Computational approaches to identify small-molecule inhibitors of non-traditional drug targets

By

Ragul Gowthaman

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

_____
Chairperson Dr. John Karanicolas

_____
Dr. Ilya Vakser

_____
Dr. Eric Deeds

_____
Dr. Joanna Slusky

_____
Dr. Liang Xu

Date Defended: 27 MAY 2015

The Dissertation Committee for Ragul Gowthaman

certifies that this is the approved version of the following dissertation:

Computational approaches to identify small-molecule inhibitors

of non-traditional drug targets

_____

Chairperson Dr. John Karanicolas

Date approved: 27 May 2015

ABSTRACT

Non-traditional targets for therapeutic intervention are those proteins that have not evolved to bind small molecules, but have instead evolved to bind other macromolecules. Such targets include protein–protein interaction sites, protein–RNA interaction sites and protein–DNA interaction sites. Modulating these biologically important targets will allow us as a community to develop novel therapeutics, but still remains a major challenge. In this thesis, I describe two different computational approaches that I have developed: one for identifying small-molecule inhibitors of protein–protein interactions, and the other for identifying small-molecule inhibitors of protein–RNA interactions.

To specifically target protein interaction sites, I have developed a docking method called DARC (*D*ocking *A*pproach using *R*ay-*C*asting). This method quantitatively measures the complementarity between the protein surface and a ligand, by using ray-casting to map and compare their shapes. I have applied DARC to carry out a virtual screen against the protein interaction site of the protein Mcl-1, allowing us to identify 6 new inhibitors of this exciting target. To specifically target protein-RNA interactions, I have developed a mimicry-inspired strategy that extracts a "hotspot pharmacophore" from the structure of a protein-RNA complex, and then uses this as a template for ligand-based virtual screening. I have applied this strategy to screen for compounds that inhibit the Musashi-1 / NUMB mRNA interaction, allowing us to identify a new class of compounds that inhibit this interaction in both biochemical and cell-based assays.

This thesis is outlined as follows. In the first chapter, I will compare the structural features of inhibitor-bound complexes of traditional versus non-traditional protein targets. In the second chapter, I will present the DARC method and its application to Mcl-1. In the third chapter, I will present various enhancements to DARC method that result in both speed and performance improvements. Finally, in the Fourth chapter I will present the "hotspot mimicry" approach for targeting protein-RNA interactions and application of this approach in identification of inhibitors for Musashi 1 / NUMB mRNA interaction.

The text of Chapter 1 is a reprint of the material from:

Gowthaman R, Deeds EJ, and Karanicolas J. (2013) The structural properties of non-traditional drug targets present new challenges for virtual screening. *J. Chem. Inf. Model*, *53*(8), p. 2073-81.

The supporting information for this chapter is included as Appendix A.1.

The text of Chapter 2 is a manuscript that has been submitted for publication as:

Gowthaman R*, Miller SA*, Rogers S, Khowsathit J, Lan L, Bai N, Johnson DK, Liu C, Xu L, Anbanandam A, Aubé J, Roy A, and Karanicolas J. (2015). DARC: mapping surface topography by ray-casting for effective virtual screening at protein interaction sites (in review)
**Authors contributed equally*

The supporting information for this chapter is included as Appendix A.2.

The text of Chapter 3 is a manuscript that has been submitted for publication as:

Gowthaman R, Lyskov S, and Karanicolas J. (2013). DARC 2.0 : Improved docking and virtual screening at protein interaction sites. (in review)

The supporting information for this chapter is included as Appendix A.3.

The text of Chapter 4 is a manuscript that has been submitted for publication as:

Xia Y*, Gowthaman R*, Lan L, Rogers S, Wolfe A, Gomez C, Ramirez O, Tsao BW, Marquez RT, Yu J, Pillai M, Neufeld KL, Aubé J, Xu L, and Karanicolas J. (2015). Rationally designing inhibitors of the Musashi protein–RNA interaction by hotspot mimicry (in review)
**Authors contributed equally*

The supporting information for this chapter is included as Appendix A.4.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. John Karanicolas. His excellent guidance, support and patience helped me overcome many crisis situations and finish this dissertation.

I would like to thank all members of my thesis committee: Dr. Ilya Vakser, Dr. Eric Deeds, Dr. Liang Xu and Dr. Joanna Slusky for their invaluable guidance. I thank Dr. Liang Xu and Dr. Joanna Slusky for willing to serve on my thesis committee at the last moment.

I am thankful to my colleagues David Johnson, Yan Xia, Sven Miller, Karen Khar and Jittasak Khowsathit for collaborating with me on various projects. I am also thankful to my colleagues Andrea Bazzoli, Jimmy Budiardjo and Shipra Malhotra for the many valuable discussions.

Finally I wish to thank my family for their love and support.

TABLE OF CONTENTS

INTRODUCTION

Non-traditional targets for therapeutic intervention:

The most common targets of modern drugs are GPCR's, ion-channels, and enzymes (kinases, hydrolases, oxidases, reductases, etc.) (1). Each of these broad protein classes have evolved to bind some cognate small-molecule, and most therapeutics disrupt activity by competitively binding to the same region of the protein surface as the natural interaction partner. These surface binding pockets are typically deep, and present a well-defined shape that often complements the natural substrate (2). In contrast, there are also a tantalizing number of well-validated protein targets that are not evolved to bind small-molecules, but rather are evolved to bind other macromolecules. These include targets involved in protein-DNA interactions, protein–RNA interactions, and protein–protein interactions. Developing novel therapeutics that modulate these interactions is an important ongoing challenge. Identifying promising new targets in these non-traditional target classes and characterizing their interaction surfaces for druggability is critical, since most of these targets are unyielding to modern approaches for computational or biochemical screening. Indeed, the performance of docking methods used for computational screening depends on the target class of protein: performance is typically poor for non-traditional protein targets in particular, suggesting that new approaches are needed in this regime. The interaction surfaces of these targets are often large and flat, lacking deep pockets typically used for small-molecule binding (2). In my thesis, I will describe two different computational approaches that I have developed: one for identifying small-molecule inhibitors of protein–protein interactions, and the other for identifying small-molecule inhibitors of protein–RNA interactions.

Protein-protein interactions:

Protein-protein interactions (PPIs) underlie almost all biological processes, making them important therapeutic targets. Targeting a protein interface using traditional approaches have met with

limited success, because PPIs are different from traditional enzyme targets (3). Enzymes have deep and well-defined active sites to bind their substrates, whereas PPIs usually have large, flat and noncontiguous interfaces. Further, PPIs do not have natural small-molecule partners that one might use for investigating druggability or as a starting point for designing inhibitors.

In many cases, protein interfaces undergo conformation changes upon binding of an inhibitor (4); hence, for many cases of PPIs the "apo" (unbound) structure is not necessarily suitable for docking, since the protein conformation may change upon binding. On the other hand, there are typically few inhibitor bound structures available for PPI drug targets, and requiring a prior small-molecule bound structure may also limit one to finding hit compounds that are similar to the known (starting) ligand. Some docking methods incorporate receptor flexibility, but these are too slow for virtual screening. There are also many ligand-based methods available for shape matching, that simply use the shape of the ligand to hop for scaffolds (5, 6): the drawback of these methods, however, is that we may miss out on potential inhibitors that do not resemble the natural protein partner.

These shortcomings of existing methods highlight the need for a receptor-based scaffold hopping approach for this challenge. To address this need, I have developed DARC, a method that matches the surface pockets to complementary small molecules.

Protein-RNA interactions:

Protein-RNA interactions play key roles in diverse cellular functions through a variety of different mechanisms, including controlling production of target proteins by post-transcriptional regulation (7-9). Interactions of proteins with their cognate RNAs are mostly driven by base stacking, electrostatic interactions and hydrogen bonding. To date there exist very few classes of compounds that target protein-RNA interactions, and the majority of them are nucleoside analogues (10). The goal of this study is to develop a robust computational method that can be utilized to design small-molecules that selectively inhibit interactions between an RNA binding protein and its cognate RNAs. Here, I will

present the "hotspot mimicry" approach for rationally designing such small-molecule inhibitors of RBPs. This approach works by identifying the chemical moieties of a given RNA that contribute critical interactions to a particular protein-RNA complex, and then identifying small molecules that can mimic the geometrical arrangement of these key moieties.

Summary of the thesis:

Chapter 1 describes the quantitative evaluation of properties that distinguish traditional and non-traditional class inhibitory complexes. I have systematically compared the existing examples of inhibitors of protein–protein interactions to inhibitors of traditional drug targets, and demonstrate that the differences in the components of the inhibitory complexes are manifest in the structural features of the complexes themselves. I also show results of experiments to study how these differences affect the performance of virtual screening tools when applied to protein–protein interaction sites. I demonstrate that the modern virtual screening methods are typically less suited for identifying inhibitors of protein interactions than for identifying inhibitors of traditional drug targets.

Chapter 2 introduces a low-resolution docking method that we call DARC (*D*ocking *A*pproach using *R*ay *C*asting) for targeting protein-protein interactions. I will show results of benchmark experiments demonstrating that we can use DARC to identify known inhibitors of Bcl-xL and XIAP from a large set of "decoy" compounds. I will then show a specific application of DARC in which we to carry out a virtual screen against Mcl-1, an anti-apoptotic member of the Bcl-2 family. We tested the top 21 hits in biochemical assays, and found that indeed 6 of these are inhibitors of Mcl-1.

Chapter 3 presents five key enhancements to the robustness, speed and accuracy of the DARC method. First, we use multiple vantage points to more accurately determine protein-ligand surface complementarity. Second, we describe a new scheme for rapidly determining optimal weights in the DARC scoring function. Third, we incorporate sampling of ligand conformers "on-the-fly" during docking. Fourth, we move beyond simple shape complementarity and introduce a term in the scoring

function to capture electrostatic complementarity. Finally, we adjust the control flow in our GPU implementation of DARC to achieve greater speedup of these calculations. For each enhancement, I will describe systematic improvement relative to the original version of DARC using a "pose recapitulation" benchmark.

Finally, Chapter 4 presents the "hotspot pharmacophore" approach for targeting protein-RNA interactions. I will describe the automated framework that extracts the pharmacophore features from the structure of a protein-RNA complex, and the protocol for using it as a template for ligand-based screening. I will describe our application of this approach to Musashi-1, which has led us to a novel class of inhibitors that disrupt the RNA-binding activity of this tumor-promoting protein.

CHAPTER 1

The structural properties of non-traditional drug targets present new challenges for virtual screening

Ragul Gowthaman[1], Eric J. Deeds[1,2], and John Karanicolas[1,2]

[1] Center for Bioinformatics and [2] Department of Molecular Biosciences,

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

1.1 Abstract

Traditional drug targets have historically included signaling proteins that respond to small-molecules and enzymes that use small-molecules as substrates. Increasing attention is now being directed towards other types of protein targets, in particular those that exert their function by interacting with nucleic acids or other proteins rather than small-molecule ligands. Here, we systematically compare existing examples of inhibitors of protein–protein interactions to inhibitors of traditional drug targets. While both sets of inhibitors bind with similar potency, we find that the inhibitors of protein–protein interactions typically bury a smaller fraction of their surface area upon binding to their protein targets. The fact that an average atom is less buried suggests that more atoms are needed to achieve a given potency, explaining the observation that ligand efficiency is typically poor for inhibitors of protein–protein interactions. We then carried out a series of docking experiments, and found a further consequence of these relatively exposed binding modes is that structure-based virtual screening may be more difficult: such binding modes do not provide sufficient clues to pick out active compounds from decoy compounds. Collectively, these results suggest that the challenges associated with such non-traditional drug targets may not lie with identifying compounds that potently bind to the target protein surface, but rather with identifying compounds that bind in a sufficiently buried manner to achieve good ligand efficiency, and thus good oral bioavailability. While the number of available crystal structures of distinct protein interaction sites bound to small-molecule inhibitors is relatively small at present (only 21 such complexes were included in this study), these are sufficient to draw conclusions based on the current state of the field; as additional data accumulate it will be exciting to refine the viewpoint presented here. Even with this limited perspective however, we anticipate that these insights, together with new methods for exploring protein conformational fluctuations, may prove useful for identifying the “low-hanging fruit” amongst non-traditional targets for therapeutic intervention.

1.2 Introduction

The majority of modern drugs modulate the function of a relatively small number of protein targets that include enzymes, G-protein coupled receptors (GPCRs), ion channels, transporters, and nuclear hormone receptors (1, 11). With the exception of proteases, each of these broad classes of protein have evolved to bind a cognate small-molecule, and most therapeutics disrupt activity by competitively binding to the same region of the protein surface as the natural interaction partner. These surface binding pockets are typically deep and present a well-defined shape that often complements the natural substrate (12). In such cases, mimicry of the natural substrate (or transition state) may serve as an attractive starting point for designing new inhibitors (13). Given that new inhibitors often have similar chemical properties as endogenous ligands, identification of one or more natural ligands with drug-like physicochemical characteristics can also be used to infer the "druggability" of a new protein target (14).

In contrast to these "traditional" protein targets for therapeutic intervention, there are also a tantalizing number of well-validated "non-traditional" potential targets. These proteins have evolved to bind not small-molecules but rather other macromolecules, and include targets involved in protein–DNA interactions (e.g., (15, 16)), protein–RNA interactions (e.g., (17, 18)), and protein–protein interactions (e.g., (19, 20)). The interaction surfaces of these proteins are often large and flat, lacking a deep pocket suitable for small-molecule binding (12). Given the size of the natural substrate in these cases, examples of mimicry by small molecules leading to potent inhibitors are few (21-26), and even in these successful cases the resulting inhibitors tend to be larger than typical orally available drugs (27, 28).

Structure-based virtual screening methods offer a means to directly identify novel inhibitory compounds that complement the target protein surface (29); these methods are not limited by the requirement for template compound(s) implicit to ligand-centric (mimicry) approaches (30). In the simplest terms, virtual screening requires some method for sequentially positioning each candidate compound from a library at its most likely position on the protein surface (i.e., "docking"), followed by a

subsequent discrimination step (i.e., "scoring") to rank each of the resulting complexes based on their likelihood of showing the desired activity.

The historical focus on inhibiting proteins evolved to bind a cognate small-molecule (both enzymes and proteins involved in signaling) has led to an understanding of many structural features exhibited by such complexes (31). These insights have been facilitated in part by databases such as MOAD (32, 33), which have enabled comparisons to reveal subtle differences in enzyme versus non-enzyme classes of "traditional" drug targets (34).

The question now arises whether the same structural features apply to inhibitors that bind at "non-traditional" sites (i.e., those not evolved for small-molecule binding). To identify any systematic differences between these two broad classes of targets is critical, since structure-based virtual screening methods absolutely require these insights to appropriately rank docked complexes and select the most promising compounds for experimental characterization. Since very few examples of direct inhibitors of nucleic acid binding sites have been described, here we instead focus on small-molecule inhibitors of protein–protein interactions. While the widespread impression in the field is that small-molecule inhibitors of protein–protein interactions tend to be larger and contribute less binding affinity per atom than inhibitors of traditional drug targets, this viewpoint is predicated largely on a study that described a relatively small number of examples (4). Here we seek to carry out a more thorough quantitative evaluation of properties that distinguish each class of inhibitor, then ask how these differences affect the performance of virtual screening tools when applied to protein–protein interaction sites.

1.3 Results

Extent of ligand burial in inhibitory complexes

For this study we compiled a set of 21 unique protein–protein interaction sites for which a crystal structure has been solved in complex with a small-molecule inhibitor (**Table 1**), which we will refer to as

the "PPI set." We compared properties of these representative complexes to those of the Astex Diverse Set from Gold (35), which contains crystal structures of proteins of pharmaceutical or agrochemical interest each bound to a small-molecule inhibitor with drug-like chemical properties. The chemical properties of the ligands in the PPI set are comparable to those of the Astex set (the construction and composition of both sets are described in the Methods section).

It has been noted anecdotally that inhibitors of protein–protein interactions bind at flatter regions of the protein surface than do inhibitors of "traditional" drug targets (4, 12). To systematically characterize whether the ligand is less buried upon binding, we defined a parameter θ_{lig} that quantifies the fraction of ligand solvent accessible surface area (SASA) (36) that remains exposed upon binding:

$$\theta_{lig} = 1 - \frac{(SASA_{protein} + SASA_{ligand}) - SASA_{complex}}{2 * SASA_{ligand}} \quad \text{(Eqn. 1)}$$

where $SASA_{protein}$ is the SASA of the protein with the ligand removed, $SASA_{ligand}$ is the SASA of the ligand with the protein removed, and $SASA_{complex}$ is the total SASA of the protein–ligand complex. We note that $SASA_{protein}$ and $SASA_{ligand}$ are each computed directly from the structures that comprise the complex, and not from other unbound crystal structures.

We computed θ_{lig} for each complex in the Astex set and the PPI set; the results of this comparison are presented in **Figure 1**. It is immediately evident that the bound inhibitors at protein interaction sites retain more exposed surface area than their counterparts which bind at sites evolved for small-molecule binding (high θ_{lig} for traditional targets); this observation further holds for other analogous sets of drug-like complexes (DUD-E (37) and SB2010 (38), **Figure S1**). Among the members of the Astex set, we find that the class of protein targets with the highest θ_{lig} values is serine proteases (βII tryptase, factor Xa, factor VIIa, thrombin, urokinase; see **Table S1**). These enzymes contribute four of the ten highest θ_{lig} values in the Astex set, corresponding to statistically significant enrichment of this target class ($p < 0.01$). Although they are enzymes, the natural substrates of serine proteases are proteins rather than small-molecules; for this reason, it is perhaps unsurprising that θ_{lig} values for such complexes resemble those associated with protein interaction sites more than those associated with traditional drug targets.

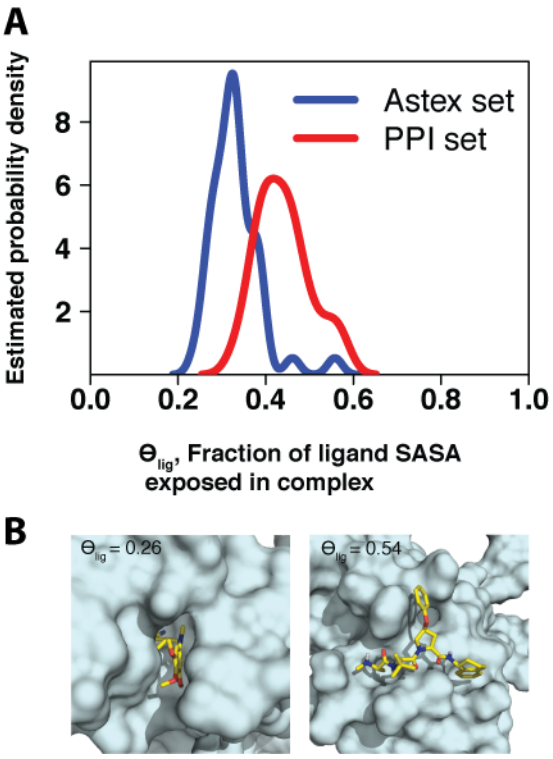


Figure 1 **(A)** Distribution of θ_{lig} values in the PPI and Astex data sets. Ligands bound at protein interaction sites (*red, median value 0.45*) tend to be more exposed than drug-like compounds (*blue, median value 0.33*), with a difference of means that is statistically significant ($p < 10^{-6}$). As described in the text, proteases are an exception in the latter set. **(B)** Representative examples of complexes with low and high θ_{lig} (PDB IDs 1s3v and 1tft, respectively).

Since the set of physical forces that underlie binding must be the same for both classes of complexes, we next asked how this difference in exposed surface area influences binding affinity. We compared the potency of each complex, and found the distributions from the two sets to be essentially the same (**Figure 2a**). Due to the larger compound size required to achieve this potency (**Figure S2**), however, we find that inhibitors acting at protein interaction sites bind with less ligand efficiency (binding energy per non-hydrogen atom (39)) than orally active compounds acting on traditional targets (**Figure 2b**). A single outlier with high ligand efficiency in the PPI set is evident, corresponding to an inhibitor of calpain (1alw in **Table 1**, with ligand efficiency 0.68 kcal/mol per non-hydrogen atom). Further inspection reveals this to be partially an artifact of the way in that ligand efficiency is defined, since this inhibitor contains a bromine atom: while the potency of this compound is high given its size, the ligand efficiency is further exaggerated because it is normalized using the number of non-hydrogen atoms rather than molecular weight. Excluding this outlier, we observe a clear relationship between θ_{lig} and ligand efficiency in the PPI set (**Figure 2c**), but not in the Astex set (**Figure S3**). Due to the variation in the molecular weight of the compounds within each set, this relationship is not apparent when simply examining potency as a function of θ_{lig} (**Figure S4**).

These results are in agreement with a previous report drawn from a much smaller set of protein targets (4), and are also consistent with our observation that the bound inhibitors at protein interaction sites retain more exposed surface area: an average atom buries less hydrophobic surface area upon binding, so the contribution of an average atom to the binding energy—the very definition of ligand efficiency—is expected to be lower (39).

Properties of complexes produced by virtual screening

We have shown above that our collection of inhibitors binding at protein interaction sites have similar potency as a traditional drug-like set, but that affinity is achieved through a structurally distinct

mode of interaction. It is therefore critical to examine how well modern energy functions perform in these contrasting regimes.

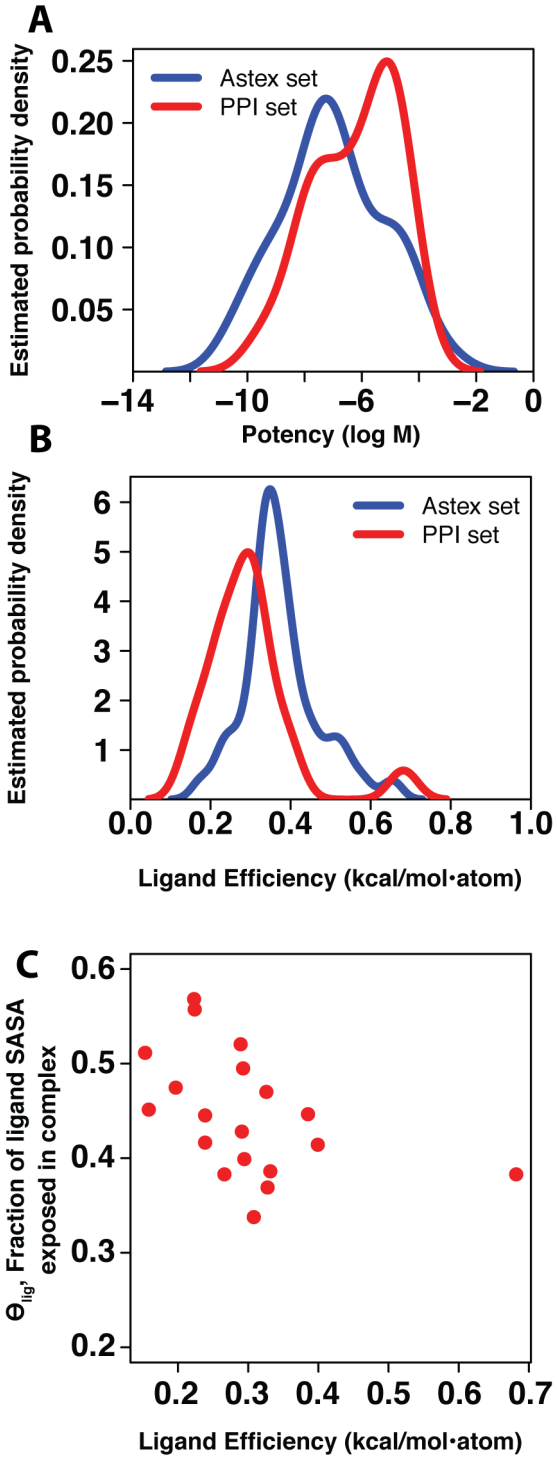


Figure 2 (A) Distribution of potency values in the Astex and PPI sets. Our set of ligands bound at protein interaction sites (*red, median value 1.0 μM*) have similar potency to drug-like compounds (*blue, median value 0.07 μM*). No statistically significant difference in means is observed ($p = 0.105$). **(B)** Distribution of ligand efficiencies in the Astex and PPI sets. The ligand efficiency of inhibitors bound at protein interaction sites (*red, median value 0.29 kcal/mol•atom*) tends to be lower than the ligand efficiency of drug-like compounds (*blue, median value 0.36 kcal/mol•atom*), and the difference in the means is statistically significant ($p < 0.007$, or $p < 0.0002$ upon removal of the single bromine-containing outlier described in the text). **(C)** The relationship between θ_{lig} and ligand efficiency in the PPI set. As expected, there is a negative correlation between these properties, with a statistically significant non-zero Spearman rank correlation coefficient ($p < 0.006$).

To evaluate the ability of a representative energy function to distinguish known inhibitors from a large number of "decoy" compounds, we carried out a mock virtual screening experiment using the FRED software package (40-42). Starting from the ZINC database (43), we constructed a "decoy library" of 10,000 compounds with chemical properties (molecular weight and XlogP) matched to the inhibitors in the Astex and PPI sets. To eliminate potential challenges associated with sampling, we used OMEGA (44-46) to build up to 300 conformers of each compound to be used for docking. We also included the active conformer of the known inhibitor (taken from the crystal structure of the protein–ligand complex) in our virtual screen. The exact active conformer of an inhibitor is generally not present in a screening library, and adding it to the set should, at least in theory, simplify the screening problem. As an indicator of how favorably a protein's known inhibitor is scored, we use its rank relative to the members of the decoy library. As an important caveat, we note that the decoy compounds are *not* necessarily inactive. Nonetheless, we expect that even if the decoy library does contain a small fraction of compounds that are active, the known inhibitor should rank with these among the top scoring compounds.

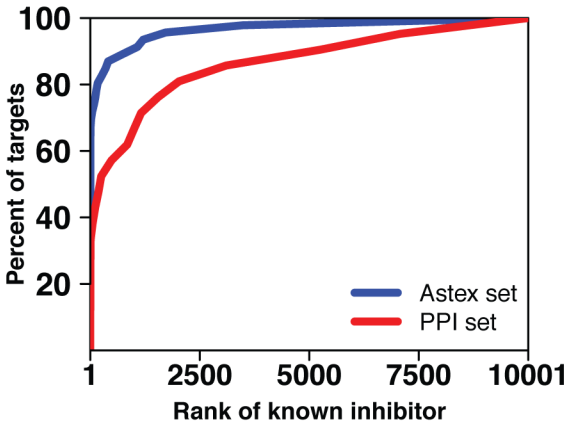


Figure 3 To test the ability of virtual screening tools to distinguish known inhibitors for the protein targets in our test sets, we embedded each inhibitor in a set of 10,000 "decoy" compounds and screened this library against each protein target. For each target, we sorted the docked scores for each member of the screening library and determined the rank of the known inhibitor. In this figure, we plot a cumulative histogram of the percent of protein targets for which the known inhibitor is ranked better than the threshold value indicated on the x-axis. Rather than comparing results for individual protein targets, this aggregate representation allows comparison of performance between the two test sets (Astex and PPI). Here we find that this virtual screening tool has difficulty identifying known inhibitors that bind at protein interaction sites (*red*) relative to its performance in the drug-like set (*blue*), and that the difference in the mean rank of the known inhibitor between the two sets is statistically significant ($p < 10^{-6}$).

We carried out this screening experiment for each complex in the Astex set and the PPI set. Within each set we collected, at increasing values of a rank threshold, the fraction of targets for which the known inhibitor was ranked better than the threshold (**Figure 3**). As can be appreciated from this figure, a perfect method, which would rank the known inhibitor at the top of the list for every target, would lead to a vertical rise at the very left of the curve. In contrast, a completely random method would be expected to rank the known inhibitor in the top 1% for 1% of the targets, in the top 10% for 10% of the targets, etc., leading to a curve that follows the diagonal of the plot area.

The first notable observation is that FRED performs exceptionally well for traditional targets (the Astex set) in this intentionally "easy" experiment. Using illustrative thresholds, the known inhibitor is ranked within the top 2% of the library (within the top 200 of 10,000 compounds) for 80% of the protein targets in the Astex set, suggesting that indeed the known inhibitor is near optimal given the scoring function. In contrast, the known inhibitor is ranked in the top 2% of the library for only 50% of the targets in the PPI set. As shown in **Figure 3**, this difference in performance holds irrespective of the threshold applied—in other words, performance on the Astex set is superior regardless of whether one counts "successes" as protein targets for which the known inhibitor is ranked in the top 5%, top 10%, top 25%, etc. Given the design of this experiment, the dramatically worse performance for the PPI set points strongly to a deficiency in discrimination of the correctly bound inhibitor from bound decoy compounds if the known inhibitor retained a large fraction of exposed surface area upon binding (high θ_{lig}).

To eliminate the possibility that this difference in performance originated from a bias in the composition of our decoy library (e.g., our library could be comprised of decoys that are viable inhibitors of protein interaction sites but easy to rule out as enzyme inhibitors), we carried out an analogous experiment in which we used the DUD-E server (37) to build a separate set of 50 decoy compounds for each target, matched to the chemical properties of the known inhibitor. We observe the same performance difference between the Astex set and the PPI set in this experiment (**Figure S5**) as in the previous screening experiment. The observed difference in performance between these two sets of target proteins

holds when another software package, DOCK 6.6 (47), is used to carry out the virtual screen (**Figure S6**), indicating that particular details specific to FRED are not responsible for the observed differences.

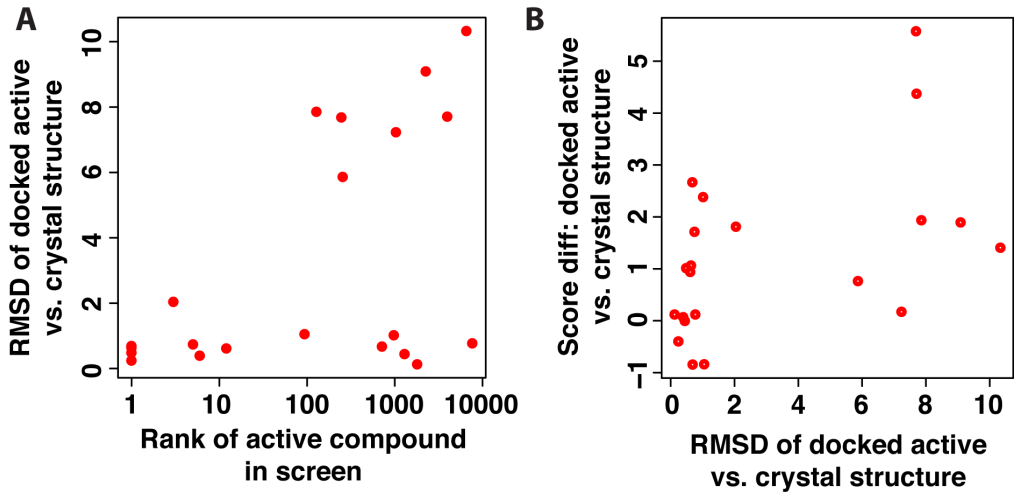


Figure 4 **(A)** For each of the active compounds in the PPI set, the root mean square deviation (RMSD) of the docked active compound is computed relative to the crystal structure, and reported as a function of the rank of the active compound in the virtual screening experiment. Approximately half the compounds that were not ranked highly were mis-docked (*high RMSD*), while the other half were correctly docked but still did not rank well relative to the decoy compounds (*low RMSD but high rank*). **(B)** The difference in score between the docked active compound and the crystal structure is shown as a function of the RMSD. The mis-docked structures scored better than the crystal structure in all cases (*positive score differences*), suggesting that the energy function did not provide the correct relative ranking of these two poses.

Since this screening experiment was explicitly designed to be "easy" with respect to sampling (for example, by including the active conformer taken from the crystal structure), we expected that the poorer performance on the PPI set stemmed from difficulty assigning the correctly docked pose a suitably favorable score. To test this, we collected the RMSD of the top-scoring docked pose of the active compound relative to its position in the crystal structure. We find the active compounds that (correctly) scored amongst the very top of the library collection were inevitably correctly docked (low RMSD) (**Figure 4a**); this is unsurprising, since one might expect that correctly predicting the pose is usually a pre-requisite for correctly identifying an active compound. Of the active compounds which did not score in the top 1% of the library, approximately half were correctly docked (RMSD less than 2 Å) with the other half dramatically mis-docked (RMSD greater than 5 Å). We then scored each of the crystal structures of the same complexes, and computed the difference in score relative to the docked pose. While in a few cases the crystal structures scored slightly better that nearly-correct docked poses, each of the mis-docked poses scored better than the crystal structure (**Figure 4b**, *high RMSD points all have positive score differences*). Collectively these results suggest that the screening challenges presented in the PPI set lie not with sampling, but rather with assigning the correctly docked ligand a suitably favorable score: relative to decoy compounds, and also relative to mis-docked poses of the ligand.

We then selected the top-ranked compound from each screen, whether it was the known inhibitor or a decoy, and evaluated θ_{lig} in the docked complex. Intriguingly, the distributions of these values match the corresponding distributions from complexes solved with the known inhibitors (**Figure 5**). In other words, though the known inhibitors from the PPI set were often not highly ranked in this screening experiment, the highly ranked decoys nonetheless retained a large fraction of exposed surface area upon binding (high θ_{lig}). Thus, the discrimination problem in the PPI set is *not* a bias in favor of low θ_{lig} complexes over high θ_{lig} complexes, but rather a failure to discriminate the known (high θ_{lig}) inhibitor from decoy compounds that also have high θ_{lig} values. In contrast, the more extensive burial of

compounds in the Astex set may lend additional clues for correctly identifying the cognate inhibitor, explaining the relative ease in correctly identifying inhibitors for these targets. These observations further suggest that the protein conformation itself is the primary determinant of θ_{lig} values in inhibitory complexes: unless a pocket suitable for extensive ligand burial is present on the protein surface, it is simply not possible to have an inhibitor with low θ_{lig}. This constraint, together with the poor ligand efficiency associated with high θ_{lig} values, suggests limited druggability of these particular protein conformations.

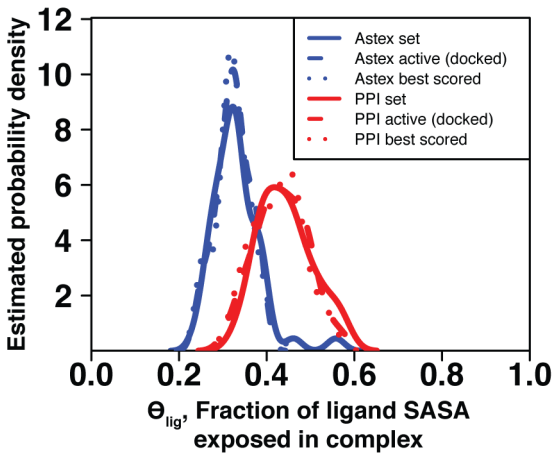


Figure 5 The distribution of ligand solvent exposure for the best-scoring compound in each test set is similar to the analogous distribution for the experimentally-derived complexes of known inhibitors in the corresponding test set. Within a set distributions are very similar for crystal structures of known inhibitors, the known inhibitor docked back to the corresponding protein target, and the top-scoring compound from among 10,000 "decoy" compounds.

We next sought to determine whether this conclusion applies to alternate conformations of these particular protein interaction sites. We have recently described an approach for exploring protein fluctuations enriched in conformations containing surface pockets suitable for small molecule binding (48). We therefore carried out the same experiment described above, this time screening against either the unbound protein structure, the structure of the protein in complex with its protein partner, or a structure generated via biased simulations. We find similar distributions of θ_{lig} for the top-scoring complex when screening against these conformations as well (**Figure 6**), suggesting that these too lack pockets suitable for extensive ligand burial that are required for complexes with low θ_{lig}. On the basis of this result, computational screening against these alternate protein conformations is unlikely to yield inhibitors with dramatically higher ligand efficiency.

1.4 Discussion

Previous analyses have shown that known to be inhibitors of protein–protein interactions can be systematically distinguished from drug-like compounds on the basis of size and shape (49, 50), and also that the deep pockets present on traditional drug targets (sites evolved for small-molecule binding) are typically absent at protein interaction sites (12, 51). Here, we demonstrate the differences in the components of the inhibitory complexes—the compound and the protein target—are manifest in the structural features of the complexes themselves.

Our results further demonstrate that modern virtual screening methods typically are less suited for identifying inhibitors of protein interactions than for identifying inhibitors of traditional drug targets. The difficulty in discriminating active compounds from decoy compounds may be due to an incomplete representation of the underlying physical forces that govern binding in this alternate structural regime; or, more simply, these methods may have been parameterized for optimal performance when applied to traditional drug targets at the expense of performance on non-traditional targets.

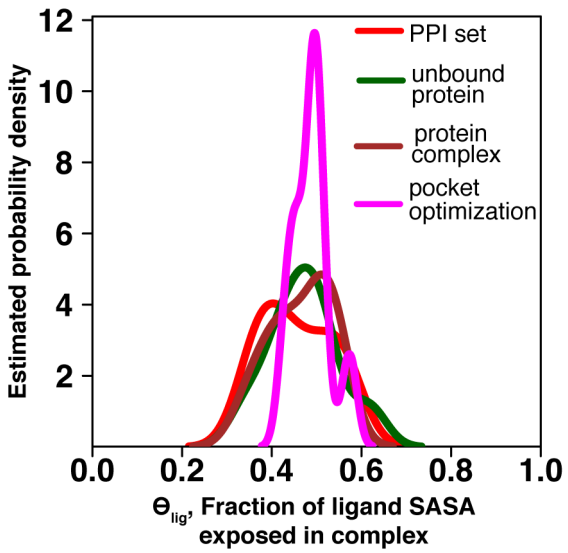


Figure 6 The distribution for the extent of ligand solvent exposure (θ_{lig}) for the best-scoring compound in a screen against PPI targets is similar regardless of whether the protein conformation used in docking corresponds to an inhibitor-bound structure (*red*), the unbound structure (*green*), a protein–bound structure (*brown*), or a structure generated from simulations biased towards pocket-containing conformations (*pink*).

Nonetheless, the results from the virtual screening experiments described here demonstrate that the protein conformations tested can *only* harbor a relatively exposed ligand, because they present relatively flat surface pockets. This feature of bound complexes does not preclude identification of potently binding compounds; rather, it simply implies that such compounds will require many atoms to achieve this potency, and therefore may violate Lipinski's "rule-of-five" criteria for oral availability (27, 28). For enzymes, the rule-of-five compliance of the natural endogenous substrate has been shown to be a good predictor of druggability (14), since inhibitors may occupy similar chemical space. For protein targets not naturally evolved to bind small-molecules, the natural binding partner cannot be used to draw such inferences. Collectively, our results do not suggest that protein interaction sites are necessarily less "bindable" than traditional drug targets, but rather that the size of compounds required for the desired potency may make protein interaction sites intrinsically less "druggable."

Our conclusions are also highly complementary to those reached in a previous study comparing inhibitory complexes of enzyme versus non-enzyme drug targets (34) through a survey of the MOAD database (32). Both types of target are "traditional" by our definition, in that both have evolved to bind small-molecules. Interestingly, the authors of this study find higher ligand efficiency in non-enzyme complexes, and propose that the origin of this difference stems from the fact that non-enzyme ligands (signaling molecules such as hormones) are typically more "encapsulated" than their enzyme counterparts (34). This suggestion is in direct agreement with our observations: ligand efficiency of inhibitors bound at protein interaction sites (median value 0.29 kcal/mol•atom) is lower than that of inhibitors of these "traditional" enzyme and non-enzyme targets (median 0.36 and 0.41 kcal/mol•atom, respectively (34)), in keeping with the limited degree to which the protein surface can accommodate these ligands.

While inhibitors bound to the protein interaction sites studied here can only achieve limited burial (and thus poor ligand efficiency), it is important to note that the protein conformation was fixed throughout these docking experiments. This restriction of the protein surface, in turn, may have influenced our observed recapitulation of θ_{lig} values. Achieving improved ligand efficiency for ligands

binding to protein interaction sites is expected to require that the protein undergo conformational changes to induce formation of alternate pocket shapes, but it is unclear *a priori* what range of pocket shapes are available at a given protein surface. A number of methods have recently been described to preferentially explore alternate conformations of protein surfaces suitable for small-molecule binding (48, 52, 53). While our studies presented here did not find that such conformations would be likely to bind a small-molecule with high ligand efficiency, such an approach may nonetheless prove useful in rapidly screening protein interaction sites for those targets which are not only "bindable", but also "druggable" in this sense. Alternatively, fragment screening approaches (54)—rapidly gaining popularity in campaigns targeting protein interaction sites (55)—may also provide a route for identifying moieties that bind with high ligand efficiency. The existence of such fragments may imply that the protein surface can provide surface pockets suitable for extensively burying at least some ligands.

The number of crystal structures of distinct protein interaction sites bound to small-molecule inhibitors accumulated in the literature to date remains small (**Table 1**), and thus this study provides a snapshot of the field at present from these available examples. Accordingly, we cannot rule out design bias in these set of examples, or insufficient exploration of chemical space in this relatively new field. It will be interesting to learn, over the upcoming years, whether the challenges described in this study can be circumvented through new approaches. Meanwhile, the number of examples of small-molecules directly inhibiting binding at other types of recognition sites is even fewer. Nonetheless, we anticipate that inhibitors of these other "challenging" targets, such as protein–nucleic acid interactions, may share the same poor ligand efficiency—and thus limited druggability—as we describe here for inhibitors of protein–protein interactions. In order to access these many diverse and biologically critical targets, then, we may need to carefully select proteins amenable for complexes with low θ_{lig}, target allosteric sites that can accommodate extensively buried ligands, or develop new approaches to improve the delivery and oral accessibility for compounds outside the "rule-of-five" chemical space.

Protein target	PDB id	Inhibitor molecular weight (Da)	Potency (μM)	Ligand efficiency (kcal/mol per non-hydrogen atom)	θ_{lig}
ZipA	1y2f	424	12	0.22	0.57
HPV E2	1r6n	608	0.18	0.22	0.56
XIAP-BIR3	1tft	535	0.005	0.29	0.52
HIV-gp41	2kp8	580	14	0.15	0.51
S100B	3gk1	279	80	0.29	0.49
IL-2	1pw6	534	6	0.2	0.47
PCNA	3vkx	651	3	0.33	0.47
Grb2-SH2	3in7	556	-NR-	-NR-	0.46
Menin	4gq4	415	0.022	0.39	0.45
VHL	3zrc	410	5.4	0.24	0.45
TNFα	2az5	548	22	0.16	0.45
calmodulin	1ctr	408	1	0.29	0.43
clathrin	4g55	475	12	0.24	0.42
Mdm2	4erf	478	0.0004	0.4	0.41
SHANK PDZ	3o5n	304	17.2	0.29	0.40
integrin	2vc2	523	-NR-	-NR-	0.40
BRD4	2yel	423	0.0155	0.33	0.39
calpain	1alw	308	0.3	0.68	0.38
Bcl-xL	1ysi	552	0.036	0.27	0.38
HIV integrase	4e1n	438	0.019	0.33	0.37
WDR5	3ur4	383	0.45	0.31	0.34

Table 1: Inhibitors bound to protein interaction sites, the PPI set. Potency is taken from reported K_d or K_i values where available; if unavailable, IC_{50} values were used instead. In two cases ("-NR-"), no measure of potency has been reported. θ_{lig} indicates the fraction of ligand SASA exposed in the complex, as defined in Equation 1.

1.5 Methods

Protein data sets

To build the PPI set, we merged entries from the 2P2I (56) and TIMBAL (57) databases and then supplemented this collection with additional recent examples from our own curation of the literature. We eliminated from consideration any complexes with inhibitors of molecular weight outside the range 200 Da – 675 Da, to discard small fragments and large peptide-like compounds. We further excluded structures containing proteins with co-factors, proteins closely related in sequence to other members of our set, and all examples of covalently bound inhibitors. In cases where more than one suitable inhibitor-bound structure has been solved, we retained the structure in complex with the most potent ligand; we have carried out the experiments described here with other complexes as well and find similar results (data not shown). Our set contains compounds with molecular weight range 279 – 651 Da (mean: 468, standard deviation: 102 Da) and XlogP range of -4.9 – 6.5 (mean: 2.6, standard deviation: 2.9). Overall our set contains structures of 21 non-redundant proteins, each in complex with a unique inhibitor bound at a protein interaction site (**Table 1**): to our knowledge this represents the largest such collection reported to date.

The Astex diverse set contains 85 protein–ligand complexes (35), intended as examples of drug-like complexes (indeed, 23 of the ligands are approved drugs). For appropriate comparisons to our PPI

set, we refined this set by removing examples containing a protein cofactor or second ligand at the binding site (typically NADPH, FAD, ATP, heme, or metal ions). We again removed any complexes with inhibitors outside the molecular weight range 200 Da – 650 Da. Overall the Astex set we used in these studies consisted of 46 binary protein–ligand complexes (**Table S1**), with molecular weight range 208 – 576 Da (mean: 362, standard deviation: 89 Da) and XlogP in the range -3.8 – 5.4 (mean: 1.9, standard deviation: 2.0).

Molecular properties were calculated using the MolProp toolkit version 2.1.5 (OpenEye Scientific Software, Santa Fe, NM) (58).

Computational screening

Starting from the ZINC database of commercially available compounds (43), we compiled a set of 10,000 randomly selected compounds with molecular weight between 200 and 750 Da. These compounds served as "decoy" compounds for the screening experiment (**Figure 3**).

In a separate experiment, we instead used the DUD-E server (37) to generate a set of 50 "decoy" compounds specifically matched to the known inhibitor of the intended target; in other words, each protein target was screened against a custom compound library built to match the chemical properties of the known inhibitor of this protein target (**Figure S5**).

For each decoy compound as well as for the known inhibitor, we used the OMEGA version 2.4.3 software (44-46) to generate up to 300 conformers, using default parameters. Charges were added using the QuacPac version 1.5.0 toolkit (OpenEye Scientific Software, Santa Fe, NM) (59). As described earlier, we included the ligand conformation observed in the crystal structure among the conformers for the known inhibitor.

We carried out computational docking using the FRED version 3.0.0 software (40-42), a part of the OEDocking suite (OpenEye Scientific Software, Santa Fe, NM). The binding site for each protein was defined based on the bound ligand in the crystal structure, using the OEDocking 'receptor_setup' utility.

All FRED docking runs were carried out using the default parameters and the resulting poses were ranked using the Chemgauss4 score.

Additional docking calculations were carried out using DOCK version 6.6 (47). The molecular surface for each protein target was calculated using the dms program, and spheres were generated in the active site using the sphgen program (both are available as DOCK utilities). All spheres within 10 Å of the bound ligand from the crystal structure were retained. A grid was generated and the ligand was docked using the default parameters; then the grid score was used to rank each docked complex.

The same set of decoy compounds were used for the experiment using FRED and the analogous experiment using DOCK.

Alternate protein conformations

Alternate pocket-containing protein conformations were generated via biased simulations as described elsewhere (48). Briefly, we applied a biasing potential proportional to the size of the surface pocket, and carried out Monte Carlo simulations adapted from refinement protocols used in comparative modeling applications. This methodology is implemented in the Rosetta software suite (60), and is freely available for academic use (www.rosettacommons.org).

Simulations were carried out for the following list of proteins (with PDB code for the unbound starting structure used in the simulation and the "target" residue at which the biasing potential was applied): Bcl-xL (1r2d, 141), BRD4 (2oss, 146), FKBP12 (2ppn, 26), HPV E2 (1r6k, 33), IL-2 (1m47, 42), HIV integrase (3l3u, 178), Mdm2 (1z1m, 61), menin (4gpq, 278), XIAP-BIR3 (1f9x, 308), ZipA (1f46, 85). In each case 1000 output structures were generated, and virtual screening was then performed using the lowest-energy pocket-containing conformation.

Statistical analysis

Probability densities were estimated using standard kernel techniques in the R statistical computing environment (61).

The statistical significance of differences in means between distributions was evaluated using a standard two-tailed nonparametric permutation test. Briefly, this test works by combining the two data sets in question, D_1 and D_2 (with n_1 and n_2 observations, respectively) into a single data set D (with $n = n_1 + n_2$ elements). A pair of data sets, D1' and D2', are then created by random sampling from D. These data sets contain the same number of elements as the original data sets (i.e., n_1 and n_2), but are constructed from a random partitioning of the original observations. For a given randomized pair, we can calculate the difference in means as $t_R = | \mu(D_1') - \mu(D_2') |$, where $\mu(X)$ is the sample mean of data set X. This procedure is repeated to generate N randomized replicates, producing a distribution of t_R values. This represents an estimate of the distribution of differences in means, given the null hypothesis that D_1 and D_2 are drawn from distributions of the same mean.

This distribution is then compared to the observed difference in means, $t_{obs} = | \mu(D_1) - \mu(D_2) |$, in order to estimate a p-value of that observed difference. In the set of N random replicates, we calculate the number of replicates N_{geq} where $t_R \geq t_{obs}$. The p-value is then defined as $p = N_{geq} / (N + 1)$. For all p-values quoted in this work, the test outlined above was performed using the "twotpermutation" method implemented in the DAAG package in R (61), with $N = 10^6$ replicates.

1.6 Acknowledgements

We are grateful to Andrea Bazzoli for critically reading the manuscript and offering helpful suggestions. We are grateful to OpenEye Scientific Software (Santa Fe, NM) for providing an academic license for the use of FRED, OMEGA, MolProp, and QuacPac. We are grateful to the developers of the DOCK software for providing an academic license for the use of this program. This work was supported

by a grant from the National Institute of General Medical Sciences of the National Institutes of Health (R01GM099959), the National Science Foundation through TeraGrid allocation TG-MCB130049, and the Alfred P. Sloan Fellowship (J.K.).

CHAPTER 2

DARC: mapping surface topography by ray-casting for effective virtual screening at protein interaction sites

Ragul Gowthaman[1†], Sven A. Miller[2†], Steven Rogers[3], Jittasak Khowsathit[2], Lan Lan[2], Nan Bai[2], David K. Johnson[1], Chunjing Liu[3], Liang Xu[2,4], Asokan Anbanandam[5], Jeffrey Aubé[3,6], Anuradha Roy[7], and John Karanicolas[1,2]

[1] Center for Computational Biology, [2] Department of Molecular Biosciences,

[3] Center of Biomedical Research Excellence, Center for Cancer Experimental Therapeutics,

[4] Department of Radiation Oncology, [5] Biomolecular NMR Laboratory,

[6] Department of Medicinal Chemistry, and [7] High Throughput Screening Laboratory

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

† indicates authors of equal contribution

2.1 Abstract

Protein-protein interactions represent an exciting and challenging target class for therapeutic intervention using small molecules. Protein interaction sites are often devoid of the deep surface pockets presented by "traditional" drug targets, and crystal structures reveal that inhibitors typically engage these sites using very shallow binding modes. As a consequence, modern virtual screening tools developed to identify inhibitors of traditional drug targets do not perform as well when they are instead deployed at protein interaction sites. To address the need for novel inhibitors of important protein interactions, here we introduce a new docking strategy specifically designed for this regime. Our method, termed DARC (*D*ocking *A*pproach using *R*ay-*C*asting), matches the topography of a surface pocket "observed" from within the protein to the topography "observed" when viewing a potential ligand from the same vantage point. We applied DARC to carry out a virtual screen against the protein interaction site of human anti-apoptotic protein Mcl-1, and found that 6 of the top-scoring 21 compounds showed clear inhibition in a biochemical assay. The Ki values for these compounds ranged from 1.2 to 21 μM, and each had ligand efficiency comparable to promising small-molecule inhibitors of other protein-protein interactions. These hit compounds do not resemble the natural (protein) binding partner of Mcl-1, nor do they resemble any known inhibitors of Mcl-1. Our results thus demonstrate the utility of DARC for identifying novel inhibitors of protein-protein interactions.

2.2 Introduction

Protein-protein interactions comprise the underlying mechanisms for cell proliferation, differentiation, and survival; manipulation of these interactions thus represents a promising avenue for therapeutic intervention in a variety of settings. Despite the urgent need for small-molecule modulators of protein interactions, however, as recently as five years ago the dearth of success stories meant that even the druggability of this target class was very much in question (3, 62). At the same time, the challenges associated with these important targets help spur refinement of innovative new techniques, including fragment-based methods (63), peptidomimetics (64), and non-peptidic mimetics of protein secondary structural elements (65, 66). By now, these advances have enabled discovery of inhibitors against many different protein-protein interactions, and have led to an impressive array of compounds in various stages of clinical trials and pre-clinical optimization (67, 68).

Modern methods for structure-based virtual screening have proven effective for a variety of applications, providing novel hits to address a very wide range of "traditional" targets such as enzymes and G protein-coupled receptors (69, 70). Unfortunately, structure-based virtual screening tools for identifying small-molecule inhibitors of protein-protein interfaces have lagged behind: due to their relatively exposed binding modes, the same docking methods that perform well for traditional drug targets struggle to correctly identify compounds that disrupt protein interactions (71). This problem is particularly acute because the natural partners of these targets – proteins – cannot easily be used for inspiration in designing small-molecule inhibitors or as templates for ligand-based virtual screening, as is possible for many enzymes (72, 73) and G protein-coupled receptors (74).

In a number of cases, protein structures have been solved both in complex with a biological protein partner and also in complex with a small-molecule inhibitor. Comparison of the unbound protein structure to the equivalent structure in complex with a small molecule shows that binding is not associated with a large conformational change; and yet, the concave pocket on the protein surface in

which the small molecule binds is often absent in the unbound structure, so that the flat surface of the unbound structure must often undergo local rearrangement to reveal the small-molecule binding site (4). This has confounded structure-based methods for identifying inhibitors – how can the protein structure be used to identify an inhibitor, if the protein structure will change upon binding?

We recently described a method for identifying protein surface pockets suitable for small-molecule binding, from either co-crystal structures or from simulations starting from the unbound protein structure (75). Here, we introduce a method that we call DARC (*D*ocking *A*pproach using *R*ay-*C*asting) for matching these surface pockets to shape-complementary small molecules. Since many inhibitors of protein interactions are found to bind in shallow flat pockets on the protein surface (4, 71), this method is specifically designed for such protein-ligand interactions. Further, our method is intrinsically low-resolution, to allow for minor conformational variation of the protein upon ligand binding. We anticipate this method will provide suitable starting points for cases in which the protein structure is not perfectly optimal for the ligand, and cases in which further small conformational changes of the protein are expected to accompany ligand binding.

2.3 Computational Approach

Computational methods are implemented in the Rosetta software suite (60); Rosetta is freely available for academic use (www.rosettacommons.org), with the new features described here included in the 3.6 release.

Scoring with DARC

To map the pocket shape on the protein surface, we use a variation of the LIGSITE (76) algorithm described fully in our earlier work (75). In its most basic form, this method involves building a three-dimensional grid around a protein (or select region of a protein), then tagging the grid points that are covered by the protein. Next, each grid point not covered by the protein is examined to determine whether

the point is bounded on two opposing sides by the protein along any direction of the grid, including diagonals. A continuous group of such special points are labeled a "pocket." We then restrict this collection of "pocket" points to those "surface pocket" points that immediately contact the protein: these points together define the concave surface with which a ligand may interact.

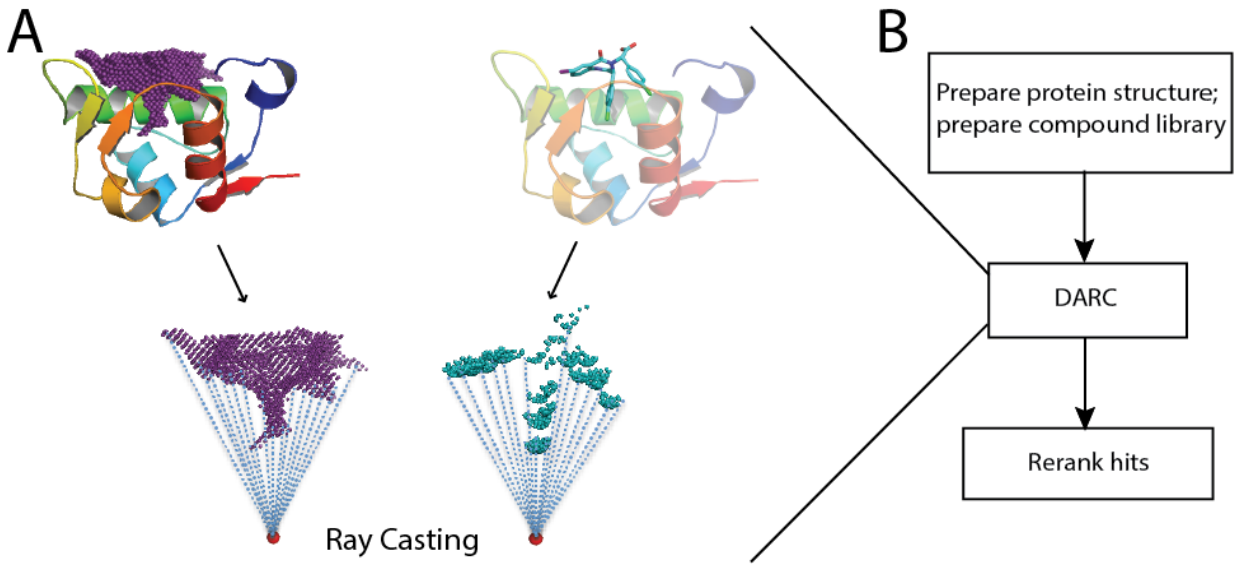


Figure 1: Docking Approach using Ray-Casting. (A) DARC first casts a set of rays emanating from an origin within the protein (*red dot*), and maps the topography of the surface pocket by monitoring intersection of these rays with the pocket (*left*). To evaluate the shape complementary of a given ligand for this pocket, DARC casts the same rays (from the same origin) and monitors their intersection with the ligand (*right*). If ligand (in its current position and orientation) is perfectly shape-complementary to the pocket, each ray will intersect the ligand at the same distance from the origin as it intersected the protein surface pocket. By moving the ligand to maximize this shape complementarity, a ligand can be docked into a protein surface pocket. **(B)** A schematic diagram of the complete workflow split into three stages: pre-DARC preparation, DARC, and post-DARC re-ranking/filtering of the docked models.

From an "origin" point within the protein (determination of this point is described in *Supplementary Methods*), we then cast a series of rays towards the surface pocket; each ray connects the origin point to one "surface pocket" point (**Figure 1a**). For each ray we express the position of the "surface pocket" point relative to the origin using spherical coordinates ($\theta,\phi,\rho_{pocket}$); collectively these spherical coordinates represent a map of the binding pocket's topography when viewed from the protein interior.

When viewed from the same vantage point (the origin), the topography of a suitably-bound ligand is expected to be highly complementary to that of the protein's binding pocket. To evaluate the degree of complementary, we therefore cast the same set of rays (defined by their angles θ and ϕ), and determine the distance at which each ray first intersects the ligand (**Figure 1a**). If the ligand is indeed complementary to the pocket, the distance to the ligand (ρ_{ligand}) for a given ray will closely match the distance to the pocket (ρ_{pocket}) for the corresponding ray.

We define the DARC "overlap score" as the degree to which the pocket and ligand topographies match, by summing over all rays as follows:

$$DARC\ overlap\ score = \sum_{rays} \left\{ \begin{array}{cc} c1*\left(\rho_{ligand} - \rho_{pocket}\right) & if\ \rho_{pocket} < \rho_{ligand} \\ c2*\left(\rho_{pocket} - \rho_{ligand}\right) & if\ \rho_{ligand} < \rho_{pocket} \\ c3 & if\ ray\ does\ not\ intersect\ ligand \\ c4 & if\ ray\ does\ not\ intersect\ pocket \end{array} \right\} \quad (1)$$

From a physical perspective, rays that reach the pocket before the ligand indicate underpacking between protein and ligand (*condition #1*); alternatively, rays that reach the ligand before the pocket indicate a steric clash (*condition #2*). Rays that do not intersect the small molecule at all are penalized as well (*condition #3*); from a physical perspective, this corresponds to the ligand not completely filling the pocket. Finally, we include a series of rays that do not intersect the pocket; intersection of these rays with the ligand indicates that the ligand is too large for the pocket (*condition #4*).

For a given set of rays, evaluating the intersection distance of each ray with the ligand is a task that is naturally very amenable to parallel computing; indeed, the need to efficiently carry out the related "ray tracing" operation provided motivation for development of graphics processing units (GPUs). We have recently demonstrated that the ray-casting approach we describe here can also be carried out on GPUs, leading to dramatic speed enhancement relative to the analogous calculation on a CPU alone (77).

Determination of the values for parameters c1/c2/c3/c4 used in these studies is described in *Supplementary Methods*.

Docking with DARC

Determining the optimal bound pose of a small molecule within a pocket (i.e. "docking") entails varying the ligand position and orientation to minimize the DARC score (**Equation 1**). Though the first derivative of the DARC score can be calculated with respect to the ligand position and orientation, we instead carry out a derivative-free minimization using the particle-swarm optimizer implemented within the Rosetta software suite (60). The ligand internal degrees of freedom are held fixed during this search; to account for ligand flexibility, we prebuild ligand "conformers" using OMEGA (44-46) and independently dock each of these. The final DARC score for a given ligand is taken to be the minimum DARC score from among each of its docked conformers.

Application of DARC for virtual screening is summarized though the workflow in **Figure 1b**. The starting pocket shape derives from a protein crystal structure, and can be an unbound structure, a bound structure in which the ligand has been discarded, or a structure generated from simulation (e.g. the "pocket optimization" simulations we have described for identifying low-energy pocket-containing conformations (75)).

2.4 Results

Building a virtual screening benchmark

Comparing DARC to other methods for virtual screening requires development of a benchmark to evaluate the performance of each method for correctly identifying known inhibitors from among "decoy" compounds. Assembling a large benchmark set is complicated by the relatively small number of available crystal structures of small-molecule inhibitors of protein interactions that have been solved in complex with their protein targets. Rather than assemble many tests with a very small number of active compounds (known inhibitors), we instead built two separate benchmark tests around protein interaction sites for which numerous chemically diverse inhibitors have been reported: Bcl-xL's interaction with Bak peptides, and XIAP's interaction with Smac/Diablo peptides. These two proteins (and their homologs) were excluded from parameterization of DARC (see **Table S1** and *Supplementary Methods*), to ensure that DARC would not inadvertently be trained to perform well against these two protein targets.

When generating a set of "decoy" compounds for this screening experiment, it is important that these compounds be suitably matched to the "active" compounds. If the decoy compounds are systematically different in some physicochemical property, for example, a docking method may successfully pick out active compounds simply due to some implicit bias (e.g. by simply picking the largest compounds, or the most hydrophobic, etc.). On the other hand, it is important to consider that decoy compounds are not necessarily inactive, but rather "presumed inactive": typically they have not been explicitly tested experimentally against the target proteins in the benchmark to ensure they are not active. If decoy compounds are too similar to active compounds in chemical structure, certain decoys may themselves be active, which will confound analysis of the results (because the "right answer" in the benchmark is to label decoy compounds as inactive).

To generate well-matched decoy compounds for these benchmarks, we drew from either i) a set of 85 drug-like compounds that comprise the Astex Diverse Set (35), or ii) a set of 6912 compounds

reported to inhibit a different protein-protein interaction than our protein targets, compiled from the TIMBAL database (78, 79).

To reduce the likelihood that "decoy" compounds might be active, we removed any decoy compounds that exhibited strong chemical similarity to a compound in the "active" set for this target (2D Fingerprint Tanimoto score (80) > 0.7). To ensure non-redundancy of the active set and non-redundancy of the decoy set, we also removed any compound with strong chemical similarity to another compound already present in the same set (2D Fingerprint Tanimoto score > 0.7).

For Bcl-xL this approach yielded 27 (diverse) active compounds, with 33 decoy compounds from the Astex set or 328 decoy compounds from the TIMBAL set. For XIAP this yielded 14 (diverse) active compounds, with 45 decoy compounds from the Astex set or 425 decoy compounds from the TIMBAL set. Though smaller than the TIMBAL decoy sets, the Astex decoys have the advantage that these compounds have typically been advanced through further optimization for their respective targets; this makes the Astex decoys less likely to exhibit (undesired) activity against Bcl-xL or XIAP.

In addition, we also generated a third set of decoy compounds using the DUD-E server (81). In this case, we identified 50 "custom" property-matched decoys from each active compound. Thus, the 27 compounds active against Bcl-xL led to a set of 1350 DUD-E decoys, and the 14 compounds active against XIAP led to a set of 700 DUD-E decoys. The careful matching of the decoys' physicochemical properties to those of specific active compounds makes this a much more challenging benchmark; at the same time, however, there is an increased likelihood that some decoy compounds may themselves be active against the target protein.

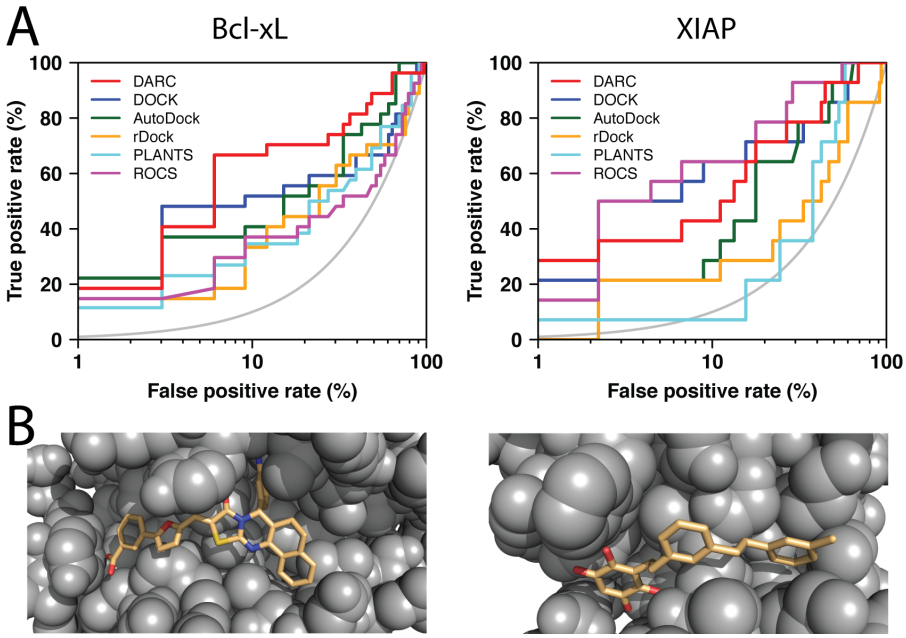


Figure 2: Virtual screening benchmark experiment using an inhibitor-bound protein structure. **(A)** This receiver operating characteristic (ROC) plot compares the performance of various methods (DARC, DOCK, AutoDock, rDock, PLANTS, and ROCS) for predicting whether a given compound is active against a particular protein interaction site. Drug-like "decoy" compounds were drawn from the Astex diverse set, then further filtered to remove any compounds that are similar in chemical structure to an active compound or any other decoy compound. The results are presented on a semi-log plot to highlight the "early" performance of the methods; the grey curve indicates the random retrieval of compounds (i.e. a random predictor). *Left*: discriminating 27 diverse compounds active against Bcl-xL from among 33 matched decoy compounds. *Right*: discriminating 14 diverse compounds active against XIAP from among 45 matched decoy compounds. **(B)** DARC-docked models of representative active compounds against Bcl-xL (*left*) and XIAP (*right*). These compounds were not chemically similar to those used in solving these crystal structures (such compounds were excluded from this benchmark).

Screening against a ligand-bound protein structure

The challenge in virtual screening entails identifying as many active compounds as possible from a library (true positives), while minimizing the number of inactive compounds incorrectly assigned as active (false positives). A very stringent score cutoff for any method will result in few compounds being assigned as active (such that some active compounds will be missed), whereas relaxing this cutoff will lead to many more compounds assigned as active (such that more false positives will be included). Since only a small fraction of the total compound library is expected to be further evaluated or validated experimentally, a typical goal for virtual screening is to rank active compounds at the very top of the sorted list (good "early" behavior), corresponding to the ability to build a small subset of the original library that is strongly enriched in active compounds (82). The "late" behavior (the fraction of the library that must be screened to ensure no active compounds are missed) is irrelevant for virtual screening applications, since a screening tool that offers to eliminate only a small fraction of (inactive) compounds from a large library is not particularly useful.

We started from the crystal structures of Bcl-xL and XIAP, each solved in complex with an inhibitor that had been excluded from our library of active compounds. In addition to DARC, we used five other methods to score and rank each compound in this benchmark experiment. Four of these (DOCK (83, 84), AutoDock (85), rDock (86), and PLANTS (87)) are receptor-based screening tools that we used to dock each compound in the protein interaction site of each protein. The fifth (ROCS (88-90)) is a ligand-based screening tool, which ranks compounds on the basis of how well their three-dimensional structure mimics the volume and chemical features of a template – in this case, the structure of the inhibitor solved in complex with the target protein. Parameters used for each of these other packages are included in the *Supplementary Methods*.

Given the rank order of each compound using each of the three methods, we prepared separate receiver operating characteristic (ROC) curves for both Bcl-xL and XIAP, corresponding to an experiment in which we screen a library comprised of active compounds and Astex decoys (**Figure 2a**).

Each point on such a plot corresponds to a given score cutoff, and each point indicates the fraction of active compounds collected at this cutoff, as a function of the inactive compounds incorrectly assigned as active at this cutoff. On this plot, the critical "early" behavior described earlier corresponds to a steep rise at the leftmost part of the plot: a useful method for screening must pick out as many active compounds as possible while accumulating a very small fraction of the inactive compounds. All three methods outperform the behavior that would result from randomly ranking the compounds in the library (*grey curve*), especially in this "early" region. We also observe similar results when carrying out the same screening experiment using decoy compounds drawn from the TIMBAL set rather than from the Astex set (**Figure S1**). With the exception of ROCS applied to XIAP, meanwhile, none of the methods perform at a comparable level when used to distinguish known active compounds from their closely-matched decoys in the DUD-E set (**Figure S2**).

As seen in representative examples of docked active compounds (**Figure 2b**), the DARC-generated models exhibit clear shape complementarity that allowed these compounds to be (correctly) identified as true ligands; experimentally-derived structures of these compounds in complex with their protein partners have not been reported to date. While low-resolution considerations of shape complementarity alone cannot be sufficient to reliably predict binding affinity or details of the interactions, it is striking that for these two benchmark tests the simple approach employed by DARC successfully distinguishes active from inactive compounds with similar accuracy as more sophisticated methods. We further note that the performance of DARC in this benchmark is remarkably robust to the location of the "origin" point from which rays emanate, provided a reasonable approach is used to define this point (**Figure S3**). Thus, these observations support the hypothesis that the crude pocket shape on the protein surface provides a strong restriction on the chemical "shape space" of inhibitory compounds.

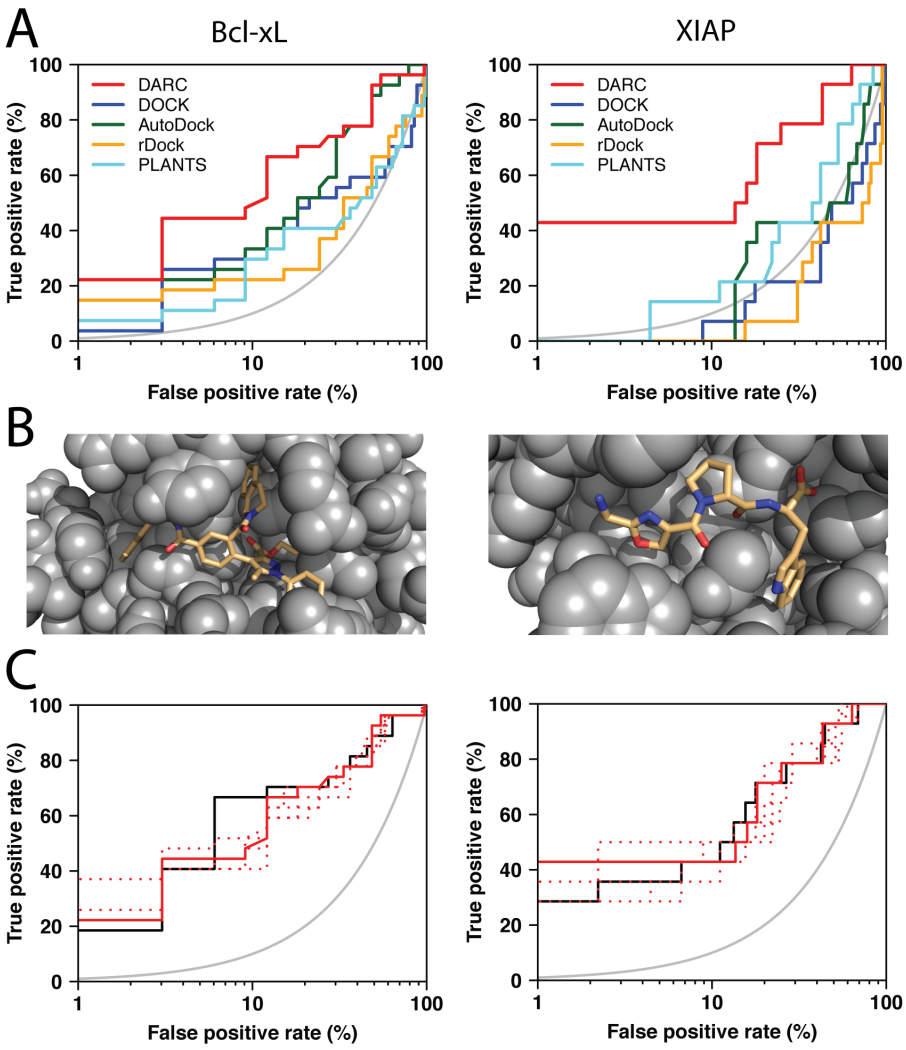

A
Bcl-xL
XIAP
DARC
DOCK
AutoDock
rDock
PLANTS
True positive rate (%)
False positive rate (%)
B
C

Figure 3: Virtual screening benchmark experiment using protein structures from "pocket optimization" simulations. (A) This benchmark experiment involves discrimination of the same active versus "decoy" compounds as in **Figure 2**, however this time compounds were screened against a protein conformation generated via "pocket optimization" simulations initiated from an unbound crystal structure (instead of protein conformations from an inhibitor-bound crystal structure). The results are again presented on a semi-log plot to highlight the "early" performance of the methods; the grey curve indicates the random retrieval of compounds (i.e. a random predictor). **(B)** Models of representative active compounds docked by DARC to the "pocket optimized" conformation of Bcl-xL (*left*) or XIAP (*right*). **(C)** The performance of DARC in this benchmark is essentially equivalent when screening against a known ligand-bound structure (*solid black line*), the lowest-energy "pocket optimized" conformations (i.e. as described in the previous panels) (*solid red line*), or any of five other low-energy "pocket optimized" conformations (*dashed red lines*); this observation demonstrates the insensitivity of the method to details of the protein structure.

Screening against a "pocket optimized" protein structure

The fact that a protein interaction site is likely to undergo a conformational change upon inhibitor binding (4) makes the choice of receptor conformation of particular interest for this target class, especially given that the performance of most approaches decreases considerably when an unbound starting structure is used in place of a ligand-bound starting structure (91). A variety of methods exist for incorporating receptor flexibility, typically by docking against multiple receptor conformations from either crystal structures (92-96) or simulation (97-99).

Rather than start from unbound crystal structures of Bcl-xL and XIAP – which lack the surface pocket required for inhibitor binding – we instead started from conformations generated using the "pocket optimization" approach we have recently described (75). Briefly, we developed a biasing potential that

drives conformational sampling in molecular mechanics simulations towards conformations containing surface pockets suitable for small molecule binding. This biasing potential operates solely on simple geometric considerations, so that the shapes of the resulting surface pockets are influenced solely by properties of the protein surface and *not* knowledge of any particular ligand. Starting from unbound crystal structures, we carried out simulations to generate "pocket optimized" conformations of Bcl-xL and XIAP: conformations containing a surface pocket, but without the precise details that would be encoded in a ligand-bound structure (our application of this approach to generate these protein conformations is described in *Supplementary Methods*).

We re-screened the same compound library described earlier, this time using a "pocket optimized" conformation of Bcl-xL or XIAP (**Figure 3a**). We did not include ROCS in this stage of the benchmark, since there is no structure of a cognate ligand available for this protein conformation that could be used as a template. Unsurprisingly, the performance of DOCK, AutoDock, rDock, and PLANTS all suffer dramatically in this more challenging regime. In contrast, the performance of DARC is remarkably similar to the earlier experiment, demonstrating that the low-resolution nature of the underlying calculations makes DARC relatively robust to slight mismatches between protein and ligand. Indeed, representative examples of active compounds docked to these alternate conformations demonstrate that the overall pocket shapes remain complementary to these ligands (**Figure 3b**). We note that these observations also hold when carrying out the same experiment using decoy compounds drawn from the TIMBAL set rather than the Astex set (**Figure S4**), but that again all of these methods exhibit diminished performance against the DUD-E set (**Figure S5**).

In summary, slight differences in these protein conformations lead to diminished performance from other receptor-based docking methods in this regime, but DARC performs equally well in this experiment as in the previous benchmark that used protein conformations from ligand-bound crystal structures. This finding is confirmed by extending this experiment to consider a series of five more distinct low-energy conformations for each protein, as generated by the "pocket optimization" approach:

we find that the performance of DARC in this benchmark is notably insensitive to the precise details of the protein conformation (**Figure 3c**).

Collectively, these experiments point to different regimes in which each class of method is expected to prove superior. Detailed docking methods, such as DOCK, AutoDock, rDock, and PLANTS, are expected to out-perform DARC if the protein structure is solved in complex with a ligand of similar chemotype to the desired compounds. By contrast, DARC allows greater discrimination if the protein conformation is not quite optimal for the ligand: this can occur when one wishes to discover inhibitors with a radically different chemotype than known ligands, or when the protein conformation derives from simulation.

Identifying novel Mcl-1 inhibitors using DARC

To evaluate the performance of DARC in a realistic application, we next applied this approach to screen for novel classes of inhibitors for Mcl-1, a member of the Bcl-2 family of proteins. Individual members of this protein family can serve either a pro- or anti-apoptotic role, and interact with one another through a structurally-conserved binding motif. Small-molecule inhibitors of Bcl-2/Bcl-xL have shown promising efficacy in overcoming chemo/radioresistance in various tumor models including prostate cancer (100-104). One such compound is ABT-737, a potent Bcl-2/Bcl-xL inhibitor (104); a more recent Bcl-2-selective derivative of this compound is currently in clinical trials (105). Recent studies have shown that cancer cells resistant to ABT-737 have high levels of Mcl-1, and that knockdown of Mcl-1 promotes ABT-737-induced apoptosis (89, 102, 103, 106, 107). Taken together, these observations motivate the pressing need for development of potent and selective inhibitors of Mcl-1 in treating of a variety of cancers, to be used either as a single agent or in combination with inhibitors of Bcl-2/Bcl-xL (108).

The strong evidence supporting the potential impact of effective Mcl-1 inhibitors has spurred intense efforts using a variety of complementary approaches (108, 109), including high-throughput screening of standard libraries using a fluorescence polarization competition assay (110, 111), and of an

sp^3-rich library using differential scanning fluorimetry (112). Aiming to build upon the success of the ABT-737 series for Bcl-2, several groups have since applied similar NMR-based fragment screening to Mcl-1; fragment hits were prioritized on the basis of ligand efficiency (binding free energy per heavyatom), first yielding weakly-binding compounds that were subsequently merged or elaborated to give inhibitors with improved potency at the expense of some ligand efficiency (113-116). Inspired by the helical conformation adopted by Mcl-1's interaction partners, a number of groups have grafted the interaction partners' sidechains onto small-molecule scaffolds that allow them to be presented in a similar geometry (117-119), or have used "stapled" derivatives of the helical peptides (120-122). Such approaches predicated on mimicry of a known helical binding partner can rapidly lead to potent inhibitors, but the large chemical scaffolds required to recapitulate the helical geometry diminish from their ligand efficiency.

Like other members of the Bcl-2 family, Mcl-1's interaction partners bind to an exposed hydrophobic groove on the protein surface (123). We recently applied our "pocket optimization" approach to this protein surface, and generated ensembles of low-energy Mcl-1 conformations that present surface pockets suitable for small molecule binding (124). Among these ensembles we find conformations similar to those observed in crystal structures of Mcl-1 bound to diverse small-molecule ligands, and we also observe "distinct" conformations that have not yet been observed in any experimentally-derived structures (124). We elected to use DARC to screen chemical libraries for compounds that would complement one of two Mcl-1 conformations: either an experimentally-derived conformation, or a conformation derived from "pocket optimization" simulations. At the time we initiated these studies no unbound structure was available (125), so for the former we used a peptide-bound crystal structure (PDB ID 2pqk) (123). For the latter, we used the lowest-energy pocket-containing conformation generated from 1000 independent trajectories.

Drawing from the ZINC database (43), we compiled a small virtual library corresponding to 62,442 highly-diverse compounds with drug-like properties (MW $\leq$ 500 Da, xlogP $\leq$ 5, etc., see

Supplementary Methods) available for immediate purchase from a commercial vendor. We then applied DARC to separately dock and rank each compound using the peptide-bound structure of Mcl-1, and also separately using the "pocket optimized" conformation of Mcl-1.

The docked models produced by DARC are intrinsically low-resolution, since they rely only on matching the protein-ligand topography. To achieve further high-resolution discrimination, we therefore included an additional final step (**Figure 1b**): for each of the top ranked 10% of the complexes produced by DARC we carried out fullatom gradient-based minimization in Rosetta, using the standard Rosetta fullatom energy function (60) and ligand parameter definitions (126). All internal dihedral angles of the protein were included as degrees of freedom, along with the ligand position and orientation relative to the protein. We then filtered the resulting models to remove those with no intermolecular hydrogen bonds or an abundance of buried unsatisfied polar groups. From the remaining models, we then purchased the top-scoring 21 compounds on the basis of DARC score (**Table S2**). Of these 21 compounds, 11 were identified by screening against the (peptide-bound) crystal structure of Mcl-1, and the other 10 were from screening against the "pocket optimized" structure.

The structural basis for the favorable scores in our DARC screen is evident from the models of these compounds in complex with Mcl-1 (**Figure 4a**). In each case the ligand – in its modeled conformation – exhibits exquisite shape complementarity for the protein surface; this is unsurprising, given that the DARC score is expressly built to identify the ligand position, orientation, and conformation that will maximize similarity to the topography of the protein surface. Nonetheless, there is also clear diversity amongst the models: the ligand shapes are different from one another, and they fill the protein surface pocket in distinctly different ways. In the case of M0, for example, a hydrophobic groove on the Mcl-1 surface is neatly complemented by the spatial arrangement of the two aromatic rings in the selected ligand conformation.

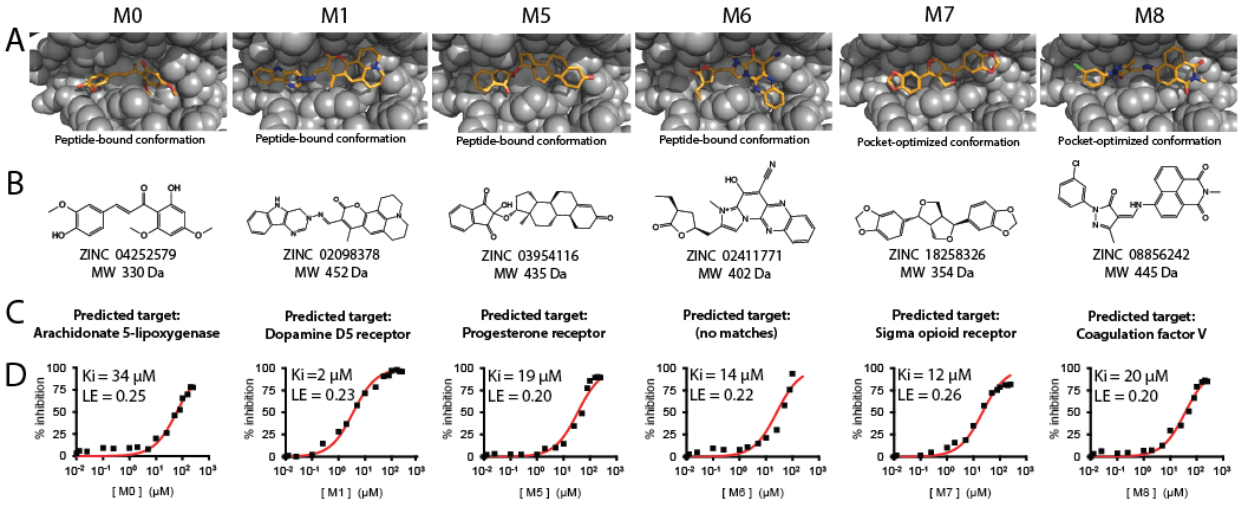


Figure 4: DARC can be used to identify novel inhibitors of human Mcl-1. Screening was carried out against the protein interaction site of Mcl-1, and 21 of the top-scoring compounds were tested in a biochemical assay. **(A)** Docked models generated by DARC for six compounds emerging from our screen. Compounds were screened either against an experimentally-derived peptide-bound structure of Mcl-1, or against an Mcl-1 conformation generated via "pocket optimization" simulations. **(B)** The chemical structures of the compounds in each model are shown. We note that some of these have reactive functional groups: such compounds could be removed from the library prior to screening, in applications where such moieties are undesirable. **(C)** The most likely activity of each compound was predicted using the Similarity Ensemble Approach (SEA). None of these compounds are assigned as likely inhibitors of Bcl-2-family proteins, underscoring their lack of similarity to known inhibitors of this family. **(D)** Each of these six compounds inhibit Mcl-1's interaction with a FITC-labeled cognate peptide, as determined via fluorescence polarization. CHAPS detergent (0.1%) was included in this assay, to ensure that these results were not due to compound aggregation; we also observe equivalent inhibition in the absence of CHAPS. Each curve is fit to a one-site inhibition model, with a single free parameter in the fitting.

The selected compounds also represent a diverse array of different chemical scaffolds (**Figure 4b**), with no evident similarity to one another. It must be noted that of the compounds shown, only M7 (a natural product) is devoid of potentially reactive functional groups. This is, of course, a reflection of the screening library: in a typical drug discovery application, such compounds should be removed from the library prior to screening since their potential for advancement is likely to be extremely limited (127). For the purpose of evaluating the performance of DARC in this virtual screening experiment, however, we did not exclude such compounds.

By visual inspection, the structures of these compounds bear no obvious resemblance to any known inhibitors of Mcl-1, or to any known inhibitors of any other Bcl-2 family members. To systematically identify biologically active compounds most resembling these compounds, we applied the Similarity Ensemble Approach (SEA); this method uses ligand similarity to identify likely receptors for a query ligand (128). In most cases SEA identifies chemical scaffolds with clear resemblance to our queries, but these related compounds have each been described in the context of very different activities (**Figure 4c**); even delving further into each list, we did not find similarity to other inhibitors of Bcl-2 family members. This observation highlights the fact that these compounds would have been very difficult to pick out using ligand-based screening, due to their lack of similarity to any known inhibitors of this protein family. The most likely activity of each compound was predicted using the Similarity Ensemble Approach (SEA) (128).

We next examined the ability of each compound to inhibit the interaction of Mcl-1 with one of its binding partners. We used a fluorescein-labeled peptide derived from the Noxa protein, which exhibits an increase in polarization upon Mcl-1 binding (**Figure S6a**). Upon addition of a known Mcl-1 inhibitor, AT-101 (i.e. R-(-)-gossypol), we observe a dose-dependent decrease in polarization that confirms this compound competes with Noxa for Mcl-1 binding, and is consistent with previously reported data (129) (**Figure S6b**). Using this assay we find that 6 of the 21 compounds from our computational screen (M0, M1, M5, M6, M7, and M8) inhibited the interaction between Mcl-1 and its cognate peptide, with Ki

values ranging from 1.2 to 21 μM (**Figure 4d, Table 1**). We also observe activity for 4 additional compounds (M11, M12, M17, and M20), with Ki values ranging from 35 to 100 μM (**Table 1**). For the remaining 11 compounds, we did not observe any activity.

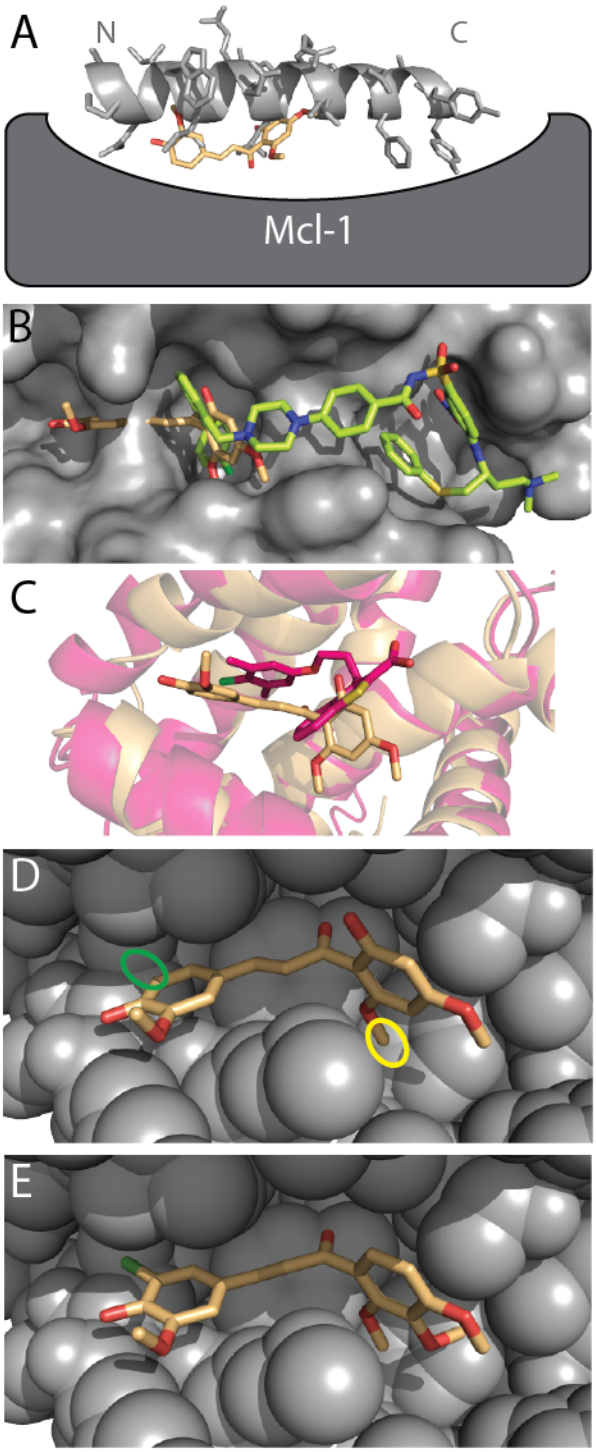


Figure 5: Evaluating DARC's model of the M0/Mcl-1 complex. (A) Though M0 was identified in a screen against a peptide-bound conformation of Mcl-1, M0 (*peach*) is distinctly *not* mimicking the interactions of the cognate peptide (*light gray*). In the DARC-generated model M0 competes with the N-terminus of the peptide for Mcl-1 binding, but fits more deeply into the Mcl-1 surface pocket than the peptide sidechains. **(B)** This M0 binding mode is distinctly different from that of known Bcl-xL inhibitors such as ABT-737 (*green*), which instead overlap with the C-terminus of the peptide-binding site. The surface of the Bcl-xL protein is shown (*grey*), solved in complex with ABT-737. **(C)** Other recently-described inhibitors of Mcl-1, including the representative example shown here (*dark pink*), bind at a similar location to M0; however, each distinct series of inhibitors induce their own unique Mcl-1 conformational change to allow deeper ligand burial. **(D)** The locations of two M0 positions that play key roles in our model of the complex: the 6'-methoxy group (*yellow*) and the unsubstituted 5-position (*green*). **(E)** A model of the D1, the most potent M0 derivative, in complex with Mcl-1. This compound preserves the crucial 6'-methoxy group, and embellishes the 5-position with a chlorine substitution to improve packing. Closely analogous compounds to D1 that lack the methoxy group at the 6'-position (D3) or that lack the chlorine at the 5-position (D5) exhibit reduced activity.

Given the chemical structures of the six most potent compounds, these observed Ki values correspond to ligand efficiency (binding free energy per heavyatom) of 0.21 to 0.27 kcal/mol·heavyatom, which is typical of other promising small-molecule inhibitors of protein-protein interactions reported in the literature (68). With respect to other Mcl-1 inhibitors in particular, these six initial hits are less active than the most potent compounds arising from screens of libraries designed to mimic the sidechain interactions of helical cognate peptides (117-119); however, the DARC hits exhibit greater ligand efficiency, due to the extensive size of the compounds designed for mimicry of the helix.

Superposition of the modeled complexes for each of the DARC hits with the structure of Mcl-1 bound to a cognate peptide reveals that these compounds are decisively *not* recapitulating the interactions of the helical peptide (**Figure 5a**). While the cognate peptide uses a cluster of aliphatic residues at its N-terminus and aromatic residues at its C-terminus, compound M0 (as a representative example) occupies only the regions of the Mcl-1 surface engaged by the N-terminus of the peptide. Given that it is not constrained to fit onto the backbone of the helix, this compound also fits more deeply into the binding pocket on the Mcl-1 surface. In contrast, the most potent of the helix-mimetics was designed to recapitulate the complete set of interactions provided by the helical template (117). The sidechains that comprise these interactions are separated by a span of 15 residues in sequence, and while this led to an inhibitor with Ki of 0.2 μM, the large chemical scaffold needed to present these interactions reduced the ligand efficiency of this compound to 0.21 kcal/mol·heavyatom.

While the underlying methodology used in our DARC screen did not explicitly select for compounds that bind in place of the N-terminus of the cognate peptide, each of the resulting hits complemented this region. This is in stark contrast to the binding modes observed in crystal structures of Bcl-xL solved in complex with potent inhibitors such as ABT-737 and WEHI-539: each of these bind to the region of the protein surface used by the C-terminus of the cognate peptide, and thus have relatively little overlap with the models of the DARC hits (**Figure 5b**). In the time since we carried out our computational screen, crystal structures of Mcl-1 solved in complex with four distinct classes of inhibitor have been reported (112-114, 130), and in each case the ligand occupies the region of the protein surface corresponding to the N-terminus of the cognate peptide. Thus, the lack of DARC hits complementing the protein surface used by inhibitors of Bcl-xL may indicate a lack of druggability for the analogous surface of Mcl-1.

In contrast to Mcl-1 inhibitors derived by mimicry of the cognate helix, the most promising hits emerging from fragment screens initially proved less potent than those identified by DARC but had superior ligand efficiency (113-116). One representative class of fragments inhibited Mcl-1 with Ki of

only ~100 μM, albeit with a promising ligand efficiency of about 0.3 kcal/mol·heavyatom (113). NMR-derived models of the binding mode for this class of compound suggested that they induce a conformational change in Mcl-1, which in turn produces a very deep pocket required for binding; upon merging with another fragment class, crystallography confirmed that the larger compound also induced this conformational change (113). In retrospect, each of the Mcl-1 crystal structures in complex with an inhibitor (solved after completion of our screen) revealed a conformational change that allowed the ligand to access a deep pocket that was not evident from the unbound or peptide-bound structures (112-114, 130) (**Figure 5c**). The observation that extensive ligand burial is related to high ligand efficiency (across many small-molecule inhibitors of protein-protein interactions) (71) is thus consistent with the impressive ligand efficiency of these (deeply buried) Mcl-1 fragment hits. However, such conformations were not among those included in our DARC screen, and accordingly the DARC hits could not have taken advantage of the deep pockets presented upon conformational rearrangement of the protein.

Evaluating the DARC model of M0 binding

To gauge the veracity of the modeled complexes that led to selection of these compounds, we sought to ask whether such a model could be used to explain the structure-activity relationship (SAR) observed across a series of related compounds. Despite the fact that M0 (2',4-dihydroxy-3,4',6'-trimethoxychalcone) was the least potent of the six hits in our original screen, the straightforward synthetic accessibility of the chalcone series prompted us to select this scaffold for further investigation.

Before designing a series of analogs, however, we noted that other chalcones are known to form covalent adducts to proteins (the chalcone α,β-unsaturated carbonyl system is a Michael acceptor and thus may react with unpaired cysteine sidechains) (131, 132). We therefore carried out analysis via surface plasmon resonance (**Figure S7**) and HSQC chemical shift mapping (**Figure S8**) to further characterize this interaction (see *Supplementary Results*). Taken together, the stoichiometric Hill coefficient (133) and insensitivity to the presence of detergent (134) in the fluorescence polarization assay, the fact that the

SPR response did not exhibit super-stoichiometric behavior (135), and the observation of a small number of distinct chemical shift differences in the HSQC spectrum (136) all provide evidence that the observed activity of M0 is not due to compound aggregation that deactivates Mcl-1, but results from specific binding. The NMR experiment also supports two features of our model for M0 binding: first that the M0 binding mode is not completely overlapping with that of the Bcl-xL inhibitors (**Figure 5b**), and second that M0 binding does not induce extensive protein conformational changes as seen in other Mcl-1 inhibitors (**Figure 5c**).

We then proceeded to produce a variety of M0 derivatives, using a simple aldol condensation reaction (**Figure S9**). We were concerned that designing these analogs from our model of the M0/Mcl-1 complex might bias the resulting series towards certain regions of chemical space, and thus not provide a fair assessment of our model. In order to evaluate the structure-activity relationship more objectively, we therefore designed a set of compounds by selecting the most readily available starting materials, based only upon the chemical structure of M0 and without consideration of our model of the complex. Using the same fluorescence polarization competition assay described earlier, we characterized a total of 27 analogs of M0 (**Table 2, Figure S10**). As described below, the SAR deduced from this series presents a compelling narrative when interpreted using our model of M0 binding (**Figure 5d**).

The most potent of these M0 analogs are D1 and D2: they each inhibit Mcl-1 about 7-fold more potently than M0. Compounds D1 and D2 differ only by a single methoxy versus fluorine substitution at the 3-position; M0 harbors a methoxy here. This methoxy group is exposed in our model of the M0/Mcl-1 complex, consistent with the observation from D1 versus D2 that its replacement with fluorine does not affect activity. Compounds D1 and D2 together share three differences relative to M0: both have a chlorine at the 5-position, both have a methoxy group at the 5'-position, and both lack the 2'-hydroxyl of M0. The latter two positions are exposed in our model of M0 binding, and are thus assumed not to strongly affect activity. In contrast, the 5-position (unsubstituted in M0) faces towards the protein, and

points towards a small cavity (**Figure 5d**): thus, the improved packing resulting from this substitution may explain the slightly enhanced potency of D1/D2 relative to M0.

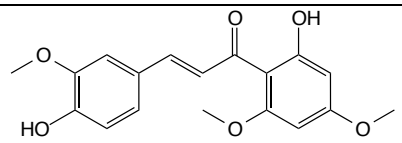
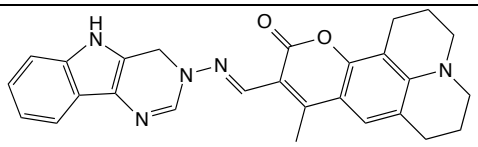
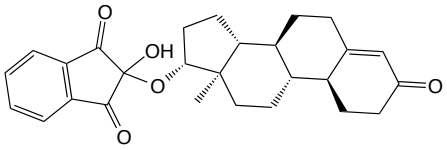
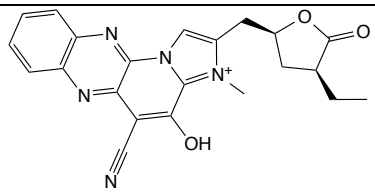
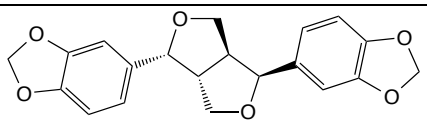
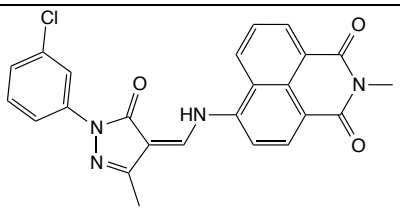
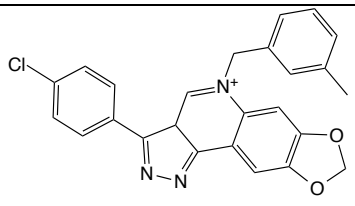
The only other two compounds in this series with potency exceeding that of M0 are D3 and D4; this equipotent pair again differs only by the same fluorine versus methoxy substitution that distinguished D1 and D2. Relative to D1/D2, compounds D3/D4 lack the 5'- and 6'-methoxy groups. While the 5'-position is unsubstituted and exposed in our model of the M0/Mcl-1 complex, the 6'-methoxy group is deeply buried in a hydrophobic cavity (**Figure 5d**); loss of this interaction may explain the reduced potency of D3/D4 relative to D1/D2.

Across this series of 27 M0 analogs, only compounds D1-D4 proved more potent than M0. While the SAR described above is consistent with our model of M0 binding, these results are also notably *inconsistent* with the behavior expected if these compounds were simply carrying out thia-Michael additions. Kinetic studies using a model thiol (cysteamine) demonstrate that the presence of the 6'-methoxy group decreases reactivity, and the 2'-hydroxy group increases reactivity (137). Through the comparisons above, we find that the most potent compounds for inhibition of Mcl-1 correspond to those expected to be *least* reactive, further implying that the inhibition we observe is not due to reactivity of these compounds with Mcl-1's unpaired cysteine sidechain.

Through the results presented in **Table 2** we demonstrate that chalcones are not "privileged scaffolds" (138, 139) for inhibition of Mcl-1, an assertion that is best supported by specific examples. Compounds D7/D12/D13/D14 lack the key 6'-methoxy group described earlier (**Figure 5d**), and accordingly are less potent than M0. Compound D5 preserves each of the M0 functional groups that interact with Mcl-1 in our model (i.e. the 6'-methoxy group and the 4-hydroxyl group), and accordingly exhibits a very similar inhibition constant as M0. Compound D8 maintains the substituents of D2 on one ring, but lacks the substituents on the other ring (including the 6'-methoxy group), and is thus less potent. Compounds D9/D10/D11 maintain the 6'-methoxy group but harbor a variety of alternate substituents on the other ring, making them less potent as well. Compounds harboring extra fused rings

(D20/D21/D22/D23) are inevitably less potent, as are compounds in which the chalcone linkage has been replaced with a flavone (D24/D25/D26/D27). Overall, the assertion that chalcones are not privileged scaffolds for inhibition of Mcl-1 or generic helix-mimetics is supported by the observation that most analogs presented here – designed without consideration of our model of binding – inhibit Mcl-1 less potently than the parent compound M0. In summary then, it is the precise complementary of select compounds for the surface of Mcl-1 that dictate their activity, and not some property of this chemical scaffold.

Ultimately, direct structural evidence will be required to confirm that these compounds are indeed binding via the designed pose. Unfortunately, our efforts to crystallize Mcl-1 in complex with a member of this chemical series have not yet proven successful. Validating the models that lead to selection of active compounds through structural biology will be important not only for retrospective evaluation of the DARC models, but also as a starting point for inspiring design of subsequent analogs. Accordingly, we used DARC to compare our model of the M0/Mcl-1 complex to a model in which M0 was replaced with D1 (**Figure 5e**). We find that DARC scores D1 even more favorably than M0, due to the improved packing resulting from substitution of chlorine at the 5-position. Thus, we expect that compounds with even better shape complementarity for Mcl-1 than those selected in our initial screen – compounds such as D1 – can be identified by screening larger chemical libraries with DARC. By further analogy to D1, we anticipate that such compounds will also exhibit superior potency relative to the initial hits described in this first screen. Despite the improved potency relative to M0, we nonetheless note that compound D2 does not exert the desired biological effect using an *in vitro* cellular assay (see *Supplementary Results*) (**Figure S11**). This underscores the need to test candidate inhibitors in cellular assays as well, since biochemical activity may not necessarily translate to cellular activity. Indeed, careful analysis has shown that a number of other inhibitors, each with comparable binding affinity for Mcl-1 as that of compound D2, also appear not to act on Mcl-1 in cells (111, 140, 141).

Compound	Chemical structure	Ki (μM)	L.E. (kcal/mol·heavyatom)
M0		21	0.27
M1		1.2	0.24
M5		12	0.21
M6		9	0.23
M7		7	0.27
M8		12	0.21
M11		35	0.20

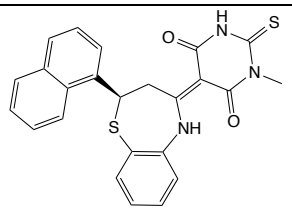
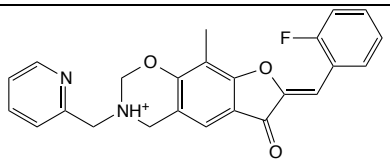
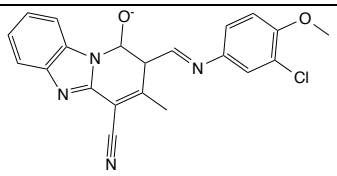
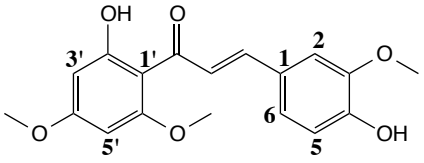
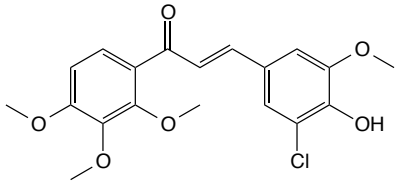
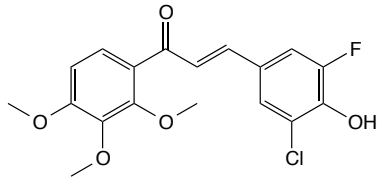
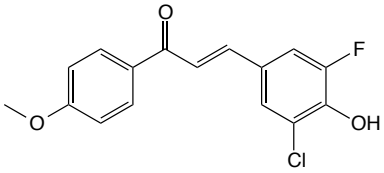
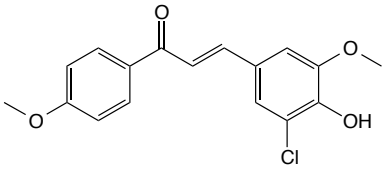
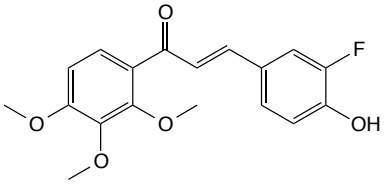
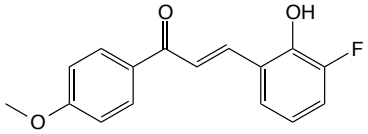
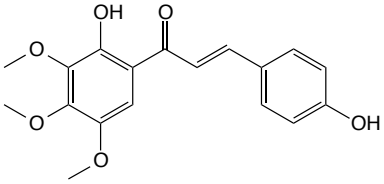
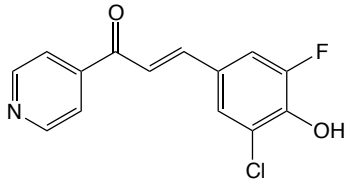
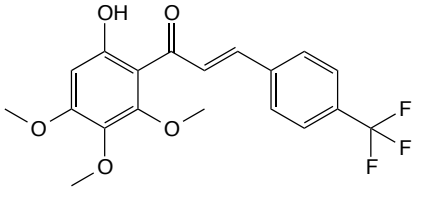
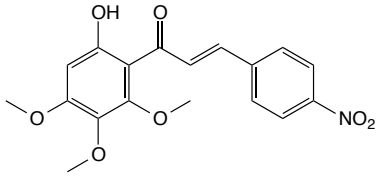
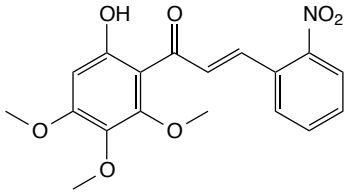
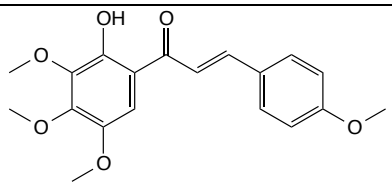
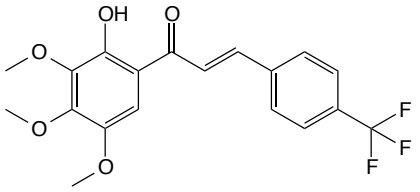
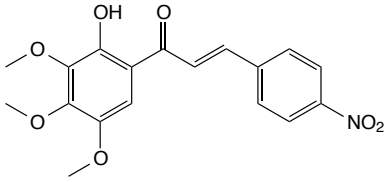
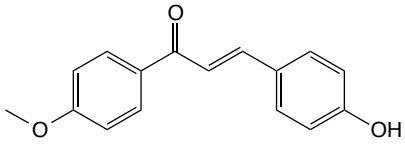
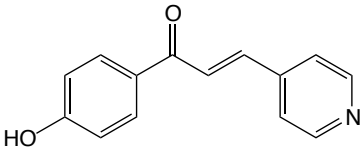
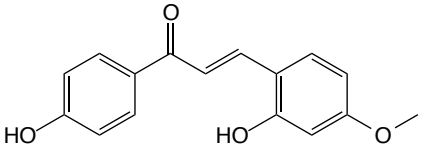
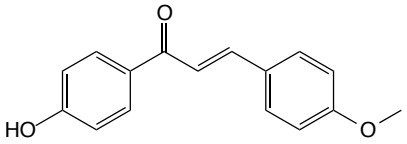
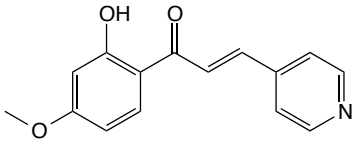
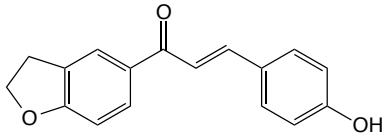
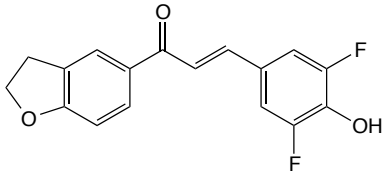
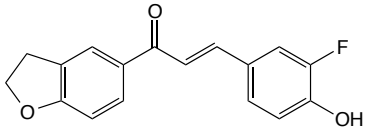
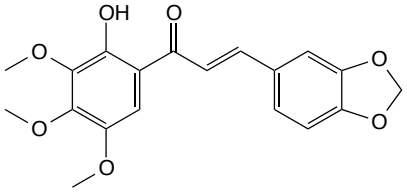
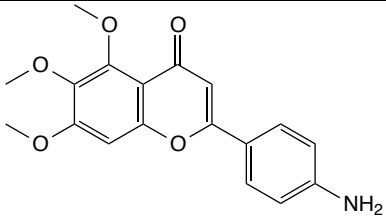
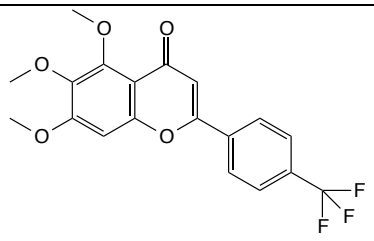
M12		~80	0.19
M17		43	0.20
M20		~100	~0.19

Table 1: DARC screening hits confirmed experimentally as inhibitors of Mcl-1. The complete set of DARC hits (including those that did not inhibit Mcl-1) are included as **Table S2**. Ki values are determined via a fluorescence polarization competition assay (see *Supplementary Methods*).

Compound	**Chemical structure**	**Ki (μM)**	**L.E. (kcal/mol•heavyatom)**
M0		21	0.27
D1		2	0.29

D2		3	0.30
D3		7	0.33
D4		10	0.31
D5		28	0.33
D6		20	0.32
D7		40	0.25
D8		>50	N.D.
D9		>50	N.D.

D10		>50	N.D.
D11		>50	N.D.
D12		>50	N.D.
D13		>50	N.D.
D14		>50	N.D.
D15		>50	N.D.
D16		>50	N.D.
D17		>50	N.D.

D18		>50	N.D.
D19		>50	N.D.
D20		>50	N.D.
D21		>50	N.D.
D22		>50	N.D.
D23		>50	N.D.
D24		>50	N.D.
D25		>50	N.D.

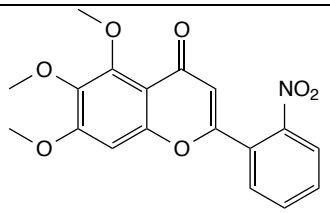
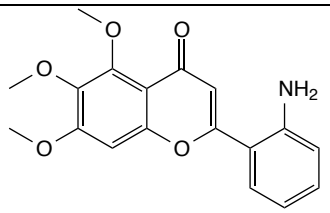
D26		>50	N.D.
D27		>50	N.D.

Table 2: Chemical analogs of M0. We designed and synthesized a series of 27 M0 derivatives, and tested each of these for inhibition of human Mcl-1 using a fluorescence polarization competition assay (see *Supplementary Methods*). This series of analogs was designed without consideration of DARC's model of the M0/Mcl-1 complex, to allow unbiased evaluation of the model.

2.5 Discussion

Screening on the basis of surface topography

Since the protein surfaces of "traditional" drug targets have evolved to bind some cognate small-molecule ligand, they typically include a deep pocket or groove that can be targeted by inhibitors. In contrast, protein interaction sites typically do not include such surface features, and accordingly small-molecule inhibitors acting at these sites rely on shallower bound poses (71). Because these shallow binding modes prove challenging to predict by conventional docking approaches, modern virtual screening tools exhibit diminished performance when used to predict inhibitors of protein interactions (relative to their performance when applied to "traditional" drug targets) (71). Here we present DARC, an entirely new approach to docking and virtual screening. By matching the topography of the protein surface to the buried face of the ligand, we find that DARC outperforms popular canonical tools (DOCK, AutoDock, rDock, and PLANTS) when screening for inhibitors of protein-protein interactions.

Given the importance of shape complementation for binding, several other fast approaches have been described to rapidly evaluate poses and enable large-scale virtual screening. These include methods based on Fourier correlation theory (142, 143), spherical harmonics (144, 145), geometric hashing (146), and negative images (147-149). Most recently, others have used variations of ray-casting to compare internal pockets in proteins to one another (150, 151). In addition to their intended usage, a subtle – but very important – difference between the latter work and DARC is the origin from which rays emanate. This other study casts rays from within the pocket (at the center of mass), resulting in a shape description that is most useful when the pocket is mostly (or completely) enclosed by the protein (150, 151). In contrast, DARC casts rays that originate from "behind" the pocket (i.e. inside the protein): this instead emphasizes shape complementarity at the deepest regions of the binding groove, and is thus better suited for describing the shallow bound poses typical of small-molecule inhibitors acting at protein interaction sites.

While shape complementarity is clearly a necessary feature of ligands that will bind to a protein surface, this alone obviously cannot be sufficient. In the present study, we used DARC to optimize and evaluate surface shape complementarity without consideration of electrostatic complementarity or solvation effects; for this reason, we included a "re-ranking" step in our virtual screen against Mcl-1 to filter out models with no intermolecular hydrogen bonds or an abundance of unsatisfied buried polar groups. In retrospect, the hydrophobicity of the Mcl-1 surface pocket provided a convenient testing ground for the ability of DARC to identify shape-complementary ligands, without the complication of polar groups on the protein surface. In future, however, we expect that incorporation of electrostatic complementarity into DARC lead to improved performance in virtual screens against protein surfaces that display surface polar groups (152).

Approaches for identifying inhibitors of protein-protein interactions

In many cases, artificial ligands that bind to a specific protein surface bear some resemblance to an endogenous ligand that also binds to that surface: analysis of precedented drug targets shows that knowledge of a substrate, product, or effector with "drug-like" physicochemical properties is a good predictor of druggability for a given protein surface (153). Analogs of endogenous ligands can also provide a starting point for designing inhibitors (154). In the case of protein interaction sites, however, the natural ligand is not a small molecule and thus does not provide an obvious template from which to start. This has driven advances in new methodologies for identifying inhibitors of these "non-traditional" targets, most notably including mimicry of key interacting groups of the protein partner (117-119, 155-157) and fragment-based approaches (113, 158-160).

In addition to chemical scaffolds that mimic secondary structural elements (117-119), exciting new computational approaches have facilitated identification of compounds that instead mimic the geometric orientation of key "hotspot" or "anchor" sidechains in a protein-protein interface (155-157). While either mimicry-based approach may provide a robust starting point for recapitulating the

interactions of a protein partner using a small-molecule scaffold, both are likely to encounter the same intrinsic limitation: ligand efficiency is unlikely to exceed that of the interacting groups from the protein partner, which in turn is usually less than that those achieved by small-molecule inhibitors that do not explicitly mimic interactions of the protein partner (4). Accordingly, an important advantage of *de novo* structure-based screening methods that do not rely on mimicry, such as DARC, is the potential to identify inhibitors that achieve superior ligand efficiency through deeply buried interactions not available to a protein-based ligand.

In contrast, fragment-based approaches often prioritize small compounds with very high ligand efficiency from the outset: this allows exploration of deep crevices that would not necessarily be evident from the structure of the protein-protein complex. Here, the challenge often lies in elaborating initial fragment hits (by growing, merging, or linking them) into larger compounds that maintain these highly productive interactions (161, 162). In our DARC screen against Mcl-1, we focused on compounds larger than traditional fragments: this led to initial hits with easily detectable activity, albeit at some expense of ligand efficiency. Much as the "multiple solvent crystal structures" method (163) uses very small fragments to probe the protein surface for productive interactions with isolated functional groups, screening a commercially-available library of "prototype" compounds – as we have done here – allows rapid evaluation of specific protein-ligand interactions predicted *in silico*, and can provide potential chemical scaffolds for further optimization.

Screening against multiple protein conformations

The compounds identified as "hits" by DARC, and by structure-based screening methods in general, naturally depend on the conformation of the protein target. With respect to inhibitors of protein-protein interactions in particular, a number of examples have been shown to bind concave surface pockets that are absent in the corresponding unbound protein structures: these "cryptic" pockets are revealed by local conformational changes associated with inhibitor binding (4). In the case of Bcl-2 family members,

the protein surface can also adopt a number of different conformations to accommodate ligands with radically different shapes and chemical properties (124).

Variation in the protein structure is most commonly included in virtual screening by first generating an ensemble of relevant conformations from crystal structures (92-96) or simulation (97-99), and then separately screening against each of these conformations. Nonetheless, identifying the optimal conformations to include in these ensembles is still a challenging task (164-166). By preferentially exploring conformations that contain a surface pocket suitable for binding some (unspecified) ligand (75), the "pocket optimization" approach provides a natural complement to DARC screening. In principle, each of these low-energy pocket-containing conformations represents a starting point for screening, and capturing the potential diversity of pockets on the protein surface will be essential for fully realizing the available diversity of potential inhibitors.

In our Mcl-1 screen, DARC identified inhibitors both when screening against the peptide-bound conformation and when screening against the pocket-optimized conformation. Due to the conformations selected for these screens, however, our models did not take advantage of deep pockets that have been observed in recent inhibitor-bound crystal structures of Mcl-1 (112-114, 130) (**Figure 5c**): the particular pocket-optimized conformation used in our screen was much more similar to the peptide-bound conformation, making these deep pockets unavailable to DARC. Careful analysis showed that similar conformations to those observed in the inhibitor-bound crystal structures were indeed represented in the ensemble produced by pocket optimization (124), but were not included here because we only screened against a single pocket-optimized conformation. In future, then, inclusion of additional pocket-containing protein conformations may allow identification of inhibitors that bind more deeply, and exhibit improved ligand efficiency.

Advantages to structure-based virtual screening

Enhanced and reliable tools for structure-based virtual screening will enable development of new tool compounds, particularly in academic settings where the costs associated with large-scale biochemical (or phenotypic) screening can be prohibitive. However, the ability to precisely complement a specific surface pocket also makes structure-based virtual screening particularly attractive in a variety of other contexts that may prove challenging for traditional biochemical screening. These include building "conformation-selective" inhibitors (such as compounds that are sensitive to the phosphorylation state of a kinase activation loop (167)), targeting sites that enable unique binding kinetics (such as a newly-discovered pocket on ERK1/2 (168)), and building allosteric inhibitors that address both the wild-type and drug-resistant isoforms of a target (such as BCR-ABL (169) or HIV-1 protease (170)).

In light of the plasticity of many protein interaction sites (4), these tools may also prove particularly useful for designing inhibitors that exhibit specific selectivity profiles. We have shown that the activity of a given compound against various members of the Bcl-2 family can be predicted from whether or not each protein samples a complementary surface pocket (124). In other words, a given compound will inhibit those Bcl-2 family members that include conformations to suitably accommodate this compound, but will not inhibit other family members that cannot adopt such conformations. By identifying "common" pockets harbored on the surfaces of multiple family members, it may be possible to identify inhibitors designed to act against multiple family members (pan-inhibitors). "Pocket optimization" simulations also reveal highly unique pockets for each family member, that are sampled by one family member but are not accessible to any other family member (124). Thus, DARC's ability to identify ligands that precisely address a given family member's unique "signature" pockets may also provide a means to identify highly selective compounds at a very early stage of development.

2.6 Methods

Detailed descriptions of computational and experimental methods are provided in the *Supplementary Methods* section, along with sample command-lines and instructions for using DARC.

2.7 Acknowledgements

We thank Amy Keating for providing the plasmid encoding Mcl-1, and we thank Anne Cooper and Philip Gao for assistance with protein production. We are grateful to OpenEye Scientific Software (Santa Fe, NM) for providing an academic license for the use of OMEGA and QuacPac. We are grateful to the developers of the DOCK, AutoDock, rDock, and PLANTS software for providing academic licenses for the use of these programs. This work was supported by grants from the National Institute of General Medical Sciences (1R01GM099959, 5P50GM069663, 8P30GM103495, and 8P20GM103420), and the National Center for Research Resources (5P30RR030926 and 5P20RR017708). This work was also supported by the National Science Foundation through XSEDE allocation MCB130049, the University of Kansas Undergraduate Research Awards (S.M.), and the Alfred P. Sloan Fellowship (J.K.).

CHAPTER 3

DARC 2.0: Improved docking and virtual screening at protein interaction sites

Ragul Gowthaman[1], Sergey Lyskov[2], and John Karanicolas[1,3]

[1] Center for Computational Biology, [3] Department of Molecular Biosciences,

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

[2] Department of Chemical and Biomolecular Engineering,

Johns Hopkins University, 3400 North Charles St., Baltimore, MD 21218

3.1 Abstract

Over the past decade, protein-protein interactions have emerged as attractive but challenging targets for therapeutic intervention using small molecules. Due to the relatively flat surfaces that typify protein interaction sites, modern virtual screening tools developed for optimal performance against "traditional" protein targets perform less well when applied instead at protein interaction sites. Previously, we described a docking method specifically catered to the shallow binding modes characteristic of small-molecule inhibitors of protein interaction sites. This method, called DARC (*D*ocking *A*pproach using *R*ay *C*asting), operates by comparing the topography of the protein surface when "viewed" from a vantage point inside the protein against the topography of a bound ligand when "viewed" from the same vantage point. Here, we present five key enhancements to DARC. First, we use multiple vantage points to more accurately determine protein-ligand surface complementarity. Second, we describe a new scheme for rapidly determining optimal weights in the DARC scoring function. Third, we incorporate sampling of ligand conformers "on-the-fly" during docking. Fourth, we move beyond simple shape complementarity and introduce a term in the scoring function to capture electrostatic complementarity. Finally, we adjust the control flow in our GPU implementation of DARC to achieve greater speedup of these calculations. At each step of this study, we evaluate the performance of DARC in a "pose recapitulation" experiment: predicting the binding mode of 25 inhibitors each solved in complex with its distinct target protein (a protein interaction site). Collectively, we find that the five enhancements described here – which together make up DARC 2.0 – lead to dramatically improved speed and performance relative to the original DARC method.

3.2 Introduction

Protein-protein interactions underlie most biological processes (171-173), and as such many of the individual proteins and networks involved in these interactions are implicated in assorted human diseases (174-177). Modulating key protein interactions using small molecules can provide exciting opportunities to develop novel therapeutics, leading to extreme interest in this target class for drug discovery (4, 178-181). Whereas almost all drugs currently in the clinic inhibit one of several "traditional" target classes (G protein-coupled receptors, enzymes, nuclear receptors, transporters, and ion channels) (1, 11), protein-protein interactions now stand among a broad new emerging class of "non-traditional" targets (182, 183).

Unlike enzymes and other traditional drug targets, protein surfaces evolved to bind other proteins typically lack the deep pockets used as small-molecule binding sites (2, 184, 185). Surveys of small-molecule inhibitors of protein interactions have revealed that these compounds tend be larger and more hydrophobic than traditional drug-like molecules, and reside in regions of chemical space that are less represented in commercial libraries (78). Analysis of crystal structures of small-molecule inhibitors bound at protein interaction sites also reveals that they tend to use shallower binding modes, leading to worse ligand efficiency (binding energy per heavyatom) than their counterparts engaging "traditional" targets (71). An ancillary pathology of these shallow binding modes is that sterics provide fewer clues for correctly docking candidate inhibitors in virtual screens, and accordingly modern virtual screening tools – tools that have been optimized over many years for their performance against "traditional" targets – do not fare as well when asked to identify compounds active against protein interaction sites (71).

To address this, we recently developed an alternative screening approach called DARC (*D*ocking *A*pproach using *R*ay-*C*asting) (186), a docking method specifically for addressing non-traditional targets such as protein interaction sites. The DARC approach is summarized schematically in **Figure 1**. We begin by defining the binding pocket around a given "target" residue (or set of "target" residues) on the

protein surface, using a pocket-finding algorithm adapted from the LIGSITE program (76) that we have implemented in the Rosetta software suite (75). To identify a ligand that complements this pocket, we begin by mapping the topography of the pocket by selecting an "origin" point within the protein interior, and casting rays from this origin at each of the pocket points that are in contact with the protein surface. Operationally, this step is equivalent to simply converting each of the pocket points into a spherical coordinate system (ρ,θ,φ) relative to this origin point, where ρ is the distance from the origin point and θ/φ are the polar/azimuthal angles. We additionally include in this step a layer of points outside the pocket, to help define the pocket's boundaries.

If a ligand docked into this surface pocket is indeed shape-complementary, the surface of this ligand when "viewed" from this vantage point (the origin) should have a very similar topography to that of the pocket. To map the topography of the ligand, we cast a collection of rays from the origin point, with each of the angles (θ,φ) used to map the pocket topography. For each ray, we determine the distance of its first intersection with the ligand (if indeed the ray intersects the ligand). We express the difference in the "observed" surface topographies of the pocket and ligand as follows:

$$Shape\ score = \sum_{rays} \begin{Bmatrix} c1*\left(\rho_{ligand} - \rho_{pocket}\right) & if\ \rho_{pocket} < \rho_{ligand} \\ c2*\left(\rho_{pocket} - \rho_{ligand}\right) & if\ \rho_{ligand} < \rho_{pocket} \\ c3 & if\ ray\ does\ not\ intersect\ ligand \\ c4 & if\ ray\ does\ not\ intersect\ pocket \end{Bmatrix} \quad (1)$$

Conceptually, each of the four conditions above represents a distinct type of deviation between the two surfaces. In the first case (c1), a ray hits the surface pocket before reaching the ligand: this indicates underpacking in the protein-ligand interface. In the second case (c2), a ray intersects the ligand before reaching the pocket: this indicates that the ligand's volume overlaps that of the protein, and thus points to a steric clash. In the third case (c3), a ray that intersects the pocket does not intersect the ligand

at all: this indicates that the ligand is too small for the pocket. Finally, in the fourth case (c4), a ray that intersects the ligand does not intersect the pocket: this indicates that the ligand extends outside the binding pocket.

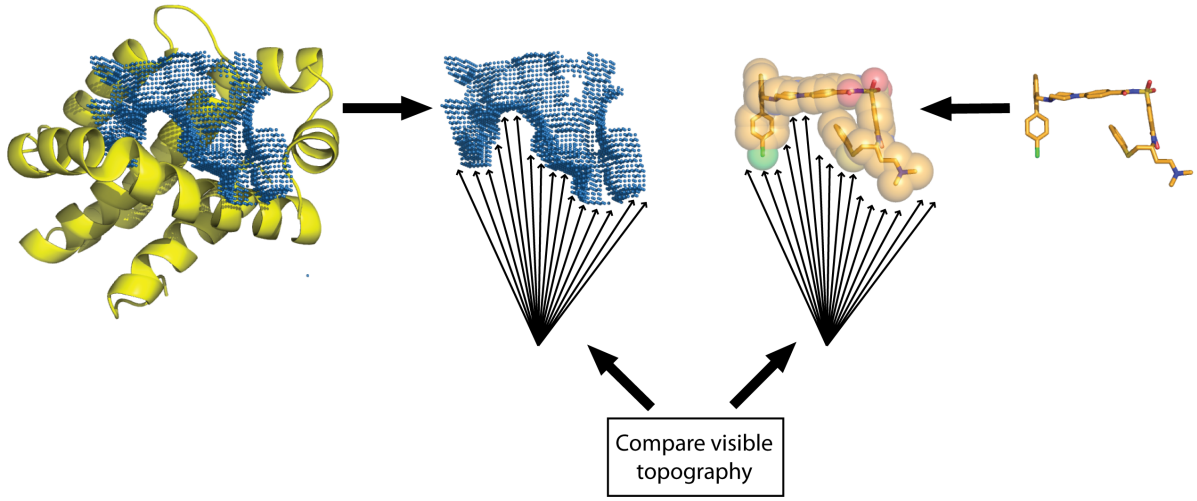


Figure 1: Schematic illustration of the DARC approach. We begin by using a geometry-based method to define the target pocket on the protein surface. Next, we define a series of rays that emanate from a vantage point inside the protein. Collectively, the distance at which these rays reach the protein surface describes the topography of the protein surface in this region, as "seen" from this vantage point. If a bound ligand is complementary to the protein surface, its surface topography will "look" similar to that of the protein surface when viewed from this vantage point; in other words, each ray will intersect the ligand at similar distance as its intersection with the protein surface.

Beyond simply evaluating the shape complementary of a protein-ligand complex, this "shape score" can also be used as an objective function for optimization. By adjusting the ligand position and orientation to minimize this score, one can use DARC to rapidly dock a ligand into a surface pocket. Extending the approach further, by sequentially docking a large number of compounds in this manner one

can use DARC to carry out a virtual screen to identify compounds that complement the shape of some surface pocket on a protein of interest.

Previously, we found that DARC could be used in virtual screening benchmarks to pick out known small-molecule inhibitors of Bcl-xL and XIAP that were hidden amongst pools of "decoy" compounds. Despite its relatively simple implementation, DARC outperformed other popular virtual screening tools at this task – tools that have been developed in the context of "traditional" target classes, rather than for inhibitors of protein interactions. We then used DARC to carry out a computational screen of 65,000 compounds to identify those that would best complement a pocket on the surface of Mcl-1, an anti-apoptotic member of the Bcl-2 protein family. We tested the top 21 DARC hits in biochemical assays, and found that indeed 10 of these are inhibitors of Mcl-1, with Ki values ranging from 2 to 34 μM for the best 6 compounds. Collectively these results validated DARC for virtual screening at protein interaction sites, and demonstrated its usefulness for identifying new inhibitors acting at these sites (186).

We have drawn upon experiences and observations from our early applications of DARC to enhance its performance through the five key ways we describe in this study: ***(1)*** We refined the ray-casting approach, such that rays emanate from multiple origin points to better map the shapes of the surface pocket and the ligand. ***(2)*** We developed a new strategy for parameterizing DARC in a faster and more robust way, thus enabling a broader and more representative set of protein complexes to be included in training. ***(3)*** We introduced a new scheme to efficiently sample small-molecule conformers "on-the-fly" in a single docking trajectory, rather than sequentially consider each conformer in a separate trajectory. ***(4)*** We incorporated electrostatics into the DARC scoring function, allowing simultaneous optimization of both shape- and electrostatic-complementarity. ***(5)*** We identified the computational performance bottleneck in our previous implementation of GPU-DARC (77), and adjusted the control flow by transferring additional calculations onto the GPU to resolve this bottleneck.

As described below, these individual enhancements collectively lead to dramatic improvements in DARC's robustness, accuracy, and speed.

3.3 Results

Previously, we assembled a set of unique proteins for which a crystal structure was available in complex with a small-molecule inhibitor of a protein-protein interaction site (71). At the time, there were 21 such structures available; since then, 4 additional examples have become available. For the studies we describe here, we make use of this new set of 25 non-redundant complexes in which a small-molecule inhibitor is bound at a protein interaction site (**Table S1**).

Enhancement #1: ray-casting using multiple origins

As described above, the topography of the protein surface is mapped from a vantage point inside the protein, using an "origin" point from which rays emanate. The placement of this origin point is critical to ensuring that the resulting topography map gives an accurate and complete view of the surface pocket – in an intuitive sense, it should be centered "behind" the pocket. For an ideal scenario, in which the surface pocket is a purely concave "dimple" on the surface of a near-spherical protein, the protein's center of mass can serve as a natural choice for the origin's location. In practice, however, the pocket shapes are never purely concave, and any ruggedness means that some parts of the protein surface cannot be "seen" from a given vantage point. An incomplete description of the protein surface, in turn, limits the ability of DARC to identify truly complementary ligands.

In order to map the protein surface topography more accurately, we therefore modified the ray casting approach such that rays emanate from multiple origins: we expected this strategy would better "illuminate" all regions of the protein surface. To do so, we begin from a single origin centered 30 Å behind the pocket (see *Methods*). Rotating about a point fixed at the pocket center, we move the origin point by $\pm 45^{o}$ in each of two orthogonal directions, to generate four new origin points. In other words, if the z-axis connects the first origin to the pocket center, we rotate in first the xz- and then the yz-plane, by $\pm 45^{o}$ each time, to place these four additional origins.

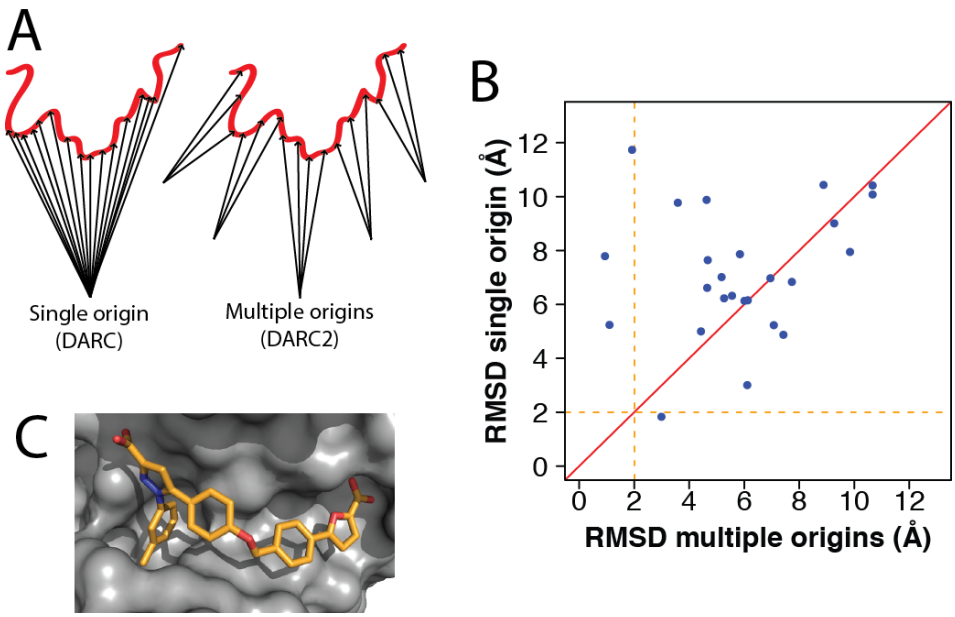


Figure 2: Ray-casting from multiple origins. (A) In DARC 2.0, rays emanate from five origin points instead of from a single origin point. This allows more of the protein surface to be "visible" to the rays, and in turn leads to a more complete representation of the surface topography. **(B)** For each protein-ligand complex in our set (**Table S1**), we used DARC to dock the ligand back into its cognate receptor. Multiple ligand conformations were considered in the search, while the receptor conformation was held fixed throughout docking. The RMSD of the docked ligand was evaluated relative to its position in the crystal structure of the complex. Each point represents a separate complex; points above the diagonal are those for which the use of multiple origins led to better pose recapitulation. **(C)** A representative example of a complex that was better-predicted using multiple origins (PDB ID 4luz). Part of the binding pocket points directly into the protein (*left side of picture*), and thus directly at the origin (if a single origin is used). The use of multiple origins allows the "walls" of this region of the binding pocket to be included in the surface topography considered for docking.

As before, the topography of the protein surface is mapped by casting rays at each of the "pocket" gridpoints that directly contact the protein. Rather than retain the distances at which rays from all five origins hits a given protein surface point, we only store the distance from the closest origin: each origin is thus "responsible" only for those regions of the protein surface pocket that are closest to it, and for which that origin is therefore likely to have an optimal vantage point (**Figure 2a**).

Having mapped the protein surface pocket using rays cast from multiple origins, we then evaluate ligand complementarity exactly as described earlier: we simply cast the rays that emanate from each origin towards the ligand, and compare the distances at which these intersect with the ligand to the distances at which they intersected the protein surface (**Equation 1**). This strategy for using multiple origins does not increase the total number of rays included in DARC's calculations, but simply distributes the same number of rays among the different origins; accordingly, this new approach does not change the time required for docking using DARC.

To examine the effect of using multiple origins, we used DARC to dock each of the 25 protein-ligand pairs in our test set (**Table S1**). In each case we used OMEGA (44, 45, 187) to pre-build a set of allowed conformations ("conformers") for each ligand, and included these in docking; the conformation of the protein was held fixed throughout each simulation. For each of these 25 protein-ligand cognate pairs, we then evaluated the RMSD of the ligand position in the docked complex relative to its position in the corresponding crystal structure.

We also carried out this "pose recapitulation" experiment using the previously described version of DARC (186), which only employed a single origin, and compared the results to those obtained using this new "multiple origins" approach (**Figure 2b**). We find that the RMSD relative to the crystal structure is better in 16 of the 25 cases (points above the diagonal), suggesting that the use of multiple origins may indeed enable more accurate matching of protein/ligand shape complementarity. Because the 25 complexes in our set represent paired samples (and are not expected to be normally distributed), we employed the (non-parametric and paired) Wilcoxon signed-rank test (see *Methods*) to compare the

results from single versus multiple origins. While this test does identify the difference in performance, it does not quite achieve statistical significance ($p < 0.054$) due to the relatively small size of the test set (which, in turn, stems from the fact that few examples of crystal structures of small-molecule inhibitors of protein-protein interactions solved in complex with their target proteins are available).

We note that before inclusion of this new feature in DARC, only one complex in our set was docked to within 2 Å RMSD of the crystal structure; using multiple origins, there are now three such cases. Examination of the crystallographic complexes for these improved cases reveals that many share a binding mode in which the ligand (or part of the ligand) faces directly into the protein (**Figure 2c**). In such cases, rays cast from a single origin would have described only the very bottom of this well; in contrast, the use of multiple origins allow the topography of the walls of this well to also be included, and thus allow the ligand to be more accurately matched to the contours of the protein surface.

Enhancement #2: fast and robust weight fitting

There are four parameters (c1/c2/c3/c4) used by DARC when evaluating shape complementarity (**Equation 1**). Since scaling all four of these by a constant would simply scale the total score, we fix c1=1.0 and determine values of the other three parameters accordingly.

In our original parameterization of DARC (186), we used a small training set of seven protein-ligand complexes to optimize these weights. We sought to identify the combination of weights that would optimally allow each of these seven ligands to be docked back into its cognate receptor such that they would match the crystal structures of these complexes. For a given set of weights, then, we assessed performance by using DARC to dock these seven pairs, and used the sum of the resulting seven ligand RMSDs as our objective function. We used simplex optimization to drive our search of parameter space, but each evaluation of this objective function required seven separate calls to Rosetta to carry out the required docking. As a result, carrying out this weight fitting procedure typically required about a week of computation on a modern CPU. Since we originally carried out this parameterization, additional crystal

structures of small-molecule inhibitors of protein interactions in complex with their targets have become available; however, our parameterization scheme was already too slow to feasibly add these new examples.

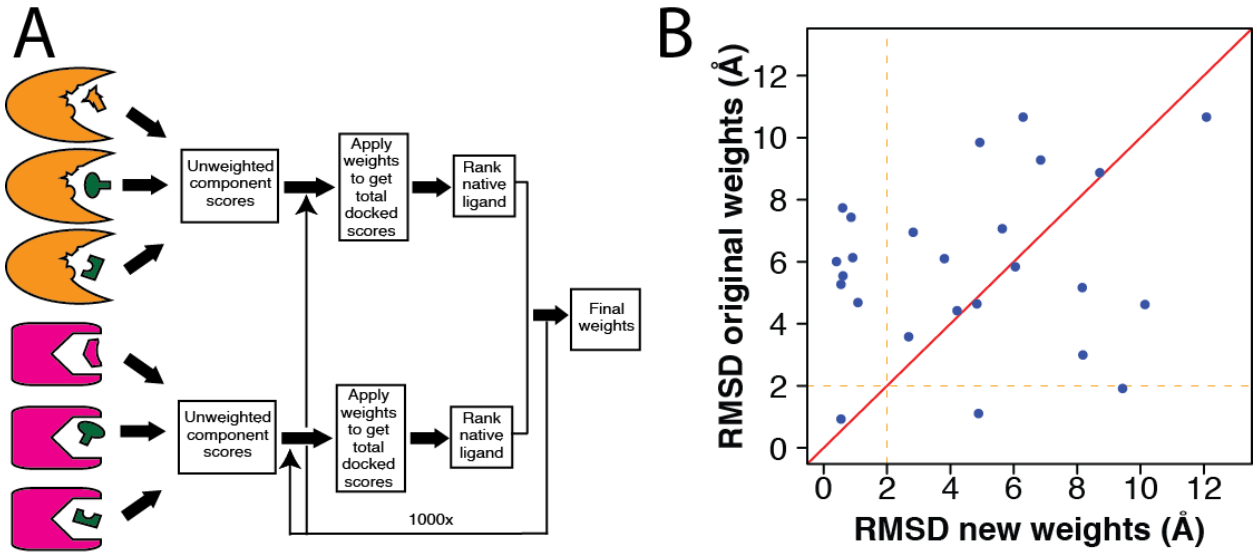


Figure 3: Updated weight-fitting strategy. (A) Schematic illustrating the strategy used in DARC 2.0 for determining values of the weights that should be applied to the terms in the DARC scoring function (**Equation 1**). For each complex used in training (**Table S1**), we generate a large set of docked complexes involving either the cognate "native" ligand (*orange/magenta*) or one of many "decoy" ligands (*green*). Using the component energies in these complexes, we identify the set of weights that optimally ranks the native ligands ahead of the "decoys". **(B)** For each protein-ligand complex in our set, we used DARC to dock the ligand back into its cognate receptor using either the original DARC weights or the newer weights derived though the approach described here. In the latter case, leave-one-out cross-validation was used to ensure the weights were not overfit to the training set. Each point represents a separate complex; points above the diagonal are those for which the new weights led to better pose recapitulation.

To speed up weight fitting, we adapted our approach such that explicit docking at every step would no longer be required. We also took this opportunity to parameterize DARC not for its ability to simply recapitulate the bound pose of a known ligand, but rather for its ability to optimally distinguish a known ligand from among a pool of "decoy" compounds. The latter represents a virtual screening scenario, such that the resulting weights may exhibit improved performance for this task.

We started by first reformulating DARC's shape complementarity score as follows:

$$Shape\ score = c1*\left(\sum_{\substack{rays\ where \\ \rho_{pocket}<\rho_{ligand}}} \rho_{ligand} - \rho_{pocket}\right) + c2*\left(\sum_{\substack{rays\ where \\ \rho_{ligand}<\rho_{pocket}}} \rho_{pocket} - \rho_{ligand}\right) + c3*\begin{pmatrix}\#\ \text{rays} \\ \text{that do not} \\ \text{intersect ligand}\end{pmatrix} + c4*\begin{pmatrix}\#\ \text{rays} \\ \text{that do not} \\ \text{intersect pocket}\end{pmatrix} \quad (2)$$

Relative to our previous formulation (**Equation 1**), we have simply gathered together the groups of rays that meet each one of the four conditions. This is a natural reformulation of this equation, since the four weights each apply to one of these four conditions.

We again turned to our non-redundant set of 25 complexes in which a small-molecule inhibitor is bound at a protein interaction site (**Table S1**); to setup our optimization as a virtual screening problem, we built a "compound library" of 650 diverse decoy ligands, and generated 1000 randomly docked poses for each compound with each protein in this set.

The key to the reformulation of DARC's shape complementarity score above (**Equation 2**) is that the result of ray-casting can be separated from the weights: for a given pose we can pre-compute (and store) each of the summations over the rays that meet each of the four conditions. Given some new set of weights, we can then apply these four weights to the four stored numbers and trivially update the score of the pose with these new weights.

Collectively, we generated more than 16 million "decoy poses" (25 proteins x 650 ligands x 1000 randomly docked poses for each protein-ligand combination). Rather than store these decoy poses explicitly, we instead simply stored the four "component energies" (the unweighted terms in **Equation 2**) for each decoy pose. Similarly, we pre-computed and stored the four "component energies" for each of the 25 native poses.

Using this data as input, we then carried out weight fitting as shown in **Figure 3a**. We used simplex optimization (as implemented in the GSL multidimensional minimization library (188)) to search parameter space. At every step, we updated the DARC score for each (decoy and native) complex using the new weights; for each protein target, we then determined the *rank* of the native pose relative to each of the 650,000 decoy poses involving this protein. As the objective function for this minimization, we used the sum of the ranks of the 25 native poses.

This approach to fitting weights required far fewer computational resources than our previous approach; this optimization (using pre-computed component scores) was typically completed in minutes (it took under a minute to run 100 iterations). To objectively examine the performance that could be expected from the weights obtained by this method, we used leave-one-out cross-validation. For each of the 25 proteins in our test set, we developed a unique set of weights trained only on the other 24 proteins; we then used this custom weight set to dock the native ligand of interest back into its cognate receptor, and evaluated the ligand RMSD relative to the crystal structure of this protein-ligand complex.

We compared the RMSD of these 25 DARC-docked examples, using our previous DARC weights (186) or using the leave-one-out cross-validated weights from our new approach. We note that results from the earlier weight-fitting approach were not subjected to cross-validation, due to the large computational requirements that would be associated with building numerous "custom" weight sets. We further note that the newer weights are not explicitly trained for docking native complexes, but rather for discrimination in virtual screening tasks. Nonetheless, we find that the newer weight set proves far superior to the original weights (**Figure 3b**): the ligand RMSD is lower for 17 of the 25 complexes when using the newer weights (points above the diagonal). Further, the newer weights perform exceptionally well in a number of testcases: there are now 8 examples for which the RMSD was less than 2 Å using the new weights, whereas this level of accuracy was only achieved in 3 cases using the older weights. Applying the Wilcoxon signed-rank test to compare the differences in RMSD associated with changing

the weights (see *Methods*) also detects this difference in performance, albeit not quite at a level achieving statistical significance ($p < 0.051$).

We attribute this improvement to the robustness of the newer weight set, which derives from training on a larger and more broadly representative set of examples; the previous weight set may have been over-fitted to the seven examples in the training set upon which it was based. This is a particularly encouraging outlook in light of the intended use for these weights: they may be far-better suited for virtual screening than the previous weights, since robust tools for this task will require the ability to rapidly and accurately evaluate many diverse ligands (and ideally should prove applicable for diverse protein targets as well).

Enhancement #3: sampling ligand conformers "on-the-fly"

Efficiently sampling the potential positions, orientations, and conformations of each compound is critical to virtual screening. As the size of "purchasable" chemical space continues to increase (189), and these compounds continue to be a useful for populating *in silico* libraries, the speed of virtual screening is likely to be of paramount importance for the foreseeable future. It is important to note that the speed of modern docking approaches generally scale not only with the size of the screening library, but also with the number of conformations considered for an average ligand in the library.

Modern docking / virtual screening tools address the problem of ligand conformational sampling in different ways. Some programs, such as FRED, pre-generate a collection of low-energy ligand conformations ("conformers"), then sample each of these individually in separate docking trajectories (190). Other programs, such as AutoDock, generate ligand conformers and evaluate their energy *in situ* during docking (85). In the DOCK6 program, ligand conformations can either be generated *in situ* during docking with "anchor-and-grow" incremental construction, or alternatively a set of rigid conformers can be pre-generated and screened sequentially (38, 83). In the original implementation of DARC, no allowance was made for ligand flexibility; alternate ligand conformations were considered by sequentially

docking pre-generated conformers, and the best-scoring member of the resulting set was taken to be the final predicted pose (186).

DARC makes use of particle swarm optimization (PSO) (191) to minimize its objective function by varying the ligand's position and orientation. PSO is a population-based optimization method that mimics swarm intelligence and applies a heuristic approach to find an optimal solution (192). Others have also used variants of PSO as a fast and efficient optimization method for protein-ligand docking (193-195). In the case of DARC, we set up the optimization problem such that the displacement and rotation relative to some "reference" position of the ligand are the six degrees of freedom included in the search.

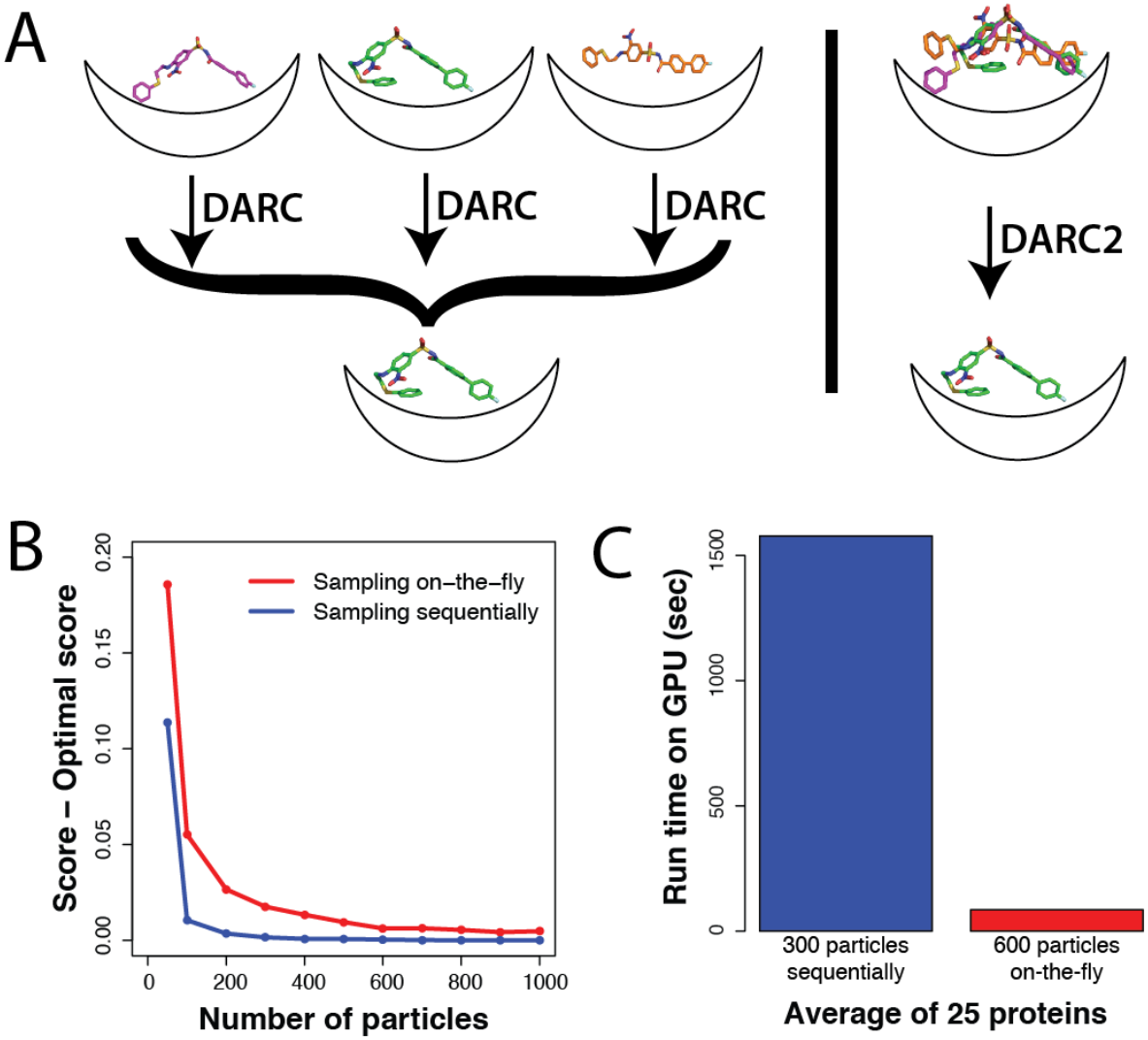


Figure 4: Screening ligand conformers "on-the-fly" during docking. (A) Previously, ligand conformations were docked sequentially through separate docking trajectories, and the ligand conformation was ultimately drawn from the best-scoring complex. In DARC 2.0, we instead sample ligand conformers *during* the docking trajectory. **(B)** Due to the extra degree of freedom associated with a single docking trajectory, docking converges more slowly when conformers are sampled "on-the-fly". Here, convergence is evaluated by the score difference relative to a "gold standard" (best achievable score) for each complex; the results shown are averaged over the 25 complexes in our test set (**Table S1**). For each point in this plot, the number of particles in PSO optimization and the number of steps in the docking trajectory were set equal to one another. **(C)** Despite the fact that individual trajectories converge more slowly when conformers are sampled "on-the-fly", docking is comprised of only a single trajectory. Across our set of 25 complexes, this ultimately makes "on-the-fly" sampling an average of 18-fold faster than sampling conformers sequentially through multiple docking trajectories.

To increase the efficiency of our sampling, we adapted our approach such that the set of (pre-generated) conformers would instead be considered on-the-fly during docking (**Figure 4a**). We reasoned that not every conformer deserved an equal amount of sampling; by focusing more of our sampling on the top-scoring conformers, the overall time needed for docking a given compound could be reduced. To achieve this, we introduced a seventh degree of freedom in our search: the "conformer index."

At the start of the simulation, we assigned each of the pre-generated conformers a unique index. During the PSO, seven parameters would be included in the optimization: one to indicate which ligand conformation should be used, and six to transform the atoms of this ligand to the appropriate position and orientation with respect to the protein. This approach is particularly suited to PSO optimization, which simultaneously maintains multiple solutions ("particles") during a docking trajectory; separate

populations that make use of different "promising" conformers can each explore their own local clusters of solution space, whereas conformers that are not used in any productive poses are sampled less frequently. This, in turn, could reduce the overall time required to run the optimization.

Given that the search space for a given trajectory is now much larger (there is an extra degree of freedom), we anticipated that using "on-the-fly" conformer sampling would lead to slower convergence than a trajectory in which only a single conformer was considered. To test this, we first used "sequential" conformer docking with very intensive sampling (1000 particles and 1000 steps) to identify the optimal score that could be obtained when docking the native ligand back into its cognate receptor, for each of the 25 complexes in our set (**Table S1**). Next, to assess convergence, we asked how closely the scores for each complex would approach these "gold standard" scores as the amount of sampling was reduced by simultaneously lowering the number of particles and steps in the search.

As expected, we indeed find that convergence to near-optimal solutions occurs more slowly with "on-the-fly" sampling instead of sequential sampling (**Figure 4b**). Whereas the optimal solutions are obtained using only 300 particles / 300 steps of sequential sampling, 600 particles / 600 steps were required for convergence when using sampling "on-the-fly". Despite the need for more sampling *per trajectory*, however, the advantage of on-the-fly sampling lies in the fact that *only a single trajectory is needed*. Since we use an average of 163 conformers for the ligands in our test set (**Table S1**), and sequential sampling requires that a separate trajectory be carried out for each conformer, the average runtime for sequential sampling is much longer (**Figure 4c**). Comparing the runtime required for equivalent sampling (300 particles / 300 steps of sequential sampling versus 600 particles / 600 steps of on-the-fly sampling), we find that on average an 18-fold speedup is achieved when on-the-fly conformer sampling is used.

Enhancement #4: inclusion of electrostatic complementarity

Complementarity between a ligand and its binding pocket on the protein surface is the guiding principle in protein-ligand docking, and the success of DARC to date is based on this fundamental principle. In its original inception, DARC was purely based on optimizing and identifying shape complementarity between the surface of the ligand and the surface of the protein (186). In addition to shape complementarity, however, the chemical complementarity of the interacting surfaces is clearly essential for protein-ligand recognition. In addition to the well-established electrostatic complementarity between evolved protein-protein binding partners (196, 197), it has more recently been recognized that small-molecule inhibitors of protein-protein interactions sometimes (inadvertently) mimic the electrostatic patterning of the natural binding partner, in order to optimally complement the charge distribution presented by the surface of the target protein (198).

Of course, other docking methods recognize the importance of electrostatic complementarity, and include its contribution through various approaches. Since most virtual screening tools do not incorporate receptor flexibility during docking, typical modern approaches pre-generate an "electrostatic grid map", and use this to calculate the electrostatic interaction energy given the position of the ligand. Broadly speaking, this is strategy utilized in both AutoDock4 (through AutoGrid) (85, 199) and the DOCK suite (83, 200).

While DARC was originally predicated on matching the surface shapes of the ligand and the protein surface, we quickly noted (by inspection of mis-docked structures) that a number of ligands exhibited pseudo-symmetry when examined purely on the basis of their shapes. In other words, docking without consideration of chemical complementarity very quickly highlighted the limitations of docking on the basis of shape complementarity alone.

To address this, we built into DARC the ability to capture electrostatic complementarity using the most common approach employed by other modern docking tools. Given the (fixed) receptor conformation, we solve the Poisson-Boltzmann equation to calculate the electrostatic potential at a series

of gridpoints that span the surface pocket of interest (**Figure 5a**). For convenience and speed, here we used the finite difference Poisson-Boltzmann solver included in OpenEye's ZAP toolkit (201) for this task (see *Methods*).

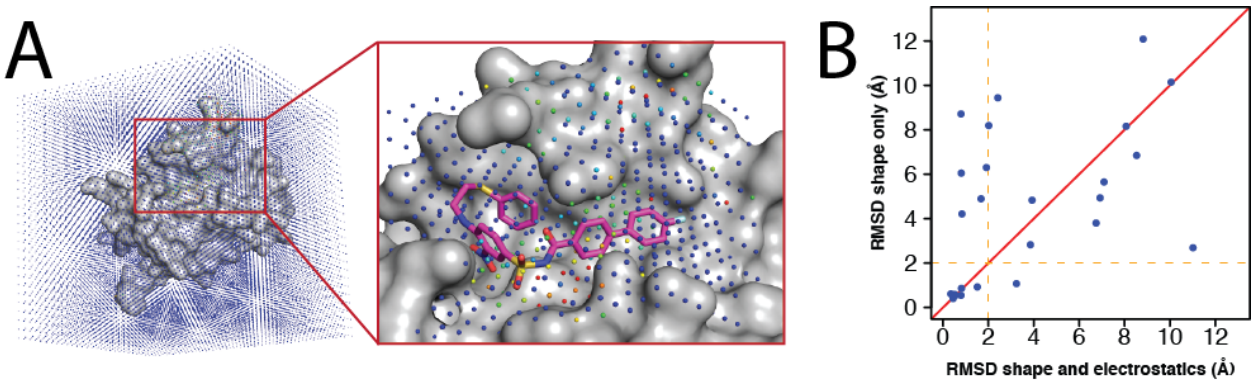


Figure 5: Incorporation of electrostatic complementarity into DARC 2.0. (A) The electrostatic potential is evaluated at a series of grid points over the whole protein using the finite difference Poisson-Boltzmann solver included in OpenEye's ZAP toolkit. We use trilinear interpolation of the closest gridpoints to determine the electrostatic potential at points corresponding to locations of ligand atoms, and then use the ligand partial charges to compute the electrostatic interaction energy (**Equation 3**). **(B)** For each protein-ligand complex in our set (**Table S1**), we used DARC to dock the ligand back into its cognate receptor either with or without including the electrostatic complementarity term. In both cases, leave-one-out cross-validation was used to ensure the weights were not overfit to the training set. Each point represents a separate complex; points above the diagonal are those for which inclusion of electrostatics led to better pose recapitulation.

Given the electrostatic potential, we evaluate the electrostatic complementarity by summing over atomic partial charges in the traditional manner as follows:

$$Electrostatics\ score = \left(\sum_{\substack{ligand \\ atoms}} q_i \phi_i \right) \tag{3}$$

To determine the electrostatic potential (ϕ_i) at the location of a given ligand atom (i), we use trilinear interpolation of the closest gridpoints that encapsulate the center of the atom. Atomic partial charges (q_i) for the ligand were determined using the "molcharge" program from OpenEye (see *Methods*). To ensure the ligand remained within the bounds of the protein surface pocket, we set the electrostatic potential to zero in the protein interior, and applied an unfavorable value of the electrostatic potential outside the defined binding site (+100 DARC units • C^{-1}).

We then used the strategy described above (*Enhancement #2*) to develop a new set of weights for DARC, this time including electrostatic complementarity as follows:

$$DARC\ score = c1 * \left(\sum_{\substack{rays\,where \\ \rho_{pocket} < \rho_{ligand}}} \rho_{ligand} - \rho_{pocket} \right) + c2 * \left(\sum_{\substack{rays\,where \\ \rho_{ligand} < \rho_{pocket}}} \rho_{pocket} - \rho_{ligand} \right) + c3 * \left(\begin{array}{c} \#\ \text{rays} \\ \text{that do not} \\ \text{intersect ligand} \end{array} \right) + c4 * \left(\begin{array}{c} \#\ \text{rays} \\ \text{that do not} \\ \text{intersect pocket} \end{array} \right) + c5 * \left(\sum_{\substack{ligand \\ atoms}} q_i \phi_i \right) \tag{4}$$

Critically, we note that the same reformulation that enabled the weight-fitting strategy described earlier (decoupling the energetic contributions from their weights) applies equally well here; this allowed us to use the same approach to derive a new set of weights that includes this electrostatic term. For the results presented below, we again used the same leave-one-out cross-validation described earlier.

As we had done after each of the previous enhancements, we returned to the 25 complexes in our test set (**Table S1**), and used this latest iteration of DARC to dock multiple conformations of each ligand against its cognate protein partner (**Figure 5b**). Previously, on the basis of shape alone, we found that the

RMSD of the docked ligand relative to the crystal structure was less than 2 Å in 8 cases. Upon inclusion of electrostatics, 7 of these remain "correctly docked" while the RMSD in one case increases above 2 Å. Of the cases that were *not* previously docked to within 2 Å RMSD, however, five new complexes were now "correctly docked" upon inclusion of this electrostatics term (for a total of 12 such cases). Applying the Wilcoxon signed-rank test to compare the differences in RMSD associated with inclusion of electrostatics (see *Methods*), the improvement is detected but again does not reach statistical significance ($p < 0.156$) due to the modest size of our test set.

Enhancement #5: improved implementation for GPU computing

The fact that graphics processing units (GPUs) were originally designed to process multithreaded 3D graphics through ray-tracing makes them extremely well-suited for the ray-casting that underlies DARC. Previously, we adapted DARC such that the ray-casting step would be carried out on the GPU; meanwhile, the central processing unit (CPU) would be responsible for updating the ligand coordinates and repeatedly passing these to the GPU. This GPU implementation proved extremely useful, because it led to a speedup of about 27-fold in typical-use cases, as compared to the time required to carry out the analogous calculations using the CPU alone (77).

Upon more recent examination of the speedup observed when carrying out various calculations in DARC, we found that the size of the ligand and the number of particles both contributed to the bottleneck in the speedup that could be achieved. As noted earlier, DARC uses particle swarm minimization to optimize the ligand's displacement and rotation (and now the "conformer index" as well, for on-the-fly sampling) relative to a saved "reference" position. While the ray-casting step was taking place on the GPU, applying the transformation to translate and rotate the ligand to its new coordinates was carried out on the CPU, and was required for every particle (at every step of the docking trajectory). Our observations of the scaling with respect to ligand size and number of particles led us to hypothesize that the performance bottleneck in the GPU-enabled calculation was either due to the time required for the

CPU to apply the appropriate transformation to every ligand atom of every particle, or because of the amount of data transferred from the CPU to the GPU.

To address this bottleneck, we devised a new scheme for splitting control flow between the CPU and the GPU (**Figure 6a**). During setup, our new approach stores the "reference" position of each ligand conformer on the GPU. At each step of a docking trajectory, we previously passed from the CPU to the GPU a message obtained by "unpacking" the information in each particle (the coordinates of each ligand atom for that particle); now, we instead pass only the seven numbers stored in each particle: the conformer index (1 number), the displacement that must be applied to the ligand's reference conformation (3 numbers), and the rotation that must be applied to the ligand's reference conformation (3 numbers). In addition to reducing the amount of information transferred, this also allows the transformations of the ligand coordinates to be carried out on the GPU in a highly parallel fashion (instead of carrying out this step sequentially on the CPU).

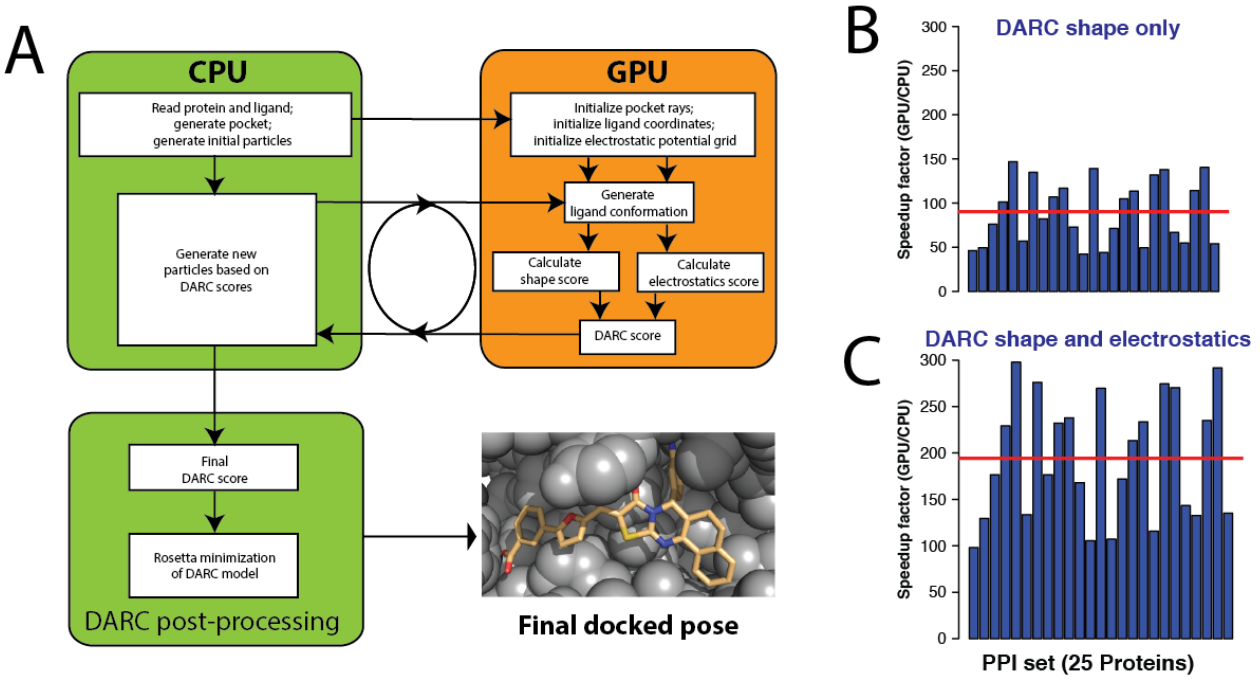


Figure 6: Updated GPU control flow. (A) Schematic illustration of CPU-GPU control flow in DARC 2.0. Previously, the ligand conformation was generated on the CPU and passed to the GPU; now, the conformer index / displacement / rotation (relative to a "reference" position) is instead passed, and the GPU is responsible for applying this transformation to the ligand's atomic coordinates. The new electrostatic complementarity term is also computed entirely on the GPU. **(B)** For each protein-ligand complex in our set (**Table S1**), we timed DARC when docking the ligand back into its cognate receptor, using either GPU+CPU or CPU alone. We find an average speedup of 90-fold when using the GPU (red line), an improvement over the 27-fold speedup we achieved in our original GPU implementation of DARC. **(C)** The GPU led to an even greater speedup over the analogous calculation on the CPU when electrostatic complementarity was included in both calculations (190-fold speedup).

To evaluate the speedup achieved by this new strategy we determined the time needed for docking each of the 25 complexes in our test set (**Table S1**), either using a CPU alone or using GPU-enabled DARC (**Figure 6b**). Unsurprisingly, we find that the ratio of the runtimes (the "speedup factor") differs for the complexes in our set: the size and shapes of the pockets differ (causing the number of rays to differ), and the ligand sizes differ. Nonetheless, on average we observe a 90-fold speedup when running on the GPU – about three times faster than our original GPU implementation. This result confirms our identification of the previous performance bottleneck, which has been successfully overcome through this new CPU/GPU control scheme.

Our scheme also proved naturally amenable for using the GPU to calculate the electrostatic part of the DARC score as well (**Figure 6a**): at setup, we simply store the electrostatic potential grid on the GPU, and use the atomic positions of the ligand to compute the electrostatic score as described earlier (**Equation 3**). This part of the calculation also benefits tremendously from GPU parallelization: when electrostatic complementarity is included in the calculation, the average speedup of GPU-enabled DARC (relative to the analogous calculation on CPU alone) reaches 190-fold (**Figure 6c**).

3.4 Discussion

Here, we present a number of enhancements to the robustness, speed and accuracy of DARC; each enhancement builds upon the previous one. These include introduction of multiple origins from which rays can emanate, a new scheme for rapidly determining optimal weights in the scoring function, the ability to rapidly screen conformers "on-the-fly" during docking, inclusion of electrostatic complementarity in the scoring function, and improved control flow for GPU computing. As a result of the linear narrative by which we have describe these enhancements, however, the overall improvement from this collection of improvements is less apparent. In **Figure 7**, we therefore re-plot the results of our docking experiment such that we compare the results from this latest, fully-enhanced version of DARC – which we call "DARC 2.0" – against the iteration of DARC described in our previous work (186) ("DARC 1.0") that marked the starting point for the current study. Whereas our starting version of DARC docked only *one* of the ligands in our test set to within 2 Å RMSD of its position in the crystal structure, "DARC 2.0" achieves this level of accuracy for *12* of the 25 complexes. The dramatic improvement in the RMSD of these docked complexes is reflected through the Wilcoxon signed-rank test, which confirms a statistically significant performance improvement from these collective enhancements ($p < 0.001$).

The ability to achieve such a dramatic improvement in DARC 2.0 relative to DARC 1.0 is striking in part due to the success of DARC 1.0 for virtual screening. As noted earlier, our initial deployment of DARC in a screen against Mcl-1 allowed us to identify 6 new inhibitors with Ki values better than 35 μM (186); given that we tested 21 compounds suggested by DARC, this corresponded to a success rate of 29% at this potency cutoff. In retrospect, the unimpressive performance of DARC for pose recapitulation did not foretell poor performance in this seemingly more challenging arena, because in fact virtual screening is – in many ways – an easier task.

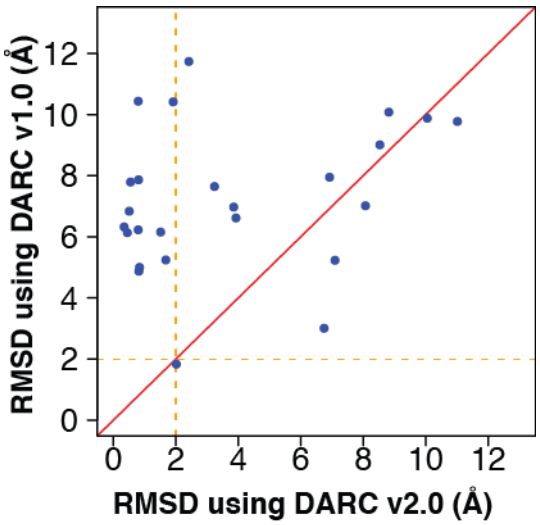


Figure 7: Summary of the collective effect of enhancements to DARC described here. For each protein-ligand complex in our set (**Table S1**), we used DARC to dock the ligand back into its cognate receptor. Here we compare performance of DARC before the enhancements described in this work ("DARC 1.0"), to its current performance ("DARC 2.0"). Each point represents a separate complex; points above the diagonal are those for which the use of multiple origins led to better pose recapitulation. Previously, only one ligand (of 25) was docked to within 2 Å RMSD of its position in the crystal structure; in contrast, "DARC 2.0" now achieves this level of accuracy in 12 cases.

In a virtual screening experiment, two types of errors can limit performance: false negatives (compounds that do not receive a high ranking, though they are in fact active) and false positives (compounds predicted to be active that are not actually active). In practice, as long as *some* hits are identified then a screening campaign is considered successful: missing out on additional active compounds in the library does not detract from this success. In other words, provided that the number of "true" hits in a library is not exceedingly small, false positives limit the perceived performance of virtual screening much more than false negatives. This is deceptive in some ways, however, since these additional hit compounds that were (incorrectly) excluded may have superior potency than the compounds that were ultimately prioritized for further characterization (i.e. in "wetlab" assays).

In a pose recapitulation benchmark, such as the one used in this study, the ligand to be used for each testcase is pre-determined, and there is a single "right answer" (i.e. the bound pose from the crystal structure). This is a far more stringent test than a virtual screen; when screening, failing to correctly dock an individual (active) compound from the library would simply lead to exclusion of this compound from among the hits (a false negative), and would go unnoticed.

In retrospect, DARC 1.0 exhibited impressive performance for virtual screening because a number of active compounds were identified – but many other compounds more potent than those we chose to characterize may have been present in our library. Because of the improved performance demonstrated by DARC 2.0 in pose recapitulation, we anticipate fewer false negatives in screening applications – leading, in turn, to improved potency of initial screening hits from DARC 2.0 relative to DARC 1.0.

3.5 Methods

Implementation in Rosetta

DARC is implemented as part of the Rosetta macromolecular modeling suite (60). Rosetta is freely available for academic use (www.rosettacommons.org), with the new features described here included in official releases 2015.05 and beyond. The Protocol Capture accompanying this manuscript (**Dataset S1**) contains all the commands required for running DARC, including sample input and output files. All results reported here were generated using *git revision 011e012* of the master source code.

Running DARC with Rosetta (no electrostatics)

Running DARC within Rosetta is a two-step process: first generating the ray file, and second docking with DARC.

In the first step we generate the protein surface pocket and map the shape of the pocket shell (points in direct contact with the protein) to a spherical coordinate file; we call this a "ray-file". To generate this ray-file we need to input the protein (in PDB format), and specify one or more target residue(s). The command to run DARC is as follows:

```
Rosetta/main/source/bin/make_ray_files.macosclangrelease -protein
4ERF.pdb
       -central_relax_pdb_num 54,99
```

To use multiple origins, we use:

```
Rosetta/main/source/bin/make_ray_files.macosclangrelease -protein
4ERF.pdb
       -central_relax_pdb_num 54,99 –multiple_origin
```

In the second step, we are actually running the docking calculations using the pre-generated ray-file. Here we give the input ligand(s) for screening against the ray-file, as follows:

```
Rosetta/main/source/bin/DARC.macosclangrelease -protein 4ERF.pdb -ligand 0R3_0001.pdb
       -extra_res_fa 0R3.params -ray_file eggshell_rosetta_4ERF_54,99.txt
```

To search conformers on-the-fly:

```
       Rosetta/main/source/bin/DARC.macosclangrelease -protein 4ERF.pdb -ligand
       0R3_0001.pdb
              -extra_res_fa 0R3.params -ray_file eggshell_rosetta_4ERF_54,99.txt
              –search_conformers true
```

Rather than center the pocket grid at the target residue(s), we can instead center it using a bound ligand (primarily for benchmarking purposes):

```
       Rosetta/main/source/bin/make_ray_files.macosclangrelease -protein
       4ERF.pdb
              -central_relax_pdb_num 54,99 -bound_ligand 0R3_0001.pdb
              -extra_res_fa 0R3.params –lig_grid
```

The output of the DARC run is a docked model of the protein-ligand complex; in this case it would be named “DARC_4ERF_0R3.pdb”

Fullatom minimization in Rosetta

Fullatom minimization of the DARC models can either be carried out separately in Rosetta, or immediately after completion of the DARC. To minimize the DARC models immediately after docking we add the flag “-minimize_output_complex” as follows:

```
       Rosetta/main/source/bin/DARC.macosclangrelease -protein 4ERF.pdb -ligand
       0R3_0001.pdb
              -extra_res_fa 0R3.params -ray_file eggshell_rosetta_4ERF_54,99.txt
```

```
-minimize_output_complex
```

This gives an additional output file named “mini_4ERF_0R3.pdb”.

Other optional flags to use when running DARC include:

```
-origin_cutoff 9 -atom_radius_scale 0.9 -num_particles 100 -num_runs 100
-missing_point_weight 13.3 -steric_weight 3.12 -extra_point_weight 8.13
–esp_weight 0.03 –use_connolly_surface
```

Multiple origin points

Whether using a single origin or multiple origins, we begin by placing the first origin point (O_1) at a distance 30 Å from the center of the pocket, and at a location centered “behind” the pocket. When using a single origin, we noted that the location of the origin is key for suitably defining the topography of the protein surface. The protein center of mass can work well for globular proteins, but can be “off-center” for many proteins that are not nearly-spherical. Below we describe several ways to define O_1: their applicability depends in part on the geometry of the pocket itself. We note, however, that the use of multiple origins provides more robust results with respect to the location of O_1.

As a first step, the user can choose whether O_1 should simply be placed in the direction of the protein’s center of mass (this is default). If so, we place O_1 30 Å away from the center of the pocket (P) along the P→Q direction, where Q is the center of the protein. If not, we offer three distinct methods to set O_1: ***(1)*** We make use of the fact that pockets at protein-interaction sites are broad and flat, and thus we find the plane that best fits the pocket points (by minimizing the least-squares distance of points to the plane). We then place O_1 along the normal to the plane passing through P, so that distance P–O_1 = 30 Å, yielding two solutions (one “above” the plane of the pocket and one “below” the plane of the pocket). We then select the solution for which the O_1–Q distance is less (i.e. the rays will emanate from within the protein rather than from far above the pocket). ***(2)*** Alternatively, for pockets that are deeper and narrower,

we define a series of vectors $\mathbf{s}_i$, each of which defines the distance and direction from the i_{th} surface point to P. We then carry out a vector summation of all these vectors $\mathbf{s}_i$, such that we determine the direction of the pocket which most faces away from solvent. We place O_1 in this direction at a distance of 30 Å. ***(3)*** Finally, we offer the user fine control over the location of the origin by placing O_1 30 Å away from P along the P→R direction, where R is a user-specified residue.

Once O_1 is defined, four more origin points (O_2–O_5) are then defined. We add O_2–O_5 as follows: O_2 and O_3 are obtained by rotating O_1 by ±45° around vector $\mathbf{w}=\mathbf{u}\times\mathbf{v}$, where $\mathbf{u}$ is the vector from P to O_1, $\mathbf{v}$ is the vector from P to a randomly chosen point, and × denotes vector product; O_4 and O_5 are obtained by rotating O_1 by ±45° around vector $\mathbf{z}=\mathbf{u}\times\mathbf{w}$.

Electrostatic potential grid

To prepare the protein, we begin by using Rosetta to fill in any missing atomic coordinates and add hydrogen atoms. We then use OpenEye's "molcharge" program to add amber99 partial charges to each atom.

To generate the electrostatic potential grid, we use OpenEye's ZAP toolkit (201) (a finite difference Poisson-Boltzmann solver). We use 0.5 Å grid spacing, with 1.0 and 80.0 for the inner and outer dielectrics, and 2 Å distance as buffer between the molecule and the edge of the grid. Once we obtain the electrostatic potential grid that encompasses the whole protein, we extract from this a smaller grid that matches the dimensions of the "pocket grid" used for ray-casting (this also matches the bounds of the search space during the docking runs). To avoid extreme values that occur at certain grid points (i.e. very close to a charged atom) during docking, we set the maximum/minimum possible value of the electrostatic potential at each point to ±10 kT/e.

Running DARC with Rosetta (including electrostatics)

To include electrostatics, we first resize the electrostatic potential grid (generated as described above) to match the size of the interface pocket grid. This step can be carried out while generating the ray file:

```
Rosetta/main/source/bin/make_ray_files.macosclangrelease -protein
4ERF.pdb
      -central_relax_pdb_num 54,99 -bound_ligand 0R3_0001.pdb
      -add_electrostatics –espGrid_file 4ERF.agd -extra_res_fa
0R3.params
```

The output from this command will be a ray-file named "eggshell_4ERF_54,99.txt" and an electrostatic potential grid file named "DARC_4ERF.agd" which we will use as input for running docking using DARC.

Then we call DARC for running the docking calculations using the pre-generated ray-file and corresponding electrostatic potential grid as follows:

```
Rosetta/main/source/bin/DARC.macosclangrelease -protein 4ERF.pdb -ligand
0R3_0001.pdb
      -extra_res_fa 0R3.params -ray_file eggshell_rosetta_4ERF_54,99.txt
```

To include electrostatics score:

```
Rosetta/main/source/bin/DARC.macosclangrelease -protein 4ERF.pdb -ligand
0R3_0001.pdb
      -extra_res_fa 0R3.params -ray_file eggshell_rosetta_4ERF_54,99.txt
–add_electrostatics -espGrid_file DARC_4ERF.agd
```

Generating conformers

For each ligand in our test set, we used the OMEGA software (44, 45, 187) to generate up to 300 conformers, using default parameters. The number of conformers used for each ligand in our study is reported in **Table S1** (these depend on the number of rotatable bonds and the ligand's geometry).

Statistical analysis

The statistical significance of the comparisons presented here was evaluated using Wilcoxon signed-rank test, as implemented in the R statistical computing environment (202).

3.6 Acknowledgements

We thank Andrea Bazzoli for valuable discussions. We are grateful to OpenEye Scientific Software (Santa Fe, NM) for providing an academic license for the use of OpenEye Toolkits, OMEGA, ZAP, and QuacPac.

CHAPTER 4

Rationally designing inhibitors of the Musashi protein-RNA interaction by hotspot mimicry

Yan Xia[1†], Ragul Gowthaman[2†], Lan Lan[1], Steven Rogers[3], Andy R. Wolfe[1], Christian L. Gomez[1], Oscar Ramirez[4], Bryan W. Tsao[1], Kelin Li[3], Jia Yu[1], Rebecca T. Marquez[1], Chunjing Liu[3], Manoj M. Pillai[4], Jeffrey Aubé[3,5], Kristi L. Neufeld[1,6], Liang Xu[1,7], and John Karanicolas[1,2]

[1] Department of Molecular Biosciences, [2] Center for Bioinformatics, [3] Center of Biomedical Research Excellence, Center for Cancer Experimental Therapeutics, [5] Department of Medicinal Chemistry, [6] Department of Cancer Biology, KU Medical Center, and [7] Department of Radiation Oncology University of Kansas, Lawrence, KS 66045

[4] Section of Hematology, Yale Cancer Center, 300 George St., New Haven, CT 06511

[†]Equal author contributions.

4.1 Abstract

RNA-binding proteins (RBPs) are key regulators of post-transcriptional gene expression, and underlie many important biological processes. Here, we develop a strategy that entails extracting a "hotspot pharmacophore" from the structure of a protein-RNA complex, and using this as a template for designing small-molecule inhibitors. With this approach we first target Musashi-1, a stem-cell marker that is upregulated in many cancers. We design and synthesize novel inhibitors that are active in biochemical and cell-based assays against Musashi-1, and then demonstrate how these inhibitors can be used as tool compounds to probe the activity of close homolog Musashi-2. This study extends the paradigm of "hotspots" from protein-protein complexes to protein-RNA complexes, supports the "druggability" of RNA-binding protein surfaces, and represents the first rationally-designed inhibitors of non-enzymatic RNA-binding proteins. Owing to its simplicity and generality, we anticipate that this approach may also be used to develop inhibitors of many other RNA-binding proteins.

4.2 Introduction

RNA-binding proteins (RBPs) play crucial roles in many diverse cellular processes. They regulate the life cycle of mRNAs by controlling splicing, polyadenylation, stability, localization and translation, and also modulate function of non-coding RNAs (8). Mammalian proteomes are thought to include upwards of 800 RBPs (203, 204), corresponding to both RNA-processing enzymes and non-enzymatic RNA-binding proteins. In light of the broad range of functions carried out by RBPs, the goal of this study is to devise a general and robust strategy for designing chemical tools that will allow precise manipulation of the interactions between RBPs and their cognate RNAs. We expect that such tools will help unravel the mechanisms of important biological processes controlled by RBPs, and may also serve as a starting point to validate RBPs as targets for therapeutic intervention (205-207).

To date, there exist few classes of compounds that target protein-RNA interactions. Inhibitors of certain RBPs have been identified via high throughput screening (208, 209), including one series from virtual screening that competes with double-stranded RNA for binding to toll-like receptor 3 (210), and a number of compounds have been reported that disrupt binding by interacting with the RNA rather than with the RBP (211, 212). Among rationally designed small-molecule inhibitors that target RBPs, however, all examples reported to date can be categorized into two general classes. The first class comprises nucleoside analogues (213-216), such as anti-HIV-1 NRTIs, that mimic the chemical structures of natural-occurring nucleosides and rely on enzymatic processing by their targets to form covalent adducts (10). While nucleoside analogues can be straightforward to design, the inability of these molecules to provide sufficient binding affinity or selectivity without covalent linkage has prevented this strategy from being extended to non-enzymatic RBPs. The second class of compounds comprises allosteric inhibitors (216, 217), such as anti-HIV-1 NNRTIs, that bind to secondary sites on the protein target and shift its conformation to an inactive state (218). In principle, allosteric inhibitors could be used to target both enzymatic and non-enzymatic RBPs; in practice, however, challenges associated with both

identifying allosteric sites and then finding small molecules to complement these sites has limited the general utility of this approach to all but a few cases. Collectively, the fact that these RNA-binding protein surfaces are not thought to have evolved to bind small molecules makes them a "non-traditional" class of drug target. Moreover, the relatively flat and polar nature of protein surfaces in this class typically leads to poor performance by structure-based virtual screening (docking) approaches (71), and given the lack of a known small-molecule binding partner it is even unclear *a priori* that such protein surfaces are suitable for inhibition by any small-molecule ligand at all (14).

Here, we present a new approach for rationally designing small-molecule inhibitors of RBPs. We draw inspiration from a related class of "non-traditional" drug targets, protein-protein interfaces. In a protein-protein complex, each of the individual interfacial residues typically do not contribute equally to the energetics of binding; rather, the majority of the binding affinity derives from a small number of "hotspot" residues (219-221). This observation, in turn, motivated several groups to mimic these key interactions when designing small-molecule inhibitors (156, 222-224). In this study, we take the "hotspot" paradigm and extend it to protein-RNA interactions.

Our approach entails identifying the chemical moieties of a given RNA that contribute critical interactions to a particular protein-RNA complex, and then identifying small molecules that recapitulate the precise geometrical arrangement of these moieties. Our underlying hypothesis is that compounds capable of mimicking the three-dimensional structure of the RNA "hotspot" will also mimic the energetically dominant interactions in the protein-RNA complex, using a much smaller chemical scaffold. By establishing a new method for reusing these protein-RNA interactions, we circumvent the challenging problem of needing to design interactions that target a flat, polar protein surface.

4.3 Computational Approach

Computational methods are implemented in the Rosetta software suite (60) unless otherwise indicated. Rosetta is freely available for academic use (www.rosettacommons.org), with the new features

described here included in the 3.6 release. Computational methods are summarized below, and presented in further detail in the *Supporting Methods* section.

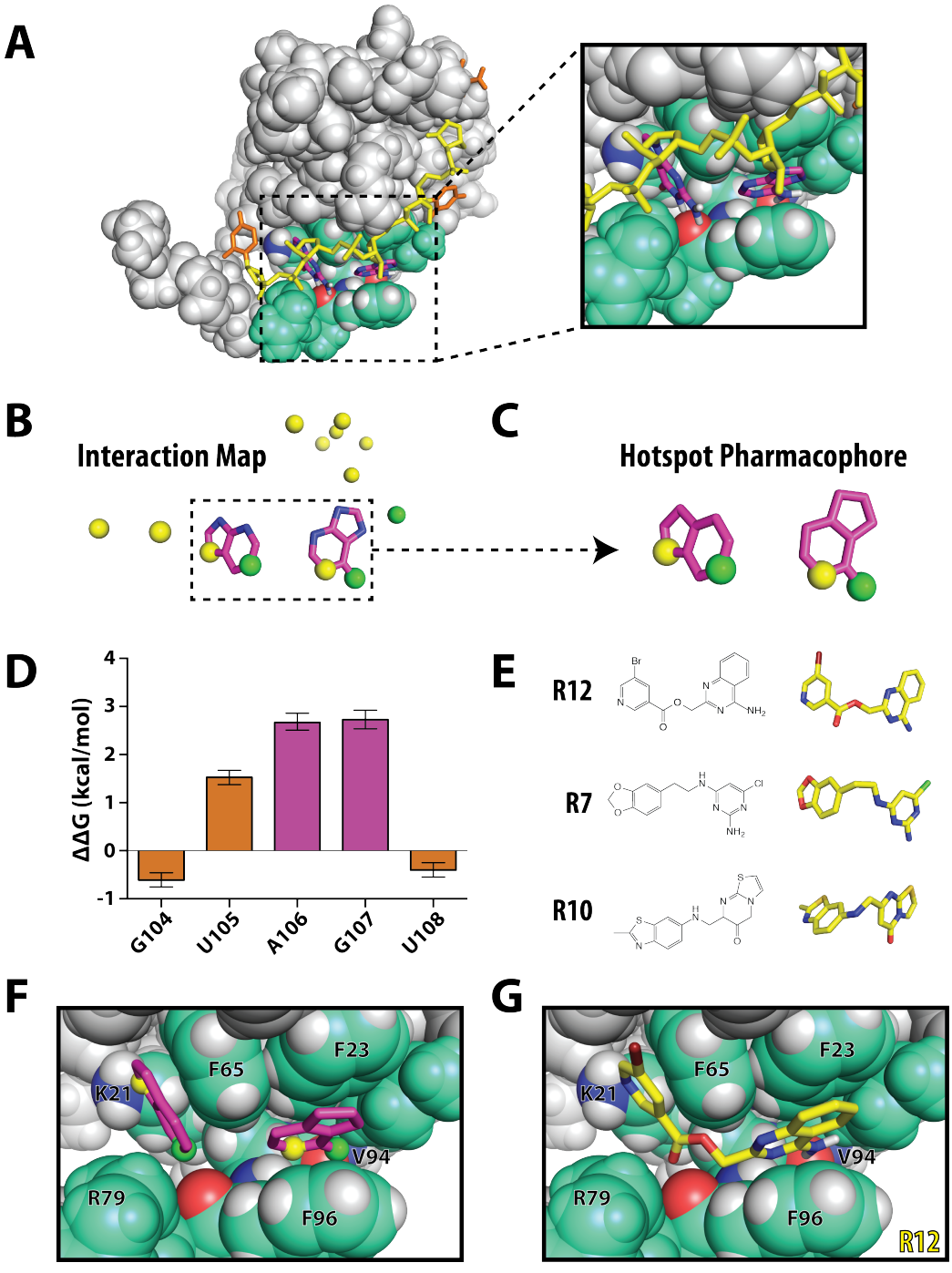


Figure 1: The hotspot mimicry approach. We demonstrate this approach by applying it to the Msi1 / *NUMB* mRNA interaction. **(A)** The structure of the Msi1 / RNA complex. The RNA (*yellow and orange sticks denoting the backbone and the bases, respectively*) wraps around the protein (*green and grey spheres*). Two adjacent bases, A106 and G107 (*magenta*), are buried in a shallow pocket on the protein surface. **(B)** An interaction map is generated from the RNA in the complex, by collecting deeply buried bases (*magenta*) and atoms involved in intermolecular hydrogen bonds (*acceptors shown in yellow, donors in green*). **(C)** Components of the interaction map are clustered in space, and atoms that do not participate in hydrogen bonding are reverted to carbon atoms; this produces a "hotspot pharmacophore." **(D)** The difference in binding free energy between an RNA harboring a single abasic site versus the wild-type *NUMB* mRNA, as determined through competition with a fluorescently-labeled RNA. Positive values indicate diminished binding when a given base is replaced with an abasic site, showing that A106 and G107 contribute more than the other nearby bases to Msi1 / *NUMB* mRNA binding affinity. **(E)** The hotspot pharmacophore serves as a template for ligand-based screening. In this case we identified three classes of hit compound that mimic the three-dimensional features of the pharmacophore, as exemplified by the representatives shown here. **(F)** Superposition of the hotspot pharmacophore back onto the protein structure illustrates the interactions that should be captured by an ideal ligand: stacking against three aromatic sidechains, and four intermolecular hydrogen bonds. **(G)** Superposition of R12 onto the protein structure shows that this compound is expected to preserve the aromatic stacking, and recapitulate three of the four hydrogen bonds.

Building "hotspot pharmacophores"

While interfaces between RBPs and their cognate RNAs are mostly flat, complexes involving segments of single-stranded RNA often include a few interfacial nucleobases that are buried much more deeply than the others (**Figure 1a**); this uneven distribution is reminiscent of "hotspot" sidechains in protein-protein complexes (219, 220). The protein has evolved to interact with these buried nucleobases through precise intermolecular aromatic stacking interactions and hydrogen bonding.

We have developed an automated framework that distills the structure of a protein-RNA complex to a "hotspot pharmacophore," which in turn can serve as a template for ligand-based screening. Our framework first picks out those RNA aromatic moieties that are deeply buried in the protein-RNA complex, as well as any RNA atoms involved in intermolecular hydrogen bonds to the protein or ordered water molecules (**Figure 1b**). Any polar atoms on the nucleobases that do not participate in hydrogen bonds are then replaced with carbon atoms, since those polar groups need not be carried forward into inhibitor design. This gives a broad spatial map of the protein-RNA interaction, which typically cannot be spanned by a single drug-like small molecule; we therefore clustered neighboring moieties, and advanced each cluster separately. Through this approach, we reduce the structure of the protein-RNA complex to a minimal "hotspot pharmacophore" that encapsulates the key interactions to be recapitulated by a small molecule (**Figure 1c**).

Identifying complementary ligands

To identify such compounds, we used this hotspot pharmacophore as a template for carrying out ligand-based virtual screening. In order to facilitate rapid characterization of compounds emerging from our screen, we restricted our search to the ~7 million compounds in the ZINC database (43) that are both commercially available, and predicted to have drug-like physicochemical properties. We used OMEGA (OpenEye Scientific Software, Santa Fe, NM) (44-46) to build low-energy conformations of each compound, then ROCS (OpenEye Scientific Software, Santa Fe, NM) (88, 225) to align each

conformation to our hotspot pharmacophore. For each of the top-scoring hits emerging from ROCS, we then used the aligned orientation to position the compound relative to the protein, and evaluated the interaction energy of the protein-ligand complex using the fullatom Rosetta energy function (60).

Musashi-1, an RRM-containing protein

The approach described above can, in principle, be applied to the structure of any protein-RNA complex. As a first test, we selected a target from the most common and well-studied of RNA-binding modules, the RNA-recognition motif (RRM) domain. Hundreds of structures of RRMs have been deposited in the Protein Data Bank, including more than fifty in complex with RNA (226). Collectively these structures show that RRMs adopt a conserved fold that packs two α-helices against one face of a four-stranded β-sheet; in most cases the opposite face of this β-sheet is then used to bind a single-stranded segment of RNA. Recognition of cognate RNA is usually driven by a cluster of three outward-facing aromatic amino acids on this β-sheet, which often form stacking interactions with a pair of adjacent RNA bases (227). Accordingly, mutations to the protein that remove these aromatic sidechains have been shown to disrupt binding in representative RRMs (227, 228), as has introduction of non-canonical bases to the RNA that alter the pattern of hydrogen bonding groups (229-231). Despite these shared features, however, the precise geometry of the dinucleotide pair in its complex with the RRM can differ very drastically across members of this family (227).

Mammalian Musashi-1 (Msi1) recognizes its cognate RNAs through a pair of RRMs, RBD1 and RBD2 (232). Together these two domains bind to the 3' UTR region of specific target mRNAs, including the mRNA encoding *NUMB*, and impede initiation of their translation (233, 234). *NUMB* mRNA encodes an inhibitor of Notch, so translational inhibition by Msi1 triggers Notch signaling and thus promotes self-renewal and cell survival (18, 234). Relative to its protein levels in normal tissue, Msi1 is over-expressed in many cancers including colon adenocarcinomas, medulloblastoma, glioma, astrocytoma, retinoblastoma, hepatoma, and endometrial, cervical, and breast carcinomas, and has particularly high

levels in later stages of cancer progression (235-241). Considering the role of Msi1 in stem cell maintenance and renewal, and its over-expression in a wide array of cancers, disrupting its RNA-binding ability may inhibit cancer stem cells that play a role in drug- and radio-resistance, and thus serve as an attractive potential anti-tumor strategy (242).

4.4 Results

Computational screening against Msi1 RBD1

We applied our "hotspot mimicry" approach to the Musashi-1 RBD1 / *NUMB* mRNA complex (232), and found a single hotspot pharmacophore derived from an adjacent pair of buried nucleobases, Adenine106 and Guanine107 (**Figure 1a**). This pharmacophore captures both the aromatic stacking and the hydrogen bonding of the RNA hotspot through its inclusion of ring moieties and donor/acceptor positions, respectively (**Figure 1c**). To test whether these particular two bases indeed serve as a hotspot of the Msi1 RBD1 / RNA interaction, we used a fluorescence polarization competition assay (see *Supporting Methods*) to measure the binding affinity of *NUMB* mRNA variants that lacked individual bases. Using this assay, we found that introduction of an abasic site at either of these two positions led to a marked decrease in binding to Msi1 RBD1 (**Figure 1d**). In contrast, introduction of an abasic site at other nearby positions affecting binding much less. Confirmation that A106 and G107 serve as hotspot bases of this interaction thus provided experimental evidence supporting the pharmacophore selection from our computational approach.

We then used this pharmacophore as a template for virtual screening, and found that the 12 top-scoring hits could each be classified into one of three diverse chemotypes (**Figure S1**). While none of these scaffolds bear any obvious resemblance in chemical structure to a nucleobase pair, the overlap in three-dimensional shape and hydrogen bonding potential between the hotspot pharmacophore and the modeled conformation of each compound is immediately evident (**Figure 1e**). Despite this strong

similarity, none of the 12 hit compounds recapitulated all four of the polar groups included in the hotspot pharmacophore, and only three hit compounds matched to three of the polar groups: R12, its close analog R4, and R7. The lack of hits that simultaneously match all four polar groups reflects a limit of the chemical space spanned by our screening library; we will discuss this in detail later.

As expected, superposition of the hit compounds back onto the hotspot pharmacophore in the context of the protein-RNA complex confirmed that these ligands might preserve the favorable interactions of the dinucleotide pair. In particular, the ring moieties in the pharmacophore represent the stacking of nucleobases against Phe23, Phe65 and Phe96 of Msi1, while the hydrogen bonding atoms indicate polar contacts with the sidechain of Lys21 and the backbones of Val94 and Phe96 of Msi1 (**Figure 1f)**. Mimicry of these interactions through the hotspot pharmacophore allows the hit compounds to recapitulate these interactions, as exemplified by R12 (**Figure 1g**). In this model R12 adopts a similar three-dimensional geometry as the hotspot pharmacophore, and thus recapitulates its aromatic stacking and polar interactions.

Biochemical characterization of computational hits

We purchased each of the compounds corresponding to these 12 top-scoring hits (**Figure S1**). We used surface plasmon resonance (SPR) to directly test for binding of each compound to Msi1, by immobilizing recombinant human Msi1 RBD1 onto an SPR chip and then passing each compound over the chip at a concentration of 50 μM (see *Supporting Methods*). The sensorgram for R12 showed a kinetic profile consistent with binding to Msi1 RBD1 (**Figure 2a**); none of the other compounds exhibited this behavior (**Figure S2**).

As noted earlier, only R12 and two other compounds matched as many as three polar groups in the hotspot pharmacophore; the lack of binding observed for the other compounds (at this concentration) may be attributable to the fact that they do not sufficiently recapitulate the interactions of the hotspot pharmacophore. While R7 matched three polar groups, retrospective analysis of the structural model

revealed that the imperfect alignment of the rings to the pharmacophore may have led to a steric clash with the protein (**Figure S3**).

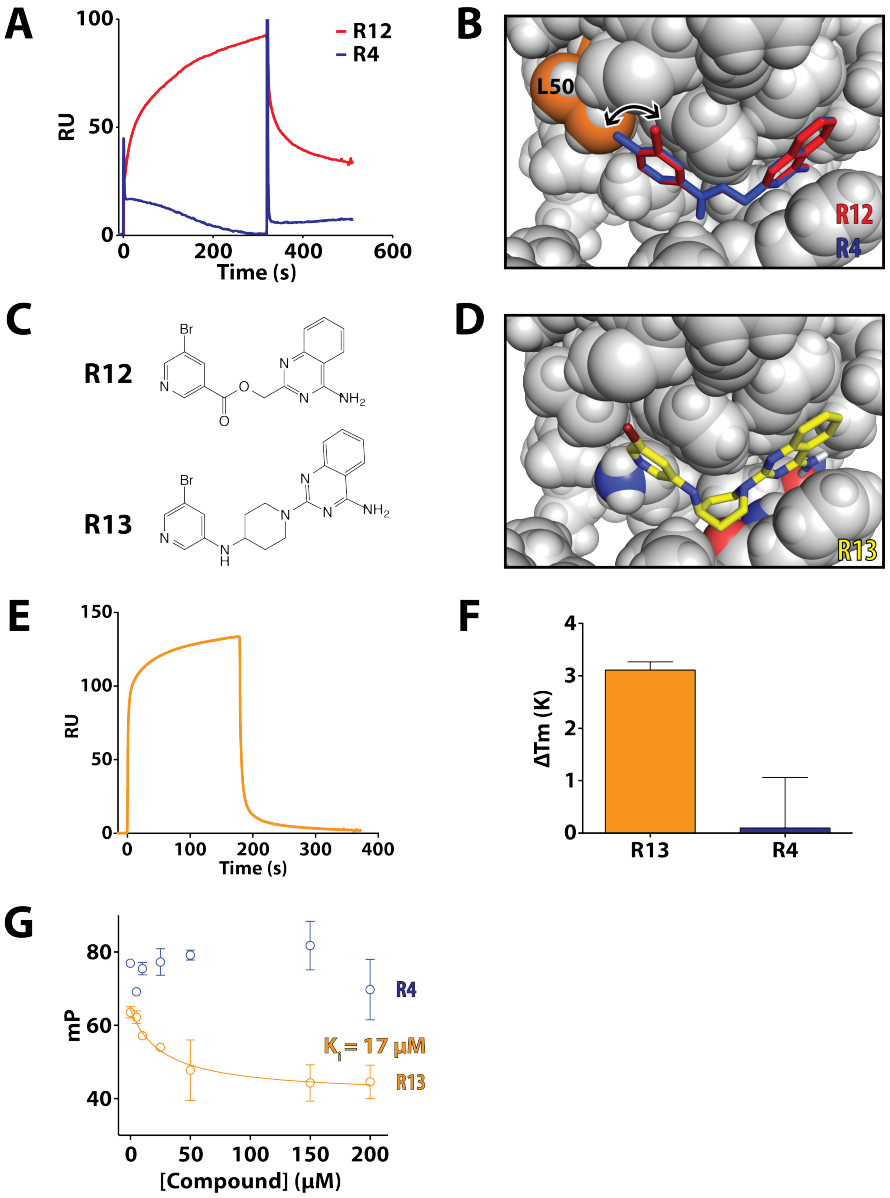


Figure 2: Biochemical characterization and optimization of computational hit compounds. **(A)** Initial screening via surface plasmon resonance: representative sensorgrams for R12 and R4 are shown. The kinetic profile of R12 (*red*) is consistent with binding to Msi1, whereas that of its close analog R4 (*blue*) shows no evidence of binding. **(B)** Comparison between the predicted binding models of R12 and R4. The top scoring conformers of R12 (*red*) and R4 (*blue*) are transferred back to the protein by alignment to the hotspot pharmacophore. The model of R4 suggests its lack of binding may stem from a steric clash with Leu50, whereas R12 avoids this steric clash since this ring is substituted at a different position. **(C)** Chemical structures of R12 and one of its derivatives, R13. **(D)** Model of R13 bound to Msi1, by alignment to the hotspot pharmacophore. R13 preserves the interactions of R12, but reduces flexibility of the linker and removes potential electrostatic repulsion with Msi1. **(E)** R13 increases the Msi1 melting temperature in a concentration-dependent manner, providing evidence of their interaction in solution. **(F)** R13 competes with fluorescein-labeled RNA for Msi1 binding, as observed through a fluorescence polarization assay.

Interestingly, the R12 class included two compounds: R12 and R4. These compounds differ only in the position and identity of a single substituent: the R12 has a bromine atom at the 3- position of the pyridine moiety, while R4 instead harbors a methyl group at the 2- position. Comparison of these compounds in the context of the protein partner immediately reveals a potential source for their differing responses in the SPR experiment: in our model, this methyl group of R4 forms a steric clash with the side chain of Leu50 on Msi1 that we had not initially recognized (**Figure 2b**); in contrast, the shifted position of the R12 substituent avoids this steric clash. This initial (inadvertent) structure-activity experiment provides strong support for the structural model of R12 binding.

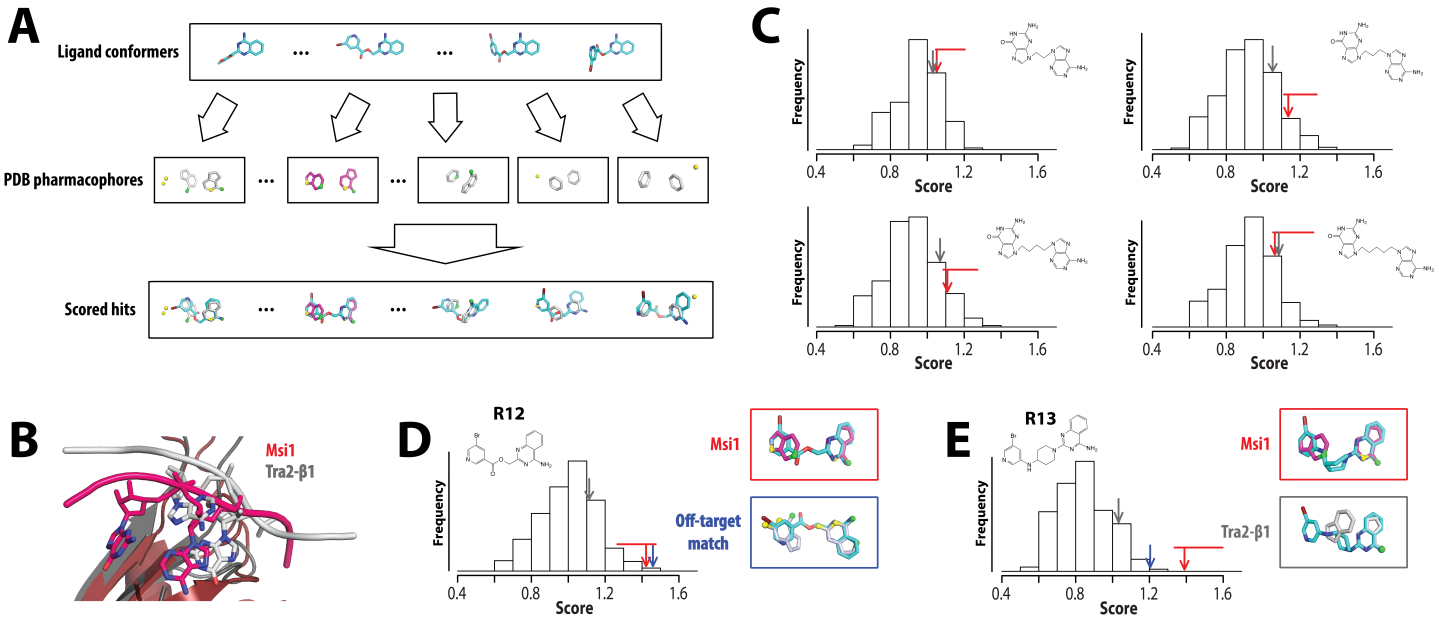

A
Ligand conformers
PDB pharmacophores
Scored hits
B
Msi1
Tra2-β1
C
Frequency
Score
0.4
0.8
1.2
1.6
D
R12
Msi1
Off-target match
E
R13
Msi1
Tra2-β1

Figure 3: Computational prediction of potential off-target interactions. (A) We screened each conformer of a given ligand against the complete set of unique hotspot pharmacophores from other protein-RNA complexes in the PDB. Hits in this screen correspond to other proteins that recognize their cognate RNAs using interaction geometry that can be mimicked by the compound of interest. **(B)** Like Msi1, the RRM from Tra2-β1 binds to specific purine-rich RNAs. The distinct geometry with which the cognate RNA is recognized, however, produces a hotspot pharmacophore that is dramatically different from that of Msi1. **(C)** Application of this approach to a series of hypothetical compounds built by connecting adenine and guanine with various flexible linkers. The distribution of scores for the complete pharmacophore library is shown, with the score of the Msi1 (*red arrow*) and Tra2-β1 (*grey arrow*) pharmacophores indicated. These artificial compounds match the Msi1 and Tra2-β1 pharmacophores equally well, and they match many other pharmacophores better than they match these two. **(D)** Application of this approach to R12. Only one "off-target" pharmacophore from the library has a score comparable to that of Msi1 (PDB ID 2MF0); this match is accessed through the conformational flexibility of R12 (*blue arrow/box*). **(E)** Application of this approach to R13. No pharmacophores from the library have scores comparable to that of Msi1; the scores of the off-target matches to R12 are now significantly worse (*blue arrow*), since R13 can no longer access these alternate conformations. Neither R12 nor R13 provides a high-quality match to the Tra2-β1 hotspot pharmacophore (*grey arrows/box*). In all cases the *red arrow* indicates the score of the hotspot pharmacophore derived from the first model in the Msi1/RNA NMR ensemble, and the *red bar* indicates the range of scores spanned by pharmacophores extracted from the other members of this NMR ensemble.

Optimization of R12

Guided by this model, we next set out to improve the potency of R12. The limited chemical space in our screening library led to two undesirable features of this compound. First, the carbonyl oxygen in the ester linker of R12 is positioned in close proximity to the Phe96 backbone carbonyl of Msi1 (**Figure 1f**); beyond simply the lost opportunity for an intermolecular hydrogen bond, we expect electrostatic repulsion between these two negatively charged moieties. Second, the two ring systems of R12 are connected by a somewhat flexible linker; rigidifying this linker might reduce the conformational entropy lost upon binding. With these two motivations in mind, we designed and synthesized a panel of five new R12 derivatives, R13-R17 (**Figure S4**) (see *Supporting Methods*).

Using our previous SPR assay, we found that all five derivatives exhibit kinetic profiles at 50 μM consistent with binding to Msi1 (**Figure S5**). Below we will present further biochemical characterization of R13, a compound that met our design goals by replacing R12's ester with a piperidine ring and a secondary amine in the linker (**Figure 2c**). Upon building models for each of the five R12 derivatives in complex with Msi1 (see *Supporting Methods*), we find that the rigidified linker in each compound allows recapitulation of R12's interactions while relieving the potential source of electrostatic repulsion; unfortunately however, none of the models include an additional hydrogen bond to Msi1 (**Figure 2d, Figure S6**).

In order to confirm binding of R13 to Msi1 in an orthogonal assay, we used differential scanning fluorimetry (DSF / Thermofluor) to determine protein thermostability as a function of ligand concentration. We found that the melting temperature of Msi1 increased upon addition of R13 in a dose-dependent manner (**Figure 2f**), up to a 3.7 ˚C increase in presence of 100 μM R13.

Finally, we directly examined the ability of R13 to not simply bind Msi1, but also to inhibit its interaction with *NUMB* mRNA. We used a fluorescein-labeled RNA oligonucleotide corresponding to the Msi1 recognition sequence of *NUMB*, which exhibits an increase in polarization upon Msi1 binding (**Figure S7**). The subsequent addition of R13 is expected to lead to a decrease in polarization, if R13

competes with *NUMB* mRNA for Msi1 binding as designed. Indeed we observe this dose-dependent behavior, and using the apparent binding affinity of the labeled RNA for Msi1 we estimate Ki for the R13-Msi1 interaction to be 17 μM (**Figure 2g**).

Predicting target selectivity

Many RRM proteins recognize their target RNAs with high sequence specificity, through additional interactions outside the central RNA dinucleotide (227). Our mimicry of the Msi1 hotspot was predicated on recapitulating the interactions solely within this dinucleotide; we therefore sought to explore the target selectivity expected for these inhibitors by searching for potential off-target interactions. Starting from every example of protein-RNA complexes in the Protein Data Bank, we used our computational approach to extract the set of all available hotspot pharmacophores (see *Supporting Methods*). For a given compound of interest, we can then screen all conformers of this molecule against this "library" of 543 unique hotspot pharmacophores (**Figure 3a**). The top-scoring hits in this experiment represent proteins that recognize their cognate RNAs through interaction patterns that can be recapitulated by the compound of interest, making these candidate proteins for off-target binding. We note that this large-scale experiment does not explicitly account for protein flexibility, which may enable further off-target interactions. To demonstrate the variation in pharmacophore structure associated with typical protein fluctuations, we have included in our studies each member of the experimentally-derived Msi1-RNA NMR ensemble (232). As a further point of comparison, we also highlight another protein domain that binds to a purine-rich segment of RNA, the RRM from Tra2-β1 (243, 244). Despite the shared domain architecture and similar recognition preference, this protein binds its cognate RNA via interactions that are quite distinct from those of Msi1 (**Figure 3b**); this, in turn, leads to a hotspot pharmacophore with little resemblance to that of Msi1.

We applied this analysis first to a series of hypothetical compounds, each one comprised of guanine and adenine attached by flexible linkers of varying lengths (**Figure 3c**). We find that each of

these constructs can adopt a conformation that aligns well to the Msi1 hotspot pharmacophore, but they also undergo rearrangements that allow them to match many of the other hotspot pharmacophores in our library, including the one from Tra2-β1. This observation is unsurprising, since one would expect these artificial ligands to mimic many guanine-adenine dinucleotide pairs with little consideration of their three-dimensional arrangement.

We next carried out the same analysis for R12 (**Figure 3d**) and each of the other hits from our initial computational screen (**Figure S8a**). Relative to the guanine-adenine pairs, R12 lacks certain polar groups (those that do not participate in the Msi1 pharmacophore). While this reduces R12's potential for mimicking some of the off-target hotspot pharmacophores, we nonetheless find one alternate match with score comparable to that of Msi1: the pharmacophore is from a complex involving the bacterial RsmE protein (245). RsmE adopts a fold completely unlike an RRM domain, and no pharmacophores from other RRM domains had comparable scores to that of Msi1. Notably, the off-target hotspot pharmacophore from RsmE was identified *not* because it bears strong resemblance to the Msi1 pharmacophore, but rather because flexibility in the R12 linker allows it to match this alternate pharmacophore by adopting a drastically different conformation from that used in mimicking the Msi1 pharmacophore (**Figure 3d**, *blue box*).

Finally, we applied this analysis to R13 (**Figure 3e**) and each of the other R12 derivatives (**Figure S8b**); we find that these match the Msi1 hotspot pharmacophore far better than any others extracted from the PDB. The conformation that allowed R12 to mimic the RsmE hotspot pharmacophore is not accessible to R13, leading to a diminished score for this comparison (*blue arrow*). R13 is also unsuitable for mimicking the alternate geometry of the purine-purine pair from the Tra2-β1 hotspot pharmacophore (*grey arrow and grey box*), and indeed superposition of R13 onto this RRM domain results in obvious steric clashes (**Figure S9**). While this experiment does not account for flexibility of any of the off-target proteins, we do note that R13 matches some of the other members of the Msi1 NMR ensemble better than the single structure that led to its design (*red bar*). This highlights the robustness of

hotspot pharmacophore matching to small changes in protein structure, and suggests that protein flexibility is unlikely to lead to alternate hotspot pharmacophores that are preferred by R13. While further experimental evidence will be necessary to explicitly determine whether these compounds engage in unanticipated interactions with other RBPs, these results suggest that the increased rigidity of the R13 linker makes it unable to access the alternate conformations that might allow R12 to mimic certain off-target pharmacophores.

Collectively, these observations point to the uniqueness of the Msi1 hotspot pharmacophore with respect to the rest of the Protein Data Bank; while many other RBPs bind to guanine-adenine pairs, only Msi1 recognizes a guanine-adenine pair in precisely this geometry. Through the use of a rigid chemical scaffold that closely mimics the three-dimensional geometry of the hotspot pharmacophore, we expect to achieve target selectivity that would not be possible by direct mimicry of chemical structure (i.e. by using nucleoside analogues).

Inhibition of cellular Msi1 activity

Through the biochemical assays presented earlier, we confirmed that R12 and its derivatives R13-R17 bind to Msi1 *in vitro*. We next set out to explore whether these compounds inhibit RNA-binding of endogenous Msi1 in cells, by monitoring Msi1's interaction with cognate mRNAs (234). The tumor suppressor APC is a target of Msi1, and we previously demonstrated that a firefly luciferase reporter containing the 3' UTR from *APC* mRNA was responsive to Msi1 levels in HCT116-βw colon cancer cells (234). Since endogenous Msi1 inhibits translation through its interaction with the *APC* 3' UTR, we expected compounds that inhibit endogenous Msi1 activity would relieve this translational inhibition, and thus result in increased luciferase activity. In each case we co-transfected with an expression construct harboring Renilla luciferase, and normalized the relative light output to the latter; by doing so we account for potential differences in transfection efficiency and monitor the possibility that loss of the firefly luciferase activity might derive from compound toxicity. We also tested the effect of each compound on

cells that express a firefly luciferase reporter lacking the *APC* 3' UTR, to ensure that any observed differences were indeed Msi1-dependent.

At a concentration of 20 μM, we find that several of the R12 derivatives lead to increased luciferase activity from the construct containing the *APC* 3' UTR (**Figure 4a**). Given that Msi1 inhibits luciferase translation via the 3' UTR of *APC*, the observed increase in luciferase activity is consistent with neutralization of this Msi1 inhibition by the R12 derivatives. This effect is most pronounced for compounds R13, R15, and R17; we see little effect from R14 or R16 in this assay. As a negative control, we also include R4 (which did not exhibit inhibition in our biochemical assays), and find no evidence of inhibition. In the case of R12 we observe increased luciferase activity in the construct lacking the *APC* 3' UTR, suggesting some off-target activity of this compound. None of the other compounds increased activity of this control construct, however, confirming that activity of these R12 derivatives is indeed through Msi1.

While Msi1 does not bind the R12 derivatives as potently as it binds to its cognate RNA, the competition for Msi1 binding is driven not only by the relative binding affinities, but also by the relative concentrations of the competing species. By testing several concentrations of R13 we found that this effect is dose-dependent, and increases monotonically up to a concentration of 40 μM (**Figure 4a**). We also find that luciferase activity is diminished at a concentration of 60 μM, which may result either from off-target activity at this higher concentration or from negative feedback in response to inhibition of Msi1.

NUMB mRNA also harbors a 3' UTR that is recognized by Msi1, and was the first identified target of translational inhibition by Msi1 (246). The designed compounds are expected to relieve the interaction of endogenous Msi1 with the 3' UTR from *NUMB*, which in turn should increase levels of Numb protein. Upon treatment of HCT116-βw cells with these compounds, we indeed observe this expected response (**Figure 4b**). Further, across this series of compounds R4 again had the least effect. These results demonstrate the ability of the designed compounds to rationally perturb endogenous signaling pathways in human cells.

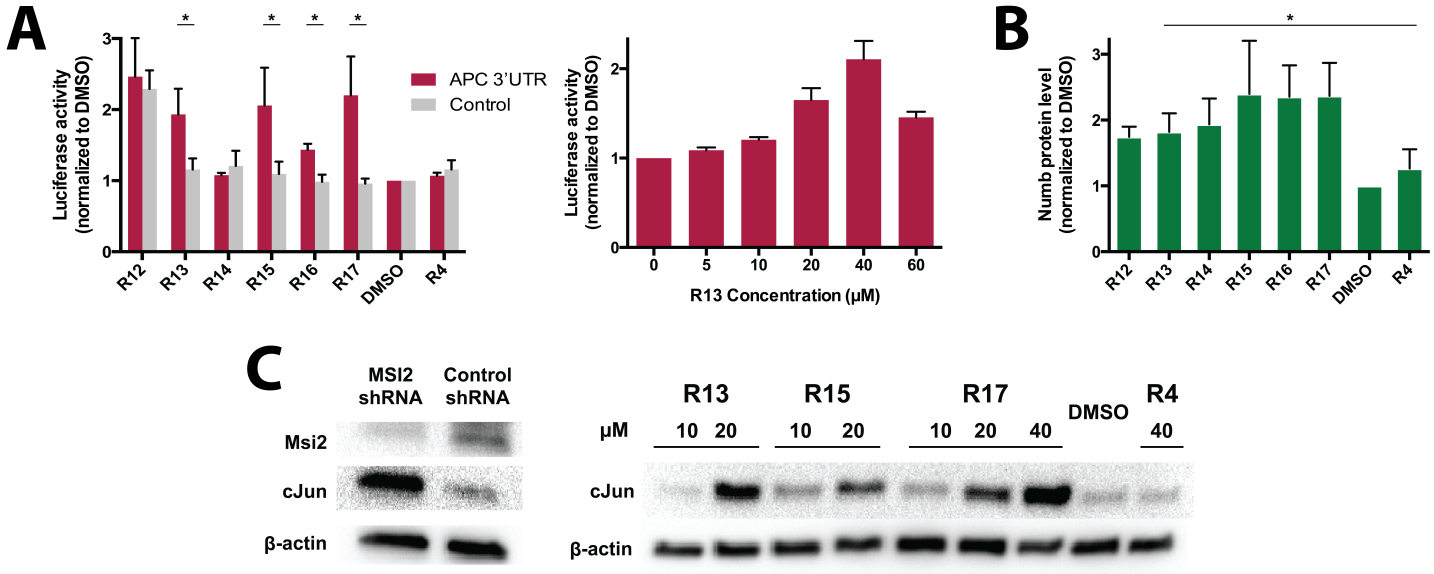

A
Luciferase activity (normalized to DMSO)
APC 3'UTR
Control
R12
R13
R14
R15
R16
R17
DMSO
R4
Luciferase activity (normalized to DMSO)
R13 Concentration (μM)
B
Numb protein level (normalized to DMSO)
C
MSI2 shRNA
Control shRNA
Msi2
cJun
β-actin
μM
10 20
10 20
10 20 40
40

Figure 4: Inhibition of Musashi's RNA-binding activity in cells. (A) The luciferase reporter construct contains the *APC* 3' UTR, and therefore responds to translation inhibition by endogenous Msi1. Luciferase activity increased after addition of R12 derivatives (each at 20 μM), consistent with Msi1 inhibition. *Control* indicates a construct lacking the *APC* 3' UTR. All values were normalized to Renilla luciferase activity and represent the average from 6 independent experiments. The response to compound R13 was also tested at increasing doses. Error bars in all cases represent SEM, and * indicates a statistically significant increase in luciferase activity relative to the construct lacking the *APC* 3' UTR ($p < 0.05$). **(B)** *NUMB* mRNA is a well-characterized target of translational inhibition by Msi1. As a second readout of endogenous Msi1 activity, levels of Numb protein were assessed via Western blot in lysates from cells treated with R12 derivatives. Normalized Numb band intensity from 4 independent experiments is presented (images of the Western blots, with quantification, are included as **Figure S10**). Error bars in all cases represent SEM, and * indicates a statistically significant increase relative to cells treated with R4 ($p < 0.05$). **(C)** Msi2 downregulates cJun in K562 cells, as demonstrated by the fact that anti-MSI2 shRNA leads to increased cJun protein. Treating these cells with R12 derivatives (R13, R15, R17) leads to increased levels of cJun (as detected in an anti-cJun Western blot). In contrast, the inactive analog R4 has levels of cJun that are comparable to vehicle alone (DMSO). Treatment with DMSO does not affect the level of cJun relative to untreated cells (**Figure S12**).

Inhibition of cellular Musashi-2 activity

While expression of Msi1 is tissue-restricted, its homolog Msi2 is ubiquitously expressed (247, 248). Like Msi1, Msi2 includes two RRM domains; the first of these shares 80% sequence identity with Msi1 RBD1. Sequence alignment of Msi1 and Msi2 reveals that with the exception of L50M, all but one of the residues that differ correspond to surface exposed positions far from the hotspot pharmacophore

(**Figure S11**); based on this model, we anticipated that the R12 derivatives would also show activity against Msi2's RBD1.

Msi2 plays a key role in hematopoietic stem cell proliferation and differentiation (247, 249) as well as in aggressive myeloid leukemia (247, 248), where it has generated extreme interest as a potential target for therapeutic intervention (250). Through polysome profiling we defined translational targets of Msi2 in K562 myelogenous leukemia cells, and identified transcription factor cJun as a potential downstream target of Msi2. This was confirmed using shRNA knock-down of Msi2 in K562 cell lines by immunoblotting (**Figure 4c**), providing evidence supporting Msi2's regulation of cJun but not indicating the mechanism by which Msi2 exerts this effect.

To test the hypothesis that Msi2 modulates cJun protein levels through its RNA-binding activity, we treated K562 cells with those R12 derivatives that were most active in our luciferase reporter assay. At a concentration of 20 μM, we found that all three of these compounds induce increased cJun protein levels (**Figure 4c**). In contrast, we did not observe any increase in cJun upon treatment with R4 even up to a concentration of 40 μM.

Leukemia cells lines, including K562, have negligible expression of Msi1 (247): this strongly suggests that indeed these R12 derivatives are active not only against Msi1, but against Msi2 as well. These observations are further consistent with the intended model of inhibition of Msi1, in which the designed compounds interact with the surface of Musashi's RBD1 that is most conserved between Msi1 and Msi2 (**Figure S11**).

4.5 Discussion

The ability to rationally design selective inhibitors of RNA-binding proteins in a robust and general way will enable development of new tool compounds to help elucidate cellular processes mediated by these interactions. Naturally-occurring examples have shown that proteins can mimic certain structural features of RNAs (251, 252); here, we instead encode a key RNA epitope on a small-molecule

scaffold. We demonstrate the application of our approach using Musashi-1, leading to a novel class of inhibitors that disrupt the RNA-binding activity of this tumor-promoting protein. By using the hotspot pharmacophore as a template for ligand-based screening, our approach circumvents the challenge of explicitly designing *de novo* interactions against a relatively flat and polar protein surface.

The major advantages of this mimicry approach are its generality and simplicity. In this first application of the RNA mimicry approach, we elected to restrict our initial screening to commercially available compounds. Though convenient, none of the resulting hit compounds provided complete recapitulation of the desired hotspot interactions. Of three compounds that each matched three of the pharmacophore's polar groups, only one compound (R12) complemented the protein surface without steric clashes. In light of the fact that this compound provided a starting point for new inhibitors of Msi1, and thus validated the computational method, in future it will be worthwhile to explore chemical space more extensively in search of hits that more effectively mimic the desired hotspot pharmacophore. A computational screening platform was recently described (155) that uses multi-component reaction chemistry (253) to build a virtual library containing tens of millions of novel compounds that can be readily accessed through proven "one-step, one-pot" reactions. While this strategy was originally used to construct a library of compounds that resemble collections of amino acid sidechains, it can be adapted to include privileged moieties that mimic patterns of hydrogen bond donors and acceptors in protein-RNA complexes, connected with rigid chemical linkers. By expanding the space of available compounds through this combinatorial strategy, and integrating computational screening with chemical synthesis, we envision discovery of compounds that more accurately match the target hotspot pharmacophores and thus exhibit improved potency prior to optimization.

The design of R12-derived compounds active against the Msi1 / *NUMB* mRNA interaction highlights the simplicity and robustness of the "hotspot mimicry" method, and also validates the "druggability" of this protein surface. We expect that the generality of this design strategy will allow it to

be applied broadly in future, to develop inhibitors of RNA-binding proteins as novel chemical probes and also as potential starting points for new therapeutics.

4.6 Methods

Detailed descriptions of computational and experimental methods are provided in the *Supporting Methods* section.

4.7 Acknowledgements

We thank David Johnson for valuable discussions. We are grateful to OpenEye Scientific Software (Santa Fe, NM) for providing an academic license for the use of ROCS, OMEGA, MolProp, and QUACPAC. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health through grant R01GM099959 (to J.K.), Institutional Development Award (IDeA) 8P30GM103495, and its support of the NIH University of Kansas Chemical Methodology and Library Development Center (P50GM069663). This work was also supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health through grant R01HL104070 (to M.M.P.), the National Science Foundation through XSEDE allocation MCB130049 (to J.K.), the Alfred P. Sloan Fellowship (to J.K.), the Kansas Bioscience Authority Rising Star Award (to L.X.), University of Kansas Cancer Center Pilot Grant (to L.X. and K.L.N.), and the University of Kansas Bold Aspiration Strategic Initiative Award (to L.X. and K.L.N.).

CONCLUSIONS

What structural properties define inhibitory complexes of non-traditional drug targets?

Over the past few decades, it has become generally believed that identifying small-molecule inhibitors for non-traditional targets such as protein-protein interactions was an impossible task (254). However, the recent discovery of several small-molecule inhibitors for protein-protein interactions demonstrated that non-traditional drug targets are in fact druggable. With these available examples, I have examined structural properties that define inhibitory complexes of non-traditional drug targets. From the results of this quantitative analysis, it is evident that small molecules binding at protein interaction sites retain more exposed surface area in comparison to small molecules bound to traditional targets. Also, the deep pockets present on traditional drug targets are typically absent at protein interaction sites. The results from the virtual screening experiment described in Chapter 1 demonstrates that these protein conformations present relatively flat surface pockets, and hence can harbor only a relatively exposed ligand. In order to address protein interaction sites with small-molecule inhibitors, we may need to carefully select proteins that are amenable for complexes with less exposed surface area, target allosteric sites that can accommodate extensively buried ligands, or develop new approaches to improve the delivery and oral accessibility for compounds outside rule-of-five chemical space. The results from this experiment further demonstrates that modern virtual screening methods typically are less suited for identifying inhibitors of protein interactions than for identifying inhibitors of traditional drug targets.

DARC: Matching pocket topography to a complementary small molecule

Next, I have presented DARC, a new approach to docking that matches the topography of a surface pocket "observed" from within the protein to the topography "observed" when viewing a potential ligand from the same vantage point. By matching the topography of the protein surface to the buried face of the ligand, DARC can outperform popular docking programs (DOCK and AutoDock) when screening

for inhibitors of protein-protein interactions. In the virtual screen carried out against the human anti-apoptotic protein Mcl-1 using DARC, 6 of the top scoring 21 compounds showed clear inhibition in a biochemical assay. The results from this virtual screen demonstrate the utility of DARC for identifying novel inhibitors of protein-protein interactions.

I have also presented several key enhancement to the DARC method; the starting version of DARC docked only one of the ligands in my test set to within 2 Å RMSD of its position in the crystal structure, whereas "DARC 2.0" achieved this level of accuracy for 12 of the 25 complexes. I anticipate these enhancements will help reduce the number of false negatives when we use DARC for real screening applications, thus allowing us to successfully identify more initial compounds with activity against the intended target.

Further potential applications of DARC:

In future, I anticipate that DARC can also be used for several other tasks related to non-traditional targets:

Focused chemical libraries:

There is great interest in designing "focused" chemical libraries: collections of compounds with enhanced hit rates against a given protein class. I believe that DARC might be used to address this challenge: by using DARC to screen against many diverse protein surfaces and compiling those compounds which are identified for these targets, one could build a collection of small-molecules that may be enriched in protein interface inhibitors.

One could also build a further target-specific library of small-molecules, for example a "Bcl-2 family focused compound library" by compiling together DARC hits for each of the Bcl-2 proteins. Target-specific, focused chemical libraries can not only increase the chances of identifying hit compounds in biochemical screens, but can also decrease the time and resources required for carrying out the screen.

Predicting binding specificity for the ligands:

Discriminating between family-nonspecific ligands and family-selective ligands is an important step in the advanced stages of the drug discovery pipeline. This is particularly important for families of proteins that perform opposing functions, like the Bcl-2 protein family . With the improved performance demonstrated by DARC 2.0 in pose recapitulation, I am enthusiastic to find out whether DARC can be used to predict which Bcl-2 family members will be inhibited by a given inhibitor.

Small molecules that mimic hotspot pharmacophores in protein-RNA complexes:

For rationally designing small-molecule inhibitors that target protein-RNA interactions, I have presented a new "pharmacophore mimicry" approach. I have demonstrated the application of this approach by identifying compounds that match the hotspot pharmacophore from the Msi1 / RNA complex, enabling development of novel inhibitors of the Musashi-1 / NUMB mRNA interaction; these inhibitors are active in both biochemical and cell-based assays.

Further potential applications of the hotspot mimicry approach:

The major advantages of this mimicry approach is its generality; accordingly, I expect that this design strategy will allow it to be applied broadly in future to develop inhibitors for other RNA-binding proteins.

Pharmacophores from dynamic protein-RNA complexes:

In this initial study targeting Msi-1, I considered only a single member of the NMR ensemble for this protein-RNA complex. However, considering hotspot pharmacophores from other members of this NMR ensemble may lead to identification of slightly different inhibitors. Moving forward, more studies will be required in order to reach general conclusions about how the dynamic nature of protein-RNA complexes can best be incorporated into this method.

Pharmacophores from homology models of protein-RNA complexes:

This initial application of the "hotspot mimicry" approach made use of an experimentally-derived structure of the protein-RNA complex. In principle, however, this approach might be extended to homology model complexes as well. With the increasing number of experimentally-derived structures of protein–RNA complexes, it may soon be possible to reliably model high quality protein-RNA complexes. Application of the hotspot mimicry approach to homology models of protein-RNA complexes will certainly prove challenging, and will require very accurate modeling in order to reliably recapitulate details of the pharmacophore. Nonetheless, if successful, this may provide new targets for structure-based drug discovery that would otherwise not be accessible by other methods.

Summary:

Collectively, the methods I have described in this thesis offer new ways to go about identifying small molecule inhibitors of non-traditional protein targets. This, in turn, may help make these exciting targets more accessible to drug discovery.

APPENDIX A.1: Supporting Information for Chapter 1

The structural properties of non-traditional drug targets present new challenges for virtual screening

Ragul Gowthaman[1], Eric J. Deeds[1,2], and John Karanicolas[1,2]

[1] Center for Bioinformatics and [2] Department of Molecular Biosciences,

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

1.1 Supplemental Tables

Protein target	PDB id	Inhibitor molecular weight (Da)	Potency (μM)	Ligand efficiency (kcal/mol per heavy atom)	θ_{lig}
βII tryptase	2bm2	402	0.015	0.35	0.56
penicillin G acylase	1gm8	350	16	0.27	0.46
urokinase	1owe	291	0.63	0.38	0.41
ADAM33	1r55	331	0.16	0.40	0.39
β-lactamase	1l2s	318	26	0.33	0.39
factor Xa	1lpz	467	0.025	0.32	0.38
Chk1	2br1	391	15.4	0.23	0.38
NS5B polymerase	1yvf	438	0.1	0.34	0.38
factor VIIa	1ygc	548	0.00035	0.34	0.37
vascular endothelial growth factor receptor 2	1y6b	479	0.038	0.30	0.37
tRNA-guanine transglycosylase	1n2v	208	83	0.37	0.36
tryptophan synthase	1k3u	290	-NR-	-NR-	0.36
cyclin-dependent kinase 2	1ke5	329	0.56	0.37	0.35
thymidine kinase	1ofl	252	4.1	0.41	0.34
thyroid hormone receptor α1	1nav	355	0.025	0.45	0.34
neuraminidase B	1vcj	351	26	0.25	0.34
thymidine phosphorylase	1uou	245	0.02	0.65	0.34
neuraminidase A	1l7f	330	0.0008	0.54	0.33
c-Abl tyrosine kinase	1opk	427	0.00015	0.46	0.33
protein kinase 5	1v0p	433	0.13	0.31	0.33
progesterone receptor	1sqn	298	0.0004	0.58	0.33
p38 kinase	1ywr	475	0.032	0.29	0.33
androgen receptor	1z95	430	0.076	0.33	0.33
thyroid hormone receptor β1	1n46	367	0.00003	0.53	0.32
purine nucleoside phosphorylase	1v48	337	0.0069	0.50	0.32
estrogen receptor α	1sj0	466	0.0008	0.37	0.32
thymidylate kinase	1w2g	242	27	0.37	0.32
c-Jun terminal kinase 3	1pmn	487	0.007	0.34	0.32
glucokinase	1v4s	349	1000	0.18	0.32
phosphodiesterase 5A	1xoz	389	0.0012	0.42	0.31
myosin II	1yv3	292	4.9	0.33	0.31
c-kit tyrosine kinase	1t46	496	0.413	0.23	0.30
heat shock protein 90	2bsm	388	0.14	0.34	0.30
thrombin	1oyt	408	0.057	0.33	0.30
cyclin-dependent kinase 5	1unl	354	0.2	0.35	0.30

glucocorticoid receptor	1m2z	392	0.06	0.35	0.29
glutamate receptor 6	1tt1	215	64.6	0.38	0.29
vitamin D nuclear receptor	1s19	413	0.0017	0.40	0.28
glycogen synthase kinase 3β	1q41	277	0.022	0.50	0.28
activated Cdc42 kinase 1	1u4d	245	-NR-	-NR-	0.28
transthyretin	1tz8	268	-NR-	-NR-	0.27
adipocyte fatty acid-binding protein	1tow	253	0.57	0.45	0.27
dihydrofolate reductase	1s3v	375	0.038	0.37	0.27
acetylcholinesterase	1gpk	244	4.3	0.41	0.27
HIV-1 reverse transcriptase	1jla	364	-NR-	-NR-	0.26
HIV-1 protease	1kzk	576	0.00004	0.34	0.24

Table S1: Inhibitors bound to traditional targets, a subset of the Astex set. Potency is taken from reported K_d or K_i values where available; if unavailable, IC_{50} values were used instead. In several cases ("-NR-"), no measure of potency has been reported. θ_{lig} indicates the fraction of ligand SASA exposed in the complex, as defined in Equation 1.

1.2 Supplemental Figures

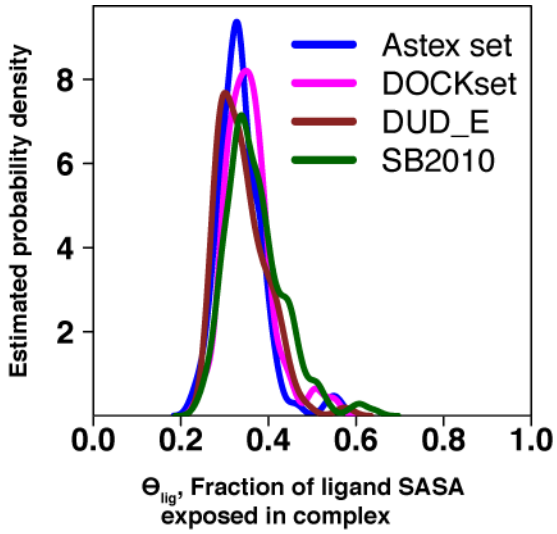


Figure S1 (complements Figure 1) The distribution of the extent of inhibitor solvent exposure (θ_{lig}) is similar across a number of drug-like sets: Astex (35) (*blue*), DOCK (255) (*magenta*), DUD-E (37) (*brown*), and SB2010 (38) (*green*).

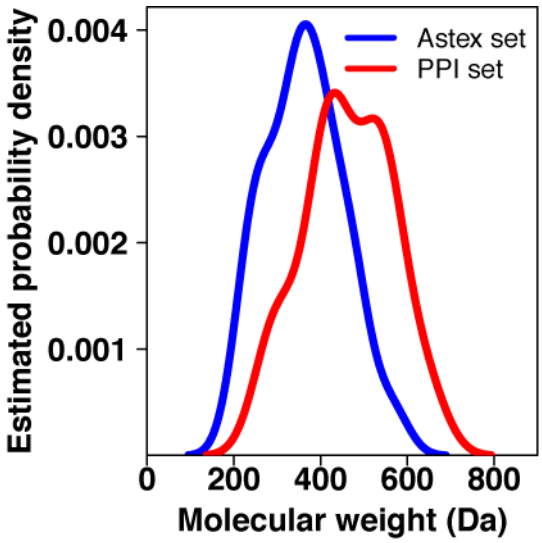


Figure S2 (complements Figure 2) The distribution of molecular weights for the inhibitors in each set underlies the observed difference in ligand efficiencies. Inhibitors binding at protein interaction sites (*red, median value 475 Da*) are typically larger than their drug-like counterparts (*blue, median value 355 Da*), and the difference in the means is statistically significant ($p < 10^{-4}$).

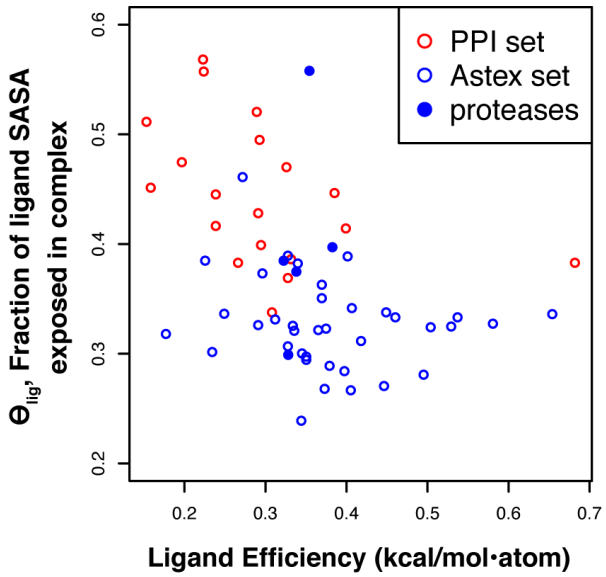


Figure S3 (complements Figure 2) The relationship between θ_{lig} and ligand efficiency for both the PPI set and the Astex set. While there is a statistically significant negative correlation between these properties for the PPI set (as noted in **Figure 2c**), no statistically significant correlation exists for the Astex set ($p = 0.27$).

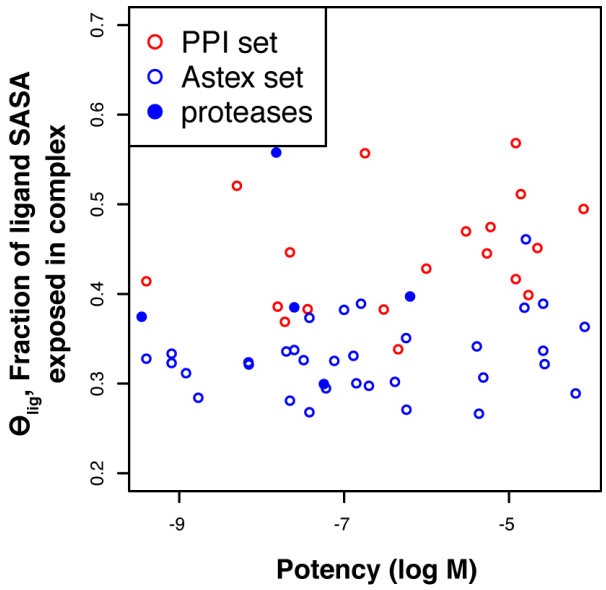


Figure S4 (complements Figure 2) The relationship between θ_{lig} and potency for both the PPI set and the Astex set. No statistically significant correlation exists between these properties for either set ($p = 0.33$ for the PPI set, $p = 0.50$ for the Astex set).

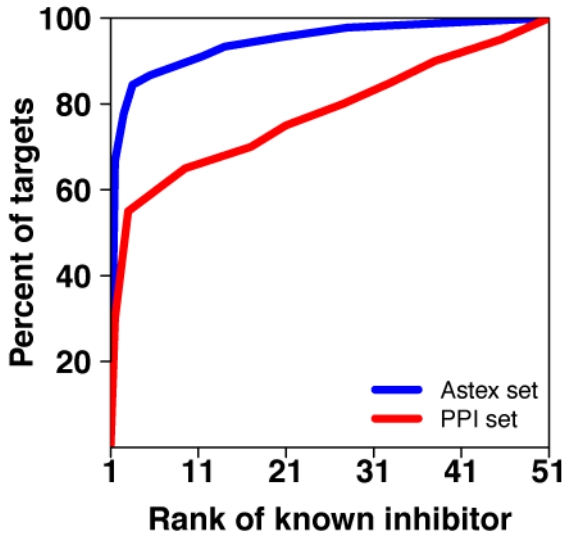


Figure S5 (complements Figure 3) A known inhibitor was embedded in a custom set of 50 "decoy" compounds selected by the DUD-E server to match the physical properties of the known inhibitor. FRED exhibits superior ability to identify the known drug-like inhibitors from the decoy compounds (*blue*), relative to inhibitors that bind at protein interaction sites (*red*), and the difference in the means is statistically significant ($p < 0.002$).

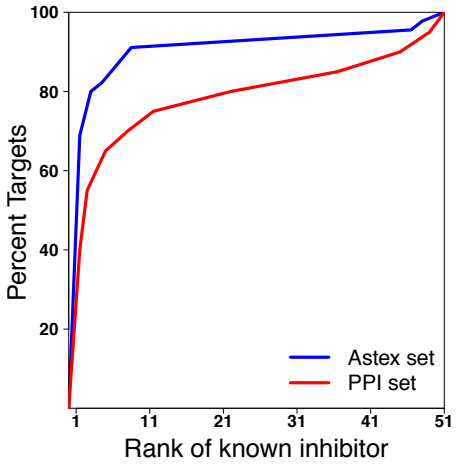


Figure S6 (complements Figure 3) The observation that virtual screening at protein interaction sites performs less well than for drug-like compounds holds for other docking software as well. DOCK 6.6 was used to identify a known inhibitor embedded in a custom set of 50 "decoy" compounds selected by the DUD-E server to match the physical properties of the known inhibitor. DOCK 6.6 exhibits superior ability to identify the known drug-like inhibitors from the decoy compounds (*blue*), relative to inhibitors that bind at protein interaction sites (*red*), though this difference is not statistically significant ($p = 0.11$).

APPENDIX A.2: Supporting Information for Chapter 2

DARC: mapping surface topography by ray-casting for effective virtual screening at protein interaction sites

Ragul Gowthaman[1†], Sven A. Miller[2†], Steven Rogers[3], Jittasak Khowsathit[2], Lan Lan[2], Nan Bai[2], David K. Johnson[1], Chunjing Liu[3], Liang Xu[2,4], Asokan Anbanandam[5], Jeffrey Aubé[3,6], Anuradha Roy[7], and John Karanicolas[1,2]

[1] Center for Computational Biology, [2] Department of Molecular Biosciences,

[3] Center of Biomedical Research Excellence, Center for Cancer Experimental Therapeutics,

[4] Department of Radiation Oncology, [5] Biomolecular NMR Laboratory,

[6] Department of Medicinal Chemistry, and [7] High Throughput Screening Laboratory

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

† indicates authors of equal contribution

2.1 Supplementary Results

Characterization of the M0/Mcl-1 interaction

To determine whether binding of M0 to Mcl-1 is reversible, we monitored their interaction via surface plasmon resonance (SPR). We immobilized Mcl-1 to the sensor surface, and then passed M0 over this surface at several different concentrations (see *Supplementary Methods*). While the resulting sensorgrams showed a kinetic profile consistent with reversible binding to Mcl-1 (**Figure S7**), the data deviate very slightly from a 1:1 Langmuir binding model. Although these data are insufficient to confidently allow quantitative comparison to the inhibition constant determined in the fluorescence polarization competition assay, these results nonetheless confirm that M0 binds reversibly to Mcl-1.

To further confirm binding of M0 to Mcl-1, we next produced isotopically-labeled Mcl-1 and collected 1H-15N-HSQC spectra in the presence and absence of M0 (**Figure S8a**). In light of concerns regarding M0 solubility, the concentration of M0 used in this experiment (100 μM) was less than the Mcl-1 concentration (300 μM): this is unfortunately expected to reduce differences in the spectra resulting from addition of M0. Nonetheless, even under these conditions we observe chemical shift differences in several peaks, providing further evidence for the interaction of M0 with Mcl-1. Further, the fact that chemical shift differences are limited to a small number of peaks suggests that residues distant from the M0 binding site are unaffected, and implies that binding does not induce a large-scale conformational change that might be expected of an allosteric inhibitor. Analogous experiments using Mcl-1's cognate peptide Bak show changes in a similarly limited subset of peaks (256), consistent with our model of binding in which M0 utilizes a very similar Mcl-1 conformation. Unfortunately, the fact that our Mcl-1 construct differs from the one used in the Bak spectra precludes direct comparison of whether the same peaks are shifted in each of these experiments.

After we collected our spectra, however, a newly assigned spectrum of unbound human Mcl-1 was made available (125); this newer spectrum corresponds much more closely to the construct used in

our studies. While this spectrum does not perfectly match our spectrum of unbound Mcl-1 (there are still slight differences between the constructs and buffer conditions), we are nonetheless able to confidently transfer assignments of several of the most disperse peaks (**Figure S8b**). While almost all of these unambiguously matched peaks do not shift upon addition of M0, two of the peaks that can be assigned *do* respond to addition of M0: Gly61 and Ser76 (**Figure S8c**). Mapping the location of these two residues to the structure of Mcl-1 reveals that they are located on opposing sites of the binding site in our model of the M0 complex, but are not in contact with the portion of the binding groove utilized by Bcl-xL inhibitors such as ABT-737 and WEHI-539 (**Figure S8d**).

In vitro cellular assay

We tested compound D2 using an *in vitro* cellular assay, by probing its effect of D2 on viability of DU-145 cells (a human prostate cancer cell line that produces high levels of Mcl-1, but little Bcl-xL and almost no Bcl-2 (257)). Despite this, compound D2 did not exhibit reduced viability at concentrations close to the 5 μM Ki observed in the FP assay (**Figure S11a**).

To explore whether this could be attributable to residual Bcl-xL and/or Bcl-2 activity rescuing cells from D2, we treating cells simultaneously with both D2 and ABT-263. The latter is an inhibitor of Bcl-xL and Bcl-2 that has been shown *not* to inhibit Mcl-1 (258, 259). Given these cells reliance on Mcl-1 for survival, ABT-263 does not begin to affect viability of DU-145 cells until concentrations of about 10 μM (**Figure S11b**); this is much higher than the concentrations needed for cells reliant on Bcl-xL or Bcl-2 for survival, or for DU-145 cells in which Mcl-1 has been knocked down via siRNA (results not shown).

If compound D2 failed to affect cell viability due to residual Bcl-xL and/or Bcl-2 activity, then treatment with ABT-263 should sensitize cells to D2. However, we find that ABT-263 does not affect cell survival in response to D2 (**Figure S11c**). This suggests that D2's lack of biological activity is *not* due to residual Bcl-xL and/or Bcl-2 activity, but rather failure of D2 to inhibit Mcl-1 in these cells.

2.2 Supplementary Methods

Implementation in Rosetta

Computational methods are implemented in the Rosetta software suite (60). Calculations were carried out using svn revision 52964 of the developer trunk source code. The new features described here are included in the Rosetta 3.6 release. Rosetta is freely available for academic use (www.rosettacommons.org).

Identifying surface pockets

As described in our earlier work (75), we have implemented into the Rosetta software suite a variation of the LIGSITE algorithm (76) for identifying potential small molecule binding sites on protein surfaces. To define pockets used in the benchmark experiments (**Figures 2, 3, S1-5**), we used a 10 Å grid around Val141 of Bcl-xL or Leu307 of XIAP. The structures used for Bcl-xL were 2yxj (inhibitor-bound) and 1r2d (unbound); the structures used for XIAP were 2jk7 (inhibitor-bound) and 1f9x (unbound). In our screen against Mcl-1, we used as a starting point the peptide-bound structure 2pqk (since no unbound structure was available at the time (125)), and defined the pocket using Phe270.

Pocket optimization

In our earlier work we developed the "pocket optimization" method for preferentially sampling protein fluctuations that contain a surface pocket (75). This method incorporates a term in the Rosetta energy function that favors pocket formation, and operates alongside the canonical energetic determinants of protein structure such as packing, hydrogen bonding, and solvation. Using this method we previously generated 1000 pocket-containing output structures starting from the unbound structures of both Bcl-xL

and XIAP (75), or the peptide-bound structure or Mcl-1. In the studies described here, we used the single lowest-energy conformation from each ensemble (energies were evaluated without the pocket-biasing term).

Choosing origin point for ray-casting

Determining where to place the "origin" (from which rays emanate) can contribute to the success of the ray-casting approach. Ideally, the rays should span the lower surface of the pocket uniformly, and should not intersect the protein again once they enter the pocket (i.e. the pocket should be purely concave, as viewed from a suitable origin). For globular proteins that are nearly spherical, it natural to choose the protein center of mass as the origin; however for proteins that are very non-spherical, an origin placed at the protein center of mass leads to rays that cover the pocket very non-uniformly.

To overcome this problem, we use an origin point directly "below" the pocket. We identify such a point by first determining the pocket center, and then identifying the two orthogonal vectors that account for the most variation around this point (the "width" of the pocket). We take a direction normal to these two vectors (corresponding to the "depth" of the pocket), and extend this vector 30 Å in the direction towards the center of mass of the protein (i.e. we place the origin "on the protein side" of the pocket, not "on the exposed side" of the pocket). While the origin identified in this manner is not necessarily located within the protein, the origin nonetheless is located at a vantage point that allows rays to span the pocket fairly uniformly.

DARC parameter optimization

The DARC score entails summing contributions for individual rays with one of four weights applied, depending on the intersections encountered by the ray (**Equation 1**). Since scaling all weights equally would not change the relative ranking of compounds / poses, there are effectively three free parameters. We fixed c1=1.0, and optimized the values of the other three weights (c2/c3/c4) as described

below. Through *ad hoc* testing of many variations of the values assigned to these weights, we find that the results of our benchmark experiments (**Figures 2 and 3**) are very robust to the particular weights used.

We compiled a training set comprised of seven different protein-protein interaction systems for which a structure had been solved in complex with a known small-molecule inhibitor (**Table S1**); none of the members of the training set are related to our test systems, Bcl-xL and XIAP (**Figures 2 and 3**), or to Mcl-1 (**Figures 4 and 5**). Starting with all four weights set to 1.0, we used simplex optimization (implemented via the GSL multidimensional minimization tools in C++, http://www.gnu.org/s/gsl/) to optimize the values of c2/c3/c4. At every step of the optimization, DARC was used to dock each inhibitor to the corresponding protein pocket with the current set of weights, and the sum of the ligand RMSDs from all seven complexes was used as the objective function. Upon completion of this optimization, we reached a set of weights that gave an average ligand RMSD of 1.2 Å when docking the seven compounds in our training set to their cognate target proteins; the value of these weights was c1=1.00 (fixed during optimization), c2=3.12, c3=13.32, and c4=8.13. We used these weights in carrying out all benchmark experiments (**Figures 2 and 3**) and also for the virtual screen against Mcl-1 (**Figures 4 and 5**).

Running DARC with Rosetta

Running DARC within Rosetta is a two-step process: first generating the ray-file from the protein surface, and second docking with DARC.

In the first step we generate the pocket at the interface and map the shape of the pocket shell into a spherical coordinate file, which is called the ray-file. To generate this ray-file we read the protein coordinates in PDB format and specify a target residue at the interface. If the protein has a bound ligand we can use the ligand to center the grid. The command to run DARC is as follows:

```
make_ray_files.macosgccrelease -protein 2YXJ.pdb -central_relax_pdb_num 141
       -bound_ligand N3C.pdb -extra_res_fa N3C.params
```

The output from this command will be a ray-file named “ray_2YXJ_141.txt” which we use as input for running docking using DARC.

In the second step, we run the docking calculations using the pre-generated ray-file. Here we give the input ligands for screening against the ray-file, as follows:

```
DARC.macosgccrelease -protein 2YXJ.pdb –ray_file ray_2YXJ_141.txt
       -ligand N3C.pdb -extra_res_fa N3C.params
```

The output for this command is a docked model of the protein-ligand complex named “DARC_2YXJ_N3C.pdb”.

Other optional flags to use when running DARC include:

```
-origin_cutoff 9 -atom_radius_scale 0.9 -num_particles 100 -num_runs
100 -missing_point_weight 13.3 -steric_weight 3.12 -extra_point_weight
8.13 -search_conformers false
```

Full documentation for these additional flags is distributed with the Rosetta source code.

Fullatom minimization in Rosetta

Fullatom minimization of the resulting DARC models can be carried out separately in Rosetta or as an additional option when invoking DARC. To minimize the DARC models immediately after docking we add the flag “-minimize_output_complex”:

```
DARC.macosgccrelease -protein 2YXJ.pdb –ray_file ray_2YXJ_141.txt
       -ligand N3C.pdb -extra_res_fa N3C.params –minimize_output_complex
```

This gives an additional output file named “mini_2YXJ_N3C.pdb”.

DOCK6 usage

For both Bcl-xL and XIAP, a surface was generated using the UCSF Chimera package (260). Spheres to define the binding pocket were selected within 10 Å of the bound ligand conformation. The same ligand conformers used to run DARC (those generated using OMEGA (44-46)) were also used to run DOCK, with charges added using QuacPac (59). The grid score based on a non-bonded molecular mechanics force field was used as the primary and secondary scoring function (to score and rank the compounds). All other DOCK parameters were set to values described elsewhere (255).

AutoDock usage

AutoDock4.2 (85) implements a Lamarckian genetic algorithm for virtual searching, by optimizing a semi-empirical energy function. For Bcl-xL, a grid box size of 40*80*60 Å with 0.375 Å spacing covering the hydrophobic groove at the interface was used. For XIAP, a grid box size of 50*50*50 Å with 0.375 Å spacing was used. Docking calculations were performed using the Lamarckian genetic algorithm with a mutation rate of 0.02 and a crossover rate of 0.8. For each ligand, 200 runs of docking were carried out with maximum 2.5 million energy evaluations and maximum 27,000 generations, to obtain 200 docked conformations. The final docked conformation was taken to be the one with the lowest energy, and this energy was used in ranking the compounds.

rDock usage

rDock version 2013.1 is a docking method that sequentially uses a genetic algorithm then minimization to generate low-energy ligand poses, using a pre-defined cavity at the binding site of the target protein (86). Docking sites at the protein interface were defined using the ligand-bound crystal structures (2YXJ for Bcl-xL, 2JK7 for XIAP) as input for the reference ligand method in the cavity mapping program “rbcavity”, with a radius of 10 Å. Docking calculations were carried out with the standard 50 runs for each ligand. All other rDock parameters were set to their default values. The final

docked conformation was taken to be the one with the lowest score, and this score was used in ranking the compounds.

PLANTS usage

PLANTS (Protein-Ligand ANT System) version1.2 is a stochastic docking method based on ant colony optimization (87). Docking sites at the protein interface were defined using the ligand-bound crystal structures (2YXJ for Bcl-xL, 2JK7 for XIAP), with a radius of 10 Å. Docking calculations were carried out using the "rigid_all" flag, to carry out multiconformer docking for each ligand. The same ligand conformers used to run DARC (those generated using OMEGA (44-46)) were also used to run PLANTS. CHEMPLP was used as the scoring function, and all other PLANTS parameters were set to their default values. The final docked conformation was taken to be the one with the lowest score, and this score was used in ranking the compounds.

ROCS usage

ROCS (88-90) ligand-based screening uses Gaussian overlap to evaluate 3-dimensional similarity of a given ligand to a template, in terms of both shape and "color" (overlap of chemical features such as rings and polar groups). In this case we used the bound co-crystallized ligands in the holo structures (2YXJ for Bcl-xL, 2JK7 for XIAP) as the templates. The same ligand conformers used to run DARC (those generated using OMEGA (44-46)) were also used to run ROCS. The TanimotoCombo score was used to score and rank the compounds. All other ROCS parameters were set to their default values.

Compound library for Mcl-1 screen

Starting from the ZINC database (43), we first downloaded the "drugs now" subset of compounds: this subset contains about 6 million compounds with molecular weight between 150 and 500 Da, xlogP $\leq$ 5, no more than 5 Hbond donors, no more than 10 Hbond acceptors, and at least one

commercial vendor that offers the compound for immediate delivery. We then filtered this library for diversity, by removing any compound with chemical similarity to any other compound in the library (2D Fingerprint Tanimoto score (80) > 0.7). Upon completion of this filtering step, 62,442 compounds remained.

Finally, we built conformers for each of these compounds using OMEGA (44-46), with a maximum of 300 conformers per compound and all other options set to default.

Expression and purification of Mcl-1

A gene encoding residues 172-327 of human Mcl-1 (this segment lacks the PEST and transmembrane regions) was incorporated into a pSV282 vector as a construct containing N-terminal TEV-cleavable hexahistidine and maltose binding protein (MBP) fusion tags. The protein was transformed into *E. coli* Rosetta2(2DE3) pLysS cells and grown in LB media (Fisher Scientific) with 100 μg/mL ampicillin at 37°C to an OD_{600} of 0.8. Expression was then induced with 1 mM IPTG for 20 hours at 15°C. The cells were pelleted by centrifugation then resuspended in lysis buffer (20 mM HEPES pH 7.55, 300 mM NaCl) and lysed by sonication. The protein was purified by nickel affinity chromatography, then the fusion tags were removed using hexahistidine-tagged tobacco etch virus protease (TEV) (provided by Dr. Philip Gao, University of Kansas) at a 1:10 (m/m) ratio. The digestion mixture was incubated at room temperature overnight in dialysis tubing against reaction buffer (20 mM HEPES pH 8.0, 300 mM NaCl). The His-tagged TEV protease and His-tagged MBP were then removed by repassaging the reaction mixture over a Ni-column and collecting the flow-through fraction (corresponding to untagged Mcl-1). The untagged Mcl-1 was further purified using gel filtration on HiLoad 16/60 Superdex 75 prep grade column in 20 mM HEPES buffer, pH 7.55, 1 mM DTT, directly followed by ion exchange chromatography on HiTrap SP FF column (GE Healthcare). Final samples were buffer-exchanged into 20 mM HEPES buffer, pH 7.55, 1 mM DTT. Mcl-1 produced in this manner was used for all assays described here, including pilot fluorescence polarization studies. For the final

fluorescence polarization experiments described in this study, we used an Mcl-1 construct that contained 17 additional residues at the N-terminus and 14 additional residues at the C-terminus.

^{15}N-labeled Mcl-1 was expressed via auto-induction in media supplemented with $^{15}NH_4Cl$ as described elsewhere (261), then purified as described above.

TEV protease used above was also produced in *E. coli* Rosetta2(2DE3) pLysS cells, using the same expression protocol. The (His-tagged) protein was purified by nickel affinity chromatography then stored in 50% (v/v) glycerol, 50 mM Tris buffer, pH 8.0, 300 mM NaCl, 500 mM imidazole at -20 ^{o}C.

Fluorescence polarization competition assay

All fluorescence polarization experiments used a carboxyfluorescein-labeled 26-mer Noxa peptide (FITC-GTMENLSRRLKVTGDLFDIMSGQTDV-CONH2) in 96-well black polystyrene flat bottom plates (Whatman), and were carried out using a Synergy Neo plate reader (Biotek). Mcl-1 samples were centrifuged prior to the assay, to remove any aggregated protein; the protein concentration in the supernatant was then estimated via Bradford assay, and this supernatant was used for fluorescence polarization experiments. All assays were carried out in 20 mM HEPES (pH 7.55), 1 mM DTT, 5% DMSO (v/v), and 0.1% CHAPS (w/v). In all cases plates were incubated at 26°C for 1 hour prior to reading. All measurements were carried out in triplicate.

To measure the apparent binding affinity of the labeled Noxa peptide with Mcl-1, a fixed concentration of Noxa peptide (50 nM) was incubated with increasing concentrations of purified Mcl-1 (from 0.05 nM to 25 μM) (**Figure S6a**). Fluorescence polarization as a function of Mcl-1 concentration was fit using GraphPad Prism, with the Hill coefficient fixed at 1, leading to a Kd estimate of 79 nM. The fluorescence polarization at the observed endpoints of this titration (i.e. peptide alone and peptide with saturating Mcl-1) were taken as 100% and 0% inhibition in the competition experiments below.

Candidate inhibitors from the DARC screen (compounds M0-M20) and M0 analogs (compounds D1-D27) were tested by mixing the compound with a pre-incubated mixture of Noxa peptide (50 nM) and

Mcl-1 (150 nM) at room temperature, at fifteen different compound concentrations ranging from 15 nM to 250 μM. Data were fit to a single-site competition model to determine IC50 using GraphPad Prism, fixing the Hill coefficient at 1 and the top and bottom values were fixed at 0 and 100% inhibition (leaving a single free parameter in each fit). Given the known experimental conditions and the binding affinity described above for the labeled Noxa peptide, the Ki was then computed from the IC50 using the method of Nikolovska-Coleska et al (262).

Bio-layer interferometry

All bio-layer interferometry (BLI) experiments were carried out using the Octet RED96 system (ForteBIO).

The human Mcl-1 construct containing N-terminal hexahistidine and maltose binding protein (MBP) fusion tags (described above) was loaded onto Ni-NTA sensor tips. The loading step was carried out using 1 μM protein in 50 mM Tris, 1 mM DTT, 0.5 mM EDTA pH 7.55, and was allowed to continue until the BLI signal was shifted by 2 nm relative to the baseline. As a negative control, we used TEV protease to first remove Mcl-1 from this construct, then loaded the resulting His-tagged MBP construct. To generate "blank" data, the program described below was also carried out using empty sensor tips and the resulting blank data were subtracted from the data collected with each protein of interest.

After loading the protein of interest, we reset the baseline in "analysis buffer" (50 mM Tris, 0.5 mM EDTA pH 7.55). The sensor tips were then introduced into wells containing the compound of interest at a concentration of 50 μM in the same "analysis buffer" (with 0.5% DMSO), and separate tips were also introduced into wells that lacked the compound. Association-phase data were collected for 300 sec, then the tips were then transferred to wells containing "analysis buffer" (with no compound), and dissociation-phase data were collected for 300 sec. All sample volumes were 250 μL, in black 96-well plates, with a shake speed of 1000 rpm. Data were collected every 0.2 seconds.

After collecting data, the sensor tips were washed with 50 mM Tris buffer, then regenerated with 10 mM Glycine at pH 1.7, then washed with ForteBIO "kinetic buffer" then activated with 10 mM $NiCl_2$.

Data were collected using the ForteBIO "Data Acquisition" software (version 8.2), and analyzed using the ForteBIO "Data Analysis" software (version 8.2). All data were analyzed by double referencing (by subtracting data from sensor tips with and without protein, and subtracting data from wells with and without the compound of interest.

Surface plasmon resonance

Human Mcl-1 was coupled to CM5 sensor chip (Biacore) at approximately 9,700 response units, with one flow cell left uncoupled as a reference cell. Mcl-1 (in 1 mM DTT, 20 mM HEPES pH 7.55) was diluted into immobilization buffer (10 mM sodium acetate pH 5.0) for this step.

Samples of M0 were prepared by carefully matching to the sample/running buffer (1 mM DTT, 20 mM HEPES pH 7.55, 5% DMSO (v/v), 0.3% TritonX-100 (v/v)). Careful matching of buffer conditions mitigated potential bulk solvent effects upon analyte injection. Data were collected using a Biacore 3000, using a flow rate of 50 µL/min, and all sample injections (250 µL) were preceded and followed by buffer carry over injections (25 µL). Data points were collected at a frequency of 1 Hz.

NMR studies

NMR studies were carried out using 300 µM of ^{15}N-labeled Mcl-1 in 20 mM Na_2HPO_4 pH 7, with 10 mM deuterated TCEP, in 90%(H_2O)/10%(D_2O). Compound M0 was added as a 10 mM deuterated stock in 100% DMSO-d_6, in three sequential additions. This led to a final concentration of 100 µM M0 and 1.2% DMSO-d_6.

All 2D-1H-15N-HSQC spectra were acquired on a BRUKER AVANCE III 600 MHz NMR instrument with z-gradient, at 25°C. Spectra were recorded with 1024 complex points in the ^{1}H (t2, acquisition) dimension and 200 complex points in the ^{15}N (t1, indirect) dimension with sweep

widths of 9615.39 Hz and 1944.90 Hz in the ^{1}H and ^{15}N dimensions. 48 scans were used per t1 increment. NMR spectra were processed using NMRPipe (263).

MTT-based cytotoxicity assay

DU-145 cell viability was determined by the MTT-based assay using WST-8 (GenScript, Piscataway, NJ). All experiments were carried out in triplicate. DU-145 cells (3500 cells/well) were plated in 96-well culture plates, and diluted compound D2 and/or ABT-263 then were added. Four days later, when control wells with no compound present reached confluence, WST-8 was added to each well and incubated for 2 hours at 37°C. Absorbance was measured with Synergy H4 plate reader (Biotek, Winooski, VT) at 450 nm. The results are expressed as the percentage of absorbance of treated wells versus that of the control in which no compound was added.

Synthesis of M0 derivatives : overview

All air and moisture-sensitive reactions were carried out in flame- or oven-dried glassware under argon atmosphere using standard gastight syringes, cannula, and septa. Stirring was achieved with oven-dried magnetic stir bars. Flash column chromatography was performed with SiO_2 from Sorbent Technology (30930M-25, Silica Gel 60 A, 40-63 mm) or by using an automated chromatography instrument with an appropriately sized column. Thin layer chromatography was performed on silica gel w/UV254 plates (1624126, sorbent technologies). ^{1}H and ^{13}C NMR spectra were recorded on instruments operating at 400 or 500 MHz and 100 or 126 MHz respectively. High-resolution mass spectrometry (HRMS) spectra were obtained on an ESI-TOF mass spectrometer. The analytical method utilized a Waters Acquity BEH C18 column (2.1 x 50 mm, 1.7 mm) eluting with a linear gradient of 95% water (modified to pH 9.8 through addition of NH_4OH) to 100% CH_3CN at 0.6 mL min^{-1} flow rate where purity was determined using UV peak area at 214 nm.

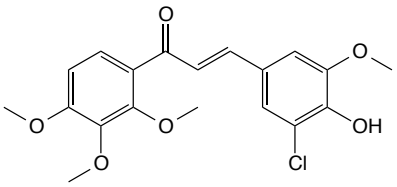
Synthesis of M0 derivatives : chalcone formation

To a round-bottomed flask equipped with a stir bar was added an aldehyde (0.819 mmol), absolute ethanol (10 mL) and a ketone (0.819 mmol). To the reaction mixture was added an aqueous solution of potassium hydroxide (20%, 10 mL), and the reaction was allowed to stir overnight at room temperature. The reaction mixture was then poured into water (100 mL), and after neutralization with hydrochloric acid (5%) to precipitate the chalcone. The product was filtered, washed with cool water. The product was either recrystallized from ethanol, or purified via normal phase (methanol/dichloromethane) and/or reverse phase (water, acetonitrile) chromatography. The generalized reaction scheme is shown in **Figure S9a**.

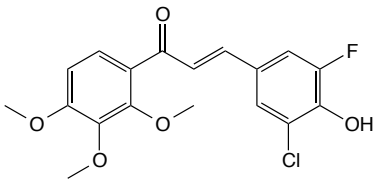
Synthesis of M0 derivatives : flavone formation

To a round-bottomed flask equipped with a stir bar was added the corresponding 2'-hydroxychalcone (numbering as in **Table 1**) (1 mmol) in DMSO (10 mL), and then I_2 (0.01 mmol) at room temperature. The reaction mixture was then refluxed for 2 h. The reaction mixture was cooled to room temperature, and then water was added. The product was extracted with ethyl acetate three times from the aqueous phase. The combined organic layers were washed twice with 5% $NaHSO_4$ aqueous solution, twice with water, once with brine, dried over sodium sulfate and then concentrated *en vacuo*. The crude product was purified via normal phase (methanol/dichloromethane, ethyl acetate/hexanes) and/or reverse phase (water, acetonitrile) chromatography. The generalized reaction scheme is shown in **Figure S9b**.

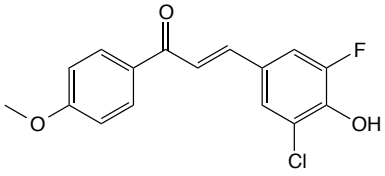
Synthesis of M0 derivatives : specific compounds



D1: (*E*)-3-(3-chloro-4-hydroxy-5-methoxyphenyl)-1-(2,3,4-trimethoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 10.23 – 9.93 (m, 1H), 7.47 – 7.29 (m, 5H), 6.93 (d, J = 8.8 Hz, 1H), 3.89 (s, 3H), 3.87 (s, 3H), 3.83 (s, 3H), 3.79 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 190.4, 156.3, 152.5, 148.5, 145.0, 142.3, 141.6, 126.5, 126.2, 125.1, 124.8, 122.8, 120.1, 109.9, 107.7, 61.6, 60.2, 56.3, 56.0. HRMS calculated for $C_{19}H_{20}ClO_6$ $[M+H]^+$: 378.0870, found 379.0955. Internal barcode: SCAP108855.

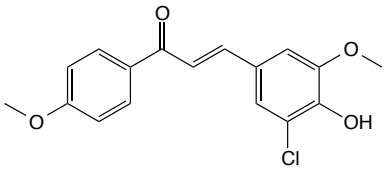


D2: (*E*)-3-(3-chloro-5-fluoro-4-hydroxyphenyl)-1-(2,3,4-trimethoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.68 (d, J = 8.8 Hz, 3H), 7.36 (d, J = 8.8 Hz, 2H), 6.93 (d, J = 8.8 Hz, 2H), 3.88 (s, 3H), 3.82 (s, 3H), 3.79 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 185.1, 180.1, 165.2, 156.5, 135.6, 134.7, 127.7, 126.1, 125.0, 119.2, 114.2, 112.4, 107.7, 101.8, 96.7, 61.7, 60.5, 56.1. HRMS calculated for $C_{18}H_{17}ClFO_5$ [M+H]+: 366.0670, found 367.0748. Internal barcode: SCAP108958.

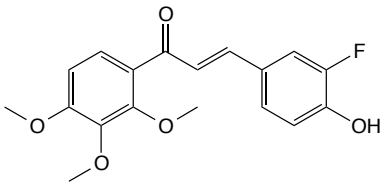


D3: (*E*)-3-(3-chloro-5-fluoro-4-hydroxyphenyl)-1-(4-methoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.19 (d, J = 8.9 Hz, 2H), 8.01 – 7.78 (m, 3H), 7.60 (d, J = 15.4 Hz, 1H), 7.09 (d, J = 8.9 Hz, 2H), 3.87 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 187.0, 163.2, 152.8, 150.9, 143.4, 141.1, 130.9,

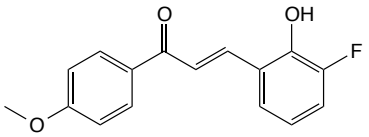
130.4, 126.3, 122.3, 121.4, 114.9, 113.9, 55.5. HRMS calculated for $C_{16}H_{13}ClO_3$ $[M+H]^+$: 306.0459, found 307.0529. Internal barcode: SCAP108760.



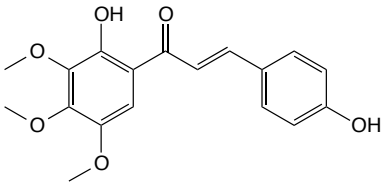
D4: (*E*)-3-(3-chloro-4-hydroxy-5-methoxyphenyl)-1-(4-methoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 10.04 (s, 1H), 8.19 (d, J = 8.9 Hz, 2H), 8.01 – 7.92 (m, 1H), 7.87 (d, J = 15.5 Hz, 1H), 7.62 (d, J = 15.5 Hz, 1H), 7.58 (d, J = 1.8 Hz, 1H), 7.50 (d, J = 1.9 Hz, 1H), 7.09 (d, J = 8.9 Hz, 1H), 7.05 (d, J = 8.9 Hz, 2H), 3.93 (s, 3H), 3.88 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 197.9, 187.1, 163.1, 162.3, 148.7, 145.0, 142.6, 130.9, 126.5, 123.1, 120.2, 113.9, 110.5, 56.4, 55.5. HRMS calculated for $C_{17}H_{16}ClO_4$ $[M+H]^+$: 318.0659, found 319.0733. Internal barcode: SCAP108835.



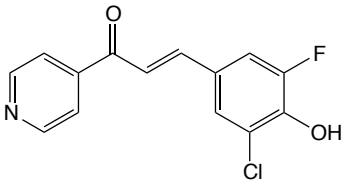
D5: (*E*)-3-(3-fluoro-4-hydroxyphenyl)-1-(2,3,4-trimethoxyphenyl)prop-2-en-1-one. ^{1}H NMR (400 MHz, DMSO-d_6) δ 10.51 (s, 1H), 7.63 (dd, J = 12.5, 2.1 Hz, 1H), 7.46 (d, J = 15.7 Hz, 1H), 7.43 – 7.39 (m, 1H), 7.38 – 7.25 (m, 2H), 7.00 (dd, J = 9.1, 8.3 Hz, 1H), 6.94 (d, J = 8.8 Hz, 1H), 3.88 (s, 3H), 3.84 (s, 3H), 3.79 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 190.1, 160.0, 156.3, 152.6, 147.4, 143.3, 141.6, 130.8, 126.8, 126.7, 124.9, 123.2, 118.7, 115.9, 107.8, 61.7, 60.8, 56.1. HRMS calculated for $C_{18}H_{18}FO_5$ $[M+H]^+$: 332.1060, found 333.1142. Internal barcode: SCAP108889.



D6: (*E*)-3-(3-fluoro-2-hydroxyphenyl)-1-(4-methoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.14 (d, J = 8.9 Hz, 1H), 7.86 – 7.51 (m, 4H), 7.45 – 7.35 (m, 1H), 7.07 (d, J = 8.9 Hz, 2H), 6.85 (t, J = 8.8 Hz, 1H), 3.87 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 186.8, 162.8, 162.3, 145.7, 143.3, 130.9, 130.6, 127.4, 125.2, 118.4, 115.3, 115.2, 113.8, 55.5. HRMS calculated for $C_{16}H_{14}FO_3$ $[M+H]^+$: 272.0849, found 273.0921. Internal barcode: SCAP108974.

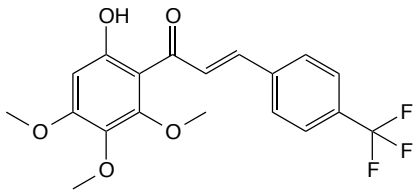


D7: (*E*)-1-(2-hydroxy-3,4,5-trimethoxyphenyl)-3-(4-hydroxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.41 (s, 1H), 10.12 (s, 1H), 7.62 – 7.55 (m, 3H), 7.42 (d, J = 15.7 Hz, 1H), 6.84 (d, J = 8.6 Hz, 2H), 6.38 (s, 1H), 3.84 (s, 3H), 3.83 (s, 3H), 3.70 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 192.4, 160.1, 158.1, 158.0, 153.2, 144.1, 134.6, 130.6, 125.6, 123.4, 116.0, 110.1, 96.4, 61.5, 60.6, 56.0. HRMS calculated for $C_{18}H_{19}O_6$ $[M+H]^+$: 330.1103, found 331.1180. Internal barcode: SCAP108798.

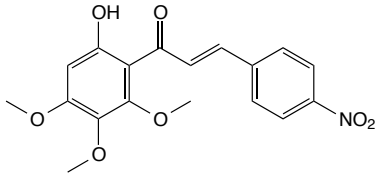


D8: (*E*)-3-(3-chloro-5-fluoro-4-hydroxyphenyl)-1-(pyridin-4-yl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 11.22 (s, 1H), 8.90 – 8.81 (m, 2H), 8.06 – 7.97 (m, 2H), 7.87 (s, 2H), 7.84 – 7.65 (m, 2H). ^{13}C NMR (126 MHz, DMSO) δ 188.4, 153.0, 151.1, 150.7, 144.0, 143.6, 128.3, 127.2, 122.5, 122.5,

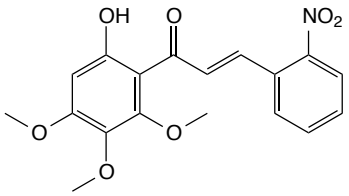
121.6, 119.9, 115.2. HRMS calculated for $C_{14}H_{10}FNO_2$ $[M+H]^+$: 277.0306, found 278.0380. Internal barcode: SCAP108810.



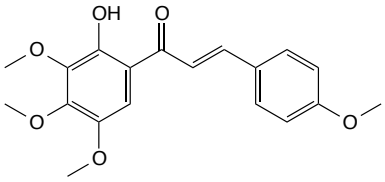
D9: (*E*)-1-(6-hydroxy-2,3,4-trimethoxyphenyl)-3-(4-(trifluoromethyl)phenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.06 (s, 1H), 7.96 (d, J = 8.0 Hz, 2H), 7.81 (d, J = 8.1 Hz, 2H), 7.66 (d, J = 4.1 Hz, 2H), 6.41 (s, 1H), 3.85 (d, J = 4.0 Hz, 6H), 3.71 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 179.8, 162.0, 159.0, 157.8, 153.2, 147.3, 140.5, 129.9, 128.9, 125.8, 122.9, 120.4, 105.9, 96.3, 61.5, 60.6, 56.0. HRMS calculated for $C_{19}H_{18}F_3O_5$ $[M+H]^+$: 382.1028, found 383.1108. Internal barcode: SKCM104238.



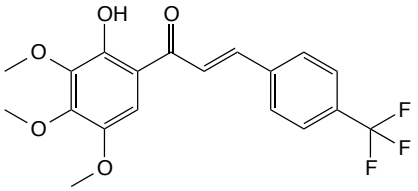
D10: (*E*)-1-(6-hydroxy-2,3,4-trimethoxyphenyl)-3-(4-nitrophenyl)prop-2-en-1-one. ^{1}H NMR (400 MHz, DMSO-d_6) δ 12.07 (s, 1H), 8.28 (d, J = 8.8 Hz, 1H), 8.02 (d, J = 8.8 Hz, 1H), 7.70 (d, J = 8.0 Hz, 2H), 6.42 (s, 2H), 3.86 (d, J = 1.6 Hz, 6H), 3.72 (s, 3H). ^{13}C NMR (101 MHz, DMSO) δ 192.1, 158.5, 157.9, 153.2, 147.7, 140.9, 139.5, 134.9, 131.2, 129.2, 123.8, 110.1, 96.2, 61.6, 60.7, 56.1. HRMS calculated for $C_{18}H_{18}NO_7$ $[M+H]^+$: 359.1005, found 360.1078. Internal barcode: SKCM104164.



D11: (*E*)-1-(6-hydroxy-2,3,4-trimethoxyphenyl)-3-(2-nitrophenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.01 (s, 1H), 8.10 (dd, J = 8.2, 1.2 Hz, 1H), 7.96 (dd, J = 7.9, 1.3 Hz, 1H), 7.89 – 7.79 (m, 2H), 7.70 (ddd, J = 8.5, 7.4, 1.4 Hz, 1H), 7.52 (d, J = 15.6 Hz, 1H), 6.41 (s, 1H), 3.85 (d, J = 0.8 Hz, 6H), 3.70 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 192.1, 158.6, 158.0, 153.2, 148.5, 137.2, 134.6, 134.0, 131.4, 130.9, 129.8, 129.0, 124.8, 110.1, 96.3, 61.6, 60.7, 56.1. HRMS calculated for $C_{18}H_{18}NO_7$ $[M+H]^+$: 359.1005, found 360.1087. Internal barcode: SKCM101933.

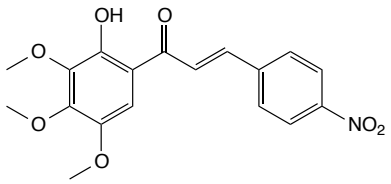


D12: (*E*)-1-(2-hydroxy-3,4,5-trimethoxyphenyl)-3-(4-methoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.32 (s, 1H), 7.72 (d, J = 8.7 Hz, 2H), 7.67 – 7.41 (m, 2H), 7.04 (d, J = 8.9 Hz, 2H), 6.40 (s, 1H), 3.89 – 3.81 (m, 10H), 3.72 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 192.4, 161.3, 158.0, 153.2, 143.4, 134.6, 130.4, 127.7, 127.2, 124.5, 114.6, 110.2, 96.4, 61.5, 60.6, 56.0, 55.4. HRMS calculated for $C_{19}H_{21}O_6$ $[M+H]^+$: 344.1259, found 345.1336. Internal barcode: SCAP108811.

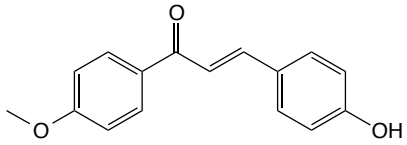


D13: (*E*)-1-(2-hydroxy-3,4,5-trimethoxyphenyl)-3-(4-(trifluoromethyl)phenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.08 (s, 1H), 7.96 (d, J = 7.9 Hz, 2H), 7.80 (d, J = 8.3 Hz, 2H), 7.72 – 7.59 (m,

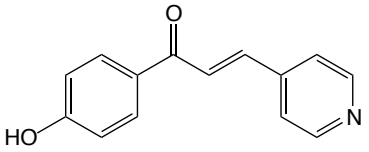
2H), 6.41 (s, 1H), 3.85 (d, J = 3.7 Hz, 6H), 3.71 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 192.4, 158.4, 157.9, 153.2, 140.5, 138.7, 134.6, 129.9, 129.0, 125.8, 125.1, 122.9, 110.3, 96.4, 61.6, 60.6, 56.1. HRMS calculated for $C_{19}H_{18}F_3O_5$ $[M+H]^+$: 382.1028, found 383.1111. Internal barcode: SCAP108799.



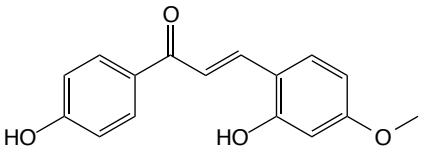
D14: (*E*)-1-(2-hydroxy-3,4,5-trimethoxyphenyl)-3-(4-nitrophenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.02 (s, 1H), 8.10 (dd, J = 8.2, 1.2 Hz, 1H), 7.96 (dd, J = 7.9, 1.3 Hz, 1H), 7.90 – 7.78 (m, 2H), 7.70 (ddd, J = 8.5, 7.5, 1.4 Hz, 1H), 7.53 (d, J = 15.6 Hz, 1H), 6.41 (s, 1H), 3.86 (d, J = 1.1 Hz, 6H), 3.70 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 192.1, 158.6, 153.2, 148.5, 137.2, 134.6, 134.0, 131.4, 130.9, 129.8, 124.8, 110.1, 96.3, 61.6, 60.7, 56.1. HRMS calculated for $C_{18}H_{18}NO_7$ $[M+H]^+$: 359.1005, found 360.1086. Internal barcode: SCAP108785.



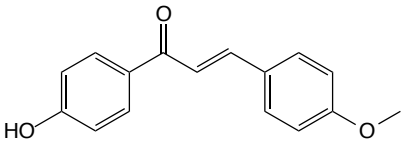
D15: (*E*)-3-(4-hydroxyphenyl)-1-(4-methoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 10.07 (s, 1H), 8.14 (d, J = 8.9 Hz, 2H), 7.84 – 7.53 (m, 4H), 7.08 (d, J = 8.9 Hz, 2H), 6.84 (d, J = 8.6 Hz, 2H), 3.87 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 187.2, 162.9, 162.3, 159.9, 143.6, 130.8, 130.7, 125.9, 118.4, 115.7, 113.9, 55.5. HRMS calculated for $C_{16}H_{15}O_3$ $[M+H]^+$: 254.0943, found 255.1014. Internal barcode: SCAP108837.



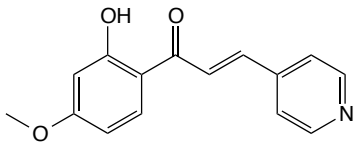
D16: (*E*)-1-(4-hydroxyphenyl)-3-(pyridin-4-yl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 10.21 (s, 1H), 8.98 – 8.70 (m, 2H), 8.11 – 7.94 (m, 2H), 7.87 – 7.55 (m, 6H), 6.85 (d, J = 8.6 Hz, 2H). ^{13}C NMR (126 MHz, DMSO) δ 150.7, 146.2, 142.9, 131.4, 121.5, 117.9, 115.9, 104.6, 102.0, 99.5. HRMS calculated for $C_{14}H_{12}ON_2$ $[M+H]^+$: 225.0789, found 226.0853. Internal barcode: SCAP108817.



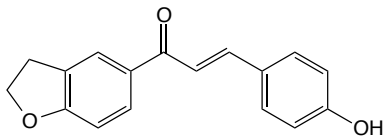
D17: (*E*)-3-(2-hydroxy-4-methoxyphenyl)-1-(4-hydroxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 13.67 (s, 1H), 10.21 (s, 1H), 8.83 (d, J = 5.4 Hz, 2H), 7.97 (d, J = 5.4 Hz, 1H), 7.78 (dt, J = 8.6, 4.6 Hz, 4H), 6.90 – 6.80 (m, 2H), 3.85 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 184.3, 156.2, 150.9, 147.9, 145.6, 138.6, 135.2, 131.9, 120.0, 118.1, 115.9, 100.4, 99.5, 58.0. HRMS calculated for $C_{16}H_{15}O_4$ $[M+H]^+$: 270.0892, found 271.0963. Internal barcode: SCAP108814.



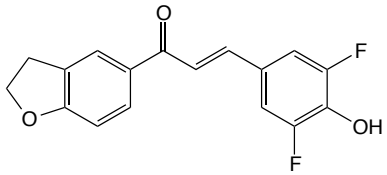
D18: (*E*)-1-(4-hydroxyphenyl)-3-(4-methoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 10.07 (s, 1H), 8.14 (d, J = 8.9 Hz, 2H), 7.81 – 7.55 (m, 4H), 7.21 – 6.99 (m, 2H), 6.93 – 6.57 (m, 2H), 3.87 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 187.2, 162.9, 159.9, 143.6, 130.8, 130.7, 130.6, 125.9, 118.4, 115.7, 113.9, 55.5. HRMS calculated for $C_{16}H_{15}O_3$ $[M+H]^+$: 254.0942, found 255.1007. Internal barcode: SCAP108927.



D19: (*E*)-1-(2-hydroxy-4-methoxyphenyl)-3-(pyridin-4-yl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 13.20 (s, 1H), 8.84 – 8.47 (m, 2H), 7.73 (d, J = 9.4 Hz, 1H), 7.56 (ddd, J = 4.5, 1.7, 0.7 Hz, 2H), 6.93 – 6.58 (m, 3H), 5.75 (dd, J = 12.4, 3.2 Hz, 1H), 3.84 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 189.3, 165.7, 162.6, 149.9, 147.6, 128.0, 122.6, 121.0, 114.4, 110.2, 107.7, 101.1, 55.9. HRMS calculated for $C_{15}H_{14}NO_3$ $[M+H]^+$: 255.0895, found 256.0969. Internal barcode: SCAP109011.

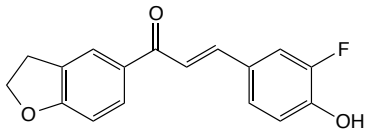


D20: (*E*)-1-(2,3-dihydrobenzofuran-5-yl)-3-(4-hydroxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 10.05 (s, 1H), 8.07 (q, J = 1.3 Hz, 1H), 8.04 – 7.92 (m, 1H), 7.82 – 7.57 (m, 4H), 6.90 (d, J = 8.4 Hz, 1H), 6.83 (d, J = 8.6 Hz, 2H), 4.66 (t, J = 8.7 Hz, 2H), 3.27 (dd, J = 9.3, 8.2 Hz, 2H). ^{13}C NMR (126 MHz, DMSO) δ 187.0, 163.7, 159.8, 143.3, 131.0, 130.8, 130.1, 128.2, 125.9, 125.8, 118.5, 115.7, 108.8, 72.1, 40.4. HRMS calculated for $C_{17}H_{15}O_3$ $[M+H]^+$: 266.0943, found 267.1014. Internal barcode: SCAP108842.

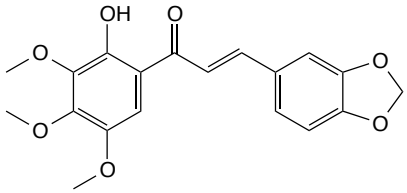


D21: (*E*)-3-(3,5-difluoro-4-hydroxyphenyl)-1-(2,3-dihydrobenzofuran-5-yl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 10.82 (s, 1H), 8.11 (q, J = 1.3 Hz, 1H), 8.04 (dd, J = 8.5, 2.0 Hz, 1H), 7.88 (d, J = 15.5 Hz, 1H), 7.69 (dd, J = 8.3, 1.7 Hz, 2H), 7.58 (d, J = 15.4 Hz, 1H), 6.91 (d, J = 8.4 Hz, 1H), 4.67 (t, J = 8.7 Hz, 2H), 3.31 – 3.25 (m, 2H). ^{13}C NMR (126 MHz, DMSO) δ 186.7, 164.0, 162.3, 153.1, 151.2,

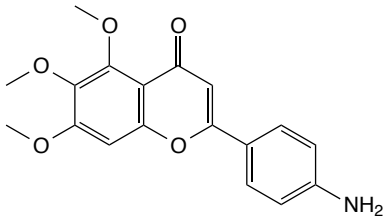
141.2, 141.2, 130.7, 130.4, 128.3, 126.0, 121.5, 112.4, 112.2, 108.8, 72.2, 28.3. HRMS calculated for $C_{17}H_{13}F_2O_3$ $[M+H]^+$: 302.0755, found 303.0828. Internal barcode: SKCM101972.



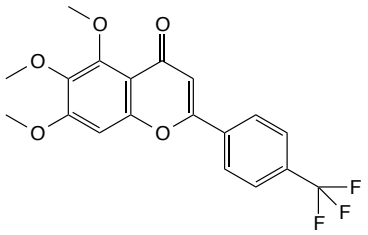
D22: (*E*)-1-(2,3-dihydrobenzofuran-5-yl)-3-(3-fluoro-4-hydroxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 10.49 (s, 1H), 8.09 (q, *J* = 1.3 Hz, 1H), 8.03 (dd, *J* = 8.4, 1.9 Hz, 1H), 7.85 (dd, *J* = 12.7, 2.1 Hz, 1H), 7.80 (d, *J* = 15.4 Hz, 1H), 7.64 – 7.57 (m, 1H), 7.51 – 7.44 (m, 1H), 6.99 (dd, *J* = 9.1, 8.3 Hz, 1H), 6.90 (d, *J* = 8.4 Hz, 1H), 4.66 (t, *J* = 8.7 Hz, 2H), 3.31 – 3.22 (m, 2H). ^{13}C NMR (126 MHz, DMSO) δ 186.9, 163.9, 162.3, 152.1, 150.2, 147.2, 142.2,130.9, 130.3, 128.2, 126.8, 120.1, 117.7, 115.6, 108.8, 72.1, 28.4. HRMS calculated for $C_{17}H_{14}FO_3$ $[M+H]^+$: 284.0849, found 285.0921. Internal barcode: SKCM104208.



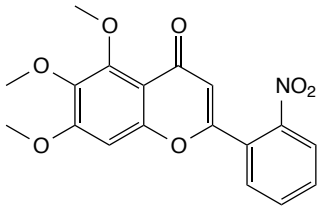
D23: (*E*)-3-(benzo[d][1,3]dioxol-5-yl)-1-(2-hydroxy-3,4,5-trimethoxyphenyl)prop-2-en-1-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.13 (s, 1H), 7.54 (d, *J* = 15.7 Hz, 1H), 7.43 (s, 1H), 7.37 (d, *J* = 1.7 Hz, 1H), 7.24 (dd, *J* = 8.3, 1.7 Hz, 1H), 6.99 (d, *J* = 8.0 Hz, 1H), 6.38 (s, 1H), 6.11 (s, 2H), 3.84 (s, 3H), 3.83 (s, 3H), 3.70 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 193.3, 157.9, 155.5, 153.1, 144.9, 130.7, 125.2, 121.0, 117.0, 113.6, 110.4, 108.7, 106.8, 106.0, 101.7, 81.8, 61.5, 60.6, 56.0. HRMS calculated for $C_{19}H_{19}O_7$ $[M+H]^+$: 358.1053, found 359.1132. Internal barcode: SCAP108867.



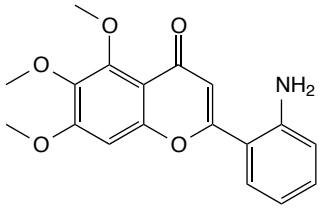
D24: 2-(4-aminophenyl)-5,6,7-trimethoxy-4H-chromen-4-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.74 (d, J = 8.7 Hz, 2H), 7.16 (s, 1H), 6.66 (d, J = 8.8 Hz, 2H), 6.49 (s, 1H), 5.95 (s, 2H), 3.94 (s, 3H), 3.79 (s, 3H), 3.76 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 195.1, 171.0, 158.9, 152.2, 131.2, 127.5, 123.1, 119.3, 113.4, 111.9, 105.9, 103.6, 102.6, 99.5, 97.2, 61.8, 60.9, 56.4. HRMS calculated for $C_{18}H_{18}NO_5$ $[M+H]^+$: 327.1107, found 328.1186. Internal barcode: SKCM101975.



D25: 5,6,7-trimethoxy-2-(4-(trifluoromethyl)phenyl)-4H-chromen-4-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.62 (dd, J = 12.4, 2.1 Hz, 1H), 7.34 (d, J = 8.7 Hz, 1H), 7.28 (d, J = 15.8 Hz, 1H), 6.99 (dd, J = 9.1, 8.3 Hz, 1H), 6.93 (d, J = 8.8 Hz, 1H), 3.87 (s, 3H), 3.83 (s, 3H), 3.79 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 190.1, 156.4, 152.7, 152.0, 150.1, 147.3, 142.0, 141.6, 126.4, 126.0, 124.9, 124.6, 117.9, 115.8, 115.7, 107.8, 61.7, 60.5, 56.0. HRMS calculated for $C_{19}H_{16}F_3O_5$ $[M+H]^+$: 380.0872, found 381.0952. Internal barcode: SKCM102002.



D26: 5,6,7-trimethoxy-2-(2-nitrophenyl)-4H-chromen-4-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.22 (dt, J = 8.1, 0.8 Hz, 1H), 7.95 – 7.92 (m, 1H), 7.89 – 7.86 (m, 1H), 6.91 (s, 1H), 6.54 (s, 1H), 3.90 (s, 3H), 3.83 (s, 3H), 3.78 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 175.2, 159.5, 157.9, 151.7, 144.8, 141.6, 140.0, 134.1, 132.6, 132.4, 131.5, 125.4, 124.9, 111.5, 96.8, 61.9, 61.0, 56.5. HRMS calculated for $C_{18}H_{16}NO_7$ $[M+H]^+$: 357.0849, found 358.0930. Internal barcode: SKCM104161.

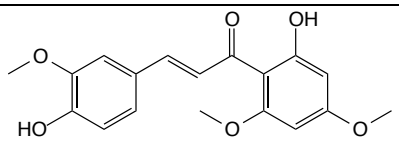
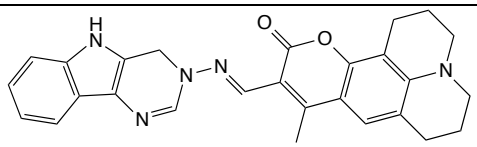
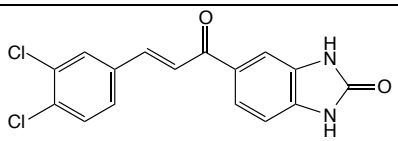
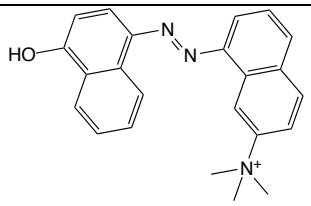
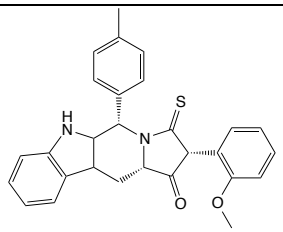
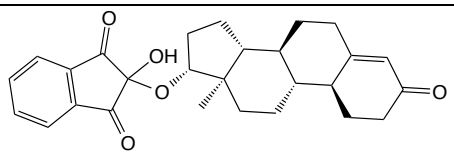
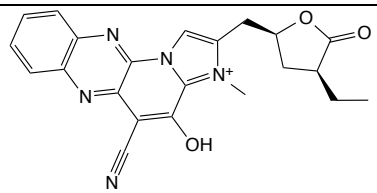


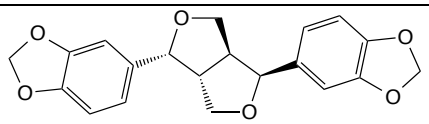
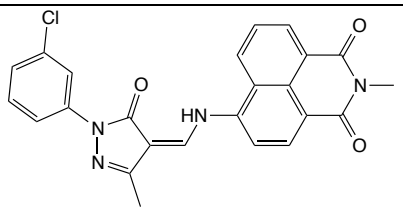
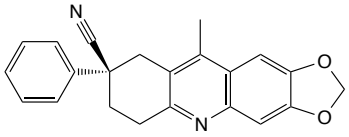
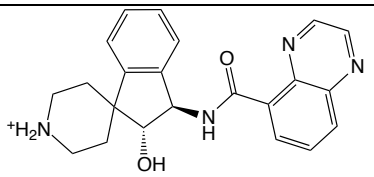
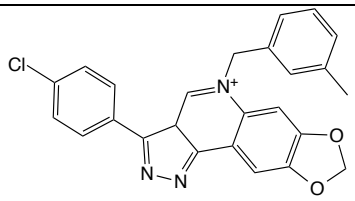
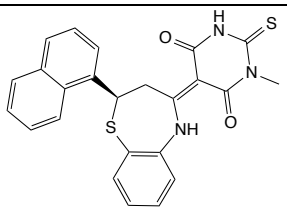
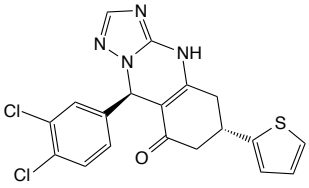
D27: 2-(2-aminophenyl)-5,6,7-trimethoxy-4H-chromen-4-one. ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.39 (dd, J = 7.8, 1.6 Hz, 1H), 7.20 (ddd, J = 8.6, 7.2, 1.6 Hz, 1H), 7.14 (s, 1H), 6.82 (dd, J = 8.3, 1.1 Hz, 1H), 6.65 (ddd, J = 8.0, 7.2, 1.1 Hz, 1H), 6.33 (s, 1H), 5.61 (s, 2H), 3.93 (s, 3H), 3.81 (s, 3H), 3.77 (s, 3H). ^{13}C NMR (126 MHz, DMSO) δ 192.9, 162.3, 157.3, 154.2, 151.5, 150.9, 147.0, 145.5, 140.9, 137.5, 131.6, 129.2, 116.5, 110.2, 97.5, 61.8, 61.0, 56.4. HRMS calculated for $C_{18}H_{18}NO_5$ $[M+H]^+$: 327.1107, found 328.1185. Internal barcode: SKCM104166.

2.3 Supplementary Tables

Protein	Inhibitor-bound PDB	Ligand RMSD after DARC docking (Å)
Clathrin	2XZG	1.35
IL-2	1PW6	2.05
gp41 peptide	2KP8	1.42
FKBP12	1J4I	0.65
ZipA	1S1S	0.56
TNFα	2AZ5	1.95
HIV-1 integrase	3LPU	0.32

Table S1: Protein-ligand complexes used in DARC parameter optimization (c1/c2/c3/c4). This training set was assembled from complexes included in the 2P2I (56) and TIMBAL (78) databases.

Compound	ZINC ID	Chemical structure
M0	04252579	
M1	02098378	
M2	12790730	
M3	04823201	
M4	08783173	
M5	03954116	
M6	02411771	

M7	18258326	
M8	08856242	
M9	12547109	
M10	72151170	
M11	20325906	
M12	49735096	
M13	18197321	

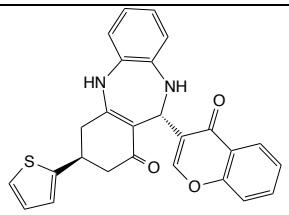
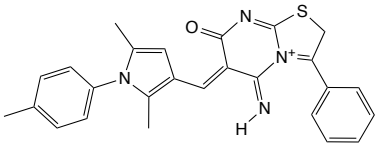
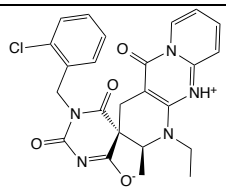
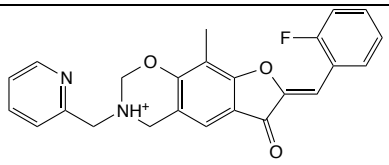
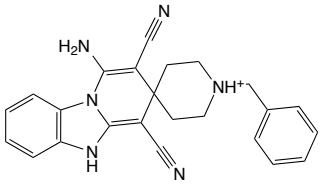
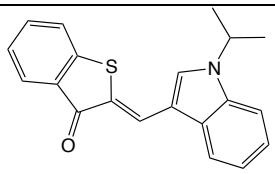
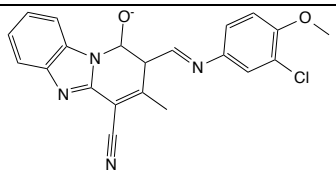
M14	18195186	
M15	57351782	
M16	08771601	
M17	12323480	
M18	04058299	
M19	03656394	
M20	05445823	

Table S2: Initial DARC screening hits tested for Mcl-1 inhibition. The chemical structures and ZINC IDs are presented for each of the initial DARC hits that were purchased and characterized.

2.4 Supplementary Figures

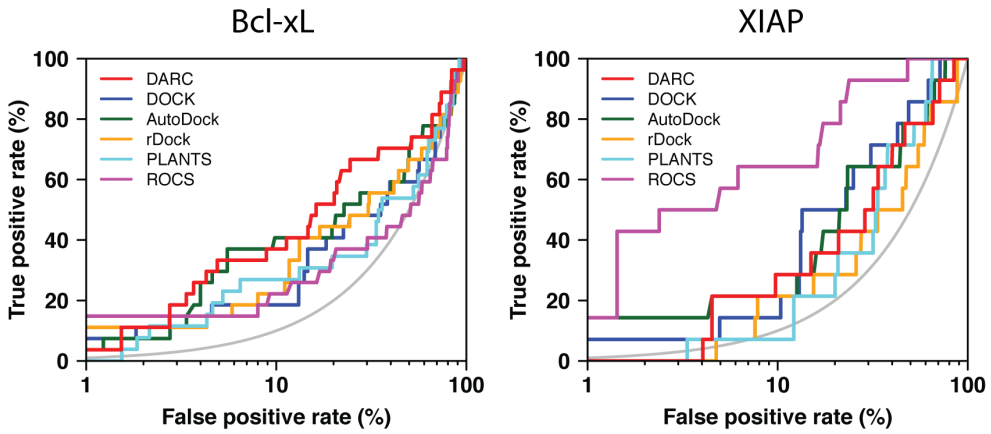


Figure S1: This receiver operating characteristic (ROC) plot compares the performance of various methods (DARC, DOCK, AutoDock, rDock, PLANTS, and ROCS) in a similar benchmark experiment to that described in **Figure 2**. Here, the same active compounds were included but decoy compounds were drawn from the TIMBAL set (78, 79) instead of from the Astex diverse set (35). This allowed the use of 328 decoy compounds in the Bcl-xL screen, and 425 decoy compounds in the XIAP screen. The results are presented on a semi-log plot to highlight the "early" performance of the methods; the grey curve indicates the random retrieval of compounds (i.e. a random predictor).

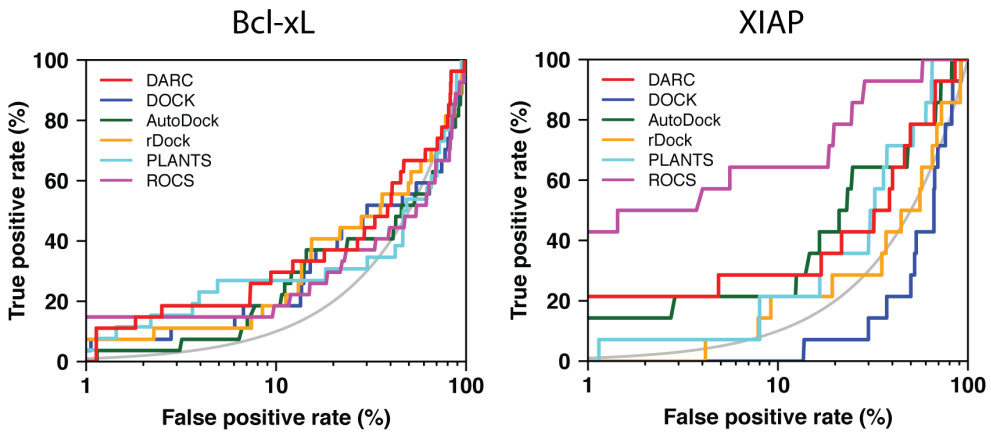


Figure S2: This receiver operating characteristic (ROC) plot compares the performance of various methods (DARC, DOCK, AutoDock, rDock, PLANTS, and ROCS) in a similar benchmark experiment to that described in **Figure 2**. Here, the same active compounds were included but decoy compounds were generated by the DUD-E server (81) to carefully match the physicochemical properties of the active compounds. There were 1350 decoy compounds in the Bcl-xL screen, and 700 decoy compounds in the XIAP screen.

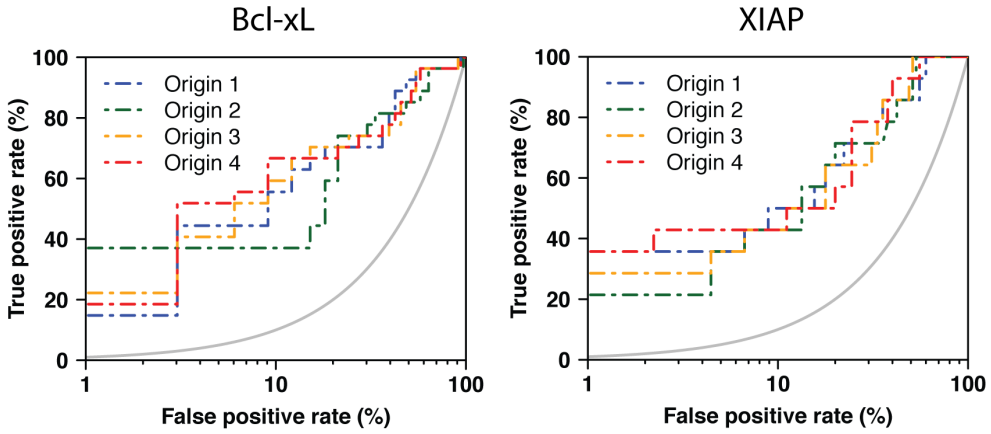


Figure S3: This receiver operating characteristic (ROC) plot examines the sensitivity of DARC's performance to the location of the "origin point" from which rays emanate. The benchmark is analogous to that described in **Figure 2**, with decoys drawn from the Astex diverse set (35). In this experiment, four different methods were used to separately determine the location of the origin point; in each case the

origin was positioned exactly 30 Å from the center of the surface pocket. For *Origin 1*, the origin was positioned by following the direction towards the protein center of mass. *Origin 2* corresponds taking a vector orthogonal to the maximal variation of the pocket points (this is the default method, as described in *Supplementary Methods*). *Origin 3* is calculated by first summing the vectors connecting each pocket point with the pocket center; for a concave pocket, this vector sum yields a vector that points "out" of the pocket. The sign of this vector is inverted, then its magnitude is scaled to give the location of the origin. Finally, *Origin 4* is determined by manually selecting an arbitrary residue in the binding site (Ser4 for Bcl-xL, and Glu282 for XIAP), then extending the vector connecting the pocket center to the center of mass of this residue to give the location of the origin. Each of these approaches gives a distinct – but reasonable – location of the origin point. For these two target proteins, the overall performance of DARC is remarkably robust to this variation of the location of the origin point.

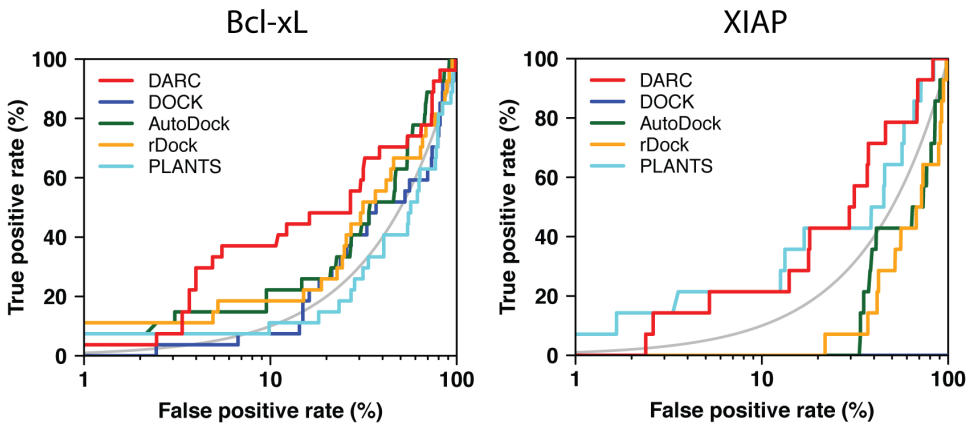


Figure S4: This benchmark experiment involves discrimination of the same active versus "decoy" compounds as in **Figure S1** (decoy compounds from the TIMBAL set (78, 79)). This time, however, compounds were screened against a protein conformation generated via "pocket optimization" simulations (instead of against an inhibitor-bound crystal structure). The protein conformation used for Bcl-xL and XIAP in this experiment was also used in carrying out the benchmark experiment shown **Figure 3**.

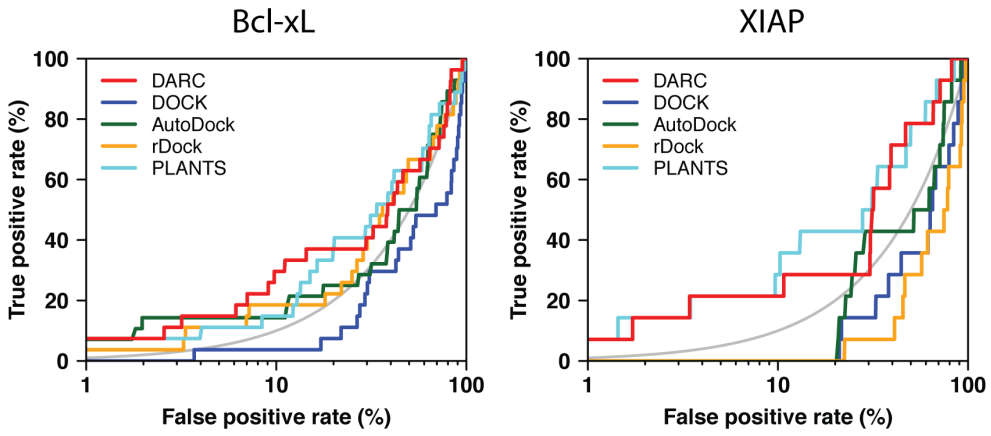


Figure S5: This benchmark experiment involves discrimination of the same active versus "decoy" compounds as in **Figure S2** (custom decoy compounds generated using the DUD-E server (81)). This time, however, compounds were screened against a protein conformation generated via "pocket optimization" simulations (instead of against an inhibitor-bound crystal structure). The protein conformation used for Bcl-xL and XIAP in this experiment was also used in carrying out the benchmark experiment shown **Figure 3**.

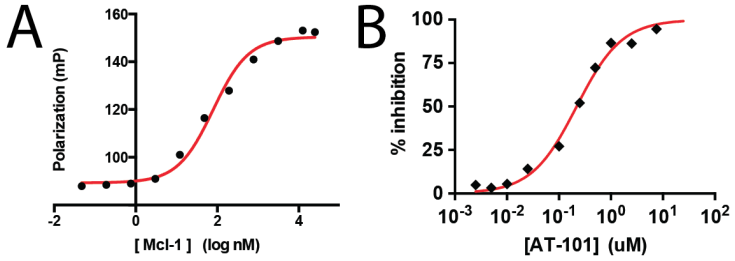


Figure S6: (A) Direct monitoring of the interaction between a FITC-labeled Noxa peptide and recombinant human Mcl-1 (at concentrations of 0.05 nM to 25 μM). Their interaction leads to an increase in fluorescence polarization (*black points*), which is well-described by a curve with Hill coefficient fixed at 1 (*red curve*); the estimated dissociation constant (Kd) for this interaction is 79 nM. **(B)** As a positive control for this assay, we tested for inhibition of Mcl-1 by AT-101 (i.e. R-(-)-gossypol), a known Mcl-1

inhibitor (129). Experimental conditions for this fluorescence polarization competition assay and fitting of the resulting data were carried out as described in **Figure 4** and in *Supplementary Results*.

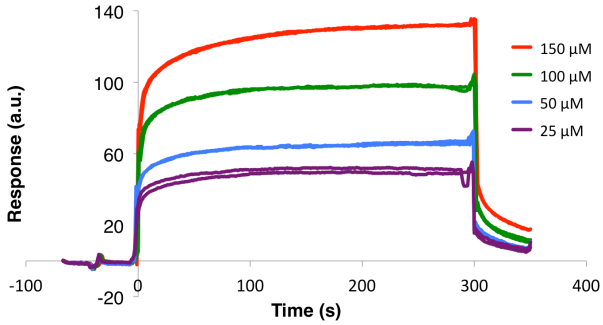


Figure S7: Surface plasmon resonance (SPR) confirms the M0/Mcl-1 interaction. Recombinant human Mcl-1 was immobilized on the surface of the chip, and increasing concentrations of M0 were flowed across the surface (concentrations of 25/50/100/150 µM are shown, each in duplicate).

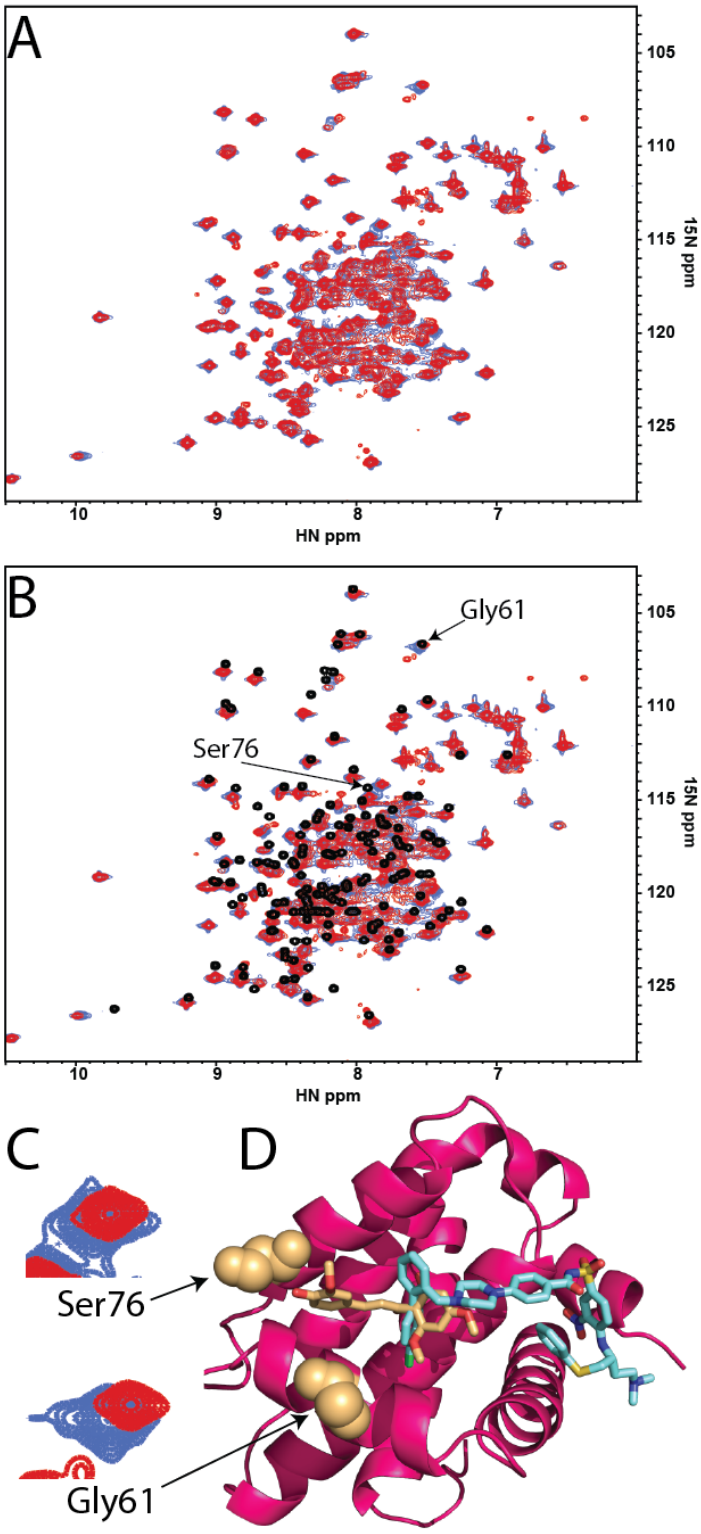


Figure S8: (A) HSQC spectra are superposed for 300 µM recombinant human Mcl-1 alone (*blue*), and in the presence of 100 µM M0 (*red*). Despite the fact that Mcl-1 remains in excess, several peaks exhibit

chemical shift differences. In separate experiments, we have confirmed that the addition of DMSO-d_6 was not responsible for these observed chemical shift differences. **(B)** Superposition of these two (unassigned) spectra with the assigned spectrum of human Mcl-1 (125) (*black*). Despite the use of slightly different constructs and buffer conditions, assignment of some peaks can be unambiguously transferred. **(C)** Among peaks that exhibit chemical shift differences upon addition of M0, assignments can be unambiguously transferred for only two of these: Ser76 and Gly61. **(D)** Mapping the location of these two residues to the DARC model of the M0/Mcl-1 complex reveals that they are located at the edge of the expected binding site, providing support for this model of M0 binding.

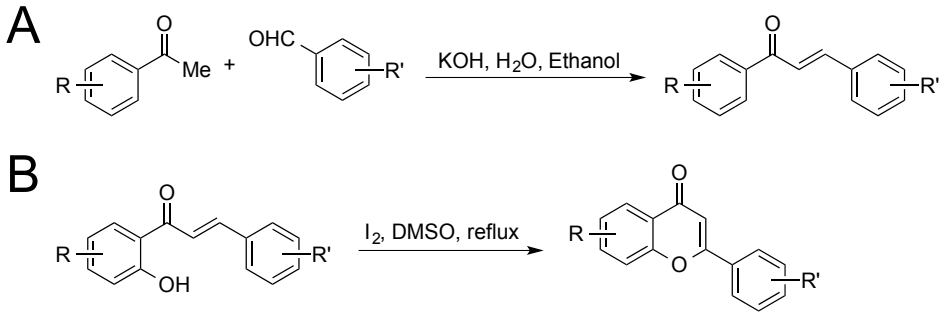


Figure S9: Generalized reaction schemes for **(A)** chalcone synthesis (compounds D1-D23) and **(B)** flavone synthesis (compounds D24-D27).

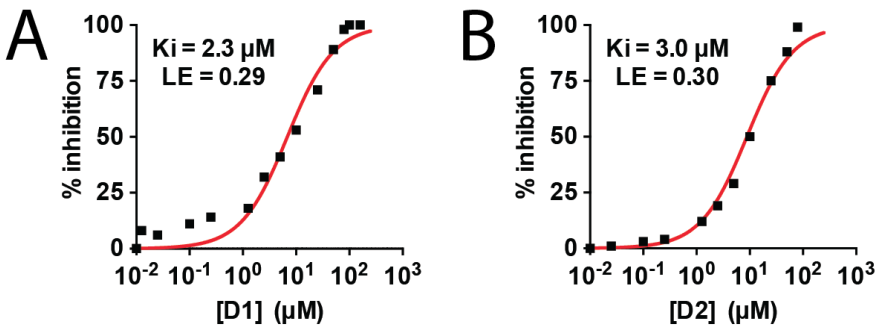


Figure S10: Inhibition of Mcl-1 by the most potent M0 derivatives, D1 and D2. Experimental conditions for this fluorescence polarization competition assay and fitting of the resulting data were carried out as described in **Figure 4**.

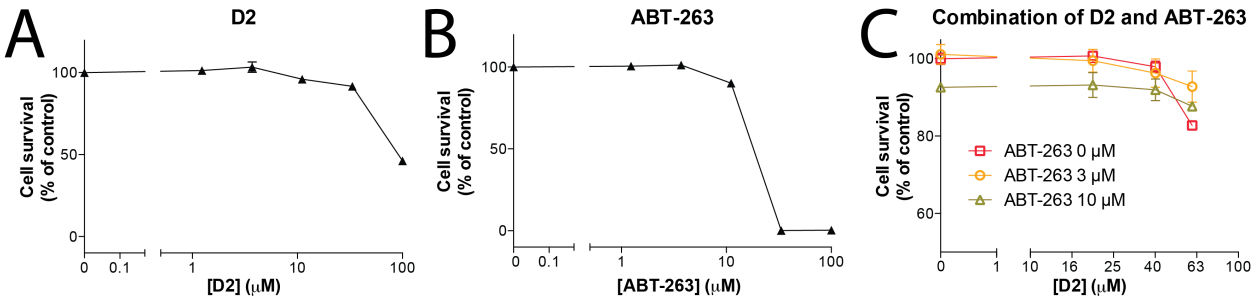


Figure S11: (A) Incubation of cells with compound D2 reduces viability of DU-145 cells only at very high concentrations. **(B)** An inhibitor of Bcl-xL and Bcl-2, ABT-263, reduces viability of DU-145 cells at concentrations above 10 µM. **(C)** Addition of ABT-263 does *not* sentitize cells to D2, suggesting lack of synergy between these compounds. Lines are drawn between adjacent points to guide the eye.

APPENDIX A.3: Supporting Information for Chapter 3

DARC 2.0: Improved docking and virtual screening at protein interaction sites

Ragul Gowthaman[1], Sergey Lyskov[2], and John Karanicolas[1,3]

[1] Center for Computational Biology, [3] Department of Molecular Biosciences,

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

[2] Department of Chemical and Biomolecular Engineering,

Johns Hopkins University, 3400 North Charles St., Baltimore, MD 21218

3.1 Supplementary Tables

Protein	Ligand-bound PDB	Ligand conformers used
Calpain	1ALW	8
calmodulin	1CTR	30
IL-2	1PW6	300
HPV E2	1R6N	50
XIAP-BIR3	1TFT	175
ZipA	1Y2F	300
Bcl-xL	1YSI	300
TNFα	2AZ5	300
HIV-gp41	2KP8	42
integrin	2VC2	300
BRD4	2YEL	250
S100B	3GK1	198
Grb2-SH2	3IN7	300
SHANK PDZ	3O5N	3
WDR5	3UR4	300
PCNA	3VKX	62
VHL	3ZRC	300
Plasminogen	4CIK	36
HIV integrase	4E1N	1
Mdm2	4ERF	10

clathrin	4G55	300
Menin	4GQ4	175
Cytohesin-2	4JMO	7
keap1	4L7D	59
RPA1	4LUZ	264

Table S1: Small-molecule inhibitors bound to protein interaction sites. We compiled a set of 25 unique protein interaction sites for which a crystal structure has been solved in complex with a small-molecule inhibitor. This set is based upon our previous set of 21 complexes that were available at the time (drawn in part from the 2P2I (56) and TIMBAL (78) databases) (71), and we now add 4 additional examples have since become available. We did not include complexes with small molecule stabilizers, or complexes with small fragments or large peptide-like compounds. We only included one representative complex from each protein family; in cases where more than one suitable inhibitor-bound structure had been solved from a given family, we retained only the structure in complex with the most potent ligand. We also report here the number of conformers for each ligand used in these studies; conformers were generated using OMEGA (see *Methods*), and an average of 163 were used for the ligands in this set.

3.2 Supplementary Data

Dataset S1. Protocol Capture. The protocol capture contains all the example input and output files, and representative command-lines and flags required for running DARC. The dataset was uploaded to the Rosetta demos revision control (Rosetta/demos/public/DARC/).

APPENDIX A.4 Supporting Information for Chapter 4

Rationally designing inhibitors of the Musashi protein-RNA interaction by hotspot mimicry

Yan Xia[1†], Ragul Gowthaman[2†], Lan Lan[1], Steven Rogers[3], Andy R. Wolfe[1], Christian L. Gomez[1], Oscar Ramirez[4], Bryan W. Tsao[1], Kelin Li[3], Jia Yu[1], Rebecca T. Marquez[1], Chunjing Liu[3], Manoj M. Pillai[4], Jeffrey Aubé[3,5], Kristi L. Neufeld[1,6], Liang Xu[1,7], and John Karanicolas[1,2]

[1] Department of Molecular Biosciences, [2] Center for Bioinformatics, [3] Center of Biomedical Research Excellence, Center for Cancer Experimental Therapeutics, [5] Department of Medicinal Chemistry, [6] Department of Cancer Biology, KU Medical Center, and [7] Department of Radiation Oncology University of Kansas, Lawrence, KS 66045

[4] Section of Hematology, Yale Cancer Center, 300 George St., New Haven, CT 06511

[†]Equal author contributions.

4.1 Supporting Methods

PDB structures used in calculations

The calculations that led to selection of R1-R12 were carried out using model 1 of the NMR structure of Musashi-1 bound to RNA (PDB ID 2RS2).

Building hotspot pharmacophores

Hotspot pharmacophores were built using a dedicated protocol implemented in the Rosetta software suite (60), and is freely available for academic use (www.rosettacommons.org).

To select deeply buried RNA bases, the solvent accessible surface area (SASA) of each base in the RNA was calculated in the presence and absence of the protein. A base was carried forward if the change in SASA upon complexation was greater than a preset cutoff value (46.81 $Å^2$ for adenine, 31.09 $Å^2$ for cytosine, 45.06 $Å^2$ for guanine and 52.66 $Å^2$ for uracil); these values correspond to the median values of 344 non-redundant protein-RNA complexes retrieved from the Protein-RNA Interface Database (PRIDB) (264) in March 2013 (http://pridb.gdcb.iastate.edu/download/RB344.txt).

Polar groups from the RNA that participate in intermolecular hydrogen bonding (as defined using the Rosetta energy function) are also included.

The Rosetta command line used to carry out this step is as follows:

```
get_rna_pharmacophore_with_water.macosgccrelease -input_rna xxx_rna.pdb -input_protein xxx_protein.pdb
```

The resulting interaction maps are then clustered using a modified version of Kruskal's minimum spanning tree algorithm. We first build a complete graph, in which vertices are the ring moieties, and the edge weights are the Euclidean distances between vertices. Then we take edges in ascending order and

cluster the end vertices of that edge if no cycle would be caused. We halt the clustering when the distance is greater than a user-specified cutoff value (default 5.0 Å). The donor/acceptor atoms are then assigned to the closest ring moieties if the distance is less than another user-specified value (default 5.0 Å). Finally, we output the pharmacophore templates if the cluster contains at least two ring moieties. The Kruskal clustering code is also implemented in Rosetta, and is carried out as follows:

```
cluster_pharmacophore.macosgccrelease –input xxx_rna.pdb –ring_cutoff xxx –da_cutoff xxx
```

Identifying complementary ligands

We used ROCS to screen large libraries for compounds that match the hotspot pharmacophore. We downloaded the standard 'drugs-now' subset of ~7 million molecules from ZINC database for screening (43). We generated up to 100 conformers for each molecule in the database using OMEGA (44-46). We screened the database using the hotspot pharmacophore (using default ROCS parameters), and carried forward the top 500 compounds ranked by 'TanimotoCombo' score. We then aligned these back to the protein using the hotspot pharmacophore, then carried out a gradient-based fullatom minimization of the complex using the Rosetta energy function (60). The top-scoring compounds were visually inspected and selected for experimental validation based on cost and availability.

Predicting target selectivity

The complete set of 1792 protein-RNA complexes were retrieved from the PDB in June 2014. Hotspot pharmacophores were extracted from each complex, and non-unique pharmacophores were removed (those with ROCS shape_tanimoto > 0.94 and color_tanimoto > 0.74). This left 543 unique pharmacophores that were comprised of at least two rings, derived from 362 different protein-RNA complexes.

Conformers for each compound were generated by OMEGA using the following command line:

```
omega2 –in xxx.pdb –strictatomtyping false –strictstereo false –strictfrags false –
searchff mmff94s –buildff mmff94s –maxconfs 500
```

For a given compound, we then used ROCS to screen conformers of this molecule against the library of hotspot pharmacophores using the following command line:

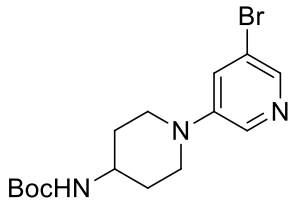
```
rocs –dbase conformer_ensemble.pdb –query hotspot.pdb –oformat pdb –rankby
FitTverskyCombo
```

Synthesis of R12 derivatives : overview

All air- and moisture-sensitive reactions were carried out in flame- or oven-dried glassware under argon atmosphere using standard gastight syringes, cannula, and septa. Stirring was achieved with oven-dried magnetic stir bars. Flash column chromatography was performed with SiO2 from Sorbent Technology (30930M-25, Silica Gel 60A, 40−63 μm) or by using an automated chromatography instrument with an appropriately sized column. Thin layer chromatography was performed on silica gel w/UV254 plates (1624126, Sorbent Technologies). ^{1}H and ^{13}C NMR spectra were recorded on instruments operating at 400 or 500 MHz and 100 or 126 MHz, respectively. High-resolution mass spectrometry (HRMS) spectra were obtained on an ESITOF mass spectrometer. The analytical method utilized a Waters Aquity BEH C18 column (2.1 × 50 mm, 1.7 μm) eluting with a linear gradient of 95% water (modified to pH 9.8 through addition of NH_4OH) to 100% CH_3CN at 0.6 mL/min flow rate where purity was determined using UV peak area at 214 nm.

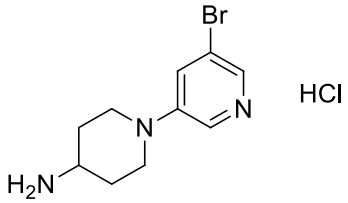
Synthesis of specific intermediates / derivatives is described in detail in the following sections, and summarized in **Figure S4**.

Synthesis of R12 derivatives : general procedure #1



Following a modified procedure outlined by Do et al. (265), a mixture of *tert*-butyl 4-aminopiperidine-1-carboxylate (0.85 g, 4.22 mmol), 3,5-dibromopyridine (1.0 g, 4.22 mmol) tris(dibenzylideneacetone)dipalladium(0) (0.077 g, 0.084 mmol), (±)-2,2'-bis(diphenylphosphino)-1,1'-binaphthalene (0.11 g, 0.17 mmol) and sodium-t-butoxide (0.61 g, 6.33 mmol) in toluene (30.2 mL) was heated to 80 °C for 16 h, then the reaction mixture was allowed to cool to ambient temperature, diluted with ether (100 mL) and washed with brine (3×30 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated under vacuum. The residue was purified by silica gel chromatography (50% EtOAc in hexanes, Rf = 0.5) to afford the title compound *tert*-butyl 4-((5-bromopyridin-3-yl)amino)piperidine-1-carboxylate (879.8 mg, 2.47 mmol, 59% yield) as a white solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.22 (d, *J* = 2.6 Hz, 1H), 8.11 (d, *J* = 1.9 Hz, 1H), 7.31 (dd, *J* = 2.6, 1.9 Hz, 1H), 4.55 (s, br. 1H), 3.70 – 3.60 (m, 3H), 2.97 – 2.90 (m, 2H), 2.14 – 2.03 (m, 2H), 1.60 – 1.49 (m, 2H), 1.47 (s, 9H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 155.10, 147.64, 140.51, 136.65, 124.72, 120.81, 47.51, 47.42, 31.88, 28.40.

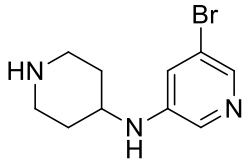
Synthesis of R12 derivatives : general procedure #2



To a solution of tert-butyl 4-((5-bromopyridin-3-yl)amino)piperidine-1-carboxylate (433.5 mg, 1.22 mmol) in dichloromethane (9 mL), was added hydrogen chloride in dioxane (15.2 mL, 60.8 mmol). The reaction was stirred at rt for 16 h. Solvents were removed to give a white solid. Yield: 445.0 mg,

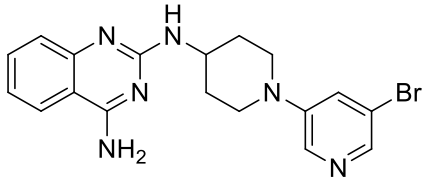
100%. ^{1}H NMR (400 MHz, MeOD) δ 8.53 (d, *J* = 2.4 Hz, 1H), 8.38 – 8.31 (m, 2H), 4.19 – 4.10 (m, 2H), 3.83 – 3.58 (m, 3H), 3.53 – 3.47 (m, 1H), 3.26 – 3.14 (m, 2H), 2.26 – 2.17 (m, 2H), 1.84 – 1.73 (m, 2H). ^{13}C NMR (101 MHz, MeOD) δ 150.00, 132.59, 131.09, 127.98, 123.69, 73.59, 72.48, 62.23, 46.60, 43.84.

Synthesis of R12 derivatives : general procedure #3



To a solution of 5-bromo-N-(piperidin-4-yl)pyridin-3-amine dihydrochloride (200.0 mg, 0.55 mmol) in methanol (4 mL), was added sodium carbonate (174.0 mg, 1.64 mmol). Solvent was removed and residue was extracted with DCM. DCM was dried over $MgSO_4$ and evaporated to dryness to give a light-yellow oil. Yield: 81.0 mg, 58%. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.20 (d, *J* = 2.6 Hz, 1H), 8.06 (d, *J* = 1.9 Hz, 1H), 7.28 (dd, *J* = 2.6, 1.9 Hz, 1H), 3.82 – 3.56 (m, 3H), 2.93 – 2.79 (m, 3H), 1.97 – 1.86 (m, 2H), 1.50 – 1.43 (m, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 147.78, 140.25, 136.61, 124.48, 120.80, 48.22, 47.37, 35.15. HRMS (m/z): calcd for $C_{10}H_{14}BrN_3$ (neutral M+H) 255.0371; found 255.0379.

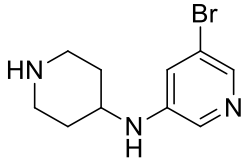
Synthesis of R12 derivatives : general procedure #4



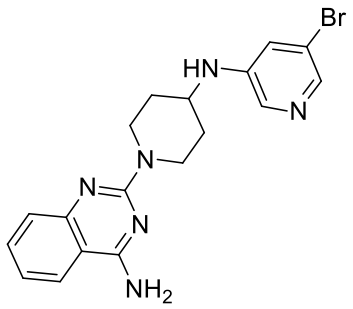
To a solution of 2-chloroquinazolin-4-amine (28.4 mg, 0.16 mmol) and 5-bromo-N-(piperidin-4-yl)pyridin-3-amine (81.0 mg, 0.32 mmol) in acetonitrile (2 mL) was heated at 180 °C for 1 h under microwave irradiation. The material was purified via reverse phase combiflash first, followed by further purification via silica gel chromatography (DCM/MeOH = 10:1, Rf = 0.3) to give 2-(4-((5-bromopyridin-

3-yl)amino)piperidin-1-yl)quinazolin-4-amine (36.4 mg, 0.091 mmol, 58% yield) as a white solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.24 (d, *J* = 2.6 Hz, 1H), 8.12 (d, *J* = 1.9 Hz, 1H), 7.65 – 7.56 (m, 2H), 7.51 – 7.48 (m, 1H), 7.33 (t, *J* = 2.3 Hz, 1H), 7.16 – 7.12 (m, 1H), 5.57 (s, br. 2H), 5.01 (s, br. 1H), 4.23 – 4.14 (m, 1H), 3.72 – 3.62 (m, 2H), 3.07 – 3.00 (m, 2H), 2.25 – 2.20 (m, 2H), 1.71 – 1.56 (m, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 162.1, 158.5, 152.3, 147.8, 140.4, 136.7, 133.4, 125.6, 124.6, 121.9, 121.5, 120.8, 110.4, 47.5, 47.3, 31.9. HRMS (m/z): calcd for $C_{18}H_{20}BrN_6$ (neutral M+H) 399.0933; found 399.0900.

Synthesis of R12 derivatives : specific compounds

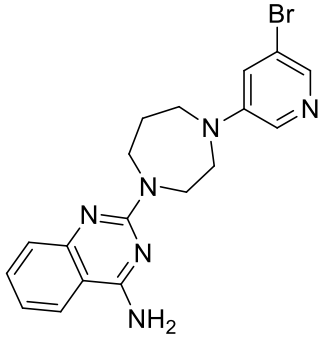


Synthesized by using general procedures #1, #2 and then #3. ^{1}H NMR (400 MHz, DMSO-d_6) δ 8.35 (d, *J* = 2.6 Hz, 1H), 8.15 – 8.02 (m, 4H), 7.72 (t, *J* = 2.2 Hz, 1H), 3.92 (d, *J* = 13.5 Hz, 2H), 3.25 (d, *J* = 12.5 Hz, 1H), 2.97 – 2.84 (m, 2H), 1.95 (d, *J* = 10.8 Hz, 2H), 1.57 (qd, *J* = 12.2, 4.1 Hz, 2H). HRMS (m/z): calcd for $C_{10}H_{14}BrN_3$ (neutral M+H) 255.0371; found 255.0366.

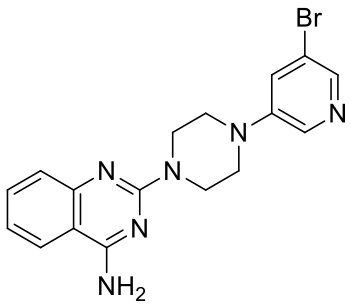


Synthesized using general procedure 4 by reacting 2-chloroquinazolin-4-amine (0.036 g, 0.203 mmol) and 5-bromo-N-(piperidin-4-yl)pyridin-3-amine (.078 g, 0.305 mmol) to give 2-(4-((5-bromopyridin-3-yl)amino)piperidin-1-yl)quinazolin-4-amine (.062 g, 0.155 mmol, 76 % yield) ^{1}H NMR (400 MHz, Acetone-d_6) δ 7.97 (d, *J* = 2.5 Hz, 1H), 7.88 (dd, *J* = 8.1, 1.4 Hz, 1H), 7.78 (d, *J* = 1.9 Hz, 1H), 7.47 (ddd, *J* = 8.4, 6.8, 1.4 Hz, 1H), 7.30 (dd, *J* = 8.5, 1.3 Hz, 1H), 7.17 (t, *J* = 2.2 Hz, 1H), 7.00

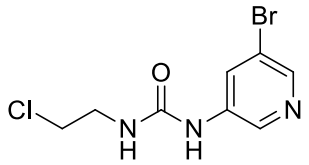
(ddd, *J* = 8.2, 6.8, 1.2 Hz, 1H), 5.38 (d, *J* = 8.1 Hz, 1H), 4.83 – 4.66 (m, 2H), 3.63 (tdd, *J* = 6.5, 4.2, 2.4 Hz, 1H), 3.08 (ddd, *J* = 13.8, 11.5, 2.7 Hz, 2H), 1.45 – 1.30 (m, 2H). ^{13}C NMR (101 MHz, Acetone) δ 206.12, 163.36, 159.98, 153.94, 146.20, 138.11, 135.70, 133.34, 126.43, 123.77, 121.26, 120.48, 111.01, 54.95, 50.64, 43.27, 32.60, 30.41, 30.22, 30.03, 29.83, 29.64, 29.45, 29.26. HRMS (m/z): calcd for $C_{18}H_{20}BrN_6$ (neutral M+H) 399.0871; found 399.0855.



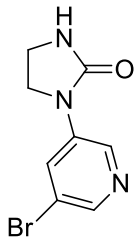
Synthesized using general procedure 4 by reacting 2-chloroquinazolin-4-amine (0.045 g, 0.248mmol) and 1-(5-bromopyridin-3-yl)-1,4-diazepane (266) (0.127 g, 0.496 mmol) to give 2-(4-(5-bromopyridin-3-yl)-1,4-diazepan-1-yl)quinazolin-4-amine (0.067g, 0.168 mmol, 68% yield) ^{1}H NMR (400 MHz, Chloroform-*d*) δ 8.05 (d, *J* = 2.7 Hz, 1H), 7.92 (d, *J* = 1.7 Hz, 1H), 7.57 – 7.45 (m, 3H), 7.13 – 7.01 (m, 2H), 4.10 – 4.00 (m, 2H), 3.72 (t, *J* = 6.2 Hz, 2H), 3.64 (t, *J* = 5.3 Hz, 2H), 3.49 (t, *J* = 6.2 Hz, 2H), 2.10 (t, *J* = 6.2 Hz, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 161.72, 158.08, 144.71, 137.58, 133.16, 132.56, 125.72, 121.87, 121.24, 121.14, 120.16, 109.69, 77.36, 77.04, 76.72, 53.75, 49.89, 47.84, 46.61, 46.22, 24.37. HRMS (m/z): calcd for $C_{18}H_{20}BrN_6$ (neutral M+H) 399.0864; found 399.0855.



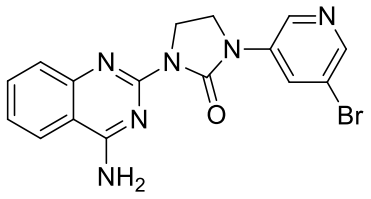
Synthesized using general procedure 4 by reacting 2-chloroquinazolin-4-amine (0.022 g, 0.124 mmol) and 1-(5-bromopyridin-3-yl)piperazine (266) (.03 g, 0.124 mmol) to give 2-(4-(5-bromopyridin-3-yl)piperazin-1-yl)quinazolin-4-amine (.031 g, 0.080 mmol, 64.9 % yield). ^{1}H NMR (400 MHz, Chloroform-*d*) δ 8.26 (d, *J* = 2.6 Hz, 1H), 8.14 (d, *J* = 1.9 Hz, 1H), 7.63 – 7.48 (m, 3H), 7.34 (dd, *J* = 2.6, 1.9 Hz, 1H), 7.13 (ddd, *J* = 8.1, 6.8, 1.3 Hz, 1H), 4.42 – 3.94 (m, 4H), 3.58 – 3.11 (m, 4H). ^{13}C NMR (126 MHz, $CDCl_3$) δ 161.68, 158.76, 152.60, 147.97, 141.01, 136.62, 133.35, 126.08, 124.64, 121.76, 121.62, 120.86, 109.91, 48.26, 43.48. HRMS (m/z): calcd for $C_{17}H_{17}BrN_6$ (neutral M+H) 384.0698; found 384.0709.



5-bromopyridin-3-amine (0.3 g, 1.734 mmol) was dissolved in acetonitrile and the reaction mixture was cooled to 0 °C. N,N'-Diisopropylethylamine (0.636 ml, 3.65 mmol) was added dropwise. Then, 1-chloro-2-isocyanatoethane (0.156 ml, 1.825 mmol) was added dropwise. The reaction was allowed to stir for 16 hours. The reaction was quenched with water, washed with brine, dried with sodium sulfate and then concentrated en vacuo and then purified via normal phase chromatography (ethyl acetate/hexanes) to give 1-(5-bromopyridin-3-yl)-3-(2-chloroethyl)urea (.49 g, 1.759 mmol, 96 % yield)^{1}H NMR (400 MHz, DMSO-d_6) δ 9.09 (s, 1H), 8.44 (d, *J* = 2.2 Hz, 1H), 8.27 (t, *J* = 2.2 Hz, 1H), 8.23 (d, *J* = 2.1 Hz, 1H), 6.67 (t, *J* = 5.9 Hz, 1H), 3.67 (t, *J* = 6.1 Hz, 2H), 3.43 (q, *J* = 6.1 Hz, 2H). ^{13}C NMR (101 MHz, DMSO) δ 149.09, 142.29, 137.94, 130.73, 126.34, 114.38, 44.06, 41.27. HRMS (m/z): calcd for $C_8H_9BrClN_3O$ (neutral M+H) 276.9609; found 276.9618.



In a round bottomed flask, 1-(5-bromopyridin-3-yl)-3-(2-chloroethyl)urea (.3 g, 1.077 mmol) was dissolved in THF, a stir bar was added and the reaction mixture was cooled to 0 °C. To the reaction mixture sodium hydride (0.078 g, 3.23 mmol) was added slowly. The reaction was allowed to slowly warm to room temperature and allowed to continue stirring for 16 hours. The reaction was cooled to 0 °C and then quenched with the dropwise addition of water. More water was added, then extracted three times with ethyl acetate, washed twice with water, twice with brine and then dried with sodium sulfate and concentrated en vacuo. The crude residue was then purified via normal phase chromatography (Methanol/DCM) to give 1-(5-bromopyridin-3-yl)imidazolidin-2-one (.217 g, 0.896 mmol, 83 % yield). ^{1}H NMR (400 MHz, DMSO-d_6) δ 8.72 – 8.64 (m, 1H), 8.34 (t, J = 2.2 Hz, 1H), 8.32 – 8.24 (m, 1H), 7.32 (s, 1H), 3.96 – 3.86 (m, 2H), 3.45 (ddd, J = 8.9, 6.8, 1.1 Hz, 2H). ^{13}C NMR (101 MHz, DMSO) δ 158.78, 142.26, 136.70, 125.52, 119.66, 43.90, 36.92. HRMS (m/z): calcd for $C_8H_8BrN_3O$ (neutral M+H) 240.9885; found 240.9851.



To a dried round bottomed flask was added 1-(5-bromopyridin-3-yl)imidazolidin-2-one (.1 g, 0.413 mmol), THF and a stir bar. The mixture was cooled to 0 °C. Then, sodium hydride (0.030 g, 1.239 mmol) was added slowly. The mixture was allowed to stir for 30 minutes. Then, 2-chloroquinazolin-4-amine (0.074 g, 0.413 mmol) was added. The reaction was allowed to warm to room temperature slowly and then continue stirring for 12 hours. The reaction was quenched with water, extracted three times with ethyl acetate, washed with brine, dried with sodium sulfate and then concentrated en vacuo. The crude residue was purified via normal phase chromatography (Methanol/dichloromethane) and tehn reverse phase chromatography (water pH = 9, acetonitrile) to yield 1-(4-aminoquinazolin-2-yl)-3-(5-bromopyridin-3-yl)imidazolidin-2-one (.11 g, 0.286 mmol, 69.1 % yield). ^{1}H NMR (400 MHz, DMSO-d_6) δ 8.83 (d, J = 2.3 Hz, 0H), 8.50 – 8.38 (m, 1H), 8.23 – 8.13 (m, 0H), 7.82 (s, 1H), 7.70 (ddd, J = 8.4,

6.9, 1.4 Hz, 0H), 7.57 – 7.51 (m, 0H), 7.34 (ddd, *J* = 8.3, 7.0, 1.3 Hz, 0H), 4.15 (dd, *J* = 9.2, 6.6 Hz, 1H), 4.04 – 3.92 (m, 1H). ^{13}C NMR (101 MHz, DMSO) δ 162.52, 154.72, 152.23, 150.58, 143.31, 137.97, 137.47, 133.05, 126.49, 126.33, 123.56, 119.65, 111.96, 102.47, 41.66, 40.64. HRMS (m/z): calcd for $C_{16}H_{13}BrN_6O$ (neutral M+H) 384.0334; found 384.0340.

Model building of R12 derivatives

Conformers of R13-R17 were generated using OMEGA. For each compound, we aimed to generate sample likely conformers that optimally matched the ring geometry in the hotspot pharmacophore. To achieve this, we used the CHARMM software (267) to carry out a biased energy minimization of the compound (in the absence of the protein). We implemented the bias using a harmonic constraint applied to the Cartesian coordinates of certain atoms, centered at the position of the corresponding atom of the hotspot pharmacophore (C4,C5,N7,C8 and N9 on Adenine106 and N1, C2, N3, C4, C5, C6, N6, N7, C8 and N9 on Guanine107) and with a scale factor of 100 (this scale factor is related to the force constant in a way that depends on the mass of individual atoms). The residue topology file and parameter file for the compounds required by CHARMM minimization were obtained from CHARMM-GUI (268).

Through this minimization of OMEGA conformers, we generated models of the R12 derivatives that maintain the ring geometry in the hotspot pharmacophore but contain a variety of geometries in the linker region. Since the resulting conformers match the hotspot pharmacophore, they are already aligned to the structure of the Msi1-RNA complex. We concluded by selecting the best model on the basis of protein-ligand interaction energy using the fullatom Rosetta energy function (60).

Expression and purification of Msi1

A gene encoding human Msi1 RBD1 domain was subcloned as a fusion protein with an N-terminal 6xHis-tagged streptococcal GB1 domain and a tobacco etch virus (TEV) protease site. The

expression plasmid was transformed into *Escherichia coli* BL21(DE3) pLysS, then a 5 mL overnight starter culture was used to inoculate a 1 L culture of Luria-Bertani (LB) media. Cells were grown at 37 ºC to an OD_{600} of 0.6–0.8 and were induced with 1 mM IPTG overnight at 15 ºC. The induced cells were resuspended in lysis buffer (50 mM Tris, 150 mM NaCl, 5 mM imidazole pH 8.0) and sonicated for 10 minutes (Fisher Scientific Sonic Dismembrator Model 100). The cell lysates were then centrifuged at 15,000g for 30 min. The GB1-RBD1 remained in the supernatant, which was purified by HPLC affinity chromatography with Ni-chelated Sepharose Fast Flow Resin (GE Healthcare), followed by a HiLoad 16/60 Superdex 75 gel filtration column (GE Healthcare). GB1 tag was digested with TEV protease (1 OD_{280} of TEV per 5 OD_{280} of fusion protein) in reaction buffer (50 mM Tris-HCl, 0.5 mM EDTA and 1mM DTT pH 8.0). All protein concentrations were determined with reference to bovine albumin standards using Bradford assays.

Fluorescence polarization competition assays

RNA oligonucleotides were ordered from Integrated DNA Technologies (Coralville, IA) and dissolved in TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8.0): sequences are included in **Table S1**. To measure the dissociation constant of Msi1 RBD1 and RNA binding, a fixed concentration (2 nM) of fluorescein-labeled RNA (FC-NUMB, **Table S1**) and increasing concentrations of Msi1 RBD1 (1 nM to 1000 nM) were mixed in binding assay buffer (20 mM HEPES, 150 mM NaCl, 0.05% F-68 pH 7.4). Fluorescence intensities were measured in replicate on the BioTek Synergy 2 plate reader (Winooski, VT) and the fluorescence polarization value (FP) was calculated by the following equation:

$$FP = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}}$$

The dissociation constant (K_D) was fit using Prism (v 6.0e, GraphPad Software, Inc., La Jolla, CA), with the Hill coefficient fixed at n=1 as follows:

$$Y = Bottom + \frac{Top - Bottom}{1 + K_D/L}$$

To test the contribution to binding affinity of each base in the NUMB RNA sequence, we purchased five oligos that each harbor an abasic site at a different position, as well as the corresponding wild-type (aNUMB0-5, **Table S1**). K_i values were then determined from a competition experiment in which serial dilution of unlabeled aNUMB oligos (5 nM – 5000 nM) were added to compete against a fixed concentration (2 nM) of fluorescein-labeled RNA (FC-NUMB) for binding to a fixed concentration of Msi1 RBD1 (75 nM). The K_i value was determined by fitting with Prism to the "Binding – Competitive – One site – Fit Ki" model (269).

To examine the displacement of FC-NUMB by R13, we performed the same competition assay using R13 as a competitor. The K_i value was determined as described above.

Surface plasmon resonance

Binding between Msi1 RBD1 and each compound were analyzed by SPR using a Biacore 3000 optical biosensor (GE Healthcare). GB1-tagged Msi1 RBD1 was covalently immobilized by amine-coupling on a carboxymethylated dextran sensor chip (CM-5, GE Healthcare). Amine-coupling reactions for immobilization of proteins were performed at approximately 5 μg⁄mL in 10 mM sodium acetate buffer pH 5.5 injected at 5 μL⁄min until 8400 response units (RU) were immobilized. An unmodified flow cell was used as reference. An unrelated protein, human Mcl-1, was immobilized on another flow cell and used to subtract out the response from unspecific binding.

All SPR runs were performed at 25 ºC using a flow rate of 50 μL/min in running buffer (HEPES buffered saline with 0.05% Tween-20 pH 7.4). Compounds were injected over sample and reference flow

cells at a concentration of 50 μM in running buffer, for 250 s. Following each injection, flow cells were regenerated with a 20 s injection of 1 M NaCl.

SPR titration data were analyzed by using Scrubber 2 software (Biologic) to zero, crop, align and subtract responses from the unmodified surface and average blank injections. Response from the Mcl-1 flow cell was also subtracted to remove the response from unspecific binding.

Differential scanning fluorimetry (Thermofluor)

Differential scanning fluorimetry (DSF) experiments were carried out using a standard protocol described by others (270). All experiments were carried out in a reaction volume of 25 μL, with 25 mM Tris-Cl pH 8.0, 120 mM NaCl, 2% DMSO, and 100x-diluted Sypro Orange dye (Invitrogen). Multiple concentrations of GB1-tagged Msi1 RBD1 (ranging from 1 μM to 15 μM) were tested to identify the lowest concentration necessary to generate a smooth melting curve. For subsequent experiments, we used a concentration of 7.5 μM.

This concentration of protein was incubated with varying concentrations of R13 (ranging from 0.1 μM to 100 μM). Testing tubes were incubated in StepOnePlus™ Real-Time PCR System (Applied Biosystems) and samples were heated from 25 ºC to 65 ºC gradually with 0.5% increase. The fluorescence emission was measured using filter for ROX (610 nm).

The melting temperature (Tm) values were determined by taking the maximum of the first derivative of the raw fluorescence intensity with respect to temperature (270), using GraphPad Prism 5.

Msi1 reporter assay

HCT-116βW cells, a generous gift from Bert Vogelstein, were cultured as previously described (271). *APC* 3' UTR-luciferase reporter was utilized to assay Msi1 activity in cell culture as previously described (234). Briefly, GeneExpresso (Lab Supply Mall, Gaithersburg, MD) was used to co-transfect cells at 30-40% confluency with *APC* 3' UTR-luciferase reporter (900 ng) or pGVP2 vector (900 ng) and

Renilla luciferase expression construct (100 ng) in 24-well plates. 24 hours after transfection the cells were treated with R12 derivatives at 20 μM for 24 hrs. The cells were lysed using 60 μl of 1X passive lysis buffer (Promega) for 10 min. Lysates were assayed using the Dual-Luciferase Assay System (Promega) and a LMAXII384 microplate reader (Molecular Devices, Sunnyvale, CA) per manufacturer's instructions.

Firefly luciferase data were normalized to Renilla luciferase. Fold change of the normalized firefly luciferase from each compound was calculated relative to the normalized firefly luciferase from DMSO. Data were collected from six independent experiments. Statistical significance was evaluated using the one-tailed Wilcoxon rank-sum test.

Msi1 Numb assay

HCT-116βw (mut ko, β-cat w/-) cells were maintained in McCoy's 5A media (Iwakata & Grace Modifications) with L-glutamine containing 10% fetal bovine serum at 37 °C with 5% CO_2. Cells were treated with R12 derivatives or vehicle to a final concentration of 20 μM. Twenty-four hours after treatment proteins from cell extracts were resolved by electrophoresis on a 4-20% SDS-PAGE gel. The proteins were then transferred to a nitrocellulose membrane (Protran BA79) and blocked in 5% milk/TBST overnight at 4 °C. Membranes were incubated with primary antibodies (Rabbit anti-Numb mAb, 1:3000, Cell Signaling Technologies C29G11; Mouse anti-β-actin mAb, 1:5000, Sigma A2228) in 1% milk/TBST overnight at 4 °C followed by incubation with horseradish peroxidase conjugated secondary antibodies in 1% milk/TBST (Goat anti-Rabbit IgG (H+L) HRP conjugate, 1:10,000, Bio-Rad 172-1019; HRP-Goat anti-mouse IgG (H+L), 1:10,000, Invitrogen 626520). The blots were developed using SuperSignal West Femto Maxiumum Sensitivity Substrate (Themo Scientific 34096) for Numb or SuperSignal West Pico Chemiluminescent Substrate (Pierce 34080) for β-actin. Blot images were captured and net band intensities quantified using the IS4000R Image Station (Kodak). Graph represents average net band intensities collected from four independent experiments, with Numb values normalized

to β-actin loading control. Statistical significance was evaluated using the one-tailed t-test (since four replicates were insufficient for the Wilcoxon rank-sum test).

Msi2 cJun assay

K562 cells were originally obtained from American Type Culture Collection (ATCC, Manassas, VA) and maintained in RPMI1640 media supplemented with 10% Fetal Bovine Serum (FBS). K562-Msi2 knock-down cells and controls were made by lentiviral shRNA vectors using the TRC 1.5 vectors available through Sigma-Aldrich through standard transduction, puromycin selection and single cell cloning by plating on methylcellulose. Clones were analyzed for Msi2 expression by immunoblot with anti-Msi2 rabbit monoclonal antibody (EMD Millipore, Billerica, MA). cJun immunoblots on K562 cells were performed with anti-JUN rabbit monoclonal antibody (Cell Signaling, Dovers, MA). β-actin antibody conjugated to horse-radish peroxidase (HRP, Genscript, Piscataway, NJ) was used as loading control. To determine effect of R12 derivatives, K562 cells were plated at $1x10^6$ cells per 6-well plate and treated with compounds at concentrations of 10, 20 or 40 μM for 48 hours along with controls (DMSO only, inactive analog R4, or no treatment). Cells were harvested at 48 hours and analysed for cJun and β-actin expression.

4.2 Supporting Tables

Name	Sequence
FC-NUMB	5’- F-GUAGU -3’
NUMB	5’- GUAGU -3’
NUMBa0 (WT)	5’- UGUAGUU -3’
NUMBa1 (G104x)	5’- UxUAGUU -3’
NUMBa2 (U105x)	5’- UGxAGUU -3’
NUMBa3 (A106x)	5’- UGUxGUU -3’
NUMBa4 (G107x)	5’- UGUAxUU -3’
NUMBa5 (U108x)	5’- UGUAUxU -3’

Table S1: Sequences of RNA oligonucleotides used in this study. “F” refers to the fluorescein label, and “x” refers to an abasic site (i.e. internal RNA spacer site). After validation to ensure binding to Msi1 RBD1 (**Figure S8**), the FC-NUMB construct was used in fluorescence competition assays (**Figure 1d**, **Figure 2g**).

4.3 Supporting Figures

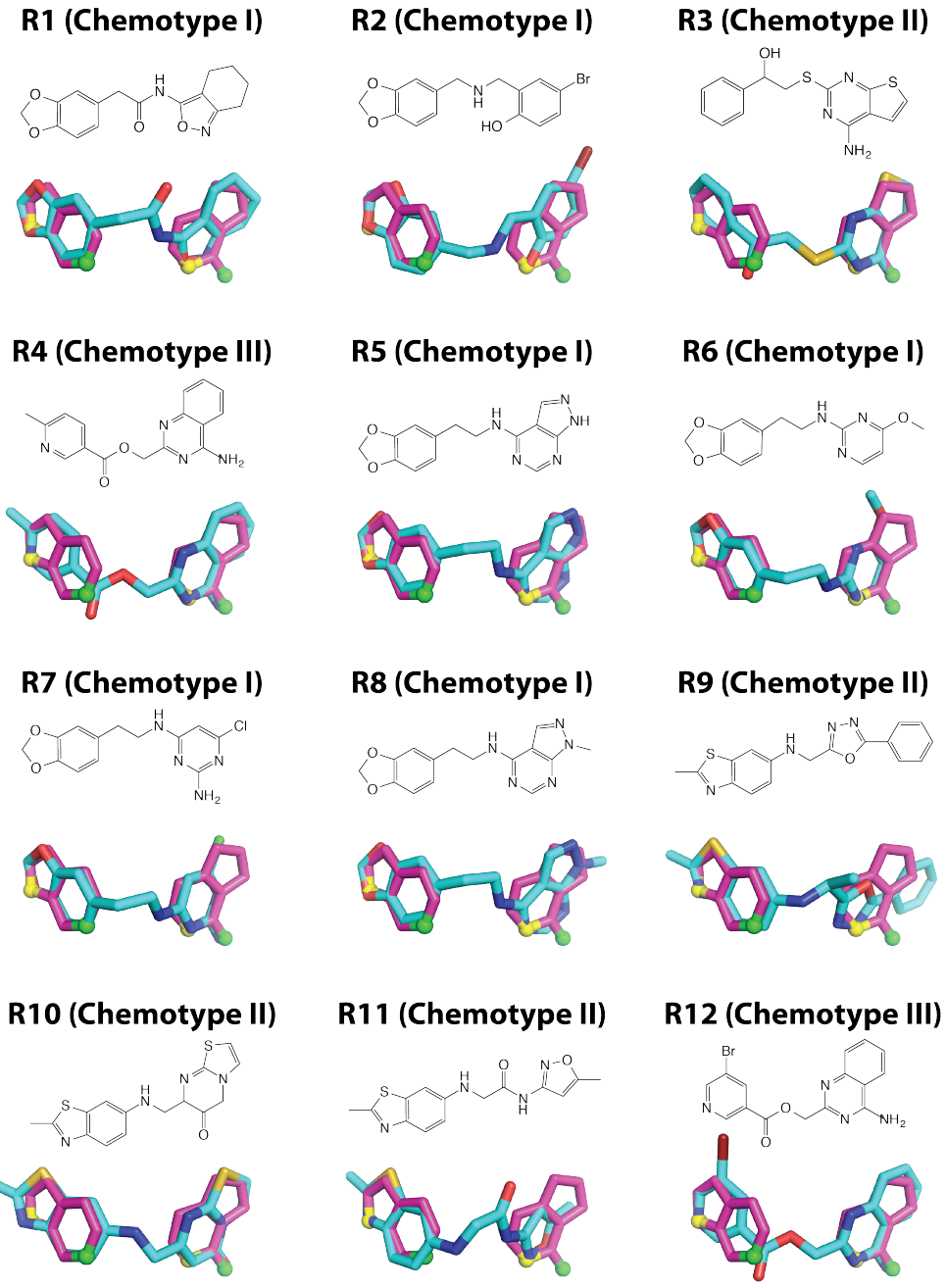


Figure S1: The 12 initial hit compounds. The chemical structure is shown for each compound, as well as a three-dimensional model of each compound (*cyan*) superposed with the Msi1 RBD1 hotspot pharmacophore (*magenta*).

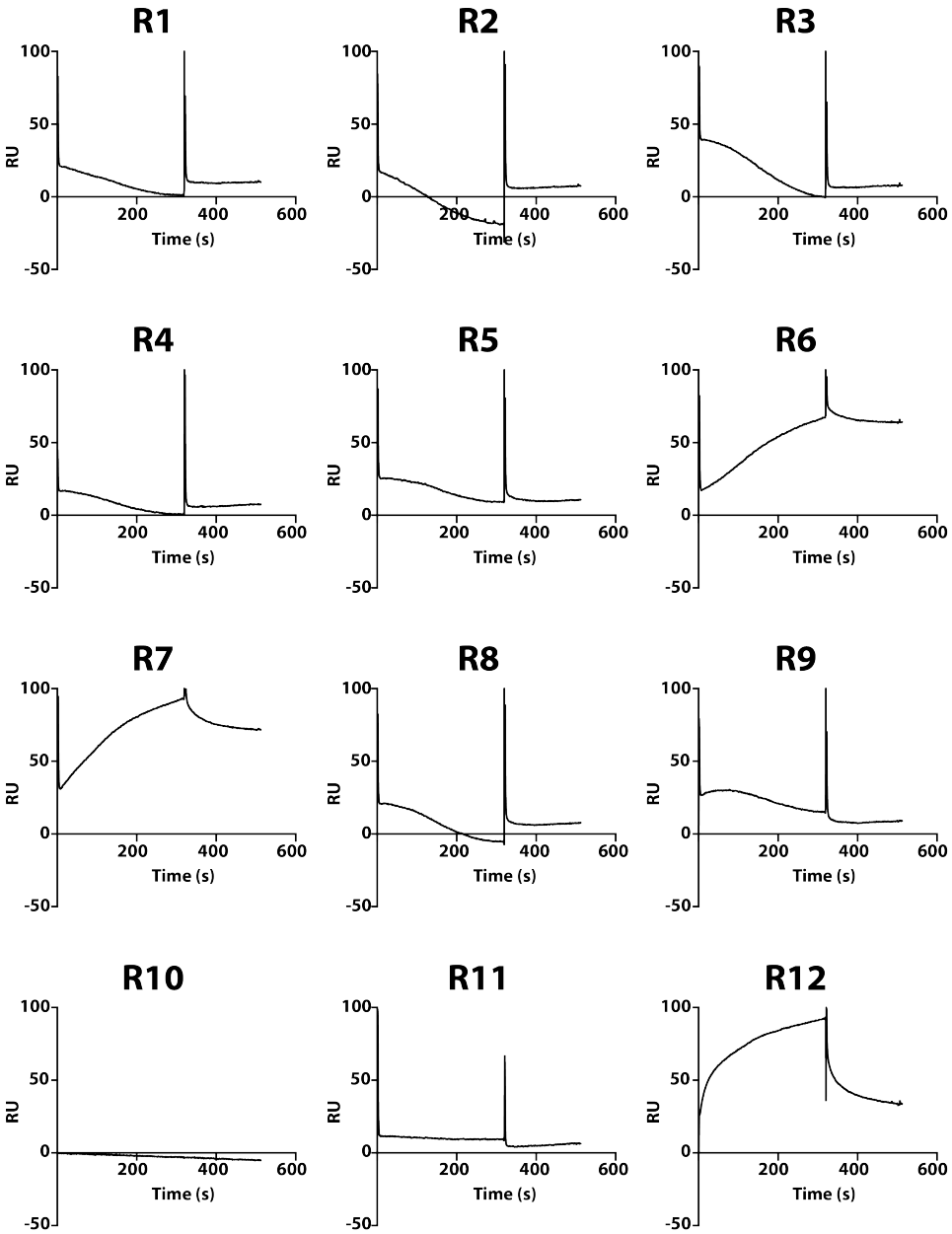


Figure S2: Initial SPR screening. Surface plasmon resonance was used to test for binding of all 12 compounds, with immobilized Msi1 RBD1. At a concentration of 50 μM, none of the compounds except

for R12 showed a kinetic profile consistent with reversible binding. All sensorgrams shown have been reference-subtracted using a flow cell with an unrelated protein immobilized (human Mcl-1).

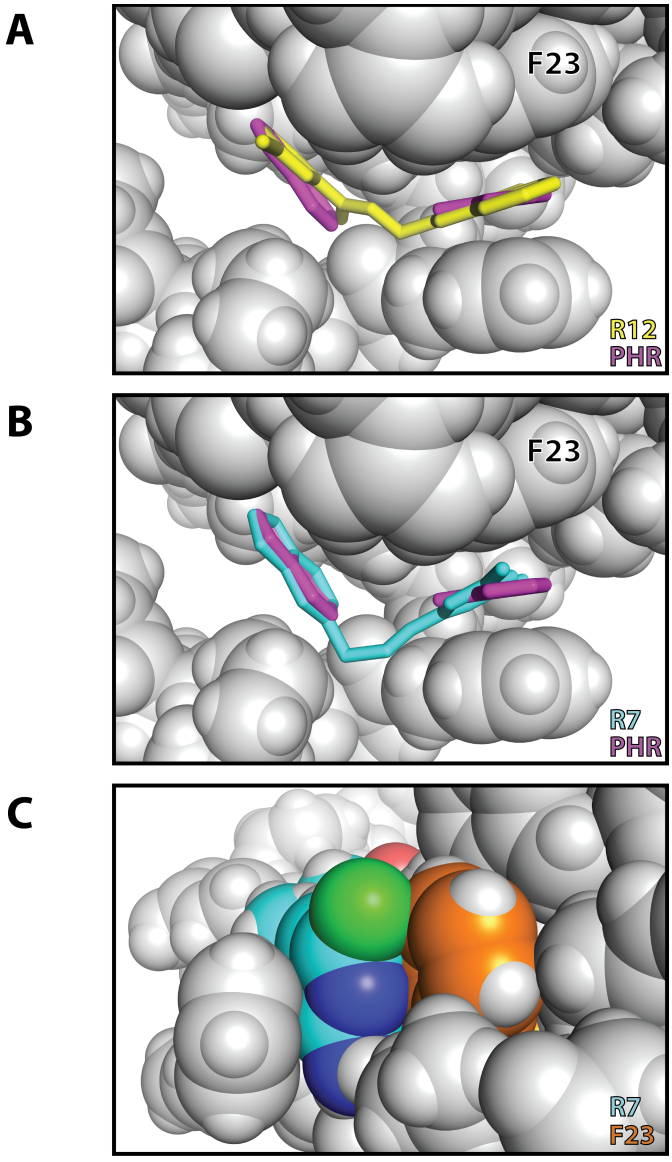


Figure S3: An inadvertent steric clash may explain the lack of binding by R7. (A) The rings in the model of R12 (*yellow*) are well-superposed with those of the hotspot pharmacophore (*magenta*), allowing for aromatic stacking with Msi1. **(B)** The relative positioning of the rings in the R7 (*cyan*) do not quite

align with the hotspot pharmacophore (*right side of this perspective*). **(C)** This difference in the positioning of the ring leads to a steric clash with Phe23 (*orange*).

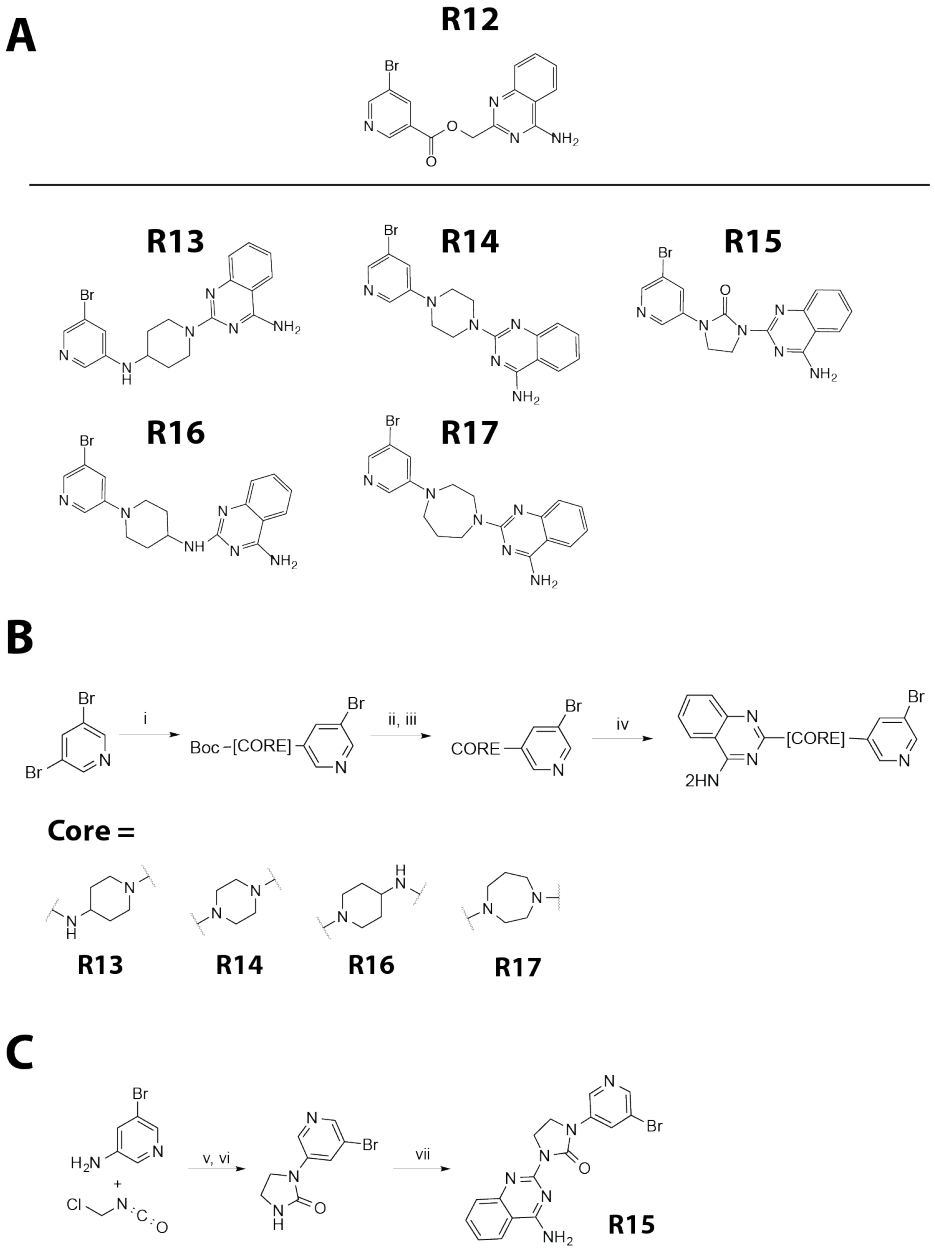


Figure S4: Synthesis of R12 derivatives. (A) Chemical structures of all derivatives R13-R17. **(B)** Scheme leading to synthesis of all derivatives R13-R17 except R15. Reagents and conditions: (i) $Pd_2(dba)_3$, *t*-BuONa, BINAP, 80 ºC, toluene, BocCOREH, 59-91%. (ii) Dioxane, HCl, >95% yield.

(iii) Aqueous $NaCO_3$, CH_2Cl_2, >95%. (iv) CH_3CN, 2-chloroquinazolin-4-amine, 180 ºC, MW, 58-71%. **(C)** Scheme leading to synthesis of R15. Reagents and conditions: (v) DIEA, CH_3CN, 96%. (vi) NaH, THF, 83%. (viii) NaH, THF, 69.1%.

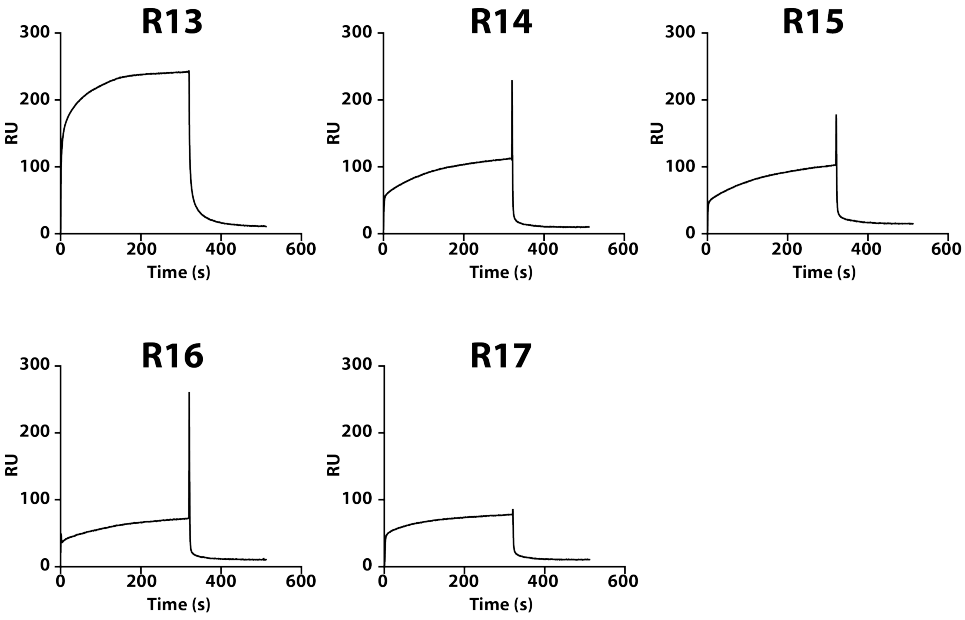


Figure S5: SPR screening of R12 analogs. Surface plasmon resonance was used to test for binding of all five R12 analogs, with immobilized Msi1 RBD1. At a concentration of 50 μM, all five showed a kinetic profile consistent with reversible binding. All sensorgrams shown have been reference-subtracted using a flow cell with an unrelated protein immobilized (human Mcl-1).

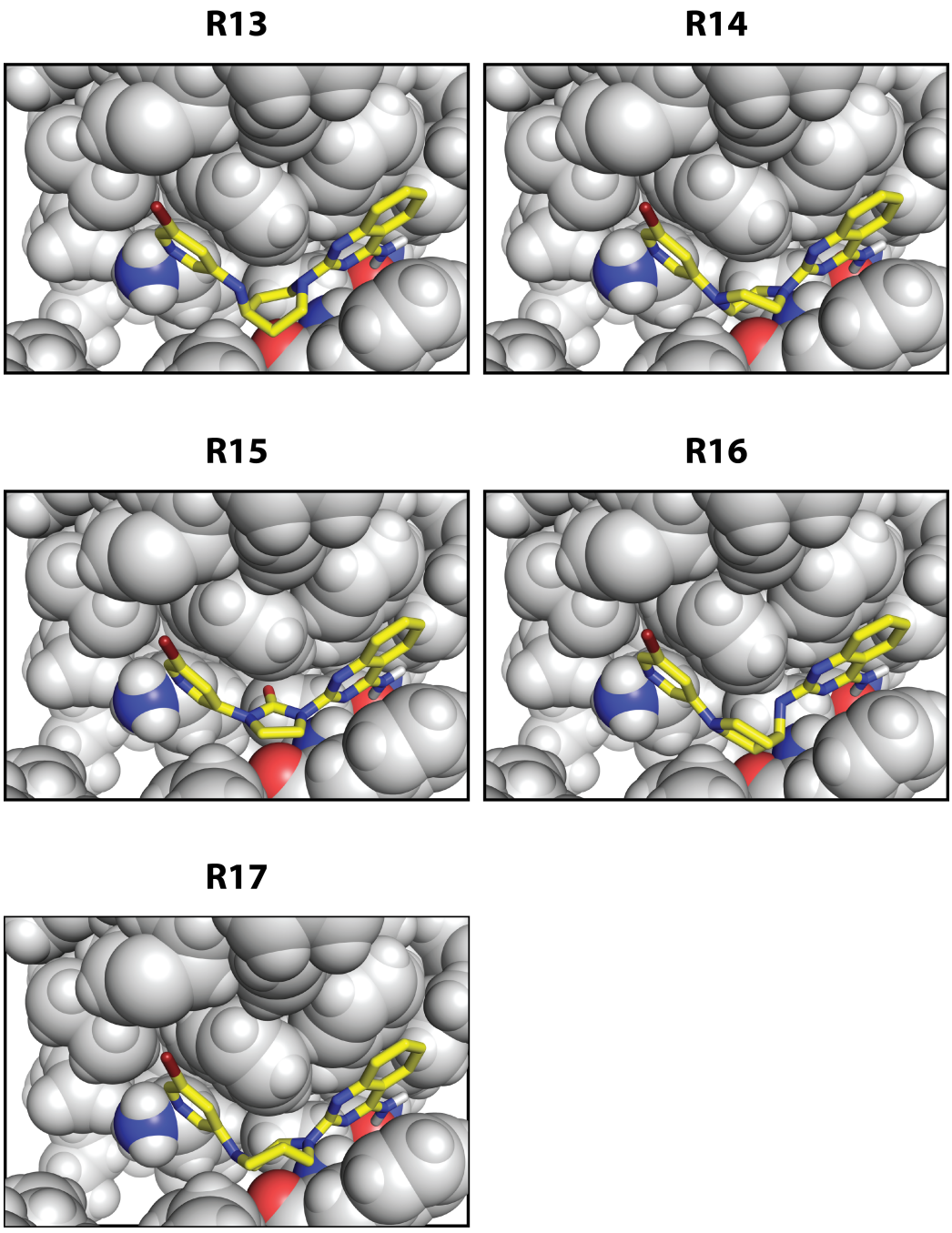


Figure S6: Models of R12 analogs in complex with Msi1. Structures were generated by building conformations of each compound that closely superpose with the rings of R12, then aligning this conformation to the hotspot pharmacophore.

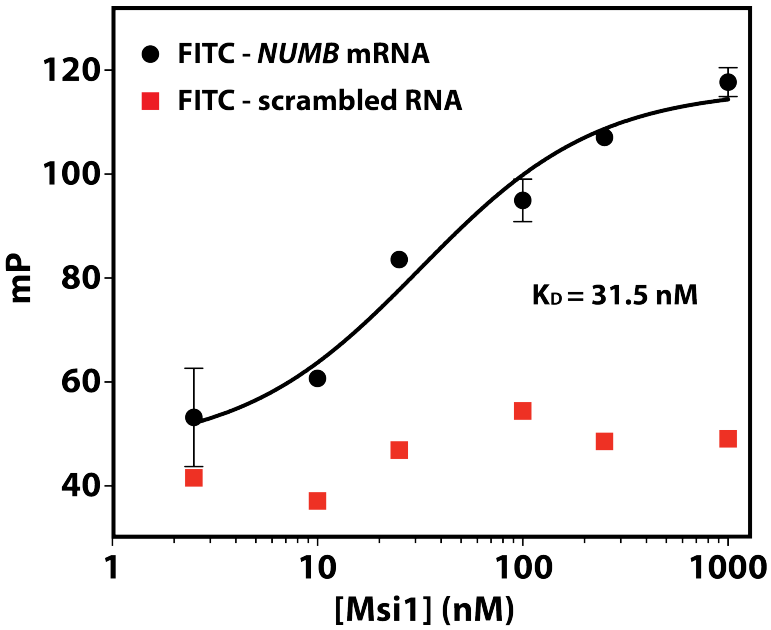


Figure S7: Binding of fluorescein-labeled RNA to Msi1. The polarization from the fluorescein tag increases upon addition of Msi1 RBD1, implying that the protein binds to the RNA. The analogous polarization change is not observed when Msi1 RBD1 is added to a mixture of fluorescein-labeled RNAs with random sequences (**Table S1**).

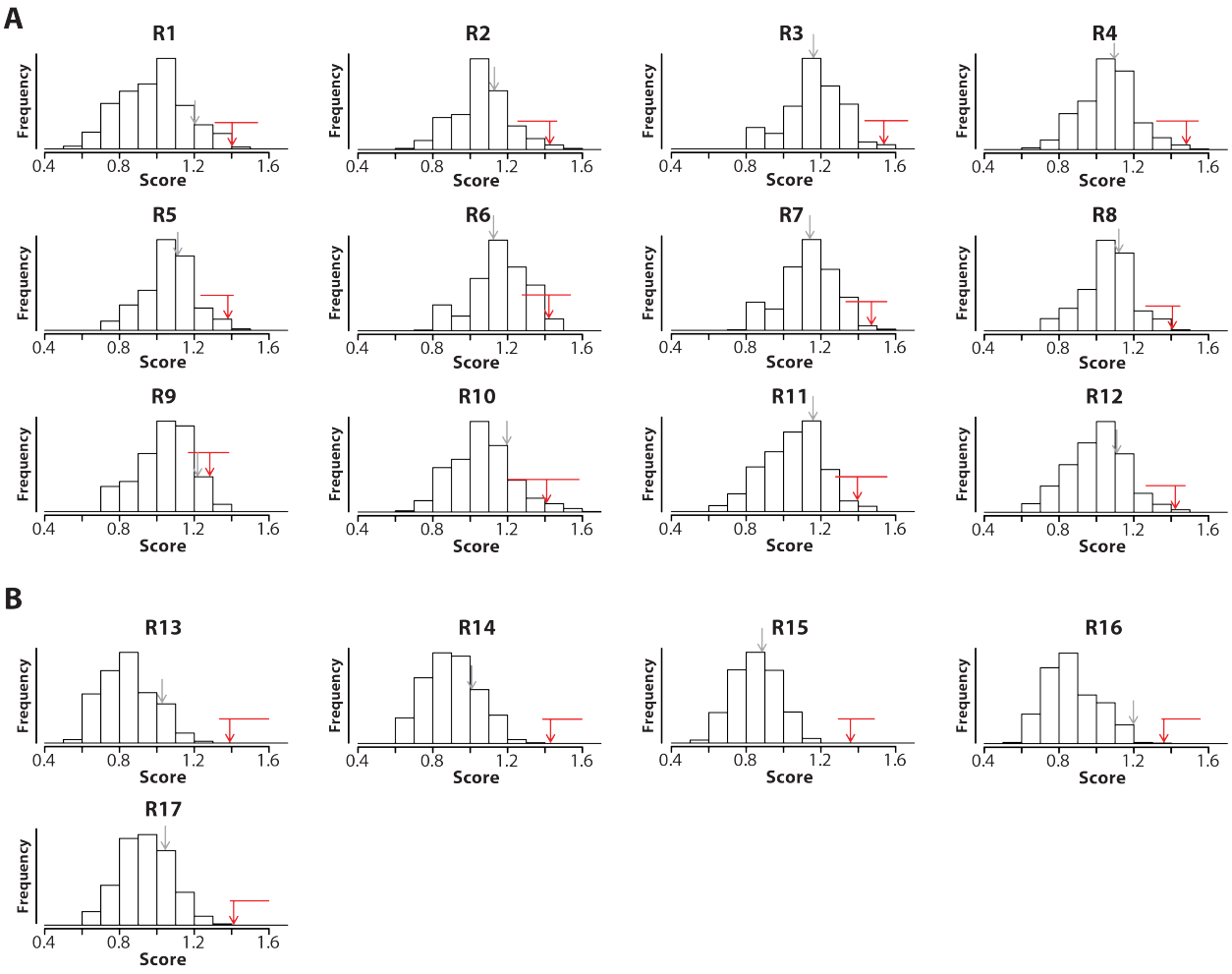


Figure S8: Computational prediction of potential off-target interactions. We screened each conformer of a given ligand against a large set of hotspot pharmacophores derived from other protein-RNA complexes, as described in **Figure 3**. **(A)** Compounds from our initial computational screen. **(B)** Derivatives of R12.

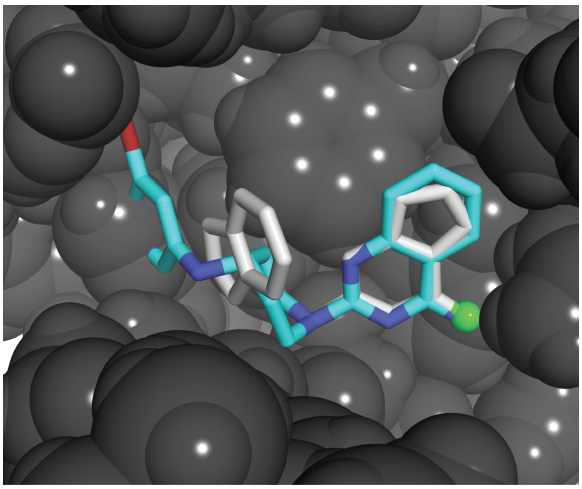


Figure S9: Superposition of R13 onto the hotspot pharmacophore from Tra2-β1. The RRM domain of Tra2-β1 recognizes a purine-purine pair in a dramatically different geometry from that used by Msi1. Superposition of R13 (*cyan*) onto the Tra2-β1 hotspot pharmacophore (*white and green*) highlights the poor match, and R13 cannot be accommodated on the surface of Tra2-β1 without obvious steric clashes.

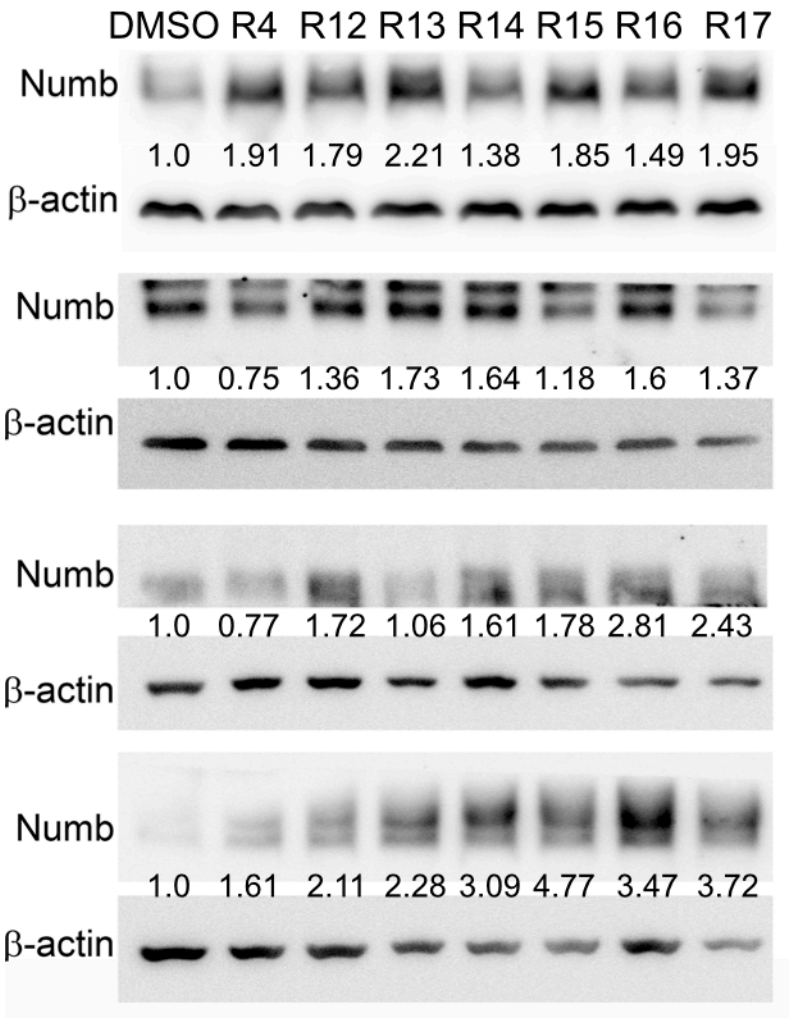


Figure S10: Quantification of Numb protein levels. Levels of Numb protein were quantified from the four Western blots shown. Values in **Figure 4b** represent the average net band intensities collected from four independent experiments, with each Numb value normalized to the corresponding β-actin loading control. Numb has 4 different isoforms (all made from mRNAs with the same 3' UTR) that resolve into 2 bands on the gels used here.

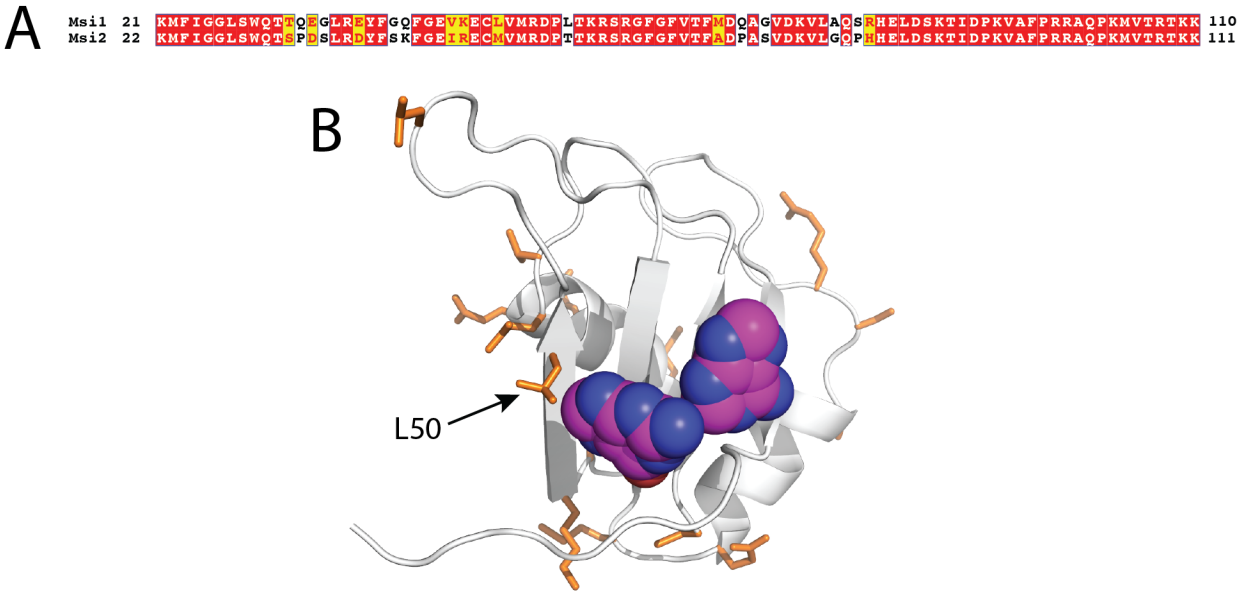


Figure S11: Comparison of Msi1 and Msi2. (A) Sequence alignment comparing RBD1 from Msi1 to RBD1 from Msi2. Over these 90 residues, only 17 positions differ (81% sequence identity) and only 9 correspond to non-conservative mutations (90% sequence similarity). This alignment was rendered using ESPript (272, 273). **(B)** The structure of Msi1 RBD1 is shown (*grey cartoons*), with the hotspot pharmacophore derived from its cognate RNA (*magenta and blue spheres*). Residues at which the sequence differs in Msi2 RBD1 are highlighted (*orange sticks*); with the exception of Leu50 (Met in Msi2), each of the residues that differ are surface exposed and located far from the hotspot pharmacophore.

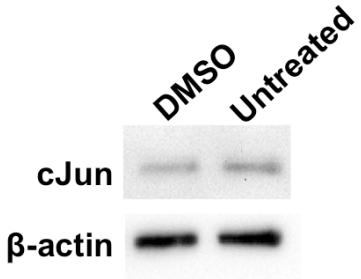


Figure S12: Effect of DMSO on cJun levels. Protein levels of cJun as determined via Western blot (as in **Figure 4c**) are unaffected by the DMSO vehicle used, relative to untreated cells.

References

1. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nature reviews Drug discovery. 2006;5(12):993-6. doi: 10.1038/nrd2199. PubMed PMID: 17139284.

2. Fuller JC, Burgoyne NJ, Jackson RM. Predicting druggable binding sites at the protein-protein interface. Drug discovery today. 2009;14(3-4):155-61. doi: 10.1016/j.drudis.2008.10.009. PubMed PMID: 19041415.

3. Whitty A, Kumaravel G. Between a rock and a hard place? Nature chemical biology. 2006;2(3):112-8. Epub 2006/02/18. doi: 10.1038/nchembio0306-112. PubMed PMID: 16484997.

4. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature. 2007;450(7172):1001-9. doi: 10.1038/nature06526. PubMed PMID: 18075579.

5. Sperandio O, Andrieu O, Miteva MA, Vo MQ, Souaille M, Delfaud F, Villoutreix BO. MED-SuMoLig: a new ligand-based screening tool for efficient scaffold hopping. J Chem Inf Model. 2007;47(3):1097-110. Epub 2007/05/05. doi: 10.1021/ci700031v. PubMed PMID: 17477521.

6. Kruger DM, Evers A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. ChemMedChem. 2010;5(1):148-58. Epub 2009/11/13. doi: 10.1002/cmdc.200900314. PubMed PMID: 19908272.

7. Varani G, Nagai K. RNA recognition by RNP proteins during RNA processing. Annual review of biophysics and biomolecular structure. 1998;27:407-45. doi: 10.1146/annurev.biophys.27.1.407. PubMed PMID: 9646873.

8. Muller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nature reviews Genetics. 2013;14(4):275-87. doi: 10.1038/nrg3434. PubMed PMID: 23478349.

9. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. Protein-RNA interactions: a structural analysis. Nucleic acids research. 2001;29(4):943-54. PubMed PMID: 11160927; PubMed Central PMCID: PMC29619.

10. Ewald B, Sampath D, Plunkett W. Nucleoside analogs: molecular mechanisms signaling cell death. Oncogene. 2008;27(50):6522-37. doi: 10.1038/onc.2008.316. PubMed PMID: 18955977.

11. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. Nature reviews Drug discovery. 2006;5(10):821-34. doi: 10.1038/nrd2132. PubMed PMID: 17016423.

12. Fuller JC, Burgoyne NJ, Jackson RM. Predicting druggable binding sites at the protein-protein interface. Drug discovery today. 2009;14(3-4):155-61. Epub 2008/12/02. doi: 10.1016/j.drudis.2008.10.009. PubMed PMID: 19041415.

13. Naylor E, Arredouani A, Vasudevan SR, Lewis AM, Parkesh R, Mizote A, Rosen D, Thomas JM, Izumi M, Ganesan A, Galione A, Churchill GC. Identification of a chemical probe for NAADP by virtual screening. Nature chemical biology. 2009;5(4):220-6. Epub 2009/02/24. doi: 10.1038/nchembio.150. PubMed PMID: 19234453; PubMed Central PMCID: PMC2659327.

14. Fauman EB, Rai BK, Huang ES. Structure-based druggability assessment--identifying suitable targets for small molecule therapeutics. Curr Opin Chem Biol. 2011;15(4):463-8. Epub 2011/06/28. doi: 10.1016/j.cbpa.2011.05.020. PubMed PMID: 21704549.

15. Turkson J, Jove R. STAT proteins: novel molecular targets for cancer drug discovery. Oncogene. 2000;19(56):6613-26. Epub 2001/06/28. doi: 10.1038/sj.onc.1204086. PubMed PMID: 11426647.

16. Newman JR, Keating AE. Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science. 2003;300(5628):2097-101. Epub 2003/06/14. doi: 10.1126/science.1084648. PubMed PMID: 12805554.

17. Ramakrishnan P, Baltimore D. Sam68 is required for both NF-kappaB activation and apoptosis signaling by the TNF receptor. Mol Cell. 2011;43(2):167-79. Epub 2011/05/31. doi: 10.1016/j.molcel.2011.05.007. PubMed PMID: 21620750; PubMed Central PMCID: PMC3142289.

18. Muto J, Imai T, Ogawa D, Nishimoto Y, Okada Y, Mabuchi Y, Kawase T, Iwanami A, Mischel PS, Saya H, Yoshida K, Matsuzaki Y, Okano H. RNA-binding protein Musashi1 modulates glioma cell growth through the post-transcriptional regulation of Notch and PI3 kinase/Akt signaling pathways. PloS one. 2012;7(3):e33431. doi: 10.1371/journal.pone.0033431. PubMed PMID: 22428049; PubMed Central PMCID: PMC3299785.

19. Rickert M, Wang X, Boulanger MJ, Goriatcheva N, Garcia KC. The structure of interleukin-2 complexed with its alpha receptor. Science. 2005;308(5727):1477-80. Epub 2005/06/04. doi: 10.1126/science.1109745. PubMed PMID: 15933202.

20. Song Z, Yao X, Wu M. Direct interaction between survivin and Smac/DIABLO is essential for the anti-apoptotic activity of survivin during taxol-induced apoptosis. J Biol Chem. 2003;278(25):23130-40. Epub 2003/03/28. doi: 10.1074/jbc.M300957200. PubMed PMID: 12660240.

21. Cochran AG. Antagonists of protein-protein interactions. Chem Biol. 2000;7(4):R85-94. Epub 2000/04/26. PubMed PMID: 10779412.

22. Chonghaile TN, Letai A. Mimicking the BH3 domain to kill cancer cells. Oncogene. 2008;27 Suppl 1:S149-57. Epub 2009/07/31. doi: 10.1038/onc.2009.52. PubMed PMID: 19641500.

23. Stewart KD, Huth JR, Ng TI, McDaniel K, Hutchinson RN, Stoll VS, Mendoza RR, Matayoshi ED, Carrick R, Mo H, Severin J, Walter K, Richardson PL, Barrett LW, Meadows R, Anderson S, Kohlbrenner W, Maring C, Kempf DJ, Molla A, Olejniczak ET. Non-peptide entry inhibitors of HIV-1 that target the gp41 coiled coil pocket. Bioorg Med Chem Lett. 2010;20(2):612-7. Epub 2009/12/17. doi: 10.1016/j.bmcl.2009.11.076. PubMed PMID: 20004576.

24. Bernal F, Wade M, Godes M, Davis TN, Whitehead DG, Kung AL, Wahl GM, Walensky LD. A stapled p53 helix overcomes HDMX-mediated suppression of p53. Cancer cell. 2010;18(5):411-22. Epub 2010/11/16. doi: 10.1016/j.ccr.2010.10.024. PubMed PMID: 21075307; PubMed Central PMCID: PMC3050021.

25. Robinson JA. Beta-hairpin peptidomimetics: design, structures and biological activities. Accounts of chemical research. 2008;41(10):1278-88. Epub 2008/04/17. doi: 10.1021/ar700259k. PubMed PMID: 18412373.

26. Harker EA, Daniels DS, Guarracino DA, Schepartz A. Beta-peptides with improved affinity for hDM2 and hDMX. Bioorg Med Chem. 2009;17(5):2038-46. Epub 2009/02/13. doi: 10.1016/j.bmc.2009.01.039. PubMed PMID: 19211253; PubMed Central PMCID: PMC2926950.

27. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods. 2000;44(1):235-49. Epub 2001/03/29. PubMed PMID: 11274893.

28. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46(1-3):3-26. Epub 2001/03/22. PubMed PMID: 11259830.

29. Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-based virtual screening for drug discovery: a problem-centric review. The AAPS journal. 2012;14(1):133-41. Epub 2012/01/28. doi: 10.1208/s12248-012-9322-0. PubMed PMID: 22281989; PubMed Central PMCID: PMC3282008.

30. Tresadern G, Bemporad D, Howe T. A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. J Mol Graph Model. 2009;27(8):860-70. Epub 2009/02/24. doi: 10.1016/j.jmgm.2009.01.003. PubMed PMID: 19230731.

31. Seifert MH. Targeted scoring functions for virtual screening. Drug Discov Today. 2009;14(11-12):562-9. Epub 2009/06/11. doi: 10.1016/j.drudis.2009.03.013. PubMed PMID: 19508918.

32. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother Of All Databases). Proteins. 2005;60(3):333-40. Epub 2005/06/23. doi: 10.1002/prot.20512. PubMed PMID: 15971202.

33. Smith RD, Hu L, Falkner JA, Benson ML, Nerothin JP, Carlson HA. Exploring protein-ligand recognition with Binding MOAD. J Mol Graph Model. 2006;24(6):414-25. Epub 2005/09/20. doi: 10.1016/j.jmgm.2005.08.002. PubMed PMID: 16168689.

34. Carlson HA, Smith RD, Khazanov NA, Kirchhoff PD, Dunbar JB, Jr., Benson ML. Differences between high- and low-affinity complexes of enzymes and nonenzymes. Journal of medicinal chemistry. 2008;51(20):6432-41. Epub 2008/10/02. doi: 10.1021/jm8006504. PubMed PMID: 18826206; PubMed Central PMCID: PMC2692211.

35. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW. Diverse, high-quality test set for the validation of protein-ligand docking performance. Journal of medicinal chemistry. 2007;50(4):726-41. Epub 2007/02/16. doi: 10.1021/jm061277y. PubMed PMID: 17300160.

36. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. Journal of molecular biology. 1971;55(3):379-400. Epub 1971/02/14. PubMed PMID: 5551392.

37. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. Journal of medicinal chemistry. 2012;55(14):6582-94. Epub 2012/06/22. doi: 10.1021/jm300687e. PubMed PMID: 22716043; PubMed Central PMCID: PMC3405771.

38. Mukherjee S, Balius TE, Rizzo RC. Docking validation resources: protein family and ligand flexibility experiments. Journal of chemical information and modeling. 2010;50(11):1986-2000. doi: 10.1021/ci1001982. PubMed PMID: 21033739; PubMed Central PMCID: PMC3058392.

39. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. P Natl Acad Sci USA. 1999;96(18):9997-10002. Epub 1999/09/01. PubMed PMID: 10468550; PubMed Central PMCID: PMC17830.

40. McGann M. FRED pose prediction and virtual screening accuracy. Journal of chemical information and modeling. 2011;51(3):578-96. Epub 2011/02/18. doi: 10.1021/ci100436p. PubMed PMID: 21323318.

41. McGann M. FRED and HYBRID docking performance on standardized datasets. Journal of computer-aided molecular design. 2012;26(8):897-906. Epub 2012/06/07. doi: 10.1007/s10822-012-9584-8. PubMed PMID: 22669221.

42. FRED version 3.0.0. OpenEye Scientific Software SF, NM. http://www.eyesopen.com.

43. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. Journal of chemical information and modeling. 2012. Epub 2012/05/17. doi: 10.1021/ci3001277. PubMed PMID: 22587354; PubMed Central PMCID: PMC3402020.

44. Hawkins PC, 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. Journal of chemical information and modeling. 2010;50(4):572-84. Epub 2010/03/20. doi: 10.1021/ci100031x. PubMed PMID: 20235588; PubMed Central PMCID: PMC2859685.

45. Hawkins PC, Nicholls A. Conformer generation with OMEGA: learning from the data set and the analysis of failures. Journal of chemical information and modeling. 2012;52(11):2919-36.

46. OMEGA version 2.4.3. OpenEye Scientific Software SF, NM. http://www.eyesopen.com.

47. Brozell SR, Mukherjee S, Balius TE, Roe DR, Case DA, Rizzo RC. Evaluation of DOCK 6 as a pose generation and database enrichment tool. Journal of computer-aided molecular design. 2012;26(6):749-73. Epub 2012/05/10. doi: 10.1007/s10822-012-9565-y. PubMed PMID: 22569593.

48. Johnson DK, Karanicolas J. Druggable protein interaction sites are more predisposed to surface pocket formation than the rest of the protein surface. PLoS computational biology. 2013;9(3):e1002951. Epub 2013/03/19. doi: 10.1371/journal.pcbi.1002951. PubMed PMID: 23505360; PubMed Central PMCID: PMC3591273.

49. Neugebauer A, Hartmann RW, Klein CD. Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods. Journal of medicinal chemistry. 2007;50(19):4665-8. Epub 2007/08/21. doi: 10.1021/jm070533j. PubMed PMID: 17705363.

50. Villoutreix BO, Labbe CM, Lagorce D, Laconde G, Sperandio O. A leap into the chemical space of protein-protein interaction inhibitors. Curr Pharm Des. 2012;18(30):4648-67. Epub 2012/06/02. PubMed PMID: 22650260.

51. Keil M, Exner TE, Brickmann J. Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. Journal of computational chemistry. 2004;25(6):779-89. Epub 2004/03/11. doi: 10.1002/jcc.10361. PubMed PMID: 15011250.

52. Kozakov D, Hall DR, Chuang GY, Cencic R, Brenke R, Grove LE, Beglov D, Pelletier J, Whitty A, Vajda S. Structural conservation of druggable hot spots in protein-protein interfaces. Proc Natl Acad Sci U S A. 2011;108(33):13528-33. Epub 2011/08/03. doi: 10.1073/pnas.1101835108. PubMed PMID: 21808046; PubMed Central PMCID: PMC3158149.

53. Lerner MG, Meagher KL, Carlson HA. Automated clustering of probe molecules from solvent mapping of protein surfaces: new algorithms applied to hot-spot mapping and structure-based drug design. Journal of computer-aided molecular design. 2008;22(10):727-36. Epub 2008/08/06. doi: 10.1007/s10822-008-9231-6. PubMed PMID: 18679808; PubMed Central PMCID: PMC2856601.

54. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW. Discovering high-affinity ligands for proteins: SAR by NMR. Science. 1996;274(5292):1531-4. Epub 1996/11/29. PubMed PMID: 8929414.

55. Valkov E, Sharpe T, Marsh M, Greive S, Hyvonen M. Targeting protein-protein interactions and fragment-based drug discovery. Top Curr Chem. 2012;317:145-79. Epub 2011/10/19. doi: 10.1007/128_2011_265. PubMed PMID: 22006238.

56. Bourgeas R, Basse MJ, Morelli X, Roche P. Atomic analysis of protein-protein interfaces with known inhibitors: the 2P2I database. PloS one. 2010;5(3):e9598. doi: 10.1371/journal.pone.0009598. PubMed PMID: 20231898; PubMed Central PMCID: PMC2834754.

57. Higueruelo AP, Schreyer A, Bickerton GR, Pitt WR, Groom CR, Blundell TL. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database.

Chem Biol Drug Des. 2009;74(5):457-67. Epub 2009/10/09. doi: 10.1111/j.1747-0285.2009.00889.x. PubMed PMID: 19811506.

58. MolProp version 2.1.5. OpenEye Scientific Software SF, NM. http://www.eyesopen.com.

59. QuacPac version 1.5.0. OpenEye Scientific Software SF, NM. http://www.eyesopen.com.

60. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YE, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popovic Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in enzymology. 2011;487:545-74. Epub 2010/12/29. doi: 10.1016/B978-0-12-381270-4.00019-6. PubMed PMID: 21187238.

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

62. Domling A. Small molecular weight protein-protein interaction antagonists: an insurmountable challenge? Current opinion in chemical biology. 2008;12(3):281-91. Epub 2008/05/27. doi: 10.1016/j.cbpa.2008.04.603. PubMed PMID: 18501203.

63. 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.

64. Akram ON, DeGraff DJ, Sheehan JH, Tilley WD, Matusik RJ, Ahn JM, Raj GV. Tailoring peptidomimetics for targeting protein-protein interactions. Molecular cancer research : MCR. 2014;12(7):967-78. doi: 10.1158/1541-7786.MCR-13-0611. PubMed PMID: 24642350.

65. Cummings CG, Hamilton AD. Disrupting protein-protein interactions with non-peptidic, small molecule alpha-helix mimetics. Current opinion in chemical biology. 2010;14(3):341-6. doi: 10.1016/j.cbpa.2010.04.001. PubMed PMID: 20430687.

66. Whitby LR, Boger DL. Comprehensive peptidomimetic libraries targeting protein-protein interactions. Accounts of chemical research. 2012;45(10):1698-709. doi: 10.1021/ar300025n. PubMed PMID: 22799570; PubMed Central PMCID: PMC3473115.

67. Cierpicki T, Grembecka J. Targeting protein-protein interactions in hematologic malignancies: still a challenge or a great opportunity for future therapies? Immunological reviews. 2015;263(1):279-301. doi: 10.1111/imr.12244. PubMed PMID: 25510283.

68. 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.

69. Kar S, Roy K. How far can virtual screening take us in drug discovery? Expert opinion on drug discovery. 2013;8(3):245-61. doi: 10.1517/17460441.2013.761204. PubMed PMID: 23330660.

70. Lionta E, Spyrou G, Vassilatis DK, Cournia Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Current topics in medicinal chemistry. 2014;14(16):1923-38. PubMed PMID: 25262799.

71. Gowthaman R, Deeds EJ, Karanicolas J. Structural properties of non-traditional drug targets present new challenges for virtual screening. Journal of chemical information and modeling. 2013;53(8):2073-81. doi: 10.1021/ci4002316. PubMed PMID: 23879197; PubMed Central PMCID: PMC3819422.

72. Cihlar T, Ray AS. Nucleoside and nucleotide HIV reverse transcriptase inhibitors: 25 years after zidovudine. Antiviral research. 2010;85(1):39-58. doi: 10.1016/j.antiviral.2009.09.014. PubMed PMID: 19887088.

73. Hodgson DR, Schroder M. Chemical approaches towards unravelling kinase-mediated signalling pathways. Chemical Society reviews. 2011;40(3):1211-23. doi: 10.1039/c0cs00020e. PubMed PMID: 21152652.

74. Jacobson KA. Structure-based approaches to ligands for G-protein-coupled adenosine and P2Y receptors, from small molecules to nanoconjugates. Journal of medicinal chemistry. 2013;56(10):3749-67. doi: 10.1021/jm400422s. PubMed PMID: 23597047; PubMed Central PMCID: PMC3701956.

75. Johnson DK, Karanicolas J. Druggable protein interaction sites are more predisposed to surface pocket formation than the rest of the protein surface. PLoS computational biology. 2013;9(3):e1002951. doi: 10.1371/journal.pcbi.1002951. PubMed PMID: 23505360; PubMed Central PMCID: PMC3591273.

76. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. Journal of molecular graphics & modelling. 1997;15(6):359-63, 89. Epub 1998/08/15. PubMed PMID: 9704298.

77. Khar KR, Goldschmidt L, Karanicolas J. Fast docking on graphics processing units via Ray-Casting. PloS one. 2013;8(8):e70661. doi: 10.1371/journal.pone.0070661. PubMed PMID: 23976948; PubMed Central PMCID: PMC3745428.

78. Higueruelo AP, Schreyer A, Bickerton GR, Pitt WR, Groom CR, Blundell TL. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. Chemical biology & drug design. 2009;74(5):457-67. doi: 10.1111/j.1747-0285.2009.00889.x. PubMed PMID: 19811506.

79. Higueruelo AP, Jubb H, Blundell TL. TIMBAL v2: update of a database holding small molecules modulating protein-protein interactions. Database : the journal of biological databases and curation. 2013;2013:bat039. doi: 10.1093/database/bat039. PubMed PMID: 23766369; PubMed Central PMCID: PMC3681332.

80. Willett P. Similarity-based approaches to virtual screening. Biochem Soc Trans. 2003;31(Pt 3):603-6. Epub 2003/05/30. doi: 10.1042/. PubMed PMID: 12773164.

81. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. Journal of medicinal chemistry. 2012;55(14):6582-94. doi: 10.1021/jm300687e. PubMed PMID: 22716043; PubMed Central PMCID: PMC3405771.

82. Nicholls A. What do we know and when do we know it? Journal of computer-aided molecular design. 2008;22(3-4):239-55. Epub 2008/02/07. doi: 10.1007/s10822-008-9170-2. PubMed PMID: 18253702; PubMed Central PMCID: PMC2270923.

83. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, Rizzo RC, Case DA, James TL, Kuntz ID. DOCK 6: combining techniques to model RNA-small molecule complexes. Rna. 2009;15(6):1219-30. doi: 10.1261/rna.1563609. PubMed PMID: 19369428; PubMed Central PMCID: PMC2685511.

84. Brozell SR, Mukherjee S, Balius TE, Roe DR, Case DA, Rizzo RC. Evaluation of DOCK 6 as a pose generation and database enrichment tool. Journal of computer-aided molecular design. 2012;26(6):749-73. doi: 10.1007/s10822-012-9565-y. PubMed PMID: 22569593; PubMed Central PMCID: PMC3902891.

85. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30(16):2785-91. Epub 2009/04/29. doi: 10.1002/jcc.21256. PubMed PMID: 19399780; PubMed Central PMCID: PMC2760638.

86. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, Barril X, Hubbard RE, Morley SD. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS computational biology. 2014;10(4):e1003571. doi: 10.1371/journal.pcbi.1003571. PubMed PMID: 24722481; PubMed Central PMCID: PMC3983074.

87. Korb O, Stutzle T, Exner TE. Empirical scoring functions for advanced protein-ligand docking with PLANTS. Journal of chemical information and modeling. 2009;49(1):84-96. doi: 10.1021/ci800298z. PubMed PMID: 19125657.

88. 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. Journal of medicinal chemistry. 2005;48(5):1489-95. doi: 10.1021/jm040163o. PubMed PMID: 15743191.

89. Hawkins PC, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. Journal of medicinal chemistry. 2007;50(1):74-82. Epub 2007/01/05. doi: 10.1021/jm0603365. PubMed PMID: 17201411.

90. ROCS version 3.2.0.3. OpenEye Scientific Software SF, NM. http://www.eyesopen.com.

91. McGovern SL, Shoichet BK. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. Journal of medicinal chemistry. 2003;46(14):2895-907. Epub 2003/06/27. doi: 10.1021/jm0300330. PubMed PMID: 12825931.

92. Claussen H, Buning C, Rarey M, Lengauer T. FlexE: efficient molecular docking considering protein structure variations. J Mol Biol. 2001;308(2):377-95. doi: 10.1006/jmbi.2001.4551. PubMed PMID: 11327774.

93. Cavasotto CN, Abagyan RA. Protein flexibility in ligand docking and virtual screening to protein kinases. J Mol Biol. 2004;337(1):209-25. doi: 10.1016/j.jmb.2004.01.003. PubMed PMID: 15001363.

94. Osguthorpe DJ, Sherman W, Hagler AT. Exploring protein flexibility: incorporating structural ensembles from crystal structures and simulation into virtual screening protocols. The journal of physical chemistry B. 2012;116(23):6952-9. doi: 10.1021/jp3003992. PubMed PMID: 22424156; PubMed Central PMCID: PMC3376248.

95. Huang SY, Zou X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. Proteins. 2007;66(2):399-421. doi: 10.1002/prot.21214. PubMed PMID: 17096427.

96. Cosconati S, Marinelli L, Di Leva FS, La Pietra V, De Simone A, Mancini F, Andrisano V, Novellino E, Goodsell DS, Olson AJ. Protein flexibility in virtual screening: the BACE-1 case study. J Chem Inf Model. 2012;52(10):2697-704. doi: 10.1021/ci300390h. PubMed PMID: 23005250; PubMed Central PMCID: PMC3523752.

97. Cavasotto CN, Kovacs JA, Abagyan RA. Representing receptor flexibility in ligand docking through relevant normal modes. Journal of the American Chemical Society. 2005;127(26):9632-40. doi: 10.1021/ja042260c. PubMed PMID: 15984891.

98. Dixit A, Verkhivker GM. Integrating ligand-based and protein-centric virtual screening of kinase inhibitors using ensembles of multiple protein kinase genes and conformations. J Chem Inf Model. 2012;52(10):2501-15. doi: 10.1021/ci3002638. PubMed PMID: 22992037.

99. Cheng LS, Amaro RE, Xu D, Li WW, Arzberger PW, McCammon JA. Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. Journal of medicinal chemistry. 2008;51(13):3878-94. doi: 10.1021/jm8001197. PubMed PMID: 18558668; PubMed Central PMCID: PMC2652358.

100. Shoemaker AR, Oleksijew A, Bauch J, Belli BA, Borre T, Bruncko M, Deckwirth T, Frost DJ, Jarvis K, Joseph MK, Marsh K, McClellan W, Nellans H, Ng S, Nimmer P, O'Connor JM, Oltersdorf T, Qing W, Shen W, Stavropoulos J, Tahir SK, Wang B, Warner R, Zhang H, Fesik SW, Rosenberg SH, Elmore SW. A small-molecule inhibitor of Bcl-XL potentiates the activity of cytotoxic drugs in vitro and in vivo. Cancer Research. 2006;66(17):8731-9. Epub 2006/09/05. doi: 66/17/8731 [pii] 10.1158/0008-5472.CAN-06-0367. PubMed PMID: 16951189.

101. Wang G, Nikolovska-Coleska Z, Yang CY, Wang R, Tang G, Guo J, Shangary S, Qiu S, Gao W, Yang D, Meagher J, Stuckey J, Krajewski K, Jiang S, Roller PP, Abaan HO, Tomita Y, Wang S. Structure-based design of potent small-molecule inhibitors of anti-apoptotic Bcl-2 proteins. Journal of medicinal chemistry. 2006;49(21):6139-42. Epub 2006/10/13. doi: 10.1021/jm060460o. PubMed PMID: 17034116.

102. van Delft MF, Wei AH, Mason KD, Vandenberg CJ, Chen L, Czabotar PE, Willis SN, Scott CL, Day CL, Cory S, Adams JM, Roberts AW, Huang DC. The BH3 mimetic ABT-737 targets selective Bcl-2 proteins and efficiently induces apoptosis via Bak/Bax if Mcl-1 is neutralized. Cancer Cell. 2006;10(5):389-99. Epub 2006/11/14. doi: S1535-6108(06)00291-1 [pii]

10.1016/j.ccr.2006.08.027. PubMed PMID: 17097561.

103. Konopleva M, Contractor R, Tsao T, Samudio I, Ruvolo PP, Kitada S, Deng X, Zhai D, Shi YX, Sneed T, Verhaegen M, Soengas M, Ruvolo VR, McQueen T, Schober WD, Watt JC, Jiffar T, Ling X, Marini FC, Harris D, Dietrich M, Estrov Z, McCubrey J, May WS, Reed JC, Andreeff M. Mechanisms of apoptosis sensitivity and resistance to the BH3 mimetic ABT-737 in acute myeloid leukemia. Cancer Cell. 2006;10(5):375-88. Epub 2006/11/14. doi: S1535-6108(06)00313-8 [pii]
10.1016/j.ccr.2006.10.006. PubMed PMID: 17097560.

104. Oltersdorf T, Elmore SW, Shoemaker AR, Armstrong RC, Augeri DJ, Belli BA, Bruncko M, Deckwerth TL, Dinges J, Hajduk PJ, Joseph MK, Kitada S, Korsmeyer SJ, Kunzer AR, Letai A, Li C, Mitten MJ, Nettesheim DG, Ng S, Nimmer PM, O'Connor JM, Oleksijew A, Petros AM, Reed JC, Shen W, Tahir SK, Thompson CB, Tomaselli KJ, Wang B, Wendt MD, Zhang H, Fesik SW, Rosenberg SH. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. Nature. 2005;435(7042):677-81. doi: 10.1038/nature03579. PubMed PMID: 15902208.

105. Souers AJ, Leverson JD, Boghaert ER, Ackler SL, Catron ND, Chen J, Dayton BD, Ding H, Enschede SH, Fairbrother WJ, Huang DC, Hymowitz SG, Jin S, Khaw SL, Kovar PJ, Lam LT, Lee J, Maecker HL, Marsh KC, Mason KD, Mitten MJ, Nimmer PM, Oleksijew A, Park CH, Park CM, Phillips DC, Roberts AW, Sampath D, Seymour JF, Smith ML, Sullivan GM, Tahir SK, Tse C, Wendt MD, Xiao Y, Xue JC, Zhang H, Humerickhouse RA, Rosenberg SH, Elmore SW. ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. Nature medicine. 2013;19(2):202-8. doi: 10.1038/nm.3048. PubMed PMID: 23291630.

106. Chen S, Dai Y, Harada H, Dent P, Grant S. Mcl-1 down-regulation potentiates ABT-737 lethality by cooperatively inducing Bak activation and Bax translocation. Cancer Research. 2007;67(2):782-91. Epub 2007/01/20. doi: 67/2/782 [pii]
10.1158/0008-5472.CAN-06-3964. PubMed PMID: 17234790.

107. Lemmen C, Lengauer T, Klebe G. FLEXS: a method for fast flexible ligand superposition. Journal of medicinal chemistry. 1998;41(23):4502-20. Epub 1998/11/07. doi: 10.1021/jm981037l. PubMed PMID: 9804690.

108. Belmar J, Fesik SW. Small molecule Mcl-1 inhibitors for the treatment of cancer. Pharmacology & therapeutics. 2015;145C:76-84. doi: 10.1016/j.pharmthera.2014.08.003. PubMed PMID: 25172548.

109. Bajwa N, Liao C, Nikolovska-Coleska Z. Inhibitors of the anti-apoptotic Bcl-2 proteins: a patent review. Expert opinion on therapeutic patents. 2012;22(1):37-55. doi: 10.1517/13543776.2012.644274. PubMed PMID: 22195752; PubMed Central PMCID: PMC3547649.

110. Abulwerdi F, Liao C, Liu M, Azmi AS, Aboukameel A, Mady AS, Gulappa T, Cierpicki T, Owens S, Zhang T, Sun D, Stuckey JA, Mohammad RM, Nikolovska-Coleska Z. A novel small-molecule inhibitor of mcl-1 blocks pancreatic cancer growth in vitro and in vivo. Molecular cancer therapeutics. 2014;13(3):565-75. doi: 10.1158/1535-7163.MCT-12-0767. PubMed PMID: 24019208; PubMed Central PMCID: PMC4174574.

111. Bruncko M, Wang L, Sheppard GS, Phillips DC, Tahir SK, Xue J, Erickson S, Fidanze S, Fry E, Hasvold L, Jenkins GJ, Jin S, Judge RA, Kovar PJ, Madar D, Nimmer P, Park C, Petros AM, Rosenberg SH, Smith ML, Song X, Sun C, Tao ZF, Wang X, Xiao Y, Zhang H, Tse C, Leverson JD, Elmore SW, Souers AJ. Structure-guided design of a series of MCL-1 inhibitors with high affinity and selectivity. Journal of medicinal chemistry. 2015;58(5):2180-94. doi: 10.1021/jm501258m. PubMed PMID: 25679114.

112. Fang C, D'Souza B, Thompson CF, Clifton MC, Fairman JW, Fulroth B, Leed A, McCarren P, Wang L, Wang Y, Feau C, Kaushik VK, Palmer M, Wei G, Golub TR, Hubbard BK, Serrano-Wu MH. Single Diastereomer of a Macrolactam Core Binds Specifically to Myeloid Cell Leukemia 1 (MCL1). ACS medicinal chemistry letters. 2014;5(12):1308-12. doi: 10.1021/ml500388q. PubMed PMID: 25516789; PubMed Central PMCID: PMC4265822.

113. Friberg A, Vigil D, Zhao B, Daniels RN, Burke JP, Garcia-Barrantes PM, Camper D, Chauder BA, Lee T, Olejniczak ET, Fesik SW. Discovery of potent myeloid cell leukemia 1 (Mcl-1) inhibitors using fragment-based methods and structure-based design. Journal of medicinal chemistry. 2013;56(1):15-30. doi: 10.1021/jm301448p. PubMed PMID: 23244564; PubMed Central PMCID: PMC3646517.

114. Petros AM, Swann SL, Song D, Swinger K, Park C, Zhang H, Wendt MD, Kunzer AR, Souers AJ, Sun C. Fragment-based discovery of potent inhibitors of the anti-apoptotic MCL-1 protein. Bioorganic & medicinal chemistry letters. 2014;24(6):1484-8. doi: 10.1016/j.bmcl.2014.02.010. PubMed PMID: 24582986.

115. Rega MF, Wu B, Wei J, Zhang Z, Cellitti JF, Pellecchia M. SAR by interligand nuclear overhauser effects (ILOEs) based discovery of acylsulfonamide compounds active against Bcl-x(L) and Mcl-1. Journal of medicinal chemistry. 2011;54(17):6000-13. doi: 10.1021/jm200826s. PubMed PMID: 21797225; PubMed Central PMCID: PMC3165075.

116. Prakesch M, Denisov AY, Naim M, Gehring K, Arya P. The discovery of small molecule chemical probes of Bcl-X(L) and Mcl-1. Bioorganic & medicinal chemistry. 2008;16(15):7443-9. doi: 10.1016/j.bmc.2008.06.023. PubMed PMID: 18603434.

117. Ding X, Li Y, Lv L, Zhou M, Han L, Zhang Z, Ba Q, Li J, Wang H, Liu H, Wang R. De novo design, synthesis and evaluation of benzylpiperazine derivatives as highly selective binders of Mcl-1. ChemMedChem. 2013;8(12):1986-2014. doi: 10.1002/cmdc.201300316. PubMed PMID: 24124106.

118. Cao X, Yap JL, Newell-Rogers MK, Peddaboina C, Jiang W, Papaconstantinou HT, Jupitor D, Rai A, Jung KY, Tubin RP, Yu W, Vanommeslaeghe K, Wilder PT, MacKerell AD, Jr., Fletcher S, Smythe RW. The novel BH3 alpha-helix mimetic JY-1-106 induces apoptosis in a subset of cancer cells (lung cancer, colon cancer and mesothelioma) by disrupting Bcl-xL and Mcl-1 protein-protein interactions with Bak. Molecular cancer. 2013;12(1):42. doi: 10.1186/1476-4598-12-42. PubMed PMID: 23680104; PubMed Central PMCID: PMC3663763.

119. Kazi A, Sun J, Doi K, Sung SS, Takahashi Y, Yin H, Rodriguez JM, Becerril J, Berndt N, Hamilton AD, Wang HG, Sebti SM. The BH3 alpha-helical mimic BH3-M6 disrupts Bcl-X(L), Bcl-2, and MCL-1 protein-protein interactions with Bax, Bak, Bad, or Bim and induces apoptosis in a Bax- and Bim-dependent manner. The Journal of biological chemistry. 2011;286(11):9382-92. doi: 10.1074/jbc.M110.203638. PubMed PMID: 21148306; PubMed Central PMCID: PMC3059047.

120. Stewart ML, Fire E, Keating AE, Walensky LD. The MCL-1 BH3 helix is an exclusive MCL-1 inhibitor and apoptosis sensitizer. Nature chemical biology. 2010;6(8):595-601. doi: 10.1038/nchembio.391. PubMed PMID: 20562877; PubMed Central PMCID: PMC3033224.

121. Muppidi A, Doi K, Edwardraja S, Drake EJ, Gulick AM, Wang HG, Lin Q. Rational design of proteolytically stable, cell-permeable peptide-based selective Mcl-1 inhibitors. Journal of the American Chemical Society. 2012;134(36):14734-7. doi: 10.1021/ja306864v. PubMed PMID: 22920569; PubMed Central PMCID: PMC3472523.

122. Muppidi A, Doi K, Ramil CP, Wang HG, Lin Q. Synthesis of cell-permeable stapled BH3 peptide-based Mcl-1 inhibitors containing simple aryl and vinylaryl cross-linkers. Tetrahedron. 2014;70(42):7740-5. doi: 10.1016/j.tet.2014.05.104. PubMed PMID: 25267861; PubMed Central PMCID: PMC4175436.

123. Fire E, Gulla SV, Grant RA, Keating AE. Mcl-1-Bim complexes accommodate surprising point mutations via minor structural changes. Protein science : a publication of the Protein Society. 2010;19(3):507-19. doi: 10.1002/pro.329. PubMed PMID: 20066663; PubMed Central PMCID: PMC2866276.

124. Johnson DK, Karanicolas J. Selectivity by small-molecule inhibitors of protein interactions can be driven by protein surface fluctuations. PLOS Comput Biol. in press.

125. Liu G, Poppe L, Aoki K, Yamane H, Lewis J, Szyperski T. High-quality NMR structure of human anti-apoptotic protein domain Mcl-1(171-327) for cancer drug design. PloS one.

2014;9(5):e96521. doi: 10.1371/journal.pone.0096521. PubMed PMID: 24789074; PubMed Central PMCID: PMC4008586.

126. Davis IW, Baker D. RosettaLigand docking with full ligand and receptor flexibility. J Mol Biol. 2009;385(2):381-92. doi: 10.1016/j.jmb.2008.11.010. PubMed PMID: 19041878.

127. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. Journal of medicinal chemistry. 2010;53(7):2719-40. doi: 10.1021/jm901137j. PubMed PMID: 20131845.

128. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nature biotechnology. 2007;25(2):197-206. doi: 10.1038/nbt1284. PubMed PMID: 17287757.

129. Wolter KG, Verhaegen M, Fernandez Y, Nikolovska-Coleska Z, Riblett M, de la Vega CM, Wang S, Soengas MS. Therapeutic window for melanoma treatment provided by selective effects of the proteasome on Bcl-2 proteins. Cell death and differentiation. 2007;14(9):1605-16. doi: 10.1038/sj.cdd.4402163. PubMed PMID: 17541428.

130. Tanaka Y, Aikawa K, Nishida G, Homma M, Sogabe S, Igaki S, Hayano Y, Sameshima T, Miyahisa I, Kawamoto T, Tawada M, Imai Y, Inazuka M, Cho N, Imaeda Y, Ishikawa T. Discovery of potent Mcl-1/Bcl-xL dual inhibitors by using a hybridization strategy based on structural analysis of target proteins. Journal of medicinal chemistry. 2013;56(23):9635-45. doi: 10.1021/jm401170c. PubMed PMID: 24215352.

131. Srinivasan B, Johnson TE, Lad R, Xing C. Structure-activity relationship studies of chalcone leading to 3-hydroxy-4,3',4',5'-tetramethoxychalcone and its analogues as potent nuclear factor kappaB inhibitors and their anticancer activities. Journal of medicinal chemistry. 2009;52(22):7228-35. doi: 10.1021/jm901278z. PubMed PMID: 19883086.

132. Battenberg OA, Yang Y, Verhelst SH, Sieber SA. Target profiling of 4-hydroxyderricin in S. aureus reveals seryl-tRNA synthetase binding and inhibition by covalent modification. Molecular bioSystems. 2013;9(3):343-51. doi: 10.1039/c2mb25446h. PubMed PMID: 23295910.

133. Shoichet BK. Interpreting steep dose-response curves in early inhibitor discovery. Journal of medicinal chemistry. 2006;49(25):7274-7. doi: 10.1021/jm061103g. PubMed PMID: 17149857.

134. Ryan AJ, Gray NM, Lowe PN, Chung CW. Effect of detergent on "promiscuous" inhibitors. Journal of medicinal chemistry. 2003;46(16):3448-51. doi: 10.1021/jm0340896. PubMed PMID: 12877581.

135. Giannetti AM, Koch BD, Browner MF. Surface plasmon resonance based assay for the detection and characterization of promiscuous inhibitors. Journal of medicinal chemistry. 2008;51(3):574-80. doi: 10.1021/jm700952v. PubMed PMID: 18181566.

136. Coan KE, Maltby DA, Burlingame AL, Shoichet BK. Promiscuous aggregate-based inhibitors promote enzyme unfolding. Journal of medicinal chemistry. 2009;52(7):2067-75. doi: 10.1021/jm801605r. PubMed PMID: 19281222; PubMed Central PMCID: PMC2664636.

137. Amslinger S, Al-Rifai N, Winter K, Wormann K, Scholz R, Baumeister P, Wild M. Reactivity assessment of chalcones by a kinetic thiol assay. Organic & biomolecular chemistry. 2013;11(4):549-54. doi: 10.1039/c2ob27163j. PubMed PMID: 23224077.

138. DeSimone RW, Currie KS, Mitchell SA, Darrow JW, Pippin DA. Privileged structures: applications in drug discovery. Combinatorial chemistry & high throughput screening. 2004;7(5):473-94. PubMed PMID: 15320713.

139. Welsch ME, Snyder SA, Stockwell BR. Privileged scaffolds for library design and drug discovery. Current opinion in chemical biology. 2010;14(3):347-61. doi: 10.1016/j.cbpa.2010.02.018. PubMed PMID: 20303320; PubMed Central PMCID: PMC2908274.

140. Varadarajan S, Vogler M, Butterworth M, Dinsdale D, Walensky LD, Cohen GM. Evaluation and critical assessment of putative MCL-1 inhibitors. Cell death and differentiation. 2013;20(11):1475-84. doi: 10.1038/cdd.2013.79. PubMed PMID: 23832116; PubMed Central PMCID: PMC3792441.

141. Eichhorn JM, Alford SE, Hughes CC, Fenical W, Chambers TC. Purported Mcl-1 inhibitor marinopyrrole A fails to show selective cytotoxicity for Mcl-1-dependent cell lines. Cell death & disease. 2013;4:e880. doi: 10.1038/cddis.2013.411. PubMed PMID: 24157874; PubMed Central PMCID: PMC3920948.

142. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proceedings of the National Academy of Sciences of the United States of America. 1992;89(6):2195-9. Epub 1992/03/15. PubMed PMID: 1549581; PubMed Central PMCID: PMC48623.

143. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol. 1997;272(1):106-20. Epub 1997/09/23. doi: S0022-2836(97)91203-1 [pii]
10.1006/jmbi.1997.1203. PubMed PMID: 9299341.

144. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics. 2005;21(10):2347-55. Epub 2005/02/25. doi: bti337 [pii]
10.1093/bioinformatics/bti337. PubMed PMID: 15728116.

145. Cai W, Shao X, Maigret B. Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. Journal of molecular graphics & modelling. 2002;20(4):313-28. Epub 2002/02/23. doi: S1093-3263(01)00134-6 [pii]. PubMed PMID: 11858640.

146. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic acids research. 2005;33(Web Server issue):W363-7. doi: 10.1093/nar/gki481. PubMed PMID: 15980490; PubMed Central PMCID: PMC1160241.

147. Lee HS, Lee CS, Kim JS, Kim DH, Choe H. Improving virtual screening performance against conformational variations of receptors by shape matching with ligand binding pocket. J Chem Inf Model. 2009;49(11):2419-28. Epub 2009/10/27. doi: 10.1021/ci9002365. PubMed PMID: 19852439.

148. Ebalunode JO, Ouyang Z, Liang J, Zheng W. Novel approach to structure-based pharmacophore search using computational geometry and shape matching techniques. J Chem Inf Model. 2008;48(4):889-901. Epub 2008/04/10. doi: 10.1021/ci700368p. PubMed PMID: 18396858.

149. Hoffmann B, Zaslavskiy M, Vert JP, Stoven V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. BMC bioinformatics. 2010;11:99. doi: 10.1186/1471-2105-11-99. PubMed PMID: 20175916; PubMed Central PMCID: PMC2838872.

150. Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D. Characterization of local geometry of protein surfaces with the visibility criterion. Proteins. 2008;71(2):670-83. doi: 10.1002/prot.21732. PubMed PMID: 17975834.

151. Chikhi R, Sael L, Kihara D. Real-time ligand binding pocket database search using local surface descriptors. Proteins. 2010;78(9):2007-28. Epub 2010/05/11. doi: 10.1002/prot.22715. PubMed PMID: 20455259; PubMed Central PMCID: PMC3009464.

152. Gowthaman R, Lyskov S, Karanicolas J. DARC 2.0 : Improved docking and virtual screening at protein interaction sitesmanuscript in preparation.

153. Fauman EB, Rai BK, Huang ES. Structure-based druggability assessment--identifying suitable targets for small molecule therapeutics. Current opinion in chemical biology. 2011;15(4):463-8. doi: 10.1016/j.cbpa.2011.05.020. PubMed PMID: 21704549.

154. Naylor E, Arredouani A, Vasudevan SR, Lewis AM, Parkesh R, Mizote A, Rosen D, Thomas JM, Izumi M, Ganesan A, Galione A, Churchill GC. Identification of a chemical probe for NAADP by virtual screening. Nature chemical biology. 2009;5(4):220-6. doi: 10.1038/nchembio.150. PubMed PMID: 19234453; PubMed Central PMCID: PMC2659327.

155. Koes D, Khoury K, Huang Y, Wang W, Bista M, Popowicz GM, Wolf S, Holak TA, Domling A, Camacho CJ. Enabling large-scale design, synthesis and validation of small molecule protein-protein antagonists. PloS one. 2012;7(3):e32839. doi: 10.1371/journal.pone.0032839. PubMed PMID: 22427896; PubMed Central PMCID: PMC3299697.

156. Koes DR, Camacho CJ. Small-molecule inhibitor starting points learned from protein-protein interaction inhibitor structure. Bioinformatics. 2012;28(6):784-91. doi: 10.1093/bioinformatics/btr717. PubMed PMID: 22210869; PubMed Central PMCID: PMC3307105.

157. Metz A, Schanda J, Grez M, Wichmann C, Gohlke H. From determinants of RUNX1/ETO tetramerization to small-molecule protein-protein interaction inhibitors targeting acute myeloid leukemia. J Chem Inf Model. 2013;53(9):2197-202. doi: 10.1021/ci400332e. PubMed PMID: 23957251.

158. Tsao DH, Sutherland AG, Jennings LD, Li Y, Rush TS, 3rd, Alvarez JC, Ding W, Dushin EG, Dushin RG, Haney SA, Kenny CH, Malakian AK, Nilakantan R, Mosyak L. Discovery of novel inhibitors of the ZipA/FtsZ complex by NMR fragment screening coupled with structure-based design. Bioorganic & medicinal chemistry. 2006;14(23):7953-61. doi: 10.1016/j.bmc.2006.07.050. PubMed PMID: 16919463.

159. Petros AM, Huth JR, Oost T, Park CM, Ding H, Wang X, Zhang H, Nimmer P, Mendoza R, Sun C, Mack J, Walter K, Dorwin S, Gramling E, Ladror U, Rosenberg SH, Elmore SW, Fesik SW, Hajduk PJ. Discovery of a potent and selective Bcl-2 inhibitor using SAR by NMR. Bioorganic & medicinal chemistry letters. 2010;20(22):6587-91. doi: 10.1016/j.bmcl.2010.09.033. PubMed PMID: 20870405.

160. Frank AO, Feldkamp MD, Kennedy JP, Waterson AG, Pelz NF, Patrone JD, Vangamudi B, Camper DV, Rossanese OW, Chazin WJ, Fesik SW. Discovery of a potent inhibitor of replication protein

a protein-protein interactions using a fragment-linking approach. Journal of medicinal chemistry. 2013;56(22):9242-50. doi: 10.1021/jm401333u. PubMed PMID: 24147804; PubMed Central PMCID: PMC3932990.

161. Villemagne B, Flipo M, Blondiaux N, Crauste C, Malaquin S, Leroux F, Piveteau C, Villeret V, Brodin P, Villoutreix BO, Sperandio O, Soror SH, Wohlkonig A, Wintjens R, Deprez B, Baulard AR, Willand N. Ligand efficiency driven design of new inhibitors of Mycobacterium tuberculosis transcriptional repressor EthR using fragment growing, merging, and linking approaches. Journal of medicinal chemistry. 2014;57(11):4876-88. doi: 10.1021/jm500422b. PubMed PMID: 24818704.

162. Murray CW, Blundell TL. Structural biology in fragment-based drug design. Current opinion in structural biology. 2010;20(4):497-507. doi: 10.1016/j.sbi.2010.04.003. PubMed PMID: 20471246.

163. Mattos C, Bellamacina CR, Peisach E, Pereira A, Vitkup D, Petsko GA, Ringe D. Multiple solvent crystal structures: probing binding sites, plasticity and hydration. J Mol Biol. 2006;357(5):1471-82. doi: 10.1016/j.jmb.2006.01.039. PubMed PMID: 16488429.

164. Sperandio O, Mouawad L, Pinto E, Villoutreix BO, Perahia D, Miteva MA. How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. European biophysics journal : EBJ. 2010;39(9):1365-72. doi: 10.1007/s00249-010-0592-0. PubMed PMID: 20237920.

165. Korb O, Olsson TS, Bowden SJ, Hall RJ, Verdonk ML, Liebeschuetz JW, Cole JC. Potential and limitations of ensemble docking. J Chem Inf Model. 2012;52(5):1262-74. doi: 10.1021/ci2005934. PubMed PMID: 22482774.

166. Sorensen J, Demir O, Swift RV, Feher VA, Amaro RE. Molecular docking to flexible targets. Methods Mol Biol. 2015;1215:445-69. doi: 10.1007/978-1-4939-1465-4_20. PubMed PMID: 25330975.

167. Hari SB, Perera BG, Ranjitkar P, Seeliger MA, Maly DJ. Conformation-selective inhibitors reveal differences in the activation and phosphate-binding loops of the tyrosine kinases Abl and Src. ACS

chemical biology. 2013;8(12):2734-43. doi: 10.1021/cb400663k. PubMed PMID: 24106839; PubMed Central PMCID: PMC3880807.

168. Chaikuad A, Tacconi EM, Zimmer J, Liang Y, Gray NS, Tarsounas M, Knapp S. A unique inhibitor binding site in ERK1/2 is associated with slow binding kinetics. Nature chemical biology. 2014;10(10):853-60. doi: 10.1038/nchembio.1629. PubMed PMID: 25195011.

169. Gray NS, Fabbro D. Discovery of allosteric BCR-ABL inhibitors from phenotypic screen to clinical candidate. Methods in enzymology. 2014;548:173-88. doi: 10.1016/B978-0-12-397918-6.00007-0. PubMed PMID: 25399646.

170. Ung PM, Dunbar JB, Jr., Gestwicki JE, Carlson HA. An allosteric modulator of HIV-1 protease shows equipotent inhibition of wild-type and drug-resistant proteases. Journal of medicinal chemistry. 2014;57(15):6468-78. doi: 10.1021/jm5008352. PubMed PMID: 25062388; PubMed Central PMCID: PMC4136727.

171. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(19):6959-64. doi: 10.1073/pnas.0708078105. PubMed PMID: 18474861; PubMed Central PMCID: PMC2383957.

172. Sevimoglu T, Arga KY. The role of protein interaction networks in systems biomedicine. Computational and structural biotechnology journal. 2014;11(18):22-7. doi: 10.1016/j.csbj.2014.08.008. PubMed PMID: 25379140; PubMed Central PMCID: PMC4212283.

173. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. Cell. 2011;144(6):986-98. doi: 10.1016/j.cell.2011.02.016. PubMed PMID: 21414488; PubMed Central PMCID: PMC3102045.

174. Gonzalez MW, Kann MG. Chapter 4: Protein interactions and disease. PLoS computational biology. 2012;8(12):e1002819. doi: 10.1371/journal.pcbi.1002819. PubMed PMID: 23300410; PubMed Central PMCID: PMC3531279.

175. Ideker T, Sharan R. Protein networks in disease. Genome research. 2008;18(4):644-52. doi: 10.1101/gr.071852.107. PubMed PMID: 18381899; PubMed Central PMCID: PMC3863981.

176. Gordo S, Giralt E. Knitting and untying the protein network: modulation of protein ensembles as a therapeutic strategy. Protein science : a publication of the Protein Society. 2009;18(3):481-93. doi: 10.1002/pro.43. PubMed PMID: 19241367; PubMed Central PMCID: PMC2760355.

177. Ivanov AA, Khuri FR, Fu H. Targeting protein-protein interactions as an anticancer strategy. Trends in pharmacological sciences. 2013;34(7):393-400. doi: 10.1016/j.tips.2013.04.007. PubMed PMID: 23725674; PubMed Central PMCID: PMC3773978.

178. Villoutreix BO, Kuenemann MA, Poyet JL, Bruzzoni-Giovanelli H, Labbe C, Lagorce D, Sperandio O, Miteva MA. Drug-Like Protein-Protein Interaction Modulators: Challenges and Opportunities for Drug Discovery and Chemical Biology. Molecular informatics. 2014;33(6-7):414-37. doi: 10.1002/minf.201400040. PubMed PMID: 25254076; PubMed Central PMCID: PMC4160817.

179. Zinzalla G, Thurston DE. Targeting protein-protein interactions for therapeutic intervention: a challenge for the future. Future medicinal chemistry. 2009;1(1):65-93. doi: 10.4155/fmc.09.12. PubMed PMID: 21426071.

180. White AW, Westwell AD, Brahemi G. Protein-protein interactions as targets for small-molecule therapeutics in cancer. Expert reviews in molecular medicine. 2008;10:e8. doi: 10.1017/S1462399408000641. PubMed PMID: 18353193.

181. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74. doi: 10.1016/j.cell.2011.02.013. PubMed PMID: 21376230.

182. Makley LN, Gestwicki JE. Expanding the number of 'druggable' targets: non-enzymes and protein-protein interactions. Chemical biology & drug design. 2013;81(1):22-32. doi: 10.1111/cbdd.12066. PubMed PMID: 23253128; PubMed Central PMCID: PMC3531880.

183. Nero TL, Morton CJ, Holien JK, Wielens J, Parker MW. Oncogenic protein interfaces: small molecules, big challenges. Nature reviews Cancer. 2014;14(4):248-62. doi: 10.1038/nrc3690. PubMed PMID: 24622521.

184. Janin J, Bahadur RP, Chakrabarti P. Protein-protein interaction and quaternary structure. Quarterly reviews of biophysics. 2008;41(2):133-80. doi: 10.1017/S0033583508004708. PubMed PMID: 18812015.

185. Smith MC, Gestwicki JE. Features of protein-protein interactions that translate into potent inhibitors: topology, surface area and affinity. Expert reviews in molecular medicine. 2012;14:e16. doi: 10.1017/erm.2012.10. PubMed PMID: 22831787; PubMed Central PMCID: PMC3591511.

186. Gowthaman R, Miller SA, Rogers S, Khowsathit J, Johnson DK, Liu C, Xu L, Anbanandam A, Aubé J, Roy A, Karanicolas J. DARC: mapping surface topography by ray-casting for effective virtual screening at protein interaction sitessubmitted.

187. OMEGA 2.5.1.4: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com. Hawkins, P.C.D.; Skillman, A.G.; Warren, G.L.; Ellingson, B.A.; Stahl, M.T.

188. http://www.gnu.org/software/gsl/.

189. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. Journal of chemical information and modeling. 2012;52(7):1757-68. doi: 10.1021/ci3001277. PubMed PMID: 22587354; PubMed Central PMCID: PMC3402020.

190. McGann M. FRED pose prediction and virtual screening accuracy. Journal of chemical information and modeling. 2011;51(3):578-96. doi: 10.1021/ci100436p. PubMed PMID: 21323318.

191. Call ST, Zubarev DY, Boldyrev AI. Global minimum structure searches via particle swarm optimization. J Comput Chem. 2007;28(7):1177-86. doi: 10.1002/jcc.20621. PubMed PMID: 17299774.

192. Kennedy J ER, editor. Particle Swarm Optimization. Proceedings of the IEEE International Conference on Neural Networks; Perth, Australia; 1995.

193. Chen HM, Liu BF, Huang HL, Hwang SF, Ho SY. SODOCK: swarm optimization for highly flexible protein-ligand docking. J Comput Chem. 2007;28(2):612-23. doi: 10.1002/jcc.20542. PubMed PMID: 17186483.

194. Liu Y, Zhao L, Li W, Zhao D, Song M, Yang Y. FIPSDock: a new molecular docking technique driven by fully informed swarm optimization algorithm. J Comput Chem. 2013;34(1):67-75. doi: 10.1002/jcc.23108. PubMed PMID: 22961860.

195. Namasivayam V, Gunther R. pso@autodock: a fast flexible molecular docking program based on Swarm intelligence. Chemical biology & drug design. 2007;70(6):475-84. doi: 10.1111/j.1747-0285.2007.00588.x. PubMed PMID: 17986206.

196. McCoy AJ, Chandana Epa V, Colman PM. Electrostatic complementarity at protein/protein interfaces. Journal of molecular biology. 1997;268(2):570-84. doi: 10.1006/jmbi.1997.0987. PubMed PMID: 9159491.

197. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. Journal of molecular biology. 1999;285(5):2177-98. PubMed PMID: 9925793.

198. Voet A, Berenger F, Zhang KY. Electrostatic similarities between protein and small molecule ligands facilitate the design of protein-protein interaction inhibitors. PloS one. 2013;8(10):e75762. doi: 10.1371/journal.pone.0075762. PubMed PMID: 24130741; PubMed Central PMCID: PMC3794991.

199. Huey R, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. J Comput Chem. 2007;28(6):1145-52. doi: 10.1002/jcc.20634. PubMed PMID: 17274016.

200. Coleman RG, Carchia M, Sterling T, Irwin JJ, Shoichet BK. Ligand pose and orientational sampling in molecular docking. PloS one. 2013;8(10):e75992. doi: 10.1371/journal.pone.0075992. PubMed PMID: 24098414; PubMed Central PMCID: PMC3787967.

201. ZAP toolkit 2.2.0: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

202. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2014.

203. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J, Hentze MW. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell. 2012;149(6):1393-406. doi: 10.1016/j.cell.2012.04.031. PubMed PMID: 22658674.

204. Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, Wyler E, Bonneau R, Selbach M, Dieterich C, Landthaler M. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular cell. 2012;46(5):674-90. doi: 10.1016/j.molcel.2012.05.021. PubMed PMID: 22681889.

205. Pascale A, Govoni S. The complex world of post-transcriptional mechanisms: is their deregulation a common link for diseases? Focus on ELAV-like RNA-binding proteins. Cellular and molecular life sciences : CMLS. 2012;69(4):501-17. doi: 10.1007/s00018-011-0810-7. PubMed PMID: 21909784.

206. Kapeli K, Yeo GW. Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. Frontiers in neuroscience. 2012;6:144. doi: 10.3389/fnins.2012.00144. PubMed PMID: 23060744; PubMed Central PMCID: PMC3462321.

207. Khalil AM, Rinn JL. RNA-protein interactions in human health and disease. Seminars in cell & developmental biology. 2011;22(4):359-65. doi: 10.1016/j.semcdb.2011.02.016. PubMed PMID: 21333748; PubMed Central PMCID: PMC3184770.

208. Ellenbecker M, Lanchy JM, Lodmell JS. Identification of Rift Valley fever virus nucleocapsid protein-RNA binding inhibitors using a high-throughput screening assay. Journal of biomolecular screening. 2012;17(8):1062-70. doi: 10.1177/1087057112448100. PubMed PMID: 22644268; PubMed Central PMCID: PMC3520603.

209. King DT, Barnes M, Thomsen D, Lee CH. Assessing specific oligonucleotides and small molecule antibiotics for the ability to inhibit the CRD-BP-CD44 RNA interaction. PloS one. 2014;9(3):e91585. doi: 10.1371/journal.pone.0091585. PubMed PMID: 24622399; PubMed Central PMCID: PMC3951440.

210. Cheng K, Wang X, Yin H. Small-molecule inhibitors of the TLR3/dsRNA complex. J Am Chem Soc. 2011;133(11):3764-7. doi: 10.1021/ja111312h. PubMed PMID: 21355588; PubMed Central PMCID: PMC3068529.

211. Gallego J, Varani G. Targeting RNA with small-molecule drugs: therapeutic promise and chemical challenges. Accounts of chemical research. 2001;34(10):836-43. PubMed PMID: 11601968.

212. Stelzer AC, Frank AT, Kratz JD, Swanson MD, Gonzalez-Hernandez MJ, Lee J, Andricioaei I, Markovitz DM, Al-Hashimi HM. Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. Nature chemical biology. 2011;7(8):553-9. doi: 10.1038/nchembio.596. PubMed PMID: 21706033; PubMed Central PMCID: PMC3319144.

213. Bitterman PB, Polunovsky VA. Attacking a nexus of the oncogenic circuitry by reversing aberrant eIF4F-mediated translation. Molecular cancer therapeutics. 2012;11(5):1051-61. doi: 10.1158/1535-7163.MCT-11-0530. PubMed PMID: 22572598; PubMed Central PMCID: PMC3349966.

214. Menendez-Arias L, Alvarez M, Pacheco B. Nucleoside/nucleotide analog inhibitors of hepatitis B virus polymerase: mechanism of action and resistance. Current opinion in virology. 2014;8C:1-9. doi: 10.1016/j.coviro.2014.04.005. PubMed PMID: 24814823.

215. James SH, Prichard MN. Current and future therapies for herpes simplex virus infections: mechanism of action and drug resistance. Current opinion in virology. 2014;8C:54-61. doi: 10.1016/j.coviro.2014.06.003. PubMed PMID: 25036916.

216. Das K, Arnold E. HIV-1 reverse transcriptase and antiviral drug resistance. Part 1. Current opinion in virology. 2013;3(2):111-8. doi: 10.1016/j.coviro.2013.03.012. PubMed PMID: 23602471; PubMed Central PMCID: PMC4097814.

217. Biswas S, Sukla S, Field HJ. Helicase-primase inhibitors for herpes simplex virus: looking to the future of non-nucleoside inhibitors for treating herpes virus infections. Future medicinal chemistry. 2014;6(1):45-55. doi: 10.4155/fmc.13.192. PubMed PMID: 24358947.

218. Das K, Arnold E. HIV-1 reverse transcriptase and antiviral drug resistance. Part 2. Current opinion in virology. 2013;3(2):119-28. doi: 10.1016/j.coviro.2013.03.014. PubMed PMID: 23602470; PubMed Central PMCID: PMC4097817.

219. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science. 1995;267(5196):383-6. PubMed PMID: 7529940.

220. Moreira IS, Fernandes PA, Ramos MJ. Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins. 2007;68(4):803-12. doi: 10.1002/prot.21396. PubMed PMID: 17546660.

221. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(31):11287-92. doi: 10.1073/pnas.0401942101. PubMed PMID: 15269345; PubMed Central PMCID: PMC509196.

222. Thanos CD, DeLano WL, Wells JA. Hot-spot mimicry of a cytokine receptor by a small molecule. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(42):15422-7. doi: 10.1073/pnas.0607058103. PubMed PMID: 17032757; PubMed Central PMCID: PMC1592646.

223. Christ F, Voet A, Marchand A, Nicolet S, Desimmie BA, Marchand D, Bardiot D, Van der Veken NJ, Van Remoortel B, Strelkov SV, De Maeyer M, Chaltin P, Debyser Z. Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. Nature chemical biology. 2010;6(6):442-8. doi: 10.1038/nchembio.370. PubMed PMID: 20473303.

224. Liu S, Wu S, Jiang S. HIV entry inhibitors targeting gp41: from polypeptides to small-molecule compounds. Current pharmaceutical design. 2007;13(2):143-62. PubMed PMID: 17269924.

225. ROCS version 3.2.0.3. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

226. Daubner GM, Clery A, Allain FH. RRM-RNA recognition: NMR or crystallography...and new findings. Current opinion in structural biology. 2013;23(1):100-8. doi: 10.1016/j.sbi.2012.11.006. PubMed PMID: 23253355.

227. Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. The FEBS journal. 2005;272(9):2118-31. doi: 10.1111/j.1742-4658.2005.04653.x. PubMed PMID: 15853797.

228. Auweter SD, Oberstrass FC, Allain FH. Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic acids research. 2006;34(17):4943-59. doi: 10.1093/nar/gkl620. PubMed PMID: 16982642; PubMed Central PMCID: PMC1635273.

229. Nolan SJS, J. C.; Tuite, J. B.; Cecere, K. L.; Baranger, A. M. Recognition of an essential adenine at a protein-RNA interface: comparison of the contribution of hydrogen bonds and a stacking interaction. J Am Chem Soc. 1999(121):2.

230. Benitex Y, Baranger AM. Recognition of essential purines by the U1A protein. BMC biochemistry. 2007;8:22. doi: 10.1186/1471-2091-8-22. PubMed PMID: 17980039; PubMed Central PMCID: PMC2203988.

231. Tuite JB, Shiels JC, Baranger AM. Substitution of an essential adenine in the U1A-RNA complex with a non-polar isostere. Nucleic acids research. 2002;30(23):5269-75. PubMed PMID: 12466552; PubMed Central PMCID: PMC137951.

232. Ohyama T, Nagata T, Tsuda K, Kobayashi N, Imai T, Okano H, Yamazaki T, Katahira M. Structure of Musashi1 in a complex with target RNA: the role of aromatic stacking interactions. Nucleic acids research. 2012;40(7):3218-31. doi: 10.1093/nar/gkr1139. PubMed PMID: 22140116; PubMed Central PMCID: PMC3326303.

233. Okano H, Kawahara H, Toriya M, Nakao K, Shibata S, Imai T. Function of RNA-binding protein Musashi-1 in stem cells. Experimental cell research. 2005;306(2):349-56. doi: 10.1016/j.yexcr.2005.02.021. PubMed PMID: 15925591.

234. Spears E, Neufeld KL. Novel double-negative feedback loop between adenomatous polyposis coli and Musashi1 in colon epithelia. The Journal of biological chemistry. 2011;286(7):4946-50. doi: 10.1074/jbc.C110.205922. PubMed PMID: 21199875; PubMed Central PMCID: PMC3037606.

235. Fan LF, Dong WG, Jiang CQ, Xia D, Liao F, Yu QF. Expression of putative stem cell genes Musashi-1 and beta1-integrin in human colorectal adenomas and adenocarcinomas. International journal of colorectal disease. 2010;25(1):17-23. doi: 10.1007/s00384-009-0791-2. PubMed PMID: 19714342.

236. Ma YH, Mentlein R, Knerlich F, Kruse ML, Mehdorn HM, Held-Feindt J. Expression of stem cell markers in human astrocytomas of different WHO grades. Journal of neuro-oncology. 2008;86(1):31-45. doi: 10.1007/s11060-007-9439-7. PubMed PMID: 17611714.

237. Seigel GM, Hackam AS, Ganguly A, Mandell LM, Gonzalez-Fernandez F. Human embryonic and neuronal stem cell markers in retinoblastoma. Molecular vision. 2007;13:823-32. PubMed PMID: 17615543; PubMed Central PMCID: PMC2768758.

238. Toda M, Iizuka Y, Yu W, Imai T, Ikeda E, Yoshida K, Kawase T, Kawakami Y, Okano H, Uyemura K. Expression of the neural RNA-binding protein Musashi1 in human gliomas. Glia. 2001;34(1):1-7. PubMed PMID: 11284014.

239. Wang XY, Penalva LO, Yuan H, Linnoila RI, Lu J, Okano H, Glazer RI. Musashi1 regulates breast tumor cell proliferation and is a prognostic indicator of poor survival. Molecular cancer. 2010;9:221. doi: 10.1186/1476-4598-9-221. PubMed PMID: 20727204; PubMed Central PMCID: PMC2939568.

240. Ye F, Zhou C, Cheng Q, Shen J, Chen H. Stem-cell-abundant proteins Nanog, Nucleostemin and Musashi1 are highly expressed in malignant cervical epithelial cells. BMC cancer. 2008;8:108. doi: 10.1186/1471-2407-8-108. PubMed PMID: 18419830; PubMed Central PMCID: PMC2387168.

241. Yokota N, Mainprize TG, Taylor MD, Kohata T, Loreto M, Ueda S, Dura W, Grajkowska W, Kuo JS, Rutka JT. Identification of differentially expressed and developmentally regulated genes in

medulloblastoma using suppression subtraction hybridization. Oncogene. 2004;23(19):3444-53. doi: 10.1038/sj.onc.1207475. PubMed PMID: 15064731.

242. Todaro M, Francipane MG, Medema JP, Stassi G. Colon cancer stem cells: promise of targeted therapy. Gastroenterology. 2010;138(6):2151-62. doi: 10.1053/j.gastro.2009.12.063. PubMed PMID: 20420952.

243. Tsuda K, Someya T, Kuwasako K, Takahashi M, He F, Unzai S, Inoue M, Harada T, Watanabe S, Terada T, Kobayashi N, Shirouzu M, Kigawa T, Tanaka A, Sugano S, Guntert P, Yokoyama S, Muto Y. Structural basis for the dual RNA-recognition modes of human Tra2-beta RRM. Nucleic acids research. 2011;39(4):1538-53. doi: 10.1093/nar/gkq854. PubMed PMID: 20926394; PubMed Central PMCID: PMC3045587.

244. Clery A, Jayne S, Benderska N, Dominguez C, Stamm S, Allain FH. Molecular basis of purine-rich RNA recognition by the human SR-like protein Tra2-beta1. Nature structural & molecular biology. 2011;18(4):443-50. doi: 10.1038/nsmb.2001. PubMed PMID: 21399644.

245. Duss O, Michel E, Yulikov M, Schubert M, Jeschke G, Allain FH. Structural basis of the non-coding RNA RsmZ acting as a protein sponge. Nature. 2014;509(7502):588-92. doi: 10.1038/nature13271. PubMed PMID: 24828038.

246. Imai T, Tokunaga A, Yoshida T, Hashimoto M, Mikoshiba K, Weinmaster G, Nakafuku M, Okano H. The neural RNA-binding protein Musashi1 translationally regulates mammalian numb gene expression by interacting with its mRNA. Molecular and cellular biology. 2001;21(12):3888-900. doi: 10.1128/MCB.21.12.3888-3900.2001. PubMed PMID: 11359897; PubMed Central PMCID: PMC87052.

247. Kharas MG, Lengner CJ, Al-Shahrour F, Bullinger L, Ball B, Zaidi S, Morgan K, Tam W, Paktinat M, Okabe R, Gozo M, Einhorn W, Lane SW, Scholl C, Frohling S, Fleming M, Ebert BL, Gilliland DG, Jaenisch R, Daley GQ. Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. Nature medicine. 2010;16(8):903-8. doi: 10.1038/nm.2187. PubMed PMID: 20616797; PubMed Central PMCID: PMC3090658.

248. Ito T, Kwon HY, Zimdahl B, Congdon KL, Blum J, Lento WE, Zhao C, Lagoo A, Gerrard G, Foroni L, Goldman J, Goh H, Kim SH, Kim DW, Chuah C, Oehler VG, Radich JP, Jordan CT, Reya T. Regulation of myeloid leukaemia by the cell-fate determinant Musashi. Nature. 2010;466(7307):765-8. doi: 10.1038/nature09171. PubMed PMID: 20639863; PubMed Central PMCID: PMC2918284.

249. de Andres-Aguayo L, Varas F, Kallin EM, Infante JF, Wurst W, Floss T, Graf T. Musashi 2 is a regulator of the HSC compartment identified by a retroviral insertion screen and knockout mice. Blood. 2011;118(3):554-64. doi: 10.1182/blood-2010-12-322081. PubMed PMID: 21613258.

250. Moore MA. A cancer fate in the hands of a samurai. Nature medicine. 2010;16(9):963-5. doi: 10.1038/nm0910-963. PubMed PMID: 20823875.

251. Nissen P, Kjeldgaard M, Nyborg J. Macromolecular mimicry. The EMBO journal. 2000;19(4):489-95. doi: 10.1093/emboj/19.4.489. PubMed PMID: 10675317; PubMed Central PMCID: PMC305586.

252. Tsonis PA, Dwivedi B. Molecular mimicry: structural camouflage of proteins and nucleic acids. Biochimica et biophysica acta. 2008;1783(2):177-87. doi: 10.1016/j.bbamcr.2007.11.001. PubMed PMID: 18068679.

253. Weber L, Illgen K, Almstetter M. Discovery of New Multi Component Reactions with Combinatorial Methods. Synlett. 1999;3:366–74.

254. Arkin Michelle R, Tang Y, Wells James A. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. Chemistry & Biology. 2014;21(9):1102-14. doi: http://dx.doi.org/10.1016/j.chembiol.2014.09.001.

255. Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N, Rizzo RC. Development and validation of a modular, extensible docking program: DOCK 5. Journal of computer-aided molecular design. 2006;20(10-11):601-19. Epub 2006/12/07. doi: 10.1007/s10822-006-9060-4. PubMed PMID: 17149653.

256. Liu Q, Gehring K. Heterodimerization of BAK and MCL-1 activated by detergent micelles. The Journal of biological chemistry. 2010;285(52):41202-10. doi: 10.1074/jbc.M110.144857. PubMed PMID: 21036904; PubMed Central PMCID: PMC3003418.

257. Lian J, Wu X, He F, Karnak D, Tang W, Meng Y, Xiang D, Ji M, Lawrence TS, Xu L. A natural BH3 mimetic induces autophagy in apoptosis-resistant prostate cancer via modulating Bcl-2-Beclin1 interaction at endoplasmic reticulum. Cell death and differentiation. 2011;18(1):60-71. doi: 10.1038/cdd.2010.74. PubMed PMID: 20577262; PubMed Central PMCID: PMC2950895.

258. Tse C, Shoemaker AR, Adickes J, Anderson MG, Chen J, Jin S, Johnson EF, Marsh KC, Mitten MJ, Nimmer P, Roberts L, Tahir SK, Xiao Y, Yang X, Zhang H, Fesik S, Rosenberg SH, Elmore SW. ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor. Cancer Res. 2008;68(9):3421-8. doi: 10.1158/0008-5472.CAN-07-5836. PubMed PMID: 18451170.

259. Shoemaker AR, Mitten MJ, Adickes J, Ackler S, Refici M, Ferguson D, Oleksijew A, O'Connor JM, Wang B, Frost DJ, Bauch J, Marsh K, Tahir SK, Yang X, Tse C, Fesik SW, Rosenberg SH, Elmore SW. Activity of the Bcl-2 family inhibitor ABT-263 in a panel of small cell lung cancer xenograft models. Clinical cancer research : an official journal of the American Association for Cancer Research. 2008;14(11):3268-77. doi: 10.1158/1078-0432.CCR-07-4622. PubMed PMID: 18519752.

260. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12. Epub 2004/07/21. doi: 10.1002/jcc.20084. PubMed PMID: 15264254.

261. Studier FW. Protein production by auto-induction in high density shaking cultures. Protein expression and purification. 2005;41(1):207-34. PubMed PMID: 15915565.

262. Nikolovska-Coleska Z, Wang R, Fang X, Pan H, Tomita Y, Li P, Roller PP, Krajewski K, Saito NG, Stuckey JA, Wang S. Development and optimization of a binding assay for the XIAP BIR3 domain using fluorescence polarization. Analytical biochemistry. 2004;332(2):261-73. doi: 10.1016/j.ab.2004.05.055. PubMed PMID: 15325294.

263. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. Journal of biomolecular NMR. 1995;6(3):277-93. PubMed PMID: 8520220.

264. Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D. PRIDB: a Protein-RNA interface database. Nucleic acids research. 2011;39(Database issue):D277-82. doi: 10.1093/nar/gkq1108. PubMed PMID: 21071426; PubMed Central PMCID: PMC3013700.

265. Do S, Hu H, Kolesnikov A, Lee W, Tsui V, Wang X, Wen Z, inventorsPyrazolo[3,4-c]pyridine compounds and methods of use. USA2012.

266. Audouze K, Nielsen EO, Olsen GM, Ahring P, Jorgensen TD, Peters D, Liljefors T, Balle T. New ligands with affinity for the alpha4beta2 subtype of nicotinic acetylcholine receptors. Synthesis, receptor binding, and 3D-QSAR modeling. Journal of medicinal chemistry. 2006;49(11):3159-71. doi: 10.1021/jm058058h. PubMed PMID: 16722635.

267. Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. Journal of computational chemistry. 2009;30(10):1545-614. doi: 10.1002/jcc.21287. PubMed PMID: 19444816; PubMed Central PMCID: PMC2810661.

268. Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: a web-based graphical user interface for CHARMM. Journal of computational chemistry. 2008;29(11):1859-65. doi: 10.1002/jcc.20945. PubMed PMID: 18351591.

269. Cheng Y, Prusoff WH. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochemical pharmacology. 1973;22(23):3099-108. PubMed PMID: 4202581.

270. Niesen FH, Berglund H, Vedadi M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nature protocols. 2007;2(9):2212-21. doi: 10.1038/nprot.2007.321. PubMed PMID: 17853878.

271. Chan TA, Wang Z, Dang LH, Vogelstein B, Kinzler KW. Targeted inactivation of CTNNB1 reveals unexpected effects of beta-catenin mutation. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(12):8265-70. doi: 10.1073/pnas.082240999. PubMed PMID: 12060769; PubMed Central PMCID: PMC123056.

272. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic acids research. 2014;42(Web Server issue):W320-4. doi: 10.1093/nar/gku316. PubMed PMID: 24753421.

273. ESPript - http://espript.ibcp.fr.