JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

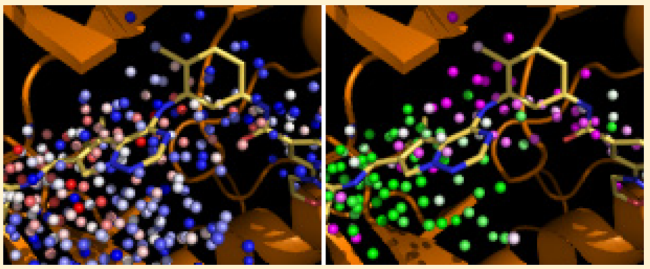
pubs.acs.org/jcim

3D Matched Pairs: Integrating Ligand- and Structure-Based Knowledge for Ligand Design and Receptor Annotation

Shana L. Posy,[†] Brian L. Claus,[†] Matt E. Pokross,[‡] and Stephen R. Johnson*[,†]

[†]Computer-Assisted Drug Design and [‡]Protein Science and Structure, Molecular Discovery Technologies, Bristol-Myers Squibb Research and Development, Princeton, New Jersey 08543, United States

ABSTRACT: We describe an extension to the matched molecular pairs approach that merges pairwise activity differences with three-dimensional contextual information derived from X-ray crystal structures and binding pose predictions. The incorporation of 3D binding poses allows the direct comparison of structural changes to diverse chemotypes in particular binding pockets, facilitating the transfer of SAR from one series to another. Integrating matched pair data with the receptor structure can also highlight activity patterns within the binding site—for example, "hot spot" regions can be visualized where changes in the ligand structure are more likely to impact activity. The method is illustrated using P38α structural and activity data to generate novel hybrid ligands, identify SAR transfer networks, and annotate the receptor binding site.



INTRODUCTION

Matched molecular pairs (MMPs) are pairs of structures that share a common core and differ in structure by only a single R group.[1−4] Analysis of the changes in compound activities that result from substituting one R group for another has identified structural factors that impact properties such as aqueous solubility,[5,6] HERG inhibition,[7] plasma protein binding,[6] and compound promiscuity.[8] In addition to describing structure−activity relationships (SAR), MMPs can also be used prospectively to design novel compounds, by transferring the SAR derived from one chemical series to a different scaffold.[9−12] However, most matched pair methods rely on two-dimensional compound structures and do not consider the three-dimensional context of ligand−receptor interactions.

In this work, we introduce a three-dimensional matched pair (3DMP) method that accounts for the orientation of ligands in a receptor binding site. The 3D approach has three advantages compared to standard matched pairs. First, if two different series contain the same substructure, 2D methods may conclude that the SAR can be transferred between the two series.[9,12] If, however, the relevant fragment projects to two different regions of the binding pocket in the two series, then incorporating structural context will prevent the inappropriate transfer of SAR between the series. Second, placing the structural fragments and related activity data in the binding site allows the matched pair SAR to inform the design of novel "hybrid" ligands that can plausibly bind to the receptor. Finally, mapping the matched pairs to the receptor structure provides a method for synthesizing and visualizing activity patterns in the binding site. For example, 3DMPs can identify "hot spot" regions where changes in the ligand structure are highly likely to impact activity.

In this study, we select kinases as an example target class for which many crystal structures are available and where interactions between ligands and the receptor binding site are well characterized. Since the 3D matched pair approach requires high confidence binding models of compounds that have not been crystallized, we first demonstrate that kinase structures provide a pathway for building reliable models. We then use models of P38α ligands to generate a 3D matched pair database and show how the database can be applied to build novel ligands, predict the activities of new analogs, and annotate the receptor binding pocket.

METHODS

CHEMBL Data Set. To illustrate the 3DMP workflow, human P38α was selected as a representative kinase target for which ample SAR and crystal structure data are publicly available. P38α activity data were downloaded from the CHEMBL database (target identifier CHEMBL260), and data points with no associated molecular structure were excluded. The data set consists of 4957 IC50's and 1117 K_i's measured for 4291 unique compounds in 385 experiments; 1427 of the compounds were tested in multiple assays.[13] For compounds that were inactive at the highest concentration tested and thus have null IC50s or K_i's, the IC50 or K_i was set to 100 000 nM for subsequent calculations that require numeric values. We refer throughout this work to IC50's alone for clarity, but all such references should be understood to mean "IC50 or K_i".

Motivation for Kinase Ligand Analysis. The 3DMP approach analyzes activity data for a set of ligands within the

Received: April 4, 2013
Published: June 30, 2013

3D context of the receptor binding pocket. In order to integrate the ligand activity data with the receptor structure, we require 3D models of each ligand bound to our target of interest, P38α. As detailed below, the orientation of the modeled ligands in the binding site determines which fragment pairs constitute valid matched pairs, so it is important to only use models where we are confident that the pose recapitulates the actual binding mode of the ligand in P38α. We assume that high quality P38α models can be built for compounds that have been crystallized in complex with a kinase (not necessarily P38α) as well as for close analogs of kinase crystal ligands. This assumption has two parts: first, we assume that a compound will bind to different kinases using the same binding mode, such that if a ligand has been crystallized with a kinase other than P38α, it is likely to adopt the same binding pose in the P38α binding pocket as in the non-P38α crystal structure. We also assume that close analogs of a crystal ligand will bind in a similar fashion: if a database compound and a crystal ligand are sufficiently similar in structure (with similarity above some threshold t), then the binding pose for the database compound in P38α will closely resemble the binding pose of the crystal ligand in its cognate receptor.

In the first part of this study, we validate these key assumptions by analyzing the relationship between kinase ligand structural similarity and binding pose similarity in a large data set of kinase−ligand crystal structures. Our goal was to first demonstrate that ligands employ consistent binding modes in different kinases and then to find the threshold t for structural similarity such that two compounds with similarity $\geq t$ are likely to share the same binding mode in kinases. We then apply the results to the P38α CHEMBL data set to build high quality binding models for compounds that are sufficiently similar to kinase crystal ligands.

Kinase Crystal Structure Data Set. Our set of kinase−ligand crystal structures consists of 1081 crystal structures from the RCSB PDB and 947 in-house crystal structures from the BMS-PDB.[14] Each protein chain and its associated ligand in the BMS-PDB is overlaid to a conserved kinase core structure, such that the ATP binding sites of all of the protein−ligand complexes are aligned.[14] For each structure, a single representative chain and its associated ligand were kept. The minimum distance between any ligand atom and the coordinates of the canonical kinase hinge residue was calculated for each ligand, and complexes with ligands >20 Å away from the hinge were excluded, since we are only interested in ligands in the vicinity of the ATP binding site. For ligands that were crystallized with the same protein multiple times, the complex with the highest resolution was selected, and the remaining data sets were filtered out. After these preprocessing steps, 1800 kinase−ligand complexes remained.

Identical Ligands Crystallized with Different Kinases. To verify that the same ligand will typically bind to different kinases in the same binding mode, we identified ligand pairs with the same canonical isomeric smiles and hence the same molecular structure. For each pair of identical ligands, the associated proteins were examined to determine whether the two proteins were the same. Cases where the ligands were identical and the receptors were splice variants of the same protein or homologues of different species were excluded. A total of 1882 pairs involving 312 different protein−ligand complexes were found where the ligands are identical and the proteins are different.

Binding Site Similarity of Identical Ligand Set. Having identified a set of identical ligands in different proteins, we next examined whether the ligands in fact bound to different kinases in a consistent manner. For each of the 1882 pairs, binding mode similarities were calculated in two ways. The root-mean-square deviation (RMSD) in atomic coordinates was calculated for all heavy atoms. The ligand structures were extracted from the prealigned kinases with no further optimization of the overlay, so the RMSD reflects the difference in ligand orientation in the two binding sites. Visual inspection of representative ligand pairs across the range of RMSD values allowed the rough determination of thresholds for binding pose similarity: ligands with RMSDs $\leq$ 2.5 Å occupied the same binding pose, while ligands with RMSDs > 4 Å tended to bind differently.

RMSDs are useful for evaluating binding site similarity of ligands that are highly similar, where atoms of ligand A can be definitively mapped to the atoms of ligand B and the deviation in coordinates can be calculated for each pair of equivalent atoms. In subsequent steps we will analyze binding similarity of less similar ligands, where no such atom-to-atom equivalency may exist. We therefore calculated a second measure of binding mode similarity, S_{ov}, that does not require a pairwise atomic mapping and calibrated the S_{ov} scores relative to RMSDs. The binding site overlap score S_{ov} was calculated by an in-house program using the Shape Toolkit[15] that uses a smooth Gaussian function to represent the molecular volume and computes the extent of shape overlap for different molecules.[16] The score has two components: a shape similarity score that quantifies the extent to which two molecules overlap in space regardless of what atom types occupy overlapping volumes; and a color score that accounts for the overlap of specific pharmacophoric groups. The overlap score S_{ov}, like the RMSD, is calculated for two ligand structures in place, with no overlay optimization, and hence the overlap score of two ligands reflects whether they occupy the same local regions of the kinase ATP binding site with similar atom types.

To determine an S_{ov} cutoff that defines similar binding modes, the overlap scores were calibrated to the RMSD values. The RMSDs were first binned according to whether they were $\leq$2.5 Å (same binding mode) or >2.5 Å (potentially different binding modes). For each S_{ov} cutoff x between 0 and 2 (at intervals of 0.1), the following values were calculated: TP_x = number of pairs with $S_{ov} \geq x$ and RMSD $\leq$ 2.5 Å (true positives); FP_x = number of pairs with $S_{ov} \geq x$ and RMSD > 2.5 Å (false positives); TN_x = number of pairs with $S_{ov} < x$ and RMSD > 2.5 Å (true negatives); FN_x = number of pairs with $S_{ov} < x$ and RMSD $\leq$ 2.5 Å (false negatives); $Sensitivity_x = TP_x/(TP_x+FN_x)$, i.e. the fraction of pairs with RMSD $\leq$ 2.5 Å that have $S_{ov} \geq x$, or the true positive rate; and $Precision_x = TP_x/(TP_x + FP_x)$, i.e. the fraction of pairs with $S_{ov} \geq x$ that have RMSD $\leq$ 2.5 Å, or the positive predictive value.

Each S_{ov} cutoff provides a different trade-off between sensitivity (the fraction of pairs with the same binding mode that score above the cutoff) and precision (the fraction of pairs above the cutoff that have the same binding mode). A threshold of 0.8 was selected for identifying with high confidence ligand pairs that have similar binding modes (sensitivity = 88% and precision = 98%).

Structural Similarity and Binding Site Similarity for Nonidentical Ligands. To validate our second assumption—that similar ligands will share the same kinase binding mode—we compared structural similarity and binding site similarity

dx.doi.org/10.1021/ci400201k | J. Chem. Inf. Model. 2013, 53, 1576−1588

across the full set of 1800 kinase ligands. For each pair of ligands, binding site similarity scores (S_{ov}) were calculated as described for the identical ligands. In contrast to the preceding analysis, which was restricted to identical ligands in different proteins, in this step the ligand pairs were not filtered, so the data set of identical ligands in different proteins described above is a subset of this larger ligand pair set. To calculate shape-based structural similarity, each ligand was converted to a 2D isomeric SMILES representation, and conformations were generated with Omega v2.4.6.[17] The parameters were set to allow fine sampling of conformational space: the RMS threshold for unique conformations was set to 0 to avoid filtering out very similar conformations, and the maximum number of conformations per molecule was set to 100 000. For a pair of ligands A and B, the conformational ensemble of A was aligned to the reference crystal structure conformation of B, where each alignment was optimized to maximize the similarity score. The final similarity score Sim_{struc} for A and B was assigned to be the maximum alignment score for any conformation of A to the crystal structure of B.

With S_{ov} and Sim_{struc} scores computed for each ligand pair, we can find the threshold t for structural similarity such that two compounds with $\mathrm{Sim}_{struc} \geq t$ are likely to have S_{ov} scores $\geq$ 0.8, indicating that they share the same kinase binding mode. For each Sim_{struc} score bin i = 0.5, 0.6, 0.7, ...1.9, the percent of ligand pairs with S_{ov} scores $\geq$ 0.8 was calculated as

$$\mathrm{pbindsim}(i) = 100(\text{number of pairs with } i \leq \mathrm{Sim}_{struc} < i + 0.1 \text{ and } S_{ov} \geq 0.8) / (\text{number of pairs with } i \leq \mathrm{Sim}_{struc} < i + 0.1)$$

To model the observed pbindsim values, a logistic regression model for pbindsim as a function of Sim_{struc} was generated using R. The resulting model is

$$\mathrm{pbindsim}_{mod}(i) = 100 \exp(-11.668 + 8.423 \times \mathrm{Sim}_{struc}) / (1 + \exp(-11.668 + 8.423 \times \mathrm{Sim}_{struc}))$$

Ninety-five percent confidence intervals for the coefficient estimates based on the profiled log-likelihood were obtained with the confint() function: they are, respectively, [−11.713, −11.622] and [8.380, 8.466]. For values of Sim_{struc} above 1.4, the model predicts $\mathrm{pbindsim}_{mod}$ values that are higher than the corresponding pbindsim values observed in the data set.

While the final selection of t is subjective, the value of 1.6 was chosen because most compounds with similarity to a kinase ligand Sim_{struc} above 1.6 are highly likely to bind kinases in the same manner as the crystal ligand (i.e., pbindsim(1.6) = 76.7%; with a 95% confidence interval for $\mathrm{pbindsim}_{mod}$ of 84.5−87.2%). The Sim_{struc} threshold is a key parameter for determining how many of the posed ligands—and therefore how much of the matched pair SAR—will be incorporated in the database. While in this study we selected a conservative threshold, for other data sets a user may wish to select a lower Sim_{struc} threshold to achieve a different balance of pose quality versus SAR comprehensiveness.

P38α CHEMBL Binding Models. We can now apply the results of the kinase ligand analysis and generate high confidence binding poses for P38α database compounds with structural similarity Sim_{struc} to a kinase ligand above 1.6. For each P38α inhibitor, we identify the best potential crystal structure template by calculating the pairwise 2D similarities with each kinase crystal ligand using atom pair fingerprints and the Tanimoto coefficient as a similarity metric.[18] The crystal ligand with the maximum similarity to the given P38α compound is designated its reference structure. We then build an ensemble of 3D conformations for each P38α database compound and overlay the conformations to the reference structure as described above. The maximum overlay score for the ensemble provides a measure of the structural similarity of the database compound and the reference crystal ligand, sim_{struc}, and the highest-scoring overlay constitutes a model of the inhibitor bound to P38α (since all of the kinase structures are in the same coordinate frame with prealigned binding sites). Of the 4291 candidate models, only 1017 models with $\mathrm{sim}_{struc} \geq 1.6$ were kept, since for these structures we are reasonably confident that the binding mode is correct. Lowering the sim_{struc} threshold for accepting a binding model would enable the inclusion of many more database compounds and therefore many additional SAR points. In practice, it may often be preferable to accept a lower degree of confidence and incorporate more data, but for the purpose of illustrating the methodology, we intentionally restrict ourselves to a smaller, higher quality set of models.

Construction of the 3DMP Database. The set of 3D binding models for P38α inhibitors and their associated activities (IC50s and/or K_i's) are processed to create the 3DMP database via the following workflow:

1. Each posed ligand is broken into sets of fragment pairs where each pair is the result of breaking a single bond in the original structure. The conversion of the molecules to fragments is performed with an in-house tool that uses a set of SMARTS rules to identify chemically accessible bonds in the complete molecule, much like RECAP,[19] including the ability for the user to specify custom bond cleavage rules. These bonds are systematically broken to enumerate all possible fragmentations of the molecule, including fragmentations that result in small functional groups. The exhaustive enumeration performed by the tool means it breaks the acyclic bond in a biphenyl, but it also retains the biphenyl intact. The tool is flexible to allow for simultaneous as well as sequential application of rules depending on the synthetic accessibility of the fragmentation positions. The fragments output by this step retain their original *xyz* coordinates. The fragments also contain an "attachment point"—a pseudoatom where the bond in the parent structure was broken. If a given fragment pair A−B is joined in the parent molecule by a bond between atom *a* of fragment A and atom *b* from fragment B, then following the fragmentation step, fragment A contains a pseudoatom with the same coordinates as atom *b* and fragment B contains a pseudoatom with the coordinates of atom *a* (Figure 1a).

2. Pairs of compounds are identified where fragment B in compound 1 (the "source" fragment) is replaced by fragment C in compound 2 (the "destination" fragment). The remainder of the structure is the "core," which is common to both compounds and is attached to fragment B in compound 1 and fragment C in compound 2. Importantly, pairs are only kept if the attachment points for the two fragments are <1 Å apart. The distance filter ensures that fragment B and fragment C project to the same region of the receptor binding site and may be considered a relevant match. The coordinates of the core pseudoatom for compound 1 are designated as the attachment point for the matched pair (Figure 1a).

Since compounds are broken into multiple fragment pairs in step 1, the same pair of compounds may give rise to multiple

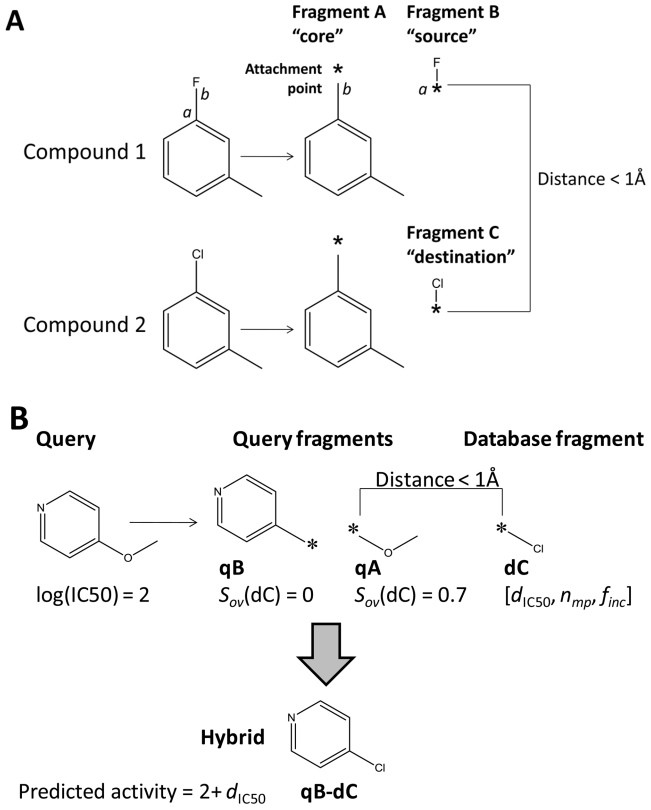


Figure 1. 3D matched pair workflow. (a) Generation of matched pairs. The bond between atoms *a* and *b* in compound 1 is broken to yield fragment A, the core fragment, and fragment B, the source fragment. The coordinates of atoms *a* and *b* are assigned to pseudoatoms (denoted *); the pseudoatom of the core fragment serves as the attachment point. Similarly, compound 2 is fragmented to yield the same core fragment and fragment C, the destination fragment. If the distance between the pseudoatoms in fragments B and C is <1 Å, then fragments A (core), B (source), and C (destination) comprise a 3D matched pair. (b) Querying the 3DMP database. The query structure is fragmented to yield fragments qA and qB. Fragment dC is retrieved from the 3DMP database since its pseudoatom is <1 Å from the pseudoatom in qA. Fragment dC overlaps in space better with qA (s_{ov} = 0.7) than with qB (s_{ov} = 0), so qA is replaced with dC, and a bond is created between fragments qB and dC to yield a novel hybrid compound. The predicted activity of the hybrid is equal to the activity of the query plus the d_{IC50} value for fragment dC, i.e. 2+ d_{IC50}.

matched pairs corresponding to different fragmentations of the parent compounds. For example, if compound 1 is 1-fluoro-3-methylbenzene and compound 2 is 1-chloro-3-methylbenzene, one transformation would consist of methylbenzene as the core, fluorine as the source fragment, and chlorine as the destination fragment (Figure 1a). A second transformation would have just methane as the core, fluorobenzene as the source fragment, and chlorobenzene as the destination fragment.

3. Matched pairs where the parent compounds were not tested in the same assay are excluded. The CHEMBL database includes data from 385 different assays. The building block of matched pair analysis is the change in activity for two compounds, which can only be assessed if the two activity values are comparable. For our analysis, we consider all K_i values to be comparable, since the inhibition constants should be independent of the substrate concentrations used,[20] and therefore matched pairs where the parent compounds' K_i's were obtained in different assays were kept. IC50s, which vary depending on substrate concentrations, can only be directly compared if they were obtained in the same assay. Once pairwise activity shifts are computed, it is possible to compare activity shifts for different compound *pairs* that were assayed in different experiments. We therefore ignore matched pairs where the IC50s were not measured in the same assay. Of >320 000 potential matched pairs from step 2, ∼80 000 passed this step.

4. The matched pairs are mapped to the activities (IC50s or K_i's) of the parent compounds. If a compound pair was tested in multiple assays, a single set of activities is selected. If both K_i's and IC50s were measured for the same compound pair, the K_i values are selected. If multiple K_i's are available, the minimum K_i for each compound is used. If multiple IC50s are present in the data set, the assays are prioritized by the number of compounds tested in each assay, and the IC50s from the assay with the maximum number of compounds is used. Of the 80 000 matched pairs, ∼22 000 are associated with K_i's and the remainder with IC50s.

5. The matched pairs are clustered in space by the location of their attachment points (as described in step 2, the attachment point for the matched pair is the core attachment point for compound 1). The *xyz* coordinates of each matched pair's attachment point are extracted, and all pairwise Euclidean distances are calculated. The distance matrix serves as input for a leader-based average-linkage clustering[21] where the distance cutoff is set to 1 Å. Each matched pair is assigned a cluster ID; matched pairs in the same cluster have attachment points within ∼1 Å of each other. Leader-based clustering was utilized for computational efficiency. We did not investigate the sensitivity of the approach to other clustering methods but expect the results to be largely consistent with those shown here.

6. The 3DMP database can be further filtered to speed up queries and to remove matched pairs with fragments that have undesirable properties. For the queries described in the next section, matched pairs with very large destination fragments (MW > 300) were removed, producing a final database of ∼30 000 matched pairs.

Querying the 3DMP Database. Given a query molecule posed in the P38α binding site, the 3DMP database can be searched for matched pairs that correspond to a fragment of the query compound. The search tool first fragments the query as described above and searches for all matched pairs with an attachment point in close proximity (the default distance is 1 Å) to a query fragment's attachment point. The ability to find relevant matched pairs based only on whether fragments occupy the same region in the binding site is an advantage of 3DMP compared to standard matched pair methods, which can only find matched pairs where the source fragment is similar in structure to a fragment in the query. In contrast, 3DMPs can extend the search to nonsimilar fragments that are nonetheless relevant since they interact with the same receptor pocket. The matched pairs can then optionally be filtered by similarity of the destination fragment to the query fragment (using either atom-pair Tanimoto coefficients or shape similarity) to tune the relevance of the retrieved data.

Once relevant matched pairs are identified, the data for each unique destination fragment in each region of the binding site is aggregated in order to quantify the localized impact of each R-group on P38α potency. Each matched pair has an associated cluster ID that maps it to a specific location. Within each cluster, matched pairs with the same destination fragment

dx.doi.org/10.1021/ci400201k | J. Chem. Inf. Model. 2013, 53, 1576−1588

(determined by canonical SMILES) are grouped together and the average change in log(IC50)—or log(K_i)—is calculated. This average activity shift, or d_{IC50}, is used to describe the fragment's aggregate effect on P38α activity at a particular site. Negative values of d_{IC50} indicate that, on average, compounds increase in potency when the given destination fragment is placed at the relevant site in the binding pocket; conversely, positive values of d_{IC50} reflect a decrease in potency. In addition to d_{IC50}, the number of associated matched pairs (n_{mp}) and the fraction of pairs for which potency increases (f_{inc}) are also tracked. A one-sided paired Student's t test is also performed to determine whether the activity shift is statistically significant.

Generating Novel Hybrid Ligands. Once a search is complete, a set of destination fragments has been found whose members are in close proximity to a fragment of the query molecule. Each destination fragment is mapped to a specific query fragment (qA). The aggregated activity data can be used to select destination fragments that we wish to incorporate into the starting molecule. For example, if the goal is to increase potency, the search results can be filtered to keep only fragments with negative d_{IC50} values where a large fraction of the associated pairs demonstrated increased potency. To hybridize the database fragment (dC) with the query compound, overlap scores (s_{ov}) are calculated for dC and qA, as well as for dC and query fragment qB, where fragments qA and qB together comprise the complete query molecule (Figure 1b). The fragment with the greater overlap score is the target for replacement, since the database fragment occupies an overlapping volume. If fragment qA has the larger overlap score, then we replace qA with dC and form a bond between qB and dC. (If instead fragment qB has the larger overlap score and the attachment points for qB and dC are <2 Å, then we replace qB with dC and form a bond between qA and dC.) The activity metrics for dC (d_{IC50}, n_{mp}, and f_{inc}) are transferred to the new hybrid qB−dC molecule; the predicted activity of the hybrid is log(IC50 of the query) + d_{IC50} (Figure 1b).

Finding SAR Transfer Networks. For designing novel hybrid ligands, the 3DMP search algorithm finds matched pairs where *different* source fragments are replaced by the same destination fragment at the same site and aggregates the relevant activity data. The matched pair database can also be used to identify cases of SAR transfer, where the *same* source fragment is replaced by the same destination fragment for two different cores at the same location in the binding pocket, and the replacement results in the same activity shift (i.e., $d_{IC50} < 0$ for both matched pairs or $d_{IC50} > 0$ for both). For each pair of 3DMPs, the criteria for SAR transfer are met if the canonical smiles of the source fragments and the destination fragments are the same, the canonical smiles of the core fragments are different, the distance between the two attachment points is <1 Å, and the sign of d_{IC50} is the same.

Annotating the Receptor Binding Pocket. The SAR data extracted in the 3DMP database can be used to annotate the receptor binding site. Although the database fragments vary in size, each matched pair is also mapped to a single point—the *xyz* coordinates of the attachment point, which is the site where the source and destination molecules diverge in structure. The coordinates of the attachment points are in turn clustered (see above), so that a group of matched pairs that cluster together can be represented visually by the coordinates of the cluster centroid. Various properties of each matched pair cluster can be computed and mapped onto the binding site at the centroid coordinates. It is important to note that since we cluster pairs coarsely using only the attachment point, the matched pairs in a given cluster may project their respective R groups along different vectors; a cluster simply represents a site where the ligand structures have been modified (for the p38 database, 92% of 3DMP pairs from the same cluster do project their R groups to a common region of space—data not shown).

The most basic property of a matched pair cluster c is the number of pairs belonging to cluster c. The number of associated matched pairs, n_c, is counted for each of the 152 clusters. To visualize the results, a pseudoatom is placed at the coordinates of each cluster centroid and colored according to log(n_c). The centroid pseudoatoms are shown within the P38α receptor binding site, with the receptor and reference ligand structures taken from the PDB entry 3BV3[22] (residues 166−171 removed for clarity in Figure 8).

In addition to n_c, we also count nimp$_c$, the number of matched pairs in cluster c where the substitution had a substantial impact on activity (absolute value of the change in log IC50 $\geq$ 1.5). For clusters with nimp$_c \geq 1$, we calculate nimp$_c/n_c$, the fraction of pairs in cluster c that meet the threshold for impact. Again, the resulting values are mapped to the cluster centroid and colored by the value of nimp$_c/n_c$ (green = 0%, magenta = 30%+). Clusters where no substitution substantially impacted activity (nimp$_c$ = 0) are not mapped (Figure 8b and d).

3DMP Database for in-House P38α and CDK2 Data. The CHEMBL database is useful for illustrating the potential applications of the 3DMP approach, but the data set is sparser than for a typical proprietary database. In the context of a drug discovery project, thousands of structurally similar compounds may be tested under the same assay conditions, such that many more matched pairs can be extracted. To demonstrate more fully the ability of 3DMPs to annotate a receptor binding site, a 3DMP database was prepared for a set of in-house compounds that were tested in the same P38α assay. A total of 3717 molecules met the criteria for high confidence model generation (sim$_{struc}$ to a crystal ligand $\geq$ 1.6), and these were processed through the same 3DMP pipeline as described above, yielding a database of >300 000 matched pairs grouped spatially into 403 clusters. The matched pair counts and percent of impactful matched pairs for each cluster were calculated and visualized as for the CHEMBL database.

We also wanted to simulate the comparison of activities for multiple kinases in a set of compounds of interest. We identified 209 000 matched pairs from the P38α database where both compounds had also been tested in an in-house CDK2 assay. These pairs cover 348 of the 403 P38α clusters. The matched pair counts and percent of impactful matched pairs for each cluster were calculated and visualized as for the P38α data. The aligned crystal structure of CDK2 (PDB 2R3Q[23]) was used for comparing the P38α and CDK2 binding sites, with residues 8−18 and 153−166 removed for clarity.

RESULTS

Binding Pose Similarity for Identical Ligands. In order to build a 3D matched pair database, we need reliable models of the compounds in the data set complexed to the target, P38α. Since most kinases form similar key interactions with ligands in the ATP binding site, it is reasonable to hypothesize that the binding mode of an inhibitor in P38α would be similar to the binding mode of a close analog in another kinase for which crystal structures are available. In the first part of this study, we first confirm that the same inhibitors employ the same binding

dx.doi.org/10.1021/ci400201k | J. Chem. Inf. Model. 2013, 53, 1576−1588

mode to interact with different kinases and that close analogs adopt similar binding poses. We then apply the results and use known inhibitor−kinase complexes to model the interactions of closely related analogs with P38α.

To determine whether ligands interact in the same manner with multiple kinases, a set of 312 aligned crystal structures was assembled in which the same ligands were crystallized in complex with different kinases (see Methods). The set contains structures of 105 unique ligands. For each pair of identical ligands in complex with different kinases, binding pose similarity was assessed by calculating the pairwise RMSD. In agreement with the hypothesis that inhibitors bind in a consistent fashion to multiple kinases, 88% of the 1882 pairs have RMSDs below 2.5 Å, indicating highly similar binding modes, and 97% have RMSDs below 4 Å, a threshold for moderately similar binding modes (Figure 2).

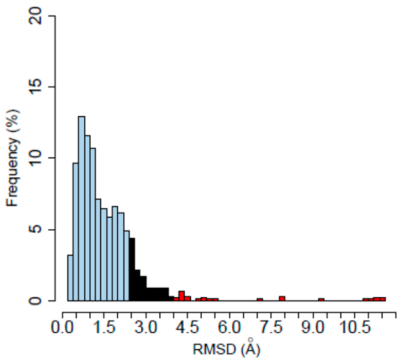


Figure 2. Histogram of pairwise RMSDs for identical ligands bound to different kinases. RMSDs for pairs where the ligand binds two kinases in the same binding mode (<2.5 Å) are shown in blue. Moderately similar binding modes (2.5−4 Å) are in black, and dissimilar binding modes (>4 Å) are in red.

In addition to RMSDs, which provide a gold standard for binding site similarity, each identical ligand pair was also scored for shape overlap, s_{ov}.[24] In the next stage, we will need to evaluate binding site similarity for less similar compounds and determine which compound pairs share the same binding mode. Unlike RMSDs, calculating s_{ov} does not require correspondence between the atoms of different compounds and hence can be applied to dissimilar compounds. We therefore calculate both RMSDs and s_{ov} scores for the identical pair set and use the RMSDs to find the s_{ov} threshold that best corresponds to an RMSD threshold of 2.5 Å. Figure 3a shows that s_{ov} scores are correlated with RMSD (r = −0.75). Ninety-eight percent of compound pairs with overlap scores of ≥0.8 have RMSDs of ≤2.5 Å; i.e., they share the same binding mode. Figure 3b plots the receiver-operator characteristic (ROC) curve for determining whether compounds have the same binding mode (RMSD ≤2.5 Å) using different s_{ov} cutoffs. Each score cutoff has a different trade-off between sensitivity (the fraction of pairs with the same binding mode that score above the cutoff) and specificity or precision (the fraction of pairs above the cutoff that have the same binding mode). An s_{ov} threshold of 0.8 was selected for identifying with high confidence ligand pairs that have similar binding modes (sensitivity = 88% and precision = 98%).

Binding Pose Similarity for Nonidentical Ligands. Having verified that a kinase ligand tends to bind different kinases with the same binding mode, we can assume that if a particular ligand has been crystallized with any kinase, we can infer its binding mode in the target of interest, P38α. However, if no crystal structures of a compound are available—but a close analog has been crystallized—can we reliably build a model of the compound bound to P38α? To answer this question, we need to know what level of ligand structural similarity allows a reasonable assumption of binding mode similarity. We therefore compared structural similarity (Sim_{struc}) and binding mode similarity (s_{ov}) across a set of 1800 kinase crystal structure ligands (see Methods).

Table 1 lists the percent of ligand pairs in each Sim_{struc} score interval, pbindsim, that have similar binding modes ($s_{ov} \geq 0.8$). Analogs that are more similar in structure are also more likely to bind in similar modes, as we might expect. For instance, only 9.4% of ligands with Sim_{struc} scores of 1.1 to 1.2 bind with similar modes, while 89% of highly similar ligand pairs (Sim_{struc} scores 1.8 to 1.9) have the same binding mode (see Methods for additional details on pbindsim).

Construction of the P38α 3D Matched Pair Database. The results for binding pose similarity of nonidentical ligands suggest that reliable binding models can be built for kinase ligands with similarity (Sim_{struc}) to a crystal structure ligand ≥1.6 (Table 1). Models for the CHEMBL P38α database compounds were built using the most similar crystal ligand analogs, and only P38α compounds with Sim_{struc} ≥1.6 were kept to ensure that the bound poses were mostly reasonable.

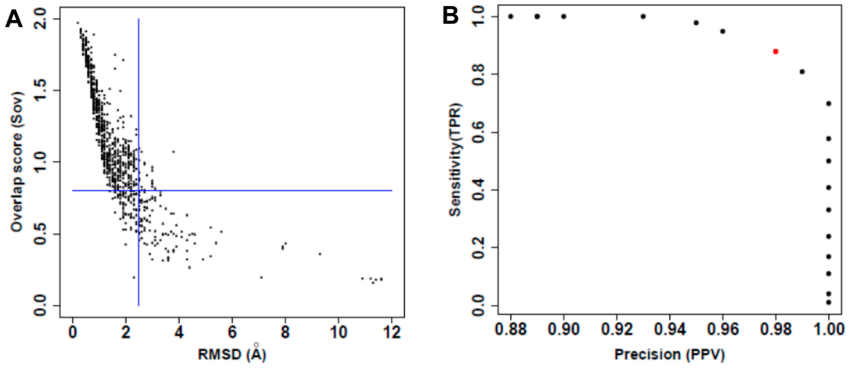


Figure 3. RMSD and overlap scores as measures of binding pose similarity. (a) Scatter plot of RMSD vs overlap score (s_{ov}) for identical ligands bound to different kinases. The vertical blue line marks RMSD = 2.5 Å, and the horizontal blue line marks overlap score = 0.8. (b) Receiver-operator characteristic (ROC) curve for overlap scores as a predictor of binding pose similarity: the sensitivity and precision are plotted for different overlap score cutoffs. The red data point indicates where the s_{ov} cutoff = 0.8.

dx.doi.org/10.1021/ci400201k | J. Chem. Inf. Model. 2013, 53, 1576−1588

Table 1. Pbindsim (% Compound Pairs with Similar Binding Modes) and the 95% confidence interval for Pbindsim for Each Structural Similarity Bin (Sim_{struc})

Sim_{struc}	Pbindsim (%)	$Confint_{95}$
1.9	94.7	98.5−98.9
1.8	88.9	96.7−97.4
1.7	87.8	92.6−94.1
1.6	76.7	84.5−87.2
1.5	60.2	70.2−74.6
1.4	45.3	50.5−55.7
1.3	31.9	30.6−35.0
1.2	19.0	16.0−18.8
1.1	9.4	7.6−9.0
1	3.8	3.4−4.1
0.9	1.7	1.5−1.8
0.8	0.5	0.7−0.8
0.7	0.2	0.3−0.3
0.6	0.1	0.1−0.1
0.5	0.0	0.1 − 0.1

This set of modeled analogs, with their associated P38α IC50s or K_i's, forms the basis of the P38α 3D matched pair database. The pipeline to generate bound models and to calculate 3DMPs is described in the Methods.

Ligand Design with 3D Matched Pairs. One application of matched pairs is the design of new analogs that are predicted to improve potency (if the receptor in question is the desired target) or decrease affinity (if the receptor is an unwanted off-target). Here, we describe a procedure to search the 3D matched pair database and create novel hybrid molecules in combination with fragments derived from a starting molecule. Given a query molecule posed in the P38α binding site, the 3D matched pair search algorithm first fragments the query and searches for all matched pairs with an attachment point in close proximity to a query fragment. Once matched pairs are identified, the data for each unique destination fragment in each region of the binding site are aggregated in order to quantify the localized impact of each R group on P38α potency (see Methods). For each collocated cluster of matched pairs with a common destination fragment, three parameters describe the fragment's aggregate effect on P38α activity: the number of associated matched pairs (n_{mp}), the fraction of pairs for which potency increases (f_{inc}), and the average log change in IC50 (d_{IC50}). Negative values of d_{IC50} indicate that, on average, compounds increase in potency when the given destination fragment is placed at the relevant site in the binding pocket; conversely, positive values of d_{IC50} reflect a decrease in potency. These parameters can also be used to predict the likely effect of adding a specific fragment to a molecule at a particular substitution vector.

To illustrate the query process, a pair of overlaid compounds (Figure 4) was selected that are structurally similar to each other but did not meet the strict criteria for inclusion in the matched pair database (both compounds had maximum shape similarity to a kinase crystal ligand Sim_{struc} = 1.48, which is below the threshold of 1.6). The binding model of the first compound, CHEMBL1807445, was used as a query to search the matched pair database. The search process generated >1000 modified compounds in which a fragment of the query was replaced by a fragment of another compound. The predicted activity of each hybrid relative to the query is the new fragment's d_{IC50} value. One of the "novel" analogs found by the search is the second compound, CHEMBL1807446, in which the pendant *N*-isopropylbenzamide of '445 is replaced by *N*-cyclopropylbenzamide. '446 is predicted to be 10 times more potent than the query (d_{IC50} = −0.9); in fact, '446 is 7-fold more potent than the query molecule (Figure 4).

The accurate activity prediction for '446 was based on 61 3D matched pairs retrieved by the search (n_{mp} = 61). In all 61 pairs, a cyclopropylamide replaced a different fragment at the same 3D site, and in 79% of the pairs the substitution resulted in a gain of potency (f_{inc} = 79%), with an average IC50 shift of −0.9 log units (d_{IC50} = −0.9). The potency shift across the 61 pairs is statistically significant, with a Student's paired *t* test p-value of 4e−9. Examples of the 3D matched pairs are shown as 2d structures in Figure 5a, and the overlaid 3D structures of the destination compounds are shown in Figure 5b. Although the scaffolds are structurally diverse, they project the cyclopropylamide to the same 3D location (Figure 5b), and in each case the cyclopropylamide at this site boosts potency. The search locates matched pairs where the source fragment occupies the same binding site region as the query fragment (here, the *N*-isopropylbenzamide), even though the source fragment may be structurally dissimilar to the query fragment (see Methods). The search can also be optionally restricted to matched pairs where the source fragment is structurally similar to the query fragment to ensure that the data included in calculating the d_{IC50} are directly relevant. In the example query result, restricting the search would have retrieved only pairs involving alkyl-substituted amides that are similar to the query's N-isopropylbenzamide, excluding the matched pair where the source fragment is benzoic acid (Figure 5a).

In addition to CHEMBL1807446 (which was not present in the 3DMP database), five other "novel" analogs predicted to increase P38α activity (d_{IC50} < 0) correspond to molecules

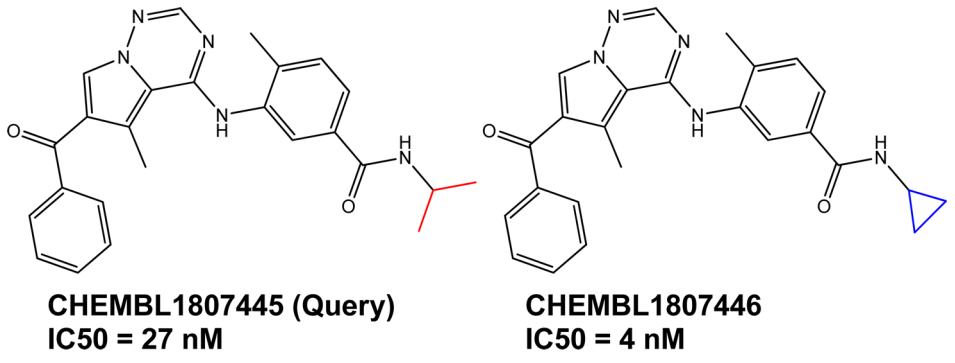


Figure 4. Structures and P38α activities of compounds used to illustrate 3DMP ligand design.

dx.doi.org/10.1021/ci400201k | J. Chem. Inf. Model. 2013, 53, 1576−1588

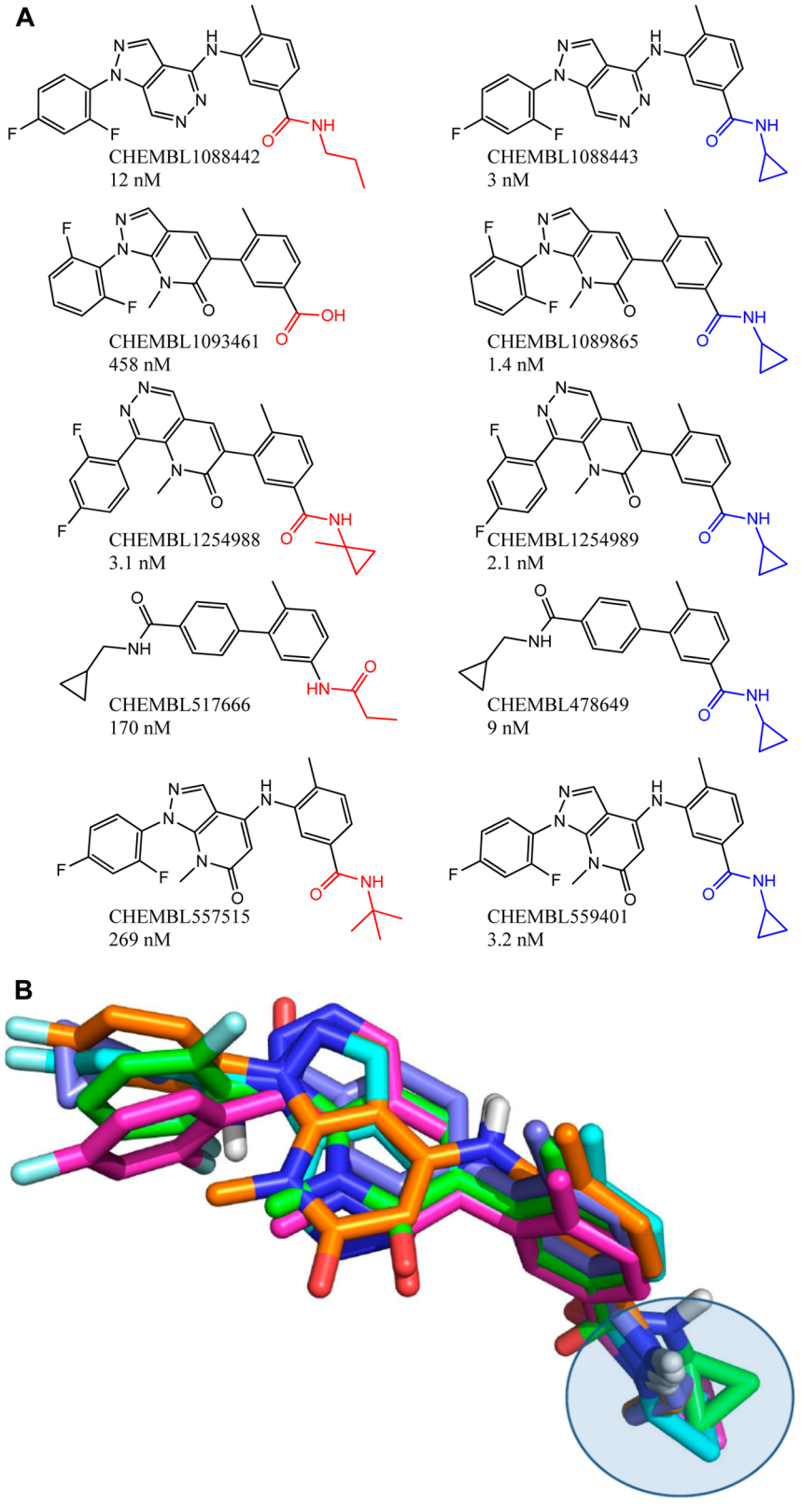
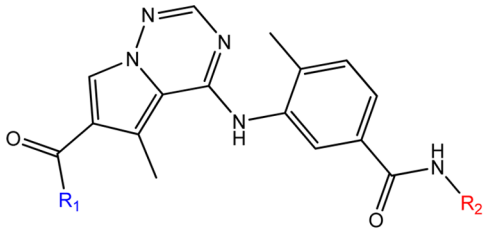


Figure 5. Examples of the 3D matched pairs that contributed to the activity prediction for the "novel" hybrid ligand. (a) 2D structures and IC50s of five representative matched pairs. (b) Overlayed 3D models of the destination compounds.

represented in the database (Figure 6). Of the five, the four with changes to the *N*-alkylamide (R_2) are more potent than the query, in agreement with the predictions. For CHEMBL1807443, the prediction of greater potency is less confident: even though the average change in activity is negative (i.e., on average potency increases), d_{IC50} is only a moderate −0.19, with a statistically insignificant p value, and for five of the eight associated matched pairs, potency actually

dx.doi.org/10.1021/ci400201k | *J. Chem. Inf. Model.* 2013, 53, 1576−1588



CHEMBL#	R1	R2	n_{mp}	f_{inc} (%)	d_{IC50}	p-value	IC50
1807445(Query)	Phe	Isopropyl	-	-	-	-	27
1807444	Phe	Ethyl	6	83	-0.46	0.07	11
1807447	Phe	3-Isoxazole	2	100	-1.49	0.02	3
1807442	Phe	H	3	67	-0.39	0.24	7
1807443	Phe	Methyl	8	37.5	-0.19	0.25	14
1254348	N-ethyl	Isopropyl	5	20	-0.22	0.30	259

Figure 6. "Novel" analogs predicted to increase P38α activity relative to the query (CHEMBL1807445). For each analog, the parameters of the predicted activity (n_{mpmp}, f_{inc}, d_{IC50}, and Student's t test p value) as well as the experimentally determined IC50 are shown.

decreases (f_{inc} = 37.5%). For '348, which replaces the R_1 phenyl of the query with *N*-ethyl, the prediction is even weaker: d_{IC50} < 0, but the p value is 0.3. For four of the five matched pairs, potency decreases (f_{inc} is only 20%). It is thus not surprising that '348 is less active than the query (IC50 = 259 nM compared to 27 nM for the query).

SAR Transfer Networks. The ligand design process finds matched pairs where *different* source fragments are replaced by the same destination fragment at the same site. To identify cases of SAR transfer, we search instead for cases where the *same* source fragment is replaced by the same destination fragment for two different cores at the same location in the binding pocket, and the replacement results in the same activity shift (i.e., IC50 moves in the same direction for both matched pairs). Of ∼30 000 matched pairs in the CHEMBL 3DMP database, 2641 met the criteria for SAR transfer. Pairwise examples of SAR transfer can be extended to identify sets of compounds that exemplify the same SAR transfer in multiple cores. We denote these compound sets "SAR transfer networks," where each compound pair is represented as a node that is linked to other nodes by an edge representing the shared transformation. A total of 506 such networks were found; the largest network comprised 11 pairs of compounds with the same SAR trend. An example network of three pairs is shown in Figure 7; in each case, substitution of F by Cl results in a 13- to 18-fold increase in potency. Of note, two of the three chemotypes contain multiple halogen-substituted aromatic rings. If we considered all cases of flourophenyl-to-chlorophenyl substitution in these chemotypes (disregarding other substituents on the phenyl ring), many 2D matched pair methods would not be able to distinguish these three examples of fluoro-to-chloro substitution, which occur on the A ring, from examples where the substitution occurs on the B ring and thus could obscure the SAR trend. Since the A and B rings project to different regions of the P38α binding site (Figure 7b), the 3DMP method can separate substitutions on the A ring from the same substitutions on the B ring and more accurately quantify the impact on activity at each site.

Annotation of the Receptor Binding Pocket. We have seen that by mapping ligand matched pairs to a common 3D coordinate system, 3DMPs can find SAR trends across chemotypes and inform the design of new analogs. The mapping also allows us to summarize an entire data set by depicting how ligand changes affect activity in each region of the receptor binding site. We do this by characterizing each cluster of collocated matched pairs and visualizing the distribution of various properties across all clusters. For example, the CHEMBL P38α 3DMPs belong to 151 different clusters, where each cluster consists of a set of matched pairs with nearby attachment points. The numbers of matched pairs belonging to each cluster are visualized in Figure 8a. The median number of matched pairs per cluster is 23, with a maximum of 2328 pairs and a minimum of 1. The pink and red spheres, which represent the "hot spots" in terms of medicinal activity across the binding site, are located around the critical hinge-binding region and the core scaffold region (the pyrrolotriazine in the sample ligand shown). The substitution vectors that extend toward the extended hinge region and the P38α specificity pocket have been less well sampled, as shown by the blue spheres that occupy these regions of the binding pocket.

Figure 8b highlights areas of the binding pocket that are most sensitive to changes in the ligand. Although a large number of matched pairs involve the core scaffold (Figure 8a), indicating that many changes to the core were made, most of the changes do not greatly impact P38α potency, and hence the clusters around the core have small values of $nimp_c/n_c$ (green spheres, see Methods). In contrast, substitution of the pendant R groups that extend toward the selectivity pocket more frequently result in significant IC50 shifts, and hence this region is more populated with magenta spheres.

The same trend is observed in a proprietary database of >300 000 P38α 3DMPs derived from a set of 3717 analogs that were tested in the same in-house assay under consistent conditions. As in the CHEMBL data set, most of the matched pair "hotspots" are located near the hinge residue Met-109 and the core scaffold, with a few hotspots extending past the hinge toward the solvent and into the specificity pocket (Figure 8c). However, P38α can accommodate many different hinge-binding structural elements, and many of the changes in structure in the hinge/core region are absorbed by the receptor with minimal effect on the compounds' IC50s. As a result, a relatively small fraction of the centroids in this area of the binding site qualify as high impact sites (magenta spheres). In contrast, the P38α selectivity pocket is highly sensitive to changes in the ligand structure, and a high fraction of MPs result in large IC50 perturbations (Figure 8d). While the same

dx.doi.org/10.1021/ci400201k | J. Chem. Inf. Model. 2013, 53, 1576−1588

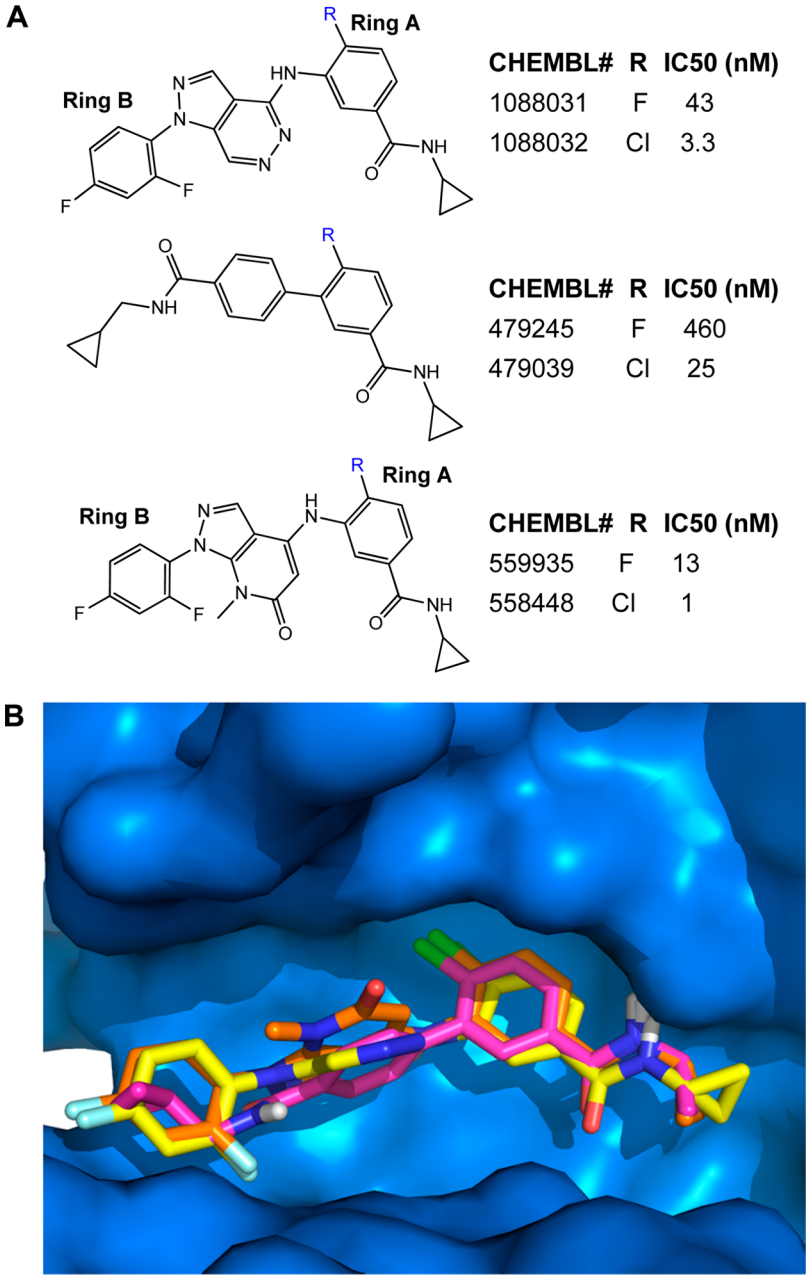


Figure 7. Example of an SAR transfer network. (a) Structures and activities of compound pairs. (b) Overlayed models of CHEMBL1088032 (yellow), CHEMBL479039 (pink), and CHEMBL558448 (orange). The P38α surface was constructed in PyMol using PDB 3HV6 (residues 167–170 removed for clarity).

difference in ligand sensitivity is seen in the CHEMBL database, the trend is greatly amplified in the larger database.

In the context of a kinase drug discovery project, where thousands of compounds designed to be active against one target may be tested against a battery of potential off-targets, it is instructive to annotate off-target SAR onto the binding site of the target of interest. In the case of our proprietary P38α data set, 2145 of the P38α compounds were also tested in a CDK2 assay, and Figure 8e shows the 3DMP clusters that frequently affect CDK2 potency (the plot of the number of matched pairs per cluster for CDK2 is not shown but closely resembles the plot for P38α in Figure 8c). None of the high-impact clusters are located in the P38α specificity pocket, since CDK2 activity is insensitive to changes in this region. Instead, different regions of the binding site are highlighted as critical for modulating CDK2 activity: the extended hinge area, which is more exposed to solvent, and the hydrophobic pocket near the gatekeeper (Phe-80 in CDK2).

DISCUSSION

Ligand Similarity and Binding Pose Similarity: Cores vs R Groups. Studies have shown that ~20–30% of close analogs of kinase inhibitors will themselves be active in the same assay.[25,26] Since ATP-competitive inhibitors tend to preserve the same key ligand–receptor interactions—most notably the hydrogen bonds to the hinge—it is reasonable to expect that the hinge interactions will be conserved across members of the same chemotype and that differences in activity of close analogs can be attributed to structural differences outside the hinge-binding core. Consistent with this notion,

dx.doi.org/10.1021/ci400201k | J. Chem. Inf. Model. 2013, 53, 1576–1588

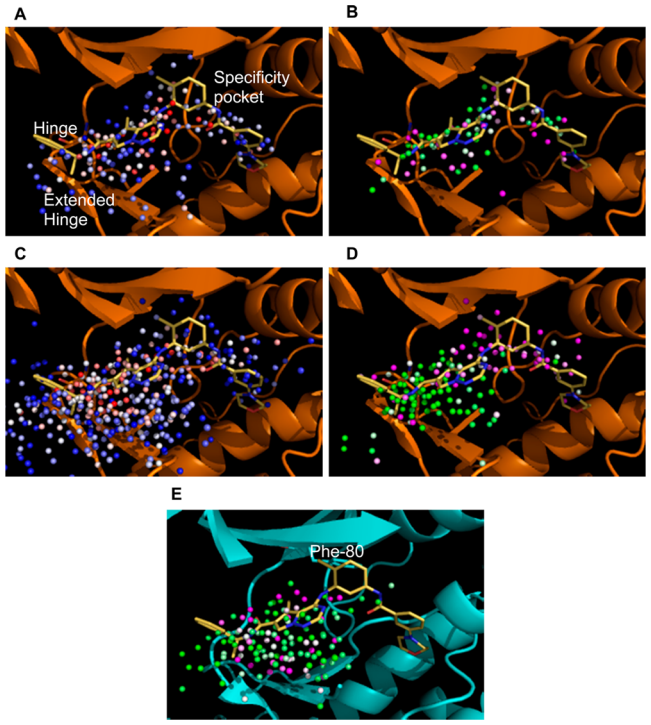


Figure 8. Properties of 3DMP clusters. (a and c) Number of matched pairs (n_c) for clusters in the P38α CHEMBL database (a) or the in-house P38α database (c). Red spheres indicate "hot spots" with large numbers of matched pairs, and blue spheres show clusters with fewer matched pairs. (b and d) Fraction of matched pairs that show a large impact on P38α potency ($nimp_c/n_c$) in the P38α CHEMBL database (b) or the in-house P38α database (d). Magenta spheres indicate clusters with large values of $nimp_c/n_c$, and green spheres show clusters with low values. Clusters with no impactful matched pairs (i.e., $nimp_c/n_c = 0$) are not shown. (e) Fraction of matched pairs that show a large impact on CDK2 potency ($nimp_c/n_c$) in the in-house data set.

Hare et al.[27] have shown that 82% of kinase core scaffolds with three or more rings bind with the core atoms assuming a single shared orientation. Here, we show that while the same ligand will bind to different kinases with the same pose, ligands that are moderately similar in structure may adopt different binding modes to interact with kinases. Even if the core hinge interactions are the same for similar ligands, the orientations of the pendant R groups can vary, which can translate into significantly different affinities for the same kinase. As a consequence, only very highly similar structures—where both the core and the R groups are structurally similar—can be confidently assumed to bind to kinases in the same way (Table 1).

These results are consistent with previous studies: Boström et al.[28] examined a set of 206 pairs of PDB structures from diverse target classes where each pair of ligands is structurally similar and bound to the same protein. They found that 80% of these highly similar ligands occupy the same region of the binding site and have similar binding poses. Similarly, Nicholls et al.[29] determined that high-confidence binding models could be constructed for ligand pairs with shape similarity scores >1.4 in a set of 363 ligand pairs bound to a diverse set of targets. Together, these studies suggest that models of a ligand pose based on a crystallographic template are reliable only when there is relatively high similarity between the template and query ligand structures.

The Importance of MMP Context. Previous studies have demonstrated that the structural context of a fragment can impact the change in activity observed when the fragment is modified.[7,30] For example, replacing hydrogen with a methoxy group is likely to increase hERG activity when the attachment point is next to an H-bond accepting aromatic ring, but the same H→OCH3 substitution tends to have the opposite effect when adjacent to an aliphatic linker.[7] When the structural context of a matched pair is considered, one can isolate effects that are specific to a particular local environment and that may be obscured when different structural contexts are assumed to be equivalent. Similarly, with 3DMPs the 3D environment of a matched pair is also used to identify context-sensitive activity trends.[31] Replacing H with OCH3 may have different consequences if the relevant fragment occupies a buried hydrophobic pocket or if it is exposed to solvent.

A recent study by Weber et al.[32] describes an alternative method for incorporating the binding site context into matched pair analysis. The VAMMPIRE algorithm identifies matched pairs in the CHEMBL database for which at least one molecule (or a close analog) has been crystallized in complex with a target protein. The resulting database stores information about amino acids that may interact with the substituted R groups and can be searched for effects resulting from changing a specific ligand substituent in a given environment.[32] Our approach, which allows for the identification of proximal R groups that have different chemical structures and integrates data from multiple matched pairs, provides a complementary path for leveraging structural information.

Incorporating the 3D context can also enable the transfer of SAR between series in cases where the source fragments (or the core[33]) are not identical but nonetheless occupy the same site in the receptor and are likely to have similar interactions. For example, if one wants to replace a cyclopropyl group with some other moiety to improve potency, 2D MMP methods can find MMPs where the source fragment is a cyclopropyl and determine which destination fragments yield improved potency relative to the cyclopropyl. 3DMPs can include SAR from MMPs where the source fragments are not cyclopropyl groups—but nonetheless project to the same region and therefore are relevant to the query. Instead of asking, "What substitutions involving a cyclopropyl improve potency?" one can ask, "What fragments at the same location as the query cyclopropyl improve potency?"

3D Ligand Design Methods. Other recent fragment-based approaches for ligand design also use crystal structures or models of ligands bound to their cognate receptors. The BREED algorithm hybridizes three-dimensional fragments to generate new compound structures in the context of the receptor binding site.[34−36] Similarly, "fragment hopping" programs such as CAVEAT[37] and BROOD[47] replace a core scaffold or an R group with a different fragment that fits the constraints imposed by the ligand geometry and the binding site. Since these methods incorporate information about the ligand binding context, they can generate hybrids from diverse starting structures. However, they do not capture the SAR learned from comparing activities of compound series. In contrast, MMP-based methods for generating novel analogs are driven by SAR trends but require the query molecule to be similar in shape or 2D structure to the MMP-database compounds in order to find relevant fragments.[38] The 3DMP method combines the advantages of both approaches—it generates diverse structures that are posed in the target's

binding site, and it predicts the activity of the novel structure using MMP-derived SAR.

Annotating Receptor Binding Sites. There are many useful ways to characterize the properties of a receptor binding site. For example, a rendering of the protein surface can be colored according to its electrostatic potential,[39] local hydrophobicity,[40] or cavity depth[41] to provide guidance about potential local interactions with ligands. More sophisticated methods render an interaction map of a binding site with different probe types, whether by calculating probe−protein interaction energies (GRID[42]) or empirical probabilities of interaction based on crystal structures (SuperStar[43]); WaterMap,[44] SZMAP,[45] and 3D-RISM[46] characterize solvation effects and can indicate regions where adding ligand−receptor interactions can be beneficial. Here, we have demonstrated a complementary method to describe the protein surface: in addition to characterizing the local environment based on the protein structure, the ligand SAR can be mapped to the binding site as well. We have shown the utility of mapping simple 3DMP properties such as the frequency of fragment substitution and the relative impact on activity at a given site to find regions of flat or steep SAR. We believe this approach is complementary to a 3D-QSAR annotation of a binding site in that it uses molecular structure explicitly as the annotation rather than the more abstract molecular interaction fields used in QSAR. Extensions might include mapping properties of fragments, such as identifying physicochemical features of fragments that result in potency changes, including the vector of attachment to more finely cluster the matched pairs, and incorporating statistical analysis of the cluster activity trends to identify the most statistically significant effects.

CONCLUSION

Understanding the interactions of small molecules with cognate receptor sites requires integrating the ligand molecular structures and activity data with the context of the protein binding site. Here, we have demonstrated a method for combining ligand and protein data to generate insights about the patterns of chemotype-activity relationships and the properties of the protein environment that drive the observed SAR. While we have illustrated the applications to kinase targets, the method could be extended to any target for which reasonable binding models of the associated ligands are available.

AUTHOR INFORMATION

Corresponding Author

*E-mail: stephen.johnson@bms.com.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Malcolm Davis for providing insight as well as technical assistance and Deborah Loughney for her support.

ABBREVIATIONS

P38α, mitogen-activated protein kinase P38 alpha; MMP, matched molecular pair; PDB, Protein Data Bank; SAR, structure−activity relationships

REFERENCES

(1) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* **2011**, *54* (22), 7739−7750.

(2) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH Verlag GmbH & Co.: Weinheim, Germany, 2005.

(3) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Res.* **2013**, *41* (D1), D1137−1143.

(4) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (1), 103−108.

(5) Zhang, L.; Zhu, H.; Mathiowetz, A.; Gao, H. Deep understanding of structure-solubility relationship for a diverse set of organic compounds using matched molecular pairs. *Bioorg. Med. Chem.* **2011**, *19* (19), 5763−5770.

(6) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* **2006**, *49* (23), 6672−6682.

(7) Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W.; Macdonald, S. J. Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of HERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.* **2010**, *50* (10), 1872−1886.

(8) Dimova, D.; Hu, Y.; Bajorath, J. Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity. *J. Med. Chem.* **2012**, *55* (22), 10220−10228.

(9) Warner, D. J.; Bridgland-Taylor, M. H.; Sefton, C. E.; Wood, D. J. Prospective Prediction of Antitarget Activity by Matched Molecular Pairs Analysis. *Mol. Inf.* **2012**, *31* (5), 365−368.

(10) Milletti, F.; Hermann, J. C. Targeted Kinase Selectivity from Kinase Profiling Data. *ACS Med. Chem. Lett.* **2012**, *3* (5), 383−386.

(11) Zhang, B.; Wassermann, A. M.; Vogt, M.; Bajorath, J. Systematic Assessment of Compound Series with SAR Transfer Potential. *J. Chem. Inf. Model.* **2012**, *52* (12), 3138−3143.

(12) Wassermann, A. M.; Bajorath, J. A data mining method to facilitate SAR transfer. *J. Chem. Inf. Model.* **2011**, *51* (8), 1857−1866.

(13) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100−1107.

(14) Finzel, B. C.; Akavaram, R.; Ragipindi, A.; Van Voorst, J. R.; Cahn, M.; Davis, M. E.; Pokross, M. E.; Sheriff, S.; Baldwin, E. T. Conserved core substructures in the overlay of protein-ligand complexes. *J. Chem. Inf. Model.* **2011**, *51* (8), 1931−1941.

(15) *Shape Toolkit*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012.

(16) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17* (14), 1653−1666.

(17) *Omega*, version 2.4.6; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012.

(18) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48* (13), 4183−4199.

(19) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 511−522.
(20) Yung-Chi, C.; Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50% inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol. (Amsterdam, Neth.)* **1973**, *22* (23), 3099−3108.
(21) Hartigan, J. A. *Clustering Algorithms*; John Wiley & Sons, Inc.: New York, 1975.
(22) Wrobleski, S. T.; Lin, S.; Hynes, J., Jr.; Wu, H.; Pitt, S.; Shen, D. R.; Zhang, R.; Gillooly, K. M.; Shuster, D. J.; McIntyre, K. W.; Doweyko, A. M.; Kish, K. F.; Tredup, J. A.; Duke, G. J.; Sack, J. S.; McKinnon, M.; Dodd, J.; Barrish, J. C.; Schieven, G. L.; Leftheris, K. Synthesis and SAR of new pyrrolo[2,1-f][1,2,4]triazines as potent p38 alpha MAP kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18* (8), 2739−2744.
(23) Fischmann, T. O.; Hruza, A.; Duca, J. S.; Ramanathan, L.; Mayhood, T.; Windsor, W. T.; Le, H. V.; Guzi, T. J.; Dwyer, M. P.; Paruch, K.; Doll, R. J.; Lees, E.; Parry, D.; Seghezzi, W.; Madison, V. Structure-guided discovery of cyclin-dependent kinase inhibitors. *Biopolymers* **2008**, *89* (5), 372−379.
(24) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48* (5), 1489−1495.
(25) Posy, S. L.; Hermsmeier, M. A.; Vaccaro, W.; Ott, K. H.; Todderud, G.; Lippy, J. S.; Trainor, G. L.; Loughney, D. A.; Johnson, S. R. Trends in kinase selectivity: insights for target class-focused library screening. *J. Med. Chem.* **2011**, *54* (1), 54−66.
(26) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45* (19), 4350−4358.
(27) Hare, B. J.; Walters, W. P.; Caron, P. R.; Bemis, G. W. CORES: an automated method for generating three-dimensional models of protein/ligand complexes. *J. Med. Chem.* **2004**, *47* (19), 4731−4740.
(28) Boström, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.* **2006**, *49* (23), 6716−6725.
(29) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.* **2010**, *53* (10), 3862−3886.
(30) Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: a novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. *J. Chem. Inf. Model.* **2010**, *50* (8), 1350−1357.
(31) Mills, J. E. J.; Brown, A. D.; Ryckmans, T.; Miller, D. C.; Skerratt, S. E.; Barker, C. M.; Bunnage, M. E. SAR mining and its application to the design of TRPA1 antagonists. *MedChemComm* **2012**, *3* (2), 174−178.
(32) Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. AMMPIRE: A Matched Molecular Pairs Database for Structure-Based Drug Design and Optimization. *J. Med. Chem.* **2013**, in press.
(33) Hessler, G.; Matter, H.; Schmidt, F.; Giegerich, C.; Wang, L.-h.; Güssregen, S.; Baringhaus, K.-H. Identification and Application of Antitarget Activity Hotspots to Guide Compound Optimization. *Mol. Inf.* **2011**, *30* (11−12), 996−1008.
(34) Pierce, A. C.; Rao, G.; Bemis, G. W. BREED: Generating Novel Inhibitors through Hybridization of Known Ligands. Application to CDK2, P38, and HIV Protease. *J. Med. Chem.* **2004**, *47* (11), 2768−2775.
(35) Vidovic, D.; Muskal, S. M.; Schurer, S. C. Novel kinase inhibitors by reshuffling ligand functionalities across the human kinome. *J. Chem. Inf. Model.* **2012**, *52* (12), 3107−3115.
(36) Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Pochet, S. EA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. *J. Med. Chem.* **2005**, *48* (7), 2457−2468.
(37) Lauri, G.; Bartlett, P. A. CAVEAT: a program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8* (1), 51−66.
(38) Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in computational medicinal chemistry: Matched molecular pair analysis. *Drug Dev. Res.* **2012**, *73* (8), 518−527.
(39) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268* (5214), 1144−1149.
(40) Young, L.; Jernigan, R. L.; Covell, D. G. A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **1994**, *3* (5), 717−729.
(41) Tan, K. P.; Varadarajan, R.; Madhusudhan, M. S. DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), W242−248.
(42) Reynolds, C. A.; Wade, R. C.; Goodford, P. J. Identifying targets for bioreductive agents: using GRID to predict selective binding regions of proteins. *J. Mol. Graphics* **1989**, *7* (2), 100.
(43) Boer, D. R.; Kroon, J.; Cole, J. C.; Smith, B.; Verdonk, M. L. Superstar: comparison of CSD and PDB-based interaction fields as a basis for the prediction of protein-ligand interactions. *J. Mol. Biol.* **2001**, *312* (1), 275−287.
(44) Wang, L.; Berne, B. J.; Friesner, R. A. Ligand binding to protein-binding pockets with wet and dry regions. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (4), 1326−1330.
(45) *SZMAP*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2011.
(46) Imai, T. Roles of water in protein structure and function studied by molecular liquid theory. *Front. Biosci.* **2009**, *14*, 1387−1402.
(47) *BROOD*, OpenEye Scientific Software, Inc.: Santa Fe, NM, 2005.