Mechanisms of G-quadruplex Unwinding and Repair

By

Andrew F. Voter

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Biochemistry)

at the

University of Wisconsin-Madison

2019

Date of final oral examination: 4/11/2019

The dissertation is approved by the following members of the Final Oral Committee

James L. Keck, Professor, Biomolecular Chemistry

Deane F. Mosher, Professor, Medicine

Michael M. Cox, Professor, Biochemistry

John M. Denu, Professor, Biomolecular Chemistry

Randal S. Tibbetts, Professor, Human Oncology

Mechanisms of G-quadruplex Unwinding and Repair

Andrew F. Voter

Under the supervision of Professor James L. Keck

University of Wisconsin – Madison

ABSTRACT

Guanine-quadruplexes (G4) are extraordinarily stable nucleic acid secondary structures with the potential to disrupt essential cellular processes. Four interlocking guanine bases can form an extensive hydrogen bond network, known as a guanine-quartet, which in turn stack upon each other to form a G4. The stability of G4s makes them hazardous to cells, and G4s can disrupt replication, transcription and translation. Yet their stability also makes G4s invaluable regulatory elements that are found throughout life, suggesting that cells have found ways to tame these potentially lethal obstructions. Systems of G4-interacting proteins and G4-resolving helicases have evolved to remove obstructing G4s, whereas damage from unresolved G4s can be repaired by dedicated DNA repair pathways. However, there is a lack of detailed mechanistic understanding of how these proteins and pathways interact with G4s, especially in bacteria. A better understanding of these processes may allow us to harness the potential of G4s as genetic control mechanisms or as therapeutic targets.

In this thesis, I describe new mechanisms of G4 unwinding and repair in bacteria and use these insights to better understand the role of G4 unwinding in *Neisseria gonorrhoeae* antigenic variation (AV). First, I propose a structural mechanism of G4 unwinding by bacterial RecQ helicases. I solved the 2.2Å X-ray crystal structure of the *Cronobacter sakazakii* RecQ bound to a resolved G4. This structure revealed a guanine specific pocket (GSP) on the surface of the RecQ helicase domain that had base flipped a guanine from the G4, leading to unfolding of the structure. Single-molecule Förster resonance energy transfer

studies demonstrated that RecQ variants with a defective GSP were incapable of unwinding G4 DNA yet retained their ability to unwind duplex DNA. I propose a base-flipping model of G4 resolution that may be broadly applicable, as other G4 resolving helicases possess similar pockets.

This mechanism paves the way for further studies of G4 by isolating G4-specific activities from other helicase functions, as exemplified by the results in Chapter 3. To avoid clearance by the immune system, the pathogen *N. gonorrhoeae* alters the composition of its immunogenic pilus proteins during infection. This process, known as antigenic variation (AV), is critical for the pathogenesis of *N. gonorrheae* and, thus, represents a potential therapeutic target. While the precise mechanisms behind AV remain unclear, the formation of the parallel *pilE* G4 is essential. RecQ is critical in AV, and its role in the process has been proposed to be unwinding of the *pilE* G4. However, RecQ also aids AV by promoting homologous recombination in a G4-independent fashion. To define the role of RecQ-mediated resolution of the *pilE* G4 in AV, we generated *N. gonorrheae* strains with GSP-defective RecQ variants and found these strains underwent AV at the same rate as those with wild type RecQ. The purified RecQ variant retained duplex helicase assay but lost the ability to resolve antiparallel G4s. Interestingly, neither RecQ nor the GSP-deficient variant were able to unwind the *pilE* G4. Together, these results demonstrate that AV occurs independently of RecQ-mediated *pilE* unwinding.

CITATIONS

This dissertation contains the following primary research articles and a review article:

Voter AF, Qiu Y, Tippana R, Myong S, Keck JL. (2018) "A guanine-flipping and sequestration mechanism *for* G-quadruplex unwinding by RecQ helicases" *Nature Communications* Article number:4201.

Voter AF, Manthei KA, Keck JL. (2016) "A high throughput screening strategy to identify protein-protein interaction inhibitors that block the Fanconi anemia DNA repair pathway" *Journal of Biomolecular Screening* 21(6), 626-633.

Voter AF*, Killoran MP*, Ananiev GE, Wildman SA, Hoffman FM, Keck JL. (2018) "A high-throughput screening strategy to identify inhibitors of SSB protein-protein interactions in an academic screening facility" *SLAS Discovery* 23(1), 94-101. *These authors contributed equally

Voter AF, Keck, JL. "Development of Protein-Protein Interaction Inhibitors for the Treatment of Infectious Diseases" In *Advances in Protein Chemistry and Structural Biology*; Academic Press: Cambridge, MA, 2018; Vol. 111, 197–222.

Liu S, Alnammi M, Ericksen SS, Voter AF, Ananiev G, Keck JL, Hoffmann FM, Wildman SA, Gitter A. (2018) "Practical model selection for prospective virtual screening" *Journal of Chemical Information and Modeling,* **2019**, *59* (1), 282–293.

Additional contents of this dissertation are submitted for publication:

Voter AF, Callaghan MM, Tippana R, Myong S, Dillard JP, Keck JL "Antigenic variation in *Neisseria gonorrhoeae* occurs independently of RecQ-mediated unwinding of the *pilE* G-quadruplex", submitted to the *Journal of Bacteriology.*

ACKNOWLEDGEMENTS

Foremost, I would like to thank Jim. Before joining a thesis lab, I consulted with a handful of people about my options. While most were equivocal, Deane Mosher was not. He assured me that Dr. Keck would be a near perfect mentor for my thesis and I am incredible grateful that Deane was right. Thank you for investing so much time and energy in my development, molding my writing into something presentable and giving me the space to grow into an independent scientist.

Thanks also to my thesis committee members, Mark Burkard, Mike Cox, John Denu, Randy Tibbetts and Deane Mosher, for challenging me, keeping up with my ever-changing projects and providing throughtful support and guidance.

I received enormous help from my scientific collaborators. I would especially like to thank Dr. Sua Myong, Dr. Ramreddy Tippana, and Dr. Peggy Qiu from Johns Hopkins; Dr. Joe Dillard and Melanie Callaghan from UW-Madison; and Dr. Hawkinson and Deepti Mudaliar at the University of Minnesota. All of you were incredibly generous with your time and pushed my work to higher levels. Thank you. I would also like to acknowledge the contributions of the members of the UW small molecule screening facility, especially Gene Ananiev, Song Guo and Scott Wildman, without whom none of the work in the appendices would have been possible. Similarly, Darrell McCaslin and Dan Stevens were instrumental in helping me collect data and pass the time during long SPR runs. Special thanks to Zach Romero, Liz Wood and Mike Cox for the generous help with strains and indulging some of my crazier ideas.

Thank you to members of the Keck lab past and present. All of this thesis builds on the work of Kelly Manthei and Michael Killoran, to whom I am indebted for preparing the way for whatever success I have had. I am grateful to my mentors in the Keck lab, Basu Bhattacharyya, Sarah Wessel, Chrissy Petzold and Tricia Windgassen for teaching me how to be a biochemist. Thanks also to Kasia Dubiel, Shawn

Laursen, Aidan McKenzie, Alex Duckworth, Brian Ferrer and Rachel Cuney for helpful discussions and making it easy to be in lab every day. Best of luck to you all.

I never would have made it to Madison without the endless love and support of my family. Mom and Dad, thank you for instilling in me the grit and curiosity needed to succeed and thanks for being there to listen, anytime I needed. Thank you, Nancy, for welcoming me into your family and moving halfway across the country to join us. Your presence in Madison has been an incredible gift and I thank you for your generosity. Finally, I would like to thank Carolyn, for not only tolerating my ever-changing and ever-expanding collection of acronyms, but even learning them. Having someone to share the highs and lows of graduate training with has been invaluable. Thank you for your patience, joy, love and boundless support. I look forward to sharing our next adventures, whatever and wherever they may be.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Guanine-quadruplexes

Guanine-quadruplexes (G4s) are extraordinarily stable nucleic acid secondary structures that form in guanine-rich DNA or RNA and can block cellular processes with devasting consequences. The first hint of these structures was reported in 1910 by Bang, who noted that concentrated solutions of guanylic acid (guanosine monophosphate or GMP) would gel upon cooling.[1] While extensive hydrogen bonding between guanine nucleobases was long suspected to be the cause, Gellert, Lipsett and Davies confirmed the structure of the interaction in 1962.[2] Drying the guanylic acid gel yielded fibers, which were shown to adopt a helical structure by X-ray diffraction. Rather than traditional Watson-Crick base pairing, each guanine base hydrogen bonds with two additional guanine bases using Hoogstein base-pairing. The bipartite binding to two adjacent nucleobases, 90° apart, yields the formation of a guanine-quartet, an interwoven ring of four hydrogen bonded guanine nucleobases (Figure 1.1). Layers of G-quartets can base-stack, yielding the stable fibers observed by Lipsett.[2] Remarkably, all of this occurs through non-covalent interactions between monomeric GMP molecules.

At approximately the same time, Khorona and co-workers observed G4s with linked G-quartets. Their work showed that short oligonucleotides (3-5 nucleotides) of polyguanine formed a secondary structure that melted at high temperatures, was destabilized by Na^+ and resisted degradation by venom phosphodiesterases. Additionally, no comparable structures were observed for oligos comprising other nucleotides.[3] These results capture many of the critical features of G4s and clearly foreshadow many later developments in the field.

Within the G-quartet, the keto group of each guanine points into the ring, and this electronegativity is countered by the presence of central cations. The identity of these cations has a marked effect on the stability of a G4. Low ionic strengths or Li^+ ions will not support G4 formation, whereas K^+ ions stabilize G4s and Na^+ ions allow for G4 folding, although with lower melting temperatures

than K^+ for the same G4. It was originally proposed that K^+ ions had an "optimal fit" for the core of the G4, leading to the preferential stabilization by K^+ over Na^+ or Li^+.[4] While intuitive, this hypothesis was shown to be over simplified. Rather, either Na^+ or K^+ ions can be readily accommodated between the G-quartet planes and the free energy of G4 binding is actually more favorable for Na^+ than K^+ (-1.7 kcal/mol at 25 °C). However, the free energy for hydration of Na^+ is much higher than K^+ (17.6 kcal/mol at 25 °C) and this drives displacement of the Na^+ ion and ultimately G4 stability.[5]

Despite sharing a core of stacked G-quartets, a range of G4 structures have been described. These are broadly classified based on the orientation of the phosphodiester bonds that link adjacent layers of G-quartets. Each set of 3 guanine nucleotides that comprise one edge of the G4 is called a "stem". In parallel G4s, typified by the *c-myc* G4, each of the stems are oriented in the same direction. Antiparallel G4s, have stems running in opposing directions and are represented by the human telomeric G4 (Figure 1.2). Additional conformations exist within each category[6]. The principle components in determining the orientation of the G4 are the length and nucleotide composition of the "loops" that link two stems. G4s are almost always parallel if the shortest loop is 1 or 2 nucleotides long because this loop is too short to reverse itself before starting the next stem. In this case, the loop must cross from top to bottom along the outside of the guanine core, yielding a parallel orientation.[7,8] If all the loops are at least 3 nucleotides in length, then the orientation is determined by composition of the loops. Loops comprising pyrimidine bases tend to be parallel, while increasing purine composition leads to antiparallel or unfolded structures.[7] This effect is likely due to the steric bulk of the purine base preventing the tight wrapping needed for a parallel orientation.

1.2. Mechanisms of G4 tolerance.

The presence of G4 structures has been shown to disrupt nearly all cellular functions that rely on interactions with nucleic acids including DNA replication,[9,10] transcription[11] and translation.[12,13] To tolerate

these insults, cells have developed an armament of G4 mitigation and DNA repair pathways. The most straightforward of these is to simply suppress G4 formation when the quadruplex is not required. G4 cannot disrupt, or even fold, when the G4-forming sequences is sequestered within duplex DNA. This is consistent with an overall cellular strategy to limit the presence of single-stranded (ss)DNA of any type within the cell, as ssDNA is susceptible to damage and formation of unwanted secondary structures.[14-16] Whenever the presence of ssDNA is required, such as for replication, repair or transcription, exposed ssDNA is coated with proteins, such as the ssDNA-binding protein (SSB) in bacteria or Replication Protein A (RPA) in eukaryotes.[16] Indeed, RPA has been shown to suppress the formation of G4 structures.[17] A comparable strategy is employed for G-quadruplexes in mRNA; eukaryotic cells contain mRNA transcripts with potential G4 sequences, but these remain unfolded *in vivo*. While the bacterial transcriptome is largely depleted of G4 forming sequences, exogenous mRNA can form G4s, although these are toxic.[18]

For instances when G4s form in genomic DNA or in RNA, cells have evolved helicases with G4 unwinding capabilities. Here, the goal is to remove the G4 before it is encounter by essential cellular processes, like replication. Three primary classes of DNA G4 resolving helicases have been described. First is the RecQ family of helicases. These superfamily-2 DNA unwinding enzymes are well conserved from bacteria to eukaryotes and unwind a diverse range of DNA substrates in a 3’ to 5’ direction to support DNA repair pathways. The bacterial[19] and yeast[20] RecQ homologs (RecQ and Sgs1, respectively) have been shown to unwind G4 DNA, as have two of the five mammalian RecQ homologs, BLM[21] and WRN.[22] All of these helicases have activities beyond G4 unwinding, so it has been difficult to isolate the physiological role of G4 unwinding specifically.[23] The second class of G4-resolving helicases is Fe-S cluster helicases, exemplified by XPD,[24] FANCJ[25] (mammals), and DinG (bacteria). Loss of the eukaryotic homologs is associated with deletions of G-rich regions of the genome, thought to be associated with a failure of lagging strand replication through stable secondary structures.[26] This phenotype has not been observed for DinG, although the protein does unwind G4 DNA.[27] The last major class of G4-resolving helicases is the

superfamily1 Pif1 helicases. Originally discovered in *Saccharomyces cerevisiae*, Pif1 was noted for its function as an inhibitor of telomerase[28] and, similar to the Fe-S helicases, for protection of genomic stability of G-rich genomic regions.[29] Subsequent *in vitro* biochemical assays validated its function as a G4-resolving helicase.[30] The closest Pif1 homolog in *E. coli* is RecD,[31] which is best known for its role in the repair of double strand DNA breaks by homologous recombination as part of the RecBCD complex. However, isolated RecD has a modest helicase activity which may limit its roles outside of the RecBCD complex.[32]

While these helicases have all been observed to unwind G4 DNA, the structural mechanisms by which they resolve G4s remain unclear. The only published mechanism to date is for the RNA G4 helicase DHX36, which requires a "DHX-specific motif".[33] While mechanistically informative, this motif is not present in other G4 resolving helicases and the proposed mechanism cannot be applied to G4-resolving helicases generally. This lack of structural mechanisms of G4 activity has hampered research progress. As noted above, all DNA G4-resolving helicases described have important G4-independent functions. With a structural mechanism of helicase activity, it should be possible to develop helicase variants that selectively disable G4 helicase activity, while sparring duplex unwinding. These variants would be invaluable for understanding the G4-specific action of helicases.

Without these G4-suppression and unwinding systems (and occasionally even with them), G4s can escape resolution and persist during replication. When this occurs, the replication fork stalls at the G4. In eukaryotes, the fork can be restarted by template switching or the structure can be bypassed by translesion synthesis[34] or primpol.[35] Either solution is problematic for the cell; translesion synthesis polymerases are mutagenic and template switching can lead to genomic rearrangements if the mediated homology search and strand invasion fails to align properly to the repetitive G-rich sequence.[36] In either case, replication proceeds beyond the G4. Occasionally, the replication fork will be unable to bypass the G4, leading to fork collapse and a double strand break. This break must be repaired to avoid the

catastrophic loss of a chromosome. Studies in *Caenorhabditis elegens* demonstrate that these breaks are repaired by a variant of non-homologous end joining known as theta-mediated end joining, which results in deletions of 50-300 base pairs.[37] However, this process is almost certainly not used by *E. coli*, which lack a non-homologous end joining system. The mechanisms by which bacterial cells tolerate the formation of genomic G4s remain unknown. However, given the relatively paucity of G4 forming sequences in bacterial genomes, they likely rely on a predominantly suppressive approach.

1.3. Physiological functions of G4s.

Given the toxicity of G4s, negative selective pressure might be expected to eliminate G4-forming sequences. However, bioinformatic studies have demonstrated that G4-forming sequences are found in all domains of life. In these studies, a genome is scanned for sequences that matches a putative G4-forming sequence, usually similar to $G_{3+}N_{1-7}$ $G_{3+}N_{1-7}$ $G_{3+}N_{1-7}$ G_{3+} (where N is any base, including guanine). By these estimates, the human genome contains ~375,000 G4 forming sequences,[38,39] while bacteria tend to have far fewer, with 52 and 94 found in the gram-negative *E. coli* and *Neisseria gonorrhoeae* respectively, and only four in the gram-positive *Bacillus subtilis*.[40] While useful as a starting point, the *in silico* nature of these surveys means that some degree of false negative and positives are observed. For instance, G4s folding has been observed in sequences that contain a bulge, or one or more non-guanine residue in one of the G4 stems.[41] Such sequences may be biologically relevant, but are unlikely to be accurately identified *in silico*. Furthermore, many of the potential sequences found computationally, such as those with especially long loops are unlikely to fold in cells due to the inherent instability of such loops and length of unprotected nucleic acid being available to fold.[8] When *bona fide* G4 forming sites were mapped using G4-seq, more G4 sites were observed than were predicted computationally, suggesting *in silico* predications may be underestimating the prevalence of genomic G4s.[42]

Rather than merely tolerating the presence of G4 forming sequences, cells seem to have tamed them and instead use G4s as regulatory elements. In support of this, G4-forming sequences have been found to be overrepresented in regulatory regions of the genome in both bacteria[43,44] and eukaryotes.[45-47] Based on their location, G4s can be both positive and negative regulators of gene expression. The simplest case is when a G4 is found within an open reading frame (ORF). In bacteria, a G4 located in the antisense strand can suppress gene expression by blocking progression of RNA polymerase.[11] If ultimately transcribed into the mRNA, an RNA G4 can block translation by the ribosome.[48] Even outside of the ORF, the presence of a G4 can modulate gene expression if located within the promotor, or 5'-UTR of a gene.[12] G4-mediated gene regulation has been observed in viruses, bacteria and eukaryotes, indicating that this is a biologically important control mechanism.[49]

Beyond the regulation of gene expression, several niche applications for G4s have been described. One of the best-studied physiological examples of G4s is their role in telomere biology. Telomeres are repeated sequences of DNA that cap the end of linear chromosomes to protect from degradation and provide a mechanism to overcome the end replication problem in eukaryotes. Because DNA replication cannot extend the extreme end of the chromosome, the chromosome is shortened during each round of replication. To avoid losing essential genetic elements, the telomere is sacrificially shortened. After ~50 rounds of division, the Hayflick limit, the telomeres have shortened extensively, and normal cells enter senescence.[50] However, stem cells continue to divide past the Hayflick limit by expressing telomerase, an enzyme that extends the telomere.[51] Telomerase is overexpressed in most cancers to bypass the Hayflick limit and avoid the extensive chromosomal loss that would otherwise result from their explosive growth.[52]

Almost all telomere sequences examined to date are guanine-rich sequences with G4 folding potential (5'-TTAGGG-3' in vertebrates).[53,54] And while G4s have been observed to form at the end of chromosomes, the precise role of the telomeric G4 remains unclear. Much of the time, the telomere is protected in a double stranded form in the shelterin complex, which is incompatible with G4 formation.[55]

However, cells have developed mechanism to regulate formation of telomeric quadruplexes[56] and one intriguing hypothesis proposed by Moye and coworkers posits that telomerase may actually recognize and partially unwind the telomeric G4 for use as a template during telomere extension.[57] Regardless of the specific function of the telomeric G4, its importance in telomere homeostasis is clear. Patients with Werner syndrome, who lack the G4-unwinding helicase WRN, have a striking progeroid (pre-mature aging) syndrome, which may to be secondary to premature telomere loss associated with a failure of WRN to unwind telomeric G4s.[58-60] Additionally, disabling telomerase in cancer tends to telomere shortening and eventual senescence or cell death.[61]

Bacteria have far fewer G4 forming sequences than eukaryotic species, but quadruplexes can still play an important role in regulation of gene expression and pathogenesis. By far, the best studied bacterial G4 is the *pilE* G4 of *Neisseria gonorrhoeae*, which orchestrates antigenic variation (AV) in the pathogen.[62] To avoid the immune system during infection, *N. gonorrhoeae* varies the composition of its immunogenic pilin using AV. The process is initiated by transcription of a small non-coding RNA. Once the strands of genomic DNA are separated by RNA polymerase, the newly single stranded *pilE* G4-forming sequence folds into a parallel G4, heralding the onset of the AV process.[63] Both the presence and parallel orientation of the G4 are essential for AV, but despite ten years of study, the specific role of the G4 remains unclear.[64] Intriguingly, it appears that G4s might be a conserved mechanism of regulating AV in other pathogens, including the causative agents of syphilis[65] and Lyme disease.[66] A deeper understanding of the mechanism of G4 regulation during these process could lead to novel strategies for the treatment of these pathogens.

1.4. G-quadruplexes as therapeutic targets.

Studies have demonstrated the importance of G4 homeostasis in diseases ranging from gonorrhea to inherited genomic instability and HIV to cancer. In infections and cancer, disrupting G4 unwinding leads to a loss of virulence, suggesting that small-molecule modulation of G4 stability could be

an attractive therapeutic target. Accordingly, a vast array of G4 stabilizing molecules have been developed for use as tool compounds and potential therapeutic.[67] Of these molecules, the most commonly used include N-methyl mesoporphyrin IX (NMM), BRACO-19 and TMPyP4 (Figure 1.3). NMM and BRACO-19 are both planer molecules that stabilize G4s by base-stacking with the topmost G-quartet (Figure 1.4), whereas BRACO-19 has additional functional groups that extent into grooves along the side of the G4.[68,69] Curiously, despite its planar structure, a crystal structure of TMPyP4 bound to the human telomeric G4 suggests that its G4 stabilization is derived from interaction with the loops that bridge adjacent stems of the G-quartet, which may account for its relatively poor selectivity for G4 DNA.[70]

These and other G4 stabilizing tool compounds have been shown to be active in *in vitro* disease models. Many virus genomes contain regulatory G4 elements. As an example, the human immunodeficiency virus (HIV) has twelve potential G4-forming sequences, which are used to modulate promoter expression and control dimerization of the viral genome. Both BRACO-19 and TMPyP4 have been shown to have some antiviral effects against HIV-1.[71,72] G4 stabilization may also represent a novel route to virulence control in pathogens that use G4s in AV. NMM treatment has been shown to suppress AV in *N. gonorroheae*,[64] which could give the immune system time to develop a protective response. Lastly, some cancers have been shown to be susceptible to G4 stabilization. G4s are found in the promotor regions of some oncogenes, such as *c-myc*, and treatment with TMPyP4 represses *c-myc* transcription and improved survival in a mouse xenograph model of breast cancer.[73] Multiple G4 stabilizers have been found to block telomerase and demonstrated activity in model systems of cancer treatment.[74-76] One therapeutic targeting G4s, quarfloxacin, was found to be safe in phase I clinical trials,[77] and entered phase II trials in 2011, although no results appear to have been released. Nevertheless, this agent showed *in vitro* activity against *Plasmodium falciparum*, and may have utility outside of cancer applications.[78]

Despite the appearance of G4 stabilizing ligands agents in clinical trials, there are major gaps in our understanding of G4 biology. While the mechanisms of G4 metabolism and repair have been identified

in eukaryotes, these pathways are almost entirely uncharacterized in bacteria. Differences in how each domain of life tolerate the presence of potentially toxic G4 could lead to new routes of infection control. Furthermore, we have almost no insights into the structural mechanisms of G4 homeostasis, either how the structures are recognized by repair factors or resolved by helicases. This gap has forced researched to rely on overly broad techniques to study G4 interactions, such as gene knockouts and treatment with G4 stabilizing ligands. These methods are informative as screening techniques, but since all known G4-resolving helicases have non-G4 functions and the human genome has over 350,000 potential G4 forming regions, off-target effects are inevitable. Interventions that are specific for G4 structures are clearly needed.

To help overcome these obstacles, I have determined the 2.2 Å resolution crystal structure of the RecQ helicase from *Cronobacter sakazkii* bound to a resolved G4, described in chapter 2. This structure revealed the presence of a novel pocket on the surface of the helicase domain of RecQ that extracts and sequesters one of the guanines from the G-quartet. This base flipping destabilizes the G4 and promotes unwinding by the helicase. I show this pocket is required for unwinding of G4 DNA but is dispensable for duplex unwinding and may be a conserved mechanism of G4 unwinding. Next, I exploited this pocket to explore the mechanisms of AV in *N. gonorroheae*. With the experiments in chapter 3, I demonstrate that RecQ-mediated unwinding of the *pilE* G4 is dispensable for AV and furthermore, that even wild type RecQ is incapable of unwinding parallel G4. With these results, I propose a new model for RecQ function in AV.

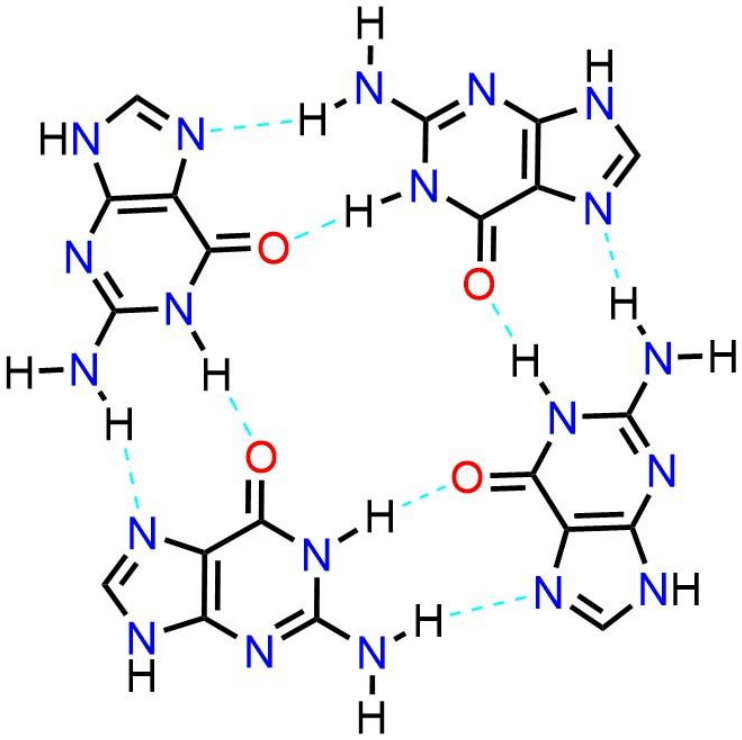


Figure 1.1. Structure of a guanine-quartet. An extensive hydrogen bonding network (dashed blue lines) links four Hoogstein base-paired guanines.

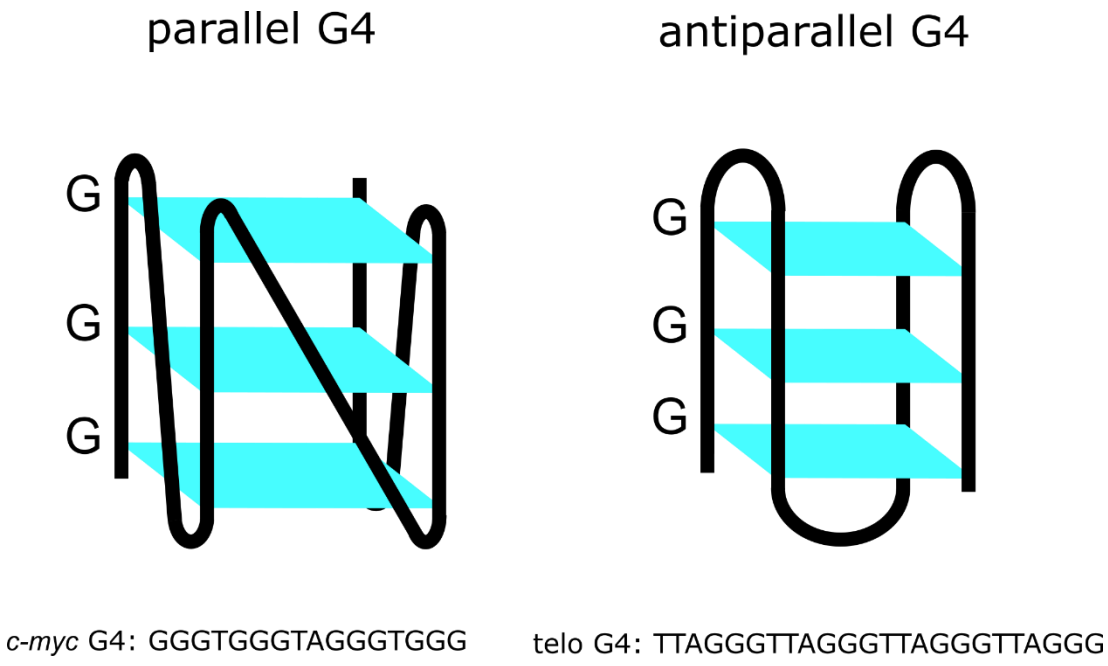


Figure 1.2. Conformations of the prototypical parallel and antiparallel G4s. Depending on the sequence and loop size, G4s can adopt either parallel or antiparallel conformations, categorized based on the orientation of the phosphodiester stem. Sequences for a prototypical example of each G4 is shown underneath the G4.

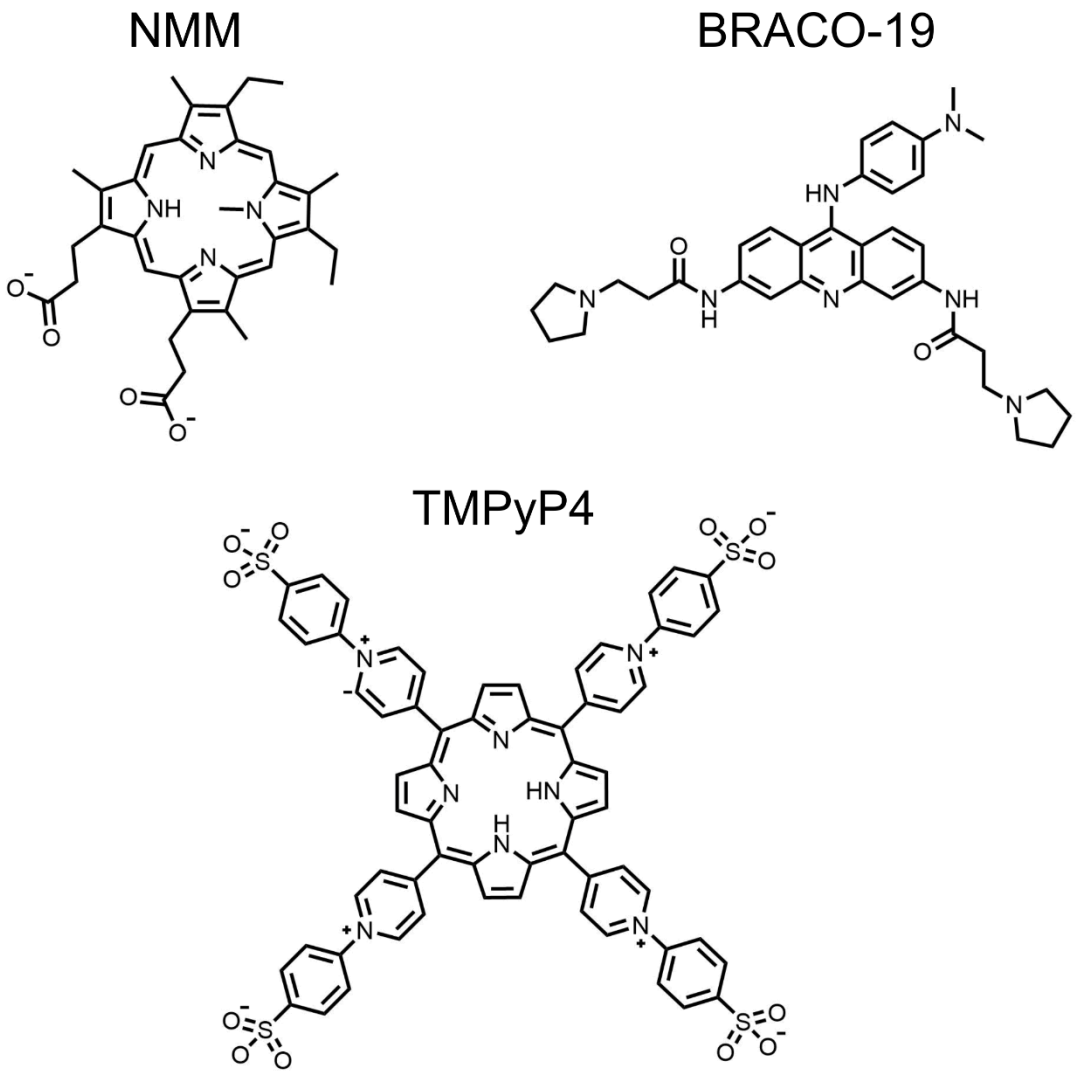


Figure 1.3. Structures of G4 stabilizing ligands.

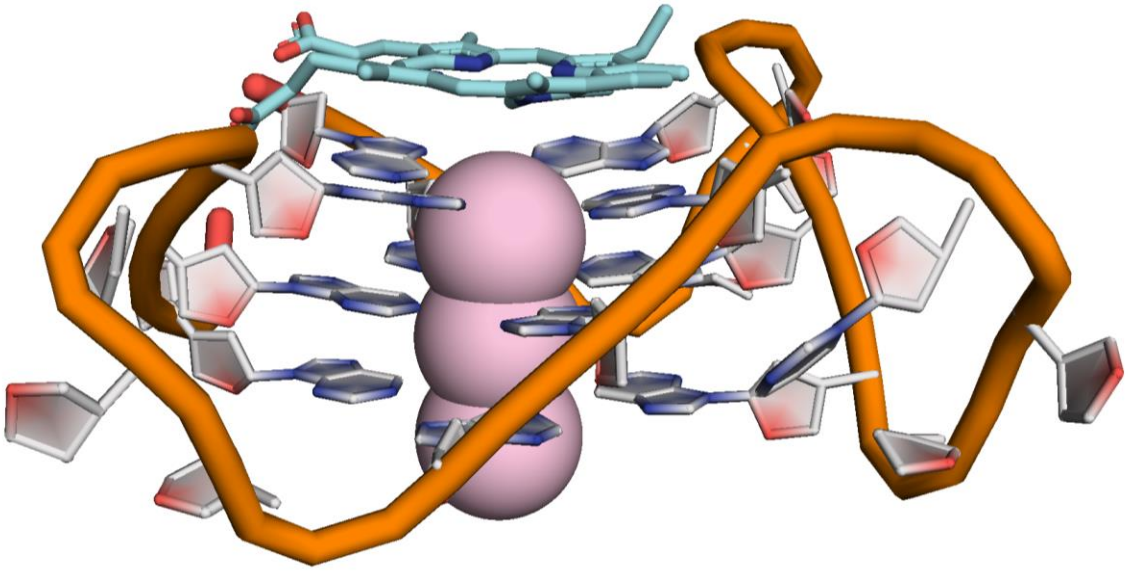


Figure 1.4. G4 stabilization by NMM. The structure of the antiparallel human telomeric G4 (white and yellow) stabilized by NMM (teal sticks) from PDB 4G0F. The central K^+ ions are represented by the pink spheres. NMM stabilizes G4s by base stacking atop the layers of G-quartets.

References.

1 Bang, I. Untersuchungen uber die Guanylsaure. *Biochem Z* **26**, 293-311 (1910).

2 Gellert, M., Lipsett, M. N. & Davies, D. R. Helix formation by guanylic acid. *Proc Nat Acad Sci* **48**, 2013-2018 (1962).

3 Ralph, R. K., Khorana, H. G. & Connors, W. J. Secondary Structure and Aggregation in Deoxyguanosine Oligonucleotides. *J Am Chem Soc* **84**, 2265-2266 (1962).

4 Sundquist, W. I. & Klug, A. Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature* **342**, 825-829 (1989).

5 Hud, N. V., Smith, F. W., Anet, F. A. L. & Feigon, J. The Selectivity for K+ versus Na+ in DNA Quadruplexes Is Dominated by Relative Free Energies of Hydration: A Thermodynamic Analysis by 1H NMR. *Biochem* **35**, 15383-15390 (1996).

6 Chaires, J. B., Trent, J. O., Gray, R. D. & Lane, A. N. Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res* **36**, 5482-5515 (2008).

7 Tippana, R., Xiao, W. & Myong, S. G-quadruplex conformation and dynamics are determined by loop length and sequence. *Nucleic Acids Res* **42**, 8106-8114, (2014).

8 Guedin, A., Gros, J., Alberti, P. & Mergny, J.-L. How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res* **38**, 7858-7868 (2010).

9 Woodford, K. J., Howell, R. M. & Usdin, K. A novel K(+)-dependent DNA synthesis arrest site in a commonly occurring sequence motif in eukaryotes. *J Biol Chem* **269**, 27029-27035 (1994).

10 Weitzmann, M. N., Woodford, K. J. & Usdin, K. The Development and Use of a DNA Polymerase Arrest Assay for the Evaluation of Parameters Affecting Intrastrand Tetraplex Formation. *J Biol Chem* **271**, 20958-20964 (1996).

11 Agarwal, T., Roy, S., Kumar, S., Chakraborty, T. K. & Maiti, S. In the Sense of Transcription Regulation by G-Quadruplexes: Asymmetric Effects in Sense and Antisense Strands. *Biochem* **53**, 3711-3718, (2014).

12 Holder, Isabelle T. & Hartig, Jörg S. A Matter of Location: Influence of G-Quadruplexes on Escherichia coli Gene Expression. *Chem & Biol* **21**, 1511-1521, (2014).

13 Arora, A. *et al.* Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. *Rna* **14**, 1290-1296 (2008).

14 Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709-715 (1993).

15 Chan, K. *et al.* Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS genetics* **8**, e1003149-e1003149, (2012).

16 Shereda, R. D., Kozlov, A. G., Lohman, T. M., Cox, M. M. & Keck, J. L. SSB as an organizer/mobilizer of genome maintenance complexes. *Crit Rev Biochem Mol Biol* **43**, 289-318, (2008).

17 Salas, T. R. *et al.* Human replication protein A unfolds telomeric G-quadruplexes. *Nucleic Acids Res* **34**, 4857-4865, (2006).

18 Guo, J. U. & Bartel, D. P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* **353**, (2016).

19 Wu, X. & Maizels, N. Substrate-specific inhibition of RecQ helicase. *Nucleic Acids Res* **29**, 1765-1771 (2001).

20 Sun, H., Bennett, R. J. & Maizels, N. The Saccharomyces cerevisiae Sgs1 helicase efficiently unwinds G-G paired DNAs. *Nucleic Acids Res* **27**, 1978-1984, (1999).

21 Sun, H., Karow, J. K., Hickson, I. D. & Maizels, N. The Bloom's syndrome helicase unwinds G4 DNA. *J Biol Chem* **273**, 27587-27592 (1998).

22 Fry, M. & Loeb, L. A. Human Werner Syndrome DNA Helicase Unwinds Tetrahelical Structures of the Fragile X Syndrome Repeat Sequence d(CGG) n. *J Biol Chem* **274**, 12797-12802 (1999).

23 Bernstein, K. A., Gangloff, S. & Rothstein, R. The RecQ DNA Helicases in DNA Repair. *Annu Rev Genet* **44**, 393-417 (2010).

24 Liu, H. *et al.* Structure of the DNA Repair Helicase XPD. *Cell* **133**, 801-812, (2008).

25 Wu, Y., Shin-ya, K. & Brosh, R. M. FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Mol Cell Biol* **28**, 4116-4128 (2008).

26 Cheung, I., Schertzer, M., Rose, A. & Lansdorp, P. M. Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. *Nature Genet* **31**, 405 (2002).

27 Thakur, R. S. *et al.* Mycobacterium tuberculosis DinG is a structure-specific helicase that unwinds G4 DNA: implications for targeting G4 DNA as a novel therapeutic approach. *J Biol Chem* **289**, 25112-25136 (2014).

28 Boulé, J.-B., Vega, L. R. & Zakian, V. A. The yeast Pif1p helicase removes telomerase from telomeric DNA. *Nature* **438**, 57 (2005).

29 Ribeyre, C. *et al.* The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS genet* **5**, e1000475 (2009).

30 Paeschke, K., Capra, John A. & Zakian, Virginia A. DNA Replication through G-Quadruplex Motifs Is Promoted by the Saccharomyces cerevisiae Pif1 DNA Helicase. *Cell* **145**, 678-691 (2011).

31 Bochman, M. L., Judge, C. P. & Zakian, V. A. The Pif1 family in prokaryotes: what are our helicases doing in your bacteria? *Mol Biol Cell* **22**, 1955-1959 (2011).

32 Chen, H. W., Ruan, B., Yu, M., Wang, J. & Julin, D. A. The RecD subunit of the RecBCD enzyme from Escherichia coli is a single-stranded DNA-dependent ATPase. *J Biol Chem* **272**, 10072-10079 (1997).

33 Chen, M. C. *et al.* Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. *Nature* **558**, 465-469 (2018).

34 Youds, J. L., O'Neil, N. J. & Rose, A. M. Homologous recombination is required for genome stability in the absence of DOG-1 in Caenorhabditis elegans. *Genet* **173**, 697-708 (2006).

35 Schiavone, D. *et al.* PrimPol Is Required for Replicative Tolerance of G Quadruplexes in Vertebrate Cells. *Mol Cell* **61**, 161-169 (2016).

36 Lopes, J. *et al.* G-quadruplex-induced instability during leading-strand replication. *EMBO J* **30**, 4033-4046 (2011).

37 Koole, W. *et al.* A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nature Commun* **5**, 3216 (2014).

38 Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* **33**, 2908-2916 (2005).

39 Todd, A. K., Johnston, M. & Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res* **33**, 2901-2907 (2005).

40 Dhapola, P. & Chowdhury, S. QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res* **44**, W277-283 (2016).

41 Mukundan, V. T. & Phan, A. T. Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J Am Chem Soc* **135**, 5017-5028 (2013).

42 Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* **33**, 877-881 (2015).

43 Rawal, P. *et al.* Genome-wide prediction of G4 DNA as regulatory motifs: Role in Escherichia coli global regulation. *Genome Research* **16**, 644-655 (2006).

44 Beaume, N. *et al.* Genome-wide study predicts promoter-G4 DNA motifs regulate selective functions in bacteria: radioresistance of D. radiodurans involves G4 DNA-mediated regulation. *Nucleic Acids Res* **41**, 76-89 (2013).

45 Hershman, S. G. *et al.* Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic Acids Res* **36**, 144-156 (2008).

46 Du, Z., Zhao, Y. & Li, N. Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Res* **18**, 233-241 (2008).

47 Balasubramanian, S. & Huppert, J. L. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res* **35**, 406-413 (2006).

48 Endoh, T., Kawasaki, Y. & Sugimoto, N. Suppression of gene expression by G-quadruplexes in open reading frames depends on G-quadruplex stability. *Angew Chem Int Ed Engl* **52**, 5522-5526 (2013).

49 Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**, 8627-8637 (2015).

50 Hayflick, L. & Moorhead, P. S. The serial cultivation of human diploid cell strains. *Exp Cell Res* **25**, 585-621 (1961).

51 Hiyama, E. & Hiyama, K. Telomere and telomerase in stem cells. *Br J Cancer* **96**, 1020-1024 (2007).

52 Feng, J. *et al.* The RNA component of human telomerase. *Science* **269**, 1236 (1995).

53 Meyne, J., Ratliff, R. L. & Moyzis, R. K. Conservation of the human telomere sequence (TTAGGG)n among vertebrates. *Proc Nat Acad Sci* **86**, 7049-7053 (1989).

54 Chen, J. J.-L. *Telomerase Database*, <http://telomerase.asu.edu/sequences_telomere.html> (2017).

55 de Lange, T. Shelterin-mediated telomere protection. *Annu Rev Genet* **52**, 223-247 (2018).

56 Paeschke, K., Simonsson, T., Postberg, J., Rhodes, D. & Lipps, H. J. Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. *Nat Struct Mol Biol* **12**, 847-854 (2005).

57 Moye, A. L. *et al.* Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. *Nat Commun* **6**, 7643 (2015).

58 Crabbe, L., Jauch, A., Naeger, C. M., Holtgreve-Grez, H. & Karlseder, J. Telomere dysfunction as a cause of genomic instability in Werner syndrome. *Proc Nat Acad Sci* **104**, 2205-2210 (2007).

59 Cheung, H. H. *et al.* Telomerase protects werner syndrome lineage-specific stem cells from premature aging. *Stem Cell Reports* **2**, 534-546 (2014).

60 Chang, S. *et al.* Essential role of limiting telomeres in the pathogenesis of Werner syndrome. *Nature Genet* **36**, 877 (2004).

61 Jafri, M. A., Ansari, S. A., Alqahtani, M. H. & Shay, J. W. Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies. *Genome Med* **8**, 69 (2016).

62 Cahoon, L. A. & Seifert, H. S. Focusing homologous recombination: pilin antigenic variation in the pathogenic Neisseria. *Mol Microbiol* **81**, 1136-1143 (2011).

63 Cahoon, L. A. & Seifert, H. S. Transcription of a cis-acting, Noncoding, Small RNA Is Required for Pilin Antigenic Variation in Neisseria gonorrhoeae. *PLOS Pathog* **9**, e1003074 (2013).

64 Cahoon, L. A. & Seifert, H. S. An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. *Science* **325**, 764-767 (2009).

65 Giacani, L. *et al.* Comparative investigation of the genomic regions involved in antigenic variation of the TprK antigen among treponemal species, subspecies, and strains. *J Bacteriol* **194**, 4208-4225 (2012).

66 Walia, R. & Chaconas, G. Suggested role for G4 DNA in recombinational switching at the antigenic variation locus of the Lyme disease spirochete. *PLoS One* **8**, e57792 (2013).

67 Islam, M. K., Jackson, P. J., Rahman, K. M. & Thurston, D. E. Recent advances in targeting the telomeric G-quadruplex DNA sequence with small molecules as a strategy for anticancer therapies. *Future Med Chem* **8**, 1259-1290 (2016).

68 Campbell, N. H., Parkinson, G. N., Reszka, A. P. & Neidle, S. Structural basis of DNA quadruplex recognition by an acridine drug. *J Am Chem Soc* **130**, 6722-6724 (2008).

69 Nicoludis, J. M. *et al.* Optimized End-Stacking Provides Specificity of N-Methyl Mesoporphyrin IX for Human Telomeric G-Quadruplex DNA. *J Am Chem Soc* **134**, 20446-20456 (2012).

70 Parkinson, G. N., Ghosh, R. & Neidle, S. Structural basis for binding of porphyrin to human telomeres. *Biochem* **46**, 2390-2397 (2007).

71 Perrone, R. *et al.* A Dynamic G-Quadruplex Region Regulates the HIV-1 Long Terminal Repeat Promoter. *J Med Chem* **56**, 6521-6530 (2013).

72 Perrone, R. *et al.* Formation of a Unique Cluster of G-Quadruplex Structures in the HIV-1 nef Coding Region: Implications for Antiviral Activity. *PLOS ONE* **8**, e73121 (2013).

73 Grand, C. L. *et al.* The Cationic Porphyrin TMPyP4 Down-Regulates c-*MYC* and Human Telomerase Reverse Transcriptase Expression and Inhibits Tumor Growth *in Vivo*. *Mol Cancer Ther* **1**, 565-573 (2002).

74 Kim, M.-Y., Vankayalapati, H., Shin-ya, K., Wierzba, K. & Hurley, L. H. Telomestatin, a Potent Telomerase Inhibitor That Interacts Quite Specifically with the Human Telomeric Intramolecular G-Quadruplex. *J Am Chem Soc* **124**, 2098-2099 (2002).

75 Riou, J. *et al.* Cell senescence and telomere shortening induced by a new series of specific G-quadruplex DNA ligands. *Proc Nat Acad Sci* **99**, 2672-2677 (2002).

76 Sun, D. *et al.* Inhibition of human telomerase by a G-quadruplex-interactive compound. *J Med Chem* **40**, 2113-2116 (1997).

77 Papadopoulos, K. *et al.* Pharmacokinetic findings from the phase I study of Quarfloxin (CX-3543): a protein-rDNA quadruplex inhibitor, in patients with advanced solid tumors. *Mol Cancer Ther* **6**, B93-B93 (2007).

78 Harris, L. M. *et al.* G-Quadruplex DNA Motifs in the Malaria Parasite *Plasmodium falciparum* and Their Potential as Novel Antimalarial Drug Targets. *Antimicrob Agents and Chemother* **62**, e01828-01817 (2018).

CHAPTER 2

A guanine-flipping and sequestration mechanism for G-quadruplex unwinding by RecQ helicases

This work has been published:

Voter, A. F., Qiu, Y., Tippana, R., Myong, S. and Keck, J. L. (2018) A guanine-flipping and sequestration mechanism for G-quadruplex unwinding by RecQ helicases. *Nature Communications* 9, Article number:4201.

Andrew Voter performed the protein purification, crystallization, fluorescence polarization and circular dichroism experiments. Single-molecule experiments were performed by Yupeng Qiu, Ramreddy Tippana and Sua Myong.

Abstract

Homeostatic regulation of G-quadruplexes (G4s), four-stranded structures that can form in guanine-rich nucleic acids, requires G4-unwinding helicases. The mechanisms that mediate G4 unwinding remain unknown. We report the structure of a bacterial RecQ DNA helicase bound to resolved G4 DNA. Unexpectedly, a guanine base from the unwound G4 is sequestered within a guanine-specific binding pocket. Disruption of the pocket in RecQ blocks G4 unwinding, but not G4 binding or duplex DNA unwinding, indicating its essential role in structure-specific G4 resolution. A novel guanine-flipping and sequestration model that may be applicable to other G4-resolving helicases emerges from these studies.

Introduction

G-quadruplexes (G4s) are highly stable nucleic acid secondary structures that can form in guanine-rich DNA or RNA[1]. G-quartets, the repeating structures within G4s, are formed by an extensive hydrogen-bonding network that links four guanine bases around a cationic core. G4 structures, in turn, comprise G-quartets stacked upon one another, stabilized by base stacking between the layers. Their stability can make G4s impediments to numerous cellular processes, including replication[2], transcription[3] and translation[4]. Despite their potential hazards, G4-forming sequences are well represented in genomes, particularly within promoter regions[5] and telomeric DNA ends[6,7], indicating cells have developed mechanisms of abating the negative consequences of G4 DNA and have even co-opted the structures as regulatory and protective genomic elements.

G4 unwinding is essential for both G4 tolerance and G4 regulatory functions. Accordingly, cells have evolved a range of helicases that can unwind G4 structures, including DHX36[8], the Pif1[2] and XPD[9,10] families of helicases, and members of the RecQ helicase family including bacterial RecQ[11], yeast Sgs1[12], and human WRN[13] and BLM[14]. The importance of these helicases is highlighted by the profound genomic instability that results from their dysfunction, observed in xeroderma pigmentosa (XPD)[15], Fanconi anemia (FANCJ (an XPD paralog))[16], Werner (WRN)[17] and Bloom (BLM)[18] syndromes. In spite of the diverse clinical presentations caused by their absence, these enzymes operate on a range of G4 substrates using an apparent shared mechanism that relies on repetitive cycles of unwinding and refolding[19,20]. However, the small number of structural studies that have provided insights into the G4 unwinding process has limited our current understanding of the physical mechanisms underlying G4 resolution.

In this study, we report the X-ray crystal structure of the RecQ helicase from *Cronobacter sakazakii* (CsRecQ) bound to a resolved G4 DNA. Surprisingly, the 3′-most guanine base, which

is the first base in the quadruplex that the 3′-5′ translocating RecQ would encounter, is bound in a guanine-specific pocket (GSP) in the helicase core. Residues within the GSP satisfy all of the hydrogen bonds that are normally formed by guanines within G-quartet structures, which highlights the remarkable guanine selectivity of the binding site. Guanine docking within the GSP is incompatible with a folded G4 structure, implying that the base must flip from the quartet to be sequestered within the GSP. Consistent with an important and selective role for the GSP in G4 unwinding, changes to the guanine-coordinating residues in RecQ block G4 DNA unwinding but do not alter duplex DNA unwinding. These data lead to a guanine-flipping and sequestration model of G4 unwinding by RecQ helicases that may also be shared with other G4 unwinding helicases.

Results

Structure of RecQ bound to a resolved G4.

To better understand how G4 structures are resolved by helicases, the catalytic core domain of CsRecQ (Fig. 2.1a) was crystallized in the presence of G4 DNA that included a 3′ single-stranded (ss) DNA loading site. An earlier structure of CsRecQ bound to duplex DNA with a 3′ ssDNA loading site showed that the enzyme's helicase and winged-helix domains closed to form backbone interactions with the duplex, whereas the 3′ ssDNA end was bound in an electropositive channel in the helicase domain (Fig. 2.2)[21]. Because G4 and duplex DNA bind to the same surface of RecQ[22], we hypothesized that RecQ would bind G4 DNA in the same orientation.

Surprisingly, the 2.2 Å-resolution structure revealed a product complex of CsRecQ bound to unwound G4 DNA rather than a folded quadruplex (Fig. 2.1b; Table 2.1). The RecQ/G4 product structure was very similar to the RecQ/duplex DNA structure, with a root mean square deviation of 0.68 Å among 511 Cα atoms (Fig. 2.2). As was seen in the RecQ/duplex structure, the 3′ ssDNA is bound in an electropositive groove across the face of the helicase domain and it extends to dock in the ATP binding site of a symmetrically related molecule. Moreover, the helicase and winged-helix domains were closed around the unfolded G4. However, electron density was only observed for the three 3′-most guanines of the G4-forming DNA with the rest of the DNA apparently disordered within the crystal lattice. The positions of the resolved guanine bases deviated significantly from their expected placement within a folded G4, indicating that the quadruplex was unwound in the structure. The structure therefore suggested that binding by RecQ was sufficient to unwind G4 DNA, despite the presence of cations that otherwise stabilize the G4 (Supplementary Fig. 2.3).

RecQ contains a guanine-specific pocket

Examination of the structure revealed an unexpectedly specific arrangement for binding to the unwound G4 product (Fig. 2.1b-d). The 3′-most guanine base of the G4-forming sequence (G21), which is the first base within the folded G4 that would be encountered by the 3′-5′ translocating RecQ enzyme, was found sequestered in a guanine-specific pocket (GSP) on the surface of RecQ. The GSP forms hydrogen bonds with the guanine base using the sidechain hydroxyl and backbone carbonyl of Ser245 and the sidechain carboxyl group from Asp312 of RecQ (Fig. 2.1c). These contacts are uniquely selective for guanine and, strikingly, they substitute for all of the hydrogen bonds that stabilize guanines within G4 structures. The base is further stabilized by base stacking against a cytosine base two nucleotides 3′ of the flipped base (C23). The GSP is capped on the 5′ end by the hydrophobic portion of the Lys248 side chain and by Trp347 on the 3′ side (Fig. 2.1d). Lys222 and Lys248 make additional contacts with the phosphodiester backbone of the unfolded DNA, anchoring it against the helicase domain (Fig. 2.4). Given this arrangement, guanine binding to RecQ is incompatible with its position within a folded G4. Instead it appears that the guanine must "flip" from within a G-quartet to be sequestered in the RecQ GSP. In both DNA-free and duplex DNA-bound bacterial RecQ structures, access to the GSP is occluded by Lys248, which folds to interact with Asp312 from the GSP[21,23]. However, the GSP is open to accept the guanine base in the RecQ/G4 product complex (Fig. 2.5). These observations suggested a possible model in which guanine flipping and GSP-mediated base-specific sequestration support RecQ unwinding of G4 DNA.

Binding of RecQ variants to duplex and G4 DNA helicase substrates

A guanine-flipping and sequestration model predicts that sequence changes in the GSP would impair G4, but not duplex, DNA unwinding. To test this prediction and allow for

comparison with prior studies, *Escherichia coli* (Ec) RecQ (92.5% similar to CsRecQ, relevant residue numbering identical to CsRecQ) and CsRecQ catalytic core domain variants with compromised GSPs (Ser245Ala and Asp312Ala) were purified. The biochemical activity of these variants was tested alongside the wild-type EcRecQ and CsRecQ catalytic core domains. The CsRecQ Asp312Ala protein was unstable and difficult to purify, therefore this protein was excluded from analysis.

Affinity for FITC-labeled duplex DNA with a 3′ ss extension was measured first for the RecQ panel (Fig. 2.6a, Table 2.2). Each variant was found to bind the DNA, although the CsRecQ protein had lower affinities relative to their EcRecQ counterparts. The EcRecQ Asp312Ala variant had a ~3-4 fold higher affinity for the partial duplex DNA, which may be due to the removal of a negative charge in the duplex DNA binding groove. The DNA affinities reported here are consistent with those reported previously for the RecQ catalytic core[24].

Next, the affinity of each variant for G4 DNA with a 3′ ss extension was measured. EcRecQ, EcRecQ Asp312Ala and CsRecQ all bound G4 DNA with very similar affinities to those measure with the partial duplex (Fig 2.6b, Table 2.2). Unfortunately, we were unable to measure the equilibrium G4 affinity for either Ser245Ala variant; both were able to bind DNA but we observed a time-dependent decrease in anisotropy that made measurement of the binding constant impossible. This is likely due to a modest instability/insolubility of the variants in the conditions tested. Nevertheless, each of the variants could bind G4 and duplex DNA, indicating that residues within the GSP are not essential for G4 binding.

Disruption of the GSP inhibits G4 but not duplex unwinding

Single-molecule (sm) FRET assays were carried out to determine the impact of GSP sequence changes on RecQ DNA unwinding. These assays were designed to test unwinding of substrates with a 3′ ss loading site that contain either a duplex structure alone or a duplex structure preceded by a G4 element (Fig. 2.7a and 2.7e, respectively). The substrates consist of an immobilized Cy5-labeled 18mer annealed to a Cy3-labeled strand comprising the complementary 18mer along with either dT_{15} or both a G4 element and dT_{15}. Unwinding of the substrate releases the Cy3-containing DNA strand and can be measured as a reduction of the number of Cy3 spots over time (Fig. 2.7b).

In this assay, both EcRecQ (Fig. 2.7c-d) and CsRecQ (Fig. 2.8) were able to unwind substrates containing either of two antiparallel G4 DNAs, TTA-T_{15} (5′-TTA GGG TTA GGG TTA GGG TTA GGG-3’) or TAA-T_{15} (5′-GGG TAA GGG TAA GGG TAA GGG-3′) (Table 2.2), using cycles of repetitive unwinding and refolding shown in the single molecule trace (Fig. 2.7c, top, Fig. 2.8). Repetitive unwinding/refolding cycles are marked by time-resolved high-amplitude FRET change signatures, such as that observed from ~30 to ~65 seconds with EcRecQ in Fig. 2.7c. Neither EcRecQ nor CsRecQ were active against cMyc, a parallel G4 DNA.

In contrast to the results with the wild-type RecQ proteins, none of the GSP variant RecQ proteins were able to unwind the G4 DNA structures. Single molecule traces (Table 2.2, Figure 2.7c, bottom, Fig. 2.8) showed that each of the GSP variants failed to elicit the repetitive unwinding/refolding FRET signature observed with the wild-type RecQ proteins and G4 unwinding was not observed, even after long (12 minute) incubation periods. These data are consistent with an essential role of the RecQ GSP in G4 unwinding.

To test whether the GSP RecQ variants retained duplex helicase activity, the assay was repeated using a substrate that lacked the G4-forming sequence (Fig. 2.7e). The single-molecule

traces (Fig. 2.7f, Fig. 2.9) and FRET histograms before and after the addition of the proteins (Fig. 2.7g), demonstrate robust helicase activity of the duplex DNA substrate by all of the variants. Each protein unwinds the DNA at rates that were very similar to those observed with EcRecQ and CsRecQ (Table 2.2). Thus, the GSP in RecQ is required uniquely for unwinding G4 DNA.

In an attempt to visualize folded G4 DNA bound to RecQ, crystals of the Ser245Ala CsRecQ catalytic core variant were generated with G4 DNA. Diffraction data were collected from over a dozen crystals and molecular replacement revealed several crystals in which the guanine base was not found in the altered GSP. In these cases, discontinuous electron density consistent with the dimensions of a folded G4 structure was observed in the cleft formed by the helicase and winged-helix domains (Fig. 2.10). Unfortunately, the fragmented nature of the electron density did not permit modeling of the full G4 structure. Nonetheless, the structural study was consistent with the significantly reduced activity of the variant predicted from the FRET experiment.

Discussion:

Despite the importance of G4 homeostasis in cells, our mechanistic understanding of quadruplex resolution has been hampered by a lack of structural information for G4-processing helicases. In this report, we have described the X-ray crystal structure of a RecQ helicase bound to a resolved G4. The structure identified a guanine-specific pocket, or GSP, in RecQ that sequesters a guanine base from the resolved G4. Guanine is selectively bound within the GSP via residues that form a pattern of hydrogen bond donors and acceptors that mimic the bonding pattern for a guanine within a G-quartet structure. As such, guanine binding to the GSP is incompatible with a folded G4 structure and instead requires the base to be flipped away from the G4. These observations suggested a possible role for the GSP in G4 unwinding. In agreement with such a role, RecQ variants with altered guanine-binding residues failed to unwind G4 DNA, but they maintained their ability to unwind duplex DNA. Our data collectively support an unexpectedly specific helicase mechanism for RecQ unwinding of G4 structures that relies on guanine base flipping and sequestration for G4 resolution.

In the G4 unwinding model, RecQ first recognizes a ssDNA/G4 junction, placing the G4 in a position adjacent to the GSP and leaving the pocket poised to receive the 3′-most guanine from a G-quartet as it flips from the folded structure (Fig. 2.11). For the structural studies described here, guanine sequestration appears sufficient to unfold a G4 with three guanine quartet planes. ATP-dependent RecQ translocation would then slide the 3′-most guanine base out of the GSP, moving it along the face of the helicase domain and allowing the next guanine to be sequestered within the GSP as the G4 structure is resolved. What then gives rise to the repetitive cycles of G4 unwinding and refolding that have been observed in single-molecule experiements[11,14]? Two possibilities may explain this phenomenon. First, since RecQ must release the first guanine to

advance along the DNA, it may be that the base can either slide along the ssDNA binding face of RecQ to promote unwinding or it can flip back and allow the G4 structure to refold (Fig. 2.11). It is possible that G4 reformation is more efficient than processive translocation, which would lead to repetitive rounds of unwinding and refolding. Second, although the GSP matches the hydrogen bonding pattern for a guanine in a folded G4, it may form a complex that is less stable than that found in the context of a G4, which includes base stacking and ionic stabilization in addition to hydrogen bonding. If RecQ transiently captures a frayed guanine from the 3′ end of the G4 and if translocation is slower than the rate at which the guanine can transition back into the folded G4, this difference could allow the captured guanine to be released and the G4 to reform, resulting in a cycle of G4 unwinding and refolding.

Base-flipping activities have been observed in several enzymes that act on nucleic acids, including polymerases[25], endonucleases[26], glycosylases[27], and methyltransferases[28]. In these enzymes, base flipping is accompanied by a distortion of B-form DNA near the flipped base, facilitating extraction of the base by the enzyme while extensive protein-DNA contacts hold the enzyme in position. Similarly to RecQ, base flipping enzymes coordinate the isolated nucleobase through a hydrogen bonding pattern that selects for the targeted base. This specificity allows repair enzymes, for example, to survey the integrity of the flipped base prior to initiation of a repair process. RecQ binding may similarly distort G4 DNA to allow guanine base flipping. It is also possible that RecQ simply traps transiently frayed guanine bases at the ssDNA/G4 interface. Additional studies are needed to examine these possibilities.

Because the RecQ GSP is specific for a canonical base, it is possible that the GSP may inadvertently sample guanines outside of G4 structures, hindering RecQ unwinding of guanine-rich duplex DNA. Indeed, RecQ pauses have been observed while unwinding GC-rich duplex

DNA[29], which could possibly result from guanine occupancy in the GSP. However, examination of the structure of the GSP reveals a mechanism that appears to counteract such non-productive base-flipping. In the absence of G4 DNA, Lys248 and Asp312 interact with one another to occlude access to the GSP (Fig. 2.5). This closure is maintained when RecQ is bound to duplex DNA[21]. However, interaction with resolved G4 DNA appears to favor GSP opening through an interaction formed between Lys248 and the phosphodiester backbone of the G4 product. This interaction could make the GSP accessible to guanine bases under conditions where resolved or, presumably, folded G4 DNA is bound to RecQ. This interaction may attenuate guanine binding by the GSP during duplex DNA unwinding while promoting it during G4 unwinding.

It remains to be seen how prevalent a guanine base-flipping mechanism is among G4 helicases. Among the bacterial RecQ helicases, the GSP sequence is conserved but not invariant. Some variability may be tolerated in the GSP while still allowing for G4 helicase activity. It may also be the case that the GSP is structurally conserved, even if the sequence homology is not invariant. For example, examination of the structure of BLM helicase, a human RecQ homolog with G4 helicase activity, reveals a potential GSP situated at the duplex/ssDNA junction comprising Ser965 and either Glu900 or Asp 997 (Fig. 2.12a-b)[30]. We are unable to assess if the other RecQ G4 helicases WRN or Sgs1 possess a GSP due to the lack of structures of their catalytic cores. However, even outside of the RecQ family, GSP-like pockets can be found. One instance is the bacterial helicase UvrD, which also contains a GSP-like structure poised to potentially receive a guanine flipped from a G4 substrate (Fig. 2.12c)[31].

While base-flipping described here provides a simple method of G4 resolution, other mechanisms may also exist. A very recent structure of G4 DNA in complex with the helicase DHX36 has been reported, suggesting a mechanism of G4 resolution in which the G4 is bound by

the extended N-terminal DHX specific motif (DSM)[32]. This binding triggers repetitive conformational shifts in the G4 that are thought to reorganize and destabilize the quadruplex before ultimately releasing the resolved DNA in an ATP-dependent manner. The broader applicability of this mechanism may be limited to proteins with a DSM or analogous domain. Furthermore, the DSM best recognizes and unfolds parallel G4s, whereas this not a requirement of the GSP mechanism. Indeed, different RecQ helicases are known to unwind both parallel and antiparallel G4s[20,33].

In summary, our studies have identified a remarkably specific mechanism for G4 DNA unwinding by RecQ DNA helicases. This model relies on base flipping in a manner that was first envisioned as a possible helicase mechanism shortly after the discovery of enzyme-mediated DNA base flipping[34], although experimental evidence for such a mechanism has been lacking prior to the structural work described here. Discovery of this novel mechanism also underscores the apparent importance of G4 regulation by helicases *in vivo*. In what ways do the G4-specific functions of RecQ helicases impact cells? Several RecQ pathways have been linked to recognition and/or processing of G4 structures, including those involved in recombination regulation[35] and telomere maintenance[36] in eukaryotes, and antigenic variation in bacteria[37]. Investigations of the cellular activities of RecQ variants with selectively-blocked G4 resolution functions could pave the way to a better understanding of the general roles of G4 structures *in vivo*.

Methods

Protein purification

The catalytic core of CsRecQ and EcRecQ and all variants were purified as previously described[21]. Briefly, proteins were overexpressed in Rosetta 2 (DE3) *E. coli* cells were transformed with pLysS (Novagen, Darmstad, Germany) and a RecQ overexpression plasmid. Cells were grown at 37°C in Luria Broth supplemented with 50 μg/mL kanamycin and 1 μg/mL chloramphenicol. Once the cells reached an OD_{600} of 0.6, protein expression was induced with 1 mM IPTG for 4 hours at 37°C before the cells were pelleted and stored at -80°C. Cell pellets were resuspended in lysis buffer (20 mM Tris·HCl (pH 8.0), 500 mM NaCl, 1 mM 2-mercaptoethanol (BME), 1 mM phenylmethane sulfonyl fluoride, 100 mM dextrose, 10% (w/v) glycerol, 15 mM imidazole), lysed by sonication and clarified by centrifugation. The supernatant was incubated with Ni-NTA agarose resin at 4°C before being washed extensively with lysis buffer. The N-terminally His-tagged proteins were eluted from the resin with elution buffer (lysis buffer containing 250 mM imidazole) before the His tag and HRDC domains were removed by overnight thrombin cleavage while the protein was dialyzed into dialysis buffer (20 mM Tris·HCl (pH 8.0), 300 mM NaCl, 1 mM BME, 10% (w/v) glycerol). The cleaved protein was diluted to 100 mM NaCl, loaded onto a HiPrep QFF ion exchange column (GE healthcare, Chicago, IL) and eluted with a 0.1-1M NaCl gradient. RecQ-containing fractions were pooled, concentrated and then further purified with an S-100 size exclusion column (GE healthcare) before dialysis into storage buffer (20 mM Tris·HCl (pH 8.0), 1 M NaCl, 4 mM BME, 40% (w/v) glycerol, 1 mM ethylenediaminetetraacetic acid) and stored at -20°C.

Structural studies

HPLC-purified DNA for crystallographic and RecQ-G4 binding studies (G4 DNA, 5′-TTA GGG TTA GGG TTA GGG TTA GGG TCG GTG CCT TAC T-3′) was purchased from Integrated DNA Technologies (Coralville, IA, USA). Oligonucleotides were resuspended in 18 MΩ H_2O and stored at -20°C. CsRecQ catalytic core or the Ser245Ala variant at 6.5 mg/mL in minimal buffer [10 mM Tris·HCl (pH 8.0), 1 M ammonium acetate] was mixed with G4 forming sequence at a 1:1.2 protein:DNA ratio. The complex was combined at a 1:1 (vol/vol) ratio with mother liquor [70 mM sodium acetate·acetic acid (pH 4.9), 30% (vol/vol) glycerol, 10% (wt/vol) PEG 4000], and crystals were formed by hanging-drop vapor diffusion then flash-frozen in liquid nitrogen.

X-ray diffraction data were collected at the Advanced Photon Source (LS-CAT beamline 21ID-F) and were indexed and scaled using HKL2000[38]. The structure of the CsRecQ/G4 DNA complex was determined by molecular replacement using the CsRecQ/duplex DNA structure (PDB ID code 4TMU)[21] as a search model in the program Phaser[39] followed by rounds of manual fitting using Coot[40] and refinement using PHENIX[41]. Coordinate and structure factor files have been deposited in the Protein Data Bank (PDB ID code 6CRM). The Ser245Ala CsRecQ variant was phased by molecular replacement using the CsRecQ/G4 product complex as a search model in the program Phaser[39] followed by rounds of manual fitting using Coot[40] and refinement using Phenix[41].

DNA-binding assay

G4 DNA containing a 3′ FAM modification (F-G4) was solubilized to 50 μM in G4 folding buffer [10 mM Tris·HCl (pH 8.0), 100 mM KCl]. Using a heat block, the DNA was heated to 95°C for 5 minutes, after which the block was removed from heat and allowed to cool to room

temperature over approximately 4 hours. Folded DNA was then stored at 4°C. RecQ proteins were serially diluted from 20,000 to 0.6 nM in G4 binding buffer [20 mM Tris·HCl (pH 8.0), 100 mM NaCl, 1 mM $MgCl_2$, 1 mM β-mercaptoethanol, 0.1 mg/mL bovine serum albumin, 4% (vol/vol) glycerol], then incubated with 5 nM F-G4 for 30 minutes at room temperature in a total volume of 100 µL. The fluorescence anisotropy of each sample was measured at 25°C with a Beacon 2000 fluorescence polarization system. Measurements are reported in duplicate and error bars represent 1 SEM. Binding affinities and uncertainties were determined using Prism version 5.0c (GraphPad Software, La Jolla, CA, USA). Duplex binding assays were performed as the G4 binding assays using a 3' fluorescein-labeled ssDNA (5′-GCG TGG GTA ATT GTG CTT CAA TGG ACT GAC-3′) annealed to an unlabeled 18-mer (5′-AAG CAC AAT TAC CCA CGC-3′) to create a substrate with an 18-bp duplex with 3' overhang of 12 nucleotides.[21] Duplex binding assays were performed in triplicate and error bars represent 1 SEM.

smFRET DNA substrates

ssDNAs with amino modifier at the labeling sites were purchased from Integrated DNA Technologies (Coralville, IA, USA). The DNAs were labeled using Cy3/Cy5 monofunctional NHS esters (GE Healthcare, Princeton, NJ, USA). Amino modified oligonucleotides (10 nmol in 50 ml ddH2O) and 100 nmol of Cy3/Cy5 NHS ester dissolved in dimethylsulphoxide were combined and incubated with rotation overnight at room temperature in the dark. The labelled oligonucleotides were purified by ethanol precipitation.

Both G4 and non-G4 substrates consist of 18 base pairs of dsDNA and a 3′ tailed ssDNA of specific sequence. For non-G4 DNA substrate, the 18mer DNA is immediately followed with a tail of dT_{18}. For G4 DNA substrates, a G4 sequence is between the 18mer dsDNA and the dT tail. A Cy5-Cy3 FRET pair are placed at the junction and the 3′ end of the ssDNA, respectively.

The sequences are as follows:

Common 18mer: 5′-Cy5-GCC TCG CTG CCG TCG CCA-biotin-3′

Non-G4 DNA, T18: 5′-TGG CGA CGG CAG CGA GGC-$(T)_{18}$-Cy3-3′

TTA-T15 DNA: 5′-TGG CGA CGG CAG CGA GGC TTA GGG TTA GGG TTA GGG TTA GGG-$(T)_{15}$-Cy3-3′

TAA-T15 DNA: 5′-TGG CGA CGG CAG CGA GGC TTG GGT AAG GGT AAG GGT AAG GG-$(T)_{15}$-Cy3-3′

c-myc-T15 DNA: 5′-TGG CGA CGG CAG CGA GGC GGG TGG GTA GGG TGG G-$(T)_{15}$-Cy3-3′

DNA substrates were annealed by mixing the biotinylated and non-biotinylated oligonucleotides in a 1:2 molar ratio in T50 buffer [10 mM Tris·HCl (pH 8.0), 50 mM NaCl]. The final concentration of the mixture is 10 μM. The mixture was then incubated at 95°C for 2 minutes followed by slow cooling to room temperature to complete the annealing reaction in just under two hours. The annealed DNAs were stored at -20°C and were diluted to 10 nM single molecule stock concentration in K100 buffer [10 mM Tris·HCl (pH 8.0), 100 mM KCl] at the time of experiment.

smFRET unwinding assays

A custom-built total internal reflect fluorescence microscope was used for the single-molecule unwinding assays. A solid state 532nm laser (75mW, Coherent CUBE) is used to excited the donor dye in the Cy3-Cy5 FRET pair used in FRET experiments. Emitted fluorescence signals collected by the microscope are separated by a dichroic mirror with a cutoff of 630nm to split the

Cy3 and Cy5 signals, which are then detect on an EMCCD camera (iXon DU-897ECS0-#BV; Andor Technology). Custom C++ programs control the camera and IDL software and are used to extract single molecule traces from the recorded data. The traces are displayed and analyzed using Matlab and Origin software. All homemade codes are in the smFRET package available at the Center for the Physics of Living Cells (https://cplc.illinois.edu/software/, Biophysics Department, University of Illinois at Urbana-Champaign).

All unwinding experiments were performed in RecQ Reaction Buffer [20 mM Tris·HCl (pH 7.5), 50 mM KCl, 3 mM $MgCl_2$, 1 mM ATP] with an oxygen scavenging system containing 0.8% v/v dextrose, 1 mg/ml glucose oxidase, 0.03 mg/ml catalase1, and 10mM Trolox. All chemicals were purchased from Sigma Aldrich (St. Louis, MO).

Biotinylated FRET DNA (50 to 100 pM) were immobilized on polyethylene glycol-coated quartz surface via biotin-neutravidin linkage. RecQ and mutant proteins (100 nM) were added at room temperature to initiate unwinding. 10-20 short movies (10 seconds) and 3-4 long movies (3 minutes) were then taken monitoring the Cy3 and Cy5 emission intensities over time. These are then analyzed to produce the FRET histograms and trajectories to monitor any unwinding activity.

To calculate the unwinding rate, note that as the DNA is unwound, the Cy3 strand is freed from the immobilized DNA substrate and the Cy3 signal disappears. Snapshots of the Cy3 spots detected in an imaging area are taken via short movies (2 seconds) and the spots counted over time. The counts are then plotted and fitted to an exponential curve to obtain the rate of disappearance of the Cy3 spots over time as the indication of unwinding. For each rate calculation, 400-500 single molecules were monitored and the standard error of the measurement was reported. During imaging, a fraction of the G4 molecules were unwound by a protein-dependent and GSP-independent mechanism. The number of G4s lost during through this process (~20% over 12

minutes) was insufficient to allow for rate calculations and the GSP-independent unwinding was assumed to be negligible relative to the GSP-dependent mechanism.

Circular dichroism

G4 DNA used for the crystallographic studies were refolded by diluting to 10 μM in either 10 mM Tris·HCl (pH 8.0) or 35 mM sodium acetate-acetic acid, 500 mM ammonium acetate, 4% (w/v) PEG 4K and 15% (v/v) glycerol by heating to 95°C for 10 minutes and slowly cooling to room temperature. These conditions represent unfolded ssDNA or crystallization conditions, respectively. CD spectra were recorded on an AVIV 420 circular dichroism spectrometer at 25°C over a range of 200-340 nm in a 1-mm path length quartz cuvette. Data were collected using a 1 nm step size with a 5 second average and a blank reading containing no DNA was subtracted from each reading.

Acknowledgments. We thank the staff of the Advance Photon Source (LS-CAT beamline) and Ken Satyshur for assistance with data collection and analysis. We also thank members of the Keck laboratory for critical reading of this manuscript. This work was funded by NIH R01 GM098885 to J.L.K. A.F.V. was supported by NIH F30 CA210465 and T32 GM008692. This research used resources of the Advanced Photon Source, a US department of Energy Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory Under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817).

Author contributions. A.F.V. purified the proteins and carried out DNA binding and circular dichroism experiments. A.F.V. and J.L.K. performed the structural analysis. Y.Q., R.T. and S.M. designed and carried out single-molecule experiments. A.F.V, Y.Q., R.T., S.M. and J.L.K participated in data analysis and wrote the article.

Author information: Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Correspondence and requests for materials should be addressed to J.L.K. (jlkeck@wisc.edu)

Data and materials availability: The RecQ/G4 product structure is available at the Protein Data Bank, PDB ID: 6CRM.

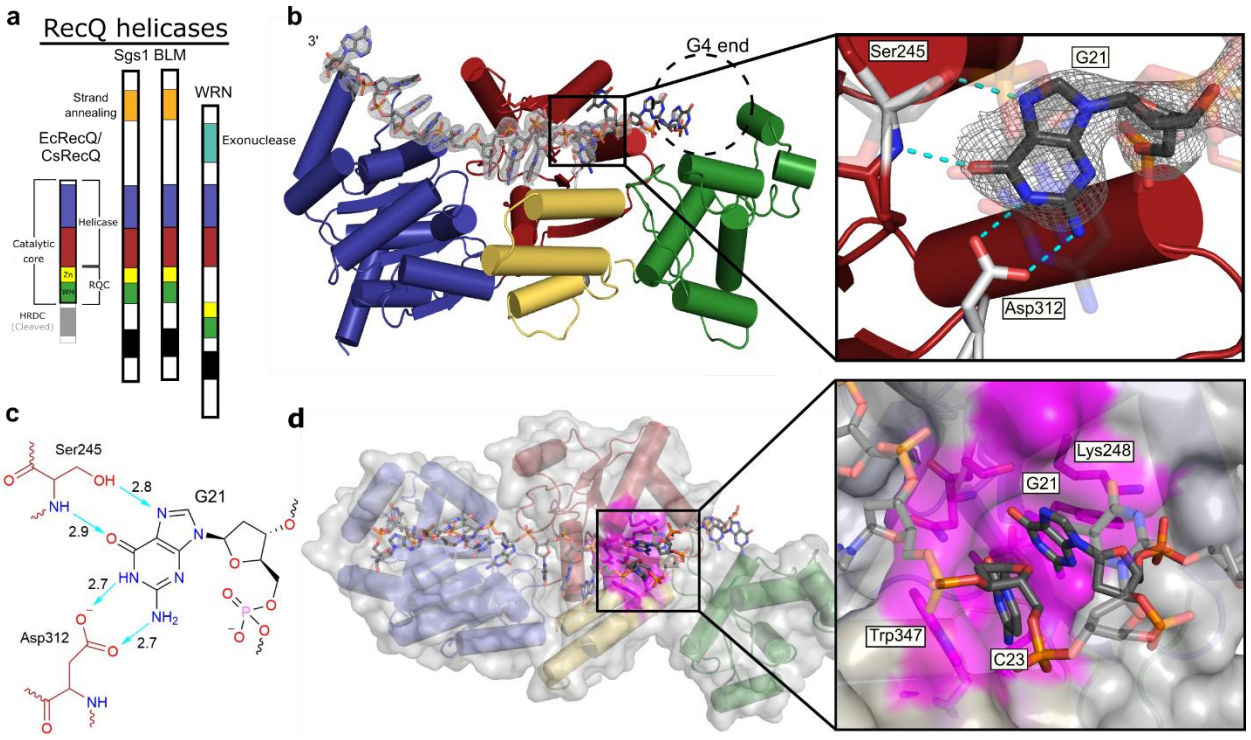


Figure 2.1. The guanine-specific pocket of the CsRecQ helicase. a) Domain schematic representation of RecQ helicase family. The C-terminal domain of the catalytic core (RQC) contains a Zn^{2+}-binding domain (Zn) and a winged-helix domain (WH). b) Crystal structure of CsRecQ bound to resolved G4 DNA. Domain colors correspond to panel 1a. F_o-F_c omit electron density contoured at 2.0σ is shown. The expected location of the G4 is highlighted. (Insert) The GSP in RecQ binds the flipped guanine with high specificity. Hydrogen bonds are represented by dashed lines. c) Ligand interaction diagram of the GSP/guanine interface. Bond distances in Å are shown. d) Surface representation of the CsRecQ bound to the resolved G4 with the GSP colored in magenta. (Insert) The flipped G21 is stabilized by hydrophobic interactions and base stacking with C23.

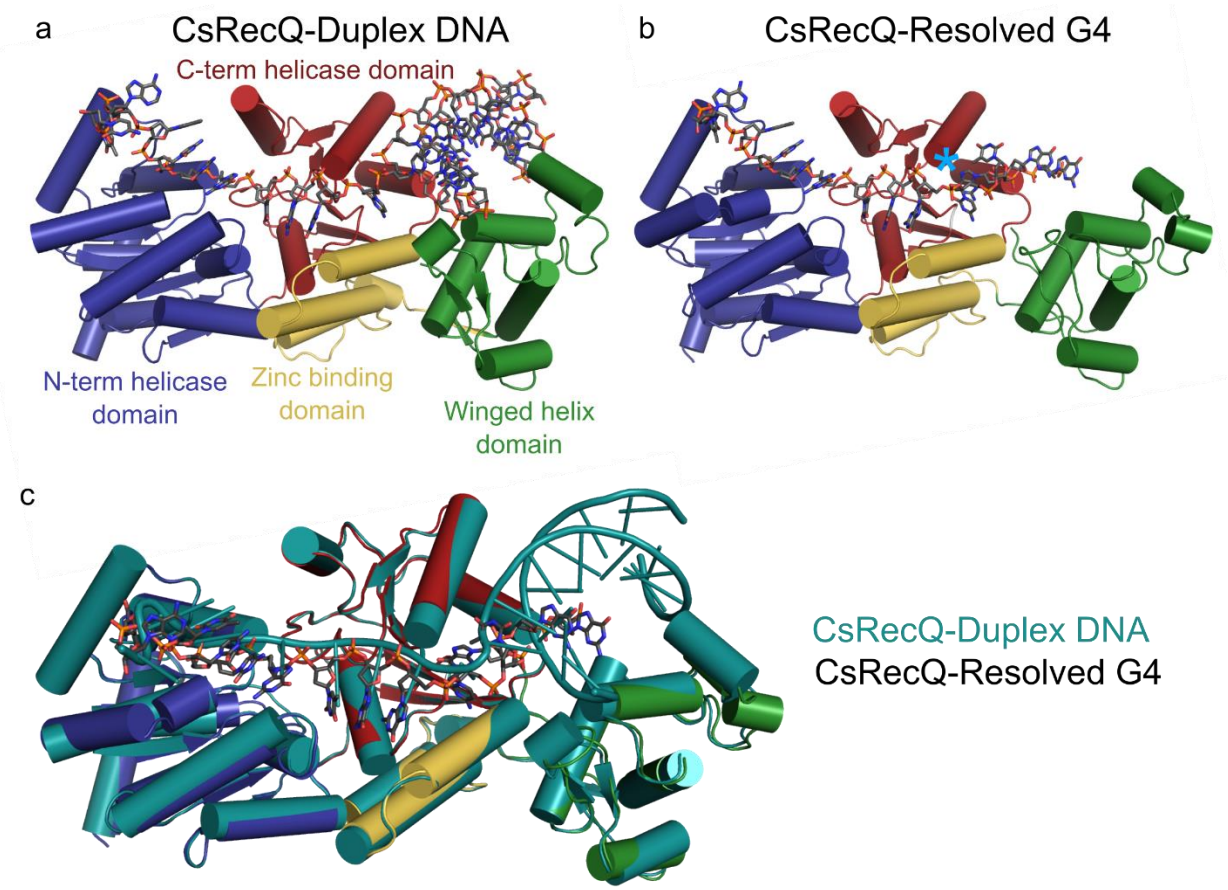


Figure 2.2. RecQ adopts a similar confirmation when bound to duplex DNA or a resolved G4. a) Structure of CsRecQ bound to duplex DNA (PDB 4TMU)[21]. b) Structure of CsRecQ bound to a resolved G4 (PDB 6CRM). The flipped guanine is marked by a teal asterisk c) Overlay of the structure from a and b. The CsRecQ-Duplex DNA is colored in teal with the CsRecQ/G4 product structure colored as in b.

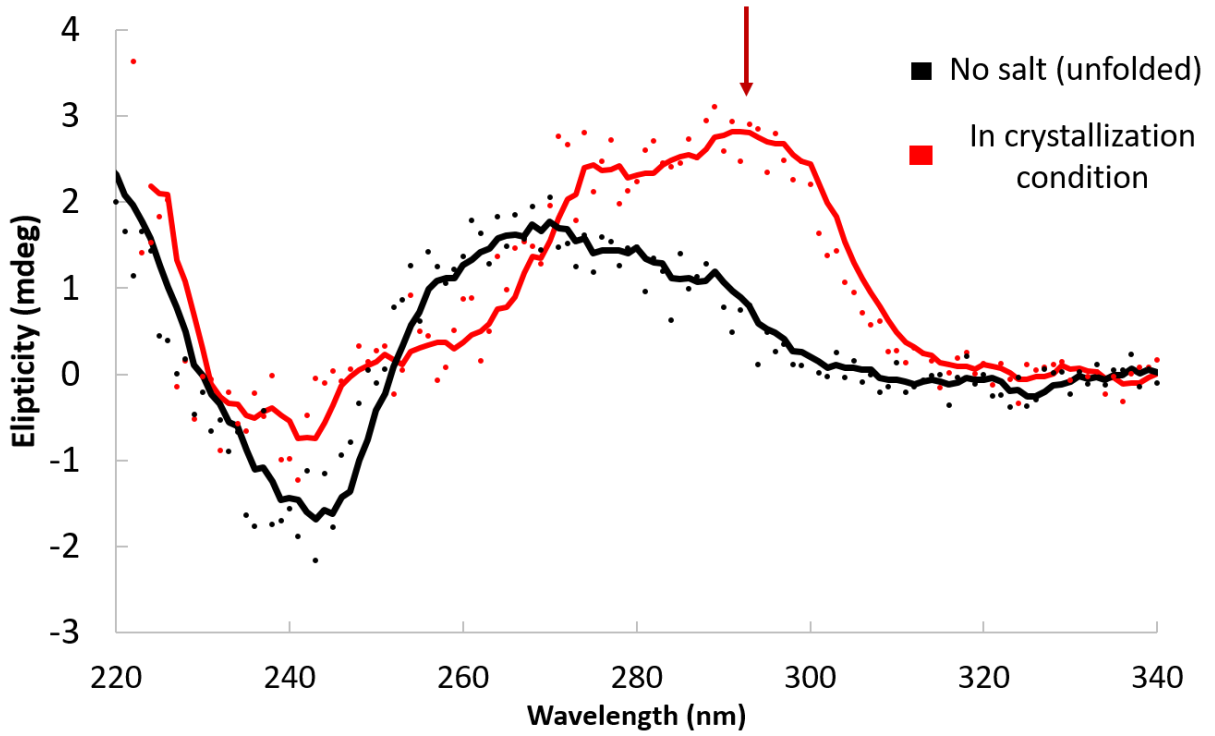


Figure 2.3. Crystallization conditions support antiparallel G4 formation. Circular dichroic spectra of G4-folding DNA in either no salt conditions (black) or the crystallization conditions (red). Line represent the rolling average of the CD values shown by the individual points. Presence of a positive CD signal at 290 nm (arrow) indicates the formation of antiparallel G4 DNA.

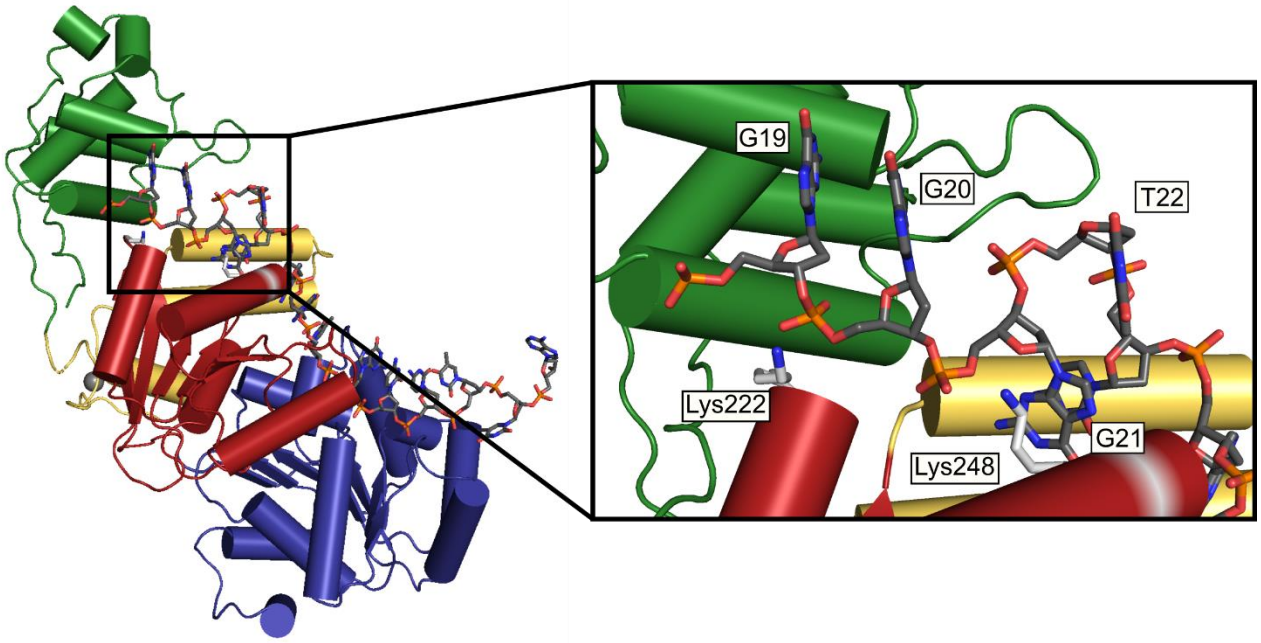


Figure 2.4. Interaction between RecQ and the resolved G4. The resolved G4 DNA (gray sticks) is held against the face of the CsRecQ helicase domain (red cartoon) by Lys222 and Lys248 (white sticks).

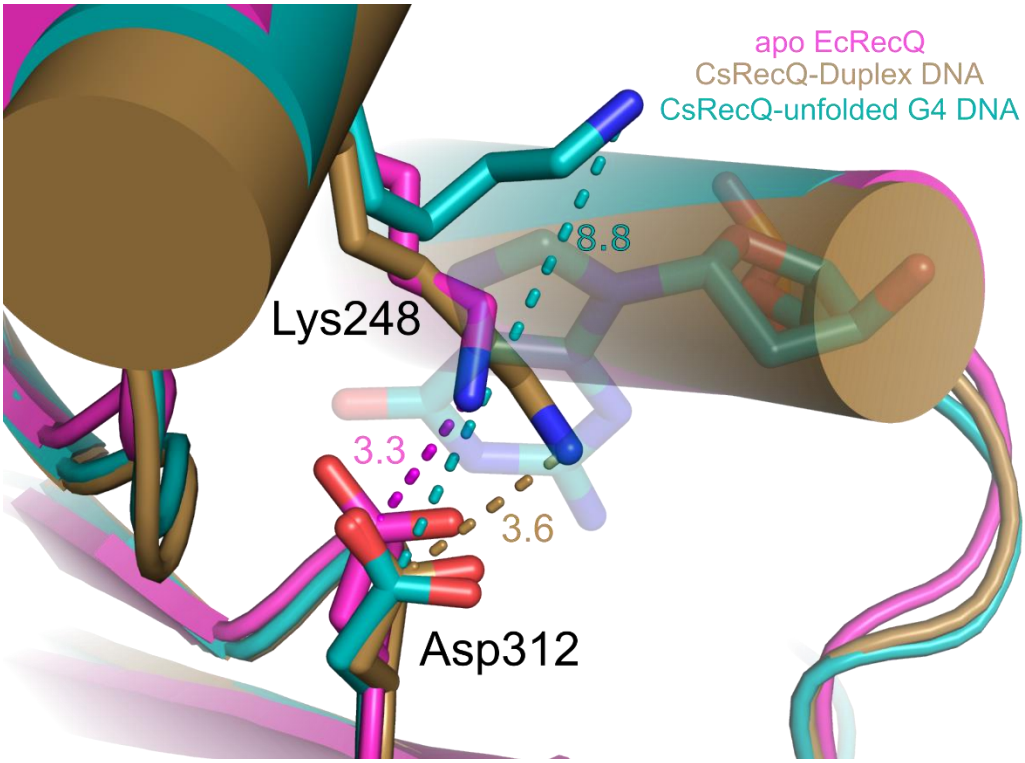


Figure 2.5. The RecQ GSP is closed unless bound to resolved G4 DNA. The GSP in both apo EcRecQ (magenta, PDB 1OYW[23]) and duplex DNA-bound CsRecQ (yellow, PDB 4TMU[21]) is closed by interaction between Lys248 and Asp312, but is opened in the G4 product complex (teal, PDB 6CRM). Dashed lines and distances (Å) denote the distances between the ε-amino group of Lys248 and the carbon of the carboxylic acid group of Asp312.

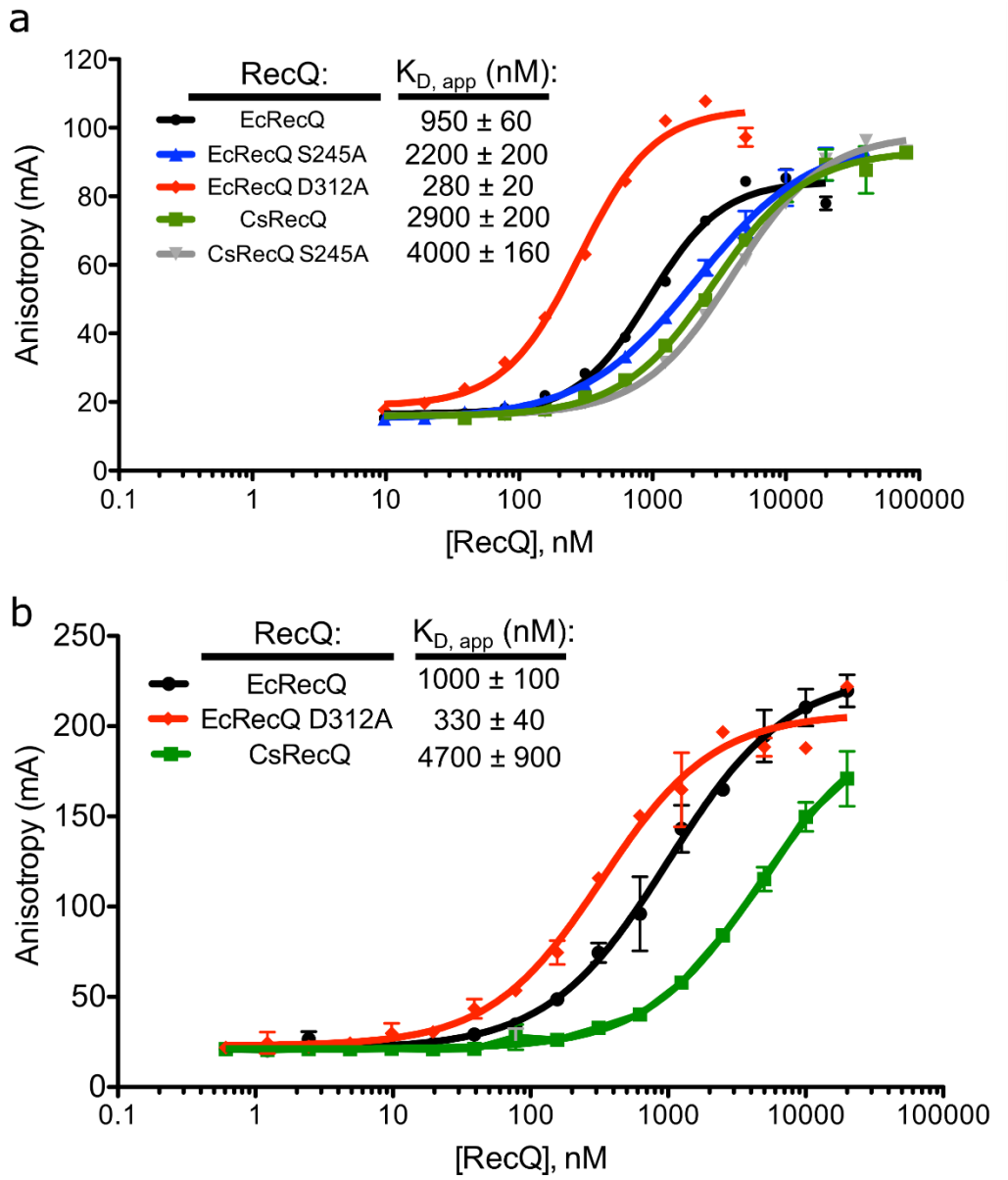


Figure 2.6. Duplex and G4 binding of RecQ variants. a) Fluorescence anisotropy of folded G4 DNA incubated with increasing concentrations of each RecQ variant. Error bars represent the SEM of 2 replicates. b) Fluorescence anisotropy of duplex DNA substrates incubated with increase concentrations of each RecQ variant. Error bars represent the SEM of 3 replicates.

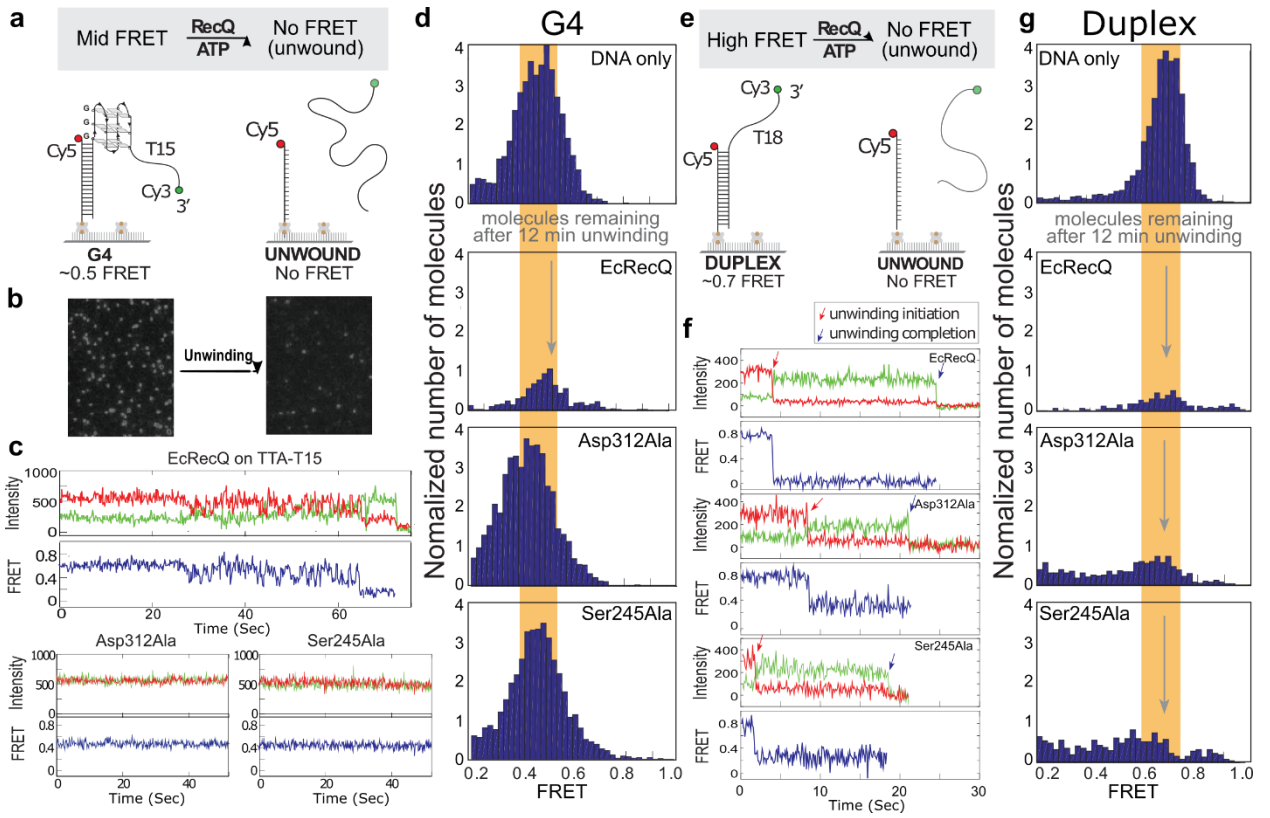


Figure 2.7. smFRET studies of RecQ helicase activity. a) smFRET strategy to monitor RecQ-mediated unwinding of G4 DNA. b) Representative field showing the loss of FRET signal following RecQ unwinding. c) Representative smFRET traces of EcRecQ and the RecQ GSP variants on G4 DNA. The top traces for each RecQ variant represent the tethered Cy5 (red) and annealed Cy3 (green) signal, while the lower blue trace denotes the FRET efficiency. d) Histograms of the smFRET signals for ~5000 G4 DNA molecules after a 12-minute incubation with the specified RecQ variant. The orange bar denotes the primary FRET peak. e) smFRET strategy to monitor RecQ-mediated unwinding of duplex DNA. f) Representative smFRET traces of the action of EcRecQ and the RecQ GSP variants on the duplex DNA substrate. Traces are colored as in panel c. g) Histograms of the smFRET signals for ~5000 duplex DNA molecules after a 12-minute incubation with the specified RecQ variant. The orange bar denotes the primary FRET peak.

Figure 2.8. Representative single molecule traces of G4 unwinding by RecQ variants.

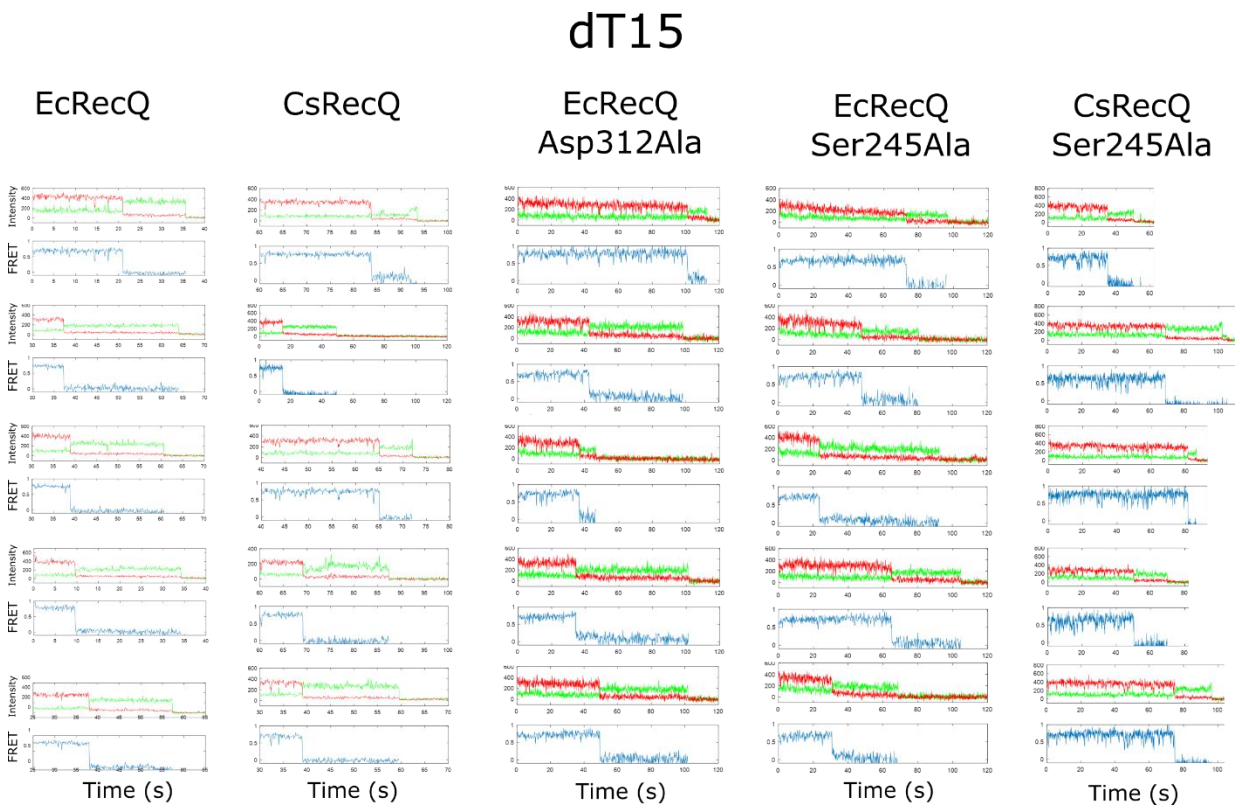


Figure 2.9. Representative single molecule traces of duplex unwinding by RecQ variants.

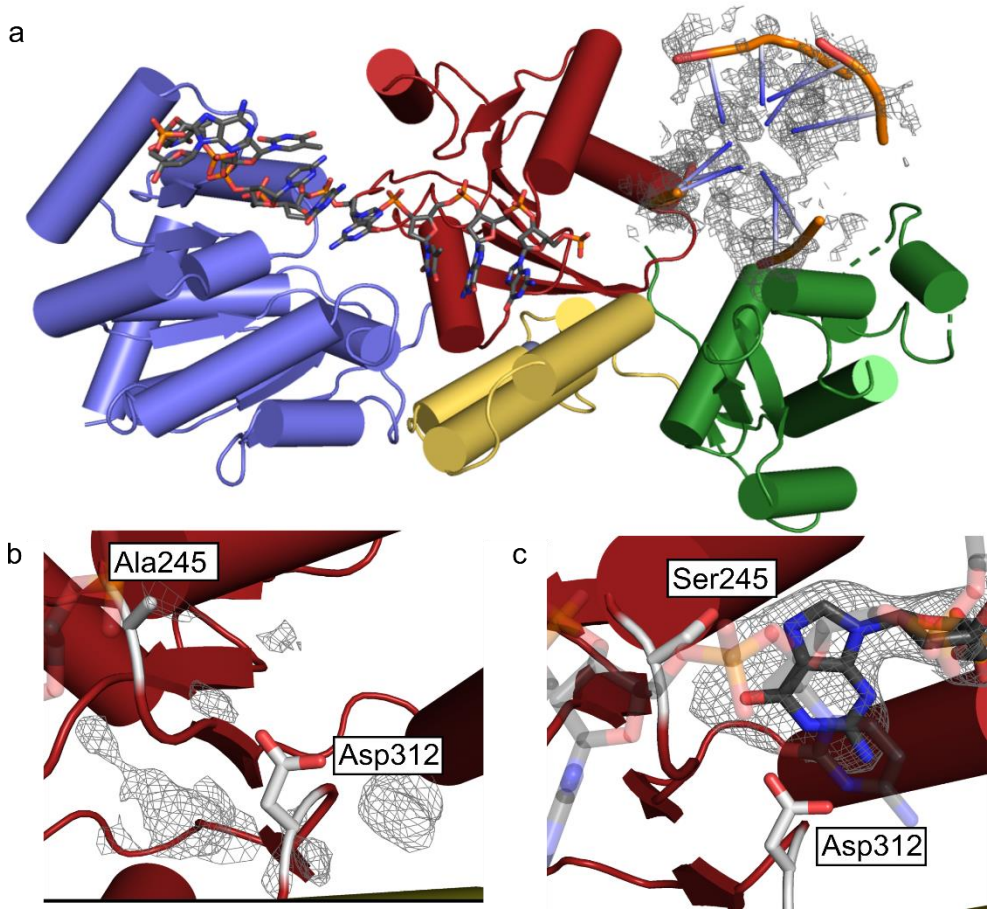


Figure 2.10. Structure of Ser245Ala CsRecQ bound to G4 DNA. a) F_o-F_c electron density map (contoured at 1.7σ) showing significant but discontinuous electron density within the helicase/winged-helix domain cleft. A G4 structure (PDB 143D)[42] is shown within the map for scale. b). F_o-F_c electron density map of the CsRecQ Ser245Ala (contoured at 2.0σ) showing that a guanine is not found in the GSP. c) F_o-F_c omit electron density map of the CsRecQ/G4 product complex (contoured at 2.0σ) with a flipped guanine for comparison.

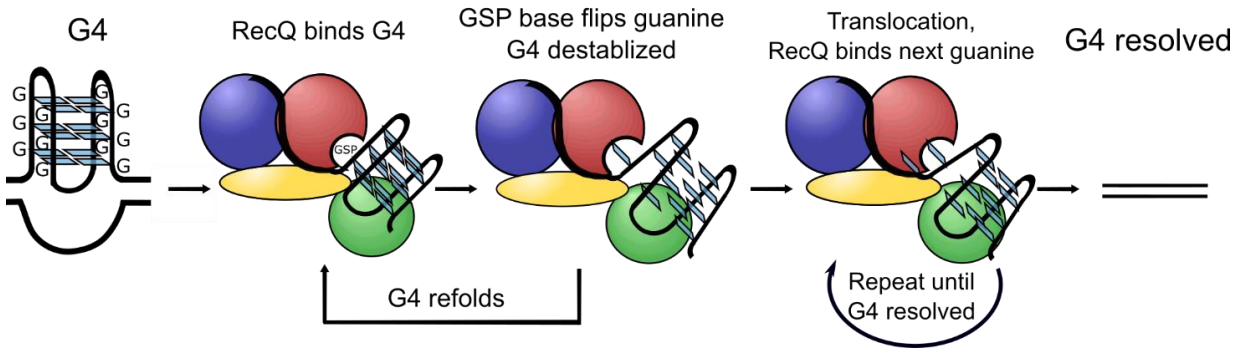


Figure 2.11. Model of RecQ-mediated G4 unwinding. RecQ binds the folded quadruplex, trapping it between the helicase and winged-helix domains. This positions the GSP near the G4, allowing for a guanine (indicated by blue squares) to be flipped out of the G-quartet and sequestered in the GSP. The guanine can either release back into the G-quartet, allowing the G4 to refold and leading to the observed repetitive FRET cycling, or RecQ can translocate to the next guanine.

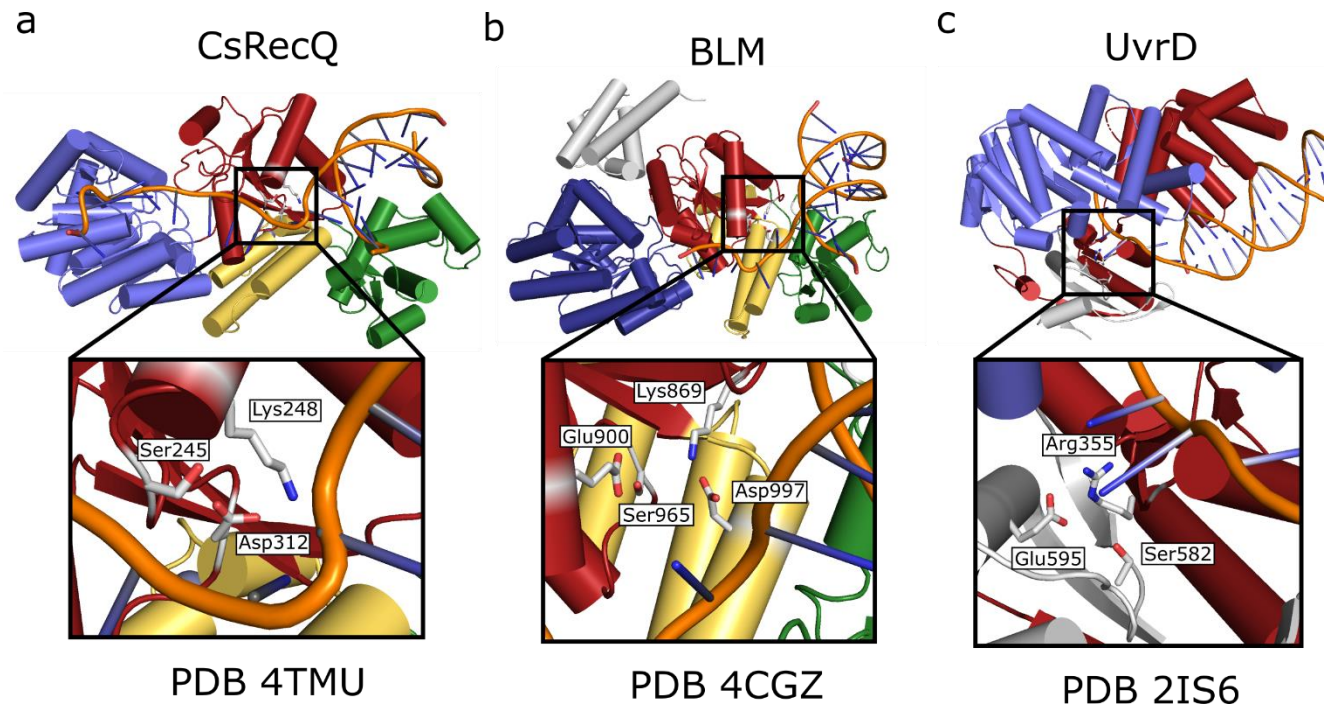


Figure 2.12. Possible structural conservation of the GSP in other G4 resolving helicases. a) The CsRecQ GSP (PDB 4TMU)21. b) Pockets with similarity to the RecQ GSP are found in comparable positions in BLM (PDB 4CGZ)30 and c) in UvrD (PDB 2IS6)31 helicases.

Table 2.1. Data collection and refinement statistics

	RecQ-G4 (PDB 6CRM)
Data collection[a]	
Space group	$P2_12_12$
Cell dimensions	
a, *b*, *c* (Å)	78.5, 94.7, 98.9
α, β, γ (°)	90, 90, 90
Resolution (Å)	43.83-2.19 (2.27-2.19)[b]
R_{sym}	0.15 (2.89)
$CC_{1/2}$	0.999 (0.377)
$I / \sigma I$	11.35 (0.76)
Completeness (%)	99.1 (92.63)
Redundancy	13.0 (11.3)
Refinement	
Resolution (Å)	43.83 – 2.19 (2.27 – 2.19)
No. reflections	38101
R_{work} / R_{free}	20.5/24.0
No. atoms	
Protein	4,035
DNA	290
Zn	1
Water	142
B-factors	
Macromolecules	87.55
Zn	54.49
Water	73.18
R.m.s. deviations	
Bond lengths (Å)	0.004
Bond angles (°)	0.67

a: A single crystal was used for data collection.

b. Values in parentheses are for highest-resolution shell.

Table 2.2 DNA binding and unwinding rates of bacterial RecQ variants

RecQ Variant	duplex binding ($K_{d,\ app}$, μM)	G4 binding ($K_{d,\ app}$, μM)	duplex unwinding rate (min^{-1})	G4 unwinding rate (TTA-T15) (min^{-1})	G4 unwinding rate (TAA-T15) (min^{-1})
EcRecQ	0.95±0.06	1.0±0.1	0.176±0.017	0.038±0.001	0.054±0.002
Ec, Ser245Ala	2.2±0.2	ND	0.19±0.06	no unwinding	no unwinding
Ec, Asp312Ala	0.28±0.02	0.33±0.04	0.088±0.005	no unwinding	no unwinding
CsRecQ	2.9±0.3	4.7±0.9	0.090±0.013	0.14±0.02	0.053±0.002
Cs, Ser245Ala	4.0±1.6	ND	0.081±0.014	no unwinding	no unwinding

ND, Not determined

References:

1 Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.* **13**, 770-780 (2012).

2 Paeschke, K., Capra, John A. & Zakian, Virginia A. DNA Replication through G-Quadruplex Motifs Is Promoted by the Saccharomyces cerevisiae Pif1 DNA Helicase. *Cell* **145**, 678-691 (2011).

3 Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci.* **99**, 11593-11598 (2002).

4 Kumari, S., Bugaut, A., Huppert, J. L. & Balasubramanian, S. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. *Nat. Chem. Biol.* **3**, 218-221 (2007).

5 Rawal, P. *et al.* Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome Res.* **16**, 644-655 (2006).

6 Koirala, D. *et al.* Intramolecular folding in three tandem guanine repeats of human telomeric DNA. *Chem. Commun.* **48**, 2006-2008 (2012).

7 Hershman, S. G. *et al.* Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic Acids Res.* **36**, 144-156 (2008).

8 Vaughn, J. P. *et al.* The DEXH Protein Product of the DHX36 Gene Is the Major Source of Tetramolecular Quadruplex G4-DNA Resolving Activity in HeLa Cell Lysates. *J. Biol. Chem.* **280**, 38117-38120 (2005).

9 Gray, L. T., Vallur, A. C., Eddy, J. & Maizels, N. G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat. Chem. Biol.* **10**, 313 (2014).

10 Wu, Y., Shin-ya, K. & Brosh, R. M. FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Mol. Cell. Biol.* **28**, 4116-4128 (2008).

11 Wu, X. & Maizels, N. Substrate-specific inhibition of RecQ helicase. *Nucleic Acids Res.* **29**, 1765-1771 (2001).

12 Sun, H., Bennett, R. J. & Maizels, N. The Saccharomyces cerevisiae Sgs1 helicase efficiently unwinds G-G paired DNAs. *Nucleic Acids Res.* **27**, 1978-1984 (1999).

13 Fry, M. & Loeb, L. A. Human Werner Syndrome DNA Helicase Unwinds Tetrahelical Structures of the Fragile X Syndrome Repeat Sequence d(CGG) n. *J. Biol. Chem.* **274**, 12797-12802 (1999).

14 Sun, H., Karow, J. K., Hickson, I. D. & Maizels, N. The Bloom's syndrome helicase unwinds G4 DNA. *J. Biol. Chem.* **273**, 27587-27592 (1998).

15 Taylor, E. M. *et al.* Xeroderma pigmentosum and trichothiodystrophy are associated with different mutations in the XPD (ERCC2) repair/transcription gene. *Proc. Natl. Acad. Sci.* **94**, 8658-8663 (1997).

16 Levitus, M. *et al.* The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. *Nat. Genet.* **37**, 934 (2005).

17 Chang, S. *et al.* Essential role of limiting telomeres in the pathogenesis of Werner syndrome. *Nat. Genet.* **36**, 877 (2004).

18 German, J. Bloom syndrome: a mendelian prototype of somatic mutational disease. *Medicine (Baltimore)* **72**, 393-406 (1993).

19 Croteau, D. L., Popuri, V., Opresko, P. L. & Bohr, V. A. Human RecQ helicases in DNA repair, recombination, and replication. *Annu. Rev Biochem.* **83**, 519-552 (2014).

20 Tippana, R., Hwang, H., Opresko, P. L., Bohr, V. A. & Myong, S. Single-molecule imaging reveals a common mechanism shared by G-quadruplex-resolving helicases. *Proc. Natl. Acad. Sci.* **113**, 8448-8453 (2016).

21 Manthei, K. A., Hill, M. C., Burke, J. E., Butcher, S. E. & Keck, J. L. Structural mechanisms of DNA binding and unwinding in bacterial RecQ helicases. *Proc. Natl. Acad. Sci.* **112**, 4292-4297 (2015).

22 Huber, M. D., Duquette, M. L., Shiels, J. C. & Maizels, N. A Conserved G4 DNA Binding Domain in RecQ Family Helicases. *J. Mol. Biol.* **358**, 1071-1080 (2006).

23 Bernstein, D. A., Zittel, M. C. & Keck, J. L. High-resolution structure of the E.coli RecQ helicase catalytic core. *EMBO J.* **22**, 4910-4921 (2003).

24 Bernstein, D. A. & Keck, J. L. Conferring Substrate Specificity to DNA Helicases: Role of the RecQ HRDC Domain. *Structure* **13**, 1173-1182 (2005).

25 Li, Y., Korolev, S. & Waksman, G. Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of Thermus aquaticus DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.* **17**, 7514-7525 (1998).

26 McCullough, A. K., Dodson, M. L., Scharer, O. D. & Lloyd, R. S. The role of base flipping in damage recognition and catalysis by T4 endonuclease V. *J. Biol. Chem.* **272**, 27210-27217 (1997).

27 Slupphaug, G. *et al.* A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA. *Nature* **384**, 87-92 (1996).

28 Klimasauskas, S., Kumar, S., Roberts, R. J. & Cheng, X. HhaI methyltransferase flips its target base out of the DNA helix. *Cell* **76**, 357-369 (1994).

29 Harami, G. M. *et al.* Shuttling along DNA and directed processing of D-loops by RecQ helicase support quality control of homologous recombination. *Proc. Natl. Acad. Sci.* **114**, E466-E475 (2017).

30 Newman, J. A. *et al.* Crystal structure of the Bloom's syndrome helicase indicates a role for the HRDC domain in conformational changes. *Nucleic Acids Res.* **43**, 5221-5235 (2015).

31 Lee, J. Y. & Yang, W. UvrD helicase unwinds DNA one base pair at a time by a two-part power stroke. *Cell* **127**, 1349-1360 (2006).

32 Chen, M. C. *et al.* Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. *Nature* **558**, 465-469 (2018).

33 Cahoon, L. A., Manthei, K. A., Rotman, E., Keck, J. L. & Seifert, H. S. Neisseria gonorrhoeae RecQ Helicase HRDC Domains Are Essential for Efficient Binding and Unwinding of the pilE Guanine Quartet Structure Required for Pilin Antigenic Variation. *J. Bacteriol.* **195**, 2255-2261 (2013).

34 Roberts, R. J. On base flipping. *Cell* **82**, 9-12 (1995).

35 Kuryavyi, V., Cahoon, L. A., Seifert, H. S. & Patel, D. J. RecA-binding pilE G4 sequence essential for pilin antigenic variation forms monomeric and 5' end-stacked dimeric parallel G-quadruplexes. *Structure* **20**, 2090-2102 (2012).

36 Singh, D. K., Ghosh, A. K., Croteau, D. L. & Bohr, V. A. RecQ helicases in DNA double strand break repair and telomere maintenance. *Mutat. Res./Fund. Mol. M.* **736**, 15-24 (2012).

37 Cahoon, L. A. & Seifert, H. S. An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. *Science* **325**, 764-767 (2009).

38 Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. D* **66**, 133-144 (2010).

39 McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658-674 (2007).

40 Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126-2132 (2004).

41 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213-221 (2010).

42 Lee, J. Y. & Yang, W. UvrD helicase unwinds DNA one base pair at a time by a two-part power stroke. *Cell* **127**, 1349-1360 (2006).

Chapter 3

Antigenic variation in *Neisseria gonorrhoeae* occurs independently of RecQ-mediated unwinding of the *pilE* G-quadruplex

The following chapter represents a first author body of work and was submitted to the *Journal of Bacteriology* in March, 2019.

Andrew Voter performed the protein purification and bulk helicase assays. *In vivo* colony phase variation assays were performed by Melanie Callaghan and Joe Dillard. Single-molecule experiments were performed by Ramreddy Tippana and Sua Myong.

Abstract

The obligate human pathogen *Neisseria gonorrhoeae* alters its cell surface antigens to evade the immune system in a process known as antigenic variation (AV). During AV, portions of the expressed pilin gene (*pilE*) are exchanged with silent pilin gene copies (*pilS*) through homologous recombination. The *pilE*/*pilS* exchange is initiated by formation of a parallel guanine quadruplex (G4) structure near the *pilE* gene, which recruits the homologous recombination machinery. The RecQ helicase, which has been proposed to aid AV by unwinding the *pilE* G4 structure, is an important component of this machinery. However, RecQ also promotes homologous recombination through G4-independent duplex DNA unwinding, leaving the relative importance of its G4 unwinding activity unclear. Previous investigations revealed a guanine-specific pocket (GSP) on the surface of RecQ that is required for G4, but not duplex, DNA unwinding. To determine whether RecQ-mediated G4 resolution is required for AV, *N. gonorrhoeae* strains that encode a RecQ GSP variant that cannot unwind G4 DNA were created. In contrast to the hypothesis that G4 unwinding by RecQ is important for AV, the RecQ GSP variant *N. gonorrhoeae* strains had normal AV levels. Analysis of a purified RecQ GSP variant confirmed that it retained duplex DNA unwinding activity but had lost its ability to unwind antiparallel G4 DNA. Interestingly, neither the GSP-deficient RecQ variant nor the wild type RecQ could unwind the parallel *pilE* G4 nor the prototypical *c-myc* G4. From these results we conclude that *N. gonorrhoeae* AV occurs independently of RecQ-mediated *pilE* G4 resolution.

Importance

The pathogenic bacteria *Neisseria gonorrhoeae* avoids clearance by the immune system through antigenic variation (AV), the process by which immunogenic surface features of the bacteria are exchanged for novel variants. RecQ helicase is critical in AV and its role has been proposed to stem from its ability to unwind a DNA secondary structure known as a guanine-quadruplex (G4) that is central to AV. In this work, we demonstrate that the role of RecQ in AV is independent of its ability to resolve G4s and that RecQ is incapable of unwinding the G4 in question. We propose a new model of RecQ's role in AV where the G4 might recruit or orient RecQ to facilitate homologous recombination.

Introduction

Over 550,000 Americans are infected annually with *Neisseria gonorrhoeae,* the causative agent of gonorrhea.[1] Untreated gonorrhea infections can lead to serious complications including septic arthritis, pelvic inflammatory disease and infertility.[2] Although gonorrhea can be treated using antibiotics, rising levels of resistance have the potential to eliminate current therapies available to patients.[3] Indeed, strains with resistance to front-line clinical antimicrobial agents have been reported.[4,5] A better understanding of the mechanisms of pathogenesis in *N. gonorrhoeae* is critical for the development of novel therapeutics and treatment strategies required to maintain our ability to treat gonorrhea.

Antigenic variation (AV) is a critical process used by *N. gonorrhoeae* and other pathogens to avoid clearance by the host immune system. During infection, antigens on the surface of the bacterial cells are detected by the host immune system, which directs production of immune cells to clear the infection. However, *N. gonorrhoeae* can evade the immune response by generating new variants of surface antigens. This variation occurs through recombination with silent copies of the antigen and alters the immunogenic epitopes. These changes force the immune system to develop new antibodies to clear the infection. An essentially limitless number of variants can be generated through iterations of AV, impairing the development of protective immunity.[6,7]

AV of several surface antigens occurs in *N. gonorrhoeae*, including lipooligosaccharides,[8] opacity proteins[9] and the type IV pilus. However, AV is most common in the pilin subunits, indicating its major role in immune system evasion.[10] Only a single pilin gene, *pilE*, is actively expressed in *N. gonorrhoeae*, whereas its genome contains 19 silent copies of the pilin gene, called *pilS*. Portions of the *pilS* loci replace portions of *pilE* through RecA-mediated homologous recombination during AV.[11] While the precise mechanistic steps that drive pilin AV remain unclear, the contributions of several major factors have been characterized.[12]

A key early step in *N. gonorrhoeae* pilin AV is formation of a guanine-quadruplex (G4) DNA structure.[13] G4s are unusual DNA secondary structures that form in guanine-rich nucleic acid sequences through extensive hydrogen bonding and stacking among the guanine bases. The interactions within G4s form extremely stable structures that can be challenging to unwind. G4s fold in either parallel or antiparallel structures based on the orientation of their phosphodiester backbone. These orientations are typified by the parallel *c*-myc G4[14] (Fig. 3.1A, nearly identical to the *N. gonorrhoeae pilE* G4 element) and the antiparallel human telomeric G4[15] (telo-G4, Fig. 3.1B). These two forms are structurally distinct, have differing stabilities, and varied susceptibilities to helicase unwinding. The *pilE* G4-forming sequence is located upstream of the *pilE* gene, and this G4 is known to be essential for AV, but not pilin expression.[13] Initiation of the AV process occurs when the *pilE* G4-forming sequence is unwound to allow transcription of a small non-coding RNA. Freed from the complementary template strand, the *pilE* G4 sequence folds into a G4 structure.[16] While it has been shown that G4 formation is required for AV and alternate G4-forming sequences fail to initiate AV, the precise role for the G4 has not been defined.[13,16] Because RecQ helicases are known to unwind G4 substrates[17] and Δ*recQ* strains have been shown to be partially deficient in AV,[12] it has been proposed that unwinding of the *pilE* G4 by the RecQ helicase is critical to the AV process.

The bacterial RecQ protein is a 3′ to 5′ DNA helicase.[18] The preferred substrate for RecQ is duplex DNA with a 3′ single-stranded (ss) DNA element, but bacterial RecQs have also been shown to unwind G4 DNA substrates.[17] Bacterial RecQs comprise a helicase motor domain made up of two RecA-like lobes, a structural Zn^{2+}-binding domain, a DNA-binding winged helix domain, and a regulatory Helicase and RNase-D C-terminal (HRDC) domain (Fig. 3.1C). The *N. gonorrhoeae* RecQ (NgRecQ) is distinct from most other RecQs in that it possesses three HRDCs rather than one. Previous studies have found that truncation of the two C-terminal-most HRDC domains from NgRecQ is sufficient to disrupt AV, and this defect was attributed to a relatively modest decrease in G4 DNA binding and unwinding by the variant.[19] However,

the truncated NgRecQ also has greatly reduced affinity for duplex and ssDNA and reduced helicase activity on Holliday junction substrates.[20] Furthermore, strains with truncated RecQ variants were found to be hypersensitive to UV irradiation, suggesting the AV deficiency could be the result of general effects on RecQ rather than a G4-specific defect.[20] A more precise isolation of the G4 unwinding activity of RecQ is needed to deconvolute the possible roles of RecQ-mediated G4 unwinding in AV.

We recently determined the crystal structure of RecQ from *Coronobacter sakazakii* (CsRecQ) in complex with an unfolded G4. This structure revealed the presence of a guanine-specific pocket (GSP) on the surface of RecQ that was essential for G4 helicase activity but dispensable for unwinding duplex substrates.[21] The GSP is conserved in NgRecQ (Fig. 3.1D), and we reasoned that mutation of the GSP in NgRecQ could be used to determine the role of RecQ-mediated G4 unwinding in AV. We therefore generated *N. gonorrhoeae* strains bearing GSP-defective *recQ*. Surprisingly, these strains underwent AV at the same rate as wild type cells, which refuted a role for RecQ G4 unwinding in AV. Wild type and GSP-defective NgRecQ proteins were purified and, although both retained duplex unwinding activity, neither was competent to unwind *pilE* G4 DNA. From these results, we conclude that antigenic variation in *N. gonorrhoeae* occurs independently of RecQ-mediated unwinding of the *pilE* G4, and the role of RecQ is limited to facilitating homologous recombination by RecA.

Materials and methods

Generation of the GSP variants in NgRecQ. Plasmids pAV305 and pAV306 were each ligated with pIDN1 after HindIII/XhoI digestion (Table 3.1).[22] Plasmids were transformed into TAM1 *E. coli*, and transformants were screened for plasmids of the expected size. Final constructs pMMC15 (containing the pAV305 NgRecQD312A mutation) and pMMC16 (containing the pAV306 NgRecQS245A mutation) were confirmed by DNA sequencing. *N. gonorrhoeae* strains MMC533 and MMC534 were generated by spot transformation of *N. gonorrhoeae* FA1090 with plasmids pMMC15 and pMMC16, respectively (Table 3.2). Screening was performed by colony PCR and digestion with either NruI (MMC533) or BssHII (MMC534).[23] The *recQ*::*ermC* interruption strain MMC536 was generated by spot transforming FA1090 with linear pPK1014.[20] Transformants were selected with 2µg/mL erythromycin and confirmed by PCR and sequencing.

Colony phase variation assay. Pilus-dependent colony morphology changes (PDCMC) assays were performed as described by Sechman, Rohrer, and Seifert.[12] Briefly, strains were grown from frozen stocks on GCB agar for 24hrs. A single piliated colony was restreaked onto GCB agar and incubated overnight. For each strain, 10 colonies were chosen. At 22, 24, 26, 28, and 30hrs, chosen colonies were analyzed using a stereomicroscope and scored by counting the nonpiliated outgrowths visible on the colony. Each new outgrowth increases the score by one, until four outgrowths have developed. All colonies with four or more outgrowths receive a score of four. FA1090 *recA6,* wherein *recA* is under the control of an IPTG-inducible promoter, was used without induction as a *recA* deficient control.[24]

Purification of the NgRecQ variants. BL21-AI *Escherichia coli* cells were transformed with overexpression plasmids encoding N-terminally His-tagged NgRecQ or the NgRecQ Asp307Ala variant. Cells were grown at 37 °C to an OD_{600} of 0.6 before protein expression was induced with 0.2% arabinose (wt/vol) and 1 mM IPTG (NgRecQ) or 0.2 % arabinose (wt/vol) (NgRecQ Asp307Ala). The cells were grown for a further 4 hours

at 37 °C, harvested by centrifugation and stored at -80 °C. Cell pellets were resuspended in lysis buffer (20 mM Tris·HCl [pH 8.0], 500 mM NaCl, 1 mM 2-mercaptoethanol (BME), 1 mM phenylmethane sulfonyl fluoride, 100 mM dextrose, 15 mM imidazole, 1 Pierce protease inhibitor tablet, 10% (vol/vol) glycerol), lysed by sonication and clarified by centrifugation. The supernatant was incubated with Ni-NTA agarose resin at 4 °C for 1 hour, then washed extensively with lysis buffer. Proteins were eluted from the resin with elution buffer (lysis buffer supplemented with 250 mM imidazole). Eluent was diluted to 50 mM NaCl using dilution buffer (5 mM Tris·HCl [pH 8.0], 1 mM BME, 10% (vol/vol) glycerol) then loaded onto a HiPrep QFF ion exchange column and eluted with a 0.05 – 1.0 M NaCl gradient. RecQ-containing fractions were identified by SDS-PAGE analysis, concentrated, and further purified with an S-100 size exclusion column before dialysis into storage buffer (20 mM Tris·HCl, 1 M NaCl, 4 mM BME, 40% (vol/vol) glycerol, 1 mM ethylenediaminetetraacetic acid) and stored at -20 °C.

Bulk dual-labelled G4 helicase assays. A dual-labelled, HPLC purified oligonucleotide with the *pilE* G4 sequence (5' FAM – GGG TGG GT TGG GTG GG – black hole quencher) was obtained from Integrated DNA Technologies (Coralville, IA, USA). The oligonucleotide was resuspended in water and diluted to 1 µM (molecules) in 20 mM Tris·HCl, 100 mM KCl, then heated to 95 °C for 10 minutes and allowed to slowly cool to room temperature to ensure the G4s were properly folded at the start of the experiment. The oligonucleotide was diluted to a final concentration of 10 nM in a reaction buffer (25 mM Tris·HCl, 0.1 mM dithiothreitol, 3 mM $MgCl_2$) containing a variable amount of the NgRecQ helicase and either 50 mM NaCl or 100 mM KCl. All measurements were taken using a Photon Technology International Inc. fluorimeter with a 490 nm excitation and the emission was measured at 520 nm. After a ten-minute incubation, ATP was added to a final concentration of 1 mM and the emission intensity was recorded. All further intensities were normalized to the first intensity measurement taken after ATP addition.

Circular dichroism. G4-forming oligonucleotides were resuspended in water and then diluted to 5 µM (molecules) in 300 µL of either a KCl (20 mM Tris·HCl [pH 7.5], 50 mM KCl, 3 mM $MgCl_2$) or NaCl (25 mM

Tris·HCl, 50 mM NaCl, 3 mM $MgCl_2$) salt buffer. The oligonucleotides were heated to 95°C for 10 minutes and then allowed to slowly cool to room temperature. Circular dichroic spectra were recorded on an AVIV 420 circular dichroism spectrometer with a step size of 2 nm and a 5 second average. A buffer matched blank lacking DNA was subtracted from each reading. Samples were equilibrated at each temperature for 5 minutes before data collection. To generate the melting curve, the ellipticity at 260 nm was measured at increasing temperatures for each salt condition. Curve fitting was performed in Prism version 5.0c.

smFRET DNA substrates. Amine-modified ssDNA substrates were purchased from Integrated DNA Technologies. Cy3/Cy5 monofunctional NHS esters were used to label the amine modified ssDNA constructs (GE Healthcare, Princeton, NJ, USA). Amino-modified oligonucleotides (10 nmol in 50 µL ddH2O) and 100 nmol of Cy3/Cy5 NHS ester dissolved in 50 µL of 0.1 M NaHCO3 were combined and incubated with rotation for four hours in the dark. The labeled oligonucleotides and unreacted dye were separated by P6 columns (Bio-Rad, USA) or ethanol precipitation.

Both G4 and non-G4 substrates consist of a stem of dsDNA with 18 base pairs and a specific sequence of 3’ tailed ssDNA (Table 3.3). A Cy5-Cy3 FRET pair are placed at the junction and the 3’ end of the ssDNA, respectively.

T50 [10 mM Tris·HCl (pH 8.0), 50 mM NaCl] buffer was used to anneal the biotinylated and non-biotinylated oligonucleotides in a 1:1.5 molar ratio to a final concentration of 10 µM duplex. The sample was heated to 95 °C for 2 min slow cooled in a thermocycler. The high concentration of annealed DNA was stored at –20 °C and was freshly diluted for each measurement to 10 nM stock concentration in K100 buffer [10 mM Tris·HCl (pH 8.0), 100 mM KCl]

smFRET unwinding assays. All single-molecule unwinding assays were measured by using a custom-built total internal reflection (TIRF) microscope. A 532 nM laser (Coherent, USA) was used to excite the donor dye in the Cy3-Cy5 FRET pair used for the single molecule measurements. Fluorescence emission was

separated by a dichroic mirror with a cutoff of 630 nm to split the Cy3 and Cy5 signals, which were then detected on an EMCCD camera (iXon DU-897ECS0-#BV; Andor Technology). Single-molecule traces from the recorded data were extracted by IDL software. Matlab and Origin software was used to display and analyze the single-molecule traces. All homemade codes are in the smFRET package available at the Center for the Physics of Living Cells (https://cplc.illinois.edu/software/, Biophysics Department, the University of Illinois at Urbana-Champaign).

RecQ unwinding experiments were performed in reaction Buffer [20 mM Tris·HCl (pH 7.5), 50 mM KCl, 3 mM MgCl2, 1 mM ATP] with an oxygen scavenging system containing 0.8% vol/vol dextrose, 1 mg/mL glucose oxidase, 0.03 mg/mL catalase1, and 10 mM Trolox. All chemicals were purchased from Sigma Aldrich (St. Louis, MO).

Biotinylated FRET DNA (50 to 100 pM) were immobilized on polyethylene glycol-coated quartz surface via biotin-neutravidin linkage for two minutes then all unbound DNA was washed away. RecQ and mutant proteins (100 nM) were added at room temperature to initiate unwinding. 10–20 short movies (10 s) with an interval of 20 sec and separately 3–4 long movies (3 min) were then taken monitoring the Cy3 and Cy5 emission intensities over time. These were then analyzed to produce the FRET histograms and trajectories to monitor any unwinding activity. Unwinding rates were calculated as previously reported.[21]

Results and discussion

The GSP of RecQ is dispensable for AV

The specific roles of the RecQ helicase in *N. gonorrhoeae* AV are unclear. Deletion of the *recQ* gene or removal of two HRDC domains from the protein diminish AV[12,13,19] but whether this effect is due to a specific loss in G4 unwinding or RecQ activities in homologous recombination has not been defined. To determine the role of RecQ-mediated G4 unwinding on AV, we generated *N. gonorrhoeae* strains

bearing NgRecQ variants that we predicted would maintain all enzyme functions except for the ability to unwind G4 DNA. Our previous biochemical analysis of the RecQ proteins from *E. coli* (EcRecQ) and *C. sakazakii* (CsRecQ) demonstrated that disruption of a GSP on the surface of the enzymes resulted in a complete loss of G4 helicase activity, whereas duplex DNA unwinding activity was retained.[21] Therefore, we mutated the *recQ* gene in *N. gonorrhoeae* to encode for single-site variants containing disabled GSPs (Ser240Ala or Asp307Ala) to distinguish between a G4-specific and G4-independent role for RecQ in the AV process. These strains were compared with *recQ::erm* or *recA6* (RecA deficient) *N. gonorrhoeae* strains that are known, respectively, to impair or eliminate AV.[12,13] We predicted that if RecQ-mediated G4 unwinding was important for AV, then AV in the single-site GSP-deficient strains would be impaired to the same extent as the *recQ::erm* strain. We recapitulated the partial AV defect of the *recQ::erm* strain and complete loss of AV in the uninduced *recA6* strain. However, in contrast to the hypothesis, the strains with GSP-defective RecQ proteins underwent AV at the same rate as wild type *N. gonorrhoeae* (Fig. 3.2).

Two possibilities could explain the observed wild type levels of AV in these strains. First, the G4 helicase activity of NgRecQ may be occurring independently of its GSP. This is unlikely as the structural and biochemical studies defining the GSP were conducted using EcRecQ and CsRecQs, which are closely related to NgRecQ (45% identical through the first HRDC domain). Additionally, the residues that form the GSP in EcRecQ and CsRecQ are exactly matched in NgRecQ (Fig. 3.1B&C). Despite this overall similarity, specific features of NgRecQ, especially the presence of two additional HRDC domains, might confer GSP-independent *pilE* G4 unwinding. Alternatively, NgRecQ-mediated unwinding of the *pilE* G4 may be dispensable for AV.

NgRecQ cannot unwind the *pilE* G4 in bulk assays

To distinguish between the possible explanations for AV function with the RecQ GSP-defective variants, we sought to determine whether the G4 helicase activity of NgRecQ occurred independently of

its GSP. NgRecQ and the NgRecQ Asp307Ala GSP variant were purified and tested for DNA unwinding *in vitro*. In this experiment, the *pilE* G4 was labelled with 5′ FAM and a 3′ black hole quencher (BHQ). In the folded state, FAM fluorescence was quenched by the nearby BHQ and unwinding was expected to result in an increase in fluorescence intensity. After allowing NgRecQ to bind to the G4, ATP was added, and the fluorescence intensity was measured over time (Fig. 3.3A). Only a modest increase (~3%) in fluorescence intensity was observed after ATP addition for either both NgRecQ and NgRecQ Asp307Ala variant (Fig. 3.3B). These results contrast with the high-magnitude, albeit slow increase in fluorescence previously observed during NgRecQ G4 helicase action.[19]

The bulk G4 unwinding experiments were conducted with Na^+ as the primary cation, which is known to destabilize G4s.[25] To measure the potential effect of cation choice on G4 unwinding, we measured the stability of the G4 by performing circular dichroism thermal denaturation assays with the substrate in the presence of either Na^+ or K^+ (Fig. 3.4). The *pilE* G4 was markedly destabilized in NaCl; a T_m of 53 °C was observed in the presence of 50 mM NaCl, whereas a T_m in excess of 80 °C was observed in the KCl buffer. In the presence of a G4 destabilizing cation such as Na^+, minor variations in helicase assay setup could result in the appearance of G4 unwinding. However, when the helicase assay experiments were repeated in the presence of KCl, we again failed to observe significant unwinding for either protein (Fig. 3.3). Because K^+ is the primary intracellular cation in bacteria, with concentrations greatly exceeding that of Na^+,[26] this result suggested that NgRecQ may not be able to unwind the *pilE* G4 under physiological conditions or even under Na^+-containing conditions in which the *pilE* G4 structure is destabilized.

NgRecQ unwinds antiparallel, but not parallel G4s

An established single-molecule helicase assay was used to further assess NgRecQ duplex and G4 DNA unwinding properties. In these experiments, a 5′ Cy5 labeled oligonucleotide was tethered to a coverslip and annealed to a test oligonucleotide containing a 5′ complementary region (Fig. 3.5). The test

oligonucleotide contained a 3′ Cy3 label, a 3′ dT_{15} region for NgRecQ loading and, if required for the experiment, a G4-forming sequence between the 3′ dT_{15} and the 5′ complementary region. Given the 3′-5′ polarity of RecQ helicases, the enzyme is expected to bind to the 3′ dT_{15} and translocate towards the 5′ end. If the enzyme is competent to unwind the G4 and duplex DNA, the G4-containing test oligonucleotide will be released leading to a loss of Cy3 fluorescence (Fig. 3.5A). To ensure both NgRecQ proteins were properly folded and active, we first tested the ability of the proteins to unwind a simple duplex substrate that lacked a G4-forming sequence. Both wild type and the NgRecQ Asp307Ala variant were competent to unwind the duplex substrate, although the Asp307Ala variant had a 2.8-fold slower unwinding rate than wild type NgRecQ (Fig. 3.5B, Table 3.4).

Having demonstrated duplex unwinding activity of both proteins, we next measured G4 unwinding ability. As was observed for EcRecQ and CsRecQ, NgRecQ was found to robustly unwind a test oligo containing the antiparallel human telomeric G4 forming sequence ($(TTAGGG)_4$). In contrast, antiparallel G4 unwinding was not observed with the NgRecQ Asp307Ala variant, consistent with the essential nature of the GSP for RecQ-mediated G4 helicase activity (Fig. 3.5C). Because the *pilE* G4 forming sequence adopts a parallel conformation rather than the antiparallel structure of the telomeric G4, we next tested if NgRecQ was competent to unwind the parallel *c-myc* G4, which differs from the *pilE* G4 by only a single nucleotide in the middle loop (Fig. 1A). Neither the NgRecQ nor the NgRecQ Asp307Ala variant unwound the *c-myc* DNA (Fig. 3D). Similar results had previously been obtained with EcRecQ and CsRecQ.[21] These results indicate that while the GSP is required for NgRecQ to unwind antiparallel G4 DNA, parallel quadruplexes such as the *c-myc* and the *pilE* G4s are not substrates of the NgRecQ helicase.

Although no unwinding was observed for the parallel G4 substrates, the addition of either NgRecQ or the NgRecQ Asp307Ala variant resulted in an ATP-independent shift to lower FRET states (~0.8 to ~0.5, Fig. 3.6). This shift likely indicates that NgRecQ and the variant can bind to and possibly alter the structure of the substrate without unwinding. Notably, this was not observed in our prior experiments using the

catalytic cores of EcRecQ and CsRecQ (which lacked HRDC domains), so this binding or reorganization may be HRDC dependent. Another possibility is that NgRecQ binds to the 3' ss dT tail, stretching the DNA and increasing the distance between the Cy3-Cy5 FRET pair.

It has been shown that G4 folding is essential for AV, and there appears to be a requirement for both the compacted structure adopted by the *pilE* G4 and its parallel orientation.[13] Indeed, *N. gonorrhoeae* strains in which the *pilE* G4 has been replaced by other G4 forming sequences (including those likely to be unwound by RecQ) cannot undergo AV.[13] These findings are consistent with our observations that NgRecQ-mediated unwinding of the *pilE* G4 is not a requirement for AV and that the isolated NgRecQ enzyme is not capable of such unwinding. Despite this, NgRecQ is involved in the AV process.[12,19,20]

The second and third HRDC domains of NgRecQ are crucial for NgRecQ's role in AV[19,20] and we propose two possible roles that are consistent with this requirement (Fig. 3.7). First, the HRDC domains might interact with the G4 to promote homologous recombination without NgRecQ G4 unwinding. In support of this, the shift to a lower FRET state observed in our FRET assays is consistent with binding to the *pilE* G4. Such binding could distort the G4 or adjacent DNA to promote access by another protein, either to unwind the G4 or facilitate RecA-mediated homologous recombination (Fig. 3.7, right). Similarly, the NgRecQ HRDC domains might bind the *pilE* G4 to orient the helicase in a manner that aids in productive RecA loading. Because deletion of the NgRecQ HRDCs have only a modest impact on either G4 binding or unwinding, the HRDC domains might instead modulate NgRecQ activity. In this scenario, the NgRecQ variant lacking the C-terminal-most HRDC domains may bind to the *pilE* G4 such that the *pilE* G4 or nearby duplex DNA cannot be unwound in preparation for RecA loading. Thus, the HRDCs might serve to recruit NgRecQ to the *pilE* G4 and properly orient the helicase (Fig. 3.7, left).

A second possibility is that the NgRecQ HRDCs might be required for efficient RecA loading via a G4-independent mechanism. To this end, the NgRecQ variant lacking the C-terminal-most HRDC domains is defective in binding and unwinding some, but not all, DNA substrates.[20] Additionally, a strain bearing truncated *recQ* was as sensitive to UV-induced DNA damage as a *recQ::erm* knockout.[20] Thus, NgRecQ dysregulation resulting solely from the loss of the HRDC domains is sufficient to inhibit NgRecQ-mediated DNA repair. The loss of the HRDCs very likely impairs NgRecQ mediated loading of RecA during AV. Further studies are needed to clarify the role of the NgRecQ HRDC domains in AV.

What then is the role of the *pilE* G4 in AV? RecA has a high affinity for the *pilE* G4 and binding to the quadruplex stimulates RecA-mediated strand exchange.[27] Furthermore, substitution of the *pilE* G4 with other G4s that RecA cannot bind blocks AV.[13] It has been proposed that the critical function of the *pilE* G4 may be to recruit and stimulate RecA.[27] Unfortunately, this is a difficult hypothesis to test; RecA-mediated homologous recombination is likely essential for AV independent of its affinity for the *pilE* G4 and there are no known RecA separation-of-function mutants that would allow for isolation of the role of the RecA-G4 interaction. The fate of the G4 remains another outstanding question. Unresolved G4s block replication fork progression and lead to toxic DNA damage. Therefore, the *pilE* G4 is likely resolved by a G4 resolving helicase, such as UvrD[28] or DinG,[29] prior to replication.

In conclusion, the results presented in this study demonstrate that AV in *Neisseria gonorrhoeae* occurs independently of NgRecQ-mediated unwinding of the *pilE* G4 and that NgRecQ is incapable of unwinding the *pilE* G4. Instead, we propose that the role of NgRecQ in AV is limited to facilitating RecA-mediated homologous recombination. Future experiments will elucidate the structure and role of the RecA G4 interaction in AV and explore how the *pilE* G4 is resolved or tolerated during DNA replication.

Acknowledgements.

We thank members of the Keck laboratory for critical reading of this manuscript. This work was funded by NIH R01 GM098885 to J.L.K. A.F.V. was supported by NIH F30 CA210465 and T32 GM008692.

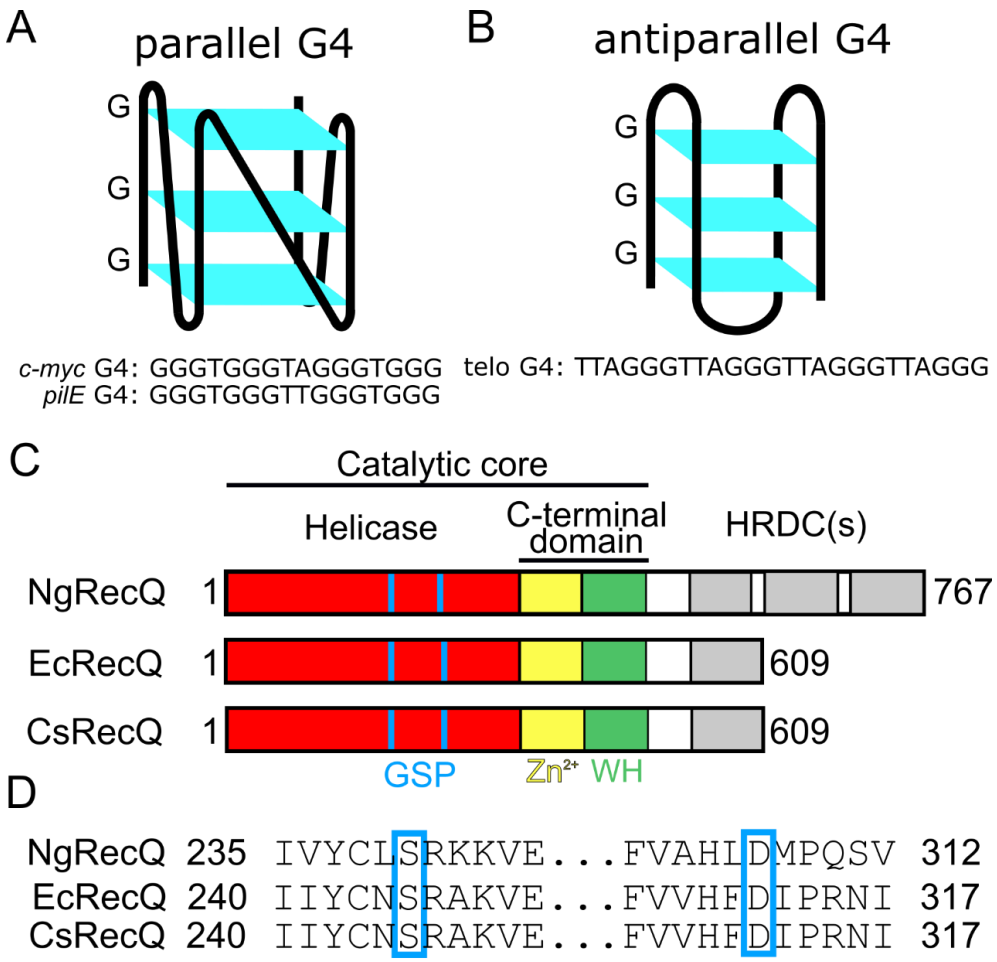


Figure 3.1. Comparison of the structures of antiparallel and parallel G4s and bacterial RecQ helicases. A) Model of a parallel G4, such as the *c-myc* and *pilE* G4s used in this study. Each blue structure represents four guanine bases in a quartet structure. G4 forming sequences are shown under each model. B) Model of an antiparallel G4 typified by the telomeric G4. C) Comparison of the domain architecture of RecQ helicases from bacterial species. The location of the GSP is denoted by the vertical blue lines. D) Sequence alignment of GSP between bacterial RecQ helicases. Residues that directly interact with a guanine base within the GSP are boxed.

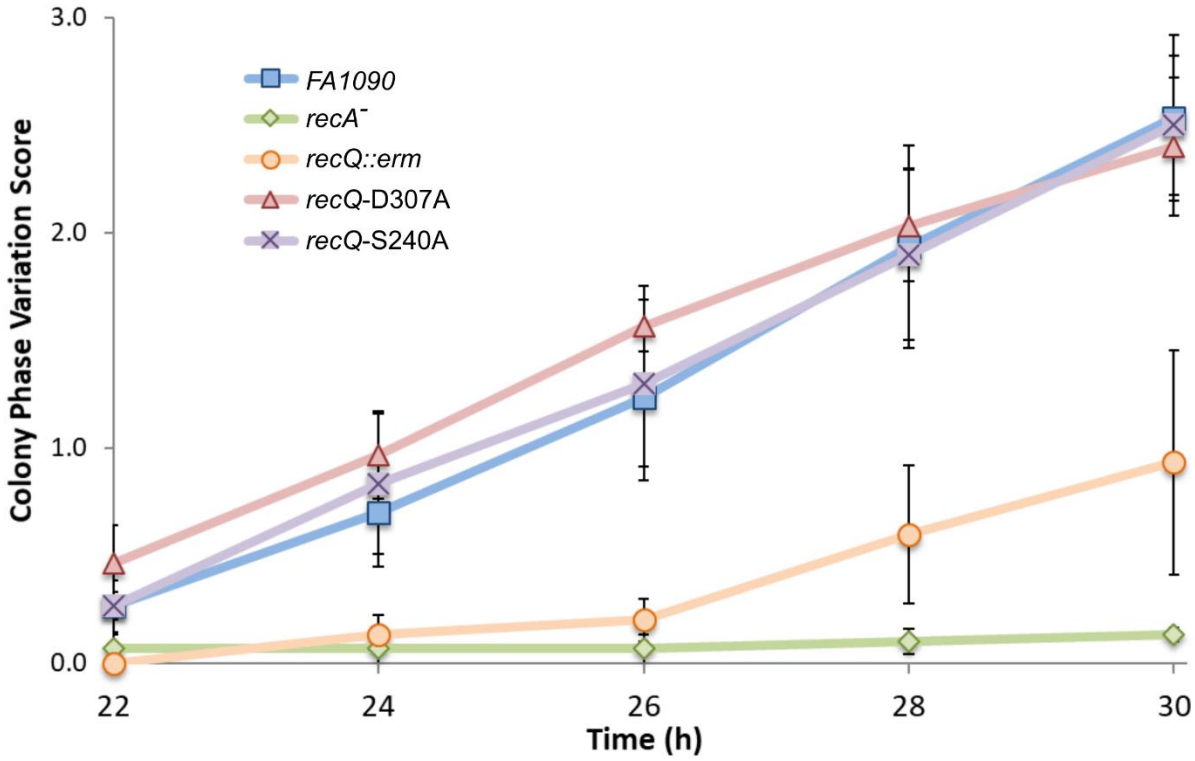


Figure 3.2. The role of the GSP in *N. gonorrhoeae* antigenic variation. Pilin-dependent colony morphology changes of *N. gonorrhoeae* variants. Error bars represent the standard error of the mean of 3 biological replicates of at least 10 colonies.

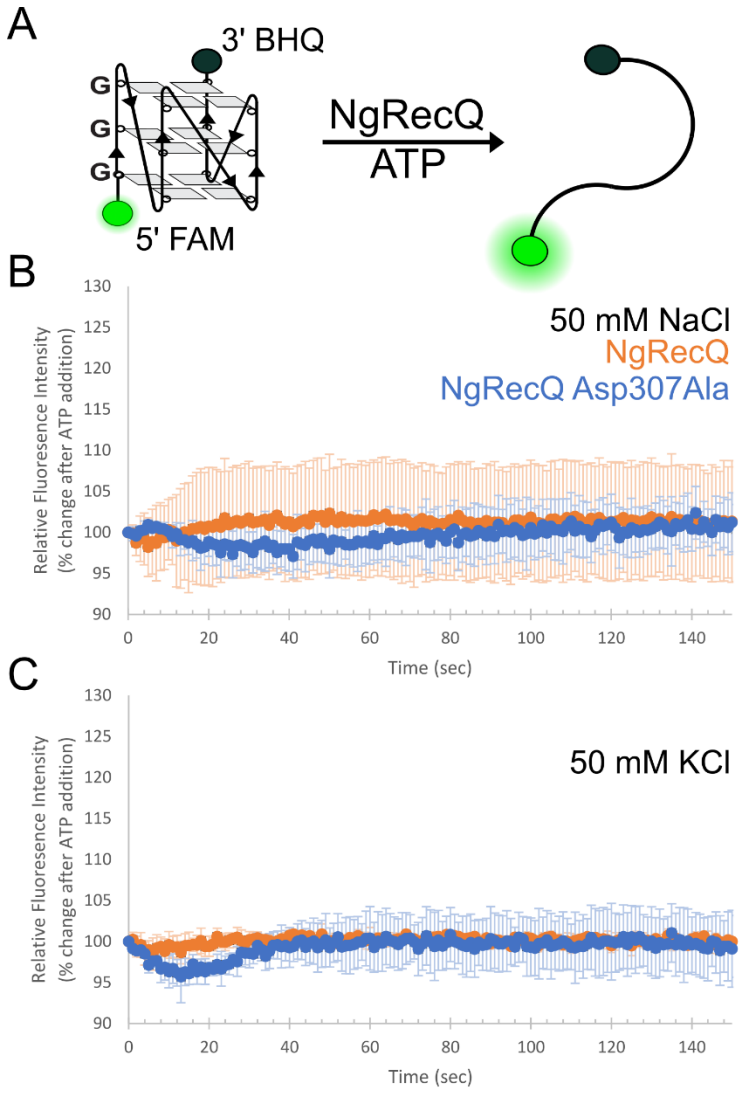


Figure 3.3. Bulk helicase assay to monitor RecQ mediated unwinding of the *pilE* G4. A) Scheme of the helicase assay. B) Relative fluorescence intensity of the *pilE* G4 after the addition of ATP in a reaction buffer containing 50 mM NaCl. The G4 was preincubated with either NgRecQ (orange) or NgRecQ Asp307Ala (blue). Error bars represent the standard deviation of 3 replicates. C) Same as in B, except the NaCl was replaced with 50 mM KCl.

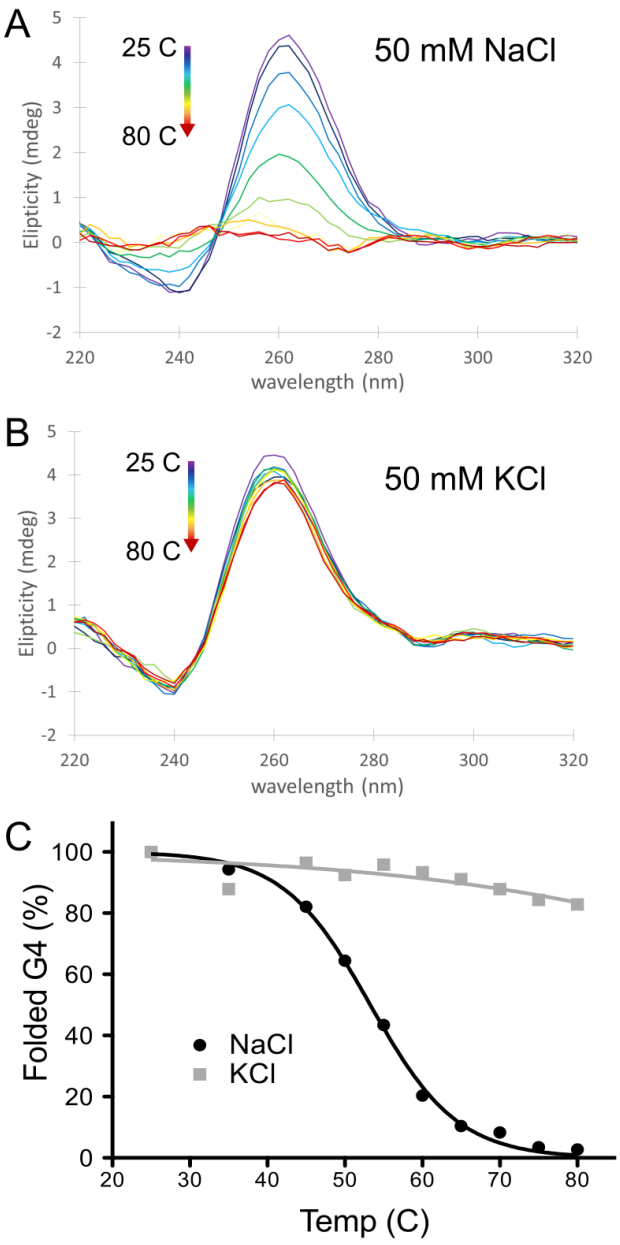


Figure 3.4. Thermal stability of the *pilE* G4 is cation dependent. A) Overlaid circular dichroism spectra of the *pilE* G4 in 50 mM NaCl at temperatures ranging from 25 °C to 80 °C. B) As in A, except with 50 mM KCl. C) Thermal melt curves of the *pilE* G4 in either KCl or NaCl containing buffers. Percentage of folded G4 was calculated from the ellipticity at 260 nm.

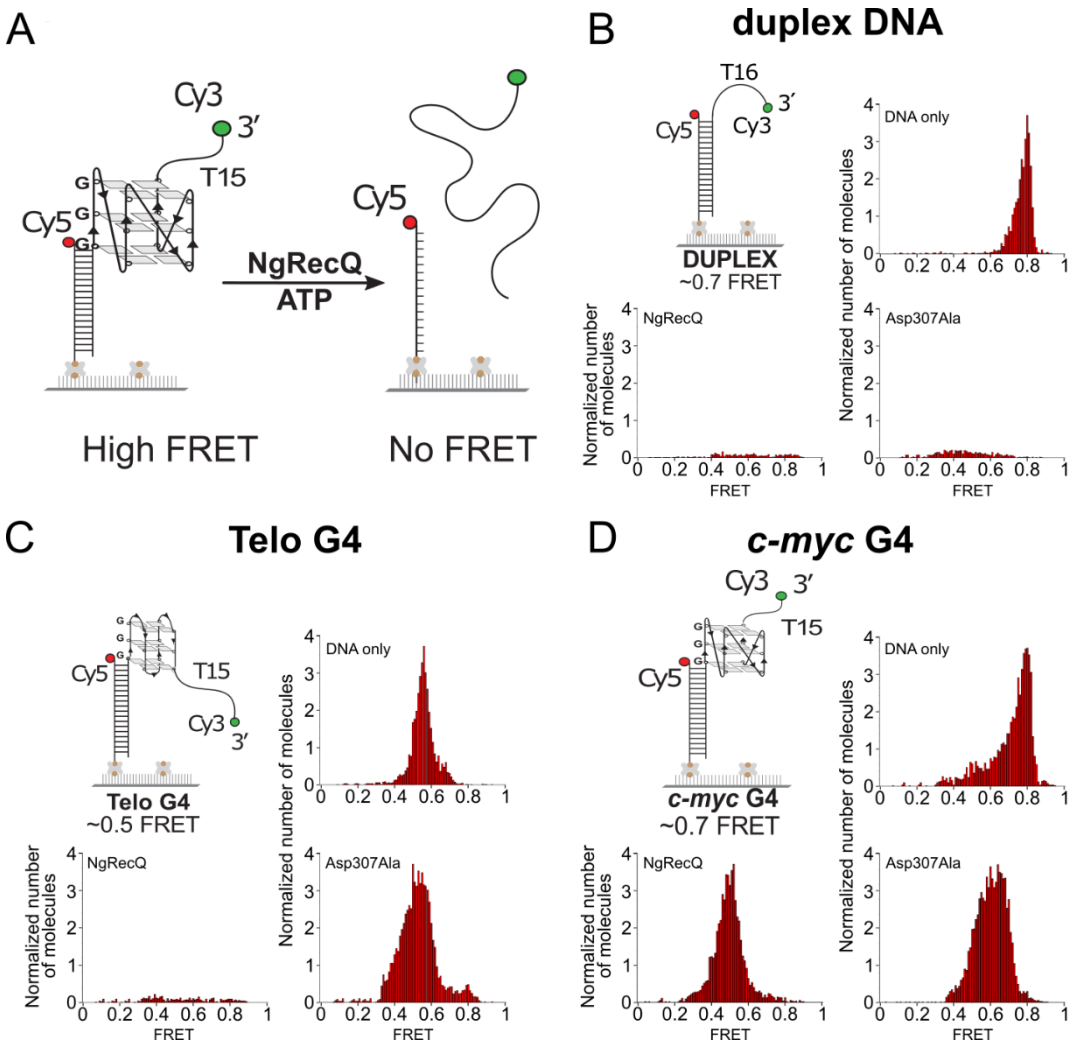


Figure 3.5. Single-molecule FRET studies of NgRecQ helicase activity. A) Scheme depicting the smFRET strategy used to monitor DNA unwinding by NgRecQ. B) NgRecQ-mediated unwinding of duplex DNA. Histograms of the smFRET signals for the DNA alone (top), or 12-min after the addition of ATP and NgRecQ or NgRecQ Asp307Ala. C) and D) Same as in B but for the antiparallel telo G4 and parallel *c-myc* G4 substrates, respectively.

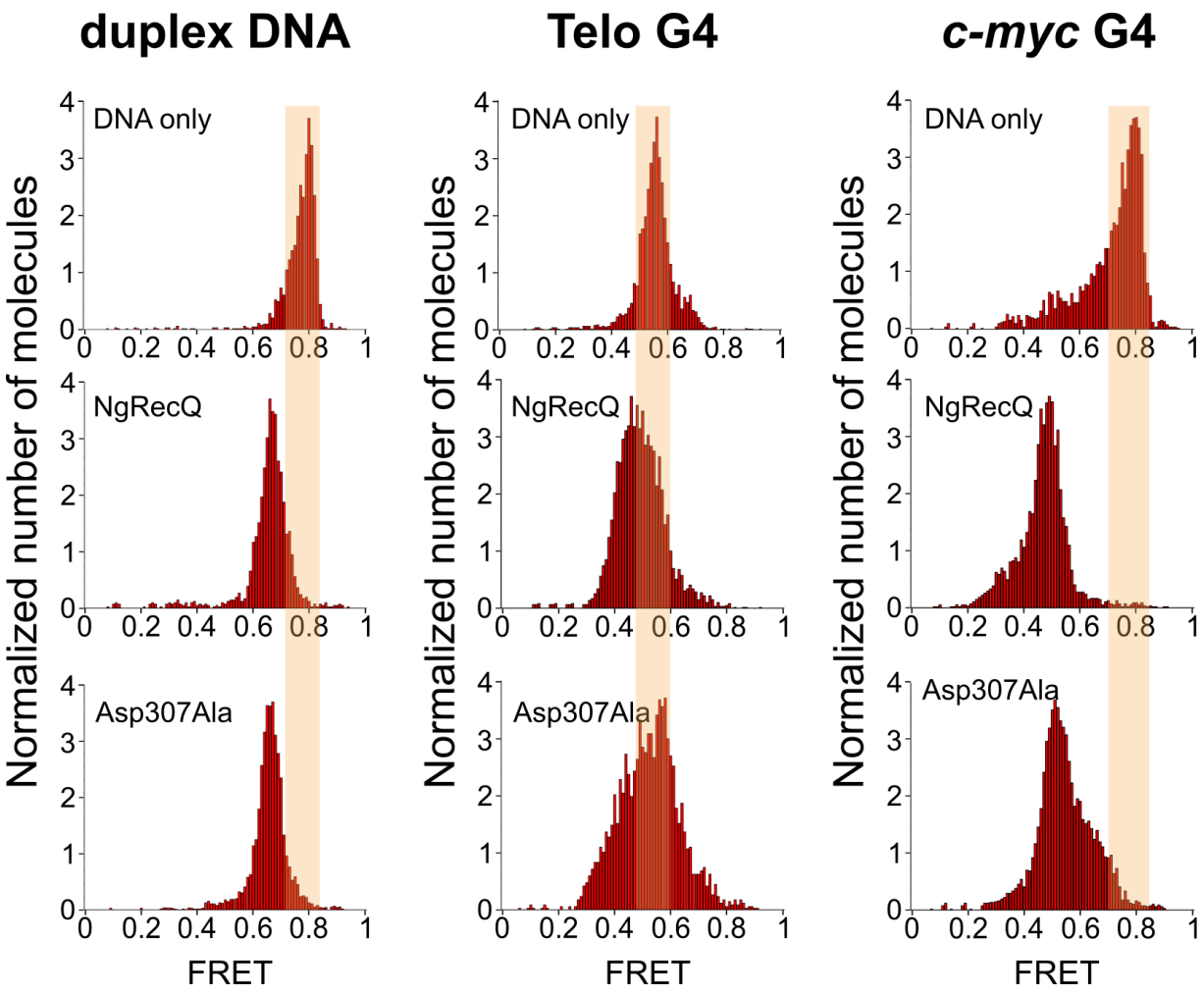


Figure 3.6. RecQ alters G4 structure in an ATP-independent manner. Histograms of the smFRET signal for each of the 3 DNA tests. The top row depicts DNA alone, the middle and bottom rows are the same DNA after incubation with NgRecQ or Asp307Ala respectively. No ATP was included in these experiments. The orange bar denotes the primary FRET peak of the DNA only condition,

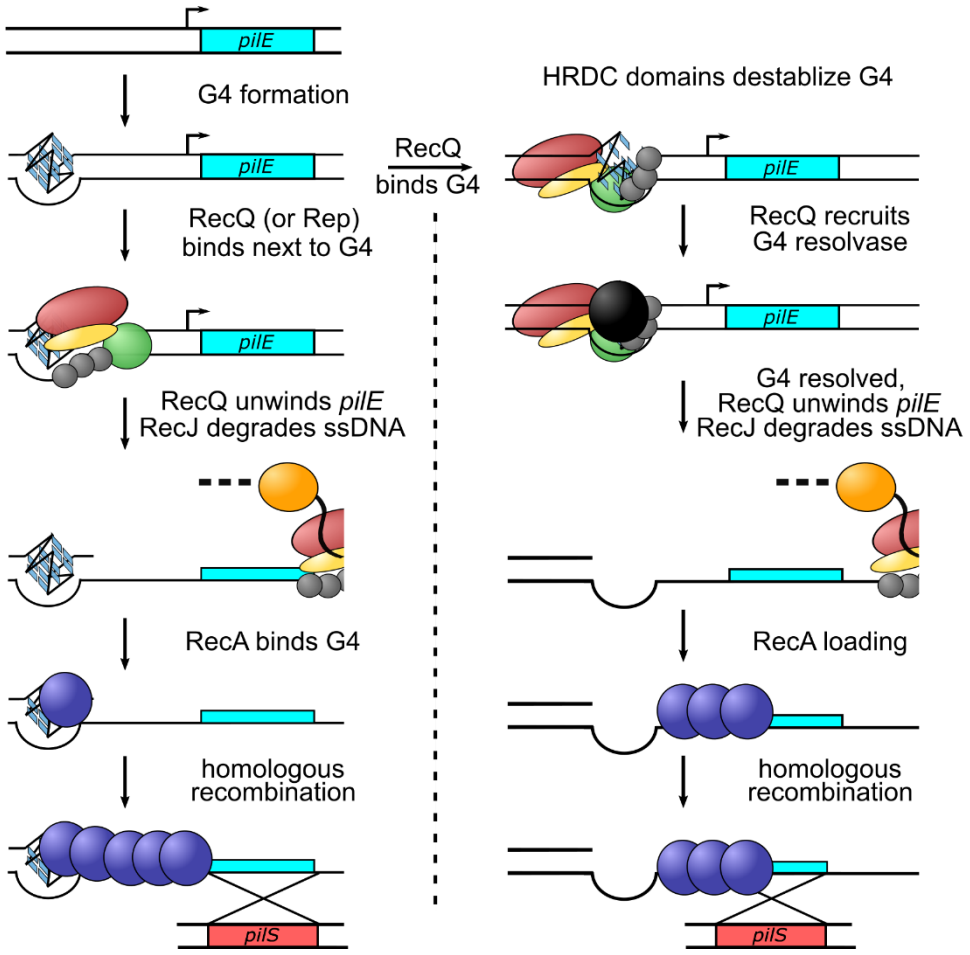


Figure 3.7. Model of the role of RecQ in *N. gonorrhoeae* AV. After formation of the *pilE* G4, RecQ (domains colored as in Figure 1) binds to the G4. **Left**. The RecQ HRDC domains bind to the G4 or nearby DNA, orienting RecQ to unwind the *pilE* gene. RecJ degrades the unwound ssDNA behind RecQ. RecQ cannot unwind the *pilE* G4, so it remains folded and recruits RecA for strand exchange by homologous recombination. **Right**. Alternatively, HRDC domains could destabilize the G4, allowing for unwinding by another G4 resolving helicase. Once the G4 obstruction is removed, RecQ unwinds the *pilE* gene to facilitate RecA loading and homologous recombination.

Table 3.1. Plasmids used in this study

Plasmid	Properties	Source or Reference
pET28.b	Overexpression vector (Kan^R)	Novagen
pMK202	*N. gonorrhoeae recQ* in pET28.b (Kan^R)	Killoran MP, Kohler PL, Dillard JP, Keck JL. 2009. Mol Microbiol 71:158-71.
pMK214	pMK202 with additional 0.8 kn of genomic DNA downstream of *recQ* (Kan^R)	Killoran MP, Kohler PL, Dillard JP, Keck JL. 2009. Mol Microbiol 71:158-71.
pAV305	pMK214 with a Asp307Ala missense mutation and an NruI site at the mutation (Kan^R)	This work
pAV306	pMK214 with a Ser240Ala missense mutation and a BssHII site at the mutation (Kan^R)	This work
pIDN1	Cloning vector (Erm^R)	Hamilton HL, Schwartz KJ, Dillard 2001. J Bacteriol 183:4718-4726.
pMMC15	pAV305 ligated into pIDN1 (Erm^R)	This work
pMMC16	pAV306 ligated into pIDN1 (Erm^R)	This work
pAV338	Codon optimized *N. gonorrhoeae recQ* Asp307Ala in a pBAD.B overexpression plasmid (Amp^R)	Thermo Fisher

Table 3.2. Neisseria gonorrhoeae strains used in this study.

Strain	Properties	Source or Reference
MMC533	FA1090 recQ D312A	This work
MMC534	FA1090 recQ S245A	This work
MMC536	FA1090 recQ::erm	Killoran MP, Kohler PL, Dillard JP, Keck JL. 2009 Mol Microbiol 71:158-71.
FA1090	Wild type *N. gonorrhoeae*	Nachamkin I, Cannon JC, Mittler RS. 1981 Infecti Immun 32(2):641-648.
FA1090 *recA6*	FA1090 IPTG-inducible *recA*	Seifert HS. 1997. Gene 188(2):215-220.

Table 3.3. Oligos used in smFRET experiments

Oligo	Sequence
Common 18-mer	5′-Cy5-GCC TCG CTG CCG TCG CCA-biotin-3′
Non G-quadruplex DNA, T16	5′-TGG CGA CGG CAG CGA GGC-$(T)_{16}$-Cy3-3′
Telo G4-T15 DNA	5′-TGG CGA CGG CAG CGA GGC TTA GGG TTA GGG TTA GGG TTA GGG-$(T)_{15}$-Cy3-3′
cmyc-G4-T15 DNA	5′-TGG CGA CGG CAG CGA GGC TTG GGT G GGTAG GGT G GG-$(T)_{15}$-Cy3-3′

Table 3.4. DNA unwinding rates of the NgRecQs.

RecQ variant	Duplex unwinding (sec^{-1})	G4 unwinding rate (telo-G4) (sec^{-1})	G4 unwinding rate (*c-myc*) (sec^{-1})
NgRecQ	0.0223 ± 0.0007	0.0093 ± 0.0005	No unwinding
NgRecQ Asp307Ala	0.0079 ± 0.0011	No unwinding	No unwinding

Values are reported as ±1 standard deviation.

References

1 Sexually transmitted disease surveillance 2017. Atlanta: Centers for Disease Control and Prevention, 2018.

2 Kraus SJ. 1972. Complications of gonococcal infection. Med Clin North Am 56: 1115-1125.

3 Zowawi HM, Harris PNA, Roberts MJ, Tambyah PA, Schembri MA, Pezzani MD, Williamson DA, Paterson DL. 2015. The emerging threat of multidrug-resistant Gram-negative bacteria in urology. Nat Rev Urol 12:570-584.

4 Fifer H, Natarajan U, Jones L, Alexander S, Hughes G, Golparian D, Unemo M. 2016. Failure of dual antimircobial therapy in treatment of gonorrhea. N Engl J Med 374(22): 2504-2506.

5 Shimuta K, Ohnishi M, Nakayama S, Morita-Ishihara T, Unemo M, Furubayashi K, Kawahata T. 2014. Treatment failure with 2 g of azithromycin (extended-release formulation) in gonorrhoea in Japan caused by the international multidrug-resistant ST1407 strain of *Neisseria gonorrhoeae*. J Antimicrob Chemother 69(8):2086-2090.

6 Cahoon LA, Seifer HS. 2011. Focusing homologous recombination: pilin antigenic variation in the pathogenic *Neisseria*. Mol Microbiol 81(5):1136-1143.

7 Rotman E, Webber DM, Seifert HS. 2016. Analyzing *Neisseria gonorrhoeae* pilin antigenic variation using 454 sequencing technology. J Bacteriol 198(18):2470-2482.

8 Burch CL, Danaher RJ, Stein DC. 1997. Antigenic variation in *Neisseria gonorrhoeae*: production of multiple lipooligosaccharides. J Bacteriol 179:982-6.

9 Stern A, Brown M, Nickel P, Meyer TF. 1986. Opacity genes in *Neisseria gonorrhoeae*: control of phase and antigenic variation. Cell 47:61-71.

10 Criss AK, Kline KA, Seifert HS. 2005. The frequency and rate of pilin antigenic variation in *Neisseria gonorrhoeae*. Mol Microbiol 58:510-9.

11 Hamrick TS, Dempsey JA, Cohen MS, Cannon JG. 2001. Antigenic variation of gonococcal pilin expression in vivo: analysis of the strain FA1090 pilin repertoire and identification of the *pilS* gene copies recombining with *pilE* during experimental human infection. Microbiol 147:839-49.

12 Sechman EV, Rohrer MS, Seifert HS. 2005. A genetic screen identifies genes and sites involved in pilin antigenic variation in *Neisseria gonorrhoeae*. Mol Microbiol 57:468-83.

13 Cahoon LA, Seifert HS. 2009. An alternative DNA structure is necessary for pilin antigenic variation in *Neisseria gonorrhoeae*. Science 325:764-7.

14 Ambrus A, Chen D, Dai J, Jones RA, Yang D. 2005. Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-Quadruplex Stabilization. Biochemistry 44:2048-2058.

15 Wang Y, Patel DJ. 1993. Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. Structure 1:263-82.

16 Cahoon LA, Seifert HS. 2013. Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in *Neisseria gonorrhoeae*. PLOS Pathog 9:e1003074.

17 Wu X, Maizels N. 2001. Substrate-specific inhibition of RecQ helicase. Nucleic Acids Res 29:1765-1771.

18 Bernstein DA, Zittel MC, Keck JL. 2003. High-resolution structure of the *E. coli* RecQ helicase catalytic core. EMBO J 22:4910-21.

19 Cahoon LA, Manthei KA, Rotman E, Keck JL, Seifert HS. 2013. *Neisseria gonorrhoeae* RecQ helicase HRDC domains are essential for efficient binding and unwinding of the *pilE* guanine quartet structure required for pilin antigenic variation. J Bacteriol 195:2255-2261.

20 Killoran MP, Kohler PL, Dillard JP, Keck JL. 2009. RecQ DNA helicase HRDC domains are critical determinants in *Neisseria gonorrhoeae* pilin antigenic variation and DNA repair. Mol Microbiol 71:158-71.

21 Voter AF, Qiu Y, Tippana R, Myong S, Keck JL. 2018. A guanine-flipping and sequestration mechanism for G-quadruplex unwinding by RecQ helicases. Nat Commun 9:4201.

22 Hamilton HL, Schwartz KJ, Dillard JP. 2001. Insertion-duplication mutagenesis of *Neisseria*: use in characterization of DNA transfer genes in the gonococcal genetic island. J Bacteriol 183:4718-4726.

23 Dillard JP. 2011. Genetic manipulation of *Neisseria gonorrhoeae*. Curr Protoc Microbiol 23:4A.2.1-4A.2.24.

24 Seifert HS. 1997. Insertionally inactivated and inducible recA allels for use in *Neisseria*. Gene 188(2):215-220.

25 Smargiasso N, Rosu F, Hsia W, Colson P, Baker ES, Bowers MT, De Pauw E, Gabelica V. 2008. G-Quadruplex DNA assemblies: loop length, cation identity, and multimer formation. J Am Chem Soc 130:10208-10216.

26 Schultz SG, Solomon AK. 1961. Cation transport in *Escherichia coli*. I. Intracellular Na and K concentrations and net cation movement. J Gen Physiol 45:355-69.

27 Kuryavyi V, Cahoon Laty A, Seifert HS, Patel Dinshaw J. 2012. RecA-binding pilE G4 sequence essential for pilin antigenic variation forms monomeric and 5′ end-stacked dimeric parallel G-quadruplexes. Structure 20:2090-2102.

28 Shukla K, Thakur RS, Ganguli D, Rao DN, Nagaraju G. 2017. *Escherichia coli* and *Neisseria gonorrhoeae* UvrD helicase unwinds G4 DNA structures. Biochem J 474:3579-3597.

29 Thakur RS, Desingu A, Basavaraju S, Subramanya S, Rao DN, Nagaraju G. 2014. *Mycobacterium tuberculosis* DinG is a structure-specific helicase that unwinds G4 DNA: implications for targeting G4 DNA as a novel therapeutic approach. J Biol Chem 289:25112-25136.

Chapter 4

Summary and Future Directions

Summary

G-quadruplexes (G4s) are remarkably stable secondary structures that can form in DNA or RNA. The stability of these structures is a potentially lethal threat to cells, as G4s have been shown to interrupt essential cellular processes, including DNA replication, transcription and translation. Given the fitness costs associated with maintaining G4 forming sequences within the genome, it might be expected that the loss of these sequences would be selected for. Nevertheless, G4 forming sequences are found in all domains of life, often associated with a regulatory function. This suggests that cells have developed extensive mechanisms to tolerate the formation of G4s. Second,

A mechanism of G4 unwinding by RecQ helicases.

In chapter 2, I described a combined structural and biochemical approach to understanding the mechanisms of G4 unwinding by RecQ helicases. I solved the X-ray crystal structure of a bacterial RecQ helicase bound to a resolved G4. Examination of this structure revealed the presence of a novel pocket adjacent to the presumed G4 binding site, which I have term the guanine-specific pocket (GSP). One of the guanines expected to be folded within the G4 was instead base flipped and sequestered within the GSP, in a position that was inconsistent with G4 folding. Through a collaboration with the Myong lab at Johns Hopkins, we found that the GSP was essential for the G4 unwinding activity of RecQ but was dispensable for unwinding duplex DNA and binding G4s. I observed similar pockets in the structures of UvrD and BLM, two other G4 unwinding helicases, suggesting that a base-flipping mechanism may be shared among G4 unwinding helicases. The discovery of this pocket allows for the generation of helicase variants that are selectively unable to unwind G4 DNA. These separation-of-function variants are ideal for isolating the specific role of the G4 resolving capacity of helicase, as exemplified in chapter 3.

RecQ-mediated G4 unwinding in *Neisseria gonorrhoeae* antigenic variation

The pathogenic bacteria *Neisseria gonorrhoeae* varies the composition of its pilin proteins to evade the immune system during an infection. This process, known as antigenic variation (AV), is dependent on the formation of the *pilE* G4. RecQ aids AV and because RecQ is known to unwind G4 DNA, it was proposed that RecQ's critical role in AV was unwinding of the *pilE* G4. However, RecQ also supports homologous recombination independent of its G4 resolving capacity. A separation-of-function variant of RecQ was needed to isolate the activities of RecQ and determine RecQ's role in AV. My structural work described in Chapter 2 allowed me to generate a *N. gonorrhoeae* RecQ variant with selectively impaired G4 unwinding capabilities. In collaboration with the Dillard lab, a *N. gonorrhoeae* strain encoding this variant was produced and was found to undergo AV at wild type levels, indicating that AV is independent of G4 resolution by RecQ. Single-molecule helicase assays conducted with the Myong lab confirmed that the NgRecQ variant was incapable of antiparallel G4s, while neither the variant nor wild type NgRecQ were competent to unwind the parallel *pilE* G4. Therefore, the role of RecQ in AV appears to be limited to promoting homologous recombination, independent of its capacity to resolve G4s.

Future directions:

Identify and characterize GSPs in other G4 unwinding helicases.

To date, only two structural mechanisms of G4 unwinding helicase have been proposed. The first, describing the mechanism of the DHX36 helicase, requires the presence of a DHX-specific domain. As implied by the name, this domain is found only in the DHX36 family of helicases and so this mechanism is not readily generalizable to other helicases. The second mechanism, described in Chapter 2, utilizes a pocket in the helicase to base flip a guanine out of the G4, destabilizing it. While I described this mechanism using the structure of a bacterial RecQ helicase, a base flipping mechanism of G4 resolution might be applicable to a broad range of helicases. Accordingly, similar pockets have been identified in other G4 resolving helicases, including BLM and UvrD (Figure 2.12). To determine if these helicases employ a comparable mechanism, the residues comprising these pockets will be mutated and the resulting proteins will be tested for their ability to unwind G4 DNA in helicase assays. Duplex unwinding capability of these proteins will also be tested to assess the specificity of the pocket for G4 DNA. If these residues are found to be unimportant for G4 unwinding, then crystallographic studies will be initiated to identify novel G4 unwinding mechanisms.

Genetic screening to identify G4 repair pathways.

Unresolved G4 DNA is a potentially lethal obstruction to cells and pathways have evolved to unfold G4s or repair the damage resulting from a replication fork collision. While eukaryotic G4 repair pathways have begun to be described, the analogous systems in bacteria are unknown. I have started to explore the mechanisms of G4 tolerance in *E. coli*, using the compound N-methylmesoporophyin IX (NMM) to block G4 unwinding and force repair of the resulting DNA lesions. A pilot transposon-sequencing screen identified the AcrAB-TolC efflux pump and the RecBCD/RecA homolgous recombination repair pathways as critical to continued *E. coli* growth on NMM.

While informative, this screening effort is necessarily limited to testing the non-essential genes in *E. coli*, as any transposon insertion that inactive essential genes are, by definition, lethal. Therefore, to assess the role of essential genes in the repair of unresolved G4s, a CRISPRi knockdown approach will be used. Additionally, a G4 pulldown approach will be used to identify novel G4-interacting proteins. Through these systematic and unbiased screens, the complete set of G4 interacting proteins and repair pathways will be uncovered.

The role of RecA in *N. gonorrhoeae* antigenic variation

The results described in chapter 3 revealed that antigenic variation (AV) in *N. gonorrhoeae* occurs independently of RecQ-mediated resolution of the *pilE* G4. However, the question remains - why is the *pilE* G4 essential to AV? One insight into this problem was observation that RecA binds to the *pilE* G4 with high affinity and that G4s in a DNA substrate aids RecA-mediated strand exchange. In this scenario, G4 formation would be necessary to recruit RecA to *pilE* and initiate AV. However, because the RecA is required for AV independent of its affinity for G4s, this hypothesis cannot be tested by a simple knockout. Instead, in parallel to chapter 4, a G4-insensitive RecA variant is needed. Unfortunately, no RecA variants without affinity for G4 substrates are currently known.

To generate these variants and test this hypothesis, better structural information is required. *In silico* studies that I have performed suggest that RecA binds to G4s using a binding site formed between the L1 and L2 loops. This is the same site that encloses a triplet of nucleotides when RecA filaments on ssDNA (which accounts for the three nucleotide site size of RecA). To explore the nature of this interaction, we will solve the X-ray crystal structure of the RecA-G4 complex. This structure will allow us to identify residues that are critical for the G4 interaction and we will test their role by making RecA variants and measuring the affinity to the G4 by fluorescence anisotropy. Variants that have been shown to retain RecA functionality in prior studies will be prioritized, with the goal of generating a variant that cannot bind G4s

yet is competent to carry out homologous recombination. Generating *Neisseria gonorrhoeae* strains with this *recA* variant will allow us to determine the precise role of the RecA-*pilE* G4 interaction in AV.

Appendix 1

A high throughput screening strategy to identify protein-protein interaction inhibitors that block the Fanconi anemia DNA repair pathway

This work has been published:

Voter, A. F., Manthei, K. A., and Keck, J. L. (2016) A high throughput screening strategy to identify protein-protein interaction inhibitors that block the Fanconi anemia DNA repair pathway. *Journal of Biomolecular Screening* **21**(6), 626-633.

Kelly Manthei and Andrew Voter performed protein purification. Kelly Manthei developed and performed the fluorescence polarization screen. Andrew Voter developed and performed the alphascreen assay, performed secondary screening and biophysical characterization of the lead compound.

Abstract

Induction of the Fanconi anemia (FA) DNA repair pathway is a common mechanism by which tumors evolve resistance to DNA crosslinking chemotherapies. Proper execution of the FA pathway requires interaction between the FA complementation group M protein (FANCM) and the RecQ-mediated genome instability protein (RMI) complex, and mutations that disrupt FANCM/RMI interactions sensitize cells to DNA crosslinking agents. Inhibitors that block FANCM/RMI complex formation could be useful therapeutics for re-sensitizing tumors that have acquired chemotherapeutic resistance. To identify such inhibitors, we have developed and validated high-throughput fluorescence polarization and proximity assays that are sensitive to inhibitors that disrupt interactions between the RMI complex and its binding site on FANCM (a peptide referred to as MM2). A pilot screen of 74,807 small molecules was performed using the fluorescence polarization assay. Hits from the primary screen were further tested using the proximity assay and an orthogonal proximity assay was used to assess inhibitor selectivity. Direct physical interaction between the RMI complex and the most selective inhibitor identified through the screening process was measured by surface plasmon resonance and isothermal titration calorimetry. Observation of direct binding by this small molecule validates the screening protocol.

Introduction

Genomic instability is a hallmark of cancer that arises from the inactivation of DNA repair pathways during tumorigenesis.[1] This defect is exploited by many cancer chemotherapeutics that act by indiscriminately damaging DNA; cancerous cells lacking robust DNA repair capacity cannot survive chemotherapeutic doses that are tolerated by healthy tissue. Although DNA damaging chemotherapies are often initially effective, reactivation of tumor DNA repair pathways can lead to treatment failure and poor patient outcomes.[2]

DNA crosslinking agents, such as cisplatin and mitomycin C, are first-line therapies for a range of malignancies including testicular,[3] lung,[4] and ovarian cancers.[5] Crosslinking agents act by covalently binding two DNA strands together, and the resulting inter-strand crosslinks (ICLs) block DNA replication and transcription, leading to cell death unless the crosslinks are promptly repaired.[6] ICLs formed during S phase stall replication forks at crosslinks, activating the Fanconi anemia (FA) repair pathway. Non-dividing cells or cells in the G1 cell cycle phase lack replication machinery and instead use nucleotide excision repair for ICL removal.[7] The FA pathway is commonly inactivated during tumorigenesis; reactivation or upregulation of FA pathway has been linked to chemotherapy resistance in multiple myeloma,[8] leukemia,[9] gliomas,[10] squamous cell head and neck tumors,[11] and ovarian cancer.[12,13] Because non-cancerous tissues maintain a functional alternative repair mechanism, reliance on the FA pathway is relatively specific for resistant tumors and its disruption is hypothesized to restore sensitivity to crosslinking agents.[14]

The FA pathway is initiated by binding of the FA complementation group M protein (FANCM) to ICL DNA at which two replication forks have collided.[15,16] FANCM subsequently recruits two DNA repair complexes, the FA core complex and the Bloom dissolvasome, to the lesion via protein-protein interactions.[17] The FA core complex directs the excision of the crosslink and bypass of one of the strands by a translesion DNA polymerase. The newly repaired strand serves as a template for homologous

recombination to repair the remaining double strand break.[15] This process results in the formation of a double Holliday junction DNA structure, which can lead to sister chromatid exchange events if not resolved by the Bloom dissolvasome.[18] The Bloom dissolvasome is comprised of the Bloom DNA helicase, topoisomerase IIIα, and a heterodimeric subcomplex of "*R*ecQ-*m*ediated genome *i*nstability" proteins, RMI1 and RMI2[19]. The RMI complex anchors the Bloom dissolvasome to FANCM by binding to a 34 amino acid motif within FANCM called MM2.[17,20]

We and others have demonstrated that the interaction between RMI1/2 and MM2 is required for repair of DNA crosslinks.[17,21,22] The introduction of point mutations in either RMI1/2 or MM2 that disrupt the association leads to genomic instability, as measured by increases in sister chromatid exchanges. Additionally, our lab has determined the X-ray crystal structures of the RMI core complex (comprised of the OB2 domain of RMI1 and the entirety of RMI2)[22] and of the RMI core complex bound to MM2.[21] Along with biochemical and cellular studies, these structures have defined a binding pocket formed by RMI1/2 that is essential for MM2 binding. Introduction of a single lysine-to-alanine mutation in the RMI core complex pocket (K121 of RMI 1) reduces the affinity for MM2 by over 80-fold, suggesting the pocket is a "hotspot" for anchoring MM2 onto the RMI1/2 complex. These data further suggest that the RMI/MM2 interaction could be amendable to disruption by small molecules that compete with MM2 for binding to this critical pocket. Such inhibitors could be of value as research probes and in the development of therapeutics that sensitize resistant tumors to DNA crosslinking chemotherapeutics.

To identify small molecule inhibitors that block MM2 interaction with the RMI proteins, we have developed two high-throughput-ready assays that measure interaction between the MM2 peptide from FANCM and the RMI core complex. A 74,807-compound library was screened using a fluorescence polarization (FP)-based assay and hits were rescreened using a proximity assay. Counter-screening against an orthogonal proximity assay led to the identification of a single compound that specifically disrupted the RMI core complex/MM2 interaction. Direct binding of this compound to the RMI core complex was

confirmed by surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC). Success of this pilot screen supports future screens against larger libraries of compounds and structure-activity relationship studies to improve potency of the identified inhibitor.

Materials and Methods

Protein purification. Expression and purification of the RMI core complex, MM2 peptide, control MM2 variant peptide incapable of binding RMI (cMM2, containing F1232A and F1236A mutations), and fluorescein labeled MM2 (F-MM2) were performed as previously described.[22,21] MM2 was biotinylated (Bio-MM2) with the EZ-link NHS-PEG_4-Biotin kit (ThermoFisher, Waltham, MA) according to the manufacture provided directions. Expression and purification of the RMI core complex with a N-terminal 6X-His tagged RMI2, was performed in an identical manner as unlabeled RMI core complex except that the thrombin protease site linking RMI2 and the His tag was mutated to prevent removal of the His tag. A peptide (SSBct) and a biotinylated variant (Bio-SSBct) containing the 8 residues from the carboxyl-terminus of *E. coli* single stranded DNA binding protein was purchased from the University of Wisconsin Biotechnology center (Madison, WI). *E. coli* PriA was purified as previously described.[23]

Fluorescence polarization. All FP measurements were carried out in black 384-well plates (ThermoFisher, Waltham, MA). For IC_{50} determinations, F-MM2 and RMI core complex were preincubated in 10 mM Tris-HCl, pH 8.8, 1 mM dithiothreitol (DTT). Unlabeled MM2 was serially diluted, added to the F-MM2/RMI core complex mixture to a final concentration of 7 nM F-MM2 and 100 nM RMI core complex and covered with a foil plate seal. After incubation for at least 20 min, FP was measured on a Tecan Biotek “synergy 2” plate reader.

To assess the suitability of the FP assay for high-throughput screen (HTS) applications, 100 nM RMI core complex and 7 nM F-MM2 in 10 mM Tris-HCl, pH 8.8, 1.0 mM DTT, 7.5% DMSO was mixed with 8 μM MM2 or SSBct peptide (positive or negative controls, respectively). After 20 minutes, the mixture was dispensed by multichannel pipet, centrifuged, and FP values were measured on a Biotek “Synergy 2” plate reader (128 wells of each peptide), independently repeated over 3 days. The Z’ score was calculated by Eq. (A1.1):[24]

$$Eq.\ (A1.1) \qquad Z' = 1 - \frac{3(\sigma_{pos} + \sigma_{neg})}{\mu_{pos} - \mu_{neg}}$$

FP HTS. Screening took place at the University of Wisconsin Small Molecule Screening and Synthesis Facility. A master mix of RMI core complex and F-MM2 (30 µL per well) was plated in black 384 shallow well plates (ThermoFisher, Waltham, MA), using a BioTek "MicroFlo Select" reagent dispenser. Compounds were added using a Beckman FX liquid handler; 0.33 µL of 10 mM stock was added for a final compound concentration of 33 µM. MM2 and cMM2 were each added to 4 wells of master mix per plate to a final concentration of 10 µM to serve as controls. Following compound addition, plates were covered, centrifuged briefly, and incubated for 20 minutes at room temperature. FP measurements were taking using a Tecan "Safire 2" microplate reader. Instrument settings were as follows: top read, EX 470, EM 525/20, G-factor 0.89947. A suitable gain was calculated from the first plate of each day. Z' scores for were calculated for each plate; plates with Z' scores <0.5 were rerun prior to analysis. All FP measurements from the primary screen will be made available on PubChem prior to publication.

Screen library composition. A total of 74,807 compounds were screened from the following compound libraries maintained by the University of Wisconsin Small Molecule Screening and Synthesis Facility: Life Chemicals library of ~50,000 compounds, Maybridge HitFinder library of ~14,400 compounds, the NIH clinical collection of 4,709 compounds, Prestwick library of 1,280 compounds, the spectrum collection of 2,000 compounds, and the JDRF TGF-β collection of 2,418 compounds. PIP-199, the most selective inhibitor discovered in the screen, was purchased from Life Chemicals (Burlington, ON, Canada)

Proximity screen (Alphascreen). For determination of AlphaScreen IC_{50} values, the inhibitor was titrated into a fixed amount of Bio-MM2 and His-tagged RMI core complex under subdued lighting conditions. The final reaction mixture contained 30 nM Bio-MM2, 100 nM His-tagged RMI core complex, 30 mM MOPS-

HCl, pH 7.2, 0.05% (v/v) Triton X-100 to a final reaction volume of 10 µL. The white 384 well plate (ThermoFisher, Waltham, MA) was sealed with foil, centrifuged and incubated for at least 2 hours prior to measurement on a Tecan "M1000" plate reader.

For validation of the AlphaScreen assay under high-throughput conditions, 10 µL of reaction mixture containing 30 nM Bio-MM2, 100 nM His-tagged RMI core complex, 30 mM MOPS-HCl, pH 7.2, 0.05% (v/v) Triton X-100, 5% (v/v) DMSO and 5 µM of either SSBct or unlableled MM2 (negative and positive controls, respectively) were added to white 384-well plates by multichannel micropipette. Plates were covered with foil seals, centrifuged, and incubated for 2 hours prior to measurement. Large edge effects were noted on the extreme rows of the plate; these rows were omitted during subsequent experiments. A Z' score was calculated using eq (A1.1).

The PriA-SSB AS was prepared and analyzed as above, except the 10 µL reaction contained 100 µM inhibitor in a final mixture of 100 nM PriA, 100 nM Bio-SSBct, 10 mM HEPES-HCl, pH 7.4, 150 mM sodium chloride, 1 mM magnesium chloride, 10 mM DTT, 1 mg/mL bovine serum albumin and 0.01% (v/v) Triton X-100.

Isothermal Titration Calorimetry. RMI core complex was dialyzed against 30 mM potassium phosphate, pH 7.0, 100 mM sodium chloride, 10% (v/v) glycerol overnight at 4°C. The sample was diluted and DMSO added to a final concentration of 1.5% (v/v) and 300 µM RMI core complex. PIP-199 dissolved in DMSO was diluted in the dialysis buffer to a final concentration of 30 µM and 1.5% (v/v) DMSO. RMI core complex was titrated into the sample cell containing PIP-199 solution maintained at 25°C using a MicroCal™ VP-ITC (GE Healthcare, Little Chalfont, United Kingdom). Five 1 µL injections were followed by 14 injections of 1.6 µL. Data analysis was performed with Origin software using a single-site binding model.

Surface Plasmon Resonance. SPR experiments were performed using a Bio-Rad "ProteOn XPR36" system with ProteOn GLH sensor chips (Bio-Rad, Hercules, CA). Phosphate buffered saline with detergent and

DMSO (137 mM sodium chloride, 2.7 mM potassium chloride, 10 mM disodium phosphate, 1.8 mM monopotassium phosphate, 0.01% (v/v) Triton X-100, 1.5% (v/v) DMSO, pH 7.2) was used as running buffer throughout. RMI core complex was immobilized unto the sensor chip by amine coupling in 10 mM $NaCH_3O_2$, pH 5.5. PIP-199 was serially diluted in running buffer containing 1.5% (v/v) DMSO from 150 µM to 9 µM using 2-fold dilutions and injected over the immobilized RMI core complex. Running buffer was injected simultaneously as a reference and subtracted from all traces. Analysis of SPR data was conducted using ProteOn Manager™ software. Data from each ligand surface were grouped to fit k_a, k_d, and R_{max} with a Langmuir kinetic model. The dissociation constant, K_d, was calculated from the equation $K_d = k_d/k_a$.

Statistical analysis. All analysis of dose response curves was carried out in Prism version 5.0c (GraphPad, La Jolla, CA) using a 4-parameter logistic fit to determine IC_{50} values.

Results

Development of the Primary FP Screen. To identify inhibitors of the RMI core complex/MM2 interaction, we adapted a previously developed FP assay[21] for use in high throughput format. In this FP assay, RMI core complex is incubated with fluorescein labeled MM2 peptide (F-MM2). After equilibration, F-MM2 is excited by polarized light. Free F-MM2 tumbles rapidly in solution and the emitted light is less polarized relative to emissions from RMI core complex-bound F-MM2. The fraction of free F-MM2 may then be calculated from the ratio of unpolarized to polarized emission intensity.

In a prior study, we determined the K_d of the RMI/F-MM2 to be <5 nM and showed that unlabeled MM2 competed with F-MM2 with an IC_{50} of 520±50 nM.[21] To adapt the assay for high-throughput screening, we transitioned to a 384 well format and evaluated assay performance. An IC_{50} of 510±20 nM was observed by titration of unlabeled MM2 into a fixed concentration of performed RMI core complex/F-MM2 (Figure A1.1A and A1.1B), in excellent agreement with the previously determined value. To further assess assay reproducibility and uniformity in a high-throughput format, RMI core complex and F-MM2 were incubated in 384-well plates in the presence of 7.5% DMSO and 8 μM of either unlabeled MM2 or an unrelated peptide, SSBct, serving as positive and negative controls, respectively. We observed a Z' score of 0.53 over 3 days (n = 128 wells per control per day, 384 total), demonstrating the suitability of our FP assay for high throughput screening (Figure A1.1C).

Development of the Secondary AS Screen. Because small molecules with intrinsic fluorescence or fluorescence quenching properties may be falsely identified as hits in FP assays, we adapted an AlphaScreen (AS) proximity assay for use with the RMI core complex/MM2 interaction to serve as a secondary screen. AS is a bead-based proximity assay using donor and acceptor beads that are tethered to the interaction partners. Stimulation of the donor bead with 680 nm light generates singlet oxygen. If the singlet oxygen encounters an acceptor bead, a chemical reaction on the acceptor bead results in the

emission of 570 nm light. The short half-life of the singlet oxygen ensures that signal is produced only when the interacting partners are in contact.

MM2 was biotinylated to allow for association with streptavidin-coated donor beads and an N-terminal 6X-His tagged version of RMI2 within the RMI core complex was bound to the Ni^{2+}-coated acceptor beads (Figure A1.2A). Titrating unlabeled MM2 into the AS assay disrupted the RMI core complex/MM2 complex with an IC_{50} of 180±20 nM, modestly lower than the FP assay (Figure A1.2B). To validate our assay for high-throughput use, the AS assay was performed in 384 well plates in the presence of either 5 μM unlabeled MM2 or SSBct as positive or negative controls, respectively. There was large day-to-day variation in the maximum signal of the AS, likely resulting from pipetting error in the addition of the AS beads to the reaction mixture. To allow for day-to-day comparison, the average maximum and minimum signals for each day were normalized to 100 and 0 respectively. Our AS assay proved suitable for HTS with Z' scores ≥ 0.7 for each day (n = 88 per control), with an overall Z' of 0.75 (n = 264 per control) (Figure A1.2C).

High throughput pilot screen. To assess the effectiveness of our screening strategy for identifying small molecule inhibitors of the RMI core complex/MM2 interaction, we conducted a pilot HTS campaign by screening 74,807 compounds at the Small Molecule Screening and Synthesis Facility at the University of Wisconsin. The primary FP screen was performed in 384 well plates, with each well containing 100 nM RMI core complex and 7 nM F-MM2. Each plate included 4 positive and 4 negative control wells for Z' calculations with each plate. Small molecules dissolved in DMSO were individually added to wells to final small molecule concentration of 32 μM and the polarization of each well was determined. Plates with individual Z' scores of <0.5 were rescreened prior to analysis. We identified 415 hits (0.55% hit rate), defined as compounds that produced FP ≥2 standard deviations below the average FP of the plate (Figure A1.3). These compounds were rescreened in the FP assay at 320, 160, 32 and 3.2 μM. Sixty-eight

compounds produced a dose dependent decrease in polarization and were advanced to the secondary AS.

Each compound was added to AS reactions at 100 μM, with 18 of the 68 compounds identified in the FP assay producing ≥ 50% decrease in AS signal. To exclude small molecules acting in a non-specific manner, we tested these compounds in an AS assay developed against an unrelated bacterial protein-protein interaction (PriA/SSBct) at 100 μM. Seven compounds were found to also inhibit the PriA-SSBct interaction and were excluded. As the eleven remaining compounds exhibited significant structural similarities, stocks of seven of the most distinct compounds were purchased for further evaluation. Upon receipt, compounds were assayed against both the RMI core complex/MM2 and PriA/SSBct AS assays. A single compound, which we have named PIP-199 (Figure A1.4A), exhibited selective inhibition of the RMI core complex/MM2 complex formation with an IC_{50} of 36±10 μM (Figure A1.4B), while the PriA-SSB AS was inhibited with an IC_{50} of 450±130 μM. Repurchased PIP-199 was rescreened against the RMI core complex/MM2 FP assay and found to inhibit with an IC_{50} of 260±110 μM (Figure A1.4C).

Confirmation of direct physical binding of PIP-199 to RMI core complex. Because of the disparate IC_{50} values obtained in the primary and secondary assays, we sought to confirm direct binding of PIP-199 to the RMI core complex. SPR has been shown to be capable of detecting small molecule binding to proteins in a semi-high throughput fashion.[25] To detect interactions via SPR, light is shined onto a gold chip bound by a receptor protein (RMI core complex) at an angle and then reflected onto a detector. A fraction of the light is not reflected but is absorbed to excite a resonant surface plasmon on the chip; the angle at which the absorbed light is reflected, or resonance angle, is highly dependent on the conditions at the chip surface. Binding of a small molecule (such as PIP-199) to the receptor alters the surface plasmon and is detected as a change in the resonance angle (Figure A1.5A).

Anticipating the need to rapidly screen for physical binding as a part of a much larger screen, we sought to determine if the RMI core complex/PIP-199 interaction could be detected by SPR. Buffer containing varying amounts of PIP-199 was flowed over the immobilized RMI core complex and a dose-dependent change in the resonance angle was observed (Figure A1.5B). A K_d of 7.3±0.8 μM (RU_{max} = 52 RU, χ^2 =15 RU) was calculated from the fit k_a and k_d (Figure A1.5B). Non-specific small molecule binding to RMI core complex limited the quality of the fit, indicated by the relatively high observed χ^2/RU_{max} (0.29, <0.1 is ideal).

To assess the reliability of the K_d obtained by SPR, we turned to isothermal titration calorimetry (ITC). In ITC, one interacting partner is titrated into a solution containing the other interacting partner. The heat evolved or absorbed from binding is measured by comparison to a reference cell lacking the interaction partners (Figure A1.5C). RMI core complex was titrated into a solution of PIP-199 and was found to bind with a K_d of 3.4±1.0 μM (Figure A1.5D), in reasonable agreement with the K_d obtained from SPR. Each PIP-199 was calculated to interact with 0.68±0.05 RMI core complexes, rather than the expected ratio of 1.0. This discrepancy likely results from the accumulation of small volumetric errors in the solubilization and dilution of the compound. Detection of a direct biophysical interaction by SPR and ITC suggests that activity in the FP and AS assays by PIP-199 is the result of true inhibition and not merely an assay artifact.

Discussion

Elevated activity of the FA DNA repair pathways has been implicated as a cause of chemotherapeutic resistance in a broad range of cancers, suggesting that targeted inhibition of the FA pathway could re-sensitize resistant tumors to ICL-forming chemotherapies.[11,8] We and others have observed that destabilization of the RMI/MM2 interaction leads to a sensitization to cross-linking agents and an increase in genomic instability in cells.[18,19] To screen for inhibitors that disrupt this interface, we have developed a HTS strategy that has identified RMI core complex/MM2 interaction inhibitors and biophysical assays showing that our most selective compound binds directly to the RMI core complex.

The first stage of our strategy uses an FP screen, followed by an orthogonal AlphaScreen to eliminate non-specific inhibition. Both assays are suitable for use in HTS campaigns with Z' scores of 0.53 and 0.75 for the FP and AS assays, respectively. Our pilot screen of 74,807 compounds yielded a single selective inhibitor of modest potency, a 0.001% overall hit rate. The low hit rate likely results from the high affinity of the RMI core complex/MM2 interaction (apparent K_d <5 nM). Only small molecules with a high affinity for the RMI pocket or an allosteric site would be capable of disrupting the RMI core complex/MM2 interaction and these are expected to occur at a low frequency in a screening library. In a previous study, we identified MM2 variants with lower affinities for the RMI core complex.[18] Interactions with these variants are more easily disrupted and could complement the primary screen as a method to identify additional scaffolds for optimization.

One limitation of the screening method described here is the use of the RMI core complex and MM2 peptide in place of the full Bloom dissolvasome and full-length FANCM. The RMI core complex and MM2 peptide are stable and easily purified, which are essential for production of reagents needed for reproducible performance in an HTS. One potential complication of using minimal domains is that sites available for inhibition in our HTS may be obscured *in vivo* where full-length proteins and complexes exist.

Activity against full-length proteins in a cellular context will be an important step in future studies that seek to determine the cellular activities of PIP-199 and related compounds.

In conclusion, our pilot screen has identified a small molecule that disrupts the protein-protein interaction between the RMI core complex and the MM2 region from FANCM. Structural studies to define the PIP-199 binding sites on the RMI core complex and structure-activity relationship experiments to improve the activity of PIP-199 are currently underway. Future studies will test whether optimized, potent RMI inhibitors are able to block the FA DNA repair pathway in human cells. Such inhibitors will be valuable tools for the study of the mechanisms underlying DNA crosslink repair and could serve as lead compounds in developing new strategies for treating chemoresistant tumors.

Acknowledgements

The authors would like to thank the Gene Ananive from the University of Wisconsin Small Molecule Screening and Synthesis Facility for his assistance in carrying out the FP screen and Michael Killoran for the development of the PriA-SSB AS used as a counterscreen in this study. The project was supported by NIH R21 CA178475 (J.L.K.) and the Clinical and Translational Science Award program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427. K.A.M. was supported in part by an NIH Training Grant in Molecular Biosciences GM07217. A.F.V. is supported by the University of Wisconsin-Madison Integrated Training for Physician-Scientists NIH Training Grant GM008692. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. AS and SPR data were obtained at the University of Wisconsin - Madison Biophysics Instrumentation Facility, which was established with support from the University of Wisconsin - Madison and grants BIR-9512577 (NSF) and S10 RR13790 (NIH). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. A.F.V. conducted assay validation, secondary screening and biophysical analysis. K.A.M. designed and performed the FP screen. A.F.V., K.A.M., and J.L.K. carried out data analysis. A.F.V., K.A.M., and J.L.K. wrote the manuscript.

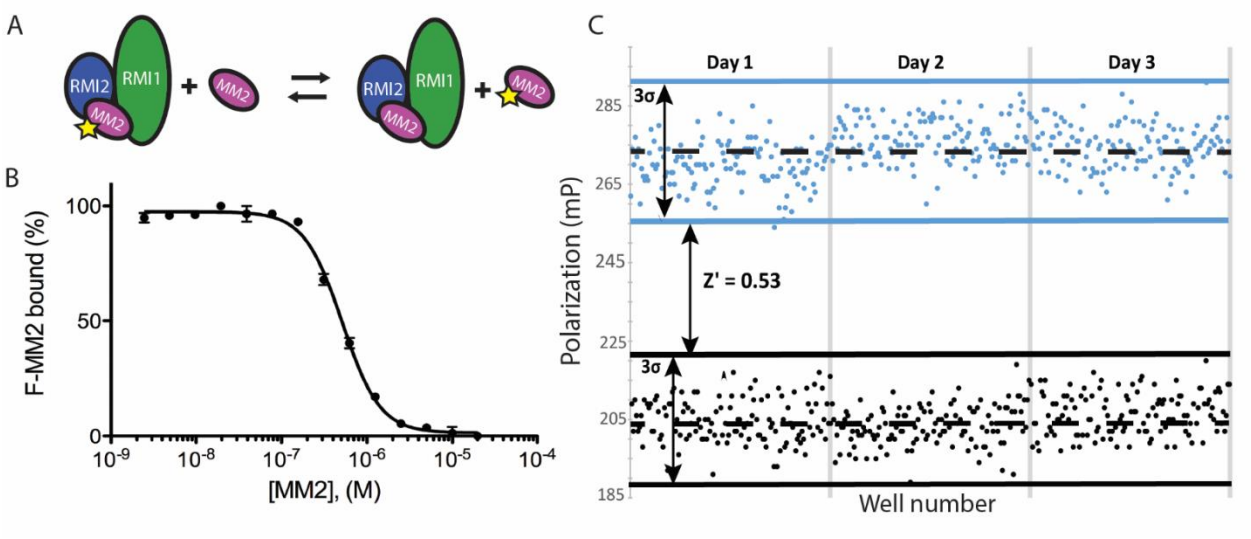


Figure A1.1. Characterization of the FP assay to identify inhibitors that disrupt interaction between the RMI core complex and the MM2 peptide from FANCM. A) Scheme of the FP assay. Preformed RMI core complex/F-MM2 complexes are incubated with increasing amounts unlabeled MM2, displacing F-MM2. **B)** Titration of unlabeled MM2 into a preformed RMI core complex/F-MM2 complex displaces F-MM2 under high-throughput conditions. Error bars represent the SEM of 3 independent reactions. **C)** Polarization of F-MM2 in the presence of RMI core complex and an excess of MM2 (black) or control peptide (blue) across 3 days. Dashed lines represent the mean FP for each condition, solid lines are 3 standard deviations from the mean.

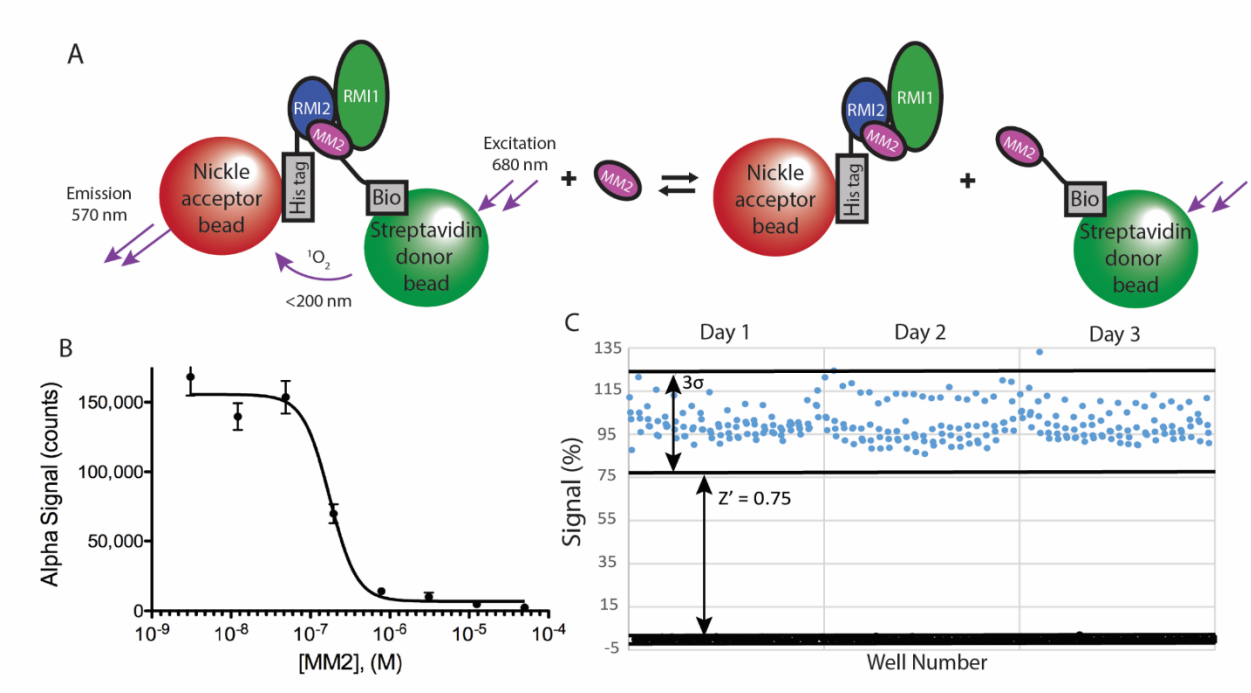


Figure A1.2. Characterization of the AS assay to identify inhibitors that disrupt interaction between the RMI core complex and the MM2 peptide from FANCM. A) Scheme depicting the RMI core complex/MM2 AS assay. **B)** Titration of unlabeled MM2 into a fixed concentration of preformed AS reaction mixture disrupts the RMI core complex/MM2 interaction. Error bars represent the SEM of 3 independent reactions. **C)** Validation of the RMI core complex/MM2 AS under high-throughput conditions. Preformed complexes of the AS beads, RMI core complex, and Bio-MM2 were incubated with an excess of MM2 (black) or control peptide (blue). Solid lines depict the mean signal of each condition and the dashed lines contain points within 3 standard deviations of the mean. Values are normalized; the daily average maximum signal is set as 100% and the average minimum as 0%.

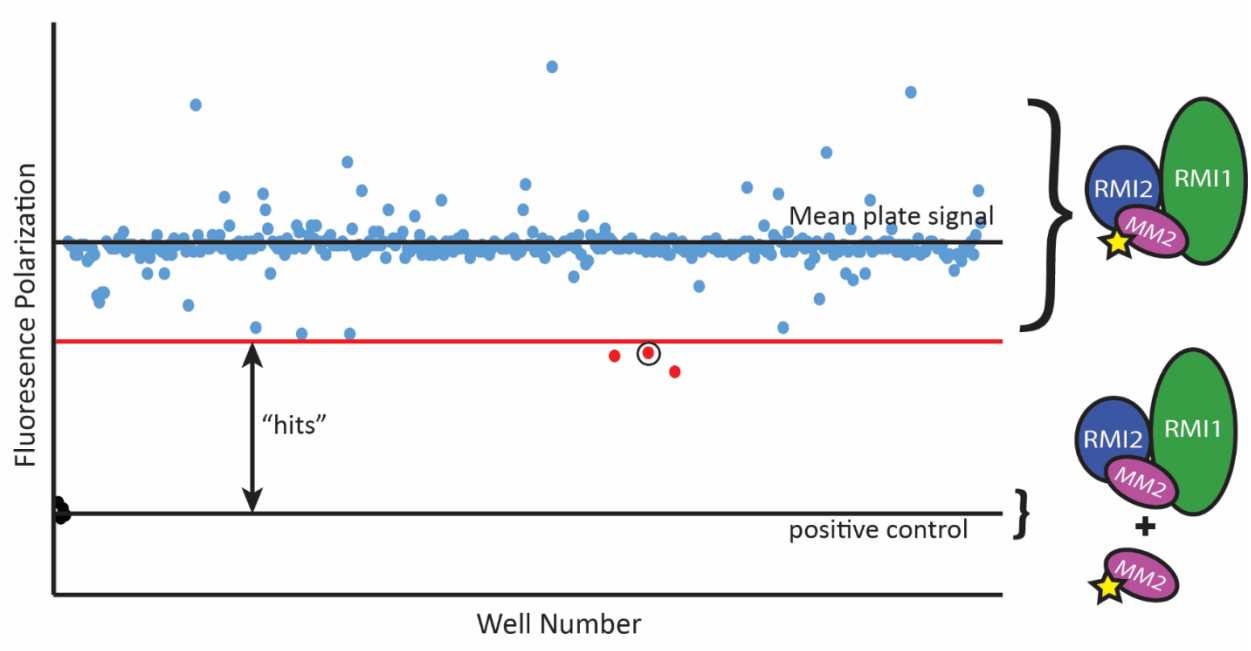


Figure A1.3. Representative plate from the high-throughput FP screen. Polarization values from each compound on the plate are reported. Compounds producing FP values ≥2 standard deviations below the mean plate polarization were advanced for further screening. The circled point is PIP-199.

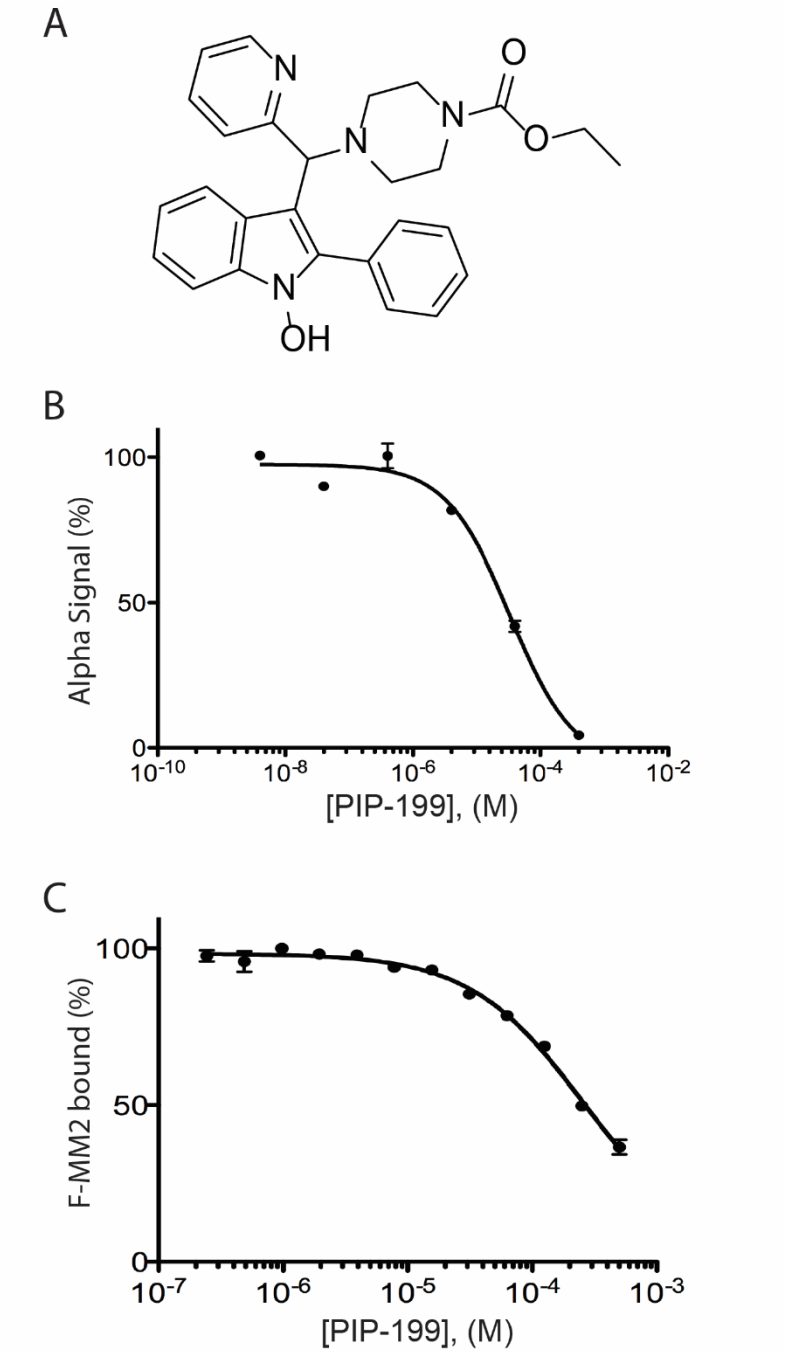


Figure A1.4. Characterization of the most selective inhibitor of the RMI core complex/MM2 interaction. **A)** Structure of PIP-199. **B)** Dose-response curve of PIP-199 in the AS assay, error bars represent the SEM of three independent experiments. **C)** Dose-response curve of PIP-199 in the FP assay, error bars represent the SEM of three independent experiments

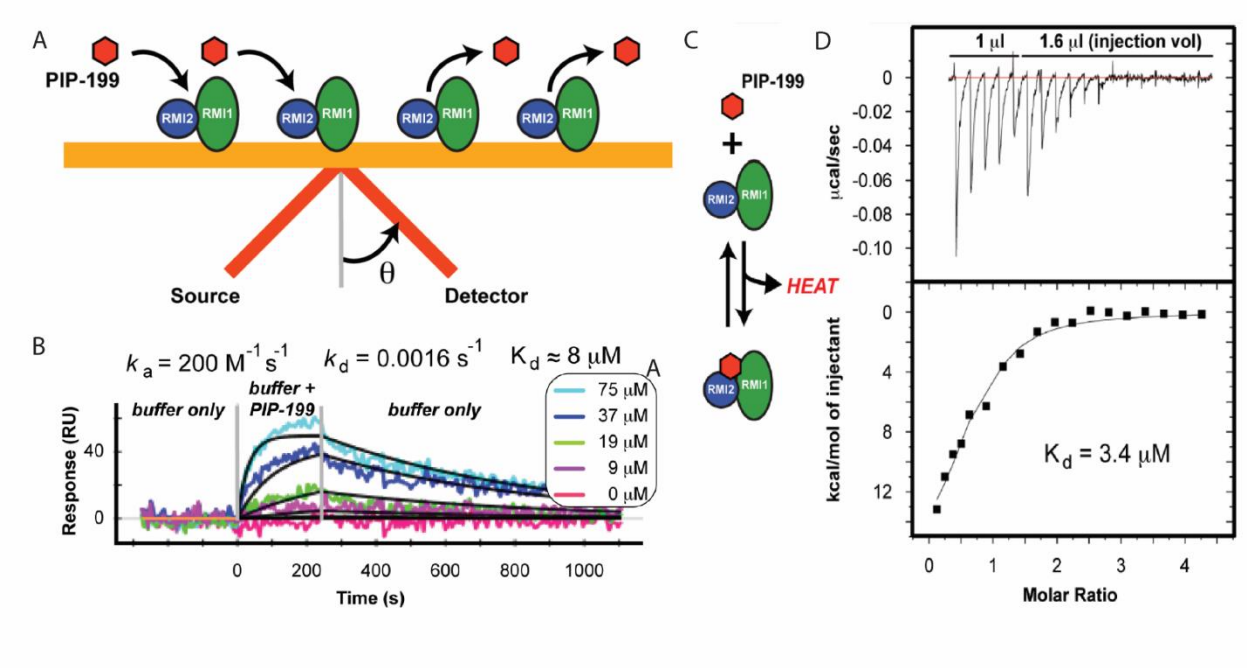


Figure A1.5. Biophysical confirmation of inhibitor binding to the RMI core complex. A) Scheme of the SPR assay. **B)** SPR results. Buffer containing indicated amounts of PIP-199 is flowed over immobilized RMI core complex (0-250 sec). Rates and binding constants are calculated from fits to data (black lines). **C)** Scheme of the ITC binding assay. **D)** Heat evolved from the titration of RMI core complex into a solution of PIP-199. Binding constant is calculated from a fit to data (black line).

References

1. Hanahan, D.; Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144* (5), 646–674.

2. Bouwman, P.; Jonkers, J. The Effects of Deregulated DNA Damage Signalling on Cancer Chemotherapy Response and Resistance. *Nat. Rev. Cancer* **2012**, *12* (9), 587–598.

3. Schmoll, H.-J.; Jordan, K.; Huddart, R.; et al. On behalf of the ESMO Guidelines Working Group. Testicular Seminoma: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* **2010**, *21* (Supplement 5), v140–v146.

4. D'Addario, G.; Fruh, M.; Reck, M.; et al. On behalf of the ESMO Guidelines Working Group. Metastatic Non-Small-Cell Lung Cancer: ESMO Clinical Practice Guidelines for Diagnosis, Treatment and Follow-Up. *Ann. Oncol.* **2010**, *21* (Supplement 5), v116–v119.

5. Oza, A. M.; Cook, A. D.; Pfisterer, J.; et al. Standard Chemotherapy with or without Bevacizumab for Women with Newly Diagnosed Ovarian Cancer (ICON7): Overall Survival Results of a Phase 3 Randomised Trial. *Lancet Oncol.* **2015**, *16* (8), 928–936.

6. Deans, A. J.; West, S. C. DNA Interstrand Crosslink Repair and Cancer. *Nat. Rev. Cancer* **2011**, *11* (7), 467–480.

7. Mouw, K. W.; D'Andrea, A. D. Crosstalk between the Nucleotide Excision Repair and Fanconi anemia/BRCA Pathways. *DNA Repair* **2014**, *19*, 130–134.

8. Chen, Q. The FA/BRCA Pathway Is Involved in Melphalan-Induced DNA Interstrand Cross-Link Repair and Accounts for Melphalan Resistance in Multiple Myeloma Cells. *Blood* **2005**, *106* (2), 698–705.

9. Yao, C.; Du, W.; Chen, H.; et al. The Fanconi anemia/BRCA Pathway Is Involved in DNA Interstrand Cross-Link Repair of Adriamycin-Resistant Leukemia Cells. *Leuk. Lymphoma* **2015**, *56* (3), 755–762.

10. Chen, C. C.; Taniguchi, T.; D'Andrea, A. The Fanconi Anemia (FA) Pathway Confers Glioma Resistance to DNA Alkylating Agents. *J. Mol. Med.* **2007**, *85* (5), 497–509.

11. Burkitt, K.; Ljungman, M. Phenylbutyrate Interferes with the Fanconi Anemia and BRCA Pathway and Sensitizes Head and Neck Cancer Cells to Cisplatin. *Mol. Cancer* **2008**, *7* (1), 24.

12. Toshiyasu T; Tischkowitz, M.; Ameziane, N.; et al. Disruption of the Fanconi Anemia-BRCA Pathway in Cisplatin-Sensitive Ovarian Tumors. *Nat. Med.* **2003**, *9* (5), 568–574.

13. Patch, A.-M.; Christie, E. L.; Etemadmoghadam, D.; et al. Whole–genome Characterization of Chemoresistant Ovarian Cancer. *Nature* **2015**, *521* (7553), 489–494.

14. Chirnomas, D. Chemosensitization to Cisplatin by Inhibitors of the Fanconi anemia/BRCA Pathway. *Mol. Cancer Ther.* **2006**, *5* (4), 952–961.

15. Räschle, M.; Knipscheer, P.; Enoiu, M.; et al. Mechanism of Replication-Coupled DNA Interstrand Crosslink Repair. *Cell* **2008**, *134* (6), 969–980.

16. Zhang, J.; Dewar, J. M.; Budzowska, M.; et al. DNA Interstrand Cross-Link Repair Requires Replication-Fork Convergence. *Nat. Struct. Mol. Biol.* **2015**, *22* (3), 242–247.

17. Deans, A. J.; West, S. C. FANCM Connects the Genome Instability Disorders Bloom's Syndrome and Fanconi Anemia. *Mol. Cell* **2009**, *36* (6), 943–953.

18. Seki, M.; Nakagawa, T.; Seki, T.; et al. Bloom Helicase and DNA Topoisomerase III Are Involved in the Dissolution of Sister Chromatids. *Mol. Cell. Biol.* **2006**, *26* (16), 6299–6307.

19. Singh, T. R.; Ali, A. M.; Busygina, V.; et al. BLAP18/RMI2, a Novel OB-Fold-Containing Protein, Is an Essential Component of the Bloom Helicase-Double Holliday Junction Dissolvasome. *Genes Dev.* **2008**, *22* (20), 2856–2868.

20. Manthei, K. A.; Keck, J. L. The BLM Dissolvasome in DNA Replication and Repair. *Cell. Mol. Life Sci.* **2013**, *70* (21), 4067–4084.

21. Hoadley, K. A.; Xue, Y.; Ling, C.; et al. Defining the Molecular Interface That Connects the Fanconi Anemia Protein FANCM to the Bloom Syndrome Dissolvasome. *Proc. Natl. Acad. Sci.* **2012**, *109* (12), 4437–4442.

22. Hoadley, K. A.; Xu, D.; Xue, Y.; et al. Structure and Cellular Roles of the RMI Core Complex from the Bloom Syndrome Dissolvasome. *Structure* **2010**, *18* (9), 1149–1158.

23. Lopper, M.; Boonsombat, R.; Sandler, S. J.; et al. A Hand-Off Mechanism for Primosome Assembly in Replication Restart. *Mol. Cell* **2007**, *26* (6), 781–793.

24. Zhang, J.-H. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J. Biomol. Screen.* **1999**, *4* (2), 67–73.

25. Navratilova, I.; Hopkins, A. L. Fragment Screening by Surface Plasmon Resonance. *ACS Med. Chem. Lett.* **2010**, *1* (1), 44–48.

Appendix 2 A high-throughput screening strategy to identify inhibitors of SSB protein-protein interactions in an academic screening facility

This work has been published:

Voter, A. F.*, Killoran, M. P.*, Ananiev, G. E., Wildman, S. A., Hoffman, F. M., Keck, J. L. (2018) A high-throughput screening strategy to identify inhibitors of SSB protein-protein interactions in an academic screening facility. *SLAS Discovery* **23**(1), 94-101. *These authors contributed equally

Andrew Voter miniaturized the high-throughput screening assays, performed the high-throughput screen, carried out data analysis and wrote the article. Michael Killoran designed and optimized the screening assays. Gene Ananiev, Scott Wildman and F. Michael Hoffman assisted with the screening and data analysis.

Abstract

Antibiotic-resistant bacterial infections are increasingly prevalent worldwide and there is an urgent need for novel classes of antibiotics capable of overcoming existing resistance mechanisms. One potential antibiotic target is the bacterial single-stranded DNA binding protein (SSB) which serves as a hub for DNA repair, recombination and replication. Eight highly conserved residues at the C-terminus of SSB use direct protein-protein interactions (PPIs) to recruit more than a dozen important genome maintenance proteins to single-stranded DNA. Mutations that disrupt PPIs with the C-terminal tail of SSB are lethal, suggesting that small molecule inhibitors of these critical SSB PPIs could be effective antibacterial agents. As a first step toward implementing this strategy, we have developed orthogonal high-throughput screening assays to identify small molecule inhibitors of the *Klebsiella pneumonia* SSB-PriA interaction. Hits were identified from an initial screen of 72,474 compounds using an AlphaScreen (AS) primary screen and their activity was subsequently confirmed in an orthogonal fluorescence polarization (FP) assay. As an additional control, an FP assay targeted against an unrelated eukaryotic PPI was used to confirm specificity for the SSB-PriA interaction. Nine potent and selective inhibitors produced concentration-response curves with IC_{50} values <40 μM and the direct binding of two compounds were observed to bind to PriA, demonstrating the success of this screen strategy.

Introduction

Antibiotic resistance among bacterial pathogens has become an increasing worldwide healthcare crisis. In the United States alone, an estimated 2 million people are infected with antibiotic-resistant bacteria annually, with over 23,000 attributed deaths each year.[1] The emergence and spread of bacterial strains with resistance to antibiotics of last resort, such as carbapenems[2] and colistin,[3] have raised fears of a post-antibiotic world in which routine infections are untreatable.[4,5] While improved prescribing practices and limitations in the use of antibiotics in livestock production are essential parts of efforts to combat antibiotic resistance,[6] novel classes of antibiotics that circumvent existing resistance mechanisms are also urgently needed.[5]

The molecular machinery essential for bacterial replication and genome maintenance is surprisingly dissimilar from analogous proteins in eukaryotes, so much so that it is thought that processes such as replication could have evolved independently.[7] From a therapeutic perspective, this difference makes genome maintenance proteins excellent antibiotic targets since the possibility of cross-reaction with functionally analogous eukaryotic proteins is minimized.[8,9] Because DNA replication, recombination and repair pathways are essential for bacterial cell viability, antibacterial drugs that target these processes have the potential to be highly effective. Fluoroquinolone topoisomerase inhibitors such as ciprofloxacin, levofloxacin, and trovafloxacin, which are commonly used to treat hospital-acquired pneumonia, urinary tract infections and other antibiotic-resistant infections,[10,11] comprise the only current antibiotics that act on bacterial genomic targets. These therapeutics inhibit type-II topoisomerases creating DNA breaks that block DNA replication fork progression[12] and are lethal unless repaired by homologous recombination and DNA replication restart processes that reload the DNA replication machinery. While the success of this class of antibiotics validates genome maintenance proteins as targets for antibacterial drug development, the therapeutic potential of numerous direct DNA replication, recombination and repair proteins remain untapped.[8,9]

In genome maintenance reactions, duplex DNA must often be separated to allow enzymes access to genomic information. However, single-stranded (ss) DNA is sensitive to chemical and nucleolytic attack and can form secondary structures that block genome maintenance reactions. To avoid these potential problems, bacteria have evolved specialized ssDNA binding proteins (SSBs) that bind and protect ssDNA as it is formed in cells. For the majority of bacteria examined to date, the functional ssDNA binding unit is a homotetrameric core of individual oligonucleotide-binding (OB) folds that are responsible for high-affinity ssDNA binding. Extending from the C-terminus of each of the monomeric constituents are intrinsically disordered linkers that terminate with an acidic and highly conserved 8 amino acid sequence called the SSBct. The SSBct interacts with a variety of critical genome maintenance proteins, positioning the SSB-DNA complex as a central hub for repair, replication and recombination. The structures of multiple SSB-interacting proteins (IPs) bound to peptides mimicking the SSBct have been determined and the SSB binding pocket of each of these proteins show striking electrostatic similarities.[13-18] Deletion or mutation of conserved residues within the SSBct disrupts its protein interactions and is lethal in *Escherichia coli*.[19-25] Given the essential nature of SSB protein interactions, we have hypothesized that small molecules capable of disrupting SSB protein interfaces could prove to be valuable antibiotic lead compounds. Previously identified inhibitors that block *E. coli* SSBct interaction with Exonuclease I support this hypothesis; however, their potential as antibiotics is limited by the fact that the activity of Exonuclease I is not essential to bacterial viability.[26,27]

As a step toward targeting SSB-protein interactions with molecules that could be used as antibacterial therapeutics, we have developed a high-throughput screening (HTS) platform to identify inhibitors that block interaction between the SSBct and the essential PriA DNA helicase from *Klebsiella pneumonia,* an important human pathogen.[28] The HTS strategy was designed with cost savings considerations as a major factor. These savings are derived from both the use of 1536-well plates to minimize reagent use and by using tip-free liquid handling. In a pilot screen of 72,474 compounds, 9

potent inhibitors that selectively disrupt the *Klebsiella* SSBct-PriA interaction were identified and 2 of these compounds were observed to bind PriA. Identification of these compounds validates this screening strategy and lays the foundation for the optimization and antibacterial activity of these inhibitors.

Materials and Methods

Protein expression

SSBct (Trp-Met-Asp-Phe-Asp-Asp-Asp-Ile-Pro-Phe), N-terminally biotinylated SSBct (Bio-SSBct), N-terminally fluorescein labeled SSBct (F-SSBct) and an SSBct lacking the C-terminal phenylalanine (ΔFSSBct) peptides were purchased from the University of Wisconsin Biotechnology center as lyophilized powders and resuspended in either dimethyl sulfoxide (DMSO) or 10 mM HEPES-HCl, pH 7.0.

Protein Purification

Rosetta 2 competent *E. coli* cells were transformed with a pET28b plasmid encoding the *Klebsiella pneumonia* PriA protein fused to an N-terminal 6xHis purification tag.[13] Single colonies were used to inoculate 100 mL of LB supplemented with kanamycin and chloramphenicol, and incubated with shacking overnight at 37 °C. The culture was used to inoculate 2 L of auto-induction media[29] supplemented with kanamycin and chloramphenicol, and allowed to grow at 37 °C with shaking for 24 hours. Cells were pelleted by centrifugation and stored at -80 °C. Cell pellets were thawed, mixed with lysis buffer (10 mM HEPES-HCl, pH 7.0, 300 mM Na_2SO_4, 100 mM glucose, 10% (v/v) glycerol, 20 mM imidazole, 1 mM β-mercaptoethanol, 1 mM phenylmethylsulfonyl fluoride and a protease inhibitor tablet (ThermoFisher, Waltham, MA)) and lysed by sonication. Insoluble components were removed by centrifugation and the supernatant was passed through a 0.22 µm filter prior to being loaded onto Ni-NTA agarose resin (Quiagen, Hilden, Germany). The column was washed with 15 column volumes (CV) of lysis buffer, then eluted with a gradient of lysis buffer containing 20 to 300 mM imidazole over 10 CV. Fractions containing PriA were pooled, diluted to 90 mM Na_2SO_4 with dilution buffer (10 mM HEPES-HCl, pH 7.0, 10 mM dithiotheritol (DTT), 10% (v/v) glycerol) and then loaded onto an SPFF column (GE Healthcare Bio-Sciences, Marlborough, MA) equilibrated with buffer A (10 mM HEPES, 100 mM Na_2SO_4, 10% glycerol, 1 mM DTT, pH 7.0). PriA was eluted from the SPFF column by gradient elution of buffer A containing 100 to 500 mM

Na_2SO_4. PriA-containing fractions were pooled, concentrated by centrifugation and then purified on a S-100 size exclusion column (GE Healthcare Bio-Sciences) equilibrated with buffer A containing 500 mM Na_2SO_4. Purified PriA was concentrated to approximately 150 µM, aliquoted and flash frozen in liquid nitrogen.

SSBct-PriA AlphaScreen assay

Screening took place at the University of Wisconsin Small Molecule Screening Facility (SMSF). Controls and small molecules (stored as 10 mM DMSO stocks) were dispensed into a 1536-well white plates (Nunc 253607, ThermoFisher) using an Echo 550 (Labcyte, Sunnyvale, CA) acoustic liquid handler. A Mantis liquid handler equipped with a high volume silicone chip (Formulatrix, Bedford, MA) was used to add 3.0 µL of a master mix containing 10 mM HEPES-HCl, pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$, 10 mM DTT, 1 mg/mL BSA, 0.01% Triton X-100, 0.1 µM PriA, 0.1 µM Bio-SSBct and 5 µg/mL AlphaScreen (AS) of both donor and acceptor beads (PerkinElmer, Waltham, MA) to each well of the plate. Master mix was prepared under diminished lighting immediately prior to dispensing. Plates were centrifuged briefly, rocked at room temperature for an hour and then read using a PheraStar (BMG Labtech, Offenburg, Germany) plate reader using the following settings: 0.1 s settling time, 0.3 s excitation, 0.6 s integration time with a 0.04 s delay between excitation and integration. Final concentrations of the positive (SSBct) and negative (ΔFSSBct) controls were 25 µM and all compounds were tested at 33.3 µM final concentration. Each screening plate contained 32 positive and negative control wells, 43 DMSO-only control wells and a control SSBct concentration-response curve conducted in triplicate with SSBct concentrations of 0.5, 1, 2, 4, 8, 16 and 32 µM. For compound concentration-response curves, the requisite amount of each compound (at 10 mM in DMSO) was added to the wells and then backfilled with 100% DMSO so that each well contained a final DMSO concentration of 0.33% (v/v). AS master mix was then added to each well as before.

Library Composition

The compound library comprised 72,474 unique small molecules originally purchased from Life Chemicals Inc. as part of their pre-plated diversity collection. The compounds were selected for diversity by the vendor and cover a large number of distinct scaffolds.

Data analysis

To reduce the effect of plate and edge effects, two 1536-well plates containing AS master mix and 0.33% DMSO in every well were read. The mean intensity of each well from the DMSO only plates was used to normalize the matching well of the assay plates. Additionally, a vertical signal gradient was noted across each plate, likely resulting from incomplete temperature equilibration inside the plate reader. To compensate for this drift without slowing the screening, a normalization process was implemented.

First, the mean value of all samples in a given row, *row i*, was calculated $(\overline{x}_{row\ i})$. The mean of all row averages was then determined $(\overline{x}_{all\ rows})$ and a ratio of each of the individual row means to overall row mean was calculated. The reading of each well in row *i* was then multiplied by this row-specific ratio to calculate the corrected reading $\left(x_{c,i}\right)$ as in equation A2.1.

$$x_{c,i} = x\left(\frac{\overline{x}_{row\ i}}{\overline{x}_{all\ rows}}\right) \qquad \text{(A2.1)}$$

After the normalization process, Z' scores for each plate were calculated and three plates with Z' values lower than 0.5 were repeated.[30] Small molecules producing more than a 35% inhibition in the SSBct-PriA AS were called hits. All hits were screened against Baell's PAINS filters to remove small molecules containing motifs suggesting PAINS activity[31] using the Drugs3 web service (fafdrugs3.mti.univ-paris-diderot.fr).[32]

Analysis of concentration-response curves was performed in Prism version 5.0c (GraphPad Software, La Jolla, CA) using a four-parameter logistic fit to determine IC_{50} values and errors.

Fluorescence polarization secondary and counter screens

To confirm the activity of compounds identified in the AS primary screen, hits were retested in duplicate at 33.3 µM in a fluorescence polarization (FP) assay. Compounds were added to black 384 shallow-well plates (ThermoFisher, 35 nL of a 10 mM DMSO stock) by an Echo acoustic liquid handler. FP master mix containing 10 mM HEPES-HCl, pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$, 1 mM DTT, 4.85 µM PriA and 0.01 µM F-SSBct was prepared, and 10 µL of master mix was added with the Mantis liquid handler using a high volume silicone chip. After incubating for at least an hour at room temperature, plates were read using a PheraStar plate reader with the following setting: Excitation: 485 nm, Emission: 520 nm, settling time of 0.2 s and 200 flashes per well. Small molecules with an average inhibition of greater than 35% were tested in the FP counter screen, which was performed as previously described.[33] Any compounds with an activity of > 25% inhibition in the RMI-MM2 FP assay were deemed promiscuous and not pursued.

Differential Scanning Fluorimetry

Differential scanning fluorimetry (DSF) was used to determine if any of the remaining compounds directly interact with PriA. In a 96-well, 0.2 µL thin-walled plate (Midsci, St. Louis, MO),10 µM PriA was incubated in 50 µL of DSF buffer containing 10 mM HEPES-HCl, 7.5, 150 mM NaCl, 1 mM $MgCl_2$, 1 mM DTT, 2% DMSO (v/v) and 5X SYPRO orange (Sigma-Aldrich, St. Louis, MO). Compounds were tested at concentrations of 20 µM and 100 µM. In positive control wells, SSBct was included at a concentration of 20 µM. Using a CFX Connect real time PCR system (Bio-Rad, Hercules, CA), reactions were held at 25 °C for 10 minutes and then ramped to 90 °C at a rate of 0.5 °C/min. Fluorescence readings were taking using the FAM channel every 0.5 °C. For melting point (T_m) determinations, readings from a well containing no PriA were subtracted from each time point. To aid in analysis, the rate of change in fluorescence signal was calculated at every temperature and the T_m was determined to be the temperature with the maximum rate of change.

Results

Assay optimization in 1536-well plates

To identify inhibitors of SSB PPIs, we developed a high-throughput AS assay to quantify the interaction between SSBct and PriA helicase from *Klebsiella pneumonia* in a low volume,1536-well plate format. In this bead-based proximity assay, Ni^{2+}-coated acceptor beads are tethered to His-tagged PriA, while streptavidin-coated donor beads bind to Bio-SSBct peptides. If SSBct and PriA interact with one another, acceptor and donor beads will be held in close proximity and singlet oxygen released from the donor bead upon excitation with 680 nm light will diffuse to the acceptor bead and trigger the emission of a 570 nm wavelength signal. Disruption of the interaction greatly reduces the amount of the short-lived singlet oxygen that can reach acceptor beads, resulting in minimal 570 nm background emission (Figure A2.1). To measure the reproducibility of the SSBct-PriA AS assay and its suitability for use in high-throughput screening, the assay was tested in a 3-day validation study. Identical reactions (10 µL) containing His-tagged PriA and Bio-SSBct were challenged with unlabeled SSBct or ΔFSSBct competitor peptide in 384-well plates on three consecutive days. We observed robust separation between the positive (unlabeled SSBct competitor peptide) and negative (unlabeled ΔFSSBct peptide) control reactions across each day of the trial, resulting in a Z' score of 0.81 and confirming its suitability for HTS. We have successfully used this assay as a counter screen in a previously published HTS.[33]

Given the difficulties associated with targeting PPIs, we anticipated a need to screen a large chemical library to successfully identify lead compounds. As protein production and AS beads are major contributors to the overall cost of the screen, we sought to minimize the assay volume without compromising assay performance. It was found that we could reduce our existing AS assay volume to 3 µL in a 1536-well format with only minimal losses in performance.

Pilot high-throughput screening campaign

Having adapted an assay for use in high-throughput format, we screened a library of 72,474 unique compounds with our primary AS assay to identify inhibitors of the SSBct-PriA interaction (Figure A2.2). A mean Z' score of 0.65 ± 0.08 was observed across 56 1536-well plates (Figure A2.3A). To allow for comparisons of potency across multiple plates, a concentration-response curve using unlabeled SSBct competitor peptide was included on each plate. Excellent consistency in the potency of the concentration-response curve was observed across plates and the IC_{50} determined from the concentration-response curve (3.0 ± 0.2 μM) was found to be in agreement with previously published values (2.4 ± 1.3 μM) of the *E. coli* SSBct-PriA interactions observed by surface plasmon resonance under similar salt concentrations (Figure A2.3B).[34] In total, the mean inhibition from members of the compound library was -2.8% and normally distributed with a standard deviation of 13.5% (Figure A2.4). With a threshold of > 35% inhibition, corresponding to approximately 3 standard deviations from the mean, 946 small molecules were found to be active in the AS for a primary hit rate of 1.3%.

Elimination of false positive hits.

Others have noted that a subset of small molecules present in screening libraries have activity against a broad range of diverse assay targets, particularly in AS assays. We used an *in silico* filter to identify and remove 276 PAINS and other compounds with suspect chemical structures from our initial list of 946 hits.[31,32]

In order to further cull false positives not detected by the *in silico* PAINS filter from our initial AS HTS, we adapted an FP assay monitoring the SSBct-PriA interaction for use in 384-well plate format as an orthogonal secondary screen.[20] In this assay, PriA is mixed with an SSBct peptide labelled with an N-terminal fluorescein (F-SSBct). The proportion of F-SSBct bound to PriA is determined by exciting the F-SSBct with polarized light and measuring the polarization of emitted fluorescence. Free F-SSBct tumbles rapidly before emission, resulting in a loss of polarization whereas the higher molecular weight PriA-F-

SSBct complex tumbles more slowly and a greater proportion of the emitted light remains polarized (Figure A2.5).

To confirm the hits selected by our primary AS assay were specific for the SSBct-PriA interaction, each of the remaining 670 compounds were tested in duplicate using the orthogonal SSBct-PriA FP assay. We expected compounds interfering with the AS chemistry rather than blocking SSBct would be inactive in the FP assay. Greater than 35% inhibition in the FP assay was observed for 35 hits (5.2%). To exclude any remaining non-specific inhibitors, each compound was tested for activity in an FP assay targeting an unrelated eukaryotic protein-protein interaction.[33] Twenty-two of the 35 small molecules had <25% inhibitory activity at a concentration of 33.3 μM in this counter screen and were deemed specific for the SSBct-PriA interaction. Concentration-response curves were obtained using the AS assay and 9 potent and selective lead compounds were identified with IC_{50} values less than 40 μM.

To assess if these compounds bind PriA, DSF was used to measure the T_m of PriA in the presence or absence of each of the 9 lead compounds. Small molecules are frequently observed to stabilize interacting proteins, thus an increase in the T_m of PriA indicates compound binding.[35] PriA was thermally denatured in the presence of a dye which fluoresces only when bound to hydrophobic patches of PriA that are exposed as the protein unfolds. Concentration-dependent stabilization of PriA was observed for 2 compounds (Figure A2.6).

Discussion

Here we describe a methodology for a low-volume AS HTS performed in 1536-well plates. While the manufacturer recommended 25 µL AS reaction volume provides for 10,000 data points per 5 mg of beads (at 20 µg/mL beads), our further miniaturization of the assay to a 3 µL final volume and reduction of bead concentration to 5 µg/mL allows for a 41-fold decrease in bead use. This allowed us to assay ~300,000 data points with an equivalent amount of reagents. Similarly, protein usage was reduced to less than 5 mg for the entire screen as compared to ~1 g estimated to be required for a comparable FP screen. Liquid handling via the Echo and Mantis instruments is both rapid (6 min per 1536-well plate ECHO and 13 min per plate Mantis) and independent of consumables. We found that the vacuum chips on the Mantis needed to be occasionally replaced, but this is only trivially added to the total cost. Of special note is the exceptionally low dead volume of AS master mix that is achieved with our liquid handling approach. On average the dead volume was only 30 µL per 1536-well plate.

During assay development we noted plate effects in the AS assay plates in the form of a signal loss gradient vertically down the plate and a significantly higher AS signal in the top row of each plate. Small temperature changes have been noted to alter AS signal and likely explain the signal gradient.[36] However, we have consistently observed the top row phenomena across different incubation times, assay plates, assay targets and plate readers. While we have been unable to eliminate this effect, its impact can be mitigated by the normalization process described herein (Figure A2.7).

We observed a relatively low hit confirmation rate between the primary AS and secondary FP assay. The low confirmation rate likely resulted from using a relatively permissive hit threshold to determine which compounds were selected from the primary AS assay. This threshold was chosen to avoid prematurely eliminating inhibitors with more modest potency which could serve as lead compounds for the future development of more potent derivatives. One major limitation of such an approach is that the

majority of these putative weak inhibitors are inactive and will need to be eliminated in downstream screening steps. However, given the historical difficulties in developing PPI inhibitors, we reasoned that the chance of identifying a *bona fide*, albeit weak, inhibitor would outweigh the additional screening burden.

Our HTS platform identified 9 selective inhibitors of the SSBct-PriA interaction with IC_{50} values < 40 μM. Of these 9 inhibitors, 2 were observed to interact with PriA by thermal stabilization in a DSF assay. While no stabilization was detected for 7 compounds, this does not exclude PriA binding or inhibition. Binding by non-stabilizing ligands or ligands that stabilize both folded and unfolded PriA to a similar extent will not be detected in this assay.[37]

Further biophysical and structural studies to characterize the interaction between these inhibitors and PriA are underway. These inhibitors are also being tested for activity against multiple SSB interacting partners to determine if they are active against other targets important for bacterial replication and genome maintenance. Compounds discriminating between specific SSB PPIs could serve as valuable chemical probes to dissect the role of individual interactions while those inhibiting SSB PPIs indiscriminately may represent a novel class of antibiotics capable of targeting pathogenic bacteria resistant to currently available antibiotics.

Acknowledgements

The authors thank Song Guo for assisting with the screening and members of the Keck lab for critical review of the manuscript.

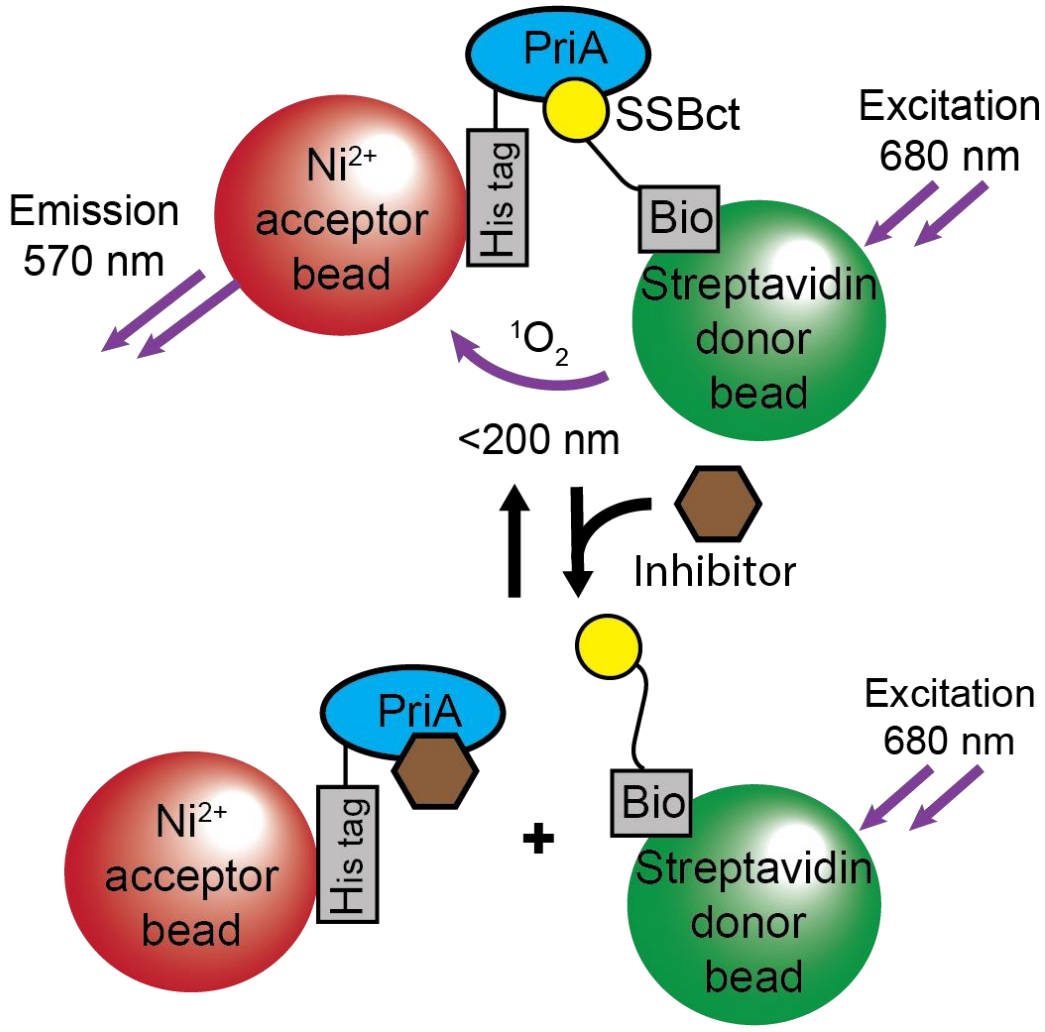


Figure A2.1 Schematic representation of the AS assay to identify inhibitors of the SSBct-PriA interaction.

SSBct and PriA are bound to donor and acceptor bead pairs that generate signal in close proximity. Disruption of the SSBct-PriA interaction by an inhibitor is detected by a loss of signal emission.

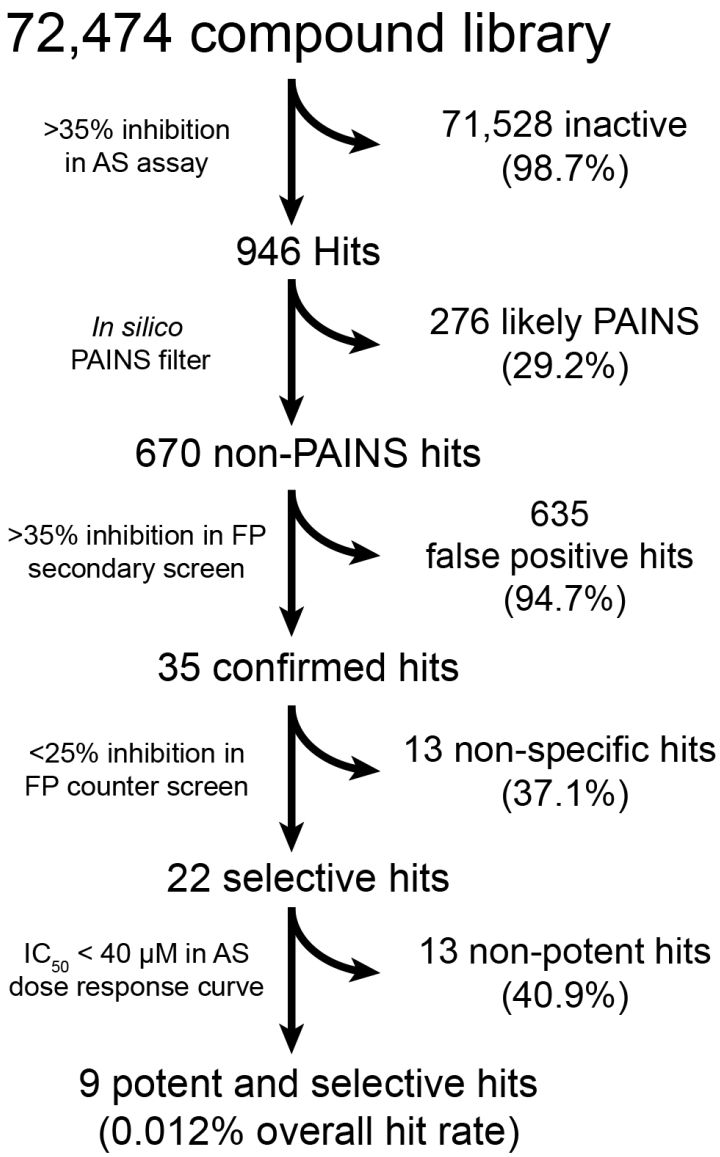


Figure A2.2. Screening and hit reduction workflow. The selection criteria and number of compounds excluded are noted for each step.

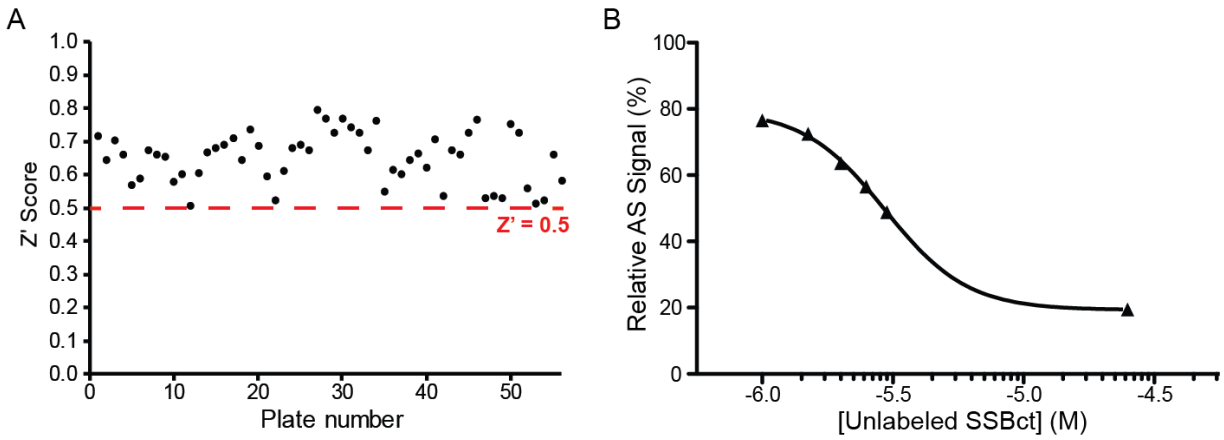


Figure A2.3 The SSBct-PriA AS assay performs well under high-throughput conditions. (A). Scatterplot of the Z' scores observed for the each of the 56 plates used in the primary screen. The dashed red line indicates the minimum acceptable Z' score of 0.5. Three plates were found to have Z' scores below 0.5 and were repeated prior to data analysis. (B) Each plate contained a titration of unlabeled SSBct peptide in triplicate, the mean AS signal relative to the negative control wells on each plate is shown. High reproducibility of this concentration-response curve was observed between the plates. Error bars representing the SEM of all of the replicates at a given concentration of SSB are obscured behind data points.

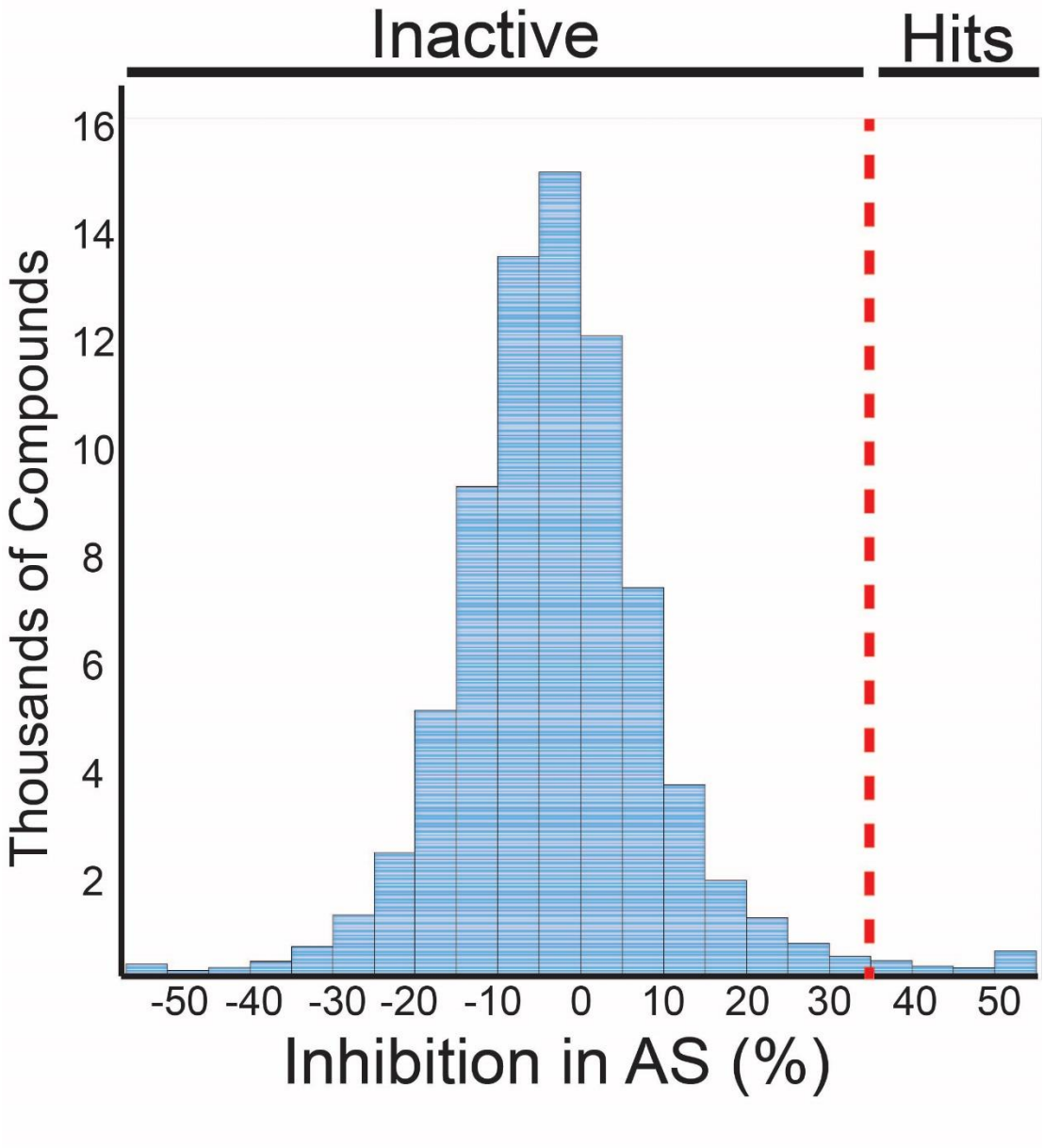


Figure A2.4. Histogram showing the activity of the compounds tested in the SSBct-PriA primary AS assay. The dashed line represents 35% inhibition in the assay. Compounds with more than 35% activity were advanced for further characterization.

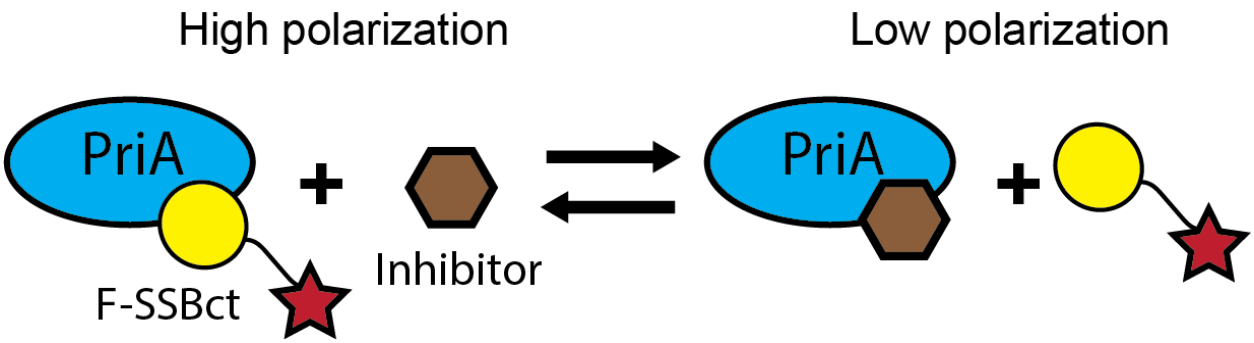


Figure A2.5. Scheme depicting the SSBct-PriA fluorescence polarization assay. Displacement of the F-SSBct from PriA by an inhibitor is detected as a decrease in polarization of light emitted from F-SSBct.

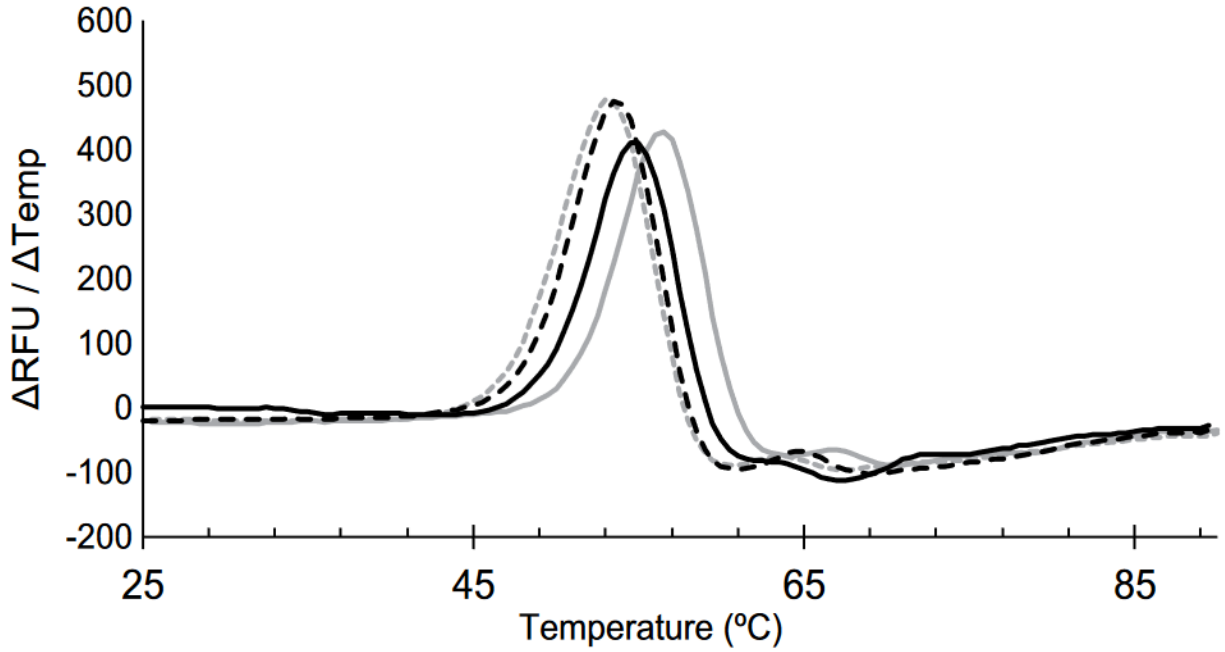


Figure A2.6. Small molecule stabilization of PriA during thermal denaturation indicates direct binding. PriA was heated in the absence of ligand (gray dashed line), or in the presence of SSBct (solid gray line), 20 µM inhibitor (dashed black line) or 100 µM inhibitor (solid black line). The T_m for each condition is the temperature corresponding to the maximum change in fluorescence.

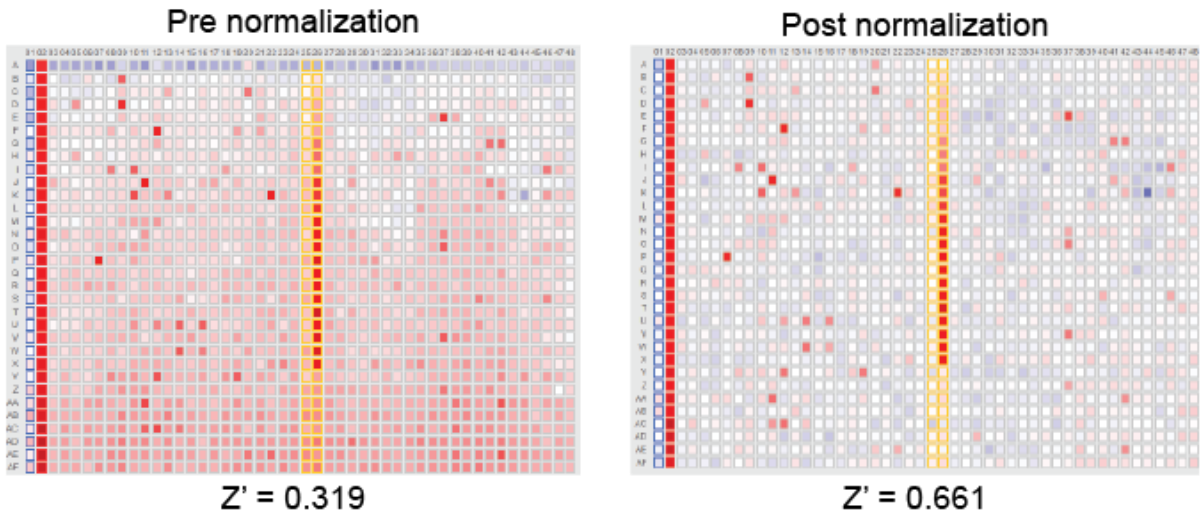


Figure A2.7 Heat map of a representative 1536-well plate before and after the normalization process. Wells outlines in blue (column 01) and red (column 02) boxes contain negative and positive controls peptides respectively, while the wells outlined in yellow contain either DMSO alone (column 25 and bottom of column 26) or a titration of the SSBct positive control peptide (column 26, top). Before normalization (left), a markedly increased signal was noted in the top row of each plate, as well as a vertical signal gradient down the plate. Both problems are resolved after normalization (right), without altering the distribution of hits.

References

1. Antibiotic Resistance Threats in the United States, 2013. (CDC).
2. Yong, D.; Toleman, M. A.; Giske, C. G.; et al. Characterization of a new metallo- -lactamase gene, blaNDM-1, and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India. Antimicrob. Agents Chemother. 2009, 53, 5046-5054.
3. Liu, Y.-Y.; Wang, Y.; Walsh, T. R.; et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect. Dis. 2016, 16, 161-168.
4. Nordmann, P.; Poirel, L.; Toleman, M. A.; et al. Does broad-spectrum -lactam resistance due to NDM-1 herald the end of the antibiotic era for treatment of infections caused by Gram-negative bacteria? J. Antimicrob. Chemother. 2011, 66, 689-692.
5. Carlet, J.; Jarlier, V.; Harbarth, S.; et al. Ready for a world without antibiotics? The pensières antibiotic resistance call to action. Antimicrob. Resist. Infect. Control 2012, 1, 1.
6. Laxminarayan, R.; Duse, A.; Wattal, C.; et al. Antibiotic resistance—the need for global solutions. Lancet Infect. Dis. 2013, 13, 1057-1098.
7. Leipe, D. D.; Aravind, L.; Koonin, E. V. Did DNA replication evolve twice independently? Nucleic Acids Res. 1999, 27, 3390-3401.
8. Shereda, R. D.; Kozlov, A. G.; Lohman, T. M.; et al. SSB as an organizer/mobilizer of genome maintenance complexes. Crit. Rev. Biochem. Mol. Biol. 2008, 43, 289-318.
9. Robinson, A. J.; Causer, R. J.; Dixon, N. E. Architecture and conservation of the bacterial DNA replication machinery, an underexploited drug target. Curr. Drug Targets 2012, 13, 352–372.
10. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated Pneumonia. Am. J. Respir. Crit. Care Med. 2005, 171, 388–416.

11. Liu, H.; Mulholland, S. G. Appropriate antibiotic treatment of genitourinary infections in hospitalized patients. Am. J. Med. 2005, 118, 14–20.
12. Drlica, K.; Zhao, X. DNA gyrase, topoisomerase IV, and the 4-quinolones. Microbiol. Mol. Biol. Rev. 1997, 61, 377-392.
13. Bhattacharyya, B.; George, N. P.; Thurmes, T.M.; et al. Structural mechanisms of PriA-mediated DNA replication restart. Proc. Natl. Acad. Sci. 2014, 111, 1373-1378.
14. Petzold, C.; Marceau, A. H.; Miller, K. H.; et al. Interaction with single-stranded DNA-binding protein stimulates Escherichia coli ribonuclease HI enzymatic activity. J. Biol. Chem. 2015, 290, 14626–14636.
15. Lu, D.; Keck, J. L. Structural basis of Escherichia coli single-stranded DNA-binding protein stimulation of exonuclease I. Proc. Natl. Acad. Sci. 2008, 105, 9169–9174.
16. Marceau, A. H.; Bahng, S.; Massoni, S. C.; et al. Structure of the SSB-DNA polymerase III interface and its role in DNA replication. EMBO J. 2011, 30, 4236–4247.
17. Ryzhikov, M.; Koroleva, O.; Postnov, D.; et al. Mechanism of RecO recruitment to DNA by single-stranded DNA binding protein. Nucleic Acids Res. 2011, 39, 6305-6314.
18. Cheng, K.; Xu, H.; Chen, X.; et al. Structural basis for DNA 5'-end resection by RecJ. eLife, 2016, 5, e14294.
19. Curth, U.; Genschel, J.; Urbanke, C.; et al. In vitro and in vivo function of the C-terminus of Escherichia coli single-stranded DNA binding protein. Nucleic Acids Res. 1996, 24, 2706–2711.
20. Lu, D.; Windsor, M. A.; Gellman, S. H.; et al. Peptide inhibitors identify roles for SSB C-terminal residues in SSB/exonuclease I complex formation. Biochemistry (Mosc.) 2009, 48, 6764–6771.
21. Kelman, Z.; Yuzhakov, A.; Andjelkvoic, J.; et al. Devoted to the lagging Strand - the χ subunit of DNA polymerase III holoenzyme contacts SSB to promote processive elongation and sliding clamp assembly. EMBO J. 1998, 17, 2436-2449.

22. Meyer, R. R.; Laine, P. S. The single-stranded DNA-binding protein of Escherichia coli. Microbio. Rev. 1990, 54, 342-380.
23. Greenberg, J.; Berends, L. J.; Donch, J.; et al. exrB: a malB-linked gene in Escherichia coli B involved in sensitivity to radiation and filament formation. Gen. Res. 1974, 23, 175-184.
24. Genschel, J.; Curth, U.; Urbanke, C. Interaction of E. coli single-stranded DNA binding protein (SSB) with exonuclease I. The carboxy-terminus of SSB is the recognition site for the nuclease. Biol. Chem. 2000. 381, 183-192.
25. Wang, T. V.; Smith, K. C.; Effects of the ssb-1 and ssb-113 mutations on survival and DNA repair in UV-irradiated ΔuvrB strains of Escherichia coli K-12. J. Bact. 1982, 151, 186-192.
26. Lu, D.; Bernstein, D. A.; Satyshur, K. A.; et al. Small-molecule tools for dissecting the roles of SSB/protein interactions in genome maintenance. Proc. Natl. Acad. Sci. 2010, 107, 633–638.
27. Marceau, A. H.; Bernstein, D. A.; Walsh, B. W.; et al. Protein interactions in genome maintenance as novel antibacterial targets. PLoS ONE 2013, 8, e58765.
28. Nordmann, P.; Cuzon, G.; Thierry, N. The real threat of Klebsiella pneumoniae carbapenemase producing bacteria. Lancet Infect. Dis. 2009, 9, 228–236.
29. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. Protein Expr. Purif. 2005, 41, 207–234.
30. Zhang, J.; Chung, T. D. Y.; Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. J. Biomol. Screen. 1999, 4, 67-73.
31. Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J. Med. Chem. 2010, 53, 2719–2740.
32. Lagorce, D.; Sperandio, O.; Baell, J. B.; et al. FAF-Drugs3: A web server for compound property calculation and chemical library design. Nucleic Acids Res. 2015, 43 (W1), W200–W207.

33. Voter, A. F.; Manthei, K. A.; Keck, J. L. A high-throughput screening strategy to identify protein-protein interaction inhibitors that block the Fanconi snemia DNA repair pathway. J. Biomol. Screen. 2016, 21, 626-633.
34. Cadman, C. J.; McGlynn, P. PriA helicase and SSB interact physically and functionally. Nucleic Acids Res. 2004, 32, 6378-6387.
35. Niesen, F. H.; Berglund, H.; Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nature Protocols. 2007, 2, 2212-2221.
36. A Practical Guide to working with AlphaScreen. 2004. PerkinElmer. https://www.urmc.rochester.edu/MediaLibraries/URMCMedia/hts/documents/AlphaScreenPracticalGuide.pdf (accessed Feb 21, 2017).
37. Simeonov, A. Recent Developments in the Use of Differential Scanning Fluorometry in Protein and Small Molecule Discovery and Characterization. Expert Opin. Drug Discov. 2013, 8, 1071-1082.

Appendix 3

Development of Protein–Protein Interaction Inhibitors for the Treatment of Infectious Diseases

This work has been published:

Voter, A. F.; Keck, J. L. “Development of Protein-Protein Interaction Inhibitors for the Treatment of Infectious Diseases” In *Advances in Protein Chemistry and Structural Biology*; Academic Press: Cambridge, MA, 2018; Vol. 111, pp 197–222.

This chapter was written by Andrew Voter.

Abstract

Protein–protein interaction (PPI) inhibitors are a rapidly expanding class of therapeutics. Recent advances in our understanding of PPIs and success of early examples of PPI inhibitors demonstrate the feasibility of targeting PPIs. This review summarizes the techniques used for the discovery and optimization of a diverse set PPI inhibitors, focusing on the development of PPI inhibitors as new antibacterial and antiviral agents. We close with a summary of the advances responsible for making PPI inhibitors realistic targets for therapeutic intervention and brief outlook of the field.

Introduction

The increasing prevalence of treatment-resistant pathogens is one of the biggest challenges facing the biomedical community today. Antiinfective agents have revolutionized medicine, allowing for curative therapies that vastly reduced the morbidity and mortality of diseases caused by pathogenic bacteria, viruses, and parasites. Unfortunately, the widespread overuse and misuse of these agents has led to increasing levels of resistance that threatens our ability to effectively treat not only infections but also the use of therapies that requires prophylactic antibiotic treatment, such as surgeries. While better stewardship is critical for extending the usable life of the existing repertoire of antibacterial and antivirals agents (Goff et al., 2017), classes of agents with novel mechanisms of action are urgently needed ("The 10 x '20 Initiative," 2010).

Traditionally, antibiotic development has focused primarily on inhibitors of essential bacterial enzymes for understandable reasons (Walsh, 2016). The mechanism of action was easily understood, biochemical assays existed to test for activity, and substrate analog inhibitors were relatively easy to produce. Unfortunately, the supply of easily inhibited targets appears to be limited as the output of new antiinfective agents has dwindled (Ventola, 2015). Deepening our understanding of the underlying biology of pathogens will undoubtedly uncover new potential enzymatic targets, but the rate of discovery has not kept pace with the development of resistance (Ventola, 2015).

In the face of a dry antiinfective pipeline, researchers and pharmaceutic companies have begun to turn their attention to a different class of targets, protein–protein interactions (PPI) (Arkin and Wells, 2004, Wells and McClendon, 2007). Many essential cellular functions rely on the precise and timely recruitment of proteins, often accomplished through a PPI. Disruption of protein interfaces, either through genetic proof-of-principle studies or small-molecule inhibitors, can kill pathogens or render them nonvirulent, making PPI inhibitors an exciting new research area in the antiinfective world. There are also

plenty of PPIs with which to work. The number of cataloged PPIs varies by database, but at a minimum there are tens of thousands human PPIs and the *Escherichia coli* interactions number in the thousands (Lehne & Schlitt, 2009). These counts do not include a sizeable number of host–pathogen interactions, with over 50,000 human–pathogen interactions cataloged in HPIDB (Kumar & Nanduri, 2010). While not all of these interactions are feasible targets for inhibition, a sizeable number are. We will explore examples of inhibitors that target several classes of PPIs: pathogen–pathogen, host–pathogen, and host–host interactions and how they might alter the treatment of infectious diseases.

Historically, PPIs were considered undruggable targets. This reputation likely stemmed from the lack of high-throughput ready screening assays as well as the thought that most PPIs are held together by large, chemically noncomplex surfaces with a lack of easily druggable pockets (Spencer, 1998). While such difficult PPI targets undoubtedly exist, it is now appreciated that many PPIs use much smaller interfaces for their interaction, frequently consisting of an unstructured peptide bound to a well-defined groove (Arkin, Tang, & Wells, 2014). Furthermore, mutagenesis studies of several PPIs have revealed that surfaces contributing to the affinity of a given PPI are not evenly distributed across the entire interface. Rather, there tends to be a “hot spot” or a small number of critical residues that anchor two proteins together (Cukuroglu, Engin, Gursoy, & Keskin, 2014). This means that a putative inhibitor would not need to displace the entirety of a given PPI, but rather only occupy the hot spot, a more tractable problem. Recent review articles have highlighted small molecules disrupting PPIs for the treatment of oncologic targets that have reached early clinical trials, demonstrating the feasibility of the approach. Because many of these inhibitors have already been reviewed in depth (Arkin et al., 2014, Sheng et al., 2015), this review will focus on PPI inhibitors for the treatment of infectious diseases.

A3.1. Antibacterial Agents

A3.1.1. ZipA-FtsZ

During bacterial cytokinesis, the cell contents must be properly partitioned between the two daughter cells and the cell wall sealed to prevent loss of cytoplasmic material or cell lysis. To accomplish this task, a ring, called the *Z*-ring, is formed at the site of division from the head-to-tail polymerization of the GTPase FtsZ (Adams & Errington, 2009). While the contribution that FtsZ and the *Z*-ring play in generating the force required to pinch the cell membrane is debated, it is clear that FtsZ plays an essential role in cytokinesis (Xiao & Goley, 2016).

To maintain contact with the cell wall throughout cytokinesis, FtsZ uses the 17 C-terminal most residues to bind to the membrane-associated protein ZipA (Mosyak et al., 2000). Loss of this interaction is lethal in the gammaproteobacteria (although it is absent in other bacteria; Hale & de Boer, 1997) likely due to the ability of ZipA to stabilize FtsZ polymers and localize them to the membrane (Kuchibhatla, Bhattacharya, & Panda, 2011). Additionally, alanine-scanning mutations of the FtsZ interaction site demonstrated that the majority of the affinity between the two protein is derived from only three hydrophobic residues, I374, F377, and L378 (Mosyak et al., 2000). Together these data suggest that a small molecule could block the FtsZ–ZipA interaction and that an inhibitor of this PPI would have antibacterial properties.

Researchers at Wyeth Research developed a high-throughput fluorescence polarization (FP) assay to screen for inhibitors of the FtsZ–ZipA interaction. During assay development, they realized that the relatively poor affinity of the PPI (7 μM K_D as determined by surface plasmon resonance) meant that a prohibitively large amount of ZipA would be required to screen an acceptable number of compounds. To circumvent this limitation, a phage display screen was conducted to identify a probe with a higher affinity to the ZipA. The resulting peptide, FtsZ-PD1, was found to have a K_D of 150 nM, a 45-fold improvement

and a FP high-throughput screen (HTS) of 250,000 compounds was conducted using a labeled version of the FtsZ-PD1 as a probe. This screening identified a pyridylpyrimidine inhibitor with a modest 12 μM K_i in the FP assay (Fig. A3.1; Kenny et al., 2003), and several additional inhibitor scaffolds with weak activities were identified in the same screen. Crystallographic studies confirmed that the inhibitor occupied the FtsZ binding pocket on ZipA.

Besides reducing the protein production burden, one can imagine two possible results of using a tighter binding probe for screening. First, the higher affinity peptide may serve to exclude low potency, but still active, inhibitor scaffolds that could be improved through medical chemistry efforts. The remaining hits are more likely to be active and potent against the native PPI, although reduced in number. The trade-offs associated with this approach likely depend on the size and quality of the chemical library to be screened. With a large enough library, the reduced hit rate is unlikely to be problematic and may even reduce the burden of secondary screening and hit validation. Presumably, the opposite approach could also be used; where a targeted interaction may be weakened to gain a foothold for later optimization. Second, the use of a higher affinity probe will introduce interaction sites not used in the wild-type interface. Ideally, a screening hit will target only the residues comprising the original interface and not the artificially introduced interactions.

Unfortunately, the pyridylpyrimidine scaffold was found to have nonspecific toxicity against both bacterial and yeast cells lines which precluded further development (Rush, Grant, Mosyak, & Nicholls, 2005). A variety of techniques were tried in attempts to develop more promising inhibitor scaffolds. First, a computational scaffold hopping method was used to identify a triazolopyridazine ring structures that mimicked the pyridylpyrimidine binding pose. While a crystal structure of a triazolopyridazine inhibitor bound to ZipA was obtained, these inhibitors were quite weak (IC_{50} ~ 80 μM) and do not appear to have been pursued (Rush et al., 2005). Second, an additional set of weak scaffolds (IC_{50} > 1 mM) were discovered in the initial FP screen. Remarkably in the face of this low affinity, crystal structures of these

inhibitors bound to the FtsZ pocket of ZipA were obtained (Jennings et al., 2004). This led to additional medicinal chemistry optimization, as well as an attempt to merge two weak inhibitors to fill the entirety of the binding pocket (Sutherland et al., 2003). While a 10-fold reduction in IC_{50} was obtained, none of these compounds displayed antibacterial activity against *E. coli*. These compounds did inhibit the growth of *E. coli* strains with compromised cell membranes, leading the authors to blame poor cell penetrance for the lack of bactericidal effect. However, a subset of these compounds was found to be active against several gram-positive organisms, which lack a ZipA homolog. This suggests that observed the antibacterial activity may have been a result of off-target effects (Sutherland et al., 2003). Finally, an NMR fragment screen was conducted to identify a more cell-penetrant inhibitor scaffold. While several new inhibitor scaffolds were identified, none possessed adequate activity in the FP assay to be pursued farther and the project appears to have been terminated (Tsao et al., 2006).

Despite the difficulties encountered in cell penetrance and nonspecific toxicity, this early effort was a successful demonstration of the ability to identify PPI disrupting inhibitor scaffolds. We are not aware of additional work against this target, although the combination of advances in screening libraries, techniques, and the ease of obtaining inhibitor-bound structures suggests that potent inhibitors could yet be discovered.

A3.1.2. Pilicides

Urinary tract infections (UTI) are one of the most prevalent infections and are responsible for nearly 10 million clinic visits per year (Schappert, 2011). Treatment of UTIs is estimated to exceed $3,500,000,000 annually (Flores-Mireles, Walker, Caparon, & Hultgren, 2015). UTIs can be caused by a wide range of pathogens, but most UTIs (65%–75%) are caused by uropathogenic *E. coli* (UPEC) (Flores-Mireles et al., 2015). Most uncomplicated UTIs will resolve spontaneously within a week; however, symptoms associated with UTIs are unpleasant (dysuria, frequency, urgency, suprapubic pain, and

hematuria; Bent, Nallamothu, Simel, Fihn, & Saint, 2002). Patients are typically treated with nitrofurantoin, trimethoprim–sulfamethoxazole, or fosfomycin (Gupta et al., 2011). While resistance to these commonly used antibiotics is relatively rare, resistance rates are increasing (Zowawi et al., 2015). Furthermore, treatment with broad spectrum antibiotics can lead to profound disruptions in the native microbiota with poorly understood consequences (Dethlefsen, Huse, Sogin, & Relman, 2008). Both of these problems could be circumvented with novel therapies targeting the virulence factors of UPEC specifically, sparing the nonpathogenic members of the microbiota.

To maintain an infection in the urinary tract despite the repeated outward flow of urine, UPEC anchor to the urothelium with the help of a variety of excreted pilus fibers (Wu, Sun, & Medina, 1996). One pilus, the type 1 pili, is capped with a FimH subunit that binds to mannose presented on urothelium glycolipids (Krogfelt, Bergmans, & Klemm, 1990). Type 1 pili also contributes toward biofilm formation and the accompanying resistance to therapy (Martinez et al., 2000, Wright et al., 2007). To ascend into the kidney and cause pyelonephritis, UPEC must express P pili, terminating with PapG, which binds to kidney specific glycolipids (Lane & Mobley, 2007).

Because bacteria remain physically anchored in the urinary tract during an infection, inhibitors of FimH binding are predicted to prophylactically block UPEC binding or allow for the washout of an existing infection. Indeed, recently reported small-molecule inhibitors of FimH binding, called mannocides, have activity as both prophylaxis or for treatment of established infections of murine models of UTI, demonstrating the utility of targeting UPEC pili (Cusumano et al., 2011, Han et al., 2010). While an exciting development for the treatment of UTIs, mannocide therapies disrupt only a specific pilus–ligand interaction, limiting the number of susceptible organisms.

Pilus formation and export to the bacteria cellular membrane occurs through the chaperone–usher pathway, which is widely conserved among gram-negative pathogens (Busch & Waksman, 2012).

Given the need for novel antibiotics targeting gram-negative pathogens, general inhibitors of the chaperone–usher pathway could find use as therapies for a range of pathogens such as *Salmonella*, *Yersinia*, and *Pseudomonas* species (Nuccio & Bäumler, 2007). The chaperone–usher pilus synthesis pathway begins with the transport of pilus subunits to the periplasmic space where they are bound by a chaperone protein, PapD, that is required for proper folding and pilin assembly. The pilin subunits contain a nearly complete β-barrel, lacking only a single strand so that donation of the missing strand by the chaperone protein allows for proper folding of the subunits. Additionally, the chaperone uses a hydrophobic groove to trap and display an unstructured N-terminal extension of the pilin subunit. To add a new subunit to the growing strand, the displayed N-terminus extension replaces the chaperone strand to complete the β-barrel and allow for the release of the chaperone (Waksman & Hultgren, 2009). Loss of binding to the chaperone blocks pilus assembly and leads to the accumulation of aggregated pilin subunits (Slonim, Pinkner, Branden, & Hultgren, 1992), suggesting that small-molecule inhibitors of the usher–chaperone pathway could have the same effect.

To test this hypothesis, a series of small molecules based on the core of the chaperone ligand, PapG, were synthesized. These so-called pilicides were observed to bind to PapD by SPR and NMR (Hedenstrom et al., 2005, Svensson et al., 2001). Initial biological results were promising, as the lead pilicide compound (Fig. A3.2A) was found to block bacterial binding to bladder epithelial cells, reduce the number of cells with developed pili, and disrupt biofilm formation, albeit at relatively high inhibitor concentrations. To confirm the pilicide mechanism of action, the authors solved the structure of a pilicide bound PapD. The pilicide was found to obstruct the binding site for FimH, leading to the observed failure of pilus assembly (Pinkner et al., 2006). Further medicinal chemistry optimization of the inhibitors yielded a derivative carrying an additional benzyl functional group, with over a 16-fold improvement in potency in a biofilm formation assay (Chorell et al., 2010). In spite of these gains, even the most active pilicides to date are relatively low potency, (7 μM IC_{50}, Fig. A3.2B) and have not been tested in animal models of

urinary tract infections. Furthermore, recent studies have indicated that the advanced pilicides may have additional, nonpilicidal, effects on the transcriptional regulation of UPEC virulence and motility, although these experiments were conducted at high inhibitor concentrations, where off-targets effects are more likely (Greene et al., 2014).

Despite the remaining challenges, these studies demonstrate exciting progress toward novels UTI therapies that target UPEC adhesion, both by mannosides and the pilicides. Since both classes of inhibitors target bacterial virulence rather than survival, it is possible that resistance would be slower to develop and nonvirulent microbiota would be spared, both welcome attributes in a therapeutic agent.

A3.1.3. DnaN-DnaE1

The development of effective therapies against *Mycobacterium tuberculosis* was a major global achievement and has made it possible to imagine that this scourge may eventually be eradicated. While the number of tuberculosis-related deaths has fallen by more than 20% since 2000, it remains one of the top 10 global causes of death (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2016). Unfortunately, there has been a remarkable increase in the extent of drug resistance in *M. tuberculosis* with the emergence of extensively (Jassal & Bishai, 2008) and totally drug-resistant strains (Udwadia, Amale, Ajbani, & Rodrigues, 2012). In 2015, over 30% of tuberculosis cases were found to be resistant to the first-line therapy rifampicin (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2016). To continue to make strides against this pathogen, antibacterial agents acting against novel targets are needed.

To identify potential inhibitors of *M. tuberculosis*, Kling and coworkers identified a promising lead compound, griselimycin (GM), that was first identified at Rhône-Poulenc in the 1960s (Kling et al., 2015). GM is a cyclic peptide natural product produced from *Streptomyces* species (Fig. A3.3), but the compound suffered from poor pharmacokinetics and the introduction of other active antituberculosis agents. Hoping

to overcome these limitations, the authors undertook a preliminary structure activity relationship (SAR) study that demonstrate that minor derivatization of the Pro^{8} residue leads to both an increased stability and potency. High doses of a cyclohexyl-derivatized GM (CGM) were found to be as active as clinical therapy isoniazid against *M. tuberculosis* in vitro and in an in vivo mouse lung TB model.

Because the mechanism of GM was not established during the initial discovery process, resistance studies were conducted to identify the cellular target of GM. A variety of genetic approaches lead to the determination that DnaN, the DNA polymerase sliding clamp, was the target, and GM binding to DnaN was confirmed by surface plasmon resonance with femtomolar dissociation constants. Both GM and CGM bind to mycoplasma DnaN with over a ~ 1000 × higher affinity than to the *E. coli* homolog or the resistance-associated protein GriR. The binding site of a methylated GM was determined by X-ray crystallographic studies of DnaN, which revealed the compounds block binding to DnaE1, the catalytic subunit of Pol III. The authors propose that disruption of this interaction leads to lethal failures in DNA replication and likely DNA breaks.

Intriguingly, previous screening campaigns have targeted the DnaN–DnaE interface with only modest results. An FP-based chemical HTS identified a 10-µM inhibitor of the interaction, and while a crystal structure of the inhibitor bound to DnaN was obtained, no measurements of antibacterial efficacy were reported. This inhibitor contains a rhodanine, a common pan-assay interference compounds motif (Baell & Holloway, 2010), suggesting off-target effects would be problematic (Georgescu et al., 2008). A later fragment screening effort identified two possible inhibitor scaffolds (Yin et al., 2014), and while further SAR efforts modestly improved the potency, only a 20-µM IC_{50} for the in vitro interaction was described (Yin et al., 2015). It seems likely that both efforts were hampered by a limited library size available for the initial screening (~ 30,000 and 352 compounds, respectively), whereas the natural product approach taken to discover GM was agnostic to the mechanism of action and only incidentally targeted a PPI.

A3.2. Antiviral PPI Inhibitors.

A3.2.1. Human Papillomavirus Ori-E1-E2 interaction

Human papillomaviruses (HPV) are double-stranded DNA viruses that infect the human epithelium and cause wart formation from the rapid growth of the epithelial tissue. One distinguishing factor of HPVs is that they are among the few known human cancer viruses, where infection of the anogenital region by high-risk strains (HPV-16 and HPV-18) can lead to the development of cervical cancer (Bosch & De Sanjosé, 2003). The increased risk of cancer results from the incorporation of the virus into the host genome, forming a provirus. To continue propagating, the host cell is then compelled to divide, simultaneously copying the provirus. This continued dysregulated replication can eventually lead to cervical cancer (Muñoz, Castellsagué, de González, & Gissmann, 2006).

The low-risk viruses (HPV-6 and HPV-11) are rarely carcinogenic, but can lead to the development of anogenital warts. Wart removal is a painful process and, because the underlying HPV infection is not cleared, warts recur frequently. An antiviral therapy for the low-risk HPV strains could allow for the ultimate clearing of the infection and a permanent wart treatment.

Rather than reproduce as a provirus, low-risk HPV strains maintain their genomes as plasmids in infected cells. HPV relies on a host polymerase to replicate this plasmid, as the HPV genome does not encode a polymerase. Initiation of viral replication depends on the formation of a ternary DNA–protein complex to recruit the polymerase to the origin of replication (Berg & Stenlund, 1997). First, the viral protein E1 binds to the origin through a DNA-binding domain (DBD), which also serves as a dimerization site. A C-terminal helicase domain of E1 mediates an interaction to the transactivation domain of E2 (Chen & Stenlund, 1998). Once located at the origin, an E2 DBD domain tethers the E1–E2–DNA complex and the recruitment of additional E1 dimers leads to the formation of a hexameric structure that encircles and melts the origin DNA (Sedman & Stenlund, 1998). During this process, E2 is displaced allowing for the

recruitment of a host DNA polymerase which replicates the plasmid (Conger, Liu, Kuo, Chow, & Wang, 1999). Disruption of any of these interactions is sufficient to block HPV replication.

To identify inhibitors of E1–E2–Ori complex formation, White and coworkers developed a bead-based scintillation proximity assay (SPA). A DNA substrate containing the HPV origin was radiolabeled and then incubated with E1 and E2. An SPA bead was tether to the complex by an anti-E1 antibody and this allowed for the recruitment of E1 to the DNA to be detected, which requires E1–E2 binding. Functionally, this allowed for the identification of inhibitors of the E1–E2, E1–DNA, or E2–DNA interactions simultaneously, since disruption of any is sufficient to block E1 recruitment. This assay was used to screen a 140,000 compound library and identified a single lead compound with an indanedione scaffold (11 μM IC_{50}) (Fig. A3.4A; White et al., 2003). Additional chlorination of the phenyl ring through SAR studies led to the discovery of compounds with approximately 20-fold greater potency (Yoakim et al., 2003). Activity against the interaction was confirmed by ELISA, and binding to E2 was confirmed by isothermal titration calorimetry (Wang et al., 2004). Unfortunately, when these compounds were tested in cellular assays, EC_{50} values were markedly higher than the activity in biochemical assays had suggested. Together with the fact that the indanedione class of compounds possessed relatively poor pharmacokinetic properties, this series was ultimately abandoned (White, Faucher, & Goudreau, 2010).

However, this well-characterized set of in vitro active compounds allowed the researchers to improve their HTS assay. A tritiated version of an indanedione compound was prepared, and a larger compound library was screened for compounds with the ability displace the tritiated probe. A novel inhibitor scaffold was found to both displace the tritiated probe and block the E1–E2 interaction. The scaffold showed a marked similarity to repaglinide, a diabetes drug with blood glucose-lowering properties (Fig. A3.4B and C). Accordingly, early members of repaglinide series shared this undesirable (for an antiviral agent) property. While no crystal structure for this class was obtained, preliminary optimization SAR suggested that antiviral activity could be maintained while minimizing the glucose-

lowering properties. Despite these promising initial results, the introduction of an effective HPV vaccine was expected to vastly reduce the future demand for HPV therapies and the project was terminated. In this particular case, the failure to generate a clinical agent was not the due to the difficultly of targeting a PPI, but rather external market forces (White et al., 2010).

A3.2.2. Human immunodeficiency Virus.

Prior to the development of antiretrovirus therapy, infection with human immunodeficiency virus (HIV), the causative agent of the acquired immunodeficiency syndrome, was a death sentence. Starting in the 1990s and continuing through today, the introduction of a broad range of HIV treatments has greatly improved patient outcomes. Unfortunately, viruses resistant to every class of HIV therapy have been observed and these can lead to treatment failure (Magambo et al., 2014, Wensing et al., 2015). Using combinations of different classes of antiretrovirals greatly diminishes the risk of developing resistance, but resistance can occur because of monotherapy or if subtherapeutic levels of antiretrovirals are maintained as a result of improper dosing. Furthermore, resistance to one member of an antiretroviral class frequently leads to cross-resistance to all members of that class, which can both reduce the number of available agents and undermine the effectiveness of preexposure prophylaxis (PrEP). In the face of these challenges, the development of novel classes of HIV treatment for both PrEP and therapy is needed (Waheed & Tachedjian, 2016).

Given the limited repertoire of viral proteins, viruses rely on hijacking host proteins to carry out functions in the viral life cycle. Viruses typically rely on protein interactions with the host cellular machinery to enter or otherwise alter the normal cellular function. As such, small-molecule blockades of these interactions can be used to disrupt the viral life cycle. Two key HIV PPIs that have been targeted for disruption as antiviral therapies play vital roles in viral–host fusion and integration into the host genome.

A3.2.3. HIV Entry Inhibitors.

To access the cellular machinery required for replication, viruses must first bypass the cellular membrane. Most viruses accomplish this by binding to a cellular receptor on the outer cellular membrane, which is used as an initiation point for membrane fusion. The structure and identity of the receptor targeted varies between viruses and determines which cell types are targeted (i.e., the tropism). To target CD4 + T cells, the gp120 subunit of the envelope protein first engages the CD4 receptor. CD4 binding causes a conformational shift that allows for binding to an additional T-cell receptor, either CXCR4 or CCR5. Once bound to both receptors, the second subunit of ENV, gp41, penetrates into the cellular membrane and begins the fusion process (Wilen, Tilton, & Doms, 2012).

There is biological support for targeting HIV fusion as a therapeutic route. In the mid-1990s, it was noted that a subset of the population lack cell surface expression of CCR5 and those lacking this receptor appear to be resistant to infection with the HIV-1 strain (Samson, Libert, Doranz, & Rucker, 1996). This is important for two principle reasons. First, this implies that blockade of the CCR5 receptor is sufficient to impede HIV infection. Second, inhibition of the normal function of CCR5 is likely to be tolerated with minimal adverse effects since the $CCR5^{-/-}$ genotype is maintained in the population.

In the early 2000s, there was much interest in developing inhibitors of CD4, CXCR4, or CCR5 receptors to block HIV entry into cells and peptide or small-molecule inhibitors of gp120 binding were developed against each of these targets (Lin et al., 2003). Despite the early success of these entry inhibitors, only two have reached the clinic, primarily due to the poor bioavailability of the CD4 antagonists (Yang et al., 2005) and limited effectiveness of the CXCR4 antagonists (Doranz et al., 2001). The first clinical agent, enfuvirtide, is a peptide that binds to gp120 and blocks entry pore formation. As a peptide inhibitor, it must be dosed intravenously, is difficult to self-administer and is not recommended as first-line therapy (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2016). But because enfuvirtide targets the viral protein gp120 rather than the cellular receptor, it is active against both CCR5 and CXCR4 trophic viruses (Matthews et al., 2004).

The second entry inhibitor approved to date is maraviroc. The lead that was eventually developed into maraviroc was originally identified from a Pfizer HTS. The compound (UK-107,543) was found to block binding of a radiolabeled substrate to CCR5. A heroic medicinal chemistry optimization campaign yielded maraviroc (Fig. A3.5), with low nanomolar IC_{50} values against three CCR5 peptide substrates. Additionally, maraviroc was found to have nanomolar antiviral activity against a variety of both lab-adapted and clinical CCR5 tropic HIV-1 strains. As expected, no activity was observed against CXCR4 tropic strains. Phase I clinical trials demonstrated that therapeutic doses could be reached safely, and after a series of phase III trials demonstrated efficacy, maraviroc was approved for clinical use. Currently, maraviroc is rarely used in treatment-naive patients but is reserved for patients with treatment resistant, CCR5-specific HIV strains (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2016). Despite these limitations, the development of maraviroc represented a major accomplishment in the field of PPI inhibitors.

A3.2.4. HIV Integrase Inhibitors.

Integration of the viral genome into the host chromosome represents another step of the HIV life cycle that has been targeted for therapy. The validity of this target has been established by the clinical success of the integrase active site inhibitor raltegravir (Eron et al., 2013). Relatively high levels of resistance to integrase inhibitors and cross-resistance within the class have limited the utility of these therapies (Malet et al., 2014). Fortunately, integrase relies on an essential PPI where lens epithelium-derived growth factor (LEDGF/p75) tethers integrase to the DNA while stimulating the integrase activity (Poeschla, 2008). This interface is essential to carry out its function and represents a potential therapeutic target.

Several groups have conducted screens to identify small-molecule inhibitors of the LEDGF/p75 interface, including both chemical and in silico screens. The first inhibitor, named D77, was found to be active in a yeast two-hybrid assay, although it proved cytotoxic to noninfected cells. This cytotoxicity likely

results from the presence of a rhodanine functional group in D77 (Fig. A3.6A; Du et al., 2008). Small molecules containing these functional groups are common hits in HTSs as a result of their ability to covalently modify proteins in a nonspecific fashion (Baell & Holloway, 2010). Further in silico screening of clinically approved drugs identified eight inhibitors with IC_{50} values in the low- to mid-micromolar range, although further improvements have not been reported (Hu et al., 2012). A more promising inhibitor scaffold was identified from an effort that used in silico pharmacophore screening to identify commercially available compounds. SAR studies allowed for the development of more potent inhibitors possessing a 2-(*tert*-butoxy) functionality, resulting in CX014442 (Fig. A3.6B), with 69 nM EC_{50} and low cytotoxicity.

Another series of inhibitors of this interface was developed at Boehringer Ingelheim. Rather than specifically targeting the LEDGF/p75 interface, this group screened for inhibitors of the 3'-processing activity of HIV integrase. After successfully screening their compound collection, they noticed that the lead compound bound integrase away from the active site, leading them to name this class of inhibitors the noncatalytic site integrase inhibitors (NCINI). Optimization of this series of compounds yielded a potent inhibitor, BI 224436 (Fig. A3.6C).

Despite the structural similarity between the NCINI and LEDGIN inhibitors, it was not immediately clear that shared a common mechanism of action since they were discovered by different methodologies. Work by Kvaratskhelia and coworkers concluded that both inhibitors blocked LEDGF binding, albeit more potently by BI 224436, and there was substantial overlap of the inhibitor binding sites (Engelman, Kessl, & Kvaratskhelia, 2013).

During in vitro analysis, BI 224436 demonstrated antiviral activity against a range of viruses resistant to other therapeutics and had a promising animal pharmacokinetic profile, which prompted the initiation of a phase I clinical trial in 2010 (Fenwick et al., 2014). While the results from this trial have not

been disclosed and no phase II trials have been started, Gilead licensed the development of the NCINI class of inhibitors in 2011 (“Boehringer license novel HIV non-catalytic integrase inhibitors to Gilead,” 2011). A paper describing the medicinal chemistry optimization of another NCINI derivative for activity against a integrase-resistant mutant and improved pharmacokinetics was published in 2016, suggesting that an integrase inhibitor may yet show clinical efficacy (Fader et al., 2016).

A3.3. Targeting Host-Host PPis to Improve Infection Survival

One of the most lethal aspects of bacterial infections can be the immune response to the insult. While important for the eventual clearance of the pathogens, widespread overactivation of the innate immune system and the resulting systemic inflammation can lead to multiorgan failure or even death (Marshall, 2005). Neutrophil exocytosis is a principle driver of this systemic inflammation, where neutrophils release a range of potent toxins such as reactive oxygen species and proteases (Marshall, 2005). Excessive release of these species does not improve control of the infection, but rather damages surrounding tissues (Narasaraju et al., 2011). In rats, blockade of neutrophil exocytosis was shown to reduce the extent of tissue injury (Uriarte et al., 2013) and this blockade has been proposed as a mechanism to allow for patient survival, while the infection is treated by the antibiotic therapy and the immune system. Related inhibitors would also be expected to be effective for the treatment of some autoimmune conditions.

A3.3.1. Neutrophil Exocytosis Inhibitors (Nexinhibs)

One of the critical regulators of neutrophil exocytosis is the interaction between JFC1 and a GTPase, Rab27a. Downregulation of either of these proteins blocks the exocytosis of neutrophil granules, but neither is known to play a role in phagosome maturation (Brzezinska et al., 2008). This raises the possibility that small-molecule inhibition of the Rab27a–JFC1 interaction would lead to a targeted inhibition of the negative effects of the neutrophil granule release, while allowing for neutrophil survival

and the continued clearance of the pathogens. Because this PPI is a host-specific interaction, resistance is unlikely to be a problem.

A high-throughput time-resolved fluorescence resonance energy transfer was used to screen 32,000 compounds, and lead compounds were confirmed by the reduction in neutrophil myeloperoxidase secretion in a cell-based secondary screen. In this assay, neutrophils were pretreated with the putative nexinhibs (*n*eutrophil *ex*ocytosis *inhib*itor) and then granule release was stimulated. A decrease in the generation of H_2O_2 indicated that the compounds successfully inhibited the release of the myeloperoxidase-containing granules. After excluding peroxide-scavenging compounds, the authors concluded that compounds active in this assay must penetrate the membrane, allowing the authors to eliminate nonpermeable compounds early in the hit reduction process (Johnson et al., 2016).

The most promising compound, Nexinhib 20 (Fig. A3.7), was found to block neutrophil degranulation in response to agonists and disrupted the Rab27a–JFC1 interaction in pulldown and ELISA assays with an IC_{50} of 2.6 μM in the ELISA. Nexinhib 20 treatment did not result in cell death or impair the ability of neutrophils to phagocytize, but did reduce exocytosis after neutrophil stimulation. To examine how nexinhibs function in a physiologic setting, mice were treated with Nexinhib 20 prior to the induction of systemic inflammation by an intraperitoneal injection of lipopolysaccharide. The total number of neutrophils and white blood cells remained unchanged relative to an untreated control, but there was a modest, albeit significant, reduction in the plasma levels of myeloperoxidase as well as neutrophil infiltration of the kidney and liver (Johnson et al., 2016). While further in vivo testing and a mouse sepsis survival study would bolster the case for nexinhibs, this preliminary study is an exciting step.

A3.4. Considerations for Targeting PPIs.

Historically, PPIs were considered undruggable therapeutic targets. One early estimate from Pfizer suggested that while inhibitors for enzymes or GPCR targets could be discovered by high-throughput screening methods at a rate 1 per 10^5 compounds screened, a discovery rate of 1 per 10^9–10^{10} compounds was expected for PPI targets (Spencer, 1998). Given this discouraging prediction, it is obvious why pharmaceutical companies limited their exposure to PPIs. The prevailing view at the time was that protein interfaces consisted of interactions between two broad, mostly hydrophobic and featureless surfaces. The lack of pockets for a hit to gain an initial foothold was also thought to complicate screening.

Two major advances have made targeting PPIs more feasible. First was the realization that while many PPI match the above description, a sizeable number are much simpler. A useful system for classifying the nature and difficulty of targeting a PPI has been developed. Briefly, PPIs can be fit into one of the three categories based on the complexity of the simplest member: (1) primary interfaces consist of an unstructured peptide that binds to a channel. (2) A single fold of secondary structure (α-helix or β-sheet) or (3) a lengthy stretch of protein that uses a tertiary structure (Fig. A3.8; Blundell et al., 2006). While there are examples of successfully targeting each class of PPI, there is a significant bias toward the simpler interfaces (Arkin et al., 2014). Careful selection of a target, ideally guided by structural characterization of the PPI, can greatly increase the odds of success. To help calibrate expectations about a potential project, an online tool has been developed to aid in assessing the “drugability” of a given PPI (Basse et al., 2013, Basse et al., 2016).

The second key development in the targeting of PPI has been improvements in screening methodologies. The introduction and miniaturization of PPI HTS assays and the advent of academic screening facilities have reduced the cost, and therefore the risk, of performing a chemical HTS. There has been debate regarding the suitability of existing screening libraries for targeting PPIs, as PPI inhibitors

tend to be more aromatic, hydrophobic, three-dimensional and have higher molecular weights than non-PPI inhibitors (Kuenemann, Labbe, Cerdan, & Sperandio, 2016). Armed with this information, PPI specific screening libraries have been introduced ("iPPI Focused Libraries," 2017; "Protein-Protein Interaction Libraries—Enamine," Protein-Protein Interaction Libraries—Enamine, 2017, Reynès et al., 2010), although more time is needed to tell if these libraries show improved hit rates. Besides chemical HTS methods, additional hit identification strategies have been successfully implemented. Some have been highlighted in this review and include in silico screening, natural product screening, rational design, scaffold hopping, and fragment screening and elaboration (Sheng et al., 2015). There is unlikely to be a single best method for lead generation, and different approaches may succeed where prior attempts have failed. Because hit identification methods rely on varied input libraries and individual libraries may not contain a bona fide inhibitor, trying multiple screening methods allows for the sampling of a broader range of chemical space than could be achieved by solely by increasing the library size. This is also demonstrated by several of the examples presented here, where different approaches against the same target yielded unique inhibitor scaffolds. Furthermore, the use of phenotypic screens or other assay designs allowing for the discovery of unpredicted mechanisms of action may prove fruitful (Fischer, Rossmann, & Hyvonen, 2015).

While the case studies presented here represent the successful inhibition of a range of protein interactions, most or all of these compounds will still fail to reach the clinic. It is difficult to assess the true success rate of PPI-targeted drug discovery efforts as publication and survivorship bias obscures the denominator of this calculation. However, from the examples presented here and elsewhere, it appears that small-molecule disruption of many PPIs is challenging yet feasible. The development of existing inhibitors appears to be hindered by nonspecific effects, poor pharmacokinetics, or cell penetrance, which are challenges common to all therapeutics. As seen with the introduction of oncologic therapies targeting PPIs, once the pipeline of antiinfective PPI inhibitors expands, we have no reason to doubt that they will find increasing clinical success.

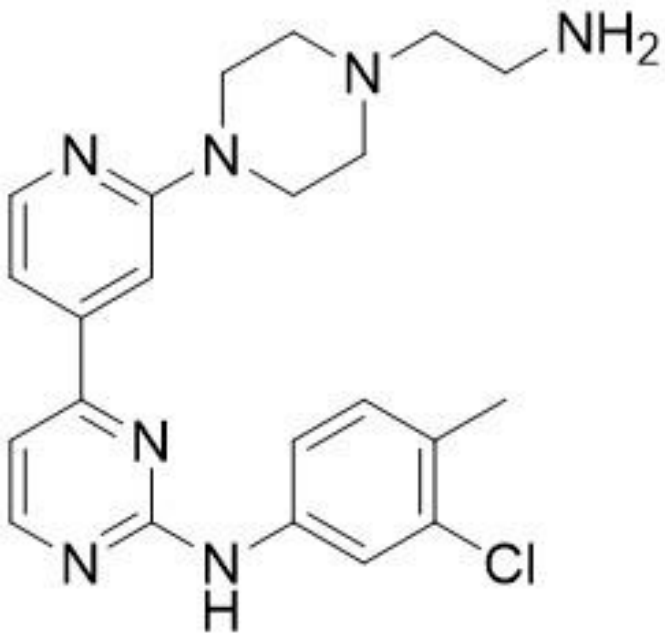


Figure A3.1. Structure of the pyridylpyrimidine HTS hit.

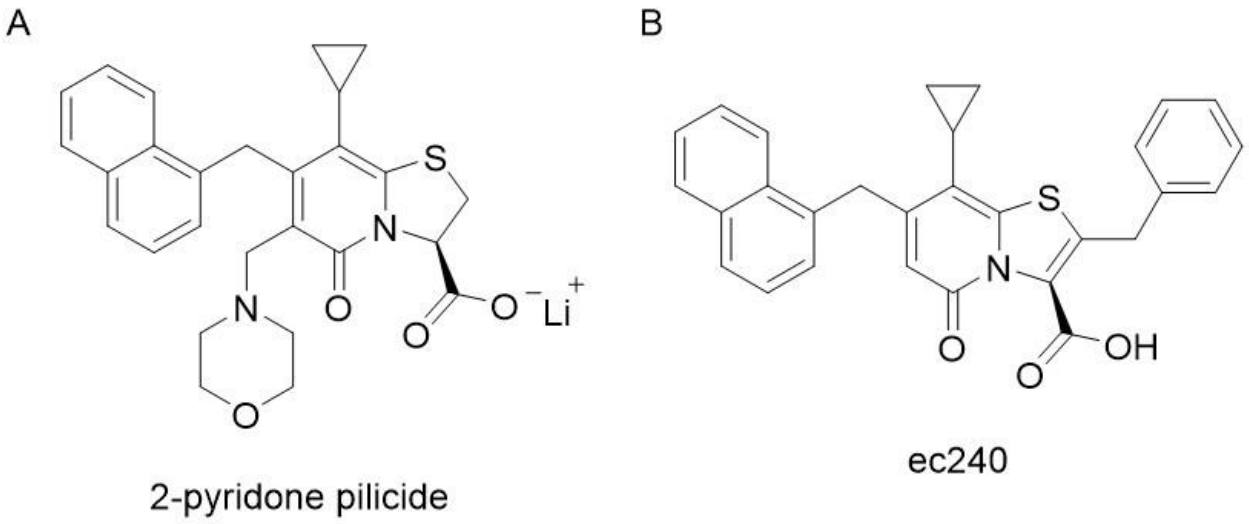


Figure A3.2. Pilicides. (A) Structure of the 2-pyridone-based pilicide. (B) Structure of the pilicide ec240.

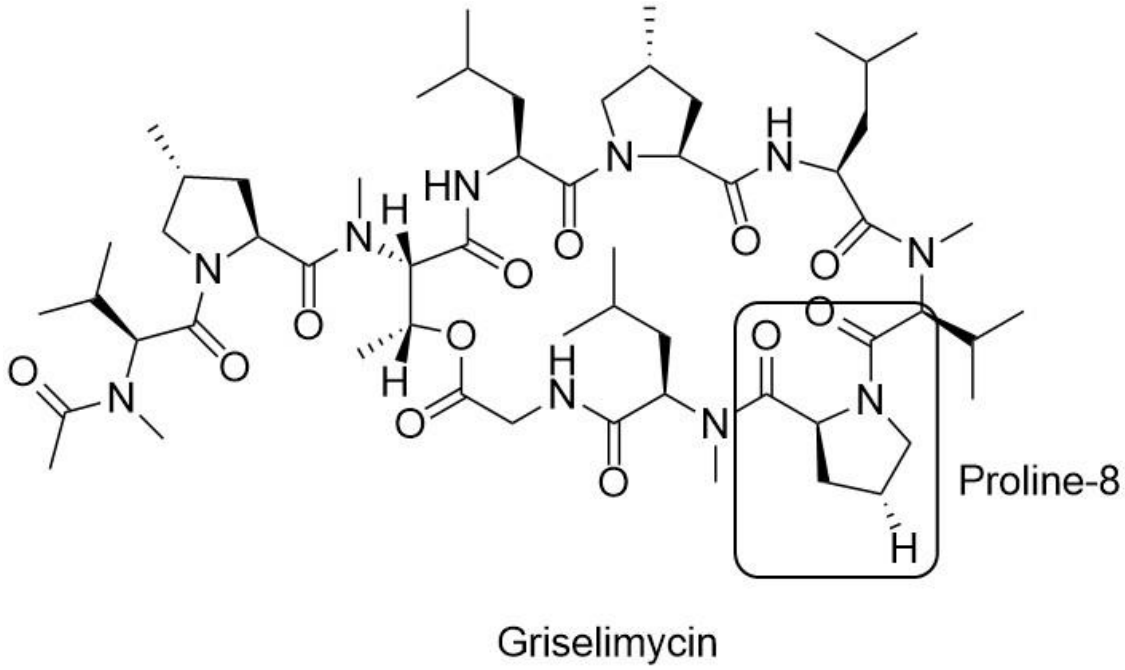


Figure A3.3. Structure of the DnaN–DnaE1 PPI inhibitor griselimycin. Methylation or cyclohexylation of the labeled hydrogen of the *boxed* residue, proline-8, leads to improved potency and pharmacokinetic properties.

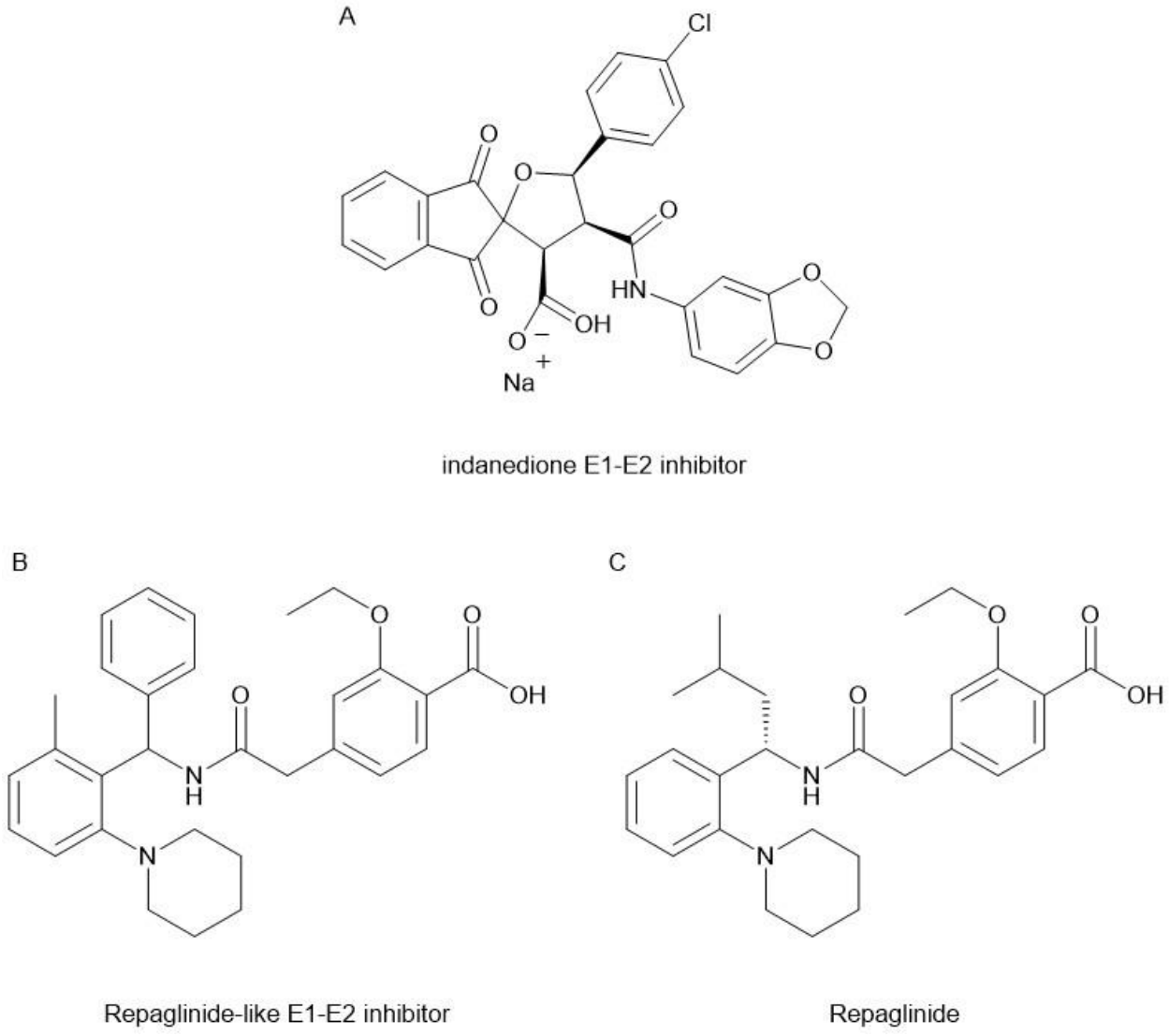


Figure A3.4. Structures of the inhibitors of the E1–E2 interface. The repaglinide like class of inhibitors was found to have undesirable hypoglycemic effects, resulting from the chemical similarity to repaglinide.

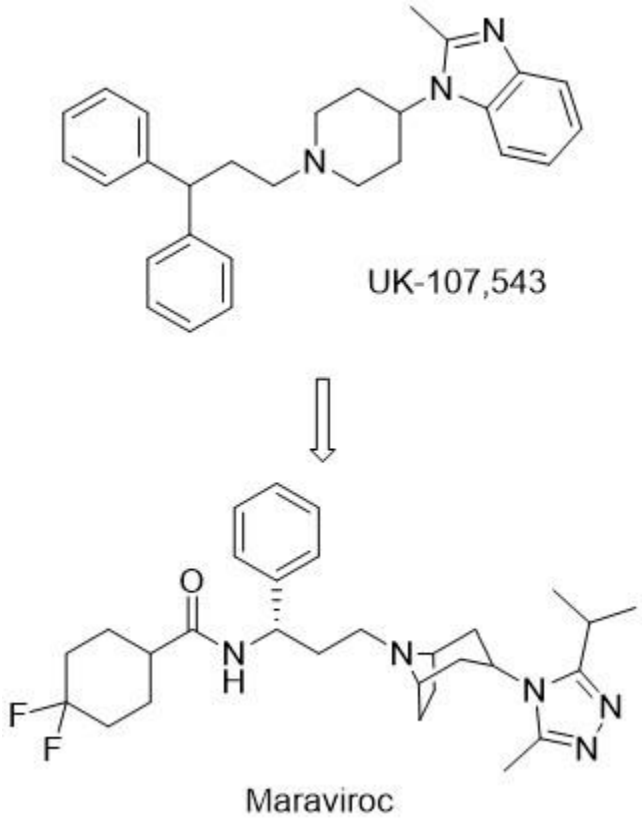


Figure A3.5. CCR5 antagonist structures. Structure of the initial screening hit for CCR5 antagonist, UK-107,543, and the optimized inhibitor, Maraviroc.

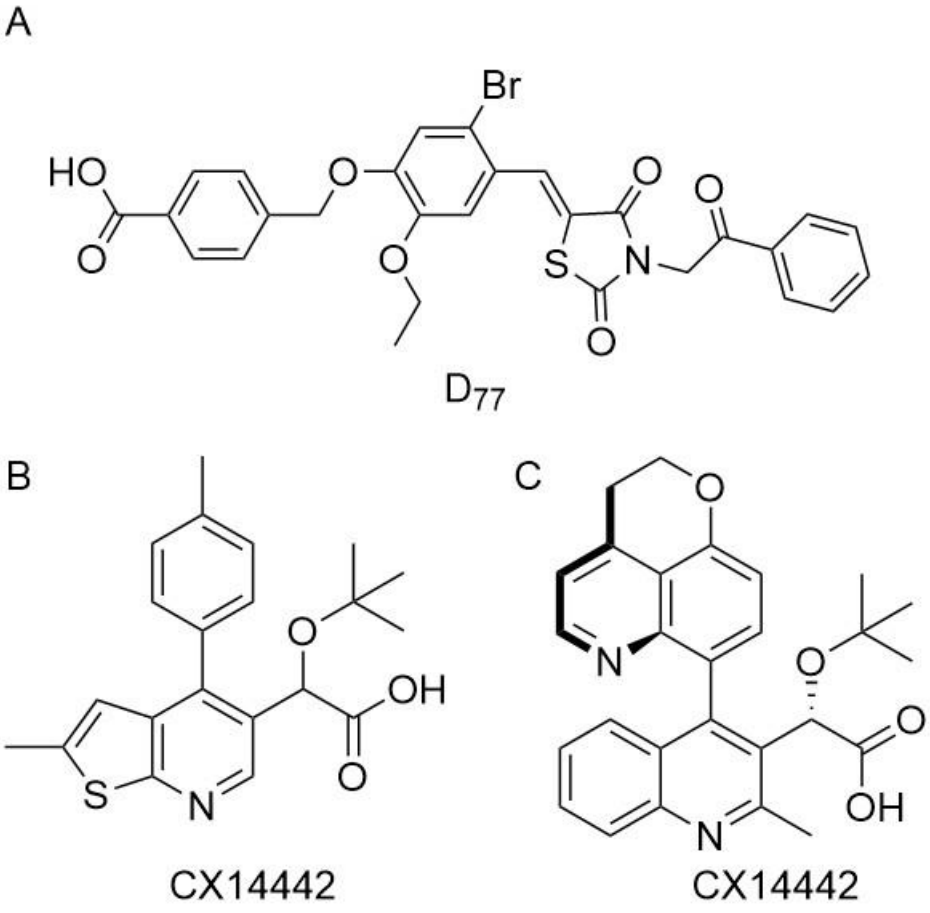
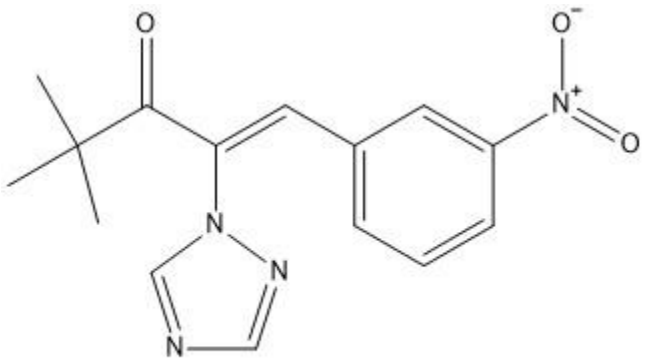


Figure A3.6. Structure of representative inhibitors of integrase/LEDGF interaction.



Nexinhib 20

Figure A3.7. Structure of Nexinhib 20, an inhibitor of the neutrophil exocytosis.

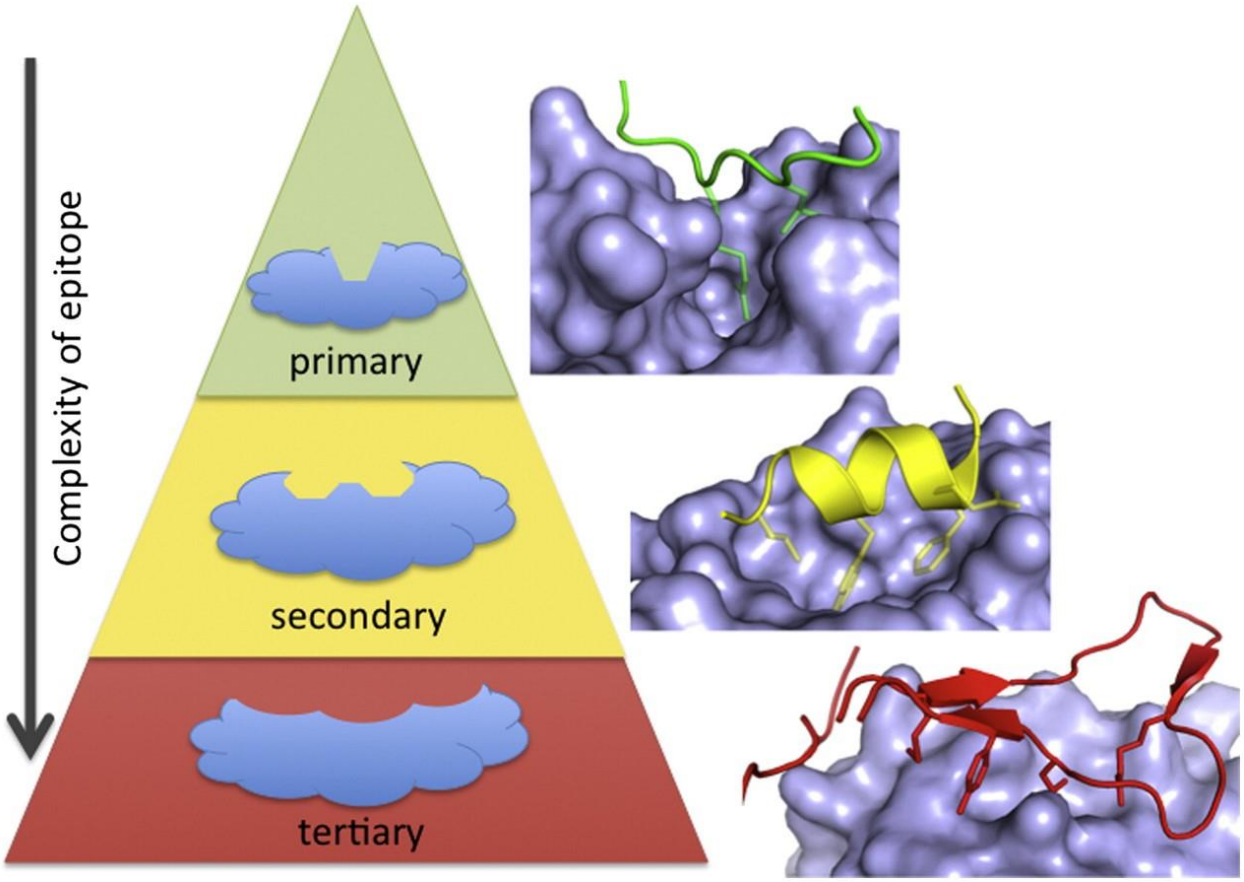


Figure A3.8. Scheme for classifying PPIs by the nature of the interaction site. Most of the successful PPI inhibitors target the simpler interactions, such as the primary or secondary ones, although there are a examples of inhibitors of tertiary interfaces. Reprinted from Arkin, M. R., Tang, Y., & Wells, J. A. (2014). Small-molecule inhibitors of protein-protein interactions: Progressing toward the reality. *Chemistry & Biology*, *21,* 1102–1114, with permission from Elsevier.

References.

The 10 x '20 Initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020. (2010). *Clin Infect Dis, 50*(8), 1081-1083. doi:10.1086/652237

A Waheed, A., & Tachedjian, G. (2016). Why Do We Need New Drug Classes for HIV Treatment and Prevention? *Current topics in medicinal chemistry, 16*(12), 1343-1349.

Adams, D. W., & Errington, J. (2009). Bacterial cell division: assembly, maintenance and disassembly of the Z ring. *Nature Reviews Microbiology, 7*(9), 642-653.

Arkin, M. R., Tang, Y., & Wells, J. A. (2014). Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. *Chem Biol, 21*(9), 1102-1114.

Arkin, M. R., & Wells, J. A. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov, 3*(4), 301-317.

Baell, J. B., & Holloway, G. A. (2010). New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of medicinal chemistry, 53*(7), 2719-2740.

Basse, M. J., Betzi, S., Bourgeas, R., Bouzidi, S., Chetrit, B., Hamon, V., . . . Roche, P. (2013). 2P2Idb: a structural database dedicated to orthosteric modulation of protein-protein interactions. *Nucleic Acids Res, 41*(Database issue), D824-827.

Basse, M. J., Betzi, S., Morelli, X., & Roche, P. (2016). 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions. *Database (Oxford), 2016*.

Bent, S., Nallamothu, B. K., Simel, D. L., Fihn, S. D., & Saint, S. (2002). Does this woman have an acute uncomplicated urinary tract infection? *JAMA, 287*(20), 2701-2710.

Blundell, T. L., Sibanda, B. L., Montalvão, R. W., Brewerton, S., Chelliah, V., Worth, C. L., . . . Burke, D. (2006). Structural biology and bioinformatics in drug design: opportunities and challenges for

target identification and lead discovery. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 361*(1467), 413-423.

Boehringer license novel HIV non-catalytic integrase inhibitors to Gilead. (2011). Retrieved from https://www.thepharmaletter.com/article/boehringer-license-novel-hiv-non-catalytic-integrase-inhibitors-to-gilead

Bosch, F. X., & De Sanjosé, S. (2003). Human papillomavirus and cervical cancer—burden and assessment of causality. *JNCI Monographs, 2003*(31), 3-13.

Brzezinska, A. A., Johnson, J. L., Munafo, D. B., Crozat, K., Beutler, B., Kiosses, W. B., . . . Catz, S. D. (2008). The Rab27a effectors JFC1/Slp1 and Munc13-4 regulate exocytosis of neutrophil granules. *Traffic, 9*(12), 2151-2164.

Busch, A., & Waksman, G. (2012). Chaperone–usher pathways: diversity and pilus assembly mechanism. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1592), 1112-1122.

Chorell, E., Pinkner, J. S., Phan, G., Edvinsson, S., Buelens, F., Remaut, H., . . . Almqvist, F. (2010). Design and Synthesis of C-2 Substituted Thiazolo and Dihydrothiazolo Ring-Fused 2-Pyridones: Pilicides with Increased Antivirulence Activity. *Journal of medicinal chemistry, 53*(15), 5690-5695.

Conger, K. L., Liu, J.-S., Kuo, S.-R., Chow, L. T., & Wang, T. S.-F. (1999). Human Papillomavirus DNA Replication: INTERACTIONS BETWEEN THE VIRAL E1 PROTEIN AND TWO SUBUNITS OF HUMAN DNA POLYMERASE α/PRIMASE. *Journal of Biological Chemistry, 274*(5), 2696-2705.

Cukuroglu, E., Engin, H. B., Gursoy, A., & Keskin, O. (2014). Hot spots in protein–protein interfaces: Towards drug discovery. *Progress in Biophysics and Molecular Biology, 116*(2–3), 165-173.

Cusumano, C. K., Pinkner, J. S., Han, Z., Greene, S. E., Ford, B. A., Crowley, J. R., . . . Hultgren, S. J. (2011). Treatment and prevention of urinary tract infection with orally active FimH inhibitors. *Sci Transl Med, 3*(109), 109ra115.

Dethlefsen, L., Huse, S., Sogin, M. L., & Relman, D. A. (2008). The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing. *PLOS Biology, 6*(11), e280. d

Doranz, B. J., Filion, L. G., Diaz-Mitoma, F., Sitar, D. S., Sahai, J., Baribaud, F., . . . Doms, R. W. (2001). Safe use of the CXCR4 inhibitor ALX40-4C in humans. *AIDS Res Hum Retroviruses, 17*(6), 475-486.

Du, L., Zhao, Y., Chen, J., Yang, L., Zheng, Y., Tang, Y., . . . Jiang, H. (2008). D77, one benzoic acid derivative, functions as a novel anti-HIV-1 inhibitor targeting the interaction between integrase and cellular LEDGF/p75. *Biochem Biophys Res Commun, 375*(1), 139-144.

Engelman, A., Kessl, J. J., & Kvaratskhelia, M. (2013). Allosteric inhibition of HIV-1 integrase activity. *Current opinion in chemical biology, 17*(3), 339-345.

Eron, J. J., Cooper, D. A., Steigbigel, R. T., Clotet, B., Gatell, J. M., Kumar, P. N., . . . Teppler, H. (2013). Efficacy and safety of raltegravir for treatment of HIV for 5 years in the BENCHMRK studies: final results of two randomised, placebo-controlled trials. *The Lancet Infectious Diseases, 13*(7), 587-596.

Fader, L. D., Bailey, M., Beaulieu, E., Bilodeau, F., Bonneau, P., Bousquet, Y., . . . Duan, J. (2016). Aligning Potency and Pharmacokinetic Properties for Pyridine-Based NCINIs. *ACS medicinal chemistry letters, 7*(8), 797-801.

Fenwick, C., Bailey, M. D., Bethell, R., Bös, M., Bonneau, P., Cordingley, M., . . . Fader, L. D. (2014). Preclinical profile of BI 224436, a novel HIV-1 non-catalytic-site integrase inhibitor. *Antimicrobial Agents and Chemotherapy, 58*(6), 3233-3244.

Fischer, G., Rossmann, M., & Hyvonen, M. (2015). Alternative modulation of protein-protein interactions by small molecules. *Curr Opin Biotechnol, 35*, 78-85.

Flores-Mireles, A. L., Walker, J. N., Caparon, M., & Hultgren, S. J. (2015). Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat Rev Micro, 13*(5), 269-284.

Georgescu, R. E., Yurieva, O., Kim, S. S., Kuriyan, J., Kong, X. P., & O'Donnell, M. (2008). Structure of a small-molecule inhibitor of a DNA polymerase sliding clamp. *Proc Natl Acad Sci U S A, 105*(32), 11116-11121.

Global tuberculosis report 2016. (2016).

Goff, D. A., Kullar, R., Goldstein, E. J., Gilchrist, M., Nathwani, D., Cheng, A. C., . . . Brink, A. (2017). A global call from five countries to collaborate in antibiotic stewardship: United we succeed, divided we might fail. *The Lancet Infectious Diseases, 17*(2), e56-e63.

Greene, S. E., Pinkner, J. S., Chorell, E., Dodson, K. W., Shaffer, C. L., Conover, M. S., . . . Hultgren, S. J. (2014). Pilicide ec240 Disrupts Virulence Circuits in Uropathogenic Escherichia coli. *mBio, 5*(6).

Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. (2016). *Panel on Antiretroviral Guidelines for Adults and Adolescents*.

Gupta, K., Hooton, T. M., Naber, K. G., Wullt, B., Colgan, R., Miller, L. G., . . . Soper, D. E. (2011). International Clinical Practice Guidelines for the Treatment of Acute Uncomplicated Cystitis and Pyelonephritis in Women: A 2010 Update by the Infectious Diseases Society of America and the European Society for Microbiology and Infectious Diseases. *Clinical Infectious Diseases, 52*(5), e103-e120.

Hale, C. A., & de Boer, P. A. (1997). Direct binding of FtsZ to ZipA, an essential component of the septal ring structure that mediates cell division in E. coli. *Cell, 88*(2), 175-185.

Han, Z., Pinkner, J. S., Ford, B., Obermann, R., Nolan, W., Wildman, S. A., . . . Janetka, J. W. (2010). Structure-Based Drug Design and Optimization of Mannoside Bacterial FimH Antagonists. *Journal of medicinal chemistry, 53*(12), 4779-4792.

Hedenstrom, M., Emtenas, H., Pemberton, N., Aberg, V., Hultgren, S. J., Pinkner, J. S., . . . Kihlberg, J. (2005). NMR studies of interactions between periplasmic chaperones from uropathogenic E. coli

and pilicides that interfere with chaperone function and pilus assembly. *Organic & biomolecular chemistry, 3*(23), 4193-4200.

Hu, G., Li, X., Sun, X., Lu, W., Liu, G., Huang, J., . . . Tang, Y. (2012). Identification of old drugs as potential inhibitors of HIV-1 integrase - human LEDGF/p75 interaction via molecular docking. *J Mol Model, 18*(12), 4995-5003.

iPPI Focused Libraries. (2017). Retrieved from http://www.otavachemicals.com/products/targeted-libraries-and-focused-libraries/protein-protein-interaction

Jassal, M., & Bishai, W. R. Extensively drug-resistant tuberculosis. *The Lancet Infectious Diseases, 9*(1), 19-30.

Jennings, L. D., Foreman, K. W., Rush, T. S., Tsao, D. H., Mosyak, L., Li, Y., . . . Kenny, C. H. (2004). Design and synthesis of indolo [2, 3-a] quinolizin-7-one inhibitors of the ZipA–FtsZ interaction. *Bioorganic & medicinal chemistry letters, 14*(6), 1427-1431.

Kenny, C. H., Ding, W., Kelleher, K., Benard, S., Dushin, E. G., Sutherland, A. G., . . . Ellestad, G. (2003). Development of a fluorescence polarization assay to screen for inhibitors of the FtsZ/ZipA interaction. *Analytical biochemistry, 323*(2), 224-233.

Kling, A., Lukat, P., Almeida, D. V., Bauer, A., Fontaine, E., Sordello, S., . . . Muller, R. (2015). Antibiotics. Targeting DnaN for tuberculosis therapy using novel griselimycins. *Science, 348*(6239), 1106-1112.

Krogfelt, K. A., Bergmans, H., & Klemm, P. (1990). Direct evidence that the FimH protein is the mannose-specific adhesin of Escherichia coli type 1 fimbriae. *Infection and immunity, 58*(6), 1995-1998.

Kuchibhatla, A., Bhattacharya, A., & Panda, D. (2011). ZipA binds to FtsZ with high affinity and enhances the stability of FtsZ protofilaments. *PLoS One, 6*(12), e28262.

Kuenemann, M. A., Labbe, C. M., Cerdan, A. H., & Sperandio, O. (2016). Imbalance in chemical space: How to facilitate the identification of protein-protein interaction inhibitors. *Sci Rep, 6*, 23815.

Kumar, R., & Nanduri, B. (2010). HPIDB - a unified resource for host-pathogen interactions. *BMC Bioinformatics, 11*(6), S16. doi:10.1186/1471-2105-11-s6-s16

Lane, M., & Mobley, H. (2007). Role of P-fimbrial-mediated adherence in pyelonephritis and persistence of uropathogenic Escherichia coli (UPEC) in the mammalian kidney. *Kidney international, 72*(1), 19-25.

Lehne, B., & Schlitt, T. (2009). Protein-protein interaction databases: keeping up with growing interactomes. *Human Genomics, 3*(3), 291.

Lin, P.-F., Blair, W., Wang, T., Spicer, T., Guo, Q., Zhou, N., . . . Colonno, R. (2003). A small molecule HIV-1 inhibitor that targets the HIV-1 envelope and inhibits CD4 receptor binding. *Proceedings of the National Academy of Sciences, 100*(19), 11013-11018.

Magambo, B., Nazziwa, J., Bbosa, N., Gupta, R. K., Kaleebu, P., & Parry, C. M. (2014). The arrival of untreatable multidrug-resistant HIV-1 in sub-Saharan Africa. *Aids, 28*(9), 1373-1374.

Malet, I., Arriaga, L. G., Artese, A., Costa, G., Parrotta, L., Alcaro, S., . . . Valantin, M.-A. (2014). New raltegravir resistance pathways induce broad cross-resistance to all currently used integrase inhibitors. *Journal of Antimicrobial Chemotherapy, 69*(8), 2118-2122.

Marshall, J. C. (2005). Neutrophils in the pathogenesis of sepsis. *Critical care medicine, 33*(12), S502-S505.

Martinez, J. J., Mulvey, M. A., Schilling, J. D., Pinkner, J. S., & Hultgren, S. J. (2000). Type 1 pilus-mediated bacterial invasion of bladder epithelial cells. *The EMBO journal, 19*(12), 2803-2812.

Matthews, T., Salgo, M., Greenberg, M., Chung, J., DeMasi, R., & Bolognesi, D. (2004). Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes. *Nat Rev Drug Discov, 3*(3), 215-225.

Mosyak, L., Zhang, Y., Glasfeld, E., Haney, S., Stahl, M., Seehra, J., & Somers, W. S. (2000). The bacterial cell-division protein ZipA and its interaction with an FtsZ fragment revealed by X-ray crystallography. *EMBO J, 19*(13), 3179-3191.

Muñoz, N., Castellsagué, X., de González, A. B., & Gissmann, L. (2006). Chapter 1: HPV in the etiology of human cancer. *Vaccine, 24, Supplement 3*, S1-S10.

Narasaraju, T., Yang, E., Samy, R. P., Ng, H. H., Poh, W. P., Liew, A.-A., . . . Chow, V. T. (2011). Excessive Neutrophils and Neutrophil Extracellular Traps Contribute to Acute Lung Injury of Influenza Pneumonitis. *The American Journal of Pathology, 179*(1), 199-210.

Nuccio, S.-P., & Bäumler, A. J. (2007). Evolution of the Chaperone/Usher Assembly Pathway: Fimbrial Classification Goes Greek. *Microbiology and Molecular Biology Reviews, 71*(4), 551-575.

Paterson, D. L. (2015). The emerging threat of multidrug-resistant Gram-negative bacteria in urology. *Nat Rev Urol, 12*(10), 570-584.

Pinkner, J. S., Remaut, H., Buelens, F., Miller, E., Aberg, V., Pemberton, N., . . . Almqvist, F. (2006). Rationally designed small compounds inhibit pilus biogenesis in uropathogenic bacteria. *Proc Natl Acad Sci U S A, 103*(47), 17897-17902.

Poeschla, E. M. (2008). Integrase, LEDGF/p75 and HIV replication. *Cell Mol Life Sci, 65*(9), 1403-1424.

Protein-Protein Interaction Libraries - Enamine. (2017). Retrieved from http://www.enamine.net/index.php?option=com_content&task=view&id=227

Reynès, C., Host, H., Camproux, A.-C., Laconde, G., Leroux, F., Mazars, A., . . . Sperandio, O. (2010). Designing Focused Chemical Libraries Enriched in Protein-Protein Interaction Inhibitors using Machine-Learning Methods. *PLOS Computational Biology, 6*(3), e1000695.

Rush, T. S., 3rd, Grant, J. A., Mosyak, L., & Nicholls, A. (2005). A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem, 48*(5), 1489-1495.

Samson, M., Libert, F., Doranz, B. J., & Rucker, J. (1996). Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature, 382*(6593), 722.

Schappert, S. M. (2011). Ambulatory medical care utilization estimates for 2007. *Vital and health statistics.*(169), 1-38.

Sheng, C., Dong, G., Miao, Z., Zhang, W., & Wang, W. (2015). State-of-the-art strategies for targeting protein-protein interactions by small-molecule inhibitors. *Chemical Society Reviews, 44*(22), 8238-8259.

Slonim, L. N., Pinkner, J. S., Branden, C. I., & Hultgren, S. J. (1992). Interactive surface in the PapD chaperone cleft is conserved in pilus chaperone superfamily and essential in subunit recognition and assembly. *EMBO J, 11*(13), 4747-4756.

Spencer, R. W. (1998). High-throughput screening of historic collections: Observations on file size, biological targets, and file diversity. *Biotechnology and bioengineering, 61*(1), 61-67.

Sutherland, A. G., Alvarez, J., Ding, W., Foreman, K., Kenny, C. H., Labthavikul, P., . . . Ruzin, A. (2003). Structure-based design of carboxybiphenylindole inhibitors of the ZipA–FtsZ interaction. *Organic & biomolecular chemistry, 1*(23), 4138-4140.

Svensson, A., Larsson, A., Emtenäs, H., Hedenström, M., Fex, T., Hultgren, S. J., . . . Kihlberg, J. (2001). Design and evaluation of pilicides: potential novel antibacterial agents directed against uropathogenic Escherichia coli. *Chem. Biochem., 2*(12), 915-918.

Tsao, D. H., Sutherland, A. G., Jennings, L. D., Li, Y., Rush, T. S., Alvarez, J. C., . . . Haney, S. A. (2006). Discovery of novel inhibitors of the ZipA/FtsZ complex by NMR fragment screening coupled with structure-based design. *Bioorganic & medicinal chemistry, 14*(23), 7953-7961.

Udwadia, Z. F., Amale, R. A., Ajbani, K. K., & Rodrigues, C. (2012). Totally Drug-Resistant Tuberculosis in India. *Clinical Infectious Diseases, 54*(4), 579-581.

Uriarte, S. M., Rane, M. J., Merchant, M. L., Jin, S., Lentsch, A. B., Ward, R. A., & McLeish, K. R. (2013). Inhibition of Neutrophil Exocytosis Ameliorates Acute Lung Injury in Rats. *Shock (Augusta, Ga.), 39*(3), 286-292.

Ventola, C. L. (2015). The Antibiotic Resistance Crisis: Part 1: Causes and Threats. *Pharmacy and Therapeutics, 40*(4), 277-283.

Waksman, G., & Hultgren, S. J. (2009). Structural biology of the chaperone-usher pathway of pilus biogenesis. *Nat Rev Micro, 7*(11), 765-774.

Walsh, C. a. (2016). *Antibiotics : challenges, mechanisms, opportunities*: Washington, DC : ASM Press, [2016] ©2016.

Wang, Y., Coulombe, R., Cameron, D. R., Thauvette, L., Massariol, M.-J., Amon, L. M., . . . Yoakim, C. (2004). Crystal structure of the E2 transactivation domain of human papillomavirus type 11 bound to a protein interaction inhibitor. *Journal of Biological Chemistry, 279*(8), 6976-6985.

Wells, J. A., & McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature, 450*(7172), 1001-1009.

Wensing, A. M., Calvez, V., Günthard, H. F., Johnson, V. A., Paredes, R., Pillay, D., . . . Richman, D. D. (2015). 2015 update of the drug resistance mutations in HIV-1. *Top Antivir Med, 23*(4), 132-141.

White, P. W., Faucher, A.-M., & Goudreau, N. (2010). Small molecule inhibitors of the human papillomavirus E1-E2 interaction *Small-Molecule Inhibitors of Protein-Protein Interactions* (pp. 61-88): Springer.

White, P. W., Titolo, S., Brault, K., Thauvette, L., Pelletier, A., Welchner, E., . . . Yoakim, C. (2003). Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1-E2 protein interaction. *Journal of Biological Chemistry, 278*(29), 26765-26772.

Wilen, C. B., Tilton, J. C., & Doms, R. W. (2012). HIV: cell binding and entry. *Cold Spring Harb Perspect Med, 2*(8).

Wright, K. J., Seed, P. C., & Hultgren, S. J. (2007). Development of intracellular bacterial communities of uropathogenic Escherichia coli depends on type 1 pili. *Cellular microbiology, 9*(9), 2230-2241.

Wu, X.-R., Sun, T.-T., & Medina, J. J. (1996). In vitro binding of type 1-fimbriated Escherichia coli to uroplakins Ia and Ib: relation to urinary tract infections. *Proceedings of the National Academy of Sciences, 93*(18), 9630-9635.

Xiao, J., & Goley, E. D. (2016). Redefining the roles of the FtsZ-ring in bacterial cytokinesis. *Current Opinion in Microbiology, 34*, 90-96.

Yang, Z., Zadjura, L., D'arienzo, C., Marino, A., Santone, K., Klunk, L., . . . Wang, T. (2005). Preclinical pharmacokinetics of a novel HIV-1 attachment inhibitor BMS-378806 and prediction of its human pharmacokinetics. *Biopharmaceutics & drug disposition, 26*(9), 387-402.

Yin, Z., Whittell, L. R., Wang, Y., Jergic, S., Liu, M., Harry, E. J., . . . Oakley, A. J. (2014). Discovery of lead compounds targeting the bacterial sliding clamp using a fragment-based approach. *J Med Chem, 57*(6), 2799-2806.

Yin, Z., Whittell, L. R., Wang, Y., Jergic, S., Ma, C., Lewis, P. J., . . . Oakley, A. J. (2015). Bacterial Sliding Clamp Inhibitors that Mimic the Sequential Binding Mechanism of Endogenous Linear Motifs. *J Med Chem, 58*(11), 4693-4702. doi:10.1021/acs.jmedchem.5b00232

Yoakim, C., Ogilvie, W. W., Goudreau, N., Naud, J., Hache, B., O'Meara, J. A., . . . White, P. W. (2003). Discovery of the first series of inhibitors of human papillomavirus type 11: inhibition of the assembly of the E1–E2–Origin DNA complex. *Bioorganic & medicinal chemistry letters, 13*(15), 2539-2541.

Zowawi, H. M., Harris, P. N. A., Roberts, M. J., Tambyah, P. A., Schembri, M. A., Pezzani, M. D., . . .

Appendix 4

Practical model selection for prospective virtual screening

This work has been published:

Liu, S., Alnammi, M., Ericksen, S. S., Voter, A. F., Ananiev, G., Keck, J. L., Hoffmann, F. M. Wildman, S. A., Gitter, A. (2018) Practical model selection for prospective virtual screening. *Journal of Chemical Information and Modeling.* **2019**, *59* (1), 282–293.

Andrew Voter performed the new PriA-SSB high-throughput screening and provided critical feedback during the preparation of the manuscript.

Abstract

Virtual (computational) high-throughput screening provides a strategy for prioritizing compounds for experimental screens, but the choice of virtual screening algorithm depends on the data set and evaluation strategy. We consider a wide range of ligand-based machine learning and docking-based approaches for virtual screening on two protein–protein interactions, PriA-SSB and RMI-FANCM, and present a strategy for choosing which algorithm is best for prospective compound prioritization. Our workflow identifies a random forest as the best algorithm for these targets over more sophisticated neural network-based models. The top 250 predictions from our selected random forest recover 37 of the 54 active compounds from a library of 22,434 new molecules assayed on PriA-SSB. We show that virtual screening methods that perform well on public data sets and synthetic benchmarks, like multi-task neural networks, may not always translate to prospective screening performance on a specific assay of interest.

Introduction

Drug discovery is time consuming and expensive. After a specific protein or mechanistic pathway is identified to play an essential role in a disease process, the search begins for a chemical or biological ligand that can perturb the action or abundance of the disease target in order to mitigate the disease phenotype. A standard approach to discover a chemical ligand is to screen thousands to millions of candidate compounds against the target in biochemical- or cell-based assays via a process called high-throughput screening (HTS), which produces vast sets of valuable pharmacological data. Even though HTS assays are highly automated, screens of thousands of compounds sample only a small fraction of the millions of commercially available drug-like compounds. Cost and time preclude academic laboratories and even pharmaceutical companies from blindly testing the full set of drug-like compounds in HTS assays. Thus, there is a crucial need for an effective virtual screening (VS) process as a preliminary step in prioritizing compounds for HTS assays.

Virtual screening comprises two categories: structure-based[1,2] and ligand-based methods.[3,4] Structure-based methods require that the target protein's molecular structure is known so that the 3D interactions between the target and each chemical compound (binding poses) may be predicted in silico. These interactions are given numeric scores, which are then used to rank compounds for potential binding to the target. These methods do not require or typically make use of historical screening data in compound scoring. In contrast, ligand-based methods require no structural information about the target. They use data generated from testing molecules in biochemical or functional assays of the target to fit empirical models that relate compound attributes to assay outcomes.

For targets with abundant assay data or where a druggable binding site is not well defined, such as the targets considered here, ligand-based methods are generally superior to structure-based methods.[5-7] Confronted with the variety of ligand-based model building methods (e.g., regression models,

random forests, support vector machines, etc.),[8] compound input representations, and performance metrics, how should one proceed with VS on a new target? The Merck Molecular Activity Challenge[9] incited the development of many ligand-based deep learning VS methods[10-14] as recently reviewed.[15-16] These methods are often assessed with cross-validation on existing HTS data, but there is presently little experimental evidence on the best option for prioritizing new compounds given a fixed screening budget.

We critically evaluated a collection of VS algorithms that include both structure-based and ligand-based methods, with a focus on the subset of quantitative structure–activity relationship ligand-based methods that use machine learning to predict active compounds for a target based on initial screening data. We present a VS workflow that first uses available HTS training data to systematically prune the specific versions of the algorithms and calculate their cross-validation performance on a variety of evaluation metrics. Based on the cross-validation results and analysis of the various evaluation metrics, we selected a single virtual screening algorithm. The selected method, a random forest model, was the best option for prioritizing a small number of compounds from a new library, as verified by experimental screening. These model selection and evaluation strategies can guide VS practitioners to select the best model for their target even as the landscape of available VS algorithms continues to evolve.

Methods

Data Sets. Our case studies were on new and recently generated data sets[17,18] for the targets PriA-SSB and RMI-FANCM. The PriA-SSB interaction is important in bacterial DNA replication and is a potential target for antibiotics.[19] The RMI-FANCM interaction is involved in DNA repair that is induced in human cancer cells to confer chemoresistance to cytotoxic DNA-cross-linking agents, making it an attractive drug target.[20] We previously screened these targets with a library of compounds obtained from Life Chemicals, Inc. (LC) in different assay formats. In addition, we screened new LC compounds on the PriA-SSB target to evaluate our VS models. The four data sets derived from these screens are described below and summarized in Table A4.1

PriA-SSB AlphaScreen. PriA-SSB was initially screened using an AlphaScreen (AS) assay in a 1536-well format[18] on 72,423 LC compounds at a single concentration (33.3 μM), with data reported as % inhibition compared to controls. We refer to these continuous values as a "PriA-SSB % inhibition". Those compounds that tested above an activity threshold (≥35% inhibition) and passed PAINS chemical structural filters[21,22] were retested in the same AS assay. PAINS filters are not a technical necessity of any VS method, and some analyses have shown they are imperfect filters of nonspecific pan assay interference.[23] Nevertheless, they are a common requirement for publication of HTS and medicinal chemistry projects. We did not remove compounds detected by PAINS filters from the data set but rather flagged them and labeled them as inactive. Compounds that were confirmed in the AS retest screen (≥35% inhibition) were marked as actives, creating the binary data set PriA-SSB AS.

PriA-SSB Fluorescence Polarization. Compounds that had PriA-SSB % inhibition ≥ 35% and passed the PAINS filters were also tested in a fluorescence polarization (FP) assay as a secondary screen. Those compounds with FP inhibition ≥ 30% were labeled as actives, creating the binary data set PriA-SSB FP, with all other compounds in the screening set labeled inactive.

RMI-FANCM Fluorescence Polarization. The RMI-FANCM interaction was initially screened with a subset of 49,796 compounds from the same LC library as PriA-SSB.[17] This FP assay was run at a single compound concentration (32 μM). We refer to these continuous values as "RMI-FANCM % inhibition". Those compounds that demonstrated activity ≥ 2 standard deviations (SD) above the assay mean and passed PAINS filters were marked as actives in the binary data set RMI-FANCM FP.

PriA-SSB Prospective. For prospective testing, we experimentally screened an additional 22,434 compounds after the VS methods predicted their activity. We removed compounds that were already included in the 72,423 LC compounds in the PriA-SSB AS data set to ensure there was no overlap between the prospective screen compounds and those used to train VS models. As with the initial library, the PriA-SSB AS assay was used in the same 1536-well format at a single concentration (33.3 μM) to test the additional 22,434 LC compounds. Actives were defined with the same criteria used for the binary data set PriA-SSB AS. Compounds with at least 35% inhibition that passed the PAINS filters were retested with the AS assay. Those with at least 35% inhibition in the AS retest were labeled as actives, creating the binary data set PriA-SSB prospective.

Because secondary screens and structural filters were used to define the active compounds, there was no single primary screen % inhibition threshold that separated the actives from the inactives. Some compounds exhibiting high % inhibition values were labeled as inactive because they did not satisfy the structural requirements or were not active in the secondary screen.

PubChem BioAssay. To help learn a better chemical representation with multi-task neural networks, we considered other screening contexts from which to transfer useful knowledge. We used a subset of 128 assays (AIDs) from the PubChem BioAssay (PCBA)[24] repository. This data set was used in previous work on multi-task neural networks.[14] This subset contained assays for which the assays were developed to probe a specific protein target and dose–response measurements were obtained for each compound (see **PCBA**

Query for other assay query filters). Potency and curve quality are factored into a PubChem Activity Score. Regardless of the assay, compounds with a PubChem Activity Score of 40 or greater (range 0–100) were assigned a PubChem Bioactivity outcome (label) of “Active”. Compounds with PubChem Activity Scores of 1–39 were labeled “Inconclusive”, and those with 0 were labeled “Inactive” (See **Data preprocessing** and Table A4.2).

PCBA Query. We downloaded additional high-throughput screening datasets from the PubChem BioAssay Database[24] using the following query: TotalSidCount from 10000, ActiveSidCount from 30, Chemi-cal, Confirmatory, Dose-Response, Target: Single, NCGC. These correspond to the search query:

(10000[TotalSidCount] : 1000000000[TotalSidCount]) AND (30[ActiveSidCount] : 1000000000[Ac-

tiveSidCount]) AND “small molecule"[filt] AND “doseresponse"[filt] AND 1[TargetCount] AND

“NCGC"[SourceName]. This follows the query from Ramsundar et al.[14]

Data preprocessing, Complex matrix composition. Each target dataset consists of compounds as rows. For each compound, it provides the biochemical features such as the fingerprint, SMILES string, interaction score, and activity label (binary or continuous). The first step was to extract the SMILES and activity label for each target and construct the data matrix for training. We used RDKit[25] for navigating and extracting information from these datasets and for generating the fingerprints. The second step was to merge the target matrices together into one consolidated matrix. We used an outer-join operation with the SMILES as the key. Given two matrices A and B with two columns, SMILES and target-activity, an outer-join operation will merge rows of A and B that have the same SMILES value into a new matrix M. If there is a row in A with SMILES values and no corresponding row in B with SMILES values, then the merge would yield a row in M with an empty target-activity for B. In the resulting matrix, each row is a compound and the columns are: SMILES, 1024-bit Morgan fingerprints, and a column for the activity outcome of each

target (5 columns for PriA-SSB AS, PriA-SSB FP, and RMI-FANCM FP and the associated % inhibition values and 128 columns for PCBA). As a result, we have two data matrices: PriA-SSB AS, PriA-SSB FP, and RMI-FANCM FP as well as PriA-SSB AS, PriA-SSB FP, and RMI-FANCM FP plus PCBA, on which we can train either single-task or multi-task learning methods. Merging all the targets introduces many empty cells for the activity outcome columns. For the distribution of active, inactive, and missing values for each target, refer to Table A4.2 We observed severe data imbalance; the ratio of positive to negative labels is very small. Inconclusive PCBA labels were 2 treated as missing. During model evaluation, we considered each task (column) separately and dropped the missing values.

Data preprocessing, Fold Splitting. The whole data set was split into 5 fixed folds for cross-validation. Label imbalance and the limited number of known active compounds is one of biggest challenges in virtual screening and must be accounted for during modeling. Stratified splitting is a way to divide data into folds while keeping the same active-to-inactive ratio for each label. For a single-target task, stratified splits can be implemented by combining folds after sampling each class of labels. But this procedure becomes more complicated in the multi-task setting. With a total of 131 binary tasks, each task has a set of molecules with activity outcomes that may or may not overlap across targets. After merging all molecules into one matrix, each row represents one molecule and each column represents one target. For each column (target), molecules can have missing, inactive, or active labels. Similarly, for each row (molecule), the molecule must have an active or inactive label for at least one of the 131 targets but can be missing for some of other targets. We construct this combined matrix of 131 targets using Algorithm A4.1 described below. We divide this matrix into 5 folds, while keeping the same data distribution.

Data preprocessing, Label Imbalance. PriA-SSB AS, PriA-SSB FP, and RMI-FANCM FP have only 79, 24, and 230 actives, respectively. To alleviate this class imbalance, one solution is to use a weighted schema. For single-task neural network models, we apply Equation A4.1.

Equation A4.1 $$weight_{(negative)} = 1,\ weight_{(positive)} = \frac{n}{p}$$

Where $weight_{(positive)}$ and $weight_{(negative)}$ are weight scalars for positive (active) and negative (inactive) compounds, respectively, and p and n represent the number of positive and negative samples on this target.

Similarly, we apply the weighted schema to multi-task models, defined as Equation A4.2.

Equation A4.2 $$weight_{(negative,i)} = t_i,\quad weight\ _{(positive,i)} = t_i \cdot \frac{n_i}{p_i}$$

Where $weight_{(positive,i)}$ and $weight_{(negative,i)}$ are weight scalars for positive and negative labels for the i^{th} target and p_i and n_i represent the number of positive and negative samples from the i^{th} target. t_i is defined as Equation A4.3 or A4.4.

Equation A4.3 $$t_i = \frac{\Sigma_i p_i}{p_i},$$ i^{th} target is in PCBA

Equation A4.4 $$t_i = \alpha \cdot \frac{\Sigma_i p_i}{p_i},$$ i^{th} target is in **PriA-SSB AS, PriA-SSB FP**, and **RMI-FANCM FP**

In the multi-task setting, we give different weights to each target, focusing more on the **PriA-SSB AS, PriA-SSB FP**, and **RMI-FANCM FP** targets and the PCBA targets that have fewer positive samples. We emphasize **PriA-SSB AS, PriA-SSB FP**, and **RMI-FANCM FP** by setting α = 100, and alleviate the data skewness among targets by the term $\frac{\Sigma_i p_i}{p_i}$.

Data preprocessing, Missing Label Imputation. For single-task machine learning models – random forest, single-task neural networks and IRV-training is done on molecules without missing labels (missing molecules are removed from the training set).

In the case of the multi-task neural networks, missing molecules were imputed as inactive. This was mainly due to Keras (at the time) not supporting sample-weighting for the multi-task case.

Compound Features. Ligand-based virtual screening methods require each chemical compound to be represented in a particular format as input to the model. We adopted two common representations. All of the ligand-based algorithms except the Long Short-Term Memory (LSTM) neural network used 1024-bit Morgan fingerprints[26] with radius 2 generated with RDKit version 2016.03.4.[25] These circular fingerprints are similar to ECFP4 fingerprints,[27] though with a slightly different implementation. For LSTM networks, we used the Simplified Molecular Input Line Entry System (SMILES) representation,[28] where the characters were treated as sequential features.

Virtual Screening Models. We selected a variety of existing virtual screening approaches for our benchmarks and prospective predictions. These included ligand-based supervised machine learning approaches, structure-based docking, and a chemical similarity baseline. Table A4.3 summarizes the types of training data used by each algorithm.

Ligand-Based Neural Networks. Deep learning is a machine learning approach that encompasses neural network models with multiple hidden layer architectures and the techniques for training these models. It represents the state of the art for many predictive tasks, which has generated extensive interest in deep learning for biomedical research, including virtual screening.[15,16] We evaluated multiple types of established neural network architectures for virtual screening.

Single-Task Neural Network (STNN). A single-task neural network (Figure A4.1a) makes a single prediction for a single target (also referred to as a task). We trained a separate model for each of the PriA-SSB AS, PriA-SSB FP, and RMI-FANCM FP data sets, taking each compound's Morgan fingerprint as the input features. We trained the neural networks using Keras[29] with the Theano backend.[30] The single-task neural networks were trained on each task to predict either the binary activity label in the classification setting

(STNN-C) or the continuous % inhibition in the regression setting (STNN-R). Because the STNN-R models were trained directly on the % inhibition, they do not depend on the PAINS filters. These neural networks used two hidden layers with 2000 hidden units each, Adam optimization,[31] 0.25 dropout rate, and other hyperparameters described in Tables A4.4 and A4.5.

Multi-Task Neural Network (MTNN). Multi-task neural networks make different predictions for multiple targets or tasks but share knowledge by training the first few hidden layers together. Each of our multi-task neural networks included one target task (PriA-SSB AS, PriA-SSB FP, or RMI-FANCM FP) and 128 tasks from PCBA. We only trained multi-task neural networks in the classification setting (MTNN-C). The MTNN-C models used two hidden layers with 2000 hidden units each, Adam optimization, 0.25 dropout rate, and other hyperparameters described in Table A4.4.

Hyperparameter Grid Search. During the hyperparameter sweeping stage, we trained models with all combinations of the hyperparameters in tables A4.4 through A4.8. For the neural networks, 80% of the 4 folds were used for training and20% for validation to select the best 2 models for each type of neural network (STNN-C, MTNN-C, STNN-R, and LSTM). For random forest, the first 3 folds were used for training and the fourth fold for validation to prune 108 models down to 8 models. In both cases, the goal was to prune the model search space before the cross-validation stage. IRV has one primary hyperparameter, the number of neighbors, so we did not need to prune the model search space before the cross-validation stage.

Based on related work[14] and our preliminary testing with the PCBA tasks, we did not consider neural networks with more than two hidden layers. Our cross-validation results confirmed that two hidden layer networks did not underfit the training data. Because random forests are resistant against overfitting as the number of trees grows,[32] we set the RF n_estimators hyperparameter to be as large as possible while still training reasonably quickly.

Model Name to Hyperparameter Mappings. We used alphabetic suffixes such as "_a" and "_b" to distinguish multiple versions of a model that use different hyperparameters. Only the best hyperparameter combinations from the hyperparameter sweeping stages were labeled with these suffixes. Hyperparameters that did not vary can be found in **Hyperparameter Grid Search** section.

Single-Task Atom-Level LSTM (LSTM). The LSTM is one of most prevalent recurrent neural network models,[33] which has been applied previously in virtual screening.[34] An LSTM assumes there exists a sequential pattern in the input string. We used a one hot encoding of the SMILES strings as input for the LSTM model. In a one hot encoding, each character in a SMILES string is replaced by a binary vector. The binary vector has one bit for each possible unique character in all SMILES strings. At each position in a SMILES string, the bit corresponding to the character at that position is set to 1, and all other bits are set to 0. We trained the LSTM model to predict the binary activity labels. The LSTM models used one or two hidden layers with 10 to 100 hidden units each, Adam optimization, 0.2 or 0.5 dropout rate, and other hyperparameters described in Table A4.6. The compounds in the cross-validation stage used SMILES generated by OpenEye Babel version 3.3. The compounds in the prospective screen were processed separately and used SMILES from RDKit version 2016.03.4.[25]

Influence Relevance Voter (IRV). IRV[35,36] is a hybrid between k-nearest neighbors and neural networks. Each compound's predicted value is a nonlinear combination of the similarity scores from its most closely related compounds in the training data set. We used Morgan fingerprints as the input and trained separate IRV models for each data set. The IRV models used 5 to 80 neighbors and other hyperparameters described in Table A4.7.

Ligand-Based Random Forest (RF). Random forests[32] are ensembles of decision trees that are often used as a baseline in virtual screening benchmarks.[37,38] We used scikit-learn[39] to train a random forest classifier for each binary label with Morgan fingerprints as features. The RF models used 4000 to 16,000 estimators,

1 to 1000 minimum samples at a leaf node, a bound on the maximum number of features, and other hyperparameters described in Table A4.8.

Protein–Ligand Docking, Target Preparation. Our structure-based VS approach involved the docking-based ranking of the LC library to the holo-form of PriA using the crystal structure (PDB: 4NL8),[40] in which it is bound to a C-terminal segment of an SSB protein. A missing loop in this structure was added from the apo-form (PBD: 4NL4), though this is not near the SSB binding site. The docking search space was limited to 8 Å from the coordinates of the cocrystallized SSB C-terminal tripeptide.

For RMI-FANCM, the RMI protein was built from both the A and B chains from the structure (PDB: 4DAY).[41] The docking search space was defined by the central five residues of the MM2 peptide (PDB: 4DAY chain C), Val-Thr-Phe-Asp-Leu, also with an 8 Å bounding box.

Protein–Ligand Docking, Compound Preparation. LC library compounds were assigned 3D coordinates and Merck Molecular Force Field partial charges using OpenEye OMEGA and Molcharge.[42] Compounds in the LC library with ambiguous stereochemistry were enumerated in all possibilities, and the best resulting docking score was retained for each.

Protein–Ligand Docking, Docking (Dock) and Consensus Docking (CD). We ran eight different docking programs and generated nine docking scores as a broad comparison to the ligand-based methods under consideration. The docking programs and names we use for their scores are AutoDock version 4.2.6[43] (Dock_ad4), Dock version 6.7[44] (Dock_dock6), FRED version 3.0.1[45] (Dock_fred), HYBRID version 3.0.1[45] (Dock_hybrid), PLANTS version 1.2[46] (Dock_plants), rDock version 2013.1[47] (Dock_rdocktot and Dock_rdockint), Smina version 1.1.2[48] (Dock_smina), and Surflex-Dock version 3.040[49] (Dock_surflex). In addition, we calculated consensus docking scores using three traditional approaches (CD_mean, CD_median, and CD_max) and two versions of the Boosting Consensus Score (CD_efr1_opt and CD_rocauc_opt).[50] The consensus docking methods were developed without any knowledge of the PriA-

SSB or RMI-FANCM assay data. Compounds with missing scores due to preparation or docking failures were not considered during evaluation.

Chemical Similarity Baseline. We introduced a compound ranking method based on chemical structure similarity to serve as a baseline for the ligand-based VS methods. The active compounds in the training set were used as seeds for similarity searching through all test set compounds. The test set compounds were ranked by their maximum Tanimoto similarity to any of the training set actives with MayaChemTools[51] using Morgan fingerprints from RDKit version 2013.09.1. Unlike the ligand-based machine learning algorithms, the similarity baseline does not consider inactive compounds in the training set.

In addition, all compounds were clustered by two separate approaches to describe chemical series. Chemical similarity-based hierarchical clusters on Morgan fingerprints using Ward's clustering are described as SIM. Maximum common substructure clusters, used to group molecules with similar scaffolds, are described as MCS. JKlustor was used for both types of clustering (JChem version 17.26.0, ChemAxon).

Evaluation Metrics. Given our goal of developing VS methods that enable very small, cost-effective, productive screens, we considered how evaluation metrics weight early active retrieval. All of the VS algorithms produce a ranked list of compounds, where compounds are ordered by the probability of being active, the continuous predicted % inhibition, the docking score, or a comparable output value. For a ranked list of compounds, we can threshold the ranked list and consider all compounds above the threshold as positive (active) predictions and those below the threshold as negative (inactive). Classification models output class probabilities. Regression models, docking, and the similarity baseline output different types of continuous scores. Thresholding on the compound rank is equivalent to thresholding on the class probability or continuous score because for each rank there is a corresponding

probability or score. By comparing those predictions to the experimentally observed activity, we can compute true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions for the ranked list at that threshold. We explored several options for summarizing how well each algorithm ranks the known active compounds. Because most of the compounds have only single-replicate measurements of % inhibition, we focus on evaluating active versus inactive compounds instead of correlation with the % inhibition.

The area under the receiver operating characteristic curve (AUC[ROC]) has been recommended for virtual screening because it is robust, interpretable, and does not depend on user-defined parameters.[52] The ROC curve plots the relationship between true positive rate (TPR, also known as sensitivity or recall) and false positive rate (FPR, equivalent to 1 – specificity), which are defined in equation A4.5 and equation A4.6, respectively. As the FPR goes to 100%, all ROC curves converge, whereas early active retrieval (a more meaningful characteristic of VS performance) can be assessed in the low FPR region of the ROC curve, which exhibits greater variability across VS methods. Thus, we also considered the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC).[53] It emphasizes the early part of the ROC curve through a scaling function α, which we set to 10 for our purposes of early enrichment up to 20%. We used the BEDROC implementation from the CROC Python package.[54]

Equation A4.5 $$TPR = \frac{TP}{TP+FN}$$

Area under the precision-recall curve (AUC[PR]) is another common metric (equation A4.2). AUC[PR] has an advantage over AUC[ROC] for summarizing classifier performance when the class labels are highly skewed, as in virtual screening where there are few active compounds in a typical library. AUC[PR] evaluates a classifier's ability to retrieve actives (recall) and which of the predicted actives are correctly classified (precision) as the prediction threshold varies. We used the PRROC R package's "auc.integral"[55] to compute AUC[PR].

EquationA4.6 $$FPR = \frac{FP}{FP+TN}$$

Another VS metric is enrichment factor (EF), which is the ratio between the number of actives found in a prioritized subset of compounds versus the expected number of actives in a random subset of same size. In other words, it assesses how much better the VS method performs over random compound selection. Let R ∈ [0%,100%] be a predefined fraction of the compounds from the total library of compounds screened.

Equation A4.7 $$EF_R = \frac{\#\ active\ in\ top\ R\ ranked\ compounds}{\#\ active\ in\ entire\ library\ X\ R}$$

Equation A4.8 $$EF_{max,R} = \frac{min\{\#active, total\ \#compounds\ X\ R\}}{\#\ actives\ in\ entire\ library\ X\ R}$$

$EF_{max,R}$ represents the maximum enrichment factor possible at R. Difficulty arises when interpreting EF scores because they vary with the data set and threshold R. We defined the normalized enrichment factor (NEF) as

Equation A4.9 $$NER_R = \frac{EF_R}{EF_{max,R}}$$

Because NEF_R ∈ [0,1], it is easier to compare performance across data sets and thresholds. Here, 1.0 is the perfect NEF. Furthermore, we can create an NEF curve as NEF_R versus R ∈ [0%,100%] and compute the area under that curve to obtain AUC[NEF] ∈ [0,1]. However, most models tend to exhibit similar late enrichment behavior. We are typically interested in early enrichment behavior, so we computed AUC[NEF] using R ∈ [0%,20%].

Finally, we considered the metric n_{hits}, which is simply the number of actives found in a selected number of tested compounds (e.g., how many hits or actives were found in 250 tested compounds). This metric represents the typical desired utility of a screening process: retrieve as many actives as possible in

the selected number of tested compounds (denoted as n_{tests}). We compared n_{hits} at various n_{tests} to the different evaluation metrics to identify which metrics best mimic the n_{hits} utility.

Pipeline. Our virtual screening workflow contains three stages: (1) Tune hyperparameters in order to prune the model search space. (2) Train, evaluate, and compare models with cross-validation to select the best models. (3) Assess the best models' ability to prospectively identify active compounds in a new set.

In contrast to most other virtual screening studies, the experimental screen was not conducted until after all models were trained and evaluated in the cross-validation stage (Figure A4.2). For the first two stages, we first split the PriA-SSB AS, PriA-SSB FP, and RMI-FANCM FP data sets into five stratified folds as described the **Data Preprocessing** section.

Hyperparameter Sweeping Stage. Hyperparameters are model configurations or settings that are set by an expert as opposed to the weights or parameters that are learned or fit during model training. For most of the ligand-based machine learning models, the hyperparameter space was too large for exhaustive searches using the full data set. Therefore, we applied a grid search on a predefined set of hyperparameters in a smaller data set and pruned those that performed poorly. We performed a single iteration of training on the first four folds of PriA-SSB AS to avoid overfitting. The hyperparameters considered are listed in Tables A4.4 – A4.8.

Cross-Validation Stage. To identify which VS algorithms are likely to have the best performance in a prospective screen, we applied a traditional cross-validation training strategy on data sets PriA-SSB AS, PriA-SSB FP, and RMI-FANCM FP after reducing the hyperparameter combinations to consider. Selecting the best model is nontrivial. Ideally, the best model would have dominant performance on all evaluation metrics, but this is rarely observed with existing models. Each evaluation metric prioritizes different performance characteristics. Our cross-validation results illustrate which models consistently perform

well over different metrics, the correspondence of metrics relative to a desired utility (n_{hits}), and how to choose models and evaluation metrics in order to successfully identify active compounds in a prospective screen.

Cross-validation is commonly used to avoid overfitting when there are few training samples. We split the training data into five folds: four folds for training and one for testing. Models like RF and IRV that do not require a hold-out data set for early stopping used four folds for training. The neural networks perform early stopping based on a hold-out set, so we iteratively selected one of the four training data folds for this purpose. This led to a nested cross-validation with 5 × 4 = 20 trained neural networks.

Prospective Screening Stage. Our prospective screen used a library of 22,434 new LC compounds that were not present in the PriA-SSB AS training set. We used each VS model to prioritize 250 of these compounds that are most likely to be active. This emulates virtual screening on much larger compound libraries, in which only a small fraction of all computationally scored compounds can be tested experimentally. When models assigned the same score to multiple compounds, we broke ties arbitrarily to obtain exactly 250 compounds.

After finalizing the models' predictions, we screened all 22,434 compounds in the wet lab and assigned actives based on a 35% inhibition threshold and structural filters (the PriA-SSB prospective data set). Finally, we evaluated how many of the experimental actives each VS method identified in its top 250 predictions, the number of distinct chemical clusters recovered, and the number of active compounds that were not in the top 250 predictions from any of the VS algorithms. The prospective screen allowed us to assess how well the cross-validation results generalized to new compounds and further verified our conclusions from the retrospective cross-validation tests.

Data and Software Availability. Code implementing our ligand-based virtual screening algorithms is available at https://github.com/gitter-lab/pria_lifechem and archived on Zenodo (DOI: 10.5281/zenodo.1257673). This GitHub repository also contains additional Jupyter notebooks to reproduce the visualizations and analyses. Our new PriA-SSB HTS data are available on PubChem (PubChem AID:1272365) along with the existing RMI-FANCM HTS data (PubChem AID:1159607). A formatted version of this data set for training virtual screening algorithms is available on Zenodo (DOI: 10.5281/zenodo.1257462).

Results

Cross-Validation Results. In the cross-validation stage, we assessed 35 models: eight neural networks (STNN-C (Table A4.9), STNN-R (Table A4.10), MTNN-C (Table A4.11), and LSTM (Table A4.12)), five IRV (Table A4.13), eight RF (Table A4.14), and 14 from docking (Dock) or consensus docking (CD). When there are multiple versions of a model that use different hyperparameters, we distinguish them with alphabetic suffixes such as "_a" and "_b". Tables A4.9-A4.14 describe the hyperparameters associated with these suffixes. We highlight the PriA-SSB AS data set as a representative example, but the VS workflow is applicable for all tasks.

Comparing Virtual Screening Algorithms. We tested all 35 models on three data sets, and the results for four evaluation metrics on the PriA-SSB AS data set are shown in Figure A4.3. Figures A4.4 – A4.11 contains the results for PriA-SSB FP and RMI-FANCM FP. The PriA-SSB AS performance using AUC[ROC] was comparable for many models. All models except LSTM, some IRV models, and docking were above 0.8 AUC[ROC]. Some of the other evaluation metrics better stratify the ligand-based VS methods. Random forest was the best model, especially for the most-relevant metrics that prioritize early enrichment. We also ran the chemical similarity-based method for PriA-SSB AS and confirmed that random forest outperformed this simple baseline.

Random forest was again the best overall method for the RMI-FANCM FP data set (Figures A4.8-A4.11). On the PriA-SSB FP data set, STNN-R achieved the highest scores over the majority of the metrics (Figures A4.4-Figure A4.7). The other types of VS models were effectively tied for most metrics.

Evaluation Metrics. Given a fixed evaluation metric, we could compare two models with a t test to assess if one statistically outperforms the other. However, we needed to make such comparisons repeatedly between each pair of models and required a statistical test that accounts for multiple hypothesis testing.

Due to unequal variances and sample sizes (Figure A4.3), we used Dunnett's modified Tukey–Kramer test (DTK)[56,57] for pairwise comparison to assess whether the mean metric scores of two models were significantly different. Using DTK results for each metric, we scored each model based on how many times it attained a statistically significantly better result than other models. For most metric-target pairs, many models have the same rank because DTK does not report a significant difference.

In a prospective screen, our goal is to maximize the number of active compounds identified by a VS algorithm given a fixed budget (number of predictions). We wanted to determine which of the VS evaluation metrics best aligns with n_{hits}. Thus, we compared the model ranking induced by each metric with the model ranking induced by n_{hits} for a varying number of tests.

To score the evaluation metrics, we used Spearman's rank correlation coefficient based on the model rankings induced by the metric of concern versus n_{hits} at a specific n_{tests}. We then ranked the metrics based on their correlation with n_{hits} (Tables A4.15-A4.20). The metric ranking varies depending on n_{tests} and the target. Some metrics overtake one another as we increase n_{tests}. For PriA-SSB AS, NEF_R consistently placed in the top ranking correlations when R coincided with n_{tests}. This is evident when we focus on a single metric and see the top ranking metrics for $n_{tests} \in [100,250,500,1000,2500]$. Only for a large enough n_{tests} do metrics like AUC[ROC] that evaluate the complete ranked list become comparable. This suggests that if we know a priori how many new compounds we can afford to screen, then NEF_R at a suitable R is a viable metric for choosing a VS algorithm during cross-validation in the hopes of maximizing n_{hits}.

Selecting the Best Model. Based on these results, we selected the VS screening models that are most likely to generalize to new compounds and identify actives in our experimental screen of 22,434 new compounds. We focused on PriA-SSB for the prospective screen using models trained on PriA-SSB AS because the assay was more readily available for us to generate data for the new compounds.

Table A4.21 compares model selection based on evaluation metric means alone versus the DTK+Mean approach for multiple evaluation metrics on the three tasks. The complete model rankings for means only and DTK+Mean can be found online. DTK+Mean ranks models by statistical significance and uses the mean value only for tie-breaking. Both strategies selected the same models for a fixed evaluation metric, except for AUC[PR] on all three tasks (Table A4.21). This is mainly due to DTK not detecting statistically significant differences among the models' evaluation scores, so tie-breaking by means selected the same models as ranking by means. Recall that PriA-SSB FP has fewer actives than PriA-SSB AS and RMI-FANCM FP (Table A4.1). Similar RF and STNN-C models were selected for PriA-SSB AS and RMI-FANCM FP. However, PriA-SSB FP prioritized STNN-R models exclusively.

In our prospective screen, each model prioritizes 250 top-ranked compounds, approximately 1% of the new LC library. In this setting where each model has a fixed budget for the predicted compounds, NEF_R is a suitable metric. Therefore, we used $NEF_{1\%}$ with DTK+Means to choose the best models from each class. The best-in-class models were RandomForest_h, SingleClassification_a, SingleRegression_b, MultiClassification_b, LSTM_b, IRV_d, and ConsensusDocking_efr1_opt, with RandomForest_h being the strongest model overall.

Prospective Screening Results. After selecting the best model from each class based on cross-validation and the $NEF_{1\%}$ metric, we retrained the models on all 72,423 LC compounds to predict PriA-SSB inhibition using the same types of data shown in Table A4.3. This provided a single version of each model instead of one for each cross-validation fold. All models then ranked 22,434 new LC compounds that were provided without activity labels. We selected the top 250 ranked new compounds from each model. Then, we experimentally screened all 22,434 new compounds to assess PriA-SSB % inhibition and defined actives based on a 35% inhibition threshold and PAINS filters. The new binary data set PriA-SSB prospective contained 54 actives.

Table A4.22 presents how many of the 54 actives were identified by each best-in-class virtual screening method and the chemical structure similarity baseline. For context, randomly selecting 250 compounds from the PriA-SSB prospective data set is expected to identify less than one active based on the overall hit rate. The VS models' PriA-SSB prospective performance for the other evaluation metrics can be found online.

Table A4.22 also lists the number of distinct chemical clusters identified by each method, with the goal of identifying as many diverse active compounds as possible. The 22,434 compounds form 124 SIM and 714 MCS clusters or chemical series. Of these, the 54 experimental actives represent 27 SIM and 35 MCS clusters. Commonly, virtual screening is followed by a medicinal chemistry effort that would be expected to identify other members of these clusters.

In general, the number of distinct chemical clusters captured in the top 250 predictions is correlated with the number of actives (Table A4.22), meaning that the methods selected structurally diverse hits. The similarity baseline identified compounds from roughly half of the SIM or MCS clusters. With the exception of docking, each of the methods in Table A4.22 found at least one cluster not present in the baseline. The machine learning techniques are not limited to finding only the chemotypes that are present in the training set (Figures A4.12, A4.13).

The ligand-based VS methods recovered many of the same actives as the chemical similarity baseline, but they also found actives that were missed by the baseline (Figure A4.14). There was a group of 11 active compounds that were identified by most ligand-based methods, including the baseline model. The compounds identified were not the most potent, either within their cluster or overall, nor did any of the methods exhibit any correspondence between the number of compounds identified from a cluster and their potency.

The similarity baseline included one active compound that was excluded from the top 250 compounds from RF (Figure A4.14), but RF recovered a different member from this active compound's SIM cluster (Figures A4.12, A4.13). Only the RF model recovered more active compounds in its top 250 predictions than the chemical similarity baseline, including two unique actives not identified by any other model. Therefore, cross-validation with $NEF_{1\%}$ as the metric successfully identified the best PriA-SSB model before the prospective screen.

Trained Models are Target Specific. As a control, we retrained the best RF model on randomized data to confirm that its strong prospective performance was due to meaningful detected patterns among the active compounds instead of biases in the data set. Similar to y-scrambling or y-randomization,[59] we randomly permuted the binary activity labels in the PriA-SSB AS data set, retrained the RF_h model on the randomized data, and evaluated the classifier on the PriA-SSB prospective data set. This procedure was repeated 100 times with different y-scrambling performed each time. The number of active compounds in the top 250 predictions for these 100 runs is summarized in Figure A4.15. The mean number of actives was 0.83, and 55 of the runs found zero actives. The best y-scrambled run found only 10 actives, far less than the 37 actives when RF_h was trained on the real data.

In addition, we assessed the performance of all models trained on RMI-FANCM FP instead of PriA-SSB AS for making PriA-SSB prospective predictions on the new 22,434 compounds. As expected, the RMI-FANCM FP models perform poorly on PriA-SSB prospective (Table A4.23), indicating that the best PriA-SSB AS models have learned compound properties that are specific to PriA-SSB.

Discussion

We followed a VS pipeline with the goal of maximizing the number of active compounds identified in a prospective screen with a limited number of predictions. From an initial pool of structure-based and ligand-based models, we pruned models in a hyperparameter search stage and conducted cross-validation with multiple evaluation metrics. We used DTK+Means with the $NEF_{1\%}$ metric to select the best models based on the cross-validation results and experimentally evaluated their top 250 prospective predictions from a new library of 22,434 compounds. The single best model from our pool, which was RandomForest_h for PriA-SSB AS, was also the top performing model on PriA-SSB prospective. Therefore, our overall pipeline successfully identified the best prospective model.

Metrics like AUC[ROC] can compare models in general, regardless of cost or other additional constraints.[52] However, for virtual screening in practice, one typically only experimentally tests a small fraction of all available compounds. In this setting, metrics like EF that capture early enrichment are preferable. In our prospective screen, STNN-R_a had higher AUC[ROC] than RF_h (Tables A4.23 and A4.24), but the random forest found eight more active compounds in its top 250 predictions (Table A4.25). Our study suggests that EF_R, or its normalized version NEF_R, are the preferred metrics for identifying the best target-specific virtual screening method that maximizes n_{hits} when there is a budget for experimental testing. Other metrics like AUC[ROC] or AUC[PR], which is more appropriate for problems where the inactive compounds far outnumber the actives,[60] may still be reasonable for benchmarking virtual screening methods on large existing data sets where the entire ranked list of compounds is evaluated.[38]

Some recent studies[3,37,61] reported that deep learning models substantially outperform traditional supervised learning approaches, including random forests. Our finding that a random forest model was the most accurate in both cross-validations, and our prospective screen does not refute those results. Rather, it reinforces that the ideal virtual screening method can depend on the training data available,

target attributes, and other factors. Therefore, careful target-specific cross-validation is important to optimize prospective performance. One cannot assume that deep learning models will be dominant for all targets and all virtual screening scenarios. We also recommend hyperparameter exploration for all models, including traditional supervised learning methods. For example, our best random forest model contained 8000 estimators, whereas a previous benchmark considered at most 50 estimators.[3]

Ramsundar et al.[14] showed that performance improved in multi-task neural networks as they added more training compounds and tasks. Furthermore, the degree of improvement varied across the data sets and was moderately correlated with the number of shared active compounds among the targets within a single data set. Task-relatedness also affects the success of multi-task learning but is difficult to quantify.[62,63] We observed that PriA-SSB AS, PriA-SSB FP, and RMI-FANCM FP have no shared actives with any of the PCBA tasks, and multi-task neural networks were not substantially better than single-task neural networks in PriA-SSB AS cross-validation (Figure A4.3). The MTNN-C model outperformed the STNN-C model in the prospective evaluation (Table A4.22), possibly because multi-task learning can help prevent overfitting,[64] but was still considerably worse than the random forest. Multi-task random forests can also be constructed by using multi-task decision trees as the base learner.[65] However, these methods have not been used widely in the context of virtual screening.

We focused on well-established machine learning models instead of more recent deep learning models, such as graph-based neural networks.[38,66-69] This is because our main goal was to investigate the virtual screening principles for choosing the best model for a specific task (PriA-SSB AS) in a practical setting instead of broadly benchmarking virtual screening algorithms. In addition, a recent benchmark showed that conventional methods outperformed graph-based methods on most biophysics data sets.[38]

Consensus docking[50] failed to recover any actives in the PriA-SSB prospective data set, even though some of the individual docking programs did. Specifically for the PriA-SSB protein–protein

interaction, docking is limited by the large, flat nature of the binding site. Many compounds that are inactive in the experimental screen have good scores and reasonable binding poses (per visual inspection) but fail to interrupt necessary specific interactions in the protein–protein interface. This will the limit overall performance by pushing true actives down the ranked list.

Our results are not intended to make general conclusions about the performance of ligand-based versus structure-based models. We use docking only for comparison to traditional structure-based VS methods and do not evaluate more sophisticated structure-based scoring functions. In addition, the individual docking and consensus docking methods do not train and optimize hyperparameters on the target-specific HTS screening data, whereas the ligand-based machine learning methods do. A more direct comparison would be to retrain a custom structure-based model or consensus scoring function to include the initial HTS data, though this effort is out of scope for this study. In addition, there are computational trade-offs between docking and ligand-based machine learning approaches. The machine learning models require substantial training time to select hyperparameters and fit models, but the trained models make predictions on new compounds very quickly. The docking programs take more time to score each new compound but have the advantage of not requiring training compounds.

The random forest model performed the best overall, but there were six active compounds identified by the other methods that the random forest missed (Figure A4.14). The single-task regression neural network recovered five of those six as well as unique active compound clusters (Figures A4.12 and A4.13). In addition, this regression model performed the best on PriA-SSB FP during cross-validation (Table A4.21), possibly because there are fewer binary actives in this data set. In future work, we will explore whether ensembling classification and regression models, potentially in combination with structure-based VS algorithms, can further improve accuracy.

We emphasize our prospective performance on the new LC library, which minimizes the biases that make evaluation with retrospective benchmarks challenging.[70] There are many sources of experimental error in HTS, and the active compounds in the prospective evaluation must still be interpreted conservatively. However, a VS algorithm that can prioritize compounds with high % inhibition in primary and retest screens is valuable for further compound optimization even if not all of the actives confirm experimentally. Our study provides guidelines for selecting a target-specific VS model and complements other practical recommendations for VS pertaining to hit identification, validation, and filtering,[71] as well as avoiding common pitfalls.[72] Having established that our best virtual screening model successfully prioritized new active compounds in the LC library, another future direction will be to test prospective performance on much larger, more diverse chemical libraries.

Acknowledgements

We acknowledge GPU hardware from NVIDIA, computing resources from the University of Wisconsin-Madison Center for High Throughput Computing, and funding from the Center for Predictive Computational Phenotyping NIH U54 AI117924, the University of Wisconsin Carbone Cancer Center Support Grant NIH P30 CA014520, and the Morgridge Institute for Research. Additional support for this research was provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. We are grateful for the assistance and feedback from Chengpeng Wang, Haozhen Wu, and many members of the Center for High Throughput Computing and the Gitter lab. We thank Julio Lopes for alerting us about duplicate compounds in a preliminary version of the PriA-SSB prospective data set.

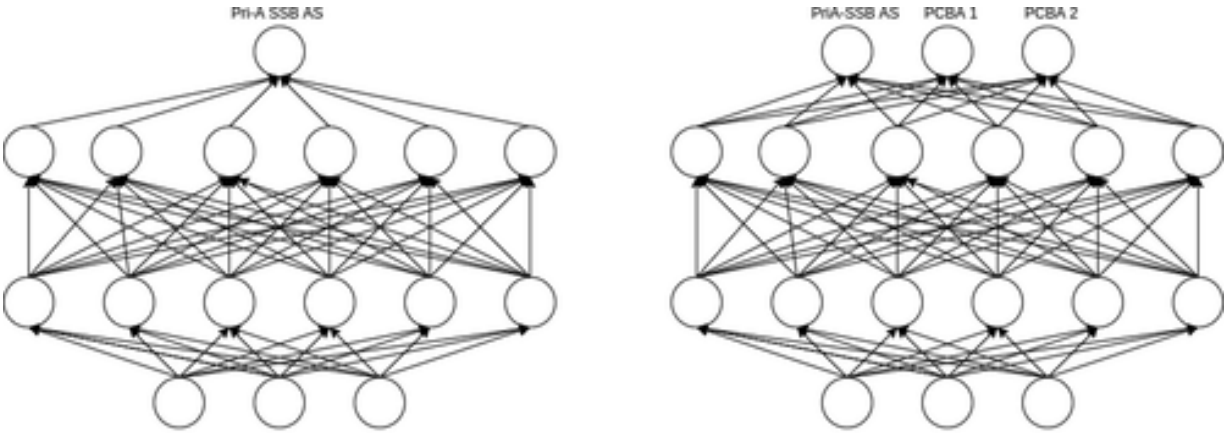


Figure A4.1. Neural network structures. The neural networks map the input features (e.g., fingerprints) in the input (bottom) layer to intermediate chemical representations in the hidden (middle) layers and finally to the output (top) layer, which makes either continuous or binary predictions. Panel (a) has only one unit in the output layer. Panel (b) has multiple units in the output layer representing different targets, one for our new target of interest and the others for PCBA targets.

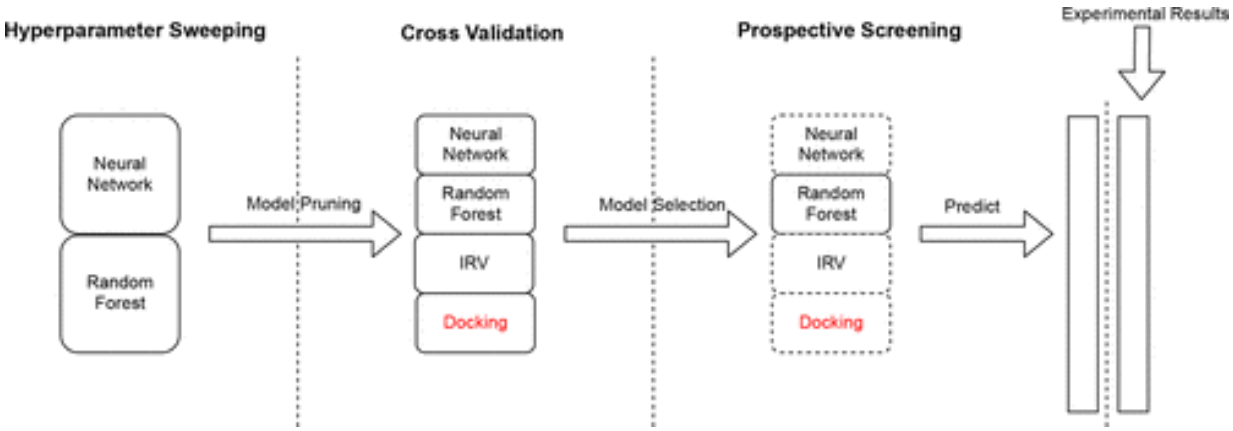


Figure A4.2. Study workflow. Initially, 258 neural network and random forest models were evaluated to eliminate poorly performing hyperparameter combinations. The models with the best hyperparameters advanced to cross-validation along with IRV and docking-based methods for a total of 35 models. Cross-validation identified a random forest as the best overall model. The VS methods and similarity baseline then predicted active compounds in the PriA-SSB prospective data set. After the predictions were finalized, we experimentally screened the compounds to evaluate the predictions. Black text denotes ligand-based machine learning models. Red text denotes docking-based models, which did not train on the target-specific HTS data.

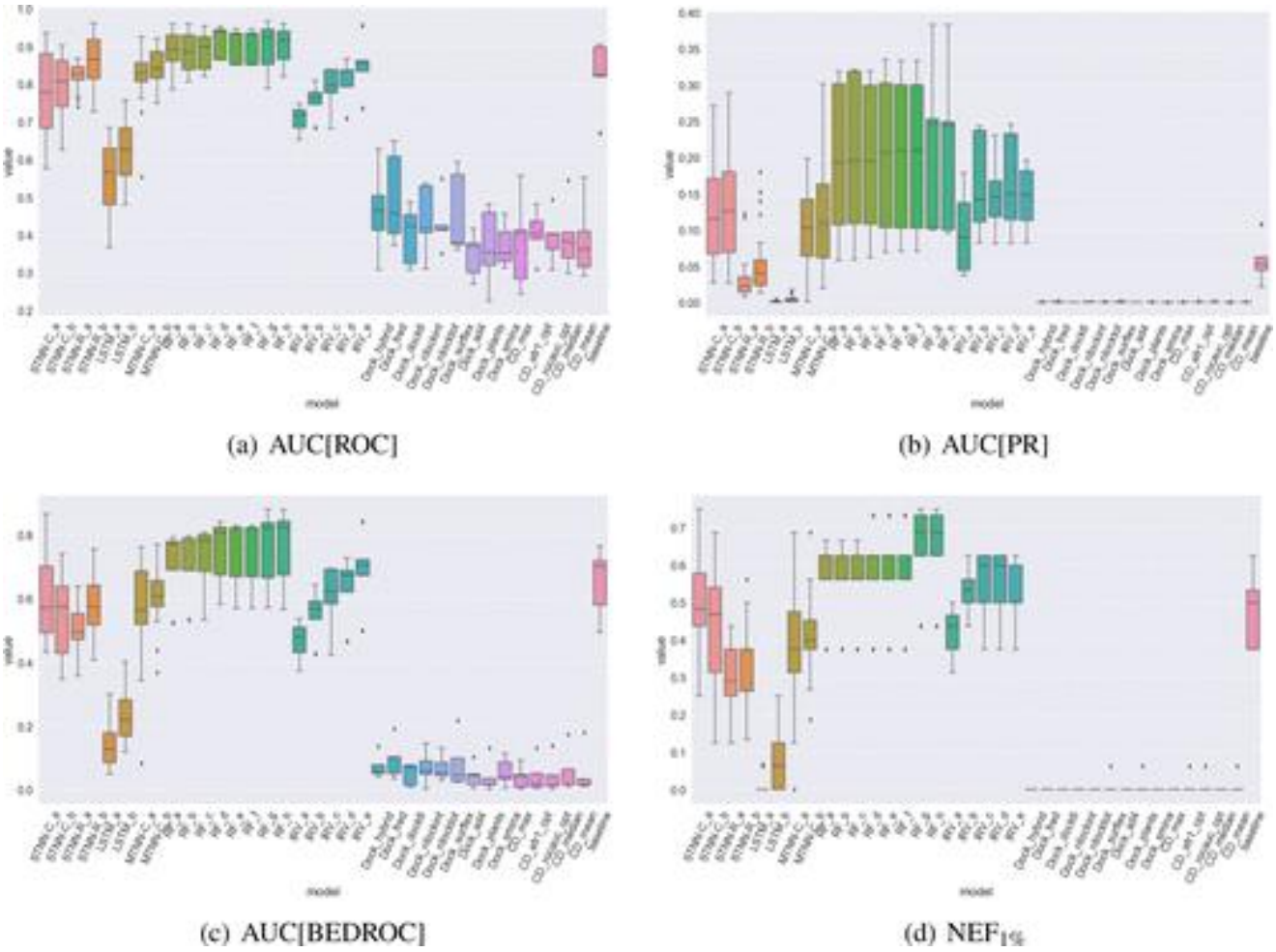


Figure A4.3. Evaluation metric distributions on PriA-SSB AS over the cross-validation folds. The metrics are (a) AUC[ROC], (b) AUC[PR], (c) AUC[BEDROC], and (d) $NEF_{1\%}$ as described in Evaluation Metrics. Unlike the ligand-based models, the docking methods do not train on the PriA-SSB AS training folds and are applied directly to the test fold during cross-validation (see Discussion).

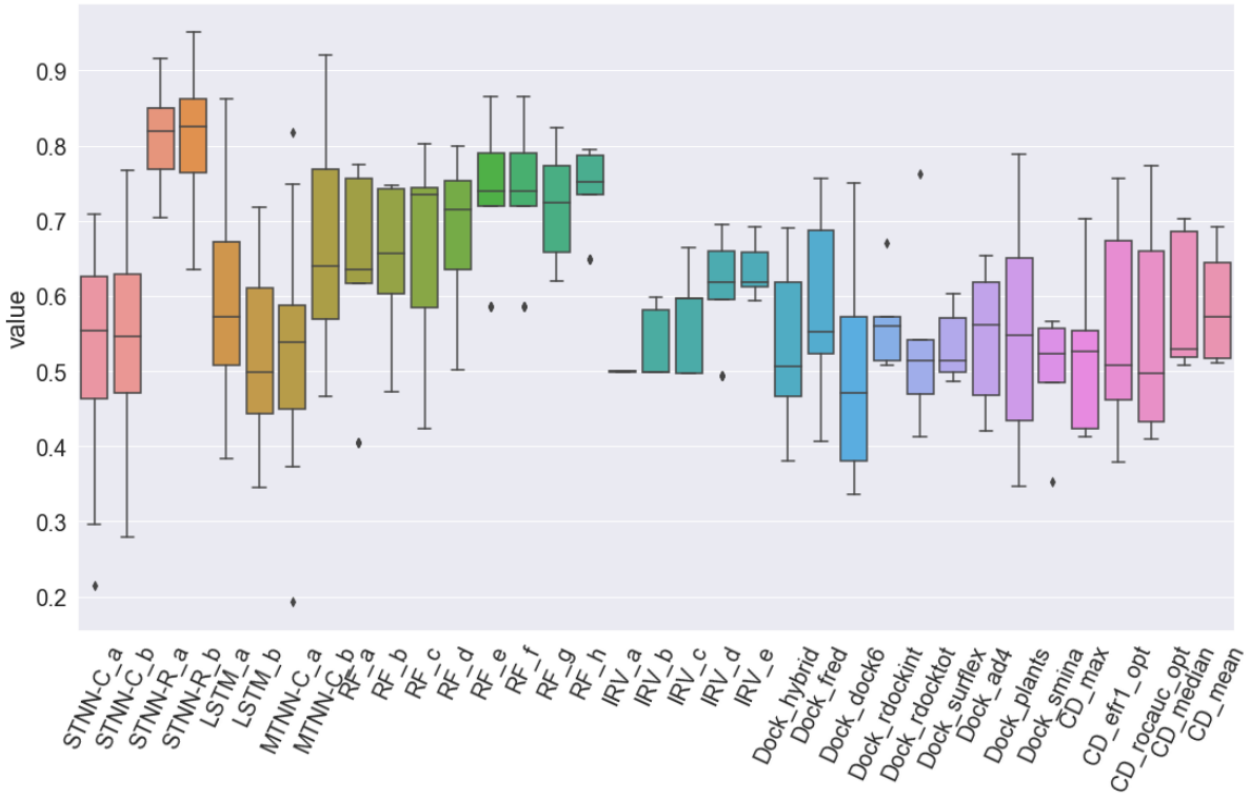


Figure A4.4. Cross-validation performance on PriA-SSB FP with AUC(ROC).

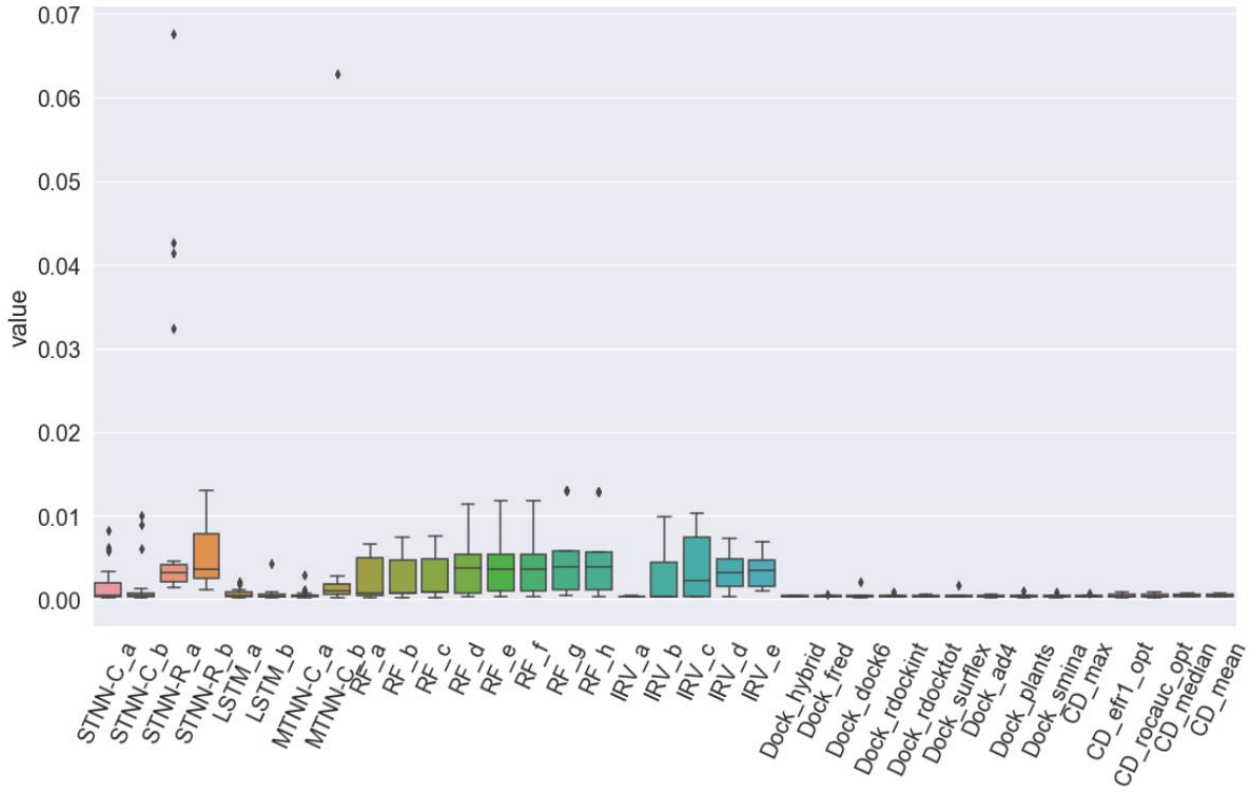


Figure A4.5. Cross-validation performance on PriA-SSB FP with AUC[PR].

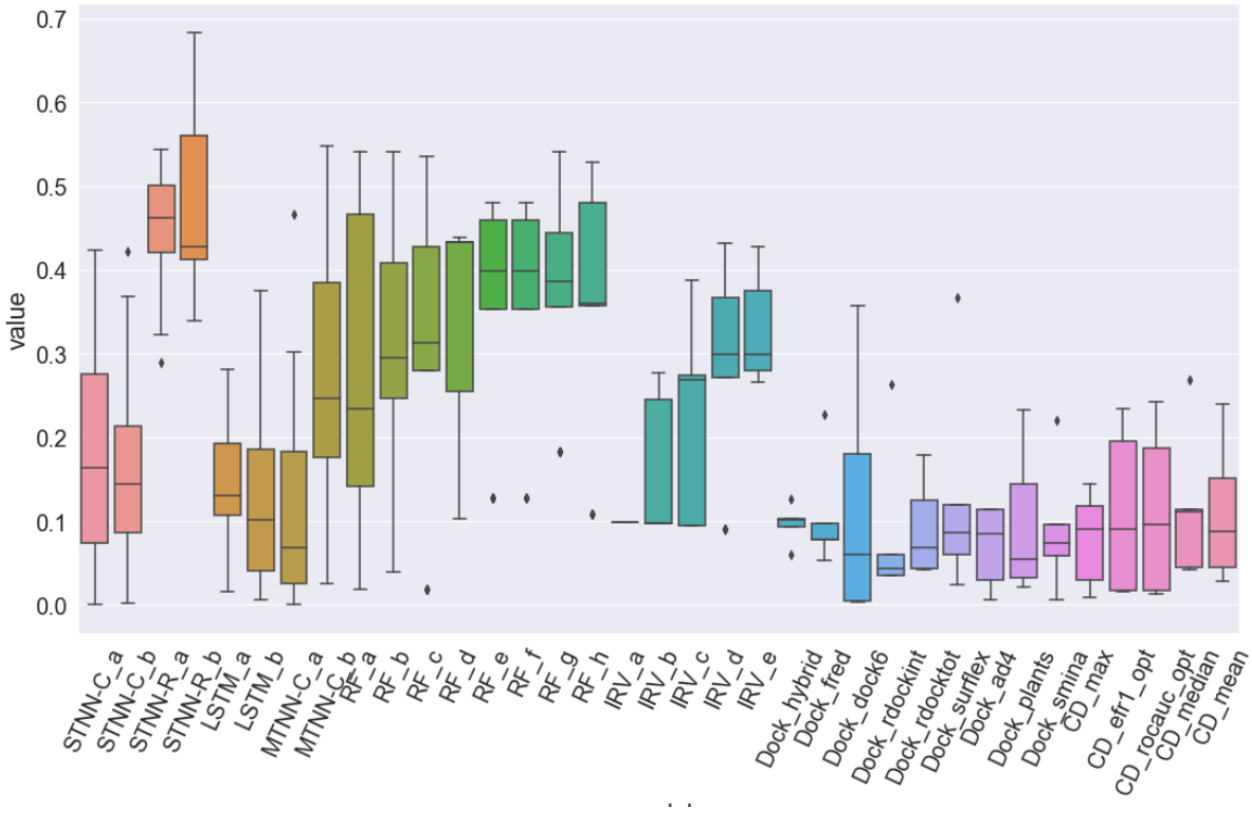


Figure A4.6. Cross-validation performance on PriA-SSB FP with AUC[BEDROC].

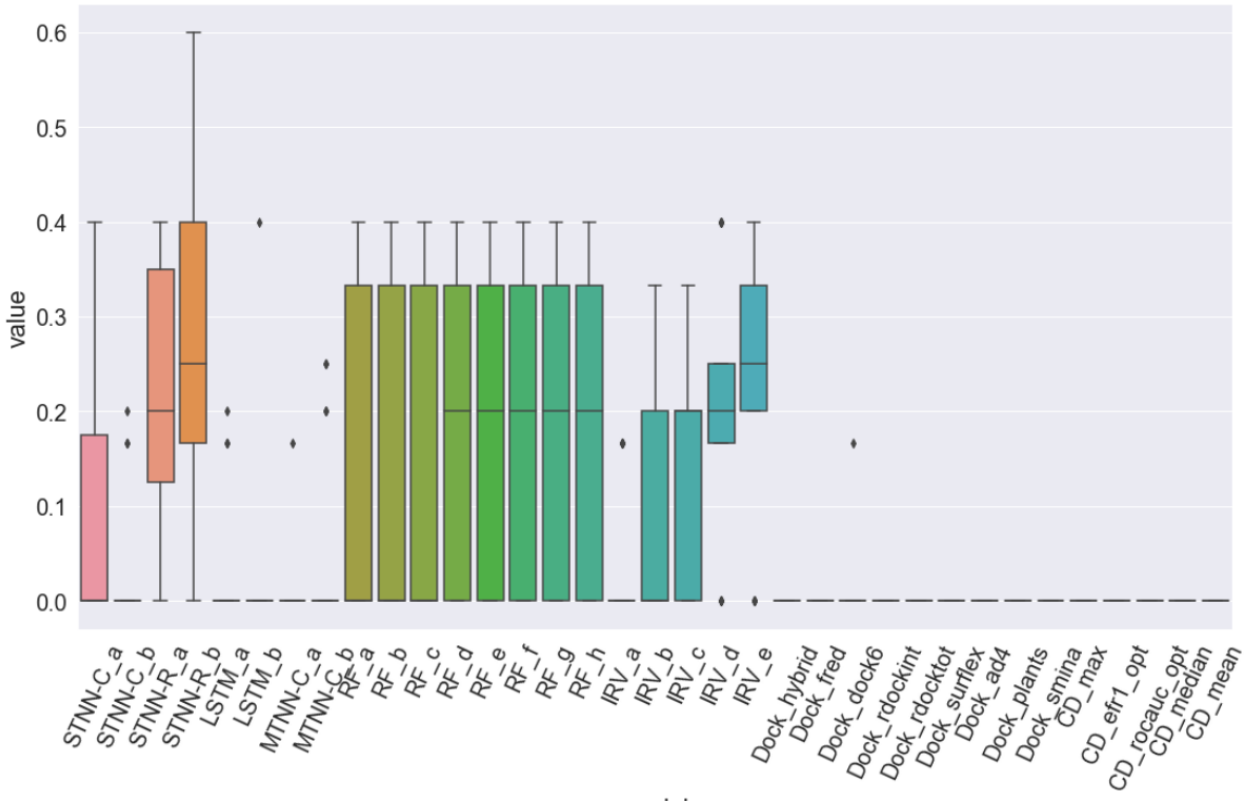


Figure A4.7. Cross-validation performance on PriA-SSB FP with $NEF_{1\%}$.

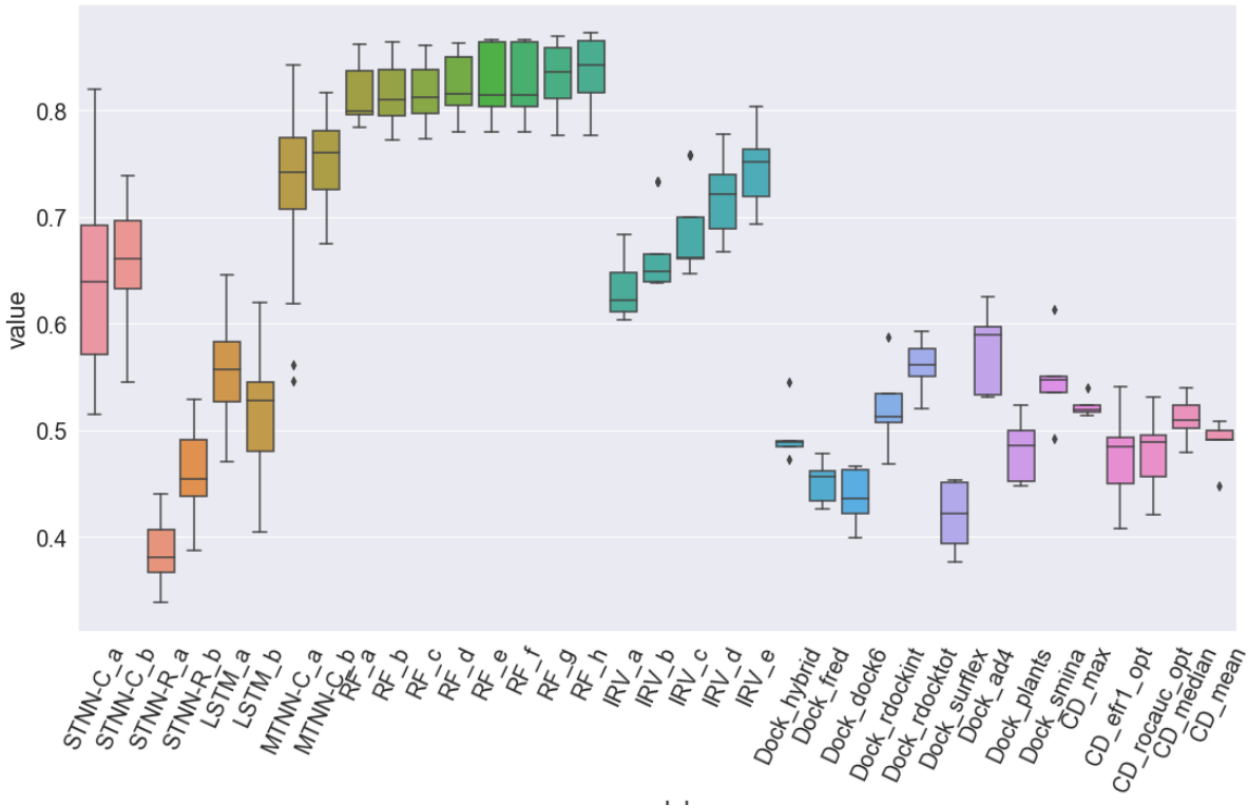


Figure A4.8. Cross-validation performance on RMI-FANCM FP with AUC[ROC].

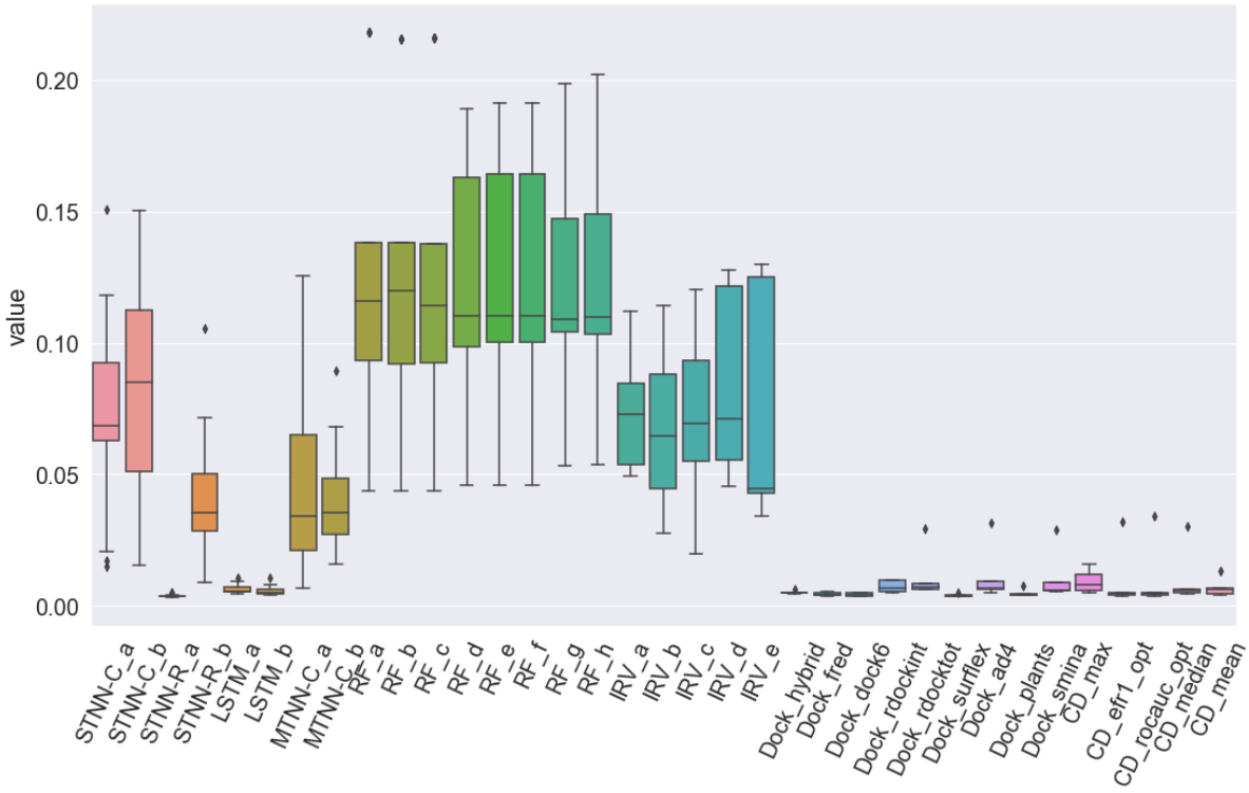


Figure A4.9. Cross-validation performance on RMI-FANCM with AUC[PR].

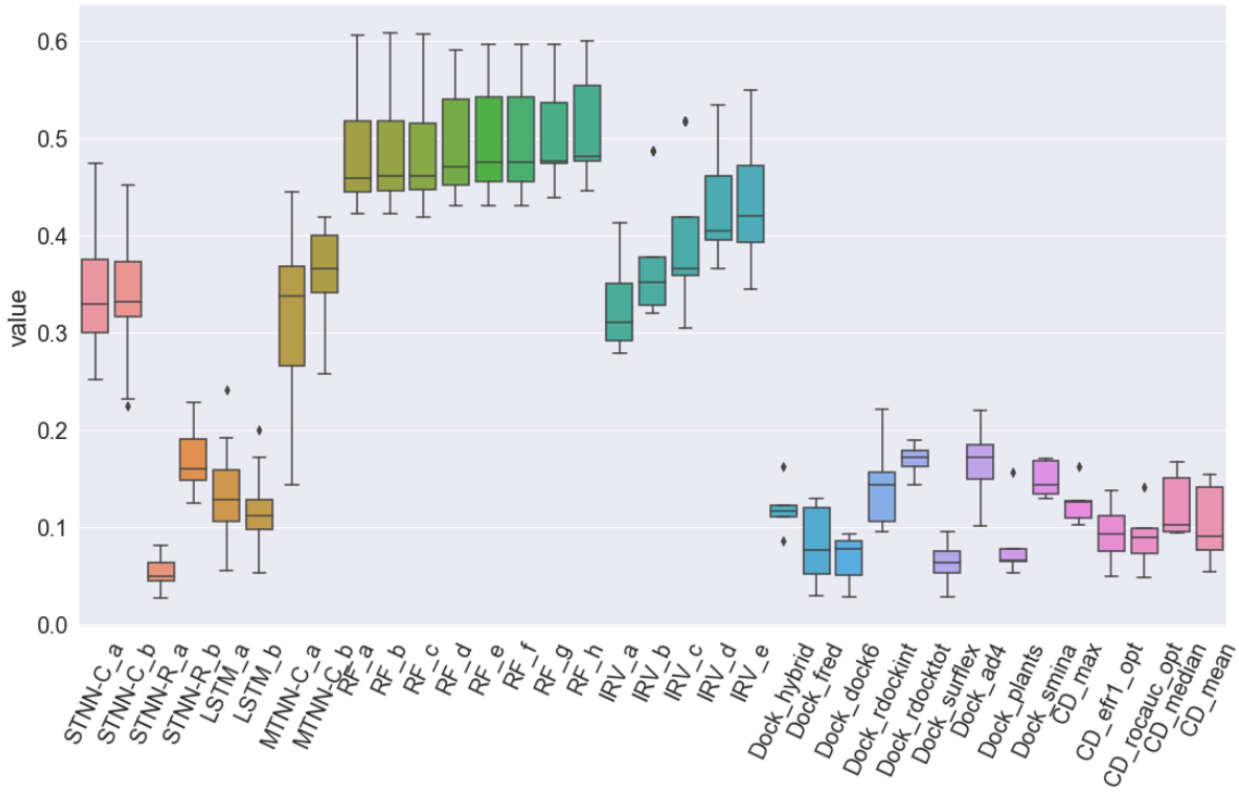


Figure A4.10. Cross-validation performance on RMI-FANCM with AUC[BEDROC].

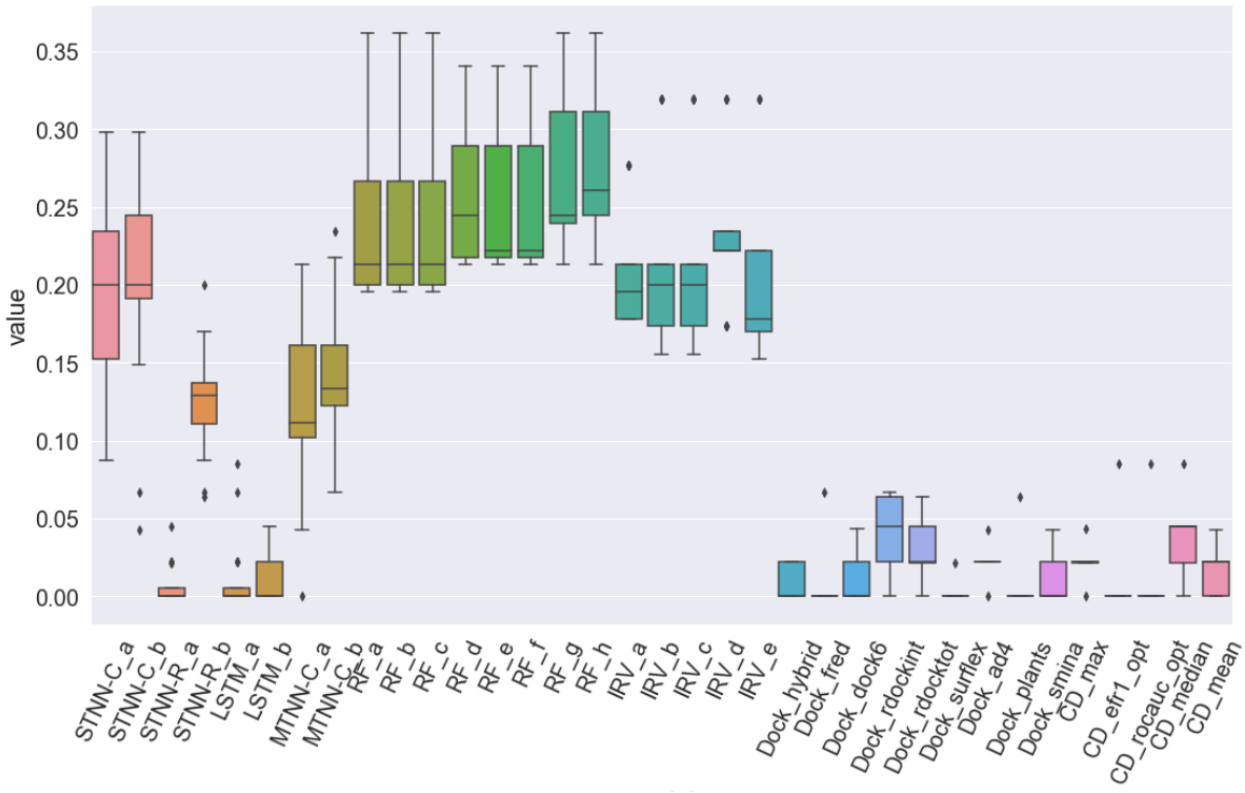


Figure A4.11. Cross-validation performance on RMI-FANCM with $NEF_{1\%}$.

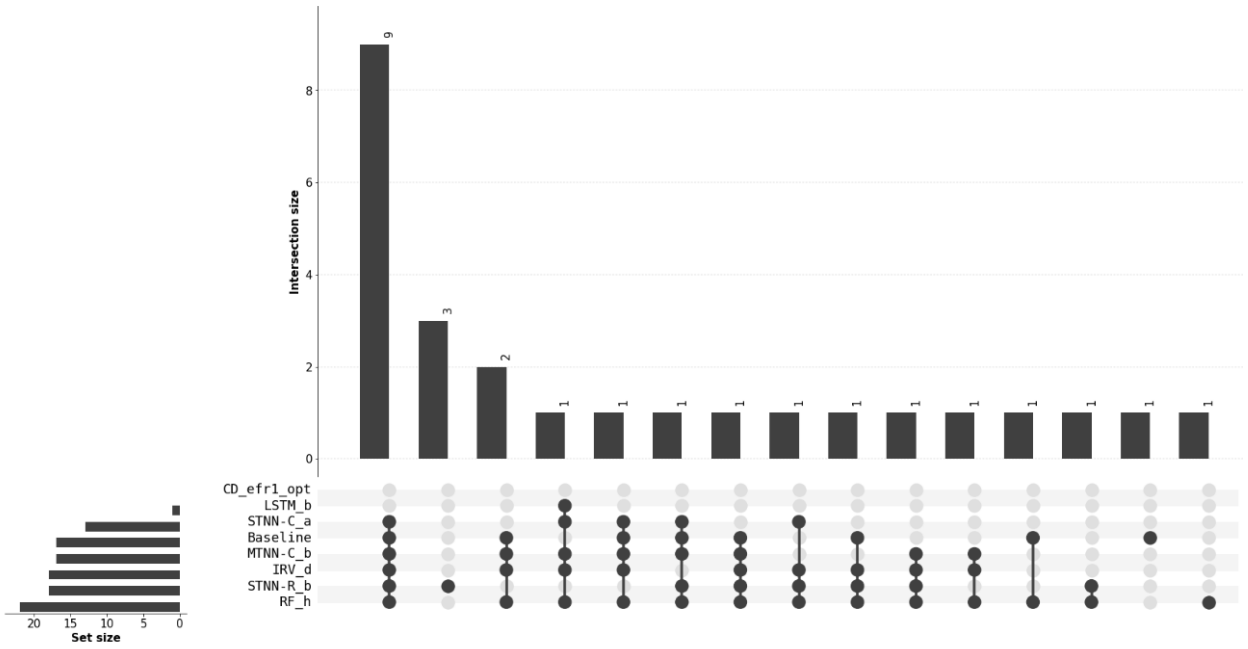


Figure A4.12. An UpSet plot showing the overlap in identified MCS clusters between the selected models and the chemical similarity baseline on PriA-SSB prospective.

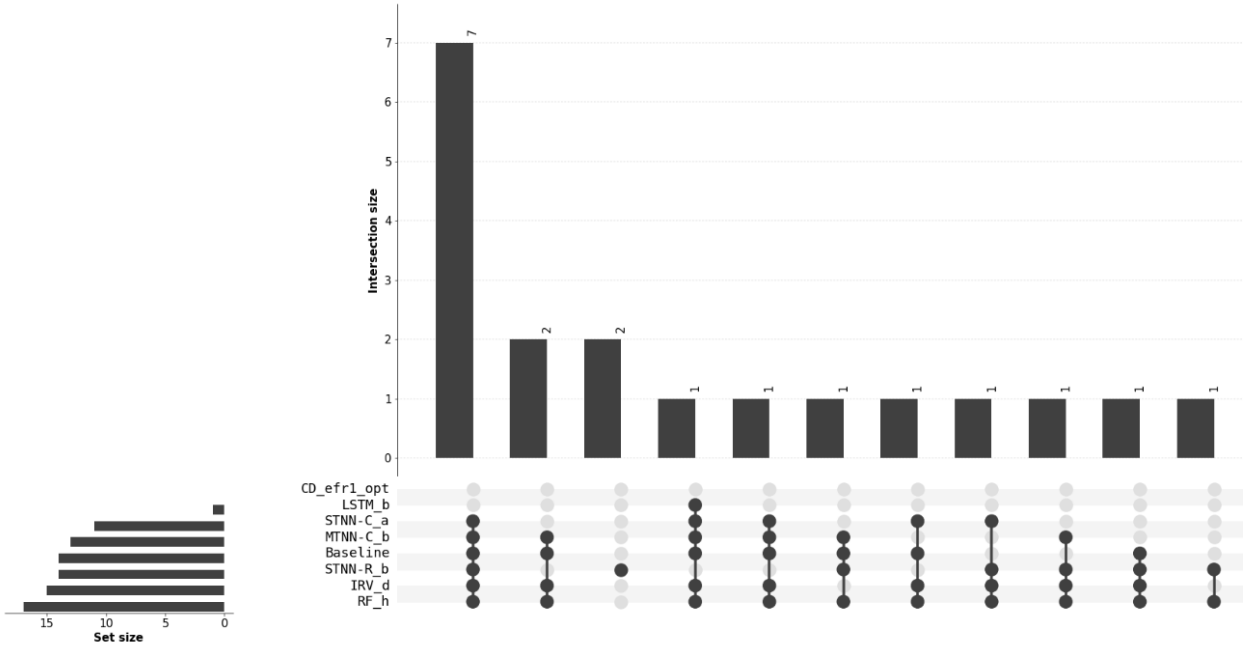


Figure A4.13. An UpSet plot showing the overlap in identified SIM clusters between the selected models and the chemical similarity baseline on PriA-SSB prospective.

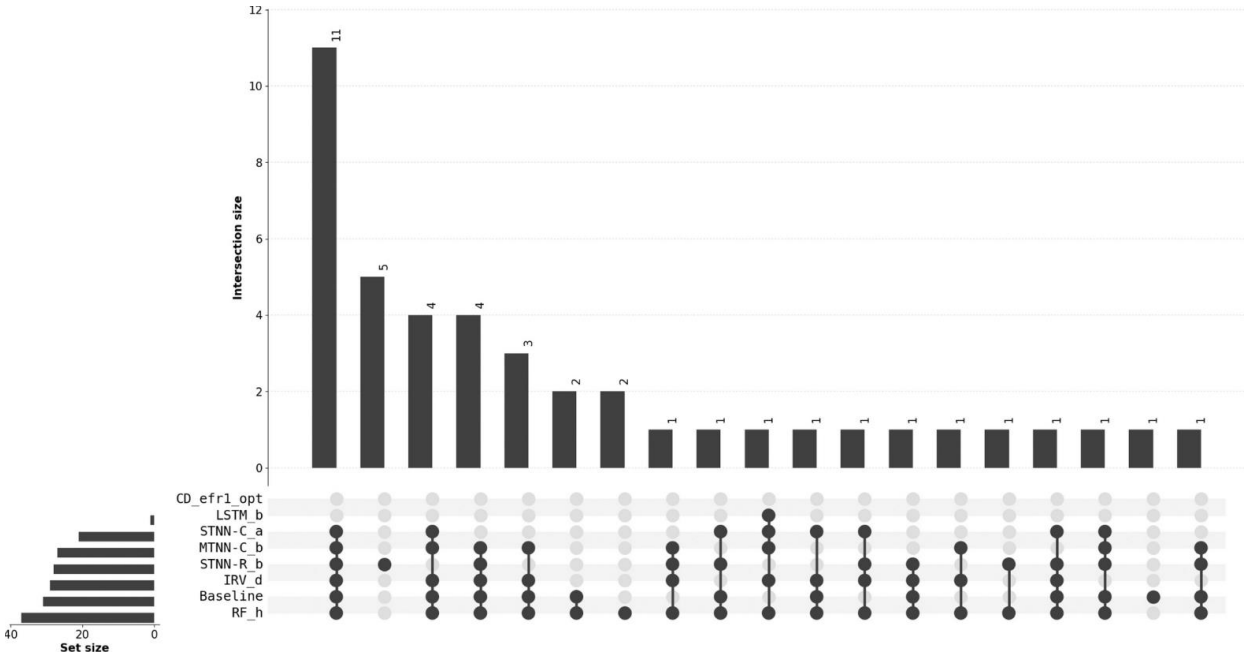


Figure A4.14. UpSet plot showing the overlap between the top 250 predictions from the selected VS models and the chemical similarity baseline on PriA-SSB prospective. The plot generalizes a Venn diagram by indicating the overlapping sets with dots on the bottom and the size of the overlaps with the bar graph.[58] Altogether, the combined predictions from the best-in-class VS methods and the baseline found 43.

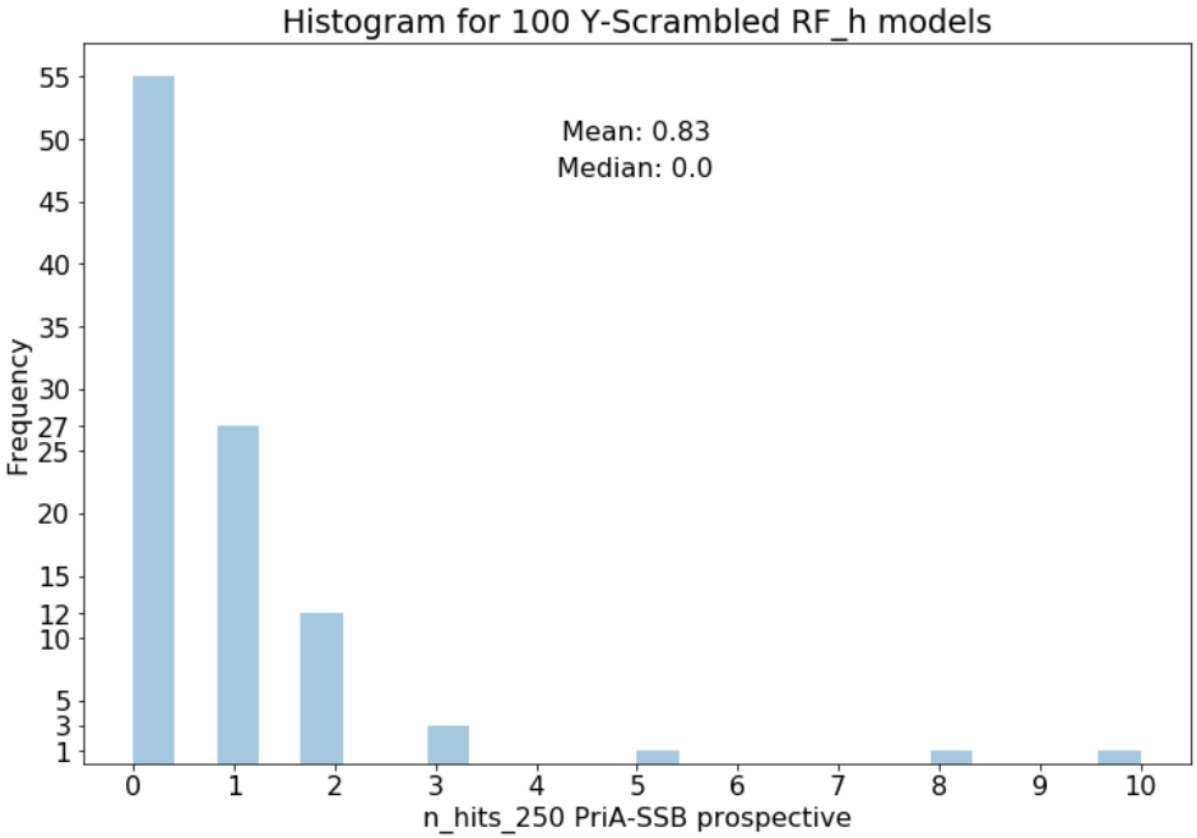


Figure A4.15. Histogram for 100 Y-Scrambled RF_h runs evaluated on the top 250 predictions for PriA-SSB prospective. 55 out of 100 runs found zero actives. Only a single run found 10 actives, far less than the 37 actives RF_h identified when trained on the real PriA-SSB AS data.

Table A4.1. Summary Statistics for the Four Binary Data Sets.

Stage	Data set	% inhibition threshold	# actives	# inactives
Cross-validation	PriA-SSB AS	≥35%	79	72,344
	PriA-SSB FP	≥30%	24	72,399
	RMI-FANCM FP	≥ mean + 2 SD	230	49,566
Prospective	PriA-SSB prospective	≥35%	54	22,380

Table A4.2. Data distribution of positive, negative, and missing molecules for each task.

Task name	Positive molecules	Negative molecules	Missing Molecules	Ratio (pos #/neg #, %)
pcba-aid1030	15932	145369	335063	10.96
pcba-aid1379	561	196368	314806	0.29
pcba-aid1452	178	149367	362573	0.12
pcba-aid1454	513	115335	395935	0.44
pcba-aid1457	720	202110	308746	0.36
pcba-aid1458	5778	188852	311888	3.06
pcba-aid1460	5650	217010	283986	2.60
pcba-aid1461	2305	206016	301670	1.12
pcba-aid1468	1038	251148	259072	0.41
pcba-aid1469	170	272533	239423	0.06
pcba-aid1471	293	218258	293452	0.13
pcba-aid1631	892	259030	251482	0.34
pcba-aid1634	154	261988	250000	0.06
pcba-aid1688	2375	201910	305636	1.18
pcba-aid1721	1087	289651	220471	0.38
pcba-aid2100	1157	291855	218127	0.40
pcba-aid2101	288	309907	201813	0.09
pcba-aid2147	3473	188764	316586	1.84
pcba-aid2242	715	183374	327492	0.39
pcba-aid2326	1065	259688	250478	0.41
pcba-aid2451	2005	271718	236568	0.74
pcba-aid2517	1138	332123	117897	0.34
pcba-aid2528	652	340938	170054	0.19
pcba-aid2546	10556	267886	223298	3.94
pcba-aid2549	1211	230450	279424	0.53
pcba-aid2551	16671	253653	225301	6.57
pcba-aid2662	110	285240	226836	0.04
pcba-aid2675	99	248789	263309	0.04
pcba-aid2676	1081	357341	152793	0.30
pcba-aid411	1563	69057	440113	2.26
pcba-aid463254	41	329171	183043	0.01
pcba-aid485281	253	314347	197443	0.08
pcba-aid485290	938	335859	174561	0.28
pcba-aid485294	148	309649	202351	0.05
pcba-aid485297	9128	301294	192746	3.03
pcba-aid485313	7569	304194	192964	2.49
pcba-aid485314	4493	312590	190720	1.44
pcba-aid485341	1729	325703	183135	0.53
pcba-aid485349	618	319466	191594	0.19
pcba-aid485353	603	322454	188636	0.19

Table A4.2 (continued). Data distribution of positive, negative, and missing molecules for each task.

Task name	Positive molecules	Negative molecules	Missing Molecules	Ratio (pos #/neg #, %)
pcba-aid485360	1485	216997	292329	0.68
pcba-aid485364	10698	331470	159430	3.23
pcba-aid485367	557	325598	185584	0.17
pcba-aid492947	80	329301	182835	0.02
pcba-aid493208	342	41294	470318	0.83
pcba-aid504327	766	370995	139769	0.21
pcba-aid504332	30264	263754	188014	11.47
pcba-aid504333	15673	310114	170836	5.05
pcba-aid504339	16859	338757	139821	4.98
pcba-aid504444	7388	282993	214527	2.61
pcba-aid504466	4169	306751	197207	1.36
pcba-aid504467	7648	235607	261393	3.25
pcba-aid504706	201	302548	209346	0.07
pcba-aid504842	101	324570	187524	0.03
pcba-aid504845	100	372270	139826	0.03
pcba-aid504847	3509	376531	128747	0.93
pcba-aid504891	34	361224	151004	0.01
pcba-aid540276	4393	192748	310762	2.28
pcba-aid540317	2129	367917	140121	0.58
pcba-aid588342	25036	301746	160478	8.30
pcba-aid588453	3904	365862	138626	1.07
pcba-aid588456	51	384356	127838	0.01
pcba-aid588579	1980	384213	124123	0.52
pcba-aid588590	3931	352947	151487	1.11
pcba-aid588591	4700	367981	134915	1.28
pcba-aid588795	1307	376247	133435	0.35
pcba-aid588855	4897	347556	154946	1.41
pcba-aid602179	364	384856	126712	0.09
pcba-aid602233	165	379055	132911	0.04
pcba-aid602310	310	393819	117857	0.08
pcba-aid602313	762	372273	138499	0.20
pcba-aid602332	69	408322	103836	0.02
pcba-aid624170	838	397756	112864	0.21
pcba-aid624171	1239	394674	115144	0.31
pcba-aid624173	487	399643	111679	0.12
pcba-aid624202	3968	362543	141817	1.09
pcba-aid624246	101	364511	147583	0.03
pcba-aid624287	423	302226	209224	0.14
pcba-aid624288	1356	323051	186533	0.42
pcba-aid624291	222	331803	180049	0.07

Table A4.2 (continued). Data distribution of positive, negative, and missing molecules for each task.

Task name	Positive molecules	Negative molecules	Missing Molecules	Ratio (pos #/neg #, %)
pcba-aid624296	9840	282428	210188	3.48
pcba-aid624297	6213	301951	197919	2.06
pcba-aid624417	6389	319289	180229	2.00
pcba-aid651635	3784	343160	161568	1.10
pcba-aid651644	748	353982	156818	0.21
pcba-aid651768	1677	355992	152950	0.47
pcba-aid651965	6346	318038	181566	2.00
pcba-aid652025	238	364167	147653	0.07
pcba-aid652104	7126	368557	129487	1.93
pcba-aid652105	4072	318365	185787	1.28
pcba-aid652106	497	362334	148968	0.14
pcba-aid686970	5948	331060	169340	1.80
pcba-aid686978	62375	236628	150918	26.36
pcba-aid686979	48532	257279	157953	18.86
pcba-aid720504	10170	340357	151599	2.99
pcba-aid720532	976	11815	498529	8.26
pcba-aid720542	733	356204	154626	0.21
pcba-aid720551	1265	341660	168106	0.37
pcba-aid720553	3259	336029	169749	0.97
pcba-aid720579	1908	280991	227489	0.68
pcba-aid720580	1508	304454	204826	0.50
pcba-aid720707	268	363257	148503	0.07
pcba-aid720708	661	356743	154231	0.19
pcba-aid720709	516	352850	158414	0.15
pcba-aid720711	290	363245	148471	0.08
pcba-aid743255	901	366915	143579	0.25
pcba-aid743266	306	398728	112956	0.08
pcba-aid875	34	73821	438407	0.05
pcba-aid881	590	103808	407308	0.57
pcba-aid883	1217	6647	503215	18.31
pcba-aid884	3396	6983	498521	48.63
pcba-aid885	160	12683	499293	1.26
pcba-aid887	1017	68423	441839	1.49
pcba-aid891	1564	6012	503156	26.01
pcba-aid899	1773	6141	502609	28.87
pcba-aid902	1865	117072	391494	1.59
pcba-aid903	338	52451	459169	0.64
pcba-aid904	528	50430	460810	1.05
pcba-aid912	453	56178	455212	0.81
pcba-aid914	221	7524	504330	2.94

Table A4.2 (continued). Data distribution of positive, negative, and missing molecules for each task.

Task name	Positive molecules	Negative molecules	Missing Molecules	Ratio (pos #/neg #, %)
pcba-aid915	421	7524	503930	5.60
pcba-aid924	1144	118813	391195	0.96
pcba-aid925	39	64140	448078	0.06
pcba-aid926	345	56230	455376	0.61
pcba-aid927	60	58565	453611	0.10
pcba-aid938	1781	60720	448014	2.93
pcba-aid995	699	65056	445842	1.07
PriA-SSB AS	79	72344	439794	0.11
PriA-SSB FP	24	72399	439849	0.03
RMI-FANCM FP	230	49566	462270	0.46

Table A4.3. Summary of Virtual Screening Methods and Which Labels Each Model Used during Training.[a]

Model	Continuous % inhibition	Binary label	PCBA binary labels
Dock			
CD			
STNN-C		√	
STNN-R	√		
MTNN-C		√	√
LSTM		√	
IRV		√	
RF		√	
Similarity baseline		√	

a. The docking and consensus docking models do not train on the PriA-SSB or RMI-FANCM data sets.

Table A4.4. Hyperparameter sweeping for classification neural networks (STNN-C and MTNN-C).

Hyperparameters	Candidate values
hidden layer sizes	[2000, 2000]
learning rate	0.00003, 0.00001, 0.003
optimizer	Adam
weighted schema	no_weight, weighted_sample
epoch patience	[epoch_size: 200, patience: 50], [epoch_size: 1000, patience: 200]
activations	[ReLU, Sigmoid, Sigmoid], [ReLU, ReLU, Sigmoid]
dropout	0.25

Table A4.5. Hyperparameter sweeping for regression neural networks (STNN-R).

Hyperparameters	Candidate values
hidden layer sizes	[2000, 2000]
learning rate	0.00003, 0.00001, 0.003
optimizer	Adam
weighted schema	no_weight
epoch	200, 1000
activations	[Sigmoid, Sigmoid, Linear], [ReLU, Sigmoid, Sigmoid]
dropout	0.25

Table A4.6. Hyperparameter sweeping for LSTM neural networks.

Hyperparameters	Candidate values
hidden layer sizes	[50], [100], [100, 10], [100, 50], [50, 10]
embedding layer size	30, 50, 100
learning rate	0.00003, 0.00001, 0.003
optimizer	Adam
epoch patience	[epoch_size: 200, patience: 50]
dropout	0.2, 0.5

Table A4.7. Hyperparameters for IRV.

Hyperparameters	Candidate values
number of neighbors	5, 10, 20, 40, 80
epoch patience	[epoch_size: 1000, patience: 20]
batch size	8192
learning rate	0.01
penalty	0.05

Table A4.8. Hyperparameter sweeping for RF.

Hyperparameters	Candidate values
n_estimators	4000, 8000, 16000
max_features	None, sqrt, log2
min_samples_leaf	1, 10, 100, 1000
class_weight	None, balanced_subsample, balanced

Table A4.9. Single-task neural network classification model name to hyperparameter mapping.

Model	weighted schema	optimizer	learning rate	early stopping	epoch patience	activations
STNN-C_a	no_weight	Adam	0.003	PR	patience: 200, epoch_size: 1000	[ReLU, ReLU, Sigmoid]
STNN-C_b	no_weight	Adam	3E-05	PR	patience: 200, epoch_size: 1000	[ReLU, ReLU, Sigmoid]

Table A4.10. Single-task neural network regression model name to hyperparameter mapping.

Model	activations	epoch size	weighted schema	optimizer	learning rate
STNN-R_a	[Sigmoid, Sigmoid, Linear]	200	no_weight	Adam	0.003
STNN-R_b	[Sigmoid, Sigmoid, Linear]	1000	no_weight	Adam	0.003

Table A4.11. Multi-task neural network classification model name to hyperparameter mapping.

Model	weighted schema	optimizer	learning rate	early stopping	epoch patience	activations
STNN-C_a	weighted_sample	Adam	0.0001	PR	patience: 50, epoch_size: 200	[ReLU, ReLU, Sigmoid]
STNN-C_b	no_weight	Adam	3E-05	PR	patience: 200, epoch_size: 1000	[ReLU, ReLU, Sigmoid]

Table A4.12. LSTM model name to hyperparameter mapping.

Model	embedding size	optimizer	dropout	early stopping	epoch patience	hidden size
LSTM_a	50	RMSprop	0.2	ROC	patience: 50, epoch_size: 200	[100, 50]
LSTM_b	30	RMSprop	0.5	ROC	patience: 50, epoch_size: 200	[50, 10]

Table A4.13. IRV model name to hyperparameter mapping.

Model	n_neighbors	epochs	patience	batch_size	learning_rate	penalty
IRV_a	5	1000	20	8192	0.01	0.05
IRV_b	10	1000	20	8192	0.01	0.05
IRV_c	20	1000	20	8192	0.01	0.05
IRV_d	40	1000	20	8192	0.01	0.05
IRV_e	80	1000	20	8192	0.01	0.05

Table A4.14. Random Forest model name to hyperparameter mapping.

Model	n_estimators	max_features	min_samples_leaf	class_weight
RF_a	4000	sqrt	1	None
RF_b	8000	sqrt	1	None
RF_c	16000	sqrt	1	None
RF_d	4000	log2	1	None
RF_e	8000	log2	1	None
RF_f	4000	None	1	balanced
RF_g	4000	log2	1	balanced
RF_h	8000	log2	1	balanced

Table A4.15. PriA-SSB AS metric comparison showing metrics ranked by their Spearman correlation.

	n hits 100	n hits 250	n hits 500	n hits 1000	n hits 2500	n hits 5000	n hits 10000
0	NEF_0.5%	NEF_1%	NEF_0.5%	NEF_5%	NEF_20%	NEF_20%	NEF_0.5%
1	NEF_1%	NEF_2%	NEF_2%	NEF_2%	NEF_10%	ROC AUC	NEF_5%
2	NEF_2%	NEF_0.5%	NEF AUC BEDROC	NEF AUC BEDROC	NEF AUC BEDROC	NEF_10%	NEF AUC BEDROC
3	NEF_5%	NEF_5%	AUC	AUC	AUC	NEF AUC BEDROC	AUC
4	NEF_0.15%	NEF AUC BEDROC	NEF_5%	NEF_10%	ROC AUC	AUC	NEF_0.15%
5	NEF AUC BEDROC	AUC	NEF_1%	NEF_1%	NEF_5%	NEF_5%	NEF_0.15%
6	AUC	NEF_0.15%	NEF_10%	NEF_0.5%	NEF_2%	NEF_2%	NEF_1%
7	NEF_10%	NEF_10%	ROC AUC	ROC AUC	NEF_1%	NEF_1%	ROC_AUC
8	NEF_20%	NEF_20%	NEF_20%	NEF_20%	NEF_0.5%	NEF_0.5%	NEF_2%
9	ROC AUC	ROC AUC	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF_20%
10	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%
11	PR auc.integral	PR auc.integral	PR auc.integral	PR auc.integral	PR auc.integral	PR auc.integral	PR auc.integral

Table A4.16. PriA-SSB FP metric comparison showing metrics ranked by their Spearman correlation.

	n hits 100	n hits 250	n hits 500	n hits 1000	n hits 2500	n hits 5000	n hits 10000
0	NEF_0.5% PR	NEF_2% BEDROC	NEF AUC	NEF_10%	NEF_5%	NEF_10%	NEF_5%
1	auc.integral	AUC	NEF_1%	NEF AUC	NEF AUC	NEF AUC	NEF AUC
2	NEF_10%	ROC AUC	NEF_10%	NEF_1%	NEF_1%	NEF_1%	NEF_1%
3	NEF AUC	NEF_5%	NEF_5%	NEF_5% BEDROC	NEF_10%	NEF_5% BEDROC	NEF_10%
4	NEF_1%	NEF AUC	NEF_20% BEDROC	AUC	NEF_20% BEDROC	AUC	NEF_2%
5	NEF_5% BEDROC	NEF_1%	AUC	NEF_20%	AUC	NEF_20%	NEF_20% BEDROC
6	AUC	NEF_10%	NEF_2%	NEF_2%	NEF_2%	NEF_2%	AUC
7	NEF_20%	NEF_20%	ROC AUC	ROC AUC	ROC AUC	NEF_0.5% PR	ROC AUC
8	NEF_2%	NEF_0.5% PR	NEF_0.5% PR	NEF_0.5% PR	NEF_0.5% PR	auc.integral	NEF_0.5% PR
9	ROC AUC	auc.integral	auc.integral	auc.integral	auc.integral	ROC AUC	auc.integral
10	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%
11	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF_0.15%

Table A4.17. RMI-FANCM FP metric comparison showing metrics ranked by their Spearman correlation.

	n hits 100	n hits 250	n hits 500	n hits 1000	n hits 2500	n hits 5000	n hits 10000
0	NEF_1%	NEF_5%	NEF_5%	NEF_10%	NEF_20%	NEF AUC	ROC AUC BEDROC
1	NEF_2%	NEF_2%	NEF_2% BEDROC	NEF AUC BEDROC	NEF AUC	NEF_20%	AUC PR
2	NEF_5%	NEF_1%	AUC	AUC	NEF_10%	ROC AUC	auc.integral
3	NEF_0.5% BEDROC	NEF_0.5% BEDROC	NEF_10%	NEF_20%	BEDROC AUC	NEF_10% BEDROC	NEF_0.1%
4	AUC	AUC	NEF AUC	NEF_5%	ROC AUC	AUC	NEF_0.15%
5	NEF_10%	NEF_10%	NEF_0.5%	ROC AUC	NEF_2%	NEF_2%	NEF_0.5%
6	NEF AUC	NEF AUC	NEF_1%	NEF_2%	NEF_5%	NEF_5%	NEF_1%
7	ROC AUC	NEF_20%	NEF_20%	NEF_1%	NEF_1%	NEF_1%	NEF_2%
8	NEF_20%	ROC AUC	ROC AUC PR	NEF_0.5% PR	NEF_0.5% PR	NEF_0.5% PR	NEF_5%
9	NEF_0.1%	NEF_0.1% PR	auc.integral	auc.integral	auc.integral	auc.integral	NEF_10%
10	NEF_0.15% PR	auc.integral	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_0.1%	NEF_20%
11	auc.integral	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF_0.15%	NEF AUC

Table A4.18. PriA-SSB AS metric comparison Spearman correlation coefficient.

	n hits 100	n hits 250	n hits 500	n hits 1000	n hits 2500	n hits 5000	n hits 10000
ROC AUC	0.8809	0.8827	0.9340	0.9164	0.9288	0.9150	0.7885
BEDROC AUC	0.9251	0.9395	0.9613	0.9744	0.9414	0.8584	0.8116
PR auc.integral	0.4550	0.4352	0.4440	0.4346	0.4363	0.4570	0.3013
NEF_0.1%	0.7653	0.7767	0.7446	0.7991	0.6760	0.5676	0.6034
NEF_0.15%	0.9323	0.9346	0.9058	0.9122	0.7454	0.6410	0.8062
NEF_0.5%	0.9999	0.9878	0.9667	0.9460	0.8480	0.7720	0.8343
NEF_1%	0.9881	1.0000	0.9563	0.9590	0.8719	0.7770	0.7954
NEF_2%	0.9749	0.9882	0.9651	0.9747	0.8868	0.8102	0.7725
NEF_5%	0.9461	0.9590	0.9598	1.0000	0.9189	0.8305	0.8251
NEF_10%	0.9160	0.9319	0.9484	0.9648	0.9524	0.8956	0.7852
NEF_20%	0.8947	0.8954	0.9239	0.9154	0.9763	0.9251	0.7087
NEF AUC	0.9251	0.9395	0.9613	0.9744	0.9414	0.8584	0.8116

Table A4.19. PriA-SSB FP metric comparison Spearman correlation coefficient.

	n hits 100	n hits 250	n hits 500	n hits 1000	n hits 2500	n hits 5000	n hits 10000
ROC_AUC	0.4918	0.8700	0.7281	0.7271	0.7284	0.4993	0.5226
BEDROC_AUC	0.5944	0.9979	0.8282	0.8285	0.8271	0.5862	0.5594
PR_auc.integral	1.0000	0.5421	0.6966	0.7174	0.6752	0.5226	-0.0290
NEF_0.1%	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NEF_0.15%	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NEF_0.5%	1.0000	0.5421	0.6966	0.7174	0.6752	0.5226	-0.0290
NEF_1%	0.6966	0.8159	1.0000	0.9996	0.9996	0.7281	0.6966
NEF_2%	0.5245	0.9997	0.8031	0.8013	0.8042	0.5654	0.5944
NEF_5%	0.6752	0.8166	0.9996	0.9983	1.0000	0.7271	0.7174
NEF_10%	0.7174	0.8144	0.9996	1.0000	0.9983	0.7284	0.6752
NEF_20%	0.5771	0.6658	0.8284	0.8281	0.8281	0.5859	0.5771
NEF_AUC	0.6966	0.8159	1.0000	0.9996	0.9996	0.7281	0.6966

Table A4.20. RMI-FANCM FP metric comparison Spearman correlation coefficient.

	n hits 100	n hits 250	n hits 500	n hits 1000	n hits 2500	n hits 5000	n hits 10000
ROC AUC	0.7391	0.8150	0.8313	0.9208	0.8917	0.7973	NaN
BEDROC AUC	0.8275	0.9036	0.9011	0.9672	0.8966	0.7632	NaN
PR auc.integral	0.4326	0.4359	0.4302	0.2172	0.2204	0.1487	NaN
NEF_0.1 %	0.5478	0.4665	0.3854	0.2062	0.1504	0.1130	NaN
NEF_0.15 %	0.4394	0.3307	0.2823	0.1149	0.0003	-0.0401	NaN
NEF_0.5 %	0.8491	0.9108	0.8844	0.7668	0.6479	0.5617	NaN
NEF_1 %	0.9963	0.9115	0.8801	0.8081	0.7264	0.7144	NaN
NEF_2 %	0.9232	0.9653	0.9163	0.8750	0.8417	0.7476	NaN
NEF_5 %	0.8926	0.9660	0.9779	0.9289	0.8412	0.7334	NaN
NEF_10 %	0.8195	0.8776	0.8973	0.9997	0.9003	0.7804	NaN
NEF_20 %	0.7281	0.8342	0.8418	0.9328	0.9622	0.8122	NaN
NEF AUC	0.8058	0.8776	0.8914	0.9868	0.9285	0.8127	NaN

Table A4.21. Top-ranked Models by Mean versus DTK+Mean on the Three Tasks. Evaluation metric means were computed over all cross-validation folds. [a]

	Best by Mean Model			Best by DTK+Mean model		
Metric	**PriA-SSB AS**	**PriA-SSB FP**	**RMI-FANCM FP**	**PriA-SSB AS**	**PriA-SSB FP**	**RMI-FANCM FP**
AUC[ROC]	RF_d	STNN-R_a	RF_h	RF_d	STNN-R_a	RF_h
AUC[BEDROC]	RF_h	STNN-R_b	RF_h	RF_h	STNN-R_b	RF_h
AUC[PR]	RF_g	STNN-R_a	RF_h	STNN-C_b	STNN-R_b	STNN-C_b
AUC[NEF]	RF_h	STNN-R_b	RF_h	RF_h	STNN-R_b	RF_h
$NEF_{1\%}$	RF_h	STNN-R_b	RF_h	RF_h	STNN-R_b	RF_h

[a]The prospective screening was only performed on PriA-SSB. Model names are mapped to their hyperparameter values in Tables A4.9-A4.13.

Table A4.22. Number of Active Compounds in the Top 250 Predictions from the Seven Selected Models and the Chemical Similarity Baseline Compared to the number of Experimentally Identified Actives. [a]

Model	Actives	Actives not in baseline	SIM clusters	MCS clusters
Experimental	54	–	27	35
Similarity baseline	31	–	14	17
CD_efr1_opt	0	0	0	0
STNN-C_a	21	2	11	13
STNN-R_b	28	8	14	18
LSTM_b	1	1	1	1
MTNN-C_b	27	3	13	17
RF_h	37	7	17	22
IRV_d	29	4	15	18

[a]These selected models are the best in each algorithm category from cross-validation. The last two columns correspond to the number of distinct chemical clusters from similarity or maximum common substructure clustering that are represented among the 54 actives. The consensus docking model CD_efr1_opt ranks the PriA-SSB prospective compounds without using information from the PriA-SSB AS training data. Prospective performance for all VS models is in Table A4.25.

Table A4.23. Off-target evaluation metrics for all models. As a control, the models trained on RMI-FANCM FP were evaluated on PriA-SSB prospective.

Model	AUC[ROC]	AUC[BEDROC]	AUC[PR]	NEF1%
ConsensusDocking_efr1_opt	0.39931	0.04559	0.00182	0.00000
ConsensusDocking_max	0.49919	0.07458	0.00227	0.00000
ConsensusDocking_mean	0.48110	0.07205	0.00217	0.00000
ConsensusDocking_median	0.47915	0.06370	0.00214	0.00000
ConsensusDocking_rocauc_opt	0.41457	0.04680	0.00186	0.00000
Docking_ad4	0.37911	0.02775	0.00173	0.00000
Docking_dock6	0.60630	0.12972	0.00318	0.00000
Docking_fred	0.40761	0.06924	0.00218	0.01852
Docking_hybrid	0.43895	0.04436	0.00196	0.00000
Docking_plants	0.45539	0.07053	0.00212	0.01852
Docking_rdockint	0.52785	0.08938	0.00245	0.00000
Docking_rdocktot	0.60125	0.13316	0.00313	0.00000
Docking_smina	0.44368	0.04307	0.00195	0.00000
Docking_surflex	0.56411	0.10643	0.00295	0.01852
IRV_a	0.52240	0.13901	0.00615	0.03704
IRV_b	0.51675	0.13037	0.00701	0.03704
IRV_c	0.53331	0.15181	0.00605	0.03704

IRV_d	0.52513	0.14306	0.00593	0.03704
IRV_e	0.52069	0.14026	0.00608	0.03704
LSTM_a	0.60099	0.15468	0.00341	0.00000
LSTM_b	0.55336	0.18084	0.00351	0.00000
MultiClassification_a	0.64870	0.20565	0.00551	0.03704
MultiClassification_b	0.54647	0.14914	0.00376	0.05556
RandomForest_a	0.50393	0.09951	0.00583	0.03704
RandomForest_b	0.52529	0.09898	0.00634	0.03704
RandomForest_c	0.52841	0.09705	0.00654	0.03704
RandomForest_d	0.49301	0.09502	0.00617	0.03704
RandomForest_e	0.49008	0.09053	0.00609	0.03704
RandomForest_f	0.61363	0.15680	0.01225	0.03704
RandomForest_g	0.51600	0.10026	0.00628	0.03704
RandomForest_h	0.53381	0.10430	0.00637	0.03704
SingleClassification_a	0.51852	0.11102	0.01917	0.03704
SingleClassification_b	0.53075	0.13639	0.01135	0.03704
SingleRegression_a	0.44809	0.07585	0.00224	0.01852
SingleRegression_b	0.44100	0.07796	0.00482	0.03704

Table A4.24. On-target evaluation metrics for all models. Models were trained on PriA-SSB AS and evaluated on the PriA-SSB prospective.

Model	AUC[ROC]	AUC[BEDROC]	AUC[PR]	NEF1%
Baseline	0.84937	0.67375	0.16167	0.55556
ConsensusDocking_efr1_opt	0.57953	0.11677	0.00293	0.00000
ConsensusDocking_max	0.57996	0.14810	0.00337	0.03704
ConsensusDocking_mean	0.55288	0.09588	0.00261	0.00000
ConsensusDocking_median	0.53129	0.07482	0.00246	0.00000
ConsensusDocking_rocauc_opt	0.58635	0.11949	0.00298	0.00000
Docking_ad4	0.36292	0.01643	0.00159	0.00000
Docking_dock6	0.55541	0.13279	0.00454	0.01852
Docking_fred	0.51009	0.12103	0.00301	0.03704
Docking_hybrid	0.49760	0.13474	0.00293	0.01852
Docking_plants	0.48162	0.06959	0.00223	0.01852
Docking_rdockint	0.56174	0.12492	0.00324	0.01852
Docking_rdocktot	0.68720	0.21658	0.00470	0.01852
Docking_smina	0.42361	0.03394	0.00188	0.00000
Docking_surflex	0.57940	0.15061	0.00341	0.01852
IRV_a	0.64669	0.35955	0.07617	0.29630
IRV_b	0.71961	0.48510	0.12394	0.44444

IRV_c	0.78292	0.59296	0.18787	0.51852
IRV_d	0.82602	0.65816	0.19050	0.51852
IRV_e	0.86718	0.71450	0.20442	0.53704
LSTM_a	0.58979	0.17634	0.00357	0.01852
LSTM_b	0.61639	0.18218	0.00440	0.01852
MultiClassification_a	0.83244	0.58368	0.18462	0.40741
MultiClassification_b	0.84750	0.61346	0.22199	0.50000
RandomForest_a	0.87578	0.73649	0.28165	0.66667
RandomForest_b	0.87065	0.74287	0.28530	0.66667
RandomForest_c	0.87524	0.74433	0.28648	0.66667
RandomForest_d	0.88677	0.75521	0.28425	0.64815
RandomForest_e	0.89007	0.75693	0.28200	0.66667
RandomForest_f	0.88105	0.68324	0.17308	0.44444
RandomForest_g	0.88903	0.76547	0.36893	0.66667
RandomForest_h	0.89689	0.76886	0.37933	0.66667
SingleClassification_a	0.76435	0.51959	0.11103	0.37037
SingleClassification_b	0.81857	0.61809	0.30469	0.55556
SingleRegression_a	0.92068	0.73424	0.18769	0.53704
SingleRegression_b	0.89712	0.68403	0.18575	0.50000

Table A4.25. Number of active compounds and unique clusters in the top 250 predictions compared to the experimental actives.

Model	Actives	Actives not in baseline	SIM clusters	MCS clusters
Experimental	54	–	27	35
Baseline	31	0	14	17
ConsensusDocking_efr1_opt	0	0	0	0
ConsensusDocking_max	2	1	2	2
ConsensusDocking_mean	0	0	0	0
ConsensusDocking_median	0	0	0	0
ConsensusDocking_rocauc_opt	0	0	0	0
Docking_ad4	0	0	0	0
Docking_dock6	1	1	1	1
Docking_fred	2	1	2	2
Docking_hybrid	1	0	1	1
Docking_plants	1	1	1	1
Docking_rdockint	1	0	1	1
Docking_rdocktot	1	0	1	1
Docking_smina	0	0	0	0
Docking_surflex	1	1	1	1
IRV_a	16	1	9	12

IRV_b	24	3	14	16
IRV_c	28	4	15	18
IRV_d	29	4	15	18
IRV_e	29	4	15	18
LSTM_a	1	0	1	1
LSTM_b	1	1	1	1
MultiClassification_a	22	1	11	13
MultiClassification_b	27	3	13	17
RandomForest_a	36	6	17	20
RandomForest_b	37	7	17	21
RandomForest_c	36	6	17	20
RandomForest_d	36	7	17	21
RandomForest_e	36	7	17	21
RandomForest_f	24	4	12	17
RandomForest_g	37	7	17	22
RandomForest_h	37	7	17	22
SingleClassification_a	21	2	11	13
SingleClassification_b	31	5	16	19
SingleRegression_a	29	5	13	18

SingleRegression_b	28	8	14	18

Algorithm A4.1: Multi-task Data Splitting

Input: Initial pre-split molecule-target matrix M, number of desired folds k
Output: k folds F[1], F[2], ..., F[k] containing stratified splits of M

```
shuffle rows of M randomly
create k folds F[1], F[2], ..., F[k] that contain the row indexes only
indexList  argsort columns of M from smallest active counts to largest
for i in indexList do
|       currColumn  <- M[:; i]
|       split active indexes of currColumn into the k folds
|       split inactive indexes of currColumn into the k folds
|       split missing indexes of currColumn into the k folds
|       take the unique compounds in each fold to remove duplicate row indexes
|_      greedily remove overlapping indexes from each fold (fold-by-fold manner)
take the unique compounds in each fold to remove duplicate row indexes
return F[1], F[2], ..., F[k]
```

References

1. Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.***2009**, *49*, 1455– 1474.
2. Lionta, E.; Spyrou, G.; Vassilatis, D.; Cournia, Z. Structure-based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.* **2014**, *14*, 1923– 1938.
3. Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharmaceutics* **2017**, *14*,4462– 4475.
4. Tseng, Y. J.; Hopfinger, A. J.; Esposito, E. X. The Great Descriptor Melting Pot: Mixing Descriptors for the Common Good of QSAR Models. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 39– 43.
5. Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-matching and Docking As Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74– 82.
6. Krüger, D. M.; Evers, A. Comparison of Structure-and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem* **2010**, *5*, 148– 158.
7. Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-based Virtual Screening Tools against the DUD Data Set Reveals Limitations of Current 3d Methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079– 2093.
8. Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdisciplinary Reviews. Computational Molecular Science* **2014**, *4*, 468– 481.
9. Merck. Merck Molecular Activity Challenge. https://www.kaggle.com/c/MerckActivity (accessed **2017**–10–01).

10. Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions. *arXiv preprint arXiv:1406.1231*, **2014**.

11. Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets As a Method for Quantitative Structure–activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263– 274.

12. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning.*Front. Environ. Sci.* **2016**, *3*, 80.

13. Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Hochreiter, S. Deep Learning as an Opportunity in Virtual Screening. *Deep Learning and Representation Learning Workshop: Neural Information Processing Systems* **2014**, *2014*, 27.

14. Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. *arXiv preprint arXiv:1502.02072*, **2015**.

15. Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.;Agapow, P.-M.; Zietz, M.; Hoffman, M. M.; Xie, W.; Rosen, G. L.; Lengerich, B. J.; Israeli, J.; Lanchantin, J.;Woloszynek, S.; Carpenter, A. E.; Shrikumar, A.; Xu, J.; Cofer, E. M.; Lavender, C. A.; Turaga, S. C.;Alexandari, A. M.; Lu, Z.; Harris, D. J.; DeCaprio, D.; Qi, Y.; Kundaje, A.; Peng, Y.; Wiley, L. K.;Segler, M. H. S.; Boca, S. M.; Swamidass, S. J.; Huang, A.; Gitter, A.; Greene, C. S. Opportunities and Obstacles for Deep Learning in Biology and Medicine. *J. R. Soc., Interface* **2018**, *15*, 20170387.

16. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241– 1250.

17. Voter, A. F.; Manthei, K. A.; Keck, J. L. A High-throughput Screening Strategy to Identify Protein-protein Interaction Inhibitors That Block the Fanconi Anemia DNA Repair Pathway. *J. Biomol. Screening* **2016**, *21*,626– 633.

18. Voter, A. F.; Killoran, M. P.; Ananiev, G. E.; Wildman, S. A.; Hoffmann, F. M.; Keck, J. L. A High-Throughput Screening Strategy to Identify Inhibitors of SSB Protein–Protein Interactions in an Academic Screening Facility. *SLAS DISCOVERY: Advancing Life Sciences R&D* **2018**, *23*, 94– 101.

19. Nordmann, P.; Cuzon, G.; Naas, T. The Real Threat of Klebsiella Pneumoniae Carbapenemase-producing Bacteria. *Lancet Infect. Dis.* **2009**, *9*, 228– 236.

20. Manthei, K. A.; Keck, J. L. The BLM Dissolvasome in DNA Replication and Repair. *Cell. Mol. Life Sci.* **2013**,*70*, 4067– 4084.

21. Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.

22. Lagorce, D.; Sperandio, O.; Baell, J. B.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs3: a Web Server for Compound Property Calculation and Chemical Library Design. *Nucleic Acids Res.* **2015**, *43*, W200– W207.

23. Capuzzi, S. J.; Muratov, E. N.; Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference CompoundS. *J. Chem. Inf. Model.* **2017**, *57*, 417– 427.

24. Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.;Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955– D963.

25. RDKit: Open-source Cheminformatics. *http://rdkit.org* accessed 2016-03-04.

26. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107– 113.

27. Rogers, D.; Hahn, M. Extended-connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742– 754.

28. Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97– 101.

29. Chollet, F. Keras. https://github.com/fchollet/keras (accessed 12/20/2016).

30. The Theano Development Team. Al-Rfou, R.; Alain, G.; Almahairi, A.; Angermueller, C.; Bahdanau, D.;Ballas, N.; Bastien, F.; Bayer, J.; Belikov, A.; Belopolsky, A.; Bengio, Y.; Bergeron, A.; Bergstra, J.;Bisson, V.; Snyder, J. B.; Bouchard, N.; Boulanger-Lewandowski, N.; Bouthillier, X.; de Brébisson, A.;Breuleux, O.; Carrier, P.-L.; Cho, K.; Chorowski, J.; Christiano, P.; Cooijmans, T.; Côté, M.-A.; Côté, M.;Courville, A.; Dauphin, Y. N.; Delalleau, O.; Demouth, J.; Desjardins, G.; Dieleman, S.; Dinh, L.; Ducoffe, M.;Dumoulin, V.; Kahou, S. E.; Erhan, D.; Fan, Z.; Firat, O.; Germain, M.; Glorot, X.; Goodfellow, I.;Graham, M.; Gulcehre, C.; Hamel, P.; Harlouchet, I.; Heng, J.-P.; Hidasi, B.; Honari, S.; Jain, A.; Jean, S.;Jia, K.; Korobov, M.; Kulkarni, V.; Lamb, A.; Lamblin, P.; Larsen, E.; Laurent, C.; Lee, S.; Lefrancois, S.;Lemieux, S.; Léonard, N.; Lin, Z.; Livezey, J. A.; Lorenz, C.; Lowin, J.; Ma, Q.; Manzagol, P.-A.;Mastropietro, O.; McGibbon, R. T.; Memisevic, R.; van Merriënboer, B.; Michalski, V.; Mirza, M.; Orlandi, A.;Pal, C.; Pascanu, R.; Pezeshki, M.; Raffel, C.; Renshaw, D.; Rocklin, M.; Romero, A.; Roth, M.;Sadowski, P.; Salvatier, J.; Savard, F.; Schlüter, J.; Schulman, J.; Schwartz, G.; Serban, I. V.; Serdyuk, D.;Shabanian, S.; Simon, É.; Spieckermann, S.; Subramanyam, S. R.; Sygnowski, J.; Tanguay, J.;van Tulder, G.; Turian, J.; Urban, S.; Vincent, P.; Visin, F.; de Vries, H.; Warde-Farley, D.; Webb, D. J.;Willson, M.; Xu, K.; Xue, L.; Yao, L.; Zhang, S.; Zhang, Y. Theano: A Python Framework for Fast Computation of Mathematical Expressions. *arXiv preprint arXiv:1605.02688*, **2016**.

31. Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, **2014**.

32. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5-32.

33. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9*, 1735– 1780.

34. Jastrzębski, S.; Leśniak, D.; Czarnecki, W. M. Learning to SMILE(S). *arXiv preprint arXiv:1602.06289*, **2016**.

35. Swamidass, S. J.; Azencott, C.-A.; Lin, T.-W.; Gramajo, H.; Tsai, S.-C.; Baldi, P. Influence Relevance Voting: an Accurate and Interpretable Virtual High Throughput Screening Method. *J. Chem. Inf. Model.* **2009**, *49*,756– 766.

36. Lusci, A.; Fooshee, D.; Browning, M.; Swamidass, J.; Baldi, P. Accurate and Efficient Target Prediction Using a Potency-sensitive Influence-relevance Voter. *J. Cheminf.* **2015**, *7*, 63.

37. Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma?. *J. Chem. Inf. Model.* **2017**, *57*, 2068– 2076.

38. Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V.MoleculeNet: a Benchmark for Molecular Machine Learning. *Chemical Science* **2018**, *9*, 513.

39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.;Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.Scikit-learn: Machine Learning in Python. *J. Machine Learning Res.* **2011**, *12*, 2825– 2830.

40. Bhattacharyya, B.; George, N. P.; Thurmes, T. M.; Zhou, R.; Jani, N.; Wessel, S. R.; Sandler, S. J.; Ha, T.;Keck, J. L. Structural Mechanisms of PriA-mediated DNA Replication Restart. *Proc. Natl. Acad. Sci. U. S. A.***2014**, *111*, 1373– 1378.

41. Hoadley, K. A.; Xue, Y.; Ling, C.; Takata, M.; Wang, W.; Keck, J. L. Defining the Molecular Interface That Connects the Fanconi Anemia Protein FANCM to the Bloom Syndrome Dissolvasome. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 4437– 4442.

42. Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572– 584.

43. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*,2785– 2791.

44. Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.;Kuntz, I. D.; Rizzo, R. C. DOCK 6: Impact of New Features and Current Docking Performance. *J. Comput. Chem.* **2015**, *36*, 1132– 1156.

45. McGann, M. FRED and HYBRID Docking Performance on Standardized Datasets. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 897– 906.

46. Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84– 96.

47. Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.;Barril, X.; Hubbard, R. E.; Morley, S. D. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **2014**, *10*, e1003571.

48. Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893– 1904.

49. Cleves, A. E.; Jain, A. N. Knowledge-guided Docking: Accurate Prospective Prediction of Bound Configurations of Novel Ligands Using Surflex-Dock. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 485– 509.

50. Ericksen, S. S.; Wu, H.; Zhang, H.; Michael, L. A.; Newton, M. A.; Hoffmann, F. M.; Wildman, S. A. Machine Learning Consensus Scoring Improves Performance across Targets in Structure-based Virtual Screening. *J. Chem. Inf. Model.* **2017**, *57*, 1579– 1590.

51. Sud, M. MayaChemTools: An Open Source Package for Computational Drug Discovery. *J. Chem. Inf. Model.* **2016**, *56*, 2292– 2297.

52. Nicholls, A. What Do We Know and When Do We Know It?. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.

53. Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488– 508.

54. Swamidass, S. J.; Azencott, C.-A.; Daily, K.; Baldi, P. A CROC Stronger Than ROC: Measuring, Visualizing and Optimizing Early Retrieval. *Bioinformatics* **2010**, *26*, 1348– 1356.

55. Grau, J.; Grosse, I.; Keilwagen, J. Grosse I PRROC: Computing and Visualizing Precision-recall and Receiver Operating Characteristic Curves in R. *Bioinformatics* **2015**, *31*, 2595– 2597.