Computational tools to address challenging targets in drug discovery: target-focused chemical libraries and structure-based machine learning.

By
© 2019
Yusuf Adeshina

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

Chair: Dr. Ilya Vakser

Co-Chair: Dr. John Karanicolas

Dr. Christian Ray

Dr. Joanna Slusky

Dr. Yinglong Miao

Dr. Liang Xu

Dr. Michael Rafferty

Date Defended: 19 December 2019

The dissertation committee for Yusuf Adeshina certifies that this is the approved version of the following dissertation:

Computational tools to address challenging targets in drug discovery: target-focused chemical libraries and structure-based machine learning.

_____

Chair: Dr. Ilya Vakser

_____

Co-Chair: Dr. John Karanicolas

Date Approved: 19 December 2019

ABSTRACT

There are many human drug targets without any known small molecule inhibitors, and a lot of these challenging targets play a crucial role in important disease-relevant processes. RNA-binding proteins (RBPs) are some of the examples of these kind of targets. While RBPs play a crucial role in countless cellular processes, including post-transcriptional regulation of genes, efforts directed at finding small molecule inhibitors for these targets have been largely unsuccessful. For this reason, I have focused my PhD studies in developing computational methods that will allow us to rapidly and robustly identify small-molecule inhibitors of RBPs. From an *in-silico* standpoint, the scoring functions that power most computational structure-based drug discovery are limited by high false positive rates. To address this challenge, I built the first false-positive-aware machine learning scoring function (vScreenML). vScreenML demonstrated a significant improvement in the false positive rate over the current state-of-the-art classical and machine learning scoring functions—both in retrospective and prospective evaluations. More broadly, existing virtual screening approaches were also built to suit traditional drug targets. By contrast, RBPs have unique structural features that differentiate them from most of these types of target classes: they have a large, shallow interface that more polar than most traditional drug targets. Since RBPs are structurally distinct, this may explain why traditional methods have struggled to find chemical matter to address these. To tackle this challenge, we built the first fully automated RBP pharmacophore extractor that identifies "hotspots" on the RNA that contribute extensively to the binding affinity of the protein-RNA interaction; these hotspots are then used as template for pharmacophoric virtual screening. This tool also powers the first PDB-wide pharmacophore analysis and selectivity profiling of RBPs and was instrumental to our success in designing the first series of rationally designed inhibitors of Musashi proteins. Finally, the compounds that comprise typical screening libraries are also biased towards the types of chemical space that are appropriate for traditional drug targets. To address this, I developed a method for building target-focused libraries of synthetically accessible compounds and applied this to build a collection of compounds enriched in likely Musashi inhibitors. To test the utility of

this library, I synthesized and tested some of the top-scoring hits and confirmed that we had identified new Musashi inhibitors from this library. Looking ahead, I envision that these three tools will collectively enable development of better inhibitors targeting Musashi, and entirely new inhibitors of other RBPs.

ACKNOWLEDGEMENTS

I want to sincerely thank my advisor, Dr. John Karanicolas, for all the support, advice and patience all through these years. Sincerely, this work wouldn't have been possible without his guidance and encouragement.

I also want to specially thank all the members of my dissertation committee (Dr. Ilya Vakser, Dr. Christian Ray, Dr. Joanna Slusky, Dr. Yinglong Miao, Dr. Liang Xu and Dr. Mike Rafferty) for their invaluable contribution to my academic and professional development. I want to also thank them for taking out time from their busy schedule to attend my annual meetings.

My gratitude will not be complete without thanking Dr. Eric Deeds, a former member of my dissertation committee for his immense contribution to my dissertation research.

Also, I want to acknowledge the enormous contribution of Karanicolas lab members—both former and current. Especially David Johnson, Andrea Bazzoli and Ragul Gowthaman—all of whom are alumni of the lab now—for showing me how to do good-quality research in my early days in the lab. The current members (Nan Bai, Sven Miller, Shipra Malhotra, Kirubakaran Palani, Chris Parry, Lei Kei, Daniel Yeggoni, Jittasak Khowsathit and Grigorii Andrianov) of the lab are also well appreciated for their contributions in my PhD journey.

I also want to thank my parents for all the encouragement, love and prayers. Thank you for encouraging me to keep going even at times when I don't feel like it.

Finally, I want to thank my family for the unending love and support. Especially, my wife Adijat Mustapha and our son Michael Yusuf for all the support, understanding that PhD is a long and hard journey.

TABLE OF CONTENTS

TITLE I

ACCEPTANCE II

ABSTRACT III

ACKNOWLEDGEMENTS V

TABLE OF CONTENTS VI

INTRODUCTION 1

CHAPTER 1: MACHINE LEARNING CLASSIFICATION CAN REDUCE FALSE POSITIVES IN STRUCTURE-BASED VIRTUAL SCREENING 4

1.1 ABSTRACT 5

1.2 INTRODUCTION 6

1.3 RESULTS 8

1.3.1 DEVELOPING A CHALLENGING TRAINING SET 8

1.3.2 A NEW CLASSIFIER FOR IDENTIFYING ACTIVE COMPLEXES: VSCREENML 12

1.3.3 BENCHMARKING VSCREENML USING INDEPENDENT TEST SETS 16

1.3.4 EVALUATING VSCREENML IN A PROSPECTIVE EXPERIMENT 21

1.4 DISCUSSION 24

1.5 METHODS 27

1.5.1 ACCESSING THESE TOOLS 27

1.5.2 BUILDING THE D-COID SET 27

1.5.3 EXTRACTING STRUCTURAL FEATURES 28

1.5.4 MACHINE LEARNING 29

1.5.5 VIRTUAL SCREENING BENCHMARKS 29

1.5.6 VIRTUAL SCREEN AGAINST ACETYLCHOLINESTERASE 30

1.5.7 ACETYLCHOLINESTERASE INHIBITION ASSAY 31

1.5.8 NOVELTY OF AC6 AS AN ACHE INHIBITOR 31

1.6 ACKNOWLEDGEMENTS 32

CHAPTER 2: RATIONALLY DESIGNING INHIBITORS OF THE MUSASHI PROTEIN-RNA INTERACTION BY HOTSPOT MIMICRY 33

2.1 ABSTRACT 34

2.2 INTRODUCTION 35

2.3 COMPUTATIONAL APPROACH 36

2.3.1 BUILDING "HOTSPOT PHARMACOPHORES" 37

2.3.2 IDENTIFYING COMPLEMENTARY LIGANDS 37

2.3.3 MUSASHI-1, AN RRM-CONTAINING PROTEIN 38

2.4 RESULTS 39

2.4.1 COMPUTATIONAL SCREENING AGAINST MSI1 RRM1 39

2.4.2 SAR STUDY OF R12 DERIVATIVES 42

2.4.3 INHIBITION OF MUSASHI-2 45

2.4.4 BIOPHYSICAL CHARACTERIZATION OF R12-8-44-3 45

2.4.5 EXPLORING SELECTIVITY OF R12-8-44-3 47

2.5 DISCUSSION 52

2.6 METHODS 54

2.7 ACKNOWLEDGEMENTS 54

CHAPTER 3: ..... DESIGNING INHIBITORS OF RNA-BINDING PROTEIN MUSASHI-1 VIA *IN SILICO* COMBINATORIAL CHEMISTRY 55

3.1 ABSTRACT 56

3.2 INTRODUCTION 57

3.3 COMPUTATIONAL APPROACH 60

3.3.1 OVERARCHING STRATEGY 60

3.3.2 BUILDING A MSI1-FOCUSED LIBRARY 63

3.3.3 SCREENING MSILIB FOR MUSASHI-1 INHIBITORS 65

3.4 RESULTS 66

3.4.1 COMPARISON OF MSILIB, ZINC AND ENAMINE 66

3.4.2 TESTING HITS FROM MSILIB AS INHIBITORS OF MUSASHI-1 69

3.5 DISCUSSION 71

3.6 METHODS 74

3.6.1 PROTEIN EXPRESSION AND PURIFICATION 74

3.6.2 FLUORESCENCE POLARIZATION COMPETITION ASSAYS 74

3.6.3 SYPRO DSF ASSAY 74

3.6.4 SYNTHESIS OF YA-SERIES COMPOUNDS: OVERVIEW 75

3.7 ACKNOWLEDGEMENTS 79

CONCLUSIONS 80

FUTURE DIRECTIONS **82**

APPENDIX A: SUPPORTING INFORMATION FOR CHAPTER 1 ..... **83**

SUPPORTING FIGURES **83**

APPENDIX B: SUPPORTING INFORMATION FOR CHAPTER 2 ..... **88**

SUPPORTING METHODS **88**

PDB STRUCTURES USED IN CALCULATIONS 88

BUILDING HOTSPOT PHARMACOPHORES 88

IDENTIFYING COMPLEMENTARY LIGANDS 89

PREDICTING TARGET SELECTIVITY 89

COMPOUNDS OPTIMIZATION 90

MODEL BUILDING OF R12 DERIVATIVES 91

PROTEIN EXPRESSION AND PURIFICATION 91

FLUORESCENCE POLARIZATION COMPETITION ASSAYS 92

DIFFERENTIAL SCANNING FLUORIMETRY (THERMOFLUOR) 93

NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY 93

SUPPORTING TABLES **94**

SUPPORTING FIGURES **96**

APPENDIX C: SUPPORTING INFORMATION FOR CHAPTER 3 ..... **102**

SUPPORTING FIGURES **102**

REFERENCES **103**

INTRODUCTION

Completion of the human genome project heralded the dawn of a new era in drug discovery, as it was believed that the availability of more drug targets would translate into an explosion in the number of FDA-approved drugs. While there has indeed been a dramatic increase in the number of possible drug targets, this was accompanied by only a small uptick in the number of FDA-approved drugs. This is because the vast majority of these newly-identified targets are considered challenging ("difficult-to-drug" targets), largely because conventional drug discovery approaches have historically failed to deliver useful compounds addressing these targets.

There are two main approaches for finding inhibitors of a new target: high-throughput screening (HTS) (experimental) and virtual screening (VS) (computational). While most drug discovery campaigns still rely primarily on HTS, it is expensive and has high initial false positive rate. Moreover, with developments in the field of synthetic chemistry, and particularly *in silico* collections of synthetically-accessible compounds, the growth in size of chemical libraries has begun to outpace HTS capabilities. As a result, research programs are increasingly turning to virtual screening. At its essence, virtual screening involves ranking the compounds in a virtual library for their ability to modulate some target protein. In order to rank the compounds in the library, some features are collected (featurization) and these features are scored (scoring) using a scoring function (SF). When the features derive solely from the chemical structures / properties of the compounds themselves, this is ligand-based virtual screening (LBVS); by contrast, other approaches entail building models of the compounds bound to the target protein to extract the relevant features, termed structure-based virtual screening (SBVS). Both classes of virtual screening have been gaining momentum in recent years, as they can provide new hits from very large libraries in a cost-effective manner. There remain three outstanding hurdles with respect to structure-based virtual screening, however, which I sought to tackle in the course of my PhD research.

The first is the accuracy of scoring function. Classical SFs assume a predetermined theory-inspired functional form; as an alternative, machine learning SFs have been proposed by several other research groups [1-7]. The latter tend to perform better than their classical counterparts in retrospective benchmark experiments, because they learn the underlying functional form from data rather than building it in ahead of time. That said, prospective experimental validation of these SFs are still very rare; in the few cases where these have been reported, prospective predictions are still fraught with high false positive rates [8]. In Chapter 1, I describe the construction of a new machine learning training/benchmark set (D-COID), and use this to develop the first false-positive-aware machine learning scoring function (vScreenML) for structure-based virtual screening. In retrospective benchmarks I find that vScreenML out-performs other machine learning SFs, and in a prospective benchmark I find that it provides an unprecedented high success rate for delivering active compounds.

The second challenge is that these tools can struggle when the target protein surface is relatively flat and polar. While this is not the case for most traditional drug targets (e.g., GPCRs and kinases), accessing many interesting biological targets – such as RBPs – would require a means to address such surfaces. In chapter 2, I describe the novel computational approach that we developed to design inhibitors of RBPs. Briefly, the method defines a hotspot pharmacophore from the RNA's interactions with the target protein and uses this to screen for putative inhibitors. By leveraging the unique structural features of the RNA-binding site in the same manner as utilized by RNA itself, we have used this method to develop the first rationally designed inhibitors of Musashi-1 and Musashi-2.

Finally, the chemical matter comprising the screening collection itself can be a limitation. It is by now accepted that screening for inhibitors of non-traditional targets is more challenging than screening against GPCRs and enzymes, in part because libraries have been designed and optimized against the backdrop of these traditional targets [9-12]. In effect, the fact that many early drugs targeted these protein classes led to biases in compound collections, which further reinforced the bias by enhancing success rates specifically for these types of targets. A proposed solution to this problem, at least for addressing non-traditional classes, has been the construction of target-focused libraries. In the past, these have

typically been constructed at the level of libraries intended for whole target classes, for example libraries purported to be enriched in inhibitors of protein-protein interactions. In Chapter 3, I describe the first Musashi-focused small molecule library, build by enumerating products of chemical reactions from building blocks that match fragment of Musashi's hotspot pharmacophore. I screened this library, synthesized the top-scoring compounds, and confirmed that this led to a scaffold for inhibition of Musashi.

Moving forward, I anticipate that these tools will enable discovery of inhibitors for additional biomedically-relevant RBPs. Moreover, I also highly optimistic that the same tools will also serve as a valuable resource for targeting other protein classes beyond RBPs, because of the generality of the underlying approaches.

CHAPTER 1: Machine learning classification can reduce false positives in structure-based virtual screening

Yusuf Adeshina[1,2], Eric Deeds[2,3], and John Karanicolas[1*]

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

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

University of Kansas, Lawrence, KS 66045

*To whom correspondence should be addressed.

E-mail: john.karanicolas@fccc.edu

1.1 Abstract

With the recent explosion in the size of libraries available for screening, virtual screening is positioned to assume a more prominent role in early drug discovery's search for active chemical matter. Modern virtual screening methods are still, however, plagued with high false positive rates: typically, only about 12% of the top-scoring compounds actually show activity when tested in biochemical assays. We argue that most scoring functions used for this task have been developed with insufficient thoughtfulness into the datasets on which they are trained and tested, leading to overly simplistic models and/or overtraining. These problems are compounded in the literature because none of the studies reporting new scoring methods have validated their model prospectively within the same study. Here, we report a new strategy for building a training dataset (D-COID) that aims to generate highly-compelling decoy complexes that are individually matched to available active complexes. Using this dataset, we train a general-purpose classifier for virtual screening (vScreenML) that is built on the XGBoost framework of gradient-boosted decision trees. In retrospective benchmarks, our new classifier shows outstanding performance relative to other scoring functions. We additionally evaluate the classifier in a prospective context, by screening for new acetylcholinesterase inhibitors. Remarkably, we find that nearly all compounds selected by vScreenML show detectable activity at 50 μM, with 10 of 23 providing greater than 50% inhibition at this concentration. Without any medicinal chemistry optimization, the most potent hit from this initial screen has an IC_{50} of 280 nM, corresponding to a Ki value of 173 nM. These results support the use of the D-COID strategy for training classifiers in other computational biology tasks, and for using vScreenML in virtual screening campaigns against other targets.

1.2 Introduction

Advances in biomedical sciences, driven especially by the advent of next-generation genome sequencing technologies, have enabled discovery of many new potential drug targets [13,14]. Ultimately, however, validating a new candidate target for therapeutic intervention requires development of a chemical probe to explore the consequences of pharmacological manipulation of this target [15]. In recent years this step has typically been carried out by using high-throughput screening (HTS) [16] as a starting point for subsequent medicinal chemistry optimization; with improvements in automation, it has become feasible to screen libraries that exceed a million compounds [17].

More recently, however, sets of robust chemical transformations from available building blocks have been used to enumerate huge virtual libraries of compounds that are readily accessible but never before synthesized [18-21]. These libraries can comprise billions of compounds, and thus remain far beyond the scale accessible to even the most ambitious HTS campaign. This expansion of chemical space in which to search, along with the high cost of setting up and implementing an HTS screen, has increasingly driven the use of complementary computational approaches.

In broad terms, virtual screening approaches can be categorized into two classes: ligand-based screens and structure-based screens [22-24]. Ligand-based screening starts from the (2D or 3D) structure of one or more already-known ligands, and then searches a chemical library for examples that are similar (in either a 2D or a 3D sense). In contrast, structure-based screening does not require *a priori* knowledge of any ligands that bind to the target protein: instead, it involves sequentially docking each member of the chemical library against the three-dimensional structure of the target protein (receptor) and using a scoring function to evaluate the "quality" of each modeled protein-ligand complex. The scoring function is intuitively meant to serve as a proxy for the expected strength of a given protein-ligand complex (i.e. its binding affinity) [25], and is typically built upon either a physics-based force-field [26,27], an empirical function [28-31], or a set of knowledge-based terms [32,33].

After docking, the scoring function is used to select the most promising compounds for experimental characterization; at this stage the accuracy of the scoring function is of paramount importance, and represents the primary determinant of success or failure in structure-based screening [3]. A snapshot of the field was captured by a review summarizing successful outcomes from 54 virtual screening campaigns against diverse protein targets [24]; for the most part, all groups screened the same 3-4 million compounds from ZINC [34,20]. Excluding GPCR's and artificial cavities designed into protein cores, the median values across the set reveal that an expert in the field – using their own preferred methods of choice, which can include various post-docking filters and human visual inspection ("expert hit-picking") – can expect about 12% of their predicted compounds to show activity. That said, the hit rate can also be higher in cases where the composition of the screening library is restricted to compounds containing a functional group with natural affinity for the target site (certain well-explored enzyme active sites). Conversely, the hit rate is typically lower when the scoring function is applied without additional filters or human intervention [24]. The median value of the most potent hit from each of the collected campaigns had K_d or K_i value of ~3 μM, although this latter result is strongly impacted by the fact that some of these K_d or K_i values are from custom compounds subsequently optimized via medicinal chemistry, rather than from the initial screening hit.

Despite extensive efforts, the reasons for which active compounds are only identified at a relatively low rate are not quite clear. In addition to factors not evident from the structure of the modeled complex (compound solubility, incorrectly modeled protonation/tautomerization states of the ligand, etc.), we and others have hypothesized that the current bounds of performance may be attributable to limitations in traditional scoring functions [1,35]: these may include inadequate parametrization of individual energy terms, exclusion of potentially important terms, and also failure to consider potential non-linear interactions between terms. For these reasons, machine learning techniques may be especially well-suited for developing scoring functions that will provide a dramatic improvement in the ability to identify active compounds without human expert intervention. However, while machine learning may offer the potential to improve on the high false positive rate of current scoring function, further analysis

has revealed that some of the methods to date reporting promising results in artificial benchmark experiments may have inadvertently overfit models to the training data [36] or achieve apparently impressive performance by detecting systematic differences in the chemical properties of active versus decoy compounds [37]; as a result, these models may not yield transferrable performance when tested in prospective evaluations [38].

Here, we report the development of a dataset aimed to promote training of a machine learning model designed to be maximally useful in real-world (prospective) virtual screening applications. To build this dataset, we compile a set of "compelling" decoy complexes: a set that mimics representative compounds that might otherwise move forward to experimental testing if generated in the course of a typical virtual screening pipeline. We then use this dataset to train a machine learning classifier to distinguish active complexes from these compelling decoys, with the rationale that this is precisely the step at which standard scoring functions must be augmented. Finally, we apply this model in a *prospective* experiment, by screening against a typical enzyme target (acetylcholinesterase) and testing the top-scoring compounds in a biochemical (wet lab) assay for inhibition of protein activity.

1.3 Results

1.3.1 Developing a challenging training set

Machine learning methods at varying levels of sophistication have already been considered in the context of structure-based virtual screening [1,35,39-41,3,4,42,5,43,6,44,45,7]. The vast majority of such studies sought to train a regression model that would recapitulate the binding affinities of known complexes, and thus provide a natural and intuitive replacement for traditional scoring functions [1,35,39-41,3-6,44,45,7]. The downside of such a strategy, however, is that the resulting models are not ever exposed to any inactive complexes in the course of training: this is especially important in the context of docked complexes arising from virtual screening, where most compounds in the library are presumably inactive. We instead anticipated that a binary classifier would prove more appropriate for distinguishing

active versus inactive compounds, and that training would prove most effective if decoy complexes closely reflected types of complexes that would be encountered during real applications.

Building first our set of active complexes, we drew examples from available crystal structures in the Protein Data Bank (PDB). Others have used collections of active compounds for which the structure of the complex is not known, and docked these to obtain a considerably larger set of active complexes [42,5]. The downside of this approach, however, is that mis-docked examples (which may be numerous) are labeled as active during training; this is problematic because mis-docked models do not have appropriate interactions with the protein target that would lead to engagement, and thus should be marked as inactive by the classifier. While restricting examples of active complexes to those available in the PDB drastically limits the number available for training, this strategy ensures that the resulting model will evaluate complexes on the basis of the protein-ligand interactions provided.

Our primary consideration in compiling active compounds for the training set was that the scope of examples should match as closely as possible those anticipated to be encountered when the model is deployed. Training the model on an overly restrictive set of examples would limit its utility (since many cases will be "out of distribution"), whereas training too broadly might limit the resulting model's performance. Accordingly, we sought to train the model on precisely the type of scenarios that match its intended application. We therefore further filtered the set of active compounds from the PDB to include only ligands that adhere to the same physicochemical properties required for inclusion in our compound library for real screening applications (see *Methods*). This led to a collection of 1383 active complexes, which were then subjected to energy minimization: this prevented us from inadvertently training a model that simply distinguished between crystal structures and models produced by virtual screening.

Turning next to the set of decoy complexes, our primary consideration in compiling the training set was that the decoy complexes should be as "compelling" as possible. If the decoy complexes can be distinguished from the active complexes in some trivial way – if they frequently have steric clashes, for example, or they are systematically under-packed, or they do not contain intermolecular hydrogen bonds – then the classifier can simply use these obvious differences to readily distinguish active versus inactive

compounds. In addition to making compelling decoys, the proportion of decoys-to-actives also has a significant effect on the performance of machine learning trained model [46]. In order to achieve a nearly balanced training set, we aimed to include only small number of (very challenging) decoy complexes.

For each active complex, we first used the DUD-E server [47] to identify fifty compounds with physicochemical properties matched to the active compound but completely unrelated chemical structure: this provided a set of compounds compatible in very broad terms for the corresponding protein's active site, and also ensured that the decoy compounds would not have systematic differences from the active compounds. We then built low-energy conformations of each candidate decoy compound, and screened these against the three-dimensional structure of the active compound using ROCS [48]. From among the fifty candidates, we selected those that best matched the overall shape and charge distribution of the active ligand. Using the structural alignment of the decoy compound onto the active compound, we placed the decoy into the protein's active site, and carried out the same energy minimization that was applied to the active complexes (**Figure 1a**).

We note that the protocol used here to build the decoy complexes doubles as an entirely reasonable approach for ligand-based (pharmacophoric) virtual screening: indeed, ROCS is typically applied to identify compounds with matched three-dimensional properties to a given template, with the expectation that the hits will themselves be active [49-51]. Thus, the unique strategy motivating construction of our training set is in essence a form of adversarial machine learning: we intentionally seek to build decoys that we anticipate would be mis-classified by most models. We named this dataset **D-COID** (**D**ataset of **CO**ngruent **I**nhibitors and **D**ecoys), and have made it publicly-available for others to use freely (see *Methods*).

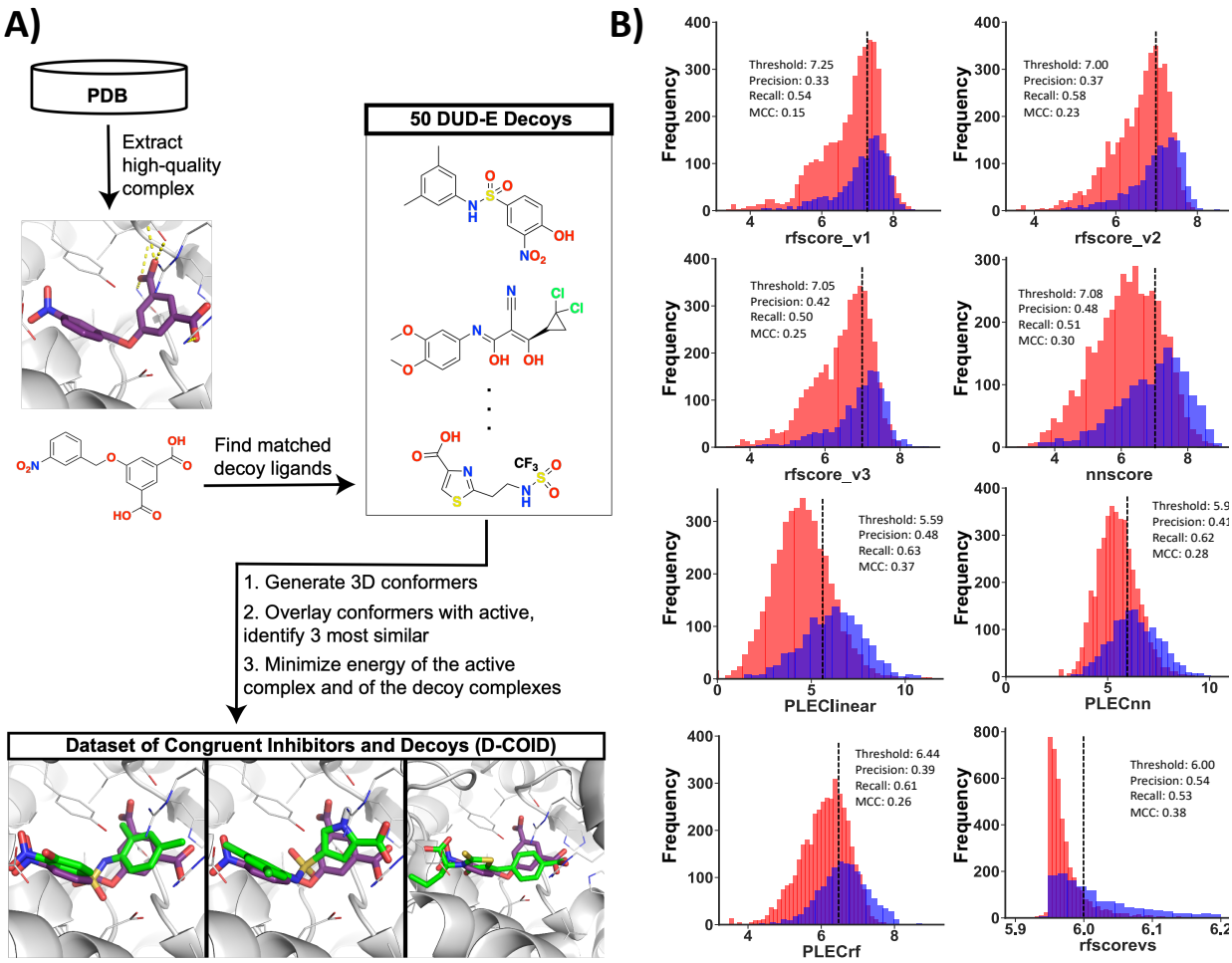


Figure 1: Developing a challenging training set (D-COID). (A) Active complexes were assembled from the PDB by filtering for ligands that match those reflected in a screening library. For each active complex, 50 physicochemically-matched compounds were selected and overlaid onto the active compounds; the three most similar compounds on the basis of overall shape and electrostatic similarity were aligned into the protein active site, and used as decoy complexes. This strategy mimics the selection of candidate (active) compounds in a realistic pharmacophore-based screening pipeline, and thus generates highly compelling decoy complexes for training/testing. **(B)** Modern scoring functions cannot distinguish active complexes from decoys in this set. Overlaid histograms are presented for scores obtained using various scoring functions when applied to active complexes (*blue*) and decoy complexes (*red*) in D-COID. For all eight methods tested, the distribution of scores assigned to active complexes strongly overlaps with the distribution of scores assigned to decoy complexes. From each model's continuous scores, 10-fold cross validation was used to obtain the classification cutoff that maximizes Matthews correlation coefficient (MCC) on each subset of the data. These cutoffs were used in calculating the precision/recall/MCC for each method. The mean of these 10 threshold values is reported with each plot.

To confirm that this decoy-generation strategy indeed led to a challenging classification problem, we applied some of the top reported scoring functions in the literature to distinguish between active and decoy complexes in the D-COID set. For all eight methods tested (nnscore [35], RF-Score v1 [1], RF-Score v2 [3], RF-Score v3 [3], PLEClinear [45], PLECnn [45], PLECrf [45], and RF-Score-VS [5]), we found that the distribution of scores assigned to active complexes was strongly overlapping with those of the decoy complexes (**Figure 1b**), indicating that these models showed very little discriminatory power when applied to this set.

Typical scoring functions report a continuous value, because they intend to capture the strength of the protein-ligand interaction. In order to use the scoring function for classification, one must define a threshold value at which complexes are predicted to be either active or inactive. To avoid over-estimating performance by selecting the threshold with knowledge of the test set, we carried out 10-fold cross validation to determine the appropriate threshold. In particular, we used 90% of the dataset to define the threshold that maximized the Matthews correlation coefficient (MCC), then applied this threshold to assign each complex in the unseen 10% as active/inactive. Using this unbiased thresholding measure to assign each complex in the D-COID set, we found the Matthews correlation coefficient (MCC) for best-performing scoring function in this experiment to be only 0.39.

1.3.2 A new classifier for identifying active complexes: vScreenML

Having developed a relevant and challenging training set, we next sought to develop a machine learning model that could discriminate between active and decoy complexes in this set. It has been pointed out in the past that machine learning models built exclusively upon protein-ligand element-element distance counts can yield apparently impressive performance in certain benchmarks without proving useful beyond these [52]. To avoid this pitfall, we used as our starting point the Rosetta energy function [53]: a classical linear combination of traditional (physics-based) molecular mechanics energy terms, alongside empirical terms added so that distributions of atomic arrangements would quantitatively mimic those observed in the PDB [54]. While we acknowledge that the Rosetta energy function is not

commonly used for virtual screening, this is primarily because it is too slow to be applied for docking large compound libraries: in one recent benchmark for classification of active versus decoy complexes [55], the Rosetta energy function showed equivalent performance as the popular FRED Chemgauss4 scoring function [56].

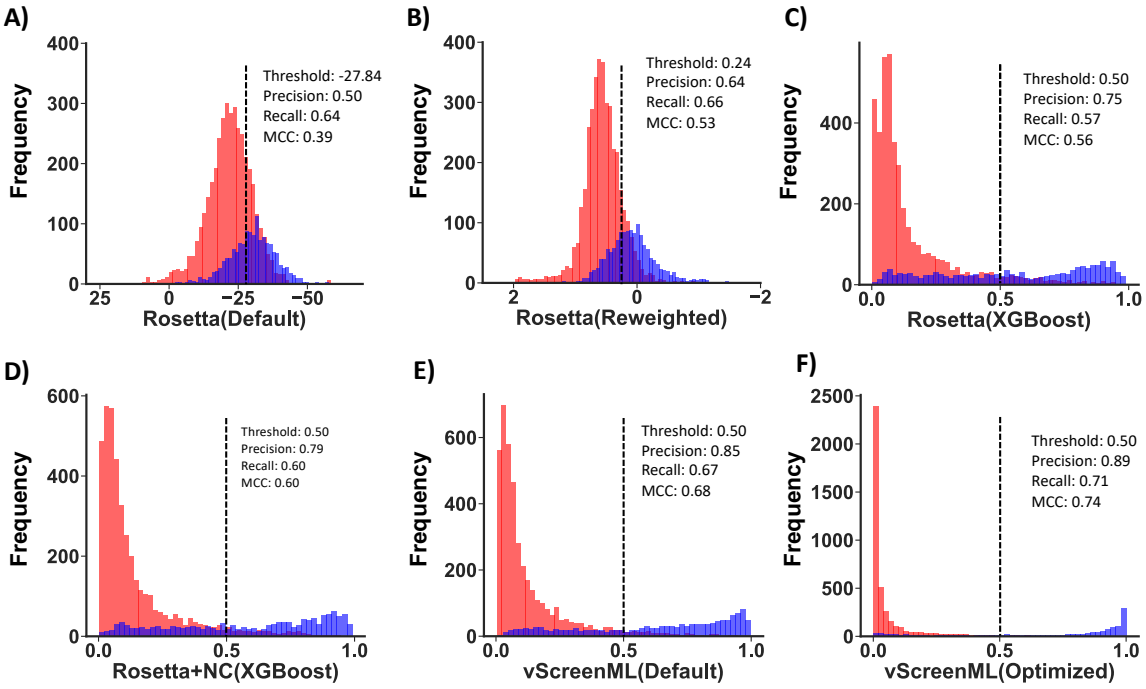


Figure 2: Development of vScreenML. Overlaid histograms are presented for scores obtained when scoring active complexes (*blue*) and decoy complexes (*red*) from D-COID. Scoring functions used were: **(A)** Default Rosetta energy function, **(B)** Linearly-reweighted Rosetta energy terms, **(C)** Rosetta energy terms combined via XGBoost, **(D)** Rosetta energy terms plus structural assessments, **(E)** Rosetta terms plus additional diverse descriptors (non-optimized vScreenML), **(F)** vScreenML after hyper-parameter tuning. Over the course of this sequence, the overlap between the active and decoy complexes is progressively reduced and MCC systematically increases. For the first two panels, 10-fold cross validation was used to obtain the classification cutoff that maximizes Matthews correlation coefficient (MCC) on each subset of the data. These cutoffs were used in calculating precision/recall/MCC, and the mean of these 10 threshold values is reported. Because the remaining panels each report results from classification models, their thresholds are fixed at 0.5.

At the outset, we found that applying Rosetta to the D-COID set did not yield results notably different than in our previous experiment (**Figure 2a**), and indeed this was confirmed quantitatively through the Matthews correlation coefficient (0.39). Next, we used 10-fold cross validation to re-weight the terms in this scoring function for improved performance in this D-COID classification task using a

perceptron [57,58] to maintain the linear functional form of the Rosetta energy function: this resulted in a modest improvement in the apparent separation of scores (**Figure 2b**), but a notable improvement in MCC (0.53). This observation is unsurprising, because the Rosetta energy function is primarily optimized for proteins rather than protein-ligand complexes, and re-training its component energies for a specific task will naturally lead to improved performance for that task. For precisely this reason, historically a separate linearly re-weighted version of the default Rosetta energy function has been used when modeling protein-ligand complexes [59] or when re-ranking complexes from virtual screening [55].

Next, we explored the performance of models that move beyond linear combinations of these energy terms, and instead use these component energies as the basis for building decision trees. Using the XGBoost framework (an implementation of gradient-boosted decision trees), we observed notable separation of the scores assigned to active/decoy complexes (**Figure 2c**), along with a slight increase in MCC (0.56).

To complement the existing terms in the Rosetta energy function, we next added a series of structural quality assessments calculated by Rosetta that are not included in the energy function (**Figure S1**); inclusion of these terms yielded a model with further improved discriminatory power (**Figure 2d**). Inspired by this improvement, we then incorporated additional structural features aiming to capture more sophisticated chemistry than that encoded in Rosetta's simple energy function, specifically from RF-Score [1] (features that count the occurrence of specific pairwise intermolecular contacts), from BINANA [60] (analysis of intermolecular contacts), from ChemAxon [61] (ligand-specific molecular descriptors), and from Szybki [62] (a term intended to capture ligand conformational entropy lost upon binding). We proceeded to train a model using this collection of features, which we denote "vScreenML", and were pleased to discover that these again increased the separation between scores assigned to active and decoy complexes (**Figure 2e**). Finally, we used hyperparameter tuning to optimize development of the model (**Figure S2**), and accordingly developed a model that provided nearly complete separation of active and decoy complexes (**Figure 2f**) and unprecedented MCC for this challenging task (0.74). We have made this model publicly-available for others to use freely (see *Methods*).

Through the course of developing of this model, we transitioned from a linear combination of six Rosetta features with clear physical basis, to a collection of 68 diverse and likely non-orthogonal features connected through a more complex underlying model (**Figure S1**). Using the complete set of features that comprise vScreenML, we tested alternate machine learning frameworks, leading us to discover that a different implementation of gradient-boosted decision trees yielded essentially identical performance, and other models built upon decision trees were only slightly worse. By contrast, other models that are not built on decision trees did not provide comparable performance (**Figure S3a**). Importantly, we note that this model has been trained to distinguish actives from decoy complexes in a context where both have been subjected to energy minimization using the Rosetta energy function: the same optimized model is not necessarily expected to recognize actives successfully if they have not been prepared this way (e.g. crystal structures).

To evaluate the contributions of each part of our feature set, we next removed one at a time all features from a given origin, and explored how the lack of these features would affect performance (**Figure S3b**). This experiment showed that only a very small deterioration in performance was observed when either the RF-Score or BINANA features were removed, but removing both had a large impact; this is unsurprising, given the fact that many of the features in these sets are correlated. Further, removal of SZYBKI's conformational entropy term had no impact on the model's performance, suggesting either that the change in ligand conformational entropy as described by SZYBKI does not help distinguish active versus decoy complexes in this dataset, or that this effect is already captured through some combination of other features. In principle, features that are unnecessary (either because they are correlated with other features or because they do not help in classification) should be removed to better avoid the risk of overtraining. In this case, however, we because XGBoost is not particularly susceptible to overtraining and our feature set remains relatively small in comparison to our training set, we elected to instead test our model immediately in orthogonal benchmarks to evaluate potential overtraining.

1.3.3 Benchmarking vScreenML using independent test sets

The DEKOIS project (currently at version 2.0) [63,64] is intended to provide a "demanding" evaluation set for testing virtual screening methods. Acknowledging that a wide variety of factors make some protein targets easier to model than others, this set includes 81 different proteins with available crystal structures. For each protein, a custom library is provided that contains 40 active compounds and 1200 decoys: thus, about 3.2% of each library is active. The crystal structures of active complexes are not provided (and indeed, most have not yet been experimentally determined). To evaluate performance of a new scoring function, one typically ranks all 1240 compounds for a given protein and selects the top-scoring 12; the enrichment factor for this subset of the library (EF-1%) corresponds to the ratio of the percent of active compounds among the selected 12 to the ratio of active compounds in the original library. Scoring perfectly for a given protein in this set would mean ranking 12 active compounds before all 1200 of the decoys: this would correspond to EF-1% = 1.00/0.032 = 31. Conversely, a method that randomly selects compounds from the library would (on average) select active compounds 3.2% of the time, and thus yield an EF-1% of 1.

Among the 81 proteins in the DEKOIS set, we noted that some were included in our training set as well. To avoid any potential information leakage that might overestimate the performance we could expect in future applications, we removed these testcases. This left a set of 23 protein targets, each of which vScreenML had never seen before. For each protein, we docked each compound in the corresponding library to the active site (see *Methods*); we note that this unavoidable step could artificially deflate the apparent performance of vScreenML or other models tested, since a mis-docked active compound should have no basis for being identified as active. Some of the compounds in the DEKOIS set could not be suitably modeled in all parts of our pipeline, and were therefore removed; each of the 23 proteins considered ultimately was used to generate 30-40 active complexes and 800-1200 decoy complexes. Each of these complexes (both actives and decoys) were then subjected to energy minimization using the Rosetta: as noted earlier, vScreenML should only be applied in the context of Rosetta-minimized structures. Along with vScreenML, we used eight other machine learning scoring

functions were then used to rank the docked-and-minimized models: nnscore [35], PLECnn [45], PLECrf [45], PLEClinear [45], RF-Score v1 [1], RF-Score v2 [3], RF-Score v3 [3] and RF-Score-VS [5]. We additionally included the (default) Rosetta energy function in this benchmark [53].

To compare performance between methods, we plot EF-1% using one method (for each of the 23 protein targets) as a function of EF-1% using the other method (**Figure 3a**). As plotted here, points below the diagonal are specific protein targets for which vScreenML outperformed the alternate method (higher EF-1% for this protein target). The importance of training on both actives and decoys for this task is immediately apparent in these comparisons, by comparing for example vScreenML against PLECnn (a neural network representing the current state-of-the-art among models trained exclusively on active complexes). For the 23 targets in this experiment, PLECnn out-performs vScreenML in 3 cases (points above the diagonal), whereas vScreenML proves superior in 12 cases (the other 8 cases were ties).

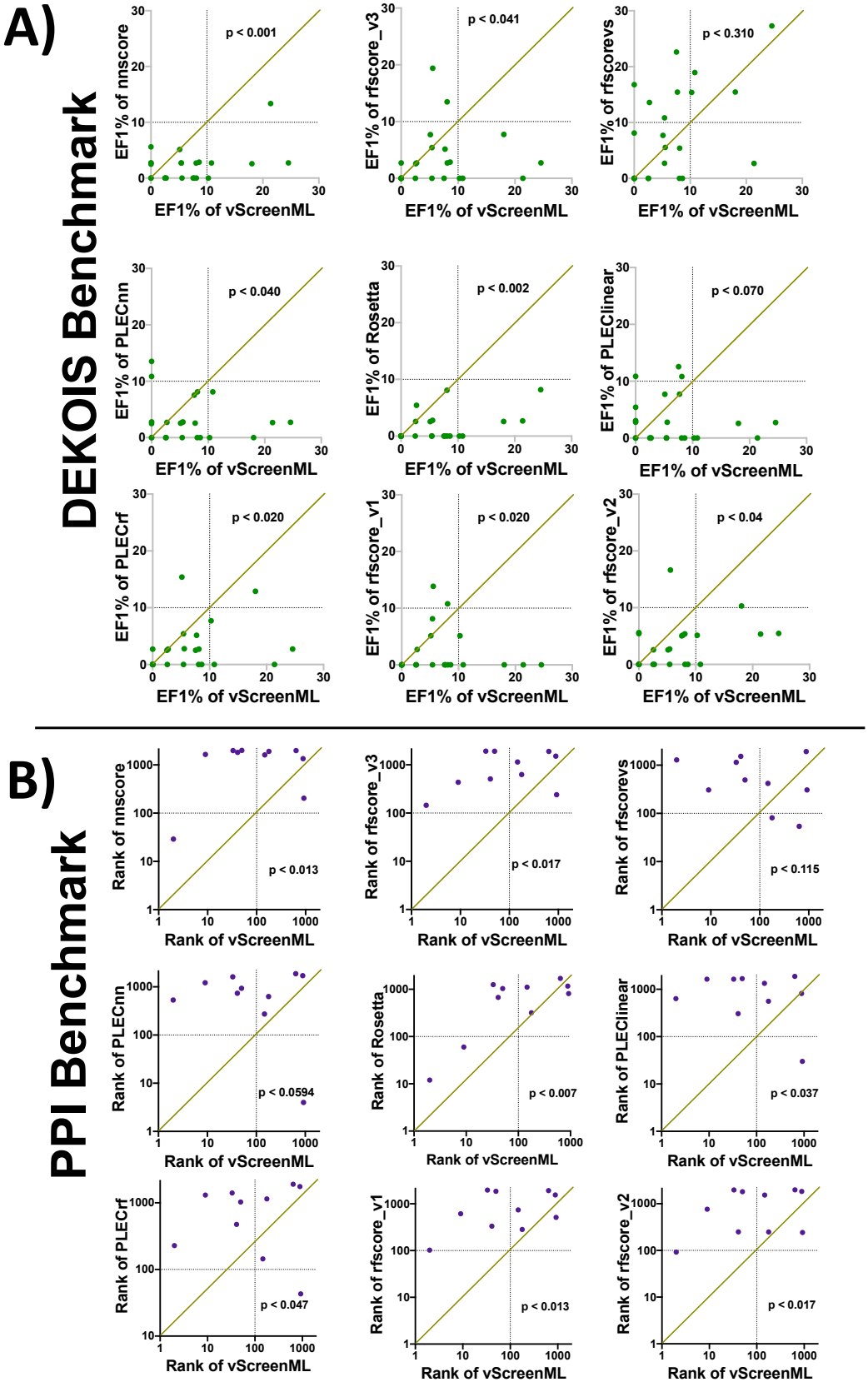

A)
DEKOIS Benchmark
EF1% of nnscore
p < 0.001
EF1% of rfscore_v3
p < 0.041
EF1% of rfscorevs
p < 0.310
EF1% of PLECnn
p < 0.040
EF1% of Rosetta
p < 0.002
EF1% of PLEClinear
p < 0.070
EF1% of PLECrf
p < 0.020
EF1% of rfscore_v1
p < 0.020
EF1% of rfscore_v2
p < 0.04
EF1% of vScreenML
B)
PPI Benchmark
Rank of nnscore
p < 0.013
Rank of rfscore_v3
p < 0.017
Rank of rfscorevs
p < 0.115
Rank of PLECnn
p < 0.0594
Rank of Rosetta
p < 0.007
Rank of PLEClinear
p < 0.037
Rank of PLECrf
p < 0.047
Rank of rfscore_v1
p < 0.013
Rank of rfscore_v2
p < 0.017
Rank of vScreenML

Figure 3: Comparing vScreenML to other scoring functions using two independent virtual screening benchmarks. Each benchmark is comprised of multiple protein targets, corresponding to points on these plots. **(A)** DEKOIS benchmark, comprised of 23 protein targets. For each target (individual dots), 30-40 active complexes and 800-1200 decoy complexes are provided. For a given target, each scoring is used to rank the set of complexes. For a given scoring function, the number of active complexes in the top 1% of all complexes is used to calculate the enrichment of actives relative to randomly selecting complexes; thus, *higher* numbers indicate better performance). When comparing vScreenML against another method, a point *below* the diagonal indicates superior performance by vScreenML for this particular target. **(B)** PPI benchmark, comprised of 10 protein targets. For each target, a single active complex is hidden amongst 2000 decoy complexes. Instead of using enrichment, the rank of the active compound (relative to the decoys) is calculated: thus, *lower* numbers indicate better performance. When comparing vScreenML against another method, a point *above* the diagonal indicates superior performance by vScreenML for this particular target. p-values in both cases were computed using the two-tailed Wilcoxon Signed-Rank test.

To evaluate in a statistically rigorous way which method was superior, we applied the (non-parametric) Wilcoxon Signed-Rank test: this paired difference test uses the rank values in the data, and thus it takes into account not just which method has higher EF-1%, but also the magnitude of the difference [55]. We used a two-tailed test, in order to assume no *a priori* expectation about what method would out-perform the other. At a threshold of $p<0.05$, this analysis shows that vScreenML out-performed 8 of the 9 alternate scoring functions to a statistically significant degree. Only RF-Score-VS was not out-performed by vScreenML at a statistically significant threshold; however, we note that about half of the 23 targets in this benchmark were included in training RF-Score-VS (and none for vScreenML), which may have provided it with a slight advantage.

To test these methods on a second independent virtual screening benchmark, we drew from our own prior studies of inhibitors of protein-protein interactions [55]. In the course of evaluating existing scoring functions, we had several years ago assembled a set of small molecules that engage protein interaction sites; 10 of these protein targets had not been included in training vScreenML. For each of these, we had previously compiled 2000 decoys with dissimilar chemical structure matched to the active compound's lipophilicity. The decoy compounds were already docked and energy minimized from our studies, making this "PPI set" a natural testbed for the newer methods that were not available at the time this benchmark was developed [55]. In contrast to the DEKOIS benchmark, the structures of the active

complexes are drawn from (energy-minimized) crystal structures, removing a potential source of variability (since mis-docked active compounds should not be labeled "correct" by a scoring function).

Because each protein target is only associated with a single active compound in this test set, we cannot meaningfully calculate enrichment factor; instead, after scoring each of the complexes we simply report the rank of the active compound. As there are 2001 complexes for each protein target, a method that performs as random would be expected to rank the active compound at position 1001, on average. After applying each of the same scoring functions used in the DEKOIS experiment, we find that for 5 of the 10 protein targets vScreenML ranks the active compound among the top 100 (i.e., top 5% of the compounds for a given target) (**Figure 3b**). The other scoring functions tested each ranked the active compound in the top 100 for at most one target, except for RF-Score-VS which met this criterion twice. Once again applying the Wilcoxon Signed-Rank test to these rankings, we once again conclude that vScreenML out-performs at a statistically significance degree all of these alternate scoring functions except for RF-Score-VS.

To determine whether vScreenML's impressive performance derived from its training on the D-COID set or from the broad collection of features it includes, we used D-COID to train a model using the features from RF-Score v1; our re-trained model preserves the same random forest framework and hyperparameters from the original model [1]. As noted earlier (**Figure 1b**), RF-Score v1 initially yields very little discriminative power when applied to the D-COID set; after re-training on this set, we find much improved separation of the scores assigned to active versus decoy complexes (**Figure S4ab**), though not close to the performance of vScreenML (**Figure 2f**). This re-trained variant of RF-Score v1 also out-performs the original RF-Score v1 on both the DEKOIS and the PPI benchmarks, albeit not to a level of statistical significance, and for the PPI benchmark it even ranks two actives in the top 100 for their corresponding protein targets (**Figure S4cd**). That said, the level of improvement is insufficient for the re-trained RF-Score v1 to out-perform vScreenML in either benchmark (**Figure S4ef**), consistent with their relative performance on D-COID set. Overall, these observations show that training using the

D-COID approach can certainly improve performance of existing scoring functions for other unrelated tasks; however, it also suggests that some part of vScreenML's power derives from the broad and diverse set of features that it uses.

1.3.4 Evaluating vScreenML in a prospective experiment

As noted earlier, it is absolutely essential to test new scoring functions in prospective experiments: this can readily determine whether performance in a given benchmark experiment is likely to extend into real future applications, and rule out any possibility that inadvertent information leakage allowed an overfit model to "cheat" in benchmark experiments. We selected as our representative target human acetylcholinesterase (AChE) because of its biomedical relevance and the availability of a straightforward functional assay (using commercially-available enzyme and substrate).

To ensure that our search for new candidate AChE inhibitors would not be limited by the chemical space present in a small screening library, we turned to a newly-available virtual library of "readily-accessible" but never-before-synthesized compounds [21]. At the time of our screen, this library was comprised of 732 million chemical entities that conform to historic criteria for drug-likeness [65,66]. Because building conformers and docking each entry in this library would be extremely computationally demanding, we instead took a two-step approach to finding candidate inhibitors. First, we explicitly docked a chemically-diverse set of 15 million representatives from the library, and applied energy minimization to the top 20,000 models from the crude docking step. We ranked each of these using vScreenML, and identified the top 100 candidates. For each of these 100 initial candidates, we returned to the complete compound library and identified 209 analogs on the basis of chemical similarity: after merging these with the parent compounds from each search, this led to a new focused library of 20,213 unique compounds. We structurally aligned each of these compounds back onto the parent docked model that led to their selection, re-minimized, and then used vScreenML to rank these second-stage candidates. We collected into a single list the 20 top-scoring compounds from the first round together with the 20 top-scoring compounds from the second round, noting that 4 compounds were included on both lists. We

eliminated compounds that were extremely close analogs of one another, and sought to purchase the remainder. Based on a standard filter [67], none of these structures were predicted to be PAINS (pan-assay interference) compounds. Ultimately 23 compounds were successfully synthesized, as selected by vScreenML without any human intervention.

We initially tested these compounds at a concentration of 50 μM for inhibition of AChE, using a colorimetric enzyme assay (**Figure 4a**). To our amazement, we found that nearly all of the 23 compounds selected by vScreenML showed detectable enzyme inhibition; of these, 10 of them provided more than 50% inhibition (indicating that these compounds' IC_{50} was better than 50 μM). Moreover, the most potent of these used a variety of diverse chemical scaffolds, although the most potent pair (AC6 and AC3) do share an extensive common substructure (**Figure 4b**). We then evaluated the activity of the most potent inhibitor, AC6: in the absence of any medicinal chemistry optimization, we found this compound to have an IC_{50} of 280 nM, corresponding to a Ki value of 173 nM (**Figure 4c**). Thus, applying vScreenML led to a much higher hit rate than observed in typical screening campaigns, and also yielded a much more potent starting point than is typically observed.

Unsurprisingly, the underlying model of the complex that was used by vScreenML to identify this compound shows extensive and nearly optimal protein-ligand interactions (**Figure 4d**). In principle, it should be the quality of these interactions that guided vScreenML to prioritize this compound for experimental validation. To rule out the possibility that vScreenML had instead somehow "recognized" AC6 as an AChE inhibitor from its training, we asked whether chemoinformatic approaches could have been used to find AC6.

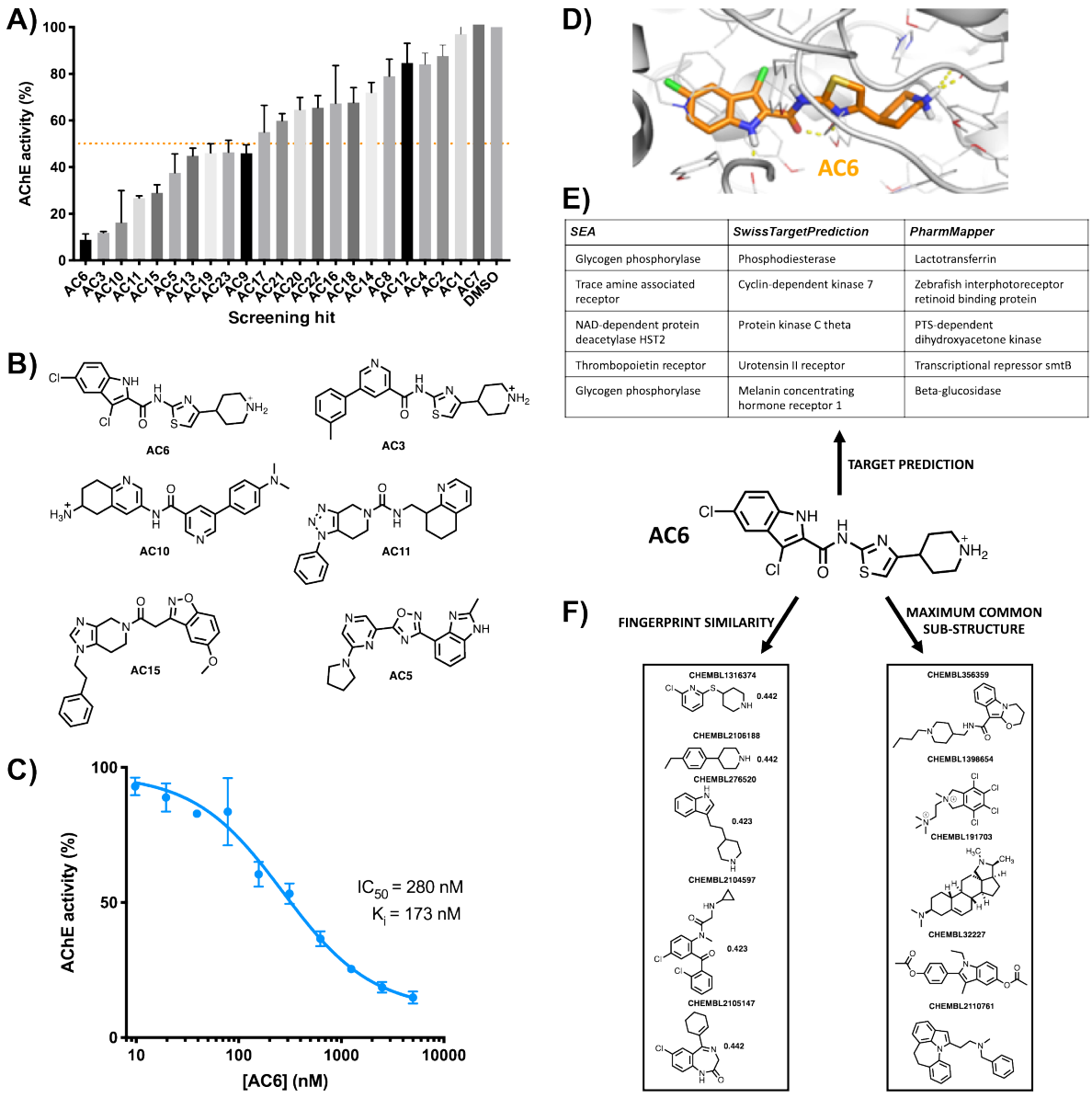


Figure 4: Prospective evaluation of vScreenML in a virtual screen against human acetylcholinesterase (AChE). (A) Of the 23 compounds prioritized by vScreenML for testing, at 50 μM nearly all of these inhibit AChE. **(B)** Chemical structures of the most potent hit compounds. **(C)** Dose-response curve for the most potent hit compound, AC6. **(D)** Model of AC6 (*orange sticks*) in the active site of the acetylcholinesterase (*light gray*). **(E)** Predicted activity of AC6 from three target identification tools: none of these identify AChE as a potential target of this compound, suggesting that this is a new scaffold for AChE inhibition. **(F)** Similarity searching against all compounds in ChEMBL designated as AChE inhibitors (either by fingerprint similarity of by shared substructure) finds no hits with discernible similarity, confirming that this is a new scaffold for AChE inhibition.

We first provided the chemical structure of AC6 to three different "reverse screening" methods: Similarity Ensemble Approach (SEA) [68], SwissTargetPrediction [69,70], and PharmMapper [71,72].

Each of these tools look for similarity of the query compound against all compounds with known bioactivity, then they rely on the fact that similar compounds have similar bioactivity to predict the likely target(s) of the query compound. SEA and SwissTargetPrediction carry out this search on the basis of 2D similarity (i.e. similar chemical structures), whereas PharmMapper evaluates 3D similarity (i.e. shared pharmacophores). We took for each method the top 5 predicted activities for AC6, but found that none of these methods included AChE among their predictions (**Figure 4e**). All of these methods do include AChE among their list of potential targets, however, as confirmed by ensuring that this prediction emerges when these servers are provided with the structure of previously-described AChE inhibitor donepezil (**Figure S5**).

To directly determine the AChE inhibitor described to date that is most similar to AC6, we compiled from ChEMBL all 2742 compounds reported to have this activity. We then screened this collection to determine their similarity to AC6, as defined by either chemical fingerprints or by shared substructure. The 5 most similar compounds as gauged by either approach bear no obvious similarity to AC6 (**Figure 4f**): collectively then, these experiments confirm that AC6 is indeed a novel chemical scaffold with respect to its inhibition of AChE, and could not possibly have been identified by vScreenML through inadvertent leakage during the model's training.

1.4 Discussion

At the outset of this work, we noted that typical virtual screening studies report hit rates of about 12%, with the most potent reported compound having K_d or K_i value of ~3 μM (with the caveat that some of these relied on additional optimization beyond the initial screen) [24]. Obviously the results of our screen against AChE using vScreenML far surpass these mileposts; in light of this, it important to carefully consider the potential contributions to vScreenML's performance in this experiment.

First, we re-emphasize the dissimilarity between AC6 and any known AChE inhibitor: this makes it exceedingly unlikely that vScreenML found AC6 simply on the basis of having been trained on some close analog.

Second, we carried out a non-standard two-step screening strategy to efficiently explore the complete Enamine collection, hoping to essentially carry out an internal round of medicinal chemistry optimization before testing any compounds explicitly. Tracking the provenance of our most potent compounds, however, we discovered that all four of our most potent compounds had already been identified in the first of the two screening steps (**Table S1**). A previous virtual screen of the Enamine library [21] explicitly docked all compounds from the library, at a time that the library comprised "only" 138 million compounds, and found through retrospective analysis that picking single representative compounds from a cluster of analogs would typically not yield sufficient docking score for the cluster to be advanced for further exploration. In essence, both our results and the observations from this previous screen suggest that the SAR landscape may not be sufficiently smooth to allow potentially promising scaffolds to be identified from a single arbitrary representative: rather, finding the best hits (on the basis of docking scores) does unfortunately require explicitly screening each member of the library individually. In this context, then, it is unlikely that the observed performance of vScreenML can be attributed to having used a two-step strategy for screening the Enamine library.

In this vein, we also note that our screening strategy was allowed to explore an unusually large chemical space comprising 732 million synthetically-accessible compounds. However, seven of our top ten compounds (those with IC_{50} values better than 50 μM) had already been identified in the first screening step (**Table S1**), owing to the ineffectiveness of identifying useful scaffolds from a single representative compound. The bulk of the success in this screen was essentially achieved by screening a library of 15 million diverse compounds, which is by no means unprecedented and has not led to such dramatic success in the past.

Importantly, we cannot rule out the prospect that the performance we observe here is a result of AChE being an unexpectedly easy target. It is certainly the case that virtual screening hit rates against

GPCRs are often much higher those obtained for other target classes [24]. Indeed, careful examination of the literature showed that some of the studies reporting virtual screens against AChE [73-76] do indeed find considerably higher hit rates and more potent compounds than the median values we quote across all target classes. In light of these other results, then, a degree of caution must be exercised before extrapolating the performance of vScreenML in this prospective AChE benchmark to other target classes; further evaluation will be needed to explicitly determine whether vScreenML affords similarly outstanding results in future screening experiments.

At the same time, however, results of retrospective benchmarks comparing vScreenML to other scoring functions are unambiguous. As described, vScreenML dramatically outperforms eight other modern machine learning scoring functions on both the DEKOIS and the PPI benchmark sets. Both benchmarks were carried out with careful vigilance to ensure that information from training could not contaminate the test data. In the past, we strongly suspect inadvertent overtraining of this type has limited the utility of other models and at the same time provided artificially inflated performance on initial (retrospective) benchmarks. Indeed, a recurrent disappointment from many past machine learning scoring functions has been their inability to translate performance from retrospective benchmarks into equivalent results in future prospective applications. For example, three years after publication of nnscore [35] this program was used in a screen against farnesyl diphosphate synthase, and only provided one hit with IC_{50} of 109 μM (from ten compounds tested) [8]. Where possible, then, we strongly urge incorporation of careful prospective evaluations alongside retrospective benchmarks, as a safeguard against potentially misleading performance from the latter. Already such prospective experiments have been included in other recent studies [39,77], strongly supporting transferability of the underlying methods.

The ability to readily compare vScreenML against other machine learning scoring functions was greatly facilitated by the Open Drug Discovery Toolkit (ODDT) [78], which provides implementations of multiple methods. Direct head-to-head evaluations of this manner are indeed critical to explore the relative strengths of different approaches, ideally across diverse types of benchmarks. While vScreenML does incorporate a broad and distinct set of features, these have been largely collected from other

approaches: there is nothing particularly unique associated with the individual features. Rather, we believe that the superior performance of vScreenML in our retrospective benchmarks is instead attributable primarily to the strategy used in training the model. Whereas scoring functions have historically focused on recapitulating binding affinities of complexes, vScreenML is unique in having been trained to distinguish active complexes from extremely challenging decoys in the D-COID set. Indeed, the overarching hypothesis of our study was using truly compelling decoys that are representative of (inactive) compounds selected from actual virtual screens we would lead to a model capable of distinguishing precisely these cases.

Thus, the D-COID set represents an important resource for driving development of improved scoring functions beyond vScreenML, and accordingly we have made this dataset freely available for this purpose (see *Methods*).

1.5 Methods

1.5.1 Accessing these tools

The D-COID dataset is available at https://data.mendeley.com/datasets/8czn4rxz68/ [79]. vScreenML is available at https://github.com/orgs/karanicolaslab/vScreenML .

1.5.2 Building the D-COID set

The overarching goal of our study was to train a model for real virtual screening applications. We therefore included in D-COID only active complexes that included representative drug-like ligands, and excluded chemical matter that did not reflect the composition of the screening libraries we prefer to use.

We downloaded from the Protein Data Bank (PDB) [80] all protein-ligand complexes (56,195 entries as of May 2018), and then restricted this set to crystal structures with resolution better than 2.5 Å (43,148 complexes). We then drew from Ligand Expo [81] to define a set of 2937 specific ligands found in the PDB that we deemed ineligible for our study: these include nucleotide-like molecules (e.g., ATP),

co-factors (e.g., NAD), metal-containing ligands (e.g., heme), crystallographic additives (e.g., PEG), and covalent ligands. We filtered to retain only complexes that included an eligible ligand, and did not have an additional ligand within 12 Å of the eligible ligand (leaving 26,271 complexes). To focus training on precisely the type of chemical matter used in our virtual screens, we then applied to this collection the same stringent filter we use when building our screening libraries: molecular weight between 300-400 Da and clogP 1-4. This filter drastically cut down the size of our collection (to 2,075 complexes). Finally, complexes with double occupancy or ambiguous density were manually excluded, leaving a high-quality collection of 1,383 active complexes.

For each of these active complexes, we extracted the ligand and used the Database of Useful Decoys Enhanced (DUD-E) [47] server to generate 50 property-matched decoys: compounds with similar physicochemical properties but dissimilar chemical topology. For each of these decoy compounds, we used OpenEye's OMEGA [82] to generate 300 low-energy conformers, and then used ROCS [83] to align each of these to the structure of the active conformer from the PDB. The three decoys that best matched the three-dimensional shape and pharmacophoric features of the active conformer were identified on the basis of their Tanimoto-Combo score; this led to a total of 4,149 decoy compounds. By virtue of having aligned the conformers of the decoys to the active conformation to evaluate their similarity, already the alignment was available for placing the decoy compound in the corresponding protein's active site. We later discovered that 39 of these decoy compounds included chemical features that could not be processed by the programs we used to extract structural features for vScreenML; these decoys were removed, leading to a total of 4,110 decoy complexes.

Finally, to present both the active and decoy complexes in a context mimicking that of a virtual screening output, we subjected all complexes to standard energy minimization in Rosetta [53].

1.5.3 Extracting structural features

For each of the minimized active and decoy complexes, structural features were extracted first using the Rosetta ("REF15") energy function [53]. Ligand properties were calculated using ChemAxon's

cxcalc[61], and the ligand's conformational entropy was estimated using OpenEye's SZYBKI tool[62]. The open source implementations of RF-Score [1] and BINANA [60] were used to calculate structural features from these two programs. The complete list of vScreenML's features is presented in **Figure S1**.

1.5.4 Machine learning

We considered a total of nine classification algorithms in this study, using the Python implementations of each: Support Vector Machine (SVM) [84], Gradient Boosting (GB) [85], Extreme Gradient Boosting (XGB) [86], Random Forest (RF) [87], Extremely Randomized Trees (ET) [88], Gaussian Naïve Bayes (GNB) [89], k-Nearest Neighbor (kNN) [89], Linear Discriminant Analysis (LDA) [89], and Quadratic Discriminant Analysis (QDA) [89].

Training was carried out using 10-fold cross-validation; splitting the dataset into 10 subsets was carried out in a stratified manner to ensure that the overall ratio of actives to decoys was preserved in each split. For XGBoost hyperparameter optimization, we carried out a grid search to find the set of parameters that gave the best cross-validation accuracy (splitting out a separate validation set from the data).

To re-train RF-Score v1 under D-COID, we used a standard random forest model with hyperparameters n_estimators=500 and max_features=5 (drawing these values from the original study describing RF-Score v1 [1]).

1.5.5 Virtual screening benchmarks

Comparisons between scoring functions was enabled by the Open Drug Discovery Toolkit (ODDT) [78], which provides implementations of nnscore (version 2), RF-Score v1, RF-Score v2, RF-Score v3, PLEClinear, PLECnn and PLECrf at https://github.com/oddt/oddt. The implementation of RF-Score-VS was obtained from https://github.com/oddt/rfscorevs.

In both the DEKOIS and the PPI benchmark experiments, we carefully sought to minimize any potential information leakage from vScreenML's training (on D-COID) and the targets present in these benchmark sets. Excluding a specific complex present in both sets is insufficient, because the structure of

a close chemical analog bound to the same target protein could still provide an unfair advantage. For this reason, we excluded from these benchmarks sets any protein targets present in D-COID (on the basis of shared Uniprot IDs). This reduced the number of DEKOIS targets from 81 to 23, and the number of PPI targets from 18 to 10.

For the DEKOIS set, we docked both the actives and the decoys to their respective target protein using OpenEye's FRED [56], then applied energy minimization in Rosetta. For the PPI set, active complexes were minimized starting from their crystal structures; decoy complexes were generated by docking with FRED then energy minimized.

Statistical analysis was carried out using the (two-tailed) Wilcoxon Signed-Rank test as implemented in Python. Comparisons were applied directly to the EF-1% values for the DEKOIS experiment, and to the $\log_{10}$ of the ranks in the PPI experiment.

1.5.6 Virtual screen against acetylcholinesterase

We began by downloading from the chemical vendor Enamine the "diverse set" of 15 million compounds representative of their REAL database (732 million compounds). For each compound we used OMEGA [82] to generate 300 low-energy conformers, and used FRED [56] to dock these against the crystal structure of human acetylcholinesterase solved in complex with potent inhibitor donepezil (PDB ID 4ey7) [90]. We carried forward the top 20,000 complexes (on the basis of FRED score) for Rosetta minimization, and used each of these minimized models as input for vScreenML.

For each of the top 100 complexes (as ranked by vScreenML), we extracted the ligand and used this to query the Enamine database for analogs. Each query returned 210 analogs; because 787 of these were redundant, this led to a new collection of 20,213 unique compounds for the second stage of screening. Each of the compounds in this new library was used to build 300 conformers, and ROCS was used to select the conformer that allowed for optimal alignment onto the ligand in the complex from the first round of screening. The resulting models were energy minimized in Rosetta, then used as input for vScreenML.

Models from both the first and second rounds of screening were collected together, and the top-ranked models from vScreenML were identified, and the top-scoring 32 compounds were requested for synthesis. Of the requested compounds, 23 were successfully synthesized and delivered for testing.

1.5.7 Acetylcholinesterase inhibition assay

Compounds were tested for inhibition of human acetylcholinesterase (AChE) using a colorimetric assay [91]. Acetylthiocholine is provided as substrate, which is hydrolyzed by AChE to thiocholine; the free sulfhydryl then reacts with Ellman's reagent (5,5'-dithiobis-(2-nitrobenzoic acid); DTNB) to yield a yellow product that we detected spectrophotometrically at 410 nm. AChE, acetylthiocholine, and DTNB were acquired together as the Amplite^{TM} Colorimetric assay kit (AAT Bioquest). Assays were carried out in 0.1 M sodium phosphate Buffer (pH 7.4), 1% DMSO, with 0.01% Triton. Assays were carried out in 96-well plates in reaction volumes of 100 μL, and absorbance was monitored for 30 min. The rate of product formation was determined by taking the slope of the absorbance as a function of time, and normalized to that of DMSO alone to yield percent inhibition for each well.

IC_{50} values were obtained from dose-response curves spanning inhibitor concentrations from 10 nM to 50 μM. To determine K_i, we first determined the K_m value for substrate acetylthiocholine under our assay conditions. This allowed the Cheng-Prusoff equation [92] to be used for obtaining K_i from IC_{50}, assuming classic competitive inhibition.

1.5.8 Novelty of AC6 as an AChE inhibitor

For each of the target identification methods (Similarity Ensemble Approach (SEA) [68], SwissTargetPrediction [69,70], and PharmMapper [71,72]), we used the corresponding web servers to generate predictions for AC6.

To find the most similar known AChE ligands, we searched ChEMBL [93] for acetylcholinesterase and downloaded all 2742 hits in SDF format. We then used ChemAxon's Standardizer tool to remove counterions from compounds in salt form. Using RDKit [94] we generated

Morgan fingerprints with radius of 2 for each of the ChEMBL ligands, then evaluated the Dice similarity of these fingerprints relative to that of AC6. We also used RDKit to evaluate the maximum common substructure (MCS) between AC6 and each of the ChEMBL ligands, setting ringMatchesRingOnly=True and completeRingsOnly=True. We ranked the resulting matches based on the number of atoms and bonds in the common substructure.

1.6 Acknowledgements

We thank Joanna Slusky for a useful suggestion regarding presentation of the figures, and Juan Manuel Perez Bertoldi for his initial application of vScreenML to the PPI benchmark set. We thank ChemAxon for providing an academic research license. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) allocation MCB130049, which is supported by National Science Foundation grant number ACI-1548562. This work was supported by grants from the National Institute of General Medical Sciences (R01GM099959, R01GM112736, and R01GM123336) and from the National Science Foundation (CHE-1836950). This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA006927.

CHAPTER 2: Rationally designing inhibitors of the Musashi protein-RNA interaction by hotspot mimicry

Nan Bai[1,4†], Yusuf Adeshina[2,4†], Lan Lan[1], Petr B. Makhov[4], Yan Xia[1], Ragul Gowthaman[2], Sven A. Miller[4], David K. Johnson[2], Yanis Boumber[4], Liang Xu[1,3], and John Karanicolas[1,2,4*]

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

[3] Department of Radiation Oncology, University of Kansas, Lawrence, KS 66045

[4] Program in Molecular Therapeutics, Fox Chase Cancer Center, Philadelphia, PA 19111

[†]Equal author contributions.

* To whom correspondence should be addressed.

E-mail: john.karanicolas@fccc.edu

215-728-7067

2.1 Abstract

RNA-binding proteins (RBPs) are key regulators of post-transcriptional gene expression, a role found to be crucial in many 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, stem-cell marker that is upregulated in many cancers. We report novel inhibitors that are active in biochemical and biophysical assays against Musashi-1, and then demonstrate how these inhibitors can also be used as tool compounds to probe the activity of close homolog Musashi-2. Finally, we explore the selectivity of these compounds, and consider the prospects of identifying potential off-target interactions by searching for other RBPs that recognize their cognate RNAs using similar interaction geometries (i.e. hotspot pharmacophores). 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 one of 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. At the same time, we additionally expect that in future these compounds may serve as warheads for new PROTACs that selectively degrade RNA-binding proteins including Musashi.

2.2 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 [95]. Mammalian proteomes are thought to include upwards of 800 RBPs [96,97], 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 [98-100].

To date, there exist few classes of compounds that target protein-RNA interactions. Inhibitors of certain RBPs have been identified via high throughput screening [101,102], including one series from virtual screening that competes with double-stranded RNA for binding to toll-like receptor 3 [103], and a number of compounds have been reported that disrupt binding by interacting with the RNA rather than with the RBP [104,105]. 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 [106-112], 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. 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 [110,113,114], such as anti-HIV-1 NNRTIs, that bind to secondary sites on the protein target and shift its conformation to an inactive state. 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 [115], 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 [116].

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 [117-119]. This observation, in turn, motivated several groups to mimic these key interactions when designing small-molecule inhibitors [120-123]. 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.

2.3 Computational Approach

New computational methods have been implemented in the Rosetta software suite [124] 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 and beyond. Computational methods are summarized below and presented in further detail in the *Supporting Methods* section.

2.3.1 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 [117,119]. 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**).

2.3.2 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 [125] that are both commercially available, and predicted to have drug-like physicochemical properties. We used OMEGA

(OpenEye Scientific Software, Santa Fe, NM) [126,82,127] to build low-energy conformations of each compound, then ROCS (OpenEye Scientific Software, Santa Fe, NM) [128,49] 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 [124].

2.3.3 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 [129]. 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 [130]. Accordingly, mutations to the protein that remove these aromatic sidechains have been shown to disrupt binding in representative RRMs [130,131], as has introduction of non-canonical bases to the RNA that alter the pattern of hydrogen bonding groups [132-134]. 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 [130].

Mammalian Musashi-1 (Msi1) recognizes its cognate RNAs through a pair of RRMs, RRM1 and RRM2 [135]. 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 [136,137]. *NUMB* mRNA encodes an inhibitor of Notch, so translational inhibition by Msi1 triggers Notch signaling and thus promotes self-renewal and cell survival [137,138]. 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 [139-145]. 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 [146-148].

2.4 Results

2.4.1 Computational screening against Msi1 RRM1

We applied our "hotspot mimicry" approach to the Musashi-1 RRM1 / *NUMB* mRNA complex [135], and found a single hotspot pharmacophore derived from an adjacent pair of buried nucleobases, Adenine106 and Guanine107 (**Figure 1a**). Because no ordered water molecules were included in this NMR structure, the resulting pharmacophore did not include any explicit contribution from solvent. 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 RRM1 / RNA interaction, we used a fluorescence polarization (FP) 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 RRM1 (**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. 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 (**Figure 1e**), its close analog R4, and R7 (**Figure S2**). Among these three, only R12 showed inhibition in FP competition assay (**Figure 1f**); that said, the binding affinity could not be reasonably quantified because of the low solubility of R12.

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 1g)**. Mimicry of these interactions through the hotspot pharmacophore allows the hit compounds to recapitulate these interactions, as exemplified by R12 (**Figure 1h**). In this model R12 adopts a similar three-dimensional geometry as the hotspot pharmacophore, and thus recapitulates its aromatic stacking and polar interactions (**Figure 1h**).

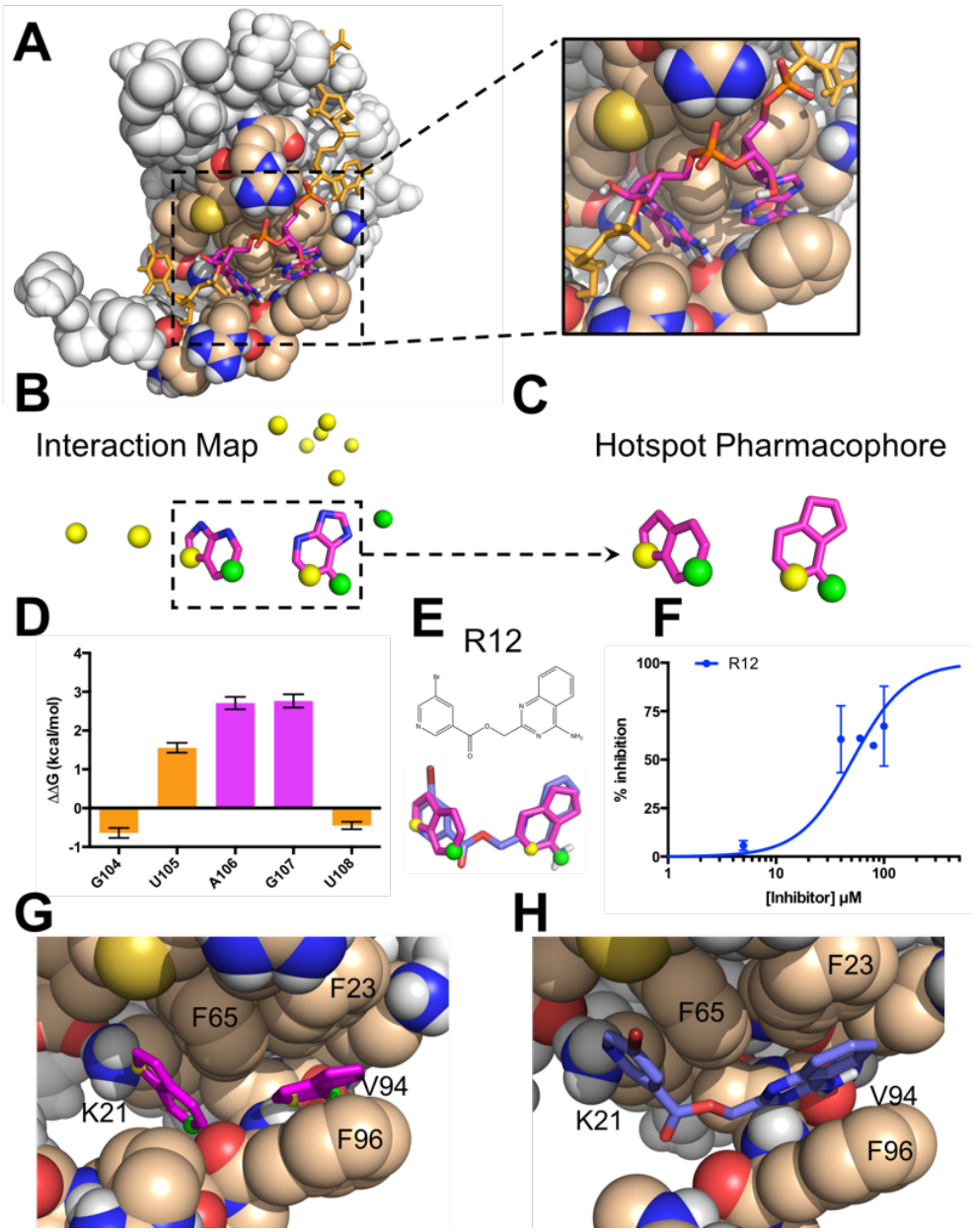


Figure 1: The hotspot mimicry approach. We demonstrate this approach by applying it to the Msi1 / *NUMB* mRNA interaction. **(A)** The structure of the Msi1 / RNA complex. The RNA (*sticks*) wraps around the protein (*spheres*). Two adjacent bases, A106 and G107 (*magenta*), are buried in a shallow pocket 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, searching for compounds that would mimic the three-dimensional features of the pharmacophore. The screen led to the identification of compound R12, which mimics the geometry of the rings and provides three of the four desired hydrogen bonding groups. **(F)** R12 competes with fluorescein-labeled RNA for Msi1 binding, as observed through a fluorescence polarization assay. These data do not allow the binding affinity to be confidently determined. **(G)** 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. (H) 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.

2.4.2 SAR study of R12 derivatives

With R12 as starting point and guided by our structural model (**Figure 2a**), we set out to improve potency of this interaction. While our initial screen had been restricted to ~7 million compounds in the ZINC database, the newly-available Enamine database included ~8 *billion* compounds: each of these not previously synthesized by Enamine, but readily available on-demand. Thus, the Enamine database afforded us an exciting opportunity to carry out "SAR-by-catalog" at a much larger scale than would otherwise have been possible.

The low solubility of R12 prompted us to begin by looking for alternatives to the bromopyridine group on the left-hand side. Amongst a set of 16 bespoke analogs that we purchased, we found that two of these, R12-7 and R12-8 (**Table S2**) demonstrated superior inhibition and solubility compared R12. As a secondary validation assay we used Differential Scanning Fluorimetry (DSF) to confirm the interaction and found that R12-7 induced noisy changes in the protein's melting temperature (**Figure S3a**), but R12-8 led to continuous stabilization with increasing dose (**Figure S3b**). Out of concern that R12-7's activity may be associated with compound aggregation, we elected to proceed with R12-8 (**Figure 2b**).

Refinement of R12-8 after aligning it to our initial model of R12 suggested a potential reason for the improved potency. In our initial model of R12, the carbonyl oxygen in its ester linker was positioned in close proximity to the Phe96 backbone carbonyl of Msi1 (**Figure 2a**); beyond simply the lost opportunity for an intermolecular hydrogen bond with the backbone, we expected electrostatic repulsion between these two negatively charged moieties. By contrast, this linker has shifted slightly in our model of R12-8, turning

the ester group upward to face solvent, and instead engaging Msi1's backbone carbonyl using R12-8's newly-added amine (**Figure 2b**). Whereas R12 matched only three of the four desired hotspot pharmacophore features, our model of R12-8 now matched all four (**Figure S3c**).

We next sought to optimize the right-hand side of this compound, and purchased another 50 custom analogs from Enamine. While the most potent compound from this second round was R12-8-46, we avoided this compound because catechols are well-known candidate PAINS (pan-assay interference) compounds. Instead, we focused our attention on R12-8-44, which provided more than ten-fold improvement in potency (**Figure 2c**) and a consistently increasing melting temperature with increasing ligand concentration (**Figure S4**). R12-8-44 retains the same polar interactions as R12-8 (and R12), but has slightly different aromatic stacking by replacing R12-8's 4-quinazolinamine group for 5,6-dimethylthieno[2,3-d]pyrimidin-4-amine.

Satisfied with this large improvement, we returned to the left side of the compound. Based on our earlier SAR, we sketched ideas for preferred compounds; however, we found that preserving the right-hand side, restricted our choices somewhat. We therefore elected to purchase the compounds with the highest 2D similarity to the ideas we had laid out (**Figure S5**). Amongst these seven compounds, we found that R12-8-44-3 yielded another 2-fold improvement in potency, ultimately providing an IC_{50} value of 9 μM (**Figure 2d**). In parallel, we also explored alternatives to the ester linker, which may represent a metabolic liability. Though we could not yet incorporate the fully-optimized left-hand side at the time of these studies, we tested four alternate linkers, and found that activity was retained when replacing the ester with a thioether in R12-8-44-lk2 (**Figure 2e**). Our strong focus in this first study on restricting our optimization to purchasable compounds eliminated certainly very natural choices, including merging the promising features of R12-8-44-lk2 with R12-8-44-3, or even replacing R12-8-44-3's ester with an amide; these ideas will be explored in subsequent studies.

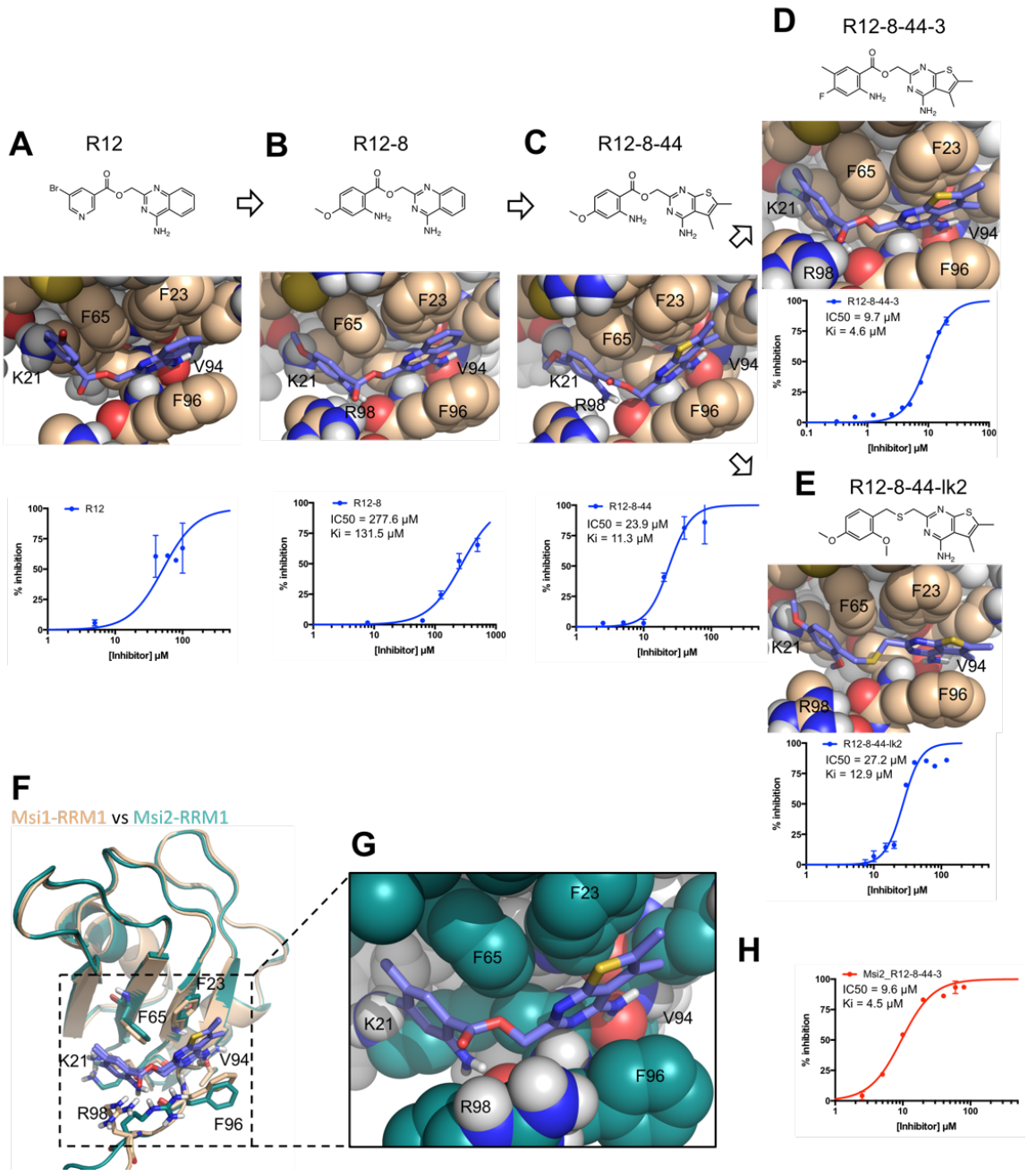


Figure 2: Optimization of R12 to the dual Msi1/Msi2 inhibitor R12-8-44-3. (A) R12 was the starting point for optimization, as identified from the computational screen of a limited library; availability of a much larger library was used to enable optimization. **(B)** The left-hand side of R12 was replaced to improve solubility and potency, yielding R12-8. **(C)** The right-hand side of R12-8 was replaced to improve potency, yielding R12-8-44. **(D)** Further exploration of the left-hand side provided improved potency, in R12-8-44-3. **(E)** Exchanging the ester linker for a thioether did not diminish activity, as found in R12-8-44-lk2. **(F)** Superposition of R12-8-44-3 models bound to Msi1 (*wheat*) and Msi2 (*green*). **(G)** Zoomed-in view of the Msi2 model confirms that the expected interactions are unchanged relative to the Msi1 model. **(H)** R12-8-44-3 competes with fluorescein-labeled RNA for Msi2 binding.

2.4.3 Inhibition of Musashi-2

While expression of Msi1 is tissue-restricted, its homolog Msi2 is ubiquitously expressed [149,150]. Moreover, functional redundancy between the two Musashi family members has led to the proposal that it would be most desirable to have a dual inhibitor that acts on both proteins [148]. Like Msi1, Msi2 includes two RRM domains; the first of these shares 80% sequence identity with Msi1 RRM1. 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 S6**); based on this model, we anticipated that the R12-8-44-3 would also show activity against Msi2.

With this hypothesis in mind, we first built a model of R12-8-44-3 bound to Msi2 by starting from our Msi1-bound model, and replacing the 17 residues that differ between Msi1 and Msi2 (**Figure 2f**). The resulting Msi2-bound model (**Figure 2g**) is essentially identical to our earlier Msi1-bound model (**Figure 2d**), implying that R12-8-44-3 should also inhibit Msi2. We tested this using the same FP competition assay, and confirmed that R12-8-44-3 indeed inhibits Msi2 with comparable activity as it inhibits Msi1 (**Figure 2h**). Thus, we have confirmed that R12-8-44-3 is a dual inhibitor of both Msi1 and Msi2, with the similar potency for each isoform.

2.4.4 Biophysical characterization of R12-8-44-3

To further characterize R12-8-44-3, we next advanced it to differential scanning fluorimetry (DSF) as an orthogonal secondary assay. Surprisingly, addition of R12-8-44-3 did not increase Msi1's melting temperature as we expected – and as we observed for its parent compounds R12-8 (**Figure S3b**) and R12-8-44 (**Figure S4a**). In fact, it acted in the opposite direction: addition of R12-8-44-3 consistently *decreased* Msi1's melting temperature (**Figure 3a**). While this could be a sign of apparent inhibition occurring through aggregation, there are indeed true inhibitors that have been shown to reduce their target protein's melting

temperature [151,152]. An alternate explanation could simply be that Msi1's folding landscape is not simply two-state, and that a partially-folded intermediate affects the unfolding transition.

To test the hypothesis that R12-8-44-3 engages Msi1 though specific binding interactions (rather than non-specific aggregation), we used HSQC chemical shift mapping. To facilitate interpretation of the spectra, we used only Msi1 RRM1 (rather than the construct with domains RRM1 and RRM2 used in the studies described above). We noted that R12-8-44-3 binds slightly less tightly to the RRM1-only construct (**Figure S7**), consistent with our model of binding at the C-terminus of RRM1. Nonetheless, the smaller RRM1-only construct facilitated collection and interpretation of the HSQC spectra.

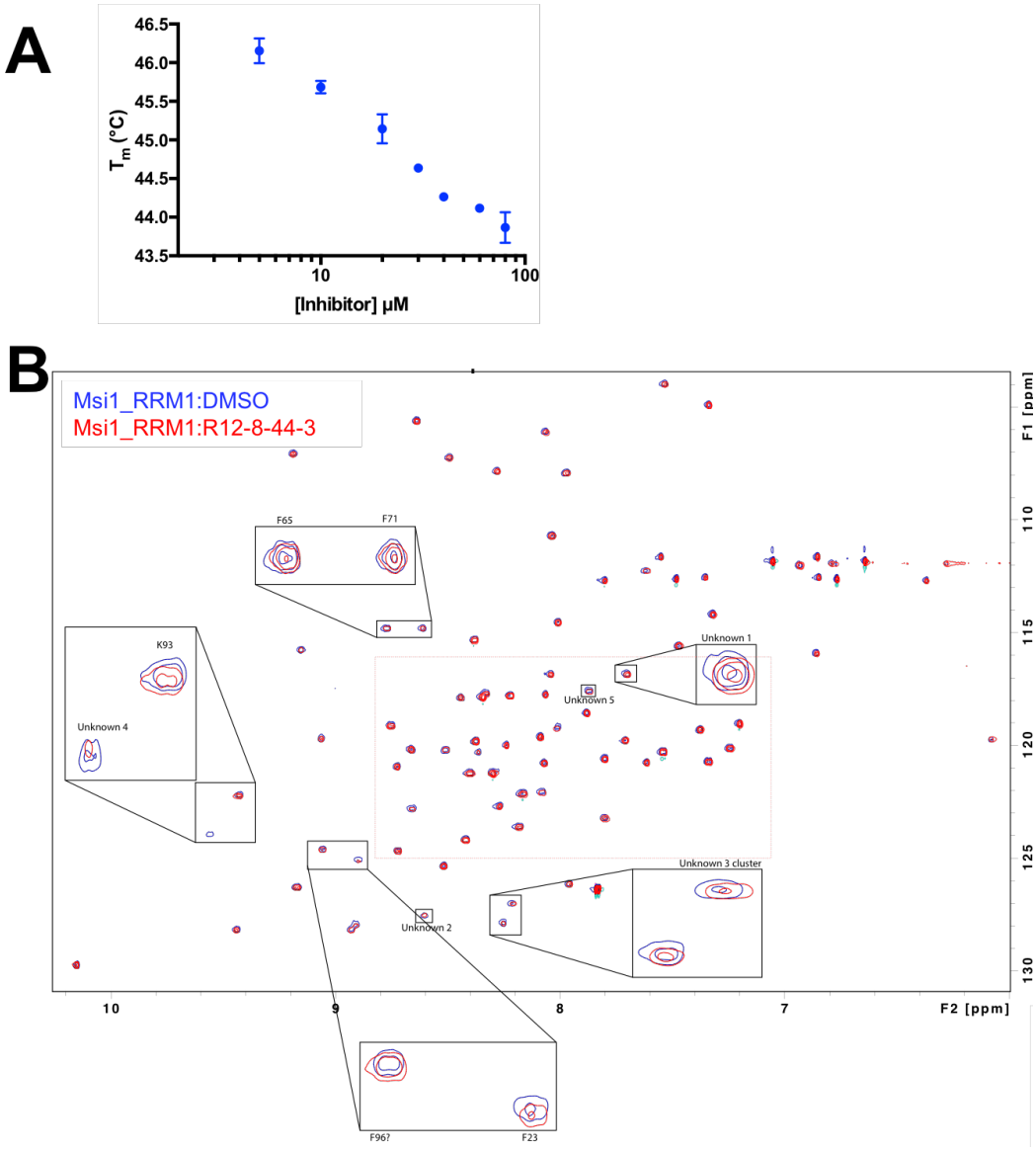


Figure 3: Biophysical characterization of R12-8-44-3. (A) Differential scanning fluorimetry shows that addition of R12-8-44-3 decreases the melting temperature of Msi1 in a concentration-dependent manner. **(B)** HSQC spectrum of Msi1 RRM1 collected in the presence and absence of R12-8-44-3. Peaks showing the strongest chemical shift difference are labeled (Phe23, Phe65, Phe96, Lys93).

We collected HSQC spectra for this construct in the presence and absence of R12-8-44-3 (**Figure 3b**), using previously-reported assignments from a very similar construct [135]. Importantly, we find that only a small number of peaks respond to addition of R12-8-44-3: this confirms specific binding to Msi1, rather than non-specific interactions that would imply aggregation. Gratifyingly, the peaks with the largest chemical shift differences were three aromatic residues (Phe23, Phe65, Phe96), all of which comprise the expecting binding site for R12-8-44-3 (**Figure 2d**). Perturbation of Lys93 is also evident, consistent with this binding site. Overall, these results strongly support interaction of R12-8-44-3 with the intended binding surface from our computational designs.

2.4.5 Exploring selectivity of R12-8-44-3

Many RRM proteins recognize their target RNAs with high sequence specificity, through additional interactions outside the central RNA dinucleotide [130]. 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 (PDB), 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 4a**). 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. In addition to Msi1, there are two other RRM-domain proteins in the PDB that recognize an A-G as the dinucleotide pair: human heterogeneous nuclear ribonucleoprotein A1 (hnRNP A1) [153], and yeast Prp24 [154].

We applied this approach first to a hypothetical compound, comprised of adenine and guanine attached by a flexible linker (**Figure 4b**). We built low-energy conformers of this compound, and then evaluated how closely this compound could mimic the three-dimensional geometry of each hotspot pharmacophore found in our library. While this artificial compound can indeed adopt a conformation that aligns well to the Msi1 hotspot pharmacophore (a score of 0.936) and hnRNP A1 (a score of 0.965), this hypothetical A-G compound can also be matched to many other hotspot pharmacophores from the PDB just at least as well as these (**Figure 4b**). Thus, this implies that such a compound would bind to many other off-target RBPs, in addition to Msi1 and hnRNP A1. In a sense, this observation underscores the *lack* of selectivity that one might expect from simply mimicking the nucleosides' chemical structure, without consideration of three-dimensional geometry.

We next carried out the same analysis for our starting compound R12 and optimized compound R12-8-44-3 (**Figure 4b**). We found that these two compounds both matched to the Msi1 hotspot pharmacophore *much* better than they match to any other hotspot pharmacophore extracted from the PDB. This result is unsurprising, given that R12 and R12-8-44-3 lack certain polar groups from the A-G pair (those that did not participate in the Msi1 pharmacophore), and also that these compounds have a restricted geometry that allows them only to mimic the specific orientation of the bases needed to complement Msi1 and hnRNP A1.

The reason for R12 and R12-8-44-3 matching well to the hotspot pharmacophore from hnRNP A1 is because of its strong similarity to Msi1's hotspot pharmacophore: although one of the bases is flipped, the structure of the central RNA dinucleotide in these two complexes is virtually superposable (**Figure 4c**). Overall, this analysis suggests that R12-8-44-3 is likely to be selective for Msi1 over the majority of other RBPs, but that hnRNP A1 could be a potential off-target interaction.

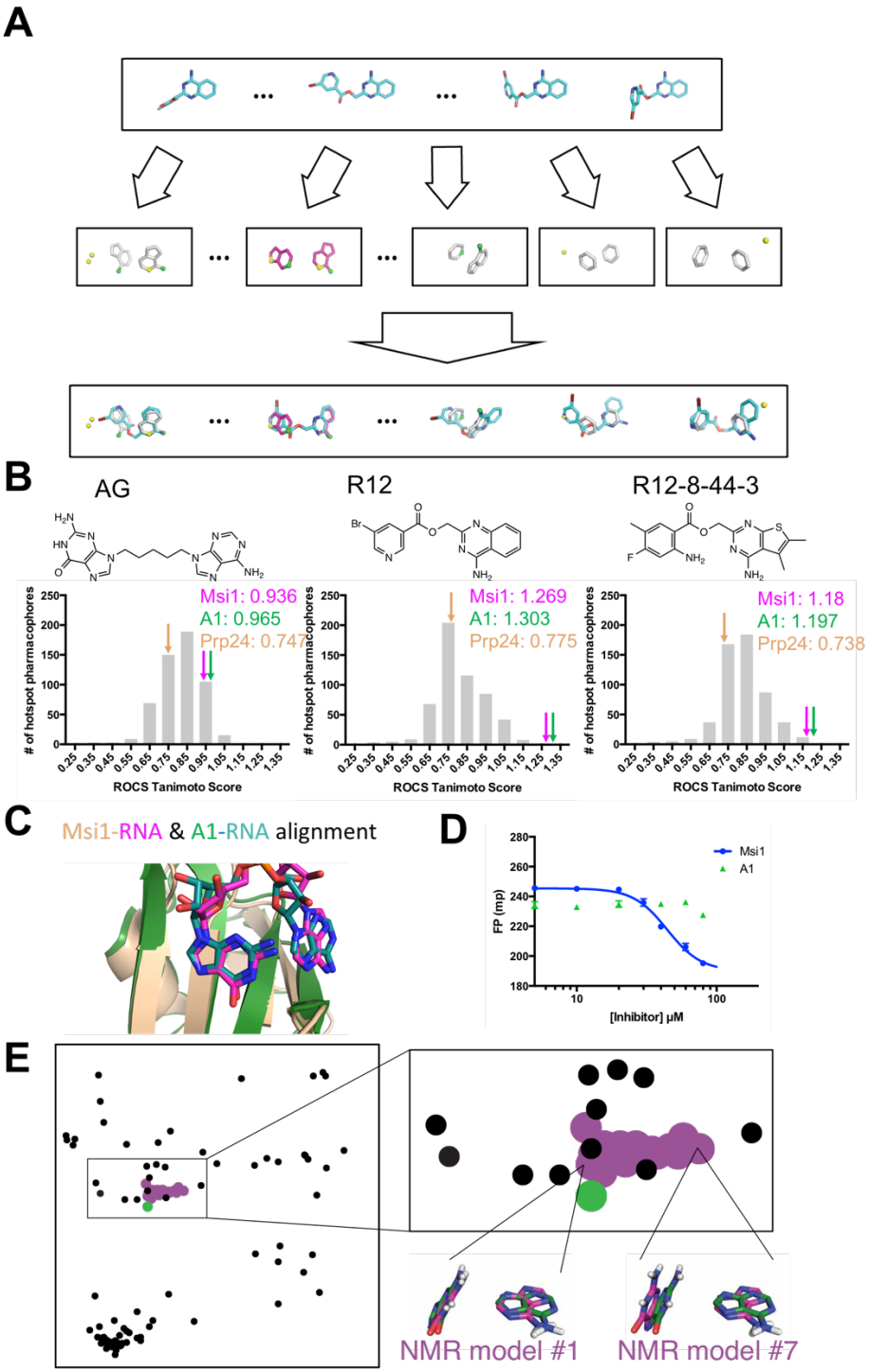


Figure 4: Predicting candidate off-target interactions of a given inhibitor. (A) We screened each conformer of a given ligand against the complete set of unique hotspot pharmacophores from other protein-RNA complexes in the PDB. Hits in this screen correspond to other proteins that recognize their cognate RNAs using interaction geometry that can be mimicked by the compound of interest. **(B)** Application of this approach to a hypothetical compound built by connecting adenine and guanine with a flexible linker, to R12, and to R12-8-44-3. High scores correspond to other proteins that recognize

their cognate RNAs using interaction geometry that can be mimicked by the compound of interest. The distribution of scores for the complete pharmacophore library is shown, with the score of the Msi1 (*pink arrow*) and hnRNP A1 (*green arrow*) pharmacophores indicated. The artificial compound matches the pharmacophores from many proteins equally well, whereas R12 and R12-8-44-3 match the Msi1 and hnRNP A1 pharmacophores much better than anything else in the library. **(C)** Comparison of the structures of the complexes reveal the basis for identification of hnRNP A1 as a candidate off-target interaction: this protein recognizes its cognate RNA (*green*) with similar positioning of functional groups as Msi1 (*magenta/wheat*), even though the adenine on the right is flipped. **(D)** Evaluation of R12-8-44-3 in an FP competition assay shows that this compound does not inhibit hnRNP A1. **(E)** A projection of the hotspot pharmacophores from all RRM/RNA complexes currently available in the PDB. Each point corresponds to an individual hotspot pharmacophore; this map was generated by using multidimensional scaling analysis to generate the 2D projection that best preserves relative distances between points. With the exception of Msi1, only a single conformation was used for complexes that were solved by NMR. Hotspot pharmacophores from individual members of the Msi1 NMR ensemble are indicated (*magenta*), along with the one from hnRNP A1 (*green*). The compounds described here were identified by computational screening using the hotspot pharmacophore from model #1 of the Msi1 NMR ensemble; this hotspot pharmacophore is very similar to extracted from the hnRNP A1 complex, and accordingly the A-G dinucleotide pair is nearly superposable. In contrast, hotspot pharmacophore from other members of the Msi1 NMR ensemble (such as model #7) are more distant from that of hnRNP A1, and indeed the A-G dinucleotide pair is recognized in a different conformation in these models. The use of highly-distinct models as templates for computational screening may lead to identification of compounds incapable of mimicking the pattern of interactions recognized by other RRM domains, and thus very likely to be selective inhibitors.

To test this, we expressed and purified RRM1 domain of hnRNP A1. We confirmed that hnRNP A1 would indeed bind the same fluorescently-labeled RNA used in our previous experiments, and then probed the effect of adding R12-8-44-3 (**Figure 4d**). In this competition experiment, we find that R12-8-44-3 does *not* inhibit RNA binding of hnRNP A1. We propose that matching to a given hotspot pharmacophore may provide some modest degree of potential binding energy, but that further fine details of the interactions must also be complementary in order to achieve potent binding. Thus, optimization of R12 to R12-8-44-3 enhanced potency for Msi1 by design, and potency for Msi2 because the two proteins are so similar, but would not have impacted the very weak starting affinity for hnRNP A1.

While further experimental evidence will be necessary to explicitly determine whether the compounds reported here engage in unanticipated interactions with any other RBPs, this computational approach provides a potential strategy to identify likely off-target interaction partners. While we cannot explicitly confirm that R12-8-44-3 does not inhibit any other RBPs in the human proteome, we can at least provide rationale for why selectivity should be expected from this compound. Looking ahead, this general

strategy may also enable predicting at an earlier stage the potential selectivity of a given compound, which in turn may help prioritize specific scaffolds and drive further focused optimization.

Taking this analysis one step further, we extracted hotspot pharmacophores from each of the 95 RRM/RNA complexes currently present in the PDB, and evaluated their similarity in an all-versus-all manner. From these pairwise distances, we then used multidimensional scaling analysis (MDS) to construct the two-dimensional projection that best reflects the pairwise distance between every pair of hotspot pharmacophores: this projection represents a visual "map" of all the hotspot pharmacophores in the PDB (**Figure 4e**).

Unsurprisingly, all of the hotspot pharmacophores built from members of the Msi1 NMR ensemble cluster into a punctate group, reflecting their shared geometric features. The pharmacophore extracted from the hnRNP A1 crystal structure also overlaps the Msi1 cluster. A single hotspot pharmacophore was used in our initial screen that led to identification R12; this hotspot pharmacophore was extracted from model #1 of the Msi1 NMR ensemble, and looking retrospectively it is evident that this is one of the pharmacophores closest to that of hnRNP A1. Because of variation between the models that comprise this NMR ensemble, different models lead to slightly different hotspot pharmacophores (as seen on this projection). Indeed, included among the set of Msi1 hotspot pharmacophores are examples (such as model #4) that are quite distinct from that of hnRNP A1. We expect that screening against a template that is more dissimilar to the hnRNP A1 hotspot pharmacophore, and all other hotspot pharmacophores, will lead to compounds with even more assurance of selectivity from the outset. Further, the "isolated" points on this map that are most distant from any other points represent the most distinctive and unique hotspot pharmacophores in the PDB: these protein targets may be particularly amenable to design of highly selective inhibitors.

2.5 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 [155,156]; here, we instead encode a key RNA epitope on a small-molecule scaffold. We demonstrate the application of our approach using Musashi-1 and Musashi-2, 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 our first application of this RNA mimicry approach, we elected to restrict our initial screening to commercially available compounds. Though none of the resulting hit compounds provided complete recapitulation of the desired hotspot interactions, we found that one of these, R12, complemented the protein surface without steric clashes and provided a starting point for new inhibitors of Msi1, thus validating the computational approach. In parallel with this work, a contemporaneous study also sought to design inhibitors of RBPs via mimicry of the cognate RNA [157]. This study reports inhibitors of a different RRM-domain protein, HuR, by manually selecting key moieties from the RNA to mimic, then using computational approaches to design compounds accessible through multicomponent reaction chemistry. While STD-NMR confirmed the interaction of some of these compounds with the protein, the binding affinities for these compounds were not reported.

To optimize R12, we then drew from an enormous new library of make-on-demand compounds to carry out a modern form of SAR-by-catalog. The availability of this resource allowed us to rapidly drive forward optimization, and ultimately led us to an inhibitor with single-digit micromolar binding affinity, R12-8-44-3. This library also enabled us to probe the effect of replacing the central ester linker in our initial series, leading to analog R12-8-44-lk2. In future, we expect that ongoing increases in the size of this (and

competitors') "virtual catalogs" will greatly facilitate medicinal chemistry optimization of potency and selectivity for many other projects as well.

With regards to selectivity, we propose that aligning candidate compounds to hotspot pharmacophores extracted from other RBPs can help identify potential off-target interactions. This can allow prioritization of select off-target RBPs for explicit biochemical testing, rather than simply collecting arbitrary off-target RBPs for evaluation. In this vein, we were especially pleased to note that R12-8-44-3 showed no inhibition for the RBP predicted as its most-likely off-target interaction, hnRNP A1.

We do note, however, that predictions of potential off-target interactions in this manner are necessarily limited: both by the incompleteness of the set of RBP complexes in the PDB, and by the fact that complexes solved using x-ray crystallography are present as single points on this map, instead of clusters that reflect conformational flexibility. Even with this limitation, however, already this utility of this approach to identify potential off-target interactions is clear.

Finally, we do acknowledge that our optimization of the R12 series did not lead to extremely potent compounds; we suspect that this may be an intrinsic limitation of the relatively flat binding site available on the protein surface. That said, PROTACs (PROteolysis TArgeting Chimeras) [158-162] have emerged as a viable strategy for addressing challenging targets, and may be exquisitely well suited for advancing these compounds. In considering development of RNA-mimicking inhibitors as warheads for development of new PROTACs, we note specifically that the binding affinity for the target has proven not to be a major determinant of effective target degradation. Thus, even if achieving highly potent direct inhibitors of RRM domains remains challenging, selective inhibitors may nonetheless offer a path forward for unlocking the tantalizing biology of RBPs, both as novel chemical probes and also as potential starting points for new therapeutics.

2.6 Methods

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

2.7 Acknowledgements

This work used the Extreme Science and Engineering Discovery Environment (XSEDE) allocation MCB130049, which is supported by National Science Foundation grant number ACI-1548562. This work was supported by a grant from the NIGMS (R01GM123336) and grants from the NCI (R01CA178831 and R01CA218802). This research was also funded in part through the NIH/NCI Cancer Center Support Grant P30 CA006927.

CHAPTER 3: Designing inhibitors of RNA-binding protein Musashi-1 via *in silico* combinatorial chemistry

Yusuf Adeshina[1,2], Nan Bai[1,3], and John Karanicolas[1*]

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

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

University of Kansas, Lawrence, KS 66045

*To whom correspondence should be addressed.

E-mail: john.karanicolas@fccc.edu

3.1 Abstract

RNA-binding proteins (RBPs) are increasing being implicated as key regulators of disease-associated pathways. Their importance has prompted investigations as potential target for therapeutic intervention, but efforts to develop inhibitors of non-enzymatic RBPs have often stalled because high-throughput screens against these targets typically yield low hit rates. We hypothesized that a potential reason for the relative difficulty in targeting RBPs could be the types of chemical matter used in these screens; in particular, many screening libraries are enriched in compounds more suitable as GPCR ligands and kinase inhibitors. To test this hypothesis, we sought to build an *in silico* chemical library of compounds specifically designed to complement the binding site of cancer-driving RBP Musashi-1 (Msi1). Starting from a pharmacophore describing the interactions of RNA with Msi1, we assembled building blocks that matched portions of this pharmacophore and could be linked together using straightforward chemistry. We then enumerated the products of reactions involving these building blocks, leading to a virtual library of 900,000 Msi1-focused compounds. We then screened this library to identify specific compounds expected to adopt three-dimensional conformations matching our target pharmacophore, and refined these in complex with Msi1. Finally, we synthesized the top-scoring 5 compounds, and found that the most potent of these inhibited Msi1 with a Ki value of 10 μM. Looking ahead, we envision that the same strategy may prove useful in designing compounds to precisely fit pharmacophores extracted from other protein classes as well.

3.2 Introduction

The number of new drug application (NDA) approvals before and after sequencing the human genome hasn't changed drastically. This is in sharp contrast with all expectations that the vast amount of data garnered from the project and the potential information it holds was going to lead to an explosion in the number of novel therapeutics. In fact, a report recently showed that more than 75% of protein related research still focuses on only 10% of proteins that were identified before the human genome project [163]. This bias is even more pronounced in small molecule drug discovery, with G-protein coupled receptors (GCPRs) and enzymes being over-represented in the list of FDA approved drugs [164,165]. The problem here is not that newer targets are not being identified and validated, it's the fact that the existing tools are not well suited for probing these new targets and this has led to a lot of these novel target being tagged "undruggable". Protein-Protein Interfaces (PPIs) were tagged "undruggable" for a long time because they were not amenable to the traditional drug discovery paradigms, that have been well validated for enzymes and GCPRs. The story changed only when radically newer approaches were employed [166,117,120,167]. In the same vein, if the "undruggable" targets of today are ever going to be drugged, it will be via a radically different approach and tools from what we have today.

Apart from the fact that the tools do not match the challenge, the chemical space of available chemical probes to investigate these targets are limited and the coverage does not extend to novel targets. In fact, screening for inhibitors are more likely to succeed for traditional target [168-170]; this is in contrast to what is observed in the case of novel target classes (hit rates usually low) [9-11,169,12] because the cognate small molecule inhibitor chemotypes are under-represented in the current libraries [171-173]. A natural solution to this challenge would be to generate focused virtual chemical libraries replete with chemotypes with features that will make them a binding partner of the target in question. In fact, it has been shown that this approach can increase the hit rate to >20% [168,174,175].

RBPs play a significant role in a number of post-transcriptional cellular processes and being able to modulate these processes will be crucial in understanding the biology a lot of diseases including

cancers. For example, more than 90% of human genes undergo pre-mRNA alternative splicing, accounting for most of the proteomic diversity in human cells [176], and this is just splicing; there are other important post-transcriptional regulation like polyadenylation, mRNA stability, mRNA stabilization, etc. In fact, increasing number of human diseases—whose etiology wasn't clear—are now being linked to dysfunctional RNA-binding proteins (RBPs) [177]. Therefore, targeting these RBPs (specifically RNA-recognition motifs, RRMs) might be a good strategy to finding potent treatment to several diseases, including cancers. Unfortunately, this venture has not been very fruitful because the current tools and methodologies have failed to yield any meaningful result, hence they are generally considered undruggable. RBPs have proven undruggable for many reasons, among these their unique structural characteristics: that is, the interfaces are relatively large, more polar than classical drug targets and lack well defined pockets. Therefore, the conventional approaches like docking used in finding small molecules are not well suited for these non-traditional targets [116]. Apart from the peculiar characteristics of these targets, another challenge is the relatively small coverage of the chemical space of libraries used in conventional screening approaches (virtual screening and high-throughput screening).

Currently, there are only few classes of compounds that target protein-RNA interactions. The majority of them include a number of compounds that have been reported to disrupt binding by interacting with the RNA rather than with the RBP [104,105]. A few inhibitors of certain RBPs have also been identified via high throughput screening [101,102,178] , also one particular series from virtual screening that competes with double-stranded RNA for binding to toll-like receptor 3 [103]. Though rational design of inhibitors of RRMs are scarce, there are plenty examples for non-RRM RBPs. These reported examples can be categorized into two general classes. The first class is nucleoside analogues [106-110,179,180] such as anti-HIV-1 NRTIs, that mimic the chemical structures of natural-occurring nucleosides and takes advantage of enzymatic processing by their targets to form covalent adducts. Though nucleoside analogues can be straightforward to design, these molecules are unable to provide sufficient binding affinity or selectivity without covalent linkage; this has prevented this strategy from being extended to non-enzymatic RBPs (including RRMs). Allosteric inhibitors [110,113,114] such as

anti-HIV-1 NNRTIs are the second class. They bind to secondary sites on the protein target and shift its conformation to an inactive state. In principle, allosteric inhibitors could inhibit both enzymatic and non-enzymatic RBPs; in practice, however, this is very challenging because of the difficulty in identifying allosteric sites and then finding small molecules to complement these sites, which has limited the general utility of this approach to all but a few cases. Volpe et.al [181] is the only reported example of rationally designed inhibitors of RRMs (HuR). While this is a useful starting point, this study did not report any binding affinities for any of the resulting compounds.

Here, we report a new pipeline for rationally designing non-enzymatic, non-allosteric small-molecule inhibitors of Msi1 by building a library specifically focused on the structure of our target protein. In this approach, we employ a modified version of a library enumeration strategy called "deconstruction-reconstruction approach" [182]. The original implementation of this approach involves splitting small molecule ligands into fragments that can be used as templates for fragment search (deconstruction), followed by chemical enumeration of library using these fragments as starting material. Our approach, on the other hand, involves splitting the pharmacophores (instead of small molecules) into smaller units, using these units as templates to find matching chemical fragments, then joining together these fragments chemically to generate a focused library of synthetically accessible compounds (MsiLib). This library is then used in virtual screening for small molecule inhibitors of Msi1. This is expected to yield a generalizable pipeline that utilizes a focused-library—built from scratch—that accounts for the peculiarity of RRMs in the search for small molecule inhibitors of Msi1. One of the outcomes expected is the rational design of novel class of Msi1 inhibitors, which could serve as a starting point for the development of novel chemical probes for a variety of diseases. Also, we present here a head-to-head comparison of our Msi1-focused library against a pair of target-agnostic libraries comprising commercially available compounds (ZINC and ENAMINE).

3.3 Computational Approach

3.3.1 Overarching strategy

The difficulty in targeting protein-RNA interactions, amongst other factors, has been attributed to the fact that the interfacial features are different from most traditional drug targets [116]. Therefore, a different approach is required to effectively understand the nature of interactions at the interface, in order to find a good inhibitor of this interaction. In separate work [183], we designed a computational method for extracting these interfacial features (the pharmacophore). The method is based on the idea that has been used in a related class of "non-traditional" drug targets, protein-protein interfaces. In a protein-protein complex, each of the individual interfacial residues typically contribute differently to the energetics of binding; the majority of the binding affinity derives from a small number of "hotspot" residues [117-119]. This observation, in turn, motivated several groups to mimic these key interactions when designing small-molecule inhibitors [120-123]. In our method, we take this "hotspot" paradigm and extend it to protein-RNA interactions. The central theme here is that even though interfaces between RBPs and their cognate RNAs are relatively flat, some of the interfacial nucleobases are buried much more deeply than the others; as shown in **Figure 1A,** the nucleobases that are the most buried are purple sticks with blue heteroatom representation, while the remaining bases are shown as pink sticks. Our algorithm first extracts 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)**. All polar atoms on the nucleobases that do not engage the protein hydrogen bonding interactions are then replaced with carbon atoms, since those polar groups are not required for inhibitor design. The resulting output from this is a broad spatial interaction map of the protein-RNA interaction, which typically cannot be spanned by a single drug-like small molecule; this is followed by clustering neighboring chemical groups and advancing each cluster individually. Using this approach, we distill the structure of the Msi1-RNA complex to a minimal "hotspot pharmacophore" that possesses the key interactions to be recapitulated by a small molecule where the hydrogen bond donor and acceptor are

represented as yellow and green spheres respectively (**Figure 1C**). Then, we used the resulting pharmacophores as templates for 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. As a proof of concept, we applied our method to Musashi-1 (Msi1), an RBP that is a post-transcriptional repressor of anticancer mRNAs including mNUMB, APC, and more. Since there is no guarantee that commercially available libraries will contain compounds that would mimic these pharmacophores, it is an attractive venture to explore focused libraries that is built from scratch. Therefore, we sought to build the first Msi1 targeted virtual library and we called it MsiLib. The idea here is translating the identified pharmacophore into chemical fragments with relevant linking functional group that could be used to join the fragments together (**Figure 1D)**. This library of fragments serves as the building block for MsiLib. As shown in **Figure 1E**, we generated low-energy 3D conformations of all the compounds in MsiLib. We then aligned each of the conformers with the pharmacophores. We placed the top scoring conformers in the binding site of the protein, minimized the energy and scored the complex. We sourced the fragments needed to build the top scoring compounds from the energy-minimized complexes, and synthesized the designed compounds (**Figure 1F**). The synthesized compounds were characterized in biochemical assays (**Figure 1G**).

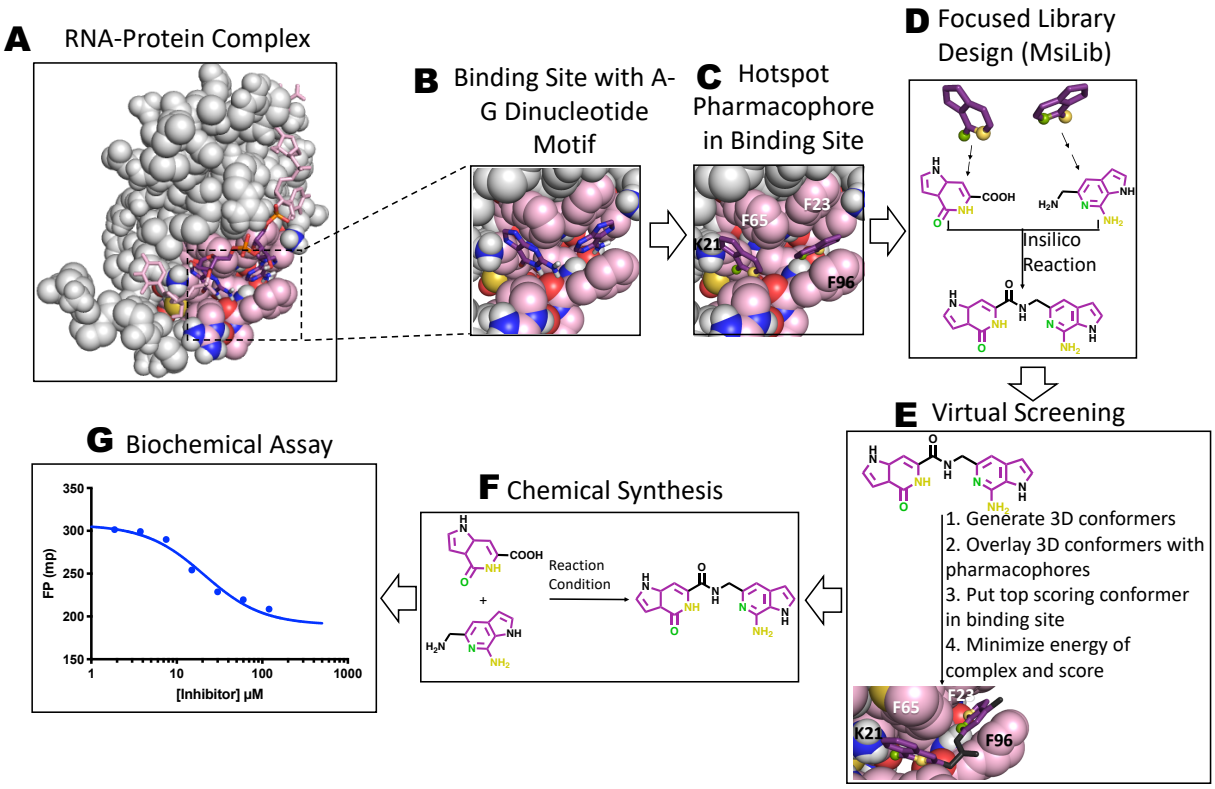


Figure 1: Workflow of our pipeline for designing inhibitors of RRM proteins. (A) Msi1 bound to mNUMB RNA with the dinucleotide motif binding site highlighted in color. Purple, red and blue sticks represent the RNA dinucleotide motif and pick, red and blue spheres represent the atoms of the protein the in the binding site. The region outside of the hotspot is shown in gray sphere and pink sticks. We used all the 20 NMR models in PDB ID 2rs2. **(B)** Binding site zoomed in. All the heteroatoms of the nucleobase are shown in blue on the rings. Not all these heteroatoms participate in interaction and our algorithm distills this into the most important interactions. **(C)** The key atoms on the ring participating are shown in small spheres (yellow for hydrogen bond donors and green for hydrogen bond acceptors). On the left ring, the ring makes a π-stacking interaction with F65; the hydrogen bond acceptor makes hydrogen bond with the side chain amine of K21, while the donor atom on the ring interacts with the backbone carbonyl of F96. For the ring on the right side, it makes a sandwiched π-stacking interaction with F23 and F96. The hydrogen bond acceptor atom makes hydrogen bond with the backbone amino group of F96 while the donor interacts with the backbone carbonyl of V94. **(D)** We used the generated pharmacophore to search for fragment with useful chemical handles (carboxylic acid and amine) and used these fragments *in silico* chemical reaction to generate a Msi1-focused library of chemical compounds. **(E)** Virtual screening process. We started by generating 3D conformers of each compound, followed by aligning each of the conformer to the pharmacophores; finally, the top scoring aligned conformers are placed in the binding, minimized and scored. **(F)** Chemical synthesis of the selected compound compounds from virtual screening. Here, we used standard amide coupling reagent to carry out the reaction. **(G)** We characterized these compounds using a fluorescence polarization competition assay to determine whether they disrupt the interaction of Msi1 with its cognate RNA.

3.3.2 Building a Msi1-focused library

The choice of chemistry is a fundamental determinant of the chemotypes in a library. First, we had to select chemical reaction for the library enumeration (**Figure 2A**). It's common practice to use a number of very different reactions to introduce diversity into the proposed library. However, for this proof-of-concept study, we elected to limit the available chemistries to a single reaction for simplicity and ease of synthesis. For this reason, we chose the most-used reaction in drug discovery campaigns [184]: amide coupling. Since the overall goal is to generate a library of compounds very similar to the ensemble of pharmacophores, we used them to inspire the kind of fragments we used in enumerating the compounds in MsiLib. After choosing the reaction, we deconstructed each of the pharmacophores in the ensemble into two rings (along with their respective heteroatoms) (**Figure 2B**). Using each of the rings from templates generated above, we exhaustively enumerated several patterns of fragments having the same spatial arrangement of the hydrogen bond donors and acceptors. We fed these patterns into PubChem, applying the following filter: commercial availability, number of heavy atoms <15, presence of the desired functional groups (-NH2 and -COOH). This returned hits fragments with different R groups but fulfilling the criteria above. Our search yielded 1498 amine fragments and 603 acid fragments (**Figure 2C**) that was then fed into ChemAxon Reactor [185], an in-silico chemical reaction software which is capable of performing up to half a million reaction in 30 minutes. The software is capable of carrying out combinatorial synthesis, as well as sequential synthesis. Overall, we generated a focused library of about ~900k compounds (MsiLib) with features similar to our pharmacophore (**Figure 2D**).

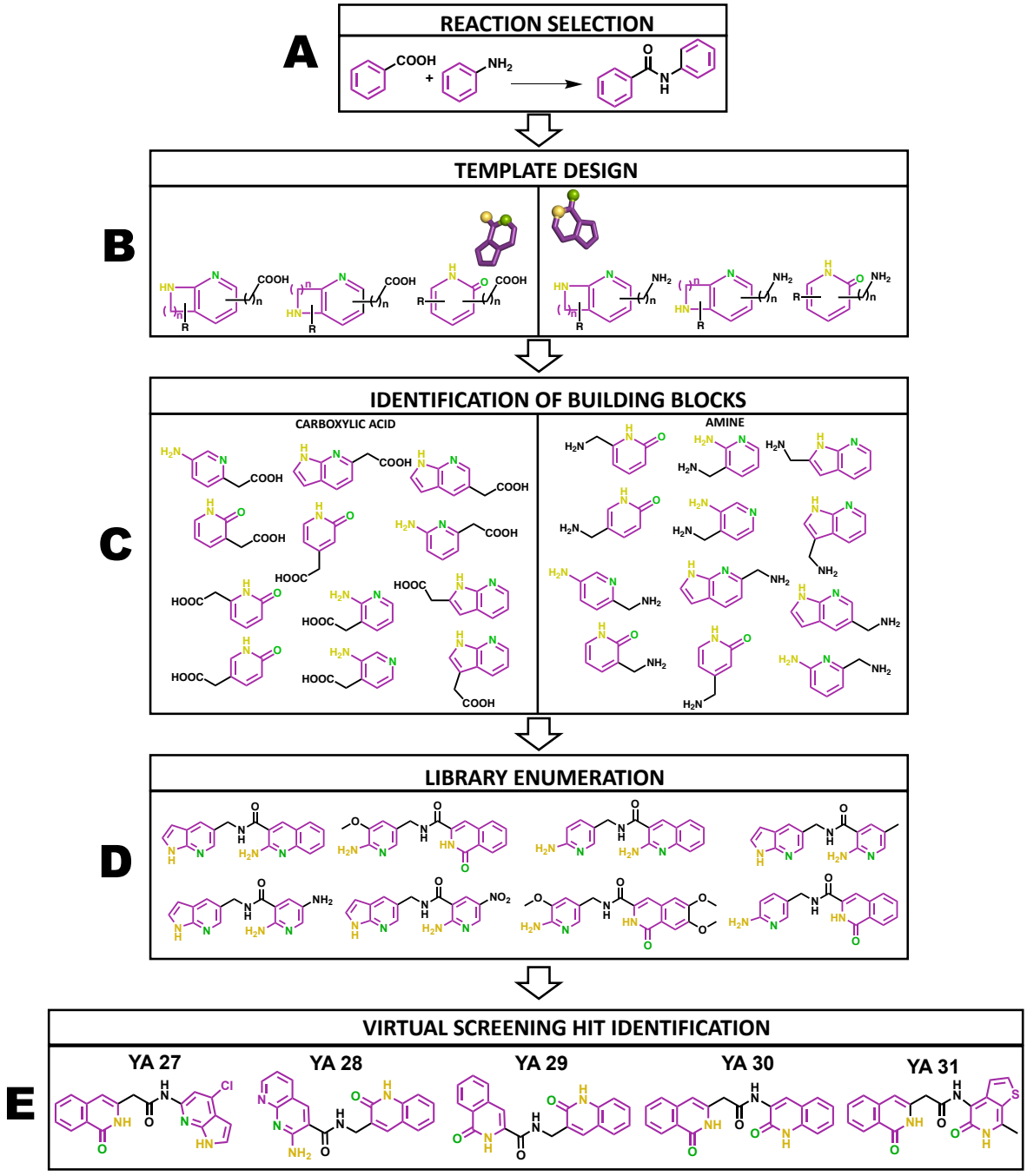


Figure 2: Strategy for building up a target-focused chemical library. (A) Reaction selection. We selected amide coupling based on its simplicity and widespread application in drug discovery. **(B)** Fragment template design. We deconstructed each of the pharmacophore in the ensemble of pharmacophores into two ring structures and we used each of these deconstructed rings to design templates that captures all the different patterns and also containing either carboxylic acid or amine functional group for joining the resulting fragments later. These were used as queries for PubChem search. **(C)** Identification of building blocks: we searched PubChem for hits to our desired templates, with the following filters: commercial availability, number of heavy atoms <15, presence of the desired functional groups (-NH2 and -COOH). **(D)** Enumeration of a library of Msi1-focused ligands. We used ChemAxon Reactor to enumerate library of compounds from the building blocks identified. With this,

we enumerated a library of ~900k compounds and used this for virtual screening. **(E)** Chemical structures of hit compounds from virtual screening. All chemical structures are color-matched to the pharmacophore, with the purple ring structures matching the ring structures of the pharmacophore. The yellow atoms are hydrogen bond donor while green atoms are hydrogen bond acceptors. The black part represents chemical structures not part of the pharmacophore, for example linkers and extra R-groups.

The novelty in this approach is that we are searching for hit compounds in a library—which we created—that we know, in theory, should have a higher density of complementary ligands. In this experiment, we extracted dinucleotide pharmacophores from the 20 NMR models (from PDB ID: 2rs2). Only 12 (models 2, 3, 9-13,15-19) out of the 20 models yielded output, as these were the ones that met the criteria for pharmacophore generation. As expected, because each of the models are slightly different structurally, the resulting pharmacophores are also different (**Figure S1)**.

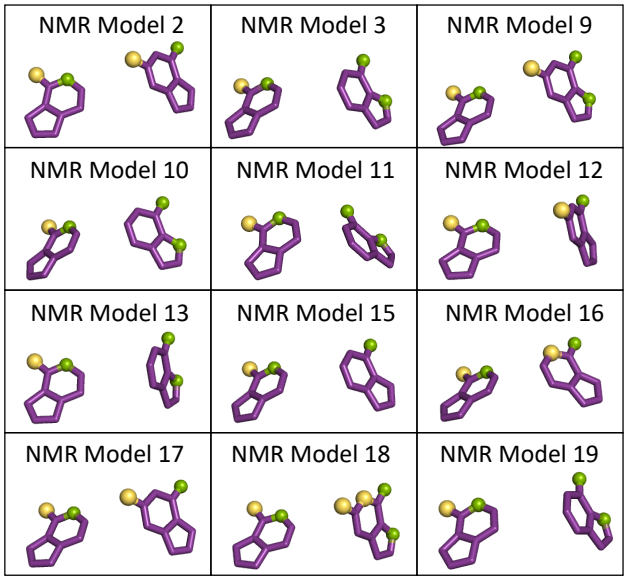


Figure S1: Pharmacophores extracted from the different models in the NMR ensemble. Our algorithm extracted 12 structurally and geometrically different pharmacophores from the NMR ensemble of Msi1. These different patterns inspired the design of the chemical building blocks used in building MsiLib.

3.3.3 Screening MsiLib for Musashi-1 inhibitors

We used OpenEye OMEGA [82] to generate low energy conformers (3D structures) for each of the compounds in the library. OMEGA is a conformer generator that works by fragmenting the 2D

structure along the sigma bonds, samples different torsional angles to generate low energy fragment, this low energy fragments are then reassembled [82]. We generated 300 conformers for each molecule the library. Following this, we overlaid each of the 3D conformers with the pharmacophores generated using the OpenEye software Rapid Overlay of Chemical Structure (ROCS) [83]. ROCS is a ligand-based virtual screening tool that matches a query molecule to a library based on shape and electrostatic features. ROCS maximizes the overlap of molecules by using atom-centered Gaussians and rotates and translates the molecules to calculate the optimal alignment [83,82]. The highest scoring matches are those that most closely match the pharmacophore features. The similarity metric used is combined volume and electrostatic Tversky score that penalizes for extra atoms on ligands larger that the query molecule (RefTverskyCombo). For this, we selected the top 1000 compounds according to RefTverskyCombo score and generated models corresponding to the RBP bound small molecules. Rosetta energy function was used minimization to resolve any clashes and the ligand interaction energy will be calculated. We selected top scoring compounds according to Rosetta interface energy, number of hydrogen bonds and number of buried unsatisfied group. From these, we selected 5 diverse compounds (**Figure 2E**) for synthesis and characterization.

3.4 Results

3.4.1 Comparison of MsiLib, ZINC and Enamine

For this, we generated conformers and carried out shape and electrostatic feature alignment for compounds in ZINC and Enamine database like we did for MsiLib. From here, we used the RefTverskyCombo scores as input for the two sets of experiment we carried out. One of the experiments involves using the scores as they are, without accounting for the difference in the library size between the three libraries compared (~900k for MsiLib, ~9 million for ZINC and ~32 million for Enamine). For this, we selected the top 1000 compounds according to RefTverskyCombo score and generated model corresponding to the RBP bound to the small molecule. The Rosetta energy function was used to

minimize the complex, and then the ligand interaction energy was calculated. In the second experiment, we normalized for library size by randomly selecting 1000 compounds in top 1% from the respective libraries and advanced these to Rosetta energy minimization. For all this, we evaluated the libraries by looking at the distribution of scores and rank of scores.

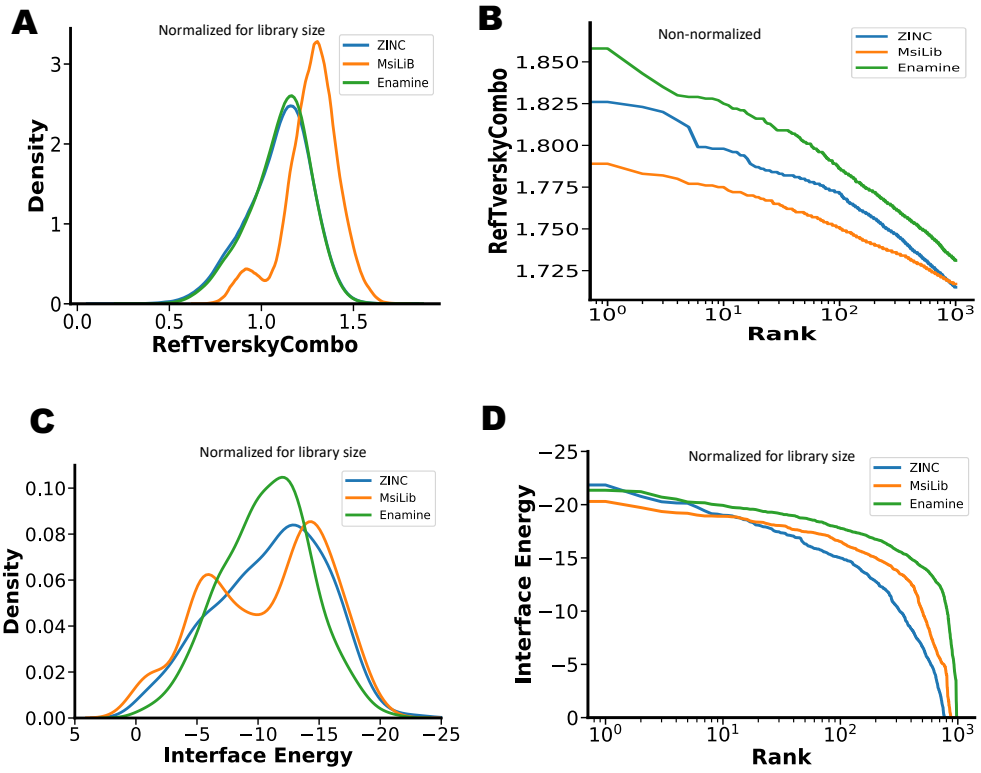


Figure 3: Quantitative comparison of the MsiLib, ZINC, and Enamine libraries. (A) Comparison of the degree to which compounds in each library match the desired pharmacophore (as evaluated via ROCS RefTverskyCombo). The distributions are normalized to the size of each library. The curve corresponding to MsiLib is significantly right shifted compared to the other two distribution (ZINC and Enamine). This means that MsiLib is more enriched in complementary ligand compared to ZINC and Enamine. Unsurprisingly, the is an almost overlapping distribution between ZINC and Enamine, showing that general purpose library are not biased towards a particular target. **(B)** We investigated the tail of the distribution in (A) (without the normalization) by ranking the scores and plotting the top 1000 scores against their ranks. Here the more north-east shifted a plot is, the more highly ranked the compounds in the top 1000 are. Even though ZINC and Enamine are not highly enriched in Msi1 complementary ligands (as shown in (Figure 3A)), they contain a few highly ranked compounds. This is not surprising since the larger a library is, the larger the chance of finding a ligand for a particular target in it. **(C)** Density plot of Rosetta interface scores for top 1000 highest ranking compounds according to ROCS RefTverskyCombo scores after library size normalization. Again, the Rosetta interface score is slightly right shifted. **(D)** Similarly, when we look at the tail of (B), ZINC and Enamine have few compounds in the tail of the distribution that ranks better.

In order to assess how focused our library really is and if it is more enriched with Msi1 binders than diverse, general-purpose commercially available libraries, we overlaid the conformers of each compound in our library to each of the pharmacophore in the pharmacophore ensemble and computed the 3D shape and electrostatic similarity. For this, we used RefTverskyCombo as metric, since compounds in our libraries are generally bigger than the pharmacophores. We also evaluated the Rosetta energies of the minimized complexes of the most similar 1000 compounds. We did the same for ZINC database and Enamine database. The results are shown in **Figure 3**. **Figure 3A** shows the density plot of the RefTverskyCombo score for MsiLib (yellow), ZINC (blue) and Enamine (Green). From the figure, it's clear that, overall, MsiLib is more enriched in compounds that matched the pharmacophores better (evident from being the rightmost shifted) and more focused (the tightest distribution). Here, the mean ± standard deviation of these distributions are 1.09 ± 0.18, 1.27 ± 0.15 and 1.10 ± 0.17 respectively ZINC, MsiLib and Enamine respectively. What this means is that the bulk of MsiLib is replete with more Msi1 complementary molecules compared to ZINC and Enamine. Since analysis like this doesn't provide much insight into what is happening in the tail of the distribution, we further investigated the right tail of the distribution (**Figure 3B**) by sorting the RefTverskyCombo score and then plotting ROCS score (RefTverskyCombo) against the logarithm (base 10) of ranks. In this figure, the top 1000 of Enamine consistently ranked better than ZINC which in turns ranked better than MsiLib. This analysis showed that ZINC and Enamine contains a small number of compounds in the extended right tail of the distribution that ranked better than compounds in the right tail of MsiLib. This is, however, unsurprising because ZINC is 9 times bigger than MsiLib, and Enamine is 32 times bigger. The bigger a library is, the higher the chance of finding a small number of complementary ligands. To correct for size difference, of the three libraries, we selected the top 1% according to RefTverskyCombo score and randomly select 1000 compounds. Often times, high ROCS scores do not translate into high interaction energy with the target protein. So, we computed the interaction energy for complexes of compounds with top 1000 ROCS score and also for the size-corrected top 1000 using Rosetta [124] and plotted these energies against the rank of

the compounds (**Figure 3C)**. Here, we observe a similar distribution to **Figure 3A**, where the MsiLib shows better enrichment in complementary ligands when the libraries are corrected for size. **Figure 3D** is a supporting figure for **Figure 3C**, and supports our claim that though ZINC and Enamine are not highly enriched in Msi1 complementary ligands, they contain a few high-ranking ligands because of the size of the library. In summary, we concluded that MsiLib is highly enriched in compounds with features that would be complementary with Msi1 binding site. Even though, ZINC and Enamine have a few compounds that ranks better than MsiLib, their relatively large size can become a challenge in a setting where computational resources are limited. With this in mind, we moved forward with virtual screening of MsiLib and experimental validation of the hits from this library.

3.4.2 Testing hits from MsiLib as inhibitors of Musashi-1

We started our VS campaign by filtering out potential PAINS compounds using OpenEye's PAINS filter implementation (see details in *Methods*). We generated 300 low energy 3D conformations for each compound in the library using OpenEye OMEGA. We overlaid these with each of the pharmacophores generated from the NMR ensemble; Then, we selected the top 1000 compounds according to combined shape and chemical feature Tversky score and placed these in the binding site and minimized with Rosetta scoring function. We found that the 20 top-scoring hits selected can be classified 5 diverse scaffolds. One representative of each scaffold was synthesized and further characterized.

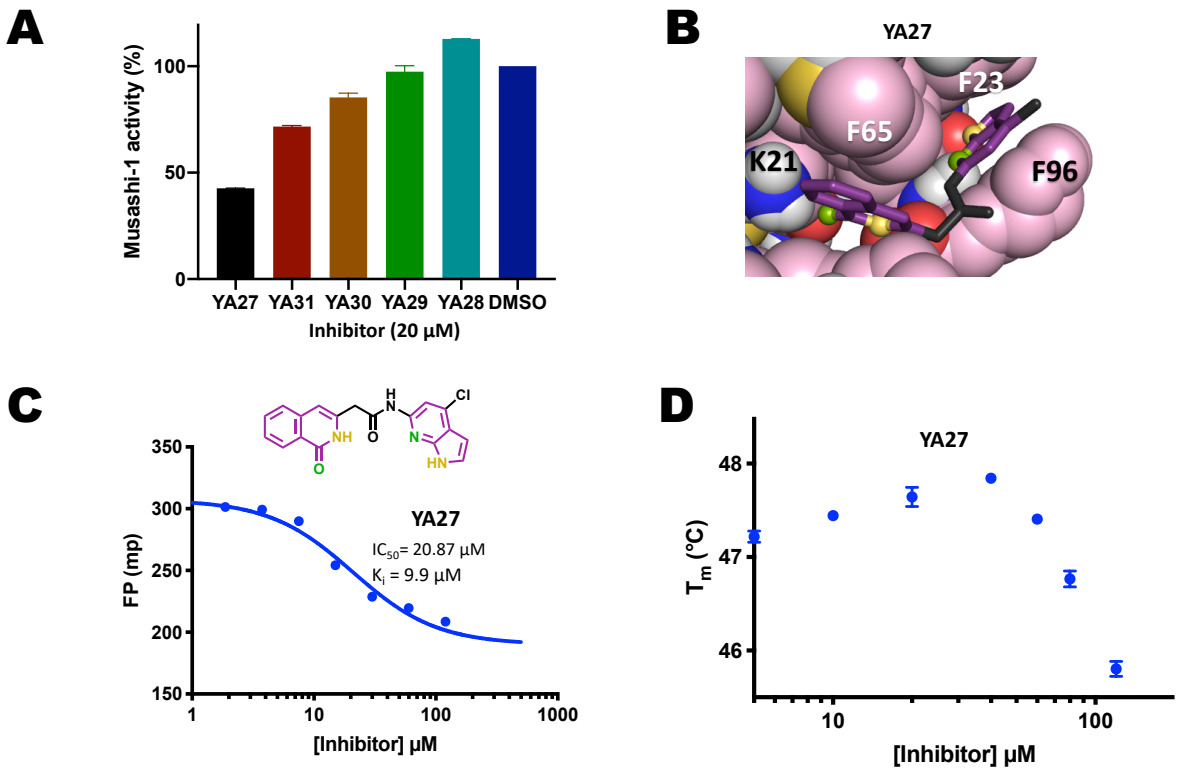


Figure 4: Biochemical characterization of designed inhibitors from the Msi-focused library. **(A)** Single point activity profile (at 20 µM) for Msi1 inhibition upon addition of YA compounds, as obtained from a fluorescence polarization competition assay. **(B)** Structural model of the most potent compound (YA27), showing how this compound matches all the desired pharmacophore interactions: in the model, YA27 makes π-stacking interactions with F23, F65, and F96 as well as four hydrogen bonds to K21, F96 and V94. **(C)** Dose-response of YA27 in the FP competition assay. **(D)** Differential scanning fluorimetry data of YA27 as a secondary orthogonal assay to confirm activity.

In order to validate our synthesized compounds experimentally, we first employed a fluorescence polarization competition assay (FP) described previously [183], to probe which of these compounds would inhibit Msi1's interaction with (labeled) cognate RNA. We found that 3 of 5 compounds showed a > 15% decrease in activity of Msi1, with YA27 being the most potent (**Figure 4A**). In our model of YA27 this compound engaged Msi1 by precisely recapitulating RNA's interactions, as specified by the hotspot pharmacophore (**Figure 4B**). To evaluate the potency of YA27 as a Msi1 inhibitor, we carried out a dose-response experiment (**Figure 4C**). This experiment confirmed increasing inhibition as a function of YA27, with a Ki value of 10 µM. Finally, to confirm a direct interaction of YA27 with Msi1 and rule out that apparent inhibition in the competition assay could occur through sequestration of the labeled RNA,

we used differential scanning fluorimetry (DSF). In this assay, the melting temperature (Tm) of protein is used to assess its thermal stability. Unbound protein unfolds at a certain temperature that is expected to increase in the event of binding to a small molecule partner. Here, we observed increased Tm as a function of YA27 at lower compound concentrations, confirming a direct and specific binding interaction, followed by a steep decrease in Tm above 40 μM that is likely attributable to non-specific interactions at these high concentrations (**Figure 4D**). Together, these two assays confirmed that YA27 is indeed an inhibitor of Musashi-1.

3.5 Discussion

As the resources required to carry out high throughput screening is beyond the reach of most academic laboratories, many labs are turning to virtual screening. While virtual screening has been widely successful in many cases, the limitation imposed by having to screen using general-purpose library limits the effectiveness of this approach, particularly in novel drug targets. To introduce some flexibility into virtual screening, we have introduced a modeling pipeline that couples target-specific library generation with structure-based virtual screening. With this methodology, we demonstrated that RBPs whire were previously labeled undruggable, can indeed be probed with a small molecule, even with the simplest possible focused library. To date, RBP drug discovery campaigns have relied mostly on high through-put screening [186,187,178,188-191]. While most of these yielded some results, it's been shown that virtual screening has better false positive rate [192,193,168]. While most academic labs would favor virtual screening over high-throughput screening, the dearth of suitable virtual screening methodology for RBPs has limited its application in this domain. A recent work reported the use of related tools (AnchorQuery and NucleoQuery tools [194]) to design candidate compounds inhibiting a different RBP, HuR [181]. In addition to the fact that this other study did not demonstrate the potency of the resulting compounds, certain other limitations are evident. The first of these is the need for human intervention in manually selecting the anchors used in designing the ligands, since it was not clear what criteria were used in

selecting the anchor bases. Since manual human design is required, the design is prone to a variety of human bias that could alter the course entire project. For example, the decision to replace one of the uracil's with phenyl ring is just one of the many paths that the work could have taken. Replacing the uracil with any other group capable of making π-π stacking interaction could have sent the work down a totally different path. In contrast, our work here relies on a general biophysical idea that the more buried a nucleobase (and its collection of warheads) is, the more it contributes to the energetics of binding. For the most buried dinucleotide pharmacophore, we considered every possible commercially available bio-isosteric fragment in building our screening library. Also, the pharmacophore template generation requires zero human input. Another consideration is how we incorporated the differential recognition of a nucleobase by RNA-recognition motifs (RRMs) into design of small molecule inhibitors. Since RNA-recognition motifs (RRMs) are highly conserved, design strategies that rely on using all the atoms on a nucleobase are bound to generate non-selective inhibitors that can be recognized by a lot of RRMs. Our approach took advantage of the geometrical difference in the placement of anchors in the binding site by only extracting atoms that had been geometrically positioned to contribute to binding and atoms not involved in interaction were not included in the pharmacophore. For example, two RBPs recognizing the dinucleotide motif AG did so in a geometrically different way and our methodology took advantage of these differences. With this idea, two different RRMs recognizing the same base but using different warheads would result in a different pharmacophore template. On the contrary, in methods that do not account for this differential recognition of nucleobases by RRMs, the same template would be used for the inhibitor design. Also, in the work of Volpe et.al, a general-purpose RBP library was screened and no binding affinity was reported. Furthermore, the only characterization method used in this study was STD-NMR (a method developed specifically for weak binder). The authors of this work suggested that building a focused library is very crucial in finding RBP inhibitors with stronger binding affinities. This makes our work the first study to report a rationally designed inhibitor of RBP, validated in orthogonal assays, with binding affinity data.

Our library uses commercially available building blocks to ensure ready access to the compounds in the library. However, one of the important considerations for the reagents used in the enumeration of the library is that access to compounds in our library depends heavily on commercial vendors being able to supply the compounds as promised. Sometimes, reagents that were previously deemed commercially available could be out-of-stock at the time of synthesis. This was the case with YA27, YA30 and YA31, where the actual carboxylic acid required for the reaction was 2-(6-fluoro-1-oxo-1,2-dihydroisoquinolin-3-yl)acetic acid, but the vendor only had the analog 2-(1-oxo-1,2-dihydroisoquinolin-3-yl)acetic acid without the fluorine substitution.

This work also provides perspective on how a simple focused library (MsiLib) enumerated with just one reaction could be more enriched in complementary compounds compared to large, diverse, general-purpose libraries (ZINC and Enamine). Even though, ZINC and Enamine have a few compounds that are better scoring compared to MsiLib, the bulk of the compounds in these libraries are compounds that are totally non-complementary to Msi1. Also, the sheer size of these library and the computational resources required can be daunting in a resource-limited setting.

Also the deconstruction-reconstruction paradigm has been reported in literature [182] and it was specifically applied to small molecule hit-to-lead optimization. Our work is the first that applied the idea to RPB pharmacophore-inspired hit identification. Even though we applied it here to RBPs, in theory, the idea is transferable to any templated interaction interfaces.

Recently, with advances in deep learning, several groups have been able to represent chemical space in continuous space and use this continuous representation to optimize search landscape. Deep learning has also been applied to creating a generative model that could spit out chemical molecules without the need for explicit library enumeration. For future work, we plan to apply these innovative technologies to RBPs.

3.6 Methods

3.6.1 Protein expression and purification

Msi1 RRM1-RRM2 was expressed and purified as described in our earlier studies [183].

3.6.2 Fluorescence polarization competition assays

For this experiment, we used RNA oligonucleotides (UAGGUAGUAGU/36-FAM/) that was ordered from Integrated DNA Technologies (Coralville, IA). Msi1 RRM1-RRM2 was expressed and purified as described in our previous work [183]. In this experiment, we measured the competition between the fluorescein-labeled RNA and the YA compounds for binding to Msi1. For each of the YA compounds, different concentrations of the compounds were titrated into wells containing a mixture of 8 nM Msi1 and 5 nM fluorescein-labeled RNA. This was done in two replicates. The resulting data from the experiment were fitted to a single-site competition model to determine IC50 using Prism 6 (GraphPad Software Inc.). Assay conditions are described fully in our previous work [183].

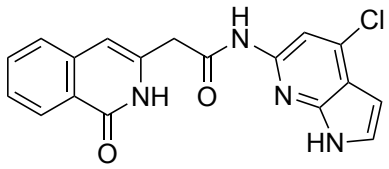
3.6.3 SYPRO DSF assay

We carried out this assay using an optical reaction plate (Thermo Fisher Scientific 4306737) sealed with optical PCR plate sheet (Thermo Fisher Scientific AB-1170). Each well in the plate was filled with 50 µL of reaction mixture containing 4 µM Msi1 and 5X SYPRO Orange (Invitrogen S6651), in triplicates. The assay buffer contains 20 mM HEPES, 150 mM NaCl, 0.01% Triton-X100, 2.5% DMSO pH 7.4. Different concentrations of the YA compounds were titrated into the final mixture, sealed with optical PCR plate sheet (Thermo Fisher Scientific AB-1170) and data was collected with Eppendorf Realplex2 Mastercycler. We used the JOE emission filter (550 nm) for measuring the fluorescent intensity, and “PTS clear plate” was set as the background for the calibration. The temperature range was set to 37 ºC to 56.6 ºC at 0.4 ºC/min. All data were analyzed with Prism 6 (GraphPad Software Inc.). Assay conditions are described fully in our previous work [183].

3.6.4 Synthesis of YA-series compounds: overview

All reactions were carried out in flame- or oven-dried glassware. Stirring was achieved with oven-dried magnetic stir bars. ^{1}H and ^{13}C NMR spectra were recorded on instruments operating at 600 MHz. High-resolution mass spectrometry (HRMS) spectra were obtained on an ESITOF mass spectrometer. The analytical method used for all compounds was the same except for YA27. This utilized a Waters Acquity T3 C18 column (2.1 × 50 mm, 1.7 μm) eluting with a linear gradient of 95% water (0.05% Formic Acid) to 99% CH_3CN at 0.6 mL/min flow rate where purity was determined using UV peak area at 214 nm. For YA27, a basic aqueous solvent was used instead (water with PH adjusted to 9.8 via Ammonium Hydroxide). We carried preparative LC/MS to purify the compounds and this utilized Waters Atlantis T3 C-18 column (19 x 150 mm, 5 μm). The gradient is the same as the analytical method.

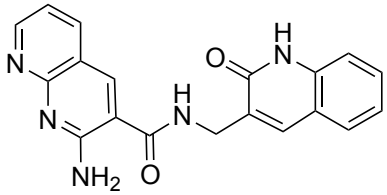
3.6.4.1 Synthesis of YA27



YA27: *N*-(4-chloro-1*H*-pyrrolo[2,3-*b*]pyridin-6-yl)-2-(1-oxo-1,2-dihydroisoquinolin-3-yl)acetamide. A 20 mL vial containing a magnetic spin bar was charged with **2-(1-oxo-1,2-dihydroisoquinolin-3-yl)acetic acid** (0.061 g, 0.2983 mmol), **1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide** (0.114 g, 0.5966 mmol), **1-Hydroxybenzotriazole** (0.0914, 0.5966 mmol) and **DMF** (5 ml). The mixture was stirred for 10 minutes under nitrogen on ice bath; then solid **4-chloro-1*H*-pyrrolo[2,3-*b*]pyridin-6-amine** (0.050 g, 0.2983 mmol) was added and the reaction vial was transferred to a mineral oil bath at 50°C. The reaction came to completion after for 24 hours. The **DMF** solvent was removed under vacuum. 1 N solution of sodium hydroxide was added to neutralize the mixture; the precipitate was filtered and purified via reverse phase chromatography (water pH = 9.8, acetonitrile) to yield *N*-(4-chloro-1*H*-pyrrolo[2,3-*b*]pyridin-6-yl)-2-(1-oxo-1,2-dihydroisoquinolin-3-yl)acetamide (0.0056 g, 0.016 mmol, 5.3 % yield) as light brown solid. ^{1}H NMR (600MHz, DMSO-d_6) δ 11.85 (s, 1H), 11.29 (s, 1H) 10.74 (s, 1H),

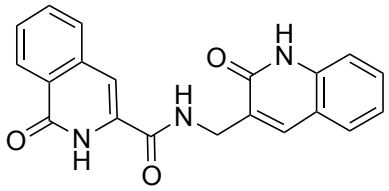
8.15 (d, J= 7.9 Hz, 1H), 8.00 (s, 1H), 7.67 (m, 1H), 7.60 (d, J= 7.92 Hz, 1H), 7.48 (t, J = 2.9 Hz, 1H), 7.45 (t, J = 7.2 Hz, 1H), 6.48 (s, 1H), 6.45 (m, 1H), 3.73 (s, 2H). ^{13}C NMR (600MHz, DMSO-d_6) δ 168.14, 162.72, 147.42, 147.12, 138.37, 136.73, 135.55, 132.89, 127.00, 126.56, 126.43, 126.41, 125.13, 115.92, 106.51, 105.06, 98.68, 40.85. HRMS calculated for $C_{18}H_{13}ClN_4O_2$ $[M+H]^+$: 353.0727 Da; found 353.0751 Da.

3.6.4.2 *Synthesis of YA28*



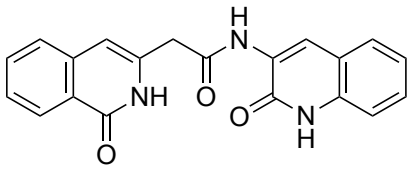
YA28: 2-amino-N-((2-oxo-1,2-dihydroquinolin-3-yl)methyl)-1,8-naphthyridine-3-carboxamide. A 20 mL vial containing a magnetic spin bar was charged with **2-amino-1,8-naphthyridine-3-carboxylic acid** (0.0675 g, 0.3561 mmol), **1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide** (0.091 g, 0.4747 mmol), **1-Hydroxybenzotriazole** (0.0727 g 0.4747 mmol) and **DMF** (5 ml). The mixture was stirred for 10 minutes under nitrogen on ice bath; then solid **3-(aminomethyl)quinolin-2(1*H*)-one** (0.050 g, 0.2374 mmol) was added and the reaction vial was transferred to a mineral oil bath at 60°C. The reaction came to completion after for 24 hours. The **DMF** solvent was removed under vacuum. 1N solution of sodium hydroxide was added to neutralize the mixture; the precipitate was filtered and purified via reverse phase chromatography (water, acetonitrile) to yield **2-amino-*N*-((2-oxo-1,2-dihydroquinolin-3-yl)methyl)-1,8-naphthyridine-3-carboxamide** (0.0366 g, 0.1061 mmol, 45% yield) as yellow solid. ^{1}H NMR (600MHz, DMSO-d_6) δ 11.96 (s, 1H), 9.23 (t, J= 5.58 Hz, 1H), 8.80 (m, 1H), 8.61 (s, 1H), 8.16 (m, 1H), 7.84 (s, 1H), 7.71 (m, 1H), 7.48 (m, 3H), 7.33 (d, J= 8.16Hz, 1H), 7.25 (m, 1H), 7.17 (m, 1H), 4.38 (d, J= 5.40, 2H). ^{13}C NMR (600MHz, DMSO-d_6) δ 167.51, 161.92, 159.05, 157.30, 154.68, 139.42, 138.37, 138.08, 135.69, 130.35, 130.27, 128.16, 122.35, 119.51, 118.47, 116.34, 116.08, 115.35, 38.94. HRMS calculated for $C_{19}H_{15}N_5O_2$ $[M+H]^+$: 346.1225 Da; found 346.1302 Da.

3.6.4.3 Synthesis of YA29



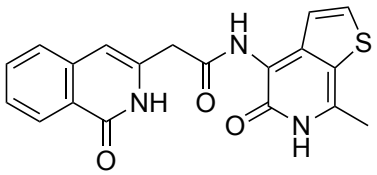
YA29: 1-oxo-*N*-((2-oxo-1,2-dihydroquinolin-3-yl)methyl)-1,2-dihydroisoquinoline-3-carboxamide. A 20 mL vial containing a magnetic spin bar was charged with **1-oxo-1,2-dihydroisoquinoline-3-carboxylic acid** (0.0675 g, 0.3561 mmol), **1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide** (0.091 g, 0.4747 mmol), **1-Hydroxybenzotriazole** (0.0727 g 0.4747 mmol) and **DMF** (5 ml). The mixture was stirred for 10 minutes under nitrogen on ice bath; then solid **3-(aminomethyl)quinolin-2(1*H*)-one** (0.050 g, 0.2374 mmol) was added and the reaction vial was transferred to a mineral oil bath at 60°C. The reaction came to completion after for 26 hours. The **DMF** solvent was removed under vacuum. 1N solution of sodium hydroxide was added to neutralize the mixture; the precipitate was filtered and purified via reverse phase chromatography (water, acetonitrile) to yield **1-oxo-*N*-((2-oxo-1,2-dihydroquinolin-3-yl)methyl)-1,2-dihydroisoquinoline-3-carboxamide** (0.0366 g, 0.034 mmol, 14% yield) as white solid. ^{1}H NMR (600MHz, DMSO-d_6) δ 11.95 (s, 1H), 10.80 (s, 1H), 9.24 (s, 1H), 8.24 (d, J= 7.92 Hz, 1H), 7.82 (s, 1H), 7.79 (m, 1H) 7.69 (d, J= 7.62 Hz, 1H), 7.62 (m, 1H), 7.48 (m, 1H), 7.42 (s, 1H), 7.39 (d, J= 8.22 Hz, 1H), 7.17 (t, J= 7.26 Hz, 1H), 4.36 (d, J= 4.74 Hz, 2H), 2.54 (s, 1H). ^{13}C NMR (600MHz, DMSO-d_6) δ 162.00, 161.87, 161.75, 150.82, 139.09, 138.39, 136.80, 135.81, 133.39, 130.09, 128.71, 128.21, 128.14, 127.44, 127.34, 122.36, 119.47, 115.34, 106.43, 40.85. HRMS calculated for $C_{20}H_{15}N_3O_3$ $[M+H]^+$: 346.1113 Da; found 346.1132 Da.

3.6.4.4 *Synthesis of YA30*



YA30: 2-(1-oxo-1,2-dihydroisoquinolin-3-yl)-*N*-(2-oxo-1,2-dihydroquinolin-3-yl)acetamide. A 20 mL vial containing a magnetic spin bar was charged with **2-(1-oxo-1,2-dihydroisoquinolin-3-yl)acetic acid** (0.0951 g, 0.468 mmol), **1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide** (0.119 g, 0.624 mmol), **1-Hydroxybenzotriazole** (0.0955, 0.624 mmol) and **DMF** (5 ml). The mixture was stirred for 10 minutes under nitrogen on ice bath; then solid **3-aminoquinolin-2(1*H*)-one** (0.050 g, 0.312 mmol) was added and the reaction vial was transferred to a mineral oil bath at 60°C. The reaction came to completion after for 48 hours. The **DMF** solvent was removed under vacuum. 1N solution of sodium hydroxide was added to neutralize the mixture; the precipitate was filtered and purified via reverse phase chromatography water, acetonitrile to yield **2-(1-oxo-1,2-dihydroisoquinolin-3-yl)-*N*-(2-oxo-1,2-dihydroquinolin-3-yl)acetamide** (0.0244 g, 0.0707 mmol, 23 % yield) as light brown solid. ^{1}H NMR (600MHz, DMSO-d_6) δ 11.95 (s, 1H), 10.79 (s, 1H), 9.24 (t, J= 5.58 Hz, 1H), 8.25 (d, J= 7.80 Hz, 1H), 7.82 (s, 1H), 7.79 (m, 2H), 7.69 (d, J= 7.32 Hz, 1H), 7.62 (m, 1H), 7.48 (m, 1H), 7.42 (s, 1H), 7.32 (d, J= 8.22 Hz, 1H), 7.17 (m, 1H), 4.36 (d, J= 5.16 Hz, 2H). ^{13}C NMR (600MHz, DMSO-d_6) δ 161.95, 161.87, 161.70, 138.39, 136.78, 135.83, 133.43, 132.93, 130.31, 130.08, 128.75, 128.23, 128.15, 127.47, 127.36, 122.37, 119.48, 115.35, 106.46, 39.13. HRMS calculated for $C_{20}H_{15}N_3O_3$ $[M+H]^+$: 346.1115 Da; found 346.1175 Da.

3.6.4.5 *Synthesis of YA31*



YA31: *N*-(7-methyl-5-oxo-5,6-dihydrothieno[2,3-*c*]pyridin-4-yl)-2-(1-oxo-1,2-dihydroisoquinolin-3-yl)acetamide. A 20 mL vial containing a magnetic spin bar was charged with **2-(1-oxo-1,2-**

dihydroisoquinolin-3-yl)acetic acid (0.075 g, 0.369 mmol), **1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide** (0.094 g, 0.492 mmol), **1-Hydroxybenzotriazole** (0.0753, 0.492 mmol) and **DMF** (5 ml). The mixture was stirred for 10 minutes under nitrogen on ice bath; then solid **4-amino-7-methylthieno[2,3-*c*]pyridin-5(6*H*)-one** (0.044 g, 0.246 mmol) was added and the reaction vial was transferred to a mineral oil bath at 60°C. The reaction came to completion after for 48 hours. The **DMF** solvent was removed under vacuum. 1N solution of sodium hydroxide was added to neutralize the mixture; the precipitate was filtered and residue was purified via reverse phase chromatography water, acetonitrile to yield ***N*-(7-methyl-5-oxo-5,6-dihydrothieno[2,3-*c*]pyridin-4-yl)-2-(1-oxo-1,2-dihydroisoquinolin-3-yl)acetamide** (0.0034 g, 0.0093 mmol, 4 % yield) as light brown solid. ^{1}H NMR (600MHz, DMSO-d_6) δ 11.31 (s, 1H), 9.65 (s, 1H), 8.44 (s, 1H), 8.15 (d, J= 7.86 Hz, 1H), 7.85 (d, J= 5.4 Hz, 1H), 7.68 (t, J= 7.14 Hz, 1H), 7.60 (d, J= 7.92 Hz, 1H), 7.45 (t, J= 7.56 Hz, 1H), 6.98 (m, 1H), 6.66 (s, 1H), 6.54 (s, 1H), 3.70 (s, 3H). HRMS calculated for $C_{19}H_{15}N_3O_3S[M+H]^+$: 366.0834 Da; found 366.0904 Da.

3.7 Acknowledgements

We thank Sven Miller for helping with collection of NMR spectra. We thank ChemAxon for providing an academic research license. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) allocation MCB130049, which is supported by National Science Foundation grant number ACI-1548562. This work was supported by a grant from the National Institute of General Medical Sciences (R01GM123336). This research was also funded in part through the NIH/NCI Cancer Center Support Grant P30 CA006927.

CONCLUSIONS

Broadly speaking, this dissertation focuses on addressing three major challenges limits the application of *in-silico* methodologies to challenging targets.

One of these challenges is the issue of false-positive-prone scoring functions; My work here introduces the first positive-aware scoring function that not only outperforms the current state-of-the-art algorithms in retrospective benchmark studies but show superior performance in prospective virtual screen. Also, my work led to the development the first benchmark dataset (D-COID) specifically created with special consideration to false positive cases. Looking ahead, I envision a wide adoption of both the dataset, as well as our novel scoring function (vScreenML) by the drug discovery community.

The second issue is fact that the unique features of RBPs makes them particularly not amenable to the traditional computational methodologies. For this, we developed a novel computational methodology for probing RNA-binding proteins. This work demonstrates a novel approach for targeting RBPs with small molecules and represents the first body of work to computationally screen for inhibitors, demonstrate binding (in multiple orthogonal assays) and report binding affinities for Msi1. Even though we used Msi1 as a proof-of-concept target, our approach here can be used for any RBP. One of the challenges we faced here is that the hit compound from our initial screen has very weak binding affinity and we had to go through several rounds of medicinal chemistry optimization to arrive at a low micromolar compound. The reason for this observation, we hypothesize, might be because off-shelf libraries are biased towards traditional drug targets.

To address the inherent bias of current screening libraries towards traditional drug targets, I built a first RBP focused library. With this, I was able to show the value of building a target specific library; Right from the first screening and testing only five compounds, we got a low micromolar inhibitor (YA27). We showed that building a Msi1 focused library obviate the need for medicinal chemistry optimization to arrive at a low micromolar binder.

Overall, the combination of all these tools is poised to change the landscape of drug discovery as it relates to challenging targets, particularly RBPs.

\

FUTURE DIRECTIONS

In my work on building D-COID and vScreenML, one possible future direction is expansion of the physicochemical property space of the compounds in the D-COID set. Also, a lot of the features used in building vScreenML come from proprietary packages like OpenEye Szybki and ChemAxon; in the future, we would like to replace these with open source tools that have the same or similar function.

In our MsiLib project, we intend to include more reactions in order to improve the diversity of our novel library (MsiLib). For example, we can incorporate all the 120 reactions used in building Enamine library to build a library that is not just huge but also replete with complementary ligands.

Finally, we intend to extend the application all the novel tools we have built to other RBPs like Serine-rich Splicing Factor 2 (SRSF2).

APPENDIX A: Supporting Information for Chapter 1

Supporting Figures

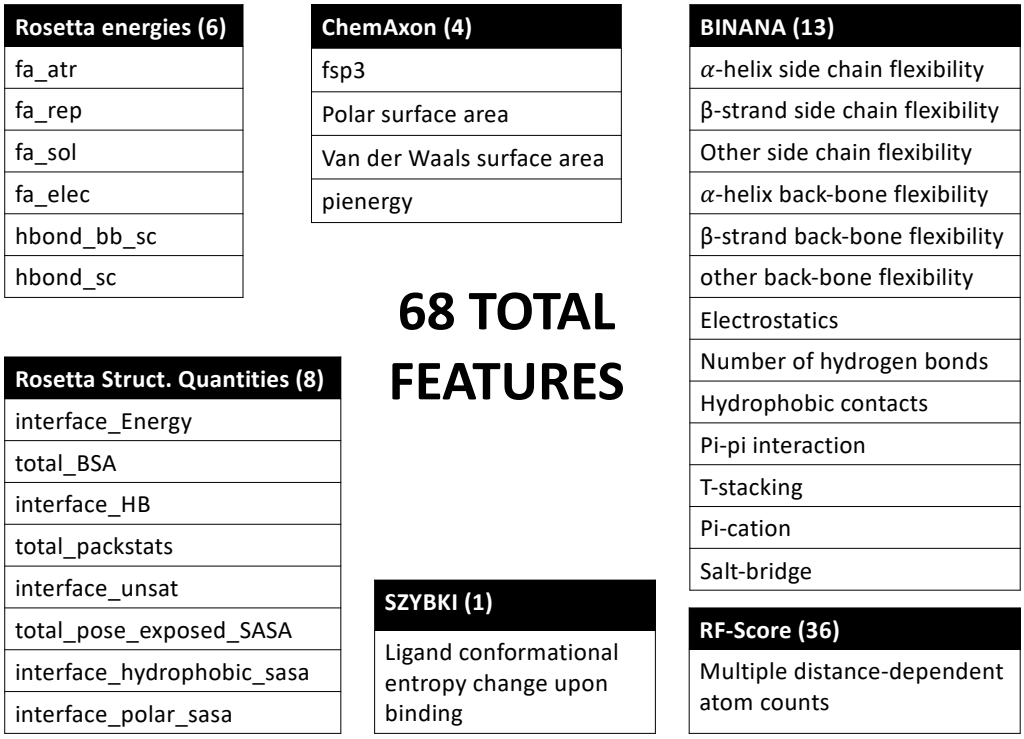


Figure S1: Features incorporated into vScreenML. These features derive from six sources: Rosetta energy terms, Rosetta structural quantifiers, RF-Score's rfscore_v1 features, BINANA's analysis of intermolecular contacts, ChemAxon's cxcalc features, OpenEye's SZYBKI conformational entropy term.

(A)

Metrics	XGB-Op	XGB
Accuracy	0.90	0.89
Precision	0.87	0.86
Recall	0.70	0.67
AUC	0.83	0.81
F1-Score	0.78	0.75
Matt	0.72	0.69

(B)

Parameters	XGB-Pre (Default)	XGB-after (Optimized)
Learning rate	0.3	0.01
Min_child_weight	1	1
Max_depth	6	7
Gamma	0	0.1
Subsample	1	0.5
Colsample_bytree	1	0.4
Lambda	1	Default
Alpha	0	Default
Scale_pos_weught	1	1
N_estimators	100	1945

Figure S2: Hyperparameter tuning of vScreenML. (A) Performance comparison of the optimized and non-optimized vScreenML models. **(B)** XGBoost parameters before and after optimization.

(A)

Features	Accuracy	Precision	Recall	AUC	Mathews
XGBoost	0.89	0.86	0.67	0.81	0.69
Gradient Boosting	0.89	0.85	0.67	0.81	0.69
Random Forest	0.86	0.83	0.55	0.76	0.61
Extra Trees	0.86	0.85	0.53	0.76	0.60
Support Vector Machine	0.75	0.00	0.00	0.50	0.00
Linear Discriminant Analysis	0.87	0.79	0.65	0.78	0.63
Quadratic Discriminant Analysis	0.38	0.28	0.96	0.56	0.17
Gaussian Naïve Bayes	0.50	0.32	0.90	0.63	0.26
K-nearest Neighbour (KNN)	0.73	0.45	0.25	0.57	0.18
DUMB	0.75	0.00	0.00	-	0.00

(B)

Features	Accuracy	Precision	Recall	AUC	Mathews
Rosetta (Reweighted)	0.85	0.75	0.57	0.76	0.56
RF (Reweighted)	0.78	0.66	0.30	0.62	0.34
BINANA (Reweighted)	0.83	0.75	0.50	0.72	0.51
Rosetta+NC	0.86	0.79	0.60	0.77	0.60
Rosetta+NC+LigProp	0.86	0.79	0.60	0.77	0.62
Rosetta+NC+LigProp+Szybki	0.86	0.79	0.60	0.77	0.62
Rosetta+NC+LigProp+Szybki+RF	0.88	0.83	0.65	0.80	0.67
Rosetta+NC+LigProp+Szybki+BINANA	0.87	0.82	0.62	0.79	0.65
Rosetta+NC+LigProp+Szybki+RF+BINANA	0.89	0.86	0.67	0.81	0.69

Figure S3: Performance of alternate models. (A) Using the complete vScreenML feature set, alternate frameworks are used for building the model. **(B)** Examination of models in which a set of features from a given origin is removed en-masse; all models are trained using XGBoost.

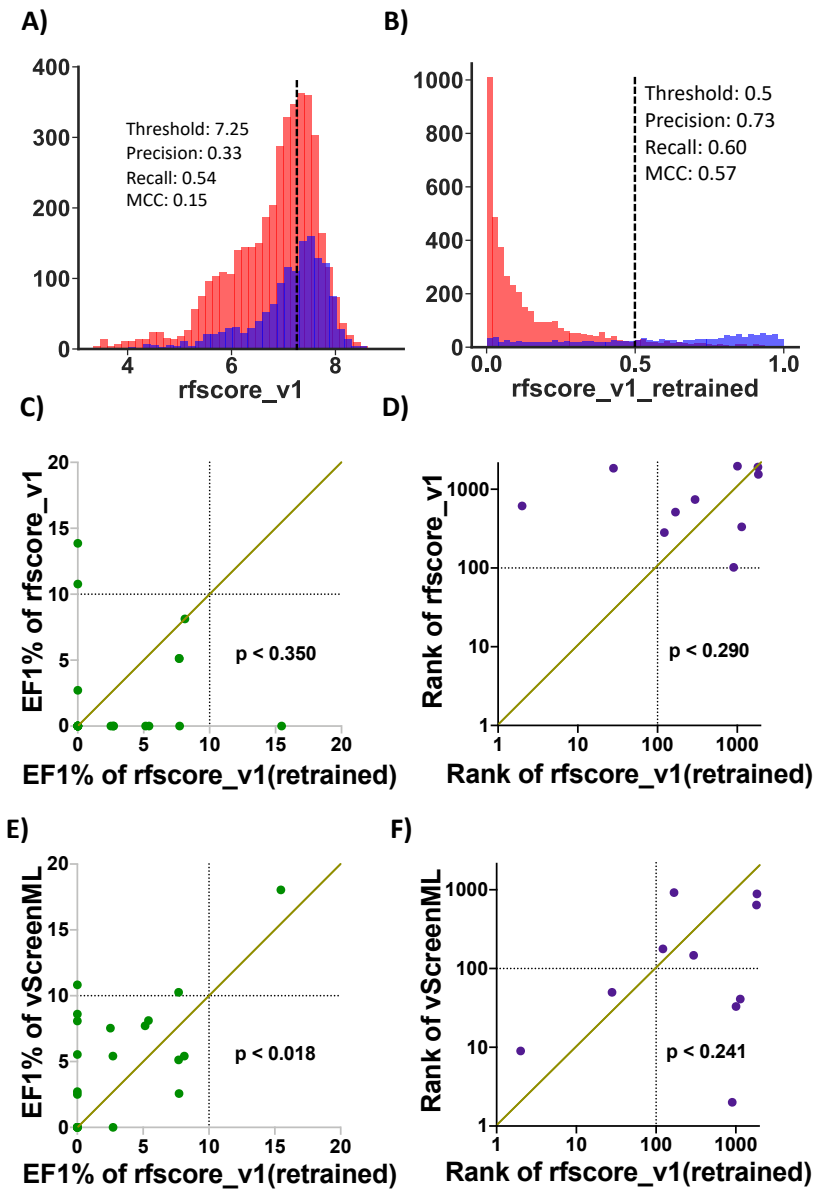


Figure S4: Retraining rfscore_v1 using D-COID. (A) Overlaid histograms for scores obtained when scoring active complexes (*blue*) and decoy complexes (*red*) from D-COID using the original rfscore_v1. **(B)** Overlaid histograms after re-training rfscore_v1. **(C)** Comparison of the original and re-weighted versions of rfscore_v1 applied to the DEKOIS benchmark. **(D)** Comparison of the original and re-weighted versions of rfscore_v1 applied to the PPI benchmark. **(E)** Comparison of re-weighted rfscore_v1 versus vScreenML on the DEKOIS benchmark. **(F)** Comparison of re-weighted rfscore_v1 versus vScreenML on the PPI benchmark.

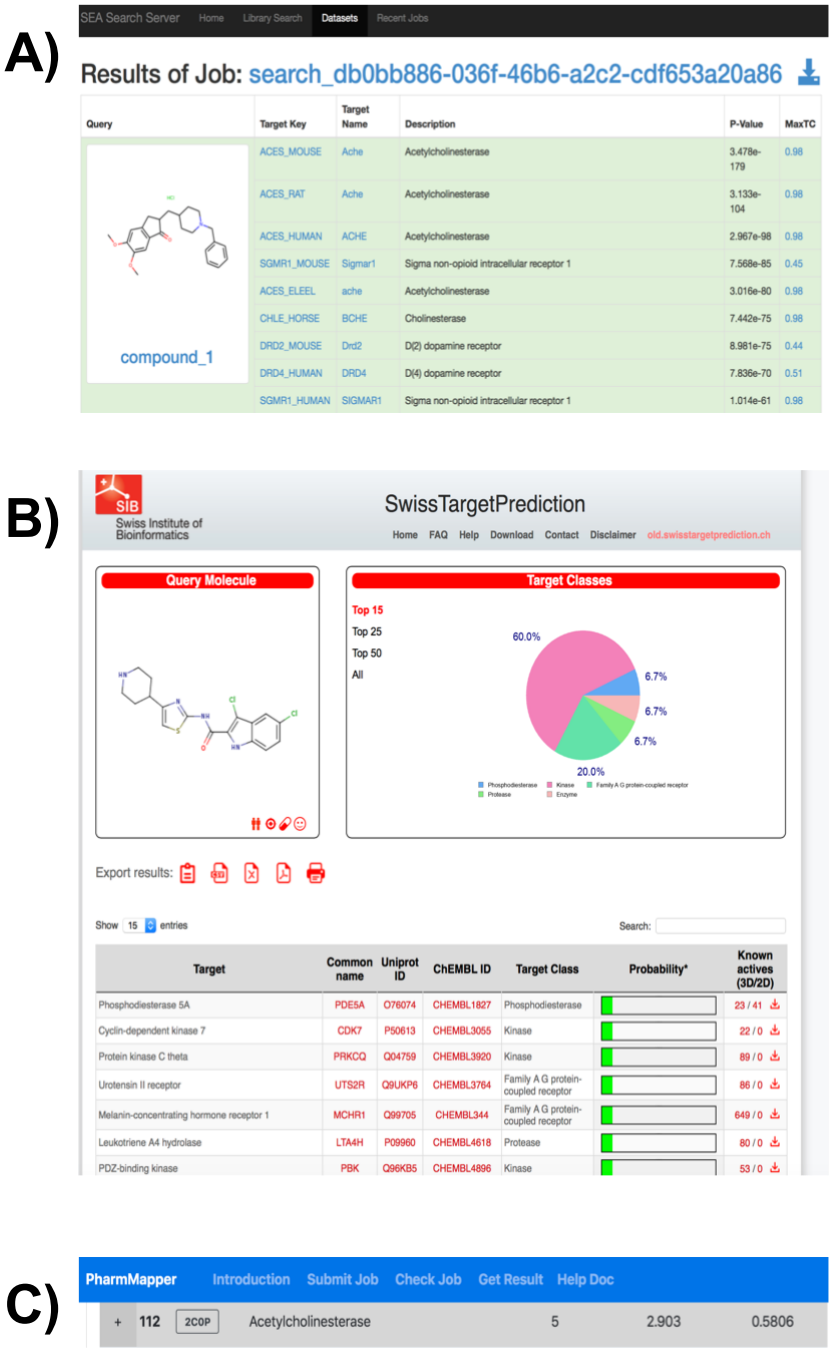


Figure S5: Positive control for target identification methods. We confirmed that all three methods would successfully identify AChE as the target of a known AChE inhibitor (donepezil, CHEMBL1678). **(A)** Similarity Ensemble Approach (SEA). **(B)** SwissTargetPrediction. **(C)** PharmMapper. We note that AChE was only ranked #112 among the PharmMapper hits because the 3D conformations it built for donepezil were not sufficiently well-matched to the active conformation to produce a better ranking.

Compound	Identified as a top-scoring hit
AC6	1st round
AC3	1st round
AC10	Both rounds
AC11	1st round
AC15	2nd round
AC5	2nd round
AC13	2nd round
AC19	Both rounds
AC23	1st round
AC9	Both rounds

Table S1: Provenance of AChE inhibitors. For each of the 10 AChE inhibitors that provided more than 50% inhibition at a concentration of 50 μM, we determined at what stage this compound was prioritized for testing. Our strategy included two stages of screening: first we screened only 15 million diverse compounds from the Enamine collection, then we expanded our search by collecting analogs for each of these hits. We note that 7 of these 10 compounds were identified in the first round of screening; after re-refinement in the second round, 3 of these were still highly-ranked whereas 4 had been surpassed by analogs (or received lower scores upon re-refinement). Only 3 of these 10 compounds would have been missed if our screening had been limited to a single round of 15 million compounds.

APPENDIX B: Supporting Information for Chapter 2

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) [135].

Building hotspot pharmacophores

Hotspot pharmacophores were built using a new dedicated protocol implemented in the Rosetta software suite [124], and is freely available for academic use (www.rosettacommons.org). The hotspot pharmacophore is extracted solely from the structure of the protein-RNA complex, and thus does not take into any account potential rearrangement of the protein upon RNA binding.

To select deeply buried RNA bases, the solvent accessible surface area (SASA) of each base in the RNA was calculated in the presence of the protein, then re-calculated after deleting the protein: this yielded the SASA that was directly buried by the protein in the complex. 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) [195] 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 new functionality 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. This new 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 [125]. We generated up to 100 conformers for each molecule in the database using OMEGA [126,82,127]. 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 [124]. This energy function includes terms that capture packing, hydrogen bonding, implicit solvation (modeled via EEF1 [196]), sidechain rotamer preferences, and backbone dihedral preferences. After minimization, 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 was 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
```

The multidimensional scaling (MDS) analysis presented in Figure 5 was carried out in R [197], using the "cmdscale" function. Pharmacophores for hnRNP A1 and Prp24 were extracted from PDB IDs 4YOE and 4N0T, respectively.

Compounds optimization

We carried out sequential two rounds of optimization via automated fingerprint-based searched, followed by a round of traditional medicinal chemistry optimization by combining desirable fragments and R-groups from the best compounds from earlier rounds of optimization. For each of the rounds of the fingerprint-based optimization, we took advantage of Enamine Real Database, a database of greater than 11 billion readily synthesizable compounds that boasts speed of compounds delivery and affordable price. We started by querying the database with our initial hit molecule from screening ZINC library (R12), based on fingerprint similarity score, we selected the top 1000 compounds. From this, we cluster based on diversity of substituent on the rings and linker length and select representative compounds of the clusters for biochemical characterization. In the first round of screening, we choose 16 compounds; In the second round, we selected 50 diverse compounds based on similarity to the best compound from the first round. Finally, the knowledge acquired from the structure activity relationship of the 66 compounds tested so far

was combined in designing new inhibitors, in a typical traditional medicinal chemistry style; but, rather synthesizing the compounds, we found the most similar compound in Enamine database and purchase them. With this, we purchased 7 additional compounds for biochemical characterization.

Model building of R12 derivatives

For each round of optimization carried out, we built structural models for the compounds to assess whether the compounds are making appropriate interaction with the protein. For this, we generated 300 low energy 3D conformations for each of the compounds using OpenEye OMEGA[126,82,127]. Then, we aligned each of the conformers to Msi1 pharmacophore (generated from model 1 of the NMR structure with PDB ID 2RS2) using OpenEye ROCS[128]; then, for each compound, we selected top 10 conformers with the highest shape and electrostatic overlap with the pharmacophore (TanimotoCombo). We placed these in the binding site of the protein and energy-minimized the complexes using Rosetta energy function. Finally, for each compound, we selected the conformer with the lowest energy.

Protein expression and purification

The RRM1-RRM2 domain of human Msi1 and Msi2 were purchased from Genewiz as a fusion protein with an N-terminal 6xHis-tag and a tobacco etch virus (TEV) protease site on vector pET28a (+). The RRM1 domain of human Msi1 with an N-terminal 6xHis-tagged streptococcal GB1 domain and hnRNP A1 with an N-terminal 6xHis-tagged maltose- binding protein (MBP) fusion proteins were purchased from Genewiz, both on vector pET28a (+). Each of these four constructs were expressed and purified as described below.

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 at 37 ºC for 4 hours. The induced cells were harvest and the pellet was resuspended in lysis buffer (20 mM HEPES, 1 M NaCl, 50 mM imidazole, 1 mM DTT, pH 7.4) and sonicated for 10 minutes (Fisher Scientific Sonic Dismembrator Model 100). The

cell lysates were then centrifuged at 15,000g for 50 min. The protein of interest remained in the supernatant, which was purified by HPLC affinity chromatography with Ni-chelated Sepharose Fast Flow Resin (GE Healthcare). The buffer was exchanged with dialysis (20 mM HEPES, 150 mM NaCl, 0.1 mM EDTA, 1 mM DTT, pH 7.4).

All protein concentrations were determined with reference to bovine albumin standards using Bradford assays.

Fluorescence polarization competition assays

RNA oligonucleotides (UAGGUAGUAGU/36-FAM/) were purchased from Integrated DNA Technologies (Coralville, IA) and dissolved in RNase free water. To measure the dissociation constant of Msi1 RRM1-RRM2 and RNA binding, a fixed concentration (5 nM) of fluorescein-labeled RNA and increasing concentrations of Msi1 RRM1-RRM2 (0 nM to 128 nM) were mixed in binding assay buffer (20 mM HEPES, 150 mM NaCl, 0.01% Triton-X100 pH 7.4). Fluorescence intensities were measured in replicate on the Molecular Devices SpectraMax® i3x (San Jose, CA) and the fluorescence polarization value (FP) was calculated by the following equation:

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

where $I_{\parallel}$ refers to the intensity of the parallel fluorescence and $I_{\perp}$ refers to the intensity of the perpendicular fluorescence. The dissociation constant (K_D) was fit using Prism 6 (GraphPad Software Inc.) as follows:

$$Y = Bottom + \frac{Top - Bottom}{1 + EC_{50}^{Hill\ Slope}/L^{Hill\ Slope}}$$

To examine the displacement of RNA by R12 derivative compounds, the competition assays were performed with 8 nM as fixed Msi1 concentration, 5 nM as fixed fluorescein-labeled RNA concentration, and a serial dilution of compounds were added. Data were fit to a single-site competition model to determine IC50 using Prism 6 (GraphPad Software Inc.). There are two RRM domains in the protein construct

therefore the Hill coefficient was set to be free. Given the known experimental conditions and the binding affinity described above, the Ki was then computed from the IC50 using the method of Nikolovska-Coleska et al [198].

To test the contribution to binding affinity of each base in the NUMB RNA sequence, we purchased eight shorter oligos that one with the fluorescein-label, five of them harbor an abasic site at a different position, as well as the corresponding wild-type (**Table S1**). The shorter fluorescein-labeled RNA was applied here to match the other RNA fragments with the abasic site. K_i values were then determined from the competition experiment described as above.

Differential scanning fluorimetry (Thermofluor)

Differential scanning fluorimetry (DSF) experiments were carried out using a standard protocol described by others [199]. The protein concentration was fixed to be 4 μM and SYPRO Orange (Invitrogen S6651) was used at a final concentration of 5X. The experiments were carried out in 20 mM HEPES, 150 mM NaCl, 0.01% Triton-X100, 2.5% DMSO pH 7.4, with Eppendorf Realplex2 Mastercycler. Each sample was divided to three 50 μL replicates. Sample solutions were dispensed into 96-well optical reaction plate (Thermo Fisher Scientific 4306737) and the plate was sealed with optical PCR plate sheet (Thermo Fisher Scientific AB-1170). Fluorescence intensity was measured via the JOE emission filter (550 nm) and "PTS clear plate" was set as the background for the calibration. Temperature was continuously increased: 0.4 ºC/min, from 37 ºC to 56.6 ºC. Melting curves were directly exported from the instrument, and then were analyzed with Prism 6 (GraphPad Software Inc.).

Nuclear magnetic resonance (NMR) spectroscopy

^{15}N-labeled protein was expressed and purified as described above then cleaved with TEV overnight at 4 ºC in 20 mM HEPES pH 6.3, 50 mM NaCl, and 2 mM DTT in a 1:20 ratio. Cleaved protein was then passed over a 5 mL HisTrap column. Pure fractions were then pooled and concentrated to 1 mL.

Buffer exchange was performed using a NAP10 column (GE Healthcare) into 20mM HEPES pH 7.0, 1 mM TCEP and 10% D_2O.

All spectra were recorded at 298K on a Bruker Ascend 600-MHz spectrometer. DMSO control spectra were prepared by adding DMSO to a final concentration of 0.5% in a 75 μM protein solution. Compounds R12.8.4.44 and R12.8.44.3 were added to final concentrations of 150 μM and 100 μM, respectively. All data were processed using TopSpin 4.0.

Supporting Tables

Name	Sequence
FC-NUMB	**5'- F-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).

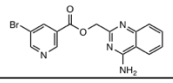
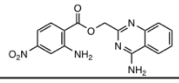
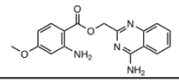
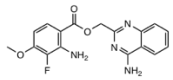
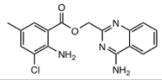
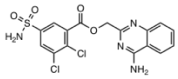
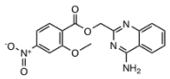
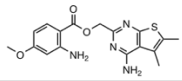
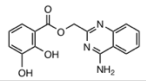
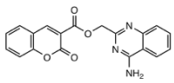
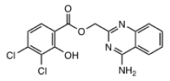
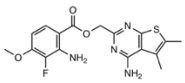
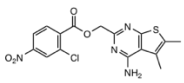
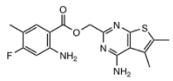
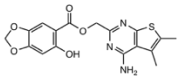
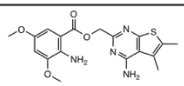
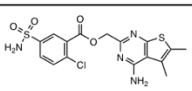
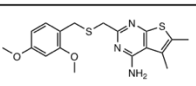
Compound	2D Structure	EC50 (μM) in Fluorescence Polarization Assay
R12		N.D.*
R12-7		49
R12-8		> 100
R12-8-19		> 100
R12-8-22		90
R12-8-29		100
R12-8-38		> 100
R12-8-44		22
R12-8-46		9
R12-8-47		> 100
R12-8-48		> 100
R12-8-44-1		16
R12-8-44-2		> 100
R12-8-44-3		9
R12-8-44-4		6
R12-8-44-6		> 100
R12-8-44-7b		> 100
R12-8-44-lk2		26

Table S2: Three rounds of compound optimization, starting from initial hit R12.

Supporting Figures

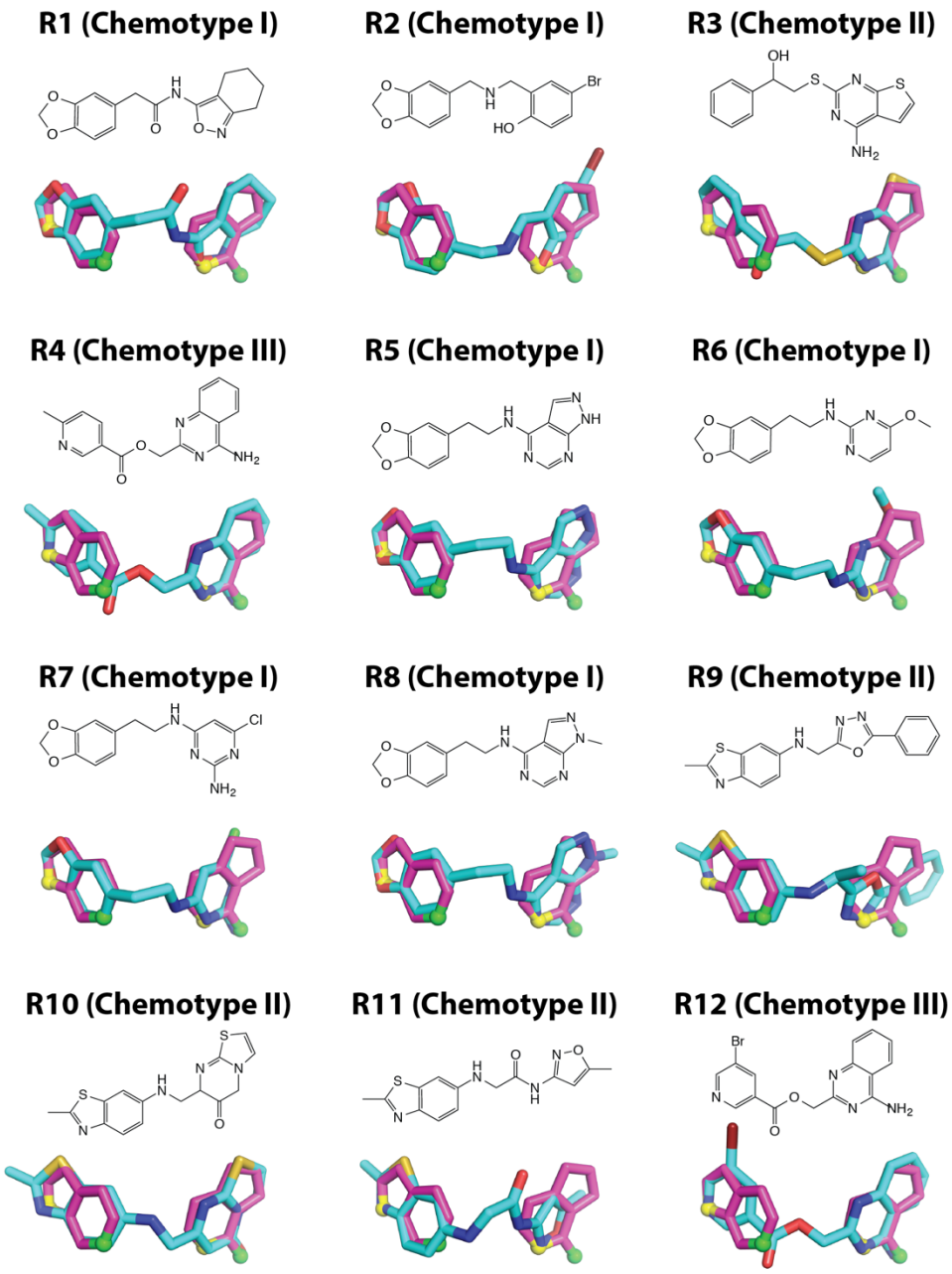


Figure S1: The 12 initial hit compounds. The chemical structure is shown for each compound, as well as a three-dimensional model of each compound (*cyan*) superposed with the Msi1 RBD1 hotspot pharmacophore (*magenta*).

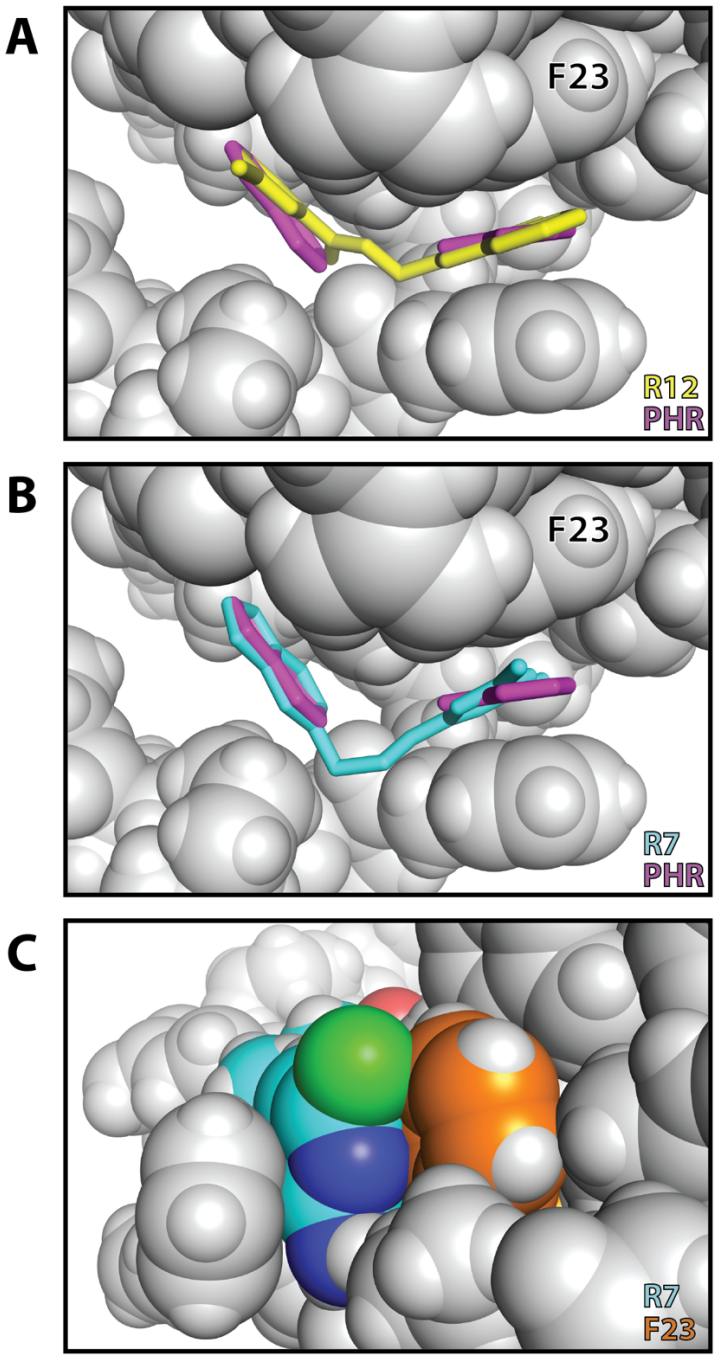


Figure S2: An inadvertent steric clash may explain the lack of binding by R7. (A) The rings in the model of R12 (*yellow*) are well-superposed with those of the hotspot pharmacophore (*magenta*), allowing for aromatic stacking with Msi1. **(B)** The relative positioning of the rings in the R7 (*cyan*) 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*).

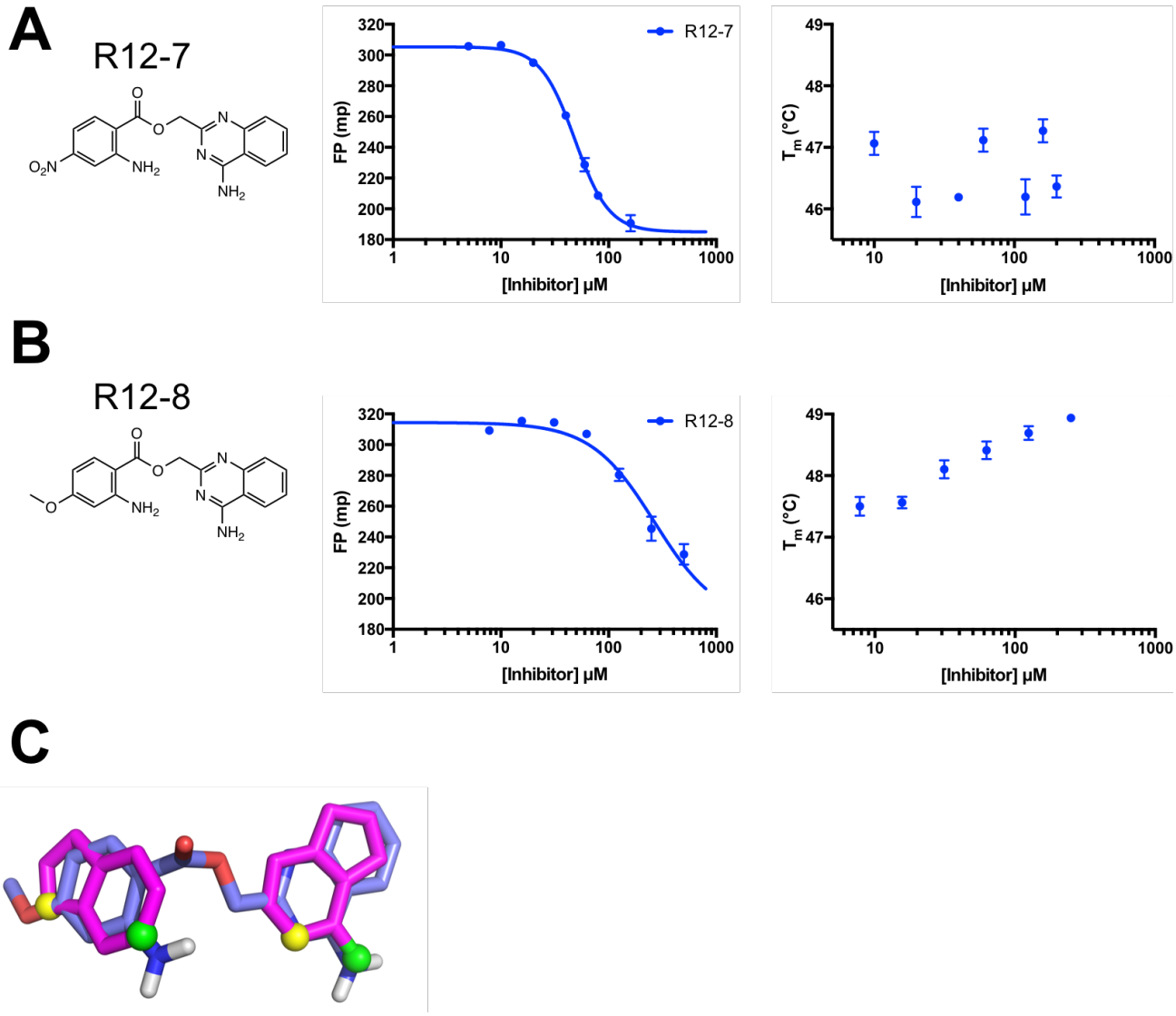


Figure S3: Inhibitors coming from the first round SAR study. (A) The chemical structure of R12-7 and the results of biochemical assays. The middle panel is the FP competition assay and the right panel is the DSF assay. **(B)** The chemical structure of R12-8 and the results of biochemical assays. The middle panel is the FP competition assay and the right panel is the DSF assay. **(C)** Superposition of R12-8 (*slate*) and the hotspot pharmacophore (*magenta*).

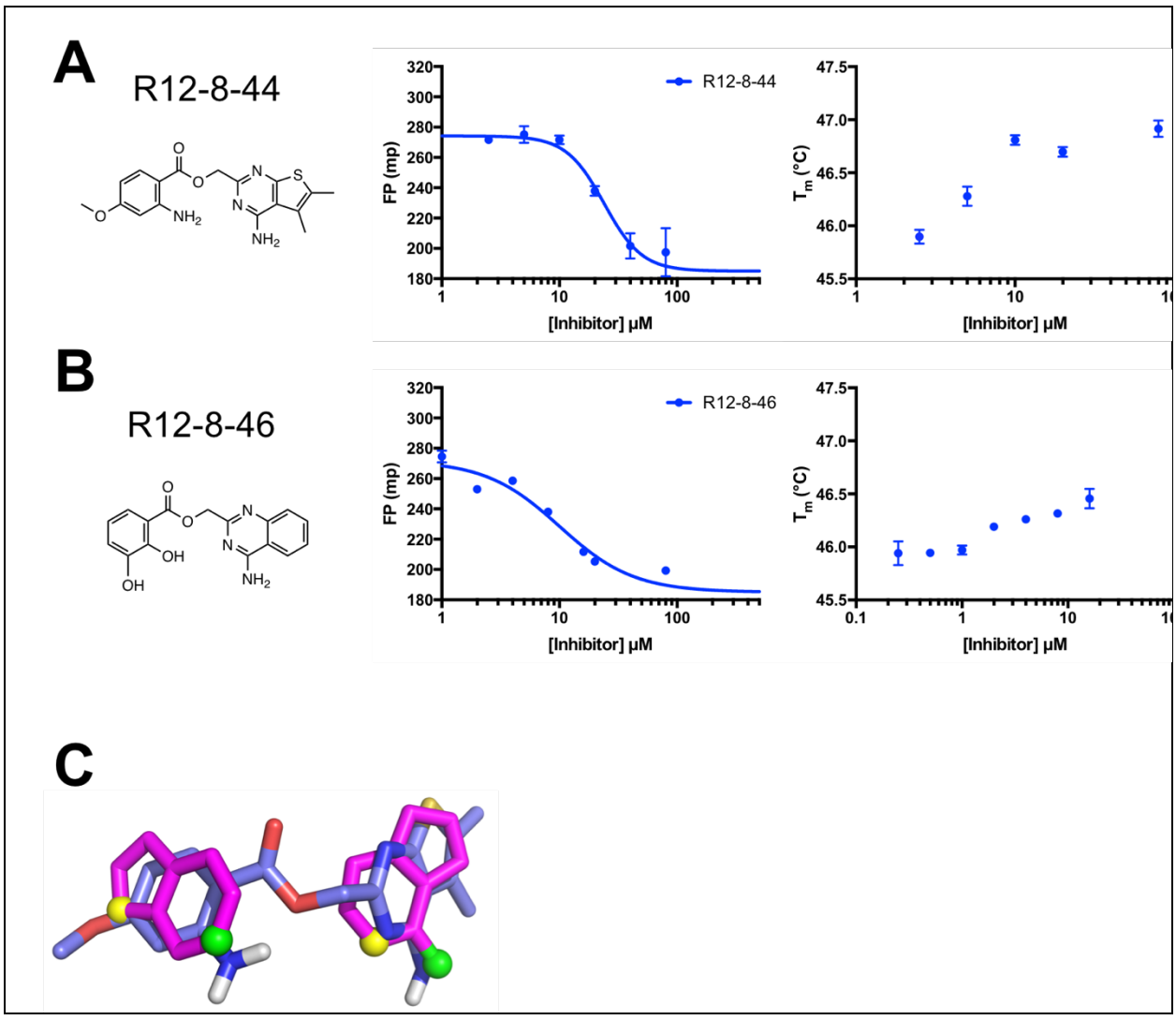


Figure S4: Inhibitors coming from the second round SAR study. (A) The chemical structure of R12-8-44 and the results of biochemical assays. The middle panel is the FP competition assay and the right panel is the DSF assay. **(B)** The chemical structure of R12-8-46 and the results of biochemical assays. The middle panel is the FP competition assay and the right panel is the DSF assay. **(C)** Superposition of R12-8-44 (*slate*) and the hotspot pharmacophore (*magenta*).

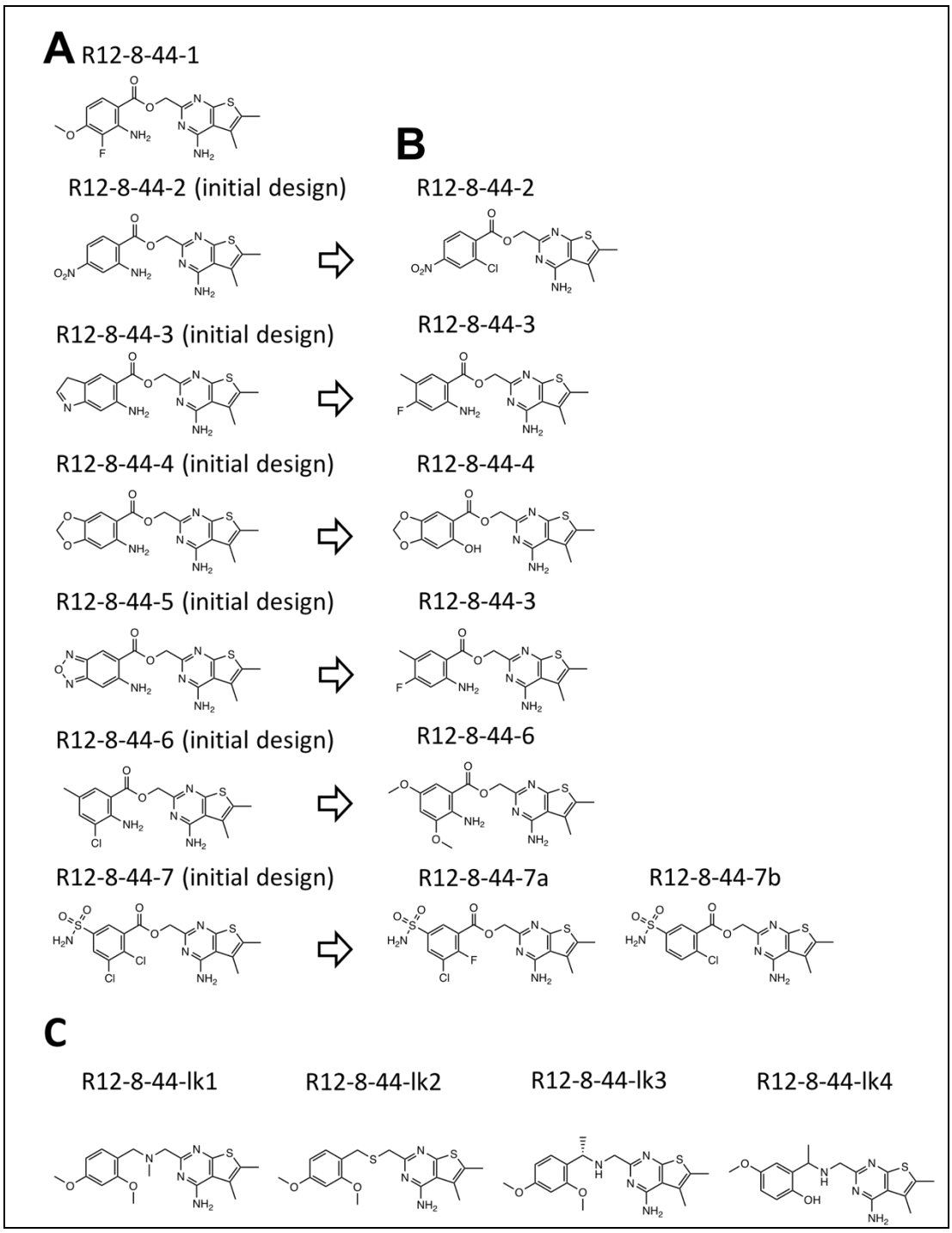


Figure S5: Compounds design of the third round SAR study. (A) Chemical structures of the initial compounds designed based on R12-8-44. **(B)** Chemical structures of the compounds which were commercially available. **(C)** Chemical structures of the compounds with similar structures as R12-8-44 but different linkers.

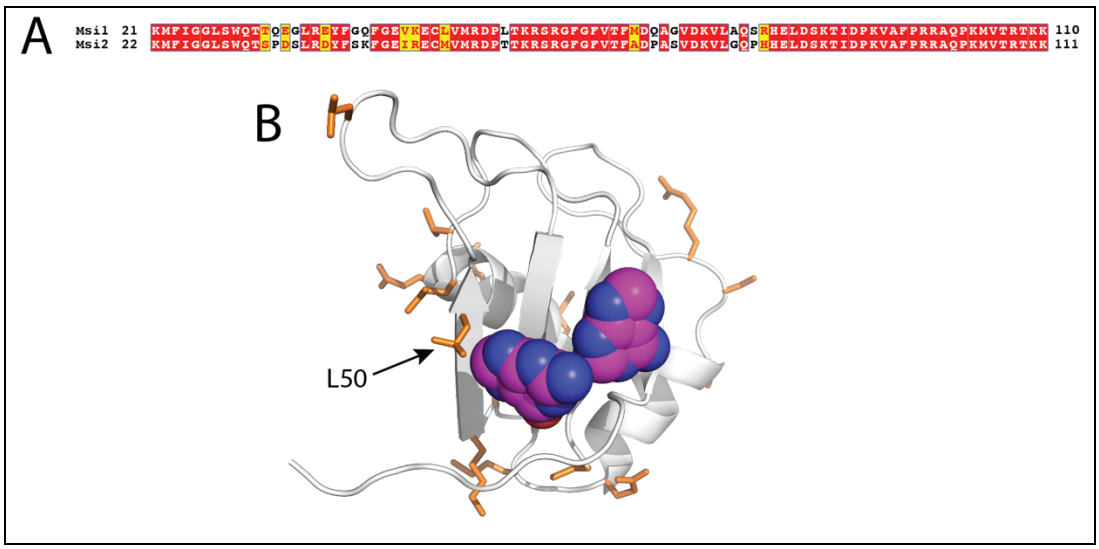


Figure S6: Comparison of Msi1 and Msi2. (A) Sequence alignment comparing RBD1 from Msi1 to RBD1 from Msi2. Over these 90 residues, only 17 positions differ (81% sequence identity) and only 9 correspond to non-conservative mutations (90% sequence similarity). This alignment was rendered using ESPript [200,201]. **(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.

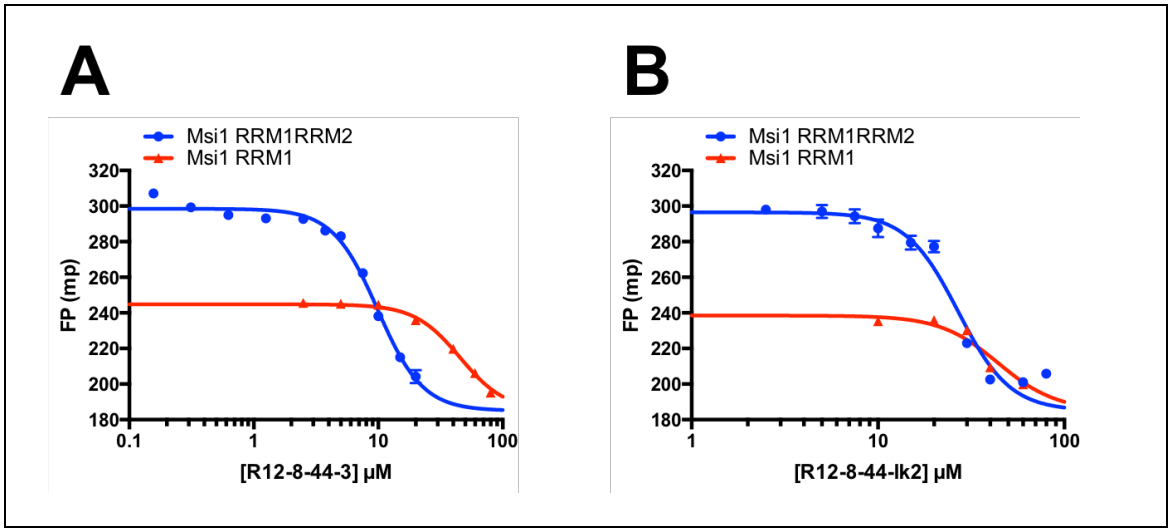


Figure S7: Compounds activity: Msi1 RRM1+RRM2 vs. Msi1 RRM1. (A) FP competition assay of R12-8-44-3 with Msi1 RRM1+RRM2 (*blue*) vs. Msi1 RRM1 only (*red*). **(B)** FP competition assay of R12-8-44-lk2 with Msi1 RRM1+RRM2 (*blue*) vs. Msi1 RRM1 only (*red*).

APPENDIX C: Supporting Information for Chapter 3

Supporting Figures

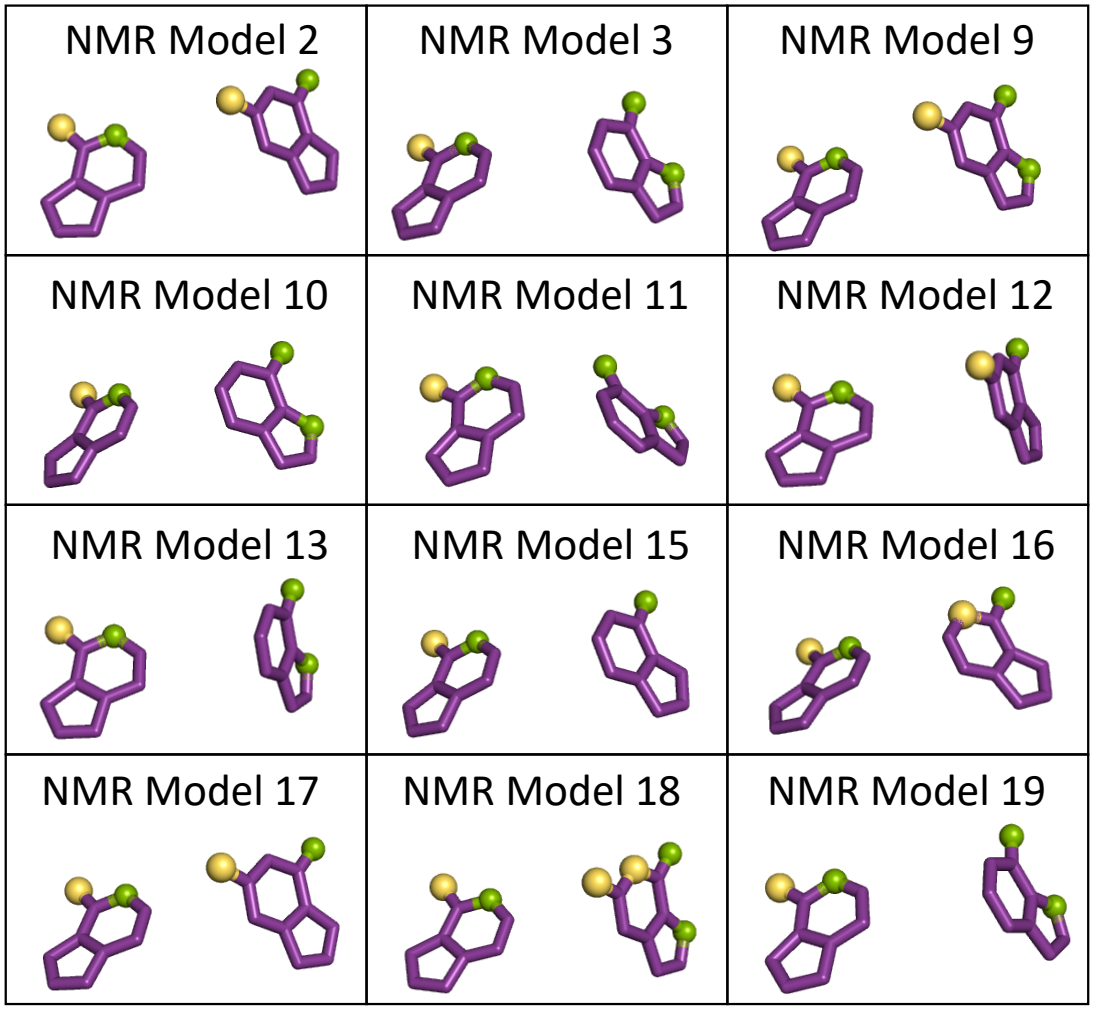


Figure S1: Pharmacophores extracted from the different models in the NMR ensemble. Our algorithm extracted 12 structurally and geometrically different pharmacophores from the NMR ensemble of Msi-1. These different patterns inspired the design of the chemical building blocks used in building MsiLib.

REFERENCES

1. Ballester PJ, Mitchell JB. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*. 2010; 26:1169-75.

2. Sato T, Honma T, Yokoyama S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *J Chem Inf Model*. 2010; 50:170-85.

3. Li H, Leung KS, Wong MH, Ballester PJ. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol Inform*. 2015; 34:115-26.

4. Heck GS, Pintro VO, Pereira RR, de Avila MB, Levin NMB, de Azevedo WF. Supervised Machine Learning Methods Applied to Predict Ligand- Binding Affinity. *Curr Med Chem*. 2017; 24:2459-70.

5. Wojcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci Rep*. 2017; 7:46710.

6. Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics*. 2018; 34:i821-i9.

7. Wojcikowski M, Siedlecki P, Ballester PJ. Building Machine-Learning Scoring Functions for Structure-Based Prediction of Intermolecular Binding Affinity. *Methods Mol Biol*. 2019; 2053:1-12.

8. Lindert S, Zhu W, Liu YL, Pang R, Oldfield E, McCammon JA. Farnesyl diphosphate synthase inhibitors from in silico screening. *Chem Biol Drug Des*. 2013; 81:742-8.

9. Spencer RW. High-throughput screening of historic collections: observations on file size, biological targets, and file diversity. *Biotechnol Bioeng*. 1998; 61:61-7.

10. Bleicher KH, Bohm HJ, Muller K, Alanine AI. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov*. 2003; 2:369-78.

11. Pritchard JF, Jurima-Romet M, Reimer ML, Mortimer E, Rolfe B, Cayen MN. Making better drugs: Decision gates in non-clinical drug development. *Nat Rev Drug Discov*. 2003; 2:542-53.

12. Jacoby E, Boettcher A, Mayr LM, Brown N, Jenkins JL, Kallen J, Engeloch C, Schopfer U, Furet P, Masuya K, Lisztwan J. Knowledge-based virtual screening: application to the MDM4/p53 protein-protein interaction. *Methods Mol Biol*. 2009; 575:173-94.

13. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou SB, Diaz LA, Kinzler KW. Cancer Genome Landscapes. *Science*. 2013; 339:1546-58.

14. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, Gao J, Socci ND, Solit DB, Olshen AB, Schultz N, Taylor BS. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol*. 2016; 34:155-63.

15. Bunnage ME, Chekler EL, Jones LH. Target validation using chemical probes. *Nat Chem Biol*. 2013; 9:195-9.

16. Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, Green DV, Hertzberg RP, Janzen WP, Paslay JW, Schopfer U, Sittampalam GS. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov*. 2011; 10:188-95.

17. Clare RH, Bardelle C, Harper P, Hong WD, Borjesson U, Johnston KL, Collier M, Myhill L, Cassidy A, Plant D, Plant H, Clark R, Cook DAN, Steven A, Archer J, McGillan P, Charoensutthivarakul S, Bibby J, Sharma R, Nixon GL, Slatko BE, Cantin L, Wu B, Turner J, Ford L, Rich K, Wigglesworth M, Berry NG, O'Neill PM, Taylor MJ, Ward SA. Industrial scale high-throughput screening delivers multiple fast acting macrofilaricides. *Nat Commun*. 2019; 10:11.

18. Hu Q, Peng Z, Kostrowicki J, Kuki A. LEAP into the Pfizer Global Virtual Library (PGVL) space: creation of readily synthesizable design ideas automatically. *Methods Mol Biol*. 2011; 685:253-76.

19. Hu Q, Peng Z, Sutton SC, Na J, Kostrowicki J, Yang B, Thacher T, Kong X, Mattaparti S, Zhou JZ, Gonzalez J, Ramirez-Weinhouse M, Kuki A. Pfizer Global Virtual Library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb Sci*. 2012; 14:579-89.

20. Sterling T, Irwin JJ. ZINC 15-Ligand Discovery for Everyone. *J Chem Inf Model*. 2015; 55:2324-37.

21. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, O'Meara MJ, Che T, Algaa E, Tolmachova K, Tolmachev AA, Shoichet BK, Roth BL, Irwin JJ. Ultra-large library docking for discovering new chemotypes. *Nature*. 2019; 566:224-9.

22. McInnes C. Virtual Screening Strategies in Drug Discovery. *Curr Opin Chem Biol*. 2007; 11:494-502.

23. Lavecchia A, Di Giovanni C. Virtual screening strategies in drug discovery: a critical review. *Curr Med Chem*. 2013; 20:2839-60.

24. Irwin JJ, Shoichet BK. Docking Screens for Novel Ligands Conferring New Biology. *J Med Chem*. 2016; 59:4103-20.

25. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL. Assessing Scoring Functions for Protein-Ligand Interactions. *J Med Chem*. 2004; 47:3032-47.

26. Meng EC, Shoichet BK, Kuntz ID. Automated docking with grid-based energy evaluation. *J Comput Chem*. 1992; 13:505-24.

27. Novikov FN, Zeifman AA, Stroganov OV, Stroylov VS, Kulkov V, Chilov GG. CSAR Scoring Challenge Reveals the Need for New Concepts in Estimating Protein–Ligand Binding Affinity. *J Chem Inf Model*. 2011; 51:2090-6.

28. Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des*. 1994; 8:243-56.

29. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J Comput Aided Mol Des*. 1997; 11:425-45.

30. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem*. 2004; 47:1739-49.

31. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. LigScore: A Novel Scoring Function for Predicting Binding Affinities. *J Mol Graph Model*. 2005; 23:395-407.

32. Muegge I, Martin YC. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *J Med Chem*. 1999; 42:791-804.

33. Mooij WT, Verdonk ML. General and targeted statistical potentials for protein-ligand interactions. *Proteins*. 2005; 61:272-87.

34. Irwin JJ, Shoichet BK. ZINC - a Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model*. 2005; 45:177-82.

35. Durrant JD, McCammon JA. NNScore 2.0: a neural-network receptor-ligand scoring function. *J Chem Inf Model*. 2011; 51:2897-903.

36. Wallach I, Heifets A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J Chem Inf Model*. 2018; 58:916-32.

37. Chen L, Cruz A, Ramsey S, Dickson CJ, Duca JS, Hornak V, Koes DR, Kurtzman T. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One*. 2019; 14:e0220113.

38. Liu S, Alnammi M, Ericksen SS, Voter AF, Ananiev GE, Keck JL, Hoffmann FM, Wildman SA, Gitter A. Practical Model Selection for Prospective Virtual Screening. *J Chem Inf Model*. 2019; 59:282-93.

39. Ballester PJ, Mangold M, Howard NI, Robinson RLM, Abell C, Blumberger J, Mitchell JBO. Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification. *J R Soc Interface*. 2012; 9:3196-207.

40. Ballester PJ, Schreyer A, Blundell TL. Does a More Precise Chemical Description of Protein-Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J Chem Inf Model*. 2014; 54:944-55.

41. Li H, Leung KS, Wong MH, Ballester PJ. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. *BMC Bioinformatics*. 2014; 15:291.

42. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-Ligand Scoring with Convolutional Neural Networks. *J Chem Inf Model*. 2017; 57:942-57.

43. Abbasi WA, Asif A, Ben-Hur A, Minhas F. Learning protein binding affinity using privileged information. *BMC Bioinformatics*. 2018; 19:425.

44. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics*. 2018; 34:3666-74.

45. Wojcikowski M, Kukielka M, Stepniewska-Dziubinska MM, Siedlecki P. Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*. 2019; 35:1334-41.

46. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One*. 2017; 12:e0177678.

47. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J Med Chem*. 2012; 55:6582-94.

48. Hawkins PC, Skillman Ag Fau - Nicholls A, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem*. 2007; 50:74-82.

49. Rush TS, 3rd, Grant JA, Mosyak L, Nicholls A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem*. 2005; 48:1489-95.

50. Muchmore SW, Souers AJ, Akritopoulou-Zanze I. The use of three-dimensional shape and electrostatic similarity searching in the identification of a melanin-concentrating hormone receptor 1 antagonist. *Chem Biol Drug Des*. 2006; 67:174-6.

51. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD. Comparison of topological, shape, and docking methods in virtual screening. *J Chem Inf Model*. 2007; 47:1504-19.

52. Gabel J, Desaphy J, Rognan D. Beware of machine learning-based scoring functions-on the danger of developing black boxes. *J Chem Inf Model*. 2014; 54:2807-15.

53. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, Labonte JW, Pacella MS, Bonneau R, Bradley P, Dunbrack RL, Jr., Das R, Baker D, Kuhlman B, Kortemme T, Gray JJ. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput*. 2017; 13:3031-48.

54. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, Gray JJ, Kortemme T, Richardson JS, Havranek JJ, Snoeyink J, Baker D, Kuhlman B. Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol*. 2013; 523:109-43.

55. Bazzoli A, Kelow SP, Karanicolas J. Enhancements to the Rosetta Energy Function Enable Improved Identification of Small Molecules that Inhibit Protein-Protein Interactions. *PLoS One*. 2015; 10:e0140359.

56. McGann M. FRED Pose Prediction and Virtual Screening Accuracy. *J Chem Inf Model*. 2011; 51:578-96.

57. Gallant SI. Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks*. 1990; 1:179-91.

58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011; 12.

59. Meiler J, Baker D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins*. 2006; 65:538-48.

60. Durrant JD, McCammon JA. BINANA: A Novel Algorithm for Ligand-Binding Characterization. *J Mol Graph Model*. 2011; 29:888-93.

61. ChemAxon (www.chemaxon.com).

62. SZYBKI 1.9.0.3 ed: OpenEye Scientific Software, Santa Fe, NM. .

63. Vogel SM, Bauer MR, Boeckler FM. DEKOIS: demanding evaluation kits for objective in silico screening--a versatile tool for benchmarking docking programs and scoring functions. *J Chem Inf Model*. 2011; 51:2650-65.

64. Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0--a public library of challenging docking benchmark sets. *J Chem Inf Model*. 2013; 53:1447-62.

65. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods*. 2000; 44:235-49.

66. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem*. 2002; 45:2615-23.

67. Sud M. MayaChemTools: An Open Source Package for Computational Drug Discovery. *J Chem Inf Model*. 2016; 56:2292-7.

68. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. 2007; 25:197-206.

69. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res*. 2014; 42:W32-8.

70. Daina A, Michielin O, Zoete V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Res*. 2019; 47:W357-W64.

71. Liu X, Ouyang S, Yu B, Liu Y, Huang K, Gong J, Zheng S, Li Z, Li H, Jiang H. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res*. 2010; 38:W609-14.

72. Wang X, Shen Y, Wang S, Li S, Zhang W, Liu X, Lai L, Pei J, Li H. PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database. *Nucleic Acids Res*. 2017; 45:W356-W60.

73. Mizutani MY, Itai A. Efficient method for high-throughput virtual screening based on flexible docking: discovery of novel acetylcholinesterase inhibitors. *J Med Chem*. 2004; 47:4818-28.

74. Sopkova-de Oliveira Santos J, Lesnard A, Agondanou JH, Dupont N, Godard AM, Stiebing S, Rochais C, Fabis F, Dallemagne P, Bureau R, Rault S. Virtual screening discovery of new acetylcholinesterase inhibitors issued from CERMN chemical library. *J Chem Inf Model*. 2010; 50:422-8.

75. Chen Y, Fang L, Peng S, Liao H, Lehmann J, Zhang Y. Discovery of a novel acetylcholinesterase inhibitor by structure-based virtual screening techniques. *Bioorg Med Chem Lett*. 2012; 22:3181-7.

76. Doytchinova I, Atanasova M, Valkova I, Stavrakov G, Philipova I, Zhivkova Z, Zheleva-Dimitrova D, Konstantinov S, Dimitrov I. Novel hits for acetylcholinesterase inhibition derived by docking-based screening on ZINC database. *J Enzyme Inhib Med Chem*. 2018; 33:768-76.

77. Durrant JD, Carlson KE, Martin TA, Offutt TL, Mayne CG, Katzenellenbogen JA, Amaro RE. Neural-Network Scoring Functions Identify Structurally Novel Estrogen-Receptor Ligands. *J Chem Inf Model*. 2015; 55:1953-61.

78. Wojcikowski M, Zielenkiewicz P, Siedlecki P. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J Cheminform*. 2015; 7:26.

79. Adeshina Y, Karanicolas J. Dataset of Congruent Inhibitors and Decoys (D-COID). *Mendeley Data*. 2019; 1.

80. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000; 28:235-42.

81. Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*. 2004; 20:2153-5.

82. 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. *J Chem Inf Model*. 2010; 50:572-84.

83. Hawkins PC, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. *J Med Chem*. 2007; 50:74-82.

84. Ivanciuc O. Applications of Support Vectors Machines in Chemistry. . *Rev Comput Chem*. 2007; 23:291-400.

85. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot*. 2013; 7:21.

86. XGBoost: A Scalable Tree Boosting System [database on the Internet]2016.

87. Breiman L. Random Forests. *Machine Learning*. 2001; 45:5-32.

88. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Machine Learning*. 2006; 63:3-42.

89. Vapnik V. Statistical Learning Theory. New York: Wiley; 1998.

90. Cheung J, Rudolph MJ, Burshteyn F, Cassidy MS, Gary EN, Love J, Franklin MC, Height JJ. Structures of human acetylcholinesterase in complex with pharmacologically important ligands. *J Med Chem*. 2012; 55:10282-6.

91. Ellman GL, Courtney KD, Andres V, Jr., Feather-Stone RM. A new and rapid colorimetric determination of acetylcholinesterase activity. *Biochem Pharmacol*. 1961; 7:88-95.

92. 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. *Biochem Pharmacol*. 1973; 22:3099-108.

93. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, Davies M, Dedman N, Karlsson A, Magarinos MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017; 45:D945-D54.

94. RDKit: Open-source cheminformatics (www.rdkit.org).

95. Muller-McNicoll M, Neugebauer KM. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nat Rev Genet*. 2013; 14:275-87.

96. 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. *Mol Cell*. 2012; 46:674-90.

97. 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; 149:1393-406.

98. Khalil AM, Rinn JL. RNA-protein interactions in human health and disease. *Semin Cell Dev Biol*. 2011; 22:359-65.

99. Kapeli K, Yeo GW. Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. *Front Neurosci*. 2012; 6:144.

100. 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. *Cell Mol Life Sci*. 2012; 69:501-17.

101. Ellenbecker M, Lanchy JM, Lodmell JS. Identification of Rift Valley fever virus nucleocapsid protein-RNA binding inhibitors using a high-throughput screening assay. *J Biomol Screen*. 2012; 17:1062-70.

102. 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:e91585.

103. Cheng K, Wang X, Yin H. Small-molecule inhibitors of the TLR3/dsRNA complex. *J Am Chem Soc*. 2011; 133:3764-7.

104. Gallego J, Varani G. Targeting RNA with small-molecule drugs: therapeutic promise and chemical challenges. *Acc Chem Res*. 2001; 34:836-43.

105. 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. *Nat Chem Biol*. 2011; 7:553-9.

106. Squires KE. An introduction to nucleoside and nucleotide analogues. *Antivir Ther*. 2001; 6 Suppl 3:1-14.

107. Ewald B, Sampath D, Plunkett W. Nucleoside analogs: molecular mechanisms signaling cell death. *Oncogene*. 2008; 27:6522-37.

108. Cihlar T, Ray AS. Nucleoside and nucleotide HIV reverse transcriptase inhibitors: 25 years after zidovudine. *Antiviral Res*. 2010; 85:39-58.

109. Bitterman PB, Polunovsky VA. Attacking a nexus of the oncogenic circuitry by reversing aberrant eIF4F-mediated translation. *Mol Cancer Ther*. 2012; 11:1051-61.

110. Das K, Arnold E. HIV-1 reverse transcriptase and antiviral drug resistance. Part 1. *Curr Opin Virol*. 2013; 3:111-8.

111. James SH, Prichard MN. Current and future therapies for herpes simplex virus infections: mechanism of action and drug resistance. *Curr Opin Virol*. 2014; 8C:54-61.

112. Menendez-Arias L, Alvarez M, Pacheco B. Nucleoside/nucleotide analog inhibitors of hepatitis B virus polymerase: mechanism of action and resistance. *Curr Opin Virol*. 2014; 8C:1-9.

113. Das K, Arnold E. HIV-1 reverse transcriptase and antiviral drug resistance. Part 2. *Curr Opin Virol*. 2013; 3:119-28.

114. 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 Med Chem*. 2014; 6:45-55.

115. Gowthaman R, Deeds EJ, Karanicolas J. Structural properties of non-traditional drug targets present new challenges for virtual screening. *J Chem Inf Model*. 2013; 53:2073-81.

116. Fauman EB, Rai BK, Huang ES. Structure-based druggability assessment--identifying suitable targets for small molecule therapeutics. *Curr Opin Chem Biol*. 2011; 15:463-8.

117. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. *Science*. 1995; 267:383-6.

118. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. *Proc Natl Acad Sci U S A*. 2004; 101:11287-92.

119. Moreira IS, Fernandes PA, Ramos MJ. Hot spots--a review of the protein-protein interface determinant amino-acid residues. *Proteins*. 2007; 68:803-12.

120. Thanos CD, DeLano WL, Wells JA. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc Natl Acad Sci U S A*. 2006; 103:15422-7.

121. Liu S, Wu S, Jiang S. HIV entry inhibitors targeting gp41: from polypeptides to small-molecule compounds. *Curr Pharm Des*. 2007; 13:143-62.

122. 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. *Nat Chem Biol*. 2010; 6:442-8.

123. Koes DR, Camacho CJ. Small-molecule inhibitor starting points learned from protein-protein interaction inhibitor structure. *Bioinformatics*. 2012; 28:784-91.

124. 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 Enzymol*. 2011; 487:545-74.

125. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. *J Chem Inf Model*. 2012.

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

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

128. ROCS version 3.2.0.3. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

129. Daubner GM, Clery A, Allain FH. RRM-RNA recognition: NMR or crystallography...and new findings. *Curr Opin Struct Biol*. 2013; 23:100-8.

130. Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J*. 2005; 272:2118-31.

131. Auweter SD, Oberstrass FC, Allain FH. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res*. 2006; 34:4943-59.

132. 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*. 1999:2.

133. Tuite JB, Shiels JC, Baranger AM. Substitution of an essential adenine in the U1A-RNA complex with a non-polar isostere. *Nucleic Acids Res*. 2002; 30:5269-75.

134. Benitex Y, Baranger AM. Recognition of essential purines by the U1A protein. *BMC Biochem*. 2007; 8:22.

135. 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 Res*. 2012; 40:3218-31.

136. Okano H, Kawahara H, Toriya M, Nakao K, Shibata S, Imai T. Function of RNA-binding protein Musashi-1 in stem cells. *Exp Cell Res*. 2005; 306:349-56.

137. Spears E, Neufeld KL. Novel double-negative feedback loop between adenomatous polyposis coli and Musashi1 in colon epithelia. *J Biol Chem*. 2011; 286:4946-50.

138. 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:e33431.

139. 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; 34:1-7.

140. 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; 23:3444-53.

141. Seigel GM, Hackam AS, Ganguly A, Mandell LM, Gonzalez-Fernandez F. Human embryonic and neuronal stem cell markers in retinoblastoma. *Mol Vis*. 2007; 13:823-32.

142. 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. *J Neurooncol*. 2008; 86:31-45.

143. 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.

144. 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. *Int J Colorectal Dis*. 2010; 25:17-23.

145. 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. *Mol Cancer*. 2010; 9:221.

146. Todaro M, Francipane MG, Medema JP, Stassi G. Colon cancer stem cells: promise of targeted therapy. *Gastroenterology*. 2010; 138:2151-62.

147. Fox RG, Park FD, Koechlein CS, Kritzik M, Reya T. Musashi Signaling in Stem Cells and Cancer. *Annu Rev Cell Dev Biol*. 2015; 31:249-67.

148. Li N, Yousefi M, Nakauka-Ddamba A, Li F, Vandivier L, Parada K, Woo DH, Wang S, Naqvi AS, Rao S, Tobias J, Cedeno RJ, Minuesa G, Y K, Barlowe TS, Valvezan A, Shankar S, Deering RP, Klein PS, Jensen ST, Kharas MG, Gregory BD, Yu Z, Lengner CJ. The Msi Family of RNA-Binding Proteins Function Redundantly as Intestinal Oncoproteins. *Cell Rep*. 2015; 13:2440-55.

149. 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; 466:765-8.

150. 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. *Nat Med*. 2010; 16:903-8.

151. Cimmperman P, Baranauskienė L, Jachimovičiūtė S, Jachno J, Torresan J, Michailovienė V, Matulienė J, Sereikaitė J, Bumelis V, Matulis D. A quantitative model of thermal stabilization and destabilization of proteins by ligands. *Biophysical journal*. 2008; 95:3222-31.

152. Layton CJ, Hellinga HW. Thermodynamic analysis of ligand-induced changes in protein thermal unfolding applied to high-throughput determination of ligand affinities with extrinsic fluorescent dyes. *Biochemistry*. 2010; 49:10831-41.

153. Morgan CE, Meagher JL, Levengood JD, Delproposto J, Rollins C, Stuckey JA, Tolbert BS. The First Crystal Structure of the UP1 Domain of hnRNP A1 Bound to RNA Reveals a New Look for an Old RNA Binding Protein. *J Mol Biol*. 2015; 427:3241-57.

154. Montemayor EJ, Curran EC, Liao HH, Andrews KL, Treba CN, Butcher SE, Brow DA. Core structure of the U6 small nuclear ribonucleoprotein at 1.7-A resolution. *Nat Struct Mol Biol*. 2014; 21:544-51.

155. Nissen P, Kjeldgaard M, Nyborg J. Macromolecular mimicry. *EMBO J*. 2000; 19:489-95.

156. Tsonis PA, Dwivedi B. Molecular mimicry: structural camouflage of proteins and nucleic acids. *Biochim Biophys Acta*. 2008; 1783:177-87.

157. Della Volpe S, Nasti R, Queirolo M, Unver MY, Jumde VK, Dömling A, Vasile F, Potenza D, Ambrosio FA, Costa G. Novel Compounds Targeting the RNA-Binding Protein HuR. Structure-Based Design, Synthesis, and Interaction Studies. *ACS medicinal chemistry letters*. 2019; 10:615-20.

158. Sakamoto KM. Protacs for treatment of cancer. *Pediatric research*. 2010; 67:505.

159. Neklesa TK, Winkler JD, Crews CM. Targeted protein degradation by PROTACs. *Pharmacology & therapeutics*. 2017; 174:138-44.

160. Fisher SL, Phillips AJ. Targeted protein degradation and the enzymology of degraders. *Current opinion in chemical biology*. 2018; 44:47-55.

161. Gu S, Cui D, Chen X, Xiong X, Zhao Y. PROTACs: an emerging targeting technique for protein degradation in drug discovery. *BioEssays*. 2018; 40:1700247.

162. Pettersson M, Crews CM. PROteolysis TArgeting Chimeras (PROTACs)—past, present and future. *Drug Discovery Today: Technologies*. 2019.

163. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. *Nature*. 2011; 470:163-5.

164. Zheng CJ, Han LY, Yap CW, Ji ZL, Cao ZW, Chen YZ. Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol Rev*. 2006; 58:259-79.

165. Rask-Andersen M, Almen MS, Schioth HB. Trends in the exploitation of novel drug targets. *Nat Rev Drug Discov*. 2011; 10:579-90.

166. Schiff PB, Horwitz SB. Taxol stabilizes microtubules in mouse fibroblast cells. *Proc Natl Acad Sci U S A*. 1980; 77:1561-5.

167. Cukuroglu E, Engin HB, Gursoy A, Keskin O. Hot spots in protein-protein interfaces: towards drug discovery. *Prog Biophys Mol Biol*. 2014; 116:165-73.

168. Shoichet BK. Virtual screening of chemical libraries. *Nature*. 2004; 432:862-5.

169. Macarron R. Critical review of the role of HTS in drug discovery. *Drug Discov Today*. 2006; 11:277-9.

170. Fedorov O, Muller S, Knapp S. The (un)targeted cancer kinome. *Nat Chem Biol*. 2010; 6:166-9.

171. Pagliaro L, Felding J, Audouze K, Nielsen SJ, Terry RB, Krog-Jensen C, Butcher S. Emerging classes of protein-protein interaction inhibitors and new tools for their development. *Curr Opin Chem Biol*. 2004; 8:442-9.

172. Neugebauer A, Hartmann RW, Klein CD. Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods. *J Med Chem*. 2007; 50:4665-8.

173. Sperandio O, Reynes CH, Camproux AC, Villoutreix BO. Rationalizing the chemical space of protein-protein interaction inhibitors. *Drug Discov Today*. 2010; 15:220-9.

174. Czarna A, Beck B, Srivastava S, Popowicz GM, Wolf S, Huang Y, Bista M, Holak TA, Domling A. Robust generation of lead compounds for protein-protein interactions by computational and MCR chemistry: p53/Hdm2 antagonists. *Angew Chem Int Ed Engl*. 2010; 49:5352-6.

175. Reynes C, Host H, Camproux AC, Laconde G, Leroux F, Mazars A, Deprez B, Fahraeus R, Villoutreix BO, Sperandio O. Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods. *PLoS Comput Biol*. 2010; 6:e1000695.

176. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008; 456:470-6.

177. Lukong KE, Chang KW, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. *Trends Genet*. 2008; 24:416-25.

178. Lan L, Appelman C, Smith AR, Yu J, Larsen S, Marquez RT, Liu H, Wu X, Gao P, Roy A, Anbanandam A, Gowthaman R, Karanicolas J, De Guzman RN, Rogers S, Aube J, Ji M, Cohen RS, Neufeld KL, Xu L. Natural product (-)-gossypol inhibits colon cancer cell growth by targeting RNA-binding protein Musashi-1. *Mol Oncol*. 2015; 9:1406-20.

179. James SH, Prichard MN. Current and future therapies for herpes simplex virus infections: mechanism of action and drug resistance. *Curr Opin Virol*. 2014; 8:54-61.

180. Menendez-Arias L, Alvarez M, Pacheco B. Nucleoside/nucleotide analog inhibitors of hepatitis B virus polymerase: mechanism of action and resistance. *Curr Opin Virol*. 2014; 8:1-9.

181. Della Volpe S, Nasti R, Queirolo M, Unver MY, Jumde VK, Domling A, Vasile F, Potenza D, Ambrosio FA, Costa G, Alcaro S, Zucal C, Provenzani A, Di Giacomo M, Rossi D, Hirsch AKH, Collina S. Novel Compounds Targeting the RNA-Binding Protein HuR. Structure-Based Design, Synthesis, and Interaction Studies. *ACS Med Chem Lett*. 2019; 10:615-20.

182. Chen H, Zhou X, Wang A, Zheng Y, Gao Y, Zhou J. Evolutions in fragment-based drug design: the deconstruction-reconstruction approach. *Drug Discov Today*. 2015; 20:105-13.

183. Bai NA, Y.; Lan, L.; Makhov, Petr P. B., Xia, Y., et al. Rationally designing inhibitors of the Musashi protein-RNA interaction by hotspot mimicry. (Manuscript submitted for publication). In: Kansas Uo, editor.2019.

184. Brown DG, Bostrom J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J Med Chem*. 2016; 59:4443-58.

185. Reactor was used for enumeration and reaction modeling. JChem 15.12.14 ed: ChemAxon; 2015.

186. Liang PI, Li WM, Wang YH, Wu TF, Wu WR, Liao AC, Shen KH, Wei YC, Hsing CH, Shiue YL, Huang HY, Hsu HP, Chen LT, Lin CY, Tai C, Lin CM, Li CF. HuR cytoplasmic expression is associated with increased cyclin A expression and poor outcome with upper urinary tract urothelial carcinoma. *BMC Cancer*. 2012; 12:611.

187. Minuesa G, Antczak C, Shum D, Radu C, Bhinder B, Li Y, Djaballah H, Kharas MG. A 1536-well fluorescence polarization assay to screen for modulators of the MUSASHI family of RNA-binding proteins. *Comb Chem High Throughput Screen*. 2014; 17:596-609.

188. Wu X, Lan L, Wilson DM, Marquez RT, Tsao WC, Gao P, Roy A, Turner BA, McDonald P, Tunge JA, Rogers SA, Dixon DA, Aube J, Xu L. Identification and validation of novel small molecule disruptors of HuR-mRNA interaction. *ACS Chem Biol*. 2015; 10:1476-84.

189. Lim D, Byun WG, Koo JY, Park H, Park SB. Discovery of a Small-Molecule Inhibitor of Protein-MicroRNA Interaction Using Binding Assay with a Site-Specifically Labeled Lin28. *J Am Chem Soc*. 2016; 138:13630-8.

190. Wang L, Rowe RG, Jaimes A, Yu C, Nam Y, Pearson DS, Zhang J, Xie X, Marion W, Heffron GJ, Daley GQ, Sliz P. Small-Molecule Inhibitors Disrupt let-7 Oligouridylation and Release the Selective Blockade of let-7 Processing by LIN28. *Cell Rep*. 2018; 23:3091-101.

191. Minuesa G, Albanese SK, Xie W, Kazansky Y, Worroll D, Chow A, Schurer A, Park SM, Rotsides CZ, Taggart J, Rizzi A, Naden LN, Chou T, Gourkanti S, Cappel D, Passarelli MC, Fairchild L, Adura C, Glickman JF, Schulman J, Famulare C, Patel M, Eibl JK, Ross GM, Bhattacharya S, Tan DS, Leslie CS, Beuming T, Patel DJ, Goldgur Y, Chodera JD, Kharas MG. Small-molecule targeting of MUSASHI RNA-binding activity in acute myeloid leukemia. *Nat Commun*. 2019; 10:2691.

192. Paiva AM, Vanderwall DE, Blanchard JS, Kozarich JW, Williamson JM, Kelly TM. Inhibitors of dihydrodipicolinate reductase, a key enzyme of the diaminopimelate pathway of Mycobacterium tuberculosis. *Biochim Biophys Acta*. 2001; 1545:67-77.

193. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem*. 2002; 45:2213-21.

194. 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:e32839.

195. Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D. PRIDB: a Protein-RNA interface database. *Nucleic Acids Res*. 2011; 39:D277-82.

196. Lazaridis T, Karplus M. Effective energy function for proteins in solution. *Proteins*. 1999; 35:133-52.

197. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

198. 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. *Anal Biochem*. 2004; 332:261-73.

199. Niesen FH, Berglund H, Vedadi M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat Protoc*. 2007; 2:2212-21.

200. ESPript - http://espript.ibcp.fr.

201. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res*. 2014; 42:W320-4.