DISSERTATION

DEVELOPMENT OF COMPUTATIONAL TOOLS TO MODEL MOLECULAR INTERACTIONS FOR MEDICINAL CHEMISTRY

Submitted by

Melissa Coates Ford

Department of Biochemistry and Molecular Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2017

Doctoral Committee:

Advisor: P. Shing Ho

Robert Cohen
Christopher Snow
Martin McCullagh

ABSTRACT

DEVELOPMENT OF COMPUTATIONAL TOOLS TO MODEL MOLECULAR INTERACTIONS FOR MEDICINAL CHEMISTRY

Medicinal chemistry has evolved over the past 40 years to rely heavily on the computationally aided design of new drugs. The work in this dissertation focuses on developing computational tools for the application of medicinal chemistry. For computational techniques to be dependable, important interactions must be properly modeled and the techniques must be rigorously tested. In this work, I first introduce an important interaction for drug design, the halogen bond (X-bond). We consider how decades of work has come closer to properly modeling the X-bond, yet there remain many unexplored areas. Two areas are addressed in this dissertation: the structure-energy relationship of 1) a Br···S^- X-bond in a DNA junction and 2) Br···O and I···O X-bonds in T4 Lysozyme (T4-L). Using these systems, we can better understand the X-bond and further test computational tools. One such tool, a molecular mechanics/dynamics package, TINKER, does not model X-bonds. Thus, I then incorporate a force field for a broad range of X-bonding molecules into TINKER, creating X-TINKER. X-TINKER reproduces the energies and geometries of the X-bond in the DNA and T4-L systems. Last, I will discuss testing a different software developed by Schrödinger, FEP+. We find FEP+ can effectively predict protein stability; however, it still has areas that need improvement. Together, the findings of this dissertation emphasize the importance of understanding molecular interactions, improving algorithms, and testing current programs to find remaining failures. By continuing to use this cycle, we hope to see the impact of computational tools in medicinal chemistry.

ACKNOWLEDGEMENTS

It is hard to not fill countless pages about how grateful I am for all of the people who have helped me get to this point. I must first thank my advisor, Shing Ho, without who I would not be where I am today. You have always known how to step that fine line of letting me learn on my own, while knowing when you need to step in and give me a hand. You have pushed me to become a better writer, presenter, and critical thinker–all invaluable traits to me as a scientist. Lastly, you have helped me find my passion in computational science and for that I will forever be indebted to you. The rest of the Ho lab has been instrumental in my pursuit for this degree. I must first thank Crystal Vander Zanden and all the help you have given me over the years, you truly are a great role model. Rhea Kay Rowe and Anna-Carin Carlsson have been amazing for all the advice and listening ears over the past few years. It is truly wonderful to leave a workplace and get to call your colleagues, friends.

I would also never be at this point without the other important scientists that have helped shape me. First and foremost, I must thank Anthony Rappé for always answering all of my questions, even if they were about the most basic quantum mechanics or molecular dynamics fundamentals. Up until the very end, you have offered me with advice and help that I will never be able to replace. I also need to show gratitude to my student advisory committee: Robert Cohen, Christopher Snow and Martin McCullagh. Each of you have always shown interest, asked great questions and answered many, many of my questions. Above all, you have shaped me into a better writer and speaker.

It is always easier to look at the more immediate, scientific help, but equally important is the solid foundation that has been provided by my family and friends. My sister, Sarah, you not only

have been an inspiration to me but have helped me through every problem I have had large or small. Of course, my parents, Johnny and Julie, there is no way this accomplishment could have happened without either of you. You have given me a platform to become the scientist I am today and for that, I am eternally grateful. All of the lifelong friendships that I have made here are some of the sole reasons I have made it through graduate school. Although I have made many great friends in this department that I am grateful for, I must thank Alison, Jenifer and Sarah for always offering some of the best conversation about science while also knowing when the distraction from science was needed. Graduate school would have been very different without you all in it.

My husband, Brian, deserves my largest thanks. There is not a single person who I would rather have cheering me on. You have been there through the worst of this process and the best. I cannot put to words what you have sacrificed for me to allow me to pursue my dreams. I am looking forward to the next stage with you.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The origins of medicinal chemistry can be traced back thousands of years to the use of natural products to treat ailments and disease. Flashing forward to the present day, medicinal chemistry has become a billion dollar industry worldwide. This transition did not happen overnight, but instead, has travelled a long path to reach where it is today. The modern pharmaceutical industry as we know it began in the 1930s, when a small number of companies began introducing new molecular entities (NMEs) or "drugs". Until the 1950s, the number of NMEs remained relatively low (below 100). Now, this number has swelled to over 1,450 NMEs approved in the United States that were produced by 275 different companies[1].

Since the 1990s, the pharmaceutical industry has spent an increasing amount of money and resources into acquiring NMEs; however, the output has been underwhelming with an average between 25 to 30 NMEs per year (Figure 1.1). As technology continues to improve, medicinal chemists have begun to incorporate computational techniques to save time and money during drug design. The use of computational techniques in drug discovery fundamentally shifts medicinal chemist's efforts to computer-aided drug design instead of a purely experimental-heavy chemical synthesis. This type of design, however, is heavily dependent on the accuracy, speed, and reliability of the computer program that is used and many programs fall short in one or all of these areas. The field has reached a crucial point–where the user must be able to identify and understand failures and possibly improve upon them. In this dissertation, I will discuss where some programs fail to predict fundamental biological or chemical aspects in an accurate,

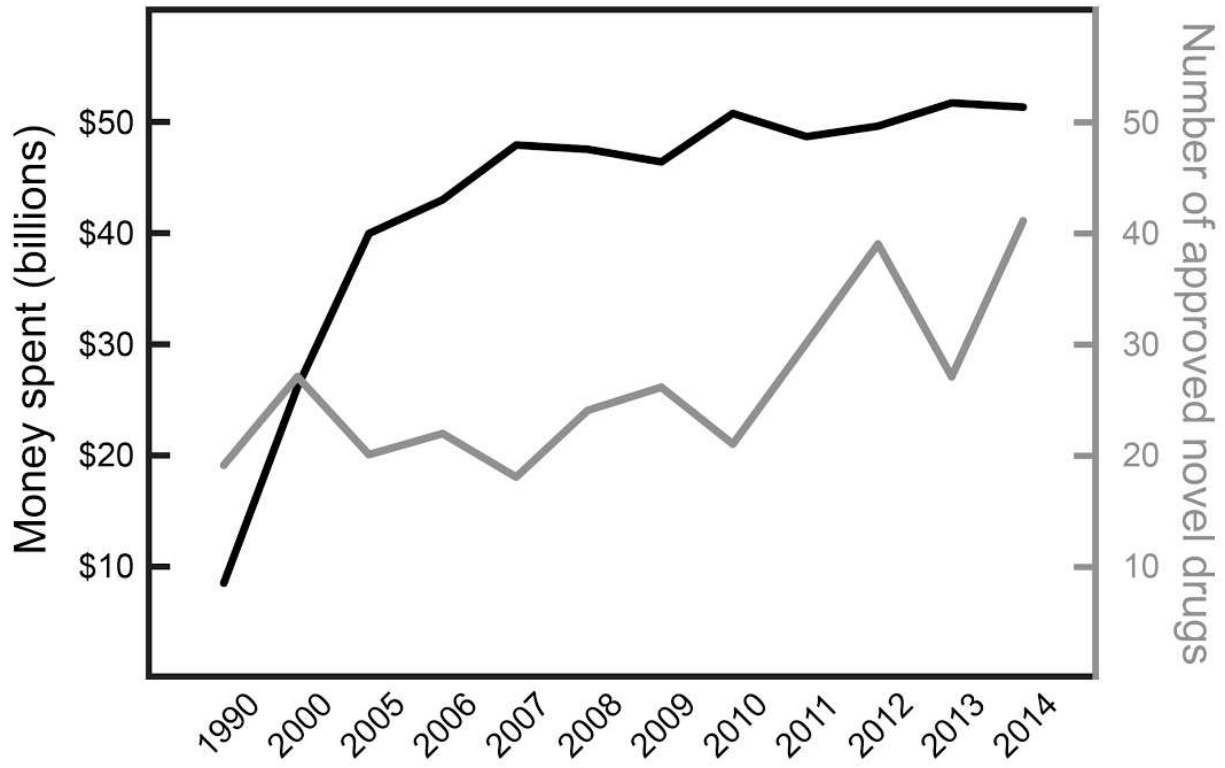


Figure 1.1. Money spent on drugs and number of novel drugs per year. Money spent (left axis) on research and discovery for drug design in billions. Number of approved novel drugs or NMEs (right axis). Data was obtained from the Food and Drug Administration (FDA)[2].

reliable and/or fast manner. I will then offer potential improvements that overcome some of these deficiencies.

It is important to understand the stages of drug discovery and the types of programs that can be applied to each stage to better appreciate the role of computers in drug design relies on first grasping the (Figure 1.2). The initial stages of drug design are disease genomics and target identification. The latter of these, target identification, can make use of many computational techniques including reverse docking and protein structure prediction. The subsequent stages of drug discovery are identifying a lead compound for the classified target and optimizing that lead. During these stages, computational programs can be very valuable due to their speed and ability to screen thousands of compounds. Some of the applications for this stage include docking, library design, quantitative structure activity relationship (QSAR), and molecular mechanics/ molecular dynamics (MM/MD). Finally, the optimized lead will undergo preclinical tests and eventually, clinical trials. Furthermore, computational programs can be used prior to these trials to predict absorption, distribution, metabolism, excretion and toxicity (ADME/T) of the compound.

There the power of rational design can be shown by taking a look at a few successful examples that save both time and money in different stages of the discovery process. One such example is the use of molecular docking during the lead identification stage to discover a potent inhibitor for protein tyrosine phosphatase 1B, a key target for type II diabetes[3]. In this study, a group from Pharmacia (now a part of Pfizer) used both a traditional, experimental high-throughput screen (HTS) with 400,000 compounds and a computational molecular docking approach with 235,000 compounds. While the more time-intensive experimental HTS found 85 potential lead compounds with an inhibitory constant (IC_{50}) of less than 100 μM (a hit rate of

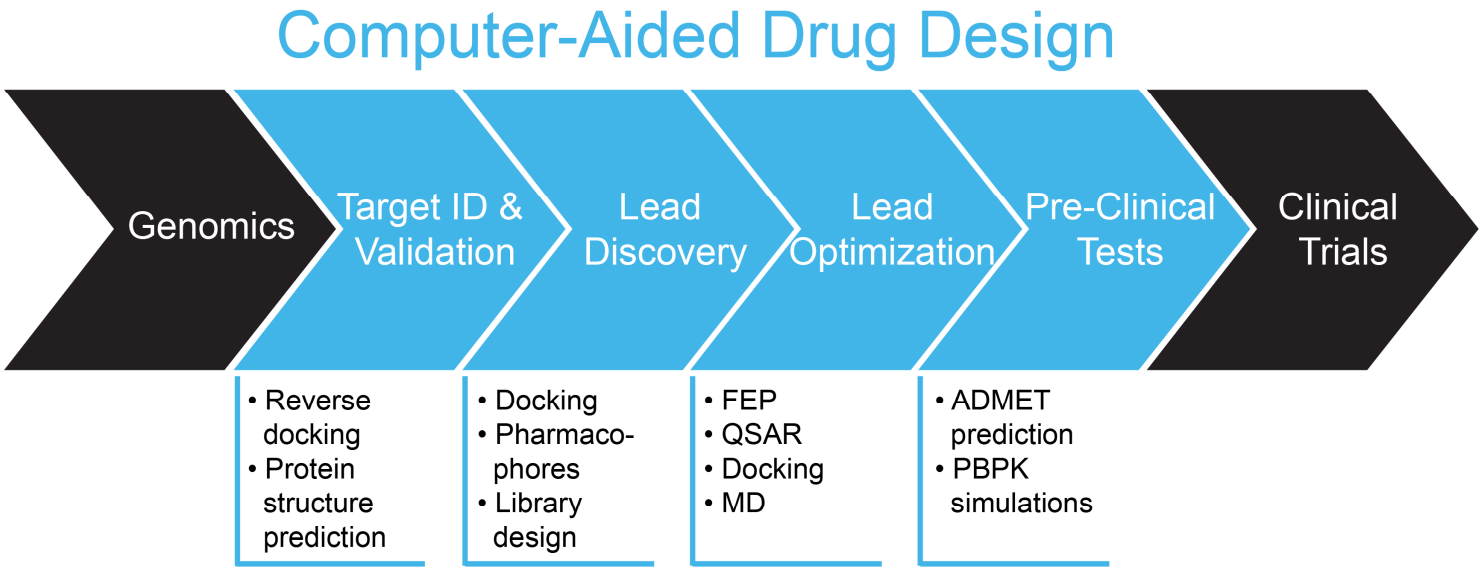


Figure 1.2. Computers in the drug design process. Each stage of drug design is shown from initial stages of disease genomics to the final stages of clinical trials. The stages that computers can be used are shown in blue with a few notable examples of simulation packages and programs bulleted for each stage.

0.021%), the computational docking found 365 potential lead compounds. These compounds were then experimentally tested and 127 of the 365 compounds were found to have an IC_{50} of less than 100 μM (a hit rate of 34.8%). This was an astonishing increase in hit rate of over 1700-fold when comparing the docking approach to the traditional experimental screening approaches. Additionally, the computational screen identified a 2.5-fold more potent lead inhibitor than the one found using HTS.

Another successful demonstration of the use of computers is in the lead optimization stage of drug discovery where the use of free energy perturbation (FEP) is applied to a non-nucleoside inhibitor of the human immunodeficiency virus reverse transcriptase (HIV-RT)[4]. Jorgensen's group had previously found a potential lead of 5 μM using common docking programs, *Glide*[5] and *GlideXP*[6], on a library of about two million compounds[7]. Here, they optimized the low micromolar lead compound to an extremely potent 55 pM using FEP[4]. In more detail, using FEP substitutions on the lead compound were made *in silico* allowing the prediction of binding energies. Only the compounds that had predicted tighter binding to HIV-RT were then synthesized, tested and crystallized. By using this method of FEP in conjunction with experimental techniques, the Jorgensen group was able to find the most potent, non-toxic anti-HIV agent ever discovered.

In addition to discovering and optimizing lead compounds, computational design can aid in how we understand the mechanism of drugs binding to their targets. Medicinal chemists can then use this understanding as a stronger platform for designing better drugs. MM/MD is a particularly promising method for this application. Once a target and a lead compound have been identified, a MM/MD simulation can be performed to study the binding event. An exceptionally interesting example of this was done by Decherchi et al by studying the binding of a transition

state analogue (TSA) to purine nucleoside phosphorylase (PNP)[8]. PNP-TSAs are promising inhibitors for many T-cell cancers and autoimmune diseases including but not limited to gout, rheumatoid arthritis, psoriasis, tissue transplant rejection, and multiple sclerosis.[9] In this study, the binding event of a strong inhibitor, DADMe-immucillin-H, is simulated and validated to experimental findings. Using these findings, they were able to elucidate the mechanism of binding and estimate binding energies while determining the kinetics of the binding event.

Although it is encouraging to look at these successful cases, many programs still fall short in important areas. Thus, it is important for the field of computational chemistry to undergo constant testing and improvements where necessary. Furthermore, it is important to note, the use of computers in drug design cannot be a full replacement for educated medicinal chemists. In this dissertation, there is an important interplay between understanding the biology and chemistry of noncovalent interactions and learning how to use this understanding to make improvements in computational programs for drug design.

Often the failures seen in computational programs occur due to their inaccuracy of modeling important noncovalent interactions. In the beginning of this dissertation, I take a look at a perspective on a common noncovalent interaction found in medicinal chemistry–the halogen bond (X-bond). The X-bond is named due to its analogous behavior to the hydrogen bond (H-bond)[10]. Although there are many competing theories to describe the X-bond[11,12], the most widely accepted of these is the σ-hole theory[13]. In this description, when the halogen participates in a σ-bond, the electron present in the halogen's p_z orbital is pulled into the σ-bond, thus, exposing the halogen's positive nucleus on the surface diametrically opposed to the bond. This causes a unique polarization on the halogen and an aspherical shape of the van der Waals surface that allows the halogen to favorably interact with a large number of both positive and negative

acceptors. In Chapter 2, I will explore the X-bonds importance in medicinal chemistry, current programs and how they improperly model X-bonds, and current attempts to fix this problem.

In order to fix these inaccuracies, we must first understand the fundamentals of the noncovalent interaction. This introduces the first scientific question that this dissertation aims to answer: Can we better understand the structure-energy relationship of the X-bond in different biological settings? In previous work in our lab, we have developed a unique experimental DNA system that allows for determining structure-energy relationships of halogens in a biological setting[14,15]. In this work, relationships were determined for the common halogens that participate in X-bonding–chlorine (Cl), bromine (Br), and iodine (I) interacting with a negative oxygen (O) acceptor. In Chapter 3, I will use this system to determine the structure-energy relationship of Br to another negative acceptor, sulfur (S). Although formally negative oxygen and sulfur are found as X-bond acceptors, they are not the most common. Instead, formally neutral oxygens make up the most prevalent acceptors[16] Moreover, in Chapter 4 of this dissertation, I will discuss how we have developed a protein system that allows us to determine structure-energy relationships of Br and I to a formally neutral O. Using both the DNA and protein systems we can better understand the structure-energy relationship of X-bonds in biological settings.

With better understanding of the fundamentals of X-bonding, we can then ask the next important question: Can we use this understanding to develop and incorporate a better model for X-bonds in programs for drug design? Most molecular simulation packages, especially MM/MD, model the average behaviors of both the overall charge and shape of atoms and, therefore, fail to properly model halogens due to their unique polarization. Chapter 5 of this dissertation attempts to use the knowledge from our experimental systems to develop a unique fix to its improper modeling in MM/MD. Our lab has previously developed an accurate model to describe X-

bonding[17,18]. Here, I will expand this model for more X-bonding systems and incorporate it into an MM/MD simulation package, TINKER. The experimental systems that have been developed offer a platform to validate the MM/MD implementation.

Although many MM/MD packages are valuable for many stages in drug design, FEP is also a very useful application for lead optimization and target identification. While FEP has been around for many decades, it has recently made great advances that prove to be highly effective in protein ligand binding[19]. Another promising application for FEP in medicinal chemistry is designing and testing protein (or drug target) stability, however, FEP has been much less explored for this application and only undergone initial testing[20,21]. Designing more stable proteins has a great potential in medicinal chemistry to help crystallize target proteins. This brings up the last question this dissertation will address: Can FEP accurately predict protein stability in a practical manner? In Chapter 6, I present the results of the testing of new software developed by Schrödinger, FEP+[19], with the application of protein stability and explore its applicability for use in academia and industry.

When using rational design, it is critical to understand the chemistry and biology behind the systems that are being modeled. As seen with many MM/MD and many other molecular simulation packages, by testing software we can discover where the software succeeds and what drawbacks still remain. This knowledge can then help build and test better software. Only with the constant interplay between–understanding the fundamentals and using this knowledge to test and improve software–will computers begin to reach their full potential in medicinal chemistry. In the conclusion of this dissertation, I will explore the scope of this work, how it fits into the field of computational medicinal chemistry, and what still lies ahead.

CHAPTER 2

COMPUTATIONAL TOOLS TO MODEL HALOGEN BONDS IN MEDICINAL CHEMISTRY[1]

2.1 SUMMARY

The use of halogens in therapeutics dates back to the earliest days of medicine when seaweed was used as a source of iodine to treat goiters. The incorporation of halogens to improve the potency of drugs is now fairly standard in medicinal chemistry. In the past decade, halogens have been recognized as direct participants in defining the affinity of inhibitors through a noncovalent interaction called the halogen bond or X-bond. Incorporating X-bonding into structure-based drug design requires computational models for the anisotropic distribution of charge and the nonspherical shape of halogens, which lead to their highly directional geometries and stabilizing energies. We review here current successes and challenges in developing computational methods to introduce X-bonding into lead compound discovery and optimization during drug development. This fast-growing field will push further development of more accurate and efficient computational tools to accelerate the exploitation of halogens in medicinal chemistry.

2.2 INTRODUCTION

The standard strategy in drug design is a multistep process that involves very time-intensive and costly steps[22], starting with the discovery of a therapeutic target, followed by identification of a lead compound that can potentially inhibit that target, optimizing the lead compound to improve its efficacy, and then performing clinical trials that will lead to eventual FDA approval for

[1] The work in this chapter was published as a review article in 2016 in *Journal of Medicinal Chemistry*. P.S.H and I conceived and co-wrote this manuscript.

Ford, M. C.; Ho, P. S. Computational Tools to Model Halogen Bonds in Medicinal Chemistry. *J. Med. Chem.* 2016, 59, 1655 –1670.

release of the new drug to market (Figure 2.1). Halogens have been classically incorporated into the development of pharmaceuticals to increase membrane permeability and decrease metabolic degradation[16]. Interestingly, 50% of the top leading drugs on the market are halogenated, and halogens persist throughout the drug development process, from initial discovery to launch (Figure 2.2)[23]. There is now a greater appreciation that halogens play a direct role in the efficacy of certain drugs through a molecular interaction that is now defined as the halogen bond (or X-bond[10], Figure 2.3a), in addition to their more traditional role as acceptors of hydrogen bonds (H-bonds). X-bonding is recognized as providing upwards of 1,000-fold increase in specificity and affinity of inhibitors toward their molecular targets; however, their contributions have been recognized primarily in hindsight. Although fluorine is the dominant substituent, the percent of fluorinated compounds decreases whereas those of the heavier halogens (Cl, Br, and I) that most commonly form X-bonds increase from 40% at the start to >60% of halogenated drugs at launch[23]. Thus, there is great potential to exploit X-bonding as a molecular tool in medicinal chemistry if the interaction can be accurately incorporated into drug design algorithms.

On average, the traditional approach to drug development stretches over 14 years and costs ~$880 million from the time a therapeutic target is discovered to the launch of a drug to market (Figure 2.1)[22]. This inordinately long time and high cost can be mitigated with the aid of computational approaches, particularly during the earlier stages of drug development. The rational design of new halogenated inhibitors, however, is greatly hampered by the inability of current computational programs to properly model X-bonds, particularly their contributions to specificity and affinity. The shortcomings of current computational models for halogens lie primarily in their inability to describe the fundamental underpinnings of the interaction– specifically how halogens, which are generally electron-rich, can direct the binding of inhibitors

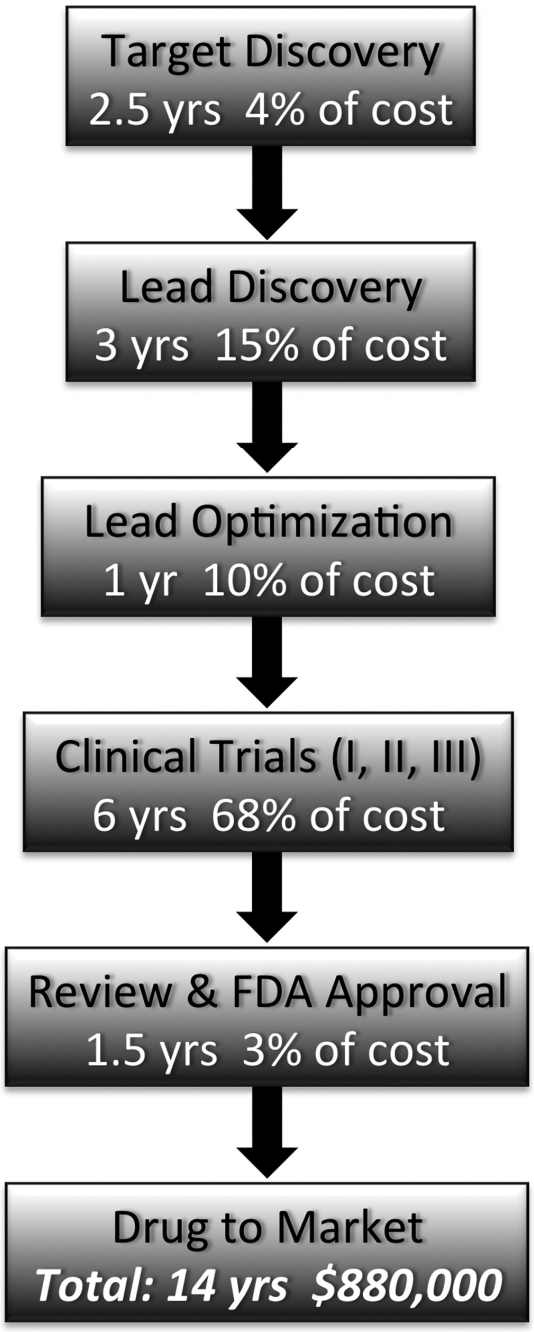


Figure 2.1. Flowchart of the time and cost in traditional drug development. The flowchart follows the process for bringing one drug to market from initial target discovery to final approval by the Food and Drug Administration (FDA), including the average times and the percentage of the overall costs associated with each step (data from Parexel[22]).

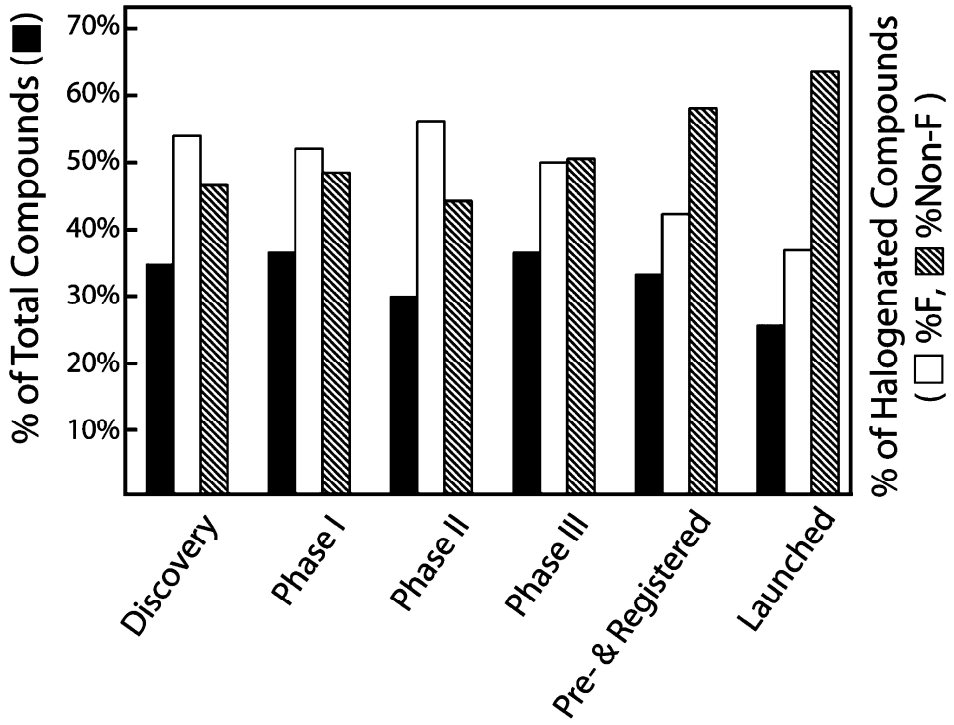


Figure 2.2. Progression of halogenated compounds through each phase of drug development. Percentages of compounds that are halogenated are shown as black bars. The percentages of halogenated compounds that are fluorinated are shown as white bars, whereas those that contain halogens other than fluorine are striped bars (data from Xu et al.[23]).

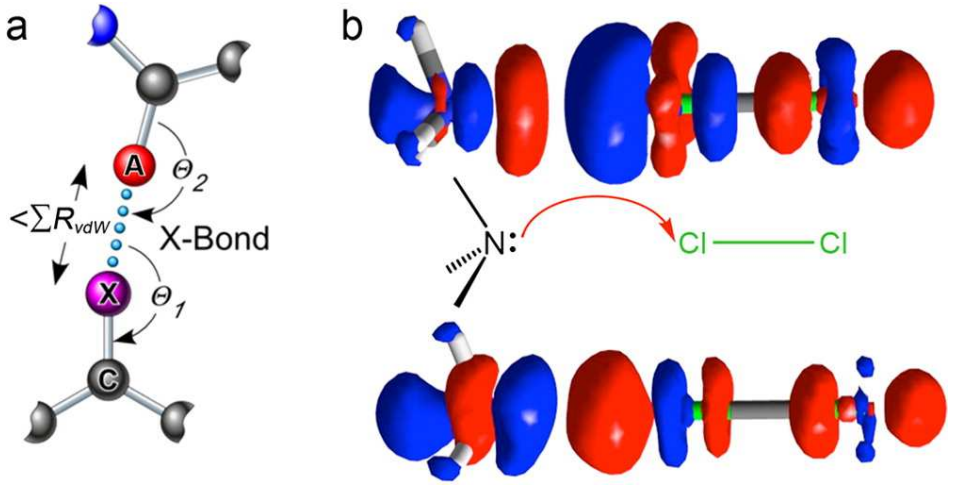


Figure 2.3. X-bonding and charge-transfer bonding model. (a) The X-bond is defined as a short, directional interaction between a halogen substituent (X) and an electron-rich acceptor atom (A). (b) The charge-transfer bond in the NH3···Cl2 complex. The top figure shows the electron density difference (EDD) map for polarization of the molecules in their Lewis states. The EDD map showing electron transfer from the nitrogen to the Cl in the complex (reproduced with permission from Wang et al.[24]).

through directional X-bonding interactions with electronegative atoms (oxygen, nitrogen, sulfurs, and delocalized π-electron systems) in their protein targets.

In this perspective, we will focus on the various approaches to applying the X-bonding concept in drug design, and what challenges still lay ahead. The most obvious stages for inserting X-bonding into this process from a traditional medicinal chemistry perspective are the early steps of lead discovery and optimization. Our discussion will focus primarily on computational methods to enhance these steps through halogenation and the introduction of X-bonds; thus, we will start by summarizing the various physicochemical models that have been presented to explain this seemingly paradoxical interaction. Fluorines serve as X-bond donors only under extraordinary circumstances[25,26] and, thus, for most medicinal chemistry applications and for the remainder of this review, we will consider only the heavier halogens (Cl, Br, and I).

Physical Chemical Basis of X-bonding.

The X-bond concept explains generally how electron-rich halogens (particularly Cl, Br, and I) can interact favorably and at short distances with nucleophilic and, in some instances, formally anionic atoms (N, O, or S). X-bonds share many similarities to hydrogen bonds (H-bonds), hence the name. The interaction poses a significant challenge to standard molecular modeling programs, particularly those that are applied to macromolecular structures, including protein-inhibitor complexes. To better appreciate how challenging this interaction is for a medicinal chemist, we must first understand the basic chemical principles of X-bonding.

Charge-Transfer Theory. The short-distance, stabilizing interaction between molecular halogens (Br_2 and I_2) and, for example, oxygen atoms in dioxane was first described as charge-transfer (CT) bonds by Odd Hassel[27] as an extension of Mulliken's CT theory[28,29]. A CT interaction is characterized as a classic electron donor/acceptor complex (Figure 2.3b) in which a

Lewis base transfers partial electron densities from its highest occupied molecular orbital to the lowest unoccupied orbital of a Lewis acid (the halogen). Such a complex is typically characterized by a unique spectroscopic CT absorption band[30]. There remains strong support for the contribution of charge transfer to X-bonds, including experimental characterization of carbon tetrabromide and bromoform[11] and recent analysis of various Lewis acid/base complexes applying valence bond and blocked-localized wave function (VB/BWF) theories[24]. Recently, however, the CT-model has been challenged by models that focus more on the electrostatic nature of the X-bond. In particular, there is evidence from electrostatic potential maps that the central atoms (C or Si) of highly halogenated Group IV compounds (such as CCl_4 and $SiCl_4$) carry an electropositive potential that could attract a nucleophilic atom (see the σ-hole theory below).[31] Experimental evidence in support of this primarily electrostatic interaction comes from the structures of various halogenated Si or Ge cages surrounding halide ions[32]–a so-called "tetral-bond"[33].

σ-Hole Theory. Perhaps the most readily accessible explanation for the X-bond is the σ-hole model espoused by Pulitzer and colleagues[13,31,34]. In this model, a halogen that forms a covalent bond to, for example, a carbon (a C–X σ-bond) will result in depopulation of the valence p_z orbital of the halogen and, consequently, an electropositive crown diametrically opposed to that σ-bond–a σ-hole (Figure 2.4). The electropositive σ-hole serves as the X-bond donor, whereas the electron-rich partner is the X-bond acceptor[10], in analogy to H-bonding nomenclature.

The model provides a simple explanation for the strong directionality and the various chemical factors that affect X-bonding energies. The electropositive crown specifies that X-bond acceptors will align primarily opposite the σ-bond. The size of the σ-hole defines the potential strength of the X-bond and intensifies with the increased size of the halogen (F < Cl < Br < I, which also follows the polarizability and conversely the electronegativity in their periodicity)

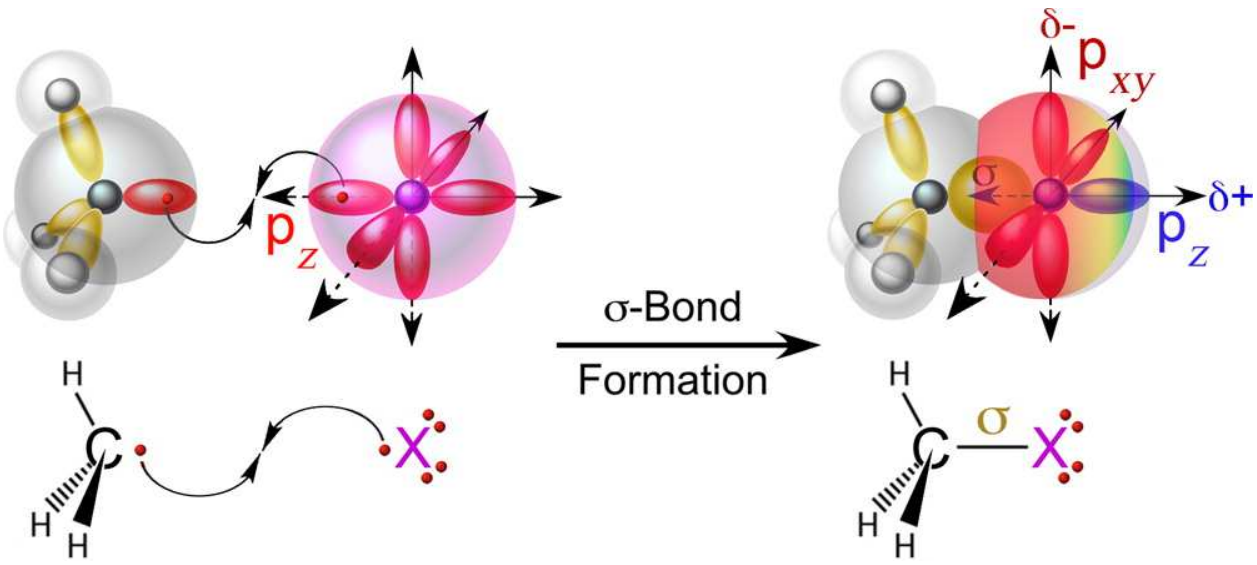


Figure 2.4. σ-Hole model: The formation of a covalent carbon−halogen bond (a C–X σ-bond) for example pairs the electrons from the valence orbitals of the two atoms. As a result, the p_z orbital of the halogen opposite the σ-bond becomes depopulated, resulting in an electropositive crown (in blue), whereas the p_{xy} orbitals retain their complement of electrons to account for the overall negative charge of the halogen (reproduced with permission from Scholfield et al.[18]).

and the increased electron withdrawing ability of the atom or molecule participating in the σ-bond with the halogen.

Lump-Hole Theory. There are alternative electrostatic models that purport to fill-in gaps or shortcomings of the σ-hole model. The lump-hole model, for example, does not rely on an explicit positive charge at the surface of the halogen but rather on a local charge depletion at the X-bond donor end (the hole), which can interact with excess charge of the acceptor (the lump) (Figure 2.5)[12,35]. This model is consistent with the local charge concentration (CC) and local charge depletion zones (CD) that result in anisotropic electron-density distribution around chlorines, as seen by Espinosa et al., in the high-resolution crystal structures of C_6Cl_6[36].

Although the physicochemical description remains under debate, it is well understood that X-bonds are most accurately modeled through rigorous QM calculations derived from these various models. However, QM is not appropriate for biological complexes, except in very rare situations where a structure is known to subatomic resolution (as in the case of aldose reductase[37]). Thus, we are left with the need to adapt what we learn from QM calculations on model systems to more tractable molecular mechanics and dynamics (MM/MD) simulation methods. A major challenge for medicinal chemists is thus to accurately incorporate X-bonding models into MM/MD methods in drug design, particularly the steps of lead discovery and optimization. The strategies will take advantage of some properties of X-bonds that appear to be unique to the interactions in biological systems[38]; therefore, we need to summarize these structural and energetic features.

X-Bonds in Biology.

In the past decade, our understanding of X-bonds in biological molecules (biomolecular X-bonds or BXBs) has increased dramatically[38]. BXBs share many of the same geometric features with their small molecule counterparts, including short donor−acceptor distances and the strong

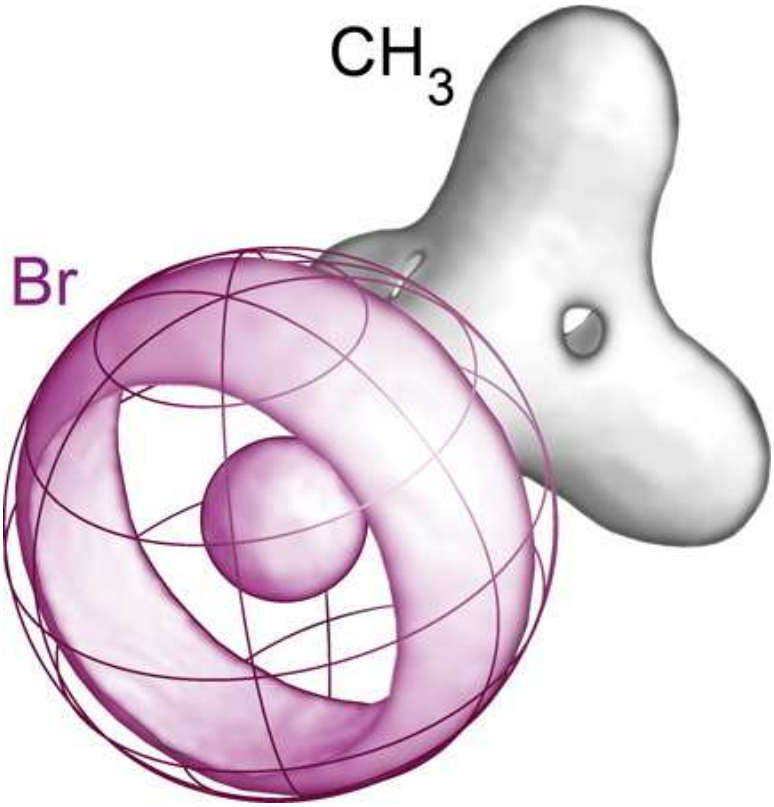


Figure 2.5. Electron distribution of atoms in CH_3Br, as predicted from the lump-hole theory. The distribution of electrons forms a ring around the bromine center (accounting for the majority of electrons at the atomic surface) and a "hole" at the surface that can interact with the "lump" of electrons from an interacting X-bond acceptor (data from Eskandari and Zariny[12]). The standard surface of the bromine atom is outlined as a spherical cage.

directional preference for the approach of the acceptor toward the electropositive σ-hole of the halogen donor. Anything that can serve as a H-bond acceptor (O, N, S, π-systems, including aromatic rings[38] and radicals[39]) can also accept X-bonds in both small molecule and biological complexes (Figure 2.6[40–42]). We are now seeing, however, complexities in the relationships of BXBs to H-bonds in proteins and to solvent.

In protein-inhibitor complexes, the most common X-bond acceptors are the carbonyl oxygen of a protein's peptide backbone[16,43], which is not entirely surprising given their prevalence in proteins. However, because most peptide bonds are also involved in H-bonds that stabilize protein structures (*α*-helices, *β*-sheets, turns, etc.), the carbonyl oxygen serves as an acceptor for both molecular interactions. In small organic molecules, H- and X-bonds can compete or complement each other with H-bonds being stronger, weaker, or similar to X-bonds[44–46]. In biological systems, however, they are observed as orthogonal interactions that intersect at a common peptide oxygen (Figure 2.6a)[40]. Such X-bonds (called hX-bonds) are geometrically perpendicular and thermodynamically independent of each other. The orthogonal relationship between H- and X-bonds are now becoming relevant also in small molecule complexes,[47] including those involving isolated amides that mimic peptide bonds[48]. The significance of this concept is that the addition of a BXB interaction to a protein can be well-defined in terms of an optimum position and will not be expected to perturb the integrity of the protein structure– important concepts in drug design, as we will see later.

The halogen itself is amphoteric[41] with the electropositive *σ*-hole being a potential X-bond donor and the electronegative annulus around the girth of the atom being a H- or X-bond acceptor. In biomolecular systems, this amphoteric nature is most evident in the interactions of halogens with water (in itself amphoteric), a unique feature of biological systems. The

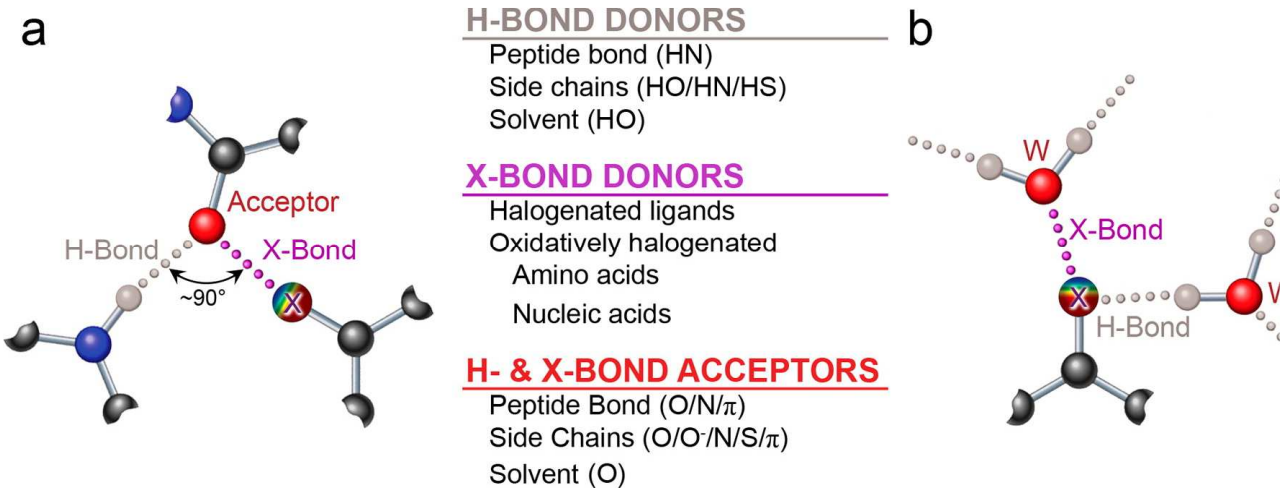


Figure 2.6. Biological X-Bonds (BXBs) and their relationship to H-bonds. The halogen (X) is shown with its electrostatic potential going from negative (red) to positive (blue), allowing it to be both an X-bond donor and a H-bond acceptor. (a) BXBs (XB, magenta dotted line) and H-bonds (HB, gray dotted line) that share a common acceptor are orthogonally related both geometrically and thermodynamically[40]. (b) The amphoteric nature of the halogen[41] also allows it to serve as an X-bond donor or H-bond acceptor to a water, which in turn bridges to other acceptors and donors[42].

relationship between halogens and water is complex. Waters can, for example, serve to bridge an X-bond donor to its acceptor[42]. Alternatively, because any H-bond acceptor can also form an X-bond, the BXB at a solvent-exposed surface will invariably displace a water molecule. Finally, with all these potentially stabilizing electrostatic-type interactions at the surface, we must come to terms with the truism that halogens are actually hydrophobic atoms, with Br, for example, having the same hydrophobic effect as a methyl group[38,49].

The stabilizing energy of a BXB can be comparable to or greater than that of a H-bond, depending on the system. The BXB energy can be tuned according to the type of halogen, becoming more favorable as the halogen becomes larger and more polarizable[50]. In addition, attaching the halogen to a more electron-withdrawing group will result in a more positive σ-hole and, consequently, a stronger BXB. For any given system, the effective BXB energy is dependent on the geometry (distances and angles of approach) that relates the X-bond donor and acceptor atoms. These general concepts are consistent with all of the basic QM models described above for X-bonds and have been confirmed experimentally using a unique four-stranded DNA junction model system[15]. The DNA studies show that the enthalpies of stabilization by BXBs follow the series F < Cl < Br < I in stability, and the more enthalpically favorable interactions result in more ideal geometries (more linear approach of acceptor toward the σ-hole and shorter distances). However, the studies also demonstrate the concept of enthalpy−entropy compensation, where the most stabilizing enthalpic interaction (that of I) comes at an entropic cost in that fitting such a large atom into a small space results in the loss of conformational dynamics[17]. As a result, we must consider not only the enthalpic contributions of the electrostatic component of the interaction, but also how the interaction affects the entropy of the entire system, including the solvent.

This detailed understanding of BXBs can help facilitate the rational design of halogenated inhibitors and drugs. We will start this discussion with how detailed knowledge of BXBs helps in the development of general computational tools for the rational design of inhibitors, starting with those used to optimize the efficacy of a lead compound, followed by how BXBs can facilitate lead compound discovery through scoring functions. Although chronologically backward relative to the progression of drug development (Figure 2.1), the computational methods required to enhance the efficacy of a molecule can be derived directly from BXB geometries and energies, and these optimization methods lend themselves logically to solving the more complex problem of deriving scoring functions.

■ 2.3 COMPUTATIONAL TOOLS TO MODEL BXBS FOR MOLECULAR DOCKING AND LEAD OPTIMIZATION

Several studies have shown how important halogens can be at enhancing the affinity of a lead compound, potentially increasing the affinity of an inhibitor by up to 1000-fold (Table 2.1). In a very systematic study, Hardegger et al.[51] demonstrated that introducing halogens that form BXBs reduces the inhibitor's IC_{50}s against two anticancer targets (human cathepsin L and MEK1 kinase). In one set of studies, the authors showed that adding X-bonding halogens to the human cathepsin L inhibitor resulted in successive increases in affinity to its target[51]. The introduction of Cl, Br, or I to the phenyl ring of the inhibitor that sits in the polar S3 pocket of the substrate binding site was seen to form BXBs with favorable geometries to the peptide oxygen of Gly61. The new interactions were attributed to reducing the IC_{50} from 0.29 μM (with no interaction) to 0.022 μM for Cl, 0.012 μM for Br, and 0.0065 μM for I X-bonds. As expected, the F analogue showed repulsion of the electronegative halogen from the Gly61 oxygen and a concomitant

Table 2.1. Effect on Affinity of Replacing H-Bond Donors with Iodine Substituents in Inhibitors against Protein Targets[51–56] with Protein Data Bank (PDB)[57] Codes, Where Available, Listed for the Single Crystal Structures Showing an X-Bonding Interaction of Each Inhibitor Along with the Fold Increase Associated with the H to I Substitution

Inhibitor Target	Inhibitor	Fold Increase in Affinity	PDB ID
membrane-anchored aspartyl protease (BACE)	(1S,2R)-N-[1-(3,5-Difluorobenzyl)-2-hydroxy-3-(3-iodobenzyl-amino)-propyl]-5-methyl -N′,N′-dipropylisophthalamide	25	2IQG[52]
Cathepsin L	(4R)-4-[(2-Chlorophenyl)sulfonyl]-N-(1-cyanocyclopropyl)-1-{[1-(2-fluoro-4- iodophenyl)cyclopropyl]carbonyl}-L-prolinamide	74	N/A[51]
Mouse double minute 2 (MDM2/HDM2)	(4-chlorophenyl)[3-(4-chlorophenyl)-7-iodo-2,5-dioxo-1,2,3,5-tetrahydro-4h-1,4-benzodiazepin-4-yl]acetic acid	100	1T4E[53]
adenosine kinase	7-iodo-7-deazaadenosine	200	1LIJ[54]
HIV RT	3-iodo-4-phenoxypyridinone	300	N/A[55]
Aldose reductase	2-[2-[(4-bromo-2-fluorophenyl)methylcarbamothioyl]-5-fluorophenoxy]acetic acid (IDD 594)	1100	1US0[56]

increase in the IC_{50} relative to the parent inhibitor. A methyl substituent (which mimics the steric and hydrophobic properties of Br) reduced the IC_{50} only slightly (to 0.13 μM).

In a parallel study, Hardegger et al.[58] also considered the potential effects of X-bonding in the apolar binding pocket of MEK1 kinase. In this case, similar trends were seen with an F substituent showing higher IC_{50} values, and systematically reduced IC_{50} going from Cl to Br to I. In each case, the halogen interaction to the carbonyl oxygen of Val127 was of the hX-bond type. For Cl, the effect on affinity was not as pronounced (with only ~2-fold reduction in IC_{50}), which may be compromised by the flexibility of the inhibitor within the binding pocket, leading to less ideal geometries for the X-bond. The studies demonstrated the importance of X-bonding geometry on the effects of halogens on affinity in either a polar or nonpolar environment.

Our understanding of the principles leading to BXBs has led us to propose a general strategy for lead compound optimization that takes advantage of the structural and thermodynamic properties of the interaction (Figure 2.7)[59]. In this approach, we start with the structure (determined either experimentally or derived from a virtual screen) of a lead compound in the binding pocket of its target. From this initial model, we can analyze the protein structure to identify the potential X-bond acceptors in the binding pocket and, from the orthogonality concept, predict the optimum position in space to place a halogen to form a BXB. This predicted position can then help inform a medicinal chemist as to what type and where to place a halogen as a substituent of the lead compound in order to form an effective BXB. The geometry and energy, and the corresponding effect on affinity of the newly halogenated compound, can then be predicted by geometry optimization/molecular dynamics simulation of the complex with the protein target. It is this last step of computational modeling that remains an obstacle to this strategy because most current molecular docking and simulation programs do not incorporate

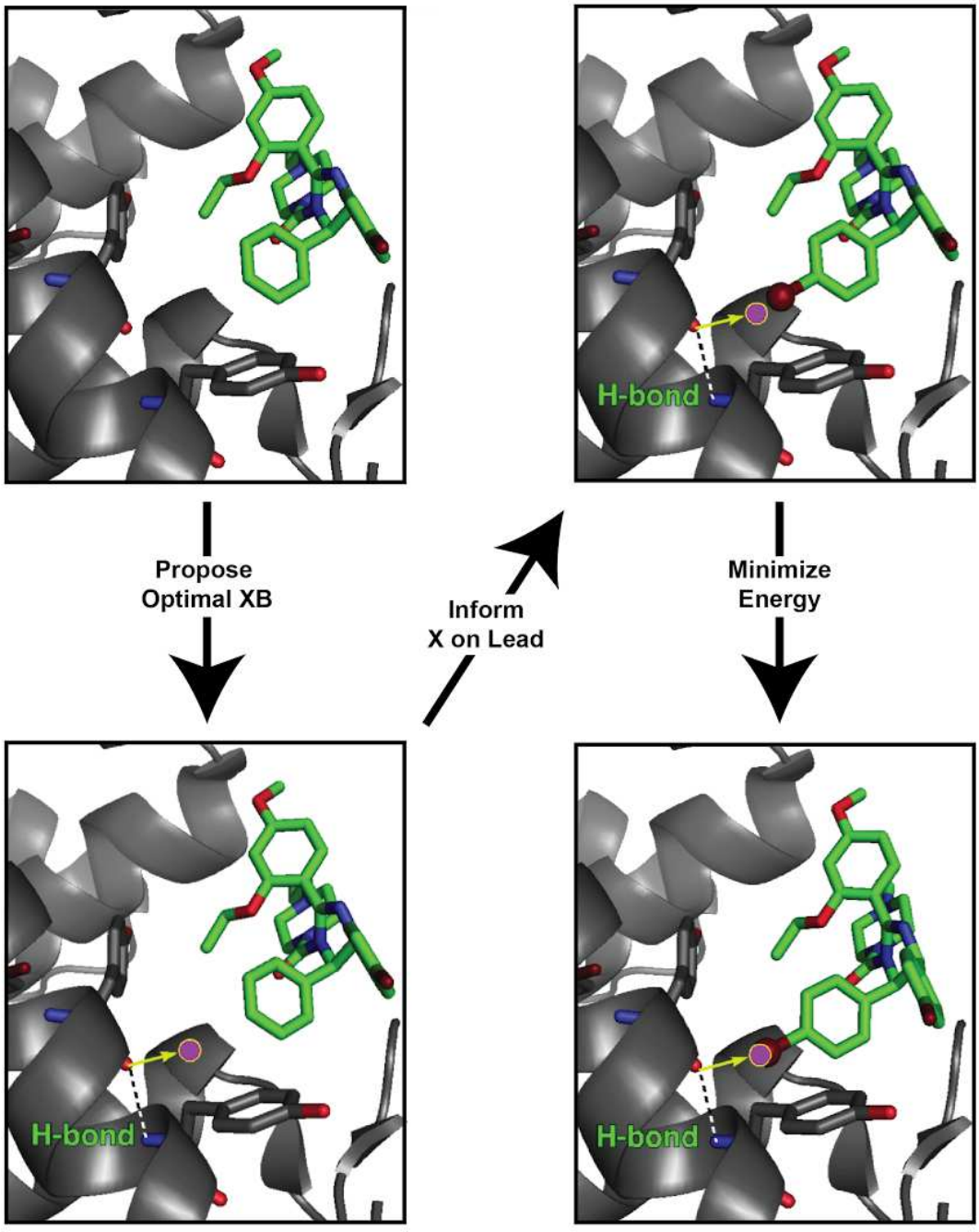


Figure 2.7. Strategy for lead optimization. An optimization strategy starts with a structure of a target in complex with a lead compound. X- bonds in protein-inhibitor complexes tend to be geometrically perpendicular and energetically independent of H-bonds that share the same acceptor (the concept of orthogonality between the interactions[40]); thus, the position for where to place a halogen to form an optimum BXB can be predicted from the geometries of the acceptors and the H-bonding pattern within the binding site. This position can then be used to inform where a halogen substituent should be added to the lead compound. Finally, MM/MD simulation predicts the final geometry and binding energy for the optimized inhibitor.

BXBs in their algorithms. In the remainder of this section, we will focus on the challenges to developing computational methods that accurately model BXB geometries and energies, which in turn facilitate the optimization of a lead compound.

QM Modeling of X-Bonding on Inhibitor Specificity and Optimization.

The most accurate computational method to study and model X-bonding is through QM calculations; however, there are very few biomolecular systems that lend themselves to this level of rigorous analysis. The first example of a QM analysis defining the role of X-bonding in inhibitor specificity was with the inhibitor of aldose reductase, 2-[2-[(4-bromo-2-fluorophenyl) methylcarbamothioyl]-5-fluorophenoxy]acetic acid (1, IDD-594)[56]. The 0.6 Å resolution structure of the reductase inhibitor in complex had shown that the Br of the inhibitor was within 3.0 Å of the hydroxyl group of a Thr side chain[56]. An analysis of the electrostatic potential of the complex, applying density functional theory (DFT), attributed the 1,000-fold increase in the specificity of 1 for the aldose reductase over the closely related aldehyde dehydrogenase to this short bromine-oxygen interaction, which we now define as an X-bond[37].

A more direct application of QM to model BXBs for inhibitor design and optimization can be seen in the studies of Xu et al.[60] on analogues of sildenafil, an inhibitor against the degradative action of phosphodiesterase type 5 (PDE5) on certain smooth muscle groups in the body. PDE5 inhibitors are important drug candidates for the treatment of male erectile dysfunction and pulmonary arterial hypertension[61]. Starting with the crystallographic structure of the PDE5-sildenafil complex[62], the authors first considered how systematically replacing a hydrogen in the inhibitor with X-bonding halogens would potentially affect affinity, using the molecular docking program GLIDE XP[6] to determine if a particular X-bond was favorable. These simulations were then used to inform the subsequent costly steps of chemically synthesizing the compounds that

had potential as leads. Because X-bonding is not incorporated in these algorithms, the docking studies were followed by hybrid QM/MM calculations, where the BXB donors and acceptors constituted the QM layer to predict the energies of any potential X-bond, and the remainder of the complex treated by classical molecular mechanics force fields as an MM layer. Those halogenated analogues that were predicted to form a favorable X-bonding interaction were then chemically synthesized, their structures determined by X-ray crystallography to confirm the presence of the X-bond, and the interaction energies determined by NMR. The resulting experimental binding energies correlated well with the QM/MM predicted energies, supporting the concept that X-bonds can be rationally incorporated into a lead compound optimization strategy. Furthermore, the study indicated that accurate computational modeling of X-bonds has the potential to reduce the time and cost of structure-based drug development. The question, however, is how to make the computational modeling of X-bonds in biomolecular complexes more readily accessible, i.e., without the need to apply rigorous and costly QM calculations.

Application of Semiempirical QM (SQM) in Drug Design.

One approach to reducing the computational cost of a QM calculation in drug design is to apply a semiempirical QM (SQM) appraoch[63]. SQM describes the electronic behavior of a molecular system by applying experimentally parametrized functions (often linear scaling) that approximate electron exchange interactions of the QM Hamiltonian equations. In the late 1970s, Michael Dewar and James Stewart developed MNDO[64], which used experimental heats of formation to help parametrize chemical bonding potentials. AM1[65] and then PM3[66] were developed to help incorporate noncovalent interactions, specifically H-bonding, into SQM calculations. More recently, PM6[67] was developed to provide consistent global parameter optimization for all main

group elements. The result is a method that maintains sufficient accuracy relative to QM, but with significant improvements in computational time to allow simulations of protein systems[68].

Despite the obvious computational time savings, there are several hurdles to incorporating SQM into drug design strategies. One particular problem is that the parametrized equations are highly dependent on the experimental system that they were derived from and thus may not be accurate for describing energies of molecules that are not closely related to the parent system. In addition, many of the SQM approaches still have difficulty describing noncovalent interactions, including dispersion[69], and more complex interactions such as H-bonding and, for this discussion, X-bonding.

Hobza's group introduced corrections for dispersion and electrostatic H-bonds into PM6 to develop PM6-DH[70]. The dispersion energy (E_{dis}, Eq. 2.1) is a classical r^{-6} function with C_6 serving as an empirical scaling coefficient. The overall function is dampened at short distances by $f_{damp}(r_{ij}, R_{ij}^0)$, which is a function of r_{ij} (the distance) and R_{ij}^0 (the sum of van der Waals radii) of interacting atoms (i and j). The various scaling terms (including the slope α and the scaling factor for the radii s_r of the dampening function) are fit empirically to replicate distance-dependent interaction energies in a benchmark data set[71] that includes only dispersion.

$$\sum E_{dis} = -\sum f_{damp}\,(r_{ij}, {R_{ij}}^0)\, C_{6ij} {r_{ij}}^{-6} \qquad (Eq.\,2.1)$$

The H-bonding function of PM6-DH was derived from a small molecule data set that contained 104 H-bonded complexes with various hybridized N and O acceptors[70]. The resulting H-bonding energy (E_{HB}, Eq. 2.2) includes two distance (r)-dependent components. The first term defines the electrostatic and directional properties of H-bonds incorporated with the charges ($q(H)$ and $q(Y)$) and the angle (θ) relating the H-bond donor and acceptor. The second $1/r$ dependent term is incorporated to account for positive correction terms associated with close

distance repulsions, particularly with −H···O hydrogen bonds. The scaling coefficients c, c_{rep}, and A are parameters that are fitted against the complete benchmark data set. The authors found that interaction energies calculated with the corrected PM6-DH had errors of <1 kcal/mol relative to QM calculations, reducing some errors from PM6 (with no correction) by as much as 8 kcal/mol and making them comparable to MP2 level calculations.

$$E_{HB} = c\left[\frac{q(H)q(Y)}{r^2\cos(\theta)} + c_{rep}A^{-r}\right] \qquad (Eq.2.2)$$

A second generation of PM6-DH (PM6-DH2) was applied in virtual design strategies to identify potential inhibitors against two important therapeutic targets: HIV protease and cyclin-dependent kinase 2 (CDK2)[72]. HIV protease is responsible for the maturation of HIV-encoded proteins in the host to generate infectious virus particles. The study by Fanfrlík et al. evaluated the ability of the molecular docking program DOCK[73] to score the potential binding affinities of 22 ligands, 11 current drugs on the market, and 11 "false" decoys against HIV protease[74]. The correlation of the DOCK- calculated enthalpies to the experimentally determined K_i values in this set improved from $R^2 = 0.3$ for the native program to 0.8 with PM6-DH2 incorporated.

CDK2 is a key serine/theorine kinase involved in cell cycle control at the G1/S phase; defects in this enzyme are associated with many forms of cancer[75]. ATP analogues that are competitive inhibitors form extensive H-bonding networks in the ATP-binding cleft of CDK2. Using a similar docking approach as with HIV protease, Dobeš et al. evaluated the ability of the AMBER force field to predict the binding affinity of 15 inhibitors against CDK2 and found that PM6-DH2 improved the correlation of the calculated total enthalpies to $\ln K_i$ from $R^2 = 0.21$ to 0.87[72]. Thus, the incorporation of noncovalent interactions (dispersion and H-bonding) into SQM methods can significantly enhance virtual lead discovery.

A correction to PM6 that incorporates the BXB concept (PM6-DH2X) takes the form of an exponential function that is dependent on the X-bond distance r (Eq. 2.3)[76]. The scaling parameters a and b were derived for each donor (Cl, Br, and I) and acceptor (O and N) atom type by fitting to MP2 calculated energies for X-bonded complexes.

$$E_{XB} = ae^{-br} \qquad (Eq. 2.3)$$

For determining how well this correction works, PM6-DH2X was applied to a test set of inhibitors against casein kinase 2 (CK2), a key Ser/Thr kinase involved in cell cycle control, DNA repair, and other cellular processes[77]. CK2 is abnormally active in tumor cells, making it a good target for the development of anticancer therapeutics. Halogenated drugs are powerful inhibitors of CK2 with X-bonds in the ATP-binding site contributing significantly to the affinity and specificity of these compounds[78,79]; thus, this class of compounds is ideal to test the performance of PM6-DH2X against PM6-DH2. This study utilized two test sets (Table 2.2, Figure 2.8), the first set with structural and binding energy data on CK2 inhibitor complexes and the second set with only information on binding (structures were calculated for complexes resulting from molecular docking procedures using the program DOCK[80]). PM6-DH2X performed very well in predicting the interaction enthalpies in the set with known structural data (R^2=0.86) and moderately well for the test set with no structural data (R^2=0.52) compared to PM6-D2H (R^2=0.00 and 0.17 for the two sets, respectively). When entropy is incorporated to calculate overall binding free energies, the correlations for the two sets decrease to R^2=0.24 and 0.19, respectively. This is the first successful implementation of BXB models into the molecular docking program DOCK.

QM-Derived Force Fields for Biomolecular X-Bonds.

The most commonly used programs for virtual lead discovery and optimization are based on classical Newtonian potential energy functions, as they are applied in molecular mechanics/

Table 2.2. Halogenated CK2 inhibitors[a]

PDB Code	Inhibitor	Number of Halogens	K_i (μM)
3KXN[81]	4,5,6,7-tetraiodo-1H-benzimidazole	4 I	0.023
1ZOE[82]	4,5,6,7-tetrabromo-benzimidazole	4 Br	0.045
N/A[79]	4,5,6,7-tetrabromo-N,N-dimethyl-1H-benzimidazol-2-amine	4 I	0.050
1ZOG[82]	4,5,6,7-tetrabromo-1H,3H-benzimidazol-2-thione	4 Br	0.070
N/A[79]	2-chloro-4,5,6,7-tetraiodobenzimidazole	4 I	0.090
N/A[79]	2-bromo-4,5,6,7-tetraiodobenzimidazole	4 I	0.120
N/A[79]	4,5,6,7-tetrabromo-1H,3H-benzimidazol-2-one	4 I	0.120
2OXD[83]	4,5,6,7-tetrabromo-2-(methylsulfanyl)-1H- benzimidazole	4 Br	0.150
2OXX[83]	5,6-dichloro-4,7-diiodo-2-methyl-1H-benzimidazole	4 Br	0.200
N/A	4,5,6,7-tetraiodo-1H,3H-dihydro-2H-benzimidazole-2-thione	4 Br	0.200
2OXY[83]	5,6-dichloro-4,7-diiodo-1H-benzimidazole	4 Br	0.300
N/A[79]	4,5,6,7-tetrabromo-1H,3H-dihydro-2H-benzimidazole-2-thione	2 Cl, 2 I	0.330
N/A	4,5,6,7-tetraiodo-2-trifluoromethyl-1H-benzimidazol	4 Br	0.370
1J91[5]	4,5,6,7-tetrabromo-2-trifluoromethyl-1H-benzimidazol	4 Br	0.400
N/A[79]	5,6-dichloro-4,7-diiodo-1H-benzimidazole	2 Cl, 2 I	0.460

[a]The halogenated inhibitors against the CK2 protein kinase[79,81–84] have been used to evaluate the positive-extra-point[85] and explicit sigma hole[77] corrections to the AMBER force field for MM/MD simulation. The number and types of halogen substituents along with their experimentally determined inhibition constants (K_i) are listed for each inhibitor.

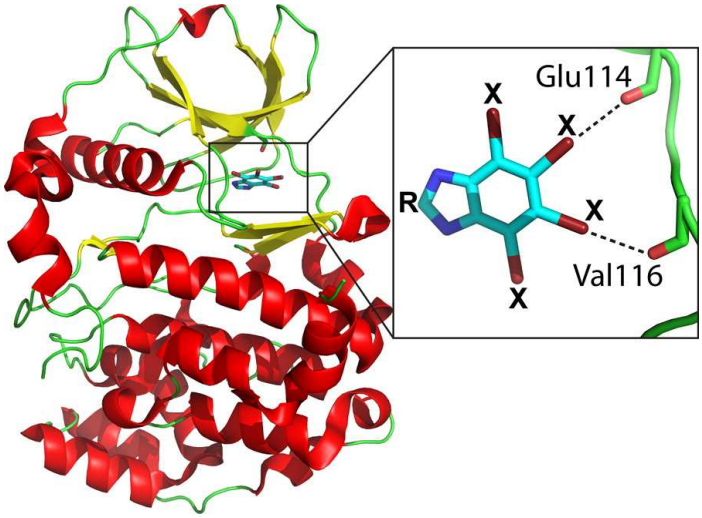


Figure 2.8. Structure of CK2 in complex with a tetrabromobenzimidizole inhibitor. The inhibitor forms two BXBs: one to the carbonyl oxygen of Glu114 and the other to Val116 (dotted lines in the inset). The X and R represent sites where halogens (X) or other substituent groups (R) are attached (see Table 2.2 for list of inhibitors)[78,79].

molecular dynamics (MM/MD) algorithms. The intra- and intermolecular interactions in folded structures of large biomolecules and their complexes with inhibitors are simulated by applying a Coulombic function ($V_{Coulomb}$) to model the electrostatic potential (Eq. 2.4, where Z_X and Z_A are the partial charges of the X-bond donor and acceptor, respectively, e is the charge of the electron, D is the dielectric constant of the medium, and r is the distance separating the interacting atoms) and a Lennard-Jones 6-12 type potential (V_{LJ}) to model the dispersion and steric repulsion effects in these interactions (Eq 2.5, where ε_X and ε_A define the energy minima of the potentials and $R_{vdW(x)}$ and $R_{vdW(A)}$ define the van der Waals radii of the X-bond donor and acceptor, respectively). Although classical in form, these energy functions are parametrized against QM calculations on complexes of small model compounds and validated against empirical data.

$$V_{Coulombic} = \frac{Z_X Z_A e^2}{Dr} \qquad (Eq. 2.4)$$

$$V_{LJ} = \sqrt{\varepsilon_X \varepsilon_A}\left[\left(\frac{R_{vdw(X)} + R_{vdw(A)}}{r}\right)^{12} - 2\left(\frac{R_{vdw(X)} + R_{vdw(A)}}{r}\right)^{6}\right] \qquad (Eq. 2.5)$$

Atoms are treated in Eqs. 2.4 and 2.5 as near hard-spheres with uniform charges and shapes. The problem is that halogens as organic substituents are electron-rich and are typically assigned an overall negative partial charge. Thus, when approaching another electron-rich atom, such as the standard H- and X- bond acceptors (O, N, or S), these atoms will repel each other. Consequently, X-bonding becomes impossible, which will lead to rejection of all halogenated ligands as viable drug candidates[77,86]. Thus, the standard MM/MD force fields must be extended to account for the potential effects of X-bonding in the affinity of a halogenated inhibitor.

Positive-Extra-Point (PEP) Approaches. One of the first extended MM/MD methods developed to model halogenated inhibitor-protein complexes is the positive-extra-point (PEP) approach, first implemented by Ibrahim[87]. This approach was adapted from a computational

strategy to more accurately describe the structure and energies of H-bonding in MM/MD force fields[88–90]. For X-bonds, the PEP method accounts for the electropositive crown of the σ-hole by placing a massless positive pseudoatom at a position diametrically opposed to the C–X bond. The specific position relative to the halogen center and charge of the pseudoatom are optimized by calculating the energy of the halogenated ligand interacting with a particular Lewis base type using density functional theory (DFT)[91,92]. Within the context of the popular AMBER MM/MD force field[93], the PEP method successfully predicted favorable X-bond energies in various protein complexes with halogenated ligands[85,94]. For example, when applied to CK2 using the same test sets as Hobza et al. (Table 2.2), the method more accurately ranked CK2 inhibitors according to their inhibition constant (K_i) than the PM6-DH2X (seen in better correlation in Figure 2.9) and much better than AMBER alone[85,94]. It should still be mentioned, however, that this particular adaptation of the PEP did not predict interaction energies in absolute terms, nor was it able to accurately recapitulate the detailed X-bonding geometries observed in the crystal structures of the inhibitor-protein complexes. Finally, the method still relies on an initial DFT calculation. Despite these shortcomings, the initial success of the PEP method has spawned several variations that attempt to improve the geometries and energies of interactions of halogens in protein complexes.

A variation on the PEP approach offered by Hobza's group is the explicit sigma hole (ESH) model[95], which attempts to determine whether the QM calculation to assign the charge and position of the pseudocharge could be circumvented to make the method more efficient. In this study, the authors compared calculations based on QM, non-QM, and combinations of these in defining the PEP mimic of the σ-hole for seven different brominated inhibitors to CK2. The ESH models, when incorporated into AMBER, resulted in energy minimized structures with

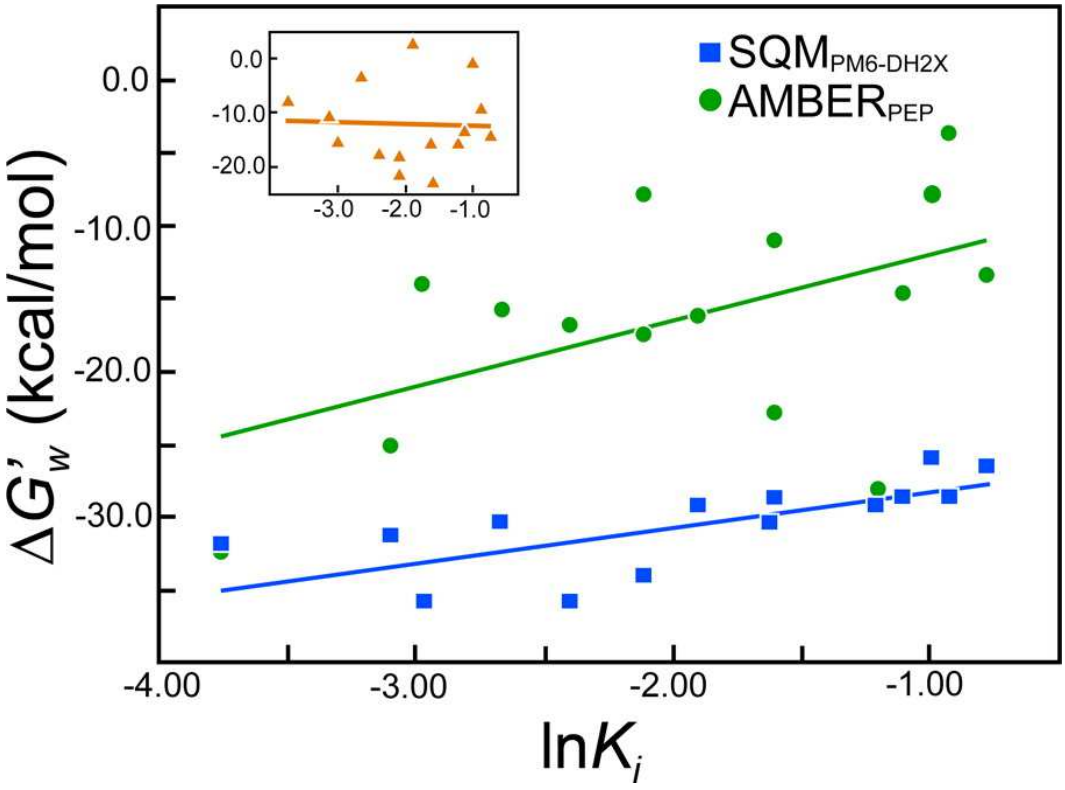


Figure 2.9. Comparison of force field calculated binding free energies ($\Delta G'_W$) to experimental inhibitor constants (ln K_i) for halogenated inhibitors against CK2. Binding free energies are calculated with either an X-bond correction implemented into AMBER with the PEP[85] method (blue squares, R = 0.69742) or an SQM PM6-DH2X[77] method (green circles, R = 0.52548). $\Delta G'_W$ calculated using an uncorrected AMBER force field are shown in the inset (amber triangles, R = 0.04950)[85].

geometries for the protein-inhibitor complexes that were closer to the experimental X-ray structures (as measured by root-mean-square deviations or RMSDs) than those simulated using the native AMBER force field.

OPLS-AA[96,97] was developed as a computationally efficient force field for molecular simulations of organic liquids and biomolecular systems[98]. In a 2011 study, Jorgensen's group applied this force field in a free energy perturbation (FEP) simulation to increase the potency of an HIV-1 reverse transcriptase (RT) from EC_{50} = 4.8 μM to 55 pM, an extraordinary example of lead optimization[4]. The design of the inhibitor started as a uracil base connected to a diphenyl-methane via an ethyl ether linkage. For this discussion, we will focus on two components of the study involving chlorine substitutions at the phenyl rings, although there were numerous other modifications, including substitution of the methyl linker between the two phenyl rings to an oxygen, resulting in a catechol diether analogue that ultimately led to the highly potent end product. FEP calculations on the catechol diether analogue with chlorines introduced systematically around the terminal phenyl group indicated that adding two halogens (at the 2,5-, 3,5-, or 2,6- positions) to this ring would greatly enhance the free energy of binding ($\Delta\Delta G_b$). Studies on potency also indicated that substituting the methyl group on the catechol ring with methyl chlorine increased the potency by an order of magnitude. For the most part, these calculations attributed the contribution of the chlorines to their abilities to fill spaces in the RT binding pocket. An interesting aspect of this study, however, was that the authors proposed a potential X-bonding interaction between the Cl at position 5 of the phenyl ring to the carbonyl oxygen of Pro95. The computed structure of the final optimized inhibitor with HIV-RT showed a distance for the atoms to be from 0.1 to 0.4 Å longer than the sum of their van der Waals radii (OPLS-AA does not incorporate X-bonding).

Soon afterwards, Jorgensen's group introduced the OPLS-AAx[99] and accompanying OPLS-CM1Ax force fields (CM1A charges are derived from AM1[65] calculations), which incorporated X-bonding using a strategy analogous to PEP; a massless pseudocharge (called an x-site) is added to each halogen as a mimic of the positive σ-hole. The resulting force fields were parametrized against QM calculated energies of aryl halide-acetone complexes in the gas phase. Comparison of FEP calculations on the optimized HIV-RT inhibitor from their 2011 study showed that $\Delta\Delta G_b$ for the protein-ligand complex became 1 to 2 kcal/mol more negative when X-bonding is included in the force field, including enthalpic, entropic, and solvent terms.

Force Field for Biological X-Bonds (ffBXB). Despite their successes, the PEP and ESH approaches are work-arounds in adapting force field potentials to accommodate the unique directional properties of X-bonds. Our lab has taken the approach that an empirical force field specific for biological X-bonds (*ff*BXB) could be derived based on fundamental atomic properties of halogens that contribute to the formation of BXB[17,18]. For this derivation, we consider the primary observations that the geometric and energetic properties of X-bonds result from the anisotropic distribution of charge and nonuniform atomic shape of halogens as covalent substituents. Incorporating these structural properties into the potential energy functions of Eqs. 2.4 and 2.5 simply required introducing angle-dependent functions to describe the charge of the halogen ($Z_X = A \cos(\pi - \theta_1) + B$, where the parameter A describes the magnitude of the charge anisotropy and B the overall charge of the halogen surface) for $V_{Coulomb}$, and the effective van der Waals radius ($R_{vdW(X)} = < R_{vdW(X)} > + \Delta R \cos(\pi - \theta_1)$, where $<R_{vdW(X)}>$ is the average effective radius, and ΔR is the perturbation from spherical) for the steric repulsive component of V_{LJ}. The resulting *ff*BXB functions were parametrized against a set of QM-calculated X-bonding energies calculated for model donor-acceptor pairs with various distances and angles of approach[18]. The

parametrized model shows one-to-one correlation between the calculated and experimentally determined energies for specific geometries observed for Cl, Br, and I halogen bonds (Figure 2.10)[18,100]. This promising correlation suggests that once the *ff*BXB has been implemented into drug design programs, it will be able to accurately model X-bonds in protein-inhibitor complexes as the last step in the virtual lead optimization strategy outlined in Figure 2.7.

■ 2.4 BXBS IN SCORING FUNCTIONS FOR VIRTUAL SCREENING

Inhibitor Screening.

Discovery of a lead compound that specifically inhibits a clinically important enzyme or protein target is typically identified through high-throughput screening of a small molecule library either experimentally or virtually. Introducing X-bonding into the virtual discovery process is not as straightforward as lead optimization but can apply many of the same computational concepts, adapting them for the additional steps of constructing halogenated libraries and docking such compounds to potential binding sites on their protein targets. Boeckler's group has taken a major step in this direction by creating a virtual fragment library, called the Halogen-Enriched Fragment Libraries (HEFLib) that is enriched in halogenated aryl compounds.[101] HEFLib takes advantage of the electron-withdrawing ability of aromatic systems to enhance the size of the σ-hole and, consequently, the strength of a potential BXB. The HEFLib was tested against the Y220C destabilizing variant of the p53 tumor suppressor (Figure 2.11a), a mutation associated with various types of cancers[102]. A previous virtual screen had identified PhiKan083 as a compound that could bind and stabilize this p53 variant (Figure 2.11b)[103]. Virtual screening using the HEFLib identified 79 compounds that could have this same stabilizing effect. Each of the lead compounds from the HEFLib virtual screen were then synthesized and physically characterized for their binding specificity. Through this process, one lead compound was

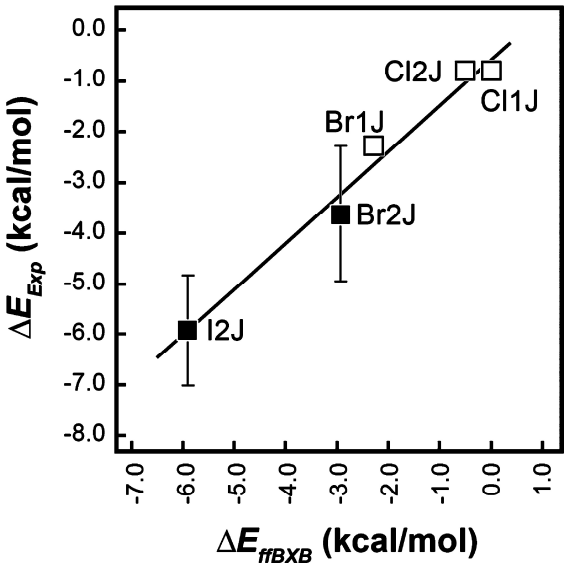


Figure 2.10. Correlation between *ff*BXB calculated (ΔE_{ffBXB}) and experimentally determined (ΔE_{Exp}) BXB energies[18]. BXB energies were measured by differential scanning calorimetry (filled squares) or by an X-ray crystallographic titration assay (open squares) in a DNA junction model system[15]. The Cl1J and Cl2J constructs competed one and two chlorine BXBs (respectively) against two H-bonds in stabilizing the junction. Similarly, the Br1J and Br2J competed one and two bromine BXBs, and the I2J competed two iodine BXBs in this system.

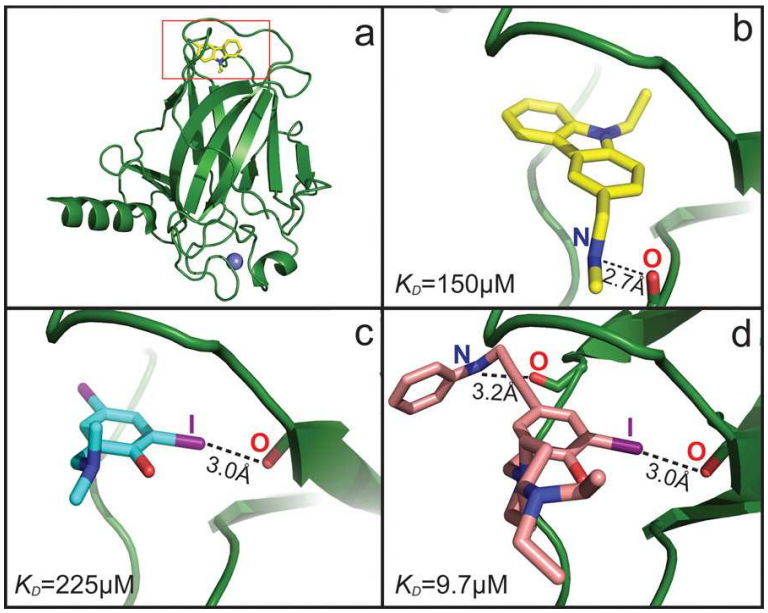


Figure 2.11. Structures of compounds from the virtual screening of X- bonding library (HEFLib)[101] against the Y220C mutant of p53. (a) Structure of the p53 protein with the inhibitor binding site highlighted by a red box. (b) Lead compound PhiKan083, found using traditional screening library, shown in the p53 binding pocket with the H-bond highlighted. (c) Halogenated compound identified using HEFLib with BXB in the inhibitor site. (d) Optimized compound from (c) with stabilizing BXB and a H-bond in the p53 pocket.

identified, which contained an iodine substituent with a K_D = 225 μM (Figure 2.11c). Further optimization resulted in a series of iodinated analogues with even higher affinities (K_D <20 μM) to p53 compared to PhiKan083 (example in Figure 2.11d). The increase in affinity and specificity can partly be attributed to the introduction of an X-bond. The early success of the HEFLib emphasizes the potential benefits of incorporating BXBs in the discovery stage of drug design, including the virtual screening of potential lead compounds.

Scoring Functions and Docking Programs with X- Bonds.

The process of virtual screening of libraries is a molecular docking problem–engaging multiple compounds with a target in virtual space and analyzing the optimum (energetically favorable) geometries for those complexes to identify potential leads. Scoring functions in docking programs offer a benefit to screening libraries by introducing a computational way to estimate binding affinities throughout the high-throughput inhibitor screening process[104]. Generally, there are three types of scoring functions applied to docking programs: knowledge-based, empirical, and force field-based (Figure 2.12). We focus here on how X-bonds have been incorporated into knowledge-based and empirical scoring functions.

Knowledge-Based Scoring Functions. Knowledge-based scoring functions are constructed through pairwise interaction networks that are defined around experimentally determined properties of molecular interactions, in particular, structural data from the Protein Data Bank[57] (PDB). Such functions sacrifice accuracy in favor of computational efficiency; however, this trade-off has particular advantages when searching across a very large virtual space, for example, during the initial screening phase. A simple example of an early knowledge-based scoring function is seen in the studies on protein folding and docking[105]. In these methods, the goal is to construct a set of data-derived potential energy functions based on what is observed in known

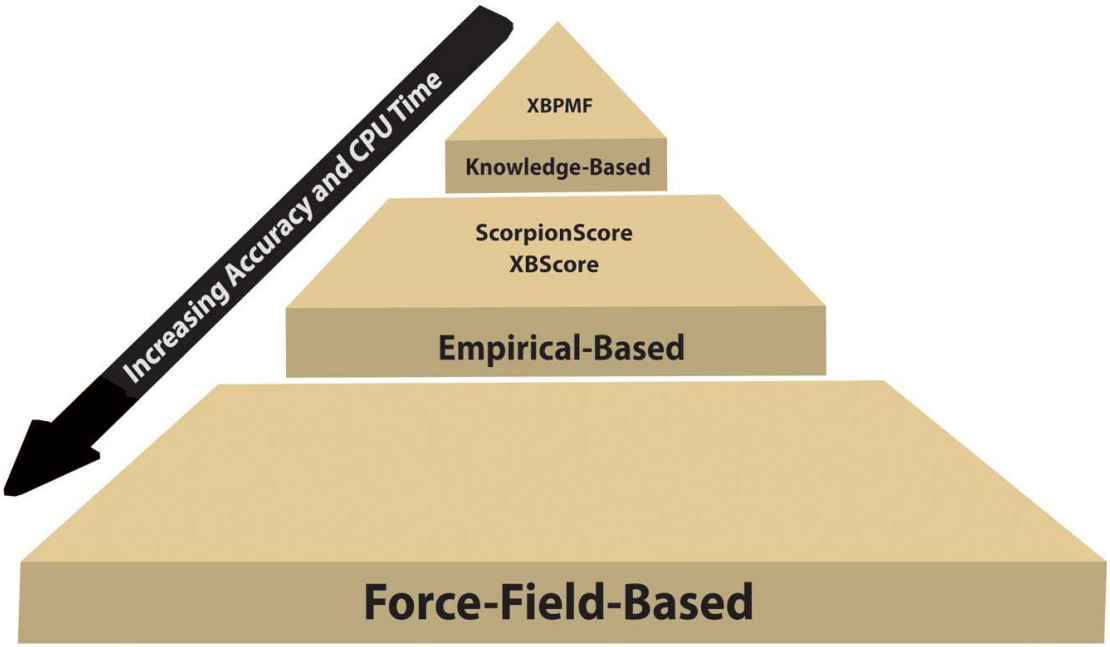


Figure 2.12. Hierarchy of scoring functions. The three common types of scoring functions to estimate binding energies during virtual screening/lead discovery stages of drug development are shown. Force field-based functions are the most accurate but computationally most expensive. These are followed by empirical- and knowledge-based with each being less accurate but more efficient. Specific scoring functions with BXBs incorporated are listed for the empirical- and knowledge-based functions.

folded protein structures to facilitate predictions of folded structures for an unknown sequence. For example, consider the interactions of amino acids in a folded protein in space. The tendency for a particular amino acid to pair with other amino acids will define a pairing potential. Thus, the frequency for which a particular pair is observed in the PDB structural database relative to an ideal reference pairing is expected to follow a Boltzmann distribution across all possible pairings, which can then be translated into a potential energy for pairing by a standard Boltzmann energy relationship. The challenge is that ideal pairing frequencies are not known and, therefore, an iterative process is used to estimate these reference frequencies. Once determined, however, these pairing energies can then help drive the folding of a test sequence, even if the physicochemical reason for the pairings are not well understood. For inhibitor screening, the strategy is extended to the atomic level, where knowledge of structural properties, such as distances, can be used to derive energies for atomic pairs.

In an early attempt at deriving a knowledge-based scoring function for ligand binding, Zhu et al.[106] developed XBPMF, which considers geometric and energetic preferences for molecular interactions, focusing specifically on incorporating X-bonds for screening halogenated compounds. The pairwise potentials considered H- and X-bonding energies and were derived using training sets composed of crystallographic structures of various protein-ligand complexes (solved to 3.0 Å resolution or better) extracted from the PDB. Because X- and H-bond energies are dependent on both distance and angle of approach, the potentials consider both geometric properties for atomic pairs, defined as 2-D potentials for the molecular interactions. The derivation of the potential energy functions require a set of known ideal distributions for the 2-D parameters, which were derived through an iterative approach, in this case against a set of decoy ligands. The XBPMF scoring functions were parametrized using training sets that consisted of all protein-

ligand complexes that met their selection criteria (31,145 structures) or those with only halogenated ligands (1,591 structures). The resulting knowledge-based potentials for X- and H-bond donor/acceptor pairs defined optimal geometries that are short-range and linear with interaction energies that vary according to the types of halogen donors and their associated acceptor atoms.

The success of XBPMF scoring functions in docking was evaluated by comparison against the several widely used scoring functions that use only one structural property (distance) in scoring and, therefore, are considered to be 1-D functions: LigScore1[107], LigScore2[107], GlideScoreSP[5], GlideScoreXP[6], DrugScore[108], Jain[109], PMF[110], and PMF04[111]. The three criteria for evaluation of the scoring functions were docking power (ability of the binding score to identify the native ligand conformation away from a set of decoys, as reflected in the RMSD of the best-scored ligand conformation), ranking power (the ability of the score to properly rank a set of ligands against a particular protein in order of affinity), and scoring power (the correlation between the binding score and the experimentally determined binding affinity of ligands to a protein target). In the docking power test, the success of the XBPMF scoring function was about average overall compared to other functions (falling in the middle group between the LigScore functions at the top and the PMF-type functions at the bottom). The ranking order evaluation was based on the ability of the functions to properly place in order the affinity of a set of ligands against eight protein targets (carbonic anhydrase II, casein kinase-1, coagulation factor X, HSP 90-alpha, HIV protease, tyrosine-protein phosphatase non-receptor type 1, beta-trypsin, and urokinase-type plasmogen activator). In this test, XBPMF was the top performer, yielding Spearman correlation coefficients for rank ≥ 0.6 for 4 of the test cases, and 2 other functions properly scored 3, 3 functions scored 2, and 1 scored only 1. Finally, the performance of XBPMF was average on the scoring power evaluation against three different test sets. Overall, it is fair to say that

incorporating the angular dependence of X-bonding into a knowledge-based scoring function, such as XBPMF, works best when the complexes show near ideal X-bonding geometries (i.e., strong X-bonds) but not markedly better than without the additional parameters for complexes that are less than ideal.

More recently, the XBPMF approach was extended by incorporating QM scan grids to account for the anisotropic electrostatic potentials of X-bonds (incorporating both distance and angle dependencies). It is the first scoring function to be incorporated in the widely used AutoDock program[112] for high throughput virtual screening. The resulting function in AutoDock showed improved performance compared to the original XBPMF and the various other scoring functions in the docking power, ranking power, and scoring power. However, AutoDock itself could account for the majority of the improvement in these evaluative measures with the X-bonding component showing slight improvements or, in some cases, slightly detriments to the parent program. In particular, the scoring function only improves the predicted binding energy slightly (from 1 to <0.5 kcal/mol difference) or for some not at all relative to AutoDock alone.

Empirical Scoring Functions. Alternatives to knowledge-based scoring functions are the empirical scoring functions, which attempt to recapitulate experimental binding energies through a series of individual parametrized functions (which may or may not be based on physical properties). The form of these functions varies drastically depending on how they are implemented into a particular program, but they are all parametrized against training sets derived from experimental data[113].

ScorpionScore is an empirical scoring function developed to consider how the environment modulates noncovalent interactions, including X-bonds, in ligand binding by proteins[114]. The authors create a "small world" interaction network (described by Watts et al.[115]), which considers

local cooperative effects instead of independent pairwise interactions. The scoring functions in ScorpionScore define favorable and unfavorable noncovalent interactions as well as local networks of interactions based on geometric (distance and angles) constraints. Several training sets of crystal structures were used to parametrize the scoring functions. We will focus on the binding of a chlorinated inhibitor to hepatitis C viral polymerase. A substitution at the para position of the aromatic ring on the inhibitor had previously been shown to decrease the IC_{50} from >32 μM for the unsubstituted to 1.2 μM for the chloro-substituted inhibitor, correlating well with the halogen contributing 2.7 to the overall favorable Scorpion score of 5.2 for the complex. Although the favorable score was attributed primarily to the close packing of the chlorine in a hydrophobic pocket, the halogen is in a geometry that can form a potentially stabilizing X-bond to the ε-nitrogen of an arginine side-chain (Figure 2.13[116]).

Boeckler's group recently derived an empirical scoring function, XBScore[117], which, unlike the ScorpionScore, is not designed as a screening mechanism for all possible inhibitors to a particular target but rather is directed to specifically identify halogenation sites on a predefined ligand scaffold that would potentially introduce X-bonding–a strategy similar to that described for lead optimization in Figure 2.7. In particular, XBScore focuses on the most common X-bonding interactions involving halogenated aromatic ligands as donors and the carbonyl oxygen along the protein backbone as acceptors[16,43]. The geometric constraints of BXBs are incorporated into XBScore through look-up tables: SigmaHoleScore as determined from polynomial functions for the directionality and SphericalScore for the available interaction sphere at various distances (Figure 2.14). XBScore is applied by first developing a scaffold decoration algorithm, where over 230,000 unsubsti- tuted and aromatic carbons were identified in the PDB (97,362 files) as potential halogenation scaffolds. After identification of the scaffolds, the algorithm

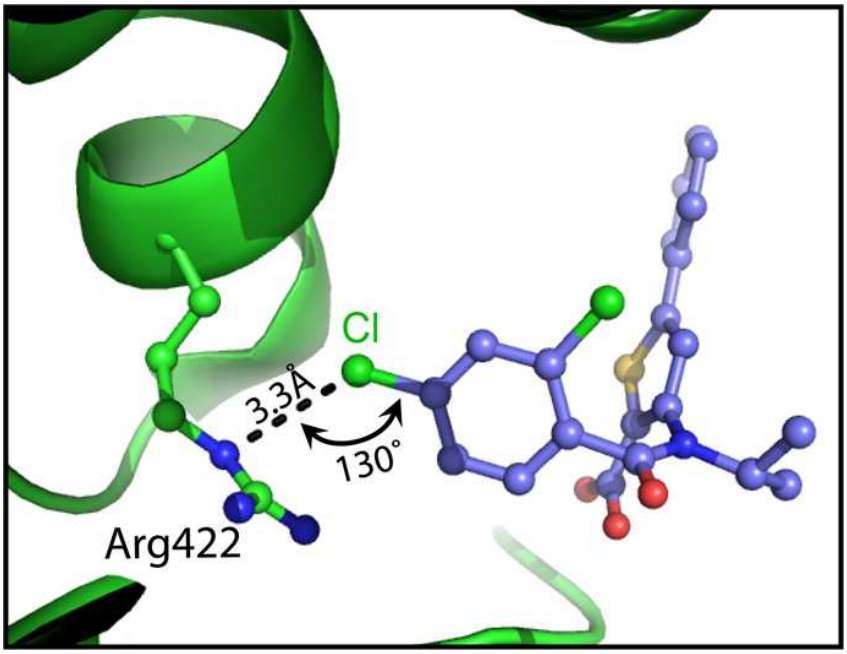


Figure 2.13. Potential stabilizing BXB in the hepatitis C viral polymerase-inhibitor complex (PDB ID: 1YVZ[116]). The para-chloro- substituent off the phenyl ring of 3-[(2,4-dichlorobenzoyl) (isopropyl)amino]-5-phenylthiophene-2-carboxylic acid, a non-nucleo- side inhibitor, is in close contact (3.3 Å) with the ε-nitrogen of the Arg422 side-chain in the polymerase binding pocket.

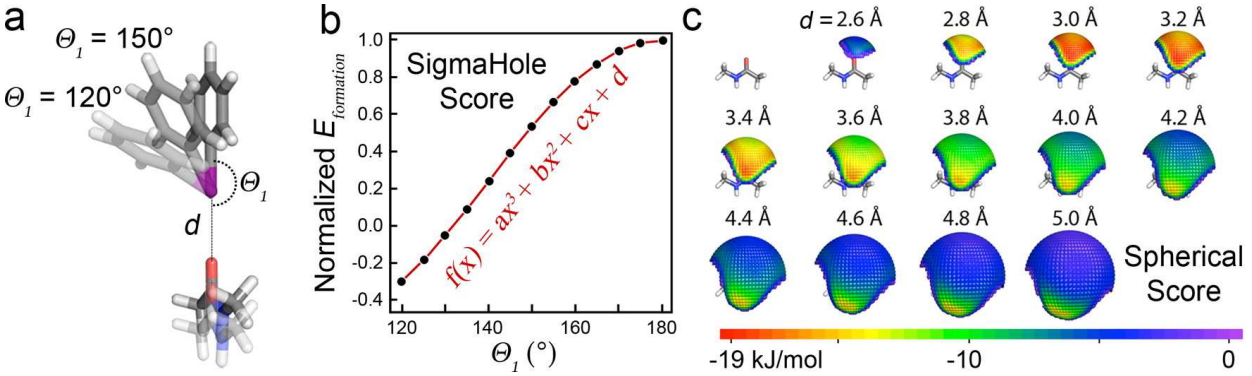


Figure 2.14. Components of the XBScore scoring function[117]. XBScore is made up of two separate components (SigmaHoleScore and SphericalScore) that approximate the angle and distance dependence of a BXB energy. (a) The interaction between the carbonyl on N-methylacetamide and the halogen on a halobenzene was used to model both components. (b) The SigmaHoleScore describes the directionality of X-bonding where the normalized complex energies at various angles relating the acceptor approach to the halogen (Θ_1) were fit with a polynomial function. (c) The SphericalScore describes the spatial position of the donor to the acceptor. Different distances (d)define different possible points of contact from the BXB acceptor to donor. Because no mathematical relationship was found, a lookup table is used to assign the SphericalScore.

"halogenates" various carbons with Cl, Br, or I. The algorithm then identifies those halogens that are near carbonyl oxygens and calculates their XBScore. Remarkably, a scan of the PDB with the XBScore algorithm was able to extract all structures with known X-bonds to the protein backbone with distances between 2.6 and 5.0 Å. Using the scaffold decoration, they found potential structures that would benefit from X-bonding. Not only do the resulting complexes recapitulate QM binding energies with good correlation, they have also been able to identify possible target sites on the scaffold for the introduction of BXBs.

■ 2.5 CONCLUSIONS AND PERSPECTIVES

Halogenated compounds have a long history in therapeutics, including the introduction of chloramphenicol as an antibiotic in 1949[118], with a more recent example being flavopiridol (HMR 1275), a chlorinated CDK2 inhibitor currently in human clinical trials as an anticancer agent[119]. For the most part, the inclusion of halogens, particularly the heavy halogens (Cl, Br, and I) into a compound for hit-to-lead and lead-to-drug steps in drug development have occurred empirically through trial and error as opposed to purposeful design. When halogens have been included in rationale design strategies, they are added primarily to improve bulk properties of the compound, such as membrane permeability or, at the molecular level, to fill spaces in binding pockets. For example, a recent systematic study on the effect of halogen substituents on the ability of compounds to penetrate cellular membranes showed that replacing a hydrogen in promazine and perazine with a chlorine to make chlorpromazine and chlorperazine, respectively, increases the permeability coefficient on average by a factor of 2[120].

In the past decade, there is a growing appreciation that halogens play direct roles in defining the affinity and specificity of inhibitors against their protein targets, specifically through a molecular interaction that is now defined as the halogen bond or X-bond[10]. The first known

application of the X-bonding concept to structure-based design of a potential therapeutic compound was in 1996 when P. Lam was inspired to replace a H-bonding amidine group with an X-bonding iodine in an inhibitor to Factor Xa[121]. Our understanding of the fundamental principles of X-bonds and their physicochemical properties have grown significantly in parallel with the exponential growth in interest in the interaction, resulting largely from the expanding application of X-bonding in materials science[122,123] and from the recognition that X- bonds play a broader role in biology[124,125]. Thus, the application of halogen chemistry to drug design now keeps one eye on how X-bonds in biomolecular systems (biomolecular X-bonds or BXBs) can facilitate the construction of more effective inhibitors, including those against important pharmacological targets such as HIV and cancer. Halogenation, however, is not always a positive or even benign modification. There have been toxicity issues associated with halogens that cause off target effects or that create toxic small molecules[126,127]. Thus, as with all drug design strategies, we need to remain aware of potentially harmful side effects associated with introducing halogens with the potential to increase specificity through X-bonds.

A more deliberate exploitation of X-bonds as molecular tools in medicinal chemistry, however, has until recently been hampered by the lack of computational tools to properly model BXBs, particularly the anisotropic charge distribution. The excitement in the area of BXBs has drawn attention to new research approaches to attacking this computational problem. There are now several approaches to incorporating BXBs into virtual lead discovery and optimization from full QM computations to hybrid QM/MM to SQM to force field to empirical and finally structure knowledge-based strategies (going from highly accurate and computationally expensive to highly approximate and computationally efficient). We have seen some impressive success, particularly in simulating the geometries and energies of known protein-ligand complexes, leading in some

cases to successful optimization of inhibitors with higher affinities and efficacies against their molecular targets. For complexes where accurate structural information is lacking, however, the success has been more moderate, and for true virtual screening applications, the introduction of BXBs appears to primarily affect the details of docking results, which is not surprising because a single molecular interaction can only add to 1−5 kcal/mol to the total binding enthalpy (which can be upwards of tens to hundreds of kcal/mol). In effect, including BXBs improves the accuracies in molecular simulations of protein complexes with halogenated ligands but cannot overcome the inherent problems associated with the more general problem of accurately predicting de novo the geometries and energies of such complexes, whether halogenated or not.

One of the primary issues in applying molecular simulations to structure-based drug design is that there are more components to binding affinity than just the final static structure seen in the crystal structure of a complex. For many of the studies testing the accuracies of BXB corrections, the evaluations have been limited to comparing the calculated X-bonding energy to the rank order of potency, as measured by K_i, IC_{50}, and so forth. This assumes, however, that the effect of the halogen on affinity is restricted only to their effects on enthalpies of interaction–certainly an improvement over not including BXBs at all. However, potency and even affinity are dependent not only on the strength of a particular molecular interaction but on the overall effect of free energy differences between the complex and the individual molecular components, including the solvent. The work from Jorgensen's group[4] has come a long way toward incorporating many of these important concepts in his studies, demonstrating that systematic and detailed simulations through free energy perturbation methods can contribute to lead optimization that results in upward of 105-fold increase in affinity with ∼103 coming from effects of halogen substituents. These studies set the bar for how to evaluate the efficacy of a BXB correction. A greater

challenge is to accurately predict binding free energies when starting with the native uncomplexed structure of a protein target in a virtual screening environment. Boeckler's strategy, with the halogen enhanced library[101] and X-bond focused scoring function[117], is an example of a comprehensive approach to incorporating BXBs at the discovery stage.

As the field continues to develop, the cache of computational tools to model X-bonds is expected to grow. Along with this growth and the improvements in accuracy in the computational models, we hope and expect that the algorithms for simulating all aspects of macromolecular structure, including the more ubiquitous H-bonding and solvent interactions, will improve in parallel. Only then will structural-based rational drug design fully reach its potential in medicinal chemistry.

CHAPTER 3

SULFUR AS A BROMINE BIOMOLECULAR HALOGEN ACCEPTOR [2]

■ 3.1 SUMMARY

The halogen bond (X-bond) has become an important noncovalent interaction for the design of pharmaceuticals. The X-bond is able to form an interaction with many electronegative acceptors, such oxygen, nitrogen, sulfur, and aromatic systems. In this study, we explore the role of sulfur as a bromine biomolecular halogen acceptor by assessing the structure and thermodynamics of a S···Br X-bond present at the core of a stacked-X Holliday junction (a four-stranded DNA structure). The sulfur incorporated into the phosphate of the DNA backbone adopts two conformations–one forming a more geometrically ideal X-bond and the other forming a less ideal X-bond with a bromine atom on a uracil base. Both conformations were, however, found to be in a less ideal geometry than a previously studied O acceptor. Surprisingly, using calorimetry, the S acceptor is found to be more thermodynamically stable than the O acceptor by 0.43 kcal/mol. Thus, for the first time, we have shown that an anionic sulfur can serve as a favorable acceptor for bromine X-bonds in a biological system. With the current understanding of sulfur as a favorable X-bond acceptor in less ideal geometries, medicinal chemists have an attractive target when designing halogenated inhibitors.

■ 3.2 INTRODUCTION

Interest in halogen bonds has grown, corresponding with its emergence as a tool in medicinal chemistry[58,86,128,129]. The term "halogen bond"[10] (X-bond) has been adopted to reflect its

[2] The work in this chapter is formatted for *an ACS journal*. P.S.H. and I planned the studies and wrote the manuscript. I performed the purifications, DSC studies, and crystallographic studies. M.S. and I performed the QM.

Formatted for submission.

analogous behavior to the better known hydrogen bond (H-bond). In particular, any electron-rich atom or functional group that serves as an H-bond acceptor can potentially be an X-bond acceptor (Figure 3.1a). Most recent studies have focused on oxygen, nitrogen, and π-electron systems as X-bond acceptors, because of their prevalence in biomolecules. Sulfur is also commonly seen as an acceptor, and is gaining attention as a potential target in drug design using the X-bond concept[43,130,131], but there is currently very little experimental data on its X-bonding potential. Here, we apply a DNA Holliday junction as a model system to determine the structure-energy relationship of S···Br X-bonds, showing that sulfur is slightly stronger as an X-bond acceptor relative to oxygen.

The X-bond, similar to the H-bond, is considered to be primarily an electrostatically driven interaction[13,31,132], although charge-transfer, dispersion, and polarization contribute significantly at very short distances[11,12,30]. In the case of the X-bond, the donor is seen as an electropositive crown created at tip of the halogen as a consequence of depletion of electron-density opposite a covalent σ-bond (*e.g.*, a C—X bond) (Figure 3.1b)[13]. The resulting "σ-hole" can form an interaction with an electron-rich acceptor, with a geometry that is characterized as being shorter than the sum of the van der Waals radii ($\sum r_{vdW}$) of the respective atoms and linear in respect to the approach of the acceptor atom to the halogen (the θ_1-angle). The σ-hole is most pronounced with the heavier halogens (Cl, Br, and I)—F is not a significant X-bond acceptor in biology—and is accentuated by electron-withdrawing groups.

The most common X-bond acceptor in biomolecules is the carbonyl oxygen of the peptide bond, primarily because of its prevalence, although the amino acid side chains collectively are nearly equivalent in their representation[16,43,130,133]. The sulfur containing amino acids, particularly Cys, are underrepresented as X-bond acceptors[133], but quantum mechanical (QM)

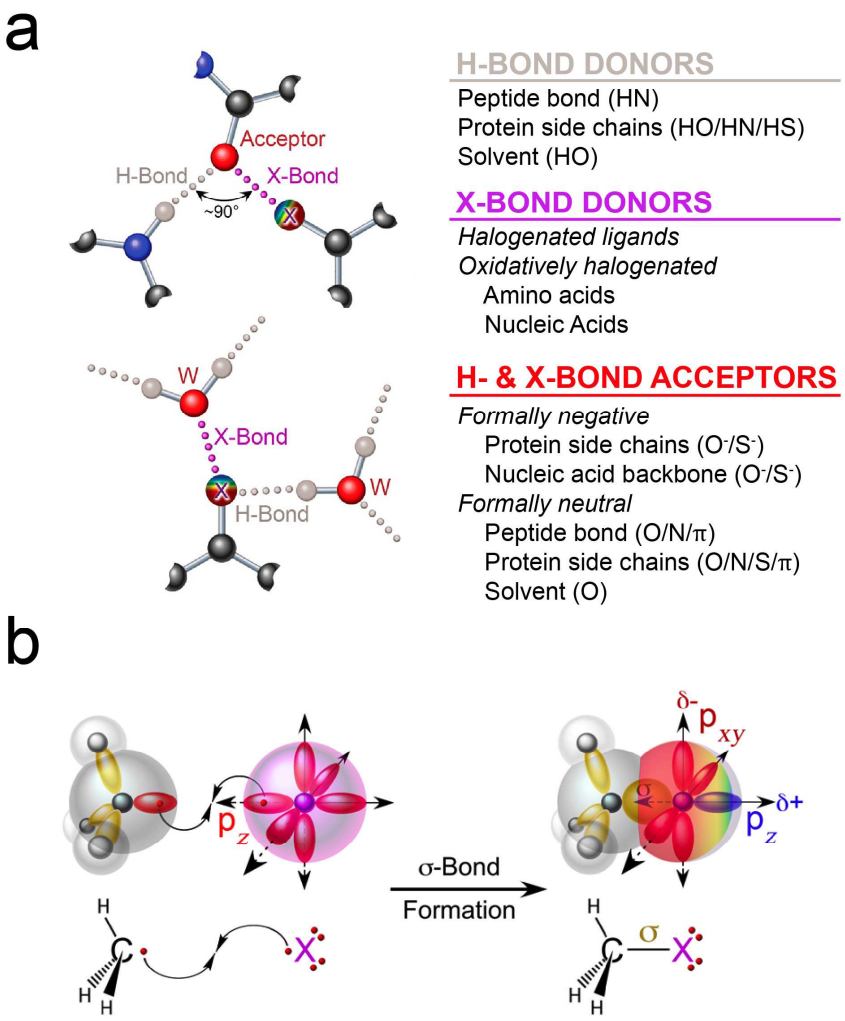


Figure 3.1. The X-bond. (a) Comparison of H- and X-bond donors and acceptors in biological systems, and relationships between the two. H-bond donors are typically the electropositive hydrogens bonded to more electronegative atoms (O, N, or S) in proteins or nucleic acids. In contrast, X-bond donors are primarily halogenated ligands that bind noncovalently to proteins or nucleic acids, but can also come from halogenated amino acids or nucleic acids that result from oxidative halogenation. Since both H- and X-bonds are primarily electrostatically driven interactions, they share a common set of electron-rich acceptors. When an H- and X-bond shares a common acceptor, such as the carbonyl oxygen of the peptide backbone in a protein, the interactions are known to be orthogonally related (both in terms of geometry and in their energy independence)[40]. The anisotropic charge distribution results in halogens being amphipathic, allowing them to serve simultaneously as X-bond donors and H-bond acceptors[134]. (b) σ-Hole model for X-bonding. A σ-hole is shown to form a halogen becomes covalently bonded to, for example, the carbon of a methyl group. The p_z-orbital of the halogen is depopulated when the valence electron is subsumed by the resulting σ-bond molecular orbital, creating a "σ-hole", which accounts for the electropositive charge and polar flattening in the direction opposite the σ-bond. The electrons that remain in the perpendicular $p_{x,y}$-orbitals creates an anisotropic charge distribution across the halogen surface. Adapted from Auffinger et al[6] and Ford and Ho[4].

analyses have shown that the S in Met has potential as a target for inhibitor design from the perspective[131,135]. Experimental and computational studies on small molecule systems show sulfur as similar to slightly more favorable than oxygen as an X-bond acceptor[136], with geometries that deviate significantly from linearity[131,136,137]. In one study, an iodobenzene ligand bound in an artificial cavity of T4 lysozyme was shown to form an X-bond to the sulfur of a methionine or selenium of a selenomethionine residue[138]. The energetic contribution of this interaction, however, was shown to be very small (~0.6 kcal/mol), leaving the question open as to whether sulfur is a reasonable X-bond acceptor in a biological system.

Our lab had previously developed a DNA system to study X-bonds in a biological context[14,15]. In this model system, a 5-bromouracil, ^{Br}U (Figure 3.2) is placed at a nucleotide position that can form an X-bond, replacing a comparable H-bond, to stabilize a decanucleotide sequence (5'-CCGGT***p***$A_6{}^{Br}U_7$CCGG-3', where the ***p*** represents a standard phosphate group linking the T_5 nucleotide with A_6) as a four-stranded Holliday junction in the stacked-X geometry. The single-crystal structures of this construct (Protein Data Bank, PDB[57] code 2ORG) reveal the detailed geometries of the molecular interaction from the bromine of $^{Br}U_7$ to the phosphate oxygen of A_6 when the DNA backbone forms a tight U-turn in the junction, while differential scanning calorimetry (DSC) allow determination of the energetics of this interaction in solution[139]. Finally, QM analyses of the interacting groups within the DNA junction helps bridge the structural details with the associated energies[17]. We apply this same DNA junction system here to determine the structure-energy relationship of an S···Br X-bond by introducing a phosphorothioate (***ps***) in place of the standard phosphate linkage between T_5 and A_6, thereby allowing for the first time a direct comparison between sulfur and oxygen as biomolecular X-bond acceptors.

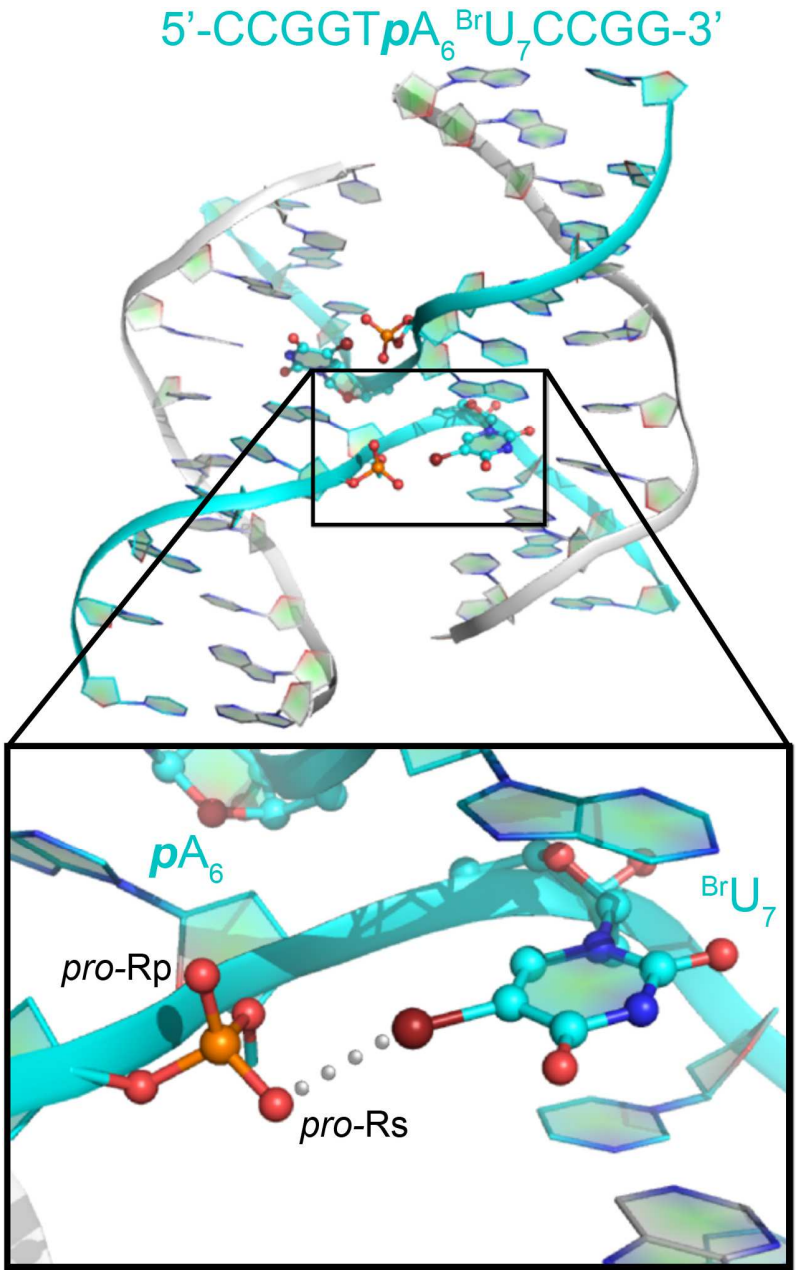


Figure 3.2. X-bonding interactions with oxygen as the acceptor at the crossover in a DNA junction. The overall junction (PDB code 2ORG[14]) is shown, with the inset depicting the 5-bromouracil base at $^{Br}U_7$ interacting with phosphate oxygen pA_6. Replacement of an oxygen with a sulfur in this phosphate group, as a phosphorothioate in the junction (***ps***-J), creates a prochiral center at the phosphorus, thereby defining the two prochiral configurations of the sulfur (labeled *pro*-Rp and *pro*-Rs).

■ 3.3 MATERIALS AND METHODS

DNA Synthesis and Purification

The DNA constructs for this study were designed as self-complementary sequences of the form 5'-CCGAT**X**^{br}UCGG-3' (where **X** has either a standard 2'-deoxyadenosine 5'-monophosphate, *p*A, or a 2-deoxyadenosine-5'-thiomonophosphate, *ps*A). Chemically synthesized DNA oligonucleotides are purchased from Midland Certified Reagent Company attached to the solid controlled-pore glass (CPG) support, with the terminal dimethoxytrityl (DMT) protecting group intact to allow for purification. The full-length product sequences were purified using high-performance liquid chromatography (HPLC), and subsequently the DMT was removed as previously described[14,15]. Phosphorothioate constructs required an additional HPLC step to separate the *pro*-Rp from *pro*-Sp enantiomers, using a protocol derived from that as described in Liu, *et al*[140]. The final product was desalted by size exclusion chromatography using a Sephadex G-25 Column.

Crystallization and Structure Determination

DNA constructs were crystallized by hanging drop vapor diffusion under conditions similar to those used previously to crystallize brominated DNA junctions[14]. The initial 6 µL crystallization drops containing 0.75 mM DNA, 25 mM sodium cacodylate buffer (pH 7.0), 10-20 mM calcium chloride, and 0.9-1.2 mM spermine. Thin diamond-shaped crystals of one of the *ps*A enanatiomers grew after 7 days equilibration against a reservoir containing 30% 2-Methyl-2,4-pentanediol and 1% ß-mercaptoethanol.

X-ray diffraction data on DNA crystals were collected using a Rigaku Compact Home Lab equipped with a copper sealed-tube microfocus source and a PILATUS detector. Data were indexed, integrated, and scaled using the HKL2000 program [141]. Crystal structures were solved

by molecular replacement using PHENIX[142], with a brominated junction (2ORG) in which the bromine had been removed, serving as the initial phasing model[14]. The actual enantiomer of the phosphorothioate of the S-construct was not known. The residual positive electron density in the initial difference Fo-Fc electron density maps indicated that the crystal of the S-construct was the *pro*-Rp enantiomer, and subsequently confirmed by the drop in R_{work} and R_{free} during refinement. DNA base pair measurements (rise, twist, slide, etc.) were performed with the CURVES+ DNA structure analysis program[143], and junction structure parameters (Jroll and Jtwist) were calculated according to the methods described by Watson et al[144].

Differential Scanning Calorimetry Studies

Differential scanning calorimetry (DSC) was performed to determine the melting energies of the sulfur-containing and oxygen-containing constructs, as previously described[15]. DNA constructs (300 μM concentration in a solution containing 50 mM sodium cacodylate buffer (pH 7.0) and 15 mM calcium chloride) were annealed in the DSC instrument (TA Instruments Nano DSC) by heating to 90 °C for 20 min then allowing to cool to 10 °C at a rate of 0.9 °C/min. DNA melting energies were collected in the DSC, starting with equilibration for 400 s and scanning from 10 °C to 95 °C at a rate of 0.9 °C/min at a pressure of 3.0 atm. Melting temperatures (T_M) and enthalpies of melting (ΔH_M) were determined by fitting the data with TA Nano Analyze software. Previous studies[15,100] had shown the presence of both junction and duplex at 300 μM DNA concentrations and, therefore, a two-component (junction and duplex), two-state scaled model was used to analyze the data (Figure 3.3). Data for each construct were measured for at least 7 replicates. Melting energies were extrapolated to a standard temperature of 25 °C, and the duplex melting energies were subtracted from the junction to determine the stabilization energy of the junction core as previously described[15,100].

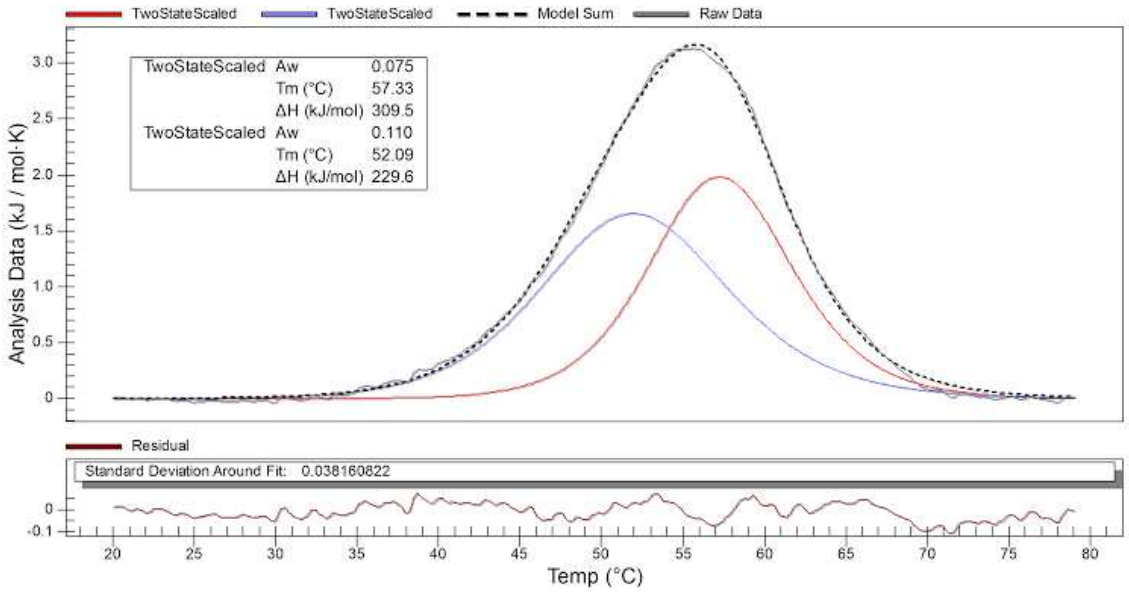


Figure 3.3. Example Differential Scanning Calorimetry (DSC) Two-Component Van't Hoff Fit. The raw data (solid gray line) was fit with a two-component, two-state scaled model, one for DNA duplex (solid blue line) and one for DNA junction (solid red line) to yield an overall fit (dotted black line).

Quantum Mechanical (QM) Calculations

Models for QM energy calculations of the X-bonded junctions consisted of a 5-brominouracil base interacting with either (dimethylphosphate) for the non-S construct or dimethylphosphorothioate (for the S-construct), with atomic position placed according to their respective crystal structures (Figure 3.4). The Møller−Plesset second-order (MP2) QM energies were calculated using Gaussian 09[145], with cyclohexane as the solvent (D = 2, relative to a vacuum) to mimic the hydrophobic environment of the junction core, applying an appropriate polarizable basis set that includes dispersion (aug-cc-PVTZ) from the EMSL Basis Set Exchange[146]. Basis set superposition errors (BSSE) were determined from a separate counterpoise gas phase calculation and directly summed into the calculated solvent phase energy[147].

■ 3.4 RESULTS

The goals of the current studies are to determine the structure-energy relationships of sulfur as an X-bond acceptor and to compare against the more prevalent oxygen acceptor in a biomolecular system. From previous studies[100], bromine was found to be overall the most stabilizing X-bond donor in the model DNA junction. We have thus designed a DNA junction in which a phosphorothioate (***ps***) linkage is introduced into the self-complementary sequence (CCGGT$_5$***ps***A$_6$^{Br}UCCGG, abbreviated ***ps***-J). The crystal structure of this ***ps***-J construct, when compared to the previous structure of 2ORG, allows the geometries of an S···Br X-bond to be compared to that of an O···Br interaction. Unlike previous studies, all four strands of ***ps***-J are identical, forcing an X-bond to form in all structures, instead of having an X-bond compete against an H-bond (as in the 2ORG structure). Thus, for the DSC melting studies, we constructed

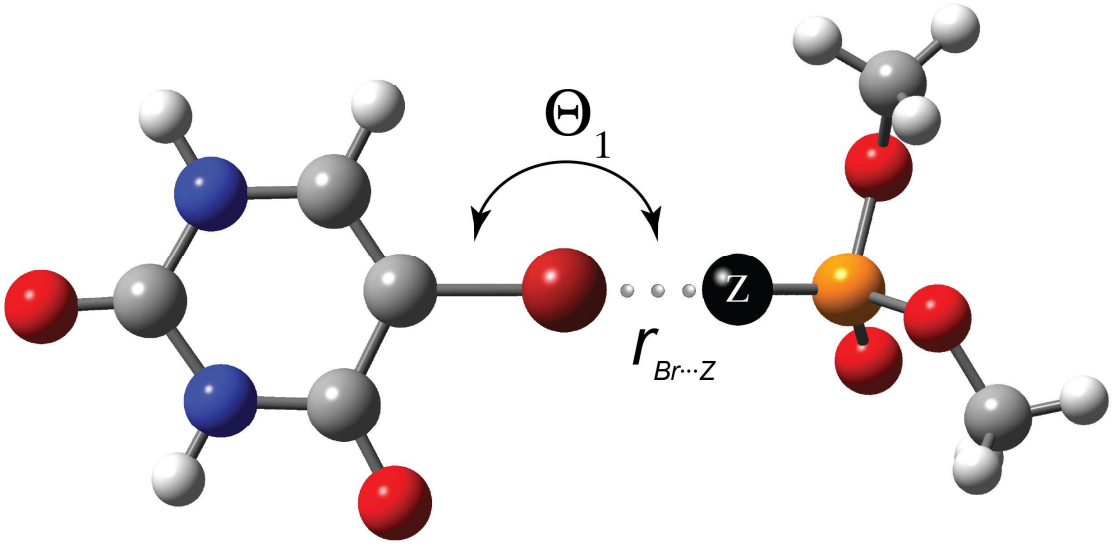


Figure 3.4. Quantum Mechanics Calculation Setup. 5-bromouracil was used as a model for the base of nucleotide 7 ($^{Br}U_7$). The bromine atom was interacting with either a dimethylphosphate (Z = oxygen) to mimic the interaction found in 2ORG or dimethylphosphorothioate (Z = sulfur) to mimic the interaction found in both conformations of 5VBJ. All three systems (2ORG, ***ps***-JA and ***ps***-JB) were modeled using geometries ($r_{Br...Z}$, Θ_1) found from the x-ray crystal structures.

a self-complementary sequence, analogous to ***ps***-J, but with a standard phosphate (***p***) diester linkage between T_5 and A_6 ($CCGGT_5\boldsymbol{p}A_6{}^{Br}UCCGG$, abbreviated ***p***-J).

X-ray Crystallography: Comparison of Structures

The overall structure of the ***ps***-J is that of the four-stranded DNA junction, similar to the stacked-X junctions seen previously (Table 3.1 and Figure 3.5a)[14], but with the sulfur from A_6 positioned as the *pro*-Rp enantiomer at the core of the junction and close to the bromine on $^{Br}U_7$. During refinement, it became clear from the electron density maps that the ***ps***-linkage assumed two alternative conformations (Figure 3.5b)—conformer A (labeled as ***ps***-JA, with 52% occupancy) and conformer B (***ps***-JB, with 48% occupancy). Evidence for the presence of two conformers came initially from careful analysis of the residual electron densities in the Fo-Fc difference map, and further supported by the reduced R_{work} and R_{free} values when both conformations were included in the final model (with R_{work} dropping from 22.5% to 21.8% and R_{free} from 27.8% to 27.6% with the addition of two conformations). Finally, support for the presence of both conformations came from analysis of the structural *B*-factors, which were higher for the T_5 and A_6 residues with only a single model, but reduced to be consistent with the remainder of the structure with two conformers (Figure 3.6).

The overall junction geometries (J_{twist}, which relates the rotational angle between the stacked arms along their helical axes, and $J_{roll,}$ which relates the stacked arms perpendicular to the helix axes) indicate that the global geometries of the two conformations of ***ps***-J are very similar to each other and to the previously determined structure of the junction (2ORG) containing O···Br X-bonds (Table 3.2). The primary difference between ***ps***-J and 2ORG is that the stacked duplex arms of the sulfur containing junction are related by a much shallower angle ($J_{twist} \approx 30°$ compared to ≥40° in all previous structures), making it a more planar junction structure.

Table 3.1. Parameters from Crystallization and Structure Solution of ***ps***-J (PDB 5VBJ)

Crystallographic Parameters	
Space group	*C2*
Unit cell dimensions	
a(Å)	64.165
b(Å)	23.981
c(Å)	37.815
ß(˚)	112.318
no. of unique reflections (for refinement)[a]	3562 (3499)
Resolution	1.94
Completeness (%)[a]	85.93 (42.57)
I/sigma (I)[a]	12.4 (1.13)
R-merge (%)[a]	8.8 (77.1)
Refinement Statistics	
R_{cryst} (%)[a,b]	21.8 (44.1)
R_{free} (%)[a,b]	27.6 (45.5)
no. of atoms of DNA (solvent)	431 (29)
<B factor> for DNA (solvent)	33.825 (37.2)
RMSD for bonds (Å)	0.009
RMSD for angles (˚)	1.17

[a]Values for highest-resolution shell are given in parentheses

[b]Values for R_{cryst} and R_{free} are very similar to previously published, 2ORG

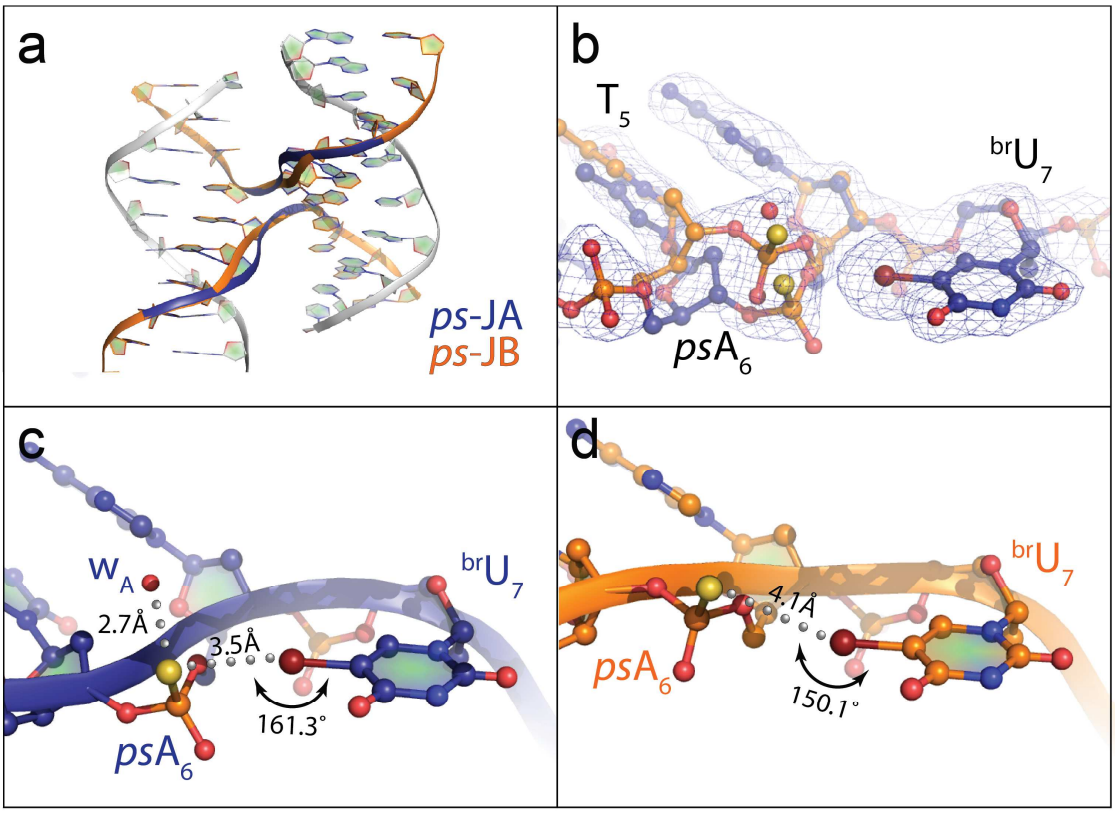


Figure 3.5. X-ray crystal structure of phosphorothioate DNA junction (*ps*-J). (a) The phosphorothioate DNA junction (*ps*-J) crystallizes as a stacked-X four-stranded junction. The junction was seen to adopt two conformations (*ps*-JA in blue and *ps*-JB in orange). (b) Electron density map at the junction cross-over. The 2Fo-Fc electron density map calculated at the 1σ level (blue wire) for the *ps*A_6 and brominated uracil ($^{br}U_7$) nucleotides is shown, indicating the presence of the *ps*-JA (blue carbons) and *ps*-JB (orange carbons) conformations (c) S from the phosphothioate of A_6 (yellow sphere) to Br of $^{br}U_7$ interaction at the junction core of *ps*-JA. The short distance and angle of approach are labeled for the S···Br interaction, along with distance to the associated water molecule (W1). (d) S to Br interaction at the junction core for *ps*-JB. The distance and angle of approach are labeled for the S···Br interaction.

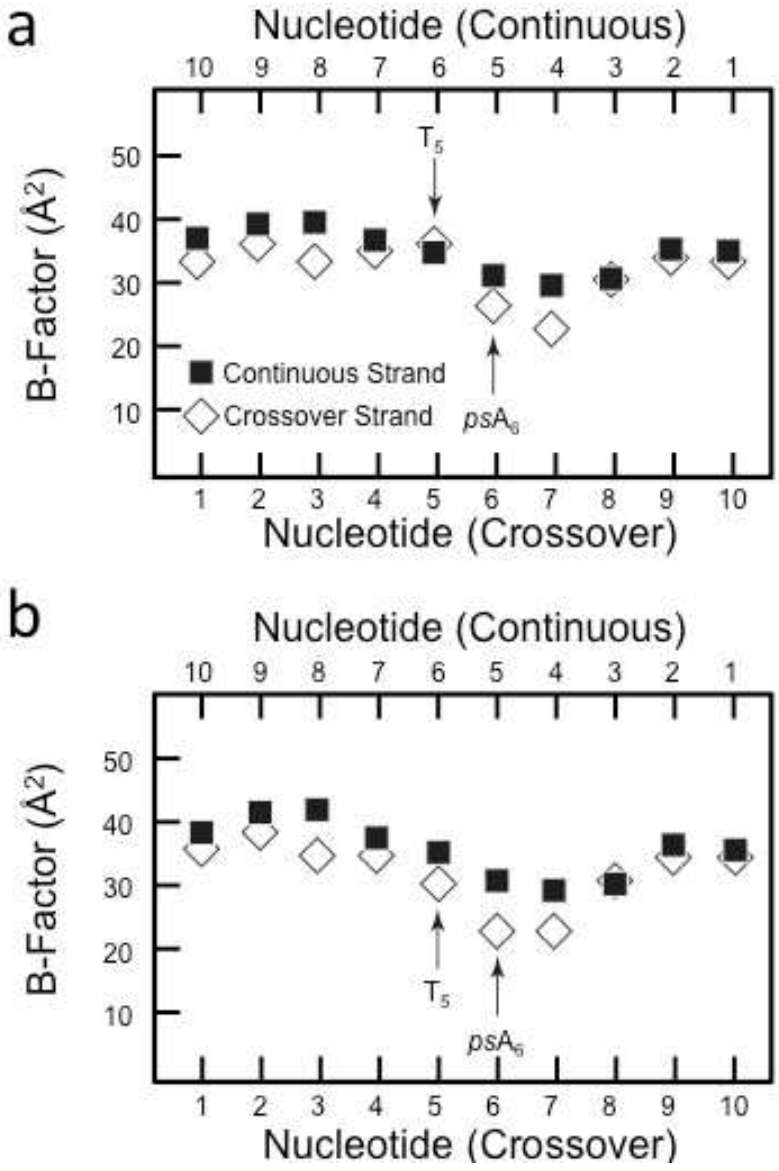


Figure 3.6. Temperature factors in ***ps*-J** crystal structure measured as (a) a single model or (b) two conformers. The average temperature factors (B factors) of the nucleotides along the continuous outside strand (□) and inside crossover strands (◆) of single model structure are compared to those when two conformers are added to T_5 and *ps*A_6 (indicated by the arrows). The B factors for the two conformers were averaged. (The standard deviations for each B factor are around or below 5.0 Å^2, which are about the size of the symbols). After the addition of the two conformers, the B-factors of the crossover nucleotides (5-7) are significantly lower than the continuous strand.

Table 3.2. Comparison of overall junction parameters and helical parameters for the $A_6 \cdot T_5$ base pair at the crossover of the phosphate (2ORG) and phosphorothioate (***ps***-JA, ***ps***-JB) junction constructs.[a]

	***p*-J (2ORG)**	***p*-JA**	***p*-JB**
	Junction parameters		
J_{twist}	40.68°	30.13°	30.16°
J_{roll}	138.70°	129.55°	124.82°
	Rotational helix parameters		
Helical Twist (34.7°)	36.85°	36.90°	36.95°
Propeller Twist (-12.0°)	-19.90°	-12.20°	-19.10°
Tilt (-0.62°)	0.15°	0.50°	1.30°
Roll (1.74°)	3.35°	2.60°	-1.50°
Buckle (-0.23°)	-7.80°	8.10°	-17.10°
Opening	3.10°	4.70°	5.80°
	Translational helix parameters		
Rise (3.30 Å)	3.29	3.21	3.27
Slide (0.66 Å)	0.75	0.79	0.19
Shear	0.30	0.59	-0.93
Stretch	0.12	-0.34	0.12
Stagger	-0.06	-0.60	0.54
Shift	0.03	0.03	0.03

[a]Helical parameters around the modified phosphate, T_5 and A_6, were calculated using Curves+[143]. Ideal values of B-DNA from Hays et. al.[148] that vary from 0 are indicated in parentheses.

The local structures of the T_5 and the A_6 nucleotides at the point of junction cross over differed among the two ***ps***-J conformers. In particular, the base pair from T_5 to its complementary A_6 base is slightly stretched in ***ps***-JA compared to ***ps***-JB (Table 3.2). A more detailed analysis showed the sulfur of ***ps***-JA approaches the Br of $^{Br}U_7$ at a closer distance ($r_{Br\cdots(S/O)}$) and more linearly (θ_1) than that of ***ps***-JB (Figure 3.5c-d, Table 3.3). It is clear from the geometries that ***ps***-JA forms an X-bond and, as a result pulls T_5 base slightly away from the A_6 base on the complementary strand.

The geometry of the X-bond in ***ps***-JA, however, is longer (in terms of the $\%\Sigma r_{vdw}$) and less linear compared to the analogous O···Br interaction in the 2ORG structure. The less linear acceptor approach to the halogen, however, is consistent with results from previous work on sulfur acceptors in small molecule assemblies[131], where the approach angles are seen to fall between 160˚ and 170˚, as compared to those of oxygen acceptors[18], which fall between 170˚ and 180˚. Furthermore, the X-bond of sulfur acceptors in the small molecule systems have been shown to be more energetically favorable when compared to oxygen acceptors[136], suggesting that the angle of approach may not be as destabilizing as initially thought. Finally, a water molecule is seen in ***ps***-JA bridging the sulfur acceptor to the Br, which could confer additional stability to this interaction in the junction (Figure 3.5d).

Differential Scanning Calorimetry: Comparison of Energies

The effect of the Br···S X-bond on the thermal stability of the junction in solution was determined through DSC melting studies. We had shown that junction formation is concentration dependent, with the DNA remaining as a duplex at low concentrations and as a four-stranded junction at higher concentrations. At the concentration of DNA (300 μM) for the current study, both junction and duplex are present, allowing the melting parameters of the two forms to be

Table 3.3. Geometries of Br X-bond interactions for sulfur and oxygen acceptors.[a]

	***ps*-JA**	***ps*-JB**	2ORG
$r_{Br\cdots(S/O)}$	3.5 Å	4.1 Å	2.9 Å
$\%\Sigma r_{vdw}$	96%	112%	86%
θ_1	161.3°	150.1°	167.2°

[a]The distances ($r_{Br\cdots(S/O)}$) and the corresponding percent of the Σr_{vdw} ($\%\Sigma r_{vdw}$) from the Br of $^{Br}U_7$ to the sulfur of conformers A and B of the of the DNA junctions containing a phosphorothioate linkage (***ps*-JA** and ***ps*-JB**, respectively) are compared to those to the oxygen of a phosphate containing junction (from a previous crystal structure, 2ORG from the PDB). The angles of approach of the acceptor to the Br—C bonds (θ_1) are also compared.

determined simultaneously. The DSC melting and cooling profiles of the ***ps***-J construct showed fully reversible melting and reannealing of the DNA, and were best fit using two two-state van't Hoff models (Figure 3.3). Subtraction of the duplex energies from the junction leaves us with the energy of stabilization associated with the interactions at the junction crossover. Subsequent comparison of the crossover energies between ***ps***-J and ***p***-J allows us to determine the specific thermodynamic parameters (ΔH, ΔS, and ΔG) of a sulfur relative an oxygen as the X-bond acceptor.

The DSC melting data show that both the junction and duplex forms of the phosphothioate ***ps***-J construct are more thermally stable than the native ***p***-J, with melting temperatures (T_M) that are higher by ~0.7 ˚C for the duplex and by 1 ˚C for the junction (Table 3.4). Furthermore, melting enthalpies (ΔH_M) for ***ps***-J in duplex form is ~6 kcal/mol higher and in junction form is ~13 kcal/mol higher than the comparable oxygen containing DNA.

In order to directly compare the stabilizing potential of the sulfur X-bond acceptor to that of oxygen, the DSC melting parameters were translated to stabilizing enthalpies and entropies ($\Delta H^{25°C}$ and $\Delta S^{25°C}$, respectively) at a reference temperature of 25 °C, applying the heat capacity (ΔC_p) in the standard relationships (Eqs. 3.1 and 3.2). Since the core structures are nearly identical for the oxygen and sulfur constructs, the difference in $\Delta H^{25°C}$ and $\Delta S^{25°C}$, reflect the difference in specific X-bonding potential for the two types of acceptors. The resulting $\Delta\Delta H$ and $\Delta\Delta S$ values (Table 3.5) show that the sulfur-bromine X-bond is enthalpically more stabilizing by -3.5 kcal/mol. As had been previously seen[100], this more thermally stable interaction comes at an entropic cost, which is reflected in the significantly lower crystallographic B-factors of the nucleotides at the junction crossing-over—those that participate in the X-bond (Figure 3.6). The

Table 3.4. Differential Scanning Calorimetry (DSC) melting temperatures (T_M), enthalpies (ΔH_M), and entropies (ΔS_M) for the duplex and junction forms of the phosphate (***p***-J) and phosphorothioate (***ps***-J) DNA constructs.

DNA Construct	Duplex			Junction		
	T_M (°C)	ΔH_M (kcal/mol)	ΔS_M (cal/mol K)	T_M (°C)	ΔH_M (kcal/mol)	ΔS_M (cal/mol K)
p-J	51.9 ± 0.12	53.2 ± 0.5	163.5 ± 1.6	57.2 ± 0.08	70.7 ± 1.0	214 ± 3
ps-J	52.6 ± 0.10	58.4 ± 0.3	179.4 ± 0.9	58.2 ± 0.13	83.8 ± 0.6	253.1 ± 1.8

$$\Delta H^{25^\circ C} = \Delta H_M + \Delta C_p(T^{25^\circ C} - T_M) \qquad (Eq.\,3.1)$$

$$\Delta S^{25^\circ C} = \Delta S_M + \Delta C_p\, ln\left(\frac{T^{25^\circ C}}{T_M}\right) \qquad (Eq.\,3.2)$$

Table 3.5. Stabilization energies at 25° C ($\Delta H^{25^\circ C}$, $\Delta S^{25^\circ C}$, and $\Delta G^{25^\circ C}$) for the junction minus duplex forms of the ***p***-J and ***ps***-J DNA constructs.

DNA Construct	$\Delta H^{25^\circ C}{}_{(J\text{-}D)}$ (kcal/mol)	$\Delta S^{25^\circ C}{}_{(J\text{-}D)}$ (cal/mol K)	$\Delta G^{25^\circ C}{}_{(J\text{-}D)}$ (kcal/mol)
p-J	5.6 ± 0.5	15.6 ± 1.5	0.97 ± 0.06
ps-J	9.1 ± 0.3	25.7 ± 0.9	1.40 ± 0.04
	$\Delta\Delta H^{25^\circ C}$ (kcal/mol)	$\Delta\Delta S^{25^\circ C}$ (cal/mol K)	$\Delta\Delta G^{25^\circ C}$ (kcal/mol)
p-J - ***ps***-J	-3.5 ± 0.6	-10.2 ± 1.8	-0.43 ± 0.07

resulting $\Delta\Delta G$, however, shows that sulfur remains overall a slightly stronger X-bond acceptor in this DNA system.

Quantum Mechanics Calculations of X-bonding energies

In order to associate the geometries of the X-bonds seen in the crystal structures with the stabilizing potentials determined by DSC, we performed quantum mechanics calculations at the MP2 level on a set of model compounds (Figure 3.3) that mimic the O···Br interaction from the previous 2ORG and the S···Br interaction of the current ***ps***-J constructs (Table 3.6). The MP2 energies show that the X-bond of the ***ps***-JA conformer is about twice as stabilizing as that of ***ps***-JB. This is not surprising, given the shorter and more linear geometry of the ***ps***-JA conformer. Contrary to the DSC results, however, the S···Br X-bond of ***ps***-JA was predicted to be slightly less stabilizing than that of O···Br. However, we note that our MP2 calculations do not take into account the explicit solvent interactions, including the water seen in ***ps***-JA that bridges the sulfur to the Br atom.

■ 3.5 DISCUSSION

In the current work, we have shown that an anionic sulfur can serve as a favorable acceptor for bromine X-bonds. Although previous studies have characterized the structure-energy relationship of sulfur X-bonds in small molecule assemblies both experimentally and computationally, there was a dearth of knowledge about sulfur acceptors in a biological context. In our DNA model system, a bromine X-bond with a sulfur acceptor was shown to be enthalpically more stabilizing than a comparable oxygen acceptor, despite what appears to be less favorable geometries. Both the energetics and broader range of geometries are consistent with prior computational studies on sulfur-halogen interactions[131,136]. The stronger interaction,

Table 3.6. Quantum mechanical energies calculated at the MP2 level (E_{MP2}), applying the (aug-cc-PVTZ) with BSSE correction for a 5-bromouracil complex with dimethylphosphate (for 2ORG) or with dimethylphosphorothioate (for ***ps***-JA and ***ps***-JB conformers of the ***ps***-J construct).

	O···Br (2ORG)	S···Br (***ps***-JA)	S···Br (***ps***-JB)
E_{MP2} **(kcal/mol)**	-2.22	-1.75	-0.94

however, comes with an entropic cost, which renders sulfur overall only slightly more stabilizing as an X-bond acceptor.

Sulfurs have been observed to comprise only 5% of X-bond acceptors in protein-ligand complexes[130], which may reflect the overall lower occurrence of the element at accessible regions in protein structures. The sulfur containing amino acids cysteine and methionine, however, are both prevalent in active sites of enzymes[149]. Cysteines, for example, are key reactive nucleophiles for many hydrolases, deubiquitinases, caspases and enzymes involved in redox reactions[150]. Methionine is a particularly interesting residue for drug design due to it's flexibility and lipophilicity. Similar to cysteine, it is found at higher rates in active sites of many enzymes including nuclear hormone receptors, catechol-*O*-methyltransferase, and herpes simplex virus type 1 thymidine kinase[151]. It has also been found to be the most common "gatekeeper", a residue that blocks the hydrophobic pocket and plays a key role in the recognition elements at the ATP binding site in human kinases[135]. Methionines could, therefore, serve as a key acceptor for halogenated inhibitors. Our understanding that sulfur is capable of forming X-bonds that are comparable to or more stable to oxygen over a broader range of geometries would allow this underrepresented element to become an attractive target when designing halogenated compounds as therapeutic agents.

CHAPTER 4

STRUCTURE-ENERGY RELATIONSHIPS OF HALOGEN BONDS IN PROTEINS[3]

■ 4.1 SUMMARY

The structures and stabilities of proteins are defined by a series of weak noncovalent electrostatic, van der Waals, and hydrogen bond (HB) interactions. In this study, we have designed and engineered halogen bonds (XBs) site-specifically to study their structure−energy relationship in a model protein, T4 lysozyme. The evidence for XBs is the displacement of the aromatic side chain toward an oxygen acceptor, at distances that are equal to or less than the sums of their respective van der Waals radii, when the hydroxyl substituent of the wild-type tyrosine is replaced by a halogen. In addition, thermal melting studies show that the iodine XB rescues the stabilization energy from an otherwise destabilizing substitution (at an equivalent noninteracting site), indicating that the interaction is also present in solution. Quantum chemical calculations show that the XB complements an HB at this site and that solvent structure must also be considered in trying to design molecular interactions such as XBs into biological systems. A bromine substitution also shows displacement of the side chain, but the distances and geometries do not indicate formation of an XB. Thus, we have dissected the contributions from various noncovalent interactions of halogens introduced into proteins, to drive the application of XBs, particularly in biomolecular design.

[3] The work in this chapter was published in 2017 in *Biochemistry*. P.S.H, M.R.S., A. –C. C. C., and I planned experimental studies and wrote the manuscript. M.R.S, A. –C. C. C. and I performed purifications, DSC studies, and crystallographic studies. M.R.S and I performed the MP2 calculations. H.B. carried out crystallographic studies. R.A.M. provided the modified proteins.

Scholfield, M. R.*; Ford, M. C.*; Carlsson, A.-C.C.; Butta, H.; Mehl, R. A.; Ho, P. S. Structure-Energy Relationships of Halogen Bonds in Proteins. *Biochemistry.* 2017.
*These authors contributed equally to this work

■ 4.2 INTRODUCTION

The halogen bond (XB)[10], which is analogous to the hydrogen bond (HB), has become widely applied as a design element in supramolecular chemistry[152]. This recently rediscovered interaction in biology, however, is recognized only in hindsight, leaving it an inchoate tool for biomolecular design. In particular, XBs are seen to be critical for defining the specificity and affinity of halogenated inhibitors against protein targets[43], making them important tools in medicinal chemistry[38,125]. Although we had previously studied the structure−energy relationships of XBs in a model DNA system[14,100], it is critical to define these relationships in a protein system. We present here the first detailed characterization of XB geometry−energy relationships in a protein, with XBs specifically designed to affect the stability of the model protein T4 lysozyme.

Although the physicochemical basis for a halogen as a covalent substituent, interacting with an electron-rich oxygen, nitrogen, or sulfur, is still being debated, a readily accessible explanation for the XB is the σ-hole model[13] (Figure 4.1a). When a halogen is covalently bonded to a carbon atom, one of its valence electron is pulled into the σ-molecular orbital, resulting in a "σ-hole", as an electropositive crown that serves as the donor in the XB and accounts for the strong directionality of the interaction (Figure 4.1b). Because the atomic orbitals perpendicular to the σ-hole retain their full complement of electrons, the halogen is amphoteric[38,41], with the halogen serving as electropositive XB donor along the σ-hole and as an electronegative HB acceptor in the perpendicular direction[153,154].

The typical XB donor in protein systems is a halogen substituent of an enzyme inhibitor, while the acceptor is most often the peptide carbonyl oxygen of the protein' s back-bone[43,51,86,130,155]. The strength of the XB depends on the size of the σ-hole, which in turn follows

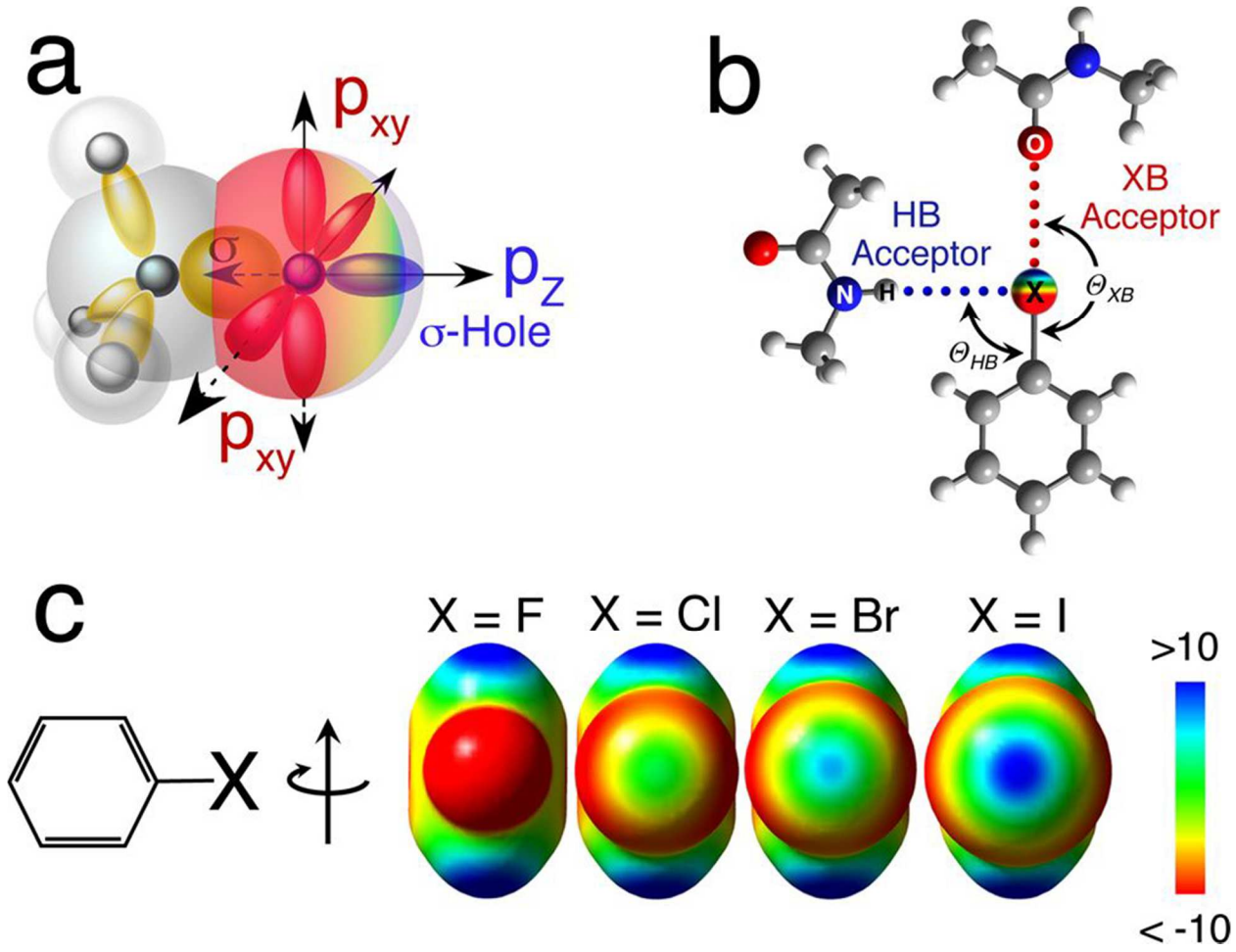


Figure 4.1. Halogen bond (XB). (a) *σ*-Hole model for X bonding[13]. The electropositive *σ*-hole that serves as the XB donor is created as a result of pairing the valence electron from a halogen (typically assigned to the p_z atomic orbital) with, in this example, an electron from a carbon to form a covalent C−X bond. The resulting *σ*-molecular orbital depopulates the outside lobe of the p_z orbital, leaving an electropositive crown (blue surface) and flattening of the atomic radius, the *σ*-hole. The $p_{x,y}$ orbitals remain fully occupied, resulting in an electronegative (red surface) ring perpendicular to the *σ*-molecular orbital. (b) Relationship between XBs and hydrogen bonds (HBs). The XB (red dotted line) is defined by contact distances that are less than the sum of the standard van der Waals radii for the halogen donor and the acceptor, and an approximate linear approach of the acceptor to the halogen ($\Theta_{XB} \approx 180°$). The $p_{x,y}$ orbitals of the halogen can also serve as an acceptor to an HB (blue dotted line), an approach that is approximately orthogonal to the XB ($\Theta_{HB} \approx 90°$). (c) Electrostatic potential of halobenzenes. The DFT-calculated electrostatic potentials at the 6-311+G (d,p) level [from >10 kcal/mol (blue) to less than −10 kcal/mol (red)] show the size and intensity of the *σ*-hole increase as the size of the halogen increases from F to Cl to Br to I.

the polarizability of the halogen [F ≪ Cl < Br < I (Figure 4.1c)]. The XB can be further tuned according to the electron withdrawing ability of the compound to which the halogen is covalently bonded[38,156], with aromatic groups accentuating XBs[78]. XB energies to anionic oxygen acceptors measured in a model DNA system are equal to or greater than that of a competing HB[15,18,139]. Furthermore, their geometries become more ideal (with shorter distances and more linear alignment to the X–C bond, Θ_{XB}) as their strength increases[18]. The structure−energy relationships of XBs to a neutral carbonyl oxygen acceptor in a protein[43,51,86,155], however, have not been previously determined.

To study the impact of XBs on protein stability, we started with T4 lysozyme, a classic model system for studying the effects of molecular interactions on protein structure and stability in crystals and in solution[157,158]. An iodinated phenylalanine had previously been incorporated into lysozyme to demonstrate the utility of noncanonical amino acids (ncAAs) to solve the phase problem in protein crystallography; however, this iodine was not positioned to form an XB[159]. For the study presented here, we have replaced tyrosines (Y) with a chemically modified phenylalanine (^{Z}F) at two different sites: the first position (Y18) has the potential to interact with the protein backbone through HBs and/or XBs, while the second (Y88) serves as a control for the effects of each substituent at a noninteracting site (Figure 4.2). Comparing the structure and energy effects separately at the two positions reveals how the engineered XB interactions affect the structure and stability of the protein.

■ 4.3 MATERIALS AND METHODS

Site directed mutagenesis and protein expression

All T4 lysozyme constructs started with the gene of the pseudo-wild-type (WT*) protein[160], with the DNA sequence encoding a six-His tag appended at the C-terminus to facilitate protein

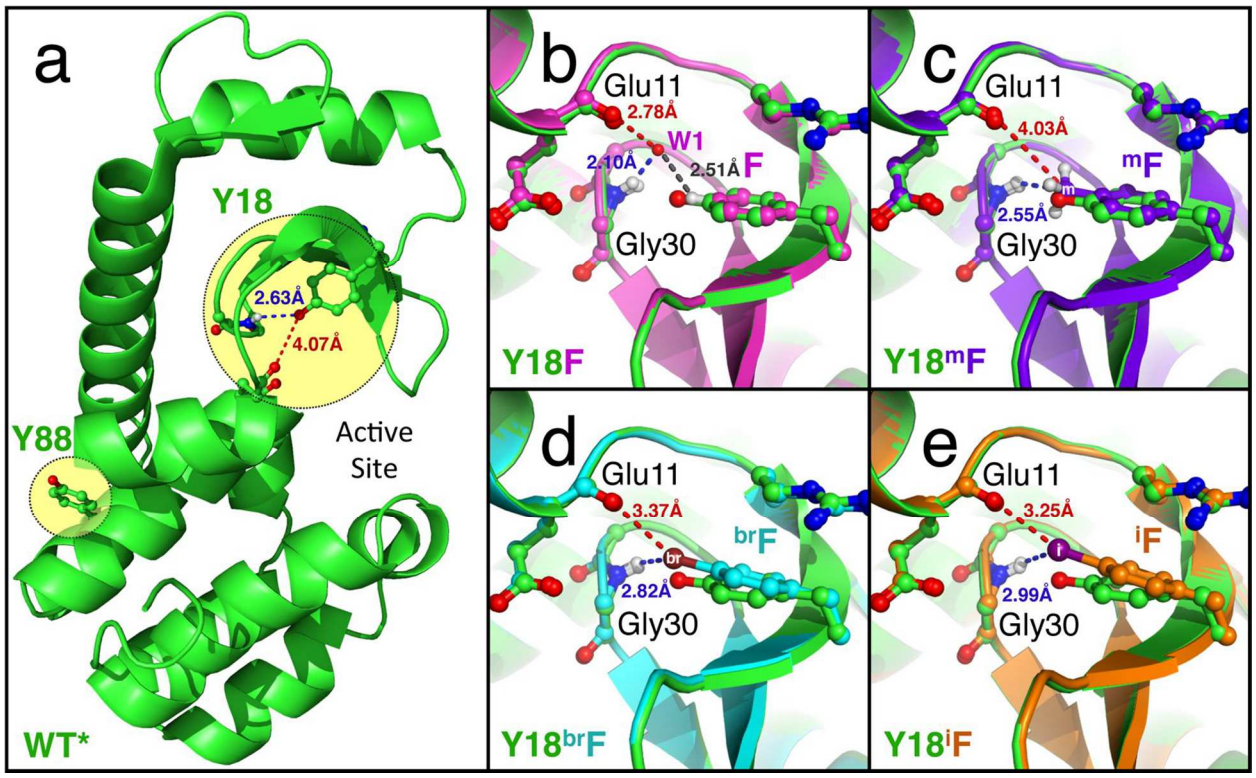


Figure 4.2. Structures of halogenated and nonhalogenated T4 lysozyme constructs. (a) Cartoon representation of the T4 lysozyme structure. The aromatic amino acid residues selected for modifications in this study (tyrosine 18 and 88) are circled and labeled (Y18 and Y88, respectively). The side chain at Y18 makes specific intramolecular interactions with the carboxyl oxygen of Glu11 (red dashed line) and the amine of Gly30 (blue dashed line). (b−e) Crystal structures of halogenated and nonhalogenated T4 lysozyme constructs at Y18. In each panel, the structures of the modified ^{Z}F residues at position 18 are superimposed on the structure of WT* (green carbons). The modified constructs are colored (b) magenta for Y18F, (c) purple for Y18mF, (d) cyan for Y18brF, and (e) orange for Y18iF. The blue dashes indicate short distances (*RN*−*H*···*Z*) of each modified substituent (Z) to the HB donor of Gly30, while the red dashes indicate short distances (*RZ*···*O*) to the potential HB or XB acceptor oxygen of Glu11.

purification. The modified DNA sequences were inserted into the pBAD vector for expression in DH5*α Escherichia coli*.

Expression vectors for WT* containing canonical amino acids were transformed into BL21 (DE3) pLysS *E. coli*. Transformed cells were grown in 2xYT medium with the appropriate antibiotic (ampicillin and chloramphenicol) and incubated at 37 °C while being shaken at 250 rpm until an OD_{600} of 0.4−0.6 was reached. The cells were induced with arabinose added directly to the cultures to a final concentration of 0.2% (w/v) and allowed to grow for an additional 3 h. Subsequently, the cells were harvested by centrifugation at 2.2K RCF; the supernatant was decanted, and the bacterial pellets were stored at − 80 ° C.

WT* constructs that incorporate ncAAs were expressed in *E. coli* cells, as previously described[161]. Briefly, for the constructs with halogenated or methylated Phe, the codons for Y18 or Y88 were replaced with an AMBER (TAG) codon. The expression vectors were cotransformed with pBAD (gene) and pDule2-pCNF (containing the orthogonal aminoacyl-tRNA synthetase and tRNA pair) into DH10*β E. coli*. The cells were grown overnight in 5 mL of noninducing medium with the appropriate antibiotics. These cultures were used to inoculate 1 L of autoinduction medium without arabinose, and with the appropriate antibiotics at 37 ° C while being shaken at 250 rpm. At an OD_{600} of 0.8, ncAAs were added to the cultures to a final concentration of 1 mM. Once cells reached an OD_{600} of 3.0, arabinose was added to a final concentration of 0.1% (w/v) and cell growth continued for an additional 3 h. The cells were harvested by centrifugation at 2.2K RCF; the supernatant was decanted, and bacterial pellets were stored at − 80 ° C.

Protein purification

Bacterial pellets were suspended in 35 mL of binding buffer [40 mM potassium phosphate (pH 7.4), 500 mM sodium chloride, 50 mM imidazole, and 0.02% (w/v) sodium azide] and thawed in a 37 °C water bath for 5 min, and all subsequent steps were performed at 4 °C. Thawed pellets were sonicated for 1.5 min and then centrifuged for 30 min at 35.3K RCF. The supernatant was decanted and directly loaded onto a HisTrap HP column on an ÄKTA FPLC system, after which His-tagged protein was eluted with an increasing imidazole gradient with elution buffer [40 mM potassium phosphate (pH 7.4), 500 mM sodium chloride, 500 mM imidazole, and 0.02% (w/v) sodium azide]. Selected fractions were concentrated in an Amicon Ultra centrifugal concentrator (NMWL, 10 kDa) to 1 mL and then loaded onto a gravity-fed Sephadex G-50 fine column equilibrated in buffer specific for crystallization or differential scanning calorimetry (DSC) experiments [crystallization buffer consisting of 500 mM sodium chloride (pH 7.4), 50 mM sodium phosphate, and 0.02% (w/v) sodium azide; DSC buffer 23 consisting of 20 mM glycine-HCl (pH 3.5), 80 mM NaCl, and 1 mM EDTA]. Selected fractions were combined and used for crystallization or DSC experiments.

Protein crystallization

Protein Crystallization. Combined fractions in crystallization buffer were concentrated to 18−20 mg/mL. T4 lysozyme crystals were grown at 18 °C using the hanging drop vapor diffusion method with a 1:1 ratio of protein (18−20 mg/mL) to precipitant solution [2.0−2.4 M potassium phosphate (pH 6.6−7.5), 50 mM 2-hydroxyethyldisulfide, and 50 mM 2-mercaptoethanol], as previously described[138,162]. Diffraction quality crystals grew in 1−7 days. Crystals were harvested using a cryo-loop, flash-frozen, and stored in liquid nitrogen until X-ray data were collected.

X-ray Data Collection and Structure Determination

X-ray diffraction data were collected on crystals held in a cryogenic nitrogen stream (100 K), on the home source (Rigaku copper anode X-ray generator, 1.54 Å, Dectris Pilatus 200K detector) or Advanced Light Source (ALS) beamline 4.2.2 at Berkeley National Laboratory (1.00 Å, Research Detectors Inc. complementary metal-oxide-semiconductor 8M detector). Diffraction data from the home source were reduced using Denzo/Scalepack[141], or for data from the ALS beamline using d*TREK and CCP4 suite[163]. X-ray data were phased by molecular replacement, applying the atomic coordinates of WT* [Protein Data Bank (PDB) entry 1L63][164] as the starting model, yielding initial models with R_{work} values that ranged from 31.8 to 35.5% and R_{free} values that ranged from 31.0 to 36.1%. The PHENIX suite of crystallographic software[142] was used for subsequent refinement, which resulted in final structures with R_{work} values that ranged from 16.0 to 20.3% and R_{free} values that ranged from 19.6 to 24.7% (Tables 4.1−4.3).

Differential Scanning Calorimetry

Combined fractions of T4 lysozyme, after gel filtration purification, were diluted to a concentration of 0.1 mg/mL using DSC buffer and stored at −80 °C. A low pH was used to help promote reversible folding[165]. Melting curves were collected on a TA Instruments Nano DSC instrument under a constant pressure (3.0 atm) with all samples matched against identical buffers in the reference cell. Samples were equilibrated for 600 s, followed by melting data collected through heating cycles from 10 to 90 °C (at scan rates of 1 °C/min). Reversibility was confirmed for all constructs by performing a cooling scan from 90 to 10 °C (at scan rates of 0.5 °C/min) and a subsequent heating cycle. A minimum of five replicate experiments was conducted for each mutant. Melting data were analyzed using NanoAnalyze Data Analysis version 3.5.0 from TA Instruments to extract the melting temperatures (T_m) and enthalpies (ΔH°_{Tm}).

Table 4.1. Crystallographic parameters for non-halogenated T4 lysozyme constructs. All constructs crystallized in the space group $P3_221$ (angles for the unit cells: $\alpha = \beta = 90.0°$; $\gamma = 120.0°$). All data were collected on the Rigaku home source, as described in the Materials and Methods section.

Parameter	WT*	Y18F	Y18^{m}F
		Crystal	
Unit Cell Lengths	$a = b = 60.248$ Å; $c = 96.454$ Å	$a = b = 59.763$ Å; $c = 95.119$ Å	$a = b = 60.197$ Å; $c = 95.969$ Å
		Data Collection	
Resolution (Å)[a]	21.98 - 1.49 (1.54 - 1.49)	29.88 - 1.46 (1.51 - 1.46)	18.76 - 1.50 (1.55 - 1.50)
#Total Reflections	298,362	283,211	310,833
#Unique Reflections[a]	31,370 (3,322)	34,257 (3,097)	31,673 (2,670)
Multiplicity	9.5	8.3	9.8
Completeness[a]	92% (99%)	98% (85%)	96% (67%)
Mean I/σ(I)[a]	28.4 (2.9)	31.0 (2.1)	23.1 (1.8)
R_{merge}[a]	0.085 (0.592)	0.065 (0.594)	0.115 (0.757)
R_{meas}[a]	0.089 (0.623)	0.069 (0.727)	0.121 (0.825)
		Structure Refinement	
		Molecular Replacement: Initial Model Statistics	
R_{work}	0.3418	0.3536	0.3384
R_{free}	0.3479	0.3438	0.3312
Final Model			
PDB Code	5KHZ	5KI1	5KI2
A.A. Residues	162	162	162
Non-Solvent atoms	1,374	1,313	1,383
Solvent atoms	383	325	367
R_{work}[a]	0.1755 (0.2164)	0.1800 (0.2919)	0.1879 (0.2760)
R_{free}[a]	0.1962 (0.2537)	0.2067 (0.343)	0.2216 (0.3188)

[a]Values in parentheses are for the highest resolution shell.

Table 4.2. Crystallographic parameters for halogenated T4 lysozyme construct and one control construct. All constructs crystallized in the space group $P3_221$ (angles for the unit cells: $\alpha = \beta = 90.0°$; $\gamma = 120.0°$).

Parameter	Y18^{br}F	Y18^{i}F	Y88F
		Crystal	
Unit Cell Lengths	$a = b = 59.589$ Å; $c = 95.150$ Å	$a = b = 60.103$ Å; $c = 96.080$ Å	$a = b = 60.368$Å; $c = 96.571$Å
		Data Collection	
Data collect at:	ALS	Rigaku home source	Rigaku home source
Resolution (Å) [a]	34.98 - 1.65 (1.71 - 1.65)	32.03 - 1.63 (1.69 - 1.63)	22.99 – 1.493 (1.546 – 1.493)
#Total Reflections	236,904	218,989	269,948
#Unique Reflections[a]	23,668 (2,084)	25,701 (2,478)	33,480 (3,216)
Multiplicity	5.3	8.5	8.1
Completeness[a]	98% (92%)	99% (100%)	99% (94%)
Mean I/σ(I) [a]	4.6 (1.0)	31.0 (2.5)	42.2 (3.7)
R_{merge}[a]	0.210 (0.759)	0.098 (0.792)	0.061 (0.339)
R_{meas}[a]	0.233 (0.913)	0.103 (0.865)	0.064 (0.470)
		Structure Refinement	
		Molecular Replacement: Initial Model Statistics	
R_{work}	0.3344	0.3396	0.3374
R_{free}	0.3379	0.3494	0.3272
Final Model			
PDB Code	5KI3	5KIO	5KIG
A.A. Residues	162	162	162
Non-Solvent atoms	1,365	1,367	1,408
Solvent atoms	248	264	343
R_{work}[a]	0.2028 (0.3547)	0.1858 (0.2232)	0.1628 (0.2336)
R_{free}[a]	0.2473 (0.3744)	0.2165 (0.2959)	0.1958 (0.2857)

[a]Values in parentheses are for the highest resolution shell.

Table 4.3. Crystallographic parameters for control T4 lysozyme constructs (Res. 88). All constructs crystallized in the space group $P3_221$ (angles for the unit cells: $\alpha = \beta = 90.0°$; $\gamma = 120.0°$).

Parameter	Y88^{m}F	Y88^{br}F	Y88^{i}F
		Crystal	
Unit Cell Lengths	$a = b = 60.183$ Å; $c = 95.834$ Å	$a = b = 60.013$ Å; $c = 95.945$ Å	$a = b = 60.283$ Å; $c = 96.197$ Å
		Data Collection	
Data collect at:	Rigaku home source	ALS	Rigaku home source
Resolution (Å)[a]	45.79 – 1.56 (1.616 – 1.560)	45.70 - 1.55 (1.61 - 1.55)	28.76 – 1.499 (1.553 – 1.499)
#Total Reflections	361,463	163,055	431,722
#Unique Reflections[1]	29,011 (2836)	28,167 (2,068)	33,083 (3,250)
Multiplicity	12.4	1.9	13.0
Completeness[a]	99% (98%)	95% (71%)	100% (99%)
Mean I/σ(I)[a]	43.6 (2.4)	19.1 (6.3)	70.4 (6.5)
R_{merge}[a]	0.065 (0.783)	0.026 (0.310)	0.058 (0.341)
R_{meas}[a]	0.068 (0.870)	0.037 (0.438)	0.060 (0.376)
		Structure Refinement	
		Molecular Replacement: Initial Model Statistics	
R_{work}	0.3430	0.3178	0.3546
R_{free}	0.3290	0.3095	0.3613
Final Model			
PDB Code	5KII	5KI8	5KIM
A.A. Residues	162	162	162
Non-Solvent atoms	1,393	1,366	1,436
Solvent atoms	342	205	368
R_{work}[a]	0.1999 (0.2893)	0.1757 (0.2757)	0.1604 (0.2117)
R_{free}[a]	0.2244 (0.3198)	0.2101 (0.3147)	0.1844 (0.2326)

[a]Values in parentheses are for the highest resolution shell.

Quantum Mechanical (QM) Calculations

The atomic coordinates for the interacting residues (11, 18, and 30) were taken directly from the refined structures of each construct. Residues 11 and 30 were reduced to N-methylacetamides, and residue 18 was reduced to **Z**-benzene to decrease the computational time (Figure 4.3). QM energies were calculated using Gaussian 09 Rev.E.01[145], with the Møller−Plesset second-order (MP2) calculations in cyclohexane as the solvent (D = 2, relative to a vacuum). Geometry optimization of the hydrogen atoms was performed with the Hartree−Fock method prior to the energy calculations. Polarizable basis sets including dispersion were applied to the calculations (aug-cc-PVTZ for WT*, Y18F, Y18^{m}F, and Y18^{br}F and aug-cc-PVTZ-PP[166] from EMSL Basis Set Exchange for Y18^{i}F). Basis set superposition errors (BSSEs)[147,167] were determined from a separate counterpoise gas phase calculation and directly summed into the calculated solvent phase energy.

Turbidity Assay

The activity of T4 lysozyme was monitored via a standard cell clearing assay[168,169]. *Micrococcus lysodeikitcus* bacteria was grown in 2xYT medium overnight and then diluted in 50 mM sodium phosphate buffer until an OD_{450} of 1.0 was reached. Purified T4 lysozyme was added to the solution at room temperature to reach a final concentration of 0.1 mg/mL, and the absorbance change over time was measured.

■ 4.4 RESULTS AND DISCUSSION

The constructs for engineering XBs into a model protein are based on the modified WT* form of T4 lysozyme, which has its two disulfide-forming Cys residues replaced and, thus, follows a classical two-state reversible folding/unfolding pathway[158]. To identify positions at which an optimal XB could be introduced into this system, we started by identifying all aromatic

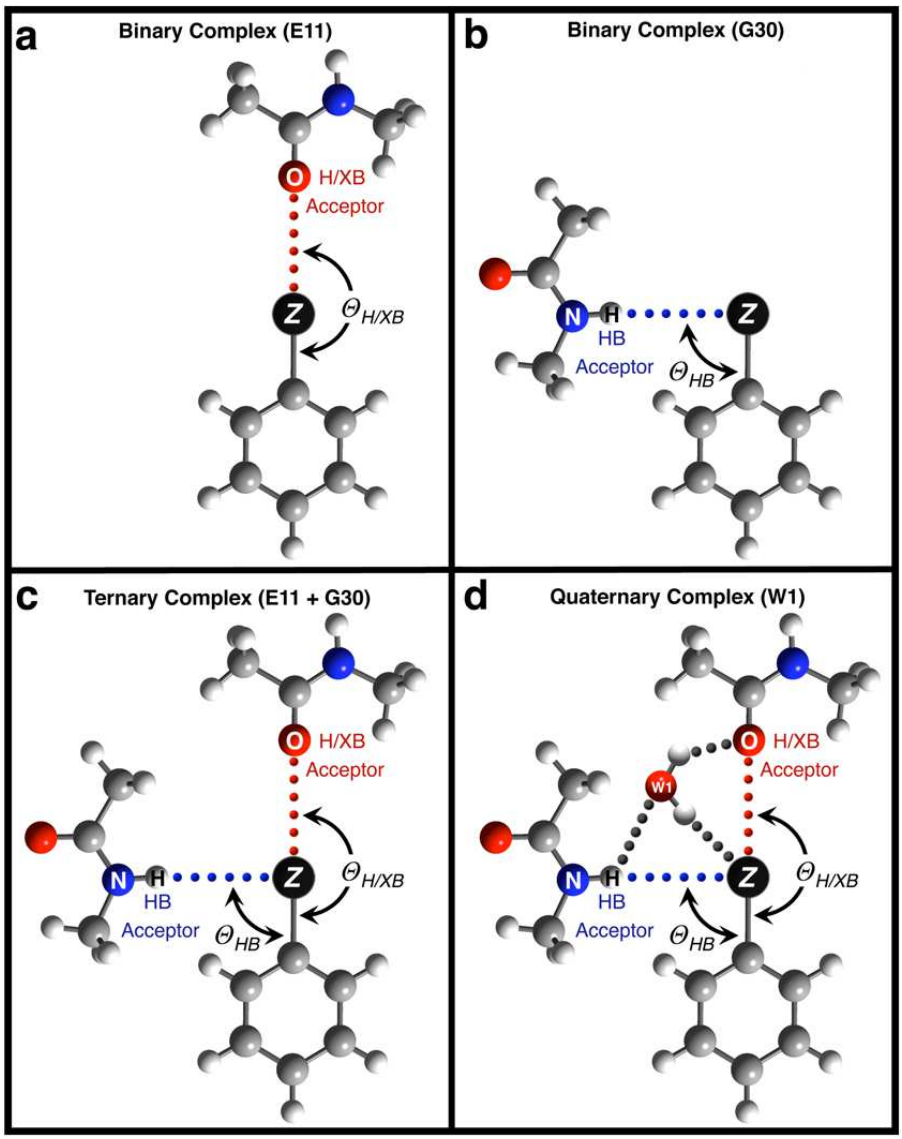


Figure 4.3. Schematic of the molecular models used in the quantum mechanical calculations of HB/XB interactions at the Y18 site of T4 lysozyme constructs. The modified side chain at the Y18 position was modeled as a benzene molecule with a single substituent (Z = H, OH, CH_3, Br, or I), while *N*-methylacetamides were used to model the HB donor from Gly30 and HB/XB acceptor of Glu11 (atomic coordinates were taken directly from the refined crystal structures of the interacting residues (11, 18, 30) of each Y18 construct). QM energies were calculated using Gaussian 09e[170], with the Møller−Plesset second-order (MP2) calculations performed in cyclohexane. The positions of the hydrogen atoms were optimized prior to the energy calculations as their positions are dependent on Z. Interaction energies were calculated for the binary complexes of two residues (a) ([Glu11(O) + Y18(C–Z)] and (b) [Gly30(N–H) + Y18(C–Z)]); for the ternary complex of three residues (c) ([Glu11(O) + Gly30(N–H) + Y18(C–Z)]); or for the quaternary complex with water W1 (d) ([Glu11(O) + Gly30(N–H) + Y18(C–Z) + (W1)]). The dashed lines in each panel indicate the interactions for which energies were calculated.

amino acids with side chains that are in the proximity of a peptide carbonyl oxygen. Tyrosine 18 (Y18) was selected as the site to engineer XB interactions; the hydroxyl group of Y18 serves as an HB acceptor to a backbone amino group at Gly30 and is close but does not form a direct HB to the peptide oxygen of Glu11 (Figure 4.2a). Constructs of Y18 were engineered as Y18^{Z}F, where ^{Z}F is a phenylalanine (F) residue with a Z substituent [hydrogen (F), bromine (^{br}F), iodine (^{i}F), or a methyl group (^{m}F)]. We expected both of the halogenated ^{Z}F analogues to form stabilizing XBs with the Glu11 oxygen. The amphoteric nature of halogens predicts that the Y18^{br}F and Y18^{i}F constructs would also maintain the orthogonal HBs to Gly30.[171] The Y18F mutant was designed as a control with no HB or XB capabilities, while methylated Y18^{m}F was designed to mimic the size and hydrophobic properties of the halogens.

Constructs at Y88 (Y88^{Z}F) serve as controls to determine the nonspecific effects of each substituent on the structure and stability of the enzyme. Y88 is similar to Y18 in that the Tyr side chain is mostly buried within the protein, with very little exposure of the hydroxyl group at the solvent accessible surface. The hydroxyl group of Y88, however, is pointed toward the surface, and thus, the potential interacting (Z) substituent of Y88^{Z}F constructs would be unable to form direct intramolecular interactions within the protein (Figure 4.2a). The effects of the engineered XB on the protein structure are characterized by comparing the single-crystal structures of the Y18^{Z}F to Y88^{Z}F constructs, while the effects on stability were determined by comparing the thermodynamics of melting by differential scanning calorimetry (DSC).

Single-Crystal Structures

The crystals from all of the Y18^{Z}F to Y88^{Z}F constructs were isomorphous and diffracted to sufficiently high resolution to provide highly accurate geometries (Tables 4.1-4.3 and Figure 4.4). The single-crystal structures of Y18F and Y18 mF place the phenyl rings in nearly identical

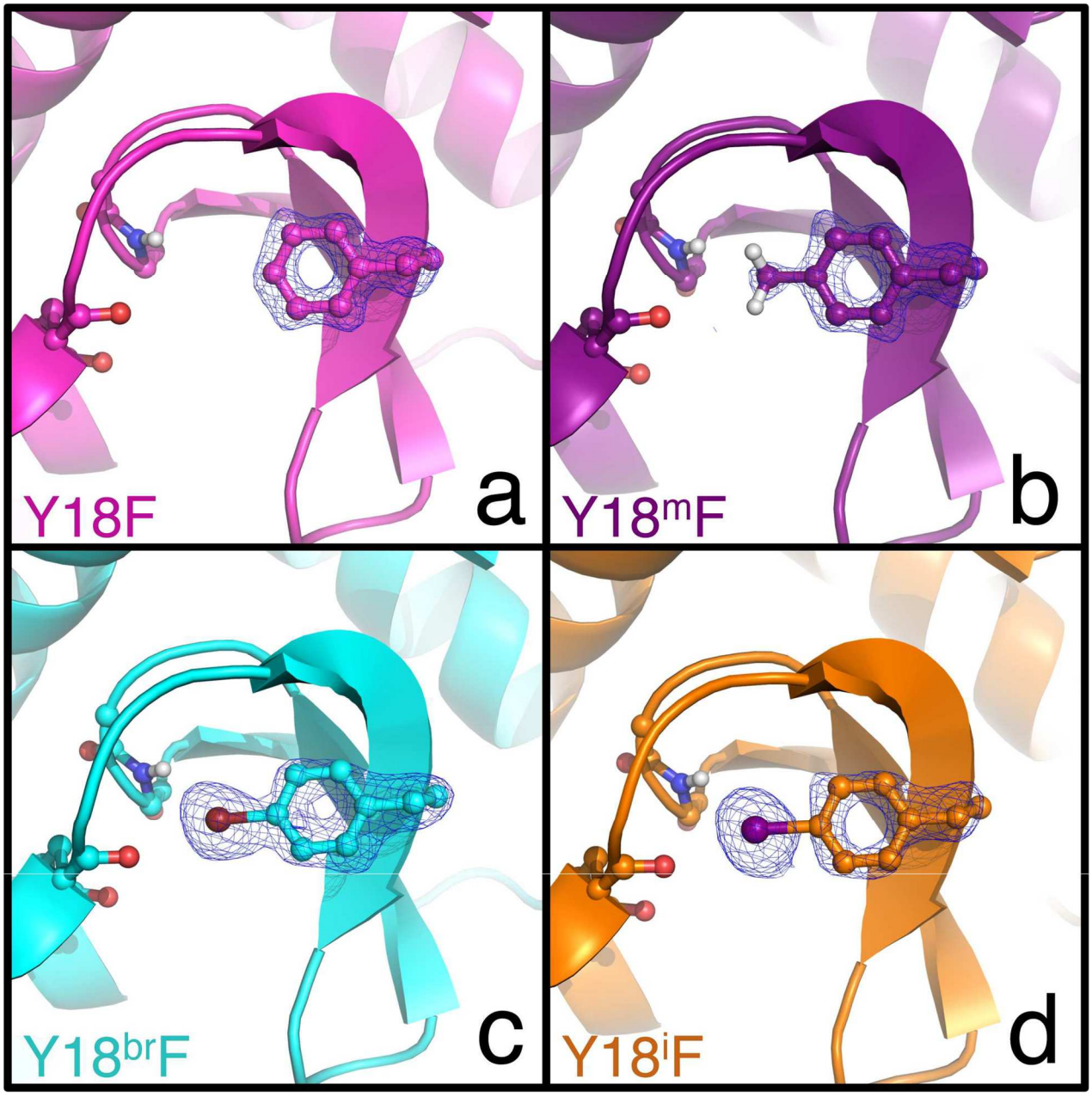


Figure 4.4. Omit electron density maps from crystal structures of halogenated Y18^{X}F T4 lysozyme constructs. Fo-Fc electron densities of structures refined with Y18 modeled as a Gly residue, after simulated annealing, for the Y18F (a), methylated Y18^{m}F (b), brominated Y18^{br}F (c), and the iodinated Y18^{i}F constructs (d) are rendered at the 2.5σ level of contours.

geometries relative to that of WT* (Figure 4.2b,c), suggesting that the local loop structure is not perturbed by interactions to Gly30 or a lack of interaction to Glu11. In contrast, the orientations of the aromatic rings in the Y18^{br}F and Y18^{i}F constructs are significantly perturbed, showing the halogens pulled toward the Glu11 oxygen (Figure 4.2d,e). The aromatic rings in all the control Y88^{Z}F constructs superimpose nearly exactly on the WT* structure (Figure 4.5), indicating that the perturbations in Y18^{br}F and Y18^{i}F are directly associated with the interactions at that site. In Y18^{br}F, the bromine is within van der Waals distance of the Glu11 oxygen [∼100% of the sum of the van der Waals radii, ΣR_{vdW} (Table 4.4)], but at an angle of approach from the oxygen of Glu11 (Θ_{E11} = 142.7°) that is nonlinear relative to the C−Br bond. The geometries suggest that the bromine is not making an XB, or at least not a very strong one. The distortion is significantly greater, however, with Y18^{i}F, where the I···O distance is within the optimal distance of ∼93% of ΣR_{vdW} for XBs in proteins.[38]This cannot be a steric effect, because Y18^{m}F remains unperturbed. The angle of approach of the Glu11 oxygen toward the I− C bond (Θ_{E11} = 150.0°) places it within the electropositive σ-hole for the iodine.[38] Thus, the geometries indicate that an XB is formed in the Y18^{i}F construct, with the iodine being more ideal and thus leading to a potentially stronger interaction than with the Y18^{br}F construct.

In addition to the interactions with Glu11, the Y18^{br}F and Y18^{i}F constructs maintain the HB to Gly30 seen in WT*, but the angles of approach of the Gly30 N− H group to each halogen (Θ_{G30} = 130.3° and 125.9° for Br and I, respectively) are significantly far from the optimum,6 suggesting HBs weaker than those in WT* .

Quantum mechanical (MP2) analyses of simple models for the molecular interactions with Glu11 and Gly30 (Table 4.5 and Figure 4.3) predict Y18^{i}F to be the most stable ternary construct [$E_{MP2\ (ternary)}$]. The XB from the iodine to the Glu11 oxygen was the strongest interaction of any

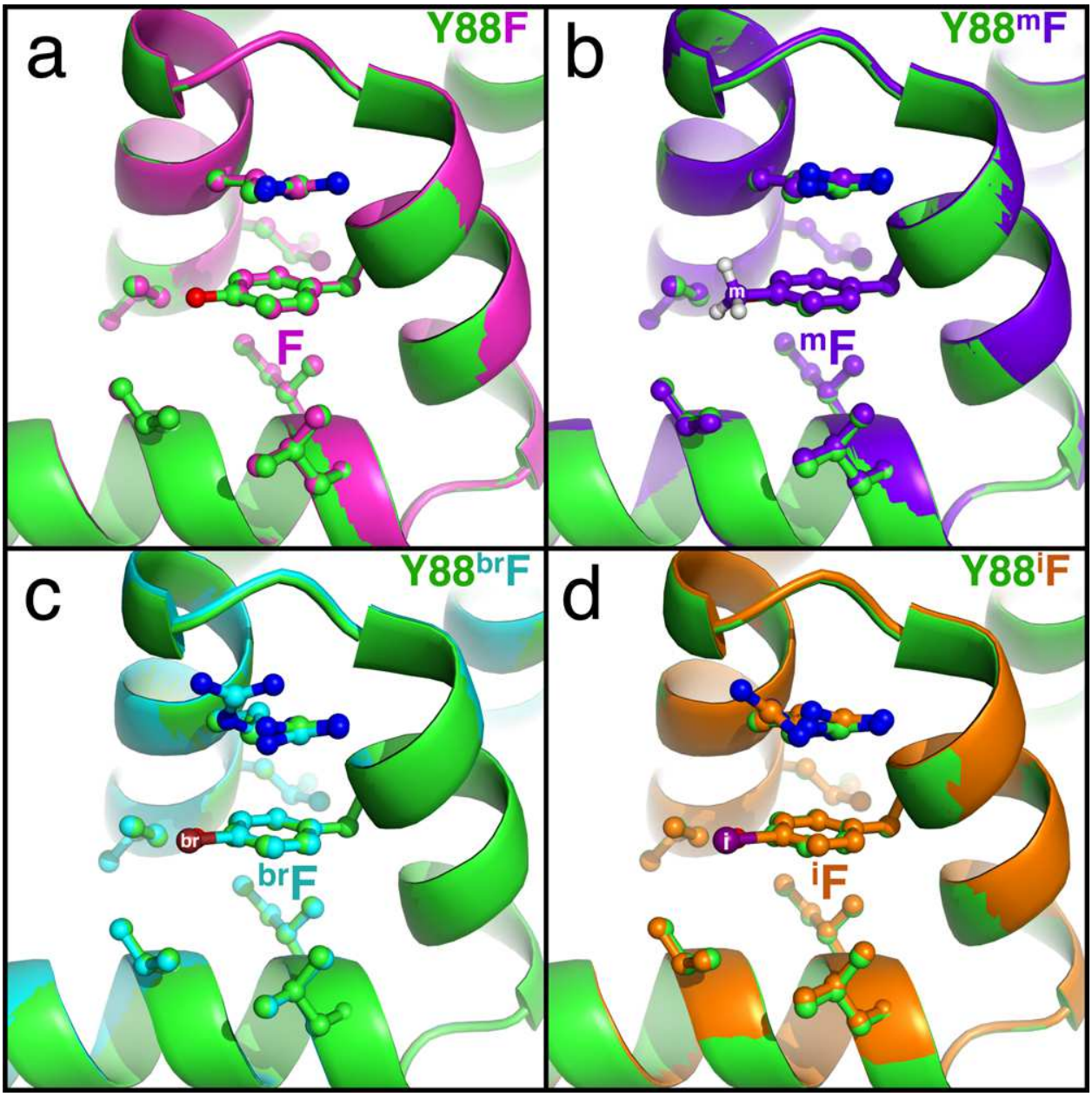


Figure 4.5. Details of the crystal structures of halogenated and non-halogenated T4 lysozyme constructs at Y88 position. In each panel, the structures of the modified residues are superimposed on the structure of the WT* enzyme (carbons and backbone trace in green). The carbons and backbone traces of the modified constructs are colored as (a) Y88F magenta, (b) Y88^{m}F purple, (c) Y88^{br}F cyan, (d) Y88^{i}F orange.

Table 4.4. Interaction geometries for T4 lysozyme constructs.

		Glu11$_{(O)}$			*Gly30*$_{(N)}$		
Construct	**Substituent**	$R_{O\cdots Z}$ **(Å)**	ΣR_{vdW} **(%)**	Θ_{E11} **(°)**	$R_{N-H\cdots Z}$ **(Å)**	ΣR_{vdW} **(%)**	Θ_{G30} **(°)**
WT*	OH	4.07	133.0	126.4	2.63	96.7	147.1
Y18F	H	5.38	175.2	142.1	3.21	108.1	155.0
Y18^{m}F	CH_3	4.03	148.2	133.4	2.55	106.3	156.0
Y18^{br}F	Br	3.37	100.0	142.7	2.82	92.5	130.3
Y18^{i}F	I	3.25	92.9	150.0	2.99	94.0	125.9

[a]The distances from the modified substituent (Z) to the potential HB or XB acceptor oxygen atom of Glu11 (*R*O···Z) or the HB donor of Gly30 (*R*N−H···Z) identify HB or XB interactions, in terms of the percent of the sum of the standard van der Waals distances of the interacting atoms (%ΣRvdW). The approach of Z to the potential HB or XB acceptor oxygen atom of Glu11 (ΘE11) or to the HB donor of Gly30 (ΘG30) determines whether the interaction is linear.

Table 4.5. Experimental and calculated interaction energies for T4 lysozyme constructs.[a]

Construct	**Substituent (*Z*)**	$\Delta H°_M$ **(kcal·mol^{-1})**	$\Delta\Delta H°_{M\ (18\text{-}88)}$ **(kcal·mol^{-1})**	T_M **(°C)**	$\Delta T_{M\ (18\text{-}88)}$ **(°C)**	$\Delta E_{MP2\ (ternary)}$ **(kcal·mol^{-1})**	$\Delta E_{MP2\ (quaternary)}$ **(kcal·mol^{-1})**
				Nonhalogenated T4 Lysozyme Constructs			
WT*$_{18}$	OH	130 ± 1	–	57.56 ± 0.02	–	-3.4	-15.0
WT*$_{88}$	OH	130 ± 1	–	57.56 ± 0.02	–	–	–
Y18F	H	122.6 ± 0.4	7 ± 1	56.65 ± 0.04	-0.12 ± 0.05	-0.3	-13.5
Y88F	H	115 ± 1	–	56.77 ± 0.04	–	–	–
Y18^{m}F	CH_3	118.7 ± 0.5	2 ± 1	55.32 ± 0.03	-1.12 ± 0.05	-1.4	-9.8
Y88^{m}F	CH_3	116.4 ± 0.7	–	56.53 ± 0.04	–	–	–
				Halogenated T4 Lysozyme Constructs			
Y18^{br}F	Br	115 ± 1	1 ± 1	55.21 ± 0.02	0.27 ± 0.07	-2.4	–
Y88^{br}F	Br	114 ± 1	–	54.94 ± 0.06	–	–	–
Y18^{i}F	I	119 ± 1	6 ± 1	56.21 ± 0.07	0.79 ± 0.08	-3.5	-11.3
Y88^{i}F	I	113 ± 1	–	55.43 ± 0.03	–	–	–

[a]The experimental melting enthalpies ($\Delta H°_M$) and melting temperatures (T_M) are measured by DSC for each Y18 and Y88 construct. The differences in $\Delta H°_M$ ($\Delta\Delta H°_{M\ (18\text{-}88)}$) and T_M ($\Delta T_{M\ (18\text{-}88)}$) between the Y18 and Y88 constructs reflect the energies associated with specific interactions at the Y18 site relative to a non-interacting substitution at Y88. Errors are the standard deviations of mean for each measurement. The H/XB interaction energies for the ternary complex of the interacting residues of ^{Z}F18 to Glu11 and Gly30 ($\Delta E_{MP2\ (E11\ \&\ G30)}$) were calculated by the Møller-Plesset 2 (MP2) method, applying the aug-cc-PVTZ basis set, with atomic coordinates for individual residues taken from the single crystal structures, in the absence of waters. The quaternary complex MP2 interaction energies of W1 (atomic coordinates taken from the single crystal structures seen in Figure 4.6) with residues Y18^{Z}F, Glu30, and Gly11 ($\Delta E_{MP2\ (W1)}$) shows how the position of this water contributes additionally to the stabilization of each Y18 construct ($\Delta E_{MP2\ W1}$ was not calculated for Y18^{br}F construct, since W1 was not observed in this structure). See Figure 4.3 for more detail.

construct, as expected from the short O⋯ I distance. However, as predicted from the N− H⋯ I angle, the iodine is not as strong as an HB acceptor as the Tyr OH in WT* (Table 4.6). Indeed, WT* was predicted to have an overall energy for direct interactions that is nearly identical to that of Y18^{i}F. Similarly, the weak interaction in Y18^{br}F (a consequence of the less positive σ-hole and less ideal geometry of the bromine) is compensated by a stronger N− H⋯ Br interaction. Not surprisingly, the Y18^{m}F construct shows only weak interactions with the amine group at this site.

An MP2 analysis of Y18F indicates that there are essentially no stabilizing interactions from the Phe to either Glu11 or Gly30. Overall, the MP2 analyses predict the stability of the Y18 constructs to decrease in the following order: Y18^{i}F > WT* > Y18^{br}F > Y18^{m}F > Y18F. However, a water [W1 (Figure 4.2b)] fills in the cavity in the Y18F construct, bridging Glu11, and Gly30 through HBs. We see that the positioning of W1 is relatively conserved for all of the Y18 constructs, except for Y18^{br}F, where it is absent, and Y18F, where it moves significantly compared to the other Y18 constructs (Figure 4.6). When W1 is taken into account, the MP2 energies [$E_{MP2\ (quaternary)}$] show that the Y18F construct becomes very stabilizing, much more so than the other Y18 halogenated or methylated constructs (Table 4.5 and Table 4.7). Thus, the MP2 analysis that includes W1 predicts the stability decreases in the following order: WT* > Y18F > Y18^{i}F ≈ Y18^{m}F > Y18^{br}F.

Thermal Melting Studies To Assess Protein Stability

DSC-determined melting temperatures (T_M) and melting enthalpies (ΔH°_M) for each construct (Table 4.5) showed that any substitution made either at position Y18 or Y88 is destabilizing relative to WT*, reinforcing the general under- standing that it is very difficult to engineer a more stable T4 lysozyme.[172] Indeed, the stability, as reflected by the T_Ms, decreases in the

Table 4.6. Quantum mechanical energies of the binary and ternary complexes.[a]

Construct	Substituent (*Z*)	$\Delta E_{MP2\ (binary,\ E11)}$ (kcal·mol^{-1})	$\Delta E_{MP2\ (binary,\ G30)}$ (kcal·mol^{-1})	$\Delta E_{MP2\ (ternary)}$ (kcal·mol^{-1})
WT*	OH	-1.0	-2.7	-3.4
Y18F	H	-0.2	-0.1	-0.3
Y18^{m}F	CH_3	-0.6	-0.8	-1.4
Y18^{br}F	Br	-0.6	-2.4	-2.4
Y18^{i}F	I	-1.6	-1.9	-3.5

[a]MP2 level quantum mechanical energies were calculated for noncovalent interactions from the *Z*-substituent of Y18^{Z}F constructs (where *Z* = H, OH, CH3, Br, or I) to the HB donor of Gly30 ($\Delta E_{MP2\ (binary,\ G30)}$), to the HB/XB acceptor of Glu11 ($\Delta E_{MP2\ (binary,\ E11)}$), or to both ($\Delta E_{MP2\ (ternary)}$). Quantum mechanical energies were calculated by MP2 method, applying the aug-cc-PVTZ basis set, in cyclohexane with BSSE corrections. Refer to Figure 4.2 and 4.3 for a reference of the schematic showing the geometries of the interacting residues. We note that the individual interaction energies ($\Delta E_{MP2\ (binary,\ E11)} + \Delta E_{MP2\ (binary,\ G30)}$) sums to approximately that calculated with both interaction groups in the model, indicating that the ternary model is fairly accurate, and can be segregated into its individual components.

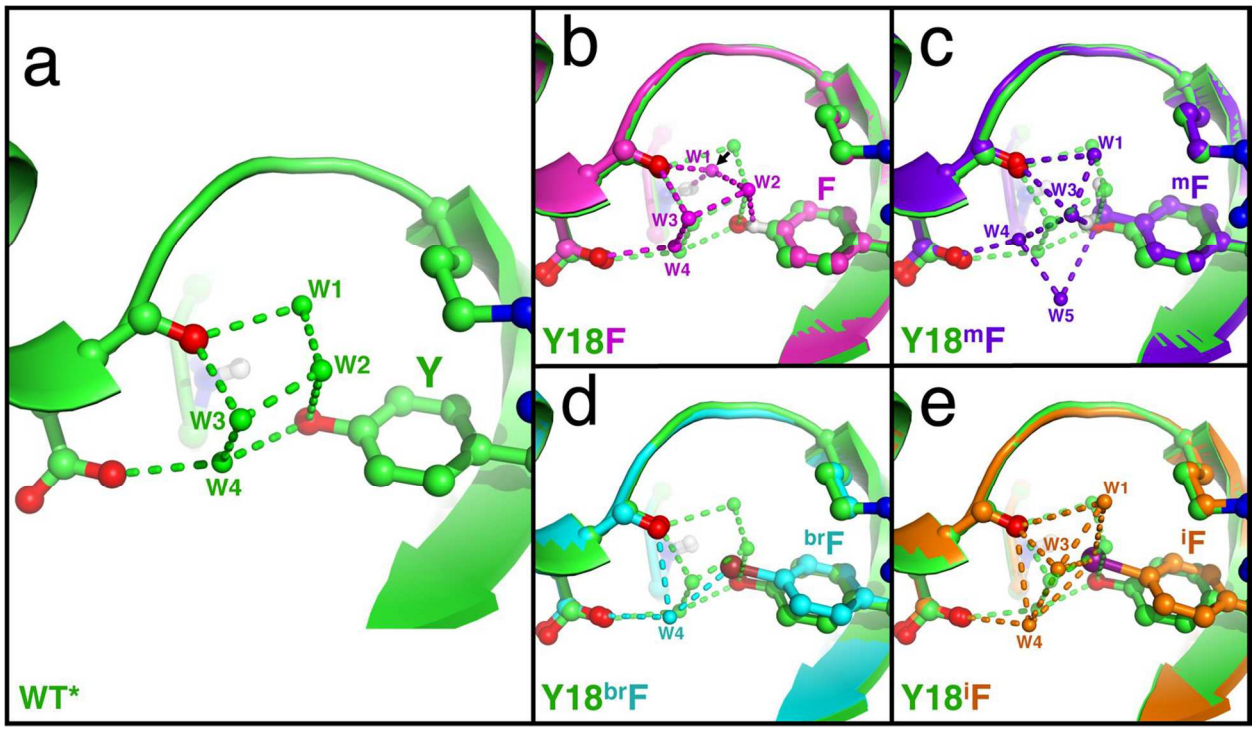


Figure 4.6. Structure of water molecules in the halogenated and nonhalogenated T4 lysozyme constructs. (a−e) Details of the crystal structures of the halogenated and nonhalogenated T4 lysozyme constructs around the Y18 residue. In each panel, the structures of the modified residues at position 18 are superimposed on the structure of WT* (a, green). The modified constructs are colored (b) magenta for Y18F, (c) purple for Y18^{m}F, (d) cyan for Y18^{br}F, and (e) orange for Y18^{i}F. The nonbonded spheres are water molecules, which are colored with their associated structure. The waters in each modified construct are aligned and labeled relative to the closest corresponding water in WT*. The arrow in panel b shows the shift in the position of W1 to accommodate the loss of the Tyr hydroxyl group in the Y18F construct.

Table 4.7. Quantum mechanical energies for HB interactions of water W1 to the oxygen (O) of Glu11, amide group (N–H) of Gly30, and/or the ring hydrogen (C–H) of Phe18 in the Y18F construct.

Interacting Residue (Functional Group)	ΔE_{MP2} **(kcal·mol^{-1})**
Glu11(O)···W1	-6.0
Gly30(N–H)···W1	-4.6
Phe18(C–H)···W1	-1.3
[Glu11(O) + Gly30(N–H)]···W1	-11.5
[Glu11(O) + Phe18(C–H)]···W1	-6.1
[Gly30(N–H) + Phe18(C–H)]···W1	-7.4
[Glu11(O) + Gly30(N–H) + Phe18(C–H)]···W1	-13.2

[a]Interaction energies (ΔE_{MP2}) were calculated by MP2 method with the aug-cc-PVTZ basis set in cyclohexane, with BSSE corrections applied. ΔE_{MP2} values were calculated for W1 paired with the individual amino acids (atomic coordinates taken from the refined crystal structures), with pairs of amino acids ([Glu11(O) + Gly30(N–H)], [Glu11(O) + Phe18(C–H)], or [Gly30(N–H) + Phe18(C–H)]), or with all three amino acids as a quaternary complex ([Glu11(O) + Gly30(N–H) + Phe18(C–H)]. The positions of the hydrogen atoms were optimized prior to the energy calculations. From this analysis, we see that the ΔE_{MP2} values for pairs that are approximately the sums of the energies to the respective individual amino acids, and the quaternary complex the sums of the individual residues, or the pairs plus the individual residues. Thus, the overall energies of interactions within the quaternary complex is accurate, and can be accurately segregated into the individual components.

following order: WT* > Y18F > Y18^{i}F > Y18^{m}F > Y18^{br}F (which initially did not follow our predictions from the MP2 analysis until W1 was added to the calculations). Analyses of the control Y88^{Z}F constructs indicate that all Z substitutions destabilize the protein, with halogenation being the most destabilizing. Comparing the thermodynamic parameters for the Y18^{Z}F to Y88^{Z}F sites [$\Delta T_{M\,(18-88)}$ and $\Delta\Delta H°_{M\,(18-88)}$] allows us to gauge the effects of the XB and HB interactions at the Y18 site on the stability of the protein relative to the general substituent effects. In general, Z substitutions at both positions destabilize the protein, with nearly equivalent effects on the T_M (lowering T_M by an average of 1.7° and 1.6° for the Y18 Z F and Y88Z F constructs, respectively). However, there is a significant difference in the effect of the substituents on $\Delta\Delta H°_{M}$, with Y88^{Z}F showing on average a loss of enthalpic stability of ~ 4 kcal/mol greater than that of the Y18^{Z}F constructs, which can be attributed to the additional intramolecular interactions seen in the latter.

Focusing on individual interactions, we see that the iodine as a substituent has a significant effect on both the differences in $\Delta T_{M\,(18-88)}$ and $\Delta\Delta H°_{M\,(18-88)}$, reflecting the contribution of the iodine XB to the enthalpic stabilization of the protein and the associated increase in the relative melting temperature. Thus, we can confidently state that the XB seen in the crystal structure also exists in solution. The Y18F shows the largest $\Delta\Delta H°_{M\,(18-88)}$, which is in agreement with the MP2 analysis of the direct intramolecular interactions at this site when W1 is included. As previously suggested, W1 of Y18F shifts to compensate for the lost hydroxyl group in WT* (Figure 4.6b). Indeed, the MP2 analysis shows that W1 becomes energetically more favorable by ~ 2 kcal/mol when shifted from its WT* position to that of the Y18F construct, which may help to account for the increased $\Delta\Delta H°_{M\,(18-88)}$ for this construct (Table 4.5). This enthalpic stabilization, however, does not translate into an increase in $\Delta T_{M\,(18-88)}$, which could be

interpreted as the entropic penalty resulting from positioning this water into a structurally fixed position.

Both Y18^{m}F and Y18^{br}F constructs show small effects on $\Delta\Delta H°_{M\,(18-88)}$; however, the methyl substituent results in significant thermal destabilization, while the bromine only slightly increases the stability. Thus, the $\Delta\Delta H°_{M\,(18-88)}$values reflect the contributions of each substituent group on the molecular interactions at the Y18 site, as quantified by the MP2 analysis of the crystal structures. The data indicate that the iodine of Y18^{i}F forms an XB that contributes significantly to the stability of the protein in solution, relative to iodination at position 88 that cannot form an XB. The effect of these interactions on the overall thermal stability of the protein, however, remains a more complex relationship.

■ 4.5 CONCLUSIONS

We have for the first time designed and engineered an XB directly into a protein and, in the process, determined the geometry−energy relationships for the interaction in this biomolecular system. We had previously introduced XBs into a model DNA junction[14] and showed that MP2 analyses of the interaction from the geometries in the crystal structures correlated with DSC-measured energies[14,15,100], resulting in the development of a force field[17] to model the interaction. The energetic contribution of XBs that we have now characterized in the T4 lysozyme helps us test this force field to more accurately model XBs in proteins, including direct applications in the design of new halogenated inhibitors against clinically important cellular targets[38,125], by informing the development of new scoring functions for lead discovery[173].

In this study, we see a linear relationship between the MP2-calculated molecular interactions and the melting enthalpies [$\Delta\Delta H°_{M\,(18-88)}$] in T4 lysozyme (Figure 4.7), but in this case, we need to consider the solvent structure for the relationship to hold. This analysis shows that the

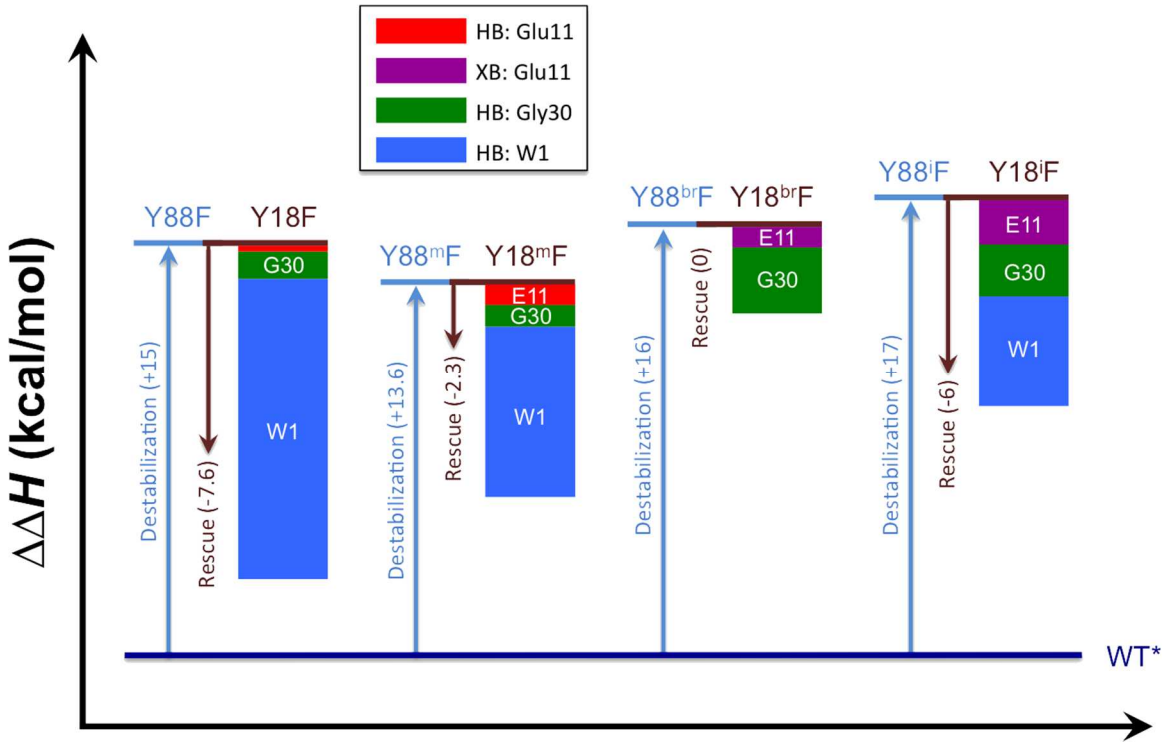


Figure 4.7. Molecular interactions at Y18 that rescue destabilizing substituents. The effect of replacing the hydroxyl group of Y88 in Y88^{Z}F constructs, where Z is H (Y88F), methyl (Y88^{m}F), bromine (Y88^{br}F), or iodine (Y88^{i}F), results in all cases in a destabilization of T4 lysozyme, as reflected in the positive $\Delta H°_{M}$ relative to that of native WT* (blue arrows). The losses in enthalpic stabilization are rescued by specific molecular interactions when the Z substituents are instead engineered at the Y18 site [maroon arrows, calculated as differences in $\Delta H°_{M}$ for the Y88 vs. the equivalent Y18 construct, $\Delta\Delta H°_{M\ (18\text{-}88)}$ (Table 4.5)]. The $\Delta\Delta H°_{M\ (18\text{-}88)}$ enthalpies are linearly correlated with the total MP2 energies from HB interactions of the substituents with the carboxyl group of Glu11 (E11, red bars), N−H of Gly30 (G30, green bars), a bridging water (W1, blue bars), or an XB to E11 (violet bars) (**R** = 90.7%). The slope of 0.87 of this linear relationship suggests that the general rescue of experimental enthalpic energies is modeled well by the interactions included in the MP2 calculations, but the **y**-intercept of 4.3 kcal/mol indicates that additional destabilizing effects have not been included in the model.

hydroxyl group in WT* is too distant to form a direct HB to the Glu11 peptide oxygen, yet WT* remains overall more stable than any of the Y18^{Z}F constructs. Again, we must consider the solvent structure. A detailed analysis of all the current crystal structures that maintain the tyrosine at the Y18 position (WT* and all the Y88 constructs) shows four very well-defined water molecules help bridge the Y18 hydroxyl to the Glu11 peptide oxygen (Figure 4.6a), one of which also connects Y18 to the side chain of Glu11 in the catalytic pocket of the enzyme (labeled W4). In all control Y88 constructs, this constellation of waters remains intact. In all cases in which Y18 is replaced, W4 remains intact, while W1 is either repositioned or displaced (Figure 4.6b−e). In the case of Y18F, W1 is repositioned to fill the void space left by the loss of the tyrosyl OH group, while in Y18^{m}F and Y18^{i}F, this water remains largely in place, suggesting that W1 is particularly important for conferring stability to the entire protein. Our MP2 calculations show that W1 positioned at the Y18F position is ∼ 2 kcal/mol more favorable than its similar position in WT*, which explains why simply removing the hydroxyl group was not as detrimental to the protein's stability as expected. Thus, when we consider how to engineer a more stable protein, it is perhaps not surprising that we must pay attention to not only the direct interactions within the protein but also how they affect the solvent structure.

The significance of the solvent in the structure also applies to the enzyme function. The hydroxyl group of Y18 sits near the substrate binding pocket and interacts with the side chain of the catalytic Glu11 residue through a bridging water (W4 in Figure 4.6). We see that the activities, as monitored by a turbidity assay[168], of the substituted Y18 constructs are all significantly diminished when the OH of the Tyr residue is replaced with a halogen or non-halogen substituent, even though W4 remains intact. The activities of constructs in which Y18 and its constellation of waters remain intact are all at least 40% of that of WT* (Figure 4.8). In

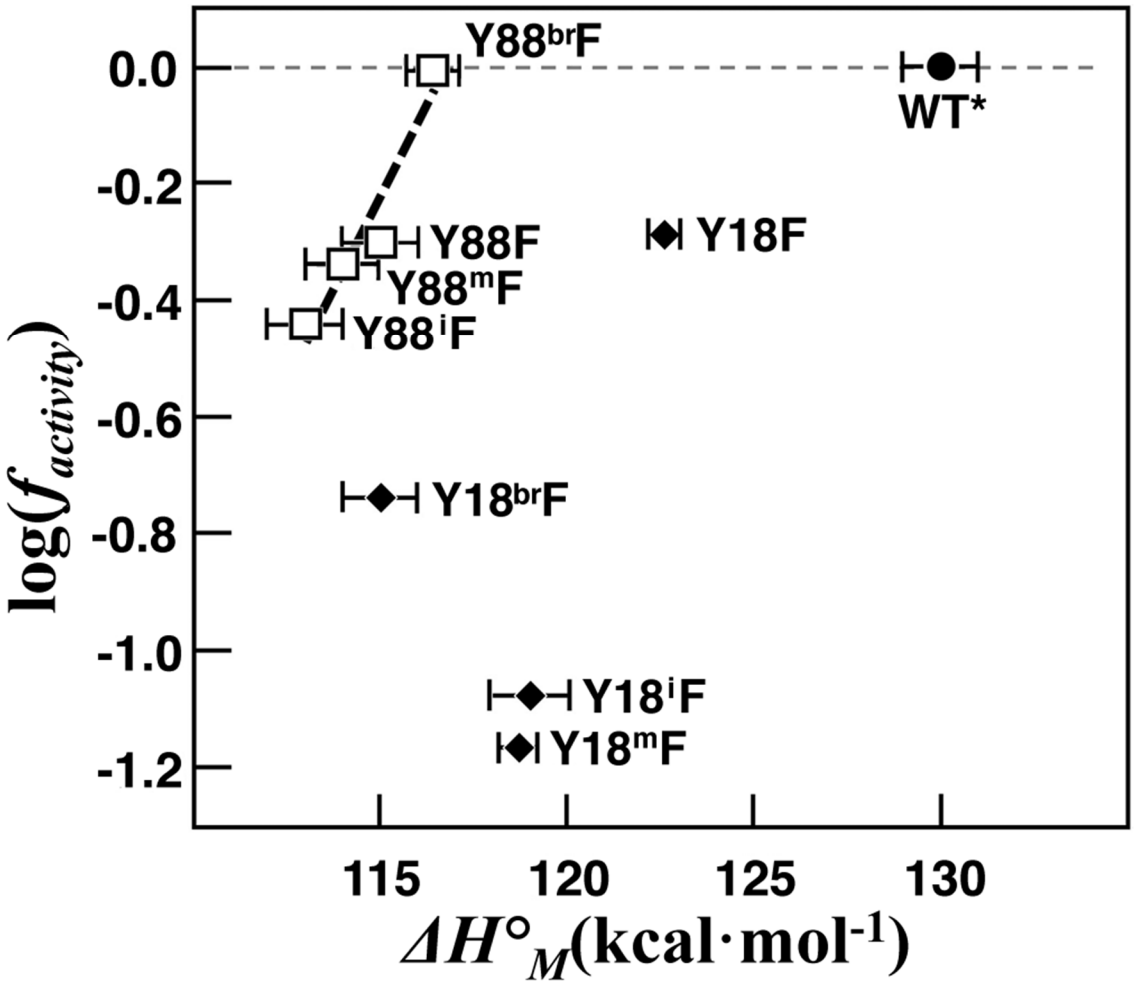


Figure 4.8. Cell clearing activity vs melting enthalpy. The base-10 logarithm of the fractions of the enzymatic activities for each construct relative to that of WT* ($\log(f_{activity})$) are plotted against their melting enthalpies ($\Delta H°_M$), as determined from the DSC measured melting data (Table 4.5). Diamonds show the T4 lysozyme constructs that have been modified at the interacting Y18 residue, while squares are the control Y88 constructs. WT* is shown as a circle. The $\log(f_{activity})$ are linearly correlated with 0H°M for the control Y88 constructs (R^2=0.92). Error bars show standard deviations of the mean for $\Delta H°_M$ measured for each construct.

contrast, the halogenated and methylated constructs, where this cluster of waters is disrupted, show diminished activities. The Y18F construct, however, retains significant activity even though it has lost the direct interactions to Glu11, suggesting that indeed W1 helps to compensate for the missing OH.

The additive nature of noncovalent interactions suggests that introducing XBs into the system can result in a more thermally stable protein. Although we did not succeed in creating an overall more thermally stable T4 lysozyme, we have shown that XBs will help rescue effects that generally cause instability in a protein (Figure 4.7). All modifications, including halogenation, to Y88 were seen to destabilize the protein to a similar degree (∼15 kcal/mol enthalpy). By moving the halogen from this noninteracting position to Y18, we can restore much of the loss of thermal stability by the relatively strong iodine XB. The XB (with an MP2 energy of − 1.6 kcal/mol), however, was not in an ideal geometry, with a Θ_{XB} of 150° (30° from the optimal linear angle of 180°). Thus, there is significant room to improve the geometry and consequently to increase the stabilizing potential of the engineered XB in this and other proteins. As ncAAs, including halogenated residues, become more widely applied to engineer proteins with new functions, the XB can provide added stability to an otherwise destabilizing substitution.

The application of XBs in protein engineering can be easily expanded beyond simply affecting protein stability. We can envision that XBs can be introduced at interfaces to engineer new protein−protein interactions, recognition sites, and even XB-dependent enzymatic catalysts. Organocatalysts have been designed in which XB donors help to accelerate halide abstraction[174,175], nucleophilic substitution[176], and aza-Diels−Alder reactions[177]. In addition, XBs are thought to facilitate iodination abstraction by iodothyronine deiodinase[178]. In such catalysts, the XB interaction helps to weaken the covalent bond to facilitate extraction of the leaving

group. The introduction of amino acids as XB donors can thus provide new catalytic capabilities that take advantage of their tunability and high directionality.

CHAPTER 5

X-TINKER: AN ALGORITHM FOR THE RATIONAL DESIGN OF BIOLOGICAL HALOGEN BONDS [4]

5.1 SUMMARY

The halogen bond (or X-bond) is a noncovalent interaction that is increasingly recognized as important in protein-ligand interactions, and is being engineered to control the structures of proteins and nucleic acids. In the past ten years, there have been significant efforts to characterize the structure-energy relationships for this interaction in macromolecules. Progress in the computational modeling of X-bonds in biological molecules, however, has lagged behind these experimental studies, with most molecular simulation algorithms, particularly molecular mechanics and dynamics (MM/MD) methods, not treating the X-bond at all. In the current study, we demonstrate how the force field for biological X-bonds (*ff*BXB), which we developed as a set of potential energy functions that describe the anisotropic charge distribution and shape properties of halogens participating in X-bonds, can be incorporated into a molecular simulation program. In doing so, we have generalized the *ff*BXB by reducing the number of variable parameters from the original seven to one for each halogen type, and show that this remaining charge variable can be accurately estimated for any new halogenated molecule, including inhibitors, through a routine Restricted Electrostatic Potential (RESP) calculation of atomic charges. As a proof of principle, we have parameterized this more reasonable *ff*BXB against the AMBER force field and incorporated it into TINKER. The resulting X-bond specific program

[4] The work in this chapter is formatted for *an ACS journal*. A.K.R., P.S.H. and I planned computational studies and wrote the manuscript. I wrote the X-TINKER routines and tested them with input from A.K.R. and P.S.H.

Formatted for Submission.

(X-TINKER) was tested against experimental studies on X-bonds incorporated into DNA junction and T4 lysozyme model systems. X-TINKER was shown to accurately predict the interaction energies and recapitulate geometries from x-ray structures of our experimental system. The approach used in creating X-TINKER is readily portable to other commonly used molecular simulation programs for biomolecular design, including those applied to the development of new inhibitors against therapeutic targets in medicinal chemistry.

5.2 INTRODUCTION

Halogen atoms are abundant substituents in pharmaceutical compounds, comprising 50% of the leading drugs currently on the market[60]. Historically, halogens have been utilized to improve absorption, distribution, metabolism, and excretion (ADME) properties like drug metabolism or bioavailability[128,179]. Recently, however, halogens have been shown to form a favorable non-covalent interaction called the halogen bond or X-bond[10], which increases affinity and specificity of inhibitors against their protein targets[52–56]. Over the past few decades, the strategies to develop new and more effective pharmaceuticals has undergone a fundamental shift from purely empirical trial and error to more rational approaches that utilize the power of molecular simulation[180]. In order to exploit the X-bond concept in rational design strategies, they must be properly modeled in molecular simulation programs—unfortunately, they currently are not. Here, we demonstrate how a set of potential energy functions developed to accurately model X-bonds (the force field for biological X-bonds, or *ff*BXB) can be incorporated into the molecular simulation program TINKER, creating X-TINKER. This strategy for developing X-TINKER is generally applicable to creating variations of other simulation programs to properly model X-bonds in biomolecular systems.

The application of computational rational design has found success in the development of biologics for catalysis[181–183], biofuels [184], and drugs[185]. The variety of computational tools that have been used in biomolecular design include (in increasing order of computational complexity and cost) molecular docking, molecular mechanics (MM), molecular dynamics (MD), and quantum mechanics (QM) algorithms. Only QM, however, has the inherent capacity of properly treating the unique aspects of X-bonds.

To understand the challenge in properly modeling for X-bonds in MM/MD molecular simulation programs, we must first understand the nature of the X-bond. Although there are several competing physical descriptions for the X-bond[11,12], the most readily accessible description comes from the σ-hole theory[34]. In this model, a halogen that is covalently bonded to another atom will have its valence electron pulled into the σ-bond, creating a positive crown called a σ-hole that sits diametrically opposed to the σ-bond (Figure 5.1a). The roots of two unique physical properties of halogen substituents—anisotropic charge distribution and flattening of the van der Waals radius—can be rationalized by this σ-hole concept. Consequently, this positive σ-hole allows halogens to interact with a variety of electron-rich atoms, serving as X-bond acceptors (Figure 5.1b). The size of the σ-hole, which helps to define the strength of the X-bond, increases with the size and thus the polarizability of the halogen, as well as the electron-withdrawing ability of the atom or molecule to which the halogen is bound (Figure 5.1c)[43]. With this relatively straightforward model in mind, we can now understand why properly modeling X-bonds is a particular challenge to molecular simulation algorithms.

With the acceptance that X-bonds affect folding and recognition, there have been several focused efforts to model the interaction in biomolecular systems. The standard force fields used in molecular mechanics (MM) and dynamics (MD) algorithms treat halogens as isotropic spheres

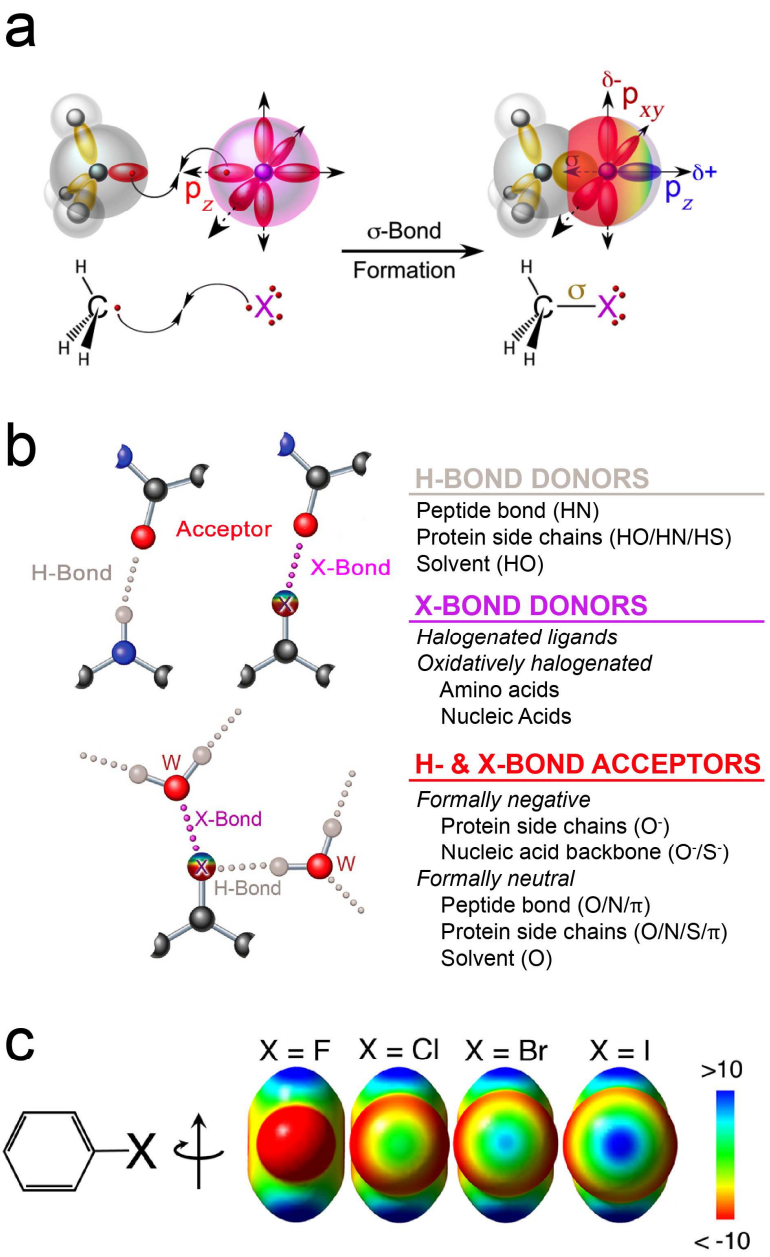


Figure 5.1. The X-bond. (a) Cartoon of the σ-hole theory. (b) Comparison of H- and X-bond donors and acceptors in biological systems. H-bond donors are typically the electropositive hydrogens bonded to more electronegative atoms (O, N, or S) in proteins or nucleic acids. Contrarily, halogenated ligands comprised most X-bond donors, but they are also found on halogenated amino acids or nucleic acids that result from oxidative halogenation. Similarly, H- and X-bonds are primarily electrostatically driven interactions, thus, share a common set of electron-rich acceptors. The anisotropic charge distribution results in halogens being amphipathic, allowing them to serve simultaneously as X-bond donors and H-bond acceptors[134]. (c) Degree of electrostatic potential variation of the σ-hole across the halogens is shown from -10kcal/mol (red) to +10kcal/mol (blue). Adapted from Ford and Ho[19] and Scholfield et al.[186]

with uniform charge distributions and shapes; thus, they cannot account for any aspect that makes the X-bond a stabilizing interaction. Quantum mechanics (QM) calculations can recapitulate the anisotropic properties of halogens and, consequently, accurately model X-bonds, but at computational costs that make them impractical for studying biopolymers. There have been some success seen with hybrid QM/MM approaches to help with inhibitor design[60,187], but the most efforts have been to incorporate X-bonds into Class I force fields of classical MM/MD algorithms. One approach in this effort is the positive extra point (PEP) strategy, where a massless pseudoatom with a defined positive charge is placed at or near the surface of the halogen to mimic the anisotropic charge distribution associated with the σ-hole. This strategy has also seen some success in reproducing the geometries and relative energies of ligand interactions with proteins and, thus, have been incorporated into various Class I force fields, including AMBER[85], CHARMM[188], and OPLS3[189]. The drawbacks are that the PEP parameters for any new ligand to be studied must be defined through QM calculations and do not account for polar flattening, which affects the important contribution of dispersion in defining the angular dependence of X-bonds[190]. Recently, multipole (MTP) electrostatics has been implemented in CHARMM for X-bonding, which has resulted in a vast improvement in energies, but at a much greater computational cost with slowdowns of 8-10x depending on the number of MTP pairs involved[191]. With these challenges in mind, we had previously derived a set of potential energy functions that model the anisotropic electrostatic and shape properties of halogens by defining their charge and van der Waals radius as being angle dependent[24]. The resulting force field for biological X-bonds (*ff*BXB) could accurately reproduce the experimental energies for X-bonds determined from a set of studies on model DNA junctions[18]. The limitations of the approach, however, were that there were seven independent parameters (four shape and three charge

variables) in the *ff*BXB and, as with the PEP, parameterization of the model required high level QM calculations for each new halogenated compound.

In this study, we have reparameterized the *ff*BXB so that there is now only a single independent charge variable that needs to be determined, and show that this variable can be readily assigned through a standard restrained electrostatic potential (RESP) calculation that is commonly used to parameterize any new compound prior to classical MM/MD simulations. As a proof of concept that the *ff*BXB can be readily integrated into MM/MD algorithms, we have incorporated the reparameterized force field into the open source program TINKER[192], creating X-TINKER. Finally, we show that MM calculations with X-TINKER can recapitulate energies and geometries of X-bonds in DNA and now model protein systems.

■ 5.3 THEORY AND METHODS

The X-bond Model for Class I MM/MD

The ultimate goal of this paper is to properly model the halogen's anisotropic charge distribution and non-spherical van der Waals surface using a classical force field without the need for additional parameterization. This model will then be implemented into Class I molecular dynamics making it important to first understand the functional form of these Class I potential energy equations (Eq. 5.1). As seen in Eq. 5.1, there are a series of equations describing the bonding terms (bonds, angles, dihedrals, and improper dihedrals) and nonbonded term (generally an electrostatic potential and a 6-12 Lennard-Jones potential).

$$V = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2$$
$$+ \sum_{dihedrals} K_\phi(1 + \cos(n\phi - \delta)) + \sum_{\substack{improper \\ dihedrals}} K_\varphi(1 + \cos(n\varphi - \varphi_0))$$
$$+ \sum_{nonbonded} \frac{Z_i Z_j e^2}{4\pi\varepsilon_0 r_{ij}} + \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] \qquad (Eq.\,5.1)$$

We have previously modified the nonbonding portion of this functional form to accommodate X-bonding. In short, the modification for halogens has an additional cosine function in both the electrostatic term and the Lennard-Jones equations to describe the angular dependence found in X-bonding (Eq. 5.2). It is important to add that Class I MM/MD simulations are generally performed on pairwise interactions, for X-bonding this would be between the halogen (X) and its acceptor (A). However, the addition of the cosine function for X-bonding causes the interaction to be a three-body term. The three bodies will be: 1) the atom bound to the halogen (Y), 2) the halogen (X), and 3) its acceptor (A).

$$V_{(r_{AX},\alpha)} = \frac{[A\cos(\nu\alpha) + B]Z_A e^2}{4\pi\varepsilon_0 D {r_{AX}}^n} + \varepsilon_A\varepsilon_X\left[\left(\frac{R_A + \langle R_X\rangle - \Delta R\cos(\nu\alpha)}{r_{AX}}\right)^{12} - \left(\frac{R_A + \langle R_X\rangle}{r_{AX}}\right)^{6}\right]$$

(Eq. 5.2)

Parameterization of Equation 5.2 using QM fitting

In standard programs, Class I MM/MD simulations (Eq. 5.1) require three non-bonding parameters as input for all new systems: 1) van der Waals radii of interaction atoms i and j (R_i and R_j), 2) minima of potential energy well of the van der Waals interaction (ε_i and ε_j), and 3)

electrostatic charges (Z_i and Z_j). Generally, the shape or van der Waals parameters are found using high-level QM, empirically, or both[93]. Charge parameters are found using either Restrained Electrostatic Potential (RESP[193,194]) charge model or AM1 with bond charge corrections (AM1-BCC[195]) charge model.

For the modified Eq. 5.2, apart from the standard parameters: $R_{A,X}$, $\varepsilon_{A,X}$, and $Z_{A,X}$, additional parameters need to be determined. Additional parameters were initially found using QM. For QM on all systems in this paper, second-order Møller-Plesset (MP2) calculations were used applying the aug-cc-PVTZ basis set for Cl, and Br; and aug-cc-PVTZ-PP for I[146]. Calculations were done using cyclohexane (D = 2) and a gas phaseBSSE correction[147] was included. For QM fitting of Eq. 5.2, a selection of systems were used, varying the type of X-bond donors and acceptors). The geometries of these systems were defined by varying the angle of approach (Θ_1) from 90° to 180°, and varying the interacting distances from 70% to 100% of the standard radius of the various halogens to the acceptors. These calculated energies generated a QM energy landscape for each system that could be used to fit Eq. 5.2. This was done using a non-linear least squares fit floating the parameters needed.

Energy minimization simulations

MM energy minimization simulations were carried out using TINKER[192] and X-TINKER, which contains 3 extra subroutines written in Fortran 95 (Supplemental). The small molecule mimic of the T4 lysozyme system (described in more detail below) was energy minimized in a vacuum using steepest descent to a final energy gradient RMS of 0.05 kcal/molÅ. The overall T4 lysozyme systems were prepared using AmberTools16[196] in a rectangular TIP3P water box with Amber99sb parameters[197] using the starting structures from Scholfield et al[186]. The systems were minimized using a multistep protocol using TINKER or X-TINKER. This protocol first

minimized the solvent using steepest descent to a final energy gradient RMS value of 0.8 kcal mol^{-1} $Å^{-1}$ with the protein inactive. The macromolecule was then minimized to a final energy gradient RMS value of 0.6 kcal mol^{-1} $Å^{-1}$.

■ 5.4 RESULTS AND DISCUSSION

Reduced and General ffBXB Parameters

We had previously fully parameterized the seven variables of the *ff*BXB (Eq. 5.2, Table 5.1) for Cl, Br, and I in terms of the experimental geometries and energies of X-bonds seen in a DNA junction system (Figure 5.2)[17,18]. The X-bonds seen in this DNA system were modeled as the interactions in a complex formed between a halogenated uracil base (^{X}U) and an anionic hypophosphite X-bond acceptor ($O_2PH_2^{-1}$, mimicking the phosphate of the DNA backbone) (Figure 5.2a inset). This minimal model allowed an accurate energy landscape to be derived from high level QM calculations, with the X-bond donors and acceptors sampling a broad range of distances and angles of approach. The *ff*BXB parameters were then globally fit to reproduce the QM landscape. The resulting parameterized *ff*BXB accurately reproduced the energies of the DNA junction system, in terms of the experimental energies associated with the X-bond geometries from the crystal structures, thereby validating both the minimal $^{X}U\cdots O_2PH_2^{-1}$ as a model for the DNA system and the *ff*BXB approach. Although accurate, the seven independent parameters of the *ff*BXB made it unwieldy. In addition, the parameters were specific for a DNA system, although most biological X-bonds involve ligands bound in a protein environment. Thus, our first goal for the current study was to reduce the number of independent variables in Eq. 5.2, and, in the process, define parameters that are general for both proteins and nucleic acids.

We started with the shape parameters that defined the average van der Waals radius ($<R_X>$), the perturbation to that radius due to polar flattening (ΔR_X), and the contribution of the halogen

Table 5.1: Parameters from Equation 5.2 fit to QM values. [a]

	Shape Parameters				Electrostatic Parameters			Fit Correlation
	$\langle R_{vdw} \rangle$ (Å)	ΔR (Å)	ε_x (kcal/mol)	v	n	A	B	R^2
				Fully Parameterized				
Chlorine	1.698 ± 0.010	0.1487 ± 0.0010	0.107 ± 0.002	2.39 ± 0.05	2.4 ± 0.3	0.17 ± 0.05	-0.08 ± 0.02	0.9995
Bromine	1.831 ± 0.015	0.160 ± 0.018	0.11 ± 0.04	2.41 ± 0.05	2.66 ± 0.19	0.19 ± 0.04	0.18 ± 0.04	0.9998
Iodine	1.933 ± 0.014	0.185 ± 0.005	0.087 ± 0.008	2.21 ± 0.04	2.28 ± 0.13	0.40 ± 0.04	0.31 ± 0.05	0.9997
				Parameters for MM/MD				
Chlorine	1.687 ± 0.012	0.15	0.107 ± 0.002	2	2	0.12 ± 0.007	-0.062 ± 0.006	0.9986
Bromine	1.798 ± 0.016	0.16	0.11 ± 0.04	2	2	0.113 ± 0.008	0.057 ± 0.010	0.9987
Iodine	1.918 ± 0.011	0.19	0.087 ± 0.008	2	2	0.330 ± 0.011	0.233 ± 0.010	0.9997

[a]The fully parameterized fit was found by allowing all parameters in Eq. 5.2 to float. The parameters for MM/MD simulations were found by fixing n, ΔR, and v .

to the van der Waals energy (ε_x). Since these parameters were determined explicitly by probing the molecular halogens Cl_2, Br_2, or I_2 with a small nonpolarizable helium atom, their values for each halogen type were already defined as fixed and, thus, Eq. 5.2 of the *ff*BXB really only has one variable shape parameter, ν. The parameter ν reflects the position of the most electronegative waist of a halogen substituent, which affects where an H-bond donor, for example will approach the atom. The non-integer value of ν (>2) derived for the fully parameterized *ff*BXB indicated that the p_{xy} orbitals were slightly tipped away from 90° relative to the σ-bond. However, these were deemed to be details that would not dramatically affect the X-bonding potential of a halogen substituent and, therefore, the value of ν was fixed to an integer value of 2, defining the position of the electronegative annulus explicitly to be perpendicular to the σ-hole. In doing so, we have also greatly simplified the derivatives of the potential energy functions in Eq. 5.2, as will be described in a latter section.

We next tackled the parameters that define the charge distribution across the atomic surface of the halogen. The parameter *n*, which reflects the power of the distance dependence and thus the form of the electrostatic interaction, was seen from the fully parameterized *ff*BXB to fall between 2 and 3. Again, we fixed *n* to be the integer 2, which explicitly defines the X-bond as a charge-dipole type interaction.

The *A* parameter defines the amplitude difference in charge between the positive σ-hole and negative waist, while *B* defines the overall charge of a halogen substituent. Together, they define the angle at which the halogen charge is neutral. We expect both *A* and *B* to be dependent on the inductive effects of the atom or molecule that the halogen is covalently bonded to, with an electron withdrawing group enhancing and electron donating group diminishing the σ-hole and, consequently, the X-bonding potential of the halogen. We thus constructed a set of halogenated

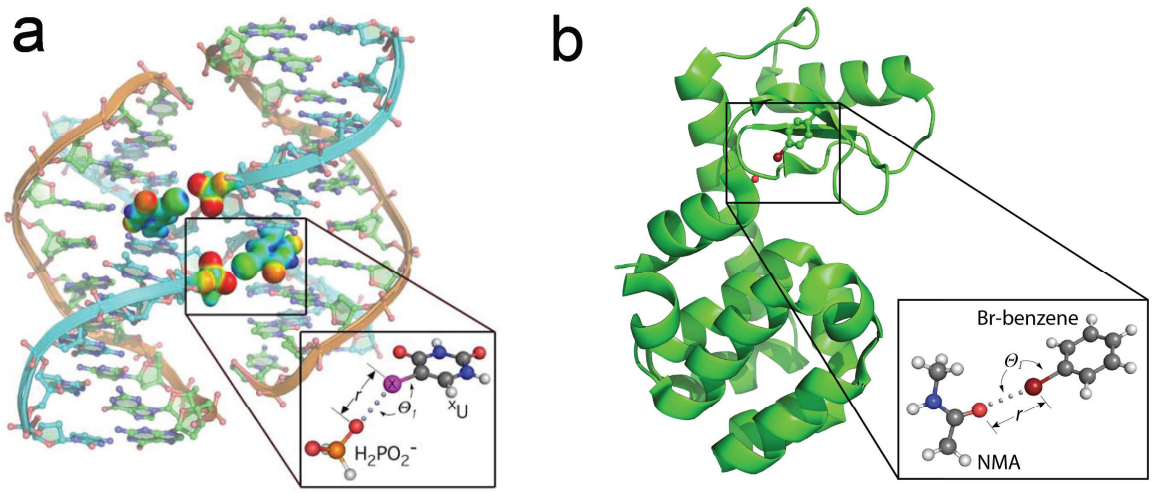


Figure 5.2. Experimental and Computational Setup. (a) The previously characterized experimental DNA junction system[15] is a four-stranded stacked-X DNA junction. The small molecule mimic of the stabilizing X-bond, $O_2PH_2^{-1}\cdots\ ^XU$, that was used in computational studies is depicted in the inset. (b) The previously solved T4 Lysozyme structure[186] is shown with its small molecule mimic, N-methylacetamide (NMA)$\cdots$ Br-benzene, shown in the inset. Adapted from Scholfield et al.[18]

benzene compounds, in which various substituents (their electron donating to electron withdrawing potential defined by Hammett constants, Table 5.2) were placed *para* to the halogen, and calculated the energies of interactions (E_{int}) to the anionic $O_2PH_2^{-1}$ and, now also to *N*-methylacetimide (NMA) as a neutral X-bond acceptor. The distances and angles of approach between the X-bond donors and acceptors were varied to develop a QM energy landscape and, as before, we used this landscape to parameterize the *ff*BXB. In this case, however, we kept the values for all of the shape parameters and *n* fixed, as described above, fitting only *A* and *B*. The results showed that for each halogen type, *B* became more positive, as expected, *para*-substituent of the donor became more electron withdrawing. However, *A* was fairly consistent within each halogen type (Table 5.2). These results suggest that the amplitude difference between positive and negative regions of the halogen are defined by the polarizability of the halogen (which in turn is defined by its atomic size), while the overall charge and the neutral point angle is variable and defined by extent to which the p_z orbital of the halogen is depopulated. We could, therefore, fix *A* to an average value for each halogen type, as determined from the QM landscape, now leaving *B* as the only parameter of the *ff*BXB as being variable.

With this set of six fixed and one variable parameters, we show that the *ff*BXB can accurately recapitulate the QM energy landscape. As seen in Table 5.1 and Figure 5.3, the reparameterized Eq. 5.2 fits the QM calculated E_{int} landscape very well, particularly at θ_1 from 120° to 180°. The primary deviations, are at 90°, where the *ff*BXB predicts a more positive E_{int} than the QM calculation, as expected when ν was fixed to 2. However, deviations at this angle of approach should not affect the ability of the *ff*BXB to predict the X-bonding potential of each halogen. Furthermore, by constructing the QM energy landscape with both the $O_2PH_2^{-1}$ and neutral NMA acceptors, we have parameterized the *ff*BXB to be applicable to both nucleic acids and proteins.

Table 5.2: *A* Parameter for each system[a]

			A Parameter		
Y	**Acceptor**	**σ_p**	**Chlorine**	**Bromine**	**Iodine**
CN	NMA	0.66	0.114	0.194	0.403
CN	hypophosphite	0.66	0.124	0.180	0.325
Cl	hypophosphite	0.23	0.104	0.177	0.302
F	hypophosphite	0.23	0.127	0.190	0.298
H	NMA	0	0.111	0.198	0.317
H	hypophosphite	0	0.129	0.187	0.286
CH_3	hypophosphite	-0.17	0.115	0.192	0.314
OH	hypophosphite	-0.37	0.115	0.192	0.337
NH_2	NMA	-0.66	0.116	0.208	0.402
NH_2	hypophosphite	-0.66	0.119	0.197	0.287
		Average	0.117±0.008	0.191±0.009	0.33±0.04

[a] Y represents the para-substituent relative to the halogen. σ_p denotes the Hammet-sigma constant, where the greater the value, the more electron-withdrawing.

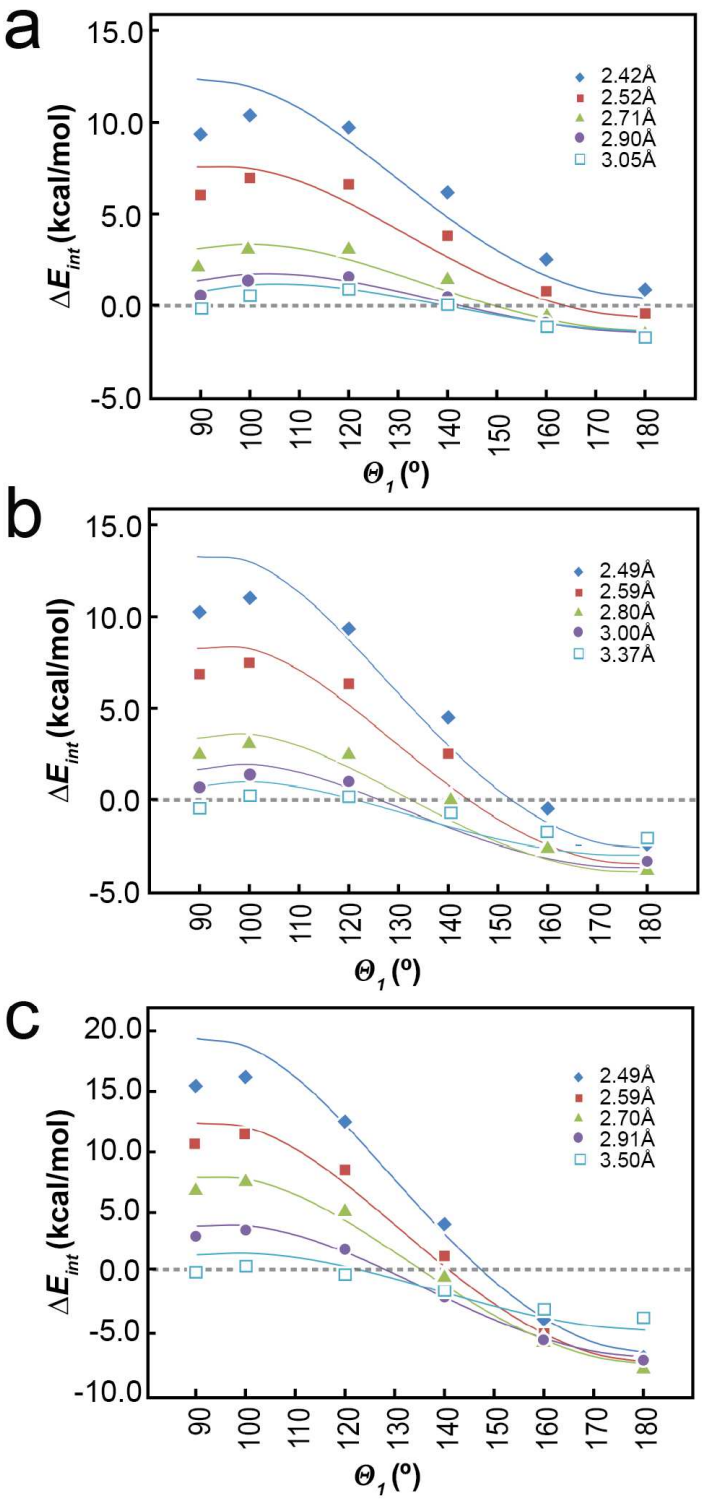


Figure 5.3. Results of QM energy landscape with *ff*BXB fit for (a) Cl, (b) Br, and (c) I. QM calculated interaction energies (ΔE_{int}) were performed with the DNA junction mimic ($^{X}U\cdots O_2PH_2^{-1}$, Figure 5.2a) at various distances (data points) and, for each distance, at angles of approach (Θ_1) from 90˚ to 180˚. The curves represent the values calculated using the parameterized *ff*BXB with fixed parameters for MM/MD implementation (Table 5.1).

Non-QM fitting of the ffBXB parameter

The *ff*BXB has now become less unwieldy, but still requires high-level QM calculations to determine the single remaining variable parameter (*B*). At this point, we asked whether *B* could be determined through a non-QM calculation. For classical Class I MM/MD simulations, an RESP or AM1-BCC protocol is used to assign the isotropic charge of atoms of a new compound or ligand. Since *B* reflects the overall charge of a halogen, it was possible that this parameter could be determined from either of these protocols. The RESP charges for the atoms of the halogenated benzene X-bond donors with various electron donating or withdrawing *para*-substituents were calculated and compared to the *B* parameters determined from the QM E_{int}. We saw that the RESP assigned charges were linearly related to the QM determined values for *B* (Figure 5.4).

Integration of the ffBXB into TINKER

With the *ff*BXB reduced to a single variable parameter without significantly affecting its ability to model the QM energy landscape of an X-bond in either a DNA or protein system, we can now integrate the force field into an MM/MD algorithm. As a proof of concept, we elected to work here with the open source MM/MD program TINKER, creating an X-bonding variant we call X-TINKER. Our goal was to develop an integration approach that would require minimal modification to the main program, so that the methodology could be generally portable to other programs. The integration methodology was constructed around three distinct routines: 1) identification of a halogen and its covalently bonded atoms, as required to define the angular components of the *ff*BXB equations; 2) assignment of *ff*BXB parameters to the halogen; and 3) energy and force calculations.

The first routine identifies and keeps track of any halogen that is present in the molecular system, along with atoms that are 1, 2, or 3 covalent bonds from that halogen. A logical is then set to "TRUE" for X-bonding, triggering the program to execute the next two X-bond only routines. The next routine assigns values to the *ff*BXB shape and charge parameters that are specific to the type of halogen being considered. Finally, the list of atoms and the assigned parameters are passed on to the final routine to calculate the potential energy and force between the halogen and any potential X-bond acceptor.

The next step is to calculate the energy of each halogen with any potential X-bonding acceptor. This routine starts by taking any atom that X-TINKER identified as interacting with the halogen atom, then subtracting out the nonbonding energy calculated by the main TINKER program for the interacting pair. Although this step slightly increases the computational costs (in that X-TINKER first calculates the energy, then subtracts it out), this approach does not require modification to the main program. We elected to sacrifice some speed for more general adaptability of the methodology. With the standard noncovalent interaction energy subtracted out, the routine then proceeds to calculate the potential X-bonding energy of the halogen to this same neighboring atom. The distance between the X-bonding pair ($r_{X\text{-}A}$) would have already been calculated by the main program. The X-bond routine calculates the angle of approach of the acceptor to the halogen (A···X—B, or θ_1). Notice that by assigning the *ff*BXB parameter *v* with the integer value of 2, the cosine of this angle is simply calculated by the Cartesian dot product in Eq. 5.3. Eq. 5.2 is then applied using the $r_{X\text{-}A}$ distance and θ_1 angle to calculate the X-bonding energy, which is then passed back to the TINKER main program to incorporate into the overall energy of the system.

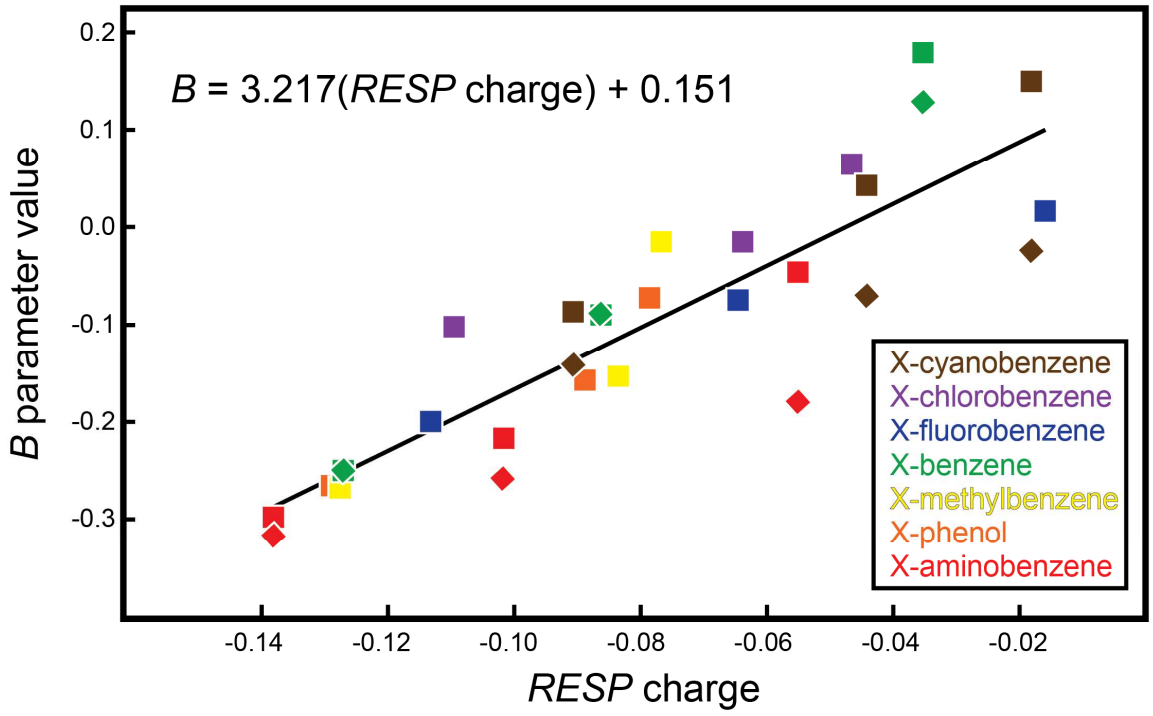


Figure 5.4. *B* parameter vs *RESP* charge of halogen. QM calculated energies of various halogenated benzenes (color in legend) interacting with either a $O_2PH_2^{-1}$ (squares) or N-methylacetamide (diamonds) were calculated for various distances and angles then fit using the *ff*BXB with all parameters fixed except *A* and *B* (Table 5.1). The *B* parameter was then plotted against the halogen's overall charge calculated using the RESP charge model.

$$\cos 2\theta = 2\left(\frac{r_{BX} \cdot r_{AX}}{|r_{BX}||r_{AX}|}\right)^2 - 1 \qquad (Eq. 5.3)$$

To test the accuracy of the energy calculations in this subroutine, the same $^{X}U\cdots O_2PH_2^{-1}$ model complex used for the QM of the DNA junction (Figure 5.2 inset) was run through the energy routine of X-TINKER. The system was fixed to the X-ray crystallography geometries in order to compare the energies calculated by X-TINKER to the experimentally determined energies. As seen in Figure 5.5, energies found using Eq. 5.2 implemented into TINKER agree very well with experimental energies, with an R^2 of 0.960 and a slope ≈ 1.

Once X-TINKER was shown to successfully calculate X-bonding energies, it needs to be able to calculate the force associated with the halogen and its nonbonded pair in order to allow energy minimization and/or MD simulations. The force equations were derived as the first derivatives of the Lennard-Jones potential (V_{LJ}) and separately the electrostatic potential (V_{Elec}) energy functions of Eq. 5.2 in respect to the *x*-, *y*-, and *z*-directions. These derivatives are shown in in Eqs. 5.4 and 5.5 in respect to the *x*-direction for simplicity, but are repeated for the *y*- and *z*-directions. Eqs. 5.4 and 5.5 were tested by comparing the force calculated from these derivatives to the finite difference of the potential energies (final minus initial coordinates) $^{X}U\cdots O_2PH_2^{-1}$ model as the positions of the molecular components were perturbed along the *x*-, *y*-, or *z*-directions. As seen in Table 5.3, the finite differences (numerical derivative) matched exactly to the derivative calculated (analytical derivative) using Eqs. 5.4 and 5.5 in TINKER.

After confirming the successful implementation of the energy and force calculations, X-TINKER was further validated using a system that was outside of the DNA junction system or the trained set—in this case, the X-bonds engineered into T4 lysozyme (T4L, Figure 5.2b). In this T4L test system, a bromo- or iodophenylalanine replaced the tyrosine at the Y18 position, to create the constructs ^{Br}F18-T4L or ^{I}F18-T4L.[186] The crystal structure of ^{Br}F18-T4L showed a

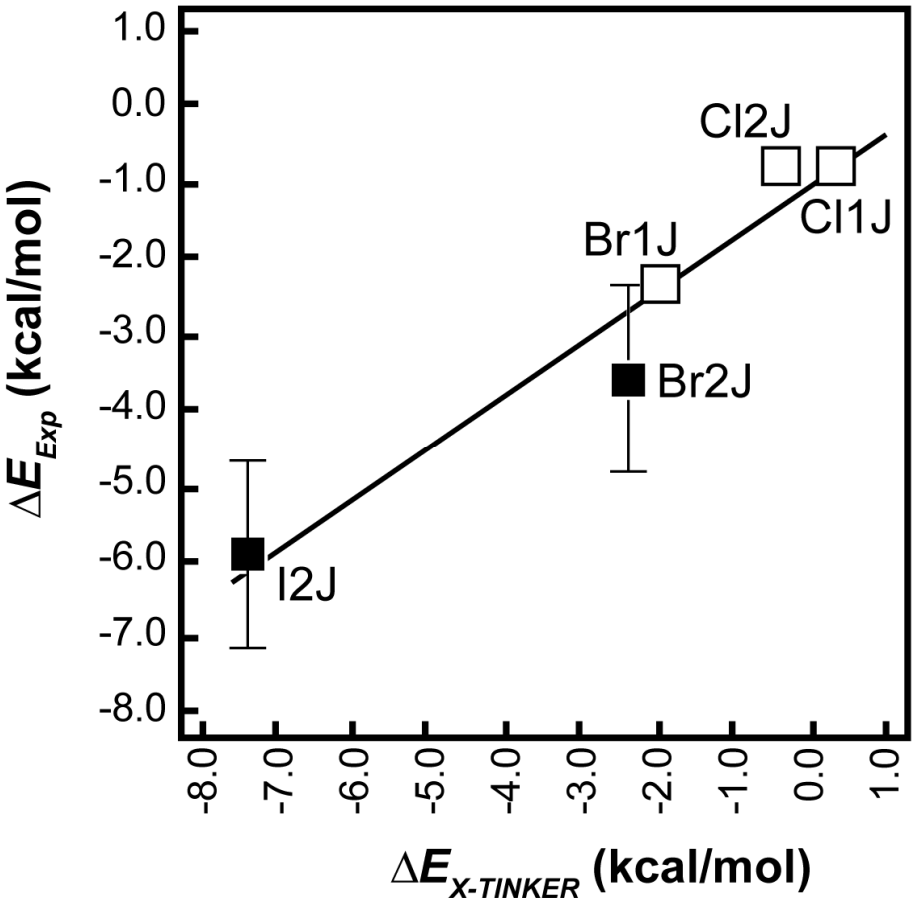


Figure 5.5. Experimental energies vs. theoretical energies from X-TINKER of DNA junction system. Energies found by either crystallographic competition assays (open squares) or differential scanning calorimetry (closed squares)[31] are compared to theoretical energies found by X-TINKER. Geometries were taken from crystal structures in Carter et al.[33], where either one halogen (X1J) or two halogens (X2J) were competed against 2 H-bonds.

Table 5.3: Analytical vs Numerical Derivatives Calculated with X-TINKER[a]

Derivative	Atom	dE/dX	dE/dY	dE/dZ
Analytical	Y	0.6568	-2.7566	-0.1392
Numerical	Y	0.6568	-2.7566	-0.1392
Analytical	X	0.2946	2.2215	0.6227
Numerical	X	0.2946	2.2215	0.6227
Analytical	A	0.1322	-1.7527	-0.0362
Numerical	A	0.1322	-1.7527	-0.0362

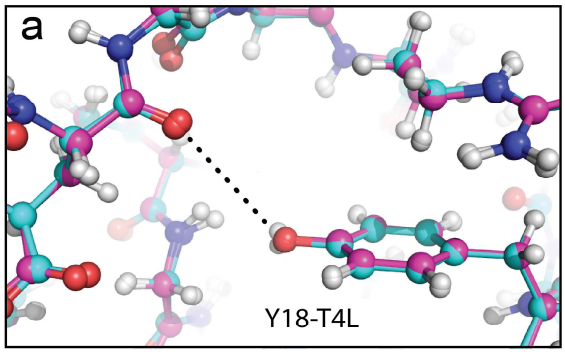
[a]Derivatives were calculated on the small molecule mimic of NMA··· Br-benzene and the 3 interacting atoms are shown: the halogen (X), the atom covalently bound to the halogen (Y), and the acceptor (A).

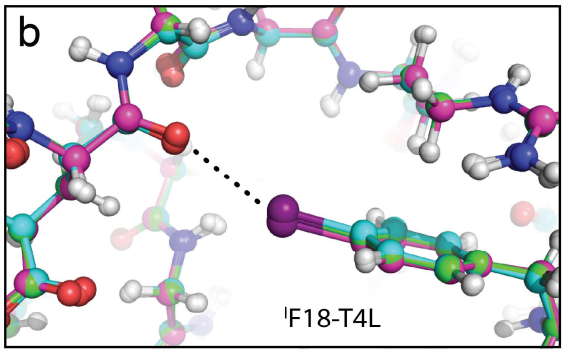
$$\frac{d}{dx}V_{LJ} = -12\varepsilon\left[\frac{R_{0(A)} + \langle R_{0(X)}\rangle - \Delta R\ cos\ 2\alpha}{r}\right]^{11} \times \left[\frac{\Delta R}{r}\frac{dcos2\alpha}{dx} + \frac{R_{0(A)} + \langle R_{0(X)}\rangle - \Delta R\ cos\ 2\alpha}{r^2}\frac{dr}{dx}\right] \quad (Eq.5.4)$$

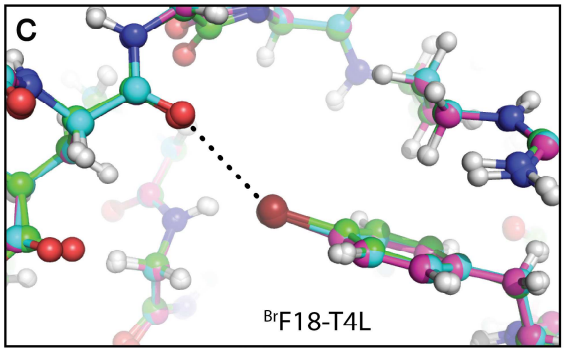
$$\frac{d}{dx}V_{Elec} = \frac{Z_A A e^2}{Dr^2}\frac{dcos2\alpha}{dx} - \frac{2[A\ cos(v\alpha) + B]Z_A e^2}{Dr^3}\frac{dr}{dx} \quad (Eq.5.5)$$

long weak Br···O X-bond interaction to the carbonyl oxygen of the protein backbone (E_{int} = -0.6 kcal/mol), while the ^{I}F18-T4L showed a shorter and stronger I···O X-bond (E_{int} = -1.6 kcal/mol). We applied TINKER and X-TINKER to perform energy minimization simulations on the X-ray structures of each of the T4L construct (the WT* and the halogenated variants Y18^{br}F and Y18^{i}F), with a convergence to a final RMS for the energy gradient of 0.6 kcal/mol Å. It was clear that both programs kept the structures of the protein essentially identical to those of their respective X-ray structures. These results reflect the very rigid conformational constraint imposed on this particular site by the protein. The overall energies, however, were seen to be much more favorable from X-TINKER as compared to TINKER for the halogenate structures (Figure 5.6), reflecting the contribution of the X-bonding potential. Finally, the T4L simulations serve as important tests for how the patch affects the timing of the TINKER program. Using one CPU, X-TINKER only slows TINKER down by ~10%, representing a reasonable trade-off between speed and accuracy.

We further modeled the Br···O X-bonding interaction as a complex between a bromobenzene (mimicking the ^{Br}F18 side chain as the X-bond donor) and NMA (mimicking the backbone carbonyl oxygen acceptor) in a system that was entirely unconstrained by the protein. Energy minimizations were performed using both TINKER and X-TINKER on this model interacting







Construct	TINKER *E* (*kcal/mol*)	X-TINKER *E* (*kcal/mol*)
Y18-T4L	-72701	-72701
^{Br}F18-T4L	-74156	-78510
^{i}F18-T4L	-76365	-79744

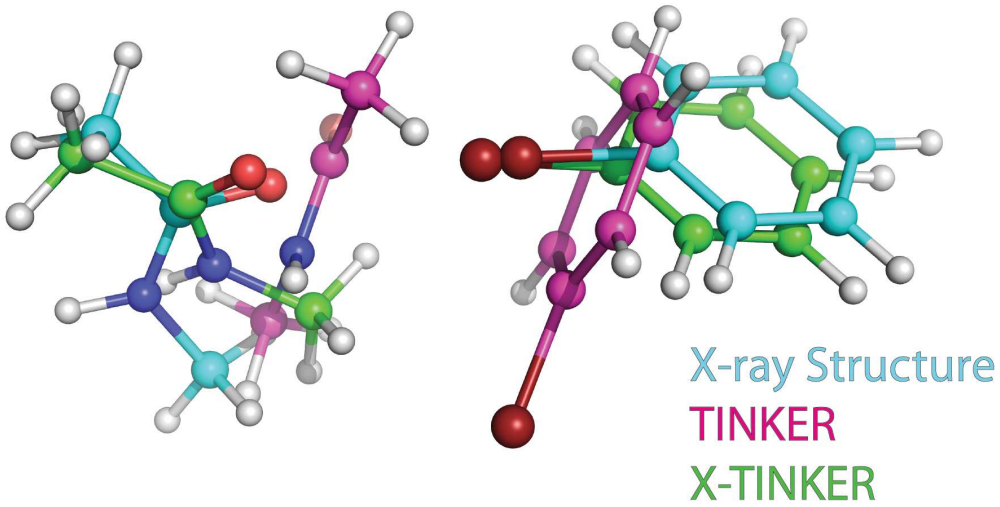
Figure 5.6. TINKER vs. X-TINKER energy minimization of T4L system. The X-ray structure (cyan) of the Y18-T4L (a) ^{Br}F18-T4L (b) the ^{i}F18-T4L (c) was energy minimized using either TINKER (magenta) for all constructs and X-TINKER (green) for the halogenated constructs. The entire structure energies for the minimizations can be found in the table.

pair, with the resulting structures compared to that of the X-ray structure (Figure 5.7). The more stabilizing energies of ~ 6 kcal/mol from the TINKER minimization compared to the X-TINKER minimization can possibly be attributed to interactions from π-systems of the NMA and the Br-benzene. It is clear, however, from these simulations that the parent TINKER program in the absence of the X-bonding potential and the protein environment could not maintain the close Br···O interaction, placing the donor parallel with the acceptor. X-TINKER, however, faithfully recapitulated the X-bonding interaction seen in the crystal structure, even in the absence of the remainder of the protein.

■ 5.5 CONCLUSIONS

X-bonding has become an increasingly important area of research, which is beginning to be used in biological applications including drug design and protein engineering. With the constant improvements in computational power over the past decades, medicinal chemists and biomolecular engineers have relied increasingly on many computational approaches for the rational design of new compounds and materials. One of the most powerful of such approaches has been the development of MM/MD simulations, which have been successfully applied to modeling macromolecular conformations, along with their binding to various ligands. Still, standard MM/MD programs are challenged by atomic systems, such as halogens, that do not behave isotropically.

There have been various approaches to incorporate X-bonding into MM/MD simulations. The PEP and MTP approaches have made great leaps towards more accurately capturing the electrostatic component of the σ-hole model for X-bonding. Each method, however, comes with its own set of disadvantages. While the PEP approach provides a very fast computational fix, it



Method	$r_{Br...O}$ (Å)	Θ_1 (deg)	*E* minimized (*kcal/mol*)
X-ray Structure	3.37	142.7	–
TINKER minimized	5.29	52.8	-25.35
X-TINKER minimized	2.96	167.2	-19.48

Figure 5.7. TINKER vs. X-TINKER energy minimization of T4L mimic. The X-ray structure (cyan) of the X-bond interaction of the ^{Br}F18-T4L was energy minimized using either TINKER (magenta) or X-TINKER (green). The resulting X-bonding geometries and overall minimized structures energies can be found in the table below.

comes at the cost of accuracy. MTP, on the other hand, more accurately describes the halogen's electratic properties, but at a high computational cost.

In this paper, we offer an alternative to implementation of X-bonding, that of the *ff*BXB into an MM/MD program. The *ff*BXB was derived in an attempt to model X-bonds based on the physical properties of halogen substituents—the anisotropic charge distribution and polar flattening associated with the σ-hole model. We have now reparameterized the potential energy functions of the *ff*BXB to become less unwieldy, reducing the seven independent variables to a single one and showing that this remaining variable can be readily obtained through a standard RESP calculation for any new or unknown compound. This optimized force field was implemented into the open source MM/MD program TINKER to create X-TINKER, which was shown to be capable of recapitulating the geometries and energies of X-bonds seen in both model DNA junctions and protein systems in the presence and absence of their respective macromolecular backbone. These initial energy minimization simulations offer a good proof of concept that the *ff*BXB can be successfully implemented into an MM/MD program. This implementation of X-TINKER is applicable to a large variety of X-bond donors without the need for costly QM calculations. Finally, since only a minor alternation needs to be made to the original source code, our approach to developing X-TINKER, including the functional subroutines, is readily portable to other more widely used MM/MD simulation programs.

We have shown here that the *ff*BXB, as implemented into X-TINKER, accurately captures the energies of well-defined systems. We have not, at this point, determined how well it can model the solvent interactions in and around halogen atoms, nor how the entropy of the overall macromolecular system is modeled. Although this seems intuitively simple, halogens are hydrophobic, making it different from the H-bond. The proper modeling of the desolvation of the

halogen atom will be vital to model the overall entropy of the system. In order to test these aspects of the force field, we would need to perform full MD simulations on complex biomolecular systems, but such a test would be best performed using an implementation in an MM/MD approach, such as free energy perturbation, that is better suited to treating all aspects of biomolecular dynamics.

Even with some of these questions, X-TINKER offers a unique MM/MD algorithm that for the first time accurately models the halogens electronic structure in many complex systems without significant computational costs, when comparing the original with the altered code. This program is now one additional computational tool that can be applied to rational design of halogenated compounds for both medicinal chemistry and biomolecular engineering applications.

CHAPTER 6

EXAMINING THE FEASIBILITY OF USING FREE ENERGY PERTURBATION (FEP+) IN PREDICTING PROTEIN STABILITY [5]

■ 6.1 SUMMARY

The importance of engineering protein stability is well known and has the potential to impact many fields ranging from pharmaceuticals to food sciences. Engineering proteins can be both a time consuming and expensive experimental process. The use of computation is a potential solution to mitigating some of the time and expenses required to engineer a protein. This process has been previously hindered by inaccurate force fields or energy equations and slow computational processors, however improved software and hardware have made this goal much more attainable. Here we find that Schrödinger's new FEP+, although still imperfect, proves more successful in predicting protein stability than other simpler methods of investigation. This increased accuracy comes at a cost of computational time and resources when compared to simpler methods. This work adds to the initial testing of FEP+ by offering options for more accurately predicting protein stability in an efficient manner.

■ 6.2 INTRODUCTION

Engineering protein stability is an ever-expanding field having direct applications that range from designing better catalysts for food sciences, chemical assays or biofuels[181,183,198] to studying structure and function of biological targets[199]. Moreover, some success has been found in creating more stable proteins for X-ray crystallography efforts using protein mutation[200],

[5] The work in this chapter was published as a result of my internship in 2016 with Merck & Co. K.B. and I planned computational studies and wrote the manuscript. I performed all FEP+ calculations.

Ford, M. C.; Babaoglu, K. Examining the Feasibility of Using Free Energy Perturbation (FEP+) in Predicting Protein Stability. *Journal of Chemical Information & Modeling.* 2017.

particularly in challenging targets like G-protein-coupled receptors[201,202]. Engineering stable proteins generally starts by using random mutagenesis to generate large libraries, screening those libraries, and finally purifying and testing for stabilizing mutants[203]. However, depending on the resources and screens available this can take anywhere from weeks to a year. The ability to computationally predict mutations that have the highest chance of being stabilizing would aid tremendously in cutting down on these significant burdens on time and resources.

There are many computational methodologies that can be applied to predicting protein stability such as molecular mechanics (MM) and molecular dynamics (MD)[204,205]. The most demanding yet potentially accurate approach available is using free energy calculations, such as free energy perturbation (FEP) or thermodynamic integration (TI). Each method employs different techniques that transform the wild-type amino acid to the mutant amino acid via chemical or alchemical transformation. The transformation is generally split into a series of non-physical intermediate states and an ensemble is generated at each state by Monte-Carlo (MC) or MD techniques[206]. Finally, the resulting ΔG is calculated by employing various statistical mechanics methods. FEP methodology has been around for many years[207–210] and has had some early success in predicting protein stability in T4 lysozyme[211,212]. In recent years, free energy calculations have gained much interest in protein-ligand binding for drug design applications[213] and protein stability calculations[214]. In particular, Schrödinger has improved upon their existing FEP protocol – creating FEP+, which has had success predicting binding affinities for protein-ligand interactions[19,215].

The main limitations to FEP historically have been poorly performing force fields, errors in the potential energy model, inadequate sampling and the high cost of computational time to run the calculations. FEP+ now uses an improved force field, enhanced sampling via replica-

exchange with solute tempering (REST)[216,217], and faster calculations using GPU hardware. In a recent publication on protein-ligand binding prediction using FEP+[19], the prior implementation of OPLS2.1 was found to be an improvement over both OPLS2005[218] and MMFF[219] with better modeling of noncovalent interactions and the inclusion of additional covalent parameters. The current force field OPLS3 includes off-point charges for halogen bonding and nitrogen lone pairs and incorporates better parameters for peptide dihedrals[189]. The REST method was shown to be a more efficient sampling technique to its predecessor, replica exchange method (REM) and demonstrated an improvement in simulation convergence[216]. Lastly, the GPU implementation of the code showed a significant speed-up over CPUs[19]. Now, with a new force field, faster processing, and enhanced sampling, FEP+ is a much more useful program for many applications.

FEP+ has shown initial success in protein-ligand studies and is gaining traction in the pharmaceutical industry[19,220,221]. In light of some of these successes, eleven pharmaceutical companies met to evaluate how ligand based FEP methods were fitting scientific needs, especially those pertinent to drug design[222].

Here, however, we present testing of the default FEP+ in the context of protein stability, compare it to other available MM methods, and discuss practical strategies around employing it in design efforts. There has been a recent, initial test of FEP+ by Schrödinger[223] on soluble proteins; we hope to add to this by offering an external review of the practicality of using FEP+ for such predictions. This includes a thorough analysis of how the program compares to MM methods, the effectiveness of speeding up the costly FEP+ calculations, and the pitfalls of the program as it is now. Additionally, this work uniquely uses a dataset with an equal distribution of stabilizing (N=30) to destabilizing (N=29) mutations as to not overly weight the less desired, but more common destabilizing mutations. This dataset also includes neutral (N=3) mutations,

although not as interesting still important to properly predict. Lastly, due to the difficulty that can arise when experimentally measuring the ΔΔG of folding, the predicted values were all compared to a much more easily attainable ΔT_m value. Ultimately, this work addresses the practicality of using FEP+ or faster MM methods when predicting protein stability in a realistic setting.

There has also recently been a thorough analysis of FEP for protein mutations including an individual force field analysis, a consensus approach using multiple force fields, sampling time analysis, and a dataset of both soluble proteins and a membrane protein[224]. Although these types of analyses are very important for the field and future iterations of software, the work presented here is not intended to do the same analysis. Instead, we aim to assess FEP+ (one of the fastest FEP programs) in its current state for protein stability calculations and how it can be used in the most practical manner.

■ 6.3 MATERIALS AND METHODS

Selecting Mutations

The ProTherm database[225] contains 311 unique proteins with experimental stability data for 12561 single mutations making it impractical to run FEP on all of them. A recent paper from Amgen and Intrexon used a smaller subset of 82 proteins and 799 mutations to develop a machine learning algorithm for stability prediction[226]. Using this dataset, the proteins were further narrowed down to 9 structures with 62 mutations total. The proteins and their respective mutations were chosen to maximize the variety in the final dataset using the following criteria. The proteins were initially filtered to 24 to include proteins that contained more than one stabilizing mutation as to not overweigh destabilizing mutations. Due to the use of alanine scans for many stability assays, the 24 were narrowed to 9 proteins to accommodate a variety of amino

acid changes. The final 62 mutations are composed of 32 destabilizing and 30 stabilizing mutations and 14 different amino acids.

Selecting Proteins and Protein Prep

The starting structures were obtained from the Protein Database (PDB)[57] with the following PDB codes: 1ARR, 1I0V, 1RN1, 1RTB, 1STN, 1WQ5, 2A36, 2AFG, 4LYZ, and 5CRO. To maintain consistency across the protein preparation, all structures were prepared using Schrödinger's Maestro 11 (2016-2) Protein Preparation Wizard. To ensure the validity of using Maestro's prepping algorithm prior to running energy calculations using alternative programs, the proteins were also prepped and reran using solely MOE. Those results were found to be worse with the R-value from the Pearson correlation decreasing from 0.504 using Maestro preparation to 0.382 using MOE preparation. Additionally, the Spearman R-value decreased from 0.417 to 0.215. Therefore, all calculations were done using the protein preparation algorithm in Maestro. All solvent molecules present in the crystal structures were removed and protonation was assigned at pH 7.0. 1ARR and 5CRO are present as homodimers in solution and prepped as such. For 1WQ5, residues 57-64 and 184 and 185 were missing from the crystal structure thus they were added using Prime. 2A36 was solved using NMR, resulting in 10 structures present, structure 5 showed the most consensus and was used going forward.

Energy Calculations

The prepped structures from Maestro were used to run energy calculations using Chemical Computing Group's (CCG) MOE 2015.1001[227]. Biovia's Discovery Studios 2017 (DS)[228], Schrödinger's Bioluminate [229], and Schrödinger's FEP+[19,189]. The default settings for MOE, DS, and Bioluminate were used for each programs protein mutation calculations. MOE was used with an AMBER ff10 forcefield. DS was used with CHARMm Polar H forcefield without pH

dependence or temperature dependence. Bioluminate was run inside Maestro 11 using OPLS3[189] and side-chain prediction with backbone minimization.

For FEP+, the protein structures were solvated in a 5 Å orthorhombic, neutral water box for complex calculations and a 10 Å box for solvent calculations. All simulations were run using an SPC water model. Using a default Desmond[230] protocol provided by Schrödinger's software, the following series of relaxation and equilibration simulations were performed employing OPLS3. The system was first minimized with the Brownie integrator and was simulated using an NVT class ensemble at 10 K for 100 ps with restraints on heavy solute atoms. This was followed with a 12 ps NVT calculation at 10 K using a Berendsen thermostat with heavy solute atoms still restrained. The last equilibrations prior to FEP were 2 NPT calculations at room temperature and 1 atm pressure, the first with restraints still on the heavy solute atoms for 24 ps and the last with no restraints for a total of 240 ps.

Once the systems were relaxed and equilibrated, the production simulation for FEP/REST was performed using NPT class ensembles for both complex and solvent. A 10 Å cutoff radius was used for both vdW and electrostatic interactions along with smooth particle mesh Ewald to calculate long-range electrostatic interactions. The simulations were done first using the default settings of 12 λ at 5000 ps for charge-conserved mutations and 16 λ windows at 5000 ps for charge-changing mutations. The calculations were accelerated by using 4, 6, or 8 λ windows at 500 ps, 750 ps, or 1000 ps respectively. Replica exchanges were attempted every 1.2 ps as described in Wang et al.[19] and the final ΔG's were calculated using the Bennett Acceptance Ratio (BAR)[231]. All FEP calculations were run on 4 NVidia Tesla K80 GPU cores. To assess protein stability over the course of the simulation, RMSD of the backbone atoms was monitored and was consistently observed as ≤ 2.0Å. For the error analysis, the cycle closure correction to

incorporate redundant information into free energy estimates was not used due to the calculations being direct wild-type-to-mutant transformations without redundancy as discussed in Steinbrecher et al.[223]

Statistical Evaulations

Each programs calculated energy values were compared to experimentally found ΔT_m values. A Pearson correlation was used to evaluate the linear relationship found between predicted values and experimental values. A p-value was found for each correlation indicating the significance of the linear relationship. Using a Fisher z-transformation, the sampling distribution from the Pearson correlation was transformed to a normal distribution allowing the calculation of confidence intervals. The confidence intervals were calculated at 95% confidence. Furthermore, a Spearman rank correlation was performed.

■ 6.4 RESULTS AND DISCUSSION

To examine the ability of computation to predict stability, 62 mutations and associated stability data from 9 different proteins were chosen (see Methods for details) from the ProTherm database[225]. A set of commonly used software platforms were used as a baseline comparison to the FEP+ protocol. Each program, Molecular Operating Environment (MOE), Bioluminate (Schrödinger), Discovery Studios (DS), and FEP+(Schrödinger), were run using default settings as an "out-of-the-box" approach (see Methods). Each of the three comparison methods estimates the change in energy upon the introduction of a mutation using classical molecular mechanics, with the major difference being the force field used during the calculations. The final predicted energy for each mutant was compared to the measured ΔT_m found experimentally. In order to use FEP+ in a prospective manner, it is essential to predict the experimental value that is much more widely used in measuring protein stability, ΔT_m. Although the predicted values of $\Delta\Delta G$ are not

absolutely correlated to ΔT_m, they are known to have a negative correlation[232] and have been successfully compared previously[224]. Therefore, scale-independent statistical methods were employed while comparing these two values, which deviate from previous literature[19,223], but give a more practical application.

Initially, the predicted energies were compared to experimental data via a linear Pearson correlation. As seen in Table 6.1, DS, Bioluminate and MOE predicted energies all had a Pearson R^2 of around 0.3 to experimental ΔT_ms. However, FEP+ performed better, predicting stability with a Pearson R^2 of 0.5. The difference in predicted and experimental variables prohibits the use of mean unsigned error (MUE), however, the p-values give validity of the correlation and the confidence intervals (CI) show the range that the population correlation falls between with 95% confidence (Table 6.1).

From a practical perspective, a more useful metric for predictive power is the use of truth tables, which explicitly illustrate the number of calculations a computational tool predicted correctly and incorrectly. To remain consistent with the error of the experimental methods, the mutants were categorized as stabilizing with an experimental $\Delta T_m \geq 1$ ˚C (N=30), destabilizing with a $\Delta T_m \leq -1$ ˚C (N=29), and neutral with a ΔT_m falling between -1 ˚C and 1 ˚C (N=3). For the predicted data, the analysis becomes more convoluted due to the unquantified error associated with each method as well as the varying range of the predicted values for each program. Thus, placing boundaries on a neutral prediction is somewhat arbitrary. In order to determine the neutral category and remain as consistent as possible for each program, the same percentage of the total range of the observed experimental data was used. For instance, a total range of experimental data of 47.5˚C would give a cutoff of ± 1˚C, which is 2% of the total. Applying the same 2% to each method, MOE (range of 4.5 kcal/mol) had a ± 0.09 kcal/mol

Table 6.1. Linear Correlation (Pearson) Statistical Analysis of Various Prediction Methods[a]

Method	FEP+	DS	Bioluminate	MOE
R^2-value	0.502	0.329	0.283	0.254
R-value	-0.708	-0.574	-0.532	-0.504
p-value	1.19E-10	1.08E-06	1.02E-05	3.00E-05
95% CI	(-0.814,-0.557)	(-0.721,-0.378)	(-0.691,-0.324)	(-0.669,-0.290)
Accuracy	0.71	0.55	0.50	0.50
Sensitivity (TPR)	0.67	0.53	0.38	0.21
Fall-out (FPR)	0.17	0.26	0.38	0.11

[a]Values reported were obtained using default settings of FEP+, DS, MOE and Bioluminate as described in methods. Statistical analysis was performed using a linear Pearson correlation. The p-value and 95% CI were calculated for the R-value. The accuarcy, sensitivity, and fall-out were calculated using Eqs. 6.1, 6.2, and 6.3 respectively with the inclusion of neutral values.

cutoff, DS (range of 8.68 kcal/mol) had a ± 0.34 cutoff, Bioluminate (range of 53.97 kcal/mol) had a ± 1.08 kcal/mol cutoff, and lastly FEP+ (range of 12.51 kcal/mol) had a ± 0.25 kcal/mol cutoff. While admittedly arbitrary, these distinctions permitted grouping of the mutants into *true positive* (maximum of 30), *false positive*, *true negative* (maximum of 29), *false negative*, *true neutral* (maximum of 3) or *false neutral* (where false neutrals can be classified into 2 categories: those that were experimentally stabilizing and those that were experimentally destabilizing). Using these classifications, accuracy, true positive rates (TPR or sensitivity), and false positive rates (FPR or fall-out) were calculated using Eqs. 6.1, 6.2 and 6.3 respectively, where a positive indicates a stabilizing mutation (for experimental: a positive ΔT_m or for predictions: a negative $\Delta\Delta G$). Therefore, a true positive is one where the predicted method has a negative energy and the experimental value has a positive ΔT_m. The false positive includes both the *false positives* and the *false neutrals* that were experimentally stabilizing. Using this analysis, FEP+ again stands alone with 71.0% accuracy, where the others fall near 50% (Table 6.1). Additionally, FEP+ provides good sensitivity of 66.7% with a low fall-out rate of 16.7%.

$$Accuracy = \frac{\sum true\ positives\ + \sum true\ negatives}{\sum all\ values} \qquad (Eq.\ 6.1)$$

$$Sensitivity = \frac{\sum true\ positives}{\sum experimental\ positives} \qquad (Eq.\ 6.2)$$

$$Fall - out = \frac{\sum false\ positives}{\sum experimental\ negatives} \qquad (Eq.\ 6.3)$$

Although the FEP+ appears to only be slightly better in terms of a linearly correlated R^2, we cannot be certain that the calculated $\Delta\Delta G$ has a linear relationship to ΔT_m. Spearman ranks are an advantageous statistical analysis when measuring a monotonic relationship without making assumptions about the shape of that relationship. Using this analysis, the experimental data is first ranked from stabilizing ΔT_m to destabilizing and given a numerical value from 1 to N (N =

62 in this dataset), then each predictive method is labeled the same. Now a linear correlation analysis (the standard Pearson) can be used when comparing the experimental rank and each predicted rank. A Spearman Rank correlation on each of the methods yielded a slight improvement of FEP+ with an R^2 of 0.53 and a detrimental effect on the relationship of the MM programs all with R^2 values near 0.2 (Table 6.2).

FEP+ was clearly more accurate than the simpler techniques but at a much greater computational cost. The MM methods were on the order of seconds to minutes per mutation on one CPU, where FEP+ took an average of 6 hours per mutation on 4 GPUs. To use this method in a more applied fashion, it was important to consider if combining or altering approaches could better predict protein stability at a faster speed. Intuitively, the use of MM programs has the potential to save a lot of computational time, yet individually their overall accuracy is underwhelming. For other computational predictive methodologies, such as molecular docking, techniques like consensus scoring have been used to overcome these types of shortcomings. Along these lines, it was hypothesized that if the more poorly performing MM programs were combined, the average rank score obtained might improve the overall predictions providing a "crude screen" to initially filter out the truly destabilizing mutants. A "refined screen" of the remaining top-scoring mutants could then be run using FEP+ to filter out the false positives predicted by the crude screen. Here, FEP+ acts as the refined screen due to the smaller number of mutants needed to run, potentially saving time.

Testing this approach on our data set (Figure 6.1a), a retrospective crude screen was performed using an average rank of MOE, DS, and Bioluminate to remove the bottom half of the mutations (N=31), where 21 were true negatives but 10 were false negative or neutral. For simplicity of analysis, experimentally defined neutral mutants were classified as *destabilizing* since in a true experiment they would not be selected for validation (Figure 6.1b). Of the 31

Table 6.2. Rank Correlation (Spearman) Statistical Analysis of Various Prediction Methods[a]

Method	FEP+	DS	Bioluminate	MOE
R^2-value	0.526	0.209	0.189	0.174
R-value	-0.726	-0.457	-0.434	-0.417
p-value	2.537E-11	1.89E-04	4.18E-04	7.49E-04
95% CI	(-0.826, -0.581)	(-0.634, -0.234)	(-0.617,-0.207)	(-0.604,-0.187)

[a]Statistical analysis was performed using a Spearman rank correlation. The p-value and 95% CI were calculated for the R-value.

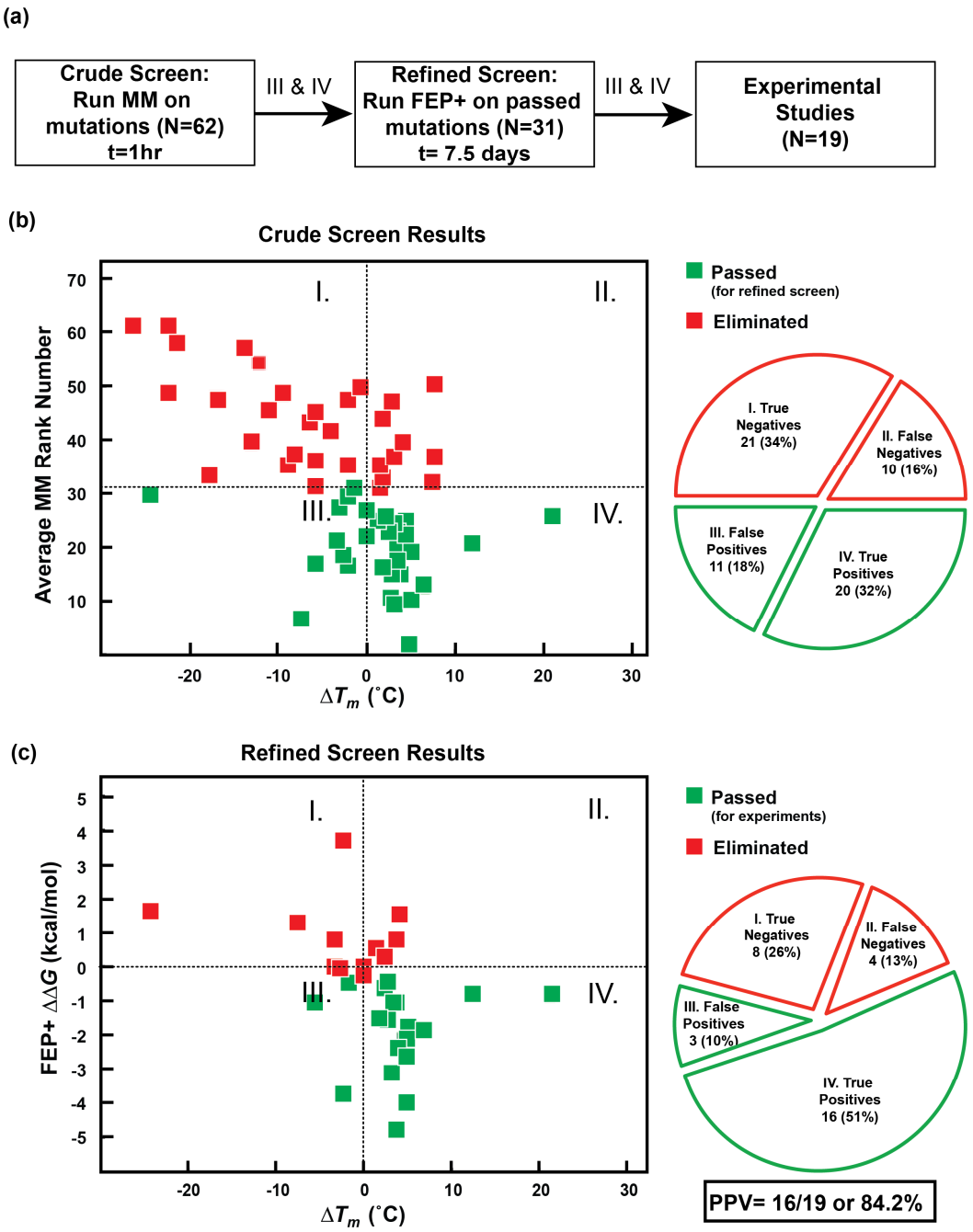


Figure 6.1. Results of protocol for predicting protein stability using default programs. (a) Overview of screen is shown with time and number of mutations for each step. (b) Results of crude screen are shown after ranking the average MM prediction and eliminating the top half (red). The resulting bottom half (green) proceeded to the refined screen. The breakdown of the negatives (destabilizing) and positives (stabilizing) are located in the pie charts to the right, where the neutral values are labeled as negative (not stabilizing) for simplicity. (c) Results of the refined screen are shown after the FEP+ ΔΔG is calculated. The unfavorable ΔΔ*G*s (red) are eliminated, proceeding with the favorable ΔΔ*G*s (green) to experimental studies. This would yield a final positive predictive value (PPV) shown.

mutations that would be carried forward to the refined screen, 20 were *true positives* and 11 were *false positives*. The best expectation was that after running the refined screen all of the *false positives* would be identified, however, only 8 of these were identified while 3 were labeled as *false positives*. Additionally, 4 mutations were incorrectly eliminated using the refined screen leaving 19 mutations that passed the selection criteria with 16 true positives. These results yielded a positive predictive value of 84.2% for the proceeding experiments (Figure 6.1c). Although a high hit rate was returned and a total of 16 stabilizing mutants were found, this method would still take 180 hours or 7.5 days to run.

The question remains: can we construct an even faster, more accurate workflow using the tools in hand? Potential approaches to answer this question are to make either the crude or refined screen more accurate, faster, or both. Generally, these goals are diametrically opposed as making the programs more accurate compromises the time of the calculation. However, it is important to further consider both approaches and the feasibility of each.

Starting with the less practical approach of making the refined screen more accurate, it is relevant to examine what went poorly in the FEP+ simulations. Generally, it is important to examine how the calculation performed including the convergence of the simulation for both the solvent and complex calculations. 10 of the 62 mutations were found to have poor convergence. Similar to Steinbrecher et al[223], after removing the non-converged data, the Spearman rank R-value increased from -0.726 to -0.813.

Although the overall statistical correlations for the entire set were mediocre, when inspecting the nine different proteins, the targets had vastly dissimilar correlations (Table 6.3). The methodologies, like most structure-based approaches, appear to show target-biased effects with some targets being 100% accurately predicted and others being essentially random. This result

Table 6.3. Comparison of Prediction Performance by Target[a]

PDB	1arr	1rn1,1i0v	1rtb	1stn	1wq5	2a36	2afg	4lyz	5cro
# Mutations	5	10	7	9	7	5	6	8	5
MW (kDa)	12.5	33.6	13.7	16.8	59.5	6.9	15.8	14.3	14.7
				FEP Results					
Spearman R^2-value	0.915	0.571	0.548	0.174	0.655	0.045	0.087	0.963	0.367
Spearman R-value	0.956	0.756	0.740	0.417	0.810	0.211	0.295	0.981	0.605
p-value	1.09E-02	1.15E-02	5.72E-02	2.64E-01	2.73E-02	7.33E-01	5.71E-01	1.65E-05	2.79E-01
95% CI	(0.93, 0.97)	(0.62, 0.85)	(0.60, 0.84)	(0.19, 0.60)	(0.70, 0.88)	(-0.04, 0.44)	(0.05, 0.51)	(0.97, 0.99)	(0.42, 0.74)
FEP Accuracy	1.000	0.600	0.714	0.556	0.714	0.800	0.500	1.000	0.600
				MM Results					
Spearman R^2-value	0.414	0.417	0.552	0.201	0.567	0.104	0.282	0.796	0.760
Spearman R-value	0.643	0.646	0.743	0.449	-0.753	-0.322	0.531	0.892	0.872
p-value	0.242	0.044	0.056	0.226	0.051	0.597	0.279	0.003	0.054
95% CI	(0.47, 0.77)	(0.47, 0.77)	(0.61, 0.84)	(0.22, 0.63)	(-0.84, -0.62)	(-0.53, -0.08)	(0.32, 0.69)	(0.83, 0.93)	(0.80, 0.92)
MM Accuracy	0.600	0.433	0.714	0.481	0.238	0.400	0.500	0.750	0.533

[a]Values reported were obtained using default settings of FEP+, DS, MOE and Bioluminate as described in methods. R^2 values were found using Spearman correlations and accuracy values were found using Eq. 6.1. The MM values were averaged from final rank values of DS, MOE, and Bioluminate.

could be due the reproducibility of experimental data. For example, lysozyme has been extensively studied in the field and seen to be a robust experimental model. In this case, FEP+ performed much better with an R^2 of 0.909 and 100% accuracy[172,233]. Furthermore, the MM methods performed the best on this protein with an average R^2 of 0.554 and an accuracy of 75.0% (Table 6.3). Again, the correlation between ΔG and ΔT_m is not linear and depends on the ΔC_p of the folded and unfolded form of the protein. The ΔC_p is highly dependent on the overall hydrophobicity of the protein and could be responsible for some of the discrepancy seen with each target.

Apart from the type of target, the types of mutations made could also account for some of the failures of FEP+. Variables such as type of amino acid, molecular weight (MW), hydropathy, number of rotatable bonds, and solvent accessible surface area (SASA) of the mutation were examined (Figure 6.2). Two potential problems were found: the amino acid composition and the change in solvent accessible surface area (SASA). Arguably, the more important failures from the population are the false negatives (including the false neutrals that are experimentally stabilizing) as these represent useful mutations that would be missed by FEP+. When looking at this particular subset, the mutation changes were examined by amino acid sequence and physical chemical properties. One potentially significant observation, given the caveat of the small sample size, is that 9 of the 10 false negatives by FEP+ had a side chain with a polar functional group; a hydroxyl (Ser, Thr, Tyr), a sulfhydryl (Cys) or an imidazole (His). As seen in Figure 6.2a, these amino acids (Tyr, Thr, Cys, and His) are overrepresented in false negatives as compared to the total population, with Ser being a noted outlier. To support this observed trend of changes in polarity being problematic, we examined the overall change in solvent accessible surface area (SASA) for each mutation in the set. Seen in Figure 6.2b, the largest discrepancy between the

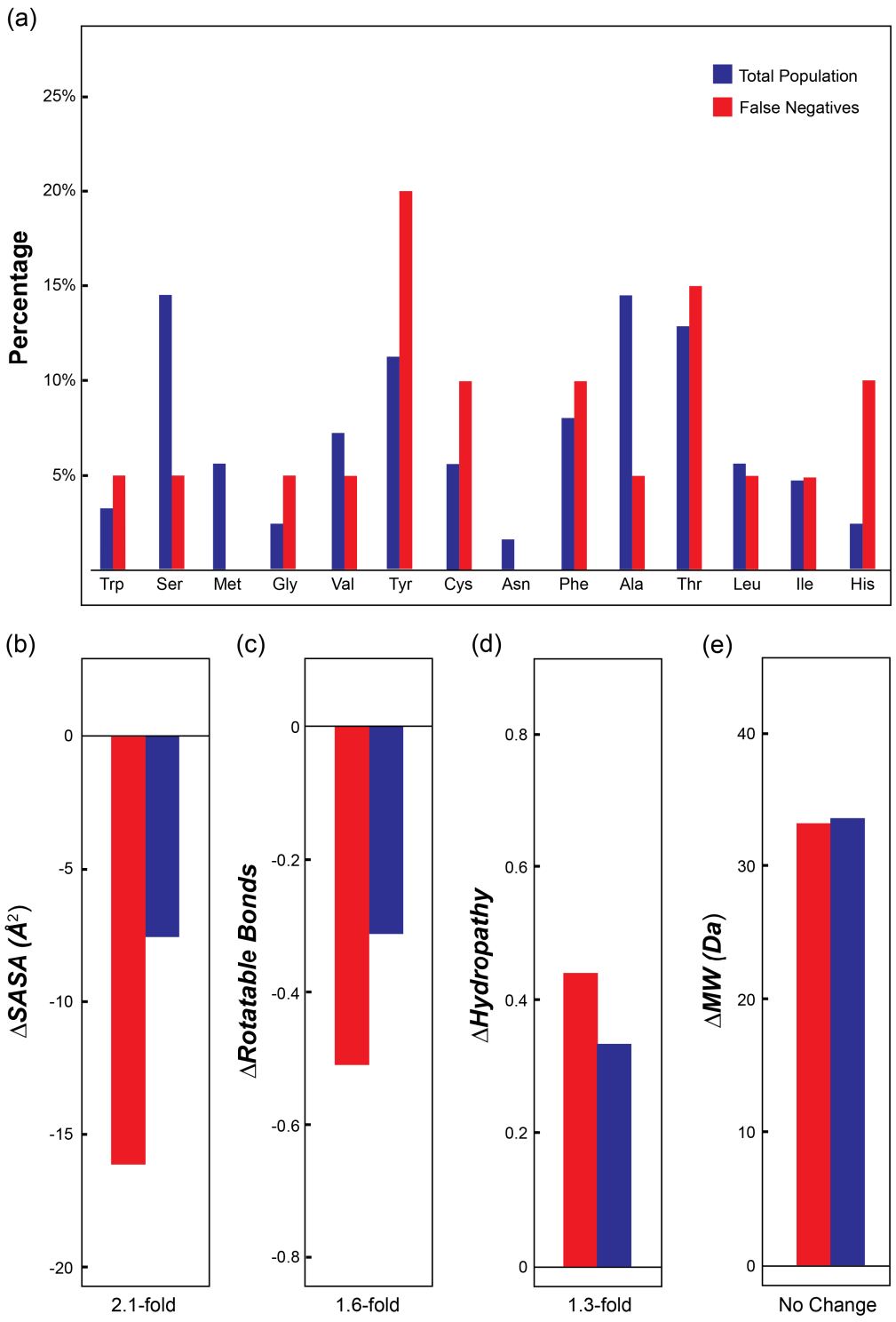


Figure 6.2. Comparison of False Negatives to Total Population. (a) Type of amino acid was compared as percentage of total population (blue) and percentage of false negatives (red). Changes in physical properties: (b) SASA, (c) number of rotatable bonds, (d) hydropathy index[234], and (e) MW were compared. These changes are the average change between the mutation and wildtype for the total population (blue) and false negatives (red).

false negatives and the total population was SASA varying by more than 2-fold from the total population. This is in stark contrast to the other physical properties: rotatable bonds, hydrophobicity and MW, which varied from the total population by 1.6-fold, 1.3-fold, and no change respectively (Figure 6.2c-e). If this change in polarity is a legitimate issue than a higher representation of the canonical uncharged polar residues Gln and Asn would be expected in the false negatives, however, these residues were not highly observed in the overall data set. Also of note, 2 out of 3 total charge to nonpolar mutations were false positives. This is not more than an anecdotal observation due to the small sample statistics.

Aside from general trends, there are more specific examples of failure. In the case of Staphylococcal nuclease (PDB 1STN) when Thr41 is mutated to the hydrophobic residues Ile or Val this is experimentally stabilizing in both cases with T_m shifts of 4.2 ˚C. However, the predictions fail to recognize this as stabilizing, in fact FEP+ predicts an increase in $\Delta\Delta G$ >1.5 kcal/mol. Looking more closely at the molecular context of this residue, the side chain is located in a fairly hydrophobic pocket (Figure 6.3). Intuitively, both Ile and Val are small hydrophobic residues and thus if placed into this hydrophobic pocket should be predicted as favorable interactions. One explanation, in this case, is the starting point may be incorrectly calculated to be more stable than it actually is. Given the trend observed of $\Delta SASA$ being problematic, perhaps the method is either incorrectly calculating desolvation of the polar side chain or assigning undue weight to the potential hydrogen bond, which in the starting structure appears to have a less than ideal geometry (Figure 6.3). To answer these questions more directly requires further study beyond the scope of this work.

With a stated goal of making the refined screen more accurate, it is important to consider the resulting FEP+ could potentially be improved by increasing lambda windows, increasing

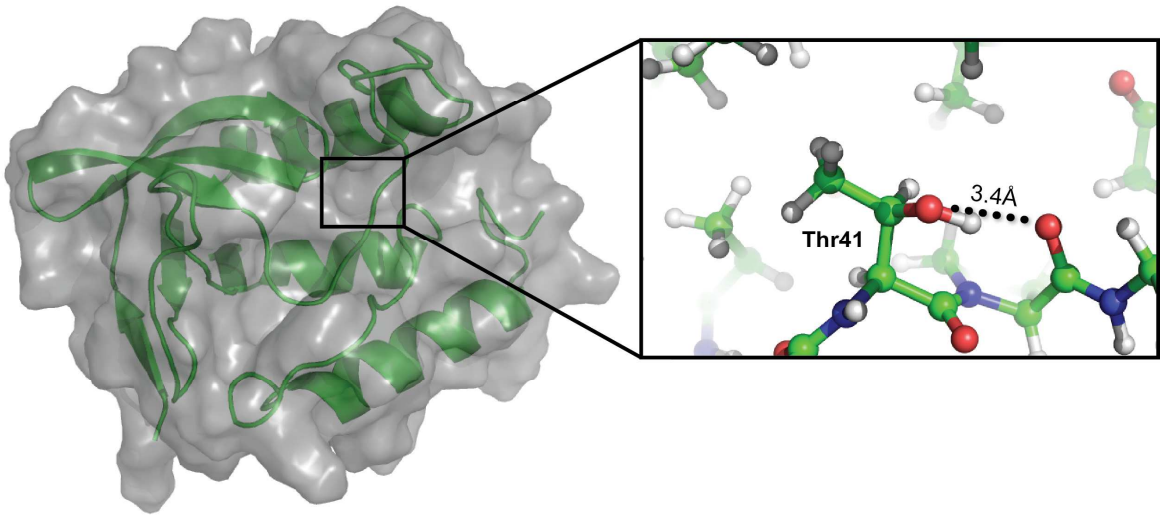


Figure 6.3. Example of False Negative from FEP+ calculations. The entire staphylococcal nuclease with surface representation is depicted on the left. The residue Thr41 is showed at closer detail in the inset with the potential H-bond interaction highlighted.

simulation time, increasing sampling at the end points or increasing initial equilibration. FEP methods have also been known to be affected by the van der Waals endpoint problem[235–237]. These problems have been heavily studied in the field[238–240] and will result in adding exponentially more simulation time. Our focus, however, is in shortening the overall computational cost whilst maintaining a level of accuracy. Additionally, this work is not an extensive analysis of how to improve FEP+, but instead aims to determine how feasible FEP+ is in its current form. Thus, the focus was turned to making either the crude screen more accurate or the refined screen faster, in essence becoming the equivalent of a more accurate crude screen. The crude screen could be improved by either altering the default settings of the MM programs or by using a faster version of the more accurate FEP itself as the crude screen. The first option is certainly viable, but would require an exhaustive study altering multiple variables in the separate software packages so attention was instead focused on the latter option. The goal was to test if a truncated FEP+ protocol could be significantly faster without sacrificing accuracy, potentially making a one step process. To make FEP+ faster, both the lambda windows and the simulation time were decreased incrementally. As seen in Table 6.4 and Figure 6.4, for our small dataset the faster FEP sacrificed little accuracy 71.0% to 67.7% for up to a 6X speed up in time. It should be noted that this could depend on the types of proteins and hardware setup making the degree of speedup variable. On this dataset, however, the computational time is sped up from more than a week to a little over 2 days to run. Additionally, using FEP+ as a filter cuts down on experimental time and resources by at least 55%.

■ 6.5 CONCLUSIONS

Overall the results here show that the ability to computationally predict protein stability with total accuracy still has a ways to go to be practically useful in protein engineering. The statistical

Table 6.4. Results of Faster FEP+[a]

No. of λ Windows	12	8	6	4
λ window length (ps)	5000	1000	750	500
Spearman R^2-value	0.526	0.434	0.476	0.472
Spearman R-value	-0.726	-0.659	-0.690	-0.687
p-value	2.54E-11	5.77E-09	5.55E-10	7.07E-10
95% CI	(-0.826,-0.581)	(-0.780,-0.490)	(-0.802,-0.532)	(-0.800,-0.528)
Accuracy	0.710	0.645	0.710	0.677
Sensitivity (TPR)	0.667	0.529	0.667	0.655
Fall-out (FPR)	0.167	0.179	0.167	0.212
Mutation Time (hrs)	5.73	2.04	1.49	0.94

[a]The number of λ windows and simulation time were varied incrementally and Spearman rank R values were reported. The mutation time reported is the total time for one mutation to run. The accuarcy, sensitivity, and fall-out were calculated using Eqs. 6.1, 6.2, and 6.3 respectively.

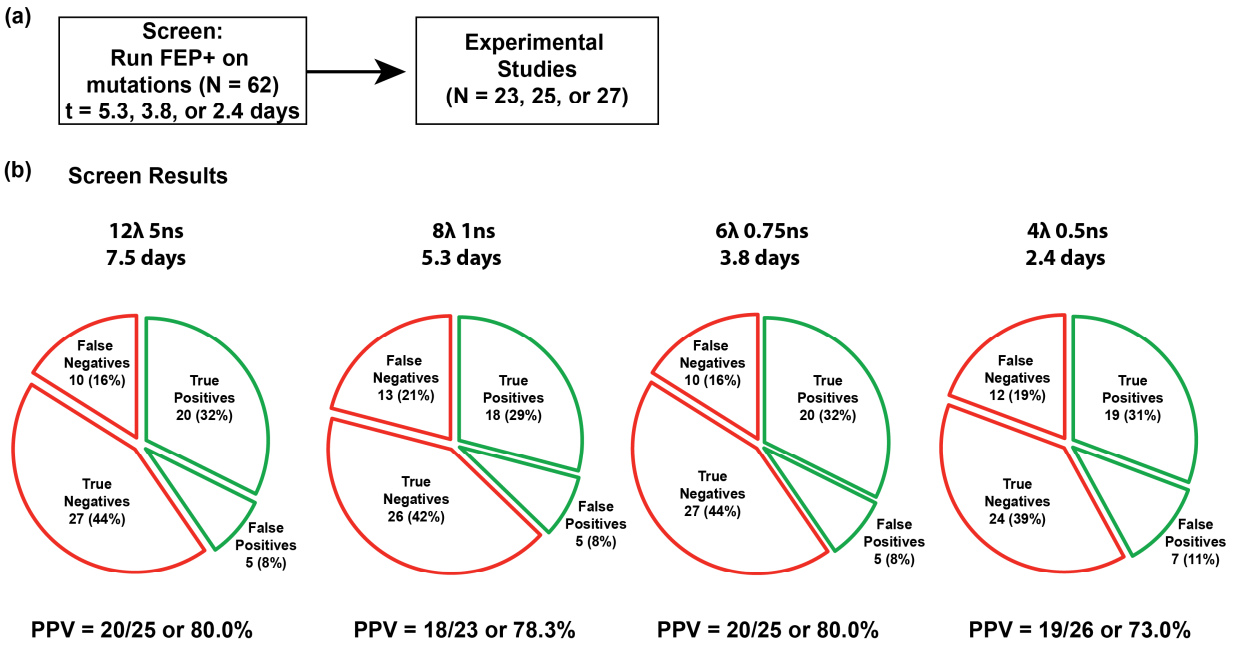


Figure 6.4. Results of FEP+ with altered settings. (a) Overview of fast FEP+ screens is shown with times and number of mutations for each step. (b) Breakdown of results from screen for varying λ windows and simulation times are depicted for each of these setups. The green slices are the mutations that would be selected for experimental testing.

correlations observed for all the methods employed here were not impressive, however when considered from a relative perspective there was a significant increase in overall performance of FEP+ when compared to the much simpler MM based methods. FEP+ showed an approximate 30% increase in accuracy, but this number is tempered by the dramatic target-by-target variation, wherein some cases show up to a 100% accuracy rate and some are essentially a coin flip. These improvements in accuracy come at a cost in speed of the calculation, in our hands slowing it down by more than 36X. How significant a 36X slowdown is however depends on the situation, timelines, and available resources. For instance, in an academic setting it may be appropriate to spend a week or more on FEP+ calculations to save time on costly experiments that need to be executed manually. On the other hand, in an industrial setting where resources are less constrained and high-throughput experimentation technologies exist, the amount of time spent on FEP+ could take much longer than the respective experiments.

In this work, we have introduced two new approaches to performing protein stability FEP+ in a more practical manner. Although FEP+ applied to protein stability has been recently explored, we sought to address two unanswered questions. First, "can you combine the faster MM methods with FEP+ into a usable workflow?" A novel consensus approach was used, where the simpler MM methods were a quick "pre-screen" and then followed by an in depth screen using FEP+ on a smaller number of potential mutations. The consensus approach maintains a very high enrichment with a final hit rate of 84.2%. This success can be very valuable to the right user, however, for some it requires too many software packages and two different methods to reproduce. This raises the next question: "could you speed up the FEP+ alone and get successful results?" After altering the default settings, we have shown that for soluble protein stability predictions, the FEP+ could be sped up by 6-fold. Under our setup, being limited to the use of 4

GPU nodes, the average time for a mutation decreased from ~6 hours to less than 1 hour. This decrease in time permits up to 168 mutations per week where with the default setting, the same number of mutations would take more than a month to run. Speeding up the calculations comes without a noticeable decrease in performance, applying the rates observed here to the hypothetical 168 mutations, the default settings would correctly identify 119 whereas switching to 4λ at 500 ps, 113 mutations would be correctly identified. This remarkable and somewhat unexpected finding merits future study. Although each user must decide if this timescale and error tolerance is feasible for their own application, the two approaches described here provide a starting point for employing FEP+ for predicting protein stability of soluble proteins.

An important question for the use of FEP is when and why does it fail. Although not definitive, this analysis suggests that changing the solvent accessible surface area of the amino acid is not always correctly captured with these methods and this should be considered when trying to predict such a change. There could be a number of explanations and potential solutions to addressing them. For instance, the desolvation of the amino acids could be improperly modeled or the parameters in the current force field, OPLS3, could be contributing to the incorrect energies seen. Some success in FEP has been found in improving protein-ligand binding by altering the number or spacing of λ windows, the simulation time or the sampling technique.[217,240,241]

With increasing understanding of protein engineering comes an increasing demand to efficiently and accurately predict protein stability. The improvement of computational programs such as FEP+ provides the field with a reliable method to predict stability. With the inclusion of OPLS3 (that includes extra parameters for halogen bonding and nitrogen atoms), GPU implementation, and REST technology, FEP+ can more readily be applied to protein stability

calculations. In this paper, we have shown how these improvements predict stability measurements in soluble proteins; however, the workflows outlined here have the potential to be applied to many systems. Although it is important to consider the amount of time and resources, FEP+ proves to be a valuable resource. Better understanding in the successes and failures is vital when continuing the improvement of an ever-rising field.

CHAPTER 7

CONCLUSION AND FUTURE PERSPECTIVE

■ 7.1 SUMMARY

The work throughout this dissertation highlights the importance of testing and improving computational models to better agree with experimental findings. Although this work has the potential to impact many fields, I turn focus to its influence on medicinal chemistry. Medicinal chemistry has been around since 3500 BC; however, it was not until the early 20th century when it began to be commercialized to today's standards. In more recent years, medicinal chemistry has begun a transition, where chemists are now incorporating computational techniques into drug discovery. The field has begun to see some initial success in this transition; however, there is still more to overcome for computers to reach their full potential in medicinal chemistry.

The start to computers becoming more successful in medicinal chemistry is to better understand the failures they have when modeling some fundamentals of biology or chemistry. The initial work of this dissertation is to characterize one of these failures, an important noncovalent interaction in medicinal chemistry–the halogen bond (X-bond). The information from these studies was then used to better model this interaction in a common simulation methodology for drug design, molecular mechanics or dynamics (MM/MD). The final important task to improving drug design software is testing current versions to better understand where failures still remain. The last work of this dissertation looks at another software, Schrödinger's Free Energy Perturbation (FEP) software package, FEP+, and determines its feasibility to being applied to protein stability predictions. With the combination of understanding fundamental

biology and chemistry and testing and improving current software, we can help reach the goal of improving computers for drug design efforts.

■ 7.2 UNDERSTANDING THE FUNDAMENTALS OF X-BONDING

Halogens have been classically incorporated into drugs to increase membrane permeability and decrease metabolic degradation[16]. More recently, halogens have been recognized to increase drug to target specificity and affinity through the X-bond[129]. Since the initial recognition of the importance of X-bonds in biological systems in 2004[43], there has been a large amount of research to better grasp this interaction. However, there are still many areas that have yet to be explored. Here, I have described how we explored the structure-energy relationships of two different types of X-bonds in biological systems that have never been shown before.

The first was an X-bond in a DNA junction system, where, more specifically, a sulfur present on the phosphate group of the DNA backbone was interacting with a bromine on a uracil base. Sulfur is an important element found in both cysteine and methionine amino acids. Cysteine is present at many active sites of enzymes, including hydrolases, deubiquitinases, caspases, and enzymes involved in redox reactions[150]. Methionine is a particularly interesting target for drug target due to its lipophilicity and flexibility and is also found at high percentages in the binding site of enzymes relative to the enzyme's surface[149]. Due to the fact that many of these enzymes are key targets for drug design, better understanding of sulfur as an X-bond acceptor is imperative. There was a large dearth in the field of X-bonding–understanding sulfur's role as an acceptor in biological systems. Although there has been some initial work done, this work focuses mostly on small molecule studies[136] and quantum mechanics (QM)[131] calculations with only a few studies done in biological systems[135,138]. Here, I explain how we use our DNA junction system to further explore sulfur's role as an X-bond acceptor in a biological molecule.

In this study, the crystallographic findings showed that the bromine was interacting with the sulfur in two different conformations. When comparing the geometries of this interaction to our previously solved DNA junction system with an oxygen acceptor, the sulfur acceptor was at a less linear angle of approach to the halogen and further distance in relation to the sum of the interacting atoms van der Waals radii. Surprisingly, the energies obtained from calorimetry studies indicated the sulfur to be a more favorable interaction with a free energy of stabilization ($\Delta\Delta G$) of -0.4 kcal/mol as compared to the previously studied negative oxygen acceptor. These findings show that sulfur is a favorable acceptor for X-bonding at a broader range of geometries. Now, using the knowledge found from this study, medicinal chemists can have an attractive target to keep in mind when designing inhibitors.

The next experimental system that was used to study the structure-energy relationship of X-bonding was an X-bond site-specifically engineered in a protein system, T4 lysozyme. Lysozyme is a heavily studied protein (as it was first enzyme to ever be crystallized[242]) providing a robust experimental system for this study. Although this system is heavily studied, the work here is one of the first time X-bonds are site-specifically engineered in a protein system with the aim to stabilize the protein. Protein stabilization can be very useful for the target identification stage of drug design. It has been found that the more stable the protein is, the more likely it is to crystallize and thus, obtain a 3-D structure of the target[200]. This study can also elucidate information about the fundamentals of another type of X-bond. In this case, a bromine or iodine was designed to form an X-bond with a formally neutral oxygen acceptor. Although this type of interaction is more studied in this field, its detailed structure-energy relationship in of a halogenated amino acid can help us design and refine our computational models.

The setup of the T4 lysozyme system, more specifically, replaced a hydroxyl group on tyrosine 18 with either a bromine or iodine in order to interact with a carbonyl oxygen in close proximity. As a control, a non-interacting site, tyrosine 88, was identified and used to compare the effects of halogenation without an X-bond forming. At site 18, the crystallography results indicate that the phenyl ring is moved to interact with the carbonyl oxygen when either a bromine or iodine is present as compared to a tyrosine, phenylalanine or methylphenylalanine alone. At site 88, the phenyl ring remains unperturbed suggesting no interaction is forming at that site. Calorimetry results showed that although modifying the tyrosine to either halogenated phenylalanine at site 18 destabilized lysozyme, they were less destabilizing modifications than when present in the control site 88. These results suggest that a favorable X-bond forming at site 18 partially rescues the overall destabilizing effect of halogenation.

Additionally, an important finding of this study is that the solvent structure was altered at site 18 causing change in stability. QM calculations on the interacting residues could only recapitulate the stabilization energy when the water structure was considered. Often times the solvent structure is not always considered and when perturbed, can have a large detrimental effect on the overall stability. So I must emphasize that while not only considering important interactions occurring within the structure, we must also consider the solvent structure. In conclusion, we are able to use this structure-energy relationship to better understand X-bonds against their most common acceptor in a protein setting and in the future, use these structures to test computational models.

■ 7.3 IMPROVING A MOLECULAR MECHANICS ALGORITHM

Despite the known importance that X-bonds have in medicinal chemistry and the rising use of computers in drug design, X-bonds are still improperly modeled using most standard

programs. There have been some attempts to rectify this problem, however, these attempts remain either inaccurate or slow and require additional costly QM parameterization. Molecular mechanics/molecular dynamics (MM/MD) simulations offer a fast, fairly accurate model when understanding how inhibitors bind to their targets, which can be incredibly useful when discovering and optimizing lead compounds. Although MM/MD is more accurate than some computational models, it does not inherently model X-bonding.

Here, I have discussed a unique solution in MM/MD for X-bonding, which incorporates a modified potential energy equation that properly models the halogen's unique polarization. Incorporating this equation into a Class I MM/MD algorithm, TINKER, energies are recapitulated from our experimental DNA junction system using the MM portion of the program. Additionally, to try to eliminate the need for costly QM parameterization, different systems were used to reparameterize the modified potential equation, finding general mathematical trends that can be incorporated into the MM/MD force fields. When parameters n, v, ΔR, $R_{X,A}$, A, and $\varepsilon_{X,A}$ were fixed to constants, the overall fit was not compromised greatly. The last parameter, B, was found to have a linear relationship to the overall *RESP* charge of the halogen. For the first time, using these relationships, we have eliminated the need for costly QM parameterization of halogen atoms for Class I MM/MD simulations.

After incorporating all of these parameters into the MM/MD package (creating X-TINKER), we can use the structures found in the T4 lysozyme studies as an external test system. As an initial proof of concept, the entire protein system was simulated in a water box. We find that the constructs with halogens remain close to the crystallography findings in TINKER and X-TINKER. Thus, to test the X-bond with no external restraints, an energy minimization was performed on a mimic of the X-bond interaction at site 18 (a bromobenzene interacting with N-

methylacetamide). In this simulation, the TINKER program completely missed the X-bonding potential, placing the donor parallel with the acceptor. X-TINKER, however, faithfully repeated the X-bonding interaction seen in the crystal structure.

This work provides the field with a new model for X-bonding in MM/MD methodology. Additionally, with the elimination of the additional QM steps, we have provided the first general model of X-bonding for multiple donors and oxygen acceptors. Using this program, we can accurately model this important, prevalent interaction present in medicinal chemistry.

■ 7.4 TESTING FREE ENERGY PERTURBATION SOFTWARE

While understanding the basics of molecular interactions and incorporating that knowledge into current programs used for drug design is important, in order to understand the knowledge of where these programs fail we must first test the programs. In this portion of my dissertation, I examined how feasible a new free energy perturbation (FEP) program, FEP+, is for predicting protein stability. FEP+ has recently shown to be highly effective when predicting protein-ligand interactions[19,243]; however, it has only undergone recent testing for protein stability[223,224]. We offer additional testing comparing FEP+ to other molecular mechanics (MM) programs and provide two novel, faster methods for applying FEP+ to protein stability.

In this study, we find that FEP+ is more successful than some current MM programs at predicting stability. We present two protocols that can be employed, one of which was found to speed up the calculations by 36X than the default, standard protocol. We also found that there were some places where FEP+ failed to properly predict protein stability. By using the knowledge of these failures, programmers can start to make improvements on FEP+. This important study gives users an understanding of the best way to use FEP+ as well as gives developers an idea of what still needs to be upgraded in regards to protein stability predictions.

■ 7.5 CONCLUSIONS AND FUTURE OUTLOOK

The work in this dissertation highlights an important interaction that has been improperly modeled in most programs–the X-bond. We characterize X-bonds in two important biological systems to help inform medicinal chemists in future design efforts. Although this work and previous work are a great beginning to better understanding this interaction, there are still many X-bonding interactions and X-bonding fundamentals that remain to be characterized in biological systems. One such poorly understood area of halogen bonds is that these polarizable atoms are mainly hydrophobic.[49] This phenomenon is in stark contrast to the analogous H-bond and causes some of the differences of the two in desolvation events in protein-ligand binding. It is important to consider the solvent structure and how X-bonds may affect it. Tools like X-TINKER allow us to study and possibly better understand the contribution of entropy, both the conformational entropy and the solvent entropy associated with the hydrophobic effect. However, a full MD set of tests would need to be performed on X-TINKER prior to this application. A more accurate way to look at the entropic effects of halogenation would be to apply the modified force field and MD simulations to FEP studies.

Using the knowledge gained in the first part of this dissertation as a better platform, we can begin to build more adequate computational algorithms for drug discovery. For example, by using some of the information found in characterizing these X-bonds, we are able to understand and improve upon failures in current MM/MD algorithms for X-bonding. For example, the solvent structure that was observed to be important in X-ray structures of T4 lysozyme studies can be used to determine if the solvent is properly modeled in X-TINKER. We might find that different solvent models may perform better than others and thus it is important to continuing

testing all aspects of this program. Additionally, it is important to continue testing and improving X-TINKER for an even larger range of X-bonding donors and acceptors.

By testing programs, we are able to discover the methodology where they perform best as well as where they still fail. In the last part of this dissertation, I examined the feasibility of using FEP+ in protein stability predictions. Although we discovered that FEP+ is the best commercially available method for predicting protein stability, we found some potential failures still exist. This realization can now be used to study and improve the software further.

In this dissertation, I investigate the field of computational medicinal chemistry–where the current field is, how it is important to understand the fundamentals of biology and chemistry to make improvements to the field, and what still remains to be studied, especially as it pertains to important chemical interactions such as X-bonds. There is a recurring theme throughout this dissertation that we must first understand the basics of biology and chemistry and leverage this knowledge to test and improve upon our current computational algorithms. Although this field has incredible potential, we must continue to test the current programs, identify failures, and learn how to overcome them–only then will we strengthen the role of computers in medicinal chemistry.

REFERENCES

(1) Kinch, M. S.; Haynesworth, A.; Kinch, S. L.; Hoyer, D. An Overview of FDA-Approved New Molecular Entities: 1827-2013. *Drug Discov. Today* **2014**, *19* (8), 1033–1039.

(2) Food & Drug Administration. U.S. Food & Drug. *https://www.fda.gov/Drugs/default.htm*.

(3) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D. T.; Shoichet, B. K. Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. *J. Med. Chem.* **2002**, *45* (11), 2213–2221.

(4) Bollini, M.; Domaoal, R. a; Thakur, V. V; Gallardo-Macias, R.; Spasov, K. a; Anderson, K. S.; Jorgensen, W. L. Computationally-Guided Optimization of a Docking Hit to Yield Catechol Diethers as Potent Anti-HIV Agents. *J. Med. Chem.* **2011**, *54* (24), 8582–8591.

(5) Friesner, R. a.; Banks, J. L.; Murphy, R. B.; Halgren, T. a.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.

(6) Friesner, R. a.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. a.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49* (21), 6177–6196.

(7) Nichols, S. E.; Domaoal, R. A.; Thakur, V. V.; Tirado-Rives, J.; Anderson, K. S.; Jorgensen, W. L. Discovery of Wild-Type and Y181C Mutant Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors Using Virtual Screening with Multiple Protein

Structures. *J. Chem. Inf. Model.* **2009**, *49* (5), 1272–1279.

(8) Decherchi, S.; Berteotti, A.; Bottegoni, G.; Rocchia, W.; Cavalli, A. The Ligand Binding Mechanism to Purine Nucleoside Phosphorylase Elucidated via Molecular Dynamics and Machine Learning. *Nat. Commun.* **2015**, *6*, 6155.

(9) Ho, M.-C.; Shi, W.; Rinaldo-Matthis, A.; Tyler, P. C.; Evans, G. B.; Clinch, K.; Almo, S. C.; Schramm, V. L. Four Generations of Transition-State Analogues for Human Purine Nucleoside Phosphorylase. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (11), 4805–4812.

(10) Desiraju, G. R.; Ho, P. S.; Kloo, L.; Legon, A. C.; Marquardt, R.; Metrangolo, P.; Politzer, P.; Resnati, G.; Rissanen, K. Definition of the Halogen Bond (IUPAC Recommendations 2013). *Pure Appl. Chem.* **2013**, *85* (8), 1711–1713.

(11) Rosokha, S.; Neretin, I.; Rosokha, T.; Hecht, J.; Kochi, J. Charge-Transfer Character of Halogen Bonding: Molecular Structures and Electronic Spectroscopy of Carbon Tetrabromide and Bromoform Complexes with Organic σ- and π-Donors. *Heteroat. Chem.* **2006**, *17* (5), 449–459.

(12) Eskandari, K.; Zariny, H. Halogen Bonding: A Lump-Hole Interaction. *Chem. Phys. Lett.* **2010**, *492* (1), 9–13.

(13) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen Bonding: The Sigma-Hole. *J. Mol. Model.* **2007**, *13*, 291–296.

(14) Voth, A. R.; Hays, F. a; Ho, P. S. Directing Macromolecular Conformation through Halogen Bonds. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (15), 6188–6193.

(15) Carter, M.; Ho, P. S. Assaying the Energies of Biological Halogen Bonds. *Cryst. Growth Des.* **2011**, *11* (11), 5087–5095.

(16) Lu, Y.; Wang, Y.; Zhu, W. Nonbonding Interactions of Organic Halogens in Biological

Systems: Implications for Drug Discovery and Biomolecular Design. *Phys. Chem. Chem. Phys.* **2010**, *12* (18), 4543–4551.

(17) Carter, M.; Rappe, A. K.; Ho, P. S. Scalable Anisotropic Shape and Electrostatic Models for Biological Bromine Halogen Bonds. *J. Chem. Theory Comput.* **2012**, *8*, 2461–2473.

(18) Scholfield, M. R.; Ford, M. C.; Vander Zanden, C. M.; Billman, M. M.; Ho, P. S.; Rappé, A. K. Force Field Model of Periodic Trends in Biomolecular Halogen Bonds. *J. Phys. Chem. B* **2015**, *119* (29), 9140–9149.

(19) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137* (7), 2695–2703.

(20) Steinbrecher, T.; Zhu, C.; Wang, L.; Abel, R.; Negron, C.; Pearlman, D.; Feyfant, E.; Duan, J.; Sherman, W. Predicting the Effect of Amino Acid Single Point Mutations on Protein Stability – Large Scale Validation of MD-Based Relative Free Energy Calculations. *J. Mol. Biol.* **2016**, 1–10.

(21) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angew. Chemie - Int. Ed.* **2016**.

(22) *PAREXEL's Pharmaceutical R&D Statistical Sourcebook*; PAREXEL International Corp: University of Chicago, 2001.

(23) Xu, Z.; Yang, Z.; Liu, Y.; Lu, Y.; Chen, K.; Zhu, W. Halogen Bond: Its Role beyond Drug-Target Binding Affinity for Drug Discovery and Development. *J. Chem. Inf. Model.* **2014**, *54*, 69–78.

(24) Wang, C.; Danovich, D.; Mo, Y.; Shaik, S. On the Nature of the Halogen Bond. *J. Chem. Theory Comput.* **2014**, *10* (9), 3726–3737.

(25) Li, W.; Zeng, Y.; Zhang, X.; Zheng, S.; Meng, L. The Enhancing Effects of Group V Sigma-Hole Interactions on the F-O Halogen Bond. *Phys. Chem. Chem. Phys.* **2014**, *16* (1), 19282–19289.

(26) Metrangolo, P.; Murray, J. S.; Pilati, T.; Politzer, P.; Resnati, G.; Terraneo, G. The Fluorine Atom as a Halogen Bond Donor, Viz. a Positive Site. *CrystEngComm* **2011**, *13* (22), 6593–6596.

(27) Hassel, O. Structural Aspects of Interatomic Charge-Transfer Bonding. *Science (80-.).* **1970**, *170* (3957), 497–502.

(28) Mulliken, R. S. Overlap Integrals and Chemical Binding. *J. Am. Chem. Soc.* **1950**, *72* (10), 4493–4503.

(29) Mulliken, R. S.; Person, W. B. Molecular Compounds and Their Spectra. XXI. Some General Considerations. *J. Am. Chem. Soc.* **1969**, *91* (13), 3409–3413.

(30) Benesi, H. a; Hildebrand, J. H. A Spectrophotometric Investigation of the Interaction of Iodine with Aromatic Hydrocarbons. *J. Am. Chem. Soc.* **1949**, *71* (8), 2703–2707.

(31) Murray, J. S.; Lane, P.; Politzer, P. Expansion of the Sigma-Hole Concept. *J. Mol. Model.* **2009**, *15*, 723–729.

(32) Taylor, P. G.; Bassindale, A. R.; El Aziz, Y.; Pourny, M.; Hursthouse, M. B.; Coles, S. J. Further Studies of Fluoride Ion Entrapment in Octasilsesquioxane Cages; X-Ray Crystal

Structure Studies and Factors That Affect Their Formation. *Dalt. Trans.* **2012**, *41* (7), 2048–2059.

(33) Bauzá, A.; Mooibroek, T. J.; Frontera, A. Tetrel-Bonding Interaction: Rediscovered Supramolecular Force? *Angew. Chemie - Int. Ed.* **2013**, *52* (47), 12317–12321.

(34) Murray, J. S.; Lane, P.; Politzer, P. A Predicted New Type of Directional Noncovalent Interaction. *Int. J. Quantum Chem.* **2007**, *107* (12), 2286–2292.

(35) Eskandari, K.; Mahmoodabadi, N. Pnicogen Bonds: A Theoretical Study Based on the Laplacian of Electron Density. *J. Phys. Chem. A* **2013**, *117* (48), 13018–13024.

(36) Bui, T. T. T.; Dahaoui, S.; Lecomte, C.; Desiraju, G. R.; Espinosa, E. The Nature of Halogen-Halogen Interactions: A Model Derived from Experimental Charge-Density Analysis. *Angew. Chemie Int. Ed.* **2009**, *48* (21), 3838–3841.

(37) Muzet, N.; Guillot, B.; Jelsch, C.; Howard, E.; Lecomte, C. Electrostatic Complementarity in an Aldose Reductase Complex from Ultra-High-Resolution Crystallography and First-Principles Calculations. *Proc. Natl. Acad. Sci. USA* **2003**, *100* (15), 8742–8747.

(38) Scholfield, M. R.; Zanden, C. M. Vander; Carter, M.; Ho, P. S. Halogen Bonding (X-Bonding): A Biological Perspective. *Protein Sci.* **2013**, *22* (2), 139–152.

(39) Wang, Y.-H.; Zou, J.; Lu, Y.; Yu, Q.; Xu, H.-Y. Single-Electron Halogen Bond: Ab Initio Study. *Int. J. Quantum Chem.* **2007**, *107* (2), 501–506.

(40) Voth, A. R.; Khuu, P.; Oishi, K.; Ho, P. S. Halogen Bonds as Orthogonal Molecular Interactions to Hydrogen Bonds. *Nat. Chem.* **2009**, *1* (1), 74–79.

(41) Nelyubina, Y. V; Antipin, M. Y.; Dunin, D. S.; Kotov, V. Y.; Lyssenko, K. a. Unexpected "amphoteric" character of the Halogen Bond: The Charge Density Study of the Co-Crystal of N-Methylpyrazine Iodide with I2. *Chem. Commun.* **2010**, *46* (29), 5325–5327.

(42) Zhou, P.; Lv, J.; Zou, J.; Tian, F.; Shang, Z. Halogen-Water-Hydrogen Bridges in Biomolecules. *J. Struct. Biol.* **2010**, *169* (2), 172–182.

(43) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. Halogen Bonds in Biological Molecules. *Proc. Natl. Acad. Sci. USA* **2004**, *101* (48), 16789–16794.

(44) Alkorta, I.; Blanco, F.; Solimannejad, M.; Elguero, J. Competition of Hydrogen Bonds and Halogen Bonds in Complexes of Hypohalous Acids with Nitrogenated Bases. *J. Phys. Chem. A* **2008**, *112* (43), 10856–10863.

(45) Li, Q.; Xu, X.; Liu, T.; Jing, B.; Li, W.; Cheng, J.; Gong, B.; Sun, J. Competition between Hydrogen Bond and Halogen Bond in Complexes of Formaldehyde with Hypohalous Acids. *Phys. Chem. Chem. Phys.* **2010**, *12* (25), 6837–6843.

(46) Aakeröy, C. B.; Fasulo, M.; Schultheiss, N.; Desper, J.; Moore, C. Structural Competition between Hydrogen Bonds and Halogen Bonds. *J. Am. Chem. Soc.* **2007**, *129* (45), 13772–13773.

(47) Takemura, A.; McAllister, L. J.; Hart, S.; Pridmore, N. E.; Karadakov, P. B.; Whitwood, A. C.; Bruce, D. W. Halogen- and Hydrogen-Bonded Salts and Co-Crystals Formed from 4-Halo-2,3,5,6-Tetrafluorophenol and Cyclic Secondary and Tertiary Amines: Orthogonal and Non-Orthogonal Halogen and Hydrogen Bonding, and Synthetic Analogues of Halogen-Bonded Biological Syst. *Chem. Eur. J.* **2014**, *20* (22), 6721–6732.

(48) Vasylyeva, V.; Nayak, S. K.; Terraneo, G.; Cavallo, G.; Metrangolo, P.; Resnati, G. Orthogonal Halogen and Hydrogen Bonds Involving a Peptide Bond Model. *CrystEngComm* **2014**, *16* (35), 8102–8105.

(49) Vallejos, M.; Auffinger, P.; Ho, P. S. Halogen Interactions in Biomolecular Crystal Structures. *Int. Tables Crystallogr.* **2012**, *F* (2), 821–826.

(50) Lommerse, J. P. M.; Stone, A. J.; Taylor, R.; Allen, F. H. The Nature and Geometry of Intermolecular Interactions between Halogens and Oxygen or Nitrogen. *J. Am. Chem. Soc.* **1996**, *118* (13), 3108–3116.

(51) Hardegger, L. a; Kuhn, B.; Spinnler, B.; Anselm, L.; Ecabert, R.; Stihle, M.; Gsell, B.; Thoma, R.; Diez, J.; Benz, J.; Plancher, J.-M.; Hartmann, G.; Banner, D. W.; Haap, W.; Diederich, F. Systematic Investigation of Halogen Bonding in Protein-Ligand Interactions. *Angew. Chemie Int. Ed.* **2011**, *50* (1), 314–318.

(52) Maillard, M. C.; Hom, R. K.; Benson, T. E.; Moon, J. B.; Mamo, S.; Bienkowski, M.; Tomasselli, A. G.; Woods, D. D.; Prince, D. B.; Paddock, D. J.; Emmons, T. L.; Tucker, J. A.; Dappen, M. S.; Brogley, L.; Thorsett, E. D.; Jewett, N.; Sinha, S.; John, V. Design, Synthesis, and Crystal Structure of Hydroxyethyl Secondary Amine-Based. *J. Med. Chem.* **2007**, *50*, 776–781.

(53) Parks, D. J.; Lafrance, L. V; Calvo, R. R.; Milkiewicz, K. L.; Gupta, V.; Lattanze, J.; Ramachandren, K.; Carver, T. E.; Petrella, E. C.; Cummings, M. D.; Maguire, D.; Grasberger, B. L.; Lu, T. 1,4-Benzodiazepine-2,5-Diones as Small Molecule Antagonists of the HDM2-p53 Interaction: Discovery and SAR. *Bioorg. Med. Chem. Lett.* **2005**, *15* (3), 765–770.

(54) Iltzsch, M.; Uber, S. S.; Tankersley, K.; Kounit, M. H. Structure-Activity Relationship for the Binding of Nucleoside Ligands to Adenosine Kinase from Toxoplasma Gondii. *Biochem. Pharmacol.* **1995**, *49* (10), 1501–1512.

(55) Benjahad, A.; Guillemont, J.; Andries, K.; Nguyen, C. H.; Grierson, D. S. 3-Iodo-4-Phenoxypyridinones (IOPY's), a New Family of Highly Potent Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Bioorg. Med. Chem. Lett.* **2003**, *13* (24), 4309–4312.

(56) Howard, E. I.; Sanishvili, R.; Cachau, R. E.; Mitschler, A.; Chevrier, B.; Barth, P.; Lamour, V.; Van Zandt, M.; Sibley, E.; Bon, C.; Moras, D.; Schneider, T. R.; Joachimiak, A.; Podjarny, A. Ultrahigh Resolution Drug Design I: Details of Interactions in Human Aldose Reductase-Inhibitor Complex at 0.66Å. *Proteins* **2004**, *55* (4), 792–804.

(57) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(58) Hardegger, L. a; Kuhn, B.; Spinnler, B.; Anselm, L.; Ecabert, R.; Stihle, M.; Gsell, B.; Thoma, R.; Diez, J.; Benz, J.; Plancher, J.-M.; Hartmann, G.; Isshiki, Y.; Morikami, K.; Shimma, N.; Haap, W.; Banner, D. W.; Diederich, F. Halogen Bonding at the Active Sites of Human Cathepsin L and MEK1 Kinase: Efficient Interactions in Different Environments. *ChemMedChem* **2011**, *6* (11), 2048–2054.

(59) Ho, P. S. Halogen Bonding I: Impact on Materials Chemistry and Life Sciences. In *Topics in current chemistry*; Metrangolo, P., Resnati, G., Eds.; Springer International Publishing: Switzerland, 2014; pp 241–276.

(60) Xu, Z.; Liu, Z.; Chen, T.; Chen, T.; Wang, Z.; Tian, G.; Shi, J.; Wang, X.; Lu, Y.; Yan, X.; Wang, G.; Jiang, H.; Chen, K.; Wang, S.; Xu, Y.; Shen, J.; Zhu, W. Utilization of Halogen Bond in Lead Optimization: A Case Study of Rational Design of Potent Phosphodiesterase Type 5 (PDE5) Inhibitors. *J. Med. Chem.* **2011**, *54*, 5607–5611.

(61) Prasad, S.; Wilkinson, J.; Gatzoulis, M. A. Sildenafil in Primary Pulmonary Hypertension. *N. Engl. J. Med.* **2000**, *343*, 1342.

(62) Sung, B.; Hwang, K.; Jeon, Y.; Lee, J.; Heo, Y.; Kim, J.; Moon, J.; Yoon, J.; Hyun, Y.; Kim, E.; Eum, S.; Park, S.; Lee, J.; Lee, T.; Ro, S.; Cho, J. Structure of the Catalytic

Domain of Human Phosphodiesterase 5 with Bound Drug Molecules. *Nature* **2003**, *425* (6953), 98–102.

(63) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M. The Role of Quantum Mechanics in Structure-Based Drug Design. *Drug Discov. Today* **2007**, *12* (17/18), 725–731.

(64) Dewar, M.; Thiel, W. Ground States of Molecules. 38. The MNDO Method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.

(65) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902–3909.

(66) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods IV: Extension of MNDO, AM1 and PM3 to More Main Group Elements. *J. Mol. Model.* **2004**, *10* (2), 155–164.

(67) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13* (12), 1173–1213.

(68) Stewart, J. J. P. Application of the PM6 Method to Modeling Proteins. *J. Mol. Model.* **2009**, *15*, 765–805.

(69) Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza. Stabilization and Structure Calculations for Noncovalent Interactions in Extended Molecular Systems Based on Wave Function and Density Functional Theories. *Chem. Rev.* **2010**, *110*, 5023–5063.

(70) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes

Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5* (7), 1749–1760.

(71) Jurecka, P.; Sponer, J.; Cerný, J.; Hobza, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Phys. Chem. Chem. Phys.* **2006**, *8* (17), 1985–1993.

(72) Dobeš, P.; Fanfrlík, J.; Rezáč, J.; Otyepka, M.; Hobza, P. Transferable Scoring Function Based on Semiempirical Quantum Mechanical PM6-DH2 Method: CDK2 with 15 Structurally Diverse Inhibitors. *J. Comput. Aided. Mol. Des.* **2011**, *25* (3), 223–235.

(73) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13* (4), 505–524.

(74) Fanfrlík, J.; Bronowska, A. K.; Rezác, J.; Prenosil, O.; Konvalinka, J.; Hobza, P. A Reliable Docking/scoring Scheme Based on the Semiempirical Quantum Mechanical PM6-DH2 Method Accurately Covering Dispersion and H-Bonding: HIV-1 Protease with 22 Ligands. *J. Phys. Chem. B* **2010**, *114* (39), 12666–12678.

(75) Kong, N.; Fotouhi, N.; Wovkulich, P.; Roberts, J. Cell Cycle Inhibitors for the Treatment of Cancer. *Drugs Future* **2003**, *28* (9), 881.

(76) Řezáč, J.; Hobza, P. A Halogen-Bonding Correction for the Semiempirical PM6 Method. *Chem. Phys. Lett.* **2011**, *506* (4–6), 286–289.

(77) Dobeš, P.; Řezáč, J.; Fanfrlík, J.; Otyepka, M.; Hobza, P. Semiempirical Quantum Mechanical Method PM6-DH2X Describes the Geometry and Energetics of CK2-Inhibitor Complexes Involving Halogen Bonds Well , While the Empirical Potential Fails Ez. *J. Phys. Chem. B* **2011**, *115*, 8581–8589.

(78) Voth, A. R.; Ho, P. S. The Role of Halogen Bonding in Inhibitor Recognition and Binding by Protein Kinases. *Curr. Top. Med. Chem.* **2007**, *7* (14), 1336–1348.

(79) Gianoncelli, A.; Cozza, G.; Orzeszko, A.; Meggio, F.; Kazimierczuk, Z.; Pinna, L. a. Tetraiodobenzimidazoles Are Potent Inhibitors of Protein Kinase CK2. *Bioorg. Med. Chem.* **2009**, *17* (20), 7281–7289.

(80) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and Validation of a Modular, Extensible Docking Program: DOCK 5. *J. Comput. Aided. Mol. Des.* **2006**, *20* (10–11), 601–619.

(81) Sarno, S.; Papinutto, E.; Franchin, C.; Bain, J.; Elliott, M.; Meggio, F.; Kazimierczuk, Z.; Orzeszko, A.; Zanotti, G.; Battistutta, R.; Pinna, L. ATP Site-Directed Inhibitors of Protein Kinase 2: An Update. *Curr. Top. Med. Chem.* **2011**, *11*, 1340–1351.

(82) Battistutta, R.; Mazzorana, M.; Sarno, S.; Kazimierczuk, Z.; Zanotti, G.; Pinna, L. a. Inspecting the Structure-Activity Relationship of Protein Kinase CK2 Inhibitors Derived from Tetrabromo-Benzimidazole. *Chem. Biol.* **2005**, *12* (11), 1211–1219.

(83) Battistutta, R.; Mazzorana, M.; Cendron, L.; Bortolato, A.; Sarno, S.; Kazimierczuk, Z.; Zanotti, G.; Moro, S.; Pinna, L. a. The ATP-Binding Site of Protein Kinase CK2 Holds a Positive Electrostatic Area and Conserved Water Molecules. *ChemBioChem* **2007**, *8* (15), 1804–1809.

(84) Battistutta, R.; De Moliner, E.; Sarno, S.; Zanotti, G.; Pinna, L. a. Structural Features Underlying Selective Inhibition of Protein Kinase CK2 by ATP Site-Directed Tetrabromo-2-Benzotriazole. *Protein Sci.* **2001**, *10* (11), 2200–2206.

(85) Ibrahim, M. A. A. AMBER Empirical Potential Describes the Geometry and Energy of Noncovalent Halogen Interactions Better than Advanced Semiempirical Quantum

Mechanical Method PM6-DH2X. *J. Phys. Chem. B* **2012**, *116* (11), 3659–3669.

(86) Lu, Y.; Shi, T.; Wang, Y.; Yang, H.; Yan, X.; Luo, X.; Jiang, H.; Zhu, W. Halogen Bonding--a Novel Interaction for Rational Drug Design? *J. Med. Chem.* **2009**, *52* (9), 2854–2862.

(87) Ibrahim, M. A. A. Molecular Mechanical Study of Halogen Bonding in Drug Discovery. *J. Comput. Chem.* **2011**, *32* (12), 2564–2574.

(88) Dixon, R. W.; Kollman, P. A. Advancing beyond the Atom-Centered Model in Additive and Nonadditive Molecular Mechanics. *J. Comput. Chem.* **1997**, *18* (13), 1632–1646.

(89) Cieplak, P.; Caldwell, J.; Kollman, P. Organic and Biological Systems Going Beyond the Atom Centered Two Body Additive Approximation : Aqueous Solution Free Energies of Methanol and Amide Hydrogen Bonding and Chloroform / Water Partition Coefficients of the Nucleic Acid Bases. *J. Comput. Chem.* **2001**, *22* (10), 1048–1057.

(90) Halgren, T. A.; Damm, W. Polarizable Force Fields. *Curr. Opin. Struct. Biol.* **2001**, *11* (2), 236–242.

(91) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136* (3B), 864–871.

(92) Kohn, W.; Sham, L. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A), 1133–1138.

(93) 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* (9), 1157–1174.

(94) Ibrahim, M. A. A. Performance Assessment of Semiempirical Molecular Orbital Methods in Describing Halogen Bonding: Quantum Mechanical and Quantum Mechanical/molecular Mechanical-Molecular Dynamics Study. *J. Chem. Inf. Model.*

2011, *51* (10), 2549–2559.

(95) Kolar, M.; Hobza, P. On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds. *J. Chem. Theory Comput.* **2012**, *8*, 1325–1333.

(96) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OLPS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (15), 11225–11236.

(97) Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666.

(98) Jorgensen, W. L.; Tirado-Rives, J. Potential Energy Functions for Atomic-Level Simulations of Water and Organic and Biomolecular Systems. *Proc. Natl. Acad. Sci. USA* **2005**, *102* (19), 6665–6670.

(99) Jorgensen, W. L.; Schyman, P. Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-HIV Agents. *J. Chem. Theory Comput.* **2012**, *8* (10), 3895–3901.

(100) Carter, M.; Voth, A. R.; Scholfield, M. R.; Rummel, B.; Sowers, L. C.; Ho, P. S. Enthalpy-Entropy Compensation in Biomolecular Halogen Bonds Measured in DNA Junctions. *Biochemistry* **2013**, *52* (29), 4891–4903.

(101) Wilcken, R.; Liu, X.; Zimmermann, M. O.; Rutherford, T. J.; Fersht, A. R.; Joerger, A. C.; Boeckler, F. M. Halogen-Enriched Fragment Libraries as Leads for Drug Rescue of Mutant p53. *J. Am. Chem. Soc.* **2012**, *134* (15), 6810–6818.

(102) Vousden, K. H.; Prives, C. Blinded by the Light: The Growing Complexity of p53. *Cell* **2009**, *137* (3), 413–431.

(103) Boeckler, F. M.; Joerger, A. C.; Jaggi, G.; Rutherford, T. J.; Veprintsev, D. B.; Fersht, A. R. Targeted Rescue of a Destabilized Mutant of p53 by an in Silico Screened Drug. *Proc. Natl. Acad. Sci.* **2008**, *105* (30), 10360–10365.

(104) Warren, G. L.; Andrews, C. V; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49* (20), 5912–5931.

(105) Thomas, P. D.; Dill, K. a. An Iterative Method for Extracting Energy-like Quantities from Protein Structures. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 11628–11633.

(106) Liu, Y.; Xu, Z.; Yang, Z.; Chen, K.; Zhu, W. A Knowledge-Based Halogen Bonding Scoring Function for Predicting Protein-Ligand Interactions. *J. Mol. Model.* **2013**, *19* (11), 5015–5030.

(107) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: A Novel Scoring Function for Predicting Binding Affinities. *J. Mol. Graph. Model.* **2005**, *23* (5), 395–407.

(108) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295* (2), 337–356.

(109) Jain, A. Scoring Noncovalent Protein-Ligand Interactions: A Continuous Differentiable Function Tuned to Compute Binding Affinities. *J. Comput. Aided. Mol. Des.* **1996**, *10* (5), 427–440.

(110) Xue, M.; Zheng, M.; Xiong, B.; Li, Y.; Jiang, H.; Shen, J. Knowledge-Based Scoring Functions in Drug Design. 1. Developing a Target-Specific Method for Kinase-Ligand Interactions. *J. Chem. Inf. Model.* **2010**, *50* (8), 1378–1386.

(111) Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **2006**, *49* (20), 5895–5902.

(112) Goodsell, D. S.; Lauble, H.; Stout, C.; Olson, A. J. Automated Docking in Crystallography: Analysis of the Substrates of Aconitase. *Proteins* **1993**, *17* (1), 1–10.

(113) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3* (11), 935–949.

(114) Kuhn, B.; Fuchs, J. E.; Reutlinger, M.; Stahl, M.; Taylor, N. R. Rationalizing Tight Ligand Binding Through Cooperative Interaction Networks. *J. Chem. Inf. Model.* **2011**, *51*, 3180–3198.

(115) Watts, D. J.; Strogatz, S. H. Collective Dynamics of "Small-World" Networks. *Nature* **1998**, *393*, 440–442.

(116) Biswal, B. K.; Cherney, M. M.; Wang, M.; Chan, L.; Yannopoulos, C. G.; Bilimoria, D.; Nicolas, O.; Bedard, J.; James, M. N. G. Crystal Structures of the RNA-Dependent RNA Polymerase Genotype 2a of Hepatitis C Virus Reveal Two Conformations and Suggest Mechanisms of Inhibition by Non-Nucleoside Inhibitors. *J. Biol. Chem.* **2005**, *280* (18), 18202–18210.

(117) Zimmermann, M. O.; Lange, A.; Boeckler, F. M. Evaluating the Potential of Halogen Bonding in Molecular Design: Automated Scaffold Decoration Using an XB-Scoring Function. *J. Chem. Inf. Model.* **2015**, *55* (3), 687–699.

(118) McLean, I.; Schwab, J.; Hillegas, A.; Schlingman, A. Susceptibility of Micro-Organisms to Chloramphenicol. *J. Clin. Invest.* **1949**, *22*, 953–963.

(119) Wang, L.; Ren, D. Flavopiridol: The First Cyclin-Dependent Kinase Inhibitor in Human Clinical Trials. *Mini-Reviews Med. Chem.* **2010**, *10* (11), 1058–1070.

(120) Gerebtzoff, G.; Li-Blatter, X.; Fischer, H.; Frentzel, A.; Seelig, A. Halogenation of Drugs Enhances Membrane Binding and Permeation. *ChemBioChem* **2004**, *5* (5), 676–684.

(121) Lam, P.; Clark, C.; Smallwood, A.; Alexander, R. Structure-Based Drug Design Utilizing Halogenbonding: Factor Xa Inhibitors. In *The 238th ACS National Meeting*; 2009; p 58.

(122) Metrangolo, P.; Resnati, G.; Pilati, T.; Liantonio, R.; Meyer, F. Engineering Functional Materials by Halogen Bonding. *J. Polym. Sci. Part A Polym. Chem.* **2007**, *45*, 1–15.

(123) Ding, X.; Tuikka, M.; Haukka, M. Halogen Bonding in Crystal Engineering. In *Recent Advances in Crystallography*; Benedict, J. B., Ed.; InTech, 2012.

(124) Scholfield, M. R.; Zanden, C. M. Vander; Carter, M.; Ho, P. S. Halogen Bonding (X-Bonding): A Biological Perspective. *Protein Sci.* **2013**, *22* (2), 139–152.

(125) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Joerger, A. C.; Boeckler, F. M. Principles and Applications of Halogen Bonding in Medicinal Chemistry and Chemical Biology. *J. Med. Chem.* **2013**, *56* (4), 1363–1388.

(126) DeWeese, A. D.; Schultz, T. W. Structure-Activity Relationships for Aquatic Toxicity to Tetrahymena: Halogen-Substituted Aliphatic Esters. *Environ. Toxicol.* **2001**, *16* (1), 54–60.

(127) Kortagere, S.; Ekins, S.; Welsh, W. J. Halogenated Ligands and Their Interactions with Amino Acids: Implications for Structure-Activity and Structure-Toxicity Relationships. *J. Mol. Graph. Model.* **2008**, *27* (2), 170–177.

(128) Hernandes, M. Z.; Cavalcanti, S. M. T.; Moreira, D. R. M.; Filgueira, W.; Junior, D. A.; Cristina, A.; Leite, L. Halogen Atoms in the Modern Medicinal Chemistry : Hints for the Drug Design. *Curr. Drug Targets* **2010**.

(129) Ford, M. C.; Ho, P. S. Computational Tools to Model Halogen Bonds in Medicinal

Chemistry. *J. Med. Chem.* **2015**, 151014111649009.

(130) Sirimulla, S.; Bailey, J. B.; Vegesna, R.; Narayan, M. Halogen Interactions in Protein-Ligand Complexes: Implications of Halogen Bonding for Rational Drug Design. *J. Chem. Inf. Model.* **2013**, *53* (11), 2781–2791.

(131) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Kirchner, B.; Boeckler, F. M. Addressing Methionine in Molecular Design through Directed Sulfur-Halogen Bonds. *J. Chem. Theory Comput.* **2011**, *7* (7), 2307–2315.

(132) Murray, J. S.; Lane, P.; Clark, T.; Riley, K. E.; Politzer, P. Σ-Holes, Π-Holes and Electrostatically-Driven Interactions. *J. Mol. Model.* **2012**, *18* (2), 541–548.

(133) Zhang, Q.; Xu, Z.; Zhu, W. The Underestimated Halogen Bonds Forming with Protein Side Chains in Drug Discovery and Design. *J. Chem. Inf. Model.* **2017**, *57*, 22–26.

(134) Metrangolo, P.; Resnati, G. *Halogen Bonding*; 2008; Vol. 126.

(135) Lange, A.; Günther, M.; Büttner, F. M.; Zimmermann, M. O.; Heidrich, J.; Hennig, S.; Zahn, S.; Schall, C.; Sievers-Engler, A.; Ansideri, F.; Koch, P.; Laemmerhofer, M.; Stehle, T.; Laufer, S. A.; Boeckler, F. M. Targeting the Gatekeeper MET146 of C-Jun N-Terminal Kinase 3 Induces a Bivalent Halogen/Chalcogen Bond. *J. Am. Chem. Soc.* **2015**, *137* (46), 14640–14652.

(136) Hauchecorne, D.; Moiana, A.; van der Veken, B. J.; Herrebout, W. A. Halogen Bonding to a Divalent Sulfur Atom: An Experimental Study of the Interactions of CF3X (X = Cl, Br, I) with Dimethyl Sulfide. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10204–10213.

(137) Lu, Y. X.; Zou, J. W.; Wang, Y. H.; Jiang, Y. J.; Yu, Q. Sen. Ab Initio Investigation of the Complexes between Bromobenzene and Several Electron Donors: Some Insights into the Magnitude and Nature of Halogen Bonding Interactions. *J. Phys. Chem. A* **2007**, *111* (42),

10781–10788.

(138) Liu, L.; Baase, W. A.; Matthews, B. W. Halogenated Benzenes Bound within a Non-Polar Cavity in T4 Lysozyme Provide Examples of I···S and I···Se Halogen-Bonding. *J. Mol. Biol.* **2009**, *385* (2), 595–605.

(139) Vander Zanden, C. M.; Carter, M.; Ho, P. S. Determining Thermodynamic Properties of Molecular Interactions from Single Crystal Studies. *Methods* **2013**, *64* (1), 12–18.

(140) Liu, J.; Déclais, A. C.; McKinney, S. a.; Ha, T.; Norman, D. G.; Lilley, D. M. J. Stereospecific Effects Determine the Structure of a Four-Way DNA Junction. *Chem. Biol.* **2005**, *12* (2), 217–228.

(141) Otwinowski, Z.; Minor, W. *Processing of X-Ray Diffraction Data Collected in Oscillation Mode*; Elsevier: Amsterdam, 1997.

(142) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: A Comprehensive Python-Based System for Macromolecular Structure Solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 213–221.

(143) Lavery, R.; Moakher, M.; Maddocks, J. H.; Petkeviciute, D.; Zakrzewska, K. Conformational Analysis of Nucleic Acids Revisited: Curves+. *Nucleic Acids Res.* **2009**, *37* (17), 5917–5929.

(144) Watson, J.; Hays, F. A.; Ho, P. S. Definitions and Analysis of DNA Holliday Junction Geometry. *Nucleic Acids Res.* **2004**, *32* (10), 3017–3027.

(145) M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li,

H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Had, and D. J. F. Gaussian 09. Gaussian, Inc.: Wallingford, CT 2009.

(146) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. Basis Set Exchange: A Community Database for Computational Sciences. *J. Chem. Inf. Model.* **2007**, *47* (3), 1045–1052.

(147) Boys, S.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *100* (1), 65–73.

(148) Hays, F. A.; Teegarden, A.; Jones, Z. J. R.; Harms, M.; Raup, D.; Watson, J.; Cavaliere, E.; Ho, P. S. How Sequence Defines Structure: A Crystallographic Map of DNA Structure and Conformation. *Proc. Natl. Acad. Sci. USA* **2005**, *107* (9), 4486–4486.

(149) Khazanov, N. A.; Carlson, H. A. Exploring the Composition of Protein-Ligand Binding Sites on a Large Scale. *PLoS Comput. Biol.* **2013**, *9* (11).

(150) Hallenbeck, K.; Turner, D.; Renslo, A.; Arkin, M. Targeting Non-Catalytic Cysteine Residues Through Structure- Guided Drug Discovery. *Curr. Top. Med. Chem.* **2017**, *17* (1), 4–15.

(151) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084.

(152) Priimagi, A.; Cavallo, G.; Metrangolo, P.; Resnati, G. The Halogen Bond in the Design of Functional Supramolecular Materials: Recent Advances. *Acc. Chem. Res.* **2013**, *46* (11), 2686–2695.

(153) Brammer, L.; Bruton, E. A.; Sherwood, P. Understanding the Behavior of Halogens as Hydrogen Bond Acceptors. *Cryst. Growth Des.* **2001**, *1* (4), 277–290.

(154) Aakeröy, C. B.; Desper, J.; Helfrich, B. A.; Metrangolo, P.; Pilati, T.; Resnati, G.; Stevenazzi, A. Combining Halogen Bonds and Hydrogen Bonds in the Modular Assembly of Heteromeric Infinite 1-D Chains. *Chem. Commun* **2007**, No. 41, 4236–4238.

(155) Wilcken, R.; Zimmermann, M. O.; Lange, A.; Zahn, S.; Boeckler, F. M. Using Halogen Bonds to Address the Protein Backbone: A Systematic Evaluation. *J. Comput. Aided. Mol. Des.* **2012**, *26* (8), 935–945.

(156) Riley, K. E.; Murray, J. S.; Fanfrlík, J.; Řezáč, J.; Solá, R. J.; Concha, M. C.; Ramos, F. M.; Politzer, P. Halogen Bond Tunability I: The Effects of Aromatic Fluorine Substitution on the Strengths of Halogen-Bonding Interactions Involving Chlorine, Bromine, and Iodine. *J. Mol. Model.* **2011**, *17* (12), 3309–3318.

(157) Matthews, B. W. Structural and Genetic Analysis of the Folding and Function of T4 Lysozyme. *FASEB J.* **1996**, *10* (1), 35–41.

(158) Carra, J. H.; Murphy, E. C.; Privalov, P. L. Thermodynamic Effects of Mutations on the Denaturation of T4 Lysozyme. *Biophys. J.* **1996**, *71* (4), 1994–2001.

(159) Xie, J.; Wang, L.; Wu, N.; Brock, A.; Spraggon, G.; Schultz, P. G. The Site-Specific Incorporation of P-Iodo-L-Phenylalanine into Proteins for Structure Determination. *Nat. Biotechnol.* **2004**, *22* (10), 1297–1301.

(160) Matsumura, M.; Matthews, B. W. Control of Enzyme Activity by an Engineered Disulfide Bond. *Science* **1989**, *243* (4892), 792–794.

(161) Peeler, J. C.; Mehl, R. A. Site-Specific Incorporation of Unnatural Amino Acids as Probes for Protein Conformational Changes. *Methods Mol. Biol.* **2012**, *794*, 125–134.

(162) Eriksson, A.; Baase, W. A.; Matthews, B. W. Similar Hydrophobic Replacements of Leu99 and Phe153 within the Core of T4 Lysozyme Have Different Structural and

Thermodynamic Consequences. *J Mol Biol* **1993**, *229*, 747–769.

(163) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the CCP4 Suite and Current Developments. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2011**, *67* (4), 235–242.

(164) Nicholson, H.; Anderson, D. E.; Dao Pin, S.; Matthews, B. W. Analysis of the Interaction between Charged Side Chains and the α-Helix Dipole Using Designed Thermostable Mutants of Phage T4 Lysozyme. *Biochemistry* **1991**, *30* (41), 9816–9828.

(165) Heinz, D. W.; Baase, W. a; Matthews, B. W. Folding and Function of a T4 Lysozyme Containing 10 Consecutive Alanines Illustrate the Redundancy of Information in an Amino Acid Sequence. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (9), 3751–3755.

(166) Peterson, K. A.; Shepler, B. C.; Figgen, D.; Stoll, H. On the Spectroscopic and Thermochemical Properties of ClO, BrO, IO, and Their Anions. *J. Phys. Chem. A* **2006**, *110* (51), 13877–13883.

(167) Simon, S.; Duran, M.; Dannenberg, J. J. How Does Basis Set Superposition Error Change the Potential Surfaces for Hydrogen-Bonded Dimers? *J. Chem. Phys.* **1996**, *105* (24), 11024–11031.

(168) Kuroki, R.; Weaver, L. H.; Matthews, B. W. Structure-Based Design of a Lysozyme with Altered Catalytic Activity. *Nat. Struct. Biol.* **1995**, *2* (11), 1007–1011.

(169) Gorin, G.; Wang, S. Assay of Lysozyme by Its Lytic on M . Lysodeikficus. **1971**, *27*, 113–127.

(170) 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, J.; Peralta, J. 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, J. M.; 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.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.; Gaussian, I. Gaussian 09, Revision E.01. Wallingford CT 2009.

(171) Yang, X.; Zhou, P. P.; Zhou, D. G.; Zheng, P. J.; Dai, Y.; Yan, C. X.; Yang, Z. Simultaneous Interactions of Amphoteric Halogen in XY (X=Cl, Br and Y=F, Cl, Br) with C and O Atoms of CO2 in Ring-Shaped CO2·X(Y)·CO2 Complexes. *Comput. Theor. Chem.* **2016**, *1076*, 32–41.

(172) Baase, W. A.; Liu, L.; Tronrud, D. E.; Matthews, B. W. Lessons from the Lysozyme of Phage T4. *Protein Sci.* **2010**, *19* (4), 631–641.

(173) Zimmermann, M.; Lange, A.; Boeckler, F. Validated Scoring of Halogen Bonding in Molecular Design. *J. Cheminform.* **2014**, *6* (Suppl 1), O17.

(174) Walter, S. M.; Kniep, F.; Herdtweck, E.; Huber, S. M. Halogen-Bond-Induced Activation of a Carbon-Heteroatom Bond. *Angew. Chemie - Int. Ed.* **2011**, *50* (31), 7187–7191.

(175) Kniep, F.; Rout, L.; Walter, S. M.; Bensch, H. K. V.; Jungbauer, S. H.; Herdtweck, E.;

Huber, S. M. 5-Iodo-1,2,3-Triazolium-Based Multidentate Halogen-Bond Donors as Activating Reagents. *Chem. Commun.* **2012**, *48* (74), 9299.

(176) Kniep, F.; Jungbauer, S. H.; Zhang, Q.; Walter, S. M.; Schindler, S.; Schnapperelle, I.; Herdtweck, E.; Huber, S. M. Organocatalysis by Neutral Multidentate Halogen-Bond Donors. *Angew. Chemie - Int. Ed.* **2013**, *52* (27), 7028–7032.

(177) Takeda, Y.; Hisakuni, D.; Lin, C. H.; Minakata, S. 2-Halogenoimidazolium Salt Catalyzed Aza-Diels-Alder Reaction Through Halogen-Bond Formation. *Org. Lett.* **2015**, *17* (2), 318–321.

(178) Bayse, C. A.; Rafferty, E. R. Is Halogen Bonding the Basis for Iodothyronine Deiodinase Activity? *Inorg. Chem.* **2010**, *49* (12), 5365–5367.

(179) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54* (10), 3451–3479.

(180) Brown, F. K.; Sherer, E. C.; Johnson, S. A.; Holloway, M. K.; Sherborne, B. S. The Evolution of Drug Design at Merck Research Laboratories. *J. Comput. Aided. Mol. Des.* **2016**, 1–12.

(181) Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Iii, C. F. B.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D. De Novo Computational Design of Retro-Aldol Enzymes. *Science* **2008**, *319* (5868), 1387–1391.

(182) Richter, F.; Leaver-Fay, A.; Khare, S. D.; Bjelic, S.; Baker, D. De Novo Enzyme Design Using Rosetta3. *PLoS One* **2011**, *6* (5), 1–12.

(183) Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.;

Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D. Kemp Elimination Catalysts by Computational Enzyme Design. *Nature* **2008**, *453* (7192), 190–195.

(184) Mukhopadhyay, A.; Redding, A. M.; Rutherford, B. J.; Keasling, J. D. Importance of Systems Biology in Engineering Microbes for Biofuel Production. *Curr. Opin. Biotechnol.* **2008**, *19*, 228–234.

(185) Jorgensen, W. L. Computer-Aided Discovery of Anti-HIV Agents. *Bioorg. Med. Chem.* **2016**, 1–11.

(186) Scholfield, M. R.; Ford, M. C.; Carlsson, A.-C. C.; Butta, H.; Mehl, R. A.; Ho, P. S. Structure-Energy Relationships of Halogen Bonds in Proteins. *Biochemistry* **2017**, acs.biochem.7b00022.

(187) El Hage, K.; Pandyarajan, V.; Phillips, N. B.; Smith, B. J.; Menting, J. G.; Whittaker, J.; Lawrence, M. C.; Meuwly, M.; Weiss, M. A. Extending Halogen-Based Medicinal Chemistry to Proteins: Iodo-Insulin as a Case Study. *J. Biol. Chem.* **2016**, No. 10, jbc.M116.761015.

(188) Soteras Gutiérrez, I.; Lin, F.-Y.; Vanommeslaeghe, K.; Lemkul, J. A.; Armacost, K. A.; Brooks, C. L.; MacKerell, A. D. Parametrization of Halogen Bonds in the CHARMM General Force Field: Improved Treatment of Ligand–protein Interactions. *Bioorg. Med. Chem.* **2016**.

(189) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12* (1), 281–296.

(190) Riley, K. E.; Hobza, P. The Relative Roles of Electrostatics and Dispersion in the Stabilization of Halogen Bonds. *Phys. Chem. Chem. Phys.* **2013**, *15* (41), 17742–17751.

(191) Bereau, T.; Kramer, C.; Meuwly, M. Leveraging Symmetries of Static Atomic Multipole Electrostatics in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9* (12), 5450–5459.

(192) Ponder, J. W.; Richards, F. An Efficient Newton-like Method for Molecular Mechanics Energy Minimization of Large Molecules. *J Comput Chem* **1987**, *8* (7), 1016–1024.

(193) Bayly, C. C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. a. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. ...* **1993**, *97*, 10269–10280.

(194) Wang, J.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **2000**, *21* (12), 1049–1074.

(195) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast , Efficient Generation of High-Quality Atomic Charges . AM1-BCC Model : II . Parameterization and Validation. *J Comput Chem* **2002**, *23* (16), 1623–1641.

(196) Case, D. A.; Betz, R. M.; Botello-Smith, W.; Cerutti, D. S.; Cheatham, T. E. I.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. AMBER2016. University of California, San Francisco 2016.

(197) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Genet.* **2006**, *65* (3), 712–725.

(198) Kim, D.-H.; Kim, M.-S. Hydrogenases for Biological Hydrogen Production. *Bioresour. Technol.* **2011**, *102* (18), 8423–8431.

(199) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (15), 5869–5874.

(200) Deller, M. C.; Kong, L.; Rupp, B. Protein Stability: A Crystallographer's Perspective. *Acta Crystallogr. Sect. Struct. Biol. Commun.* **2016**, *F* (72), 72–95.

(201) Srivastava, A.; Yano, J.; Hirozane, Y.; Kefala, G.; Gruswitz, F.; Snell, G.; Lane, W.; Ivetac, A.; Aertgeerts, K.; Nguyen, J.; Jennings, A.; Okada, K. High-Resolution Structure of the Human GPR40 Receptor Bound to Allosteric Agonist TAK-875. *Nature* **2014**, *513* (7516), 124–127.

(202) Hollenstein, K.; Kean, J.; Bortolato, A.; Cheng, R. K. Y.; Doré, A. S.; Jazayeri, A.; Cooke, R. M.; Weir, M.; Marshall, F. H. Structure of Class B GPCR Corticotropin-Releasing Factor Receptor 1. *Nature* **2013**, *499* (7459), 438–443.

(203) Giver, L.; Gershenson, A.; Freskgard, P.-O.; Arnold, F. H. Directed Evolution of a Thermostable Esterase. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 12809–12813.

(204) Wijma, H. J.; Floor, R. J.; Jekel, P. A.; Baker, D.; Marrink, S. J.; Janssen, D. B. Computationally Designed Libraries for Rapid Enzyme Stabilization. *Protein Eng. Des. Sel.* **2014**, *27* (2), 49–58.

(205) Funahashi, J.; Sugita, Y.; Kitao, A.; Yutani, K. How Can Free Energy Component Analysis Explain the Difference in Protein Stability Caused by Amino Acid Substitutions?

Effect of Three Hydrophobic Mutations at the 56th Residue on the Stability of Human Lysozyme. *Protein Eng.* **2003**, *16* (9), 665–671.

(206) Chipot, C.; Pohorille, A. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Chipot, C., Pohorille, A., Eds.; Springer, 2007; Vol. 86.

(207) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 585–590.

(208) Wong, C. F.; McCammon, J. A. Dynamics and Design of Enzymes and Inhibitors. *J. Am. Chem. Soc.* **1986**, *108*, 3830–3832.

(209) Kollman, P. Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93* (7), 2395–2417.

(210) Jorgensen, W.; Ravimohan, C. Monte Carlo Simulation of Differences in Free Energies of Hydration. *J Chem Phys* **1985**, *83* (6), 3050–3054.

(211) Dang, L. X.; Merz Jr, K. M.; Kollman, P. A. Free Energy Calculations on Protein Stability: Thr-157. -> Val-157 Mutation of T4 Lysozyme. *J. Am. Chem. Soc.* **1989**, *111* (22), 8505–8508.

(212) Pitera, J. W.; Kollman, P. a. Exhaustive Mutagenesis in Silico: Multicoordinate Free Energy Calculations on Proteins and Peptides. *Proteins* **2000**, *41* (3), 385–397.

(213) Goldfeld, D. a; Murphy, R.; Kim, B.; Wang, L.; Beuming, T.; Abel, R.; Friesner, R. a. Docking and Free Energy Perturbation Studies of Ligand Binding in the Kappa Opioid Receptor. *J. Phys. Chem. B* **2015**, *119* (3), 824–835.

(214) Seeliger, D.; de Groot, B. L. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophys. J.* **2010**, *98* (10), 2309–2316.

(215) Cappel, D.; Hall, M. L.; Lenselink, E. B.; Beuming, T.; Qi, J.; Bradner, J. E.; Sherman, W.

Relative Binding Free Energy Calculations Applied to Protein Homology Models. *J. Chem. Inf. Model.* **2016**, Article ASAP.

(216) Liu, P.; Kim, B.; Friesner, R. a; Berne, B. J. Replica Exchange with Solute Tempering: A Method for Sampling Biological Systems in Explicit Water. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (39), 13749–13754.

(217) Wang, L.; Berne, B. J.; Friesner, R. A. On Achieving High Accuracy and Reliability in the Calculation of Relative Protein–ligand Binding Affinities. *Proc. Natl. Acad. Sci. USA* **2012**, *109* (6), 1937–1942.

(218) Banks, J. A. Y. L.; Beard, H. S.; Cao, Y.; Cho, A. R. T. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; R, J. O. N.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Y, L.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **2005**, *26* (16), 1752–1780.

(219) Halgren, T. A. Merck Molecular Force Field. *J. Comput. Chem.* **1996**, *17* (5 & 6), 490–519.

(220) Lenselink, E. B.; Louvel, J.; Forti, A. F.; van Veldhoven, J. P. D.; de Vries, H.; Mulder-Krieger, T.; McRobb, F. M.; Negri, A.; Goose, J.; Abel, R.; van Vlijmen, H. W. T.; Wang, L.; Harder, E.; Sherman, W.; IJzerman, A. P.; Beuming, T. Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation. *ACS Omega* **2016**, *1* (2), 293–304.

(221) Ciordia, M.; Pérez-Benito; Delgado, F.; Trabanco, A. A.; Tresadern, G. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. *J. Chem. Inf. Model.* **2016**, *56*, 1856–1871.

(222) Sherborne, B.; Shanmugasundaram, V.; Cheng, A. C.; Christ, C. D.; DesJarlais, R. L.;

Duca, J. S.; Lewis, R. A.; Loughney, D. A.; Manas, E. S.; McGaughey, G. B.; Peishoff, C. E.; van Vlijmen, H. Collaborating to Improve the Use of Free-Energy and Other Quantitative Methods in Drug Discovery. *J. Comput. Aided. Mol. Des.* **2016**, *30* (12), 1139–1141.

(223) Steinbrecher, T.; Zhu, C.; Wang, L.; Abel, R.; Negron, C.; Pearlman, D.; Feyfant, E.; Duan, J.; Sherman, W. Predicting the Effect of Amino Acid Single Point Mutations on Protein Stability – Large Scale Validation of MD-Based Relative Free Energy Calculations. *J. Mol. Biol.* **2016**.

(224) Gapsys, V.; Michielssens, S.; Seeliger, D.; de Groot, B. L. Accurate and Rigorous Prediction of the Changes in Protein Free Energies in a Large-Scale Mutation Scan. *Angew. Chemie - Int. Ed.* **2016**, *55*, 7364–7368.

(225) Gromiha, M. M.; An, J.; Kono, H.; Oobatake, M.; Uedaira, H.; Sarai, A. ProTherm: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Res.* **1999**, *27* (1), 286–288.

(226) Jia, L.; Yarlagadda, R.; Reed, C. C. Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. *PLoS One* **2015**, *10* (9), 1–19.

(227) Molecular Operating Environment (MOE). Chemical Computing Group Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7 2016.

(228) Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment. San Diego: Dassault Systèmes.

(229) Beard, H.; Cholleti, A.; Pearlman, D.; Sherman, W.; Loving, K. A. Applying Physics-Based Scoring to Calculate Free Energies of Binding for Single Amino Acid Mutations in

Protein-Protein Complexes. *PLoS One* **2013**, *8* (12), 1–11.

(230) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of Absolute Solvation Free Energies Using Molecular Dynamics Free Energy Perturbation and the Opls Force Field. *J. Chem. Theory Comput.* **2010**, *6* (5), 1509–1519.

(231) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comput. Phys.* **1976**, *22* (2), 245–268.

(232) Becktel, W. J.; Schellman, J. A. Protein Stability Curves. *Biopolymers* **1987**, *26* (11), 1859–1877.

(233) Matthews, B. W. Studies on Protein Stability with T4 Lysozyme. In *Advances in Protein Chemistry*; 1995; pp 249–278.

(234) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157* (1), 105–132.

(235) Simonson, T. Free Energy of Particle Insertion. *Mol. Phys.* **1993**, *80* (2), 441–447.

(236) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations. *Chem. Phys. Lett.* **1994**, *222* (6), 529–539.

(237) Zacharias, M.; Straatsma, T. P.; McCammon, J. a. Separation-Shifted Scaling, a New Scaling Method for Lennard-Jones Interactions in Thermodynamic Integration. *J. Chem. ...* **1994**, *100* (12), 9025.

(238) Jorgensen, W. L.; Thomas, L. L. Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J Chem Theory Comput* **2008**, *4* (6), 869–876.

(239) Pearlman, D. a. A Comparison of Alternative Approaches to Free Energy Calculations. *J Phys Chem* **1994**, *98*, 1487–1493.

(240) Lim, N. M.; Wang, L.; Abel, R.; Mobley, D. L. Sensitivity in Binding Free Energies Due to Protein Reorganization. *J. Chem. Theory Comput.* **2016**, *12*, 4620–4631.

(241) Price, D. J.; Jorgensen, W. L. Improved Convergence of Binding Af Nities with Free Energy Perturbation: Application to Nonpeptide Ligands with pp60 SH2 Domain. *J. Comput. Aided. Mol. Des.* **2001**, *15*, 681–695.

(242) Giegé, R. A Historical Perspective on Protein Crystallization from 1840 to the Present Day. *FEBS J.* **2013**, *280* (24), 6456–6497.

(243) Hu, Y.; Sherborne, B.; Lee, T.-S.; Case, D. A.; York, D. M.; Guo, Z. The Importance of Protonation and Tautomerization in Relative Binding Affinity Prediction: A Comparison of AMBER TI and Schrödinger FEP. *J. Comput. Aided. Mol. Des.* **2016**, *30* (7), 533–539.

APPENDIX

SOURCE CODE

Supplemental Subroutine 1: Identifying the X-bond.

```
c
c
c     ##################################################
c     ##  COPYRIGHT (C)  1990  by  Jay William Ponder  ##
c     ##              All Rights Reserved              ##
c     ##################################################
c
c     ############################################################
c     ##                                        ##
c     ##  subroutine xbonds  --  locate and store bxb atoms  ##
c     ##                                        ##
c     ############################################################
c
c
c     "xbonds" finds the total number of halogen bonding atoms
c     and stores the atom numbers of the atoms defining each bxb
c
c
      subroutine xbonds
      implicit none
      include 'sizes.i'
      include 'action.i'
      include 'atoms.i'
      include 'atmtyp.i'
      include 'couple.i'
      include 'xbond.i'
      include 'potent.i'
      integer i,j,k,m,nint
c
c     loop over all atoms, finding the halogens
c     when halogen is found, an array is made (ixbnd) with index c      (nint,nexbond) nint is 1,2 1,3
interaction and nexbnd
c     is the number of xbond
c
      nexbnd = 0
      do i = 1, n
        nint = 1
        j=atomic(i)
        if((j.eq.17).or.(j.eq.35).or.(j.eq.53)) then
          nexbnd=nexbnd+1
          ixbnd(nint,nexbnd)=i
          if(n12(i).gt.1) then
            write(6,*) "oops ",i," bonded to more than one atom"
```

```
         end if
         xatmnm=j
         do k = 1, n12(i)
            nint=nint+1
            ixbnd(nint,nexbnd)=i12(k,i)
         enddo
         do k = 1, n13(i)
            nint=nint+1
            ixbnd(nint,nexbnd)=i13(k,i)
         enddo
       end if
     end do

c
c    set use_xbond logical to true if xbond present
c
     use_xbond=.false.
     if (nexbnd.ne.0) then
       use_xbond=.true.
     end if
     return
     end
```

Supplemental Subroutine 2. Assigning the X-bond parameters.

```
c
c
c    ##################################################
c    ##  COPYRIGHT (C)  1990  by  Jay William Ponder  ##
c    ##              All Rights Reserved              ##
c    ##################################################
c
c    ############################################################
c    ##                                        ##
c    ##  subroutine kxbond  --  assign xbond parameters    ##
c    ##                                        ##
c    ############################################################
c
c
c    "kxbond" assigns parameters of ffbxb depending on halogen
c
c
     subroutine kxbond
     implicit none
     include 'sizes.i'
     include 'atmtyp.i'
     include 'atoms.i'
     include 'xbond.i'
     include 'charge.i'
     integer i,j
```

```
c
c
c
      do i = 1, nexbnd
        if(xatmnm.eq.17) then
           xbdr=0.1487
           xba=0.117
        endif
        if(xatmnm.eq.35) then
           xbdr=0.160
           xba=0.191
        endif
        if(xatmnm.eq.53) then
           xbdr=0.185
           xba=0.33
        endif
      enddo
      return
      end
```

Supplemental Subroutine 3. Calculating the X-bond energies and forces.

```
c
c
c     ################################################
c     ##  COPYRIGHT (C)  1990  by  Jay William Ponder  ##
c     ##              All Rights Reserved              ##
c     ################################################
c
c     ##############################################################
c     ##                                          ##
c     ##  subroutine exbond1  --  xbond energy & derivatives   ##
c     ##                                          ##
c     ##############################################################
c
c
c     "exbond1" calculates the xbond energy and
c     first derivatives with respect to Cartesian coordinates
c
c
      subroutine exbond1
      implicit none
      include 'sizes.i'
      include 'atoms.i'
      include 'xbond.i'
      include 'vdw.i'
      include 'charge.i'
      include 'deriv.i'
      include 'energi.i'
```

```
      include 'group.i'
      include 'chgpot.i'
      include 'usage.i'
      include 'virial.i'
      include 'couple.i'
      integer i,ia,ib,ic,iaclass, o, p
      integer icvdw,mvdw,mion,icion
      integer k,l,j,jc,mm,break
      integer*4 ict(maxatm),icclass,id(3)
      real*8 delr,qafac,qcfac,qfac,rad,eps,qa
      real*8 a(3),amod,b(3),c(3),d(3),dmod,bmod,abdot
      real*8 da(3),db(3),damod,dbmod,dcostheta,dcos2theta
      real*8 dexvdw,dexcc,dlocal(3,3)
      real*8 ordexvdx,ordexvdy,ordexvdz
      real*8 ordexcdx,ordexcdy,ordexcdz
      real*8 costheta,cos2theta
      real*8 rvdw,p12,orp12
      real*8 p6,oridec,oridev,scale14
      real*8 qc,ees,oriees,rb,r,fik,f,fi
      real*8 expterm,bde,fgrp
      logical proceed
c
c
c     zero out the xbond energy and first derivatives
c
      orev = 0.0d0
      orec = 0.0d0
      exv = 0.0d0
      exc = 0.0d0
      do i = 1, n
        dexv(1,i) = 0.0d0
        dexv(2,i) = 0.0d0
        dexv(3,i) = 0.0d0
        dexc(1,i) = 0.0d0
        dexc(2,i) = 0.0d0
        dexc(3,i) = 0.0d0
        ordev(1,i) = 0.0d0
        ordev(2,i) = 0.0d0
        ordev(3,i) = 0.0d0
        ordec(1,i) = 0.0d0
        ordec(2,i) = 0.0d0
        ordec(3,i) = 0.0d0
      end do
c
c     calculate the xbond energy term
c
c     ia = xbond center (center of the angle)
c     ib = center bonded to the xbond center
c     delr = distance displacement factor
c     qafac = A charge factor
c     qcfac = C charge factor
```

```
c
      do i = 1, nexbnd
        ia = ixbnd(1,i)
        ib = ixbnd(2,i)
        id(1) = ib
        id(2) = ia
        iaclass = jvdw(ia)
        delr = xbdr
        qafac = xba
        qa = pchg(ia)
        qcfac = (3.2174*pchg(ia)+0.1512)
        a(1) = x(ia) - x(ib)
        a(2) = y(ia) - y(ib)
        a(3) = z(ia) - z(ib)
        amod = dsqrt(a(1)*a(1) + a(2)*a(2) + a(3)*a(3))
c
c     loop over all atoms for non-bond terms
c       van der Waals first
c       generate a list that doesn't have 1,2 1,3 terms
c       need separate variables to keep consistent with loops of elj1
c
        mvdw = 0
        do icvdw = 1,nvdw
          break = 0
          ic = ivdw(icvdw)
          do k = 1, 10
            if ((ic.eq.ixbnd(k,i))) then
               break = 1
            endif
          enddo
          if (break.ne.1) then
             mvdw = mvdw+1
             ict(mvdw) = ic
          endif
        enddo
        do icvdw = 1,mvdw
          ic = ivdw(icvdw)
          ic = ict(icvdw)
          if (ic.gt.ia) then
             o = ia
             p = ic
          elseif (ic.lt.ia) then
             o = ic
             p = ia
          endif
          icclass = jvdw(ic)
          id(3) = ic
          rad = radmin(iaclass,icclass)
          eps = epsilon(iaclass,icclass)
          scale14 = 1.0d0
c
```

```
c     need to loop over atoms and scale 1,4 interactions
c
          do j = 1, n14(ia)
             if ((i14(j,ia).eq.ic)) then
                scale14 = 2.0d0
             endif
          enddo
          b(1) = x(ia) - x(ic)
          b(2) = y(ia) - y(ic)
          b(3) = z(ia) - z(ic)
          c(1) = x(ib) - x(ic)
          c(2) = y(ib) - y(ic)
          c(3) = z(ib) - z(ic)
          d(1) = x(o) - x(p)
          d(2) = y(o) - y(p)
          d(3) = z(o) - z(p)
          bmod = dsqrt(b(1)*b(1) + b(2)*b(2) + b(3)*b(3))
          abdot = a(1)*b(1) + a(2)*b(2) + a(3)*b(3)
          costheta = abdot/(amod*bmod)
          cos2theta = (2.0d0*(costheta**2))-1.0d0
          rvdw = rad - delr*cos2theta
          p12 = ((rvdw**12/bmod**12)*eps)/scale14
          orp12 = ((rad**12/bmod**12)*eps)/scale14
          p6 = (2.0d0*eps*(rad**6/bmod**6))/scale14
          exv = exv + p12
          oridev = ((orp12)*(-12.0d0/bmod))/bmod
          orev = orev - orp12
          ordexvdx = d(1)*oridev
          ordexvdy = d(2)*oridev
          ordexvdz = d(3)*oridev
c
c      loop over center(in id) & coordinate to generate derivatives
c      zero derivative array
c
       do l = 1,3
          do j = 1,3
             dlocal(j,l) = 0.0d0
          enddo
       enddo
       do l = 1,3 ! center
          mm = id(l)
          do j = 1,3 ! coordinate
             da(1) = 0.0d0
             da(2) = 0.0d0
             da(3) = 0.0d0
             if(mm.eq.ia) da(j) = 1.0d0
             if(mm.eq.ib) da(j) = -1.0d0
             damod = (a(1)*da(1)+a(2)*da(2)+a(3)*da(3))/amod
             db(1) = 0.0d0
             db(2) = 0.0d0
             db(3) = 0.0d0
```

```
            if(mm.eq.ia) db(j) = 1.0d0
            if(mm.eq.ic) db(j) = -1.0d0
            dbmod = (b(1)*db(1)+b(2)*db(2)+b(3)*db(3))/bmod
            dcostheta = (-((abdot)*(amod*dbmod+bmod*damod))/
     &        ((amod*bmod)**2))+((da(1)*b(1)+da(2)*b(2)+da(3)*b(3))
     &        +(a(1)*db(1)+a(2)*db(2)+a(3)*db(3)))/(bmod*amod)
            dcos2theta = 4*costheta*dcostheta
            dexvdw=((-12.0d0*eps*(rvdw/bmod)**11)
     &        *(((delr*dcos2theta)/bmod)+((rvdw/(bmod**2))
     &        *dbmod)))/scale14
            dlocal(j,l) = dlocal(j,l)+dexvdw
          enddo
        enddo
c
c     check original derivative array, note that o is i and p is k in elj1.f
c
          ordev(1,o) = ordev(1,o) - ordexvdx
          ordev(2,o) = ordev(2,o) - ordexvdy
          ordev(3,o) = ordev(3,o) - ordexvdz
          ordev(1,p) = ordev(1,p) + ordexvdx
          ordev(2,p) = ordev(2,p) + ordexvdy
          ordev(3,p) = ordev(3,p) + ordexvdz
c
c     fill derivative array
c
          dexv(1,ia) = dexv(1,ia) + dlocal(1,2)
          dexv(2,ia) = dexv(2,ia) + dlocal(2,2)
          dexv(3,ia) = dexv(3,ia) + dlocal(3,2)
          dexv(1,ib) = dexv(1,ib) + dlocal(1,1)
          dexv(2,ib) = dexv(2,ib) + dlocal(2,1)
          dexv(3,ib) = dexv(3,ib) + dlocal(3,1)
          dexv(1,ic) = dexv(1,ic) + dlocal(1,3)
          dexv(2,ic) = dexv(2,ic) + dlocal(2,3)
          dexv(3,ic) = dexv(3,ic) + dlocal(3,3)
c
c     generate virial contributions
c
          vir(1,1) = vir(1,1)+(b(1)*dlocal(1,1)+c(1)*dlocal(1,2))
     &                -(d(1)*ordexvdx)
          vir(2,1) = vir(2,1)+(b(2)*dlocal(1,1)+c(2)*dlocal(1,2))
     &                -(d(2)*ordexvdx)
          vir(3,1) = vir(3,1)+(b(3)*dlocal(1,1)+c(3)*dlocal(1,2))
     &                -(d(3)*ordexvdx)
          vir(1,2) = vir(1,2)+(b(2)*dlocal(1,1)+c(2)*dlocal(1,2))
     &                -(d(2)*ordexvdx)
          vir(2,2) = vir(2,2)+(b(2)*dlocal(2,1)+c(2)*dlocal(2,2))
     &                -(d(2)*ordexvdy)
          vir(3,2) = vir(3,2)+(b(3)*dlocal(2,1)+c(3)*dlocal(2,2))
     &                -(d(3)*ordexvdy)
          vir(1,3) = vir(1,3)+(b(3)*dlocal(1,1)+c(3)*dlocal(1,2))
     &                -(d(3)*ordexvdx)
```

```
      vir(2,3) = vir(2,3)+(b(3)*dlocal(2,1)+c(3)*dlocal(2,2))
     &           -(d(3)*ordexvdy)
      vir(3,3) = vir(3,3)+(b(3)*dlocal(3,1)+c(3)*dlocal(3,2))
     &           -(d(3)*ordexvdz)
    enddo
c
c     loop over electrostatics
c     generate a list that doesn't have 1,2 1,3 terms
c
    mion = 0
    do icion = 1,nion
      break = 0
      ic = ivdw(icion)
      do k = 1, 10
        if ((ic.eq.ixbnd(k,i))) then
          break = 1
        endif
      enddo
      if (break.ne.1) then
        mion = mion+1
        ict(mion) = ic
      endif
    enddo
    do icion = 1,mion
      ic = ict(icion)
      if (ic.gt.ia) then
        o = ia
        p = ic
      elseif (ic.lt.ia) then
        o = ic
        p = ia
      endif
      id(3) = ic
      qc = pchg(ic)
      scale14 = 1.0d0
      do j = 1, n14(ia)
        if ((i14(j,ia).eq.ic)) then
          scale14 = 1.2d0
        endif
      enddo
      b(1) = x(ia) - x(ic)
      b(2) = y(ia) - y(ic)
      b(3) = z(ia) - z(ic)
      c(1) = x(ib) - x(ic)
      c(2) = y(ib) - y(ic)
      c(3) = z(ib) - z(ic)
      d(1) = x(o) - x(p)
      d(2) = y(o) - y(p)
      d(3) = z(o) - z(p)
      bmod = dsqrt(b(1)*b(1) + b(2)*b(2) + b(3)*b(3))
      abdot = a(1)*b(1) + a(2)*b(2) + a(3)*b(3)
```

```
          costheta = abdot/(amod*bmod)
          cos2theta = (2.0d0*(costheta**2))-1.0d0
          qfac = (qafac*cos2theta) + qcfac
          ees = ((qfac*qc*electric)/(dielec*(bmod**2+ebuffer)))
     &          /scale14
          f =electric/dielec
          fi = f*pchg(ic)
          fik = fi*qa
          r = bmod
          rb = bmod+ebuffer
          oriees= (fik/rb)/scale14
          orec = orec - oriees
          exc = exc + ees
          oridec = -(fik/rb**2/r)/scale14
          ordexcdx = d(1)*oridec
          ordexcdy = d(2)*oridec
          ordexcdz = d(3)*oridec
c
c         loop over center(in id) & coordinate to generate derivates
c         zero derivative array
c
        do l = 1,3
          do j = 1,3
            dlocal(j,l) = 0.0d0
          enddo
        enddo
        do l = 1,3 ! center
          mm = id(l)
          do j = 1,3 ! coordinate
            da(1) = 0.0d0
            da(2) = 0.0d0
            da(3) = 0.0d0
            if (mm.eq.ia) da(j) = 1.0d0
            if (mm.eq.ib) da(j) = -1.0d0
            damod=(a(1)*da(1)+a(2)*da(2)+a(3)*da(3))/amod
            db(1) = 0.0d0
            db(2) = 0.0d0
            db(3) = 0.0d0
            if (mm.eq.ia) db(j) = 1.0d0
            if (mm.eq.ic) db(j) = -1.0d0
            dbmod = (b(1)*db(1)+b(2)*db(2)+b(3)*db(3))/bmod
            dcostheta = (-1.0d0*((abdot)*(amod*dbmod+bmod*damod))/
     &        ((amod*bmod)**2))+((da(1)*b(1)+da(2)*b(2)+da(3)*b(3))
     &        +(a(1)*db(1)+a(2)*db(2)+a(3)*db(3)))/(bmod*amod)
            dcos2theta = 4*costheta*dcostheta
            dexcc = (((qafac*qc*dcos2theta*electric)/
     &        (dielec*(bmod**2)))-((2.0d0*qfac*qc
     &        *electric*dbmod)/(dielec*(bmod**3))))/scale14
c           print *, "this is dexcc", dexcc
            dlocal(j,l) = dlocal(j,l) + dexcc
c           print *, "this is dlocal for cc", dlocal
```

```
          enddo
        enddo
c
c     check original derivative array, note that o is i and p is k in echarge1.f
c
        ordec(1,o) = ordec(1,o) - ordexcdx
        ordec(2,o) = ordec(2,o) - ordexcdy
        ordec(3,o) = ordec(3,o) - ordexcdz
        ordec(1,p) = ordec(1,p) + ordexcdx
        ordec(2,p) = ordec(2,p) + ordexcdy
        ordec(3,p) = ordec(3,p) + ordexcdz
c
c     fill derivative array
c
        dexc(1,ia) = dexc(1,ia) + dlocal(1,2)
        dexc(2,ia) = dexc(2,ia) + dlocal(2,2)
        dexc(3,ia) = dexc(3,ia) + dlocal(3,2)
        dexc(1,ib) = dexc(1,ib) + dlocal(1,1)
        dexc(2,ib) = dexc(2,ib) + dlocal(2,1)
        dexc(3,ib) = dexc(3,ib) + dlocal(3,1)
        dexc(1,ic) = dexc(1,ic) + dlocal(1,3)
        dexc(2,ic) = dexc(2,ic) + dlocal(2,3)
        dexc(3,ic) = dexc(3,ic) + dlocal(3,3)
c
c     generate virial contributions
c
        vir(1,1)=vir(1,1)+(b(1)*dlocal(1,1)+c(1)*dlocal(1,2))
     &             -(d(1)*ordexvdx)
        vir(2,1)=vir(2,1)+(b(2)*dlocal(1,1)+c(2)*dlocal(1,2))
     &             -(d(2)*ordexvdx)
        vir(3,1)=vir(3,1)+(b(3)*dlocal(1,1)+c(3)*dlocal(1,2))
     &             -(d(3)*ordexvdx)
        vir(1,2)=vir(1,2)+(b(2)*dlocal(1,1)+c(2)*dlocal(1,2))
     &             -(d(2)*ordexvdx)
        vir(2,2)=vir(2,2)+(b(2)*dlocal(2,1)+c(2)*dlocal(2,2))
     &             -(d(2)*ordexvdy)
        vir(3,2)=vir(3,2)+(b(3)*dlocal(2,1)+c(3)*dlocal(2,2))
     &             -(d(3)*ordexvdy)
        vir(1,3)=vir(1,3)+(b(3)*dlocal(1,1)+c(3)*dlocal(1,2))
     &             -(d(3)*ordexvdx)
        vir(2,3)=vir(2,3)+(b(3)*dlocal(2,1)+c(3)*dlocal(2,2))
     &             -(d(3)*ordexvdy)
        vir(3,3)=vir(3,3)+(b(3)*dlocal(3,1)+c(3)*dlocal(3,2))
     &             -(d(3)*ordexvdz)
      enddo
     enddo
     return
     end
```