

Implementation. The steps described in this section have been implemented using Scientific Vector Language (SVL) within Molecular Operating Environment (MOE).[5] The algorithms used are either new to this work or exist as standard callable SVL functions. The application described in this work has been a part of the MOE platform since the 2006.08 version. In more recent versions, the protein−ligand diagrams can be accessed programmatically as well as through an interactive panel. The only necessary inputs are the protein and the ligand, and they can originate from any format which the MOE platform supports (e.g., PDB and mmCIF).[6] The pictorial output can be captured as a raster bitmap image or in a vector format. The typical calculation time for the layout of a single ligand interaction diagram is on the order of 1 s, using a contemporary 2 GHz x86 processor.

Classification and Interactions. Prior to any of the layout and rendering steps, it is necessary to categorize the molecular ensemble, decide which pieces belong in the diagram, and identify the interactions which will be prominently annotated.

For purposes of this discussion, it is assumed that a single ligand is clearly identified and consists of a monoconnected component; that protein residues are indicated and named according to conventional amino acid codes; and that other extraneous fragments can be recognized as ions, solvent molecules, or cofactors. This information may be provided as a parameter, or established methods can be used to distinguish between them.

All protein residues containing atoms which approach within a certain distance of the ligand are flagged for inclusion in the diagram by virtue of proximity, even if there is no obvious interaction. The default cutoff is 4.5 Å (heavy atoms only).

Atom pairs are then studied for the possible existence of hydrogen bonds. These are evaluated using an empirical type-based scoring function, which was trained using statistics derived from examining contacts in the RCSB Protein Data Bank. Each plausible atom pair is scored in terms of the percentage likelihood of being a geometrically ideal hydrogen bond. The score is expressed as a function of atom type and immediate bonding environment, distance, and angular orientation. A user-supplied threshold, which defaults to 10%, is used to determine which interactions qualify and which are not considered important. For purposes of subsequent annotation, these interactions are classified according to whether the protein atom was part of the amide backbone or the side chain and to the direction of the hydrogen bond.

Hydrogen-bonding interactions between ligand atoms and solvent molecules are discarded unless the solvent molecule also has a qualifying interaction with a protein residue. Solvent molecules which are not anchored in such a way are assumed to be transient and fluxional and are not explicitly included in the diagram.

Ligand/metal interactions are scored in a similar manner to hydrogen bonds, using an empirical type-based scoring function, trained by examining statistics from the RCSB Protein Data Bank. Ionic species which are not strongly ligated to a metal, and any other nonsolvent molecular species in proximity to the ligand, are included in the diagram but are not classified as having a specific strong interaction for purposes of the layout. Their interactions with the ligand or with other residues may however be annotated in some way, as is described subsequently.

Ligand Layout. In order to produce an aesthetically high-quality ligand layout which preserves 3D conformational information, we extend the 2D depiction layout method which we have described previously.[7] The 2D depiction layout algorithm operates by sampling and refining a set of locally ideal constraints in order to find the most disperse combination. We have enhanced the original implementation to allow the scoring function to be modified by the calling function. To the original score is added a supplementary offset which is comprised of

$$\sum_{ij} w_i w_j \frac{(d_{ij} - D_{ij})^2}{D_{ij}^2} + 10 \sum_{ijkl} |t_{ijkl} - T_{ijkl}|$$

where ij consists of all unique pairs of nonbonded heavy atoms, w_i is the weight of each atom (10 if involved in a hydrogen bond with a nonligand entity, 1 otherwise), d_{ij} is the distance between two atoms in the proposed flat layout, D_{ij} is the corresponding distance in the original geometry, $ijkl$ consists of all bonded torsion sets of four heavy atoms, t is the torsion angle in the flat layout, and T is the torsion angle in the original geometry, the latter two both in radians.

The distances are scaled such that the 3D bond distances are comparable to the default bond distances used by the 2D layout algorithm. Only nonbonded atom pairs are included in the term, and the values are weighted inversely according to the original 3D distance, in order to prevent extremely far atoms from dominating the equation. The overall offset is balanced such that it is sufficiently large in magnitude to override the intrinsic weak repulsion terms,

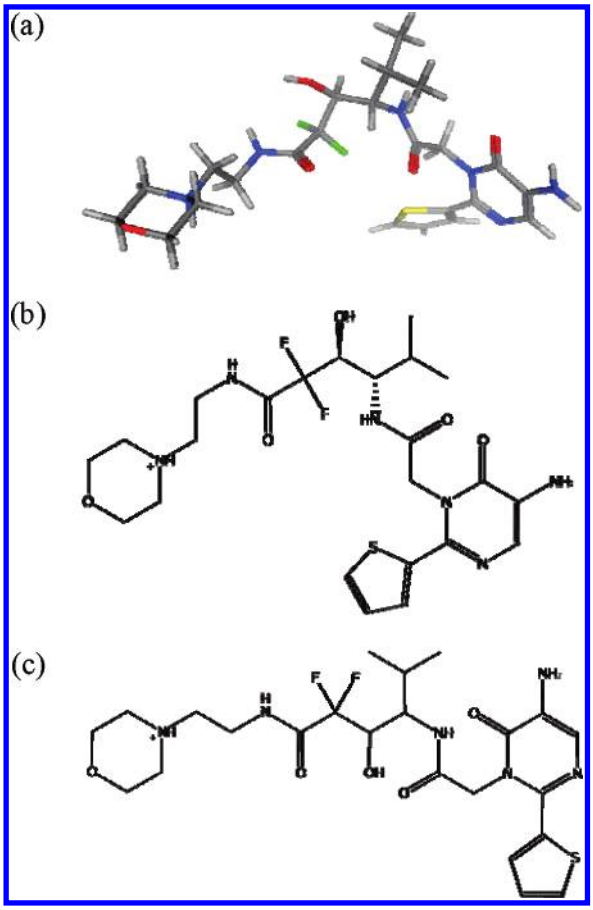


Figure 2. Production of a flattened version of a 3D ligand conformation: (a) 3D conformation of bound ligand (PDB code 1EAU); (b) layout produced by 2D depiction constrained by original 3D geometry; (c) layout produced by 2D depiction without additional constraints.

but not so large as to force the depiction layout algorithm to select aesthetically unacceptable combinations.

This method is a generally effective way to find the best compromise between the two competing goals. Figure 2a shows a ligand with a significant amount of flexibility, which adopts a slightly bunched-up conformation in its bound state. The constrained 2D layout, shown in Figure 2b, matches the original conformation to the extent possible given that no aesthetically displeasing layout choices were made. The output from the depiction layout process without any extra constraints is shown in Figure 2c, which maximizes the dispersity of the structure and is therefore far less suitable for representing the ligand in its bound environment.

Receptor Layout. The compromises involved in arranging the receptor residues around the ligand in 2D are typically much more severe than those required for the ligand, and so it is necessary to prioritize to a larger extent. The method which we describe treats each distinct receptor residue as a single point, which will be subsequently rendered in a cartoonish manner, rather than a molecular drawing. This approach allows a larger number of residues to be placed around the ligand and permits more freedom to convey relative proximity. Nonetheless, while most of the ligand−receptor distances can be qualitatively shown, many of the receptor−receptor distances can not, and there is no notion of "above" or "below" the ligand.

In order to accomplish these tradeoffs, the collection of residues, solvent molecules, ions, and cofactors requiring placement is divided up into two groups: those which have a qualifying hydrogen-bonding interaction with the ligand and those which do not. The former take priority and are placed first.

The placement method for each group proceeds in two steps: (1) reasonable best guess initial positioning and (2) refinement by continuous optimization. Initial positions are selected by dividing the space around the ligand into grid points, using a fairly coarse mesh (100×100 is sufficient). The grid points are each assigned an initial value of zero. For each of the atoms in the ligand, and for each ring center, a Gaussian function is added to the grid points, centered at the atom:

$$f(r) = \exp\left(-\frac{1}{2}r^2\right)$$

where r is the distance from the atom or ring center. The grid point with the largest value is defined as 1 and the remainder scaled accordingly. The result is a grid map which demarcates areas where residues should *not* be placed, as can be seen in Figure 3a, which shows the magnitude of the grid values placed around the rolipram ligand.[8]

For the placement of each of the individual residues, the initial grid map is overlaid with a further function for each hydrogen bond to the ligand:

$$f(r) = \frac{1}{r}$$

where r is the distance from the point of attachment. Each of these overlays is scaled so that the largest magnitude is equal to 1 divided by the total number of attachments, which introduces a balanced preference for placing the residue as close as possible to the interaction points. Figure 3b shows the initial placement position for Gln443, which has hydrogen-bonding interactions with two ether oxygen atoms, and Figure 3c shows the initial placement of His234, which has an interaction with the carbonyl oxygen. Figure 3d shows the ligand itself, as well as the final positions of both of the hydrogen-bonded residues.

The residues hydrogen-bonded to the ligand are initially placed separately without taking into account inter-residue interactions, and a reasonable set of starting points is obtained, albeit with overlap. A twice continuously differentiable fitness function is then composed, in order to take into account inter-residue repulsion, satisfaction of the original 3D inter-residue distances, and the actual distances between the ligand and the residue interaction points:

$$\text{energy} = \sum_{ij} w_{ij}(d_{ij}^2 - D_{ij}^2) \quad (1)$$

$$+ \sum_{ij} \exp\left(-\frac{1}{2}d_{ij}^2\right) \quad (2)$$

$$+ \sum_{ik} (d_{ik}^2 - D_{ik}^2) \quad (3)$$

$$+ \sum_{ik} \exp\left(-\frac{1}{2}d_{ik}^2\right) \quad (4)$$

$$+ \sum_{\theta} \exp\left[-\frac{5}{2}(1 - \cos\theta)^2\right] \quad (5)$$

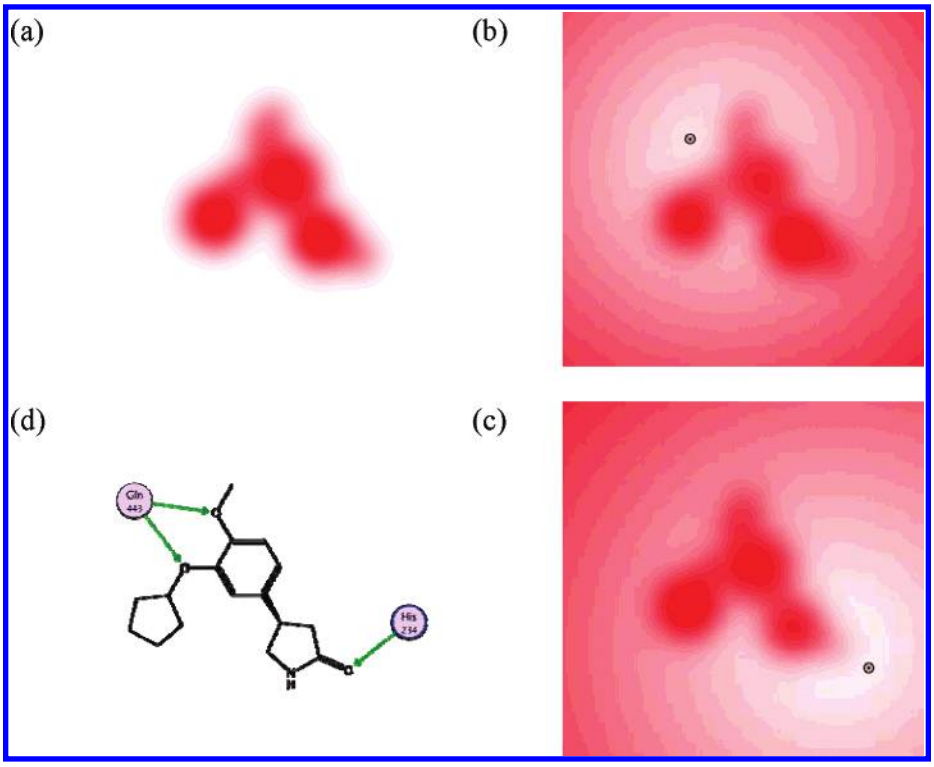


Figure 3. Initial placement of residues using grid-based method (PDB code 1RO6): (a) plot of repulsion from atoms and ring centers; (b) initial placement of Gln443; (c) initial placement of His234; (d) final placement of both hydrogen-bonded residues. In b and c, the locations where repulsion from the ligand and distance to the point of attachment is minimum are denoted with a dot.

where *ij* combinations refer to pairs of residues whose positions are the subject of the optimization and *ik* refers to a residue currently subject to optimization (*i*) with a fixed point such as a ligand atom (*k*); d_{ij} values refer to the distance between the two objects in the 2D diagram, while D_{ij} is the distance in the original 3D geometry, scaled so that the average heavy−heavy bond distance is 1.5 Å. Terms 1 and 3 encourage residue positions to conform to the distances in the 3D source, while terms 2 and 4 become significantly repulsive at short distances. For term 1, the value of w_{ij} is set to $0.05/(D_{ij} + 1)$, multiplied by 10 if the residues *i* and *j* share a covalent bond (peptidic or other, such as disulfide) or have a strong hydrogen-bonding interaction. The inverse distance term reflects the decreasing importance of positioning distant residues relative to one another. Term 3 lacks this weighting, because distances between receptor residues and the ligand atoms to which they are hydrogen-bonded are always highly important. Term 5 is an angular repulsion term which is used whenever two residues are anchored to the same ligand atom, forming the angle θ, which should not be small.

The energy function is optimized using the truncated Newton method.[9] Once the optimization is finished, the positions of the residues are considered to be fixed and inviolable.

The remaining residues, those not having strong hydrogen bonds to the ligand, are placed in successive waves, such that local proximity to other residues is honored to the extent possible within the constraints of an aesthetic diagram. An iterative selection process is used: if any of the remaining residues have strong bonding interactions (e.g., peptidic links, disulfide bonds, or significant hydrogen-bonding interactions) to residues which have already been placed, they are included in the next group. If there are none, all remaining residues are placed. The process is repeated until all residues have been placed.

These secondary residues use the same initial placement method as the original group, except that residues which have already been placed are treated as honorary ligand atoms; that is, interactions with the already-placed entities are mapped onto the preplacement grid, and they are treated as static points in the subsequent optimization. A similar refinement function is used:

$$\text{energy} = 0.01 \sum_{ij} (d_{ij}^2 - D_{ij}^2) \quad (1)$$

$$+ \sum_{ij} \exp\left(-\frac{1}{2} d_{ij}^2\right) \quad (2)$$

$$+ 0.001 \sum_{ik} (d_{ik}^2 - D_{ik}^2) \quad (3)$$

$$+ 5 \sum_{ik} \exp\left(-\frac{1}{2} d_{ik}^2\right) \quad (4)$$

the differences being that term 1 only includes pairs of residues which are connected (by a covalent or a hydrogen bond), term 3 is instead used as a weak tethering of each residue to its initial placement position, term 4 includes repulsion from already-placed residues, and there is no angular repulsion term.

Coordinated Metals. Metals which are observed to have one or more metal−ligand interactions with the ligand are placed using a different method. The metal−ligand interactions are temporarily converted into covalent bonds, and the metals are included as a part of the ligand during the 2D depiction layout process then subsequently disconnected.

Ligated metals are therefore placed before any of the other nonligand entities.

Rendering. Once the positions of the ligand atoms, protein residues, solvent molecules, cofactors, and loose ions are known, the primary features of the diagram can be drawn. The ligand is drawn in a conventional way, according to stylistic norms. If the ligand molecule contains explicit hydrogen atoms, then those which are not identified as hydrogen-bond donors are folded into the diagram (e.g., the −OH mnemonic), while those which are hydrogen-bond donors are drawn as explicit atoms (e.g., −O−H···residue).

Protein residues and solvent molecules are drawn as circles, with coloring dependent on the type. By default, protein ligands are coded according to the residue types: polar residues have a mauve interior, while hydrophobic residues are green. Basic residues are further annotated with a blue rim, and acidic residues with a red rim. Solvent molecules and cofactors are plain and colorless, and metal ions are drawn in plain text if coordinated to the ligand or in a gray disc if not.

Hydrogen bonds are drawn as dotted lines with arrows denoting the direction of the bond (double-headed for bidirectional interactions). The default color of the interaction line is green for interactions with the residue side chain, blue for interactions with the backbone, yellow for interactions with solvent ions, and purple for metal−ligand contacts or disconnected covalent linkages. In order to make sure ambiguity does not arise when diagrams are rendered in monochrome, interactions to residue backbones are distinguished from interactions with residue side chains by a solid dot on the residue end of the interaction line.

Solvent Exposure. Because the ability to convey 3D proximity in a 2D diagram by spatial arrangement is restricted, it is desirable to annotate the diagram entities with properties calculated from the original 3D structure. The degree of exposure to a solvent (assumed to be primarily water) is one such property.

Solvent exposure per atom is calculated by forming a sphere around each non-hydrogen atom of van der Waals radius +1.4 Å (the effective radius of a water molecule), and calculating the fraction of the surface area of each sphere which overlaps with that of none of the other atoms in the ensemble. The calculation is approximated by plotting individual points (122 per atom is sufficient) around the surface of the spheres and counting the proportion of points which approach within the spacing limits of those of other atoms.

Figure 4 shows a schematic example of two distinct molecular species which are in close contact. The inner gray bounding area shows the van der Waals radii, whereas the dotted bounding area, color-coded according to the atom type, shows the solvent-exposed surface area. The closest

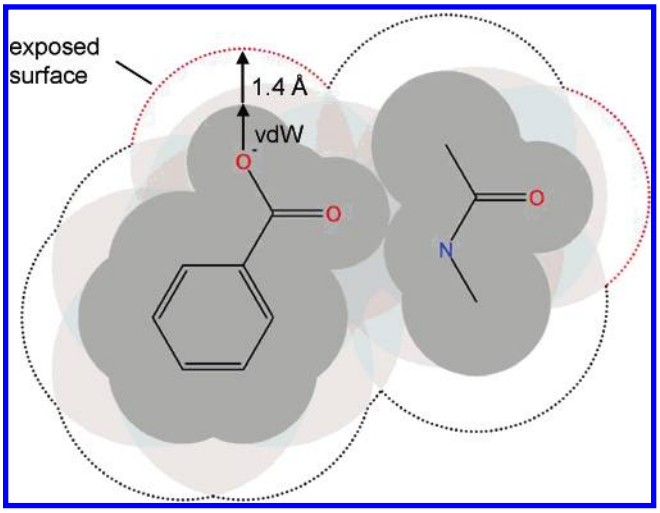
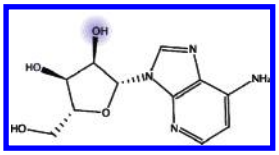


Figure 4. Schematic illustration of the solvent-exposed surface area calculation. The dark-gray shaded area is enclosed by the van der Waals radii of the constituent atoms, while the light-gray areas are enclosed by the van der Waals radii + 1.4 Å. The non-overlapping surface area is traced with a dotted line, which is color-coded according to the contributing atom.

point of contact between the amide and benzoate functional groups is shown to leave no solvent exposure between the two, whereas the anionic oxygen atom of the benzoate group has a significant amount of exposure to solvent, which can be seen by the red dots which remain on its outer surface.

Because the spheres plotted around each atom are large, atoms with two heavy neighbors are mostly covered and those with three almost completely. Any remaining exposure quickly dissipates as the more exposed parts of the ligand approach the interior boundary of the active site. The presence of any residual surface area implies that there is some room for a water molecule to reside between the ligand and the receptor. Applied to the ligand atoms, this suggests that there is empty space immediately nearby which may well be occupied by solvent, or that the region of the ligand projects outside of the active site. In the following example:

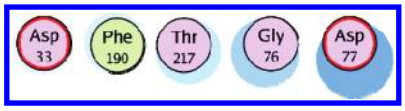


there is one hydroxyl functional group which has a very large amount of exposure to water in its local environment, and it is shown with a heavy blue smudge surrounding it, which is drawn with a radius and intensity proportional to its magnitude. The other atoms are mostly well-contained within the boundary of the active site and have little if any additional annotation.

While the absolute magnitude of solvent exposure of individual protein atoms is not especially interesting for the diagram, the difference in solvent exposure due to the presence of the ligand is. For each individual amino acid residue, the sum of the solvent exposure of all constituent heavy atoms without the ligand is calculated, followed by the same calculation with the ligand atoms included. The difference between the two sums is the extent to which the presence of the ligand has caused the residue to be shielded from exposure to solvent. If the ligand made no difference,

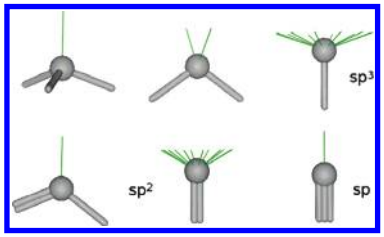
or the residue had little exposure to begin with, the value would be low. The reduction of solvent exposure is of particular interest for residues which have hydrophobic side chains, since reduced solvent exposure can indicate a favorable stabilizing interaction.

A reduction of solvent exposure induced by the ligand is drawn as a halolike disc around the residue:



In the series shown above, the arbitrarily placed residues are increasing in reduction of solvent-exposed surface area from left to right, as can be seen by the turquoise disc growing in size and intensity. The position of the surrounding disc is situated away from the ligand, so as not to obscure any other annotations.

Substitution Contour. Often an important reason for visually scrutinizing a protein−ligand complex is to perceive ways in which the ligand could be modified in order to better fit the binding pocket. By drawing a contour about the ligand, in which the distance from the atoms is representative of the available space, an easily interpretable indication can be given as to where empty regions exist. This property is similar in some respects to solvent exposure but is calculated by measuring along vectors of potential substitution points on each ligand atom, using the original 3D coordinates, depending on type:



The angles are chosen on the basis of geometric positions which would be suitable for a hydrogen atom. A "cannonball" of radius 0.8 Å is fired down each of the vectors, and the distance which it is able to travel before touching the van der Waals surface of any other non-hydrogen atom in the system is recorded. The distances for those atom types with multiple vectors are taken as the average. The minimum distance is 0, and the maximum distance is taken to be 4 Å. If more than half of the vectors for an atom exceed the maximum distance, the atom is also marked as having "unlimited" room in which to expand.

In this way, each heavy atom of the ligand is ascribed a numerical value. A grid is prepared (approximately 100 × 100), and the magnitudes of each of the atoms are additively overlaid onto the grid according to the function:

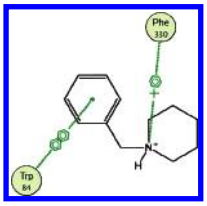
$$f(r) = \begin{cases} 1 & :r < D - 1 \\ \frac{1}{2}\left\{1 + \cos\left[\frac{1}{2}\pi(D - r)\right]\right\} & :D - 1 \le r \le D \\ 0 & :r > D \end{cases}$$

where D is the trajectory extent for the heavy atom being considered. The contour is formed by tracing a line about the exterior at a value = 0.5. Regions of the outline which are most close to an atom which has been marked as "unlimited" are colored invisibly, and hence the line is broken to denote exposure to the outside of the pocket, or to a large empty region.

The function is scaled such that, if the distance from a ligand atom to the contour is approximately one unit bond length, then there is roughly enough room for the substitution of one extra heavy atom (e.g., methyl, hydroxyl, etc.) within the context of the binding mode of the ligand. The contour also provides an overall indication of how tightly the ligand is contained within the active site.

$\pi-\pi$ and $\pi-$Cation Interactions. A significant number of protein−ligand complexes exist where at least one part of the binding region is oriented such that aromatic rings of the ligand and receptor stack in parallel, or a cation is conspicuously located directly above an aromatic ring center. Rigorous prediction of the magnitude of such interactions has not been attempted in this work; rather, a simple geometry-based estimate has been devised, which is sufficient for calling to attention the presence of $\pi-\pi$ and $\pi-$cation interactions. Arenes are denoted as having a Gaussian sphere on either side of the ring, defined by the function $f(r) = 0.78 \exp(-|r - V|^2)$, where r is a point in space, and two values of V exist for each aromatic ring, which are computed as the ring center, extended 1 Å along the normal to the ring, in either direction. Cations are denoted as having a single diffuse Gaussian sphere centered at the atom which bears the charge, and defined by the function $f(r) = 7.8 \exp(-0.05|r - P|^2)$ where P is the position of the atom. The strength of the interaction between two arenes, or between an arene and a cation, is taken to be the integral of the respective functions over space. The scaling of the arene function is chosen such that the integral of the functions belonging to two benzene atoms directly facing each other at a distance of 3.8 Å would evaluate to 1. The function for cations has been scaled empirically, in order to reproduce literature reports of such interactions.[10]

Visual annotation of $\pi-\pi$ interactions is shown by a dotted green line from the residue to the center of the aromatic ring, with an insignia in the middle which shows the sizes of the rings involved in the interaction. $\pi-$cation interactions are drawn in a similar way, both of which are illustrated below:



Concurrent Ligand Alignment. Preparing a series of diagrams with a layout that shows consistent geometrical trends is accomplished in two distinct steps, the first of which is to obtain a realistic coalignment of the ligands on the plane. It is fortuitous that druglike molecules have a tendency to adopt binding conformations that can be approximately transformed to a 2D manifold and also that such a manifold typically applies similarly well to all of the ligand binding conformations in a series. Representation of a series of complexes is predicated on an already extant alignment in 3D, which can be accomplished by superposition based on protein residue identity.

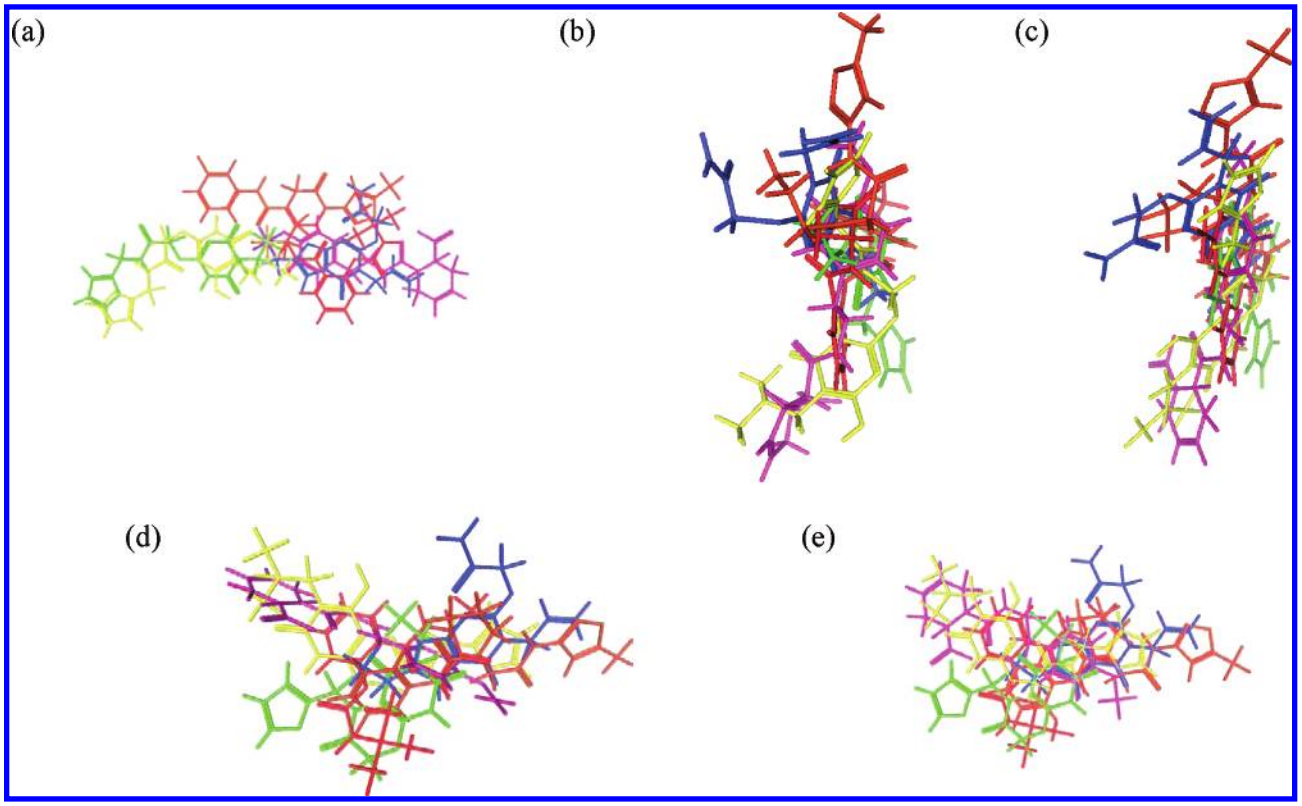


Figure 5. Coalignment of flattened ligands: (a) flattened ligands with no common orientation (F_{2D}); (b) ligand ensemble in original 3D space (F_{3D}); (c) flattened ligands superimposed onto original 3D positions; (d) rotation of the ensemble in c to get positions P_{2D}, close to the plane of the paper; (e) tidyup superimposition of P_{2D} to P'_{2D} to achieve final positions L_{2D}.

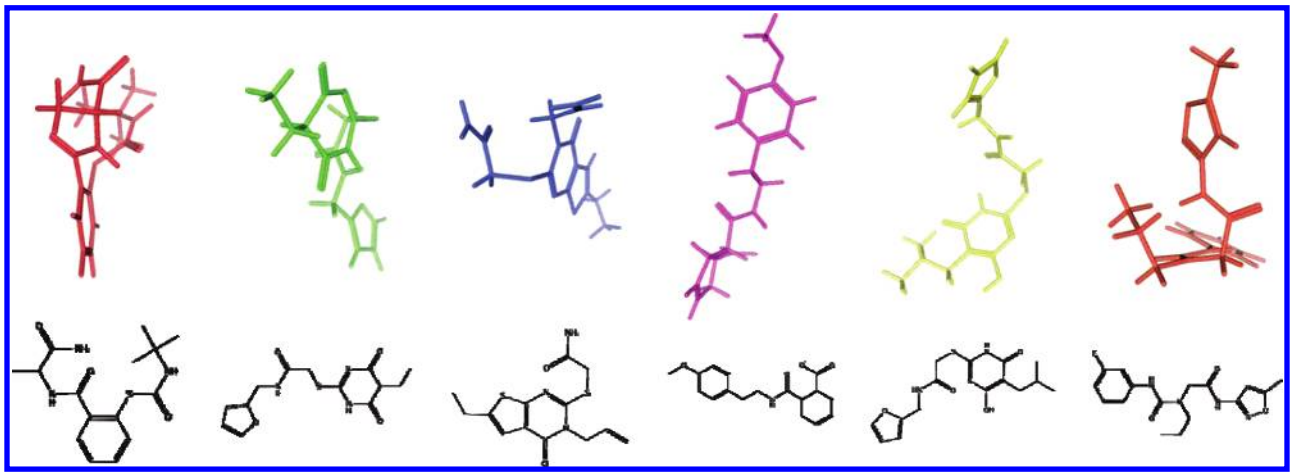


Figure 6. Original 3D positions of each ligand (above), and final aligned 2D layout (below).

Figure 5 illustrates the process of transforming a collection of flattened structures onto the page in a way that reflects the original poses. First, the flattened structures (F_{2D}, shown in Figure 5a) are individually superposed into their original positions (F_{3D}, shown in Figure 5b) to get the positions shown in Figure 5c. For this ensemble, a rotation matrix is found which best aligns these positions collectively onto the flat plane using principle moments, and this is applied to all of the ligands to get a set of points P_{2D}, shown in Figure 5d. At this point, the ligand coordinates P_{2D} are typically almost aligned on the plane of the paper but, depending on the variability of the poses in the original binding mode, will be misaligned to some extent. The final aesthetic tidyup is accomplished by defining P'_{2D} as P_{2D}, in which the *z* axis has been set to zero for all points, and superposing each ligand from P_{2D} to P'_{2D}, to get the final ligand positions, L_{2D}, which are shown in Figure 5e.

This method is effective as long as the 3D conformations of the ligands are not completely antagonistic to a planar layout, and their binding modes are not highly orthogonal to one another. In practice, this is a very effective way to show the relative binding modes of ligands.

Figure 6 shows the same ligands undergoing the same alignment as is shown in Figure 5, with the original and final positions indicated separately for clarity.

Concurrent Receptor Layout. The second part of concurrent series layout is to force the residue positions to be identical in each case, which is accomplished by performing a single residue layout step which applies simultaneously to all of the diagrams.

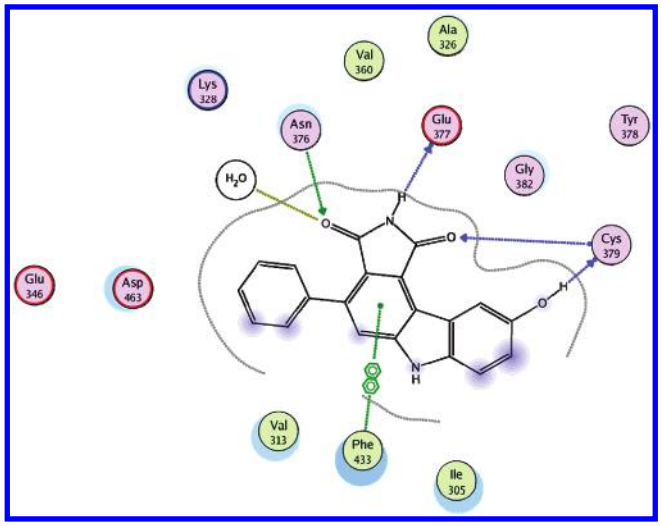
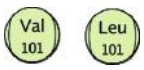


Figure 7. Wee1 inhibitor (PDB code 1X8B), showing hydrogen bonds to side chains of Asn376, to the peptide backbone of Glu377 and Cys379, and to one water molecule. An arene interaction with Phe433 is also shown.

To facilitate this, the collection of ligands is merged into a single composite entity by overlaying a grid, of spacing 0.5 Å, over the aligned ligands. Each atom is classified as belonging to the closest grid point. Grid points which own at least one atom from a contributed ligand are considered pseudo-atoms for the purpose of initial placement of the residues. Hydrogen-bonding interactions between residues and constituent ligands are reassigned to the corresponding grid point instead of the atom and weighted proportionately to their frequency of occurrence throughout the series.

The layout procedure for the residues, solvent molecules, cofactors, and ions is done in the same way as for the nonseries algorithm, except using the pseudo-atom grid points instead of the ligands.

Series Annotation. A series of protein−ligand complexes will typically have a different set of hydrogen-bonding interactions between the ligand and the protein residues, and in some cases, the residues at certain positions may differ. On each diagram, a hydrogen bond which is not consistently present throughout the series is annotated by a bar behind the arrowhead. A hydrogen bond which is present in other cases but not the current one is indicated by a broken interaction from the residue to halfway to the average position of those ligand atoms which comprised the other end of the interaction. Residues which are present in the current diagram, but are missing or of a different type in other entries in the series, are drawn with a curved sidebar on either side:



In this way, hydrogen bonds and residue identities which vary across the series are brought to attention without overly detracting from the features of the diagram.

RESULTS AND DISCUSSION

A number of case studies have been selected from recent literature reports and used to highlight particular capabilities of the diagrams which can be produced using the method we describe. In each case, the source data is available in the Protein Data Bank, and the corresponding codes are provided.

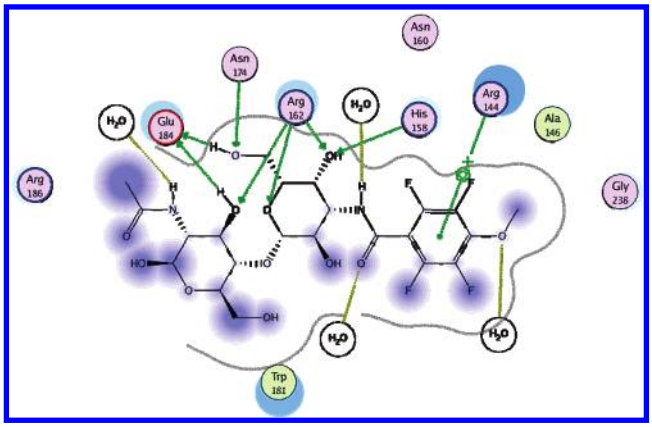


Figure 8. Galectin-3C inhibitor (PDB code 1KJR), showing hydrogen bonds to four protein residues and four water molecules, as well as an arene−cation interaction with Arg144. A large amount of solvent exposure is shown on one side of the ligand, which is bound in a wedgelike cavity.

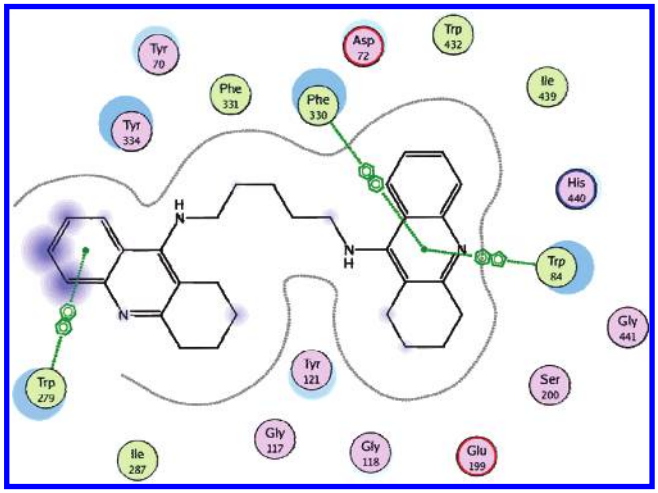


Figure 9. AChE inhibitor (PDB code 2CMF), showing three distinct arene-stacking interactions and no significant hydrogen-bonding interactions. Most of the ligand is deeply buried and shows a tight proximity contour and negligible solvent exposure, except for the exposed corner.

A relatively straightforward example is shown in Figure 7. The Wee1 inhibitor is small and largely planar, which makes it feasible to lay out the nearby residues in a manner that is qualitatively representative of the actual 3D environment. The ligand is anchored into the active site by five hydrogen bonds, including a donor and an acceptor interaction with the backbone of Cys379, a donor interaction with the Glu377 backbone, and an acceptor interaction with Asn376. One of the carboxyl groups has a hydrogen bond to a crystallographically ordered water molecule, and the central aromatic ring is positioned for a possible aromatic stacking interaction with Phe433. The interactions shown on the diagram correspond to those described in the literature.[11]

Figure 8 shows a carbohydrate-based Galectin-3C inhibitor, which presents a more challenging layout problem due to the nonplanarity of the ligand, and the relatively high density of perceived hydrogen bonds. The layout of the ligand reaches an agreeable compromise between the original 3D geometry, which is qualitatively preserved, and a legible

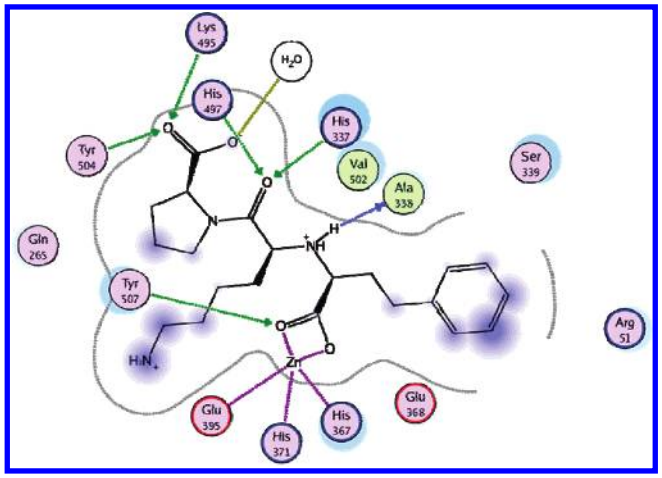


Figure 10. Angiotensin inhibitor (PDB code 1J36), showing a variety of hydrogen-bonding interactions, as well as a 5-coordinated zinc ion, which binds both oxygen atoms of a carboxylate group of the ligand, as well as Glu395, His371, and His367.

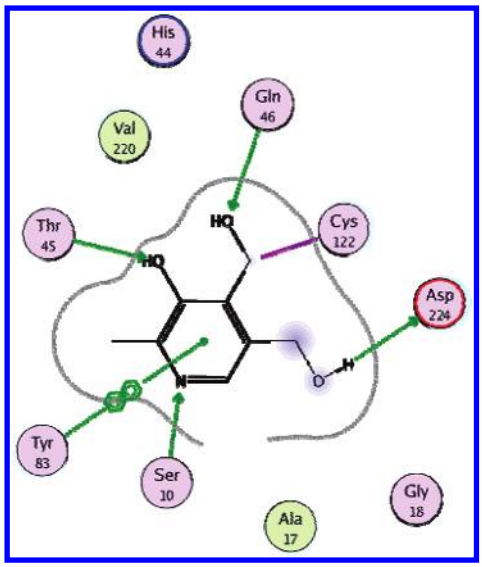


Figure 11. Covalent pyridoxal cofactor bound to PdxY (PDB code 1TD2), which shows the covalent bond to Cys122.

molecular sketch. The Glu184, Arg162, Asn174, and His158 residues present a convoluted set of side-chain interactions to the saccharide units, particularly that which is located in the center of the ligand. There are four ordered water molecules, included in the diagram because they are anchored to protein residues as well as to the ligand, further complicating the residue layout, because two of them interact with the ligand in the most busy region. It can be seen from the diagram that the aromatic end of the ligand possesses an interaction between the arene ring and the positively charged side chain of Arg144. It is also interesting to note that the diagram shows most of the receptor residues concentrated on one side of the ligand, which is an accurate impression, since the binding site of the monomeric protein is a shallow groove which leaves most of the ligand exposed to solvent.[12]

An example of a ligand/receptor complex where nonpolar interactions are the dominant method of stabilization is shown in Figure 9. The symmetrical bis-tacrine ligand is a strong binder of *Torpedo californica* acetylcholinesterase, yet it exhibits no contacts which might obviously be interpreted as hydrogen bonds.[13] Apart from an aromatic ring at one

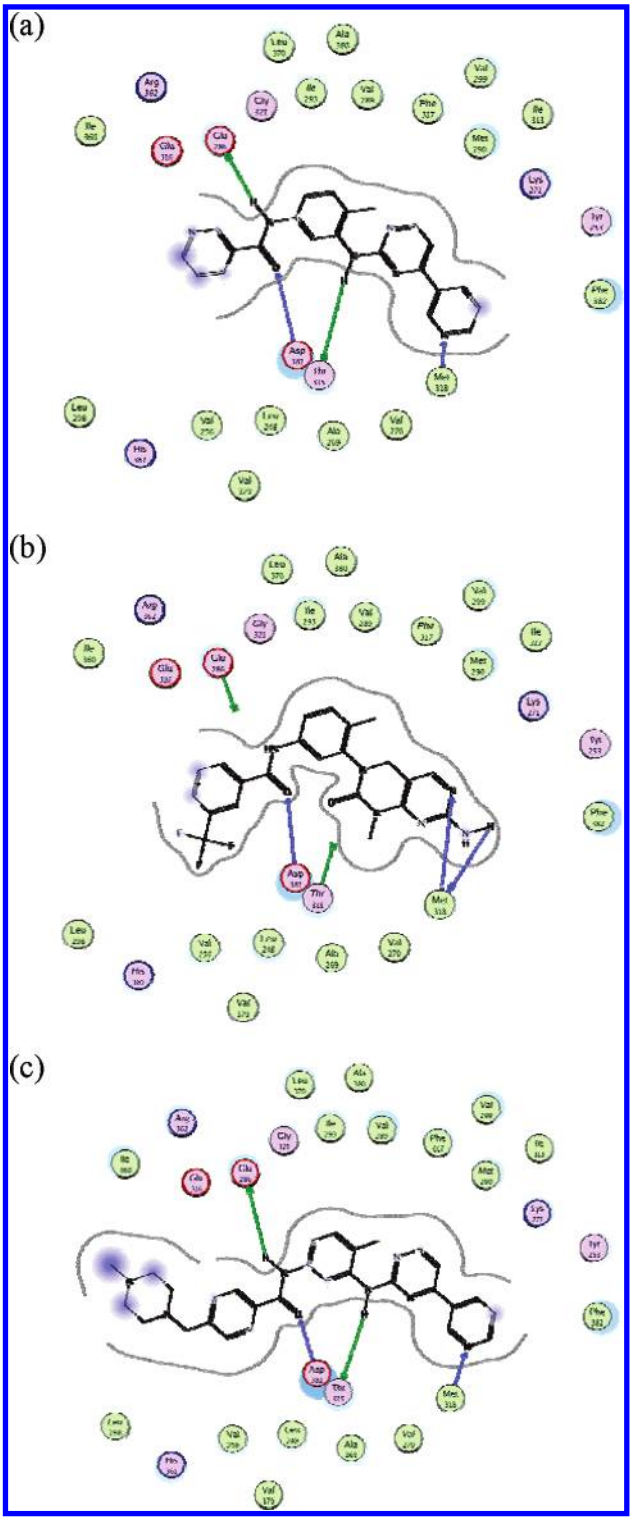


Figure 12. Abl inhibitors: (a) PDB code 1FPU; (b) PDB code 2HIW; (c) PDB code 1IEP. The three complexes are aligned as a series, in which complex a shows a significant amount of open space about the 3-pyridyl substituent shown on the left. In complexes b and c, ligands with a similar binding motif exploit the available space in the meta and para positions, respectively.

end of the ligand which projects out slightly from the active site, the ligand forms a snug fit within the cavity, which is shown by the continuity of the substitution contour line and negligible solvent exposure of the ligand atoms. The interior heteroaromatic ring is annotated with $\pi-\pi$ stacking interactions with both Phe330 and Trp84. The annotation also shows a high degree of reduction of solvent-exposed surface area

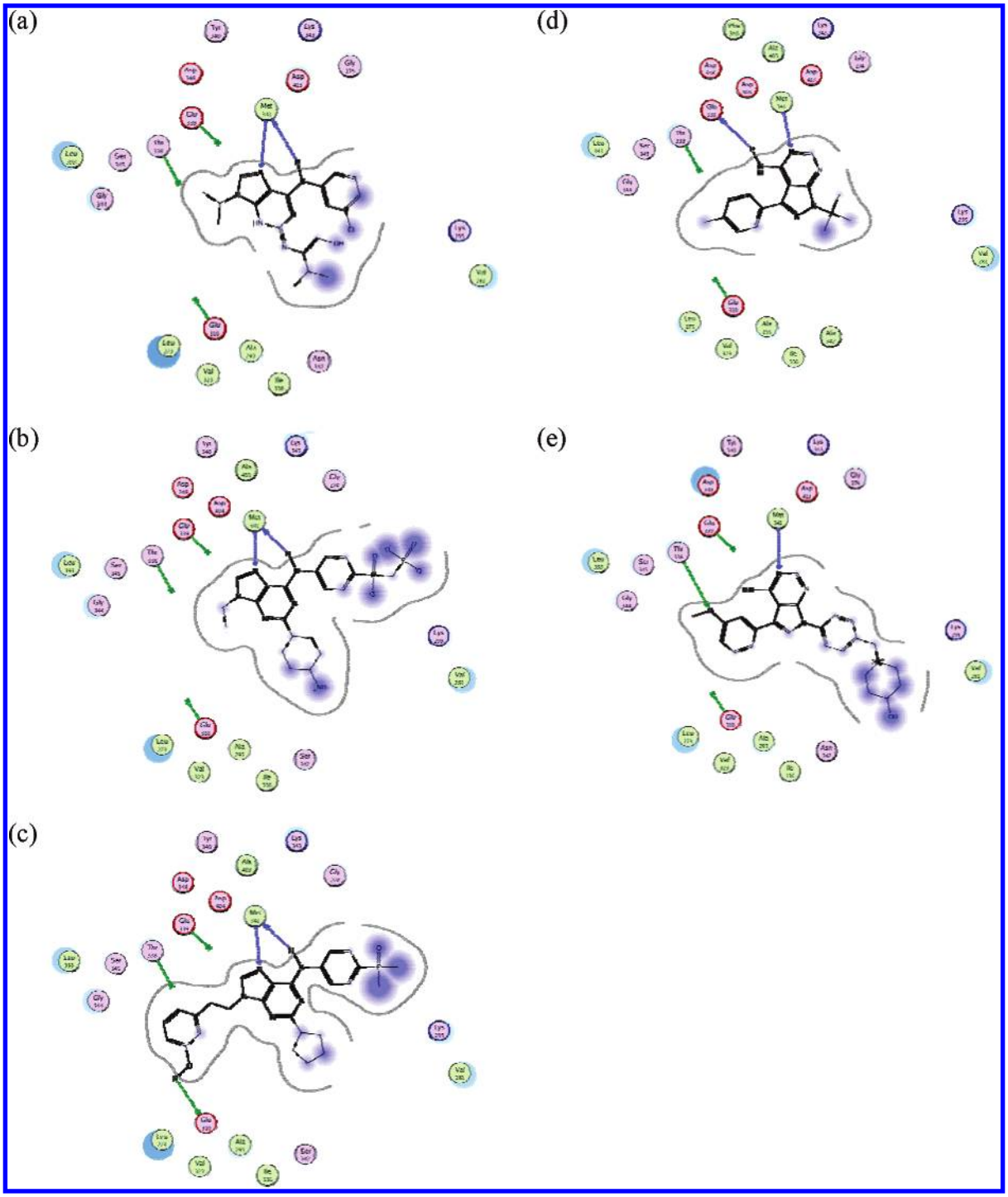


Figure 13. Inhibitors of Src based on purine-core: (a) PDB code 1YOM; (b) PDB code 2BDF; (c) PDB code 2BDJ. Inhibitors of Src based on non-purine core: (d) PDB code YOL; (e) PDB code 1QCF. The substitutued purines a, b, and c show a common binding motif involving hydrogen-donor and -acceptor bonds to Met341, which are invariant to modification of the substituents. Switching to the non-purine core alters the binding mode, as can be seen in d and e.

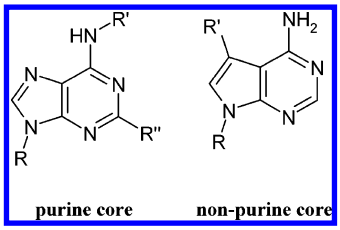
on these two hydrophobic residues. The side of the ligand which is not fully enclosed within the active site features a single aromatic interaction with Trp279.

Metal ligation is an important feature in many active sites, such as for the *Drosophila* angiotensin-I inhibitor lisinopril, which is shown in Figure 10. As can be seen in the diagram, the carboxylate group situated near the center of the ligand is chelated to a zinc ion, which in turn is also held in place by close contacts with Glu395, His371, and His367. The ligand displaces two water molecules which previously made up the coordination sphere of the metal ion.[14]

Schematic diagrams are applicable also to ligands which are covalently bound to the receptor. Figure 11 shows one of two binding environments for the dimeric protein structure of *Escherichia coli* PdxY crystallized with two pyridoxal ligands. The diagram shown is the peptide in which the ligand is postulated to be bound by a thiohemiacetal linkage with Cys122.[15] The divider between protein and ligand is shown with a dotted purple line, while the diagram is otherwise produced in the usual way. Four significant hydrogen-bonding interactions are detected, in addition to a $\pi-\pi$ stacking effect with Tyr83.

The utility of the substitution proximity feature is demonstrated by the series of Abl kinase inhibitors shown in Figure 12. Each of these three inhibitors occupy the central activation loop of the target protein. All of these ligands form significant hydrogen bonds between the common amide carboxyl group and the backbone of Asp181 in the central region, and a backbone interaction with Met318 at the deep end of the pocket. One of the most striking features of Figure 12a is that the 3-amidopyridyl terminus is jutting out into a large unoccupied region of space, which is clearly indicated by the broken contour line and the high degree of solvent exposure at the 5 and 6 positions on the pyridine ring.[16] Figure 12b shows a related structure with a similar binding mode which introduces a trifluoromethyl substituent meta to the amide functional group in order to fill a small hydrophobic binding pocket specific to the type-II protein conformation.[17] Figure 12c shows the structure of STI-571, also known as Gleevec, which is almost identical to "a" in both its composition and binding mode, except with the addition of a methylpiperazine moiety at the para position of the terminal aromatic ring, which fills a largely hydrophobic groove.[18]

The concurrent alignment of ligands in a related series is an effective way to spot certain trends in binding environments. Figure 13 shows a total of five active sites, all of which feature inhibitors of the Src target,[19] and all of which are based on one of two similar cores:



It can be seen in the diagrams for Figure 13a−c that the duo of hydrogen bonds between the purine core and amine substituent to Met341 is a key feature which anchors all three of these ligands in place. All of these ligands have significantly different attachments to the base core, some of them bulky and capable of forming a variety of interactions with the protein, yet the core binding mode remains consistent. The diagrams shown in Figure 13d and e are based on a similar ligand framework, yet the key binding motif is different, and so the ligands adopt completely different binding modes. For this system, the information which can be gained by studying the available structural data is likely to lead to significant insight into drug design efforts. The presentation of the series in its aligned schematic form makes this insight both immediate and obvious.

Comparison between series of complexes only requires that sufficient sequence identity exists in order to align the peptide backbone onto a common space. The diagrams in Figure 14 show three CDK inhibitor complexes. Figure 14a is a complex with CDK2, while Figure 14b and c are complexes with CDK6. The common alignment clearly shows that the core structural feature is a bidentate hydrogen bond with the backbone of residue 101, which is Leu in CDK2 and Val in CDK6. In each of the diagrams, residue 101 is drawn with two curved sidebars, to bring attention to the fact that it is not conserved throughout the series. Note

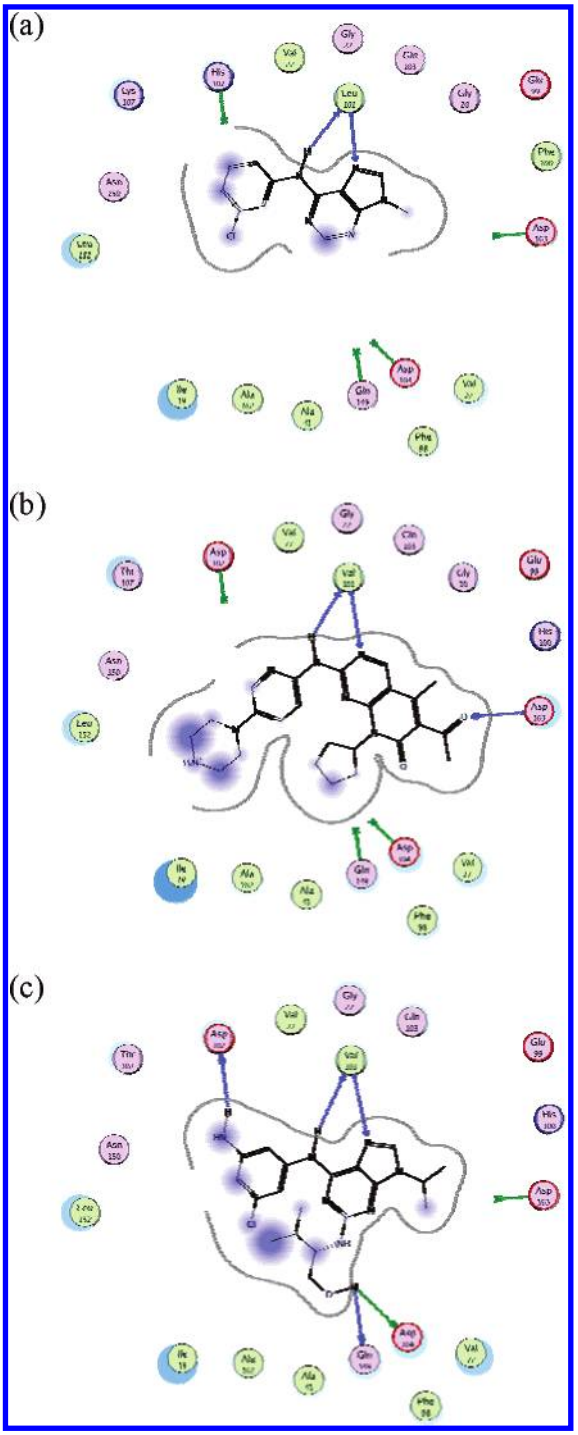


Figure 14. Inhibitors of CDK: Part a is a complex with CDK2 (PDB code 1CKP), while b and c are complexes with CDK6 (PDB codes 2EUF and 2F2C, respectively). The diagrams show a series of ligands bound to three different proteins, with high conserved residue identity. The nonconserved residues in the active site region are 100, 101, 102, and 107, which are annotated with sidebars. The interactions with Val101/Leu101 are consistent throughout the series and are drawn normally, while other interactions are not consistent and are annotated with a bar behind the arrowhead, or a broken interaction.

however that the hydrogen bonds are marked as being consistent throughout the series, because this is a conserved feature despite the change in residue type, which happens to be fairly unimportant. The other pertinent differences between these complexes are revealed by the various hydrogen-bonding interactions which are not preserved

throughout all three systems. Missing hydrogen bonds are shown with a dotted line leading in the general direction of where the ligand atom might be situated, as determined from the other members of the series, but terminated with an "X". The ligand in Figure 14b, PD0332991, is reported to be highly specific for CDK6, which is possibly due to the piperazinylpyridine substituent oriented outward of the binding pocket in the so-called "specificity region", in addition to a hydrogen bond with Asp163. The ligand in Figure 14c, aminopurvanol, does not provide such selectivity, and apart from the common core binding motif, it otherwise shows quite a different bonding environment, with hydrogen bonds to Asp102, Asp104, and Gln149, which are not observed in the other two cases.[20]

CONCLUSION

We have presented a method for the generation of protein−ligand interaction diagrams, which features robust layout algorithms for the ligand and the surrounding residues, as well as techniques for summarizing the important geometric features. The combination of a finely tuned layout methodology and mnemonic representation of active site features, which parallel the information which can be gleaned by studious examination using 3D software, leads to diagrams which are visually sparse and easy to behold, yet surprisingly information-rich on closer examination. Important features can be shown using minimalistic visual cues, highlighting the features of an active site using flat static media, which can be easily shared. Furthermore, the generation method is automatic, requiring only routine preparation of the protein−ligand complex, and optional modification of default parameters, such as hydrogen-bond thresholds or various annotation options. The ability to conserve relative alignments throughout a series of complexes, and diagrammatically highlight the differences, imparts significant perception advantages. We anticipate that renewed interest in automated generation of 2D diagrams will significantly reduce the burden of communicating structural results, and make insights into structure-based drug design available to a wider audience.

REFERENCES AND NOTES

(1) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **1995**, *8*, 127−134.

(2) Stierand, K.; Maa*β*, P. C.; Rarey, M. Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics* **2006**, *22*, 1710−1716.

(3) Sung, B.-J.; Hwang, K. Y.; Jeon, Y. H.; Lee, J. I.; Heo, Y.-S.; Kim, J. H.; Moon, J.; Yoon, J. M.; Hyun, Y.-L.; Kim, E.; Eum, S. J.; Park, S.-Y.; Lee, J.-O.; Lee, T. G.; Ro, S.; Cho, J. M. Structure of the catalytic domain of human phosphodiesterase 5 with bound drug molecules. *Nature* **2003**, *425*, 98−102.

(4) RCSB Protein Data Bank; http://www.rcsb.org (accessed Feb 15, 2007). All protein-ligand structures described can be obtained free of charge from this site.

(5) The source code for the algorithms described within, and those referred to, is packaged as part of MOE and may be examined and used under the terms of the MOE user license, which is available from the Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montréal, Québec, Canada. http://www.chemcomp.com (accessed Feb 15, 2007).

(6) (a) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimamouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542. (b) Bourne, P. E.; Berman, N. M.; McMahon, B.; Watenpaugh, K. D.; Westbrook, J.; Fitzgerald, R. M. D. The Macromolecular Crystallographic Information File (mmCIF). *Methods Enzymol.* **1997**, *277*, 571−590.

(7) Clark, A. M.; Labute, P.; Santavy, M. 2D Structure Depiction. *J. Chem. Inf. Model.* **2006**, *46*, 1107−1123.

(8) Xu, R. X.; Rocque, W. J.; Lambert, M. H.; Vanderwall, D. E.; Nolte, R. T. Crystal Structures of the Catalytic Domain of Phosphodiesterase 4B Complexed with AMP, 8-Br-AMP, and Rolipram. *J. Mol. Biol.* **2004**, *337*, 355−365.

(9) Gill, P.; Murray, W.; Wright, M. K. *Practical Optimization*; Academic Press: New York, 1981.

(10) Meyer, E. A.; Castellano, R. K.; Diederich, F. Interactions with Aromatic Rings in Chemical and Biological Recognition. *Angew. Chem., Int. Ed.* **2003**, *42*, 1210−1248.

(11) Palmer, B. D.; Thompson, A. M.; Booth, R. J.; Dobrusin, E. M.; Kraker, A. J.; Lee, H. H.; Lunney, E. A.; Mitchell, L. H.; Ortwine, D. F.; Smaill, J. B.; Swan, L. M.; Denny, W. A. 4-Phenylpyrrolo[3,4-*c*]-carbazole-1,3(2H,6H)-dione Inhibitors of the Checkpoint Kinase Wee1. Structure-Activity Relationships for Chromophore Modification and Phenyl Ring Substitution. *J. Med. Chem.* **2006**, *49*, 4896−4911.

(12) Sömme, P.; Arnoux, P.; Kahl-Knutsson, B.; Leffler, H.; Rini, J. M.; Nilsson, U. J. Structural and Thermodynamic Studies on Carion-P Interactions in Lectin-Ligand Complexes: High-Affinity Galectin-3 Inhibitors through Fine-Tuning of an Arginine-Arene Interaction. *J. Am. Chem. Soc.* **2005**, *127*, 1737−1743.

(13) Rydberg, E. H.; Brumshtein, B.; Greenblatt, H. M.; Wong, D. M.; Shaya, D.; Williams, L. D.; Carlier, P. R.; Yuan-Ping, P.; Silman, I.; Sussman, J. L. Complexes of Alkylene-Linked Tacrine Dimers with *Torpedo californica* Acetylcholinesterase: Binding of Bis(5)-tacrine Produces a Dramatic Rearrangement in the Active-Site Gorge. *J. Med. Chem.* **2006**, *49*, 5491−5500.

(14) Kim H. M.; Shin, D. R.; Yoo, O. J.; Lee, H.; Lee, J.-O. Crystal structure of *Drosophila* angiotensin I-converting enzyme bound to captopril and lisinopril. *FEBS Lett.* **2003**, *538*, 65−70.

(15) Safo, M. K.; Musayev, F. N.; Hunt, S.; di Salvo, M. L.; Scarsdale, N.; Schirch, V. Crystal Structure of the PdxY Protein from *Escherichia coli*. *J. Bacteriol.* **2004**, *186*, 8074−8082.

(16) Schindler, T.; Bornmann, W.; Pellicena, P.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Structural Mechanism for STI-571 Inhibition of Abelson Tyrosine Kinase. *Science* **2000**, *289*, 1938−1942.

(17) Okram, B.; Nagle, A.; Adrian, F. J.; Lee, C.; Ren, P.; Wang, X.; Sim, T.; Xie, Y.; Wang, X.; Xia, G.; Spraggon, G.; Warmuth, M.; Liu, Y.; Gray, N. S. A General Strategy for Creating "Inactive-Conformation" Abl Inhibitors. *Chem. Biol.* **2006**, *13*, 779−786.

(18) Nagar, B.; Bornmann, W.; Pellicena, P.; Schindler, T.; Veach, D. R.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res.* **2002**, *62*, 4236−4243.

(19) (a) Breitenlechner, C. B.; Kairies, N. A.; Honold, K.; Scheiblich, S.; Koll, H.; Greiter, E.; Koch, S.; Schäfer, W.; Huber, R.; Engh, R. A. Crystal Structures of Active Src Kinase Domain Complexes. *J. Mol. Biol.* **2005**, *353*, 222−231. (b) Schindler, T.; Sicheri, F.; Pico, A.; Gazit, A.; Levitzki, A.; Kuriyan, J. Crystal structure of Hck in Complex with a Src Family-Selective Tyrosine Kinase Inhibitor. *Mol. Cell.* **1999**, *3*, 639−648. (c) Dalgarno, D.; Stehle, T.; Narula, S.; Schelling, P.; van Schravendijk, M. R.; Adams, S.; Andrade, L.; Keats, J.; Ram, M.; Jin, L.; Grossman, T.; MacNeil, I.; Metcalf, C.; Shakespeare, W.; Wang, Y.; Keenan, T.; Sundaramoorthi, R.; Bohacek, R.; Weigele, M.; Sawyer, T. Structural Basis of Src Tyrosine Kinase Inhibition with a New Class of Potent and Selective Trisubstituted Purine-based Compounds. *Chem. Biol. Drug Des.* **2006**, *67*, 46−57.

(20) Lu, H.; Schulze-Gahmen, U. Toward Understanding the Structural Basis of Cyclin-Dependent Kinase 6 Specific Inhibition. *J. Med. Chem.* **2006**, *49*, 3826−3831.

CI7001473