Predicting the Most Tractable Protein Surfaces in the Human Proteome for Developing New Therapeutics

By

Shipra Malhotra

Submitted to the graduate degree program in Computational Biology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chair: Dr. Ilya Vakser

Co-Chair: Dr. John Karanicolas

Dr. Christian Ray

Dr. Joanna Slusky

Dr. Yinglong Miao

Prof. Roberto De Guzman

Dr. Michael Rafferty

Date Defended: 01 May 2019

The dissertation committee for Shipra Malhotra certifies that this is the approved version of the following dissertation:

Predicting the Most Tractable Protein Surfaces in the Human Proteome for Developing New Therapeutics

Chair: Dr. Ilya Vakser

Co-Chair: Dr. John Karanicolas

Date Approved: 01 May 2019

ABSTRACT

A critical step in the target identification phase of drug discovery is evaluating druggability, i.e., whether a protein can be targeted with high affinity using drug-like ligands. The overarching goal of my PhD thesis is to build a machine learning model that predicts the binding affinity that can be attained when addressing a given protein surface. I begin by examining the lead optimization phase of drug development, where I find that in a test set of 297 examples, 41 of these (14%) change binding mode when a ligand is elaborated. My analysis shows that 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. My proposed measure of RMAC (rmsd after minimization of the aligned complex) can help determine whether a given ligand can be reliably elaborated without changing binding mode, thus enabling straightforward interpretation of the resulting structure-activity relationships. Moving forward, I next noted that a very popular machine learning algorithm for regression tasks, random forest, has a systematic bias in the predictions it generates; this bias is present in both real-world datasets and synthetic datasets. To address this, I define a numerical transformation that can be applied to the output of random forest models. This transformation fully removes the bias in the resulting predictions, and yields improved predictions across all datasets. Finally, taking advantage of this improved machine learning approach, I describe a model that predicts the "attainable binding affinity" for a given binding pocket on a protein surface. This model uses 13 physiochemical and structural features calculated from the protein structure, without any information about the ligand. While details of the ligand must (of course) contribute somewhat to the binding affinity, I find that this model still recapitulates the binding affinity for 848 different protein-ligand complexes (across 230 different proteins) with correlation coefficient 0.57. I further find that this model is not limited to "traditional" drug targets, but rather that it works just as well for emerging "non-traditional" drug targets such as inhibitors of protein-protein interactions. Collectively,

I anticipate that the tools and insights generated in the course of my PhD research will play an important role in facilitating the key target selection phase of drug discovery projects.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. John Karanicolas for his invaluable guidance, continued support and patience throughout the duration of my research. You appreciated my hard work were always willing to help me.

I would like to thank the members of my dissertation committee: Dr. Ilya Vakser, Dr. Christian Ray, Dr. Joanna Slusky, Dr. Yinglong Miao, Prof. Roberto De Guzman, Dr. Michael Rafferty and my former committee member Dr. Eric Deeds for their support of my academic career.

I am thankful to my colleagues Dr, Andrea Bazzoli, Dr. David Johnson, Dr. Ragul Gowtahman, Yusuf Adeshina, Jittasak Khowsathit and Nan Bai for their collaborations and discussions.

Finally, I would like to acknowledge the support and love of my parents: Dr. Kiran Malhotra and and Dr. Rajesh Malhotra and my fiancé Sanchit Arora during this time.

ABBREVATIONS

AUC: Area Under the Curve ---------- 24, 26, 27

COS: Combined Overlap Score ---------- 8, 9, 11, 41, 42, 98

FO: Fraction Overlap ---------- 19, 27, 43, 44, 47

GBM: Gradient Boosting Method -------- viii, 2, 63, 76, 77, 78, 79, 80, 82, 83, 86, 87, 88, 90, 91, 92, 93, 94, 96

HTS: High Throughput Screening ---------- 35

MCS: Maximum Common Substructure ---------- 39, 43, 44, 45

MW: Molecular Weight ---------- 47, 86, 108

PDB: Protein Data Bank ---------- 7, 16, 40, 43, 73, 79, 82, 86, 88, 90, 95, 100, 105, 121

PPI: Protein-Protein Interactions ---------- viii, 2, 79, 80, 88, 90, 96

RF: Random Forest ---------- 3, 82, 126

RMAC: rmsd after minimization of the aligned complex ---- iii, vii, 1, 22, 24, 27, 28, 29, 31, 32, 34, 45, 47, 105

RMSD: Root Mean Square Deviation ---------- 6, 20, 22, 43, 45, 47

RMSE: Root-Mean-Square Error ---------- 65, 66, 67, 77, 78, 79, 80, 82, 84, 87, 88

ROC: Receiver Operating Characteristic ---------- 23, 24, 26, 48, 97, 104

SASA: Solvent Accessible Surface Area ---------- 16, 27, 42, 47

SBDD: Structure Based Drug Design ---------- 1

Table of Contents

INTRODUCTION

Structure based drug discovery (SBDD) has become an important tool in fast and efficient drug discovery, design and optimization. From designing selective ATP competitors to modulate protein kinase activity to modulating GPCRs through small molecules, SBDD has seen it all. However, small-molecule drug discovery is quite a challenging multidimensional problem in terms of which various characteristics of small molecules like efficacy, pharmacokinetics and safety need to be optimized in parallel to provide drug candidates. While selecting or designing small molecules medicinal chemists routinely face complex multidimensional optimization. This happens due to the assumption that in a traditional hit-to-lead optimization the ligand retains its binding mode relative to the receptor upon chemical elaboration of the structure. In Chapter 1, we find that for 41 of 297 pairs (14%), the binding mode changes upon elaboration of the smaller ligand. We observed that 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. As, 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. In chapter 1, we designed a new measure, “RMAC” (“rmsd after minimization of the aligned complex”) objective of which was 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.

Another hurdle that can make drug discovery a challenging problem is the druggability of the protein surface. A binding site a pocket on protein surface, where small molecules can best fit or bind to activate the receptor and/or target and produce the desirable effect. A target is considered druggable if it has high affinity and specificity for small drug-like molecules. In Chapter 3, we build a machine model that can quantitatively assess a protein target and all its potential ligand binding sites for druggability without relying on information from the small molecules. This can be an important step in target identification phase of drug discovery. For this study, we assembled a refined set of 230 proteins targets and their small molecule bound binding pocket from 848 crystal structures in Protein Data Bank. We trained our GBM model based 13 physiochemical and structural features of these pockets. The binding data from PDBbind database was used as label. Predicted "attainable binding affinity" for these targets are their binding pockets showed a high correlation to its actual binding affinity. Not just traditional ligands, GBM also showed that even large and flat protein-protein interfaces can be modulated through small molecules assembly. This model can potentially aid in selecting sub-pockets on large PPI interfaces that contains high affinity pockets.

Machine learning approaches like ours are gaining importance since last two decades. Essentially, it is a technique aimed at building systems that analyze data, identify patterns in the data, and then make decisions without explicit human intervention. This field of data science has help drug discovery evolve dramatically. Random forest models have emerged as a popular and robust method for high-dimensional complex data. In the field of computational biology, for example, early random forest models proved especially effective for analysis of genetic data that sought to predict a continuous phenotype or clinical trait (disease state, viral replication capacity, emergence of resistance, or necessary dose of a drug) using variations in gene sequence (1-4). In other domains of computational biology, random forest have shown success in probing the influence of biological and physical properties for thousands of recombinant Protein Epitope Signature Tags (PrESTs) used as antigens and aid in generating antibodies for profiling tissue microarrays(5). Random forest based predictors are also used in predicting the protein-proteins interface, one by incorporating the information of “chaos game representation” into the PseAAC (Pseudo Amino Acid Composition)(6). In other tool, RF based classifier is used to classify protein–protein interfaces into biological and crystallographic interfaces(7).

Through development of random forest models for a variety of different regression problems, we have observed that the resulting models were inevitably too conservative with their predictions. In Chapter 2, we demonstrate that this is a systematic pathology of random forest regressors, and we show that it applies even to artificial data for which the target variable is trivially calculable from the features. Through these studies we developed a numerical transformation that can be applied to the output values from a random forest regressor. In chapter 2, we show that this improves the accuracy of the model’s predictions.

CHAPTER 1: 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 and [3] Department of Molecular Biosciences,
University of Kansas, 2030 Becker Dr., Lawrence, KS 66045-7534

[*]To whom correspondence should be addressed. E-mail: **john.karanicolas@fccc.edu**, 215-728-7067

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.

1.1 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 (8). 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 (9). 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 (10). The Nutlin series that inhibits the MDM2/p53 interaction was also decomposed into its component fragments, and these were shown to retain detectable activity (11) – 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 (12). 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 (13). Motivation for these two studies stemmed primarily from the growing popularity of fragment-based drug discovery (14), 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 (15, 16). 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 (14).

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.

1.2 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 (12) 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**).

1.2.1 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 (17). 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 (18, 19) 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.

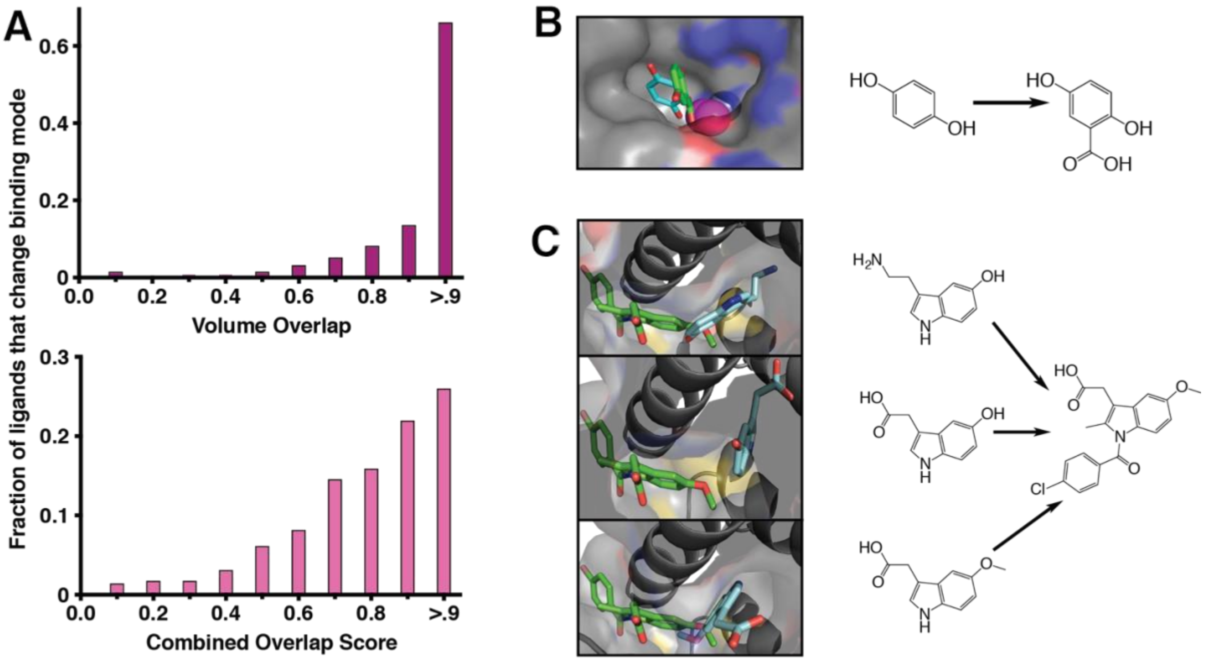


Figure 1: Identifying ligand pairs with alternate binding modes, from the Protein Data Bank. **(A)** In most cases, almost all of the smaller ligand's volume is contained within the volume of the larger ligand; however, there are a surprising number of cases for which this is not the case. The use of a combined overlap score "COS" that captures overlap of both volume and chemical types (see *Methods*) provides 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**.**

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 (20-22), 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 (23).

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 (23). 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 (24), versus IC_{50} = 5 mM for 2,5-dihydroxybenzoic acid (23)).

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) (25). 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 (25). 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.

1.2.2 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 (26) (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.

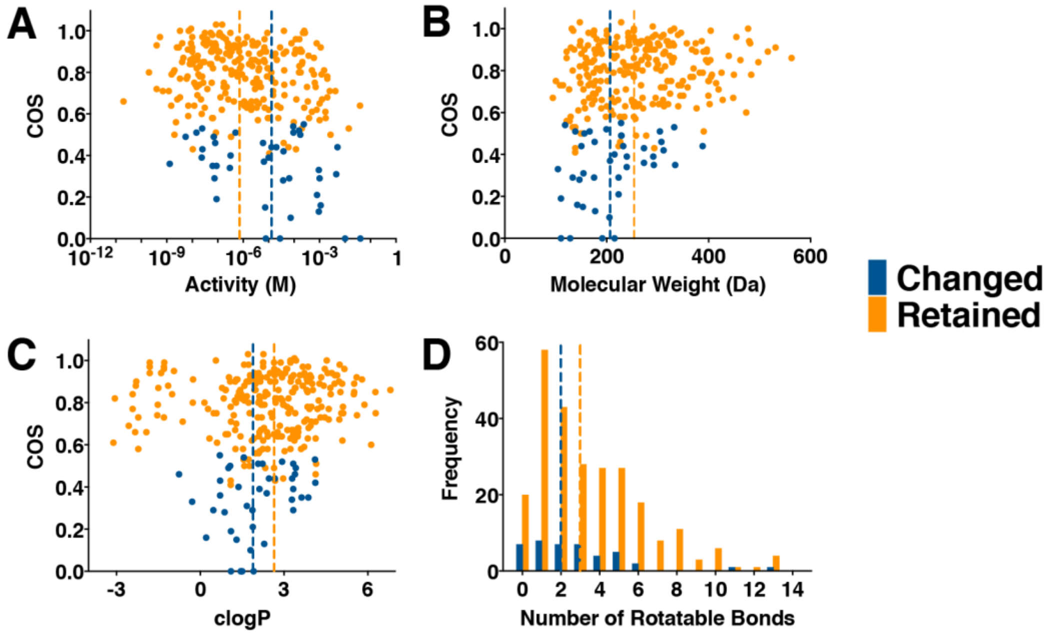


Figure 2: Certain properties of the smaller ligand correlate with increased likelihood of changing binding mode when elaborated. In each case, *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is preserved. 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 ($p < 0.1$).

Given that binding affinity typically becomes stronger with increasing molecular weight (27), 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 ($p < 4 \times 10^{-4}$). 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 ($p < 4 \times 10^{-4}$) (**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 ($p < 0.02$).

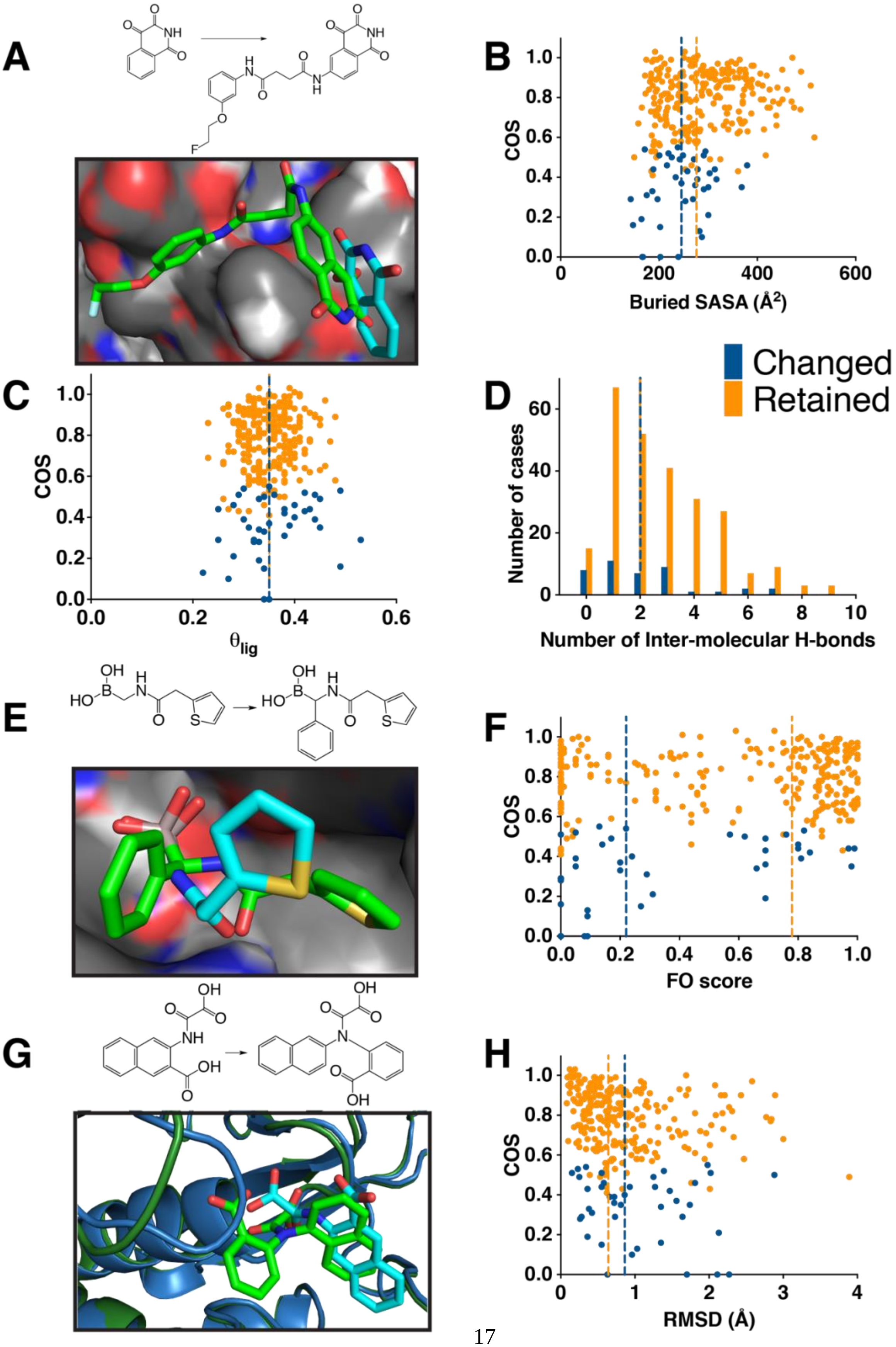
Finally, we anticipated that a ligand's flexibility might also contribute to it's potential for adopting a new binding mode. If a functional group were added to a ligand at a position that is incompatible with the existing binding mode, the opportunity to slightly vary the ligand's internal degrees of 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.

1.2.3 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 (28), 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}) (29); 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 (30), 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.



A
O
NH
O
O
NH
O
H
N
O
O
N
H
O
F
B
1.0
0.8
0.6
0.4
0.2
0.0
COS
0
200
400
600
Buried SASA (Å²)
C
1.0
0.8
0.6
0.4
0.2
0.0
COS
0.0
0.2
0.4
0.6
θlig
D
Changed
Retained
60
40
20
0
Number of cases
0
2
4
6
8
10
Number of Inter-molecular H-bonds
E
OH
HO
B
H
N
O
S
OH
HO
B
H
N
O
S
F
1.0
0.8
0.6
0.4
0.2
0.0
COS
0.0
0.2
0.4
0.6
0.8
1.0
FO score
G
OH
O
O
NH
OH
O
OH
O
O
N
O
OH
H
1.0
0.8
0.6
0.4
0.2
0.0
COS
0
1
2
3
4
RMSD (Å)

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).

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" (31). 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 (32). 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 (12)) from examples in other systems for which the parent compound's binding mode is conserved by the fragment (31).

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 (32) (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 (14). 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 (14), 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.

1.2.4 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**).

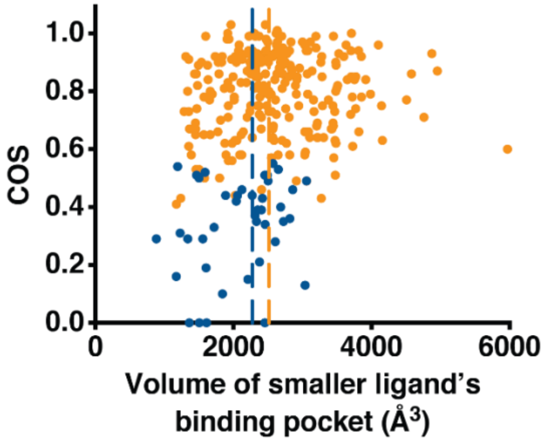


Figure 4: The size of the initial binding pocket correlates with the likelihood of changing binding mode upon elaboration. Here, *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is preserved. Vertical lines 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 ($p < 9\times10^{-5}$).

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" (33, 34). 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.

1.2.5 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 "RMAC" ("*r*msd after *m*inimization of the *a*ligned *c*omplex") (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.

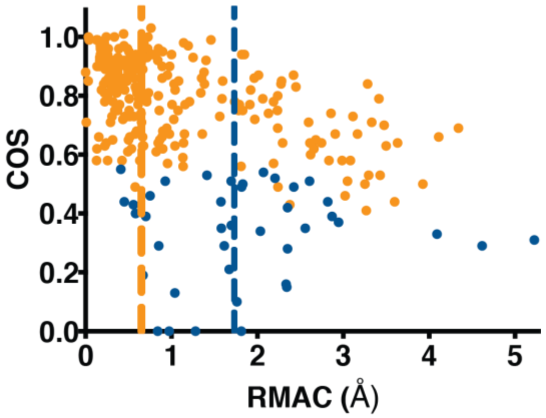


Figure 5: Directly probing whether the larger ligand can be accommodated without changing binding mode. "RMAC" is a measure of whether the structure of the smaller ligand's complex can be used to model the larger ligand: the larger ligand is aligned to the smaller ligand, and then its RMSD is measured 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.

1.2.6 <u>Predicting the presence of alternate binding modes based on these properties</u>

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 (35), 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**).

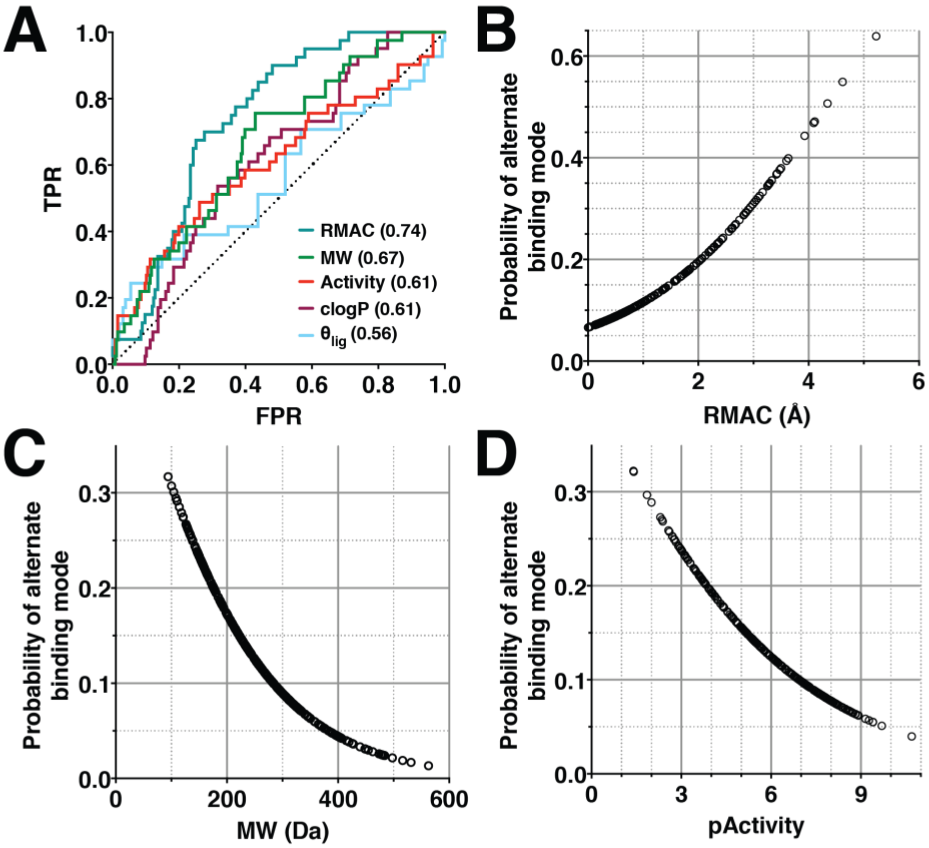


Figure 6: Primary determinants of chemical substitutions that lead to new binding modes.

(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.

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.

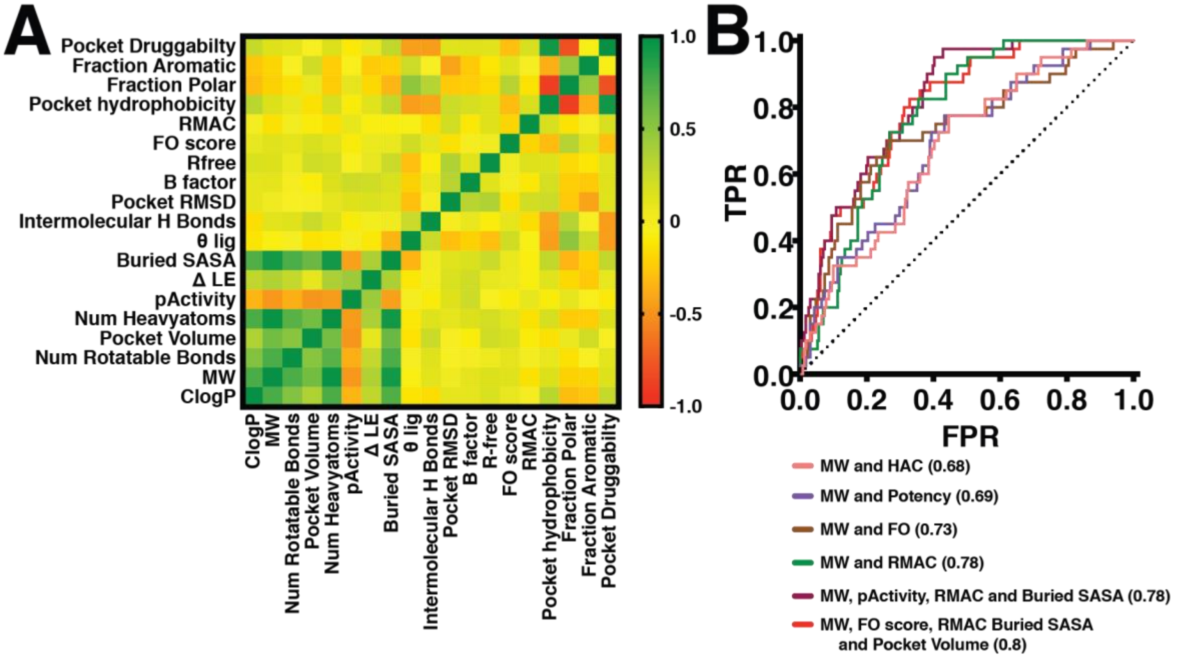


Figure 7: Combining properties leads to a model with more predictive power. **(A)** Correlation in our test set between each of the properties considered: using this color gradient, uncorrelated properties are *yellow*. **(B)** Receiver operating characteristic (ROC) plots comparing the multiple-regression 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.

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.

1.3 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.

1.3.1 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.

1.3.2 <u>Substitutions that enable new interactions</u>

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 (36). The lack of additional contacts leads to dual occupancies observed for the ring, though the positioning of the hydroxyl group is preserved in both (23). 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 (23).

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 (37).

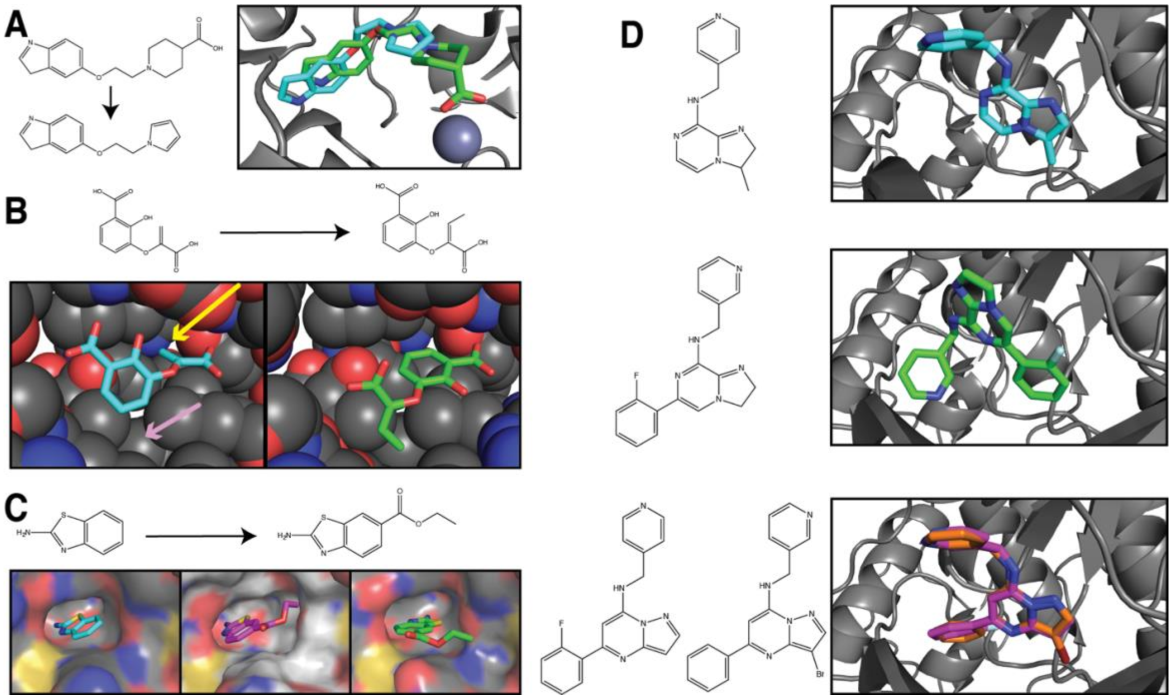

A
N
O
OH
B
HO
OH
O
C
H₂N
S
N
O
D
N
HN
F
Br

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.

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 (38). 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 (39). 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 (40). 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) (41). 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.

1.3.3 Fragments adopting alternate binding modes

The past two decades have been marked by a broad and enthusiastic adoption of fragment-based drug discovery (42). 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 (43). 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) (44), the same authors subsequently refined their recommendation to compounds under 230 Da (45). 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 (43); 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 (14). 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) (46). 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 (47), 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^o 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.

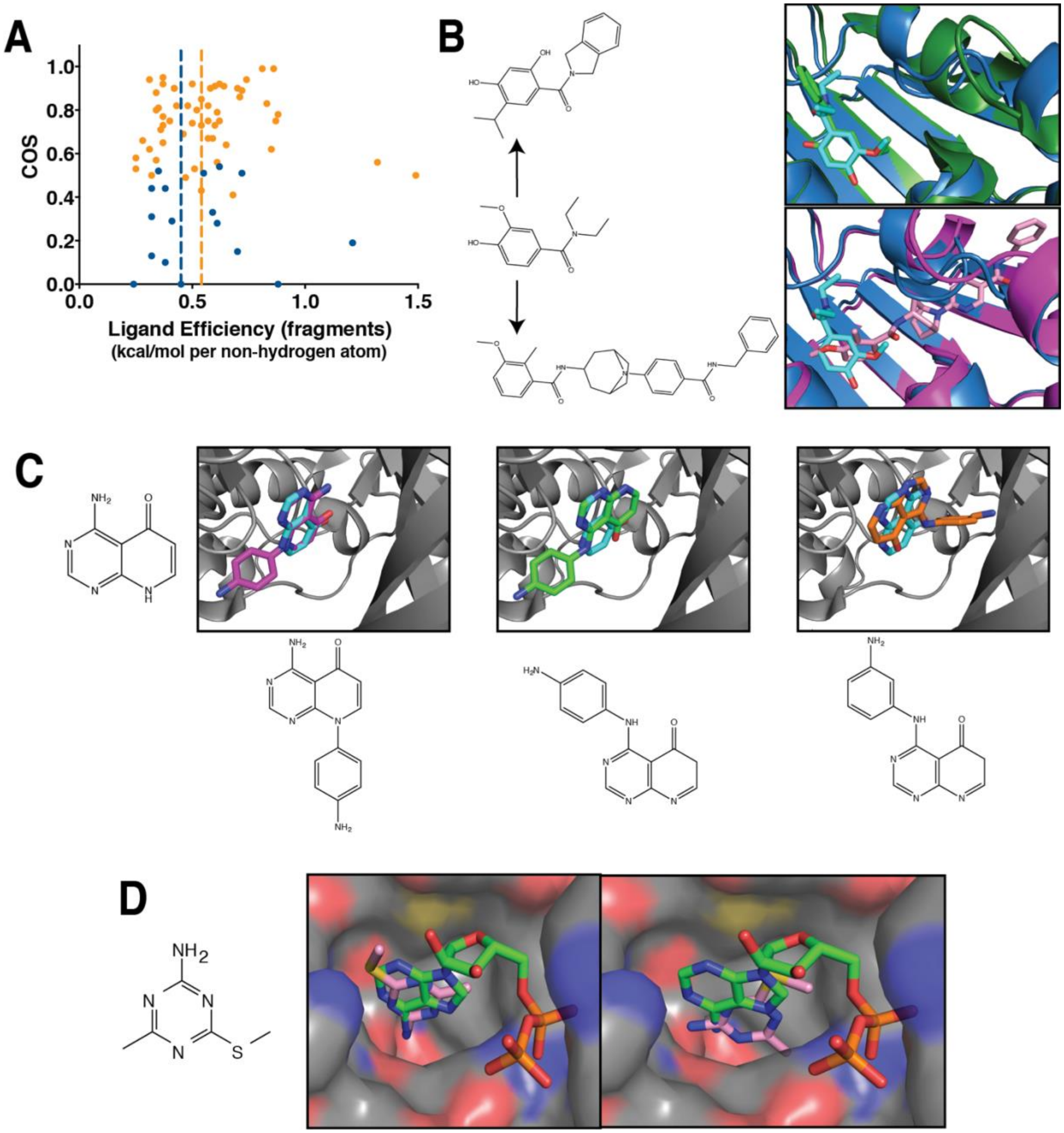


Figure 9: Fragments are particularly prone to alternate binding modes. (A) Among fragment starting points, there is no statistically significant difference in ligand efficiency between those that change binding mode upon elaboration versus those that retain their binding mode ($p < 0.3$). **(B)** Fragment screening for ATP-competitive Hsp90 inhibitors yielded an initial hit (*cyan*, PDB ID 2xdl) 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 4x0m), 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 4x2f), whereas substituting this anilino group at two other positions yielded two more distinct binding modes (*middle*, *green*, PDB ID 4x2g; *right*, *orange*, PDB ID 4x2j). **(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.

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 (48). 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 (49) (**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 (50) (**Figure 9d**).

1.4 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.

1.5 Experimental Section

1.5.1 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) (51, 52). 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 (53) 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 (54). 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:

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

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 (55). 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 (56). 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 (57), 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 (58) 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**).

1.5.2 *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 (17). 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 (18, 19). 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 the "combined overlap score" (COS) as follows:

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

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.

1.5.3 Properties collected for individual complexes

For each ligand in our set, we used OpenBabel (54) to determine the molecular weight and to estimate the octanol-water partition coefficient (clogP). We used OMEGA (59-61) to calculate the number of rotatable bonds. We drew the number of ligand heavyatoms and the potency from the PDBbind database itself (52). 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 (26). The ligand efficiency (LE) was calculated from the potency (K) and heavy atom count (HAC) as follows:

$$LE = -\frac{0.596 lnK}{HAC} \qquad \text{(Eqn. 3)}$$

From the structure of each complex, we used the Rosetta macromolecular modeling suite (62) 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}) (29). 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 (32). 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 (55) 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 (31). 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:

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

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 (31), 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 (33, 34).

1.5.4 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 (62), 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 (55) 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 (63), 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.

1.5.5 Statistical Analysis

We used the Mann Whitney U-test (as implemented in the R statistical computing environment (64)) 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) (35) as implemented in the R statistical computing environment (64).

1.6 Tables

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 (31)	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	
RMSD$_{pocket}$	RMSD difference of binding site residues between the two ligand-bound	0.03	0.60	✓

	structures (Å)			
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	

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.

CHAPTER 2: Numerical Transformation of Random Forest Outputs yield Fruitful Continuous Predictions

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 and [3] Department of Molecular Biosciences,

University of Kansas, 2030 Becker Dr., Lawrence, KS 66045-7534

[*]To whom correspondence should be addressed. E-mail: **john.karanicolas@fccc.edu**, 215-728-7067

Over the past decade, random forest models have become widely used as a robust method for high-dimensional data regression tasks. In part, the popularity of these models arises from the fact that they require little hyperparameter tuning and are not very susceptible to overfitting. Random forests are comprised of an ensemble of decision trees that independently predict the value of a dependent variable; predictions from each of the trees are ultimately averaged to yield an overall predicted value from the forest. Using a suite of representative real-world datasets, we find a systematic bias in predictions from random forest models. Further, we find that this bias is recapitulated in synthetic datasets, regardless of whether or not they include irreducible error (noise) in the data. Here we define a numerical transformation that can be applied to the output of random forest models in order to fully remove this bias. Application of this transformation yields improved predictions across all real world and synthetic datasets evaluated in our study.

2.1 Introduction

Machine learning is a technique aimed at building systems that analyze data, identify patterns in the data, and then make decisions without explicit human intervention. Over the past two decades this field of data science has evolved dramatically, now occupying the position of a highly practical technology with widespread commercial benefit. Its transformative impact is felt most directly in data-intensive fields such as logistics, financial modeling, marketing, cosmology, bioinformatics, social science and many others. Machine learning also powers many of the algorithms underlying our everyday lives, such as credit-card fraud detection, recommendations for online content, image recognition, autonomous vehicle control, and natural language processing (65).

Conceptually, the goal of machine learning is to learn a function *f* that optimally maps input variables *x* ("features" or "attributes") to an output variable *y* ("response variable" or "target variable"), i.e. $y = f(x)$. Machine learning approaches vary greatly, both in how they algorithmically represent *f* (e.g., decision trees, mathematical functions, and general programming languages) and in how they optimize the parameters intrinsically contained within *f*. There are many factors that govern the selection of an appropriate machine learning algorithm for a particular problem, such as the nature of the available training data (including the size and quality of available data) and the ultimate objective for the model. Examples of differing objectives can include finding patterns in a dataset (unsupervised learning), versus seeking to predict a specific output value from new data (supervised learning) (66).

Supervised learning problems can be further divided into classification and regression problems. In both cases the goal is to build a model that predicts the value of some dependent attribute from a collection of input variables; the difference is that classification problems have a categorical (discrete) target variable, whereas regression problems have a continuous target variable. Success in a classification problem is thus gauged by the accuracy of assigning new data points with the correct label, whereas success in a regression problem is gauged by how closely the predicted outputs match their true values.

Random forest models have emerged as a popular and robust method for high-dimensional complex data. In the field of computational biology, for example, early random forest models proved especially effective for using variations in gene sequence to predict a continuous phenotype or clinical trait (e.g. disease state (1), viral replication capacity (3), emergence of resistance (4), or necessary dose of a drug (2)). More recently, models have been developed for predicting protein's expression and solubility (5), predicting proteins that will interact with one another (6), and predicting the biological relevance of protein-protein interactions (7).

A random forest model is comprised of a collection of decision trees. Starting from the root, each (non-leaf) node of a decision tree compares the value of the current features to reference values obtained during training. These comparisons generate a path to a specific node, which in turn contains the predicted target value: in a classification tree these values are discrete labels, whereas in a regression tree these values are continuous predictions of the output value. Whereas individual decision trees often overfit their training data (67), random forests avoid this problem by building many trees that each use a subset of the training data (bootstrap aggregating, aka "bagging") and a subset of the available features ("feature bagging"). In a regression problem the output from all trees is averaged, and this value is returned as the random forest's predicted value for the target variable.

Through development of random forest models for a variety of different regression problems, we have observed that the resulting models were inevitably too conservative with their predictions. Here we demonstrate that this is a systematic pathology of random forest regressors, and we show that it applies even to artificial data for which the target variable is trivially calculable from the features. Through these studies we develop a numerical transformation that can be applied to the output values from a random forest regressor, and we show that this improves the accuracy of the model's predictions.

2.2 Methods

All calculations were carried out using the R statistical computing environment (68) (v3.4.2).

2.2.1 Publicly-available regression datasets

While Kaggle (66) does provide a source for relevant datasets, their licenses do not always allow the data to be used outside of the competitions. We therefore collected seven other standard datasets from the UC Irvine Machine Learning Repository (65), from De Cock (69), and from Lantz (70). Each of these multivariate datasets are intended as regression tasks, i.e. prediction of a continuous output variable. These datasets are:

1. Airfoil Self-Noise Dataset (from UCI): This dataset entails predicting the sound pressure level for various airfoils, based on their physical properties and those of the wind (71-73). There are 6 features and 1503 data points.

2. Concrete Slump Test Dataset (from UCI): This dataset entails predicting the compressive strength of concrete, based on its ingredients (other output variables were also available, but were not used in our study) (74-78). There are 7 features and 103 data points.

3. Bike Sharing Dataset (from UCI): This dataset entails predicting the number of rental bikes in Washington DC on a given day, based on the weather and season (79). We pre-processed the data by converting all categorical variables into factors (using the factor function R package base), yielding a dataset with 11 features and 731 data points.

4. Combined Cycle Power Plant Dataset (from UCI): This dataset entails predicting the output from a power plant, based on the weather (80, 81). There are 4 features and 9568 data points.

5. Online Video Characteristics and Transcoding Time Dataset (from UCI): This dataset entails predicting the transcoding time for YouTube videos, based on the input and output videos'

characteristics (82). The dataset contains 168,286 data points, of which we selected only those with transcoding time less than 10 secs. This yielded a dataset with 20 features and 50,945 data points.

6. Ames Housing Dataset (from De Cock): This dataset entails predicting house sale prices in Ames IA (an updated version of the classic Boston dataset), based on 79 features that describe the houses (69). Because the dataset contained missing values for many feature variables, we employed some exploratory analysis for the feature space using R package ggplot. We used R package base to find variables with missing values and converted all categorical variables into factors. Using the R package stats we computed the Pearson correlation coefficient for all 79 features with one another as well as with the target variable. We then removed features that contained missing values or that had no correlation with the target variable. This preprocessing yielded a dataset with 16 features and 1460 data points.

7. Insurance Cost Dataset (from Lantz): This dataset entails predicting the medical insurance costs billed by health insurance for a set of individuals, based on the peoples' physical and geographic attributes (70). We curated the dataset by removing outliers corresponding to the costliest conditions (those costing more than $16,000). This preprocessing yielded a dataset with 6 features and 1070 data points.

All pre-processing that led to the datasets used in this study are described fully in the *Supplemental Methods* section.

2.2.2 Synthetic regression datasets

In order to clearly observe artifacts of the random forest models we sought to study, we also generated synthetic datasets that would be free of subtle systematic errors potentially lurking in the real-world datasets.

We generated a noise-free dataset, by defining the target value as a linear combination of eight features, *A-H*:

$$Target = 2A + 3B + 4C + 5D + 6E + 7F + 8G + 9H \qquad (1)$$

We populated this dataset with 50,000 points, using values for the features drawn from a normal distribution (μ=0, σ=1). In this case, the target value is completely determined by the eight features that are presented to the model.

We use the same approach to build a second dataset, this time defining the target value as:

$$Target = 2A + 3B + 4C + 5D + 6E + 7F + 8G + 9H + n_1 + n_2 + n_3 \qquad (2)$$

In this case the target value now additionally depends on three variables (normally distributed with μ=0, σ=1) that are not included among the features *A-H* that are presented to the model. Thus, the target value is no longer fully determined by the features (i.e. the model includes irreducible error), making this model correspond more closely to a real-world scenario.

2.2.3 Building random forest models

For datasets comprised of less than 10,000 points (i.e., all except the synthetic datasets and the UCI Online Video dataset), 80% of the points were randomly assigned to the training set and the other 20% comprised the test set. For datasets with more than 10,000 points, we randomly assigned the data to training set (60%), validation set (20%), and test set (20%).

All random forest models described in this study were built using R’s randomForest package (v3.4.2) (83), which implements Breiman’s algorithm (84). This implementation uses two key adjustable parameters. The parameter **ntree** (the number of trees to include in the random forest model) was set to

500 for models trained with less than 15,000 points, and to 1000 for models trained on larger datasets. The parameter **mtry** (the number of candidate features that are considered when building a given split into the component decision trees) was set to 1/3 of the total number of features (the default value) in all cases. The minimum size for terminal nodes (**nodesize**) was set to its default value of 5, and the maximum number of terminal nodes (**maxnodes**) was set to its default value such that trees are grown to their maximal extent possible.

2.2.4 Numerical transformation applied to model outputs

As described in the *Results* section, we observed that random forest predictions showed characteristic curve shapes that resembled the inverse of a sigmoid function. Given the classic logistic function (shifted so that it passes through the origin):

$$y = \frac{1}{1+e^{-x}} - \frac{1}{2} \tag{3}$$

We inverted this function to obtain:

$$y = -\log\left(\frac{1}{x+{}^{1}/_{2}} - 1\right) \tag{4}$$

We additionally tested other functional forms with similar curve shapes, in particular:

$$y = sinh(x) = \frac{e^{x}-e^{-x}}{2} \tag{5}$$

And:

$$y = tan(x) \tag{6}$$

In each case we included 4 free parameters in order to fit the curves: a linear scaling and an offset for *x*, and a linear scaling and an offset for *y*. In the latter case, for example, the following gives the full fitting equation (where *a*-*d* are the fitting parameters):

$$y = d * tan\left(\frac{x-b}{a}\right) + c \tag{7}$$

By analogy, in the case of the inverted logistic function (Eqn. 4) we use the full fitting equation:

$$y = -d * \log\left(\frac{1}{\frac{x-b}{a} + {}^{1}\!/_{2}} - 1\right) + c \tag{8}$$

All fitting was carried out using GraphPad Prism (v8.0.2). The outputs from applying the random forest model to the data in the training or validation set were fit to the known (ground truth) values for this set, then the corresponding fit was used to transform the model outputs generated for the data in the test set.

2.3 Results

2.3.1 *Random Forest regression models give overly conservative predictions*

Using each of seven standard regression datasets we first processed the datasets to remove points with missing features and remove any highly co-correlated features (see *Methods*). For each dataset, we then trained a random forest regression model using standard techniques (see *Methods*), and then applied this model to predict the values of the target variable for points in the test set.

Results from this first experiment are presented as **Figure 1**. In all seven cases the resulting models show good performance, yielding useful predictions of the target variable when applied to the test set. Unsurprisingly, the specific quality of the models varies somewhat, due to the size of the training set and the extent to which the features fully caption the variation in the target variable. That said, in all seven cases – for each of these very diverse datasets – predictions are found to contain common artifactual characteristics. In particular, while each model's predictions are indeed correlated with the ground truth values of the target variable, predictions for values far from the mean tend to be too "conservative". In other words, all seven models tend to make predictions that are overly close to the mean, which in turn leads to slopes greater than 1 when the data are plotted as in **Figure 1**. The same artifact is also evident when examining points in the dataset furthest from the middle of the distribution, because it reduces the range of prediction values. In Airfoil Self-Noise dataset (**Figure 1a**), for example, the ground truth (desired) values range from 104 to 140 dB but the model only predicts values between 110 and 132 dB.

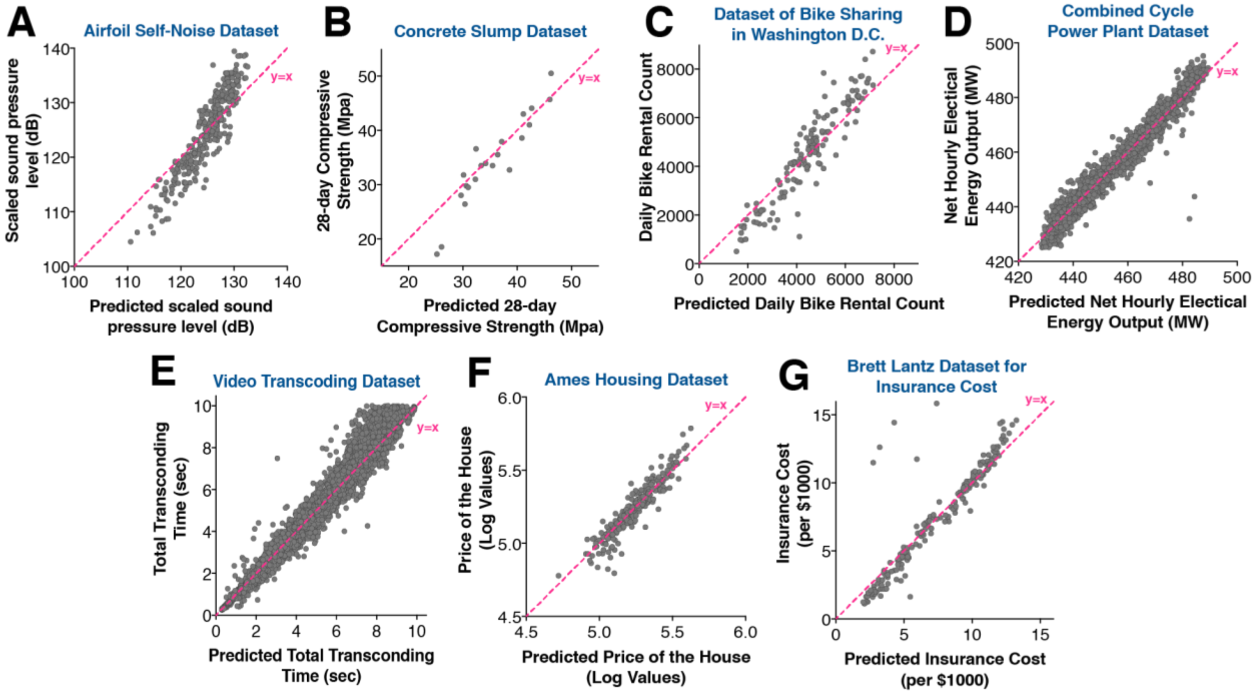


Figure 1: Regression predictions from random forest models built for seven publicly available datasets. Real-world datasets were obtained from UCI Machine Learning Repository (A-E), De Cock (F) and Brett Lantz's Machine Learning in R (G). In all cases the pink dashed line represents the identity line (y=x).

In order to better explore the basis for this behavior, we created a series of synthetic datasets (see *Methods*). Because we could fully control the generation of these sets we could guarantee that no potential problems with the datasets themselves were producing this artifact: we could directly ensure by construction that the features are entirely uncorrelated with one another, and also that the training and test sets are drawn from the same distributions. Further, we could also make the datasets arbitrarily large, and thus exclude artifacts that may stem from a lack of training data. Each model includes eight features with values drawn from a normal distribution; the target variable for each point in the set corresponds to a linear combination of its features, optionally with an additional (random) contribution that is not explained by the features to mimic a real-world scenario.

We trained a random forest regression models for this synthetic dataset (**Figure 2**), both using a target variable that is fully calculable from the features and using a target variable that includes irreducible error (variation that is not explained by the features, i.e. noise). Surprisingly, both synthetic datasets exhibited the same behavior observed in the real-world datasets examined earlier. With the ability to probe larger datasets, the precise shape of this artifactual behavior becomes even more apparent: these plots are not simply linear with a slope greater than 1, but rather have a non-linear "S" shape that resembles the inverse of a sigmoid function.

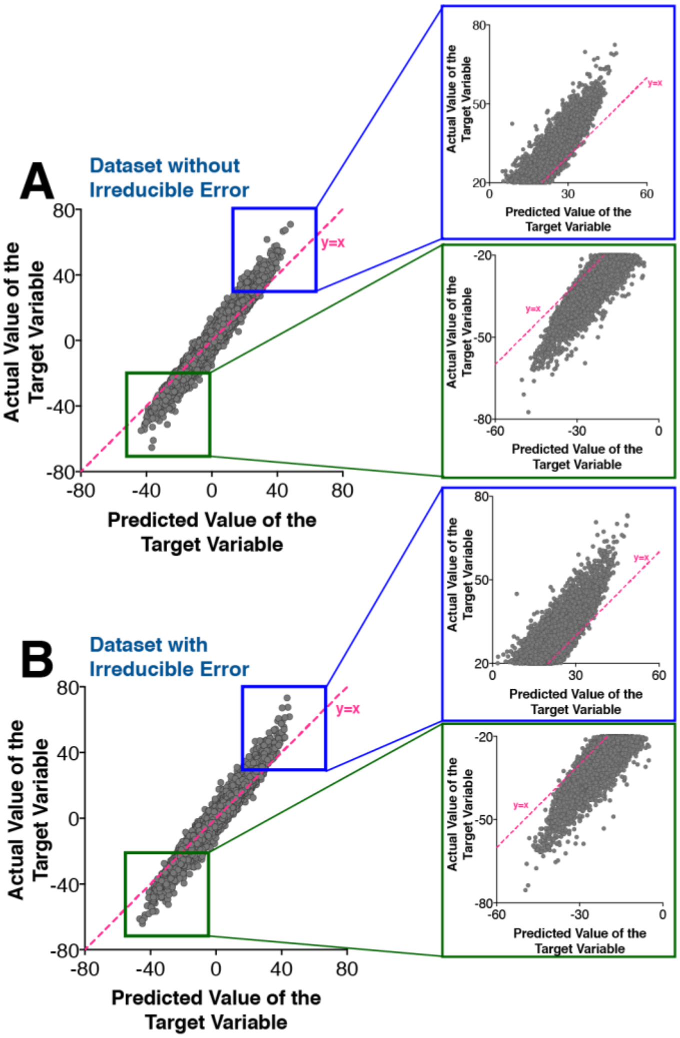


Figure 2: Prediction of target variable from Random Forest Models for two synthetic datasets. Large synthetic datasets were used in which **(A)** the response variable is fully described by the features (i.e. no irreducible error), or else **(B)** an additional noise term is included in the response variable that is not encode by the features. In all cases the pink dashed line represents the identity line (y=x).

To confirm this non-linearity we fit the curves with a least-squares regression line and applied the Wald–Wolfowitz runs test (85): using synthetic test sets of size 100,000 in which irreducible error is either present or absent, both have statistically significant deviation from linearity, with $p<0.02$ in both cases.

The popularity of random forest models derives in part from the ease of building robust models, due to the relative dearth of adjustable parameters. Indeed, it is typical to vary only two parameters when building models: the number of trees **ntree**, and the number of features to consider when split at each node of the component decision trees **mtry** (the feature that best explains the variation in the node's data is then selected from among these choices). Of these, increasing **ntree** indefinitely comes at computational cost, but has been shown not to affect model performance beyond a certain point (86-88); knowing that there is no model performance downside in including a large number of trees in our study, each of the models presented here uses a large number of trees (in results not shown, we have ensured that we are beyond the point at which additional trees would improve the models). The model's response to **mtry**, on the other hand, can be important (especially when features are correlated with one another) and its optimal value is not straightforward to predict (89); for this reason, the value of this parameter is typically selected through cross-validation.

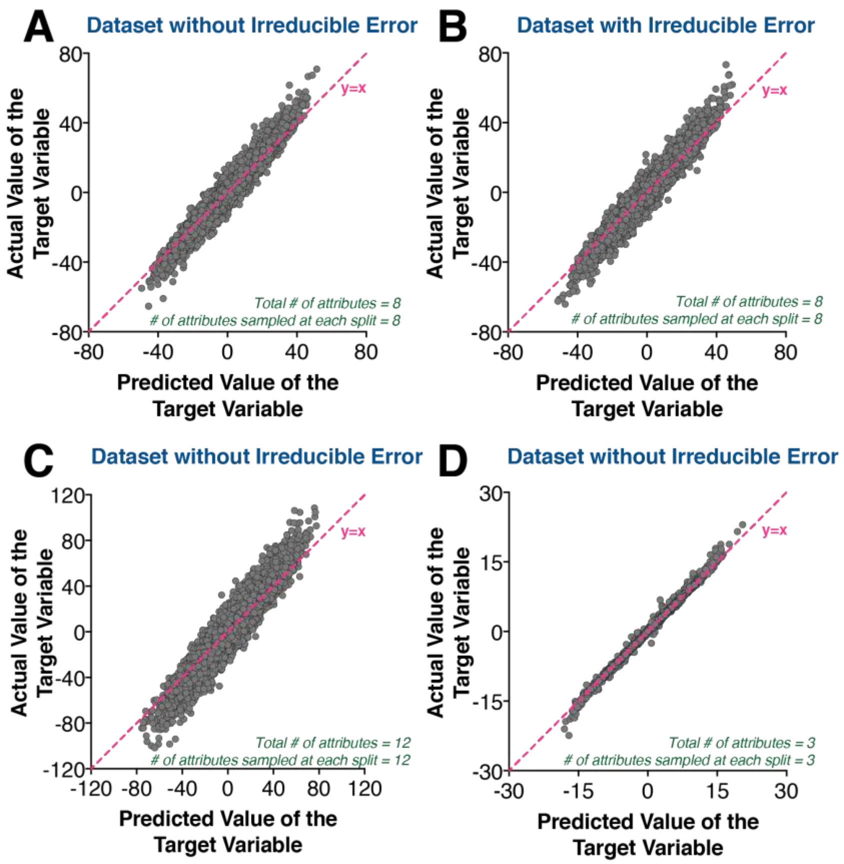


Figure 3: Tuning parameters of the random forest model. When the all the attributes are used for splitting decision trees during training of random forest model (i.e. feature bagging is removed), the same pathology is observed. This is true irrespective of whether irreducible error is **(A)** absent or **(B)** present in the dataset, and still holds as the number of features is **(C)** increased or **(D)** decreased.

From the default value used earlier (1/3 of the total number of features), we therefore increased **mtry** to match the total number of features. This strategy eliminated variation between the underlying decision trees associated with the order in which features were selected ("feature bagging"), and retained only the variation between trees arising from the fact that individual trees are each built using a subset of the training data ("bagging"). Ordinarily this could increase a random forest model's susceptibility to overtraining or to artifacts from correlations between features, but the size of the datasets and the purely orthogonal nature of the features in our synthetic datasets ruled out such concerns. Through this experiment we confirmed that the same pathology observed earlier persisted even when all features are made available at every stage of tree-building (**Figure 3ab**). We further found that increasing the number of features that contribute to the target value made this behavior slightly more pronounced (**Figure 3c**), whereas decreasing the number of features to a nearly trivial problem reduced – but did not eliminate – the non-linearity of the predictions (**Figure 3d**).

2.3.2 The use of boosting eliminates this artifact

Given that the response variable in our synthetic datasets was simply calculated as a linear combination of the features, we sought to ensure that purely linear regression models should serve as effective methods for these datasets, particularly in the absence of irreducible error (in which case multiple linear regression should yield a solution complete free of error). To confirm this hypothesis, we applied both classic multiple linear regression (**Figure 4a**) and a non-parametric method that generates predictions using a series of linear splines (MARS, Multivariate Adaptive Regression Splines (90)) (**Figure 4b**); in both cases, as anticipated, the resulting models yielded perfect predictions when applied to the test set.

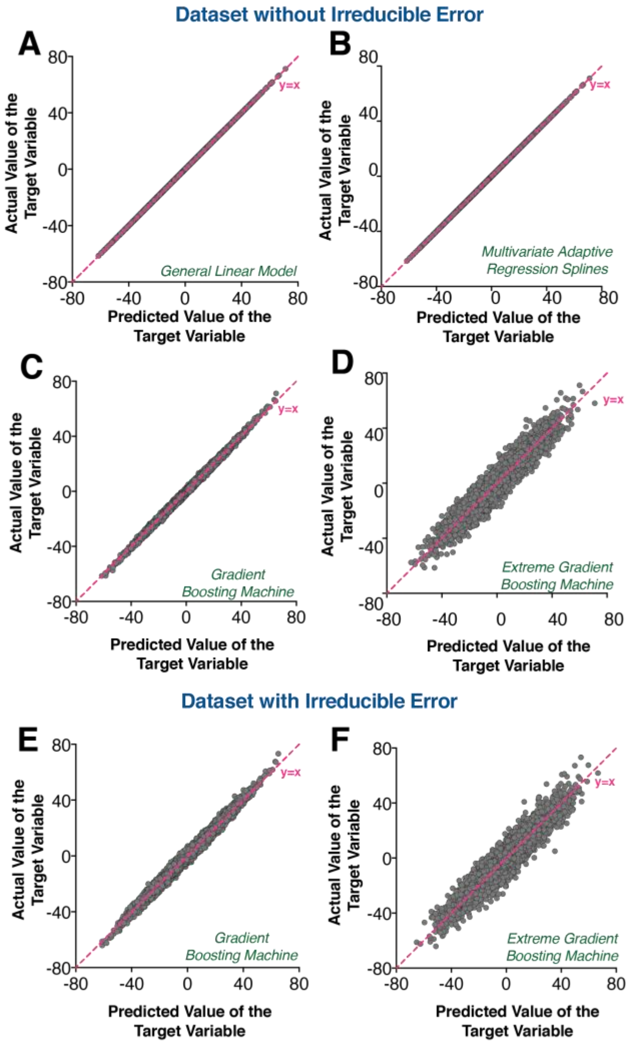


Figure 4: This artifact is absent in predictions from other regression methods. Using our (linear) synthetic dataset, we find that methods build on linear models fit the test set data exactly. Linear models examined were: **(A)** multiple linear regression and **(B)** multivariate adaptive regression splines. We next tested two implementations of tree-based models that use gradient boosting **(C,E)** GBM and **(D,F)** XGB. Neither model led to test set output with the previously-observed pathology, whether applied to (**C,D**) a dataset with no irreducible error, or (**E,F**) a dataset that included irreducible error.

An important class of random forest models do not build ensembles of independent decision trees, but instead build trees sequentially using a technique known as "boosting". In this approach, a single decision tree is first built to explain the training set data. The residuals of the predictions from this tree (from the training set data) are then fit to a new decision trees, which is added to the model. These steps are iteratively repeated, building up an ensemble of trees that sequentially and collectively reduces the residuals (error) in the overall model's recapitulation of the training set data.

Given the persistent and systematic pathology present in the random forest models, we speculated that a boosting approach may recognize – and eliminate – this bias through its presence in the residuals of early models. To test this, we applied to our synthetic dataset two slightly different implementations of gradient boosted machines, from the GBM (**Figure 4c**) and XGBoost (**Figure 4d**) packages. It was immediately apparent that both boosting methods fit this synthetic data, with no hint of the previous-observed artifact. Even upon addition of irreducible error, both GBM (**Figure 4e**) and XGBoost (**Figure 4f**) yielded useful models.

2.3.3 Numerical transformation of predictions leads to improved accuracy

If the observed pathology in these predictions stemmed from some form of overtraining or from some difference between the training and test sets, then the same pathology would not be observed when the model was applied back to the data in the training set. In contrast, if this artifact indeed arises as a general property of the random forest model itself, then this pathology *will* be observed when the model is applied back to the data in the training set. The fact that boosting eliminated this artifact implies that the pathology can be quantitatively identified (and thus corrected) by applying the model to the training set data.

Starting with our synthetic data sets, we first confirmed that indeed the same characteristic behavior is observed when the model is applied to data from the training set (**Figure 5a**). Further analysis showed that the nature of this artifact was quantitatively the same in the training set and in the test set: this led us to anticipate that a numerical transformation could be developed using the training set, and then applied to correct the predictions in the test set. Ideally, our hypothesis regarding the origin of this artifact could be used to develop an analytical formulation that quantitatively explains this behavior and provides a means to correct it. We have not been successful in developing such an approach, however; the simple algorithmic procedure behind building the random forest is not readily translated to mathematical modeling, and indeed studies seeking to analytically explain or optimize specific behaviors of random forest models are still in the early stages of progress (91).

Given the characteristic curve shape in each of the plots presented here, we attempted to fit these data with several different functional forms: y=tan(x), y=sinh(x), and the inverse of a classic logistic function (see *Methods*). We applied each functional form to the random forest predictions generated using synthetic data, in each case using a fit that includes four free parameters (corresponding to a linear scaling and an offset for x, and a linear scaling and an offset for y). While all three approaches yield very similar curves with reasonable fits (**Figure 5b**), we found that the inverse logistic function fit the curve shape very slightly better than the other two options.

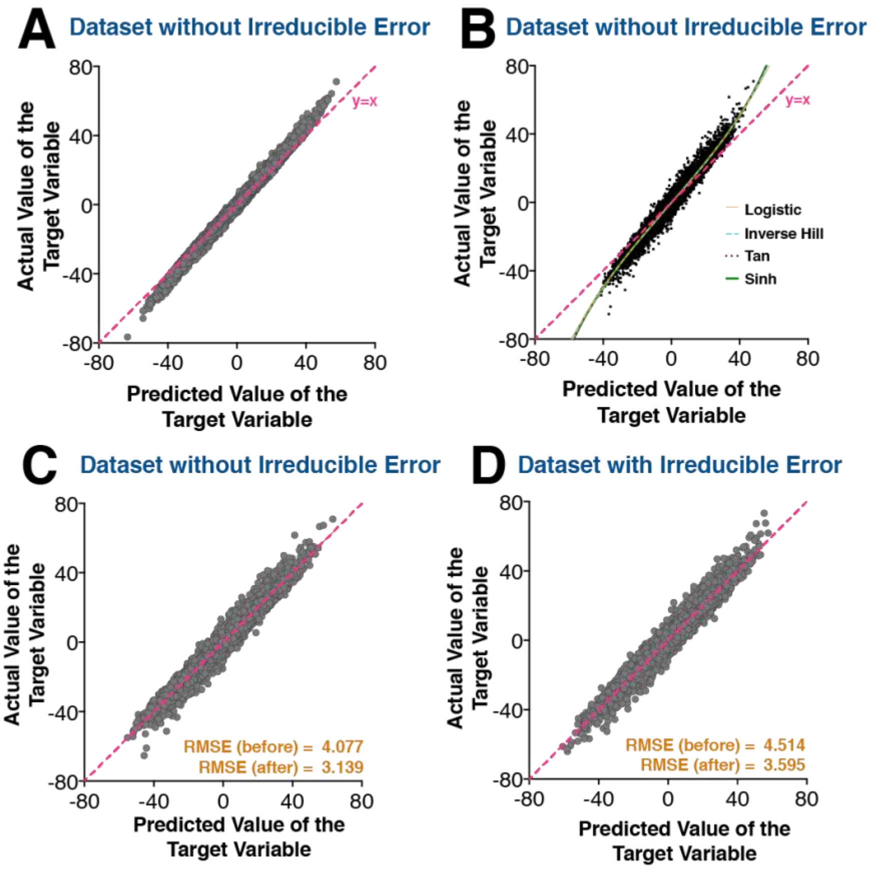


Figure 5: Developing a transformation to remove the observed bias. A: After training a random forest model, the model was applied back to the data in the training set (rather than to the test set). The same previous-observed pathology persists here, suggesting that a numerical transformation can be developed from the training data and later applied to the test set. **B:** Several alternate functional forms were evaluated (see *Methods*), all of which yielded suitable fits to the underlying data. **C:** Fitting the transformation from the training data then applying it to test set data (in the absence of irreducible error) removes the previously-observed pathology and leads to improved predictions as gauged by root-mean-square error (RMSE). **D:** In the presence of irreducible error, again the transformation can be effectively developed using the training data and applied to the test set. In all cases the pink dashed line represents line of identity (y=x).

We envisioned that the inverse of these curve fits obtained from the training data could then be applied to each prediction in the test set, as means to undo the effect of this underlying artifact present in the model. Importantly, the fit parameters are obtained using only the training data, and are thus determined without any knowledge of the data in the test set; simply put, the training data is simply used to develop an additional “post-processing” function that is applied to each prediction made by the model. To explore this approach, we used the same random forest model trained on this synthetic dataset, and applied these fitting parameters to transform predictions for the test set data: gratifyingly, the non-linearity observed previously was no longer present, and points were distributed evenly above and below the identity line (**Figure 5cd**). Most importantly, the overall accuracy of predictions was improved by removing this artifact: relative to the ground truth value of the target variable, the root-mean-square error (RMSE) of the predictions was reduced once this transformation had been applied.

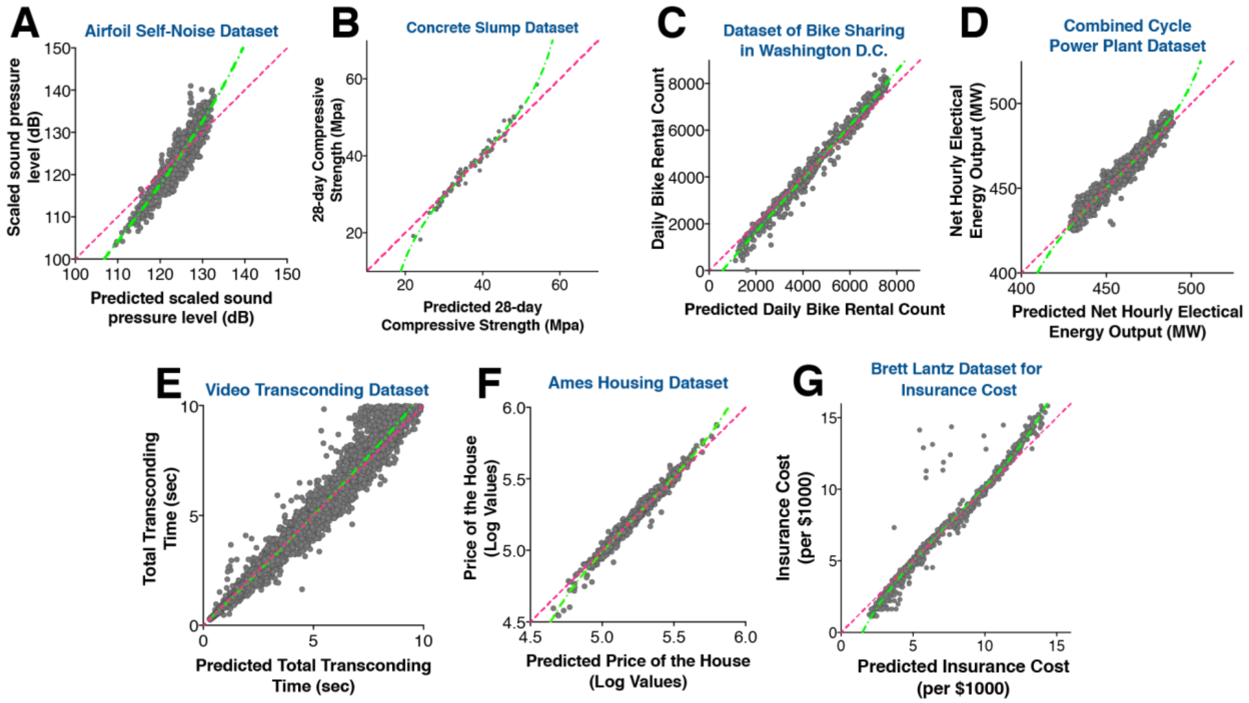


Figure S1: Determining the fit parameters for the inverse logistic function to be used in transforming predictions. For each of the 7 real-world datasets in our study, a random forest model was trained then used to generate predictions for the data in the training set. As expected, the same systematic bias was observed in this experiment. These data were then fit using an inverse logistic function (*green*), and this fit was used to transform output from applying the model to the test set data (presented in **Figure 6**). In all cases the pink dashed line represents line of identity (y=x).

Next, we applied this strategy to each of the seven "real-world" regression datasets presented earlier. By applying the random forest model back to the data in the training set, in each case we fit an inverse logistic function that captured the artifact present in each random forest model (**Figure S1**). When the inverse of this function is now applied to the previously-collected data from the test set (**Figure 1**), we find that the characteristic pathologies are no longer present (**Figure 6**): points are now distributed evenly above and below the identity line, with no evidence of non-linearity, and the range of output predictions matches the range of ground truth values. The overall accuracy is also improved, as measured by the root-mean-square error (RMSE) across all the predictions in each dataset (**Table 1**). Adjusting the slope of the predictions to better match those of the ground truth values – making the random forest predictions less conservative – affects many of the predictions, and is thus expected to account for the majority of the improvement (relative to removing the non-linearity, which affects only the most extreme points). Unsurprisingly, the extent to which the accuracy is improved can be explained by the degree to which the original predictions suffered from this artifact: the airfoil self-noise predictions were dramatically improved, whereas the Ames housing predictions were only marginally affected. Nonetheless, the RMSE was reduced in every case, implying that this strategy should be applied in all cases where a random forest model is being used to predict the value of a continuous target variable (i.e. regression problems).

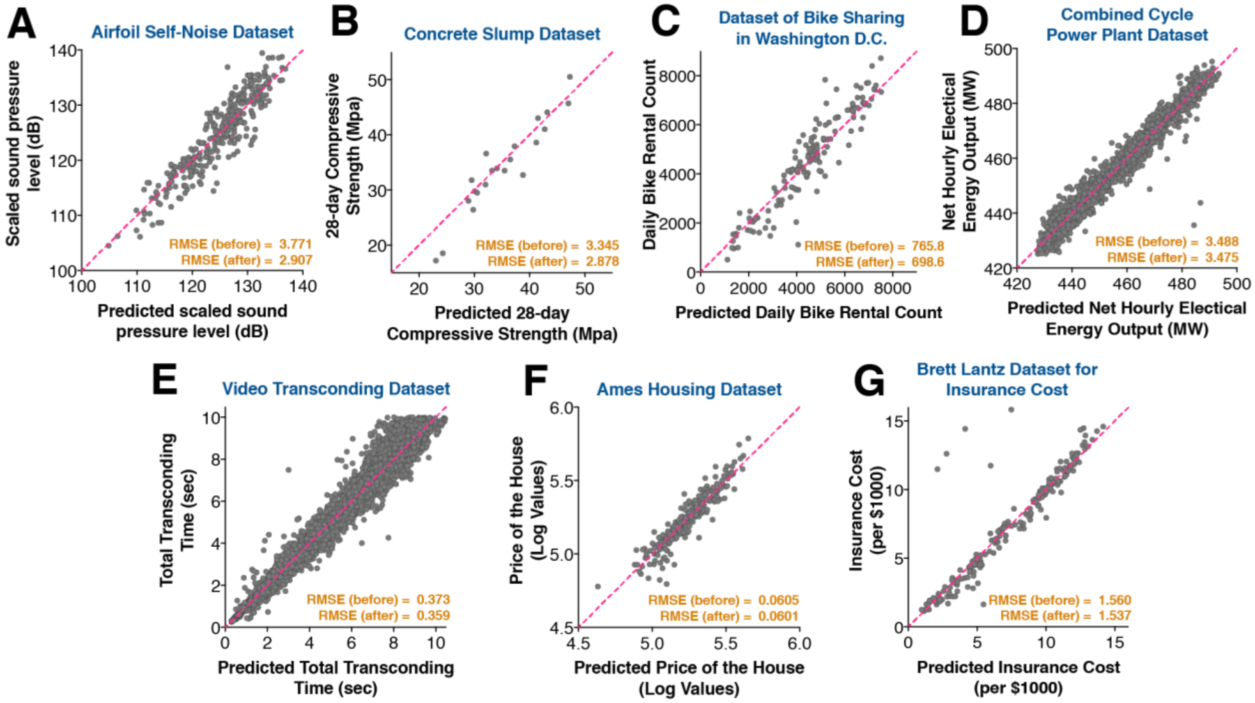


Figure 6: Updated test-set regression predictions using transformations obtained from the training data. For each of the 7 real-world datasets in our study, applying a transformation that corrects the training data removes systematic bias in the test set predictions that was observed earlier (**Figure 1**), thus leading to improved predictions.

Dataset	Original RMSE	RMSE after transformation
Airfoil self-noise	3.77	2.91
Concrete slump test	3.35	2.88
Bike sharing	766	699
Combined cycle power plant	3.49	3.475
Video transcoding	0.373	0.359
Ames housing	0.0605	0.0601
Insurance cost	1.56	1.54

Table 1: Effect of transformations obtained from the training data. For each of the 7 real-world datasets in our study, applying a transformation that corrects the training data leads to improved predictions in the test set (as determined by RMSE before/after transforming the data).

2.4 Discussion

Over the past couple decades, machine learning has established itself as a mainstream technology tool for companies and a great aid for science and research. Its application can be seen in GPS-based traffic predictions, online transportation networks, video surveillances, spam filtering and computational biology. Spurred by Kaggle and Cortana Intelligence competitions, but especially by the meaningful consequences of using machine learning for real and important problems, multiple approaches are often tested for a given dataset and have clearly demonstrated that no single method is optimal for every problem.

Random forest models remain popular not only for their ease of use, but also for their relative insensitivity to outliers and resistance to overtraining. Here, we have identified a systematic bias present in regression output from random forest models; this bias persists even in synthetic datasets, and even in the absence of irreducible error. Because the same bias is present when the random forest model is applied to the training set, we have demonstrated that a correction can be developed from the training set and subsequently be applied to predictions of the test set data.

Interestingly, we have also shown that boosting models do not suffer from the same pathology: by applying models back to the training set and iteratively correct their systematic errors, such classes of models are immune to systematic biases such as the one we demonstrate here. That said, the same self-correcting strategy that allows such models to avoid this bias also contributes to their well-established tendency to focus on outliers in the data and overfit small datasets. Because of this, and because of the relative dearth of hyperparameters that require tuning in separately validation sets, random forest models (without boosting) remain a preferred choice in scenarios in which huge datasets cannot be provided, such as many computational and structural biology applications.

Here we have shown that correcting random forest predictions using a transformation “learned” from the training set data can yield improved predictions in real-world examples. In particular, we find dramatic improvements in some examples, and little improvement (but no negative impact) in other examples (**Table 1**). For this reason, we advocate for the use of this approach as a standard post-processing step in development of any random forest regression model.

CHAPTER 3: Quantitatively Predicting Druggability of a Protein Surface

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 and [3] Department of Molecular Biosciences,

University of Kansas, 2030 Becker Dr., Lawrence, KS 66045-7534

[*]To whom correspondence should be addressed. E-mail: **john.karanicolas@fccc.edu**, 215-728-7067

Protein druggability is an important consideration in the target identification and validation phase of drug discovery. Unfortunately, experimentally determining druggability of a protein target is quite difficult: typical methods include using hit rates from NMR fragment screening or other biochemical screening as a measure of druggability. Complementary computational methods based on pocket features, ligand features, or a combination of the two have also been used to give a qualitative sense of druggability for a protein target. Here, we report a model that seeks to predict the "attainable binding affinity" for a given binding pocket on a protein; this model relies on 13 physiochemical and structural features calculated using the protein structure. To train and then test the model, we have developed a benchmark experiment in which we seek to predict the binding affinity of known ligands, from the structure of the ligand-bound protein without knowing any information about the ligand. While details of the ligand must inevitably contribute to the ground truth (experimentally determined) binding affinity, we nonetheless find that in a leave-on-out experiment our model quantitatively recapitulates the activity for 230 different

protein (848 ligand binding pockets) with correlation coefficient 0.57. When applied to a set of proteins that change conformation upon ligand binding (cryptic pockets) we find that the model marks the bound protein structures as more druggable than the unbound structures. Through a series of representative examples, we can rationalize the basis for the model’s predictions and also explain cases in which the observed binding affinity is likely dominated by properties of the specific ligand rather than by features of the binding pocket. In conclusion, this model offers a means to quantitatively identify and interpret binding pockets likely to support binding of high-affinity drug-like ligands, either from crystal structures or from simulations that sample alternate conformations of proteins.

3.1 Introduction

Target validation plays a pivotal role in the success of drug discovery projects. Traditionally it tries to assess whether or not alteration of the normal activity of a potential target can have some significant therapeutic effect. The druggability concept adds a structural dimension and evaluates the likelihood that small drug-like molecules can bind a given target with sufficient potency to alter its activity. These small molecules bind to preferred sites of action, which are in the majority of cases pockets located at proteins surface. Such ligand binding sites are known to have significant difference in terms of structure and composition as compared to rest of the protein surface. Research groups have thus been developing tools that compute a ‘druggability index’(92) for this sites. Druggability of a protein target can directly be assessed via biochemical screen of small molecules libraries against it. A target is druggable if it binds multiple structurally diverse hit molecules. NMR-based screening that employs 2D heteronuclear correlation to identify ligand binding has a potential to screen large fragment libraries in less time(92). A further improvement to this was when instead of performing an NMR screen for every new potential target, characteristics of sites are mapped with the sites tagged druggable after NMR screen in a similar fashion(93). For instance, an index termed ‘propensity for ligand binding’(94) has been developed based of the specific amino acid composition surrounding the concavities on the surface of proteins. There is another type of druggability score that has been developed and implemented in SiteMap(95). The druggability of targets has also been addressed through simplified shape descriptions of the binding site (e.g. the hydrophobic nature and an additional term taking into account the concavity of the pocket, among others)(96). More advanced models that use several global properties, describing the size, shape, depth as well as biophysical and chemical features of the predicted pockets are gathering more attention recently for their better prediction accuracy(97, 98).

All the methods mentioned above provide a confidence score-based classification as “druggable” or “less druggable” targets. A different approach has been developed where first-principle molecular simulations are used to estimate the maximal binding achievable between a target and an orally available compound(96, 99). These methods need computationally expensive simulation times and thus are developed using a smaller set of protein-ligand complexes. Ventured by this idea, we decided to build a machine model that solely depends on the geometric and physiochemical properties of the protein surface to predict the binding affinity a ligand can achieve when binding to that pocket. With the growth of X-ray crystallography, we have relatively more number of high-quality structures for protein targets. We used the “refined set” of PDBbind database to match these structures with binding data. This method can used to either select more druggable surfaces on protein targets based on what range of activity it is achieving or to get the precise activity for pockets classified as “druggable”.

3.2 Experimental Section

3.2.1 Dataset Collection and Curation

PDBbind(100) is a database of experimentally measured binding affinity data for the biomolecular complexes in the Protein Data Bank (PDB). We used PDBbind (v.2017) that contained binding data for 124,962 entries in PDB including 14,761 Protein-Ligand complexes. PDBbind provides a sub-set named as “refined set” with considerations of (i) the quality of the complex structures, (ii) the quality of the binding data and (iii) the biological/chemical nature of the complex. We extracted the refined set with 4154 Protein-Ligand complexes. We applied a few filters in order to further refine this dataset with cases where ligand bound to protein falls in a more drug-like space and where the binding data is only responsible for ligand binding. We removed 7 PDB structures which were present in SAbDab(101), a structural database for antibodies. We obtained the smiles for metabolites from the Human Metabolome Database(102) (HMDB 3.0) and compared them with all the ligands in refined set of PDBbind. We then excluded the 139 protein-metabolite complexes. We then identified ligands with polypeptide and polynucleotides and removed 286 such complexes and excluded these. From the remaining complexes we filtered cases where ligand’s molecular weight is between 300-750 kDa and logP 0-6 in order to form a dataset focused on protein interaction with drug-like small molecule. We used Open Babel v.2.3.1(103) to calculate these 3 properties of the ligand and excluded 1572 complexes outside molecular weight and 407 complexes outside logP bounds. Metals can in some cases facilitate ligand's binding to the protein. We thus excluded 395 complexes where a metal was present with 4 Å of any non-hydrogen atom on ligand in the complex 3D structure. We also excluded 230 cases where either 2 copies of same ligand or some other ligand was present in the binding site. Small molecule peptidomimetics like Saquinavir are intended to mimic protein interaction with polypeptide or protein partner. Similarly binding of lipids or sugars or nucleotides are different in terms of interaction with proteins when compared to drug-like small molecules. We used number of rotatable bonds calculated through mol_charcterize script in the silico package(104) and het-ids for ligands in PDB to filter out flexible peptidomimetics, lipids, nucleotides like ATP and solvents like glyercol (GOL) if present. This gave us 848 protein-ligand complexes.

We then clustered these proteins into groups based on sequence identity with a cut-off of 0.65. We used Clustal Omega (v.1.2.3)(105, 106) to perform pairwise sequence alignment for all the 848 proteins to obtain an identity matrix. We then used hclust function of package stats in R statistical computing environment(68) (v3.4.2) to hierarchically cluster these sequences and cutree function with sequence identity cut-off of 0.65 to cluster the proteins into groups. This resulted in 230 groups of protein and their protein-ligand complex 3D structures in each of the 230 bins.

3.2.2 Collecting Pockets

For estimating ligand binding pockets for each of the 848 protein-ligand complexes, we collected all heavy atoms with 4 Å from a bound small-molecule. This was inspired from the Prox4(107) method that is employed in PockDrug to get pockets for holo proteins.

3.2.3 Characterizing Pockets

The ligand pockets were then characterized using the 8 physiochemical and 5 structural features of each pocket as detailed in Table 1. The 'hydrophobicity_pocket' was obtained from NACCESS software(v.2.1.1)(108) and 4 of the structural features, 'SURFACE_HULL', 'VOLUME_HULL', 'SMALLEST_SIZE' and 'INTERIA_3' are calculated using the RADI software (v.4.0.1)(109). Rest of the physiochemical features and 'C_RESIDUE' were computed using their definition in a perl script.

Property Name	**Relative Importance**	**Description(107)**
VOLUME_HULL	12.594	volume of convex hull computed using RADI software(109)
hydrophobicity_kyte	10.787	hydrophobicity based properties of residues(110)
SMALLEST_SIZE	10.172	distance separating the two closest slabs enclosing the hull computed using RADI software(109)
INERTIA_3	9.453	smallest eigenvalue of inertia matrix computed using RADI software(109)
p_N_atom	8.493	frequency of N atoms in pocket(111)
hydrophobicity_pocket	7.798	hydrophobicity pocket estimated with solvent accessibility computed using NACCESS software(108)
p_aliphatic_residues	6.93	frequency of positive residues in pocket (I, L, V) (111)
p_aromatic_residues	6.803	frequency of aromatic residues in pocket (F, Y, H, W) (111)
SURFACE_HULL	6.684	surface of convex hull computed using RADI software(109)
p_negative_residues	5.692	frequency of negative residues in pocket (D, E) (111)
C_RESIDUE	5.301	number of residues in pocket

p_Ooh_atom	4.95	frequency of Ooh atoms in pocket(111)
p_Ccoo_atom	4.342	frequency of Ccoo atoms in pocket(111)

Table 1: A list of the 13 features used to characterizing the Pockets and construct the machine learning models along with their feature importance for building GBM model.

3.2.4 *Obtaining pActivity Values*

Binding data for each of the 848 protein-ligand complex was obtained from PDBbind database (v.2017)(100). Binding data is reported as IC50 or Ki or Kd values in PDBbind. Based upon the type of the activity data, pActivity value for each case was determined using equation 1 below. In order to maintain the comparability of binding data we used a factor or 2 in cases where IC50 was reported(112).

$$pActivity = \begin{cases} -log10(activity), & activity\ as\ Ki\ or\ Kd \\ -log10(activity * 0.5), & activity\ as\ IC50 \end{cases} \qquad \text{-------------------------------(1)}$$

3.2.5 *Building machine model*

The final dataset consisted of 848 data instances of 13 pocket features and their respective pActivity values as labels. We used Gradient Boosting(113-115) algorithm in R statistical computing environment(68) (v3.4.2) to build a model that will predict pActivity for the respective pockets using their features. We used 1000 trees that were allowed to have variable interactions to the highest level of 10. A minimum of 10 observations were required at the terminal nodes of the trees. We also used a step-size reduction, also known as learning rate, of 0.02. We used the whole dataset as the training set for our final GBM model. This model is available through supplemental material.

$$\begin{aligned} formula < -\, Activity \sim VOLUME_{HULL} + INERTIA_3 + SMALLEST_{SIZE} + hydrophobic_{kyte} \\ + C_{RESIDUE} + p_{Ooh} + p_{N_{atom}} + SURFACE_{HULL} + p_{aliphatic_{residue}} + p_{negative_{residue}} + p_{Ccoo} \\ + hydrophobicity_{pocket} + p_{aromatic_{residue}} \end{aligned}$$

$$\begin{aligned} model\ < -\, gbm(\, formula, distribution\ = \ "gaussian", \\ data\ = \ train, \\ n.trees\ = \ 1000, \end{aligned}$$

$$
\begin{aligned}
&shrinkage\ =\ 0.02,\\
&interaction.depth\ =\ 20,\\
&n.minobsinnode\ =\ 10)
\end{aligned}
$$

3.2.6 Testing performance of the GBM model

We followed the “Leave Group Out” protocol to evaluate the performance of the GBM model. We built 230 independent GBM model in 230 iterations where each iteration is using the whole protein group as the test set and rest of the dataset as the training set. In the following iteration, next protein group is put in a new test set and rest of the dataset goes into the training set. This loop runs until each of the 230 protein groups has been left out of training the model. A schematic presentation of this protocol is

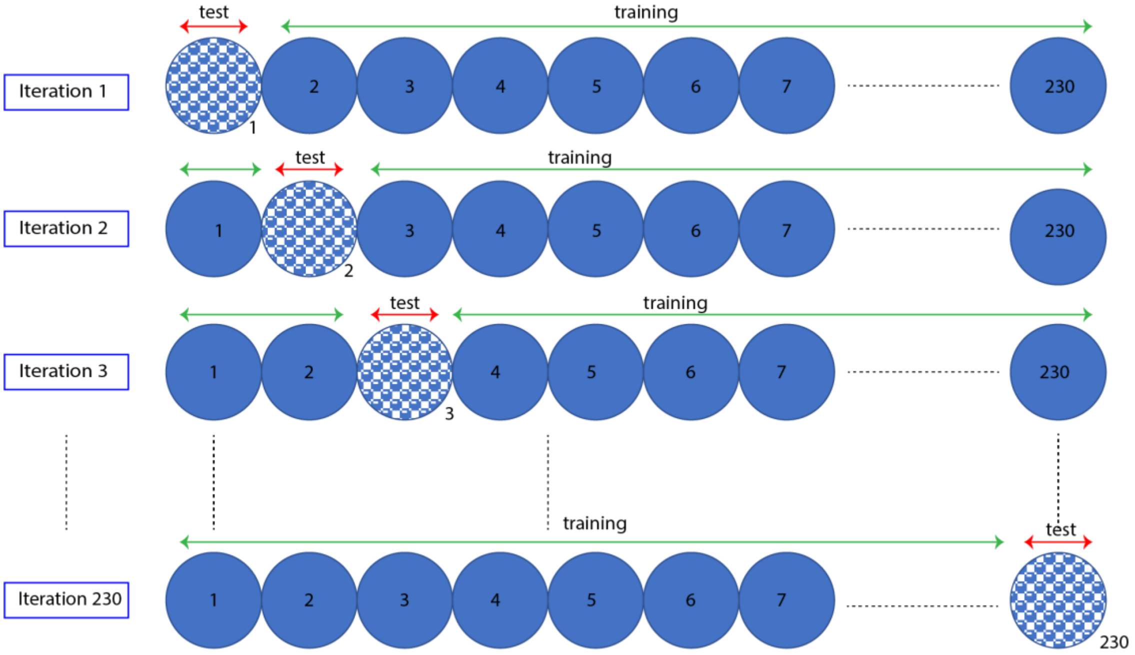


Figure S1: Schematic for the “leave group out” protocol. In the first iteration, protein group 1 is used as test set and rest 229 protein groups are used as training set to train the GBM model. This GBM model is then applied to the test set to get predictions for pActivity for all the proteins 3D structures in group 1. In the following iteration, next protein group is set aside as test side and another GBM model is trained on the remaining data to get predictions for the pActivity values for all the structures in the test set. This step is repeated 230 times until all the protein groups are tested on once.

show in **Figure S1**. This gave us predictions for pActivity for all 848 protein pockets via 230 different models. We used Pearson correlation coefficient(116) and RMSE to evaluate the performance for this ensemble of 230 GBM models.

3.2.7 *Comparing GBM performance with other established methods for predicting druggability*

The GBM model's comparisons concerning predicted pActivity were made against four druggability prediction models: PockDrug(107), SiteMap Dscore(117), fpocket druggability score(98), and DoGSiteScorer Simple as well as DoGscore(97). These models do not rely on any ligand information to identify the boundaries of the binding pockets. For PockDrug, we used the same Prox4 pockets for the proteins-ligand complexes. However, for the other three methods, we used the druggability score of the pocket that had the maximum overlap in residues with prox4 pockets for the respective complex. We used Pearson correlation coefficient(116) to compare the performance in each case.

3.2.8 *Analyzing druggability of Cryptic Sites*

We used the "CryptoSite Set"(118), a set of 93 proteins which was assembled after analyzing all the proteins in Protein Data Bank that have a ligand-bound and unbound structures available. 41 of the proteins in this set had binding data that was reported with the dataset (originally derived from BindingMOAD(119)). In order to maintain comparability, we further extracted only the cases where one ligand molecule was binding to the pocket and there was no metal in vicinity to facilitate the binding affinity which gave us a set of 26 proteins. We then used Prox4 method to obtain pockets and calculated the 13 features for these 26 cryptic pockets. There is no overlap between the GBM training set and 21 of these proteins. So, we used our original single model to obtain the predictions for pActivity for each of the 21 pockets. The remaining five proteins in structures, 2BYS, 2H4K, 2JDS, 2OT1 and 3CFN were carefully matched with the protein in GBM dataset and predictions were obtained from the GBM model which was trained without that particular protein. We used Pearson correlation coefficient(116) and RMSE to evaluate the performance of our GBM model on this test set. As the pocket only forms in the ligand bound conformation of the protein, we would hypothesize that the "closed up" ligand binding site in the apo conformation will be less druggable. For each of the 26 proteins, we structurally aligned the apo protein to the ligand bound conformation using lsqfit script in MMTSB toolset(120). We then extracted all the heavy atoms in the apo protein within 4 A of any non-hydrogen atom of the ligand in ligand bound structures. All other steps of pocket feature collection and application of GBM model remains the same. We used sign test to evaluate the success of the method in distinguishing cryptic pockets to the absence of pockets in apo conformation.

3.2.9 Collecting PTPN set for druggability analysis

We assembled all the crystal structures from PDB for PTP protein. Out of 55 that we collected, 50 of the structures had binding data reported in PDBbind. In order to maintain comparability, we removed 3 cases where inhibitors were bound to an allosteric site. We used Prox4 method to obtain pockets and calculated the respective 13 features in a similar manner. Thus, we assembled a test set consisting of 47 data instances for PTPN and their respective pocket features and pActivity. There were 20 common PDB structures in our training set that that we used for our GBM model. We removed these 20 cases and retrained a gbm model on 828 instances. We then used this new GBM model to obtain predictions for pActivity for each of these 47 pockets. We used Pearson correlation coefficient(116) and RMSE to evaluate the performance of our GBM model on this test set.

3.2.10 Collecting dataset of inhibitors of LFA-1-ICAM for druggability analysis

We assembled eight crystal structures from PDB of LFA-1 bound to 8 different inhibitors of LFA-1 interaction with ICAM. We obtained pActivity for each of these structures from PDBbind. We used Prox4 method to obtain pocket and calculated the 13 features in a similar manner. Thus, we assembled a test set consisting of 8 data instances for LFA and their respective pocket features and pActivity. As there is no overlap between the GBM training set and these 8 instances, we used our original single model to obtain predictions for pActivity for each of the 8 pockets. We used Pearson correlation coefficient(116) and RMSE to evaluate the performance of our GBM model on this test set.

3.2.11 Analyzing druggability of PPI-inhibitors

We assembled a list of 368 Protein-Protein Interaction non-polymer inhibitors from P2PI(121) that have a crystal structure reported in PDB and binding data for the inhibitor in PDBbind. This set consisted of 27 proteins. We then eliminated the cases where either another molecule or metal was present within 4 A of the inhibitor. We then selected the high affinity inhibitor's protein-ligand complex as a representative of the PPI inhibition. We also excluded HIV-Integrase because the inhibitor is known to bind the protein at the interface of the dimer conformations and the biological assemblies obtained for this protein from PDB had single chain. Thus, the pockets so obtained were not the complete representative of the inhibitor binding site. This gave us a set of 24 structures. We then obtained pockets, calculated the 13 pocket features and obtained pActivity from PDBbind. Five proteins (KRAS, BRDT-1, BRD2-2, WDR5, KEAP1) in this set were common to out GBM training set. We carefully matched the protein with the GBM dataset and predictions were obtained from the model which was trained without that particular protein. For rest of the 19 instances, we used our original single model to obtain the predictions for pActivity for each of the pockets. We used Pearson correlation coefficient(116) and RMSE to evaluate the performance of our GBM model on this test set.

3.3 Results

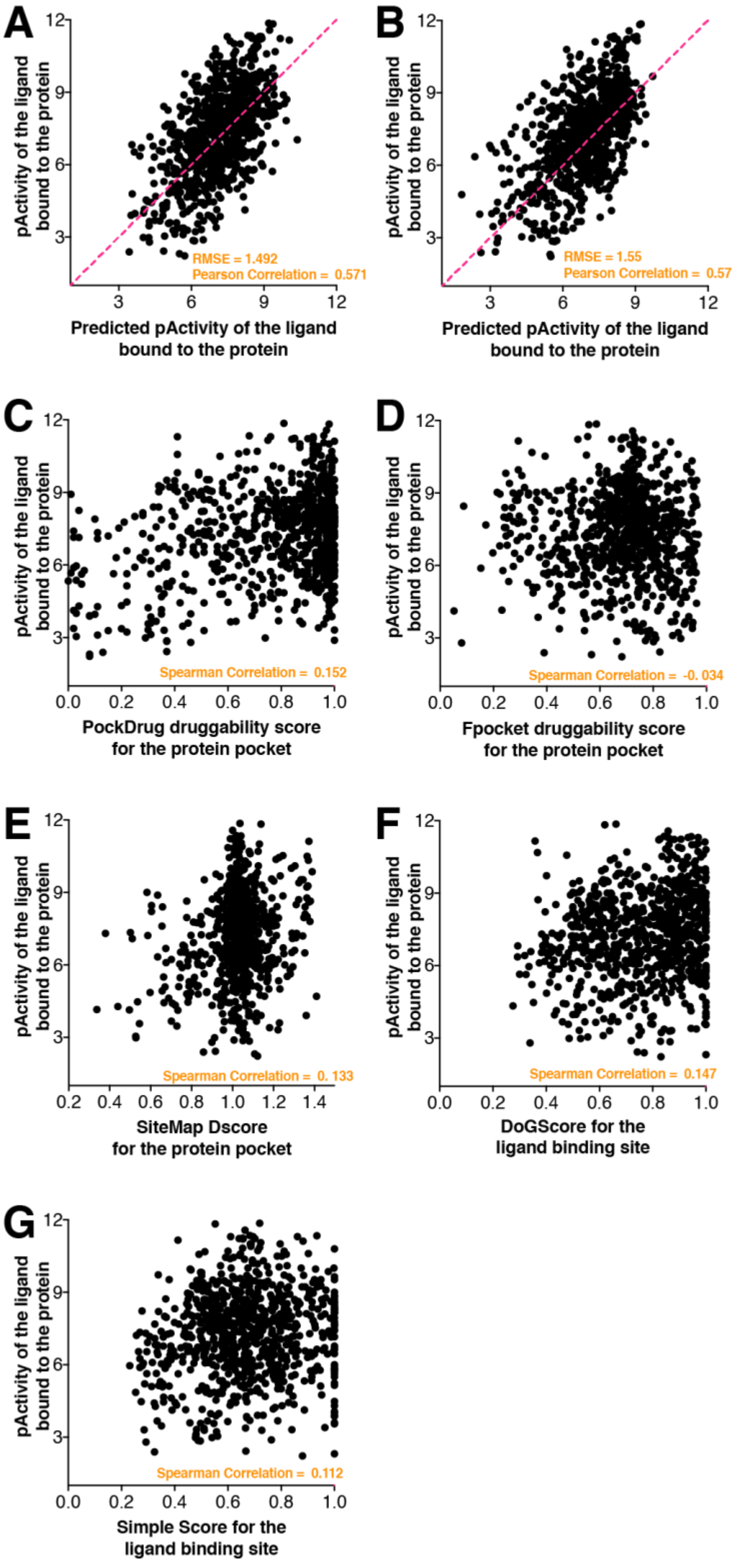


Figure 1: Prediction of pActivity and druggability estimate from the features of ligand binding protein surfaces. GBM model was used to predict pActivity that a ligand can achieve when binding to protein surface of interest (A). Numerically transformed predictions for pActivty Random Forest Model show similar performance (B). In these cases the pink dashed line represents the identity line (y=x). Druggability estimates using PockDrug (C), Fpocket (D), SiteMap (E) and DoGSiteScorer (F-G) show little to no correlation to the quantitative binding data information. The RMSE in case of GBM results and Pearson correlation in all cases is shown on each plot in orange.

3.3.1 GBM Model Performance

The final dataset used to build the machine model was obtained by filtering PDBbind dataset "refined set" to include only the protein-ligand complexes where (i) the ligand is not a poly-peptide or poly-nucleotide or a lipid or a metabolite or a no solvent; (ii) ligand's molecular weight falls into the range of 300-750 kDa, logP with 0-6 and number of rotatable bonds less than 14; (iii) there is no metal or any other ligand in 4 Å proximity; and (iv) protein is not an antibody. Binding data obtained for such 848 protein-ligand complexes from PDBbind is used as label in terms of pActivity values, for training Gradient Boosting Method (GBM) model. Based on 65% sequence identity, the proteins in this dataset are clustered into 230 protein groups. This sequence identity cut-off was carefully chosen to cluster PDB structures into groups where one group contains all the structures of the same proteins extracted from different organisms or isoforms of same proteins.

We then applied the "Leave Group Out" protocol as illustrated in **Figure S1**, to build 230 GBM model where each model is trained on 229 groups of protein and tested on the remaining group. In the first iteration of this protocol, protein group 1 is used as test set and rest 229 protein groups are used as training set to train the GBM model. This GBM model is then applied to the test set to get predictions for pActivity for all the proteins structures in group 1. In the following iteration, next protein group is set aside as test set and another GBM model is trained on the remaining data to get predictions for the pActivity values for all the structures in the test set. This step is repeated 230 times until all the protein groups are tested on once. In the end we have pActivity predictions for all 848 pockets structures.

Predicted pActivity values of each of the pocket in all the Protein-Ligand complexes correlated well with their actual values obtained from PDBbind with a Pearson Correlation coefficient of 0.571 (p-value <0.0001) and a low RMSE of 1.492 (**Figure 1A**). We used another popular tree based algorithm, random forest(84) to build a model in a similar fashion as GBM. The prediction so obtained for pactivity were numerically transformed using the method proposed in (RF paper). Both the models give very similar predictions (**Figure 1B**). The popular methods, PockDrug(107), SiteMap(117), fpocket(98), and

DoGSiteScorer(97), provide a qualitative estimate of druggability of a protein surface without relying on any ligand information to identify the boundaries of the binding pockets. PockDrug and Fpocket druggability, DoGSiteScorer DoGScore and Simple score greater than 0.5 represents druggable pockets. There is no cap for the SiteMap Dscore, but a score greater than 1.0 shows particular promise. Even though these algorithms can predict the confidence for druggability, these can't be co-opted to make quantitative predictions of binding affinity as evident from **Figure 1C-G**.

3.3.2 Druggability Analysis of Cryptic Pockets

It has long been well-known that proteins being the dynamic objects may change conformation upon ligand binding(122) at the binding site. These are known as "cryptic sites" that are formed in a ligand-bound structure, but not in the unbound protein structure(123). Pocket formation at such a site is

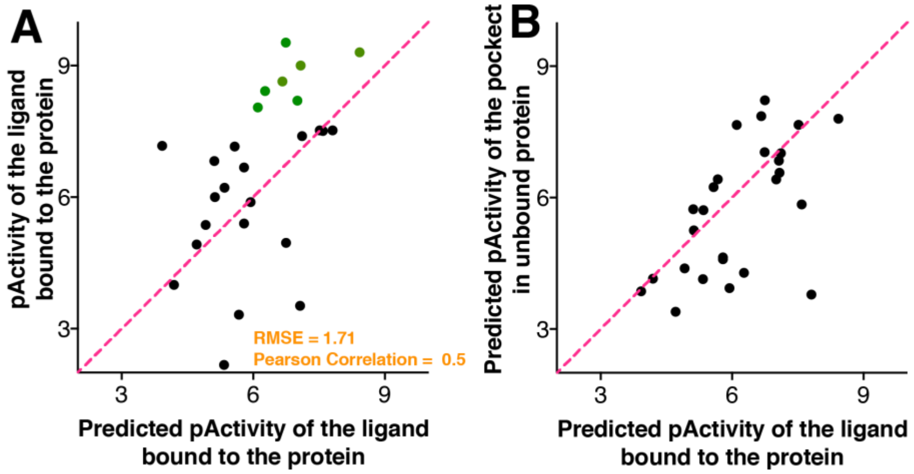


Figure 2: Druggability Analysis of Cryptic Sites. (A) GBM model was used to predict pActivity that a ligand can achieve when binding to a pocket in cryptic sites. High affinity pockets where ligand binds with an affinity greater than 10 nM, is shown in green. The RMSE in this case of GBM results and Pearson correlation in all cases is shown on each plot in orange. (B) GBM model was used to predict pActivity for the ligand binding site in the apo conformation of the protein (Signs Test: p< 0.163 & Wilcoxon signed rank test: p<0.104). In these cases, the pink dashed line represents the identity line (y=x).

driven by conformational changes in the structure. Even though it can be flexible in nature, cryptic site is conformationally conserved regardless of the ligand type(124). We next asked how well our GBM model performs while predicting pActivity of pockets at such cryptic sites. We used the “CryptoSite Set”(118) a set of 93 proteins which was assembled after analyzing all the proteins in Protein Data Bank that have a ligand-bound and unbound structures available. 41 of these proteins in this set had binding data that was available in BindingMOAD(119) (as used in originally paper). In order to maintain comparability, we further extracted only the cases where one ligand molecule was binding to the pocket and there was no metal in vicinity to facilitate the binding affinity. This gave us a set for 26 proteins including 7 the validated high affinity (pActivity >8) cryptic sites. The 7 high affinity pockets (**Figure 2**, green data points) were predicted to have higher pActivity than others with a correlation of 0.59. Our model was able distinguish between low affinity pockets from high affinity pocket with a correlation of 0.5 (p-value < 0.0049) and RMSE of 1.71 in predictions (**Figure 2A**). As the pocket only forms in the ligand bound conformation of the protein, we would hypothesize that the “closed up” ligand binding site in the apo conformation will be less druggable. For each of the 26 proteins, we aligned the apo protein to the ligand bound conformation. We then exported all the heavy atoms in the apo protein within 4 A of any non-hydrogen atom of the ligand in ligand bound structures. This set of atoms are then used to predict pActivity using GBM model. In 16 of the structures our model gave a less pActivity value for the apo conformation (**Figure 2B**).

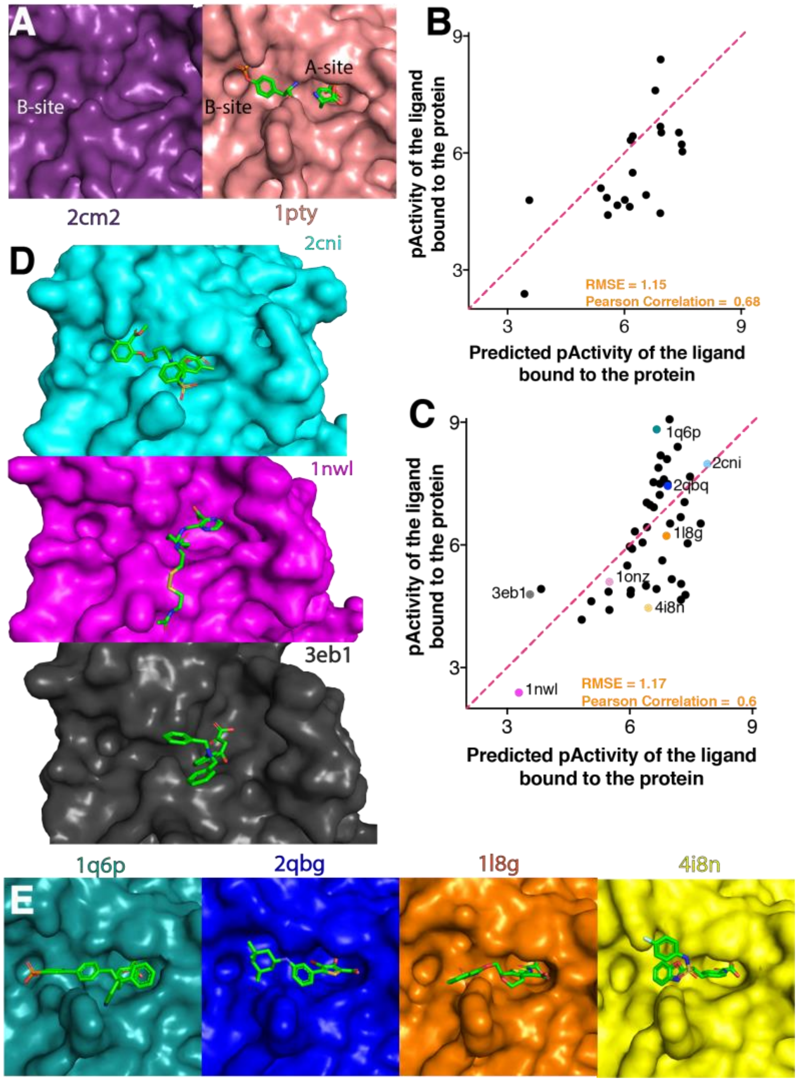


Figure 3: GBM Model based Druggability Analysis of PTP-1B. (A) PTP-1B protein changes its conformation from open (PDBid: 2cm2) conformation [top left] to closed (PDBid: 1pty) conformation [top right] upon ligand binding at catalytic A-site. An adjacent non-catalytic B-site doesn't change conformation. Shift of flap region Trp179 - Glu186 [bottom] is required for catalysis. GBM model was used to predict pActivity that a ligand can achieve when binding to protein pockets in refined dataset in this study (B) and whole PDB (C). In these cases, the pink dashed line represents the identity line (y=x). The RMSE in case of GBM results and Pearson correlation in all cases is shown on each plot in orange. (D) A high affinity ligand bound to A-site and B-site of PTP1B (Top), a low affinity ligand bound to open conformation (middle), a medium affinity ligand that bind weakly to non-druggable open conformation of PTP1B (bottom). (E) High affinity for a ligand can be achieved when it binds tightly with the highly druggable A-site in open conformation as well as to B-site.

3.3.3 GBM Model based Druggability Analysis of PTP-1B

Protein tyrosine phosphatases belong to a super family of enzymes that hydrolytically remove phosphate groups from proteins. Protein tyrosine phosphatases 1B (PTP1B) is a well validated target for its involvement in type-II diabetes. This enzyme is known to negatively regulate the insulin signaling by dephosphorylating phosphotyrosine (Tyr(P)) residues in the kinase regulatory domain of the insulin receptor (IR)(125). It is a well-known example where the protein opts a closed conformation (**Figure 3A**, top right) after the 8 Å shift of flap region Trp179 - Glu186, upon ligand binding(126, 127). Flap movement is important for catalysis because it brings the general acid Asp181 into position to protonate the tyrosyl leaving group and positions the flap over the active site (A-site) where it shields the catalytic machinery from the solvent during catalysis(127). A secondary non-catalytic site (B-site) adjacent to A-site is present in both the conformations (**Figure 3A**). This site has been explored to help inhibitors achieve higher binding affinity(128).

After keeping aside, the 20 structures of PTP proteins from our dataset, we built a GBM model trained on all but PTP proteins and their pocket features. When this model was applied to the 20 PTP pockets, it gave a highly correlated (Pearson correlation coefficient of 0.68, p-value < 0.0005) predictions for pActivity for each pocket (**Figure 3B**).

We next asked, whether the GBM model can predict pActivity for all the protein structures regardless of whether it has gone through structural refinement or not. This would test the capabilities of our model when dealing with data with noise. So, we assembled all the ligand bound structures from PDB of PTP protein. We found 27 more PDB structures where a single copy of ligand molecule that was in drug-like space in terms of MW, LogP and number of rotatable bonds, was bound at the catalytic site. The predictions for pActivity that our model made for all 47 pockets is shown in **Figure 3C**. The model maintained a similar correlation (Pearson correlation coefficient of 0.6, p-value <.0001) and RMSE in predictions. The millimolar inhibitor in 1NWL binds superficially to open conformation (**Figure 3D** bottom). Our model's predicted activity for open conformation pocket is in the low millimolar range. Ligand binding sites in case of 1NWL (**Figure 3D**, bottom), 1ONZ (**Figure 3E**, left) and 3EBL (**Figure 3E**, right) is similar to the unbound protein in 2CM2 (**Figure 3A**. left) and thus were predicted to be non-druggable. On the other hand, the loop 179-187 moves in case of ligand bound structure 2CNI and becomes very similar to the loop in 1PTY, forming the high affinity pTyr binding A-site accompanied by B-site. This binding pocket (A-site and B-site) was predicted to have an activity value in range of high micromolar range (pActivity 6.5-8). As evident from the **Figure 3F**, all the pockets in that range are very similar to each other as well to 1PTY pocket. The difference in actual pActivity values can be explained by the how much the ligand is interacting with B-site while making strong interactions with A-site. The ligand in 1Q6P is interacting with both A and B site (**Figure 3F**) and thus has the highest pActivity value of the lot. Our model predicted that the maximum pActivity these pockets with both A and B-site could achieve is around 8.00 which is very close the maximum pActivity of inhibitors reported for this protein till date (ligand in 2CNE with pActivity of 9.071).

3.3.4 GBM Model based Druggability Analysis of LFA-1 ICAM Interaction

Many proteins undergo local conformational changes at the interface upon interacting with their protein partner. In some cases, the conformational change happens away from the protein-protein interface. One of the best understood examples of such protein-protein interaction is Integrin alpha L (LFA-1) interaction with ICAM molecules. LFA-1 changes its conformation from "bent-closed" to "extended-open" by moving the α-l helix in the I-domain(129-131). This conformation has high affinity for its protein partner. This conformational change can be visualized through its solved unbound crystal structure, 1LFA and protein-bound structure in 1T0P (**Figure 4A,** top**)**. A pocket enclosed by α-l helix at the I-domain present in unbound conformation is missing in the protein-bound conformation for LFA-1.

This offers an attractive opportunity for intervention of this PPI through ligand binding at allosteric site. Till date 10 such inhibitors have been discovered where binding of allosteric ligand locks in the structure of the unbound protein conformation and hinder the conformational change required for binding its protein partner. **Figure 4A** (bottom) shows the difference in the binding site of ligand-bound and protein-bound conformations of LFA-1.

We assembled 7 inhibitor bound structures reported in PDB and collected binding data from PDBbind. Prox4 pockets of all the inhibitors had similar shape and composition. There was a subtle difference in 2 of the pockets (in case of 1CQP and 1XDG) where Lys305 and Tyr307 flips away(132) creating a shallower surface in the pockets (**Figure 4D**). Also, Lys 287 shifts in and the pocket loses concavity (**Figure 4D**). It is evident in **Figure 4B** that our GBM model could even differentiate based on these subtle differences and gave correlated predictions for pActivity (Pearson Correlation Coefficient = 0.64, p-value <0.062). Our model predicted the well-formed inhibitor bound pocket to have a pActivity value of around 7.5. As our model makes predictions without relying on any ligand information, differences in binding of structurally diverse compounds (**Figure 4C**) like in this case remains unaccounted for and thus our model gives a RMSE of value 1.7 in predictions. This indicates that our model is predicting the ligand binding pocket to be druggable and it is up to the ligand to achieve that binding affinity.

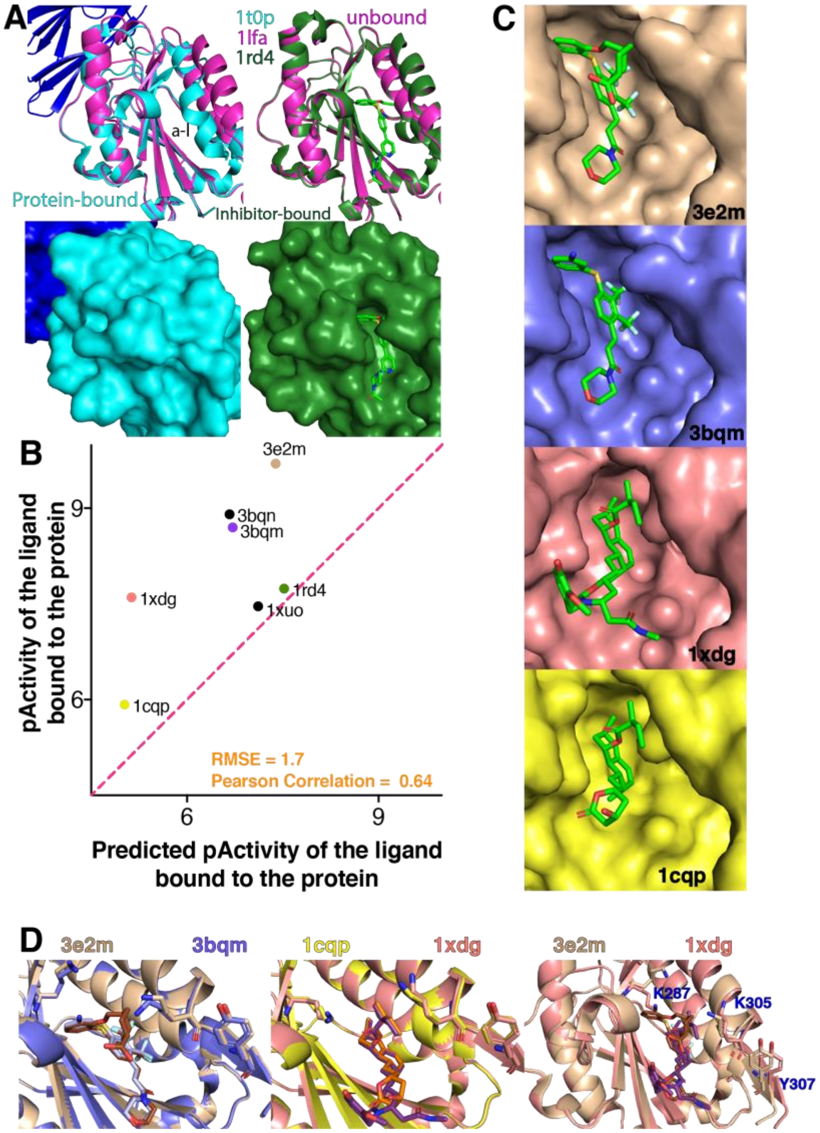


Figure 4: GBM Model based Druggability Analysis of LFA-1 ICAM Interaction. (A) LFA-1 changes its conformation from bent closed (PDBid: 1lfa) [right] to extended open (PDBid: 1t0p) [left] upon protein partner (ICAM) binding. This change is achieved by the shift of α-L helix in I domain of LFA-1. (B) GBM model was used to predict pActivity for all the inhibitor bound structures of LFA-1. In these cases, the pink dashed line represents the identity line (y=x). The RMSE in case of GBM results and Pearson correlation in all cases is shown on each plot in orange. (C) Prox4 pockets of all the inhibitors have similar shape and composition. (D) There was a subtle difference in conformation of pockets between high and low affinity inhibitors. In case of 1XDG, Lys 287 shifts in and the pocket loses concavity. Lys305 and Tyr307 flips away creating a shallower surface in the pockets.

3.3.5 *GBM Model based Druggability Analysis of Protein-Protein Interactions (PPIs)*

Protein-protein interactions (PPIs) have a pivotal role in most biological processes and thus provide an excellent opportunity for therapeutic intervention. Despite being of great interest for drug discovery, developing small molecule modulators of protein-protein interactions with desired potency, selectivity, and physicochemical properties unfortunately remains very difficult. PPIs display a wide range of diverse shapes and sizes; however, they tend to be flat, featureless, and rather large(133).

We assembled a list of Protein-Protein Interaction inhibitors that have a structure reported in PDB

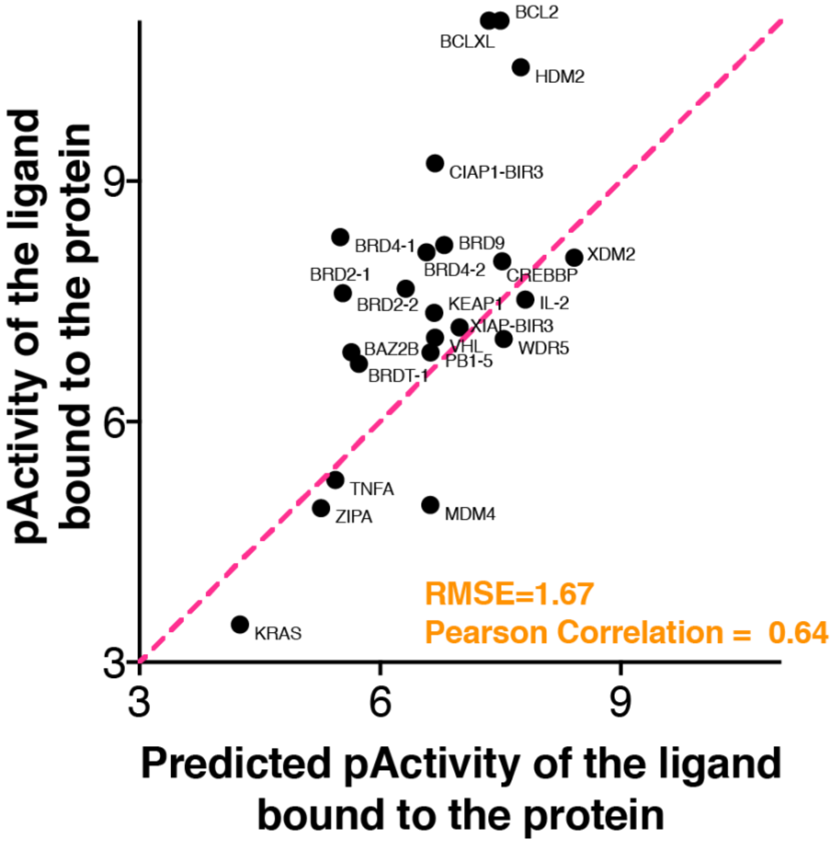


Figure 5: Druggability Analysis of Protein-Protein Interactions. GBM model was used to predict pActivity that a ligand can achieve when binding to a pocket at Protein-Protein Interaction interface. In these cases, the pink dashed line represents the identity line (y=x). The RMSE in case of GBM results and Pearson correlation in all cases is shown on each plot in orange.

and binding data for the inhibitor in PDBbind. We next asked whether PPI can be modulated through drug-like molecules. Even though our GBM model is trained on pockets where bound ligand had a molecular weight in the range of 300-750 kDa, logP in range of 0-6, we decided not to use the same restrictions while selecting the PPI structures and their inhibitors for their druggability analysis using our GBM model. For these pockets at Protein-Protein interface, our model could predict highly correlated pActivity with a Pearson Correlation coefficient of 0.64 ($p<0.0006$) (**Figure 5**). This result indeed conforms our hypothesis that contrary to the notion that PPI inhibitors require a special chemical space, this result suggests that PPI can be modulated with drug-like compounds, so it is not always true that larger compounds are needed for PPI inhibition.

3.4 Discussion

The ability to predict whether a protein target can be modulated when a small molecule binds to its surface is a longstanding goal in the field of drug discovery. A target can be classified as druggable if it has high affinity and specificity for small drug-like molecules. The goal of this study is to quantitatively assess a protein target and all its potential ligand binding sites for druggability without relying on information from the small molecules. This tool can be used before virtual screening for hit molecules and can thus guide the screen towards a particular set of chemical space. This would be a great alternative to the methods that provide qualitative or “True” – “False” druggability estimates based on NMR fragment hits. Druggability indices by hajduk have shown to be correlated with NMR hit rates(92, 93). However, this method has low accuracy when assessing druggability of high affinity pockets. Inspired by methods like this, amongst others, we build a machine model that predict protein target’s “attainable binding affinity” based on the structural and physiochemical feature of the binding pocket.

We assembled a refined set of 230 proteins targets and their small molecule bound binding pocket from 848 crystal structures in Protein Data Bank. We trained our GBM model based 13 physiochemical and structural features of these pockets. The binding data from PDBbind database was used as label.

Predicted “attainable binding affinity” for these targets are their binding pockets showed a high correlation to its actual binding affinity. We consider a pocket druggable if it is predicted to attain high affinity.

Over the past year much effort has been seen in assessing the binding-site druggability both computationally and experimentally. Currently available computational predictors for classical receptor targets are based upon the use of three-dimensional structural information of the target proteins to define the concavity of putative binding sites. These methods use geometrical scanning as in ConCavity(134) or Voronoi tessellation in Fpocket(98). Many approaches identify favorable interaction energies, often by using a van der Waals probe to explore the protein-binding site as in Q-SiteFinder(135). Alternatively, they use random forest classifiers and residue-based properties in Site Predict(136) and LDA based pocket druggability model, PockDrug(107). A comparison of Finder, Fpocket, PocketFinder, and SiteMap using structures from a huge collection of protein-ligand complexes and the respective unbound conformations demonstrated that on the average these identified 95% of the binding sites(137). In another research, after a careful analyses of protein ligand-binding sites, combined evolutionary-sequence conservation and three-dimensional structures it was shown that protein-ligand-binding sites can be predicted successfully(138, 139). With an intent to distinguish a pocket from a pocket that can bind to a drug-like molecule, method like “drug-like density” (DLID) that depends on the local density in pocket space of pockets containing drug-like ligands vs the density of total pockets, has been developed. Some experimental approaches used to indicate ligandability of a target measure hit rates, for example, in NMR-based or X-ray-based fragment screens(92, 93).

Amongst these popular methods, PockDrug(74), SiteMap(84), fpocket(64), and DoGSiteScorer(63), that just like our GBM model does not rely on any ligand information but instead give a qualitative confidence based scores for druggability. As these methods are aimed to find and define pockets that can potentially bind a small molecule, these do not show any correlation with “attainable binding affinity" (**Figure 1**). On the other hand, GBM model can be used to analyze druggability of target or to obtain the value for "attainable binding affinity" for the target known to be druggable.

In order to assess how well this model, we tested its performance against protein targets that change conformation upon ligand binding, a well characterized protein-target for therapeutic modulation, PTP-1B, and even protein-protein interaction sites including LFA-ICAM interacting where small molecule allosterically modulates the interaction. In 16 out 26 cases where the protein forms a cryptic site upon ligand binding, GBM model was able to distinguish these with unbound conformation, more so for high affinity pockets than others (**Figure 2B**). This can be further extended to protein targets that are not yet been explored through small molecules modulation. Through exploring low-energy fluctuations of such the protein surface, a computational tool can reveal cryptic druggable pockets as explained by Johnson et.al(140). Employment of our GBM model to get druggability for all such pocket opened structures can be a beneficial analysis before carrying out virtual screen. PTP-1B is one of the well-studied examples where protein changes its conformation upon ligand binding at catalytic A-site. GBM model was able to distinguish between these conformations. As the prediction here is the binding affinity that a pocket can attain, GBM model gives similar predictions for similar pockets. As our methods does not rely on any information from ligand, the actual binding data for such cases can thus be a little different than our predictions, owing to the interactions small molecule in question could achieve that are required for binding. High affinity pocket of PTP-1B consists of fully established A-site as well as B-site. Our model considers the 4 pockets in **Figure 3F** as similar pockets which are predicted to have high affinity for ligand. Now it is up to the ligand to interact in a way that it fully covers A-site and B-site.

Protein-Protein Interactions interface are considered undruggable owing to its flatness spread over large area. Our GBM model that was trained on pockets of single protein-ligand (drug-like) complexes, was able to detect surfaces on known PPIs that can bind high affinity ligands. This can be further extended to find the more druggable sub-pockets in cases where the protein-protein interface is large.

CONCLUSIONS

In this research, we address three different challenges related to structure-based drug discovery. First chapter addresses the question of when does chemical elaboration during medicinal chemistry optimization lead to change in binding mode of ligand. This will aid in designing and prioritizing the hit molecules for lead discovery. In the field of structural and computational biology, due to scarcity of structural information, proper handling of data with noise becomes very important. In second chapter, after identifying the systematic pathology of random forest regressors we developed a numerical transformation that can be applied to the output values from the regressor in order to improve the accuracy of the model’s predictions. This helped us understand how to train a machine for different application. We then build a Gradient Boosting Method (GBM) based machine model in chapter 3. This tool can predict druggable protein surface on whole proteome that can then be used in several therapeutic intervention regimes.

FUTURE RESEARCH

Allosteric Modulation of Protein-Protein Interactions

Proteins that undergo conformational changes upon binding offer an attractive opportunity for intervention of PPIs through ligand binding at a distant site. In this case a small molecule binding at an allosteric site might lock the protein into its unbound conformation, thus inhibiting the conformational changes needed for binding to the natural protein partner. We have found that allosteric inhibitors of PPIs can operate via two distinct mechanisms. First, binding of an allosteric ligand can induce a series of conformational changes that disrupt the protein interaction site and thus hinders protein partner binding. We observe this mechanism for compounds that inhibit interaction the protein CDC4 with CyclinE(141). It is very challenging to model these conformational changes. As, a result, it is difficult to apply this as a general approach for designing allosteric inhibitors for novel targets. On the other hand, there are also cases in which protein binding is accompanied by a significant conformational change away from the protein-protein interface. We observe this mechanism for compounds that inhibit the interaction of LFA1 with its protein partners in ICAM family of proteins(129-132, 142). In such cases, binding of allosteric ligand can "lock in" the structure of the unbound protein and hinder the conformational change required for binding its protein partner. There are 750 unique proteins in PDB for which structures are available both unbound and in complex with a protein partner and for which the structure has not been solved in complex with any small molecule bound. These complexes are potential candidates for development of LFA1-like allosteric inhibitors only if protein-binding induces a conformational change that includes residues remote from the protein binding site. Once the fraction of PDB that has potential to be modulated via "LFA-1 like

Allostery” has been deduced, our GBM model can be used to rank the protein in the order their druggability. This will aid in developing new (allosteric) inhibitors for a substantial number of PPIs, which would otherwise be undruggable.

APPENDIX A: Supporting Information

A.1 Supporting Information for Chapter 1

Supporting Information for this chapter includes figures showing quantitative criteria for alternate binding modes, figures showing representative electron density to confirm altered binding modes, plots analysis of several other properties with respect to whether binding modes are charged or retained, ROC plots for additional features included in this analysis. The complete set of ligand pairs compiled here, along with quantitative properties used in analysis (Dataset_S1.xlsx) can be obtained from https://doi.org/10.1021/acs.jmedchem.6b00725.

A.1.1 *Figures*

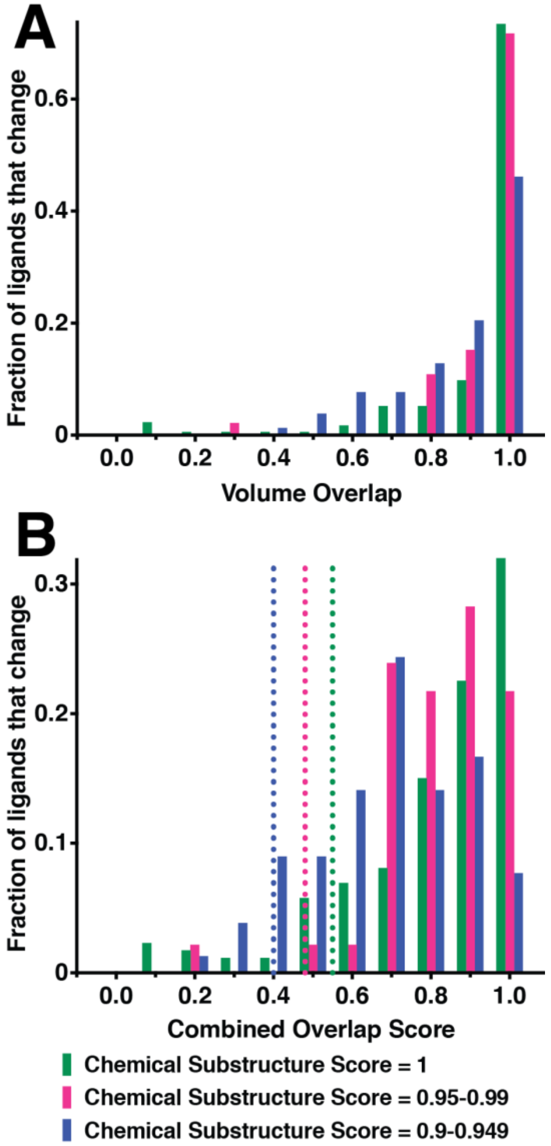


Figure S1: Defining quantitative criteria for alternate binding modes. Ligand pairs are first binned on the basis of their chemical substructure matches. For each of these ranges, **(A)** the volume overlap and **(B)** the COS overlap are plotted separately; thus, these histograms represent a more fine-grained analysis of the same data presented in **Figure 1a**. For defining alternate binding modes, the COS score has the advantage that it (correctly) detects cases in which the larger ligand subsumes the smaller ligand's volume, but does so using different functional groups; these would be missed when identifying alternate

binding modes on the basis of volume alone. Upon careful manual inspection, it became evident that any cases meeting the following criteria could be confidently assigned as "altered": (i) chemical substructure score of 1.0 and less than 0.55 COS score, (ii) chemical substructure score within the range of 0.95-0.99 and less than 0.48 COS score, and (iii) chemical substructure score within the range of 0.9-0.949 and less than 0.4 COS score. These cutoffs, marked on the COS histogram above, were used to define the set of alternate binding modes used in all subsequent studies; all other ligand pairs were assigned as "unchanged".

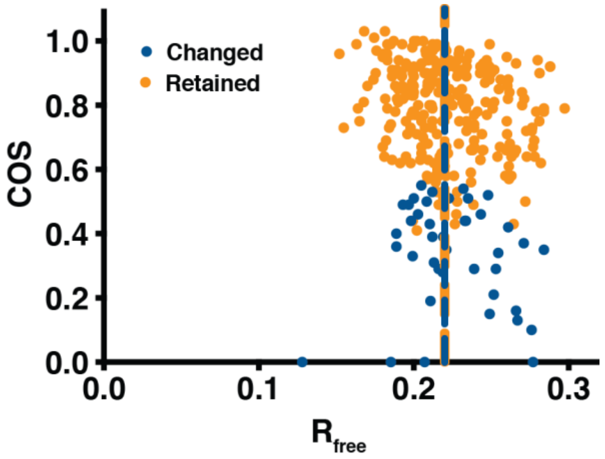


Figure S2: Dependence of alternate binding modes on crystallographic properties. Here *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is preserved. Vertical lines indicate the medians of each distribution (they essentially overlap in this case). There is no statistically significant difference between the distributions of the R_{free} values of the crystal structures in the two sets (we use only the worse of the two R_{free} values for a given pair). If these examples of alternate binding modes arose due to ambiguities in structure determination, one might expect higher R_{free} values among the structures with alternate binding modes.

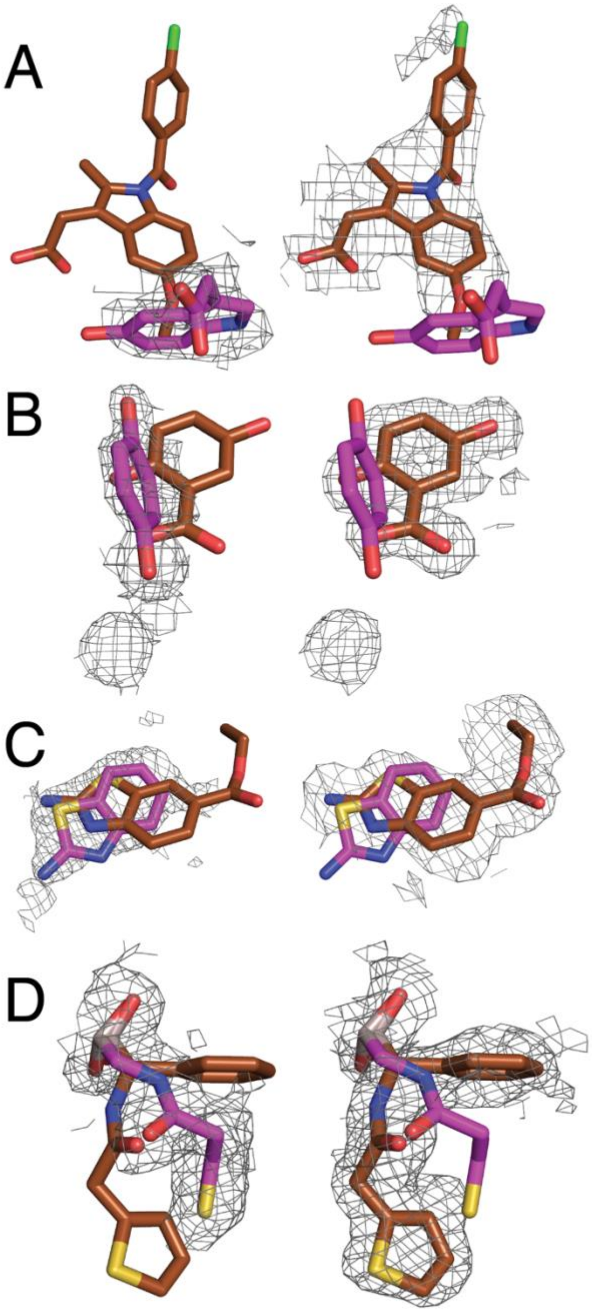


Figure S3: Examination of electron density confirms that the observed alternate binding modes are not artifacts of bad crystallographic refinement. In each case the protein structures were aligned, and the electron density for the smaller ligand (*magenta*) is shown on the left, and the density for the larger ligand (*brown*) is on the right. In each case the electron density for one ligand clearly does not fit the other ligand, providing unambiguous evidence supporting these alternate binding modes. CCP4 maps

were generated using the EDS server (143). Here we show four representative examples (listing first the smaller ligand, then the larger ligand), in each case contoured at 0.5σ: **(A)** PDB IDs 4e3h and 4e3d. **(B)** PDB IDs 3adt and 3ads. **(C)** PDB IDs 3mhw and 3kid. **(D)** PDB IDs 1fsw and 1my8.

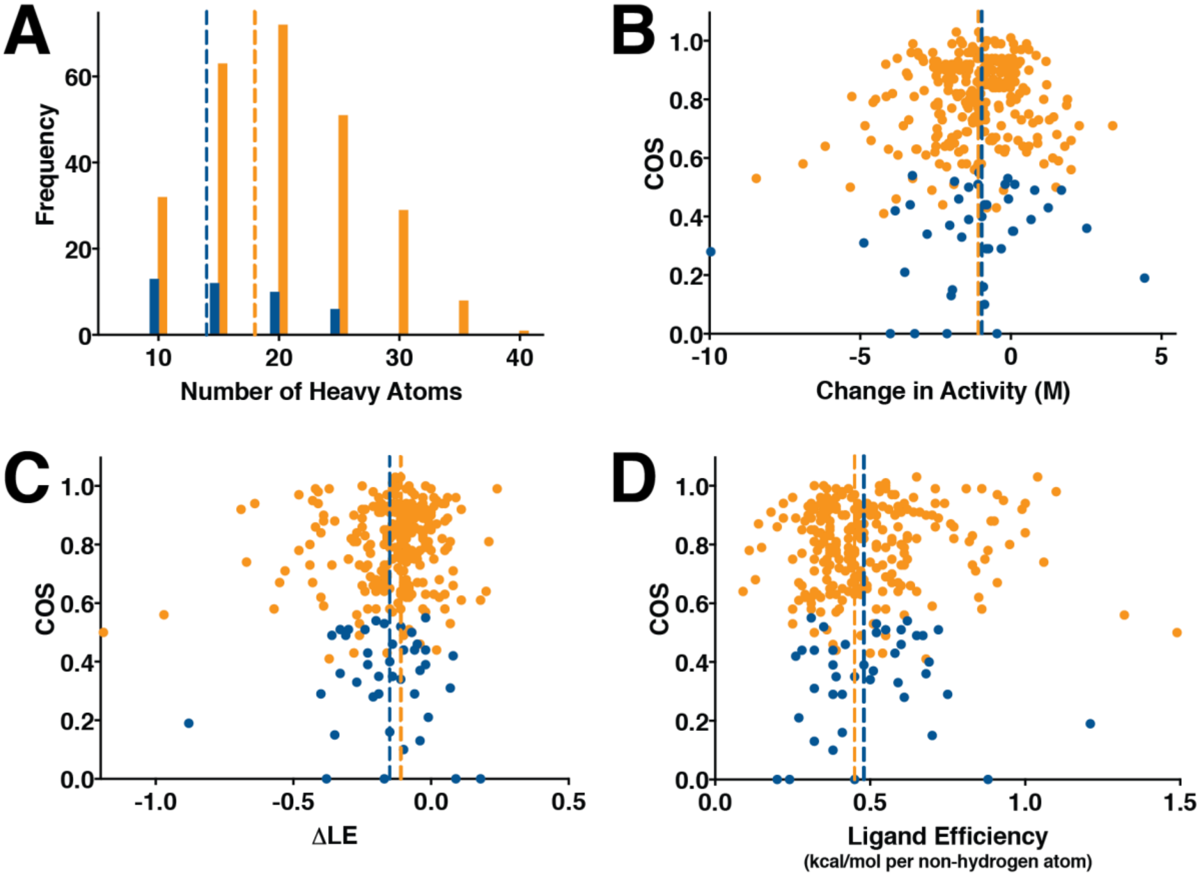


Figure S4: Further analysis of properties that correlate with increased likelihood of alternate binding modes. In each case, *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is preserved. Vertical lines indicate the medians of each distribution. **(A)** Compounds that change binding mode upon elaboration typically have fewer (non-hydrogen) atoms than compounds that retain their binding mode ($p < 4\times10^{-4}$). **(B)** There is no statistically significant difference in the distributions of the change in binding affinity upon elaboration, for ligand pairs that change binding mode versus those that retain their binding mode

(negative numbers indicate tighter binding for the larger ligand). **(C)** There is no statistically significant difference in the distributions of the change in ligand efficiency upon elaboration, for ligand pairs that change binding mode versus those that retain their binding mode (positive numbers indicate higher ligand efficiency for the larger ligand). **(D)** There is no statistically significant difference in the ligand efficiency of the smaller ligand, for ligand pairs that change binding mode versus those that retain their binding mode.

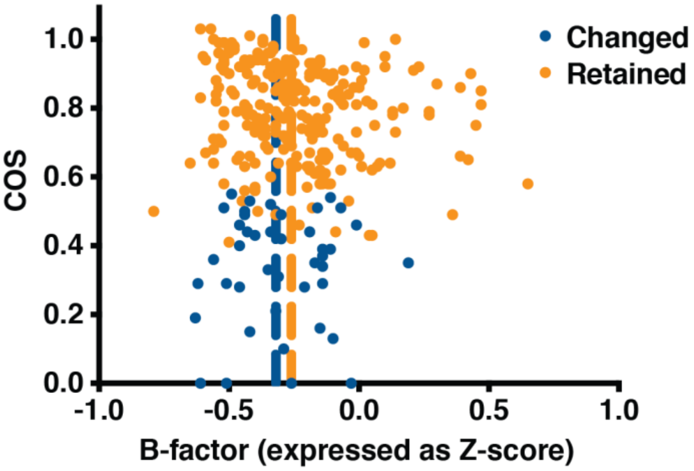


Figure S5: Analysis of binding site flexibility, in the crystal structure of the smaller ligand. Here *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is retained. Vertical lines indicate the medians of each distribution. To allow fair comparison of B-factors from crystal structures solved at many different resolutions in our set, in each case we expressed the B-factors of binding site residues as a Z-score normalized to the B-factors for all residues in the protein. Thus, negative numbers are interpreted to mean that the binding site is less flexible than an average residue in the protein, and positive numbers indicate more flexibility. There is no statistically significant difference in the distributions from ligand pairs that change binding mode versus those that retain their binding mode.

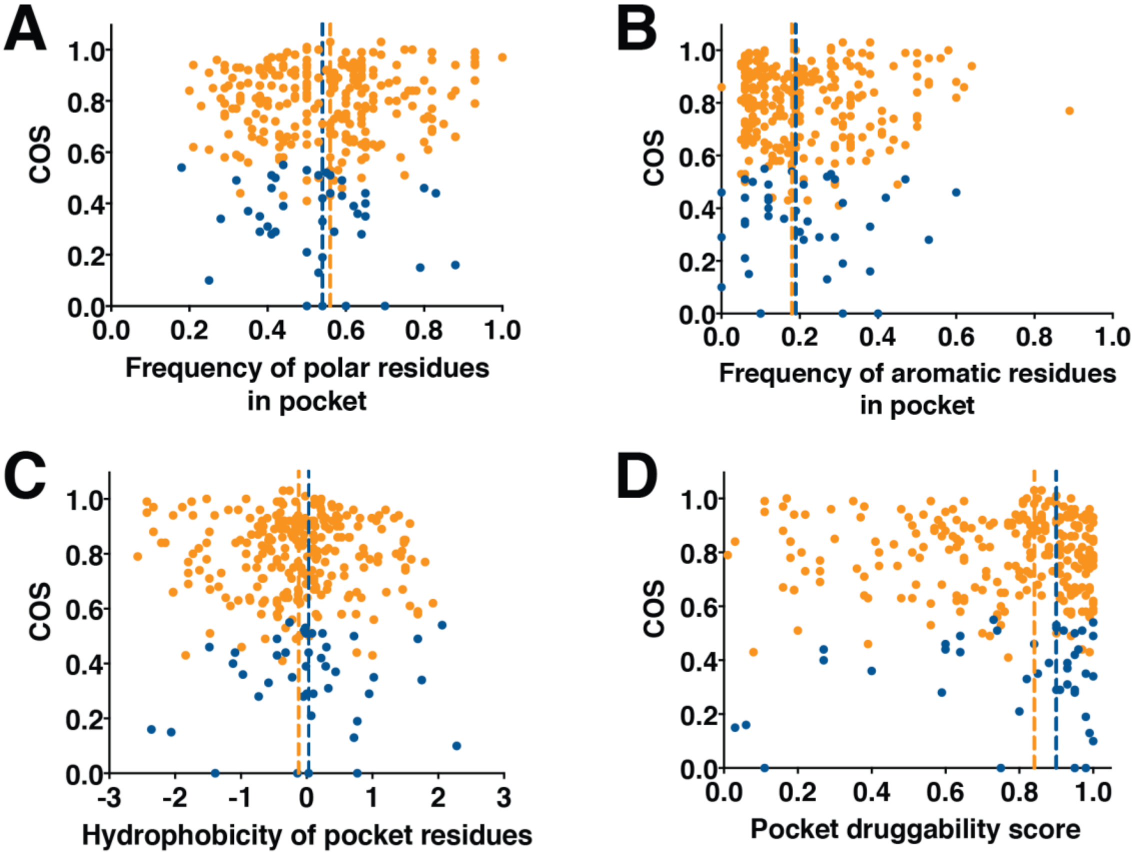


Figure S6: Analysis of the smaller ligand's binding pocket properties. In each case, *blue* indicates cases in which the smaller ligand changes binding mode upon elaboration, and *orange* is used for cases in which the binding mode is preserved. Vertical lines indicate the medians of each distribution. There is no significant difference in distributions for ligand pairs that change binding mode versus those that retain their binding mode for **(A)** frequency of polar residues in the binding pocket, **(B)** frequency of aromatic residues in the binding pocket **(C)** hydrophobicity of residues in the binding pocket, or **(D)** predicted druggability score of the binding pocket.

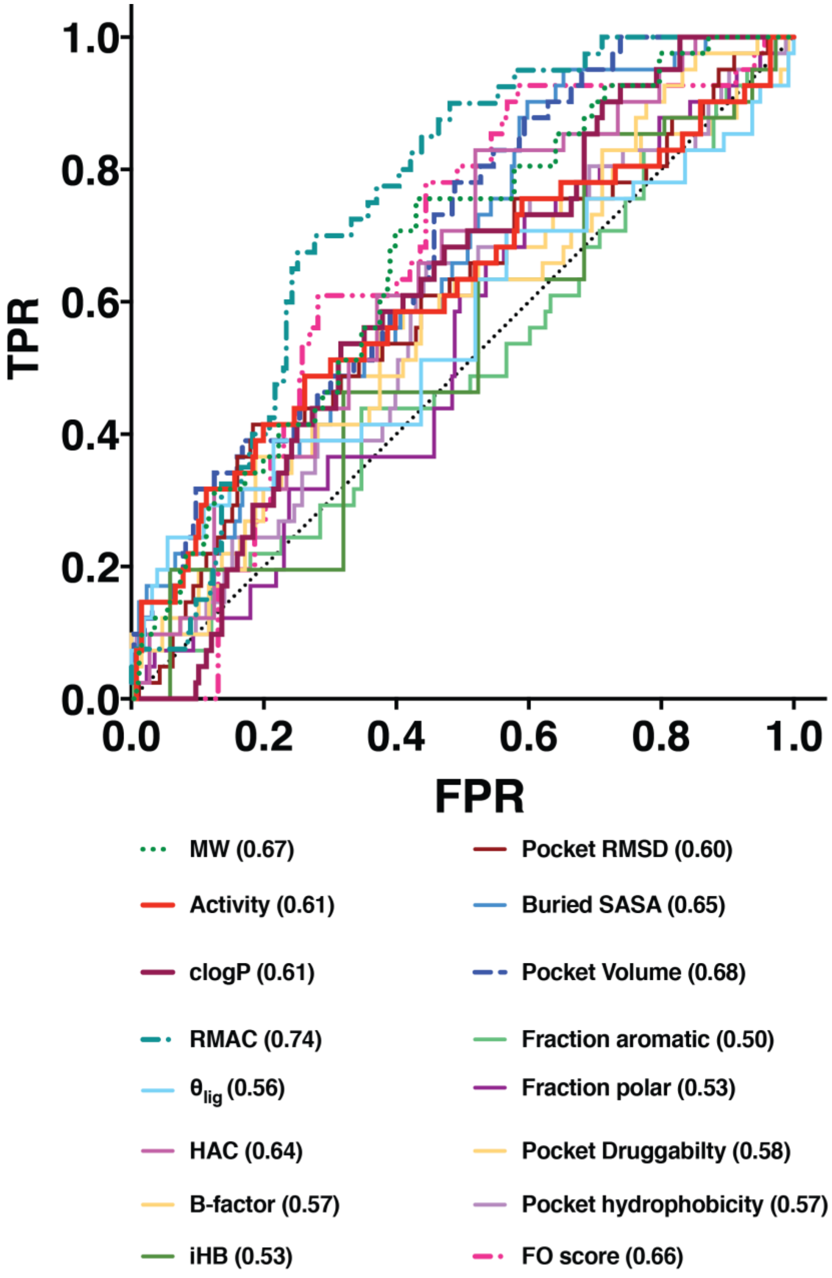


Figure S7: ROC plots for each of the properties included in this study. Receiver operating characteristic (ROC) plots comparing the utility of several different properties for predicting whether a ligand will change binding mode upon chemical elaboration; the performance of a random classifier is denoted by the black dotted line. The area under curve for each property is included for each feature, and also in **Table S1**.

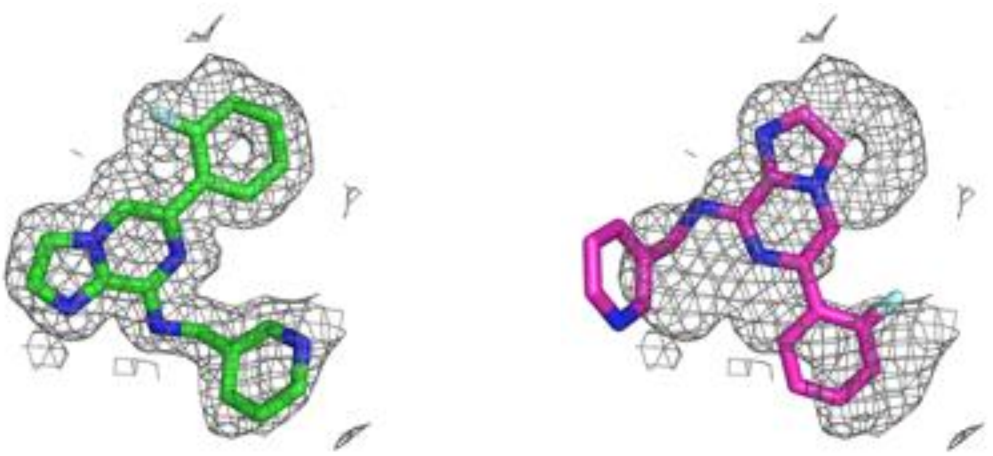


Figure S8: Examination of electron density confirms that the alternate binding mode in CDK2 inhibitor 2r3g is not simply modeled incorrectly. The electron density is contoured at 0.5σ, using a CCP4 map generated using the EDS server (143). **(Left)** The ligand coordinates as deposited in the PDB *(green)*, which closely match the density. **(Right)** The ligand coordinates from the RMAC model *(magenta)*, which shows the ligand in the binding mode shared by other compounds in this class of 5,6-bicyclic heterocyclic inhibitors.

A.2 Supporting Information for Chapter 2

A.2.1 Methods: Publicly-available regression datasets

Because the original datasets were in a crude form, we applied some basic data preprocessing methods to each one (e.g. processing missing data, removing highly co-correlated attributes and converting categorical data into factors). The specific pre-processing for each dataset is as follows:

1. <u>Airfoil Self-Noise Dataset (from UCI):</u> This dataset was originally collected by Thomas F.

Brooks, D. Stuart Pope and Michael A. Marcolini in a Technical Report to NASA(71). This dataset was donated to UCI repository in 2014. This dataset comprises different size NACA 0012 airfoils at various wind tunnel speeds and angles of attack. The span of the airfoil and the observer position were the same in all of the experiments. Target output for this data is the Scaled sound pressure level, in decibels and the attributes of this data are listed below:

- Frequency, in Hz
- Angle of attack, in degrees
- Chord length, in meters
- Free-stream velocity, in meters per second
- Suction side displacement thickness, in meters

After basic exploratory data analysis performed in R base package, we found out that this data set contains 1503 instances without any missing values for any of the attributes.

2. <u>Concrete Slump Test Dataset (from UCI):</u> This dataset was originally collected by I-Cheng Yeh at the Department of Information Management, Chung-Hua University (Republic of China)(74-78). This dataset was donated to UCI repository in 2009. The data set includes 103 data points. There are 7 input variables (listed below), and 3 output variables in the data set. The values of input variables include amount in kg of seven components included in 1 m^3 of concrete.

- Cement
- Slag
- Fly ash
- Water
- SP
- Coarse Aggregate.
- Fine Aggregate.

Out of the 3 output variables, Slump (cm), Flow (cm) and 28-day compressive strength (Mpa), for the purpose of this study we chose strength as our target variable for training random forest model. Slump and Flow are serviceability criteria which can be measured at the time of concrete pour. However, 28-day compressive strength is a design criterion that needs to be predicted before-hand. After basic exploratory data analysis performed in R base package, we found out that this data set contains 1503 instances without any missing values for any of the attributes.

3. Bike Sharing Dataset (from UCI): This dataset was originally collected by Hadi Fanaee-T and contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information (79). This dataset was donated to UCI repository in 2013. The dataset has the following fields:

- instant: record index
- dteday: date
- season: season (1: springer, 2: summer, 3: fall, 4: winter)
- yr: year (0: 2011, 1:2012)
- mnth: month (1 to 12)
- hr: hour (0 to 23)
- holiday: weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- weekday: day of the week workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit:

  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy

  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- temp: Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users

Using "day" and "year" function of lubridate package (v.1.7.4) we converted date in day and year column. We then used "factor" function of the base R package (v3.4.2) to convert categorical features season, month, holiday, weekday, weathersit and day into factors. The target variable here is the count of total rental bikes including both casual and registered. This dataset was deduced in bike counts for 731 days and 11 features for weather and seasonal information.

4. Combined Cycle Power Plant Dataset (from UCI): This dataset was donated to UCI repository in 2014. It contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the plant was set to work with full load. Features consist of hourly average ambient variables

- Temperature (T) in the range 1.81°C and 37.11°C,
- Ambient Pressure (AP) in the range 992.89-1033.30 milibar,
- Relative Humidity (RH) in the range 25.56% to 100.16%
- Exhaust Vacuum (V) in the range 25.36-81.56 cm Hg

The target variable is the net hourly electrical energy output (EP) 420.26-495.76 MW. For comparability with our baseline studies, and to allow 5x2 fold statistical tests be carried out, we provide the data shuffled five times. For each shuffling 2-fold CV is carried out and the resulting 10 measurements are used for statistical testing. The averages are taken from various sensors located around the plant that record the ambient variables every second. The variables are given without normalization.

5. Online Video Characteristics and Transcoding Time Dataset (from UCI): Authors provided a separate dataset to gain insight in characteristics of consumer videos on youtube. This file contains 10 columns of fundamental video characteristics for 1.6 million youtube videos. Authors show that the distribution of video transcoding times on a set of randomly selected YouTube videos with randomly selected but valid transcoding parameters show a heavy-tailed distribution in transcoding time values (82). As most of the jobs (peak of distribution) were complete around 10sec, we chose 10 transcoding time to find the subset to train the model on. A second dataset containing 20 columns which include input and output video characteristics along with their transcoding time and memory resource requirements while transcoding videos to different but valid formats was provide for building and testing machine learning model. This dataset was collected based on experiments on an Intel i7-3720QM CPU through randomly picking two rows from the first dataset and using these as input and output parameters of a video transcoding application, ffmpeg 4. We selected a subset (50945 instances) of this dataset based on a cut-off of maximum total transcoding time set at 10 secs. There are 20 input and output video characteristics which forms the feature space of this dataset with total transcoding time in seconds as the target variable.

6. Ames Housing Dataset (from De Cock): This dataset was compiled by Dean De Cock and contains 79 explanatory variables describing many aspects of residential homes in Ames, Iowa with an aim to predict the final sale price of the house. Following is this list of these features,

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house

- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms

- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality

- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

For the purpose of this study we excluded all those features that had missing data. This gave us 36 features. 12 of these features were categorical variables, we then used “factor” function of the base R package (v3.4.2) to convert these into factors. With an intent to eliminate the multi colinear features, we used ggplot plot correlation heatmap in the reshape2 package (v.1.4.3) for plotting correlation heatmap for SalePrice. We also used ggplot with geom_smooth method “lm” to establish correlation between Sale Price and numeric variables. We finally selected 16 features ('SalePrice', 'OverallQual', 'OverallCond', 'YearBuilt', 'ExterCond2', 'TotalBsmtSF', 'HeatingQC2', 'CentralAir2', 'GrLivArea', 'BedroomAbvGr', 'KitchenAbvGr', 'TotRmsAbvGrd', 'Fireplaces', 'GarageArea', 'OpenPorchSF', 'PoolArea' and 'YrSold') that were highly correlated with Sale Price. After this preprocessing the dataset contained 1460 instances of sale prices for the houses and their 16 features.

7. <u>Insurance Cost Dataset (from Lantz):</u> This dataset consists of 6 features listed below, including physique related and region-related features for 1339 patients.

   - sex: insurance contractor gender, female, male
   - bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg /m^2) using the ratio of height to weight, ideally 18.5 to 24.9
   - children: Number of children covered by health insurance / Number of dependents

- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

We used “factor” function of the base R package (v3.4.2) to convert ‘sex’ and ‘region’ into factors. As the costliest conditions are rare and thus difficult to predict, we opted to keep the records for the 1070 patients who paid less than $16000 as medical charges.

REFERENCES

1. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC genetics. 2010;11(1):49.

2. Cosgun E, Limdi NA, Duarte CW. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. Bioinformatics. 2011;27(10):1384-9.

3. Segal MR, Barbour JD, Grant RM. Relating HIV-1 sequence variation to replication capacity via trees and forests. Statistical applications in genetics and molecular biology. 2004;3(1):1-18.

4. Cummings MP, Segal MR. Few amino acid positions in rpoB are associated with most of the rifampin resistance in Mycobacterium tuberculosis. BMC bioinformatics. 2004;5(1):137.

5. Sastry A, Monk J, Palsson BO, Brunk E, Tegel H, Rockberg J, Uhlen M. Machine learning in computational biology to accelerate high-throughput protein expression. Bioinformatics. 2017;33(16):2487-95. doi: 10.1093/bioinformatics/btx207.

6. Jia J, Li X, Qiu W, Xiao X, Chou K-C. iPPI-PseAAC(CGR): Identify protein-protein interactions by incorporating chaos game representation into PseAAC. Journal of Theoretical Biology. 2019;460:195-203. doi: https://doi.org/10.1016/j.jtbi.2018.10.021.

7. Da Silva F, Desaphy J, Bret G, Rognan D. IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein–Protein Interfaces. Journal of Chemical Information and Modeling. 2015;55(9):2005-14. doi: 10.1021/acs.jcim.5b00190.

8. Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H. Chemical predictive modelling to improve compound quality. Nat Rev Drug Discov. 2013;12(12):948-62. doi: 10.1038/nrd4128. PubMed PMID: 24287782.

9. 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(3):295-301. doi: http://dx.doi.org/10.1016/j.chembiol.2008.02.015.

10. Stout TJ, Sage CR, Stroud RM. The additivity of substrate fragments in enzyme–ligand binding. Structure. 1998;6(7):839-48. doi: http://dx.doi.org/10.1016/S0969-2126(98)00086-0.

11. 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 medicinal chemistry letters. 2013;4(7):660-5. doi: 10.1021/ml400062c. PubMed PMID: 24900726; PMCID: 4027557.

12. Babaoglu K, Shoichet BK. Deconstructing fragment-based inhibitor discovery. Nat Chem Biol. 2006;2(12):720-3. doi: http://www.nature.com/nchembio/journal/v2/n12/suppinfo/nchembio831_S1.html.

13. Barelier S, Pons J, Marcillat O, Lancelin JM, Krimm I. Fragment-based deconstruction of Bcl-xL inhibitors. J Med Chem. 2010;53(6):2577-88. doi: 10.1021/jm100009z. PubMed PMID: 20192224.

14. Murray CW, Verdonk ML, Rees DC. Experiences in fragment-based drug discovery. Trends Pharmacol Sci. 2012;33(5):224-32. doi: http://dx.doi.org/10.1016/j.tips.2012.02.006.

15. 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(7):e102300. doi: 10.1371/journal.pone.0102300.

16. Murray CW, Blundell TL. Structural biology in fragment-based drug design. Curr Opin Struct Biol. 2010;20(4):497-507. doi: http://dx.doi.org/10.1016/j.sbi.2010.04.003.

17. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302-9. doi: 10.1093/nar/gki524.

18. 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(5):1489-95. doi: 10.1021/jm040163o. PubMed PMID: 15743191.

19. ROCS version 3.2.0.3. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com (accessed January 9, 2015).

20. Liebeschuetz J, Hennemann J, Olsson T, Groom CR. The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. Journal of computer-aided molecular design. 2012;26(2):169-83. doi: 10.1007/s10822-011-9538-6. PubMed PMID: 22246295; PMCID: 3292722.

21. Reynolds CH. Protein-ligand cocrystal structures: we can do better. ACS medicinal chemistry letters. 2014;5(7):727-9. doi: 10.1021/ml500220a. PubMed PMID: 25050154; PMCID: 4094245.

22. 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(Pt 3):179-93. doi: 10.1107/S2052252514005442. PubMed PMID: 25075337; PMCID: 4086436.

23. Martin DP, Cohen SM. Nucleophile recognition as an alternative inhibition mode for benzoic acid based carbonic anhydrase inhibitors. Chem Commun. 2012;48(43):5259-61. doi: 10.1039/c2cc32013d. PubMed PMID: 22531842; PMCID: PMC3674230.

24. 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(15):7424-8. doi: 10.1016/j.bmc.2008.06.013. PubMed PMID: 18579385.

25. 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(19):3395-407. doi: 10.1038/emboj.2010.197.

26. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of Mixed IC50 Data – A Statistical Analysis. PLoS ONE. 2013;8(4):e61007. doi: 10.1371/journal.pone.0061007.

27. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. Proceedings of the National Academy of Sciences of the United States of America. 1999;96(18):9997-10002. PubMed PMID: 10468550; PMCID: 17830.

28. 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(44):30205-15. doi: 10.1074/jbc.M803347200.

29. 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(8):2073-81. doi: 10.1021/ci4002316.

30. 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(40):4872-97. doi: 10.1002/anie.200200565.

31. 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. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(20):E2585-E94. doi: 10.1073/pnas.1501567112.

32. 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(5):733-55. doi: 10.1038/nprot.2015.043 http://www.nature.com/nprot/journal/v10/n5/abs/nprot.2015.043.html#supplementary-information.

33. 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(W1):W436-42. doi: 10.1093/nar/gkv462. PubMed PMID: 25956651; PMCID: 4489252.

34. 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(4):882-95. doi: 10.1021/ci5006004. PubMed PMID: 25835082.

35. Cox DR. The Regression Analysis of Binary Sequences. J R Statist Soc B. 1958;20(2):215-42.

36. 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(8):3659-60. doi: 10.1021/ja00087a086.

37. 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(15):4694-715. doi: 10.1021/jm900259h. PubMed PMID: 19618939; PMCID: PMC2722745.

38. Manos-Turvey A, Bulloch EM, Rutledge PJ, Baker EN, Lott JS, Payne RJ. Inhibition studies of Mycobacterium tuberculosis salicylate synthase (MbtI). ChemMedChem. 2010;5(7):1067-79. doi: 10.1002/cmdc.201000137. PubMed PMID: 20512795.

39. 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(24):4868-79. doi: 10.1021/bi3002067. PubMed PMID: 22607697.

40. 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(11):1427-32.

41. 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(5):372-9. doi: 10.1002/bip.20868. PubMed PMID: 17937404.

42. Zartler ER. Fragonomics: the -omics with real impact. ACS medicinal chemistry letters. 2014;5(9):952-3. doi: 10.1021/ml5003212. PubMed PMID: 25221648; PMCID: 4160759.

43. Murray CW, Rees DC. The rise of fragment-based drug discovery. Nature chemistry. 2009;1(3):187-92. doi: 10.1038/nchem.217. PubMed PMID: 21378847.

44. Congreve M, Carr R, Murray C, Jhoti H. A ‘Rule of Three’ for fragment-based lead discovery? Drug Discov Today. 2003;8(19):876-7. doi: http://dx.doi.org/10.1016/S1359-6446(03)02831-9.

45. 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(8):644-. doi: 10.1038/nrd3926-c1.

46. 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(16):5942-55. doi: 10.1021/jm100059d. PubMed PMID: 20718493.

47. 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(17):5396-404. doi: 10.1016/j.bmcl.2012.07.052. PubMed PMID: 22877636.

48. 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(7):e102300. doi: 10.1371/journal.pone.0102300. PubMed PMID: 25025339; PMCID: 4099364.

49. 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(1):457-65. doi: 10.1021/jm501597j. PubMed PMID: 25437144.

50. 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(15):4794-809. doi: 10.1021/jm900357y.

51. 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(12):2977-80. doi: 10.1021/jm030580l.

52. 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(3):405-12. doi: 10.1093/bioinformatics/btu626. PubMed PMID: 25301850.

53. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204-12. doi: 10.1093/nar/gku989. PubMed PMID: 25348405.

54. O'Boyle N, Banck M, James C, Morley C, Vandermeersch T, Hutchison G. Open Babel: An open chemical toolbox. J Cheminform. 2011;3(1):33. PubMed PMID: doi:10.1186/1758-2946-3-33.

55. Englert P, Kovacs P. Efficient heuristics for maximum common substructure search. J Chem Inf Model. 2015;55(5):941-55. doi: 10.1021/acs.jcim.5b00036. PubMed PMID: 25865959.

56. Kleywegt GJ, Harris MR, Zou J-y, Taylor TC, Wahlby A, Jones TA. The Uppsala Electron-Density Server. Acta Crystallographica Section D. 2004;60(12 Part 1):2240-9. doi: doi:10.1107/S0907444904013253.

57. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

58. 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):1-12. doi: 10.1186/1758-2946-6-12.

59. 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(4):572-84. doi: 10.1021/ci100031x.

60. Hawkins PC, Nicholls A. Conformer generation with OMEGA: learning from the data set and the analysis of failures. J Chem Inf Model. 2012;52(11):2919-36.

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

62. 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. In: Michael LJ, Ludwig B, editors. Meth Enzymol: Academic Press; 2011. p. 545-74.

63. Feig M, Karanicolas J, Brooks III CL. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graph Model. 2004;22(5):377-95. doi: http://dx.doi.org/10.1016/j.jmgm.2003.12.005.

64. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

65. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255-60. doi: 10.1126/science.aaa8415.

66. Trevor H, Robert T, JH F. The elements of statistical learning: data mining, inference, and prediction. New York, NY: Springer; 2009.

67. Mitchell T, Buchanan B, DeJong G, Dietterich T, Rosenbloom P, Waibel A. Machine learning. Annual review of computer science. 1990;4(1):417-33.

68. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.

69. De Cock D. Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education. 2011;19(3):1-15.

70. Lantz B. Machine learning with R: Packt Publishing Ltd; 2013.

71. Brooks TF, Pope DS, Marcolini MA. Airfoil self-noise and prediction. Technical report, NASA RP-1218, 1989.

72. Lau K. A neural networks approach for aerofoil noise prediction. London, United Kingdom: Master's thesis, Department of Aeronautics, Imperial College of Science, Technology and Medicine, 2006.

73. Lopez R. Neural Networks for Variational Problems in Engineering. PhD Thesis, Technical University of Catalonia, 2008.

74. Yeh I-C. Modeling slump flow of concrete using second-order regressions and artificial neural networks. Cement and Concrete Composites. 2007;29(6):474-80.

75. Yeh I-C. Modeling slump of concrete with fly ash and superplasticizer. Computers and Concrete. 2008;5(6):559-72.

76. Yeh I-C. Prediction of workability of concrete using design of experiments for mixtures. Computers and Concrete. 2008;5(1):1-20.

77. Yeh I-C. Exploring concrete slump model using artificial neural networks. J of Computing in Civil Engineering. 2006;20(3):217-21.

78. Yeh I-C. Simulation of concrete slump using neural networks. Computers and Concrete. 2009;162(1):11-8.

79. Fanaee-T H, Gama J. Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence. 2014;2(2-3):113-27.

80. Tüfekci P. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. International Journal of Electrical Power & Energy Systems. 2014;60:126-40.

81. Kaya H, Tüfekci P, Gürgen FS, editors. Local and global learning methods for predicting power of a combined gas & steam turbine. Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE; 2012.

82. Deneke T, Haile H, Lafond S, Lilius J, editors. Video transcoding time prediction for proactive load balancing. 2014 IEEE International Conference on Multimedia and Expo (ICME); 2014: IEEE.

83. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18-22.

84. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32. doi: 10.1023/A:1010933404324.

85. Bradley JV. Distribution-free statistical tests. 1968.

86. Latinne P, Debeir O, Decaestecker C. Limiting the Number of Trees in Random Forests. Multiple Classifier Systems2001. p. 178-87.

87. Oshiro TM, Perez PS, Baranauskas JA. How Many Trees in a Random Forest? Machine Learning and Data Mining in Pattern Recognition2012. p. 154-68.

88. Paul A, Mukherjee DP, Das P, Gangopadhyay A, Chintha AR, Kundu S. Improved Random Forest for Classification. IEEE Transactions on Image Processing. 2018;27(8):4012-24. doi: 10.1109/tip.2018.2834830.

89. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9(1):307. doi: 10.1186/1471-2105-9-307.

90. Friedman JH. Multivariate Adaptive Regression Splines. The Annals of Statistics. 1991;19(1):1-67. doi: 10.1214/aos/1176347963.

91. Coeurjolly J-F, Scornet E, Leclercq-Samson A. Tuning parameters in random forests. ESAIM: Proceedings and Surveys. 2017;60:144-62. doi: 10.1051/proc/201760144.

92. Hajduk PJ, Huth JR, Fesik SW. Druggability indices for protein targets derived from NMR-based screening data. Journal of medicinal chemistry. 2005;48(7):2518-25.

93. Hajduk PJ, Huth JR, Tse C. Predicting protein druggability. Drug discovery today. 2005;10(23-24):1675-82.

94. Soga S, Shirai H, Kobori M, Hirayama N. Use of amino acid composition to predict ligand-binding sites. Journal of chemical information and modeling. 2007;47(2):400-6.

95. Halgren T. New method for fast and accurate binding-site identification and analysis. Chemical biology & drug design. 2007;69(2):146-8.

96. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES. Structure-based maximal affinity model predicts small-molecule druggability. Nature biotechnology. 2007;25(1):71.

97. Volkamer A, Kuhn D, Rippmann F, Rarey M. DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. Bioinformatics. 2012;28(15):2074-5. doi: 10.1093/bioinformatics/bts310.

98. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10(1):168. doi: 10.1186/1471-2105-10-168.

99. Vukovic S, Brennan PE, Huggins DJ. Exploring the role of water in molecular recognition: predicting protein ligandability using a combinatorial search of surface hydration sites. Journal of Physics: Condensed Matter. 2016;28(34):344007.

100. Liu Z, Su M, Han L, Liu J, Yang Q, Li Y, Wang R. Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. Accounts of Chemical Research. 2017;50(2):302-9. doi: 10.1021/acs.accounts.6b00491.

101. Fuchs A, Georges G, Dunbar J, Leem J, Shi J, Krawczyk K, Baker T, Deane CM. SAbDab: the structural antibody database. Nucleic Acids Research. 2013;42(D1):D1140-D6. doi: 10.1093/nar/gkt1043.

102. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorndahl T, Perez-Pineiro R, Eisner R, Allen F, Neveu V, Greiner R, Scalbert A. HMDB 3.0--The Human Metabolome Database in 2013. Nucleic acids research. 2013;41(Database issue):D801-D7. Epub 2012/11/17. doi: 10.1093/nar/gks1065. PubMed PMID: 23161693.

103. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of Cheminformatics. 2011;3(1):33. doi: 10.1186/1758-2946-3-33.

104. Chalmers DKR, B.P. Silico: A Perl Molecular Toolkit [cited 2018 June, 2018]. Available from: http://silico.sourceforge.net.

105. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011;7(1):539.

106. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R. A new bioinformatics analysis tools framework at EMBL–EBI. Nucleic acids research. 2010;38(suppl_2):W695-W9.

107. Borrel A, Regad L, Xhaard H, Petitjean M, Camproux A-C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. Journal of Chemical Information and Modeling. 2015;55(4):882-95. doi: 10.1021/ci5006004.

108. Burgoyne NJ, Jackson RM. Predicting protein interaction sites: binding hot-spots in protein–protein and protein–ligand interfaces. Bioinformatics. 2006;22(11):1335-42.

109. Petitjean M. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. Journal of chemical information and computer sciences. 1992;32(4):331-7.

110. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology. 1982;157(1):105-32.

111. Milletti F, Vulpetti A. Predicting polypharmacology by binding site similarity: from kinases to the protein universe. Journal of chemical information and modeling. 2010;50(8):1418-31.

112. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of mixed IC_{50} data - a statistical analysis. PloS one. 2013;8(4):e61007-e. doi: 10.1371/journal.pone.0061007. PubMed PMID: 23613770.

113. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics. 2000;28(2):337-407.

114. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002;38(4):367-78.

115. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001:1189-232.

116. Pearson K. Royal society proceedings, 581895.

117. Halgren TA. Identifying and Characterizing Binding Sites and Assessing Druggability. Journal of Chemical Information and Modeling. 2009;49(2):377-89. doi: 10.1021/ci800324m.

118. Beglov D, Hall DR, Wakefield AE, Luo L, Allen KN, Kozakov D, Whitty A, Vajda S. Exploring the structural origins of cryptic sites on proteins. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(15):E3416-E25. Epub 2018/03/26. doi: 10.1073/pnas.1711490115. PubMed PMID: 29581267.

119. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA. Binding MOAD, a high-quality protein-ligand database. Nucleic acids research. 2008;36(Database issue):D674-D8. Epub 2007/11/30. doi: 10.1093/nar/gkm911. PubMed PMID: 18055497.

120. Feig M, Karanicolas J, Brooks III CL. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. Journal of Molecular Graphics and Modelling. 2004;22(5):377-95.

121. Basse M-J, Betzi S, Morelli X, Roche P. 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein–protein interactions. Database. 2016;2016. doi: 10.1093/database/baw007.

122. Laskowski RA, Gerick F, Thornton JM. The structural basis of allosteric regulation in proteins. FEBS letters. 2009;583(11):1692-8.

123. Vajda S, Beglov D, Wakefield AE, Egbert M, Whitty A. Cryptic binding sites on proteins: definition, detection, and druggability. Current Opinion in Chemical Biology. 2018;44:1-8. doi: https://doi.org/10.1016/j.cbpa.2018.05.003.

124. Cimermancic P, Weinkam P, Rettenmaier TJ, Bichmann L, Keedy DA, Woldeyes RA, Schneidman-Duhovny D, Demerdash ON, Mitchell JC, Wells JA. CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. Journal of molecular biology. 2016;428(4):709-19.

125. Saltiel AR, Kahn CR. Insulin signalling and the regulation of glucose and lipid metabolism. Nature. 2001;414(6865):799.

126. Jia Z, Barford D, Flint AJ, Tonks NK. Structural basis for phosphotyrosine peptide recognition by protein tyrosine phosphatase 1B. Science. 1995;268(5218):1754-8.

127. Ala PJ, Gonneville L, Hillman MC, Becker-Pasha M, Wei M, Reid BG, Klabe R, Yue EW, Wayland B, Douty B, Polam P, Wasserman Z, Bower M, Combs AP, Burn TC, Hollis GF, Wynn R. Structural Basis for Inhibition of Protein-tyrosine Phosphatase 1B by Isothiazolidinone Heterocyclic Phosphonate Mimetics. Journal of Biological Chemistry. 2006;281(43):32784-95. doi: 10.1074/jbc.M606873200.

128. Barr AJ. Protein tyrosine phosphatases as drug targets: strategies and challenges of inhibitor development. Future Medicinal Chemistry. 2010;2(10):1563-76. doi: 10.4155/fmc.10.241. PubMed PMID: 21426149.

129. Walling BL, Kim M. LFA-1 in T Cell Migration and Differentiation. Frontiers in Immunology. 2018;9(952). doi: 10.3389/fimmu.2018.00952.

130. Pflugfelder SC, Stern M, Zhang S, Shojaei A. LFA-1/ICAM-1 Interaction as a Therapeutic Target in Dry Eye Disease. Journal of ocular pharmacology and therapeutics : the official journal of the Association for Ocular Pharmacology and Therapeutics. 2017;33(1):5-12. Epub 2017/01/01. doi: 10.1089/jop.2016.0105. PubMed PMID: 27906544.

131. Crump MP, Ceska TA, Spyracopoulos L, Henry A, Archibald SC, Alexander R, Taylor RJ, Findlow SC, O'Connell J, Robinson MK, Shock A. Structure of an Allosteric Inhibitor of LFA-1 Bound to the I-Domain Studied by Crystallography, NMR, and Calorimetry. Biochemistry. 2004;43(9):2394-404. doi: 10.1021/bi035422a.

132. Weitz-Schmidt G, Welzenbach K, Dawson J, Kallen J. Improved lymphocyte function-associated antigen-1 (LFA-1) inhibition by statin derivatives: molecular basis determined by x-ray analysis and monitoring of LFA-1 conformational changes in vitro and ex vivo. Journal of Biological Chemistry. 2004;279(45):46764-71. doi: 10.1074/jbc.M407951200.

133. Arkin MR, Tang Y, Wells JA. Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. Chemistry & biology. 2014;21(9):1102-14. doi: 10.1016/j.chembiol.2014.09.001. PubMed PMID: 25237857.

134. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS computational biology. 2009;5(12):e1000585.

135. Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. Bioinformatics. 2005;21(9):1908-16.

136. Bordner AJ. Predicting protein-protein binding sites in membrane proteins. BMC bioinformatics. 2009;10(1):312.

137. Schmidtke P, Souaille C, Estienne F, Baurin N, Kroemer RT. Large-scale comparison of four binding site detection algorithms. Journal of chemical information and modeling. 2010;50(12):2191-200.

138. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM. A method for localizing ligand binding pockets in protein structures. PROTEINS: Structure, Function, and Bioinformatics. 2006;62(2):479-88.

139. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape variation in protein binding pockets and their ligands. Journal of molecular biology. 2007;368(1):283-301.

140. Johnson DK, Karanicolas J. Selectivity by small-molecule inhibitors of protein interactions can be driven by protein surface fluctuations. PLoS computational biology. 2015;11(2):e1004081.

141. Orlicky S, Tang X, Neduva V, Elowe N, Brown ED, Sicheri F, Tyers M. An allosteric inhibitor of substrate recognition by the SCF Cdc4 ubiquitin ligase. Nature biotechnology. 2010;28(7):733.

142. Zhang H, Astrof NS, Liu J-H, Wang J-h, Shimaoka M. Crystal structure of isoflurane bound to integrin LFA-1 supports a unified mechanism of volatile anesthetic action in the immune and central nervous systems. The FASEB Journal. 2009;23(8):2735-40.

143. 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(12 Part 1):2240-9. doi: doi:10.1107/S0907444904013253.