MODULATING PROTEIN FUNCTION WITH SMALL MOLECULES THROUGH COMPUTATIONAL AND EXPERIMENTAL DESIGN TECHNIQUES

By

Yan Xia

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 (John Karanicolas, Ph.D.)

____

(Susan Egan, Ph.D.)

____

(Mark Richter, Ph.D.)

____

(Scott Hefty, Ph.D.)

____

(Liang Xu, Ph.D.)

____

(Emily Scott, Ph.D.)

Date Defended: Aug/15/2014

The Dissertation Committee for Yan Xia certifies that this is the approved version of the following dissertation:

MODULATING PROTEIN FUNCTION WITH SMALL MOLECULES THROUGH COMPUTATIONAL AND EXPERIMENTAL DESIGN TECHNIQUES

Chairperson (John Karanicolas, Ph.D.)

Date approved: Aug/15/2014

ABSTRACT

The ability to modulate protein function using exogenous small molecules is a longstanding goal in chemical biology. Selective activation or inhibition of a particular protein function can help elucidate crucial molecular mechanisms and enables important advances in cell biology. Small-molecule controlled molecular systems also possess tremendous value in bioengineering and biomedical applications: activation of protein function allows the construction of protein switches and biosensor proteins, whereas inhibition of protein function contributes to the development of novel therapeutic agents.

The discovery of small-molecule modulators of function is greatly aided by computational modeling methodologies. By utilizing structural information obtained through X-ray crystallography or NMR spectroscopy, these tools allow efficient and affordable examination of large small-molecule databases and provide quantitative evaluation of the likelihood that a given protein-ligand interaction occurs. Advances in computer algorithms and hardware development continue to accelerate and scale up the computation and lower the cost of this discovery process.

The primary focus of this thesis is the development of structure-based computer-aided methodologies for designing small-molecule modulators of protein function. To this end I explored two parallel paths, one to study activation and one to study inhibition of protein functions. Taken together, my work aims to not only apply rational design strategies to specific proteins, but also demonstrate their general applicability.

The first project, focused on activation of protein function, is built on an approach developed by our laboratory that designs a *de novo* allosteric binding site directly into the catalytic domain of an enzyme. This approach achieves modulation of function by a

novel “chemical rescue of structure approach”: a tryptophan-to-glycine mutation disrupts local structure and induces conformational changes that distort the geometry at the active site; the subsequent binding of exogenous indole then reverts this conformational change and restores the native enzyme structure. The main challenge of generalizing this approach, however, is the difficulty of rationally designing analogous conformational changes in other proteins. It is therefore important to study the possible mechanisms that can be utilized by chemical rescue of structure. Through collaborative and multidisciplinary efforts, we find that the switchable proteins built via the chemical rescue of structure are frequently controlled indirectly by modulating protein stability, rather than discrete conformational changes. Since energetic evaluation of protein stability is far more tractable than designing and/or predicting allosteric conformational changes, this finding demonstrates how chemical rescue of structure can be applied to other systems for building a variety of new protein switches.

To further generalize the applicability of chemical rescue of structure, I sought to extend it to include multiple amino acids, rather than just one. I chose ChxR, a homodimeric response regulator in *Chlamydia*, as the model protein to examine the feasibility of this strategy. I mutated a pair of tryptophans at the dimer interface to glycine in order to disrupt the dimerization of ChxR. To enable the subsequent functional rescue, I used the removed structural elements as a template for ligand-based virtual screening and discovered a set of candidate small molecules that mimic the three-dimensional geometry and chemical properties of the removed chemical moieties. Biophysical characterization of these compounds suggests that the majority of them selectively bind to the engineered ChxR variant. This observation shows promises in extending this generalized design strategy to allow alternate activating ligands.

In parallel to these efforts I carried out studies aimed at inhibition of protein function, as exemplified by my project that uses small molecules to disrupt a protein-RNA interaction. Conventional methods of inhibitor design mostly target RNA-processing enzymes and cannot be generalized to the majority of RNA-binding proteins (RBPs). I contributed to the development of a general strategy of designing competitive inhibitors targeting RBPs. This method involves identifying “hotspot pharmacophores” from the protein-RNA interaction and using it as a template in ligand-based virtual screening. To evaluate the performance of this approach, my collaborators and I applied it to Musashi-1 (Msi1), a protein that upregulates Notch and Wnt signaling pathway and promotes cell cycle progression. Our “hotspot mimicry” approach led us to discover compounds that match the hotspot pharmacophore, and thus enabled the development of novel inhibitors to the Msi1/RNA interaction that we validated in both biochemical and cell-based assays. This approach extends the “hotspot” paradigm from protein-protein complexes to protein-RNA complexes, and helps establish the “druggability” of RNA-binding interfaces. It is the first example of a rationally-designed competitive inhibitor for a non-enzymatic RNA-binding protein. Owing to the simplicity and generality, I anticipate that the hotspot mimicry approach may lead to the identification of inhibitors of other protein-RNA complexes, which in future may serve as starting points for the development of a novel class of therapeutic agents.

To the memory of my grandfather, Changming Duan (1931 - 2014)

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help of many individuals, and I would like to offer my sincerest gratitude to all of them.

First and foremost, I offer my utmost gratitude to my mentor, Dr. John Karanicolas, for supporting me over the past five years. John has given excellent guidance, unconditional encouragement and endless patience, and provided a wonderful environment for doing research. John has always been a role model to me, and I hope that someday I could be as caring, supportive, enthusiastic and energetic as him.

I am grateful to all the wonderful labmates, past and present, for the valuable discussions. Special thanks to Jimmy Budiardjo, Jittasak Khowsathit and Ragul Gowthaman for their friendship and all the thought-provoking advice.

I am very thankful to my committee members: Dr. Susan Egan, Dr. Mark Richter, Dr. Scott Hefty, Dr. Liang Xu and Dr. Emily Scott for their time and suggestions on the thesis. I thank Dr. Liang Xu and Dr. Emily Scott for willing to serve on my thesis committee at the last moment. Dr. Eric Deeds has been a source of knowledge in mathematics and statistics, and I greatly appreciate all his help through the years.

I would like to convey my deepest appreciation and affection to my family, especially my parents, Jianguo Xia and Lin Duan, and my grandfather, Changming Duan, who passed away earlier this year. I wish my grandpa were here with me to see and celebrate my graduation.

Finally, I would like to thank Shengkai Sun for standing by my side through the good time and the bad. Shengkai believed in me even when I did not have faith in myself.

It has been a wonderful journey with you, and I am looking forward to the roads ahead of us.

TABLE OF CONTENTS

Chapter I.
Introduction

The interaction between proteins and small-molecule ligands is a fundamental molecular phenomenon. The binding of a small molecule can result in the selective modulation of protein function, which in turn can induce many important biological processes (*1-4*). Thus, the ability to systematically manipulate protein function with small molecules is highly desirable and in great demand in many facets of biological research. In the realm of fundamental biological studies, the selective activation or inhibition of protein function by small molecules has helped unravel the mechanism of many important biological processes (*4-7*). From a biotechnological perspective, small molecule-directed activation of engineered proteins can enable the development of biosensor proteins (*8, 9*) and genetic circuits (*10*); inhibition of protein activity by a small molecule, meanwhile, can serve as an indispensable method for prevention and treatment of diseases (*11-13*).

Rational design of selective small-molecule modulators relies on structural insights, typically gained from X-ray crystallography or NMR spectroscopy. Computational tools and methodologies can further utilize the structural information and accelerate the design and identification processes. Together, this structural-based computer-aided design paradigm provides great robustness and generality for systematically designing protein structures, functions and interactions, and has enabled development of various small molecule modulators in different biological systems.

“Turning on” Function - Designing Protein Switches and Biosensor Proteins

The use of small molecules to selectively activate the function of engineered proteins gives rise to the construction of protein switches and biosensor proteins that serve as powerful chemical biology tools. In cell biology, protein switches have been applied to control various cellular processes, ranging from cell morphology to reprogramming cellular signaling pathways (*2, 4, 14-16*). Biosensor proteins have also resulted in a plethora of practical applications, such as detecting chemicals in food or the environment (*17*).

Conventional Approaches

The conventional strategies of engineering small-molecule dependent activity into proteins utilize allosteric protein domains that naturally bind to specific small molecules as the starting points. The conformational change induced by small molecule binding is coupled to the activation of the same or a separate domain. Catalytic domains are commonly chosen since the enzymatic reaction can amplify a single event of small molecule binding to the catalysis of multiple molecules of fluorescent or colorimetric products, allowing a more sensitive and quantitative detection of the analyte.

Allosteric changes induced by small molecule binding can be relayed to increase catalytic activity via several methods (*18*): domain insertion is one such strategy that entails fusing the allosteric domain to a separate catalytic domain (*19, 20*). To achieve optimal spatial arrangement of two domains and preserve their native structures, circular permutation is frequently used to create multiple positions at which the guest domain can be inserted into the host domain. Successful designs can transduce the structural change from the allosteric domain to the catalytic domain and modulate its activity. A different strategy couples small molecule binding to the correct assembly of enzyme structures: the allosteric domain is placed in the middle of the catalytic domain and separates it into two inactive halves. The allosteric changes upon small molecule binding, in selected cases, can bring the two halves to close proximity and lead to the assembly of active enzymes (*21*). A closely related approach utilizes protein splicing and constructs an inactive intein fragment by inserting an allosteric domain to intein (*6*). The engineered intein module is then inserted into a catalytic domain and inactivates it. The binding of small molecule activates the intein activity so that it self-cleaves out of the catalytic domain, which ultimately restores enzymatic activity. In addition to fusing separate allosteric domain and catalytic domains, *de novo* catalytic function can be directly engineered into an allosteric domain. However, this strategy is extremely limited with respect to the protein system, and to date only one successful application of this approach has been reported (*22*).

The main challenge for these strategies, though, is the difficulty associated with predicting the detailed mechanism by which small-molecule binding modulates protein activity. Computational approaches are limited by the inaccuracies in biophysical representation of the molecular interactions and the complexity of conformation space. For this reason, these conventional approaches usually require the screening of multiple domain arrangements and various linker compositions to identify the optimal spatial arrangements that lead to allosteric regulation of enzyme activity upon small molecule binding.

Chemical Rescue of Structure

Rather than using a natural allosteric transition, small molecule-dependent function can be *de novo* engineered into a protein domain that is not naturally allosteric. This entails constructing inactive protein variants, typically via single-point mutations; the subsequent addition of exogenous ligand that complements the removed structural elements can then restore the missing active-site moiety, and in turn regain the function. This *de novo* method is generally referred to as "chemical rescue" (*23*). The existing examples mainly focused on mutations at enzyme active sites and most notably used to rescue the function of a Src kinase variant in living cells (*24*). However, the chemical rescue of enzyme active sites often occurs through unanticipated mechanisms: in the Src kinase example, the arginine-to-alanine variant is most effectively rescued by imidazole, rather than guanidinium, which is the intuitive structural complement of the deleted chemical group.

Inspired by the observation that internal cavities caused by mutations can sometimes be complemented by the binding of hydrophobic small molecules (*25*), our laboratory developed a different paradigm of chemical rescue, termed "chemical rescue of structure" (*26*). This method seeks to indirectly activate protein function by mutating residues at a separate buried site, instead of immediately at the active sites. A buried tryptophan residue is typically selected and mutated to glycine, with the expectation that the removal of the tryptophan sidechain may result in a more substantial energetic destabilization than other amino acids due to its size and hydrophobicity.

When using this method, we seek to identify a buried tryptophan sidechain that serves as a structural "buttress", in that it supports the structural integrity of the nearby protein architecture. The deletion of a buttressing tryptophan can potentially result in various structural consequences, ranging from a discrete conformation change to varying extents of protein unfolding and leads to the loss of protein function. The subsequent restoration of the buttress by the addition of exogenous indole, which complements the deleted tryptophan sidechain, can reinstate the original protein structure, and thus reactivate protein function. Chemical rescue of structure has been applied to β-glycosidase and led to the construction of an indole-controlled enzyme switch in living cells (*26*).

Chemical rescue of structure is not limited to catalytic domains; in fact, it has also been successfully applied to GFP and multiple essential proteins from *E.coli* (*27*). Another natural extension of chemical rescue of structure is mutating multiple amino acids, instead of a single tryptophan. In this scenario, deleting a constellation of atoms from neighboring hydrophobic sidechains creates the larger buried cavity, and the binding of small molecules that resemble the structural geometry and chemical property of the removed chemical groups can potentially restore the protein structure and function.

“Turning off” Function - Designing Small Molecule Inhibitors

Discovery and characterization of small molecules that inhibit cellular function of specific proteins is an active research area that has received great attention from the scientific community. In basic cell biology, small molecule inhibitors that can permeate the cell membrane provide a means of hindering the normal function of a specific protein target and are indispensable tools for elucidating the intracellular role of that protein (*28, 29*). In biomedical applications, identifying small molecule inhibitors and transforming them into potent lead compounds are essential activities in modern drug discovery. These small molecules frequently serve as invaluable starting points for developing novel and selective therapeutic agents that lead to the repression of aberrant cellular functions by mutated or dis-regulated protein targets in various diseases (*13, 30*).

Conventional Strategies

High-throughput screening (HTS) is a method for discovering small molecule inhibitors and/or optimizing lead series (*31, 32*). HTS entails testing a library of chemically synthesized or naturally occurring small molecules, and enables the rapid identification of active compounds for a protein target via biochemical or cellular assays. Using robotics and data processing technologies, it is possible to test libraries containing millions of compounds. However, HTS is generally very expensive and time consuming to perform, and the success rate of HTS is limited by the compounds in theses selected libraries, which are biased towards chemotypes that have proven successful against popular therapeutic targets, and thus may not be suitable for different classes of protein targets (*33*).

For these reasons, traditional HTS has been complemented by structure-based computer-aided methodologies in recent years, which provides a rational means to design and/or identify small molecule inhibitors. Structure-based approaches rely on the X-ray crystallographic or NMR-based structures of a protein target to rationally design small molecules that are likely to bind and inhibit the protein function. Computational tools and algorithms often aid the discovery process to optimize potential hit compounds identified from HTS or structure-based methods, or to screen for novel inhibitors from larger virtual libraries.

Virtual screening can be performed using two general paradigms: ligand-based or structure-based screening (*34*). Ligand-based methods collectively consider the structures of diverse small-molecule ligands that are known to bind to the protein targets, and extract crucial features that the new compounds need to possess. These structural features are included in models known as "pharmacophore" and are compared to a library of candidate small molecules to identify the ones compatible with the pharmacophore model. In contrast, structure-based screening does not require the complex structure of protein and small-molecule ligand. This methodology involves docking different conformations of candidate small molecules into postulated binding sites on protein. The likelihood of ligand binding is then predicted by score functions that model molecular interactions and estimate binding affinities.

Traditional vs. Non-Traditional Drug Targets

The primary targets in traditional inhibitor discovery are protein receptors that naturally bind to small-molecule ligands. These targets include signaling proteins whose functions are regulated by small molecules, and enzymes using small molecules as substrates (*33*). In recent years, however, increased attention has been shifted to proteins interacting with other types of macromolecules, such as nucleic acids or other proteins (*35, 36*). These interactions are crucial to a wide variety of biological processes, and therefore offer more potential drug targets for therapeutic intervention.

These non-traditional interactions remain enticing, but extremely difficult, targets for developing small molecule inhibitors. The rare instances of small molecules that bind to the non-traditional interfaces result in the lack of starting points for inhibitor development. Compared to the deep and hydrophobic binding pockets in traditional targets, the general flatness of the interface further complicates the problem, as it is difficult to identify druggable sites on this type of interface (*37*).

These challenges imply that preexisting small molecule libraries and computational methodologies, which were developed and parameterized for traditional targets, may not be appropriate for these non-traditional targets. Recently, a systematic examination from our laboratory revealed the differences between inhibitors for traditional drug targets and protein-protein interactions (PPI) (*33*). Due to the flat interface, inhibitors for PPI are less buried than the traditional counterparts, and owing to this reduce ligand efficiency more atoms are required to achieve a given potency. Small molecules included in the screening libraries have been preselected to cater to the traditional classes of protein targets, and therefore these libraries may be ill-suited for non-traditional proteins in ligand-based screening. The systematic differences between these two classes of targets also lead to complications in structure-based virtual screening. The parameters in popular docking tools are trained and optimized against the deep binding pockets from traditional targets. Therefore, their performance deteriorates when dealing with the relatively exposed binding modes of non-traditional inhibitors (*11, 37*).

Fortunately, advances in protein-protein interactions provide a promising strategy for designing inhibitors to non-traditional targets. In protein-protein interactions, binding affinity is not evenly distributed over the binding interface. Rather, a small cluster of "hotspot" residues contribute to most of the binding energy (*38-40*). This observation has naturally led to the idea of using these "hotspot" interactions as the templates in drug discovery (*11, 40, 41*): small molecule inhibitors can be rationally designed by mimicking the structural geometry and chemical properties of “hotspot” residues. Retrospective inspection reveals that certain inhibitors identified via irrational HTS actually resemble the structure of a cluster of “hotspot” residues at the interaction interface, which validate the feasibility of this “hotspot” mimicry approach (*42, 43*). To predict or identify the hotspot residues in PPI, a number of computational methods, along with the experimental alanine scanning, have been developed (*40, 44, 45*), and enabled the development of inhibitors for several PPI targets (*43, 46-48*).

Targeting Protein-RNA Interactions

Protein-RNA interactions represent a class of non-traditional targets that offer great promise and opportunity for drug development. RNA-binding proteins (RBPs) bind to single or double stranded RNA in cells and play crucial roles in diverse cellular processes. RBPs participate in post-transcriptional control of RNAs, which is one of the major ways of gene regulation during development (*49, 50*). The post-transcriptional regulation occurs at many different stages in RNA metabolism, including splicing (*51*), polyadenylation (*52*), mRNA stability (*53*), mRNA localization (*54*) and translation (*55*), and these regulatory functions are achieved through either the nucleic acid processing by RNA-binding enzymes (such as RdRPs and reverse transcriptase) (*56, 57*) or the specific RNA-binding by non-enzymatic RBPs (such as HuR and Musashi 1) (*58-60*). Due to the versatile functions carried out by RBPs, modifying and controlling their interactions with the cognate RNAs is an essential means for elucidating the mechanisms of important biological processes, and developing therapeutic interventions of various diseases.

To date, there exist only limited examples of rationally designed small molecule inhibitors that target protein-RNA interactions. Based on the design strategies, they can be categorized into three general classes: The first class contains nucleoside analogues (eg. NRTIs for HIV), which mimic the chemical structures of natural-occurring nucleosides (*61-63*). They can competitively bind to the orthosteric sties and interfere with the synthesis of nucleic acids through diverse mechanisms (*64-66*). The main advantage of nucleoside analogues is the extremely straightforward design scheme. However, nucleoside analogues alone provide only inadequate binding affinities due to limited sizes of compounds. The implication is that the spontaneous binding and correct functioning of nucleoside analogues require the coupling to the energy provided by enzymatic reactions. Therefore, their usage is strictly confined to nucleic acid-processing enzymes and cannot be generalized to non-enzymatic instances. Furthermore, nucleoside analogues are not specific to the designed targets due to the close resemblance to naturally occurring nucleosides and the off-target binding can lead to severe side effects (*67, 68*). The second class consists of allosteric inhibitors which bind to secondary sties on the protein targets and shift the conformation ensemble towards an inactive state (eg. NNRTIS for HIV) (*63, 69, 70*). They can be used to target both enzymatic and non-enzymatic instances of RBPs and deliver desirable binding affinity and selectivity. The main disadvantage, though, is that the rational design of allosteric inhibitors is extremely difficult, as it entails the identification of both allosteric sites and small molecules that bind to such sites and transition the protein into inactive conformations. The third class of inhibitors are RNA-binding small molecules originated from docking small molecules onto RNA structures, instead of RBPs (*71, 72*). This strategy also provides a rational way of designing inhibitors targeting non-enzymatic RBPs, but the binding of small molecules may affect the normal functioning of the RNAs. In addition, without the structural information of bound RNAs there is no guarantee that the targeted RNA conformation cannot bind to the protein partner.

The non-traditional nature of protein-RNA interactions may most likely be the cause for the limited success in previous rational designs. RBPs have naturally evolved to bind to RNAs, in contrast to the traditional protein targets that bind to small molecules (*33*). This difference poses challenges to ligand-based virtual screenings because of the lack of templates as the starting points. There exists few instances of natural small molecules known to bind at protein–RNA interfaces (except for nucleotides) and the natural binding partners (cognate RNA) are too large to guide the design of small molecule inhibitors and cannot be used as a whole to infer their druggability (*11, 33*). Structure-based virtual screening (i.e. docking) is also expected to be troublesome when applying to the shallow and polar interfaces of the non-traditional interactions (*11, 37*).

Computational Tools

Rapid Overlay of Chemical Structure

Rapid overlay of chemical structure (ROCS) is employed in ligand-based virtual screening to identify small molecule conformers that mimic the 3D geometry and chemical properties of the template. ROCS is a ligand centric 3D method that aligns two compounds by their similarity in shape (*73*). ROCS can be used to match (search) the shape of a large collection of compounds in a database to a query molecule. The input and query molecule are aligned and optimized very rapidly using atom-centered Gaussian functions to maximize the overlap of volume between them. The similarity of two molecules can be quantified by Tanimoto score with the following equation:

$$Tanimoto_{f,g} = \frac{O_{f,g}}{I_f + I_g - O_{f,g}}$$

The I terms are self-volumes of each molecule, while the O term is the overlap between two molecules.

Along with shape matching (ShapeTanimoto), ROCS also includes pharmacophoric features (ColorTanimoto), such as hydrogen bond donor, acceptor, and aromatic rings, to score the alignment. Previous studies shows that the combination of both shape and color score (TanimotoCombo) gives better enrichment (*74-76*).

Rosetta Software Suite

The structure evaluation and prediction in my studies have been primarily carried out using the Rosetta software suite (*77*). Rosetta is a popular object-oriented software package that provides versatile and robust tools for predicting and designing protein structures, and their interaction with other macromolecules.

The general strategy of Rosetta is to capture the natural variation observed in protein structures. Protein systems can be described and modeled using sets of degree of freedoms (DOFs), such as φ/ψ/ω/χ angles in protein folding and translation/rotation in docking. A given conformational manipulation results in changes in the DOF space, which in turn determine the coordinates of atoms in 3D Cartesian space. The energetic consequence of such structrual manipulations can be evaluated by the Rosetta energy function using the Cartesian coordinates of atoms. In computational modeling of protein structures, this workflow of "DOFs space (torsion space) → Cartesian space → Energy" is applied for a large number of iterations to explore the relevant conformational space and locate the conformation(s) with the lowest energy. Minimization methodologies, such as Monte Carlo simulation and gradient descent, are used in the process to guide the trajectories of the simulation.

In Rosetta, a molecular system is modeled by an object called `Pose`. The `Pose` object represents a certain state of a molecular system, and contains all the structural information that is required to completely describe the system in both torsion and Cartesian space. In addition to the storage of structural information, the `Pose` object contains two essential components that facilitate the modeling on the molecular system: `Conformation` and `Energy`. Upon a structural movement, the `Conformation` component updates the DOFs in the torsion space and translates the changes to the Cartesian coordinates. Given an updated conformation, the `Energy` component is responsible for evaluating and storing the energy of the current structure.

Structural manipulations in Rosetta are achieved by `Mover` objects. The `Mover` objects operate on `Pose` objects and apply different types of conformational perturbation by instructing the update of the torsion space. `Mover` objects allow a variety of manipulations to the protein conformation, ranging from the perturbation of backbone torsion angle to the minimization of structures. In a single iteration of modeling, these `Mover` objects can either executed for multiple cycles or be combined together to perform a composite modification. The conformational search space can be limited and controlled using the `MoveMap` object, which specifies the DOFs that need be held rigid during simulation.

Rosetta score function evaluates the energy of `Pose` objects. The Rosetta score function is a weighted sum of independent energy terms. These energy terms captures the likelihood of a particular conformation, as well as the fitness of sequence given a protein conformation. Over 20 energy terms are available in the score function, and they mainly describe van der Waals attractive/repulsive interactions, solvation energy and hydrogen bonding and electrostatic energy. Each energy component is calculated by a certain "energy method". For example, the van der Waals interactions are modeled by Lennard-Jones potential. The attractive and repulsive components are separated with different weights. The solvation energy is described by Lazaridis–Karplus solvation free energy. The user can customize the score function by either adding additional energy terms or setting the weights of any unwanted term(s) to zero.

Collaborators' Contribution

The cI repressor assay in Chapter II was performed by Dr. Nina DiPrimio (Department of Bioengineering, University of California, Berkeley). The crystal structure of W57G +5 GFP was solved by Dr. Scott Lovell (Protein Structure Laboratory, University of Kansas). The hydrogen-deuterium exchange experiment in Chapter II was performed by Dr. Theodore Keppel (Department of Chemistry, University of Kansas).

The "get_rna_pharmacophore" Rosetta application in Chapter IV was developed together with Ragul Gowthaman. Optimized small molecules in Chapter IV were synthesized by Dr. Steven Rogers (Department of Medicinal Chemistry, University of Kansas). The DSF assay in Chapter IV was performed by Dr. Lan Lan (Department of Molecular Biosciences, University of Kansas).

Chapter II.

The Designability of Protein Switches by Chemical Rescue of Structure: Mechanisms of Inactivation and Reactivation

Abstract

The ability to selectively activate function of particular proteins via pharmacological agents is a longstanding goal in chemical biology. Recently, we reported an approach for designing a *de novo* allosteric effector site directly into the catalytic domain of an enzyme. This approach is distinct from traditional chemical rescue of enzymes in that it relies on disruption and restoration of structure, rather than active site chemistry, as a means to achieve modulated function. However, rationally identifying analogous *de novo* binding sites in other enzymes represents a key challenge for extending this approach to introduce allosteric control into other enzymes. Here we show that mutation sites leading to protein inactivation via tryptophan-to-glycine substitution and allowing (partial) reactivation by the subsequent addition of indole are remarkably frequent. Through a suite of methods including a cell-based reporter assay, computational structure prediction and energetic analysis, fluorescence studies, enzymology, pulse proteolysis, x-ray crystallography and hydrogen-deuterium mass spectrometry we find that these switchable proteins are most commonly modulated *indirectly*, through control of protein stability. Addition of indole in these cases rescues activity not by reverting a discrete conformational change, as we had observed in the sole previously reported example, but rather rescues activity by restoring protein stability. This important finding will dramatically impact the design of future switches and sensors built by this approach, since evaluating stability differences associated with cavity-forming mutations is a far

more tractable task than predicting allosteric conformational changes. By analogy to natural signaling systems, the insights from this study further raise the exciting prospect of modulating stability to design optimal recognition properties into future *de novo* switches and sensors built through chemical rescue of structure.

Introduction

Important advances in cell biology have been enabled through the ability to selectively activate proteins involved in key processes (*5, 7, 15, 78-81*). We recently described an approach for introducing allosteric control into enzymes via a strategy termed “chemical rescue of structure” (*26*). This strategy entails introducing one or more cavity-forming mutations into a protein core at “buttressing” locations, i.e. where specific sidechains are critical for maintaining the structural integrity of the active site. Deletion of these “buttressing” residues leads to distortion of the active site geometry, and accordingly loss of enzyme activity. The subsequent addition of an exogenous compound that matches the deleted moiety is then expected to restore the “buttress” by binding in the cavity, and thus restore protein structure and activity.

Our previous studies (*26*) focused on β-glycosidase from *S. solfataricus* as a model enzyme. We introduced a tryptophan-to-glycine mutation (W33G) at a site close to (but distinct from) the active site, and found the ratio k_{cat}/K_m for this mutant to be about 730-fold worse than that of the wild-type enzyme. Upon solving the crystal structure of this mutant, we found that a very local conformational change distinguished it from the wild-type structure: a single nearby active site residue had shifted away from the active site to fill the cavity produced by the mutation. The change in position of this active site residue led to a loss of contact with the substrate, explaining the loss of function. We then found that exogenous indole could be used to completely restore activity to the mutant, with both k_{cat} and K_m reaching the corresponding values of the wild-type enzyme. The crystal structure of the mutant enzyme in complex with indole revealed that indole occupied exactly the cavity created by the mutation. This in turn perfectly restored the active site geometry, explaining the complete rescue of enzyme activity.

In contrast to chemical rescue of structure, most approaches for building ligand-dependent activity into enzymes have involved fusing a gene encoding some naturally-occurring allosteric “binding domain” (for the desired ligand) into a gene encoding some naturally-occurring “output domain” (for the desired activity) (*82*). By using screens or selections to sift through the large number of potential insertion points and linkers, these fusions of existing protein domains have led to a variety of synthetic “switchable” proteins that are activated through allostery by the binding of an effector ligand (*1, 15, 83-86*). The chemical rescue of structure approach is unique in that it introduces a ligand-binding site *directly* into the “output domain,” rather than rely on allosteric coupling to a separate “binding domain.” This alleviates the need for a naturally-occurring allosteric binding domain as a starting point, but instead requires that ligand binding alters *intradomain* function.

In the β-glycosidase example described above, the structural consequences of the cavity-forming mutation were indeed transduced to the active site, leading to loss and subsequent rescue of function. However, identifying cavity-forming mutations that induce analogous conformational changes in other proteins represents a key challenge in building further *de novo* switches and sensors by chemical rescue of structure. Here, we seek to explore the general considerations that make this approach possible. In particular, we aim to address the following questions: How frequently does a single W→G cavity-forming mutation induce loss of function? How might one select sites that will lead to protein inactivation and reactivation by indole? And most importantly, must we explicitly tackle the challenge of modeling conformational changes resulting from cavity-forming mutations in order to predict sites at which chemical rescue of structure may be applied?

Materials and Methods

Plasmids and Cloning

For the cI repressor assay, genes encoding the homodimeric protein targets were obtained from *E.coli* K12 MG1655 genome via PCR. The gene encoding the cI lambda repressor DNA-binding domain was acquired from lambda phage DNA. We cloned the N-terminal DNA-binding domain of cI (residues 1-132) into pth7035K (R6K origin and Kanamycin R). Driving the cI truncation was a constitutive promoter generated in-house. Additionally, at the C-terminus of the cI truncation we included a 6aa flexible linker. We then subcloned in amplicons of the genes encoding the homodimeric protein targets to generate cI–target gene fusions. A summary of the homodimeric protein targets is included in **Table 2.1.**

To study dimerization of these target proteins, we engineered reporter cells containing the Pr promoter driving GFP expression. The Pr promoter segment was obtained from lambda phage DNA and was subcloned into pth7033C containing the ColE2 origin of replication and chloramphenicol resistance marker. The Pr reporter plasmid was transformed into *E. coli* DIAL strain JI (*87*).

A plasmid containing the *E.coli* β-glucuronidase (β-gluc) gene was generously provided by Bret Wallace and Matt Redinbo (University of North Carolina). A plasmid containing the +5 GFP gene was generously provided by David R. Liu (Harvard University).

+5 GFP is a 3-point mutant (G65T/R80Q/V206A) of superfolder GFP (*88*). Genes encoding β-gluc and +5 GFP were amplified using PCR. The DNA fragments were cut with the *Ssp*I restriction enzyme and cloned into a ligation-independent vector pTBSG1 generously provided by F. P. Gao (University of Kansas). The final construct encodes an N-terminal 6xHis tag, a 17-amino-acid linker, and the gene of interest (β-gluc or +5 GFP) under control of a T7 promoter.

Protein expression and purification

Recombinant β-gluc and +5 GFP were expressed from a pET28a vector in *E. coli* Rosetta 2(DE3)pLysS cells at 15ºC overnight. The cells were resuspended in Lysis Buffer (50 mM Tris, 150 mM NaCl, 5 mM imidazole pH 8.0) and sonicated for 10 minutes (Fisher Scientific Sonic Dismembrator Model 100). The cell lysates were then centrifuged at 15,000g for 30 min. The β-gluc remained in the supernatant, which was purified by HPLC affinity chromatography with Ni-chelated Sepharose Fast Flow Resin (GE Healthcare), followed by a HiLoad 16/60 Superdex 75 gel filtration column (GE Healthcare).

Point mutations were introduced using the QuikChange methodology (Stratagene), and mutant proteins were purified as described for the corresponding wild-type (WT) protein. All protein concentrations were determined with reference to bovine albumin standards using Bradford assays.

cI repressor reporter assay

Expression plasmids for gene fusions of cI with the target protein (wild-type or W→G mutant) were transformed into the reporter JI PrGFP *E.coli* strain, along with the cI wild-type (constitutive dimer) as a control. Colonies were picked in triplicate, inoculated in 1 mL 2YT Kan/Cam media and grown to saturation at 37˚C. The following day, 1 µL of saturated culture from each sample was seeded into 999 µL of 2YT Kan/Cam media with or without 1 mM indole (final concentration). Cultures were grown to saturation at 37˚C, and 200 µL of each sample was transferred to a Costar 96-well flat black bottom plate for TECAN analysis. Final growth and fluorescence (RFU) time-point reads were taken at OD_{600nm} and 481nm, respectively. All RFU data were normalized for cell density. A summary of the three protein targets is presented in **Table 2.1**.

Table 2.1: A summary of homodimeric *E. coli* proteins included in the reporter gene assay.

Gene	Activity of protein product	PDB entry	Number of Trp
yeaZ	unknown function (hypothetical protease)	1OKJ (*89*)	7
orn	exoribonuclease	2IGI (*90*)	4
tadA	tRNA adenosine deaminase	1Z3A (*91*)	3

Calculation of distance from mutation site to dimer interface

For each dimeric protein target, the distance from the W→G mutation site to the dimer interface was defined to be the Euclidean distance between the Cα atom at the mutation site and the closest Cα atom on the other chain. Distances were calculated using PyRosetta (*92*).

Rosetta refinement protocol and estimated stability differences

Estimates of protein stability differences were computed out using the Rosetta macromolecular modeling suite (*77*). All calculations were carried out using svn revision 54048 of the developer trunk source code. Rosetta is freely available for academic use (www.rosettacommons.org).

The Rosetta command line used to carry out refinement simulations is as follows:

```
relax.linuxgccrelease -s input_pdb –relax:fast –in:file:fullatom
```

We generated the starting structure of each W→G mutant from the cI repressor assay by manually modifying of the wild-type PDB file: removing the sidechain atoms from tryptophan and changing the amino acid identity to glycine. During the simulation, this refinement protocol entails optimization of both backbone and sidechain degrees of freedom in a Monte Carlo search.

We performed 1000 independent simulations for each mutant and computed the energy for each output structure. Energies for a given protein construct—whether the dimer interface energy or the total energy of the (dimeric) protein—were taken to be the average over the 100 lowest-energy output structures. From each of these averages we subtracted the average value from analogous simulations of the corresponding wild-type protein.

GFP fluorescence assays

Fluorescence studies were carried out using protein concentration of 9.5 μg/ml in 20mM Tris-HCl (pH 8.0), 20 mM NaCl, 5% DMSO (with or without 1 mM indole).

GFP pulse proteolysis

Subtilisin (P5380) was acquired from Sigma-Aldrich. Proteolysis experiments were carried out in 20mM Tris-HCl (pH 8.0), 20 mM NaCl, using 0.9 mg/ml of +5 GFP (wild-type, W57G, or W57A) and increasing concentrations of subtilisin. The protease inhibitor control experiment included 5 mM PMSF (phenylmethanesulfonylfluoride). Reactions were incubated at 37˚C for 1 hour.

Band intensities were quantified using ImageJ (http://imagej.nih.gov/ij/). The gel image was converted to grayscale and the colors were inverted. Next, a strip of bands were selected in a rectangle and the intensity of each band in the region was computed using a built-in feature of ImageJ.

GFP crystal structures

A purified sample of the +5 GFP point mutant (W57G or W57A) concentrated to 4.0 mg/mL in 100 mM NaCl, 20 mM Tris pH 8.0 was used for crystallization screening. All crystallization experiments were conducted using Compact Jr. (Emerald Biosystems) sitting drop vapor diffusion plates at 20˚C using equal volumes of protein and crystallization solution equilibrated against 75 μL of the latter.

For W57G, yellow crystals that displayed a prismatic morphology formed in one week from the IndexHT screen (Hampton Research) condition D6 (25% (v/v) PEG 3350, 0.1 M Bis-Tris 5.5). Streak seeding was conducted which resulted in the production of higher quality crystals. Samples were transferred to a fresh drop composed of 75% crystallization solution and 25% PEG 400, and stored in liquid nitrogen. X-ray diffraction data were collected at the Advanced Photon Source beamline 17-ID using a Dectris Pilatus 6M pixel array detector.

Intensities were integrated using XDS (*93*) and the Laue class analysis and data scaling was performed with Aimless (*94*), which confirmed that the highest probability Laue class was *mmm* and space group was likely $P2_12_12_1$. The Matthew's coefficient (*95*) indicated that the asymmetric unit contained four independent molecules (Vm=2.1 $Å^2$/Da, 40.7% solvent). Structure solution was conducted by molecular replacement with Phaser (*96*) using a previously determined structure of the superfolder GFP (PDB: 2B3P) as the search model. All space groups with 222 point symmetry were tested for the molecular replacement searches and the top solution was obtained for four molecules in the asymmetric unit in the space group $P2_12_12_1$. Structure refinement and manual model building were conducted with Phenix (*97*) and Coot (*98*) respectively. TLS refinement (*99*) was incorporated in the later stages to model anisotropic atomic displacement parameters. Structure validation was conducted with Molprobity (*100*).

For W57A, yellow crystals that displayed a prismatic morphology formed in 2–4 days from various conditions in the Wizard 3&4 screen (Emerald Biosystems). Crystals obtained from condition H9 (40% (v/v) isopropanol, 0.1 M imidazole / hydrochloric acid 6.5, 15% (w/v) PEG 8000) were used for data collection. Samples were transferred to a fresh drop composed of 80% crystallization solution and 20% ethylene glycol, and stored in liquid nitrogen. Initial X-ray diffraction data were collected in-house at 93K using a Rigaku RU-H3R rotating anode generator (Cu-Kα) equipped with Osmic Blue focusing mirrors and an R-Axis IV^{++} image plate detector. Higher resolution diffraction data were collected at the Advanced Photon Source beamline 17-ID using a Dectris Pilatus 6M pixel array detector.

Intensities were integrated using XDS (*93*) via the XDSAPP (*101*) interface. Indexing suggested a monoclinic *C* lattice with *a*=88.51 Å, *b*=46.43 Å, *c*=69.36 Å, β=123.4^{o}. The Laue class analysis and data scaling were performed with Aimless (*94*) which confirmed that the highest probability Laue class was 2/*m* and space group *C*2. The unit cell was transformed to the non-standard body-centered setting *I*2 with *a*=69.36 Å, *b*=46.43 Å, *c*= 76.69 Å, β=105.6^{o} using the reindexing operator (-l, k, h+l). Structure solution was conducted by molecular replacement with Phaser (*96*) via the Phenix (*102*) interface using the in-house diffraction data scaled to 1.37 Å resolution. A previously determined structure of superfolder GFP (PDB: 2B3P) served as the search model. A clear solution for a single molecule in the asymmetric unit was obtained. Structure refinement using anisotropic atomic displacement parameters and manual model building were conducted with Phenix and Coot (*98*), respectively. Structure validation was conducted with Molprobity (*100*).

Figures were prepared using the CCP4MG package (*103*). Relevant crystallographic data are provided in **Table 2.3**. The Cα RMSD for residues within 4 Å of the chromophore was computed using the following 21 residues: 44, 46, 61, 62, 63, 64, 68, 69, 94, 96, 112, 121, 145, 148, 150, 165, 167, 203, 205, 220, 222.

Table 2.3: Crystallographic data for superfolder +5 GFP W57G refined to 1.6 Å resolution, and for +5 GFP W57A refined to 1.1 Å resolution.

	W57G	**W57A**
Data Collection		
Unit-cell parameters (Å, °)	a=96.07, b=97.25, c=98.16	a=69.36, b=46.43, c=76.69, β=105.6
Space group	$P2_12_12_1$	$I2$
Resolution (Å)*	48.03-1.60 (1.63-1.60)	44.02-1.10 (1.12-1.10)
Wavelength (Å)	1.0000	0.8265
Temperature (K)	100	100
Observed reflections	803,212	313,428
Unique reflections	121,582	93,517
<I/σ(I)>*	15.3 (2.0)	12.1 (1.9)
Completeness (%)*	100 (100)	98.3 (98.4)
Multiplicity*	6.6 (6.6)	3.4 (3.2)
R_{merge} (%)*#	7.5 (95.5)	4.5 (48.9)
R_{meas} (%)*^	8.1 (104.5)	5.4 (67.6)
R_{pim} (%)*^	3.1 (40.5)	2.9 (37.2)
$CC_{1/2}$ *@	0.999 (0.704)	0.999 (0.767)
Refinement		
Resolution (Å)	43.07-1.60	24.70-1.10
Reflections (working/test)	115,373/6,107	88,532/4,689
R_{factor} / R_{free} (%)&	15.5/18.2	13.4/14.6
No. of atoms (Protein/Water)	7,214/633	1,919/257
Model Quality		
R.m.s deviations		
Bond lengths (Å)	0.009	0.008
Bond angles (°)	1.075	1.278
Average B-factor (Å^2)		
All Atoms	24.5	15.4
Protein	23.8	13.5
Water	33.1	29.0

Coordinate error (maximum likelihood) (Å)	0.16	0.10
Ramachandran Plot		
Most favored (%)	98.9	99.2
Additionally allowed (%)	1.1	0.8

* Values in parenthesis are for the highest resolution shell.

$R_{\mathrm{merge}} = \Sigma_{hkl}\Sigma_i \, |I_i(hkl) - \langle I(hkl)\rangle| \, / \, \Sigma_{hkl}\Sigma_i \, I_i(hkl)$, where $I_i(hkl)$ is the intensity measured for the *i*th reflection and $\langle I(hkl)\rangle$ is the average intensity of all reflections with indices hkl.

& $R_{\mathrm{factor}} = \Sigma_{hkl} \, ||F_{\mathrm{obs}}(hkl)| - |F_{\mathrm{calc}}(hkl)|| \, / \, \Sigma_{hkl} |F_{\mathrm{obs}}(hkl)|$; R_{free} is calculated in an identical manner using 5% of randomly selected reflections that were not included in the refinement

^ R_{meas} = redundancy-independent (multiplicity-weighted) R_{merge} (*94*).

^ R_{pim} = precision-indicating (multiplicity-weighted) R_{merge} (*104, 105*).

@ CC1/2 is the correlation coefficient of the mean intensities between two random half-sets of data (*106, 107*).

Hydrogen-deuterium exchange and mass spectrometry

Hydrogen–deuterium exchange experiments for β-gluc were carried out at a protein concentration of 0.5 µg/µL in 50 mM sodium phosphate, 100 mM sodium chloride, 10% ethanol, pH 7.4. Experiments were carried out the absence of indole, or in the presence of 5 mM indole.

All deuterium exchange experiments were carried out by a LEAP HDX-PAL robot (LEAP Technologies, Carrboro, NC). The robot was used to mix 2 µL of protein solution with 40 µL deuterium exchange buffer, followed by the addition of 40 µL quench buffer after the desired labeling time. Deuterium exchange labeling times were: 10 sec, 30 sec, 1 min, 2 min, 5 min, 20 min, 1 hour, and 12 hours. The deuterium exchange buffer was 50 mM sodium phosphate, 100 mM sodium chloride, with and without 5 mM indole in 90% D_2O, 10% ethanol, pD 7.4. Undeuterated control samples were prepared using analogous buffers containing 90% H_2O instead of D_2O, at pH 7.4. All deuterium exchange buffers were kept at ambient room temperature. The exchange quench buffer was 0.75 M hydrochloric acid, kept at 1°C. Following the labeling, quench buffer was added to sample at a 1:1 volume ratio. Quenched samples were immediately injected onto a 100 µL loop prior to pepsin digestion and peptide separation.

An isocratic pump flowed 0.1% formic acid at 0.2 mL/min through the sample loop to inject samples through an immobilized pepsin column, prepared in-house, for online proteolysis of labeled samples (*108*). Peptides were captured on a C12 trap, packed in-house. A 4.5–40.5% acetonitrile gradient over 10 minutes was used to separate peptides on a 50 mm × 1 mm Ascentis Express ES-C18 column (2.7 µm particle size, 160 Å pore size, Supelco Analytical, Bellefonte, PA). All HPLC components were kept at 0°C to reduce back-exchange. Eluted peptides were analyzed by TOF-MS (Agilent Technologies, Santa Clara, CA). HDExaminer (Sierra Analytics, Modesto, CA) was used to obtain average mass values for peptide spectra at each labeling time and for the undeuterated controls.

Peptides were assigned by mapping the masses onto the protein sequence with a mass tolerance of +/- 10 ppm. Peptides that had ambiguous assignments were not included in the analysis.

Analysis of hydrogen-deuterium exchange and mass spectrometry data

For each assigned peptide we computed the mass increase Δ*m* at each time point. To compare data from different conditions (i.e. WT versus mutant, with or without indole), we divided the difference in mass increase at each time point (ΔΔ*m*) by the number of exchangeable amide hydrogens to normalize for peptide length, and then averaged this value over all time points. This yielded a "normalized deuterium difference" for each peptide, *NDD*, that can range from -1 to 1. The normalized deuterium difference for a given residue in the protein sequence was taken to be the average *NDD* value over all peptides covering that residue.

Determining what constitutes a statistically significant difference between two states of a single peptide requires uncertainty calculations provided by replicate sample analyses. A single standard deviation of ±0.14 Da for a replicated time point (i.e. a Δ*m* measurement) with a 98% confidence interval of about ±0.5 Da has been reported in previous works (*109, 110*). The confidence interval appears to be unaffected by labeling time, peptide length, or mass difference caused by deuterium uptake (*109*). The ±0.14 Da standard deviation may be used to calculate a statistical significance threshold between multiple datasets that contain all time points, $\Sigma\Delta\Delta m$, using the Student's t-distribution. In a previous study (*109*), 5 time points were sampled to give a 98% confidence interval of ±1.1 Da. Because of single trial sampling conducted for our study, we instead adopt the uncertainty from this previous study (*109*). If uncertainty and artificial triplicate sampling are applied to this work and are assumed to be similar to those of the previous study (*109*), the significance threshold would be about ±1.7 Da for the 9 sampled time points.

Statistical analysis

All statistical analysis (Spearman rank correlation, Welch's t-test, etc.) was carried out using the R statistical computing environment (*111*).

Results

Reporter gene assay for loss of function and indole rescue

To explore the frequency at which a W→G mutation leads to loss of function, we developed a reporter gene assay to monitor the loss and rescue of protein homodimerization *in vivo*. As a starting point we used the cI repressor from λ phage, which is comprised of an N-terminal DNA-binding domain and a C-terminal dimerization domain (*112*). Upon homodimerization induced by the C-terminal domain, the N-terminal domain recognizes a pR promoter to repress downstream gene transcription (*113*). To explore homodimerization in several different proteins, we created chimeras by separately fusing each protein to the N-terminal domain of cI. We expressed each chimera in *E. coli* that harbor a GFP gene under control of a pR promoter, and monitored GFP fluorescence in these cell cultures. By coupling the target construct dimerization (“function”) to transcriptional repression, this assay provides a straightforward readout of the protein’s oligomeric state (**Figure 2.1A**).

We applied this assay to test a total of 14 W→G mutations in three separate functionally unrelated *E.coli* genes that encode homodimeric proteins with available crystal structures: yeaZ, orn, and tadA (**Table 2.1**). As controls we used the reporter gene plasmid without cI repressor to monitor GFP fluorescence in the absence of repression (high RFUs), and we used the wild-type cI repressor to estimate the expected maximal repression (low RFUs); neither is strongly indole dependent. Chimeras produced by replacing the C-terminal domain of cI with any of the three wild-type homodimeric proteins led to repression comparable to that of the intact full-length wild-type cI repressor (**Figure 2.1B, Table 2.2**).

Upon introducing W→G mutations into these genes, we found that at least half disrupted repression of the GFP gene (**Figure 2.1B, Table 2.2**). The extent of repression from these mutants varied broadly: for example, yeaZ W134G and orn W9G had fluorescence intensities 96-fold and 59-fold higher than their wild-type counterparts (**Table 2.2**). In contrast, other mutants, such as yeaZ W169G and tadA W34G, maintained repressor activity nearly equivalent to that of their wild-type counterparts. Subsequent addition of 1 mM indole to the cell cultures appeared to rescue repression in a number of cases: for example, yeaZ W123G, yeaZ W134G, yeaZ W159G, and orn W9G (**Figure 2.1B, Table 2.2**). For the mutant showing the greatest indole-induced relative difference in repression, yeaZ W123G, we further found that this enhanced repression responded smoothly to the concentration of indole (**Figure 2.2**). Though these results suggest that indole may restore dimerization in these mutants, the addition of indole did not result in complete repression of fluorescence back to the wild-type levels, most likely because higher concentrations of indole may be required for complete rescue (*26*). Furthermore, despite the unchanged the fluorescence levels of reporter plasmid alone and wild-type protein chimeras upon addition of indole (**Figure 2.1B, Table 2.2**), we also cannot fully rule out the possibility that indole may cause the observed decrease in fluorescence through some other unrelated mechanism, such as unanticipated alterations in *E. coli* metabolism.

While this experiment does not explicitly normalize for possible changes in expression levels of our chimeric repressors, the observed differences in the behavior of W→G single-point mutants within the same construct are unlikely to be attributable to altered expression levels. To further investigate how structural changes upon incorporation of a W→G mutation may lead to inactivation in this experiment, we turned to simulation studies of these protein variants.

Figure 2.1: Schematic of the cI repressor assay. Various homodimeric proteins (*pink*) are fused to the DNA-binding domain of cI repressor (*yellow*), enabling binding at the Pr promoter and repression of the GFP gene. A W→G mutation that disrupts dimerization will lead to loss of repression, and thus increased expression of GFP. If the subsequent addition of indole rescues dimerization, repression will be restored and GFP expression will decrease. ***(B)*** Effect of individual W→G mutations, and the subsequent addition of 1 mM indole, determined by GFP expression in the cI repressor assay (relative fluorescence units, *RFU*). More than half of the mutations disrupted repression of the GFP gene; repression could then be partially rescued by addition of indole in a number of cases. Notable examples exhibiting loss of repression and subsequent rescue include yeaZ W123G, yeaZ W134G, yeaZ W159G, and orn W9G.

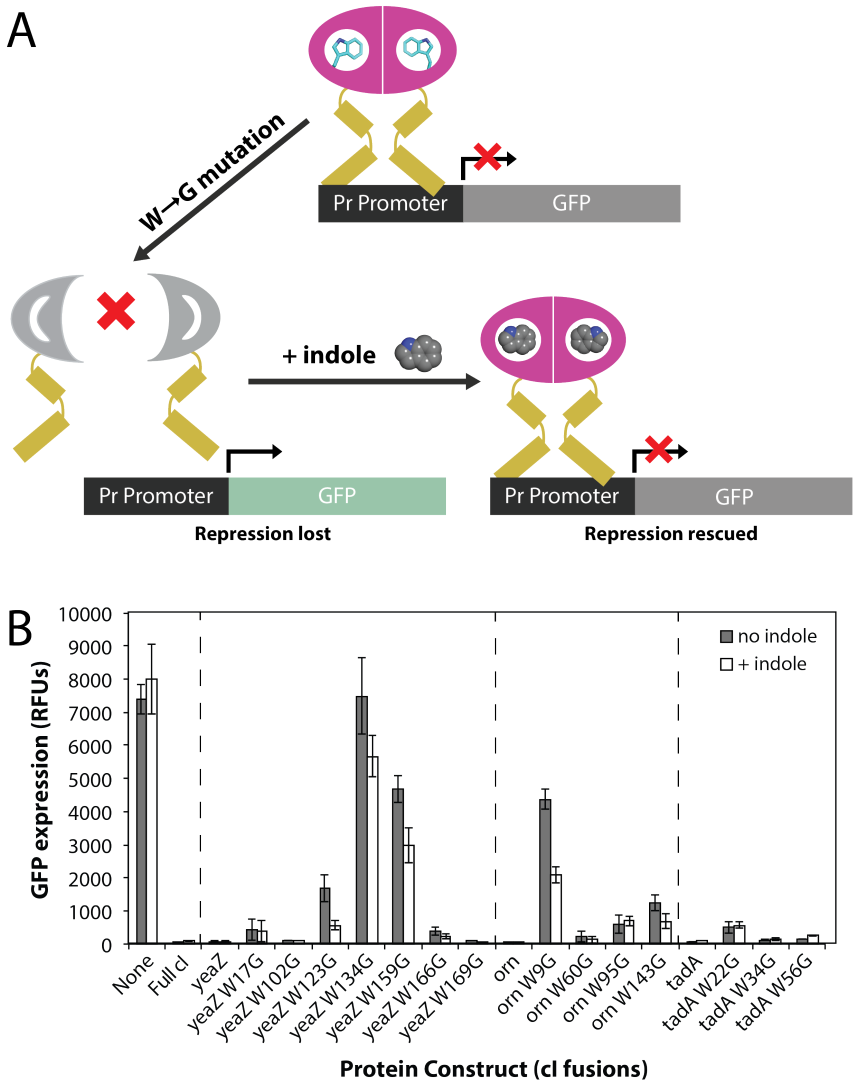

A
W→G mutation
Pr Promoter
GFP
+ indole
Pr Promoter
GFP
Repression lost
Pr Promoter
GFP
Repression rescued
B
GFP expression (RFUs)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
■ no indole
□ + indole
None
Full cI
yeaZ
yeaZ W17G
yeaZ W102G
yeaZ W123G
yeaZ W134G
yeaZ W159G
yeaZ W166G
yeaZ W169G
orn
orn W9G
orn W60G
orn W95G
orn W143G
tadA
tadA W22G
tadA W34G
tadA W56G
Protein Construct (cI fusions)

Table 2.2: Summary of data generated from the cI reporter gene assay and the computational analysis of W→G mutations in homodimeric proteins from *E. coli*.

# Results from the cI reporter assay. Values are expressed in RFUs (relative fluorescence intensity units). The reported uncertainty is the standard error of the mean from 3 experimental replicates.

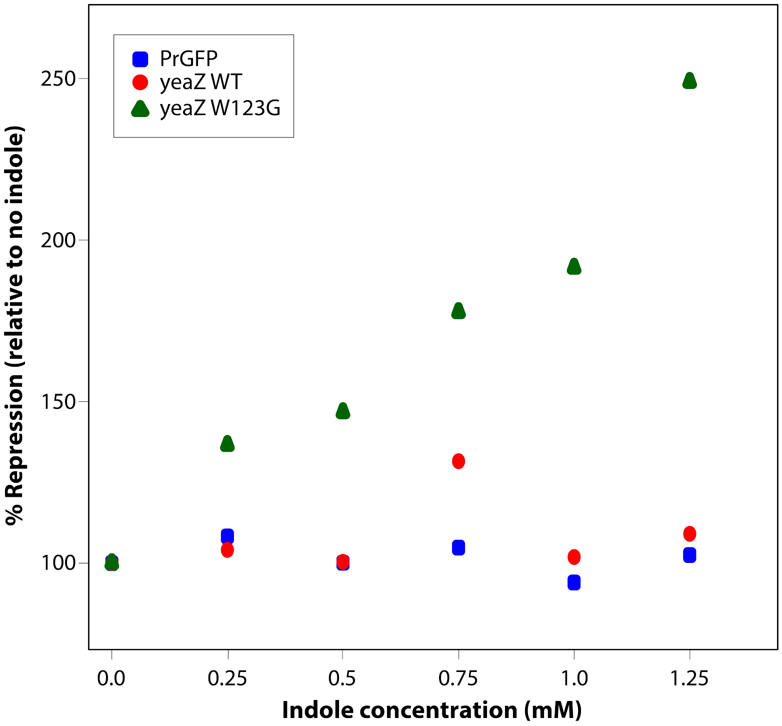
^ Computed from published crystal structures of the wild-type protein.

@ Results computed from the 100 best-scoring output structures from Rosetta refinement. Values are expressed in REUs (Rosetta energy units). The reported uncertainty is the standard error of the mean.

** mutations leading to at least 6-fold loss of repression in the cI reporter assay relative to the corresponding wild-type construct (i.e. "positive" for loss-of-function).

Protein construct	**GFP expression in cI assay, no indole #**	**GFP expression in cI assay, 1 mM indole #**	**Distance from mutation site to dimer interface (Å) ^**	**Difference dimer interface energy @**	**Estimated stability difference @**
None	7389 ± 454	7993 ± 1047	--	--	--
Full cI	69 ± 13	86 ± 9	--	--	--
yeaZ	78 ± 12	61 ± 26	--	--	--
yeaZ W17G	425 ± 311	386 ± 315	17.9	-0.4 ± 8.2	4.8 ± 3.4
yeaZ W102G	115 ± 6	102 ± 7	31.3	1.5 ± 8.0	3.1 ± 3.3
yeaZ W123G **	1689 ± 415	567 ± 143	24.0	1.9 ± 8.3	8.8 ± 3.0
yeaZ W134G **	7486 ± 1145	5671 ± 618	28.2	4.6 ± 8.0	15.3 ± 3.4
yeaZ W159G **	4678 ± 402	2980 ± 528	36.2	-0.8 ± 8.0	11.8 ± 3.1
yeaZ W166G	389 ± 138	228 ± 69	27.8	-0.1 ± 7.8	3.7 ± 3.3
yeaZ W169G	93 ± 5	55 ± 1	28.5	2.4 ± 7.5	8.9 ± 3.1
orn	74 ± 11	49 ± 8	--	--	--
orn W9G **	4361 ± 305	2089 ± 251	13.9	-0.5 ± 8.1	16.2 ± 3.0
orn W60G	219 ± 169	134 ± 105	15.8	-0.2 ± 8.0	-2.1 ± 3.1
orn W95G **	607 ± 282	696 ± 147	18.3	-0.6 ± 7.9	11.6 ± 3.0
orn W143G **	1249 ± 241	684 ± 228	7.2	5.7 ± 8.2	7.3 ± 3.2
tadA	75 ± 4	99 ± 3	--	--	--
tadA W22G **	493 ± 165	569 ± 86	22.6	-3.0 ± 5.5	11.7 ± 2.3
tadA W34G	124 ± 32	150 ± 56	17.9	-0.1 ± 5.4	2.3 ± 2.4
tadA W56G	159 ± 7	258 ± 15	15.6	-0.1 ± 5.6	8.0 ± 2.2

Figure 2.2: Dose-dependent rescue of yeaZ W123G in the cI assay using indole. All repression values for a given construct are normalized to the GFP fluorescence for that construct in the absence of indole.



Structural analysis of mutations affecting dimerization

In order to develop a structure-based approach that would allow us to identify which tryptophan sidechains would lead to loss of function when mutated to glycine, we first labeled each tryptophan sidechain as "buttressing" (with respect to the dimer interface) or "not buttressing." Sites were labeled as "buttressing" if mutation to glycine led to at least 6-fold loss of repression in the cI reporter assay; 7 of the 14 mutation sites met this criterion (**Table 2.2**). We note that each of the proteins included in the cI assay has a different fold, and that the mutation sites are dispersed across each protein (**Figure 2.3A**).

On the basis of our studies of indole rescue in β-glycosidase (*26*), we expected that protein inactivation would again result from an allosteric conformational change that coupled disruption at the mutation site to distortion at the functional site (in this case, the dimer interface). We further reasoned that such allosteric conformational changes—if not explicitly evolved or designed—would be more likely to occur locally than over long distances through the protein. As a first indirect test of this hypothesis, we therefore computed the distance of each mutation site to the dimer interface, with the expectation that the mutations closest to the dimer interface would most frequently be those producing loss of repression in the cI assay.

To evaluate the accuracy of this approach for predicting the effect of these cavity-forming mutations, we turned to receiver operating characteristic (ROC) analysis. Using the distance to the interface as our predictor, we plotted the fraction of true positives identified in our set (sites that are "buttressing" and are correctly classified as such) versus the fraction of false positives (non-buttressing sites that are incorrectly classified as “buttressing”), for increasing values of the discrimination distance threshold. Using this analysis, the curve for a perfect predictor will rise vertically to the upper left corner of the plot; in contrast, a method that makes predictions at random have a curve that approximately follows the diagonal (*red dashed line*). While mutations to either of the two tryptophan sites closest to the dimer interface indeed led to loss of repression (orn W9G and orn W143G), this approach failed to readily identify the other five buttressing sites (**Figure 2.3B**); overall, this predictor performed essentially as a random predictor.

To further explore the hypothesis that disruption at the mutation site could be coupled to distortion of the dimer interface through some distinct conformational change, we used structure prediction tools in the Rosetta macromolecular modeling suite (*77, 114, 115*) to probe the structural consequences of each mutation. We treated prediction of each mutant structure as a comparative modeling task, using the crystal structure of the wild-type dimer as a template for refinement (see *Methods*). For each of the resulting output structures, we evaluated the interaction energy between the two subunits and compared it to the corresponding energy in the wild-type structure: our hypothesis was that specific structural changes resulting from mutations at buttressing residues might lead to disruption of interactions in the protein-protein interface. However, this approach also performed essentially as a random predictor (**Figure 2.3C**), suggesting that direct consideration of interface energetics—predicated on building structural models from the wild-type template—was incapable of explaining why certain W→G mutations led to loss of repression while others did not.

We next surmised that perhaps these drastic cavity-forming mutations had destabilized the protein to the point of inducing local or global unfolding (*116-119*), in which case the crystal structure of the wild-type dimer may not prove to be suitable template structure prediction. Starting from the premise that the likelihood of a long-range allosteric conformational change in response to an arbitrary mutation is rare, we postulated that a protein could respond in three other ways to a W→G mutation: absorb the energetic cost of maintaining a cavity in the hydrophobic core of the protein, undergo local collapse of nearby structure to minimize the occupied volume in the core, or unfold. Given that the structural response to mutations that decrease sidechain volume can vary substantially depending on context (*116*), we returned to the comparative models we had previously built. Using these models, in which local reorganization may have been captured by our refinement protocol, we used Rosetta to estimate the stability difference of each mutant (dimeric) protein relative to the corresponding wild-type dimer (see *Materials and Methods*).

In stark contrast to the previous approaches, the estimated stability difference proved to be an outstanding predictor of which W→G mutations would lead to loss of repression in the cI assay (**Figure 2.3D**). We further note that the difference in average energy associated with each mutant came not from a small number of outlying conformations, but rather from a systematic shift in energy over the entire ensemble (**Figure 2.4**); while there is variation from averaging over the set of conformations, the nature of these differences thus highlights the robustness of this method for estimating stability differences.

In addition being a powerful binary classifier, the estimated stability difference also gave *quantitative* correlation with the relative fluorescence measured in the cI reporter assay, with Spearman rank correlation coefficient $\rho=0.69$, a statistically significant non-zero value ($p < 0.008$). The excellent predictive power of this approach supports the hypothesis that the loss of dimerization in the cI repressor assay was caused by loss of protein stability rather than a discrete conformational change. To test this novel mechanism for inactivation and rescue, we next turned to direct biochemical characterization.

Figure 2.3: Structural analysis of mutations affecting dimerization. ***(A)*** Distribution of W→G mutation sites over the three homodimeric proteins used in the cI assay. Mutation sites that led to loss of repression are shown in *magenta*, the other mutation sites are shown in *yellow*. The dimer subunits are colored *green* and *blue*, respectively. ***(B)*** A receiver operating characteristic (ROC) plot for predicting whether a given mutation will lead to loss of repression in the cI assay, using the distance from the mutation site to the dimer interface as the predictor. The area under the curve is 0.51, indicating that this method performs about as well as making predictions purely at random (the *red dashed line* in each ROC plot corresponds to a random predictor). ***(C)*** An analogous ROC plot generated by using the difference in interface energy of comparative models to predict whether a given mutation will lead to loss of repression in the cI assay. The area under the curve is 0.41, indicating that this method is not predictive of the data. ***(D)*** An analogous ROC plot generated by using the estimated stability difference from the same set of comparative models. The area under the curve is 0.94, indicating that this method performs much better than a random predictor; the difference from a random predictor is statistically significant ($p < 0.004$). The identification of stability difference as a successful predictor for loss of function suggests that, at least in this experiment, changes in protein stability may underlie inactivation/reactivation.

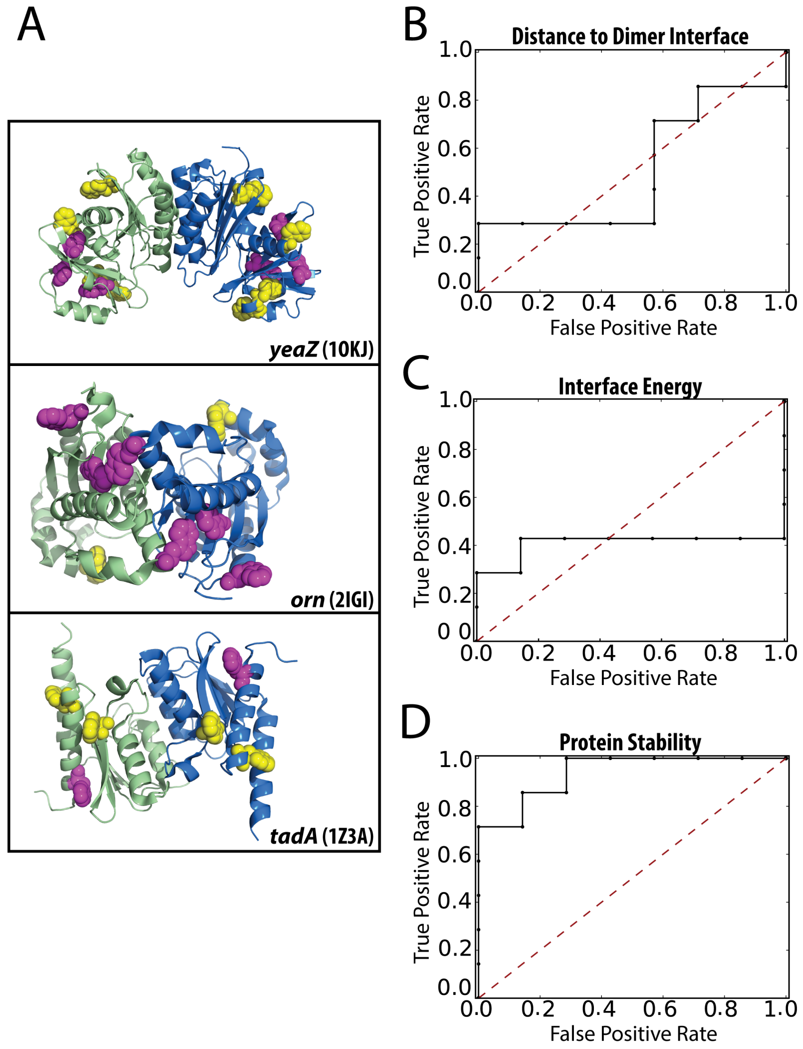
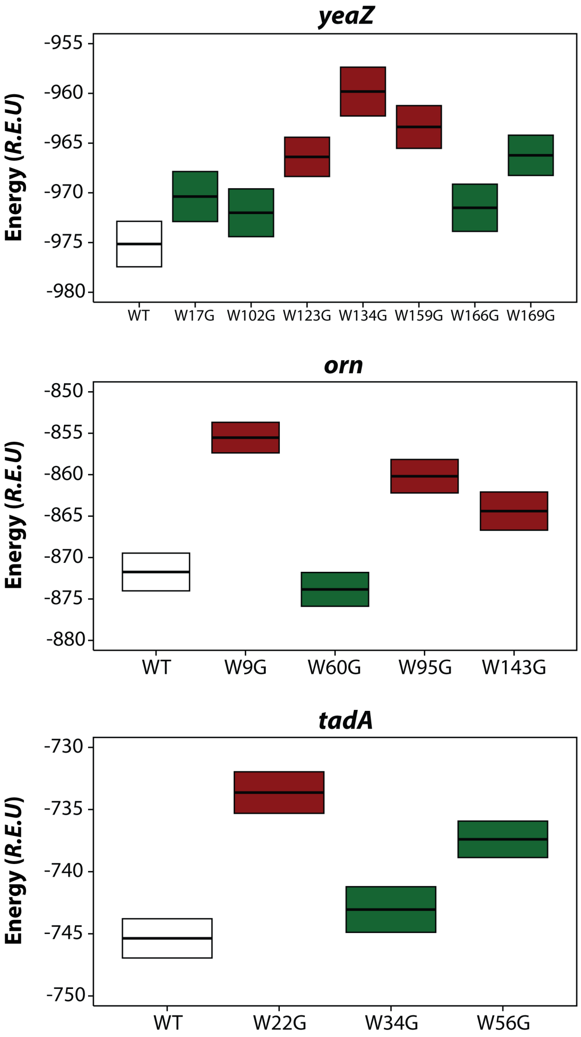

A
yeaZ (1OKJ)
orn (2IGI)
tadA (1Z3A)
B
Distance to Dimer Interface
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
C
Interface Energy
True Positive Rate
False Positive Rate
0 0
0.2
0.4
0.6
0.8
1.0
D
Protein Stability
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Figure 2.4: Energy distribution for each protein construct in the cI reporter assay, of the 100 top-scoring output structures from Rosetta refinement simulations. The mean energy is shown as a horizontal bar across the box, while the height of the box indicates two times the standard error of the mean. Boxes filled in red denote mutations that led to at least 6-fold loss of repression in the cI reporter assay relative to the corresponding wild-type construct. Boxes filled in green denote mutations that did not lead to at least 6-fold repression.



Mechanism of inactivation and rescue in +5 GFP

Due to the inherent challenges associated with the biochemical and structural characterization of homodimers, we elected to explore whether the same stability-mediated mechanism of inactivation and rescue occurred in a model system more naturally amenable to these *in vitro* techniques. We selected +5 GFP for these studies, a variant of "superfolder" GFP (*88*). Like most GFP constructs, +5 GFP folds into a β-barrel harboring a single tryptophan residue (Trp57) on the central helix, 10 Å from the chromophore (*88*). Simulations analogous to those described above gave an estimated stability difference of 4.5 Rosetta energy units associated with this W57G mutation; this value nearly, but not quite, reaches the threshold of 5.0 over which we regularly observed loss of function in the cI reporter assay (**Table 2.2**).

We measured the fluorescence intensity of wild-type +5 GFP and its W57G mutant, and found that deletion of this tryptophan sidechain reduced the fluorescence intensity by 50% (**Figure 2.5A**). While addition of 1 mM indole led to a slight decrease in fluorescence intensity for the wild-type protein, indole instead *rescued* fluorescence in the W57G mutant, back to 63% of the wild-type value (the difference in fluorescence intensity upon addition of indole to +5 GFP W57G is statistically significant, $p < 0.001$ using Welch's t-test). Rescue of W57G fluorescence by indole increases in a dose-dependent manner (**Figure 2.6**).

It is well established that slight structural rearrangements close to the GFP chromophore can lead to dramatic spectral differences (*120, 121*); the fluorescence properties can thus serve as a sensitive readout of the local environment surrounding the chromophore. We therefore carried out excitation and emission wavelength scans for both GFP constructs (**Figure 2.5B**). The shapes of the wild-type and W57G spectra are identical, notwithstanding a 46% decrease in intensity upon mutation (consistent with **Figure 2.5A**). The addition of 1 mM indole did not change either curve shape, save the same intensity differences observed previously (**Figure 2.5A**). Collectively, the lack of peak shifts or additional peaks in these spectra suggests that the partial inactivation and rescue we observed was not coupled to reorganization of the packing around the chromophore.

Based on the unchanged excitation and emission maxima, we formulated the hypothesis that in the absence of indole, +5 GFP W57G populates two states. The first, comprised of 46–50% of the population, is characterized by a conformation very similar to that of wild-type +5 GFP and accounts for the native-like excitation and emission spectra. The second state, accounting for the remaining 50–54% of the population, may be partly unfolded or have changes in conformational dynamics that disrupt the chromophore and result in loss of fluorescence.

Figure 2.5: Mechanism of inactivation and rescue in +5 GFP. ***(A)*** Fluorescence intensity of +5 GFP constructs, with excitation at 485 nm and emission at 528 nm. The indole concentration was 1 mM. Error bars indicate the standard error of the mean from 10 replicate measurements. *** statistically significant difference at $p < 0.001$. ***(B)*** Excitation and emission spectra of +5 GFP constructs. The indole concentration was 1 mM. *(C)* Crystal structure of W57G +5 GFP refined to 1.6 Å resolution (*green* and *magenta*), superposed with wild-type superfolder GFP (*gray* and *blue*). ***(D)*** Coomassie-stained SDS-PAGE gel showing products of pulse proteolysis reactions. Incubation with subtilisin led to more extensive degradation of W57G +5 GFP (*third lane*) than of wild-type +5 GFP (*fourth lane*).

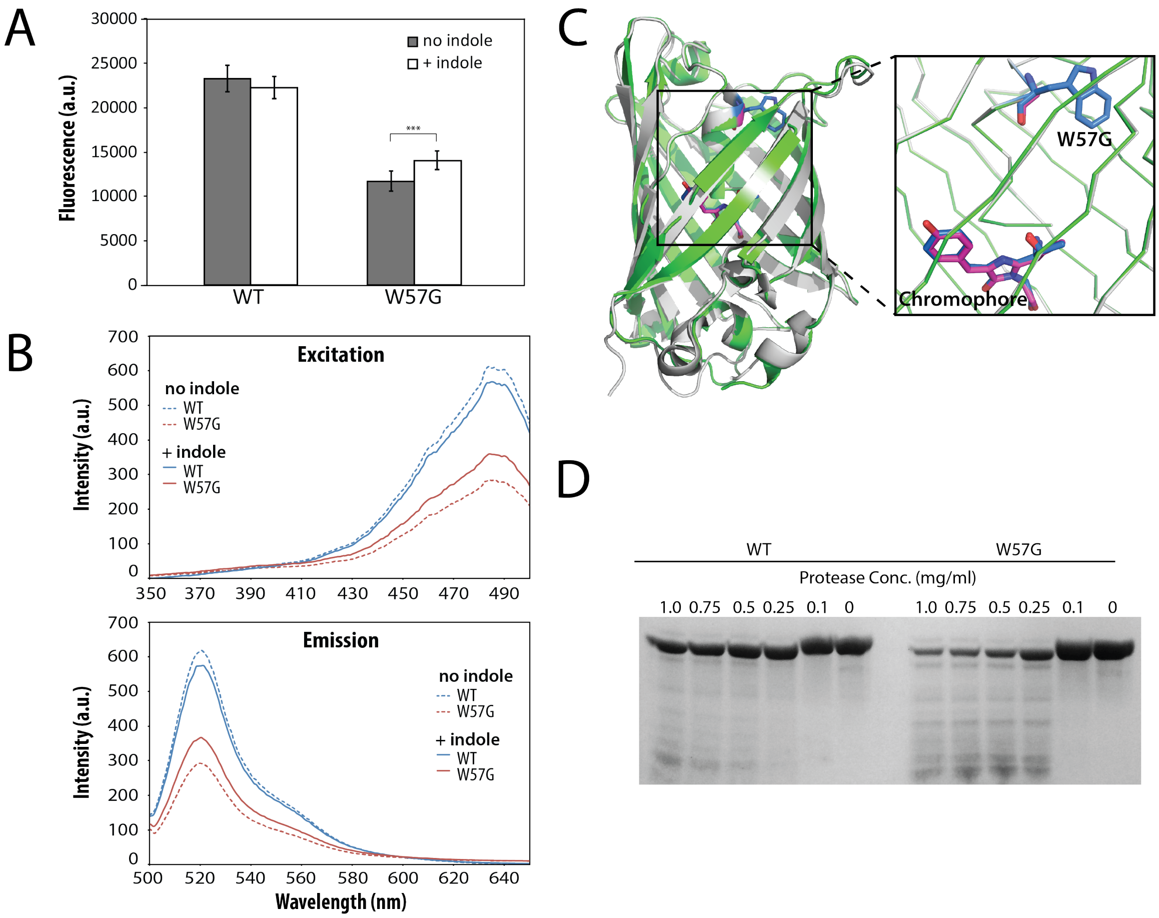
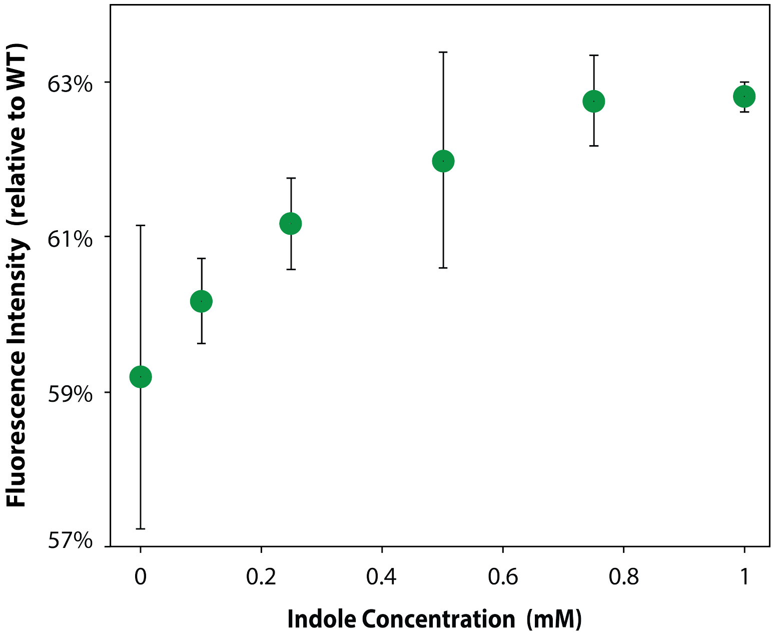

A
30000
25000
20000
15000
10000
5000
0
Fluorescence (a.u.)
no indole
+ indole
***
WT
W57G
B
Excitation
no indole
WT
W57G
+ indole
WT
W57G
Intensity (a.u.)
700
600
500
400
300
200
100
0
350
370
390
410
430
450
470
490
Emission
no indole
WT
W57G
+ indole
WT
W57G
Intensity (a.u.)
500
520
540
560
580
600
620
640
Wavelength (nm)
C
W57G
Chromophore
D
WT
W57G
Protease Conc. (mg/ml)
1.0
0.75
0.5
0.25
0.1
0
1.0
0.75
0.5
0.25
0.1
0

Figure 2.6: Dose-dependent rescue of +5 GFP W57G fluorescence using indole. Original fluorescence measurements indicated that +5 GFP W57G fluorescence was 50% that of WT, and reached 63% that of WT upon addition of 1 mM indole (**Figure 2.5A,B**). The data shown here were collected using samples of both W57G and WT produced separately. While these samples reach 63%, W57G has 59% the fluorescence of the WT prior to addition of indole. We cannot explain the origin of this difference, but speculate it may be due to endogenous cellular indole that remained in this sample through our purification.



To test this hypothesis, we solved the crystal structure of +5 GFP W57G to 1.6 Å resolution (Table 2-3). While it was somewhat surprising to obtain crystals from the heterogeneous population we anticipated, we postulate that the (non-equilibrium) process of crystallization allowed us to capture the native-like (fluorescent) state. Accordingly, our solved structure of +5 GFP W57G closely resembles the structure of wild-type +5 GFP previously determined (*88*), with overall Cα RMSD of 0.84 Å (229 residues), Cα RMSD for residues within 4 Å of the chromophore of 0.25 Å (21 residues), and no structural differences evident in response to the mutation (**Figure 2.5C**). We also found that +5 GFP W57A exhibited similar fluorescence properties as +5 GFP W57G including rescue by indole (**Figures 2.7, 2.8**), and yielded crystals that diffracted to 1.1 Å resolution. Like +5 GFP W57G, the crystal structure of +5 GFP W57A showed no structural differences relative to the wild-type structure, including the backbone at the site of the mutation (**Figures 2.9, 2.10**). Interestingly, the +5 GFP W57A structure revealed a water molecule located exactly at the position previously occupied by the indole nitrogen of Trp57, recapitulating the hydrogen bond to a nearby aspartate observed in the wild-type structure (**Figure 2.11**). While both the W57G and the W57A structures contain a large cavity previously filled by the tryptophan sidechain, this cavity is neither completely occluded from solvent nor completely hydrophobic; this makes it unsurprising that water occupies the space vacated by either mutation (*25*).

Figure 2.7: Fluorescence intensity of +5 GFP constructs, with excitation at 485 nm and emission at 528 nm. The indole concentration was 1 mM. Error bars indicate the standard error of the mean from 10 replicate measurements. The same data for WT and W57G are shown in **Figure 2.4A**.

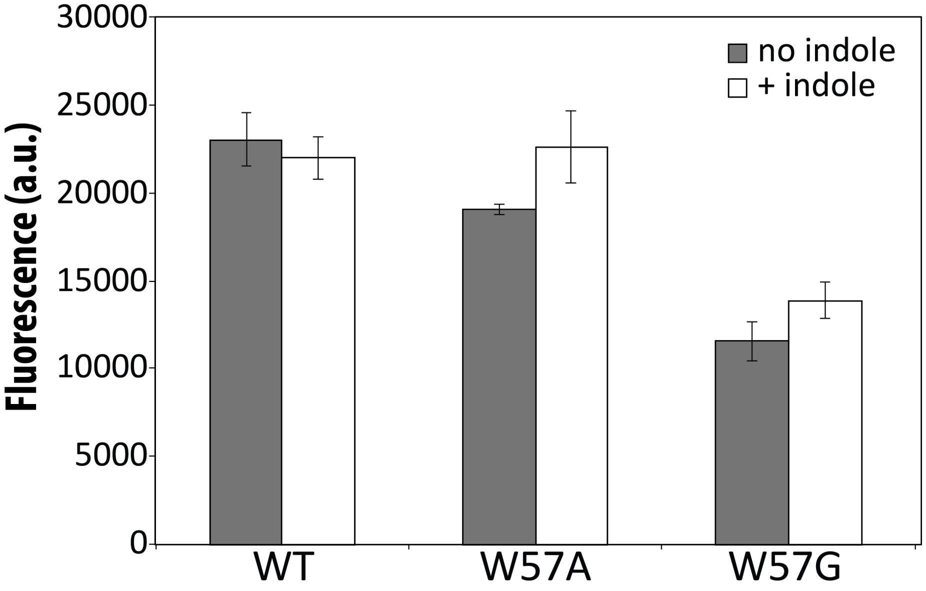


Figure 2.8: Excitation and emission spectra of +5 GFP constructs. The indole concentration was 1 mM. The same data for WT and W57G are shown in **Figure 2.4B**.

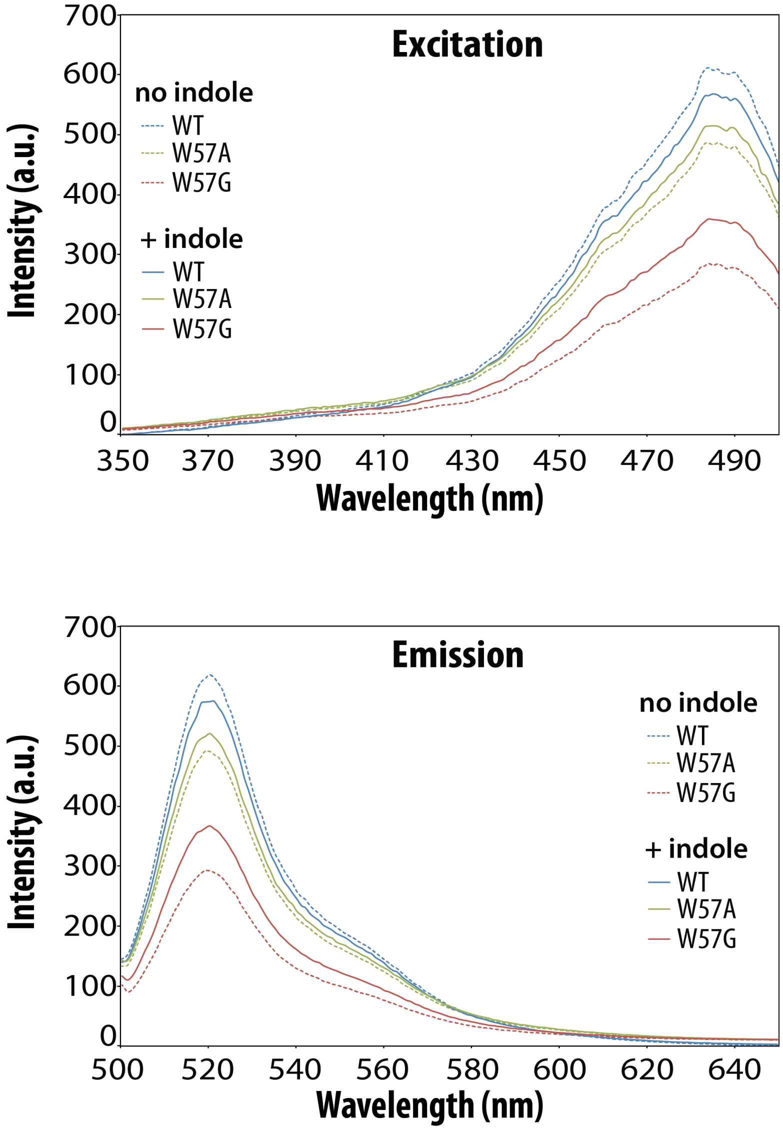


Figure 2.9: F_o-F_c omit maps contoured at 3σ showing residue 57 of +5 GFP W57G (*left*) and of +5 GFP W57A (*middle*), and the chromophore from the +5 GFP W57A structure (*right*).

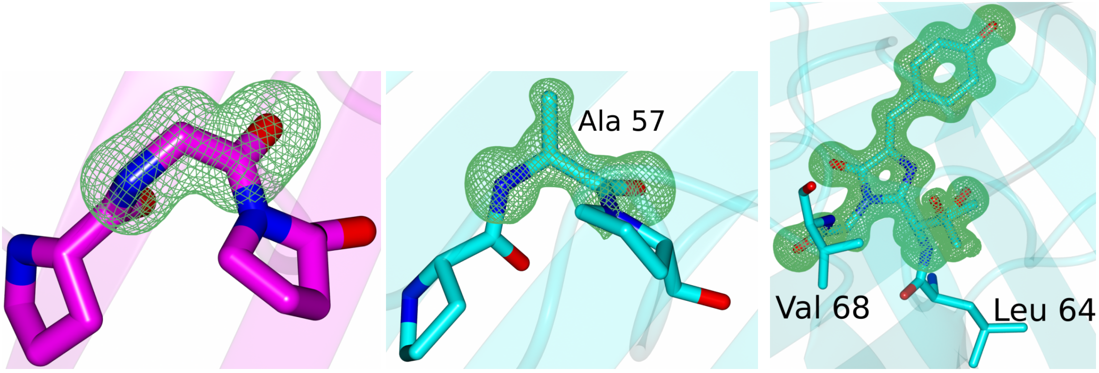


Figure 2.10: *Left:* Comparison of +5 GFP W57G and +5 GFP W57A. The overall Cα RMSD is 0.32 Å (over 225 residues). *Right:* Comparison of WT superfolder GFP, +5 GFP W57G, and +5 GFP W57A showing the similarity around residue 57 amongst all three structures. In both panels the chromophore is colored *grey*, WT superfolder GFP is colored *green*, +5 GFP W57G is colored *magenta*, and +5 GFP W57A is colored *cyan*. The structure of WT superfolder GFP is from PDB ID 2B3P (*88*).

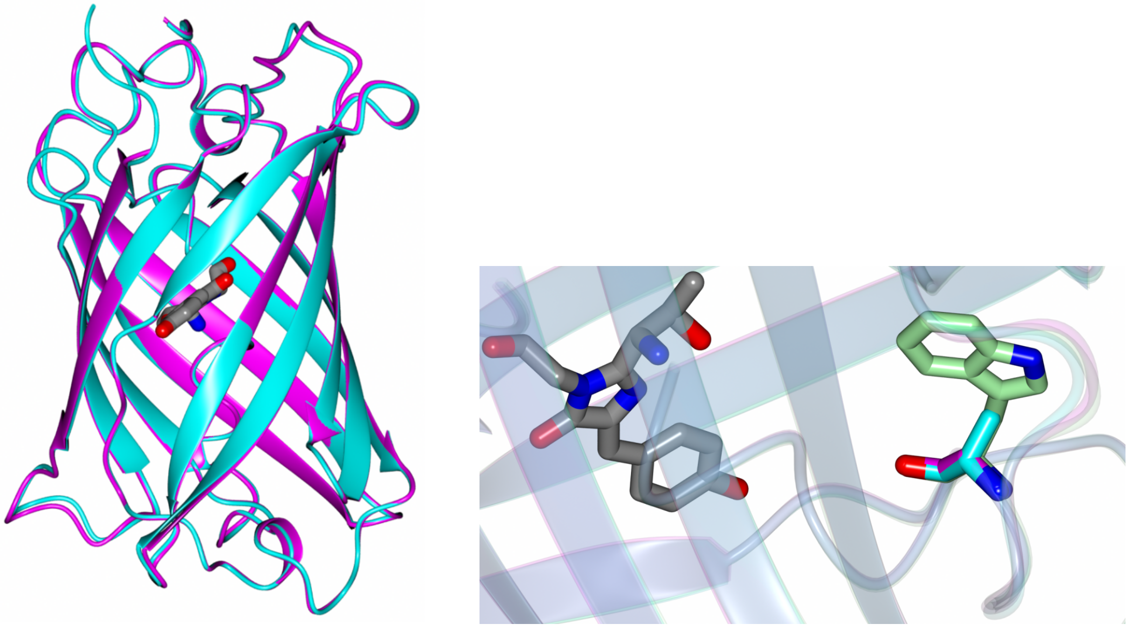
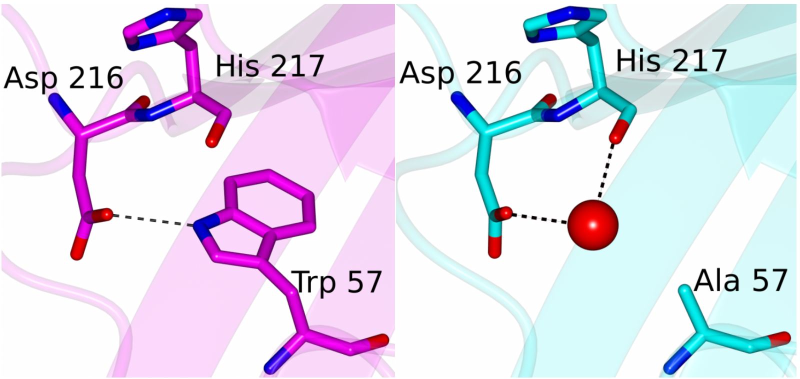
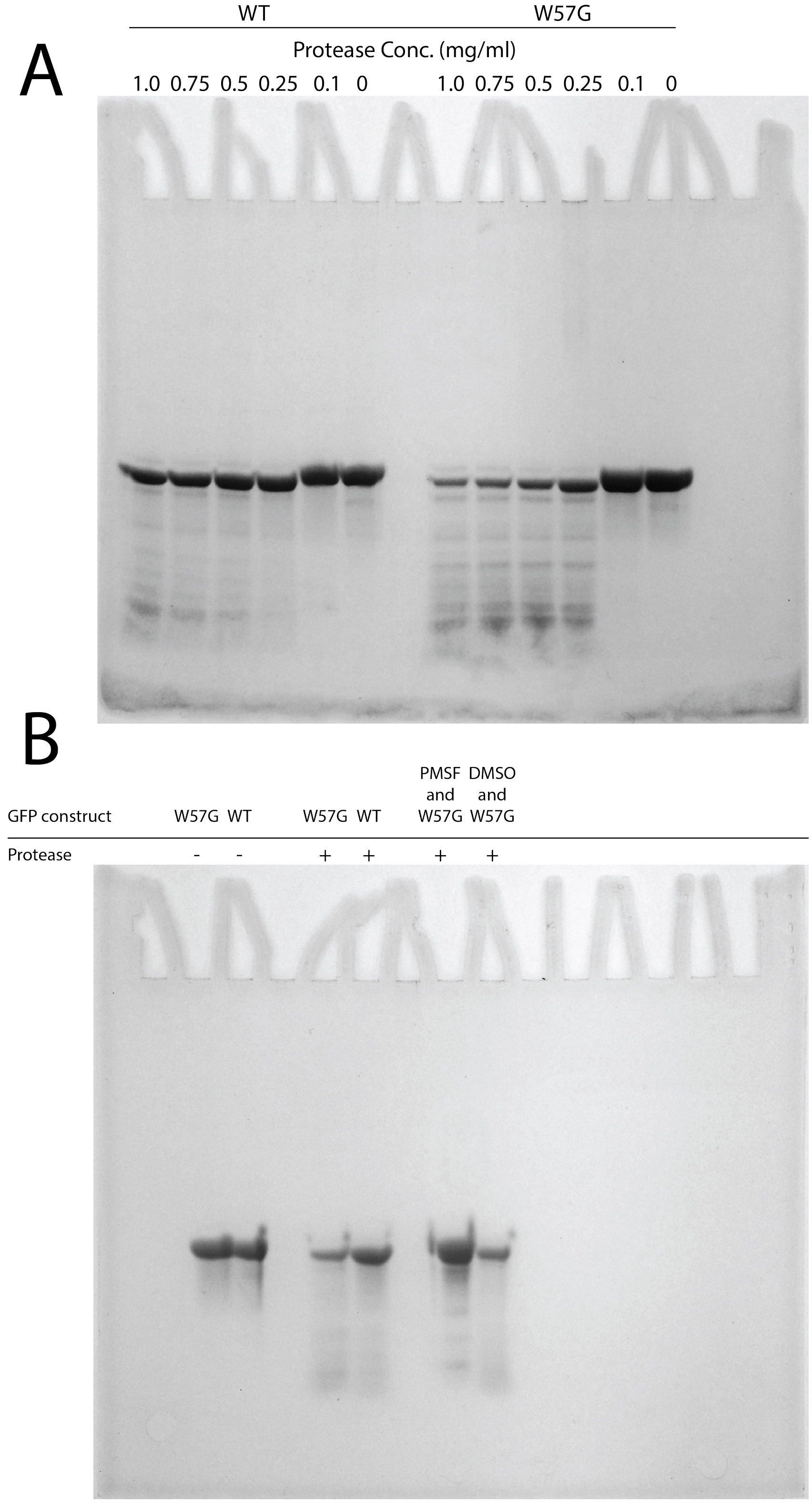


Figure 2.11: *Left:* WT superfolder GFP (PDB ID 2B3P), showing the hydrogen bond from Trp57 to Asp216. *Right:* +5 GFP W57A showing the analogous region. A water molecule (*red sphere*) adopts the position previously occupied by the indole nitrogen of Trp57, and forms a hydrogen bond with Asp216 and the backbone carbonyl of His217. The structure of WT superfolder GFP is from PDB ID 2B3P (*88*).



With this evidence that fluorescence in +5 GFP W57G derives from a species having essentially the wild-type structure, we next sought evidence for an alternate state comprising the remainder of the population. To probe for such a state we carried out a pulse proteolysis experiment, incubating either of wild-type +5 GFP or its W57G mutant with subtilisin. We found that while the folded native structure of wild-type +5 GFP renders it largely protected from proteolysis, the W57G mutant is extensively digested almost immediately (**Figure 2.5D**). We further found that inclusion of a protease inhibitor (PMSF) in the reaction prevents loss of +5 GFP W57G, while DMSO (used as a vehicle for PMSF) does not (**Figure 2.12**). The fact that PMSF prevents the disappearance of W57G +5 GFP serves to confirm that indeed proteolysis is responsible, and not some other process such as aggregation. The susceptibility of +5 GFP W57G to proteolysis supports the hypothesis that in addition to a state that strongly resembles wild-type +5 GFP, this mutant also populates a state in which subtilisin cleavage sites are more exposed than in its native-like (fluorescent) conformation. While we speculate that addition of indole would confer enhanced subtilisin resistance to W57G +5 GFP, we found through separate control experiments (not shown) that indole itself inhibits this protease directly; this made it impossible to test for indole rescue of W57G +5 GFP subtilisin resistance.

Figure 2.12: ***(A)*** An uncropped image of the Coomassie-stained SDS-PAGE gel showing products of proteolysis reactions from the +5 GFP constructs. ***(B)*** An uncropped image of a Coomassie-stained SDS-PAGE gel showing addition of the protease inhibitor PMSF prevented W57G +5 GFP degradation, while DMSO vehicle alone did not.


A
WT
W57G
Protease Conc. (mg/ml)
1.0 0.75 0.5 0.25 0.1 0
1.0 0.75 0.5 0.25 0.1 0
B
GFP construct
W57G WT
W57G WT
PMSF and W57G
DMSO and W57G
Protease
- -
+ +
+ +

Collectively, these observations point to a model in which incorporation of the W57G mutation into +5 GFP induces unfolding or enhanced fluctuations in a subset of the population (loss of fluorescence intensity), followed by a shift in this population back to the native-like state upon addition of indole (rescue of fluorescence intensity). This model is qualitatively distinct from the mechanism of inactivation and rescue we observed in our characterization of β-gly W33G (*26*).

Mechanism of inactivation and rescue in β-glucuronidase

Motivated by this stability-mediated model for inactivation and rescue of +5 GFP W57G, we returned to the *E. coli* β-glucuronidase (β-gluc) W492G mutant described previously (*26*). We had characterized this enzyme only in passing as part of our initial studies of indole rescue, showing that indole could be used to partially restore activity to this mutant in a dose-dependent manner. Though the structure of the wild-type enzyme has been solved via X-ray crystallography (*122*), we found that the W492G mutant was not amenable to crystallization. We further applied the Rosetta refinement tools (*77, 115*) to build comparative models of the W492G mutant, with the structure of the wild-type enzyme as a template; none of these models, however, included a conformational change linking the mutation site to the active site. In the absence of any structural insights we were, at the time, unable to explain the basis for the loss of enzyme activity due to this mutation (*26*), particularly given that the mutation site lies 13 Å from the enzyme active site.

In light of the studies we reported above, we formulated the hypothesis that the indole-dependent activity of β-gluc W492G may also be modulated by enhanced fluctuations or local/global unfolding, which are then reverted upon addition of indole.

This hypothesis could explain our inability to form crystals of the W492G mutant, and also our inability to build a compelling model of the structure of this protein. This hypothesis was further supported by the stability difference of 6.5 Rosetta energy units estimated for this W57G mutation, above the threshold of 5.0 that proved predictive in the cI reporter assay (**Table 2.2**).

To directly test this hypothesis, we used hydrogen/deuterium (H/D) exchange experiments to probe local fluctuation events in the protein. Upon incubation with deuterium-containing solvent, amides that are not strongly hydrogen bonded are more rapidly isotopically labeled than amides involved in intramolecular hydrogen bonds (*123-126*). Consequently, hydrogen–deuterium exchange allows us to localize conformational differences between β-gluc variants or upon addition of indole. The large size of this enzyme precluded straightforward residue-level localization of deuterium exchange information via NMR. For this reason, we instead quenched the exchange reaction, used pepsin to digest the protein, and then quantified the extent of exchange for each peptide fragment via mass spectrometry (*127*). A total of 147 peptides, of average length 13 residues, collectively covered 82% of the whole protein sequence excluding proline residues (**Figure 2.13**); this included good coverage near the active site and the mutation site **(Figure 2.14A**), and extensive overlap in many regions. We separately incubated wild-type β-gluc and the W492G variant in deuterium-containing buffer, both in the absence of indole and in the presence of 5 mM indole. Aliquots at multiple time points were digested and analyzed by mass spectrometry to determine the degree to which the protein environment conferred protection from exchange at specific regions of the protein **(Figure 2.15**).

Relative to the wild-type enzyme, a number of segments from the W492G variant showed enhanced deuterium uptake, corresponding to less protection by the protein environment. Upon addition of indole, many of the *same* peptides exhibited decreased deuterium uptake, suggesting that indole reverted the effect of this mutation (**Figure 2.16**). To allow direct comparison between peptides of different sizes, we calculated for each peptide the "normalized deuterium difference", *NDD*, defined as the difference in peptide mass increase per exchangeable amide hydrogen, averaged over all time points (see *Materials and Methods*). To further localize the effect of mutation and indole rescue, we then returned to the mapping of each peptide to the protein sequence. At every position in the protein sequence, we assigned the normalized deuterium difference for the residue as the average *NDD* value for all peptides covering its position in the sequence. While this does allow calculation of an *NDD* value for all residues covered by at least one peptide, we note that these *NDD* values are not truly residue-resolved, since each peptide represents information integrated from adjacent residues as well as the residue of interest. Relative to the wild-type enzyme, we observe enhanced deuterium uptake in the W492G mutant that is localized to specific regions of the protein sequence (**Figure 2.14B**). Addition of indole to the wild-type enzyme does not result in appreciable changes in deuterium uptake (**Figure 2.14C**); in contrast, addition of indole to β-gluc W492G leads to protection against deuterium uptake (**Figure 2.14D**). We further note that most of the regions in this mutant that exhibit increased protection upon addition of indole are the *same* as those that showed enhanced deuterium uptake as a result of the mutation. Upon comparing deuterium uptake between wild-type β-gluc and W492G with 5 mM indole present in both, we find that indole does not change the pattern but slightly reduces the magnitude of the differences (**Figure 2.17**). Our observation that the enhanced deuterium uptake of the mutant is not fully abrogated by addition of 5 mM indole is not surprising, given that we previously observed only partial rescue of enzyme activity at this indole concentration (*26*).

Mapping *NDD* values to the structure of the wild-type enzyme reveals a cohesive picture of inactivation and rescue. First, introduction of the W492G mutation leads to less protection from deuterium uptake in a nearby region that includes two helices and several intervening loops, indicating that loss of function in response to this cavity-forming mutation occurs due to enhanced fluctuations or local unfolding (**Figure 2.14E**). Addition of indole then restores protection from deuterium uptake at the same regions (**Figure 2.14F**), suggesting rigidification or refolding of these regions around the indole. These changes induced by addition of indole (partially) shift the conformational ensemble back towards that of the wild-type enzyme, thus providing a structural explanation for the previously-unexplained (partial) rescue of enzyme activity (*26*).

Figure 2.13: Unambiguously assigned peptic peptides of β-glucuronidase span most of the protein. Residues with an exchangeable backbone amide hydrogen that were not covered by at least one peptide are colored *red*; the remaining residues are colored *yellow*. Excluding proline residues (which lack an exchangeable backbone amide hydrogen), coverage comprises 82% of the protein sequence. The location of Trp492 is indicated using *orange spheres* and a substrate analog is shown in *blue spheres*.

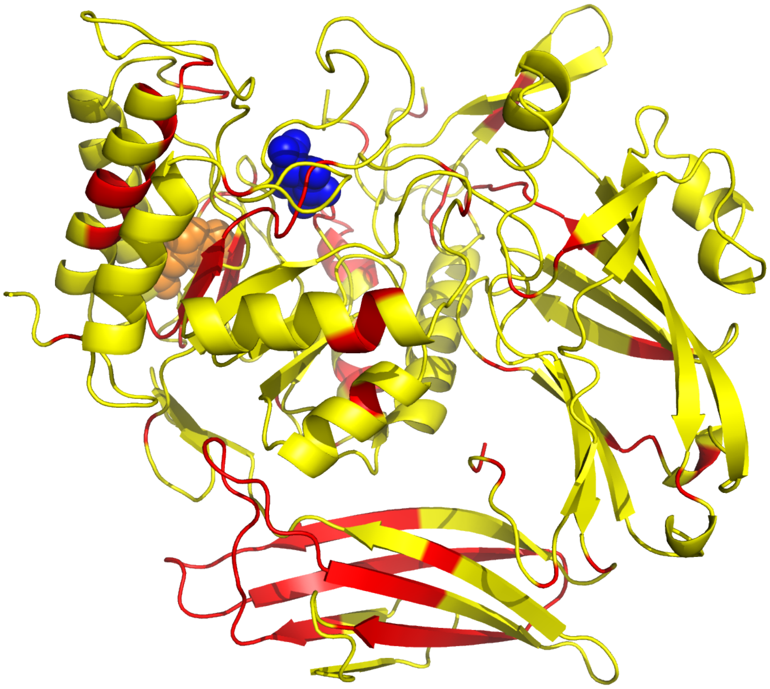


Figure 2.14: Mechanism of inactivation and rescue in β-glucuronidase as revealed by hydrogen–deuterium exchange analysis. ***(A)*** Peptic peptides provide thorough coverage of the β-gluc active site. Residues with an exchangeable backbone amide hydrogen that were not covered by at least one peptide are indicated in *red*; the remainder of the protein is shown in *yellow*. The locations of Trp492 (*orange*) and a substrate analog (*blue*) are shown in spheres. ***(B)*** Comparison of deuterium uptake ("normalized deuterium difference", *NDD*) between wild-type β-gluc and W492G; positive values indicate enhanced deuterium uptake in the mutant**.** ***(C)*** Effect of adding 5 mM indole to wild-type β-gluc. ***(D)*** Effect of adding 5 mM indole to β-gluc W492G; negative values indicate increased protection from deuterium uptake upon addition of indole.
(E) Mapping the mutant versus wild-type *NDD* to the β-gluc protein structure reveals a spatial localization of residues that undergo enhanced deuterium uptake in β-gluc W492G relative to the wild-type. The color of each residue reflects the normalized deuterium difference between the mutant and wild-type, using a gradient from *purple* (most enhanced deuterium uptake in the mutant, relative to the wild-type) to *green* (most protected in the mutant, relative to the wild-type)**.** ***(F)*** Mapping the absence versus presence of indole *NDD* to the β-gluc protein structure reveals the pattern of changes that occur upon addition of indole. Each residue is colored using a gradient from *purple* (most enhanced deuterium uptake upon addition of indole) to *green* (most protected upon addition of indole).

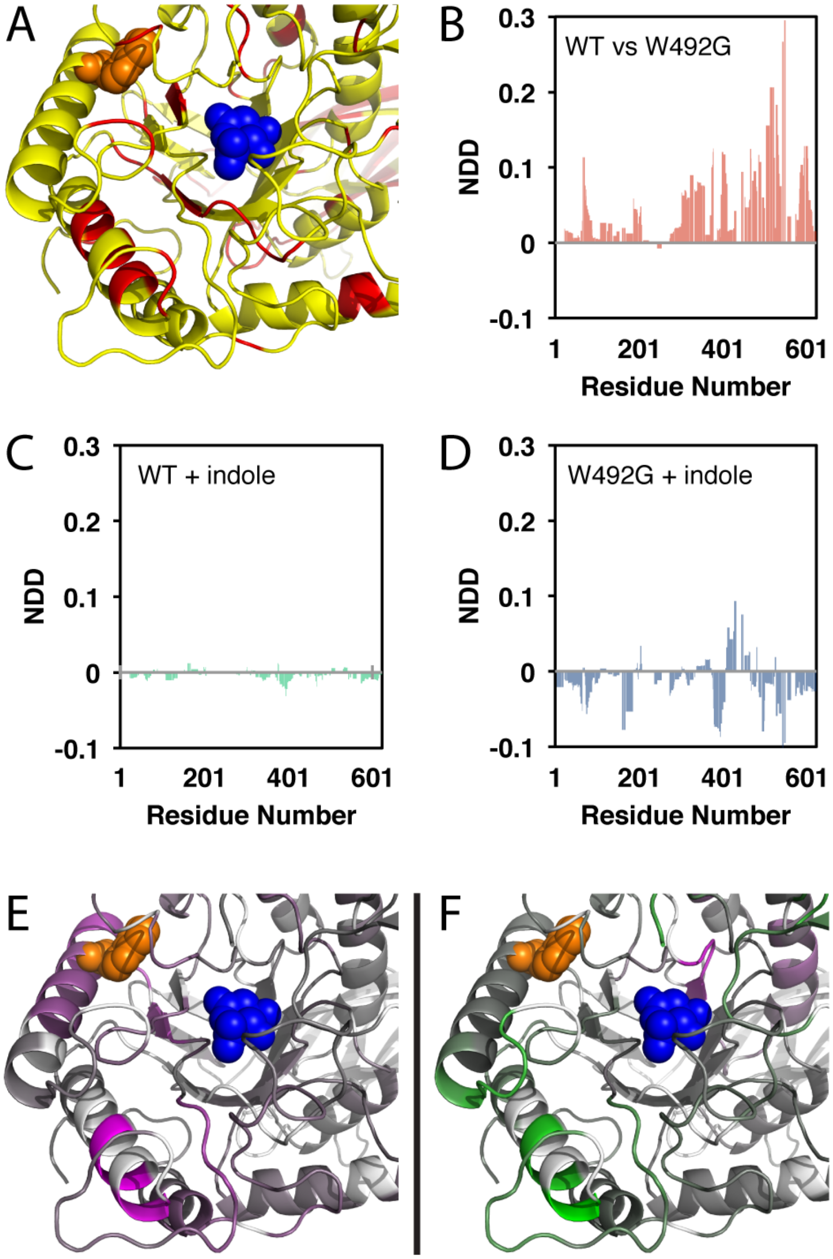

A
B
WT vs W492G
NDD
0.3
0.2
0.1
0
-0.1
1
201
401
601
Residue Number
C
WT + indole
NDD
0.3
0.2
0.1
0
-0.1
1
201
401
601
Residue Number
D
W492G + indole
NDD
0.3
0.2
0.1
0
-0.1
1
201
401
601
Residue Number
E
F

Figure 2.15: Deuterium uptake curves for four representative peptides of β-glucuronidase. Each y-axis is scaled to the maximum number of exchangeable amide hydrogens.

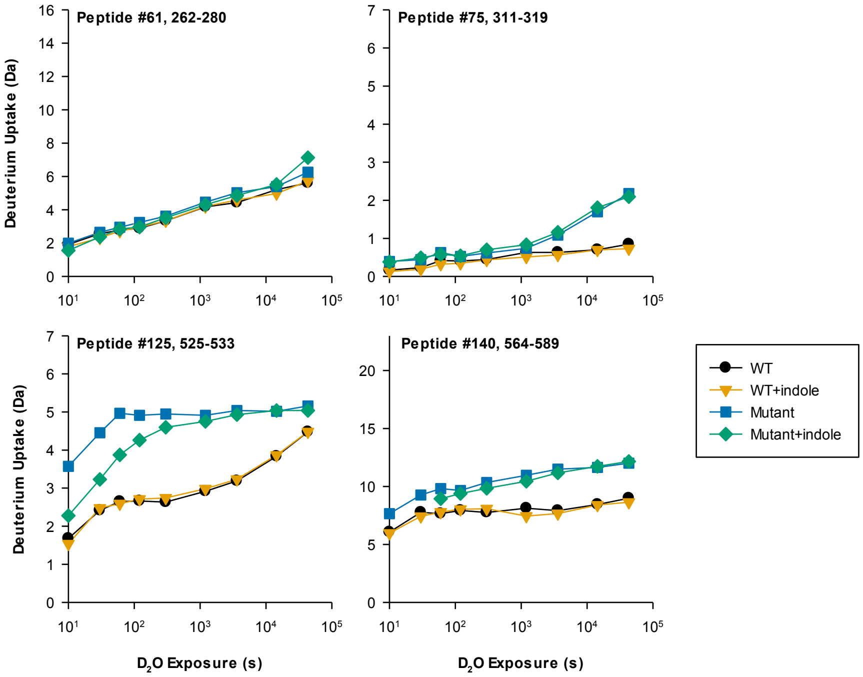


Figure 2.16: Hydrogen–deuterium exchange analysis of β-glucuronidase. The W492G mutation to β-gluc leads to enhanced deuterium uptake for many peptides relative to the corresponding peptides from the wild-type enzyme (*red*). Addition of indole to β-gluc W492G leads to decreased deuterium uptake for many of the same peptides (*blue*). The magnitude of deuteration differences is reported as the sum of the differences over all time points, ΣΔΔ*m*. Peptides are listed in order of their sequence midpoint relative to the whole protein (i.e. the x-axis does *not* refer to residue numbering directly).

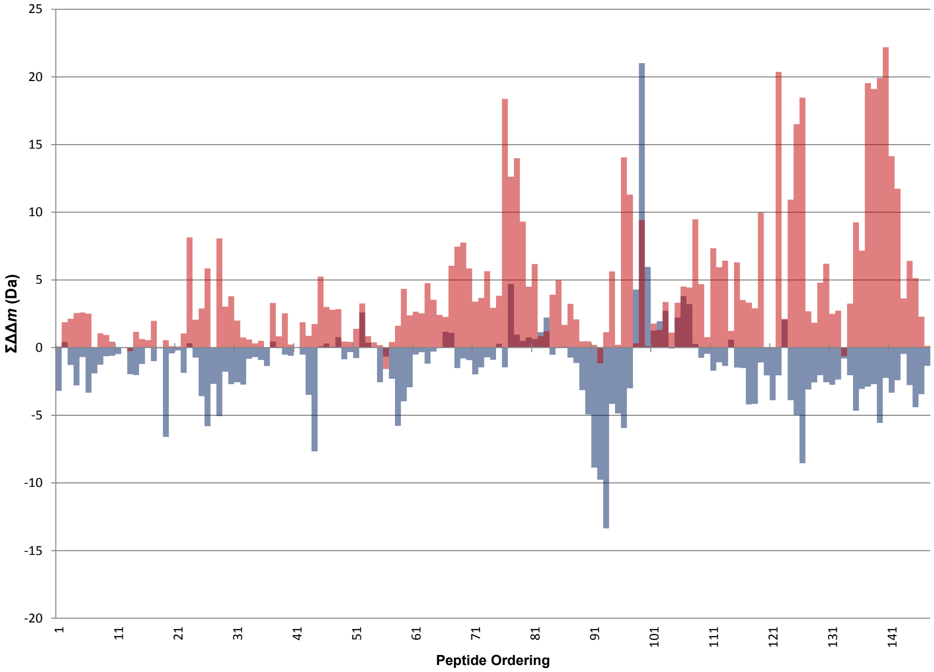
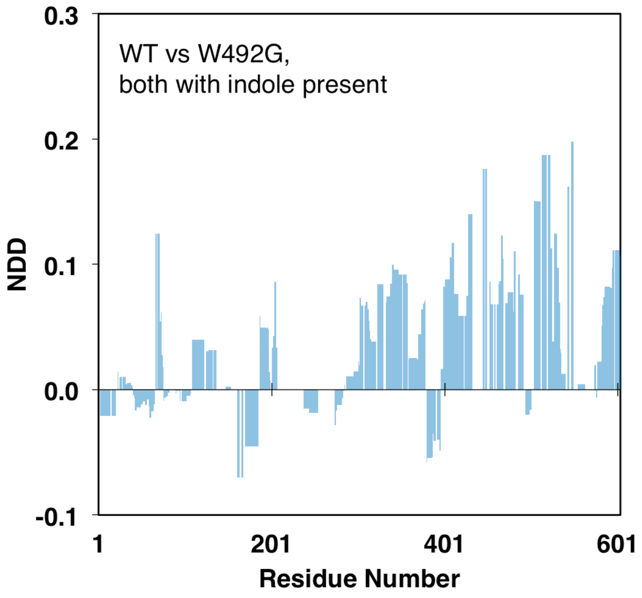


Figure 2.17: A comparison of deuterium uptake ("normalized deuterium difference", *NDD*) between wild-type β-gluc and W492G, with 5 mM indole present in both. Positive values indicate enhanced deuterium uptake in the mutant. The pattern of protected residues is similar to the analogous comparison carried out in the absence of indole (**Figure 2.14B**), but the magnitude is decreased.



Discussion

In our earlier work (*26*) we identified two examples of residues required for buttressing the nearby active site: removal (via cavity-forming mutation) of a sidechain playing this key role in maintaining the protein architecture results in collapse of the active site geometry, and thus loss of function. Our structural studies of β-gly W33G revealed a distinct conformational change induced by the cavity-forming mutation, which fortuitously transduced this disruption to the active site. Predicting the long-range effects of structural variations in general represents a very challenging problem (*128-132*), making it exceedingly unlikely that such predictions can be routinely used to introduce analogous mutations for building allosteric control into other proteins.

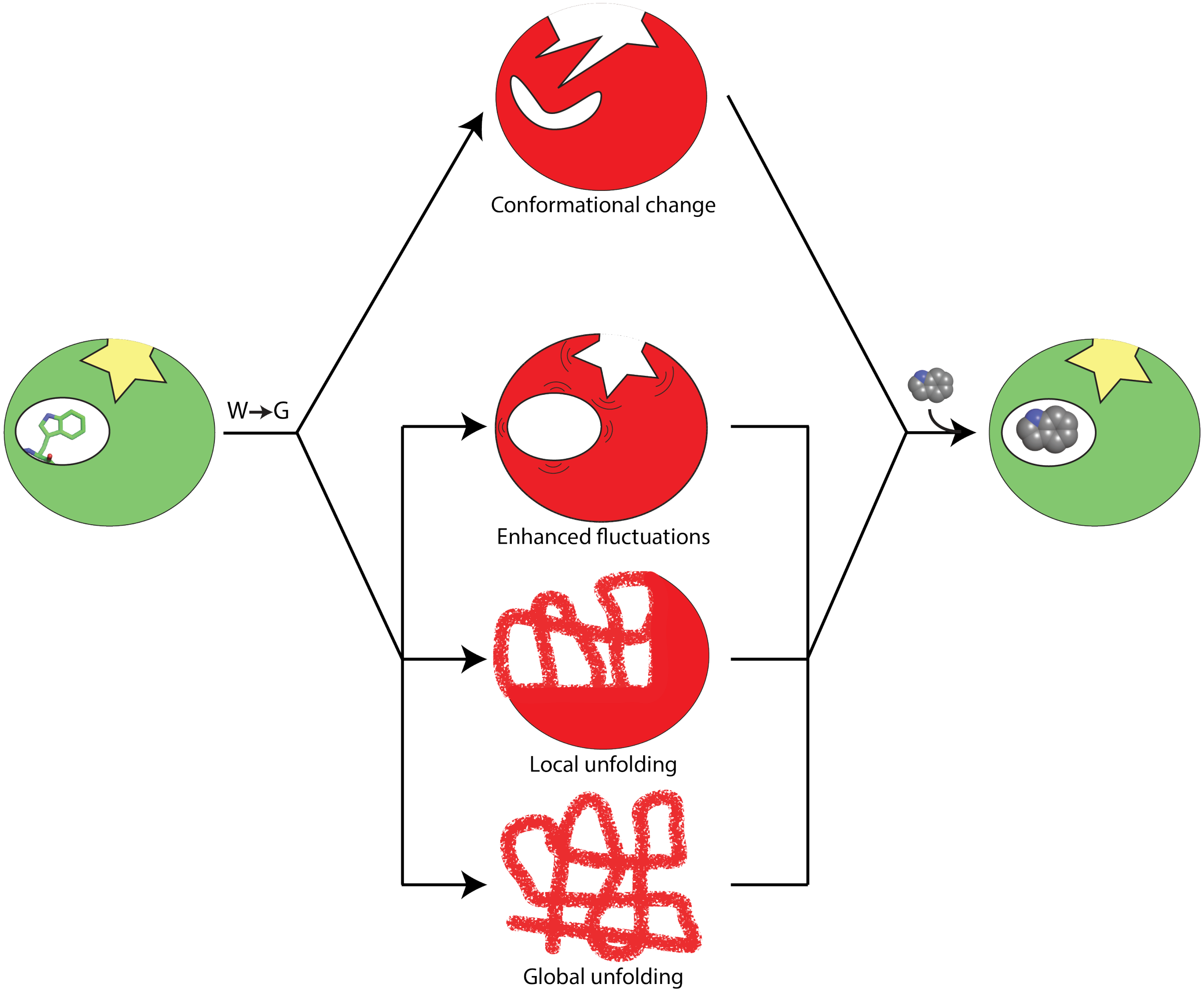
The systematic evaluation of a larger test set in our cI repressor assay (**Figure 2.1**) and the subsequent computational analysis **(Figure 2.3**), implied that protein structure and function could instead be modulated *indirectly*, through control of protein stability. In both examples for which we subsequently carried out detailed biochemical studies (**Figure 2.5, 2.14**), we found strong evidence pointing to enhanced fluctuations or unfolding resulting from destabilization as the mechanism underlying loss of function upon mutation. Accordingly, reactivation by indole may occur not only by reversion of a discrete conformational change (as in β-gly W33G), but alternatively by rigidifying or refolding the protein to its active state.

It is also noteworthy that all of the proteins characterized here derive from mesophilic organisms, whereas the β-glycosidase we studied previously derives from a hyperthermophilic organism (*Sulfolobus solfataricus*). The extreme stability of β-gly may have rendered it resistant to unfolding, allowing it to instead respond to the cavity-forming mutation via the conformational change we described earlier (*26*). In light of the results presented here, we expect that modulation of function via chemical rescue of structure will rely, for most proteins from mesophilic organisms, on stability-mediated mechanisms.

In order to build small-molecule dependence into a protein domain via chemical rescue of structure, the cavity-forming mutation must induce the protein to undergo a transition to some non-functional state (**Figure 2.18**); however, the precise details of this inactive state need not be explicitly designed. Attempting to rationally identify cavity-forming mutations to inactivate some protein of interest via a discrete conformational change would prove exceedingly challenging; on the other hand, evaluating the stability difference associated with cavity-forming mutations represents a far more tractable task. Accordingly, we expect that the insights offered here will immediately enable rational design of a variety of new protein switches that rely on activation by indole-induced protein stabilization.

Natural systems make use of small-molecules to encode a broad range of signals, whose diversity is reflected in the wide variety of mechanisms that are used to transduce ligand binding into downstream activity (*97*). These mechanisms range from discrete conformational changes (*133, 134*), to population shifts (*135-137*), to induced folding (*138-141*). There are specific design advantages associated with using each distinct mechanism: these may include intrinsic differences in dynamic range (*142*), selectivity (*143, 144*), kinetics (*141, 145*), and the ability to modulate signals by altering cellular accumulation through resistance to proteolysis (*140, 146*). Using chemical rescue of structure, we have already observed a similar range of mechanisms for recognition and activation in our set of designed protein switches. This leads to the prospect of “designing in” the desired signaling mechanism, by carefully selecting a protein of appropriate structure and stability. By analogy to natural systems, this in turn may allow us to tune the specific properties of *de novo* switches and sensors built through chemical rescue of structure, to optimally cater the designed system to the unique criteria associated with any new application.

Figure 2.18: United model of inactivation and rescue. A protein may respond to a cavity-forming W→G mutation by undergoing a discrete conformational change, as seen in our previous study (*26*), or through stability-mediated mechanisms, as described here. Addition of indole re-activates the protein irrespective of the underlying cause for loss of function.



Chapter III.
Generalizing Chemical Rescue of Structure to Neighboring Amino Acids

Abstract

The ability of engineering switch-like regulation of protein function by small-molecule signal is a long-standing goal in basic biological research and biotechnology. Our laboratory has recently developed an approach, termed “chemical rescue of structure”, which enables the modulation of protein function by disrupting and restoring protein structure. The existing designs of protein switches through this approach require a single tryptophan-to-glycine mutation, and the subsequent rescue by indole. However, these requirements restrain the selection of applicable protein scaffolds, and result in relatively insensitive protein switches due to the limited binding affinity provided by indole.

I wished to test whether the chemical rescue of structure approach might be generalized to utilize multiple neighboring mutations, such that small molecules mimicking the three-dimensional geometry of the removed moieties could be used to restore the protein structure. This generalization would resolve the limitation from indole-based systems, extending the applicability of chemical rescue of structure to more protein scaffolds. In addition, matching to multiple removed sidechains should enable the identification of more diverse and hydrophobic small molecules, which may increase the binding affinity and membrane permeability when used in cell-based assays. As a first test of the generalized chemical rescue of structure, I applied it to ChxR, a homodimeric two-component response regulator in *Chlamydia trachomatis*. Using the deleted chemical groups as a template for ligand-based screening, I successfully identified small molecules

that exhibited specific binding to the engineered ChxR variant. These preliminary results support the feasibility of generalized chemical rescue of structure approach, and can potentially enable the design of diverse protein switches and sensor proteins targeting various small molecules, especially the ones that lack the natural binding proteins.

Introduction

The switch-like modulation of protein function via small-molecule signals is a common theme in natural-occurring signaling systems. The engineering of these systems has also helped elucidate the mechanism of key processes. We recently described an approach for engineering such switch-like control into proteins via a strategy termed "chemical rescue of structure" (*26, 27*). This strategy entails introducing one or more cavity-forming mutation at locations critical for the local or global structural integrity. Deletion of these critical sidechains induces the protein to undergo diverse structural changes, and lead to loss of protein function. The subsequent addition of exogenous compounds that resemble the deleted moieties is then expected to complement the deleted structural elements, and thus restore protein structure and function.

Previous protein switches designed through chemical rescue of structure solely rely on using a single tryptophan to glycine (W→G) mutation to inactivate the enzymatic activity. Exogenous indole, which perfectly complements the deleted sidechain, restores the original protein conformation, and thus rescues the function. However, there are intrinsic limitations to the indole rescue paradigm: its application is restricted to protein scaffolds in which a single W→G mutation leads to loss-of-function. The selection of rescuing compounds is also confined to indole and its simple derivatives, whose small size and hydrophobic surface limit the binding affinity. In addition, the metabolic indole existing endogenously in cells may cause leaky activation of the engineered protein switches *in vivo*.

In principle, however, the chemical rescue of structure approach can be generalized to use multiple amino acids, rather than a single tryptophan. In this case, the removal of a constellation of atoms from neighboring sidechains creates a larger buried cavity and induces the loss of protein function. In the following rescuing stage, small molecules resembling the structural geometry and chemical property of the removed chemical groups can be identified through ligand-based screening and used to complement the cavity and restore the protein structure and function.

This generalization vastly increases the versatility of chemical rescue of structure in designing protein switches: it extends the application to target principally all protein hosts, including the ones lacking buried tryptophan residues. By selecting different combinations of neighboring amino acids as templates, the chemical space of candidate compounds is enlarged and contains more diverse chemical scaffolds. The matching to multiple removed sidechains also increases the size and hydrophobicity of candidate small molecules, which may provide more favorable binding affinity and better membrane permeability in cell-based assays. Furthermore, “bio-orthogonal” small molecules (those distinct from endogenous metabolites) may reduce the potential for inadvertent activation in indole-based systems. Furthermore, the generalized chemical rescue of structure approach can also be used in a converse direction to enable the construction of sensor proteins. In this case, we can select mutations to match the three-dimensional structure of a given small molecule, rather than selecting small molecules to complement the cavity space.

In this study, I applied this approach to a model protein and identified small molecules that can specifically bind to the engineered protein host. This study establishes the feasibility of a generalized chemical rescue of structure approach, and I anticipate that these findings will help the construction of novel protein switches and sensor proteins through the proposed approach.

Materials and Methods

Plasmid, site-direct mutagenesis, and protein expression

A pET28b plasmid containing the full length Chlamydia trachomatis ChxR gene was generously provided by Dr. Scott Hefty (University of Kansas). Point mutations were introduced using the QuikChange methodology (Stratagene). Primers used to introduce the F75A, W89G and F75A/W89G mutations were designed by the QuickChange Primer Design server (http://www.genomics.agilent.com/primerDesignProgram.jsp).

Recombinant wild-type and mutant ChxR were expressed from the pET28b vector in *E. coli* Rosetta 2(DE3)pLysS cells at 15ºC overnight. The cells were resuspended in Lysis Buffer (50 mM Tris, 150 mM NaCl, 5 mM imidazole pH 8.0) and sonicated for 10 minutes (Fisher Scientific Sonic Dismembrator Model 100). The cell lysates were then centrifuged at 15,000g for 30 min. Wild-type and mutant ChxR remained in the supernatant, which was purified by HPLC affinity chromatography with Ni-chelated Sepharose Fast Flow Resin (GE Healthcare), followed by a HiLoad 16/60 Superdex 75 gel filtration column (GE Healthcare). All protein concentrations were determined with reference to bovine albumin standards using Bradford assays.

ROCS screening

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

Surface Plasmon Resonance

Binding between W89G ChxR and each compound were analyzed by SPR using a Biacore 3000 optical biosensor (GE Healthcare). His-tagged ChxR was immobilized by Ni^{2+}/NTA chelation on a NTA sensor chip (GE Healthcare). An unmodified flow cell was used as reference. Wild-type ChxR was also immobilized on another flow cell as a negative control. All SPR runs were performed at 25 ºC using a flow rate of 50 μL/min in running buffer (HEPES buffered saline with 0.05% Tween-20 pH 7.4). Compounds were injected over sample and reference flow cells at a concentration of 100 μM in running buffer for 100 s, followed by a dissociation phase for 200 s. Sensorgram data were processed using BIAEvaluation (GE Healthcare Life Sciences) and the figures were prepared by Prism (GraphPad, La Jolla, CA).

EMSA

IR800-labeled Oligonucleotides containing ChxR binding sites DR2 were was generously provided by Dr. Scott Hefty (University of Kansas). Binding reactions (20 µl) contained DNA and ChxR at their respective concentrations, as listed in Results and were performed. The reactions were incubated at 25°C for 20 min and separated by electrophoresis in 5% TBE gel, 0.5× TBE buffer at 50 V for 2 h.. After native PAGE, IR800-labeled DNA fragments were visualized by using the Odyssey Infrared Imaging System (LI-COR Biosciences, Lincoln, NE).

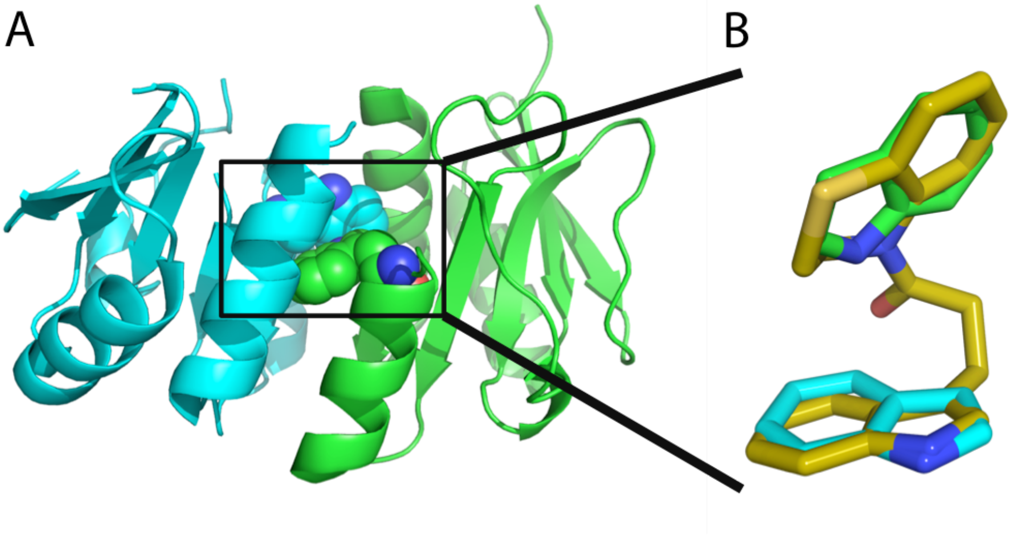
Results

ChxR: a model protein for testing the generalized chemical rescue of structure

We selected a dimeric protein, ChxR, to test the feasibility of applying chemical rescue of structure with multiple neighboring mutations. ChxR is a signal transduction response regulator of the OmpR/PhoB subfamily encoded by *Chlamydia trachomatis* (*147*). ChxR is composed of a receiver domain and an effector domain: the receiver domain drives the homodimerization of two monomeric subunits, and consequently positions the effector domain to the structural arrangement optimal for DNA binding (*148, 149*). This homodimerization is an essential structural feature that determines the correct functioning of ChxR: disruption of ChxR dimeric interface greatly reduces the binding affinity to the target DNA sequences (*147*). Visual inspection of the homodimeric interface of the receiver domain reveals the structural importance of a pair of neighboring tryptophans on opposing sides of the interface (W89 on chain A and W89 on chain B), which correspond to the same residue in the primary structure of a ChxR subunit (**Figure 3.1A**). These symmetry-related tryptophan residues are in direct contact and compactly buried in the critical position of the dimeric interface.

The critical location of the pair of tryptophan residues in the dimeric interface and the direct coupling of structural dimerization and DNA binding activity make ChxR an ideal model protein for testing the generalized chemical rescue of structure approach. A single W89G mutation on the monomeric subunit leads to the removal of two neighboring tryptophan sidechains in the native dimeric structure. These cavity-forming mutations, in combination with other background mutations if necessary, can perturb the dimeric interface and induce ChxR to undergo a transition to some nonfunctional conformation. Subsequently, we can identify small molecules complementing the deleted structural elements and investigate their ability of restoring native dimeric structure, and in turn, rescuing DNA-binding activity (**Figure 3.1B**).

Figure 3.1: Generalized chemical rescue of structure using ChxR. ***(A)*** Crystal structure of ChxR receive domain (PDB: 3Q7R) showing the symmetry-related pair of tryptophan residues (W89) at the dimeric interface (spheres representation). Monomeric subunits are colored in green and cyan, respectively. ***(B)*** Small molecules (yellow sticks) that can potentially mimic the three-dimensional structure and chemical property of the pair of W89 sidechains. Compound C2 (**Figure 3.2**) is shown here for illustrative purpose.

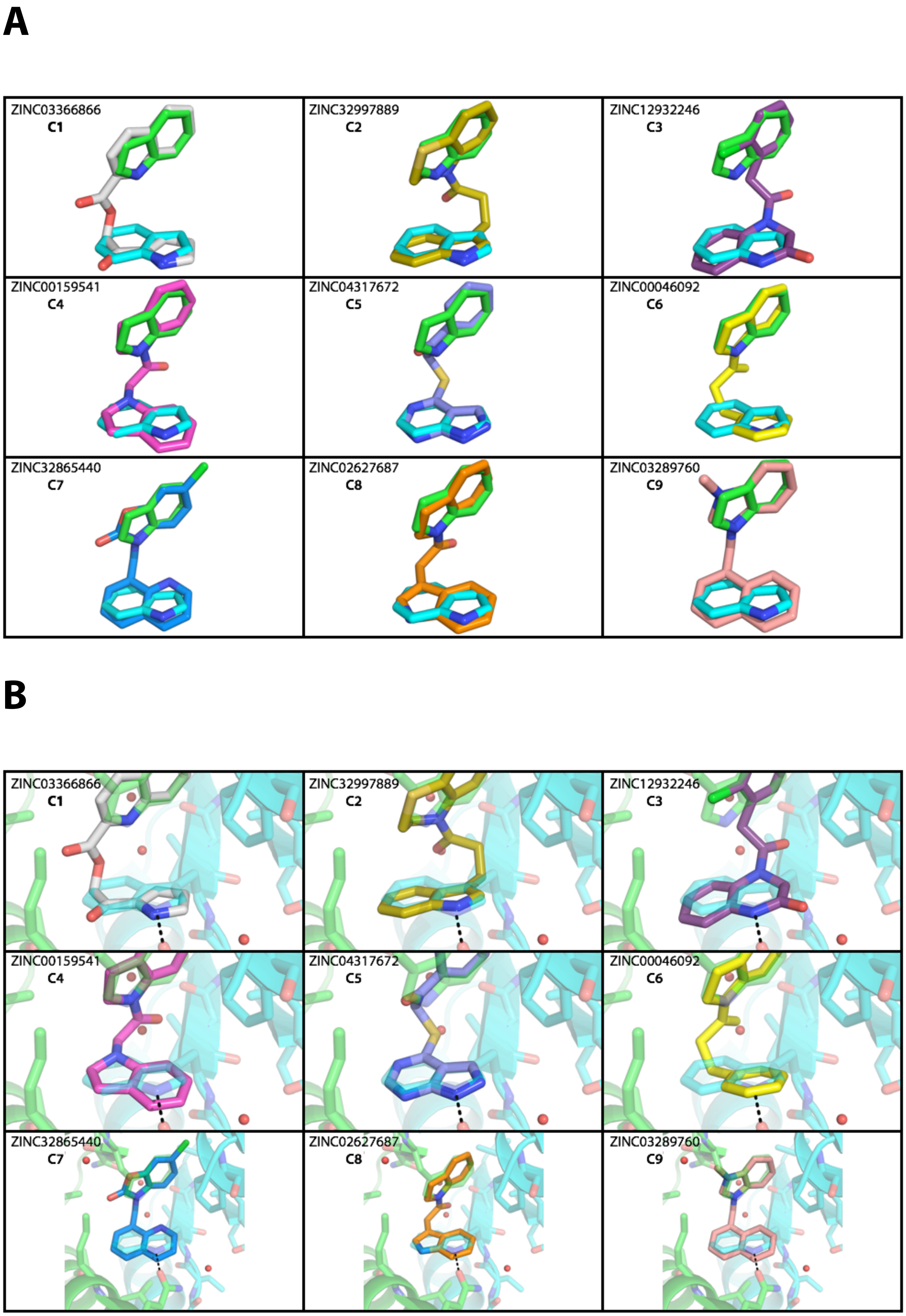


Identification and biochemical characterization of computational hits

The first step is to identify small molecules that recapitulate the precise three-dimensional structure of the deleted structural elements. For this purpose, I used the geometric arrangement of the pair of indole moieties in the native ChxR structure as a template for carrying out ligand-based virtual screening. To facilitate rapid characterization of compounds emerging from the virtual screen, I restricted the search to around seven million compounds in the ZINC database (*150*) that are both commercially available and predicted to have drug-like physicochemical properties. I used OMEGA (OpenEye Scientific Software, Santa Fe, NM) to build low-energy conformations of each compound, then ROCS (OpenEye Scientific Software, Santa Fe, NM) to align each conformation to the pair of indole moieties and quantitatively measure the three-dimensional similarity between the template and small molecule. For each of the top-scoring hits emerging from ROCS, I then used the aligned orientation to position the compound relative to the protein, and evaluated the potential binding models.

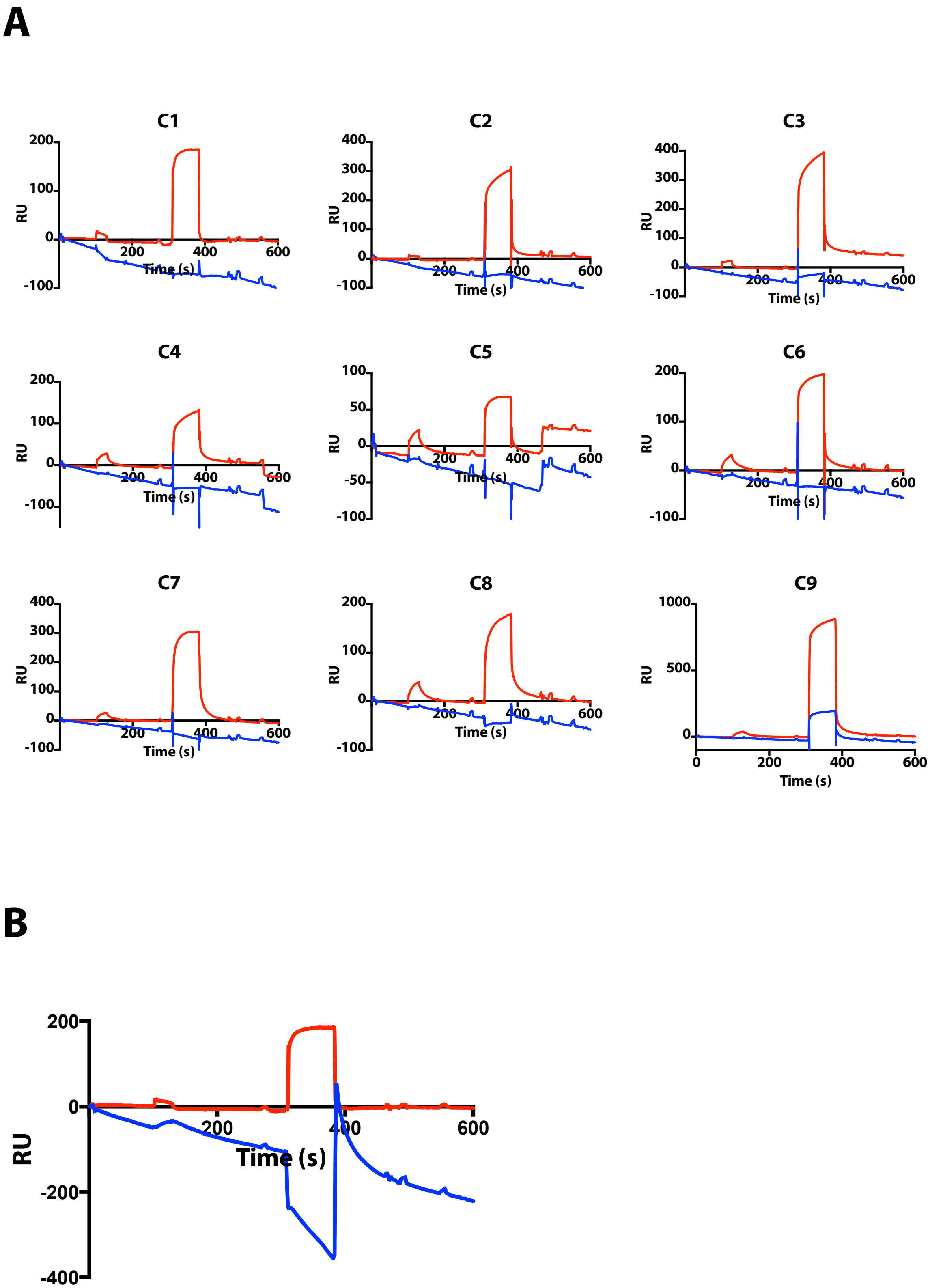
Through the virtual screening, I identified nine hit compounds (**Figure 3.2**). These compounds consist of different ring systems, connected by various linker fragments. However, they can potentially adopt conformations that resemble the three-dimensional geometry of the removed pair of indole moieties (**Figure 3.2A**). Because of this structural similarity, superposition of these hit compounds back to the removed tryptophan sidechains at the ChxR dimeric interface demonstrates that these compounds can potentially recapitulate the hydrophobic packing and, in certain examples, hydrogen bonding interaction, from the deleted tryptophan sidechains (**Figure 3.2B**).

Figure 3.2: The nine computational hit compounds. ***(A)*** The three-dimensional model of hit compounds (C1 – C9, assorted colors) superposed with the template (W89 sidechains, colored in green and cyan). ***(B)*** Potential binding models of each hit compound built from positioning the aligned structures of hit compounds back to the ChxR dimeric interface. Monomeric subunits of ChxR are colored in green and cyan, respectively. Hit compounds are colored using the same scheme as panel A.


A
ZINC03366866
C1
ZINC32997889
C2
ZINC12932246
C3
ZINC00159541
C4
ZINC04317672
C5
ZINC00046092
C6
ZINC32865440
C7
ZINC02627687
C8
ZINC03289760
C9
B
ZINC03366866
C1
ZINC32997889
C2
ZINC12932246
C3
ZINC00159541
C4
ZINC04317672
C5
ZINC00046092
C6
ZINC32865440
C7
ZINC02627687
C8
ZINC03289760
C9

As the first test, I directly examined the binding of each candidate compound to W89G ChxR mutant using surface plasmon resonance (SPR) (**Figure 3.3**). The ChxR variant was immobilized onto an SPR chip and then I sequentially passed each compound over the chip at a concentration of 100 μM. All of the nine top-scoring compounds showed a kinetic profile consistent with binding to the W89G ChxR. In contrast, except for the slight increase of signal for C9, none of the other eight compounds bind to the wild-type ChxR, suggesting that the binding of hit compounds is a direct consequence of W89G mutation (**Figure 3.3A**). However, considering the hydrophobicity of the hit compounds, another possibility is that these compounds may promiscuously bind to the exposed hydrophobic surface caused by local or global unfolding in the W89G mutant. Therefore, I also included an unrelated compound of similar size and hydrophobicity as a further negative control, This compound does not show any evidence of binding to either protein construct, suggesting that the positive signal from SPR is not a result of non-specific binding to the "sticky" hydrophobic surface of the ChxR mutant, but rather specific binding arising from the three-dimensional complementarity of the small molecule hits (**Figure 3.3B**).

Figure 3.3: Initial screening via surface plasmon resonance. ***(A)*** SPR sensorgrams testing the binding of each hit compound to either W89G (red traces) and wild-type (blue traces) ChxR. ***(B)*** SPR sensorgrams of unrelated compound (blue trace) serving as a negative control. Sensorgram of C1 (red trace) is included as an example of kinetic profile that is indicative of binding events.

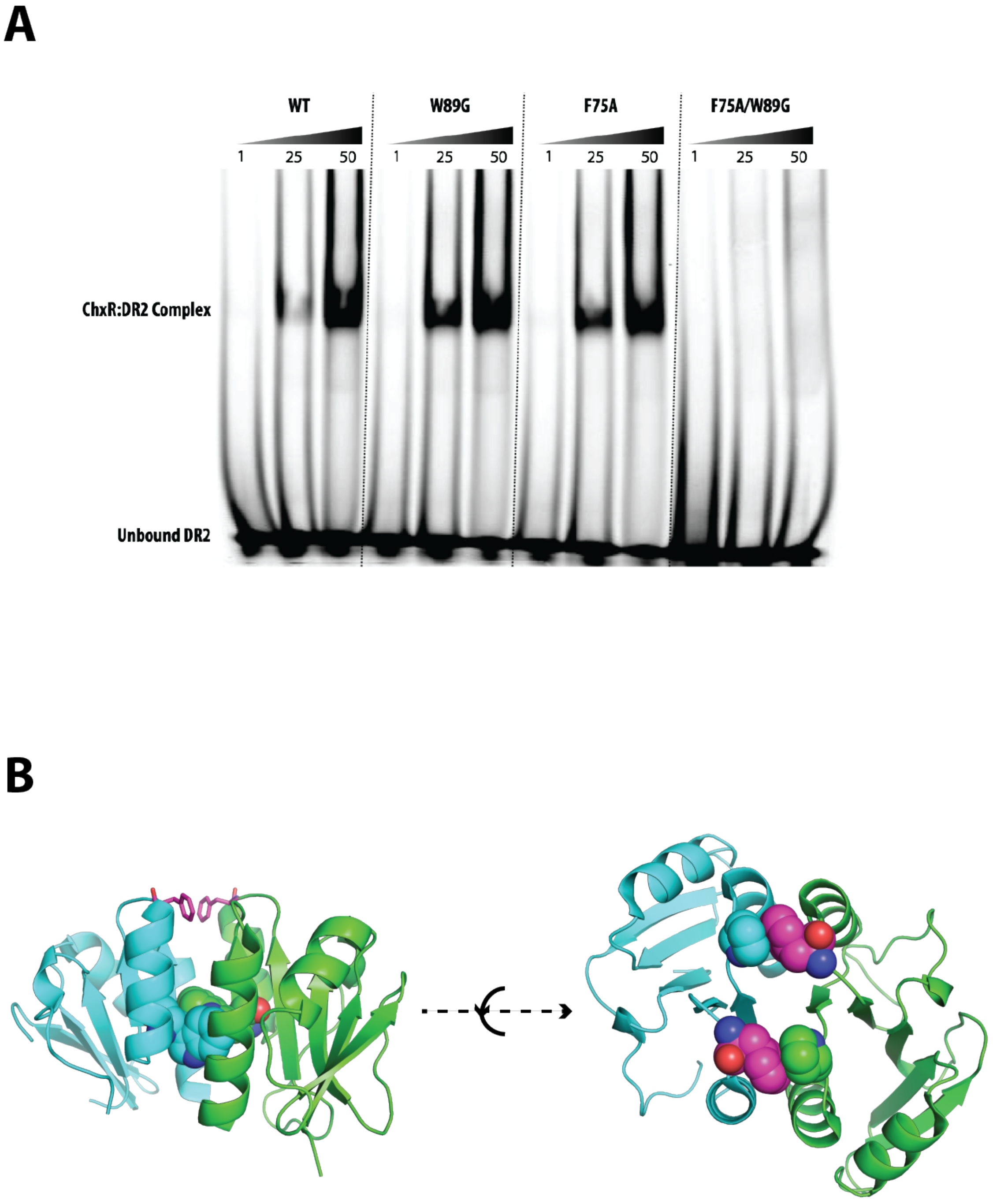

A
C1
200
100
0
-100
RU
200
400
600
Time (s)
C2
400
300
200
100
0
-100
RU
200
400
600
Time (s)
C3
400
300
200
100
0
-100
RU
200
400
600
Time (s)
C4
200
100
0
-100
RU
200
400
600
Time (s)
C5
100
50
0
-50
-100
RU
200
400
600
Time (s)
C6
200
100
0
-100
RU
200
400
600
Time (s)
C7
400
300
200
100
0
-100
RU
200
400
600
Time (s)
C8
200
100
0
-100
RU
200
400
600
Time (s)
C9
1000
500
0
RU
0
200
400
600
Time (s)
B
200
0
-200
-400
RU
200
400
600
Time (s)

Designing suitable ChxR variant for testing functional rescue

The specific binding of hit compounds to W89G ChxR led me to further investigate whether this binding could lead to structural restoration and functional rescue. A prerequisite, however, is to examine if the W89G ChxR mutant is a loss-of-function variant. To this end, I employed an electrophoretic mobility shift assay (EMSA) to directly detect the protein:DNA complex between the W89G variant and DR2 sequence, a DNA fragment recognized by ChxR with high affinity (*147*). Unfortunately, the DR2-binding profile exhibited by the W89G variant is indistinguishable from the wild-type counterpart, showing that the W89G mutation alone is insufficient to cause the loss of DNA-binding activity **(Figure 3.4A**).

This finding prompted me to construct a ChxR variant harboring an auxiliary background mutation that is distal from the pair of tryptophans at the dimeric interface. This mutation, when made alone, should still possess the wild-type level of DNA-binding activity, but cause a significant loss of binding affinity in combination with W89G. Such ChxR variant can serves as a “pseudo wild-type” platform to incorporate the W89G mutation and test the subsequent rescue by complementing small molecules. Taking heed of these requirements, further examination of the dimeric interface suggests F75A to be an excellent choice as the auxiliary mutation **(Figure 3.4B)**. The F75 residue interacts with P94 from the opposing subunit, and is located remotely from W89 with a shortest distance of 11 Å. Therefore, the removal of the F75 sidechain will cause a disconnected void volume from the W89G cavity, and most likely have no influence to the engineered small-molecule binding location. In addition, the peripheral position of F75 at the dimeric interface suggests that this F→A mutation conveys less energetic penalty to the stability of the ChxR dimer comparing to W89G. These observations lead to the hypothesis that the F75A mutation alone is insufficient to cause detectable loss of DNA-binding activity. However, the W89G mutation in combination with this background mutation will lead to a significant loss of function. As expected, the F75A mutant still forms the protein:DNA complex, with levels similar to the wild-type and the W89G variant. The W89G to this pseudo wild-type construct, however, completely lost the DNA-binding activity (**Figure 3.4A**). These observations suggest that the F75A ChxR variant can be used as a pseudo wild-type construct to incorporate the W89G mutation and examine the subsequent functional rescue of complementary small molecules.

Figure 3.4: Construction of pseudo wild-type ChxR variant. ***(A)*** W89G ChxR variant still demonstrates DNA binding activity identical to that of the wild type. F75A mutation alone does not lead to loss-of-function. Combining the two mutations, however, the F75A/W89G ChxR variant shows complete abolishment of DNA-binding activity. ***(B)*** Left panel: The position of F75 is distant from W89. W89 residues are shown in spheres; F75 residues are shown in magenta sticks. Right panel: P94 (green and cyan spheres) interacts with F75 (magenta spheres) at the periphery of the dimeric interface.


A
WT
W89G
F75A
F75A/W89G
1
25
50
1
25
50
1
25
50
1
25
50
ChxR:DR2 Complex
Unbound DR2
B

Discussion and Future Steps

In this study, I evaluated the generalized chemical rescue of structure approach using ChxR as a model structure, and identified small molecules that can specifically bind to the engineered ChxR mutant. The immediate next step is to characterize whether the binding of hit compounds identified by SPR can rescue the DNA binding activity of F75A/W89G ChxR. Biochemical assays, such as EMSA and fluorescence polarization assay, have been developed and optimized to test the rescue of DNA-binding of the double-mutation variant. Furthermore, our collaborator is currently developing an *in vivo* assay to examine the rescue in living *Chlamydia* cells. Recently, a shuttle vector based system has been discovered to allow a tight control of target gene expression (*151*). Utilizing this shuttle vector to transform *Chlamydia* cells, we can express the ChxR double-mutation variant in *Chlamydia* cells that harbor a reporter gene under control of a ChxR-specific promoter, and monitors reporter gene expression upon the addition of the hit compounds. By coupling the rescue of ChxR structure to transcriptional activation, this assay provides a straightforward readout of the modulation of ChxR cellular function by small molecules. One major advantage of this *in vivo* assay is that the rescue of one single ChxR molecule leads to multiple events of DNA transcription, and this amplification of signal may be able to detect the rescue by compounds with either low binding affinity or weak partial agonistic effect, which may have been difficult to conclusively demonstrate in the proposed biochemical assays.

The usage of ROCS as a virtual screening tool is also rather unconventional in this study: the ROCS software was originally developed as a ligand-based screening tool, for using a known drug lead to identify other potentially active compounds with similar volume and shape. The preliminary results from this study suggest that ROCS is also a suitable tool for matching small molecules to constellations of disconnected atoms arising from multiple cavity-forming mutations.

The preliminary results in this study provide a thorough evaluation of the generalized chemical rescue of structure, and strongly suggest the feasibility of employing chemical rescue of structure using multiple mutations. Accordingly, we expect that the insights offered here will enable rational design of a variety of new protein switches that utilizing diverse protein scaffold through this approach. In parallel, this methodology can also be utilized in the converse direction: rather than screen for a compound that matches some pre-selected constellation, we will instead screen for a constellation that matches our pre-selected compound of interest, which enables the design of selective sensor proteins to certain small molecules, especially those for which no naturally-occurring protein binding partner is known.

Chapter IV.
Identifying inhibitors of the Musashi-1 protein-RNA interaction by hotspot mimicry

Abstract

RNA-binding proteins (RBPs) are key regulators of post-transcriptional gene expression. This makes understanding and controlling interactions between RBPs and their cognate RNAs critical for decoding mechanisms that underlie many important biological processes. Small-molecule inhibitors of specific RBPs would be extremely useful, but conventional approaches to design nucleoside analogues or allosteric ligands are not suitable for targeting the majority of RBPs. Drawing inspiration from inhibitors of protein-protein interactions, here we develop a strategy that entails extracting a "hotspot pharmacophore" from the structure of a protein-RNA complex, then using this as a template for ligand-based screening. For a first application of this approach we have selected the RNA-binding protein Musashi-1 (Msi1), a stem cell marker that positively regulates the Notch and Wnt signaling pathways, promotes cell cycle progression, and is upregulated in many cancers. Using this "hotspot mimicry" strategy we identified compounds that match the hotspot pharmacophore from the Msi1 / RNA complex, enabling development of novel inhibitors of the Musashi-1 / *NUMB* mRNA interaction that are active in both biochemical and cell-based assays. This study extends the paradigm of "hotspots" from protein-protein complexes to protein-RNA complexes, supports the "druggability" of RNA-binding protein surfaces, and represents the first example of a rationally-designed small-molecule that inhibits a non-enzymatic RNA-

binding protein without relying on allostery. Owing to its simplicity and generality, we anticipate that the same approach may be used to develop inhibitors of many other RNA-binding proteins as well, thus enabling design of a broad new class of chemical tools and potential starting points for novel therapeutic agents.

Introduction

RNA-binding proteins (RBPs) play crucial roles in 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 (*152*). Mammalian proteomes are thought to include upwards of 800 RBPs (*153, 154*), 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 (*155-157*).

To date, there exist few classes of compounds that target protein-RNA interactions. Inhibitors of certain RBPs have been identified via high throughput screening (*158, 159*), including one series from virtual screening that competes with double-stranded RNA for binding to toll-like receptor 3 (*160*), and a number of compounds have been reported that disrupt binding by interacting with the RNA rather than with the RBP (*71, 72*). 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 is comprised of nucleoside analogues, such as anti-HIV-1 NRTIs, that mimic the chemical structures of natural-occurring nucleosides (*61-63*) and rely on enzymatic processing by their targets to form covalent adducts (*161*). While nucleoside analogues can be straightforward to design, the inability of these molecules to provide sufficient binding affinity or selectivity without covalent linkage precludes this strategy from being extended to non-enzymatic RBPs. The second class of compounds is comprised of allosteric inhibitors, such as anti-HIV-1 NNRTIs, that bind to secondary sites on the protein target and shift its conformation to an inactive state (*63, 69, 70*). In principle, allosteric inhibitors could be used to target both enzymatic and non-enzymatic RBPs; in practice, however, challenges associated with both identifying allosteric sites and then finding small molecules to complement these sites has limited the general utility of this approach to all but a few cases. Collectively, the fact that these RNA-binding protein surfaces are not thought to have evolved to bind any small-molecule makes them a "non-traditional" class of drug target. The relatively flat and polar nature of protein surfaces in this class typically leads to poor performance by structure-based virtual screening (docking) approaches (*33*), 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 (*162*).

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

(*38-40*). This observation, in turn, motivated several groups to mimic these key interactions when designing small-molecule inhibitors (*43, 46, 48, 163*). 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.

Materials and Methods

PDB structures used in calculations

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

Building hotspot pharmacophores

Hotspot pharmacophores were built using a dedicated protocol implemented in the Rosetta software suite (*77*), and is freely available for academic use (www.rosettacommons.org).

To select deeply buried RNA bases, the solvent accessible surface area (SASA) of each base in the RNA was calculated in the presence and absence of the protein. A base was carried forward if the change in SASA upon complexation was greater than a preset cutoff value (46.81 Å^2 for adenine, 31.09 Å^2 for cytosine, 45.06 Å^2 for guanine and 52.66 Å^2 for uracil); these values correspond to the median values of protein-RNA complexes in the PRIDB database.

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

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

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

The resulting interaction maps are then clustered using a modified version of Kruskal's minimum spanning tree algorithm. We first build a complete graph, in which vertices are the ring moieties, and the edge weights are the Euclidean distances between vertices. Then we take edges in ascending order and cluster the end vertices of that edge if no cycle would be caused. We halt the clustering when the distance is greater than a user-specified cutoff value (default 5.0 Å). The donor/acceptor atoms are then assigned to the closest ring moieties if the distance is less than another user-specified value (default 5.0 Å). Finally, we output the pharmacophore templates if the cluster contains at least two ring moieties. The Kruskal clustering code is also implemented in Rosetta, and is carried out as follows:

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

Identifying complementary ligands

We used ROCS to screen large libraries for compounds that match the hotspot pharmacophore. We downloaded the standard ‘drugs-now’ subset of ~7 million molecules from ZINC database for screening (*150*). We generated up to 100 conformers for each molecule in the database using OMEGA (*164-166*). 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 (*77*). The top-scoring compounds were visually inspected and selected for experimental validation based on cost and availability.

Predicting target selectivity

Hotspot pharmacophores were extracted from each protein-RNA complex in the “RB344” dataset, which contains a non-redundant set of 344 protein-RNA complexes from the PDB extracted in March 2013. The dataset was retrieved from the Protein-RNA Interface Database (PRIDB) v2.0 (http://pridb.gdcb.iastate.edu/download/RB344.txt). Conformers for each compound were generated by OMEGA using the following command line:

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

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

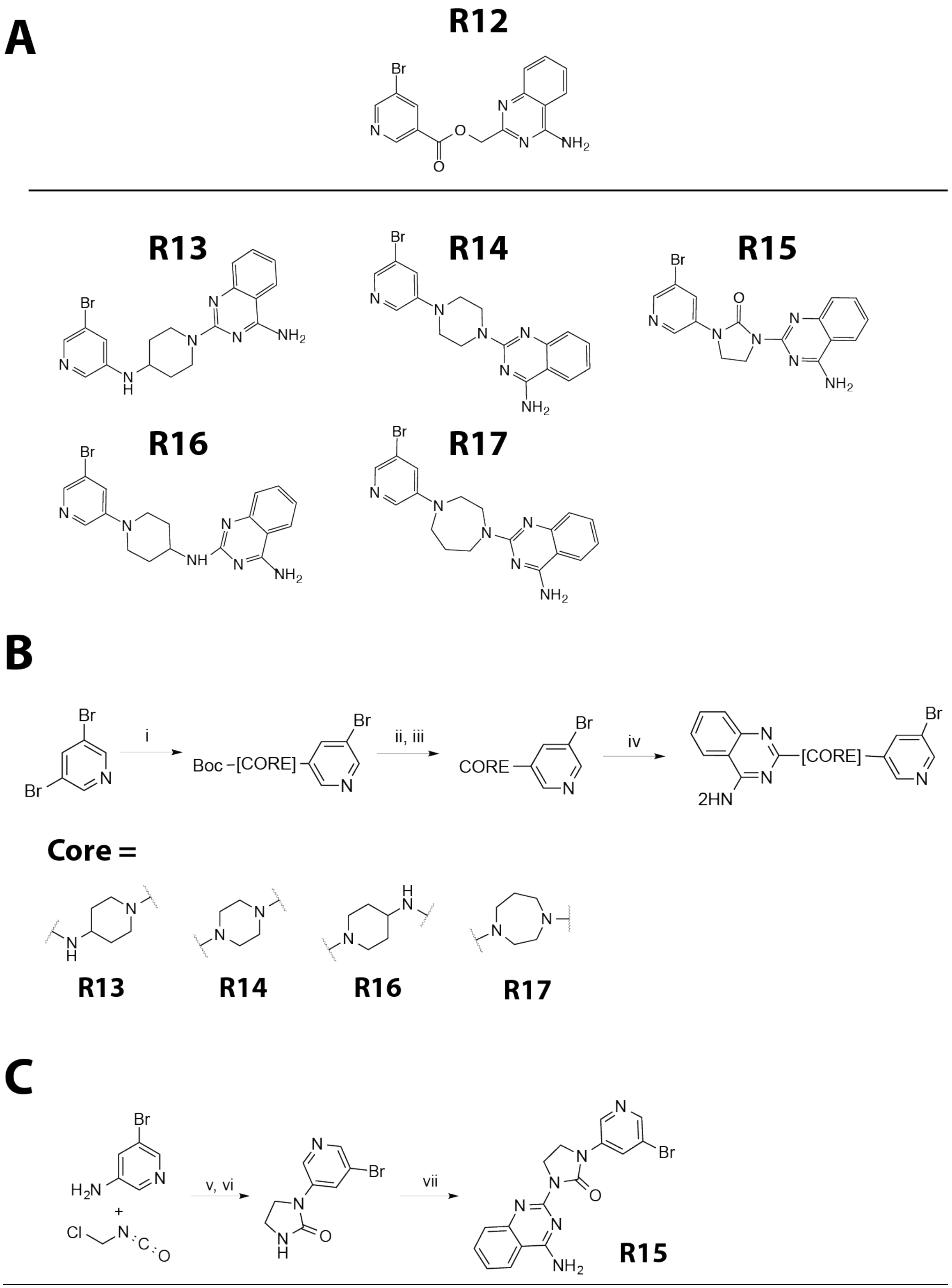
```
rocs –dbase conformer_ensemble.pdb –query hotspot.pdb –oformat pdb –rankby
FitTverskyCombo
```

Synthesis of R12 derivatives : overview

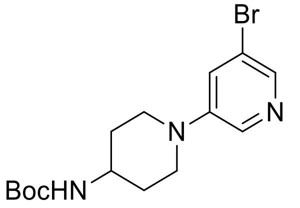
All air- and moisture-sensitive reactions were carried out in flame- or oven-dried glassware under argon atmosphere using standard gastight syringes, cannula, and septa. Stirring was achieved with oven-dried magnetic stir bars. Flash column chromatography was performed with SiO2 from Sorbent Technology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

Synthesis of specific intermediates / derivatives is described in detail in the following sections, and summarized in **Figure 4.1**.

Figure 4.1: Synthesis of R12 derivatives. (A) Chemical structures of all derivatives R13-R17. **(B)** Scheme leading to synthesis of all derivatives R13-R17 except R15. Reagents and conditions: (i) $Pd_2(dba)_3$, *t*-BuONa, BINAP, 80 ϒC, toluene, BocCOREH, 59-91%. (ii) Dioxane, HCl, >95% yield. (iii) Aqueous $NaCO_3$, CH_2Cl_2, >95%. (iv) CH_3CN, 2-chloroquinazolin-4-amine, 180 ϒC, MW, 58-71%. **(C)** Scheme leading to synthesis of R15. Reagents and conditions: (v) DIEA, CH_3CN, 96%. (vi) NaH, THF, 83%. (viii) NaH, THF, 69.1%.

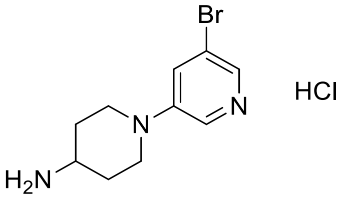

A
R12
R13
R14
R15
R16
R17
Br
N
O
NH
NH2
B
i
Boc –[CORE]
ii, iii
CORE
iv
[CORE]
2HN
Core =
R13
R14
R16
R17
C
H2N
+
Cl
N
C
O
v, vi
vii
R15

Synthesis of R12 derivatives : general procedure #1



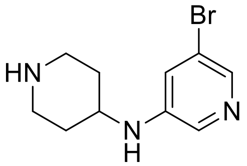
Following a modified procedure outlined by Do et al. (*167*), a mixture of *tert*-butyl 4-aminopiperidine-1-carboxylate (0.85 g, 4.22 mmol), 3,5-dibromopyridine (1.0 g, 4.22 mmol) tris(dibenzylideneacetone)dipalladium(0) (0.077 g, 0.084 mmol), (±)-2,2'-bis(diphenylphosphino)-1,1'-binaphthalene (0.11 g, 0.17 mmol) and sodium-t-butoxide (0.61 g, 6.33 mmol) in toluene (30.2 mL) was heated to 80 °C for 16 h, then the reaction mixture was allowed to cool to ambient temperature, diluted with ether (100 mL) and washed with brine (3×30 mL). The organic layer was dried over $MgSO_4$, filtered and concentrated under vacuum. The residue was purified by silica gel chromatography (50% EtOAc in hexanes, Rf = 0.5) to afford the title compound *tert*-butyl 4-((5-bromopyridin-3-yl)amino)piperidine-1-carboxylate (879.8 mg, 2.47 mmol, 59% yield) as a white solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.22 (d, *J* = 2.6 Hz, 1H), 8.11 (d, *J* = 1.9 Hz, 1H), 7.31 (dd, *J* = 2.6, 1.9 Hz, 1H), 4.55 (s, br. 1H), 3.70 – 3.60 (m, 3H), 2.97 – 2.90 (m, 2H), 2.14 – 2.03 (m, 2H), 1.60 – 1.49 (m, 2H), 1.47 (s, 9H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 155.10, 147.64, 140.51, 136.65, 124.72, 120.81, 47.51, 47.42, 31.88, 28.40.

Synthesis of R12 derivatives : general procedure #2



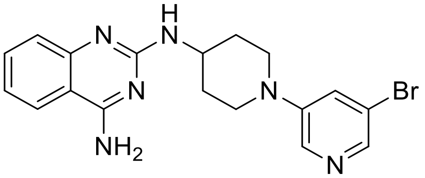
To a solution of tert-butyl 4-((5-bromopyridin-3-yl)amino)piperidine-1-carboxylate (433.5 mg, 1.22 mmol) in dichloromethane (9 mL), was added hydrogen chloride in dioxane (15.2 mL, 60.8 mmol). The reaction was stirred at rt for 16 h. Solvents were removed to give a white solid. Yield: 445.0 mg, 100%. ^{1}H NMR (400 MHz, MeOD) δ 8.53 (d, *J* = 2.4 Hz, 1H), 8.38 – 8.31 (m, 2H), 4.19 – 4.10 (m, 2H), 3.83 – 3.58 (m, 3H), 3.53 – 3.47 (m, 1H), 3.26 – 3.14 (m, 2H), 2.26 – 2.17 (m, 2H), 1.84 – 1.73 (m, 2H). ^{13}C NMR (101 MHz, MeOD) δ 150.00, 132.59, 131.09, 127.98, 123.69, 73.59, 72.48, 62.23, 46.60, 43.84.

Synthesis of R12 derivatives : general procedure #3



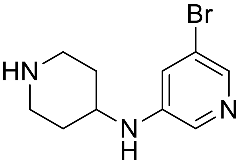
To a solution of 5-bromo-N-(piperidin-4-yl)pyridin-3-amine dihydrochloride (200.0 mg, 0.55 mmol) in methanol (4 mL), was added sodium carbonate (174.0 mg, 1.64 mmol). Solvent was removed and residue was extracted with DCM. DCM was dried over $MgSO_4$ and evaporated to dryness to give a light-yellow oil.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

Synthesis of R12 derivatives : general procedure #4



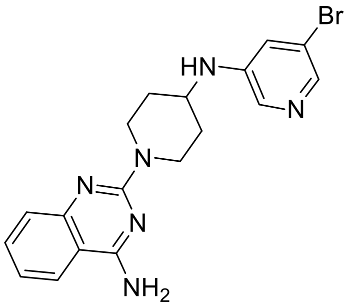
To a solution of 2-chloroquinazolin-4-amine (28.4 mg, 0.16 mmol) and 5-bromo-N-(piperidin-4-yl)pyridin-3-amine (81.0 mg, 0.32 mmol) in acetonitrile (2 mL) was heated at 180 °C for 1 h under microwave irradiation. The material was purified via reverse phase combiflash first, followed by further purification via silica gel chromatography (DCM/MeOH = 10:1, Rf = 0.3) to give 2-(4-((5-bromopyridin-3-yl)amino)piperidin-1-yl)quinazolin-4-amine (36.4 mg, 0.091 mmol, 58% yield) as a white solid. ^{1}H NMR (400 MHz, CDCl$_3$) δ 8.24 (d, *J* = 2.6 Hz, 1H), 8.12 (d, *J* = 1.9 Hz, 1H), 7.65 – 7.56 (m, 2H), 7.51 – 7.48 (m, 1H), 7.33 (t, *J* = 2.3 Hz, 1H), 7.16 – 7.12 (m, 1H), 5.57 (s, br. 2H), 5.01 (s, br. 1H), 4.23 – 4.14 (m, 1H), 3.72 – 3.62 (m, 2H), 3.07 – 3.00 (m, 2H), 2.25 – 2.20 (m, 2H), 1.71 – 1.56 (m, 2H). ^{13}C NMR (101 MHz, CDCl$_3$) δ 162.1, 158.5, 152.3, 147.8, 140.4, 136.7, 133.4, 125.6, 124.6, 121.9, 121.5, 120.8, 110.4, 47.5, 47.3, 31.9. HRMS (m/z): calcd for $C_{18}H_{20}BrN_6$ (neutral M+H) 399.0933; found 399.0900.

Synthesis of R12 derivatives : specific compounds

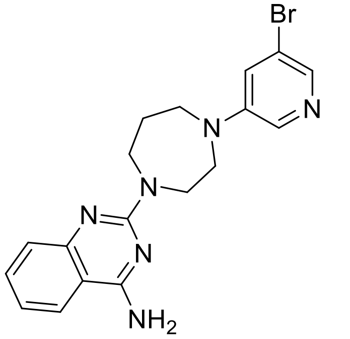


Synthesized by using general procedures #1, #2 and then #3. ^{1}H NMR (400 MHz, DMSO-d_6) δ 8.35 (d, *J* = 2.6 Hz, 1H), 8.15 – 8.02 (m, 4H), 7.72 (t, *J* = 2.2 Hz, 1H), 3.92

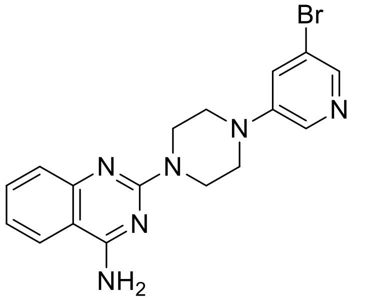
(d, *J* = 13.5 Hz, 2H), 3.25 (d, *J* = 12.5 Hz, 1H), 2.97 – 2.84 (m, 2H), 1.95 (d, *J* = 10.8 Hz, 2H), 1.57 (qd, *J* = 12.2, 4.1 Hz, 2H). HRMS (m/z): calcd for $C_{10}H_{14}BrN_3$ (neutral M+H) 255.0371; found 255.0366.



Synthesized using general procedure 4 by reacting 2-chloroquinazolin-4-amine (0.036 g, 0.203 mmol) and 5-bromo-N-(piperidin-4-yl)pyridin-3-amine (.078 g, 0.305 mmol) to give 2-(4-((5-bromopyridin-3-yl)amino)piperidin-1-yl)quinazolin-4-amine (.062 g, 0.155 mmol, 76 % yield) ^{1}H NMR (400 MHz, Acetone-*d*$_6$) δ 7.97 (d, *J* = 2.5 Hz, 1H), 7.88 (dd, *J* = 8.1, 1.4 Hz, 1H), 7.78 (d, *J* = 1.9 Hz, 1H), 7.47 (ddd, *J* = 8.4, 6.8, 1.4 Hz, 1H), 7.30 (dd, *J* = 8.5, 1.3 Hz, 1H), 7.17 (t, *J* = 2.2 Hz, 1H), 7.00 (ddd, *J* = 8.2, 6.8, 1.2 Hz, 1H), 5.38 (d, *J* = 8.1 Hz, 1H), 4.83 – 4.66 (m, 2H), 3.63 (tdd, *J* = 6.5, 4.2, 2.4 Hz, 1H), 3.08 (ddd, *J* = 13.8, 11.5, 2.7 Hz, 2H), 1.45 – 1.30 (m, 2H). ^{13}C NMR (101 MHz, Acetone) δ 206.12, 163.36, 159.98, 153.94, 146.20, 138.11, 135.70, 133.34, 126.43, 123.77, 121.26, 120.48, 111.01, 54.95, 50.64, 43.27, 32.60, 30.41, 30.22, 30.03, 29.83, 29.64, 29.45, 29.26. HRMS (m/z): calcd for $C_{18}H_{20}BrN_6$ (neutral M+H) 399.0871; found 399.0855.

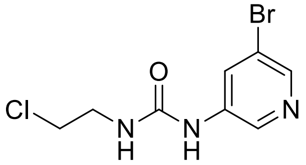


Synthesized using general procedure 4 by reacting 2-chloroquinazolin-4-amine (0.045 g, 0.248mmol) and 1-(5-bromopyridin-3-yl)-1,4-diazepane (*168*) (0.127 g, 0.496 mmol) to give 2-(4-(5-bromopyridin-3-yl)-1,4-diazepan-1-yl)quinazolin-4-amine (0.067g, 0.168 mmol, 68% yield) ^{1}H NMR (400 MHz, Chloroform-*d*) δ 8.05 (d, *J* = 2.7 Hz, 1H), 7.92 (d, *J* = 1.7 Hz, 1H), 7.57 – 7.45 (m, 3H), 7.13 – 7.01 (m, 2H), 4.10 – 4.00 (m, 2H), 3.72 (t, *J* = 6.2 Hz, 2H), 3.64 (t, *J* = 5.3 Hz, 2H), 3.49 (t, *J* = 6.2 Hz, 2H), 2.10 (t, *J* = 6.2 Hz, 2H). ^{13}C NMR (101 MHz, CDCl$_3$) δ 161.72, 158.08, 144.71, 137.58, 133.16, 132.56, 125.72, 121.87, 121.24, 121.14, 120.16, 109.69, 77.36, 77.04, 76.72, 53.75, 49.89, 47.84, 46.61, 46.22, 24.37. HRMS (m/z): calcd for $C_{18}H_{20}BrN_6$ (neutral M+H) 399.0864; found 399.0855.

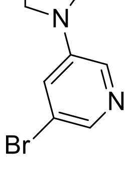


Synthesized using general procedure 4 by reacting 2-chloroquinazolin-4-amine (0.022 g, 0.124 mmol) and 1-(5-bromopyridin-3-yl)piperazine (*168*) (.03 g, 0.124 mmol) to give 2-(4-(5-bromopyridin-3-yl)piperazin-1-yl)quinazolin-4-amine (.031 g, 0.080 mmol, 64.9 % yield). ^{1}H NMR (400 MHz, Chloroform-*d*) δ 8.26 (d, *J* = 2.6 Hz, 1H), 8.14

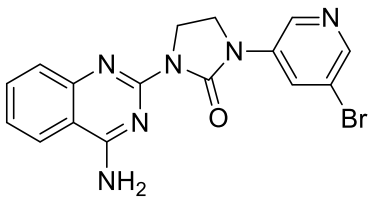
(d, *J* = 1.9 Hz, 1H), 7.63 – 7.48 (m, 3H), 7.34 (dd, *J* = 2.6, 1.9 Hz, 1H), 7.13 (ddd, *J* = 8.1, 6.8, 1.3 Hz, 1H), 4.42 – 3.94 (m, 4H), 3.58 – 3.11 (m, 4H). ^{13}C NMR (126 MHz, $CDCl_3$) δ 161.68, 158.76, 152.60, 147.97, 141.01, 136.62, 133.35, 126.08, 124.64, 121.76, 121.62, 120.86, 109.91, 48.26, 43.48. HRMS (m/z): calcd for $C_{17}H_{17}BrN_6$ (neutral M+H) 384.0698; found 384.0709.



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



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



To a dried round bottomed flask was added 1-(5-bromopyridin-3-yl)imidazolidin-2-one (.1 g, 0.413 mmol), THF and a stir bar. The mixture was cooled to 0 °C. Then, sodium hydride (0.030 g, 1.239 mmol) was added slowly. The mixture was allowed to stir for 30 minutes. Then, 2-chloroquinazolin-4-amine (0.074 g, 0.413 mmol) was added. The reaction was allowed to warm to room temperature slowly and then continue stirring for 12 hours. The reaction was quenched with water, extracted three times with ethyl acetate, washed with brine, dried with sodium sulfate and then concentrated en vacuo. The crude residue was purified via normal phase chromatography (Methanol/dichloromethane) and tehn reverse phase chromatography (water pH = 9, acetonitrile) to yield 1-(4-aminoquinazolin-2-yl)-3-(5-bromopyridin-3-yl)imidazolidin-2-one (.11 g, 0.286 mmol, 69.1 % yield). ^{1}H NMR (400 MHz, DMSO-d_6) δ 8.83 (d, J = 2.3 Hz, 0H), 8.50 – 8.38 (m, 1H), 8.23 – 8.13 (m, 0H), 7.82 (s, 1H), 7.70 (ddd, J = 8.4, 6.9, 1.4 Hz, 0H), 7.57 – 7.51 (m, 0H), 7.34 (ddd, J = 8.3, 7.0, 1.3 Hz, 0H), 4.15 (dd, J = 9.2, 6.6 Hz, 1H), 4.04 – 3.92 (m, 1H). ^{13}C NMR (101 MHz, DMSO) δ 162.52, 154.72, 152.23, 150.58, 143.31, 137.97, 137.47, 133.05, 126.49, 126.33, 123.56, 119.65, 111.96, 102.47, 41.66, 40.64. HRMS (m/z): calcd for $C_{16}H_{13}BrN_6O$ (neutral M+H) 384.0334; found 384.0340.

Model building of R12 derivatives

Conformers of R13-R17 were generated using OMEGA. For each compound, we aimed to generate sample likely conformers that optimally matched the ring geometry in the hotspot pharmacophore. To achieve this, we used the CHARMM software (*169*) to carry out a biased energy minimization of the compound (in the absence of the protein). We implemented the bias using a harmonic constraint applied to the Cartesian coordinates of certain atoms, centered at the position of the corresponding atom of the hotspot pharmacophore (C4,C5,N7,C8 and N9 on Adenine106 and N1, C2, N3, C4, C5, C6, N6, N7, C8 and N9 on Guanine107) and with a scale factor of 100 (this scale factor is related to the force constant in a way that depends on the mass of individual atoms). The residue topology file and parameter file for the compounds required by CHARMM minimization were obtained from CHARMM-GUI (*170*).

Through this minimization of OMEGA conformers, we generated models of the R12 derivatives that maintain the ring geometry in the hotspot pharmacophore but contain a variety of geometries in the linker region. Since the resulting conformers match the hotspot pharmacophore, they are already aligned to the structure of the Msi1-RNA complex. We concluded by selecting the best model on the basis of protein-ligand interaction energy using the fullatom Rosetta energy function (*77*).

Expression and purification of Msi1

A gene encoding human Msi1 RBD1 domain was subcloned as a fusion protein with an N-terminal 6xHis-tagged streptococcal GB1 domain and a tobacco etch virus (TEV) protease site. The expression plasmid was transformed into *Escherichia coli* BL21(DE3) pLysS, then a 5 ml overnight starter culture was used to inoculate a 1 L culture of LB media. Cells were grown at 37 ºC to an OD600 of 0.6–0.8 and were induced with 1 mM IPTG overnight at 15 ºC. The induced cells were resuspended in lysis buffer (50 mM Tris, 150 mM NaCl, 5 mM imidazole pH 8.0) and sonicated for 10 minutes (Fisher Scientific Sonic Dismembrator Model 100). The cell lysates were then centrifuged at 15,000g for 30 min. The GB1-RBD1 remained in the supernatant, which was purified by HPLC affinity chromatography with Ni-chelated Sepharose Fast Flow Resin (GE Healthcare), followed by a HiLoad 16/60 Superdex 75 gel filtration column (GE Healthcare). GB1 tag was digested with TEV protease (1 OD_{280} of TEV per 5 OD_{280} of fusion protein) in reaction buffer (50 mM Tris-HCl, 0.5 mM EDTA and 1mM DTT pH 8.0). All protein concentrations were determined with reference to bovine albumin standards using Bradford assays.

Fluorescence polarization competition assays

RNA oligonucleotides were ordered from Integrated DNA Technologies (Coralville, IA) and dissolved in TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8.0): sequences are included in **Table 4.1**. To measure the dissociation constant of Msi1 RBD1 and RNA binding, a fixed concentration (2 nM) of fluorescein-labeled RNA (FC-NUMB, **Table 4.1**) and increasing concentrations of Msi1 RBD1 (1 nM to 1000 nM) were mixed in binding assay buffer (20 mM HEPES, 150 mM NaCl, 0.05% F-68 pH 7.4). Fluorescence intensities were measured in replicate on the BioTek Synergy 2 plate reader (Winooski, VT) and the fluorescence polarization value (FP) was calculated by the following equation:

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

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

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

To test the contribution to binding affinity of each base in the NUMB RNA sequence, we purchased five oligos that each harbor an abasic site at a different position, as well as the corresponding wild-type (aNUMB0-5, **Table 4.1**). K_i values were then determined from a competition experiment in which serial dilution of unlabeled aNUMB oligos (5 nM – 5000 nM) were added to compete against a fixed concentration (2 nM) of fluorescein-labeled RNA (FC-NUMB) for binding to a fixed concentration of Msi1

RBD1 (75 nM). The K_i value was determined by fitting with Prism to the "Binding – Competitive – One site – Fit Ki" model (*171*).

To examine the displacement of FC-NUMB by R13, we performed the same competition assay using R13 as a competitor (5 μM – 150 μM). The K_i value was determined as described above.

Table 4.1: Sequences of RNA oligonucleotides used in this study. "FC" refers to the fluorescein label, and "x" refers to an abasic site (i.e. internal RNA spacer site). After validation to ensure binding to Msi1 RBD1 (**Figure 4.11**), the FC-NUMB construct was used in fluorescence competition assays (**Figure 4.2D**, **Figure 4.4G**).

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

Surface plasmon resonance

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

All SPR runs were performed at 25 ºC using a flow rate of 50 μL/min in running buffer (HEPES buffered saline with 0.05% Tween-20 pH 7.4). Compounds were injected over sample and reference flow cells at a concentration of 50 μM in running buffer, for 250 s. Following each injection, flow cells were regenerated with a 20 s injection of 1 M NaCl.

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

Differential scanning fluorimetry (Thermofluor)

Differential scanning fluorimetry (DSF) experiments were carried out using a standard protocol described by others (*172*). All experiments were carried out in a reaction volume of 25 μL, with 25 mM Tris-Cl pH 8.0, 120 mM NaCl, 2% DMSO, and 100x-diluted Sypro Orange dye (Invitrogen). Multiple concentrations of GB1-tagged

Msi1 RBD1 (ranging from 1 μM to 15 μM) were tested to identify the lowest concentration necessary to generate a smooth melting curve. For subsequent experiments, we used a concentration of 7.5 μM.

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

The melting temperature (Tm) values were determined by taking the maximum of the first derivative of the raw fluorescence intensity with respect to temperature (*172*), using GraphPad Prism 5. ΔTm values were calculated by comparing the Tm of a particular R13 concentration to that of the DMSO control.

Computational Approach

Computational methods are implemented in the Rosetta software suite (*77*) 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.

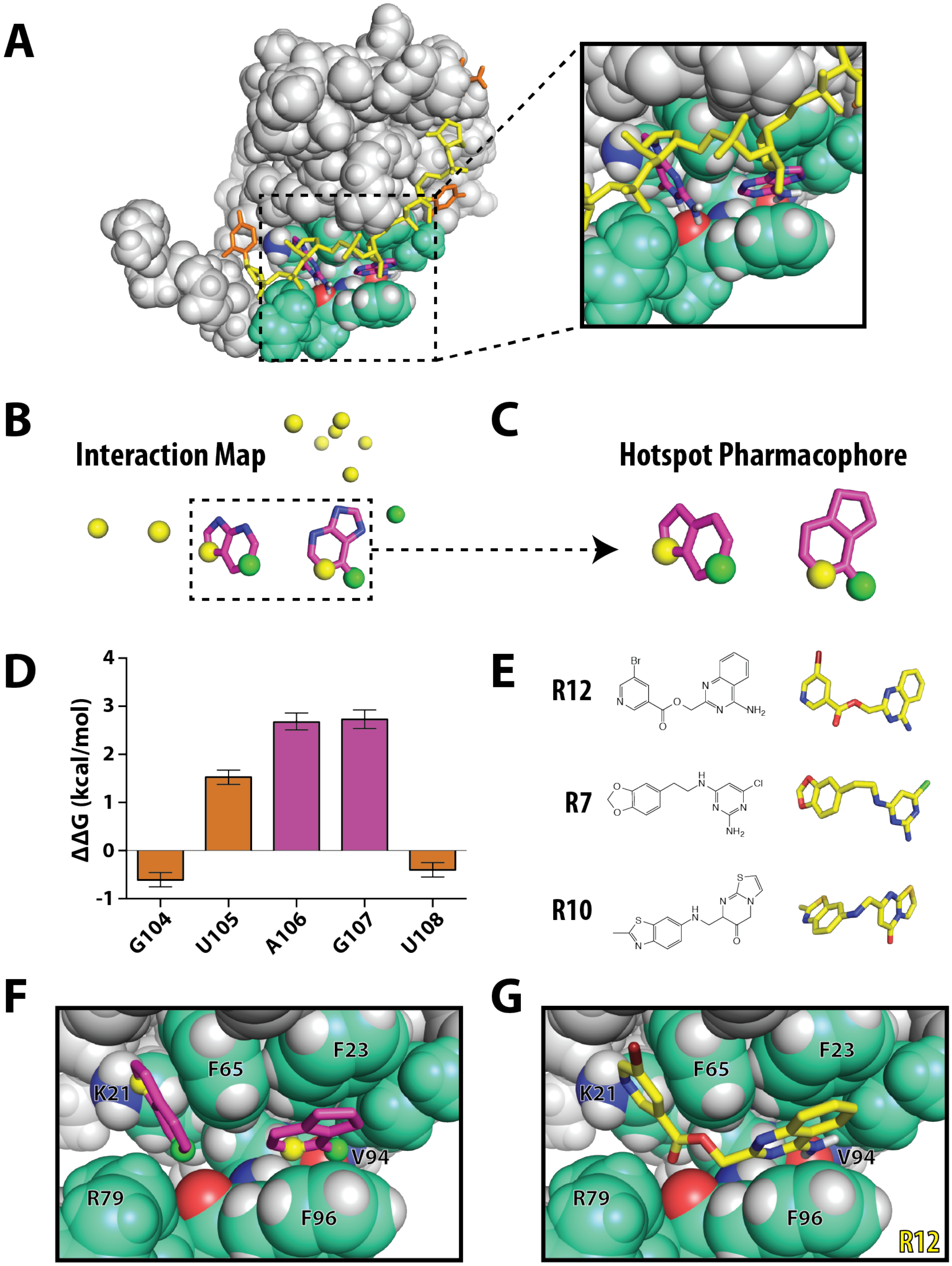
Building "hotspot pharmacophores"

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 4.2A**); this uneven distribution is reminiscent of "hotspot" sidechains in protein-protein complexes (*38, 39*).

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 4.2B**). Any polar atoms on the nucleobases that do not participate in hydrogen bonds are then replaced with carbon atoms, since those polar groups need not be carried forward into inhibitor design. This gives a broad spatial map of the protein-RNA interaction, which typically cannot be spanned by a single drug-like small molecule; we therefore clustered neighboring moieties, and advanced each cluster separately. Through this approach, we reduce the structure of the protein-RNA complex to a minimal “hotspot pharmacophore” that encapsulates the key interactions to be recapitulated by a small molecule (**Figure 4.2C**).

Figure 4.2: The hotspot mimicry approach. We demonstrate this approach by applying it to the Msi1 / *NUMB* mRNA interaction. **(A)** The structure of the Msi1 / RNA complex. The RNA (*yellow and orange sticks denoting the backbone and the bases, respectively*) wraps around the protein (*green and grey spheres*). Two adjacent bases, A106 and G107 (*magenta*), are buried in a shallow pocket on the protein surface. **(B)** An interaction map is generated from the RNA in the complex, by collecting deeply buried bases (*magenta*) and atoms involved in intermolecular hydrogen bonds (*acceptors shown in yellow, donors in green*). **(C)** Components of the interaction map are clustered in space, and atoms that do not participate in hydrogen bonding are reverted to carbon atoms; this produces a "hotspot pharmacophore." **(D)** The difference in binding free energy between an RNA harboring a single abasic site versus the wild-type *NUMB* mRNA, as determined through competition with a fluorescently-labeled RNA. Positive values indicate diminished binding when a given base is replaced with an abasic site, showing that A106 and G107 contribute more than the other nearby bases to Msi1 / *NUMB* mRNA binding affinity. **(E)** The hotspot pharmacophore serves as a template for ligand-based screening. In this case we identified three classes of hit compound that mimic the three-dimensional features of the pharmacophore, as exemplified by the representatives shown here. **(F)** Superposition of the hotspot pharmacophore back onto the protein structure illustrates the interactions that should be captured by an ideal ligand: stacking against three aromatic sidechains, and four intermolecular hydrogen bonds. **(G)** Superposition of R12 onto the protein structure shows that this compound is expected to preserve the aromatic stacking, and recapitulate three of the four hydrogen bonds.


A
B
Interaction Map
C
Hotspot Pharmacophore
D
ΔΔG (kcal/mol)
4
3
2
1
0
-1
G104
U105
A106
G107
U108
E
R12
R7
R10
F
K21
F65
F23
V94
R79
F96
G
K21
F65
F23
V94
R79
F96
R12

Identifying complementary ligands

To identify such compounds, we used this hotspot pharmacophore as a template for carrying out ligand-based virtual screening. In order to facilitate rapid characterization of compounds emerging from our screen, we restricted our search to the ~7 million compounds in the ZINC database (*150*) that are both commercially available, and predicted to have drug-like physicochemical properties. We used OMEGA (OpenEye Scientific Software, Santa Fe, NM) (*164-166*) to build low-energy conformations of each compound, then ROCS (OpenEye Scientific Software, Santa Fe, NM) (*173, 174*) to align each conformation to our hotspot pharmacophore. For each of the top-scoring hits emerging from ROCS, we then used the aligned orientation to position the compound relative to the protein, and evaluated the interaction energy of the protein-ligand complex using the fullatom Rosetta energy function (*77*).

Musashi-1, an RRM-containing protein

The approach described above can, in principle, be applied to the structure of any protein-RNA complex. As a first test, we selected a target from the most common and well-studied of RNA-binding modules, the RNA-recognition motif (RRM) domain. Hundreds of structures of RRMs have been deposited in the Protein Data Bank, including more than fifty in complex with RNA (*175*). 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 (*176*). Accordingly, mutations to the protein that remove these aromatic sidechains have been shown to disrupt binding in representative RRMs (*176, 177*), as has introduction of non-canonical bases to the RNA that alter the pattern of hydrogen bonding groups (*178-180*). 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 (*176*).

Mammalian Musashi-1 (Msi1) recognizes its cognate RNAs through a pair of RRMs, RBD1 and RBD2 (*181*). 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 (*59, 60*). *NUMB* mRNA encodes an inhibitor of Notch, so translational inhibition by Msi1 triggers Notch signaling and thus promotes self-renewal and cell survival (*35, 60*). Considering the role of Msi1 in stem cell maintenance and renewal, distrupting its RNA-binding ability may inhibit cancer stem cells that play a role in drug- and radioresistance, and thus serve as an attractive potential anti-tumor strategy (*182*).

Results

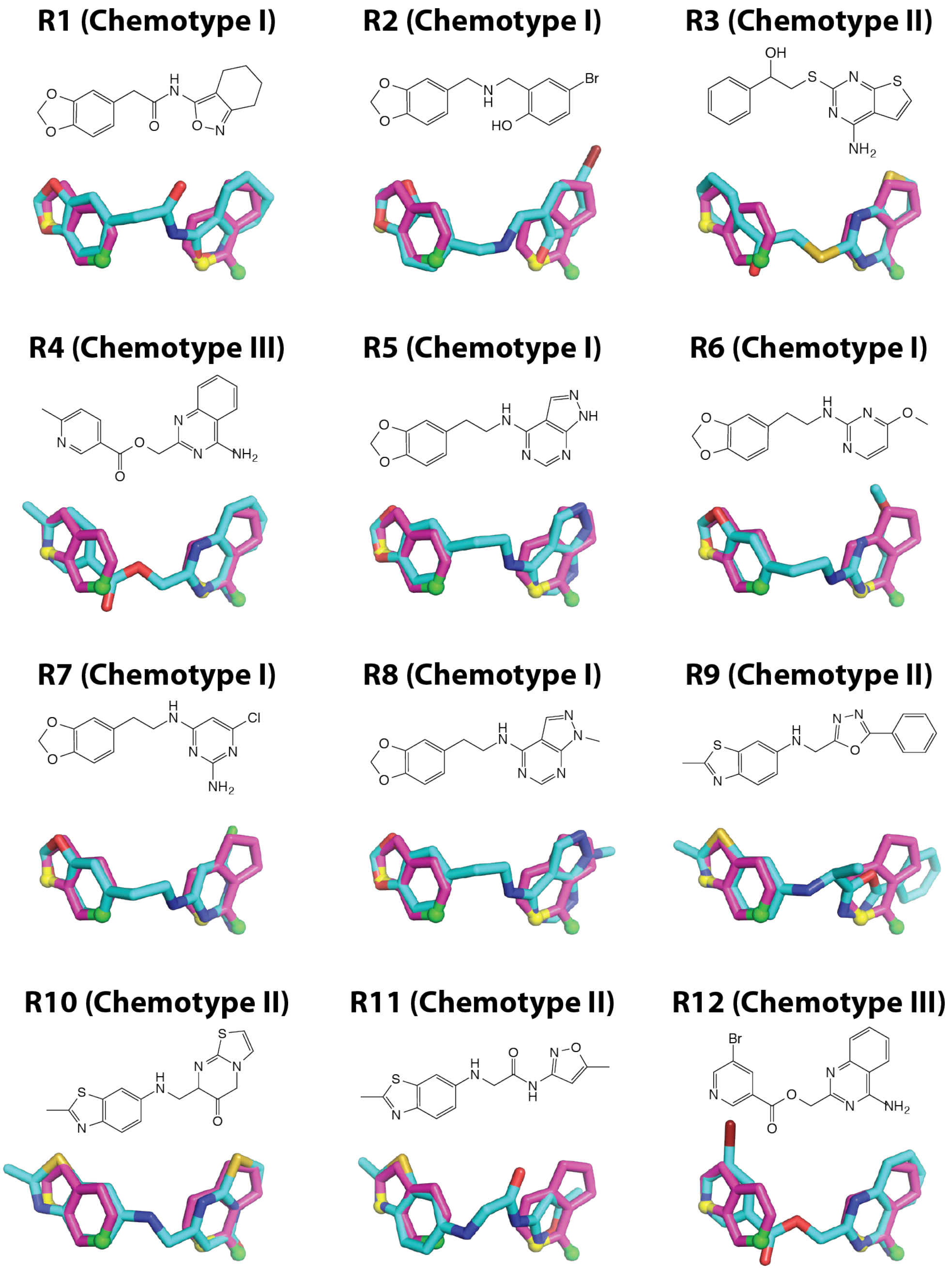
Computational screening against Msi1 RBD1

We applied our "hotspot mimicry" approach to the Musashi-1 RBD1 / *NUMB* mRNA complex (*181*), and found a single hotspot pharmacophore derived from an adjacent pair of buried nucleotides, Adenine106 and Guanine107 (**Figure 4.2A**). 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 4.2C**). To test whether these particular two bases indeed serve as a hotspot of the Msi1 RBD1 / RNA interaction, we used a fluorescence polarization competition assay to measure the binding affinity of *NUMB* mRNA variants that lacked individual bases. Using this assay, we found that introduction of an abasic site at either of these two positions led to a marked decrease in binding to Msi1 RBD1 (**Figure 4.2D**). In contrast, introduction of an abasic site at other nearby positions affecting binding much less. Confirmation that A106 and G107 serve as hotspots bases for 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 4.3**). While none of these scaffolds bear any obvious resemblance in chemical structure to a nucleobase pair, the overlap in three-dimensional shape and hydrogen bonding potential between the hotspot pharmacophore and the modeled conformation of each compound is immediately evident (**Figure 4.2E**). Despite this strong similarity, none of the 12 hit compounds recapitulated all four of the polar groups included in the hotspot pharmacophore, and only three hit compounds matched to three of the polar groups: R12, its close analog R4, and R7. The lack of hits that simultaneously match all four polar groups reflects a limit of the chemical space spanned by our screening library; we will discuss this in detail later.

As expected, superposition of the hit compounds back onto the hotspot pharmacophore in the context of the protein-RNA complex confirmed that these ligands might preserve the favorable interactions of the dinucleotide pair. In particular, the ring moieties in the pharmacophore represent the stacking of nucleobases against Phe23, Phe65 and Phe96 of Msi1, while the hydrogen bonding atoms indicate polar contacts with the sidechain of Lys21 and the backbones of Val94 and Phe96 of Msi1 (**Figure 4.2F)**. Mimicry of these interactions through the hotspot pharmacophore allows the hit compounds to recapitulate these interactions, as exemplified by R12 (**Figure 4.2G**). In this model R12 adopts a similar three-dimensional geometry as the hotspot pharmacophore, and thus recapitulates its aromatic stacking and polar interactions.

Figure 4.3: 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*).



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)

Biochemical characterization of computational hits

We purchased each of the compounds corresponding to these 12 top-scoring hits (**Figure 4.3**). We used surface plasmon resonance (SPR) to directly test for binding of each compound to Msi1, by immobilizing recombinant human Msi1 RBD1 onto an SPR chip and then passing each compound over the chip at a concentration of 50 μM (see *Materials and Methods*). The sensorgram for R12 showed a kinetic profile consistent with binding to Msi1 RBD1 (**Figure 4.4A**); none of the other compounds exhibited this behavior (**Figure 4.5**).

As noted earlier, only R12 and two other compounds matched as many as three polar groups in the hotspot pharmacophore; the lack of binding observed for the other compounds (at this concentration) may be attributable to the fact that they do not sufficiently recapitulate the interactions of the hotspot pharmacophore. While R7 matched three polar groups, retrospective analysis of the structural model revealed that the imperfect alignment of the rings to the pharmacophore may have led to a steric clash with the protein (**Figure 4.6**).

Interestingly, the R12 class was comprised of two compounds: R12 and R4. These compounds differ only in the position and identity of a single substituent: the R12 has a bromine atom at the meta position of the pyridine moiety, while R4 instead harbors a methyl group at the para position. Comparison of these compounds in the context of the protein partner immediately reveals a potential source for their differing responses in the SPR experiment: in our model, this methyl group of R4 forms a steric clash with the side chain of Leu31 on Msi1 that we had not initially recognized (**Figure 4.4B**); in contrast, the shifted position of the R12 substituent avoids this steric clash. This initial (inadvertent) structure-activity experiment provides strong support for the structural model of R12 binding.

Figure 4.4: Biochemical characterization and optimization of computational hit compounds. (A) Initial screening via surface plasmon resonance: representative sensorgrams for R12 and R4 are shown. The kinetic profile of R12 (*red*) is consistent with binding to Msi1, whereas that of its close analog R4 (*blue*) shows no evidence of binding. **(B)** Comparison between the predicted binding models of R12 and R4. The top scoring conformers of R12 (*red*) and R4 (*blue*) are transferred back to the protein by alignment to the hotspot pharmacophore. The model of R4 suggests its lack of binding may stem from a steric clash with Leu31, whereas R12 avoids this steric clash since this ring is substituted at a different position. **(C)** Chemical structures of R12 and one of its derivatives, R13, that replaces R12's ester with a piperidine ring and a secondary aimine in the linker. **(D)** Model of R13 bound to Msi1, by alignment to the hotspot pharmacophore. R13 preserves the interactions of R12, but reduces flexibility of the linker and removes potential electrostatic repulsion with Msi1. (**E**) R13 shows kinetic profile consistent with reversible binding using surface plasmon resonance at the concentration of 50 μM. **(F)** R13 increases Msi1 melting temperature in a concentration-dependent manner, providing evidence of their interaction in solution. **(G)** R13 competes with fluorescein-labeled RNA for Msi1 binding, as observed through fluorescence polarization.

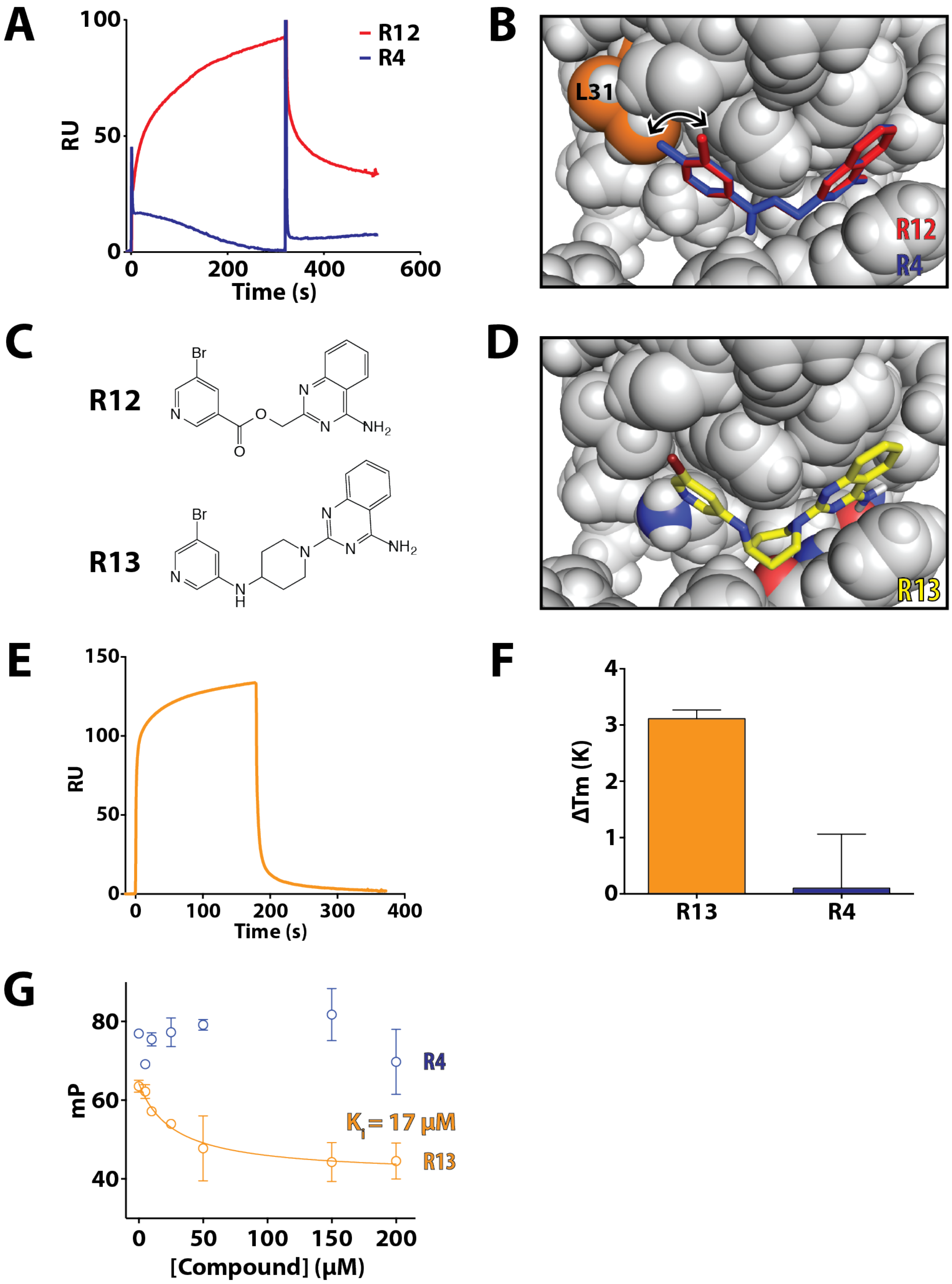

A
100
50
0
RU
0
200
400
600
Time (s)
R12
R4
B
L31
R12
R4
C
R12
Br
N
O
O
N
N
NH2
R13
Br
N
N
H
N
N
N
NH2
D
R13
E
150
100
50
0
RU
0
100
200
300
400
Time (s)
F
4
3
2
1
0
ΔTm (K)
R13
R4
G
80
60
40
mP
0
50
100
150
200
[Compound] (μM)
R4
K1 = 17 μM
R13

Figure 4.5: Initial SPR screening. Surface plasmon resonance was used to test for binding of all 12 compounds, with immobilized Msi1 RBD1. At a concentration of 50 μM, none of the compounds except for R12 showed a kinetic profile consistent with reversible binding. All sensorgrams shown have been reference-subtracted using a flow cell with an unrelated protein immobilized (human Mcl-1).

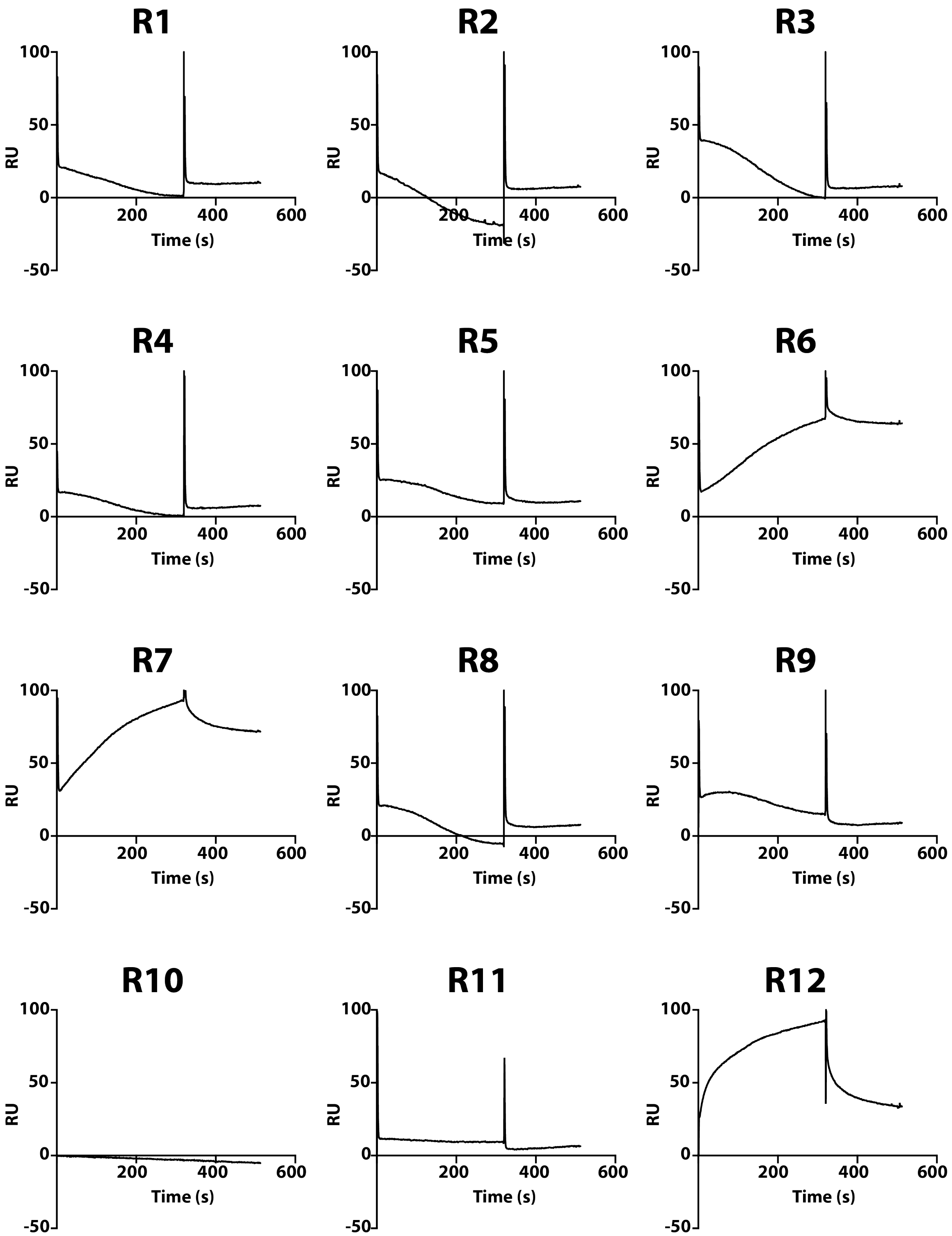
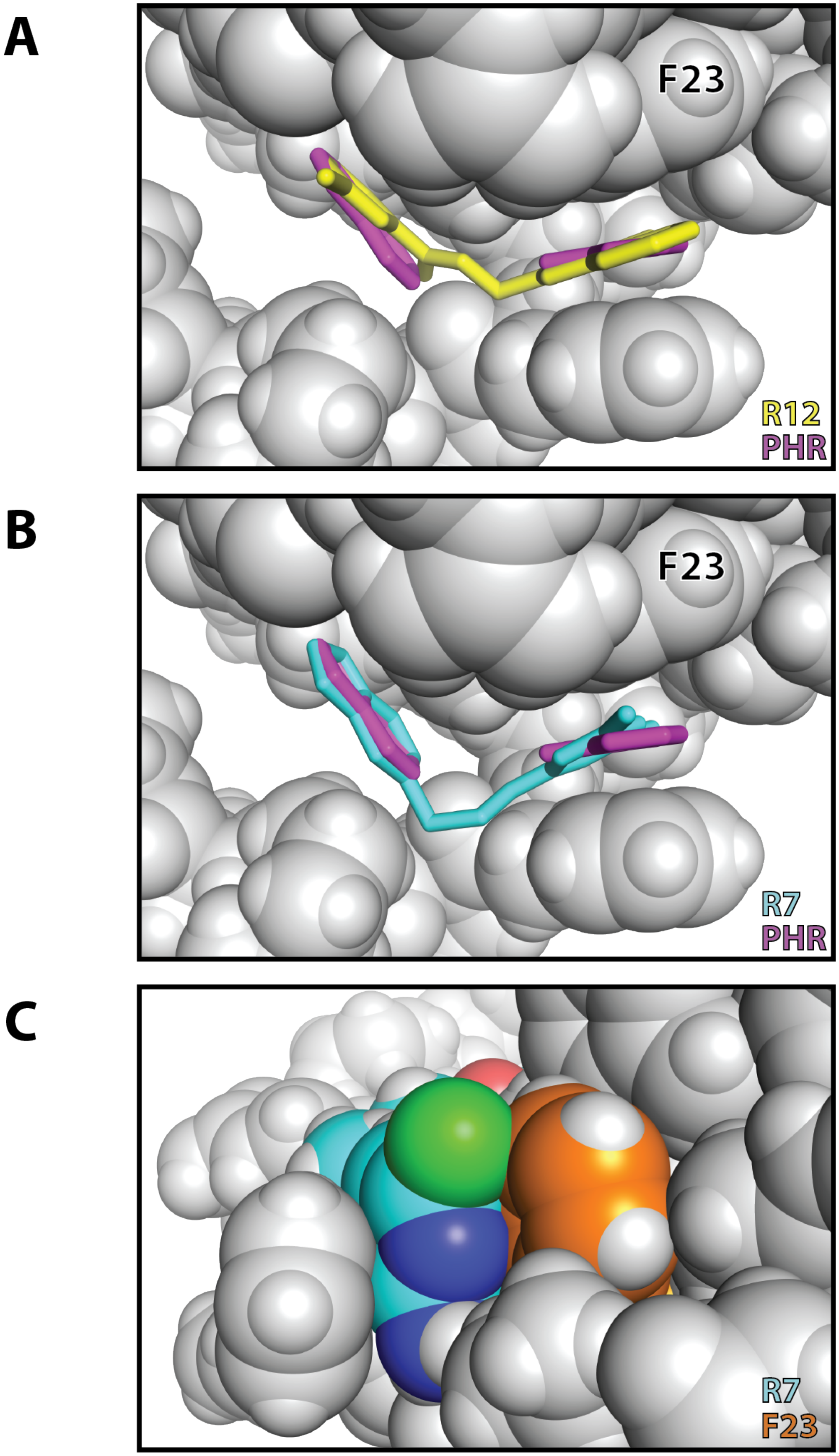

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
100
50
0
-50
RU
200
400
600
Time (s)

Figure 4.6: 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*).


A
F23
R12
PHR
B
F23
R7
PHR
C
R7
F23

Optimization of R12

Guided by this model, we next set out to improve the potency of R12. The limited chemical space in our screening library led to two undesirable features of this compound. First, the carbonyl oxygen in the ester linker of R12 is positioned in close proximity to the Phe96 backbone carbonyl of Msi1 (**Figure 4.2F**); beyond simply the lost opportunity for an intermolecular hydrogen bond, we expect electrostatic repulsion between these two negatively charged moieties. Second, the two ring systems of R12 are connected by a somewhat flexible linker; rigidifying this linker might reduce the conformational entropy lost upon binding. With these two motivations in mind, we designed and synthesized a panel of five new R12 derivatives, R13-R17 (**Figure 4.1**) (see *Materials and Methods*).

Using our previous SPR assay, we find that all five derivatives exhibit kinetic profiles at 50 μM consistent with binding to Msi1 (**Figure 4.4E, Figure 4.7**). Below we will present further biochemical characterization of R13, a compound that met our design goals by replacing R12's ester with a piperidine ring and a secondary amine in the linker (i.e. homopiperazine) (**Figure 4.4C**). Upon building models for each of the five R12 derivatives in complex with Msi1, we find that the rigidified linker in each compound allows recapitulation of R12's interactions while relieving the potential source of electrostatic repulsion; unfortunately however, none of the models include an additional hydrogen bond to Msi1 (**Figure 4.4D, Figure 4.8**).

In order to confirm binding of R13 to Msi1 in an orthogonal assay, we used differential scanning fluorimetry (DSF / Thermofluor) to determine protein thermostability as a function of ligand concentration. We find that the melting temperature of Msi1 increases upon addition of R13 in a dose-dependent manner (**Figure 4.4F**), up to a 3 ˚C increase in presence of 30 μM R13.

Finally, we directly examined the ability of R13 to not simply bind Msi1, but also to inhibit its interaction with *NUMB* mRNA. We used a fluorescein-labeled RNA oligonucleotide corresponding to the Msi1 recognition sequence of *NUMB*, which exhibits an increase in polarization upon Msi1 binding (**Figure 4.9**). The subsequent addition of R13 is expected to lead to a decrease in polarization, if R13 competes with *NUMB* mRNA for Msi1 binding as designed. Indeed we observe this dose-dependent behavior (**Figure 4.4G**), and using the apparent binding affinity of the labeled RNA for Msi1 we estimate the Ki for the R13-Msi1 interaction to be 17 μM.

Figure 4.7: SPR screening of R12 analogs. Surface plasmon resonance was used to test for binding of all five R12 analogs, with immobilized Msi1 RBD1. At a concentration of 50 μM, all five showed a kinetic profile consistent with reversible binding. All sensorgrams shown have been reference-subtracted using a flow cell with an unrelated protein immobilized (human Mcl-1).

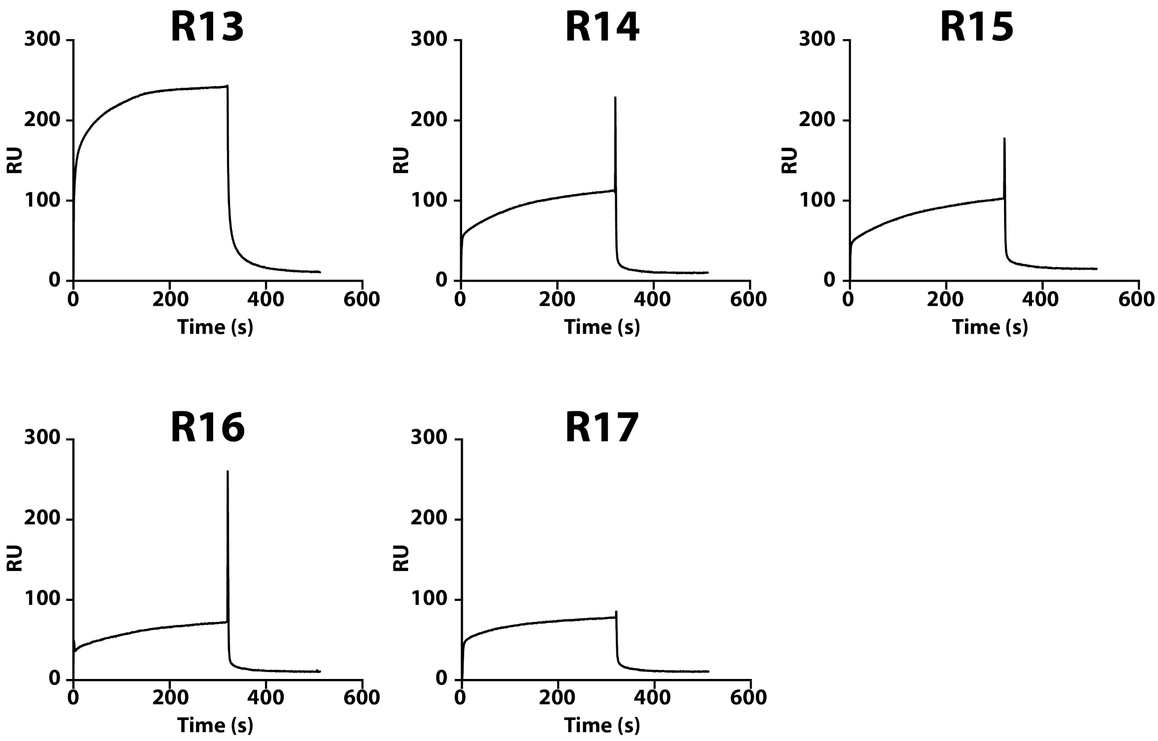
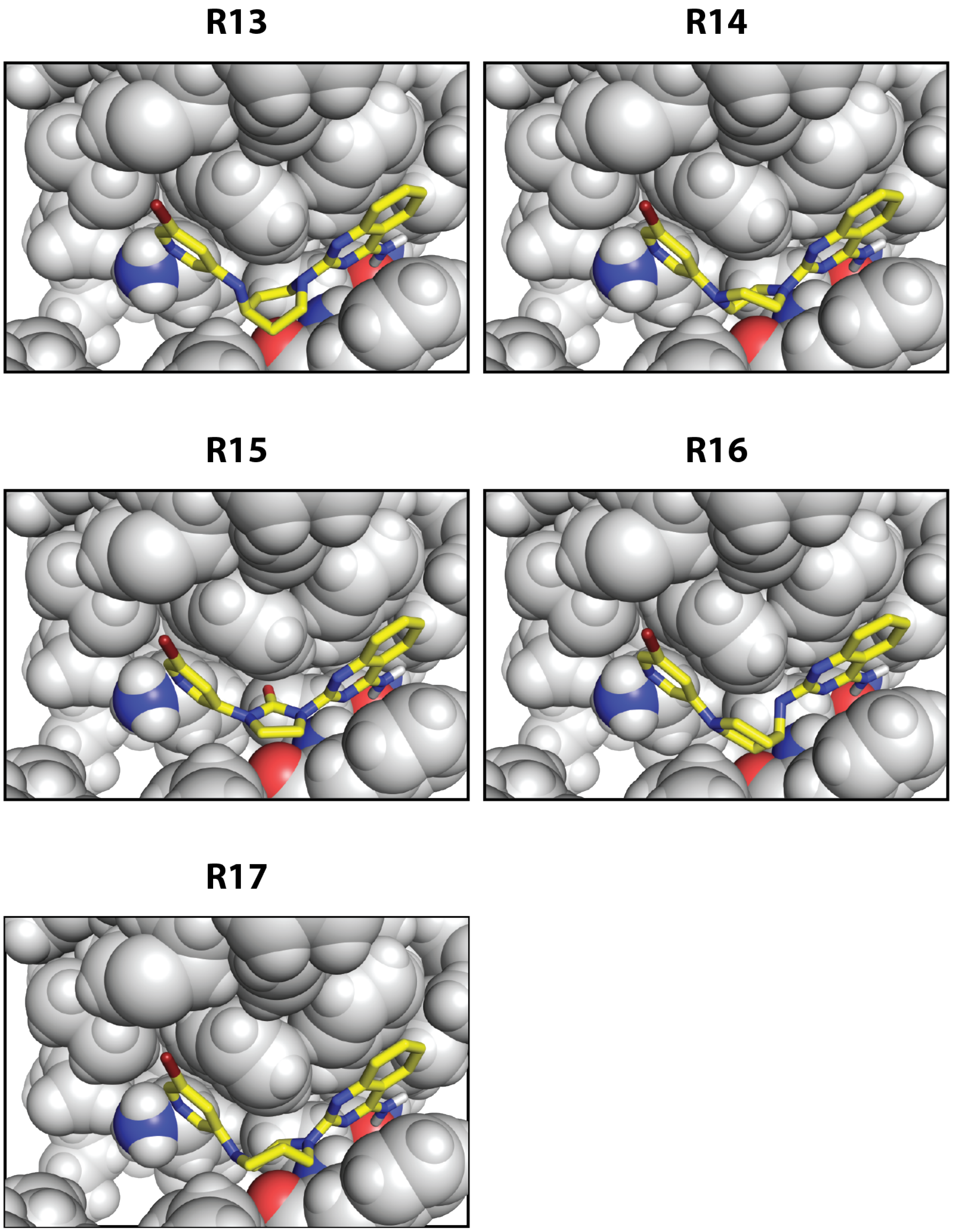


Figure 4.8: Models of R12 analogs in complex with Msi1. Structures were generated by building conformations of each compound that closely superpose with the rings of R12, then aligning this conformation to the hotspot pharmacophore.



R13

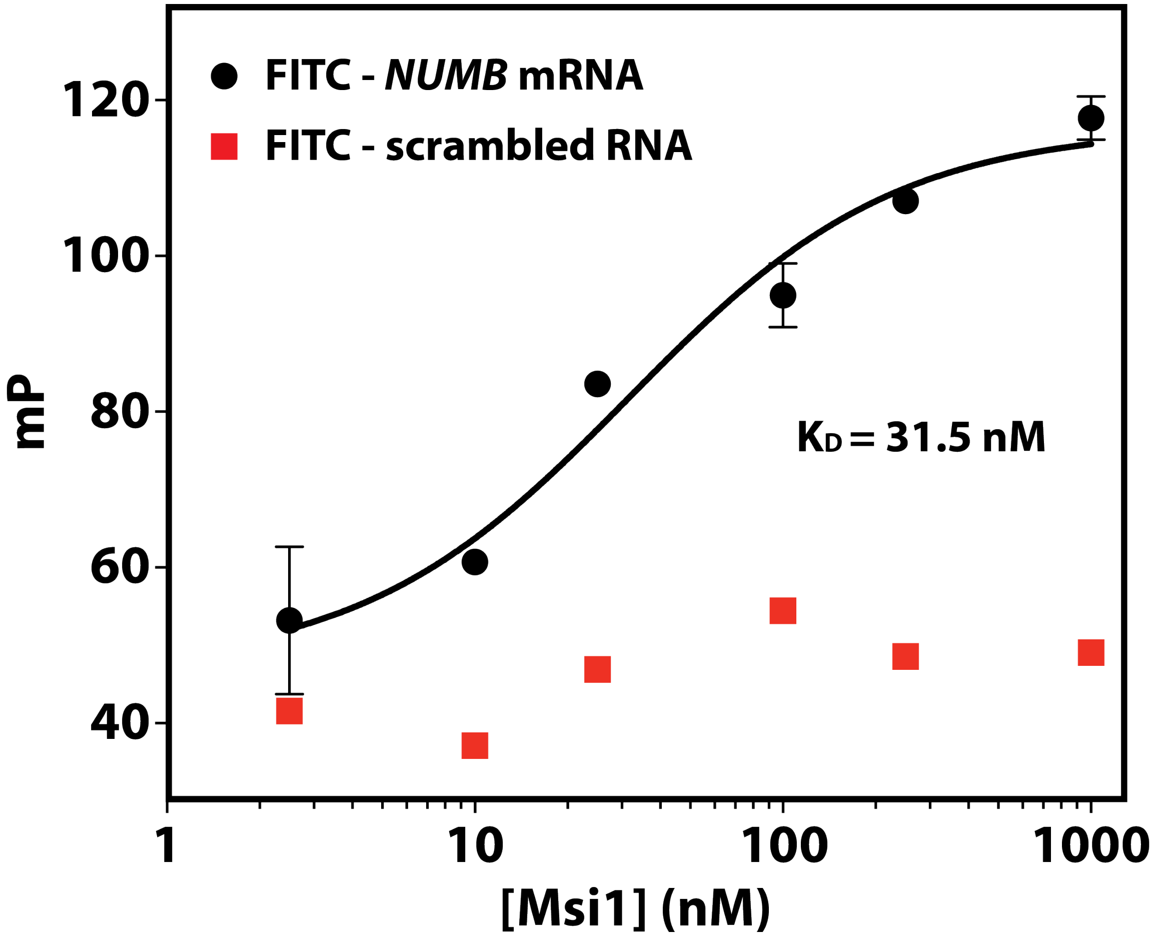
R14

R15

R16

R17

Figure 4.9: Binding of fluorescein-labeled RNA to Msi1. The polarization from the fluorescein tag increases upon addition of Msi1 RBD1, implying that the protein binds to the RNA. The analogous polarization change is not observed when Msi1 RBD1 is added to a mixture of fluorescein-labeled RNAs with random sequences (**Table 4.1 in Materials and Methods**).



Predicting target selectivity

Many RRM proteins recognize their target RNAs with high sequence specificity, through additional interactions outside the central RNA dinucleotide (*176*). 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. Staring with all the protein-RNA complexes deposited in Protein Data Bank, we used our computational approach to extract the set of all hotspot pharmacophores in the PDB (see *Materials and Methods*). For a given compound of interest, we can then screen all conformers of this molecule against this "library" of hotspot pharmacophores (**Figure 4.10A**). The top-scoring hits in this experiment represent proteins that recognize their cognate RNAs through interaction patterns that can be recapitulated by the compound of interest, making these candidate proteins for off-target binding. We note that this large-scale experiment does not explicitly account for protein flexibility, which may enable further of off-target interactions. To demonstrate the variation in pharmacophore structure associated with typical protein fluctuations, we have included in our studies each member of the experimentally-derived Msi1-RNA NMR ensemble (*181*).

We applied this analysis first to a series of hypothetical compounds, each one comprised of guanine and adenine attached by flexible linkers of varying lengths (**Figure 4.10B**). We find that each of these constructs can adopt a conformation that aligns well to the Msi1 hotspot pharmacophore, but they also undergo rearrangements that allow them to match many of the other hotspot pharmacophores in our library. This observation is unsurprising, since one would expect these artificial ligands to mimic many guanine-adenine dinucleotide pairs with little consideration of their three-dimensional arrangement.

We next carried out the same analysis for R12 (**Figure 4.10C**) and each of the other hits from our initial computational screen (**Figure 4.11A**). Relative to the guanine-adenine pairs, R12 lacks certain polar groups (those that do not participate in the Msi1 pharmacophore). While this reduces R12's potential for mimicking some of the off-target hotspot pharmacophores, we nonetheless find one alternate match with score comparable to that of Msi1, which is derived from a KH domain from the bacterial methyltransferase RsmE (**Figure 4.10C**, *blue arrow*). Notably, this off-target hotspot pharmacophore is recognized *not* because they bear strong resemblance to the Msi1 pharmacophore; rather, flexibility in the R12 linker allows it to match these alternate pharmacophores by adopting a different conformation (**Figure 4.10C**, *blue boxes*).

Finally, we applied this analysis to R13 (**Figure 4.10D**) and each of the other R12 derivatives (**Figure 4.11B**); we find that these match the Msi1 hotspot pharmacophore far better than any of the others extracted from the PDB. Constraint by the more rigid linker, R13 demonstrates much worse matching to the potential off-target pharmacophore for R12. This result suggests that the specific three-dimensional structure of the Msi1 hotspot pharmacophore, which is derived from the nucleobases of the buried adenine-guanine dinucleotides, may be sufficient to deliver selectivity into the matching compounds. To illustrate this point, we further examined another hotspot pharmacophore that is derived from an adenine-guanine pair in complex with a different RRM domain (**Figure 4.12**). Though the general domain topology and the binding site of the dinucleotides are similar in both RRMs, the detailed interaction of these dinucleotides is drastically different

(**Figure 4.12A**). Inheriting the uniqueness of Msi1 hotspot pharmacophore, R13 exhibits low structural similarity to the other adenine-guanine derived pharmacophore, which results in a potential steric clash to the binding site when transferred to the binding site of the other RRM domain (**Figure 4.12B**).

While this experiment does not account for flexibility of any of the off-target proteins, we do note that R13 matches some of the other members of the Msi1 NMR ensemble better than the single structure that led to its design (*red bar*). This highlights the robustness of hotspot pharmacophore matching to small changes in protein structure, and suggests that protein flexibility is unlikely to lead to alternate hotspot pharmacophores that are preferred by R13. While further experimental evidence will be necessary to explicitly determine whether these compounds engage in unanticipated interactions with other RBPs, these results suggest that the increased rigidity of the R13 linker makes it unable to access the alternate conformations that might allow R12 to mimic certain off-target pharmacophores.

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

Figure 4.10: Computational prediction of potential off-target interactions. (A) We screened each conformer of a given ligand against a large set of hotspot pharmacophores derived from other protein-RNA complexes. Hits in this screen correspond to other proteins that recognize their cognate RNAs using interactions that could be mimicked by the compound of interest. **(B)** Application of this approach to a series of hypothetical compounds built by connecting adenine and guanine with various linkers. The distribution of scores for the complete pharmacophore library is shown, with the score of the Msi1 pharmacophore indicated (*red arrow*). These artificial compounds match many other pharmacophores better than they match the Msi1 pharmacophore. **(C)** Application of this approach to R12. One pharmacophore from the library have scores comparable to that of Msi1; this match is accessed through the conformational flexibilty of R12 (*blue arrows/boxes*). **(D)** Application of this approach to R13. No pharmacophores from the library have scores comparable to that of Msi1; the scores of the off-target matches to R12 are now significantly worse (*blue arrows*), since R13 can no longer access these alternate conformations. In all cases the *red arrow* indicates the score of the hotspot pharmacophore derived from the first model in the Msi1/RNA NMR ensemble, and the *red bar* indicates the range of scores spanned by pharmacophores extracted from the other members of this NMR ensemble.

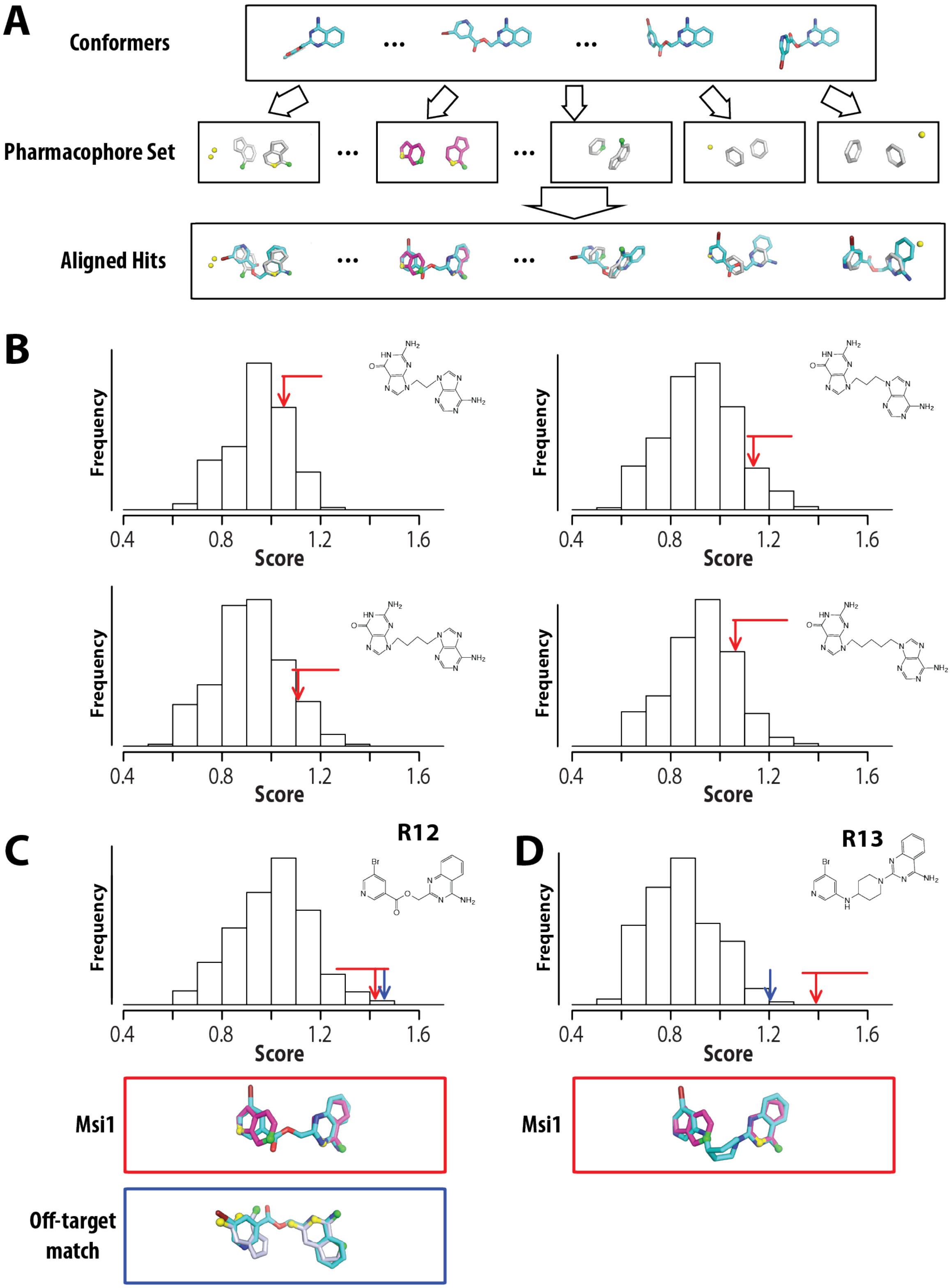

A
Conformers
Pharmacophore Set
Aligned Hits
B
Frequency
Score
0.4
0.8
1.2
1.6
C
R12
D
R13
Msi1
Off-target match

Figure 4.11: Computational prediction of potential off-target interactions. We screened each conformer of a given ligand against a large set of hotspot pharmacophores derived from other protein-RNA complexes, as described in **Figure 4.10**. **(A)** Compounds from our initial computational screen. **(B)** Derivatives of R12.

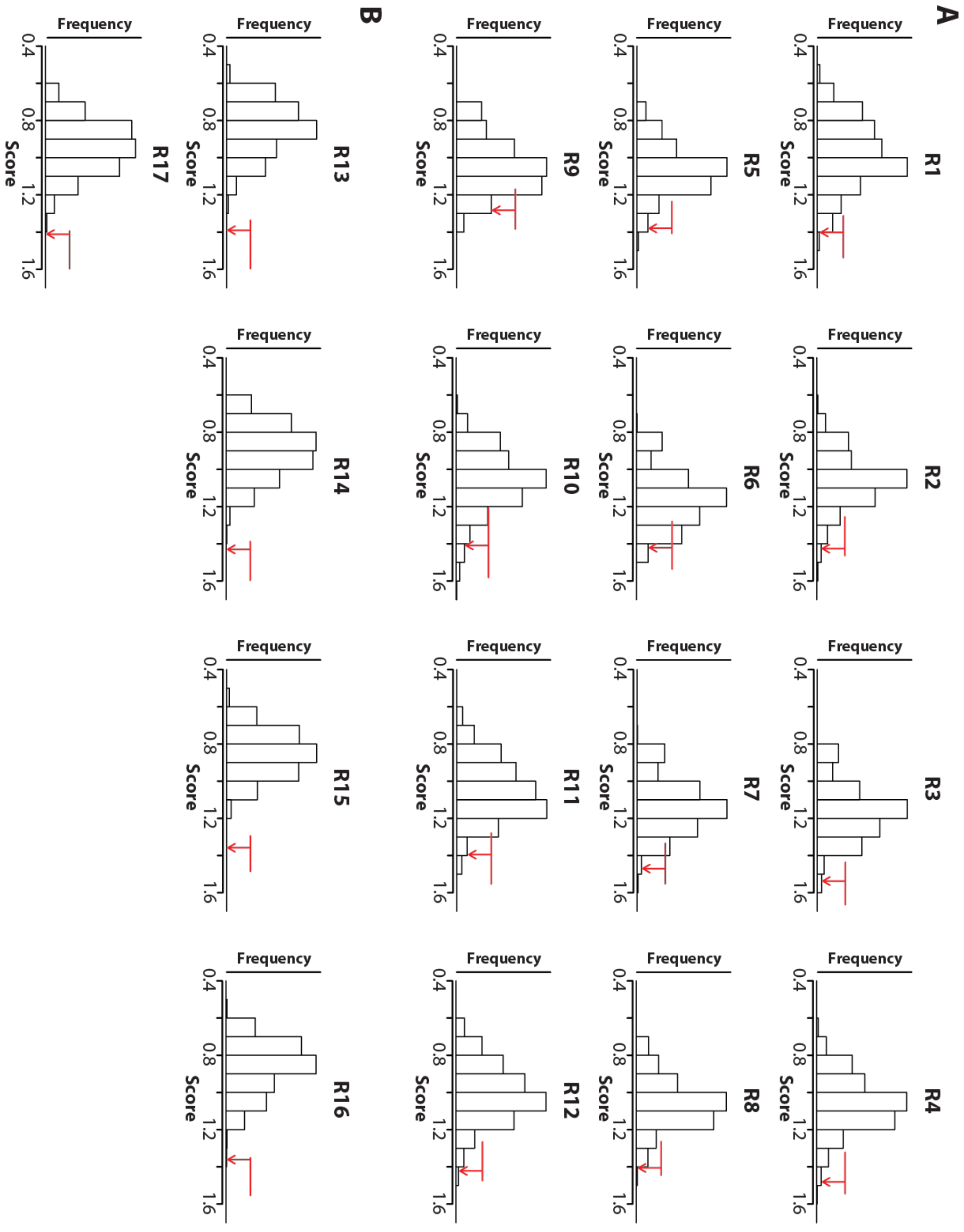
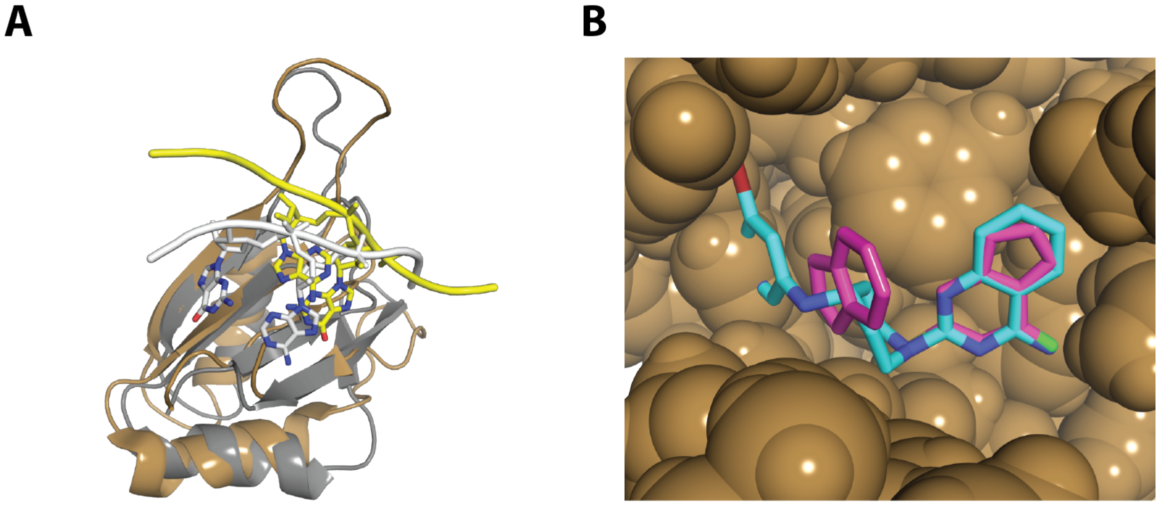

A
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
B
R13
R14
R15
R16
R17
Frequency
Score
0.4
0.8
1.2
1.6

Figure 4.12: Detailed geometry of hotspot pharmacophore delivers selectivity. We identified another adenine-guanine derived pharmacophore from another RRM domain (Tra2-β1, PDB ID: 2KXN), and examined the similarity between this alternative pharmacophore and R13. **(A)** The superposition of Msi1 RBD1 and the RRM domain from Tra2-β1. Protein and adenine-guanine nucleobases from Msi1 complex are shown in different shades of gray, while they are in different shades of yellow for Tra2-β1 complex. **(B)** The superposition of R13 (cyan stick) to pharmacophore of Tra2-β1 (magenta stick) demonstrates the lack of similarity between them. An obvious steric clash is revealed when transferring the aligned R13 to the binding site of Tra2-β1.



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 (*183, 184*); here, we instead encode a key RNA epitope on a small-molecule scaffold. We demonstrate the application of our approach using Musashi-1, leading to a novel class of inhibitors that disrupt the RNA-binding activity of this tumor-promoting protein. By using the hotspot pharmacophore as a template for ligand-based screening, our approach circumvents the challenge of explicitly designing *de novo* interactions against a relatively flat and polar protein surface.

The major advantages of this mimicry approach are its generality and simplicity. In this first application of the RNA mimicry approach, we elected to restrict our initial screening to commercially available compounds. Though convenient, none of the resulting hit compounds provided complete recapitulation of the desired hotspot interactions. Of three compounds that each matched three of the pharmacophore's polar groups, only one compound (R12) complemented the protein surface without steric clashes. In light of the fact that this compound provided a starting point for new inhibitors of Msi1, and thus validated the computational method, in future it will be worthwhile to explore chemical space more extensively in search of hits that more effectively mimic the desired hotspot pharmacophore. A computational screening platform was recently described that uses multi-component reaction chemistry (*185*) to build a virtual library containing tens of millions of novel compounds that can be readily accessed through proven “one-step, one-pot” reactions (*42*). While this strategy was originally used to construct a library of compounds that resemble collections of amino acid sidechains, it can be adapted to include privileged moieties that mimic patterns of hydrogen bond donors and acceptors in protein-RNA complexes, connected with rigid chemical linkers. By expanding the space of available compounds through this combinatorial strategy, and integrating computational screening with chemical synthesis, we envision discovery of compounds that more accurately match the target hotspot pharmacophores and thus exhibit improved potency prior to optimization.

The design of R12-derived compounds active against the Msi1 / *NUMB* mRNA interaction highlights the simplicity and robustness of the “hotspot mimicry” method, and also validates the “druggability” of this protein surface. We expect that the generality of this design strategy will allow it to be applied broadly in future, to develop inhibitors of RNA-binding proteins as chemical probes and potentially as starting points for developing unique new therapeutic agents.

Chapter V.
Discussion and Future Steps

The overall objective of this thesis is to evaluate and develop structure-based strategies to modulate protein function. Within this goal, I have demonstrated the general designability of “chemical rescue of structure”, a design strategy to achieve functional activation developed by our laboratory. Furthermore, I explored the feasibility of extending this approach to a more general setting. Towards the inhibition of protein function, I contributed to the development of a “hotspot mimicry” strategy to target RNA-binding proteins, and discovered novel compounds targeting the Msi1-1 / *NUMB* mRNA complex. The findings from these projects exhibit successful examples for both chemical rescue and hotspot mimicry approaches, demonstrate their general applicability, and provide valuable guidelines for the future applications of these approaches.

Underlying Mechanism of Chemical Rescue of Structure and Hotspot Mimicry

The fundamental principle underlying both chemical rescue and hotspot mimicry approaches is the exploitation of energetic contributions by buried hydrophobic moieties in protein structures and macromolecular interactions. From a thermodynamic standpoint, the desolvation of hydrophobic chemical groups, i.e. the hydrophobic effect, is the dominant driving force in protein folding and interactions (*186*). So structural manipulations involving these hydrophobic elements naturally become the starting points for developing rational approaches to modulate protein function.

Chemical rescue of structure deactivates protein function by deleting the sidechain of a buried tryptophan residue, which creates a cavity in the hydrophobic core of protein structure. This removal of hydrophobic moieties decreases the free energy of protein folding and leads to a wide variety of structural consequences, ranging from discrete conformational changes to local/global unfolding. The hydrophobic packing of exogenous indole compensates for the loss of hydrophobicity and restores the native protein structure. Analogous processes are also frequently observed in nature: hydrophobic packing upon the binding of small molecule also leads to a similar range of mechanisms to relay ligand binding to protein function in naturally-occurring systems (*81*).

The hotspot mimicry approach demonstrates and benefits from the deeply buried residues, i.e. hotspots, in protein-RNA interactions. The similar strategy has been extensively studied in protein-protein interactions. Seminal work by Jim Wells' group demonstrated the uneven distribution of binding free energies in protein-protein interface via alanine scanning (*38, 187*), and proposed the idea of mimicking the small cluster of key residues to design small molecule inhibitors (*11, 41*). The probing of protein-RNA interface via abasic RNA oligos is analogous to the usage of alanine scanning in protein-protein interactions, and its finding directly establishes the major contribution from the hydrophobic effects of buried nucleobases to the free energy of binding. This is the first definitive example proving the existence of hotspot residues on RNA molecules. The successful usage of these nucleobases as pharmacophore in ligand-based screening further supports the druggability of RNA-binding protein surface.

Generality of Chemical Rescue of Structure and Hotspot Mimicry

The design strategies of both chemical rescue of structure and hotspot mimicry rely on straightforward usages of hydrophobic effect, and therefore possess excellent applicability to other protein scaffolds.

Chemical rescue of structure can be achieved indirectly by the control of protein stability via small molecule binding. This observation immediately expands the applicability of the approach: future applications only involve the identification of cavity-forming mutations that inactivate the protein; the detailed mechanism, however, need not be explicitly considered. We have already demonstrated the additional application of chemical rescue of structure to GFP and some essential proteins in *E.coli*, and expect more future constructions of protein switches via this approach.

Chemical rescue of structure demonstrates a wide range of activation mechanisms similar to naturally occurring systems. In nature, specific mechanism possesses distinct advantages to satisfy different functional requirements, such as kinetics, sensitivity, and dynamic range (*133, 138, 188*). This observation suggests that it is possible to engineer desired mechanism into protein scaffolds through careful evaluation of protein structure and stability, which will in turn confer the associated advantages that meet to the unique criteria presented in specific biological applications.

Chemical rescue of structure is not limited solely to tryptophan-to-glycine mutation and indole rescue. The exploration with ChxR illustrates the feasibility of mutating a constellation of atoms from multiple buried residues and rescuing the function by more complex small molecules mimicking the three-dimensional structure of the missing structural elements. This extension further expands the versatility of the approach: it allows the application to, in principle, all protein hosts, even the ones lacking buried tryptophan residues; it also enables the construction of selective sensor proteins to a given small molecule by carefully selecting mutations to match the 3D structure of that compound (discussed more in the following sections). In addition, the increased hydrophobicity of these larger small molecules provides more favorable binding affinity and, when applying to living cells, likely improves the membrane permeability.

The hotspot mimicry approach can, in principle, be applied to any protein-RNA interaction that utilizes base stacking. This hydrophobic stacking between nucelobases and aromatic amino acids is a predominant mechanism used by the majority of naturally occurring RNA-binding domains, including RNA recognition motif (RRM), Zinc finger and KH domain (*177*). I applied the hotspot mimicry approach to a RRM domain, as it is one of the most abundant protein domains in eukaryotes and associated with many important cellular processes (*176*). The hotspot mimicry approach sifted out two buried nucleobases, whose major contribution to binding free energy was experimentally confirmed. The interaction pattern involving a pair of adjacent nucleobases and three outward-facing aromatic amino acids is a common feature to all RRMs, but the precise geometry of the dinucleotide in complex with the RRM can differ very drastically across members of this family. This finding not only corroborates the robustness of our computational framework, but also demonstrates the advantages of applying hotspot mimicry approach to RRMs: we can effectively identify small molecule inhibitors for other members of RRMs via a similar way, and these compounds will most likely be selective to the intended targets, as they mimic the 3D geometry of the dinucleotides, not merely the 2D chemical properties. Via homology modeling, the conserved interaction pattern of RRMs can be further utilized to generalize the hotspot mimicry approach to RRMs that lack the holo-structure with cognate RNAs.

Future steps of activation via chemical rescue

The possibility of employing chemical rescue of structure using multiple mutations can be utilized in the converse manner to achieve the design of selective sensor proteins to certain small molecules, especially the ones for which no naturally-occurring protein binding partner is known. To discover protein structures suitable of harboring the small-molecule binding site, we can perform the chemical rescue of structure in reverse. The procedure starts by partitioning the structure of a given small molecule into in substructures corresponding to the deleted chemical moieties from any “large-to-small” single point mutation. If such partition is feasible, we can examine all protein structures to identify protein hosts that have the corresponding “large” amino acids in close proximity, and build a library containing these constellations of the deleted chemical moieties from all suitable proteins. This library can then be screened using the small molecule of interest as the template for constellations that mimic the structural geometry and chemical property of the small molecule. The corresponding protein host can then be mutated to harbor a *de novo* binding site specific for that small molecule. In order to achieve sensing activity, we can either solely use catalytic domains in the screening process, or fuse the designed domain with a catalytic domain and resort to conventional design strategies, such as domain insertion, to achieve the coupling between small molecule binding and catalytic activity.

Future steps of inhibition via hotspot mimicry

The computational examination of selectivity demonstrates that small molecule inhibitors identified via hotspot mimicry approach are generally specific to the intended target. The procedure, however, did identify potential off-target proteins. It will greatly corroborate the robustness of the computational procedure if we can experimentally demonstrate that the predicted off-target proteins can indeed bind to the corresponding small molecule. The ability of reliably predicting the selectivity of a hotspot pharmacophore will enable the discovery of most unique hotspot pharmacophore or a group of isolated but closely related pharmacophores: we can cluster the hotspot pharmacophores from available structures of disease-related RBPs basing on 3D similarity, and identify the most isolated pharmacophore, or a isolated cluster of pharmacophores. The former study enables the development of highly selective drugs by targeting the most unique hotspot pharmacophore; the latter help the identification of small molecules that concomitantly inhibit multiple protein targets.

One limitation of the examination of selectivity is that it only considers the static structure of protein-RNA complexes, which may result in an overestimation of the selectivity. The inclusion of twenty Msi-1 pharmacophores obtained from NMR structure suggests that the selectivity of a hotspot pharmacophore may not be sensitivity to the dynamic range observed in NMR. However, thorough studies are required to draw more definitive conclusions. Computational tools have been recently developed and are available in Rosetta software suite to allow effective modeling of the dynamic movement in protein-RNA complexes, and enable the generation of dynamic ensemble of protein-RNA complexes from a static structure. Including these dynamic structures in the selectivity examination provides a more rigorous evaluation of potential off-target binders.

In the application to Msi-1, I am aware of the low hit rate and the relative weak bindings of the hit compounds, and we can improve the methodology via various strategies. As the first test, we elected to restrict the virtual screening to commercially available compounds in the "Drugs-Now" database. Though convenient, none of the resulting compounds completely recapitulates the intended hotspot interactions. Furthermore, molecules in the "Drugs-Now" database are selected following the Lipinski's "rule of five" criteria. It is shown, however, that the inhibitors for non-traditional targets may violate these criteria (*189*). These observations strongly suggest the need of a larger and focused small molecule library in future applications to enable a more extensive exploration of the chemical space in search of hits that can more effectively mimic that desired hotspot pharmacophore. One method of constructing such virtual libraries is to include the compounds obtainable from the multi-component reactions (*42*). While this strategy was originally used to construct a library of compounds that resemble collections of amino acid sidechains, it can be adapted to include privileged moieties that mimic patterns of hydrogen bond donors and acceptors in protein-RNA complexes, connected with rigid chemical linkers. By expanding the space of available compounds through this combinatorial strategy, and integrating computational screening with chemical synthesis, we envision discovery of compounds that more accurately match the target hotspot pharmacophores and thus exhibit improved potency prior to optimization.

Reference

1. G. Guntas, T. J. Mansell, J. R. Kim, M. Ostermeier, Directed evolution of protein switches and their application to the creation of ligand-binding proteins. *Proc Natl Acad Sci U S A* **102**, 11224-11229 (2005); published online EpubAug 9 (10.1073/pnas.0502673102).

2. B. J. Yeh, R. J. Rutigliano, A. Deb, D. Bar-Sagi, W. A. Lim, Rewiring cellular morphology pathways with synthetic guanine nucleotide exchange factors. *Nature* **447**, 596-600 (2007); published online EpubMay 31 (10.1038/nature05851).

3. B. Chen, M. E. Dodge, W. Tang, J. Lu, Z. Ma, C. W. Fan, S. Wei, W. Hao, J. Kilgore, N. S. Williams, M. G. Roth, J. F. Amatruda, C. Chen, L. Lum, Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer. *Nature chemical biology* **5**, 100-107 (2009); published online EpubFeb (10.1038/nchembio.137).

4. S. L. Schreiber, Small molecules: the missing link in the central dogma. *Nature chemical biology* **1**, 64-66 (2005); published online EpubJul (10.1038/nchembio0705-64).

5. L. A. Banaszynski, L. C. Chen, L. A. Maynard-Smith, A. G. Ooi, T. J. Wandless, A rapid, reversible, and tunable method to regulate protein function in living cells using synthetic small molecules. *Cell* **126**, 995-1004 (2006); published online EpubSep 8 (10.1016/j.cell.2006.07.025).

6. A. R. Buskirk, Y. C. Ong, Z. J. Gartner, D. R. Liu, Directed evolution of ligand dependence: small-molecule-activated protein splicing. *Proc Natl Acad Sci U S A* **101**, 10505-10510 (2004); published online EpubJul 20 (10.1073/pnas.0402762101).

7. D. M. Spencer, T. J. Wandless, S. L. Schreiber, G. R. Crabtree, Controlling signal transduction with synthetic ligands. *Science* **262**, 1019-1024 (1993); published online EpubNov 12 (

8. R. M. de Lorimier, J. J. Smith, M. A. Dwyer, L. L. Looger, K. M. Sali, C. D. Paavola, S. S. Rizk, S. Sadigov, D. W. Conrad, L. Loew, H. W. Hellinga, Construction of a fluorescent biosensor family. *Protein Sci* **11**, 2655-2675 (2002); published online EpubNov (10.1110/ps.021860).

9. H. W. Hellinga, J. S. Marvin, Protein engineering and the development of generic biosensors. *Trends Biotechnol* **16**, 183-189 (1998); published online EpubApr (

10. J. Hasty, D. McMillen, J. J. Collins, Engineered gene circuits. *Nature* **420**, 224-230 (2002); published online EpubNov 14 (10.1038/nature01257).

11. M. R. Arkin, J. A. Wells, Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature reviews. Drug discovery* **3**, 301-317 (2004); published online EpubApr (10.1038/nrd1343).

12. M. A. Fabian, W. H. Biggs, 3rd, D. K. Treiber, C. E. Atteridge, M. D. Azimioara, M. G. Benedetti, T. A. Carter, P. Ciceri, P. T. Edeen, M. Floyd, J. M. Ford, M. Galvin, J. L. Gerlach, R. M. Grotzfeld, S. Herrgard, D. E. Insko, M. A. Insko, A. G. Lai, J. M. Lelias, S. A. Mehta, Z. V. Milanov, A. M. Velasco, L. M. Wodicka, H. K. Patel, P. P. Zarrinkar, D. J. Lockhart, A small molecule-kinase interaction map for clinical kinase

inhibitors. *Nat Biotechnol* **23**, 329-336 (2005); published online EpubMar (10.1038/nbt1068).

13. J. Zhang, P. L. Yang, N. S. Gray, Targeting cancer with small molecule kinase inhibitors. *Nature reviews. Cancer* **9**, 28-39 (2009); published online EpubJan (10.1038/nrc2559).

14. G. Guntas, S. F. Mitchell, M. Ostermeier, A molecular switch created by in vitro recombination of nonhomologous genes. *Chem Biol* **11**, 1483-1487 (2004); published online EpubNov (10.1016/j.chembiol.2004.08.020).

15. A. V. Karginov, F. Ding, P. Kota, N. V. Dokholyan, K. M. Hahn, Engineered allosteric activation of kinases in living cells. *Nat Biotechnol* **28**, 743-747 (2010); published online EpubJul (10.1038/nbt.1639).

16. C. L. Tucker, S. Fields, A yeast sensor of ligand binding. *Nat Biotechnol* **19**, 1042-1046 (2001); published online EpubNov (10.1038/nbt1101-1042).

17. J. Kirsch, C. Siltanen, Q. Zhou, A. Revzin, A. Simonian, Biosensor technology: recent advances in threat agent detection and medicine. *Chemical Society reviews* **42**, 8733-8768 (2013); published online EpubNov 21 (10.1039/c3cs60141b).

18. O. V. Makhlynets, I. V. Korendovych, Design of catalytically amplified sensors for small molecules. *Biomolecules* **4**, 402-418 (2014)10.3390/biom4020402).

19. M. Ostermeier, Engineering allosteric protein switches by domain insertion. *Protein Eng Des Sel* **18**, 359-364 (2005); published online EpubAug (10.1093/protein/gzi048).

20. M. Ostermeier, Designing switchable enzymes. *Curr Opin Struct Biol* **19**, 442-448 (2009); published online EpubAug (10.1016/j.sbi.2009.04.007).

21. S. B. Kim, Y. Otani, Y. Umezawa, H. Tao, Bioluminescent indicator for determining protein-protein interactions using intramolecular complementation of split click beetle luciferase. *Anal Chem* **79**, 4820-4826 (2007); published online EpubJul 1 (10.1021/ac0621571).

22. I. V. Korendovych, D. W. Kulp, Y. Wu, H. Cheng, H. Roder, W. F. Degrado, Design of a switchable eliminase. *Proc Natl Acad Sci U S A* **108**, 6823-6827 (2011); published online EpubApr 26 (1018191108 [pii]
10.1073/pnas.1018191108).

23. M. D. Toney, J. F. Kirsch, Direct Bronsted analysis of the restoration of activity to a mutant enzyme by exogenous amines. *Science* **243**, 1485-1488 (1989); published online EpubMar 17 (

24. Y. Qiao, H. Molina, A. Pandey, J. Zhang, P. A. Cole, Chemical rescue of a mutant enzyme in living cells. *Science* **311**, 1293-1297 (2006); published online EpubMar 3 (10.1126/science.1122224).

25. B. W. Matthews, L. Liu, A review about nothing: are apolar cavities in proteins really empty? *Protein Sci* **18**, 494-502 (2009); published online EpubMar (10.1002/pro.61).

26. K. Deckert, S. J. Budiardjo, L. C. Brunner, S. Lovell, J. Karanicolas, Designing allosteric control into enzymes by chemical rescue of structure. *J Am Chem Soc* **134**, 10055-10060 (2012); published online EpubJun 20 (10.1021/ja301409g).

27. Y. Xia, N. DiPrimio, T. R. Keppel, B. Vo, K. Fraser, K. P. Battaile, C. Egan, C. Bystroff, S. Lovell, D. D. Weis, J. C. Anderson, J. Karanicolas, The designability of protein switches by chemical rescue of structure: mechanisms of inactivation and

reactivation. *J Am Chem Soc* **135**, 18840-18849 (2013); published online EpubDec 18 (10.1021/ja407644b).

28. C. M. Crews, J. B. Shotwell, Small-molecule inhibitors of the cell cycle: an overview. *Progress in cell cycle research* **5**, 125-133 (2003).

29. A. J. Firestone, J. S. Weinger, M. Maldonado, K. Barlan, L. D. Langston, M. O'Donnell, V. I. Gelfand, T. M. Kapoor, J. K. Chen, Small-molecule inhibitors of the AAA+ ATPase motor cytoplasmic dynein. *Nature* **484**, 125-129 (2012); published online EpubApr 5 (10.1038/nature10936).

30. S. Shangary, S. Wang, Small-molecule inhibitors of the MDM2-p53 protein-protein interaction to reactivate p53 function: a novel approach for cancer therapy. *Annual review of pharmacology and toxicology* **49**, 223-241 (2009)10.1146/annurev.pharmtox.48.113006.094723).

31. R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, G. S. Sittampalam, Impact of high-throughput screening in biomedical research. *Nature reviews. Drug discovery* **10**, 188-195 (2011); published online EpubMar (10.1038/nrd3368).

32. K. P. Mishra, L. Ganju, M. Sairam, P. K. Banerjee, R. C. Sawhney, A review of high throughput technology for the screening of natural products. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* **62**, 94-98 (2008); published online EpubFeb (10.1016/j.biopha.2007.06.012).

33. R. Gowthaman, E. J. Deeds, J. Karanicolas, Structural properties of non-traditional drug targets present new challenges for virtual screening. *Journal of chemical*

information and modeling **53**, 2073-2081 (2013); published online EpubAug 26 (10.1021/ci4002316).

34. S. Mandal, M. Moudgil, S. K. Mandal, Rational drug design. *European journal of pharmacology* **625**, 90-100 (2009); published online EpubDec 25 (10.1016/j.ejphar.2009.06.065).

35. J. Muto, T. Imai, D. Ogawa, Y. Nishimoto, Y. Okada, Y. Mabuchi, T. Kawase, A. Iwanami, P. S. Mischel, H. Saya, K. Yoshida, Y. Matsuzaki, H. Okano, RNA-binding protein Musashi1 modulates glioma cell growth through the post-transcriptional regulation of Notch and PI3 kinase/Akt signaling pathways. *PloS one* **7**, e33431 (2012)10.1371/journal.pone.0033431).

36. J. Turkson, R. Jove, STAT proteins: novel molecular targets for cancer drug discovery. *Oncogene* **19**, 6613-6626 (2000); published online EpubDec 27 (10.1038/sj.onc.1204086).

37. E. Kellenberger, J. Rodrigo, P. Muller, D. Rognan, Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **57**, 225-242 (2004); published online EpubNov 1 (10.1002/prot.20149).

38. T. Clackson, J. A. Wells, A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383-386 (1995); published online EpubJan 20 (

39. I. S. Moreira, P. A. Fernandes, M. J. Ramos, Hot spots--a review of the protein-protein interface determinant amino-acid residues. *Proteins* **68**, 803-812 (2007); published online EpubSep 1 (10.1002/prot.21396).

40. D. Rajamani, S. Thiel, S. Vajda, C. J. Camacho, Anchor residues in protein-protein interactions. *Proc Natl Acad Sci U S A* **101**, 11287-11292 (2004); published online EpubAug 3 (10.1073/pnas.0401942101).

41. J. A. Wells, C. L. McClendon, Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **450**, 1001-1009 (2007); published online EpubDec 13 (10.1038/nature06526).

42. D. Koes, K. Khoury, Y. Huang, W. Wang, M. Bista, G. M. Popowicz, S. Wolf, T. A. Holak, A. Domling, C. J. Camacho, Enabling large-scale design, synthesis and validation of small molecule protein-protein antagonists. *PloS one* **7**, e32839 (2012)10.1371/journal.pone.0032839).

43. D. R. Koes, C. J. Camacho, Small-molecule inhibitor starting points learned from protein-protein interaction inhibitor structure. *Bioinformatics* **28**, 784-791 (2012); published online EpubMar 15 (10.1093/bioinformatics/btr717).

44. T. Kortemme, D. E. Kim, D. Baker, Computational alanine scanning of protein-protein interfaces. *Science's STKE : signal transduction knowledge environment* **2004**, pl2 (2004); published online EpubFeb 10 (10.1126/stke.2192004pl2).

45. Y. Bromberg, B. Rost, Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics* **24**, i207-212 (2008); published online EpubAug 15 (10.1093/bioinformatics/btn268).

46. C. D. Thanos, W. L. DeLano, J. A. Wells, Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc Natl Acad Sci U S A* **103**, 15422-15427 (2006); published online EpubOct 17 (10.1073/pnas.0607058103).

47. F. Christ, A. Voet, A. Marchand, S. Nicolet, B. A. Desimmie, D. Marchand, D. Bardiot, N. J. Van der Veken, B. Van Remoortel, S. V. Strelkov, M. De Maeyer, P. Chaltin, Z. Debyser, Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. *Nature chemical biology* **6**, 442-448 (2010); published online EpubJun (10.1038/nchembio.370).

48. S. Liu, S. Wu, S. Jiang, HIV entry inhibitors targeting gp41: from polypeptides to small-molecule compounds. *Current pharmaceutical design* **13**, 143-162 (2007).

49. R. J. Bandziulis, M. S. Swanson, G. Dreyfuss, RNA-binding proteins as developmental regulators. *Genes & development* **3**, 431-437 (1989); published online EpubApr (

50. D. Curtis, R. Lehmann, P. D. Zamore, Translational regulation in development. *Cell* **81**, 171-178 (1995); published online EpubApr 21 (

51. A. J. Matlin, F. Clark, C. W. Smith, Understanding alternative splicing: towards a cellular code. *Nature reviews. Molecular cell biology* **6**, 386-398 (2005); published online EpubMay (10.1038/nrm1645).

52. J. D. Richter, Cytoplasmic polyadenylation in development and beyond. *Microbiology and molecular biology reviews : MMBR* **63**, 446-456 (1999); published online EpubJun (

53. J. Guhaniyogi, G. Brewer, Regulation of mRNA stability in mammalian cells. *Gene* **265**, 11-23 (2001); published online EpubMar 7 (

54. J. E. Wilhelm, R. D. Vale, RNA on the move: the mRNA localization pathway. *The Journal of cell biology* **123**, 269-274 (1993); published online EpubOct (

55. I. Abaza, F. Gebauer, Trading translation with RNA-binding proteins. *Rna* **14**, 404-409 (2008); published online EpubMar (10.1261/rna.848208).

56. P. Ahlquist, RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science* **296**, 1270-1273 (2002); published online EpubMay 17 (10.1126/science.1069132).

57. W. M. Kati, K. A. Johnson, L. F. Jerva, K. S. Anderson, Mechanism and fidelity of HIV reverse transcriptase. *J Biol Chem* **267**, 25988-25997 (1992); published online EpubDec 25 (

58. W. J. Ma, S. Cheng, C. Campbell, A. Wright, H. Furneaux, Cloning and characterization of HuR, a ubiquitously expressed Elav-like protein. *J Biol Chem* **271**, 8144-8151 (1996); published online EpubApr 5 (

59. H. Okano, H. Kawahara, M. Toriya, K. Nakao, S. Shibata, T. Imai, Function of RNA-binding protein Musashi-1 in stem cells. *Experimental cell research* **306**, 349-356 (2005); published online EpubJun 10 (10.1016/j.yexcr.2005.02.021).

60. E. Spears, K. L. Neufeld, Novel double-negative feedback loop between adenomatous polyposis coli and Musashi1 in colon epithelia. *J Biol Chem* **286**, 4946-4950 (2011); published online EpubFeb 18 (10.1074/jbc.C110.205922).

61. K. E. Squires, An introduction to nucleoside and nucleotide analogues. *Antiviral therapy* **6 Suppl 3**, 1-14 (2001).

62. S. S. Carroll, D. B. Olsen, Nucleoside analog inhibitors of hepatitis C virus replication. *Infectious disorders drug targets* **6**, 17-29 (2006); published online EpubMar (

63. H. Mitsuya, R. Yarchoan, S. Broder, Molecular targets for AIDS therapy. *Science* **249**, 1533-1544 (1990); published online EpubSep 28 (

64. D. Sampath, V. A. Rao, W. Plunkett, Mechanisms of apoptosis induction by nucleoside analogs. *Oncogene* **22**, 9063-9074 (2003); published online EpubDec 8 (10.1038/sj.onc.1207229).

65. T. Robak, E. Lech-Maranda, A. Korycka, E. Robak, Purine nucleoside analogs as immunosuppressive and antineoplastic agents: mechanism of action and clinical activity. *Current medicinal chemistry* **13**, 3165-3189 (2006).

66. P. A. Jones, S. M. Taylor, Cellular differentiation, cytidine analogs and DNA methylation. *Cell* **20**, 85-93 (1980); published online EpubMay (

67. J. J. Kohler, W. Lewis, A brief overview of mechanisms of mitochondrial toxicity from NRTIs. *Environmental and molecular mutagenesis* **48**, 166-172 (2007); published online EpubApr-May (10.1002/em.20223).

68. S. Lutz, M. Ostermeier, G. L. Moore, C. D. Maranas, S. J. Benkovic, Creating multiple-crossover DNA libraries independent of sequence identity. *Proc Natl Acad Sci U S A* **98**, 11248-11253 (2001); published online EpubSep 25 (10.1073/pnas.201413698).

69. E. De Clercq, The role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. *Antiviral research* **38**, 153-179 (1998); published online EpubJun (

70. V. J. Merluzzi, K. D. Hargrave, M. Labadia, K. Grozinger, M. Skoog, J. C. Wu, C. K. Shih, K. Eckner, S. Hattox, J. Adams, et al., Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor. *Science* **250**, 1411-1413 (1990); published online EpubDec 7 (

71. J. Gallego, G. Varani, Targeting RNA with small-molecule drugs: therapeutic promise and chemical challenges. *Accounts of chemical research* **34**, 836-843 (2001); published online EpubOct (

72. A. C. Stelzer, A. T. Frank, J. D. Kratz, M. D. Swanson, M. J. Gonzalez-Hernandez, J. Lee, I. Andricioaei, D. M. Markovitz, H. M. Al-Hashimi, Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. *Nature chemical biology* **7**, 553-559 (2011); published online EpubAug (10.1038/nchembio.596).

73. S. Ostermeier, Evaluation of a barrier dental sealant in dogs. *J Vet Dent* **22**, 215; author reply 215 (2005); published online EpubDec (

74. V. I. Perez-Nueno, D. W. Ritchie, O. Rabal, R. Pascual, J. I. Borrell, J. Teixido, Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. *Journal of chemical information and modeling* **48**, 509-533 (2008); published online EpubMar (10.1021/ci700415g).

75. G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J. F. Truchon, W. D. Cornell, Comparison of topological, shape, and docking methods in virtual screening. *Journal of chemical information and modeling* **47**, 1504-1519 (2007); published online EpubJul-Aug (10.1021/ci700052x).

76. J. Kirchmair, S. Distinto, P. Markt, D. Schuster, G. M. Spitzer, K. R. Liedl, G. Wolber, How to optimize shape-based virtual screening: choosing the right query and including chemical information. *Journal of chemical information and modeling* **49**, 678-692 (2009); published online EpubMar (10.1021/ci8004226).

77. A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. Mandell, F. Richter, Y. E. Ban, S. J. Fleishman, J. E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popovic, J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, P. Bradley, ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology* **487**, 545-574 (2011)10.1016/B978-0-12-381270-4.00019-6).

78. A. R. Buskirk, A. Landrigan, D. R. Liu, Engineering a ligand-dependent RNA transcriptional activator. *Chem Biol* **11**, 1157-1163 (2004); published online EpubAug (10.1016/j.chembiol.2004.05.017).

79. M. R. Pratt, E. C. Schwartz, T. W. Muir, Small-molecule-mediated rescue of protein function by an inducible proteolytic shunt. *Proc Natl Acad Sci U S A* **104**, 11209-11214 (2007); published online EpubJul 3 (0700816104 [pii]
10.1073/pnas.0700816104).

80. J. Lee, M. Natarajan, V. C. Nashine, M. Socolich, T. Vo, W. P. Russ, S. J. Benkovic, R. Ranganathan, Surface sites for engineering allosteric control in proteins. *Science* **322**, 438-442 (2008); published online EpubOct 17 (10.1126/science.1159052).

81. J. H. Ha, S. N. Loh, Protein conformational switches: from nature to design. *Chemistry* **18**, 7984-7999 (2012); published online EpubJun 25 (10.1002/chem.201200348).

82. M. Kanwar, R. C. Wright, A. Date, J. Tullman, M. Ostermeier, Protein switch engineering by domain insertion. *Methods in enzymology* **523**, 369-388 (2013)10.1016/B978-0-12-394292-0.00017-5).

83. M. M. Stratton, S. N. Loh, Converting a protein into a switch for biosensing and functional regulation. *Protein Sci* **20**, 19-29 (2011); published online EpubJan (10.1002/pro.541).

84. J. E. Dueber, B. J. Yeh, K. Chak, W. A. Lim, Reprogramming control of an allosteric signaling switch through modular recombination. *Science* **301**, 1904-1908 (2003); published online EpubSep 26 (10.1126/science.1085945).

85. G. E. Meister, N. S. Joshi, An engineered calmodulin-based allosteric switch for Peptide biosensing. *Chembiochem : a European journal of chemical biology* **14**, 1460-1467 (2013); published online EpubAug 19 (10.1002/cbic.201300168).

86. O. Dagliyan, D. Shirvanyants, A. V. Karginov, F. Ding, L. Fee, S. N. Chandrasekaran, C. M. Freisinger, G. A. Smolen, A. Huttenlocher, K. M. Hahn, N. V. Dokholyan, Rational design of a ligand-controlled protein conformational switch. *Proc Natl Acad Sci U S A* **110**, 6800-6804 (2013); published online EpubApr 23 (10.1073/pnas.1218319110).

87. J. T. Kittleson, S. Cheung, J. C. Anderson, Rapid optimization of gene dosage in E. coli using DIAL strains. *J Biol Eng* **5**, 10 (2011)10.1186/1754-1611-5-10).

88. J. D. Pedelacq, S. Cabantous, T. Tran, T. C. Terwilliger, G. S. Waldo, Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79-88 (2006); published online EpubJan (10.1038/nbt1172).

89. S. Jeudy, M. Stelter, B. Coutard, R. Kahn, C. Abergel, Preliminary crystallographic analysis of the Escherichia coli YeaZ protein using the anomalous signal of a gadolinium derivative. *Acta crystallographica. Section F, Structural biology and*

crystallization communications **61**, 848-851 (2005); published online EpubSep 1 (10.1107/S1744309105025856).

90. T. J. Fiedler, H. A. Vincent, Y. Zuo, O. Gavrialov, A. Malhotra, Purification and crystallization of Escherichia coli oligoribonuclease. *Acta Crystallogr D Biol Crystallogr* **60**, 736-739 (2004); published online EpubApr (10.1107/S0907444904002252).

91. J. Kim, V. Malashkevich, S. Roday, M. Lisbin, V. L. Schramm, S. C. Almo, Structural and kinetic characterization of Escherichia coli TadA, the wobble-specific tRNA deaminase. *Biochemistry* **45**, 6407-6416 (2006); published online EpubMay 23 (10.1021/bi0522394).

92. S. Chaudhury, S. Lyskov, J. J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689-691 (2010); published online EpubMar 1 (10.1093/bioinformatics/btq007).

93. W. Kabsch, Automatic indexing of rotation diffraction patterns. *J Appl Cryst* **21**, 67-72 (1988).

94. P. R. Evans, An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr* **67**, 282-292 (2011); published online EpubApr (10.1107/S090744491003982X).

95. B. W. Matthews, Solvent content of protein crystals. *J Mol Biol* **33**, 491-497 (1968); published online EpubApr 28 (

96. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007); published online EpubAug 1 (10.1107/S0021889807021206).

97. P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010); published online EpubFeb (10.1107/S0907444909052925).

98. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010); published online EpubApr (10.1107/S0907444910007493).

99. J. Painter, E. A. Merritt, Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr D Biol Crystallogr* **62**, 439-450 (2006); published online EpubApr (10.1107/S0907444906005270).

100. V. B. Chen, W. B. Arendall, 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, D. C. Richardson, MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010); published online EpubJan (10.1107/S0907444909042073).

101. M. Krug, M. S. Weiss, U. Heinemann, U. Mueller, XDSAPP: a graphical user interface for the convenient processing of diffraction data using XDS. *J Appl Crystallogr* **45**, 568-572 (2012).

102. S. Adams, M. Ostermeier, Retinopathy associated with pegylated interferon and ribavirin treatment for chronic hepatitis C. *Optometry* **81**, 580-586 (2010); published online EpubNov (10.1016/j.optm.2010.04.094).

103. L. Potterton, S. McNicholas, E. Krissinel, J. Gruber, K. Cowtan, P. Emsley, G. N. Murshudov, S. Cohen, A. Perrakis, M. Noble, Developments in the CCP4 molecular-graphics project. *Acta Crystallogr D Biol Crystallogr* **60**, 2288-2294 (2004); published online EpubDec (10.1107/S0907444904023716).

104. K. Diederichs, P. A. Karplus, Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nat Struct Biol* **4**, 269-275 (1997); published online EpubApr (

105. M. S. Weiss, Global indicators of X-ray data quality. *J Appl Cryst* **34**, 130-135 (2001).

106. P. A. Karplus, K. Diederichs, Linking crystallographic model and data quality. *Science* **336**, 1030-1033 (2012); published online EpubMay 25 (10.1126/science.1218231).

107. P. Evans, Biochemistry. Resolving some old problems in protein crystallography. *Science* **336**, 986-987 (2012); published online EpubMay 25 (10.1126/science.1222162).

108. L. Wang, H. Pan, D. L. Smith, Hydrogen exchange-mass spectrometry: optimization of digestion conditions. *Mol Cell Proteomics* **1**, 132-138 (2002); published online EpubFeb (

109. D. Houde, S. A. Berkowitz, J. R. Engen, The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies. *J Pharm Sci* **100**, 2071-2086 (2011); published online EpubJun (10.1002/jps.22432).

110. I. A. Kaltashov, C. E. Bobst, R. R. Abzalimov, S. A. Berkowitz, D. Houde, Conformation and dynamics of biopharmaceuticals: transition of mass spectrometry-

based tools from academe to industry. *J Am Soc Mass Spectrom* **21**, 323-337 (2010); published online EpubMar (10.1016/j.jasms.2009.10.013).

111. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2010).

112. C. E. Bell, P. Frescura, A. Hochschild, M. Lewis, Crystal structure of the lambda repressor C-terminal domain provides a model for cooperative operator binding. *Cell* **101**, 801-811 (2000); published online EpubJun 23 (

113. I. B. Dodd, K. E. Shearwin, J. B. Egan, Revisited gene regulation in bacteriophage lambda. *Current opinion in genetics & development* **15**, 145-152 (2005); published online EpubApr (10.1016/j.gde.2005.02.001).

114. B. Qian, S. Raman, R. Das, P. Bradley, A. J. McCoy, R. J. Read, D. Baker, High-resolution structure prediction and the crystallographic phase problem. *Nature* **450**, 259-264 (2007); published online EpubNov 8 (10.1038/nature06249).

115. S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B. H. Kim, R. Das, N. V. Grishin, D. Baker, Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* **77 Suppl 9**, 89-99 (2009)10.1002/prot.22540).

116. A. E. Eriksson, W. A. Baase, J. A. Wozniak, B. W. Matthews, A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. *Nature* **355**, 371-373 (1992); published online EpubJan 23 (10.1038/355371a0).

117. A. E. Eriksson, W. A. Baase, X. J. Zhang, D. W. Heinz, M. Blaber, E. P. Baldwin, B. W. Matthews, Response of a protein structure to cavity-creating mutations and its

relation to the hydrophobic effect. *Science* **255**, 178-183 (1992); published online EpubJan 10 (

118. B. Xu, Q. X. Hua, S. H. Nakagawa, W. Jia, Y. C. Chu, P. G. Katsoyannis, M. A. Weiss, A cavity-forming mutation in insulin induces segmental unfolding of a surrounding alpha-helix. *Protein Sci* **11**, 104-116 (2002); published online EpubJan (10.1110/ps.32102).

119. J. Xu, W. A. Baase, E. Baldwin, B. W. Matthews, The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Sci* **7**, 158-177 (1998); published online EpubJan (10.1002/pro.5560070117).

120. T. Ehrig, D. J. O'Kane, F. G. Prendergast, Green-fluorescent protein mutants with altered fluorescence excitation spectra. *FEBS letters* **367**, 163-166 (1995); published online EpubJun 26 (

121. G. J. Palm, A. Zdanov, G. A. Gaitanaris, R. Stauber, G. N. Pavlakis, A. Wlodawer, The structural basis for spectral variations in green fluorescent protein. *Nat Struct Biol* **4**, 361-365 (1997); published online EpubMay (

122. B. D. Wallace, H. Wang, K. T. Lane, J. E. Scott, J. Orans, J. S. Koo, M. Venkatesh, C. Jobin, L. A. Yeh, S. Mani, M. R. Redinbo, Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science* **330**, 831-835 (2010); published online EpubNov 5 (10.1126/science.1191175).

123. A. Hvidt, K. Linderstrom-Lang, Exchange of hydrogen atoms in insulin with deuterium atoms in aqueous solutions. *Biochimica et biophysica acta* **14**, 574-575 (1954); published online EpubAug (

124. S. W. Englander, N. R. Kallenbach, Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Quarterly reviews of biophysics* **16**, 521-655 (1983); published online EpubNov (

125. J. J. Skinner, W. K. Lim, S. Bedard, B. E. Black, S. W. Englander, Protein dynamics viewed by hydrogen exchange. *Protein Sci* **21**, 996-1005 (2012); published online EpubJul (10.1002/pro.2081).

126. J. J. Skinner, W. K. Lim, S. Bedard, B. E. Black, S. W. Englander, Protein hydrogen exchange: testing current models. *Protein Sci* **21**, 987-995 (2012); published online EpubJul (10.1002/pro.2082).

127. Z. Zhang, D. L. Smith, Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. *Protein Sci* **2**, 522-531 (1993); published online EpubApr (10.1002/pro.5560020404).

128. C. Machicado, M. Bueno, J. Sancho, Predicting the structure of protein cavities created by mutation. *Protein Eng* **15**, 669-675 (2002); published online EpubAug (

129. O. N. Demerdash, M. D. Daily, J. C. Mitchell, Structure-based predictive models for allosteric hot spots. *PLoS Comput Biol* **5**, e1000531 (2009); published online EpubOct (10.1371/journal.pcbi.1000531).

130. A. Dixit, G. M. Verkhivker, Computational modeling of allosteric communication reveals organizing principles of mutation-induced signaling in ABL and EGFR kinases. *PLoS Comput Biol* **7**, e1002179 (2011); published online EpubOct (10.1371/journal.pcbi.1002179).

131. B. A. Kidd, D. Baker, W. E. Thomas, Computation of conformational coupling in allosteric proteins. *PLoS Comput Biol* **5**, e1000484 (2009); published online EpubAug (10.1371/journal.pcbi.1000484).

132. E. Laine, C. Goncalves, J. C. Karst, A. Lesnard, S. Rault, W. J. Tang, T. E. Malliavin, D. Ladant, A. Blondel, Use of allostery to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor. *Proc Natl Acad Sci U S A* **107**, 11277-11282 (2010); published online EpubJun 22 (10.1073/pnas.0914611107).

133. T. Amemiya, R. Koike, S. Fuchigami, M. Ikeguchi, A. Kidera, Classification and annotation of the relationship between protein structural change and ligand binding. *J Mol Biol* **408**, 568-584 (2011); published online EpubMay 6 (10.1016/j.jmb.2011.02.058).

134. S. Flores, N. Echols, D. Milburn, B. Hespenheide, K. Keating, J. Lu, S. Wells, E. Z. Yu, M. Thorpe, M. Gerstein, The Database of Macromolecular Motions: new features added at the decade mark. *Nucleic acids research* **34**, D296-301 (2006); published online EpubJan 1 (10.1093/nar/gkj046).

135. H. N. Motlagh, V. J. Hilser, Agonism/antagonism switching in allosteric ensembles. *Proc Natl Acad Sci U S A* **109**, 4134-4139 (2012); published online EpubMar 13 (10.1073/pnas.1120519109).

136. J. B. Bruning, A. A. Parent, G. Gil, M. Zhao, J. Nowak, M. C. Pace, C. L. Smith, P. V. Afonine, P. D. Adams, J. A. Katzenellenbogen, K. W. Nettles, Coupling of receptor conformation and ligand orientation determine graded activity. *Nature chemical biology* **6**, 837-843 (2010); published online EpubNov (10.1038/nchembio.451).

137. J. O. Wrabl, J. Gu, T. Liu, T. P. Schrank, S. T. Whitten, V. J. Hilser, The role of protein conformational fluctuations in allostery, function, and evolution. *Biophysical chemistry* **159**, 129-141 (2011); published online EpubNov (10.1016/j.bpc.2011.05.020).

138. V. J. Hilser, E. B. Thompson, Structural dynamics, intrinsic disorder, and allostery in nuclear receptors as transcription factors. *J Biol Chem* **286**, 39675-39682 (2011); published online EpubNov 18 (10.1074/jbc.R111.278929).

139. H. N. Motlagh, J. Li, E. B. Thompson, V. J. Hilser, Interplay between allostery and intrinsic disorder in an ensemble. *Biochemical Society transactions* **40**, 975-980 (2012); published online EpubOct (10.1042/BST20120163).

140. M. M. Babu, R. van der Lee, N. S. de Groot, J. Gsponer, Intrinsically disordered proteins: regulation and disease. *Curr Opin Struct Biol* **21**, 432-440 (2011); published online EpubJun (10.1016/j.sbi.2011.03.011).

141. P. E. Wright, H. J. Dyson, Linking folding and binding. *Curr Opin Struct Biol* **19**, 31-38 (2009); published online EpubFeb (10.1016/j.sbi.2008.12.003).

142. A. Vallee-Belisle, F. Ricci, K. W. Plaxco, Engineering biosensors with extended, narrowed, or arbitrarily edited dynamic range. *J Am Chem Soc* **134**, 2876-2879 (2012); published online EpubFeb 15 (10.1021/ja209850j).

143. V. N. Uversky, Intrinsically disordered proteins may escape unwanted interactions via functional misfolding. *Biochimica et biophysica acta* **1814**, 693-712 (2011); published online EpubMay (10.1016/j.bbapap.2011.03.010).

144. E. Hazy, P. Tompa, Limitations of induced folding in molecular recognition by intrinsically disordered proteins. *Chemphyschem : a European journal of chemical*

physics and physical chemistry **10**, 1415-1419 (2009); published online EpubJul 13 (10.1002/cphc.200900205).

145. Y. Huang, Z. Liu, Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: a critical assessment of the "fly-casting" mechanism. *J Mol Biol* **393**, 1143-1159 (2009); published online EpubNov 13 (10.1016/j.jmb.2009.09.010).

146. J. H. Choi, A. San, M. Ostermeier, Non-allosteric enzyme switches possess larger effector-induced changes in thermodynamic stability than their non-switch analogs. *Protein Sci* **22**, 475-485 (2013); published online EpubApr (10.1002/pro.2234).

147. J. M. Hickey, L. Weldon, P. S. Hefty, The atypical OmpR/PhoB response regulator ChxR from Chlamydia trachomatis forms homodimers in vivo and binds a direct repeat of nucleotide sequences. *J Bacteriol* **193**, 389-398 (2011); published online EpubJan (10.1128/JB.00833-10).

148. M. L. Barta, J. M. Hickey, A. Anbanandam, K. Dyer, M. Hammel, P. S. Hefty, Atypical response regulator ChxR from Chlamydia trachomatis is structurally poised for DNA binding. *PloS one* **9**, e91760 (2014)10.1371/journal.pone.0091760).

149. J. M. Hickey, S. Lovell, K. P. Battaile, L. Hu, C. R. Middaugh, P. S. Hefty, The atypical response regulator protein ChxR has structural characteristics and dimer interface interactions that are unique within the OmpR/PhoB subfamily. *J Biol Chem* **286**, 32606-32616 (2011); published online EpubSep 16 (10.1074/jbc.M111.220574).

150. J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, R. G. Coleman, ZINC: A Free Tool to Discover Chemistry for Biology. *J Chem Inf Model*, (2012); published online EpubJun 15 (10.1021/ci3001277).

151. J. Wickstrum, L. R. Sammons, K. N. Restivo, P. S. Hefty, Conditional gene expression in Chlamydia trachomatis using the tet system. *PloS one* **8**, e76743 (2013)10.1371/journal.pone.0076743).

152. M. Muller-McNicoll, K. M. Neugebauer, How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nature reviews. Genetics* **14**, 275-287 (2013); published online EpubApr (10.1038/nrg3434).

153. A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld, M. W. Hentze, Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393-1406 (2012); published online EpubJun 8 (10.1016/j.cell.2012.04.031).

154. A. G. Baltz, M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, M. Landthaler, The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell* **46**, 674-690 (2012); published online EpubJun 8 (10.1016/j.molcel.2012.05.021).

155. A. Pascale, S. Govoni, 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* **69**, 501-517 (2012); published online EpubFeb (10.1007/s00018-011-0810-7).

156. K. Kapeli, G. W. Yeo, Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. *Frontiers in neuroscience* **6**, 144 (2012)10.3389/fnins.2012.00144).

157. A. M. Khalil, J. L. Rinn, RNA-protein interactions in human health and disease. *Seminars in cell & developmental biology* **22**, 359-365 (2011); published online EpubJun (10.1016/j.semcdb.2011.02.016).

158. M. Ellenbecker, J. M. Lanchy, J. S. Lodmell, Identification of Rift Valley fever virus nucleocapsid protein-RNA binding inhibitors using a high-throughput screening assay. *J Biomol Screen* **17**, 1062-1070 (2012); published online EpubSep (10.1177/1087057112448100).

159. D. T. King, M. Barnes, D. Thomsen, C. H. Lee, Assessing specific oligonucleotides and small molecule antibiotics for the ability to inhibit the CRD-BP-CD44 RNA interaction. *PloS one* **9**, e91585 (2014)10.1371/journal.pone.0091585).

160. K. Cheng, X. Wang, H. Yin, Small-molecule inhibitors of the TLR3/dsRNA complex. *J Am Chem Soc* **133**, 3764-3767 (2011); published online EpubMar 23 (10.1021/ja111312h).

161. B. Ewald, D. Sampath, W. Plunkett, Nucleoside analogs: molecular mechanisms signaling cell death. *Oncogene* **27**, 6522-6537 (2008); published online EpubOct 27 (10.1038/onc.2008.316).

162. E. B. Fauman, B. K. Rai, E. S. Huang, Structure-based druggability assessment--identifying suitable targets for small molecule therapeutics. *Curr Opin Chem Biol* **15**, 463-468 (2011); published online EpubAug (10.1016/j.cbpa.2011.05.020).

163. C. Becher, R. Huber, H. Thermann, L. Ezechieli, S. Ostermeier, M. Wellmann, G. von Skrbensky, Effects of a surface matching articular resurfacing device on tibiofemoral contact pressure: results from continuous dynamic flexion-extension cycles. *Arch Orthop*

Trauma Surg **131**, 413-419 (2011); published online EpubMar (10.1007/s00402-010-1201-5).

164. P. C. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, M. T. Stahl, Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J Chem Inf Model* **50**, 572-584 (2010); published online EpubApr 26 (10.1021/ci100031x).

165. P. C. Hawkins, A. Nicholls, Conformer generation with OMEGA: learning from the data set and the analysis of failures. *J Chem Inf Model* **52**, 2919-2936 (2012).

166. S. F. OMEGA version 2.4.3. OpenEye Scientific Software, NM. http://www.eyesopen.com.

167. S. Do, H. Hu, A. Kolesnikov, W. Lee, V. Tsui, X. Wang, Z. Wen. (USA, 2012), vol. US20130039906 A1.

168. K. Audouze, E. O. Nielsen, G. M. Olsen, P. Ahring, T. D. Jorgensen, D. Peters, T. Liljefors, T. Balle, New ligands with affinity for the alpha4beta2 subtype of nicotinic acetylcholine receptors. Synthesis, receptor binding, and 3D-QSAR modeling. *J Med Chem* **49**, 3159-3171 (2006); published online EpubJun 1 (10.1021/jm058058h).

169. B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, CHARMM: the biomolecular simulation program. *Journal of computational chemistry* **30**, 1545-1614 (2009); published online EpubJul 30 (10.1002/jcc.21287).

170. S. Jo, T. Kim, V. G. Iyer, W. Im, CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry* **29**, 1859-1865 (2008); published online EpubAug (10.1002/jcc.20945).

171. Y. Cheng, W. H. Prusoff, Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical pharmacology* **22**, 3099-3108 (1973); published online EpubDec 1 (

172. F. H. Niesen, H. Berglund, M. Vedadi, The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature protocols* **2**, 2212-2221 (2007)10.1038/nprot.2007.321).

173. T. S. Rush, 3rd, J. A. Grant, L. Mosyak, A. Nicholls, A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* **48**, 1489-1495 (2005); published online EpubMar 10 (10.1021/jm040163o).

174. Z. Guo, D. Zhou, P. G. Schultz, Designing small-molecule switches for protein-protein interactions. *Science* **288**, 2042-2045 (2000); published online EpubJun 16 (

175. G. M. Daubner, A. Clery, F. H. Allain, RRM-RNA recognition: NMR or crystallography...and new findings. *Curr Opin Struct Biol* **23**, 100-108 (2013); published online EpubFeb (10.1016/j.sbi.2012.11.006).

176. C. Maris, C. Dominguez, F. H. Allain, The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS journal* **272**, 2118-2131 (2005); published online EpubMay (10.1111/j.1742-4658.2005.04653.x).

177. S. D. Auweter, F. C. Oberstrass, F. H. Allain, Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic acids research* **34**, 4943-4959 (2006)10.1093/nar/gkl620).

178. S. J. S. Nolan, 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.*, 2 (1999).

179. Y. Benitex, A. M. Baranger, Recognition of essential purines by the U1A protein. *BMC biochemistry* **8**, 22 (2007)10.1186/1471-2091-8-22).

180. J. B. Tuite, J. C. Shiels, A. M. Baranger, Substitution of an essential adenine in the U1A-RNA complex with a non-polar isostere. *Nucleic acids research* **30**, 5269-5275 (2002); published online EpubDec 1 (

181. T. Ohyama, T. Nagata, K. Tsuda, N. Kobayashi, T. Imai, H. Okano, T. Yamazaki, M. Katahira, Structure of Musashi1 in a complex with target RNA: the role of aromatic stacking interactions. *Nucleic acids research* **40**, 3218-3231 (2012); published online EpubApr (10.1093/nar/gkr1139).

182. M. Todaro, M. G. Francipane, J. P. Medema, G. Stassi, Colon cancer stem cells: promise of targeted therapy. *Gastroenterology* **138**, 2151-2162 (2010); published online EpubJun (10.1053/j.gastro.2009.12.063).

183. P. Nissen, M. Kjeldgaard, J. Nyborg, Macromolecular mimicry. *The EMBO journal* **19**, 489-495 (2000); published online EpubFeb 15 (10.1093/emboj/19.4.489).

184. P. A. Tsonis, B. Dwivedi, Molecular mimicry: structural camouflage of proteins and nucleic acids. *Biochimica et biophysica acta* **1783**, 177-187 (2008); published online EpubFeb (10.1016/j.bbamcr.2007.11.001).

185. L. Weber, K. Illgen, M. Almstetter, Discovery of New Multi Component Reactions with Combinatorial Methods. *Synlett* **3**, 366–374 (1999).

186. R. S. Spolar, J. H. Ha, M. T. Record, Jr., Hydrophobic effect in protein folding and other noncovalent processes involving proteins. *Proc Natl Acad Sci U S A* **86**, 8382-8385 (1989); published online EpubNov (

187. A. A. Bogan, K. S. Thorn, Anatomy of hot spots in protein interfaces. *J Mol Biol* **280**, 1-9 (1998); published online EpubJul 3 (10.1006/jmbi.1998.1843).

188. H. Pan, J. C. Lee, V. J. Hilser, Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble. *Proc Natl Acad Sci U S A* **97**, 12020-12025 (2000); published online EpubOct 24 (10.1073/pnas.220240297).

189. C. A. Lipinski, Drug-like properties and the causes of poor solubility and poor permeability. *Journal of pharmacological and toxicological methods* **44**, 235-249 (2000); published online EpubJul-Aug (