Designing Small Molecule Inhibitors of RNA-Binding Protein Musashi Using New Biochemical and Computational Approaches

By

Nan Bai

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

Chairperson: Dr. Liang Xu

Co-Chairperson: Dr. John Karanicolas

Dr. Michael F. Rafferty

Dr. Roberto N. De Guzman

Dr. Joanna Slusky

Dr. William Picking

Date Defended: December 19th, 2019

The dissertation committee for Nan Bai certifies that this is the approved version of the following dissertation:

Designing Small Molecule Inhibitors of RNA-Binding Protein Musashi Using New Biochemical and Computational Approaches

____

Chairperson Dr. Liang Xu

Date Approved: December 19th, 2019

ABSTRACT

RNA-binding proteins (RBPs) are key regulators of post-transcriptional gene expression, and underlie many disease-relevant processes. However, they have historically been challenging to target with drug-like compounds. Inspired by the "anchor residues" of protein-protein interactions, we developed a computational approach for rationally designing small-molecule inhibitors of RBPs. In this dissertation, we first selected Musashi-1 and Musashi-2 to apply our "RNA mimicry" approach. Both Musashi proteins are well-studied RBPs, known principally as stem-cell markers that are upregulated in many cancers. In the future, we hope our "RNA mimicry" approach can be generally applied to inhibitor design of diverse target RBPs.

To design inhibitors of Musashi proteins, we applied our strategy by mimicking the three-dimensional interactions in the protein-RNA complex. As described in Chapter II, by using pharmacophoric screening, we searched for drug-like compounds that can present the same geometric arrangement of functional groups as the RNA in the complex. We hypothesized that such ligands would engage Musashi in a similar manner as the RNA binds to Musashi. Since the interaction geometries can be quite distinct from one another for different RBPs, we anticipated that this strategy would lead to inhibitors that were selective for Musashi.

To facilitate characterization of these candidate Musashi inhibitors, I developed the "isothermal analysis" approach. As described in Chapter III, this method allows us to calculate quantitative binding constants by using differential scanning fluorimetry (DSF) data. The method requires only the protein unfolding information at a given temperature as a function of ligand concentration, and thus no thermodynamic parameters are included in the calculation.

Finally, I describe the use of computational docking to better understand the basis for PROTAC-mediated degradation of target proteins. PROteolysis TArgeting Chimeras (PROTACs) are heterobifunctional small molecules which can induce target protein degradation through cell ubiquitination process. Rational design of PROTACs is still challenging, however, because of the limited structural

understanding of their mechanism. In Chapter IV, I seek to predict the formation of the ternary structure complex by including both effects of the protein-protein interaction and effects of the chemical linker. Looking ahead, I hope to use these ternary structure models to explain the activity and selectivity of the given PROTAC molecules, and ultimately to use our designed Musashi inhibitors as a starting point for building new PROTACs to degrade Musashi.

The text of Chapter II is a manuscript that is in preparation for publication as:

> Bai N[‡], Adeshina Y[‡], Lan L, Makhov PB, Xia Y, Gowthaman R, Miller SA, Johnson DK, Boumber Y, Xu L, Karanicolas J. Rationally designing inhibitors of the Musashi protein-RNA interaction by hotspot mimicry.
> [‡]*equally contributing co-authors*
>
> The supporting information for this chapter is included as Appendix A.1.

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

> Bai N, Roder H, Dickson A, Karanicolas J. Isothermal analysis of ThermoFluor data can readily provide quantitative binding affinities. *Sci. Rep.* 9, p. 2650 (2019).
> Note: the software disseminated with this paper has accumulated >1000 downloads in the 9 months since publication
>
> The supporting information for this chapter is included as Appendix A.2.

The text of Chapter IV is a manuscript that is in preparation for publication as:

> Bai N, Karanicolas J. Predicting PROTAC-mediated ternary complex formation using Rosetta.
>
> The supporting information for this chapter is included as Appendix A.3.

ACKNOWLEDGEMENTS

I would like to offer my greatest gratitude to my mentor, Dr. John Karanicolas for his support, guidance, enthusiasm, and patience during my PhD study.

I would like to thank all the lab members, past and present, for their great help and valuable discussion: Dr. David Johnson, Dr. Jittasak Khowsathit, Dr. Sven Miller, Dr. Kirubakaran Palani, Dr. Ragul Gowthaman, Dr. Karen Khar, Dr. Daniel Pushparaju Yeggoni, Dr. Jimmy Budiardjo, Dr. Andrea Bazzoli, Dr. Yan Xia, Yusuf Adeshina, Grigorii Andrianov, Lei Ke, Chris Parry and Jin Niu. Special thanks to Dr. Shipra Malhotra for our sweet friendship and her warm, kindly and cheerful support.

I am grateful to all my committee members: Dr. Liang Xu, Dr. Michael F. Rafferty, Dr. Roberto De Guzman, Dr. Joanna Slusky, Dr. William Picking and Dr. Eric Deeds, for their helpful guidance, and always understanding.

Finally, I wish to thank my parents, my mom Caihong Sun and my dad Guangsi Bai for their love and support. My deepest appreciation to my husband, Yangjie Wei, for everything. Every second we spend together is my most valuable treasure, always.

Table of Contents

Chapter I: Introduction

More than 1500 human RNA-binding proteins (RBPs) have been discovered and studied to varying degrees (1-5). Many of these RBPs have been confirmed as key players in important, diverse biological processes by interacting with different types of target RNAs, including both double-stranded RNAs (6) and single-stranded RNAs (7). Mis-regulation of these RBPs has been implicated as related to multiple serious diseases, and in some cases this makes them potential therapeutic targets (8-10).

Compared with their variety of functions and their diversity of cognate RNAs, the structures of RBPs are actually surprisingly conserved and modular (11, 12). Most RBPs are built from only a few functional domains, for example, RNA-recognition motif (RRM), hnRNP K homology domain (KH), and Zinc-Finger domain (ZnF). RRMs are the most abundant RNA-binding domains in humans, and has become a key focus of attention in the past decades (7, 13-17). RRMs are comprised of 80-90 amino acids, folded into a β-α-β-β-α-β topology. On the sequence level, there are two highly conserved short sequences: RNP1 on β-sheet 3 and RNP2 on β-sheet 1. In most cases, these two RNPs contain several conserved residues: an Arg or Lys, and two or three aromatic residues (11, 12, 17). The aromatic residues often form π stacking interactions with cognate RNA bases, and it has been shown that mutations to these aromatic sidechains disrupt binding to cognate RNA (7, 14).

Mammalian Musashi (Msi) proteins are RBPs with two homologs: Musashi-1 (Msi1) and Musashi-2 (Msi2). Each of them contains two RRMs, and these two RRMs are about 75% sequence identical (18). Their function was originally identified in the context of stem cell proliferation and differentiation, by binding to target mRNAs and regulating their stability and translation (18-22). Over-expression of Msi1 and Msi2 has been linked to multiple serious diseases, including Alzheimer's disease (23) and several certain cancers (24-34). Their rich biology makes both Msi1 and Msi2 attractive therapeutic targets, and already several independent groups have described small molecule inhibitors targeting Musashi proteins

(35-38). Each of these groups began with similar fluorescence polarization (FP) competition assays to screen their own distinct small molecule libraries, and ultimately each group arrived at distinctive inhibitors.

Although each group discovered a potentially promising inhibitor, the size of the libraries they used and the biochemical assay they applied may have limited their outcomes from their screens. As a more robust and general approach, I instead developed a computational method for large-scale *in silico* screening against Musashi. I also developed a comprehensive biochemical assay for testing potential inhibitors by introducing a new analysis framework for the differential scanning fluorimetry (DSF) assay, a commonly-used assay in drug discovery. Finally, to enhance the efficacy of the Musashi inhibitors, I began to develop a computational approach that will allow me to rationally design new PROteolysis TArgeting Chimeras (PROTACs) predicated on using these inhibitors as warheads for inducing Musashi degradation in cells.

RNA Mimicry Approach

Musashi proteins, like many other RBPs, have been considered “non-traditional” targets for therapeutic intervention because of their relatively flat binding surface (18, 38). Inspired by rational design of inhibitors against protein-protein interactions (39-45), another class of “non-traditional” target, we elected to adapt this “mimicry” idea for RBPs.

To mimic the three-dimensional interaction of Msi-RNA, we start with identifying the essential chemical moieties of RNA fragment from Msi-RNA complex structure. We define these essential functional groups as “hotspot pharmacophores”, which is similar to the concept of hotspots in protein-protein interactions (41, 43, 44). The “hotspot pharmacophore” will then be applied as a template for *in silico* pharmacophore-based screening. Our hypothesis is that small molecules which have similar structures to the “hotspot pharmacophore” extracted from the protein-RNA complex would mimic the interaction of the RNA with the protein, and thus competitively inhibit the protein-RNA interaction.

Importantly, this underlying conceptual framework can also allow for consideration of selectivity of the resulting inhibitors. Because RBPs are so plentiful in humans (2, 3) and their structures are so highly conserved (11, 14, 17), it is critical to incorporate selectivity as much as possible in the design process. In our approach, we therefore have multiple steps of at which we consider selectivity. During the initial computational design, the "hotspot pharmacophore" not only includes the identities of the polar and stacking interactions but also include the geometries of these interactions: our hypothesis is that this approach will lead to selective inhibitors of our target RBP, by virtue of mimicking the structure of the (bound) cognate RNA and not simply mimicking the chemical structure of the bases. Second, we also built a "hotspot pharmacophore library" using available RBP-RNA complex structures from the PDB, and we screened our candidate inhibitors against this library. If we find that one of our candidate inhibitors matches an off-target hotspot pharmacophore with a similar score as it matches the Musashi hotspot pharmacophore, this off-target RBP is prioritized for explicit testing in biochemical assays. Finally, we optimized our Musashi inhibitors with several rounds of structure activity relationship (SAR) studies to improve on-target potency and also the selectivity.

While we have developed and used this pipeline to identify new Musashi inhibitors, we expect that in the future the same approach can be used for identifying inhibitors of other biologically important RBPs as well.

Isothermal Analysis of ThermoFluor Data

As we began to identify new Musashi inhibitors and test these in biochemical assays, we sought a complementary and efficient biophysical assay to validate our initial promising hit compounds. Differential scanning fluorimetry (DSF), also known as ThermoFluor or Thermal Shift Assay, can detect ligand-protein interactions by measuring the shift of the protein melting temperature (46, 47). This method meets our criteria because of its many advantages, including that it is label-free, is low cost, has low sample requirement, and is fast to deploy (48). With all these benefits, DSF has been frequently applied in drug discovery for ligands screening (49-53). Although the protocol for a typical DSF assay is by now well-developed, it is challenging to calculate accurate binding affinities from the resulting data. Because of the complicated underlying thermodynamic process involved, it is typical to either use a rigorous but difficult "thermodynamic model" to calculate binding affinity (46, 52, 54-56), or use a so-called "Boltzmann model" to estimate the shift of the melting temperature resulting from ligand-protein binding but not try to obtain a binding constant (57-60). Overall, DSF is more popular as a qualitative approach rather than a quantitative method.

Motivated by the opportunity to provide a straightforward method for obtaining binding constants, we sought to develop a novel analysis isothermal approach that would allow us to fit DSF data and readily achieve accurate binding constants. The basic idea of our isothermal analysis is to avoid the complexity of fitting to underlying thermodynamic process of heating a given sample. Instead of calculating binding constant from the whole melting curve with the traditional "thermodynamic model", we select a single temperature which should be in the range of the protein's unfolding transition, and we measure the change of the unfolded protein fraction along with the ligand titration at this specific temperature. We then calculate the binding constant by fitting the ligand concentrations and unfolded protein fractions into a protein folding-unfolding reaction as a competitive coupled equilibrium with ligand binding. Because all the unfolded protein fractions are collected at the same temperature, no thermodynamic parameters would be required in the fitting.

Combining all the well-known advantages of DSF and our straightforward isothermal analysis approach, we now have in place a well-developed and optimized biophysical assay for validating our Musashi inhibitors. In the future, we expect that this DSF assay coupled with our isothermal analysis will also prove useful in efficiently testing inhibitors designed against other target proteins as well.

PROTAC Prediction

So far, the best compounds emerging from our Msi-RNA inhibitor design have exhibited single-digit micromolar binding constants, and it has proven difficult to improve them further: this is probably a fundamental limitation of the relatively flat binding site available on the protein surface. Inspired by the emergence and popularity of the PROTAC concept, we were inspired to use our top Musashi inhibitors as a starting point for developing Musashi-degrading PROTACs.

PROteolysis TArgeting Chimeras (PROTACs) are heterobifunctional small molecules designed to recruit an E3 ubiquitin ligase to degrade some protein of interest (POI). A PROTAC molecule contains three parts: a warhead ligand which binds the POI, an E3 ligase ligand which recruits the E3 ligase, and a chemical linker that joins these two parts. When a PROTAC simultaneously recruits the POI and the E3 ligase, this can induce ubiquitination and subsequent degradation of the POI. Because of their unique active mechanism, PROTAC molecules have multiple advantages over traditional small molecule drugs, including longer cellular lifetime (61); no specific requirement of binding to the active site (62-64); addressing "non-traditional" targets (65-67); and low requirements for the binding affinity (61, 68, 69).

Although PROTACs have been successfully designed to selectively degrade many broad and diverse classes of target proteins (70-85), PROTACs are still typically designed by trial-and-error: once the substrate-binding warhead and the E3-recruiting moieties have been selected, they are attached using tens or hundreds of different linkers to determine which one(s) yield efficient degradation. Thus, the field as a whole has not even begun to reach the point where PROTACs can be rationally designed.

As a first step towards this goal, we propose to develop a computational approach for modeling the ensemble of ternary complexes that can be formed by a given POI / PROTAC / E3 ligase. This method involves protein-protein docking to identify complementary binding modes, followed by screening of low-energy linker conformations to determine which docked binding modes are accessible to a given PROTAC linker. By modeling the ternary complexes in this manner, we seek to predict the activity and selectivity of a given PROTAC. To test this, we benchmarked our method using a variety of PROTACs reported in the literature.

Having developed this new approach for predicting PROTAC ternary complexes, we are now eager to deploy it for building Musashi-degrading PROTACs. While the warheads developed in my earlier studies lack the potency to serve as useful chemical probes for *in vivo* applications, we expect that developing these into PROTACs may help overcome this limitation. Further, the fact that these inhibitors already appear to be selective for Musashi over other RRM proteins suggests that their selectivity may be retained once we elaborate them into the PROTAC format.

CONTRIBUTIONS BY COLLABORATORS

The RNA mimicry approach in Chapter II was developed in part by Ragul Gowthaman and Yan Xia. The pharmacophore library and MDS plot in Chapter II was generated by David Johnson. The NMR data in Chapter II were collected by Sven Miller.

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

Abstract

RNA-binding proteins (RBPs) are key regulators of post-transcriptional gene expression, and underlie many important biological processes. Here, we develop a strategy that entails extracting a “hotspot pharmacophore” from the structure of a protein-RNA complex, and using this as a template for designing small-molecule inhibitors. With this approach we first target Musashi-1, 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.

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 (4). Mammalian proteomes are thought to include upwards of 800 RBPs (1, 5), 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 (8-10).

To date, there exist few classes of compounds that target protein-RNA interactions. Inhibitors of certain RBPs have been identified via high throughput screening (86, 87), including one series from virtual screening that competes with double-stranded RNA for binding to toll-like receptor 3 (88), and a number of compounds have been reported that disrupt binding by interacting with the RNA rather than with the RBP (89, 90). 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 (91-97), 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 (96, 98, 99), 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 have 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 RBPs 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 (100), 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 (101).

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 (40, 43, 44). This observation, in turn, motivated several groups to mimic these key interactions when designing small-molecule inhibitors (39, 41, 42, 45). 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.

Computational Approach

New computational methods have been implemented in the Rosetta software suite (102) 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.

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 (40, 43). 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 cluster neighboring moieties, and advance each cluster separately. Through this approach, we reduce the structure of the protein-RNA complex to a minimal "hotspot pharmacophore" that encapsulates the key interactions to be recapitulated by a small molecule (**Figure 1c**).

Identifying complementary ligands

To identify such compounds, we use 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 restrict our search to the ~7 million compounds in the ZINC database (103) that are both commercially available, and predict to have drug-like physicochemical properties. We use OMEGA

(OpenEye Scientific Software, Santa Fe, NM) (104-106) to build low-energy conformations of each compound, then ROCS (OpenEye Scientific Software, Santa Fe, NM) (73, 107) to align each conformation to our hotspot pharmacophore. For each of the top-scoring hits emerging from ROCS, we then use the aligned orientation to position the compound relative to the protein, and evaluate the interaction energy of the protein-ligand complex using the fullatom Rosetta energy function (102).

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 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 (108). 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 (14). Accordingly, mutations to the protein that remove these aromatic sidechains have been shown to disrupt binding in representative RRMs (14, 109), as has introduction of non-canonical bases to the RNA that alters the pattern of hydrogen bonding groups (13, 15, 16). 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 (14).

Mammalian Musashi-1 (Msi1) recognizes its cognate RNAs through a pair of RRMs, RRM1 and RRM2 (110). 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 (111, 112). *NUMB* mRNA encodes an inhibitor of Notch, so translational inhibition by Msi1 triggers Notch signaling and thus promotes self-renewal and cell survival (112, 113). 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 (24, 28-30, 32-34). 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 (19, 27, 31).

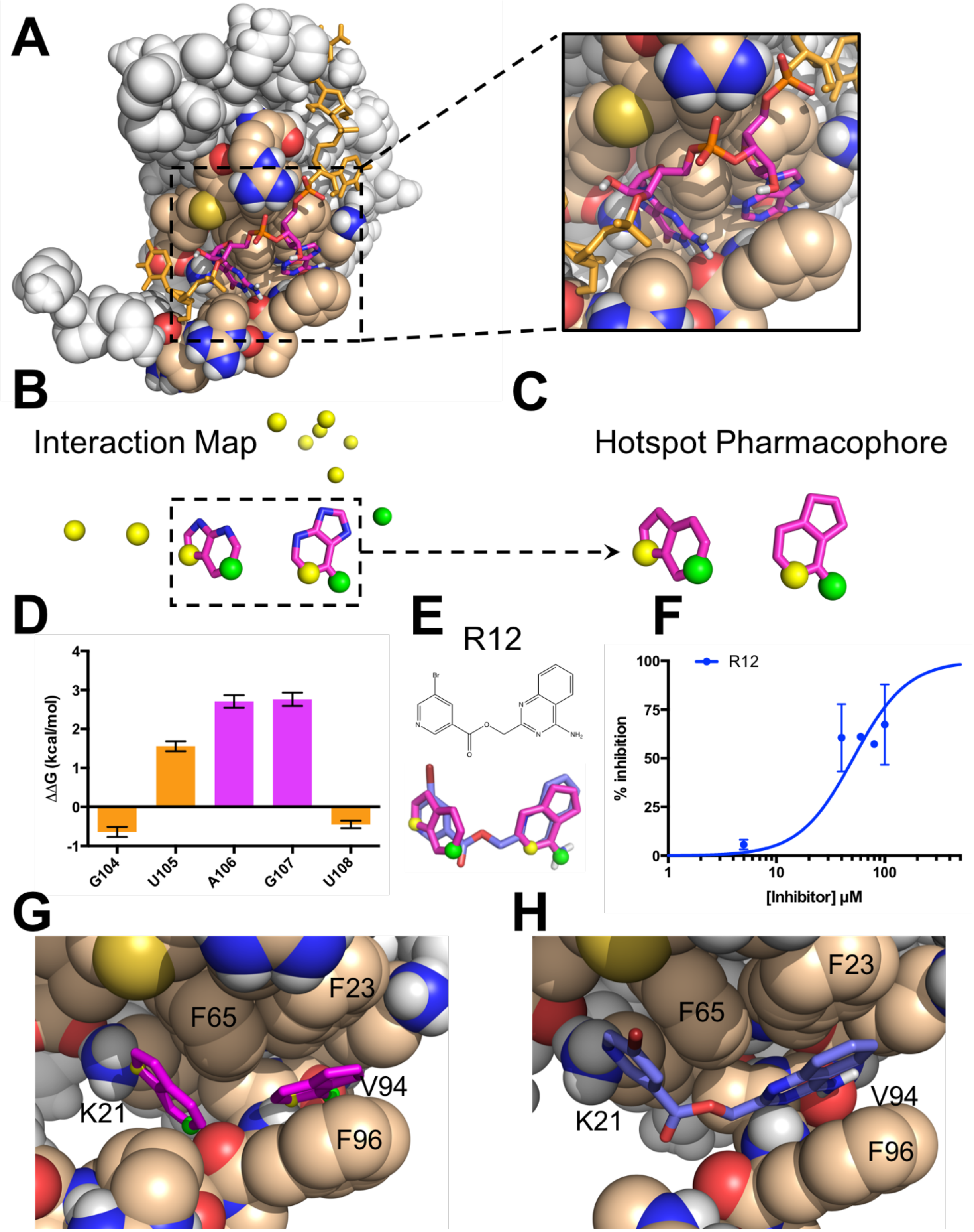

A
B
Interaction Map
C
Hotspot Pharmacophore
D
ΔΔG (kcal/mol)
4
3
2
1
0
-1
G104
U105
A106
G107
U108
E
R12
Br
N
O
O
N
N
NH2
F
R12
100
75
50
25
0
% inhibition
1
10
100
[Inhibitor] μM
G
F65
F23
K21
V94
F96
H
F23
F65
K21
V94
F96

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.

Results

Computational screening against Msi1 RRM1

We applied our "hotspot mimicry" approach to the Musashi-1 RRM1 / *NUMB* mRNA complex (110), 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 does 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 affected 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**).

SAR study of R12 derivatives

With R12 as a starting point, and guided by our structural model (**Figure 2a**), we set out to improve potency of this interaction. While our initial screening 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 than R12. As a secondary validation assay we used Differential Scanning Fluorimetry (DSF) to confirm the interaction, and found that R12-7 induced inconsistent changes in the protein's melting temperature (**Figure S3a**), but R12-8 led to consistently 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 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 closest 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.

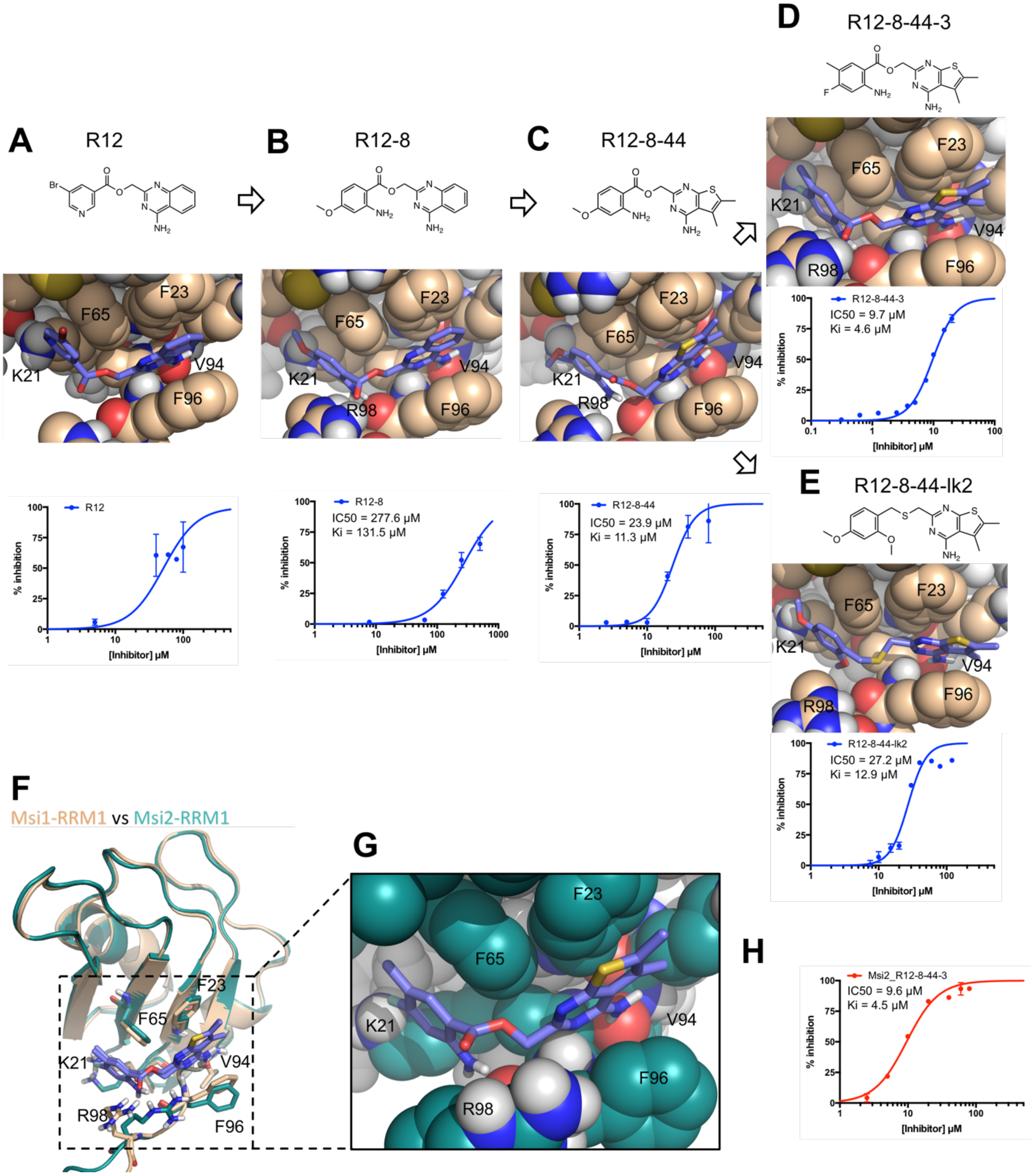

A
R12
K21
F65
F23
V94
F96
R12
% inhibition
[Inhibitor] µM
B
R12-8
F65
F23
K21
R98
V94
F96
R12-8
IC50 = 277.6 µM
Ki = 131.5 µM
% inhibition
[Inhibitor] µM
C
R12-8-44
F23
F65
K21
R98
V94
F96
R12-8-44
IC50 = 23.9 µM
Ki = 11.3 µM
% inhibition
[Inhibitor] µM
D
R12-8-44-3
F23
F65
K21
V94
R98
F96
R12-8-44-3
IC50 = 9.7 µM
Ki = 4.6 µM
% inhibition
[Inhibitor] µM
E
R12-8-44-lk2
F23
F65
K21
V94
R98
F96
R12-8-44-lk2
IC50 = 27.2 µM
Ki = 12.9 µM
% inhibition
[Inhibitor] µM
F
Msi1-RRM1 vs Msi2-RRM1
F23
F65
K21
V94
R98
F96
G
F23
F65
K21
V94
R98
F96
H
Msi2_R12-8-44-3
IC50 = 9.6 µM
Ki = 4.5 µM
% inhibition
[Inhibitor] µM

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.

Inhibition of Musashi-2

While expression of Msi1 is tissue-restricted, its homolog Msi2 is ubiquitously expressed (25, 26). 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 (27). 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.

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: additional 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 (114, 115). 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.

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 (110). 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.

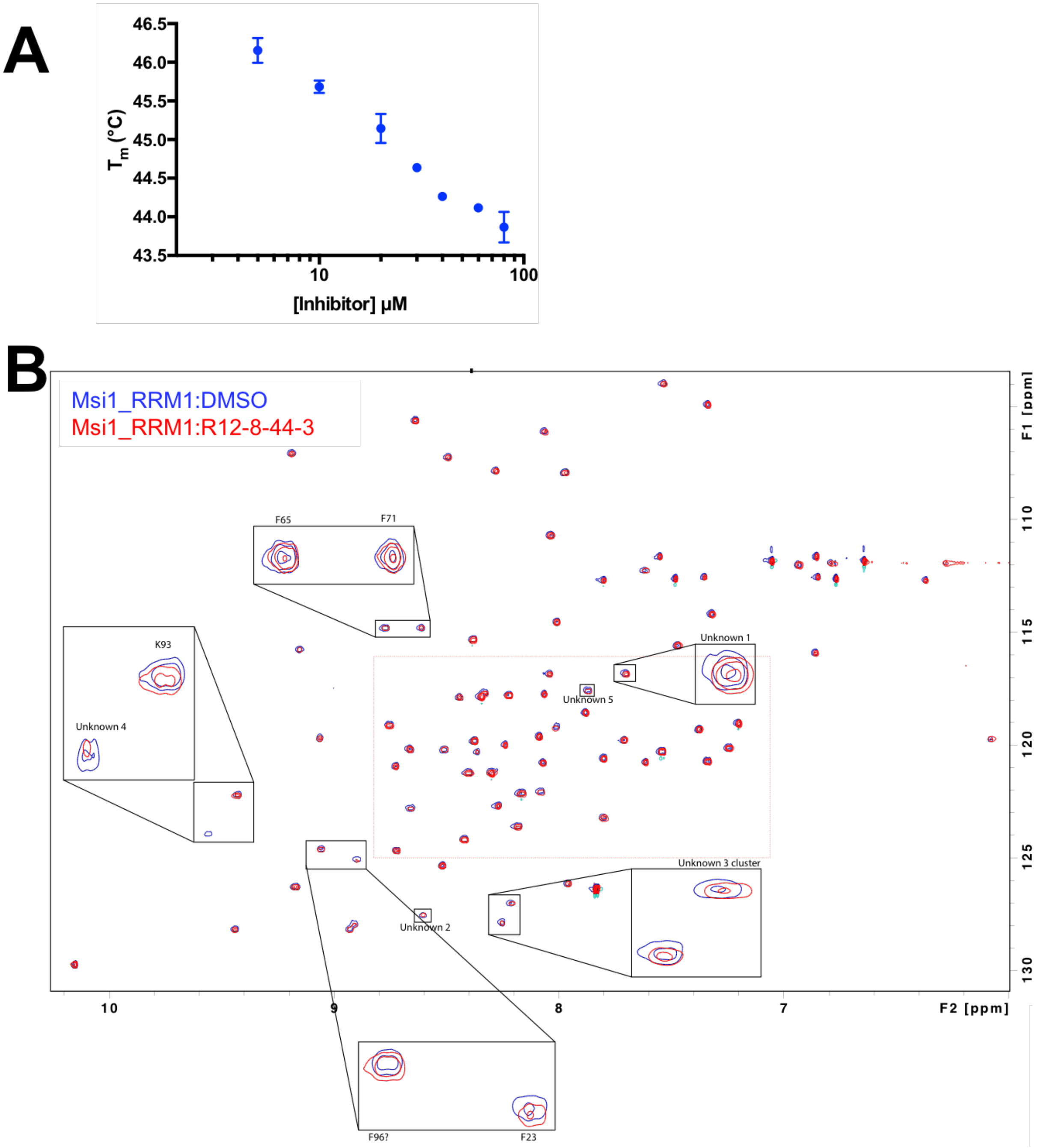


Figure 3: Biophysical characterization of R12-8-44-3. (A) Differential scanning fluorimetry shows that addition of R12-8-44-3 decreases the melting temperature of Msi1 in a concentration-dependent manner. **(B)** HSQC spectrum of Msi1 RRM1 collected in the presence and absence of R12-8-44-3. Peaks showing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

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 (14). 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) (116), and yeast Prp24 (117).

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) but not Prp24 (a score of 0.747), 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.

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 hnRNP A1's RNA binding. 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 #7) 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.

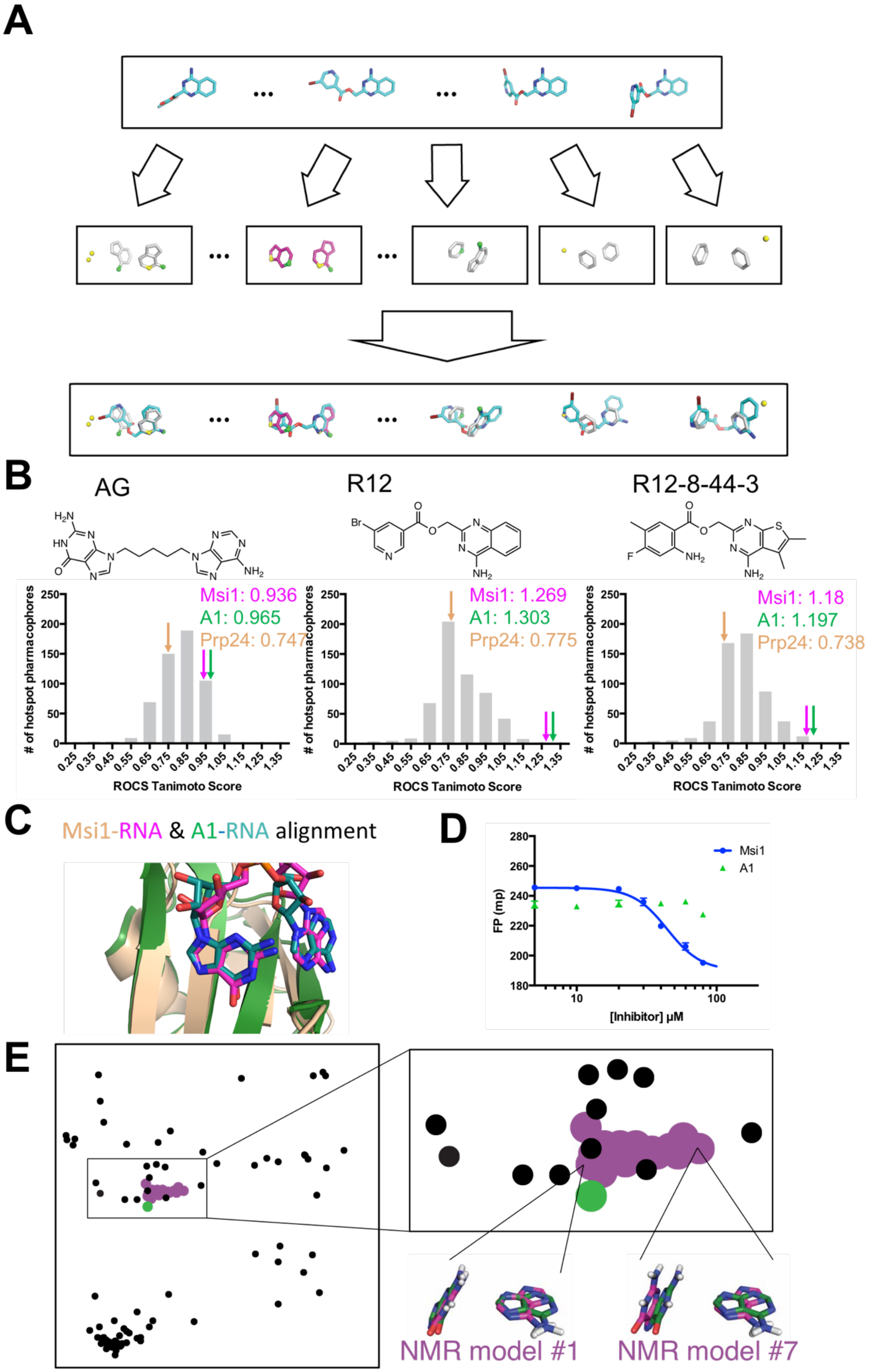

A
B
AG
R12
R12-8-44-3
H_2N
HN
N
O
N
N
N
N
NH_2
Br
O
O
N
N
N
NH_2
O
O
F
NH_2
N
N
S
NH_2
Msi1: 0.936
A1: 0.965
Prp24: 0.747
Msi1: 1.269
A1: 1.303
Prp24: 0.775
Msi1: 1.18
A1: 1.197
Prp24: 0.738
# of hotspot pharmacophores
250
200
150
100
50
0
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
1.05
1.15
1.25
1.35
ROCS Tanimoto Score
C
Msi1-RNA & A1-RNA alignment
D
FP (mp)
280
260
240
220
200
180
10
100
[Inhibitor] µM
Msi1
A1
E
NMR model #1
NMR model #7

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*), hnRNP A1 (*green arrow*) and Prp24 (*yellow 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.

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 (118, 119); 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 (120). 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 selecting 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) (62-64, 121, 122) 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.

Methods

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

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 III: Isothermal Analysis of ThermoFluor Data can readily provide Quantitative Binding Affinities

Abstract

Differential scanning fluorimetry (DSF), also known as ThermoFluor or Thermal Shift Assay, has become a commonly-used approach for detecting protein-ligand interactions, particularly in the context of fragment screening. Upon binding to a folded protein, most ligands stabilize the protein; thus, observing an increase in the temperature at which the protein unfolds as a function of ligand concentration can serve as evidence of a direct interaction. While experimental protocols for this assay are well-developed, it is not straightforward to extract binding constants from the resulting data. Because of this, DSF is often used to probe for an interaction, but not to quantify the corresponding binding constant (K_d). Here, we propose a new approach for analyzing DSF data. Using unfolding curves at varying ligand concentrations, our "isothermal" approach collects from these the fraction of protein that is folded at a single temperature (chosen to be temperature near the unfolding transition). This greatly simplifies the subsequent analysis, because it circumvents the complicating temperature dependence of the binding constant; the resulting constant-temperature system can then be described as pair of coupled equilibria (protein folding/unfolding and ligand binding/unbinding). The temperature at which the binding constants are determined can also be tuned, by adding chemical denaturants that shift the protein unfolding temperature. We demonstrate the application of this isothermal analysis using experimental data for maltose binding protein binding to maltose, and for two carbonic anhydrase isoforms binding to each of four inhibitors. To facilitate adoption of this new approach, we provide a free and easy-to-use Python program that analyzes thermal unfolding data and implements the isothermal approach described herein (https://sourceforge.net/projects/dsf-fitting).

Introduction

Differential scanning fluorimetry (DSF), also known as ThermoFluor or Thermal Shift Assay, has become an important label-free technique for biophysical ligand screening and protein engineering (123-127). Briefly, this method makes use of a dye – typically either SYPRO Orange or 1-anilino-8-naphthalenesulfonate (ANS) – that is quenched in an aqueous environment but becomes strongly fluorescent when bound to exposed hydrophobic groups of a protein. By heating one's protein of interest in the presence of such a dye, the thermal unfolding transition can be monitored spectrophotometrically. Because ligands that interact with proteins typically stabilize the folded protein, this leads to a shift in the midpoint of the unfolding transition (i.e. the melting temperature, T_m) (128, 129).

The simplicity of this assay makes DSF very straightforward to implement using an RT-PCR thermocycler, it can be inexpensive and fast, and it requires relatively little sample (48). These advantages have made this approach attractive for screening applications in drug discovery – particularly for moderately-sized fragment libraries (50, 123, 124) – and also for protein stability formulation (51, 130). Meanwhile, the fact that this method is label-free and well-suited to detect binding over a wide range of affinities has made DSF one of the most popular approaches in drug discovery for fragment screening (128, 131-134) and for evaluating the "ligandability" of a target protein (135). While it would be desirable to obtain binding constants at an early stage, for example to prioritize fragment hits on the basis of their ligand efficiency (136), the magnitudes of the observed T_m-shifts (at a given ligand concentration) have been shown to correlate only weakly with compounds' potency measured in other orthogonal assays (137).

Typical DSF data are shown in **Figure 1a**. Here, SYPRO dye is used as a reporter for the extent of unfolding of maltose binding protein (MBP), and the melting temperature from each curve is determined. Using this method, MBP is observed to have a T_m of approximately 52.5 ºC in the absence of its ligand, maltose. Upon addition of increasing concentrations of maltose, the unfolding transition is shifted to increasingly higher temperatures: this implies that maltose stabilizes MBP, by binding to the natively folded protein.

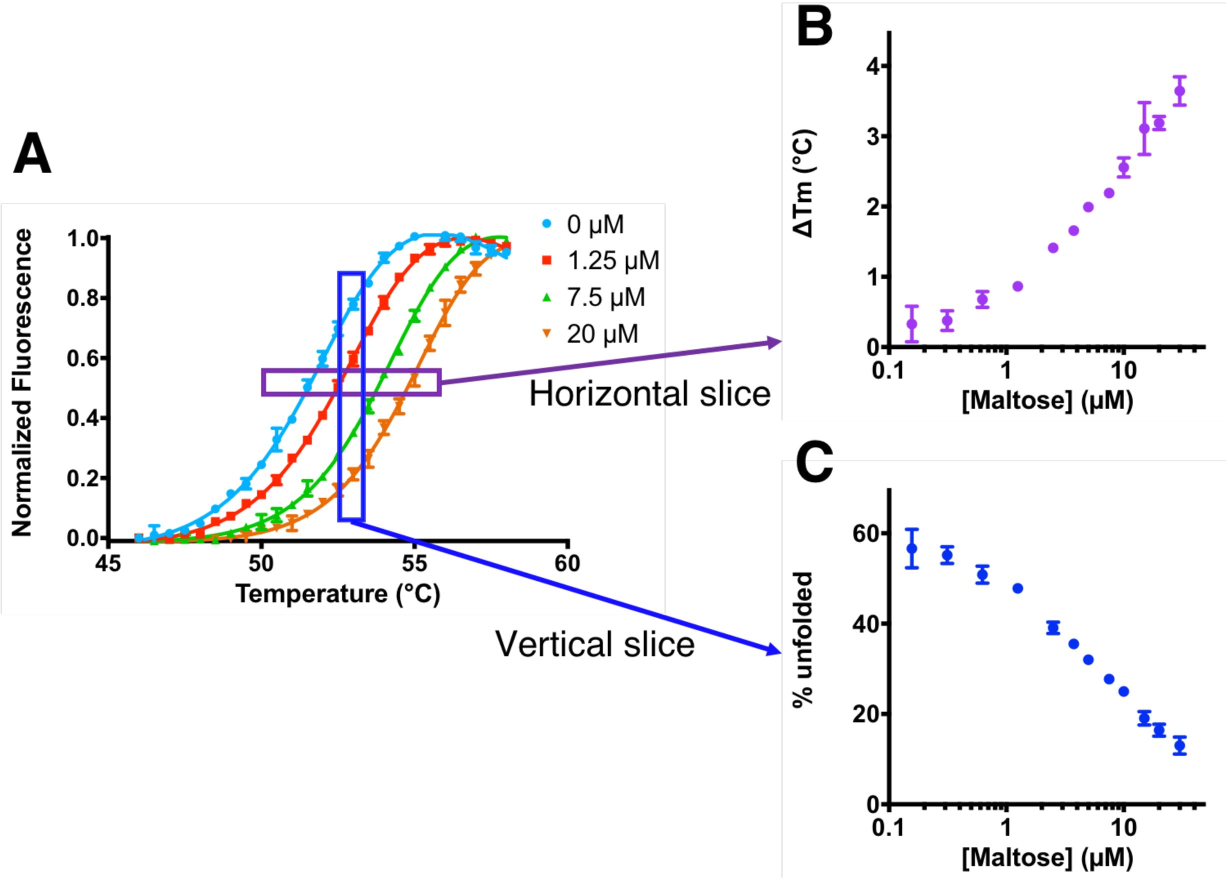


Figure 1: Maltose binding to MBP, as probed via DSF. (A) Thermal unfolding of MBP is monitored using SYPRO Orange. Data were collected in the presence of increasing maltose concentrations, leading to a rightward shift in the unfolding transition. **(B)** The T_m-shift (ΔT_m) is determined by plotting the increase in temperature at which each curve has 50% relative fluorescence, corresponding to a horizontal "slice" of the original data. However, this analysis does not provide the binding affinity of the protein/ligand pair. **(C)** Instead, here we use vertical "slices" of the original data. By plotting – *at a single temperature* – the fraction of protein that is unfolded as a function of ligand concentration (here at 53 °C), the binding affinity can then be easily determined. All data are collected in triplicate, and error bars correspond to the standard error of the mean (some are too small to be seen).

Dose-response data in DSF experiments are typically presented by showing the T_m-shift as a function of ligand concentration (**Figure 1b**), and there are a number of ways to determine T_m from the fluorescence data. One simple method is to take the first derivative of the observed fluorescence data with respect to temperature, and to then identify the maximum value (corresponding to the steepest part of the transition). Other methods instead smoothly fit the whole melting curve, either by using a so-called

Boltzmann model (57, 58, 60, 138), or by using a more rigorous "thermodynamic model" (123, 125, 128, 139, 140), or occasionally by using other arbitrary polynomials (141-143).

The Boltzmann model is the most widely-used approach, in part because it is very simple (58). The fluorescence at a given temperature is linearly related to the fraction of unfolded protein, which takes the form $F_{unfolded}(T) = 1 + \frac{1}{1+e^{\frac{Tm-T}{a}}}$, where Tm is the melting temperature and a is a parameter that reflects the steepness of the thermal unfolding transition. This model is applied primarily because it provides a sigmoidal shape that can be fit quite well to experimental data, especially when additional fitting parameters are included to account for the fact that the dye itself often has some temperature dependence (**Figure S1**). Despite its name, however, this equation does not explicitly model the thermodynamic transition (138): for this reason, the Boltzmann model is not used to garner any information beyond accurately identifying the midpoint of the protein unfolding transition (T_m) (58, 138), and studies that use this model simply report the presence/absence of binding rather than using this data to determine binding constants (58, 126, 144-148).

In studies to date seeking quantitative binding constants, "thermodynamic models" have been used. The simplest of such models write the fraction of unfolded protein as $F_{unfolded}(T) = 1 + \frac{1}{1+e^{\frac{\Delta H(1-{}^{T}/_{Tm})-\Delta C_p\left(Tm-T+Tln({}^{T}/_{Tm})\right)}{RT}}}$, where ΔH is the enthalpy change of protein unfolding and ΔCp is the change in heat capacity enthalpy change of protein unfolding (both assumed to be temperature-independent) (125, 149, 150). Typically ΔCp is under-determined given the available experimental data, and therefore determined through separate complementary experiments (139) or estimated from the buried surface area of the folded protein (151, 152), then fixed when fitting the thermal unfolding data. Though more complicated to write down, these models in fact have the same number of the effective free parameters (when ΔCp is fixed at a pre-determined value). Further, these models also have the advantage of using physically meaningful parameters.

Simply determining the T_m-shift as a function of ligand concentration is not sufficient to provide the binding affinity, however. Although some groups have simply fit these curves using the Hill equation (58, 153) – treating the T_m as an arbitrary "observable" that depends on the ligand concentration – this is not a physically reasonable approach. The Hill equation is only applicable when the observable is linearly proportional to the fraction of one of the species that is bound/unbound in solution, and T_m is not such a variable. The ΔT_m data are also (by definition) drawn from different temperatures: the binding affinity cannot be assumed to be constant at different temperatures, further making the Hill equation inappropriate for this usage. This point is further underscored by the fact that these experimental data do not correspond to a simple saturation-based ligand titration method (126): rigorous thermodynamic simulations show that ΔT_m should change monotonically with increasing ligand concentration (139, 154), even if this behavior is not always observed in real cases due to artifacts like irreversible protein aggregation (154).

Instead, correct binding constants have thus far been determined using a more rigorous approach that explicitly considers the temperature-dependent enthalpy, entropy, and heat capacity of both protein folding and ligand binding (125, 127). Using these thermodynamic parameters determined from the complete unfolding transitions, binding constants can subsequently be determined at the T_m. The means to do so was presented several decades ago (127), and also in the context of screening for ligands that bind a particular protein (123). In the earliest cases, these equations were formulated for the weak-binding regime (i.e. high dissociation constants), such that the free ligand concentration can be approximated by the total ligand concentration; these equations have since been extended to avoid the latter assumption (139, 154, 155). In all cases, though, the binding constant is determined at the T_m; together with the binding enthalpy, the van't Hoff equation can then be used to extrapolate binding constants at other temperatures. Because the binding enthalpy is difficult to determine from the unfolding transition data, this most commonly comes from a knowledge-based estimate (123) or is measured directly using other techniques like isothermal titration calorimetry (ITC) (156).

While details of the model have been iteratively improved since the original formulation, the two key elements of the "thermodynamic model" have remained unchanged: a fit of the melting curves is used to obtain multiple thermodynamic parameters, then these are used to calculate the binding constant at T_m and potentially (via extrapolation) at other temperatures (125). These elements of the model also remain the two key practical limitations of DSF. Because of the complexity associated with correctly replicating this analysis, it is often cited in modern studies but not frequently used: DSF is most popular as a qualitative test rather than a quantitative test, with the majority of literature reports reporting T_m-shifts as shown in **Figure 1b** but not attempting to extract binding constants (48, 130, 144, 145, 157-160). Collectively this has led to a general consensus that the observed T_m shifts "cannot be readily transformed into binding affinities" (161).

Here, we develop and describe a new *isothermal* strategy for analysis of DSF data. Rather than determine the T_m values from the raw fluorescence data at each ligand concentration, we instead select a *single* temperature of interest, and *at this temperature* we evaluate the fraction of protein that is folded/unfolded at each ligand concentration (**Figure 1c**). Because all of the data used corresponds to the same temperature, no thermodynamic parameters are required; instead, a very simple model of coupled equilibria (protein folding/unfolding and ligand binding/unbinding) describe our system. Furthermore, because we only require the fraction of protein that is unfolded (for a given ligand concentration, at the temperature of interest), the raw data can be fit either with the simple Boltzmann model or with the more rigorous thermodynamic model (**Figure S1**). Other studies have similarly used isothermal slices of unfolding data, for example in analysis of cellular thermal shift data (CETSA) (162) and other protein-ligand interactions (163, 164); however, each of these stopped short of using these data to quantitatively determine binding constants. As demonstrated below, here we show that this approach leads to a very simple formulation for determining the binding affinity near the protein's unfolding temperature, and it provides values consistent with those measured in other orthogonal assays.

Theory

Isothermal analysis of ThermoFluor data

DSF experiments, specifically those in which large compound collections are screened, yield melting temperatures that shift either higher or lower when various compounds are added (125, 154, 155). Most non-covalent drug-like ligands stabilize their protein target upon selective binding, and accordingly they increase the protein's T_m (58, 156, 165-167). Conversely, compounds that decrease the protein's T_m are thought to operate by binding the unfolded protein more tightly than the folded protein, by competing with an endogenous (stabilizing) co-factor, or through potentially non-specific effects (134, 154, 155, 168); some metal ions, like Zn^{2+}, can also destabilize proteins (169). We have excluded from the present analysis cases in which the ligand destabilizes the protein, and we focus solely on the scenario in which the ligand exclusively binds the natively-folded protein with a 1:1 stoichiometry.

Accordingly, we write the protein folding-unfolding reaction as a competitive coupled equilibrium with ligand binding, as follows:

$$[U] + [L] \underset{\mathrm{K_U}}{\rightleftharpoons} [F] + [L] \underset{\mathrm{K}_d}{\rightleftharpoons} [FL] \tag{1}$$

where [U] is the concentration of the unfolded protein, [L] is the concentration of free ligand, [F] is the concentration of the folded and unbound protein, and [FL] is the concentration of the protein-ligand complex. K_U is the equilibrium constant for the protein unfolding reaction, and K_d is the equilibrium constant for the unbinding reaction. Both K_U and K_d depend on temperature, but both are constant at fixed temperature (and fixed buffer conditions). Intuitively from this scheme, we see that the concentration of unfolded protein goes to zero as the ligand concentration becomes large and drives the equilibrium to the right. Importantly, this scheme assumes each reaction (folding and binding) has no intermediates, and thus can be represented in this two-state manner; we will consider further the implications of this assumption in the Discussion section. We also note that the presence of the reporter dye is not included in our model.

From the conservation of mass and the definitions of these two equilibrium constants, we write the following:

$$[P]_T = [F] + [U] + [FL] \tag{2}$$

$$[L]_T = [L] + [FL] \tag{3}$$

$$K_U = [U]/[F] \tag{4}$$

$$K_d = ([F] \times [L])/[FL] \tag{5}$$

where [U] is the concentration of the unfolded protein, [L] is the concentration of free ligand, [F] is the concentration of the folded and unbound protein, and [FL] is the concentration of the protein-ligand complex. In **Equation 4** we define K_U as the equilibrium constant between the *unbound* unfolded and folded states ([U] and [F]). This equilibrium constant is therefore independent of ligand concentration, and reflects the overall fraction of protein that is unfolded/folded *only* when no ligand is present (since inclusion of ligand shifts some of [U] and [F] into the [FL] state). K_d is the equilibrium constant for the unbinding reaction. $[P]_T$ is the total protein concentration, and $[L]_T$ is the total ligand concentration (both of which are known). We note that the interaction between the reporter dye and the protein is not explicitly included in this model, though the presence of the dye presumably does contribute to stabilizing the unfolded protein.

Once the raw data have been normalized, fluorescence intensity in the DSF experiment (**Figure 1a**) is linearly related to the fraction of the unfolded protein f_u. Starting from the definition of f_u, we simplify using **Equations 2-5** and obtain the following expression:

$$f_u = \frac{[U]}{[U] + [F] + [FL]} = \frac{1}{1 + ((1/K_U) \times (1 + [L]/K_d))} \tag{6}$$

This provides the fraction of unfolded protein in terms of the *free* ligand concentration [L], whereas the known quantity in this experiment is the total ligand concentration $[L]_T$. From **Equations 2-5** we obtain the following quadratic equation for [L]:

$$[L]^2 + ([P]_T - [L]_T + K_d(1 + K_U))\,[L] - [L]_T K_d(1 + K_U) = 0 \tag{7}$$

Thus, [L] can be written in terms of the *total* ligand concentration $[L]_T$ as follows:

$$[L] = \frac{1}{2}\left[([L]_T - [P]_T - K_d(1 + K_U)) + \sqrt{([P]_T - [L]_T + K_d(1 + K_U))^2 + 4[L]_T K_d(1 + K_U)}\right] \quad (8)$$

We note that this expression corresponds to only one root of the quadratic equation, since the other root is unphysical.

Together, **Equations 6 and 8** provide a single expression to write f_u in terms of $[L]_T$, $[P]_T$, K_U, and K_d. As expected for the limiting case where $[L]_T$ becomes large, we see from this set of equations that f_u goes to zero. Conversely in the limiting case when $[L]_T$ goes to zero, we see that [L] goes to zero and thus **Equation 6** reduces to the definition of the equilibrium constant for unfolding. Together, these two limits correspond to the endpoints of the data shown in **Figure 1c**.

$[L]_T$ and $[P]_T$ are known experimental parameters; our expression for f_u therefore uses only two free parameters (K_U and K_d). These two parameters can be fit to the normalized data at the same time (as we will demonstrate), or alternatively K_U can be first determined at the temperature of interest from the thermal unfolding curve in the absence of ligand; this allows fitting of the data in **Figure 1c** to be subsequently carried out with a single free parameter (K_d).

A simpler approximate solution

Monitoring the fraction of unfolded protein in this competitive coupled equilibrium (**Equation 1**) is very much analogous to detecting the fraction of labeled probe molecule in a competitive binding assay. In the latter case, one uses increasing concentrations of the unlabeled inhibitor of interest to explore the effect on a labeled probe that binds at the same site. The concentrations of all species, as well as the binding affinity of the probe ligand, can then be used to determine the inhibition constant for the unlabeled species from its IC_{50} (170).

Inspired by this analogy, we explored whether the same strategy could be applied here. We summarize our solution for these equations below, and elaborate further in the Appendix.

We again start from **Equations 2-5**, but this time we solve these equations for the specific scenario in which the total ligand concentration matches the EC_{50}. By definition, the EC_{50} is the ligand concentration at which the fraction of unfolded protein is half of that observed in the absence of ligand (note: the EC_{50} is *not* defined by the ligand concentration at which half of the protein is unfolded, since this can happen even before ligand is added, depending on the temperature). For this special case:

$$[L]_T = [L]_{50} + [FL]_{50} = EC_{50} \tag{9}$$

$$[P]_T = [F]_{50} + [U]_{50} + [FL]_{50} \tag{10}$$

$$K_U = [U]_{50}/[F]_{50} \tag{11}$$

$$K_d = ([F]_{50} \times [L]_{50})/[FL]_{50} \tag{12}$$

where $[U]_{50}$, $[L]_{50}$, and $[F]_{50}$ are the concentrations of unfolded protein, free ligand, and folded unbound protein at the condition when $[L]_T = EC_{50}$. Recall from **Equation 4** that K_U is defined to be the equilibrium constant between the only the *unbound* unfolded/folded states (not the overall fraction of protein that is unfolded/folded), and thus for this reason **Equation 11** does not include any contribution from $[FL]_{50}$.

Correspondingly, in the absence of ligand we write:

$$[P]_T = [F]_0 + [U]_0 \tag{13}$$

$$[L]_T = [L]_0 = [FL]\ _0 = 0 \tag{14}$$

$$K_U = [U]_0/[F]_0 \tag{15}$$

From **Equations 13-15** we can solve for the fraction unfolded in the absence of ligand (f_{u0}):

$$f_{u0} = \frac{[U]_0}{[U]_0 + [F]_0} = \frac{1}{1 + 1/K_U} \tag{16}$$

From the definition of EC_{50}, we write:

$$[U]_{50} = \frac{[U]_0}{2} \tag{17}$$

From **Equations 15 and 17**, we can write $[F]_{50}$ in terms of $[U]_0$ and K_U. Substituting this into **Equation 10** yields an expression for $[FL]_{50}$ in terms of $[P]_T$, $[U]_0$ and K_U; simplifying this with **Equations 15 and 16**, we find that at the ligand concentration corresponding to the EC_{50}, half of the total protein concentration has ligand bound to it:

$$[FL]_{50} = [P]_T/2 \tag{18}$$

This allows solution of **Equations 12 and 16** to yield a simple expression for $[L]_{50}$ as well:

$$[L]_{50} = \frac{K_d}{1 - f_{\mathrm{u0}}} \tag{19}$$

Combining **Equations 18 and 19** back into **Equation 9**, we obtain a simple expression that relates the EC_{50} to K_d:

$$EC_{50} = \frac{K_d}{1 - f_{\mathrm{u0}}} + \frac{[P]_T}{2} \tag{20}$$

There are no additional assumptions required to reach this equation (e.g. no need to assume that $[L] \approx [L]_T$). This expression is intuitively gratifying, and it highlights the fact that the EC_{50} observed in this experiment cannot be simply interpreted as the K_d. Most notably, in the limit where ligand binding is very tight (low K_d), the observed EC_{50} is driven essentially by stoichiometry (enough ligand must be added to match half the number of available sites on the protein); this makes the EC_{50} very insensitive to changes in the K_d in this regime, and it suggests that our approach may not be well-suited to determining the binding affinity for very tight interactions. This implication is borne out in real experimental data, as presented at the end of the following section.

Finally, rearranging **Equation 20** yields:

$$K_d = (1 - f_{\mathrm{u0}}) \times \left(EC_{50} - \frac{[P]_T}{2}\right) \tag{21}$$

$[P]_T$ is a known experimental parameter. f_{u0} corresponds to the fraction of protein that is unfolded (at the temperature of interest) in the absence of ligand, and thus it can be determined directly from the thermal unfolding curve in the absence of ligand. Even using a very simple and arbitrary fit of f_{u0} as a function of ligand concentration (e.g. the Hill equation), we can still easily estimate the midpoint of this transition (the ligand's EC_{50} value): thus, **Equation 21** provides a rapid means to estimate the K_d when it is undesirable to fit the complete curve using **Equations 6 and 8**. That said, fitting with the functional form presented in **Equations 6 and 8** leads to the most accurate estimate of the midpoint (since the complete curve is used to determine the fitting parameters), and is thus preferred.

Results

To test the utility of this isothermal fitting approach, we wrote a program in Python that fits the thermal unfolding curves and uses these to solve **Equations 6 and 8** presented above. All of the analysis presented below was carried out using this program, and it is freely available for others to use (https://sourceforge.net/projects/dsf-fitting).

<u>*Accuracy and robustness of isothermal analysis*</u>

We first sought to test the accuracy of binding affinity values resulting from this isothermal approach. To do so, we generated realistic simulated thermal unfolding curves. The rigorous approach referenced earlier (139, 154) allows the fraction of unfolded protein to be calculated as a function of temperature and ligand concentration, provided thermodynamic parameters that describe protein unfolding in the absence of ligand (ΔH_U^{Tm}, ΔCp_U^{Tm}, T_m and K_U^{Tm}) and thermodynamic parameters that describe ligand binding (ΔH_b^{Tm}, ΔCp_b^{Tm} and K_d^{Tm}). We selected values for each of these parameters by using values for maltose/MBP from the literature where possible, and then assigning reasonable values to the remaining terms such that the resulting curves were qualitatively similar to those observed experimentally for maltose/MBP.

In our simulations we set T_m in the absence of ligand to be 50 ºC, and set the ligand's dissociation constant (K_d) to be 1 μM at this temperature. By definition, the unfolding constant (K_U) at the T_m is 1. Using the formulation laid out by others (139, 154) we then generated the corresponding simulated experimental data (**Figure 2a**). To make the simulated data suitably approximate the type of experimental data that would be produced in a real experiment, we generated data with temperature intervals and ligand concentrations drawn from the real experimental protocol used earlier (**Figure 1**). Reassuringly, analyzing this data using the isothermal approach presented above yielded values for both K_d and K_U that matched those used to generate the simulated data. The previous formulation (139, 154) makes it straightforward to generate unfolding curves from thermodynamic parameters, but the inverse problem is more challenging to solve; by demonstrating that our isothermal methods recovers the underlying K_d and K_U, we show that our method is indeed compatible with the previous formulation of this system.

To explore the robustness of our isothermal analysis, we next introduced noise into the simulated experimental data. Having already normalized each of the simulated unfolding curves to range from 0 to 1, we added to each point a random number drawn from a normal distribution defined by a given standard deviation (σ). We find that analysis of the resulting data yields K_d and K_U values that closely match the true value for σ up to 0.05; only once the data becomes noisier than this ($\sigma = 0.1$) do the estimates start to differ from those used to generate the unfolding data. The amount of noise in the simulated data at $\sigma = 0.1$ is more than observed in typical experiments, suggesting that indeed this isothermal analysis is robust to the random error present in most real experimental data.

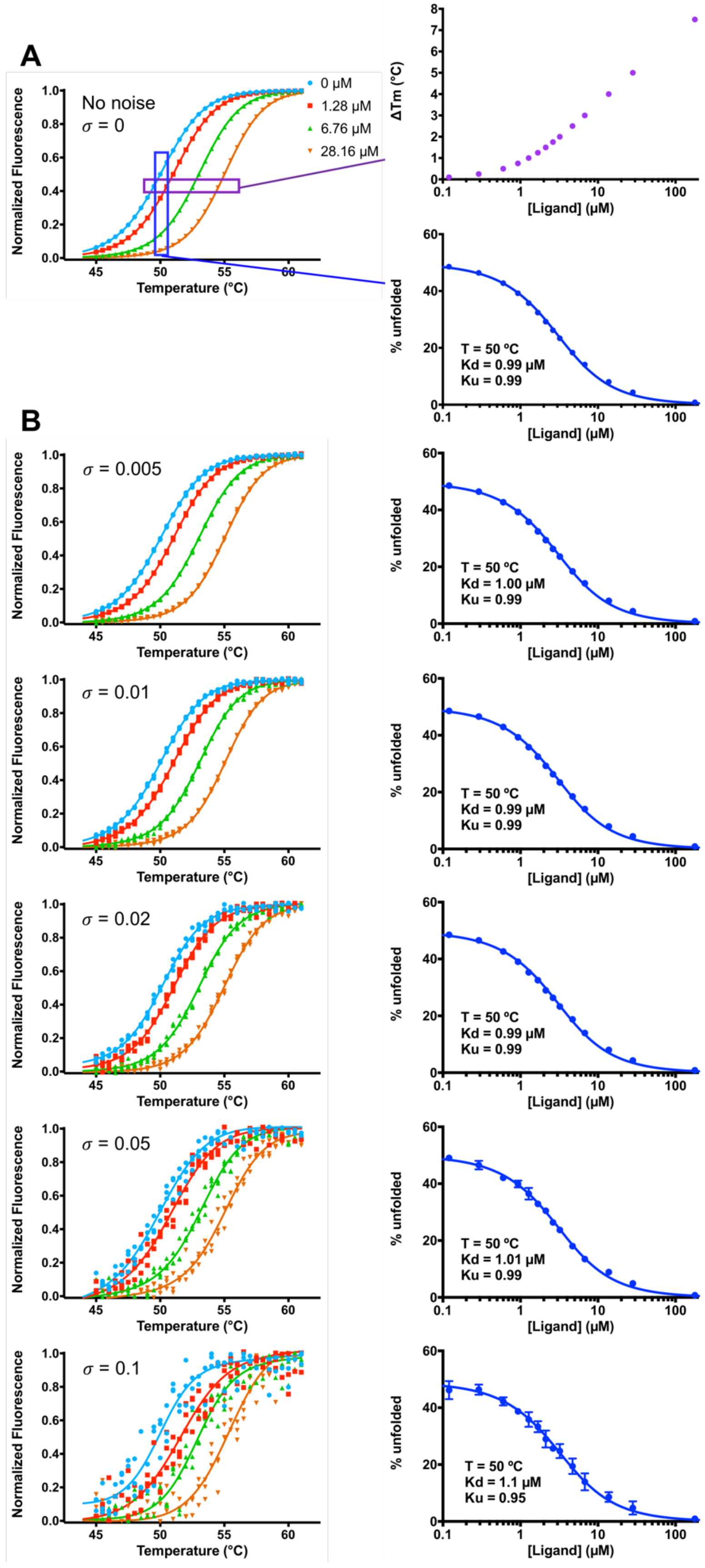

A
B
No noise
σ = 0
0 μM
1.28 μM
6.76 μM
28.16 μM
Normalized Fluorescence
Temperature (°C)
ΔTm (°C)
[Ligand] (μM)
% unfolded
T = 50 ºC
Kd = 0.99 μM
Ku = 0.99
σ = 0.005
T = 50 ºC
Kd = 1.00 μM
Ku = 0.99
σ = 0.01
T = 50 ºC
Kd = 0.99 μM
Ku = 0.99
σ = 0.02
T = 50 ºC
Kd = 0.99 μM
Ku = 0.99
σ = 0.05
T = 50 ºC
Kd = 1.01 μM
Ku = 0.99
σ = 0.1
T = 50 ºC
Kd = 1.1 μM
Ku = 0.95

Figure 2: Simulations to explore the consistency and robustness of isothermal analysis. **(A)** Simulated thermal unfolding curves were generated using a thermodynamic model for unfolding and binding. Parameters were set as follows: T_m = 50 ºC, K_d^{Tm} = 1 μM, ΔH_U^{Tm} = 120 kcal mol^{-1}, ΔH_b^{Tm} = -10 kcal mol^{-1}, ΔCp_U^{Tm} = 4 kcal mol^{-1} K^{-1}, ΔCp_b^{Tm} = -0.5 kcal mol^{-1} K^{-1}, and total protein concentration = 2 μM. By definition, K_U^{Tm} = 1. Fitting this simulated data using the simpler isothermal approach yields K_U^{Tm} = 0.99, and K_d^{Tm} = 0.99 μM. **(B)** Upon addition of increasing random noise to the simulated unfolding data, the isothermal approach still leads to accurate estimates of K_U^{Tm} and K_d^{Tm}, up to values exceeding the noise typically present in real experimental data.

Application to maltose/MBP

As a first test of this approach, we analyzed in further detail the maltose/MBP interaction. This interaction has been frequently studied using many different forms of calorimetry (151, 171), in part because both the ligand and the protein are soluble to very high concentrations. We returned to the same DSF experimental design described earlier, with 12 increasing concentrations of maltose (**Figure 3a**). Given MBP's T_m of about 52.5 ºC in the absence of maltose, we first elected to determine the binding affinity for this pair at 53 ºC. From individual fits to the complete thermal unfolding curves, we used the thermodynamic model to determine the fraction of unfolded MBP (at 53 ºC) at each maltose concentration. We also separately used the Boltzmann model to determine the fraction of unfolded MBP from each thermal unfolding curve, and we found that both methods yielded essentially identical results (**Figure S1**).

We then used the expressions presented in **Equations 6 and 8** to fit the fraction of unfolded MBP at each maltose concentration (**Figure 3b**). From the fraction of unfolded protein at each maltose concentration, there are only two free parameters to be fit: the fraction of unfolded MBP in the absence of maltose (K_u) and the dissociation constant for ligand binding (K_d). At low maltose concentration, the curve does not go to 100% unfolded, but rather plateaus at about 50% (the first free parameter): this is expected, given that we carried out our analysis at a temperature only slightly above the T_m. The K_d at 53 ºC, derived directly from this fit, is 2.7 μM, and the K_U value at 53 ºC is 1.3. Because we have defined K_U as the unfolding constant *in the absence of ligand*, we can also compare the value to that obtained directly from the thermal unfolding curve collected in the absence of ligand: the latter yields a value of 1.5 (at 53 ºC), in very close agreement with the value obtained from fitting the binding curve.

We additionally fit these data using the approximate solution shown in **Equation 21**: given the EC_{50} value of 6.8 μM (estimated by arbitrarily using the Hill equation to fit this curve), this expression yields a K_d value of 2.5 μM, in agreement with the more rigorous fit.

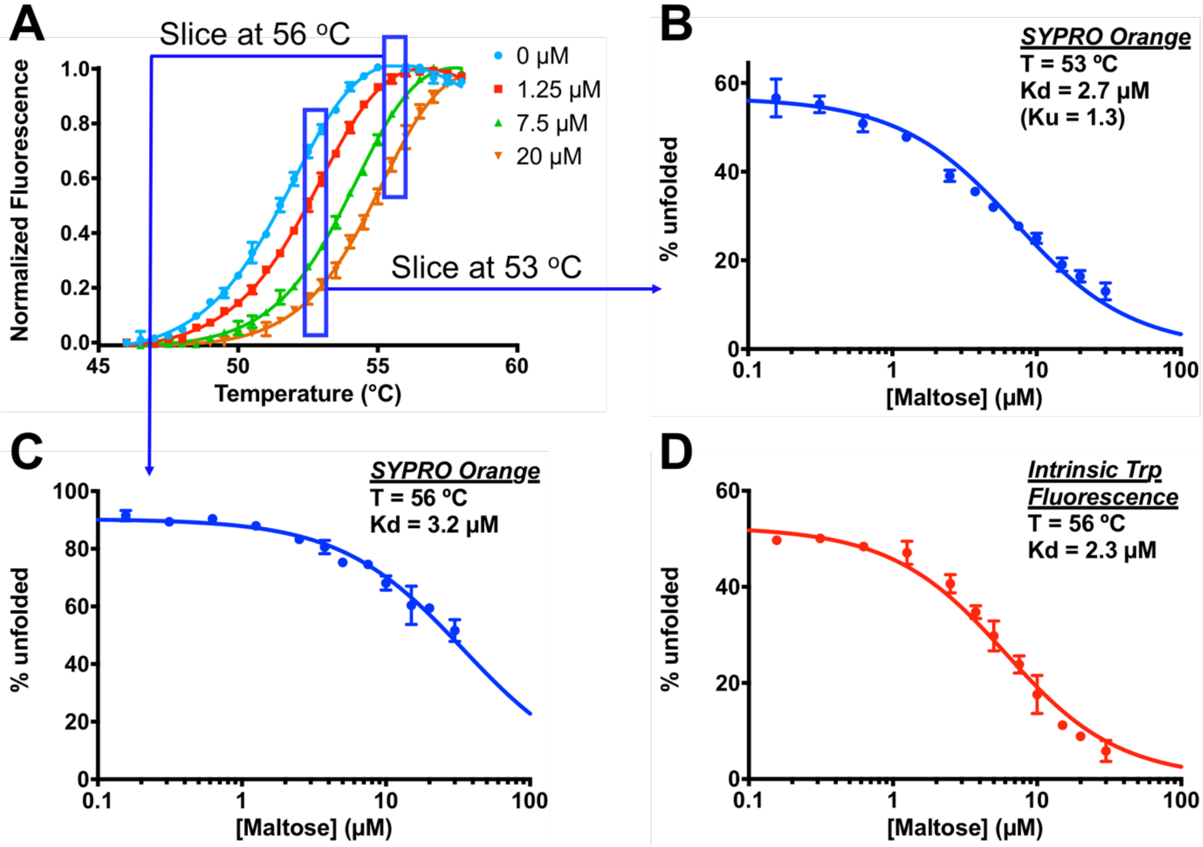


Figure 3: Determination of maltose/MBP binding affinity using isothermal analysis of thermal unfolding data. (A) Thermal unfolding of MBP is monitored using SYPRO Orange. Data were collected using 12 increasing maltose concentrations, each in triplicate; 4 representative unfolding curves are shown, after normalization using the Boltzmann equation. **(B)** The fraction of unfolded protein is calculated at 53 ºC for each maltose concentration. Fitting using Equations 6 and 8 yields a K_d value of 2.7 μM and a K_u value of 1.3. **(C)** Extracting instead the fraction of unfolded protein at 56 ºC yields a K_d value of 3.2 μM. **(D)** The thermal unfolding transition was instead monitored using MBP's intrinsic tryptophan fluorescence, and the fraction of unfolded protein was calculated at 56 ºC for each maltose concentration. Two replicates were carried out for each maltose concentration. Fitting this complementary experimental data using Equations 6 and 8 yields a K_d value of 2.3 μM.

One advantage of the isothermal fitting approach is that the binding constants can be directly determined at other temperatures close to the T_m, provided that there are sufficient differences in the fraction of unfolded protein. As a demonstration of this, we carried out the corresponding analysis using a slightly higher temperature, at 56 ºC (**Figure 3c**); as expected, the curve from this fit has a higher fraction of unfolded MBP in the absence of maltose. Binding at this slightly elevated temperature yields a very similar K_d value of 3.2 μM.

This general approach for extracting dissociation constants is by no means specific to the DSF format; while this is a convenient method for monitoring protein unfolding, the analysis presented here can also be applied to data collected via using other experimental techniques. While DSF is label-free, in principle the presence of SYPRO Orange (or other analogous dyes) may shift the folding equilibrium by preferentially binding to the unfolded state (155); still, given the analysis outlined above, a systematic shift in protein stability (due to the dye, for example) is not expected to affect the resulting binding affinity. To further explore the effect of the dye, we repeated the experiment described above, this time in the absence of SYPRO Orange and instead relying on MBP's intrinsic tryptophan fluorescence to monitor unfolding. From an initial experiment in the absence of maltose, we noted that the T_m was now about 2.5 ºC higher: this confirmed our expectation (and previous reports (155)) that the presence of the dye slightly destabilizes the protein.

Using thermal unfolding traces collected via intrinsic tryptophan fluorescence, we plotted the fraction of unfolded protein at 56 ºC, as a function of maltose concentration (**Figure 3d**). We again fit these data using the expression from **Equation 6 and 8**, and again we find that this expression (with two free parameters) appropriately describes the underlying data. The K_d value resulting from this fit at 56 ºC is 2.3 μM, in close agreement with the value obtained at this temperature using the DSF data. In contrast, the value of K_U (in the absence of ligand) at 56 ºC is 1.1 using intrinsic tryptophan fluorescence versus 9.4 using SYPRO Orange: this is consistent with the fact that this temperature is very close to the T_m determined via intrinsic tryptophan fluorescence but above the T_m determined using SYPRO Orange, and again implies that the dye destabilizes the protein.

Using this pair of complementary detection modalities, we have thus confirmed that the general approach laid out above is applicable for analysis of thermal unfolding data, regardless of the experimental means by which the protein's foldedness is monitored. Inevitably, however, this analysis reports on the binding affinity at a temperature near the protein's T_m (e.g. ± 4 ºC in this case). In the data presented above we obtain the binding affinity for maltose/MBP at 53 and 56 ºC, whereas other techniques to directly probe binding such as ITC and SPR can be used at more physiological temperatures (e.g. room temperature or 37 ºC (151, 172, 173)). In general, extrapolation of binding data from DSF to lower temperatures will require knowledge of the thermodynamic contributions to binding: these may be derived from applying the isothermal approach multiple times over a small temperature range, or from complementary calorimetry experiments (156) / knowledge-based estimates (123) as described elsewhere. We will test the feasibility of these strategies in future work; here, instead, we next sought to explore whether addition of chaotropic agents would allow us to probe this interaction at lower temperature.

Using denaturants to access binding constants at lower temperature

Although thermal unfolding may be monitored over a large temperature range, accurate determination of the binding affinity by this method requires that there is a well-resolved range in the fraction of protein that is unfolded; thus, it is natural to carry out this analysis at temperatures close to the T_m measured in the absence of ligand. In many cases, however, it is desirable to probe binding at lower, more physiologically-relevant temperatures.

To shift MBP's T_m to the desired temperature, we added denaturant to our system. Guanidine hydrochloride (GdnHCl) has been shown not to greatly affect the binding affinities for most protein-ligand interactions, with the exception of strongly ionic ligands (161, 174-176). As a starting point, we used DSF experiments to monitor the MBP's T_m in the presence of increasing denaturant (**Figure 4a**); based on these results, we elected to study maltose binding at a GdnHCl concentration of 0.7 M. At this denaturant concentration, we then carried out the same DSF experiments with increasing concentrations of maltose. Having shifted the transition temperature into the physiological range, we now determined the fraction of unfolded MBP at 35 ºC (**Figure 4b**). Under these conditions, the fit once again appropriately describes the data, and yields a K_d value of 2.3 μM: this estimate is consistent with previous studies reporting of values ranging from 0.5 to 2 μM for this interaction (171, 172, 177).

By adding this chaotropic agent, then, we have demonstrated that the thermal unfolding transition can be rationally shifted to allow determination of binding affinity at a specific temperature. Importantly, we also observe that – at least in this model system – the presence of GdnHCl does not significantly affect the resulting binding affinity, in agreement with previous reports (161, 174-176).

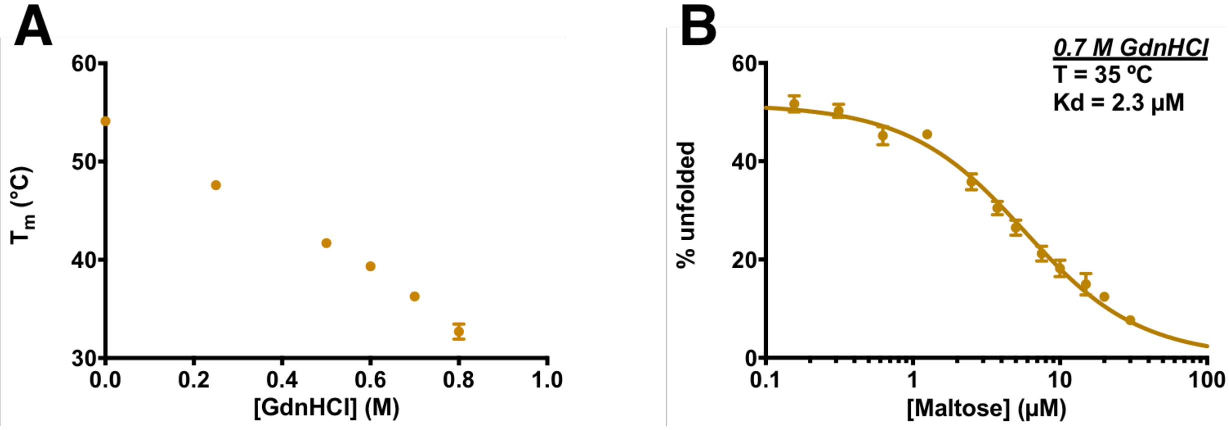


Figure 4: Denaturant effect of MBP unfolding and MBP-maltose binding. **(A)** T_m of MBP decreases with increasing GdnHCl concentration. **(B)** MBP-maltose binding with 0.7 M GdnHCl at 35 ºC. The value of K_d is 2.3 μM. All the experiments were carried out in triplicate. The protein concentration in all the assays was fixed to 2 μM. All assays were taken in the buffer: 120 mM NaCl, 20 mM NaH_2PO_4/Na_2HPO_4, pH 7.4 with 1% DMSO and the melting program was set to 0.5 ºC/min.

Applying this approach to other protein-ligand pairs (without denaturant)

We next applied this approach to study a different protein, with multiple ligands spanning a broad range of binding affinities. We selected another model system that has been frequently used in calorimetric studies (139, 140, 154, 161, 178), carbonic anhydrase (isoforms I and II). From among commercially-available inhibitors of these two enzymes we selected the weak inhibitor sulfanilamide (SULFA, mM K_i) and the potent inhibitor trifluoromethanesulfonamide (TFMSA, nM K_i). We also selected two inhibitors with intermediate inhibition constants (μM K_i), acetazolamide (ACTAZ) and methazolamide (METAZ). The chemical structures of all four inhibitors are shown in **Figure S2**. There is no chemical denaturant used in these assays.

From thermal unfolding data collected in the absence of inhibitor (not shown), we observed that *b*-CA II was slightly more stable than *h*-CA I. For this reason, we evaluated isothermal binding data for the two isoforms at 60 ºC and 57 ºC, respectively. From the resulting binding curves (**Figure 5a-d**), the relative activities of each inhibitor are clear: SULFA is the weakest, followed by ACTAZ and METAZ, and TFMSA is the most potent. Importantly, this experiment also distinguishes between the two isoforms, with tighter binding observed for *b*-CA II rather than *h*-CA I in all four cases. Overall, this isothermal analysis of the underlying thermal unfolding data provides a range of binding affinities between CA isoforms and their four inhibitors.

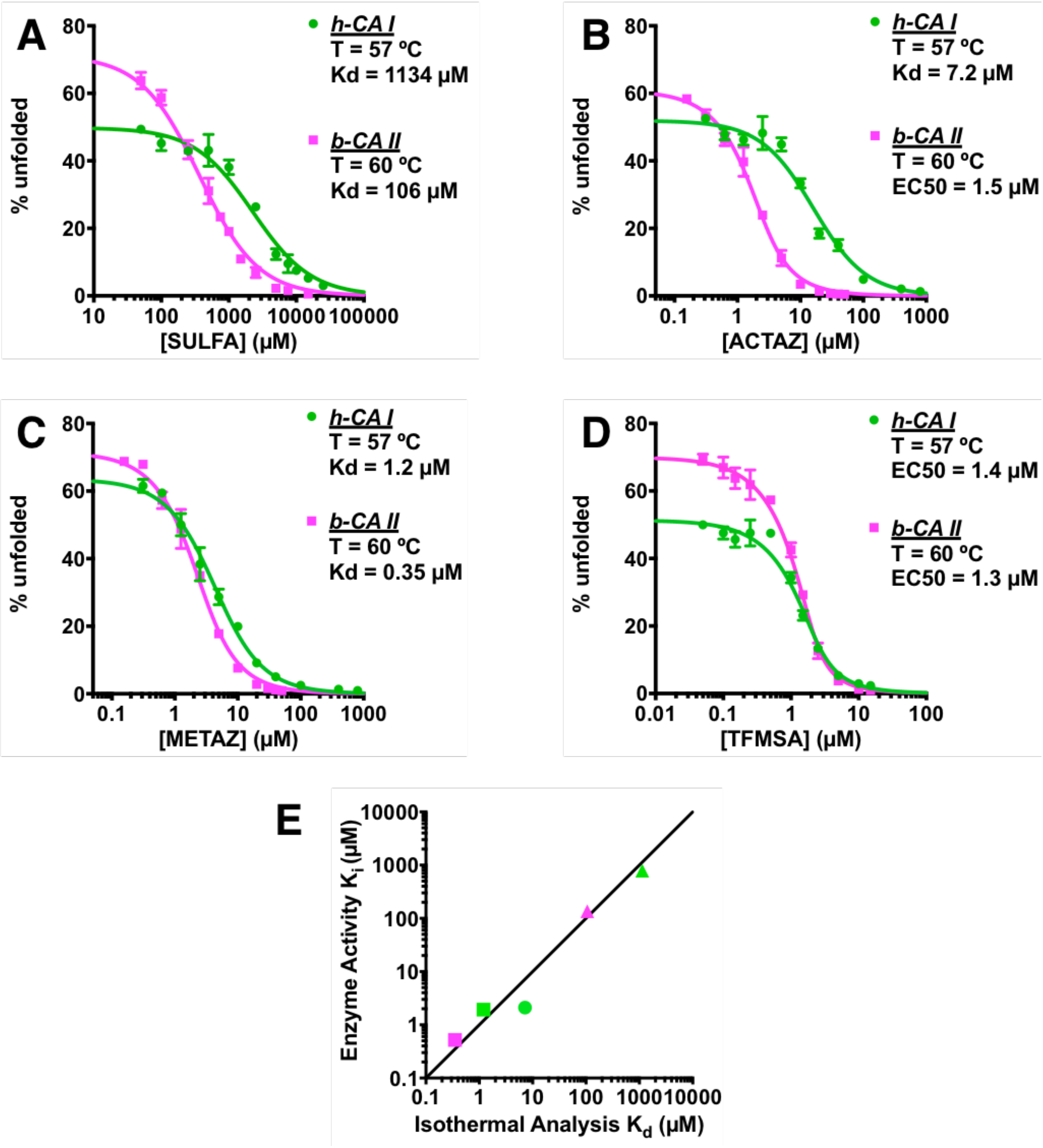


Figure 5: Determination of binding affinities for carbonic anhydrase inhibitors using isothermal analysis of thermal unfolding data. Each inhibitor was characterized with two carbonic anhydrase isoforms, *h*-CA I (*green*) and *b*-CA II (*pink*). **(A)** Analysis of SULFA yielded binding constants of 1.1 mM and 0.1 mM for isoforms *h*-CA I and *b*-CA II. **(B)** ACTAZ gave binding constants of 7.2 μM for *h*-CA I and 1.5 μM as EC_{50} for *b*-CA II. **(C)** METAZ gave binding constants of 1.2 μM and 0.35 μM for the two isoforms. **(D)** TFMSA gave EC_{50} of 1.4 μM and 1.3 μM for the two isoforms. **(E)** Comparison of the binding constants obtained from isothermal analysis of thermal unfolding data versus inhibition constants obtained in an enzyme inhibition activity. The TFMSA/*h*-CA I, TFMSA/*b*-CA II and ACTAZ/*b*-CA II pairs are not included here, because they all have less than 2 μM EC_{50}. All experiments were carried out in triplicate.

As noted earlier with regards to our discussion of **Equation 20**, the dissociation constant is difficult to obtain from the isothermal analysis of DSF data if K_d is much lower than the protein concentration, because binding becomes stoichiometric in this regime. The protein concentration used in these experiments was 2 μM (to allow robust detection of the unfolding transition), and thus **Equation 21** shows that this will make it difficult to interpret the K_d for EC_{50} values below about 2 μM. For this reason, we used the EC_{50} value to guide interpretation of the results: for cases with EC_{50} values greater than 2 μM we report the K_d, whereas for cases with EC_{50} less than 2 μM we simply conclude that the K_d is less than 0.5 μM (as seen from **Equation 20**, when $[P]_T = 2$ μM and f_{u0} is 0.5, a K_d value of 0.5 μM will lead to $EC_{50} = 2$ μM).

Interaction (ligand/protein)	K_d (μM), from DSF	K_i (μM), from enzyme assay
SULFA / *h*-CA I	1134 ± 106	786 ± 60
SULFA / *b*-CA II	106 ± 9	136 ± 5
ACTAZ / *h*-CA I	7.2 ± 0.8	2.1 ± 0.2
ACTAZ / *b*-CA II	< 0.5	0.48 ± 0.03
METAZ / *h*-CA I	1.2 ± 0.1	1.9 ± 0.2
METAZ / *b*-CA II	0.35 ± 0.03	0.52 ± 0.03
TFMSA / *h*-CA I	< 0.5	< 1
TFMSA / *b*-CA II	< 0.5	0.12 ± 0.04

Table 1: Binding/inhibition constants derived from isothermal analysis of DSF and from enzyme activity assays.

To obtain an independent measure of these interactions under identical conditions (temperature and buffer composition), we applied an esterase activity assay and determined inhibition constants (K_i) for each isoform/inhibitor pair (**Table 1**). Of the eight protein/ligand pairs, we could not accurately determine the inhibition constant for TFMSA with *h*-CA I due to its potency relative to the enzyme concentration used in our assay: standard Michaelis-Menten analysis cannot be used to determine the inhibition constant for a nanomolar inhibitor at an enzyme concentration of 2 μM. With the exception of this pair, we compared these inhibition constants to the binding constants obtained via DSF: overall there is excellent agreement between the inhibition constants and the binding constants, for activities in the sub-micromolar to millimolar range (**Figure 5e**).

Discussion

<u>*Limitations of using thermal unfolding to monitor ligand binding*</u>

The simplicity and practical advantages of the DSF format have made this experimental approach very popular, particularly for fragment screening. Nonetheless, there are important considerations that can limit its application with respect to certain ligands and/or proteins.

With respect to ligand screening, certain ligands can naturally interfere with the reporter dye through their own fluorescent properties (123). In addition, certain ligands may interact with the unfolded protein (125, 154), or promiscuously form non-specific (or covalent) interactions with the protein (179). Other ligands may also interact with the protein via a stoichiometry other than 1:1, or alternatively, and particularly for small fragment-like compounds, the ligand may interact with a single site on the protein surface using multiple binding modes with comparable affinities (180). The isothermal formulation we present here does not yet consider any such scenarios, and it is currently restricted to stoichiometric 1:1 binding.

The protein to be studied is also subject to important restrictions. Most importantly, any equilibrium analysis involving protein folding assumes reversibility: this can be difficult to establish conclusively, and many proteins aggregate at elevated temperature. In the course of initiating the studies described here, we explicitly tested whether thermally-unfolded protein could be cooled, and then once again heated to yield the same thermal unfolding transition. In light of well-justified concerns about non-reversible unfolding and aggregation, a number of strategies have been proposed: these include using faster speeds for the unfolding process (to minimize the potential for aggregation) (181, 182) and including in the reaction dyes that explicitly detect protein aggregation (183, 184).

Further, our formulation also makes the important assumption that protein folding/unfolding is two-state, and that there are no substantially-occupied partly-folded intermediates. This is a pervasive assumption, because it greatly simplifies analysis of folding/unfolding data: however, it is also widely-understood that this assumption is not valid in all cases (185). Relative to traditional thermodynamic characterizations of ligand binding, we expect that the isothermal nature of our analysis will somewhat mitigate the effect of partially-folded states science they will simply be lumped into the folded or unfolded state, depending on their ability to bind ligand. That said, the presence of such states can certainly confound this analysis if they bind the ligand with a different affinity than the folded state (e.g. leading to inadvertent determination of some ensemble-weighted average of the binding constant), or if they lead to errors in calculating the fraction of unfolded protein at the temperature of interest.

Thermal versus denaturant-induced unfolding

The binding constants derived from this isothermal approach can be accurately determined only in the vicinity of the target protein's melting temperature, which may not correspond to a temperature of real biological (physiological) interest. We have shown that in such cases denaturant can be used to shift the melting temperature to the desired range; an alternative approach, however, is simply to extract binding constants from the ligand-dependence of the denaturant-induced stability differences.

Indeed, previous studies have laid the groundwork for determining binding constants – at room temperature – based on the denaturation midpoint of protein stability (50, 161). In both of these studies the extent of protein unfolding was monitored by intrinsic tryptophan fluorescence, obviating the need for a reporter dye. That said, an important drawback of this detection modality is the potential for interference from many drug-like compounds, which may limit the range of applicability of this technique. By contrast, typically used reporter dyes such as SYPRO Orange are much less likely to exhibit spectral overlap with potential ligands of interest. Additionally, whereas the data collection for denaturation profiles is usually more rapid, data are typically fit using closely-spaced increments in denaturant concentrations which necessitates more liquid handling to setup the assay.

Overall, we envision that thermal and chemical unfolding can serve as complementary assays, depending on available instrumentation, the importance of obtaining binding constants at a specific temperature, and the spectral nature of the ligands of interest.

Sensitivity of detecting protein unfolding

As described in the context of **Equations 20 and 21**, the fraction of protein that is folded depends on the protein concentration relative to the ligand's K_d. Under circumstances in which the dissociation constant is much smaller than the protein concentration (i.e. very tight binding), addition of ligand leads to stoichiometric binding and makes it difficult to determine the binding constant. Indeed, we encountered precisely this scenario in our characterization of the carbonic anhydrase inhibitor TFMSA.

The natural solution to this problem is to use very low protein concentration, so that the EC_{50} observed for unfolding is most sensitively dependent on the K_d rather than on the protein concentration. This raises a practical consideration, however, because the protein concentration to be used in the assay is determined by the sensitivity with which unfolding can be monitored. The observed fluorescence signal upon dye binding is related to protein size, with larger proteins yielding more signal: for this reason, experiments typically use similar protein concentrations in mass units (i.e. mg/ml) rather than in molar concentrations. For larger proteins, then, lower molar concentrations are accessible for this experiment (provided they unfold in a single cooperative transition), which in turn may allow for characterization of tighter-binding ligands using this approach.

Thermodynamic models versus our isothermal model

Monitoring protein thermal unfolding transitions is a highly attractive means to access ligand binding, because in principle it can be rapidly setup and deployed for many different protein systems. In addition to DSF / ThermoFluor, analogous data can be collected using other experimental modalities: most notably, probing the protein directly via intrinsic tryptophan fluorescence or circular dichroism (CD) spectroscopy. Regardless of the method by which protein unfolding is monitored, however, the analysis is the same; and indeed, the same thermodynamic models described earlier in the context of DSF have also been applied to thermal unfolding probed via CD (181, 186, 187). Unsurprisingly, the challenges associated with applying thermodynamic models to directly quantify ligand binding at different temperatures apply to these other experimental formats as well (188). Here we have demonstrated that our isothermal analysis can equally well be applied to thermal unfolding transitions monitored via intrinsic tryptophan fluorescence, and we expect this framework to apply equally for data collected using any technique for monitoring protein unfolding.

Materials and Methods

Materials

His-tagged maltose-binding protein (MBP) was expressed from a plasmid in *E. coli* and then purified through Ni-chelated Sepharose Fast Flow Resin (GE Healthcare) and HiLoad 16/60 Superdex 75 gel filtration column (GE Healthcare). The protein was exchanged into assay buffer (120 mM NaCl, 20 mM NaH_2PO_4/Na_2HPO_4, pH 7.4) by dialysis. Both carbonic anhydrases were obtained from a commercial vendor (*h*-CA I (Sigma C4396) and *b*-CA II (Sigma C2522)). All protein concentrations were determined with Quick Start^{TM} Bradford Protein Assay Kit (Bio-Rad, catalog no. 5000201).

Ligands were all obtained from commercial vendors, as follows: maltose (EMD Millipore 105910), acetazolamide (Sigma 97582), methazolamide (Sigma SML0720), sulfanilamide (Sigma 46874), trifluoromethanesulfonamide (Sigma 638455), and 4-nitrophenyl acetate (Sigma N8130).

Generating simulated unfolding data

Simulated experimental data were generated using the formulation laid out by others (139, 154) that allows the fraction of unfolded protein to be calculated at a given temperature and ligand concentration, provided a set of thermodynamic parameters that describe protein unfolding and ligand binding. Values for these thermodynamic parameters are reported in the caption of **Figure 2**. Data were calculated near the T_m value in 0.25 ºC increments.

SYPRO DSF assay: experimental protocol

All proteins (MBP, *h*-CA I and *b*-CA II) were used at a final concentration of 2 µM for this assay. SYPRO Orange (Invitrogen S6651) was used at a final concentration of 20X for MBP, and at 10X for the carbonic anhydrases. MBP experiments were carried out in 120 mM NaCl, 20 mM NaH_2PO_4/Na_2HPO_4, 1% DMSO, pH 7.4. Carbonic anhydrase experiments were carried out in 100 mM NaCl, 20 mM TRIS, 1% DMSO, pH 6.1.

All DSF experiments were carried out 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:

0.5 ºC/min for MBP, and 1 ºC/min for carbonic anhydrase. In the MBP-maltose-denaturant systems, 0.7 M guanidine hydrochloride (GdnHCl) was added into each sample and the reaction was carried out exactly as described above. Melting curves were directly exported from the instrument, and then were analyzed with Prism 6 (GraphPad Software Inc.).

SYPRO DSF assay: data analysis

DSF data were analyzed in three steps. First, raw fluorescence data as a function of temperature were fit to a modified form of the thermodynamic equation, as follows:

$$\Delta G = \Delta H\left(1-\frac{T}{T_m}\right) - \Delta C_p\left(T_m - T + T\ \ln\left[\frac{T}{T_m}\right]\right) \qquad (22)$$

$$K_U = e^{-\frac{\Delta G}{RT}} \qquad (23)$$

$$Y(T)\ = \left[\frac{1}{1+K_U}(m_F T + b_F)\right] + \left[\frac{K_U}{1+K_U}(m_U T + b_U)\right] \qquad (24)$$

Here **Equation 24** relates the observed fluorescence signal, Y, as a function of temperature (T). The two terms in this equation correspond to contributions from folded and unfolded protein, respectively. Each term consists of the fraction of folded/unfolded protein, with a term that depends linearly on temperature (due to the temperature-dependence of the dye); thus, m_F and b_F capture this dependence of the dye when the protein is folded, as observed in the baselines before the thermal unfolded transition. The fraction of folded/unfolded protein at a given temperature depends on the “effective” unfolding/folding equilibrium constant (K_U), which is dependent on both the temperature and the ligand concentration. As noted earlier, interaction between the dye and the protein is not explicitly included in this model of unfolding.

Throughout all of the analysis presented here, the value of ΔCp is held fixed. Thus, an individual thermal unfolding curve is fit using six free parameters: Tm, ΔH, m_F, b_F, m_U, and b_U. Because the temperature dependence of the dye is the same regardless of the ligand concentration, however, we found that the fitting was be improved by using a single shared global parameter for the slopes of the baselines (m_F and m_U).

From these fits, the fraction of unfolded protein can be determined at any temperature: collecting together data collected at different ligand concentrations for a single temperature of interest thus allows construction of the "isothermal" plots presented above.

To facilitate adoption of this approach, software is provided that carries out all of the analysis described herein. The software, and its associated user guide, is freely available for download via SourceForge (https://sourceforge.net/projects/dsf-fitting).

Intrinsic Trp fluorescence

MBP was used at a final concentration of 2 µM in the buffer: 120 mM NaCl, 20 mM NaH_2PO_4/Na_2HPO_4, pH 7.4 and 1% DMSO. Data were collected with sample size 800 µL, in triplicate. All experiments were carried out using Photon Technology International (PTi) spectrophotometer with 4 X 10 mm quartz cuvettes. The excitation wavelength was set to 290 nm with 1 nm light pass-width and the emission wavelength was set to 337 nm (where Trp has the highest fluorescence intensity), with 6 nm light pass-width. The sample was pre-incubated for 10 minutes prior to measurement for each different temperature, from 25 ºC to 72 ºC. Fluorescence was measured continuously for 60 seconds at every temperature, and intensity values were averaged over this interval. The average fluorescence intensity over this interval was plotted as a function of temperature to obtain the thermal unfolding curve. Based on this unfolding curve, the T_m was estimated to be 56 ºC.

This temperature was then used to measure fluorescence as a function of maltose concentration. Serial dilutions of maltose were prepared with 2 µM MBP (in the same buffer described above), and fluorescence was determined as described above. Data were analyzed via the same isothermal approach used for DSF data.

Esterase activity assay

Carbonic anhydrase activity (*h*-CA I and *b*-CA II) was measured using a spectrophotometer (Molecular Devices, SpectraMax® i3x) as described elsewhere (189). *h*-CA I was used at a final concentration of 2 μM and *b*-CA II was used at a final concentration of 0.5 μM in the buffer: 100 mM NaCl, 20 mM TRIS, pH 6.1, 1% DMSO. The substrate, 4-nitrophenyl acetate, was titrated through 0 to 3 mM from a freshly-prepared 3.2 mM stock. All reactions of *h*-CA I took place at 57 ºC and reactions of *b*-CA II took place at 60 ºC. The change in absorbance was measured at 348 nm. Enzyme initial velocity was plotted with different substrate concentration using Prism6. Data were collected for each of the four inhibitors, and the change in initial velocities were analyzed with the "Enzyme-noncompetitive inhibition" equation in Prism6.

Acknowledgements

We thank Dr. Mark Andrake for the use of the Eppendorf Realplex2 Mastercycler instrument, and for assistance and useful discussions. This work was supported by a grant from the National Institute of General Medical Sciences of the National Institutes of Health (R01GM112736). AD acknowledges funding from the Human Frontiers Science Program grant RGP0022/2017.

Chapter IV: Predicting PROTAC-mediated ternary complex formation using Rosetta

Abstract

Recent years have brought a flood of interest in developing compounds that selectively degrade protein targets in cells. These compounds have been exemplified by PROTACs (PROteolysis TArgeting Chimeras), heterobifunctional molecules that combine a target-binding warhead with an E3 ligase-recruiting moiety: if the PROTAC recruits both proteins into a ternary complex, this can induce ubiquitination and proteolytic degradation of the target protein. An important hurdle in this field, however, has been the rational design of effective PROTACs: specifically, how to identify a suitable linker between the two protein-recruiting moieties in order to enable formation of the ternary complex. Modern development of a PROTAC typically requires synthesis and evaluation many tens – or even hundreds – of candidate linker lengths, compositions, or attachment sites. Here, we describe a structure-based computational method to build models of candidate ternary complexes, and evaluate whether a given linker suitably bridges the protein-recruiting moieties. Briefly, the method entails docking the two proteins (with their respective fragments of the PROTAC in place) using the Rosetta software, then using pre-built low-energy conformations of the linkers to evaluate which of the resulting models can be spanned by the desired linker. This allows many crude models of the ternary complex to be built, and these are subsequently refined. We have applied this approach to retrospectively evaluate the activity and selectivity of PROTACs reported in the literature, which address diverse targets (Brd4^{BD1}, c-Met/EphA2/Stk10, and CDK4/CDK6) using different E3 ligases (VHL and CRBN). We find that this computational approach can indeed explain the observed activity and selectivity of these PROTACs, and further that explaining PROTAC activity is best carried out using an ensemble of structures of the ternary complex rather than a single district conformation.

Introduction

PROteolysis TArgeting Chimeras (PROTACs) are heterobifunctional small molecules containing two functional ligand-moieties: a warhead ligand binding target protein of interest (POI); and an E3 ligase recruiting ligand binding an E3 ubiquitin ligase. Combining these two functional groups through certain choices of linker can induce formation of a POI – PROTAC – E3 ligase ternary complex. This complex can then lead to ubiquitination and degradation of the target protein. PROTACs have gained increasing attention in recent years starting from its first application in degradation of the target protein in 2001 (190). In contrast with genetic methods to decrease cellular abundance of a specific protein (e.g., CRISPR/Cas9, RNAi, antisense oligonucleotides), PROTACs retain the advantages associated with traditional pharmacological approaches, including oral bioavailability, superior tissue penetrance, and better stability (23, 61, 63, 121). PROTACs also provide multiple advantages over traditional small molecule inhibitors, including: 1) low requirement of binding affinity (61, 68-70); 2) prolonged cellular effects (61); 3) no requirement to bind at the target protein's active site (23, 61, 63, 64), which facilitates addressing non-druggable targets such as scaffolding proteins (61, 64-67); and 4) the ability to simultaneously diminish (or eliminate) the activity of multiple functions spread over a multi-domain protein (61, 64, 65, 191-193).

Over the past two decades, multiple successful PROTACs have been reported with good activity and selectivity. Among these promising PROTAC molecules, most target either BET proteins (74, 76, 78, 82, 84) or enzymes (and kinases specifically) (65, 68, 70-73, 75, 77, 79-81, 83, 85), as these well-studied targets often have useful inhibitors that can be immediately adapted for use as warheads. Beyond these two classes, PROTACs have also been designed against other intriguing targets that include multi-functional proteins (e.g., TRIM24 (191), SMARCA2 (192) and tau (66, 67)), nuclear receptors (e.g., AR (194)), and HaloTag7 fusion proteins (engineered proteins to be degraded by HaloPROTACs (195)). Therefore, we can see this PROTAC method have been successfully applied for many different types of POIs.

With respect to selecting an E3 ligase, since the very beginning much of the focus has been on two candidates: Von Hippel-Lindau protein (VHL) and Cereblon (CRBN). Spurred by extreme interest in development of additional PROTACs, particularly for therapeutic applications, more potential E3 ligases have been explored recently, including MDM2 (83), IAP (81), RNF4 (196), beta-TRCP (197), parkin (197), and DCAF16 (198). As the field progresses, there is certainly reason for optimism that the availability of moieties to recruit other E3 ligases will further promote development of new PROTACs.

Even with these rapidly accumulating success stories, many crucial challenges remain before this technology can be broadly applied. Foremost among these, the underlying features that contribute to effective cellular degradation are still not established. Certainly it has become evident that high binding affinity between the PROTAC and POI does not necessarily translate to efficient degradation of the POI (68, 70). Somewhat surprisingly, the selectivity does not directly translate from the inhibitor to the PROTAC either: a non-selective kinase inhibitor can lead to a highly selective PROTAC, given precisely the right linker (70, 73, 75, 77). To explain these observations, the field has generally concluded from published crystal structures of ternary complexes (74, 76, 192) and in-cell fluorescence assays (71, 199) that the ability to form the ternary complex (POI / PROTAC / E3 ligase) may be the key determinant. That said, it is currently controversial whether cooperativity in forming the ternary complex is necessary, since both cooperative (74, 85) and anti-cooperative (76) ternary complexes have been found to induce efficient target degradation.

To optimize and explore the role of ternary complex structure formation, the linker between the warhead ligand and the E3 ligase ligand has naturally been the focus of attention (85, 200). In previous studies, a traditional but inefficient way to design PROTACs for a given pair of POI / E3 ligase is to synthesize a batch of PROTAC molecules with fixed warhead / E3 ligase ligands but many different linkers, and to test all these molecules in cell assays (78, 84). To improve the efficiency of PROTAC design, others have very recently applied computational approaches including protein-protein docking (76, 201), and also tried to use information from solved PROTAC crystal structures (192). Although these attempts have improved understanding into the role of the linker, and have provided very tentative clues for PROTAC design, there has not yet been reported a carefully-benchmarked computational approach for predicting the activity and selectivity of PROTACs.

Here, we develop and describe a computational strategy for evaluating the “fit” of a given linker in the context of a POI – PROTAC – E3 ligase ternary complex. In this approach, we combine protein-protein docking, screening of linker conformations, and analysis of protein-protein interactions associated with a given binding mode. Together these result in a collection of models for the ternary complex, which in turn allow the evaluation of a given POI / PROTAC / E3 ligase’s activity. By applying this approach to a family of linkers, one can test the effect on activity of varying the linker; conversely, by varying the POI one can test the effect on target selectivity arising from the chosen linker.

To validate this approach, we have compiled data from a series of reports describing optimization and characterization of PROTACs addressing diverse targets using multiple E3 ligases. These include Brd4^{BD1} – CRBN (76, 78), c-Met/EphA2/Stk10 – CRBN/VHL (70), and CDK4/CDK6 – CRBN (71, 79). We find first that this computational method can indeed explain the activity of related PROTACs in response to different linker lengths. Next, we find that these models not only recapitulate the selectivity of a given PROTAC with different (kinase) targets, but that they can also explain different degradation efficiencies resulting from pairing a given POI with different E3 ligases.

Computational Approach

The PROTAC molecule contains three parts: the warhead ligand which can bind to the protein of interest (POI); the E3 ligase ligand which can recruit the E3 ligase; and the linker which joins these two parts. Among these three components, the two ligands are always well studied and modified, and also both typically have strong binding affinity for their respective protein targets. Thus, the key missing piece of this PROTAC design puzzle is the linker: how to find the suitable linker to connect these two ligands leading both activity and selectivity. When the POI / warhead ligand and E3 ligase / E3 ligand are determined, the question that if a given PROTAC can selectively work or not would be simplified to predict if its linker is properly designed. To answer this question, one straightforward idea is to build a ternary structure model with this PROTAC by trying its different linker conformations. A PROTAC should be considered as a promising one if we can find a linker conformation to build a reasonable ternary structure model that has a good energy score and also shows strong interaction among the POI / PROTAC / E3 ligase. Here, we summarize our computational pipeline of building ternary structure model and model analysis below, and present the further details in the *Supporting Methods* section.

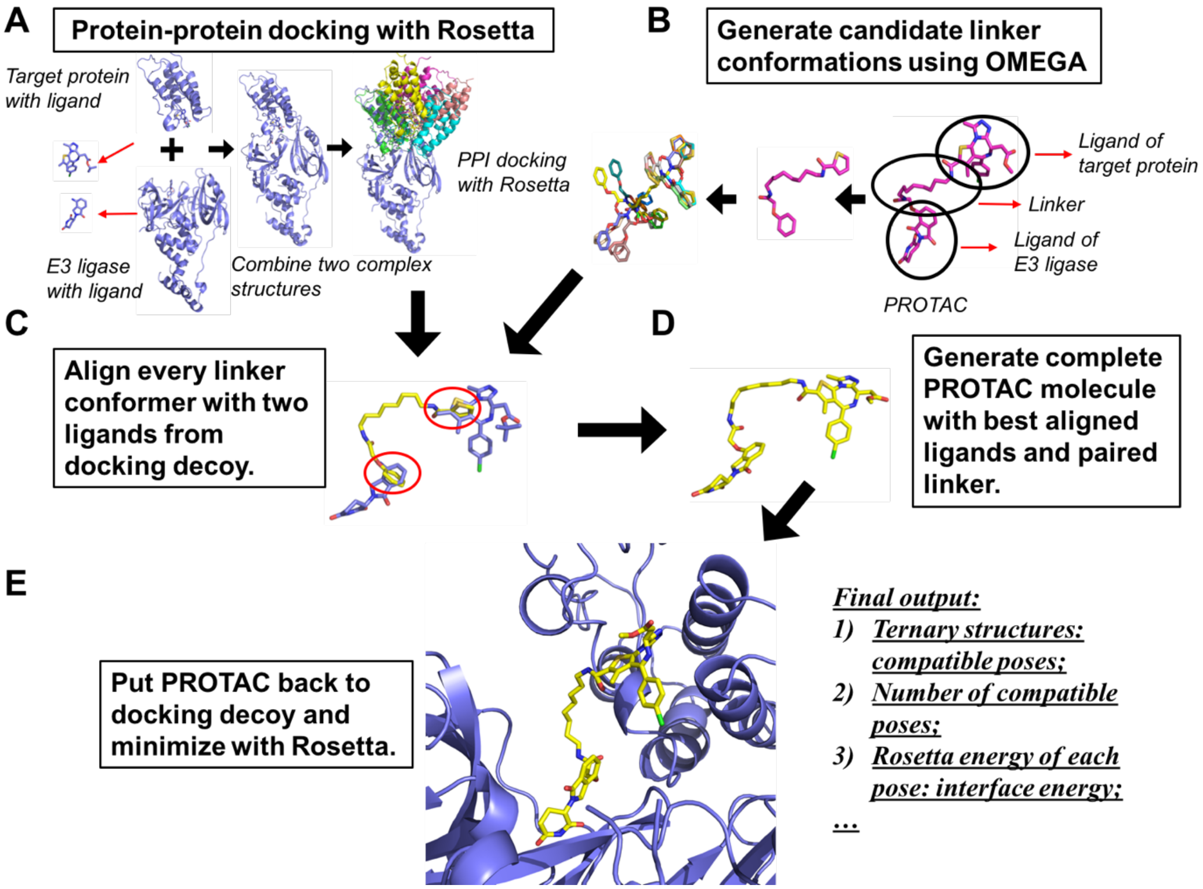

A
Protein-protein docking with Rosetta
Target protein with ligand
E3 ligase with ligand
Combine two complex structures
PPI docking with Rosetta
B
Generate candidate linker conformations using OMEGA
Ligand of target protein
Linker
Ligand of E3 ligase
PROTAC
C
Align every linker conformer with two ligands from docking decoy.
D
Generate complete PROTAC molecule with best aligned ligands and paired linker.
E
Put PROTAC back to docking decoy and minimize with Rosetta.
Final output:
1) Ternary structures: compatible poses;
2) Number of compatible poses;
3) Rosetta energy of each pose: interface energy;
…

Figure 1: Overview of the computational approach. Proteins are shown using cartoon representation, with PROTACs (or their component moieties) as sticks. **(A)** Protein-protein docking using Rosetta. **(B)** Generation of candidate linker conformations using OMEGA. **(C)** For each docking pose, the collection of linker conformations is screened in search of a compatible geometry. **(D)** If a match is found, the complete PROTAC molecule is built from the (compatible) component parts. **(E)** Because the protein-bound parts of the PROTAC have not moved when the complete PROTAC is built, it can be trivially aligned back into the proteins' binding mode to yield a model of the ternary complex.

Step 1: Protein-protein docking

To generate the ternary structure models by testing different linker conformations for a given PROTAC, we need to separate this ternary complex to three groups: POI and its cognate ligand, E3 ligase and its paired ligand, and the linker. After the separation, we download the PDB structures for the two protein-ligand complexes: the POI and ligand complex, and the E3 ligase and ligand complex (**Fig. 1A, left**). We manually combine these two complexes structures together with PyMol (202), such that the two ligands face one another (**Fig. 1A, middle**). Then we apply protein-protein docking with Rosetta software suite (203) (**Fig. 1A, right**). During the docking process, we fix the relative positions of the two ligands with their cognate proteins and only allow them move together with the proteins. Total 50,000 decoys are generated in this case. We are interested in the docking decoys that have good interaction between the POI and the E3 ligase. Therefore, we rank them with the interface score (I_sc scores from Rosetta protein-protein docking output score file) and pass the top 5,000 decoys (top 10%) to the next stage.

Step 2: Identifying complementary linker conformations

Next, we seek to identify low-energy linker conformations that appropriately join the two ligands in a given pose. To identify the linker conformer for a given PROTAC molecule, we first pre-build a library of linker conformers using OMEGA (OpenEye Scientific Software, Santa Fe, NM) (204-206). Because of the different linker lengths, the maximum available conformer numbers are different. We generated the all available conformers for a short linker with a maximum conformer number less than 1000. For the longer linkers, we generated 1000 conformers. When we generate this linker conformers library, we include not only the linker itself but also one piece from each ligand where the linker is attached, like the five-member ring from the POI ligand and the six-member ring from the E3 ligase ligand shown in **Figure 1B**. These two pieces are used for the alignment in the next step (**Fig. 1C**). To check if a linker conformer can successfully link the two ligands from a docking decoy, we align it to the two ligands from the docking decoy by calculating the optimized RMSD (root mean square deviation) of the two parts at both ends of the linker and the same two pieces from the two ligands (**Fig. 1C**). If the RMSD is less than the cutoff value, in this case we use 0.4 Å, we keep this linker / docking decoy pair and pass it to the next step.

Step 3: Building models of the ternary complex

To build the ternary structure model, we first generate the whole PROTAC molecule by combining the linker conformer and its paired two ligands from the docking decoy with the small RMSD value, coming from the previous step (**Fig. 1D**). If a whole PROTAC molecule can be built by these three pieces without any distorted bond geometries, we then put this completed PROTAC molecule back to the docking decoy and use Rosetta to minimize this ternary structure model (**Fig. 1E**). If a ternary structure model can be successfully minimized by Rosetta, we consider it as a successful model and count this binding mode as a "compatible pose" for the PROTAC of interest.

Step 4: Analysis

Overall, the approach entails building a large set of docked poses, and then distilling these down to a collection of complete models of the ternary complex. From the collection of complete models, we evaluate the following: 1) the number of complete models resulting from this pipeline; 2) the interaction energy among the three components of any given model, and 3) the interaction energy in a given model between the POI and E3 ligase (i.e., excluding contributions from the PROTAC). We will discuss and analyze the utility of these quantitative outputs in the following sections. Finally, we also note that these models contain complete structure information, and thus can be used to evaluate interaction details further in select cases.

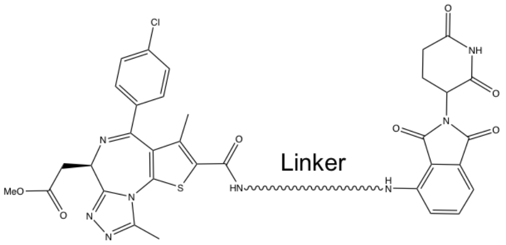
Results

To evaluate the utility of our computational approach, we turn to extensive data available in the literature describing the cellular activity and selectivity of various PROTACs. While crystal structures of a small number of PROTAC ternary complexes have been reported, the number is insufficient for thorough evaluation of our method. Further, it remains unclear whether a single static snapshot is sufficient to understand the structural basis for PROTAC activity – as will be re-visited in the *Discussion* section.

PROTAC activity versus linker length

One of the most common approaches to PROTAC design is trying a series of linkers with different lengths while the warhead ligand and E3 ligase ligand are fixed. Therefore, we started to test our approach with two such scenarios drawn from the literature (76, 78): in each case the warhead ligand and the E3 ligase ligand are fixed, and only the linker length changes.

In **Table 1**, we have five PROTAC molecules, all drawn from a single study (76). Among these five PROTACs, zxh-3-27 has the shortest linker length and didn't induce any target protein, Brd4^{BD1}, degradation in the cell assay. zxh-2-147 and zxh-2-184 have a little bit longer linker and started to show some activity, but the EC50 values of these two are worse than zxh-3-26 and dbet70, which have the second and the first longest linkers among these five PROTACs. Comparing zxh-3-26 and dbet70: they showed similar EC_{50} in cell protein degradation assay while zxh-3-26 also showed selectivity between Brd4^{BD1} and Brd4^{BD2} while dbet70 can induce degradation for both proteins, meaning low selectivity. More interestingly, the activity is better and better when the linker is longer and longer from zxh-3-27 to zxh-3-26 but this trend stops at dbet70. Furthermore, the selectivity starts to get worse at the same point. zxh-3-26 has a five-carbon linker which is shorter than the eight-carbon linker of dbet70, but zxh-3-26 has the similar EC50 as dbet70 and better selectivity than dbet70. Based on this interesting observation from the literature, we applied our approach to these five PROTACs in order to determine if our approach could identify this same trend and selectivity pattern as shown in their cell assays. We also want to explore and explain the reasons of these results.



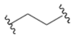
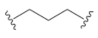
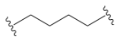
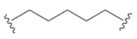
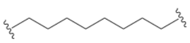
PROTAC	Linker	Conformers #*	Compatible poses #	SR(%)**	EC$_{50}$ (nM) in cellular degradation of Brd4^{BD1} ***
zxh-3-27		237	36	0.00	> 10000
zxh-2-147		606	205	0.01	10 - 100
zxh-2-184		1000	1750	0.04	10 - 100
zxh-3-26		1000	2535	0.05	1 - 10
dbet70		1000	3219	0.06	1 - 10

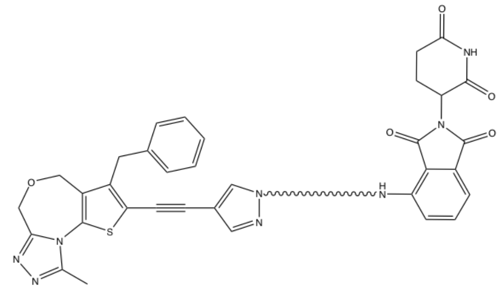
Table 1: Evaluation of the Brd4 – zxh/dbet – CRBN complexes.

*: The linker and the rings at either end are combined to generate conformers.
**: SR (%) means success rate of compatible pose generation. SR (%) = (compatible poses #) / ((docking decoys #) X (conformers #)) X 100%; docking decoys # equals 5000 in this case.
***: All the cellular data in this Table are from Nowak, R. P. *et al*, 2018

The first and the most intuitive output result we got from our prediction pipeline is the number of the successful compatible poses for each PROTAC molecule. In **Table 1**, we can see the values of this number increasing from 36 of zxh-3-27 to 2535 of zxh-3-26, along with the increase of the linker length. To account for the fact that different linkers have different numbers of conformers (e.g., zxh-3-27 has the fewest conformers because it is the shortest), we calculated the success rate (SR) of compatible pose generation: this corresponds to the number of compatible poses normalized to the total number of decoys generated (a constant) and the number of available conformers for the linker provided (not constant) (**Table 1**). After this transformation, SR still shows the same trend: zxh-3-27 has the lowest SR value, while zxh-3-26 and dbet70 have the highest SR values (**Table 1** and **Figure 2B**). This SR pattern also matches the activity results from the literature: zxh-3-27 has the worst EC50 and zxh-3-26 and dbet70 have the best EC50. We then thought there might be a positive correlation between these SR values and the activity of PROTAC molecules.

To check whether the compatible poses we generated from this approach were reasonable or not, we also went through the top compatible poses, which have the best minimization scores from Rosetta. In **Figure 2A**, we can see one of the best compatible poses of zxh-3-26. In the zoom in view of this ternary structure model (**Figure 2C**), we can see the target protein ($Brd4^{BD1}$) shows good interaction with the E3 ligase (CRBN) with several hydrogen bonds forming at the interaction face. Based on the good energy score and the interactions on the model interface, we are more confident about the ternary structure models we generated through our pipeline. These observations also provide some clues that the protein-protein interaction between POI and E3 ligase may contribute in the formation of the ternary structure and promote this process.

To test the idea of the SR value versus the activity of the PROTAC and the protein-protein interaction during the ternary structure formation, we applied our method to the second group of PROTAC molecules with different linker lengths (**Table 2**). These four PROTACs are from a separate study (78), and again the only difference is the linker length: from two carbons to five carbons. The SR values of these four PROTACs are shown in **Table 2** and **Figure 2E**. The SR values increased from 0.05% of PROTAC-28 with a two-carbon linker to 0.54% of PROTAC 30 with a four-carbon linker and these results match the cell data from the literature. The interesting observation is the decrease of the SR values between the PROTAC 30 and PROTAC 31. PROTAC 31 has a five-carbon linker which is longer than PROTAC 30, but the SR value is worse. We thought this may be because it is easier to induce the clashing during the ternary structure formation process when the linker is longer than some certain value as longer linker would require more space while the space between two proteins is limited. We also went into details of ternary structure models (**Figure 2D** and **2F**), and observed the good protein-protein interactions in this structure model as described in the previous case.



PROTAC	Linker	Conformers #*	Compatible poses #	SR(%)**	IC_{50} (nM) in cell growth inhibition***
28		46	116	0.05	0.47
29		102	1414	0.28	0.138
30		241	6561	0.54	0.0654
31		618	3052	0.10	0.092

Table 2: Evaluation of the Brd4 – 28/29/30/31 – CRBN complexes.

*: The linker and the rings at either end are combined to generate conformers.
**: SR (%) means success rate of compatible pose generation. SR (%) = (compatible poses #) / ((docking decoys #) X (conformers #)) X 100%; docking decoys # equals 5000 in this case.
***: All the cellular data in this Table are from Qin C. *et al*, 2018, here the cell line is MV4; 11

Combining the results of these two groups of the PROTAC molecules, we suspected there might be a positive correlation between the SR values and the activity of the PROTAC molecules. We put forward this hypothesis because SR values reflect the success rate of building a ternary structure model for a given PROTAC, and the higher the value is, the easier to build a model. The easier a model can be generated might reflect it will be also easier to form the ternary structure in the real case. Furthermore, the formation of the ternary structure is related to the activity of PROTAC. All these together would suggest that the SR value may be positively correlated to the PROTAC activity. For future applications, we therefore plan to use 0.05% as a standard cutoff value (**Figure 2B** and **2E**). This cutoff value would be tested multiple times in the following parts. As we also observed the good protein-protein interactions in both cases, suggesting that this may need to be considered as an essential part in identifying whether a PROTAC will have promising activity. This hypothesis will be explored further in the *Discussion*.

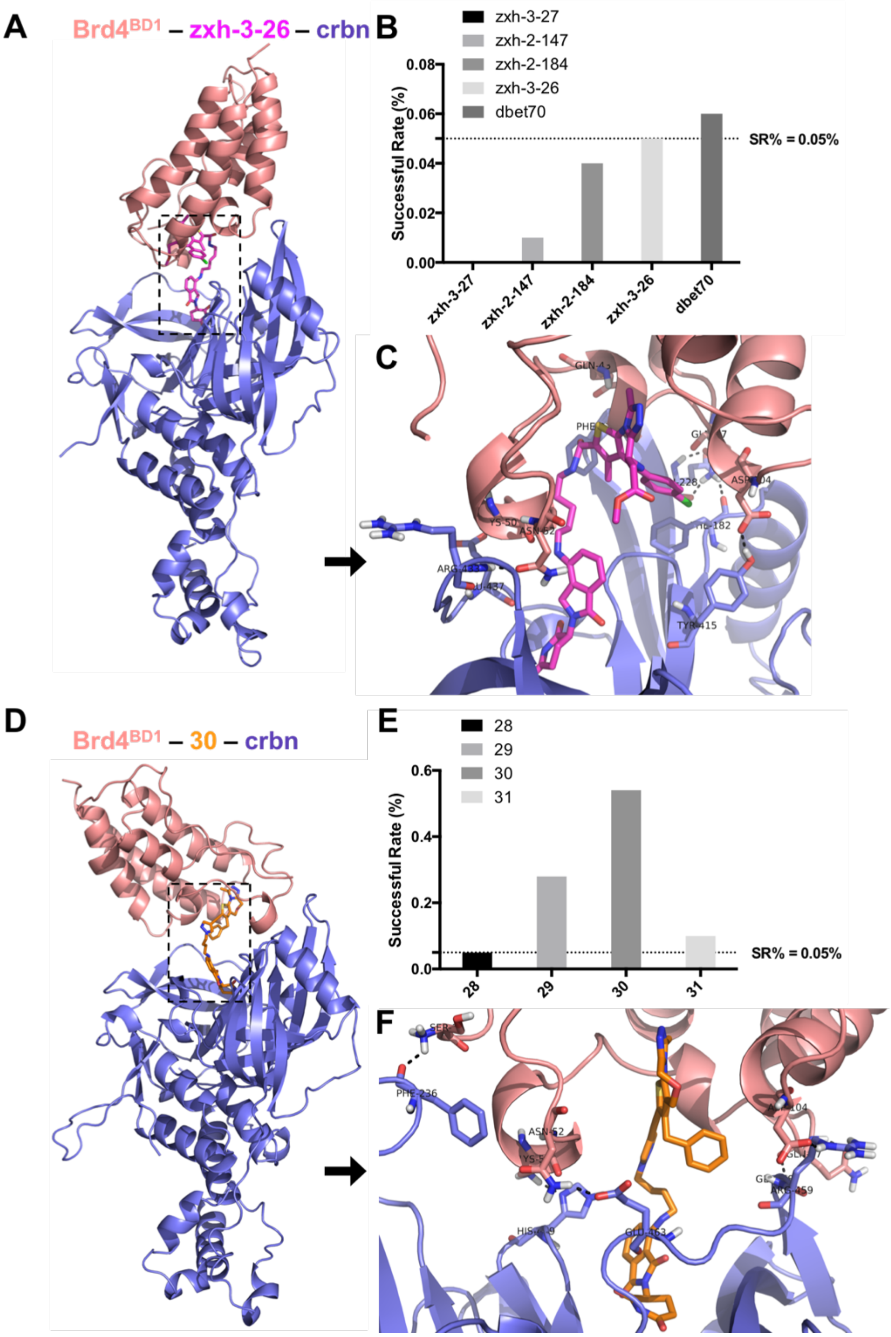


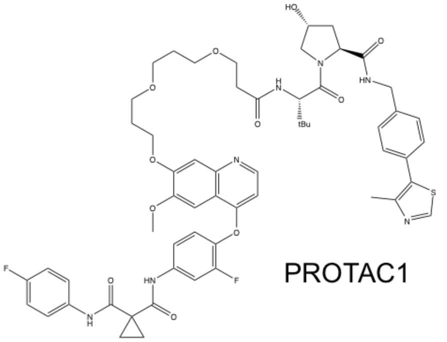
Figure 2: Top compatible poses of PROTACs inducing a ternary complex for degradation of Brd4BD1 by CRBN. (A) Model of the Brd4BD1/zxh-3-26/CRBN ternary complex. **(B)** Success rate of compatible poses for the zxh/dbet series of PROTACs. **(C)** Zoomed-in view of the Brd4BD1/zxh-3-26/CRBN interface from our model. **(D)** Model of the Brd4BD1/30/CRBN ternary complex. **(E)** Success rate of compatible pose generation for the 28-31 series of PROTACs. **(F)** Zoomed-in view of the Brd4BD1/30/CRBN interface from our model.

PROTAC activity and selectivity prediction of c-Met/EphA2/Stk10 – PROTAC1/2 – CRBN/VHL

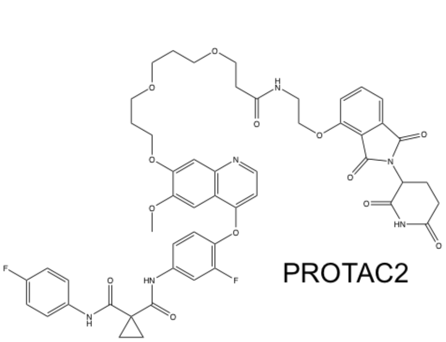
We next applied our approach to predict the selectivity with a more complicated group of PROTAC molecules: c-Met/EphA2/Stk10 – PROTAC1/2 – VHL/CRBN (70). These two PROTACs (PROTAC-1 and PROTAC-2) were designed to recruit VHL and CRBN, respectively. The warhead ligand in this case is a well-known multi-kinase inhibitor, foretinib. This warhead ligand can bind all three kinases studied here (c-Met, EphA2 and Stk10), and all three inhibitor/kinase complex structures are available in PDB (**Figure 3A**). At the outset, one might expect PROTAC-1 and -2 to induce degradation of all these three target kinases by recruiting different E3 ligases. From the experimental data shown in **Table 3**, however, we see that PROTAC-1 only degrades c-Met, and PROTAC-2 degrades both c-Met and EphA2 but not Stk10. Thus, both PROTACs have selectivity for different kinases even though the warhead itself does not. We want to test our computational approach with this interesting and challenging case. We hope our computational approach can successfully predict the selectivity of these two PROTAC molecules and can explain the reasons of this interesting experimental observation.

As described in the two previous cases, we first calculated SR values of all these six kinase / PROTAC / E3 ligase pairs and we were very excited to find that SR values match the cell data very well if we used 0.05% as the cutoff number (**Table 3** and **Figure 3B**). The three pairs that worked in the cell assay are c-Met / PROTAC-1 / VHL, c-Met / PROTAC-2 / CRBN and EphA2 / PROTAC-2 / CRBN, and all and only these three pairs have the SR values higher than 0.05%. In the ternary structure models (**Figure 3C** and **3D**), we can observe multiple hydrogen bonds on the kinase-E3 ligase interfaces of all these three ternary structure models.

As the whole PROTAC molecule is the same for all three different kinases, including the linker part, and the warhead ligand doesn't have the selectivity, this selectivity may come from the linker and the protein-protein interaction between the POI and the E3 ligase. Our computational approach can successfully catch this information: our method not only predicted that PROTAC-1 can only induce the degradation of c-Met and PROTAC-2 can only work for c-Met and EphA2, but also predicted that EphA2 can only be degraded by CRBN but not VHL. This makes the results more reasonable and reliable as this EphA2 case can confirm the importance of protein-protein interaction: EphA2 may only have the good interaction with CRBN but not VHL. As our predictions are based on the protein-protein docking, our pipeline can extract and keep the information of protein-protein interaction from the docking step. The further analysis of this hypothesis will be described in the *Discussion*.



PROTAC1



PROTAC2

Target – PROTAC - E3 ligase	Conformers #*	Compatible poses #	SR(%)**	Protein level remaining upon PROTAC treatment (%)***	
				100 nM PROTAC	1 μM PROTAC
cmet – PROTAC1 - vhl	1000	7858	0.16	70	62
epha2 – PROTAC1 - vhl	1000	795	0.02	93	96
stk10 – PROTAC1 - vhl	1000	1625	0.03	100	99
cmet – PROTAC2 - crbn	1000	2285	0.05	71	65
epha2 – PROTAC2 - crbn	1000	3586	0.07	71	59
stk10 – PROTAC2 - crbn	1000	239	0.00	99	102

Table 3: Evaluation of the c-Met/EphA2/Stk10 – CRBN/VHL PROTAC-1/2 complexes.

*: The linker and the rings at either end are combined to generate conformers.
**: SR (%) means success rate of compatible pose generation. SR (%) = (compatible poses #) / ((docking decoys #) X (conformers #)) X 100%; docking decoys # equals 5000 in this case.
***: All the cellular data in this Table is from Bondeson, D. P. *et al*, 2018.

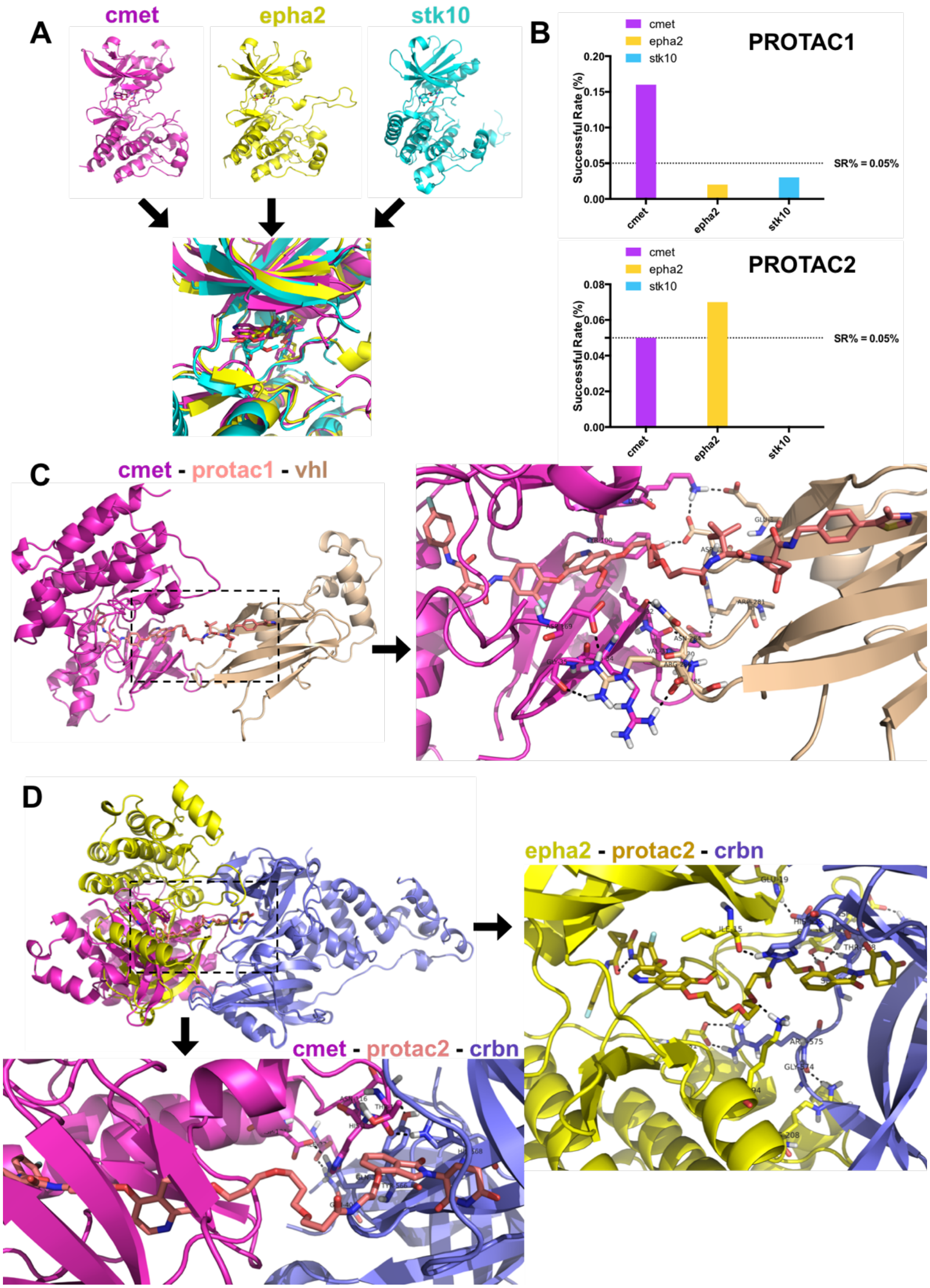

A
cmet
epha2
stk10
B
PROTAC1
cmet
epha2
stk10
Successful Rate (%)
0.20
0.15
0.10
0.05
0.00
SR% = 0.05%
cmet
epha2
stk10
PROTAC2
cmet
epha2
stk10
Successful Rate (%)
0.08
0.06
0.04
0.02
0.00
SR% = 0.05%
cmet
epha2
stk10
C
cmet - protac1 - vhl
D
epha2 - protac2 - crbn
cmet - protac2 - crbn

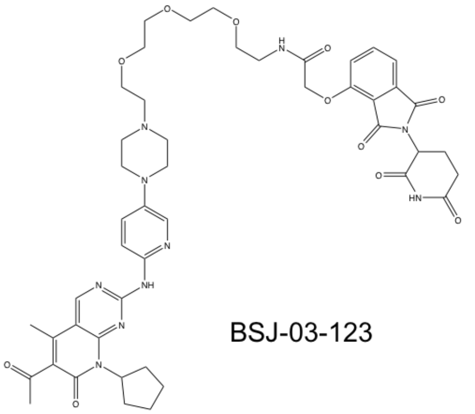
Figure 3: Modeling results of c-Met/EphA2/Stk10 in complex with PROTAC-1/2. (A) c-Met, EphA2 and Stk10 crystal structures solved with the same inhibitor, foretinib. **(B)** Success rates for each PROTAC (with the corresponding E3 ligase) with all three kinases. **(C)** Top compatible pose of c-Met/PROTAC-1/VHL. **(D)** Top compatible poses of c-Met and EphA2 complex with PROTAC-2/CRBN.

After successfully our approach with these kinases, we decided to pursue and even more difficult challenge for which we selected CDK4/6 as the target proteins. Even though CDK4 and CDK6 are 94% identical in their ATP binding pocket (71) there are several PROTAC molecules that selectively to target CDK6 (71, 79), including BSJ-03-123 and PROTAC 6.

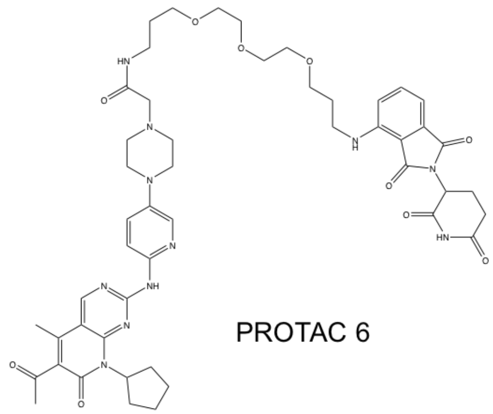
PROTAC selectivity prediction of CDK4/CDK6 – BSJ-03-123/PROTAC 6 – CRBN

Our principle objective was to determine whether our method can predict the selectivity of these two PROTACs against CDK6 over CDK4 and explain the possible reasons. We calculated the SR values for each of these four systems: CDK6 / BSJ-03-123 / CRBN, CDK4 / BSJ-03 / 123 / CRBN, CDK6 / PROTAC 6 / CRBN and CDK4 / PROTAC 6 / CRBN. In our initial hypothesis, we would expect to see the higher SR values, larger than 0.05%, for both CDK6 systems and the small SR values for the other two CDK4 systems. From the data shown in **Table 4**, the SR values don't match our expectation and the SR values of CDK4 are even higher than the values of CDK6 systems (**Figure 4B**). These results can infer that we only successfully predict the positive cases. Our method predicted that these two PROTAC molecules would have activity for CDK6; but our method didn't give the correct prediction of the negative cases. These two PROTAC molecules would also have activity for CDK4 in our prediction, which would not match the experimental observations (**Table 4**).

Our next hypothesis would be that the two ternary structures of CDK6 should have good protein-protein interactions but the other two of CDK4 would not have these good interactions, or not as good as the ones of CDK6 systems. In **Figure 4C** and **4D**, we can see that both CDK4 ternary structures seem to show good protein-protein interactions with CRBN, just like CDK6 ones, which means our hypothesis still did not fit this situation. Even though neither of our first two hypotheses fit the real data at this point, we did notice something interesting in these top ternary structure models: the CDK4 poses were different from CDK6 poses for both PROTAC molecules (**Figure 4C** and **4D**). We know that CDK4 is very similar to CDK6, both sequence and structure. If the E3 ligase and the PROTAC is the same, we would expect to get the very similar ternary structure models for both CDK4 and CDK6 coming from our prediction pipeline. In reality though, we obtained totally different poses for CDK4 and CDK6 when we aligned with CRBN. We wondered why this situation would happen and we thought this may explain the selectivity: the CDK6 poses made better protein-protein interactions than the CDK4 poses. Put another way, we suspected that perhaps simply searching for PROTAC-compatible poses was missing information from the preferred conformations (and thus the underlying energetics).



BSJ-03-123



PROTAC 6

CDK – PROTAC - CRBN	Conformers #*	Compatible poses #	SR(%)**	EC_{50} (nM) in cellular degradation of CDK4/6 ***
CDK6 – BSJ-03-123 - CRBN	1000	3557	0.07	< 50
CDK4 – BSJ-03-123 - CRBN	1000	11601	0.23	> 1000
CDK6 – PROTAC 6 - CRBN	1000	3610	0.07	10 - 100
CDK4 – PROTAC 6 - CRBN	1000	5169	0.10	> 5000

Table 4: Evaluation of the CDK4/CDK6 – BSJ-01-123/PROTAC 6 – CRBN complexes.

*: The linker and the rings at either end are combined to generate conformers.
**: SR (%) means success rate of compatible pose generation. SR (%) = (compatible poses #) / ((docking decoys #) X (conformers #)) X 100%; docking decoys # equals 5000 in this case.
***: The cellular data in this Table for BSJ-03-123 is from Brand M. *et al*, 2019. The cell data of PROTAC 6 in this Table is from Rana S. *et al*, 2019.

To check if our top compatible poses indeed match those with the best protein-protein interactions, we compared our top ternary structure models with top docking decoys. We removed the ligands from the docking decoys and did minimization for each of these 5000 decoys in each system (CDK6/CDK4 with BSJ-03-123/PROTAC 6 and CRBN) and rank them with the Rosetta minimization scores. As only the two proteins were kept in the decoys, the top decoys after ranking theoretically would be the ones with the best protein-protein interaction. We then calculated RMSD of our final top ternary structure models, top 1% in this case, and these top docking decoys, also top 1% in this case. The low RMSD values mean that the poses of the top ternary structure models match the poses of the top docking decoys and keep the best protein-protein interactions. After plotting these RMSD values in the distribution histograms (**Figure 4E**), we found that the top ternary structure models clearly have a good overlapping area (RMSD < 2 Å) with the top docking decoys in CDK6 cases for both BSK-03-123 and PROTAC 6; while this overlapping area disappeared in CDK4 cases, for both two PROATC molecules. This result confirmed our hypothesis that the CKD4 ternary models didn't achieve the best protein-protein interactions. Therefore, this observation may also explain the selectivity of these two PROTAC molecules between CDK6 and CDK4. We will re-visit this idea further in the *Discussion*.

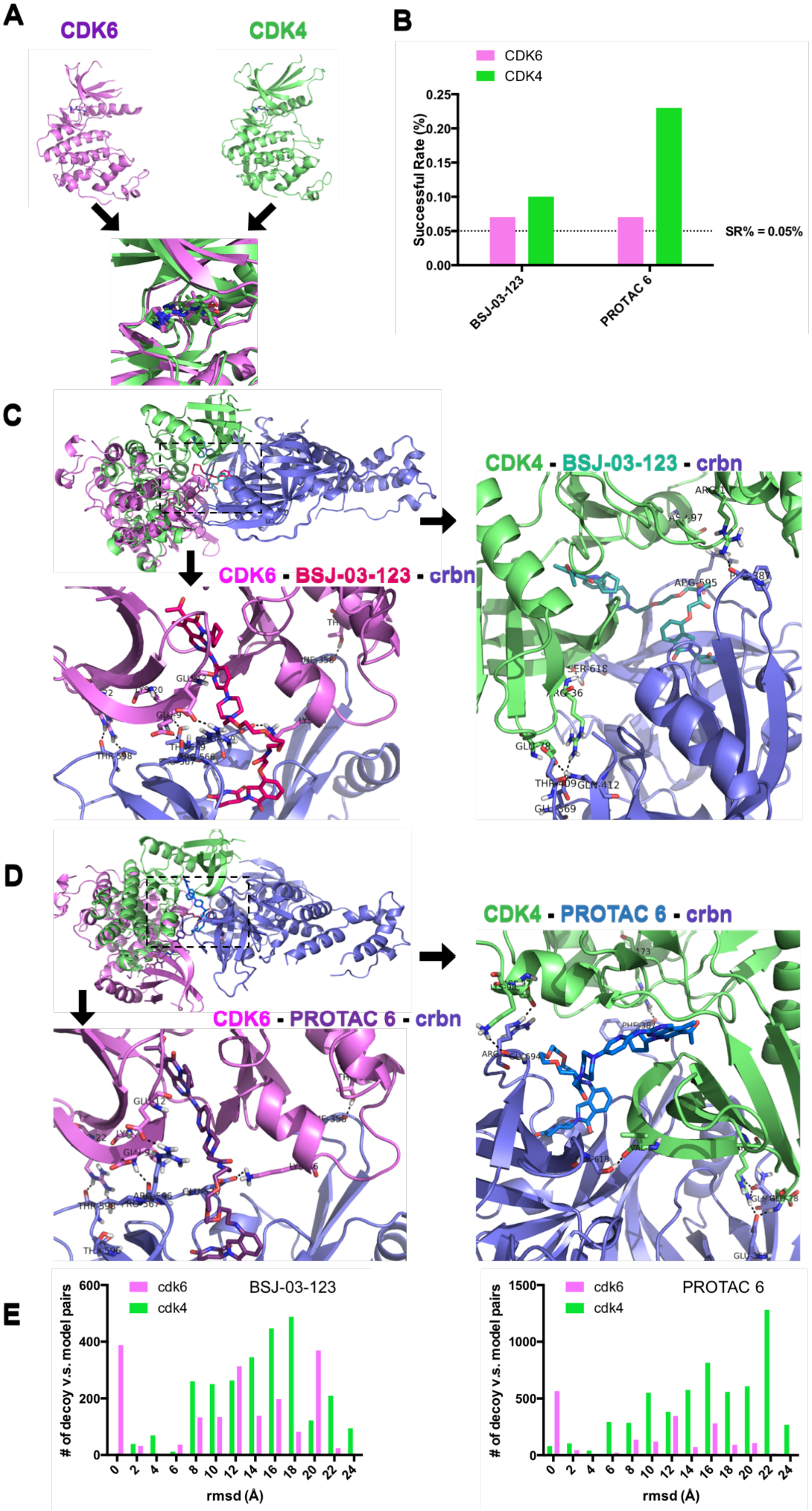

A
CDK6
CDK4
B
CDK6
CDK4
Successful Rate (%)
0.25
0.20
0.15
0.10
0.05
0.00
SR% = 0.05%
BSJ-03-123
PROTAC 6
C
CDK4 - BSJ-03-123 - crbn
CDK6 - BSJ-03-123 - crbn
ARG-595
SER-618
GLU-78
THR-409
GLN-412
THR-598
PHE-358
D
CDK4 - PROTAC 6 - crbn
CDK6 - PROTAC 6 - crbn
PHE-38
GLU-12
PHE-358
THR-598
PRO-567
GLU-78
E
cdk6
cdk4
BSJ-03-123
PROTAC 6
# of decoy v.s. model pairs
600
400
200
0
1500
1000
500
0
2
4
6
8
10
12
14
16
18
20
22
24
rmsd (Å)

Figure 4: Modeling results of CDK4/CDK6 in complex with BSJ-03-123/PROTAC 6 and CRBN. **(A)** Starting structures of CDK4 and CDK6 in complex with the same inhibitor, palbociclib. **(B)** Success rates of compatible pose generation for each of CDK4 and CDK6 with BSJ-03-123 and PROTAC 6. **(C)** Top compatible poses of CDK6 and CDK4 with BSJ-03-123 and CRBN. **(D)** Top compatible poses of CDK6 and CDK4 with PROTAC 6 and CRBN. **(E)** RMSD distribution of top docking decoys versus top compatible poses, for each PROTAC.

Discussion

Crystal structure: the only true answer or just one possible snapshot?

In our computational approach, instead of using the existed crystal ternary structures as benchmark, we chose to validate our prediction with the cell protein degradation data directly. The reason underlying this choice was the concern that crystal structures may be biased in some cases, either by presenting just one out of a collection of relevant PROTAC-compatible poses, or instead by showing an artificial ternary complex stabilized by crystal packing. This idea is also mentioned in previous literature (76, 201). As monitored by the real-time fluorescence assay (199), the formation of the ternary structure is a dynamic process, the crystal structure may just be one possibility or snapshot of the ternary structure but not the only true answer. Therefore, we prefer to use the cell protein degradation data as the validation, but not to calculate the RMSD between the computational models from our approach and the crystal structures.

Top decoy versus top ternary structure pose

In the CDK4/CDK6 case, we introduced a sequential step following SR (success rate) value comparison: calculate RMSD of our top ternary models versus the top docking decoys. The docking decoys are minimized only including the two proteins and ranked by the minimization interface scores reflecting the protein-protein interactions meaning the top docking decoys have the best protein-protein interactions. The ternary structure models are minimized including both proteins and the PROTAC molecules, and ranked by the minimization interface scores containing both protein-protein interaction energies and proteins-PROTAC interaction energies. Thus, a successful model of the ternary structure should not only have the top score as a ternary complex, but also use the best possible protein-protein interaction (i.e. leading to a similar pose as that found from docking without the PROTAC). If a ternary structure model only has the top score as a ternary complex but doesn't not include the top docked decoy, we suspect there is a risk that it may not truly be PROTAC-compatible pose, because the protein-protein interaction pose is not sufficiently favorable.

In **Figure 4E**, we see the CDK6 top ternary models of the two different PROTAC molecules both show good overlaps with their top docking decoys, low RMSD values. On the other hand, the top ternary models built for CDK4 along with the same two PROTAC molecules did not show this overlap. We propose that this may explain the CDK6 selectivity of these two PROTAC molecules.

We also applied this idea to the Brd4$^{BD1/2}$ / zxh-3-26/dbet70 / CRBN system. Based on the cell data (**Table 1**), zxh-3-26 and dbet70 both can induce degradation of Brd4^{BD1}. Furthermore, zxh-3-26 also has good selectivity for Brd4^{BD1} versus Brd4^{BD2}, while dbet70 doesn't. From the SR values, we can see neither zxh-3-26 nor dbet70 appeared to be selective (**Figure S1A**). Thus we calculated the RMSD of the top ternary models and the top docking decoys. Unexpectedly, none of the Brd4$^{BD1/2}$ / zxh-3-26 or dbet70 / CRBN pairs shows good overlap between the top ternary structure models and the top docking decoys (**Figure S1B** and **S1C**). The reason for these results may be because Brd4$^{BD1/2}$ both have anti-cooperative interactions with CRBN. This anti-cooperative interaction may lead to a more complicated situation, in which the linker is forced to contribute more in the process of forming ternary complex structure, as compared with scenarios in which the POI / E3 ligase already have some natural affinity for one another, leading to a cooperative interaction.

Cooperativity in the ternary complex

Protein-protein interaction between POI and E3 ligase should be considered as a factor in the ternary complex structure formation process, but neither critical nor alone. As described in the previous studies, both cooperative (74, 85) and anti-cooperative (76) ternary complexes have been observed in different PROTAC systems. Although both cooperative and anti-cooperative protein-protein interactions can work, the roles of protein-protein interaction and linker may slightly different in these two different situations. In the case that POI and E3 ligase have a cooperative interaction, this cooperativity may contribute more to promote the formation of the ternary structure and reduce the effort of the linker. Under this situation, the proper linker may be chosen because it can fit the best POI and E3 ligase interaction pose and let the cooperativity between two proteins can contribute most. On the other hand, when the POI and E3 ligase engage in an anti-cooperative interaction, there may be a different requirement for the linker choice. In this case, the linker may need to contribute more so that the ternary structure can still form even with the anti-cooperative POI / E3 ligase interaction.

Direct contributions of the linker

As we described above, we need to consider more about the linker when POI and E3 ligase interaction is anti-cooperative, like Brd4$^{BD1/2}$ / zxh-3-26 or dbet70 / CRBN. To promote the ternary structure formation in such a case, the linker may also need to contribute to the interface energy of the ternary complex. To test this hypothesis, we calculated the interface energies of the top 1% ternary structure models, including the PROTAC molecules. We also calculated the interface energies of these top models but without the PROTAC molecules and compare the difference between these two energies. In **Figure S1D** and **S1E**, the negative energy difference reflects the interface energy is better when the PROTAC molecule is included meaning the linker or the PROTAC molecule can benefit the formation of the ternary structure. A positive value would refer to a punishment coming from the linker or the PROTAC molecule during the ternary structure formation. If we compare the effects of zxh-3-26 / Brd4^{BD1} to zxh-3-26 / Brd4^{BD2}, we observe that this PROTAC linker confers benefit with the interface energy more in Brd4^{BD1} than Brd4^{BD2} (**Figure S1D**). In contrast, there was no difference observed between the dbet70 / Brd4^{BD1} and dbet70 / Brd4^{BD2} interactions resulting from linker contributions (**Figure S1E**). We propose that this difference may explain the selectivity of zxh-3-26, and the lack of selectivity from dbet70.

Conclusions

We report here the development of a new computational pipeline for structure-based modeling of PROTAC-mediated ternary complexes. Through the resulting models, we demonstrate that it is possible to retrospectively explain relationships between linker length/composition and cellular activity. Further, these models can also explain the surprising observation that the target-selectivity of a given PROTAC is not simply transferred from the target-binding warhead: rather, interactions with the E3 ligase in the ternary complex can shift the selectivity of the PROTAC. Moving forward, we expect that this approach can certainly be used to facilitate design of new PROTACs: by computationally screening small libraries of candidate linkers, we envision prioritizing a very small number of compounds for synthesis and characterization. Given the current landscape, which relies on synthesis of extensive collections of candidate PROTACs in order to screen for those yielding at least some target degradation, we are confident that if successful, this new approach will strongly contribute to many more groups designing PROTACs for their diverse targets of interest.

Acknowledgements

We thank Palani Kirubakaran for providing the model of inhibitor-bound CDK4. 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 (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 V: Conclusions and Future Work

The overarching goal of my PhD thesis work was to develop a general and robust approach for rationally designing inhibitors of RNA-binding proteins (RBPs). With this motivation in mind, we proposed the "RNA mimicry" approach and demonstrated its feasibility by applying it to inhibitor design of Musashi proteins: Msi1 and Msi2. Starting with the *de novo* inhibitors coming from our initial RNA mimicry approach, we optimized them based on SAR studies built on fluorescence polarization competition assays and differential scanning fluorimetry (DSF). To take the advantage of the efficiency of DSF while determining binding constants, we developed a novel isothermal analysis approach to calculate binding affinities from DSF data. After several rounds of optimization of our Musashi inhibitors, we struggled to achieve further improvements in the compounds' potency. In the long term, we envision improving these compounds further by building them into PROteolysis TArgeting Chimeras (PROTACs). As a first step towards this goal, we developed a computational pipeline for modeling PROTAC ternary complexes, to help predict the efficacy and selectivity expected for a given PROTAC molecule. Ultimately, in addition to building and testing new PROTACs based on the current Musashi inhibitors, in the future we look forward to applying this pipeline for rationally designing inhibitors of other RBPs as well.

Rationally design selective inhibitors of Musashi

With our "RNA mimicry" approach, we successfully designed single-digit micromolar inhibitors of Musashi1 and Musashi2, that are also selective (at least, in the studies we have carried out to date). We applied our "RNA mimicry" approach and initially obtained compound R12; our initial screening confirmed its inhibition of Msi1. After a first round of optimization, we demonstrated the binding mode of R12 and its derivatives (R12-8): these inhibitors bind the same position as Msi1's cognate RNA, and they mimic the interaction of Msi1-RNA. To improve potency further, we went through a couple more rounds of optimization and ended with the single-digit micromolar inhibitor R12-8-44-3. In keeping with our initial expectation, we find that R12-8-44-3 is a dual inhibitor of both Msi1 and Msi2, showing similar binding constants for both. At the same time, explored the selectivity of R12-8-44-3 using both computational prediction and *in vitro* experiments. We generated a pharmacophore library by extracting these from all structures of RBP-RNA complexes from the PDB, and screened R12-8-44-3 against this library. This computational experiment implied that R12-8-44-3 would be selective on the basis of pharmacophore matching, but also suggested a potential off-target interaction with hnRNP A1. We explicitly tested this in a biochemical assay, and confirmed that R12-8-44-3 is indeed selective for Msi1/Msi2 over hnRNP A1. Our hypothesis is that optimization of the R12-series for Msi1/Msi2 may have provided late-stage selectivity beyond that recognized through our pharmacophoric matching experiment.

In the process of designing R12-8-44-3, several opportunities became apparent for how we might improve this approach in future. The first is the size of the compound library: at the start of these studies we used the ZINC database (103), which contained ~7 million commercially-available compounds at the time of our screen. This initial screen did not yield any compounds that fully matched all the chemical moieties of the template hotspot pharmacophore; perhaps relatedly, the only active compound from this initial screen (R12) showed both poor solubility and weak potency. We believe that the limited library size may be the reason, and therefore switched to a much larger library (Enamine), for our optimization. The Enamine library was not available when we began this project, but we expect that searching this much larger collection (currently 11 billion compounds) may yield more promising starting points. Second, we note that we carried out our selectivity studies only once we had reached the final inhibitor described in these studies, R12-8-44-3. In retrospect, though we did ultimately find this compound to be selective (to the extent tested), characterizing selectivity at every iteration of our SAR studies may lead optimization along a different path.

Having thoroughly used this pipeline and optimized the resulting compounds, we are confident about the future utility of this approach for targeting other RBPs. We expect that these may prove useful both as chemical probes for exploring the biological role of RBPs, and also perhaps as a starting point for developing new therapeutics.

Application of isothermal analysis approach

The isothermal analysis approach for DSF data was developed not only for our Musashi project, but also with an eye towards studying many other ligand / protein binding interactions. During development of the approach, we tested this novel analysis method with multiple different protein/ligand systems including maltose/MBP and multiple inhibitors of carbonic anhydrase (isoforms I and II). We demonstrated that this isothermal approach can be applied to characterize diverse ligand / protein interactions, and works well over a large range of binding affinities: from sub-micromolar level to millimolar level. Using simulated data, we further evaluated the accuracy and robustness of this isothermal analysis. Through these studies, we found that our isothermal analysis can provide highly accurate binding constant and can also tolerate extremely high amounts of noise in the data (much higher than in a realistic experimental setting).

Our isothermal analysis approach is predicated on three important conditions: 1) protein/ligand has a 1:1 stoichiometry; 2) ligand binding stabilizes the protein; and 3) protein folding/unfolding and ligand binding/unbinding are both two-state and reversible. These conditions do cause limitations in certain real-world cases. Already others have contacted us for help in scenarios in which unfolding is partially non-reversible: for these, the lack of saturation in the unfolding curves (as a function of ligand concentration) may be addressed by carrying out the analysis at lower temperature (while requiring that the temperature is still in the range of the protein unfolding transition). Another strategy can be using faster heating rates, which make the process more non-equilibrium but can reduce the extent of protein aggregation (182, 207).

In the future, we are eager to optimize this isothermal approach to make it suit wider situations, including when the ligands behave like a destabilizer instead of a stabilizer, or ligand / protein is not 1:1 stoichiometry binding. For example, the top compound from among our designed Musashi inhibitors is not a stabilizer, which prevents us from using this method to determine its binding affinity. With this motivation in mind, we are exploring ways to make this approach more general, and also considering separate analysis frameworks that could be used for these alternate cases.

Pipeline of PROTAC prediction

Our PROTAC prediction pipeline yields computational results that are consistent with multiple different systems: Brd4^{BD1} – CRBN (76, 78), cmet/epha2/stk10 – CRBN/VHL (70), and CDK4/6 – CRBN (71, 79). Through studies of each of these systems, we found that our PROTAC prediction approach gives ternary structure models that explain the activity and selectivity for a given POI / PROTAC / E3 ligase system. In the Brd4^{BD1} – CRBN cases, we selected two groups of PROTAC molecules with different linker lengths (from two different papers) and were able to rationalize the different cellular activities of these compounds. In the cmet/epha2/stk10 – CRBN/VHL and CDK4/6 – CRBN cases, we challenged our simplest approach with selectivity predictions. In the cmet/epha2/stk10 – CRBN/VHL case we found that our simplest method did match the reported cellular activities, but it did not match in the CDK4/6 – CRBN case. In this latter case we found that additional considerations of the protein-protein interaction were necessary to explain selectivity of these PROTACs.

Despite the success of our method in these early benchmarks, there still remains ample opportunity to improve this approach. Additional features from the structural models could certainly be incorporated into these predictions, but at present further development is limited by the lack of availability of large datasets describing activity (and especially ternary complex formation) for many diverse PROTACs acting on many diverse protein targets. Given enough data, we ultimately envision using this structure-based modeling to analyze conformations of tertiary complexes, and then using very simple machine learning approaches to provide a way of combining these features with appropriate importance. We have recently begun to acquire much more data describing both cellular activity and ternary complex formation, which we hope will enable these improved predictions in future.

Future work

In the short term, we are eager to discover the cellular effects of treating cells with our top Musashi inhibitor, R12-8-44-3. We envision testing R12-8-44-3 in multiple cell lines, to probing effects of both

Msi1 and Msi2, and using multiple read-outs. In parallel, we envision additional optimization of this compound. Our primary target is now the linker: R12-8-44-3 includes an ester linker which may not be stable enough for future animal studies, and therefore we seek to identify more stable linker. Amides or ethers may be good options to investigate. Further, while the initial SAR studies we describe here were dramatically enabled by the availability of the Enamine collection, these nonetheless did impact the chemical space that we probed. Moving instead to custom-synthesis at this point in the project will slow the pace at which we can test new ideas, but may allow for development of improved compounds.

In the long term, we are excited that now having each of these three pieces in place will enable the ability to start with an arbitrary RBP, and use rational design to develop and validate warheads and then ultimately PROTACs addressing this exciting and non-traditional class of target.

References

1. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J, Hentze MW. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell. 2012;149(6):1393-406. doi: 10.1016/j.cell.2012.04.031. PubMed PMID: 22658674.

2. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nature Reviews Genetics. 2014;15(12):829.

3. Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. Nature Reviews Molecular Cell Biology. 2018;19(5):327.

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

5. Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, Wyler E, Bonneau R, Selbach M, Dieterich C, Landthaler M. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular cell. 2012;46(5):674-90. doi: 10.1016/j.molcel.2012.05.021. PubMed PMID: 22681889.

6. Masliah G, Barraud P, Allain FH-T. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. Cellular and Molecular Life Sciences. 2013;70(11):1875-95.

7. Auweter SD, Oberstrass FC, Allain FH-T. Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic acids research. 2006;34(17):4943-59.

8. Kapeli K, Yeo GW. Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. Frontiers in neuroscience. 2012;6:144. doi: 10.3389/fnins.2012.00144. PubMed PMID: 23060744; PMCID: 3462321.

9. Khalil AM, Rinn JL. RNA-protein interactions in human health and disease. Seminars in cell & developmental biology. 2011;22(4):359-65. doi: 10.1016/j.semcdb.2011.02.016. PubMed PMID: 21333748; PMCID: 3184770.

10. Pascale A, Govoni S. The complex world of post-transcriptional mechanisms: is their deregulation a common link for diseases? Focus on ELAV-like RNA-binding proteins. Cellular and molecular life sciences : CMLS. 2012;69(4):501-17. doi: 10.1007/s00018-011-0810-7. PubMed PMID: 21909784.

11. Burd CG, Dreyfuss G. Conserved structures and diversity of functions of RNA-binding proteins. science. 1994;265(5172):615-21.

12. Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. Nature reviews Molecular cell biology. 2007;8(6):479-90. doi: 10.1038/nrm2178. PubMed PMID: 17473849.

13. Benitex Y, Baranger AM. Recognition of essential purines by the U1A protein. BMC biochemistry. 2007;8:22. doi: 10.1186/1471-2091-8-22. PubMed PMID: 17980039; PMCID: 2203988.

14. Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. The FEBS journal. 2005;272(9):2118-31. doi: 10.1111/j.1742-4658.2005.04653.x. PubMed PMID: 15853797.

15. Nolan SJS, J. C.; Tuite, J. B.; Cecere, K. L.; Baranger, A. M. Recognition of an essential adenine at a protein-RNA interface: comparison of the contribution of hydrogen bonds and a stacking interaction. J Am Chem Soc. 1999(121):2.

16. 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(23):5269-75. PubMed PMID: 12466552; PMCID: 137951.

17. Daubner GM, Cléry A, Allain FH. RRM–RNA recognition: NMR or crystallography… and new findings. Current opinion in structural biology. 2013;23(1):100-8.

18. Kudinov AE, Karanicolas J, Golemis EA, Boumber Y. Musashi RNA-binding proteins as cancer drivers and novel therapeutic targets. Clinical Cancer Research. 2017;23(9):2143-53.

19. Fox RG, Park FD, Koechlein CS, Kritzik M, Reya T. Musashi Signaling in Stem Cells and Cancer. Annual review of cell and developmental biology. 2015;31:249-67. doi: 10.1146/annurev-cellbio-100814-125446. PubMed PMID: 26566113.

20. Gunter KM, McLaughlin EA. Translational control in germ cell development: A role for the RNA-binding proteins Musashi‐1 and Musashi‐2. IUBMB life. 2011;63(9):678-85.

21. Kharas MG, Lengner CJ. Stem cells, cancer, and MUSASHI in blood and guts. Trends in cancer. 2017;3(5):347-56.

22. Sutherland JM, McLaughlin EA, Hime GR, Siddall NA. The Musashi family of RNA binding proteins: master regulators of multiple stem cell populations. Transcriptional and Translational Regulation of Stem Cells: Springer; 2013. p. 233-45.

23. Sengupta U, Montalbano M, McAllen S, Minuesa G, Kharas M, Kayed R. Formation of Toxic Oligomeric Assemblies of RNA-binding Protein: Musashi in Alzheimer's disease. Acta neuropathologica communications. 2018;6(1):113.

24. Fan LF, Dong WG, Jiang CQ, Xia D, Liao F, Yu QF. Expression of putative stem cell genes Musashi-1 and beta1-integrin in human colorectal adenomas and adenocarcinomas. International journal of colorectal disease. 2010;25(1):17-23. doi: 10.1007/s00384-009-0791-2. PubMed PMID: 19714342.

25. Ito T, Kwon HY, Zimdahl B, Congdon KL, Blum J, Lento WE, Zhao C, Lagoo A, Gerrard G, Foroni L, Goldman J, Goh H, Kim SH, Kim DW, Chuah C, Oehler VG, Radich JP, Jordan CT, Reya T. Regulation of myeloid leukaemia by the cell-fate determinant Musashi. Nature. 2010;466(7307):765-8. doi: 10.1038/nature09171. PubMed PMID: 20639863; PMCID: 2918284.

26. Kharas MG, Lengner CJ, Al-Shahrour F, Bullinger L, Ball B, Zaidi S, Morgan K, Tam W, Paktinat M, Okabe R, Gozo M, Einhorn W, Lane SW, Scholl C, Frohling S, Fleming M, Ebert BL, Gilliland DG, Jaenisch R, Daley GQ. Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. Nature medicine. 2010;16(8):903-8. doi: 10.1038/nm.2187. PubMed PMID: 20616797; PMCID: 3090658.

27. 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 reports. 2015;13(11):2440-55. doi: 10.1016/j.celrep.2015.11.022. PubMed PMID: 26673327.

28. Ma YH, Mentlein R, Knerlich F, Kruse ML, Mehdorn HM, Held-Feindt J. Expression of stem cell markers in human astrocytomas of different WHO grades. Journal of neuro-oncology. 2008;86(1):31-45. doi: 10.1007/s11060-007-9439-7. PubMed PMID: 17611714.

29. Seigel GM, Hackam AS, Ganguly A, Mandell LM, Gonzalez-Fernandez F. Human embryonic and neuronal stem cell markers in retinoblastoma. Molecular vision. 2007;13:823-32. PubMed PMID: 17615543; PMCID: 2768758.

30. Toda M, Iizuka Y, Yu W, Imai T, Ikeda E, Yoshida K, Kawase T, Kawakami Y, Okano H, Uyemura K. Expression of the neural RNA-binding protein Musashi1 in human gliomas. Glia. 2001;34(1):1-7. PubMed PMID: 11284014.

31. Todaro M, Francipane MG, Medema JP, Stassi G. Colon cancer stem cells: promise of targeted therapy. Gastroenterology. 2010;138(6):2151-62. doi: 10.1053/j.gastro.2009.12.063. PubMed PMID: 20420952.

32. Wang XY, Penalva LO, Yuan H, Linnoila RI, Lu J, Okano H, Glazer RI. Musashi1 regulates breast tumor cell proliferation and is a prognostic indicator of poor survival. Molecular cancer. 2010;9:221. doi: 10.1186/1476-4598-9-221. PubMed PMID: 20727204; PMCID: 2939568.

33. Ye F, Zhou C, Cheng Q, Shen J, Chen H. Stem-cell-abundant proteins Nanog, Nucleostemin and Musashi1 are highly expressed in malignant cervical epithelial cells. BMC cancer. 2008;8:108. doi: 10.1186/1471-2407-8-108. PubMed PMID: 18419830; PMCID: 2387168.

34. Yokota N, Mainprize TG, Taylor MD, Kohata T, Loreto M, Ueda S, Dura W, Grajkowska W, Kuo JS, Rutka JT. Identification of differentially expressed and developmentally regulated genes in

medulloblastoma using suppression subtraction hybridization. Oncogene. 2004;23(19):3444-53. doi: 10.1038/sj.onc.1207475. PubMed PMID: 15064731.

35. Clingman CC, Deveau LM, Hay SA, Genga RM, Shandilya SM, Massi F, Ryder SP. Allosteric inhibition of a stem cell RNA-binding protein by an intermediary metabolite. Elife. 2014;3:e02848.

36. 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. Molecular oncology. 2015;9(7):1406-20. doi: 10.1016/j.molonc.2015.03.014. PubMed PMID: 25933687; PMCID: 4523432.

37. 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. Combinatorial chemistry & high throughput screening. 2014;17(7):596-609.

38. Dinares GM, Albanese SK, Chow A, Schurer A, Park SM, Rotsides CZ, Taggart J, Rizzi A, Naden LN, Chou T. Small-Molecule Targeting of Musashi RNA-Binding Activity in Acute Myeloid Leukemia. Am Soc Hematology; 2018.

39. Christ F, Voet A, Marchand A, Nicolet S, Desimmie BA, Marchand D, Bardiot D, Van der Veken NJ, Van Remoortel B, Strelkov SV, De Maeyer M, Chaltin P, Debyser Z. Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. Nature chemical biology. 2010;6(6):442-8. doi: 10.1038/nchembio.370. PubMed PMID: 20473303.

40. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science. 1995;267(5196):383-6. PubMed PMID: 7529940.

41. Koes DR, Camacho CJ. Small-molecule inhibitor starting points learned from protein-protein interaction inhibitor structure. Bioinformatics. 2012;28(6):784-91. doi: 10.1093/bioinformatics/btr717. PubMed PMID: 22210869; PMCID: 3307105.

42. Liu S, Wu S, Jiang S. HIV entry inhibitors targeting gp41: from polypeptides to small-molecule compounds. Current pharmaceutical design. 2007;13(2):143-62. PubMed PMID: 17269924.

43. Moreira IS, Fernandes PA, Ramos MJ. Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins. 2007;68(4):803-12. doi: 10.1002/prot.21396. PubMed PMID: 17546660.

44. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(31):11287-92. doi: 10.1073/pnas.0401942101. PubMed PMID: 15269345; PMCID: 509196.

45. Thanos CD, DeLano WL, Wells JA. Hot-spot mimicry of a cytokine receptor by a small molecule. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(42):15422-7. doi: 10.1073/pnas.0607058103. PubMed PMID: 17032757; PMCID: 1592646.

46. Kranz JK, Schalk-Hihi C. Protein thermal shifts to identify low molecular weight fragments. Methods in enzymology: Elsevier; 2011. p. 277-98.

47. Scott DE, Spry C, Abell C. Differential scanning fluorimetry as part of a biophysical screening cascade. Fragment-based drug discovery lessons and outlook, Erlanson, DA and Jahnke W, Eds. 2016.

48. Ehrhardt MK, Warring SL, Gerth ML. Screening Chemoreceptor–Ligand Interactions by High-Throughput Thermal-Shift Assays. Bacterial Chemosensing: Springer; 2018. p. 281-90.

49. Carver TE, Bordeau B, Cummings MD, Petrella EC, Pucci MJ, Zawadzke LE, Dougherty BA, Tredup JA, Bryson JW, Yanchunas J. Decrypting the biochemical function of an essential gene from Streptococcus pneumoniae using ThermoFluor® technology. Journal of Biological Chemistry. 2005;280(12):11704-12.

50. Mahendrarajah K, Dalby PA, Wilkinson B, Jackson SE, Main ER. A high-throughput fluorescence chemical denaturation assay as a general screen for protein-ligand binding. Anal Biochem. 2011;411(1):155-7. doi: 10.1016/j.ab.2010.12.001. PubMed PMID: 21138727.

51. Niedziela-Majka A, Kan E, Weissburg P, Mehra U, Sellers S, Sakowicz R. High-throughput screening of formulations to optimize the thermal stability of a therapeutic monoclonal antibody. J Biomol Screen. 2015;20(4):552-9. doi: 10.1177/1087057114557781. PubMed PMID: 25385011.

52. Pantoliano MW, Petrella EC, Kwasnoski JD, Lobanov VS, Myslik J, Graf E, Carver T, Asel E, Springer BA, Lane P. High-density miniaturized thermal shift assays as a general strategy for drug discovery. Journal of biomolecular screening. 2001;6(6):429-40.

53. Seabrook SA, Newman J. High-throughput thermal scanning for protein stability: making a good technique more robust. ACS combinatorial science. 2013;15(8):387-92.

54. Matulis D, Kranz JK, Salemme FR, Todd MJ. Thermodynamic Stability of Carbonic Anhydrase: Measurements of Binding Affinity and Stoichiometry Using ThermoFluor. Biochemistry. 2005;44(13):9.

55. Zhang R, Monsma F. Fluorescence-based thermal shift assays. Current opinion in drug discovery & development. 2010;13(4):389-402.

56. Zubrienė A, Matulienė J, Baranauskienė L, Jachno J, Torresan J, Michailovienė V, Cimmperman P, Matulis D. Measurement of nanomolar dissociation constants by titration calorimetry and thermal shift assay–radicicol binding to Hsp90 and ethoxzolamide binding to CAII. International journal of molecular sciences. 2009;10(6):2662-80.

57. Ericsson UB, Hallberg BM, Detitta GT, Dekker N, Nordlund P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. Anal Biochem. 2006;357(2):289-98. Epub 2006/09/12. doi: 10.1016/j.ab.2006.07.027. PubMed PMID: 16962548.

58. Niesen FH, Berglund H, Vedadi M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat Protoc. 2007;2(9):2212-21. doi: 10.1038/nprot.2007.321. PubMed PMID: 17853878.

59. Schulz MN, Landström J, Hubbard RE. MTSA—A Matlab program to fit thermal shift data. Analytical biochemistry. 2013;433(1):43-7.

60. Sorrell FJ, Greenwood GK, Birchall K, Chen B. Development of a differential scanning fluorimetry based high throughput screening assay for the discovery of affinity binders against an anthrax protein. J Pharm Biomed Anal. 2010;52(5):802-8. Epub 2010/04/09. PubMed PMID: 20376913.

61. Scheepstra M, Hekking KF, van Hijfte L, Folmer RH. Bivalent ligands for protein degradation in drug discovery. Computational and structural biotechnology journal. 2019.

62. Gu S, Cui D, Chen X, Xiong X, Zhao Y. PROTACs: an emerging targeting technique for protein degradation in drug discovery. BioEssays. 2018;40(4):1700247.

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

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

65. Burslem GM, Smith BE, Lai AC, Jaime-Figueroa S, McQuaid DC, Bondeson DP, Toure M, Dong H, Qian Y, Wang J. The advantages of targeted protein degradation over inhibition: an RTK case study. Cell chemical biology. 2018;25(1):67-77. e3.

66. Chu T-T, Gao N, Li Q-Q, Chen P-G, Yang X-F, Chen Y-X, Zhao Y-F, Li Y-M. Specific knockdown of endogenous tau protein by peptide-directed ubiquitin-proteasome degradation. Cell chemical biology. 2016;23(4):453-61.

67. Lu M, Liu T, Jiao Q, Ji J, Tao M, Liu Y, You Q, Jiang Z. Discovery of a Keap1-dependent peptide PROTAC to knockdown Tau by ubiquitination-proteasome degradation pathway. European journal of medicinal chemistry. 2018;146:251-9.

68. Crew AP, Raina K, Dong H, Qian Y, Wang J, Vigil D, Serebrenik YV, Hamman BD, Morgan A, Ferraro C. Identification and characterization of Von Hippel-Lindau-recruiting proteolysis targeting chimeras (PROTACs) of TANK-binding kinase 1. Journal of medicinal chemistry. 2017;61(2):583-98.

69. Paiva S-L, Crews CM. Targeted protein degradation: elements of PROTAC design. Current opinion in chemical biology. 2019;50:111-9.

70. Bondeson DP, Smith BE, Burslem GM, Buhimschi AD, Hines J, Jaime-Figueroa S, Wang J, Hamman BD, Ishchenko A, Crews CM. Lessons in PROTAC design from selective degradation with a promiscuous warhead. Cell chemical biology. 2018;25(1):78-87. e5.

71. Brand M, Jiang B, Bauer S, Donovan KA, Liang Y, Wang ES, Nowak RP, Yuan JC, Zhang T, Kwiatkowski N. Homolog-Selective Degradation as a Strategy to Probe the Function of CDK6 in AML. Cell chemical biology. 2019;26(2):300-6. e9.

72. Buhimschi AD, Armstrong HA, Toure M, Jaime-Figueroa S, Chen TL, Lehman AM, Woyach JA, Johnson AJ, Byrd JC, Crews CM. Targeting the C481S ibrutinib-resistance mutation in Bruton's tyrosine kinase using PROTAC-mediated degradation. Biochemistry. 2018;57(26):3564-75.

73. Burslem GM, Song J, Chen X, Hines J, Crews CM. Enhancing antiproliferative activity and selectivity of a FLT-3 inhibitor by proteolysis targeting chimera conversion. Journal of the American Chemical Society. 2018;140(48):16428-32.

74. Gadd MS, Testa A, Lucas X, Chan K-H, Chen W, Lamont DJ, Zengerle M, Ciulli A. Structural basis of PROTAC cooperative recognition for selective protein degradation. Nature chemical biology. 2017;13(5):514.

75. Huang H-T, Dobrovolsky D, Paulk J, Yang G, Weisberg EL, Doctor ZM, Buckley DL, Cho J-H, Ko E, Jang J. A chemoproteomic approach to query the degradable kinome using a multi-kinase degrader. Cell chemical biology. 2018;25(1):88-99. e6.

76. Nowak RP, DeAngelo SL, Buckley D, He Z, Donovan KA, An J, Safaee N, Jedrychowski MP, Ponthier CM, Ishoey M. Plasticity in binding confers selectivity in ligand-induced protein degradation. Nature chemical biology. 2018;14(7):706.

77. Olson CM, Jiang B, Erb MA, Liang Y, Doctor ZM, Zhang Z, Zhang T, Kwiatkowski N, Boukhali M, Green JL. Pharmacological perturbation of CDK9 using selective CDK9 inhibition or degradation. Nature chemical biology. 2018;14(2):163.

78. Qin C, Hu Y, Zhou B, Fernandez-Salas E, Yang C-Y, Liu L, McEachern D, Przybranowski S, Wang M, Stuckey J. Discovery of QCA570 as an exceptionally potent and efficacious proteolysis targeting chimera (PROTAC) degrader of the bromodomain and extra-terminal (BET) proteins capable of inducing complete and durable tumor regression. Journal of medicinal chemistry. 2018;61(15):6685-704.

79. Rana S, Bendjennat M, Kour S, King HM, Kizhake S, Zahid M, Natarajan A. Selective degradation of CDK6 by a palbociclib based PROTAC. Bioorganic & medicinal chemistry letters. 2019;29(11):1375-9.

80. Sun Y, Zhao X, Ding N, Gao H, Wu Y, Yang Y, Zhao M, Hwang J, Song Y, Liu W. PROTAC-induced BTK degradation as a novel therapy for mutated BTK C481S induced ibrutinib-resistant B-cell malignancies. Cell research. 2018;28(7):779.

81. Tinworth CP, Lithgow H, Dittus L, Bassi ZI, Hughes SE, Muelbaier M, Dai H, Smith IE, Kerr WJ, Burley GA. PROTAC-Mediated Degradation of Bruton's Tyrosine Kinase Is Inhibited by Covalent Binding. ACS chemical biology. 2019;14(3):342-7.

82. Zengerle M, Chan K-H, Ciulli A. Selective small molecule induced degradation of the BET bromodomain protein BRD4. ACS chemical biology. 2015;10(8):1770-7.

83. Zhao Q, Lan T, Su S, Rao Y. Induction of apoptosis in MDA-MB-231 breast cancer cells by a PARP1-targeting PROTAC small molecule. Chemical communications. 2019;55(3):369-72.

84. Zoppi V, Hughes SJ, Maniaci C, Testa A, Gmaschitz T, Wieshofer C, Koegl M, Riching KM, Daniels DL, Spallarossa A. Iterative design and optimization of initially inactive proteolysis targeting chimeras (PROTACs) identify VZ185 as a potent, fast, and selective von Hippel–Lindau (VHL) based dual degrader probe of BRD9 and BRD7. Journal of medicinal chemistry. 2018;62(2):699-726.

85. Zorba A, Nguyen C, Xu Y, Starr J, Borzilleri K, Smith J, Zhu H, Farley KA, Ding W, Schiemer J. Delineating the role of cooperativity in the design of potent PROTACs for BTK. Proceedings of the National Academy of Sciences. 2018;115(31):E7285-E92.

86. 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(8):1062-70. doi: 10.1177/1087057112448100. PubMed PMID: 22644268; PMCID: 3520603.

87. King DT, Barnes M, Thomsen D, Lee CH. Assessing specific oligonucleotides and small molecule antibiotics for the ability to inhibit the CRD-BP-CD44 RNA interaction. PLoS One. 2014;9(3):e91585. doi: 10.1371/journal.pone.0091585. PubMed PMID: 24622399; PMCID: 3951440.

88. Cheng K, Wang X, Yin H. Small-molecule inhibitors of the TLR3/dsRNA complex. J Am Chem Soc. 2011;133(11):3764-7. doi: 10.1021/ja111312h. PubMed PMID: 21355588; PMCID: 3068529.

89. Gallego J, Varani G. Targeting RNA with small-molecule drugs: therapeutic promise and chemical challenges. Accounts of chemical research. 2001;34(10):836-43. PubMed PMID: 11601968.

90. Stelzer AC, Frank AT, Kratz JD, Swanson MD, Gonzalez-Hernandez MJ, Lee J, Andricioaei I, Markovitz DM, Al-Hashimi HM. Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. Nature chemical biology. 2011;7(8):553-9. doi: 10.1038/nchembio.596. PubMed PMID: 21706033; PMCID: 3319144.

91. Squires KE. An introduction to nucleoside and nucleotide analogues. Antiviral therapy. 2001;6 Suppl 3:1-14. PubMed PMID: 11678469.

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

93. Bitterman PB, Polunovsky VA. Attacking a nexus of the oncogenic circuitry by reversing aberrant eIF4F-mediated translation. Molecular cancer therapeutics. 2012;11(5):1051-61. doi: 10.1158/1535-7163.MCT-11-0530. PubMed PMID: 22572598; PMCID: 3349966.

94. Menendez-Arias L, Alvarez M, Pacheco B. Nucleoside/nucleotide analog inhibitors of hepatitis B virus polymerase: mechanism of action and resistance. Current opinion in virology. 2014;8C:1-9. doi: 10.1016/j.coviro.2014.04.005. PubMed PMID: 24814823.

95. James SH, Prichard MN. Current and future therapies for herpes simplex virus infections: mechanism of action and drug resistance. Current opinion in virology. 2014;8C:54-61. doi: 10.1016/j.coviro.2014.06.003. PubMed PMID: 25036916.

96. Das K, Arnold E. HIV-1 reverse transcriptase and antiviral drug resistance. Part 1. Current opinion in virology. 2013;3(2):111-8. doi: 10.1016/j.coviro.2013.03.012. PubMed PMID: 23602471; PMCID: 4097814.

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

98. Biswas S, Sukla S, Field HJ. Helicase-primase inhibitors for herpes simplex virus: looking to the future of non-nucleoside inhibitors for treating herpes virus infections. Future medicinal chemistry. 2014;6(1):45-55. doi: 10.4155/fmc.13.192. PubMed PMID: 24358947.

99. Das K, Arnold E. HIV-1 reverse transcriptase and antiviral drug resistance. Part 2. Current opinion in virology. 2013;3(2):119-28. doi: 10.1016/j.coviro.2013.03.014. PubMed PMID: 23602470; PMCID: 4097817.

100. Gowthaman R, Deeds EJ, Karanicolas J. Structural properties of non-traditional drug targets present new challenges for virtual screening. J Chem Inf Model. 2013;53(8):2073-81. doi: 10.1021/ci4002316. PubMed PMID: 23879197; PMCID: 3819422.

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

102. 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. Epub 2010/12/29. doi: 10.1016/B978-0-12-381270-4.00019-6. PubMed PMID: 21187238.

103. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. J Chem Inf Model. 2012. Epub 2012/05/17. doi: 10.1021/ci3001277. PubMed PMID: 22587354; PMCID: 3402020.

104. 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(4):572-84. Epub 2010/03/20. doi: 10.1021/ci100031x. PubMed PMID: 20235588; PMCID: 2859685.

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

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

107. Rush TS, 3rd, Grant JA, Mosyak L, Nicholls A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J Med Chem. 2005;48(5):1489-95. Epub 2005/03/04. doi: 10.1021/jm040163o. PubMed PMID: 15743191.

108. Daubner GM, Clery A, Allain FH. RRM-RNA recognition: NMR or crystallography...and new findings. Curr Opin Struct Biol. 2013;23(1):100-8. doi: 10.1016/j.sbi.2012.11.006. PubMed PMID: 23253355.

109. Auweter SD, Oberstrass FC, Allain FH. Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Res. 2006;34(17):4943-59. doi: 10.1093/nar/gkl620. PubMed PMID: 16982642; PMCID: 1635273.

110. 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(7):3218-31. doi: 10.1093/nar/gkr1139. PubMed PMID: 22140116; PMCID: 3326303.

111. Okano H, Kawahara H, Toriya M, Nakao K, Shibata S, Imai T. Function of RNA-binding protein Musashi-1 in stem cells. Experimental cell research. 2005;306(2):349-56. doi: 10.1016/j.yexcr.2005.02.021. PubMed PMID: 15925591.

112. Spears E, Neufeld KL. Novel double-negative feedback loop between adenomatous polyposis coli and Musashi1 in colon epithelia. J Biol Chem. 2011;286(7):4946-50. doi: 10.1074/jbc.C110.205922. PubMed PMID: 21199875; PMCID: 3037606.

113. Muto J, Imai T, Ogawa D, Nishimoto Y, Okada Y, Mabuchi Y, Kawase T, Iwanami A, Mischel PS, Saya H, Yoshida K, Matsuzaki Y, Okano H. RNA-binding protein Musashi1 modulates glioma cell growth through the post-transcriptional regulation of Notch and PI3 kinase/Akt signaling pathways. PLoS One. 2012;7(3):e33431. doi: 10.1371/journal.pone.0033431. PubMed PMID: 22428049; PMCID: 3299785.

114. 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(7):3222-31.

115. 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(51):10831-41.

116. 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(20):3241-57. doi: 10.1016/j.jmb.2015.05.009. PubMed PMID: 26003924; PMCID: 4586317.

117. 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. Nature structural & molecular biology. 2014;21(6):544-51. doi: 10.1038/nsmb.2832. PubMed PMID: 24837192; PMCID: 4141773.

118. Nissen P, Kjeldgaard M, Nyborg J. Macromolecular mimicry. The EMBO journal. 2000;19(4):489-95. doi: 10.1093/emboj/19.4.489. PubMed PMID: 10675317; PMCID: 305586.

119. Tsonis PA, Dwivedi B. Molecular mimicry: structural camouflage of proteins and nucleic acids. Biochimica et biophysica acta. 2008;1783(2):177-87. doi: 10.1016/j.bbamcr.2007.11.001. PubMed PMID: 18068679.

120. 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(4):615-20.

121. Sakamoto KM. Protacs for treatment of cancer. Pediatric research. 2010;67(5):505.

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

123. Pantoliano MW, Petrella EC, Kwasnoski JD, Lobanov VS, Myslik J, Graf E, Carver T, Asel E, Springer BA, Lane P, Salemme FR. High-density miniaturized thermal shift assays as a general strategy

for drug discovery. J Biomol Screen. 2001;6(6):429-40. Epub 2002/01/15. doi: 10.1177/108705710100600609. PubMed PMID: 11788061.

124. Carver TE, Bordeau B, Cummings MD, Petrella EC, Pucci MJ, Zawadzke LE, Dougherty BA, Tredup JA, Bryson JW, Yanchunas J, Jr., Doyle ML, Witmer MR, Nelen MI, DesJarlais RL, Jaeger EP, Devine H, Asel ED, Springer BA, Bone R, Salemme FR, Todd MJ. Decrypting the biochemical function of an essential gene from Streptococcus pneumoniae using ThermoFluor technology. J Biol Chem. 2005;280(12):11704-12. Epub 2005/01/07. doi: 10.1074/jbc.M413278200. PubMed PMID: 15634672.

125. Zhang R, Monsma F. Fluorescence-based thermal shift assays. Curr Opin Drug Discov Devel. 2010;13(4):389-402. Epub 2010/07/03. PubMed PMID: 20597025.

126. Holdgate GA, Anderson M, Edfeldt F, Geschwindner S. Affinity-based, biophysical methods to detect and analyze ligand binding to recombinant proteins: matching high information content with high throughput. J Struct Biol. 2010;172(1):142-57. Epub 2010/07/09. doi: 10.1016/j.jsb.2010.06.024. PubMed PMID: 20609391.

127. Brandts JF, Lin LN. Study of strong to ultratight protein interactions using differential scanning calorimetry. Biochemistry. 1990;29(29):6927-40.

128. Kranz JK, Schalk-Hihi C. Protein thermal shifts to identify low molecular weight fragments. Methods Enzymol. 2011;493:277-98. Epub 2011/03/05. doi: 10.1016/B978-0-12-381274-2.00011-X. PubMed PMID: 21371595.

129. Scott DE, Spry C, Abell C. Differential Scanning Fluorimetry as Part of a Biophysical Screening Cascade. Fragment-based Drug Discovery Lessons and Outlook2016. p. 139-72.

130. Seabrook SA, Newman J. High-throughput thermal scanning for protein stability: making a good technique more robust. ACS Comb Sci. 2013;15(8):387-92.

131. Göth M, Badock V, Weiske J, Pagel K, Kuropka B. Critical Evaluation of Native Electrospray Ionization Mass Spectrometry for Fragment-Based Screening. ChemMedChem. 2017;12(15):1201-11.

132. Rombouts FJ, Alexander R, Cleiren E, De Groot A, Carpentier M, Dijkmans J, Fierens K, Masure S, Moechars D, Palomino-Schätzlein M. Fragment Binding to β-Secretase 1 without Catalytic Aspartate Interactions Identified via Orthogonal Screening Approaches. ACS Omega. 2017;2(2):685-97.

133. Schiebel J, Radeva N, Köster H, Metz A, Krotzky T, Kuhnert M, Diederich WE, Heine A, Neumann L, Atmanene C. One question, multiple answers: biochemical and biophysical screening methods retrieve deviating fragment hit lists. ChemMedChem. 2015;10(9):1511-21.

134. Dai R, Geders TW, Liu F, Park SW, Schnappinger D, Aldrich CC, Finzel BC. Fragment-based exploration of binding site flexibility in Mycobacterium tuberculosis BioA. J Med Chem. 2015;58(13):5208-17. Epub 2015/06/13. doi: 10.1021/acs.jmedchem.5b00092. PubMed PMID: 26068403; PMCID: PMC4687966.

135. Chilton M, Clennell B, Edfeldt F, Geschwindner S. Hot-Spotting with Thermal Scanning: A Ligand- and Structure-Independent Assessment of Target Ligandability. J Med Chem. 2017;60(12):4923-31.

136. Hopkins AL, Keseru GM, Leeson PD, Rees DC, Reynolds CH. The role of ligand efficiency metrics in drug discovery. Nat Rev Drug Discov. 2014;13(2):105-21. Epub 2014/02/01. doi: 10.1038/nrd4163. PubMed PMID: 24481311.

137. Rudolf AF, Skovgaard T, Knapp S, Jensen LJ, Berthelsen J. A comparison of protein kinases inhibitor screening methods using both enzymatic activity and binding affinity determination. PLoS One. 2014;9(6):e98800. Epub 2014/06/11. doi: 10.1371/journal.pone.0098800. PubMed PMID: 24915177; PMCID: PMC4051630.

138. Schulz MN, Landström J, Hubbard RE. MTSA—A Matlab program to fit thermal shift data. Anal Biochem. 2013;433(1):43-7.

139. Matulis D, Kranz JK, Salemme FR, Todd MJ. Thermodynamic Stability of Carbonic Anhydrase: Measurements of Binding Affinity and Stoichiometry Using ThermoFluor. Biochemistry. 2005;44(13):5258-66.

140. Zubrienė A, Matulienė J, Baranauskienė L, Jachno J, Torresan J, Michailovienė V, Cimmperman P, Matulis D. Measurement of nanomolar dissociation constants by titration calorimetry and thermal shift assay–radicicol binding to Hsp90 and ethoxzolamide binding to CAII. Int J Mol Sci. 2009;10(6):2662-80.

141. Yeh AP, McMillan A, Stowell MH. Rapid and simple protein-stability screens: application to membrane proteins. Acta Crystallogr D Biol Crystallogr. 2006;62(Pt 4):451-7. Epub 2006/03/23. doi: 10.1107/S0907444906005233. PubMed PMID: 16552147.

142. Crowther GJ, Napuli AJ, Thomas AP, Chung DJ, Kovzun KV, Leibly DJ, Castaneda LJ, Bhandari J, Damman CJ, Hui R, Hol WG, Buckner FS, Verlinde CL, Zhang Z, Fan E, van Voorhis WC. Buffer optimization of thermal melt assays of Plasmodium proteins for detection of small-molecule ligands. J Biomol Screen. 2009;14(6):700-7. Epub 2009/05/28. doi: 10.1177/1087057109335749. PubMed PMID: 19470714; PMCID: PMC2819745.

143. Wang CK, Weeratunga SK, Pacheco CM, Hofmann A. DMAN: a Java tool for analysis of multi-well differential scanning fluorimetry experiments. Bioinformatics. 2012;28(3):439-40. Epub 2011/12/03. doi: 10.1093/bioinformatics/btr664. PubMed PMID: 22135419.

144. Milton ME, Allen C, Feldmann EA, Bobay BG, Jung DK, Stephens MD, Melander RJ, Theisen KE, Zeng D, Thompson RJ. Structure of the Francisella response regulator QseB receiver domain, and characterization of QseB inhibition by antibiofilm 2-aminoimidazole-based compounds. Mol Microbiol. 2017;106(2):223-35.

145. Milton ME, Minrovic BM, Harris DL, Kang B, Jung D, Lewis CP, Thompson RJ, Melander RJ, Zeng D, Melander C. Re-sensitizing multidrug resistant bacteria to antibiotics by targeting bacterial response regulators: characterization and comparison of interactions between 2-aminoimidazoles and the response regulators BfmR from Acinetobacter baumannii and QseB from Francisella spp. Front Mol Biosci. 2018;5:15.

146. Molledo MM, Quistgaard EM, Flayhan A, Pieprzyk J, Löw C. Multispecific Substrate Recognition in a Proton-Dependent Oligopeptide Transporter. Structure. 2018;26(3):467-76. e4.

147. Larsson EA, Jansson A, Ng FM, Then SW, Panicker R, Liu B, Sangthongpitag K, Pendharkar V, Tai SJ, Hill J. Fragment-based ligand design of novel potent inhibitors of tankyrases. J Med Chem. 2013;56(11):4497-508.

148. Huynh K, Partch CL. Analysis of protein stability and ligand interactions by thermal shift assay. Curr Protoc Protein Sci. 2015;79:28.9.1-14.

149. Robertson AD, Murphy KP. Protein Structure and the Energetics of Protein Stability. Chem Rev. 1997;97(5):1251-68. Epub 1997/08/05. PubMed PMID: 11851450.

150. Becktel WJ, Schellman JA. Protein stability curves. Biopolymers. 1987;26(11):1859-77. Epub 1987/11/01. doi: 10.1002/bip.360261104. PubMed PMID: 3689874.

151. Novokhatny V, Ingham K. Thermodynamics of maltose binding protein unfolding. Protein Sci. 1997;6(1):141-6. Epub 1997/01/01. doi: 10.1002/pro.5560060116. PubMed PMID: 9007986; PMCID: PMC2143499.

152. Privalov PL, Makhatadze GI. Heat capacity of proteins. II. Partial molar heat capacity of the unfolded polypeptide chain of proteins: protein unfolding effects. J Mol Biol. 1990;213(2):385-91. Epub 1990/05/20. doi: 10.1016/S0022-2836(05)80198-6. PubMed PMID: 2160545.

153. Vivoli M, Novak HR, Littlechild JA, Harmer NJ. Determination of protein-ligand interactions using differential scanning fluorimetry. J Vis Exp. 2014(91):51809. doi: 10.3791/51809. PubMed PMID: 25285605; PMCID: PMC4692391.

154. Cimmperman P, Baranauskiene L, Jachimoviciute S, Jachno J, Torresan J, Michailoviene V, Matuliene J, Sereikaite J, Bumelis V, Matulis D. A quantitative model of thermal stabilization and destabilization of proteins by ligands. Biophys J. 2008;95(7):3222-31. doi: 10.1529/biophysj.108.134973. PubMed PMID: 18599640; PMCID: PMC2547457.

155. 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(51):10831-41. doi: 10.1021/bi101414z. PubMed PMID: 21050007.

156. Lo M-C, Aulabaugh A, Jin G, Cowling R, Bard J, Malamas M, Ellestad G. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. Anal Biochem. 2004;332(1):153-9.

157. Abbott JA, Livingston NM, Egri SB, Guth E, Francklyn CS. Characterization of aminoacyl-tRNA synthetase stability and substrate interaction by differential scanning fluorimetry. Methods. 2017;113:64-71.

158. Booth WT, Schlachter CR, Pote S, Ussin N, Mank NJ, Klapper V, Offermann LR, Tang C, Hurlburt BK, Chruszcz M. Impact of an N-terminal Polyhistidine Tag on Protein Thermal Stability. ACS Omega. 2018;3(1):760-8.

159. Choudhary D, Kumar A, Magliery TJ, Sotomayor M. Using thermal scanning assays to test protein-protein interactions of inner-ear cadherins. PLoS One. 2017;12(12):e0189546.

160. Hallett ST, Pastok MW, Morgan RML, Wittner A, Blundell KL, Felletar I, Wedge SR, Prodromou C, Noble ME, Pearl LH. Differential regulation of G1 CDK complexes by the Hsp90-Cdc37 chaperone system. Cell reports. 2017;21(5):1386-98.

161. Schon A, Brown RK, Hutchins BM, Freire E. Ligand binding analysis and screening by chemical denaturation shift. Anal Biochem. 2013;443(1):52-7. doi: 10.1016/j.ab.2013.08.015. PubMed PMID: 23994566; PMCID: PMC3809086.

162. Molina DM, Jafari R, Ignatushchenko M, Seki T, Larsson EA, Dan C, Sreekumar L, Cao Y, Nordlund P. Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. Science. 2013;341(6141):84-7.

163. Sarver RW, Rogers JM, Epps DE. Determination of ligand-MurB interactions by isothermal denaturation: application as a secondary assay to complement high throughput screening. J Biomol Screen. 2002;7(1):21-8. Epub 2002/03/19. doi: 10.1177/108705710200700104. PubMed PMID: 11897052.

164. Senisterra GA, Soo Hong B, Park HW, Vedadi M. Application of high-throughput isothermal denaturation to assess protein stability and screen for ligands. J Biomol Screen. 2008;13(5):337-42. Epub 2008/05/02. doi: 10.1177/1087057108317825. PubMed PMID: 18448703.

165. Fedorov O, Marsden B, Pogacic V, Rellos P, Muller S, Bullock AN, Schwaller J, Sundstrom M, Knapp S. A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(51):20523-8. Epub 2007/12/14. doi: 10.1073/pnas.0708800104. PubMed PMID: 18077363; PMCID: PMC2154464.

166. Filippakopoulos P, Qi J, Picaud S, Shen Y, Smith WB, Fedorov O, Morse EM, Keates T, Hickman TT, Felletar I, Philpott M, Munro S, McKeown MR, Wang Y, Christie AL, West N, Cameron MJ, Schwartz B, Heightman TD, La Thangue N, French CA, Wiest O, Kung AL, Knapp S, Bradner JE. Selective inhibition of BET bromodomains. Nature. 2010;468(7327):1067-73. Epub 2010/09/28. doi: 10.1038/nature09504. PubMed PMID: 20871596; PMCID: PMC3010259.

167. Waldron TT, Murphy KP. Stabilization of proteins by ligand binding: application to drug screening and determination of unfolding energetics. Biochemistry. 2003;42(17):5058-64.

168. Dai R, Wilson DJ, Geders TW, Aldrich CC, Finzel BC. Inhibition of Mycobacterium tuberculosis transaminase BioA by aryl hydrazines and hydrazides. Chembiochem. 2014;15(4):575-86. Epub 2014/02/01. doi: 10.1002/cbic.201300748. PubMed PMID: 24482078; PMCID: PMC4020011.

169. Garnier C, Devred F, Byrne D, Puppo R, Roman AY, Malesinski S, Golovin AV, Lebrun R, Ninkina NN, Tsvetkov PO. Zinc binding to RNA recognition motif of TDP-43 induces the formation of amyloid-like aggregates. Sci Rep. 2017;7(1):6812.

170. 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(2):261-73. doi: 10.1016/j.ab.2004.05.055. PubMed PMID: 15325294.

171. Thomson J, Liu Y, Sturtevant JM, Quiocho FA. A thermodynamic study of the binding of linear and cyclic oligosaccharides to the maltodextrin-binding protein of Escherichia coli. Biophys Chem. 1998;70(2):101-8. Epub 1998/04/16. PubMed PMID: 9540203.

172. Telmer PG, Shilton BH. Insights into the conformational equilibria of maltose-binding protein by analysis of high affinity mutants. J Biol Chem. 2003;278(36):34555-67. doi: 10.1074/jbc.M301004200. PubMed PMID: 12794084.

173. Walker IH, Hsieh PC, Riggs PD. Mutations in maltose-binding protein that alter affinity and solubility properties. Appl Microbiol Biotechnol. 2010;88(1):187-97. Epub 2010/06/11. doi: 10.1007/s00253-010-2696-y. PubMed PMID: 20535468; PMCID: PMC2940430.

174. Greenfield NJ. Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism. Nat Protoc. 2006;1(6):2733-41. doi: 10.1038/nprot.2006.229. PubMed PMID: 17406529; PMCID: PMC2728349.

175. Monera OD, Kay CM, Hodges RS. Protein denaturation with guanidine hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interactions. Protein Sci. 1994;3(11):1984-91. Epub 1994/11/01. doi: 10.1002/pro.5560031110. PubMed PMID: 7703845; PMCID: PMC2142645.

176. Myers JK, Pace CN, Scholtz JM. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci. 1995;4(10):2138-48. Epub 1995/10/01. doi: 10.1002/pro.5560041020. PubMed PMID: 8535251; PMCID: PMC2142997.

177. Miller DM, 3rd, Olson JS, Pflugrath JW, Quiocho FA. Rates of ligand binding to periplasmic proteins involved in bacterial transport and chemotaxis. J Biol Chem. 1983;258(22):13665-72. Epub 1983/11/25. PubMed PMID: 6358208.

178. Smirnov A, Zubrienė A, Manakova E, Gražulis S, Matulis D. Crystal structure correlations with the intrinsic thermodynamics of human carbonic anhydrase inhibitor binding. PeerJ. 2018;6:e4412.

179. Baell JB, Nissink JWM. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017-Utility and Limitations. ACS Chem Biol. 2018;13(1):36-44. Epub 2017/12/05. doi: 10.1021/acschembio.7b00903. PubMed PMID: 29202222; PMCID: PMC5778390.

180. Malhotra S, Karanicolas J. When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode? J Med Chem. 2017;60(1):128-45. Epub 2016/12/17. doi: 10.1021/acs.jmedchem.6b00725. PubMed PMID: 27982595; PMCID: PMC5525026.

181. Holdgate GA, Ward WH. Measurements of binding thermodynamics in drug discovery. Drug Discov Today. 2005;10(22):1543-50. Epub 2005/11/01. doi: 10.1016/S1359-6446(05)03610-X. PubMed PMID: 16257377.

182. Wang W. Protein aggregation and its inhibition in biopharmaceutics. Int J Pharm. 2005;289(1-2):1-30. Epub 2005/01/18. doi: 10.1016/j.ijpharm.2004.11.014. PubMed PMID: 15652195.

183. Hoeser J, Gnandt E, Friedrich T. Low cost, microcontroller based heating device for multi-wavelength differential scanning fluorimetry. Sci Rep. 2018;8(1):1457.

184. Navarro S, Ventura S. Fluorescent dye ProteoStat to detect and discriminate intracellular amyloid-like aggregates in Escherichia coli. Biotechnol J. 2014;9(10):1259-66. Epub 2014/08/13. doi: 10.1002/biot.201400291. PubMed PMID: 25112199.

185. Seelig J, Schönfeld H-J. Thermal protein unfolding by differential scanning calorimetry and circular dichroism spectroscopy two-state model versus sequential unfolding. Q Rev Biophys. 2016;49:e9.

186. Jones CL, Fish F, Muccio DD. Determination of RNase A/2′-cytidine monophosphate binding affinity and enthalpy by a global fit of thermal unfolding curves. Anal Biochem. 2002;302(2):184-90.

187. Mayhood TW, Windsor WT. Ligand binding affinity determined by temperature-dependent circular dichroism: cyclin-dependent kinase 2 inhibitors. Anal Biochem. 2005;345(2):187-97.

188. Greenfield NJ. Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nat Protoc. 2006;1(6):2527-35. doi: 10.1038/nprot.2006.204. PubMed PMID: 17406506; PMCID: PMC2752288.

189. Verpoorte JA, Mehta S, Edsall JT. Esterase activities of human carbonic anhydrases B and C. J Biol Chem. 1967;242(18):4221-9. Epub 1967/09/25. PubMed PMID: 4964830.

190. Sakamoto KM, Kim KB, Kumagai A, Mercurio F, Crews CM, Deshaies RJ. Protacs: chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. Proceedings of the National Academy of Sciences. 2001;98(15):8554-9.

191. Gechijian LN, Buckley DL, Lawlor MA, Reyes JM, Paulk J, Ott CJ, Winter GE, Erb MA, Scott TG, Xu M. Functional TRIM24 degrader via conjugation of ineffectual bromodomain and VHL ligands. Nature chemical biology. 2018;14(4):405.

192. Farnaby W, Koegl M, Roy MJ, Whitworth C, Diers E, Trainor N, Zollman D, Steurer S, Karolyi-Oezguer J, Riedmueller C. BAF complex vulnerabilities in cancer demonstrated via structure-based PROTAC design. Nature chemical biology. 2019:1.

193. Owen D. Choosing your druggability battle. 2019.

194. Han X, Wang C, Qin C, Xiang W, Fernandez-Salas E, Yang C-Y, Wang M, Zhao L, Xu T, Chinnaswamy K. Discovery of ARD-69 as a highly potent proteolysis targeting chimera (PROTAC) degrader of androgen receptor (AR) for the treatment of prostate cancer. Journal of medicinal chemistry. 2019;62(2):941-64.

195. Buckley DL, Raina K, Darricarrere N, Hines J, Gustafson JL, Smith IE, Miah AH, Harling JD, Crews CM. HaloPROTACS: use of small molecule PROTACs to induce degradation of HaloTag fusion proteins. ACS chemical biology. 2015;10(8):1831-7.

196. Ward CC, Kleinman JI, Brittain SM, Lee PS, Chung CYS, Kim K, Petri Y, Thomas JR, Tallarico JA, McKenna JM. Covalent ligand screening uncovers a RNF4 E3 ligase recruiter for targeted protein degradation applications. ACS chemical biology. 2019.

197. Ottis P, Toure M, Cromm PM, Ko E, Gustafson JL, Crews CM. Assessing different E3 ligases for small molecule induced protein ubiquitination and degradation. ACS chemical biology. 2017;12(10):2570-8.

198. Zhang X, Crowley VM, Wucherpfennig TG, Dix MM, Cravatt BF. Electrophilic PROTACs that degrade nuclear proteins by engaging DCAF16. Nature chemical biology. 2019:1.

199. Riching KM, Mahan S, Corona CR, McDougall M, Vasta JD, Robers MB, Urh M, Daniels DL. Quantitative live-cell kinetic degradation and mechanistic profiling of PROTAC mode of action. ACS chemical biology. 2018;13(9):2758-70.

200. Cyrus K, Wehenkel M, Choi E-Y, Han H-J, Lee H, Swanson H, Kim K-B. Impact of linker length on the activity of PROTACs. Molecular BioSystems. 2011;7(2):359-64.

201. Drummond ML, Williams CI. In Silico Modeling of PROTAC-Mediated Ternary Complexes: Validation and Application. Journal of chemical information and modeling. 2019;59(4):1634-44.

202. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.

203. Leaver-Fay A, Tyka M, Lewis S, Lange O, Thompson J, Jacak R, Kaufman K, Renfrew P, Smith C, Sheffler W. ROSETTA3: This article is licensed under a Creative Commons Attribution 3.0 Unported Licence. an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545-74.

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

205. Hawkins PC, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. Journal of medicinal chemistry. 2007;50(1):74-82.

206. Omega V. 3.1.1.2, OpenEye Scientific Software. Inc: Santa Fe, NM. 2019.

207. Holdgate GA, Anderson M, Edfeldt F, Geschwindner S. Affinity-based, biophysical methods to detect and analyze ligand binding to recombinant proteins: matching high information content with high throughput. Journal of structural biology. 2010;172(1):142-57.

208. 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(Database issue):D277-82. doi: 10.1093/nar/gkq1108. PubMed PMID: 21071426; PMCID: 3013700.

209. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins. 1999;35(2):133-52. Epub 1999/05/01. doi: 10.1002/(SICI)1097-0134(19990501)35:2<133::AID-PROT1>3.0.CO;2-N [pii]. PubMed PMID: 10223287.

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

211. Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res. 2014;42(Web Server issue):W320-4. doi: 10.1093/nar/gku316. PubMed PMID: 24753421.

Appendix A: Supporting Information

A.1: Supporting Information for Chapter II

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

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) (110).

Building hotspot pharmacophores

Hotspot pharmacophores were built using a new dedicated protocol implemented in the Rosetta software suite (102), 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) (208) 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 (103). We generated up to 100 conformers for each molecule in the database using OMEGA (104-106). 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 (102). This energy function includes terms that capture packing, hydrogen bonding, implicit solvation (modeled via EEF1 (209)), 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 (210), 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(104-106). 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(73); 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 (170).

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 (58). 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 ^{o}C 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).

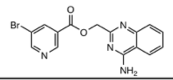
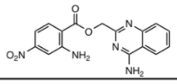
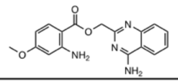
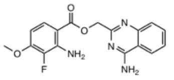
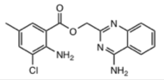
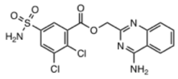
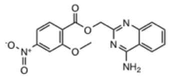
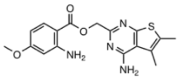
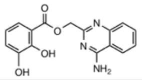
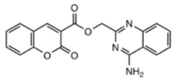
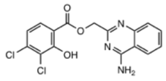
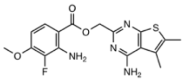
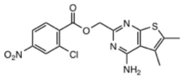
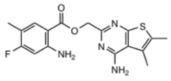
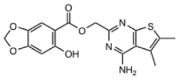
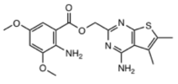
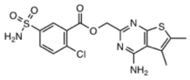
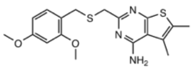
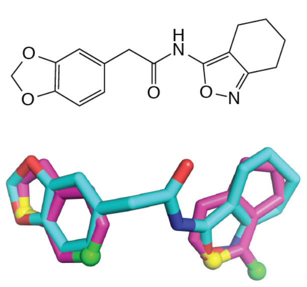
Compound	2D Structure	EC50 (μM) in Fluorescence Polarization Assay
R12		N.D.*
R12-7		49
R12-8		> 100
R12-8-19		> 100
R12-8-22		90
R12-8-29		100
R12-8-38		> 100
R12-8-44		22
R12-8-46		9
R12-8-47		> 100
R12-8-48		> 100
R12-8-44-1		16
R12-8-44-2		> 100
R12-8-44-3		9
R12-8-44-4		6
R12-8-44-6		> 100
R12-8-44-7b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

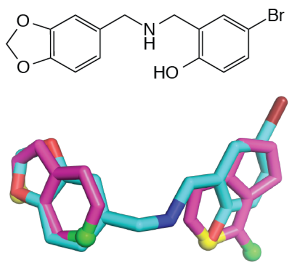
Table S2: Three rounds of compound optimization, starting from initial hit R12.

Supporting Figures

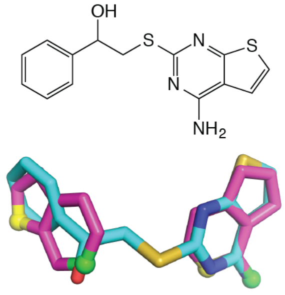
R1 (Chemotype I)



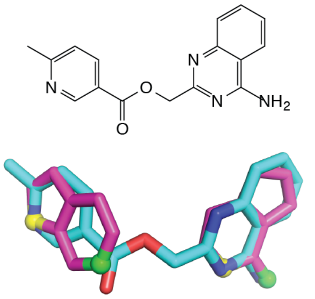
R2 (Chemotype I)



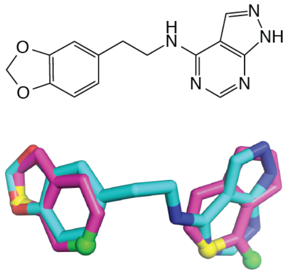
R3 (Chemotype II)



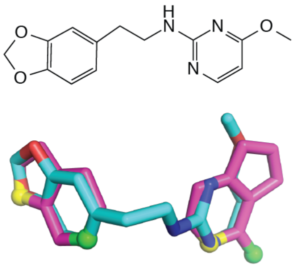
R4 (Chemotype III)



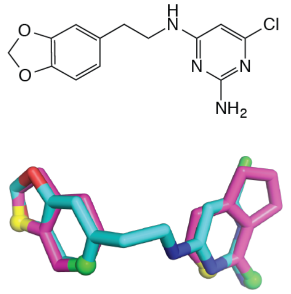
R5 (Chemotype I)



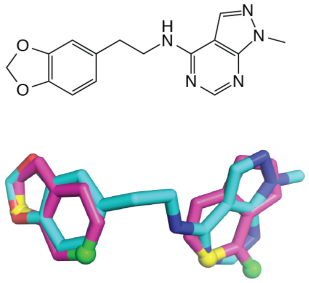
R6 (Chemotype I)



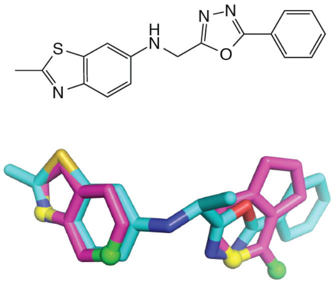
R7 (Chemotype I)



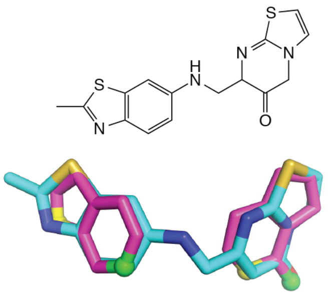
R8 (Chemotype I)



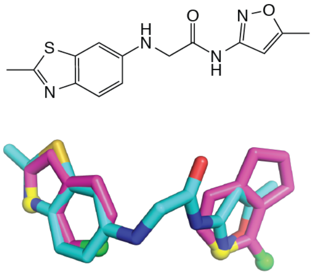
R9 (Chemotype II)



R10 (Chemotype II)



R11 (Chemotype II)



R12 (Chemotype III)

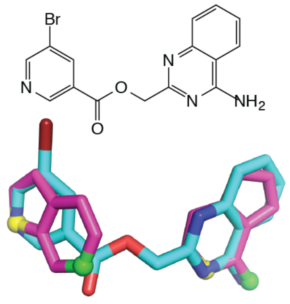


Figure S1: The 12 initial hit compounds. The chemical structure is shown for each compound, as well as a three-dimensional model of each compound (*cyan*) superposed with the Msi1 RBD1 hotspot pharmacophore (*magenta*).

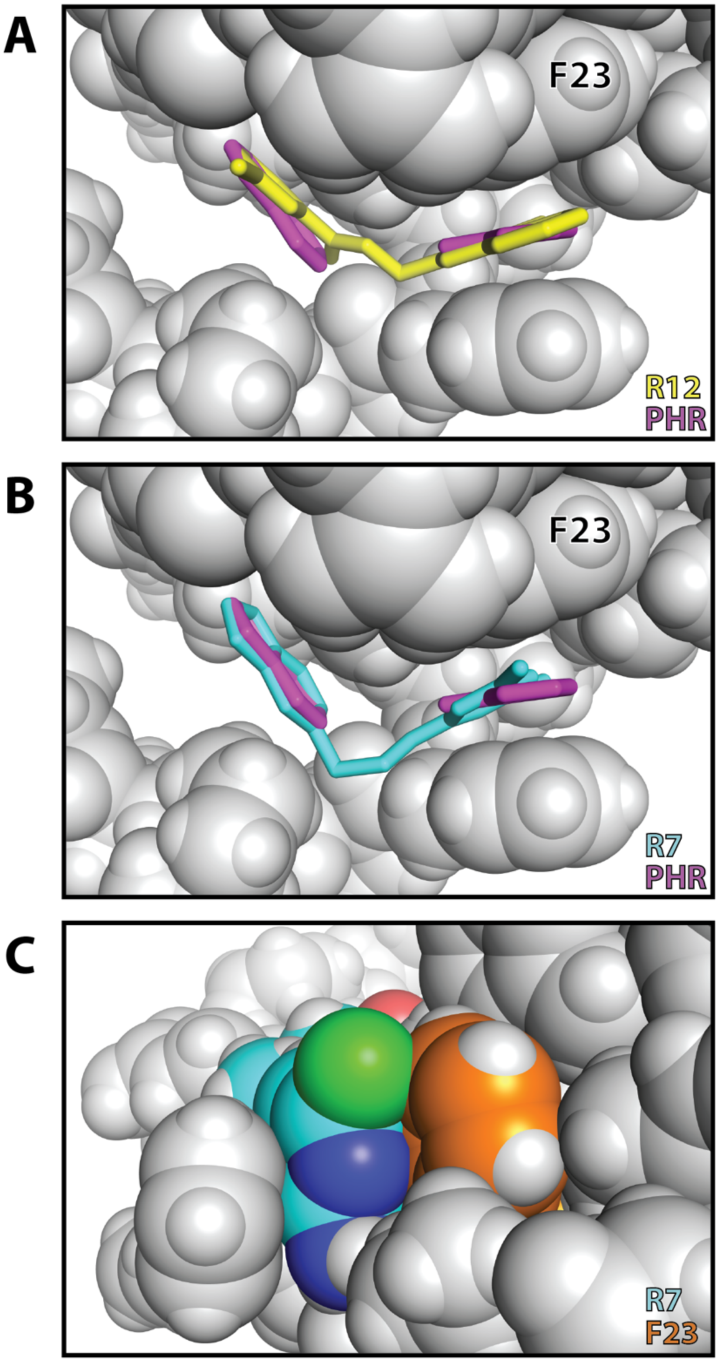


Figure S2: An inadvertent steric clash may explain the lack of binding by R7. (A) The rings in the model of R12 (*yellow*) are well-superposed with those of the hotspot pharmacophore (*magenta*), allowing for aromatic stacking with Msi1. **(B)** The relative positioning of the rings in the R7 (*cyan*)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

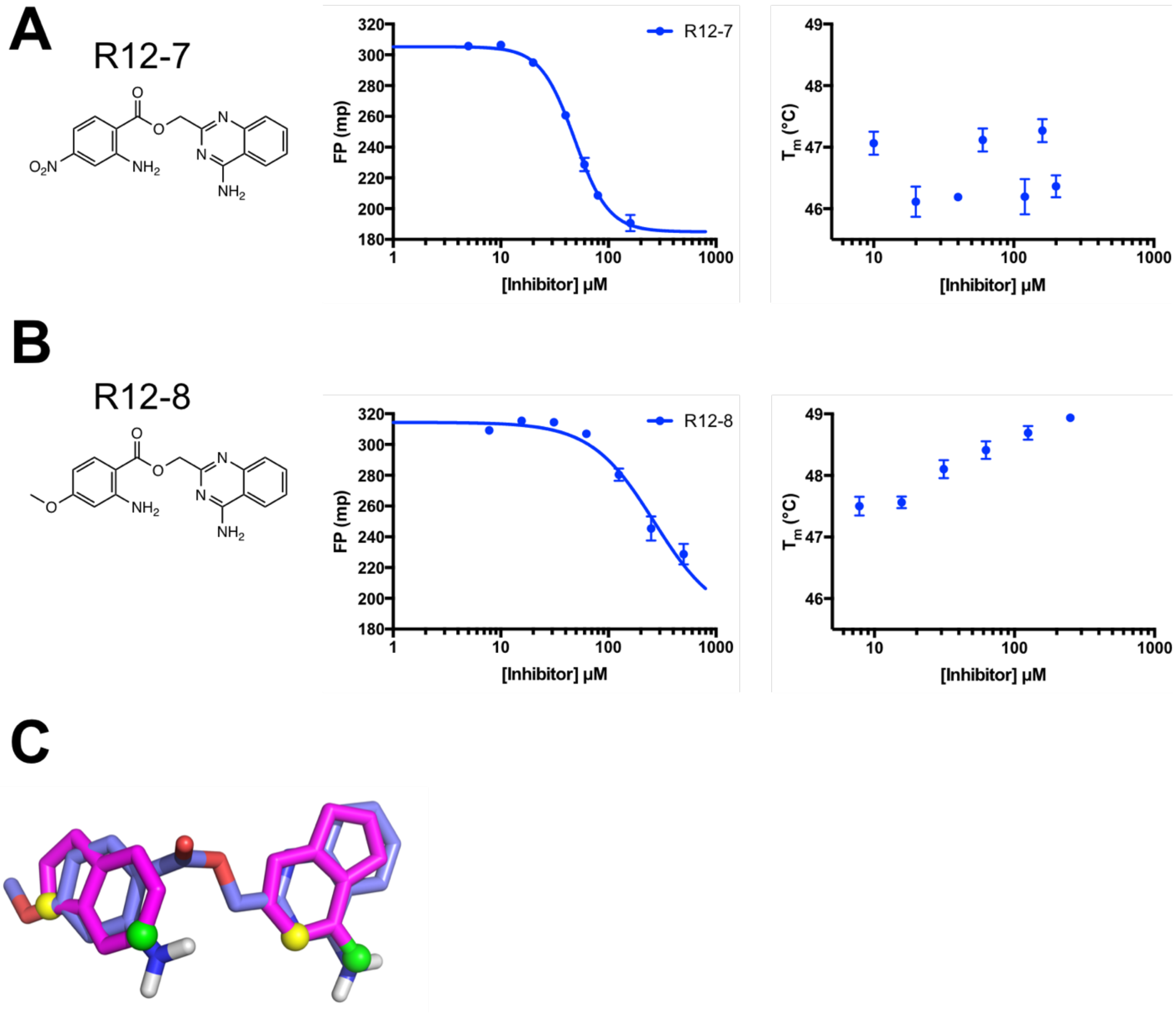


Figure S3: Inhibitors coming from the first round SAR study. (A) The chemical structure of R12-7 and the results of biochemical assays. The middle panel is the FP competition assay and the right panel is the DSF assay. **(B)** The chemical structure of R12-8 and the results of biochemical assays. The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

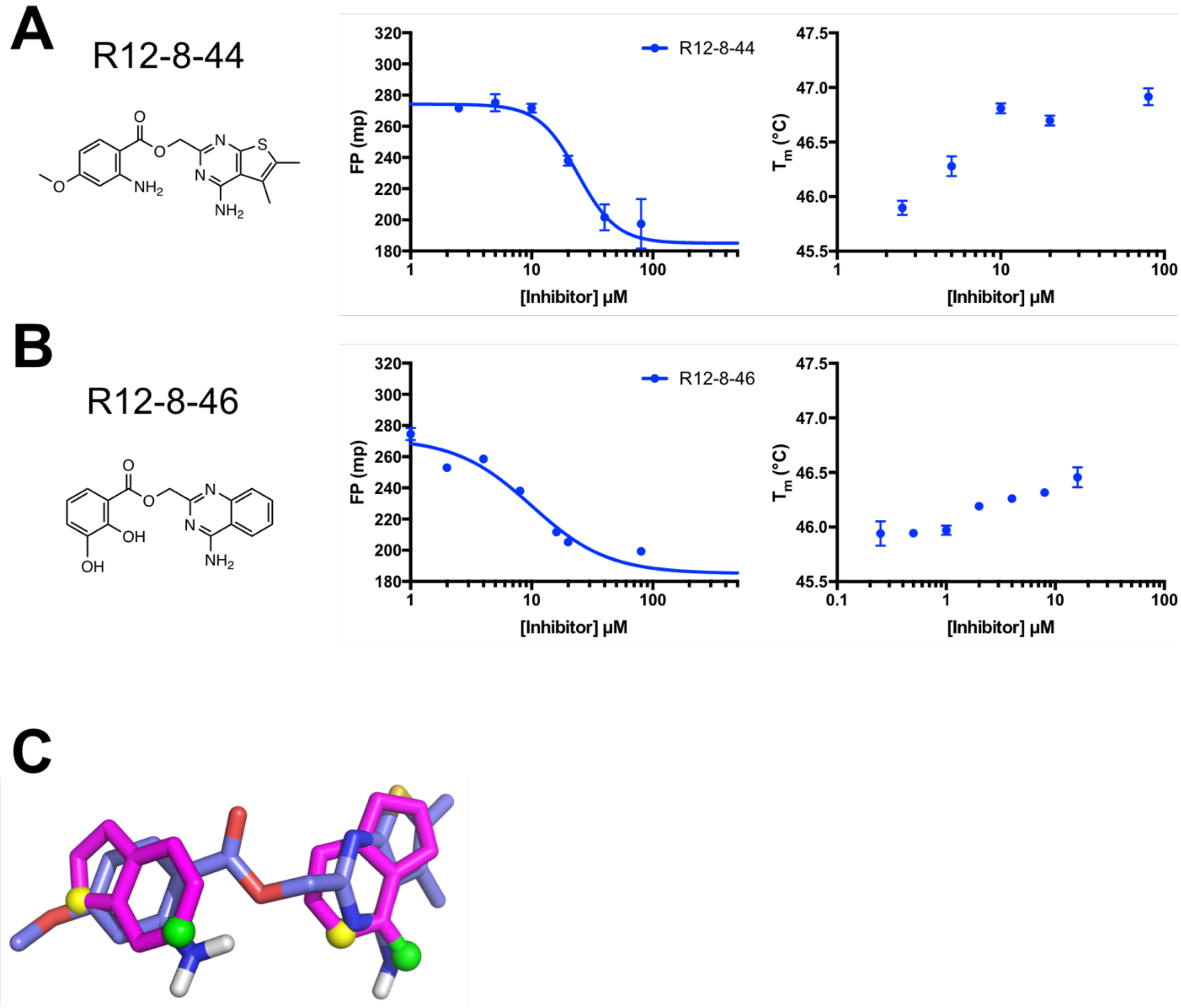


Figure S4: Inhibitors coming from the second round SAR study. (A) The chemical structure of R12-8-44 and the results of biochemical assays. The middle panel is the FP competition assay and the right panel is the DSF assay. **(B)** The chemical structure of R12-8-46 and the results of biochemical assays.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

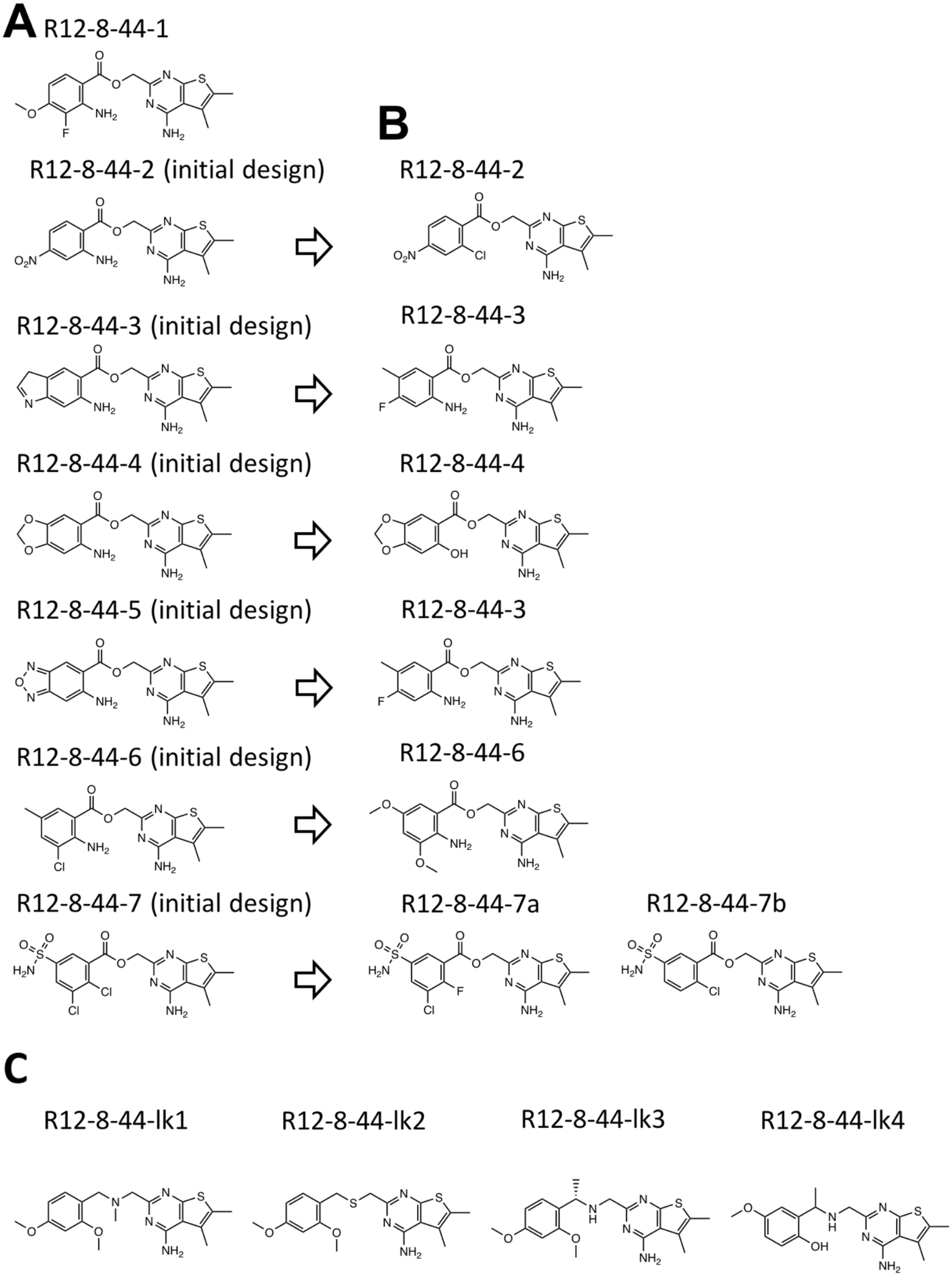

A
R12-8-44-1
R12-8-44-2 (initial design)
B
R12-8-44-2
R12-8-44-3 (initial design)
R12-8-44-3
R12-8-44-4 (initial design)
R12-8-44-4
R12-8-44-5 (initial design)
R12-8-44-3
R12-8-44-6 (initial design)
R12-8-44-6
R12-8-44-7 (initial design)
R12-8-44-7a
R12-8-44-7b
C
R12-8-44-lk1
R12-8-44-lk2
R12-8-44-lk3
R12-8-44-lk4

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.

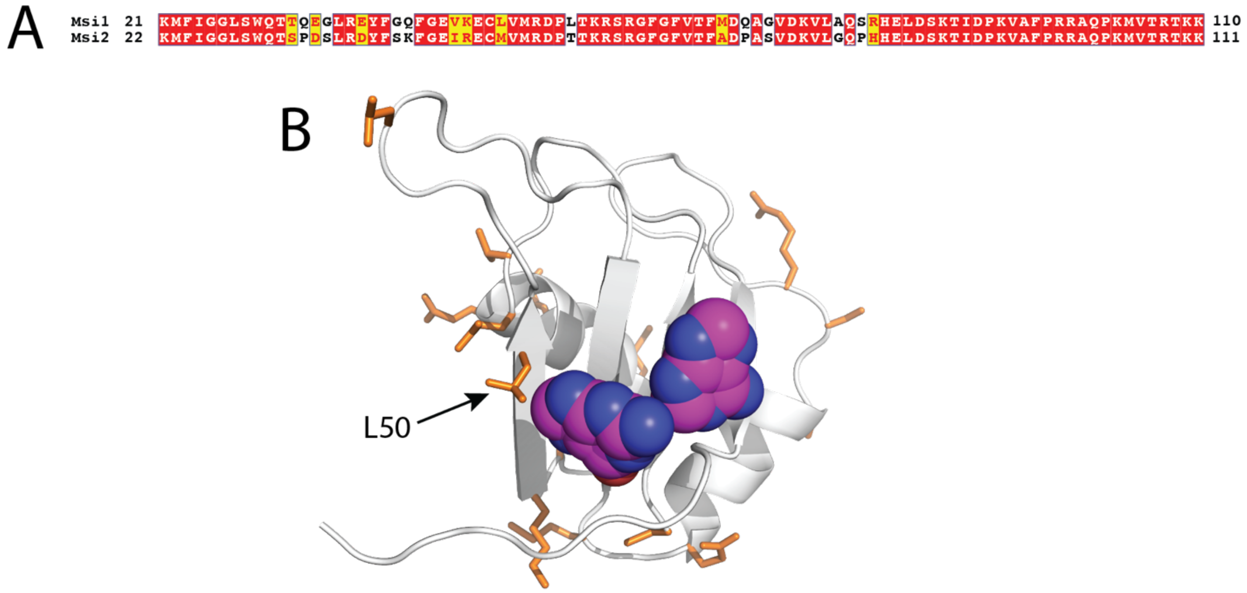

A
Msi1 21 KMFIGGLSWQTTQEGLREYFGQFGEVKECLVMRDPLTKRSRGFGFVTFMDQAGVDKVLAQSRHELDSKTIDPKVAFPRRAQPKMVTRTKK 110
Msi2 22 KMFIGGLSWQTSPDSLRDYFSKFGEIRECMVMRDPTTKRSRGFGFVTFADPASVDKVLGQPHHELDSKTIDPKVAFPRRAQPKMVTRTKK 111
B
L50

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 (165, 211). **(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.

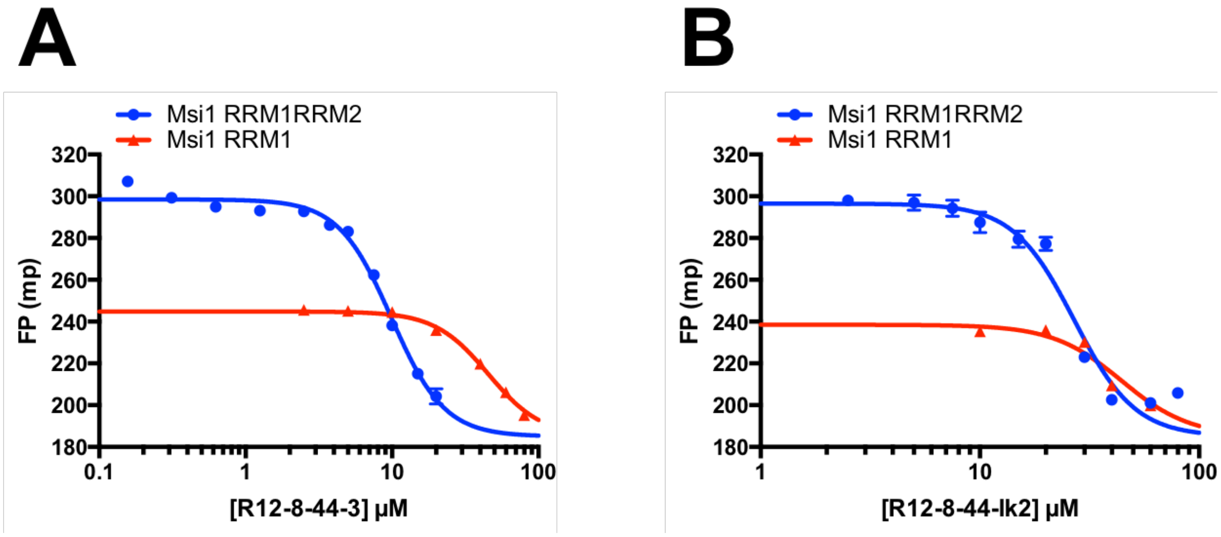
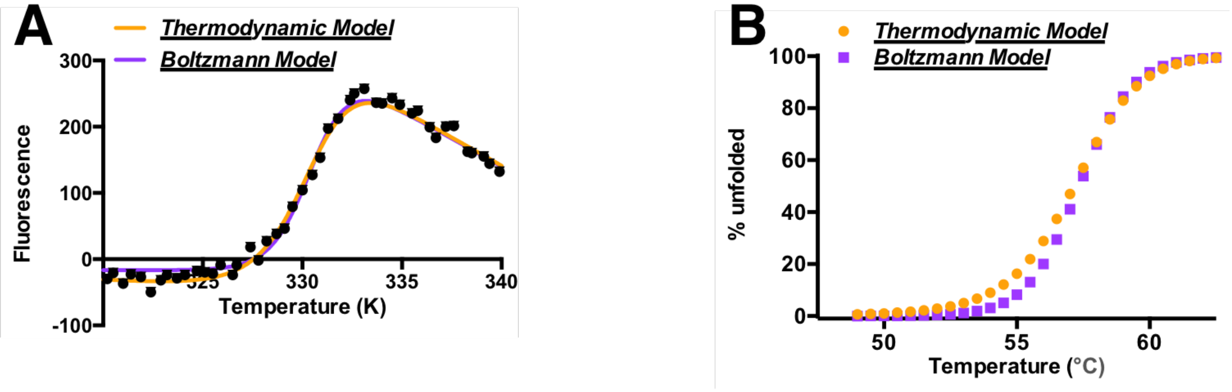


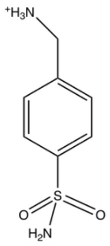
Figure S7: Compounds activity: Msi1 RRM1+RRM2 vs. Msi1 RRM1. (A) FP competition assay of R12-8-44-3 with Msi1 RRM1+RRM2 (*blue*) vs. Msi1 RRM1 only (*red*). **(B)** FP competition assay of R12-8-44-lk2 with Msi1 RRM1+RRM2 (*blue*) vs. Msi1 RRM1 only (*red*).

A.2: Supporting Information for Chapter III

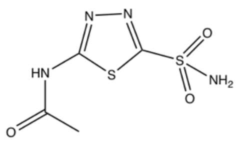
Supplemental Figures



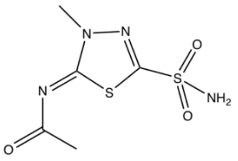
Supplemental Figure S1: The fraction of protein unfolding, as calculated from raw fluorescence data using various models to fit melting curve. **(A)** For analysis using both the thermodynamic model and the Boltzmann model, we included in the fitting linear (non-flat) baselines which necessary due to the temperature dependence of the dye. Both models fit the data well, but they differ slightly at the lower baseline. **(B)** These fits can be used to calculate the fraction of unfolded protein at different temperatures: results match closely between the two models, but they do exhibit a slight difference just below the transition region.



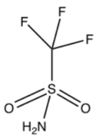
SULFA



ACTAZ



METHZ



TFMSA

Supplemental Figure S2: Chemical structures of the carbonic anhydrase inhibitors used in this study.

Appendix

*Detailed derivation of **Equation 21***

We start from the basic equilibrium between protein folding-unfolding and protein-ligand binding presented in the main text:

$$[U] + [L] \underset{\mathrm{K_U}}{\rightleftharpoons} [F] + [L] \underset{\mathrm{K}_d}{\rightleftharpoons} [FL]$$

From here, we will consider the two specific conditions.

1) when no ligand is present:

$$[P]_T = [F]_0 + [U]_0 \qquad (A.1)$$

$$[L]_T = [L]_0 = [FL]_{\ 0} = 0 \qquad (A.2)$$

$$K_U = [U]_0/[F]_0 \qquad (A.3)$$

$$f_{u0} = \frac{[U]_0}{[U]_0 + [F]_0} = \frac{1}{1 + 1/K_U} \qquad (A.4)$$

Here, f_{u0} is the fraction unfolded protein in the absence of ligand.

Combining **Equation A.1** and **A.4** gives:

$$[U]_0 = f_{u0} \times [P]_T \qquad (A.5)$$

$$K_U = \frac{f_{u0}}{1 - f_{u0}} \qquad (A.6)$$

2) at the EC$_{50}$:

The EC$_{50}$ is defined as the ligand concentration at which the initial fraction unfolded (f_{u0}) is reduced to half its original value. So:

$$[P]_T = [F]_{50} + [U]_{50} + [FL]_{50} \qquad (A.7)$$

$$[L]_T = [L]_{50} + [FL]_{50} = EC_{50} \qquad (A.8)$$

$$K_U = [U]_{50}/[F]_{50} \qquad (A.9)$$

$$K_d = ([F]_{50} \times [L]_{50})/[FL]_{50} \qquad (A.10)$$

$$[U]_{50} = \frac{[U]_0}{2} \qquad (A.11)$$

where **Equation A.11** is derived from the definition of the EC$_{50}$ (the fraction unfolded is half its original value). Combining **Equations A.9** and **A.11** gives:

$$[F]_{50} = \frac{[U]_{50}}{K_U} = \frac{[U]_0}{2\,K_U} \qquad (A.12)$$

Combining **Equations A.7**, **A.11** and **A.12** gives:

$$[FL]_{50} = [P]_\mathrm{T} - [U]_{50} - [F]_{50} = [P]_\mathrm{T} - \frac{[U]_0}{2} - \frac{[U]_0}{2K_U} = [P]_\mathrm{T} - \frac{[U]_0}{2}\left(1 + \frac{1}{K_U}\right) \qquad (A.13)$$

Combining **Equations A.5**, **A.6** and **A.13** gives:

$$[FL]_{50} = [P]_\mathrm{T} - \frac{[U]_0}{2} \times \left(1 + \frac{1 - f_{u0}}{f_{u0}}\right) = [P]_\mathrm{T} - \frac{[U]_0}{2} \times \frac{1}{f_{u0}} = [P]_\mathrm{T} - \frac{[P]_\mathrm{T}}{2} = \frac{[P]_\mathrm{T}}{2} \qquad (A.14)$$

Combining **Equations A.5**, **A.6, A.10, A.12** and **A.14** gives:

$$[L]_{50} = [FL]_{50}\times\frac{K_d}{[F]_{50}} = \frac{[P]_\mathrm{T}}{2}\times K_d\times\frac{1}{\frac{[U]_0}{2K_U}} = \frac{[P]_\mathrm{T}}{2}\times K_d\times\frac{2K_U}{[U]_0}$$

$$= \frac{[P]_\mathrm{T}}{2}\times K_d\times\frac{\frac{f_{u0}}{1 - f_{u0}}}{\frac{1}{2}\,f_{u0}\,[P]_T} = \frac{[P]_\mathrm{T}}{2}\times K_d\times\frac{1}{\frac{1}{2}[P]_T(1 - f_{u0})}$$

$$= \frac{K_d}{1 - f_{u0}} \qquad (A.15)$$

Combining **Equations A.8**, **A.14** and **A.15** gives:

$$EC_{50} = [L]_{50} + [FL]_{50} = \frac{K_d}{1 - f_{\mathrm{u}0}} + \frac{[P]_T}{2} \tag{A.16}$$

$$K_d = (1 - f_{\mathrm{u}0}) \times \left(EC_{50} - \frac{[P]_T}{2}\right) \tag{A.17}$$

The latter two equations (**A.16** and **A.17**) correspond to **Equations 20** and **21** in the main text.

A.3: Supporting Information for Chapter IV

Supplemental Figures

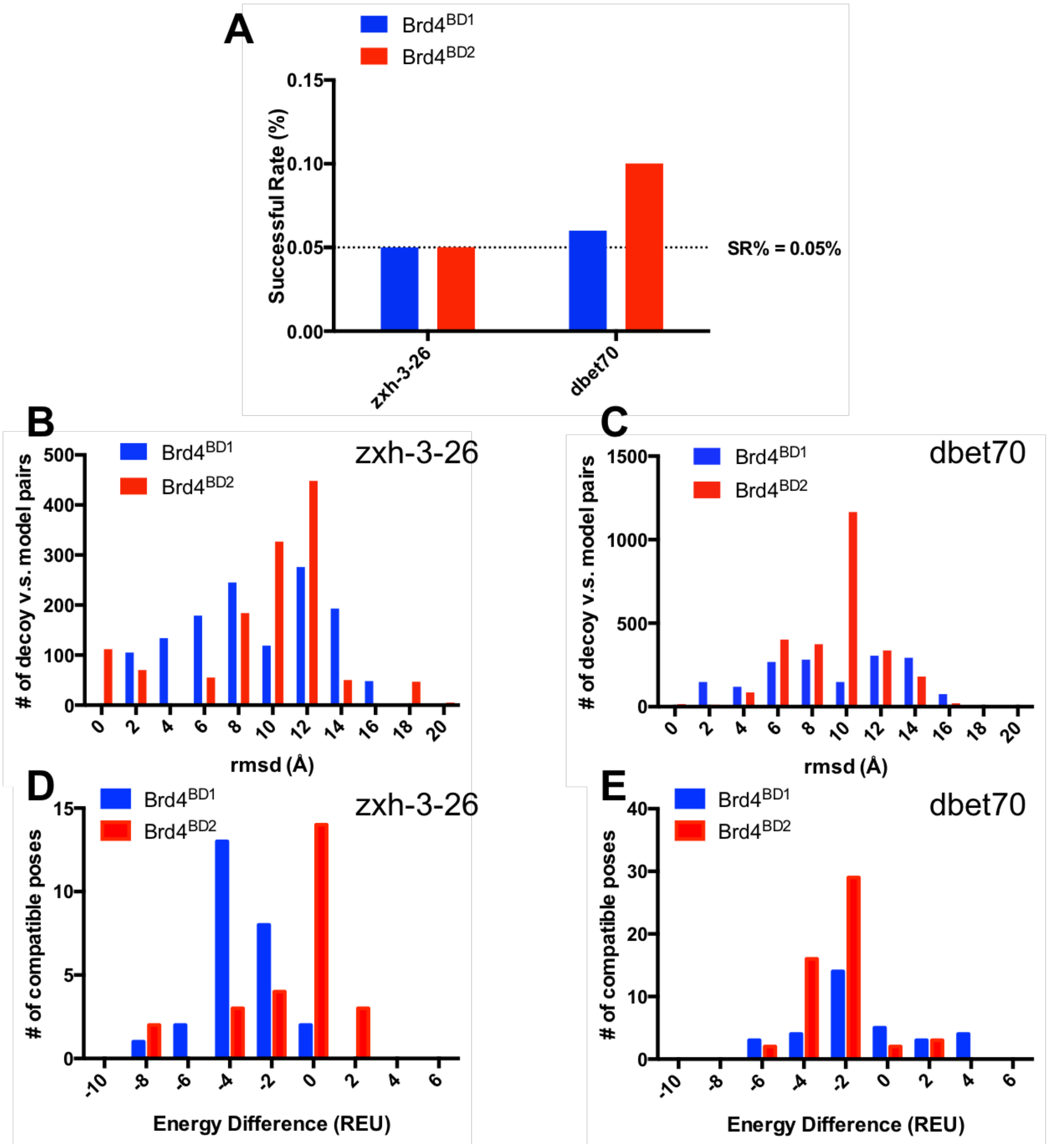


Figure S1: Sequential analysis of BD4$^{BD1/2}$ – zxh-3-26/dbet70 – CRBN. (**A**) Success rate of compatible pose generation of BD4$^{BD1/2}$ – zxh-3-26/dbet70 – CRBN. (**B**) Top 1% compatible poses of BD4$^{BD1/2}$ – zxh-3-26 – CRBN vs top 1% docking decoys of BD4$^{BD1/2}$ – CRBN RMSD distribution. (**C**) Top 1% compatible poses of BD4$^{BD1/2}$ – dbet70 – CRBN vs top 1% docking decoys of BD4$^{BD1/2}$ – CRBN RMSD distribution. (**D**) Top 1% compatible poses vs. initial docking decoys energy difference distribution of BD4$^{BD1/2}$ – zxh-3-26 – CRBN. (**E**) Top 1% compatible poses vs. initial docking decoys energy difference distribution of BD4$^{BD1/2}$ – dbet70 – CRBN. Brd^{BD1} related data is shown in blue and Brd4^{BD2} related data is shown in red.

Supporting Methods

PDB structures used in calculations

The structures were applied to study Brd4 – PROTACs – CRBN cases are Brd4^{BD1} – ligand complex (PDB ID 3mxf), Brd4^{BD2} – ligand complex (PDB ID 5ueu), and CRBN – ligand complex (PDB ID 4tz4).

The structures were applied to study c-Met/EphA2/Stk10 – PROTACs – CRBN/VHL cases are c-Met – ligand complex (PDB ID 3lq8), EphA2 – ligand complex (PDB ID 5ia4), Stk10 – ligand complex (PDB ID 6i2y), CRBN – ligand complex (PDB ID 4tz4) and VHL – ligand complex (PDB ID 4w9h).

The structures were applied to study CDK4/CDK6 – PROTACs – CRBN/VHL cases are CDK6 – ligand complex (PDB ID 5l2i) and CRBN – ligand complex (PDB ID 4tz4). There is no available CDK4 – palbociclib complex structure in the PDB, therefore this CDK4 – ligand complex structure was generated by using CDK4 sequence, complex structures that CDK4 with other ligands, and CDK6 – ligand complex structure.

Computational approach

The two proteins – ligand complex structures (POI – ligand1 complex and E3 – ligand2 complex) were downloaded from PDB and minimized using the Rosetta software suite. The two minimized complex structures were manually combined together with PyMol (202) and the two ligands were oriented towards one another. This combined structure with POI and E3 ligands and their ligands was prepared for docking with running Rosetta pre-pack command:

```
$ pathwaytoRosetta/Rosetta/main/source/bin/docking_prepack_protocol.linuxgccrelease –
database pathwaytoRosetta/Rosetta/main/database –s POI_ligand1_E3ligase_ligand2.pdb –
use_input_sc –extra_res_fa ligand1.params ligand2.params
```

A new complex structure with POI and E3 ligase and two ligands was generated through pre-pack step. This prepacked complex structure was then applied in Rosetta protein-protein docking with command:

```
$ pathwaytoRosetta/Rosetta/main/source/bin/docking_protocol.linuxgccelease –database
pathwaytoRosetta/Rosetta/main/database –s POI_ligand1_E3ligase_ligand2_prepacked.pdb –
nstruct 50000 –use_input_sc –spin –dock_pert 5 20 –partners XY_MN –ex1 –ex2aro –
extra_res_fa ligand1.params ligand2_params –out:file:scorefile score.sc –
score:docking_interface_score 1
```

where X and Y are the chain IDs of POI and its ligand1 and M and N are the chain IDs of E3 ligase and its ligand2. The "`–partners XY_MN`" flag was used to make the ligands only move together with their paired proteins. The "`–nstruct`" flag as used to set the docking decoy number and 50000 was used in this manuscript. The "`–dock_pert`" flag as used to set the freedom of the docking (distance and the rotation angle) and 5 Å and 20 degrees were applied in this manuscript. Finally, top 10% of the docking decoys were picked based on their I_sc values, referring the interface energy between POI – ligand1 complex and E3 ligase – ligand2 complex.

For each linker, maximum number or 1000 conformers were generated using OMEGA (204-206), with command:

```
$ oeomega classic –in linker_name.smi –out linker_name.oeb.gz –maxconfs 1000
```

where the flag "`–maxconfs`" was used to set the conformer number. When the conformers were generated, two extra pieces from two ligands (**Figure 1B**) need to be added on each end of the linker and was used to align these conformers to every docking decoy.

The alignment between one linker conformer and docking decoy pair was performed using one python script written by us, with command:

```
$ python decoy_linker_alignment.py –decoy_pdb decoy.pdb –linker_conformer_pdb
linker_conformer.pdb –linker_atoms_alignment linker_atoms.txt –decoy_atoms_alignment
decoy_atoms.txt –cut_off 0.4 –optimized_linker_conformer linker_conformer_op.pdb
```

where the "`–decoy_pdb decoy.pdb`" was used to pass the docking decoy structure, and "`–linker_conformer_pdb linker_conformer.pdb`" was used to pass the linker conformer. The "`–linker_atoms_alignment linker_atoms.txt`" was used to let user apply a text file with the atom names which would be used to do the alignment, and the same as "`–decoy_atoms_alignment decoy_atoms.txt`".

There are example files offered as Supplemental Documents which can be used as templates to write these two linker/decoy_atoms.txt files. The "–cut_off" flag was used to set the cut off value and 0.4 Å is the default setting. If one linker conformer and docking decoy pair can have a good alignment, RMSD value is less than the cut off value, an optimized linker conformer pdb file would be generated and the file can be named with the flag "–optimized_linker_conformer". If the alignment RMSD didn't pass the cut off value, no optimized linker conformer file would be generated.

A PROTAC file was generated with this new optimized linker conformer file and the two ligands from its paired docking decoy using another python script written by us, with command:

```
$ python buildmodel.py –decoy_pdb decoy.pdb –linker_pdb linker_conformer_op.pdb –
warheads_atom_delete warheads_delete.txt –linker_atom_delete linker_delete.txt –
warheads_pdb warheads.pdb –protac_pdb protac.pdb –protac_rename_pdb protac_rename.pdb –
protein_pdb decoy_protein.pdb
```

where the "–decoy_pdb decoy.pdb" was used to pass the docking decoy structure, and "–linker_pdb linker_conformer_op.pdb" was used to pass the optimized linker conformer. As the linker conformer was generated with two extra parts from two ligands, these two parts would be repeated when linker conformer and two ligands were combined. To exclude these repeated atoms, "–warheads_atom_delete" and "–linker_atom_delete" flags were used to pass two text files "warheads_delete.txt" and "linker_delete.txt", with the atom names which would be deleted from ligand part and linker conformer part. There are example files offered as Supplemental Documents which can be used as templates to write these two warheads/linker_delete.txt files. The "–warheads_pdb warheads.pdb" flag was used to extract the two ligands from the docking decoy and "–protac_pdb protac.pdb" was used to generated the PROTAC file. The "–protac_rename_pdb" flag was used to rename and renumber the atoms in PROTAC file to exclude any repeated atom names or numbers, then the final PROTAC file, "protac_rename.pdb", was generated and was used to combine with two proteins, POI and E3 ligase, extracted from the same docking decoy with flags "–protein_pdb decoy_protein.pdb".

The file with final PROTAC molecule, POI and E3 ligase was minimized with Rosetta. After removing the bad ternary structure models which had the bad Rosetta minimization scores, all the left ternary structures were considered reasonable models and the number of these ternary structures was used to calculate the success rate (SR).