HHS Public Access

Author manuscript

J Med Chem. Author manuscript; available in PMC 2018 January 12.

Published in final edited form as:

J Med Chem. 2017 January 12; 60(1): 128–145. doi:10.1021/acs.jmedchem.6b00725.

When does chemical elaboration induce a ligand to change its binding mode?

Shipra Malhotra[1,2] and **John Karanicolas**[1,2,3,*]

[1]Program in Molecular Therapeutics, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA, 19111

[2]Center for Computational Biology, University of Kansas, 2030 Becker Dr., Lawrence, KS 66045-7534

[3]Department of Molecular Biosciences, University of Kansas, 2030 Becker Dr., Lawrence, KS 66045-7534

Abstract

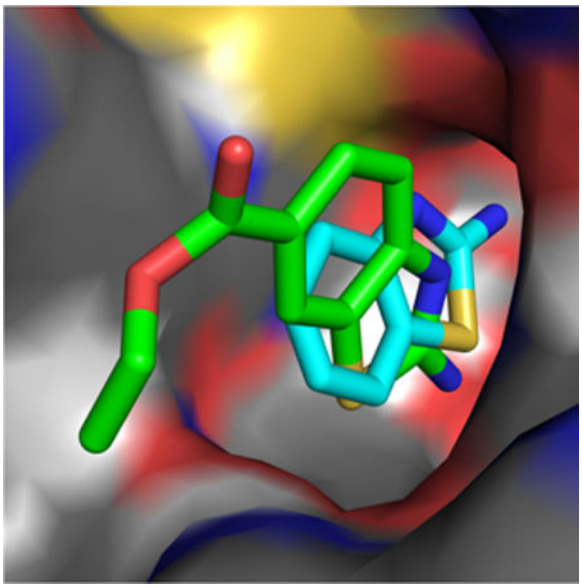
Traditional hit-to-lead optimization assumes that upon elaboration of chemical structure, the ligand retains its binding mode relative to the receptor. Here, we build a large-scale collection of related ligand pairs solved in complex with the same protein partner: we find that for 41 of 297 pairs (14%), the binding mode changes upon elaboration of the smaller ligand. While certain ligand physiochemical properties predispose changes in binding mode, particularly those properties that define fragments, simple structure-based modeling proves far more effective for identifying substitutions that alter the binding mode. Some ligand pairs change binding mode because the added substituent would irreconcilably conflict with the receptor in the original pose, whereas others change because the added substituent enables new, stronger interactions that are available only in a different pose. Scaffolds that can engage their target using alternate poses may enable productive structure-based optimization along multiple divergent pathways.

TOC image

*To whom correspondence should be addressed. john.karanicolas@fccc.edu, 215-728-7067.

Supporting Information Availability

Supporting Information. Figures showing quantitative criteria for alternate binding modes (Figure S1), plots analysis of several other properties with respect to whether binding modes are charged or retained (Figures S2, S4, S5, and S6), figures showing representative electron density to confirm altered binding modes (Figures S3 and S8), ROC plots for additional features included in this analysis (Figure S7). The complete set of ligand pairs compiled here, along with quantitative properties used in analysis (Dataset_S1.xlsx).



Introduction

Elaborating an initial hit compound to improve its biological activity is a fundamental goal of medicinal chemistry. In building up structure-activity relationships (SAR), one compiles information on how substitutions at different positions of a molecule affect activity [1]. By collecting together the optimal substituents at each available position, one expects to maximize the activity that can be achieved from a given chemical scaffold. This approach, however, relies upon an important implicit assumption: that the binding mode (the position and orientation of the ligand with respect to the receptor) is conserved across each of these individual representative compounds. The ability to explain the effect of individual substitutions solely through changes in interactions from the altered chemical moiety – a simple framework of functional group additivity – will clearly work only if the interactions separate from the substitutions are preserved.

Directly testing this pillar of medicinal chemistry requires determination of crystal structures of multiple related compounds in a chemical series, each in complex with their protein target. One such study has been carried out retrospectively by decomposing a natural product cyclopentapeptide, argifin, that inhibits a chitinase: upon trimming the starting inhibitor to a linear tetrapeptide, then a tripeptide, then a dipeptide, monopeptide, and finally a single sidechain, the authors showed that the binding mode used to recognize key interacting groups on the enzyme was conserved at every step [2]. An analogous study has also been carried out using substrates of thymidylate synthase, by sequentially removing pieces from its natural substrate dUMP. Here again, a series of crystal structures showed that the location and orientation of fragments drawn from dUMP were nearly identical to that of the corresponding groups in the complete ligand [3]. The Nutlin series that inhibits the MDM2/p53 interaction was also decomposed into its component fragments, and these were shown to retain detectable activity [4] – once again implying that the Nutlin molecule could, in principle, have been designed from these fragments.

This assumption has also been challenged, however, by other studies carrying out similar decompositions. A known β-lactamase was broken into two parts, each corresponding to half of the starting compound. Remarkably, crystal structures showed that *neither* of these two fragments engaged the receptor using the same interactions as the parent compound [5]. Similar observations by NMR have been reported for nine inhibitors of the Bcl-x_L protein-protein interaction, further noting that even the *location* at which deconstructed ligand fragments engage their receptor may not be conserved [6]. Motivation for these two studies stemmed primarily from the growing popularity of fragment-based drug discovery [7], prompting the authors to ask – retrospectively – whether these particular "mature" inhibitors could have been derived by linking, merging, or growing their constituent fragments. The surprising behavior of the fragments in this study provided a cautionary note when using structural approaches to rationally elaborate fragments, and underscored the need to confirm via crystallography or NMR that each ligand's binding mode is conserved over the course of optimization [8, 9]. In contrast, a retrospective analysis of 39 Astex fragments that were ultimately advanced into leads showed that these inevitably preserved their original binding modes, with the shared substructure changing by less than 1.5 Å RMSD in all cases [7].

Here, we explore the frequency at which the position and/or orientation of a bound ligand changes upon chemical elaboration. By carrying out a large-scale survey of available crystal structures, we have compiled a diverse set of paired ligands: in each case the smaller ligand is a substructure of the larger ligand, and in each case the two ligands have been independently solved in complex with the same protein structure. While the smaller of the two ligands did not (in most cases) serve as a starting point for design of the larger ligand, these pairs nonetheless represent examples in which the smaller ligand *could* have feasibly been optimized to yield the larger ligand. As described below, this set provides a means to ask how often the binding mode is expected to change upon elaboration of chemical structure, and what types of protein-ligand complexes are most likely to exhibit this behavior.

Results

Starting from the complete set of crystal structures in the Protein Data Bank (PDB), we grouped together crystal structures in which a given protein was separately solved in complex with multiple different ligands. For each protein, we then extracted any pairs of ligands in which the smaller compound corresponded to a chemical substructure of the larger compound, with a size increase typical of chemical elaboration in the course of hit-to-lead optimization. We also filtered such that no compound was used as the "smaller" ligand more than once (in other words, we did not allow multiple "derivatives" from a single "parent compound"). Ultimately this approach – described fully in *Methods* – produced a non-redundant set of 297 pairs of crystal structures with related ligands.

These pairs of structures represent a collection of examples for which a given protein has been solved in complex with some ligand, and this protein has also been separately solved in complex with a larger ligand that elaborates upon the first ligand. Again, we must point out that the larger ligand was not necessarily identified by rationally designing derivatives of the smaller ligand: indeed, in many examples this was not the case. In addition to examples of

ligands that *were* designed in this manner, our set includes pairs of synthetic analogs identified fortuitously, synthetic compounds paired with the natural endogenous ligands they mimic, and even pairs of endogenous ligands that naturally happen to meet the criteria laid out above. Gratifyingly, the set of paired structures resulting from this approach included the β-lactamase inhibitor and its "deconstruction" fragment described earlier [5] that motivated our study.

We have made this complete set of paired structures is available to the broader community for further study, as Supporting Information (Dataset S1).

Alternate binding modes are surprisingly common

To determine whether the binding mode of the smaller ligand was preserved by the larger ligand, we began by carrying out a global structural alignment of two protein conformations using TM-align [10]. Given that the smaller ligand is a chemical substructure of the larger ligand, we then examined the extent to which the region occupied by the larger ligand subsumes the volume occupied by the smaller ligand, in their bound crystal structures. If indeed the bound pose is preserved, the shared parts of the chemical scaffold will be superposed in the binding site, and the smaller ligand's volume will be fully covered by that of the larger ligand. In this scenario, the additional moieties that distinguish the larger ligand will extend into new regions of the binding site (presumably making additional interactions with the protein).

We used the ROCS software [11, 12] to compute the fraction of the smaller ligand's volume that is included within the volume of the larger ligand, fixing the relative position of the two ligands observed in the aligned protein structures. In addition we computed a combined overlap score ("COS") (see *Methods*), that considers not only the extent to which the larger ligand subsumes the volume of the smaller ligand, but also the extent to which the larger ligand recapitulates placement of chemical types (e.g. hydrogen bond donors/acceptors) from the smaller ligand.

As expected, in most cases the larger ligand indeed covers almost all of the volume occupied by the smaller ligand, and in most cases the larger ligand also recapitulates the positioning of equivalent chemical groups in the smaller ligand: these correspond to overlap scores close to 1 (Figure 1a). In quite a number of cases however, we find remarkably low overlap scores; this holds when considering only volume overlap, and also when considering "combined" (volume and chemotypes) overlap.

Noting that our definition of chemical substructures did not require that the larger ligand include absolutely all features of the smaller ligand, we next examined whether the lack of overlap typically occurred for pairs in which the smaller ligand was not a perfect chemical substructure of the larger ligand. Indeed, overlap scores are very slightly diminished if the match of chemical structures is imperfect (Figure S1), but this effect is small. By visual inspection of individual cases (described below) and by comparison with these histograms, we determined a COS cutoff value such that scores less than this cutoff corresponded unambiguously with a dramatic change in binding mode between the larger and smaller ligands. In light of the relationship between COS and chemical substructure scores, different

COS cutoff values were used depending on the chemical substructure score (Figure S1) (see *Methods*).

Using these conservative cutoff values we found that 41 of the 297 ligand pairs (~14%) were marked as cases in which the larger ligand's binding mode had dramatically changed relative to that of the smaller ligand. In light of this extraordinarily high number of altered binding modes, we manually confirmed every one of these 41 cases by visual inspection of the complexes; no false positives were present in this set. In this first comparison, we compared the pose of the smaller ligand to a single (arbitrary) larger ligand. We next asked what fraction of the smaller ligands had a larger ligand in an alternate binding mode if we instead searched through *all* of the partners for the given smaller ligand; we found that a larger ligand in an alternate binding mode could be identified for ~15% of the smaller ligands in our set.

In light of recent studies highlighting examples of incorrectly refined ligand geometries in published crystal structures [13–15], we recognize the importance of excluding the possibility that the ligand pairs we identified were not simply a collection of errors in crystal structures and/or differences in crystallographic conditions. As a first test, we asked whether ligand pairs with altered binding modes were observed more frequently in crystal structures with poor resolution, or crystal structures with poor R_{free}; this proved not to be the case (Figure S2). To further rule out errors from placing the ligand within crystallographic electron density, we only used structures in which the electron density was available and matches the binding mode unambiguously (Figure S3), and for which crystal contacts were unlikely to have affected the binding mode (see *Methods*). Finally, for the cases in which the ligand's binding mode had changed, we examined the pH at which the structures were solved. The pH was unchanged in 17 of these 41 structures, and the ligand protonation state was predicted not to have changed in the other 24 (see *Methods*). Thus, we are confident that none of the examples of pairs with altered binding modes included in our set are due to errors or experimental artifacts.

Below we will present two representative examples of ligands that adopt a new binding mode upon chemical elaboration, as identified in this set. In these examples, the crystal structures exhibiting variation in binding mode for a given protein were solved by the same research group; consequently, these two groups each recognized that a distinct binding mode had emerged. Throughout the rest of this study we will also present further examples to illustrate specific points, mostly drawing from new examples of ligand pairs with alternate binding modes that were identified in our set.

As a first example, we selected a ligand pair in which the smaller ligand was a perfect substructure of the larger ligand (i.e. chemical substructure score = 1). Indeed, we identified the highest fraction of ligands with altered binding modes from among the ligand pairs with perfect chemical substructural matches (Figure S1), since this allowed the most stringent COS cutoff to be used (see *Methods*). In this first representative example (Figure 1b) the smaller ligand is hydroquinone (benzene-1,4-diol), and the larger ligand simply adds to this a carboxylic acid (2,5-dihydroxybenzoic acid). Both compounds inhibit carbonic anhydrase, and the structures of each have been separately solved in complex with this enzyme [16].

The carbonic anhydrase active site includes a catalytic Zn(II) ion coordinated by three histidine sidechains and an activated hydroxide ion. Each of these two inhibitors does not bind directly to the zinc ion, but rather each one forms a hydrogen bond to the active site water molecule, and in doing so occludes the rest of the active site [16]. Despite this similarity, however, the binding modes of these two ligands are notably distinct (Figure 1b). While the protein active site does not undergo extensive reorganization, these two ligands engage their shared target through completely different contacts. The smaller ligand (hydroquinone) uses a hydroxyl group to form a hydrogen bond to the active site water, allowing the aromatic ring to be packed into a shallow hydrophobic surface cleft. In contrast, the larger ligand (2,5-dihydroxybenzoic acid) uses the carboxylic acid to engage the active site water and also a nearby threonine sidechain. As a result, the aromatic ring is less well packed, and one of the hydroxyl groups faces into a hydrophobic region of the active site. Interestingly, given these distinct modes of interaction, it is the smaller ligand that exhibits more potent inhibition than the larger ligand (K_i = 90 nM for hydroquinone [17], versus IC_{50} = 5 mM for 2,5-dihydroxybenzoic acid [16]).

As a second representative example, we present a series of peroxisome proliferator-activated receptor γ (PPARγ) structures solved with various ligands (Figure 1c). This nuclear receptor appeared in our set as three distinct ligand pairs, corresponding to three different “smaller” ligands that could each be elaborated to yield indomethacin as the “larger” ligand. Two of the smaller ligands are perfect substructures of indomethacin (5-hydroxyindole acetate and 5-methoxyindole acetate), while the other (serotonin) is only a near match because it contains a primary amine not found in indomethacin. In this case the larger ligand, indomethacin, is a synthetic analog of these three smaller natural metabolites that serve as full/partial agonists of PPARγ – and remarkably, each of these four ligands engage the protein in different orientations.

Comparison of these crystal structures highlights the importance of the binding mode for biological activity. Remarkably, each of these four ligands confers different activity in cells: their degree of agonism differs, as does their preference for binding a second simultaneous ligand (different fatty-acids) [18]. Substantial efforts had been directed towards establishing a structure-function relationship for receptor activation by various ligands, but these inevitably proved challenging to interpret. This series of crystal structures provided a clear explanation for this behavior, and – in retrospect – demonstrated that it was the lack of conservation of ligand binding mode that confounded earlier efforts to derive a simple structure-function relationship [18]. Here, the changes in binding mode lead to altered biological activity that cannot be understood or predicted solely on the basis of changes in binding affinity – which in turn makes it extremely difficult to rationally design new agonist ligands.

We selected these two examples of alternate binding modes partly because there is no accompanying conformational change of the protein, making it very straightforward to visually recognize the changes in the ligand’s pose. In addition to these two, our dataset contains examples of ligand pairs with distinct binding modes from many medicinal chemistry campaigns, with target classes that include kinases, phosphatases, proteases, and β-lactamases. We next used this set to explore what physiochemical properties of the ligand,

the protein-ligand complex, or the protein surface might suggest that a given bound ligand may be predisposed to change its pose upon chemical elaboration.

Chemical properties that correlate with alternate binding modes

We began from the hypothesis that the initial (smaller) ligand might be more likely to change its binding mode upon chemical elaboration if it does not initially engage its target with very high affinity. This can be rationalized by considering that a weakly-binding ligand may exhibit only a slight preference for the observed binding mode, and thus be pre-disposed to adopt an alternate binding mode in response to even a relatively small perturbation (adding one or more substituents). In contrast, a tightly-binding ligand may be "locked" into place, and thus be more likely to accommodate structural changes even if they are suboptimal in the context of this complex.

To test this hypothesis, we plot the distribution of potency for the smaller ligand across our collection of paired complexes, separating examples of pairs that changed binding mode from those that did not (Figure 2a). We note that "potency" in this context can be either K_d, K_i, or IC_{50}; to compare them in this manner, IC_{50} values were first scaled as described elsewhere [19] (see *Methods*). Comparing these distributions, we find that indeed the median potency for pairs that retain their binding mode is stronger than the median value for pairs that did change binding mode upon elaboration, and that this difference between the distributions is statistically significant ($p < 0.005$) (see *Methods*). We also note that assay differences can lead to large disparities in IC_{50} values reported for the same compound, and this may contribute noise to our dataset: were it not for this source of error, the difference between the two distributions may be even more striking.

Given that binding affinity typically becomes stronger with increasing molecular weight [20], one might further expect that smaller ligands would be more likely to change binding mode upon chemical elaboration. Indeed, upon collecting the molecular weight of the smaller ligand in each pair, we find that the median value of ligands that changed binding mode was smaller than the median value for ligands in which the binding mode was preserved (Figure 2b), and that the difference between the distributions was again statistically significant ($p < 4\times10^{-4}$). Given the natural relationship between molecular weight and number of atoms, it is unsurprising that we also observe a difference when comparing distributions of the number of (non-hydrogen) atoms ($p < 4\times10^{-4}$) (Figure S4a). Interestingly, however, the change in potency (Figure S4b) or the change in ligand efficiency (Figure S4c) between the smaller and larger ligand does not yield a statistically significant difference in the distributions of the two data sets, and neither does the initial ligand efficiency of the smaller ligand (Figure S4d).

Next, we hypothesized that polar ligands would be less likely to preserve their binding modes than hydrophobic ligands: because hydrogen bonding requires more precise geometry than non-polar interactions, binding modes that rely on hydrogen bonding may not be sufficiently robust to allow slight perturbations needed to accommodate the larger ligand. Using the computed octanol-water partition coefficient (clogP) as a measure of hydrophobicity, we indeed find a difference between median clogP of ligands that adopted a

new binding mode versus those that did not (Figure 2c); the difference between these distributions was again statistically significant ($p < 0.02$).

Finally, we anticipated that a ligand's flexibility might also contribute to it's potential for adopting a new binding mode. If a functional group were added to a ligand at a position that is incompatible with the existing binding mode, the opportunity to slightly vary the ligand's internal degrees of freedom may allow this new substituent to be accommodated without dramatically changing the binding mode. Because such reorganization may not be possible for a purely rigid ligand, in contrast, the substitution might prevent binding altogether, or – if binding still takes place – require adoption of a new binding mode. We therefore compared the distributions of the number of rotatable bonds in ligands that change binding mode versus those that do not across our complete set of ligand pairs (Figure 2d). We indeed find fewer rotatable bonds in the ligands that change binding mode, however the difference in their distributions did not achieve statistical significance ($p < 0.1$).

Collectively, then, we have demonstrated here that several properties of the smaller ligand are correlated with the likelihood that the larger ligand will not preserve its interactions: weak binding, low molecular weight, polar, and rigid compounds appear most likely to change binding mode upon elaboration.

Properties of the initial complex that correlate with changing binding mode

Beyond simply a ligand's chemical structure, we also hypothesized that its interactions with the receptor may contribute to whether alternate binding modes might be adopted. In particular, we anticipated that ligands bound to deep pockets might have fewer opportunities to explore other poses, relative to ligands that occupy shallow surface grooves. One representative example from our survey of the PDB is a pair of isoquinoline-1,3,4-trione derivatives [21], which bind superficially to the surface of caspase-3 (Figure 3a).

As a starting point, we computed the solvent accessible surface area (SASA) buried in each protein-ligand complex. Indeed, the median SASA buried by ligands that change their binding mode upon chemical elaboration is less than those for which the binding mode is preserved ($p < 9\times10^{-4}$) (Figure 3b). We also computed the fraction of ligand's surface area that remains exposed upon complexation (θ_{lig})[22]; surprisingly, ligands that change binding mode are *not* systematically bound using shallower binding modes (Figure 3c). The fact that the extent to which the ligand is buried is not correlated with its propensity to change binding mode suggests that the observation of fewer altered binding mode for ligands with high SASA may be an indirect effect: high SASA is naturally correlated with larger and more potent ligands [23], and we showed earlier that each of these make the binding mode less likely to change. We will return to the complication of correlations between features later in our analysis.

Given that polar compounds are more likely to change binding mode, we next counted the number of intermolecular hydrogen bonds in each complex. Although the median value for this discrete variable is the same in both sets, the distributions are not the same (Figure 3d): there are fewer intermolecular hydrogen bonds involving the ligands that change binding mode, and this difference in the distributions is statistically significant ($p < 0.02$). Here

again, the properties of a given compound cannot be assumed to be independent from the properties of the complex: while polar compounds are more likely to change binding mode, this may be especially true if the hydrogen bonding potential of the ligand is not fully satisfied in the original binding mode. Further, given that hydrogen bonds generally stabilize protein-ligand complexes, the observation that complexes with more hydrogen bonds are less likely to change binding mode might be explained – at least partially – by our earlier observation that tighter binding complexes are less likely to change binding mode. Accordingly, in this vein, we also note that there are examples in which the binding mode changes, but it does so in a manner that preserves certain key hydrogen bonds (Figure 3e).

Very recently, the hypothesis was put forward that a substructure pulled from a larger ligand would retain its original position and orientation if the fragment is located at a "binding energy hot spot" [24]. To test this, the authors developed a "fraction overlap score" that quantifies how much of the fragment falls within the primary hotspot region that is determined from the protein structure using computational solvent mapping [25]. By examining eight classic ligand-deconstruction testcases, the authors found that indeed this "FO score" distinguishes a series of fragments that do not preserve the parent compound's binding mode (the β-lactamase case introduced earlier [5]) from examples in other systems for which the parent compound's binding mode is conserved by the fragment [24].

To explore the generality of this observation beyond these eight testcases, we sought to compute the FO score for each of the ligand pairs in our dataset, exactly as described by these authors [25] (see *Methods*). However, we found that in about a quarter of the cases examined, the primary hotspot did not coincide with the ligand binding site, which in turn precluded a meaningful FO score from being calculated. To address this, we instead defined the primary hotspot as the largest cluster that overlaps with the larger ligand in our pair, rather than simply the largest cluster anywhere on the protein. This allowed us to calculate the FO score for 293 ligand pairs (41 changed binding mode and 252 did not; in 4 cases computational solvent mapping did not yield any probes near the larger ligand).

Our much larger set now allows for a more rigorous evaluation of the FO score, and indeed confirms a statistically significant difference ($p < 5\times10^{-4}$) in the distribution of FO scores for ligand pairs in which the binding mode is preserved, versus those for which the ligand adopts an alternate pose (Figure 3f). Nonetheless, we note that the FO score is computed from the crystal structure of the larger ligand, and thus cannot be applied prospectively to assess the effect of elaborating the smaller ligand: its intended use is rather to determine whether *reducing* the size of a ligand might alter its binding mode.

In a related vein, a recent retrospective analysis of Astex screening campaigns revealed three cases in which elaboration led to new binding modes, and in each case this change was accompanied by a corresponding conformational change of the protein [7]. Each of the specific examples of alternate binding modes we have presented thus far include minimal changes to the protein's binding site, since these can make it more difficult to visually compare the two poses. However, our set does include examples of alternate binding modes that are accompanied by protein conformational changes, such as the tyrosine phosphatase 1B active site (Figure 3g). Overall, and unsurprisingly, the RMSD of the protein's binding

site when comparing the pair of ligand-bound structures is typically higher if the two ligands engage the protein with a different binding mode (Figure 3h), with a statistically significant difference between the distributions ($p < 0.03$). To examine whether the flexibility of the binding site was evident from the structure of the smaller ligand we compared the crystallographic B-factors of the two sets: however, this revealed no difference (Figure S5). It is important to remember that protein binding sites are malleable and may adapt to bind different ligands – in the Astex survey [7], 17 of 25 cases included a protein conformational change greater than 1 Å RMSD, even when the ligand pose did not change. However, it is not clear at this point how one might anticipate such changes ahead of time, to predict whether a given ligand will change its binding mode upon chemical elaboration.

Properties of the initial binding pocket that correlate with changing binding mode

The physicochemical properties of a ligand (e.g. size and hydrophobicity) are somewhat predisposed by the physicochemical properties of the binding pocket on the protein surface. Having determined that certain ligand properties are correlated with increasing propensity for changing binding mode upon elaboration, we next asked whether analogous features of the protein surface pocket would be similarly predictive.

Indeed, we find that cases in which the initial ligand occupies a smaller pocket volume are more likely to change binding mode upon chemical elaboration (Figure 4), and that this difference is statistically significant ($p < 9\times10^{-5}$). This is unsurprising, given that small pockets can typically only accommodate small and weak-binding ligands, and we have already shown that these are more likely to change binding mode (Figure 2a and 2b).

Intriguingly however, we do not find analogous differences between the distributions for changed versus preserved binding modes when considering frequency of polar residues in the binding pocket, frequency of aromatic residues in the binding pocket, binding pocket hydrophobicity, or binding site “druggability” [26, 27]. Thus, with the exception of pocket size, it appears that it is primarily the physicochemical properties and activity of the ligand, rather than the binding site properties, that dictate whether the binding mode is likely to be preserved.

Analysis of chemical substitutions in the structure of the complex

Our analysis thus far has focused on details of the initial protein-ligand complex to explore how often the binding mode is preserved upon chemical elaboration: thus far we have not yet considered the location or identity of the substituent(s) that are to be added. Clearly we expect that this will be important: one would expect that building on a new group that extends into solvent may not alter the binding mode, whereas adding in a direction that faces into the protein may cause a steric clash that forces a new binding mode to be adopted.

To directly address this question, we developed a new tool for rapidly probing whether the larger ligand could be accommodated in the protein without changing binding mode. Briefly, we align the shared substructure from the larger ligand onto the corresponding region in the smaller ligand, using the crystal structure of the smaller ligand’s complex: this provides us with an initial model of the large ligand’s complex, built in a manner that completely preserves the binding mode of the smaller ligand. We then carry out energy minimization of

this model, and monitor the RMSD difference of the large ligand relative to the initial pose. If the ligand can be accommodated in the model with only minor rearrangement, the RMSD difference will be small; if, however, there is egregious incompatibility between the protein and the large ligand in this binding mode, we will observe a much larger RMSD difference. In our analysis below, we will refer to this RMSD difference as "*r*msd after *m*inimization of the *a*ligned *c*omplex" (RMAC) (see *Methods*).

We computed RMAC for the ligand pairs in our set, and as anticipated we found that RMAC values are typically higher for ligand pairs in which the binding mode is not preserved (Figure 5). This difference is statistically significant, with much lower p-value than any other property we have examined thus far ($p < 6\times10^{-7}$). It natural to expect that the specific chemical substitution plays a role in dictating whether the binding mode will change, and thus perhaps unsurprising that this property exhibited such a large difference between ligand pairs that change binding mode versus those that do not – even given the simplicity of the approach.

Predicting the presence of alternate binding modes based on these properties

Thus far, we have evaluated various properties to determine which ones correlate with ligand pairs that change binding mode upon chemical elaboration from those that do not; we have summarized these properties, and the statistical significance of the observed differences, in Table 1.

Applying this analysis, the predictive power of the properties considered above can best be compared using receiver operating characteristic (ROC) plots. In this case we seek to predict the value of a binary classifier – will the ligand change its binding mode upon elaboration? – using a known quantitative property. For a given property (e.g. molecular weight) at a given stringency (e.g. 250 Da), we plot the fraction of cases in our test set that would be correctly assigned as changed binding modes (true positives recovered), as a function of the fraction of preserved binding modes that would be incorrectly assigned as changed (false positives). Points on this plot corresponds to increasing the stringency at which assignments are made; for a truly random classifier, the true positives and the false positives will accumulate at an equal rate.

We have generated ROC plots (Figure 6a, Figure S7) for each of the properties described in the preceding sections. Consistent with our earlier analysis, certain properties are useful in anticipating the likelihood that a ligand will change binding mode: these include RMAC, pocket volume, molecular weight, lipophilicity, and potency. Meanwhile, other properties (such as θ_{lig}) have no predictive power at all. The area under the curve (AUC) for each property is also reported in Table 1. As expected, these are largely aligned with the p-values describing the statistical significance of the difference between distributions; the exceptions are discrete variables (such as the number of intermolecular hydrogen bonds), where there are differences between the distributions but the large number of "ties" among these values limit the predictive power of such variables.

The dependence on this single descriptor can be summarized most intuitively through logistic regression [28], since this allows estimation of a binary output (changed or preserved

binding mode) based on the value of a continuous variable. Through the resulting model, we emphasize that low RMAC values are highly indicative of ligand pairs in which the binding mode will be preserved: values below 0.65 Å, the median value for pairs that did not change binding mode in Figure 5, are found to change binding mode less than 10% of the time. In contrast, ligand pairs with RMAC values of 4 Å are 45% likely to change binding mode (Figure 6b). Molecular weight is also predictive, though less able to confidently identify ligand pairs that change binding mode: a compound of 400 Da molecular weight has a probability of about 5% of changing binding mode upon chemical elaboration; this probability increases to 17% if the starting compound is 200 Da, and to 30% if the starting compound is only 100 Da (Figure 6c). Given the correlation between molecular weight and potency, we also observe analogous behavior as a function of the smaller ligand's binding affinity (Figure 6d).

While some of the properties we consider here are trivially correlated with one another (e.g. molecular weight with number of heavyatoms), many more are known to correlate in practice (e.g. molecular weight with activity). We therefore systematically evaluated the correlation between all properties used in this study, using the Spearman correlation coefficient; even among the properties that are predictive of whether a ligand will change binding mode, many not are correlated with one another (Figure 7a). This observation suggests that by using multiple properties in tandem, further predictive power can be achieved.

To combine these properties into a more powerful tool for predicting whether a given ligand will change binding mode upon elaboration, we applied multiple logistic regression with several different combinations of these properties as inputs. For each model, we express the results as ROC plots, with the AUC value indicative of the model's ability to predict whether a given ligand pair will change binding mode (Figure 7b).

As expected, adding a highly correlated property to the molecular weight, such as number of heavyatoms or activity, does little to improve performance: the AUC value is essentially the same as that of molecular weight alone. On the other hand, including molecular weight alongside FO score or RMAC (AUC values 0.73 and 0.78, respectively) provides improved discrimination relative to FO score or RMAC alone (AUC values 0.66 and 0.74, respectively). Finally, incorporating further correlated properties into one of these models does not improve them any further: adding activity and buried SASA into the model built using RMAC and molecular weight does not provide any noticeable benefit.

Ultimately then, RMAC and molecular weight together offer the ability to make fairly accurate predictions regarding whether adding a specific new substituent will cause a ligand to change its binding pose, given the crystal structure of the initial ligand.

Discussion

Using our conservative definition, we found that 41 of the 297 ligand pairs in our set (~14%) clearly and unambiguously changed binding mode upon elaboration. However, this fraction most certainly does *not* reflect the likelihood that addition of any arbitrary substituent will

lead to a new pose: it is certainly dependent on the nature of the initial ligand, and the compatibility of the binding site to accommodate the new substituent. Below, we will consider in detail the factors that further tune the likelihood of a derivative adopting a new binding mode.

Substitutions incompatible with the original binding mode

It must be immediately noted that cases in which elaborated ligands are designed arbitrarily are increasingly rare: modern medicinal chemistry relies heavily on structural biology to help identify useful vectors at which to add substituents. In circumstances when the elaborated ligand is designed with explicit consideration of how the additional groups might interact with its receptor, the rationally designed substituents are expected to reinforce the original binding mode. Thus, in these cases we expect that the likelihood of an altered binding mode would be lower than the frequency observed across our complete set, since many examples in our test set originate from different research groups (rather than by explicit optimization of a known ligand), and thus the larger ligand was identified without knowledge of the smaller ligand's interactions.

In essence, our application of RMAC crudely mimics the human expertise typically underlying structure-based design of new analogs. Accordingly, most substitutions that preserve the original binding mode have very low RMAC values (Figure 5). This serves as validation of our simple modeling approach, by demonstrating that it usually generates quite accurate models of the larger ligand when the binding mode is unchanged. Nonetheless, ligand pairs with unchanged binding modes are assigned higher RMAC values: these correspond to failures of the modeling approach, typically arising from slight adjustments of the protein conformation that are not adequately recapitulated.

At the lowest RMAC values – cases in which the larger ligand is predicted to be highly compatible with the receptor – the extrapolated probability of a new binding mode drops below 7% (Figure 6b). Relative to the frequency of alternate binding modes in the complete set, this lower value confirms that some pairs in the complete set were elaborated in ways that simply wouldn't make sense given the structure of the smaller ligand. Such pairs – which are assigned much higher RMAC values – add substituents that produce irreconcilable conflict between the protein and the elaborated ligand. If the ligand is elaborated in this manner, there are three potential outcomes: either the ligand will no longer bind (or will bind much less potently), or it will induce a dramatic conformational change in the protein, or it must find a different pose that avoids this conflict.

In some sense, crystal structures from ligand pairs with high RMAC values are already somewhat surprising: intuitively, most such "nonsensical" substitutions would presumably lead to loss of ligand binding. Our data cannot speak to the frequency of substitutions that lead to loss of ligand binding, since we have collected data only where crystal structures of complexes are available. Inspection of pairs of crystal structures with high RMAC values confirm that these arise either because of structural rearrangements not captured by our simplistic modeling approach (e.g. large conformational changes to the protein or the ligand), or else because the ligand adopts an alternate binding mode. At the highest RMAC values, resolution via each of these two possibilities is about equally likely (Figure 6b). In

these examples, the alternate binding modes are available to the original (smaller) ligand, but are less favorable in binding free energy: they become populated only when addition of a new substituent makes the original binding mode unavailable.

Substitutions that enable new interactions

Aside from examples in which a clear structural conflict induces the binding mode to change (i.e. high RMAC), there are also a surprising number of examples that change binding mode despite having low RMAC values (Figure 5). Here, our modeling is absolutely able to build apparently-reasonable complexes of the larger ligand using the same pose as the smaller ligand, and yet the crystallographic binding mode reveals an alternate pose. Through individual inspection of these cases, we find that modeling error may have been responsible for a few – the protein was adjusted slightly to accommodate the larger ligand, and our approach may have underestimated the energetic consequences of the rearrangement. For the most part, however, these represent “opportunistic” changes in binding mode: addition of an new substituent draws the ligand into a new pose, to allow the new substituent to participate in new, favorable interactions. Certain themes emerge amongst these cases, which are best demonstrated through select examples: those drawn from ligand pairs with low RMAC values (< 0.7 Å), strongly suggesting that the larger ligand could have been accommodated without changing its binding mode.

The first theme is the addition of a substituent that naturally enables a single, very strong interaction. Indeed, two separate examples of alternate binding modes arise from adding a carboxylic acid near a metal ion. In the case of carbonic anhydrase, described briefly earlier (Figure 1b), the binding site consists of a Zn(II) ion bound by three histidine residues and a bound water molecule. The smaller ligand, hydroquinone, engages the activated water using a hydroxyl group, with very few additional contacts to the protein; a phenol-bound structure also shows a very similar bound pose [29]. The lack of additional contacts leads to dual occupancies observed for the ring, though the positioning of the hydroxyl group is preserved in both [16]. When hydroquinone is elaborated to include a carboxylic acid, however, the opportunity for a stronger interaction with the activated water molecule leads to an altered binding mode. As noted earlier, despite this new stronger interaction the potency of 2,5-dihydroxybenzoic acid is surprisingly worse than that of the hydroquinone parent [16].

A similar interaction induces the conformational rearrangement observed in a series of metabolite-inspired leukotriene A4 hydrolase (LTA4H) inhibitors. Starting from a weak 5-hydroxyindole fragment hit, linking a pyrrolidine group yielded improved activity. Subsequently replacing the pyrrolidine with a piperidine carboxylic acid moiety, however, shifted the ligand towards active site Zn(II) (Figure 8a). This allows the larger ligand to form a direct interaction between the carboxylic acid and the metal ion; however, this strong interaction again comes at the expense of overall activity, which is decreased in the elaborated ligand [30].

A second theme among these “surprising” changes (low RMAC) in binding mode is the addition of substituent that inadvertently helps optimize shape complementary for the receptor. The salicylate synthase enzyme from M. tuberculosis, MbtI, converts chorismate to salicylate through an isochorismate intermediate; this allowed design of an isochorismate

mimic to inhibit the enzyme. Elaborating the enolpyruvyl side chain with substituents ranging from a methyl group to a phenyl group all improved potency at least 10-fold; the authors' docking studies – as well as our RMAC values – suggested these could be accommodated in the original binding mode [31]. Surprisingly though, crystal structures of these derivatives revealed that the binding mode had changed (Figure 8b). While the original unsubstituted isochorismate scaffold binds in a manner analogous to the substrate, in retrospect there remains a buried cavity that was present in the crystal structure of this complex. In the alternate binding mode, the additional substituents are well-positioned to fill this cavity, leading to improved packing overall [32]. Importantly, the ligand's pseudosymmetry, arising from carboxylic acids at either end of the molecule, may also facilitate the altered binding mode: upon elaboration the ligand flips over, such that the interactions of the two carboxylic acids are nearly perfectly exchanged with one another. Thus, we propose that the pseudosymmetric ligand already had similar binding free energy in both orientations, and the new substituents preferentially stabilized the "flipped" orientation.

A pre-formed cavity is also evident in the structure of 2-aminobenzothiazole bound to urokinase [33]. Here, the receptor presents a pair of nearly identical small pockets at the base of a deep, narrow cleft in the active site; the initial fragment engages one of these small pockets, but not the other. The modeling that underlies RMAC calculations confirms that elaboration with an ethyl ester at the other side of the ligand could be accommodated through very minor changes to the protein surface; instead, however, the crystal structure of this derivative shows that the compound instead shifts to fill the other binding site pocket (Figure 8c). Here again, we propose that the two 2-aminobenzothiazole orientations are close in energy to one another; thus, the shift in conformation may be driven by the opportunity for the new substituent to interact with the shallow surface groove that hosted a crystallographic sulfate ion in the original crystal structure.

Throughout these examples, the interactions of the new substituent allow rationalization of the energetic benefits afforded by the alternate binding mode. However, the structural basis for adopting an alternate binding mode need not necessarily be so clear. This point is well illustrated through a series of 5,6-bicyclic heterocyclic inhibitors of cyclin-dependent kinase 2 (pyrazolopyrimidines and imidazopyrazines) [34]. Each of these compounds engages the kinase at the ATP binding site, through hydrogen bonds to the protein backbone in the hinge region. Among 11 structures reported in this study, 10 share a binding mode previously observed in other pyrazolopyrimidines/purine-based cores: one compound, however, adopted a completely different binding mode, with unambiguous crystallographic density supporting this binding mode (Figure S8). Intriguingly, comparison of the chemical structure of this compound to its various analogs shows that there is not a single substitution that determines the binding mode. Relative to the compound with a distinct binding mode, there are individual analogs that share either its core, or its various substituents: and yet, none of these analogs adopt the distinct new binding mode (Figure 8d). Ultimately, the authors of this study conclude that it is a precise combination of each of these contributions – the imidazopyrazine core, with a fluorophenyl substituent, and with another position that must remain unsubstituted – that collectively induces the binding mode to change.

Fragments adopting alternate binding modes

The past two decades have been marked by a broad and enthusiastic adoption of fragment-based drug discovery [35]. This technique seeks to sample chemical space more efficiently, by elaborating low-molecular weight ligands (~150 Da) identified typically through biophysical screens designed to detect very weak binding [36]. Indeed, whereas custom fragment libraries were original constructed to obey a "rule of three" (less than 300 Da, clogP less than 3, and no more than 3 hydrogen-bond donors/acceptors) [37], the same authors subsequently refined their recommendation to compounds under 230 Da [38]. One might expect that prioritizing fragment hits on the basis of ligand efficiency may give a preference for compounds that are "locked in" with respect to their binding modes; however, our data do *not* show that fragments with higher ligand efficiency are statistically more likely to retain their binding mode upon elaboration (Figure 9a). Rather, on the basis of the analysis presented here, fragments are precisely the types of compounds that are overall most likely to adopt new binding modes upon chemical elaboration.

To highlight this point, we compiled from our test set the 73 smaller ligands that are rule-of-three compliant: the larger ligand adopts an alternate binding mode in 23% of these cases (17 changed versus 56 unchanged). That said, a key tenet of fragment-based drug discovery is the deployment of structural biology to guide optimization [36]; when analogs are designed with knowledge of the fragment's binding mode, they are often intended to probe specific vectors that are available based on the initial binding mode, and structural insights were certainly used in designing many of the larger ligands derived from fragments. Among the 39 fragment-to-lead pairs described in Astex's study, none changed their binding modes upon elaboration; however, all substitutions were carefully designed to stabilize the pose observed for the fragment hit [7]. We have shown that growing the ligand in "reasonable" directions based on structural considerations (i.e. low RMAC value) greatly reduces the likelihood of alternate binding modes; therefore, the chances of identifying a new pose by growing a fragment in a single *arbitrary* way is presumably much greater than 23%.

One potential limitation of growing fragments exclusively in directions expected to reinforce the existing binding mode is that this design may needlessly limit the space of analogs that could otherwise be productively explored; we will illustrate this point through a pair of Hsp90 inhibitors acting at the ATP binding site. A fragment screen carried out at Astex yielded four validated hits; impressively, one these was advanced into a compound more than a million times more potent than the fragment (the K_d dropped from 0.8 mM to 0.5 pM) [39]. Moreover, the binding mode was essentially identical to the initial fragment hit (Figure 9b) – as we have now come to expect from careful structure-guided changes. While the hydrogen bonding interactions of these compounds do not mimic those of ATP, the plane of the fragment's ring does overlap with the adenine moiety in an ADP-bound structure.

In parallel, a different group separately identified a tropane from a high-throughput screen of 4.1 million compounds [40], and ultimately advanced this to a derivative yielding tumor regression in a mouse xenograft model. There is structural similarity between this HTS hit and the previous fragment hit, and indeed this ring once again binds at the location of ADP's adenine moiety. Superposition to the fragment-bound structure, however, reveals that the HTS hit binds in an orientation rotated 90° relative to the fragment (Figure 9b). While it

appears on the basis of chemical structure that the HTS hit could have resulted from optimization of the fragment, this would be exceedingly unlikely on the basis of iterative structure-guided design: rational substituents intended to reinforce the fragment's binding mode are unlikely to enable discovery of alternate ways in which the fragment might engage the receptor.

Naturally, a fragment captured in a different bound orientation will inspire completely divergent pathways for structure-guided optimization: in essence, the same scaffold can have the "value" of more than one starting point if it can be used in more than one way. Potential alternate binding modes for a scaffold might, in principle, be identified by a combination of docking and experiments such as STD-NMR, which were recently used to identify different binding modes from within across a family of analogous fragments [41]. Alternatively, a series of crystal structures of the fragment core harboring substituents at different positions might also be used to find potential binding modes: this strategy was recently used to explore a fragment against TGF-β receptor type-1. In this study, crystal structures of five different fragments were solved, yielding three distinct binding modes with different patterns of interactions by which the same core could engage the hinge region of this kinase [42] (Figure 9c). Remarkably, it appears that even changing the crystallization conditions can sometimes reveal new ways in which a fragment can be used. A triazine fragment was co-crystallized with Hsp90, and found to mimic the interactions observed in an ADP-bound structure; however, the same compound yielded a different binding mode when it was instead soaked into Hsp90 crystals [43] (Figure 9d).

Conclusions

By building a large-scale collection of ligand pairs solved in complex with the same protein partner, and in which the larger ligand could have arisen by elaboration of the smaller ligand, we have laid the groundwork for better understanding – and predicting – the pose that a given ligand will adopt. Ultimately, the binding mode must reflect the lowest free energy state for a particular ligand. As the ligand is chemically modified, or as the conditions change, the relative free energies of each binding mode change with respect to one another.

In some cases, chemical substitutions lead to clear incompatibility between the ligand and the receptor: in these cases, major conformational reorganization of the protein, ligand, or both is required for the ligand to bind to the receptor at all. In other cases a specific substitution may enable formation of a new, strong interaction, such as those involving metal ions. Alternatively, a specific substitution may inadvertently stabilize an alternate pose; this is most common among pseudosymmetric ligands, because the alternate pose can mimic many of the interactions in the original pose.

Structure-based medicinal chemistry entails carefully selecting new compounds expected to improve interactions with the receptor; thus the optimization trajectory is strongly reliant on the binding mode. For fragments that are capable of adopting alternate binding modes, prosecuting each pose in a parallel and orthogonal manner may allow effective exploration into new chemical space.

Experimental Section

Building the set of complexes with paired ligands

We began with the contents of the PDBbind database that include corresponding experimental measurements of binding affinity (currently 10,776 protein complexes) [44, 45]. Structures were then removed if the ligand did not have between 6 and 55 non-hydrogen atoms, or if the ligand was a common crystallographic additive or detergent. We also removed all NMR structures. We then identified the Uniprot ID [46] for the protein component(s) in each structure, and grouped together all crystal structures with the same Uniprot ID. For all the complexes of a given protein (i.e. a unique Uniprot ID), we collected all pairwise combination of ligands in which the larger of the two has molecular weight at least 1.3 times that of the smaller; this arbitrary cutoff was designed to reflect a size increase typical of chemical elaboration. Finally, any redundant pairs were removed.

To determine cases in which one ligand could feasibly have been used as a starting point to develop the other ligand, we identified those pairs in which the smaller ligand is a chemical substructure of the larger ligand. We did so by first generating a “fingerprint” (an ordered binary string of chemical moieties that are either present or absent) for each ligand, using OpenBabel [47]. Given a larger ligand “A” and a smaller ligand “B”, we counted the number of “on” bits in B’s fingerprint that were also “on” in A’s fingerprint: these correspond to shared chemical moieties. We then normalized this to the total number of “on” bits in B, yielding a “chemical substructure score” as follows:

$$chemical\ substructure\ score = \frac{A \cap B}{B} \quad \text{(Eqn. 1)}$$

We first sought to use this score to eliminate any clear derivatives of amino acids, sugars and nucleoside analogs from the set. We generated fingerprints from 14 amino acids (Asp/Glu/Phe/His/Ile/Lys/Leu/Met/Asn/Pro/Gln/Arg/Trp/Tyr), 5 nucleobases (adenine/cytosine/thymine/guanine/uracil), and 5 representative monosaccharides (glucose/fructose/ribose/mannose/galactose). Any compounds with substructure score above 0.95 were removed from our set. We excluded the other 6 amino acids (Ala/Cys/Gly/Ser/Thr/Val) from this step because they did not contain sufficiently descriptive fingerprints to allow derivatives to be identified in this manner (e.g. there are many unrelated compounds include all of the functional groups present on alanine).

We note that this a definition of substructures does not *guarantee* that B is a chemical substructure of A; it simply reports on the number of B’s moieties that are also present in A, but does not ensure identical connectivity. By examination of a number of chemical structures with high substructure scores, we determined that B was almost inevitably a substructure of A if the score was above 0.9. To test this cutoff value, we later evaluated each pair using the MCS (Maximum Common Substructure) tool implemented in ChemAxon [48]. The MCS is defined as the largest subgraph shared by graphs representing the chemical structures of the large and small ligands; similarity between the graph of the smaller ligand and the shared subgraph implies that the smaller ligand is indeed a

substructure of the larger ligand. Using this approach, we confirmed the suitability of a 0.9 for “chemical substructure score” as a cutoff to select pairs of ligands.

Filtering using this cutoff value led to a set of 1454 pairs of ligands. Of these, only 383 unique “smaller” ligands were reflected: there were examples for which crystal structures of multiple large ligands had be solved in complex with a given protein, and each these ligands could have derived from a single smaller ligand. In such cases we retained only a single pair, by keeping only a single (randomly-selected) representative from the larger ligands. While each “smaller” ligand was only paired with a single “larger” ligand, however, we did not require that a “larger” ligand be used only once. Thus, our set does include cases in which more than one small ligand could be elaborated to produce the same larger ligand.

The chemical structures for all 383 ligand pairs were each manually examined, and a single “false positive” was identified in the set: a case in which the larger ligand indeed contained all the functional groups of the smaller ligand, but on a completely different chemical scaffold. This example was removed from further consideration, leaving 382 paired PDB structures.

To ensure that no false positives were included in our set due to missing or ambiguous electron density, we downloaded 2mFo-DFc electron maps in CCP4 format from the Electron Density Server [49]. For 87 of the 382 paired PDB structures, electron density data was unavailable for one of the two structures; however, in 9 of these cases we were able to identify a replacement ligand for which this data was available. We manually examined each of the resulting 304 pairs of PDB structures using PyMOL [50], to ensure that the ligand position and orientation were unambiguously determined by the electron density, and to check for crystal contacts at the ligand binding site. We removed 4 cases with ambiguous electron density, and 1 case with crystal contacts near the ligand binding site. We also removed two more pairs: one was a set of covalent inhibitors bound to a co-factor, and was a ligand with highly unusual geometry.

Next, we examined the pH at which the pairs of structures were solved, to rule out potential cases in which difference in pH may have led to an observed difference in binding mode. Among the 41 cases in which the binding mode was altered, 17 cases were comprised of pairs of structures solved at the same pH. For the other 24 cases, we used the PROTOSS server [51] to determine the most probable ligand protonation/tautomerization state in the context of the protein-ligand complex: for all 24 cases, the region common to the smaller and larger ligands was identical in this regard. Thus, there is no evidence supporting a pH-driven change of binding mode for any of the examples included in our set.

This process ultimately led to a set of 297 paired PDB structures. The complete set of paired structures is included (Dataset S1).

Comparing bound poses

To compare the position and orientation of different ligands relative to the protein, we began by using TM-align to carry out a global structural alignment between the two proteins [10]. The same transformation was applied to the respective ligands, so that the ligands were

shifted into the corresponding reference frame. We manually examined the effect on the binding site produced by this alignment for each pair in our dataset, and concluded that no individual adjustment was needed in any of the cases.

The ROCS software was originally developed as a ligand-based virtual screening tool, for using a known drug lead to identify other potentially active compounds [11, 12]. Briefly, the underlying algorithm uses a summation of Gaussians to represent the shape density function of a molecule; the intersection volume of two molecules can then be rapidly computed to align one with respect to the other. While traditional virtual screening applications of ROCS require this alignment step, here we did *not* align one ligand with respect to the other: we simply used ROCS to evaluate their overlap, given their relative positions and orientations from the aligned crystal structures.

In addition to evaluating volume overlap, ROCS can also report on spatial overlap of chemical "color" features (hydrogen bond donors, hydrogen bond acceptors, cations, anions, and aromatic rings). Overlap of these features is computed as with the volume, but a given "type" of feature may only contribute through overlap with the same feature "type" on the comparison ligand.

When used for virtual screening, it is assumed that the size of the template ligand should be similar to that of the hits that are generated. Accordingly, by default ROCS penalizes both ligands equally for containing volume (or chemical features) not shared by the other ligand. Here, however, we wish to penalize the larger ligand for failing to cover the smaller ligand, but we do *not* wish to penalize the larger ligand for including extra volume not present in the smaller ligand. For this reason, we did not use the complete ROCS scores in our analysis, but instead defined COS (the combined overlap score) as follows:

$$COS = 0.5\frac{O_{ls}}{O_{ss}} + 0.5\frac{C_{ls}}{C_{ls}} \quad \text{(Eqn. 2)}$$

Here O_{ls} represents the volume overlap between the two ligands (i.e. the shared volume), and O_{ss} represents the volume overlap of the smaller ligand with itself (i.e. the total volume of the smaller ligand, as normalization). By direct analogy, C_{ls} represents the "color" overlap between the two ligands (i.e. chemical features present in analogous locations), and C_{ss} represents the "color" overlap of the smaller ligand with itself (as a normalization). For the purposes of this study, the two terms are given equal weight. Thus, the value of COS ranges from 0 (if the volume and features of the larger ligand do not overlap with those of the smaller ligand at all) to 1 (if the larger ligand fully contains the volume of the smaller ligand and also perfectly recapitulates the positioning of the smaller ligand's chemical groups). Values between 0 and 1 may be interpreted as the fraction of the smaller ligand's volume/features that are preserved by the larger ligand.

After visually inspecting all ligand pairs with low overlap scores, we defined the criteria by which we could be confident that the binding mode of the larger ligand differed from that of the smaller ligand (Figure S1): (i) chemical substructure score of 1.0 and COS less than

0.55, (ii) chemical substructure score within the range of 0.95–0.99 and COS less than 0.48, or (iii) chemical substructure score within the range of 0.9–0.949 and COS less than 0.4.

Properties collected for individual complexes

For each ligand in our set, we used OpenBabel [47] to determine the molecular weight and to estimate the octanol-water partition coefficient (clogP). We used OMEGA [52–54] to calculate the number of rotatable bonds. We drew the number of ligand heavyatoms and the potency from the PDBbind database itself [45]. Noting that potency values collected in PDBbind can derive from either K_d, K_i, or IC_{50} data, we applied a factor of 2 to each of the IC_{50} values so that they may be most appropriately compared against K_d and K_i values [19]. The ligand efficiency (LE) was calculated from the potency (K) and heavy atom count (HAC) as follows:

$$LE = -\frac{0.596 ln K}{HAC} \quad \text{(Eqn. 3)}$$

From the structure of each complex, we used the Rosetta macromolecular modeling suite [55] to calculate the change in solvent accessible surface area (SASA) upon complexation and the fraction of the ligand's SASA that remains exposed upon complexation (θ_{lig}) [22]. We also used Rosetta to count the number of protein-ligand hydrogen bonds in the complex. Crystallographic water molecules were discarded prior to carrying out these Rosetta calculations.

As with PDBbind, we defined the protein's binding site to be the collection of residues that had at least one (non-hydrogen) atom within 4.5 Å of any (non-hydrogen) ligand atom. For a pair of ligands bound to the same protein, we defined the collective binding site as the union of the sets of residues defined in each structure. The RMSD between structures for all non-hydrogen atoms of residues involved in the collective binding site was computed after global structural alignment using TM-align, without adjusting the alignment to minimize RMSD of the binding site.

The resolution and R_{free} for each crystal structure was drawn directly from the PDB files. We collected the lower resolution of each PDB structure comprising a given ligand pair. Among our set, R_{free} was not reported in one of the PDB structures for 49 ligand pairs; we did not include these pairs in our examination of the effect of R_{free}.

The B-factor of residues comprising the collective binding site was also drawn directly from the PDB files. To account for differences in resolution between different crystal structures, we expressed the average B-factor of the residues in the binding site as a Z-score, computed from the mean and standard deviation of the B-factors for the whole protein. Thus, negative values indicate that the B-factors in the binding site are lower than the overall average for this protein, and positive values indicate that the B-factors in the binding site are higher than the overall average for this protein.

In contrast to most values used in this study, calculation of the FO score is based on the crystal structure of the larger ligand, not the smaller one (because FO score was developed for studies of ligand deconstruction, not chemical elaboration). To compute FO score across our complete benchmark set, we first used the crystal structure of the larger ligand as input for the FTMap server [25]. This computational solvent mapping server carries out global docking using 16 distinct small molecule probes, then reports consensus clusters at which many overlapping probe molecules are found. In parallel, we used the MCS (maximum common substructure) tool implemented in ChemAxon [48] to identify the portion of the larger ligand's chemical structure that is common to both the larger and smaller ligands in our pair.

The next step in determining the FO score was to identify the protein's "main" hotspot: the probe cluster with the highest total number of probe atoms [24]. Applying this to our test set, however, produced a "main" hotspot that did not coincide with the larger ligand in 109 of 382 cases, which in turn cannot be used to calculate a meaningful FO score. For this reason, we instead defined the "main" hotspot as the probe cluster with the highest total number of probe atoms that was within 2 Å of the larger ligand.

Using this "main" hotspot, the FO score is defined as:

$$FO = \frac{N_f}{N_t} \quad \text{(Eqn. 4)}$$

where N_t is the total number of non-hydrogen atoms for all probe-molecules in the main hot spot, and N_f is the number of these atoms that are within 2 Å of the shared substructure (using its position in the crystal structure of the larger ligand). Thus, the docked probe compounds together map a "hotspot" volume that includes the shared substructure, and FO reports on the fraction of this volume that is covered by the shared substructure.

In summary, FO scores were computed exactly as described in the original study [24], except for the (automated) MCS step that replaced manual identification of the atoms in the larger ligand that corresponded to the smaller ligand. We note that in 4 cases the FTMap server did not generate any clusters within 2 Å of the larger ligand. For this reason, we only report the FO score for the remaining 293 pairs in our dataset (41 changed binding mode, and 252 did not).

Properties of the binding pocket (pocket volume, frequency of polar residues, frequency of aromatic residues, pocket hydrophobicity, and pocket druggability score) were obtained from the PockDrug Server [26, 27].

Modeling the effect of a specific chemical substitution (RMAC)

The objective of the RMAC ("rmsd after minimization of the aligned complex") measure is to rapidly determine of whether the structure of the smaller ligand's complex can be used to model the larger ligand, without extensive changes to the binding mode. Our protocol is implemented in the Rosetta macromolecular modeling suite [55], and takes place as follows.

We begin by carrying out an unconstrained gradient-based energy minimization of the crystal structure of the smaller ligand's complex: this ensures that any changes that occur in our model of the larger ligand are indeed due to the chemical substitution, and not due to unfavorable interactions in the starting model. Indeed, we found that for 22 (of 297) cases the starting complex involving the smaller ligand moved by more than 1 Å, preventing further analysis. Thus, our analysis continued using only the other 275 ligand pairs.

We next used the MCS tool implemented in ChemAxon [48] to identify the pairs of corresponding atoms that comprise the maximum common structure (MCS) between the chemical structures of the two ligands. Using a custom script developed in the MMTSB toolset [56], we superposed the three-dimensional structure of the larger ligand onto the smaller ligand, by RMSD alignment of the shared substructure. By removing the smaller ligand, we were then left with an initial model of the larger ligand bound to the (minimized) protein structure from the smaller ligand's complex.

Finally, we carried out an energy minimization of this model complex in Rosetta, this time including a "coordinate constraint" that provides a small energetic bias to hold the atoms to their starting positions. The intention of this approach is to determine whether the initial model built from the smaller ligand's complex can be minimized to yield an energetically reasonable model of the larger ligand's complex: if so, the larger ligand will move only slightly from its starting position. If, on the other hand, the larger ligand is completely incompatible with the smaller ligand's pose, then we expect the large ligand to move upon energy minimization. Accordingly, then we report RMAC as the RMSD of the larger ligand relative to its starting position.

Statistical Analysis

We used the Mann Whitney U-test (as implemented in the R statistical computing environment [57]) to compute the significance of differences between distributions each property between the paired ligands that change binding mode versus those that did not change binding mode. Since we had an expectation ahead of time for whether increases in the value of each property would lead to an increase or decrease in preservation of binding mode, we used one-tailed tests in all cases to test the corresponding hypothesis.

To obtain the predicted probability of an alternate binding mode given a single property (or a combination of properties), we applied logistic regression (or multiple logistic regression) [28] as implemented in the R statistical computing environment [57].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Christian Ray and Eric Deeds for helpful suggestions regarding statistical analysis, Scott Lovell for help with interpretation of electron density, and Hongjing Ma for her preliminary examination of the PDBbind set. We are grateful to OpenEye Scientific Software (Santa Fe, NM) for providing an academic license for the use of ROCS and OMEGA. This work was supported by a grant from the National Institute of General Medical Sciences of the

National Institutes of Health (R01GM099959), the National Science Foundation through XSEDE allocation MCB130049, and the Alfred P. Sloan Fellowship (J.K.).

Abbreviations

COS	combined overlap score
SASA	solvent accessible surface area
FO	fraction overlap
ROC	receiver operating characteristic
RMAC	rmsd after minimization of the aligned complex
TPR	true positive rate
FPR	false positive rate

References

1. Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H. Chemical predictive modelling to improve compound quality. Nat Rev Drug Discov. 2013; 12:948–962. [PubMed: 24287782]
2. Andersen OA, Nathubhai A, Dixon MJ, Eggleston IM, van Aalten DMF. Structure-based dissection of the natural product cyclopentapeptide chitinase inhibitor argifin. Chem Biol. 2008; 15:295–301. [PubMed: 18355729]
3. Stout TJ, Sage CR, Stroud RM. The additivity of substrate fragments in enzyme–ligand binding. Structure. 1998; 6:839–848. [PubMed: 9687366]
4. Fry DC, Wartchow C, Graves B, Janson C, Lukacs C, Kammlott U, Belunis C, Palme S, Klein C, Vu B. Deconstruction of a Nutlin: dissecting the binding determinants of a potent protein-protein interaction inhibitor. ACS Med Chem Lett. 2013; 4:660–665. [PubMed: 24900726]
5. Babaoglu K, Shoichet BK. Deconstructing fragment-based inhibitor discovery. Nat Chem Biol. 2006; 2:720–723. [PubMed: 17072304]
6. Barelier S, Pons J, Marcillat O, Lancelin JM, Krimm I. Fragment-based deconstruction of Bcl-xL inhibitors. J Med Chem. 2010; 53:2577–2588. [PubMed: 20192224]
7. Murray CW, Verdonk ML, Rees DC. Experiences in fragment-based drug discovery. Trends Pharmacol Sci. 2012; 33:224–232. [PubMed: 22459076]
8. Aguirre C, Brink Tt, Guichou J-F, Cala O, Krimm I. Comparing binding modes of analogous fragments using NMR in fragment-based drug design: application to PRDX5. PLoS ONE. 2014; 9:e102300. [PubMed: 25025339]
9. Murray CW, Blundell TL. Structural biology in fragment-based drug design. Curr Opin Struct Biol. 2010; 20:497–507. [PubMed: 20471246]
10. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005; 33:2302–2309. [PubMed: 15849316]
11. Rush TS 3rd, Grant JA, 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; 48:1489–1495. [PubMed: 15743191]
12. ROCS version 3.2.03. OpenEye Scientific Software; Santa Fe, NM: http://www.eyesopen.com (accessed January 9, 2015).
13. Liebeschuetz J, Hennemann J, Olsson T, Groom CR. The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. J Comput Aided Mol Des. 2012; 26:169–183. [PubMed: 22246295]
14. Reynolds CH. Protein-ligand cocrystal structures: we can do better. ACS Med Chem Lett. 2014; 5:727–729. [PubMed: 25050154]

15. Dauter Z, Wlodawer A, Minor W, Jaskolski M, Rupp B. Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining. IUCrJ. 2014; 1:179–193.
16. Martin DP, Cohen SM. Nucleophile recognition as an alternative inhibition mode for benzoic acid based carbonic anhydrase inhibitors. Chem Commun. 2012; 48:5259–5261.
17. Innocenti A, Vullo D, Scozzafava A, Supuran CT. Carbonic anhydrase inhibitors: inhibition of mammalian isoforms I-XIV with a series of substituted phenols including paracetamol and salicylic acid. Bioorg Med Chem. 2008; 16:7424–7428. [PubMed: 18579385]
18. Waku T, Shiraki T, Oyama T, Maebara K, Nakamori R, Morikawa K. The nuclear receptor PPARγ individually responds to serotonin- and fatty acid-metabolites. EMBO J. 2010; 29:3395–3407. [PubMed: 20717101]
19. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of mixed IC50 data – a statistical analysis. PLoS ONE. 2013; 8:e61007. [PubMed: 23613770]
20. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. Proc Natl Acad Sci USA. 1999; 96:9997–10002. [PubMed: 10468550]
21. Du J-Q, Wu J, Zhang H-J, Zhang Y-H, Qiu B-Y, Wu F, Chen Y-H, Li J-Y, Nan F-J, Ding J-P, Li J. Isoquinoline-1,3,4-trione derivatives inactivate caspase-3 by generation of reactive oxygen species. J Biol Chem. 2008; 283:30205–30215. [PubMed: 18768468]
22. Gowthaman R, Deeds EJ, Karanicolas J. Structural properties of non-traditional drug targets present new challenges for virtual screening. J Chem Inf Model. 2013; 53:2073–2081. [PubMed: 23879197]
23. Houk KN, Leach AG, Kim SP, Zhang X. Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes. Angew Chem Int Ed Engl. 2003; 42:4872–4897. [PubMed: 14579432]
24. Kozakov D, Hall DR, Jehle S, Luo L, Ochiana SO, Jones EV, Pollastri M, Allen KN, Whitty A, Vajda S. Ligand deconstruction: Why some fragment binding positions are conserved and others are not. Proc Natl Acad Sci USA. 2015; 112:E2585–E2594. [PubMed: 25918377]
25. Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, Xia B, Beglov D, Vajda S. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. Nat Protocols. 2015; 10:733–755. [PubMed: 25855957]
26. Hussein HA, Borrel A, Geneix C, Petitjean M, Regad L, Camproux AC. PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins. Nucleic Acids Res. 2015; 43:W436–442. [PubMed: 25956651]
27. Borrel A, Regad L, Xhaard H, Petitjean M, Camproux AC. PockDrug: A model for predicting pocket druggability that overcomes pocket estimation uncertainties. J Chem Inf Model. 2015; 55:882–895. [PubMed: 25835082]
28. Cox DR. The regression analysis of binary sequences. J R Statist Soc B. 1958; 20:215–242.
29. Nair SK, Ludwig PA, Christianson DW. Two-site binding of phenol in the active site of human carbonic anhydrase II: Structural implications for substrate association. J Amer Chem Soc. 1994; 116:3659–3660.
30. Davies DR, Mamat B, Magnusson OT, Christensen J, Haraldsson MH, Mishra R, Pease B, Hansen E, Singh J, Zembower D, Kim H, Kiselyov AS, Burgin AB, Gurney ME, Stewart LJ. Discovery of leukotriene A4 hydrolase inhibitors using metabolomics biased fragment crystallography. J Med Chem. 2009; 52:4694–4715. [PubMed: 19618939]
31. Manos-Turvey A, Bulloch EM, Rutledge PJ, Baker EN, Lott JS, Payne RJ. Inhibition studies of Mycobacterium tuberculosis salicylate synthase (MbtI). ChemMedChem. 2010; 5:1067–1079. [PubMed: 20512795]
32. Chi G, Manos-Turvey A, O'Connor PD, Johnston JM, Evans GL, Baker EN, Payne RJ, Lott JS, Bulloch EM. Implications of binding mode and active site flexibility for inhibitor potency against the salicylate synthase from Mycobacterium tuberculosis. Biochemistry. 2012; 51:4868–4879. [PubMed: 22607697]
33. Jiang L-G, Yu H-Y, Yuan C, Wang J-D, Chen L-Q, Meehan EJ, Huang Z-X, Huang M-D. Crystal structures of 2-aminobenzothiazole-based inhibitors in complexes with urokinase-type plasminogen activator. Chin J Struct Chem. 2009; 28:1427–1432.

34. Fischmann TO, Hruza A, Duca JS, Ramanathan L, Mayhood T, Windsor WT, Le HV, Guzi TJ, Dwyer MP, Paruch K, Doll RJ, Lees E, Parry D, Seghezzi W, Madison V. Structure-guided discovery of cyclin-dependent kinase inhibitors. Biopolymers. 2008; 89:372–379. [PubMed: 17937404]
35. Zartler ER. Fragonomics: the -omics with real impact. ACS Med Chem Lett. 2014; 5:952–953. [PubMed: 25221648]
36. Murray CW, Rees DC. The rise of fragment-based drug discovery. Nat Chem. 2009; 1:187–192. [PubMed: 21378847]
37. Congreve M, Carr R, Murray C, Jhoti H. A 'rule of three' for fragment-based lead discovery? Drug Discovery Today. 2003; 8:876–877.
38. Jhoti H, Williams G, Rees DC, Murray CW. The 'rule of three' for fragment-based drug discovery: where are we now? Nat Rev Drug Discov. 2013; 12:644–644. [PubMed: 23845999]
39. Murray CW, Carr MG, Callaghan O, Chessari G, Congreve M, Cowan S, Coyle JE, Downham R, Figueroa E, Frederickson M, Graham B, McMenamin R, O'Brien MA, Patel S, Phillips TR, Williams G, Woodhead AJ, Woolford AJ. Fragment-based drug discovery applied to Hsp90. Discovery of two lead series with high ligand efficiency. J Med Chem. 2010; 53:5942–5955. [PubMed: 20718493]
40. Bussenius J, Blazey CM, Aay N, Anand NK, Arcalas A, Baik T, Bowles OJ, Buhr CA, Costanzo S, Curtis JK, DeFina SC, Dubenko L, Heuer TS, Huang P, Jaeger C, Joshi A, Kennedy AR, Kim AI, Lara K, Lee J, Li J, Lougheed JC, Ma S, Malek S, Manalo JC, Martini JF, McGrath G, Nicoll M, Nuss JM, Pack M, Peto CJ, Tsang TH, Wang L, Womble SW, Yakes M, Zhang W, Rice KD. Discovery of XL888: a novel tropane-derived small molecule inhibitor of HSP90. Bioorg Med Chem Lett. 2012; 22:5396–5404. [PubMed: 22877636]
41. Aguirre C, ten Brink T, Guichou JF, Cala O, Krimm I. Comparing binding modes of analogous fragments using NMR in fragment-based drug design: application to PRDX5. PLoS One. 2014; 9:e102300. [PubMed: 25025339]
42. Czodrowski P, Holzemann G, Barnickel G, Greiner H, Musil D. Selection of fragments for kinase inhibitor design: decoration is key. J Med Chem. 2015; 58:457–465. [PubMed: 25437144]
43. Brough PA, Barril X, Borgognoni J, Chene P, Davies NGM, Davis B, Drysdale MJ, Dymock B, Eccles SA, Garcia-Echeverria C, Fromont C, Hayes A, Hubbard RE, Jordan AM, Jensen MR, Massey A, Merrett A, Padfield A, Parsons R, Radimerski T, Raynaud FI, Robertson A, Roughley SD, Schoepfer J, Simmonite H, Sharp SY, Surgenor A, Valenti M, Walls S, Webb P, Wood M, Workman P, Wright L. Combining hit identification strategies: Fragment-based and in silico approaches to orally active 2-aminothieno[2,3-d]pyrimidine inhibitors of the Hsp90 molecular chaperone. J Med Chem. 2009; 52:4794–4809. [PubMed: 19610616]
44. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: Collection of binding affinities for protein−ligand complexes with known three-dimensional structures. J Med Chem. 2004; 47:2977–2980. [PubMed: 15163179]
45. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, Nie W, Liu Y, Wang R. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics. 2015; 31:405–412. [PubMed: 25301850]
46. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–212. [PubMed: 25348405]
47. O'Boyle N, Banck M, James C, Morley C, Vandermeersch T, Hutchison G. Open Babel: An open chemical toolbox. J Cheminform. 2011; 3:33. [PubMed: 21982300]
48. Englert P, Kovacs P. Efficient heuristics for maximum common substructure search. J Chem Inf Model. 2015; 55:941–955. [PubMed: 25865959]
49. Kleywegt GJ, Harris MR, Zou J-y, Taylor TC, Wahlby A, Jones TA. The Uppsala Electron-Density server. Acta Crystallogr D Biol Crystallogr. 2004; 60:2240–2249. [PubMed: 15572777]
50. The PyMOL Molecular Graphics System. Version 1.8 Schrödinger, LLC.
51. Bietz S, Urbaczek S, Schulz B, Rarey M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. J Cheminform. 2014; 6:1–12. [PubMed: 24397863]

52. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J Chem Inf Model. 2010; 50:572–584. [PubMed: 20235588]

53. Hawkins PC, Nicholls A. Conformer generation with OMEGA: learning from the data set and the analysis of failures. J Chem Inf Model. 2012; 52:2919–2936. [PubMed: 23082786]

54. OMEGA version 2.4.3. OpenEye Scientific Software; Santa Fe, NM: http://www.eyesopen.com (accessed January 9, 2015)

55. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman KW, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban Y-EA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011; 487:545–574. [PubMed: 21187238]

56. Feig M, Karanicolas J, Brooks CL III. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graph Model. 2004; 22:377–395. [PubMed: 15099834]

57. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2010.

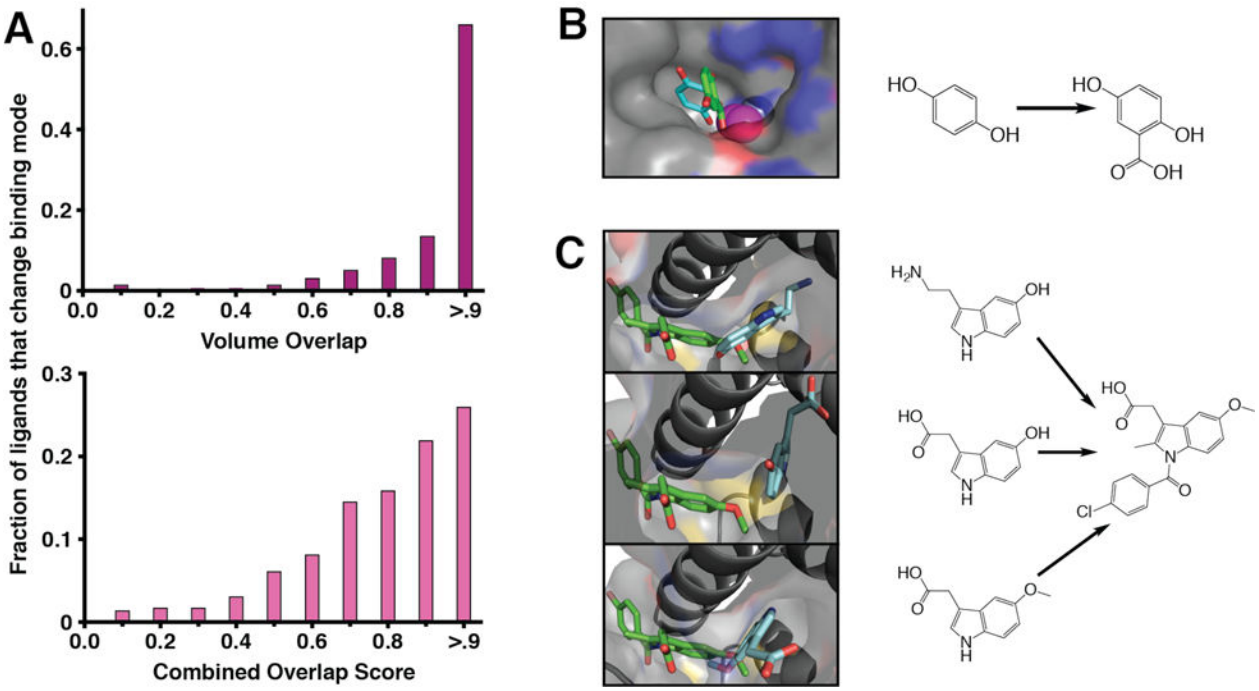


Figure 1. Identifying ligand pairs with alternate binding modes, from the Protein Data Bank

(A) In most cases, almost all of the smaller ligand's volume is contained within the volume of the larger ligand; however, there are a surprising number of cases for which this is not the case. The use of COS ("combined overlap score") captures overlap of both volume and chemical types (see *Methods*), providing additional accuracy in identifying alternate binding modes: cases in which position of the smaller ligand does not match the position of this substructure in the larger ligand. **(B)** An example of one such alternate binding mode: upon elaboration, the position of the ring in the larger ligand (*green*, PDB ID 4e3d) no longer matches the position of the ring in the smaller ligand (*cyan*, PDB ID 4e3h). In this case, the smaller ligand is a perfect substructure of the larger ligand. In this case the enzyme active site includes a Zn(II) ion (*grey*) that activates a bound water molecule (*pink*). **(C)** Another example of an alternate set of binding modes, this time across a chemical series. The largest ligand (*green*, PDB ID 3ads) is shown in each panel, for reference. Though the smaller ligands are very similar to one another (*cyan*, PDB IDs 3adv/3adt/3adu), they each adopt different binding modes – and none of them match that of the corresponding structural element in the larger ligand.

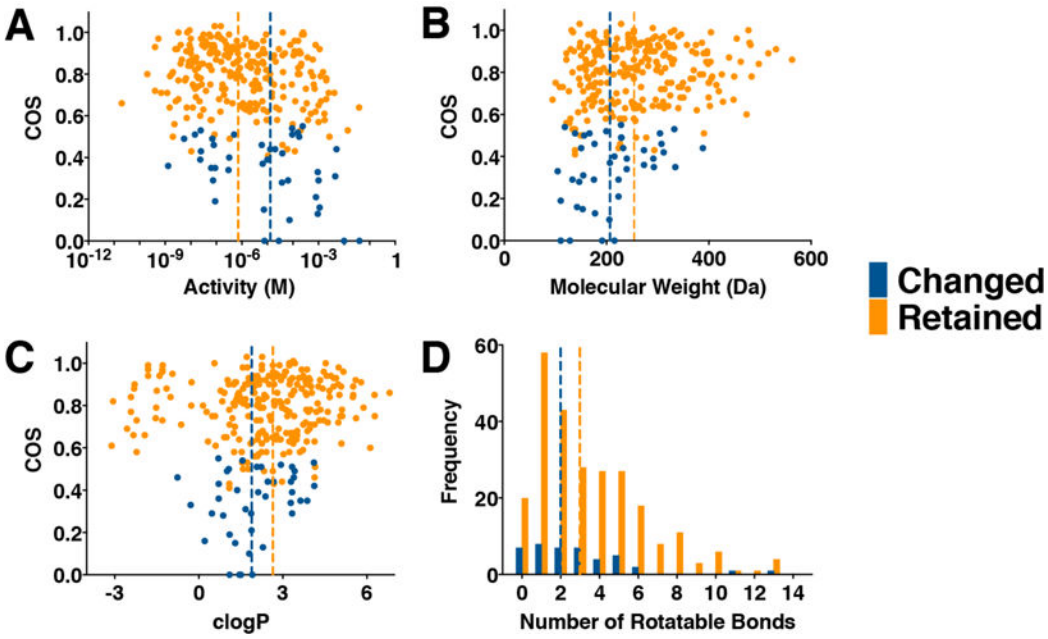


Figure 2. Certain properties of the smaller ligand correlate with increased likelihood of changing binding mode when elaborated

In each case, *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is preserved. Vertical lines indicate the medians of each distribution. **(A)** Compounds that change binding mode upon elaboration are typically less potent than compounds that retain their binding mode ($p < 0.005$). **(B)** Compounds that change binding mode upon elaboration are typically smaller than compounds that retain their binding mode ($p < 4\times10^{-4}$). **(C)** Compounds that change binding mode upon elaboration are typically less lipophilic than compounds that retain their binding mode ($p < 0.02$). **(D)** Compounds that change binding mode upon elaboration typically have fewer rotatable bonds relative to compounds that retain their binding mode, but this difference is not statistically significant ($p < 0.1$).

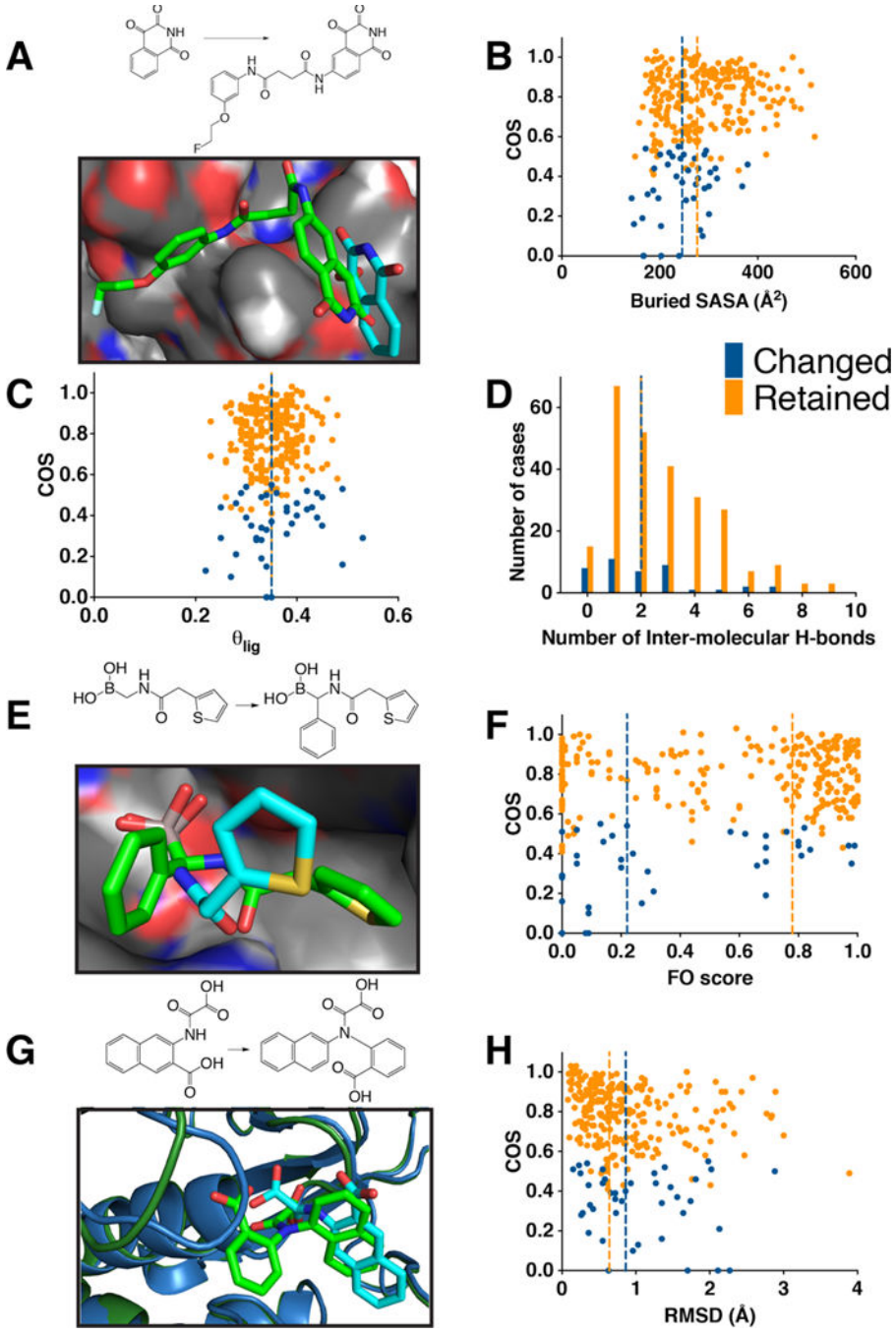


Figure 3. Certain properties of the smaller ligand's complex correlate with increased likelihood of changing binding mode when elaborated

In each case, *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is preserved. Vertical lines indicate the medians of each distribution. **(A)** In this example of alternate binding modes, the smaller ligand (*cyan*, PDB ID 3deh) uses a very shallow binding mode on the surface of caspase-3; upon elaboration, the binding mode of the larger ligand retains the position of this structural element, but has reversed the relative orientation of the

arrangement of the polar and non-polar sides of this fragment (*green*, PDB ID 3dek). **(B)** Compounds that change binding mode upon elaboration typically bury less solvent accessible surface area than compounds that retain their binding mode ($p < 9\times10^{-4}$). **(C)** Compounds that change binding mode upon elaboration do *not* bind with more shallow binding modes, which would correspond to higher θ_{lig} values (θ_{lig} is the fraction of the ligand's SASA that remains exposed upon complexation). **(D)** Compounds that change binding mode upon elaboration typically have fewer intermolecular hydrogen bonds ($p < 0.02$), even though the median value is the same for this discrete variable. **(E)** In this example of alternate binding modes (*cyan*, PDB ID 1fsw; *green*, PDB ID 1my8), both β-lactamase inhibitors make identical hydrogen bonds using their boronic acid groups; upon addition of an extra phenyl ring, however, the amide linker flips over to position the thiophene in a very different location. **(F)** Compounds that change binding mode upon elaboration typically have lower FO scores (a measure of the extent to which the smaller ligand fills the larger ligand's "binding energy hot spot") than compounds that retain their binding mode ($p < 5\times10^{-4}$). **(G)** In this example of alternate binding modes, the smaller ligand forms stacking interactions with a phenylalanine sidechain in the binding site (*cyan*, PDB ID 1c84). Elaboration with a benzoic acid group pushes away this phenylalanine sidechain, and instead forms new hydrogen bonds that require the ligand to move within the binding site (*green*, PDB ID 1no6). Meanwhile, this larger ligand pushes away a loop that previously covered the binding site (*left*), which is primarily responsible for the RMSD difference between the two protein structures. **(H)** Compounds that change binding mode upon elaboration are more often accompanied by conformational rearrangement of the protein's binding site, relative to compounds that retain their binding mode ($p < 0.03$).

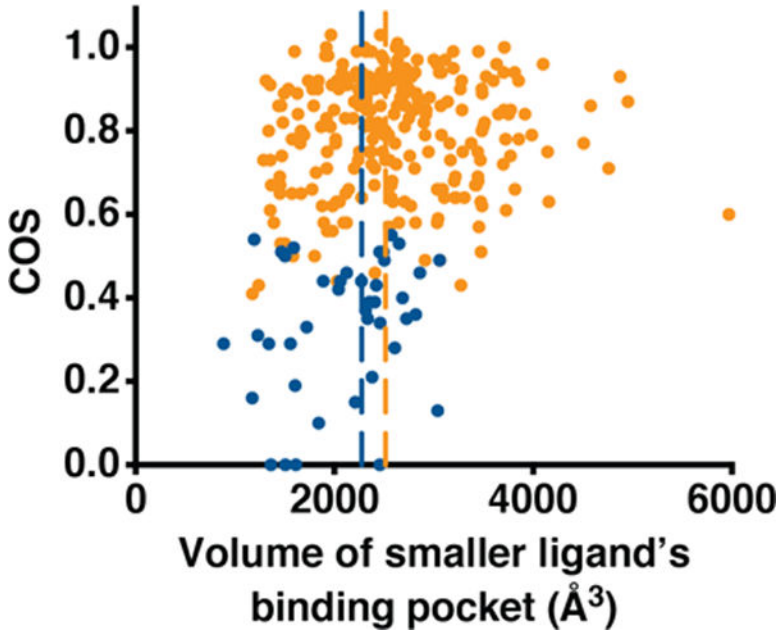


Figure 4. The size of the initial binding pocket correlates with the likelihood of changing binding mode upon elaboration

Here, *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is preserved. Vertical lines indicate the medians of each distribution. Compounds that change binding mode upon elaboration typically have a smaller binding pocket than compounds that retain their binding mode ($p < 9\times10^{-5}$).

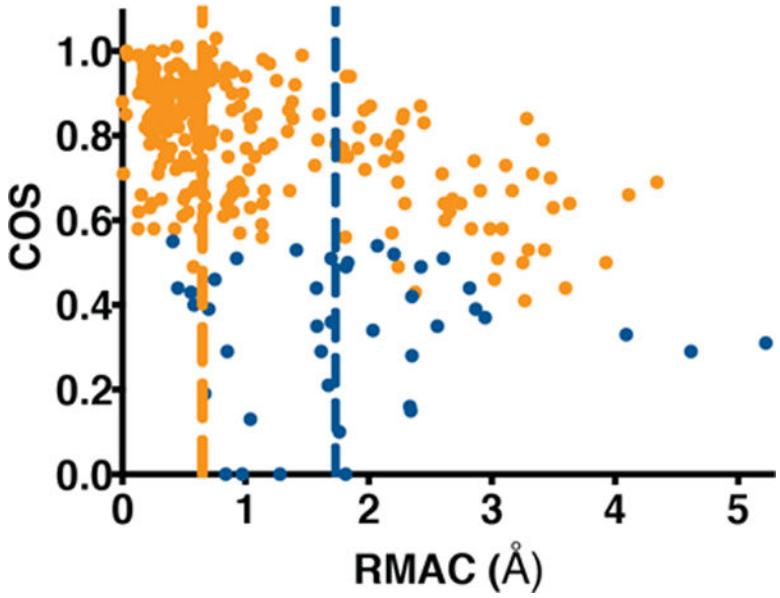


Figure 5. Directly probing whether the larger ligand can be accommodated without changing binding mode

RMAC is a measure of whether the structure of the smaller ligand's complex can be used to model the larger ligand: the larger ligand is aligned to the smaller ligand, and then its RMSD is measured after energy minimization of the complex. Here, *blue* indicates cases in which the binding mode changes upon elaboration, and *orange* is used for cases in which the binding mode is preserved. Vertical lines indicate the medians of each distribution. We find that substitutions that cannot be accommodated in the original binding mode (high RMAC) are more likely to change binding mode ($p < 6\times10^{-7}$), and that RMAC distinguishes ligand pairs that with alternate binding modes better than any other single individual property considered in this study.

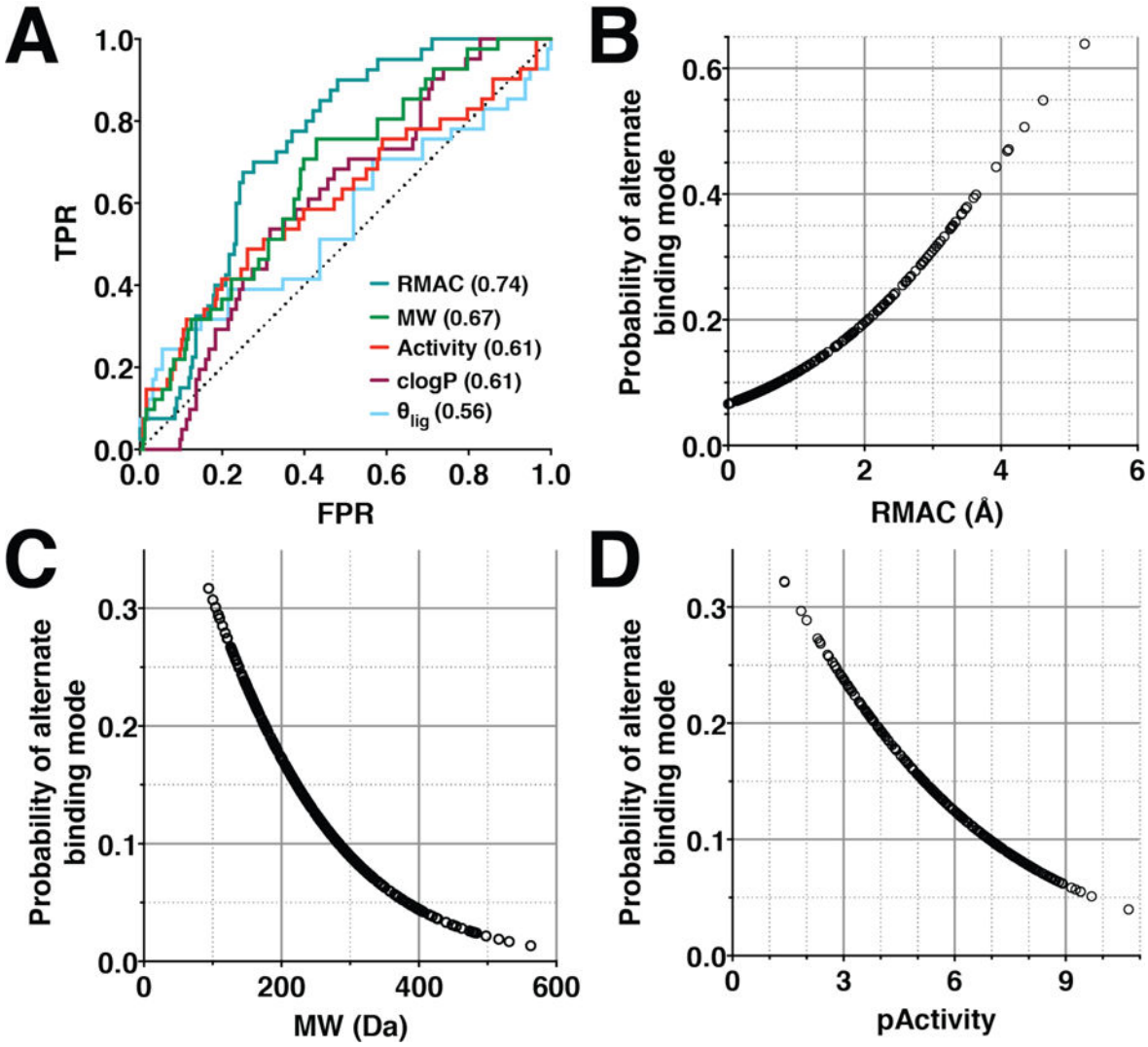


Figure 6. Primary determinants of chemical substitutions that lead to new binding modes

(A) Receiver operating characteristic (ROC) plots comparing the utility of several different properties for predicting whether a ligand will change binding mode upon chemical elaboration, by plotting the true positive rate (TPR) as a function of the false positive rate (FPR). The performance of a random classifier is denoted by the black dotted line. The area under curve (AUC) for each of these properties is indicated. AUC values for all properties in this study are included in Table 1, and the corresponding ROC plots are included as Figure S7. **(B)** Using logistic regression, we estimate the probability that a given substitution will lead to a change in binding mode, as a function of RMAC. We also estimate the probability that a given ligand will change its binding mode upon chemical elaboration as a function of the initial compound's **(C)** molecular weight and **(D)** potency.

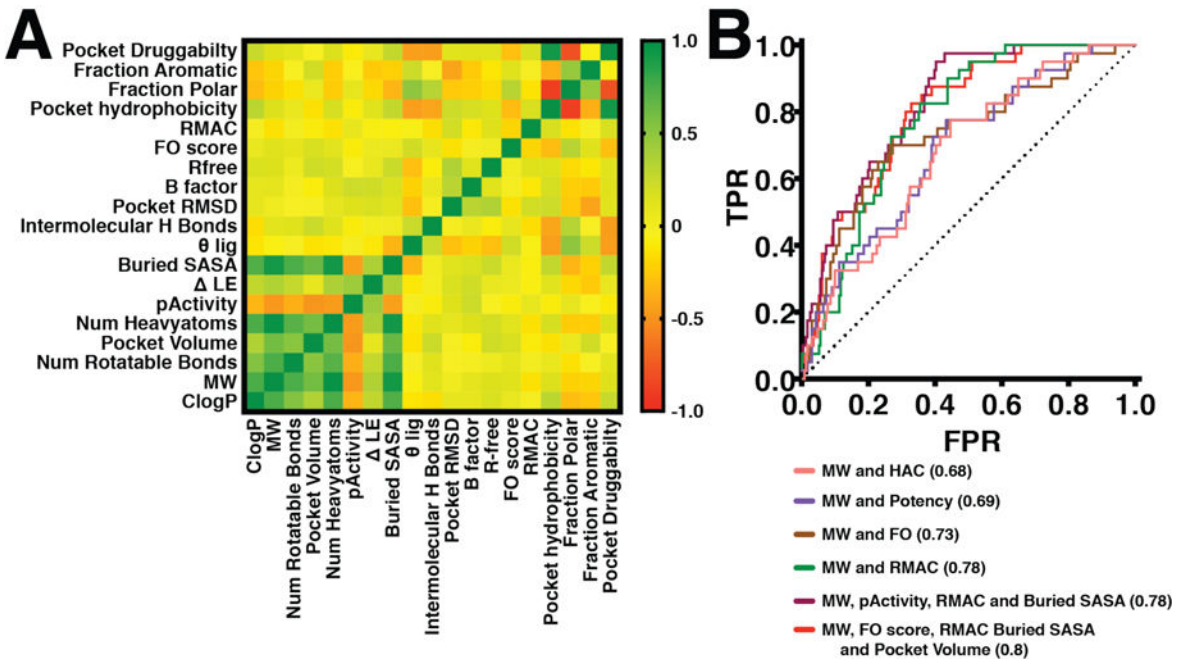


Figure 7. Combining properties leads to a model with more predictive power

(A) Correlation in our test set between each of the properties considered: using this color gradient, uncorrelated properties are *yellow*. **(B)** Receiver operating characteristic (ROC) plots comparing the multiple-regression analysis based predictive powers of several different properties for predicting whether a ligand will change binding mode upon chemical elaboration. The performance of a random classifier is denoted by the black dotted line. The area under curve (AUC) for each of these properties is indicated.

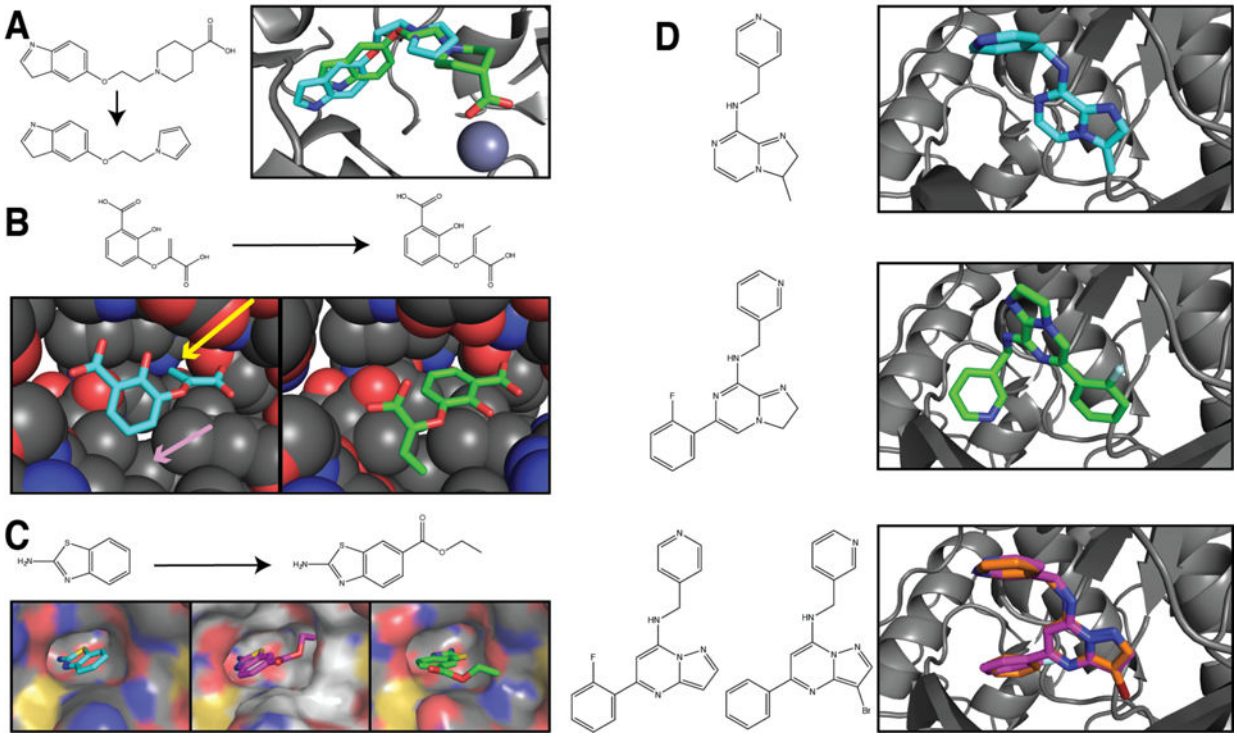


Figure 8. Examples of alternate binding modes adopted despite the lack of a conflict for the larger ligand in the original binding mode

(A) The LTA4H active site contains a Zn(II) ion (*grey*) that is not engaged by the smaller ligand (*cyan*, PDB ID 3fuj). Elaboration with a carboxylic acid shifts the ligand position to allow a direct interaction with the metal ion (*green*, PDB ID 3fuk), but diminishes potency. **(B)** The crystal structure of an isochorismate mimic inhibitor of MbtI (*left*, *cyan*, PDB ID 3st6) reveals a cavity that is not filled by the ligand (*pink arrow*). Elaboration with a methyl group at this terminal alkene (*yellow arrow*) induces the ligand to flip over, preserving the interactions of the two carboxylic acid groups and positioning the methyl group to fill this cavity (*right*, *green*, PDB ID 3veh). **(C)** The crystal structure of 2-aminobenzothiazole in complex with urokinase reveals two small pockets at the base of the binding site (*left*, *cyan*, PDB ID 3mhw). Modeling shows that the larger ligand can be accommodated using this binding mode, through slight adjustment of surface sidechains (*middle*, *magenta*). However, a crystal structure of this complex reveals that the ligand has instead shifted to engage the other small pocket at the base of the binding site (*right*, *green*, PDB ID 3kid). **(D)** Most 5,6-bicyclic heterocyclic inhibitors of CDK2 use a common binding mode (*top*, *cyan*, PDB ID 2r3h). A crystal structure of one specific compound, however, shows the ligand rotated in the binding site (*middle*, *green*, PDB ID 2r3g). Multiple analogs that collectively test different cores and substituents each retain the more common binding mode (*bottom*, *magenta/orange*, PDB IDs 2r3i/2r3j), suggesting that the alternate binding mode arises not because of a single change to the structure, but rather due to a specific combination of the core and the substituents.

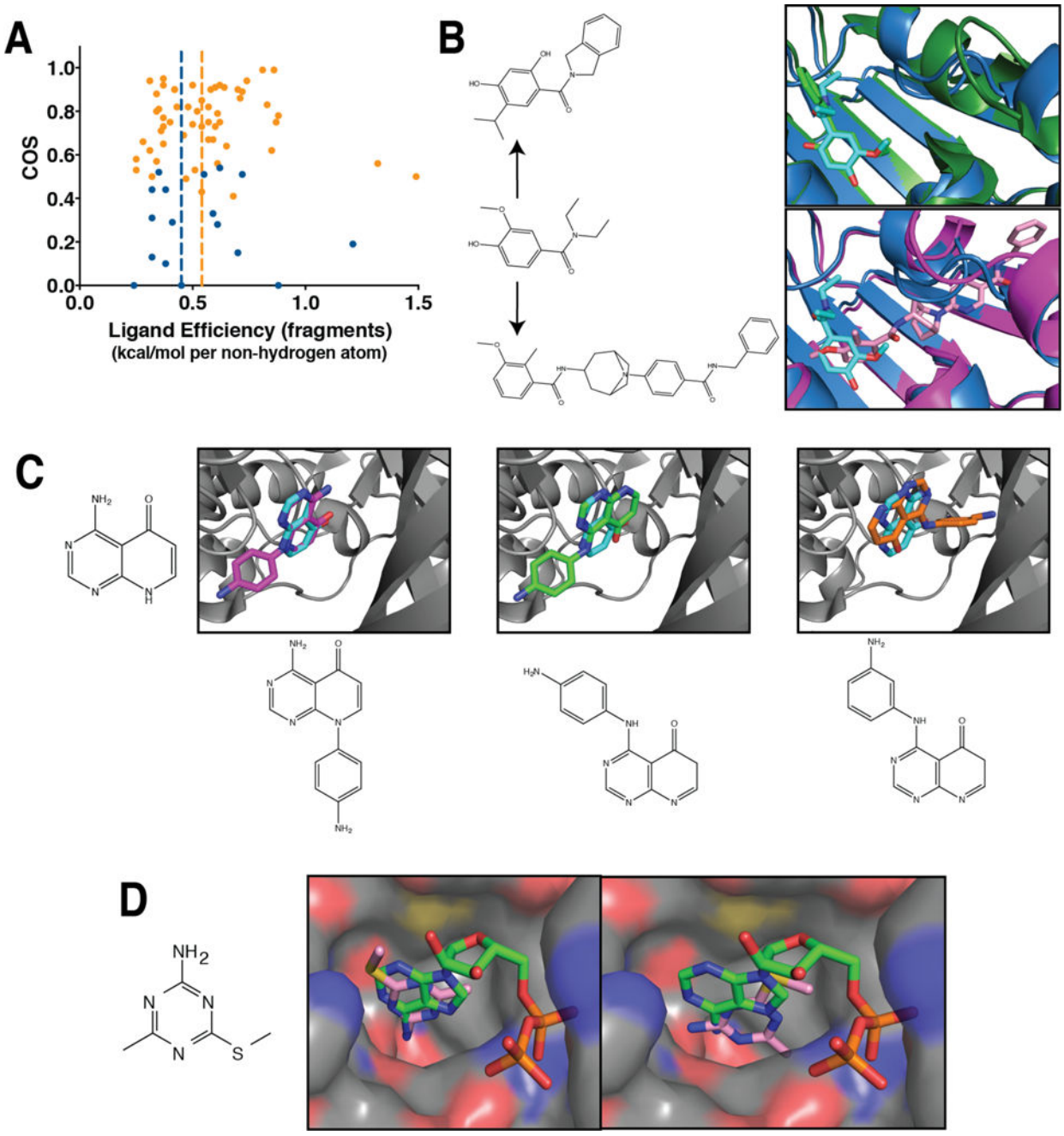


Figure 9. Fragments are particularly prone to alternate binding modes

(A) Among fragment starting points, there is no statistically significant difference in ligand efficiency between those that change binding mode upon elaboration versus those that retain their binding mode ($p < 0.3$). **(B)** Fragment screening for ATP-competitive Hsp90 inhibitors yielded an initial hit (*cyan*, PDB ID 2×dl) that was elaborated into a more potent lead while perfectly preserving the binding mode (*top, green*, PDB ID 2xab). Separately, a high-throughput screen yielded a compound with related chemical structure that positions the corresponding ring in a completely different orientation from that of the fragment (*bottom, pink*, PDB ID 4awq). **(C)** The structure of the 4-amino-8H-pyrido[2,3-d]pyrimidin-5-one

core compound was solved in complex with TGFBR1 (*cyan*, PDB ID 4×0m), and found to engage with the kinase hinge region through a specific set of hydrogen bonds. Elaborating with an anilino group at one position preserved the binding mode (*left, magenta*, PDB ID 4×2f), whereas substituting this anilino group at two other positions yielded two more distinct binding modes (*middle, green*, PDB ID 4×2g; *right, orange*, PDB ID 4×2j). **(D)** Fragment screening for ATP-competitive Hsp90 inhibitors led to 4-methyl-6-(methylsulfanyl)-1,3,5-triazin-2-amine. When this compound is co-crystallized with the protein (*left*, PDB ID 2wi2), it closely mimics the interactions of ADP. However, soaking the same compound into protein crystal yields a different binding mode (*right*, PDB ID 2wi3), which makes different interactions and offers distinct opportunities for optimization.

Table 1
Summary of properties collected in the course of this study

The specific values for each ligand pair included in our study are available as Dataset S1. p-values values refer to the statistical significance of the difference between distributions of each property between the paired ligands that change binding mode versus those that did not change binding mode, in all cases evaluated using the one-tailed Mann Whitney U-test. AUC_{ROC} values refer to the area under the curve for the corresponding ROC plots (Figure S7). Certain properties (indicated) cannot be calculated without a crystal structure solved in complex with the larger ligand; thus, they are not immediately useful for predicting whether a small ligand will preserve its binding mode upon chemical elaboration.

Property	Description	p-value	AUC_{ROC}	Requires crystal structure of larger ligand
RMAC	RMSD after minimization of the large ligand, when aligned onto the small ligand's complex (Å)	6×10^{-7}	0.74	
Pocket Volume	Volume of the pocket in the structure of the smaller ligand ($Å^3$)	9×10^{-5}	0.68	
MW	Molecular weight of the smaller ligand (Da)	4×10^{-4}	0.67	
FO score	Fraction overlap of the smaller ligand with the "binding energy hot spot" from the larger ligand, defined in [24]	5×10^{-4}	0.66	✓
Buried SASA	Solvent accessible surface area buried upon binding of the smaller ligand ($Å^2$)	9×10^{-4}	0.65	
Num heavyatoms	Number of non-hydrogen atoms in the smaller ligand	4×10^{-4}	0.64	
clogP	Computed octanol-water partition coefficient	0.02	0.61	
pActivity	$-\log_{10}$ of the smaller ligand's Kd/Ki	0.005	0.61	
RMSDpocket	RMSD difference of binding site residues between the two ligand-bound structures (Å)	0.03	0.60	✓
B-factor	Crystallographic B-factors of the binding site residues, relative to the rest of the (smaller ligand's) protein structure	0.06	0.57	
Pocket druggability	Predicted druggability score (from PockDrug)	0.07	0.58	
Pocket hydrophobicity	Hydrophobicity of pocket residues	0.08	0.57	
θ_{lig}	Fraction of the smaller ligand's SASA that remains exposed upon binding to the protein	0.2	0.56	
Intermolecular Hbonds	Number of intermolecular hydrogen bonds in the smaller ligand's complex	0.02	0.53	
Fraction polar	Frequency of polar residues in the smaller ligand's binding pocket	0.2	0.53	
Fraction aromatic	Frequency of aromatic residues in the smaller ligand's binding pocket	0.4	0.50	
Num rotatable bonds	Number of rotatable bonds in the smaller ligand	0.1	0.50	