Development and Application of Structural Prediction Methods for Flexible Protein–Ligand Interactions

by

James M.B. McFarlane
Diploma of Applied Chemistry and Biotechnology, Camosun College, 2011
B.Sc. (Hons), University of Victoria, 2013

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Chemistry

We acknowledge with respect the Lekwungen peoples on whose traditional territory the university stands and the Songhees, Esquimalt, and W̱SÀNEĆ peoples whose historical relationships with the land continue to this day.

Development and Application of Structural Prediction Methods for Flexible Protein–Ligand Interactions

by

James M.B. McFarlane
Diploma of Applied Chemistry and Biotechnology, Camosun College, 2011
B.Sc. (Hons), University of Victoria, 2013

Supervisory Committee

Dr. Irina Paci, Supervisor
(Department of Chemistry)

Dr. Fraser Hof, Departmental Member
(Department of Chemistry)

Dr. Dennis Hore, Departmental Member
(Department of Chemistry)

Dr. Patrick Nahirney, Outside Member
(Division of Medical Sciences)

ABSTRACT

This dissertation presents a collection of biological simulations and predictions in collaboration with experiment to support and elucidate the trends observed in various protein–ligand systems. Within the model systems, there is strong focus on the support for development of peptidomimetic inhibitors for post-translational reader proteins (CBX proteins). The systems studied throughout this document each present their own unique challenges but fall under the general theme of protein flexibility and the difficulties of sampling such systems. As part of this work, methodological advances were made to address the challenges of structural prediction on flexible proteins and ultimately form the method *Selective Ligand-Induced Conformational Ensemble (SLICE)*. The development, validation, and future directions of the SLICE method are also discussed. Ultimately, the collaborative efforts presented in this dissertation bring forward a greater understanding of the drug design challenges on the CBX proteins as well a new methodology in the field of structure-based drug design.

Contents

List of Tables

List of Figures

GLOSSARY

AMBER	Assisted Model Building with Energy Refinement	A software suite for molecular dynamics simulation and analysis.
CADD	Computer-Assisted Drug Design	The use of computer software in the design and discovery of new drugs.
CBX	Chromobox Homolog	Post-translational reader subunit of the Polycomb Repressive Complex
CS	Conformational Selection	A type of drug binding event that requires correct protein conformation prior to binding.
IF	Induced-Fit	A type of drug binding event that induces a correct host conformation upon binding.
MD	Molecular Dynamics	An all-atom molecular simulation technique.
MBP	Maltose Binding Protein	Escherichia coli protein responsible for maltodextrin uptake with high disparity between apo and holo states.
SBDD	Structure-Based Drug Design	The use of host-protein structural information in drug design.
LBDD	Ligand-Based Drug Design	The use of known ligand activity in drug design.
PRC2	Polycomb Repressive Complex	Protein complex involved with histone methylation and methyl recognition.
PTM	Post-Translational Modification	Post-translationally modified amino acids, e.g., trimethylated lysine.
H3K9Me$_3$	Trimethylated Lysine 9 Histone 3 Tail	A methylated histone protein tail with a methylation site on Lysine 9
SLICE	Selective Ligand-Induced Conformational Ensemble	An iterative mixed stochastic and determination molecular simulation method.
FEP	Free Energy Perturbation	A free energy calculation method used in computational drug design.
TI	Thermodynamic Integration	A free energy calculation method used in computational drug design.
MMPBSA	Molecular Mechanics Poisson-Boltzmann Solvent Accessible	A free energy calculation method used in computational drug design.
MC	Monte-Carlo	A stochastic molecular simulation technique.
QSAR	Quantitative Structure–Activity Relationship	A predictive model for drug binding based on molecular descriptors.

ACKNOWLEDGEMENTS

I would like to thank:

Chelsea and Gavin, for giving me a reason to better myself as a father, husband, and a person.

Irina Paci, for giving me this opportunity, believing in me, and being a mentor and a friend over the past several years.

Various donors over the past several years, for funding that has helped me continue my studies and help support my new family.

Fraser Hof and Natasha Milosevich, for their collaborations and fruitful discussions on the research presented in this dissertation.

I got satisfaction out of doing things that were difficult. It was an incredible feeling. The pain was there, but the pain didn't matter.

Terry Fox

DEDICATION

To my brothers and sisters who share a love for these two trees:
41°8'29.14"N, 119°57'11.31"W
48°29'37.57"N, 124°17'45.36"W

Chapter 1

Introduction

The research and topics discussed in this dissertation revolve around the general theme of the prediction and analysis of molecular interactions between an organic molecule ligand and its protein host—more specifically, the use of tools and methods common to the field of structure-based drug design (SBDD) to gain insight into complex structural interactions. Included here are a set of introductory chapters that aim to guide the reader through the background of the proteins of interest as well as the methods used throughout this dissertation. The research content of this document is presented through several selected joint publications along with supplemental method descriptions and inferences of the data through a SBDD lens. All together, this dissertation aims to tell a story of collaborations between experiment and theory that not only worked to elucidate more than each could provide alone, but that also carved a path for the development of new a method in the structural prediction of flexible protein–ligand interactions.

This first chapter introduces molecular recognition of protein–ligand interactions and the role that these interactions play as challenges in the development of tools in computer-assisted drug design (CADD). The current role of CADD in drug discovery and where it needs to go as a field is also discussed. Throughout the selected publications of later chapters, the protein models within them fall under the category of epigenetic reader proteins and present unique modelling challenges that require new methodologies and knowledge of the frontier techniques used in SBDD. For this reason, individual chapters dedicated to the introduction of the protein models as well as current methods in SBDD are included as well.

Throughout this read, I would like the reader to maintain a healthy skepticism regarding the information gleaned from *all* theoretical research and molecular simulation but at the same time try to understand the usefulness of the models used with respect to the problem at hand. To let George Box put it more plainly,

"All models are wrong, but some are useful." We will be discussing the probing of biological interactions that exist in a reality far more complicated than we can depict in a computer simulation. However, with the hope that we have made the correct approximations, we may still extract useful information about a protein–ligand interaction for further exploitation. What are these approximations? What can we try to extract? Let us explore what we understand of protein–ligand interactions and how we try to digitally simulate perhaps the most important type of molecular recognition with regards to human health.

1.1 Protein–Ligand Interactions

Molecular recognition processes are indisputably regarded as the foundation for biological processes in all living organisms. The specificity and affinity of biological macromolecules interacting with other macromolecules or small compounds allows for fine control in an overwhelmingly complex system of potential interactions. Despite contributing to the complex and vast biochemical network in a living organism, individual molecular recognition processes themselves can be viewed as any other host–guest interaction and share the same features illustrated in Figure 1.1.

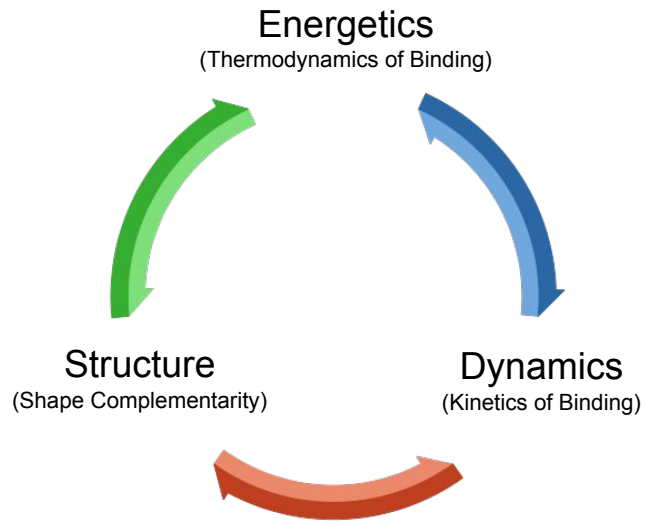


Figure 1.1: Aspects of molecular recognition. Kinetics, shape complementarity, and free energy of binding are core aspects of molecular recognition and important considerations in molecular modelling and computer-assisted drug design.

This supramolecular approach of zooming in on the driving components of specificity and affinity gives us a workable lens to study or exploit various pharmacologically relevant systems. In assessing host–guest interactions, it can be

useful to further partition aspects of binding into the thermodynamics (the relative strength of binding), kinetics (the rate at which the interaction occurs), and the shape of the interaction (the structural factors that determine the interaction's specificity). While this is a convenient classification, the three components are intrinsically linked. But for now, we will use this as a starting point for the discussion on why and how protein–ligand binding occurs with an emphasis on both shape complementarity and the strength of the interactions at hand.

1.1.1 Molecular Shape and Specificity

For over a century, going back as early as 1894, we have understood that the shape of a molecule acts as the figurative and metaphorical key in molecular recognition processes between a ligand and its protein host. Emil Fisher's early lock and key model [1] to describe enzyme specificity conjures images of unique molecular shapes inserting themselves into their mated active site. This model as depicted in Figure 1.2, simple yet robust, stood the test of time for nearly seventy years.

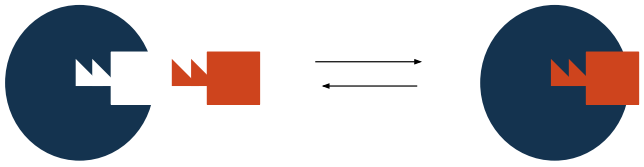


Figure 1.2: Lock and Key Diagram of Protein–Ligand Binding. The lock-and-key analogy of protein–ligand binding suggests that the host protein contains a pre-existing cavity amenable to the binding of its ligand guest.

Despite its profound impact on our understanding of molecular recognition with proteins, the potential of this concept went largely unrealized throughout its lifetime—as to design a key, one must know the shape of the lock. Structural information about binding sites (the lock) through crystallographic techniques would not be available until the later half of the 20th century and it would be around this time that our view of proteins would also change. A switch from proteins as static objects to dynamic and flexible macromolecules conflicts with Fisher's hypothesis. For this reason, Fisher's model had to evolve and was improved upon by the Koshland-Nementhy-Filmer theory of induced-fit in 1958 [2]. In this seminal work, Koshland et al. project ligand–protein binding through the analogy of a glove changing shape as a hand slips into it and describes it as a cooperative process wherein the host conforms to its guest upon binding.

It was not long after the introduction of the induced-fit model that protein–ligand binding dynamics was again challenged with an alternative. Changeaux and colleagues postulated that a bound configuration of the protein pre-exists in a conformational ensemble, and a population shift to the bound state occurs when the ligand is present. A contrast between these two paradigms is illustrated in Figure 1.3.

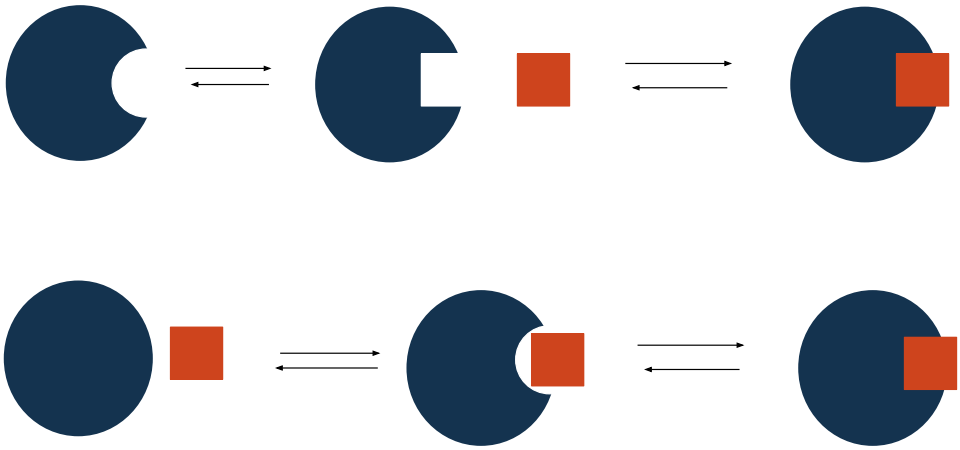


Figure 1.3: Conformational Selection and Induced Fit Protein–Ligand Binding Models. Two models of protein–ligand interactions assume different routes for how the host protein shape adapts to create a complimentary cavity for its ligand guest. Induced fit of the ligand implies that the host change is caused by direct interactions of the ligand molecule. On the other hand, conformational selection assumes that throughout the natural motions of the protein, a state of the protein exists in which the pocket is temporarily formed and then exploited.

In other words, the shape of the bound protein naturally exists only some of the time and ligand binding is an opportunistic event where the bound state is caught by the ligand. Nearly half a century forward to 2011, in a boldly titled paper *"Conformational selection or induced-fit? 50 years of debate resolved"* [3], Changeaux presents several concrete examples of protein–ligand systems where the bound configuration is observed through numerous experimental techniques in the absence of its respective ligand. As Changeaux suggests in his title, there has been a continued debate of the existence of one mechanism over the other. Despite his upfront statements (including a brazen title), Changeaux still posits in the conclusions that induced-fit mechanisms may work cooperatively to expedite the conformational selection of a protein conformation. Why the disclaimer? Perhaps it was the case studies presented in the work by Karplus et al. [4], or perhaps it was the nagging reality that molecules in contact with one another will always

exhibit a force on one another?

The debate of induced-fit versus conformational ensemble promotes a dichotomy that may not entirely exist. Several flavours and combinations of these theories exist, including a more widely accepted notion that both models may apply at various stages of binding and is dependent on the energetic and kinetic barriers involved in the recognition process such as studies done on the large clasp-like binding mechanism of Maltose Binding Protein [5]. Mixed models of CS and IF are clearly more universal in their descriptions of protein–ligand interactions and more importantly, open the door for us to think about binding events in stages and the various energetic contributions/penalities that both ligand and host incur.

The comparison of these theories and how the shape of the protein (and or ligand) come to be may not seem immediately important. However, if we are interested in how to predict the bound structure of a ligand with a flexible protein host, this distinction for each part of binding is paramount. The various models of host reorganization offer very different paths in how we would sample the configurational space of the host: An induced-fit model would require an interaction between host and ligand whereas the conformational selection model would allow us to sample a variety of host configurations generated in the absence of its guest. A mixed model would require us to do both. Either way, no matter how we get there, the consideration of the binding pocket shape allows us to probe the likely intermolecular interactions between ligand and host with the hope of later quantifying the strength of the interaction.

1.1.2 Thermodynamics of Protein–Ligand Binding

Similar to any physical or chemical process, the spontaneity and strength of protein–ligand interactions are governed by the energy of reactants (unbound ligand and host) in comparison to the products (the ligand–host complex). One incredibly important feature to note early in our discussions is that the difference in energy is the sum of both destabilizing and the stabilizing interactions that occur during binding. For instance, we may increase the number of favourable interactions between the protein and its guest, but if those new changes come at the cost of reorganizing the protein host to a higher energy state, the new interaction energy may be significantly offset (see Figure 1.4).

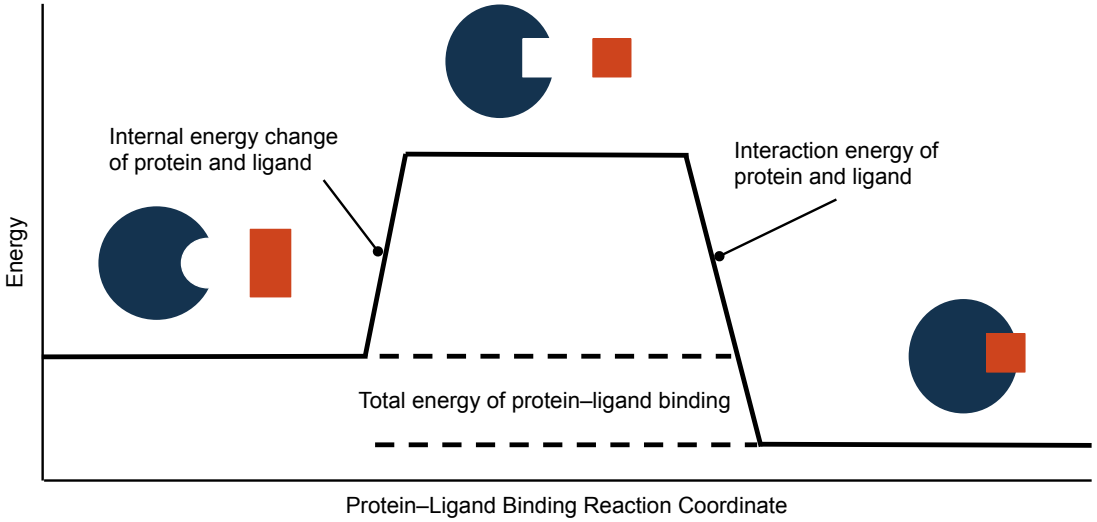


Figure 1.4: Protein–Ligand Reorganization Energy and Interaction Energy Compromise. Total binding energy of a protein–ligand complex is the result of balanced interaction energies with a number of energetic penalties. These penalties may include unfavourable solvation changes, loss of entropy, or in the case of this figure, a change in internal energy of the host and guest molecules.

To discuss the magnitude of these energy changes and the total free energy of binding, we use ΔG, the Gibb's free energy.

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

where ΔH equals the change in enthalpy, ΔS is the change in entropy, and T is the temperature.

The balance of stabilizing and destabilizing interactions can be explored by further splitting ΔG into the various energetic changes that occur during binding as shown in Equation 1.2.

$$\Delta G_{Binding} = \Delta G_{Desolvation} + \Delta G_{Motion} + \Delta G_{Configuration} + \Delta G_{Interaction} \tag{1.2}$$

where $\Delta G_{Desolvation}$ represents the energy change associated with the displacement of solvent molecules as the protein–ligand complex is formed, ΔG_{Motion} accounts for the change in entropic loss as two flexible entities form a single less flexible unit, $\Delta G_{Configuration}$ represents the change in energy as both host and ligand structurally rearrange to form the required binding geometries, and $\Delta G_{Interaction}$ is the enthalpic stabilization of the presence of the ligand caused by the intermolecular interactions between host and guest.

From an experimental approach, the strength and selectivity of a ligand can

be determined by accessing the heat of formation, equilibrium concentrations, or dissociation constants of the protein–ligand complex—all parameters leading to the $\Delta G_{Binding}$ through the relationships in Equations 1.3 and 1.4. Assays such as differential scanning calorimetry (DCS) or isothermal calorimetry (ITC) provide insight into the strength of intermolecular interactions (enthalpy) as well as entropic contributions to binding. On the other hand, methods such as fluorescence polarization (FP) provide the total free energies of binding through the extraction and application of equilibrium constants (K_d in Equation 1.3) for the protein–ligand interaction.

$$\Delta G_{\text{Binding}} = RT \ln K_d \tag{1.3}$$

$$K_d = \frac{[\text{Ligand}][\text{Host}]}{[\text{Complex}]} \tag{1.4}$$

Using experimental methods such as those listed above to access information about a protein–ligand interaction (both enthalpic and entropic) is paramount to the drug discovery process [6]. However, finding out the overall binding free energy can only go so far in the optimization of a protein–ligand interaction. To be able to finely tune ΔG, we need to think about the individual contributions of binding with respect to the mutable intermolecular interactions of a ligand with its protein host as illustrated in Eqn 1.5. In other words, how can we enhance the interaction energy from Equation 1.2? $\Delta G_{Interaction}$ is largely driven by the enthalpic contributions of intermolecular interactions listed below in Equation 1.5.

$$\Delta H = \Delta H_{\text{H-Bonding}} + \Delta H_{\text{VDW}} + \Delta H_{\text{Electrostatic}} + \Delta H_{\text{Hydrophobic}} \ldots \tag{1.5}$$

The endeavour to rationally tune the magnitude of the individual terms in equation 1.5 is at the heart of structure-based drug design and requires knowledge of both the positions of the atoms involved in the interactions and the equations that predict their strength—Enter theoretical chemistry. In later chapters, methods to calculate the energies of atoms based on their positions for simulation purposes will be described. For our current discussion, we will go over what intermolecular interactions are that contribute to the enthalpy of binding between a protein and its ligand guest.

Binding Energy Decomposition

As illustrated in equation 1.5, the enthalpic contributions of binding can be categorized into various types of intermolecular interactions. Throughout this dissertation, the interpretation of molecular simulations and structural prediction models is heavily supported with qualitative and quantitative discussions of the types of intermolecular interactions that are occurring. Such interactions include π-π, cation-π, hydrogen bonding, and numerous others. However, all interactions can be further broken down into fundamental intermolecular interactions: coulombic, dispersive, and partially covalent.

Coulombic interactions between biomolecules can be binned into charge-charge, charge-dipole, and dipole-dipole interactions. Some canonical amino acids contain charged or polar side-chains that interact electrostatically with each other, as well as surrounding solvent, with an inverse charge–distance dependance.

$$V(r_{12})_{\text{Coulombic}} = \frac{-1}{4\pi\epsilon_o}\frac{q_1 q_2}{r_{12}} \tag{1.6}$$

where q_1 and q_2 represent two point charges, k_e is Coulomb's constant, and r_{12} is the distance between the point charges with an example given in Figure 1.5.

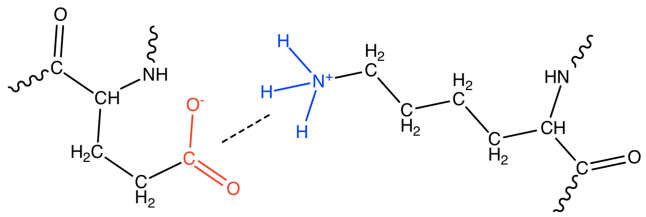


Figure 1.5: Lysine–Glutamate Salt-bridge. Salt bridging is a type of charge–charge electrostatic interaction between charged amino acid side chains.

Hydrogen bonding is often described as a polarizable electrostatic interaction. However, theoretical studies [7], preferred geometry of hydrogen bonds, and interatomic distances suggest a sharing of electrons—details that suggest chemical bonding and partial covalent character. Aside from being one of the more interesting intermolecular interactions, its importance in the formation of important biological complexes is unrivalled.

Lastly, dispersion forces or van der Waals forces are those created by the instantaneous dipoles of non-polar molecules. This interaction is seen primarily with non-polar or hydrophobic side chains of amino acids. The attractive dispersion forces are often described along with nuclear repulsion forces in what is called the

Lennard-Jones potential [8].

$$V_{\text{Repulsion}} + V_{\text{VDW}} = V_{\text{Lennard-Jones}} = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{1.7}$$

where r represents the distance between two atoms and σ represents the interatomic distance at the most stable interatomic distance, and ϵ as the corresponding minimum energy at the distance σ as illustrated in Fig. 1.6.

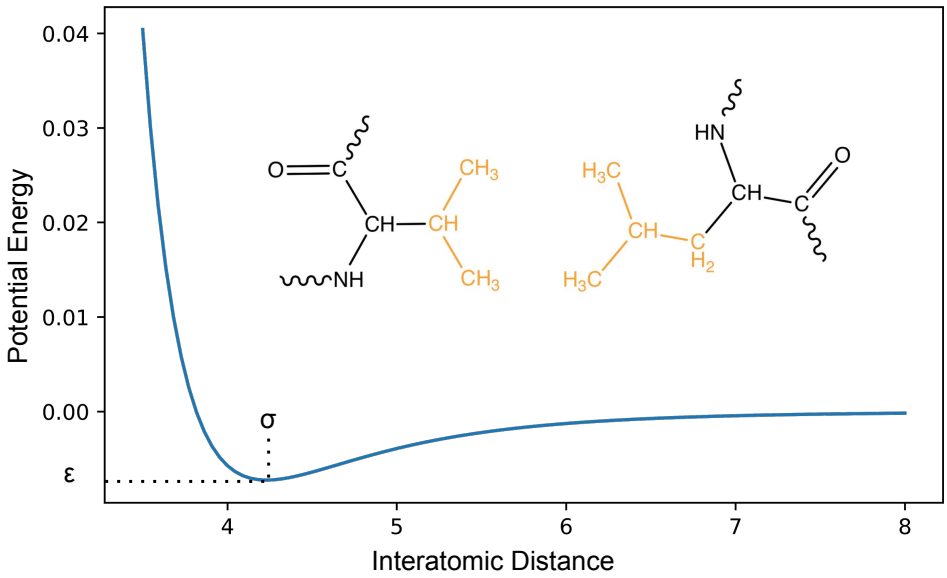


Figure 1.6: Lennard-Jones Potential and Valine–Leucine Interaction. Interactions between hydrophobic residues such as Valine and Leucine include dispersion forces or van der Waals forces that can be expressed by a Lennard-Jones potential such as in Equation 1.7.

The introduction of these enthalpic contributions to binding may seem trivial. However, it forms the basis for breaking down more complicated intermolecular interactions that are exploited in structure-based drug design. For instance, π-stacking is a balance between dispersive and electrostatic interactions, of which the balance of the contributions may change depending on substitutions on the interacting species [9]. This concept of balancing interactions becomes even more convoluted when entropy is taken into consideration and even more so when the enthalpy and entropy of the surrounding solvent become involved. However, all these interactions add to the total free energy of binding and in the case of the research presented in later chapters, sometimes remain elusive despite our best efforts in the breakdown of these terms.

The entropic gain or loss during binding comes from a number of sources including the conformational and translational freedom of the host and guest. Entropic changes are also affected by the number of bound solvent molecules around

the binding site before and after protein–ligand binding. For experimental methods, ΔS can be parsed from the free energy of binding if the change in enthalpy is known through calorimetric methods. However, in theoretical models, ΔS is much more elusive, as comprehensive information on the positions of atoms over a significant amount of time is needed, whereas enthalpies may be approximated from instantaneous positions of the atoms. To include entropy in theoretical models, entropic changes are spread over a number of terms. The conformational entropy is evaluated using a normal mode analysis [10] whereas the hydrophobic entropy is evaluated via a non-polar solvation term using empirical models based on the surface areas of binding between protein and ligand [11].

As an energy decomposition example to illustrate the subtle balance between all these interactions, let us look at the cation–π interaction. The interaction occurs between a positively charged species and the electronegative regions of an aromatic ring (See Fig. 1.7).

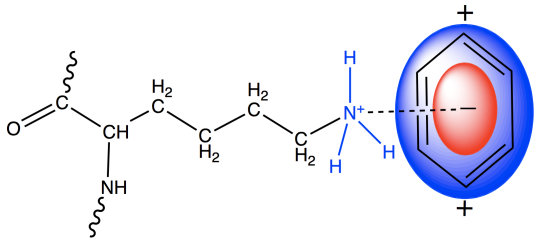


Figure 1.7: Cation-π interaction between Lysine and Benzene.

At first glance, the electrostatic interaction appears to be the driving force and is supported through computational studies in the gas phase [12]. The trends observed in the gas phase for the alkali cations show that smaller and more densely charged atoms produce a higher binding enthalpy, with binding energies ordered $Li^+ > Na^+ > K^+ > Rb^+$ for the binding to benzene. However, the introduction of a polar solvent such as water significantly changes the order to $K^+ > Rb^+ > Na^+ > Li^+$ [13]. This shuffling is interesting for a number of reasons. First, the order is not reversed and shows potassium is in an optimal position. Of each ion–benzene pair, the energy is the result of the difference between the electrostatics between ion–solvent and ion–host interactions as well as the gain in entropy as the ion solvation shells are displaced. As ionic radii become larger, dispersive energies are also likely to become more important. The message here is trends in atomic descriptions for even the simplest of systems present unsuspecting changes to binding free energies. The subtleties of solvation structures and the potential roles of enthalpy-entropy compensation [14] can complicate even the simplest of models, let alone an interaction as complex as a protein–peptide binding.

1.1.3 Peptide-Protein Binding and Drug Design

In the human body, it is estimated that 15–40 % of all the host–guest interactions occurring are comprised of either a peptide–protein (pepPI) or protein–protein interaction (PPI) [15]. This large (and naturally selected) contribution to physiological control in the body via peptides is inspiring. Clearly, there are advantages to using peptides as protein binders that the body has leaned into. Given this knowledge, why is it then that peptide-based drugs have not traditionally been sought out as first line candidates in pre-clinical discovery phases of drug development? Through a combination of synthetic challenges, pharmacokinetic limitations, and difficulty with theoretical prediction, peptide-based therapies have traditionally been steered clear of. However, innovations in drug delivery [16] and modular synthesis [17] involving non-standard amino acids [18] have significantly aided in the growing interest in peptides as potential drugs. Furthermore, the computational technology for predicting the interactions between proteins and peptides has significantly advanced within the last few decades [19].

To be totally fair, peptide based therapies have actually been around since the advent of insulin therapy. The use of endogenous human peptides as a peptide replacement has been a long-standing practice in medicine—Oxytocin and Calcitonin are just a few other examples of this. However, modern peptide-based therapies extend far beyond the use of synthetic or naturally sourced endogenous peptides and can now be either be a synthetic analog of a natural peptide or more excitingly, a novel chemical entity. Positive trends in the number of cumulative peptide approvals as well as peptides entering clinical trials show that not only are we overcoming challenges with designing peptides as selective binders, but overcoming the pharmacokinetic challenges associated with them as well [20].

Throughout this dissertation, peptide-based ligands are repeatedly explored as potential inhibitors for a protein class called the CBX proteins. In doing so, many of the typical challenges associated with peptides are encountered. Ignoring the synthetic challenges, peptides are also riddled with pharmacokinetic challenges such as issues with protease degradation as well as poor absorption and distribution including cell permeability problems. However, these challenges are beyond the modelling work presented here but still set the tone for the difficult path in the rational design of peptide inhibitors. For the peptide ligand work presented here, challenges in peptide design at the peptide-protein interface and the structural prediction of these complexes are our greatest concerns.

Peptides (in proportion to their size) have an outstanding range of conformational flexibility depending on their amino acid [21]. The most obvious issue arising

from this feature would be the entropic cost of binding given the enormous loss of conformational freedom [22, 23]. Even though peptides are notoriously flexible, peptides as small as eight residues in length have been shown to exhibit secondary structure features or at the very least, have intramolecular interactions that in turn would increase the cost of reorganization prior to binding [24]. It would seem that on either ends of the scale of flexibility, the enthalpic and entropic costs of ligand reorganization (not even considering hydrophobic contributions) pose a significant obstacle in optimizing the binding free energy. All this again begs the question: Why are we interested in using peptides?

The hidden costs of reorganization are partially buffered by the fact that these large molecules can contain inherently spaced hydrogen bond networks that match those of their protein targets. Main-chain to main-chain hydrogen bonding networks of peptide–protein interactions are seen to typically dominate the enthalpic binding contributions of endogenous as well as synthetic peptide ligands [22]. The hydrogen bonding networks along with the usual suspects of salt bridging and hydrophobic surface interactions actually create a sizeable enthalpy of binding. Furthermore, the large surface area and extended hydrogen bond networks also allow peptide ligands to occupy shallower binding pockets on their protein targets. In summary, peptide–protein interactions involve more intermolecular interactions than a typical small molecule ligand, and while this can be advantageous for selectivity and binding affinity, it poses significantly more structural prediction challenges. These challenges will be discussed in later chapters with emphasis on implications of structure-based drug design and the methods used to search through the conformational space of both ligand and host.

1.2 Computer-Assisted Drug Design (CADD)

Moore's law: Moore's perception that the number of transistors on a microchip doubles every two years, though the cost of computers is halved, inferring that we can expect the speed and capability of our computers to increase every couple of years, and we will pay less for them.

Eroom's Law: The observation that drug discovery is becoming slower and more expensive over time, despite improvements in technology, a trend first observed in the 1980s. The cost of developing a new drug roughly doubles every nine years.

The total cost of drug development from discovery to approval is highly varied

but figures for the years 2016 to 2018 are estimated at anywhere between 800 million to 2.6 billion dollars [25]. The accuracy of these costs is questionable and based on companies developing multiple drugs at the same time. However, the magnitude of these figures is not up for debate. This incredible cost is also coupled with a development time spanning up to a decade (and in some cases even more). In accordance to Eroom's Law above, these figures are expected to become even higher in the future. Needless to say, there are significant implications outside of the profit margins of companies doing drug development. In a world where antibiotic resistance is growing and antibiotic drug development is stagnating, high development costs into drugs that are admittedly not the most profitable can only worsen the situation. Drug development costs are also intrinsically linked to other economic issues—rising healthcare costs for the public and the prohibitive costs for pharmaceutical startups are both examples of this.

Breakdowns of drug development costs indicate as much as one third of the total development costs are wrapped into pre-clinical discovery and development [26]. Discovery phases to identify new molecular entities for development are met with the challenge of the vastness of chemical space. It is estimated that the chemical space occupied by drug-like molecules (adhering to Lipinski's *Rule of Five* [27]) contains up to 10^{60} possible compounds [28]. Once a protein target has been validated, it is then the goal to cleverly carve out a selection of this immense space for further testing. Exhaustive testing through combinatorial chemistry and high throughput screening methods are a popular means for attempting to tackle this problem. Needless to say, this falls incredibly short. One may think computers are the solution to this problem, but even then, if we were to computationally evaluate each of the compounds in a 10^{60} chemical space with the most basic methods, this is still a highly intractable problem. This is one of the fundamental problem of drug discovery; accessing the few interesting compounds that contain our desired set of properties out of an unfathomable amount of atomic combinations. Figure 1.8 illustrates the current number of tractable compounds at the various stages of narrowing chemical space.

The argument for the use of CADD is not to completely replace the traditional medicinal chemist. The use of any tool that can potentially speed up the exploration of this chemical space in regards to how it's sampled as well as how it's tested is just another tool in the toolbox. The use of computers for the automated enrichment of chemical space such as chemical similarity searching and machine learning methods is an intensely studied field garnering much interest but lie outside the scope of this dissertation. For all publications presented in this dissertation involving a library of potential ligands, the compounds have been

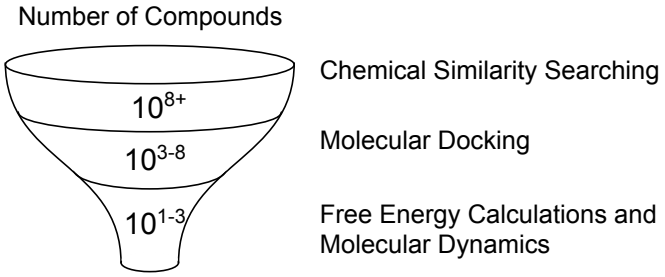


Figure 1.8: Narrowing Chemical Space with CADD Methods. From low computational cost to high computational cost, methods are used to funnel chemical space into a tractable number of testable compounds. The relative size of space for each method continues to grow as computational resources improve. Recently, docking experiments have hit the 10^8 mark for number of compounds docked on a single protein [29].

generated through a manual selection via rational design or resulting from another an *in vitro* high-throughput assay. Therefore, this dissertation is mostly primarily focussed on the testing and predictive applications of CADD. Two types of CADD are relevant here: ligand-based and structure-based methods.

1.2.1 Ligand-Based Design (LBDD)

Ligand-based design allows the prediction of a molecule's pharmacological activity by utilizing information about a molecule's physical features in reference to similar molecules with a known activity. The special and somewhat surprising feature of ligand-based approaches are that they do not require structural information about the binding location on the host protein. One of the most common forms of this type of prediction is a *quantitative structure activity relationship* (QSAR). QSARs aim to compartmentalize features of the molecular structure with respect to the overall activity of the molecule. This compartmentalization can be any number of physical attributes: Number of hydrogen bond acceptors, distance between two functional groups, the existence of a functional group, molecular weight, length of a particular alkyl chain, and a number of other molecular descriptors [30]. With the advent of machine learning, these sorts of intuitive physical parameters are replaced with convoluted relationships between atomic connections, and occupy a much higher dimension of parameterization [31].

Both QSARs and ligand-based machine learning approaches are essentially no more than complicated regression models fit to a set of experimental data. As a consequence, errors involving extrapolation to molecules far removed from the chemical space of the training set can be unpredictable. It turns out that

completely ignoring the geometry of the binding site or other chemical and physical properties can quickly lead to unpredictable cutoffs of predicted activity [32, 33]. Despite the clear limitations of ligand-based designs, QSARs have been especially useful in the past several decades for pre-clinical discovery leading into successful drug candidates and numerous examples in the literature can be found [34, 35]. As well, the addition of machine learning applications is incredibly promising and gaining grounds in a variety of drug discovery projects [36, 37].

Advantages of ligand-based design methods arise when the data exists to support the predictive models. In the cases where compounds for comparison have yet to be tested, we are left in a lurch. However, if we are structurally privileged and structural information of protein target exists, we can take the route of a structure-based design path. However, the use of ligand-based or structure-based methods are not exclusive, and in fact, there are several advantages to combining the two sets of methodologies in terms of the chemical space they are able to explore [38].

1.2.2 Structural-Based Design (SBDD)

Similar to ligand-based methods, the objective of structure-based methods is to design and optimize a compound to elicit a physiological response. However, structure-based methods utilize information of the biological target as a guide to compound design; a kind of space filling strategic placement of features with chosen intermolecular forces (See Figure 1.9).

As mentioned above, the power behind the lock and key concept of molecular recognition went largely unrealized until structural information of ligand-host systems could be characterized. It was not until the late 80s/early 90s that the first reported successes of drug development were partly attributed to a structure-based approach. Some of these first applications were focussed on inhibitor development for HIV proteases[40–42], and relied on various structural information including crystal structures of the apo-host protein, inferences about the binding site from previous ligand-based approaches, and crystal structures of other bound inhibitors.

1.3 Goals

In a general sense, this dissertation is focussed on the use of SDBB methods including the use of molecular docking, molecular dynamics, and combinations of the two. As such, the use of SBDD methods and the science and theory behind them are fully presented in a later chapter. However, before we explore these methods

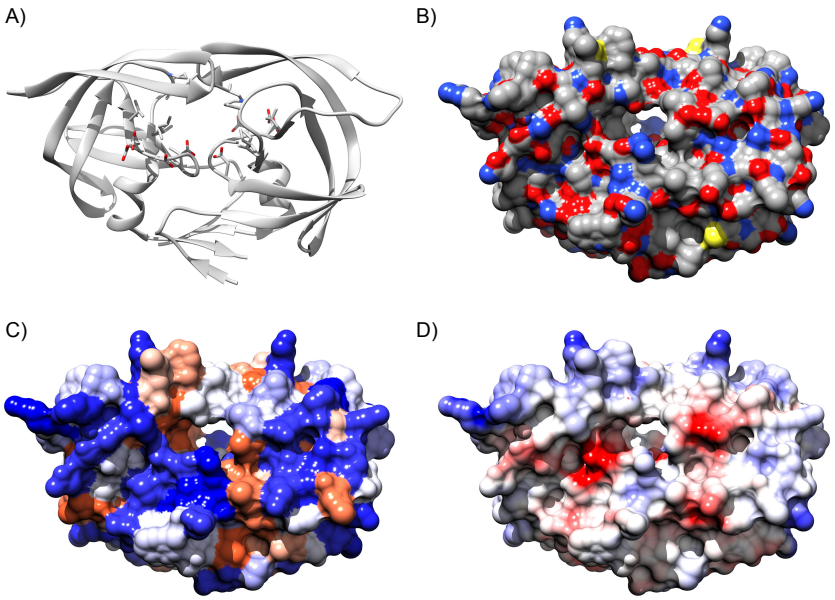


Figure 1.9: Structural Representations of a Host Protein. Using HIV protease as an example(PDB:4LL3 [39]), several structural representations of the protein are shown: (a) Secondary structure features, (b) atom types, (c) Location of hydrophobic residues (orange) and polar residues (blue), and (d) an electrostatic surface potential. Together, these representations lay out a map of potential interactions with a ligand and enable the rational design of a molecule for binding.

in detail, an introduction to the model systems studied in this thesis is warranted, specifically the CBX proteins and the inhibitor development efforts of the Hof group. The aim for the next chapter on the model systems is to provide context as to how SBDD methods are employed in this research but as well as the unique challenges involved in the CBX systems and how new methodology is required to understand the structure-activity relationships provided by experimental data.

The remaining content of this dissertation presents a chronology of collaborations and publications that explore specific CBX–peptide inhibitor complexes as well as a handful of other host–guest systems. The molecular modelling in each chapter uncovers a new facet and challenge associated with the systems at hand. For example, the first publication is a structural prediction problem where six potential binding sites on the Hen Egg White Lysozyme protein are present for a calixarene ligand. Through various docking, MD, and free energy methods, we were able to uncover the potential binding site. However, our initial work on this project was misdirected in that we were naive to the reorganizational energies of the host protein. These lessons learned and the changes to our methodology ultimately guided us to the development of our own structural prediction method also presented as a later chapter.

Bibliography

[1] Raymond U. Lemieux and Ulrike Spohr. How Emil Fischer was led to the lock and key concept for enzyme specificity. 203rd National Meeting of the American Chemical Society, Division of Carbohydrate Chemistry, San Francisco, California, april 5–10, 1992. In *Advances in Carbohydrate Chemistry and Biochemistry*, pages 1–20. Elsevier, 1994.

[2] Daniel E. Koshland. The Key–Lock Theory and the Induced–Fit Theory. *Angewandte Chemie International Edition in English*, 33(2324):2375–2378, January 1995.

[3] Jean-Pierre Changeux and Stuart Edelstein. Conformational selection or induced-fit? 50 years of debate resolved. *F1000 Biology Reports*, 3, September 2011.

[4] Qiang Cui and Martin Karplus. Allostery and cooperativity revisited. *Protein Science*, 17(8):1295–1307, August 2008.

[5] Denis Bucher, Barry J. Grant, and J. Andrew McCammon. Induced fit or conformational selection? The role of the semi-closed state in the maltose binding protein. *Biochemistry*, 50(48):10530–10539, dec 2011.

[6] Xing Du, Yi Li, Yuan-Ling Xia, Shi-Meng Ai, Jing Liang, Peng Sang, Xing-Lai Ji, and Shu-Qun Liu. Insights into protein–ligand interactions: Mechanisms, models, and methods. *International Journal of Molecular Sciences*, 17(2):144, January 2016.

[7] Sławomir J. Grabowski, W. Andrzej Sokalski, and Jerzy Leszczynski. The possible covalent nature of n-h···o hydrogen bonds in formamide dimer and related systems: an ab initio study. *The Journal of Physical Chemistry A*, 110(14):4772–4779, April 2006.

[8] J. E. Jones. On the determination of molecular fields. II. from the equation of state of a gas. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 106(738):463–477, October 1924.

[9] Mutasem Omar Sinnokrot and C. David Sherrill. Substituent effects in π–π interactions: Sandwich and T-shaped configurations. *Journal of the American Chemical Society*, 126(24):7690–7697, June 2004.

[10] Samuel Genheden, Oliver Kuhn, Paulius Mikulskis, Daniel Hoffmann, and Ulf Ryde. The normal-mode entropy in the MM/GBSA method: Effect of system truncation, buffer region, and dielectric constant. *Journal of Chemical Information and Modeling*, 52(8):2079–2088, August 2012.

[11] Samuel Genheden, Paulius Mikulskis, LiHong Hu, Jacob Kongsted, Pr Sderhjelm, and Ulf Ryde. Accurate predictions of nonpolar solvation free energies require explicit consideration of binding-site hydration. *Journal of the American Chemical Society*, 133(33):13081–13092, August 2011.

[12] Dennis A. Dougherty. The cation-π interaction. *Accounts of Chemical Research*, 46(4):885–893, December 2012.

[13] Justin P. Gallivan and Dennis A. Dougherty. A computational study of cation-pi interactions vs salt bridges in aqueous media: implications for protein engineering. *Journal of the American Chemical Society*, 122(5):870–874, February 2000.

[14] John D. Chodera and David L. Mobley. Entropy-enthalpy compensation: Role and ramifications in biomolecular ligand recognition and design. *Annual Review of Biophysics*, 42(1):121–142, May 2013.

[15] Victor Neduva, Rune Linding, Isabelle Su-Angrand, Alexander Stark, Federico de Masi, Toby J Gibson, Joe Lewis, Luis Serrano, and Robert B Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biology*, 3(12):e405, November 2005.

[16] Benjamin J Bruno, Geoffrey D Miller, and Carol S Lim. Basics and recent advances in peptide and protein drug delivery. *Therapeutic Delivery*, 4(11):1443–1467, November 2013.

[17] Raymond Behrendt, Peter White, and John Offer. Advances in fmoc solid-phase peptide synthesis. *Journal of Peptide Science*, 22(1):4–27, January 2016.

[18] Seok Hoon Hong, Yong-Chan Kwon, and Michael C. Jewett. Non-standard amino acid incorporation into proteins using escherichia coli cell-free protein synthesis. *Frontiers in Chemistry*, 2, June 2014.

[19] Tayebeh Farhadi and Seyed MohammadReza Hashemian. Computer-aided design of amino acid-based therapeutics: a review. *Drug Design, Development and Therapy*, Volume 12:1239–1254, May 2018.

[20] Jolene L. Lau and Michael K. Dunn. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & Medicinal Chemistry*, 26(10):2700–2707, June 2018.

[21] Fang Huang and Werner M. Nau. A conformational flexibility scale for amino acids in peptides. *Angewandte Chemie International Edition*, 42(20):2269–2272, May 2003.

[22] Nir London, Dana Movshovitz-Attias, and Ora Schueler-Furman. The structural basis of peptide-protein binding strategies. *Structure*, 18(2):188–199, February 2010.

[23] Benjamin J. Killian, Joslyn Yudenfreund Kravitz, Sandeep Somani, Paramita Dasgupta, Yuan-Ping Pang, and Michael K. Gilson. Configurational entropy in protein–peptide binding:. *Journal of Molecular Biology*, 389(2):315–335, June 2009.

[24] Bosco K. Ho and Ken A. Dill. Folding very short peptides using molecular dynamics. *PLoS Computational Biology*, 2(4):e27, 2006.

[25] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47:20–33, May 2016.

[26] Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, February 2010.

[27] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3):3–25, January 1997.

[28] Peter Kirkpatrick and Clare Ellis. Chemical space. *Nature*, 432(7019):823–823, December 2004.

[29] Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmachev, Brian K. Shoichet, Bryan L. Roth, and John J. Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, February 2019.

[30] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Wiley, September 2000.

[31] Angélica Nakagawa Lima, Eric Allison Philot, Gustavo Henrique Goulart Trossini, Luis Paulo Barbour Scott, Vinícius Gonçalves Maltarollo, and Kathia Maria Honorio. Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery*, 11(3):225–239, February 2016.

[32] Mark T.D. Cronin and T.Wayne Schultz. Pitfalls in QSAR. *Journal of Molecular Structure: THEOCHEM*, 622(1-2):39–51, March 2003.

[33] Gerald M. Maggiora. On outliers and activity cliffs: Why QSAR often disappoints. *Journal of Chemical Information and Modeling*, 46(4):1535–1535, July 2006.

[34] Tao Wang, Xin song Yuan, Mian-Bin Wu, Jian-Ping Lin, and Li-Rong Yang. The advancement of multidimensional QSAR for novel drug discovery - where are we headed? *Expert Opinion on Drug Discovery*, pages 1–16, June 2017.

[35] Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR modeling: Where have you been? where are you going to? *Journal of Medicinal Chemistry*, 57(12):4977–5010, January 2014.

[36] Alex P. Lind and Peter C. Anderson. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLOS ONE*, 14(7):e0219774, July 2019.

[37] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackerman, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, February 2020.

[38] Malgorzata N. Drwal and Renate Griffith. Combination of ligand- and structure-based methods in virtual screening. *Drug Discovery Today: Technologies*, 10(3):e395–e401, September 2013.

[39] K. Grantz Saskova, P. Rezacova, J. Brynda, M. Kozisek, and J. Konvalinka. Structure of wild-type HIV protease in complex with darunavir, April 2014.

[40] N. Roberts, J. Martin, D Kinchington, A. Broadhurst, J. Craig, I. Duncan, S. Galpin, B. Handa, J Kay, A Krohn, and al. et. Rational design of peptide-based HIV proteinase inhibitors. *Science*, 248(4953):358–361, April 1990.

[41] J Erickson, D. Neidhart, J VanDrie, D. Kempf, X. Wang, D. Norbeck, J. Plattner, J. Rittenhouse, M Turon, N Wideburg, and al. et. Design, activity, and 2.8 A crystal structure of a c2 symmetric inhibitor complexed to HIV-1 protease. *Science*, 249(4968):527–533, August 1990.

[42] Bruce D. Dorsey, Rhonda B. Levin, Stacy L. McDaniel, Joseph P. Vacca, James P. Guare, Paul L. Darke, Joan A. Zugay, Emilio A. Emini, and William A. Schleif. L-735, 524: The design of a potent and orally bioavailable HIV protease inhibitor. *Journal of Medicinal Chemistry*, 37(21):3443–3451, October 1994.

Chapter 2

Models

This chapter focusses on providing the background information on the protein systems presented in later chapters—specifically, the CBX proteins and their relevance as pharmaceutical targets. Throughout this thesis, simulation work involving the CBX proteins is largely aimed at providing insight into the structure–activity relationships (SAR) of peptidic inhibitors with the various CBX isoforms. As we will see in this chapter, selectivity between the CBX isoforms (of which there are several) has potential implications as both cancer therapeutics as well as chemical probes for studies involving stem cell differentiation. The study of CBX proteins in relation to disease states has garnered sufficient attention in that inhibitor development from a number of research groups has led to isoform-specific peptide-based ligands. A brief description of the current state of CBX inhibitor development and the challenges faced are described herein.

2.1 CBX Protein Biology

CBX proteins are associated with chromatin reorganization through the recognition of post-translational modifications (PTMs) on histone proteins and their interaction within a larger complex known as the Polycomb Repressive Complex 1 (PRC1). To best describe where CBX proteins fit into the big picture (both physically and functionally), let us take a bottom-up approach starting with the CBX substrate, chromatin.

DNA exists in a structural heirarchy beginning at the double helix wrapping around octamers of histone proteins (H2A, H2B, H3, H4) to form nucleosomes. (See Figure 2.1) These nucleosomes are connected by both DNA as well as an additional histone protein (H1). The sequence of these DNA-wrapped nucleosomes is known as chromatin, and depending on the structural modifications of

the histone proteins, can exist in either a condensed and less accessible structure known as heterochromatin, or a more "loose" and transcriptionally active form known as euchromatin [1]. These structural modifications are commonly referred to as post-translational modifications (PTMs) and include a variety of chemical changes. Of these changes, the most relevant to the CBX proteins include lysine methylation and ubiquitination. This concept that not only the DNA code but *how* DNA is presented to transcriptional mechanisms is an active field of study known as epigenetics. As one can imagine, the implications of controlling or at the very least understanding this complex and subtle control of DNA through PTMs and their related proteins appeals to pharmaceutical and general biology interests.

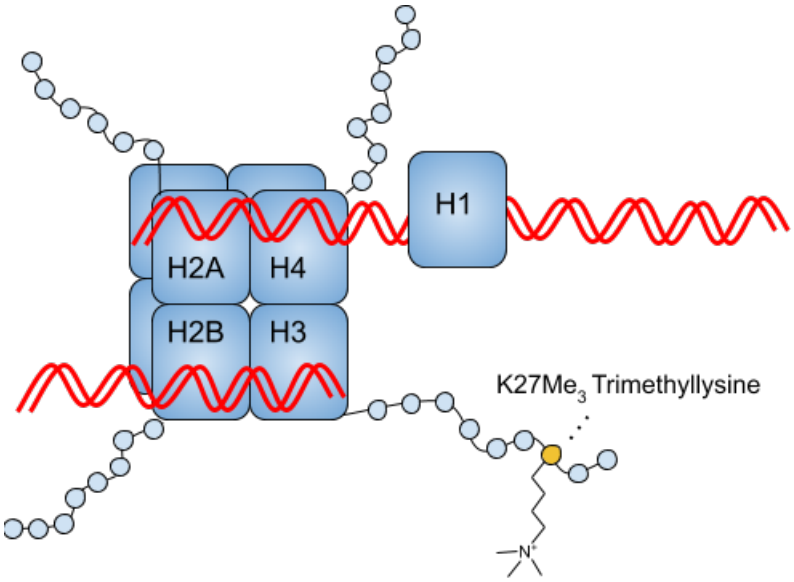


Figure 2.1: Post-translationally Modified Nucleosome. Chromatin structure showing trimethyllation site H3K27Me$_3$

One of the earliest families of proteins found to be involved in such chromatin modifications are those in the Polycomb Group (Pc) [2]. The Pc proteins form what are known as the Polycomb repressive complexes (PRC) of which two main forms exist. PRC2 functions primarily as a means to methylate lysine residues located on histone protein H3. Methylations on H3 lead to transcriptionally inactive portions of DNA and therefore PRC2 functions as a gene inactivator [3]. PRC1 has also traditionally assumed a role as a repressor through a PRC2 dependent ubiquitination of the H2A histone protein as illustrated in Figure 2.2. However, more recent insights into the diversity of the Pc proteins paints a more complicated picture in terms of structure and function. Various protein subunits of PCR1 can be swapped out (See Figure 2.3) creating a combinatorial arrangement

of 180 possible versions. Therefore, it's not surprising that the role of PCR1 is not just limited to a single function, but dependent on the particular combination of subunits [4] and extends beyond ubiquitination and methyl lysine recognition.

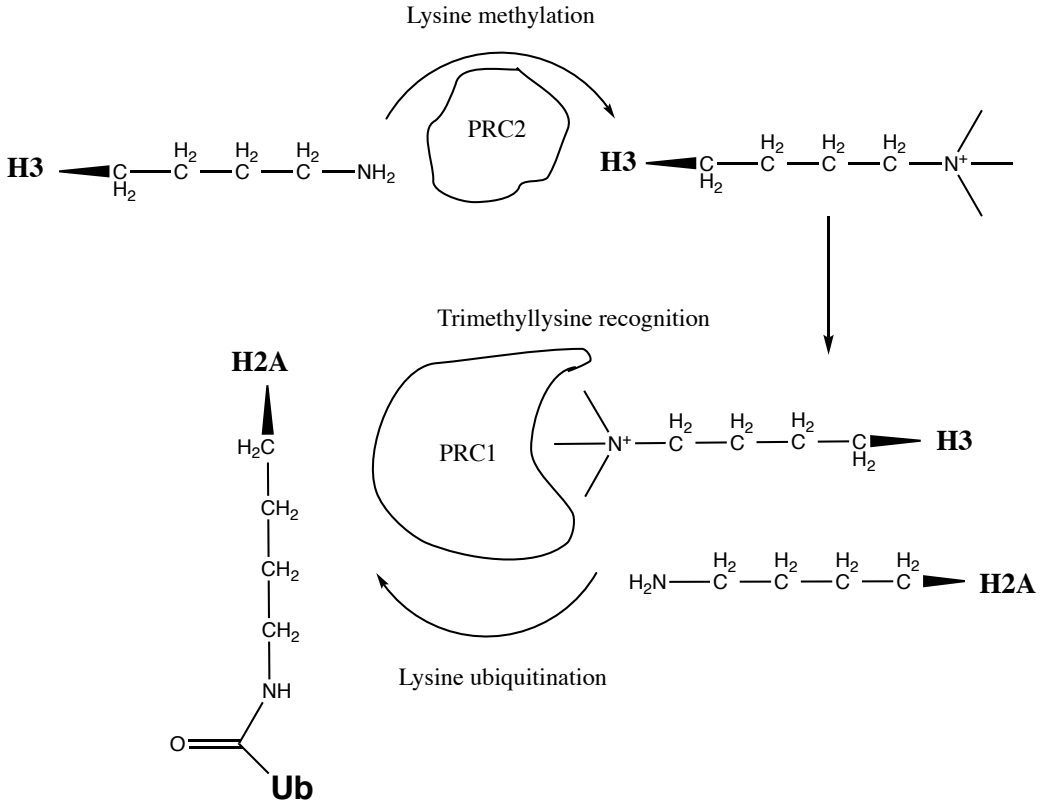


Figure 2.2: Classic PRC2 Dependent Ubiquitination via PRC1. Ubiquitination of H2A via a PRC2 dependent pathway. PRC2 is responsible for the trimethyllation of H3 which is later recognize by the CBX proteins on PRC1. RING subunits on PRC1 then ubiquitinate K119 on the H2A histone protein.

Of the four different PRC1 subunits, there is a particular interest in the CBX proteins due to correlations with with various disease states as well as their known physical function as the methyl lysine recognition portion of PRC1. The CBX protein sizes range between 251 aa (CBX7) and 560 aa (CBX2) containing two domains: the chromodomain and the polycomb domain. The chromodomain is a relatively conserved sequence throughout the isoforms with few distinctions between them, but seem to have large impacts on their form and function as observed in various knockout studies (See Table 2.1). The CBX chromodomain is approximately 50 amino acids in length and contains the trimethyllysine recognition site. For clarity, the models in this thesis refer to the CBX chromodomain when discussing the various CBX isoforms.

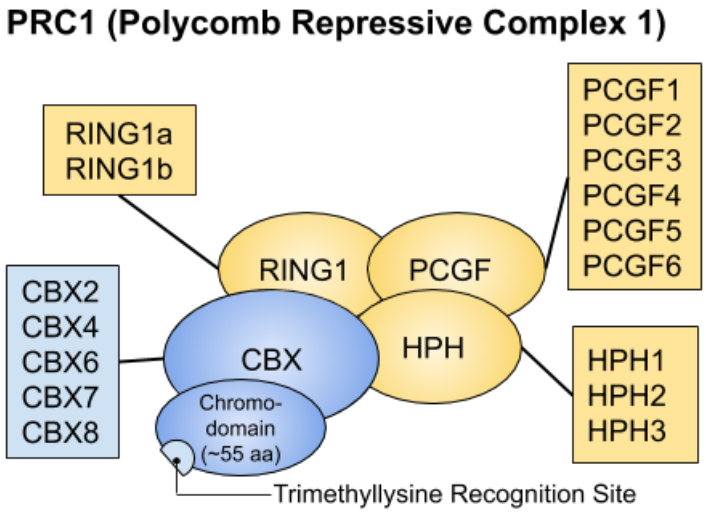


Figure 2.3: Polycomb Repressive Complex 1. The CBX Chromodomain contributes a small but important part to the PRC1 Complex. RING proteins (Really Interesting New Gene), PCGF (Polycomb Group RING Finger Protein), HPH (Human Polyhomeotic Homolog), and polycomb/chromo domains of CBX form the larger chromatin repressing PRC1 complex.

CBX Isoform	Knockout Observations on Mice Models	Recognition Activity	Studies
CBX2	Effects on sexual development Spleen and adrenal gland abnormalities Skeletal deformations	$H3K27Me_3$	[5] [6]
CBX4	Neonatal lethality Thymic hypoplasia	$H3K27Me_3$ $H3K9Me_3$	[7] [8]
CBX6	Decrease in body fat Metabolic defects Decreased heart weight	$H3K27Me_3$	[9]
CBX7	Increased body length Increased chance to develop liver and lung cancer	$H3K27Me_3$	[10]
CBX8	Abnormal cell physiology of marrow cells	$H3K27Me_3$	[11]

Table 2.1: CBX knockout Studies and Reported Trimethyllysine Recognition Sites. Phenotypic expression of various CBX isoform knockout mice and experimentally determined chromodomain recognition sites [4].

One thing that stands out in Table 2.1 is the overlapping recognition sites of the chromodomain. Despite overlap, the isoforms have different roles in cellular development in *in vivo* studies. Until recently, much of this work has relied on immunoprecipitation assays and other *in vitro* studies. Unfortunately, as more information becomes available about the CBX proteins, the knowledge gap appears to grow even larger. Conflicting information such as the presence of crystal

structures of CBX8 with H3K9[12], CBX6 association with proteins outside of the canonical PRC1[13], and the discovery of DNA binding sites on CBX8[14] cast a major shadow of doubt on the current understanding of the actual full role of CBX proteins. Furthermore, discrepancy between in vivo and in vitro associations of CBX8 highlight the importance of the consideration of CBX proteins in biologically relevant context [14]. Despite the functional complexity of these proteins, the observed correlations in both stem cell differentiation as well as disease development still stand (See Table 2.2). Therefore, the possibility of isoform-specific inhibitors as either a chemical probe into CBX functionality or as chemotherapeutic agents remains a worthy pursuit.

CBX Isoform	Disease Relation (Expression Levels)
CBX2	Breast cancer (Elevated) [15]
CBX4	Hepatocellular carcinoma (Elevated)[16]
CBX6	Gliobastoma (Declined) [17]
CBX7	Prostate cancer (Elevated) [18]
	Lymphoma (Elevated) [19]
	Gastric cancer (Elevated)[20]
	Lung cancer (Declined) [21]
	Colon cancer (Declined) [22]
CBX8	Glioblastoma (Elevated) [17]
	Breast cancer (Elevated) [23]

Table 2.2: CBX Isoforms and Associated Cancers. Both increased and decreased levels of CBX isoforms are tied to several cancer indications with apparent overlapping phenotypic expression.

2.2 Structural Challenges in Inhibitor Design for CBX Proteins

The family of CBX chromodomains associated with PRC1 exhibit high sequence similarity leading to multiple conserved features of the native peptide binding site. Furthermore, the differences in sequence between the isoforms are largely outside of the binding regions and likely impose energetic and structural complexities to binding not observable in crystallographic studies. To put it plainly, the CBX proteins are difficult targets from a structure-based drug design perspective. However, as we will see in the following chapters, the pursuit of isoform selectivity is not impossible—just immensely challenging. Throughout the chapters, two structural

similarities (See Figure 2.4) will be constantly referenced: (i) an aromatic cage consisting of a phenylalanine and two tryptophan residues with a preference for various alkylated lysine residues and (ii), a hydrophobic clasp consisting of valine and leucine residues wrapping over the bound ligand.

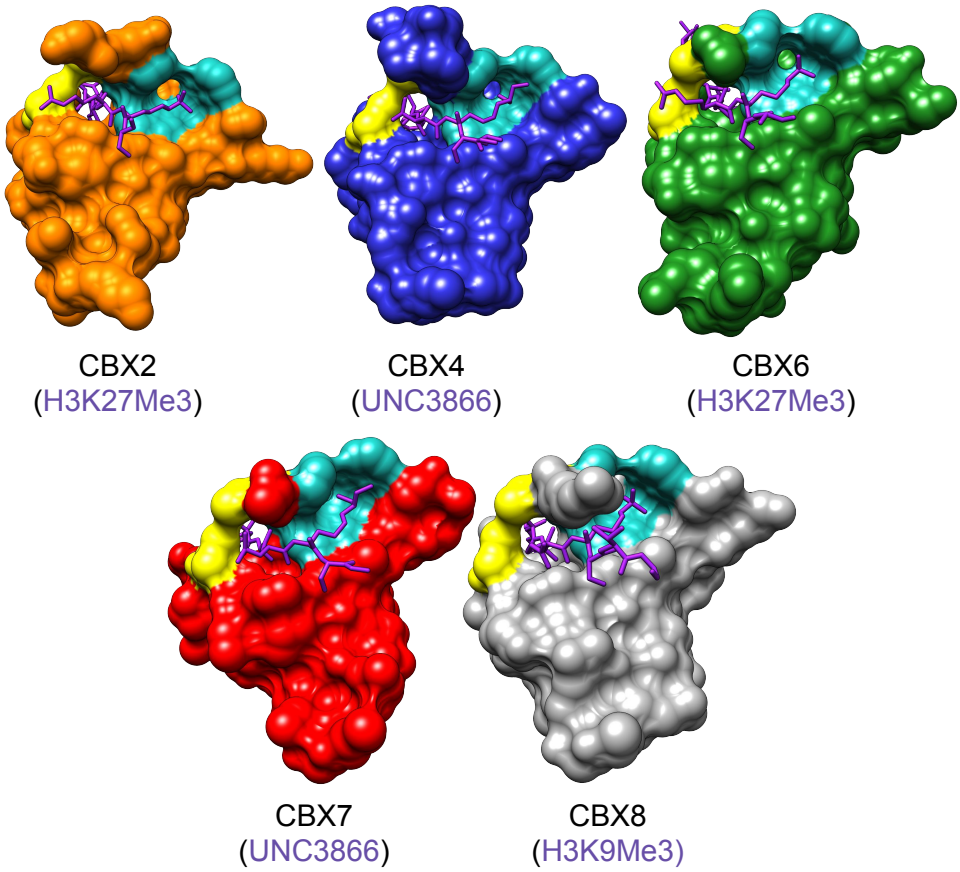


Figure 2.4: Polycomb Group CBX Chromodomain Structural Similarities. CBX2,4,6,7,8 all contain a trimethyllysine recognition pocket (teal) consisting of a a phenylalanine and two tryptophan residues, commonly referred to as the *aromatic cage*. Attached to the phenylalanine of this pocket, a clasp (yellow) containing hydrophobic residues valine and leucine wrap over the bound ligand, and is referred to as the *hydrophobic clasp*. Each isoform is presented along with a bound ligand (purple) containing an alkylated lysine residue in the aromatic cage. PDB access codes include CBX2 (3H91 [24]), CBX4 (5EPL [25]), CBX6 (3I90 [26]), CBX7 (5EPJ [27]), and CBX8 (3i91 [28]).

Figure 2.5: CBX Chromodomain Conserved Sequences. Sequences taken from PDB access codes presented in Figure 2.4. Highlighted teal features include residues contributing to the aromatic cage whereas yellow represents those involved in the hydrophobic clasp.

The sequences shown Figure 2.5 present another interesting challenge that isn't immediately apparent ? regions of ligand contact are *highly* similar in sequences. As well, dynamic features with respect to the binding event are also evident when the crystal structure is taken into consideration. For example, it apparent that the clasp is a dynamic feature and has to open and close upon binding. This is evident by the large steric clashes that would occur in trying to remove the ligand in the bound pose without changing the host structure. As the clasp is a dynamic feature, the proximity of the clasp residues to the aromatic cage phenylalanine suggests a concerted fit where the cage is optimally oriented when the clasp is properly in place. This induced-fit feature was found in molecular dynamics studies from our own research as well as others [29, 30]

To leverage this induced-fit mechanism, variations in pocket size between isoforms under the clasp have been exploited to create a tipping point for selectivity [29]. The pocket under the clasp is referred to as the –2 pocket due to the location of the ligand residue with respect to the ligand's trimethyllysine. Natural peptide ligands H3K9Me and K3K27Me3 present an alanine residue in this location. Different isoforms have been found to accept larger residues such as cyclopentyl groups and have been the basis for creating selectivity with isoforms like CBX8 [31]. Unfortunately, the –2 pocket like the other parts of the protein, is seen to exhibit flexibility. Direct placements of the ligands using computational methods on the crystal structures produces large steric clashes, whereas from both experimental and more advanced molecular simulations, ligands are seen to fit under the clasp.

Along with the –2 pocket, regions containing a continued hydrogen bonding network with the natural ligand known as the β groove and extended β -groove are also the focus for rational design (See Figure 2.6). However, reasons for the binding affinity created and lost by ligand substitutions in this region are still unclear at this time and are potentially subject to non-additive effects caused by allosteric changes in the protein [31].

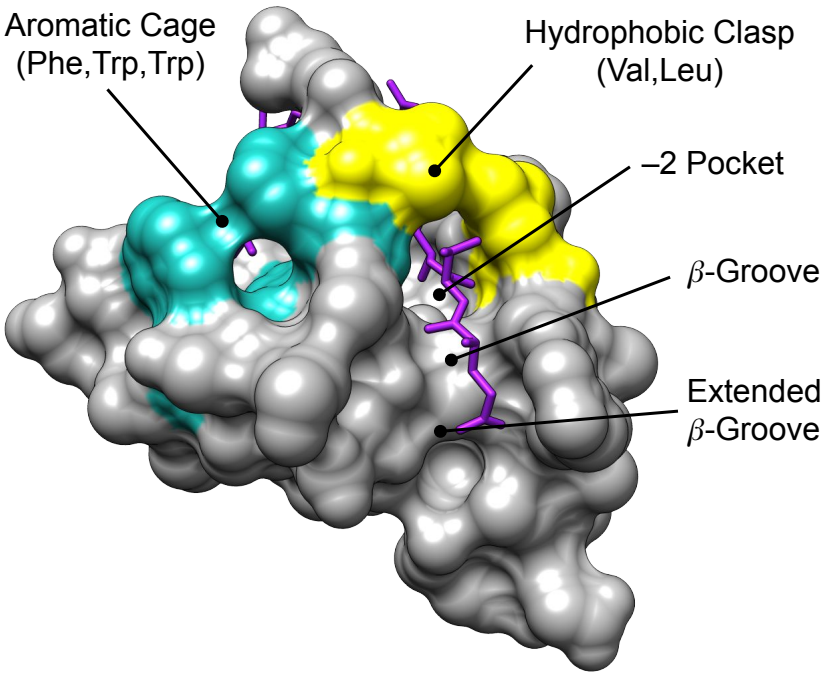


Figure 2.6: Crystal Structure of CBX8 bound to H3K9Me3 Peptide. Various regions of the CBX proteins have been the focus of rational ligand design. Regions under the clasp are exploited through steric bulk, whereas the β groove regions are much more unclear with respect to binding energy contributions and preferred ligand binding geometries.

In summary, the exact role of CBX proteins in human biology has yet to be fully defined. However, the impact of CBX proteins in cellular development and disease cannot be ignored and the pursuit of inhibitors is a worthy cause. The CBX proteins themselves are as challenging to model and target as they are functionally complex. From previous molecular simulations and experimental work compared to crystal structures, we see that CBX protein binding events are riddled with the classic strifes of induced-fit mechanisms. To model these proteins successfully, considerations of full protein flexibility need to addressed. In the following chapter, we will discuss the computational methods used to tackle such problems and create a structural prediction method fit for the rational design of CBX protein inhibitors.

Bibliography

[1] Taiping Chen and Sharon Y. R. Dent. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nature Reviews Genetics*, 15(2):93–106, December 2013.

[2] T C James and S C Elgin. Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. *Molecular and Cellular Biology*, 6(11):3862–3872, November 1986.

[3] J.N. Nichol, D. Dupéré-Richer, T. Ezponda, J.D. Licht, and W.H. Miller. H3k27 methylation. In *Advances in Cancer Research*, pages 59–95. Elsevier, 2016.

[4] Jesús Gil and Ana O'Loghlen. PRC1 complex diversity: where is it taking us? *Trends in Cell Biology*, 24(11):632–641, November 2014.

[5] Yuko Katoh-Fukui, Reiko Tsuchiya, Toshihiko Shiroishi, Yoko Nakahara, Naoko Hashimoto, Kousei Noguchi, and Toru Higashinakagawa. Male-to-female sex reversal in M33 mutant mice. *Nature*, 393(6686):688–692, June 1998.

[6] Yuko Katoh-Fukui, Kanako Miyabayashi, Tomoko Komatsu, Akiko Owaki, Takashi Baba, Yuichi Shima, Tomohide Kidokoro, Yoshiakira Kanai, Andreas Schedl, Dagmar Wilhelm, Peter Koopman, Yasushi Okuno, and Ken ichirou Morohashi. Cbx2, a polycomb group gene, is required forSryGene expression in mice. *Endocrinology*, 153(2):913–924, February 2012.

[7] Nuno Miguel Luis, Lluis Morey, Stefania Mejetta, Gloria Pascual, Peggy Janich, Bernd Kuebler, Guglielmo Roma, Elisabete Nascimento, Michaela Frye, Luciano Di Croce, and Salvador Aznar Benitah. Regulation of human epidermal stem cell proliferation and senescence requires polycomb- dependent and -independent functions of CBX4. *Cell Stem Cell*, 9(3):233–246, September 2011.

[8] B. Liu, Y.-F. Liu, Y.-R. Du, A. N. Mardaryev, W. Yang, H. Chen, Z.-M. Xu, C.-Q. Xu, X.-R. Zhang, V. A. Botchkarev, Y. Zhang, and G.-L. Xu. Cbx4 regulates the proliferation of thymic epithelial cells and thymus function. *Development*, 140(4):780–788, January 2013.

[9] William C. Skarnes, Barry Rosen, Anthony P. West, Manousos Koutsourakis, Wendy Bushell, Vivek Iyer, Alejandro O. Mujica, Mark Thomas, Jennifer

Harrow, Tony Cox, David Jackson, Jessica Severin, Patrick Biggs, Jun Fu, Michael Nefedov, Pieter J. de Jong, A. Francis Stewart, and Allan Bradley. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351):337–342, June 2011.

[10] K. White et al. Jacqueline. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell*, 154(2):452–464, July 2013.

[11] Jiaying Tan, Morgan Jones, Haruhiko Koseki, Manabu Nakayama, Andrew G. Muntean, Ivan Maillard, and Jay L. Hess. CBX8, a polycomb group protein, is essential for MLL-AF9-induced leukemogenesis. *Cancer Cell*, 20(5):563–575, November 2011.

[12] Lilia Kaustov, Hui Ouyang, Maria Amaya, Alexander Lemak, Nataliya Nady, Shili Duan, Gregory A. Wasney, Zhihong Li, Masoud Vedadi, Matthieu Schapira, Jinrong Min, and Cheryl H. Arrowsmith. Recognition and specificity determinants of the human cbx chromodomains. *Journal of Biological Chemistry*, 286(1):521–529, November 2010.

[13] Alexandra Santanach, Enrique Blanco, Hua Jiang, Kelly R. Molloy, Miriam Sansó, John LaCava, Lluis Morey, and Luciano Di Croce. The polycomb group protein CBX6 is an essential regulator of embryonic stem cell identity. *Nature Communications*, 8(1), November 2017.

[14] Katelyn E Connelly, Tyler M Weaver, Aktan Alpsoy, Brian X Gu, Catherine A Musselman, and Emily C Dykhuizen. Engagement of DNA and H3K27me$_3$ by the CBX8 chromodomain drives chromatin association. *Nucleic Acids Research*, 47(5):2289–2305, December 2018.

[15] S. Zheng, P. Lv, J. Su, K. Miao, H. Xu, and M. Li. Overexpression of CBX2 in breast cancer promotes tumor progression through the PI3K/AKT signaling pathway. *Am J Transl Res*, 11(3):1668–1682, 2019.

[16] Boqing Wang, Jianjun Tang, Dan Liao, Gang Wang, Meifang Zhang, Yi Sang, Jingying Cao, Yuanzhong Wu, Ruhua Zhang, Shengping Li, Wei Ding, Guoqing Zhang, and Tiebang Kang. Chromobox homolog 4 is correlated with prognosis and tumor cell growth in hepatocellular carcinoma. *Annals of Surgical Oncology*, 20(S3):684–692, August 2013.

[17] Gang Li, Charles Warden, Zhaoxia Zou, Josh Neman, Joseph S. Krueger, Alisha Jain, Rahul Jandial, and Mike Chen. Altered expression of polycomb

group genes in glioblastoma multiforme. *PLoS ONE*, 8(11):e80970, November 2013.

[18] David Bernard, Juan F Martinez-Leal, Sian Rizzo, Dolores Martinez, David Hudson, Tapio Visakorpi, Gordon Peters, Amancio Carnero, David Beach, and Jesus Gil. CBX7 controls the growth of normal and tumor-derived prostate cells by repressing the ink4a/arf locus. *Oncogene*, 24(36):5543–5551, May 2005.

[19] C. L. Scott, J. Gil, E. Hernando, J. Teruya-Feldstein, M. Narita, D. Martinez, T. Visakorpi, D. Mu, C. Cordon-Cardo, G. Peters, D. Beach, and S. W. Lowe. Role of the chromobox protein CBX7 in lymphomagenesis. *Proceedings of the National Academy of Sciences*, 104(13):5389–5394, March 2007.

[20] Xiao-Wei Zhang, Li Zhang, Wei Qin, Xiao-Hong Yao, Lei-Zhen Zheng, Xin Liu, Jin Li, and Wei-Jian Guo. Oncogenic role of the chromobox protein CBX7 in gastric cancer. *Journal of Experimental & Clinical Cancer Research*, 29(1):114, 2010.

[21] Floriana Forzati, Antonella Federico, Pierlorenzo Pallante, Adele Abbate, Francesco Esposito, Umberto Malapelle, Romina Sepe, Giuseppe Palma, Giancarlo Troncone, Marzia Scarfò, Claudio Arra, Monica Fedele, and Alfredo Fusco. CBX7 is a tumor suppressor in mice and humans. *Journal of Clinical Investigation*, 122(2):612–623, February 2012.

[22] Pierlorenzo Pallante, Luigi Terracciano, Vincenza Carafa, Sandra Schneider, Inti Zlobec, Alessandro Lugli, Mimma Bianco, Angelo Ferraro, Silvana Sacchetti, Giancarlo Troncone, Alfredo Fusco, and Luigi Tornillo. The loss of the CBX7 gene expression represents an adverse prognostic marker for survival of colon carcinoma patients. *European Journal of Cancer*, 46(12):2304–2313, August 2010.

[23] Sang Hyup Lee, Soo-Jong Um, and Eun-Joo Kim. CBX8 suppresses sirtinol-induced premature senescence in human breast cancer cells via cooperation with SIRT1. *Cancer Letters*, 335(2):397–403, July 2013.

[24] M.F. Amaya, M. Ravichandran, P. Loppnau, I. Kozieradzki, A.M. Edwards, C.H. Arrowsmith, J. Weigelt, C. Bountra, A. Bochkarev, J. Min, and H. Ouyang and. Crystal structure of the complex of human chromobox homolog 2 (CBX2) and h3k27 peptide, August 2009.

[25] Y. Liu, W. Tempel, J.R. Walker, J.I. Stuckey, B.M. Dickson, L.I. James, S.V. Frye, C. Bountra, C.H. Arrowsmith, A.M. Edwards, and J. Min and. Crystal structure of chromodomain of CBX4 in complex with inhibitor UNC3866, December 2015.

[26] M.F. Amaya, M. Ravichandran, P. Loppnau, I. Kozieradzki, A.M. Edwards, C.H. Arrowsmith, J. Weigelt, C. Bountra, A. Bochkarev, J. Min, and H. Ouyang and. Crystal structure of human chromobox homolog 6 (CBX6) with h3k27 peptide, September 2009.

[27] Y. Liu, W. Tempel, J.R. Walker, J.I. Stuckey, B.M. Dickson, L.I. James, S.V. Frye, C. Bountra, C.H. Arrowsmith, A.M. Edwards, and J. Min and. Crystal structure of chromodomain of CBX7 in complex with inhibitor UNC3866, December 2015.

[28] M.F. Amaya, M. Ravichandran, P. Loppnau, I. Kozieradzki, A.M. Edwards, C.H. Arrowsmith, J. Weigelt, C. Bountra, A. Bochkarev, J. Min, and H. Ouyang and. Crystal structure of human chromobox homolog 8 (CBX8) with h3k9 peptide, September 2009.

[29] Natalia Milosevich, Michael C. Gignac, James McFarlane, Chakravarthi Simhadri, Shanti Horvath, Kevin D. Daze, Caitlin S. Croft, Aman Dheri, Taylor T. H. Quon, Sarah F. Douglas, Jeremy E. Wulff, Irina Paci, and Fraser Hof. Selective inhibition of CBX6: A methyllysine reader protein in the polycomb family. *ACS Medicinal Chemistry Letters*, 7(2):139–144, December 2015.

[30] Jacob I Stuckey, Bradley M Dickson, Nancy Cheng, Yanli Liu, Jacqueline L Norris, Stephanie H Cholensky, Wolfram Tempel, Su Qin, Katherine G Huber, Cari Sagum, Karynne Black, Fengling Li, Xi-Ping Huang, Bryan L Roth, Brandi M Baughman, Guillermo Senisterra, Samantha G Pattenden, Masoud Vedadi, Peter J Brown, Mark T Bedford, Jinrong Min, Cheryl H Arrowsmith, Lindsey I James, and Stephen V Frye. A cellular chemical probe targeting the chromodomains of polycomb repressive complex 1. *Nature Chemical Biology*, 12(3):180–187, January 2016.

[31] Sijie Wang, Kyle E. Denton, Kathryn F. Hobbs, Tyler Weaver, James M. B. McFarlane, Katelyn E. Connelly, Michael C. Gignac, Natalia Milosevich, Fraser Hof, Irina Paci, Catherine A. Musselman, Emily C. Dykhuizen, and Casey J. Krusemark. Optimization of ligands using focused DNA-encoded li-

braries to develop a selective, cell-permeable CBX8 chromodomain inhibitor. *ACS Chemical Biology*, December 2019.

Chapter 3

Methods in Structure-Based Drug Design

"All models are wrong, but some are useful." — George Box

This chapter aims to provide the background information behind some of the current methodologies in structure-based drug design (SBDD) and more specifically, how they are applied to the systems studied in this thesis. A background on molecular dynamics, molecular docking, and combined approaches are included as they serve as both independent methods used throughout this thesis, but also as tools within a more complex combined method named "SLICE" which is presented later chapters.

To give credit where it is due, SBDD methods owe their existence to protein crystallography. Structural information from experimental methods has allowed the SBDD field to flourish alongside the growing availability of protein structures. Even when there are no structures for a given target, homology modelling algorithms to generate a target structure are almost entirely based off the sequence-to-structure correlations from large protein structure databases [1]. However, simulation tools such as molecular dynamics (MD) reveal additional information about a target proteins; specifically, as dynamic, hydrated, and complicated pieces of biological machinery [2]. Dynamic information about target structures is realized as the frontier challenge for SDBB and provide us with better starting points regarding predictions of binding sites and the ability to rank libraries of ligands within them. It's no surprise that the implementation of these techniques (MD and docking) is ubiquitous in the CADD community and great strides are being made to push us further. For instance, there are dozens of known molecular docking techniques that exist today [3].

With so much focus in SDBB and the growing availability of target proteins for

a number of untreated diseases, how is it that these techniques have not noticeably curbed *Eroom's Law* (see Chapter 1), despite computational resources becoming more abundant? How much progress has been made and where do we need to improve? To understand the current direction of SBDD methods, we must first take a look under the hoods of these techniques.

3.1 Molecular Dynamics

Molecular dynamics (MD) is a powerful simulation tool with broad applications across physics, biology, and chemistry. Like many theoretical applications, the concept was born before it could be fully realized into the usable tool it is today. Some of the first MD simulations of biological events were published as early as 1975 by recent nobel laureates, Michael Levitt and Arieh Warshel. Their work was aimed at simulating folding events of linear polypeptides—requiring significant approximations to be computationally feasible at the time.[4] Since then, MD simulations have tackled increasingly complex systems at larger and larger timescales. Milestones of MD such as simulations of an entire viral capsid of the satellite tobacco mosaic virus (1 M atom, 50 ns, 2006)[5] and the first observed ligand binding event from random position (35 µs, 2011)[6][1] foreshadow fully atomistic approaches to currently intractable problems.

The simulation method is based on a classical mechanical approach and describes systems at the atomic level but is still within the current computational resources to simulate large biomolecules. MD is often described as a "ball and spring" model; an underwhelming description that detracts from the power of this simulation tool. The ball and spring analogy comes from the fact that there are no electrons in MD simulations and electronic properties of the atoms are described as point charges on the nucleii. Yet, an enormous effort over several decades of parameterization and software design has made MD an accessible and reliable simulation tool—assuming it is used correctly. In assessing whether molecular dynamics is the right tool for a job in SDBB, let us consider the timescale on which features in a potential protein-ligand event occur [7].

The time scales illustrated in Figure 3.1 are derived from observables from spectroscopic techniques such as NMR and represent time averaged data of many interactions at once. The molecular dynamics timescale may appear applicable here (barely in the case of ligand binding), but has the severe disadvantage of being a microscopic simulation of what we hope to observe as ensemble average

[1] Video of this process occurring: https://pubs.acs.org/doi/abs/10.1021/ja202726y

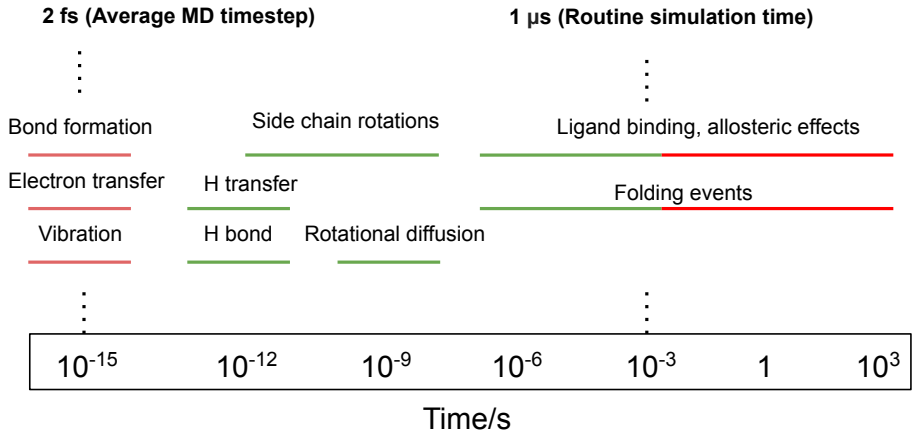


Figure 3.1: Protein Event Timescale. Various protein motions and solution phase processes compared to conventional accessible molecular dynamics timeframes.

properties. This microscopic take on simulation presents a number of hurdles with regards to how we sample within MD. As an example, Pan et al. studied a number of protein–protein interactions within the experimental timescale shown above and observed only a fraction of the parallel simulations to converge to the expected structure [8]. Clearly, questions such as "how many replicate simulations?" and "how long do they need to be?" need to be addressed when using MD.

Assuming processing speed and resources are not a problem (solving the sampling problem above), we are still left with issues regarding the accuracy of the simulation. Remembering that MD is an approximated method with no electrons, proper descriptions of the atoms and the intermolecular forces driving the simulation is still an active field of research, with new descriptions and simulation methods underway. The limitations of translating fully electronic effects such as reactivity, electron transfer, and polarization (only recently implemented in MD packages) into a mechanical force field are just some of the current obstacles. Even if parameterized properly, descriptions of molecular species outside of the training set can fall short. For example, many force fields are parameterized to canonical biological residues such as the standard amino acids and classic nucleotides. Inclusion of non-standard amino acids such a methyllated lysines compromise the integrity of the force field being used—this specific example will be discussed in later chapters. For now, let's dive into how the approximations in MD are applied and discuss the mechanics of how the simulations model atoms in time.

3.1.1 Theoretical Background

Molecular Mechanics: Atom Positions to Potential Energies

The term force field refers to the parameterized features of interacting atoms at both the intra and intermolecular level (the types of balls and the strengths of the springs). These collections of parameters describe the potential energy of a molecular system given its atomic coordinates—similar to those discussed in Chapter 1—and come in many variations. To name just a few, CHARMM22, CHARMM22*, Amber ff03, Amber ff03*, Amber ff09SB-ILDN, OPLS, GAFF, SMIRNOFF, are all examples of collections of parameterized potential energy functions and atom types used to calculate the forces on each atom in a simulation.

The variation between force fields is due both to *how* they are parameterized, but also *what* they are parameterizing. Forcefields are built using a variety of experimental and quantum mechanical data to which they are fitted to. The inclusion of some molecular properties such as polarization occur in only a number of forcefields [9]. Reactivity in forcefields is even further specialized and exists in just a handful of proof-of-concept examples [10]. However, the the classic potential functions are found throughout most force fields.

Intramolecular forces (Bonded potentials):

$$U_{\text{Stretching}} = \frac{1}{2} \sum_{\text{Bonds}} k_{ij}^{r} (r_{ij} - r_{eq})^2 \tag{3.1}$$

$$U_{\text{Bending}} = \frac{1}{2} \sum_{\text{BondAngles}} k_{ijk}^{\theta} (\theta_{ijk} - \theta_{eq})^2 \tag{3.2}$$

$$U_{\text{Torsions}} = \frac{1}{2} \sum_{\text{TorsionAngles}} \sum_{n} k_{ijkl}^{\phi,n} \left[1 + \cos(n\phi_{ijkl} - \psi_n)\right], \tag{3.3}$$

where the various k terms represent the magnitude of energy depending on the atom types involved and r_{eq}, θ_{eq}, and ψ_n represent equilibrium distances or angles representing minima for the potential energies. Terms are halved in this case to avoid double counting of the potential energies. See Figure 3.2 below for examples of the various distances and angles.

Intermolecular forces (Non-bonded potentials):

$$U_{\text{Lennard-Jones}} = \frac{1}{2} \sum_{ij} \varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \tag{3.4}$$

where ε_{ij} represents the well depth, r_{ij} as the interatomic distances, and σ_{ij} is the distance between atoms at the well depth.

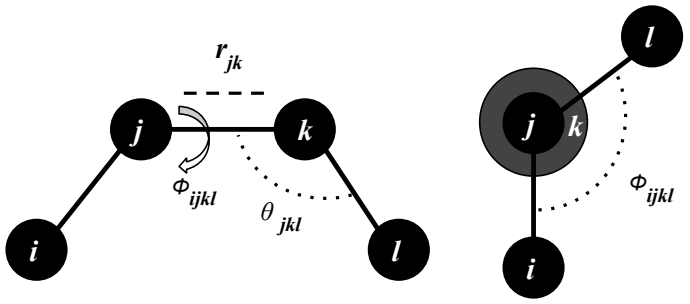


Figure 3.2: Atomistic diagram of intramolecular forces

$$U_{\text{Electric}} = \frac{1}{2}\sum_{ij} \varepsilon_{ij} \frac{q_i q_j}{\varepsilon_o r_{ij}}, \tag{3.5}$$

where ε is the effective dielectric constant and q_i,q_j, and r_{ij} represent partial charges of atoms i and j and their respective interatomic distance.

Newton's Law of Motion: Potential Energies to Forces:

The functions listed above describe the potential energy of the atomic interactions in the simulation. Given a potential energy experienced on a particle, we can use that information to calculate the acceleration imparted onto it using Newton's equations of motion:

$$\boldsymbol{F}_i = m_i \boldsymbol{a}_i, \tag{3.6}$$

where $\boldsymbol{F}_i$ is the force exerted on an atom, a_i is the acceleration of the atom, and m_i is the mass. To re-write this in terms of atomic positions, we can express this as the following:

$$\boldsymbol{F}_i = \frac{\mathrm{d}^2 \boldsymbol{r}_i}{\mathrm{d}t^2} m_i \tag{3.7}$$

To now relate the force on the atoms to potential energy, can express the force as the gradient of the potential energy with respect to position:

$$\boldsymbol{F}_i = -\frac{\mathrm{d}U}{\mathrm{d}\boldsymbol{r}_i}, \tag{3.8}$$

where U is the potential energy on the atom i at position r. By combining 3.7 and 3.8, we produce a differential equation that describes the acceleration of a particle

related to the potential energy of its position:

$$-\frac{\mathrm{d}U}{\mathrm{d}\boldsymbol{r}_i} = \frac{\mathrm{d}^2\boldsymbol{r}_i}{\mathrm{d}t^2}m_i \tag{3.9}$$

With a bit of shuffling and using U as the potential energy functions above dependent on the atomic coordinates, $U(r)$, we can rearrange this equation to represent the acceleration on a particle as:

$$\boldsymbol{a}_i(t) = -\frac{dU(r(t))}{m_i d\boldsymbol{r}_i(t)} \tag{3.10}$$

where this acceleration can be then used to recalculate both the velocities and positions of the particles at time t shown in the next section.

Integration: Forces to New Atom Positions:

Numerical methods for solving the many-body problem utilized in molecular dynamics are termed integrators, of which several flavours exist. The purpose of integrators is to approximately inch forward the atomic positions in the simulation by utilizing the forces/potential energies calculated on each atom. Presented below is the *velocity Verlet* approach [11] using the following equations of motion:

$$r_i(t+\delta t) = r_i(t) + v_i(t)\delta t + \frac{1}{2}a_i\delta t^2 \tag{3.11}$$

$$v_i(t+\delta t) = v_i(t) + \frac{1}{2}[a_i(t) + a(t+\delta t)]\delta t, \tag{3.12}$$

where $r(t)$,$v(t)$, and $a(t)$, represent position, velocity, of acceleration of each atom at time t and, in the case of $t+\delta t$, at the following time step. From Eqn. 3.10, we can calculate the acceleration needed for the velocity calculation above.

Now, with information about the acceleration, velocity, and initial positions, the Velocity Verlet approach can then be applied stepwise in the following manner:

i) Calculate $x(t+\delta t)$ using existing velocity and accelerations
ii) Calculate $a(t+\delta t)$ from the potential functions, $U(r)$
iii) Calculate new $v(t+\delta t)$
iii) Repeat

This algorithm is just one approach, and will contain errors relating to changes of the potential energy midway through the time step. For this reason, variations

of the algorithm that implement half steps and forward and backward approaches to the approximations are also found, such as the *half-step velocity Verlet* and the *leap frog* algorithm [12]. The errors occurring between time steps are the result of overshooting or undershooting the trajectories of the atoms. As the potential functions are only calculated at distinct times, the changes in potential energy mid-step are missed along with any influence on the change in acceleration between time steps. Along with this error in the time step calculation, the accuracy of the simulation is also affected by how the energies of the molecules are distributed for a given temperature.

Physical Relevancy in Simulations

As we have briefly covered the drivers for motion in molecular dynamics, it is also worth mentioning what makes MD a thermodynamically relevant simulation. More specifically, what role does temperature play in the simulation and how accurate is it? Without temperature, you can imagine starting a molecular dynamics simulation without any atomic velocities. The only energy in the system would be the potential energies between atoms as defined by the forcefield. Allowing the simulation to progress would generate velocities and relax the potential energy of the system. The remaining velocities would be the only kinetic energy in the system, and due to errors in the stepwise movement of atoms, could significantly diminish or climb rapidly. To address these issues, the use of a thermostat is required to ensure that the kinetic energy of the system is representative of the temperature the user desires. The use of a thermostat manages the velocities of the atoms such that the sum of kinetic energy is representative of the temperature:

$$\frac{3}{2}NkT = \sum_i \frac{m_i v_i^2}{2}, \tag{3.13}$$

where N is the number of particles in the system, T is temperature, k is the Boltzmann constant, and m and v are mass and velocity.

The actual monitoring and adjustment of the velocities of the atoms can be done using a variety of methods [13–15] such as the temperature exchange with a bath. The same applies to barostats in MD. The actual mechanisms of these tools are not important for this discussion, but it is worth acknowledging they exist as a key component in the accuracy of an MD simulation.

Along with temperature control, there are a number of other issues that are tackled by clever routine maintenance of the atom positions and velocities. One of the more amusing problems known as the "flying ice cube"[16] results when there

is an overall non-zero momentum in the system.

$$P = \sum_{i=1}^{N} m_i \mathbf{v}_i = 0 \tag{3.14}$$

Net momentum in any direction causes the simulation box to translate in that direction and on top of this, the potential energies from vibrations and rotations are eventually all converted into translational energy. The net result is a frozen cube (see Figure 3.3) flying through space with the kinetic energy proportional to the set temperature. Amusing, right? Therefore it is important at the beginning of the simulation when initial velocities are assigned as well as periodically through the simulation, to somehow adjust velocities to produce a net zero momentum. Despite these approximations and the dangers of flying ice cubes, MD is still a workhorse in SDBB with numerous applications beyond simple movies of ligands interacting with proteins.

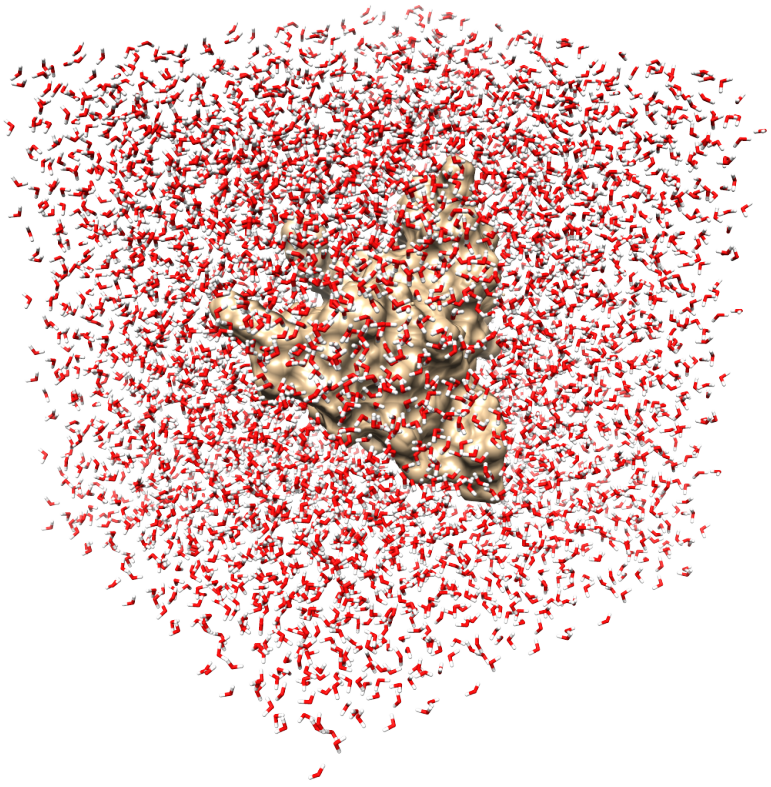


Figure 3.3: Molecular dynamics trajectories are set up to initially include the solute, followed by additional counterions, and finally the solvent. Periodic boundary conditions are then applied and the system is ready for simulation.

3.1.2 Molecular Dynamics as a Tool in CADD

Although amusing to watch, the use of MD as a tool in SDBB is seldom a straight forward simulation of start-to-finish protein-ligand binding. Rather, to investigate ligand activity, the trajectories produced in MD are used as *post facto* approaches, requiring some level of structural knowledge of the binding mode. Successful unbiased approaches of protein–ligand binding are amazing and rare, and currently limited to only a small handful of examples [6, 17]. Given what we have discussed above in Fig 3.1, the need for a predicted bound pose as a starting point should not come as a surprise as the timescale of which ligand binding occurs is generally outside the accessible simulation time for average researchers. In the cases where bound poses are known, molecular dynamics runs for small libraries of compounds are within the realm of possibility, but still pose a problem regarding ergodic sampling and how to generate any sort of meaningful simulation data. Therefore, as a primary *in silico* screening tool involving hundreds of thousands of compounds, MD is currently unfit as the choice methodology. Despite the timescale limitations, MD is still widely used in CADD [18]. So what is the useful information we get out of MD if we already need to know the mode of binding?

Free Energies of Binding

The force field energy description of the atoms in an MD simulation lends itself to a set of useful observables within SBDD. Picking apart the energetic contributions of certain atoms and monitoring the total potential of the system allows us to calculate the energies involved with binding events and useful allosteric changes. A number of free energy calculations specific to ligand binding exist: Some of these methods include free energy perturbation (FEP) [19], thermodynamic integration (TI) [20], and the Molecular Mechanics/ Poisson Boltzmann Solvent Accessible (MMPBSA) [21] approaches—each with their own unique strengths and weaknesses but all confined to the same practical limitations [22]:

i) Correct force field description for the atoms and intermolecular forces
ii) Entropic approximations for both complex and solvent
iii) Proper sampling.

Throughout later chapters, MMPBSA calculations are done to estimate the free energy of various binding events. The MMPBSA method was chosen specifically for usability purposes, as it's integrated into the AMBER molecular dynamics package. The method relies on a simple thermodynamic cycle incorporating im-

plicit solvation energies of each component in the binding complex as shown in Figure 9.1 and Eqn. 3.15 below.

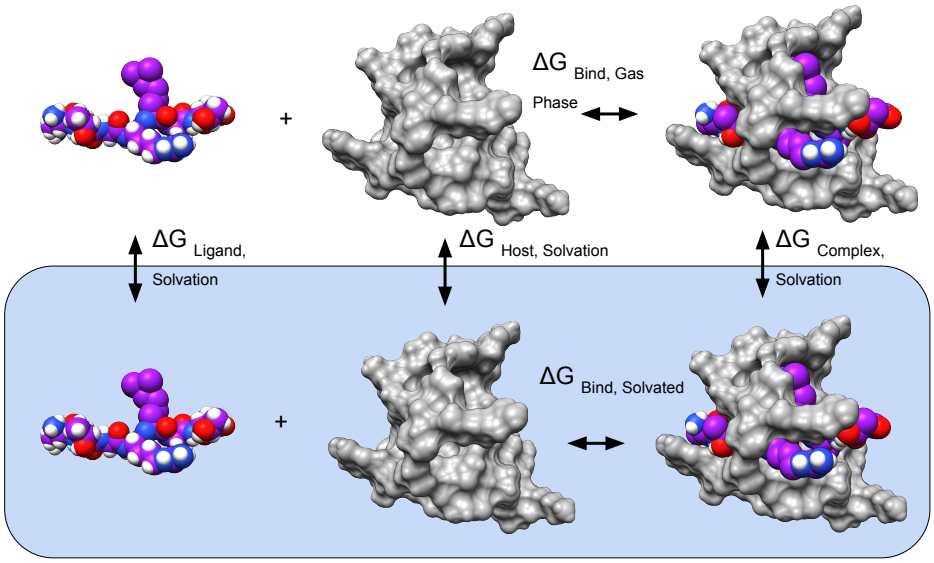


Figure 3.4: MMPBSA.py Thermodynamic Cycle Example with CBX8/H3K9Me$_3$ (PDB: 3i91)

$$\begin{aligned} \Delta G_{binding,solvated} = \Delta G_{binding,vacuum} \\ + \Delta G_{solvation,complex} \\ - (\Delta G_{solvation,ligand} + \Delta G_{solvation,host}), \end{aligned} \tag{3.15}$$

where $\Delta G_{solvation}$ terms represent the free energies of solvation using implicit solvation models (no actual explicit water molecules), and free energies between complexes ($\Delta G_{binding,vacuum}$) are calculated using the potential energies described by the force field terms, $U(r)$. As we are talking about ΔG, it is worth mentioning that entropy contributions in both solvation and complexation terms are estimated through other approximations. In the case of solvation entropy, empirically derived expressions dependent on the surface charge and size are used [23]. In the case of the entropy of complexation, other approximation methods such as normal mode analysis are employed [24].

The usage of MMPBSA (or any free energy method for that matter) requires a proper conformational sampling to be reliable. More simply put, one needs to generate an actual statistically sound and realistic representation of the binding event. In the case of MMPBSA, standard usage of the program involves a "single trajectory approach" where ligand, host, and complex coordinates for the calcu-

lation are taken from just the complex coordinates. This is done by using the trajectory coordinates of the complex and stripping away either ligand or host, resulting for example, in a trajectory of the host induced by the ligand but assumed to be the same as the unbound state. Furthermore, the calculation assumes the free energy of solvation is properly described by the implicit solvent model. In the case of structural waters in and around the binding site, this will likely lead to further error.

These issues relate back to our discussions on configurational changes upon binding and the energetic costs of host reorganization. Without a multiple trajectory approach with independent ligand and host states for reference, the cost of any induced binding or loss of conformational freedom is unaccounted for. Despite the absence of any induced-fit energy, single trajectory calculations remain as the common and default usage of MMPBSA [25, 26].

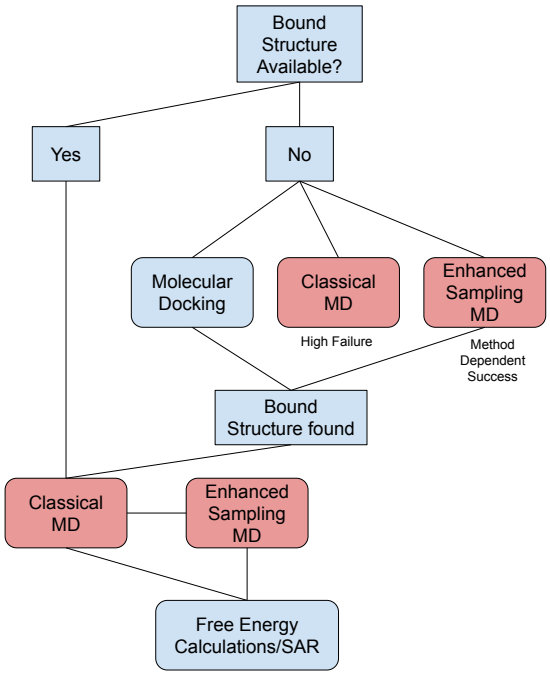


Figure 3.5: Example Use Diagram for MD in SBDD

Structural Metrics and Ensemble Generation

Reasonable starting structures followed by MD can provide information regarding the structure activity relationships of a protein-ligand complex. Conserved interactions such as hydrogen bonding or salt bridges are useful metrics for features that can be exploited in further ligand optimization, and can be monitored over time within a simulation. Small allosteric changes as well as ligand stability in

the binding site are also within the MD timescale window, provided the ligand is optimally placed and the conformational changes are not too large.

Aside from free energy calculations, MD also plays a role in the preparation of files for other SDBB methods, namely docking and other enhanced sampling techniques. Molecular dynamics is particularly useful for generating alternative host poses to be used in docking. The structural space searched in MD, even though limited and non-equilibrium, can prove incredibly useful. For instance, MD has been used to find hidden binding pockets not observed in experimental methods [27]. These combined methods fall under MD generated ensemble docking techniques and will be introduced in more detail after a discussion on molecular docking.

3.2 Molecular Docking

Molecular docking techniques take aim at producing predicted protein-ligand complexes with a characteristic binding energy or "score". Structural accuracy, correct ranking compared to other ligands, and absolute free energies of binding represent the typical metrics for success [28]. To distinguish docking from MD, the algorithms used are not reflective of any realistic binding event. No kinetics, temperature, or diffusion of the ligand occur within these methods. In fact, the absence of these characteristics is what makes docking such a useful technique—by exploring the conformational space of the ligand in a non-deterministic way, docking softwares are able to explore this space with incredible speed. For this reason, the ability of docking software to quickly scan through chemical space has made it a heavily used tool in SBDD and is increasingly leading to success in the identification of new compounds for further study [29–31].

The ability to quickly rank large libraries of ligands in terms of binding affinity *in silico* has the potential to significantly decrease the cost of pre-clinical drug development. Unsurprisingly, the past several decades of software development have expanded molecular docking techniques into dozens of available software options with varying strengths and weaknesses. The correct choice and application of docking software is generally a system-specific problem and without a doubt, the most important consideration is the level of protein flexibility of the system at hand [32, 33]. This isn't to say there aren't other factors though: inclusion of structural waters [34] in the binding event, and a *systems in context* approach are also important considerations when tackling a structural prediction. The concept of a system in context is paramount for meaningful molecular modelling. For example, membrane proteins without a membrane [35], removal of co-factors [36], and exclusion of quaternary structures play a large role in allosteric effects as well as solvent accessible surface area for docking on the target protein. Without the unique physical environment of the system in question, the impact on target proteins is unknown and has the potential to dramatically alter the structural prediction of a protein–ligand complex.

To emphasize the drive in the SBDD community to incorporate features such as flexibility as well as the opportunity for improvement, a small survey of molecular docking software is presented here, along with their strengths and weaknesses. As there are literally dozens of molecular docking applications [37], this will not be a comprehensive discussion of all available methods, but more of a broad array of popular techniques one may find when choosing a docking program. More specifically, the framework presented below serves to show how the method, "SLICE",

fits into the conventional arsenal of SBDD methods.

Classifications of docking approaches are typically defined by three attributes: (i) The search algorithm for the ligand/host conformations, (ii) The scoring functions that rank the best pose as well as relative binding affinity and (iii) the degree of host flexibility and how it is incorporated. Definitions between semi-flexible and flexible are also blurred depending on what is being discussed and *the level of flexibility* as a quantitative or even qualitative measurement is ambiguous. For this discussion, we will use *semi-flexible* to include any methods that involve partial flexibility of the host, whereas fully flexible methods incorporate at least some aspect of host backbone flexibility.

Conformational Search Algorithms

Protein: Meet ligand. Ligand: Meet protein.

For the majority of search algorithms, generating the final coordinates of a best fit involves the evaluation of numerous configurations. The process of how the algorithm generates and sorts through these intermediate structures to a final pose can contain both systematic and stochastic elements. In the case of deterministic methods, the method can contain a set of instructions for generating poses using pre-defined moves involving rotations and translations of the ligand in an exhaustive manner. Stochastic elements can involve the occasional random placement of the ligand as well as a random element in accepting a new pose. Both deterministic and stochastic methods have advantages to how quickly or rigorously they are able to search the ligand conformational space.

As an example, a *heavily* simplified algorithm using a Monte Carlo (a type of stochastic) approach and random move generator is shown below.

ExampleDock: A Conformational Search Algorithm Example

Let us first create a pose for the ligand using random selections of torsions, overall rotation of the molecule, and a translation of coordinates for it's centre of mass in the search space. The same algorithm can be used to make small or large changes in the position and orientation. Each move would contain rotations of the whole molecule as well as internal torsions along with moving in one direction.

```
for i in len(rotatable_bond):
```

```
      rotatable_bond[i] = random(0,360)
ligand_rotation[phi,psi] = random(phi),random(psi)
ligand_translation[x,y,z] = random((0,xmax),0,ymax),0,zmax))
```

The above provides the basic information to place the ligand somewhere in the search space. If this is our first move, we have no choice but to accept it or if it is completely unreasonable as defined per the user, ie. $\Delta G > 0$, then we may try again. Let's call this $pose_0$. We then repeat this process to generate a $pose_1$, both with their own respective scores, E_0 and E_1.

At this point, we need to make a decision. Do we keep this move or not? If the new pose contains a better score, we can accept the new move. Continuing on this process will cause us to follow a path of minimization, continuously accepting better poses until we cannot get any better. Depending on the initial start coordinates and the magnitude of allowed translation, the risk of finding a local minimum is high. To avoid this type of trapping, we can add another stochastic element to how we accept poses. Enter a Monte Carlo probability factor.

$$P = exp\left[\frac{-(E_1 - E_0)}{k_B T}\right], \tag{3.16}$$

where E_1 and E_0 represent new potential move and current state binding energies, respectively, and $k_B T$ represents the product of the Boltzmann constant and temperature.

The variable P in this case, a type of Bolztmann factor, results in a number between 0 and 1. This factor is then compared to a randomly generated number between 0 and 1. If the calculated factor is higher than the random number, the move is accepted. Some bookkeeping is required as well. Lowest energy or "best" poses are saved throughout this type of search. The search continues on for a number of defined steps or until another convergence criteria is met. In the end, we should at the very least have a pose lower in energy than the starting point.

The example above is *incredibly* basic compared to the diverse techniques applied in modern docking software, but the same fundamental steps of generating, evaluating, and accepting poses exist. For example just within the realm of the Monte Carlo selection in ligand docking, several flavours of this algorithm exist such Hamiltonian [38] or Replica Exchange Monte Carlo [39, 40]. Outside of the Monte Carlo scheme, another group of popular search/acceptance methods are the evolutionary algorithms implemented in GOLD [41] and AutoDock4 [42]. As for

generating new poses, some programs even break the ligand into fragments and rejoin them after docking, as is done in FlexX [43].

Scoring Functions

In the example above, the variables E_0 and E_1 were briefly described as "scores" or binding energies. The terms are often interchangeably used but both describe a value for ligand binding. Scoring functions are arguably the heart of a docking program. Even if the program was able to generate every single possible configuration of the ligand with the host, the scoring function decides what pose is selected and how that pose compares to those of other ligands.

From previous discussions on molecular mechanics in Section 3.1.1, we've seen that we can use a force field approach that uses sets of equations to describe the potential energy using atomic coordinates. Similar to explicit force fields, empirically based scoring functions rely on interatomic distances but expand on the basic force field description and include an array of empirically parameterized corrections. For the most part, these corrections are intuitive elements missing from a basic force field description; namely, special treatment of hydrogen bonds [44, 45], for co-factors [46], halogen bonding [47], and most importantly, desolvation and entropy terms [48]. These special treatments in scoring functions can lead to high accuracy for specific systems. However, some users opt to perform consensus docking [49] to take averages from different scoring functions in attempts to have a more broadly effective evaluation.

Example of the Generalized AutoDock Scoring Function [50]:

$$c = \sum_{i<j} f t_i t_j(r_{ij}), \tag{3.17}$$

where c, the final score, is composed of a linear combination of various functions, $f t_i t_j$, describing the interatomic potential dependent on the distance, r_{ij}. In the case of AutoDock Vina, scoring function terms are split up in terms of interatomic potentials with specially added treatments for hydrophobic and hydrogen bonding interactions.[50, 51]

More recently, machine learning approaches have emerged in scoring function development. The resulting scoring functions show promise but can lack an interpretative picture of what the function physically means [52]. Machine learning approaches result in highly complicated systems of weightings and numerous internal functions relating to every atom as opposed to the discrete forces split up in force field and semi-empirical methods. Despite this black box element, the

promise is apparent in the results, and machine learning for scoring function development is a highly active field with numerous recent successes [53–55].

Inclusion of Host Flexibility

Host flexibility is frequently cited as the most important challenge for docking software going into the future [56–59]. The treatment of proteins as rigid or semi-rigid entities can lead to enormous disparities in predicted activities depending on the actual flexibility of the host. One solution is to include flexibility within amino acid side chains by allowing their torsional freedom and searching both ligand space as well as side chain torsional space. Even this small detail can combinatorially explode the search space required to find the lowest energy pose. To address protein flexibility but still stay within the computational means of its users, developers have incorporated a number of tricks.

The use of soft potentials, or softer docking penalties for steric interactions is one such method [60] and is used by the program, GOLD [61]. Other programs include a set of predefined side chain rotamers such as in the program ICM [62]. The use a stochastic side chain generator like in AutoDock4 could also be employed. Either way, the current methods in the most widely used programs today address protein flexibility in terms of side chain reorganization. Full protein flexibility remains the holy grail in molecular docking.

The emerging use of ensemble techniques tackle aspects of full protein flexibility by attempting to start with a collection of relevant structures for docking i.e. representative configurations of host with different backbone and side chain orientations. This can be done by acquiring a number of crystal structures and using the flexible side chain methods listed above, or alternatively by using molecular dynamics to generate poses for docking in a combined approach.

3.3 Combined MD/Docking Approaches and Ensemble Generation Methods

At this point, we have made the distinction between docking and MD and discussed their respective strengths and weaknesses. At numerous points, we have also alluded to the potential combination of both methods as improvements to either approach. Surveying the current literature would make it seem as if combined MD/docking methods or *dynamic* docking are a recent paradigm shift [63, 64], however, there are early examples of successfully applied combinations dating to the late 90's. In a seminal paper, Carlson et al. describe one of the first uses of an

ensemble docking technique where they first acknowledge their target receptor as occupying multiple configurational states. They then go on to use MD to generate a number of poses for subsequent docking [65]. This work eventually led to the identification of a novel class of HIV-1 integrase inhibitiors [66]. So why is it that docking software for almost 15 years continued to work with static or semi-flexible methods when ensemble docking techniques showed such promise?

The work done by Carlson et al. involved performing 1 ns simulations of the target and taking a small handful of structures as their "dynamic pharmacophore". Remembering back to the timescales for MD, these are very short simulations and would have captured minimal backbone reorganization but a significant amount of side chain rotamers, and for the HIV-1 integrase, this was apparently enough to generate a useful ensemble. At the time, the 1 ns simulations were expensive and subject to the accuracy limitations of their time. The process of docking on multiple conformers was also an expensive task. No doubt, the current re-emergence of ensemble docking techniques is due to the accessibility of molecular dynamics software and sufficient hardware to process timescales of reasonable length. Moreover, recent enhanced sampling methods within molecular dynamics have allowed for the generation of more useful ensembles, providing the variation of flavours in ensemble docking techniques today.

As a clarification, enhanced sampling techniques encompass the methods used to explore a larger configurational space within molecular dynamics itself, whereas ensemble generation is the process of turning MD trajectories into a set of host poses for docking. Variations on both of these techniques allow for an enormous amount of method exploration and represent a significant challenge in terms of best practices. An example of just a few routes within ensemble docking are presented in Figure 3.6 and illustrate the numerous paths one may follow in choosing the component methods in an ensemble docking experiment.

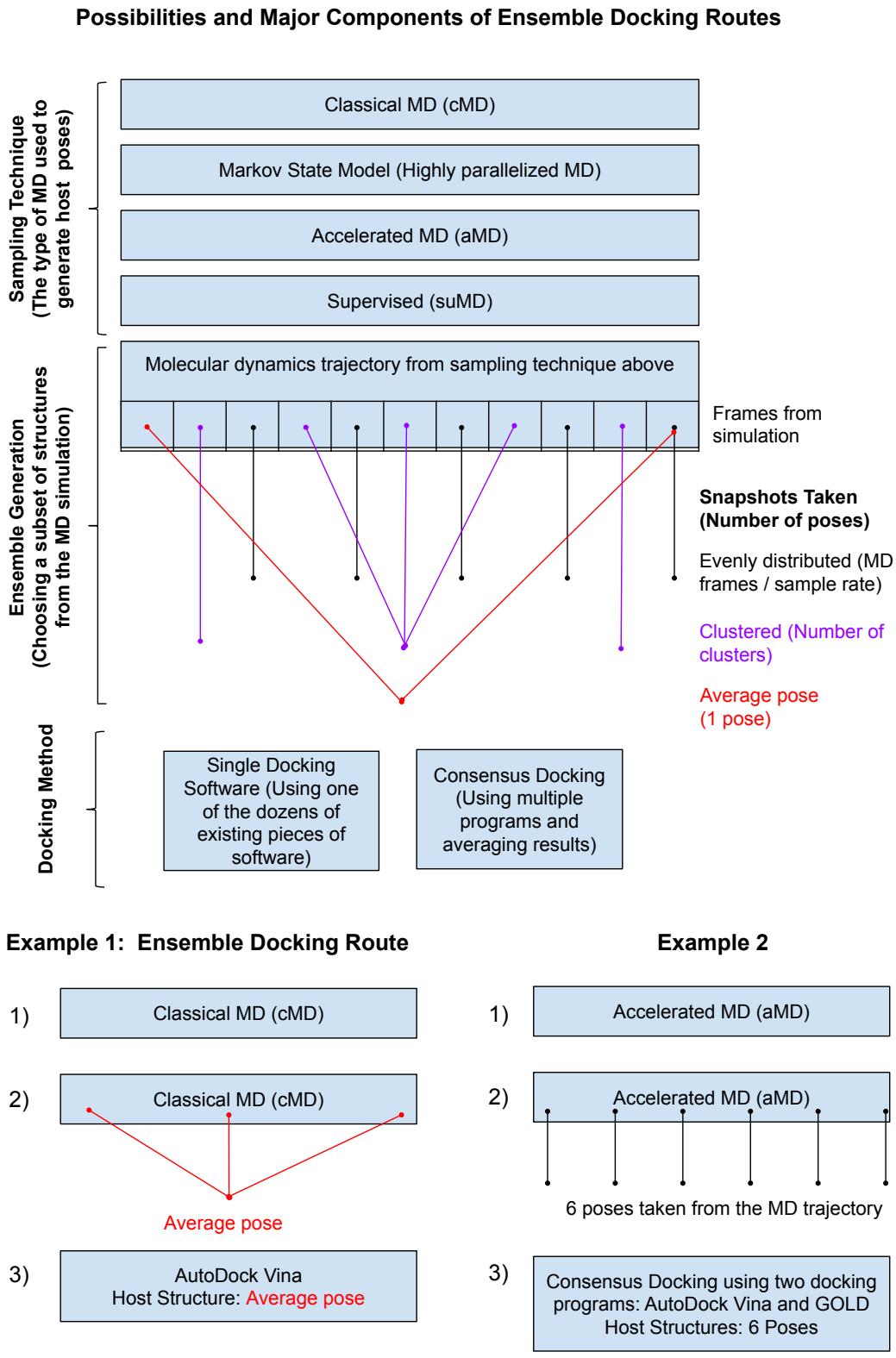


Figure 3.6: Possible Ensemble Docking Routes. The top panel illustrates the three major steps in an ensemble docking experiment: (i) simulation, (ii) parsing for host structures, and (iii) docking on the host structures. The two examples below are just two of numerous possibilities given the available methods for each step.

Attempts to consolidate methods and gauge their effectiveness has also provided conflicting results. As an example, work by Evangelista-Falcon et al. compares the use of evenly spaced snapshots versus clustered poses as the host structures for an ensemble docking experiment—both sets taken from the same microsecond long MD simulation [67]. In this particular case, the evenly spaced snapshots were marginally more successful but incurred a much higher computational cost. Clustering methods are explored in a number of studies [68–70] and opinions regarding effectiveness seem generally mixed, stemming from one prevalent issue: Do the clusters represent any meaningful pose for the binding mode? From our discussions regarding induced-fit and conformational selection, it makes sense that there would be a subset of systems where if the binding mode matched a prevalent pose in the actual ensemble, then this method may work. But what about systems where the pose is available in the ensemble, but only at a very small fraction? This is the point where the effectiveness of clustering is likely to fall off. Furthermore, in the case of induced-fit where the bound pose is completely absent from the ensemble, clustering methods (and evenly spaced snapshots) will completely fail [71].

This search for relevant ensembles calls to enhanced sampling techniques in MD where through a number of methods, the sampled configurational space of the host is increased. Enhanced sampling methods typically involve some manipulation of the energy of transitions such as in accelerated MD (aMD) [72] or the energy available to make transitions such as in replica exchange MD [73]. In doing so, they are able to overcome barriers and explore the conformational space within the simulation much faster and more thoroughly than classical MD (cMD) without the perils of being trapped in local minima. This leads to a higher diversity of structures and hopefully still represents a meaningful distribution of poses. For conformational selection problems, this is advantageous to cMD in that bound poses have a higher chance of being found and subsequently docked on, and has shown a number of successes [74]. However, they are still limited to the potential energy surfaces of the apo host and ignore any induced effects by the ligand.

Inclusion of induced effects (ligand or other) on the potential energy surface have been cleverly applied through various approaches. One interesting approach developed by Uehara et al. involves generating host poses with MD simulations containing an organic solvent. The solvent is added at various concentrations and formulation (benzene, methanol, hexane etc.). This creative approach aims to expose hydrophobic regions of the protein surface that may contain cryptic or induced pockets for binding [75]. In essence, this is still an enhanced sampling technique that specifically aims to change the potential energy surface of the host—

only in this case, it is done by adding potential energy in the form of hydrophobic interactions that are generated in a somewhat realistic manner. Does a ligand induce the same behaviour?

Studies of ligand-induced effects on conformational changes of the protein are *numerous*. This is especially true in the age of rationally designed allosteric modulators [76] and is widely seen in studies involving G-protein coupled receptors (GPCRs) [77–80]. However, studies of ligand induced effects in early stages of binding aimed at discerning the energetic profiles of binding events and important early binding configurational changes on the host are rare. Why is this? Upon investigation, the few examples that do exist show these types of studies require an intense amount of calculation and therefore only a limited number of systems have been studied [81–84]. That's not to say there aren't important lessons to be derived from them:

i) Binding of a ligand can occur in multiple stages.
ii) Both induced-fit and conformational selection type binding can occur for a single system.
iii) Binding of a ligand can occur in multiple pathways.
iv) The presence of ligands on protein surfaces can destabilize other local interactions on the protein surface, leading to poses more conducive to binding.

If we take these points to be true, can we use the lessons learned as a platform for a conformational ensemble generating method that includes partial ligand binding effects? The methods used in the following publications in this thesis arguably illustrate that perhaps this is a sound approach.

Selective Ligand-Induced Conformational Ensembles (SLICE)

As a comparison to other ensemble docking methods, our method SLICE (Chapter 6) takes in account a conformational ensemble induced by the presence of the ligand. Similar to the co-solvent work previously discussed, the aim is to capture host configurations outside of the distribution of unbound protein poses, but still relevant to the actual mode of binding. In the ideal case, the host poses captured in a ligand-induced ensemble will be conducive to either induced-fit (the ligand actively plays a role in changing the available protein structures) *or* a conformational ensemble type of binding (the ligand was there with no impact). Figure 3.7 illustrates three pathways to binding and how SLICE would be relevant in capturing host structures in each binding pathway.

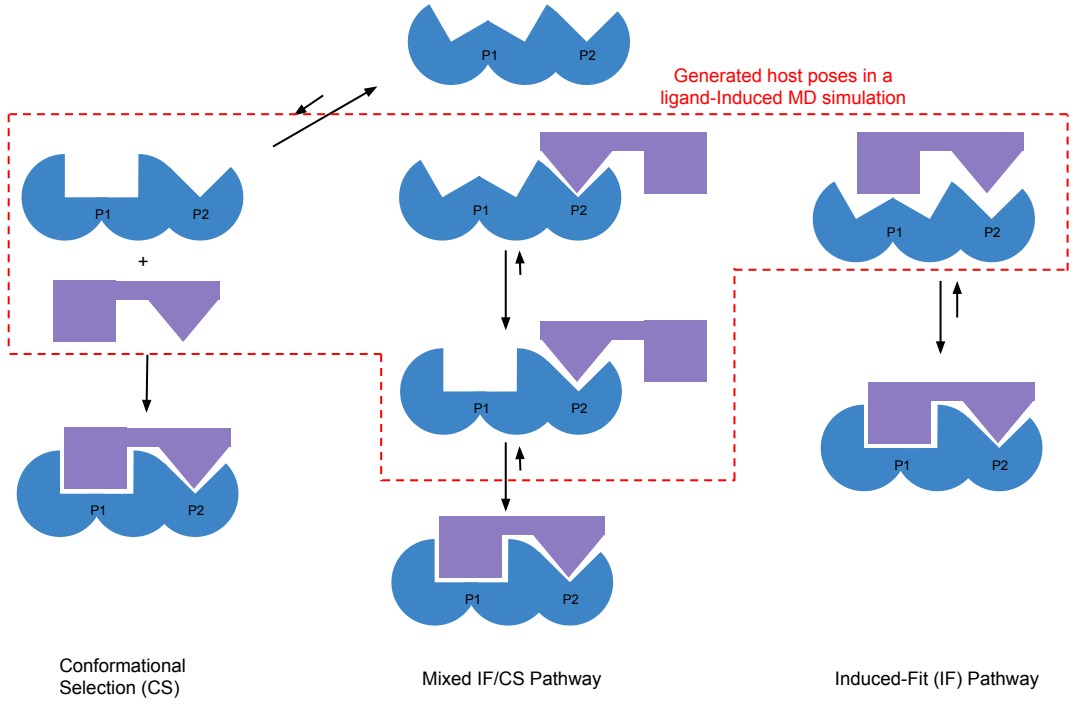


Figure 3.7: Ensemble Selection in Various Binding Schemes. The conformational selection mechanism assumes the bound host pose is present at some point and the protein?host complex is formed when it is sampled. The presence of the ligand in this case would not interfere with the pose generation. On the right, the induced-fit generated host poses are sampled by non-specific contacts of the ligand and the host that change the available conformations that can be sampled. A mixed mode of binding would also be captured by the presence of the ligand. The red box illustrates the type of protein conformations that would be sampled in the presence of a ligand.

Throughout the chapter, we briefly surveyed dozens of methods relating to both docking and MD and introduced the benefits of combined methodologies. The takeaway message from this chapter should be that we are now in an exciting era where computational resources have allowed us to expand quickly into method space. We are truly standing on the shoulders of giants and now have the ability to wield the tools they have created in novel ways. These new approaches in SDBB are opening up new doors to previously intractable systems in pharmaceutical development and will continue to bring advances to medicine and our general understanding of human health.

Bibliography

[1] Yang Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, June 2008.

[2] R Elber and M Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235(4786):318–321, January 1987.

[3] Nataraj S. Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical Reviews*, 9(2):91–102, January 2017.

[4] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, February 1975.

[5] Peter L. Freddolino, Anton S. Arkhipov, Steven B. Larson, Alexander McPherson, and Klaus Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–449, March 2006.

[6] Yibing Shan, Eric T. Kim, Michael P. Eastwood, Ron O. Dror, Markus A. Seeliger, and David E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, June 2011.

[7] Gabriel Ortega, Miquel Pons, and Oscar Millet. Protein functional dynamics in multiple timescales as studied by NMR spectroscopy. In *Dynamics of Proteins and Nucleic Acids*, pages 219–251. Elsevier, 2013.

[8] Albert C. Pan, Daniel Jacobson, Konstantin Yatsenko, Duluxan Sritharan, Thomas M. Weinreich, and David E. Shaw. Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of Sciences*, 116(10):4244–4249, February 2019.

[9] Christopher M. Baker. Polarizable force fields for molecular dynamics simulations of biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 5(2):241–254, January 2015.

[10] Thomas P Senftle, Sungwook Hong, Md Mahbubul Islam, Sudhir B Kylasa, Yuanxia Zheng, Yun Kyung Shin, Chad Junkermeier, Roman Engel-Herbert, Michael J Janik, Hasan Metin Aktulga, Toon Verstraelen, Ananth Grama, and Adri C T van Duin. The ReaxFF reactive force-field: development,

applications and future directions. *npj Computational Materials*, 2(1), March 2016.

[11] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, January 1982.

[12] Stephen P. Molner. The art of molecular dynamics simulation. *Journal of Chemical Education*, 76(2):11–43, 1999.

[13] William G. Hoover and Brad Lee Holian. Kinetic moments method for the canonical ensemble distribution. *Physics Letters A*, 211(5):253–257, February 1996.

[14] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, October 1984.

[15] Hans C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, February 1980.

[16] Stephen C. Harvey, Robert K.-Z. Tan, and Thomas E. Cheatham. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *Journal of Computational Chemistry*, 19(7):726–740, May 1998.

[17] I. Buch, T. Giorgino, and G. De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 108(25):10184–10189, June 2011.

[18] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of molecular dynamics and related methods in drug discovery. *Journal of Medicinal Chemistry*, 59(9):4035–4061, February 2016.

[19] Robert W. Zwanzig. High-temperature equation of state by a perturbation method involving non-polar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954.

[20] Ilja V. Khavrutskii and Anders Wallqvist. Improved binding free energy predictions from single-reference thermodynamic integration augmented with hamiltonian replica exchange. *Journal of Chemical Theory and Computation*, 7(9):3001–3011, August 2011.

[21] Bill R. Miller, T. Dwight McGee, Jason M. Swails, Nadine Homeyer, Holger Gohlke, and Adrian E. Roitberg. MMPBSA.py: An efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation*, 8(9):3314–3321, August 2012.

[22] Niels Hansen and Wilfred F. van Gunsteren. Practical aspects of free-energy calculations: A review. *Journal of Chemical Theory and Computation*, 10(7):2632–2647, June 2014.

[23] Qin Cai, Jun Wang, Meng-Juei Hsieh, Xiang Ye, and Ray Luo. Poisson–boltzmann implicit solvation models. In *Annual Reports in Computational Chemistry Volume 8*, pages 149–162. Elsevier, 2012.

[24] Samuel Genheden, Oliver Kuhn, Paulius Mikulskis, Daniel Hoffmann, and Ulf Ryde. The normal-mode entropy in the MM/GBSA method: Effect of system truncation, buffer region, and dielectric constant. *Journal of Chemical Information and Modeling*, 52(8):2079–2088, August 2012.

[25] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the MM/PBSA and MM/GBSA methods. *Journal of Chemical Information and Modelling*, 51(1):69–82, November 2010.

[26] Changhao Wang, D'Artagnan Greene, Li Xiao, Ruxi Qi, and Ray Luo. Recent developments and applications of the MMPBSA method. *Frontiers in Molecular Biosciences*, 4, January 2018.

[27] Shaoyong Lu, Wenkang Huang, and Jian Zhang. Recent computational advances in the identification of allosteric sites in proteins. *Drug Discovery Today*, 19(10):1595–1600, October 2014.

[28] Yan Li, Minyi Su, Zhihai Liu, Jie Li, Jie Liu, Li Han, and Renxiao Wang. Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nature Protocols*, 13(4):666–680, March 2018.

[29] Andrea Savarino. In-silico docking of HIV-1 integrase inhibitors reveals a novel drug type acting on an enzyme/DNA reaction intermediate. *Retrovirology*, 4(1):21, 2007.

[30] Juswinder Singh, Claudio E. Chuaqui, P.Ann Boriack-Sjodin, Wen-Cherng Lee, Timothy Pontz, Michael J. Corbley, H.-Kam Cheung, Robert M. Arduini, Jonathan N. Mead, Miki N. Newman, James L. Papadatos, Scott Bowes, Serene Josiah, and Leona E. Ling. Successful shape-based virtual screening: The discovery of a potent inhibitor of the type i TGFβ receptor kinase (tβRI). *Bioorganic & Medicinal Chemistry Letters*, 13(24):4355–4359, December 2003.

[31] Oren M. Becker, Dale S. Dhanoa, Yael Marantz, Dongli Chen, Sharon Shacham, Srinivasa Cheruku, Alexander Heifetz, Pradyumna Mohanty, Merav Fichman, Anurag Sharadendu, Raphael Nudelman, Michael Kauffman, and Silvia Noiman. An integrated in silico 3d model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1aagonist (PRX-00023) for the treatment of anxiety and depression. *Journal of Medicinal Chemistry*, 49(11):3116–3135, June 2006.

[32] M. Teodoro and L. Kavraki. Conformational flexibility models for the receptor in structure based drug design. *Current Pharmaceutical Design*, 9(20):1635–1648, August 2003.

[33] N Moitessier, P Englebienne, D Lee, J Lawandi, and C R Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British Journal of Pharmacology*, 153(S1):S7–S26, January 2009.

[34] Niu Huang and Brian K. Shoichet. Exploiting ordered waters in molecular docking. *Journal of Medicinal Chemistry*, 51(16):4862–4865, August 2008.

[35] Panagiotis I. Koukos, Inge Faro, Charlotte W. van Noort, and Alexandre M.J.J. Bonvin. A membrane protein complex docking benchmark. *Journal of Molecular Biology*, 430(24):5246–5256, December 2018.

[36] Pavel Pospisil, Thomas Kuoni, Leonardo Scapozza, and Gerd Folkers. Methodology and problems of protein-ligand docking: Case study of dihydroorotate dehydrogenase, thymidine kinase, and phosphodiesterase 4. *Journal of Receptors and Signal Transduction*, 22(1-4):141–154, January 2002.

[37] Leonardo Ferreira, Ricardo dos Santos, Glaucius Oliva, and Adriano Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, July 2015.

[38] Manuel P. Luitz and Martin Zacharias. Protein–ligand docking using hamiltonian replica exchange simulations with soft core potentials. *Journal of Chemical Information and Modeling*, 54(6):1669–1675, June 2014.

[39] Zhe Zhang, Christina E. M. Schindler, Oliver F. Lange, and Martin Zacharias. Application of enhanced sampling monte carlo methods for high-resolution protein-protein docking in rosetta. *PLOS ONE*, 10(6):e0125941, June 2015.

[40] Jens Meiler and David Baker. ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics*, 65(3):538–548, November 2006.

[41] Marcel L. Verdonk, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, and Richard D. Taylor. Improved protein-ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, August 2003.

[42] Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, December 2009.

[43] Matthias Rarey, Bernd Kramer, Thomas Lengauer, and Gerhard Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3):470–489, August 1996.

[44] Richard A. Friesner, Robert B. Murphy, Matthew P. Repasky, Leah L. Frye, Jeremy R. Greenwood, Thomas A. Halgren, Paul C. Sanschagrin, and Daniel T. Mainz. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21):6177–6196, October 2006.

[45] Hans-Joachim Bohm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, 8(3):243–256, June 1994.

[46] Christoph A. Sotriffer, Paul Sanschagrin, Hans Matter, and Gerhard Klebe. SFCscore: Scoring functions for affinity prediction of protein-ligand complexes. *Proteins: Structure, Function, and Bioinformatics*, 73(2):395–419, April 2008.

[47] Gautam R. Desiraju, P. Shing Ho, Lars Kloo, Anthony C. Legon, Roberto Marquardt, Pierangelo Metrangolo, Peter Politzer, Giuseppe Resnati, and Kari Rissanen. Definition of the halogen bond (IUPAC recommendations 2013). *Pure and Applied Chemistry*, 85(8):1711–1713, July 2013.

[48] Ashish Gupta, Neha Chaudhary, Kumar Reddy Kakularam, Reddanna Pallu, and Aparoy Polamarasetty. The augmenting effects of desolvation and conformational energy terms on the predictions of docking programs against mPGES-1. *PLOS ONE*, 10(8):e0134472, August 2015.

[49] Rajeev Jaundoo, Jonathan Bohmann, Gloria Gutierrez, Nancy Klimas, Gordon Broderick, and Travis Craddock. Using a consensus docking approach to predict adverse drug reactions in combination drug therapies for gulf war illness. *International Journal of Molecular Sciences*, 19(11):3355, October 2018.

[50] Oleg Trott and Arthur J. Olson. AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, pages NA–NA, 2009.

[51] Ajay N. Jain. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *Journal of Computer-Aided Molecular Design*, 10(5):427–440, October 1996.

[52] Joffrey Gabel, Jérémy Desaphy, and Didier Rognan. Beware of machine learning-based scoring functions—on the danger of developing black boxes. *Journal of Chemical Information and Modeling*, 54(10):2807–2815, September 2014.

[53] Cheng Wang and Yingkai Zhang. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *Journal of Computational Chemistry*, 38(3):169–177, November 2016.

[54] Janaina Cruz Pereira, Ernesto Raúl Caffarena, and Cicero Nogueira dos Santos. Boosting docking-based virtual screening with deep learning. *Journal of Chemical Information and Modeling*, 56(12):2495–2506, November 2016.

[55] José Jiménez, Miha Škalič, Gerard Martínez-Rosell, and Gianni De Fabritiis. KDEEP: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296, January 2018.

[56] Xiaotian Kong, Huiyong Sun, Peichen Pan, Feng Zhu, Shan Chang, Lei Xu, Youyong Li, and Tingjun Hou. Importance of protein flexibility in molecular recognition: a case study on type-i1/2 inhibitors of ALK. *Physical Chemistry Chemical Physics*, 20(7):4851–4863, 2018.

[57] Claudio Cavasotto and Narender Singh. Docking and high throughput docking: Successes and the challenge of protein flexibility. *Current Computer Aided-Drug Design*, 4(3):221–234, September 2008.

[58] Pierre Tuffery and Philippe Derreumaux. Flexibility and binding affinity in protein–ligand, protein–protein and multi-component protein interactions: limitations of current computational approaches. *Journal of The Royal Society Interface*, 9(66):20–33, October 2011.

[59] Simon J. Teague. Implications of protein flexibility for drug discovery. *Nature Reviews Drug Discovery*, 2(7):527–541, July 2003.

[60] Anna Maria Ferrari, Binqing Q. Wei, Luca Costantino, and Brian K. Shoichet. Soft docking and multiple receptor conformations in virtual screening. *Journal of Medicinal Chemistry*, 47(21):5076–5084, October 2004.

[61] Gareth Jones, Peter Willett, Robert C Glen, Andrew R Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking 1 1edited by f. e. cohen. *Journal of Molecular Biology*, 267(3):727–748, April 1997.

[62] Marco A. C. Neves, Maxim Totrov, and Ruben Abagyan. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *Journal of Computer-Aided Molecular Design*, 26(6):675–686, May 2012.

[63] Veronica Salmaso and Stefano Moro. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in Pharmacology*, 9, August 2018.

[64] Luca Pinzi and Giulio Rastelli. Molecular docking: Shifting paradigms in drug discovery. *International Journal of Molecular Sciences*, 20(18):4331, September 2019.

[65] Heather A. Carlson, Kevin M. Masukawa, and J. Andrew McCammon. Method for including the dynamic fluctuations of a protein in computer-aided drug design. *The Journal of Physical Chemistry A*, 103(49):10213–10219, December 1999.

[66] Heather A. Carlson, Kevin M. Masukawa, Kathleen Rubins, Fredric D. Bushman, William L. Jorgensen, Roberto D. Lins, James M. Briggs, and J. Andrew McCammon. Developing a dynamic pharmacophore model for HIV-1 integrase. *Journal of Medicinal Chemistry*, 43(11):2100–2114, June 2000.

[67] Wilfredo Evangelista Falcon, Sally R. Ellingson, Jeremy C. Smith, and Jerome Baudry. Ensemble docking in drug discovery: How many protein configurations from molecular dynamics simulations are needed to reproduce known ligand binding? *The Journal of Physical Chemistry B*, 123(25):5189–5195, January 2019.

[68] Edward Lyman and Daniel M. Zuckerman. Ensemble-based convergence analysis of biomolecular trajectories. *Biophysical Journal*, 91(1):164–172, July 2006.

[69] Sheng-You Huang and Xiaoqin Zou. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 66(2):399–421, November 2006.

[70] Renata De Paris, Christian Vahl Quevedo, Duncan D. Ruiz, Furia Gargano, and Osmar Norberto de Souza. A selective method for optimizing ensemble docking-based experiments on an InhA fully-flexible receptor model. *BMC Bioinformatics*, 19(1), June 2018.

[71] Jung-Hsin Lin, Alexander L. Perryman, Julie R. Schames, and J. Andrew McCammon. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *Journal of the American Chemical Society*, 124(20):5632–5633, May 2002.

[72] Levi C.T. Pierce, Romelia Salomon-Ferrer, Cesar Augusto F. de Oliveira, J. Andrew McCammon, and Ross C. Walker. Routine access to millisecond time scale events with accelerated molecular dynamics. *Journal of Chemical Theory and Computation*, 8(9):2997–3002, August 2012.

[73] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151, November 1999.

[74] Yinglong Miao, Dahlia Anne Goldfeld, Ee Von Moo, Patrick M. Sexton, Arthur Christopoulos, J. Andrew McCammon, and Celine Valant. Accelerated structure-based design of chemically diverse allosteric modulators of a

muscarinic g protein-coupled receptor. *Proceedings of the National Academy of Sciences*, 113(38):E5675–E5684, September 2016.

[75] Shota Uehara and Shigenori Tanaka. Cosolvent-based molecular dynamics for ensemble docking: Practical method for generating druggable protein conformations. *Journal of Chemical Information and Modeling*, 57(4):742–756, April 2017.

[76] John Christopher, Andrew Doré, and Benjamin Tehan. Potential for the rational design of allosteric modulators of class c GPCRs. *Current Topics in Medicinal Chemistry*, 17(1):71–78, November 2016.

[77] C Hoffmann, A Zürn, M Bünemann, and M J Lohse. Conformational changes in g-protein-coupled receptors-the quest for functionally selective conformations is open. *British Journal of Pharmacology*, 153(S1):S358–S366, January 2009.

[78] Xavier Deupi, Xiao-Dan Li, and Gebhard F.X. Schertler. Ligands stabilize specific GPCR conformations: But how? *Structure*, 20(8):1289–1290, August 2012.

[79] Mac Kevin E. Braza, Jerrica Dominique N. Gazmen, Eizadora T. Yu, and Ricky B. Nellas. Ligand-induced conformational dynamics of a tyramine receptor from sitophilus oryzae. *Scientific Reports*, 9(1), November 2019.

[80] Davide Provasi, Marta Camacho Artacho, Ana Negri, Juan Carlos Mobarec, and Marta Filizola. Ligand-induced modulation of the free-energy landscape of g protein-coupled receptors explored by adaptive biasing techniques. *PLoS Computational Biology*, 7(10):e1002193, October 2011.

[81] Geraldo Rodrigues Sartori, Andrei Leitão, Carlos A. Montanari, and Charles A. Laughton. Ligand-induced conformational selection predicts the selectivity of cysteine protease inhibitors. *BioRxiv*, August 2019.

[82] Alex Dickson and Samuel D. Lotz. Multiple ligand unbinding pathways and ligand-induced destabilization revealed by WExplore. *Biophysical Journal*, 112(4):620–629, February 2017.

[83] Denis Bucher, Barry J. Grant, and J. Andrew McCammon. Induced fit or conformational selection? the role of the semi-closed state in the maltose binding protein. *Biochemistry*, 50(48):10530–10539, dec 2011.

[84] Polo C.-H. Lam, Ruben Abagyan, and Maxim Totrov. Ligand-biased ensemble receptor docking (LigBEnD): a hybrid ligand/receptor structure-based approach. *Journal of Computer-Aided Molecular Design*, 32(1):187–198, September 2017.

Chapter 4

Publication: Structural study of a small molecule receptor bound to dimethyllysine in lysozyme

4.1 Preface

The following publication attempts to understand the behaviour of chemical probes designed for the use of post-translational detections — specifically to detect various methylated states of lysine. The work itself is split into three approaches: (i) NMR titration data that first presented the problem at hand. (ii) the x-ray crystallography so show the preferential binding site for the molecular probe, and (iii), the computational work to break down the potential cause of selectivity observed. Computational studies were performed by James McFarlane under the supervision of Irina Paci. The computational work performed in this publication tackles the structural prediction through classical molecular dynamics in a "mechanical bull" approach to assess the trajectory stability. Calculations involving solvent-accessible surface areas of the unbound and bound binding sites were also performed over the trajectories as a measure of the induced changes on the protein surface by the ligand. Prior to the production trajectories, parameterization steps for both dimethyllysine (intermediate and N-terminus) as well as the calixarene ligand were also performed using the AMBER12 molecular dynamics tool suite. Additional work involving MMPBSA free energy methods and molecular docking using AutoDock 4.2 was also performed; however, due to methodological challenges apparent in the results, the work was not presented in the following publication.

4.2 Publication

Full publication including links to supplementary information may be found at the following link:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4266562/

Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2015, **6**, 442

Structural study of a small molecule receptor bound to dimethyllysine in lysozyme†

Róise E. McGovern,[a] Brendan D. Snarr,[b] Joseph A. Lyons,[c] James McFarlane,[b] Amanda L. Whiting,[b] Irina Paci,[b] Fraser Hof[b] and Peter B. Crowley*[a]

Received 6th August 2014
Accepted 10th October 2014

DOI: 10.1039/c4sc02383h

www.rsc.org/chemicalscience

Lysine is a ubiquitous residue on protein surfaces. Post translational modifications of lysine, including methylation to the mono-, di- or trimethylated amine result in chemical and structural alterations that have major consequences for protein interactions and signalling pathways. Small molecules that bind to methylated lysines are potential tools to modify such pathways. To make progress in this direction, detailed structural data of ligands in complex with methylated lysine is required. Here, we report a crystal structure of *p*-sulfonatocalix[4]arene ($sclx_4$) bound to methylated lysozyme in which the lysine residues were chemically modified from Lys-NH_3^+ to Lys-$NH(Me_2)^+$. Of the six possible dimethyllysine sites, $sclx_4$ selected Lys116-Me_2 and the dimethylamino substituent was deeply buried in the calixarene cavity. This complex confirms the tendency for Lys-Me_2 residues to form cation–π interactions, which have been shown to be important in protein recognition of histone tails bearing methylated lysines. Supporting data from NMR spectroscopy and MD simulations confirm the selectivity for Lys116-Me_2 in solution. The structure presented here may serve as a stepping stone to the development of new biochemical reagents that target methylated lysines.

Introduction

Lysine is one of the most abundant residues on protein surfaces. With four methylenes and an epsilon amino group it is a cation of substantial conformational flexibility. Although native lysine is generally excluded from protein–protein interfaces[1] numerous post-translational modifications[2,3] produce a variety of functional groups with altered interaction properties. In particular, lysine methylation to the mono-, di- or trimethylated amine yields hotspots for protein interactions. Prominent examples include the methyllysines of histone tails that insert into aromatic cage motifs of chromatin remodelling enzymes.[2,4–6] In recent years many other (non-histone) proteins have been shown to contain methylated lysines,[7,8] though the roles of these modifications remain largely uncharacterized.

There is growing interest in the development of small molecule receptors that bind to lysine[9–19] and its methylated derivatives and analogues.[20–30] Synthetic ligands for methylated lysines have potential applications as inhibitors of protein–protein interactions and can be used as reagents in biochemical assays[26,31] and cell biology.[32] In certain cases it has been shown that lysine receptors bind to the methylated side chain with an affinity that is greater by several orders of magnitude and equilibrium dissociation constants (K_d) of ~10 μM have been reported for peptides containing trimethyllysine.[15,24,29] The anionic *p*-sulfonatocalix[4]arene[33] ($sclx_4$) and its analogues[34] have proven to be particularly useful for protein surface recognition[17,35] and/or complexation with lysine and methylated lysine.[10,15] A recent study of $sclx_4$ interactions with cytochrome *c* provided some of the first structural evidence of lysine recognition by a small molecule receptor.[17] And the structure of a phosphate-tweezers bound to a lysine in the 14-3-3 protein[18] further corroborated the use of supramolecular receptors for protein surface recognition.[36,37] Despite the growing literature on ligand binding to methylated lysines,[27–32] the structural characterization of a synthetic receptor bound to methylated lysine in a protein is completely lacking.

To gain structural knowledge of the interaction between a small molecule ligand and a protein bearing post-translationally modified lysines we solved the crystal structure of $sclx_4$ in complex with dimethylated lysozyme (lysozyme-KMe_2). The complex was further characterized by NMR spectroscopy and molecular dynamics simulations. We identified a surprisingly selective binding of the calixarene at one of six possible dimethyllysine residues. This selectivity was rationalized in terms of

[a]*School of Chemistry, National University of Ireland Galway, University Road, Galway, Ireland. E-mail: peter.crowley@nuigalway.ie; Tel: +353 91 49 24 80*
[b]*Department of Chemistry, University of Victoria, British Columbia, V8W 3V6, Canada*
[c]*School of Biochemistry and Immunology, Trinity College Dublin, Dublin, Ireland*

† Electronic supplementary information (ESI) available: Fig. S1: crystals of the lysozyme-KMe_2:$sclx_4$ complex grown at different $sclx_4$ concentrations. Fig. S2: crystals of the complex grown in the presence of chloride- and sulfate-containing salts. Fig. S3: 1D ^{1}H NMR spectra of lysozyme-KMe_2 in buffer and DMSO mixtures. Table S1: summary of crystallization conditions, data collection and refinement statistics. Movie S1: MD simulation snapshots of $sclx_4$ binding to Lys116-Me_2. See DOI: 10.1039/c4sc02383h

the local chemical environments of the dimethyllysines. A second binding site at Arg14 was also found in the crystal structure.

Results and discussion

Choice of model system

Lysozyme is a highly-characterized model protein that is frequently used for ligand binding studies.[38–41] Moreover it is a workhorse for structural studies of lysine methylation[42–46] with well-established protocols for dimethylation by reductive alkylation which modifies lysines and the N-terminus. With a high proportion of lysine/arginine side chains and an overall positive electrostatic potential (pI ~10, Fig. 1) lysozyme is suited to binding the anionic $sclx_4$.[35]

Calixarene binding in solution

The presence of $sclx_4$ resulted in the immediate precipitation of lysozyme-KMe_2. Precipitation occurred at μM – mM protein concentrations and crystals grew at protein and ligand concentrations as low as 20 and 1 μM, respectively (Fig. S1†) and in the presence of different sulfate- and chloride-containing salts (Fig. S2†). Notably, crystal growth occurred in the absence of precipitants such as PEG or ammonium sulfate. These data suggested a relatively high affinity interaction (K_d ~μM).

Attempts to characterize the complex in water/buffer were thwarted by precipitation. Thus, NMR spectroscopy was performed on protein samples in water–DMSO mixtures. Apart from small changes, the ^{1}H NMR spectrum of lysozyme was largely unaffected by 20% DMSO (Fig. S3†) indicating that the protein was stably-folded under these conditions[47] (>50% DMSO is required to unfold lysozyme).[48] Titrations were performed by the addition of μL volumes of a stock solution of $sclx_4$ and complex formation was monitored by collecting 1D ^{1}H and 2D ^{1}H-^{13}C HSQC spectra (Fig. 2). A single resonance at ~2.92 ppm, assigned to the $N^{\varepsilon}Me$ protons of Lys116-Me_2,[46] demonstrated a large upfield chemical shift. The nature of this chemical shift

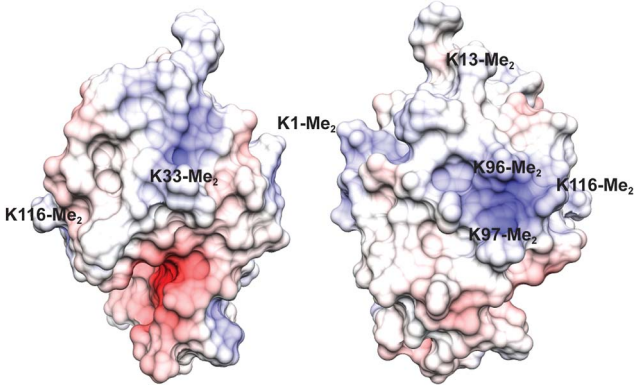


Fig. 1 The electrostatic potential surface (adaptive Poisson–Boltzmann solver) of lysozyme-KMe_2 with positive and negative patches coloured blue and red, respectively (the two views are related by a 180° rotation). Labels indicate the approximate locations of each of the six dimethyllysine residues.

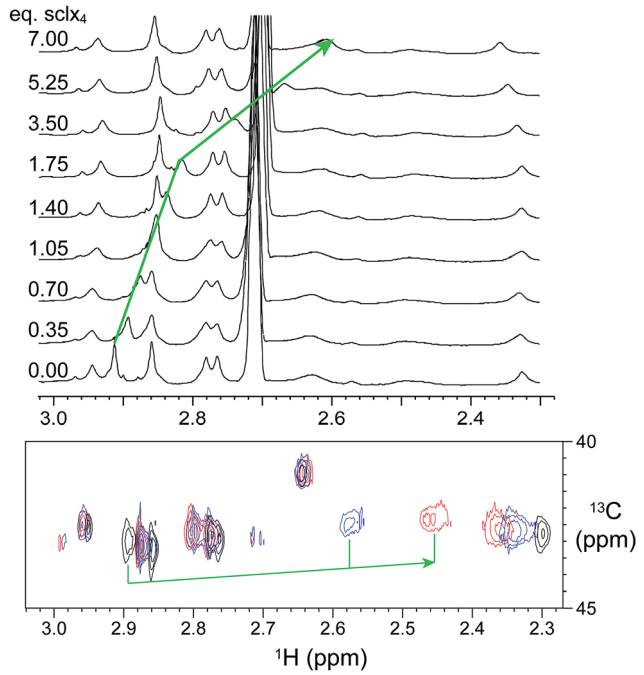


Fig. 2 NMR spectroscopic characterization of $sclx_4$ interactions with lysozyme-KMe_2. Upper panel, 1D ^{1}H NMR spectra (showing the region corresponding to $-NMe_2$ resonances) of dimethylated lysozyme in the presence of 0–7 equivalents of $sclx_4$. Lower panel, ^{1}H-^{13}C HSQC spectra of ^{13}C-labeled dimethylated lysozyme (black contours) in the presence of 7 (blue) and 10 (red) equivalents of $sclx_4$. The green arrows indicate the upfield shifts of the resonance assigned to Lys116-Me_2. A smaller downfield shift was observed for the Lys1-Me_2 resonance. The signal at ~2.7 ppm corresponds to DMSO. Samples were in 40 mM sodium phosphate, 10% D_2O, and 20% DMSO-d_6, pH 7.4.

perturbation was consistent with ring current effects induced by the phenyl rings of the calixarene cavity and suggests that the dimethylamino group was buried inside $sclx_4$.[15] A plot of the chemical shift changes as a function of the ligand concentration resulted in a shallow curve (data not shown) that was unsuited to an accurate K_d determination. It was not possible to reach saturation as sample precipitation occurred at >10 equivalents of ligand. Similar chemical shift perturbations were observed in 10% DMSO, although precipitation occurred at lower $sclx_4$ concentrations. This indicates that DMSO serves to reduce precipitation without impacting the binding selectivity.

A small downfield shift for the resonance assigned to Lys1-Me_2 (~2.32 ppm)[46] was observed at >3 equivalents of $sclx_4$ suggesting that a weaker interaction occurred at this site. However, the downfield shift indicated that encapsulation of the dimethylamino did not occur in this case. It is reasonable to assume that the probability of weak interactions at the N-terminal Lys1-Me_2 is greater than at the other dimethyllysines due to the relatively higher accessibility of this residue and the presence of two dimethylamines (at the N^{ε} and the N^{α} atoms).

Crystal structure of the lysozyme-KMe_2:$sclx_4$ complex

Crystals of the $sclx_4$ complex with lysozyme-KMe_2 grew under similar conditions and in the same space group as native

lysozyme[35] but with a ~50% smaller unit cell (Table S1†). Two almost identical structures (1.9 and 2.2 Å) were refined. The asymmetric units comprised two molecules of lysozyme-KMe_2 and four molecules of $sclx_4$ (Fig. 3A). Similar protein–ligand interactions were observed in each lysozyme-KMe_2 molecule. In agreement with the NMR data (Fig. 2) the side chain of Lys116-Me_2 was encapsulated by $sclx_4$. A second ligand was observed to bind Arg14. As noted in previous structures, the calixarene appeared to act like "molecular glue" at protein–protein interfaces in the crystal.[17,35]

The $sclx_4$–dimethyllysine interaction involves one of the methyl groups of Lys116-Me_2 inserted into the core of the calixarene, which behaves like a four-walled aromatic cage (Fig. 4A). The distance between the methyl carbon (C^{η}) and the centroids of the calixarene phenyl rings (3.5–3.9 Å) is consistent with cation–π interactions.[6,24,49] Short range contacts occur between the second methyl group of Lys116-Me_2 and two of the phenyl rings. Interestingly, this methyl is also in van der Waals contact (3.5 and 3.8 Å) with oxygen atoms of two sulfonates,

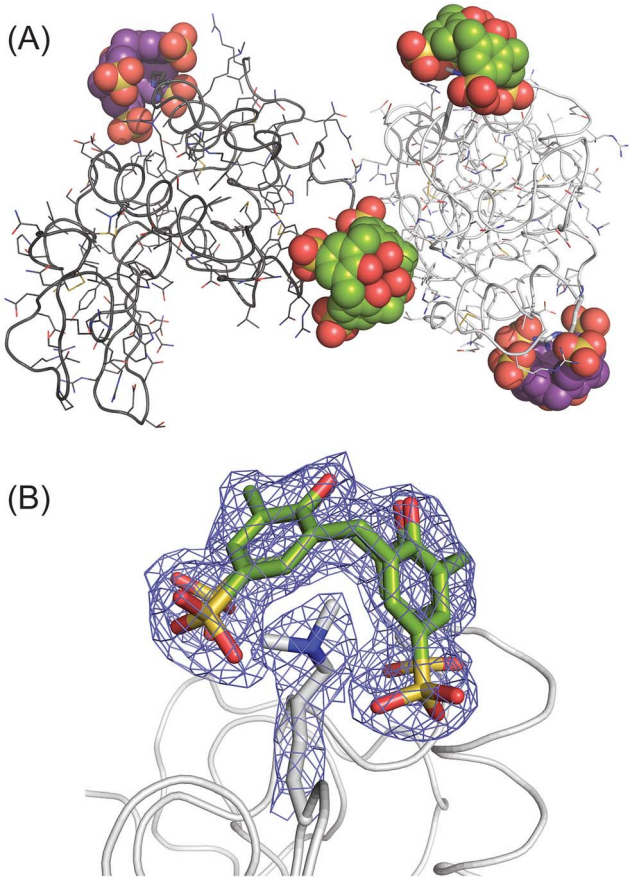


Fig. 3 The lysozyme-KMe_2:$sclx_4$ complex. (A) The asymmetric unit comprises two molecules of lysozyme-KMe_2 (rendered as light and dark grey ribbons) and four molecules of $sclx_4$. The dimethyllysine-binding and the arginine-binding calixarenes are coloured green and purple, respectively. (B) The calixarene binding site at Lys116-Me_2 showing the $2F_o - F_c$ electron density map around the Lys116-Me_2 side chain and $sclx_4$ (contoured at 1.0σ). See Fig. 4A for a detailed view of the $sclx_4$–Lys116-Me_2 interaction. The crystals used for this structure were grown from a 1 : 5 protein–ligand mixture (Table S1†).

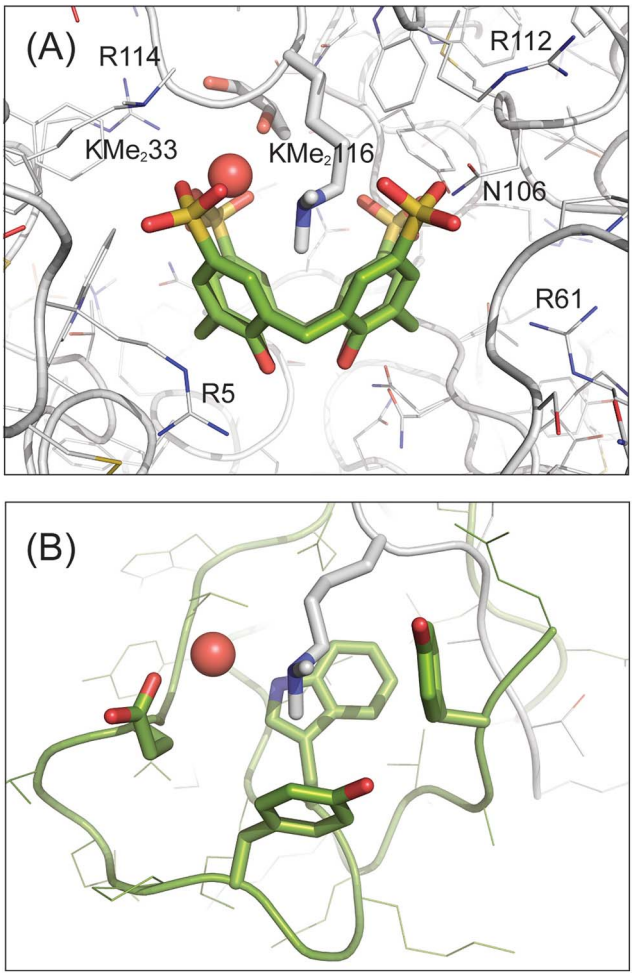


Fig. 4 $Sclx_4$ mimics the aromatic cage motif for binding dimethyllysine. (A) Detailed view of the $sclx_4$ complex with Lys116-Me_2. Neighbouring cationic side chains and Asn106 are labelled, see main text for details. (B) The aromatic cage (green) in the chromodomain of HP1 bound to dimethylysine in a histone H3 tail peptide (PDB 1KNA).[5] The chromodomain side chains Tyr24, Trp45, Tyr48 and Glu52 are shown as sticks. In both structures the dimethyllysine amino is solvated by a water molecule (red spheres), which is hydrogen bonded to one or more acidic substituents on the receptor. The proteins are shown as C^{α} traces with side chains as lines or sticks and the C atoms of $sclx_4$ are green.

hinting at the possibility of CH···O salt bridges. Water too plays a role in the coordination environment of the dimethyllysine. The tertiary amino nitrogen is 2.7 Å from a water molecule, which is neatly positioned between two sulfonates, indicative of water-mediated salt bridge interactions. The energetic contributions to $sclx_4$–dimethyllysine binding are expected to be dominated by cation–π interactions (rather than salt bridges)[6] and a contribution from the hydrophobic effect is also to be expected considering the water-bearing capacity of the calixarene cavity.[17,33] These features of the $sclx_4$–dimethyllysine interaction are further interesting in terms of their resemblance to how proteins such as chromodomains bind to methylated lysines in histone tails. An aromatic cage, typically comprising three aromatic side

chains, provides a pocket for the methylated lysine side chain, which remains partially solvated (Fig. 4B).[2,5]

Remarkably the conformation of Lys116-Me_2 bound to $sclx_4$ was almost identical to the side chain conformation observed in the original structure of dimethylated lysozyme (PDB 132L)[43] even though the crystals were grown from completely different conditions [PEG and low salt at pH 6.0 (Table S1†) *versus* 1.5–2.2 M $MgSO_4$ at pH 8.0]. The only difference was a rotation about the C^{δ}–C^{ε} bond, which increases the accessibility of the dimethylamino group in the $sclx_4$-bound structure.

Lysine *versus* dimethyllysine binding and selectivity

Substantial differences were observed for the binding of $sclx_4$ to lysine and to dimethyllysine. In the case of cytochrome *c*, the lysine side chains were fully encapsulated by the calixarene cavity such that all four methylene groups were in van der Waals contact with one or more of the calixarene phenyl rings.[17] To accommodate the entire side chain in this fashion the calixarene adopted a pinched cone conformation and the lysine amino group was positioned off-centre bringing it close to two of the sulfonates. In the case of dimethyllysine, one of the methyl groups is positioned in the centre of the calixarene such that the amino nitrogen is equidistant from all four sulfonates. Only the C^{η} and C^{ε} atoms make van der Waals contact with the calixarene phenyls while the remainder of the side chain protrudes from the cavity (Fig. 4A). Thus it appears that this complex favours a regular cone conformation of $sclx_4$, which maximises cation–π bonds with the dimethylamino group.[6,49] Similar interactions were found previously in complexes of $sclx_4$ with tetramethylammonium cations.[22]

These observations help to explain the selectivity for Lys116-Me_2, which projects out from the protein surface and has Asn106, Arg112 and Gly117 as neighbours. A hydrogen bond (Lys116 to Asn106) in the native protein is absent in the dimethylated protein where the Lys116-Me_2 side chain flips into the calixarene cavity and Asn106 hydrogen bonds to two of the sulfonates (Fig. 4A). The steric accessibility of most of the other Lys-Me_2 side chains is significantly lower, which may preclude $sclx_4$ binding. Lys1-Me_2 forms a cation–π interaction with Phe3, and salt bridges with Glu7 and the $sclx_4$ bound to Arg14; Lys13-Me_2 is screened by the carboxylates of Asp18 and *C*-terminal Leu129; Lys33-Me_2 is flanked by Asn37 and the bulky aromatics Phe34, Phe38 and Trp123, only the amine is accessible and it forms a salt bridge with a $sclx_4$ sulfonate; Lys96-Me_2 is buried and forms weak cation–π interactions with both Tyr20 and the $sclx_4$ bound to Arg14; while Lys97-Me_2 forms a salt bridge with Asp101. Similar conformations of the lysine side chains are present in the native lysozyme structure. To substantiate these observations the solvent accessible surface area (ASA) was calculated for each lysine in a dataset of 15 high resolution structures of lysozyme.[35] On average, Lys116 was the most accessible lysine (Fig. 5). While Lys97 has a similar ASA, it may be the differences in the local charge that tips the scales in favour of $sclx_4$ binding to Lys116. Considering charged groups within an 8 Å radius, Lys97-Me_2 forms a salt bridge with Asp101 while Lys116-Me_2 is neighboured by Arg112. The higher positive charge of the latter region will afford a greater attraction for the anionic $sclx_4$. To investigate this hypothesis we performed MD simulations of $sclx_4$ binding to Lys-Me_2 side chains.

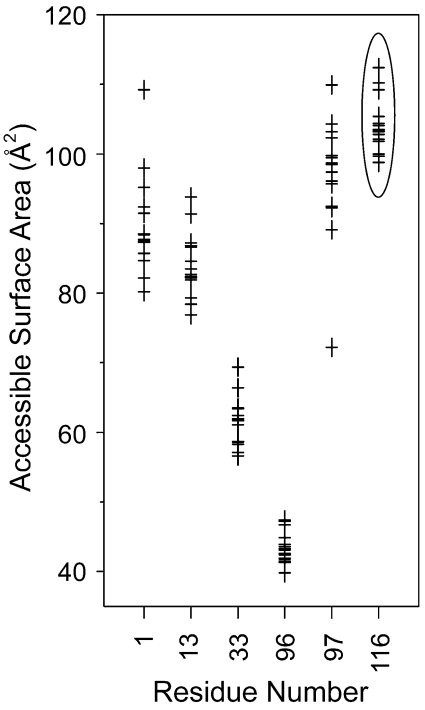


Fig. 5 The accessible surface area of the lysine residues in 15 high resolution crystal structures of lysozyme.[35] Lys116 (highlighted by an ellipse) was on average the most accessible residue. The dimethylated form of Lys116 was selectively bound by $sclx_4$ (Fig. 2 and 4A).

Molecular dynamics of $sclx_4$ binding to Lys-Me_2

Duplicate simulations (10 ns duration) were performed on $sclx_4$ binding to each of the six Lys-Me_2 side chains. The goal was to identify the structural features that distinguish Lys116-Me_2 from the other five potential binding sites. Two main features were considered; (1) the accessibility of the side chain, and (2) the local interactions, including those between the ligand substituents and peripheral residues. Site-specific information was determined from the molecular dynamics trajectories, which were examined primarily in terms of the "binding distance" as a means to quantify the degree of encapsulation of the Lys-Me_2 in the $sclx_4$ cavity. This was defined as the distance from the Lys-Me_2 N^{ε} atom to the best-fit plane through the methylene bridge carbons of the calixarene. Fig. 6A shows the evolution of the binding distance at the six sites, during two simulation trials. The trends explicitly illustrate the large differences in the potential for complex formation at the different sites. Lys13-Me_2, Lys33-Me_2 and Lys96-Me_2 did not form temporally stable complexes with $sclx_4$. In contrast, Lys97-Me_2 and Lys116-Me_2 were stably bound for almost the entire trajectory of each simulation consistent with the greater accessibility of these residues (Fig. 5). At a binding distance of 4–5 Å the dimethylamino is positioned deep within the calixarene cavity where it forms cation–π interactions. The complex

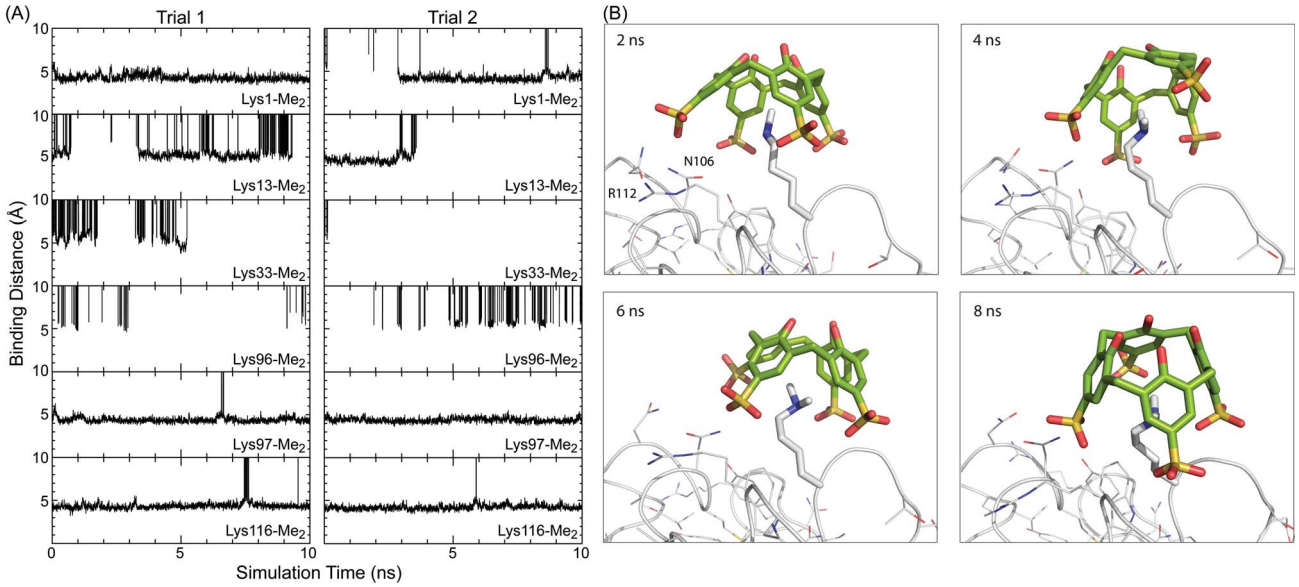


Fig. 6 (A) Binding distance (see main text for description) between sclx$_4$ and individual Lys-Me$_2$ side chains over the course of two MD simulations of 10 ns duration. Time points when the ligand was unbound are off the scale (0–10 Å). (B) Representative snapshots of ligand binding at Lys116-Me$_2$. The protein is represented as the C$^\alpha$ trace with side chains and Lys116-Me$_2$ shown as lines and sticks, respectively. Refer to the movie (ESI†) for a more comprehensive view of the binding conformations.

at Lys1-Me$_2$ was similar though it was unbound over ~3 ns of trial 2, suggesting a lower stability interaction. Attempts to discriminate the sites at Lys1-Me$_2$, Lys97-Me$_2$ and Lys116-Me$_2$ using calculated binding energies were unsuccessful, likely due to the well-documented incompatibility of the MM-PBSA end-state free energy method with highly charged systems.[50,51]

Further analysis was focused on the complexes at Lys97-Me$_2$ and Lys116-Me$_2$. Representative snapshots of the complexes at Lys116-Me$_2$ (Fig. 6B, see also the MD movie, ESI†) show the potential role for both Arg112, which can form salt bridge interactions with one of the calixarene sulfonates, and Asn105, which can hydrogen bond to the sulfonates. The two residues interact with the calixarene outer rim alternatively, with the Arg112 in binding position 25–40% of the time. These observations were corroborated by the crystal structure. In protein chain B there is a salt bridge between Arg112 and one of the calixarene sulfonates, while in chain A an alternative rotamer is present and Arg112 forms an intramolecular hydrogen bond with the protein backbone. The amide of Asn106 is positioned equidistantly from two sulfonates in the site at chain A, while in chain B only the amide N can form a hydrogen bond with sclx$_4$. Similarly, calixarene binding at Lys97-Me$_2$ can involve salt bridge interactions with Arg21 (80–90% of the simulation time) and hydrogen bonds with Asn93 with a lower incidence. However, as noted above Lys97-Me$_2$ can also salt bridge with Asp101. The MD data suggests that, once complexed, Lys116-Me$_2$ and Lys97-Me$_2$ can interact with sclx$_4$ in a similar fashion. Selectivity of the calixarene to Lys116-Me$_2$ is due to the greater steric accessibility at that site. This is supported by the relative values of ASA averages collected during the solution-phase simulations (data not shown), which were in agreement with the crystallographic data (Fig. 5), although ASA values in solution tend to be larger than in solid state due to the enhanced conformational freedom.

Arginine binding by sclx$_4$

While the selectivity of sclx$_4$ for Lys116-Me$_2$ can be rationalised in terms of steric accessibility, there remains the question of why binding also occurred at Arg14 in the crystal structure (Fig. 3). This question is interesting for two reasons. (1) NMR-derived binding curves for the interaction of sclx$_4$ with free amino acids have revealed a ~50-fold greater affinity for Lys-Me$_2$ over Arg.[15] Thus, it might be expected that sclx$_4$ would bind only Lys-Me$_2$ side chains in lysozyme-KMe$_2$. However, complex formation at Arg14 (instead of, for example, at Lys13-Me$_2$) reinforces the fact that protein surfaces, with their complex topologies and chemistries, can greatly alter the affinity of ligand binding. (2) In a crystal structure of native lysozyme and sclx$_4$ the ligand was bound at Arg128, the most sterically accessible arginine residue.[35] In lysozyme-KMe$_2$, Arg14, the second most accessible Arg residue was selected by sclx$_4$. Arg128 provides additional longer range (~8 Å) interactions to the sulfonates. It can be concluded that the affinities of sclx$_4$ for Arg14 and Arg128 are closely matched and the particular complex that prevails in the crystal structure depends on the crystal packing environment where the calixarene mediates protein–protein contacts.[17,35]

Conclusions

The data presented here illustrate how a protein containing dimethyllysine can be non-covalently modified by a small molecule receptor. Using a combination of X-ray crystallography, NMR spectroscopy and MD simulations we have shown

how the symmetric and anionic $sclx_4$ selectively binds to a single dimethylated lysine on the surface of a globular, folded protein. We note structural and chemical similarities between the complexes of dimethyllysine bound to the *simple* calixarene or to the aromatic cage motif of a chromodomain. This data will likely benefit the design of synthetic receptors for proteins (including histones) that contain methylated lysines.

Experimental

Materials

Hen egg white lysozyme (62971 Fluka) was dimethylated by using dimethylamine borane complex and formaldehyde according to published methods.[43–46] Electrospray ionization mass spectrometry data (Waters LCT Premier XE) for lysozyme (14,302.2 Da) and dimethylated lysozyme (14,498.0 Da) indicated complete dimethylation of all six lysines and the N-terminus. ^{13}C-formaldehyde was used to prepare dimethylated samples for ^{13}C NMR spectroscopy. Chemically-modified protein was purified by carboxymethyl (GE Healthcare) ion exchange chromatography prior to the crystallization experiments.

NMR spectroscopy

1D ^{1}H and 2D ^{1}H-^{13}C HSQC spectra were acquired on a Bruker AV500 operating at 500 MHz and 25 °C. Protein samples of 0.3–10 mM lysozyme-KMe_2 in 40 mM sodium phosphate, 10% D_2O, and 20% DMSO-d_6 at pD = 7.0 (pH 7.4) were titrated with μL volumes of a 0.65 M stock of $sclx_4$ in the same solution.

Crystallization and X-ray structure determination

The hanging drop vapour diffusion method was used for crystallization at 20 °C. Co-crystals of lysozyme-KMe_2 and $sclx_4$ were grown from similar conditions to those reported for lysozyme.[35] The drops were prepared by combining 1 μL volumes of protein, $sclx_4$ and the reservoir solution (Table S1†). Diffraction data for the lysozyme-KMe_2:$sclx_4$ single crystals were collected at the ESRF (BM14, MarCCD detector, φ scans of 1° over 180° to a resolution of 1.9 Å) and at the Swiss Light Source (X10SA, 10 μm minibeam, Pilatus 6M detector, φ scans of 0.5° over 180°, to a resolution of 2.2 Å). Data processing and scaling were performed in MOSFLM[52] and SCALA,[53] respectively or in xia2 (ref. 54) using XDS,[55] XSCALE and SCALA. See Table S1† for the data collection and refinement statistics. The structures were solved by molecular replacement in PHASER. Refinement and manual rebuilding were performed in REFMAC5 as implemented in CCP4 (ref. 56) and COOT,[57] respectively. Solvent molecules were placed automatically using ARP/wARP[58] and refinement was continued until no features remained in the $F_o - F_c$ difference maps. Molprobity[59] was used to check the structure quality. Coordinates and structure factors were deposited in the Protein Data Bank with the accession codes 4PRU (2.2 Å) and 4N0J (1.9 Å). The protein–ligand and protein–protein interfaces were analysed in COOT.

Molecular dynamics of protein–calixarene interactions

Binding dynamics were followed using classical molecular dynamics over 10 ns intervals, at a temperature of 300 K. To reduce the introduction of bias a structure of native lysozyme (PDB 3RZ4) was used for the initial coordinates. This structure was modified with newly parameterized Lys-Me_2 residues replacing all six of the lysines. Dimethyllysine was not available in the standard AMBER residue library so the parameters were retrieved from the literature.[60] Partial charges were derived from gas phase optimized HF/6-31*G calculations in Gaussian09, and fit in the preparatory program Antechamber using the RESP charge fitting method. The remaining parameters for the nonstandard amino acids were obtained from Antechamber and fit to the AMBER ff10 force field. Parameters for the calixarene were fit to the general AMBER force field for small organic molecules. The ligand structure was minimized in explicit TIP3P water prior to being placed with the protein for simulation. The calculations used explicit TIP3P water and Cl^- counter-ions added to charge neutrality.[61,62] Duplicate protein–calixarene complexes were generated for each candidate Lys-Me_2 site by combining the equilibrated protein and the minimized $sclx_4$ structure. The ligand was placed approximately 7–10 Å above the Lys-Me_2 side chain and the complex was allowed to equilibrate. The binding energy at each site was calculated using the MM-PBSA method (implicit solvent).[50,51] Docking was also monitored through calculations of the average binding distance between the dimethylamino group and the plane of the calixarene methylene bridge carbons.

Acknowledgements

This research was supported by NUI Galway (college scholarship to REM, Millennium Fund to PBC), NSERC grants to IP and FH (Canada Research Chair), and Science Foundation Ireland (10/RFP/BIC2807 to PBC). JAL was funded by grants 07/IN.1/B1836 (SFI) and GM75915 (NIH) to M. Caffrey. ALW was funded by a Michael Smith Foundation for Health Research fellowship. We acknowledge the Swiss light source and the European synchrotron radiation facility for beam time allocation, and the staff of beam lines X10SA (Villigen) and BM14 (Grenoble) for assistance with data collection. Computational work was performed using WestGrid, funded in part by the Canada Foundation for Innovation, Alberta Innovation and Science, BC Advanced Education, and the participating research institutions. For their assistance we thank colleagues at TCD; V. Pye, A. R. Khan, and UVic; M. Beatty, L. Netter, C. Bohne, C. Greenwood, C. Barr, O. Granot.

References

1 J. Janin, R. P. Bahadur and P. Chakrabarti, *Q. Rev. Biophys.*, 2008, **41**, 133.
2 S. D. Taverna, H. Li, A. J. Ruthenburg, C. D. Allis and D. J. Patel, *Nat. Struct. Mol. Biol.*, 2007, **14**, 1025.
3 Z. Zhang, M. Tan, Z. Xie, L. Dai, Y. Chen and Y. Zhao, *Nat. Chem. Biol.*, 2011, **7**, 58.

4 S. Rea, F. Eisenhaber, D. O'Carroll, B. D. Strahl, Z. W. Sun, M. Schmid, S. Opravil, K. Mechtler, C. P. Ponting, C. D. Allis and T. Jenuwein, *Nature*, 2000, **406**, 593.
5 S. A. Jacobs and S. Khorasanizadeh, *Science*, 2002, **295**, 2080.
6 R. M. Hughes, K. R. Wiggins, S. Khorasanizadeh and M. L. Waters, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 11184.
7 X. J. Cao, A. M. Arnaudo and B. A. Garcia, *Epigenetics*, 2013, **8**, 477.
8 K. E. Moore, S. M. Carlson, N. D. Camp, P. Cheung, R. G. James, K. F. Chua, A. Wolf-Yadlin and O. Gozani, *Mol. Cell*, 2013, **50**, 444.
9 Y. Hamuro, M. C. Calama, H. S. Park and A. D. Hamilton, *Angew. Chem., Int. Ed. Engl.*, 1997, **36**, 2680.
10 M. Selkti, A. W. Coleman, I. Nicolis, N. Douteau-Guevel, F. Villain, A. Tomas and C. de Rango, *Chem. Commun.*, 2000, **2**, 161.
11 M. Fokkens, T. Schrader and F.-G. Klärner, *J. Am. Chem. Soc.*, 2005, **127**, 14415.
12 C. Renner, J. Piehler and T. Schrader, *J. Am. Chem. Soc.*, 2006, **128**, 620.
13 O. Hayashida, N. Ogawa and M. Uchiyama, *J. Am. Chem. Soc.*, 2007, **129**, 13698.
14 K. Kano and Y. Ishida, *Angew. Chem., Int. Ed. Engl.*, 2007, **46**, 727.
15 C. S. Beshara, C. E. Jones, K. D. Daze, B. J. Lilgert and F. Hof, *ChemBioChem*, 2010, **11**, 63.
16 M. Florea, S. Kudithipudi, A. Rei, M. J. González-Álvarez, A. Jeltsch and W. M. Nau, *Chem.–Eur. J.*, 2012, **18**, 3521.
17 R. E. McGovern, H. Fernandes, A. R. Khan, N. P. Power and P. B. Crowley, *Nat. Chem.*, 2012, **4**, 527.
18 D. Bier, R. Rose, K. Bravo-Rodriguez, M. Bartel, J. M. Ramirez-Anguita, S. Dutt, C. Wilch, F. G. Klärner, E. Sanchez-Garcia, T. Schrader and C. Ottmann, *Nat. Chem.*, 2013, **5**, 234.
19 C. J. Li, J. W. Ma, L. Zhao, Y. Y. Zhang, Y. H. Yu, X. Y. Shu, J. Li and X. S. Jia, *Chem. Commun.*, 2013, **49**, 1924.
20 J. L. Atwood, L. J. Barbour, P. C. Junk and G. W. Orr, *Supramol. Chem.*, 1995, **5**, 105.
21 G. Arena, A. Casnati, A. Contino, G. G. Lombardo, D. Sciotto and R. Ungaro, *Chem.–Eur. J.*, 1999, **5**, 738.
22 E. Da Silva, F. Nouar, M. Nierlich, B. Rather, M. J. Zaworotko, C. Barbey, A. Navaza and A. W. Coleman, *Cryst. Eng.*, 2003, **6**, 123.
23 L. Vial, R. F. Ludlow, J. Leclaire, R. Pérez-Fernández and S. Otto, *J. Am. Chem. Soc.*, 2006, **128**, 10253.
24 L. A. Ingerman, M. E. Cuellar and M. L. Waters, *Chem. Commun.*, 2010, **46**, 1839.
25 M. Dionisio, G. Oliviero, D. Menozzi, S. Federici, R. M. Yebeutchou, F. P. Schmidtchen, E. Dalcanale and P. Bergese, *J. Am. Chem. Soc.*, 2012, **134**, 2392.
26 S. A. Minaker, K. D. Daze, M. C. Ma and F. Hof, *J. Am. Chem. Soc.*, 2012, **134**, 11674.
27 K. D. Daze, T. Pinter, C. S. Beshara, A. Ibraheem, S. A. Minaker, M. C. F. Ma, R. J. M. Courtemanche, R. E. Campbell and F. Hof, *Chem. Sci.*, 2012, **3**, 2695.
28 M. A. Gamal-Eldin and D. H. Macartney, *Org. Biomol. Chem.*, 2013, **11**, 488.
29 L. I. James, J. E. Beaver, N. W. Rice and M. L. Waters, *J. Am. Chem. Soc.*, 2013, **135**, 6450.
30 N. K. Pinkin and M. L. Waters, *Org. Biomol. Chem.*, 2014, **12**, 7059.
31 S. Tabet, S. F. Douglas, K. D. Daze, G. A. Garnett, K. J. H. Allen, E. M. M. Abrioux, T. H. Quon, J. E. Wulff, F. Hof and F. Biorg, *Med. Chem.*, 2013, **21**, 7004.
32 H. F. Allen, K. D. Daze, T. Shimbo, A. Lai, C. A. Musselman, J. K. Sims, P. A. Wade, F. Hof and T. G. Kutateladze, *Biochem. J.*, 2014, **459**, 505.
33 K. Fucke, K. M. Anderson, M. H. Filby, M. Henry, J. Wright, S. A. Mason, M. J. Gutmann, L. J. Barbour, C. Oliver, A. W. Coleman, J. L. Atwood, J. A. Howard and J. W. Steed, *Chem.–Eur. J.*, 2011, **17**, 10259.
34 K. D. Daze, M. C. F. Ma, F. Pineux and F. Hof, *Org. Lett.*, 2012, **14**, 1512.
35 R. E. McGovern, A. A. McCarthy and P. B. Crowley, *Chem. Commun.*, 2014, **50**, 10412.
36 X. Salvatella, M. Martinell, M. Gairí, M. G. Mateu, M. Feliz, A. D. Hamilton, J. De Mendoza and E. Giralt, *Angew. Chem., Int. Ed. Engl.*, 2004, **43**, 196.
37 J. M. Chinai, A. B. Taylor, L. M. Ryno, N. D. Hargreaves, C. A. Morris, P. J. Hart and A. R. Urbach, *J. Am. Chem. Soc.*, 2011, **133**, 8810.
38 T. M. Hunter, I. W. McNae, X. Liang, J. Bella, S. Parsons, M. D. Walkinshaw and P. J. Sadler, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 2288.
39 J. Muldoon, A. E. Ashcroft and A. J. Wilson, *Chem.–Eur.J.*, 2010, **16**, 100.
40 T. Muraoka, K. Adachi, M. Ui, S. Kawasaki, N. Sadhukhan, H. Obara, H. Tochio, M. Shirakawa and K. Kinbara, *Angew. Chem., Int. Ed. Engl.*, 2013, **52**, 2430.
41 M. Calvaresi, F. Arnesano, S. Bonacchi, A. Bottoni, V. Calò, S. Conte, G. Falini, S. Fermani, M. Losacco, M. Montalti, G. Natile, L. Prodi, F. Sparla and F. Zerbetto, *ACS Nano*, 2014, **8**, 1871.
42 T. A. Gerken, J. E. Jentoft, N. Jentoft and D. G. Dearborn, *J. Biol. Chem.*, 1982, **257**, 2894.
43 W. R. Rypniewski, H. M. Holden and I. Rayment, *Biochemistry*, 1993, **32**, 9851.
44 M. A. Macnaughtan, A. M. Kane and J. H. Prestegard, *J. Am. Chem. Soc.*, 2005, **127**, 17626.
45 S. J. Abraham, T. Kobayashi, R. J. Solaro and V. Gaponenko, *J. Biomol. NMR*, 2009, **43**, 239.
46 S. T. Larda, M. P. Bokoch, F. Evanics and R. S. Prosser, *J. Biomol. NMR*, 2012, **54**, 199.
47 M. S. Lehmann and R. F. D. Stansfield, *Biochemistry*, 1989, **28**, 7028.
48 S. Bhattacharjya and P. Balaram, *Proteins*, 1997, **29**, 492–507.
49 D. A. Dougherty, *Acc. Chem. Res.*, 2013, **46**, 885.
50 B. Kuhn, P. Gerber, T. Schulz-Gasch and M. Stahl, *J. Med. Chem.*, 2005, **48**, 4040.
51 B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke and A. E. Roitberg, *J. Chem. Theory Comput.*, 2012, **8**, 3314.
52 A. G. W. Leslie, Recent changes to the MOSFLM package for processing film and image plate data, *Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography*, 1992.

53 P. Evans, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2006, **62**, 72.
54 G. Winter, *J. Appl. Crystallogr.*, 2010, **43**, 186.
55 W. Kabsch, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 125.
56 Collaborative Computational Project, N. *The CCP4 suite: programs for protein crystallography, Acta Crystallogr. D*, 1994, **50**, 760.
57 P. Emsley and K. Cowtan, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2004, **60**, 2126.
58 E. J. van Asselt, A. Perrakis, K. H. Kalk, V. S. Lamzin and B. W. Dijkstra, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1998, **54**, 58.
59 V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 2010, **66**, 12.
60 G. V. Papamokos, G. Tziatzos, D. G. Papageorgiou, S. D. Georgatos, A. S. Politou and E. Kaxiras, *Biophys. J.*, 2012, **102**, 1926.
61 I. S. Joung and T. E. Cheatham III, *J. Phys. Chem. B*, 2008, **112**, 9020.
62 I. S. Joung and T. E. Cheatham III, *J. Phys. Chem. B*, 2009, **113**, 13279.

4.3 Additional Data and Inferences

The theoretical predictions in this publication were almost exclusively done using short (10 ns) replicate molecular dynamics trajectories with the hopes that the complex stability would systematically rule out non-binding locations. This was surprisingly effective for this problem. Earlier attempts to dock the calixarene using a blind docking approach with AutoDock 4.2 placed the calixarene upside down in the dense postively charged region between K97Me$_2$ and K96Me$_2$. The analysis of complex stability was done using the best-fit plane as described in the publication. An illustration of the reference plane is shown in Fig 4.1.

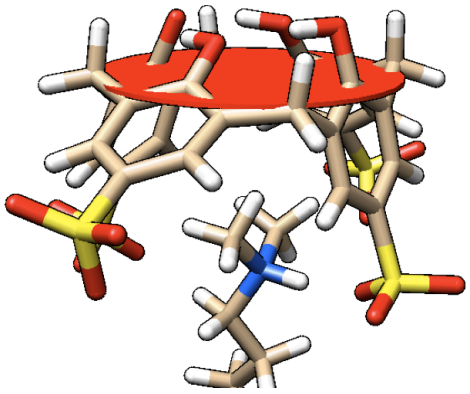


Figure 4.1: Best fit plane for Calixarene to KMe$_2$ Nitrogen

The method of following along the complex stability along with preliminary free energy calculations originally led us to believe that K97Me$_2$ was the most likely binding site. This was until we had considered the cost of the host reorganization. Energetic contributions of this cost were not directly assessed, however, the displacement of the methylated lysine sites upon binding was used as a way to assess this potential cost. Table 4.3 shows the differences between surface areas of the various host binding site residues. K97 shows one of the larges changes in solvent accessible surface area, the likely source of binding penalty that lends K116 its selectivity.

Residue Number	Unbound SASA/ Å^2	Bound SASA/ Å^2	Δ SASA/ Å^2
1	226.4	165	61.4
13	153	150.4	2.6
33	102.2	100.7	1.5
96	106.5	101.6	4.9
97	163	138.6	24.4
116	182	164.7	17.3

Table 4.1: HEWL Unbound and Bound Dimethyllysine Surface Areas. Differences between residue SASAs were averaged over the MD frames where the calixarene ligand was bound. The differences between the SASAs indicate that K97 is one of the more strained residues upon binding.

Chapter 5

Publication: Selective Inhibition of CBX6: A Methyllysine Reader Protein in the Polycomb Family

5.1 Preface

In this publication, selective features of inhibitors for CBX6 versus CBX7 were explored by fluorescence polarization, SPR, as well as molecular dynamics. All computational simulations and interpretations of the theoretical results were performed solely by James McFarlane. This work included the parameterization of the non-standard amino acids in the peptide ligands and construction of the ligands as well. Computational work also included molecular docking and molecular dynamics simulations and subsequent data visualization for structural metric changes during binding. For remaining author contributions, please refer to page 142 of the following publication.

5.2 Publication

Full publication including links to supplementary information may be found through the following link:

https://doi.org/10.1021/acsmedchemlett.5b00378

ACS Medicinal Chemistry Letters

Letter

pubs.acs.org/acsmedchemlett

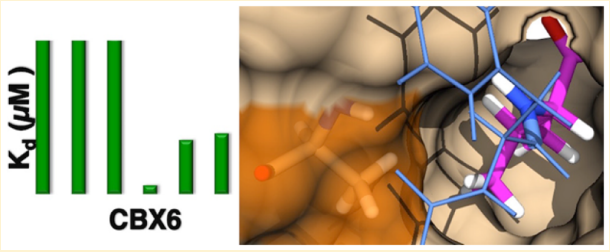
Selective Inhibition of CBX6: A Methyllysine Reader Protein in the Polycomb Family

Natalia Milosevich, Michael C. Gignac, James McFarlane, Chakravarthi Simhadri, Shanti Horvath, Kevin D. Daze, Caitlin S. Croft, Aman Dheri, Taylor T. H. Quon, Sarah F. Douglas, Jeremy E. Wulff, Irina Paci,* and Fraser Hof*

Department of Chemistry, University of Victoria, Victoria, V8W 3V6, Canada

Supporting Information

ABSTRACT: The polycomb paralogs CBX2, CBX4, CBX6, CBX7, and CBX8 are epigenetic readers that rely on "aromatic cage" motifs to engage their partners' methyllysine side chains. Each CBX carries out distinct functions, yet each includes a highly similar methyllysine-reading chromodomain as a key element. CBX7 is the only chromodomain that has yet been targeted by chemical inhibition. We report a small set of peptidomimetic agents in which a simple chemical modification switches the ligands from one with promiscuity across all polycomb paralogs to one that provides selective inhibition of CBX6. The structural basis for this selectivity, which involves occupancy of a small hydrophobic pocket adjacent to the aromatic cage, was confirmed through molecular dynamics simulations. Our results demonstrate the increases in affinity and selectivity generated by ligands that engage extended regions of chromodomain binding surfaces.



KEYWORDS: *Epigenetics, methyllysine reader proteins, Polycomb Group proteins, CBX6, peptidomimetics*

Post-translational methylation exerts critical control over multiple gene expression pathways.[1] Histone methylation is among the most prominent and diverse of post-translational modifications (PTMs) in which differences in the location and degree of methylation dictate the engagement of different partners and give varying downstream biological effects.[2,3] The *Drosophila* protein polycomb protein reads trimethylation marks on histone tails and is the namesake of the polycomb group (PcG), a set of functionally diverse proteins that coordinate to modify gene expression at hundreds of loci.[4] There are five human paralogs of polycomb, CBX2, CBX4, CBX6, CBX7, and CBX8, each with distinct functions in cellular differentiation during development, cancer progression, and stem cell maintenance.[5−8] The canonical functions of CBX proteins involve participation in variations of polycomb repressive complex 1 (PRC1), within which they serve as readers of the mark histone 3, lysine 27 trimethylated (H3K27me3). In spite of functional differences, all polycomb paralogs rely on a common methyllysine reader module (chromodomain) with high structure and sequence similarity.[9−11]

Epigenetic reader proteins are growing as a class of potential drug targets. Inhibitors of acetyllysine-binding bromodomains (BRD) are well-known; clinical trials for diverse malignancies are underway, and their promise in control of inflammation, cancer, and viral infection is being actively explored.[12−24] There are hundreds of methyl reader proteins, but progress in inhibiting methyllysine readers has been comparatively slow. The first examples were inhibitors of methyllysine-binding Malignant Brain Tumor (MBT) domains.[25−29] Other recent reports of methyllysine reader protein inhibitors include agents targeting one tudor domain[30] and two PHD fingers.[31,32]

In the family of chromatin organization modifier domains (chromodomains), the initial focus has been on CBX7, one of the five human polycomb paralogs. (The proteins CBX1, CBX3, and CBX5 are a more distantly related set that are heterochromatin protein 1 (HP1) paralogs). Our group reported peptidomimetic inhibitors developed from peptide leads,[33] and the group of Zhou recently reported on small molecule inhibitors discovered through high-throughput screening.[34] The initial focus on CBX7 is partly because it is strongly associated with many disease phenotypes, and partly also because it tends to give higher in vitro affinities for its native ligands than the other CBX proteins.[9,10] This characteristic has also been identified in a computational analysis that predicted that CBX7 would have a relatively "druggable" binding site among methyl reader proteins.[35]

We sought to identify selective chemical or peptidic tools that would overcome this bias and target other members within the polycomb CBX family of epigenetic modifiers. The sequence and structural similarities within the human polycomb chromodomains are very high (Figure S1). Their highly diverse

Special Issue: Epigenetics

Received: September 29, 2015
Accepted: December 7, 2015
Published: December 7, 2015

DOI: 10.1021/acsmedchemlett.5b00378
ACS Med. Chem. Lett. 2016, 7, 139−144

in vivo functions (above) are partially understood as arising from their divergent domain architecture outside of their methyl-reading chromodomains.[11] However, there are no known sites for chemical binding outside of the chromodomains, so any efforts to create selective ligands must rely on being able to discriminate among the five highly similar CBX chromodomains.

The overall similarity of chromodomains makes their selective inhibition challenging. The aromatic cage pockets that bind to the histone's Kme3 side chain are essentially indistinguishable among the five polycomb paralogs (Figure 1A,B). A small, hydrophobic adjacent pocket binds the

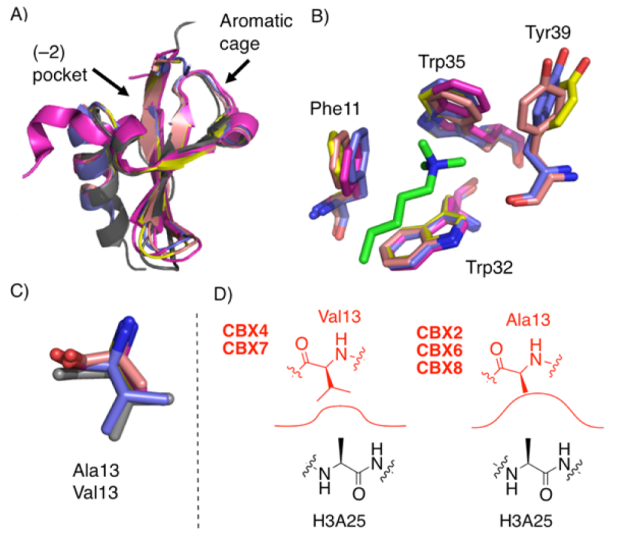


Figure 1. Overview of the chromodomains of human polycomb paralogs CBX2, CBX4, CBX6, CBX7, and CBX8. (A) Structural alignments show the overall similarity of CBX proteins (magenta = CBX2, pdb code 3H91; gray = CBX4, pdb code 3I8Z; yellow = CBX6, pdb code 3I90; purple = CBX7, pdb code 4MN3; salmon = CBX8, pdb code 3I91). (B) Overlay of the aromatic cages with the Kme3 native ligand in green. (CBX4 not shown, as the only available structure lacks bound Kme3 ligand.) (C) Overlay of the (−2) pocket floor Val/Ala residues. (D) Depiction of (−2) pocket size in each CBX protein with histone 3 alanine 25 as the native ligand. CBX7 numbering used throughout.

conserved histone Ala side chains that uniformly occur (−2) to the trimethyllysine sites H3K9me3 and H3K27me3, the two sites known to be targeted in vitro by CBX proteins. The (−2) pocket is created and shielded from solvent by the closure of two residue side chains over the ligand in a motif called the hydrophobic clasp.[9,33] The floor of the (−2) pocket is partially defined by a Val residue in CBX4 and CBX7 that is replaced by an Ala residue in CBX2, -6, and -8 (Figure 1C,D). It was recently shown that a Val/Ala exchange at this position can make CBX7 display CBX2-like binding affinities and functions, demonstrating the importance of these residues at the floor of the (−2) pocket in the CBX proteins' intrinsic biological functions.[8]

We report here that varying the ligand structure within the (−2) pocket has a large influence on CBX selectivity and in fact allows for the simple creation of potent CBX6-selective inhibitors.

We first established a panel of proteins for use in fluorescence polarization assays of ligand affinities and selectivities. Recombinant chromodomains for each polycomb paralog (CBX2, CBX4, CBX6, CBX7, and CBX8) were expressed and purified using minor modifications of reported protocols.[9] We also prepared one member of the related HP1 family (CBX1; HP1β) as a representative from this more distantly related set of CBX proteins.

In spite of the canon that defines H3K9me3 and/or H3K27me3 as the targets of CBX proteins, H3K9me3 and H3K27me3 peptides have been shown not to bind measurably with multiple members of the CBX family.[9,10] In order to ensure strong baseline affinity for our ligands, we started with a peptidic sequence (**1**) that we previously identified as a moderate-strength CBX7 binder (IC_{50} = 73 μM) and a chemically modified version (**2**) that has improved CBX7 affinity (IC_{50} = 1.7 μM) arising from a *p*-bromobenzamide group (Figure 2 and Figure S11).[33] Competitive fluorescence

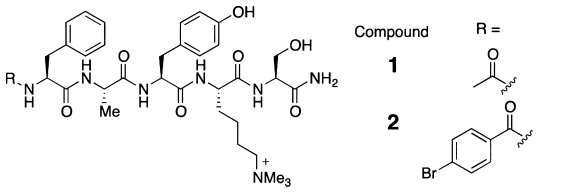


Figure 2. Peptide sequences Ac-FAYKme$_3$S-NH$_2$ (**1**) and pBr-FAYKme$_3$S-NH$_2$ (**2**) identified as CBX7 binders.

polarization (FP) studies are limited in their ability to measure K_d values for potent ligands, *especially for comparison across different proteins that have different intrinsic affinities for the FP probes.*[36] To overcome this limitation by using direct FP titrations of an entire panel of proteins into ligands (which gives reliable K_d values for all complexes), we modified **2** by adding a linking residue and fluorescent dye at the C-terminus. Compounds **3**, **4**, and **5** are such compounds that vary only in the identity of the side chain directed into the (−2) pocket.

Figure 3 shows K_d values arising from direct FP titrations of all six chromodomains into all three ligands (see also Figures S13−15). The (−2) substitutions in **3**−**5** have dramatic effects on potencies and selectivity for CBX proteins. Compound **3**, bearing the methyl substituent at the (−2) position, was potent and promiscuous. Its K_d values are, from strongest to weakest, CBX4/7 < CBX2/6 < CBX8. However, even the affinity for the weakest partner, at 1 μM, is >25-fold stronger than the affinities of any small molecule inhibitor for any chromodomain yet reported. Inhibitor **3** is moderately selective for all polycomb paralogs (0.1 to 1 μM) over the HP1 paralog CBX1 (5 μM).

Addition of the ethyl substituent in **4** weakened binding to CBX1 (HP1β) by >10-fold, while generating smaller decreases in binding potency to CBX2/4/6/7 and no change in binding to CBX8. The isopropyl substitution in compound **5** decreased binding affinity of the peptide to all of CBX2/4/7/8, while not significantly changing binding to CBX6. Compound **5** is 90×, 20×, 18×, 6×, and 7× selective for CBX6 over CBX1/2/4/7/8, respectively. Analogues of compounds **3** and **5** lacking FITC labels were tested using a competitive FP assay (Figure S12). The IC_{50} values determined showed 7-fold selectivity for CBX6 over CBX7 for unlabeled **5**, demonstrating that the FITC tag alone is not the source of CBX6 selectivity.

We further confirmed the affinities and selectivities of **3** and **5** by preparing **3**-biotin and **5**-biotin to enable orthogonal characterization of the complexes by surface plasmon resonance (SPR). The results echo the selectivity trends obtained by

DOI: 10.1021/acsmedchemlett.5b00378
ACS Med. Chem. Lett. 2016, 7, 139−144

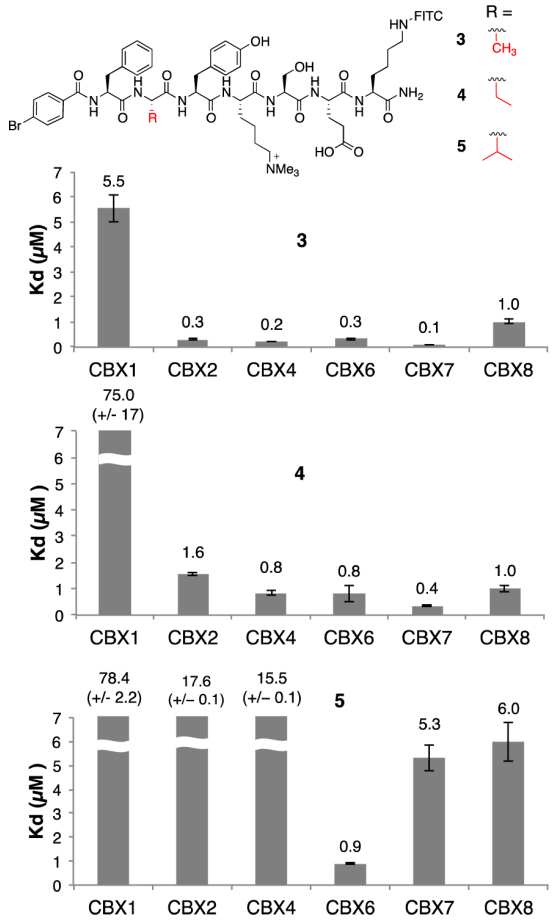


Figure 3. Series of CBX ligands with varying alkyl substitutions at the (−2) position and binding affinities determined by direct FP. Standard errors are shown as error bars or in parentheses for values that exceed axis limits.

direct FP, while also generally agreeing with the absolute K_d values (Table 1 and Figures S16 and S17).

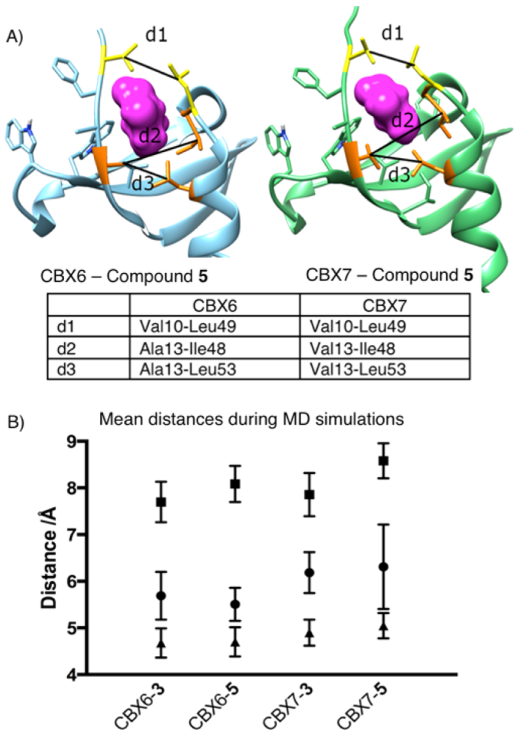
Table 1. Binding Affinities Determined by SPR

compd	protein	K_d (μM)
3-biotin	CBX6	0.11 ± 0.01
	CBX7	0.16 ± 0.01
	CBX7-V13A	0.29 ± 0.05
5-biotin	CBX6	0.91 ± 0.05
	CBX7	34 ± 8
	CBX7-V13A	24 ± 2

To probe the role of Val/Ala substitutions at the floor of the (−2) pocket (Figure 1C,D) in defining CBX protein selectivity, the CBX7-V13A mutant was prepared and tested by SPR (Table 1). The weak binding to compound **5** by this CBX6-like mutation of CBX7 shows that the V13A substitution alone is not sufficient to drive CBX6-like selective binding of **5**. This is consistent with the observed low potency of **5** for binding CBX2 and CBX8, which also have the V13A substitution but differ in other residues.

We carried out energy minimizations (Moloc) and ligand docking (SeeSAR) for ligands **3** and **5** in order to gain more insight into this SAR. Both methods suggested that the large side chain on ligand **5** could be accommodated in identical modes and with identical energies in the (−2) pockets of CBX6 and CBX7, confounding the simple picture shown in Figure 1D. The overall picture is that the (−2) pocket is critical for the native binding preferences of CBX proteins, but that the way in which ligand **5** provides for CBX6 selectivity by occupying this pocket is too complicated to be understood using static X-ray structures.

To gain a dynamic view of the complexes, we carried out MD simulations of the complexes of CBX6 and CBX7 with each of **3** and **5**. The trajectories show that the complex of CBX7 with ligand **5** (mismatched) undergoes opening of the hydrophobic clasp that envelopes the ligand. In contrast, the complex of CBX6 with **5** (matched) remains completely wrapped around the ligand throughout the simulation, as illustrated by clasp distance d1 in Figure 4 (see also Supporting Information for



	CBX6	CBX7
d1	Val10-Leu49	Val10-Leu49
d2	Ala13-Ile48	Val13-Ile48
d3	Ala13-Leu53	Val13-Leu53

Figure 4. Molecular dynamic simulation results showing the change in distances within CBX6 and CBX7 when in complex with compounds **3** and **5**. (A) Representative snapshots from MD trajectories, showing only those parts of the ligands that occupy the (−2) pocket as magenta surfaces. The distances d1, d2, and d3 define changes in pocket shape that can be compared between simulations (see text). (B) Mean values for distances d1 (circle), d2 (square), and d3 (triangle) show the changes induced in and around the (−2) pocket when the bulkier compound **5** is bound to CBX6 and CBX7. See also the movies of complete MD simulations (Supporting Information).

movies). Figure 4 includes other geometric parameters that quantify how CBX6 and CBX7 differ in their engagement of ligands. The ligand side chains always stay in the (−2) pockets, but the pocket shapes respond in different ways to different ligands. The overall "external pocket width" (d3 in Figure 4) is the same in all complexes, showing that the mouth of the pocket is (surprisingly) the same regardless of the Val13Ala swap at one edge of this measured distance. The "internal pocket width" (d2 in Figure 4) and hydrophobic clasp distance

DOI: 10.1021/acsmedchemlett.5b00378
ACS Med. Chem. Lett. 2016, 7, 139−144

(d1) both show significant increases for CBX7-**5** that are not observed for CBX6-**5**, showing increased strains that account for the observed experimental selectivity.

The aromatic cage of CBX proteins are known to be important binding sites, but our results show how ligand affinity also depends on occupying other binding subsites. The histone tail ligands in native CBX−histone complexes occupy the beta groove and interdigitate between the existing two protein strands in order to make a short three-strand beta-sheet motif. The previously reported small molecule ligands for CBX7 engage only the region around the aromatic cage and have modest potencies (28−67 μM).[34] The modified peptidic ligands we report here reach affinities as strong as 0.1 μM (for the complex CBX7-**3**) and are routinely sub-μM. We attribute this to their occupation of the beta groove, which includes the aforementioned (−2) pocket as well as a hydrophobic cleft that extends further from the Kme3 binding site. Comparison of **1** and **2** shows that occupying that cleft provides >40-fold enhanced potency.

The origins of selectivities for the small molecule vs the peptidic ligands are more subtle. The small-molecule ligand of Zhou shows impressive 3- to 22-fold selectivities for CBX7 over CBX2/4/6/8 even though it mainly binds the aromatic cage region.[34] The aromatic cages of all CBXs are highly similar to each other in structure, so we infer that the aromatic cage of CBX7 has better preorganization and/or reduced solvent accessibility relative to its family members, rather than a large difference in protein−ligand interactions in the bound state. Promiscuous compound **3** shows stronger binding to CBX7 than any other CBX protein, suggesting that it is benefiting from similar effects.

Large groups in the (−2) pocket are able to overcome the inherent bias toward CBX7 binding. The particular CBX6 selectivity is not simply explained by the Val/Ala difference at position 13 of the chromodomains, as would have been predicted both by simple modeling and by the CBX2/7 results reported earlier this year. Kaustov et al. showed using a peptide array that CBX8 (but not CBX7) could accommodate a valine side chain in the (−2) position of a histone tail sequence,[9] but their qualitative array-blotting result did not include CBX6 for comparison. Our solution phase data agree with this result to an extent, in that they show similar solution-phase affinities for valine-containing ligand **5** binding to CBX7 or CBX8 (ca. 5 μM each). However, the affinities for CBX6 are higher, providing selectivity for this polycomb paralog that would not have been anticipated based on prior studies of these proteins.

These results also uncover the previously unknown binding preferences of CBX6's chromodomain. While it has been assumed to be a canonical polycomb reader of H3K9me3 and/or H3K27me3, the *in vitro* affinities of CBX6 for these marks are in fact unmeasurably weak.[9,10] Each of these histone marks has an alanine residue occupying the (−2) pocket and, according to our results, would be poorly suited to bind CBX6. Our results suggest that CBX6 might be a reader of a different, as-yet undetermined trimethyllysine site.[37] Identifying the unknown native target(s) of CBX6 would help prove that the SAR for CBX6 selectivity identified here also persists in a cellular context.

Chromodomain-containing proteins are increasingly suggested as targets for therapeutic intervention, and the functional biology of polycomb paralogs is an important frontier of epigenetics and stem cell biology. Studying the chemical biology and therapeutic potential of CBX proteins requires selective ligands that, until now, have not been available. Our results provide new inhibitors for CBX6, which has been the least studied of the human polycombs. They also inform the general design requirements of the next-generation of potent and selective small-molecule ligands for CBX proteins.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsmedchemlett.5b00378.

Synthesis, characterization, protein expression and purification, FP and SPR binding data for protein−ligand complexes, and MD simulation methods (PDF)
Captured from MD simulation of CBX6 in complex with compound **5** (MPG)
Captured from MD simulation of CBX7 in complex with compound **5** (MPG)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: ipaci@uvic.ca.
*E-mail: fhof@uvic.ca.

Author Contributions

The manuscript was written by N.M. and J.M. with guidance from I.P., J.W., and F.H. N.M. carried out compound design and synthesis, with assistance of C.C., K.D., and C.S. M.G. carried out protein expression and binding studies with assistance of A.D. S.H. carried out SPR analyses. S.D. and T.Q. carried out CBX7 mutagenesis. J.M. carried out computational studies under supervision of I.P. All authors have given approval to the final version of the manuscript.

Funding

This research was supported by a CIHR fellowship and WestCoast Motorcycle Ride to Live award to N.M., Cancer Research Society Grant 19284, and Prostate Cancer Canada Movember Discovery Grant D2013−18,

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Plasmids were gifts of Cheryl Arrowsmith (Addgene plasmids 25245, 25158, 25237, 25296, 25241, and 62514). We thank Anna Hirsch and Milon Mondal for help with SeeSAR docking. F.H. and J.W. thank the Canada Research Chairs program. The computational component of this work was enabled in part by WestGrid (www.westgrid.ca) and Compute Canada/Calcul Canada (www.computecanada.ca).

■ ABBREVIATIONS

PTM, post-translational modifications; CBX, chromobox; HP1, heterochromatin protein; PRC1, polycomb repressive complex 1; BRD, bromodomain; MBT, malignant brain tumor; PHD, plant homeodomain; FP, fluorescence polarization; SPR, surface plasmon resonance; MD, molecular dynamics

■ REFERENCES

(1) Berger, S. L.; Kouzarides, T.; Shiekhattar, R.; Shilatifard, A. An Operational Definition of Epigenetics. *Genes Dev.* **2009**, *23*, 781−783.
(2) Jenuwein, T.; Allis, C. D. Translating the Histone Code. *Science* **2001**, *293*, 1074−1080.

DOI: 10.1021/acsmedchemlett.5b00378
ACS Med. Chem. Lett. 2016, 7, 139−144

(3) Huang, Y.; Fang, J.; Bedford, M. T.; Zhang, Y.; Xu, R. M. Recognition of Histone H3 Lysine-4 Methylation by the Double Tudor Domain of Jmjd2a. *Science* **2006**, *312*, 748−751.
(4) Gieni, R. S.; Hendzel, M. J. Polycomb Group Protein Gene Silencing, Non-Coding Rna, Stem Cells, and Cancer. *Biochem. Cell Biol.* **2009**, *87*, 711−746.
(5) O'Loghlen, A.; Munoz-Cabello, A. M.; Gaspar-Maia, A.; Wu, H. A.; Banito, A.; Kunowska, N.; Racek, T.; Pemberton, H. N.; Beolchi, P.; Lavial, F.; Masui, O.; Vermeulen, M.; Carroll, T.; Graumann, J.; Heard, E.; Dillon, N.; Azuara, V.; Snijders, A. P.; Peters, G.; Bernstein, E.; Gil, J. Microrna Regulation of Cbx7Mediates a Switch of Polycomb Orthologs During Esc Differentiation. *Cell Stem Cell* **2012**, *10*, 33−46.
(6) Morey, L.; Aloia, L.; Cozzuto, L.; Benitah, S. A.; Di Croce, L. Rybp and Cbx7 Define Specific Biological Functions of Polycomb Complexes in Mouse Embryonic Stem Cells. *Cell Rep.* **2013**, *3*, 60−69.
(7) Klauke, K.; Radulovic, V.; Broekhuis, M.; Weersing, E.; Zwart, E.; Olthof, S.; Ritsema, M.; Bruggeman, S.; Wu, X.; Helin, K.; Bystrykh, L.; de Haan, G. Polycomb Cbx Family Members Mediate the Balance between Haematopoietic Stem Cell Self-Renewal and Differentiation. *Nat. Cell Biol.* **2013**, *15*, 353−362.
(8) Tardat, M.; Albert, M.; Kunzmann, R.; Liu, Z.; Kaustov, L.; Thierry, R.; Duan, S.; Brykczynska, U.; Arrowsmith, C. H.; Peters, A. H. Cbx2 Targets Prc1 to Constitutive Heterochromatin in Mouse Zygotes in a Parent-of-Origin-Dependent Manner. *Mol. Cell* **2015**, *58*, 157.
(9) Kaustov, L.; Ouyang, H.; Amaya, M.; Lemak, A.; Nady, N.; Duan, S.; Wasney, G. A.; Li, Z.; Vedadi, M.; Schapira, M.; Min, J.; Arrowsmith, C. H. Recognition and Specificity Determinants of the Human Cbx Chromodomains. *J. Biol. Chem.* **2011**, *286*, 521−529.
(10) Bernstein, E.; Duncan, E. M.; Masui, O.; Gil, J.; Heard, E.; Allis, C. D. Mouse Polycomb Proteins Bind Differentially to Methylated Histone H3 and Rna and Are Enriched in Facultative Heterochromatin. *Mol. Cell. Biol.* **2006**, *26*, 2560−2569.
(11) Senthilkumar, R.; Mishra, R. K. Novel Motifs Distinguish Multiple Homologues of Polycomb in Vertebrates: Expansion and Diversification of the Epigenetic Toolkit. *BMC Genomics* **2009**, *10*, 549.
(12) Bamborough, P.; Diallo, H.; Goodacre, J. D.; Gordon, L.; Lewis, A.; Seal, J. T.; Wilson, D. M.; Woodrow, M. D.; Chung, C. W. Fragment-Based Discovery of Bromodomain Inhibitors Part 2: Optimization of Phenylisoxazole Sulfonamides. *J. Med. Chem.* **2012**, *55*, 587−596.
(13) Banerjee, C.; Archin, N.; Michaels, D.; Belkina, A. C.; Denis, G. V.; Bradner, J.; Sebastiani, P.; Margolis, D. M.; Montano, M. Bet Bromodomain Inhibition as a Novel Strategy for Reactivation of Hiv-1. *J. Leukocyte Biol.* **2012**, *92*, 1147−1154.
(14) Chung, C. W.; Dean, A. W.; Woolven, J. M.; Bamborough, P. Fragment-Based Discovery of Bromodomain Inhibitors Part 1: Inhibitor Binding Modes and Implications for Lead Discovery. *J. Med. Chem.* **2012**, *55*, 576−586.
(15) Fish, P. V.; Filippakopoulos, P.; Bish, G.; Brennan, P. E.; Bunnage, M. E.; Cook, A. S.; Federov, O.; Gerstenberger, B. S.; Jones, H.; Knapp, S.; Marsden, B.; Nocka, K.; Owen, D. R.; Philpott, M.; Picaud, S.; Primiano, M. J.; Ralph, M. J.; Sciammetta, N.; Trzupek, J. D. Identification of a Chemical Probe for Bromo and Extra C-Terminal Bromodomain Inhibition through Optimization of a Fragment-Derived Hit. *J. Med. Chem.* **2012**, *55*, 9831−9837.
(16) Gehling, V. S.; Hewitt, M. C.; Vaswani, R. G.; Leblanc, Y.; Cote, A.; Nasveschuk, C. G.; Taylor, A. M.; Harmange, J. C.; Audia, J. E.; Pardo, E.; Joshi, S.; Sandy, P.; Mertz, J. A.; Sims, R. J., 3rd; Bergeron, L.; Bryant, B. M.; Bellon, S.; Poy, F.; Jayaram, H.; Sankaranarayanan, R.; Yellapantula, S.; Bangalore Srinivasamurthy, N.; Birudukota, S.; Albrecht, B. K. Discovery, Design, and Optimization of Isoxazole Azepine Bet Inhibitors. *ACS Med. Chem. Lett.* **2013**, *4*, 835−840.
(17) Gehling, V. S.; Hewitt, M. C.; Vaswani, R. G.; Leblanc, Y.; Cote, A.; Nasveschuk, C. G.; Taylor, A. M.; Harmange, J. C.; Audia, J. E.; Pardo, E.; Joshi, S.; Sandy, P.; Mertz, J. A.; Sims, R. J., 3rd; Bergeron, L.; Bryant, B. M.; Bellon, S.; Poy, F.; Jayaram, H.; Sankaranarayanan, R.; Yellapantula, S.; Bangalore Srinivasamurthy, N.; Birudukota, S.; Albrecht, B. K. Discovery, Design, and Optimization of Isoxazole Azepine Bet Inhibitors. *ACS Med. Chem. Lett.* **2013**, *4*, 835−840.
(18) Mirguet, O.; Gosmini, R.; Toum, J.; Clement, C. A.; Barnathan, M.; Brusq, J. M.; Mordaunt, J. E.; Grimes, R. M.; Crowe, M.; Pineau, O.; Ajakane, M.; Daugan, A.; Jeffrey, P.; Cutler, L.; Haynes, A. C.; Smithers, N. N.; Chung, C. W.; Bamborough, P.; Uings, I. J.; Lewis, A.; Witherington, J.; Parr, N.; Prinjha, R. K.; Nicodeme, E. Discovery of Epigenetic Regulator I-Bet762: Lead Optimization to Afford a Clinical Candidate Inhibitor of the Bet Bromodomains. *J. Med. Chem.* **2013**, *56*, 7501−7515.
(19) Puissant, A.; Frumm, S. M.; Alexe, G.; Bassil, C. F.; Qi, J.; Chanthery, Y. H.; Nekritz, E. A.; Zeid, R.; Gustafson, W. C.; Greninger, P.; Garnett, M. J.; McDermott, U.; Benes, C. H.; Kung, A. L.; Weiss, W. A.; Bradner, J. E.; Stegmaier, K. Targeting Mycn in Neuroblastoma by Bet Bromodomain Inhibition. *Cancer Discovery* **2013**, *3*, 308−323.
(20) Asangani, I. A.; Dommeti, V. L.; Wang, X.; Malik, R.; Cieslik, M.; Yang, R.; Escara-Wilke, J.; Wilder-Romans, K.; Dhanireddy, S.; Engelke, C.; Iyer, M. K.; Jing, X.; Wu, Y. M.; Cao, X.; Qin, Z. S.; Wang, S.; Feng, F. Y.; Chinnaiyan, A. M. Therapeutic Targeting of Bet Bromodomain Proteins in Castration-Resistant Prostate Cancer. *Nature* **2014**, *510*, 278−282.
(21) Smith, S. G.; Sanchez, R.; Zhou, M. M. Privileged Diazepine Compounds and Their Emergence as Bromodomain Inhibitors. *Chem. Biol.* **2014**, *21*, 573−583.
(22) Filippakopoulos, P.; Qi, J.; Picaud, S.; Shen, Y.; Smith, W. B.; Fedorov, O.; Morse, E. M.; Keates, T.; Hickman, T. T.; Felletar, I.; Philpott, M.; Munro, S.; McKeown, M. R.; Wang, Y.; Christie, A. L.; West, N.; Cameron, M. J.; Schwartz, B.; Heightman, T. D.; La Thangue, N.; French, C. A.; Wiest, O.; Kung, A. L.; Knapp, S.; Bradner, J. E. Selective Inhibition of Bet Bromodomains. *Nature* **2010**, *468*, 1067−1073.
(23) Muller, S.; Knapp, S. Discovery of Bet Bromodomain Inhibitors and Their Role in Target Validation. *MedChemComm* **2014**, *5*, 288−296.
(24) Garnier, J. M.; Sharp, P. P.; Burns, C. J. Bet Bromodomain Inhibitors: A Patent Review. *Expert Opin. Ther. Pat.* **2014**, *24*, 185−199.
(25) Herold, J. M.; Wigle, T. J.; Norris, J. L.; Lam, R.; Korboukh, V. K.; Gao, C.; Ingerman, L. A.; Kireev, D. B.; Senisterra, G.; Vedadi, M.; Tripathy, A.; Brown, P. J.; Arrowsmith, C. H.; Jin, J.; Janzen, W. P.; Frye, S. V. Small-Molecule Ligands of Methyl-Lysine Binding Proteins. *J. Med. Chem.* **2011**, *54*, 2504−2511.
(26) Camerino, M. A.; Zhong, N.; Dong, A.; Dickson, B.; James, L.; Baughman, B.; Norris, J.; Kireev, D.; Janzen, W. P.; Arrowsmith, C.; Frye, S. V. The Structure-Activity Relationships of L3mbtl3 Inhibitors: A Second Series of Potent Compounds Which Bind the L3mbtl3 Dimer. *MedChemComm* **2013**, *4*, 1501−1507.
(27) James, L. I.; Barsyte-Lovejoy, D.; Zhong, N.; Krichevsky, L.; Korboukh, V. K.; Herold, J. M.; MacNevin, C. J.; Norris, J. L.; Sagum, C. A.; Tempel, W.; Marcon, E.; Guo, H. B.; Gao, C.; Huang, X. P.; Duan, S. L.; Emili, A.; Greenblatt, J. F.; Kireev, D. B.; Jin, J.; Janzen, W. P.; Brown, P. J.; Bedford, M. T.; Arrowsmith, C. H.; Frye, S. V. Discovery of a Chemical Probe for the L3mbtl3Methyllysine Reader Domain. *Nat. Chem. Biol.* **2013**, *9*, 184−191.
(28) James, L. I.; Korboukh, V. K.; Krichevsky, L.; Baughman, B. M.; Herold, J. M.; Norris, J. L.; Jin, J.; Kireev, D. B.; Janzen, W. P.; Arrowsmith, C. H.; Frye, S. V. Small-Molecule Ligands of Methyl-Lysine Binding Proteins: Optimization of Selectivity for L3mbtl3. *J. Med. Chem.* **2013**, *56*, 7358−7371.
(29) Gao, C.; Herold, J. M.; Kireev, D.; Wigle, T.; Norris, J. L.; Frye, S. Biophysical Probes Reveal a "Compromise" Nature of the Methyl-Lysine Binding Pocket in L3mbtl1. *J. Am. Chem. Soc.* **2011**, *133*, 5357−5362.
(30) Perfetti, M. T.; Baughman, B. M.; Dickson, B. M.; Mu, Y.; Cui, G.; Mader, P.; Dong, A.; Norris, J. L.; Rothbart, S. B.; Strahl, B. D.; Brown, P. J.; Janzen, W. P.; Arrowsmith, C. H.; Mer, G.; McBride, K. M.; James, L. I.; Frye, S. V. Identification of a Fragment-Like Small

DOI: 10.1021/acsmedchemlett.5b00378
ACS Med. Chem. Lett. 2016, 7, 139−144

Molecule Ligand for the Methyl-Lysine Binding Protein, 53bp1. *ACS Chem. Biol.* **2015**, *10*, 1072.
(31) Miller, T. C. R.; Rutherford, T. J.; Birchall, K.; Chugh, J.; Fiedler, M.; Bienz, M. Competitive Binding of a Benzimidazole to the Histone-Binding Pocket of the Pygo Phd Finger. *ACS Chem. Biol.* **2014**, *9*, 2864−2874.
(32) Wagner, E. K.; Nath, N.; Flemming, R.; Feltenberger, J. B.; Denu, J. M. Identification and Characterization of Small Molecule Inhibitors of a Plant Homeodomain Finger. *Biochemistry* **2012**, *51*, 8293−8306.
(33) Simhadri, C.; Daze, K. D.; Douglas, S. F.; Quon, T. T. H.; Dev, A.; Gignac, M. C.; Peng, F. N.; Heller, M.; Boulanger, M. J.; Wulff, J. E.; Hof, F. Chromodomain Antagonists That Target the Polycomb-Group Methyllysine Reader Protein Chromobox Homo Log 7 (Cbx7). *J. Med. Chem.* **2014**, *57*, 2874−2883.
(34) Ren, C.; Morohashi, K.; Plotnikov, A. N.; Jakoncic, J.; Smith, S. G.; Li, J.; Zeng, L.; Rodriguez, Y.; Stojanoff, V.; Walsh, M.; Zhou, M. M. Small-Molecule Modulators of Methyl-Lysine Binding for the Cbx7 Chromodomain. *Chem. Biol.* **2015**, *22*, 161−168.
(35) Santiago, C.; Nguyen, K.; Schapira, M. Druggability of Methyl-Lysine Binding Sites. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 1171−1178.
(36) Huang, X. Fluorescence Polarization Competition Assay: The Range of Resolvable Inhibitor Potency Is Limited by the Affinity of the Fluorescent Ligand. *J. Biomol. Screening* **2003**, *8*, 34−38.
(37) This is indirectly anticipated by a massive parallel computational docking effort that was backed up by microarray binding data. Li, N.; Stein, R. S. L.; He, W.; Komives, E.; Wang, W. Identification of methyllysine peptides binding to chromobox protein homolog 6 chromodomain in the human proteome. *Mol. Cell. Proteomics* **2013**, *12*, 2750−2760. The authors identify 50 putative Kme3-containing sequences that should bind the CBX6 chromodomain. Two have Val and two have Ala at the (−2) position relative to Kme3 and are therefore potential candidates for binding CBX6. The other 46 candidates have larger and/or polar groups at the (−2) position that could not be accommodated by CBX6 or any other CBX chromodomain, based on extensive knowledge of X-ray structures and solution phase binding data for many CBX−peptide complexes. In the absence of solution phase or cellular binding studies, we are skeptical that these are true CBX6 chromodomain binders in vitro or in vivo.

DOI: 10.1021/acsmedchemlett.5b00378
ACS Med. Chem. Lett. 2016, 7, 139−144

5.3 Additional Simulation Details

Initial attempts to produce meaningful simulations of compound 5 with CBX6 and 7 was done through molecular dynamics and set up by placing the anchor trimethyllysine residues within the aromatic cages while making sure the remainder of the ligand did not sterically clash with host (See Fig. 5.1). Simulations were performed in triplicate under the same conditions as presented in the publication. The trajectories produced significant changes of the host protein in this short timeframe, but little progression of the ligand towards the assumed pose under the hydrophobic clasp.

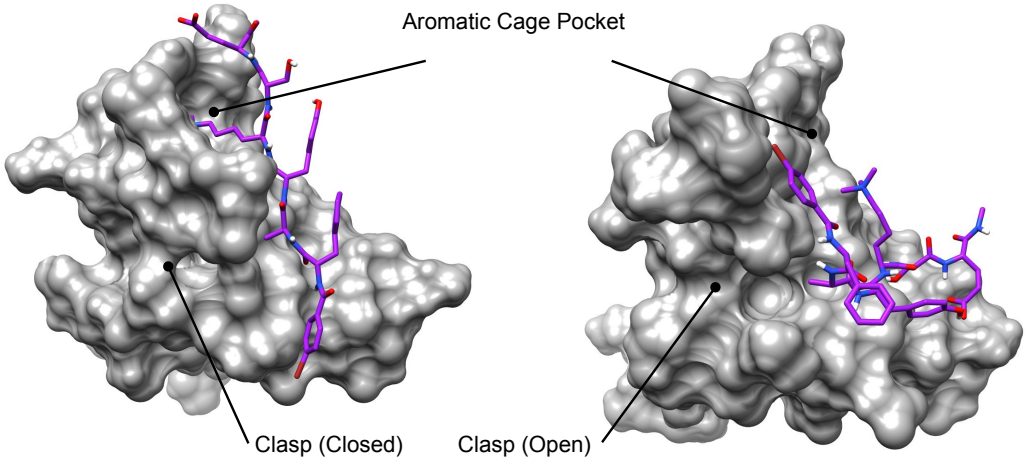


Figure 5.1: CBX6–Compound 5 Initial Simulation Attempt. Trimethllysine anchor residues were used to attach ligands to their respective hosts in attempts to start MD simulations and progress to bound structures. The simulation techniques showed no progress to convergence or the canonical binding poses seen with native ligands.

As a last resort, a template based docking approach was performed to set up the initial MD coordinates. To do this, crystal structures of the various CBX isoforms with their native peptide ligands were used as a template for compound 5's backbone coordinates. We were fortunate in that there were no major clashes when this was done. Molecular dynamics starting from these structures were then used for the data presented in the paper. The MD simulations for these systems presented an interesting structural feature of the CBX complexes which were not described yet at the time of results. Correlating features of the hydrophobic clasp and the effect on the aromatic cage structure were observed. This was the first piece of evidence we had structurally to discuss non-additive effects of binding that are suggested in ligand-based structural activity relationships.

Chapter 6

Publication: Accelerated Structural Prediction of Flexible Protein–Ligand Complexes: The SLICE Method

6.1 Preface

The following work describes the method development and validation for a combined molecular docking/dynamics technique that aims to capture host flexibility in the prediction of protein–ligand complexes. As previously described in introductory chapters, the use of combined molecular dynamics and molecular docking is not a novel concept. However, the combination of the two through an iterative process as demonstrated in this publication, yields surprising results. The accelerated structural prediction for the CBX system using the SLICE method was successful using accessible molecular dynamics timeframes under 50 ns of combined simulation time, whereas the comparison of single trajectory dynamics over a microsecond proved unsuccessful.

The method and concept were developed by James McFarlane and the writing of the manuscript was shared between Irina Paci, James McFarlane. Katherine Krause aided in the development of python scripts for automated job submissions used to generate data in the validation section. The current scripts are currently being developed into a usable python library containing all the necessary tools for file conversion and execution of the various docking and molecular dynamics components.

6.2 Publication

Full publication including links to supplementary information may be found at the following link:

https://pubs.acs.org/doi/10.1021/acs.jcim.9b00688

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING
Cite This: J. Chem. Inf. Model. XXXX, XXX, XXX−XXX
Article
pubs.acs.org/jcim

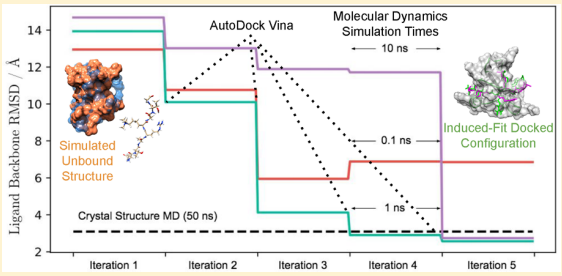
Accelerated Structural Prediction of Flexible Protein−Ligand Complexes: The SLICE Method

James M. B. McFarlane,* Katherine D. Krause, and Irina Paci*

Department of Chemistry, University of Victoria, Victoria, British Columbia V8W 3V6, Canada

Supporting Information

ABSTRACT: Using existing and academically available software, we present a new method for the structural prediction of binding events containing flexible protein targets. SLICE (Selective Ligand-Induced Conformational Ensemble) combines opportunistic stochastic jumps of ligand position with standard molecular dynamics to model the induced-fit binding of ligands starting with unbound host coordinates. To induce the structural adaptations of the complex at the binding site, conformational jumps in ligand position are selected in SLICE from structures generated by a docking software. Multiple binding trajectories from the docking set are followed using molecular dynamics for a set time to relax the host structure and generate new host poses. A new configurational jump is made on the set of newly generated host poses. The process is then repeated. The method was implemented with AutoDock Vina as the docking method, Vina scores as the selection criterion, and Amber code for molecular dynamics and applied to several test systems. A system consisting of Chromobox protein homologue 8 (CBX8) and its small peptide ligand, H3K9Me$_3$, for which the final (bound) configuration is known, is used for verifying SLICE in the present setup. The setup was also applied to several nonpeptide molecules on known difficult flexible targets exhibiting a large disparity between apo and holo host states. The SLICE simulations provide a promising approach to generate induced-fit configurations compared to existing long (microsecond) classical and accelerated dynamics approaches in all the test systems considered here. However, further optimization of SLICE parameters is required for replicating crystal structure coordinates for some systems. We discuss in the following pages the various SLICE parameters and how they can be optimized for the system at hand.



INTRODUCTION

The development of structural prediction tools for computer-assisted drug design has continued to grow since its inception decades ago. The computational power available today permits us to explore a chemical space orders of magnitude greater in size, with increased accuracy, compared to early pioneering docking techniques. Despite these advances, there is still much room for improvement for systems that exhibit conformational changes in the host upon binding. Such systems must sample a much larger potential energy landscape than is currently available through rigid protein docking simulations, even when followed by standard molecular dynamics simulations.

Descriptions of host-based conformational changes upon binding have classically fallen into two categories: the induced-fit (IF) model and the conformational selection model (CS). Both models subscribe to the theory of a protein-folding landscape occupying multiple states at the bottom of a well[1−3] but differ in their description of how proteins access those states. The CS model assumes that the bound pose is thermally accessible even in the absence of a ligand and that a population shift occurs once the ligand is bound. The IF model assumes that the presence of a ligand lowers the potential energy barrier between the unbound and bound states, thus inducing a conformational change in the host.

Despite a clear distinction in their definition, attempts to classify host−guest systems into either of these models using free energy methods are often inconclusive, as many systems possess properties of both IF and CS.[4,5] As an example, Bucher et al.[6] explore a mixed model of conformational change with the maltose binding protein. Their findings uncover a number of semiclosed states of the host binding around the ligand. From this, the authors suggest a binding event where IF and CS models are applicable at various stages of binding. From a structural prediction point of view, Bucher's study raises the question: Do we need to dock on a conformational ensemble or should we instead allow the ligand to deterministically change the binding site? Better yet, is it possible or beneficial to do both? Furthermore, given the structural complexity of the protein/ligand complex, how do we ensure efficient sampling of a configurational space riddled with local minima and barriers to structural adaptation?

The successful implementation of such structural prediction techniques would open up swaths of systems previously found intractable in docking applications,[7−9] particularly when significant host flexibility is encountered at the binding site.

Received: August 19, 2019
Published: November 6, 2019

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

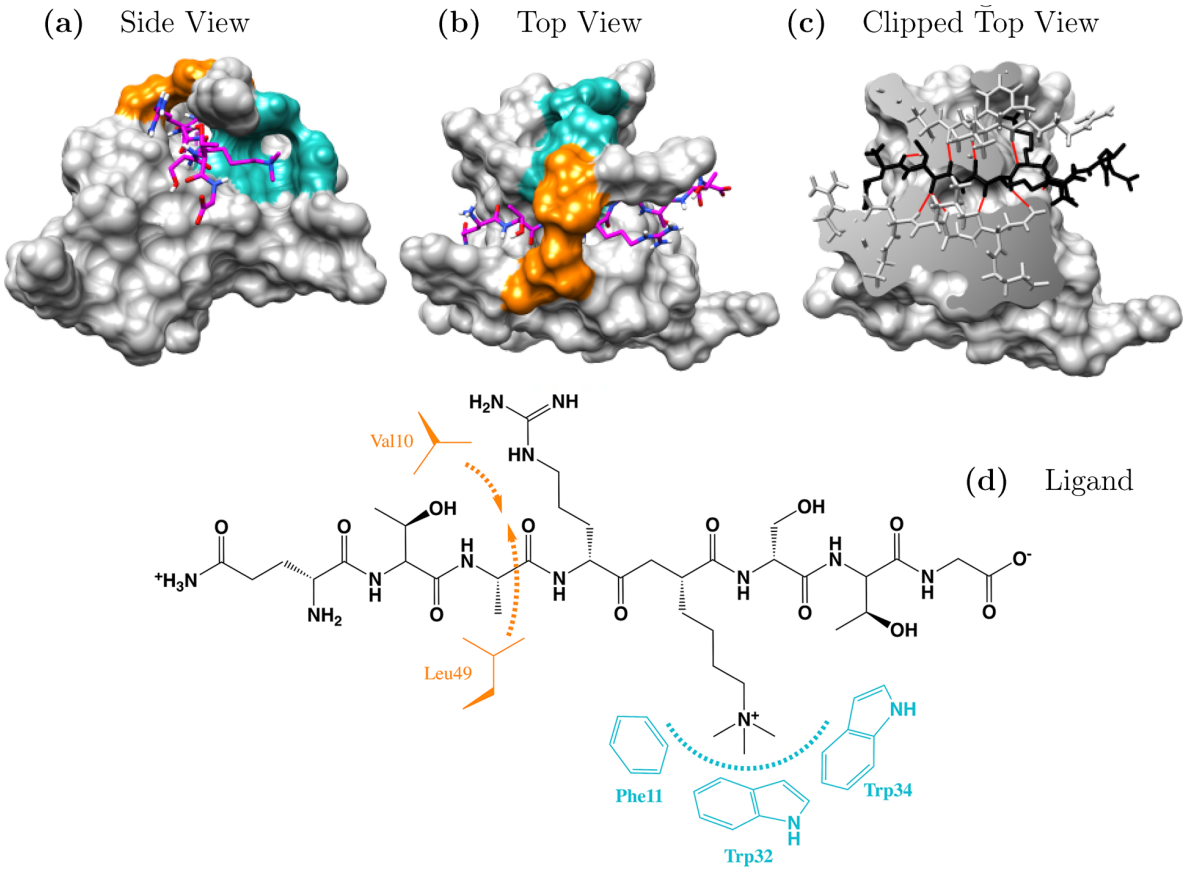


Figure 1. CBX8/H3K9Me$_3$ model system. (a) H3K9Me$_3$ trimethyllysine residue, shown atomistically excluding hydrogens, nested in the aromatic cage (teal shading at top right of figure). (b) Hydrophobic clasp formation (orange shaded foreground) bridged over ligand. (c) Hydrogen bonds (highlighted in red) of the intermolecular β sheet network between host and guest near N-terminus of ligand. (d) Atomistic structure of the H3K9Me$_3$ ligand with putative contacts.

This unmet need has drawn significant attention from the modeling community and much headway has been made into addressing the issue of host conformational change upon binding. At the present time, the majority of methods addressing host flexibility belong to single software approaches wherein host flexibility is coded into the structural sampling through a number of approaches. However, there are emerging examples of the application of multiple software approaches that utilize a stochastic search method followed by molecular dynamics simulation.

Single-software approaches that address protein flexibility have taken a number of routes to predict induced-fit configurations. One approach is to allow the molecular docking process to overcome steep conformational barriers using Lennard-Jones potentials with a softer repulsive wall (such as the soft-potential GOLD[10] method). Another approach is to sample a conformational ensemble of host structures through stochastic placement of side chains such as in AutoDock4.[11] The issue of side-chain flexibility can also be addressed by a pregenerated side-chain rotamer library such as in ICM.[12] Going beyond side-chain configurations, FlexE[13] docks on an ensemble of host configurations which includes both changes in the backbone structure and the side-chain rotamers. All of these methods include some degree of conformational sampling and steer away from the classic static host docking regime. For small-molecule docking and small deviations in host structures, these methods have shown numerous structural prediction successes. A single-software approach that uses HADDOCK[14] to combine IF and CS aspects of docking has been reported by Trellet et al.[15] In this method, both structural refinement and ensemble docking contribute to the increased docking accuracy for bound and unbound proteins with their peptide guests.

Using combinations of existing softwares—specifically, sequentially applying stochastic docking methods and deterministic structural minimizations—has shown promise as an IF/CS approach. For example, Schrodinger's induced-fit docking algorithm[16] uses the molecular docking program GLIDE followed by protein minimizations using the homology-based utility, Prime. An iterative algorithm using Rosetta Fold and MD simulations increased structural match success in protein folding applications.[17] Furthermore, accelerated sampling techniques for complex potential energy surfaces have been developed over the years, with varying degrees of generality: annealing, tempering and replica exchange,[18−20] bias or deformation potentials,[21−23] and genetic or machine learning algorithms,[24−27] as well as numerous combinations and derivative methodologies. Here, we develop a combination approach that fuses stochastic and deterministic features in an iterative methodology that effectively tackles induced-fit docking challenges, with a specific focus on

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

accelerating structural convergence of the protein/ligand complex.

The proposed method (SLICE) exploits the large stochastic potential energy surface movements of rigid-host molecular docking with the deterministic MD interaction between ligand and host. This is achieved through an iterative, multisoftware approach that takes advantage of widely available programs: the Amber software suite[28] and AutoDock Vina.[11] Overall, the method builds an accelerated dynamics pathway conducive to both IF and CS models. The conformational ensemble is sampled in stochastic docking steps, resembling a CS model, whereas any IF is tackled via MD simulation.

The SLICE methodology is applied to a validation system involving the binding of the native oligopeptide H3K9Me$_3$ onto an unbound CBX8 protein. The process requires structural changes along the protein backbone in multiple regions of the binding site to accommodate the peptide guest. The flexibility of the ligand and host, as well as the large conformational changes arising in the docking process, likely qualify this system as an induced-fit binding scenario. The resulting docked configuration is validated against the crystal structure of the CBX8/H3K9Me$_3$ complex.[29] Three other test sets are investigated, each presenting induced fit challenges to standard and accelerated dynamics simulations, and all three belong to the "Class III" binding site category in the Gunasekaran classification:[30] the maltose binding protein complex, the retinol binding protein complex, and the HIV-2 protease complex. These systems present specific challenges for the SLICE method and allow an exploration of its applicability and necessary adaptations.

As an added benefit beyond proof of concept, the CBX proteins (of which there are several isoforms) are currently a speculative clinical target due to their relationship to hepatocellular carcinoma (CBX4),[31] prostate cancer (CBX7),[32,33] and glioblastoma multiforme (CBX6).[34] Development of inhibitors selective for each isoform is currently underway, and several selective inhibitors have been developed.[35,36] However, the discerning structural features of the ligand that determine isoform selectivity have not been definitively identified or quantified. Successfully docking ligands to the flexible binding sites of these proteins would enable rational design of selective CBX inhibitors.

MODEL AND METHOD

Validation System. CBX (Chromobox protein homologue) proteins are epigenetic regulators whose role is to recognize trimethyllysine residues on histone proteins. The various isoforms share a common binding motif, involving an aromatic cage pocket, a hydrophobic clasp, and an extended intermolecular hydrogen bond network. The features are illustrated in Figure 1(a–c), respectively, for the model system CBX8 with its natural H3K9Me$_3$ ligand (shown in Figure 1(d)).

The aromatic cage pocket that traps the KMe$_3$ residue for CBX8 consists of a single phenylalanine (Phe11) and two tryptophan (Trp35,32) residues. The hydrophobic clasp is composed of a leucine (Leu49) and valine (Val10) residue on either side of the binding cleft that wrap over the ligand on binding. The extended hydrogen bond network is an intermolecular β sheet that runs between the ligand and host under the clasp. Together, the hydrophobic clasp and hydrogen bond network form the induced-fit configuration for all known CBX complexes and impart ligand selectivity on the host. Ligand selectivity has been studied *in silico* for a number of CBX complexes,[29,35−38] where ligand variations have been found to affect steric interactions under the clasp and the hydrogen bond network.

Due to the difficulties in modeling this flexible system, previous studies have used either direct starting coordinates from the crystal structures or template docking such as a superposition of ligand backbones to generate initial ligand positions. However, in the case of known binders without a crystal structure analogue, steric clashes arise in these docking methods and limit the ability to explore new chemical space. By applying the SLICE procedure to the unbound structure of the protein, we expect that the host adaptation to these steric clashes will be more easily sampled by the simulation.

To generate the unbound host structure, we isolated the CBX8 coordinates from PDB 3i91,[39] which contains both the CBX8 host and its H3K9Me$_3$ guest. The complex in 3i91 exists in a dimer system with contacts between monomers in a region outside of the binding site for the H3K9Me$_3$ guest. In order to ensure this point of contact does not alter the interaction between ligand and host, we ran independent MD simulations of the bound crystal structure monomer and guest and observed no significant changes in the binding motif. The exact coordinates taken were from chains A (CBX8) and C (H3K9Me$_3$). Furthermore, a RMSD comparison to a non-dimer CBX8 structure complexed with a peptidomimetic inhibitor (PDB: 5EQ0)[40] with the 3i91 PDB was done. Results show good structural overlap with both host and ligand backbone and upon visual inspection appear sufficiently aligned in the binding site around the ligand. The host structure thus obtained is solvated without ligand and undergoes a 50 ns MD simulation, during which large backbone displacements and side-chain movements produce a drastically different host configuration. These structural changes indicate a strong departure of the host configuration from its initial bound state, allowing the use of the 3i91 crystal structure of the CBX8/H3K9Me$_3$ complex as validation for the final structural prediction target.

Overarching Principles. Protein−ligand complexes evolve on high-dimensional, highly structured potential energy surfaces (PES). Classical challenges in computer simulations for complex PESs apply: high dimensionality, dependence on initial configuration, and restricted sampling by any given method due to the PES complexity. Stochastic methods such as Monte Carlo (MC) have been developed specifically to promote ergodic sampling of the PES, but pure MC methods are unable to process highly flexible systems. MD methods are useful in capturing and describing molecular flexibility for very large molecules, which is essential in describing the induced-fit adaptation of a protein host to a bound ligand. However, MD relies on the deterministic evolution of the system on minimum potential energy paths and is therefore prone to evolving in areas of the PES neighboring that of the initial configuration.

To tackle the PES complexity, stochastic moves in the SLICE methodology carry the system into distinct regions of its PES between the combined short (10 ns) MD bursts. The stochastic moves in SLICE arise from selecting docked poses that reset the position of the ligand onto previously ligand-adapted host configurations. MD bursts of these poses fulfill three roles: (i) probing the interaction environment in the local PES region for binding effectiveness, (ii) disruption of inhibitory surface features on the host, and (iii) generation of a

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

new conformational ensemble, preparing the host for a new generation of docking poses, and closing the iterative loop.

Short simulations (10 ns) are sufficient to model conformational changes in the active sites, in part due to the effectiveness of parallel simulations[41] and the conformational freedom explored in this simulation time.[42] For these reasons, we have chosen to do 10 separate 10 ns MD trajectories to generate host ensembles for docking, yielding a combined 100 ns simulation time for each iteration. We expect that the required simulation time will be highly dependent on the specifics of the protein/ligand pair. The nature and accessibility of local minima sampled, as well as the potential energy barriers involved in the binding site reorganization, will determine the required length of simulation bursts. While the ligand is allowed stochastic placement, the host protein is still subject to a deterministic walk along its potential energy surface—more complex backbone changes will likely require longer MD simulation runs. As a final caveat, the usage of AutoDock Vina scores for pose selection between rounds of MD could potentially steer the simulation into irrelevant local minima as docking scores can be poor indicators of binding potential energy. This is especially true for molecules used outside of the scope of the docking software such as the case with excessive torsional ligand freedom or situations where a major contribution from the host internal energy is ignored in the evaluation of the potential energy of the interaction between host and guest. We found Vina scores to be an easy to use yet satisfactory selection criterion for the systems considered here, but Rosenbluth factors and MM/PBSA scores are valid alternatives for more potential-constrained systems. Details about the selection criteria in the SLICE method are presented in the Ensemble Generation, Selection, and Iteration section.

Unbound Structure and Starting Poses. Initial stages of the SLICE method require a set of ligand−host poses to submit for the first set of MD simulations. To obtain these poses, we first simulate the target protein without a ligand in explicit water, generating a single unbound host pose. As shown in Figure 2, in the absence of a ligand, the hydrophobic clasp initially trends to an open geometry. The change occurs within 10 ns, suggesting a relatively high energetic penalty for the clasp to close over the ligand.

From the MD generated unbound host structure, AutoDock Vina is used to dock the H3K9Me$_3$ ligand. The 10 top scoring docked poses are selected for further simulation, concluding the preliminary sequence in Figure 3.

Ensemble Generation, Selection, and Iteration. The 10 initial poses of H3K9Me$_3$ docked on the unbound (open clasp) structure were submitted for 10 ns of MD in explicit solvent (Figure 4). Over the course of the simulations, 50 evenly spaced snapshots of each trajectory were selected. Host coordinates were isolated from the snapshots, generating a new ensemble of 500 overall host structures for docking in Autodock Vina. Selection of the top 10 scoring poses on each host structure yield 5000 docked poses overall.

The quick expansion of poses generated through iterations of the method (5000 per iteration) quickly becomes an intractable problem. To limit the number of simulations, we employed an adaptive sampling technique, wherein only the overall top 10 poses are submitted for the next round of molecular dynamics and generating the new host ensemble.

Docked poses are ranked in order of scores where "highest-scored" poses are lowest in potential energy and quantified by

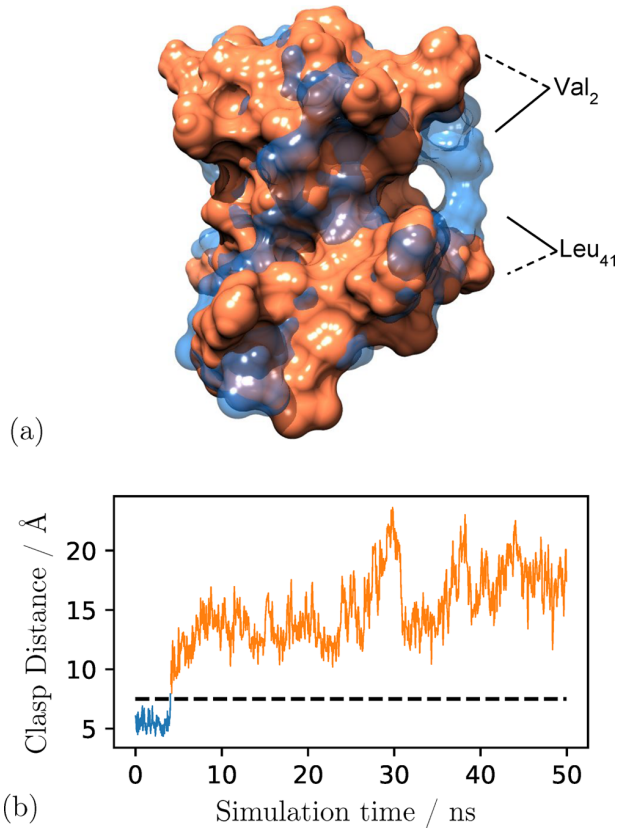


Figure 2. Simulated unbound structure. (a) Comparison of crystal structure PDB 3i91 (blue) with generated unbound structure (orange). In the absence of a ligand, the protein undergoes opening of the hydrophobic clasp and exposure of the binding groove, as clasp residues Val10 and Leu49 move apart. (b) Distance plot of 50 ns simulation time from crystal structure to structurally dissimilar unbound state.

ΔG (kcal/mol) energy by the Autodock Vina scoring function. In keeping with standard stochastic simulation methods, low-energy (high scoring) structures are pursued as likely routes to well-bound configurations, and higher-energy (lower scored) poses may open the way out of local PES minima. We found that for the CBX8/H3K9Me$_3$ complex discussed here, the choice of the top 10 poses at each docking event provided a broad enough selection of starting coordinates for MD simulations, while keeping the combined MD time to a manageable 100 ns per iteration.

The adaptive sampling selection of the 10 new poses concludes one iteration of the SLICE method in Figure 4. The loop can be iterated until the docked structures converge or there is no further improvement in docking scores. Criteria to terminate the iterations are discussed in the Validation and Figures of Merit section.

AutoDock Vina was chosen as the docking software for this method for a number of reasons. The automatic grid generation and command line executable made the software exceptionally easy to script into the method. It is also generally well known that of the available software, Vina is a robust and fast docking option for small organic molecules and has also been reported as a competitive method for docking small peptides given an appropriate exhaustiveness within the program usage.[43−45] However, since the convergence of ligand

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

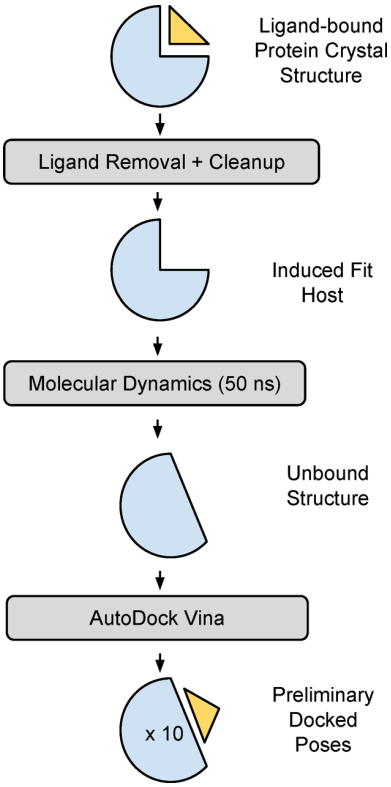


Figure 3. Preliminary docking steps. Clean up and equilibration of the host PDB structure were performed, providing an unbound host structure for redocking, in order to generate initial poses for host−ligand molecular dynamics.

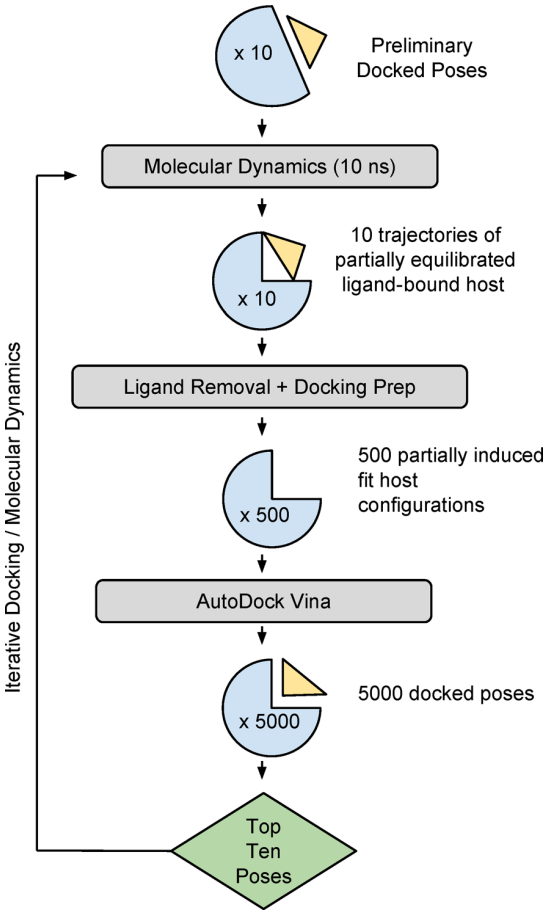


Figure 4. Iterative docking procedure. Initially docked poses are equilibrated via molecular dynamics. These trajectories are then parsed for new host structures for redocking. Top-scored poses are then re-equilibrated in MD for the next iteration.

poses rely on the docking portion of this method, we suspect the usual pitfalls of any docking method may come into play, namely, lack of host internal energy contributions, lack of torsional exploration, and lack of structural waters. Vina scores are in some cases insensitive to large conformational changes and should be replaced by MMPBSA energies or total potential in such cases. A Rosenbluth factor can be an effective criterion in the selection of poses that make it to the next iteration on more highly structured potential energy surfaces. For systems containing larger peptides spanning more torsional freedom, we believe a more peptide-specific docking software such as Rosetta's FlexPepDock or GalaxyPepDock would better serve this purpose.

COMPUTATIONAL DETAILS

File Preparation. To produce the unbound structure of CBX8, the PDB file was first cleaned by removing excess atoms (ligand, ions, water).

Molecular dynamics were done using the Amber16 suite with the ff14SB/GAFF force field.[46,47] All simulations coordinates were construction in the tleap building environment. Counterions were added to neutrality, and the CBX systems were solvated with approximately 8000 TIP3PBOX waters with a distance buffer of 14 Å. Simulations were minimized using gradient descent for 10,000 steps, heated to 300 K over 200 ps, and then run for 10 ns. Each simulation was run with a 2 fs time step with the SHAKE algorithm applied to bonds containing hydrogen atoms.[48] The systems were then run under constant pressure dynamics (ntp = 2) with a Langevin thermostat (ntt = 3) and set with periodic boundary conditions. Electrostatic cutoffs were set to 8 Å and coordinates were set to print 50 times over the course of the 10 ns simulations. Original input files for the molecular dynamics trajectories are available in the Supporting Information. As the purpose of these MD trajectories is to provide small and immediate deterministic responses in the host structure by the presence of the ligand, we opted to use these NPT equilibrations as our trajectory for which host poses are selected. For validation runs including both 250 ns classical and accelerated MD, NVT ensembles were run as production runs after equilibration under NPT conditions.

The ligand, H3K9Me$_3$, contained a trimethyllysine residue not available in the ff14SB force field. For this nonstandard residue, partial charges were obtained using Gaussian 09[49] at the HF-6-31G* level of theory and fitted using a restricted electrostatic potential using the *residuegen* utility. Remaining parameters were generated using the *parmchk* utility. Both *residuegen* and *parmchk* are available in AmberTools 17. UCSF Chimera's[50] AutoDock Vina utility was used to produce the PDBQT files required. The host and guest docking input files were then used for all subsequent docking.

Docking on each of the host poses generated from the MD trajectories was done using AutoDock Vina via command line and scripted to automatically parse MD trajectories, generate

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Docking Method Comparison

	Static Holo-Host Docking	Static Apo-Host Docking	cMD	aMD	SLICE
Host Backbone RMSD/Å	–	1.55	1.47	2.45	0.86
Ligand Backbone RMSD/Å	0.866	12.13	9.51	12.76	1.09
Ligand All-Atom RMSD/Å	2.30	12.17	8.81	13.01	3.056
AutoDock Vina Score /kcal mol^{-1}	−7.8	−5.2	−5.7	−5.8	−7.5

the docking box, and submit jobs to our cluster. The box was generated automatically by the inclusion of predetermined residues by the user as well as a 0.5 nm buffer in every axis. In the case of CBX8, our residue list included several residues in the known binding site that on average created a box that included more than 75% of the protein volume. A full list of box residues and descriptions of the required files for docking are provided in the Supporting Information. Docking exhaustiveness was set to 7, and grid generation was done through AutoDock Vina defaults.

Resources. Each 10-pose iteration required 12 h on 280 cores (10 nodes with 2 Intel Xeon E5-2680 v4 processors each). Between MD runs, each docking iteration was performed using a single 2.8 GHz Intel Core i5, averaging 12 h of processing time. Therefore, the total computing time for four iterations of the process from start to finish was approximately 4 days. We estimate this process could have been drastically accelerated by running AutoDock Vina on our cluster in parallel and automating the steps between iterations. Current versions of the automated software (developed after the production of the results reported below) achieve the same result in half the time.

VALIDATION AND FIGURES OF MERIT

Two considerations are paramount when evaluating trajectories in the application of the SLICE method: (i) devising a criterion for convergence to a predictive structure and (ii) the efficiency of sampling within the SLICE method relative to exhaustive molecular dynamics or other enhanced sampling algorithms. Above all, in the current validation work, the agreement of the predicted structure to the initial, experimentally bound crystal structure is an important consideration.

The solid-state crystal structure of the host does not fully describe the host's ensemble of thermally accessible solution-state conformations. Therefore, we solvated and ran MD simulations of the crystal structure, generating an equilibrated ensemble, to be used as a validation target. A further complication particular to systems like CBX8/H3K9Me_3 is that conformational freedom in regions unaffected by the binding process can have detrimental effects on RMSD convergence. As a result, the ligand RMSD of the target ensemble fluctuates around 3.5 Å. This value is greater than conventional measures for docking success,[51,52] a further argument for consideration of protein flexibility in docking calculations for CBX systems in general.

The SLICE iterations in the present validation set are thus considered to have converged when the ligand RMSD median data matches that from the validation ensemble, a value that aims to account for these effects. Qualitative features such as vicinity to putative contacts are also discussed below: (i) These contact points are maintained throughout the simulation of the solvated target complex, which underlines their stability. (ii) Optimization of the putative contacts is important in the rational design of peptidomimetic inhibitors targeting this complex.

In practical applications of the SLICE methodology, a reference structure is not available, requiring the definition of a convergence criterion intrinsic to the simulation. RMSD monitoring can serve this purpose,[53] but there are several other metrics that could be also be used. Two such metrics are considered below, and we discuss their merits as convergence criteria: energetic factors such as the potential energy of the system or a statistical structural description of the complex such as RMSD clustering.

The efficiency of this method boils down to a combination of resource cost and software usability. On the subject of resource efficiency, we have performed direct comparisons to molecular dynamics trajectories on the microsecond time scale as a way to show how SLICE overcomes the intractability of the system with MD alone. A user interface consolidating the python scripts that implement the SLICE method and tie together its associated software is under development and will be the subject of a future publication. Partial automation of the process at the present time has already significantly reduced the time required for input and analysis of results.

COMPARISON TO OTHER ENSEMBLE DOCKING METHODS

To evaluate the performance of SLICE relative to other ensemble docking methods, we considered both a classical MD trajectory and an accelerated trajectory for ensemble generation. Static host docking on both the crystal structure coordinates and the starting simulated unbound structure were also performed for comparison. In the static host docking experiment where the crystal host structure was used for docking, the ligand was successfully docked in the binding site. A backbone RMSD of 0.866 Å and successful putative contacts such as the trimethyllysine pocket interaction were observed. Static docking on the simulated host pose used as the initial target structure in the SLICE simulations (the unbound host structure) did not produce a ligand orientation with the putative contacts and acceptable ligand RMSD. Inspection of the structure suggests that the Arg9-Glu43 salt bridge blocks the ligand from forming the hydrogen bond network within the binding site.

Both ensemble methods were preceded by running a pre-equilibration of CBX8 with 50 ns of simulation time as described previously. The final frame of the simulation was then split into further simulations including a 250 ns classical MD simulation and 250 ns of accelerated MD simulation.[54] Full descriptions of these simulations as well as their analyses are available in the Supporting Information.

Both ensemble docking techniques yielded unsatisfactory results, comparable to the unbound static host docking, despite accessing 50 conformations over a 250 ns trajectory. Table 1 presents relevant RMSD and AutoDock Vina scores for the static and ensemble docking techniques, as well as the five SLICE iterations discussed below.

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

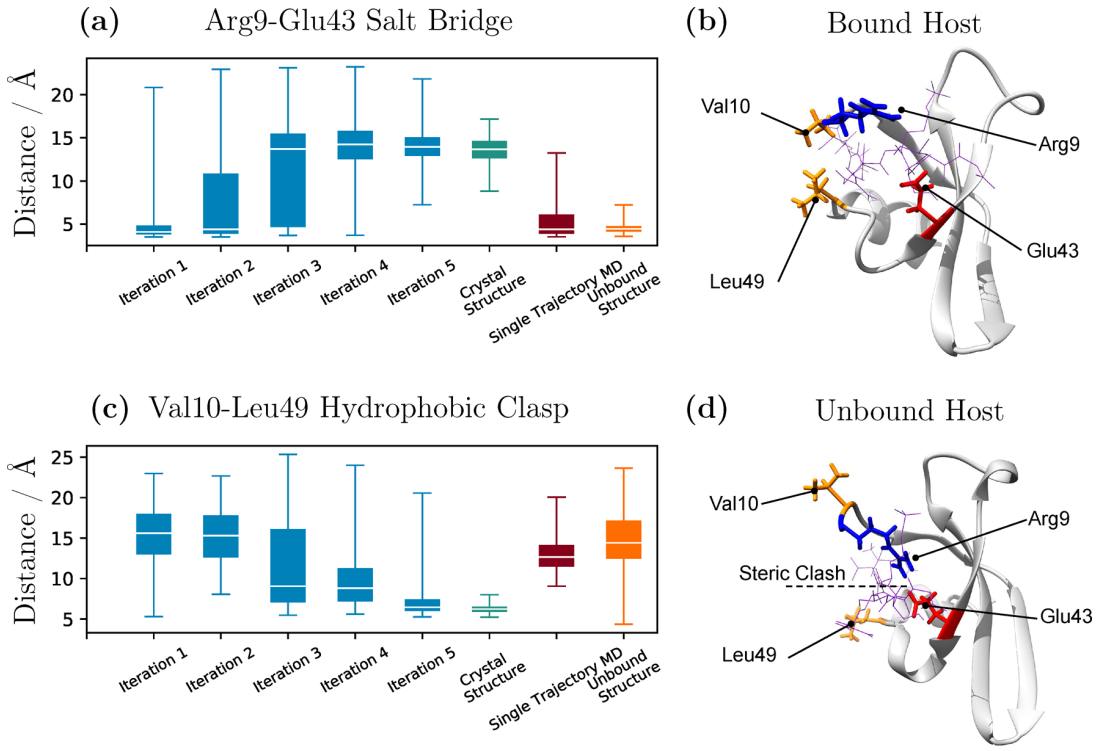


Figure 5. Structural figures of merit. (a) Arg9-Glu43 salt bridge distances. (b) Initial (crystal) structure of the CBX8 protein, with the residues relevant to the figure of merit plotted in (a) and (c). The ligand's crystal structure is superimposed as a purple wire frame. (c) Hydrophobic clasp distance, represented by the separation between the Val10-Leu49 terminal groups. (d) Simulated unbound structure of CBX8 with superimposed ligand crystal structure. In (a) and (c), median values are shown as white lines between boxes. Filled boxes include the second and third quartiles of the data values. Error bars indicate the full range of observed distances. The plots show the evolution of the figures of merit throughout the SLICE iterations (blue), with data for a simulation of the solvated crystal structure (teal), the unbound molecular dynamics simulation (orange), and direct docking molecular dynamics (red) included for comparison.

RESULTS AND DISCUSSION

Barriers in Docking and Induced-Fit Evolution. The simulation of the solvated unbound protein leads to the opening of the hydrophobic clasp (as shown by the orange surface representation in Figure 2(a)), accompanied by the deformation of the nearby aromatic cage and the creation of a salt bridge between the Arg9 and Glu43 residues adjacent to the clasp. This concerted behavior of the hydrophobic clasp and the aromatic cage has been previously reported experimentally by Stuckey and co-workers[36] and theoretically by ourselves.[35]

These changes in the host structure create steric barriers for the binding of H3K9Me$_3$. On the other hand, as putative contacts are reformed during the SLICE iterations, they provide measures of progress along the binding process. Particularly, the newly formed Arg9-Glu43 salt bridge at the base of the aromatic cage initially clashes with the ligand. As well, the clasp closure over H3K9Me$_3$ represents a large but necessary structural change that is only allowed to occur when the ligand is optimally aligned. Therefore, as a way to monitor the progression to a bound structure throughout the iterations of the SLICE method, we followed the breaking of the salt bridge as well as the eventual closure of the hydrophobic clasp by means of their representative distances (Figure 5).

Several notable points arise from the analysis of Figure 5: while average values of the figures of merit in the unbound protein structure are characteristic of an open clasp and a bound salt bridge, large variations around these figures are indicative of a highly dynamic system (Figure 5(b)). In particular, the clasp opens and closes periodically in the unbound structure simulation, spending most of its time in the open position, whereas the salt bridge is mostly bound and rarely dissociates. In contrast, a 50 ns single-trajectory MD simulation presents a partially closed clasp around an incorrectly bound ligand, sterically repelled from closing by the ligand. Other single trajectory simulations on the microsecond time scale were similarly unable to properly position the ligand and close the clasp. The local minimum that the system is trapped in also prohibits the opening of the salt bridge and further optimization of the docked pose in these systems.

The Iteration 1 data in the SLICE simulation mirrors the behavior of the single-trajectory MD. The SLICE methodology, however, bypasses such trapped states by removing and redocking the ligand. Further SLICE steps exhibit a gradual closing of the clasp and opening of the salt bridge to values analogous to those of the solvated crystal structure.

The recognition events responsible for the proposed binding process are reminiscent of both the IF and CS mechanisms: the salt bridge's stability suggests that the formation of the salt bridge has a higher potential energy barrier than that of the hydrophobic clasp. Dynamics studies done on similar arginine−glutamate interactions were found to contribute up to 7 kJ/mol of potential energy to the system,[55] a considerable energetic penalty stabilized by ligand binding (thus indicative of IF character). On the other hand, the docked ligand favors

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

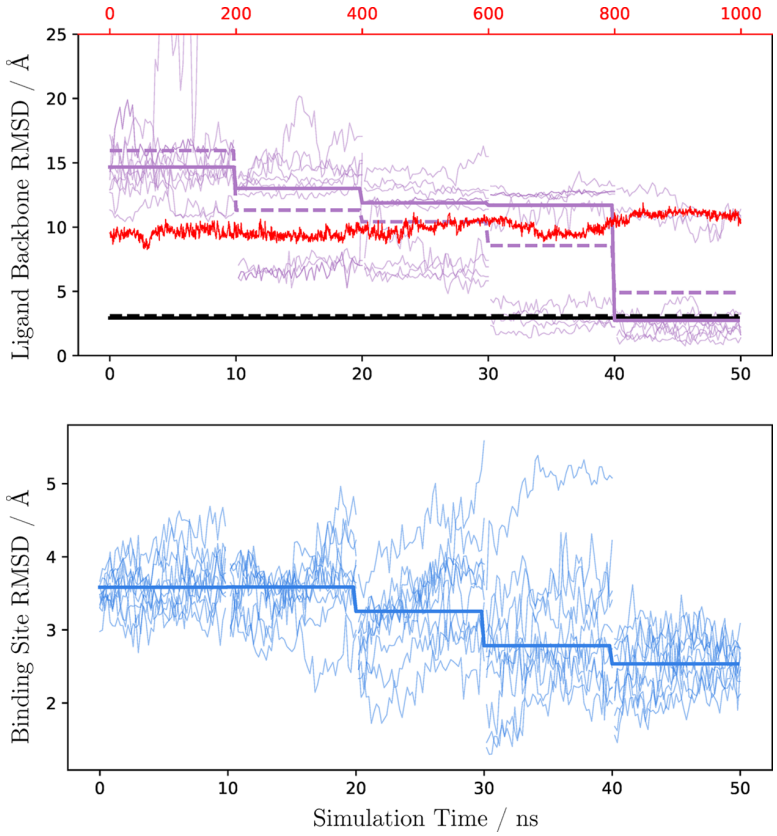


Figure 6. Ligand backbone RMSD evolution in SLICE iterations (top). RMSDs calculated for each 10 ns trajectory are presented for each iteration (thin purple lines). Mean and median RMSD values for each iteration (calculated over combined time and trajectory data) are presented in bold dashed and solid purple lines, respectively. For comparison, a single trajectory MD simulation from a top docked pose on the unbound structure is presented in red (scale on top). The direct MD run extends over 1 μs of simulation time. Bottom dashed and solid black lines show solvated crystal structure simulation mean and median data, respectively. Binding site RMSD (bottom) illustrates the decrease in all-atom RMSD of the binding site residues Arg9, Val10, Leu49, Glu43, Phe11, Typ32, and Trp34 as a combined metric for overall binding site similarity in reference to the crystal structure 3i91.

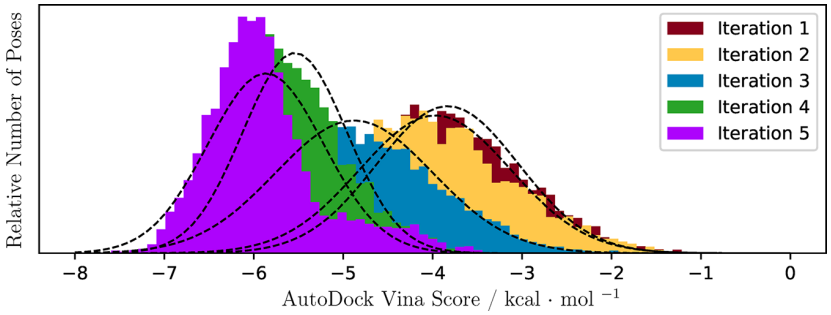


Figure 7. Distributions of docked scores over the SLICE iterations. Vina docking poses show a progression to narrower distributions of lower scores (indicating stronger binding). Top docked scores from the fourth round in green going into the final production of MD are comparable to docking scores performed on the crystal structure (approximately −7.6 kcal/mol). Final docked poses from the final MD production (purple) show further increase in binding affinity of the docked poses.

the closed form of the clasp among the geometries sampled by the unbound protein—this behavior has CS characteristics. However, the effect of the clasp position on the structure of the aromatic cage makes such an assignment more ambiguous.

The significantly larger variance of Iteration 1 data than that of single trajectory MD highlights another important aspect. The figure of merit averages for the two data sets are similar. However, each part of the SLICE methodology follows 10 distinct docking poses over 10 ns, in contrast to a single pose simulated over 50 ns in single trajectory MD. This results in the increased variability of the Iteration 1 data, with a crucial advantage: creating configurations that can be exploited to swiftly advance optimization.

RMSD Convergence to Validation Structure. Ligand backbone RMSDs (Figure 6) were also calculated for each of the replicate trajectories throughout the SLICE iterations. As observed for the structural figures of merit discussed in the Validation and Figures of Merit section, structural descriptors

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

maintain high variability throughout SLICE iterations. For example, in Iteration 5, several converged structures and several higher-energy structures are observed in the cluster. Several key features of this data highlight the advantages and challenges of the SLICE method.

From the first iteration, bound poses with sufficient potential energy for the ligand to leave the complex were produced, highlighting the statistical safety in starting the MD from multiple ligand poses. Furthermore, a top docked pose on the unbound structure was simulated for 1000 ns as a comparison (red line in Figure 6) and was found to show no significant trend toward convergence. Notably, the short simulation bursts do not individually converge toward smaller RMSD values—instead, large downward RMSD jumps are observed at the large stochastic moves in the SLICE method. This underlines the essential role of the docking steps as drivers to PES convergence.

The decrease in RMSD is coupled with a decrease in both mean and minimum Vina binding scores within the poses generated. For instance, top poses from each redocking yielded Vina scores of −5.9, −6.5, −7.0, −7.6, and −7.5 kcal/mol from the first to last iterations of SLICE, respectively. These changes in docking scores are illustrated in Figure 7. Meanwhile, the MD components between stochastic jumps serve two purposes: identifying which PES locations are likely to lead to successful minimization and facilitating the diversity of stochastic movements by offering lateral, though local, changes in the position along the PES, thus providing new sampling opportunities for the next stochastic move.

The challenge inherent in the SLICE methodology is also a consequence of this variability: in the final iteration, many replicates have converged to the crystal structure MD RMSD. However, a small number of replicates with high potential energy remain. Where a final reference structure is not available (as in the crystal structure of this test system), clustering of replicates with low potential energy within a narrow RMSD interval can be indicative of convergence. In the current case, we note the formation of an RMSD cluster along the crystal structure simulation in Figure 6 in Iteration 5. As an example of the structural convergence, a snapshot comparison between the crystal structure simulation and Iteration 5 is presented in Figure 8. The snapshots show good backbone alignment of the protein and ligand, despite variability in the ligand C- and N-termini due to the conformational freedom of these regions in an aqueous environment.

The ligand backbone RMSD is an effective tool to follow convergence of the SLICE simulation in the CBX8/H3K9Me$_3$ complex because the protein/peptide interaction region is overwhelmingly mapped along the peptide backbone. The peptide arginine side chain is outside of this interaction region, thus the high conformational variability of this residue in the aqueous environment (Figure 8). In many protein/ligand systems, ligand side chains are central to the interaction and would need to be accounted for in convergence descriptors. Methods other than ligand backbone RMSD such as interface RMSD and percentage of native contacts have been shown to be effective metrics in these cases[56,57] (using the contacts method available in CPPTRAJ in the AmberTools suite of programs, for example).

Simulation Time Effects. To gain an understanding of the effect of MD simulation time on the efficiency of the SLICE procedure, SLICE iterations on the CBX8/H3K9Me$_3$ complex were repeated using 1 and 0.1 ns simulation burst times. We maintained all other numerical details of the procedure: number of iteration steps, replica selection numbers, and number of host structures sampled. Figure 9 shows the evolution of median ligand backbone RMSD values over iterations in the three SLICE applications.

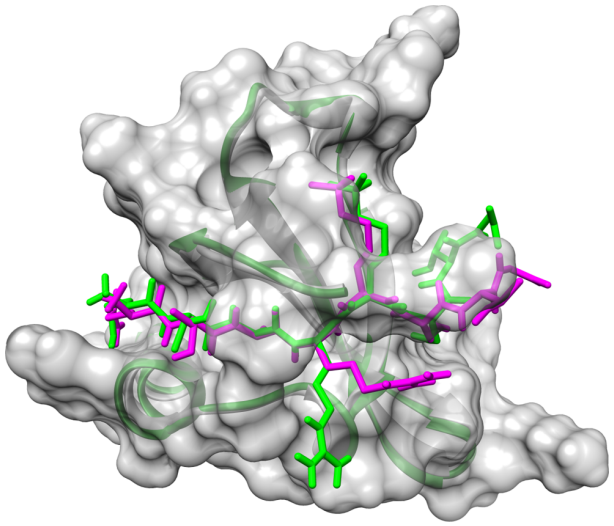


Figure 8. Final predicted bound pose. Comparison between a low-RMSD snapshot taken from the converged trajectories in Iteration 5 (green) and a crystal structure simulation starting point host (gray) and ligand (magenta). Side-chain deviations are most noticeable close to the more flexible C-terminus region of the ligand, whereas good ligand backbone overlap is observed throughout the binding motif.

As shown in Figure 9, the 1 ns simulation burst SLICE run converged to the crystal structure simulation in four iterations—as opposed to the five iterations required by the 10 ns SLICE run. Attempts to further decrease simulation time with 0.1 ns bursts led to clustered structures in a nonrelevant local minimum with high potential energy host configurations. This highlights the importance of the MD steps in allowing the adaptation of the host structure around the docked ligand.

As the selection criteria for following rounds of dynamics is solely based on the interaction energy between the surface of the protein and the ligand of previous docked poses, key energetic terms to fully describe the true potential energy surface are missing, specifically, the internal energy of the host as well as change in entropy of both host and ligand. Longer MD simulation times such as the 1 and 10 ns bursts partially negate this error by offering a conformational ensemble for redocking that was generated in part by these missing contributions. On the other hand, 0.1 ns simulation times are barely long enough for the host to locally adjust to the newly placed ligands and certainly not long enough to produce any significant backbone reorganization.

Inferences for Docking on CBX8. The comparison between bound and unbound CBX8 dynamics presented here raises several possible avenues toward designing higher affinity inhibitors targeting the aromatic cage pocket, as well as its downstream region. In particular, the Arg9-Glu43 salt bridge is a highly enthalpic barrier to binding in CBX8 and can be partially overcome by placing a hydrogen bond donor in the vicinity of Glu43. H3K9Me$_3$ addresses this through the presence of a serine in the region, but the flexible C-terminus of the ligand leads to instability in the Glu43-serine H-bond. A larger hydrogen donating residue, such as homoserine (serine

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

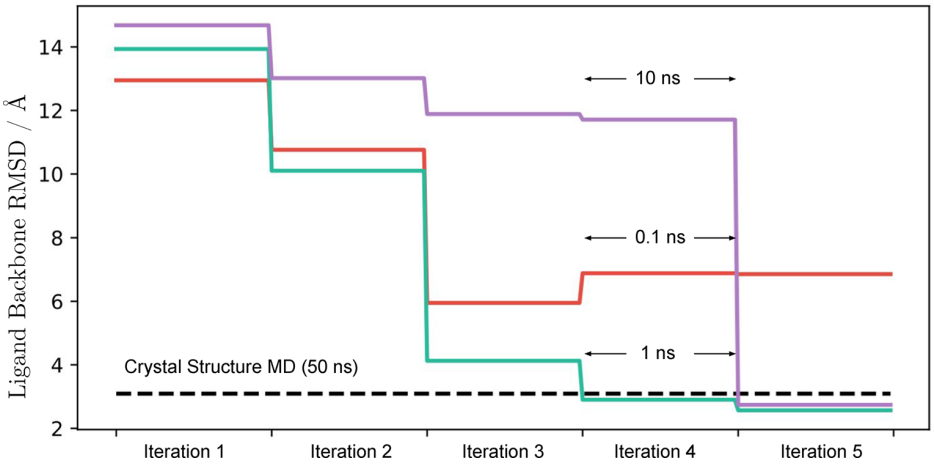


Figure 9. Simulation time burst comparisons. Median RMSD values over the length of the iteration and over all replicates is presented for SLICE series of 0.1 ns (red), 1 ns (teal), and 10 ns (purple) simulation bursts.

Table 2. Flexible Host Test Systems

Receptor Ligand (Holo-PDB, Apo-PDB)	Static Apo-Host Docking	SLICE Iterations/Vina Docking Score (Ligand RMSD/Å)				
		1	2	3	4	5
Maltose Binding Protein Maltose (1anf,1omp)	−6.0 (15.40)	−7.3 (31.83)	−8.0 (30.14)	−8.1 (28.84)	−8.0 (15.31)	−8.1 (27.40)
Retinol Binding Protein Retinol (1brp,1brq)	−7.9 (4.35)	−9.2 (4.48)	−9.3 (4.94)	−9.7 (4.51)	−9.7 (4.39)	−9.8 (4.46)
HIV-2 Protease CGP 53820 (1hii A,1hsi)	−6.7 (8.13)	−7.6 (8.39)	−8.8 (9.71)	−9.2 (9.49)	−9.0 (9.91)	−9.2 (3.04)

with an additional methylene unit), could better allow this interaction to achieve a consistent optimal distance by further removing the flexible C-terminus from the H-bonding region.

On the other side of the binding site, stronger binding via a continued hydrogen bond network could be provided by the substitution of the H3K9Me$_3$ N-terminus glutamine. Substitutions of this residue should include hydrogen bond donors targeting Ala13 and Glu14. The latter are in the vicinity of the terminal glutamine already but appear to be nonspecifically interacting in the region and could be directly targeted to augment binding.

Other Flexible Test Systems. To further test the applicability range of the SLICE method, a few challenging test systems were chosen based on protein flexibility and large differences in binding site RMSD between known apo and holo states. The systems are described by Gunasekaran and Nussinov as "Class III" binding sites, the highest degree of difficulty in their survey of flexible binding sites, and include apo- to holo-binding site RMSD differences of 2.0 Å or larger.[30] The systems were treated identically to that of CBX8/H3K9Me$_3$, using 1 ns MD bursts used instead of 10 ns. The decision to use 1 ns was made after observing comparable convergence using 1 ns with the CBX8 test system as shown in Figure 9.

Among all test systems, the trend of increasing binding scores was observed. The HIV-2 Protease system achieved full structural convergence of the ligand RMSD compared to the holo state within the parameter constraints of this SLICE application, as shown in Table 2.

The maltose binding protein (MBP) and retinol binding protein (RBP) systems experienced trapping in local minima within the SLICE parameter set considered here. Trapping is a known issue in simulations of maltose binding protein due to the multiple potential energy basins between holo and apo states.[58] In SLICE, the short 1 ns simulation bursts led to the convergence of available host structures to an incorrect pose in both systems, though still locally optimized for ligand interactions. Possible reasons for this are as follows: (i) MD simulation times were too short to escape local minima. (ii) There were an insufficient number of iterations. (iii) The funnelling of top-docked poses was too narrow, leading to local trapping.

By extending the number of SLICE iterations in the MBP system to 10, the formation of holo-like MBP structures was observed. However, the system was still trapped in a semiclosed state. By purposefully including a subset of lower scored poses between iterations, fully holo-like states of MBP as well as low ligand RMSD structures were simulated by iteration 10. Similar convergence behavior was observed in initial tests for the RBP system. This indicates that the selection criterion used in the SLICE implementation described here (namely, the Vina scores) may be insufficient for systems evolving on potential energy surfaces that are highly structured in the binding region. Optimization of the SLICE iteration parameters addressing the issues listed above for MBP, RBP, and a larger set of Class III systems is underway. A set of scoring functions that consider changes in the internal energy of the host are also under development, and these results will be communicated in a new manuscript addressing refinement of SLICE options.

The parameters for the SLICE methodology selected in this work are applicable to a subset of small molecule and small peptide ligands that have been identified as challenging flexible-host docking systems. Additional options can be added to the SLICE methodology, as required, to broaden it is applicability to a range of known challenging systems. Here, we initiated an exploration of improvements related MD simulation times, number of SLICE iterations, and inclusion of lower-scored poses into successive SLICE rounds. We also explored alternative sampling techniques for the stochastic

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

moves and alternative scoring options for the generated poses—these options are showing great promise in our initial investigations for both the MBP and the RBP problems discussed above and will make the topic of further communications on the SLICE method. The SLICE methodology, in its fundamental concept of successive stochastic and dynamic moves to investigate ligand and host flexibility and its impact on binding structures, forms a powerful backbone that can be enhanced and adapted by addition of a range of options involving alternate docking software and MD suites beyond the options discussed above.

CONCLUSIONS

We have presented an accelerated dynamics method capable of matching crystal structure simulation data for a peptide bound to a flexible protein host. The combination of molecular dynamics with a ligand position exchange performed by molecular docking was shown to effectively produce a bound structure when done iteratively. This method was shown to outperform microsecond length molecular dynamics or molecular docking alone. In doing so, we also uncovered a mixture of important structure features of the CBX8 binding motif that can be further exploited for the development of high affinity binders. The method was also tested on other flexible host proteins with nonpeptide guests. The results of these test systems varied in success, spurring efforts toward further development in scoring function selection, as well as the need for optimization of SLICE parameters for the specific system under consideration. These efforts are currently being undertaken in our group and will be the subject of a forthcoming publication.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00688.

Input files for molecular dynamics, description of the box generation for AutoDock Vina, and list of the required files for the SLICE method. Example docking input files from the CBX8 system test as well as cpptraj trajectory parsing files. Figures depicting the structural changes of the CBX8 system during both 250 ns of accelerated and classical MD. (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jmbm87@uvic.ca.
*E-mail: ipaci@uvic.ca.

ORCID

James M. B. McFarlane: 0000-0002-2505-1408

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Funding was provided by the National Science and Engineering Research Council of Canada and the University of Victoria. This research was performed in part using the Compute Canada, WestGrid, and CAMTEC computing resources. Special thanks to Belaid Moa for computing support and Fraser Hof for discussions on the model system.

REFERENCES

(1) Onuchic, J. N.; Wolynes, P. G. Theory of Protein Folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70−75.
(2) Frauenfelder, H.; Sligar, S.; Wolynes, P. The Energy Landscapes and Motions of Proteins. *Science* **1991**, *254*, 1598−1603.
(3) Changeux, J.-P.; Edelstein, S. Conformational Selection or Induced-fit? 50 Years of Debate Resolved. *F1000 Biol. Rep.* **2011**, *3*, 19−34.
(4) Morando, M. A.; Saladino, G.; D'Amelio, N.; Pucheta-Martinez, E.; Lovera, S.; Lelli, M.; López-Méndez, B.; Marenchino, M.; Campos-Olivas, R.; Gervasio, F. L. Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase. *Sci. Rep.* **2016**, *6*, 24439−24448.
(5) Liu, W.; Huang, B.; Kuang, Y.; Liu, G. Molecular Dynamics Simulations Elucidate Conformational Selection and Induced Fit Mechanisms in the Binding of PD-1 and PD-L1. *Mol. BioSyst.* **2017**, *13*, 892−900.
(6) Bucher, D.; Grant, B. J.; McCammon, J. A. Induced Fit or Conformational Selection? The Role of the Semi-closed State in the Maltose Binding Protein. *Biochemistry* **2011**, *50*, 10530−10539.
(7) Antunes, D. A.; Devaurs, D.; Kavraki, L. E. Understanding the Challenges of Protein Flexibility in Drug Design. *Expert Opin. Drug Discovery* **2015**, *10*, 1301−1313.
(8) Baumgartner, M. P.; Evans, D. A. Lessons Learned in Induced Fit Docking and Metadynamics in the Drug Design Data Resource Grand Challenge 2. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 45−58.
(9) Kirys, T.; Ruvinsky, A. M.; Singla, D.; Tuzikov, A. V.; Kundrotas, P. J.; Vakser, I. A. Simulated Unbound Structures for Benchmarking of Protein Docking in the Dockground Resource. *BMC Bioinf.* **2015**, *16*, 243−249.
(10) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609−623.
(11) Cosconati, S.; Forli, S.; Perryman, A. L.; Harris, R.; Goodsell, D. S.; Olson, A. J. Virtual screening with AutoDock: theory and practice. *Expert Opin. Drug Discovery* **2010**, *5*, 597−607.
(12) Fernández-Recio, J.; Totrov, M.; Abagyan, R. ICM-DISCO Docking by Global Energy Optimization with Fully Flexible Side-Chains. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 113−117.
(13) Perez, A.; Yang, Z.; Bahar, I.; Dill, K. A.; MacCallum, J. L. FlexE: Using Elastic Network Models to Compare Models of Protein Structure. *J. Chem. Theory Comput.* **2012**, *8*, 3985−3991.
(14) van Zundert, G.; Rodrigues, J.; Trellet, M.; Schmitz, C.; Kastritis, P.; Karaca, E.; Melquiond, A.; van Dijk, M.; de Vries, S.; Bonvin, A. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **2016**, *428*, 720−725.
(15) Trellet, M.; Melquiond, A. S. J.; Bonvin, A. M. J. J. A Unified Conformational Selection and Induced Fit Approach to Protein-Peptide Docking. *PLoS One* **2013**, *8*, No. e58769.
(16) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *J. Med. Chem.* **2006**, *49*, 534−553.
(17) Lindert, S.; Meiler, J.; McCammon, J. A. Iterative Molecular Dynamics—Rosetta Protein Structure Refinement Protocol to Improve Model Quality. *J. Chem. Theory Comput.* **2013**, *9*, 3843−3847.
(18) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers. *Biopolymers* **2001**, *60*, 96−123.
(19) Hukushima, K.; Nemoto, K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.
(20) Frenkel, D.; Smit, B. *Understanding Molecular Simulations: From Algorithms to Applications*; Academic Press, ISBN:0122673514, 9780122673511, 2002.
(21) de Oliveira, C. A. F.; Hamelberg, D.; McCammon, J. A. Coupling Accelerated Molecular Dynamics Methods with Thermody-

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

namic Integration Simulations. *J. Chem. Theory Comput.* **2008**, *4*, 1516−1525.
(22) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated Molecular Dynamics: A Promising and Efficient Simulation Method for Biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919−11929.
(23) Piela, L.; Kostrowicki, J.; Scheraga, H. A. On the Multiple-Minima Problem in the Conformational Analysis of Molecules: Deformation of the Potential Energy Hypersurface by the Diffusion Equation Method. *J. Phys. Chem.* **1989**, *93*, 3339−3346.
(24) Lee, J.; Scheraga, H. A.; Rackovsky, S. New Optimization Method for Conformational Energy Calculations on Polypeptides: Conformational Space Annealing. *J. Comput. Chem.* **1997**, *18*, 1222−1232.
(25) Bernardi, R. C.; Melo, M. C.; Schulten, K. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850*, 872−877.
(26) Matsunaga, Y.; Sugita, Y. Linking Time-Series of Single-Molecule Experiments with Molecular Dynamics Simulations by Machine Learning. *eLife* **2018**, *7*, No. e32668.
(27) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, No. e1701816.
(28) Case, D. A.; Cerutti, D. S.; Cheatham, T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Greene, D.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. *Amber 2016*; University of California, San Francisco, 2016.
(29) Kaustov, L.; Ouyang, H.; Amaya, M.; Lemak, A.; Nady, N.; Duan, S.; Wasney, G. A.; Li, Z.; Vedadi, M.; Schapira, M.; Min, J.; Arrowsmith, C. H. Recognition and Specificity Determinants of the Human Cbx Chromodomains. *J. Biol. Chem.* **2011**, *286*, 521−529.
(30) Gunasekaran, K.; Nussinov, R. How Different are Structurally Flexible and Rigid Binding Sites? Sequence and Structural Features Discriminating Proteins that Do and Do not Undergo Conformational Change upon Ligand Binding. *J. Mol. Biol.* **2007**, *365*, 257−273.
(31) Wang, B.; Tang, J.; Liao, D.; Wang, G.; Zhang, M.; Sang, Y.; Cao, J.; Wu, Y.; Zhang, R.; Li, S.; Ding, W.; Zhang, G.; Kang, T. Chromobox Homolog 4 Is Correlated with Prognosis and Tumor Cell Growth in Hepatocellular Carcinoma. *Ann. Surg. Oncol.* **2013**, *20*, 684−692.
(32) Bernard, D.; Martinez-Leal, J. F.; Rizzo, S.; Martinez, D.; Hudson, D.; Visakorpi, T.; Peters, G.; Carnero, A.; Beach, D.; Gil, J. CBX7 Controls the Growth of Normal and Tumor-Derived Prostate Cells by Repressing the Ink4a/Arf Locus. *Oncogene* **2005**, *24*, 5543−5551.
(33) Yap, K. L.; Li, S.; Muñoz-Cabello, A. M.; Raguz, S.; Zeng, L.; Mujtaba, S.; Gil, J.; Walsh, M. J.; Zhou, M.-M. Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. *Mol. Cell* **2010**, *38*, 662−674.
(34) Li, G.; Warden, C.; Zou, Z.; Neman, J.; Krueger, J. S.; Jain, A.; Jandial, R.; Chen, M. Altered Expression of Polycomb Group Genes in Glioblastoma Multiforme. *PLoS One* **2013**, *8*, No. e80970.
(35) Milosevich, N.; Gignac, M. C.; McFarlane, J.; Simhadri, C.; Horvath, S.; Daze, K. D.; Croft, C. S.; Dheri, A.; Quon, T. T. H.; Douglas, S. F.; Wulff, J. E.; Paci, I.; Hof, F. Selective Inhibition of CBX6: A Methyllysine Reader Protein in the Polycomb Family. *ACS Med. Chem. Lett.* **2016**, *7*, 139−144.
(36) Stuckey, J. I.; Dickson, B. M.; Cheng, N.; Liu, Y.; Norris, J. L.; Cholensky, S. H.; Tempel, W.; Qin, S.; Huber, K. G.; Sagum, C.; Black, K.; Li, F.; Huang, X.-P.; Roth, B. L.; Baughman, B. M.; Senisterra, G.; Pattenden, S. G.; Vedadi, M.; Brown, P. J.; Bedford, M. T.; Min, J.; Arrowsmith, C. H.; James, L. I.; Frye, S. V. A Cellular Chemical Probe Targeting the Chromodomains of Polycomb Repressive Complex 1. *Nat. Chem. Biol.* **2016**, *12*, 180−187.
(37) Liu, H.; Li, Z.; Li, L. The Molecular Selectivity of UNC3866 Inhibitor for Polycomb CBX7 Protein from Molecular Dynamics Simulation. *Comput. Biol. Chem.* **2018**, *74*, 339−346.
(38) Ren, C.; Morohashi, K.; Plotnikov, A. N.; Jakoncic, J.; Smith, S. G.; Li, J.; Zeng, L.; Rodriguez, Y.; Stojanoff, V.; Walsh, M.; Zhou, M.-M. Small-Molecule Modulators of Methyl-Lysine Binding for the CBX7 Chromodomain. *Chem. Biol.* **2015**, *22*, 161−168.
(39) Amaya, M.; Ravichandran, M.; Loppnau, P.; Kozieradzki, I.; Edwards, A.; Arrowsmith, C.; Weigelt, J.; Bountra, C.; Bochkarev, A.; Min, J.; Ouyang, H. Crystal Structure of Human Chromobox Homolog 8 (CBX8) with H3K9 peptide. *Protein Data Bank*, 2009. DOI: 10.2210/pdb3i91/pdb.
(40) Liu, Y.; Tempel, W.; Walker, J.; Stuckey, J.; Dickson, B.; James, L.; Frye, S.; Bountra, C.; Arrowsmith, C.; Edwards, A.; Min, J. Crystal Structure of Chromodomain of CBX8 in Complex with Inhibitor UNC3866. *Protein Data Bank*, 2015; DOI: 10.2210/pdb5eq0/pdb.
(41) Chodera, J. D.; Noé, F. Markov State Models of Biomolecular Conformational Dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.
(42) Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. J. Principles of Flexible Protein-Protein Docking. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 271−289.
(43) Rentzsch, R.; Renard, B. Y. Docking Small Peptides Remains a Great Challenge: An Assessment Using AutoDock Vina. *Briefings Bioinf.* **2015**, *16*, 1045−1056.
(44) Hauser, A. S.; Windshügel, B. LEADS-PEP: A Benchmark Data Set for Assessment of Peptide Docking Performance. *J. Chem. Inf. Model.* **2016**, *56*, 188−200.
(45) Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein−Peptide Docking: Opportunities and Challenges. *Drug Discovery Today* **2018**, *23*, 1530−1537.
(46) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(47) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(48) Lambrakos, S.; Boris, J.; Oran, E.; Chandrasekhar, I.; Nagumo, M. A Modified Shake Algorithm for Maintaining Rigid Bonds in Molecular Dynamics Simulations of Large Molecules. *J. Comput. Phys.* **1989**, *85*, 473−486.
(49) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, P. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision E.08 ; Gaussian, Inc.: Wallingford, CT, 2009.
(50) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera: A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(51) Jain, A. N. Bias, Reporting, and Sharing: Computational Evaluations of Docking Methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 201−212.
(52) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing Protein-Ligand Docking Programs is Difficult. *Proteins: Struct., Funct., Genet.* **2005**, *60*, 325−332.

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(53) Knapp, B.; Frantal, S.; Cibena, M.; Schreiner, W.; Bauer, P. Is an Intuitive Convergence Definition of Molecular Dynamics Simulations Solely Based on the Root Mean Square Deviation Possible? *J. Comput. Biol.* **2011**, *18*, 997−1005.
(54) Pierce, L. C.; Salomon-Ferrer, R.; de Oliveira, C. A. F.; McCammon, J. A.; Walker, R. C. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 2997−3002.
(55) White, A. D.; Keefe, A. J.; Ella-Menye, J.-R.; Nowinski, A. K.; Shao, Q.; Pfaendtner, J.; Jiang, S. Free Energy of Solvated Salt Bridges: A Simulation and Experimental Study. *J. Phys. Chem. B* **2013**, *117*, 7254−7259.
(56) Lensink, M. F.; Wodak, S. J. Docking and Scoring Protein Interactions: CAPRI 2009. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 3073−3084.
(57) Lensink, M. F.; Velankar, S.; Wodak, S. J. Modeling Protein-Protein and Protein-Peptide Complexes: CAPRI 6th edition. *Proteins: Struct., Funct., Genet.* **2017**, *85*, 359−377.
(58) Wang, Y.; Tang, C.; Wang, E.; Wang, J. Exploration of Multi-State Conformational Dynamics and Underlying Global Functional Landscape of Maltose Binding Protein. *PLoS Comput. Biol.* **2012**, *8*, No. e1002471.

DOI: 10.1021/acs.jcim.9b00688
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

6.3 Further Optimization Directions

The publication presented above laid the groundwork for the validation of the SLICE method. From the extended test systems, it's clear that more validation must be done to tune the method or at the very least benchmark it against a variety of system types. The RMSD differences between apo- and holo-host structures is indicative of large changes, but much of the challenge remains in the kinetics of this change as limited by the deterministic (MD) portion of the method. With the extended test systems—with Maltose Binding Protein in particular—local minima effects were observed despite initial convergence. To mitigate these local mimina effects, a stochastic selection in form of a Rosenbluth selection of the docked poses between iterations were performed with the selection factor weighted as shown in Figure 6.1.

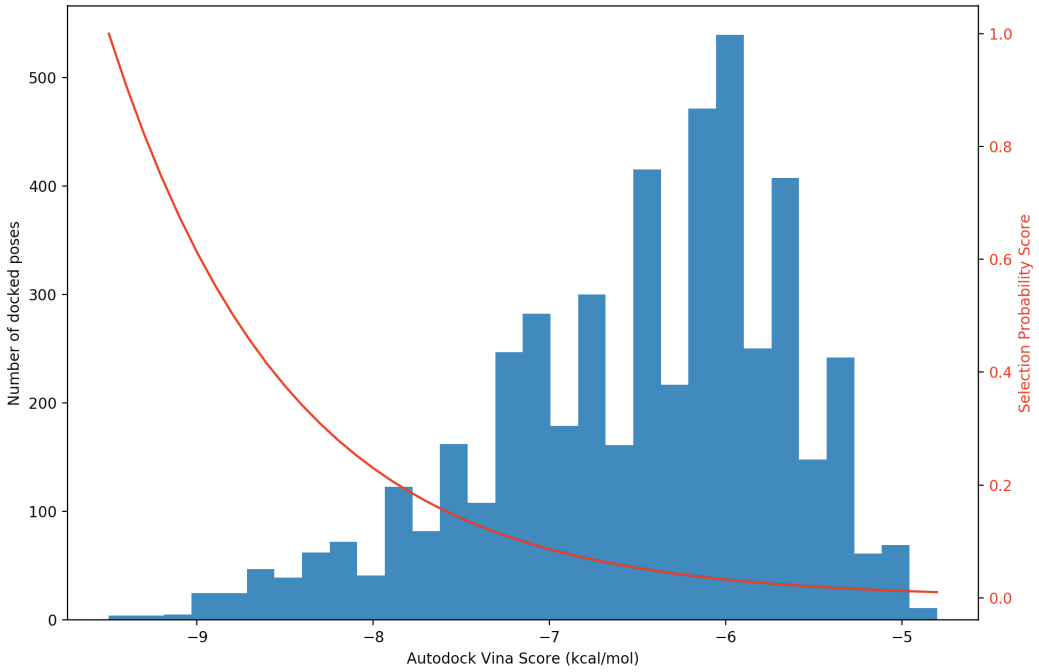


Figure 6.1: Rosenbluth Selection Scheme and Selection Probability. The distribution of docked poses for one iteration of the Maltose Binding Protein with overlaid Boltzmann-like factor (red line) for Rosenbluth selection.

Without the Rosenbluth selection factor applied, the data in Figure 6.2 shows a local minimum effect at the fifth iteration of the protocol whereas the effect of the Rosenbluth selection in Figure 6.3 appeared to slow down convergence of the host structure but at the same time continued to frame the data such that the crystal structure reference host structure was similar in the final iteration.

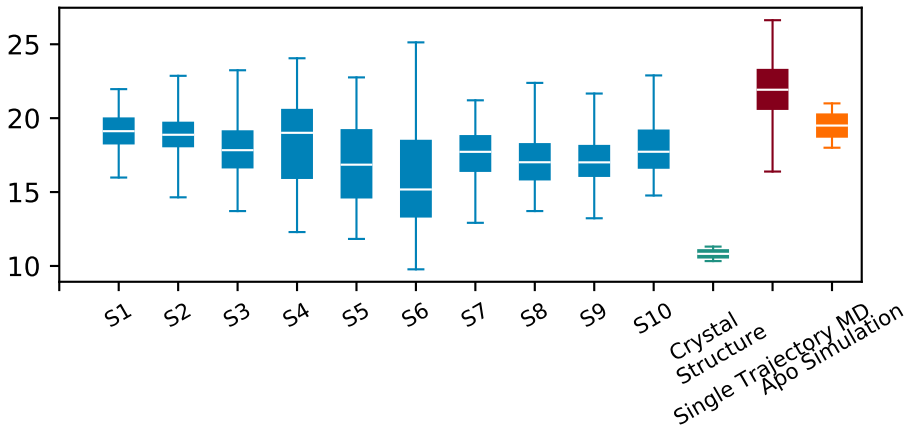


Figure 6.2: Maltose Binding Protein Inter-domain Convergence. The SLICE iterations of this system show vulnerabilities to local minimum effects near the fifth iteration. Data onward from this point show the ligand is wedged between a pivoting section of the two domains, preventing the closure to the correct host pose. Distance on the y-axis is represented in Å.

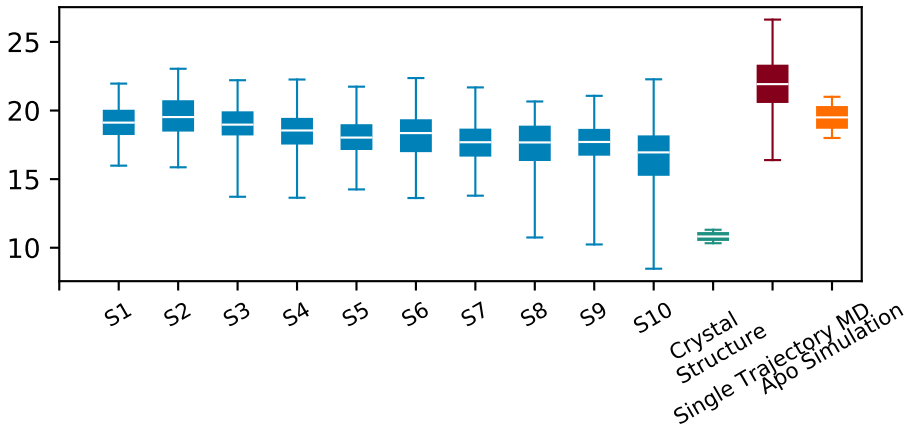


Figure 6.3: Applied Rosenbluth Selection Scheme on Maltose Binding Protein. Application of the selection scheme appears to slow down the convergence relative to the skimming technique in the previous example. However, the data continues to capture the correct interdomain distance at the end of the iterations. Distance on the y-axis is represented in Å.

Chapter 7

Publication: Pan-specific and partially selective dye-labeled peptidic inhibitors of the polycomb paralog proteins

7.1 Preface

This publication explores the selective features between the CBX7 and CBX8 isoforms. Peptidomimetic designs were synthesized and tested experimentally by Milosevich and coworkers. All computational work and analysis was performed by James McFarlane. Computational work for this publication included the parameterization of non-standard residues for the peptide ligands using the AMBER16 molecular dynamics suite. Structural prediction prior to production trajectories utilized the SLICE method from previous chapters and was also performed by James McFarlane. This work also includes the presented trajectory analysis and free energy decompositions. For full author contributions, please refer to Page 9 of the following publication.

7.2 Publication

Full publication including links to supplementary information may be found at the following link:

https://doi.org/10.1016/j.bmc.2019.115176

Bioorganic & Medicinal Chemistry xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Pan-specific and partially selective dye-labeled peptidic inhibitors of the polycomb paralog proteins

Natalia Milosevich[a], James McFarlane[a], Michael C. Gignac[a], Janessa Li[a], Tyler M. Brown[a], Chelsea R. Wilson[a], Lindsey Devorkin[b], Caitlin S. Croft[a], Rebecca Hof[a], Irina Paci[a], Julian J. Lum[b], Fraser Hof[a,*]

[a] Department of Chemistry, University of Victoria, Victoria, BC V8W 3V6, Canada
[b] BC Cancer – Trev and Joyce Deeley Research Centre, Victoria, BC V8R 6V5, Canada

ARTICLE INFO

Keywords:
CBX
Methyllysine reader proteins
Polycomb paralogs
Peptidomimetics
Epigenetics

ABSTRACT

Epigenetic regulation of gene expression is in part controlled by post-translational modifications on histone proteins. Histone methylation is a key epigenetic mark that controls gene transcription and repression. There are five human polycomb paralog proteins (Cbx2/4/6/7/8) that use their chromodomains to recognize trimethylated lysine 27 on histone 3 (H3K27me3). Recognition of the methyllysine side chain is achieved through multiple cation-pi interactions within an 'aromatic cage' motif. Despite high structural similarity within the chromodomains of this protein family, they each have unique functional roles and are linked to different cancers. Selective inhibition of different CBX proteins is desirable for both fundamental studies and potential therapeutic applications. We report here on a series of peptidic inhibitors that target certain polycomb paralogs. We have identified peptidic scaffolds with sub-micromolar potency, and will report examples that are pan-specific and that are partially selective for individual members within the family. These results highlight important structure-activity relationships that allow for differential binding to be achieved through interactions outside of the methyllysine-binding aromatic cage motif.

1. Introduction

Post-translational modifications on histones control the functions of chromatin through the actions of various epigenetic protein complexes.[1] Methyllysine reader proteins bind to post-translationally methylated lysine residues via an aromatic cage motif.[2] The five human polycomb paralog proteins (CBX2/4/6/7/8) recognize trimethyllysine residues on histone 3 and each participate in the multi-protein Polycomb Repressive Complex 1 (PRC1).[3] PRC1 serves to activate or silence genes by altering accessibility and compaction of chromatin.[4]

Each CBX protein has unique functional roles and displays distinct activities in different stages of cancer and in different tissues.[5–8] To better understand the biology of these proteins and to test their potential as drug targets, chemical tools are needed to understand the phenotypes generated by inhibition. Significant progress has been made in understanding the biological impacts of inhibiting CBX7,[9,10] but comparatively much less is understood about the other CBX proteins.[11]

Typical approaches to generating small-molecule inhibitors have proven very challenging for CBX proteins. Early efforts in our group at virtual and small molecule screening for CBX7 did not yield potent inhibitors (unpublished results). The challenges associated with screening small molecules against the CBX proteins have also been reported by others.[9,11,12]

We have previously reported a peptide-driven approach to identify a series of sub-micromolar inhibitors targeting CBX4/CBX7 and CBX6.[12–14] Potent peptidic inhibitors of CBX4/CBX7 have also been identified by the Frye group and have shown activity in cell based studies.[9,15] The first small molecule inhibitors of any CBX protein also targeted CBX7.[10,16] Recent work on new approaches to target the CBX proteins has identified inhibitors of CBX7 and CBX8 using a DNA-encoded library.[17]

No selective inhibitors have been reported for the other CBX proteins. Relatively little is known about the many biological roles of CBX2/4/6/8,

Abbreviations: Ac, acetyl; arg, arginine; CBX, chromobox; CDY, chromodomain protein Y; Chromodomain, chromatin organization modifier domain; DNA, deoxyribonucleic acid; FITC, fluorescein isothiocyanate; Fmoc, fluorenylmethoxyloxycarbonyl; FP, fluorescence polarization; Glu, glutamic acid; H3, histone 3; H3K9me3, trimethylated lysine 9 on histone 3; H3K27me3, trimethylated lysine 27 on histone 3; HP1, heterochromatin protein 1; IC_{50}, inhibitory concentration that reduces effect by 50 percent; K, lysine; Kme3, trimethyllysine; K_d, dissociation constant; Leu, leucine; MeCN, acetonitrile; MD, molecular dynamics; MORF4L1, mortality factor 4-like protein

* Corresponding author.
E-mail address: fhof@uvic.ca (F. Hof).

https://doi.org/10.1016/j.bmc.2019.115176
Received 25 June 2019; Received in revised form 9 October 2019; Accepted 17 October 2019

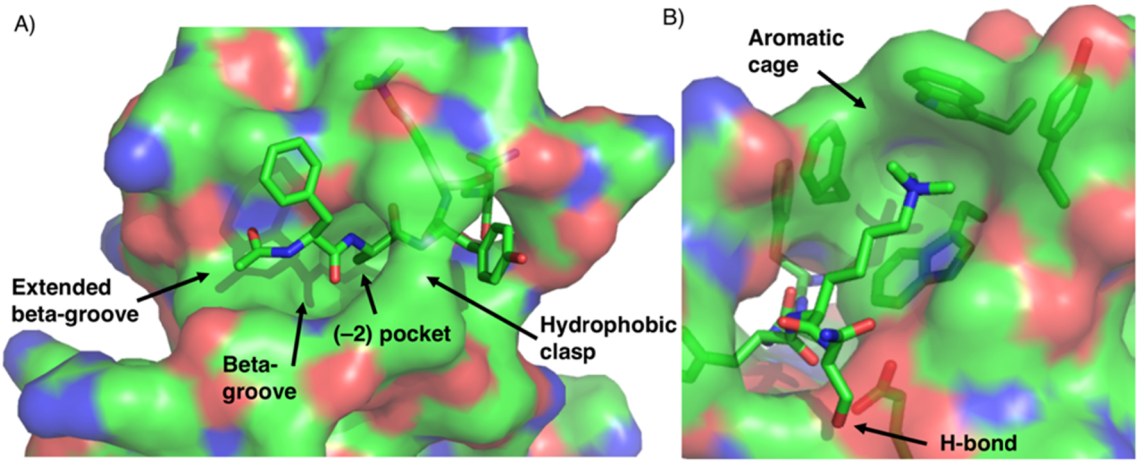


Fig. 1. Key regions of the CBX proteins responsible for binding. A and B) Co-crystal structure of Ac-FAYKme3S in complex with CBX7 (pdb: 4MN3). Key structural features of the proteins binding pocket are labeled with black arrows.

and nothing is known about the impacts of chemical inhibition of any of these proteins. The delay in progress is because of the many challenges in targeting the CBX proteins. CBX chromodomains bind to their native substrates with weak affinity,[3] undergo an induced fit mode of binding,[9] and are between 79 and 98% similar within the family (Fig. 1, Table 1).

Significant structural differences exist between the HP1 homologs (CBX1/3/5) and the polycomb paralogs (CBX2/4/6/7/8),[3] but the differences within the polycomb paralog family are subtle. The polycomb paralog proteins bind Kme3 through cation-pi interactions in their conserved aromatic cage (Fig. 1A, B). The aromatic cage is nearly identical within the family, but differences exist in the protein structure in the nearby beta-groove (Fig. 1A). Discovery and exploitation of these structural differences outside of the methyllysine binding motif are necessary for the development of selective inhibitors.

Our initial reports of CBX inhibitors first established the pentapeptide sequence Ac-FALKme3S and related analogs as inhibitors of CBX7 (Table 1).[12,13] A co-crystal structure of the Ac-FAYKme3S complex with CBX7 shows the peptide ligand forming several key binding interactions in the peptide-binding groove (Fig. 1). The *N*-terminus of the ligand sits in the protein's beta-groove, and the ligand's (−2) Ala residue points into a small hydrophobic pocket called the (−2) pocket. The hydrophobic clasp of the CBX proteins, made up of residues Val10 and Leu49, fold and clamp around the ligand (Fig. 1A). The Kme3 group of the ligand forms several cation-pi interactions in the protein's aromatic cage, and the C-terminal Ser residue of the ligand peptide hydrogen bonds with the carboxylate side chain of Glu43 (Fig. 1B).

We have previously reported a CBX6 ligand that contains a valine residue at the (−2) position relative to the ligand's Kme3 residue.[14] This substitution exploited differences within the family's (−2) hydrophobic pockets in order to generate a CBX6-selective ligand. In this work, we report novel SAR of peptidic ligands that explore multiple other regions of the CBX-ligand binding interface.

The goal of this work was to further study the structural determinants of recognition, so that we can better target each family member. A secondary goal was to create novel dye-labeled inhibitors as chemical tools that would allow for biochemical and biophysical studies of the CBX proteins. To this end, we synthesized a small library of peptidomimetic compounds and tested each compound with a panel of CBX proteins. The peptides synthesized are labeled with the fluorescent dye fluorescein isothiocyanate (FITC). Labeled inhibitors were used in for multiple forms of testing, including direct fluorescence polarization assays (to determine affinity) and microarray testing (to determine selectivity).

Table 1
Binding affinities for previously reported peptidic ligands with CBX7 and CBX8.[12]

	IC_{50} values (μM) determined by CBX7-H3K27me3 disruption	K_d values (μM) determined by ITC	
Ligand	CBX7	CBX7	CBX8
Ac-FALKme3S	11 ± 0.4	2 ± 0.2	14.2 ± 2
Ac-FAYKme3S	6 ± 0.4	2 ± 0.1	12 ± 1.4
pBr-FALKme3S	5 ± 0.3	0.3 ± 0.05	5 ± 0.3

2. Results

2.1. Synthesis of peptides

Peptides were synthesized using standard Fmoc solid-phase peptide synthesis protocols. Peptides contained either a beta-alanine residue at the *N*-terminus or a Lys(Mtt) residue at the C-terminus to allow dye labeling while still on resin. Fmoc-beta alanine was deprotected using standard protocols and the peptide was then reacted with FITC to produce compounds **1** and **2**. Peptides containing a C-terminal Lys-(Mtt) residue were selectively deprotected under mildly acidic conditions, followed by a reaction with FITC to give compounds **3**, **4**, **8**, **9**, **11**, **12** (Scheme 1).

2.2. Fluorescence polarization-driven studies to understand polycomb paralog SAR

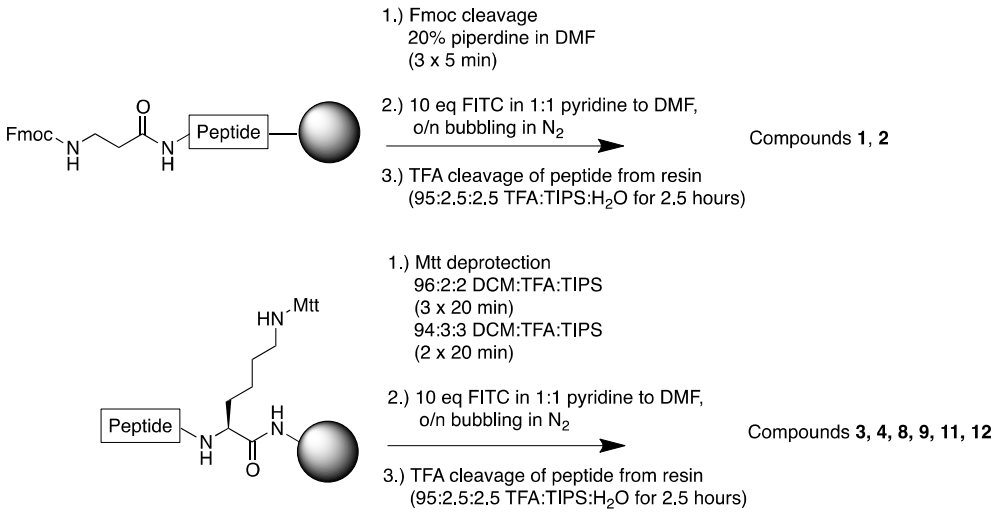
2.2.1. Dye labeling CBX7 inhibitor and N-terminal modifications

A dye-labeled analog of our previously identified CBX7 inhibitor FALKme3S was synthesized with a *N*-terminal beta-alanine residue (compound **1**) for covalent attachment to FITC. Compound **1** was screened against all CBX polycomb paralogs and the HP1 homolog CBX1. Compound **1** displayed low micromolar affinity to all CBX proteins tested (0.69–4.5 μM), with nearly equipotent binding to CBX2/4/6/7 and weaker binding to CBX1 (an HP1 paralog) and CBX8 (Fig. 2A). The previously reported K_d values of Ac-FALKme3S with CBX7 and CBX8 are similar and within the same magnitude of those reported for the dye-labeled analog **1** (Table 1, Fig. 2A).

The addition of a second phenylalanine in the *N*-terminal region of **1** resulted in compound **2**, which showed an increase in binding to CBX8 (2.6-fold) and decrease in binding to CBX7 (2.3-fold) (Fig. 2B, **2**). Compound **2** also showed a 19-fold decrease in binding to CBX1, 2 to 3-fold decrease in binding to CBX2/4 and no significant change in binding to CBX6.

2.2.2. Salt bridge interactions between ligands and CBX6 and CBX8

Another key structural difference is the presence of an Arg9 residue in CBX6 and CBX8 that is not present in CBX7, and that is located near the



Scheme 1. Synthesis of peptidic compounds **1**–**4** and **8**, **9**, **11**, **12**.

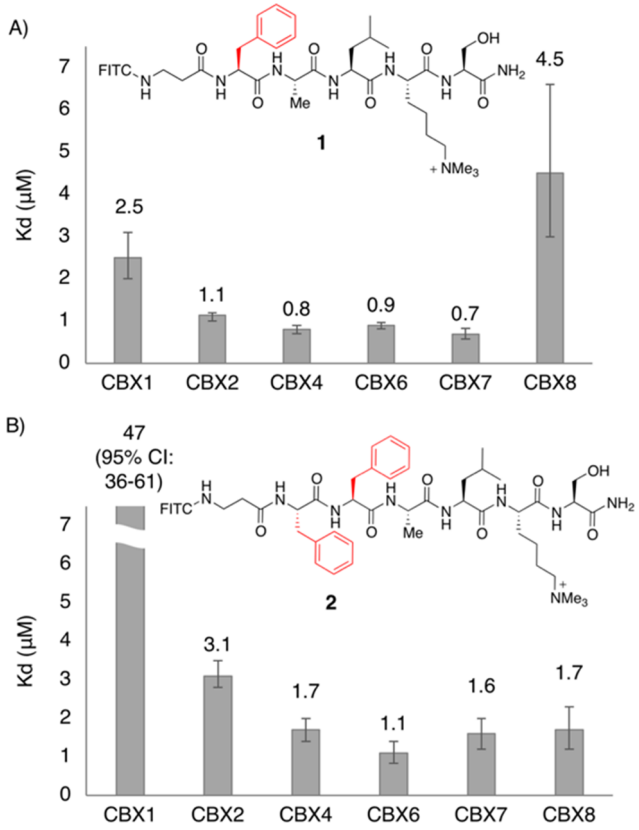


Fig. 2. Peptidic inhibitors for CBX proteins and corresponding dissociation constants for CBX1/2/4/6/7/8. A) Binding affinities and chemical structure of compound **1**, B) binding affinities and chemical structure of compound **2**. Error bars are reported as asymmetric 95% confidence intervals from experiments done in triplicate.

C-terminus of the ligand (Fig. 3A). We predicted that an anionic substitution at the (+2) position would improve binding to CBX6/8 but not CBX7. Where the *N*-terminus is free of dye modification, we used the *p*-bromobenzamide end-capping group that we had previously shown to provide a boost in potency (pBr-FALKme3S, Table 1). Two analogous compounds were synthesized containing either a Leu residue at the (+2) position (compound **3**) or a Glu residue (compound **4**) (Fig. 3). Compound **3** was a potent inhibitor of CBX6 and CBX7 with K_d values of 78 and 11 nM. Addition of a Glu residue in the (+2) position provided compound **4** (previously reported).[14] The (+2) Glu gave a small increase in binding to CBX6 (1.6-fold) and CBX8 (1.4-fold), no change in binding to CBX7 and a decrease in binding to CBX1 (2.7-fold) (Fig. 3C, D).

To investigate further the role of the Glu residue and to understand the effects of the dye on binding, we synthesized analogues of compounds **3** and **4** lacking the C-terminal lysine and FITC label. A competitive FP assay was used to determine the IC_{50} values of the unlabeled compounds **5**, **6** and **7** (Fig. 4). Compound **4** was used as the dye-labeled probe in the competitive FP assay because of its good solubility and low K_d values for all CBX proteins. Compound **5**, lacking a residue at the (+2) position, displayed binding to CBX7 with 7- and 18-fold selectivity over CBX6 and 8 respectively. The addition of a Leu residue at the (+2) position[6] did not significantly change binding to CBX6, a slight increase in binding to CBX7 was observed and a 2-fold increase in binding to CBX8. Compound **7** with a (+2) Glu residue gave a 2-fold increase in binding to CBX6 and no significant change in binding to CBX8. The trends in binding affinities observed with the dye-labeled compounds **3** and **4** are consistent with the trends seen for the unlabeled compounds **6** and **7**. These competitive assays demonstrate that the modifications that are C-terminal to the trimethyllysine residue in general improve affinity, including that the C-terminal FITC dye contributes significantly to the potency of compounds **3** and **4**. Each of these sets of binding data suggest that the ligands' (+2) Glu residue makes small improvements for binding to CBX6 and CBX8.

2.3. Docking and MD simulations of compound 6 and 7

The salt bridge between CBX8's Arg9 and the ligands' Glu residues seemed to be a small but significant provider of increased affinity. We sought to further examine the importance of the salt bridge interaction to binding of compounds **6** and **7**, using an accelerated sampling molecular dynamics technique developed in our group.[19] The SLICE method is a hybrid, iterative stochastic deterministic methodology that fuses AutoDock Vina[20] and the Amber16 molecular dynamics suite.[21] The method allows for fast structural identification of binding complexes of highly flexible hosts with peptidomimetic ligands. Binding of compounds **6** and **7** with CBX8 was investigated using SLICE.

Two 1-ns SLICE iterations were used to generate partially-induced-fit host configurations of CBX8. A third docking round of the ligands on the resulting host configurations provided starting points for a 100 ns molecular dynamics production run. After the two SLICE iterations, only the glutamate-containing compound **7** was observed to bind in the presumed correct orientation with the trimethyllysine placed in the aromatic cage. Template docking for compound **6**, using bound configurations of compound **7**, was performed to generate initial

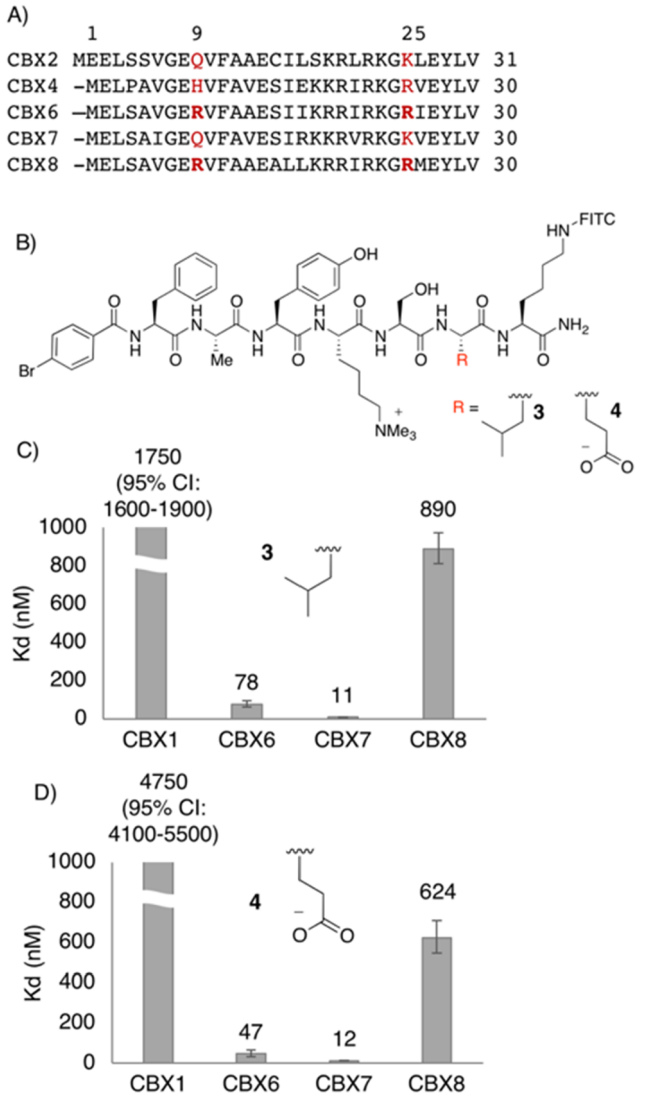


Fig. 3. Interactions between anionic ligand substituent and cationic protein residue in CBX6/CBX8. A) Sequence alignments highlighting residues predicted to interact with anionic ligand substituents determined using ClustalW2 alignment, B) Chemical structure of compounds **3** and **4**, B) and C) Binding affinities of **3** and **4** to CBX1/6/7/8. K_d values are reported in nM and error bars are reported as asymmetric 95% confidence intervals from experiments done in triplicate.

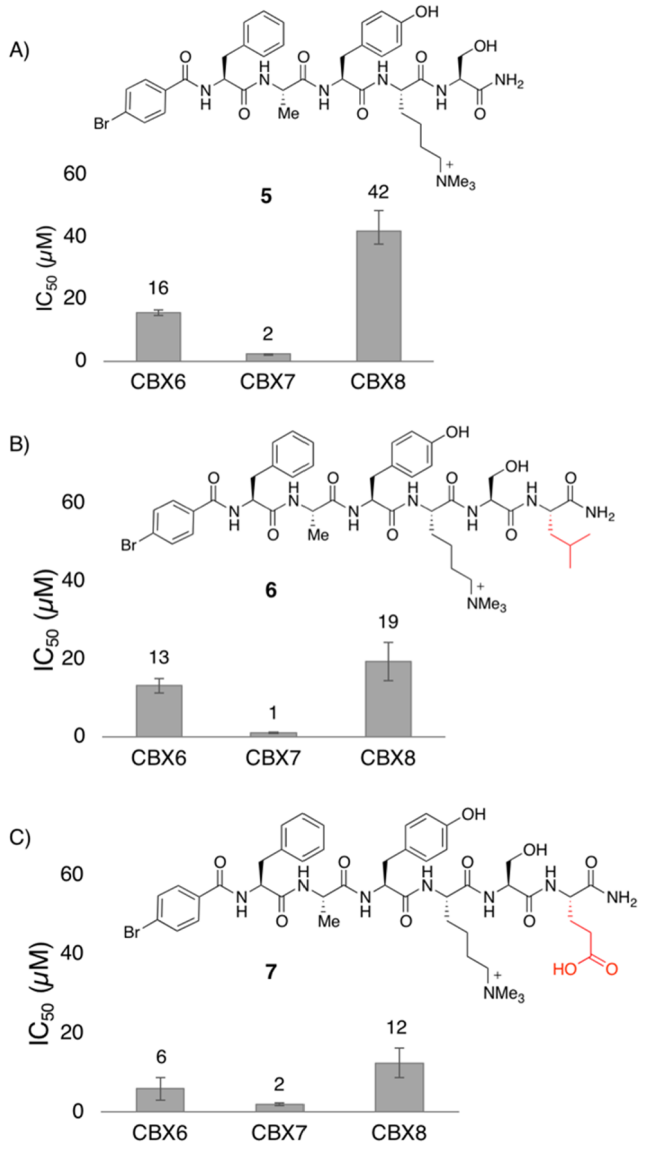


Fig. 4. IC_{50} values of unlabeled ligands **5–7** for CBX6/7/8 determined by competitive FP. A) Binding affinities and chemical structure of compound **5**, B) compound **6**, C) compound **7**. Error bars are reported as asymmetric 95% confidence intervals from experiments done in triplicate with 1 to 3 biological replicates (see supplementary Figs. 19–21).

coordinates for the final production molecular dynamics of bound compound **6**. The simulation was required to examine the behavior of a leucine in the electropositive region of CBX8 depicted in Fig. 5.

The initial docked pose of compound **7** (purple) glutamate (red) predicts a salt bridge with Arg25 instead of the proposed Arg9. The electrostatic potential surface area (APBS[22]of CBX8 shows the entire region adjacent to the aromatic cage as an electropositive environment (blue surface) consisting of several potentially competing arginine residues

The SLICE application to compounds **6** and **7** consistently showed a preference of CBX8 for **7**. However, the docked initial structure for the final production MD run presented a salt bridge of the Glu residue with Arg25, instead of the hypothesized Arg9 (Figs. 3A and 6B). This is a consequence of two factors: (i) The protein surfaces of CBX6/8 in the area surrounding the aromatic cage present significant positive charge, for which both Arg9 and Arg25 are jointly responsible;[3] (ii) the docking procedure is insufficient to tease out effectively the binding differences between Arg9 and Arg25 in the complex solvation and thermal environment in which the binding process takes place. However, the Arg25 salt bridge dissociated in the initial stages of the production MD run.

In fact, monitoring the 100 ns trajectories for interaction distances showed the docked Arg25 salt bridge to be unstable compared to the interaction with Arg9. The latter was maintained for nearly half of the simulation, as shown in Fig. 6B. Hydrophobic clasp distances, (−2) pocket placement, and aromatic cage interactions were stable for both compounds throughout the simulations and suggest that the key differences between the binding of **6** and **7** are the local interactions of the Leu/Glu substitutions (see Fig. 7).

To explore the contributions of competing interactions to the binding energy, we calculated the per-residue free energies of binding using MMPBSA.py.[23] Coordinates for the calculation for host/ligand/complex were all taken from the 100 ns trajectory.[1] Further energy

[1] This single trajectory approach ignores the induced-fit energy penalty of the host, which is likely significant, based on our experience with apo-protein simulations of CBX8. However, this contribution is also likely the same for **6** and **7**, given their structural similarities, and can be neglected for the purpose of this comparison.

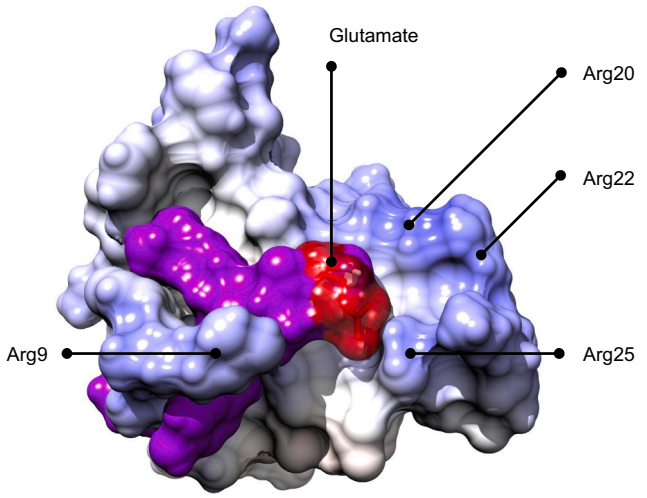


Fig. 5. Final SLICE structure of compound **7** with CBX8. The initial docked pose of compound **7** (purple) glutamate (red) predicts a salt bridge with Arg25 instead of the proposed Arg9. The electrostatic potential surface area (APBS(22)) of CBX8 shows the entire region adjacent to the aromatic cage as an electropositive environment (blue surface) consisting of several potentially competing arginine residues.

decomposition, specifically the electrostatic and solvation energies of the residues involved in the Arg9 salt-bridge and surface Glu43 hydrogen bond (Table 3), support our structural findings regarding the competition between salt bridge and hydrogen bond formation for compound **7**. Overall, the hydrogen bond interaction for compound **7** is weaker than that of compound **6**. This is an overall result that arises from significantly more stable electrostatic contributions for compound **6**, due to residence time, balanced out by a proportionally large solvent exclusion effect. On the other hand, the salt bridge interaction favors compound **7**, as initially designed: The interactions in the salt bridge region are slightly destabilizing for compound **6**, whereas the intermittent compound **7** Glu-Arg9 interaction provides an overall stabilizing effect. The difference of the total MMPBSA binding free energies for compounds **6** and **7**, using the Poisson Boltzmann solvent, was found to be 3.79 kcal/mol in favor of compound **7** using a single trajectory approach. The large size of this value compared to the observed binding differences almost certainly arise from shortcomings in the MMPBSA.py method. Differences in uncomplexed ligand stability including intramolecular enthalpic terms as well as entropic differences are ignored by MMPBSA.py.[24] Despite this known pitfall, the method is useful here for allowing a calculation of the various local contributors to binding, at a qualitative level.

2.4. Ligand substitutions at the (−1) and (−2) position

Returning to synthesis and testing of new ligands, we sought to

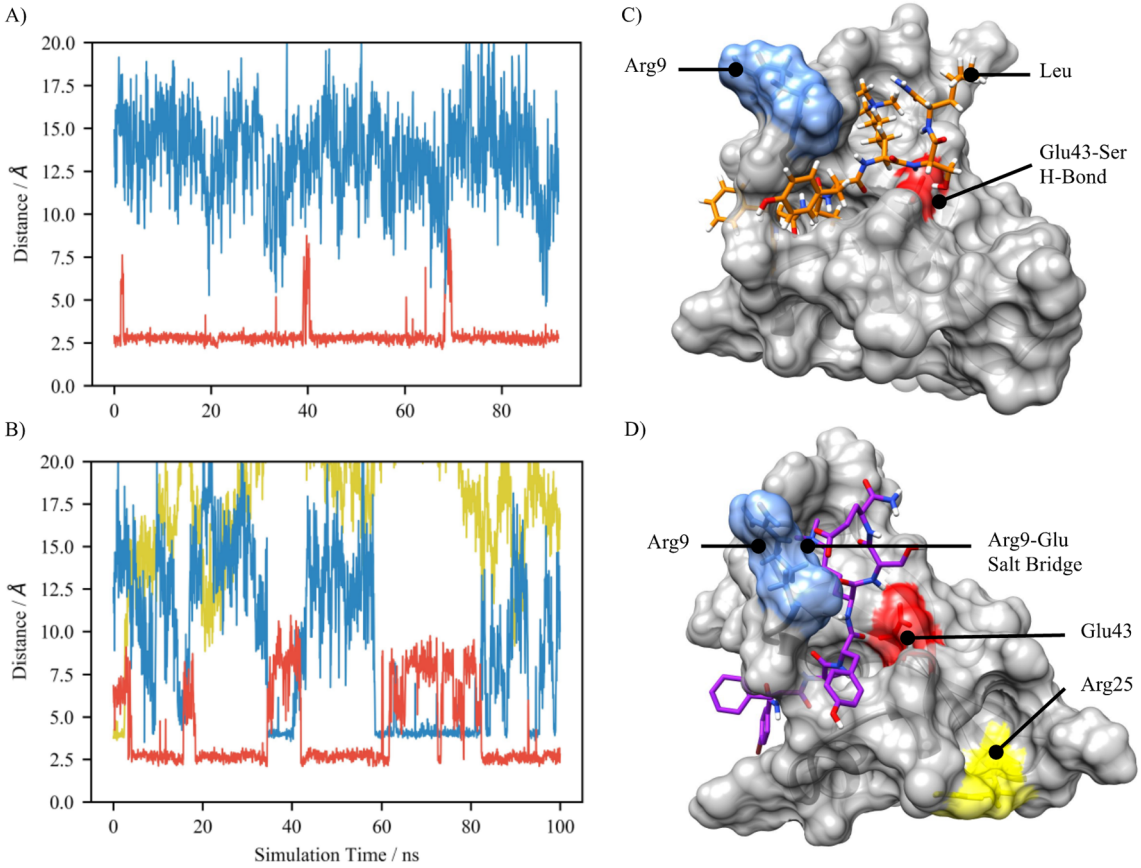


Fig. 6. Highlighted interactions of CBX8 with compounds **6** and **7**. A) The evolution of the distance between compound **6** leucine δ-carbon and CBX8 Arg9 guanidinium carbon through the 100 ns simulation is shown in blue. The compound **6** serine OH – Glu43 δ-carbon distance is shown in red, highlighting the stability of the hydrogen bond in this complex. In (B), distances shown are for: compound **7** glutamate δ-carbon – Arg25 guanidinium carbon (yellow), glutamate δ-carbon – Arg9 guanidinium carbon (blue), and compound **7** serine OH – Glu43 δ-carbon (red). C) MD Snapshot of compound **6** (stick representation) with the relevant residues on CBX8 highlighted in red and blue. D) MD Snapshot of compound **7** (stick representation) with CBX8 (with the relevant residues highlighted in blue, red and yellow).

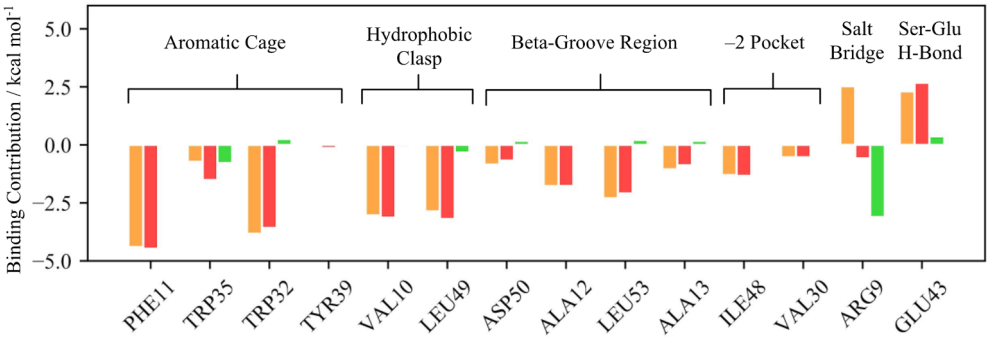


Fig. 7. MMPBSA per-residue binding free energies for CBX8 residues in the interaction region. Orange and red bars show energy values for the interaction with compound **6** (E**6**) and **7** (E**7**), respectively. The relative stabilization contribution of compound **7**, (E**7**-E**6**), is presented in green. Residues that did not contribute to binding for either compounds **6** and **7** were not included.

explore substitutions of the (−2) residue within the scaffold of **4**. An additional difference between CBX7 and CBX8 is the hydrophobic (−2) pocket. The (−2) pocket is different sizes in CBX7 and CBX8. We sought to exploit this difference by adding a larger alkyl substituent at the (−2) position. The methyl substituent at the (−2) position was replaced with an ethyl group to produce compound **8** (Fig. 8A).[14] This subtle change weakened binding to all partners except CBX8. Significantly weaker binding was observed for **8** with CBX1 (31-fold), between a 2- to 5-fold decrease in binding for CBX2/4/6/7 and no change in binding to CBX8.

We next replaced the Phe residue with a cyclopentyl moiety at the (−1) position to give compound **9** (Fig. 8B). This swap produced the most potent CBX8 inhibitor reported to date with a K_d value of 120 nM. In addition to improving potency for CBX8, the cyclopentyl side chain in the (−1) position increased binding to all CBX proteins tested. Compound **9** is 26-fold selective for CBX8 over CBX1, 3-fold selective over CBX2/6, and 5-fold selective over CBX4. The difference in affinity of **9** for CBX7 and CBX8 is not significant. To verify the observed binding affinity trends, we synthesized a dye-free analog of **9** (compound **10**) and tested it against CBX6, CBX7 and CBX8. Unlabeled **10** binds with similar affinity to CBX6 and CBX8, and is 2-fold selective for CBX8 over CBX7 (Fig. 8C, Table 4).

We hoped to push further toward CBX8 selectivity by adding a second Phe residue to **9**, as in compound **2**. To this end, we synthesized compound **11** containing an acetylated *N*-capping Phe residue in combination with the (−2) ethyl group and (−1) cyclopentyl side chain (Fig. 9A). Within the series, compound **11** exhibits the greatest difference in affinity favoring CBX8 over CBX7. **11** is most potent for CBX6, and is 4 and 2-fold selective for CBX6 and 8 over CBX7. In respect to CBX8, **11** is between 14x and 135x selective over CBX1/2/4.

Replacing the (−2) and (−1) substituents of **11** to Ala-Leu residues known to be favored by CBX1/2/4/7 did not significantly change binding to these proteins, but did decrease binding to CBX6 and 8 by a factor of 2.7 and 3.5 (Fig. 9B). Compounds **10** and **11** are selective for CBX6/7/8 over CBX1/2/4.

Selective inhibition of CBX7 over CBX4 has not yet been reported. The CBX4 chromodomain is the most similar to CBX7 (similarity score of 90%),[3,14] both bind the native histone substrate with similar affinity,[3] and almost all CBX7 ligands reported to date have had similar affinities for CBX4.[9,10,12,13] Interestingly, both **10** and **11** show significantly weaker binding to CBX4 compared to CBX7. Future efforts on selective inhibition of CBX7 may benefit from extended engagement of the peptide-binding groove.

Taken together, the studies of unlabeled inhibitors (Tables 2 and 4) and dye-labeled inhibitors (Table 5) provide many consistent structural lessons on how to tilt the selectivity among CBX proteins in the direction of one polycomb paralog or another.

2.5. Selectivity studies using a methyl reader protein microarray

The tagged peptides also allow us to use a protein microarray, which provides both validation of our FP results and more diverse

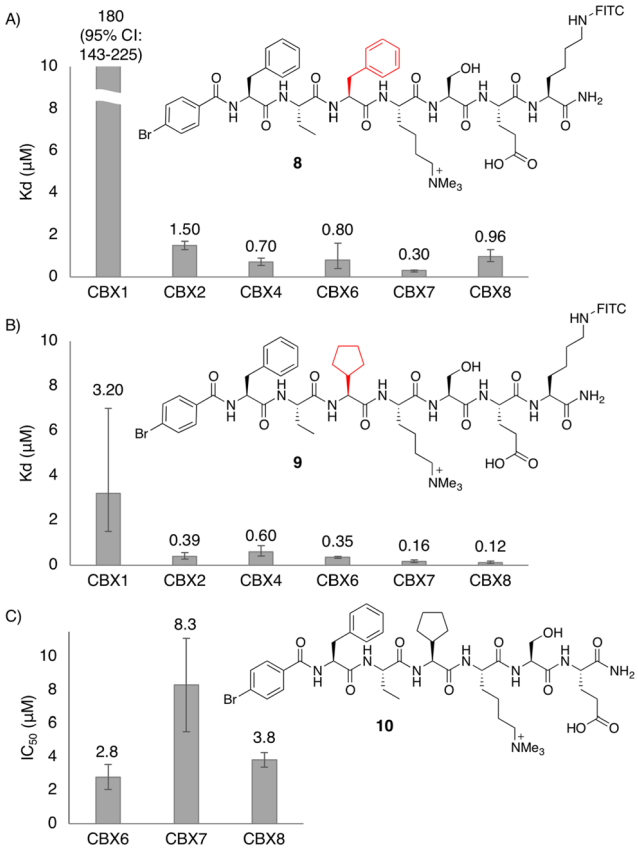


Fig. 8. Peptidic inhibitors **8–10** and corresponding dissociation constants and IC_{50} values. A) Binding affinities and chemical structure of compound **8**, B) compound **9**, and C) compound **10**. Error bars in A and B are reported as asymmetric 95% confidence intervals from experiments done in triplicate. Data in C is the average of two biological replicates each performed in triplicate with the error bars representing one standard deviation.

knowledge on selectivity beyond the CBX family of proteins. We utilized a protein microarray made up of 98 different recombinant human methyl reader proteins (including all CBX proteins) arrayed in duplicate.[25] Initial testing of tetramethylrhodamine isothiocyanate (TRITC) dye-labeled inhibitors produced the expected binding trends but with a high degree of background fluorescence. To prevent this, a biotinylated analog of **3** was synthesized (compound **13**, Fig. 10). The microarray was incubated with the probe, and the binding of the probe was imaged using a fluorescent streptavidin reagent, providing a very low background signal (Fig. 10A).

The inhibitor tested in the microarray showed excellent selectivity for CBX proteins over a broad selection of other methyl readers. Some off-target binding to the chromodomain Y like (CDYL) proteins was observed,[26] along with weak off-target binding to the chromodomain-

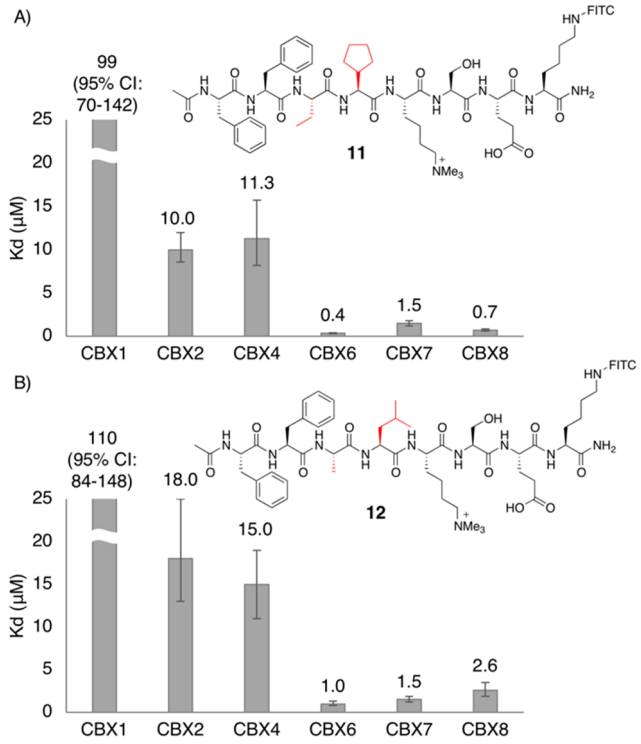


Fig. 9. Peptidic inhibitors **11** and **12** and corresponding dissociation constants for CBX1/2/4/6/7/8. A) Binding affinities and chemical structure of compound **11**, B) and compound **12**. Error bars are reported as asymmetric 95% confidence intervals from experiments done in duplicate for CBX1/2/4/6/7 and triplicate for CBX8.

containing mortality factor 4-like protein (MORF4L1 or MRG15). Compound **13** showed the highest selectivity for CBX4 and 7, with weaker binding observed to CBX2/6/8 (Fig. 10).

To ensure this data was in agreement with K_d values we collected from our FP assays, ImageJ software was used to quantify the brightness of each microarray spot and this was plotted against the K_d values for analogous dye-labeled compound of **13** (Fig. 10C). In general, it is clear that proteins with weak K_d values also have weak pixel intensity showing that the microarray data is in qualitative agreement with our solution-phase results. CBX6 shows significantly less pixel brightness than would be expected from its K_d for this probe ligand, suggesting that microarrayed CBX6 does not accurately represent solution activity.

2.6. Cell-based data

We sought to determine the ability of the dye-labeled peptidic agents to be used in cell-based studies. Our efforts to study the inhibitors in cells included live cell imaging, flow cytometry, and MTT-based viability studies.

Cell-based studies showed varying degrees of cellular uptake and minimal changes cell cycle distribution as well as viability. Live cell imaging and immunofluorescence confocal imaging using TOV21G ovarian carcinoma cells and PC3 prostate cancer cells, treated with compounds **2** and **4** did not show compounds entering the cytoplasm or nucleus. Inhibitors were seen in characteristic punctate dots in the cytoplasm at concentrations above 10 μM, suggesting endosomal entrapment (data not shown). Flow cytometry experiments with TOV21G cells treated with compound **2** showed uptake of the inhibitor, supported by the formation of a population of fluorescent cells in each case (Fig. S1). In the presence of the inhibitor, there was no observable change in the distribution of cells across Go, G1, S or G2/M (Table S1). This is also consistent with compounds being taken up in endosomes, but not able to escape to cytoplasm or nucleus in order to have a biological effect. We also carried out MTT assays, used to measure metabolic activity and cell viability, with TOV21G cells treated with inhibitor **4**, and unlabeled analogs **6** and **7**. A slight decrease in cell viability was seen for the cells treated with **4**, however we did not observe a dose-response for the unlabelled analogs **6** and **7** (Fig. S2). We conclude that these compounds are unfortunately not suitable for cell-based activity studies.

Table 3
MMPBSA.py per-residue energies for selected residues in salt bridge and adjacent glu43 hydrogen bond.

		Eel[α]	Esol[β]	Etot[γ]	Eint[ε]
CBX8/compound 6					
Hydrogen Bond	Glu43 (CBX8)	−37.32	40.61	2.31	−4.16
	Ser1 (compound **6**)	−13.01	7.07	−6.47	
Salt Bridge Region	Arg9 (CBX8)	33.13	−29.38	2.52	2.05
	Leu2 (compound **6**)	−0.49	0.65	−0.47	
CBX8/compound 7					
Hydrogen Bond	Glu43 (CBX8)	−16.08	19.71	2.68	−2.11
	Ser1 (compound **7**)	−9.97	5.71	−4.79	
Salt Bridge Region	Arg9 (CBX8)	−22.58	23.66	−0.58	−2.81
	Glu2 (compound **7**)	−68.53	66.89	−2.23	
	Glu2 (Compound **7**)	−68.53	66.89	−2.23	

Electrostatic potential[α], PBSA solvation energy[β], total per-residue energy[γ], and combined total energies[ε] (sum of E_{tot} for the two residues) are presented for the interacting residues on CBX8 and the two ligands. All energies reported in kcal/mol. Values listed are overall per-residue values, and not pairwise energies.

Table 4
IC_{50} values and calculated K_i values for compounds **10**.

	IC_{50} values (μM) determined by CBX-**4** disruption	Calculated K_i (μM)[18]
CBX6	2.8 ± 0.8	0.093 ± 0.027
CBX7	8.3 ± 2.8	0.14 ± 0.048
CBX8	3.8 ± 0.4	0.043 ± 0.006

3. Discussion

Inhibitors were developed that are either pan-specific or partially selective within the polycomb paralog family. While none of the inhibitors developed were selective for a single CBX protein, we have

Table 2
IC_{50} values and calculated K_i values for compounds **5**–**7**.

	IC_{50} values (μM) determined by CBX7-**4** disruption			Calculated K_i (μM)[18]		
Compound	CBX6	CBX7	CBX8	CBX6	CBX7	CBX8
5	16 95% CI: (15–17)	2.3 95% CI: (2.1–2.7)	42 95% CI: (39–47)	0.7 ± 0.07	0.02 ± 0.02	5.2 ± 0.66
6	13 95% CI: (11–15)	1.1 95% CI: (0.9–1.3)	19 95% CI: (15–24)	0.57 ± 0.07	0.007 ± 0.0006	2.1 ± 0.52
7	6 95% CI: (3.1–8.9)	2 95% CI: (1.6–2.4)	12 95% CI: (8–16)	0.24 ± 0.09	0.024 ± 0.004	1.2 ± 0.28

Table 5
All K_d values arising from titrations of CBX proteins into dye labeled peptides **1**–**4**, **8**, **9**, **11**, **12**. Errors are reported as 95% confidence intervals.

	K_d (μM)					
Compound	CBX1	CBX2	CBX4	CBX6	CBX7	CBX8
1	2.5 95% CI: 2.0–3.1	1.1 95% CI: 1.0–1.2	0.8 95% CI: 0.7–0.9	0.9 95% CI: 0.82–0.0.96	0.7 95% CI: 0.57–0.83	4.5 95% CI: 3.0–6.8
2	47 95% CI: 36–61	3.1 95% CI: 2.8–3.5	1.7 95% CI: 1.4–2.0	1.1 95% CI: 0.8–1.4	1.6 95% CI: 1.2–2.0	1.7 95% CI: 1.2–3.2
3	1.75 95% CI: 1.6–1.9	N.D.	N.D.	0.078 95% CI: 0.063–0.097	0.011 95% CI: 0.009–0.013	0.89 95% CI: 0.81–0.98
4	4.75 95% CI: 4.1–5.5	N.D.	N.D.	0.047 95% CI: 0.035–0.062	0.012 95% CI: 0.0097–0.015	0.624 95% CI: 0.57–0.68
8	180 95% CI: 143–225	1.5 95% CI: 1.3–1.7	0.7 95% CI: 0.57–0.93	0.8 95% CI: 0.4–1.6	0.3 95% CI: 0.23–0.33	0.96 95% CI: 0.7–1.3
9	3.2 95% CI: 1.5–7.0	0.39 95% CI: 0.27–0.56	0.60 95% CI: 0.41–0.87	0.35 95% CI: 0.3–0.4	0.16 95% CI: 0.11–0.23	0.12 95% CI: 0.08–0.19
11	99 95% CI: 70–142	10 95% CI: 8.6–12	11.3 95% CI: 8.2–15.7	0.4 95% CI: 0.3–0.4	1.5 95% CI: 1.2–1.8	0.7 95% CI: 0.63–0.84
12	110 95% CI: 84–148	18 95% CI: 13–25	15 95% CI: 11–19	1.0 95% CI: 0.8–1.3	1.5 95% CI: 1.2–1.9	2.6 95% CI: 1.9–3.5

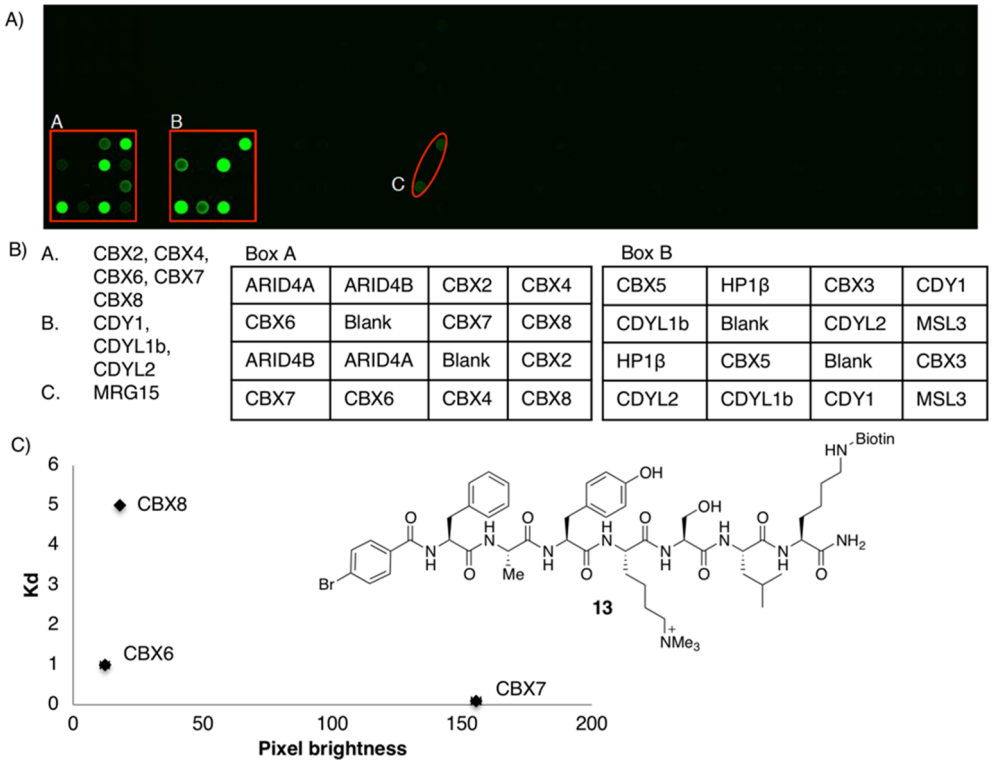


Fig. 10. Protein microarray made up of 98 human methyl reader proteins shows peptidic inhibitors are selective against a broad set of methyl reader targets. Proteins were coated onto the wells in each plate with each square representing a family of proteins. A) Protein microarray with compound **13**. B) Legend of proteins in box A and B. Box A contains polycomb paralogs, box B contains HP1 homologs and CDYL proteins, and box C highlights weak off target binding to MRG domains. The full map of the protein microarray is available in Supporting Fig. S28. C) K_d values of compound **3** (dye labeled analog of **13**) with CBX proteins plotted against brightness of each spot from protein microarray (quantified using ImageJ) and chemical structure of biotin-tagged compound **13**.

identified several compounds with sub-micromolar affinity including two inhibitors that bind CBX6/7 with IC_{50} values < 100 nM[3,4].

Several compounds displayed class selectivity, with preferential binding to CBX6/7/8 over CBX1/2/4. We also identified inhibitors with affinity for CBX6/8 over CBX7, which is unusual because CBX7 consistently gives higher affinity binding to peptidic and small molecule ligands.[9,10,12,13,15–17] This includes the most potent CBX8 inhibitor to date (**9**, 120 nM) and an inhibitor that is 2-fold selective for CBX8 over CBX7[10]. Our work also produced the first peptidic inhibitor that is 10-fold selective for CBX7 over the highly similar chromodomain of CBX4 (compound **10**).

We have identified areas of the peptide-binding groove that are different within the family of CBX proteins and form distinct interactions with the peptide ligands reported. Substitutions at the (−1) position effect the hydrophobic clasp of the CBX proteins and alter binding affinities differentially within the family. Addition of a cyclopentyl moiety in the (−1) position increases binding to all CBX members with the greatest increase observed for binding to CBX8. The addition of two Phe residues at the *N*-terminus of the ligand diminishes binding to CBX1/2/4 to give inhibitors that are selective for CBX6/7/8[9,10]. We predict that future efforts targeting the extended beta-groove, (−1) and (−2) position of the protein will aid in the discovery of selective inhibitors.

Substitutions at the (+2) position of the ligand that participate in salt-bridge interactions alter the binding to CBX6 and CBX8. Our prediction that a Glu residue at the (+2) position could improve binding affinity to CBX6 and 8 by interacting with Arg9 was partially correct. We did observe favorable salt-bridge interactions with a ligand (+2)

Glu interacting with the protein residues Arg9 and Arg25 in MD simulations with CBX8. However, this interaction destabilized a key hydrogen bond between the (+1) serine in the ligand with Glu43 in CBX8, and the overall observed tilt in selectivity was < 2-fold.

The protein microarray studies with **13** show the inhibitors to be highly selective for the CBX polycomb paralogs over many other methyllysine readers. Potent off-target binding was observed with the CDY proteins. The peptidomimetic inhibitor UNC3866 targeting CBX4/CBX7, was discovered to have off target binding to the CDY proteins.[9] The authors followed up this work by repurposing the scaffold to develop a combinatorial peptide library resulting in the discovery of potent inhibitors of CDYL1/CDYL2,[26] and small molecule CDYL inhibitors have also been recently reported, with varying degrees of selectivity over CBX proteins.[27] Differentiating binding between the two highly similar families of chromodomains will be a persistent challenge as inhibitors continue to be developed.

While the dye-labeled agents gave access to new biochemical assays, our efforts to use them in cell-based studies were unsuccessful. Several earlier CBX inhibitors have shown poor cellular activities due to low permeability, and the addition of FITC is unlikely to have improved this situation. Poor solubility of the peptidic inhibitors was another challenge and limited our ability to test higher concentrations of the inhibitors. Future efforts to use the reported inhibitors in cell-based studies will require alternative delivery strategies.

4. Conclusion

The goals of this work were to study the structural determinants of recognition for the CBX proteins and to create dye-labeled inhibitors as tools for biochemical and biophysical studies of the CBX proteins. We have successfully created potent inhibitors for each CBX polycomb paralog protein. The inhibitors reported are useful tools for biochemical assays and for future competitive based screens for the discovery of new ligands.

The SAR learned from this work provides new insights into the structure and molecular recognition properties of these proteins outside of the previously well explored aromatic cage binding pocket. This SAR lays the foundation for creating highly selective and cell-permeable chemical tools to study the role of CBX proteins in epigenetic regulation, which we will report in due course.

Author Contributions

The manuscript was written by N.M with revisions provided by F.H. N.M carried out the experimental design, synthesis, and characterization of peptides and binding affinities. J.M wrote and conceived of the modeling and MD simulations with input for revisions provided by I.P. C.C and C.W assisted with peptide synthesis. M.G, T.B and R.B grew and purified all CBX proteins. N.M, M.G and T.B carried out the FP assays. J.L and L.D conducted the cell-based studies. All authors have given approval to the final version of the manuscript.

Funding Sources

N.M and F.H thank the Prostate Cancer Foundation of British Columbia, West Coast Motorcycle Ride to Live, and CCSRI. Funding for the computational studies was provided by the National Science and Engineering Research Council of Canada, the Canada Foundation for Innovation, and the British Columbia Knowledge Development Fund. This research was in part performed in part using the Compute Canada and WestGrid computing resources.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Acknowledgment

F.H thanks the Canada Research Chairs program. Probing of arrayed methyl reader domains was made possible via the UT MDACC Protein array and analysis core (PAAC) CPRIT Grant RP180804 (Directed by Mark T. Bedford).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bmc.2019.115176.

References

1. Rothbart SB, Strahl BD. Interpreting the language of histone and DNA modifications. *Biochim Biophys Acta.* 2014;1839:627–643.
2. Musselman CA, Khorasanizadeh S, Kutateladze TG. Towards understanding methyllysine readout. *Biochim Biophys Acta.* 2014;1839:686–693.
3. Kaustov L, Ouyang H, Amaya M, et al. Recognition and specificity determinants of the human cbx chromodomains. *J Biol Chem.* 2011;286:521–529.
4. Aranda S, Mas G, Di Croce L. Regulation of gene transcription by Polycomb proteins. *Sci Adv.* 2015;1:e1500737.
5. Richly H, Aloia L, Di Croce L. Roles of the Polycomb group proteins in stem cells and cancer. *Cell Death Dis.* 2011;2:e204.
6. Vincenz C, Kerppola TK. Different polycomb group CBX family proteins associate with distinct regions of chromatin using nonhomologous protein sequences. *Proc Natl Acad Sci U S A.* 2008;105:16572–16577.
7. Ma RG, Zhang Y, Sun TT, Cheng B. Epigenetic regulation by polycomb group complexes: focus on roles of CBX proteins. *J Zhejiang Univ Sci B.* 2014;15:412–428.
8. Koppens M, van Lohuizen M. Context-dependent actions of Polycomb repressors in cancer. *Oncogene.* 2015;35:1341.
9. Stuckey JI, Dickson BM, Cheng N, et al. A cellular chemical probe targeting the chromodomains of Polycomb repressive complex 1. *Nat Chem Biol.* 2016;12:180–187.
10. Ren C, Morohashi K, Plotnikov AN, et al. Small-molecule modulators of methyl-lysine binding for the CBX7 chromodomain. *Chem Biol.* 2015;22:161–168.
11. Milosevich N, Warmerdam Z, Hof F. Structural aspects of small-molecule inhibition of methyllysine reader proteins. *Future Med Chem.* 2016;8:1681–1702.
12. Simhadri C, Daze KD, Douglas SF, et al. Chromodomain antagonists that target the polycomb-group methyllysine reader protein chromobox homolog 7 (CBX7). *J Med Chem.* 2014;57:2874–2883.
13. Simhadri C, Gignac MC, Anderson CJ, et al. Structure-activity relationships of Cbx7 inhibitors, including selectivity studies against other Cbx proteins. *ACS Omega.* 2016;1:541–551.
14. Milosevich N, Gignac MC, McFarlane J, et al. Selective inhibition of CBX6: A Methyllysine reader protein in the polycomb family. *ACS Med Chem Lett.* 2016;7:139–144.
15. Stuckey JI, Simpson C, Norris-Drouin JL, et al. Structure-activity relationships and kinetic studies of peptidic antagonists of CBX chromodomains. *J Med Chem.* 2016;59:8913–8923.
16. Ren C, Smith SG, Yap KL, et al. Structure-guided discovery of selective antagonists for the chromodomain of polycomb repressive protein CBX7. *ACS Med Chem Lett.* 2016;7:601–605.
17. Denton KE, Wang S, Gignac MC, et al. Robustness of In vitro selection assays of DNA-encoded peptidomimetic ligands to CBX7 and CBX8. *SLAS Discov.* 2018;23:417–428.
18. Nikolovska-Coleska Z, Wang R, Fang X, et al. Development and optimization of a binding assay for the XIAP BIR3 domain using fluorescence polarization. *Anal Biochem.* 2004;332:261–273.
19. James M, Katherine K, Irina P. Accelerated Molecular dynamics for structural prediction in protein/peptide binding: The SLICE method. ChemRXiv, deposited June 19, 2019, https://doi.org/10.26434/chemrxiv.8297129.v1.
20. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem.* 2010;31:455–461.
21. Case IYB-S DA, Brozell SR, Cerutti DS, et al. *AMBER 2018.* San Francisco: University of California; 2018.
22. Jurrus E, Engel D, Star K, et al. Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* 2018;27:112–128.
23. Miller BR, McGee TD, Swails JM, et al. MMPBSA.py: an efficient program for end-state free energy calculations. *J Chem Theory Comput.* 2012;8:3314–3321.
24. Wang C, Nguyen PH, Pham K, et al. Calculating protein–ligand binding affinities with MMPBSA: Method and error analysis. *J Comput Chem.* 2016;37:2436–2446.
25. Kim J, Daniel J, Espejo A, et al. Tudor, MBT and chromo domains gauge the degree of lysine methylation. *EMBO Rep.* 2006;7:397–403.
26. Barnash KD, Lamb KN, Stuckey JI, et al. Chromodomain ligand optimization via target-class directed combinatorial repurposing. *ACS Chem Biol.* 2016;11:2475–2483.
27. Yang L, Liu Y, Fan M, et al. Zhang L Identification and characterization of benzo[d]oxazol-2(3H)-one derivatives as the first potent and selective small-molecule inhibitors of chromodomain protein CDYL. *Eur J Med Chem.* 2019;182:111656.

Chapter 8

Publication: Optimization of Ligands Using Focused DNA-Encoded Libraries To Develop a Selective, Cell-Permeable CBX8 Chromodomain Inhibitor

8.1 Preface

In this publication, DNA encoded libraries along other experimental techniques were used to drive the discovery of selective CBX8 peptide-based inhibitors. All computational modelling and analysis was performed by James McFarlane. Computational work for this publication includes residue parameterization, docking via the SLICE method, submission of MD trajectories, and subsequent analysis for structural metrics including hydrogen bond formation and RMSD fluctuations. For full author contributions, please refer to Page 129 of the following publication.

8.2 Publication

Full publication including links to supplementary information may be found at the following link:

https://doi.org/10.1021/acschembio.9b00654

ACS chemical biology

Articles

Cite This: ACS Chem. Biol. 2020, 15, 112−131

pubs.acs.org/acschemicalbiology

Optimization of Ligands Using Focused DNA-Encoded Libraries To Develop a Selective, Cell-Permeable CBX8 Chromodomain Inhibitor

Sijie Wang,[†] Kyle E. Denton,[†] Kathryn F. Hobbs,[§] Tyler Weaver,[§] James M. B. McFarlane,[‡] Katelyn E. Connelly,[†] Michael C. Gignac,[‡] Natalia Milosevich,[‡] Fraser Hof,[‡] Irina Paci,[‡] Catherine A. Musselman,[§] Emily C. Dykhuizen,*[,†] and Casey J. Krusemark*[,†]

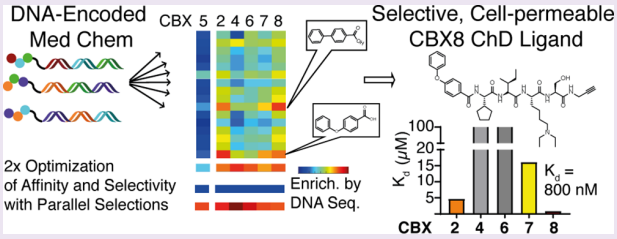
[†]Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University and Purdue University Center for Cancer Research, 575 Stadium Mall Drive, West Lafayette, Indiana 47906, United States

[‡]Department of Chemistry, University of Victoria, Victoria V8W 3V6, Canada

[§]Department of Biochemistry, Carver College of Medicine, University of Iowa, 51 Newton Road, Iowa City, Iowa 52242, United States

Supporting Information

ABSTRACT: Polycomb repressive complex 1 (PRC1) is critical for mediating gene expression during development. Five chromobox (CBX) homolog proteins, CBX2, CBX4, CBX6, CBX7, and CBX8, are incorporated into PRC1 complexes, where they mediate targeting to trimethylated lysine 27 of histone H3 (H3K27me3) *via* the N-terminal chromodomain (ChD). Individual CBX paralogs have been implicated as drug targets in cancer; however, high similarities in sequence and structure among the CBX ChDs provide a major obstacle in developing selective CBX ChD inhibitors. Here we report the selection of small, focused, DNA-encoded libraries (DELs) against multiple homologous ChDs to identify modifications to a parental ligand that confer both selectivity and potency for the ChD of CBX8. This on-DNA, medicinal chemistry approach enabled the development of SW2_110A, a selective, cell-permeable inhibitor of the CBX8 ChD. SW2_110A binds CBX8 ChD with a K_d of 800 nM, with minimal 5-fold selectivity for CBX8 ChD over all other CBX paralogs *in vitro*. SW2_110A specifically inhibits the association of CBX8 with chromatin in cells and inhibits the proliferation of THP1 leukemia cells driven by the MLL-AF9 translocation. In THP1 cells, SW2_110A treatment results in a significant decrease in the expression of MLL-AF9 target genes, including HOXA9, validating the previously established role for CBX8 in MLL-AF9 transcriptional activation, and defining the ChD as necessary for this function. The success of SW2_110A provides great promise for the development of highly selective and cell-permeable probes for the full CBX family. In addition, the approach taken provides a proof-of-principle demonstration of how DELs can be used iteratively for optimization of both ligand potency and selectivity.



Polycomb group (PcG) proteins are transcriptional repressors required for proper body segmentation in *Drosophila*[1] and for maintaining progenitor cell populations in mammals.[2] PcG proteins are part of two distinct complexes, polycomb repressive complex 1 (PRC1) and polycomb repressive complex 2 (PRC2) (Figure 1A).[3] Canonical polycomb function, as defined in *Drosophila*, begins with PRC2-mediated trimethylation of lysine 27 of histone 3 (H3K27me3), which recruits PRC1 *via* the chromodomain (ChD) of the chromobox homolog (CBX) subunit. PRC1 then compacts chromatin and ubiquitinates lysine 119 on histone H2A to promote transcriptional repression.[4,5]

In mammals, PRC1 subunits are represented by multiple, mutually exclusive paralogs that combinatorially assemble into dozens of distinct PRC1 complexes (Figure 1A).[6,7] The CBX subunit is represented by five paralogs (CBX2, CBX4, CBX6, CBX7, and CBX8) that shift in expression during development,[8−12] as well as cancer progression.[13,14] Studies in multiple cell types show that individual paralogs, even when expressed simultaneously, have unique and non-overlapping functions in development and disease.[8,10] In particular, CBX8 has recently emerged as a potential oncogenic target in multiple malignancies. It drives growth in lymphoma,[15] hepatocellular carcinoma,[16] breast cancer,[17] and leukemia with MLL translocations.[18] While a dependency on CBX8 has been defined at the genetic level, potential druggable sites on the protein have not been explored, and chemical probes specifically targeting CBX8 have not been developed.

The high flexibility of the apo structures, the shallow, extended nature of the peptide binding site, and the high sequence similarity among paralogs present significant challenges for development of potent and selective inhibitors

Received: August 9, 2019
Accepted: November 22, 2019
Published: November 22, 2019

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

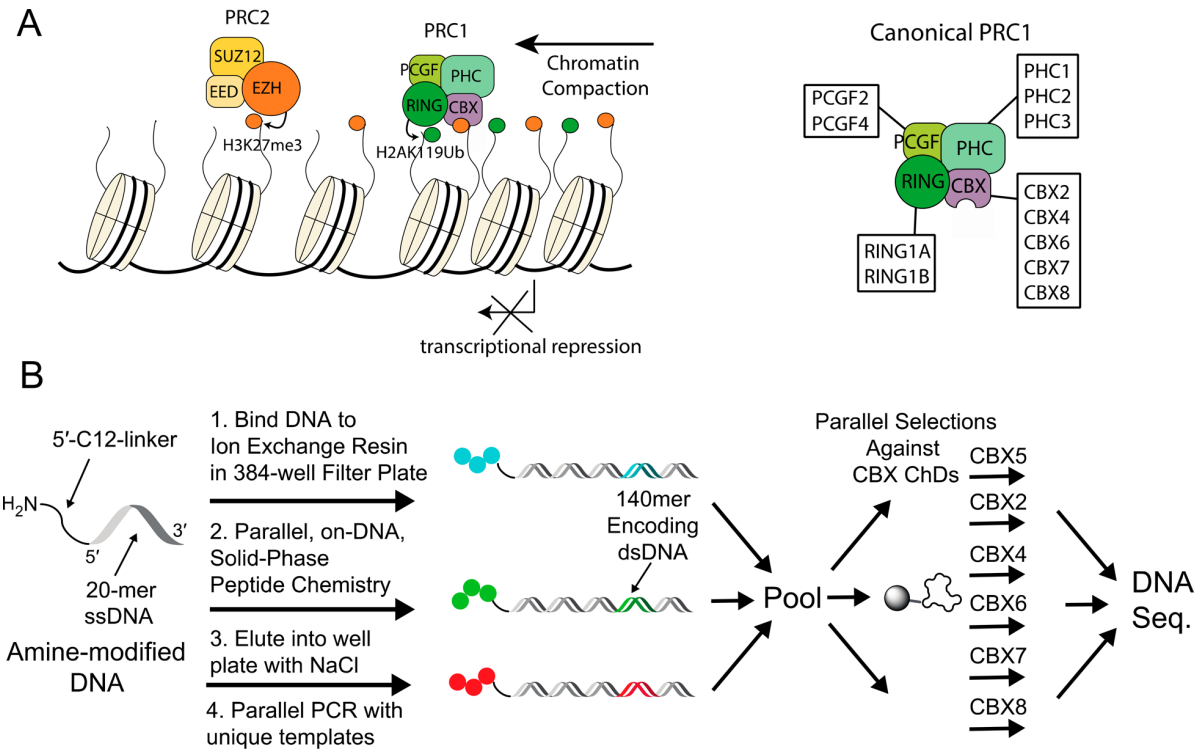


Figure 1. (A) (left) Canonical polycomb-mediated gene repression. Trimethyllysine marks installed by PRC2 on H3K27 are recognized by ChDs in the CBX subunit of PRC1. This is followed by monoubiquitination at H2AK119, chromatin compaction, and transcriptional repression. (right) Paralogous subunits assemble combinatorially to produce distinct PRC1 complexes with unknown differential function. (B) Preparation and selection of DNA-encoded chemical libraries for CBX ChD ligands. Positional scanning libraries were prepared by parallel chemical modification of an amine-modified oligonucleotide immobilized on DEAE Sepharose as a solid support. Subsequent encoding was performed by parallel PCRs with unique templates. Libraries were pooled, split, and selected against immobilized CBX ChDs. Enriched libraries were then pooled for DNA sequencing.

to CBX ChDs. Computational analysis of the peptide-bound polycomb CBX ChD structures suggests that the aromatic cage forming the trimethyllysine binding site is "druggable" based on binding site volume, enclosure, and hydrophobicity.[19] Yet, prior work has shown this site to be particularly difficult to target with traditional small molecules. Two reported small-molecule ChD inhibitors target CBX7,[20,21] yet these molecules display weak (~20 μM) affinity for CBX7 and over 10-fold weaker affinity for CBX8. Larger molecular weight trimethyllysine-containing peptidomimetics (5–6-mers) developed for CBX4,[22] CBX6,[23] and CBX7[22,24] ChDs display much greater affinity (<1 μM); however, they have limited cell permeability.

An additional challenge lies in developing ChD inhibitors with specificity for one paralog over another. There is high sequence similarity among the CBX ChDs, particularly among the polycomb CBX ChDs (CBX2, CBX4, CBX6, CBX7, and CBX8) that recognize H3K27me3 (>67% conserved residues).[25] Moderate selectivity has been achieved for CBX6 ChD,[23] and CBX7/CBX4 ChDs.[22,24] No ligand has been developed with selectivity for CBX8 ChD.

To address these challenges, we have employed DNA-encoded chemical libraries, which have numerous advantages over conventional ligand optimization approaches.[26,27] In a previous study, we described an approach for synthesizing and selecting small DNA-encoded libraries (DELs) of peptidic compounds against a panel of targets (see Figure 1B for overview). We used previously published CBX7 ChD ligands with over 10-fold selectivity over CBX8 ChD to develop quantitative metrics for affinity selection assays of DNA-encoded libraries against CBX ChDs.[22] We demonstrated that selection assays are capable of faithfully replicating known structure–activity relationships (SARs) of CBX7 and CBX8 ChD ligands and identified five monomers that increased affinity and selectivity to CBX8 ChD.[28] In this article, we utilize DNA-encoding and affinity selection with on-DNA medicinal chemistry optimization to obtain CBX8 ChD inhibitors with high affinity (3–800 nM), selectivity (>5–20-fold over other paralogs), and cell permeability. We used these ligands as chemical probes to define the CBX8 ChD as a therapeutic target in MLL-AF9 leukemia.

RESULTS AND DISCUSSION

***In Vitro* Selection Assays of Ligands to CBX Chromodomains *via* DNA-Encoded Positional Scanning Library (PSL).** *First-Generation DNA-Encoded Positional Scanning Library (PSL1).* We performed selections of a DNA-encoded positional scanning library (PSL1) reported in Denton *et al.*[28] against all five ChDs of the polycomb CBX paralogs and the ChD of CBX5, an HP1 protein (Supporting Information, Figure SI 1). As observed previously with selections against CBX8 and CBX7 ChDs, position −2 (P(−2), see Figure SI 1 for overview) was the most critical position for determining selectivity for binding to CBX paralogs due to differences in the size of a hydrophobic binding pocket lined by two valines and a leucine in CBX4 and CBX7, but by a valine, leucine, and alanine in CBX2 and

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112–131

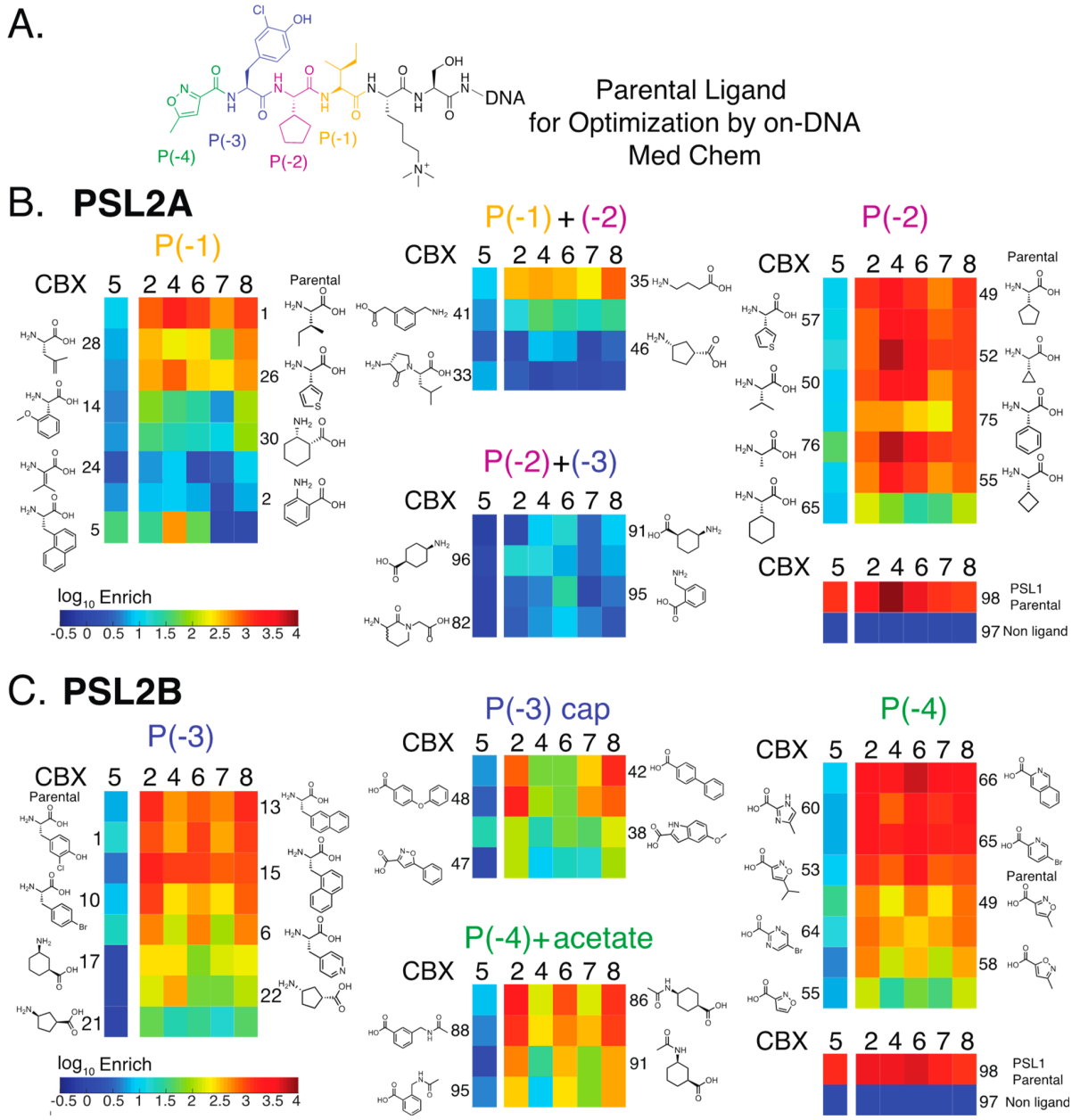


Figure 2. Select enrichment data from affinity-based selection of 192 compounds (PSL2) against six CBX ChD isoforms by DNA sequencing. (A) Compound KED98 was utilized as the parental ligand for PSL2. Unique monomers (188) were tested in total at four positions along with four replicates of the parental peptide in (B) PSL2A and (C) PSL2B. Parallel selections against all five polycomb CBX ChD and CBX5 ChD using PSL2 library were performed, and selected DNA pools sequenced. A color map designates the enrichment of each library member to a particular CBX ChD. For each position, representative synthons are ordered by enrichment of CBX8 in decreasing order. Enrichment was calculated as the fold change of sequencing reads from selected DNA over sequencing reads of input DNA and was normalized to a non-ligand control.

CBX8.[22−29] We synthesized two off-DNA ligands (KED97 and KED98) composed of monomers that gave improved affinity and selectivity for CBX8 over CBX7. Both peptides are highly selective for CBX8 over CBX7, with KED98 showing the best CBX8 ChD selectivity over all other CBX ChDs (Figure SI 2A−C). Unfortunately, a diethyllysine variant of KED97, KED97L, displayed no activity in published CBX8-dependent assays of cell viability,[18] transcription,[30] and chromatin binding (Figure SI 2D−F).[31] Although the diethyllysine substitution has shown to increase cellular permeability without compromising binding affinity for CBX ChDs,[15] this ligand still displays poor cellular permeability (Figure SI 2G) in the chloroalkane penetration assay (CAPA).[32] While this is the first reported CBX8-specific ligand, the low selectivity and cell permeability severely limit its utility.

Second-Generation DNA-Encoded Positional Scanning Library (PSL2). To identify probes of CBX8 ChD with improved affinity, selectivity, and cellular permeability, we designed and synthesized a second-generation positional scanning library (PSL2) around KED98, the most selective CBX8 ligand derived from PSL1. For PSL2, we again varied the four positions to the N-terminal side of the trimethyllysine.

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

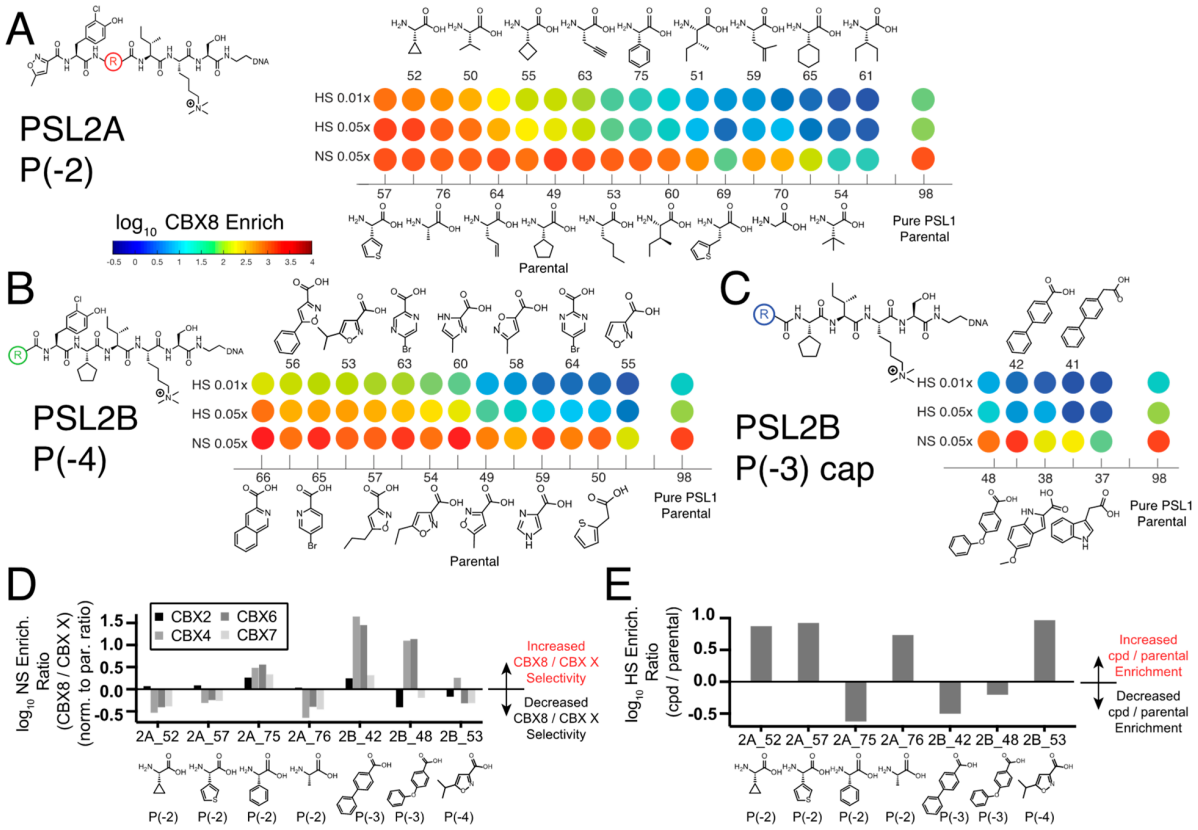


Figure 3. Increased stringency selections of PSL2 against CBX8 ChD. Two on-bead effective protein concentrations (0.01x: ~0.5 μM, 0.05x: ~2.5 μM, calculated based on bead capacity), and two washing stringencies (HS: high stringency, increased washing cycles and time; NS: normal stringency) were applied in the selection of PSL2 against CBX8 ChD. Select enrichment data for library molecules with various building blocks in (A) PSL2A at P(−2), (B) PSL2B at P(−4), and (C) PSL2B with capping at P(−3). A color map indicates the enrichment of selected library members to CBX8 ChD under the indicated conditions (increasing stringency bottom to top). Values are ordered by CBX8 enrichment at the highest stringency in decreasing order left to right. (D) For select compounds, enrichment values from the normal stringency (NS 0.05x) CBX8 selections were divided by the enrichments of four additional ChDs under the same conditions and were normalized using the ratio of enrichments for the parental ligand. (E) For select compounds, enrichment values from the high stringency (HS 0.01x) selections to CBX8 were divided by the enrichment of the parental ligand under the same conditions. Enrichment was calculated as fold change of sequencing reads from selected DNA over sequencing reads of input DNA, normalized to the enrichment of a non-ligand control.

Several monomers were chosen to expand upon the SARs observed with PSL1.[28] In particular, we explored a number of P(−2) side chains with sizes between the valine and the cyclopentyl group to optimize binding in the ChD hydrophobic pocket. Based on the methyl isoxazole hit, a number of 5-substituted isoxazoles were included at the P(−4) cap along with additional heterocyclic aromatics.[17] To facilitate the discovery of cell permeable molecules, we sought to reduce the number of amide bonds and molecular weight through truncation and incorporation of dipeptide mimetics or simple linkers (both rigid and flexible) to bridge side chain binding sites. Therefore, we included a number of monomers that might substitute for two sequential monomers from the parental molecule (either [−1 + −2], [−2 + −3], or [−3 + −4]) (Figure 2; see Figure SI 3 for full monomer set).

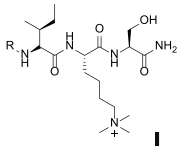
Using 96 unique 140-mer dsDNA constructs, we prepared the 192-membered PSL2 in two sets: PSL2A and PSL2B. PSL2A incorporated 32 synthons each for positions −1 and −2 and 16 synthons each for combined [−1 + −2] and [−2 + −3] positions. Likewise, PSL2B included 32 synthons each for positions −3 and −4, as well as 16 synthons for combined [−3 + −4] positions and 16 amino acid synthons for position −4 together with an acetate cap.

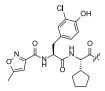
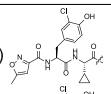
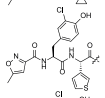
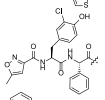
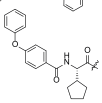
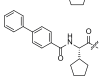
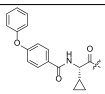
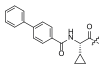
For position −1 (PSL2A_1−32), a wide variety of lipophilic amino acids were enriched for all paralogs, with the exception of non-α-amino acids (e.g., PSL2A_2 and PSL2A_30) and phenylglycine derivatives (e.g., PSL2A_14). In general, there was little indication that modification of the P(−1) residue could increase selectivity of the parental ligand for CBX8, or any paralog, with the notable exception of compound PSL2A_5, where a naphthyl Phe derivative decreased binding to all paralogs except CBX4 (Figures 2B and SI 3). For position −2 (PSL2A_49−80), synthons with small hydrophobic groups were well tolerated by all polycomb (Pc) CBXs, while non-α-amino acids were not tolerated by any CBX paralogs at this position. The tolerance of larger side chains at P(−2) by CBX8 in particular was reiterated, and synthons as large as phenylglycine (PSL2A_75) were tolerated for CBX8. For position −3, Phe derivatives (e.g., PSL2B_10, PSL2B_13, PSL2B_15) similar to parental synthon Cl-Tyrosine were all favored for binding. For position −4, the substituted isoxazole derivatives, as well as particular additional heterocyclic

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

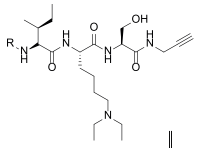
Table 1. IC_{50} Values for Off-DNA Ligands in a Ligand Displacement Fluorescence Polarization Assay: (A) Ligands Containing Trimethyllysine and (B) Ligands Containing Diethyllysine[a]

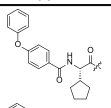
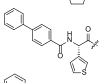
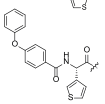
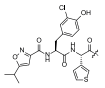
A



Compound	R =	IC_{50} (μM) CBX6	CBX7	CBX8
KED98 (Parental)		34 ± 1	83 ± 2	14 ± 1
SW2_90 (PSL2A_52)		0.36 ± 0.05	0.8 ± 0.1	2.1 ± 0.5
SW2_101E (PSL2A_57)		4.8 ± 0.6	83 ± 1	3.1 ± 0.8
SW2_49B (PSL2A_75)		ND	ND (>100)	44 ± 8
SW2_101B (PSL2B_48)		NB	ND (>100)	12 ± 1
SW2_89 (PSL2B_42)		ND (Aggreg.)	22 ± 1	13 ± 1
SW2_101F		ND (>100)	1.6 ± 0.3	4.0 ± 1
SW2_101A		98 ± 26	35 ± 9	40 ± 8

B



Compound	R =	IC_{50} (μM) CBX6	CBX7	CBX8
SW2_110A		ND (Aggreg.)	ND (Aggreg.)	>7.0 (Aggreg.)
SW2_104B		ND (Aggreg.)	ND (Aggreg.)	>15 (Aggreg.)
SW2_110B		ND (Aggreg.)	ND (Aggreg.)	>5.7 (Aggreg.)
SW2_104A		2.7 ± 0.7	4.8 ± 0.9	1.6 ± 0.6

[a]IC_{50} values for each ligand against CBX6, CBX7, and CBX8 ChDs were measured using 100 nM fluorescent probe and 1 μM CBX6, 0.4 μM CBX7, and 4 μM CBX8 ChDs. Reported values are the average of quadruplicates ± s.d. NB: No binding observed. ND: Value not determined. In some cases, full curves could not be determined due to IC_{50} > 100 μM or compound aggregation (Aggreg.). Full binding curves are shown in Figures SI 6 and SI 4.

aromatics were favored or well tolerated by the polycomb CBX paralogs (Figures 2C and SI 3).

Monomers included to substitute for both P[−1 + −2] positions (compounds PSL2A_33−48) were not tolerated, with the exception of γ-aminobutyric acid (PSL2A_35), which could not be confirmed in off-DNA follow-up studies (Figure SI 4B). Similarly, monomers intended to substitute for both P[−2 + −3] residues (PSL2A_81−96) were not tolerated. Gratifyingly, two monomers among those included to substitute for both P[−3 + −4] monomers (PSL2B_33−48) were tolerated without a large loss in CBX8 binding. Specifically, ligands with a biphenylcarboxylic acid (PSL2B_42) and phenoxybenzoic acid (PSL2B_48) acyl caps at the [−3 + −4] position demonstrated retained affinity and improved selectivity toward CBX8 (Figures 2C and SI 3).

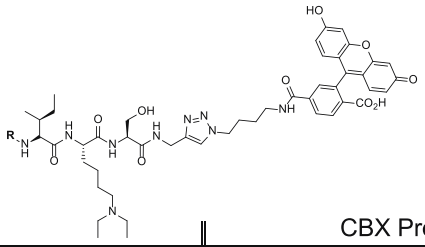
Several trends in the PSL2 data sets suggested that selection conditions were not sufficiently stringent to yield differential enrichments among high affinity ligands. In prior work with PSL1-derived compounds KED97 and KED98, we observed greater enrichments in the PSL1 selections and greater affinity in displacement assays of valine over cyclopentyl glycine at P(−2) for all Pc ChDs (Figures SI 1 and SI 2), yet enrichments observed for PSL2A_49 and PSL2A_50 were similar. Additionally, the majority of the acyl cap monomers at P(−4) displayed similarly high levels of enrichment for all the Pc CBX ChDs despite their varied structures (Figure SI 3). In order to more effectively differentiate the top binders, we performed selections against CBX8 a second time using more stringent conditions.

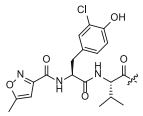
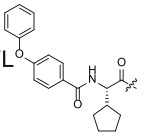
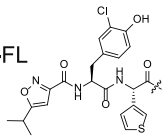
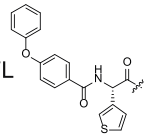
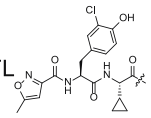
Optimized High-Stringency (HS) Selections against CBX8 ChD. To increase the stringency of the affinity selection assay, we increased the number and time of bead wash cycles, and further reduced on-bead protein concentration (Figure 3). Under these conditions, valine at P(−2) (PSL2A_50) now showed higher enrichment than cyclopentylglycine (PSL2A_49), consistent with prior PSL1 selections and off-DNA competitive FP assays (Figures SI 1 and SI 2A). Overall, few substitutions to the parental ligand at either the P(−1) or P(−3) positions suggested that significant gains in affinity could be achieved (Figure SI 5). Several substitutions at the P(−2) and P(−4) positions gave increased enrichment over the parental ligand (Figure 3).

Among additional substitutions at the −2 position, we now observe the highest enrichment of cyclopropylglycine (PSL2A_52), 3-thienylglycine (PSL2A_57), and L-alanine (PSL2A_76) (Figure 3A). Enrichment of other monomers at P(−2) decreased roughly with increasing side chain size. Similar enrichment was observed for cyclobutaneacetic acid

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

Table 2. K_d Values of Fluorescein-Conjugated Ligands to PcG CBX ChDs[a]



Compound	R =	CBX Protein Chromodomain (K_d (μM)) CBX2	CBX4	CBX6	CBX7	CBX8
KED97L-FL		1.1 ± 0.4	2.7 ± 0.4	0.3 ± 0.1	4.7 ± 0.9	0.24 ± 0.03
SW2_110A-FL		4.6 ± 0.9	NB	NB	>16 ± 5	0.8 ± 0.2 (MST: 0.7 ± 0.1) (TSA*: 0.5 ± 0.3)
SW2_104A-FL		0.05 ± 0.04	0.17 ± 0.04	0.014 ± 0.004	0.8 ± 0.4	0.0029 ± 0.0008 (MST: 0.005 ± 0.001)
SW2_110B-FL		ND	NB	NB	NB	ND (MST: 0.8 ± 0.1)
SW2_90L-FL		0.79 ± 0.01	0.17 ± 0.01	0.030 ± 0.001	0.140 ± 0.008	0.110 ± 0.008

[a]Values were determined by direct fluorescence polarization and are displayed as the average of $n = 4 \pm$ s.d. NB: No binding. ND: Not determined due to an inability to acquire full binding curves due to low affinity or compound aggregation. MST: MicroScale Thermophoresis. TSA: Thermal Shift Assay. *The non-derivatized ligand was used for K_d determination by TSA. Full binding curves are provided in Figure SI 7 for FP data, Figure SI 8 for MST data, and Figure SI 9 for TSA data.

(PSL2A_55), propargylglycine (PSL2A_63), and allylglycine (PSL2A_64), and decreased enrichment was found for allo-isoleucine (PSL2A_51), isoleucine (PSL2A_60), norleucine (PSL2A_53), and phenylglycine (PSL2A_75) compared to the parental compound.

As with P(−2), the high-stringency (HS) selection results for modification at P(−4) showed greater differentiation of high affinity ligands (Figure 3B). Several 3-substituted isoxazole carboxylic acids with various substituents (PSL2B_53, 54, 56, 57) gave improved enrichment with enrichment increasing roughly with the size of the substituent. In agreement with this SAR was the low enrichment of the desmethyl isoxazole (PSL2B_55). In addition, 5-bromo-2-pyrazinecarboxylic acid (PSL2B_63), 5-bromopyridine-2-carboxylic acid (PSL2B_65), and isoquinoline-3-carboxylic acid (PSL2B_66) gave higher enrichment to CBX8 than the parental ligand, while 2-thiopheneacetic acid (PSL2B_50) and 1*H*-imidazole-4-carboxylic acid (PSL2B_59) showed lower enrichment. These results were consistent with both published and PSL1 SAR of ligands to CBX8 (Figure SI 1), which have shown increased affinity for benzoyl caps with lipophilic para substituents.[28]

For the single building blocks intended to substitute for both P[−3 + −4] monomers in the parental ligand (PSL2B_33−48), the increased stringency selections showed a marked decrease in enrichment (Figure 3C). This was particularly the case for the ligands with phenoxybenzoic acid (PSL2B_48) or biphenylcarboxylic acid (PSL2B_42) acyl caps at the [−3 + −4] position, which showed high enrichment under the normal stringency selection conditions.

Selection, Validation, and Structural Optimization of Off-DNA Ligands. To facilitate decision making for off-DNA follow up, we used results from the normal stringency selections for all Pc ChDs as indicators of selectivity and the high stringency selections to CBX8 as indicators of affinity (Figure 3D,E). To assess improvements in selectivity, we plotted the ratio of CBX8 enrichment to the enrichment for each of the other isoforms and normalized this value to the same ratio observed for each isoform with the parental compound. Compounds with the P(−3) capping with phenoxybenzoic acid (PSL2B_48) or biphenylcarboxylic acid

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

(PSL2B_42) showed the largest gains in CBX8 selectivity. PSL2B_42 showed improved selectivity against all isoforms, and PSL2B_48 showed improvements in selectivity over CBX4 and CBX6. Compound PSL2A_75 with phenylglycine at P(−2) showed more modest selectivity gains but for all isoforms.

Among the P(−2) substitutions that indicated increased affinity, all showed lower selectivity for CBX 4, 6, and 7. These decreases roughly tracked with decreasing size of the side chain with P(−2) alanine (PSL2A_76) demonstrating the largest decrease in selectivity. For selectivity over CBX4 and CBX7, this observation is consistent with the structural differences in the binding pocket for this side chain.[23] Also, prior work has shown high affinity binding of ligands with L-alanine at this position for all Pc CBX ChDs.[23,28] For the P(−4) substitutions that showed high enrichment (3-isopropyl isoxazole was selected as an example), only modest, if any, changes in selectivity were observed.

To facilitate comparison of affinity across the two library selections, which were conducted separately, we plotted the ratio of the enrichment of a compound to the enrichment of the parental ligand under the highest stringency conditions (Figure 3E). For the molecules that indicated improved selectivity, decreased enrichment relative to the parental compound was observed at varying levels, with the greatest decrease due to the P(−2) phenylglycine (PSL2A_75). Similar ratios were observed for the improved affinity P(−2) substitutions, as well as the exemplary isopropyl isoxazole.

Based on this analysis, we selected five molecules for off-DNA synthesis on solid phase for subsequent determination of IC_{50} values using a competition fluorescence polarization (FP) assay against CBX6, CBX7, and CBX8 (Table 1).[23] Consistent with sequencing results, both cyclopropylglycine (SW2_90) and 3-thienylglycine (SW2_101E) substitutions at P(−2) have greater affinity to CBX8 than the cyclopentylglycine parental compound (Table 1A). Selectivity for CBX8 with these compounds suffers, particularly against CBX6, however. The larger phenylglycine at this position was tolerated by CBX8 and not CBX7, but this compound (SW2_49B) shows a large decrease in affinity. For the P(−3) capped compounds PSL2B_48 and PSL2B_42 with high selectivity in the sequencing data, this selectivity was confirmed with the off-DNA compounds. The phenoxyphenyl compound SW2_101B shows very high selectivity over both CBX6 and CBX7. The biphenyl compound SW2_89 shows high selectivity over CBX6. For both of these compounds, affinity to CBX8 is largely unchanged from the parental compound.

In addition, we combined monomers that were able to substitute for both the −3 and −4 positions (either the phenoxybenzyl or biphenyl) together with the high affinity monomers identified for the −2 position (cyclopropyl, thienyl) (Table 1). SW2_101F (P(−2) cyclopropyl, P(−3) phenoxylphenyl) had increased affinity for CBX8, but gave decreased selectivity for CBX8 over CBX7. Interestingly, the biphenyl cap gave significantly reduced affinity for both CBX8 and CBX7 when paired with the P(−2) cyclopropyl glycine (SW2_101A), while in the context of the P(−2) cyclopentyl glycine (SW2_89), there was little change in affinity. For diethyl compounds SW2_110A, SW2_110B, and SW2_104B, full IC_{50} curves could not be obtained due to aggregation issues with the fluorescein-containing probe at high compound concentrations. Lastly, we also combined the 5-isopropyl-3-isoxazole at P(−4) with the 3-thienylglycine at P(−2) in compound SW2_104A. Tight binding to CBX8 ChD was observed, albeit with low selectivity.

To measure K_d values, fluorescence polarization assays of fluorescein conjugates were conducted (Table 2) with titration of each of the Pc ChDs. The K_d of SW2_110A-FL for the CBX8 ChD was determined to be ∼800 nM, which was similar to the ∼700 nM obtained by microscale thermophoresis (MST) (Figure SI 8) and the ∼500 nM value obtained with underivatized SW2_110A in a thermal shift assay (Figure SI 9). The affinity of SW2_110A-FL to CBX8 is decreased slightly compared to that of KED97L-FL; however, SW2_110A-FL displayed dramatic improvements in selectivity. This compound is completely selective for CBX8 over CBX4 and CBX6, while maintaining 20-fold selectivity over CBX7 and 5-fold selectivity over CBX2. In contrast, the selectivity of SW2_104A for CBX8 ChD was only modestly improved over KED97L while the affinity to all paralogs was increased significantly. The K_d of SW2_104A is 2.9 nM for CBX8 (5.4 nM by MST, Figure SI 8), making it the tightest binding CBX ChD ligand reported to date. Intriguingly, combining the phenoxyphenyl cap with the thienylglycine at P(−2) in compound SW2_110B-FL did not increase binding compared to SW2_110A-FL with cyclopentyl at P(−2), which differs from the difference observed within the 6-mer ligand context (Table 1), suggesting potential crosstalk between the P(−2) pocket and the biphenyl binding region.

We selected the phenoxyphenyl compound SW2_110A for follow up studies. While the fluorescein-conjugate of this compound displayed modest affinity for CBX8 (800 nM K_d), the selectivity profile was far more favorable compared to the high affinity SW2_104A (3 nM K_d). In addition, SW2_110A includes one less amino acid, which will improve the physiochemical properties of this compound as a probe. As SW2_110A, in particular, demonstrated some aggregation effects in the displacement FP assays (Table 1), we tested the solubility of both SW2_110A and SW2_104A in a shake-flask equilibrium assay.[33] With 24 h incubation in PBS, SW2_110A showed solubility at 130 μM, and SW2_104A was soluble to 41 μM.

Structural Basis of SW2_110A Association with CBX8 ChD. To investigate the structural basis of inhibitor binding, we utilized NMR spectroscopy. We collected a series of $^{1}H,^{15}N$ heteronuclear single quantum coherence (^{15}N-HSQC) spectra on ^{15}N-labeled CBX8 ChD upon the titration of SW2_110A. Addition of the inhibitor led to substantial changes in the CBX8 ChD spectrum, including chemical shift perturbations (CSPs) and disappearance of resonances, indicating binding (Figure 4A). Mapping these CSPs onto the solved structure of the ChD-H3K9me3 complex reveals a cluster of residues with significant CSPs cluster in and around the canonical histone binding pocket, indicating that the inhibitor can directly compete with histone tail binding (Figure 4B,C), which was confirmed using a competitive fluorescence polarization of SW_110A_L-FL with H3K27me3 peptides (Figure SI 10A). Note that a ChD-H3K27me3 structure is currently unavailable, but that we have previously demonstrated that H3K9me3 and H3K27me3 bind in the same pocket.[34] Importantly, the limited solubility of the inhibitor at high concentrations required introduction of DMSO. A control titration with DMSO only showed minimal CSPs, suggesting that DMSO alone does not significantly alter the ChD structure (Figure SI 10B).

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

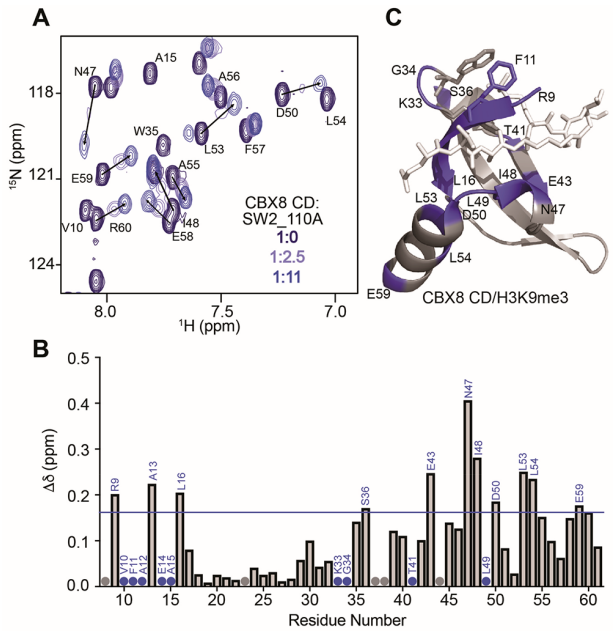
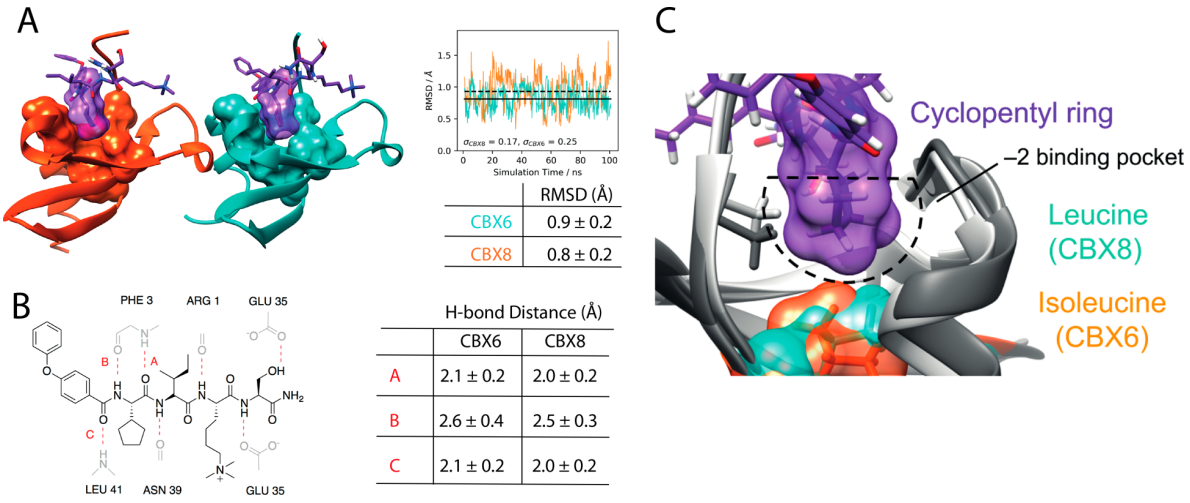


Figure 4. Structural basis of SW2_110A binding. (A) Overlay of ^{15}N-HSQC spectra of CBX8 CD upon addition of increasing concentration of SW2_110A. Molar ratios are color coded as indicated in the legend inset. (B) Chemical shift perturbations (CSPs) as a function of CBX8 CD residue. Residues for which resonances disappear upon binding are denoted by a blue sphere, and residues missing resonances entirely are denoted with gray spheres. Perturbations were considered significant if they were greater than the average plus one standard deviation (denoted by the blue line), not including the highest 10% of CSP values. (C) Residues with significant CSPs or for which resonances broadened to the point of disappearing upon binding are colored blue on a cartoon representation of CBX8 in complex with H3K9me3 (PDB ID 3I91). The peptide is shown as white sticks, and the aromatic cage residues important for coordinating the methyllysine are shown as gray sticks.

We have previously determined the structural basis of H3K27me3 and H3K9me3 binding using NMR spectroscopy.[34] Comparing the CSPs induced upon addition of SW2_110A to those of an H3K27me3 peptide (residues 23−34) reveals the largest differences are in resonances corresponding to E43, N47, and I48, as well as D50, L53, and L54 (Figure SI 10C). Both subsets of resonances are significantly more perturbed upon inhibitor binding as compared to H3K27me3 and were not significantly perturbed upon addition of DMSO alone (Figure SI 10B,C). Residues E43, N47, and I48 are in the expected location of the P(−1) and P(−2) side chains, which are Arg(P−1) and Ala(P−2) in the H3K27me3 peptide. In addition, residues D50, L53, and L54 lie where the phenoxyphenyl group is expected to bind. Notably, these residues are in the hydrophobic pocket of the CBX8 ChD, which is a key determinant for CBX8 ChD specificity for H3K27me3. A number of resonances (corresponding to V10, F11, A12, E14, A15, K33, G34, T41, and L49) disappear upon binding (Figure 4B). These residues lie in the β1 strand and the β1−β2 loop, which contain the aromatic cage residues. Based on the crystal structures of the apo CBX8 ChD and the ChD in complex with H3K9me3, the N-terminal portion of the β1-strand is stabilized upon histone tail binding.[25] The disappearance of these peaks upon addition of SW2_110A suggests that, in contrast, inhibitor binding may not fully stabilize this region instead leading to conformational exchange in the bound state on the intermediate NMR time scale. Together, the NMR analysis reveals the structural basis of SW2_110A binding and suggests the determinants of histone binding inhibition.

Interestingly, although the inhibitor is selective for CBX8, it associates with regions of the ChD that are conserved between homologs. Indeed, most of the residues that differ between the ChD homologs do not exhibit CSPs upon SW2_110A binding with two exceptions: A15, which is a Ser in CBX6 and CBX7, and S36, which is an Ala in CBX6 and a Pro in CBX7. Neither residue, however, is expected to make direct contact with the inhibitor. To gain further insight into the potential source of



	RMSD (Å)
CBX6	0.9 ± 0.2
CBX8	0.8 ± 0.2

H-bond Distance (Å)

	CBX6	CBX8
A	2.1 ± 0.2	2.0 ± 0.2
B	2.6 ± 0.4	2.5 ± 0.3
C	2.1 ± 0.2	2.0 ± 0.2

Figure 5. Molecular dynamics simulations of inhibitor (SW2_101B) binding to CBX8 and CBX6 ChDs. (A) Root mean square deviation (RMSD) plots of surface area portrayed residues at bottom of −2 binding pocket in CBX6 (orange) and CBX8 (teal). RMSDs were taken relative to highest clustered pose taken from the 100 ns trajectories. (B) Hydrogen-bonding map for CBX8. H-bonds labeled to their respective panel. A−C hydrogen bonding distances contributed by cyclopentyl alanine and the N-terminus diphenyl ether residues on CBX6 and CBX8. (C) The cyclopentyl ring in the −2 pocket. The differing residue at the bottom of the −2 pocket (leucine (CBX8), isoleucine (CBX6)) gives a slight change in pocket shape.

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

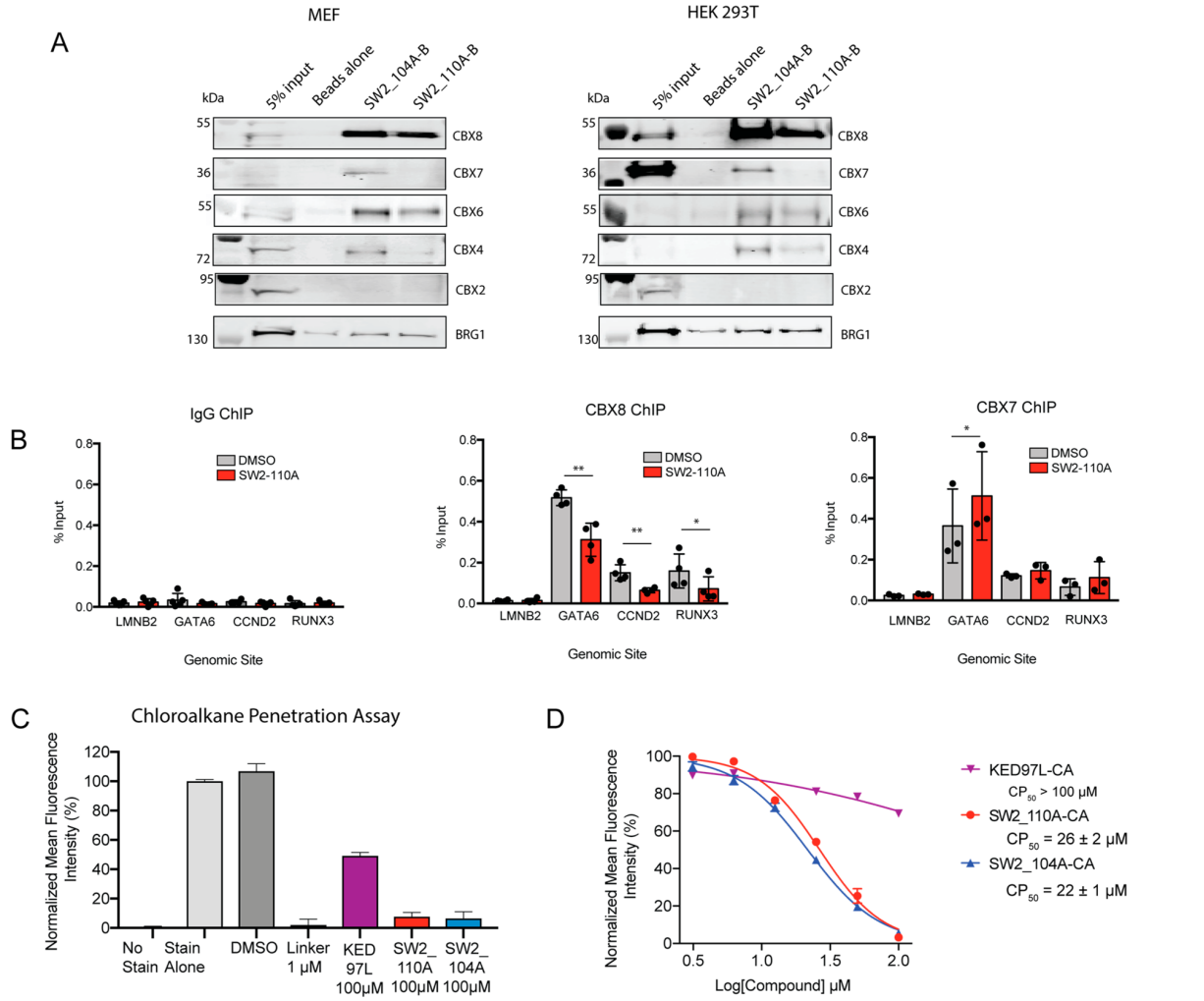


Figure 6. Cellular activity of CBX8 ChD ligands. (A) Chemoprecipitations from Mouse Embryonic Fibroblast (MEF) (left) and HEK293T (right) nuclear lysates using biotin-labeled SW2_104A (SW2_104A-B) and SW2_110A (SW2_110A-B) were analyzed using immunoblot analysis. (B) Chromatin immunoprecipitation (ChIP) followed by quantitative PCR of genomic regions with CBX8 and CBX7 binding in Hs68 fibroblast cell line. ChIP-qPCR was used to evaluate the ability of SW2_110A to disrupt endogenous CBX protein associations with chromatin in cells. Cells were treated with 100 μM SW2_110A for 4 h prior to harvest. ChIP-qPCR of CBX7 and CBX8 at *LMNB2* (negative locus), *RUNX3*, *GATA6*, and *CCND2*. For all qPCR, error bars represent SEM $n = 3$ biological replicates; p-values were calculated using two-tailed Student's t test, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$. (C) Relative cytosolic access of chloroalkane-modified ligands, SW2_110A-CA, SW2_104A-CA and KED97L-CA, was evaluated by chloroalkane penetration assay (CAPA). (D) Dose-dependent cytosolic access of SW2_110A-CA, SW2_104A-CA, and KED97L-CA was assessed by pre-incubation of designated concentrations of CA-molecules, followed by HT-TAMRA dye. CP_{50} of CA-molecules were evaluated and averaged from three independent curve fits.

selectivity, we investigated binding to two compounds, which demonstrated low selectivity between homologs. Titration of KED97L or KED98L led to very similar subsets of CSPs as compared to SW2_110A, revealing a similar binding pocket (Figure SI 10D). The non-selective KED97L and KED98L, however, led to far fewer disappearance of resonances as compared to SW2_110A, suggesting that both KED97L and KED98L lead to greater stabilization of the CBX8 ChD. This is especially evident in the β1 strand. Together, these data suggest that rather than differences in the manner in which each CD directly coordinates the inhibitor, selectivity likely arises from a difference in the accessible conformational ensemble available to each CBX ChD. This could be modulated by small differences in the ChD sequence ultimately leading to differences in the size and nature of chemical groups that can be accommodated.

Molecular Dynamics Simulations of Ligand Association with CBX8 ChD and CBX6 ChD. To gain insight into the ligand selectivity for CBX8 over CBX6, we carried out MD simulations of CBX8 ChD and CBX6 ChD with the trimethyllysine version of SW2_110A (SW2_101B). Due to the increased selectivity of cyclopentyl glycine at −2 position for CBX8 over CBX6 in off-DNA validations (Figure SI 2A), we first investigated the interaction of cyclopentyl ring in the −2 position. The closest residue difference within this region is a leucine (CBX8)/isoleucine (CBX6) at the bottom of the −2 binding pocket. We evaluated RMSD traces (Figure 5A) for two sets of five residues at the bottom of the −2 binding pocket in CBX6 (orange) and CBX8 (teal) (residues V30, A13, L53, L16, I48 for CBX8 and V30, A13, L53, I16, I48 for CBX6). This area is directly involved in enclosing the cyclopentyl ring within the pocket (Figure 5C). The RMSDs

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

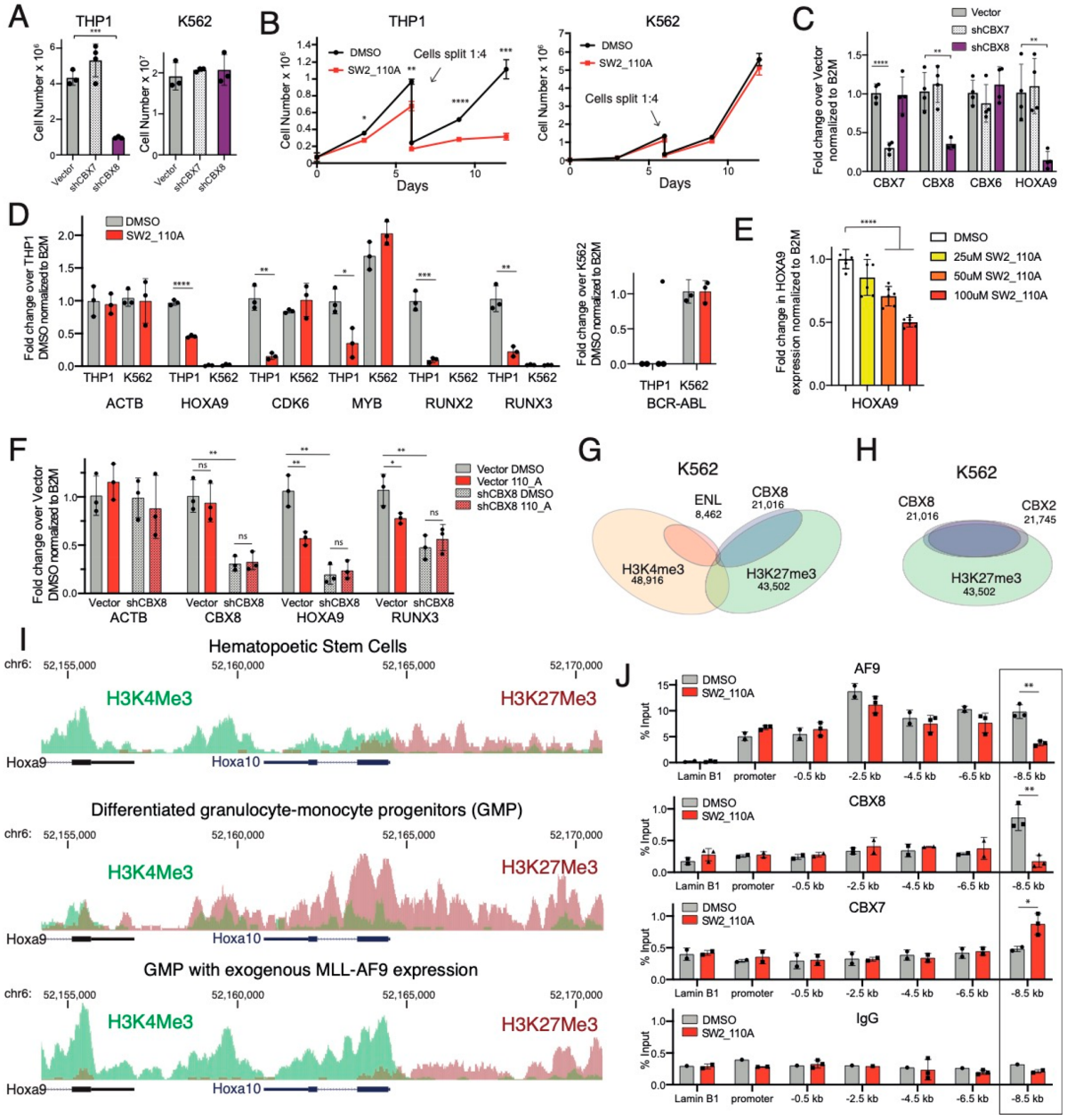


Figure 7. Dependency on CBX8 for MLL-AF9-mediated cell proliferation and gene transcription. (A) Cell viability of THP1 and K562 cells upon *CBX8* and *CBX7* knockdown. (B) SW2_110A treatment (100 μM) inhibits proliferation of THP1 cells but not K562 cells. Cells were split at 1:4 on day 6 to avoid overconfluency. s.d. is represented by error bars (n = 3, three biological replicates with three technical replicates for each biological replicate). (C) qRT-PCR analysis of *CBX8*, *CBX7*, *CBX6*, and *HOXA9* gene expression in THP1 cells with knockdown of CBX8 and CBX7. (D) qRT-PCR analysis of gene expression for MLL-AF9 target genes (*HOXA9*, *CDK6*, *MYB*, *RUNX2*, and *RUNX3*) and controls (*ACTB* and *BCR-ABL*) in THP1 and K562 cells after 48 h of 100 μM SW2_110A treatment. (E) qRT-PCR analysis of *HOXA9* gene expression in THP1 cells after 48 h of SW2_110A treatment at the indicated dose. (F) qRT-PCR analysis of MLL-AF9 target gene expression in THP1 cells treated with SW2_110A (100 μM for 24 h) after *CBX8* knockdown. (G) CBX8 and AF9 genome-wide localization analysis using publicly available ChIP-Seq data sets from K562 cells.[51] (H) Genome-wide analysis of CBX8, CBX2 and H3K27me3 peak overlaps in K562 cells using published ChIP-Seq data sets.[51] Numbers indicate number of called peaks, overlap percent of total CBX8 peaks. (I) Previously published ChIP-Seq tracks of H3K4me3 and H3K27me3 enrichment at HOXA locus in hematopoetic stem cell (HSC) (top) in HSCs that have undergone differentiation to granulocyte monocyte progenitors (GMP) (middle) and in GMP cells with exogenous MLL-AF9 expression (bottom). Data obtained from Bernt *et al.*[52] and visualized using UCSC genome browser. (J) ChIP-qPCR analysis of AF9, CBX8, and CBX7 enrichment at sites upstream of the *HOXA9* transcription start site. Inhibitor treatment was 100 μM for 24 h. Error bars represent s.e.m. (n = 3); p-values were calculated using Student's two-tailed t test: * = $p < 0.05$, ** = $p < 0.01$,*** = $p < 0.001$, **** = $p < 0.0001$.

presented in Figure 5A indicate persistent flexibility of the CBX6-ligand system, and a relative binding stability for CBX8-ligand complex. The increased flexibility in the CBX6 −2 pocket leads to higher cyclopentyl mobility and a coupled

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

disruption of the hydrogen bonding network (illustrated by larger variance of the hydrogen bond distances presented in Figure 5B).

The diaryl ether group determined experimentally to promote SW2_110A selectivity for CBX8 does not appear to preferentially interact with either isoform in the MD simulation data. The terminal phenyl ether group oscillates back and forth to either side of the beta groove for both CBX8 and CBX6, as described *via* the calculated distances of the terminal phenyl ring to the side of beta-groove region (Figure SI 11D). The intermediate phenyl ring, however, adopts distinct orientations emerging from under the clasp; this likely carries a steric penalty, as well as further contribution to the destabilization of the hydrogen bond network discussed above. Despite the instability below the clasp, the hydrophobic clasps at the −1 position for both isoforms are closed around the ligand in both simulations. The binding geometries in this region are highly similar for CBX6 and CBX8, as shown in the distances between residues V10 and L49 (Figure SI 11B).

Overall, the MD data present a complex binding motif for SW2_110A to CBX6 and CBX8, which is complicated by the collective motion of the two isoforms. However, decreased stability of the binding configuration around the −2 pocket (including the collective effects of binding of the cyclopentyl ring, the hydrogen bonding network and steric effects in the accommodation of the diphenyl ether group) appears to disfavor CBX6 complexation of the SW2_110A ligand.

Cellular Selectivity Studies. *Chemoprecipitations.* The CBX ChDs demonstrate significant structural flexibility *in vitro*, making it necessary to determine whether recombinant CBX ChDs accurately recapitulate binding properties of ChDs found within the context of fully formed PRC1. Therefore, we utilized biotinylated derivative SW2_104A-B and SW2_110A-B to enrich CBX8 and its paralogs from mouse embryonic fibroblast (MEF) lysates (Figure 6A, left) and HEK293T lysates (Figure 6A, right) using affinity purification. Both SW2_104A-B and SW2_110A-B robustly enrich CBX8 in support of the *in vitro* results. Further, enrichment of CBX4, CBX6 and CBX7 is higher for SW2_104A-B than SW110A-B, supporting the *in vitro* FP assay results with recombinant CBX ChDs that suggest better CBX8 selectivity for SW2_110A. In contrast to the results from both *in vitro* assays, however, we did observe some enrichment of CBX6 with SW2_110A-B. Considering the high homology between these two ChDs (86% sequence identity), this is not surprising, but indicates that *in vitro* properties for the ChDs may not completely be recapitulated in a cellular environment.

ChIP-qPCR. In order to evaluate the ability of SW2_110A to disrupt CBX8 association with chromatin, we used chromatin immunoprecipitation (ChIP) followed by quantitative PCR (ChIP-qPCR) (Figure 6B). Using ChIP-Seq data sets for CBX6, CBX7 and CBX8 in Hs68 fibroblast cells,[31] we selected representative target loci with detectable enrichment of CBX7 and CBX8 using ChIP-qPCR.[34] Upon incubation of cells with SW2_110A, we observe significant reduction of CBX8 binding at these sites, while CBX7 binding was unaffected or even increased upon treatment with SW2_110A, a phenomenon similarly observed with CBX8 knockdown (Figure SI 12B). Since CBX paralogs primarily localize together at H3K27me3 sites,[8,31] it is possible that selective reduction of CBX8 at a genomic locus could increase the enrichment of other paralogs at the same site. To confirm that SW2_110A can disrupt global CBX8 chromatin binding, we also performed sequential salt extraction (SSE), which can determine the impact of chromodomain inhibition on the bulk chromatin binding affinity of CBX paralogs.[34−36] We confirmed that SW2_110A abrogates CBX8, but not CBX7, binding to bulk chromatin (Figure SI 12B).

Evaluation of Cytosolic Access Using Chloroalkane Penetration Assay (CAPA). The ChIP-qPCR results imply that SW2_110A has increased cell permeability compared to KED97L. To confirm this difference, we used CAPA, a cell penetration assay recently developed by the Kritzer lab.[32] CAPA is easy to perform, quantitative, and measures compound availability in the cytosol without interference from molecules trapped in endosomes. CAPA utilizes a HeLa cell line stably transfected with a cytosolic HaloTag protein[37] in a pulse-chase experiment. Cells are incubated with ligands conjugated to a chloroalkane (CA) (pulse), which will covalently react with the HaloTag protein when/if the ligand reaches the cytoplasm. The cells are then treated with chloroalkane-TAMRA dye (chase), which reacts with any remaining, unblocked HaloTag. The red fluorescence is quantified using flow cytometry, which is inversely proportional to the CA-molecule cytosolic concentration. SW2_110A and SW2_104A were conjugated to the chloroalkane (denoted SW2_110A-CA and SW2_104A-CA) and compared to KED97L-CA. A significant increase in permeability was observed for both SW2_110A-CA and SW2_104A-CA with CP_{50} values of 26 ± 2 μM and 22 ± 1 μM respectively, compared to CP_{50} > 100 μM for KED97L-CA (Figure 6C,D).

Cellular Activity in MLL-AF9 Transformed Leukemia. We next wished to evaluate the activity of SW2_110A in a CBX8-dependent cell line. Previous studies identified that CBX8 is required for leukemogenesis in an MLL-AF9 mouse model of leukemia, as well as for the viability of human leukemia cell lines with MLL-AF9 translocations.[18] Supporting a requirement for CBX8 in the viability of MLL-AF9 transformed cell lines, we were unable to isolate a CBX8 knockout in the THP1 leukemia cell line using CRISPR-Cas9. As an alternative approach, we used lentiviral-mediated shRNA knockdown of *CBX8* and *CBX7* and confirmed that CBX8, but not CBX7, is required for maintaining the growth of THP1 cells (Figure 7A). In contrast, the viability of K562 leukemia cells driven by a BCR-ABL translocation[38] was not affected by either *CBX7* or *CBX8* knockdown. To determine whether CBX8 ChD inhibitors can inhibit MLL-AF9 mediated oncogenesis, we measured the viability of THP1 (MLL-AF9 translocation) and K562 (BCR-ABL translocation) leukemia cells cultured with SW2_110A. We observed a significant decrease in the proliferation of THP1 cells starting at 3 days, while the growth of control leukemia cell line K562 was not affected (Figure 7B). Similar proliferation effects were observed with SW2_104A (Figure SI 13A) and the IC_{50} of the two inhibitors after 12 days of THP1 treatment were similar with IC_{50} = 26 μM for SW2_110A, and IC_{50} = 25 μM for SW2_104A (Figure SI 13B).

Gene Expression Analysis. The MLL-AF9 translocation is a fusion of the N-terminus of the MLL H3K4 histone methyltransferase with the C terminus of AF9, a subunit of the Super Elongation Complex (SEC). The MLL portion is missing methyltransferase activity and is primarily responsible for targeting, while the AF9 portion drives transcriptional activation.[39−41] Previous studies identified a paralog-specific interaction between CBX8 and AF9 (or its paralog ENL), which is conserved in the MLL-AF9 (or MLL-ENL)

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

translocation product.[18,42,43] MLL-AF9 mediated oncogenesis is mediated primarily through aberrant gene activation,[44−46] most notably the activation of *HOXA9*, which drives "de-differentiation" to a hematopoetic stem cell-like state.[47,48] Previous work defined CBX8 as required for *HOXA9* activation by MLL-AF9,[18] which we confirmed using qRT-PCR analysis of *HOXA9* gene expression in THP1 cells with knockdown of CBX8 (Figure 7C).

To define whether CBX8 ChD inhibition results in a decrease in the transcription of MLL-AF9 target genes, we performed qRT-PCR for *HOXA9*, as well as *CDK6*, *MYB*, *RUNX2*, and *RUNX3*, which have been defined as MLL-AF9 transcriptional targets in THP1 cells.[49] All MLL-AF9 targets genes displayed significantly decreased expression after 48 h of SW2_110A treatment, while non-MLL-AF9 target genes *B2M* and *ACTB* were unchanged in the presence of compound. Neither these gene targets, nor *BCR-ABL*, were altered upon compound treatment in the K562 cell line, consistent with their different oncogenic drivers (Figure 7D). To demonstrate the on-target activity of the inhibitor, SW2_110A treatment displayed significantly decreased *HOXA9* expression with increased inhibitor concentration (Figure 7E). To confirm that the expression changes for MLL-AF9 targets observed with compound treatment are in fact mediated through CBX8, we further analyzed gene expression of MLL-AF9 target genes in THP1 cells treated with SW2_110A after *CBX8* knockdown. We found that while both compound treatment and *CBX8* knockdown reduced the expression of MLL-AF9 target gene expression, there was no additive effect, which supports that compound effects on gene expression are mediated through CBX8 (Figure 7F).

AF9 and CBX8 Genome-Wide Localization. The specific association between CBX8 and the AF9/ENL proteins is well characterized;[18,42,43] however, it is still unclear how CBX8, a known transcriptional repressor,[50] mediates AF9-mediated gene activation. Additionally confounding is how the ChD of CBX8 specifically is involved in gene activation, as the ChD's established role is binding H3K27me3, a mark associated with gene repression.[6] Using publicly available ChIP-Seq data sets from K562 cells,[51] we confirmed that CBX8 is almost exclusively localized at sites with H3K27me3, and that ENL is almost exclusively localized at sites with H3K4me3, a mark associated with gene activation (Figure 7G). Further, ENL binding displays almost no overlap with sites of H3K27me3 or CBX8 binding, providing evidence that CBX8 association does not recruit ENL to CBX8-bound sites. In addition, CBX8 binding almost completely overlaps with that of CBX2, a paralog that does not associate with AF9/ENL, providing evidence that AF9/ENL association conversely does not recruit CBX8 to ENL/AF9-bound sites (Figure 7H).

CBX8 Binding at the HOXA9 Locus. During hematopoetic stem cell (HSC) differentiation to granulocyte monocyte progenitors (GMPs), *HOXA9* expression decreases and the repressive H3K27me3 mark spreads toward the *HOXA9* coding region (Figure 7I, visualization of data from Bernt *et al.*[52]). Transformation of GMPs using exogenous MLL-AF9 re-establishes H3K4me3 and pushes the H3K27me3 boundary back upstream to allow for *HOXA9* reactivation (Figure 7I, bottom). Using ChIP-qPCR in THP1 cells, we confirmed robust enrichment of DNA across the HOXA9 locus with AF9 IP, but with the CBX8 IP, no significant enrichment of DNA was observed until ∼8.5 kb upstream of the HOXA9 start site where the H3K27me3 boundary begins. This enrichment is reduced upon 24 h treatment with SW2_110A, which also reduces AF9 enrichment at the boundary region (Figure 7J). Previous studies have identified an antagonistic relationship for CBX8 and AF9/ENL where overexpression of CBX8 represses AF9 target gene expression,[30] and overexpression of ENL activates CBX8 repressed genes.[53] This supports a model by which the direct association between AF9 and CBX8 allows for MLL-AF9 to antagonize CBX8 and prevent repressive chromatin from spreading into the *HOXA9* locus. Release of CBX8 from the boundary region using SW2_110A likely allows other CBX paralogs to bind, which we observed using ChIP-qPCR of CBX7 (Figure 7J). CBX7 cannot be antagonized by MLL-AF9 so the increase in CBX7 binding can facilitate repressive memory and further deposition of H3K27me3, a function observed for CBX7 in other cell types.[54]

CONCLUSIONS

Using two generations of directed DNA-encoded chemical libraries, we identified selective, cell-permeable, peptidomimetic ligands for the CBX8 ChD. While increased potency and partial selectivity for CBX8 were achieved after the selection in PSL1, a subsequent library led to compounds with further increases in potency, cell permeability, and selectivity. While DELs are often used for hit generation to initiate traditional optimization efforts, this work highlights the ability to use DNA-encoded chemistry within the design-make-test-analyze cycles (DMTA cycles) of medicinal chemistry.[55] Compared with traditional synthesis, purification, and discrete screening of individual molecules, we found this DNA-encoded approach to be lower cost and less labor-intensive for the identification and optimization of ligands. These benefits largely arise from the nature of the *in vitro* selection assay. The ease of this assay allowed concurrent optimization of affinity and selectivity against multiple protein targets. In this assay, it is the concentration of the protein target (not the synthetic ligand) that drives the binding event. Thus, the concentration of the DNA-encoded molecules is insignificant and can be very low, which allows synthesis on a very small scale, permitting the incorporation of monomers that would be too expensive to include using traditional approaches. Several studies have shown how enrichment values from selection assays can correlate to ligand affinity.[28,56−58] An additional advantage is the low requirement for purity of the synthetic ligands using this approach. While ligand purity can complicate the relationship between the observed enrichment of a molecule and its affinity to the protein target, this can be addressed by performing selections at multiple concentrations, as performed here.[59] Likewise, the scale of this approach requires only small amounts of the protein target for the selection assays.

Using this methodology, we have identified the tightest binding and most selective ligands to date for the CBX8 ChD, indicating that new combinations of monomers can improve the affinity of ligands for this target class. Of particular note, we found that a truncated scaffold lacking the position −4 monomer retains affinity to CBX8 and displays improved selectivity. SW2_110A, a ligand containing the truncated phenoxyphenyl cap, is the first CBX ChD ligand to demonstrate complete selectivity over certain isoforms; however, the structural basis for this selectivity is not straightforward, as indicated by both NMR and MD analysis. These structural analyses suggest that unique conformations of the ChDs that are accessible for ligand binding are largely

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

responsible for selectivity, highlighting the difficulty in applying structure-based design for optimizing ligands to these highly dynamic domains.

Importantly, the removal of a single amino acid reduced the molecular weight, reduced the number of rotatable bonds, increased hydrophobicity, and showed improved cell permeability compared to the parental peptidomimetic. A major hurdle for developing peptidomimetics as chemical probes is cell permeability. Identification of SW2_110A as a selective and cell-permeable CBX8 ligand indicates that further improvements can still be made to increase "druglikeness". The chemical tractability of the phenoxyphenyl group will facilitate further improvements in probe properties as will additional minimization of the ligand through removal of amide bonds.

While there are examples of redundant repressive functions for CBX paralogs,[60,61] there are also examples of individual paralogs acting in non-redundant roles.[8−12] In particular, CBX8 has been implicated in transcriptional activation in both development and disease.[18,62] Previous studies have demonstrated the significance of *HOXA9* gene activation for MLL-AF9 leukemogenesis, and the requirement for CBX8 in *HOXA9* gene activation; however, the mechanism by which CBX8 activates gene expression has been elusive. In fact, it has been suggested to be through a PRC1-independent function,[18,15] which may or may not require the chromodomain. Using our cell permeable CBX8 ChD inhibitors, we have determined that the ChD of CBX8 is required for CBX8-mediated *HOXA9* gene activation in MLL-AF9 leukemia and likely works through canonical binding of CBX8 at H3K27me3 to maintain chromatin boundaries *via* interaction with AF9.[52] Based on this model, the gene repression observed upon CBX8 inhibition requires the activity of other CBX paralogs, highlighting the necessity for highly selective CBX inhibitors for determining paralog-specific function.

METHODS

Materials. Oligonucleotides were purchased from IDT (Coralville, IA) or Bioneer (Alameda, CA) and used as provided. Analytical high-performance liquid chromatography (HPLC) separations were completed using an Agilent 1100 system with detection at 260 nm using a water/MeCN gradient containing 100 mM triethylammonium acetate, pH 5.5. Preparative HPLC separations were completed using a Varian ProStar system with detection at 260 and 280 nm using a water/MeOH gradient containing 0.75% hexafluoroisopropanol, 0.0035% triethylamine, pH 7.0. Reagents and solvents were used as received from commercial sources.

Preparation of 96 Single 140-Mer dsDNA Constructs. The integrated polymerase chain assembly (PCA)−PCR experiments were used to generate 96 single-gene barcode DNA constructs using a modified procedure.[63] For each reaction, six pairs of complementary 40-mer DNA oligonucleotides were used.[64] Six 40-mer oligos were pooled and used as templates for PCA. Each 5.0 μL PCA reaction contained 0.2 μM of each template 40-mer, with the following: 1.0 mM dNTPs, 0.1 U/μL of Vent DNA polymerase in 1x DNA polymerase buffer (NEB). All thermocycle procedures were as follows: 3 min at 94 °C, then cycling for denaturation at 94 °C for 15 s, annealing at 58 °C for 15 s, extension at 72 °C for 30 s, and a final extension of 72 °C for 5 min after 20 cycles. Each 50 μL PCR reaction contained 5 μL of PCA product, 0.2 mM each dNTP, 0.4 μM of each end primer (Z_A and Z_D'), and 0.025 U/μL DreamTaq DNA polymerase in 1X DreamTaq buffer (Thermo Fisher). The successive PCR went for 20 cycles using the same thermocycling conditions as PCA. Following PCR, each reaction was purified using SeraMag Carboxylate-Modified Magnetic SpeedBeads (GE Healthcare, Pittsburgh, PA) as previously reported[65] and quantified by UV absorbance at 260 nm.

CBX ChD Protein Expression and Purification. CBX chromodomain constructs (Addgene plasmids no. 25158 (CBX2), no. 25237 (CBX4), no. 25296 (CBX6), no. 25241 (CBX7), and no. 62514 (CBX8), provided by Cheryl Arrowsmith)[25] were transformed into chemically competent BL21 CodonPlus RIL *E. coli* cells (Stratagene, La Jolla, CA) as N-terminal His6-tagged proteins. Bacterial growth was completed at 37 °C in LB media to OD_{600} = 2.0, followed by reducing the temperature to 16 °C over 30−60 min and induced with 1 mM IPTG for 16 h. Cells were collected by centrifugation at 6000 rpm for 20 min and resuspended in ChD binding buffer (20 mM Tris, pH 8, 150 mM NaCl, 0.01% Tween 20, 20 mM imidazole) with 1.0 mM PMSF. Bacteria pellets were stored at −80 °C until needed. Pellets were thawed on ice for 10 min in ChD binding buffer and shaken at 4 °C supplemented with 100 μg/mL lysozyme, 1 mg mL^{-1} CHAPS and 1 mM PMSF, for 30 min. Cells were subsequently lysed by sonication (2x: 15 W for 30 s on, 30 s off, followed by 1x: 20 W for 1 min). The solubilized fraction was collected by centrifugation at 15000 rpm for 40 min at 4 °C. Meanwhile, Ni-NTA Agarose resin (QIAGEN, Venlo, The Netherlands) was washed with H_2O and equilibrated with ChD binding buffer. The soluble fraction was incubated with the pre-washed Ni-NTA agarose resin at 4 °C for 2 h. The resin was then washed three times with ChD purification buffer (20 mM Tris, pH 8, 150 mM NaCl, 0.01% Tween 20, 1 mM PMSF). Proteins were eluted by the addition of 0.5 M imidazole to ChD purification buffer. The elution was diluted with 30% glycerol, flash frozen, and stored at −80 °C until needed. Protein purity was assessed by SDS-PAGE and concentration was determined by the Pierce 660 kit (Thermo Scientific).

Preparation of Kme3-Ser-CPF. The first two residues of the CBX consensus sequence were synthesized in bulk as previously described.[28] Briefly, 150 nmol of NH_2-5′-CP_F in DEAE binding buffer (10 mM HOAc and 0.005% Triton X-100) was split between six cartridges. Each contained 220 μL of 50% DEAE Sepharose slurry in 50% ethanol and was pre-washed with DEAE binding buffer. The DNA-loaded cartridges were washed three times with 3 mL of MeOH. Fmoc-amino acid coupling was achieved by incubating the cartridges in 1 mL of 50 mM Fmoc-amino acid, 50 mM EDC-HCl, and 5 mM HOAt in 40% DMF/60% MeOH for 30 min at RT, with double couplings. After couplings, the cartridges were washed three times with 3 mL of MeOH and three times with 3 mL of DMF. Fmoc deprotection was achieved by incubating the cartridges in 1 mL of 20% piperidine in DMF for 30 min at RT and then washed three times with 3 mL of DMF, three times with 3 mL of MeOH, and once with 1 mL of DEAE binding buffer after the final coupling. The DNA was eluted and collected by passing 1 mL of DEAE elution buffer (1.5 M NaCl and 0.005% Triton X-100) through each cartridge. The crude conjugate was desalted and concentrated to dryness.

Positional Scanning Library Synthesis. The purified Kme_3-Ser-CP_F conjugate was suspended in 4.8 mL of DEAE binding buffer. To 96 wells in a 384-well filter plate, 20 μL of DEAE Sepharose was added and washed three times with 90 μL of DEAE binding buffer. To each well, 50 μL of Kme_3-Ser-CF_F solution (approximately 1 nmol conjugate per well) was added and washed three times with 90 μL of MeOH. Briefly, Fmoc-amino acids were coupled using 50 mM Fmoc-amino acid, 50 mM EDC-HCl, 5 mM HOAt in 40% DMF/60% MeOH for 30 min at RT with double coupling and deprotected by 20% piperidine in DMF for 30 min at RT. Wells were washed three times with 90 μL of MeOH and three times with 90 μL of DMF between each step. Following the final chemistry step, wells were washed three times with 90 μL of DMF, three times with 90 μL of MeOH, and once with 90 μL of DEAE binding buffer. DNA conjugates were eluted by incubating two times with 40 μL of DEAE elution buffer in each well for 5 min at RT and then collected by centrifugation. Each conjugate was then attached to a unique 140-mer dsDNA template sequence by PCR individually (1X DreamTaq Buffer, 0.5 μM CF_F-conjugate (PSL library member), 0.5 μM CP_R, 0.2 mM dNTPs, 0.05 ng/μL template, and 0.025 U/μL). All PCRs were

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

pooled and purified by SPRI and quantified by UV absorbance at 260 nm.

Positional Scanning Library Selection against PcG CBX ChDs. A frozen pellet from an induced 5 mL *E. coli* culture with CBX-His_6 ChD was suspended in 300 μL of ice-cold lysis buffer (20 mM Tris, pH 8, 150 mM NaCl, 100 μg/mL, 1 mg mL^{-1} CHAPS, 0.02% Tween-20, 1 mM PMSF) and lysed by sonication for 2 min (3 s on, 3 s off) at 30% power while on ice. The lysate was collected after centrifugation at 4000*g* at 4 °C. Meanwhile, 21 μL of His Mag Sepharose Ni (Ni-NTA-Sepharose MBs) were pre-washed three times with 21 μL of purification buffer (20 mM Tris, pH 8.0, 150 mM NaCl, 20 mM imidazole, 1 mM PMSF, 0.02% Tween-20). The soluble lysate was then combined with 12 μL of pre-washed Ni-NTA-Sepharose-MBs and incubated at 4 °C for 1 h. The MBs were separated and washed five times in 11 μL of purification buffer. After the last wash, the MBs were suspended in 12 μL of purification buffer. The CBX-bound MBs were split and diluted to yield 10 μL of 1X (~50 μM CBX), 10 μL of 1/10X (~5 μM CBX), and 10 μL of 1/20X (~2.5 μM CBX). The MBs were separated and 10 μL of the DNA pre-mix (50 nM benzylamide (Bz) on DNA construct 97 (non-ligand), 0.5 nM 4-BrBA-F-A-I-Kme3-S on DNA construct 98 (high-affinity ligand), and 50 nM CBX positional scanning library-DNA conjugates (approximately 0.5 nM of each library member) in 20 mM Tris, pH 8, 150 mM NaCl, 10 mM $MgCl_2$, 0.02% Tween-20, 1 mg mL^{-1} BSA, 1 mg mL^{-1} sheared salmon sperm DNA) was added to all four samples (mock [no protein/MBs only], 50 μM CBX, 5 μM CBX, and 2.5 μM CBX) and allowed to incubate at RT for 1 h. The MBs were then separated and washed five times in 10 μL of the above buffer. DNA conjugates and protein were eluted by incubating the MBs for 5 min at RT in the above buffer with 0.5 M imidazole. Each elution was collected and prepared for PCR and next-generation sequencing (NGS). The above procedure was applied to selections against all PcG CBX paralog.

Solid-Phase Peptide Synthesis (SPPS). Off-DNA peptides were prepared using traditional SPPS methods. All couplings and deprotections were monitored by ninhydrin tests. Briefly, 50 mg of Rink Amide MBHA resin was swelled for 20 min in 1,2-dichloroethane and 20 min in DMF before suspension in 20% piperidine in DMF for the initial Fmoc deprotection for 30 min at RT. Couplings were completed using 5.0 equiv (relative to the capacity of the resin) of Fmoc-AA (or carboxylic acid), 5.0 equiv of HOAt, and 5.0 equiv of DIC in DMF (approximately 0.1 M) and pre-activated for 20 min at RT before being added to the resin. Fmoc deprotections were achieved by incubating the resin for 30 min at RT in 20% piperidine in DMF. Peptides were cleaved and deprotected by incubating in 95% TFA, 2.5% triisopropylsilane, and 2.5% H_2O for 3 h at RT. The crude peptide was collected by precipitation out of ice-cold diethyl ether and then suspended in 50% MeOH/50% H_2O and concentrated to dryness. The residue was dissolved in DMSO and purified on a semi-prep HPLC using a H_2O/MeOH + 0.1% TFA gradient with detection at 215 and 254 nm. Yield was determined by mass of the dried, purified peptide as the TFA salt relative to the equivalents as determined by the mass of resin used. Purity was confirmed to be >95% by HPLC.

Synthesis of Diethyllysine Derivatives. Crude peptides were synthesized as described above. After purification of the peptide, reductive amination was accomplished *via* dissolvation of peptides in 80% MeOH, and 20% DMSO, followed with 100 equiv of acetaldehyde and 50 equiv of $NaCNBH_3$ (final peptide concn 0.1 M) and incubated at 37 °C overnight. The mixture was concentrated and HPLC purified as described above.

Synthesis of C-Terminal Alkyne Peptide. Modified methods from previously reported procedure[66] were used to synthesize C-terminal alkyne peptides. Polystyrene-linked aldehyde resin (FMPB AM resin, 100 mg, 1.08 mmol/g) was added to a round-bottom flask and gently stirred for 30 min at RT in DCM. DCM was gently evaporated and 5 mL of DMF with 5 mL of MeOH was added to the resin. To this, 10 equiv of glacial AcOH was added with 10 equiv of propargyl amine and 10 equiv of $NaCNBH_3$ and gently stirred under light reflux for 3 h at 80 °C. The mixture was cooled and washed with MeOH, DCM, and DMF and re-swelled for 30 min in 1,2-dichloroethane prior to the first acylation. To the resin, 5.0 equiv of Fmoc-Ser(OtBu)-OH with 5.0 equiv of DIC, 8.0 equiv of HOAt in DMF was added and incubated at 37 °C overnight. The remaining synthesis, purification, and reductive amination were completed as described above. Purity was confirmed to be >95% by HPLC.

Synthesis of 5-/6-FAM. To 10.0 mg of 5-/6-FAM NHS ester (ThermoFisher), 422 μL of THF was added. Once dissolved, 7.2 mg (3.0 equiv) of 4-azido-1-aminobutane was added and mixed vigorously. A precipitate initially formed but dissolved upon mixing and then the reaction was incubated at RT, protected from light, and incubated at RT for 16 h. The reaction was then concentrated and purified by semi-prep HPLC with H_2O/MeOH 0.1% TFA gradient.

Synthesis of FAM−Peptide Conjugates. To 150 μL of 100 mM alkyne peptides in DMSO was added 13.1 mg of a single isomer of 4-azido-5/6-FAM (0.5 equiv). To this were added 5.0 μL of 2 M TEAA, pH 5.5, and 10 μL of 0.1 M aminoguanidinium-HCl. Separately, 25 μL of CuBr-saturated DMSO was suspended in 50 μL of 50 mM THPTA and then added to the azide/alkyne mixture. The mixture was incubated at RT overnight and then 10 μL of 0.5 M EDTA, pH 8 was added to the mixture. The FAM-peptide conjugate was purified as described above. Purity was confirmed to be >95% by HPLC.

Synthesis of Biotin−Peptide Conjugate. To 150 μL of 100 mM alkyne peptides in DMSO was added 13.1 mg of Biotin-PEG3-azide (2.0 equiv). To this were added 5.0 μL of 2 M TEAA, pH 5.5, and 10 μL of 0.1 M aminoguanidinium-HCl. Separately, 25 μL of CuBr-saturated DMSO was suspended in 50 μL of 50 mM THPTA and then added to the azide/alkyne mixture. The mixture was incubated at RT for overnight, and then 10 μL of 0.5 M EDTA, pH 8, was added. The peptide−biotin conjugate was purified as described above. Purity was confirmed to be >95% by HPLC.

Synthesis of Chloroalkane Linker. 2-(2-Azidoethoxy)ethanol was prepared as previously described.[67]

Competitive Fluorescence Polarization (FP) Assay of PSL Hits against CBX6 ChD, CBX7 ChD, and CBX8 ChD. Fluorescence Polarization (FP) was measured by titration of CBX ChDs to a FITC-labeled probe as previously reported.[24] Binding and competition FP assays were performed in black 384-well plates with optical bottoms. Buffer used in FP assays consists of 20 mM Tris, pH 8, 150 mM NaCl, 0.01% Tween 20. The FITC-labeled probe was kept constant at 100 nM with 1 μM CBX6 ChD, 0.4 μM CBX7 ChD, or 4 μM CBX8 ChD, concentrations selected based on the reported relative affinity of the CBX ChD protein for the FITC probes. Two-fold dilutions of ligand were used, starting with 500 μM as highest peptide concentration to 0.488 μM as the lowest. Four replicates were tested at each concentration. Raw data were analyzed using GraphPad Prism 7 following a "one site-Fit $logIC_{50}$" competition model with any outliers (95% confidence interval) being excluded.

Competitive Fluorescence Polarization (FP) Assay with H3K27me3. FP assays were performed in black 384-well plates with 100 nM FAM-labeled SW2_110A (SW2_110AL-FL) and 3 μM CBX8 ChD CBX8. Two-fold dilutions of H3K27me3(21−44)) peptide (Active Motif 81052) were added starting at 1 mM as highest peptide concentration. Raw data were analyzed using GraphPad Prism 7 following a "one site-Fit $logIC_{50}$" competition model.

Direct Fluorescence Polarization (FP) Binding Assay of PSL Hits against PcG CBX ChDs. FP assays were conducted as above with slight modifications. The FAM-labeled peptide was kept constant at 100 nM except for high-affinity ligands (SW2_104A-FL), where the concentration was reduced to 10 nM. The CBX ChD proteins were titrated by 2-fold series dilutions in the assays with varying protein concentrations, depending on the binding affinity of the ligands. Four replicates were used for each ligand. Raw data were analyzed for determinations of K_d, using GraphPad Prism 7 following a "one-site" total binding model with any outliers (95% confidence interval) being excluded.

Thermal Shift Assay (TSA). The TSA was performed according to a previously reported protocol.[68] In brief, the reaction was run in 20 μL using a standard qPCR machine with a ROX filter (Applied

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

Biosystems). The reaction was run in the following reaction buffer: 10 mM HEPES 7.0, 150 mM NaCl, 8X SYPRO Orange S6651 Invitrogen (5000X stock), 0.2 mg mL^{-1} CBX8 ChD, 5% DMSO containing SW2_110A at designated concentrations. Melt curves were obtained using a temperature gradient of 25−75 °C in 40 min with readings every 0.5 °C. Melt curves for CBX8 ChD were obtained for four replicates at each ligand concentration and the T_m values were calculated using non-linear least-squares fit on Prism 8. The approximate K_d was calculated from T_m values using non-linear least-squares fit on Prism 8.

Solubility Assay. Aqueous solubility assays (Eurofins, ITEM 435) involved assessment of solubility (from 10 mM DMSO stock solution) using a shake-flask method with UV−vis detection.[33] The DMSO stock solution was diluted to 200 μM with phosphate buffered saline (pH 7.4) at RT and shaken for 24 h. After centrifugation, the concentration of corresponding soluble compound was determined by HPLC-UV, comparing the peak area obtained with that from a series of reference compounds in DMSO. The results presented are the average of duplicate measurements.

NMR analysis of SW2_110A Binding. The CBX8 CD construct was a gift from Cheryl Arrowsmith (Addgene plasmid no. 62514). GST-tagged CBX8 CD was created using Infusion and the pGSTag vector, provided by Gerald Crabtree. The CBX8 CD were expressed in BL21 (DE3) pLysS *E. coli* cells. Cells were grown in LB media at 37 °C to an A_{600} OD of ∼1.0. Cells were pelleted using centrifugation at 4000 rpm and 18 °C for 10 min. Cells were then resuspended in M9 minimal media (4 L of LB cells per 1 L of M9) supplemented with ^{15}N-NH_4Cl. Cells were allowed to recover at 18 °C and 210 rpm for up to 1 h before induction with 1 mM IPTG for 16−18 h. Cells were pelleted *via* centrifugation at 6000 rpm and 18 °C for 20 min. Cells were resuspended in 40 mL of Low Salt Buffer (25 mM Tris-HCl (pH 7.5), 50 mM NaCl) with DNase I and a protease inhibitor tablet. Resuspended pellets were lysed using the Emulsiflex. Cell lysate was cleared at 15000 rpm and 4 °C for 1 h.

^{15}N-CBX8 CD was purified according to the following protocol. Clarified cell lysate containing GST-tagged ^{15}N-CBX8 CD was rocked with glutathione agarose resin for 1 h at 4 °C. GST-tagged ^{15}N-CBX8 CD bound beads were purified using a gravity flow column. Bound beads were rinsed thoroughly with high salt buffer (25 mM Tris-HCl (pH 7.5), 1 M NaCl), followed by low salt buffer. GST-tagged ^{15}N-CBX8 CD was eluted from the beads using 50 mM glutathione in Low Salt Buffer, adjusted to a pH of 7.5. GST-tagged CBX8 CD was concentrated to a volume of 2 mL using a 10000 MWCO filter. The GST tag was cleaved using TEV protease at RT (25 °C) for 3 h. The cleaved ^{15}N-CBX8 CD was then purified using cation exchange chromatography and size exclusion (Superdex S75, 300/10). All ^{15}N-CBX8 CD were stored in a final buffer containing 40 mM NaPi (pH 6.8) and 100 mM NaCl.

SDS-PAGE was used to confirm the identity and purity of ^{15}N-CBX8 CD samples. Quantification of ^{15}N-CBX8 CD was performed using the calculated extinction coefficient (ε = 19 480 M^{-1} cm^{-1}) and measured A_{280} value. All ^{15}N-CBX8 CD samples were concentrated to 25−50 μM for NMR and flash frozen in liquid nitrogen for long-term storage at −80 °C. Prior to collection of HSQC data, ^{15}N-CBX8 CD samples were thawed overnight at 4 °C.

^{15}N-HSQC spectra were collected on 25−50 μM ^{15}N-CBX8 CD at 25 °C on a Bruker Avance II 800 MHz spectrometer equipped with a cryogenic probe. Titration with DMSO was performed by subsequently adding 1%, 2%, and 5% (v/v%) DMSO. Titration with SW2_110A was performed by addition of 0.5 (1% DMSO), 1.0 (2% DMSO), 2.5 (5% DMSO), or 11 (6% DMSO) molar ratios of compound to protein. All spectra were processed using NMRPipe and ccpNmr.

Normalized chemical shift perturbation values ($\Delta\delta$) were calculated for the DMSO and SW2_110A in Excel using the following equation:

$$\Delta\delta = \sqrt{(\Delta\delta_H)^2 + (0.20\Delta\delta_N)^2}$$

where $\Delta\delta$ is the chemical shift perturbation in parts per million (ppm). $\Delta\delta$ values were considered significant when greater than the average plus one standard deviation after trimming the 10% of residues with the largest $\Delta\delta$ value.

MD Simulation. Molecular dynamics using the Amber16 suite[69] were carried out for CBX6 (PDB: 3GV6)[25] and CBX8 (PDB: 3I91)[25] hosts bound to compound SW2 in the presence of approximately 8000 TIP3PBOX waters along with charge neutralizing chloride ions using the ff14S force field.[70] Simulations were minimized, heated to 300 K over 200 ps, and then equilibrated for 100 ns to yield the data presented in MD related figures. Equilibration trajectories were done using a 2 fs time step at 300 K under NPT conditions. Interaction cutoffs were set to 8 Å and SHAKE hydrogen constraints were applied. Initial pose generation for MD was done using AutoDock Vina[71] in combination with UCSF Chimera.[72] Parameterization of non-standard residues was in part done using Gaussian09[73] and the AmberTools16[69] suite of preparatory programs.

File Preparation and Simulation Setup. Non-standard residue parametrization was done by construction in Avogadro with C and N terminal caps. This includes the diphenyl ether n-terminus, the cyclopentyl alanine, and the trimethyllysine. Residue structures were minimized in Gaussian09 at the HF 6-31G* level of theory. AmberTools16 residuegen utility was then used to construct the preparatory files for use in tLeAP molecular dynamics preparation environment. The SW2_101B ligand was also loaded into UCSF Chimera to generate pdbqt files required for AutoDock Vina. Host coordinates were taken from the crystal structures of PDB 3I91 and 3GV6 by stripping extraneous atoms such as ligand, solvent, and ions.

A total of 500 ligand poses for both CBX6 and CBX8 were generated using AutoDock Vina (Exhaustiveness 7) on an MD generated ensemble of 50 host configurations. The top 10 docked poses were then selected for molecular dynamics. This process was done iteratively until converged structures were found. RMSD clustering was done on the final replicate trajectories to provide a starting pose for the 100 ns production trajectories.

Computational Resources. Each 10-pose iteration required 12 h on 280 cores (10 nodes with 2 Intel Xeon E5-2680 v4 processors each). The 100 ns trajectories required approximately 120 h using 28 cores each. Total simulation time for both systems including starting pose generation was approximately 180 h.

Cell Culture. HEK293T cells were cultured in Dubecco's Modified Essential Media (DMEM), 10% fetal bovine serum (FBS, JR Scientific), 1% glutagro (Corning), 1% penicillin/streptomycin (Corning), 1% sodium pyruvate (Corning). Human THP1 cells were cultured in RPMI (Gibco), 10% FBS (J R Scientific), 1% sodium pyruvate (Invitrogen), 1% Pen/Strep (Invitrogen), 1% Glutamax (Thermo Scientific), 0.1% 2-mercaptoethanol. Human K562 cells were cultured in RPMI (Gibco), 10% FBS (JR Scientific), 1% sodium pyruvate (Invitrogen), 1% Pen/Strep (Invitrogen), 1% Glutamax (Thermo Scientific). Halo-GFP-Mito HeLa cells were cultured in DMEM (Gibco), 10% FBS (JR Scientific), 1% Sodium pyruvate (Invitrogen), 1% Pen/Strep (Invitrogen), 1% Glutamax (Thermo Scientific), 1 μg/mL puromycin. Human Hs68 cells were cultured in DMEM (Gibco), 10% FBS (JR Scientific), 1% Sodium pyruvate (Invitrogen), 1% Pen/Strep(Invitrogen), 1% Glutamax (Thermo Scientific). Human G401 cells were cultured in McCoy's (Corning), 10% FBS (Thermo Scientific), 1% Pen/Strep (Invitrogen), 1% Glutamax (Thermo Scientific), 1% MEM NEAA (Invitrogen). Human LNCaP cells were cultured in RPMI (Gibco), 10% FBS (J R Scientific), 1% Pen/Strep (Invitrogen), 1% Glutamax (Thermo Scientific). All cells were grown at 37 °C and 5% CO_2. For generation of CBX8 and control CRISPR knockout lines, 200 000 THP1 cells were plated in 6-well 24 h prior to transfection. The respective vector (3.3 μg) was co-transfected with 13 μL of Fugene 6 (Promega). Media was changed 24 h post-transfection. Transfected cells underwent puromycin selection (2 μg/mL) for 3 days, 48 h post-transfection.

Lentiviral Transduction. HEK293T cells were co-transfected with pLKO.1 constructs and viral packaging vectors (pMD2.G and psPAX2). Short hairpin constructs for knockdown are below: CBX8

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

(TRCN0000021896), CBX7 (TRCN0000019144). Viral supernatant was harvested 72 h after transfection and concentrated by ultracentrifugation at 17 300 rpm for 2 h. Virus was resuspended in 100 μL of PBS and 5 μL was added to K562 or THP1 cells in a minimal volume of media for 1 h on an orbital shaker. Additional media was added to bring cells to 0.1 × 10^6 cells/mL and cells were grown at 37 °C and 5% CO_2. Twenty-four hours after infection, cells were selected with puromycin (2 μg/mL) for 48 h.

Cell Proliferation Assays. Leukemia cell lines were seeded at 0.1 × 10^6 cells/mL in 24-well flat bottom cell culture plate (no. 353047, Corning). Peptides SW2_110A and SW2_104A (100 μM in DMSO) were added to the cells. Cells were grown for 3 days, the cells were counted and the media was exchanged with fresh media containing DMSO, SW2_110A, or SW2_104A. At day 6, cells in each well were split 1:4 and transferred to new wells to avoid over confluency, and media was replaced with fresh compounds in new media. The cells were counted at 12 days, with additional cell counting and fresh compounds replenishment at days 3, 6, and 9.

CellTiter-Glo Luminescent Dose-Dependent Cell Viability Assay. The effects of SW2_110A and SW2_104A on cell viability were determined using a CellTiter-Glo ATP detection system (no. G7573, Promega). THP1 cells were seeded in 0.1 × 10^6 cells/mL density in 96-well clear bottom white microplate (no. 655098, Greiner Bio-One). Cells were treated with compounds SW2_110A and SW2_104A for 12 days, with fresh compounds replenishment at days 3, 6, and 9. For dose−response studies, IC_{50} was derived from an eight-point 2-fold titration ranging from 100 μM to 1.56 μM of SW2_110A or SW2_104A. CellTiter-Glo reagent was added to cells, and incubated with gentle shake for 15 min in dim light at RT. Luminescence was read on a GloMax microplate reader. Luminescence was normalized to DMSO-treated groups. The IC_{50} was calculated using the "log[inhibitor] vs the normalized response-variable slope" equation in GraphPad Prism 7.

Sequential Salt Extraction (SSE). SSE was performed as previously described.[35] 2.5 × 10^6 293T cells were seeded in 10 cm cell culture dish (no. 353003, Corning) overnight. Next day, media was removed and cells were washed with PBS. Cells were then pre-treated with peptides SW2_110A or DMSO on plate (100 μM, 1% DMSO) in 3 mL of media for 4 h in 37 °C. After the 4 h pre-treatment, media was removed and cells were washed again with PBS. Cells were harvested and washed with PBS. [Note: It is critical to balance the cell numbers the same in the peptide treated group and the DMSO-treated control group.] Cells were resuspended in 1 mL of Buffer A (25 mM HEPES, 25 mM KCl, 5 mM $MgCl_2$, 0.1% NP-40, 10% glycerol, 0.05 M EDTA, pH 7.8, plus protease inhibitor) was added to the cell pellet from the centrifuge and rotated at 4 °C for 10 min. Cells were spun down at 6500 × g for 5 min at 4 °C. Supernatant was removed and cell pellet was resuspended with 500 μL of mRIPA (Modified Radioimmunoprecipitation Assay) buffer (50 mM Tris, 1% Nonidet P-40, 0.25% sodium deoxycholate, plus protease inhibitors) by pipetting up and down 15 times and incubated on ice for 5 min. [Note: It is critical to maintain the consistence for all washing steps with the subsequent NaCl containing mRIPA buffers.] The sample was then centrifuged for 3 min at 6500*g*. The supernatant was saved in a separate tube, labeled as "0 mM fraction" - 0 mM sequential salt extraction washing supernatant. [Note: DMSO or the peptide SW2_110A was also added to the mRIPA washing buffers, in order to maintain the peptide in binding to the protein through the assay.] The pellet was sequentially resuspended in 500 μL of mRIPA Buffer with increasing NaCl concentrations (100, 200, 300, 400, and 500 mM). The procedures for 0 mM was repeated for each salt concentrations in the subsequent washes. All washing supernatants were saved and labeled. Next, 4X Bolt LDS sample buffer (Invitrogen) with 10% β-mercaptoethanol (AMRESCO LLC, Solon, OH) was added to each sample, and 50 μL of each fraction was loaded onto a 4−12% gradient gel (Invitrogen) for immunoblotting analysis of the proteins of interest. ImageJ was employed to quantitate the protein bands.

Chemoprecipitation (Peptide Pull-Down). Cells were grown to confluency in a 15 cm cell culture dish and washed with PBS. Cells were scraped into with 2 mL of Buffer A (25 mM HEPES, 5 mM KCl, 25 mM $MgCl_2$, 0.05 mM EDTA, 10% glycerol, 0.1% NP-40, plus protease inhibitors) and lysed on ice for 15 min. The nuclei were pelleted and resuspended in 2 mL of peptide pull-down buffer (20 mM HEPES pH 7.9, 250 mM NaCl, 0.6% NP-40, protease inhibitor and 0.5 mM DTT). Benzonase (200 U) was added and the lysate was incubated at 37 °C for 10 min and RT for 10 min with agitation. The samples were centrifuged at 13000 rpm at 4 °C for 10 min and lysates were transferred to separate tubes. Meanwhile, 30 μL of streptavidin M-270 Dynabeads (Solulink, San Diego, CA) was washed three times with peptide pulldown buffer. Biotinylated ligands in DMSO (0.3 μL of 10 mM stock), were incubated with pre-equilibrated beads at RT for 1 h. Extra unbound ligands were removed by washing 2× with peptide pulldown buffer, and 300 μL (∼300 μg) of nuclear lysate supernatant was added to immobilized biotinylated peptide (SW2_110A-B, SW2_104A-B) or beads alone. The mixture was rotated at 4 °C overnight, the depleted lysate was removed, and the beads were washed with peptide pulldown buffer 3 × 5 min at RT. The bound proteins were eluted from the beads with 1× Bolt LDS Sample Buffer (Invitrogen) with 10% β-mercaptoethanol (AMRESCO LLC, Solon, OH). The samples, along with 10% input samples were heated at 95 °C for 5 min and loaded onto a 4−12% gradient gel (Invitrogen) for immunoblotting analysis of the proteins of interest. ImageJ was employed to quantitate the protein bands.

Immunoblot and Antibodies. Lysates were boiled and loaded on a 4−12% SDS-PAGE gel (Invitrogen). Gels were transferred to PDVF membranes (Millipore) and incubated in 5% bovine serum albumin (BSA) in PBS-t (PBS with 0.1% Tween-20) prior to primary antibody. Blots were incubated at 4 °C overnight in primary antibody. Blots were washed with PBS-t and incubated for 1 h at RT in goat anti-rabbit or mouse conjugated to IRDye 800CW or IRDye 680 (LI-COR) secondary antibody. Blots were imaged on the Licor Odyssey. Primary antibodies used: CBX8 (rabbit, 1:1000, Bethyl Cat. No. A300-882A), CBX4 (rabbit, 1:500, Bethyl Cat. No. A302-355A), CBX7 (rabbit, 1:1000, Bethyl Cat. No. A302-525A), CBX2 (mouse, 1:400, Santa Cruz Cat. No. sc-136387), CBX6 (mouse, 1:400, Santa Cruz Cat. No. sc-86354), TBP (mouse, 1:1000, Abcam Cat. No. ab818).

Chromatin Immunoprecipitation-qPCR (ChIP-qPCR). Cells were grown to confluency in 100 mm cell culture plates (∼1−2 × 10^7 cells). Hs68 cells were treated with 100 μM SW2-110A or DMSO for 4 h and THP1 cells were treated with 100 μM SW2-110A or DMSO for 24 h. ChIP was performed as previously described.[34] Briefly, cells were washed with PBS and fixed with 1% formaldehyde in PBS for 10 min at RT. Cross-linking was quenched with 0.125 M glycine for 5 min at 4 °C. Cells were washed once with PBS and resuspended in CiA NP Rinse buffer 1 (50 mM HEPES pH 8.0, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X) for 10 min. Cells were pelleted at 400*g* for 5 min at 4 °C. The supernatant was removed, and the cells were resuspended in CiA NP rinse buffer 2 (10 mM Tris pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 200 mM NaCl). Cells were collected by centrifugation at 400*g* for 5 min. Supernatant was removed and the cells were washed twice with shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris HCl, pH 8.0). Cells were resuspended in 1 mL of shearing buffer and sonicated with a probe sonicator (Branson) to obtain ∼300 bp DNA fragments. Lysate was centrifuged at 21000*g* for 15 min to remove debris. Supernatant was collected and pre-cleared 2 h with Protein A Dynabeads (Thermo Fisher). For immunoprecipitation, 250 μL of pre-cleared cell lysate was incubated with 2 μg of antibody overnight and 10% input was saved. The IPs were washed three times for 3 min at RT with IP buffer (50 mM HEPES/KOH pH 7.5, 300 mM NaCl, 1 mM EDTA, 1% Triton X, 0.1% DOC, 0.1% SDS) followed by 3 min with DOC (10 mM Tris pH 8.0, 0.25 M LiCl, 0.5% NP-40, 0.5% DOC, 1 mM EDTA) and once with TE. Protein was eluted from beads with 300 μL of elution buffer (1% SDS, 0.1 M $NaHCO_3$) for 20 min at RT with agitation. Samples, including saved input, were treated with RNase A (100 U) at 37 °C for 30 min followed by proteinase K (2 μg) digestion for 3 h at 55 °C. Samples were reverse-cross-linked overnight at 65 °C and extracted with phenol chloroform followed by isopropanol precipitation. The isolated DNA was resuspended in 80

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

Table 3. Primers Used in the qPCR

gene name	forward primer (5′-3′)	reverse primer (5′-3′)
LMNB2	CCGAATCTCTGAAATGAAAGTCCATGC	TTAAAGATCTGAGGGACTCCTCAGTC
CCND2 (Pemberton *et al.*, 2016)	ACTGTCTGAAATGAAGGTGAAGC	GATTTGATGGACACTTGGTTTGT
GATA6 (Pemberton *et al.*, 2016)	GCCTCTCCATTCCAGAGTTTT	TCCAGAAACCGTTCTCATCC
RUNX3 (Pemberton *et al.*, 2016)	TCAAAAGGCATCCGCCTCTCCGT	AAGGATGCACCTGCCGGGAATTG
HOXA9 (−0.5 kb)	AAGTCGGAAACGACCAACAGA	TTACAGGGAGCTCGCCAAC
HOXA9 (−2.5 kb)	ACAAAGCTGCAGCGAATGTC	ATGATCACGACCGGATGGC
HOXA9 (−4.5 kb)	GCAAAATAACCGGCCTCTGC	CCTATAGCCCTGGTGCCGTA
HOXA9 (−6.5 kb)	GACGCTGCGGCTAATCTCTA	AAGAGTGGTCGGAAGAAGCG
HOXA9 (−8.5 kb)	CTGATGAGCGAGTCGACCAA	GGAAACTCTGGCTCGGGATT

Table 4. Primers Used in the qRT-PCR

gene name	forward primer (5′-3′)	reverse primer (5′-3′)
HOXA9	GGCCCAGGACCGAGATACTT	CGCTCACGGACAATCTAGTTGT
B2M	TGCTGTCTCCATGTTTGATGTATCT	TCTCTGCTCCCCACCTCTAAGT
ACTB	GCACCACACCTTCTACAATGA	GTCATCTTCTCGCGGTTGGC
MYB	TCAGGAAACTTCTTCTGCTCACA	AGGTTCCCAGGTACTGCT
CDK6	TGGAGACCTTCGAGCACC	CACTCCAGGCTCTGGAACTT
RUNX2	TCTTAGAACAAATTCTGCCCTTT	TGCTTTGGTCTTGAAATCACA
RUNX3	GTTCAACGACCTTCGCTTC	GTCCACGGTCACCTTGATG
BCR-ABL	ACTCCAGACTGTCCACAGCA	TTGGGGTCATTTTCACTGG
CBX6	AAACGGCGGATCCGAAAGGGAC	GCTGCAATGAGCCGCGAGTC
CBX7	CGTCATGGCCTACGAGGA	TGGGTTTCGGACCTCTCTT
CBX8	CAACATGGAGCTTTCAGCGG	GTGCTGTACTTCTGCGACCA

μL of water for qPCR analysis (4 μL used per well). Antibodies used for IP were CBX8 (Bethyl Cat. No. A300-882A, rabbit), IgG (CST Cat. No. 2729, rabbit), CBX7 (Bethyl Cat. No. A302-525A, rabbit), and AF9 (Bethyl Cat. No. A300-595A, rabbit).

Genome-Wide Data Analysis. Published annotated data sets were downloaded from Encyclopedia of DNA Elements (ENCODE) or Gene Expression Omnibus (GEO) as BED files.[74,75] Peaks for CBX8, CBX2, H3K4me3, H3K27me3, and ENL (MLLT1) were called in reference to GrCh38. All data sets were imported into R Studio. Peak overlaps were determined using the ChIPpeakAnno package.[76,77] Overlaps were defined as being within 150 base pairs of each other and on the same strand. Accession numbers for CBX8: GSM830987, GSM1295078, GSM1295089, ENCFF001SYX; H3K27me3: GSM1295084, GSM1295094, ENCFF001SZF; H3K4me3: GSM1295085, GSM1295095, ENCFF001SZJ; MLLT1: GSM2423844, GSM2423845, ENCSR675LRO.

ChIP-Seq Visualization. Published ChIP-Seq coordinate data sets from Bernt *et al.*[52] (GSE29130) were uploaded to the UCSC genome browser in reference to mm8 genome for visualization. Accession numbers: GSM721213, GSM721214, GSM721215, GSM721216, GSM721217, and GSM721218.

Quantitative Polymerase Chain Reaction (qPCR). qPCR was performed on the isolated ChIP DNA (4 μL) using SYBR master mix (Thermo) and run on the BioRad CFX thermo cycler. Three biological replicates were performed in technical triplicate. Enrichment was determined as percent of input DNA. Primers shown in Table 3 are used in the qPCR.

Quantitative Reverse Transcriptase Polymerase Chain Reaction (qRT-PCR). First, 1× 10^6 THP1 and K562 cells were treated with SW2_110A at 100 μM or DMSO (1%) for 24 h. Cells were harvested after 24 h for RNA extraction. After homogenization of THP1 or K562 cells using TRIzol reagent (Thermo Scientific), RNA was extracted from the aqueous phase in the phase separation step. RNA pellet was washed with 75% ethanol and concentrated for subsequent reverse transcription. Two μg RNA was then converted into cDNA using Verso cDNA synthesis kit (Thermo Scientific). SYBR Green Mastermix (Thermo Scientific) was used for quantitative PCR. Primers shown in Table 4 are used in the qRT-PCR.

Chloroalkane Penetration Assay (CAPA). CAPA is a recently developed cell penetration assay for measuring relative cytosolic access without interference from endosomally trapped peptides.[32] Halo-GFP-Mito HeLa cells were cultured and seeded at a 1× 10^5 cells/well in a 24- or 48-well plate the day before experiments. Cells were rinsed by PBS and treated with chloroalkane conjugated CBX8 peptidomimetic ligands SW2_110A-CA or KED97L-CA in acidified Opti-MEM (0.15% 6 N HCl) for 4 h. Next, media was removed, and cells were washed by phenol red-free Opti-MEM for 30 min, followed by incubation with 5 μM HT-TAMRA (HTag-TMR, Promega) for another 30 min. Then, cells were washed for 15 min by phenol red-free DMEM + 10% FBS + 1% pen/strep, followed with PBS wash and trypsin incubation. Cells were transferred to a new microcentrifuge tube and pelleted by centrifuge, with two times of PBS washes. Cell pellets were resuspended in 250 μL of PBS, and 200 μL was used for flow cytometry analysis. Live cells were gated and 10000 cells were measured per sample. Mean fluorescence intensity was calculated from raw data, and these values were normalized to the samples with no dye (0% red signal) and with dye but no HT-molecule (100% red signal).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acschembio.9b00654.

Supplemental figures (Figures SI 1−SI 14) and key resources; full selection results with all building block structures; off-DNA hit validation including FP assays, MST and TSA, NMR spectra, molecular dynamic simulations, cellular activity data, and LCMS chromatograms (PDF)

Data for enrichment relative to non-ligand (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: cjk@purdue.edu.

*E-mail: edykhui@purdue.edu.

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

ORCID

James M. B. McFarlane: 0000-0002-2505-1408
Fraser Hof: 0000-0003-4658-9132
Catherine A. Musselman: 0000-0002-8356-7971
Casey J. Krusemark: 0000-0003-2964-3520

Author Contributions

Conceptualization: C.J.K., E.C.D., S.W., K.E.D. Methodology: C.J.K., E.C.D., S.W., K.E.D. Software: C.J.K., E.C.D., S.W., K.E.D., K.E.C. Formal analysis: S.W., K.E.D., K.E.C. Investigation: S.W., K.E.D., J.M.B.M., K.F.H., T.W., K.E.C., M.C.G., N.M. Resources: C.J.K., E.C.D., C.A.M., I.P., F.H. Data curation: C.J.K., E.C.D., S.W., K.E.C., K.E.D. Writing - original draft preparation: S.W., C.J.K., E.C.D. Writing – review and editing preparation: S.W., F.H., C.A.M., I.P., C.J.K., E.C.D. Visualization: S.W., K.E.C., C.J.K., E.C.D. Supervision: C.J.K., E.C.D. Project administration: C.J.K., E.C.D. Funding acquisition: C.J.K., E.C.D., F.H., C.A.M., I.P.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank the lab of J. Kritzer for providing Halo-GFP-Mito HeLa cell lines, as well as advice on CAPA assay. We thank A. Alpsoy for helpful discussions on pulldown assays, and A. Dhiman for technical assistance in flow cytometry. We thank B. Peters and C. Maschinot for their help in NMR. We thank Y. Sun for assistance with off-DNA hit synthesis, purification, and LC-MS characterization. Support for this research was provided by the Purdue Center for Cancer Research, NIH grant P30 CA023168, in the form of the Genomics Core and the Pilot Grant Program. K.E.C. is supported by Indiana Clinical and Translational Science Institute Predoctoral Fellowship (UL1TR001108 to A.S., P.I., 9/26/2013–4/30/2018) and Purdue University Bilsland Dissertation Fellowship Aug 2018–Dec 2018. N.M. is supported by fellowships from the WestCoast Ride to Live and PCFBC. F.H. is supported by CCSRI (IG703789). Funding for the computational studies was provided by the National Science and Engineering Research Council of Canada, the Canada Foundation for Innovation, the British Columbia Knowledge Development Fund, and the University of Victoria. This research was performed in part using the Compute Canada and WestGrid computing resources. C.A.M. is supported by the National Science Foundation (CAREER-1452411) and National Institutes of Health (GM128705). C.J.K. is supported by NIH grants (R35GM128894, 1NS101535). E.C.D. is supported by NIH grants (CA207532) and the V Foundation for Cancer Research (V2014-004 and D2016-030).

■ DEDICATION

This article is dedicated to Dr. Laura L. Kiessling on the occasion of her 60th birthday.

■ REFERENCES

(1) Kennison, J. A. (1995) The Polycomb and Trithorax Group Proteins of *Drosophila*: Trans-Regulators of Homeotic Gene Function. *Annu. Rev. Genet. 29*, 289–303.
(2) Sauvageau, M., and Sauvageau, G. (2010) Polycomb Group Proteins: Multi-Faceted Regulators of Somatic Stem Cells and Cancer. *Cell Stem Cell 7*, 299–313.
(3) Di Croce, L., and Helin, K. (2013) Transcriptional Regulation by Polycomb Group Proteins. *Nat. Struct. Mol. Biol. 20*, 1147–1155.
(4) Francis, N. J., Kingston, R. E., and Woodcock, C. L. (2004) Chromatin Compaction by a Polycomb Group Protein Complex. *Science 306*, 1574–1577.
(5) Lavigne, M., Francis, N. J., King, I. F. G., and Kingston, R. E. (2004) Propagation of Silencing; Recruitment and Repression of Naive Chromatin in Trans by Polycomb Repressed Chromatin. *Mol. Cell 13*, 415–425.
(6) Connelly, K. E., and Dykhuizen, E. C. (2017) Compositional and Functional Diversity of Canonical PRC1 Complexes in Mammals. *Biochim. Biophys. Acta, Gene Regul. Mech. 1860*, 233–245.
(7) Gil, J., and O'Loghlen, A. (2014) PRC1 Complex Diversity: Where Is It Taking Us? *Trends Cell Biol. 24*, 632–641.
(8) Klauke, K., Radulović, V., Broekhuis, M., Weersing, E., Zwart, E., Olthof, S., Ritsema, M., Bruggeman, S., Wu, X., Helin, K., Bystrykh, L., and de Haan, G. (2013) Polycomb Cbx Family Members Mediate the Balance between Haematopoietic Stem Cell Self-Renewal and Differentiation. *Nat. Cell Biol. 15*, 353–362.
(9) Kloet, S. L., Makowski, M. M., Baymaz, H. I., Van Voorthuijsen, L., Karemaker, I. D., Santanach, A., Jansen, P. W. T. C., Di Croce, L., and Vermeulen, M. (2016) The Dynamic Interactome and Genomic Targets of Polycomb Complexes during Stem-Cell Differentiation. *Nat. Struct. Mol. Biol. 23*, 682–690.
(10) Morey, L., Pascual, G., Cozzuto, L., Roma, G., Wutz, A., Benitah, S. A., and Di Croce, L. (2012) Nonoverlapping Functions of the Polycomb Group Cbx Family of Proteins in Embryonic Stem Cells. *Cell Stem Cell 10*, 47–62.
(11) Morey, L., Santanach, A., Blanco, E., Aloia, L., Nora, E. P., Bruneau, B. G., and Di Croce, L. (2015) Polycomb Regulates Mesoderm Cell Fate-Specification in Embryonic Stem Cells through Activation and Repression Mechanisms. *Cell Stem Cell 17*, 300–315.
(12) O'Loghlen, A., Muñoz-Cabello, A. M., Gaspar-Maia, A., Wu, H. A., Banito, A., Kunowska, N., Racek, T., Pemberton, H. N., Beolchi, P., Lavial, F., Masui, O., Vermeulen, M., Carroll, T., Graumann, J., Heard, E., Dillon, N., Azuara, V., Snijders, A. P., Peters, G., Bernstein, E., and Gil, J. (2012) MicroRNA Regulation of Cbx7Mediates a Switch of Polycomb Orthologs during ESC Differentiation. *Cell Stem Cell 10*, 33–46.
(13) Mills, A. A. (2010) Throwing the Cancer Switch: Reciprocal Roles of Polycomb and Trithorax Proteins. *Nat. Rev. Cancer 10*, 669–682.
(14) Koppens, M., and Van Lohuizen, M. (2016) Context-Dependent Actions of Polycomb Repressors in Cancer. *Oncogene 35*, 1341–1352.
(15) Béguelin, W., Teater, M., Gearhart, M. D., Calvo Fernández, M. T., Goldstein, R. L., Cárdenas, M. G., Hatzi, K., Rosen, M., Shen, H., Corcoran, C. M., Hamline, M. Y., Gascoyne, R. D., Levine, R. L., Abdel-Wahab, O., Licht, J. D., Shaknovich, R., Elemento, O., Bardwell, V. J., and Melnick, A. M. (2016) EZH2 and BCL6 Cooperate to Assemble CBX8-BCOR Complex to Repress Bivalent Promoters, Mediate Germinal Center Formation and Lymphomagenesis. *Cancer Cell 30*, 197–213.
(16) Zhang, C. Z., Chen, S. L., Wang, C. H., He, Y. F., Yang, X., Xie, D., and Yun, J. P. (2018) CBX8 Exhibits Oncogenic Activity via AKT/ b-Catenin Activation in Hepatocellular Carcinoma. *Cancer Res. 78*, 51–63.
(17) Chung, C. Y., Sun, Z., Mullokandov, G., Bosch, A., Qadeer, Z. A., Cihan, E., Rapp, Z., Parsons, R., Aguirre-Ghiso, J. A., Farias, E. F., Brown, B. D., Gaspar-Maia, A., and Bernstein, E. (2016) Cbx8 Acts Non-Canonically with Wdr5 to Promote Mammary Tumorigenesis. *Cell Rep. 16*, 472–486.
(18) Tan, J., Jones, M., Koseki, H., Nakayama, M., Muntean, A. G., Maillard, I., and Hess, J. L. (2011) CBX8, a Polycomb Group Protein, Is Essential for MLL-AF9-Induced Leukemogenesis. *Cancer Cell 20*, 563–575.
(19) Santiago, C., Nguyen, K., and Schapira, M. (2011) Druggability of Methyl-Lysine Binding Sites. *J. Comput.-Aided Mol. Des. 25*, 1171–1178.

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112–131

(20) Ren, C., Morohashi, K., Plotnikov, A. N., Jakoncic, J., Smith, S. G., Li, J., Zeng, L., Rodriguez, Y., Stojanoff, V., Walsh, M., and Zhou, M. (2015) Small-Molecule Modulators of Methyl-Lysine Binding for the CBX7 Chromodomain. *Chem. Biol. 22*, 161−168.
(21) Ren, C., Smith, S. G., Yap, K., Li, S., Li, J., Mezei, M., Rodriguez, Y., Vincek, A., Aguilo, F., Walsh, M. J., and Zhou, M. (2016) Structure-Guided Discovery of Selective Antagonists for the Chromodomain of Polycomb Repressive Protein CBX7. *ACS Med. Chem. Lett. 7*, 601−605.
(22) Stuckey, J. I., Dickson, B. M., Cheng, N., Liu, Y., Norris, J. L., Cholensky, S. H., Tempel, W., Qin, S., Huber, K. G., Sagum, C., Black, K., Li, F., Huang, X., Roth, B. L., Baughman, B. M., Senisterra, G., Pattenden, S. G., Vedadi, M., Brown, P. J., Bedford, M. T., Min, J., Arrowsmith, C. H., James, L. I., and Frye, S. V. (2016) A Cellular Chemical Probe Targeting the Chromodomains of Polycomb Repressive Complex 1. *Nat. Chem. Biol. 12*, 180−187.
(23) Milosevich, N., Gignac, M. C., McFarlane, J., Simhadri, C., Horvath, S., Daze, K. D., Croft, C. S., Dheri, A., Quon, T. T. H., Douglas, S. F., Wulff, J. E., Paci, I., and Hof, F. (2016) Selective Inhibition of CBX6: A Methyllysine Reader Protein in the Polycomb Family. *ACS Med. Chem. Lett. 7*, 139−144.
(24) Simhadri, C., Daze, K. D., Douglas, S. F., Quon, T. T. H., Dev, A., Gignac, M. C., Peng, F., Heller, M., Boulanger, M. J., Wulff, J. E., and Hof, F. (2014) Chromodomain Antagonists That Target the Polycomb-Group Methyllysine Reader Protein Chromobox Homolog 7 (CBX7). *J. Med. Chem. 57*, 2874−2883.
(25) Kaustov, L., Ouyang, H., Amaya, M., Lemak, A., Nady, N., Duan, S., Wasney, G. A., Li, Z., Vedadi, M., Schapira, M., Min, J., and Arrowsmith, C. H. (2011) Recognition and Specificity Determinants of the Human Cbx Chromodomains. *J. Biol. Chem. 286*, 521−529.
(26) Goodnow, R. A., Dumelin, C. E., and Keefe, A. D. (2017) DNA-Encoded Chemistry: Enabling the Deeper Sampling of Chemical Space. *Nat. Rev. Drug Discovery 16*, 131−147.
(27) Neri, D., and Lerner, R. A. (2018) DNA-Encoded Chemical Libraries: A Selection System Based on Endowing Organic Compounds with Amplifiable Information. *Annu. Rev. Biochem. 87*, 479−502.
(28) Denton, K. E., Wang, S., Gignac, M. C., Milosevich, N., Hof, F., Dykhuizen, E. C., and Krusemark, C. J. (2018) Robustness of In Vitro Selection Assays of DNA-Encoded Peptidomimetic Ligands to CBX7 and CBX8. *SLAS Discov. 23*, 417−428.
(29) Tardat, M., Albert, M., Kunzmann, R., Liu, Z., Kaustov, L., Thierry, R., Duan, S., Brykczynska, U., Arrowsmith, C. H., and Peters, A. H. F. M. (2015) Cbx2 Targets PRC1 to Constitutive Heterochromatin in Mouse Zygotes in a Parent-of-Origin-Dependent Manner. *Mol. Cell 58*, 157−171.
(30) Malik, B., and Hemenway, C. S. (2013) CBX8, a Component of the Polycomb PRC1 Complex, Modulates DOT1L-Mediated Gene Expression through AF9/MLLT3. *FEBS Lett. 587*, 3038−3044.
(31) Pemberton, H., Anderton, E., Patel, H., Brookes, S., Chandler, H., Palermo, R., Stock, J., Rodriguez-Niedenführ, M., Racek, T., de Breed, L., Stewart, A., Matthews, N., and Peters, G. (2014) Genome-Wide Co-Localization of Polycomb Orthologs and Their Effects on Gene Expression in Human Fibroblasts. *Genome Biol. 15*, R23.
(32) Peraro, L., Zou, Z., Makwana, K. M., Cummings, A. E., Ball, H. L., Yu, H., Lin, Y. S., Levine, B., and Kritzer, J. A. (2017) Diversity-Oriented Stapling Yields Intrinsically Cell-Penetrant Inducers of Autophagy. *J. Am. Chem. Soc. 139*, 7792−7802.
(33) Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997) Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev. 23*, 3−25.
(34) Connelly, K. E., Weaver, T. M., Alpsoy, A., Gu, B. X., Musselman, C. A., and Dykhuizen, E. C. (2019) Engagement of DNA and H3K27me3 by the CBX8 Chromodomain Drives Chromatin Association. *Nucleic Acids Res. 47*, 2289−2305.
(35) Porter, E. G., and Dykhuizen, E. C. (2017) Individual Bromodomains of Polybromo-1 Contribute to Chromatin Association and Tumor Suppression in Clear Cell Renal Carcinoma. *J. Biol. Chem. 292*, 2601−2610.
(36) Marian, C. A., Stoszko, M., Wang, L., Leighty, M. W., de Crignis, E., Maschinot, C. A., Gatchalian, J., Carter, B. C., Chowdhury, B., Hargreaves, D. C., Duvall, J. R., Crabtree, G. R., Mahmoudi, T., and Dykhuizen, E. C. (2018) Small Molecule Targeting of Specific BAF (MSWI/SNF) Complexes for HIV Latency Reversal. *Cell Chem. Biol. 25*, 1443−1455.
(37) Los, G. V., Encell, L. P., McDougall, M. G., Hartzell, D. D., Karassina, N., Zimprich, C., Wood, M. G., Learish, R., Ohana, R. F., Urh, M., Dan, S., Jacqui, M., Kris, Z., Paul, O., Gediminas, V., Ji, Z., Aldis, D., Dieter, H. K., Robert, F. B., Keith, V. W., et al. (2008) HaloTag: A Novel Protein Labeling Technology for Cell Imaging and Protein Analysis. *ACS Chem. Biol. 3*, 373−382.
(38) de Groot, R. P., Raaijmakers, J. A., Lammers, J. W., Jove, R., and Koenderman, L. (1999) STAT5 Activation by BCR-Abl Contributes to Transformation of K562 Leukemia Cells. *Blood 94*, 1108−1112.
(39) Zeisig, D. T., Bittner, C. B., Zeisig, B. B., García-Cuéllar, M.-P., Hess, J. L., and Slany, R. K. (2005) The Eleven-Nineteen-Leukemia Protein ENL Connects Nuclear MLL Fusion Partners with Chromatin. *Oncogene 24*, 5525−5532.
(40) Monroe, S. C., Jo, S. Y., Sanders, D. S., Basrur, V., Elenitoba-Johnson, K. S., Slany, R. K., and Hess, J. L. (2011) MLL-AF9 and MLL-ENL Alter the Dynamic Association of Transcriptional Regulators with Genes Critical for Leukemia. *Exp. Hematol. 39*, 77−86.
(41) Mueller, D., Bach, C., Zeisig, D., Garcia-Cuellar, M. P., Monroe, S., Sreekumar, A., Zhou, R., Nesvizhskii, A., Chinnaiyan, A., Hess, J. L., and Slany, R. K. (2007) A Role for the MLL Fusion Partner ENL in Transcriptional Elongation and Chromatin Modification. *Blood 110*, 4445−4454.
(42) García-Cuéllar, M. P., Zilles, O., Schreiner, S. A., Birke, M., Winkler, T. H., and Slany, R. K. (2001) The ENL Moiety of the Childhood Leukemia-Associated MLL-ENL Oncoprotein Recruits Human Polycomb 3. *Oncogene 20*, 411−419.
(43) Hemenway, C. S., De Erkenez, A. C., and Gould, G. C. D. (2001) The Polycomb Protein MPc3 Interacts with AF9, an MLL Fusion Partner in t(9;11)(P22;Q23) Acute Leukemias. *Oncogene 20*, 3798−3805.
(44) Sitwala, K. V., Dandekar, M. N., and Hess, J. L. (2008) HOX Proteins and Leukemia. *Int. J. Clin. Exp. Pathol. 1*, 461−474.
(45) Yokoyama, A., and Cleary, M. L. (2008) Menin Critically Links MLL Proteins with LEDGF on Cancer-Associated Target Genes. *Cancer Cell 14*, 36−46.
(46) Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. (2002) MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia. *Nat. Genet. 30*, 41−47.
(47) Ayton, P. M., and Cleary, M. L. (2003) Transformation of Myeloid Progenitors by MLL Oncoproteins Is Dependent on Hoxa7 and Hoxa9. *Genes Dev. 17*, 2298−2307.
(48) Kumar, A. R., Hudson, W. A., Chen, W., Nishiuchi, R., Yao, Q., and Kersey, J. H. (2004) Hoxa9 Influences the Phenotype but Not the Incidence of Mll-AF9 Fusion Gene Leukemia. *Blood 103*, 1823−1828.
(49) Prange, K. H. M., Mandoli, A., Kuznetsova, T., Wang, S. Y., Sotoca, A. M., Marneth, A. E., Van Der Reijden, B. A., Stunnenberg, H. G., and Martens, J. H. A. (2017) MLL-AF9 and MLL-AF4 Oncofusion Proteins Bind a Distinct Enhancer Repertoire and Target the RUNX1 Program in 11q23 Acute Myeloid Leukemia. *Oncogene 36*, 3346−3356.
(50) Dietrich, N., Bracken, A. P., Trinh, E., Schjerling, C. K., Koseki, H., Rappsilber, J., Helin, K., and Hansen, K. H. (2007) Bypass of Senescence by the Polycomb Group Protein CBX8 through Direct Binding to the INK4A-ARF Locus. *EMBO J. 26*, 1637−1648.
(51) Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M., Durham, T., Zhang, X., Donaghey, J., Epstein, C. B., Regev, A., and Bernstein, B. E. (2011) Combinatorial Patterning of Chromatin Regulators Uncovered by

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

Genome-Wide Location Analysis in Human Cells. *Cell 147*, 1628−1639.
(52) Bernt, K. M., Zhu, N., Sinha, A. U., Vempati, S., Faber, J., Krivtsov, A. V., Feng, Z., Punt, N., Daigle, A., Bullinger, L., Pollock, R. M., Richon, V. M., Kung, A. L., and Armstrong, S. A. (2011) MLL-Rearranged Leukemia Is Dependent on Aberrant H3K79 Methylation by DOT1L. *Cancer Cell 20*, 66−78.
(53) Maethner, E., Garcia-Cuellar, M. P., Breitinger, C., Takacova, S., Divoky, V., Hess, J. L., and Slany, R. K. (2013) MLL-ENL Inhibits Polycomb Repressive Complex 1 to Achieve Efficient Transformation of Hematopoietic Cells. *Cell Rep. 3*, 1553−1566.
(54) Moussa, H. F., Bsteh, D., Yelagandula, R., Pribitzer, C., Stecher, K., Bartalska, K., Michetti, L., Wang, J., Zepeda-Martinez, J. A., Elling, U., Stuckey, J. I., James, L. I., Frye, S. V., and Bell, O. (2019) Canonical PRC1 Controls Sequence-Independent Propagation of Polycomb-Mediated Gene Silencing. *Nat. Commun. 10*, 1931.
(55) Wesolowski, S. S., and Brown, D. G. (2016) The Strategies and Politics of Successful Design, Make, Test, and Analyze (DMTA) Cycles in Lead Generation, in *Lead Generation* (Holenz, J., Ed.), pp 487−512, Wiley-VCH, Weinheim.
(56) Franzini, R. M., Ekblad, T., Zhong, N., Wichert, M., Decurtins, W., Nauer, A., Zimmermann, M., Samain, F., Scheuermann, J., Brown, P. J., Hall, J., Gräslund, S., Schüler, H., and Neri, D. (2015) Identification of Structure−Activity Relationships from Screening a Structurally Compact DNA-Encoded Chemical Library. *Angew. Chem., Int. Ed. 54*, 3927−3931.
(57) Rogers, J. M., Passioura, T., and Suga, H. (2018) Non-proteinogenic Deep Mutational Scanning of Linear and Cyclic Peptides. *Proc. Natl. Acad. Sci. U. S. A. 115*, 10959−10964.
(58) Jenson, J. M., Xue, V., Stretz, L., Mandal, T., Reich, L., and Keating, A. E. (2018) Peptide Design by Optimization on a Data-Parameterized Protein Interaction Landscape. *Proc. Natl. Acad. Sci. U. S. A. 115*, E10342−E10351.
(59) Satz, A. L. (2015) DNA Encoded Library Selections and Insights Provided by Computational Simulations. *ACS Chem. Biol. 10*, 2237−2245.
(60) Bracken, A. P., Kleine-Kohlbrecher, D., Dietrich, N., Pasini, D., Gargiulo, G., Beekman, C., Theilgaard-Mönch, K., Minucci, S., Porse, B. T., Marine, J.-C., Hansen, K. H., and Helin, K. (2007) The Polycomb Group Proteins Bind throughout the INK4A-ARF Locus and Are Disassociated in Senescent Cells. *Genes Dev. 21*, 525−530.
(61) Maertens, G. N., El Messaoudi-Aubert, S., Racek, T., Stock, J. K., Nicholls, J., Rodriguez-Niedenfuhr, M., Gil, J., and Peters, G. (2009) Several Distinct Polycomb Complexes Regulate and Co-Localize on the INK4a Tumor Suppressor Locus. *PLoS One 4*, No. e6380.
(62) Creppe, C., Palau, A., Malinverni, R., Valero, V., and Buschbeck, M. A. (2014) Cbx8-Containing Polycomb Complex Facilitates the Transition to Gene Activation during ES Cell Differentiation. *PLoS Genet. 10*, No. e1004851.
(63) TerMaat, J. R., Pienaar, E., Whitney, S. E., Mamedov, T. G., and Subramanian, A. (2009) Gene Synthesis by Integrated Polymerase Chain Assembly and PCR Amplification Using a High-Speed Thermocycler. *J. Microbiol. Methods 79*, 295−300.
(64) Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Güldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kötter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C., Ward, T. R., Wilhelmy, J., Winzeler, E. a., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., and Johnston, M. (2002) Functional Profiling of the Saccharomyces Cerevisiae Genome. *Nature 418*, 387−391.
(65) Jetson, R. R., and Krusemark, C. J. (2016) Sensing Enzymatic Activity by Exposure and Selection of DNA-Encoded Probes. *Angew. Chem., Int. Ed. 55*, 9562−9566.
(66) Ten Brink, H. T., Meijer, J. T., Geel, R. V., Damen, M., Löwik, D. W. P. M., and Van Hest, J. C. M. (2006) Solid-Phase Synthesis of C-Terminally Modified Peptides. *J. Pept. Sci. 12*, 686−692.
(67) Cai, B., Kim, D., Akhand, S., Sun, Y., Cassell, R. J., Alpsoy, A., Dykhuizen, E. C., Van Rijn, R. M., Wendt, M. K., and Krusemark, C. J. (2019) Selection of DNA-Encoded Libraries to Protein Targets within and on Living Cells. *J. Am. Chem. Soc. 141*, 17057−17061.
(68) Vivoli, M., Novak, H. R., Littlechild, J. A., and Harmer, N. J. (2014) Determination of Protein-Ligand Interactions Using Differential Scanning Fluorimetry. *J. Visualized Exp. 91*, No. e51809.
(69) Draughn, G. L., Milton, M. E., Feldmann, E. A., Bobay, B. G., Roth, B. M., Olson, A. L., Thompson, R. J., Actis, L. A., Davies, C., and Cavanagh, J. (2018) The Structure of the Biofilm-Controlling Response Regulator BfmR from Acinetobacter Baumannii Reveals Details of Its DNA-Binding Mechanism. *J. Mol. Biol. 430*, 806−821.
(70) Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015) Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput. 11*, 3696−3713.
(71) Trott, O., and Olson, A. J. (2009) Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem. 31*, 455−461.
(72) Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem. 25*, 1605−1612.
(73) Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, G. A., Nakatsuji, H., Li, X., Caricato, M., Marenich, A., Bloino, J., Janesko, B. G., Gomperts, R., Mennucci, B., Hratchian, H. P., Ortiz, J. V., Izmaylov, A. F., Sonnenberg, J. L., Williams-Young, D., Ding, F., Lipparini, F., Egidi, F., Goings, J., Peng, B., Petrone, A., Henderson, T., Ranasinghe, D., Zakrzewski, V. G., Gao, J., Rega, N., Zheng, G., Liang, W., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Throssell, K., Montgomery, J. A., Jr., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Keith, T., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Millam, J. M., Klene, M., Adamo, C., Cammi, R., Ochterski, J. W., Martin, R. L., Morokuma, K., Farkas, O., Foresman, J. B., and Fox, D. J. (2016) *Gaussian 09*, Revision A.02, Gaussian, Inc., Wallingford, CT.
(74) Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., Gabdank, I., Narayanan, A. K., Ho, M., Lee, B. T., Rowe, L. D., Dreszer, T. R., Roe, G., Podduturi, N. R., Tanaka, F., Hong, E. L., and Cherry, J. M. (2016) ENCODE Data at the ENCODE Portal. *Nucleic Acids Res. 44*, D726−D732.
(75) ENCODE Project Consortium (2012) An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature 489*, 57−74.
(76) Zhu, L. J. (2013) Integrative Analysis of ChIP-Chip and ChIP-Seq Dataset. *Methods Mol. Biol. 1067*, 105−124.
(77) Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., and Green, M. R. (2010) ChIPpeakAnno: A Bioconductor Package to Annotate ChIP-Seq and ChIP-Chip Data. *BMC Bioinf. 11*, 237.

DOI: 10.1021/acschembio.9b00654
ACS Chem. Biol. 2020, 15, 112−131

Chapter 9

Virtual Screening and Optimization of Peptidomimetic Ligands for CBX6 and CBX8 Selectivity

The work presented in this chapter was produced through a team effort by James McFarlane and Katherine Krause. Katherine's notable contributions to this project include several python scripts used to automate previously developed methods outlined in Chapter 6 (SLICE). Additionally, Katherine also wrote several scripts to generate the structural metrics of binding from molecular dynamics simulations. Her work in this project is greatly appreciated. The compounds studied in this work were suggested by Fraser Hof and Natasha Milosevich whom I thank for the ongoing collaboration and useful discussions on this matter.

9.1 Introduction

Isoform selectivity between the CBX proteins as outlined in Chapter 3 is of particular interest due to the correlation each isoform has with a disease state. In the cases between isoforms such as 6 and 7, sequence similarities in regions for known binding with the canonical methyllated histone tails are different enough to rationally exploit changes on the peptide to enforce selectivity. An example of this is the work presented in Chapter 7 performed by Milosevich and coworkers [1]. In this case, an Arg9 residue present on both CBX6 and CBX8 is exploited with a salt-bridging glutamate substitution on the peptide inhibitor, ideally producing selectivity over isoforms that contain a Gln9 instead, such as CBX7 and

CBX2. However, to further increase selectivity between similar isoforms such as CBX6 and CBX8, the overall sequence similarity makes this a less trivial task. Yet, there is still evidence that selectivity between CBX6/8 can be achieved by making substitutions on the peptide ligand that interact with the β-groove region of binding [2]. Results such as these are especially confusing when we look at the sequence similarity of this region and how there there appears to be no difference in residues that would directly contact the peptide inhibitor (See Figure 9.1).

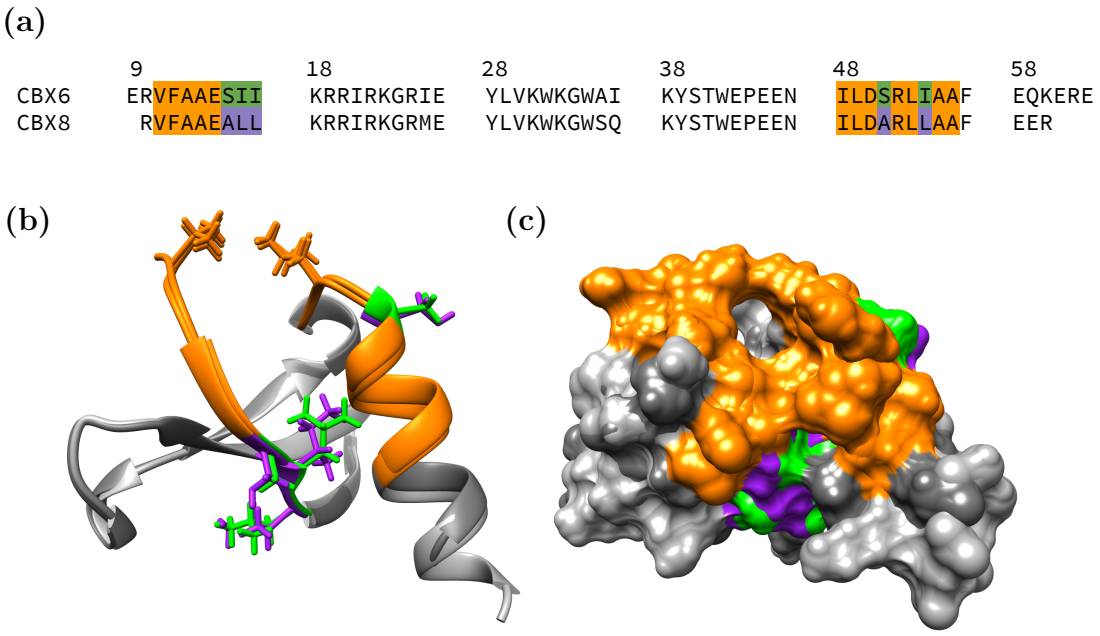


Figure 9.1: CBX8 and CBX6 β-Groove Structural Similarities. A sequence comparison (top) between the two isoforms highlights several differences in residues within the beta-groove region including two substitutions of alanine for serine, potentially adding new hydrogen bond donor interactions. However, upon inspection of the crystal structure (a) overlaid with a solvent-accessible surface (b), these residue differences are inaccessible to the static structure until much further down the extended beta-groove. Slight differences in pocket shape contributed to by the residue substitutions or other allosteric effects are still present, but are not caused by differences in side-chains within the region of binding.

To investigate the role of the β groove region in the selectivity between CBX6/8, a combinatorial library of 88 potential inhibitors shown in Figure 9.2 were constructed *in silico* under the recommendation and collaboration of the Hof group at UVic. By performing docking and MD with the library of compounds, our aims were to uncover consistent and differentiating features of the β groove region resulting from the interaction with the peptide ligands. To examine the induced-fit features of the ligands, we utilized the SLICE protocols outlined in Chapter 6. The resulting MD trajectories show the extended β-groove region between CBX6/8 can

be exploited to create an optimized hydrogen bond network but may not be the best route to the best potential binder in this region.

9.2 Methodology

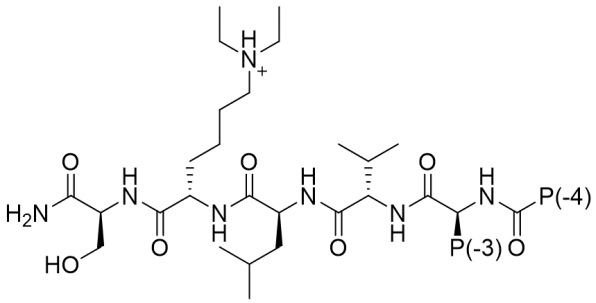
In this study, a combinatorial approach was employed to derivatize a CBX7 inhibitor described by Stuckey et al. [3]. Series of unnatural amino acids were introduced at the (–3) and (–4) positions yielding a library of 88 peptidomimetic compounds. Throughout this work, each inhibitor is denoted by its (–4) and (–3) residues (e.g. compound **1A** contains residue **1** at the (–4) position and residue **A** at the (–3) position).

Initial structures for CBX8 and CBX6 were taken from PDBs 3i91 [4] and 3GV6 [5], respectively. Bound H3K9Me$_3$ ligands and excess water molecules were removed and in the case of a dimer, the single CBX unit bound to the ligand was used. To generate the ligands, parameterization of the non-standard residues was done using the *residuegen* utility in the AmberTools16 [6] suite of programs. Initial calculations for generating the electrostatic potential maps were done using an optimization of the N- and C-capped versions of the residues in Gaussian09 using the a HF/6-31G* level of theory. Newly parameterized residues were then fit to the Amber ff14SB force field using the AmberTools *parmchk* utility. Residue connectivity information was then manually edited in the *tleap* molecular builder environment and saved for further use. A step-by-step protocol for generating the residues from methyl capped structures and electronic structure calculations are available in the supplementary information. The construction of each peptide was then done through a script that combinatorially builds the 88 peptides with the sequence command in *tleap*. Each ligand was then prepared for docking as individual PDBQT files using UCSF Chimera[7] and AutoDock Vina[8].

9.2.1 SLICE docking

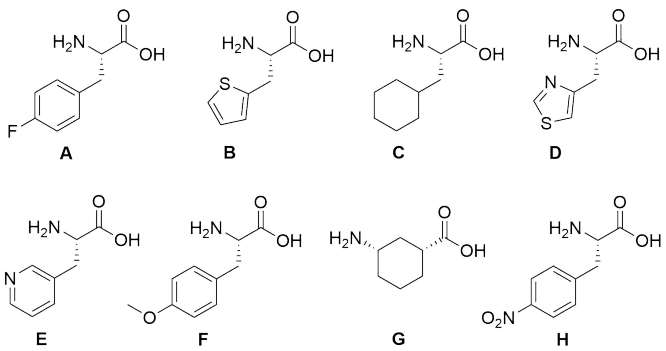
To generate individual 10 ns trajectories for every system, initial starting coordinates were generated through a round of induced-fit docking in a protocol previously described by our group called the SLICE method [9]. The method first docks on the crystal structure and a pose is selected for a round of molecular dynamics. The newly generated ensemble of host configurations through the MD trajectory is then re-docked on to yield new MD starting coordinates. The molecular docking aspect of the protocol is done using AutoDock Vina with exhaustiveness 7 and 10 poses generated per dock. Box sizes for docking were automatically generated

(a)



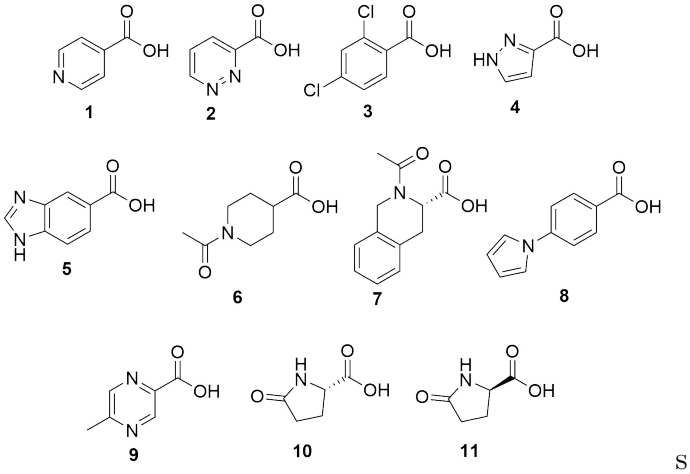
(b)

Parent Compound



P(–3) Intermediate Residues

(c)



P(–4) N-terminus Residues

Figure 9.2: Combinatorial Library for Peptide Screening against CBX6 and CBX8. The diethyl-lysine containing parent compound contains two positions that were combinatorially explored with non-standard amino acids. Position –3 is associated with the region emerging from the hydrophobic clasp into the beta-groove region whereas –4 position substitutions span into the extended beta-groove. Overall, 88 compounds were constructed *in silico*.

using a residue list to contain all residues within known contact region of the peptide plus a 10 angstrom buffer. Initial docking was done 50 times to yield 500 structures on the same host configuration. As the the next round of docking was done from MD generated coordinates, the next round produced 500 structures on 50 different partially induced-fit host configurations. The top scored poses from these docked configurations were then selected for a final 10 ns trajectory.

Molecular dynamics were done using the Amber16 suite with the ff14SB force field. All simulation coordinates were construction in the *tleap* building environment. Counter-ions were added to neutrality and the systems were solvated with approximately 8000 TIP3PBOX waters with a distance buffer of 14 Å. Simulations were minimized using gradient descent for 10,000 steps, heated to 300K over 200 ps and then run for 10 ns. Each simulation was run with a 2 fs time step with SHAKE. The systems were run using Langevin dynamics and constant pressure dynamics (ntp=2) with a Langevin thermostat (ntt = 3) and set with periodic boundary conditions. Electrostatic cutoffs were set to 8 angstroms and coordinates were set to print 50 times over 10 ns of simulation time.

9.2.2 MMPBSA.py

Binding free energy contributions were calculated using the MMPBSA.py [10] utility and was chosen for its accessibility with the AMBER suite as well as a per-residue contribution feature to explore the (–3) and (–4) residue effects. The free energy of binding according to MMPBSA.py is broken down via the following equation:

$$\begin{aligned} \Delta G_{\text{bind,solv}} = \Delta G_{\text{bind,vacuum}} \\ + \Delta G_{\text{solv,complex}} \\ - (\Delta G_{\text{solv,ligand}} + \Delta G_{\text{solv,receptor}}) \end{aligned} \tag{9.1}$$

where $\Delta G_{\text{bind,solv}}$ is the total binding energy in solvent, $\Delta G_{\text{bind,vacuum}}$ is the binding energy in vacuum, $\Delta G_{\text{solv,complex}}$ is the solvation energy of the protein–ligand complex, and $\Delta G_{\text{solv,ligand}}$ and $\Delta G_{\text{solv,host}}$ are the solvation energies of the peptide ligand and host protein, respectively.

Parameters for the MMPBSA.py calculation were set to include Poisson-Boltzmann implicit solvation with an istrng value of 0.100 and a per-residue breakdown. Structural coordinates for the host and ligand of each system were extracted from the complex during the 10 ns MD runs over 50 frames.

9.3 Results

9.3.1 Molecular Docking

Molecular docking results on the crystal structures of the host proteins shows a bias to the CBX8 isoform over CBX6 with nearly all peptides. This can be seen in Figure 9.3 in addition to visual trends for particular substitutions in both the (–4) and (–3) substitutions. The selectivity for the CBX8 crystal structure with all ligands was unexpected as the residue substitutions were chosen due to prior CBX6 selectivity shown in unpublished DNA encoded libraries by Krusemark et al.

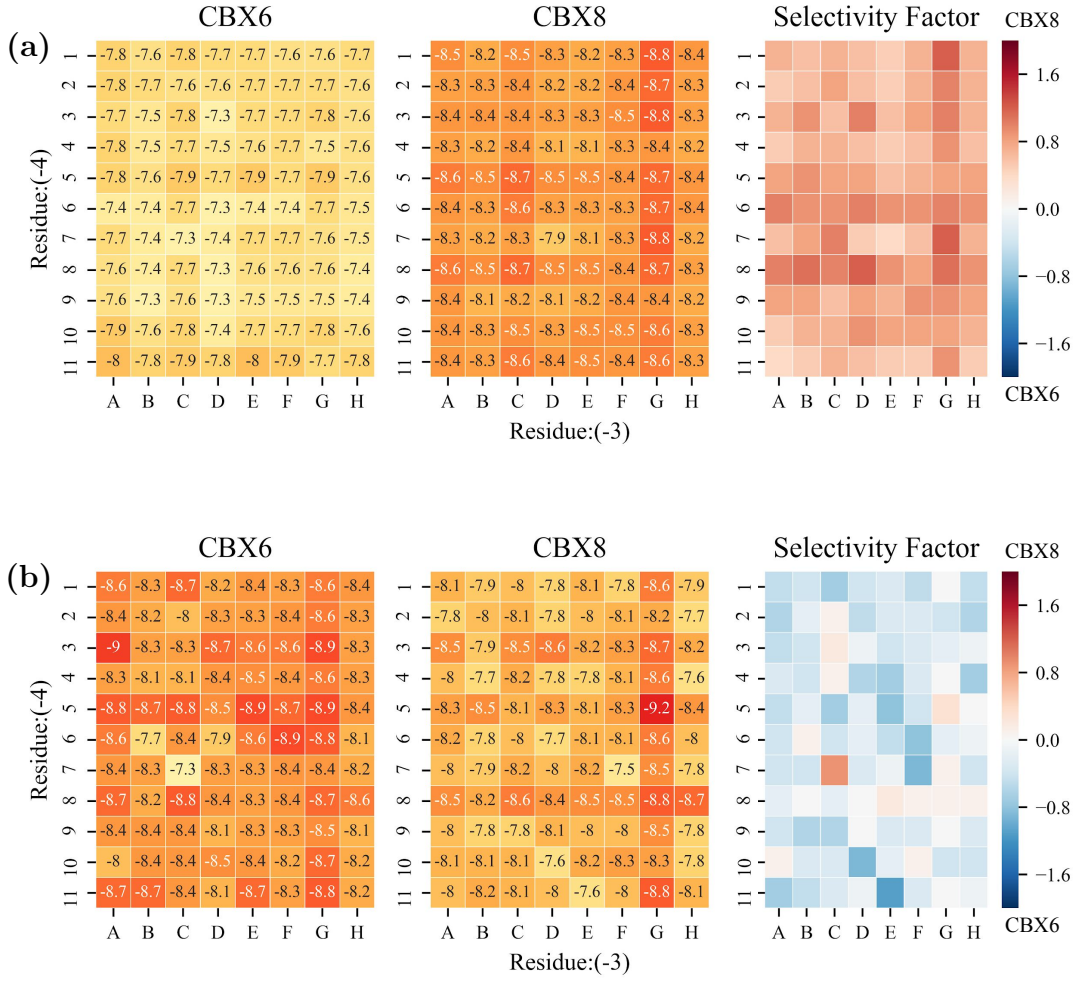


Figure 9.3: CBX6/8 Docking Results before and after SLICE. (a) A comparison of docking scores for the peptide library on crystal structure coordinates of CBX6 and CBX8 and their calculated differences in the right panel as a selectivity factor. (b) Post-SLICE re-docking scores prior to final MD trajectories. AutoDock Vina scores are reported in units of kcal mol^{-1} and the selectivity factor on the right panel is difference between scores with red and blue values showing preference for CBX8 and CBX6, respectively.

Taking the initial docked poses of the best score and submitting them for both MD and re-docking (one iteration of SLICE) dramatically shifted the docking scores and selectivity towards CBX6. This shift in selectivity was the result of two contributing factors: (i) CBX6 scores improved with the elimination of steric clashes and (ii) CBX8 scores worsened slightly through ensemble effects. Investigation into the systematic steric clash revealed a small clash with the peptide ligand's (–1) position isoleucine. The isoleucine clashes with a CBX6 glutamate residue that spans over the binding site parallel to the hydrophobic clasp. This second clasp is illustrated in Figure 9.4 and was not present after MD relaxation for a number of possible reasons including solvation and induced-fit effects of the ligand. Similar changes in CBX8 structure during MD had the opposite effect on docking scores.

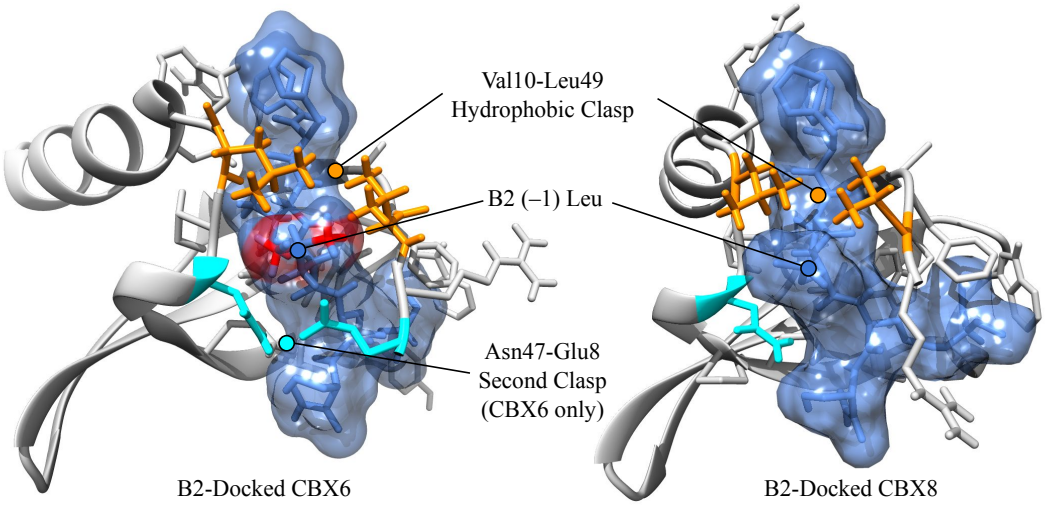


Figure 9.4: CBX6/8 Crystal Dock Steric Clash. Using compound B2 as an example, the steric clash on CBX upon docking with the crystal structure 3GV6 (left) is shown in red. Docking the same ligand on CBX8 produced no steric clash and avoiding the docking score penalty associated with it. Space in the (–2) binding pocket under the clasp was able to fully accommodate both ligands without steric clash penalties.

The worsening of the CBX8 scores is an interesting feature and is a result of the variety of host poses that are generated using the SLICE method. At first glance, the induced fit of the H3K9Me$_3$ peptide was conducive to binding the panel of substituted peptides. As the system was solvated and subjected to the substituted ligands, differences in interaction energies and points of contact from the native peptide change the preferred configuration of the host. Changes in host conformation may also be caused by solvation and the loss of crystal packing effects that may have been present—typical caveats of crystal structure docking.

However, the second docking experiment still contains the crystal structure at the start of the trajectory. As the docking process is still stochastic and occurs on each frame of the trajectory, the chance to produce the same score/pose on the crystal structure is decreased as the additional poses from the MD dilute the presence of the crystal structure coordinates in the sampling space. The extreme version of this effect would be when the MD produces a trajectory where the ligand immediately dissociates from the host protein, producing a sampling set of apo-host poses, but would still include the start of the simulation (the crystal structure coordinates) at just a small fraction of what is sampled. This systematic false positive result is a cross-docking dilemma that highlights the importance of induced-fit mechanisms and the skepticism that should be used when docking on crystal structures.

Upon visual inspection of the docked poses, every docked ligand fortunately adhered to what we would consider the canonical binding pose with the diethyllysine in the aromatic cage, the (–2) residue under the hydrophobic clasp, and the N-terminus somewhere in the β-groove region. However, the (–4) and (–3) residue orientations show two distinct poses for which some of the peptide (–3) residue ψ angles are rotated such that the (–4) and (–3) residue side-chains are switched as shown in Figure 9.5.

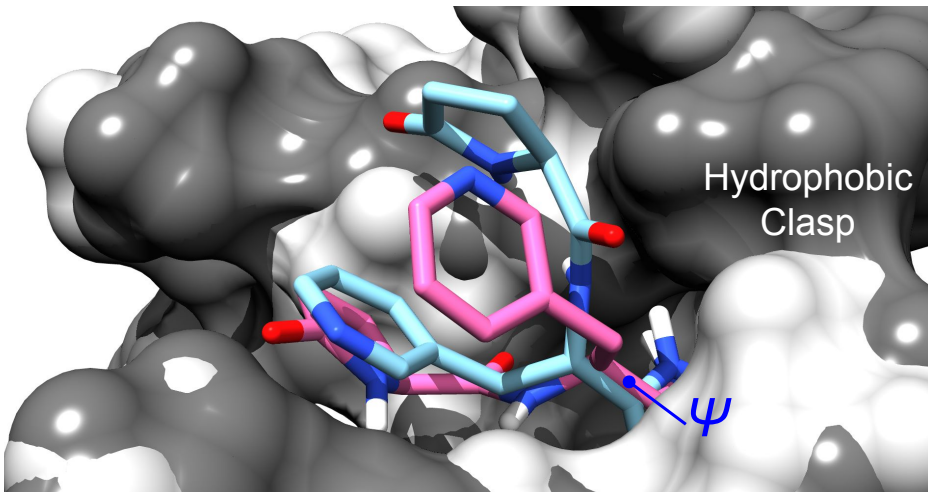


Figure 9.5: ψ-Rotated β-Groove Orientations of (–3) and (–4) Residues. Compound 10-E is used as an example show the flipped orientation of binding on CBX8 (light grey with blue ligand) and a normal linear orientation on CBX6 (dark grey pink ligand). ψ angles of the (–3) residues significantly impact the location of both (–3) and (–4) Residues side-chain locations.

The existence of a binary pose adds an extra challenge to the development of any sort of SAR for the two substitutions. For example, looking to see which (–4)

residue is able to continue the hydrogen bonding network with the host, the data contains poses where the orientation is flipped depending on which (–3) residue is present. Even though the residue is optimally designed to continue the hydrogen bonding in that region, the residue side-chain is located elsewhere. How real this binary pose problem is, the energetic barrier, and the influence on the (–4) or (–3) residue on this flipped mode of binding are ambiguous at this time. The extent of this flipped orientation in the data is presented in Figure 9.6.

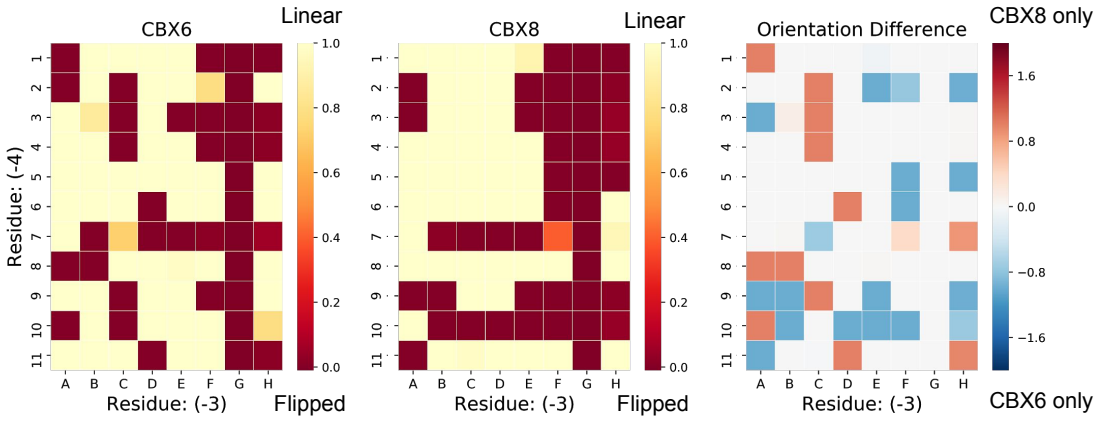


Figure 9.6: ψ-Rotated and Linear β-Groove Orientations of (–3) and (–4) Residues. Using the (–3) ψ angle as a metric for ligand orientation in the β-groove region shows similar residue orientations for both CBX6 and CBX8, but also a few distinguishing trends such as residue 10 almost entirely in the flipped position for CBX8. An example of this is shown previously in Figure 9.5 with compound 10-E. The data for residue G in this case should not be considered either position due to the unique orientation that the cyclohexyl linker incurs on the ligand and is set manually in this case.

Both docking experiments showed preference for Residue 5 with respect to CBX6 selectivity and an interesting trend involving residue G was also observed. How much attention should be paid to docking scores alone here is debatable though. As mentioned previously with the worsening of CBX8 scores, these values only portray the interaction energy between ligand and protein and avoid the energetic cost of reorganization of the host or the ligand. There is also the issue of the flipped orientations and the question of whether flipped poses should be excluded from the data when analyzing trends. Therefore the docking step shouldn't be considered a substitute for a free energy method, but more of a first step in the structural prediction for further analysis. However, seeing entire rows or columns improve is encouraging and further insight into the role that the (–4) and (–3) positions have on score and orientation can explored via MD and other free energy analyses.

9.3.2 Molecular Dynamics

Molecular dynamics of the protein–ligand complexes were monitored for structural metrics to assess the stability throughout the trajectories. The role of (–4) and (–3) position substitutions in the total binding free energy were also explored with a per-residue energy breakdown using MMPBSA.py. Other structural metrics of the (–4) and (–3) residues were monitored including both the flipped/linear orientations described previously as well hydrogen bonding in the β and extended β-region. Similar to the docking results, the theme of non-additive and downstream is seen throughout the MD data.

Our metrics for complex stability included the diethllysine–aromatic cage interaction and the inter-clasp distance (hydrophobic clasp) over the (–2) position of the ligand. No dissociations of the peptide–ligand complexes were observed and all diethyllysine residues with the exception of complex 8E/CBX6 remained within the aromatic cage pocket for the duration of the simulations. However, in monitoring the maximum inter-clasp distance between Val10 and Leu49 (Figure 9.7), several complexes appeared to have fully opened clasps (beyond 4 $\mathring{A}$) for at least some part of the 10 ns simulation time.

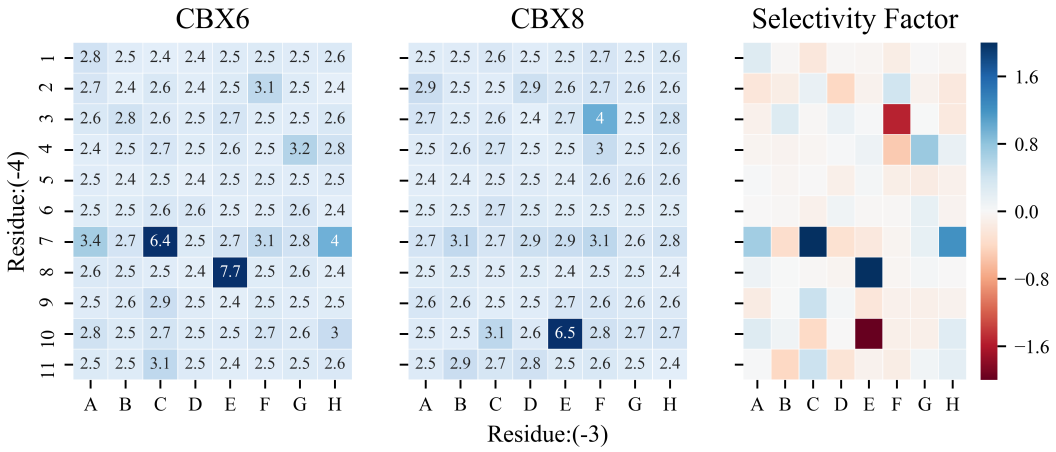


Figure 9.7: Maximum Clasp Distances. Maximum Val10 and Leu49 α-carbon distances over the 10 ns runs for each complex show several complexes had a clasp opening event at some point. One of the most notable and odd features of these plots is the possible effect of the (–4) substituent "Residue 7" on the clasp distance.

The fact that the clasp had been opened at some point may have been a result of the unique substitution combination or perhaps was a random occurrence. Multiple trajectories with longer simulation times for these particular systems may be needed to draw any conclusions regarding the effect of certain compounds on the

clasp stability. However, faint trends across the heat maps indicate that particular residues have an effect regardless of its paired substitution.

The (–4) substitution, Residue 7, appears to be responsible for the expansion of the hydrophobic clasp with all substitution combinations. Upon inspection, residue 7 is unique to the other substitutions in that there is a methyl amide present in the β position relative to the (–4)–(–3) amide bond. On top of this, the connection to the (–3) amide bond is offset whereas the other residues are para-oriented with respect to the (–3) residue. The steric influence of residue 7 is also apparent with the "flipped" orientation being most prevalent, as the side-chain cannot be accommodated in the extended β-groove. These combined factors point to a residue size and direction limit that is yet to be determined quantitatively. With what is known about the clasp/aromatic cage structural relationship discussed previously, the possibility of a downstream steric effect of the (–4) residue on the shape of the aromatic cage is yet another confounding variable in the SAR development for these compounds.

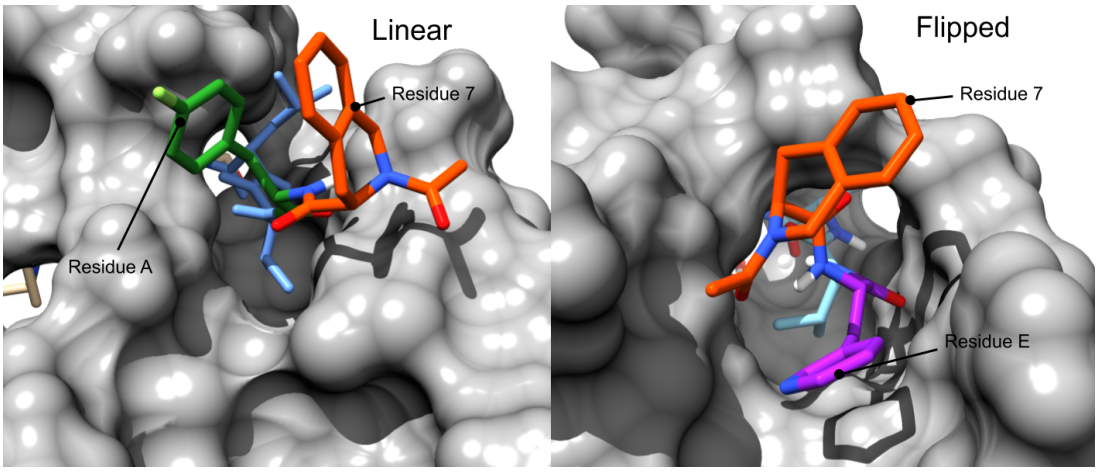


Figure 9.8: Residue 7 Steric Effects. Compounds 7A (left) and 7E (right) on CBX6 (dark grey) show the positioning of Residue 7 at the end of the MD simulation times. Compound 7A is in the linear orientation but still places the (–4) Residue 7 outside of the β groove. This orientation also causes the clasp to open for this compound. For the remaining compounds containing Residue 7, the flipped orientation is preferred and is demonstrated with 7E where the (–3) residue occupies the bottom of the β groove.

For the peptides that are in the linear position and don't present prohibitive steric clashes we can still try to understand the driving intermolecular interactions in the β-groove regions. We first hypothesized from previous data that the continuation of the hydrogen bond network would be the largest enthalpic driver of the residues in this region. This information was based on previous unpublished modelling with methyl isoxazole residues in the (–4) position that showed

the formation of hydrogen bonds with Ala12 and Ala13 on the CBX8 protein.

Analysis of this hydrogen bonding network for the library of compounds in the β and extended β-region reveal interesting relationships between the (–3) and (–4) substitutions. Figure 9.9 shows that for both isoforms, compounds containing the cyclohexyl linking peptide (G) allows for extra hydrogen bonding occupied by this (–3) position residue. Furthermore, the cyclohexyl linker has consistent downstream effects on the (–4) position as well.

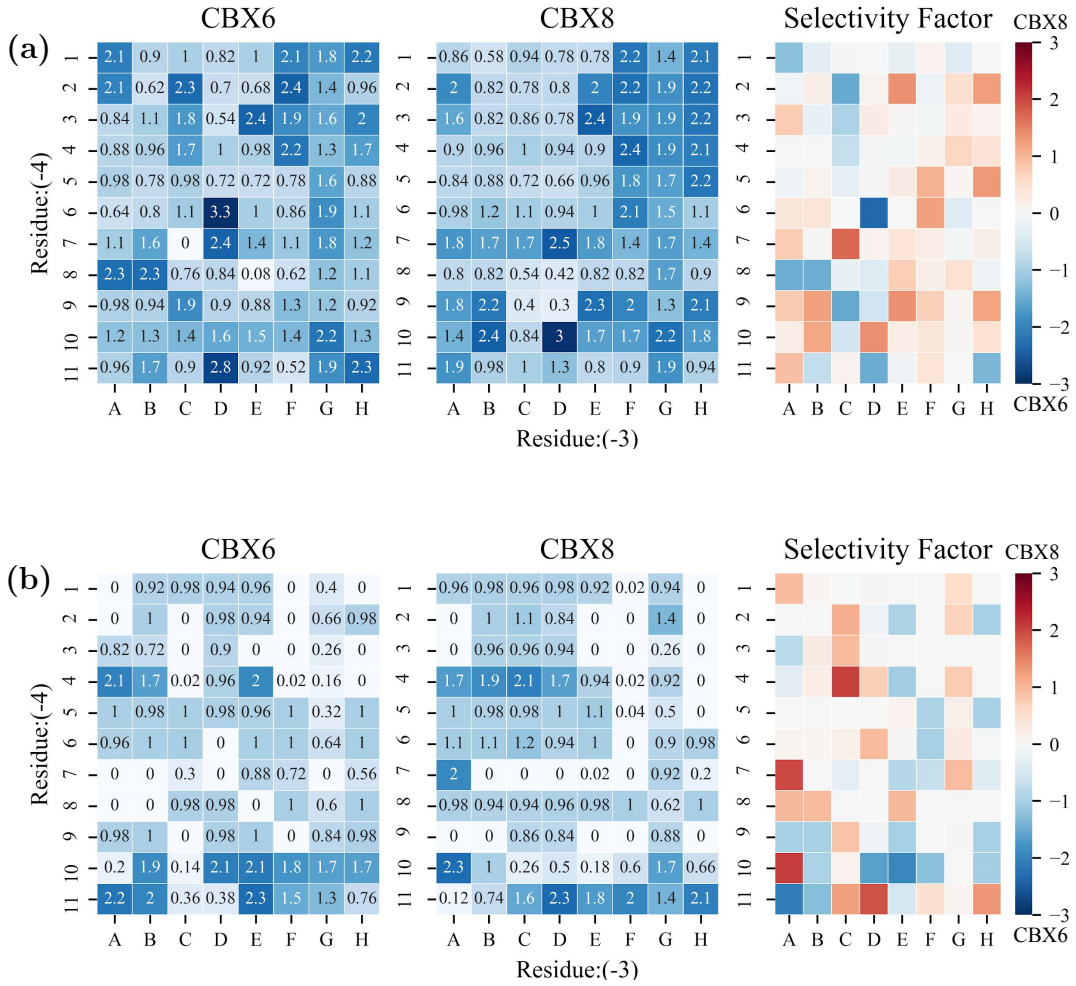


Figure 9.9: Average hydrogen bond contribution per residue for residues in (a) the –3 position and (b) –4 position. Hydrogen bond values were calculated using averages from number of hydrogen bonds formed from each of the 50 MD snapshots from the 10 ns simulations. Criteria for hydrogen bond detection was adapted from work by Harrison et al. [11]

The presence of Residue G in the (–3) position also causes CBX6 to lose hydrogen bonding ability for nearly all (–4) substitutions whereas this is not the case for the CBX8 isoform. Additionally, there is also a clear trend with residues 10 and 11, where only CBX6 is permitted to form a hydrogen bond network with

the (–4) substitution. This is also reflected in the "flipped" behaviour for the majority of Residue 10 containing compounds on CBX8. Together, this suggests the intermolecular hydrogen bond network is a potential path towards selectivity. The selectivity with compound 10 between CBX6/8 was a fortuitous result. Residue 11 was originally designed to continue the hydrogen bond network with CBX8 as shown in Figure 9.10. The addition of the L-isomer was intended to show the specific affinity caused by hydrogen bonding geometries, but surprisingly resulted in a selectivity between the isoforms.

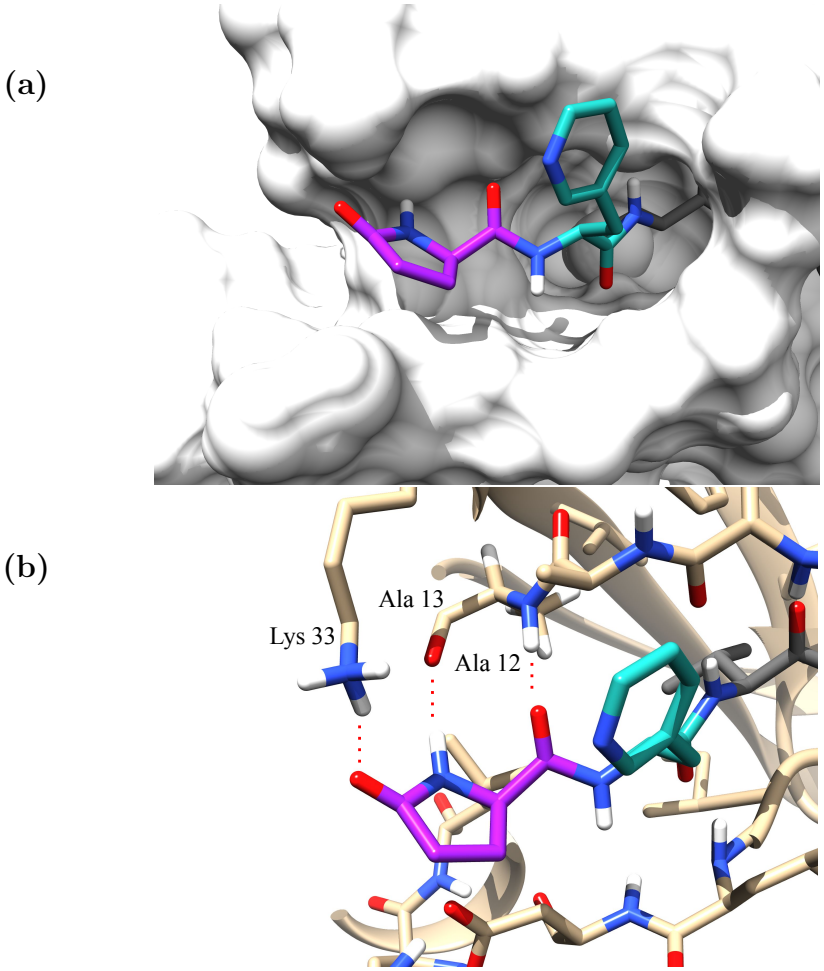


Figure 9.10: (–4) Residue Hydrogen Bonding with Compound 10-E on CBX6. The original intended geometry of the hydrogen bond network was only seen with the CBX8–10E complex. (a) The solvent-accessible surface area of the extended β-groove region occupied by Residues 10 (purple) and E (teal) shows the linear positioning of the compound with no discernible steric clashes. Panel B (stick representation) shows the intended and successfully docked hydrogen bond network with Ala13 and Ala12. Lys33 appears to also form intermittent hydrogen bonding with the Residue as well.

9.3.3 Binding Energetics

In the ideal case, docking scores, structural metrics, and total binding free energies would support each other. At the very least, enthalpic contributions of the individual residue substitutions should reflect the trends previously discussed. A per-residue breakdown of the binding free energies in Figure 9.11 seem to partly conserve the trends previously discussed.

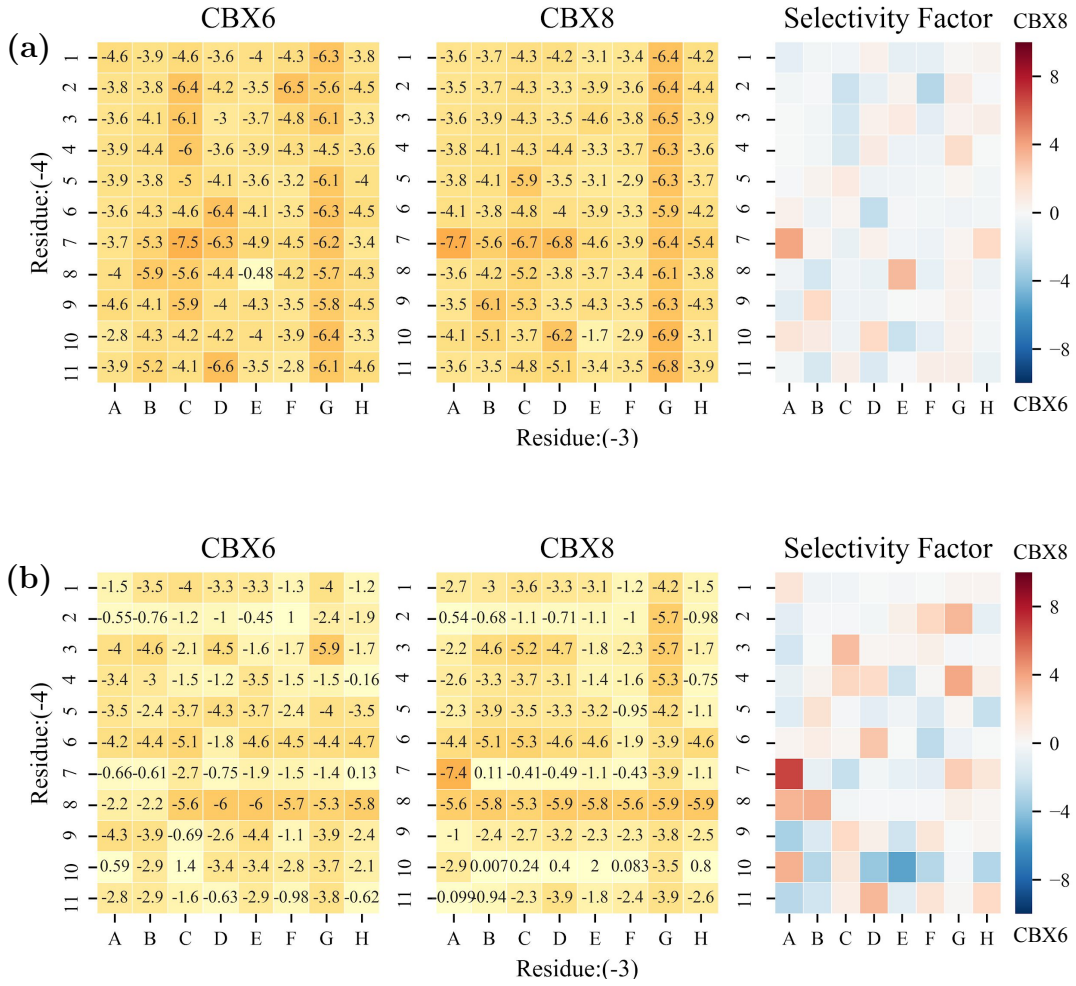


Figure 9.11: MMPBSA.py Per-Residue Binding Energies for (a) the (–3) and (b) the (–4) substitution locations. Energies are reported in kJ mol^{-1}. Selectivity factor is calculated as the difference between the contributions between the CBX8 and CBX6 systems.

Residue G remains preferred in the (–3) position but also increases the affinity towards the (–4) residue. The lack of the contribution from Residue 7 is also observed. However, the docking data is partly conflicting with the loss of Residue 5 affinity but affinity for Residue 8 remaining on both isoforms. The missing H-bond network with Residue 10 on CBX8 is also apparent in the breakdown. However, it is surprising that that magnitude of the energies are not higher for

compounds 4, 10, and 11 given the higher number of H-bonds formed. As we have seen in the past, this is likely due to desolvation penalties associated with H-bond formation with charged host residues.

Up to this point, local effects of the substitutions from a structural and energetic lens has been understandable and even consistent. However, our investigations of the total binding free energies of the complex using the MMPBSA.py single and multi-trajectory approaches yield conflicting information about which compound is the overall best binder. Some of the previously discussed trends could arguably be derived from the single trajectory panel in Figure 9.12, but the multi-trajectory approach appears to have had insufficient running parameters for any sort of meaningful calculation. On the other hand, the randomness associated with these values could also be a reflection of the variation in re-organization energies of both peptide ligand as well as the host protein.

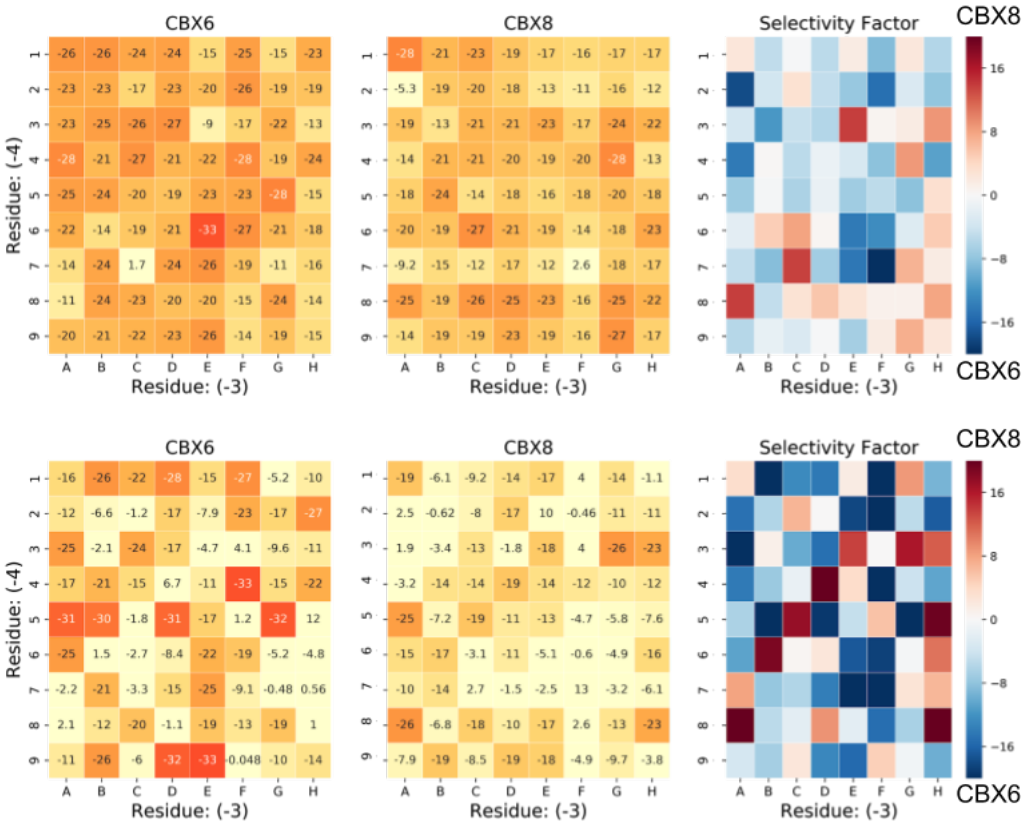


Figure 9.12: Single and Multi-Trajectory MMPBSA.py Total Binding Free Energies. Single trajectory analysis (top) shows trends similar to those of the docking scores. However, the multi-trajectory data (bottom) indicates further validation required with respect to the reference energies of the peptides as well as the hosts. Energies are reported in kJ mol^{-1}. Selectivity factor is calculated as the difference between the contributions between the CBX8 and CBX6 systems.

The concept that the peptide re-organizational cost has a large influence on total free binding energy is not surprising, but would significantly alter the approach to the rational design of future compounds. Reorganizational costs of ligands from small molecules to short peptides have shown to display energetic penalties in the 0 to 9 kcal range [12]. On top of this, the induced-fit effect of each peptide compound on the various isoforms would likely produce a unique host reorganizational cost. The purpose of the multi-trajectory MMPBSA.py approach is to capture both of these energetic changes. To successful perform this method, equilibrated states of both ligand and host are required and were likely outside of the simulation times used in this experiment.

The take-aways from the total binding free energy still support the further investigation of Residue 8 as a substitution in the (–4) position. Unfortunately, total binding free energies for both multi and single-trajectory approaches of compounds containing Residue 11 and 12 are unavailable at this time. However, from the per-residue contributions discussed earlier, we suspect that there will be little contribution to the overall total binding energies but Residue 10 may still promote a small role in isoform selectivity.

As a final metric for binding strength, we again utilized the AutoDock Vina scoring function but took the route of scoring each ligand pose during the MD trajectories (See Figure 9.13).

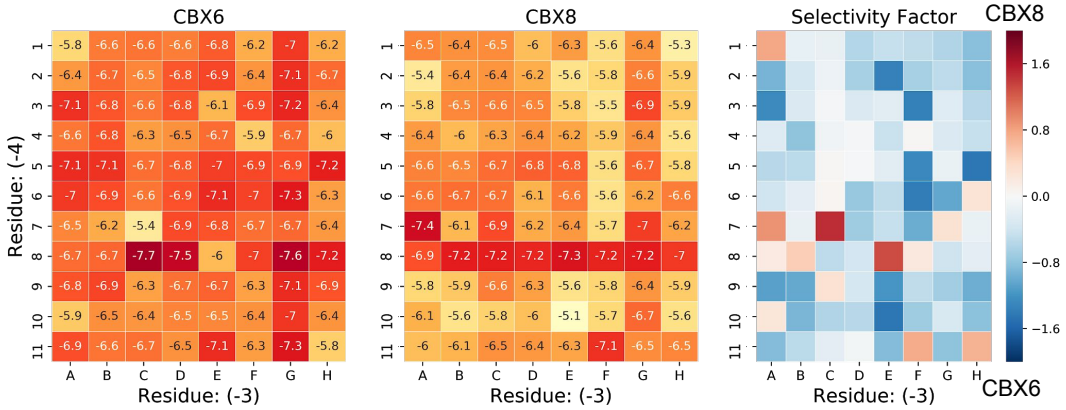


Figure 9.13: MD Frame Vina Scoring. Similar to the AutoDock Vina heatmaps previously shown, scores are shown using the AutoDock scoring function. However, the results shown here are not individual pose scores generated from the Vina seach algorithm, but instead parsed from the 10 ns trajectories used for the MMBPSA.py binding energies. The scores are averaged and the selectivity factor is shown as the difference between the scores. All values reported in kcal mol^{-1}

This is opposed to using the Vina search algorithm to place and score the ligand. The technique of pose scoring from MD snapshots is used in the Schrodinger

Induced-Fit Docking protocol [13] and aims at averaging scores over an ensemble of representative host poses. The trends seen in this MD/score method reflect those of the previous docking results as well as the free energy methods excluding the multi-trajectory approach work. This is an interesting approach that allows the interaction energy to be reported on during the MD simulation while providing a level of uncertainty to the pose scores which is unavailable in the regular usage of AutoDock Vina.

9.4 Future Considerations

Throughout the series of data collected, Residue 8 was consistently scored better than the other residues in the (–4) position yet was not part of the hydrogen bonding network initially assumed to be the driver for binding affinity in the extended β-groove region. Inspection of one of the docked poses (Figure 9.14) on CBX8 and the per-residue free energy contributions indicate that preference for the residue is caused by a number of factors.

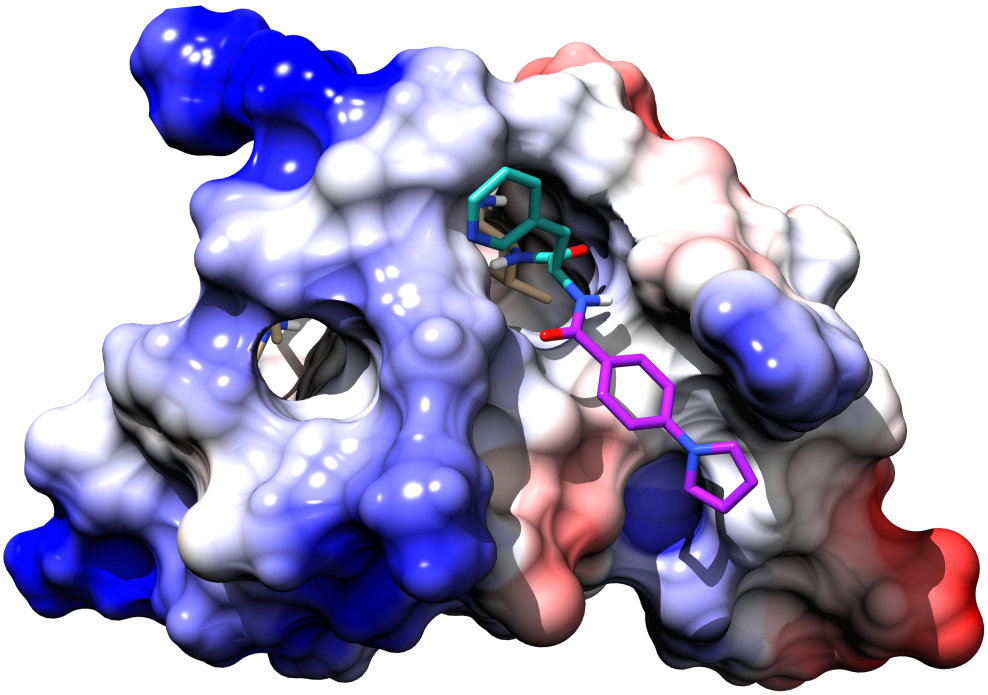


Figure 9.14: CBX8/Compound 8E Complex. Compound 8E showed consistently better docking scores as well as free energy contributions. Analysis of the docked structures indicate the n-phenylpyyrole group is situated on the bottom of the β-groove region. Despite the lack of hydrogen bonding contributed by the residue, the substitution is a promising jumping off point as a new potential scaffold.

Firstly, the location of the phenyl group is situated on the bottom of the extended β-groove region without causing a need for the upstream peptide backbone to compromise its existing hydrogen bond network. Secondly, the pyrrole group continues outwards with no steric clash. Together, the overall fit of the ligand is also coupled with modest per-residue contributions to van Der Waals (-6.3 kJ/mol) and electrostatic (-1.0 kJ/mol) terms with little penalty to solvation energies (+3.66 kJ/mol). There appears to not be one single driver for the success of this residue, but more a combination of relatively small enthalpic contributions while also lacking any significant penalty within its binding zone as well as upstream in the remainder of the ligand.

From the geometries of the docked poses, it appears that the 2-position of the pyrrole group could be modified to exploit what looks like a continued cavity in the extended β-groove region. However, the negative and positive regions near the positions are caused by a glutamate and arginine residue, respectively. Placement of a polar substitution such as an acetamide group in this region could potentially form favourable electrostatics with both charged regions, but at the cost of a much larger desolvation enthalpy penalty. However, it would be interesting to explore this substitution region on the 8E backbone as a more clear reference point in comparison to the diverse panel currently explored.

Regarding the initial interest in the hydrogen bonding network, we were able to rationally design two substitutions and to exploit this feature as well as observe selectivity between the isoforms. However, we have now seen that the presence of the various charged residues in the region, although capable of forming hydrogen bonds with the (–4) substitution, do not significantly increase the binding strength of the substitution likely again due to the solvation penalty incurred.

9.5 Conclusions

The virtual screen of 88 compounds exploring combinations of residue substitutions in the extended β and β-groove region was conducted through a combination of MD and molecular docking. Throughout the screen, the theme of non-additive effects between substitutions prevailed as well as the implications of downstream effects as far as the aromatic cage. The extent of these effects appear to be unpredictable from a ligand-based perspective and were only able to be inferred after monitoring of the MD simulations. From the screen, important lessons were still learned regarding the limitations of substitutions. Such examples include the use of Residue 4 and the consistent destabilization of the hydrophobic clasp associated

with its use. The relative size of the (–4) and (–3) residue also appear to cause a potential switching made possible by the (–3) ψ-angle rotations. This feature warrants further investigation via MD and crystallography.

Compounds of interest going forward include those containing the N-phenylpyrrole Residue 8. Despite not continuing the hydrogen bonding network originally focussed on, the residue appears to satisfy steric requirements without incurring a large desolvation penalty as most of the hydrogen bonding compounds such as those containing Residues 4,10,11 likely did. Residue G is also an interesting (–3) substitution in that one of its features included increasing the binding in the (–4) position. This is the most consistent downstream effect observed that can be positively exploited. This begs the question of how to further add variation to the (–2) to (–4) connection.

From a methodological perspective, several lessons were also learned. The initial use of crystal structures for docking showed the importance of protein flexibility and ensemble-based techniques to overcome large systematic errors caused by small steric clashes. Regarding the MMPBSA.py single and multi-trajectory approaches, useful information was gleaned from the per-residue breakdown of the single trajectory approach, but the multi-trajectory approach yielded questionable results. The multi-trajectory approach likely required longer equilibration times for increased accuracy, but at the same time, alerted us to the possibility of significant ligand reorganizational costs that were not accounted for in the docking or single-trajectory method.

Overall, this virtual screening offers jumping-off points regarding possible exploitable regions and further substitutions. Furthermore, the screen validates of the usefulness of the methods previously developed to handle such protein–ligand systems. However, it is even more clear now that the investigation of substituted peptides on these flexible hosts should not fall under a classical SAR model given the diversity of substitutions and effects but require full dynamics and energetic analysis to be properly discussed.

Bibliography

[1] Natalia Milosevich, James McFarlane, Michael C. Gignac, Janessa Li, Tyler M. Brown, Chelsea R. Wilson, Lindsey Devorkin, Caitlin S. Croft, Rebecca Hof, Irina Paci, Julian J. Lum, and Fraser Hof. Pan-specific and partially selective dye-labeled peptidic inhibitors of the polycomb paralog proteins. *Bioorganic & Medicinal Chemistry*, 28(1):115176, January 2020.

[2] Sijie Wang, Kyle E. Denton, Kathryn F. Hobbs, Tyler Weaver, James M. B. McFarlane, Katelyn E. Connelly, Michael C. Gignac, Natalia Milosevich, Fraser Hof, Irina Paci, Catherine A. Musselman, Emily C. Dykhuizen, and Casey J. Krusemark. Optimization of ligands using focused DNA-encoded libraries to develop a selective, cell-permeable CBX8 chromodomain inhibitor. *ACS Chemical Biology*, 15(1):112–131, November 2019.

[3] Jacob I Stuckey, Bradley M Dickson, Nancy Cheng, Yanli Liu, Jacqueline L Norris, Stephanie H Cholensky, Wolfram Tempel, Su Qin, Katherine G Huber, Cari Sagum, Karynne Black, Fengling Li, Xi-Ping Huang, Bryan L Roth, Brandi M Baughman, Guillermo Senisterra, Samantha G Pattenden, Masoud Vedadi, Peter J Brown, Mark T Bedford, Jinrong Min, Cheryl H Arrowsmith, Lindsey I James, and Stephen V Frye. A cellular chemical probe targeting the chromodomains of polycomb repressive complex 1. *Nature Chemical Biology*, 12(3):180–187, jan 2016.

[4] A. Dong, M.F. Amaya, Z. Li, P. Loppnau, I. Kozieradzki, A.M. Edwards, C.H. Arrowsmith, J. Weigelt, C. Bountra, A. Bochkarev, J. Min, and H. Ouyang and. Crystal structure of human chromobox homolog 6 (CBX6) with h3k9 peptide, April 2009.

[5] M.F. Amaya, M. Ravichandran, P. Loppnau, I. Kozieradzki, A.M. Edwards, C.H. Arrowsmith, J. Weigelt, C. Bountra, A. Bochkarev, J. Min, and H. Ouyang and. Crystal structure of human chromobox homolog 8 (CBX8) with h3k9 peptide, September 2009.

[6] S.R. Brozell D.S. Cerutti T.E. Cheatham III V.W.D. Cruzeiro T.A. Darden R.E. Duke D. Ghoreishi M.K. Gilson H. Gohlke A.W. Goetz D. Greene R Harris N. Homeyer S. Izadi A. Kovalenko T. Kurtzman T.S. Lee S. LeGrand P. Li C. Lin J. Liu T. Luchko R. Luo D.J. Mermelstein K.M. Merz Y. Miao G. Monard C. Nguyen H. Nguyen I. Omelyan A. Onufriev F. Pan R. Qi D.R. Roe A. Roitberg C. Sagui S. Schott-Verdugo J. Shen C.L. Simmerling J. Smith

R. Salomon-Ferrer J. Swails R.C. Walker J. Wang H. Wei R.M. Wolf X. Wu L. Xiao D.M. York D.A. Case, I.Y. Ben-Shalom and P.A. Kollman. Amber 2016, university of california, san francisco, September 2016.

[7] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF chimera?a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[8] Oleg Trott and Arthur J. Olson. AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, pages NA–NA, 2009.

[9] James M. B. McFarlane, Katherine D. Krause, and Irina Paci. Accelerated structural prediction of flexible protein–ligand complexes: The SLICE method. *Journal of Chemical Information and Modeling*, 59(12):5263–5275, November 2019.

[10] Bill R. Miller, T. Dwight McGee, Jason M. Swails, Nadine Homeyer, Holger Gohlke, and Adrian E. Roitberg. MMPBSA.py: An efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation*, 8(9):3314–3321, August 2012.

[11] Ivan Y. Torshin, Irene T. Weber, and Robert W. Harrison. Geometric criteria of hydrogen bonds in proteins and identification of `bifurcated' hydrogen bonds. *Protein Engineering, Design and Selection*, 15(5):359–363, May 2002.

[12] Emanuele Perola and Paul S. Charifson. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of Medicinal Chemistry*, 47(10):2499–2510, May 2004.

[13] Edward Miller, Robert Murphy, Daniel Sindhikara, Ken Borrelli, Matthew Grisewood, Fabio Ranalli, Steven Dixon, Steven Jerome, Nicholas Boyles, Tyler Day, Phani Ghanakota, Sayan Mondal, Salma B. Rafi, Dawn M. Troast, Robert Abel, and Richard Friesner. A reliable and accurate solution to the induced fit docking problem for protein-ligand binding. March 2020.

Chapter 10

Conclusions

Through several experimental studies coupled with theoretical investigations, the general theme of protein flexibility emerged as both an obstacle as well as a driver in method development in our structural prediction efforts. Arguments such as "induced-fit" versus "conformational selection" or the magnitude of internal host energy changes upon binding became important in how theoretical work was conducted as well as how results were interpreted. The need for new tools and methods to conduct work on flexible protein systems became apparent during the course of this research and we have made an earnest attempt to bring something to the computer-assisted drug design community that may be used to help deal with such flexible protein–ligand systems.

The model systems studied throughout this dissertation primarily focussed on epigenetic reader proteins, and specifically the CBX methyllysine readers. Along the way, features of CBX peptide binding were discovered and explored with maybe more new questions than answers. The CBX proteins as targets for inhibition consistently presented unique challenges regarding dynamic temporary structures and potential intermolecular interactions not deemed structurally possible when investigating the crystal structures alone. Binding features and regions such as the (–2) binding pocket, the hydrophobic clasp, (+2) salt-bridging interactions, and the extended β-groove regions have been exposed to fall under this dynamic binding scheme.

Furthermore, investigations into selectivity between the CBX isoforms added another layer of difficulty in that the isoforms present high sequence similarity in regions which binding occurs. We have learned that non-additive SAR features and allosteric effects play a significant role in binding when modified histone tail peptides are used as a base structure for ligand design. The unique challenges posed by the CBX proteins eventually led to the design of a new structural pre-

diction method (SLICE) that may now be extended to other systems involving similar challenges.

Throughout this work, the simulation technique, SLICE, was a constant work-in-progress evolving from a full manual execution to near full automation using cluster computing and a variety of python scripts. Work on this method will continue with further validation, optimization, and automation to eventually provide access to researchers outside of the Paci Group. The success of the SLICE method to accelerate the prediction of induced-fit structures provides support for interesting concepts relating to the potential energy surfaces of protein–ligand binding such as the concept of binding funnels—analogous to protein folding. The method also highlights the usefulness of deterministic and stochastic combinations as well as the timeframes required for useful structural changes to occur on proteins in induced-fit mechanisms. However, it is still clear that how the method is applied is still very system specific, and a way to generate the running parameters for the program standardized to some feature of the system being studied is still required. While there is much investigation remaining on the parameter space for the SLICE method, it offers a novel iterative approach with a demonstrated benefit in the exploration of protein–ligand binding energy landscapes.

In summary, the role of the CBX proteins in this research remains more of an inspirational challenge than a fully elucidated drug target for inhibitor design. Many new insights into the nature of CBX–ligand interactions were gleaned post-facto but leave many questions unanswered. However, the lessons/skills learned and methods developed to tackle these difficult systems provide an added benefit to this research and have given me a great respect for the pioneers in the fields of computer-assisted drug design and structural biology. Most importantly, the insight gained into the current developments in these fields has given me a profound optimism and confidence with respect to their future impact on human health.

Appendix A

Non-standard Residue Parameterization

A.1 Pre-Amber

This assumes the amino acid is an intermediate residue containing connections both at the C and N-termini to the rest of the peptide ligand. The residue for this tutorial will be called SE1.

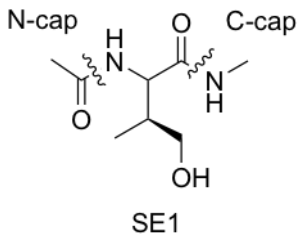


Figure A.1: Intermediate Residue SE1 Example

1) Build the desired residue in Avogadro and attach N-cap and C-caps to the molecule as shown in Fig. A.1.

2) Prepare the Gaussian input file. On the tool bar, click:

Extensions → Gaussian → Generate

This will save a .com file containing xyz and headers for a Gaussian input. Edit the new .com file to contain the appropriate header with the following format. Save the new file as a .gjf

Default Gaussian Input File:

```
n B3LYP/6-31G(d) Opt

Title

0 1
C -1.60303 -1.29234 -0.71386
C -1.43836 1.20905 -0.39070
 ...
```

Edited Gaussian Input File:

```
chk=SE1.chk
#HF/6-31G* SCF=tight Test Pop=MK iop(6/33=2) iop(6/42=6) opt

SE1

0  1
C -1.60303 -1.29234 -0.71386
C -1.43836 1.20905 -0.39070
 ...
```

3) Run Gaussian. Transfer .gjf file to new directory into an environment that can run Gaussian? etc. Make .scr file and submit optimization.

Example submission script:

```
#!/bin/csh
#
#$ -M [insert email]

# When to notify the user.
#$ -m es

# Execute the job from the current working directory
#$ -cwd
#
set echo

echo ‘pwd‘
set currentdir = ‘pwd‘
echo $currentdir
set scratch=/state/partition1
echo ’Running on ‘hostname‘’
set nameroot = SE1

#create scratch directory
mkdir $scratch/$JOB_ID

#set up Gaussian scratch
set GAUSS_SCRDIR=$scratch/$JOB_ID

cp * $scratch/$JOB_ID
cd $scratch/$JOB_ID

g09 < $nameroot.gjf> $nameroot.out

cp * $currentdir

echo "Removing scratch directory"
rm -r $scratch/$JOB_ID
```

4) Generate ESP. Example for Amber 16. Type into command line where gaussian output is present:

```
> module load amber/16
> espgen -i SE1.out -o SE1.esp
```

This will now generate an electrostatic potential file for the residue containing the caps. This file will serve as an input for a restricted electrostatic potential map to just the intermediate residue without the caps in the next step.

5) Using the Residuegen Utility and Partial Charge Assignment

5.1) Call Antechamber to prepare .ac file by typing:

```
> antechamber -at amber -i SE1.out -fi gout -fo ac -o SE1.ac
```

This will provide a base .ac file containing atom types, names, and initial parameters based off the amber force field. The input for this file is the SE1.out, the Gaussian output file from the previous section.

5.2) Create or edit a resgen.in file and execute the resiudegen utility.

This file will be the input for the residuegen utility. The purpose of this utility is to create a restricted electrostatic potential map using the newly created SE1.esp file along with the partial charge assignment for the N and C-terminal caps.

The file "resgen.in" should appear as follows:

```
INPUT_FILE      SE1.ac          (made in 5.1)
CONF_NUM        1
ESP_FILE        SE1.esp     (made in step 4)
SEP_BOND        N1 C11
SEP_BOND        C2 N2
ATOM_CHARGE     C11  0.5972
ATOM_CHARGE     O2 -0.5679
ATOM_CHARGE     C12 -0.3662
ATOM_CHARGE     H12 0.1123
ATOM_CHARGE     H13 0.1123
ATOM_CHARGE     H14 0.1123
ATOM_CHARGE     N2 -0.4157
ATOM_CHARGE     H15 0.2719
ATOM_CHARGE     C10 -0.1490
ATOM_CHARGE     H8 0.0976
ATOM_CHARGE     H9 0.0976
ATOM_CHARGE     H10 0.0976
NET_CHARGE      0
PREP_FILE       SE1.prepi  (output)
RESIDUE_FILE_NAME       SE1.res
RESIDUE_SYMBOL  SE1
```

For the partial charges, opening the Gaussian output in Avogadro and turning atom names on is helpful for this step.

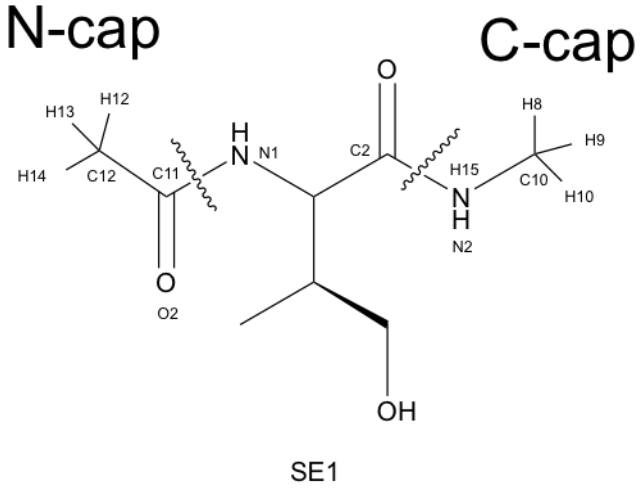


Figure A.2: Intermediate Residue SE1 Partial Charge Legend

6) Check new .prepi file for missing parameters and build supporting parameter file. Type the follow command into the terminal:

```
> parmchk -i SE1.prepi -f prepi -o SE1.frcmod
```

7 Use xleap environment to edit new residue connectivity and permanently save as a new residue that can be used to construct peptides. Open xleap (Amber module must be loaded at this point)

```
> xleap
```

The following commands need to be typed into the xleap command line:

```
> source build.scr'  (This assumes build.scr is in the directory )
build.scr is a file that contains the start up loading of all previously
parameterized residues as well as the
base AMBER forcefield for proteins and biomolecules.

> loadamberprep SE1.prepi
> loadamberparams SE1.frcmod
> desc SE1.1          (This will give you the atom legend)
> edit SE1       * click show names on the drop down menu*
> set SE1 head SE1.1.nitrogen_connection_atom_number
> set SE1.1 connect0 SE1.1.nitrogen_connection_atom_number
> set SE1 tail SE1.1.carbon_connection_atom_number
> set SE1.1 connect1 SE1.1.carbon_connection_atom_number
> check SE1
> ligand = sequence{ build whatever you want, make sure to throw SE1 in there}
> check ligand (You might see some close contacts but no missing parameters)
> saveoff SE1 SE1.dat
> savepdb ligand ligand.pdb
```

Appendix B

SLICE Configuration File Example

```
#This file controls the parameters for the execution of the SLICE method.
Placing this file in the running directorywill override the default settings
in the installation directory.

[System]
ligand_name = RBP_ligand.pdbqt (Ligand .pdbqt file including active torsions)
host_name = RBP_receptor.pdbqt (Host .pdbqt file)

[General]
SLICE_num=1 (Number of SLICE Iterations)
Replicates_skim=5 (Number of docked poses selected by highest rank)
Replicates_rose=5 (Number of Rosenbluth scheme poses)

[Docking]  (AutoDock Vina Configuration Input Section)
Exhaustiveness=7
num_poses=10
box_buffer=1
box_residues=[12 11 25 42 22 19 1] (Host residues in docking search box)
cpu=7

[Minimization]
Run_MIN = True (Run minimization before MD)

[Heating]
Run_HEAT = True (Run temperature ramping before MD)

[Equilibration]
Run_EQUIL = False (Run pre-equilbration before MD)
```

```
###Config.ini Continued ###

[Dynamics Production]
simulation_time = 10 (Nanoseconds of MD production time)
output_frames = 50 (Number of MD frames for redocking)

[Paths] (Path locations for executables/input files)
VINA = "/Path"
SANDER = "/Path"
CPPTRAJ = "/Path"
TLEAP = "/Path"
TLEAP_SOURCE= /storage/home/jmbm87/SLICE_dev/materials/scrs/build.scr
MIN_script = /storage/home/jmbm87/SLICE_dev/materials/scrs/MIN.scr
TEMP_script = /storage/home/jmbm87/SLICE_dev/materials/scrs/HEAT.scr
PRES_script = /storage/home/jmbm87/SLICE_dev/materials/scrs/PRES.scr
PROD_script = /storage/home/jmbm87/SLICE_dev/materials/scrs/PROD.scr
```

Appendix C

SLICE Execution and Application

To submit a SLICE calculation, AMBER topology files such as .dat or .off files must be prepared as per AmberTools16 documentation and referenced in the build.scr file used to load the environment in tleap. See Appendix 1 regarding the parameterization of non-standard residues for instructions on how to do this. Ligand and Host pdbqt files can be prepared by removing extraneous PDB information, loading into UCSF Chimera and executing AutoDock Vina. The PDBQT files will be generated in the user-defined directory and can be transferred to the executable directory for the SLICE method.

Execution of SLICE is currently limited to the co-installations of Amber16 and AutoDock Vina and a SLURM PBS submission system. The SLICE.py executable can be remotely executed by the following command.

```
> python installation_directory/SLICE.py
```

All flags and input parameters are controlled by the config.ini file in the executable directory. If no .ini file is present, it will default to the .ini in the installation directory. The execution will create a file directory as illustrated in the next appendix. To download the current version of SLICE, files can be copied from James McFarlane's public git repo located at the following URL: https://github.com/JMB-McFarlane/SLICE

Appendix D

SLICE Development

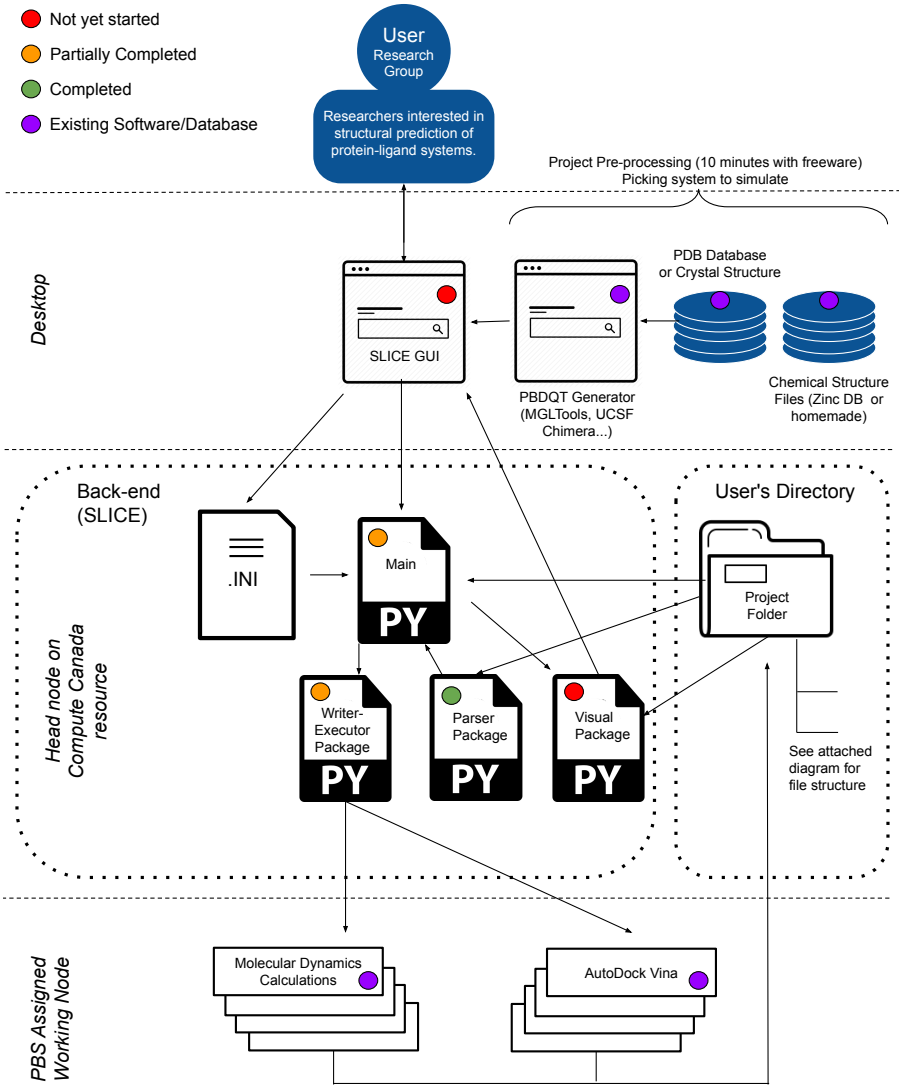


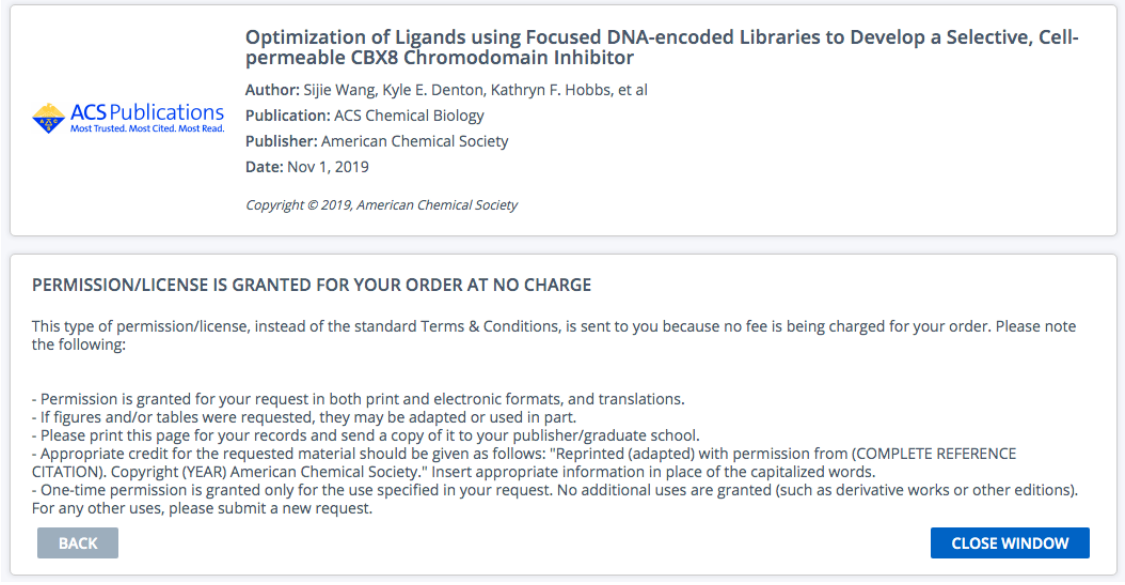
Figure D.1: SLICE Software Architecture Design. Current software architecture design for the SLICE utility package for future development.

```
Project (single ligand/host system)
|
|-------Results_File (Current status of the calculation and appended final results)
|
|-------First_Dock
|       |
|       |       [projectname].ligand.pdbqt
|       |       [projectname].receptor.pdbqt
|       '-------1
|               |
|               |
|               '-------PROD
|                               [projectname].ligand.pdbqt
|                               [projectname].receptor.pdbqt
|                               OUT.pdb.1 (Cleaned host pdbs) ...
|
|-------SLICE_1
|       |
|       |-------1
|       |       |
|       |       |-------MIN
|       |       |               MIN.in
|       |       |               pep.top (Topology for MD)
|       |       |               pose.crd (Initial MD coordinates)
|       |       |
|       |       |-------HEAT
|       |       |               HEAT.in
|       |       |
|       |       '-------PROD
|       |                       [projectname].ligand.pdbqt
|       |                       [projectname].receptor.pdbqt
|       |                       OUT.pdb.1 (Cleaned host pdbs) ...
|       |                       PROD.in (MD input for trajectory)
|       |                       mdcrd (Trajectory for MD)
|       |-------2
|       |       |...
|       |...
|
|-------SLICE_2
|       |
|       |...
|
|...
```

Figure D.2: SLICE File Structure. File system for a SLICE project.

Appendix E

Rights and Permissions



ACS Publications
Most Trusted. Most Cited. Most Read.

Optimization of Ligands using Focused DNA-encoded Libraries to Develop a Selective, Cell-permeable CBX8 Chromodomain Inhibitor

Author: Sijie Wang, Kyle E. Denton, Kathryn F. Hobbs, et al
Publication: ACS Chemical Biology
Publisher: American Chemical Society
Date: Nov 1, 2019

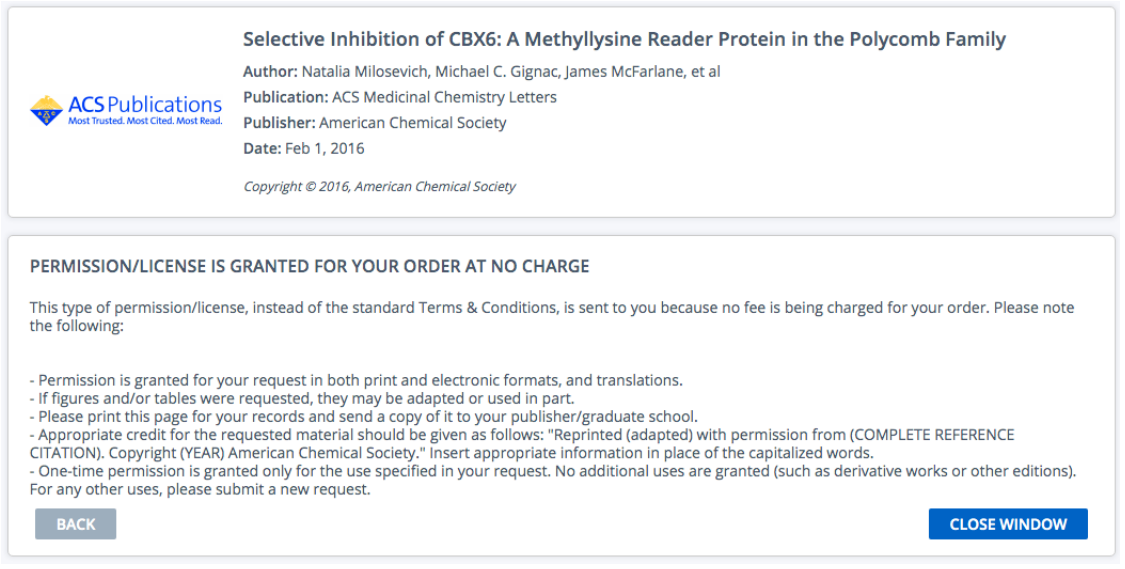
Copyright © 2019, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK CLOSE WINDOW



ACS Publications
Most Trusted. Most Cited. Most Read.

Selective Inhibition of CBX6: A Methyllysine Reader Protein in the Polycomb Family

Author: Natalia Milosevich, Michael C. Gignac, James McFarlane, et al
Publication: ACS Medicinal Chemistry Letters
Publisher: American Chemical Society
Date: Feb 1, 2016

Copyright © 2016, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK CLOSE WINDOW

ACS Publications
Most Trusted. Most Cited. Most Read.

Accelerated Structural Prediction of Flexible Protein–Ligand Complexes: The SLICE Method

Author: James M. B. McFarlane, Katherine D. Krause, Irina Paci
Publication: Journal of Chemical Information and Modeling
Publisher: American Chemical Society
Date: Nov 1, 2019

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK CLOSE WINDOW

ACS Publications
Most Trusted. Most Cited. Most Read.

An Information-Rich Graphical Representation of Catalytic Cycles

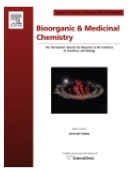
Author: James McFarlane, Brett Henderson, Sofia Donnecke, et al
Publication: Organometallics
Publisher: American Chemical Society
Date: Nov 1, 2019

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK CLOSE WINDOW



Pan-specific and partially selective dye-labeled peptidic inhibitors of the polycomb paralog proteins

Author:
Natalia Milosevich,James McFarlane,Michael C. Gignac,Janessa Li,Tyler M. Brown,Chelsea R. Wilson,Lindsey Devorkin,Caitlin S. Croft,Rebecca Hof,Irina Paci,Julian J. Lum,Fraser Hof

Publication: Bioorganic & Medicinal Chemistry
Publisher: Elsevier
Date: Available online 9 November 2019

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW

Dear James

The Royal Society of Chemistry (RSC) hereby grants permission for the use of your paper(s) specified below in the printed and microfilm version of your thesis. You may also make available the PDF version of your paper(s) that the RSC sent to the corresponding author(s) of your paper(s) upon publication of the paper(s) in the following ways: in your thesis via any website that your university may have for the deposition of theses, via your university's Intranet or via your own personal website. We are however unable to grant you permission to include the PDF version of the paper(s) on its own in your institutional repository. The Royal Society of Chemistry is a signatory to the STM Guidelines on Permissions (available on request).

Please note that if the material specified below or any part of it appears with credit or acknowledgement to a third party then you must also secure permission from that third party before reproducing that material.

Please ensure that the thesis states the following:

Reproduced by permission of The Royal Society of Chemistry

and include a link to the paper on the Royal Society of Chemistry's website.

Please ensure that your co-authors are aware that you are including the paper in your thesis.

Regards

Gill Cockhead
Publishing Contracts & Copyright Executive

Gill Cockhead
Publishing Contracts & Copyright Executive
Royal Society of Chemistry,
Thomas Graham House,
Science Park, Milton Road,
Cambridge, CB4 0WF, UK
Tel +44 (0) 1223 432134